Bioinformatics Analyses of Genomic Imprinting

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von
Barbara Hutter

Saarbrücken
2009

Tag des Kolloquiums:	08.12.2009
Dekan:	Prof. Dr.-Ing. Stefan Diebels
Berichterstatter:	Prof. Dr. Volkhard Helms
	Priv.-Doz. Dr. Martina Paulsen
Vorsitz:	Prof. Dr. Jörn Walter
Akad. Mitarbeiter:	Dr. Tihamér Geyer

Table of contents

Index of figures

Index of tables

Index of listings

Summary

In the present thesis, bioinformatics analyses of genomic DNA sequences identified a number of features that distinguish imprinted genes from normal, biallelically expressed genes. Despite species-specific differences, which particularly complicate identification of functional CpG islands, imprinted genes of human and mouse are enriched in intronic CpG islands and tandem repeats. Together with conserved LINE-1 repeats they might be involved in the establishment of the allele-specific marks in the germ line. Striking in comparison to non-imprinted genes is also the enrichment of CpG-rich motifs as well as a decreased estimated deamination ratio in conserved sequences, which hints at unanticipated effects of differential methylation. Genome-wide analyses showed that highly conserved elements in exons of imprinted genes are less conserved and shorter than those of normal genes. Maternally expressed genes and the proteins encoded by them are more divergent between rodents and other mammals, whereas paternally expressed genes are conserved above average between mouse and rat. The associated opposite patterns of selection suggest that imprinted genes played a role in the evolution of early rodents. The existence of conserved paralogs with similar functions may have facilitated divergence.

Zusammenfassung

In der vorliegenden Arbeit wurde durch bioinformatische Untersuchungen von genomischen DNS-Sequenzen eine Reihe von Merkmalen bestimmt, die elterlich geprägte Gene gegenüber normalen, biallelisch exprimierten Genen auszeichnen. Trotz artenspezifischer Unterschiede, die insbesondere die Identifizierung von funktionalen CpG-Inseln erschweren, besitzen geprägte Gene in Mensch und Maus vermehrt intronische CpG-Inseln und Tandemrepeats. Zusammen mit konservierten LINE-1-Repeats könnten diese zur Einrichtung der allelspezifischen Markierungen in der Keimbahn beitragen. Auffällig im Vergleich zu nicht geprägten Genen sind auch die Anreicherung von CpG-reichen Motiven und eine erniedrigte geschätzte Desaminierungsrate in konservierten Sequenzabschnitten, was auf unvorhergesehene Effekte differentieller Methylierung schließen lässt. Genomweite Analysen ergaben, dass hochkonservierte Elemente in Exons bei geprägten Genen weniger konserviert und kürzer sind als bei normalen Genen. Maternal exprimierte Gene und von ihnen codierte Proteine zeigen erhöhte Divergenz zwischen Nagetieren und anderen Säugetieren, wohingegen paternal exprimierte Gene zwischen Maus und Ratte einen überdurchschnittlich hohen Konservierungsgrad aufweisen. Die damit verbundenen entgegengesetzten Selektionsmuster lassen darauf schließen, dass geprägte Gene eine Rolle in der Evolution früher Nagetiere spielten. Möglicherweise erleichterte die Existenz von konservierten Paralogen mit ähnlicher Funktion die Divergenz.

Acknowledgements

First of all I want to thank my parents and my brother PD Dr. Michael Hutter for always encouraging me with my studies.

I thank my supervisors Professor Dr. Volkhard Helms and PD Dr. Martina Paulsen who gave me the opportunity to work on this interdisciplinary project. They made a successful grant proposal possible and constantly provided me with stimulating ideas and suggestions.

Furthermore, I am grateful to the two Bioinformatics students who assisted me with the project. I appreciate the help of Siba Ismael with the identification of orthologous genes. Matthias Bieg has been indispensable to programming the UCSC database.

I want to thank the whole Computational Biology group, current as well as former members and associated students, for talks, cake, occasional help with programming, and providing such a pleasant working atmosphere.

For valuable discussions I also thank the Epigenetics group, namely Professor Dr. Jörn Walter, and Dr. Christoph Bock and M.Sci. Lars Feuerbach from the Max Planck Institute for Computer Science.

I highly appreciate the work of numerous sequencing centers that made the genomic DNA sequences used in this study publicly available and I am much obliged to the NCBI, Ensembl, and UCSC bioinformatics teams for providing their data and answering questions.

Finally, I thank the Saarland University and the Deutsche Forschungsgemeinschaft (PA 750/3-1) for funding.

"Imagination is more important than knowledge. For knowledge is limited …"

("What Life Means to Einstein", The Saturday Evening Post, 26. Oktober 1929)

Abbreviations

AS	Angelman Syndrome
bp	base pair
BWS	Beckwith-Wiedemann Syndrome
cDNA	complementary DNA (mature mRNA reverse transcribed to DNA)
CGI	CpG island
chr	chromosome
CpA	dinucleotide CA
CpG	dinucleotide CG
CpG_{obs}/CpG_{exp}	ratio of observed CpGs to expected CpGs
CTCF	CCCTC binding factor (transcription factor, insulator protein)
dbSNP	database of single nucleotide polymorphisms
DMR	differentially methylated region
DNMT	DNA methyltransferase
ECR	evolutionary conserved region
GGF	CpG island criteria of Gardiner-Garden and Frommer (1997)
H_0	null hypothesis
H_A	alternative hypothesis
IAP	intracisternal A particle
IC	imprinting center
ICR	imprinting control region
ID	identity
IQR	interquartile range
JSD	Jensen-Shannon divergence
Ka	rate of nonsynonymous substitutions
kb	kilobase (1000 bp)
Ks	rate of synonymous (silent) substitutions
L1	LINE-1 subfamily of long interspersed nuclear elements
L2	LINE-2 subfamily of long interspersed nuclear elements
LINE	long interspersed nuclear element
LTR	long terminal repeat
MAR	matrix attachment region
MER	medium reiterated frequency repeat
MIR	mammalian-wide interspersed repeat
mPCS	mammalian-specific *phastCons* sequence
Mya	million years ago
NCBI	National Center for Bioinformatics Technology
ORegAnno	Open Regulatory Annotation database
p	p value (error probability)
PCS	*phastCons* most conserved sequence
pers. comm.	personal communication
PSSM	position-specific scoring matrix
PWS	Prader-Willi Syndrome
SINE	short interspersed nuclear element
SNP	single nucleotide polymorphism
std.dev.	standard deviation
TFBS	transcription factor binding site
TJ	CpG island criteria of Takai and Jones (2002)
TpG	dinucleotide TG
TSS	transcriptional start site
UCSC	University of Santa Cruz, California
YY1	Yin-Yang 1 (transcription factor)

Space for notes

Chapter 1 - Introduction

Genomic imprinting is a special epigenetic mechanism of gene regulation in mammals and flowering plants. In contrast to the vast majority of genes that are biallelically expressed, i.e. from the alleles of both chromosomes, imprinted genes are monoallelically expressed depending on whether they were inherited from the mother or from the father (Fig. 1.1). The Imprinted Gene Catalogue at the University of Otago[1] (Morison et al. 2001, Glaser et al. 2006) and the Mouse Imprinting Website at the Mammalian Research Center Harwell[2] provide records of imprinted genes identified in human and mouse. Since 2005, marking the beginning of the studies presented here, when there were about 40, their lists have been slowly but steadily growing to approximately 90 as of June 2009. It is estimated that a few hundred genes may be subject to imprinting (Reik and Walter 2001, Morison et al. 2005, Luedi et al. 2005, 2007). As, due to their monoallelic expression, the alleles of imprinted genes are quasi dominant, any disturbance on the expressed allele immediately shows its consequences. Mutations in imprinted genes or their regulatory elements, which cause either over- or underexpression of the genes, result in severe growth anomalies, organ malfunctions, behavior anomalies, and cancer. Therefore, they are of particular interest for research on human diseases.

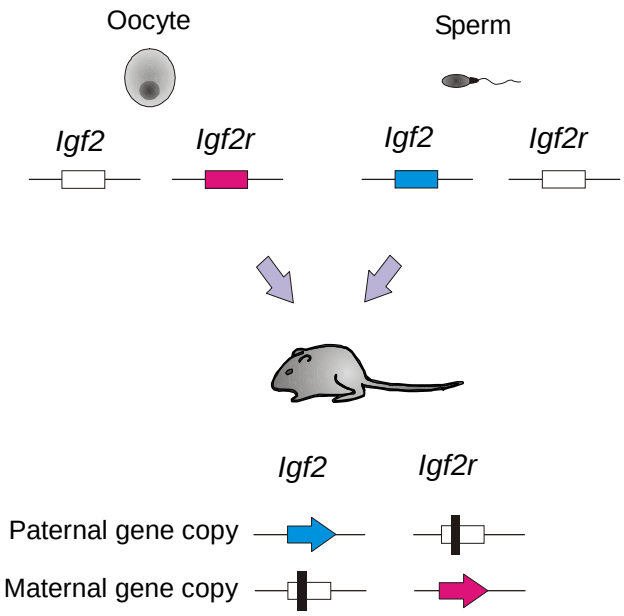


Figure 1.1: Imprinting of *Igf2* and *Igf2r* in the mouse
In mice, the insulin like growth factor gene (*Igf2*), which encodes the IGF2[3] protein, is only transcribed from the chromosome transmitted by the father. In contrast, the insulin like growth factor receptor gene (*Igf2r*), coding for the IGF2R protein, is only expressed from the chromosome inherited from the mother. The figure was kindly provided by M. Paulsen.

[1] http://igc.otago.ac.nz/home.html
[2] http://www.har.mrc.ac.uk/research/genomic_imprinting/
[3] The nomenclature in this thesis follows the recommendations of the Mouse Genome Informatics Nomenclature Committee (http://www.informatics.jax.org/mgihome/nomen/gene.shtml): Gene symbols are always italicized. Mouse gene symbols begin with an uppercase letter followed by all lowercase letters whereas human ones are all uppercase. Protein symbols correspond to gene symbols using all uppercase letters and are not italicized.

After beginning with a short overview of the most important terms and concepts related to imprinting, this chapter will give more detailed explanations in the following sections. The specific patterns of regulation are presented, including the establishment and maintenance of imprints, and the genomic organization of imprinted regions. In view of the roles performed by their protein products, functional implications and the evolution of these genes are discussed. Special emphasis is put on what bioinformatics research could reveal about the features that distinguish imprinted genes from normal, biallelically expressed genes. Our own contributions, which are presented in chapters 3 and 4, are shortly referred to in the corresponding sections.

1.1 Important terms and concepts related to genomic imprinting

Epigenetics describes inheritance patterns that, as the Greek prefix "epi" implies, are "on top of" the DNA sequence. While genetic information is encoded in the DNA sequence, epigenetic information consists of DNA methylation and histone modifications. Thus, without changing the nucleotide sequence, these epigenetic modifications influence gene expression and can be transmitted to the next generation, thereby representing a mechanism of long-term gene regulation. In the special case of genomic imprinting, differential marking of paternal and maternal chromosomes – the imprint – results in the repression of gene copies depending on their parental origin.

Paternal and maternal alleles are distinguished by different epigenetic modifications of the DNA, such as methylation of cytosines followed by guanines in CpG dinucleotides (Bestor and Tycko 1996). 5-methylcytosine is sometimes referred to as "the fifth base" of the DNA. So-called differentially methylated regions (DMRs, see section 1.3) are highly methylated (hypermethylated) on one chromosome but more or less unmethylated (hypomethylated) on the other (Umlauf et al. 2004, Kobayashi et al. 2006). These DMRs often overlap with CpG islands (CGIs, see section 1.2). CpG islands are enriched in CpG dinucleotides that are otherwise rare in mammalian genomes. They were found to be frequently associated with promoter regions of biallelically expressed genes (Bird 1986, Larsen et al. 1992). Normally, CpG islands are unmethylated; if one happens to become methylated, though, the chromatin structure of its associated promoter region is thought to become dense, hindering the access of the transcription machinery and causing silencing of the associated gene (Bird 2002; see also Fig. 1.7). At differentially methylated CpG islands, i.e. DMRs, transcription can only be initiated on the unmethylated allele of the two chromosomes. Methylated regions are established by the *de novo* DNA methyltransferases DNMT3A and DNMT3B and maintained by DNMT1 which, during DNA replication, transfers methyl groups onto CpG cytosines on the newly synthesized strand.

DMRs constitute central regulatory regions at imprinted loci around which most of the imprinted genes are clustered (section 1.3). The influence of a DMR can extend over several thousands of base pairs so that deleting it causes loss of the correct expression patterns in the affected region. Therefore, such regions are also referred to as imprinting control regions (ICRs) or imprinting centers (ICs).

Another important factor involved in epigenetics are repetitive elements, shortly called repeats. As the name suggests, repetitive elements are nucleotide patterns that occur multiple times in the genome. In the case of tandem repeats, these patterns are repeated directly following each other either as perfect copies or with slight variations. Repeats attract methylation, which might interfere with the establishment and maintenance of DMRs. Thus, it is not surprising that the vicinity of

imprinted genes shows a special distribution of different repeats (section 1.6). Since none of the above mentioned sequence features alone is sufficient for the establishment of a DMR, a combination of different factors appears to be necessary, such as transcription in the oocyte, histone modifications, and a specific pattern of CpG spacing in the DMR sequences (Chotalia et al. 2009).

The main task of proteins encoded by imprinted genes seems to provide nutrients to the developing embryo. Many paternally expressed genes encode growth factors. Maternally expressed genes include antagonistic growth-suppressing functions as well as ion channels and transporters. There is also a number of imprinted genes that do not encode proteins but regulatory RNAs (section 1.7). According to the parental conflict hypothesis (Moore and Haig 1991, see section 1.8), genes from the father "want" to become their offspring as large as possible whereas the maternal ones try to save the mother's resources for further offspring, probably fathered by different males.

To elaborate on strategies for identifying imprinted genes is beyond the focus of this introduction. Most research is done on the mouse. It involves the generation of parthenogenetic and androgenetic embryos (Nikaido et al. 2003), uniparental disomies, chromosomal translocations, and reciprocal crosses of different strains with single nucleotide polymorphisms to determine which parent the transcribed allele stems from (Babak et al. 2008, Wang et al. 2008). For human, mainly pedigrees and linkage analysis are used since research on human material is a problematic matter. Nowadays, bioinformatics approaches help to narrow down the search space for promising candidates (Oakey and Beechey 2002, Luedi et al. 2005, 2007, Ruf et al. 2007). Genome-wide screens for methylation (Smith et al. 2003) and histone modifications (Wen et al. 2008, see also section 1.5) can also facilitate finding new imprinted genes.

1.2 CpG islands as regulatory elements

The following section is mainly taken from the introduction of Hutter et al. (2009), which recapitulates what is known and hypothesized about CpG islands (CGIs) according to the literature.

In mammalian genomes the CpG dinucleotide is depleted towards 20-25% of the frequency expected by the G+C content (Lander et al. 2001, Waterston et al. 2002). The cytosines of CpG dinucleotides are usually methylated and 5-methylcytosine can easily deaminate to thymine (Fig. 1.2) so that, if this mutation is not repaired, the affected CpG is permanently converted to TpG, or CpA on the complementary DNA strand (Bird 1980, Bestor and Tycko 1996, Jones et al. 1998). Thus, 5-methylcytosines represent mutational hot spots that can cause diseases (Bestor and Tycko 1996). If such mutations occur in the germ line, they become heritable. A constant loss of CpGs over thousands of generations can explain the scarcity of this special dinucleotide.

Nevertheless, some genomic regions maintain a high CpG content close to the frequency of other dinucleotides. These so-called CpG islands (CGIs) are believed to escape methylation at least in the germ lines. CpG islands were originally described as *HpaII* tiny fragments (Bird 1986), i.e. CpG-rich sequences cut by the methylation-sensitive restriction enzyme *HpaII*. Showing frequently absence of DNA methylation, and presence of histone modifications that are characteristic for an open chromatin structure, CGIs have a commonly acknowledged potential to act as regulatory elements.

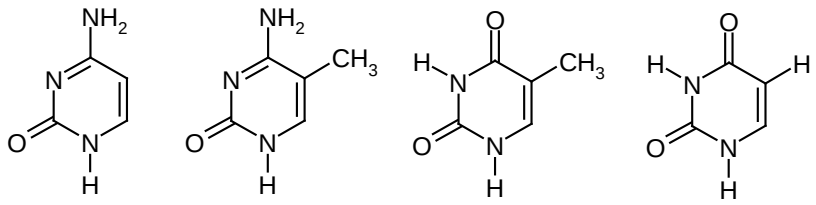


Figure 1.2: Pyrimidine nucleobases
From left to right: structures of cytosine, 5-methylcytosine, thymine, and uracil. Hydrolytic deamination converts methylated cytosines into thymines. In contrast to the deamination of unmethylated cytosines into uracil, the RNA counterpart of thymine, this mutation is not easily recognized and removed by DNA repair enzymes.

Based on the nucleotide composition of *HpaII* tiny fragments, Gardiner-Garden and Frommer (1987) introduced the original criteria for the computational identification of CGI sequences. About half of the sequences identified as CGIs with these parameters in the human genome coincide with repetitive elements (Lander et al. 2001). Being normally methylated and transcriptionally silenced (Jones et al. 1998, Meissner et al. 2008), such CGIs do not obey the original definition as an unmethylated sequence providing an open chromatin structure (Bird 1986, Tazi and Bird 1990). Therefore, more stringent parameters were developed that have nowadays been widely adopted for the identification of CGIs in genomic sequences (Takai and Jones 2002).

Having a CGI in the promoter region was first believed to be a feature limited to housekeeping genes (Larsen et al. 1992). Ponger and coworkers (2001) found that most genes that are expressed in the early embryo have promoter CGIs as well and concluded that transcription prevents them from being methylated. This is consistent with the assumption that CGIs allow binding of ubiquitous transcription factors, thereby facilitating expression of the corresponding gene (Bird 1986, Cross et al. 2000, Hannenhalli and Levy 2001, Antequera 2003). More precisely, genes with a single transcriptional start site (TSS) in their CGI-associated promoters are predominantly involved in basic cellular functions whereas those with several TSSs exhibit tissue-specific expression (Carninci et al. 2006, Saxonov et al. 2006, Baek et al. 2007). Although G+C and CpG contents are generally increased in the vicinity of transcriptional start sites (Yamashita et al. 2005), promoters with a CGI can be clearly distinguished from promoters without a CGI (Saxonov et al. 2006). As at least half of all human genes possess CGIs in their promoter regions (Gardiner-Garden and Frommer 1987, Larsen et al. 1992, Cross et al. 2000, Ponger et al. 2001), these elements are successfully being used for detecting transcriptional start sites in genomes (Ioshikhes and Zhang 2000, Hannenhalli and Levy 2001). CGIs have also been hypothesized to coincide with origins of replication (Antequera and Bird 1999).

Since it was discovered that CGIs can be methylated in some cases, they have come of increasing interest in epigenetic research. Tumor suppressor genes are silenced in cancer cells by *de novo* methylation of their promoter CGIs (Robertson and Wolffe 2000, Jones and Baylin 2002). Somatic methylation of CGIs has also been observed in normal tissues (Strichman-Almashanu et al. 2002, Yamada et al. 2004, Song et al. 2005). Similarly, methylation plays a role in X chromosome deactivation (Hellman and Chess 2007). CGIs that are not located at promoters but at intronic or intergenic locations were shown to function as regulatory elements, e.g. in imprinted genes (Reik and Walter 2001). Moreover, recent bioinformatics approaches have revived the notion of CGIs as transcriptionally active sites, finding differences between somatically methylated and

unmethylated CGIs in the DNA structure (Bock et al. 2006). Thus, computationally identified CGIs can be classified into CGIs with open chromatin features and false positive ones that are probably associated with heterochromatin (Bock et al. 2007).

Despite their important roles, there is only limited experimental data available for detection of transcriptionally active CGIs so that especially for non-human species one is restricted to mere sequence criteria. For the mouse, however, which serves as model organism in molecular biology, the commonly used parameters designed for use in human sequences (Takai and Jones 2002), may be too strict for identifying important functional CGIs. In the mouse genome, CpG is even more depleted than in the human genome (Zhao and Zhang 2006a, 2006b). This difference can be explained to some extent by the insertion of CpG-rich, primate-specific repetitive elements (i.e. *Alu* repeats) into the human genome. Additionally, in comparison to the human genome, the mouse genome exhibits an elevated accumulation of C to T transitions and single nucleotide substitutions in general (Waterston et al. 2002). These species-specific patterns of sequence conservation appear to be influenced by various factors such as differences in recombination rates, the shorter generation times in rodents, and weight-specific metabolic rates resulting in increased oxidative DNA damage and elevated DNA replication rates (Hwang and Green 2004). As a consequence, rodent CGIs are supposed to undergo a faster erosion due to the loss of CpGs that is reflected in the lower number of CGIs identified in the mouse genome (Aïssani and Bernardi 1991, Antequera and Bird 1993, Matsuo et al. 1993, Cuadrado et al. 2001, Jiang et al. 2007). Antequera and Bird (1993) estimated that the mouse genome lacks about 20% of the human CGIs. Nevertheless, out of the 27,000 CGIs identified in the human genome (Lander et al. 2001) and 15,500 in the mouse, approximately 10,000 have been found to be significantly conserved with respect to sequence between human and mouse (Waterston et al. 2002). A substantial part of the remaining ones may be structural analogs since for orthologous genes, the presence or the absence of a promoter CGI, respectively, is highly correlated (Yamashita et al. 2005). CpG-rich promoters are characterized by a lower conservation than CpG-poor ones and have been described as plastic and fast-evolving (Carninci et al. 2006, Baek et al. 2007), possibly because of the rather unspecific binding of transcription factors (Antequera 2003).

Coming back to the issue of genomic imprinting, CpG islands are of special interest because most DMRs overlap with CGIs. The hypothesis that imprinted genes possess more CGIs than biallelically expressed genes (Reik and Walter 2001, Paulsen et al. 2000, Paulsen and Ferguson-Smith 2001) stimulated further research and initiated the work presented in this thesis. Various strategies for the identification of CGI candidate sequences by bioinformatics methods are explained in chapter 2.2. Sections 3.1 and 3.2 treat the association of imprinted and biallelically expressed genes in human and mouse with CGIs identified by different computational criteria. Their performance is evaluated and advice on their use is given in section 4.2. Although the original hypothesis was invalidated (Ke et al. 2002a, 2002b, Allen et al. 2003, Hutter et al. 2006), our analyses confirmed that intergenic CGIs, which were often shown to give rise to alternative or antisense transcripts, are key feature of imprinted genes. Moreover, CGIs in imprinted regions are enriched in tandem repeats (Hutter et al. 2006).

1.3 Differentially methylated regions and imprinting clusters

CpG-rich regions at imprinted loci that show differential methylation are the key regulatory elements that convey the parent-of-origin dependent monoallelic expression of these genes. The specific DNA methylation patterns that differ between maternal and paternal chromosomes are established during germ cell development and maintained after fertilization (Tucker et al. 1996, Olek and Walter 1997, Hajkova et al. 2002). This is peculiar since the genome of the early embryo is subject to global demethylation. More precisely, the paternal genome is actively demethylated shortly after fertilization, when the parental genomes are still separated, whereas the maternal genome undergoes passive demethylation due to exclusion of DNMT1. Afterwards, there is a wave of methylation introduced by the de novo methyltransferases DNMT3A and DNMT3B. In contrast, parental imprints are resistant to this epigenetic reprogramming; they are only erased and re-set in primordial sperm cells and maturing oocytes to reflect the transmitting sex (Reik and Walter 2001, Morgan et al. 2005; Fig. 1.3). Little is known about the protein complexes involved in establishment of the associated DMRs. They require DNMT3A and a methyltransferase related protein, DNMT3L, which is thought to recruit and direct the methyltransferases (Bourc'his and Bestor 2004, Kaneda et al. 2004).

Interestingly, transcription causes the intragenic chromatin structure to open up so that methyltransferases can access the DNA. This might be responsible for the paradox situation of heavy methylation inside highly transcribed genes whereas the promoters are shielded by transcription factors and remain unmethylated (Jones 1999, Hellman and Chess 2007). DMRs might be established by transcription in oocytes, which often use alternative promoters upstream of the somatic ones (Chotalia et al. 2009). Of the known primary DMRs in mouse, 17 are set in oocytes and only three in sperm cells (Chotalia et al. 2009). Some secondary DMRs are established later, after fertilization, through chromatin interactions (Murrell et al. 2004). By knockout experiments on the three paternally derived primary DMRs it has been shown that DNMT3A alone is sufficient for methylation at the *Igf2*/*H19* and *Dlk1*/*Gtl2* loci whereas both DNMT3A and DNMT3B are required for establishment of the *Rasgrf1* DMR (Kato et al. 2007). DNMT3L is indispensable at either locus.

Since methylated CpGs are prone to deaminate to TpG, germ line DMRs are expected to lose CpGs. Indeed, compared to maternally derived primary DMRs, the paternal ones are CpG-depleted (Kobayashi et al. 2006, Bourc'his and Bestor 2006). This is attributed to the fact that imprints are set much earlier in male germ cells (before meiosis, around birth, Davis et al. 2000) than in oocytes (after meiotic recombination, just prior to ovulation, Lucifero et al. 2004) so that the methylated CpGs have a higher probability to deaminate (Bourc'his and Bestor 2006). Differential methylation could affect whole imprinted regions, most likely by leading to increased CpG deamination. Taking measures for CpG deamination into account (see chapter 3), we observed that in conserved and protein-coding regions, there seems to be no prevalent loss due to deamination but rather retention or even enrichment of CpG.

As already mentioned, imprinted genes are predominantly found in clusters around DMRs, which regulate the correct expression of genes over distances up to several kb. Isolated imprinted genes are usually associated with a DMR of their own, even if it is very distant. So far, imprinting clusters have been identified on several chromosomes (Fig. 1.4). There is no evident pattern of direction or expression of the individual genes inside these clusters. Similarly eluding a common scheme, DMRs are found at different locations: in promoter regions (*Commd1*, *Cdkn1c*, *Gtl2*, *H19*,

Ndn, *Nnat*, *Peg10/Scge*, *Plagl1*, *Snrpn*, *Impact*), at alternative intragenic promoters *(Gnas* locus, *Grb10*, *Igf2*, *Mest*), and in introns where they give rise to antisense or downstream transcripts (*Kcnq1*, *Igf2r*, *Peg3*, *Dlk1*). Whereas all paternally methylated DMRs reside between genes (*H19*/*Igf2*, *Rasgrf1*/*A19*, *Dlk1*/*Gtl2*), the maternally methylated primary ones are active promoters on the paternal allele (Wood and Oakey 2006).

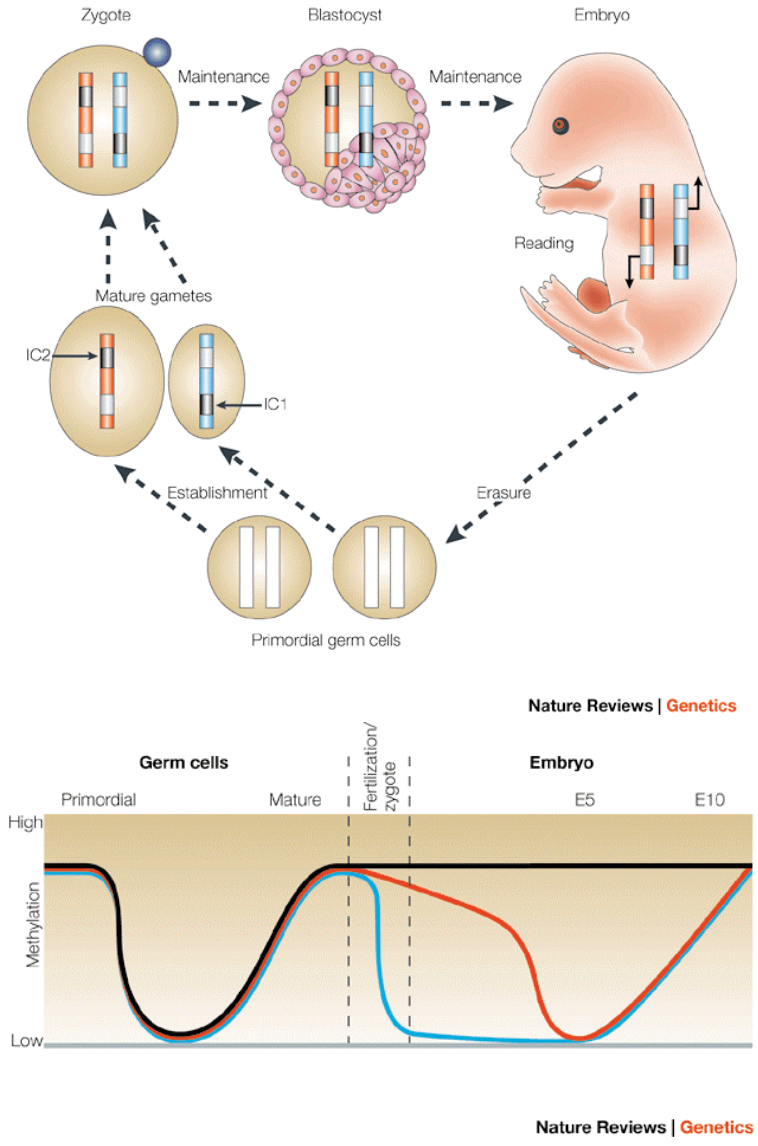


Figure 1.3: DNA methylation maintenance, erasure and establishment
The simplified depiction taken from Reik and Walter (2001) shows the cycle of DNA methylation at imprinted loci as an overview (top) and in comparison to the rest of the genome (bottom). Methylation at the DMRs of imprinted loci is altered at different time intervals than in the genome (black). Blue represents paternal and red maternal chromosomes or alleles, respectively. Differential methylation is reset in the germ cells according to the sex of the developing embryo, which will be the parent of the next generation. In its somatic cells, the existing imprints determine the expression of the associated genes. The DMR at the IC1 imprinting center is paternally methylated, that at IC2 is maternally methylated (shown by black marks).

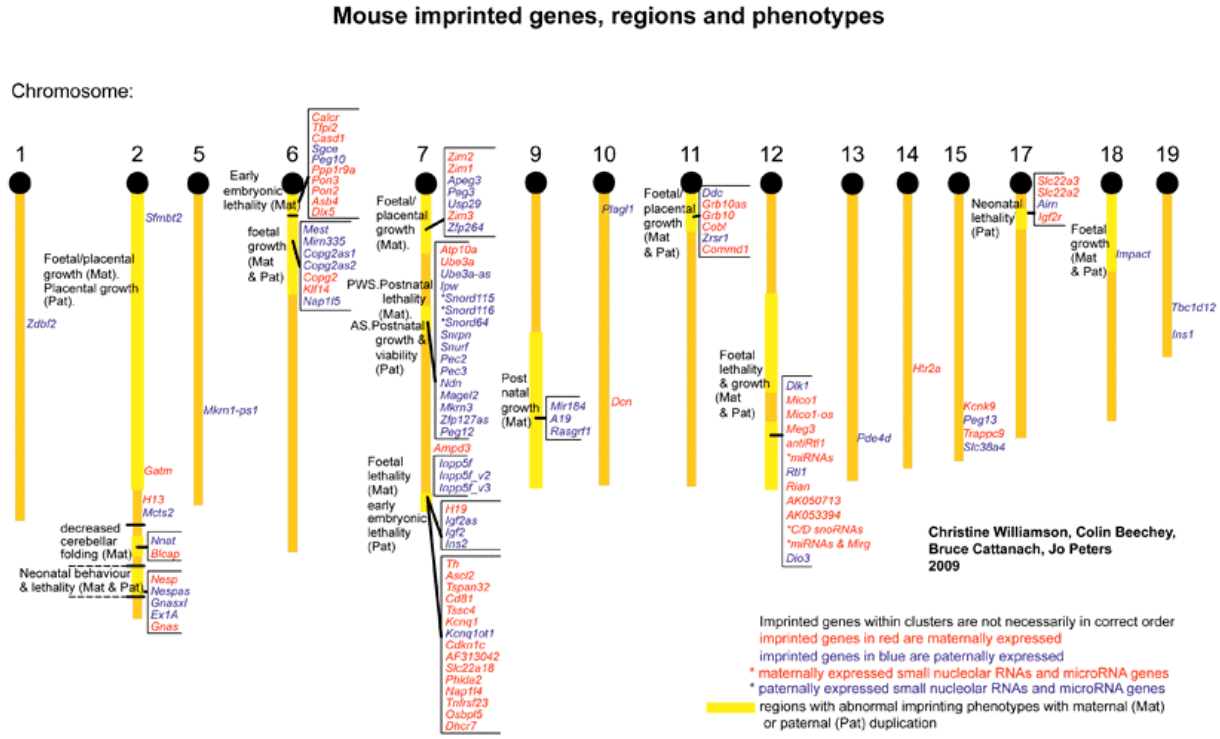


Figure 1.4: Harwell[4] imprinting map of the mouse
Most imprinted genes reside in clusters around their control centers, the differentially methylated regions. They are unevenly distributed throughout the genome. Mouse chromosome 7 harbors most of the imprinted genes known in this species in three separated clusters. Their orthologs are found on human chromosomes 11, 15, and 19. The genes studied in the present thesis are listed in Appendix B Tab. B1 and Appendix D Tab. D3.

1.4 Reading the imprint

After the imprint – the DMR – has been set in the germ line, it is converted into differential gene expression in somatic cells, which is, however, not an easy on-off mechanism. Morison et al. (2005) suggest to use the terms maternally or paternally *repressed*, respectively, because silencing of one allele may be only partial. In the case of "leaky" imprinting, which may affect a lot of yet undetected imprinted genes, expression does not always takes place exclusively but rather predominantly from one of the two alleles. Expression can additionally require the presence of tissue-specific transcription factors. Some genes possess several alternative promoters of which not all are subject to imprinting. Thus, a gene may only be imprinted in certain tissues but biallelically expressed in others (*Igf2*, Moore et al. 1997), or even switch the allele (*Grb10*, Hikichi et al. 2003, Sanz et al. 2008; *Gnas* locus, Coombes et al. 2003). Consistent with these facts, the most upstream promoter regions of imprinted genes do not show an enrichment of special sequence patterns and exhibit similar conservation profiles as biallelically expressed genes (section 3.3).

Moreover, intragenic CpG islands can act as promoters of antisense transcripts which disturb the expression of neighboring genes in imprinting clusters (Pauler et al. 2007), notably the

[4] http://www.har.mrc.ac.uk/research/genomic_imprinting/

untranslated *Kcnq1ot1* that controls the Beckwith-Wiedemann Syndrome region (Engemann et al. 2000, Paulsen et al. 2000, Mancini-DiNardo et al. 2003). The presence of an intronic DMR in the human *IGF2R* does not make the gene imprinted, probably because of the lack of an antisense transcript originating at this CpG island as in mouse (Smrzka et al. 1995, Wutz and Barlow 1998). Antisense transcripts, however, are not peculiar to imprinting but seem to be quite common in mammalian genomes (Lehner et al. 2002, Kiyosawa et al. 2003, Yelin et al. 2003, Lavorgna et al. 2004, Chen et al. 2005, Zhang et al. 2006). More recently, the repressing function of numerous small noncoding RNAs in imprinted regions is being investigated (reviewed in Peters and Robson 2008, Royo and Cavaillé 2008). They play important roles in gene regulation by RNA interference, predominantly post-transcriptional but also directly by inducing DNA methylation (He and Hannon 2004).

To further complicate the matter, there is also competition between imprinted genes. On the maternal allele, the unmethylated imprinting control center of the *Igf2-H19* region is bound by the methylation-sensitive transcription factor CTCF (CCCTC binding factor). CTCF inhibits the interaction of the *Igf2* promoter with the enhancers downstream of *H19* (Bell and Felsenfeld 2000, Hark et al. 2000). Consequently, *Igf2* is silenced and *H19* is active. On the paternal allele, this pattern is reversed. The process involves chromatin loops inside which the thereby inactivated genes cannot be accessed by the transcription machinery (Murrell et al. 2004, Kurukuti et al. 2006; Fig. 1.5).

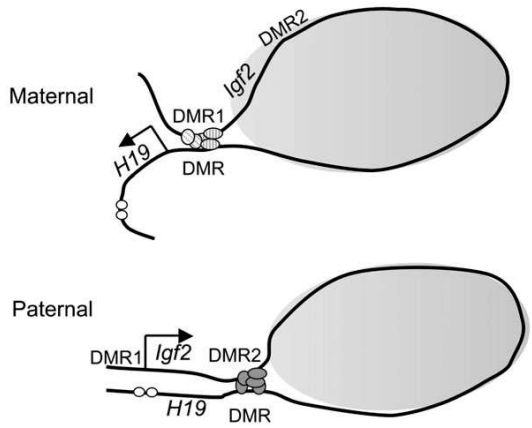


Figure 1.5: Chromatin loop model for the *Igf2-H19* locus
Differing chromatin organization on the maternal and paternal chromosome can explain imprinted expression of *Igf2* and *H19* in the mouse (Murrell et al. 2004, from where the figure is taken). The two genes are separated by approximately 70 kb. On the maternal chromosome, CTCF proteins are bound to the unmethylated DMR. They interact with other proteins like cohesin (Rubio et al. 2008) and the secondary, also unmethylated DMR1. This results in shutting off *Igf2* in a loop where it cannot be accessed by the RNA polymerase II complex that transcribes protein-coding genes and regulatory RNAs. In contrast, *H19* can interact with the enhancers and be expressed at a high level. On the paternal chromosome, the methylated DMR does not bind CTCF. Instead, it interacts with another secondary, methylated DMR2. Thus, the *Igf2* promoter is brought into contact with the enhancers, activating its expression whereas *H19* is silenced with its promoter hypermethylated. More recent research suggests that, on the maternal allele, instead of one large loop there are two tight ones, divided shortly after *Ifg2* (Kurukuti et al. 2006).

It is assumed that at the other two paternally methylated DMRs, pairs of protein-coding genes and regulatory RNAs compete for an enhancer like at the *Igf2*/*H19* locus (*Dlk1*/*Gtl2*, *Rasgrf1*/*A19*, Wood and Oakey 2006). *Rasgrf1* and *A19* are, however, both paternally expressed (de la Puente et al. 2002). CTCF binding sites have been identified here as well as at several other imprinted loci (Paulsen et al. 2001, Hikichi et al. 2003, Yoon et al. 2005, Fitzpatrick et al. 2007) and it is conceivable that they induce similar chromatin loops. As dynamic epigenetic elements guided by protein interactions, they can easily change in different developmental stages and cell types and thus lead to altered expression patterns not only of imprinted genes.

Loop formation may be responsible for the silencing of genes with an unmethylated promoter CpG island within imprinting clusters (*Dlk1*, *Ascl2*, *Klf14*, Parker-Katiraee et al. 2007; *Ppp1r9a*, *Asb4*, *Calcr*, Monk et al. 2008). So far, CTCF is the only known insulator protein. As the name implies, it establishes an epigenetic boundary between adjacent genomic regions. This may, as mentioned above, occur by separation of genes and enhancers into different chromatin loops (Yusufzai et al. 2004). Interactions of CTCF with mostly yet unidentified proteins may explain its varying role as insulator, enhancer blocker, repressor and activator of transcription (Ohlsson et al. 2001). CTCF can multimerize (Yusufzai et al. 2004), interact with the also methylation-sensitive transcription factor Yin-Yang 1 (YY1; Donohoe et al. 2007), and directly recruit the largest subunit of RNA polymerase II (Chernukhin et al. 2007). CTCF depletion in mouse oocytes results in the misregulation of zygotic gene expression, including *Gtl2*, *Grb10*, *Slc22a18*, and *Phlda2*, with subsequent apoptosis (Wan et al. 2008). Therefore, the authors suggested that *Ctcf* is a maternal effect gene, although the functions of its protein are not limited to imprinting.

1.5 Chromatin marks at imprinted regions

Chromosome structure is greatly influenced by the organization of histones, the proteins around which the DNA is wrapped. A considerable number of different modifications of histone tails have been described mostly at lysine residues. In general, methylated and deacetylated tails induce tight packing associated with transcriptional repression whereas acetylation enables expression (Fig. 1.6). Some methyl-CpG binding proteins and methyltransferase DNMT1 complex with histone deacetylases (Reik and Walter 2001). The resulting dense chromatin packing may not only affect transcription but also initiation of DNA replication during the S phase of the cell cycle. DNA in imprinted regions replicates in an asynchronous manner, the paternal allele before maternal one, at *Igf2r*, *Igf2*/*H19* and *Snrpn* (see references in Paulsen and Ferguson-Smith 2001).

Primary DMRs have been shown to present allele-specific histone modifications: H3K4me3 on the unmethylated DMR, H3K9me3 on the other one (Mikkelsen et al. 2007b, Parker-Katiraee et al. 2007, Meissner et al. 2008, Wen et al. 2008). Consequently, imprinted regions exhibit bivalent chromatin marks in genome-wide histone analyses. The roles of other histone modifications are less clear. H3K27me3 is associated with silencing (Mikkelsen et al. 2007b, Barski et al. 2007) and found on the inactive maternal allele of *Rasgrf1*, excluding DNA methylation (Lindroth et al. 2008). It is also present in the unmethylated paternal promoter region of *Grb10* in the absence of expression but reduced during neural cell development, concomitant with induction of transcription (Sanz et al. 2008). Repressed paternal alleles at the *Kcnq1* domain show H3K27me3 as well (Monk et al. 2006, Lewis et al. 2006). Hence, this modification seems to silence at least part of the imprinted genes that do not possess a promoter DMR. It may be established by YY1 which recruits

the histone H3K27me3 methyltransferase complex including the polycomb-group protein EED and the zinc finger protein SUZ12, which is a component of the polycomb repressive complex 2 (Mager et al. 2003, Ferguson-Smith and Reik 2003, Kim et al. 2003, Kim et al. 2006, Kim J et al. 2007, Kim 2008).

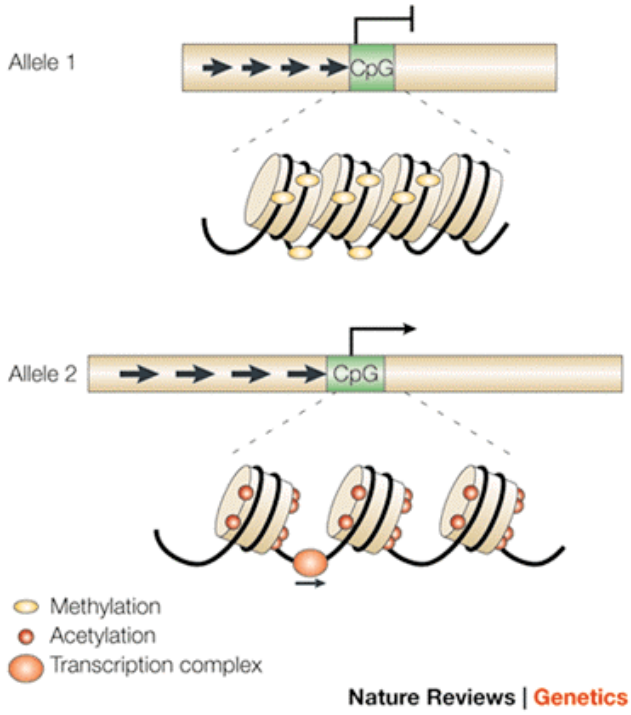


Figure 1.6: Putative chromatin structure at differentially methylated promoters
Methylation of the cytosine of CpG dinucleotides in promoter regions (Allele 1) is associated with methylation and deacetylation of histones, a dense chromatin structure and transcriptional silencing. At unmethylated promoters (Allele 2), histones are acetylated, chromatin structure is loose and transcription can take place. The arrows depict repeats that are a characteristic feature of DMRs. CpG stands for CpG island. The figure is taken from Reik and Walter 2001.

Another mediator of higher order chromatin architecture are matrix attachment regions (MARs). They have the potential to form heterochromatin and sequester genomic regions onto the nuclear matrix where they are inaccessible for transcription. On the other hand, they are frequently associated with enhancers. Conserved MARs have been identified at *Zfp127* (Greally et al. 1999) as well as at the *Igf2*/*H19* and *Dkl1*/*Gtl2* loci where they interact with the DMR in a tissue-specific manner (Kurukuti et al. 2006, Braem et al. 2008). Enhancer blocking functions of CTCF seem to be related to matrix attachment via interactions with nucleophosmin and other proteins with roles in subnuclear architecture (Yusufzai et al. 2004), thus combining chromatin loops and MARs into a joint mechanism for gene inactivation (Kurukuti et al. 2006). Such or similar complexes bound to the nucleolar surface might also stop the spreading of methylation by excluding methyltransferases, consistent with the role of CTCF as a boundary element at imprinted regions.

Lastly, human imprinted regions are enriched in recombination hot-spots where chromosomes cross over during meiosis (Reik and Walter 2001, Sandovici et al. 2006, Luedi et al. 2007). Special

DNA structures during recombination might attract methylation (Bestor and Tycko 1996) or transfer methylation (Bird 2002). Thus, interactions between the homologous regions on two chromosomes may contribute to the establishment of DMRs. Since CTCF interacts with cohesin, which holds together sister chromatids, it is possible that it might act on an intra- and interchromosomal level (Rubio et al. 2008). Furthermore, germ line specific proteins might be involved in chromatin interactions and protecting DMRs from methylation. Candidates are the transcription factors BORIS (Brother of the Regulator of Imprinted Sites), which stands in for its paralog CTCF in the male germ line and is therefore also known as CTCF-like protein (Loukinov et al. 2002, Hore et al. 2008), and SP1 that is ubiquitous at active CpG island promoters (Macleod et al. 1994, Brandeis et al. 1994). Overrepresented CpG-rich motifs and an enrichment of CTCF binding sites in intronic and intergenic regions at imprinted loci confirm the special role of this protein in imprinting (chapter 3.3). Ongoing research on chromatin structure, which is still in its infancy, will reveal valuable insight into the special protein-DNA and protein-protein interactions at imprinted loci, especially in the germ lines.

1.6 Roles of repetitive elements

Most repetitive elements belong to the category of so-called interspersed repeats that occur in a dispersed fashion. They are virus-derived sequences that became integrated into the genome and there developed into mobile elements. RNA interference seems to be responsible for their transcriptional silencing. On the other hand, transposable elements can be recruited for gene regulation and even give rise to genes (Jordan et al. 2003, Oei et al. 2004, Lowe et al. 2007, Slotkin and Martienssen 2007). Short interspersed transposable elements (SINEs) depend on long ones (LINEs) for transposition via an RNA intermediate. The broader term "repeat" includes low complexity regions with a biased base composition, e.g. polypurine or AT-rich, and simple repeats. The latter are short tandem repeats or microsatellites with motifs of 1-6 nucleotides. Long terminal repeats (LTRs) are retrotransposons that contain tandem repeats.

Repeats convey a high mutational potential. Apart from transcriptional interference and insertion events, recombination between homologous repetitive elements can cause translocations and other rearrangements (Yoder et al. 1997). It is debated whether CpG methylation was invented by evolution as a protection against the transcriptional activity of interspersed repeats or for gene regulation in general (Bestor and Tycko 1996, Yoder et al. 1997, Suzuki and Bird 2008). Anyhow, repetitive sequences are usually heavily methylated. Similar to centromeric repeats, tandem repeats are thought to attract methyltransferases by assuming an unusual structure (Bestor and Tycko 1996). In *Arabidopsis thaliana*, they are methylated by means of RNA interference (Zilberman et al. 2007), which is assumed to be the case also for animals (Martienssen 2003). Tandem repeat arrangements can also attract DNA methylation in meiotic processes in filamentous fungi (Malagnac et al. 1997). Although a similar connection could not yet be established for mammals, tandem repeats are likely to be involved in various epigenetic silencing and heterochromatin formation processes (Volpe et al. 2002, Saveliev et al. 2003). Curiously, arrays of tandem repeat motifs are frequently found in DMRs and throughout imprinted regions (Neumann et al. 1995, Ke et al. 2002a, Walter et al. 2006, Khatib et al. 2007; Appendix B Tab. B5), leading to the tandem repeat hypothesis of imprinting (Neumann et al. 1995). Some are necessary for correct differential methylation (*Rasgrf1*) but at other loci, they are dispensable (Lewis et al. 2004). Analogous DMRs

contain mostly divergent tandem repeats, only a few possess conserved motifs that occur in different numbers and at variable locations (Paulsen et al. 2001, Kim et al. 2003, Lewis et al. 2004, Paulsen et al. 2005, Khatib et al. 2007; Fig. 1.7). Our systematic investigations revealed that compared to biallelically expressed genes, there are significantly more imprinted genes that possess at least one CpG island that is associated with a tandem repeat (Hutter et al. 2006).

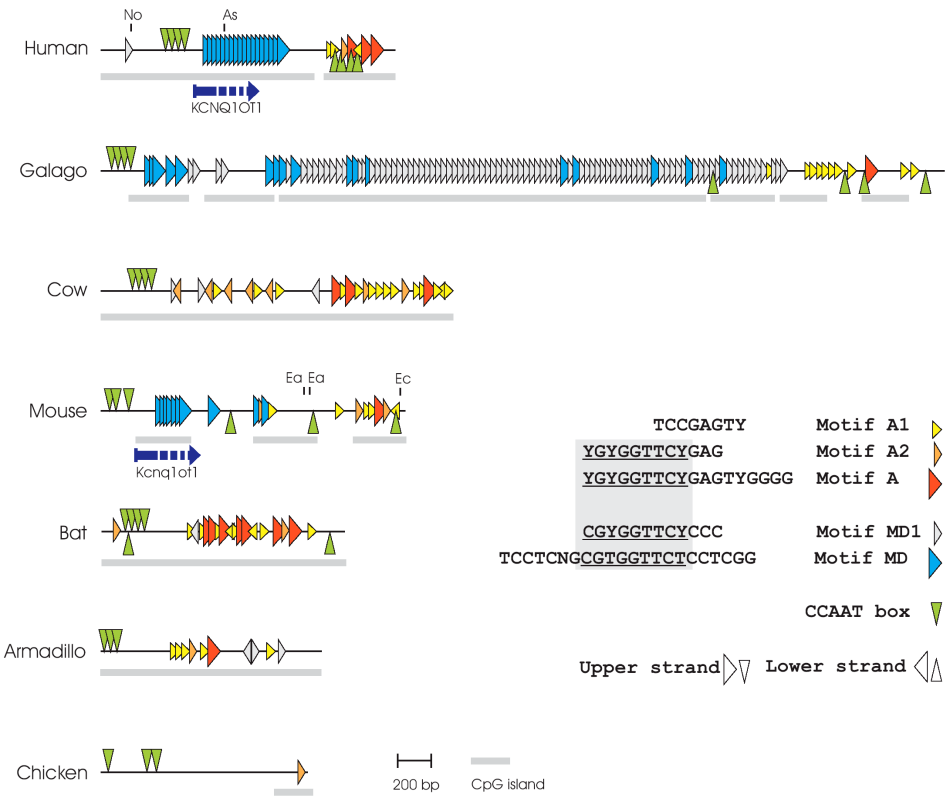


Figure 1.7: Repetitive DNA elements at IC2
The figure from Paulsen et al. 2005 shows the differentially methylated region of IC2, which is located in intron 10 of human *KCNQ1* and mouse *Kcnq1*. It acts as a promoter for the *KCNQ1OT1* (*Kcnq1ot1*) antisense transcript. CpG islands, CCAAT boxes, and various highly repetitive DNA elements were identified by Paulsen and coworkers (2005) in the human and mouse IC2 sequences as well as in homologous regions of four additional mammals. Motif MD had been previously reported by Mancini-DiNardo et al. (2003). The chicken sequence lacks tandem repeats and only contains one small CpG island, indicating that both features are important for imprinted expression.

Other repetitive elements show a particular behavior as well. SINEs are reduced in the vicinity of human and mouse imprinted genes whereas LINEs, especially from the L1 subfamily, LTRs, simple repeats, and low complexity regions occur more frequently (Greally 2002, Ke et al. 2002a, 2002b, Allen et al. 2003, Walter et al. 2006, Khatib et al. 2007). In combination with other sequence features, the distinguishing distribution of repetitive elements has been used for prediction of putative imprinted genes in the mouse (Luedi et al. 2005) and human genomes (Luedi et al. 2007). Our findings are in line with these observations and argue for a conservation of intronic and intergenic LINEs in imprinted regions (Hutter et al. 2006, chapter 3 sections 2 and 3).

SINE methylation seems to interfere severely with differential methylation so that there is most likely purifying selection against these elements in imprinted regions. In contrast, the mechanism for other repeats, namely L1, might have gained regulatory functions in the context of imprinting, possibly for spreading methylation from DMRs like on the X chromosome (Lyon 2006; see also chapter 3.3). Orthologous regions in platypus contain fewer LTRs and DNA elements and more SINEs than eutherian imprinting clusters, indicating that the distribution of repetitive elements is indeed linked to the evolution of imprinting (Warren et al. 2008, Pask et al. 2009; Fig. 1.10).

Interestingly, L1 repeats and retroviruses are inherited in a hypermethylated state on the paternal allele, but are hypomethylated on the maternal one whereas *Alu* elements, which are the most abundant SINEs in the human genome, behave exactly the other way round (Howlett and Reik 1991, Hellmann-Blumberg et al. 1993, Rubin et al. 1994, Chesnokov and Schmid 1995). Although this scenario reminds of imprinted genes, the methylation status of repeats is not maintained. After fertilization, L1 repeats are demethylated and *de novo* methylated during embryogenesis like all other interspersed elements (Yoder et al. 1997, Walter et al. 2006). The LTRs of murine Intracisternal A Particle (IAP) elements maintain most of the methylation acquired in both sperm and oocytes (Lane et al. 2003). Knockout studies revealed that different combinations of DNA methyltransferases are responsible for methylation of different repeat classes. DNMT3A target satellite repeats, DNMT3B act on B1, the murine *Alu* homolog, and both are required for correct methylation of L1 and IAP (Kato et al. 2007).

1.7 Functional implications of imprinted genes

Since several known human diseases have been linked to imprinting disorders on certain chromosomes, the corresponding imprinting clusters are named after them. Proteins encoded by imprinted genes take part in many pathways and interactions, including regulatory cascades and metabolic pathways (Grandjean et al. 2000, Arima et al. 2005, Varrault et al. 2006; Fig. 1.8). Notably, there are many transcription factors that have the potential to regulate other genes. Most imprinted genes are connected with growth regulation[5]. Others are brain-specifically imprinted or highly expressed in the brain (Tierling et al. 2006, Freed et al. 2008). Misregulation of imprinted genes at other loci is known to lead to neuronal defects, e.g. in Angelman Syndrome (AS) and Prader-Willi syndrome (PWS). PWS results from maternal disomy of chromosome 15 or microdeletions on the paternal one that result in silencing of paternally expressed genes, namely *SNPRN*, which encodes a small nuclear ribonucleoprotein, as well as *NDN*, a gene coding for a neuronal growth suppressor, and various noncoding RNAs (Nicholls et al. 1998, Paulsen and Ferguson-Smith 2001, Reik and Walter 2001, Constância et al. 2004). The inverse scenario is responsible for AS. It leads to biallelic expression of these genes and silencing of the ubiquitin protein ligase gene *UBE3A* and the ATPase gene *ATP10A*, the only maternally expressed genes identified at this locus. Both syndromes are characterized by mental retardation, PWS also by undergrowth, muscular hypotony, and eating disorders resulting in severe obesity. Unlike AS patients, which exhibit a peculiar motion pattern and frequent laughing (thence the name "happy puppet syndrome"), PWS patients are easily frustrated. The orthologous region in the mouse genome is on chromosome 7.

Mouse chromosome 7 also contains imprinted loci that are found on other human chromosomes,

[5] http://www.har.mrc.ac.uk/research/genomic_imprinting/function.html

namely the Beckwith-Wiedemann syndrome (BWS) region (chr. 11) and *PEG3* region (chr. 19; Fig. 1.4). BWS is another well-known imprinting disorder with an overgrowth phenotype, predominantly caused by biallelic expression of *IGF2*. Silencing of the normally maternally expressed tumor suppressor gene *CDKN1C* makes the patients susceptible to tumors (Constância et al. 2004). Mouse models suggest that loss of maternal methylation at IC2 is responsible for aberrant expression (Fitzpatrick et al. 2002). In general, cancers often exhibit *IGF2* overexpression and silencing of *CDKN1C*. Research on mice showed that other imprinted transcripts from the BWS region influence the morphology and function of the placenta. In contrast, the transcription factor gene *Peg3* influences not only fetal growth, but also suckling and maternal behavior (Constância et al. 2004), similar to *Mest* (Lefebvre et al. 1998).

Silver-Russell syndrome is characterized by growth restriction that already starts before birth and is caused by maternal disomy and duplications of a region on human chromosome 7 (mouse chr. 6). Mutations in *PLAGL1* (human chr. 6, mouse chr. 10), another transcription factor gene, are responsible for transient neonatal diabetes and the mouse ortholog also has role in bone formation (Varrault et al. 2006). *GNAS* (human chr. 20, mouse chr. 2), encoding a G protein subunit, is involved in metabolism disorders (Constância et al. 2004). Besides, a number of other human disorders including autism, bipolar affective disorders, and schizophrenia have been linked to imprinting effects, but it is not known yet which genes are involved.

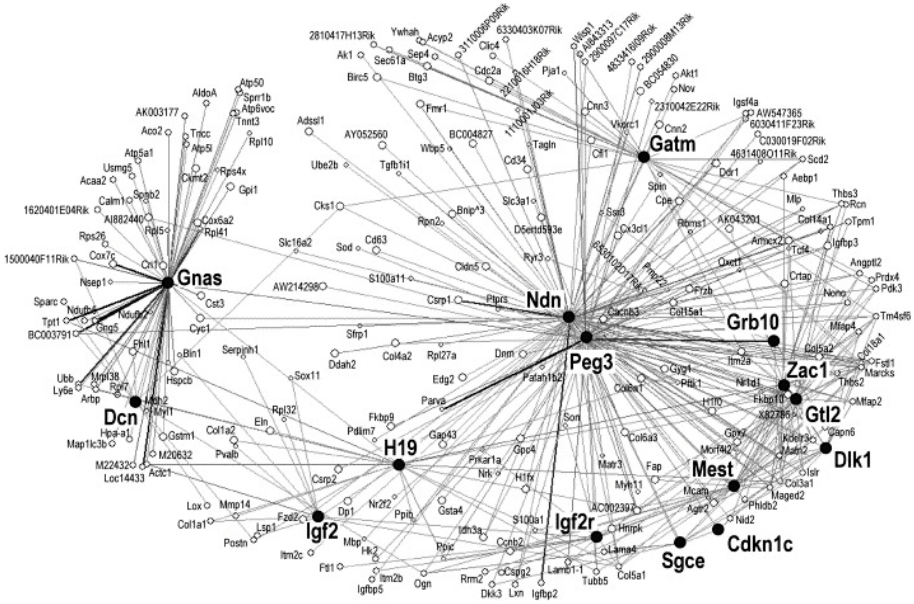


Figure 1.8: Network of imprinted genes
A gene network involved in the regulation of embryonic growth and differentiation was derived from microarray data by Varrault et al. (2006), from where the figure is taken. 246 genes are linked with at least three imprinted genes (bold style) by similar expression patterns and thus assumed to be coregulated. Zac1 is a synonymous name for the transcription factor encoded by the imprinted *Plagl1* gene. Additionally, IGF2 influences the expression of *Cdkn1c* (Grandjean et al. 2000). The human *PLAGL1* gene is involved in the activation of *KCNQ1OT1*, which in turn represses *KCNQ1* and *CDKN1C* (Arima et al. 2005).

1.8 Evolution and parental conflict

1.8.1 Occurrence of imprinting

The first mammalian imprinted genes were identified in the 1990s. To date, about 90 imprinted genes have been detected in human and mouse. Most genes that are imprinted in one species have been shown to be also imprinted in the other, but there are some discrepancies (Morison et al. 2005). For example, *Commd1* and *Impact* are imprinted in mouse but not in human; for *TRPM5* and *L3MBTL*, the situation is reversed. A few genes show opposite expression patterns in the two species such as *Copg2*, *Grb10*, and *Zim2* that are maternally expressed in mouse and paternally in human. This may be due to a different organization at these loci. In human, *ZIM2* is a transcriptional variant of *PEG3* as opposed to two separate genes, *Zim2* and *Peg3*, in mouse. Conflicting data on *IGF2R* arose because of polymorphic imprinting as imprinting was lost in the primate lineage but is still present to some extent in the human population (Killian et al. 2001). Since the necessary experimental procedures are difficult and time-consuming, the imprinting status of some orthologs remains unknown. For the same reason, data on other mammals are very limited. The Otago Catalogue currently lists a number of entries for cow, a few for pig, rat, and sheep, as well as one entry each for dog (*IGF2R*, O'Sullivan et al. 2007) and rabbit (*Impact*, Okamura et al. 2005). Recently, a large study on imprinted genes in the pig was published (Bischoff et al. 2009). Additionally, imprinting has been detected for some marsupial genes (*H19*, *IGF2*, *IGF2R*, *INS*, *MEST*, *PEG10*; Weidman et al. 2006, Ager et al. 2007, Smits et al. 2008, Suzuki et al. 2005, 2007).

IGF2, which was one of the first imprinted genes discovered, has become something like the "standard test gene", showing that among the vertebrates, imprinting is limited to the placental mammals (Fig. 1.9). Interestingly, orthologs of most imprinted genes reside in syntenic regions in the genomes of not only mammalian, but also other vertebrate species, where they are often arranged in clusters as well, for example in platypus (Warren et al. 2008), chicken, and even fish (Paulsen et al. 2005, Dünzinger et al. 2007). Thus, their existence and arrangement predate the evolution of their special regulation. Providing an example of how the function of a gene co-evolves with its regulation, *IGF2R* in non-therian species is not imprinted and its protein lacks the IGF2 binding domain (Killian et al. 2000, 2001).

Convergent evolution took place in flowering plants, which also have established imprinting mechanisms, however of completely unrelated genes. Although imprinting effects were observed in plants prior to their discovery in mammals, to date the number of identified imprinted genes is limited to ten in Arabidopsis and four in maize (Feil and Berger 2007, Gehring et al. 2009). Nevertheless, the epigenetic marks, namely differential methylation and histone modifications, are strikingly similar to those present in mammals (Feil and Berger 2007).

1.8.2 Embryonic development and parental conflict

What do plants and placental mammals have in common so that parent-of-origin dependent monoallelic expression could be evolutionary advantageous for such different organisms? The answer is likely that both have a direct connection between mother and embryo, leading to the post-zygotic extraction of maternal resources (Feil and Berger 2007). Just like the mammalian placenta, the endosperm of plants acts as an interface through which nutrients are transferred from the

mother to the embryo. Compared to spore plants or egg-laying species, seed plants and viviparous animals invest considerable amounts of maternal resources in their embryos.

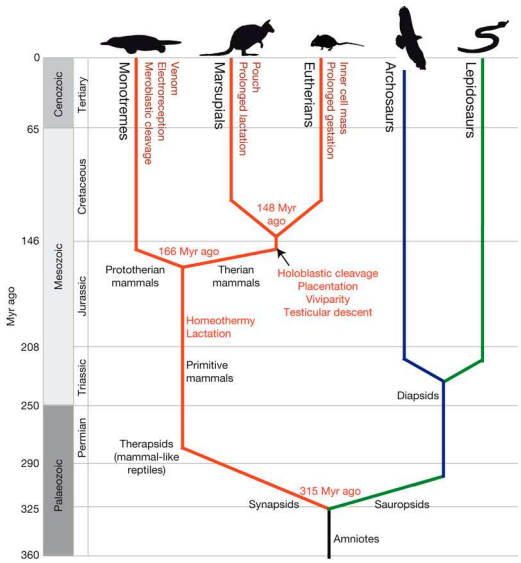


Figure 1.9: Evolution of imprinting
Imprinting supposedly came into being before the "invention" of the distinct placenta and long gestation in eutheria, but after that of milk supply by the mother, as imprinting is apparently absent in echidna and platypus. Eutherians (the "real" placental mammals) are commonly grouped together with the marsupials, whose embryonic development largely takes place in the pouch, to form the taxon theria. It is opposed to the prototheria, i.e. monotremes, also called egg-laying mammals. The figure is taken from Warren et al. 2008.

Based on the finding that the earliest identified murine imprinted genes *Igf2* and *Igf2r* encode a growth factor and its receptor, which targets IGF2 to lysosomes for degradation, it became apparent that embryonic growth was one of the key elements for the evolution of imprinting. As a prominent example for humans, the phenotype of the Beckwith-Wiedemann Syndrome is characterized by fetal and postnatal overgrowth and caused by biallelic expression of *IGF2*. Notably, *Igf2* is expressed from the paternal allele and *Igf2r* from the maternal one. This antagonism led to the nowadays widely accepted parental conflict hypothesis of imprinting, a concept describing the conflicting maternal and paternal interests within offspring, often also called kinship theory (Moore and Haig 1991). In polygamous species, siblings are on average more related to each other through the mother because they can have different fathers. Thus, whereas maternally expressed genes act for treating all her children equally in a trade-off between the fitness of individual offspring and the costs for the siblings, also future ones, paternally expressed genes aim at extracting a maximal amount of maternal resources for each of their children at a time. The paternal conflict hypothesis is supported by the functions of numerous other imprinted genes as well as the contrasting phenotypes of embryos with two paternal chromosomal sets (so-called androgenotes) and those with two maternal chromosomal sets (parthenogenotes or gynogenotes). Parthenogenetic embryos have a small placenta whereas androgenetic ones develop large extraembryonic tissues. Neither of them develop beyond mid-gestation. Thereby it was initially shown that maternal and paternal

chromosomes are unequal and both are needed for correct embryonic development (McGrath and Solter 1984, Surani et al. 1984).

Acquisition of imprinting may have been vital to the evolution of the placenta as an interface of parental conflict (Wood and Oakey 2006; Fig. 1.9). Remarkably, some imprinted genes are highly expressed in the placenta and important for its morphology. It must also be mentioned that special placental genes act in protecting the mother herself from overly demanding offspring. For egg-laying species, imprinting would make no sense according to the parental conflict theory because the amount of yolk is fixed around fertilization, long before the expression of embryonic genes, so that the embryo has no access to additional maternal resources (Moore and Haig 1991). Quantitative trait loci with parent-of-origin effects and reciprocal effects have been linked to imprinting in chicken but neither monoallelic expression nor differential methylation of the examined orthologous genes was detected (Tuiskula-Haavisto and Vilkki 2007).

When parental conflicts are reduced, as it is the case in self-fertilizing plants like *Arabidopsis thaliana* (Spillane et al. 2007) or in monogamous mammals, relaxation of imprinting would be expected (Feil and Berger 2007). Taking several evolutionary steps into account, the imprinting coevolution of *Igf2* and *Igf2r* has been modeled by Wilkins and Haig (2001, 2002). After *Igf2* became maternally silenced, expression from the active paternal allele increased and with it patrilinear fitness. Acquisition of the IGF2 binding site by the mannose phosphate receptor, which thereby became IGF2R, was a beneficial mutation in terms of the parental conflict theory, allowing for fast degradation of the growth factor. As a countermove to this, imprinting of *Igf2r* arose, leading in turn to elevated expression from the maternal allele to increase matrilinear fitness. Expression levels of *Igf2* and *Igf2r* eventually reached an evolutionary equilibrium. In general, the allele of the parent that benefits from a high production of a gene is predicted to produce its favored amount while the other allele is silent (Wilkins and Haig 2003). Thus, dosage compensation (which could as well be implemented by random monoallelic expression such as it is the case with olfactory receptor genes) is brought into agreement with parental interests. In primates and their closest relatives, reactivation of paternal *IGF2R* may be related to increased paternal contribution. Reduced interest in exploiting maternal resources should decrease *IGF2* expression to the level of the maternal optimum (Wilkins and Haig 2001). According to the models, imprinted demand inhibitors like *IGF2R* are more likely to be reactivated than demand enhancers like *IGF2* since degrading proteins requires energy, which is in any case unfavorable for the offspring (Wilkins and Haig 2001).

1.8.3 Evolution of imprinting regulatory elements

Having discussed a probable reason for why imprinting came into being, the next question is how it was established on a genomic level. The regulatory elements that modulate parent-of-origin dependent monoallelic expression must have arisen in a common ancestor of the species that show imprinting of the respective genes. Starting with the first appearance of genomic imprinting, assumed to have taken place in the late Jurassic before the marsupial-eutherian split (Fig. 1.9), genesis of imprinting centers seems to be an ongoing process since some genes are imprinted in a lineage-specific way. Paulsen et al. (2005) suppose that the Beckwith-Wiedemann syndrome region evolved from the ancestral non-imprinted state (as syntenic in fish) by gaining a DMR. Comparison to ancient states represented by the organization of the corresponding marsupial and monotreme loci can reveal the evolution of eutherian imprinting clusters (Hore et al. 2007). As an example,

before the split between marsupials and eutherian mammals the retroposon-derived gene *PEG10* brought with it a DMR leading to its imprinting. Later on, its influence expanded to neighboring genes that are now imprinted in eutherians but not in marsupials. In other cases, rearrangements conferred imprinting to loci that are only imprinted in eutherians. At the *DLK1-DIO3* domain, this happened by integration of *MEG3* (*Gtl2*), and in the Prader-Willi/Angelman Syndrome region by *SNRPN*, which is a paralog to ancestral *SNRPB*. At both regions, noncoding RNAs and more transposed genes were added (Hore et al. 2007). The recently reported, maternally expressed genes *Klf14* and *KLF14* are likely retrotransposons integrated into an existing imprinted domain after the marsupial-eutherian divergence (Parker-Katiraee et al. 2007). This example shows that if paralogs of biallelically expressed genes are inserted into an imprinting cluster, they can become imprinted as well. Also the murine-specific *Peg12* is such a proposed "innocent bystander" (Chai et al. 2001).

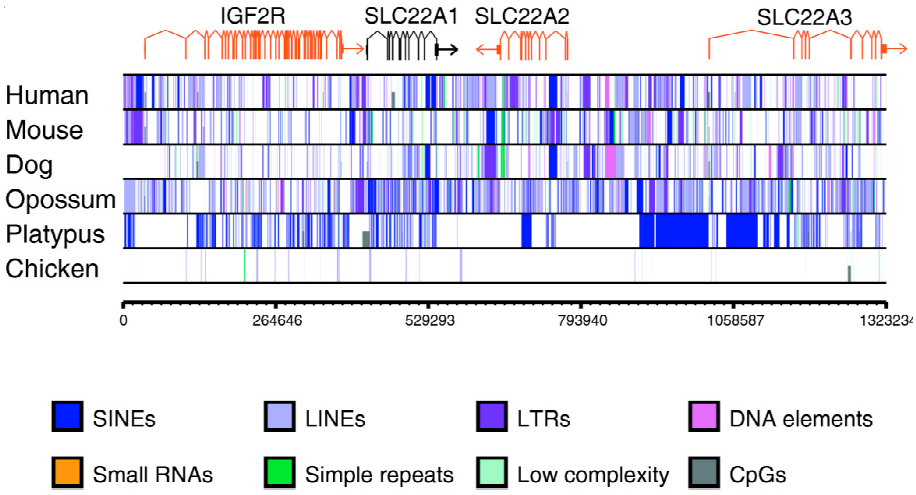


Figure 1.10: Distribution of repeats and CpG islands in orthologous sequences
Presumably in the process of becoming subject to imprinting, the corresponding regions – here orthologous sequences of the *IGF2R* region as an example, taken from Pask et al. (2009) – accumulated different repetitive elements. Whereas SINEs are reduced and LINEs prevail in species that show imprinting, the platypus shows enrichment of SINEs, which is, however, not significant (Pask et al. 2009). CpGs denote CpG islands (gray bars of half the size of repeat bars).

In contrast to *PEG10* and *IGF2/H19*, no DMRs were identified at the imprinted marsupial *IGF2R*, *MEST*, and *INS* loci (Killian et al. 2000, Suzuki et al. 2005, Ager et al. 2007). Although marsupials have not yet been investigated for the presence of inactivating histone modifications, it is assumed that these are sufficient for imprinting in these species. The histone-based silencing mechanism has been proposed to be more ancient than DNA methylation. For murine *Igf2r*, allele-specific histone variants in the absence of methylation at the inactive allele have been reported (Vu et al. 2004). Histones as an epigenetic memory that provides differentiating chromatin structures after demethylation might be responsible for the asynchronous methylation of maternally and paternally inherited alleles during imprint establishment: The originally methylated allele of *H19* becomes re-methylated earlier in murine sperm cells (Davis et al. 2000), likewise *Snrpn* in oocytes (Lucifero et al. 2004). Involvement of histones seems very probable since in primordial germ cells, DNA

demethylation occurs before histone replacement (Hajkova et al. 2008). Thus, DNA methylation might be a secondary trait to make the imprint more stable.

Bourc'his and Bestor (2006) assume that, after the CpG island character of paternally derived primary DMRs had been lost due to CpG depletion, a new mechanism with paternally expressed noncoding RNAs evolved "to counter the erosion of paternally methylated regions". Similarly, Reik and Walter (2001) argue that paternally derived DMRs should be unstable because of the demethylation of the paternal genome. Antisense transcripts might have gained increasing influence in imprinting clusters. Alternatively, they may have been the original mechanism. The complex *Gnas* locus represents an example of an evolutionary arms race in terms of the parental conflict hypothesis in which maternal genes switch off paternal ones and vice versa (Coombes et al. 2003).

A suspicious concentration of imprinted genes on rat chromosome 1 (M. Paulsen, unpublished data) suggests that the genes in question might have been distributed from a few ancestral regions, together with their regulatory elements, possibly introducing imprinting effects into new regions. Furthermore, duplications may have played an essential role in the establishment of imprinting since many genes that are subject to this special kind of regulation possess non-imprinted paralogs (Walter and Paulsen 2003, Wood et al. 2007) and the Arabidopsis *MEDEA* gene is a lineage-specific imprinted duplicate that acquired new functions (Spillane et al. 2007). Systematic investigation of paralogous genes has been one topic of this thesis, thus the mechanisms underlying duplications are explained in more detail in chapter 2.7 and the implications for imprinting are treated extensively in sections 3.4, 4.5, and 4.8.

1.8.4 Natural selection on imprinted genes

The evolution of species is considered as a dynamic interplay of mutations and different kinds of selection. If a mutation has strong negative consequences on fitness, it will not be propagated in the population. This purging is the action of the purifying selection (also called negative or stabilizing selection) that, on the genomic level, results in high conservation. Relaxation of purifying selection tolerates mutations that do not have severe effects, so-called slightly deleterious mutations. In contrast, mutations that lead to a beneficial phenotype are maintained by positive selection and, consistent with the alternative notation "Darwinian selection", can give rise to new species. Mutations affect both genes and their regulatory regions; they can influence expression as well as the sequence of the encoded proteins and even posttranslational events.

The monoallelic expression of imprinted genes and their connection with DMRs suggest that they may also be subject to a different selective pressure than biallelically expressed genes, resulting in different patterns of sequence conservation. Strong purifying selection on regulatory elements that convey imprinting would be expected whereas positive selection could be mirrored in species-specific features. The actual picture is complicated. Despite being the key regulatory elements, imprinting centers are little conserved with respect to their DNA sequences (Paulsen et al. 2001, Paulsen et al. 2005, Walter et al. 2006; compare Fig. 1.7). In some cases, existence of structural analogs may be sufficient. For example, the DMR in an intronic CpG island of *IGF2R* consists of completely unrelated sequences in human, mouse, and cow (Riesewijk et al. 1996). Nevertheless, highly conserved elements have been identified in imprinted regions outside of genes or CpG islands (Engemann et al. 2000, Paulsen et al. 2001, Tierling et al. 2006), and some of them act as additional regulatory elements (Ishihara et al. 2000, Takada et al. 2002, Lin et al. 2003).

There are few reports on the evolution of proteins encoded by imprinted genes. On the evolutionary most ancient level, the marsupial and eutherian *IGF2R* evolved from the mannose-6-phosphate receptor gene by gaining an IGF2 binding site (Killian et al. 2001). Studies involving a limited set of mouse and rat imprinted genes did not provide evidence for positive selection in the rodent lineage (McVean and Hurst 1997, Smith and Hurst 1999). Unlike the IGF2R-IGF2 interface region, which is highly conserved, the signal sequence of *IGF2R* that determines its location in the cell is strikingly divergent between mouse and rat as well as between human and cow (McVean and Hurst 1997, Smith and Hurst 1998). Consequently, interactions that are vital for protein function are likely preserved whereas the protein concentration may be altered by transporting it with increased or decreased efficiency. *KLF14* shows an enrichment of single nucleotide polymorphisms (SNPs) and accelerated evolution in the human lineage (Parker-Katiraee et al. 2007). As this gene encodes a transcription factor, that is, a member of a class of highly evolvable proteins jointly responsible for species diversity (Gibbs et al. 2004, Mikkelsen et al. 2007a), imprinting might not be involved as a crucial mechanism. It is, however, intriguing to remember that the set of imprinted genes contains many transcription factors. Research in Arabidopsis species identified the *MEDEA* gene as an evolutionary recent gene under positive selection (Spillane et al. 2007). Interestingly, all three genes mentioned above are maternally expressed, which brings evolution in context with the parental conflict hypothesis. *Ascl2*, *Cdkn1c*, and *Phlda2*, genes important for placental development, and genes encoding organic cation transporters involved in nutrient transfer to the embryo (*Slc22a2*, *Slc22a3*, and *Slc22a18*), are maternally expressed as well. Also due to their transcription in the oocyte (Chotalia et al. 2009), maternally expressed genes might be subject to special evolutionary patterns related to female-specific beneficial mutations.

Generally speaking, imprinting results in functional haploidy of the affected gene. If there are two different alleles, heterozygotes behave like homozygotes of the expressed allele and reciprocal heterozygotes differ with respect to phenotype and fitness (Patten and Haig 2008). Since imprinted genes also have direct effects on reproduction, imprinting is expected to sharpen selective elimination (Wilkins and Haig 2003). In the case of deleterious mutations on the expressed allele, purifying selection would eradicate any haploinsufficient individual and with it the mutated gene whereas beneficial mutations would provide a successful phenotype and promote positive selection. On the other hand, the inactive allele may accumulate mutations that remain unexposed as long as the sex of the transmitting parent does not switch (Wilkins and Haig 2001). As a consequence, different kinds of selection might act more efficiently on different sets of imprinted genes, similar to the situation on the X chromosome, of which the second copy is silenced in female mammals (Vicoso and Charlesworth 2006).

It seems likely that lineage-specific evolution of imprinting regulation contributed to the speciation of mammals (Reik and Walter 2001). Embryonic lethality of crosses between two species of deer mice is related to imbalanced expression of placental imprinted genes (Duselis and Vrana 2007) and strongly suggests that imprinting is involved in reproductive isolation between species. Imprinting might evolve quickly in certain groups of eutherians, namely in those where selective pressure is highest, whereas it might be lost in monogamous species (Feil and Berger 2007). In line with these hypotheses, we detected increased divergence of maternally expressed rodent genes from their human orthologs (chapter 3.4). Actually quite a number of genes that are imprinted in mouse show no or only developmentally or tissue-specifically restricted imprinting in human (Monk et al. 2006). Genes reported to be imprinted in a placenta-specific manner in mice but not in humans (*Tspan32*, *Cd81*, *Tssc4*) might be false positives. The placenta combines

maternal and embryonic tissues which are hard to separate because of the small size of this organ (M. Paulsen, pers. comm.). Thus, if there is high expression from the homozygous maternal tissue, but little or none from the heterozygous embryonic part, it seems that the maternal allele is preferentially expressed. Nevertheless, there are even some species-specific imprinted genes without any ortholog (*DIRAS3* and *TCEB3C* in human, *Peg12*, *Peg13*, *Tnfrsf23*, and *Zim1* in mouse). Ongoing Darwinian selection may therefore be restricted to evolutionary young imprinted genes or their regulatory elements.

Different imprinting equipment might be explained in the light of the conflict hypothesis. Having usually only one offspring a time would abolish intrauterine competition as well as limit necessity to conserve maternal resources and thus not require stringent imprinting (Morison et al. 2005, Monk et al. 2006). It is unclear if imprinting has relaxed in humans or expanded in mice, which have a very short gestation time and many litters. Our finding that mouse and rat imprinted genes, contrasting with their divergence from human, are highly conserved between the two modern rodents argues for the acquisition of beneficial mutations in a common ancestor with subsequent purifying selection in extant species (chapter 3.4). Unfortunately, experimental data on other species like cow, dog, pig, and rabbit are too limited. They would permit validation of this singleton pregnancy hypothesis and might reveal the relevance of multiple paternity.

1.9 Previous bioinformatics research related to imprinting

Experimental studies on imprinted genes are nowadays more often than not accompanied by bioinformatics analyses. The numerous relevant references were already mentioned in the corresponding previous sections. On the other hand, genome-wide studies that address gene expression patterns (Su et al. 2004) or histone modifications and CpG methylation (Mikkelsen et al. 2007b, Meissner et al. 2008) often cast a glance at imprinted regions. A limited number of large-scale bioinformatics studies were specially dedicated to imprinting, above all to the analysis of repetitive elements (Greally 2002, Ke et al. 2002a, 2002b, Allen et al. 2003, Walter et al. 2006; see section 1.6). Indeed, locations and orientation of repeats are among the most important discriminative features for the prediction of imprinted genes (Luedi et al. 2005, 2007). Besides repetitive elements, other specific DNA sequences might mark genes that show parent-of-origin specific monoallelic expression. An attempt to derive an "imprinting signature" from regions conserved in 24 human and mouse orthologs found 14 motifs significantly enriched in their non-exonic, non-repetitive sequences (Wang Z et al. 2004). Using a them in a logistic regression model, only eight imprinted genes of a test set of twelve were predicted correctly. None of the motifs has been linked to a regulatory element. One of them is similar to the tandem repeat motif in the intronic DMR of *KCNQ1* (Mancini-DiNardo et al. 2003, Paulsen et al. 2005; Fig. 1.7).

Comparisons between imprinted genes and genes that are subject to random monoallelic expression revealed that they are similar in some respects (Allen et al. 2003). Both classes show a depletion of SINEs and enrichment of LINEs. For randomly monoallelically expressed genes, however, the younger LINE subclasses prevail. This group differs from both imprinted and biallelically expressed genes by a lower G+C content and fewer CpG islands. Little is known about the mechanisms leading to random monoallelic expression, which are assumed to be similar to those of genomic imprinting and X inactivation. The common origin of silencing one allele may be related to dosage compensation. Having functions that are quite different from those of imprinted

genes, randomly monoallelically expressed genes are involved in odorant perception and the immune system, interestingly classes where also Darwinian selection acts (Gimelbrant et al. 2007). Since the phenomenon is widespread, functional hemizygosity seems to be evolutionary favorable. The parental conflict hypothesis gives a comprehensible explanation why in certain cases, such as those of genes related to embryonic development, the choice of the expressed allele is not left to chance.

Expression patterns of imprinted genes were investigated by Steinhoff et al. 2009. They found distinctive expression profiles and transcription factor binding site signatures for imprinted genes that are expressed in hormone producing tissues and the placenta. Kang et al. 2009 performed a systematic analysis on conserved predicted CTCF and YY1 binding sites near imprinted genes with subsequent experimental validation, emphasizing the special role of these proteins. Finally, evolution of imprinted genes was addressed for protein-coding sequences of mouse and rat (McVean and Hurst 1997, Smith and Hurst 1999), for individual genes (Parker-Katiraee et al. 2007, Spillane et al. 2007), and with reference to imprinting clusters (Paulsen et al. 2005, Hore et al. 2007, Pask et al. 2009).

Christoph Bock, an affiliated researcher at the Max Planck Institute for Computer Science, performed sequence analysis on CpG islands. He found that CGIs that are prone to methylation (Yamada et al. 2004) are characterized by T+G-rich sequence patterns, specific DNA repeats, and a particular DNA structure (Bock et al. 2006). Training support vector machines on epigenetic data – including DNA methylation, histone modifications, and chromatin accessibility – resulted in so-called epigenetic scores that for each CGI in the human genome predict the probability of having an open and transcriptionally active chromatin structure (Bock et al. 2007). CGIs with an intermediate score probably correspond to tissue-specific DMRs.

Initiated by several hypotheses, investigating the sequence characteristics of imprinted genes has been the main topic of this thesis. Their association with CpG islands, repeats, conserved elements, DNA motifs, and paralogs as well as possible effects of CpG deamination and evolutionary aspects are dealt with in detail in chapter 3 and discussed in chapter 4.

Chapter 2 – Materials and Methods

In the first section of the following chapter, I will briefly describe the three public genome databases used in my studies. The second section elaborates on various approaches for the identification of CpG islands, followed by an introduction to repeat detection. How to extract evolutionary conserved elements from alignments is explained in section 2.4. Strategies used for the annotation of regulatory sites in the UCSC database are shortly explained subsequently. Section 2.6 is dedicated to motif search. After that, I describe methods for finding patterns of evolution and custom Perl scripts. The last section gives an introduction into statistical methods.

2.1 Molecular databases and annotation resources

DNA sequences along with their annotations are provided by various sources. They share most of the available sequenced genomes but differ with respect to builds (versions of genome assemblies) and annotation of genes and other features.

2.1.1 NCBI

The RefSeq database of the United States National Bioinformatics Institute[1] (NCBI) contains amino acid sequences, genomic DNA and cDNAs (spliced mRNAs reverse transcribed into DNA) of a great variety of organisms. Identifiers starting with "NC_" or "NT_" are genomic contigs, "NM_", " NR_", and "NP_" refer to manually curated entries for cDNAs, noncoding RNAs, or protein sequences, respectively; those with an X instead of the N are predicted ones based on genomic DNA. Genes are predicted with the *gnomon* method[2] by using alignments of the sequences in the RefSeq database and *ab initio* models. cDNAs are aligned to the genome directly with *BlastN*[3] (Altschul et al. 1997), protein sequences onto the translated genome sequence using *BlastX*. After filtering the resulting heuristic local alignments for compartments in which the gene is approximately located, splice signals are taken into account to construct the exon-intron structure. A Hidden Markov transcript model, for which a schematic overview is given in figure 2.1, generates putative genes which are evaluated against the database of existing proteins.

The *MapViewer*[4] is useful for visualizing the organization of genes, markers, etc. on individual chromosomes. Genomic sequences of single genes or larger regions can be retrieved there via the download link in fasta format, which only contains a header line followed by the raw sequence, or in GenBank format, which also provides annotations of genes, transcriptional start sites and ends, coding exons, transcriptional and splice variants. This format is the default for RefSeq entries. Accession numbers for two groups of control genes (G1 and G2) were generated by appending random numbers[5] to the NM_ prefix.

Another useful source is HomoloGene[6] (see sections 2.7 and 3.4). The Entrez Programming Utilities[7] provide an interface to the NCBI Entrez query and database system so that records can be

[1] http://www.ncbi.nlm.nih.gov

[2] http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml

[3] http://www.blast.ncbi.nlm.nih.gov/Blast.cgi

[4] http://www.ncbi.nlm.nih.gov/mapview

[5] http://www.random.org

[6] http://www.ncbi.nlm.nih.gov/homologene

[7] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

retrieved by automated online queries. Last but not least the NCBI *PubMed* database has been indispensable for literature research. Entries displayed in the MEDLINE format can be imported directly into reference management programs.

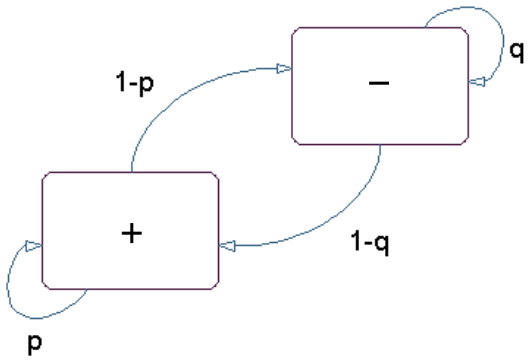


Figure 2.1: State diagram of a Markov Model
A Markov Model consists of different states, here two. Every time it is in its "+" state, it produces a nucleotide that belongs to the region with a distinguishing feature, e.g. protein-encoding, CpG island, or conserved. The "–" state produces background bases. By either staying in one of the states or alternating between them with so-called transition probabilities p and q, the Markov Model generates a DNA sequence. In a Hidden Markov Model, the production and transition probabilities are given but it is unknown which sequence of states (called a path) produced a given DNA sequence. The most probable path that assigns a state to each single nucleotide can be determined with the Viterbi algorithm (Durbin et al. 1998). To separate the regions of interest from the background, the sequence is divided into stretches of "+" and "–" states.

2.1.2 UCSC Genome Browser

Developed and provided by the University of Santa Cruz in California, the UCSC Genome Browser[8] comprises a great amount of so-called "tracks" derived from genome-wide bioinformatics analyses: genes and gene prediction, expression, comparative genomics, variation and repeats, etc. In the "Genomes" graphic interface, users can choose to display items of genomic features on the chromosome level. The "Table Browser" allows to download sequences and annotations. It is also possible to combine annotations from different tracks, e.g. intersecting CpG islands with conserved elements retrieves all CpG islands that have at least one base pair conserved. By choosing a cutoff, the regions may be restricted to a minimum overlap. Additionally, one can apply filters. Unfortunately, due to a condensed format, the results of these operations lack most of the original annotations. Therefore it was necessary to develop our own strategy including the overlap program from UCSC, custom scripts, and database queries (see section 2.8.2).

The general data format for tables inside UCSC tracks is a list of genomic coordinates followed by optional annotations with all items separated by tabs. To define regions of interest for a genome, a similarly structured "custom track" must be provided (Tab. 2.1). The so-called browser extensible data (BED) format is intended for the graphic interface; it allows different colors to be assigned to specific positions.

[8] http://genome.ucsc.edu

Table 2.1: Example custom track for UCSC

chrom	chromStart[a]	chromEnd	annotation (optional)
chr11	2246304	2248758	ASCL2
chr15	43440613	43458272	GATM
chr19	62015614	62043876	PEG3

[a] In the UCSC database, all start coordinates are zero-based. To convert coordinates shown on the graphic genome browser to the correctly corresponding region in the database, it is necessary to subtract one from the start. Zero-base becomes tricky when calculating overlaps: For example, "chr1 3000 4000" and "chr1 4000 5000" do *not* overlap by one nucleotide because the first sequence ends with base number 4000, but the second sequence starts with base number 4001.

In contrast to NCBI MapViewer and Ensembl (see 2.1.3), UCSC does not fuse genes into the longest possible transcripts but lists all variants. There are several types of genes available. For the RefSeq ones, cDNAs of protein-coding genes from NCBI RefSeq are projected onto the human genome using *Blat*[9], UCSC's *Blast*-like alignment program (Kent 2002). *Blat* keeps an index of all non-overlapping 11-mers with their positions on the genome. It is designed to quickly find sequences of 95% and greater identity of length 40 bp or more, as to map a sequence onto a chromosome[10]. Other gene sets comprise information on transcripts from various sources, including protein databases. We used data from the March 2006 human genome assembly (hg18, NCBI build 36.1) and the mouse genome assemblies February 2006 (mm8, NCBI build 36.1) and July 2007 (mm9, NCBI build 37.1). An additional control group (G3) was composed of randomly chosen genes from the human autosomes with orthologous mouse genes in mm8. Transcriptional variants were merged into genes beforehand by taking the most upstream transcriptional start site and the most downstream transcriptional termination site (compare section 2.8.1). The *liftOver* tool was used to map genomic coordinates between different genome assemblies. The hg18 annotation database and part of the mm9 one were downloaded and applied for genome-wide analyses presented in chapter 3.3 and 3.4.

2.1.3 Ensembl Genome Browser

The European counterpart to NCBI and UCSC, Ensembl[11], is run by the Sanger Institute[12] and the European Bioinformatics Institute[13]. They use essentially the same base data but a different gene build process favoring data from EMBL[14] so that genes have their own identifiers, starting with ENSHUM for human, ENSMUS for mouse, and so on. In contrast to UCSC, all Ensembl gene predictions are based on experimental evidence, that is records from RefSeq and protein sequence databases. Although Ensembl genes are linked to external identifiers like RefSeq accession numbers and gene names, the connection is incomplete. This imposed problems for relating gene

[9] http://genome.ucsc.edu/cgi-bin/hgBlat
[10] http://genome.ucsc.edu/FAQ/FAQblat
[11] http://www.ensembl.org
[12] http://www.sanger.ac.uk
[13] http://www.ebi.ac.uk
[14] http://www.ensembl.org/info/about/docs/genome_ annotation.html

data between Ensembl and NCBI (chapter 3.4). Sequences and annotations can be downloaded with BioMart[15]. This database systems, like the UCSC one, allows for different kinds of queries.

2.2 CpG islands

CpG islands (CGIs) are unmethylated CpG-rich islands in mammalian genomes that are otherwise depleted in CpG. They can mark promoter regions of genes and are found experimentally by cutting the DNA with methylation-sensitive restriction enzymes. There are several definitions and even more programs dedicated to the identification of CGIs in DNA sequences. According to the original definition (Gardiner-Garden and Frommer 1987), a CGI must have a G+C content of 50% or more, be at least 200 bp long, and have a ratio of observed CpGs to expected CpGs, CpG_{obs}/CpG_{exp}, of ≥ 0.6. There is some inconsistency between different programs on whether the values must be strictly greater or at least equal to the thresholds. More stringent parameters have been suggested to prevent detection of CpG-rich repetitive elements (Takai and Jones 2002). They require G+C content $\geq 55\%$, $CpG_{obs}/CpG_{exp} \geq 0.65$ and length ≥ 500 bp. Most simply, the G+C content and CpG_{obs}/CpG_{exp} for candidate CGIs are calculated in a sliding window. This approach and three basically different methods used by the programs applied in our studies will be presented in more detail below. The first section will present the calculation and implications of the CpG_{obs}/CpG_{exp} ratio, as elaborated in Hutter et al. 2009.

2.2.1 CpG_{obs}/CpG_{exp}, the margin effect and artifact CpG islands

To calculate the enrichment of CpG, its expected frequency is taken into account because the CpG content is highly correlated with the G+C content. Thus, for a sequence of the length *n* considered, the ratio CpG_{obs}/CpG_{exp} is defined as the frequency of CpG, CpG/*n*, divided by the product of the frequencies of G and C, G/*n* and C/*n*:

$$CpG_{obs}/CpG_{exp} = (CpG/n) \ / (G/n \cdot C/n) = (CpG \cdot n) / (G \cdot C)$$

A small example for highlighting the mathematical background: The sequence CGCGCCAG has a G+C content of 7/8 = 87.5%, and $CpG_{obs}/CpG_{exp} = (2 \cdot 8) / (4 \cdot 3) = 1.33$.

The same sequence, extended by an A+T-rich stretch on the right to CGCGCCAGAATAT, has a G+C content of 7/13 = 53.8%, and $CpG_{obs}/CpG_{exp} = (2 \cdot 13) / (4 \cdot 3) = 2.17$.

This example also shows how the CpG_{obs}/CpG_{exp} ratio can be artificially elevated by extending the central G+C rich core region by margins where G and C are underrepresented, as long as the G+C content stays above the required threshold – a phenomenon we termed the margin effect.

A high content of either G or C in low complexity sequences often results in a low expected CpG frequency (Fig. 2.2). In order to avoid that such sequences are identified as mathematical CGIs, the method implemented in the *CpG Island Searcher*[16] program (Takai and Jones 2002) requires the number of CpG dinucleotides per CGI candidate window to be at least seven. This minimum results from the Gardiner-Garden and Frommer (1987) criteria. According to these values, in a 200 bp sequence with a G+C content of 50% one would expect $200 \cdot 1/16 = 12.5$ CpGs

[15] http://biomart.org

[16] http://cpgislands.usc.edu

if all dinucleotides were equally distributed (CpG_{obs}/CpG_{exp} = 1). With CpG being still slightly depleted in CGIs, causing CpG_{obs}/CpG_{exp} = 0.6, this ratio reduces to 12.5 · 0.6 = 7.5 CpGs. Rounded to 7, this corresponds to a fraction of 3.5%.

ccccttcccctcccttcctcttcccctcccccttcccctcccttcccctcccctcccct
tcccctcccctctccttcc**cg**ttcc**cg**ttcccctcccctcccctt**cg**ctcccccttccc
ttcccctcccctcccttcccctcccttttccctcccctcccctttctcttcccctcccc
ttcctcttccc

Figure 2.2: Effect of low complexity regions
This 200 bp C- and C+T-rich sequence in the human *ANKS1* genomic sequence is identified as a CpG island by the program *CPGed* (Luque-Escamilla et al. 2005). It has a G+C content of 63%, 123 Cs and only 3 Gs which provide the 3 CpGs (bold). This yields an CpG_{obs}/CpG_{exp} ratio of 1.626. However, the resulting CpG content of 1.5% is far from the enrichment to be expected in a per definition CpG-rich island.

It must be kept in mind that computationally detected CGIs do not necessarily have regulatory functions. The identification of epigentically relevant CGIs is further complicated by CpG-rich repetitive elements like the human *Alu* elements (see also sections 1.2 and 1.6). They frequently fulfill the sequence criteria but, since their CpGs are generally methylated, do not obey the biological definition of providing an open chromatin structure. Using repeat masked sequences, in which repetitive elements are replaced by stretches of Ns, supposedly avoids detecting such repeat-dependent CGIs. However, we observed a severe drawback of this approach as artifact CGIs are created due to a phenomenon we call the N-effect. When C- and/or G-rich sequences are converted to Ns by repeat masking (see section 2.3.1), the G+C content decreases compared to original DNA. Provided that enough CpG dinucleotides are preserved, CpG_{obs}/CpG_{exp} is elevated and a CGI may be reported that would not have been identified in the unmasked sequence (Fig. 2.3). This effect is also seen for CGI detection in genome draft sequences that contain gaps of unsequenced nucleotides which are filled with stretches of Ns as a placeholder. Omitting Ns from the calculations as it is done by the program *CpGProD* (Ponger and Mouchiroud 2001) raises the G+C content, shortens the CGI (maybe even below the length threshold), and reduces CpG_{obs}/CpG_{exp} (Fig. 2.3).

AGCGGTCCATGGCATGGGCTGAG**gctgctgctgctgctgcttctg**AGGCGATCACTGGAG
CAGGCACCGCATCAGCACTGTTCCCAAGCACTCCTCCTGGGCTCCTGCGACTTTCAACCCAAG
AGTGAGGAGCAGCACTTGAGCACCGCGTCCCTTGCATACGGGATGTAGGACAGACTGTACCAG
GTCATCGACAA

Figure 2.3: N-effect
When the simple (CTG)n repeat (bold, lower case) in the murine *Adam19* genomic sequence is converted to Ns, *CpG Island Searcher* using the GGF criteria reports a CGI of 200 bp length with a G+C content of 51.5% and CpG_{obs}/CpG_{exp} = 0.603. The according section of the unmasked DNA has a G+C content of 59.5% and CpG_{obs}/CpG_{exp} 0.45, which would never meet the constraints for being a CGI. Excluding Ns from the calculations would result in a length of 175 bp, a G+C content of 58.86%, and CpG_{obs}/CpG_{exp} = 0.53, discarding this region because of a too low CpG_{obs}/CpG_{exp} ratio and being too short.

Unfortunately, omitting Ns has a drawback of its own: CpG_{obs}/CpG_{exp} can also be elevated compared to the original DNA and G+C content can even rise when As and Ts are masked, resulting in the above described margin effect, i.e. the discovery of CpG rich sequences that are actually too short to be CGIs (Fig. 2.4). For these reasons, we did not use repeat masked sequences for CGI detection but excluded CGIs that critically overlap with repetitive sequences afterwards (see chapter 3.1).

TTGGAATTGAGCATCATCACACTTAACCCCGACCACAGGCTATGTGAGTGGCCGGATGAGTCC
TTTTAGATGACCTCCATGCCAGCTGGTGTGGCTCATTCCGTCGGTCATTTGAAGCTAGTGCTC
ATCAAAGCTAGCACACTGAGCACGCCCCTCGATCGCCTGCAGTGCTTTGTATGTTTGGCGCGA
GTCTTCAGACTCTTAGTGTTCCCCAGAACCA**aagtgctgctgttgtctctgttggtgtaggga**
acagaggcacagggcagctcagtaactggtcctatgt

Figure 2.4: Inverse N-effect and margin effect
Omitting the SINE/MIR repeat section (bold, lower case) from this 289 bp sequence in the *Apbb2* genomic sequence from mouse, it is assigned a G+C content of 52.2% and CpG_{obs}/CpG_{exp} = 0.602, thus being labeled a CGI by GGF criteria. Note the margin effect that is particularly pronounced on the right side: The first CpG is at position 30 and the last one at position 187 of 289. The values calculated including the nucleotides provided by the repeat are G+C = 52.24% and CpG_{obs}/CpG_{exp} = 0.456. There are 9 CpGs, which corresponds to 3.11% CpG, but the actual CpG-rich part is rather small (159 bp) and would therefore not fulfill the GGF criteria although reaching 55.35% G+C, 5.66% CpG and CpG_{obs}/CpG_{exp} = 0.739.

2.2.2 The sliding window method

For determining CGIs in the most classical way, we used the Perl script command line version 1.3 cpgi130.pl of the *CpG Island Searcher* (Takai and Jones 2002). Starting at the beginning of the sequence, this program scans the sequence in a window of the minimum length a CGI must possess (200 bp for Gardiner-Garden and Frommer criteria and 500 bp for Takai and Jones parameters) that is moved forward in steps of 1 bp. As soon as a window meets the criteria, the next window is immediately shifted by 200 bp (or 500 bp, respectively). If the criteria are not fulfilled by the last, potentially enlarging window, it is shifted back towards the 5' end in steps of 1 bp until the new 200 (500) bp window meets the criteria. Then, G+C content and CpG_{obs}/CpG_{exp} are evaluated for the resulting large candidate CGI. If necessary, it is trimmed from both sides by 1 bp until it fulfills the criteria. Candidate windows that overlap or are less than 100 bp apart are fused under the condition that the merged region still fulfills the criteria. In a few cases, an unidentified program bug caused sequences to be reported as CGIs although they failed to fulfill the CpG_{obs}/CpG_{exp} criterion. As a result of the above mentioned margin effect, the sequences of CGIs determined with *CpG Island Searcher* usually have CpG-depleted margins.

2.2.3 Segmentation methods

Other programs split a sequence into CpG-rich and CpG-poor parts and afterwards check if the CpG-rich segments are CGIs according to the respective criteria. The Perl script *cpg* (Li et al. 2002) was obtained from the website[17] and the executable program *CPGed* (Luque-Escamilla et al. 2005) Version beta, 06-feb-2006 was kindly provided by J. Martínez-Arosa. Both methods perform entropic segmentation based on the calculation of the Jensen-Shannon divergence (JSD) of the

[17] http://www.nslij-genetics.org/wli/dnaseg/

CpG distribution in two adjacent windows *U* and *V*:

$$JSD = n/N \cdot (H[CpG_U/n] + H[\neg CpG_U/n]) + (N - n)/N \cdot (H[CpG_V/(N - n)] + H[\neg CpG_V/(N - n)])$$

where N is the length of the sequence, n the length of the left window U, CpG_U the number of CpG dinucleotides in U, $\neg CpG_U$ the number of non-CpGs in U (analogous for V), and H is the entropy, calculated as $H(x) = x \cdot \log_2(x)$ (Shannon 1948). The border between two segments – one the CGI and the other the non-CGI flanking sequence – is fixed where JSD is maximal.

In *cpg*, the left window is enlarged in steps of 1 bp from 1 to $N - 1$ while the right one shrinks from $N - 1$ to 1. At the first maximum of the Jensen-Shannon divergence, which is determined using the Bayesian information criterion, the sequence is split in two segments. Then, the recursive segmentation is applied again and again separately for the two segments until no significant maximum is found anymore. The runtime of this method is therefore $O(n \log n)$. The segmentation strength was set to 0.5. By an extension of the original script, I determined segments constituting CGIs as having a CpG content of ≥ 3.5%, which also corresponds to the threshold imposed by Takai and Jones (2002) of 7 CpGs in 200 bp for avoiding "mathematical" CGIs. Neighboring CGI candidate segments were merged and their CpG and G+C contents were recalculated as well as the CpG_{obs}/CpG_{exp} ratio. CGIs often consist of subdomains with different CpG content. In particular, the margins display lower values than the core region. Setting the CpG content to ≥ 6%, as proposed by Matsuo et al. (1993), excludes such regions. In contrast, the permissive ≥ 3.5% CpG criterion allows detection of weak CGIs as well as merging of otherwise separated CGIs. All candidate CGI segments smaller than 200 bp were excluded afterwards. Since borders between CGIs and their CpG-depleted flanking sequences are determined at individual CpG dinucleotides, there is no margin effect.

The method implemented in *CPGed* calculates the Jensen-Shannon divergence between two windows moved by steps of 1 bp which are of the same size except for the beginning and the end of the sequence. Thus, the algorithm conducts a local comparison. The estimated distance between two local maxima is called the sample interval. First, the Jensen-Shannon divergence is determined for the whole sequence. On both sides of its global maximum, the so-called jumping dwarf search algorithm searches for further maxima in jumps of the sample interval size. These must also exceed a divergence threshold of 0.001 to be significant. Default values are search window size 200 bp, $CpG_{obs}/CpG_{exp} \geq 0.65$ and G+C content > 55%, a sample interval of 5, and minimum length 200 bp for CGIs. All segments determined by segmentation are afterwards checked for their CpG_{obs}/CpG_{exp} ratio, G+C content, and length. Those fulfilling the criteria are assigned CGIs and merged if possible. The CpG content itself is not used in these steps, nor is a minimum number of CpGs required for a CGI. Thus, the program reports some "mathematical" CGIs that, being rich in either G or C, have a sufficiently high G+C content and CpG_{obs}/CpG_{exp} ratio but are in fact depleted in CpGs (Fig. 2.2). Some reported CGI sequences have an actual G+C content or CpG_{obs}/CpG_{exp} ratio below the given threshold, which is probably due to using a slightly different formula for the calculation of CpG_{obs}/CpG_{exp} (Matsuo et al. 1993) that slightly overestimates the enrichment of CpG:

$$(CpG \cdot n^2) / (G \cdot C \cdot (n - 1))$$

Despite the use of stringent criteria derived from those of Takai and Jones (2002), the authors also recommend repeat masking of the sequences, which, however, results in artifacts. Margins of CGIs

determined with *CPGed* can be CpG-depleted but these are rarely longer than 10 bp.

2.2.4 CpG clustering

CpGcluster[18] (Hackenberg et al. 2006) uses a very different approach without relying on established criteria. Instead, it determines clusters in which CpGs follow more closely after each other than expected if they were randomly distributed. The Perl script requires two input parameters that are not related to conventional CGI characteristics, namely the percentile and the p value (definitions see section 2.9). The first step of the *CpGcluster* algorithm is to calculate the distances between all neighboring CpG dinucleotides in the sequence. Immediately neighboring CpGs are assigned a distance of 1. The 50th percentile or median is the distance threshold that separates the distances into a lower half of relatively densely neighboring CpGs and an upper half of more distantly distributed ones (see also section 2.9). The median is the recommended value for the percentile parameter. Alternatively, the 75th percentile allows CpGs to be further apart up to the value that comprises 75% of all observed distance pairs. The program adds CpGs to an initial cluster as long as their distance corresponds to at most the selected percentile. Consequently, with the 75th percentile, clusters can become longer than with the 50th percentile. After finishing the extension, the p value of the resulting protoisland is calculated. It depends on both the CpG cluster length and the number of CpGs in it and serves for the distinction between a CpG enrichment by chance and a significant clustering. Only clusters that have a p value lower than or equal to the selected limit (the recommended value we used is 10^{-5}) are reported as CpG islands. Such CpG clusters begin and end with a CpG.

2.2.5 The UCSC elongation method

The method used at UCSC for the generation of their CpG islands track[19] is kind of an intermediate between sliding window and clustering. The UCSC program scans the repeat masked sequence one base at a time. It incorporates an additional scoring: If it finds a G after a C, it adds 17; in all other cases, the score is decreased by 1. At the end of the sequence (or earlier, if the score becomes 0), the segment from the start until the position with the maximum score is evaluated. Unless it meets the requirements of G+C > 50%, CpG_{obs}/CpG_{exp} > 0.6, and length > 200 bp, it is discarded. The same is done with the segment from the maximal score position up to the current one. Then search continues from the last position onwards. This strategy finds CpG islands with a very high CpG content that start and end with a CpG but it does not report stretches that would qualify by the traditional criteria only. The CpG islands track reports the percentage CpG as twice the number of CpGs divided by the sequences length. I recalculated it as the more commonly used ratio of the given number of occurrences to the length. Since the CpG dinucleotide is symmetric, the opposite strand needs not be taken into account. The arbitrary score of 17 is based on a heuristic that yields a similar number of CGIs and genes in the human genome as well as an enrichment of CGIs in promoter regions.

[18] http://bioinfo2.ugr.es/CpGcluster

[19] http://genomewiki.cse.ucsc.edu/index.php/CpG_Islands

2.3 Repetitive elements

As already mentioned in section 1.6, repetitive elements are nucleotide patterns that, in contrast to unique sequences, occur multiple times in the genome. In fact, a substantial part of mammalian genomes consists of repeats. Masking them is necessary before making alignments or performing motif search because they cause false positive hits. On the other hand, the occurrence of specific repeats can give hints on the evolution of genomic loci.

2.3.1 RepeatMasker

The standard tool for detection and masking of known repetitive elements in genomic sequences is *RepeatMasker*[20] (Smit et al., unpublished). It creates consensus sequences of known interspersed repeats from the Repbase libraries (copyrighted by the Genetic Information Research Institute[21]) and aligns them to the query sequence – that is, the genomic sequence to scan for repeats – with the *Blast*-like program *cross_match*[22]. *RepeatMasker* is also capable of finding short simple repeats and low complexity regions (compare Fig. 2.2, Fig. 2.3). Only di- to pentameric and some hexameric tandem repeats are scanned for whereas simple repeats shorter than 20 bp are ignored. By default, repeats are replaced by stretches of Ns of the same size. UCSC, which applies the most up-to-date and not yet publicly available versions of *RepeatMasker* and repeat data, makes use of so-called softmasked sequences in with repetitive elements are represented in lower case. Sometimes it is of interest to know what kinds of repeats are in the sequence. Therefore, *RepeatMasker* also provides their annotation including data on their alignment scores, divergence from the consensus, orientation, and repeat class, as well as a summary table for the classes. By using the annotations, individual repeat types can be marked in the sequence.

During the repeat masking process, different classes of repeats are searched for one after the other. Since an old repeat can be split by a younger one integrated into it, *RepeatMasker* removes already detected repeats before it scans the sequence for such "tough" elements. Options for the program include different sensitivity modes (fast for rough scan, sensitive and therefore slow for detecting highly diverged elements), limiting to subsets of repeats that have a certain maximum divergence from the consensus, and fragmentation of large sequences. Adjusting for G+C content takes isochores into account and allows using the most appropriate scoring matrices for the alignment. There are sometimes discrepancies between the masked sequence and the annotation, which can lead to more or less masked bases in either of them: Unmasked regions between flanking identical simple repeats are annotated as one stretch if fewer than 10 bases separate them, and fragments of repeats shorter than 10 bp are not annotated but are masked. When reconstructing masked sequences according to the annotation tables, we found that they contained more nucleotides assigned to repetitive elements than the original masked sequences.

UCSC *RepeatMasker* tracks were taken for the analyses presented in sections 3.3 and 3.4 whereas for the studies in sections 3.1 and 3.2, repeat masking was performed with local installations of *RepeatMasker* open version 3.08, the alternative alignment program *WuBlast*[23], and the Repbase *RepeatMasker* Libraries July 2004.

[20] http://www.repeatmasker.org
[21] http://www.girinst.org
[22] http://www.phrap.org
[23] http://blast.wustl.edu

2.3.2 Tandem Repeats Finder

A tandem repeat array consists of a sequence pattern or motif that is repeated several times in a head-to-tail fashion without being interrupted by unrelated sequence. The individual copies need not be 100% identical. Dependent on the identification method, already an incomplete second instance of the motif is sufficient to form a tandem repeat. Since motif length, number of repetitions, and deviation of individual copies induce a quite huge parameter space, the detection of tandem repeats is a complex bioinformatics problem. Algorithms applied to this task include suffix trees (Delcher et al. 2003), alignments (Parsons 1995), data compression and others. Most programs are slow with runtimes of $O(n^2 \cdot \log n)$ or worse and require a high *a priori* knowledge, e.g. length of the repeat, or even the motif (see references in Benson 1999).

A widely applied tool with fast runtime is *Tandem Repeats Finder*[24] (Benson 1999). UCSC applies it for the generation of the simple tandem repeats track. All bases in lower case that do not overlap with the *RepeatMasker* track are tandem repeats (see also section 3.3). For identifying tandem repeats in CpG islands (chapter 3.2), I used a local installation of version 3.21. The current version is 4.00. *Tandem Repeats Finder* is based on a probabilistic approach and therefore instead of numbers requires probabilities for matches and gaps, as well as penalties for mismatches and indels. The chosen values are adapted from the *Tandem Repeats Finder* description page as follows: match score = 2, mismatch score = 5, indel score = 7, match probability = 80, indel probability = 10, minscore to report = 100, maxperiod (maximal possible motif length) = 2000. All values are upper bounds, that means the given parameters can find copies that are at most this divergent.

First, the program generates all patterns (probes) of a certain size k whose positions in the sequence serve as a kind of anchors. Since this means locating 4^k possible probes, k must be chosen as a sufficiently small number. Whereas it is unlikely that a large pattern is exactly repeated, small probes occur too often to be informative. Therefore the probabilistic model checks a set of different ks where each value of k is optimal for a range of pattern sizes, e.g. $k = 4$ for 1-29 bp. Motifs of up to 500 bp can be detected with $k = 7$. The maximal possible motif length is currently 2000 bp. All probes are looked up by scanning the sequence only once. The positions of all identified probes are saved in a history list. As soon as the same probe is found at two sites i and j, a possible pattern size has been determined as the distance between them, d. Next, the program checks if any probes between i and j also occur at the same distance. Hits are stored in a distance list which is updated with the new position i every time a match at a distance d is detected. At the same time, lists for slightly different values of d are updated likewise because the distance between two copies (which eventually converges to the pattern size) may vary due to indels.

After checking if the information in the distance list passes several statistical tests (described below), all candidate patterns from j+1 to j+i are aligned with their surrounding sequence. The resulting alignments yield a consensus pattern (compare Tab. 2.5) that is used for realignment. These steps are the most time-consuming ones. A tandem repeat is modeled as a Bernoulli sequence, that is, the result of a series of coin tosses. Runs of heads (matches) may be interrupted by tails (mismatches or indels). Figure 2.5 shows an example.

[24] http://tandem.bu.edu/trf/trf.html

A	**B**
TCATGT	TCATGT
TCATGT	TCTTGT
HHHHHH	HHTHHH

Figure 2.5: Tandem repeats as Bernoulli sequences
For two adjacent copies of a repeat pattern, H indicates a match, T a mismatch (including indels). The two copies in A can be detected by looking for identical probes. The divergent copies in B will be identified in the alignment step.

The sum of heads provides information about the percent identity between two copies. From its distribution, approximated by a normal distribution with the given match and mismatch parameters, a minimal expected number of heads per pattern size can be calculated. The distance variation between two copies is estimated from a random walk distribution. Random Bernoulli sequences are used to simulate the distribution of pattern sizes in order to distinguish real tandem repeats from direct repeats that are not arranged in a head-to-tail manner. The range of probe sizes is taken from the geometric distribution, which is also known as waiting time distribution: How many coin tosses does it take until the first occurrence of a run of k heads? E.g. for a match probability of 0.75, a run of 5 heads is seen with a 95% chance after at least 31 trials. That means, for $k = 5$ we expect to find the next identical probe at a distance of at least 31 bp if there is a tandem repeat. For detecting tandem repeats with smaller or considerably larger motifs, respectively, different sizes of k have to be chosen.

Finally, if the score calculated from the alignment is high enough compared to random sequences, match statistics, base composition and entropy (see below) are calculated. In the case of very long motifs, 1.8 approximate repetitions are already sufficient for a tandem repeat to be reported. Repeats with many copies are often redundantly reported for different motif sizes, which yield slightly different scores. Also start and end points may be shifted. Sometimes, a tandem repeat in which the copies diverge gradually is split into two or even more overlapping ones. Therefore, the output was manually edited: Redundant repeat arrays, starting and ending at the same position with different single copy length, were counted as one. Two repeat arrays were regarded as a single one if the smaller one was overlapped by the larger one by more than 50%. The length was counted from the start of the first to the end of the second array. As a consensus sequence, the smaller motif was chosen.

The entropy H of a sequence is a measure of the sequence complexity. It is calculated as already mentioned in section 2.2.3, thus for a sequence

$$H(j) = -\sum_{i=1}^{j} n_{ij} \cdot \log_2(n_{ij})$$

where n_{ij} is the frequency of nucleotide i at position j. For example, a poly-A sequence has entropy 0, a simple repeat like (AT)n has entropy 1 and a sequence in which all four nucleotides are equally frequent would have a value of 2. $\log_2$ is used because it requires two bits of information to determine which nucleotide is at a position: The first bit encodes the group, the second one then

differentiates between the two possible bases in each group. An example is given in table 2.2.

Table 2.2: Possible encoding of nucleotides with two bit information

bit1/bit2	1 (3 hydrogen bonds)	0 (not 3 but 2 hydrogen bonds)
1 (pyrimidine)	11 = C	01 = T
0 (not pyrimidine = purine)	10 = G	00 = A

2.4 Alignments and conserved elements

In order to detect functional elements, comparative genomics analyzes the conservation of orthologous genomic sequences. This approach is also known as phylogenetic footprinting since evolution has left its traces on the DNA sequences, keeping protein-coding or regulatory regions conserved while allowing unimportant ones to diverge. The prerequisite for identifying conserved elements is to produce reliable alignments. Another critical issue is the definition of conservation. Conserved elements may be detected in global alignments by sliding window approaches considering a minimum of identical alignment positions, or they may be derived by combining local alignments. In this section, we will first look at the alignment tool used by UCSC and then discuss two methods of identifying differently defined conserved elements.

2.4.1 Blastz

Similar to *Blast* (Altschul et al. 1997), *Blastz*[25] is a high-performance program for generating local alignments (Schwartz et al. 2003). It is, however, not designed for database searches but for the fast and sensitive alignment of two DNA sequences. It can use quite sophisticated substitution matrices which account for the different probabilities of transitions and transversions (Tab. 2.3). One sequence is the reference that is kept fixed to determine the order of the hits. Per default, *Blastz* first looks for all 19-mers which must contain at least 12 identical positions between the two sequences. Not allowing transitions reduces the number of possible 19-mers and makes the program run faster. Alternatively, one can, as in *Blast*, set a specific word length to determine all possible *w*-mers that have a similar score as the original word. After a query word has been found in the sequence to align, the hit is extended on both sides. *Blast* does this without gaps, adding pairs of residues as long as the score does not drop below a certain threshold which is called *K* in *Blastz* (default 3000) with its gapped alignment. This results in high scoring segment pairs, of which the best ones, called diagonals, are combined using dynamic programming with gaps. Since indels are usually longer than one bp and do not alternate with match positions, there is a strict gap open penalty to set the first gap symbol (O = 400); subsequent gaps come "cheaper" with a gap extension penalty (E = 30). Alignments are not allowed to start in repetitive elements that are indicated by lower case letters, but they may extend into them, thus aligning conserved repeats.

[25] http://www.bx.psu.edu/miller_lab/

Table 2.3: Default substitution matrix of *Blastz*

	A	C	G	T
A	91	-114	-31	-123
C	-114	100	-125	-31
G	-31	-125	100	-114
T	-123	-31	-114	91

Replacing a purine nucleoside by the other purine (A-G), or pyrimidine by pyrimidine (C-T, e.g. by cytosine deamination), is called transition and more likely to occur in nature than mutation of a purine into a pyrimidine or vice versa, called transversion. Consequently, transitions are less severely penalized in the alignment. The bases that form three hydrogen bonds (C and G) are slightly underrepresented in mammalian genomes.

2.4.2 Pairwise evolutionary conserved elements

Evolutionary conserved elements (ECRs) are defined as sequence stretches that have at least 70% identity over at least 100 bp (Loots et al. 2002, Loots and Ovcharenko 2005). Our identification procedure used in sections 3.3 and 3.4 emulates the strategy of *PipMaker*[26] (Schwartz et al. 2000). After building *Blastz* alignments with the B = 0 and C = 2 options, the program *single_cov* from the *MultiPipMaker* package (Schwartz et al. 2003) was applied to get unique hits for ambiguous local alignments. This program only keeps the hit with the best score and discards hits with ≤ 15 bp. This trimming invalidates information about the percent identity of the first and last gap-free segment. ECRs were then extracted and combined from local alignments with a modified version of the Perl program *strong-hits* from the *PipTools* package[27] (Elnitski et al. 2002): Adjacent gap-free segments that must fulfill the ≥ 70% identity threshold are joined if they have maximally 100 gaps between them. The original script discounted gaps that in fact reduce the identity, therefore I modified it to discount gaps from the length of the ECR. Additionally, an ECR is allowed to contain at most 10% gaps in its alignment. Only ECRs of at least 100 bp (excluding gaps) are regarded.

2.4.3 PhastCons most conserved elements

To date, the genomes of about 30 vertebrate species have been more or less completely sequenced and assembled. These data can be used to detect conserved elements shared by all of them or by specific lineages, for example mammals. In essence, the genome-wide conserved elements provided by UCSC are generated with a combination of *Blastz* pairwise alignments into a multiple alignment, from which the highly conserved sections are extracted with a phylogenetic Hidden Markov model. In chapters 3.3 and 3.4, we worked with the subset for the 18 placental mammals.

After masking or even removing repetitive elements, the threaded blockset aligner (Blanchette et al. 2004) calls *Blastz* to create all possible pairwise alignments on the forward and reverse complementary strand. To improve the alignments, they are filtered based on reciprocal best hits and synteny to reduce both paralogs and suspect regions from low-quality assemblies. Next, the *Multiz* program creates a progressive multiple local alignment. In contrast to the algorithm of *ClustalW* (Thompson et al. 2002, Larkin et al. 2007), where the order of the sequences is

[26] http://pipmaker.bx.psu.edu/pipmaker/

[27] http://www.bx.psu.edu/miller_lab/

determined by their pairwise identities, *Multiz* is guided by a given phylogenetic tree. Moreover, whereas in *ClustalW* each resulting global alignment is fixed, the *Multiz* local one is iteratively improved as more species are added. The final alignment can be projected onto a reference genome.

Eventually, the phylogenetic Hidden Markov Model of *phastCons* (Siepel et al. 2005) divides the alignment into regions that are probably generated by the conserved state and such which are more likely to be generated by the non-conserved state. Conservation scores are calculated in a way that weighs the contribution of each species according to its position in the phylogenetic tree. Species with a high evolutionary distance from the others, like rodents (compare Fig. 2.11), therefore gain greater influence than closely related ones. The raw score of a sequence *S* is the log-odds ratio (lod):

$$lod = \log \frac{P(S \mid \text{conserved state})}{P(S \mid \text{nonconserved state})}$$

It is converted into a normalized score by dividing it by the length of the conserved region and thus is independent of any reference genome. High scores can be reached by short but highly conserved regions as well as by long but not so highly conserved ones. The set of *phastCons* mammalian most conserved regions (PCSs) given in the "Mammal (phastConsElements28wayPlacMammal)" track comprises those that receive a normalized score of at least 189 in the alignment subset of the 18 placental mammals. For the mouse genome version mm9, the updated 30way alignments were used.

bin	chrom	chromStart	chromEnd	name	score
585	chr1	1865	1948	lod=16	260
585	chr1	2873	2916	lod=16	260
585	chr1	3081	3132	lod=18	275
585	chr1	3996	4038	lod=18	275
585	chr1	4565	4651	lod=16	260
585	chr1	4826	4904	lod=17	268
585	chr1	5646	5794	lod=36	360

Table 2.4: Sample of the phastConsElements28wayPlacMammal track
There are 2,040,420 *phastCons* mammalian most conserved elements in the entire human genome, including sequences whose exact location on the chromosome is unknown as well as alternative haplotypes for some regions. The "bin" field contains an index used by programs to speed chromosome range queries. In the "name" column, the log-odds ratio of the *phastCons* Markov Model is given. Normalized by the length of the conserved regions, it is reported as the score. The highest possible score in this set is 999.

A shortcoming of the *phastCons* method is the treatment of gaps as missing data: "*PhastCons* does not model indels, and its conclusions about conservation depend purely on aligned bases." (A. Siepel, pers. comm.). This means that in regions where only sequences of a few species are present, their conservation immediately yields a high score. Additionally, the conserved regions are only available as projections onto individual genomes. If there is a insertion or deletion in the species of

interest on which the PCS is projected onto, information on the length of the conserved element is invalidated. Thus, we required the PCSs to be ≥ 20 bp long.

2.5 Annotations of regulatory regions and polymorphisms

UCSC offers several genome-wide annotations of (putative) regulatory elements. CpG islands (section 2.2) belong to them as well. Additionally, chromatin structure, variations like single nucleotide polymorphisms (SNPs), and repetitive elements (section 2.3) influence gene expression. Here, I will present some of the data that were used for the analyses in chapter 3.3.

The tfbsConSites table of the TFBS Conserved track contains computationally predicted transcription factor binding sites (TFBSs) that are conserved in human (hg18), mouse (mm8), and rat (November 2004 assembly rn4, Rat Genome Sequencing Consortium build 3.4). For creating it, multiple local alignments of the three genomes are scanned with 258 matrices for TFBS of transcription factors known in human, mouse, or rat. They are taken from the Transfac Matrix Database version 7.0 created by Biobase[28] (Matys et al. 2003). Table 2.5 shows the TATA box as an example for such a matrix. An alternative representation is in the form of a logo (Schneider and Stephens 1990, Shaner et al. 1993), given for the TATA box in Fig. 2.6.

Table 2.5: Matrix for the TATA box

nucleotide/ position	1	2	3	4	5	6	7
A	0	14	0	13	9	14	7
T	14	0	14	1	4	0	6
G	0	0	0	0	0	0	1
C	0	0	0	0	1	0	0
consensus[a]	**T**	**A**	**T**	**A**	**W**	**A**	**A**

[a] To calculate the consensus sequence, at each position the most frequent nucleotide is taken. In case of equal occurrence, there is the possibility to choose a IUPAC ambiguous nucleotide character (Tab. 2.6). Note that there are no gaps allowed in alignments of TFBSs.

To convert the matrix into a position-specific scoring matrix (PSSM), the score at each position can be calculated with the formula

$$score = \ln \frac{(n_{ij} + p_i)/(N+1)}{p_i}$$

where n_{ij} = frequency of nucleotide i at position j, p_i = *a priori* frequency of the nucleotide (given in background table; 0.25 if there is no compositional bias), and N = number of sequences.

The score of the matrix itself is just the sum of the scores of its consensus sequence. The score of a sequence is that of its nucleotide scores when aligned with the matrix. To be conserved, the

[28] http://www.gene-regulation.com/pub/databases.html

putative TFBS must exceed a significant score for the sequences all three species. Summing them up results in the score of the TFBS. An additional Z score, which gives the significance of the binding site, is computed by comparison with TFBSs in the 5000 bp upstream regions of all RefSeq genes. Since some matrices are redundant, at overlapping sites the factor with the highest Z score is chosen as the representative one. There is no filtering with respect to coding exons, which are highly conserved but not expected to harbor TFBSs. The matrix names can be linked to their accession numbers, factor names, and species given in an additional table, tfbsConsFactors.

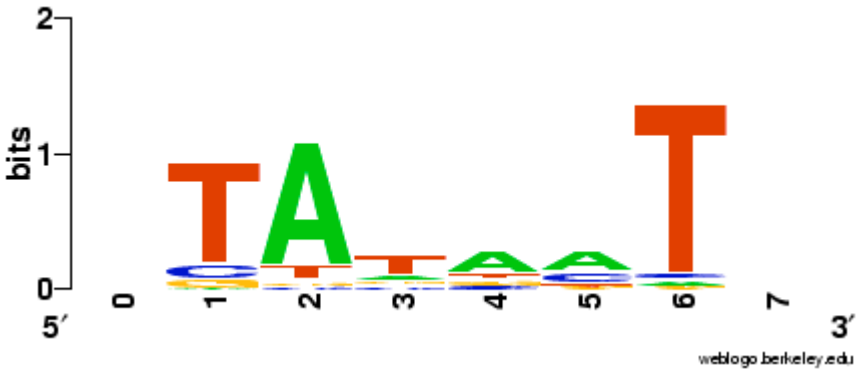


Figure 2.6: Sequence logo of the TATA box
Alignments of transcription factor binding sites or similar motifs can be transformed into logos by calculating their information content. The height of base n at position j is equal to its information content, $n_{ij} \cdot R(j)$, where $R(j) = 2 - H(j) - e(n)$ with the entropy H (see section 2.2.3) and $e(n)$ as a correction factor for short sequences. The logo was created with *WebLogo*[29].

Other data available at UCSC are adapted from public sources. The Open Regulatory Annotation database ORegAnno[30] (Montgomery et al. 2006, Griffith et al. 2008) provides data of experimentally validated regulatory regions, regulatory polymorphisms, and transcription factor binding sites (most of them CTCF binding sites) curated from the literature. Regions are reported with ≥ 40 bp of the flanking sequence in the UCSC oreganno track. The identifiers, which are composed of the prefix OREG and a number, can be linked to their attributes via the oregannoAttr table.

Single nucleotide polymorphisms (SNPs) are of special interest in medicine and pharmacogenomics since they have influence on gene expression and even protein sequences. In coding regions, SNPs can change amino acids or induce stop codons; at synonymous sites or in untranslated regions, they act on mRNA stability; intronic SNPs can abolish or create splice sites; SNPs in promoter regions may affect the binding of transcription factors. Basic information for the current UCSC snp129 track is taken from the NCBI dbSNP database[31] (Sherry et al. 2001). Besides SNPs in the strict sense (exchange of one nucleotide by another), also short deletion and insertion

[29] http://weblogo.berkeley.edu
[30] http://www.oreganno.org/oregano/Index.jsp
[31] http://www.ncbi.nlm.nih.gov/SNP/

polymorphisms are deposited in the repository. The submitted SNPs are mapped onto the genome of the respective species to compare them to the reference assembly and to infer their location relative to genes. Their estimated frequency in the population is calculated as average heterozygosity. For the human genome, UCSC provides SNP masked sequences in which the polymorphic bases are represented as the corresponding IUPAC ambiguous characters (Cornish-Bowden 1985; Tab. 2.6). Indels are not included. Interestingly, SNPs frequently occur in conserved regions: Only four out of 57 protein-encoding imprinted genes analyzed in chapter 4.4 do not have a SNP in highly conserved parts of their coding sequence.

Other regulatory data from experiments are not actually available genome-wide but only for individual chromosomes in different tissues, mostly cancer derived. The ENCODE project (Birney et al. 2007) has generated chromatin immunoprecipitation data for DNA binding proteins like TATA box binding protein, RNA polymerase II, the transcription factors c-Fos, c-Jun, chromatin remodeling proteins, and histone modifications. Additional resources of chromatin structure comprise replication origins (Uva DNA Rep track) and DnaseI hypersensitive sites (Duke/NHGRI Dnase track). The scarcity of the annotated regions seems to be a problem. Although asynchronous replication times have been reported for imprinted genes, preliminary analyses with 482 origins of late replication only yielded one overlap in *IGF2* and none of the 582 early replication origins occurred near an imprinted gene.

Table 2.6: IUPAC ambiguous nucleotide symbols

symbol	nucleotides	mnemonic
B	C, G, T	not A
D	A, G, T	not C
H	A, C, T	not G
K	G, T	**K**eto group
M	A, C	a**M**ino group
N	A, C, G, T	a**N**y
R	A, G	pu**R**ine
S	C, G	**S**trong
V	A, C, G	not T
W	A, T	**W**eak
Y	C, T	p**Y**rimidine

2.6 Motif search

In a set of sequences, overrepresented oligonucleotides (*k*-mers) may represent relevant motifs like unknown repeats or transcription factor binding sites, or be responsible for special DNA structures. Algorithms for motif search were first developed for promoter regions of coexpressed yeast genes (van Helden et al. 1998, Hughes et al. 2000, Sinha and Tompa 2002) and then extended to other organisms. Phylogenetic footprinting, i.e. search in orthologous regulatory regions, has proven useful for vertebrate sequences to reduce the proportion of false positive predictions (Blanchette and Tompa 2002, Sinha et al. 2004). A significant enrichment of motifs must be statistically verified. Numerous different approaches exist that compare the motif frequencies in the set of interest to those in a background set, which can be generated by Markov Models, consist of genomic DNA not supposed to be under functional constraints, or be a set of control sequences provided by the user.

Although the procedure of identifying regulatory elements seems simple in theory, it is not for metazoan genomes (Wasserman and Sandelin 2004). During this work I tried out several programs and found that motif search is rich in complications. First, a sequence can be decomposed into oligonucleotides in different ways. For example, AAAAAA may generate the 4-mer AAAA three times if overlaps are allowed (else only one occurrence), and TTTT three times if also the reverse complement is taken into account. Masking of low complexity regions should be considered to avoid that they introduce a bias. Similarly, interspersed repeats and tandem repeats should be removed from the sequences before motif search (see section 2.3). Another problematic issue is that motifs, particularly if they correspond to transcription factor binding sites, may deviate at some positions or overlap in parts. Although this may be resolved by clustering (van Helden et al. 1998), this step is rarely realized as it requires additional parameters and cutoffs. Additionally, some sequences may contribute many occurrences of the same motif whereas in others it is absent. Simple enumeration programs that just report the *k*-mers sorted by frequency, like *wordcount* from the EMBOSS package (Rice et al. 2000), do not adjust for the resulting bias. More advanced methods require each motif to be present in at least a certain number of sequences.

Most tools are designed for finding TFBSs. An overview of some current methods and their performance is given in Tompa et al. 2005. Since proteins interact with the DNA in a structure-dependent way and any indel completely displaces the spatial configuration of the bases, there are no gaps allowed in the motifs (compare Tab. 2.5 and Fig. 2.6). However, transcription factors often form dimers which bind to the DNA like a clothespin, with two specific interaction points separated by a stretch of nucleotides that need not be conserved (Fig. 2.7). The partial motifs can be found as separate patterns but may be too short and have to be re-combined by specifying a spacer of certain length. The dyad analysis tool (van Helden et al. 2000), which is part of the Regulatory Sequence Analysis Tools suite[32], is optimized to find pairs of conserved 3-mers spaced by a non-conserved region of fixed width.

Two methods were applied in chapter 3.3. *EpiGraph*[33] (Bock et al. 2009) considers all 4-mers including the reverse complement and statistically checks for overrepresentation in each sequence. However, 4-mers are likely not very informative as there are only $4^4 = 256$ possible 4-mers, each of which is statistically expected to occur once in 256 bp. Nevertheless, overrepresented ones may give a clue about the composition of the sequences. *K-Factor* (Lee et al. 2007) determines the

[32] http://rsat.ulb.ac.be/rsat/

[33] http://epigraph.mpi-inf.mpg.de/WebGRAPH/

normalized frequencies of all *k*-mers in the input and background sets as well as the numbers of sequences that contain them. From these values it calculates an enrichment score and a Z score that incorporates average and standard deviation of the frequencies. The program can use genome-wide precomputed values for comparison. Alternatively, it allows to input a control set. It reports the *k*-mers in descending score order for each set but does not offer a method to determine which of them are significantly overrepresented in which set. The recommended *k* is 6 since smaller *k*-mers are rather uninformative and larger ones increase the computational effort.

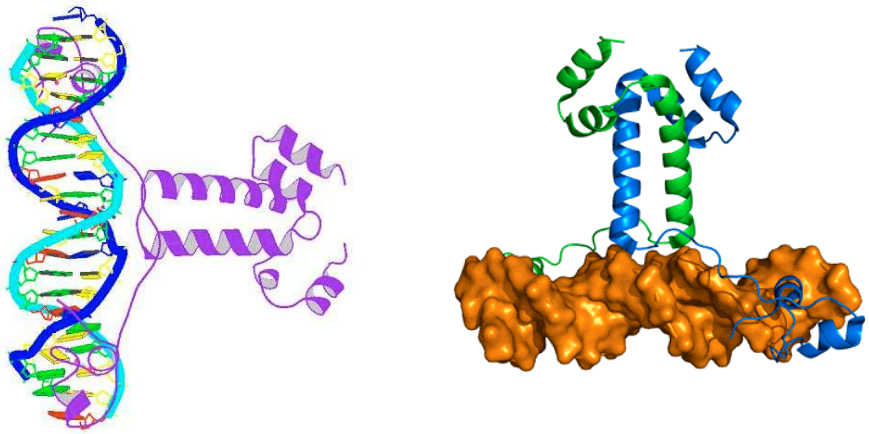


Figure 2.7: Dimer of the transcription factor Gal4 binding to DNA
Transcription factor dimers bind to the DNA at so-called dyads, i.e. short conserved sequence pairs or half-sites separated by a nonconserved spacer region. For homodimers, these are palindromic sequences. The consensus sequence of the depicted yeast Gal4 homodimer (pdb ID 3coq, A) binding site is $CGG(N)_{11}CCG$. The interaction of Gal4 with the DNA is mediated through the coordination of zinc atoms with cysteine residues (Pan and Coleman 1990), similar to zinc finger proteins, which can also use histidine. Homodimers are formed by coiled-coil and helical bundle superstructures (Hong et al. 2008, B). Another possibility for dimerization are interacting basic leucine zipper domains, amphipatic alpha helices in which every 7th residue is a leucine.

A completely different approach is to find motifs as local multiple alignments with heuristic algorithms for unsupervised learning: Expectation Maximization, as done by *MEME*[34] (Bailey and Elkan 1994, 1995a, 1995b), and Gibbs sampling (Lawrence et al. 1993), as implemented in *AlignACE* (Hughes et al. 2000). Both use essentially the same strategy which is also applicable for amino acid sequences and consists of two alternating steps. First, preliminary residue frequency matrices (also called position-specific scoring matrices, short PSSMs) for the motif and the background are estimated. This is done by regarding every *k*-mer substring of either all sequences or a randomly selected one as a possible instance of the motif and from this deriving a motif matrix by including pseudocounts to avoid zero entries. The background model is estimated from the rest or by a Markov Model. The E-step of expectation maximization corresponds to the sampling step of Gibbs sampling. Given the estimated matrices, calculate the probability for each substring to

[34] http://meme.sdsc.edu/meme4/

have been generated by the motif or the background model, respectively. Choose the one with the maximal likelihood for being a motif (compare section 2.4.3, log-odds ratio). In the following M-step, which is equal to the predictive update step, the matrices are newly estimated based on the preliminary motif. The two steps are repeated until convergence. How to calculate the likelihood and estimates is the main difference between the two algorithms. Both of them perform well on poorly conserved motifs that escape detection by word-based methods and are able to find multiple motifs simultaneously. Gibbs sampling has linear runtime but needs the exact motif width as an input parameter and is limited to the detection of motifs that occur at least once per sequence. *MEME* has quadratic runtime but can find motifs with variable numbers of occurrences and in a range of sizes (both between 2 and 300). The number of occurrences per sequence can be chosen as one, zero or one, or any number and influences the minimal and maximal number of sites. The huge parameter space has been expanded further by considering different background models, special options for palindromes, and the possibility of giving weights to individual sequences.

Since the algorithms are heuristic, different runs on the same data can yield different results and there is the problem of getting stuck in a local maximum instead of finding the global one, which would be the true motif. Although *MEME* reports *E*-values for the motifs (given in alignment form), manual inspection reveals that unspecific motifs with low *E*-values can be found for any set of sequences and that they could often be improved by shifting or extending the alignments. Whereas highly conserved patterns are readily detected, it is almost impossible to tell poorly conserved ones from noise. The authors recommend to try different parameters, or shuffle the nucleotides in the input sequence before repeating the analysis with the same parameters, to assess the significance of the obtained motifs. However, as this requires manual inspection, *MEME* is not suited for large-scale analyses. For biologists interested in finding transcription factor binding sites, there are web-based applications that determine potential ones by scanning sequences with known matrices and reporting those that belong to conserved sites in multiple alignments, for instance *rVISTA*[35] (Loots et al. 2002) and the tools at *DCODE*[36] (Loots and Ovcharenko 2005), where pre-compiled data are available as well.

It must be kept in mind that all motif search programs should report similar motifs if there are somewhat conserved patterns, which one would assume for promoter regions of genes that are regulated by the same transcription factors. More or less random input sequences are not expected to show such similarities, as I found it to be the case for computationally identified CpG islands from imprinted genes (compare chapter 3.2). If previously unknown patterns are found, they should give a clue for experimental studies to confirm their functional importance. Combinations of TFBSs, so-called cis-regulatory modules, have become of increasing interest as they regulate gene expression at metazoan promoters (Wasserman and Sandelin 2004, Blanchette et al. 2006) and can serve to predict enhancers (Pennacchio et al. 2007). Nevertheless, the performance of the methods is still rather poor due to the noise-creating complexity of vertebrate genomes. Additionally, comparative genomics has some limitations because of TFBSs turnover (Dermitzakis and Clark 2002, Frith et al. 2006), alternative promoters (Carninci et al. 2006, Baek et al. 2007), and divergent expression profiles between primates and rodents (Liao and Zhang 2006, Steinhoff et al. 2009).

[35] http://www-gsd.lbl.gov/vista/

[36] http://www.dcode.org

2.7 Homology and evolution

2.7.1 Orthologs and paralogs

Homologous genes are believed to have derived from a common ancestor gene by speciation or gene duplication (Fig. 2.8). A homologous gene with the same function in different species is an ortholog. In contrast, paralogous genes are found in the same organism; they can diverge considerably to assume new functions but may also be conserved (Studer and Robinson-Rechavi 2009).

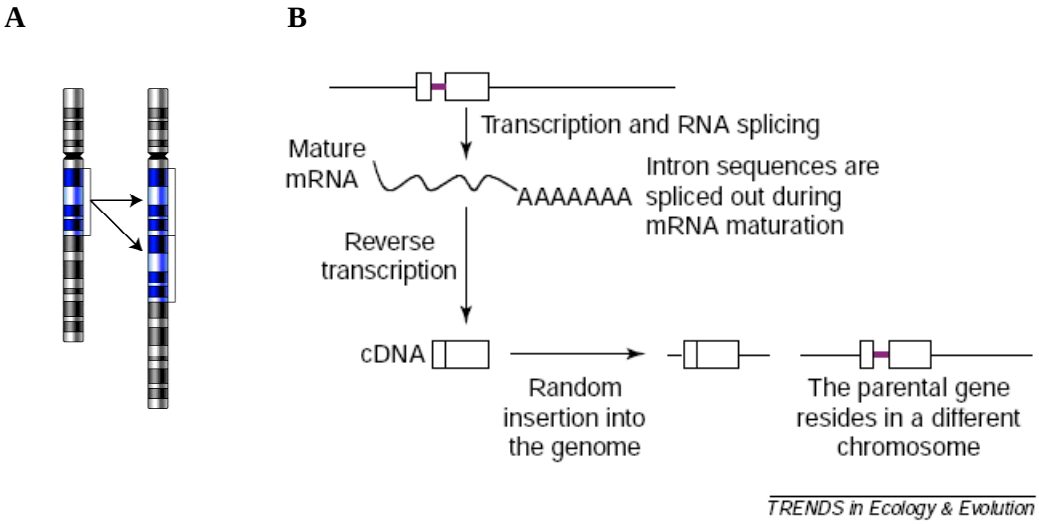


Figure 2.8: Origin of duplicated genes
Paralogous genes share a common ancestor. They can arise by chromosomal or whole genome duplications, from segmental duplication events that can be mediated by unequal crossing over (A, figure from Wikipedia), or via retroposition (B, figure taken from Zhang 2003). Retrogenes normally lack introns and the original regulatory regions since they are reverse transcribed from mature mRNA. They may, however, carry downstream promoter sequences associated with alternative transcriptional start sites or they recruit existing CpG islands at the new genomic location they are integrated into (Kaessmann et al. 2009).

Due to the high degree of sequence duplications in mammalian genomes it is difficult to distinguish orthologs from interspecies paralogs. Strategies include best reciprocal *Blast* hits, synteny, and Ks and Ka ratios of the cDNAs (see below) but are not failsafe. The Ensembl homology prediction method[37] works on the longest translation of each gene in the genomes of all species. Via clustering into gene families, multiple alignments of the protein sequences and phylogenetic trees derived from the corresponding cDNAs, pairwise relations of orthology and paralogy are inferred. Paralogs are sorted by taxonomy level, which means that the one reported first in the list is the most recent. Usually it corresponds to the best one in terms of identity of both the query and the target sequence, therefore we chose this hit as the representative one for the analyses presented in chapter 3.4.

[37] http://www.ensembl.org/info/docs/compara/homology_method.html

A similar method is used by HomoloGene at NCBI[38] for identifying homologs, the difference being that proteins from the RefSeq database are used directly and the one producing the best alignment is chosen. Close paralogs are sometimes included in the groups that otherwise consist of orthologous sequences.

2.7.2 Estimation of selection

When analyzing homologous genes, it is often interesting to know which kind of selection has acted on them in order to get a clue about the evolution of their function. Both orthologs and paralogs can substantially diverge on the DNA sequence level but still encode highly similar proteins. On the other hand, the exchange of a single nucleotide might result in the replacement of a functionally important amino acid. A commonly used method for estimating selection is estimating the rate of synonymous (Ks, also called dS) and nonsynonymous substitutions (Ka or dN) per site in alignments of coding DNA, Ka/Ks (or dN/dS, often abbreviated as ω). A Ka/Ks ratio < 1 signifies purifying selection whereas a value > 1 is indicative of Darwinian (positive) selection. For closely related species, high ratios are predominantly observed for transcription factors and genes related to reproduction, olfaction, and the immune system (Gibbs et al. 2004, Mikkelsen et al. 2007a). They contribute to speciation by positive diversifying selection. For distant species, evolutionary patterns are obfuscated: Episodes of initial Darwinian selection are followed by strict purifying selection and synonymous substitutions accumulate faster than nonsynonymous ones so that with increasing divergence, Ka/Ks decreases. Ks estimates tend to become unreliable because of so-called saturation, which occurs when the third codon positions are more than 40% different. In general, a range of 0-2 is reasonable for Ks (Z. Yang, pers. comm.).

With the exception of methionine and tryptophan, all amino acids are encoded by more than one RNA triplet (Fig. 2.9). Thus, multiple substitutions can occur when transforming one codon into another. Consequently, there are several methods to calculate the differences per codon as weighed sums of the possible mutation steps and to convert these estimates into synonymous and nonsynonymous substitutions per site. Ka and Ks rates given in the HomoloGene database are calculated using the method of Nei and Gojobori (1986), which neglects weighting and estimates the number of substitutions per site with a Jukes-Cantor model.

PAML[39] (Yang 2007) is a software package for phylogenetic analysis by maximum likelihood. Among its applications are simulations of sequence evolution, reconstruction of ancestral sequences, estimation of species divergence times, and estimation of synonymous and nonsynonymous substitution rates. The latter application, implemented in the program *codeml*, was used in section 3.4. The Ka/Ks ratio between two sequences is estimated by a maximum likelihood method that includes transition/transversion ratios and codon frequencies (Yang and Nielsen 2000). An initial value for Ka/Ks, which is used to calculate the substitution rate per codon, is iteratively refined by maximizing the log likelihood function. This function involves a transition probability matrix derived from the substitution rate matrix which for each codon *i* gives the probability to become a specific codon *j* after time *t*. For simplification *t* is set equivalent to the sequence distance. Ka and Ks are calculated as counts of sites and differences and base frequencies at

[38] http://www.ncbi.nlm.nih.gov/HomoloGene/

[39] http://abacus.gene.ucl.ac.uk/software/paml.html

synonymous and nonsynonymous sites with corrections for multiple mutations at the same site.

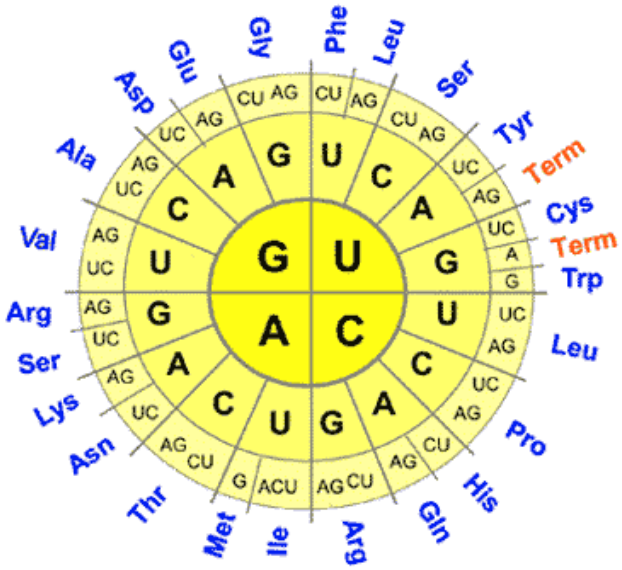


Figure 2.9: Codon sun
Due to the degeneracy of the genetic code, most amino acids are encoded by more than one mRNA triplet. The respective codon is determined by reading from the inner to outer circles. Note that with a few exceptions, the third codon position is not needed to specify the amino acid (fourfold degenerate site). The figure is taken from Wikipedia.

The input for *codeml* is an alignment of the coding parts of the cDNA, which was performed with the longest open reading frames using *transAlign*[40] (Bininda-Emonds 2005). This tool looks for the longest open reading frames, translates them into the corresponding amino acid sequences, calls *ClustalW* (Thompson et al. 2002, Larkin et al. 2007) on them and back-translates the resulting protein alignment into a DNA alignment by assigning the original codons to the amino acids. Thus, frameshifts due to gaps that are not multiples of three are avoided, which makes more biological sense most of the time, although manual inspection reveals that sometimes a frameshift would be more probable (Fig. 2.10). Codon-based DNA alignment also cannot take into account exceptions that seem to occur with elevated frequency in the well-studied imprinted genes: non-AUG (CUG) translation initiation site (*WT1*), selenocysteine encoded by the stop codon UGA (*DIO3*), and ribosomal frameshift which is also seen in retroviruses (*PEG10*). Although it is recommended to manually improve automatic alignments, this is of course not applicable for genome-wide analyses. *codeml* has an option to remove gap positions and ambiguous nucleotides (see Tab. 2.6) since there are no evolutionary models for insertions and deletions.

[40] http://www.personal.uni-jena.de/~b6biol2/ProgramsMain.html#Sequences

A

>human *COMMD1*

```
...agtatcagcacactgatcagccagcctaactga
... S  I  S  T  L  I  S  Q  P  N  *
```

>mouse *Commd1*

```
...agtatcaacaggctgatgcaggcagcctaa---[ctga]
... S  I  N  R  L  M  Q  A  A  *  -  [outside alignment]
```

>dog *COMMD1*

```
...agtatcagcacactgatgcagccagcctag[ctga]
... S  I  S  T  L  M  Q  P  A  *  [outside alignment]
```

B

>human *COMMD1* with G inserted

```
...agtatcagcacactgatgcagccagcctaa[ctga]
... S  I  S  T  L  M  Q  A  A  *  [outside alignment]
```

>mouse *Commd1* with G deleted

```
...agtatcaacaggctgatcaggcagcctaactga
... S  I  N  R  L  I  A  Q  P  N  *
```

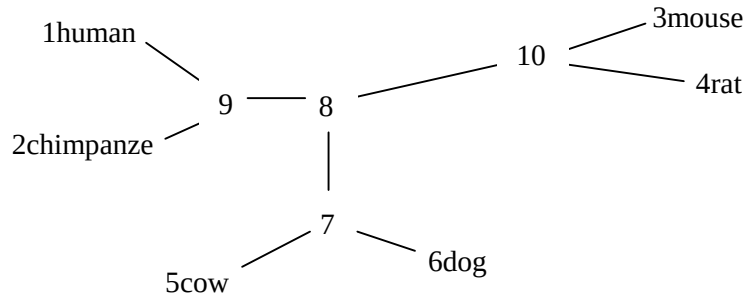
Figure 2.10: Amino acid guided cDNA alignment and frameshift
(A) Alignment of three sequences in multiFasta format. Translated amino acids are centered below their codons. * denotes the stop codon. "outside alignment" means that the corresponding noncoding part of the cDNA was not taken into account. Note that identical amino acids, such as the glutamine (Q), are not always aligned because a mismatch is "cheaper" than inserting a gap. It seems likely that at the underlined positions, insertion or deletion of a single base pair has occurred instead of that of a complete codon. The dog sequence favors the deletion of guanine in the human lineage.
(B) Insertion of a single guanine (underlined) in the human *COMMD1* sequence or deletion of a single guanine (between the two underlined bases) in mouse *Commd1* would induce a frameshift and produce a sequence that can be better aligned to its ortholog in (B).

codeml can estimate pairwise Ka and Ks rates between all species in an alignment as well as lineage-specific ones. The latter application is of interest to test the hypothesis of selection in specific lineages or branches, therefore the applied methods are called branch models (Yang 1998). They require a phylogenetic tree in which the branch(es) in question is/are marked. The one-ratio model assigns the same Ka/Ks ratio to all branches whereas a two-ratios model estimates a ratio for the marked branch that is different from the background. It is possible to assign individual Ka/Ks ratios to all branches, however the increase in parameter space makes the estimations unreliable. A modification of the two-ratios model additionally allows to fix Ka/Ks in the branch of interest to a specific value, e.g. 1. Ka and Ks are then estimated in a way to fulfill the given Ka/Ks.

Using a two-ratios model with the tree in figure 2.11, a different Ka/Ks ratio is assigned to the branch marked with #1 that leads from the euarchontoglires ancestor (8) to the rodent one (10). All

other ratios are kept fixed. *codeml* reports Ka and Ks rates and Ka/Ks ratios for all adjacent sequence pairs, e.g. rodent ancestor (10) and mouse (3), or mammalian ancestor (7) and euarchontoglires ancestor (8), or primate ancestor (9) and human (1).

A



B

```
(((1human, 2chimpanzee), (3mouse, 4rat)#1), 5cow, 6dog)
```

Figure 2.11: Phylogenetic tree of six mammalian species
A dendrogram drawing (A) displays a phylogenetic tree in a human-readable way; parenthesis notation (B) is used by computer programs. Ancestral nodes can be represented by reconstructed sequences. Note that this is an unrooted tree and that there is a trifurcation at the mammalian ancestor (7) that leads to three lineages. Branch lengths, which correspond to the evolutionary distance and are usually included in the tree, are approximately to scale in this example. Although primates are phylogenetically close to rodents, together forming the euarchontoglires, sequence identity is generally lower between them than between primates and cow or dog.

For the genome-wide analysis performed in chapter 3.4, human, mouse, rat, and cow sequences given as accession numbers in the HomoloGene XML file were retrieved from the NCBI web site with Entrez Programming Utilities and were aligned with *transAlign*. Branch models were constructed with *codeml* for each alignment with the unrooted tree (human, (mouse, rat)#1, cow). Genes with Ks = 0 and Ks > 2.5, which is a result of saturation, were omitted from further analyses because these data are unreliable. To see if the two-ratios model provided a better fit than the one-ratio model – in our example, if the Ka/Ks ratio in the rodent ancestor branch was significantly higher than that of the other lineages –, twice the difference of the two reported log likelihood ratios was compared to a χ^2 distribution with one degree of freedom (see section 2.9.1). The critical value of ≥ 2.71 must be reached for a significance level of $p < 0.05$. Comparing the log likelihood ratios of the two-ratios model with and without fixed Ka/Ks likewise results in the probability that the Ka/Ks ratio is significantly different from 1. If this is the case, one can infer at least relaxed constraints, if not positive selection, on the rodent ancestor's gene.

2.8 Custom Perl scripts

Perl[41] is a scripting language of invaluable utility. Easy file handling, string processing, and the existence of regular expressions make it especially suitable for bioinformatics applications that involve DNA sequences. I found the only drawback in inaccuracy when adding small fractions to large numbers (e.g. when summing up CpG contents for a set of sequences to calculated the mean). I used Perl extensively for writing a large variety of custom scripts: automatically call programs on files from a folder, parse their outputs, create lists of genes, their exons and coding regions from GenBank files or UCSC tables, extract DNA sequences from GenBank files, reverse complement them, extract substrings, convert lower case letters to Ns, find given motifs and patterns, calculate (di)nucleotide frequencies or fractions of something of interest, construct the longest open reading frame from a cDNA, extract certain lines from a file (e.g. those that have the name of an imprinted gene in them), process alignments, sort and select desired data from whatever format, calculate distances and find overlaps, create random data sets, conduct statistical tests (see section 2.9), and many more. To this collection student research assistant Matthias Bieg contributed a versatile script capable of efficiently calculating overlaps between two large sets of genomic regions given by chromosome coordinates using binary search (see below, section 2.8.2) as well as other scripts to parse the HomoloGene XML data set (chapter 3.4) and to work with the UCSC database.

It would be too tedious to describe all these useful scripts that are sometimes not even mentioned in the publications because they seem trivial for bioinformaticians. I will, however, elaborate on the algorithms underlying two more advanced ones.

2.8.1 Merging transcript variants into genes

UCSC RefSeq genes are listed as all their transcriptional variants in the refGene table, which is not particularly useful if one would like to download the genomic sequence of the whole gene. To construct the longest possible transcribed region, as it is done in other databases, it is necessary to find the most upstream transcriptional start site and the most downstream transcriptional termination site as well as all the exons. The latter may differ in length due to alternative splicing, in which case the longest version has to be determined by merging.

Genes that partially overlap with their neighbors or reside inside introns of other genes are easy to handle by taking the gene name as an identifier. On the other hand, genes that have the same name (and even the same RefSeq identifier) but are located on either different chromosomes or the same one, but without an overlap of their transcripts, have to be differentiated from actual alternative transcripts. In some rare cases, the directions of transcription do not match, in others, two non-overlapping transcripts can be merged via a third one that overlaps with both. The Perl script, for which the algorithm is given in pseudocode below, considers all those possibilities. It generates output in *PipMaker* format (Fig. 2.12).

[41] http://www.perl.org

Listing 2.1: Pseudocode for merging transcript variants into genes

```
for each current transcript in the file
{
   # Is there already an entry with the same name in the gene list?
   # Then check if they both belong to the same gene because
   # sometimes the two transcripts do not overlap
   # and some gene names appear on different chromosomes (paralogs).
   # Entries for each gene have their start coordinates as a unique key
   if (current_name exists in gene_list)
   {
        problem = true;
        for all entries in gene_list where (current_name == entry_name)
        # (simplified in so far that, if there is more than one entry, the names have a
number
        appended; see below)
        {
               # Is there is an overlap of the start and end coordinates?
               # If yes, it is a new version of the gene and they can be merged.
               if (current_chrom == old_variant_chrom && overlap((current, old_variant)))
               {
                      problem = false;
                      merged++;
                      merge_variants(current, old_variant);
                      # That subroutine updates the start and end coordinates and merges
                      # exons of the old and new versions to construct the longest ones.
                      # Additional exons can be gained as well. The procedure is
                      # similar to the present one for transcripts.
               }
               # current may overlap with several old entries, thereby merging them.
               # Thus, update counter of occurrences of non-overlapping
               # transcripts that have the same name
               if (merged > 1)
               {
                      counterhash(current_name)--;
               }
        }
        # if there is no overlap, the problem remains
   }
   else
   # the name does not exist yet
   {
        # make a new entry in the gene list
        new_gene(current, current_name);
   }
   if (problem == true)
   {
        counterhash(current_name)++;
        # make a new entry, distinguish it by appending a number to its name:
        number = lookup(counterhash, current_name);
        new_gene(current, current_name."_".number);
   }
}
```

2.8.2 Calculating overlaps with binary search

It is frequently necessary to calculate the intersection of two data sets: to compare CpG island reported by two different programs to see how many were detected by both of them, or to link regions given as genomic coordinates, for instance conserved elements with genes to find out which of them coincide with promoter regions. One of the data sets represents the fixed intervals to which hits found in the other list are appended for the output. Table 2.7 shows an example in which conserved regions are fixed and locations relative to imprinted genes are appended. If start or end coordinates were the same in both data sets, the task would be easily done by comparison of two sorted lists. However, this is not the case as most regions do not exactly coincide but overlap in parts. Comparing each entry of one list with each in the other to see if there is an overlap of at least

1 bp causes a runtime of $O(n^2)$. While this is still well practicable for a few dozen intervals like the CpG islands of a gene, it becomes very time-consuming and thus nearly infeasible for 10,000s of entries.

The solution is binary search on the sorted lists per chromosome. Binary search is a divide-and-conquer algorithm with logarithmic runtime, $O(\log n)$. First, both lists are sorted in ascending order. An entry *i* from the fixed list is chosen and compared to the entry in the middle of the second list, *m*. In our case, this means checking for an overlap of the two intervals $[start_i, end_i]$ and $[start_m, end_m]$. If there is a match, it is appended. If the middle entry is smaller than the sought one ($end_m < start_i$), matches can only be found in the right half of the second list. Otherwise, if it is greater ($start_m > end_i$), we expect matches only in the left half. Therefore, the middle element of the respective half is chosen for the next comparison. Search continues in this way until either a match has been found or no overlap could be determined for the only remaining middle value. In case that there are overlaps between the entries of the second list, which often happens for genes, the vicinity of a hit interval is also checked for overlaps. Since the lists must be sorted and search is done for each entry on the fixed interval list, the total runtime of the algorithm is $O(n \log n)$, considerably faster than the quadratic one before.

```
> 58954 59871 OR4F5
+ 58954 59871
58954 59871

> 357522 358460 OR4F29
+ 357522 358460
357522 358460

> 357522 358458 OR4F3
+ 357522 358458
357522 358458

> 357522 358458 OR4F16
+ 357522 358458
357522 358458

< 610959 611897 OR4F29_1
+ 610959 611897
610959 611897
```

```
...

> 850984 869825 SAMD11
+ 851185 869396
850984 851043
851165 851256
855398 855579
856282 856332
861015 861139
864283 864372
864518 864703
866387 866549
867379 867494
867653 867731
867802 868301
868496 868620
868941 869051
869151 869825

...
```

Figure 2.12: List of human genes in *PipMaker* format
All coordinates are 1-based. Direction of transcription is indicated by > for the forward strand (+), and < for the reverse complementary strand (–). The lines beginning with + show the extent of the coding region, all others represent exon intervals. Non-overlapping genes that share the same name are distinguished by appending an _ and the counter for the number of additional occurrences.

Table 2.7: Example of overlap script output

chrom	chromStart	chromEnd	score	length	location[a]	gene	gene chr	geneStart	geneEnd	strand
chr2	80384553	80385083	745	530	promoter	LRRTM1	chr2	80382513	80384998	–
chr7	129720208	129720229	320	21	promoter	CPA4	chr7	129720229	129751249	+

[a] promoter was defined as overlap with the most upstream transcriptional start site of the gene.

2.9 Statistical Tests

Descriptive statistics pools and represents numerical data whereas statistical analysis makes conclusions based on the observations. If data are pooled into size categories and plotted as frequencies, this results in a histogram for their distribution (Fig. 2.13). Normal distributions are described by their mean (average) μ and standard deviation (std.dev.) σ. The mean partitions the values into a lower and an upper half; the standard deviation (calculation see section 2.9.2) describes the average deviation from it. Approximately 95% of the data lie inside the twofold standard deviation area around the mean.

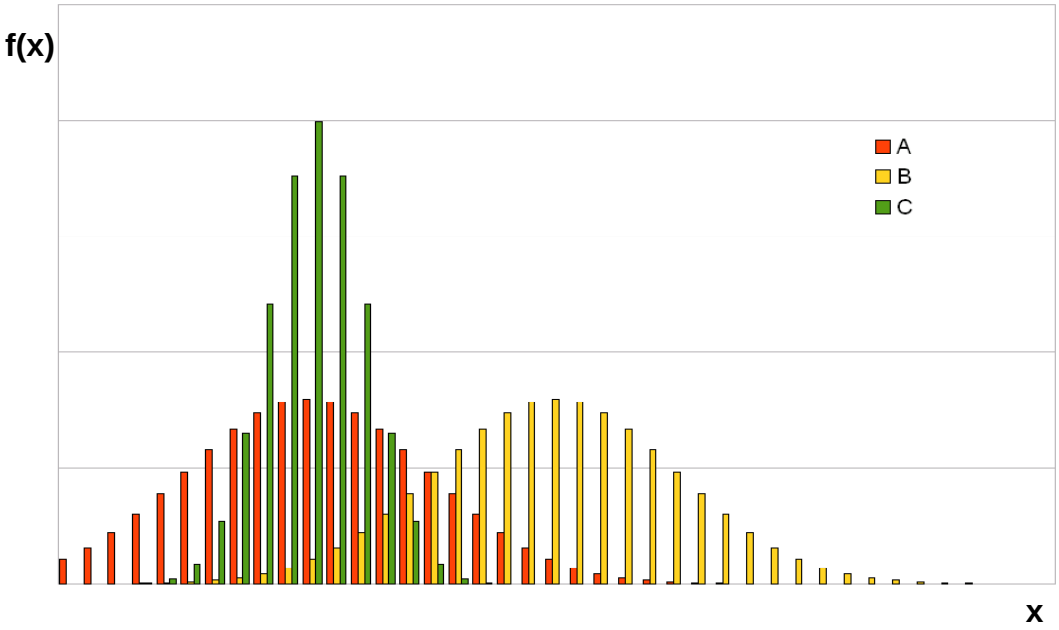


Figure 2.13: Differently shaped distributions
A normal distribution corresponds to a Gaussian curve with a bell shape determined by its mean μ and standard deviation σ:

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$$

(A) μ = 10, σ = 5; (B) μ = 20, σ = 5; (C) μ = 10, σ = 3
A and B would represent different populations with significantly different means.

For non-Gaussian distributions, the appropriate partitioning into a lower and an upper half is done by the median, which is found at the middle position in a list of the sorted values. (Of course, for normal distributions the median is equal to the mean.) Instead of the standard deviation, percentiles are given, e.g. the 25th percentile (lower quartile) as the value below which 25% of all values lie. The 75th percentile is likewise called the upper quartile. Quartiles are depicted in boxplots (also called box-and-whisker plots; Fig. 2.14).

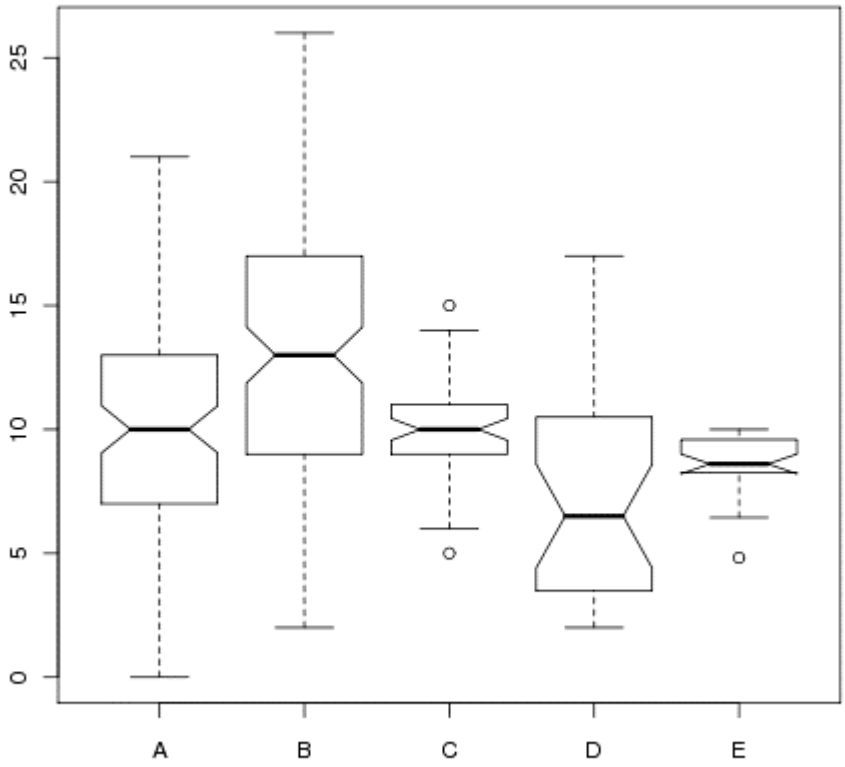


Figure 2.14: Boxplots
The box extends from the 25th to the 75th percentile with a line representing the median. The interquartile range (IQR) is the difference between the 75th and the 25th percentile. Data that lie more than 1.5 · IQR below the lower quartile or above than the upper quartile, respectively, are labeled as outliers and represented as small circles (C, E). Whiskers commonly extend from the box ends to the smallest or highest value that is not an outlier, but sometimes, with a different definition of outliers, represent other percentiles or even the minimum and maximum. Constricting the box with notches is optional. Notches extend to +- 1.58 · IQR/√*n* (with *n* being the number of data points), giving sort of a 95% confidence interval around the median. If the notches of two plots do not overlap, this is a hint for a significant difference of the medians, e.g. in the case of A and B. Sometimes, if *n* is small, notches go outside of the box so that the rims are folded up (E). Despite their usefulness, especially for skewed distributions (D, E), boxplots are not available within Excel or gnuplot, but in R[42]. For normally distributed data, boxplots are symmetric (A, B, C).

[42] http://www.r-project.org

The so-called null hypothesis (H_0) says that two distributions are essentially equal and the observed difference between them is just caused by random fluctuations, so that both groups belong to one population (or two populations that cannot be distinguished via the measured feature). The alternative hypothesis (H_A) claims that the difference is due to a really different distribution of the values in the two groups, which therefore represent samples from two different populations.

Statistical tests serve to calculate a test statistics that has a certain distribution if H_0 is true. As figure 2.15 shows, the probability of obtaining a certain value of the test statistics is calculated as the integral between this value and the "end of the curve" and then mapped into a table. One then needs to check whether the probability of the observed result, i.e. the value *t* (in German: Prüfgröße) reported by the test applied to our data, is at least as high as a critical value t_{crit}. If it is not, this means that the probability of a real difference is low and one should reject H_A in favor of H_0. Else, it is appropriate to reject H_0 in favor of H_A.

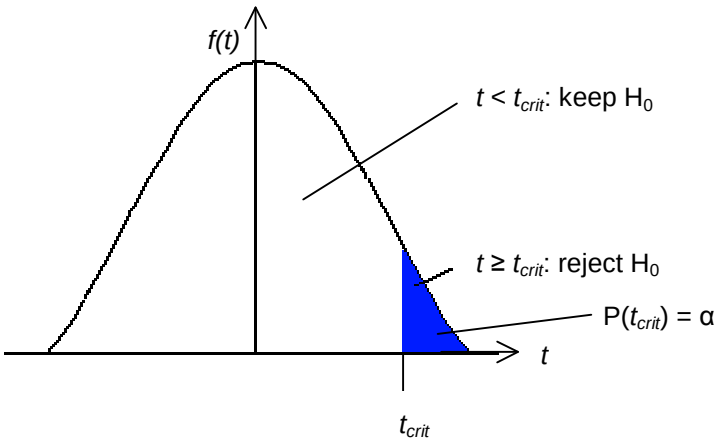


Figure 2.15: Test statistics
The probability of obtaining a test statistics t_{crit} is equal to the error probability α, indicated by the blue area. H_0 is rejected in favor of H_A if the obtained value $t \geq t_{crit}$. If $t < t_{crit}$, H_0 is kept. For a two-tailed test, α is divided between the two tails of the curve instead of being applied to the right tail only; hence, two-tailed tests require smaller error probabilities.

t_{crit} can be determined beforehand by fixing a so-called error probability α, for biological data commonly 5%. Alternatively, *t* can be looked up in the appropriate table to determine the associated p value, which is given in decimals. Thus, a difference yielding a *t* with a p value of 0.05 or less is called "significant" (on the 5% level). The lower the p value, the less likely the differences are just random. However, as it is just probabilities, the decision may still be incorrect. To falsely reject H_0 is called error of the first kind. To falsely keep H_0 is an error of the second kind; this is likely to happen if (for fear of an error of the first kind) the significance level is chosen too strictly. Often, $0.05 < p < 0.1$ is called a trend or tendency to indicate that the result is not highly significant but the two samples still look different enough that calling them similar is not justified.

If *m* different features on the same data set serve for statistical comparison (multiple hypothesis testing), the threshold of each test must be adjusted to prevent accumulation of first kind errors. The classical method is the Bonferroni one, where the threshold of each test is chosen as $\alpha' \leq \alpha/m$, where α is the total significance level. It can be easily seen that the larger *m*, the lower the probability to reject one of the null hypotheses. A slightly less conservative alternative is the

Bonferroni-Holm method. Here, the individual p values are sorted in ascending order and compared to their respective α_i, where $\alpha_i = \alpha/(m - i + 1)$. H_0 must be kept as soon as a p value becomes larger than its α_i. The false discovery rate (FDR) is suited for large data sets like microarray data.

p values become smaller with increasing size of the data sets since the degrees of freedom enlarge with the sample sizes, which in turn have mayor influence on the calculation of the test statistics. For large sample sizes, the distributions of the test statistics approximate a normal distribution. Most statistical tests are optimized for samples from a normal distribution, as shown by biological populations. There are also statistical tests for testing whether a sample has a normal distribution. Not normally distributed data can be transformed, e.g. by taking their logarithms. However, this might not always make biological sense.

2.9.1 Chi-square test

The χ^2 (chi-square) test is a nonparametric test as it does not assume the data to have a normal distribution. Table 2.8 shows the example of a classical fourfold test or cross tab, for which the calculations are given below. It is applied for determining if a certain feature is significantly enriched or decreased in one group relatively to another group.

Table 2.8: Cross tab for fourfold test

group	feature present (S)	feature absent (F)	row sum
A	*a*	*b*	$a + b = N_A$
B	*c*	*d*	$c + d = N_B$
column sum	$a + c = N_S$	$b + d = N_F$	$N = a + b + c + d$

N is the sample size. The test statistics, which obeys a χ^2 distribution with one degree of freedom, is calculated as follows:

$$\chi^2 = \frac{N \cdot (a \cdot d - b \cdot c)^2}{N_s \cdot N_F \cdot N_A \cdot N_B}$$

This χ^2 value is then compared to the table to infer its p value. For example, $\chi^2 > 3.841$ is significant on the 5% level. A high value corresponds to a small error probability. For small data sets ($N < 30$), it is appropriate to apply a Yates' correction to avoid rejection of H_0 (equal proportions of the feature between the two groups). It reduces χ^2 and thus increases its p value. The formula changes into:

$$\chi^2 = \frac{N \cdot (\left| a \cdot d - b \cdot c \right| - \frac{N}{2})^2}{N_s \cdot N_F \cdot N_A \cdot N_B}$$

If the sample size is very small (one of the four possible combinations of row sum · column sum/*N* is below five), one should use Fisher's exact test. The χ^2 test can also be expanded to more than two categories (data grouped in bins) and more than two groups. Then, the degrees of freedom are equal to (number of rows – 1) · (number of columns – 1). The higher the number of the categories, the higher the χ^2 value must become to obtain a low p value. I implemented both the fourfold and the multiple χ^2 test as Perl scripts. Another application of the χ^2 test is to check if data obey a special distribution (normal, uniform, ...). In this case, group B is assigned the expected frequencies for each category. An alternative to χ^2 is the Kolmogorow-Smirnow test, which is however not very exact and was not used in the analyses.

2.9.2 t test

The Student's t test, named after the pen name of its developer, also called two-sample t test, is used to test if the means of two data sets are equal or if there is a location shift. The samples need to be (at least approximately) normally distributed because only then the mean and standard deviation are appropriate to describe their shapes and both are used in the calculation of the t value. Thus, the t test is a parametric test. It is very sensitive to outliers since they have a large influence on the mean. The t value shows a Student's (t) distribution with $n_A + n_B - 2$ degrees of freedom, where n_A and n_B are the sample sizes of set A and set B, respectively.

$$t = \frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}$$

where μ_A and μ_B are the means (averages) and σ^2_A and σ^2_B are the variances, calculated as the sum of squares divided by the sample size – 1:

$$\sigma^2 = \sum_{x=1}^{n} \frac{(x-\mu)^2}{n-1}$$

The square root of the variance σ^2 is the standard deviation σ. The original formula assumes equality of variances, but if this is not the case, there is a modification for pooled standard deviation. It requires the degrees of freedom to be calculated with the Welch-Satterthwaite equation. I implemented this variant in a Perl script.

Depending on whether or not there is a beforehand assumption which group's mean is greater, one applies the one-tailed test or the two-tailed test, which requires *t* to be greater for the same significance level (the error probability is halved to account for both ends of the curve). In practice, this just means looking up *t* in different parts of the same table. The F test is an extension for more than two groups. It yields significance if at least one of the groups is different from the others. In order to find out which one that is, all pairwise t tests have to be performed.

2.9.3 Wilcoxon test

The Wilcoxon two sample test, also called Mann-Whitney-Wilcoxon test, Wilcoxon rank-sum test, Mann-Whitney-U test, or short U test, is not to be confused with the Wilcoxon signed rang test that is applied for dependent groups. As it compares medians, it belongs to the nonparametric order statistics (German: Rangstatistik) and is an alternative to the t test if the distributions are not Gaussian. It is less sensitive to outliers but also slightly less likely to detect a location shift than the t test. The U test orders the values of the two data sets and assigns ranks to them: the smallest one gets rang 1, the next smallest rang 2, and so on. If two or more identical values occur (ties), ranks are counted on and the tied values are assigned either the average or the median rank.

To determine significance, the rank sum of the smaller sample (with n_a values) is compared to a distribution of rank sums from a pooled sample. Alternatively, a U value is calculated for either the smaller sample (one-tailed test) or both (two-tailed test) by

$$U_a = n_a \cdot n_b + ½ \cdot n_a \cdot (n_a + 1) - \text{rank\_sum(sample a)}$$

(analogous for sample b) which is then compared to the test statistics,

$$U = \frac{n_a \cdot n_b}{2} - u(\alpha) \cdot \sqrt{\frac{1}{12} n_a \cdot n_b \cdot (n_a + n_b + 1)}$$

where α is the desired significance level (α/2 for two-tailed test). $u(\alpha)$ is to be taken from a table. H_0 is rejected if the either U_a or U_b (or, for a two-tailed test, the smaller of them) is smaller than U. U can also be taken from tables. The extension to more than two groups is the Kruskal-Wallis test. Wilcoxon tests were performed with R and a script obtained by R. van Son[43].

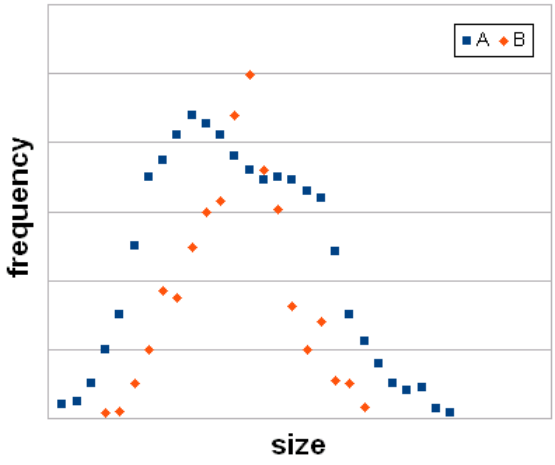


Figure 2.16: Differences in variation
If one of the samples (A) has a larger variation than the other (B), it receives both the highest and the lowest ranks so that its rank sum is similar to that of the less variable set. In this case, the result of the Wilcoxon test will not be significant. The difference between the groups then consists in the shapes of the distributions rather than a difference in the centers of location.

[43] http://www.fon.hum.uva.nl/rob/SignedRank/WlcxTest.pl

2.9.4 Correlation

If two features show interdependence, they are correlated. Their connection can be linear or more complex (Fig. 2.17). To determine the degree of correlation and its significance, there are three common methods. Their correlation coefficients each ranges between -1 (perfect negative or anticorrelation) and 1 (perfect correlation). If it is near 0, there is no correlation.

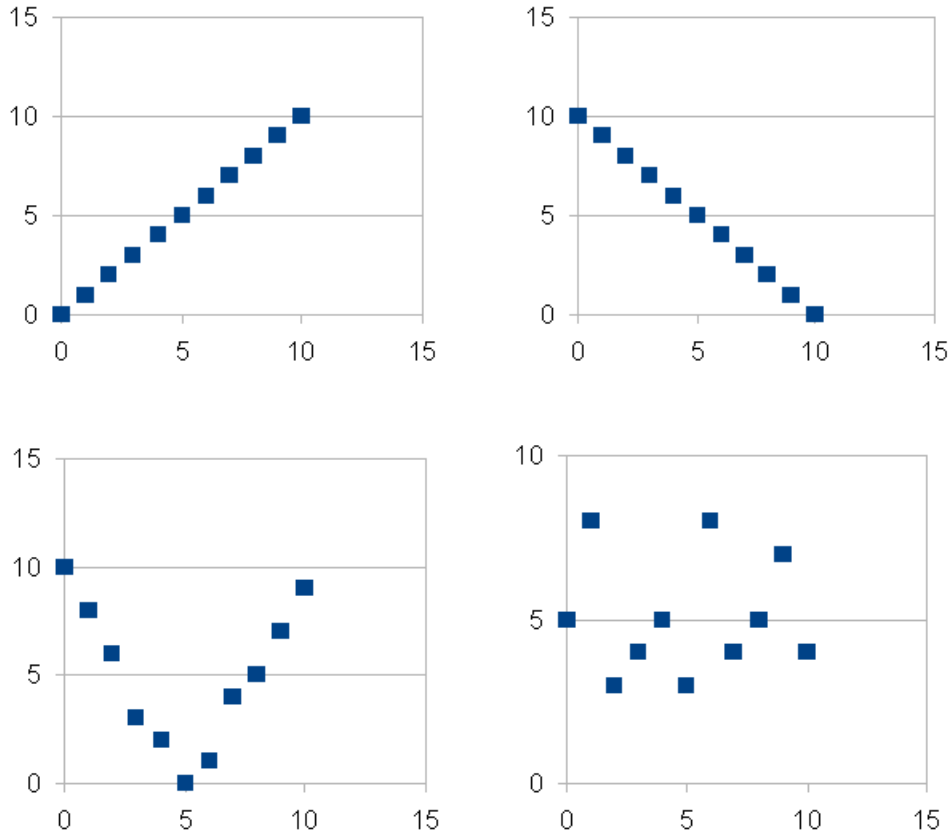


Figure 2.17: Correlations
Upper left: perfect linear correlation (r = 1). Upper right: prefect linear anticorrelation or negative correlation (r = -1). Lower left: non-linear (parabolic) correlation (r= 0.96, τ = ρ = 1). Lower right: no correlation. For the explanation of the different correlation coefficients see text.

The most commonly applied method is Pearson's correlation, abbreviated as r, which measures the strength of a linear connection:

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$$

where σ^2 is the variance and the covariance is defined as

$$\mathrm{cov}(x, y) = \sum_{i=1}^{n} \frac{(x_i - \mu_x) \cdot (y_i - \mu_y)}{n-1}$$

with n the number of data pairs, x_i and y_i individual values and μ_x and μ_y means. Sometimes, r is reported as its squared value, r^2. Significance is determined by calculating the t value as

$$t = r \cdot \sqrt{(n-2)} \cdot \sqrt{(1-r^2)}$$

that can be looked up in the t distribution table for n – 2 degrees of freedom. No significant linear correlation can mean that there is either no interdependence or at least not a linear one. Although low values like r = 0.2 can be highly significant if there are many data points, scatter plots often reveal a point cloud for which any biological interpretation of correlation is highly speculative.

Nonlinear correlation can be detected with the two other methods: Spearman's rank correlation, which performs Pearson's correlation on the ranks of the values instead of the values themselves, or Kendall's correlation, which compares all pairs with each other and therefore has a runtime of $O(n^2)$. Its coefficient τ is calculated from the numbers of concordant pairs (*x1 < x2* and *y1 < y1*, or *x1 > x2* and *y1 > y1*), discordant pairs (*x1 < x2* and *y1 > y2* or vice versa), and pairs in which either the *x* values (extra_x) or the *y* values (extra_y), but not both, are identical:

$$\tau = \frac{\text{concordant - discordant}}{\sqrt{(\text{concordant} + \text{discordant} + \text{extra\_x}) \cdot (\text{concordant} + \text{discordant} + \text{extra\_y})}}$$

The number of discordant pairs is known as the Kendall distance or bubble sort distance of two lists. It is equivalent to the number of swaps the bubble sort algorithm would make. The higher it is, the more dissimilar the values are. To make it comparable between different lists, it is normalized to the interval [-1,1]:

$$\text{dist} = 2 \cdot \frac{\text{discordant}}{n \cdot (n-1)}$$

In this work, both Pearson's and Kendall's correlation were computed with my own Perl scripts. For Spearman's correlation, I used a script from J. Karlgren[44].

[44] http://www.sics.se/~jussi/Vertyg/spearman.html

Chapter 3 – Results

The genomic sequence features of individual imprinted genes and imprinting clusters have been intensely investigated by various researchers. However, in these studies little attention has been paid to comparisons with non-imprinted genes. Moreover, the existing data might be biased because of taking regions on certain chromosome as controls or extracting data from resources that focus on specific groups of genes. Therefore, it often remains unclear whether the reported enrichment or depletion of features is really statistically significant. We tried to perform unbiased comparisons by randomly choosing several groups of control genes and applying the same procedures to them and the imprinted sets. First concentrated on human and mouse genomic sequences, the analyses were subsequently expanded to a genome-wide scale. The results presented here compile the work reported in two publications, two manuscripts in preparation, and additional findings. In the first section, we tested the performance of different methods for the identification of CpG islands in human and mouse (Hutter et al. 2009). Then, we investigated the relationship between imprinted genes and CpG islands and repetitive elements (Hutter et al. 2006). The last two sections explore sequence conservation on the levels of genomic DNA and protein-coding sequences in the context of imprinting, including potential effects of CpG deamination, substitution patterns, and paralogous genes.

3.1 Characteristics of human and mouse CpG islands

CpG islands (CGIs) are commonly regarded as epigenetic key regulatory elements. Intriguingly, differences between the CpG islands of human and mouse have been reported several times (Aïssani and Bernardi 1991, Antequera and Bird 1993, Matsuo et al. 1993, Cuadrado et al. 2001) but later on it was suggested that the genomic G+C distribution pattern has effects on the commonly used algorithms (Waterston et al. 2002). Therefore, we applied several methods for computational identification of CGIs on three sets of orthologous genomic sequences from the two species. Truly species-specific differences should be indicated by consistency between all three groups. We also investigated the influence of repetitive elements on CGI detection since interspersed repeats, namely SINEs, often fulfill the CpG islands criteria but do not obey the definition of regulatory elements (Jabbari et al. 1997, Lander et al. 2001, Ponger et al. 2001, Takai and Jones 2002, Oei et al. 2004). The study, of which an abbreviated version is presented here, was published in Hutter et al. (2009).

3.1.1 Effects of different algorithms and repetitive sequences on CpG island identification

Two groups of 79 randomly selected autosomal genes each were collected by converting random integer numbers into accession numbers for the NCBI RefSeq nucleotide database. If the corresponding entry was from human or mouse and possessed an orthologous gene in the other species, the gene pair was included under the condition that both genomic sequences were available at the NCBI Map Viewer. Numbers for the first group (G1) cover a range from 1 to 300,000; for the second group (G2), random numbers were restricted to a range from 1 to 16,000. A third group (G3) contains 79 human genes randomly taken from the UCSC hg18 RefSeq annotation and their murine orthologs in mm8 as determined with BioMart at Ensembl. The entire genomic sequence of each gene including 10 kb upstream of the transcriptional start site and 10 kb downstream of the end of the last exon was downloaded from MapViewer human build 35.1 and mouse build 33.1 (G1, G2), or UCSC hg18 and mm9 (G3), in the direction of transcription. Since the assembly of the mouse genome used for this study is still not finished, some mouse sequences contain stretches of

undefined nucleotides (represented as stretches of Ns). The murine sequences of genes containing over 5% of undefined nucleotides were replaced by improved assemblies of build 34.1. The locations of the genomic sequences with respect to contigs, their G+C and CpG contents, and chromosomal origins are given in Appendix A Tab. A1.

Gene lengths are similarly distributed throughout the groups; murine genes are insignificantly shorter than human ones (Wilcoxon test, $p > 0.4$). CpG and G+C content are highly positively correlated (Pearson's r between 0.71 and 0.94, $p < 0.0001$). The correlation is always higher in human sequences than in mouse sequences. All groups show an overall similar tendency towards a bimodal distribution of G+C contents. However, the G+C content of human sequences is more variable than that of mouse sequences, indicating that the G+C content of some human sequences is more extreme, consistent with published data (Waterston et al. 2002). Human sequences have a mean CpG content of 1.71 ± 0.86% and are thus significantly enriched in CpG compared to mouse with 1.33 ± 0.49% (t test, $p < 0.001$), reflecting the the more pronounced CpG scarcity in the mouse genome (Waterston et al. 2002, Zhao and Zhang 2006a, 2006b). Note that one should rather speak of a smaller depletion in human as the CpG content is in both cases much lower than that of other dinucleotides.

For identifying CGIs in the genomic sequences, we first applied the traditional sliding window methods implemented in the *CpG Island Searcher* (Takai and Jones 2002) using two widely known different parameter sets: The Gardiner-Garden and Frommer (1987) criteria (GGF) detect CGIs of at least 200 bp length with G+C content ≥ 50%, and $CpG_{obs}/CpG_{exp} \geq 0.6$. In contrast, with the more stringent criteria from Takai and Jones (2002), abbreviated as TJ, only stronger CpGIs of at least 500 bp length with G+C content ≥ 55%, and $CpG_{obs}/CpG_{exp} \geq 0.65$ are reported. Additionally, we identified CGIs with three alternative methods. Two programs implement segmentation methods that divide sequences into CpG-depleted and CpG-enriched segments based on the CpG content in two adjacent windows. A CGI segment from *cpg* (Li et al. 2002) has to fulfill CpG content ≥ 3.5% and length ≥ 200 bp. Requiring a CpG content of at least 6%, as originally recommended (Matsuo et al. 1993, Li et al. 2002), drastically reduced the number of CGIs (data not shown). For CGI detection with *CPGed* (Luque-Escamilla et al. 2005), G+C ≥ 55%, $CpG_{obs}/CpG_{exp} \geq 0.65$, and a minimal size of 200 bp were required. The third method, *CpGcluster* (Hackenberg et al. 2006), does not identify conventional CGIs but so-called CpG clusters in which CpG dinucleotides are more tightly spaced than expected by chance. Its main parameter, the maximal distance between two CpGs up to which a CpG cluster is extended, is determined from the distribution of pairwise CpG distances in the input sequence. It was set to the 75^{th} percentile in order to obtain longer CpG clusters that overlap more frequently with transcriptional start sites than when using the recommended median, thus being more suitable for comparison with conventional CpG islands.

Most CGIs were identified applying GGF parameters. The lowest numbers were obtained when taking TJ parameters and with the segmentation program *cpg* (see Tab. 3.1, column "CGIs"). As to be expected from the parameter choice, the numbers of *CPGed* CGIs range between GGF and TJ. They are similar to those of CpG clusters, as are those of *cpg* and TJ CGIs. In general, the number or length of CpG islands are uncorrelated with the sequence length. Thus long genes do not have proportionally more CGIs and CGIs are not equally distributed throughout a sequence. The CGI length distributions are highly left-skewed, shaped by the minimal length criteria of 200 bp (GGF, *CPGed*, *cpg*) or 500 bp (TJ). CpG-rich segments identified by *CPGed* and *cpg* can nevertheless be up to several 1000 bp long. Although *CpGcluster* does not use a length limit, the distribution of

CGI lengths has a similar shape as for the other methods. At the short extreme, CpG clusters can be eight bp long only, consisting of four CpGs. G+C and CpG contents are approximately Gaussian distributed. Depending on the parameters, the medians are lowest for GGF (52% G+C, 4.4% CpG) and highest for *cpg* (65% G+C, 8.2% CpG). CGIs with 100% G+C, a maximum peculiar to *CpGcluster*, consist of low complexity sequences with only Gs and Cs.

Since repetitive elements are not expected to provide an open chromatin structure, all CGIs critically depending on them have to be regarded as false positives. Using sequences in which repeats are replaced by Ns, however, changes the overall distribution of CpG distances and thereby identification thresholds for *CpGcluster*, results in extensive splitting of CGIs especially for segmentation methods, and can even induce artificial CGIs due to the elevation of the CpG_{obs}/CpG_{exp} ratio (see chapter 2.2). Alternatively, omitting the nucleotides that coincide with repetitive elements, we recalculated G+C content, CpG_{obs}/CpG_{exp}, number of CpGs, and length of sections identified as CGIs in the original sequences. A CGI that still fulfills all respective criteria can be regarded as a single copy, unique sequence ("unique" in Tab. 3.1) whereas one that fails to meet them is classified as repeat-dependent ("repeat" in Tab. 3.1). In some cases, repeat-dependent and unique CGIs do not sum up to the total number because program bugs caused a few CGIs to be reported despite not fulfilling the required criteria.

Table 3.1: Numbers of CpG islands and overlaps with repetitive elements

	human			mouse		
method, group	**CGIs**	**repeat (%)**	**unique (%)**	**CGIs**	**repeat (%)**	**unique (%)**
GGF G1	1106	62.30	37.70	659	26.56	73.14
GGF G2	1486	61.17	38.76	645	27.75	71.47
GGF G3	933	65.70	34.19	582	29.04	70.96
TJ G1	140	22.14	77.86	97	10.31	89.69
TJ G2	179	24.58	75.42	118	11.02	87.29
TJ G3	131	24.43	75.57	101	14.85	85.15
CPGed G1	447	54.81	37.81	212	19.34	75.47
CPGed G2	542	57.93	35.61	206	24.27	70.39
CPGed G3	366	65.03	29.51	190	25.79	68.95
cpg G1	142	22.54	77.46	88	5.68	94.32
cpg G2	146	20.55	79.45	99	6.06	93.94
cpg G3	129	26.36	73.64	91	5.49	94.51
Cluster G1	364	34.89	65.11	252	14.68	85.32
Cluster G2	428	32.24	67.76	238	15.97	84.03
Cluster G3	346	40.46	59.54	241	14.52	85.48

Most repeat-dependent CGIs are detected with GGF criteria in human sequences where they constitute approximately 60% of the total number. Most of them are caused by overlap with SINEs, namely the CpG-rich human *Alu* and its murine homolog, the B1 element (Appendix A Tab. A2). Repeat classes like simple repeats, low complexity regions, and DNA elements contribute a minor amount of CGIs. B1 elements are shorter and CpG-poorer than *Alu* elements (Quentin 1994) and

less abundant in the mouse genome than *Alu* elements in the human one (Waterston et al. 2002). Therefore, it is not surprising that in mouse more than 70% of the GGF CGIs are unique. Their numbers come close to those of human unique GGF CGIs. *CPGed* has a similarly low rate of unique CGIs.

Unique CGIs show higher median values for length, G+C and CpG content than repeat-dependent ones. Nevertheless, there is a large overlap between the ranges so that making stricter requirements for these parameters not only discards false positive CGIs but also unique, possibly functional ones, as can be seen when comparing the numbers of unique GGF CGIs with that of TJ CGIs in both species. Although more than 75% of the TJ CGIs are unique, repeat-dependent CGIs cannot be completely excluded even with these stringent criteria. Similar to TJ, only a few repeat-dependent CGIs are identified with *cpg*. *CpGcluster* does not use simple numerical criteria, therefore it is unclear when a cluster would have to be called repeat-dependent. Since the smallest clusters without any repeat overlap contain four CpGs, for simplicity a cluster that possess at least four CpGs outside of repetitive elements is regarded as unique. Under these conditions, 64% of human CpG clusters are unique and in mouse 85%.

3.1.2 Promoter CpG islands possess pronounced characteristics and are reliably detected

Only very few promoter CGIs are declared as repeat-dependent due to G+C-rich simple repeats or low complexity regions and even fewer comprise interspersed elements. The number of promoter CGIs is similar throughout the methods (χ^2 test, $p > 0.1$) and, in contrast to CGIs at other locations, most promoter CGIs are reliably detected by all programs (Fig. 3.1). Regarding homologous promoter pairs, the CGI tends to be missing for the mouse in more cases than vice versa ($p < 0.1$; Appendix A Tab. A3). Thus, the existence or absence of an analogous promoter CGI is mostly conserved, consistent with the results of Yamashita et al. (2005).

Promoter CGIs are in general longer and show higher G+C and CpG contents than CGIs that do not overlap with transcriptional start sites. Their length is approximately Gaussian distributed with both GGF and TJ parameters (see also Takai and Jones 2002). Typical promoter CGIs have been reported to be around 1000 bp or longer (Gardiner-Garden and Frommer 1987, Larsen et al. 1992). The length difference between human and mouse is neither pronounced, nor is the observed trend consistent over the three sequence groups. Considering pairs of homologous promoter CGIs (Appendix A Tab. A3), human ones are longer in most cases ($p < 0.001$), except for CpG clusters ($p > 0.1$). The G+C content is only marginally lower in mouse than in human for GGF and TJ (Wilcoxon test, $p < 0.05$) but similar for the other methods. CpG_{obs}/CpG_{exp} and CpG content are also not significantly different between both species.

Non-promoter CGIs identified in similar positions with different methods or criteria sets are promising candidates for possessing regulatory functions as well. As to be expected, virtually all TJ, *cpg*, and *CPGed* CGIs as well as more than 90% of CpG clusters overlap with GGF CGIs but there are many GGF CGIs, even unique ones, without a match (Tab. 3.2). If compared to each other, TJ, *cpg*, *CPGed*, and *CpGcluster* do not consistently identify the same CGIs (Appendix A Tab. A4, A5). Exonic CGIs are frequently identified by different programs due to the conservation of CpGs in protein-encoding sequences. These CpG-rich segments are not expected to fulfill additional regulatory functions but can be easily excluded if the exon annotation is known.

Apparently, the identification of truly functional intronic CGIs is the most challenging task for the tested programs since the numbers of intronic CGIs are highly variable between the groups (Fig. 3.1) and their recovery rates are low (Tab. 3.2). The ones in common may represent

promoters of alternative or antisense transcripts. In contrast, considerable numbers of CGIs with high recovery rates are located in intergenic regions, i.e. in the portions of upstream and downstream sequences that do not overlap with neighboring genes. This indicates that these regions might encompass additional yet unidentified epigenetic regulatory elements, such as alternative upstream promoters, enhancers or silencer elements.

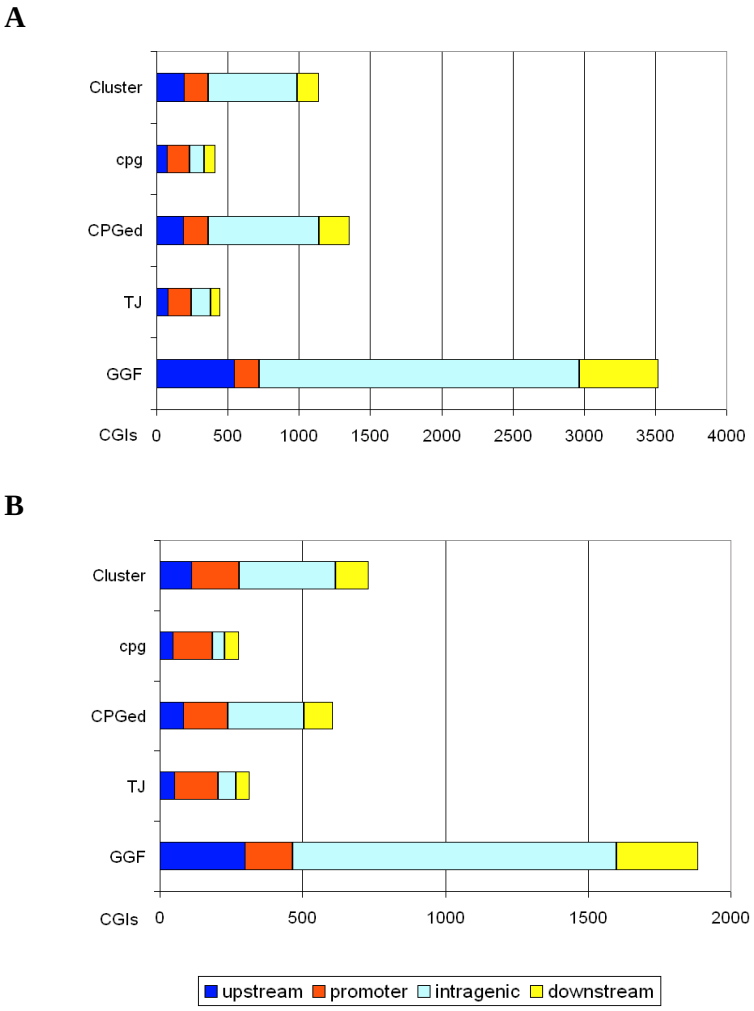


Figure 3.1: Distribution of CGIs under various criteria in different locations
The numbers of CGIs reported by *CpG Island Searcher* with GGF or TJ parameters, *CpGed*, *cpg*, and *CpGcluster* show large differences between the methods as well as between human (A) and mouse (B). Upstream CGIs reside in a 10 kb window 5' of the transcriptional start site of the reference gene but do not overlap with it; promoter CGIs were defined as overlapping with the most upstream transcriptional start site of the reference gene as annotated in the database; intragenic CGIs are located between the transcriptional start site and the 3' end of this gene; downstream CpG islands have to reside completely in the 10 kb window downstream of the annotated transcriptional termination site of the respective gene. The high numbers of intragenic, upstream and downstream GGF CGIs are mainly caused by repetitive elements.

Table 3.2: Common unique CGIs

human

method, group	total	promoter	exonic	intronic	intergenic
GGF G1	417	56	133	79	67
TJ G1	27%	86%	19%	8%	19%
CPGed G1	46%	75%	52%	29%	39%
cpg G1	44%	88%	29%	22%	46%
Cluster G1	42%	96%	43%	24%	27%
GGF G2	576	64	187	192	59
TJ G2	24%	92%	13%	8%	27%
CPGed G2	39%	63%	39%	27%	47%
cpg G2	28%	94%	17%	5%	46%
ClusterG2	37%	98%	35%	16%	41%
GGF G3	319	57	96	69	44
TJ G3	31%	93%	15%	6%	11%
CPGed G3	38%	67%	35%	25%	25%
cpg G3	32%	86%	11%	7%	30%
Cluster G3	47%	100%	30%	23%	36%

mouse

method, group	total	promoter	exonic	intronic	intergenic
GGF G1	482	51	149	134	66
TJ G1	18%	88%	9%	4%	12%
CPGed G1	35%	78%	34%	14%	32%
cpg G1	21%	82%	9%	2%	24%
Cluster G1	41%	98%	36%	21%	41%
GGF G2	461	55	142	131	69
TJ G2	23%	87%	8%	4%	23%
CPGed G2	33%	62%	33%	15%	29%
cpg G2	23%	87%	7%	4%	29%
Cluster G2	39%	96%	30%	21%	32%
GGF G3	413	58	122	117	55
TJ G3	22%	88%	8%	3%	15%
CPGed G3	35%	79%	31%	16%	33%
cpg G3	25%	84%	13%	3%	22%
Cluster G3	43%	98%	34%	32%	24%

The numbers refer to unique GGF CGIs. The percentages show their rate of overlap with unique CGIs identified with the other listed methods.

3.1.3 General differences between human and mouse CpG islands

Table 3.3 shows the results for comparing the properties of human and mouse CGIs detected with different programs. Detailed data are given in Appendix A Tab. A6. In consensus, mouse sequences have fewer CGIs than human ones and these are G+C poorer with a tendency to be shorter, which is consistent with published data. The genome of the mouse contains fewer CGIs than the human one (Antequera and Bird 1993, Waterston et al. 2002). In genome-wide analyses using TJ criteria, which found similar average lengths as our study, human CGIs were found to be slightly longer than those of mouse (Zhao and Zhang 2006a, 2006b). A higher G+C content of

human CGIs was confirmed experimentally as well as computationally (Gardiner-Garden and Frommer 1987, Aïssani and Bernardi 1991, Antequera and Bird 1993, Matsuo et al. 1993, Zhao and Zhang 2006a, 2006b). In the literature, conflicting data have been reported on the CpG_{obs}/CpG_{exp} ratios of murine and human CGIs (Gardiner-Garden and Frommer 1987, Matsuo et al. 1993, Yamashita et al. 2005, Hackenberg et al. 2006, Zhao and Zhang 2006a, 2006b). For the sequences and methods used here, there are no consistent differences of the CpG_{obs}/CpG_{exp} ratio between the two species. The obtained data indicate that this measure more likely depends on the algorithms than on species-specific features. Likewise, the CpG content of CGIs is in general similar in both species.

In addition, I later analyzed the data for the autosomal CGIs annotated by UCSC for the human (hg18) and mouse (mm8) genomes and found them to be consistent with the results for most other methods, although this investigation was not performed on orthologous genes. Interestingly, the minimal CpG content is 5.8% for human and 6.0% for mouse CGIs, respectively, which should be indicative of general hypomethylation of the UCSC CpG islands (Matsuo et al. 1993).

Table 3.3: Comparison of human and mouse CGIs detected with different methods

method	number	length	G+C content	CpG_{obs}/CpG_{exp}	CpG content
GGF	+	0	+	0	0
unique GGF	0	+	+	0	+
TJ	0	+	+	0	+
cpg	+	0	0	-	-
CpGed	+	-	0	0	0
CpG cluster	+	-	+	0	+
UCSC genome	+	+	+	-	+

"+" signifies that the value for human CGIs is greater than that of mouse CGIs; "-" refers to the opposite; 0 means that they are similar or differences are not consistent throughout all groups.

*3.1.4 The (TpG+CpA)/(2*CpG) ratio is correlated to epigenetic properties of CpG islands*

According to the original definition, CGIs are CpG rich, unmethylated genomic regions. Therefore, in contrast to functional, unique CGIs, CGIs depending on repetitive elements that are usually methylated should accumulate more mutations due to hydrolytic deamination of 5-methylCpG, which converts CpG to TpG, or CpA on the reverse complementary DNA strand (compare Fig. 1.2). A positive correlation between the 5-methylCpG content and the enrichment of TpG and CpA in CGIs has been reported (Gardiner-Garden and Frommer 1987, Matsuo et al. 1993). If the dinucleotides were equally distributed, one would expect the ratio (TpG+CpA)/(2·CpG) to equal one. Overrepresentation of CpA and TpG as potential mutation products would markedly increase the value whereas a relative enrichment of CpG would lower it. Thus, this ratio could serve as an estimated deamination rate to distinguish between different types of CGIs. Indeed, in human and mouse, the (TpG+CpA)/(2·CpG) ratio is significantly higher in repeat-dependent CGIs compared with unique CGIs in TJ, *cpg*, and GGF ($p < 0.001$; Fig. 3.2). For *CPGed* and *CpGcluster*, no significant differences between unique and repeat-dependent CGIs could be observed. This might be due to the generally low CpG content of *CPGed* CGIs and the definition of repeat-dependence for CpG clusters.

Murine CGIs have been suggested to erode due to CpG deamination in the germ line (Antequera and Bird 1993, Matsuo et al. 1993). In line with this hypothesis, both the mouse genome and computationally identified CGIs therein have a higher content of CpA and TpG and a lower content of CpG compared to human (Zhao and Zhang 2006a, 2006b). In order to assess the differences between the species with respect to the estimated deamination ratio and to derive a threshold for using the (TpG+CpA)/(2·CpG) ratio as a filter for epigenetic relevant CGIs, analyses were concentrated on GGF criteria, where there is a sufficiently large number of CGIs of each category to compare. Unique, repeat-dependent, and the totality of mouse CGIs have a significantly higher (TpG+CpA)/(2·CpG) ratio than their human counterparts (Wilcoxon test, $p < 0.001$). As can be seen in figure 3.2, the estimated deamination rate is more variable in mouse than in human. Interestingly, promoter CGIs, which are most probably always unmethylated, behave similarly in both species. 90% of the unique promoter CGIs in human have a ratio below or equal to 1.0. For mouse, the 90^{th} percentile is 1.2. While promoter CGIs only constitute 13.5% of the three unique GGF CGI sets in human, their proportion is 45% of all unique human GGF CGIs that fulfill (TpG+CpA)/(2·CpG) ≤ 1.0. Using the corresponding value of 1.2 as a filter for mouse, a similar enrichment is observed: 46% instead of only 12% of all unique GGF CGIs overlap with the annotated transcriptional start sites of the genes. The total number of CGIs fulfilling the respective criteria is similar to that of TJ CGIs and most of the filtered GGF CGIs are also detected with TJ criteria (Appendix A Tab. A7). On the other hand, the filtered GGF CGIs capture a lower percentage of TJ CGIs (Appendix A Tab. A8). Thus, applying a species-specific ratio of (TpG+CpA)/(2·CpG) ≤ 1.0 for human or ≤ 1.2 for mouse GGF CGIs, respectively, as an additional filter appears to be suitable for highlighting the most promising candidate regulatory CGIs.

Experimental data on the epigenetic properties of the CGIs investigated here are not available. However, predictions have been established for all GGF CGIs in the human genome (Bock et al. 2007). They are based on epigenetic data for CGIs on human chromosomes 21 and 22, which were used to train a support vector machine. Histone and DNA modifications characteristic for an open chromatin structure were found to correlate with certain features of the DNA sequence (Bock et al. 2006). Using these attributes, an epigenetic score can be calculated that predicts the likelihood of absence of DNA methylation, promoter activity and open chromatin structure (Bock et al. 2007). The epigenetic score ranges between 0 and 1. A value of 0.5 corresponds to a CGI that is equally likely to be either transcriptionally active or inactive whereas a CGI with a score ≥ 0.67 has a high confidence of possessing an open chromatin structure. The unique human GGF CGIs were assigned epigenetic scores by mapping their sequences to the UCSC hg17 genome, for which the pre-calculated scores are available. Sometimes, a CGI of one set overlapped with more than one CGI of the other set, which happens because starting from different points on the genomic sequence may result in splitting or merging of some CGIs that were identified when using the whole chromosomal sequence. In such cases, the unique GGF CGI was assigned the score of the CGI with the larger overlap. If a CGI was assigned to more than one gene, it was counted only for the CGI with the larger overlap.

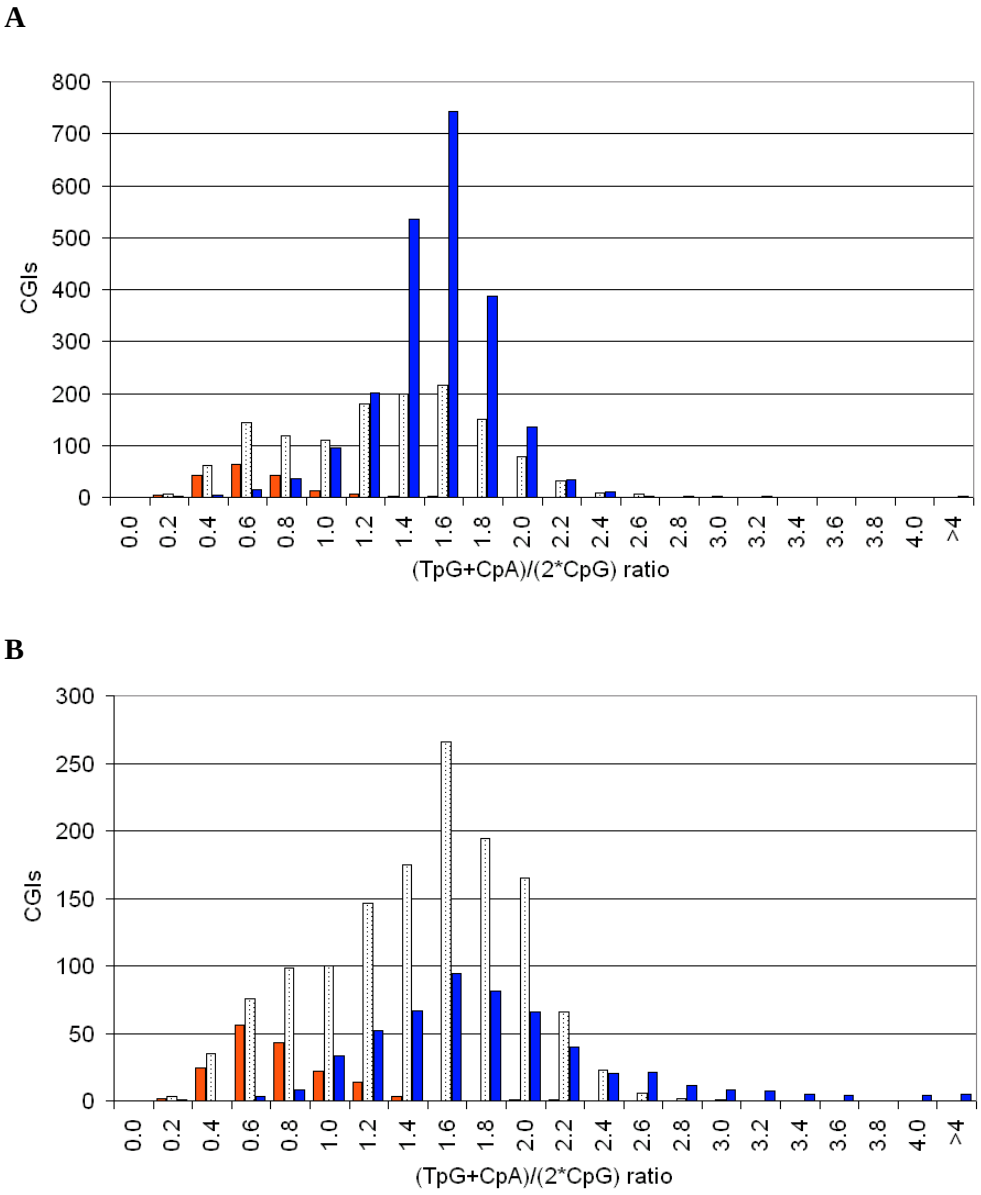


Figure 3.2: (TpG+CpA)/(2·CpG) ratio in Gardiner-Garden and Frommer CGIs
The ratio (TpG+CpA)/(2·CpG) serves as an estimation of the deamination rate in CGIs of human (A) and mouse (B). Its distribution in presumably unmethylated promoter CGIs (red) is centered at much lower values than that of unique (dotted) or repeat-dependent, most probably methylated CGIs (blue). Apart from a few outliers, the ratios of promoter CGIs are below 1.4. Having a 90th percentile of 1.0 for human and 1.2 for mouse, they can be well separated from repeat-dependent CGIs with little overlap between both groups.

The length of a CGI has been suggested as a strong predictor for an open chromatin structure (Bock et al. 2007) and there is indeed a significant correlation between the lengths of the GGF CGIs and their epigenetic scores (Pearson's r = 0.65; Kendall's τ = 0.41). The correlation, however, is not linear (Fig. 3.3 A). CGIs that are very long (more than 2000 bp) have a high probability to be transcriptionally active whereas the scores of CGIs below 1000 bp almost span the entire range. Although most of the CGIs near the lower length limit of 200 bp are assigned values below 0.5,

which means that they are more probably inactive, some achieve very high scores. A considerable number of short CGIs with high scores would therefore be discarded by focusing only on long CGIs. A much better, negative correlation can be observed between the (TpG+CpA)/(2·CpG) ratios of unique human GGF CGIs and their predicted epigenetic scores ($r = -0.74$, $\tau = -0.52$; Fig. 3.3 B). CGIs with a low (TpG+CpA)/(2·CpG) ratio are concentrated in the upper range of the epigenetic score, consistent with the assumption that they are unmethylated and transcriptionally active. Thus, the lower the (TpG+CpA)/(2·CpG) ratio, the more likely the CGI is transcriptionally active.

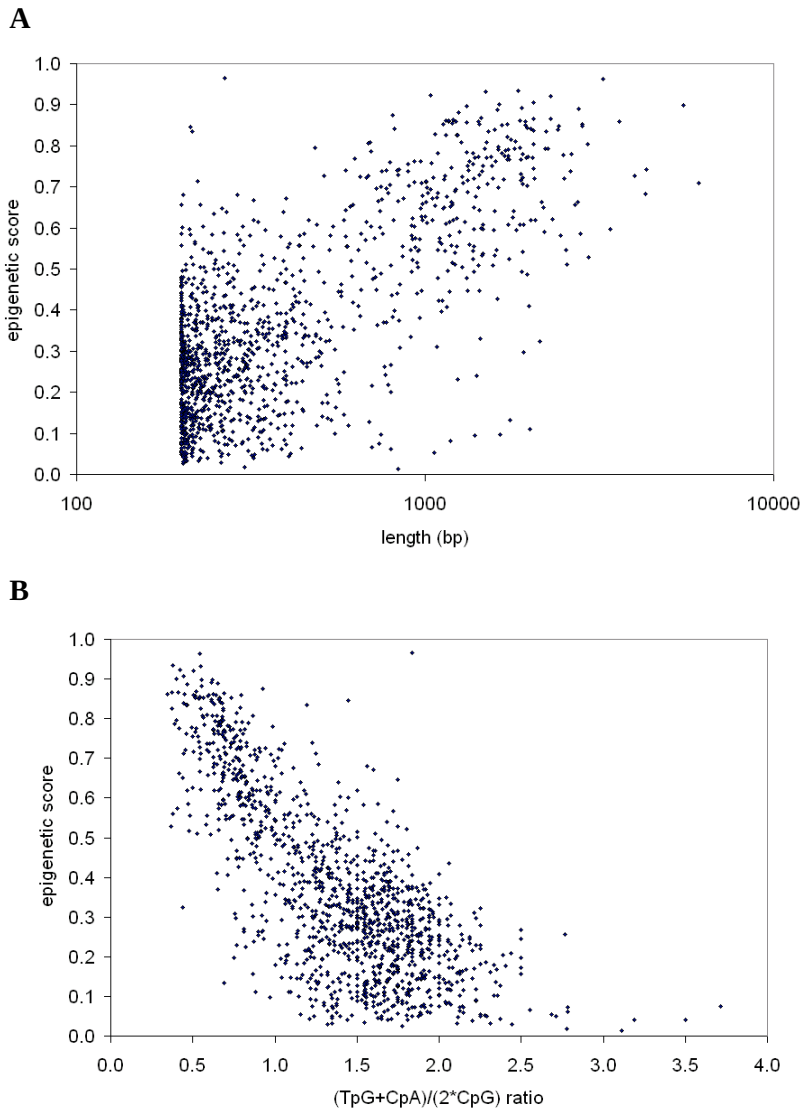


Figure 3.3: Correlation of CGI length and (TpG+CpA)/(2·CpG) ratio with the assigned predicted epigenetic score
(A) Correlation between the lengths of the unique human GGF CGIs and their assigned epigenetic scores. Note that the x-axis is in logarithmic scale. (B) The predicted epigenetic score of the unique human GGF CGIs shows a marked negative correlation with their (TpG+CpA)/(2·CpG) ratio.

The human promoter CGIs identified in this study achieve median scores of 0.70, consistent with their transcriptionally active state. For all GGF CGIs that fulfill the above mentioned threshold of (TpG+CpA)/(2·CpG) ≤ 1.0, the median score is 0.66. In both groups, the lower quartile (25^{th} percentile) is above 0.5, which means that more than 75% of these CGIs have a high probability to function as open chromatin CGIs. This finding supports the choice of the estimated deamination rate cutoff. For the mouse, predictions of epigenetic scores are not yet available. By transferring the obtained results from human, taking a (TpG+CpA)/(2·CpG) ratio of ≤ 1.2 into account might help to identify CGI with regulatory functions in this species.

3.1.5 Summary and conclusions of chapter 3.1

By comparing different methods for CGI identification, we found that promoter-associated CGIs are quite reliably detected by all methods but their performance differs largely for CGIs at other locations. Most programs identify fewer CGIs in mouse than in human sequences, and these CGIs are shorter and G+C poorer. CpG content turns out to be a highly variable feature among the programs. Repetitive elements are differently connected with human and mouse CpG islands. Moreover, an estimation of the CpG deamination rate provides means to distinguish functional regulatory CGIs from probably methylated, non-functional ones.

3.2 CpG islands in imprinted and non-imprinted regions

The key question of the following analysis was whether imprinted genes stand out against biallelically expressed ones in terms of sequence features, particularly G+C and CpG content and CpG islands. Especially CGIs might be influenced by the presence of differentially methylated regions. In order to get a statistically verified answer, imprinted genes of human and mouse were compared to two random control groups of double their size, namely the before mentioned G1 and G2. The use of two control groups allowed us to check whether they both differed from the imprinted set. If they did not agree on this but differed from each other, there would be a bias in one of them which would go undetected when only taking one control group. Human and mouse genes were treated separately to differentiate between features that are imprinting-specific and those that are species-specific. Most of the results presented here are published in Hutter et al. (2006).

3.2.1 Different sequence properties of human and mouse imprinted and non-imprinted genes

As the imprinted group, we selected 38 human and 39 murine imprinted genes for which imprinting effects have been reported in at least one of the two species. Thirty-four genes are orthologous: *ASB4*, *ASCL2*, *ATP10A*, *CDKN1C*, *CD81*, *COPG2*, *DIO3*, *DLK1*, *DLX5*, *GATM*, the *NESP-GNAS* locus, *GRB10*, *MEG3* (murine ortholog: *Gtl2*), *H19*, *HTR2A*, *IGF2*, *KCNQ1*, *MAGEL2*, *NAP1L55*, *NDN*, *NNAT*, *PEG3*, *PEG10*, *PHLDA2*, *PLAGL1*, *RASGRF1*, *SGCE*, *SLC22A2*, *SLC22A3*, *SLC22A18*, *SLC38A4*, *SNRPN*, *UBE3A*, and *WT1*. Due to some genes being unavailable in the mouse genome assemblies and others having no ortholog or being reported as not imprinted in human, the human imprinted group additionally includes *MEG8*, *MEST*, *USP29*, and *ZNF264* whereas the mouse imprinted group is completed by *Peg12*, *Commd1*, *Ins2*, *Igf2r*, and *Impact*. The *Copg2* and *MEG8* sequences were taken from the Ensembl Genome Browser release 30, and the *Grb10* sequence from release 32. Sequences were downloaded just as the control groups G1 and

G2 (see section 3.1.1) and processed with *RepeatMasker*. Detailed data for the imprinted sequences can be found in Appendix B Tab. B1. In table 3.4 below, an overview of the general sequence properties is given, listing means and standard deviations.

Table 3.4: Lengths, G+C and CpG contents of imprinted and control sequences

group	median gene length (bp)	total G+C content (%)	G+C content in repetitive elements (%)	total CpG content (%)	CpG content in repetitive elements (%)
human imprinted	26,583	46.93 ± 7.68	44.96 ± 5.65	1.82 ± 0.97	1.54 ± 0.79
human G1	27,992	46.27 ± 6.37	45.82 ± 4.18	1.66 ± 0.83	1.68 ± 0.88
human G2	24,270	46.02 ± 6.14	46.29 ± 4.37	1.70 ± 0.83	1.72 ± 0.87
mouse imprinted	20,697	45.43 ± 4.65	43.90 ± 3.18	1.28 ± 0.46	0.87 ± 0.42
mouse G1	21,603	46.37 ± 4.22	45.46 ± 2.43	1.35 ± 0.47	0.93 ± 0.29
mouse G2	15,103	45.87 ± 4.11	45.59 ± 2.30	1.35 ± 0.52	0.93 ± 0.26

The imprinted and the control groups have similar gene lengths (Wilcoxon test, $p > 0.4$). Ranging from 0 to 4.97% with a total average of 0.71 ± 1.22%, the ratio of Ns is lowest in the murine imprinted group. G+C content and the rate of CpG dinucleotides are similar between imprinted and control groups for both species. More interestingly, G+C and CpG content are more variable in the imprinted groups than in the control groups and the significant contrast between human and mouse sequences is more pronounced. This may be caused by an overrepresentation of imprinted human genes in the category of very CpG- and G+C-rich genes, particularly genes from the BWS region (chr. 11) and the *DKL1-GTL2* domain (chr. 14), and *NNAT*. Their murine orthologs do not reach comparably extreme CpG contents. For the control groups, the difference seems to be predominantly caused by a significantly higher proportion of CpGs in human-specific repeat elements: When omitting the repetitive elements identified with *RepeatMasker*, the difference is reduced to a tendency ($p < 0.1$). Sequences in the imprinted group, on the other hand, comprise a lower content of repetitive elements than control sequences ($p < 0.01$), especially with regard to the SINE Alu/B1 family ($p < 0.00005$; Greally 2002, Ke et al. 2002a, 2002b, Allen et al. 2003, Walter et al. 2006; Tab. 3.5). The reported enrichment of LINE-1 repeats (Walter et al. 2006) can only be observed on a larger genomic scale.

Table 3.5: Contents of repetitive elements

group	total (%)	SINE (%)	B1/Alu (%)	LINE (%)	LINE-1 (%)
human imprinted	32.01 ± 13.36	9.73 ± 5.70	7.34 ± 5.51	12.66 ± 8.39	10.18 ± 7.99
human G1	38.10 ± 12.20	18.37 ± 9.60	14.96 ± 9.08	11.26 ± 9.54	8.35 ± 9.33
human G2	39.81 ± 12.09	20.11 ± 10.24	16.90 ± 10.34	10.44 ± 7.35	7.30 ± 6.67
mouse imprinted	23.36 ± 11.70	7.02 ± 5.04	2.23 ± 2.06	6.82 ± 8.13	6.61 ± 8.13
mouse G1	27.80 ± 9.03	11.91 ± 6.33	4.61 ± 3.16	5.52 ± 6.43	5.21 ± 6.27
mouse G2	28.74 ± 9.38	13.16 ± 7.85	5.06 ± 3.64	5.50 ± 5.85	5.13 ± 5.74

3.2.2 General CpG island properties do not distinguish imprinted genes

Several authors previously reported that CpG islands (CGIs) seemed to be enriched in imprinted regions (Paulsen et al. 2000, Paulsen and Ferguson-Smith 2001, Reik and Walter 2001). Thus, we searched to investigate whether this hypothesis can be statistically verified by comparison with biallelically expressed genes. For simplification, we only looked at CGIs reported by the *CpG Island Searcher* with Gardiner-Garden and Frommer (1987) or Takai and Jones (2002) parameters, respectively, both before and after masking repetitive elements. With respect to length, G+C and CpG content of CGIs and the fractions of the sequence covered by CGIs (Appendix B Tab. B2), no consistent differences could be found between sequences of imprinted and randomly chosen biallelically expressed genes. Table 3.6 lists average CGI numbers. Using the Gardiner-Garden and Frommer criteria (GGF), more CGIs per gene were identified in human than in mouse also in the imprinted (t test, $p < 0.05$), but the difference is not as pronounced as in the control groups ($p < 0.001$). When masking repetitive elements beforehand (GGF_mask), the CGI number decreases significantly in human control sequences, but not in the imprinted group. This can be easily explained by the lower SINE content of imprinted genes (Greally 2002; see above, Tab. 3.5), which particularly results in fewer *Alu*-dependent CGIs. Likewise, mouse repetitive elements rarely coincide with CGIs in general so that both GGF and GGF_mask identify a similar number of CGIs. Thus, in contrast to the control sets, only 44% of the human and 16.5% of the murine GGF CGIs in imprinted regions are repeat-dependent (Tab. 3.7). Nevertheless, the difference of this rate between the two species is still highly significant (χ^2 test, $p < 0.0001$). The Takai and Jones parameters exclude most of the rather small CpG islands in repetitive elements, resulting in similar values without (TJ) or with (TJ_mask) masking of repetitive elements for all groups.

Table 3.6: Average numbers of CpG islands per gene

group	GGF	GGF_mask	TJ	TJ_mask
human imprinted	14.9	9.9	2.5	2.4
human G1	14.0	5.3	1.8	1.7
human G2	18.8	8.2	2.3	2.0
mouse imprinted	6.5	5.9	1.5	1.6
mouse G1	8.3	6.6	1.2	1.0
mouse G2	8.2	6.3	1.5	1.5

When classifying CGIs in single-copy sequences by their locations, only intragenic ones stand out as distinguishing. Human control groups contain significantly more sequences with intragenic CGIs than mouse control groups (χ^2 test, $p<0.05$). In contrast, for the groups of imprinted genes both species behave similarly, suggesting that they share the feature of strong intragenic CpG islands. Concentrating on those identified in repeat masked sequences with TJ parameters (TJ_masked), the imprinted group seems to contain a higher proportion of intragenic CpGIs satisfying the most stringent criteria (Fig. 3.4). For the mouse, this is mainly the result of more imprinted genes than control genes having at least one TJ intragenic CpG island (χ^2 test, $p < 0.05$). In human, however, χ^2 tests only yield a tendency ($p < 0.1$) for TJ_mask. Besides, there is a trend for intragenic CpG islands in imprinted genes to be longer than those in the control groups (Wilcoxon test, $p < 0.1$). Both weak effects and the presence of additional intronic CGIs provided by genes like *KCNQ1* add up to the difference of average CGI lengths at intragenic positions observed in figure 3.4. Such

intronic CGIs may be associated with the promoter regions of antisense or alternative transcripts (Reik and Walter 2001).

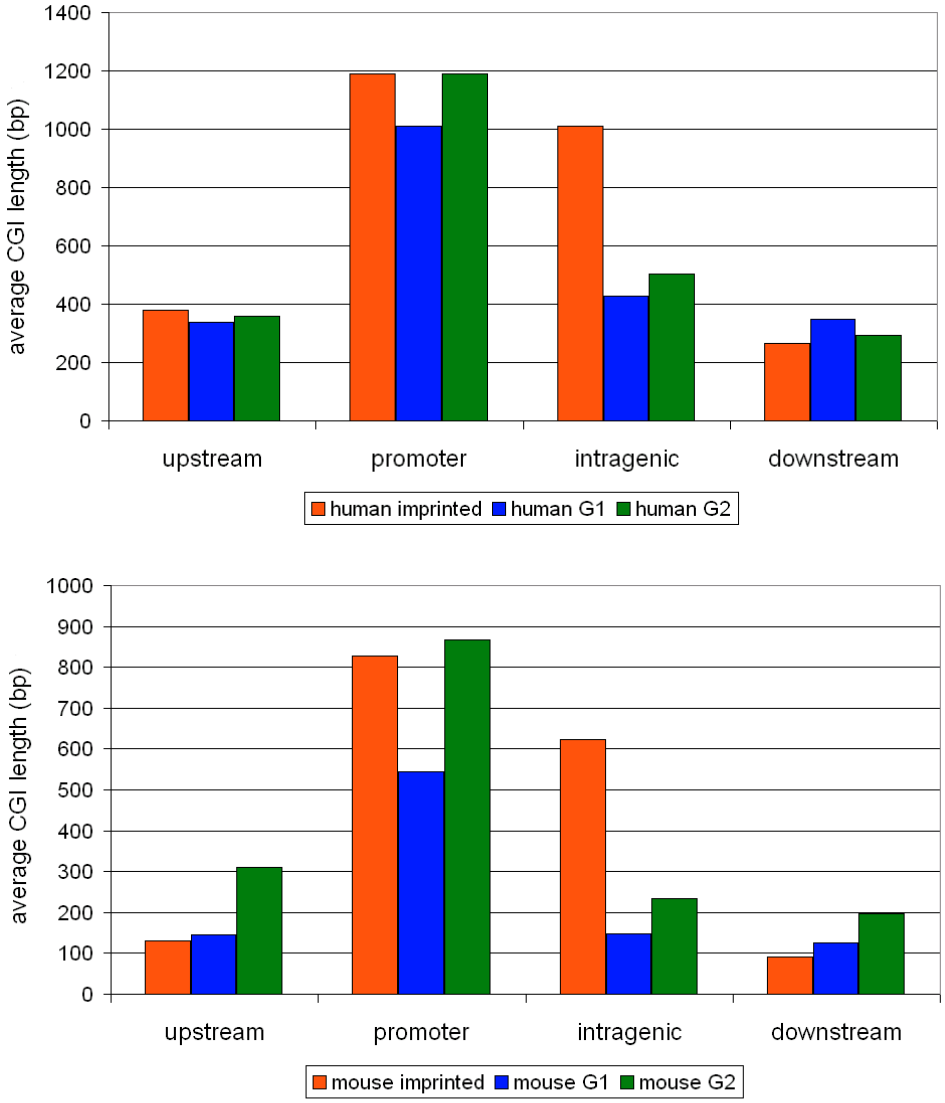


Figure 3.4: Average lengths of CpG islands per gene
CpG islands in repeat masked sequences at four different locations (definitions see Fig. 3.1) were identified according to the Takai and Jones parameters for human and mouse imprinted and two groups of control genes (G1, G2). Precisely, the CpG island lengths were summed up per sequence, group, and location, then divided by the number of sequences in the respective group.

3.2.3 Estimating CpG deamination effects on CpG islands in imprinted regions

In order to unravel the species-specific differences of CGIs from the potential effects of differential methylation, I investigated the estimated CpG deamination ratio of the CGIs identified in the imprinted group. Also here, unique GGF CGIs have a lower (TpG+CpA)/(2·CpG) ratio than repeat-dependent ones (Wilcoxon test, $p < 0.002$), indicating that deamination acts on CpG rich repetitive elements independent of differential methylation. With regard to all GGF CGIs, the estimated deamination rate is similar in human and mouse ($p > 0.2$). Unique GGF CGIs, however, and even more significantly repeat-dependent ones show a higher ratio in mouse ($p < 0.05$ and $p < 0.001$, respectively). Compared to control sets, the (TpG+CpA)/(2·CpG) ratio is not significantly reduced in CGIs in imprinted regions ($p > 0.2$) and the epigenetic scores are not lowered ($p > 0.1$).

Promoter CGIs, which are all unique, have essentially the same ratios in both species ($p > 0.6$). Here, the 90th percentile of the ratio is 1.2 for human and 1.1 for mouse. Filtering with the (TpG+CpA)/(2·CpG) limits proposed above elevates the portions of promoter CGIs from 9% to 38% for human and from 12% to 47% for mouse (Tab. 3.7). Only before filtering is the lower ratio for the human imprinted group compared to control groups significant (χ^2 test, $p < 0.05$), which may be due to the enrichment of intragenic CGIs. The epigenetic score of unique human GGF CGIs in imprinted regions correlates with their length ($r = 0.6531$, $\tau = 0.3534$) and (TpG+CpA)/(2·CpG) ratio ($r = -0.6965$, $\tau = -0.4893$) to a slightly lower degree than it is the case for biallelically expressed genes. The median epigenetic score is 0.73 for promoter CGIs and 0.63 for CGIs with (TpG+CpA)/(2·CpG) ≤ 1.0. In summary, all observations are highly similar to those made when analyzing CGIs of biallelically expressed genes. This is confirmed by analyses with *EpiGraph* (Bock et al. 2009) in which no consistent differences were found between the CGIs of imprinted regions and those of control groups G1 and G2. Apparently, the effects of differential methylation have less influence on the CGI characteristics of imprinted regions than species-specific peculiarities.

Table 3.7: GGF CGIs in imprinted regions of human and mouse

	GGF CGIs	unique GGF CGIs	unique CGIs that fulfill (TpG+CpA)/(2·CpG) threshold	promoter CGIs	promoter CGIs that fulfill (TpG+CpA)/(2·CpG) threshold
human	566	316	58	29	22
mouse	254	212	51	26	24

3.2.4 Supporting evidence from genome-wide conservation studies

Later, parts of the analyses were extended to a larger set of orthologous imprinted genes (see Appendix D Tab. D3) and to all autosomal genes in the human and mouse genome. The CpG islands annotations from UCSC used here are close to the original criteria (Gardiner-Garden and Frommer 1987) but exclude repetitive elements and have a higher CpG content. Although imprinted genes do not show an enrichment of CGIs in general, 15 out of 57 protein-encoding imprinted genes (26.32%) possess at least one intronic CGI in human and 11 out of 53 (20.75%) in mouse, which is significantly more than the 8.64% and 3.77% for autosomal human and mouse genes, respectively (χ^2 test, $p < 0.001$).

To further investigate whether CGIs differ in features that, in contrast to length, G+C and CpG

content etc., do not depend on algorithmic parameters, we analyzed the overlap of mammalian *phastCons* highly conserved sequences (PCSs; Siepel et al. 2005) of length ≥ 20 bp and CpG islands (see also section 3.3.3, Tab. 3.9). Eight percent of the PCSs in human imprinted regions overlap with CGIs, whereas the genome-wide ratio is only 4%. Regarding the mouse, the percentages are lower in both groups but nevertheless the difference is highly significant. In both species, the enrichment is most prominent for the group of intronic PCSs. This is in line with the observation that intragenic CGIs occur more frequently and are slightly longer in imprinted genes (Fig. 3.4, section 3.2.2).

Interestingly, CpG islands in imprinted regions are not significantly less conserved than genome-wide (χ^2 test, $p > 0.6$): 65% of the 133 human and 83% of the 59 murine CGIs in imprinted regions overlap with PCSs, the genome-wide rates are 68% and 86%, respectively. If also PCSs of size < 20 bp are included, the rates increase to 71% and 73% for the human imprinted regions and autosomes, respectively, and to 85% and 90%, respectively, for mouse imprinted regions and autosomes. The coverage of CGIs by PCSs as well as their conservation score – calculated as the average of the scores of the overlapping PCSs – are virtually identical for all groups ($p > 0.8$). Thus, the only clear difference between imprinted genes and biallelically expressed ones is the increased association with intronic CGIs.

3.2.5 Enrichment of tandem repeats

Since the existence of tandem repeats has been reported for many imprinted genes, leading to the tandem repeat hypothesis of imprinting (Neumann et al. 1995), and most known repeats were found in CpG islands, we examined if they are significantly enriched in comparison to randomly chosen biallelically expressed genes. Direct tandem repeats were identified in the CGI sequences with *Tandem Repeats Finder* (Benson 1999). Tandem repeats are not only found in single-copy sequences but also in microsatellites and repeats of retroviral origin. Human and murine control CGI sequences contain several tandem repeats connected to LTRs, SINEs, and LINEs. In contrast, murine imprinted sequences do not have such repeats in their CGIs. They are, however, enriched in simple repeats (Waterston et al. 2002, Luedi et al. 2005). In human control sequences there are some tandem repeats associated with the hominid-specific SVA (SINE-VNTR-Alu) retroposon, which contains a variable number of tandem repeats (VNTR) region (Strichman-Almashanu et al. 2002).

Depending on the CGI criteria and repeat masking, variable numbers of tandem repeats are found in a subset of the sequences (Appendix B Tab. B3, B4). As figure 3.5 shows, significantly more imprinted than control sequences contain at least one tandem repeat array in one of their CGIs (χ^2 test, $p < 0.01$), except for human GGF. When excluding retroelements and microsatellites, the enrichment in imprinted genes is even more pronounced ($p < 0.005$). The number of tandem repeats for TJ is very small in mouse, due to fewer CpG islands with stringent criteria and probably more diverged sequences. By concentrating on unique sequences, several tandem repeats were found in or in the vicinity of imprinted genes that had not been reported before (Tab. 3.8). A more detailed comparison of the repeats to the literature can be found in Appendix B Tab. B5.

The general structures of repeat arrays, i.e. motif length (10-140 bp), number of repetitions (1.9-50.5 times) and array length (50-1618 bp) are similar in imprinted and control sequences as well as in mouse and human (Appendix B Tab. B4). Analysis of the consensus sequences with *MEME* and *wordcount* revealed that they do not share a common motif, nor were any motifs found to be over- or underrepresented in the imprinted group. As to be expected for CpG islands, a number of their

tandem repeats contains patterns similar to the binding site motif of the ubiquitous transcription factor SP1, GGGGCGGGG.

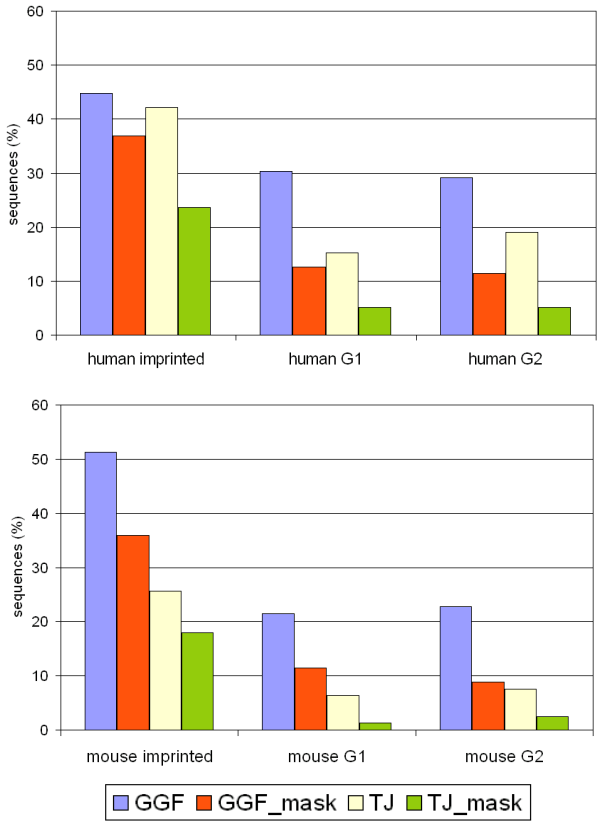


Figure 3.5: Percentage of sequences that possess at least one tandem repeat array in one of their CpG islands

In order to assess conservation, the consensus sequences of each repeat motif were concatenated to a dimer and searched against the genomic sequences of all genes with a local installation of *BLAST*. The murine *Peg3* repeat matches the human *PEG3* gene twice with an e value < 0.001 and more than half of the nucleotides aligned. This conserved repeat has been reported to contain binding sites for YY1 (Kim et al. 2003). A truncated version of the repeat motif in the mouse control sequence *Trrap* was found in human *TRRAP*. The lack of further matches indicates that the identified tandem repeats are generally not conserved and exactly tandemly arranged repeat motifs may be a unique feature of each sequence. Within *phastCons* most conserved elements in imprinted genes, simple tandem repeats are absent for human and only present for four murine PCSs, confirming that the motifs of tandem repeats in the CpG islands of imprinted genes are not conserved between human and mouse.

Table 3.8: Tandem repeat arrays in imprinted sequences

human

gene	location	motif length (bp)	repeti-tions	array length (bp)	consensus sequence	previously identified by
ATP10A	intragenic	53	1.9	103	ATGGCTGAAAACATGGGTGGGGCCCCTCCCCACCTCGC CCGGGTCGTGTTGTG	n
CD81	intragenic	17	15.6	269	TCTTGTGGGGTGGGGCG	Paulsen et al. 2000
	intragenic	31	7.0	221	GCACCCGTGCTGTGGCGTGCCCGTCGTCTGT	Paulsen et al. 2000
	down-stream	30	2.3	69	CTCGGCCTCACCCAGGTGCTCCCGCTTGTG	
GNAS	intragenic	27	8.7	238	GCAGCCCCAGCCGATCCCGACTCCGGG	Coombes et al. 2003
GRB10	down-stream	41	12.2	498	TACTCACACGTGGGACACAGATCCACTCTGCCGTAGCTG TA	
H19	upstream	29	14.5	426	TGTCCCACCCGGGTGACGTGCCGTACCCG	Lewis et al. 2004
	down-stream	39	3.3	127	GGGTGTGCGGGCGATGGGGGAGATGGACAACAGGACC GA	
KCNQ1	intragenic	44	3.6	160	TCCACATGCCCGTCTGCAGCTCGAGAATTAGACGTGCCC TGGGC	
	intragenic	40	1.9	76	GGAATCCTGGGCTGGAACCGGAAACTTCCCCGAGTACA TA	
	intragenic	50	3.9	193	CCTGACTCAGAACCACAACGTGGATTCCCAACTCCGATC CCAATTCGGGC	Paulsen et al. 2005
	intragenic	26	5.8	152	GGGAGGGCCGCGCTGAGGAGCCCCCA	Paulsen et al. 2005
	intragenic	26	4.0	102	AGAACCGCGCCGAAGAACCCCCGGGG	Paulsen et al. 2005
	intragenic	27	9.0	235	CCGAGGAGAACCGCGCTGAGGGGCGC	Paulsen et al. 2005
MAGEL2	upstream	30	6.8	204	TCCCCCTCCGGGGACACCGATGGCTCATCC	
	upstream	21	12.2	257	CCCACCACCGATCCGACAGGC	Boccaccio et al. 1999
MEG3	upstream	13	8.2	106	CAGGCAGCGGTGG	
MEST	intragenic	20	2.9	57	CCTGTGGGGTTTGTGGGCAG	n
	intragenic	37	2.3	85	TTAGGATTTTTAGACCCCGGCATCCCTCTGGTGCGAT	
PEG3	intragenic	84	1.9	162	CAAGCCCCACCCACCTGGGCGCCATCTTTAATGAAAGAG CTTGAGATTTGCCGCGCAGGCGCTGCCCGAATTGGTTG GGCGAGA	Kim et al. 2003
PHLDA2, SLC22A18	down-stream	43	7.0	303	CCGGGGATGGGCTCGGTGGGACAGGCTCGGCCGCAGG CTGCTC	n
	down-stream	36	14.7	529	CTGCCAGCCACCCGAACCCCAGAACCGCACCAGACA	
SNRPN	intragenic	16	6.9	113	GTGGGCATTGGCGGCG	Huq et al. 1997
ZNF264	intragenic	40	2.3	90	GGCGGCGGCCCTGCGTCTGGAACGCCGTTGCCACCGA GGA	n

The location indicates the position of the CpG island with reference to the transcribed portion of the gene (see also Fig. 3.1). The consensus sequence is given as reported by *Tandem Repeats Finder* (Benson 1999). Previous identifications are indicated by the original reference. An "n" means that no tandem repeats have been reported for this genes.

mouse

gene	location	motif length (bp)	repeti-tions	array length (bp)	consensus sequence	previously identified by
Commd1	intragenic	38	5.0	194	CCTGCGCAGTTACCCGGTTATCCGCAGTACGTAGCCAG	Pearsall et al. 1996
	intragenic	45	2.3	106	CTGCGCAGTTACCCGATTATCCAGTTATCCGCAGTACAG GCCTGC	Pearsall et al. 1996
Gnas	intragenic	36	3.3	120	GCCGAGCCTGCCTCCGAGGCAGTCCCTGCCACCCAG	Coombes et al. 2003
Grb10	intragenic	10	32.1	321	GCGTGTCGGC	Arnaud et al. 2003; Hikichi et al. 2003
Gtl2	intragenic	23	3.3	79	GAGGACCCCAGGAAGCCCAGCGC	
H19	upstream	11	9.5	109	GGGGGTATAGT	Lewis et al. 2004
Igf2r	intragenic	31	2.8	88	TCTCCTGCAACGTGGCACTTTTGAGCTCACC	Reinhart et al. 2002
	intragenic	24	11.1	265	CACACACCCACGGCATGGCGGTCT	
Impact	intragenic	140	2.5	362	GCTTTGCTGCATTGTCACATGAGCAGGCCCGGCCCACTC GGCTCGGCTCGGCACAGCTCGGCTGTTGCGTCACTGGC GCCTGCTCGGCTGCGTTGTCACATGTTAGCAAGGCCGA CTAGGCTGCTGCGTCACACGAGCAG	Okamura et al. 2000
Magel2	upstream	33	3.4	113	GCTGAGAGCTGCGGTGCCAGCCAGGCAGCGCTC	
Peg10, Sgce	upstream /promoter	17	3.8	65	CTCCCACCTCCCATCAT	n
	promoter	29	10.8	309	ACTAATGGGCGCTTCATGCGCTACAAAAT	
Peg3	intragenic	21	9.2	196	ATGGAGAGGCTGAAGAGCCAG	
Slc22a18	intragenic	31	2.0	62	AATACACCCACTCTCTCCCGGAGAAAGCAGG	n
Slc22a3	intragenic	44	2.0	88	AGACACACGGGGACATATATGACAGACGGAAGGAAGCT AGCGAC	n
Slc38a4	intragenic	37	9.9	367	GGGATCGGGCTGGGGTTCCCGTGGAGGGACCCTCGCG	R. Smith, pers. commun.

3.2.6 Summary and conclusions of chapter 3.2

In general, the CpG islands of imprinted loci do not differ from those of non-imprinted regions in terms of numbers, lengths, G+C and CpG content, which is in agreement with the literature (Ke et al. 2002a, 2002b, Allen et al. 2003). They also show similar levels of conservation. Applying an estimation of the CpG deamination rate to the CpG islands in imprinted regions revealed that the potential effects of differential methylation are subordinate to species-specific differences. Imprinted genes, however, are enriched in intronic CGIs, indicating that these constitute important regulatory elements. Additionally, significantly more imprinted genes possess tandem repeats in their associated CGIs. Their apparent lack of conservation hints at a recent appearance and suggests that similar epigenetic functions can be triggered by independently evolving DNA sequences. Consequently, the proposed special structure adopted by tandem repeats seems to be more important for epigenetic regulation than conserved sequences.

3.3 Sequence conservation at imprinted loci

In order to investigate other potential regulatory elements of imprinted genes, the focus of further studies was directed on comparative genomics. The analyses concentrate on the well-annotated human and mouse genomes since sequence retrieval for other mammalian species resulted in complications that may be related to a specific chromatin structure of imprinted regions. Starting with an enlarged imprinted set and the three control groups, we identified conserved elements from pairwise alignments of orthologous genes (ECRs). For extending the investigations onto a genome-wide scale, we used annotations of RefSeq genes and *phastCons* most conserved sequences (PCSs) provided by UCSC.

3.3.1 Low recovery rates of orthologs of imprinted genes

The updated set of imprinted genes comprises 58 protein-encoding genes (see Appendix D Tab. D3) and three noncoding large RNAs (*H19*, *MEG3*, *MEG8*) that are orthologous between human and mouse. The three control sets used before for CpG island analyses were reduced to 78 orthologous gene pairs each (omitting *LRRC6*, *TUBA2*, and the *SLC2A14-Slc2a3* pair due to annotation difficulties). The work of Siba Ismael, who worked as a student research assistant on this project, revealed that, using the Ensembl ortholog annotations and *Blast*, fewer imprinted than random control genes from cow, dog, opossum, and platypus could be assigned to chromosomal coordinates, although the difference was not significant (χ^2 test, $p > 0.1$; Appendix C Tab. C1). If an imprinted region is insufficiently sequenced or assembled, several genes are missed at once, e.g. the PWS/AS region in cow, the genes of which are mostly located on unplaced contigs. Sequencing and assembly might generally be more difficult in imprinted regions because of special DNA structures, e.g. chromatin loops and tandem repeats. Another issue are paralogs: The *PEG3* region in dog can be located to chromosome 1 via *Blast* but since there is a larger number of the highly similar zinc finger proteins in the corresponding region than in other species, it is unclear which are the actual orthologs. Opossum and platypus sequences are mostly located on contigs and annotations seem unreliable. For these reasons, the analyses concentrate on human and mouse. Genomic sequences of the longest transcript including 300 kb upstream and downstream were taken from the updated genome builds hg18 and mm8 at UCSC. The cow and dog sequences with known chromosomal location were downloaded as well.Weak conservation of imprinted genes between human and mouse has been reported (Engemann et al. 2000, Paulsen et al. 2000, Paulsen et al. 2001). However, no systematic comparison to biallelically expressed genes has been performed yet. Therefore, I generated *BlastZ* alignments of the human sequence as a reference with mouse, cow, and dog sequences, respectively. Coverage, or alignability, is calculated as the percentage of the human gene that has matches with the aligned species. Sequence identity is the percentage of identical aligned bases. There may be substantial amounts of non-conserved repetitive elements inside the noncoding regions of a gene which cannot be aligned, resulting in low total alignability and identity, although the coding portion may be well conserved. Low alignability can be caused by alternative promoters, differing UTRs and/or introns, species-specific indels like different extends of repetitive elements (Paulsen et al. 2000), and long unsequenced stretches in the aligned species. Nevertheless, human imprinted genes are 4 to 100% alignable to their murine orthologs, with a median of 64%, which is slightly higher than for control genes (Wilcoxon test, $p < 0.1$). The noncoding transcripts show surprisingly good coverage: *H19* is 100% alignable, *MEG3* 88%, and *MEG8* 51%. Without them, maternally expressed genes tend towards a lower coverage and identity than paternally expressed ones ($p < 0.08$), which in turn achieve better

values than control genes ($p < 0.05$). However, their conservation is more variable.

Cow and dog sequences give a better coverage of the human sequences than mouse ones and the identity of the genes is higher ($p < 0.05$). Here, a slightly increased identity of the 21 paternally expressed genes compared to genes in the control sets ($p < 0.1$) is the only trend for a difference between the groups. G+C content and CpG_{obs}/CpG_{exp} are also similar between the groups, with G3 being strikingly CpG poor. Interestingly, the G+C content of the genomic sequence is significantly correlated with the identity of the alignment ($r = 0.5$, $p < 0.001$), making G+C-rich sequences more conserved than G+C-poor ones. Thus, the CpG-rich imprinted genes mentioned in section 3.3.1 may introduce a bias in the imprinted group. In summary, there is no pronounced difference between imprinted and randomly selected genes with respect to the conservation of genomic sequences.

3.3.3 Properties of pairwise and genome-wide conserved elements

Next, we determined pairwise evolutionary conserved regions (ECRs) in the genomic sequences. ECRs are defined as sequence stretches of ≥ 70% identity over at least 100 bp (Elnitski et al. 2002, Loots and Ovcharenko 2005). They were identified from the *BlastZ* alignments mimicking the approach of *PipMaker* (Schwartz et al. 2000; see section 2.4.2). The number of ECRs per sequence is very variable, depending on multiple factors like gene length, numbers and size of introns and coding regions, conservation, and algorithmic effects. Reasons for no ECR being found in a region may either be due to weak or absent conservation, the presence of unsequenced parts, or a shorter transcript in the aligned species. Consistent with their lower conservation, there are fewer ECRs in mouse-human alignments than if cow and dog sequences are aligned to human ones and they are shorter (Wilcoxon test, $p < 0.001$). However, there are still substantial parts unsequenced in both species. ECRs have similar amounts of gaps in all groups ($p > 0.2$). The percentage of ECRs that overlap with repetitive elements is similar between the imprinted and the control groups; it is significantly lower in alignments with mouse (20%) compared to those with cow or dog (42%; χ^2 test, $p < 0.001$). On the one hand, this can be explained by the fast mutation rate, especially deletions, in the rodent lineage (Waterston et al. 2002). Additionally, since there are more ECRs in cow and dog, the probability of extending an ECR into a repetitive sequence is higher.

Most of the imprinted genes are located in clusters. Thus there is a large overlap of their +/-300 kb genomic environments, resulting in the repeated detection of the same ECRs. If overlapping regions are omitted, the sequences in the imprinted group sum up to barely 50% of the total sequence length in any of the control groups, which rarely have an overlap. Consequently, the number of non-redundant intergenic ECRs is lower in the imprinted group. Imprinted genes, however, are not closer to each other than control genes to their neighbors (Wilcoxon test, $p > 0.1$). In order to exclude ECRs in promoter and coding regions of neighboring genes, intergenic ECRs must be at least 1 kb distant from the next gene (Fig. 3.6). Under these conditions, the density of ECRs is similar in all groups: There are two human-mouse ECRs per 10 kb in intergenic regions and four ECRs per 10 kb in intragenic regions. The latter include both intronic and exonic sequences.

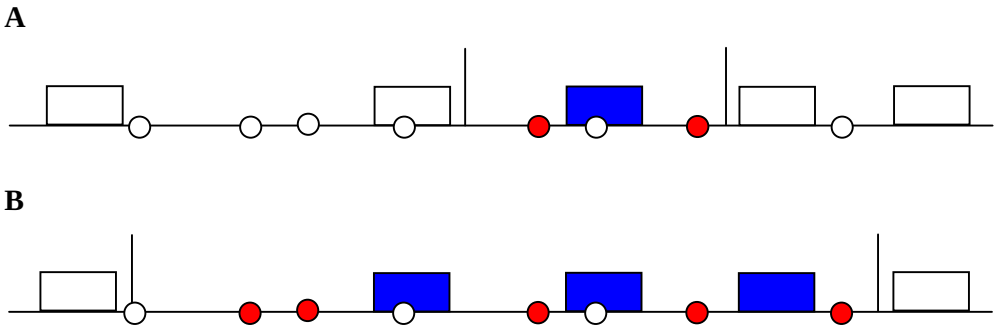


Figure 3.6: Classification of conserved elements
Boxes represent genes, circles stand for conserved elements. (A) To be counted as an intergenic conserved element (red circle) associated with a reference gene (blue box), the element must be at least 1000 bp distant from the next gene (cutoff shown by vertical bars). (B) In the case of imprinting clusters, the cutoff refers to the non-imprinted neighboring genes.

Attempts to find ECRs in common between all four mammals were hampered by incomplete sequences. For example, a considerable part of intronic human-mouse ECRs in imprinted genes did not correspond to human-cow or human-dog ECRs. Additionally, it became desirable to perform analyses on a genome-wide scale. The obvious choice for this purpose were the *phastCons* 28wayPlacMammal most conserved sequences (PCSs; Siepel et al. 2005) provided by UCSC. PCSs are determined from multiple alignments of the genomes of 18 placental mammals via a Hidden Markov model and projected onto each genome. Thus, their conservation score is independent of any reference, but the resulting length may deviate from that in the original alignment due to insertions or deletions in the species the PCS is projected onto. As a part of his work on the project, student research assistant Matthias Bieg implemented a binary search algorithm to calculate overlaps of genomic coordinates. Using it, the 1,271,956 PCSs with length ≥ 20 bp were mapped to all 17,916 autosomal protein-coding human (hg18) genes from the UCSC RefSeq Genes track, for which the longest possible transcripts were constructed. PCSs inside genes are divided into intronic, coding exonic and non-coding exonic. A PCS overlapping the transcriptional start site of a gene is termed promoter PCS. A conserved element is called exonic if it overlaps by at least 1 bp with an exon, else, it is intronic. To be coding, it must cover at least 1 bp of coding sequence. Intergenic PCSs reside between genes and are assigned to the next neighboring gene. All sequence features are calculated based on the human genome. Since *PEG3* is a transcriptional variant of *ZIM2* in the human genome, the effective gene number of the imprinted group is 57.

Compared to genome-wide data, similar numbers of PCSs per gene were detected in the imprinted regions, which possess 3969 PCSs (Wilcoxon test, $p > 0.8$). The median number of PCSs per gene is 16 for the imprinted group and 14 genome-wide. PCSs associated with imprinted genes stand out as exceptionally G+C and CpG-rich (Fig. 3.7, Appendix C Tab. C2). The number of PCSs that contain at least one CpG is also higher in the imprinted group than for the genome (42% vs. 36%, χ^2 test, $p < 0.001$).

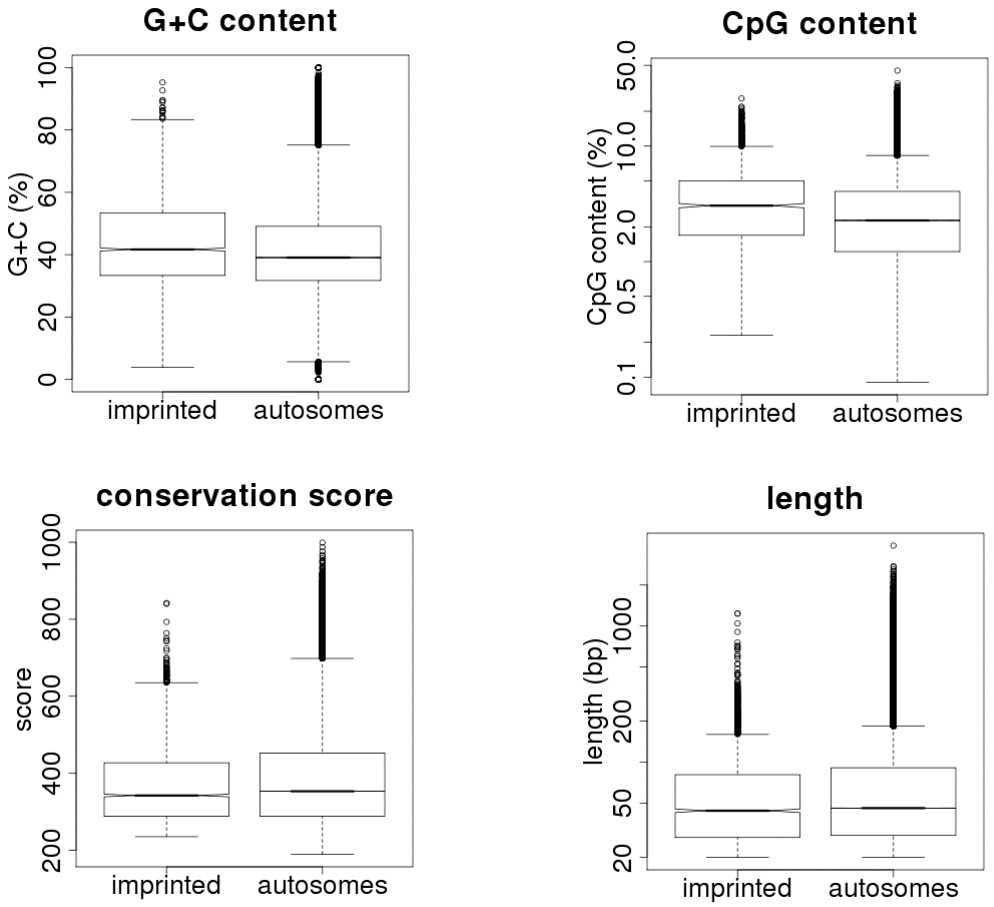


Figure 3.7: Boxplots for discerning features of human *phastCons* sequences
PCSs in imprinted regions are G+C and CpG-richer, less conserved and shorter than PCSs on human autosomes. Although the differences of the medians are rather small, they are significant (Wilcoxon test, $p < 0.001$). For the boxplot of the CpG content, PCSs without CpG were omitted. Note that for CpG content and length, features where the majority of PCSs is concentrated at small values, the y axes are in logarithmic scale to transform the highly skewed distributions into a more convenient form.

Unexpectedly, PCSs in imprinted regions have lower conservation scores and are shorter than those in the entire autosomal genome (Fig. 3.7, Appendix C Tab. C2). The correlation between G+C content and conservation score is, unlike the high correlation between G+C content and the proportion of identical alignment positions of genomic sequences observed in section 3.3.1, very low for PCSs ($r = 0.13$). Additionally, the (TpG+CpA)/(2·CpG) ratio of all PCSs in imprinted regions is lower than for PCSs in the whole genome (Wilcoxon test, $p < 0.001$; Fig. 3.8). This effect is mostly caused by lower ratios in intronic and intergenic regions.

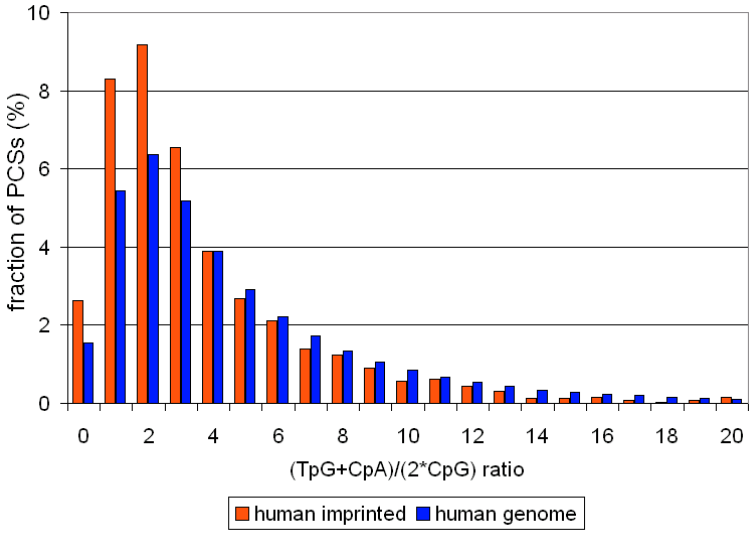


Figure 3.8: (TpG+CpA)/(2·CpG) ratio in human *phastCons* sequences
The (TpG+CpA)/(2·CpG) ratio serves as an estimation of the CpG deamination rate. Since it is not defined for sequences with 0 CpGs, PCSs that do not contain a CpG dinucleotide were omitted.

To verify that these results are not biased by the properties of the human genome, the analyses were repeated for the mouse genome (mm9), which is CpG-poorer. 18,772 genes including 53 imprinted mouse orthologs are annotated as RefSeq genes on autosomes. In the corresponding phastCons30wayPlacMammal track there are 1,268,568 highly conserved PCS elements ≥ 20 bp, of which 3502 reside in the vicinity of the imprinted genes. Although some differences exist, the general features are essentially the same (Appendix C Tab. C3). Likewise, PCSs in intergenic regions and introns behave similarly to all PCSs (Appendix C Tab. C4).

The significantly elevated association of PCSs with CpG islands in imprinted regions (Tab. 3.9) has already been mentioned in section 3.2.4. Whereas for introns it can be attributed to an enrichment of intronic CGIs, the accumulation of intergenic CGIs in the imprinted group is not significant compared to the human and mouse genomes (χ^2 test, $p > 0.2$). The estimated deamination ratio of PCSs in CGIs is slightly though not significantly higher in imprinted regions (median 0.73) than the genome-wide ratio (median 0.68; Wilcoxon test, $p > 0.3$). Hence, the CpG island character of these PCSs is apparently not caused by reduced CpG deamination rates due to germ line specific methylation patterns. Moreover, imprinted regions possess more PCSs that overlap with repetitive elements, indicating that these are better conserved than in non-imprinted regions. The enrichment is pronounced in intergenic and intronic regions as well as in both species (χ^2 test, $p < 0.01$; Tab. 3.9).

Table 3.9: Overlap of *phastCons* sequences with CpG islands and repetitive elements

human	**PCSs in imprinted regions**			**PCSs genome-wide**		
location	total	with repeat overlap (%)	with CpG island overlap (%)	total	with repeat overlap (%)	with CpG island overlap (%)
intronic	1120	14.0	7.1	365,258	9.9	2.2
intergenic	1787	13.3	4.6	588,309	9.8	2.6
coding exons	1015	4.8	13.0	306,508	3.3	9.4
all	3969	11.5	8.1	1,271,956	8.4	4.4
mouse	**PCSs in imprinted regions**			**PCSs genome-wide**		
location	total	with repeat overlap (%)	with CpG island overlap (%)	total	with repeat overlap (%)	with CpG island overlap (%)
intronic	1141	14.3	4.3	348,063	8.3	1.7
intergenic	1230	12.0	3.0	606,271	8.8	2.0
coding exons	1072	3.6	7.4	302,787	3.1	5.4
all	3502	10.3	5.7	1,268,568	7.4	3.1

As figure 3.9 shows, LINE elements (L1, L2 and other LINEs) are the most abundant repeat class with regard to overlap with PCSs. Compared to the autosomal genomes, significantly more PCSs in the imprinted regions of both human and mouse are located inside LINEs (χ^2 test, $p < 0.005$). The same holds for the subset of intergenic PCSs ($p < 0.05$ for human, $p < 0.005$ for mouse, respectively). There is also a marked enrichment of the L1 subclass of LINEs for human imprinted regions ($p < 0.001$) whereas for the mouse, it is only tentative ($p < 0.1$). Additionally, overlap with SINEs and LTRs is reduced for PCSs in intergenic and all murine imprinted regions ($p < 0.001$ and $p < 0.05$, respectively) but not in human ($p > 0.1$). Other repetitive elements are not discerningly distributed.

The data obtained for ECRs are similar to the described results for the PCSs: The densities of intergenic and intragenic ECRs associated with imprinted genes are similar to those of the control groups. ECRs of imprinted genes exhibit reduced (TpG+CpA)/(2·CpG) ratios and significantly more ECRs overlap with CpG islands in the imprinted group than in the control groups. The ECRs that overlap with CpG islands show similar (TpG+CpA)/(2·CpG) ratios in imprinted genes and in the control groups. In contrast, the ratio of repeat-overlapping ECRs is not significantly elevated in the imprinted group for human-mouse (21% vs. 18-205; χ^2 test, $p > 0.1$). However, there is an enrichment of LINE-1 in intronic ($p < 0.01$) and, more pronouncedly, in intergenic ECRs ($p < 0.005$). We also identified the PCSs for the control genes and and found the results consistent with genome-wide analyses. Thus, the differences depend on the alternative method rather than on the choice of the control groups.

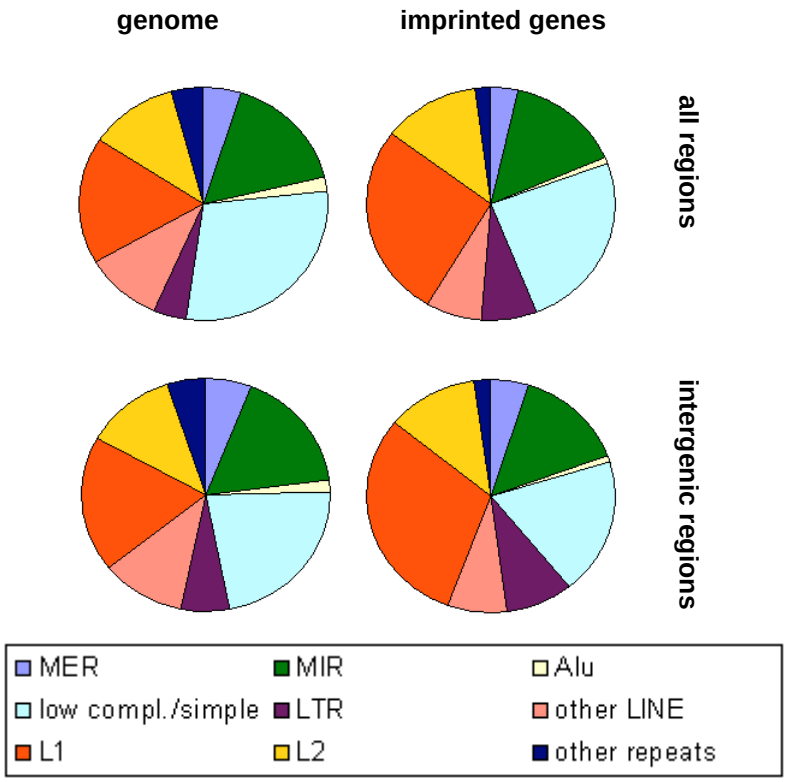


Figure 3.9: Distribution of repetitive elements in *phastCons* sequences
PCSs from the human genome that contain at least 1 bp of a repetitive element were summed up in categories according to the first overlapping repeat. The most prominent enrichment is that of PCSs with LINE-1 repeats (L1, red) in intergenic regions of imprinted genes relative to the genome.

3.3.4 Features of the promoter regions of imprinted genes

Promoter regions harbor transcriptional signals, which may be distinguishing for imprinted genes. Not all of the promoters in the imprinted group are actually imprinted. Sometimes, the imprinted transcript arises from an alternative promoter further downstream as it is the case with *WT1* alternative transcript (Hancock et al. 2007), or *Begain* (Tierling et al. 2009). In other cases, tissue-specific promoters are found far upstream (Chotalia et al. 2009). Even imprinted promoters are not always differentially methylated in clusters with shared regulatory elements (Lin et al. 2003, Parker-Katiraee et al. 2007, Ruf et al. 2007). Thus, taking the most 5' annotated transcriptional start site (TSS) as for biallelically expressed genes should yield unbiased results for the investigation of typical promoter features.

In total, 9794 (55%) of the most upstream annotated transcriptional start sites coincide with CpG islands but only 3711 (38%) of these CpG islands also overlap with PCSs. Imprinted genes show virtually identical ratios: 54% of their most upstream promoters are located in CpG islands, of which 42% overlap with PCSs. Only 30% of the human genes, likewise 16 imprinted ones, have a conserved region (PCS) overlapping their most 5' TSS. This is not surprising since turnover at transcription start sites is a general property of mammalian genomes (Frith et al. 2006).

As the requirements for ECRs are less stringent than those for *phastCons* sequences, an ECR can consist of several PCSs or not coincide with a PCS. Regarding ECRs, the number of conserved

promoters in human-mouse alignments is higher than that of PCSs (Tab. 3.10), thus they are better suited for further analysis. Far more than half of the promoter ECRs coincide with CpG islands in all groups. Promoter ECRs are of highly variable length and can even extend throughout highly conserved intronless genes like *KLF14*. Nevertheless, neither PCSs nor ECRs in promoter regions of imprinted genes do stand out from those of the genome or control genes with respect to lengths, CpG and G+C content, CpG_{obs}/CpG_{exp}, (TpG+CpA)/(2·CpG) ratio, or conservation scores (Appendix C Tab. C4). Also when analyzing conserved promoter regions defined as the part of human sequences from -1000 to the transcriptional start site that could be aligned to mouse with *BlastZ*, these features were similar or showed no consistent differences between the imprinted and control groups.

Table 3.10: Conserved promoter regions

group	genes	PCS	ECR	alignable
imprinted	61/57[a]	16	28/25[a]	56/53[a]
G1	78	47	35	73
G2	78	30	47	74
G3	78	21	33	74

[a] The numbers on the left refer to data with the noncoding RNAs *H19*, *MEG3*, and *MEG8*, the numbers on the right to protein coding imprinted genes only.

When applying *EpiGraph* (Bock et al. 2009) on the complete -1000 to TSS regions, the recombination rate was found to be significantly elevated in the imprinted group compared to all controls. Imprinted regions have indeed been reported to contain recombination hot spots (Sandovici et al. 2006). Overlap of the examined regions with the histone modification H2A.Z, which is found in promoter regions as well as in the body of transcriptionally silenced genes (Barski et al. 2007), was not consistently reduced. No sequence patterns or DNA structure predictions showed up with discriminative features. Patterns of four nucleotides as analyzed by *EpiGraph* may be too short to be meaningful. Thus, for 6-mers reported by *wordcount* to have a large difference of occurrences in the imprinted set versus all control sets, the presence or absence in individual conserved promoter regions was counted. Although some, especially GA-rich, 6-mers seem to be depleted in the imprinted group, the distribution is not significantly different (χ^2 test, $p > 0.1$). However, cccccc is overrepresented ($p < 0.05$). The program *K-Factor* (Lee et al. 2007) did not report any 6-mers to be consistently enriched in the imprinted set compared to control sets. Additionally, the occurrence of TATA boxes as well as that of consensus binding sites of SP1 and the zinc finger protein encoded by the *Plagl1* gene, which has been reported as binding at C4G4 motifs in the *H19* and *KCNQ1* regions (Varrault et al. 2006, Arima et al. 2005), were examined but did not show significant differences. Intriguingly, the occurrence of specific motifs is not conserved between the species.

On a genome-wide scale, we investigated the association of the human most upstream promoter regions with regulatory elements from the ORegAnno database of known regulatory regions (Montgomery et al. 2006, Griffith et al. 2008). Only 8.6% of all autosomal genes have experimentally determined transcription factor binding sites (TFBSs) annotated. Similarly, this is

the case for five of the 58 imprinted genes (*IGF2*, *INS*, *PEG10*, *PHLDA2*, and *SGCE*). With the exception of *INS*, there are exclusively CTCF binding sites annotated for them. The ratio is almost identical to the genome, where 1255 of the 1541 genes that have an ORegAnno TFBS in their promoter possess at least one CTCF binding site. Thus, promoter-associated CTCF binding sites are not discerningly distributed. Investigating the annotated putative TFBSs that are conserved in human, mouse, and rat for the 1000 bp upstream of the most upstream transcriptional start site, we assigned at least one conserved TFBS to 29 (50%) of the human imprinted genes, which is an insignificantly lower ratio compared to the genome-wide one (10,805 genes, 60%). Promoter regions of both groups contain on average two TFBSs. TATA boxes and predicted YY1 binding sites are similarly rarely represented in imprinted and autosomal genes (χ^2 test, $p > 0.4$). According to *K-Factor*, there are two motifs with a significant enrichment (*K-Factor* score ≥ 3.5) in the regions 1000 bp upstream of the transcriptional start site in human imprinted genes (tgcgta and gcgtat) and seven different ones in mouse imprinted genes (atagcg, atcgca, cgtacg, ctacga, tgcgtg, tgtcga, ttggcg). Most of these motifs share the feature of having both a TpG and a CpG dinucleotide, indicating their CpG island association and possible deamination effects.

3.3.5 CpG-rich motifs in intragenic and intronic conserved regions

Conserved elements in the introns of imprinted genes are exceptionally G+C rich (median 38.89% compared to 36.54% based on the human genome; Wilcoxon test, $p < 0.0001$). The imprinted group has more intronic PCSs that contain a CpG dinucleotide than the genome (32% and 25% in human, 40% and 33% in mouse, respectively; χ^2 test, $p < 0.001$) and their CpG content is significantly higher (Wilcoxon test, $p < 0.001$; Appendix C Tab. C4), consistent with the before mentioned enrichment of intronic CpG islands. To investigate whether there are sequence patterns distinguishing for imprinted genes, I concatenated the sequences of all intronic PCSs per gene, separated by 6 Ns each to prevent artificial sequence combinations. *K-Factor* identified a large number of 6 bp motifs that are overrepresented in the imprinted set compared to both the pre-calculated genomic background and the autosomal conserved intronic sequences. All of them contain at least one CpG whereas TpG and CpA are rare, which is in accordance with the lowered (TpG+CpA/(2·CpG) ratios of intronic PCSs. Converting repetitive elements to Ns to exclude potential motifs in repeats did not alter the motifs and only marginally influenced their scores. Table 3.11 shows the ten 6-mers that show a significant enrichment (*K-Factor* score > 3.5) in both human and mouse imprinted sets.

Similar motifs were detected for the concatenated intragenic PCSs, which are likewise enriched in G+C and CpG (Appendix C Tab. C4). In general, there were no genome-wide overrepresented 6-mers that were underrepresented in the imprinted groups. Of the 4^8 possible 8-mers, there are as many as 1,486 overrepresented in the conserved intronic sequences of both human and mouse imprinted genes. They mostly contain two or even three CpGs. Since they might overlap in the sequences, clustering would be required for a better interpretation. It is most likely that CpG rich sequences, just as CpG islands, provide an open chromatin structure associated with promoters of alternative and antisense transcripts. Although comparisons of the imprinting-specific 6-mers to CpG rich binding site motifs of transcription factors like SP1, CTCF, and YY1 did not show congruencies, they might represent alternative patterns. Kim et al. (2007) reported that 75% of the CTCF binding sites identified by ChIP-on-Chip experiments share a common motif (Fig. 3.10) – to which several of the overrepresented 6-mers can be aligned – whereas the rest is highly divergent. This is not surprising because CTCF possesses multiple zinc fingers which can alternate in

contacting the DNA, making its possible binding sites very diverse (Loukinov et al. 2002). Nevertheless, evolutionary conservation allows to predict the seemingly most crucial ones (Xie et al. 2007, Kang et al. 2009).

Table 3.11: 6-mers enriched in intronic *phastCons* sequences of imprinted genes

	human imprinted		**mouse imprinted**	
6-mer	score against genomic background	score against autosomes	score against human genomic background	score against autosomes
cgc**cg**c	6.21	4.08	5.91	4.04
cgcgac	3.51	4.31	3.64	3.82
gc**cgcg**	5.79	4.47	5.46	4.39
gc**cg**tc	3.62	4.32	5.37	4.70
g**cg**c**cg**	6.02	3.65	9.36	5.92
g**cg**t**cg**	4.31	6.65	5.96	6.52
gggc**cg**	3.82	4.01	3.70	3.99
gggg**cg**	5.30	4.28	4.75	3.61
gt**cgcg**	10.53	7.32	10.92	5.60
tc**cgcg**	4.51	4.18	4.68	3.96

Analyses of transcription factor binding sites not restricted to conserved elements revealed that 20 human imprinted genes (34.48%) and 25.76% of the autosomal ones have at least one experimentally verified TFBS annotated by ORegAnno in one of their introns (χ^2 test, $p > 0.1$). The rates for intergenic TFBSs are 48.28% and 47.99%, respectively ($p > 0.9$). CTCF binding sites are present in the introns of 20 imprinted genes, which is a slight enrichment compared to 22.43% of the autosomal genes ($p < 0.05$). With regard to intergenic regions, 27 imprinted genes (46.55%) and 7588 autosomal genes (42.35%) have a nearby CTCF binding site ($p > 0.6$). In total, CTCF binding sites are found in or in the vicinity of 66% of the human imprinted genes and 54% of the autosomal ones ($p > 0.1$). Thus, intronic CTCF binding sites might be most distinguishing. Unfortunately, genome-wide CTCF data are missing for the mouse and, due to the lack of a binding site matrix for CTCF, not available in the UCSC tfbsCons track that annotates putative TFBSs that are conserved in human, mouse, and rat.

Predicted YY1 binding sites from tfbsCons are found in introns of ten imprinted imprinted genes, including both previously reported and new cases, and 1744 autosomal genes (17.24% vs. 9.73%, $p < 0.1$). In intergenic regions, another 20 imprinted genes have such a site, a notable enrichment compared to the autosomes with 24% ($p < 0.005$). If all locations are taken into account, the ratio increases to 52% for imprinted and 31% for autosomal genes ($p < 0.005$). Since CTCF and YY1 interact (Donohoe et al. 2007), a combined occurrence of TFBS for both proteins might be particularly meaningful. Indeed, this is the case for 40% of the human imprinted genes as opposed to 21% on autosomes ($p < 0.005$). Ongoing research is focussing on the overlap of conserved regions with annotated regulatory elements and histone modifications.

Figure 3.10: CTCF binding site motifs
The logo for CTCF binding sites, taken from Kim TH et al. (2007), is derived from ChIP-on-Chip experiments and compared to the previous consensus sequence (Bell and Felsenfeld 2000). Notably, the positions of CpGs – methylation of which abolishes CTCF binding – are different between the previous and the new consensus.

3.3.6 Weak conservation of exonic sequences

Coding regions are supposed to be subject to selective constraints due to protein function. As imprinted genes encode proteins that are important for embryonic development, one would expect them to be highly conserved. Nevertheless, the 1024 PCSs overlapping with coding exons of imprinted genes are significantly shorter and achieve lower conservation scores than all 309,941 coding PCSs as well as randomly sampled PCS groups of the same size (Wilcoxon test, $p < 0.0002$). Closer investigation as shown in figure 3.10 revealed that these differences are especially caused by the subset of 538 coding PCSs of the 28 genes with maternal expression in human. In contrast, those in the 29 paternally expressed genes only tend to have lower scores compared to the genome-wide values ($p < 0.09$) and higher scores than maternally expressed ones ($p < 0.08$), but are not significantly different from both in terms of length ($p > 0.2$). Length, however, might be biased by projection onto the human genome as the length of a PCS sometimes deviates from that in the original alignment due to insertions or deletions in the species the PCS is projected onto. Therefore, we repeated the analyses for mouse PCSs where the set of paternally expressed genes is reduced because some are not annotated as RefSeq genes and *Grb10* and *Copg2* show maternal expression. Although paternally expressed murine genes show similar scores but increased length compared to maternally expressed ones, the other findings are completely in line with the results based on human, thus excluding a projection bias.

Suspecting that conservation of non-mammalian orthologs could also be different between imprinted and biallelically expressed genes, we investigated if there was an enrichment of mammalian-specific PCSs. Such PCS that are only conserved in the 18 mammals subset, but not in the whole 28 vertebrates set were termed mPCSs and identified as depicted in Fig. 3.12. The simple approach has a disadvantage: If orthologous genes are absent in some species, e.g. non-mammals, high conservation of the existing ones is sufficient to give rise to PCSs (compare section 2.4.2). Thus, the selected mPCSs are truly mammalian-specific but exclude an unknown number of false negatives. In the human genome, there are 37,629 mPCSs, 48% of them in intergenic and 42% in intronic regions. Only 3,532 (6%) overlap with coding exons of 17,916 genes. With a similar distribution, 66 mPCSs are assigned to imprinted genes. Since only six mPCSs are located in the coding regions of six imprinted genes, there is neither an enrichment nor a significant depletion of mammalian-specific coding PCSs (χ^2 test, $p > 0.4$).

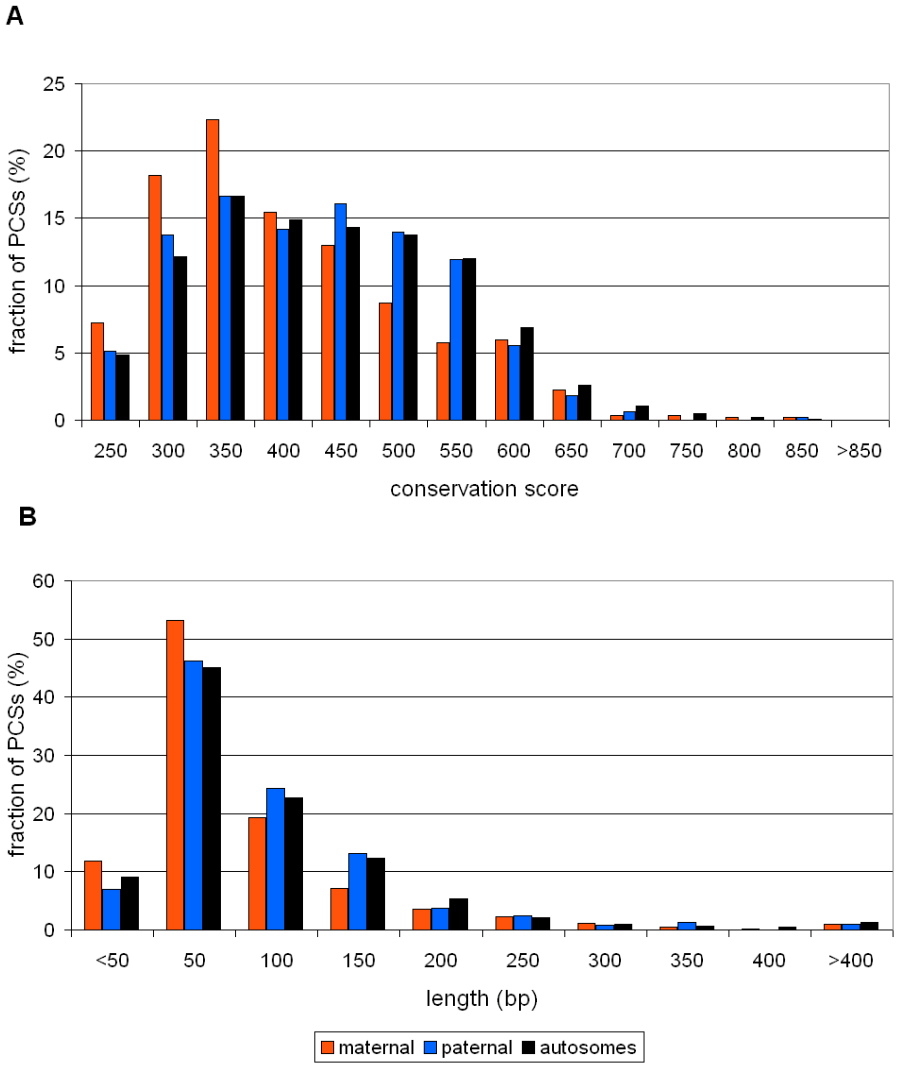


Figure 3.11: Conservation score and length of exonic *phastCons* sequences
Using the human genome as a reference, the conservation score (A) and length (B) of PCSs that overlap with coding exons were determined. Compared to genome-wide data (black bars), the PCSs of maternally expressed genes (red bars) are shorter and achieve lower scores whereas PCSs of paternally expressed ones (blue bars) are similar to PCSs of biallelically expressed genes.

We next asked whether the low conservation could be attributed to increased divergence in a particular mammalian lineage. Rodents spring to mind as they are fast-evolving (Waterston et al. 2002, Gibbs et al. 2004) and seem to exhibit particularly strict imprinting that might affect conservation (Morison et al. 2005, Monk et al. 2006). Moreover, in the *phastCons* Hidden Markov model, where sequences are weighed according to their evolutionary distances, the contribution of rodent sequences is relatively high. Using the existing pairwise ECRs is a logical way to peruse the hypothesis. There is no significant difference seen between imprinted and control genes with respect to coding ECRs derived from human-cow and human-dog alignments (Wilcoxon test, $p > 0.2$). In contrast, coding ECRs in human-mouse imprinted genes, namely in the maternally

expressed ones, have lower identities than the control ECRs ($p < 0.02$). They are not significantly shorter ($p > 0.4$), which may result from the procedure of finding ECRs with a lower limit of 100 bp for the length. This is considerably above the median length of PCSs, 67 bp, whereas the median ECRs length is 200 bp. Paternally expressed genes are similar to the control genes. The results obtained for PCSs in the control genes are consistent with genome-wide analyses. The ECR approach demonstrates that, although decreased conservation of imprinted genes in other species may also contribute to the reduced PCS scores, maternally expressed murine and human genes are highly diverged.

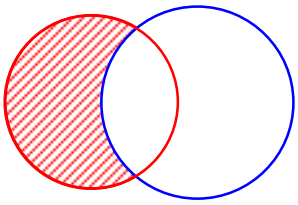


Figure 3.12: Identification of mammalian-specific *phastCons* elements
The set of mammalian-specific coding *phastCons* elements (hatched) is the complement of the phastConsElements28way set (blue circle) relative to the phastConsElements28wayPlac-Mammal set (red circle). mPCSs comprise all elements that are highly conserved in the 18 mammals but do not fulfill the conservation requirements in the whole 28 vertebrates set. From these mPCSs, we chose the subset with a length of at least 20 bp.

Interestingly, the overlap of the analyzed PCSs with coding exons is higher for imprinted genes compared to the rate for all protein-coding human genes (Wilcoxon test, $p < 0.0001$). In order to distinguish between the contribution of protein-coding sequences and adjacent intronic parts to PCSs, I looked at the subsets of PCSs that are completely located in coding exons. They comprise 51% of those in the imprinted group and 41% of the autosomal ones (χ^2 test, $p < 0.001$). Here, the weak conservation of PCSs in all imprinted genes and in maternally expressed genes is less significant (Wilcoxon test, $p < 0.02$). Only a trend remained for the difference in length for maternally expressed genes ($p < 0.06$). Together with the increased exon overlap rate this implies that intronic sequences near exon boundaries, for example splice signals, contribute substantially to the differences between PCSs in coding exons of imprinted genes and those of biallelically expressed genes.

3.3.7 Summary and conclusions of chapter 3.3

Imprinted regions have a similar content of evolutionarily conserved elements as the whole human genome contents but these elements have lower conservation scores and are shorter. PCSs that are associated with imprinted genes also stand out as being G+C and CpG-rich. They overlap more frequently with CpG islands and ancient repetitive elements, particularly LINEs, which indicates that not CpG islands and the distribution of repeats *per se*, but conserved ones constitute important elements in imprinting. The estimated deamination rate of conserved elements in the imprinted group is lowered compared to the genome. Conserved intergenic and intronic regions are enriched in CpG-rich motifs, arguing for an open chromatin structure and an enrichment of antisense or alternative transcripts. Most interestingly, the low conservation of conserved elements in coding exons of genes with a maternal expression pattern may hint at a specific mode of evolution.

3.4 Divergence and conservation of protein-coding imprinted genes

Reduced conservation of conserved elements in the coding regions of imprinted genes implies that protein conservation should also be reduced. In order to find out why they might be so divergent, the obvious next step was trying to assess the evolutionary history of protein-coding regions by using genome-wide data from HomoloGene. Furthermore, since imprinted genes have been frequently linked to paralogs, we investigated whether the existence of paralogs had an influence on the conservation of protein-coding genes.

3.4.1 Contrasting evolution of rodent imprinted genes

The evolutionary history of imprinted genes in placental mammals was investigated on a genome-wide scale using data from HomoloGene release 62, a database that provides information on the conservation of orthologous RefSeq cDNA and protein sequences derived from pairwise alignments. Matthias Bieg wrote a parser for the XML file to extract identity values of human DNA and amino acid sequences, respectively, with orthologs of mouse, rat, chimpanzee, dog, cow, chicken, and zebrafish. As orthologous sequences are not always available, comparably low numbers were obtained for statistical analyses, which did not reach the high significance level observed for the conserved elements.

Table 3.12 shows the data for human-mouse orthologous gene pairs. Compared to genome-wide data, maternally expressed genes are less conserved in terms of cDNA identity (Wilcoxon test, $p < 0.05$) and the proteins encoded by them show a trend towards reduced identity ($p < 0.06$). The comparisons to control groups are slightly more significant, which is not surprising because the proteins encoded by genes in G1 and G2 are more conserved than the genomic average ($p < 0.05$). In contrast, paternally expressed genes are not significantly worse conserved ($p > 0.15$). This is in accordance with what one would expect from the reduced scores and identities of exonic conserved elements. There is a high correlation between the average PCS score of the imprinted genes and the identity of their human-mouse orthologs ($r = 0.64$).

Table 3.12: HomoloGene data for human-mouse orthologous gene pairs

group	genes	protein ID ± std.dev. (%)	cDNA ID ± std.dev. (%)	Ka/Ks ± std.dev.	Ks ± std.dev.
imprinted	53	83.7±11.3(*)	83.4±6.4(*)	0.148±0.113(*)	0.655±0.232
maternal expr.	26	82.5±10.1*(**)	82.5±5.9**(**)	0.161±0.116*(*)	0.674±0.179
paternal expr.	27	84.8±12.4	84.3±6.9	0.136±0.110	0.639±0.272
genome	16,582[a]	85.6±11.7	84.4±6.5	0.129±0.109	0.641±0.227
X chromosome	538	85.4±13.6	85.2±7.6	0.141±0.127	0.584±0.255***
G1	75	88.7±9.2*	85.6±5.0	0.100±0.084*	0.654±0.225
G2	75	88.1±10.9*	85.4±6.1	0.104±0.102*	0.638±0.211
G3	76	87.5±8.8	85.3±4.5	0.105±0.074	0.651±0.163

[a] Ka/Ks rates are undetermined for 21 genes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$ (Wilcoxon test for comparison of the respective group to genome data). The significance level in comparison to control groups is given in parentheses.

Conservation between human and chimpanzee, dog, cow, chicken, and zebrafish is similar for imprinted and biallelically expressed genes ($p > 0.1$, Appendix D Tab. D1). Consequently, increased divergence at imprinted loci did not take place in the human lineage but in the mouse or

in rodents in general. Further evidence for the latter hypothesis is given by human-rat gene pairs, for which the data are consistent with those from mouse (Tab. 3.13). Here, the reduced conservation of maternally expressed genes is even more significant, which can be attributed to a higher number of substitutions in the rat (Gibbs et al. 2004).

When basing the same analyses on mouse as the reference, also the conservation of imprinted genes with cow is lower than that of non-imprinted genes on protein and cDNA level ($p < 0.04$; Appendix D Tab. D1). Comparisons with other species are inconclusive, probably due to a high number of predicted sequences in their genomes and the rather large evolutionary distance. However, they do not contradict the suggested increased divergence of murine imprinted genes. Dog and chimpanzee data comprise more than 95% of predicted sequences and thus have to be considered with caution. Genome-wide, mouse-chimpanzee orthologs are less conserved than mouse-human orthologs ($p < 0.02$) but in the limited sets of imprinted or control genes, they show virtually identical values ($p > 0.6$). For G1 and G2, the results of mouse-chimpanzee conservation are consistent with those obtained for human-mouse, showing a trend for reduced protein and DNA identity of genes with maternal expression in mouse ($p < 0.07$). For G3 and the genome, there are no significant differences.

Intriguingly, when comparing mouse and rat orthologs (Tab. 3.14), it is obvious that, whereas the protein identity is not significantly elevated ($p > 0.6$), DNA conservation of the imprinted group is above the genome-wide level (median 93.80%, $p < 0.02$). This is caused by the paternally expressed genes (median 94.45%) whereas maternally expressed ones (median 94.1%) do not exhibit higher conservation. Also here, control groups proved different from genome-wide data by exhibiting higher protein identities ($p < 0.08$). The distributions of both protein and cDNA identity are similar between control groups and imprinted genes ($p>0.18$). Notably, in all comparisons the standard deviation is always smallest for imprinted genes. This indicates that this group represents a quite homogeneous set on which evolution seems to have acted equivalently. Given the marked, if not even elevated conservation of imprinted genes between the two rodent species as opposed to their divergence from other mammals, most of the discerning DNA changes must have taken place before the split of rat and mouse.

Table 3.13: HomoloGene data for human-rat orthologous gene pairs

group	genes	protein ID ± std.dev. (%)	cDNA ID ± std.dev. (%)	Ka/Ks ± std.dev.	Ks ± std.dev.
imprinted	47/46[a]	83.2±11.6*	83.1±6.4	0.152±0.113*	0.186±0.074***
maternal expr.	24	81.7±10.5**	82.0±5.7**	0.164±0.115*	0.192±0.069*
paternal expr.	23	84.6±12.7	84.2±7.1	0.139±0.111	0.178±0.081**
genome	15,146/15,131[a]	85.5±11.7	84.2±6.4	0.128±0.109	0.229±0.107

[a] The second number refers to sequences available in the HomoloGene database for Ks and Ka/Ks analyses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$ (Wilcoxon test for comparison of the respective group to genome data)

3.4.2. Divergence at the base of rodent imprinting

Relaxation of the selective constraints that act on protein-coding sequences implies that evolution might have favored changes in imprinted genes of a common ancestor of mouse and rat. The different kinds of selection are commonly estimated by comparing the rates of synonymous (Ks) and non-synonymous substitutions (Ka) per site derived from pairwise alignments of coding DNA (Nei and Gojobori 1986, Yang and Bielawski 2000). In general, Ka/Ks ratios of below 0.25 indicate strict purifying or negative selection, by which in spite of changes at the DNA level the amino acid sequence is largely maintained. Ka/Ks > 1 is indicative of positive selection, also known as Darwinian selection, which favors DNA mutations that lead to changes of the protein sequence. A value of 1 suggests neutral evolution with relaxed constraints. It must be considered that for very distantly related species, Ka/Ks is often undefined due to unreliable estimates of the rate of synonymous changes. Therefore the number of sequences available for comparisons is low with chicken and zebrafish.

Synonymous substitutions have been used before to assess mutation rates in imprinted genes (McVean and Hurst 1997, Smith and Hurst 1999). For human-mouse and mouse-cow gene pairs, Ks rates are essentially similar in all groups ($p > 0.2$), indicating that there is no specific selective pressure on silent changes. Ka/Ks is thus controlled by the rate of nonsynonymous changes. In line with the high correlation between Ka/Ks and Ka ($r = 0.86$ for human-mouse and $r = 0.85$ for mouse-cow orthologs, respectively), sequences with a high Ka achieve high Ka/Ks ratios. Between human and mouse, Ka/Ks tends to be elevated for the group of 26 maternally expressed genes ($p < 0.08$) but not for the 27 paternally expressed ones ($p > 0.15$; Tab. 3.12). The Ka/Ks ratios of the 38 imprinted orthologs available for mouse and cow are tentatively elevated as well ($p < 0.1$ genome-wide; Appendix D Tab. D1). Ka/Ks ratios in G1 and G2 are lower than in the genome ($p < 0.03$), implying that their above mentioned strong conservation on protein level results from strict purifying selection. Thus, these control groups are obviously not representative for the genome. Taken separately, the 22 pairs of genes showing maternal expression in mouse and their cow orthologs show lower significance, indicating that the 16 paternally expressed ones, although not significantly different from non-imprinted genes, tend to be less conserved between mouse and cow as well. It must be noted that the variance of paternally expressed genes is very high as this set comprises both the most conserved and the most divergent genes. The results support the conclusion that imprinted genes have evolved faster than biallelically expressed genes in the mouse.

Table 3.14: HomoloGene data for mouse-rat orthologous gene pairs

group	genes	protein ID ± std.dev. (%)	cDNA ID ± std.dev. (%)	Ka/Ks ± std.dev.	Ks ± std.dev.
imprinted	46	94.9±3.3	94.4±2.1	0.137±0.091	0.186±0.074***
maternal expr.	26	94.5±3.4	94.1±2.0	0.147±0.092	0.192±0. 069**
paternal expr.	20	95.5±3.1	94.8±2.3**	0.124±0.091	0.178± 0.081**
genome	16,800[a]	93.2±7.1	92.9±4.0	0.147±0.149	0.229±0.107
X chromosome	533	92.8±8.3	93.4±4.6***	0.167±0.170	0.207± 0.175***

[a] Ka/Ks rates are undetermined for 49 genes

** $p < 0.05$, *** $p < 0.001$ (Wilcoxon test for comparison of the respective group to genome data)

Between mouse and rat (Tab. 3.14), the Ka/Ks ratio of imprinted genes (median 0.12) is not significantly higher than genome-wide (median 0.10). By concentrating on genes with human orthologs, we found a lower Ka/Ks median of 0.09 for genome-wide data. Compared to this probably more appropriate set, there is a trend for an increased Ks/Ks ratio in imprinted genes ($p < 0.09$). Contrary to the pattern observed for mouse-human and mouse-cow gene pairs, mouse-rat imprinted orthologs have a decreased rate of synonymous substitutions, Ks ($p < 0.008$), whereas the rate of nonsynonymous changes, Ka, is not significantly elevated ($p > 0.15$). This holds for both paternally and maternally expressed genes and is in agreement with the results of Smith and Hurst (1999). Selection on silent sites is related to alternative splicing and RNA secondary structure requirements (Xing and Lee 2006) and has been reported for the rodent X chromosome (Smith and Hurst 1999). Indeed, the Ks rate is significantly lower for 533 X-linked rodent orthologs as opposed to 16,268 autosomal genes ($p < 10^{-10}$). This finding explains why the DNA identity of X-linked genes is higher despite not significantly elevated protein identity (compare also Tab. 3.14). Rather unexpectedly, mouse-rat imprinted gene pairs behave like X-chromosomal genes with respect to Ks as well as protein and cDNA identity ($p > 0.9$) and Ka/Ks ($p > 0.7$). (Paternally and maternally expressed ones separately: $p > 0.4$).

Previous reports on a limited data set observed this phenomenon as well and it has been attributed to hemizygous expression as a common feature of both types of genes (Smith and Hurst 1999). A similar connection is neither seen for human-mouse orthologs, where imprinted genes are less conserved than X-chromosomal genes and have a higher Ks rate ($p < 0.05$, Tab. 3.12), nor for human-chimpanzee orthologs, which show similar Ks rates in the two sets ($p > 0.1$, Appendix D Tab. D1). Compared to the autosomes, the Ks rate is significantly reduced for X-chromosomal genes in the case of human-mouse orthologs (median 0.617, $p < 10^{-10}$) but not for human-chimpanzee orthologs (median 0.015, $p > 0.6$). The latter may be related to the low number of available genes (388 only). In a study on a different data set of human-chimpanzee sequences, reduced Ks rates on the X chromosome were reported (Lu and Wu 2005). However, consistently with published results (Chimpanzee Sequencing and Analysis Consortium 2005, Lu and Wu 2005, Vicoso and Charlesworth 2006), the data used here show a highly significant elevation of Ka/Ks on the X chromosome compared to the autosomes ($p < 10^{-5}$) and imprinted genes ($p < 0.05$). In contrast, human-mouse orthologs on the X chromosome do not show an elevated Ka/Ks ratio ($p > 0.1$) but higher conservation on the cDNA level ($p < 10^{-5}$, Tab. 3.12). This complex pattern probably results from the X chromosome comprising genes both under positive and purifying selection (Vicoso and Charlesworth 2006). In addition, although the evolutionary distance between mouse and rat is larger than that between humans and chimpanzee, it has been argued that there is stronger purifying selection in extant rodents due to their population sizes being larger than those of humans or chimpanzees (Chimpanzee Sequencing and Analysis Consortium 2005).

The highest mouse-rat Ka/Ks ratio in the imprinted set is 0.372 for *Igf2*. Between mouse and human the maximum is 0.465 for *Cdkn1c*. For the 45 imprinted genes present in all three species, Ka/Ks values are not significantly increased in mouse-human compared to mouse-rat ($p > 0.8$), nor are they in the control groups. Genome-wide, there are no values above 0.780 for mouse-human while mouse-rat reaches up to 1.623 (1.739 including rodent-specific genes) with some genes that are clearly under Darwinian selection in rodents (Gibbs et al. 2004). Human-chimpanzee Ka/Ks entries in HomoloGene can go as high as 16 for *HES2* (encoding a transcription factor) and 14 for *GYPE* (glycophorin E; see also Chimpanzee Sequencing and Analysis Consortium 2005). In summary, imprinted genes are obviously not under ongoing Darwinian selection in extant

mammals. However, increased divergence levels compared to biallelically expressed genes hint at relaxed constraints early in the rodent lineage.

3.4.3 Reconstruction of ancestral evolutionary patterns

The divergence of human-rodent orthologs in the imprinted group as opposed to their conservation between mouse and rat indicates that an evolutionary period of relaxed constraints in the rodent ancestor was followed by a phase dominated by purifying selection (Fig. 3.13).

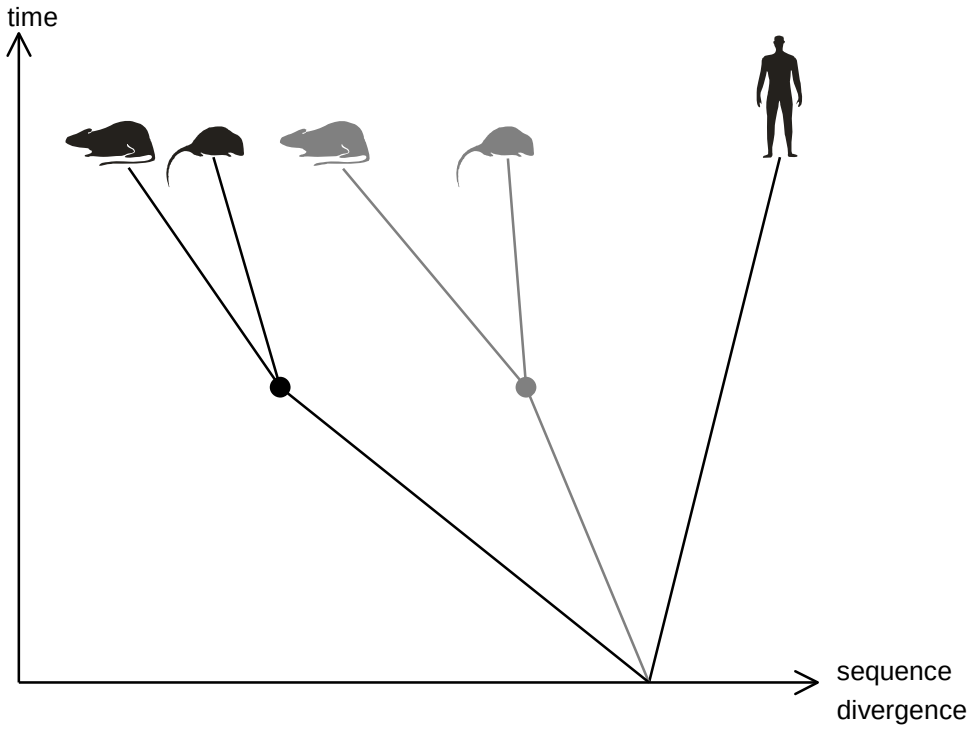


Figure 3.13: Different patterns of divergence
Imprinted genes (black) supposedly evolved faster than biallelically expressed genes (gray) in a common ancestor of extant rodents. They may have gained beneficial functions related to the production of many offspring. After the split of rat and mouse (dots), imprinted genes seem to have been subject to stricter purifying selection than biallelically expressed genes. Both mechanisms counteract and thus obscure the pattern in the sequences of extant species. The present conservation data reflect a high divergence of imprinted human-rodent orthologs as opposed to high conservation between mouse and rat.

In order to get estimates about this branch of the phylogeny, two simple formulas were derived to approximately reconstruct protein and DNA identity, Ka and Ks between human and the rodent ancestor from pairwise alignment data on human-mouse (hs_mm), human-rat (hs_rn) and mouse-rat (mm_rn), using formula (1) for protein or cDNA identity given in percent and formula (2) for Ka or Ks rates:

$$hs\_rodent = \max(hs\_mm, hs\_rn) + \frac{1}{2}(100 - (mm\_rn) - |hs\_mm - hs\_rn|) \quad (1)$$

$$hs\_rodent = \min(hs\_mm, hs\_rn) - \frac{1}{2}(mm\_rn - |hs\_mm - hs\_rn|) \quad (2)$$

With regard to this reconstructed rodent ancestor, sequence identities become higher but more discriminative between imprinted and biallelically expressed genes. When comparing the obtained values of 46 imprinted genes to 14,517 genome-wide reconstructions, results are consistent with the observations described above and the increased divergence of maternally expressed genes becomes slightly more significant. Elevation of Ka in maternally expressed genes ($p < 0.03$) and, tentatively, in the whole imprinted set ($p < 0.06$) as well as a trend towards higher Ks and Ka/Ks rates for maternally expressed genes ($p < 0.09$) implies that extensive mutation processes took place in the coding regions of imprinted genes.

Additionally, I estimated Ka/Ks ratios for 12,143 genes for which human, mouse, rat, and cow sequences were available in HomoloGene. I inferred the cDNAs of the longest open reading frames and aligned them with *transAlign* (Bindinda-Emonds 2005). Branch models were constructed with *codeml* from the *PAML* package (Yang 2007) for each alignment. Since the correct rooted tree ((human, (mouse, rat)#1), cow) caused a warning, I applied the unrooted tree (human,(mouse, rat)#1, cow). The null model assigns the same Ka/Ks ratio to each branch, the alternative model estimates a different ratio for the rodent ancestor branch marked by #1. Ks and Ka rates are calculated separately by *codeml* to fulfill the respective Ka/Ks. Genes with $Ks = 0$ and $Ks > 2.5$, which is a result of saturation, were omitted from further analyses because these data are unreliable.

In the imprinted group, the lineage leading to the rodent ancestor has similar Ka/Ks ratios as other lineages except for eleven out of 34 genes (32%), for which the two-ratios model assuming a different Ka/Ks ratio is significantly more likely than the one-ratio model assuming the same Ka/Ks ratio for all branches ($p < 0.05$). For four genes (*Cdkn1c*, *Igf2r*, *Magel2*, and *Ndn*), the Ka/Ks calculated by the two-ratios model is elevated in comparison to the one-ratio model. However, as it is lower than 1, there is no sign for positive selection at the base of the rodent lineage. On a genome-wide scale, the two-ratios model fits better for 3588 of 12032 genes (30%). For those genes, Ka/Ks is higher in 679 cases and Ks is elevated in 2913 cases. Only for two genes (*FAM100A* and *PTMS*) is Ka/Ks significantly greater than 1 in the early rodent linaege. With both models the Ka/Ks ratios are highly similar for the imprinted, the maternally expressed, and the genome-wide sets ($p > 0.4$). Imprinted genes (likewise the 20 maternally expressed ones among them) tend towards elevated Ka ratios ($p < 0.03$) compared to genome-wide data. Ks rates of the whole imprinted set (median 0.424) and maternally expressed genes (median 0.442) are significantly elevated compared to genome-wide Ks rates (median 0.313) under the one-ratio model ($p < 0.002$) whereas they are only tentatively elevated under the two-ratios model ($p < 0.03$). The 14 paternally expressed ones do not behave significantly different ($p>0.2$).

For 21 imprinted genes present in six species, the correct unrooted tree (((human, chimpanzee), (mouse, rat)#1), cow, dog) (see Fig. 2.11) could be applied. Only in two cases the two ratio model fits better: the Ka/Ks ratio is lowered in the *Slc22a18* early rodent branch whereas it is elevated for *Ndn*. For the other candidate genes mentioned above (*Cdkn1c*, *Igf2r*, and *Magel2*), there are no dog and chimpanzee orthologous sequences available. Although both Ka and Ks are elevated in the rodent ancestor compared to the other lineages, Ka/Ks is always lower than 1. Hence, there is no

indication for positive selection.

3.4.4 Assessing ongoing evolution with single nucleotide polymorphisms

Although most imprinted genes seem to have developed before mammalian radiation, evolution is still ongoing for at least some of them. For example, enrichment of single nucleotide polymorphisms (SNPs) indicating accelerated evolution in the human lineage has been reported for *KLF14* (Parker-Katiraee et al. 2007). We investigated whether there might be indications for more sequence variants associated with imprinted genes by using human-specific single nucleotide exchanges and indels of dbSNP version 129. 14,774 autosomal genes are associated with at least one SNP in their coding region, including 50 imprinted ones, which thus have a similar ratio (χ^2 test, $p > 0.4$). The median number of SNPs per 1 kb of coding sequence is 3.94 for the 50 imprinted genes and 3.19 genome-wide, thus insignificantly elevated in the imprinted set (Wilcoxon test, $p > 0.4$). The annotations of missense and nonsense SNPs outnumber those of synonymous ones. Whereas the density of nonsynonymous SNPs does not distinguish the imprinted group from the human autosomal genes, synonymous SNPs are tentatively enriched ($p < 0.08$). The subsets of maternally and paternally expressed genes differ neither from each other nor from biallelically expressed genes. Since the SNP database comprises data on both individual genes and whole-genome projects, it may be biased in favor of well-studied genes. Curiously enough, for *KLF14* there is only one synonymous coding SNP annotated in the snp129 track whereas Parker-Katiraee and coworkers (2007) found several and also nonsynonymous ones by analyzing genome project data for 826 genes. Nevertheless, the results affirm that recent sequence variations do not seem to be a common feature of human imprinted genes (Parker-Katiraee et al. 2007).

The same analyses based on the mouse genome, for which fewer SNPs are annotated, show different results. 29 of the 53 murine imprinted genes (54.72%) contain SNPs in their coding regions, a similar percentage as the autosomal 57.79% (χ^2 test, $p > 0.8$). Whereas the latter contain a median of 3.57 SNPs per 1 kb of coding sequence, this rate is significantly lower for the imprinted group with only 2.35 SNPs (Wilcoxon test, $p < 0.04$). Synonymous SNPs are tentatively depleted as well ($p < 0.06$), otherwise there are no significant differences. Since a possible bias resulting from intensive investigations of imprinted genes would result in a relative enrichment of SNPs, the observed depletion provides clear evidence that this group actually contains fewer SNPs. Therefore, accelerated recent evolution can be ruled out. Additionally, SNP depletion strongly argues for ongoing purifying selection on murine imprinted genes.

3.4.5 Other factors influencing the low conservation of imprinted genes

As imprinted genes are associated with particular, allele-specific DNA methylation patterns, the observed divergence of their protein-encoding sequences might be due to an increased rate of CpG to TpG transitions. Silent CpG mutability, i.e. CpG deamination that does not change the encoded amino acid, has been proposed as a measure of germline methylation density (Smith and Hurst 1998, Smith and Hurst 1999). To investigate whether it might be increased in the imprinted group, we made use of the four-species alignments mentioned above. After splitting them into the respective pairwise alignments, I calculated the numbers of CpG pairs and CpG-TpG pairs with C at the third codon position (Smith and Hurst 1998, Smith and Hurst 1999). Similarly, exchanging CpG to CpA due to 5-methylcytosine deamination on the reverse complementary strand does not change the amino acid if the G/A transition is at the third position. Irrespective of the method, the imprinted group shows lower levels of CpG-TpG mismatches (Appendix D Tab. D2), which is in

agreement with reports on mouse-rat orthologs (Smith and Hurst 1999). As with conservation data, there is no apparent relation of CpG mutations with localization of genes in a cluster, parental expression, or conserved expression patterns between human and mouse.

Human-mouse alignments of the protein sequences encoded by the 20 maternally expressed genes contain more gaps than those corresponding to the 12,143 autosomal genes (Wilcoxon test, $p < 0.05$). The whole imprinted group also exhibits a trend towards an enrichment of long gaps corresponding to insertions or deletions of ≥ 10 amino acids ($p < 0.1$). The difference is, however, not significant for maternally or paternally expressed genes taken separately. Thus, besides mismatches, also long regions without a counterpart in the orthologous protein contribute to the weak conservation of maternally expressed genes.

3.4.6 Paralogous genes may facilitate divergence

Gene duplication war reported as an important factor in the evolution of imprinting (Walter and Paulsen 2003). Genes that possess paralogs may diverge while acquiring new functions. Thus, we searched to analyze whether there is indeed an enrichment of paralogs in the imprinted gene group. For the autosomes of the human genome, Ensembl release 52 annotates 19,950 human autosomal protein-coding genes. The X and Y chromosomes, which were not included, differ from the autosomes by having a much higher number of paralogs that also achieve higher identities.

As the representative paralog, we chose the one that is listed first by Ensembl, which is the evolutionary youngest and in most cases also the one with the highest identity to the query gene. For some imprinted genes the approach used here yields different results than the literature (Appendix D Tab. D3) and for some genes that have a paralog according to the literature (*NAP1L5*, *Dlk1*, *MAGEL2*, *Ndn*, and *Peg10*), it does not report one (Walter and Paulsen 2003, Paulsen et al. 2005, Wood et al. 2007). Nevertheless, 60.71% of the imprinted genes possess at least one paralog, which is a higher percentage than that of the control groups as well as the genome-wide ratio (48.22%; χ^2 test, $p < 0.1$). Compared to genome-wide data and control groups, there is no significant difference of the paralog numbers per gene (Wilcoxon test, $p > 0.2$). Paralogs located on the same chromosome occur at a similar rate for imprinted genes as for biallelically expressed ones, excluding a bias for segmental duplications (Tab. 3.15). Several imprinted genes were linked to paralogs on the X chromosome (Walter and Paulsen 2003, Morison et al. 2005, Wood et al. 2007) and it has been speculated that imprinting and X chromosome inactivation may have co-evolved (Ferguson-Smith and Reik 2003, Reik and Lewis 2005, Pauler et al. 2007, Wood et al. 2007). Here, we find three imprinted genes with their youngest paralog on the X chromosome (*USP29*, *DCN*, *HTR2A*) and an additional three (*SLC38A4*, *L3MBTL*, *UBE3A*) that have a paralog on the X chromosome which is not the highest scoring one. Compared to non-imprinted genes, the resulting 10.71% is no significant enrichment ($p > 0.1$) since as much as 5.19% of all autosomal genes possess X-chromosomal paralogs (Tab. 3.15).

Using data from the mouse, we found essentially the same patterns although a higher percentage of genes is linked to paralogs. They include the imprinted genes *Ins1* and *Pon2* that are not part of the analyzed set. Compared to genome-wide data, human imprinted-paralog gene pairs tend to be less conserved on the protein level (Wilcoxon test, $p < 0.06$, Tab. 3.15). In the mouse this relaxation in paralog conservation is more pronounced ($p < 0.007$), and is probably caused by the stronger divergence of imprinted genes in the rodent ancestor as described above.

Table 3.15: Pairs of genes and their paralogs

group	genes[a]	with paralog	median number of paralogs	paralog on same chromo-some	youngest paralog on X	has paralog on X	protein identity based on original gene (%)
human imprinted	56	34*	2	6	3	6	47.12±15.24**
human autosomes	19,950	9619	2	2986	288	1035	56.85±22.73
mouse imprinted	54	33*	2*	7	2	4	44.94±18.02***
mouse autosomes	21,871	10919	3	4270	309	1029	60.68±24.07

[a] Deviating numbers compared to the previously mentioned RefSeq genes are due to a different gene annotation procedure at Ensembl.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$ (Wilcoxon test for comparison of the respective group to genome data)

To investigate whether the existence of paralogs might influence the evolution of protein-coding imprinted genes, we linked the paralogs to their entries in HomoloGene. Interestingly, the paralogs show a higher conservation than their imprinted counterparts and biallelically expressed genes and show remarkably lower Ka/Ks ratios between human-mouse, and human-rat (Tab. 3.16). Unexpectedly, mouse-rat pairs of imprinted genes are also less conserved than their paralogs on protein level ($p < 0.005$) although they have similar DNA identities ($p > 0.1$) and Ks rates ($p > 0.6$). At the same time, the Ka/Ks ratio of the paralogs is only half as high (median 0.0535, $p < 0.001$; Tab. 3.16). These relations are confirmed by comparison of the 28 pairs with a χ^2 test. Purifying selection seems to act far more strictly on the paralogs of rodent imprinted genes than on the imprinted genes themselves.

Table 3.16: HomoloGene data for paralogs of imprinted genes

orthologs of	genes	protein ID ± std.dev. (%)	DNA ID ± std.dev. (%)	Ka/Ks ± std.dev.	Ks ± std.dev.
human-mouse	32	90.04±10.42**	86.90±6.01***	0.0940±0.0925*	0.5696±0. 1847
human-rat	28	89.85±11.04**	86.56±6.25**	0.0888±0.0827**	0.6025±0.1934
mouse-rat	28	97.08±2.19***	94.81±1.68***	0.0646±0.0415***	0.2060±0.0651

* $p<0.1$, ** $p<0.05$, *** $p<0.01$ (Wilcoxon test for comparison of the respective group to genome data)

In the entire genome and also in case of imprinted genes, orthologs that possess paralogs are significantly more conserved between human and mouse or between mouse and rat than those without a paralog and they have lower Ka/Ks ratios (Wilcoxon test, $p < 0.001$; Tab. 3.17). A higher divergence of genes that do not possess paralogs has been noted before (Jordan et al. 2004, Brunet

et al. 2006). Comparing imprinted genes with or without paralogs, respectively, to the corresponding groups of autosomal genes, reveals that imprinted genes with paralogs are subject to decreased conservation between human and rodents ($p < 0.04$) and tend towards a higher Ka/Ks ratio ($p < 0.09$) whereas there is no significant difference between genes without paralogs in both groups ($p > 0.3$). Between mouse and rat, genes with paralogs behave similarly in both groups ($p > 0.2$) but imprinted ones without paralogs show increased conservation on DNA level and a lower Ks ratio ($p < 0.002$). In all comparisons maternally and paternally expressed genes behaved similarly.

In summary, there are different patterns of evolution: The orthologs of genes without paralogs are in general more divergent than those of genes with paralogs. Over a large evolutionary distance, as is the case between human and rodents, imprinted genes appear to diverge more than biallelically expressed genes even if they have paralogs. In contrast, between extant rodents, imprinted genes without paralogs are more conserved. This implies that after initial divergence, imprinted genes without a paralog that might take over at least part of its functions were subject to purifying selection. As an additional level of complexity, paternal and maternal expression come into play.

Table 3.17: HomoloGene data for genes with or without paralogs

group	genes[a]	protein ID ± std.dev. (%)	DNA ID ± std.dev. (%)	Ka/Ks ± std.dev.	Ks ± std.dev.
imprinted human-mouse with paralogs	32	85.0±10.8**	83.6±5.9**	0.131±0.106*	0.678±0.189
imprinted human-mouse without paralogs	20	82.7±11.6	83.3±7.3	0.165±0.116	0.623±0.294
genome human-mouse with paralogs	7235/ 7228	88.2±10.5	85.7±5.9	0.105±0.096	0.625±0.229
genome human-mouse without paralogs	7765/ 7756	83.7±11.9	83.4±6.4	0.145±0.112	0.656±0.212
imprinted mouse-rat with paralogs	28	94.8±3.4	93.9±2.0	0.120±0.081	0.211±0.066
imprinted mouse-rat without paralogs	18/17	95.1±3.1	95.2±2.2***	0.166±0.102	0.148±0.072***
genome mouse-rat with paralogs	7638/ 7636	94.3±6.4	93.5±3.4	0.125±0.138	0.224±0.096
genome mouse-rat without paralogs	7509/ 7505	93.0±6.4	92.9±3.4	0.155±0.143	0.229±0.082

[a] The second number refers to sequences available in the HomoloGene database for Ka/Ks analyses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.002$ (Wilcoxon test for comparison of the respective group to the corresponding genome data)

3.4.7 Summary and conclusions of chapter 3.4

Imprinted, especially maternally expressed, mouse and rat genes show reduced conservation with their non-rodent orthologs on cDNA and protein level. Most of the divergence seems to have taken place at the base of rodent evolution as opposed to purifying selection in extant rodents. The divergence at both silent codon position and those resulting in amino acid changes is apparently not caused by increased mutation of CpG positions. It is conceivable that, according to the parental conflict hypothesis, adaptations related to increasing maternal demand occurred in the muride ancestor. Moreover, there is an enrichment of paralogs in the imprinted group and imprinted-paralog gene pairs are less conserved on protein level compared to genome-wide data. More interestingly, the paralogs are subject to strict purifying selection. Thus, the presence of highly conserved paralogs may allow relaxation on the selective constraints acting on imprinted genes. Since increased divergence might cause altered functions and interactions of the proteins encoded by imprinted genes, the findings of this study have potential implications on the suitability of murine models for studies related to human imprinting disorders.

Chapter 4 – Discussion

The aim of this work was to find features of imprinted genes that distinguish them significantly from biallelically expressed genes. Analyses of CpG islands, repetitive elements, conservation on the levels of genomic DNA and protein-coding sequences, substitution patterns, putative effects of CpG deamination, and paralogous genes revealed unexpected similarities and differences. To the best of our knowledge, this is the first comprehensive study of the sequences of imprinted genes on a genome-wide level. The results of this work contribute to a better understanding of the implications of genomic imprinting and will stimulate further research.

4.1 Imprinted genes versus control genes and the genome

4.1.1 Choosing appropriate control groups

A crucial point of any statistical analysis is the choice of an appropriate control group. Since most imprinted genes reside in clusters, the first question is whether to compare complete imprinted regions to chromosomal regions of similar size and G+C content (Greally 2002) or to treat the genes separately, taking into account the complications of overlaps if also adjacent genomic sequences are considered. In previous work, control groups for the latter approach comprised genes near imprinted regions for which monoallelic expression has been excluded (Luedi et al. 2005, 2007), randomly selected BACs (Walter et al. 2006), random genes (Luedi et al. 2005, 2007), or the whole genome. As our studies showed, any choice may be biased. For example, randomly generated RefSeq accession numbers yield genes with higher conservation levels than the genomic average. Especially control group G2, where the numbers were restricted to a range of 1 to 16,000, seems to contain a large fraction of housekeeping genes that are expressed in a wide range of tissues. Such genes are supposed to be under strong selective constraints. In the public sequence databases, genes that have been studied for a long time are represented by various transcriptional variants, whereby the probability of a random number hitting one of them is increased compared to a gene with a single transcriptional variant. The human or mouse genome itself, however, features many genes of unknown function and even predicted ones that are, in contrast to the selected imprinted ones, not necessarily present in other species. In part, transcripts are also redundant, namely in the mouse, where the RefSeq list includes Ensembl genes. Randomly selecting genes of a list comprising all RefSeq genes on human autosomes and taking those with orthologs in mouse resulted in control group G3, which seems to be more representative of the genome by being relatively CpG-poor and having a similar conservation level.

Furthermore, we recommend to perform bioinformatics studies on imprinted genes for both human and mouse in order to distinguish species-specific features from those related to epigenetic effects. For analyses limited to rather small sequence sets, pairwise comparisons emerged as sufficient to highlight the most important features. The conservation between human and cow or dog is in general higher than between human and mouse so that little additional information is gained by including these species. Unfortunately, most genomes are not sequenced, assembled and annotated well enough to be useful for unbiased sequence retrieval of individual genes. On a genome-wide scale, this disadvantage is circumvented in part by multiple alignments and conserved elements derived thereof. However, the large amount of data only averages out problems caused by missing sequences and incorrect alignments.

In summary, it appears useful to have several control groups. If the statistical analyses are consistent for all of them, this provides stronger evidence than the results of comparing the imprinted set to just one control group. On the other hand, contradicting results allow inference on whether imprinted genes share features of, for example, housekeeping genes. When there are genome-wide data available for analyses, randomly chosen control groups (which, in contrast to expectations, are likely not representative of the genome) should rather be replaced by specially selected sets of genes with known properties. Our studies revealed a new group of possible control genes that had not been taken into consideration previously: Paralogs of imprinted genes. They may also be well suited for experimental analyses because they presumably have highly similar functions but are not subject to monoallelic expression. Thereby, epigenetic influences could be separated from functional constraints.

4.1.2 Imprinting candidates

Expecting at least 1-3% of all genes to be imprinted, random selection might by chance have included a few candidates in the control sets, which represent roughly one hundredth of all human or mouse genes, respectively. Indeed, G2 contains one gene, *PRIM2A*, that was reported as maternally expressed in humans (Pant et al. 2006). Additionally, there are some control genes predicted as maternally expressed in either human or mouse, or both (Luedi et al. 2005, 2007). For these and some additional genes that possess pronounced tandem repeats in their CpG islands – a feature we found to be significantly enriched for imprinted genes – information on their conservation and that of their evolutionary youngest paralogs is collected in table 4.1. These features were selected because, according to our studies, imprinted genes are highly diverged between human and mouse but highly conserved between extant rodents. Furthermore, imprinted genes and their paralogs show increased sequence divergence, as opposed to the strong conservation of the paralogs between all species (compare also sections 3.4 and 4.8). Judging from these data, no candidate fulfills all criteria, but neither do all imprinted genes.

It is interesting to note that only one gene in this candidate set, *GDNF*, is predicted as imprinted for both species. A genome-wide discordance has also been stated by Luedi and coworkers and has been attributed to species-specific differences (Luedi et al. 2007). Since for more than half of the predicted genes expression is expected from the maternal allele, it is little surprising that the cases present here do not include a predicted paternally expressed gene. From our studies we conclude that paternally expressed genes, which for various features show a higher variability than maternally expressed ones, are also a more heterogeneous group that cannot be easily separated from non-imprinted genes. Thus, their prediction is likely more challenging.

Table 4.1: Potential imprinting candidates among the control genes

gene (group)	number of human paralogs, youngest one, identity	tandem repeat in +-10 kb	human-mouse protein ID (%), cDNA ID (%)	human-mouse Ka, Ks, Ka/Ks	mouse-rat protein ID (%), cDNA ID (%)	mouse-rat Ka, Ks, Ka/Ks	paralog human-mouse protein ID (%), cDNA ID (%)	paralog human-mouse Ka, Ks, Ka/Ks	evidence
PRIM2A (*PRIM2*) (G2)	–	human only (simple repeat)	NA	NA	96.3 93.6	0.017 0.281 0.060	–	–	Pant et al. 2006
GDNF (G2)	3 NRTN 28%	–	92.9 89.7	0.036 0.429 0.084	99.1 97.0	0.004 0.126 0.032	87.2 83.1	0.073 0.621 0.118	Luedi et al. 2005, 2007
ADARB1 (G2)	3 ADARB2 49%	human only	94.9 87.7	0.026 0.653 0.040	99.0 96.2	0.004 0.165 0.024	83.5 80.8	0.093 0.856 0.109	Luedi et al. 2007
CACNA1B (G1)	–	human and mouse	93.2 87.1	0.036 0.645 0.056	98.1 95.1	0.009 0.206 0.044	–	–	Luedi et al. 2005
CNTNAP1 (G2)	–	human and mouse	93.4 87.9	0.035 0.562 0.062	98.4 95.2	0.008 0.199 0.040	–	–	Luedi et al. 2005
*FASTK** (G3)	–	human and mouse	89.5 86.7	0.055 0.490 0.112	97.2 95.7	0.013 0.145 0.090	–	–	Luedi et al. 2007
OLIG2 (G2)	1 OLIG3 46%	–	96.3 88.1	0.018 0.571 0.032	99.1 95.9	0.004 0.160 0.025	98.5 91.2	0.007 0.454 0.015	Luedi et al. 2007
*PPAP2C** (G2)	7 PPPAP2A 46%	human only	90.5 84.5	0.049 0.745 0.066	96.0 94.1	0.018 0.211 0.085	75.1 76.5	0.156 1.010 0.154	Luedi et al. 2007
ASB13 (G1)	3 ASB5 39%	human and mouse	97.8 87.3	0.010 0.826 0.012	96.8 94.1	0.014 0.235 0.060	93.6 87.4	0.033 0.648 0.051	–
CSNK1D (G1)	6 CSNK1E 82%	human only	99.7 89.8	0.001 0.638 0.002	97.8 97.2	0.015 0.076 0.197	98.8 90.7	0.006 0.518 0.012	–
CUX1 (*CUTL1*) (G2)	–	human only	85.6 81.9	0.097 0.752 0.129	97.3 96.0	0.012 0.143 0.084	–	–	–
DPYSL4 (G1)	5 DPYSL2 75%	human only	93.0 86.3	0.033 0.741 0.045	99.0 95.7	0.005 0.185 0.027	98.8 91.5	0.006 0.459 0.013	–
FBLN1 (G2)	1 FBLN2 45%	human only	85.5 84.7	0.076 0.659 0.115	96.5 94.6	0.016 0.216 0.074	83.1 82.2	0.097 0.729 0.133	–
POFUT2 (G1)	–	human only	92.5 84.1	0.039 1.070 0.036	98.4 95.2	0.007 0.214 0.033	–	–	–
TRRAP (G2)	–	human and mouse	99.1 87.9	0.006 0.855 0.007	99.6 94.5	0.002 0.277 0.007	–	–	–

NA: not applicable

* high confidence candidate (Luedi et al. 2007)

4.2 CpG islands associated with human and mouse imprinted and biallelically expressed genes

4.2.1 Performance of alternative methods for CpG island identification

One of the key questions at the beginning of my studies was which criteria to use for the identification of CpG islands (CGIs) since they constitute important regulatory elements and may be associated with differentially methylated regions. Testing different criteria and programs on the orthologous control sequences lead to the conclusion that murine CGIs are shorter, G+C-poorer, and more affected by CpG deamination than human ones (Hutter et al. 2009), which is again in agreement with published data (Aïssani and Bernardi 1991, Antequera and Bird 1993, Matsuo et al. 1993, Cuadrado et al. 2001, Waterston et al. 2002, Yamashita et al. 2005, Zhao and Zhang 2006a, 2006b, Jiang et al. 2007). Nevertheless, promoter-associated CGIs are reliably detected by all programs in both species. In contrast, at other locations, especially in introns, there is a high variability. Therefore, the identification of potentially regulatory CGIs outside of promoter regions is a challenging task. It is complicated by CpG-rich repetitive elements, namely *Alu* elements, which often coincide with CGIs, making repeat-dependent CGIs much more frequent in human than in mouse. Such CGIs are not expected to be functional for gene regulation since they are usually methylated. Consistent with their rapid cytosine deamination, they show a notably higher (TpG+CpA)/(2·CpG) ratio than unique and promoter CGIs. When discarding repeat-dependent CGIs, orthologous sequences of both species are assigned similar numbers of CGIs. Since using repeat masked sequences as input results in artifacts, CGI identification in original genomic sequences with subsequent removal of critically repeat-overlapping CGIs is highly recommendable.

Although species-specific parameters might be desirable, determining them requires detailed analyses. It is still unknown whether the recently defined CpG clusters (Hackenberg et al. 2006, Glass et al. 2007) manage to identify regulatory CGIs better than traditional sliding window methods as implemented in the *CpG Island Searcher* program (Takai and Jones 2002). Unfortunately, experimental evidence is lacking for CGIs identified with the more recent methods. Han and Zhao (2009) conclude from their studies that *CpG Island Searcher* is more appropriate for identifying promoter-associated CpG islands in vertebrate genomes than *CpGcluster*. We found that the approach of *CpGcluster* (Hackenberg et al. 2006) is especially problematic because its parameters depend on the input sequence. A CpG-rich sequence in which CpGs are more or less well clustered will generate very low 50^{th} and 75^{th} percentiles compared to a CpG-poor sequence. As a consequence, the detection threshold will be more stringent for a CpG-rich sequence than for a CpG-poor one. Moreover, if one identifies CpG clusters in a certain gene and in the same sequence with an additional genomic neighborhood, different clusters can be found (Hutter et al. 2009).

Rather than demonstrating advantages, also the other alternative programs come with their own limitations. The segmentation algorithm of *cpg* (Li et al. 2002) is likely to miss a CGI that can be found with traditional methods if the borders between CpG-rich and CpG-depleted regions are blurred by a gradually increasing/decreasing CpG content rather than a sharp boundary. This scenario is quite likely for CGIs in the process of eroding due to CpG deamination (Matsuo et al. 1993). Recently, such CGI erosion was supported by the findings of Jiang and coworkers (2007) who reported enrichment of TpG and CpA at the edges of CGIs and that human CGIs comprise relatively CpG-poor margins whereas the shorter mouse CGIs display sharper borders. Increasing

the required CpG content to the originally recommended minimum of 6% (Li et al. 2002) instead of the 3.5% applied in our analyses may exclude repeats but also discards many possibly functional CpG-rich segments, especially in murine sequences. Besides, *cpg* lacks a user interface, which makes it hardly attractive for experimentalists. The same holds for the program *CPGed* (Luque-Escamilla et al. 2005) which additionally involves many parameters. When choosing the directly CGI-related ones similar to those of Gardiner-Garden and Frommer (1997), there is an equivalently high number of repeat-dependent CGIs and G+C and CpG content of the identified promoter CGIs are considerably lower than those reported by the other methods.

4.2.2 Recommendable strategies for detection of functional CpG islands

Intriguingly, a higher rate of conserved CGIs in the mouse compared to the human genome implies that non-conserved CGIs have been lost preferentially. Since conservation is indicative of functionality, there seems to be a high selective pressure on maintaining the present CGIs. According to the literature, CGI loss is prominent in rodents (Antequera and Bird 1993, Matsuo et al. 1993) but also present in human on a smaller scale (Jiang et al. 2007). For identifying promoter CGIs in mammalian species for which annotations of genes or repetitive elements are scarce, the rather strict Takai and Jones (2002) parameters are more appropriate than the ones of Gardiner-Garden and Frommer (1987) since they exclude short CGIs that have a low probability of being functional and reduce overlap with CpG-rich repeats. If a more detailed analysis of all epigenetically relevant CGIs is planned, it should be performed on unique Gardiner-Garden and Frommer CGIs in noncoding regions. A further focussing on CGIs that most likely possess regulatory functions can be achieved by using a filter that takes into account differences in the deamination rate. In contrast to repeat-dependent CGIs, the (TpG+CpA)/(2·CpG) ratio is low in promoter CGIs, indicating the expected absence of DNA methylation in the germ line and, consequently, reduced CpG deamination. Moreover, human CGIs with a low (TpG+CpA)/(2·CpG) ratio have a high probability of being transcriptionally active (Bock et al. 2007). This correlation is not surprising because the calculation of the so-called epigenetic score includes TpG and CpA patterns (Bock et al. 2006, 2007). Calculating the (TpG+CpA)/(2·CpG) ratio is a simple alternative, especially in case genome-wide epigenetic data are not available for epigenetic score prediction. A (TpG+CpA)/(2·CpG) ratio pattern comparable to human is also seen in the mouse, suggesting that similar cytosine deamination effects take place in methylated CGIs in all mammalian species. For genome-wide use, we found the CGI annotations provided by UCSC to be well suited. Besides their easy availability, the results obtained based on them agree with previous findings. Moreover, the program circumvents the repeat overlap problem and reports very CpG-rich islands, thus considerably decreasing the portion of CGIs that probably do not fulfill regulatory functions.

4.2.3 Special features of CpG islands in imprinted regions

CGIs were suspected to be enriched in imprinted regions (Paulsen et al. 2000, Paulsen and Ferguson-Smith 2001, Reik and Walter 2001). However, their numbers and extends in orthologous imprinted regions of various mammals are quite divergent (Paulsen et al. 2001, Paulsen et al. 2005). Do these differences result from imprinting or are they species-specific? The analyses on human and mouse genomic sequences presented in this study revealed that neither the number nor the length, G+C or CpG content of CpG islands differ between imprinted and randomly chosen

biallelically expressed genes (Hutter et al. 2006). These findings support previous reports (Ke et al. 2002a, 2002b, Allen et al. 2003).

Maintaining CGIs might be especially important in imprinted regions. Compared to biallelically expressed genes, a significantly larger number of imprinted genes possesses intronic CpG islands, which probably act as promoters for antisense transcripts (Reik and Walter 2001) or contain alternative transcriptional start sites. Indeed, the strong intragenic CGIs of *Igf2* (Sasaki et al. 1996, Moore et al. 1997) and the *Nesp-Gnas* locus (Coombes et al. 2003) are associated with start sites of alternative sense transcripts. Even more such CGIs coincide with the promoter regions of previously identified antisense transcripts (*Commd1*, Wang Y et al. 2004; *Igf2*, Sasaki et al. 1996, Moore et al. 1997; *Igf2r*, Wutz and Barlow 1998; *Kcnq1*, Mancini-DiNardo et al. 2003; *Nesp-Gnas* locus, Coombes et al. 2003). Hence, although antisense transcripts have recently been identified as a widespread feature of gene regulation (Kiyosawa et al. 2003, Yelin et al. 2003, Lavorgna et al. 2004, Chen et al. 2005, Zhang et al. 2006), they are obviously concentrated as key regulatory elements in imprinting.

Another difference is that, due to the decreased content of SINEs, there are fewer repeat-dependent CGIs in imprinted regions. However, epigenetic scores and estimated deamination ratios of unique Gardiner-Garden and Frommer CGIs in imprinted regions are essentially the same as for biallelically expressed genes. On the other hand, CpG island associated conserved elements in imprinted regions do not exhibit a lowered (TpG+CpA)/(2·CpG) ratio compared to genome-wide levels, indicating that CpG deamination rates are not specially reduced in them. Also their conservation corresponds to genome-wide conditions. It must be kept in mind that, although DMRs often share little or no sequence similarity between species and thus do not contribute to conserved regions, only some CGIs in imprinted regions are affected by epigenetic modifications in the germ line so that most of them should be subject to the same mechanisms as "conventional" CGIs. Consequently, species-specific differences dominate over putative effects of imprinting

4.3 Influence of CpG deamination in imprinted regions

Since methylation spreads over several thousands of base pairs from germline methylated regions (DMRs), imprinted regions may be prone to increased CpG deamination. However, since even in germ cells CpGs in the bulk genome are methylated, allele-specific hypomethylation could show a contrary effect. Indeed, the (TpG+CpA)/(2·CpG) ratio is reduced in conserved elements of imprinted regions. DNA methylation also appears to be the primary cause for elevated mutation rates at CpG positions. Hence, monoallelic DNA methylation patterns might influence mutation rates. We did not observe a relation between reduced sequence conservation and estimated deamination rates in the coding regions of imprinted genes. In contrast, CpG deamination related mismatches have even less influence on the protein-encoding regions of imprinted genes than on those of biallelically expressed genes. Consequently, imprinted regions are less affected by loss of CpG due to methylation and subsequent deamination in the germ line than might naively be expected.

One possible explanation is that methylation spreads somatically whereas in the germ line, it is restricted to DMRs that reside mainly outside of coding regions. Thus, somatic mutations related to CpG deamination in imprinted regions would not be inherited. Imprinted genes rather seem to be hypomethylated relatively to the rest of the genome. Such a scenario has been observed for the inactive X chromosome that is hypermethylated only at CpG islands but hypomethylated in regions

outside of these regulatory elements (Hellman and Chess 2007). In general, whereas the rather small promoter regions of active genes are unmethylated, there is a high level of methylation in their intragenic regions. This paradox (Jones 1999) has been noted long before methylome studies in various higher eucaryotes reported gene body methylation as very frequent event on the human X chromosome (Hellman and Chess 2007) and in *Arabidopsis thaliana* (Zilberman et al. 2007). The open chromatin structure provided by the transcription machinery also allows methyltransferases to access the DNA. Being more than a by-product, gene body methylation seems to inhibit the generation of intragenic transcripts that could interfere with the regulation of the main gene. This mechanism was first investigated for imprinted genes but is, as it since turned out, far from being limited to them. Still, imprinted genes have the special feature that transcription in the oocyte leads to the establishment of DMRs in the maternal germ line (Chotalia et al. 2009).

Alternatively, a potential bias of the T-G mismatch repair mechanism would both maintain most of the existing CpG dinucleotides and create new ones. This may explain the lack of conservation of DMRs and CGI-associated tandem repeats as well as the G+C- and CpG-richness of conserved elements associated with imprinted genes, which are at the same time less conserved and shorter than the genome-wide average. Interestingly, a DNA repair-based model for active cytosine demethylation has been proposed (Hajkova et al. 2008) and is subject of current research. Since DMRs escape the genome-wide demethylation in the early embryo (Hajkova et al. 2002), different factors involved in the putative shielding of DMRs or their establishment may subject them to a special mode of evolution.

4.4 Possible epigenetic functions of tandem repeats

Depending on the criteria used for CpG island identification, 24-51% of the imprinted genes are associated with at least one tandem repeat in one of their CGIs (see Fig. 3.5). Although this percentage is significantly higher than that of random control genes, it would be clearly incorrect to infer that possessing a tandem repeat is a general feature of any imprinted gene. Some tandem repeats known from the literature could not be identified in this study because of several reasons: They are either located outside of the analyzed 10 kb genomic environment (*Gtl2*, Paulsen et al. 2001), are not associated with a CGI (*Igf2*, Sasaki et al. 1996, Moore et al. 1997; *Magel2*, Boccaccio et al. 1999; *Mest*, Lefebvre et al. 1997), do not constitute tandem repeats in the strict sense (*H19*, Bell and Felsenfeld 2000, Hikichi et al. 2003; *PEG3*, Kim et al. 2003; *Kcnq1*, Paulsen et al. 2005), or are too short and divergent to achieve the lower score limit (*Nesp-Gnas* locus, Coombes et al. 2003). Taking into account such elements would require to treat the control genes in the same way to avoid biases but most probably would not change the relative proportions. Still, not even every imprinting domain was found to harbor a tandem repeat that is associated with a CGI.

There has been much speculation about which functions tandem repeats might convey for imprinting. Arnaud et al. (2003) proposed that CpG-rich repeats might be expansion events that counteract the loss of CpGs by deamination. These expansions might be species-specific, consistent with the finding that virtually no conserved elements in CGIs of imprinted regions overlap with simple tandem repeats. Repeat sequences can even be polymorphic between different mouse strains as in *Impact* (Okamura et al. 2000). Nevertheless, some conserved repeats that are not tandemly arranged in the strict sense have been shown to constitute binding sites for the transcription factors YY1 (in the DMRs of *Peg3* and at the *Gnas* locus, Kim et al. 2003, Kim et al. 2006, Kim J et al.

2007, Kim 2008) and CTCF (*H19*/*Igf2* locus, Bell and Felsenfeld 2000, Hikichi et al. 2003, Szabó et al. 2004). These examples may be exceptions since most tandem repeats do not share apparent sequence motifs. Instead, they are thought to assume a special DNA structure (Neumann et al. 1995, Constância et al. 1998) which might provide a signal for special DNA-binding proteins without requiring sequence conservation. Thus, tandem repeats may play an important part in the interplay of different mechanisms that lead to the establishment of DMRs (Chotalia et al. 2009).

Interestingly, the *IGF2R* gene, whose imprinting got lost in human (Killian et al. 2001, Weidman et al. 2004), is associated with a tandem repeat in its nonconserved, intronic CGI, which is still a DMR (Smrzka et al. 1995, Riesewijk et al. 1996). But, in contrast to the mouse *Igf2r* locus (Wutz and Barlow 1998), it does not act as a promoter of an antisense transcript (Vu et al. 2004). For the orthologs of *Impact*, it has been hypothesized that existence of a tandem repeat determines over its imprinting (Okamura et al. 2000). This was later revised as rabbit *Impact* does not possess a repeat but is imprinted (Okamura et al. 2005). Since some imprinting domains possess several tandem repeats, these may be redundant to some extent, explaining that deletion of the single tandem repeat in the *Rasgrf1* regions abolished imprinting (Pearsall et al. 1999) whereas deleting one out of several ones in the *H19*/*Igf2* regions did not show any consequences (Reed et al. 2001).

Tandem repeats originate by unequal crossing over or by retrotransposition. Thus, it is tempting to establish a connection between them and the increased recombination rate at imprinted loci (Reik and Walter 2001, Sandovici et al. 2006, Luedi et al. 2007) as well as the integration of novel genes into imprinting clusters (Walter and Paulsen 2003). The latter association is supported by the existence of tandem repeats and a L2 repeat at the integration site of *Nnat* in the intron of *Blcap* (Evans et al. 2005). In support of the former connection, Sigurdsson and coworkers (2009) recently found recombination and DNA methylation to be highly correlated in the male germ line and conclude that either DNA methylation could attract recombination events or methylation could mark a region after recombination. Lastly, proteins binding to tandem repeat structures might induce chromatin loops as it was shown for the *Igf2/H19* region (Murrell et al. 2004). They may also be responsible for general epigenetic regulation since some genes from the control groups possess tandem repeats as well and it is quite improbable that these should all be imprinted.

4.5 Connections between imprinted genes and the X chromosome

Genome-wide studies have shown that sex chromosomes are different from autosomes in terms of chromosomal organization and evolution (Waterston et al. 2002, Gibbs et al. 2004, Chimpanzee Sequencing and Analysis Consortium 2005, Vicoso and Charlesworth 2006). Thus, they were largely excluded from the analyses. Imprinted genes and the X chromosome, however, have some striking similarities. First, in female mammals the second copy of the X chromosome is largely inactivated via the extremely long noncoding RNA *Xist*. *Xist* induces repressive histone modifications and CpG island methylation. Thus, the X inactivation center is similar to an imprinting center. Second, determining which copy to silence requires pairing of the two X chromosomes and was shown to be mediated by CTCF (Xu et al. 2007), a transcription factor that also has a prominent role in imprinting. Third, in marsupials and in the extraembryonic parts of the placenta of rodents and cattle, it is always the paternal X chromosomes that is inactivated. Forth, both X chromosome and imprinted regions show particular distributions of repetitive elements. Numerous studies report a depletion of SINEs and enrichment of LINEs for imprinted genes in human and mouse (Greally 2002, Ke et al. 2002a, 2002b, Allen et al. 2003, Walter et al. 2006,

Hutter et al. 2006). Taking into account that most SINEs are either primate or rodent specific, it is not surprising that the overlap of conserved elements with SINEs in imprinted regions corresponds to that of the whole human and mouse genomes. The evident purifying selection against such evolutionary young elements indicates that SINE methylation seems to interfere severely with the establishment of DMRs. In contrast, conserved elements in imprinted regions show substantial enrichment for overlaps with LINEs. These ancient repetitive elements were likely integrated early in the evolution of imprinted regions (Warren et al. 2008, Pask et al. 2009) and might have gained regulatory functions, possibly for spreading methylation from DMRs as on the X chromosome, where there is a similar enrichment of LINEs and depletion of *Alu* elements (Waterston et al. 2002, Lyon 2006). Repetitive elements can also gain regulatory roles in the human genome (Jordan et al. 2003, Oei et al. 2004). So-called exapted repeats – regulatory mobile elements that are subject to purifying selection – are often found near developmental genes (Lowe et al. 2007). Since they include all classes of repetitive elements, the observed enrichment of L1 repeats in imprinted regions and on the X chromosome implies a special role in epigenetics.

Due to the similarity of X inactivation and imprinting it has been suggested that the two mechanisms may have co-evolved or share a common ancestry and imprinting may be related to dosage compensation (Ferguson-Smith and Reik 2003, Reik and Lewis 2005, Pauler et al. 2007, Wood et al. 2007). Since according to the Ensembl annotations only six imprinted genes possess X-linked paralogs (*DCN*, *HTR2A*, *L3MBTL*, *SLC38A4*, *UBE3A*, and *USP29*), which is no significant enrichment in comparison to the whole human and mouse genomes, it does not seem that they constitute a major factor in the evolution of imprinting. Nevertheless, a few duplicates may have taken regulatory elements from the X chromosome with them and thus initiated the development of an imprinting domain. Interestingly, the six genes with X-chromosomal paralogs are distributed over different imprinting domains. Similar observations have been reported in the literature for a different set of genes. Wood et al. (2007) examined twelve murine imprinted genes, all of which are paternally expressed, that show characteristic features of retrotransposition, e.g. lack of introns. Four of them are paralogous to genes on the X chromosome (*Mcts2*, *Inpp5f*, *Nap1l5*, and *U2af1-rs1*; Walter and Paulsen 2003, Morison et al. 2005, Wood et al. 2007). They form so-called microimprinted domains with an oocyte-derived DMR each. As these genes are absent in marsupials, *Mcts2* even in cow and dog, and *U2af1-rs1* is rodent-specific, they might present initial states of imprinting domains. The newly integrated DMR might expand its regulatory influence to neighboring genes during the course of time. In an alternative scenario, whole genomic regions including their regulatory elements could have been translocated from an ancestral X chromosome by genome rearrangement (Rapkins et al. 2006).

Additionally, it has been argued that evolutionary processes should act likewise on imprinted genes and genes on the X chromosome because they have hemizygous expression as a common feature (Smith and Hurst 1999). Analysis of mouse and rat orthologous genes showed that the rate of silent substitutions, Ks, is reduced in imprinted genes as in X chromosomal genes, which has been reported before (Smith and Hurst 1999). Also with respect to protein and cDNA identity as well as the rate of nonsynonymous substitutions, Ka, and Ka/Ks, imprinted genes behave highly similar to those on the X chromosome in rodents. In contrast, the evolutionary patterns for human-mouse and human-chimpanzee genes differ between both types of genes. Consequently, as seen with respect to CpG islands, there seems to be a complex interplay of species-specific, general epigenetic, and imprinting-specific effects. The reduced Ks rate might again be connected to a possible relative hypomethylation. If this was the case, silent CpG mutability – as measured by the

ratio of CpG deamination related exchanges at silent codon positions in protein-coding regions – should be low. However, this ratio has been reported to be high for the X chromosome (Smith and Hurst 1999). In contrast, the present study showed that imprinted genes have a lower estimated CpG deamination ratio than the genome-wide one. Thus, although the X chromosome and imprinted genes possess quite a number of similarities, they differ in detail. This might be related to the fact that there are always two copies of imprinted genes present in germ cells before meiosis, so that they are equally transmitted by males and females, whereas only one copy of the X chromosome is transmitted through the male germ line.

4.6 Is there an "imprinting transcription factor"?

Their parent-of-origin dependent monoallelic expression suggests that imprinted genes possess similar regulatory elements. The simplest scenario would be that all imprinted genes are transcribed upon the presence of certain transcription factors for which binding sites should be detectable as conserved motifs in their promoter regions. Such specific sequence patterns and transcription factor binding sites (TFBSs) should be enriched in the imprinted group compared to biallelically expressed genes. However, no discerning difference could be seen in the analyses presented here. Despite sharing the feature of parent-of-origin dependent monoallelic expression and often occurring in genomic clusters, imprinted genes do not seem to possess similar regulatory motifs and, as to be expected judging from their various functions, they constitute a very inhomogeneous group.

For both imprinted and biallelically expressed genes, the most upstream TSS region rarely overlaps with conserved elements, but frequently with CpG islands. Consistent with these observations, it has been reported that especially genes with tissue-specific expression and roles in development have several transcriptional start sites in their CpG-rich promoter regions (Carninci et al. 2006, Baek et al. 2007). Such genes are likely regulated by specific combinations of transcription factors, so-called cis-regulatory modules. By using data on transcription factor binding sites annotated based on experimental evidence or conservation, only a low number of hits were obtained for the most upstream promoter regions, making it impossible to derive cis-regulatory modules. To see whether genes in the imprinted group share common motifs will require other approaches.

An alternative, promising strategy is the analysis of potential regulatory elements ouside of promoter regions, namely in conserved intronic and intergenic regions. Here, the significant enrichment of CpG-rich motifs may indicate the presence of protein binding sites. One candidate for such a putative imprinting-specific transcription factor is CTCF because binding sites have been reported for several DMRs (*Dlk1*/*Gtl2*, *Grb10* and *GRB10*, *H19*/*Igf2*, *GNAS*, *Kcnq1*, *PEG10*/*SGCE*, *Rasgrf1*, *WT1*; Bell and Felsenfeld 2000, Hark et al. 2000, Paulsen et al. 2001, Hikichi et al. 2003, Szabó et al. 2004, Yoon et al. 2005, Fitzpatrick et al. 2007, Hancock et al. 2007, Monk et al. 2008). Additional occurrences have been reported for *MAGEL2*, *CDKN1C* and *GNAS* (Kang et al. 2009). These sites might also be bound by the CTCF-like protein (Loukinov et al. 2002). The DMRs of *Peg3*, at the *Gnas* locus and in the PWS/AS region are associated with binding sites of a second candidate, YY1 (Kim et al. 2003, Rodriguez-Jato et al. 2005, Kim et al. 2006, Kim J et al. 2007, Kim 2008).

Genome-wide experimental analyses identify CTCF and YY1 as ubiquitous transcription factors (Barski et al. 2007, Kim TH et al. 2007). Nevertheless, CTCF binding sites appear to be

overrepresented in imprinted regions (Lindroth et al. 2008). We found that only 66% of the analyzed imprinted genes are associated with experimentally verified CTCF binding sites. Thus, just like tandem repeats, this feature is overrepresented in comparison to biallelically expressed genes but neither unique to imprinted genes nor strictly necessary. Similarly, less than half of the imprinted genes investigated by Wen et al. (2008) are conneted to triple hits of DNA methylation, H3K4me2, and the presence of CTCF binding sites. These results support the hypothesis that there may be several ways to establish imprints and an interplay of different factors may be required. For example, the finding that CTCF and YY1 form protein complexes with a critical role in X inactivation suggests that such interactions might also occur at imprinted loci (Donohoe et al. 2007), possibly in association with chromatin loops (Murrell et al. 2004, Kurukuti et al. 2006). Indeed, only 21% of the autosomal genes possess both experimentally verified CTCF and conserved predicted YY1 binding sites in introns or in intergenic regions but 40% of the imprinted genes: *ATP10A*, *BEGAIN*, *CALCR*, *CDKN1C*, *DCN*, *DIO3*, *DLK1*, *GNAS*, *HTR2A*, *IGF2*, *KCNK9*, *KCNQ1*, *L3MBTL*, *MAGEL2*, *NDN*, *OSBPL5*, *RASGRF1*, *SLC22A2*, *SLC38A4*, *TRPM5*, *USP29*, *WT1*, and *ZIM2*. Experimental analyses of their *in vitro* YY1 binding capacity and potential interactions would be needed to gain more insight into the role of these transcription factors in imprinting.

4.7 Distinguishing patterns of conservation and divergence

4.7.1 Possible contributions to murine speciation

The observed low recovery rate of orthologous imprinted genes in the genomes of other mammalian species than human and mouse initiated a closer investigation of their general conservation. With respect to genomic sequences, the conservation is more variable but essentially similar to that of randomly chosen genes. Imprinted regions also contain similar amounts of conserved elements as the whole genome, with the notable exception of protein-coding regions. Conserved elements that overlap by at least one base pair with coding exons of imprinted genes have significantly lower conservation scores and are shorter, which indicates a different pattern of evolution. The cDNA and protein based investigations comprise less than 60 imprinted genes for statistical comparison. For species other than human and mouse, there is an even smaller number of sequences available. Thus, analyses on protein and cDNA level could only reveal the most prominent differences, namely the decreased conservation between human and mouse or rat, respectively, and mouse and cow, as opposed to the high conservation between the two rodents. In summary, our analyses support a special role in rodent evolution. We cannot exclude that distinguishing processes acted on some imprinted genes also in other mammalian lineages, thereby contributing to the low conservation scores for mammalian conserved elements. Notably, compared to their chicken and zebrafish orthologs, imprinted genes show similar sequence identities and Ka/Ks distributions as biallelically expressed genes. This suggests that at the split of mammals from other vertebrates, there has been no specific pattern of evolution on protein-encoding genes that later became subject to imprinting. It is highly probable that instead, distinguishing regulation and expression patterns were established.

Early studies on the evolution of imprinted genes in mouse and rat did not reveal indications for conspicuous mutation rates or positive selection in the rodent lineage (McVean and Hurst 1997, Smith and Hurst 1999). From this it was concluded that imprinted genes in general did not show

special patterns of evolution. However, with the sequences of only 15 imprinted genes available at that time, the authors might have looked at the wrong place. Purifying selection in extant rodents is stronger than for example in humans or chimpanzees, presumably due to their larger population sizes (Chimpanzee Sequencing and Analysis Consortium 2005). Thus, although rodents are fast-evolving and thus prone to show increased divergence (Waterston et al. 2002), genes that acquired beneficial mutations in a common ancestor should be highly conserved between mouse and rat. In the millions of years that have passed since the split of the rodent lineage from other mammals, signs of possible initial Darwinian selection would have become obscured by the counteracting marks of purifying selection (see also Fig. 3.13). The results presented in this study strongly support increased protein evolutionary rates of imprinted genes in a rodent ancestor followed by purifying selection in modern rodents. The latter is also supported by a low number of SNPs in murine imprinted genes. The observed reduced Ks rates in rodent orthologous imprinted genes may either hint at different mutations rates or at purifying selection not only on protein function, but also on post-transcriptional regulation (Xing and Lee 2006, Resch et al. 2007). In other mammalian lineages there is no apparent difference compared to biallelically expressed genes. Nevertheless, conserved elements that are fully located in coding exons of imprinted genes do not exhibit the same weak conservation as those for which an overlap of only 1 bp is required. Thus, intronic sequences near exon boundaries, for example splice signals, also seem to be less conserved in imprinted genes than on a genome-wide level.

Initial divergence with subsequent purifying selection is assumed to be typical for the evolution of new functions, for example of duplicated genes (Jordan et al. 2004, Brunet et al. 2006, Conant and Wolfe 2008), and for the evolution of new species. A role for imprinted genes in speciation has been suggested by Reik and Walter (2001) who rather refer to regulation than to protein-coding genes. Indeed, mutations affecting expression are much more common than those affecting protein function (Wilkins and Haig 2001). This might explain that, despite their important roles in embryonic development, the correlation of expression profiles between human and mouse imprinted genes is not higher than that of random orthologs (Steinhoff et al. 2009). In the mouse, more genes are imprinted and their regulation is stricter than in humans (Morison et al. 2005, Monk et al. 2006). Other typical features, such as an enrichment of intronic CpG islands, are also more pronounced in murine than in human imprinted genes (Hutter et al. 2006). It is conceivable that, consistent with the paternal conflict theory (Moore and Haig 1991), short gestation time, frequent pregnancies and offspring of different paternity not only favored strict imprinting in rodents but also changes in the associated protein-coding sequences.

4.7.2 Reconstruction of ancient evolutionary patterns

The results of the branch models (see section 3.4.3) contradict the assumption of accelerated evolution in early rodents. Although at least some genes are expected to have been under positive selection in the ancestral rodent lineage (Gibbs et al. 2004), our attempts with *codeml* failed to detect more than two genes where Ka/Ks would have been significantly greater than 1. A possible explanation for this is that using the simplified unrooted tree (human,(mouse, rat)#1, cow) is inappropriate because it compares the cow directly to the euarchontoglires ancestor instead of a mammalian ancestor (Fig. 4.1). As a consequence, Ka/Ks is likely artificially elevated in this lineage. Since it corresponds to the background ratio, the overall Ka/Ks ratio becomes similar to or even higher than the early rodent ratio, which is allowed to differ in the two-ratios model.

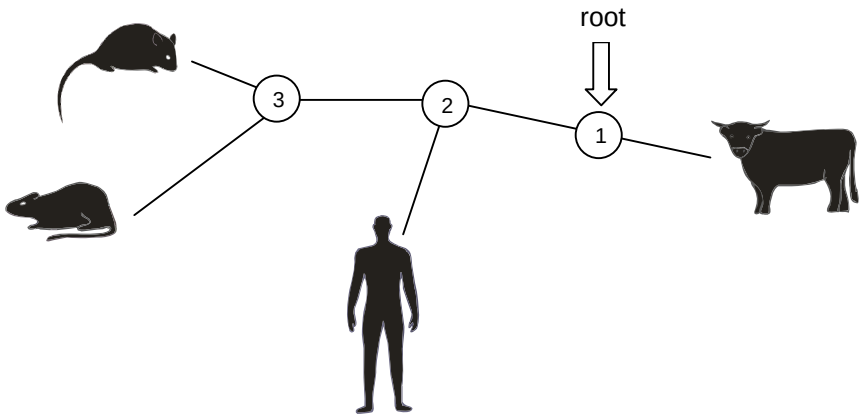


Figure 4.1: Phylogenetic tree of rat, mouse, human, and cow
The correct unrooted tree can be rooted at the mammalian ancestor node (1). From this common ancestor, the cow and the euarchontoglires (humans and rodents, with a common ancestor of mouse and rat represented by node 3) evolved. An unrooted tree in which no such mammalian ancestor is constructed compares the cow as an outgroup directly to the euarchontoglires ancestor node (2), from which the evolutionary distance is substantially higher than if the mammalian ancestor was taken into account.

The most severe drawback is surely the low number of four sequences per gene because maximum likelihood methods profit from the integration of more information. With the HomoloGene data, however, it was not possible to obtain the recommended ten or more sequences per gene. Only for 21 imprinted genes the cDNAs of all six mammalian species were available. When analyzing them with the correct unrooted tree (Fig. 2.7), just two genes had a Ka/Ks ratio that was significantly different in the early rodent branch. Only *Ndn*, a gene of the Prader-Willi/Angelman syndrome region that encodes a neural growth suppressor, has an elevated rodent Ka/Ks ratio. It seems that branch models are well suited for the analysis of genes that exhibit positive selection in one terminal branch but may come to their limit when analyzing genes with opposite patterns of ancient divergence and recent conservation. Overall, the branch models imply an excess of both silent and nonsilent substitutions in imprinted genes throughout the lineages whereas Ka/Ks is similar to the genome-wide level. This at least still argues for extensive mutational processes on imprinted genes.

From the pairwise comparisons it became clear that in the branches leading from the rodent ancestor to mouse and rat, respectively, the Ka/Ks ratio should be low due to purifying selection. On the other hand, for some genes there are indications of adaptations in other lineages, e.g. *PLAGL1*, where Ka/Ks is elevated in the branch leading to human. These issues would call for applying models that assign individual ratios to each branch. Their disadvantages are a huge parameter space (Yang 2007), which increases the probability to get stuck in a local minimum, and that they require an even larger computational effort than constructing the simpler models. Analyzing 10,000 alignments with the *codeml* program already takes several hours and downloading and aligning the sequences more than doubles the time needed. In contrast, the Perl script for calculating divergence patterns of a putative rodent ancestor and human from pairwise HomoloGene data (see section 3.4.3) runs in less than one minute and gives additional data for protein and cDNA identity. The reconstructions are consistent with both the pairwise data and the branch models in that Ka and Ks rates tend to be elevated in maternally expressed genes. Although

we found a trend for increased Ka/Ks compared to biallelically expressed genes, relaxed constraints or even Darwinian selection must be ruled out.

4.7.3 Maternally expressed genes and female-specific benefits

A possible explanation why especially maternally expressed genes show increased divergence might be gain of function. As two prominent examples, *IGF2R* developed from the mannose-6-phosphate receptor gene by gaining an IGF2 binding site (Killian et al. 2001), and the Arabidopsis *MEDEA* gene assumed a crucial role in seed development (Spillane et al. 2007). In a "feminist" view of imprinting evolution, changes in the proteins encoded by maternally expressed genes may have provided female-specific benefits. Supporting this hypothesis, the highest Ka/Ks ratios between human-mouse imprinted genes (Appendix D, Tab. D3) are exhibited by *Cdkn1c* and *Phlda2*, maternally expressed genes that fulfill important functions in the mouse placenta. Also *Tspan32* and *Ascl2*, showing placenta-specific maternal expression in the mouse, have Ka/Ks ratios over 0.3. Apart from being able to control embryonic growth, maternally expressed genes are transcribed in the oocyte from alternative promoters (Chotalia et al. 2009). Here, they may have yet unknown special functions. A dominant maternal role in imprinting is further insinuated by the facts that maternally expressed noncoding RNAs influence the expression of paternally expressed genes (Lin et al. 2003) and that most DMRs are maternally methylated. Thus, paternal repression rather than maternal expression might be the driving force of imprinting evolution.

Also paternally expressed genes have placental functions and influence maternally expressed genes (Varrault et al. 2006). They comprise both the most conserved and the most diverged genes (Appendix D Tab. D3). Whereas there is no nonsynonymous substitution in *Snprn*, *Usp29* is by far the least conserved gene between human and mouse and has the third highest Ka/Ks ratio. The high divergence might be related to the different genomic organization of the *Peg3* domain (Kim J et al. 2007). Among rodents, paternally expressed genes are particularly well conserved. Loss of function is probably devastating if it affects (paternally expressed) growth factors that are involved in crucial cellular pathways. In contrast, (maternally expressed) growth inhibitor genes might degenerate with less deleterious effects. Under special circumstances, their putative loss of function might even be associated with increased fitness. Divergent growth factors could also have acquired new function. In viviparous fish species that developed placenta-like structures, *IGF2* shows indications of Darwinian selection (O'Neill et al. 2007), indicating that evolutionary adaptations of growth factors might predate or be alternative to imprinting.

4.7.4 A critical look on sequence-based methods to keep track of protein evolution

Use of the nonsynonymous to synonymous substitution ratio as a means for detecting Darwinian selection is controversial (Hughes 2007). Since Ka/Ks decreases with evolutionary distance, it is virtually impossible to discover any patterns of ancient selection (O'Neill et al. 2007). Positive selection on individual lineages is especially hard to detect (Kosiol et al. 2008) and efforts to identify such genes are traditionally concentrated on evolutionary recent events, especially when comparing human sequences to chimpanzee orthologs (Bustamante et al. 2005, Nielsen et al. 2007). Low quality of genomic sequences can lead to false positive mismatches, as has been reported for chimpanzee (Mallick et al. 2009). This might explain that mouse-chimpanzee HomoloGene data are not fully consistent with those on mouse-human and that our results for

human-chimpanzee Ks rates differ from published observations (Lu and Wu 2005).

Since the Ka/Ks ratio is calculated as an average over a whole protein-coding sequence, a few but decisive nonsynonymous substitutions can easily be outweighed by an enrichment of (supposedly neutral) synonymous ones. One exchanged amino acid, however, can already be sufficient for altering the phenotype, as the well-known example of sickle cell anemia shows. On the other hand, if functional sites exhibited frequent changes, it would become impossible to find orthologs (Hughes 2007). Moreover, the common assumption that synonymous substitutions are neutral is questioned by findings that the third codon position is important for alternative splicing and RNA secondary structure (Xing and Lee 2006). Consistently, both purifying and positive selection on silent sites is a widespread phenomenon independent of protein evolution (Resch et al. 2007).

Another disadvantage of codon-based models is that they cannot account for drastic mutational events like exonization of intronic sequences (O'Neill et al. 2007) because alignment gaps are discarded or treated as missing data. Our analyses revealed that gaps are enriched in human-mouse alignments of the protein sequences encoded by maternally expressed genes and thus contribute substantially to the observed divergence. In murine amino acid sequences that align with <50% identity to their human counterparts (e.g. *Plagl1*, *Tspan32*, and *Usp29*), unmatched sequence stretches containing large amounts of proline, glutamine, methionine, and histidine can be observed which might indicate intron retention.

Without experimental evidence, it remains unclear whether changes in the amino acid sequence induce altered protein functions and interactions. As the example of *IGF2R* shows, divergence did not take place at the IGF2 binding site but at the signal sequence (McVean and Hurst 1997, Smith and Hurst 1998), indicating that protein-protein interactions must be preserved whereas the concentration of the functional protein may be changed by transporting it with altered efficiency.

In conclusion, Darwinian selection can probably only be detected for recently established imprinted genes such as *MEDEA* (Spillane et al. 2007). Notably, when contrasting the divergence of human and mouse imprinted genes with the existence of orthologs in non-mammalian species (Paulsen et al. 2005, Dünzinger et al. 2007, Pask et al. 2009), there is a suspicious concentration of evolutionary young genes among those that have low similarity or even lack HomoloGene data (Appendix D Tab. D3). We concluded that initial divergence and subsequent fixation may be a common pattern of imprinted genes. However, changes that affect gene expression are much more frequent and important than mutations of protein-encoding sequences (Wilkins and Haig 2001). Rearrangement events like gene duplications and insertions are a typical feature in the evolution of imprinting clusters (Paulsen et al. 2005, Rapkins et al. 2006, Hore et al. 2007, Warren et al. 2008). Most remarkably, evolution is still going on with the appearance of lineage-specific imprinted genes.

4.8 Paralogous genes and the evolution of imprinting

The existence of paralogs may have enabled a greater divergence of imprinted genes by relaxation of purifying selection. One might expect that due to functional redundancy, the orthologs of genes with paralogs are more diverged from each other than those of genes that do not possess paralogs, so-called singletons. Strangely enough, exactly the opposite is the case (Jordan et al. 2004, Brunet et al. 2006, Chain and Evans 2006, Conant and Wolfe 2008, Studer and Robinson-Rechavi 2009). Retention of gene duplicates seems to be favored for genes with important functions, whereas less

important genes tend to remain singletons (Jordan et al. 2004). Gene duplication appears as the driving force for the evolution of new functions and species (He and Zhang 2005, Han et al. 2009). He and Zhang (2005) estimated that by this means, about 10,000 new protein interactions evolved in humans since they diverged from mice. Notably, new duplicates still have the same targets as the original, but by divergence, the type of interactions may switch from activating to repressing (Bridgham et al. 2008).

Most mammalian paralogs date back to a whole genome duplication in Euteleostomi, after which many gene duplicates were deleted (McLysaght et al. 2002, Dehal and Boore 2005). Younger ones arose by tandem duplications and interchromosomal rearrangements mediated by transposons, which is frequent in mammals (Jaillon et al. 2004). In the mouse genome, more duplicates have been retained that became processed pseudogenes in the human genome (Shiu et al. 2006), which is reflected in our observation that a higher percentage of murine genes has duplicates. Whereas most duplicates became nonfunctional and eventually so mutated that they cannot be recognized any more, other gene pairs underwent combined evolution. If the original protein had several functions, they can be split up between the two copies, a process termed subfunctionalization. In the case of neofunctionalization, one copy develops a new function. Apparently, these two processes cannot be clearly separated, leading to the concept of subneofunctionalization (He and Zhang 2005). In terms of evolution, such duplication events are characterized by two phases. First, immediately after duplication, relaxed constraints – sometimes also positive selection – act on the two genes, leading to their fast divergence. After new functions of the paralogs have been established, they are again subject to purifying selection (Jordan et al. 2004, Brunet et al. 2006, He and Zhang 2005, Conant and Wolfe 2008, Studer and Robinson-Rechavi 2009). Still, it can be assumed that pairs of paralogs are similar in function to a comparable degree as orthologs in different species (Studer and Robinson-Rechavi 2009).

Interestingly, when studying the conservation of orthologs, divergence of one paralog is balanced by conservation of the other (Jordan et al. 2004, Brunet et al. 2006, Chain and Evans 2006, Conant and Wolfe 2008). As to imprinted genes, we made the striking observation that they are obviously concentrated in the faster evolving group (Fig. 4.2) and are more diverged from their paralogs as the average pair. The strict conservation of their paralogs indicates that these may have maintained the original functions and thus may have facilitated changes in the sequences of imprinted genes. In the mouse, divergence between imprinted genes and their paralogs is higher than in human, which, given the strong conservation of the paralogs, again speaks for an increased divergence of imprinted genes in the rodent ancestor. Increased conservation of singletons between mouse and rat, in turn, corroborates purifying selection due to lack of a buffering paralog.

It has further been suggested that many imprinted genes arose from transposons (Walter and Paulsen 2003, Wood et al. 2007). Retroposed paralogs are integrated in a new genomic and epigenomic context, hence they will likely exhibit different expression patterns. In other tissue types than the original ones, they will experience fast evolution leading to new functions. This scenario is strongly supported by the finding that many young, lineage-specific duplicates have undergone positive selection whereas the original genes remained subject to strict constraints (Han et al. 2009). For the imprinted genes and their paralogs analyzed in this study, direction of the duplication has not been assessed. Since most events date back to the Euteleostomi whole genome duplication (Appendix D Tab. D3), retrotransposition does not seem to be a major factor for their divergence. However, as already mentioned in section 4.5, a few more recently duplicated genes might have introduced regulatory elements that became imprinting control centers into previously

non-imprinted regions and thus may have influenced the evolution of whole gene clusters.

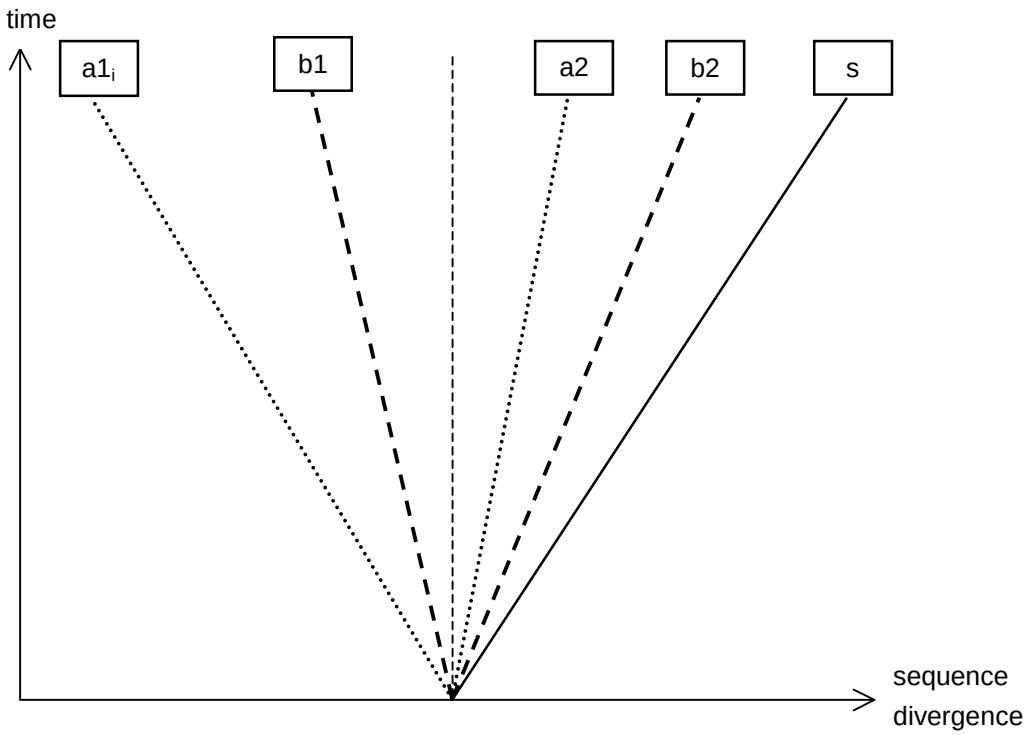


Figure 4.2: Complementary divergence
The orthologs of single copy genes (s) are more diverged than the orthologs of genes that possess paralogs. Regarding paralogous pairs of biallelically expressed genes (b1, b2), one is usually more diverged than the other. If an imprinted gene ($a1_i$) has a paralog (a2), the imprinted gene itself is in most cases the more divergent one.

4.9 Conclusions and outlook

Imprinted genes are an enigma because they represent a very heterogeneous group that has little in common apart from their monoallelic expression dependent on their parental origin and their important role in embryonic development. They differ in terms of their regulation, sequence properties, function, conservation and evolutionary history. Nevertheless, they share some features that distinguish them from biallelically expressed genes. The studies presented in this thesis could statistically verify or refute existing hypotheses on such properties. Moreover, the association of imprinted genes with tandem repeats in their CpG islands, enrichment of conserved ancient repetitive elements, as well as the lineage-specific evolutionary patterns of the proteins encoded by them and their paralogs, might provide alternative means for the identification of new imprinting candidates (compare also section 4.1.2). In contrast to large-scale predictions involving dozens of features (Luedi et al. 2005, 2007), detailed analysis of the apparently most prominent ones allows a better biological interpretation.

In summary, we could shed some light onto imprinting mechanisms – only to see that the

implications are larger than anticipated. As figure 4.3 shows, different features participate in a complex network of interactions that regulate parent-of-origin dependent monoallelic expression. Future research will concentrate on more detailed analyses of intronic and intergenic conserved elements in order to find the most promising potential regulatory elements that can be experimentally assessed. Additionally, related projects will investigate the coexpression patterns of imprinted genes and interactions of proteins encoded by them to gain a better understanding about their involvement in metabolic pathways.

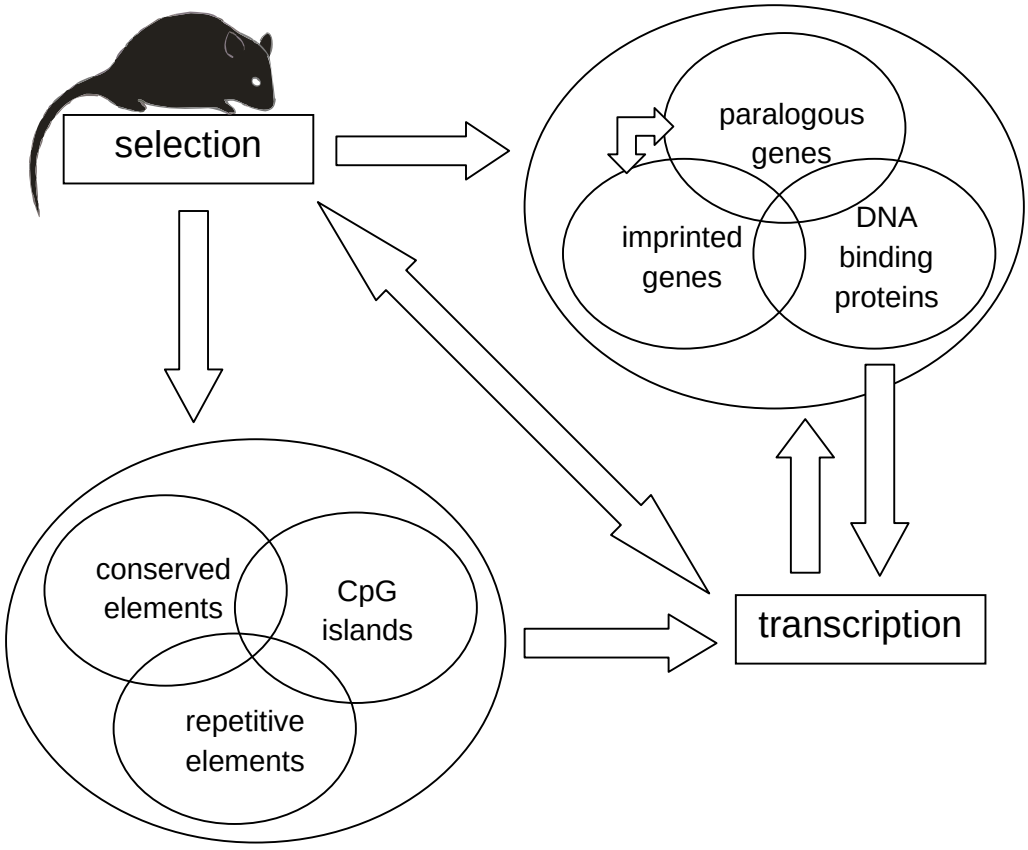


Figure 4.3: Overview of the imprinting regulatory network
Genomic sequences of imprinted regions contain several features that influence DNA structure and, either directly or indirectly, transcription. These features are highly interconnected and may have in part redundant functions. Moreover, natural selection acts via the phenotype (represented by the mouse) on several levels, in which the existence of paralogous genes may have decisive influence. Features not included here are DMRs, which may coincide with conserved elements, CGIs, or repeats; histone modifications, which influence DNA structure as well; and antisense transcripts, which are involved in transcriptional regulation and are a target of selection.

Appendices

Appendix A

Table A1: Locations and data of genomic sequences
The abbreviation dir. stands for the direction of transcription (+: on forward strand, C: on reverse complement). Ns in the mouse represent stretches of undefined nucleotides. All data refer to the genomic sequences of the genes with 10 kb of upstream and downstream sequences each. UCSC coordinates are zero-based.

human G1, NCBI build 35.1

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	G+C content (%)	CpG content (%)
ADAM19	NT_023133.12	C	1703872	118455	5	46	1.2
ANKRD1	NT_030059.12	C	11410373	29176	10	40	0.8
AQP8	NT_010393.15	+	16531364	31969	16	47	1.4
ARF4	NT_022517.17	C	57487130	45989	3	41	1.3
ARHGAP24	NT_016354.17	+	11336131	92396	4	37	0.7
ARL2BP	NT_010498.15	+	10883237	28508	16	46	1.6
ASB13	NT_077569.2	C	34584	46854	10	48	2.0
ATP5G2	NT_029419.10	C	16192143	27596	12	45	1.7
BACE1	NT_033899.7	C	20708833	50556	11	47	1.5
BCKDHB	NT_007299.12	+	18626535	259624	6	37	0.6
CACNA1B	NT_024000.16	+	1545244	264391	9	50	1.8
CHRNB3	NT_007995.14	+	12862951	59648	8	46	1.6
CIR	NT_005403.15	C	25412297	67564	2	39	1.2
CLDN12	NT_007933.14	+	15257072	32473	7	39	0.8
CPT1B	NT_011526.6	C	217353	29707	22	55	2.7
CRYBB3	NT_011520.10	+	4976394	27500	22	48	1.4
CSNK1D	NT_010663.14	C	406349	49332	17	56	3.4
CSTF1	NT_011362.9	+	20010527	31489	20	43	1.7
DDX20	NT_019273.17	+	8374369	31946	1	42	1.1
DPYSL4	NT_024040.14	+	262877	38863	10	58	3.5
ELOVL3	NT_030059.12	+	22724669	23204	10	49	2.2
FAIM2	NT_029419.10	C	12393986	57041	12	53	1.7
GLP1R	NT_007592.14	+	29864867	58904	6	50	1.2
GMPPA	NT_005403.15	+	70563028	28099	2	59	2.6
GPR142	NT_010641.15	+	6279917	25095	17	55	2.2
GSTM1	NT_019273.17	+	6306557	25926	1	46	1.3
GTF3C5	NT_035014.4	+	2673053	47775	9	53	2.4
HIST4H4	NT_009714.16	C	7672628	20412	12	42	1.9
HOXB6	NT_010783.14	C	5316391	29222	17	54	3.6
HSBP1	NT_010498.15	+	37445827	23647	16	45	1.7
HSD11B1	NT_021877.17	+	3312657	68746	1	41	0.6
HTR3B	NT_033899.7	+	17328005	61695	11	43	1.2
IREB2	NT_010194.16	+	49510946	80842	15	41	1.2
IRF3	NT_011109.15	C	22421019	26286	19	57	3.3
KIF13B	NT_023666.17	C	7289083	215846	8	43	1.3
KLF3	NT_016297.15	+	5804460	54585	4	43	1.5
LPL	NT_030737.9	+	7631706	47992	8	42	1.3
LRG1	NT_011255.14	C	4467228	22809	19	56	3.2
LRRC6	NT_008046.15	C	46792628	123365	8	40	0.7
LTBP2	NT_026437.11	C	55954640	134148	14	51	1.6
MAFG	NT_010663.14	C	80256	24945	17	61	4.4
MBNL1	NT_005612.14	+	58470983	217741	3	36	0.7

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	G+C content (%)	CpG content (%)
MCSP (*SMCP*)	NT_004487.17	+	3331153	26726	1	40	0.7
MOS	NT_008183.18	C	8868854	21041	8	44	2.0
MPP2	NT_010783.14	C	596537	52158	17	50	1.5
NKX6-1	NT_016354.17	C	9899141	24952	4	46	2.8
NOL1	NT_009759.15	C	6510290	31056	12	51	2.3
OSBPL9	NT_032977.7	+	5892157	191378	1	39	0.9
OSMR	NT_006576.15	+	38808893	108267	5	41	0.9
PDCD6IP	NT_022517.17	+	33770086	88417	3	37	0.9
PDE6D	NT_005403.15	C	82796562	68828	2	43	1.4
PLAT	NT_007995.14	C	12343144	52440	8	47	1.7
POFUT2	NT_011515.11	C	1991188	43949	21	54	3.3
POLK	NT_006713.14	+	25391939	107553	5	37	0.8
PPGB (*CTSA*)	NT_011362.9	+	9563112	27253	20	52	2.0
PPP4R1	NT_010859.14	C	9526842	87726	18	41	1.3
PRDM1	NT_025741.13	+	10693624	43619	6	42	1.3
PRKCBP1 (*ZMYND8*)	NT_011362.9	C	10881289	167094	20	47	1.6
PTPRA	NT_011387.8	+	2774830	194486	20	44	1.3
PTPRO	NT_009714.16	+	8224461	294849	12	38	0.7
PTPRU	NT_004538.16	+	745980	110289	1	51	1.6
RAB17	NT_005120.15	C	4405719	36772	2	53	1.9
RBMS3	NT_022517.17	+	29253029	729731	3	36	0.6
REL	NT_022184.14	+	39914733	61171	2	39	1.1
RGS3	NT_008470.17	+	23535583	155640	9	49	1.2
SAMD10	NT_011333.5	C	1332098	25525	20	56	3.5
SC4MOL	NT_016354.17	+	90733989	35010	4	39	1.0
SGCA	NT_010783.14	+	6886669	29899	17	53	2.0
SLC14A2	NT_010966.13	+	24673868	88307	18	46	1.0
SLC1A4	NT_022184.14	+	44022527	54406	2	45	1.4
SPRR2A	NT_004487.17	C	3508952	21392	1	37	0.6
STAT3	NT_010755.15	C	4179639	95171	17	46	1.6
STX16	NT_011362.9	+	22269235	48255	20	44	1.7
TCF19	NT_007592.14	+	21975498	23292	6	52	2.0
TRIM14	NT_008470.17	C	8157842	54852	9	47	1.5
WNT11	NT_033927.7	C	6110119	40205	11	58	2.8
YTHDF1	NT_011333.5	C	553416	40754	20	53	2.4
ZDHHC3	NT_022517.17	C	44896665	70954	3	47	1.2
ZHX1	NT_008046.15	C	37468881	45845	8	41	1.5

human G2, NCBI build 35.1

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	G+C content (%)	CpG content (%)
ABCF1	NT_007592.14	+	21387421	39920	6	48	1.7
ADARB1	NT_011515.11	+	1801851	171960	21	46	1.6
AGR2	NT_007819.15	C	16117109	33173	7	39	1.1
ANKS1	NT_007592.14	+	25705291	222150	6	45	1.2
APBB2	NT_006238.10	C	508591	419851	4	42	1.2
BCL2	NT_025028.13	C	8571425	215467	18	43	1.2
C9	NT_006576.15	C	39247763	99650	5	38	0.6
CABP2	NT_033903.7	C	12582213	24450	11	52	1.7
CALB1	NT_008046.15	C	4279011	44270	8	36	0.7
CALM3	NT_011109.15	+	19362787	29471	19	52	2.2
CDH6	NT_006576.15	+	31156553	151442	5	38	0.8

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	G+C content (%)	CpG content (%)
CHAF1B	NT_011512.10	+	23409559	51437	21	47	1.9
CLCN2	NT_005612.14	C	90549552	34874	3	54	2.4
CLK1	NT_005403.15	C	51917221	31552	2	44	1.8
CNTNAP1	NT_010755.15	+	4548928	37201	17	55	2.8
COMMD2	NT_005612.14	C	55943686	31747	3	38	0.8
CPN1	NT_030059.12	C	20540598	59523	10	44	1.0
CTSK	NT_004487.17	C	1249042	32126	1	43	1.1
CUL1	NT_007914.14	+	8961949	122270	7	42	1.5
CUTL1 (*CUX1*)	NT_007933.14	+	26632352	487959	7	49	2.1
DCT	NT_009952.14	C	8171519	60081	13	40	0.8
DHFR	NT_006713.14	C	30506406	48769	5	41	1.2
EFS	NT_026437.11	C	4815452	29231	14	54	2.0
ENPP1	NT_025741.13	+	36223589	103191	6	38	0.9
EVC	NT_006051.17	+	1790369	137861	4	48	1.5
FBLN1	NT_011522.5	+	1155300	118296	22	53	2.2
FUT1	NT_011109.15	C	21509458	24580	19	56	3.2
GDNF	NT_006576.15	C	37778510	44030	5	46	1.9
GPR2 (*CCR10*)	NT_010755.15	C	4545731	22411	17	59	4.3
GSPT1	NT_010393.15	C	3269618	63270	16	43	1.7
H1FX	NT_005612.14	C	35518769	21503	3	54	2.2
HBXIP	NT_019273.17	C	7019994	26668	1	44	1.6
HIST1H4H	NT_007592.14	C	17133605	20374	6	42	1.5
IL1R1	NT_022171.13	+	4826426	45933	2	42	1.0
IL1RAP	NT_005612.14	+	96717049	157411	3	38	0.9
ITGB1	NT_008705.15	C	15154645	77503	10	39	1.0
KLF4	NT_008470.17	C	17558340	24621	9	46	2.6
MADCAM1	NT_011255.14	+	426490	28853	19	60	4.4
MASP1	NT_005612.14	C	93421100	93529	3	44	0.9
MFAP5	NT_009714.16	C	1547514	36894	12	43	1.2
MPDU1	NT_010718.15	+	7074514	24345	17	53	2.4
MX1	NT_011512.10	+	28450024	52985	21	48	1.6
NCF2	NT_004487.17	C	33923637	55314	1	45	1.0
NFE2L1	NT_010783.14	+	4769024	33097	17	49	1.7
NPY1R	NT_016354.17	C	88730248	28632	4	37	0.8
NR2F1	NT_023148.12	+	1224514	29188	5	50	3.5
OGG1	NT_022517.17	+	9720705	37638	3	51	1.8
OLIG2	NT_011512.10	+	20050163	23209	21	54	3.5
PDHB	NT_022517.17	C	58343398	26197	3	45	1.4
PIK3R3	NT_032977.7	C	315206	112501	1	42	1.1
PPAP2C	NT_011255.14	C	211046	30390	19	54	2.9
PRIM2 (*PRIM2A*)	NT_007592.14	+	48030653	350955	6	38	0.7
PRKDC	NT_008183.18	C	529021	207075	8	42	1.4
PSTPIP1	NT_010194.16	+	48067874	62118	15	54	1.6
RBM9	NT_011520.10	C	15520162	116782	22	39	0.9
RGNEF	NT_006713.14	+	23506341	335836	5	39	0.8
RHOD	NT_033903.7	+	12120116	35164	11	55	2.3
RNF44	NT_023133.12	C	20753286	30722	5	55	3.0
RPL27	NT_010755.15	+	4864742	24531	17	48	1.9
SEPHS2	NT_010393.15	C	21758032	22272	16	47	1.8
SOCS5	NT_022184.14	+	25732032	83829	2	38	0.9
SOCS7	NT_010755.15	+	222424	67935	17	46	1.5
SOX17	NT_008183.18	+	7213847	22437	8	47	3.3
SUMO1	NT_005403.15	C	53270325	52402	2	42	1.5
SVIL	NT_008705.15	C	11711602	298454	10	44	1.4

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	G+C content (%)	CpG content (%)
TACC2	NT_030059.12	+	42487215	285369	10	48	1.6
TAX1BP3	NT_010718.15	C	3159562	25788	17	56	3.0
TFEB	NT_007592.14	C	32499967	71082	6	54	1.7
TIMM23	NT_008583.16	C	133245	51247	10	40	1.1
TJP1	NT_010194.16	C	772916	142348	15	39	1.1
TNFRSF8	NT_021937.17	+	6650801	100831	1	51	1.8
TNFSF18	NT_004487.17	C	23409390	29606	1	37	0.5
TOPORS	NT_008413.16	C	32520543	32059	9	41	1.5
TRRAP	NT_007933.14	+	23700389	154727	7	47	1.9
TUBA2	NT_024524.13	C	717920	28017	13	49	1.5
UNG	NT_009775.15	+	94924	33384	12	47	1.9
VPS24	NT_022184.14	C	65536493	79937	2	41	1.2
YPEL5	NT_022184.14	+	9175761	33572	2	42	1.4
ZNHIT1	NT_007933.14	+	26034045	26487	7	55	2.7

human G3, UCSC hg18

gene symbol (synonym)	chr.	start	end	length (bp)	dir.	G+C content (%)	CpG content (%)
ABCD2	chr12	38222813	38310237	87424	C	35	0.6
ABCG5	chr2	43883115	43929462	46347	C	44	1.2
AFM	chr4	74556325	74598581	42256	+	37	0.5
ALDH4A1	chr1	19060512	19111659	51147	C	54	2.1
APH1B	chr15	61346843	61395165	48322	+	41	1.3
ATAD1	chr10	89492854	89577897	85043	C	37	0.7
ATRN	chr20	3389675	3589760	200085	+	41	1.0
BACH1	chr21	29583090	29666086	82996	+	41	1.1
BCMP11 (*AGR3*)	chr7	16855555	16898138	42583	C	37	0.7
BNIP1	chr5	172494050	172533996	39946	+	44	1.4
C16orf53	chr16	29725028	29751317	26289	+	57	3.3
CA8	chr8	61253976	61366508	112532	C	39	0.8
CD207	chr2	70900855	70926461	25606	C	47	1.0
CHD1	chr5	98208808	98300138	91330	C	36	1.0
CLCA2	chr1	86652412	86704826	52414	+	39	0.6
COLEC10	chr8	120138626	120198376	59750	+	37	0.5
CREB3L4	chr1	152197020	152223456	26436	+	51	2.5
CXXC5	chr5	138998701	139053110	54409	+	57	2.3
DMAP1	chr1	44441711	44468938	27227	+	47	1.3
DSCR1L1 (*RCAN2*)	chr6	46286427	46411490	125063	C	40	0.6
DUS2L	chr16	66604704	66680684	75980	+	48	1.5
EDEM1	chr3	5194432	5246642	52210	+	43	1.3
EPS15L1	chr19	16323407	16453762	130355	C	51	2.1
FAM104A	chr17	68705086	68750128	45042	C	45	1.7
FAM21C	chr10	45532672	45618417	85745	+	42	1.3
FASTK	chr7	150394640	150418884	24244	C	60	4.1
FLJ22313 (*HERPUD2*)	chr7	35628796	35711254	82458	C	39	0.9
GRRP1	chr1	26348097	26371705	23608	+	54	2.7
GTF3C1	chr16	27369435	27478752	109317	C	49	1.7
HDCMA18P (*LARP7*)	chr4	113767568	113808190	40622	+	39	1.2
HOXA6	chr7	27141640	27163893	22253	C	53	4.1
HYAL2	chr3	50320243	50345146	24903	C	56	2.7

gene symbol (synonym)	chr.	start	end	length (bp)	dir.	G+C content (%)	CpG content (%)
KBTBD10	chr2	170064457	170101018	36561	+	39	1.0
KIAA0523 (*WSCD1*)	chr17	5904657	5978469	73812	+	50	1.7
LOC136242	chr7	141172554	141197690	25136	C	39	0.6
LOC285148 (*IAH1*)	chr2	9522120	9556041	33921	+	45	1.8
LRRC6	chr8	133643628	133766995	123367	C	40	0.7
LSDP5	chr19	4463545	4496208	32663	C	56	3.0
LSM1	chr8	38130014	38163183	33169	C	45	1.9
MRPL24	chr1	154963717	154987547	23830	C	52	2.1
MRPL40	chr22	17790035	17813594	23559	+	48	2.0
MTNR1B	chr11	92332436	92365596	33160	+	44	0.9
MTRR	chr5	7912216	7964233	52017	+	38	1.0
NDUFS2	chr1	159425728	159460806	35078	+	50	1.6
OCIAD2	chr4	48572163	48613572	41409	C	42	0.8
OR8A1	chr11	123935174	123956154	20980	+	38	0.7
PARL	chr3	185019867	185095387	75520	C	44	1.7
PDE4D	chr5	58290622	59235378	944756	C	37	0.6
PDZD6 (*INTU*)	chr4	128763569	128867380	103811	+	37	0.7
PLCZ1	chr12	18717382	18792185	74803	C	36	0.6
PLD3	chr19	45536446	45586229	49783	+	52	2.1
POLM	chr7	44068373	44098607	30234	C	53	1.9
PRMT3	chr11	20355678	20497349	141671	+	37	0.8
PRPH	chr12	47965175	47988747	23572	+	50	2.4
PSMB7	chr9	126145565	126227542	81977	C	45	1.5
PUS7L	chr12	42398678	42448863	50185	C	37	0.7
RAB21	chr12	70424924	70477417	52493	+	38	1.0
RAP1GDS1	chr4	99391549	99594035	202486	+	37	0.8
RBM22	chr5	150040548	150070817	30269	C	42	1.2
REG1B	chr2	79155658	79178627	22969	C	41	0.8
RSAD1	chr17	45901188	45928335	27147	+	50	2.0
SCN2B	chr11	117528729	117562546	33817	C	52	1.7
SERTAD4	chr1	208462817	208493061	30244	+	44	1.7
SESN1	chr6	109404339	109531970	127631	C	38	0.9
SH2D5	chr1	20908811	20941720	32909	C	54	2.7
SLC2A14	chr12	7847664	7926762	79098	C	45	1.7
SMAD9	chr13	36310320	36361966	51646	C	42	1.1
TANC1	chr2	159523391	159807414	284023	+	43	1.1
TGM6	chr20	2299553	2371399	71846	+	45	1.1
TMEM44	chr3	195779691	195845402	65711	C	50	2.4
TPSAB1 (*TPSB2*)	chr16	1220678	1242555	21877	+	57	3.3
TUBG1	chr17	38005219	38030775	25556	+	48	1.9
TUFM	chr16	28751612	28775144	23532	C	50	2.2
TXNDC5	chr6	7816752	8019596	202844	C	42	1.2
UBE2V1	chr20	48121067	48175901	54834	C	46	1.6
UBL3	chr13	29226543	29332160	105617	C	37	0.8
VCP	chr9	35036560	35072564	36004	C	48	1.8
XRCC5	chr2	216672376	216789248	116872	+	40	0.9
ZNF629	chr16	30687270	30716024	28754	C	55	2.7

mouse G1, NCBI builds 33.1 and 34.1

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	Ns (%)	G+C content (%)	CpG content (%)
1700023B02Rik	NT_039207.3	C	14108761	48632	2	0.0	42	1.6
3632413B07Rik	NT_039210.3	C	16172156	131642	2	0.0	50	1.7
Adam19	NT_096135.1	+	11156817	111292	11	0.0	46	1.2
Ankrd1	NT_039689.3	C	2362099	27880	19	0.0	43	1.2
Aqp8	NT_039433.3	+	41540093	25675	7	0.0	45	0.9
Arf4	NT_039597.3	+	1642472	39062	14	0.0	39	0.9
Arhgap24[a]	NT_039308.4	+	6780727	443371	5	3.5	42	0.9
Arl2bp	NT_078575.2	+	19707068	27668	8	0.0	47	1.4
Asb13[a]	NT_039573.4	+	508457	37704	13	0.0	44	1.1
Atp5g2[a]	NT_039621.4	C	63923802	28180	15	0.0	44	1.2
Bace1	NT_039473.3	+	5275698	43913	9	0.5	47	1.2
Bckdhb	NT_039475.3	+	3523653	196217	9	4.0	42	1.0
Cacna1b	NT_039206.3	C	2025021	176778	2	0.0	47	1.0
Chrnb3	NT_039456.3	+	5831769	51853	8	2.7	44	1.1
Cldn12	NT_039297.3	C	2501828	29765	5	0.0	42	1.2
Cpt1b	NT_039621.3	C	50936601	28667	15	4.3	49	1.9
Crybb3	NT_080546.2	C	3055117	42795	5	3.5	52	1.7
Csnk1d	NT_039521.3	C	32353905	47898	11	0.0	49	1.4
Cstf1	NT_039210.3	+	22757855	30084	2	0.0	49	1.9
D630039A03Rik	NT_039260.3	C	31445342	27733	4	0.0	47	1.1
Ddx20	NT_039239.3	C	7496201	29298	3	0.0	44	1.4
Dpysl4	NT_039436.3	+	13932	35697	7	0.3	52	1.6
Elovl3	NT_039692.3	+	6506766	23796	19	0.0	48	1.6
Faim2	NT_039621.3	C	61086916	47899	15	0.2	52	1.4
Glp1r	NT_039649.3	+	7430097	54613	17	4.3	50	1.3
Gmppa[a]	NT_039171.4	+	306942	27190	1	0.0	55	2.0
Gpr142	NT_039521.3	+	26189407	27822	11	0.0	52	1.6
Gstm1	NT_039239.3	C	9892371	25717	3	0.0	47	1.1
Gtf3c5	NT_039206.3	C	5984961	36932	2	0.0	51	1.4
Hist4h4	NT_039359.3	C	4413381	20312	6	0.0	45	1.9
Hoxb6	NT_039521.3	+	7689648	22399	11	0.0	51	2.7
Hsbp1	NT_078575.2	+	44467205	24392	8	3.1	49	1.8
Hsd11b1	NT_039190.3	C	1158186	62357	1	0.0	45	0.8
Htr3b	NT_039473.3	C	8383753	49643	9	0.8	45	0.9
Ireb2	NT_039474.3	+	1137655	68706	9	0.6	40	0.8
Irf3	NT_039420.3	+	1828282	25063	7	0.0	53	2.2
Kif13b	NT_039606.3	+	10957710	213501	14	1.2	45	0.9
Klf3	NT_039305.3	+	26646917	46126	5	1.0	46	1.7
Lpl	NT_039462.3	+	4123874	46244	8	0.6	42	0.8
Lrg1	NT_039656.3	C	2193208	22267	17	0.4	55	2.2
Lrrc6	NT_039621.3	C	27775452	142231	15	1.1	41	0.6
Ltbp2	NT_039551.3	C	29721411	113320	12	0.5	50	1.3
Mafg	NT_039521.3	C	32018831	25200	11	0.0	53	2.5
Mbnl1	NT_039230.3	+	9820111	147920	3	2.5	38	0.8
Mcsp (*Smcp*)	NT_039238.3	C	671213	25128	3	0.0	40	0.3
Mos	NT_039258.3	C	787930	21175	4	0.0	41	1.1
Mpp2	NT_039521.3	C	13447496	51484	11	0.0	50	1.7
Nkx6-1	NT_039308.3	C	24165277	25514	5	0.0	47	2.5
Nol1	NT_039356.3	+	599259	33015	6	1.7	51	2.0
Osbpl9	NT_039264.3	C	9174906	60943	4	0.0	42	1.0
Osmr	NT_039617.3	C	3632082	80887	15	0.2	43	1.2
Pdcd6ip	NT_095756.1	C	4248391	72756	9	2.2	43	1.1
Pde6d	NT_039173.3	C	886368	59697	1	0.5	44	0.9

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	Ns (%)	G+C content (%)	CpG content (%)
Plat	NT_039456.3	+	1152303	45077	8	0.0	47	1.2
Pofut2	NT_039496.3	+	3088740	29390	10	2.5	52	1.9
Polk	NT_039590.3	C	4479924	81796	13	0.0	41	1.1
Ppgb (*Ctsa*)	NT_039210.3	+	15220653	27227	2	0.0	51	2.0
Ppp4r1	NT_039657.3	+	9317186	41703	17	0.2	46	1.3
Prdm1	NT_039492.3	C	22096825	41456	10	1.3	45	1.5
Ptpra	NT_039207.3	+	71183216	70554	2	0.0	45	1.0
Ptpro	NT_039359.3	+	4884128	230813	6	0.0	43	1.2
Ptpru	NT_078435.2	C	1194106	89832	4	0.0	51	1.7
Rab17[a]	NT_039173.4	C	11858036	31486	1	0.0	50	1.2
Rbms3	NT_039482.3	C	2263424	696072	9	1.0	43	0.9
Rel	NT_039515.3	C	20627558	48980	11	0.0	39	0.9
Rgs3	NT_039260.3	+	35156058	162489	4	0.0	49	1.3
Samd10	NT_039212.3	C	3871045	23927	2	0.0	49	1.9
Sc4mol	NT_039461.3	C	9294915	35361	8	0.3	42	1.0
Sgca	NT_039521.3	C	6353685	33120	11	0.0	49	1.1
Slc14a2	NT_039676.3	C	207063	470866	18	0.0	43	0.9
Slc1a4	NT_039515.3	C	17187777	50465	11	0.0	46	1.2
Sprr2a	NT_039238.3	+	338078	23955	3	0.0	39	0.6
Stat3	NT_039521.3	C	12278713	71169	11	0.0	47	1.2
Tcf19	NT_039650.3	C	35835	24021	17	3.3	52	2.0
Trim14	NT_039260.3	C	20020648	49295	4	0.0	49	1.8
Wnt11	NT_039433.3	+	16708280	34904	7	0.0	54	1.9
Ythdf1	NT_039212.3	C	3180292	36558	2	0.0	47	1.1
Zdhhc3	NT_039482.3	C	8806958	58940	9	0.0	46	1.0
Zhx1	NT_039621.3	C	19370726	42348	15	0.0	41	1.0

[a] NCBI build 34.1

mouse G2, NCBI builds 33.1 and 34.1

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	Ns (%)	G+C content (%)	CpG content (%)
Abcf1	NT_039650.3	C	476788	32931	17	0.3	48	1.6
Adarb1	NT_039496.3	C	3123871	85020	10	0.7	46	1.2
Agr2	NT_039548.3	+	28009364	31157	12	0.0	41	0.9
Anks1	NT_039649.3	+	4420635	173298	17	0.0	50	1.6
Apbb2	NT_039305.3	C	28115450	336517	5	0.0	45	1.3
Bcl2	NT_039174.3	C	853211	196084	1	0.1	43	0.8
C9	NT_039617.3	+	3259655	73393	15	2.8	40	0.8
Cabp2	NT_082868.2	+	861960	23851	19	0.0	49	1.3
Calb1	NT_039258.3	+	12798364	45385	4	0.0	38	0.8
Calm3	NT_039395.3	C	1151805	28577	7	0.8	52	2.0
Cdh6	NT_039618.3	C	3989926	77048	15	1.2	42	1.2
Chaf1b	NT_039625.3	+	28794369	42026	16	3.6	47	1.5
Clcn2	NT_039624.3	C	9412228	33278	16	3.9	51	1.9
Clk1	NT_039170.3	C	36091169	32011	1	0.6	42	1.1
Cntnap1	NT_039521.3	+	12566594	33731	11	0.0	53	2.2
Commd2	NT_039230.3	C	6912791	26312	3	2.4	41	0.9
Cpn1	NT_039692.3	C	4327415	50214	19	0.0	47	1.4
Ctsk	NT_039238.3	+	3370403	30077	3	0.0	43	0.9
Cul1	NT_039341.3	+	16747462	92538	6	1.7	40	0.7
Cutl1 (*Cux1*)	NT_080526.2	C	1582466	339106	5	0.0	46	1.0
Dct	NT_039609.3	C	19911069	59379	14	0.0	44	1.3
Dhfr	NT_039590.3	+	327786	51428	13	1.2	42	1.0

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	Ns (%)	G+C content (%)	CpG content (%)
Efs	NT_039606.3	C	1247802	29273	14	1.2	52	2.1
Enpp1	NT_039492.3	C	2141746	90873	10	1.8	42	1.2
Evc	NT_039303.3	C	634119	63039	5	1.8	50	1.2
Fbln1	NT_039621.3	+	46689116	99646	15	0.0	50	1.4
Fut1	NT_039420.3	+	2456171	21134	7	2.5	51	2.0
Gdnf	NT_039617.3	+	4655284	46590	15	0.6	47	1.6
Gpr2 (*Ccr10*)	NT_039521.3	C	12563475	22446	11	0.0	54	3.0
Gspt1	NT_096986.1	C	7904894	53596	16	0.0	41	1.0
H1fx[a]	NT_039353.4	C	2175226	21062	6	0.0	50	2.0
Hbxip	NT_039239.3	+	9159626	25225	3	3.9	44	1.0
Hist1h4h	NT_039578.3	+	10217292	20470	13	0.0	42	1.9
Il1r1	NT_039170.3	+	17835687	111803	1	4.0	44	1.0
Il1rap	NT_039624.3	+	15364256	153043	16	2.0	40	0.7
Itgb1	NT_078575.2	+	53797525	67059	8	1.5	40	0.8
Klf4	NT_039260.3	C	29041018	25202	4	0.0	47	2.2
Madcam1	NT_039496.3	+	5506711	23954	10	1.3	52	1.7
Masp1	NT_039624.3	C	12230587	91051	16	0.6	45	0.9
Mfap5	NT_094510.1	+	13706977	35935	6	2.3	45	1.1
Mpdu1	NT_096135.1	C	34730290	25939	11	0.0	50	2.2
Mx1	NT_039627.3	C	1212176	33913	16	3.6	46	0.6
Ncf2	NT_039184.3	+	5159970	49102	1	3.8	47	1.1
Nfe2l1	NT_039521.3	C	8207909	30524	11	0.0	48	1.5
Npy1r	NT_039462.3	+	1902238	22261	8	0.0	41	1.1
Nr2f1	NT_039589.3	C	1581516	29319	13	4.8	47	2.4
Ogg1	NT_094510.1	+	4406984	27285	6	4.3	49	1.4
Olig2	NT_039625.3	+	26084706	20972	16	0.0	52	2.9
Pdhb	NT_081916.2	C	954503	27955	14	0.7	45	1.2
Pik3r3	NT_039264.3	+	16335170	101143	4	0.0	42	0.9
Ppap2c	NT_039496.3	C	5367540	26440	10	0.0	51	1.1
Prim2	NT_039170.3	C	11023971	235293	1	1.8	42	1.0
Prkdc	NT_039624.3	+	4349384	223531	16	1.6	40	0.8
Pstpip1	NT_039474.3	+	2353791	59433	9	4.0	50	1.2
Rbm9	NT_039621.3	C	38537062	244592	15	0.0	44	0.9
Rgnef	NT_039590.3	C	5898237	327692	13	1.5	44	1.2
Rhod	NT_082868.2	C	1203927	33968	19	0.0	49	1.2
Rnf44	NT_039586.3	C	314390	34540	13	0.6	51	1.8
Rpl27	NT_039521.3	+	12832892	23119	11	0.0	46	1.4
Sephs2	NT_039433.3	C	45340998	22178	7	0.0	47	1.4
Socs5	NT_039658.3	+	20814688	49907	17	0.0	45	1.6
Socs7	NT_039521.3	+	8753028	51082	11	0.0	47	1.4
Sox17	NT_039169.3	C	1493577	25485	1	0.0	43	1.5
Sumo1	NT_039170.3	C	37329961	51258	1	1.0	40	0.7
Svil	NT_039674.3	+	2080977	218694	18	0.0	45	1.0
Tacc2	NT_081265.2	+	349024	187270	7	0.4	48	1.3
Tax1bp3	NT_096135.1	+	38250668	24964	11	0.0	50	1.6
Tfeb (*Tcfeb*)	NT_039655.3	+	7489536	26465	17	0.4	53	1.5
Timm23	NT_039598.3	C	5396696	41704	14	0.2	40	1.0
Tjp1	NT_039428.3	C	5311549	95146	7	0.8	41	1.0
Tnfrsf8	NT_039267.3	C	12572266	66172	4	0.0	48	0.9
Tnfsf18[a]	NT_039185.4	+	5323735	30636	1	0.0	39	0.6
Topors	NT_039260.3	C	14619727	30224	4	0.0	43	1.6
Trrap	NT_080570.2	+	879272	108084	5	0.5	49	1.8
Tuba2	NT_039621.3	C	60520371	22993	15	3.5	47	1.6
Ung	NT_078458.2	+	584737	28122	5	0.0	51	2.4
Vps24	NT_039350.3	+	1497863	57721	6	1.5	41	0.8

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	Ns (%)	G+C content (%)	CpG content (%)
Ypel5	NT_039658.3	+	6378473	34970	17	1.6	42	1.3
Znhit1	NT_080526.2	C	2321160	22765	5	2.4	52	2.1

[a] NCBI build 34.1

mouse G3, UCSC mm8

gene symbol (synonym)	chr.	start	end	length (bp)	dir.	Ns (%)	G+C content (%)	CpG content (%)
1700016G05Rik	chr6	40434430	40459115	24685	C	0.0	39	0.4
2310076L09Rik	chr17	55694894	55720591	25697	C	0.0	56	1.9
2900092E17Rik	chr7	126796198	126818500	22302	C	0.0	55	2.7
4833421E05Rik (*Iah1*)	chr12	21552881	21580095	27214	+	0.0	48	1.5
5031400M07Rik (*Herpud2*)	chr9	24848539	24911918	63379	C	0.5	40	0.8
Abcd2	chr15	90963638	91029574	65936	C	0.0	39	0.9
Abcg5	chr17	84556559	84601249	44690	C	0.0	47	1.4
Afm	chr5	91584148	91638744	54596	+	0.0	42	1.1
Agr3	chr12	36425958	36469996	44038	+	0.0	39	0.4
Aldh4a1	chr4	138885085	138931757	46672	+	0.0	52	1.8
Aph1b	chr9	66564630	66604567	39937	C	0.0	42	0.8
Atad1	chr19	32728560	32788295	59735	C	0.0	40	0.9
Atrn	chr2	130587936	130731765	143829	+	0.0	42	0.7
Bach1	chr16	87578149	87632541	54392	+	0.0	47	1.9
BC030477 (*Wscd1*)	chr11	71556897	71615839	58942	+	0.0	49	1.3
Bnip1	chr17	26498813	26530248	31435	+	0.0	45	1.1
Car8	chr4	8058640	8176188	117548	C	0.0	42	0.9
Cd207	chr6	83626872	83653515	26643	C	0.0	47	0.7
Chd1	chr17	15399979	15485054	85075	+	0.0	40	1.2
Clca5	chr3	144997650	145046427	48777	C	0.0	42	1.1
Colec10	chr15	54230856	54306441	75585	+	0.0	38	0.5
Creb3l4	chr3	90313426	90339439	26013	C	0.0	47	1.5
Cxxc5	chr18	35945791	35997661	51870	+	0.0	53	1.8
D11Wsu99e	chr11	113467408	113510241	42833	C	0.0	48	1.7
D6Wsu116e	chr6	116163675	116238284	74609	+	0.0	44	1.2
Dmap1	chr4	117162618	117190157	27539	C	0.0	47	1.4
Dscr1l1 (*Rcan2*)	chr17	43254905	43512571	257666	+	0.0	43	0.9
Dus2l	chr8	108890635	108952948	62313	+	0.0	46	1.3
Edem1	chr6	108784417	108835133	50716	+	0.0	44	1.2
Eps15l1	chr8	75259987	75360449	100462	C	0.0	48	0.9
Fastk	chr5	23941104	23965300	24196	C	0.0	53	2.5
Grrp1	chr4	133513185	133536182	22997	C	0.0	53	2.3
Gtf3c1	chr7	125422102	125508836	86734	C	0.0	49	1.2
Hoxa6	chr6	52125947	52148207	22260	C	0.0	50	3.1
Hyal2	chr9	107417263	107440879	23616	+	0.0	51	2.0
Intu	chr3	40722315	40811289	88974	+	5.6	43	1.1
Kbtbd10	chr2	69460958	69495078	34120	+	0.0	43	1.3
Larp7	chr3	127518737	127555371	36634	C	0.0	41	1.1
Lrrc6	chr15	66199527	66340510	140983	C	0.0	41	0.6
Lsm1	chr8	27241135	27279520	38385	+	0.0	44	1.3
Mrpl24	chr3	87995470	88019360	23890	+	0.0	48	1.6
Mrpl40	chr16	18775780	18800200	24420	C	0.0	44	1.2
Mtnr1b	chr9	15603059	15634853	31794	C	0.0	44	0.8
Mtrr	chr13	69018150	69059492	41342	C	0.0	46	1.6

gene symbol (synonym)	chr.	start	end	length (bp)	dir.	Ns (%)	G+C content (%)	CpG content (%)
Ndufs2	chr1	173061534	173093787	32253	C	0.0	49	1.7
Ociad2	chr5	73591335	73627774	36439	C	0.0	45	1.3
Olfr160	chr9	37451019	37471948	20929	C	0.0	39	0.4
Parl	chr16	20183363	20225905	42542	C	0.0	43	1.0
Pde4d	chr13	109765040	111082352	1317312	+	0.0	40	0.8
Plcz1	chr6	139942127	140013823	71696	C	0.0	41	0.9
Pld3	chr7	27230778	27271872	41094	C	0.0	50	1.6
Polm	chr11	5717863	5747763	29900	C	0.0	48	1.2
Prmt3	chr7	49636378	49736286	99908	+	0.0	43	0.9
Prph1 (*Prph*)	chr15	98873240	98896703	23463	+	0.0	49	1.5
Psmb7	chr2	38400054	38475915	75861	C	0.0	44	0.9
Pus7l	chr15	94340819	94381670	40851	C	0.0	45	1.7
Rab21	chr10	114683972	114729701	45729	C	0.0	44	1.6
Rap1gds1	chr3	138853292	139022586	169294	C	0.0	40	1.0
Rbm22	chr18	60676154	60708098	31944	+	0.0	46	1.5
Reg1	chr6	78345491	78368175	22684	+	0.0	40	0.4
Rsad1	chr11	94345888	94375289	29401	C	0.0	48	1.5
Scn2b	chr9	44858870	44891065	32195	+	0.0	50	1.4
Sertad4	chr1	194535212	194566445	31233	C	0.0	47	1.6
Sesn1	chr10	41565863	41606838	40975	+	0.0	43	1.2
Sh2d5	chr4	137512486	137543044	30558	+	0.0	52	2.0
Slc2a3	chr6	122683444	122718138	34694	C	0.0	45	1.3
Smad9	chr3	54833510	54899186	65676	+	0.0	46	1.6
Tanc1	chr2	59402882	59656988	254106	+	0.0	46	1.4
Tgm6	chr2	129804725	129855666	50941	+	0.0	46	0.8
Tmem44	chr16	30421603	30480325	58722	C	0.0	48	1.6
Tpsab1	chr17	25060845	25083127	22282	C	0.0	49	1.1
Tubg1	chr11	100926220	100952507	26287	+	0.0	47	1.7
Tufm	chr7	126268579	126291878	23299	+	0.0	48	1.6
Txndc5	chr13	38497740	38545926	48186	C	0.0	48	1.5
Ube2v1	chr2	167288843	167333210	44367	C	0.0	52	1.8
Ubl3	chr5	148804979	148873137	68158	C	0.0	43	1.3
Vcp	chr4	42991063	43031534	40471	C	0.0	45	1.1
Xrcc5	chr1	72230727	72338155	107428	+	0.0	44	1.1
Zfp629	chr7	127388182	127415581	27399	C	0.0	51	2.1

Table A2: Overlap of CpG islands with selected repetitive elements

	human			mouse		
	SINE (%)	Alu (%)	LINE (%)	SINE (%)	B1 (%)	LINE (%)
GGF G1	54.34	53.07	5.24	11.23	3.95	2.58
GGF G2	53.50	51.82	4.98	11.47	6.67	1.40
GGF G3	57.98	57.56	5.47	12.54	6.19	5.84
TJ G1	9.29	9.29	0.71	1.03	0.00	3.09
TJ G2	12.85	12.29	1.68	2.54	0.00	0.00
TJ G3	12.98	11.45	5.34	0.00	0.00	3.96

Table A3: Ananlogous promoter CpG islands

group	method	both have promoter CGIs	both have no promoter CGIs	mouse has a promoter CGI, human not	human has a promoter CGI, mouse not	mouse promoter CGI longer than human one	mouse promoter CGI shorter than human one
G1 (76 pairs)	GGF	45 (59%)	17 (22%)	5	9	7	38
	TJ	40 (53%)	19 (25%)	5	12	6	34
	CPGed	40 (53%)	19 (25%)	4	13	6	34
	cpg	32 (42%)	20 (26%)	8	16	5	27
	Cluster	44 (58%)	17 (22%)	7	8	20	24
	GGF filter	38 (50%)	21 (28%)	6	11	4	34
G2 (79 pairs)	GGF	54 (68%)	11 (14%)	3	11	17	37
	TJ	50 (63%)	15 (19%)	3	11	19	31
	CPGed	51 (65%)	13 (16%)	3	12	13	38
	cpg	44 (56%)	16 (20%)	3	16	11	33
	Cluster	49 (62%)	13 (16%)	6	11	30	19
	GGF filter	42 (53%)	17 (22%)	8	12	11	31
G3 (79 pairs)	GGF	52 (66%)	16 (20%)	6	5	15	37
	TJ	45 (57%)	16 (20%)	9	9	12	33
	CPGed	52 (66%)	16 (20%)	5	6	12	40
	cpg	39 (49%)	19 (24%)	10	11	8	31
	Cluster	56 (71%)	17 (22%)	4	2	25	31
	GGF filter	43 (54%)	18 (23%)	12	6	9	25

GGF filter: CGIs identified with GGF criteria that do not depend on repetitive elements and that fulfill a ratio of (TpG+CpA)/(2*CpG) $\leq$ 1.0 for human and $\leq$ 1.2 for mouse. In group G1, the three mouse Riken sequences and their human counterparts were omitted from the analysis which is thus concentrated on truly orthologous gene pairs.

Table A4: Overlap of TJ CGIs with *cpg* CGIs

species/group					
human	**total**	**promoter**	**exonic**	**intronic**	**intergenic**
TJ G1	140	54	25	16	19
cpg G1	82%	91%	68%	69%	79%
TJ G2	179	61	29	29	27
cpg G2	77%	97%	62%	34%	89%
TJ G3	131	54	15	18	17
cpg G3	76%	91%	60%	28%	71%
mouse	**total**	**promoter**	**exonic**	**intronic**	**intergenic**
TJ G1	97	46	17	7	8
cpg G1	75%	87%	41%	29%	50%
TJ G2	118	53	12	10	19
cpg G2	84%	94%	67%	60%	74%
TJ G3	101	54	13	5	10
cpg G3	84%	89%	77%	60%	70%

The numbers refer to Takai and Jones (TJ) CpG islands in different locations. The percentages show their rate of overlap with CGIs identified by *cpg*.

Table A5: Overlap of *cpg* CGIs with TJ CGIs

species/group					
human	**total**	**promoter**	**exonic**	**intronic**	**intergenic**
cgp G1	142	50	18	22	28
TJ G1	73%	100%	78%	36%	46%
cpg G2	146	60	17	16	28
TJ G2	79%	98%	82%	56%	54%
cpg G3	129	50	12	17	10
TJ G3	73%	100%	67%	24%	60%
mouse	**total**	**promoter**	**exonic**	**intronic**	**intergenic**
cgp G1	88	41	11	3	11
TJ G1	78%	98%	55%	67%	36%
cpg G2	99	47	10	6	12
TJ G2	92%	100%	100%	67%	75%
cpg G3	91	49	8	6	5
TJ G3	90%	96%	88%	83%	60%

The numbers refer to CpG islands indentified with *cpg* in different locations. The percentages show their rate of overlap with Takai and Jones (TJ) CGIs. CpG-rich segments identified with *cpg* can be several 1000 bp long and thus comprise several TJ CGIs.

Table A6: Median values for CpG islands groups

	length (bp)	**G+C content (%)**	**CpG_{obs}/CpG_{exp}**	**CpG content (%)**

human	all	unique	prom.	all	unique	prom.	all	unique	prom.	all	unique	prom.
GGF G1	263	300	1422	53	59	65	0.61	0.61	0.78	4.4	5.4	7.7
GGF G2	207	272	1430	52	56	64	0.61	0.61	0.77	4.3	4.8	7.6
GGF G3	258	272	1163	52	58	65	0.61	0.62	0.77	4.3	5.4	7.8
TJ G1	1206	1368	1515	62	62	63	0.72	0.74	0.77	6.8	7.1	7.4
TJ G2	1011	1269	1481	61	62	64	0.69	0.72	0.77	6.7	7.0	6.8
TJ G3	1154	1328	1441	61	62	62	0.71	0.75	0.75	6.9	7.3	7.3
CPGed G1	285	417	2416	57	59	56	0.67	0.66	0.66	5.7	5.9	5.5
CPGed G2	281	363	2855	57	57	55	0.68	0.67	0.66	5.5	5.7	5.1
CPGed G3	283	738	2186	57	57	56	0.68	0.68	0.67	5.5	5.7	5.3
cpg G1	1167	1363	1523	64	64	65	0.76	0.77	0.79	7.5	7.9	8.7
cpg G2	1209	1360	1653	64	65	64	0.76	0.77	0.79	7.2	7.7	6.6
cpg G3	883	1057	1337	63	66	67	0.78	0.80	0.81	7.6	8.5	8.4
Cluster G1	511	675	1340	60	65	65	0.77	0.78	0.80	6.5	7.8	8.2
Cluster G2	448	540	1315	61	64	66	0.75	0.78	0.81	6.6	7.8	8.8
Cluster G3	494	714	1202	59	64	61	0.73	0.77	0.79	6.3	7.7	7.4
mouse	all	unique	prom.	all	unique	prom.	all	unique	prom.	all	unique	prom.
GGF G1	241	252	1009	53	54	62	0.61	0.61	0.79	4.3	4.5	7.1
GGF G2	205	243	1278	52	54	62	0.61	0.61	0.77	4.2	4.4	7.6
GGF G3	249	265	997	52	53	63	0.62	0.62	0.76	4.3	4.4	7.7
TJ G1	999	1053	1186	59	59	60	0.70	0.74	0.76	6.5	6.6	6.7
TJ G2	1080	1117	1369	60	60	61	0.73	0.74	0.76	6.6	6.6	7.3
TJ G3	1095	1144	1230	60	60	61	0.70	0.71	0.74	6.4	6.6	6.8
CPGed G1	365	371	1590	57	57	56	0.67	0.67	0.67	5.5	5.5	5.6
CPGed G2	430	430	2080	56	57	57	0.67	0.70	0.67	5.5	5.5	5.1
CPGed G3	477	477	1537	56	56	56	0.67	0.67	0.67	5.5	5.5	5.4
cpg G1	847	850	1000	64	64	65	0.81	0.82	0.89	7.8	8.1	9.0
cpg G2	995	1053	1195	64	65	66	0.81	0.81	0.83	8.2	8.3	7.1
cpg G3	840	867	902	65	64	65	0.81	0.82	0.83	8.2	8.3	8.8
Cluster G1	640	720	1171	58	59	60	0.72	0.68	0.76	5.9	5.8	6.7
Cluster G2	695	759	1322	59	60	61	0.76	0.75	0.79	6.5	6.5	7.3
Cluster G3	671	791	1204	58	58	61	0.74	0.73	0.77	6.2	6.2	6.9

prom. = promoter

Table A7: Overlap of filtered unique GGF CGIs with TJ CGIs

human	total	promoter	exonic	intronic	intergenic
GGF G1, 1.0	114	50	20	6	14
TJ G1	85%	100%	70%	50%	64%
GGF G2, 1.0	122	54	17	7	14
TJ G2	87%	100%	65%	71%	93%
GGF G3, 1.0	101	49	12	5	11
TJ G3	91%	100%	83%	40%	73%
mouse	**total**	**promoter**	**exonic**	**intronic**	**intergenic**
GGF G1, 1.2	105	45	15	6	15
TJ G1	73%	96%	60%	50%	27%
GGF G2, 1.2	112	50	11	9	17
TJ G2	85%	100%	64%	67%	65%
GGF G3, 1.2	99	51	14	4	7
TJ G3	82%	98%	43%	25%	100%

The numbers refer to unique GGF CGIs that do not depend on repetitive elements and that fulfill a ratio of (TpG+CpA)/(2·CpG) ≤ 1.0 for human and ≤ 1.2 for mouse. The percentages show their rate of overlap with CGIs identified with Takai and Jones (TJ) criteria.

Table A8: Overlap of TJ CGIs with filtered unique GGF CGIs

human	total	promoter	exonic	intronic	intergenic
TJ G1	114	50	20	6	19
GGF G1, 1.0	69%	96%	52%	25%	47%
TJ G2	122	54	17	7	27
GGF G2, 1.0	58%	90%	38%	14%	44%
TJ G3	131	54	15	18	17
GGF G3, 1.0	72%	91%	67%	11%	53%
mouse	**total**	**promoter**	**exonic**	**intronic**	**intergenic**
TJ G1	105	45	15	6	8
GGF G1, 1.2	78%	93%	59%	29%	38%
TJ G2	112	50	11	9	19
GGF G2, 1.2	81%	94%	50%	60%	63%
TJ G3	101	54	13	5	10
GGF G3, 1.2	79%	94%	46%	20%	50%

The numbers refer to TJ CGIs. The percentages show their rate of overlap with unique GGF CGIs that do not depend on repetitive elements and that fulfill a ratio of (TpG+CpA)/(2·CpG) ≤ 1.0 for human and ≤ 1.2 for mouse.

Appendix B

Table B1: Locations and data of imprinted genes
The abbreviation dir. stands for the direction of transcription (+: on forward strand, C: on reverse complement). Ns in the mouse represent stretches of undefined nucleotides. All data refer to the genomic sequences of the genes with 10 kb of upstream and downstream sequences each.

human imprinted genes, NCBI build 35.1

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	G+C content (%)	CpG content (%)
ASB4	NT_007933.14	+	20339560	71788	7	38	0.7
ASCL2	NT_009237.17	C	1066970	22454	11	56	3.5
ATP10A	NT_026446.13	C	2347219	206451	15	46	1.6
CD81	NT_009237.17	+	1175981	39888	11	60	3.1
CDKN1C (*P57*)	NT_009237.17	C	1682117	22100	11	57	3.3
COPG2	NT_007933.14	C	55319658	90791	7	40	1.0
DIO3	NT_026437.11	+	83017441	22098	14	57	3.5
DLK1	NT_026437.11	+	82183006	28208	14	54	2.6
DLX5	NT_007933.14	C	21873988	24432	7	47	2.7
GATM	NT_010194.16	C	16433879	37388	15	40	1.2
GNAS	NT_011362.9	+	22457691	91453	20	46	2.2
GRB10 (*MEG1*)	NT_033968.5	C	239414	210220	7	45	1.3
H19	NT_009237.17	C	794154	25788	11	61	3.6
HTR2A	NT_024524.13	C	28377514	82663	13	39	0.7
IGF2	NT_009237.17	C	931197	26047	11	60	3.7
KCNQ1	NT_009237.17	+	1243462	424120	11	52	1.8
MAGEL2	NT_026446.13	C	313494	22295	15	46	2.0
MEG3 (GTL2)	NT_026437.11	+	82283462	21935	14	54	2.7
MEG8[a]	NT_026437.11	+	82350958	61132	14	41	1.1
MEST	NT_007933.14	+	55299622	40088	7	42	1.3
NAP1L5	NT_016354.17	C	14101771	21914	4	40	0.9
NDN	NT_026446.13	C	355353	21897	15	43	1.4
NNAT	NT_011362.9	+	1192522	22486	20	50	2.1
PEG10	NT_007933.14	+	19510603	32678	7	40	1.3
PEG3	NT_011109.15	C	29581993	48262	19	43	1.3
PHLDA2	NT_009237.17	C	1726751	21141	11	59	3.3
PLAGL1 (*ZAC1*)	NT_025741.13	C	48355868	88115	6	40	0.9
RASGRF1	NT_010194.16	C	50034606	148555	15	49	1.3
SGCE	NT_007933.14	C	19438818	90937	7	37	0.8
SLC22A18	NT_009237.17	+	1698192	45526	11	61	2.9
SLC22A2	NT_007422.12	C	2915135	62166	6	42	0.8
SLC22A3	NT_007422.12	+	3046766	126590	6	42	1.0
SLC38A4	NT_029419.10	C	9291850	81196	12	37	0.7
SNRPN	NT_026446.13	+	1493593	615816	15	42	1.0
UBE3A	NT_026446.13	C	2007195	121733	15	37	0.8
USP29	NT_011109.15	+	29898201	23284	19	45	1.6
WT1	NT_009237.17	C	31186566	67753	11	47	1.9
ZNF264	NT_011109.15	+	29961077	47639	19	45	1.3

[a] sequence taken from Ensembl release 30 (Vega gene OTTHUMG00000029055)

mouse imprinted genes, NCBI builds 33.1 and 34.1

gene symbol (synonym)	contig, version	dir.	start	length (bp)	chr.	Ns (%)	G+C content (%)	CpG content (%)
Asb4	NT_039340.3	+	2105836	62470	6	0.2	42	1.0
Ascl2	NT_039437.3	C	1213793	21072	7	0	49	1.6
Atp10a	NT_039424.3	+	7816525	190579	7	0.3	42	0.7
Cd81	NT_039437.3	+	1298670	35130	7	0	51	1.2
Cdkn1c	NT_039437.3	C	1704215	22657	7	0	52	2.3
Commd1 (*Murr1*)	NT_039515.3	C	19785331	99233	11	0	43	0.9
Copg2[a]	NT_039353.3	+	2083716	45500	6	0.2	47	1.2
Dio3	NT_039553.3	+	6423115	21863	12	0	53	2.4
Dlk1	NT_039553.3	+	5600361	27113	12	0	51	1.9
Dlx5[b]	NT_039340.4	C	3840563	24264	6	0	46	2.0
Gatm	NT_039207.3	C	63274218	36788	2	0	43	1.1
Grb10[c] (*Meg1*)	NT_039515.4	C	8815301	126892	11	0	47	2.0
Gtl2 (Meg3)	NT_039553.3	+	5692251	34979	12	0	43	1.0
H19	NT_039437.3	C	821407	22615	7	0	50	1.6
Htr2a	NT_039606.3	+	20763944	85855	14	0	52	1.4
Igf2	NT_039437.3	C	896643	31093	7	1.6	42	0.9
Igf2r	NT_039638.3	C	5193580	107465	17	0	51	1.9
Impact	NT_039674.3	+	10328112	40697	18	0.7	48	1.3
Ins2	NT_039437.3	C	924536	21048	7	0	42	1.1
Kcnq1	NT_039437.3	+	1353323	339595	7	0	47	0.6
Magel2	NT_039428.3	+	2340325	22177	7	0	47	0.8
Nap1l5	NT_039343.3	C	10625880	21833	6	0	42	1.1
Ndn	NT_039428.3	+	2309209	21581	7	1.4	41	1.1
Nnat	NT_039210.3	+	7947206	22382	2	0	38	0.8
Peg10[b]	NT_039340.4	+	1690266	33097	6	0	50	1.9
Peg12 (*Frat3*)	NT_039428.3	C	2422732	22640	7	0	43	1.5
Peg3	NT_095092.1	+	12921	24146	7	0	41	1.1
Phlda2	NT_039437.3	C	1747425	20961	7	0	44	1.1
Plagl1 (*Zac1*)	NT_039491.3	+	5544235	58980	10	0	52	1.5
Rasgrf1	NT_039476.3	+	653914	137152	9	0	43	1.1
Sgce[b]	NT_039340.4	C	1617236	92750	6	0	46	1.0
Slc22a18	NT_039437.3	+	1719664	45534	7	0	41	1.3
Slc22a2[b]	NT_039638.4	+	4270125	64296	17	0	52	1.4
Slc22a3	NT_039638.3	C	4930247	107728	17	0	45	0.9
Slc38a4	NT_039621.3	C	58553511	81218	15	0	45	1.1
Snrpn	NT_039428.3	C	71251	42104	7	0.2	44	1.6
Ube3a	NT_039424.3	+	8389056	96135	7	0.2	36	0.5
Wt1	NT_039207.3	+	45950064	67084	2	5.0	34	0.5

[a] sequence taken from Ensembl release 30
[b] NCBI build 34.1
[c] sequence taken from Ensembl release 32

Table B2: Sequence portions covered by CpG islands

group	GGF (%)	GGF_mask (%)	TJ (%)	TJ_mask (%)
human imprinted	7.68	5.60	3.74	3.32
human G1	7.31	3.49	2.94	2.63
human G2	8.27	4.43	3.10	2.63
human all	7.79	4.31	3.17	2.77
mouse imprinted	5.19	4.80	3.10	2.81
mouse G1	4.15	3.44	1.75	1.29
mouse G2	4.65	3.83	2.39	2.22
mouse all	4.54	3.84	2.25	1.93

The CpG island lengths were summed up per sequence and divided by the accumulated sequence lengths of the respective group.

Table B3: Total numbers of tandem repeat arrays

group	sequences	GGF	GGF_mask	TJ	TJ_mask
human imprinted	38	35	28	26	16
human G1	79	59	31	15	4
human G2	79	58	20	25	6
mouse imprinted	39	29	18	13	8
mouse G1	79	32	12	5	1
mouse G2	79	32	7	7	2

Table B4: Distribution of tandem repeat arrays in CGIs identified in repeat masked sequences with GGF criteria

group	tandem repeat arrays	genes with tandem repeats	minimum motif length (bp)	maximum motif length (bp)	mean motif length (bp)	minimum no. of repetitions	maximum no. of repetitions	mean no. of repetitions	minimum array length (bp)	maximum array length (bp)	mean array length (bp)
human imprinted	24	14	13	84	34.17	1.9	20.5	7.25	57	596	221.63
human G1	30	10	16	92	45.37	2	50.5	8.05	60	1618	320.1
human G2	21	9	17	79	39.71	2	45.4	6.61	91	790	207.48
mouse imprinted	16	14	10	140	35.63	2	32.1	7.06	65	367	177.75
mouse G1	12	9	16	69	28.33	1.9	16.9	5.53	50	397	141.42
mouse G2	7	7	19	48	35.14	2.7	7.5	4.17	55	206	136.4

Table B5: Repeat arrays in imprinted genes according to the literature

human

gene symbol (synonym)	not found	motif length (bp)	no. of repeti-tions	consensus sequence	reference
CD81	c	31	10.1	GTCTCCCTCAGCCCCCACCCCCAGGGTCCACA	Paulsen et al. 2000
CD81		17	15.9	ACCCCACAAGCCGTCCC	
CD81		14	16.6	ACAGACGACGGGCA	
GRB10	s	5-7	5	ACCGCCC	Hikichi et al. 2003
H19	i	45-48	7	GGTTGTAgyTGTGGAATCgGAAGTGGCCGCGCGGCGGCAGTGCAGGCT	Bell and Felsenfeld 2000
H19		31	19	AYaGYgCYgTaCCCgYgTCYCtAYCCgGGTG	Lewis et al. 2004
IGF2R	*	8	4	CCCTNGNG	Smrzka et al. 1995
IGF2R		38	3	CCCCCTCGCGCCTCCCTGTACCCTGCATGCCCCGTGTG	
IGF2R		26	3	GTGCGCCTGCTGCGCCCCACGCGCCT	
IGF2R		18	2	GGAGCTGTCCAGGCGCGG	
INS	*	14	40-157	ACAGGGGTGTGGGG	Kennedy et al. 1995
KCNQ1		8	6	TCCGAGTY	Paulsen et al. 2005
KCNQ1		12	6	YGYGGTTCYGAG	
KCNQ1		18	3	YGYGGTTCYGAGTYGGGG	

gene symbol (synonym)	not found	motif length (bp)	no. of repeti-tions	consensus sequence	reference
KCNQ1		12	17	CGYGGTTCYCCC	
KCNQ1		22	18	TCCTCnGCGTGGTTCTCCTCGG	
KCNQ1		27-32	4	YYCTCnGCgYGGTYCTCCTCGGyGygyyYc	Mancini-DiNardo et al. 2003
KCNQ1	c	72	3	CACCAGGAACNCCAGCTTGGGCCAGAGGGCGTCCCACGCCAGGAACCCNAGCNTGGG TCAGAGGGGTCCCA	Paulsen et al. 2000
KCNQ1	c	18	28.2	CCCCCAGGATGGACGTCA	
MAGEL2		21	9.5	AGGCCCCACCKGTGATCCGCC	Boccaccio et al. 1999
MEG3 (*GTL2*)	r	18	9	GTTGCCYGYGGCTCACCA	Paulsen et al. 2001
NESP55	s	12	11	GAGACCGAGccC	Coombes et al. 2003
PEG3		11	10	GGCGCCATCTT	Kim et al. 2003
PHLDA2-NAP1L4	c	15	~70	TATTCACACYRAGCR	Engemann et al. 2000
SNRPN		21-43	nd	nd	Huq et al. 1997
TSSC6-ASCL2	r	56	20.4	CGGGTGGCACGCCTCTGCGAATATACTAAAGCGGGGAGTTGTTTTTGGGGGTGCTG	Paulsen et al. 2000
XLalphaS	c	60	4	TGaCCAgCCaGGCCTGGGAGGcTtCnGnCCWcCACTcgwRSAGSCYggAgcCYTYAgTgg	Coombes et al. 2003
XLalphaS	s	21	7	CRGCCCCCCnRTCRAGATnGA	
XLalphaS		27	4	TCCGGGGCRGCCCCAGCCGATCCCGAC	
XLalphaS		36	6	TCCGGGGCRGCCCCTGACGCCCCAGCCGATCCCGAC	

mouse

gene symbol (synonym)	not found	motif length (bp)	no. of repeti-tions	consensus sequence	reference
Cdkn1c (*p57*)	m	12	15	GCMGGgCGAGGA	Hatada et al. 1995
Gnasxl		36	6	GCSGAGCCnGCCnCCGGGGCAGTCCCTGYCACCCYn	Coombes et al. 2003
Gnasxl	s	18	5	GCCCGSGCAGCCyCTGCY	
Grb10		10	12	GGCGCGTGYT	Hikichi et al. 2003
Grb10	c	27-28	9.2	GCCCATCACCTCCCCCATCCTCACCCCA	Arnaud et al. 2003
Grb10	r	19-20	3	CGCGGCAACACGCGCCAACA	
Grb10	r	15-20	3	CAACACAGGCCGGCACGCGC	
Grb10	r	21	3.6	AACGCGAGCCCGGCACGCGCC	

gene symbol (synonym)	not found	motif length (bp)	no. of repeti-tions	consensus sequence	reference
Grb10	r	9-10	10	GCCGACACGC	
Gtl2 (*Meg3*)	r	24	7	TAGTGCCGCGGtTCGCCgTGgACT	Paulsen et al. 2001
Gtl2 (*Meg3*)	r	43	11	CTACGGTATAaGCCAAGTGyYtcGCgGCACAGnTAygTGgTA	
H19	i	45-48	4/5	GyTgCCGCGyGGyGGCAGYAaYnT	Bell and Felsenfeld 2000, Hark et al. 2000
H19		8-9	32	(G)GGGGTATA	Reed et al. 2001, Lewis et al. 2004
H19	c	30	22	AcACYYCTGTGYCCATgTCCYATcYatGTG	Lewis et al. 2004
Igf2	c	11	35	AGGCCTGAGCC	Sasaki et al. 1996, Moore et al. 1997
Igf2r	i	8	10	CCCTNGNG	Smrzka et al. 1995
Igf2r		30-32	3	TCTCCTGCAACGTGGCACTTTTGAGCTTnn	Reinhard et al. 2002
Igf2r	i	172-180	3	GAACCCTccGAAtCCtccCCTTGTGCAGCTTtgCACCCTcAGgaTayCTCGgAAcCtccgagcYytcyTtcCcytyCccTcgcNgyaNttcNNaaaacCyNagNaycagggcaNNNggggNNgygctNCgaaCccyCgAgCaycyygGCNgcgcNgyNcCgggGNaccCyNc	
Impact		57	5	GCACTAGCTTTGCCGCATTGTCACATGAGCAGGCCCGGCCCACTCGGCYnGGCTcGG	Okamura et al. 2000
Impact	m	60-93	6	TCGGC-rich	
Kcnq1	c	nd	nd	poly-A	Engemann et al. 2000
Kcnq1	i	8	7	TCCGAGTY	Paulsen et al. 2005
Kcnq1	i	12	4	YGYGGTTCYGAG	
Kcnq1	i	18	1	YGYGGTTCYGAGTYGGGG	
Kcnq1	i	12	1	CGYGGTTCYCCC	
Kcnq1	i	22	10	TCCTCnGCGTGGTTCTCCTCGG	
Kcnq1	i	27-32	5	YYCTCnGCgYGGTYCTCCTCGGyGygyyYc	Mancini-DiNardo et al. 2003
Magel2	c	18	11.5	GGTGCCACAGGAGCTCCC	Boccaccio et al. 1999
Mest (*Peg1*)	c	5	24	WGGGG	Lefebvre et al. 1997
Nesp	s	12	10	GAGaCCGAGCCn	Coombes et al. 2003
Peg3	i	11	14	GGCGCCATCTT	Kim et al. 2003
Snrpn	i	28-31	5/8	TGGgCTCCaGGATGCngGAGCTCTGTTGcCGCAGCCTGyGGGCT	Reinhart et al. 2002
Rasgrf1	r	42	41	CTGCCCCTGCCCCAGCCGCTACTGCTGCCCCTGCCCCnCCA	Pearsall et al. 1999
U2afbp-rs (in *Commd1*)		nd	nd	nd	Pearsall et al. 1996

Consensus sequences are given. At less conserved positions the following IUPAC symbols were used: K: G or T; M: A or C; R: A or G; S: G or C; W: A or T; Y: C or T (see also Tab. 2.6). Lower cases indicate a weak preference for the nucleotide at the respective position. nd: not defined
The *NESP55* and *XLαS* genes are included in our human *GNAS* sequence; *Nesp* and *Gnasxl* genes are included in mouse *Gnas*.
Some repeats could not be identified in this study. Reasons for this are marked as follows: (r) sequence outside of range, (c) not in a CpG island, (m) microsatellite repeat, (s) score under threshold, (i) not a direct tandem repeat.
* *IGF2R* (Smrzka et al. 1995) was excluded from the set because there is strong support that, although the tandem repeat region is differentially methylated, the gene is not imprinted in humans (Riesewijk et al. 1996, Vu et al. 2004). *INS* (Kennedy et al. 1995) was excluded as well since the length of the minisatellite repeat regulates transcription by binding PUR-1. Thus, different levels of expression from paternal and maternal alleles may not be related to imprinting effects but result from different numbers of repetitions.

Appendix C

Table C1: Retrieval of orthologs of imprinted genes and control genes

species	imprinted (61 genes)	G1 (78 genes)	G2 (78 genes)	G3 (78 genes)
cow (bosTau3)	40 + 5 + 12	65 + 1 + 7	66 + 4 + 8	67 + 4
dog (canFam2)	48[a] + 0 + 12	72 + 0 + 2	73 + 0 + 3	68 + 1
opossum (monDom5)	25 + 4 + 7	64 + 1 + 6	51+ 2 + 5	60 + 2 + 7
platypus (ornAna1)	7 + 24 + 17	9 + 46 + 10	10 + 41 + 16	9 + 36 + 25
cow and dog[b]	38 (62%; 66%[c])	61 (78%)	65 (83%)	58 (74%)
all four species[d]	22 (36%; 38%[c])	41 (53%)	40 (51%)	35 (45%)

Numbers indicate: Ensembl Biomart annotations confirmed by *Blast* with genes placed on chromosomes + unplaced genes + recovered by *Blast* only. *Blast* hits were filtered with respect to synteny.
[a] Hits for *ZIM2* and *PEG3* are identical
[b] Ensembl Biomart annotations confirmed by *Blast* with genes placed on chromosomes
[c] without *H19*, *MEG3*, *MEG8*
[d] Ensembl Biomart annotations confirmed by *Blast* with genes placed on chromosomes or, in case of platypus, contigs

Table C2: Properties of *phastCons* sequences for the human genome

hg18	conservation score		length (bp)		G+C content (%)		CpG content (%)		CpGobs/CpGexp		repeat overlap (%)	
	imprinted	genome	imprinted	genome	imprinted	genome	imprinted	genome	imprinted	genome	imprinted	genome
mean	365.5	381.89	67.62	75.98	43.64	40.94	1.69	1.18	0.25	0.2	8.25	5.73
std.dev.	101.7	114.44	72.91	85.03	14.46	13.43	3.05	2.49	0.39	0.38	25.93	2.84
minimum	235	189	20	20	3.85	0	0	0	0	0	0	0
lower quart.	288	288	28	29	33.33	31.71	0	0	0	0	0	0
median	342	353	44	46	41.67	39.07	0	0	0	0	0	0
upper quart.	427	452	81.5	91	53.44	49.09	2.44	1.43	0.43	0.31	0	0
maximum	842	999	1236	3895	95.24	100	25.81	45	7.33	23	100	100

quart. = quartile (lower quart. = 25th percentile, upper quart. = 75th percentile)

Table C3: Properties of *phastCons* sequences for the mouse genome

mm9	conservation score		length (bp)		G+C content (%)		CpG content (%)		CpGobs/CpGexp		repeat overlap (%)	
	imprinted	genome	imprinted	genome	imprinted	genome	imprinted	genome	imprinted	genome	imprinted	genome
mean	363	381.11	70.51	77.01	44.47	42.67	1.71	1.34	0.28	0.24	6.49	4.6
std.dev.	104.1	116.69	77.92	86.63	13.03	12.47	2.76	2.4	0.43	0.4	22.54	19.24
minimum	228	196	20	20	3.45	0	0	0	0	0	0	0
lower quart.	282	288	29	29	35.29	34.09	0	0	0	0	0	0
median	337	355	45	47	43.48	41.89	0	0	0	0	0	0
upper quart.	432	453	83	92	53.22	50.84	2.6	2.01	0.47	0.39	0	0
maximum	854	999	1268	4733	91.89	100	20.59	38.1	4.5	28	100	100

quart. = quartile (lower quart. = 25th percentile, upper quart. = 75th percentile)

Table C4: Properties of human *phastCons* sequences at different locations

	intron imprinted (1120)	intron genome (365,258)	promoter imprinted (16)	promoter genome (5337)	intergenic imprinted (1787)	intergenic genome (588,309)	exon imprinted (1015)	exon genome (306,508)
conservation score								
mean	346.61	359.51	482.94	479.1	348.32	366.59	413.91	436.06
std.dev.	94.27	108.73	133.32	127.18	92.51	110.52	106.84	110.1
minimum	235	235	310	235	235	189	235	235
lower quart.	275	275	391.5	379	275	281	329	349
median	320	329	462.5	474	325	338	396	429
upper quart.	398	412	589.5	568	398	423	493	517
maximum	764	937	745	895	693	967	842	999
length (bp)								
mean	60.81	66.84	207.75	175.21	60.63	72.22	84.03	92.3
std.dev.	60.61	76.31	174.86	191.02	60.87	81.45	89.31	94.4
minimum	20	20	21	20	20	20	20	20
lower quart.	27	27	63.5	54	28	28	32	36
median	40	40	164.5	114	40	43	60	67
upper quart.	70	73	324.5	220	71	82	106	118
maximum	706	1977	600	2101	905	2402	1236	3895
G+C content (%)								
mean	41.37	38.06	66.77	66.18	40.99	37.88	50.06	49.55
std.dev.	14.64	12.29	14.11	12.98	13.61	12.18	13.09	12.55
minimum	3.85	0	33.33	20	7.14	0	11.43	0
lower quart.	30.86	30	60.39	58.33	31.82	30	40.695	40.28
median	38.89	36.54	68.62	68.52	39.06	36.36	50	50
upper quart.	50	44.72	76.295	75.86	49.41	44.12	59.21	58.97
maximum	95.24	100	85.41	100	92.68	100	89.58	100
CpG content (%)								
mean	1.31	0.69	9.31	9.21	1.19	0.76	2.77	2.37
std.dev.	2.83	1.84	5.51	5.83	2.57	2.05	3.36	2.98
minimum	0	0	0	0	0	0	0	0
lower quart.	0	0	4.78	4.11	0	0	0	0

	intron imprinted (1120)	intron genome (365,258)	promoter imprinted (16)	promoter genome (5337)	intergenic imprinted (1787)	intergenic genome (588,309)	exon imprinted (1015)	exon genome (306,508)
median	0	0	8.705	9.57	0	0	1.75	1.5
upper quart.	1.37	0.34	14.36	13.575	1.54	0.56	4.17	3.57
maximum	18.85	32.35	18.75	30.95	25.81	45	20.69	32.56
CpGobs/CpGexp								
mean	0.19	0.14	0.83	0.77	0.2	0.15	0.36	0.33
std.dev.	0.35	0.36	0.36	0.39	0.41	0.37	0.36	0.35
minimum	0	0	0	0	0	0	0	0
lower quart.	0	0	0.65	0.53	0	0	0	0
median	0	0	0.785	0.86	0	0	0.32	0.28
upper quart.	0.3	0.1	0.985	1.03	0.33	0.16	0.6	0.52
maximum	2.1	21	1.61	3.12	7.33	23	2.07	12
repeat overlap (%)								
mean	10.72	7.4	12.71	10.68	10.2	7.15	1.96	0.98
std.dev.	29.36	24.89	22.33	19.67	28.87	24.38	11.26	7.58
minimum	0	0	0	0	0	0	0	0
lower quart.	0	0	0	0	0	0	0	0
median	0	0	0	0	0	0	0	0
upper quart.	0	0	14.245	15.725	0	0	0	0
maximum	100	100	74.11	100	100	100	100	100

quart. = quartile (lower quart. = 25th percentile, upper quart. = 75th percentile)

For each location, the numbers of *phastCons* sequences (PCSs) are given in parentheses. exon: overlapping by at least 1 bp with the coding region of a gene. PCSs in coding regions that also overlap with the most upstream TSS of a gene are excluded from the "exon" class in favor of the "promoter" category. There are too few PCSs in 5' UTR and 3' UTR regions of imprinted genes for analyses.

Just as all PCSs, those in introns of imprinted genes have a lower conservation score than those in introns of autosomal genes ($p < 0.005$), and reduced length ($p < 0.005$). The same holds for intergenic PCSs (score $p < 0.0001$, length $p < 0.0005$). In contrast, G+C and CpG contents as well as the repeat overlap is elevated whereas the estimated deamination ratio is reduced. PCSs in coding exons are shorter and less conserved ($p < 0.0005$); PCSs in promoter regions behave similarly for imprinted genes and autosomal ones.

Appendix D

Table D1: HomoloGene data for additional orthologous gene pairs

pairs	group	genes[a]	protein identity (%)	cDNA identity (%)	Ka/Ks	Ks
human-chimpanzee	imprinted	45/43	97.5±5.39	98.0±4.2	0.284±0.331	0.037±0.068
	genome	15,848/14,661	98.4±3.5	98.7±2.7	0.318±0.374	0.028±0.090
human-cow	imprinted	40	86.2±10.6	86.3±6.6	0.165±0.130	0.484±0.253
	genome	14,647/14,635	87.9±10.3	87.7±5.8	0.152±0.131	0.439±0.187
human-dog	imprinted	48	90.6±7.9	89.4±4.9	0.126±0.102	0.384±0.128
	genome	14,933/14,924	89.2±9.4	88.6±5.4	0.142±0.123	0.408±0.172
human-chicken	imprinted	40/31	74.8±12.5	72.9±7.9	0.125±0.096	1.470±0.500
	genome	11,201/9613	75.2±15.2	73.4±8.8	0.116±0.098	1.650±0.737
mouse-cow	imprinted	38	80.6±12.2**	80.6±7.5**	0.154±0.118*	0.820±0.383
	genome	14,682/14,648	84.2±12.3	82.8±6.9	0.127±0.106	0.740±0.256

[a] The second number refers to sequences available in the HomoloGene database for Ks and Ka/Ks analyses. For 100% identical sequences as well as for those with Ks reported as -1, Ka/Ks is not defined. Due to saturation, Ks becomes very high for human-chicken sequences. Average and standard deviation are given. p values (Wilcoxon test) refer to comparison of genome-wide data with the respective imprinted group: * $p<0.1$, ** $p<0.05$

Table D2: Silent CpG substitutions

group	mean±std.dev.	mimimum	lower quartile	median	upper quartile	maximum	p value
human-mouse imprinted	2.83±2.74	0.22	0.80	1.70	3.79	11.00	0.06420
human-mouse genome	3.92±4.30	0.00	1.43	2.60	4.75	85.00	
human-rat imprinted	2.71±2.40	0.39	0.99	2.00	4.07	10.25	0.04807
human-rat genome	4.04±4.40	0.00	1.5	2.71	5.00	61.00	
mouse-rat imprinted	0.83±0.77	0.04	0.275	0.56	1.24	3.00	0.00077
mouse-rat genome	1.37±1.34	0.00	0.59	1.00	1.67	20.00	
human-cow imprinted	2.04±2.88	0.15	0.51	1.00	2.27	15.00	0.01635
human-cow genome	2.67±3.23	0.00	0.92	1.67	3.12	58.00	

Silent CpG substitutions are calculated from pairwise alignments as the ratio of CpG pairs divided by CpG deamination-related mismatches at silent positions.

Table D3: Conservation and existence of orthologs and paralogs of protein-coding imprinted genes

imprinted gene (synonym)	human-mouse protein ID (%)	human-mouse DNA ID (%)	human-mouse Ka/Ks	opossum ortholog	platypus ortholog	chicken ortholog	zebrafish ortholog	number of human paralogs	youngest human paralog	gene-paralog protein ID (%)	paralog ancestor
USP29	48.2	64.6	0.384	NO	NO	NO	NO	2	*USP26*	46	Eutheria
TSPAN32 (*PHEMX*, *TCCS6*)	64.2	75.1	0.363	NO	NO	NO	NO				
MAGEL2	64.4	76.4	0.362	NO	NO	NO	NO				
CDKN1C (*P57*)	64.7	74.0	0.465	NO	NO	NO	chr7				
PEG3	64.8	71.5	0.220	NO	NO	NO	NO				
TSSC4	65.9	72.4	0.207	unplaced	NO	chr5	chr7				
KLF14	66.3	74.3	0.298	NO	NO	NO	NO	7	*KLF16*	48	Eutheria
MKRN3	69.4	77.9	0.340	NO	NO	NO	NO	2	*MKRN1*	47	Eutheria
PHLDA2	71.1	78.1	0.390	NO	Contig17670	chr5	NO	2	*PHLDA1*	40	Euteleostomi
PEG10	71.1	76.0	0.302	NO	NO	NO	NO				
KCNK9	73.5	73.8	0.200	chr3	chr4	chr2	chr19	2	*KCNK3*	61	Euteleostomi
ASCL2	74.1	77.7	0.327	NO	NO	NO	chr7	1	*ASCL1*	45	Euteleostomi
CALCR	78.3	79.7	0.161	NO	NO	chr2	chr19	12	*CALCRL*	51	Euteleostomi
SLC22A18	78.9	78.6	0.205	NO	NO	chr5	chr7				
DCN	81.6	80.7	0.149	chr8	Ultra443	chr1	chr4	2	*BGN*	54	Euteleostomi
INS	81.8	82.4	0.156	not assembled	not assembled	chr5	chr14	1	(*Ins1*)		Murinae
IGF2R	81.8	80.5	0.113	chr2	Ultra61	chr3	chr20				
PON1	82.0	83.1	0.155	unplaced	NO	chr2	chr16	2	*PON2*	65	Theria
ATP10A	82.2	81.5	0.136	chr7	Ultra292	chr1	chr6				
IMPACT	83.0	82.5	0.107	chr3	Contig20839	chr2	chr22	1	*AC020937.6*	43	Euarchontoglires
NDN	83.2	85.0	0.194	NO	NO	NO	NO				
SLC22A2	83.8	83.1	0.117	not assembled	unplaced	chr3	chr20	2	*SLC22A1*	69	Mammalia
IGF2	83.9	86.9	0.221	not assembled	not assembled	chr5	chr7				
NAP1L5	84.0	80.2	0.068	NO	NO	NO	NO				
CPA4	84.0	85.1	0.141	chr8	10	chr14	chr25	8	*CPA2*	63	Mammalia
L3MBTL	84.9	85.3	0.155	chr1	Ultra337	chr20	chr23	9	*L3MBTL4*	34	Euteleostomi

imprinted gene (synonym)	human-mouse protein ID (%)	human-mouse DNA ID (%)	human-mouse Ka/Ks	opossum ortholog	platypus ortholog	chicken ortholog	zebrafish ortholog	number of human paralogs	youngest human paralog	gene-paralog protein ID (%)	paralog ancestor
RASGRF1	85.0	84.4	0.125	chr1	Contig26251	chr10	chr13	5	*RASGRF2*	64	Bilateria
OSBPL5	85.3	83.2	0.102	chr5	NO	chr5	NO	1	*OSBPL8*	54	Euteleostomi
TRPM5	85.9	83.1	0.110	NO	NO	chr5	chr7	1	*TRPM4*	43	Euteleostomi
DLK1	86.2	87.0	0.188	chr1	Ultra378	chr5	NO				
COMMD1	86.4	85.9	0.118	chr1	Ultra56	chr3	chr1				
GRB10 (*MEG1*)	86.5	81.4	0.070	chr6	chr4	chr2	chr19	1	*GRB14*	49	Euteleostomi
SLC38A4	87.0	82.5	0.070	chr8	chr2	chr1	chr4	5	*SLC38A2*	57	Euteleostomi
SLC22A3	87.6	85.1	0.129	chr2	Contig2038	chr3	NO	2	*SLC22A2*	48	Euteleostomi
BEGAIN (*KIAA1446*)	87.6	86.7	0.135	chr1	Ultra378	chr5	chr20				
KCNQ1	89.8	85.2	0.085	unplaced	NO	chr5	chr7	4	*KCNQ4*	34	Bilateria
PPP1R9A	90.6	86.9	0.068	NO	NO	chr2	chr19	1	*PPP1R9B*	34	Euteleostomi
INPP5F	90.7	87.0	0.074	chr1	Ultra272	chr6	chr13	6	*SACM1L*	15	Bilateria
TP73	90.9	85.9	0.080	chr4	Ultra178	chr21	chr8	2	*TP63*	55	Euteleostomi
HTR2A	91.5	87.5	0.073	chr4	Ultra336	chr1	chr9	16	*HTR2C*	46	Euteleostomi
CD81	91.9	86.7	0.066	NO	NO	chr5	chr7	2	*CD9*	31	Euteleostomi
ASB4	92.5	85.7	0.038	unplaced	not assembled	chr2	chr19	3	*ASB18*	33	Euteleostomi
DIO3	95.7	87.8	0.032	chr1	not assembled	chr5	chr17	2	*DIO2*	33	Euteleostomi
GATM	95.7	90.9	0.056	chr1	chr2	chr10	chr18				
SGCE	96.1	89.7	0.030	unplaced	Contig1947	chr2	chr19	1	*SGCA*	36	Euteleostomi
UBE3A	96.2	93.6	0.057	chr7	Ultra222	chr1	chr6	17	*HECTD2*	22	Euteleostomi
DLX5	96.5	93.0	0.049	unplaced	not assembled	chr2	chr19				
LRRTM1	97.3	92.2	0.042	chr1	Contig8748	chr4	chr1	3	*LRRTM2*	46	Euteleostomi
WT1	97.5	91.1	0.026	chr5	Ultra222	chr3	chr25				
COPG2	97.6	91.0	0.026	chr8	chr10	chr14	chr4	1	*COPG*	83	Euteleostomi
MEST (*PEG1*)	97.9	90.8	0.021	chr8	chr10	chr14	chr4				
NNAT	98.8	96.7	0.050	NO	NO	NO	NO				
SNRPN	100.0	91.5	0.000	1(*SNRPB*)	NO	NO	NO	1	*SNRPB*	81	Theria
ZIM3	NA	NA	NA	NO	NO	NO	NO				

imprinted gene (synonym)	human-mouse protein ID (%)	human-mouse DNA ID (%)	human-mouse Ka/Ks	opossum ortholog	platypus ortholog	chicken ortholog	zebrafish ortholog	number of human paralogs	youngest human paralog	gene-paralog protein ID (%)	paralog ancestor
ZNF264	NA	NA	NA	NO	NO	NO	NO				
ZIM2	NA	NA	NA	NO	NO	NO	NO				
GNAS	NA	NA	NA	chr1	Ultra516	chr20	chr6	15	*GNAL*	28	Euteleostomi
PLAGL1 (*ZAC1*)	NA	NA	NA	chr2	chr2	chr3	chr17	2	*PLAGL2*	44	Euteleostomi

NA: not applicable (no data in HomoloGene)
NO: no ortholog. Orthology information was taken from the literature (Rapkins et al. 2006, Dünzinger et al. 2007, Hore et al. 2007, Pask et al. 2009) and completed by Ensembl release 52 data.
Genes are listed in ascending order of their protein identity. Maternally expressed genes are marked in red color, paternally expressed ones in blue. The three genes marked in violet (*COPG2*, *GRB10*, and *ZIM2*) show paternal expression in human but their murine orthologs are maternally expressed.

References

Bibliography

Ager E, Suzuki S, Pask A, Shaw G, Ishino F, and Renfree MB (2007) Insulin is imprinted in the placenta of the marsupial, *Macropus eugenii*. *Dev Biol.* **309:** 317-28.

Aïssani B and Bernardi G (1991) CpG islands: features and distribution in the genomes of vertebrates. *Gene* **106:** 173-183.

Allen E, Horvath S, Tong F, Kraft P, Spiteri E, Riggs AD, and Maharens Y (2003) High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proc Natl Acad Sci.* **100:** 9940-9945.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res.* **25:** 3389-3402.

Antequera F and Bird A (1993) Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci.* **90:** 11995-11999.

Antequera F and Bird A (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr Biol.* **9:** R661-R667.

Antequera F (2003) Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci.* **60:** 1647-1658.

Arima T, Kamikihara T, Hayashida T, Kato K, Inoue T, Shirayoshi Y, Oshimura M, Soejima H, Mukai T, and Wake N (2005) *ZAC*, *LIT1* (*KCNQ1OT1*) and *p57*^{KIP2}(*CDKN1C*) are in an imprinted gene network that may play a role in Beckwith-Wiedemann syndrome. *Nucl Acids Res.* **33:** 2650-2660.

Arnaud P, Monk D, Hitchins M, Gordon E, Dean W, Beechey CV, Peters J, Craigen W, Preece M, Stanier P, et al. (2003) Conserved methylation imprints in the human and mouse *GRB10* genes with divergent allelic expression suggests differential reading of the same mark. *Hum Mol Genet.* **12:** 1005-1019.

Babak T, Deveale B, Armour C, Raymond C, Cleary MA, van der Kooy D, Johnson JM, and Lim LP (2008) Global survey of genomic imprinting by transcriptome sequencing. *Curr Biol.* **18:** 1735-41.

Baek D, Davis C, Ewing B, Gordon D, and Green P (2007) Characterization and predictive discovery of evolutionary conserved mammalian alternative promoters. *Genome Res.* **17:** 145-155.

Bailey TL and Elkan CP (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the second International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California, 28-36.

Bailey TL and Elkan CP (1995a) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21:** 51-83.

Bailey TL and Elkan CP (1995b) The value of prior knowledge in discovering motifs with MEME. *Proceedings of the third Iinternational Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California, 21-29.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, and Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823-37.

Bell AC and Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405:** 482-485.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl Acids Res.* **27:** 573-580.

Bestor TH and Tycko B (1996) Creation of genomic methylation patterns. *Nature Genet.* **12:** 363-367.

Bininda-Emonds OR (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* **6:** 156.

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucl Acids Res.* **8:** 1499-1504.

Bird AP (1986) CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209-213.

Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.* **16:** 6-21.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799-816.

Bischoff SR, Tsai S, Hardison N, Motsinger-Reif AA, Freking BA, Nonneman D, Rohrer G, and Piedrahita JA (2009) Characterization of conserved and nonconserved imprinted genes in swine. *Biol Reprod.* **1:** 1.

Blanchette M and Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739-748.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708-15.

Blanchette M, Bataille AR, Chen X, Poitras C, Laganière J, Lefèbvre C, Deblois G, Giguère V, Ferretti V, Bergeron D, et al. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16:** 656-668.

Boccaccio I, Glatt-Deeley H, Watrin F, Roëckel N, Lalande M, and Muscatelli F (1999) The human *MAGEL2* gene and its mouse homologue are paternally expressed and mapped to the Prader-Willi region. *Hum Mol Genet.* **8:** 2497-2505.

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, and Walter J (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence patterns, repeat frequencies and predicted DNA structure. *PLoS Genetics* **2:** e26.

Bock C, Walter J, Paulsen M, and Lengauer T (2007) CpG island mapping by epigenome prediction. *PLoS Comput Biol.* **3:** e110.

Bock C, Halachev K, Büch J, and Lengauer T (2009) EpiGRAPH: A user-friendly software for statistical analysis and prediction of (epi-)genomic data. *Genome Biology* **10:** R14.

Bourc'his D and Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431:** 96-99.

Bourc'his D and Bestor TH (2006) Origins of extreme sexual dimorphism in genomic imprinting. *Cytogenet Genome Res.* **113:** 36-40.

Braem C, Recolin B, Rancourt RC, Angiolini C, Barthes P, Branchu P, Court F, Cathala G, Ferguson-Smith AC, and Forne T (2008) Genomic matrix attachment region and chromosome conformation capture quantitative real time PCR assays identify novel putative regulatory elements at the imprinted *Dlk1*/*Gtl2* locus. *J Biol Chem.* **283:** 18612-20.

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, and Cedar H (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature* **371:** 435-8.

Bridgham JT, Brown JE, Rodriguez-Mari A, Catchen JM, and Thornton JW (2008) Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet.* **4:** e1000191.

Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, and Robinson-Rechavi M (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol.* **23:** 1808-16.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153-7.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38:** 626-635.

Chai JH, Locke DP, Ohta T, Greally JM, and Nicholls RD (2001) Retrotransposed genes such as *Frat3* in the mouse Chromosome 7C Prader-Willi syndrome region acquire the imprinted status

of their insertion site. *Mamm Genome* **12:** 813-21.
Chain FJ and Evans BJ (2006) Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. *PLoS Genet.* **2:** e56.
Chen J, Sun M, Hurst LD, Carmichael GG, and Rowley JD (2005) Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet.* **21:** 326-9.
Chernukhin I, Shamsuddin S, Kang SY, Bergstrom R, Kwon YW, Yu W, Whitehead J, Mukhopadhyay R, Docquier F, Farrar D, et al. (2007) CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. *Mol Cell Biol.* **27:** 1631-48.
Chesnokov IN and Schmid CW (1995) Specific *Alu* binding protein from human sperm chromatin prevents DNA methylation. *J Biol Chem.* **270:** 18539-18542.
Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69-87.
Chotalia M, Smallwood SA, Ruf N, Dawson C, Lucifero D, Frontera M, James K, Dean W, and Kelsey G (2009) Transcription is required for establishment of germline methylation marks at imprinted genes. *Genes Dev.* **23:** 105-17.
Conant GC and Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* **9:** 938-50.
Constância M, Kelsey G, and Reik W (2004) Resourceful imprinting. *Nature* **432:** 53-57.
Coombes C, Arnaud P, Gordon E, Dean W, Coar EA, Williamson CM, Feil R, Peters J, and Kelsey G (2003) Epigenetic properties and identification of an imprint mark in the *Nesp-Gnasxl* domain of the mouse *Gnas* imprinted locus. *Mol Cell Biol.* **23:** 5475-5488.
Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucl Acids Res.* **13:** 3021-30.
Cross S, Clark VH, Simmen MW, Bickmore WA, Maroon H, Langford CF, Carter NP, and Bird AP (2000) CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. *Mamm Genome* **11:** 373-383.
Cuadrado M, Sacristán M, and Antequera F (2001) Species-specific organization of CpG island promoters at mammalian homologous genes. *EMBO reports* **2:** 586-592.
Davis TL, Yang GJ, McCarrey JR, and Bartolomei MS (2000) The *H19* methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. *Hum Mol Genet.* **9:** 2885-94.
de la Puente A, Hall J, Wu YZ, Leone G, Peters J, Yoon BJ, Soloway P, and Plass C (2002) Structural characterization of *Rasgrf1* and a novel linked imprinted locus. *Gene* **291:** 287-97.
Dehal P and Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3:** e314.
Delcher AL, Salzberg SL, and Phillippy AM (2003) Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* **Chapter 10:** Unit 10.3.
Dermitzakis ET and Clark AG (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol.* **19:** 1114-1121.
Donohoe ME, Zhang LF, Xu N, Shi Y, and Lee JT (2007) Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol Cell* **25:** 43-56.
Durbin R, Eddy S, Krogh A, and Mitchison G (1998) *Biological sequence analysis: probability models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.
Duselis AR and Vrana PB (2007) Assessment and disease comparisons of hybrid developmental defects. *Hum Mol Genet.* **16:** 808-19.
Dünzinger U, Haaf T, and Zechner U (2007) Conserved synteny of mammalian imprinted genes in chicken, frog, and fish genomes. *Cytogenet Genome Res.* **117:** 78-85.
Elnitski L, Riemer C, Petrykowska H, Florea L, Schwartz S, Miller W, and Hardison R (2002) PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. *Genomics.* **80:** 681-690.
Engemann S, Strödicke M, Paulsen M, Franck O, Reinhardt R, Lane N, Reik W, and Walter J (2000) Sequence and functional comparison in the Beckwith-Wiedemann region: implications

for a novel imprinting centre and extended imprinting. *Hum Mol Genet.* **9:** 2691-2706.
Evans HK, Weidman JR, Cowley DO, and Jirtle RL (2005) Comparative phylogenetic analysis of *Blcap/Nnat* reveals Eutherian-specific imprinted gene. *Mol Biol Evol.* **22:** 1740-1748.
Feil R and Berger F (2007) Convergent evolution of genomic imprinting in plants and mammals. *Trends Genet.* **23:** 192-9.
Ferguson-Smith AC and Reik W (2003) The need for Eed. *Nature Genet.* **33:** 433-434.
Fitzpatrick GV, Soloway PD, and Higgins MJ (2002) Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nature Genet.* **32:** 426-31.
Fitzpatrick GV, Pugacheva EM, Shin JY, Abdullaev Z, Yang Y, Khatod K, Lobanenkov VV, and Higgins MJ (2007) Allele-specific binding of CTCF to the multipartite imprinting control region KvDMR1. *Mol Cell Biol.* **27:** 2636-47.
Freed WJ, Chen J, Backman CM, Schwartz CM, Vazin T, Cai J, Spivak CE, Lupica CR, Rao MS, and Zeng X (2008) Gene expression profile of neuronal progenitor cells derived from hESCs: activation of chromosome 11p15.5 and comparison to human dopaminergic neurons. *PLoS ONE* **3:** e1422.
Frith MC, Ponjavic J, Fredman D, Kai C, Kawai J, Carninci P, Hayashizaki Y, and Sandelin A (2006) Evolutionary turnover of mammalian transcription start sites. *Genome Res.* **16:** 713-722.
Gardiner-Garden M and Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol.* **196:** 261-282.
Gehring M, Bubb KL, and Henikoff S (2009) Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **324:** 1447-51.
Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493-521.
Gimelbrant A, Hutchinson JN, Thompson BR, and Chess A (2007) Widespread monoallelic expression on human autosomes. *Science* **318:** 1136-40.
Glaser RL, Ramsay JP, and Morison IM (2006) The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. *Nucl Acids Res.* **34:** D29-31.
Glass JL, Thompson RF, Khulan B, Figueroa ME, Olivier EN, Oakley EJ, Van Zant G, Bouhassira EE, Melnick A, Golden A, et al. (2007) CG dinucleotide clustering is a species-specific property of the genome. *Nucl Acids Res.* **35:** 6798-6807.
Grandjean V, Smith J, Schofield PN, and Ferguson-Smith AC (2000) Increased IGF-II protein affects *p57*^{KIP2} expression *in vivo* and *in vitro*: Implications for Beckwith-Wiedemann syndrome. *Proc Natl Acad Sci.* **97:** 5279-5284.
Greally JM, Gray TA, Gabriel JM, Song L, Zemel S, and Nicholls RD (1999) Conserved characteristics of heterochromatin-forming DNA at the 15q11-q13 imprinting center. *Proc Natl Acad Sci U S A.* **96:** 14430-5.
Greally JM (2002) Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci.* **99:** 327-332.
Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, et al. (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucl Acids Res.* **36:** D107-13.
Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Arosa J, and Oliver JL (2006) *CpGcluster*: a distance-based algorithm for CpG-islands detection. *BMC Bioinformatics* **7:** 446.
Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, Walter J, and Surani MA (2002) Epigenetic reprogramming in mouse primordial germ cells. *Mech Dev.* **117:** 15-23.
Hajkova P, Ancelin K, Waldmann T, Lacoste N, Lange UC, Cesari F, Lee C, Almouzni G, Schneider R, and Surani MA (2008) Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature* **452:** 877-81.
Han L and Zhao Z (2009) CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? *BMC Bioinformatics* **10:** 65.
Han MV, Demuth JP, McGrath CL, Casola C, and Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. *Genome Res.* **19:** 859-67.
Hancock AL, Brown KW, Moorwood K, Moon H, Holmgren C, Mardikar SH, Dallosso AR,

Klenova E, Loukinov D, Ohlsson R, et al. (2007) A CTCF-binding silencer regulates the imprinted genes *AWT1* and *WT1-AS* and exhibits sequential epigenetic defects during Wilms' tumourigenesis. *Hum Mol Genet.* **16:** 343-54.

Hannenhalli S and Levy S (2001) Promoter prediction in the human genome. *Bioinformatics* **17 Suppl. 1:** S90-S96.

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, and Tilghman SM (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature* **405:** 486-489.

Hatada I and Mukai T (1995) Genomic imprinting of $p57^{KIP2}$, a cyclin-dependent kinase inhibitor, in mouse. *Nature Genet.* **11:** 204-206.

He L and Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet.* **5:** 522-31.

He X and Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169:** 1157-1164.

Hellman A and Chess A (2007) Gene body-specific methylation on the active X chromosome. *Science* **315:** 1141-3.

Hellmann-Blumberg U, Hintz MFM, Gatewood JM, and Schmid CW (1993) Developmental differences in methylation of human *Alu* repeats. *Mol Cell Biol.* **13:** 4523-4530.

Hikichi T, Kohda T, Kaneko-Ishino T, and Ishino F (2003) Imprinting regulation of the murine *Meg1/Grb10* and human *GRB10* genes; roles of brain-specific promotors and mouse-specific CTCF-binding sites. *Nucl Acids Res.* **31:** 1398-1406.

Hong M, Fitzgerald MX, Harper S, Luo C, Speicher DW, and Marmorstein R (2008) Structural basis for dimerization in DNA recognition by Gal4. *Structure* **16:** 1019-26.

Hore TA, Rapkins RW, and Graves JA (2007) Construction and evolution of imprinted loci in mammals. *Trends Genet.* **23:** 440-8.

Hore TA, Deakin JE, and Marshall Graves JA (2008) The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes. *PLoS Genet.* **4:** e1000169.

Howlett SK and Reik W (1991) Methylation levels of maternal and paternal genomes during preimplantation development. *Development* **113:** 119-127.

Hughes JD, Estep PW, Tavazoie S, and Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol.* **296:** 1205-1214.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99:** 364-73.

Huq AHMM, Sutcliffe JS, Nakao M, Shen Y, Gibbs RA, and Beaudet AL (1997) Sequencing and functional analysis of the *SNPRN* promoter: in vitro methylation abolishes promoter activity. *Genome Res.* **7:** 642-648.

Hutter B, Helms V, and Paulsen M (2006) Tandem repeats in the CpG islands of imprinted genes. *Genomics* **88:** 323-332.

Hutter B, Paulsen M, and Helms V (2009) Identifying CpG islands by different computational techniques. *Omics* **13:** 153-164.

Hwang DG and Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci.* **101:** 13994-4001.

Ioshikhes IP and Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. *Nature Genet.* **26:** 61-63.

Ishihara K, Hatano N, Furuumi H, Kato R, Iwaki T, Miura K, Jinno Y, and Sasaki H (2000) Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in *Igf2/H19* imprinting. *Genome Res.* **10:** 664-71.

Jabbari K, Caccio S, Pais de Barros JP, Desgres J, and Bernardi G (1997) Evolutionary changes in CpG and methylation levels in the genome of vertebrates. *Gene* **205:** 109-118.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. (2004) Genome duplication in the teleost fish *Tetraodon*

nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431:** 946-57.
Jiang C, Han L, Su B, Li WH, and Zhao Z (2007) Features and trend of loss of promoter-associated CpG islands in the human and mouse genomes. *Mol Biol Evol.* **24:** 1991-2000.
Jones PA (1999) The DNA methylation paradox. *Trends Genet.* **15:** 34-37.
Jones PA and Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nat Rev. Genet.* **3:** 415-426.
Jones PL, Veenstra GJC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, and Wolffe AP (1998) Methylated DNA and MeCP2 recruit histone deacetlyase to repress transcription. *Nature Genet.* **19:** 187-191.
Jordan IK, Rogozin IB, Glazko GV, and Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19:** 68-72.
Jordan IK, Wolf YI, and Koonin EV (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol.* **4:** 22.
Kaessmann H, Vinckenbosch N, and Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet.* **10:** 19-31.
Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E, and Sasaki H (2004) Essential role for *de novo* DNA methyltranferase Dnmt3a in paternal and maternal imprinting. *Nature* **429:** 900-903.
Kang K, Chung JH, and Kim J (2009) Evolutionary Conserved Motif Finder (ECMFinder) for genome-wide identification of clustered YY1- and CTCF-binding sites. *Nucl Acids Res.* **37:** 2003-13.
Kato Y, Kaneda M, Hata K, Kumaki K, Hisano M, Kohara Y, Okano M, Li E, Nozaki M, and Sasaki H (2007) Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Hum Mol Genet.* **16:** 2272-80.
Ke X, Thomas SN, Robinson DO, and Collins A (2002a) A novel approach for identifying candidate imprinted genes through sequence analysis of imprinted and control genes. *Hum Genet.* **111:** 511-520.
Ke X, Thomas NS, Robinson DO, and Collins A (2002b) The distinguishing sequence characteristics of mouse imprinted genes. *Mamm Genome* **13:** 639-645.
Kennedy GC, German MS, and Rutter WJ (1995) The minisatellite in the diabetes susceptibility locus *IDDM2* regulates insulin transcription. *Nature Genet.* **9:** 293-298.
Kent WJ (2002) BLAT - the BLAST-like alignment tool. *Genome Res.* **12:** 656-64.
Khatib H, Zaitoun I, and Kim ES (2007) Comparative analysis of sequence characteristics of imprinted genes in human, mouse, and cattle. *Mamm Genome.* **18:** 538-47.
Killian JK, Byrd JC, Jirtle JV, Munday BL, Stoskopf MK, MacDonald RG, and Jirtle RL (2000) M6P/IGF2R imprinting evolution in mammals. *Mol Cell* **5:** 707-16.
Killian KJ, Nolan CM, Wylie AA, Vu TH, Li T, Hoffman AR, and Jirtle RL (2001) Divergent evolution in *M6P/IGF2R* imprinting from the Jurassic to the Quaternary. *Hum Mol Genet.* **10:** 1721-1728.
Kim J, Kollhoff A, Bergmann A, and Stubbs L (2003) Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, *Peg3*. *Hum Mol Genet.* **12:** 233-245.
Kim J, Bergmann A, Choo JH, and Stubbs L (2007) Genomic organization and imprinting of the *Peg3* domain in bovine. *Genomics* **90:** 85-92.
Kim J (2008) Multiple YY1 and CTCF binding sites in imprinting control regions. *Epigenetics* **3:** 115-8.
Kim JD, Hinz AK, Bergmann A, Huang JM, Ovcharenko I, Stubbs L, and Kim J (2006) Identification of clustered YY1 binding sites in imprinting control regions. *Genome Res.* **16:** 901-11.
Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, and Ren B (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231-45.
Kiyosawa H, Yamanaka I, Osato N, Kondo S, and Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13:** 1324-34.

Kobayashi H, Suda C, Abe T, Kohara Y, Ikemura T, and Sasaki H (2006) Bisulfite sequencing and dinucleotide content analysis of 15 imprinted mouse differentially methylated regions (DMRs): paternally methylated DMRs contain less CpGs than maternally methylated DMRs. *Cytogenet Genome Res.* **113:** 130-7.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, and Siepel A (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* **4:** e1000144.

Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, Reik W, and Ohlsson R (2006) CTCF binding at the *H19* imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to *Igf2*. *Proc Natl Acad Sci.* **103:** 10684-10689.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FithHugh W, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J, and Reik W (2003) Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis* **35:** 88-93.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947-2948.

Larsen F, Gundersen G, Lopez R, and Prydz H (1992) CpG islands as gene markers in the human genome. *Genomics* **13:** 1095-1107.

Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, and Casari G (2004) In search of antisense. *Trends Biochem Sci.* **29:** 88-94.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, and Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262:** 208-214.

Lee J, Li Z, Brower-Sinning R, and John B (2007) Regulatory circuit of human microRNA biogenesis. *PLoS Comput Biol.* **3:** e67.

Lefebvre L, Viville S, Barton SC, Ishino F, and Surani MA (1997) Genomic structure and parent-of-origin-specific methylation of *Peg1*. *Hum Mol Genet.* **6:** 1907-1915.

Lefebvre L, Viville S, Barton SC, Ishino F, Keverne EB, and Surani MA (1998) Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene *Mest*. *Nature Genet.* **20:** 163-9.

Lehner B, Williams G, Campbell RD, and Sanderson CM (2002) Antisense transcripts in the human genome. *Trends Genet.* **18:** 63-65.

Lewis A, Kohzoh M, Constância M, and Reik W (2004) Tandem repeat hypothesis in imprinting: deletion of a conserved direct repeat element upstream of *H19* has no effect on imprinting in the *Igf2-H19* region. *Mol Cell Biol.* **24:** 5650-5656.

Lewis A, Green K, Dawson C, Redrup L, Huynh KD, Lee JT, Hemberger M, and Reik W (2006) Epigenetic dynamics of the *Kcnq1* imprinted domain in the early embryo. *Development* **133:** 4203-4210.

Li T, Vu TH, Lee K-O, Yang Y, Nguyen CV, Bui HQ, Zeng Z-L, Nguyen BT, Hu J-F, Murphy SK, et al. (2002) An imprinted *PEG1/MEST* antisense expressed predominantly in human testis and in mature spermatozoa. *J Biol Chem.* **277:** 13518-13527.

Li W, Bernaola-Galván P, Haghighi F, and Grosse I (2002) Applications of recursive segmentation to the analysis of DNA sequences. *Comput Chem.* **26:** 491-510.

Liao BY and Zhang J (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol.* **23:** 530-540.

Lin S-P, Youngson N, Takada S, Seitz H, Reik W, Paulsen M, Cavaille J, and Ferguson-Smith AC (2003) Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the *Dlk1-Gtl2* imprinted cluster on mouse chromosome 12. *Nature Genet.* **35:** 97-102.

Lindroth AM, Park YJ, McLean CM, Dokshin GA, Persson JM, Herman H, Pasini D, Miro X, Donohoe ME, Lee JT, et al. (2008) Antagonism between DNA and H3K27 methylation at the imprinted *Rasgrf1* locus. *PLoS Genet.* **4:** e1000145.

Loots GG, Ovcharenko I, Pachter L, Dubchak I, and Rubin EM (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12:** 832-839.

Loots GG and Ovcharenko I (2005) Dcode.org anthology of comparative genomics tools. *Nucl Acids Res.* **33:** W56-W64.

Loukinov DI, Pugacheva E, Vatolin S, Pack SD, Moon H, Chernukhin I, Mannan P, Larsson E, Kanduri C, Vostrov AA, et al. (2002) BORIS: a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. *Proc Natl Acad Sci.* **99:** 6806-6811.

Lowe CB, Bejerano G, and Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci.* **104:** 8005-10.

Lu J and Wu CI (2005) Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc Natl Acad Sci.* **102:** 4063-7.

Lucifero D, Mann MRW, Bartolomei MS, and Trasler J (2004) Gene-specific timing and epigenetic memory in oocyte imprinting. *Hum Mol Genet.* **13:** 839-849.

Luedi PP, Hartemink AJ, and Jirtle RL (2005) Genome-wide prediction of imprinted murine genes. *Genome Res.* **15:** 875-884.

Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, and Hartemink AJ (2007) Computational and experimental identification of novel human imprinted genes. *Genome Res.* **17:** 1723-1730.

Luque-Escamilla P, Martínez-Aroza J, Oliver JL, Gómez-Lopera JF, and Román-Roldán R (2005) Compositional searching of CpG islands in the human genome. *Phys Rev E Stat Nonlin Soft Matter Phys.* **71:** 061925.

Lyon MF (2006) Do LINEs have a role in X-chromosome inactivation? *J Biomed Biotechnol.* **2006:** 59746.

Macleod D, Charlton J, Mullins J, and Bird AP (1994) Sp1 sites in the mouse *aprt* gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* **8:** 2282-92.

Mager J, Montgomery ND, de Villena FP-M, and Magnuson T (2003) Genome imprinting regulated by the mouse Polycomb group protein Eed. *Nature Genet.* **33:** 502-507.

Malagnac F, Wendel B, Goyon C, Faugeron G, Zickler D, Rossignol JL, Noyer-Weidner M, Vollmayr P, Trautner TA, and Walter J (1997) A gene essential for de novo methylation and development in Ascobolus reveals a novel type of eukaryotic DNA methyltransferase structure. *Cell* **91:** 281-290.

Mallick S, Gnerre S, Muller P, and Reich D (2009) The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* **19:** 922-33.

Mancini-DiNardo D, Steele SJS, Ingram RS, and Tilghman SM (2003) A differentially methylated region within the gene *Kcnq1* functions as an imprinted promoter and silencer. *Hum Mol Genet.* **12:** 283-294.

Martienssen RA (2003) Maintenance of heterochromatin by RNA interference of tandem repeats. *Nature Genet.* **35:** 213-214.

Matsuo K, Clay O, Takahashi T, Silke J, and Schaffner W (1993) Evidence for erosion of mouse CpG islands during mammalian evolution. *Som Cell Mol Gen.* **19:** 543-555.

Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucl Acids Res.* **31:** 374-378.

McGrath J and Solter D (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell* **37:** 179-83.

McLysaght A, Hokamp K, and Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31:** 200-4.

McVean GT and Hurst LD (1997) Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. *Proc Biol Sci.* **264:** 739-46.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE,

Nusbaum C, Jaffe DB, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766-770.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, et al. (2007a) Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167-77.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. (2007b) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553-560.

Monk D, Arnaud P, Apostolidou S, Hills FA, Kelsey G, Stanier P, Feil R, and Moore GE (2006) Limited evolutionary conservation of imprinting in the human placenta. *Proc Natl Acad Sci.* **103:** 6623-8.

Monk D, Wagschal A, Arnaud P, Muller PS, Parker-Katiraee L, Bourc'his D, Scherer SW, Feil R, Stanier P, and Moore GE (2008) Comparative analysis of human chromosome 7q21 and mouse proximal chromosome 6 reveals a placental-specific imprinted gene, *TFPI2/Tfpi2*, which requires EHMT2 and EED for allelic-silencing. *Genome Res.* **18:** 1270-81.

Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, and Jones SJ (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* **22:** 637-40.

Moore T and Haig D (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet.* **7:** 45-9.

Moore T, Constância M, Zubair M, Bailleul B, Feil R, Sasaki H, and Reik W (1997) Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse *Igf2*. *Proc Natl Acad Sci.* **94:** 12509-12514.

Morgan HD, Santos F, Green K, Dean W, and Reik W (2005) Epigenetic reprogramming in mammals. *Hum Mol Genet.* **14 Spec No 1:** R47-58.

Morison IM, Paton CJ, and Cleverley SD (2001) The imprinted gene and parent-of-origin effect database. *Nucl Acids Res.* **29:** 275-6.

Morison IM, Ramsay JP, and Spencer HG (2005) A census of mammalian imprinting. *Trends Genet.* **21:** 457-65.

Murrell A, Heeson S, and Reik W (2004) Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nature Genet.* **36:** 889-893.

Nei M and Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* **3:** 418-26.

Neumann B, Kubicka P, and Barlow DP (1995) Characteristics of imprinted genes. *Nature Genet.* **9:** 12-13.

Nicholls RD, Saitoh S, and Horsthemke B (1998) Imprinting in Prader-Willi and Angelman syndromes. *Trends Genet.* **14:** 194-200.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, and Clark AG (2007) Recent and ongoing selection in the human genome. *Nat Rev Genet.* **8:** 857-68.

Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M, Kono T, Morris GA, Lyons PA, Oshimura M, et al. (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res.* **13:** 1402-1409.

O'Neill MJ, Lawton BR, Mateos M, Carone DM, Ferreri GC, Hrbek T, Meredith RW, Reznick DN, and O'Neill RJ (2007) Ancient and continuing Darwinian selection on insulin-like growth factor II in placental fishes. *Proc Natl Acad Sci.* **104:** 12404-9.

O'Sullivan FM, Murphy SK, Simel LR, McCann A, Callanan JJ, and Nolan CM (2007) Imprinted expression of the canine *IGF2R*, in the absence of an anti-sense transcript or promoter methylation. *Evol Dev.* **9:** 579-89.

Oakey RJ and Beechey CV (2002) Imprinted genes: identification by chromosome rearrangements and post-genomic strategies. *Trends Genet.* **18:** 359-366.

Oei S-L, Babich VS, Kazakov VI, Usmanova NM, Kropotov AV, and Tomilin NV (2004) Clusters

of regulatory signals for RNA polymerase II transcription associated with *Alu* family repeats and CpG islands in human promoters. *Genomics* **83:** 873-882.

Ohlsson R, Renkawitz R, and Lobanenkov V (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17:** 520-527.

Okamura K, Hagiwara-Takeuchi Y, Li T, Vu TH, Hirai M, Hattori M, Sakaki Y, Hoffman AR, and Ito T (2000) Comparative genome analysis of the mouse imprinted gene *Impact* and its nonimprinted human homolog *IMPACT*: toward the structural basis for species-specific imprinting. *Genome Res.* **10:** 1878-1889.

Okamura K, Sakaki Y, and Ito T (2005) Comparative genomics approach toward critical determinants for the imprinting of an evolutionary conserved gene *Impact*. *Biophys Res Comm.* **329:** 824-830.

Olek A and Walter J (1997) The pre-implantation ontogeny of the *H19* methylation imprint. *Nature Genet* **17:** 275-6.

Pan T and Coleman JE (1990) GAL4 transcription factor is not a "zinc finger" but forms a Zn(II)2Cys6 binuclear cluster. *Proc Natl Acad Sci.* **87:** 2077-81.

Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, and Frazer KA (2006) Analysis of allelic differential expression in human white blood cells. *Genome Res.* **16:** 331-339.

Parker-Katiraee L, Carson AR, Yamada T, Arnaud P, Feil R, Abu-Amero SN, Moore GE, Kaneda M, Perry GH, Stone AC, et al. (2007) Identification of the imprinted *KLF14* transcription factor undergoing human-specific accelerated evolution. *PLoS Genet.* **3:** e65.

Parsons JD (1995) Miropeats: graphical DNA sequence comparisons. *CABIOS* **11:** 615-619.

Pask AJ, Papenfuss AT, Ager EI, McColl KA, Speed TP, and Renfree MB (2009) Analysis of the platypus genome suggests a transposon origin for mammalian imprinting. *Genome Biol.* **10:** R1.

Patten MM and Haig D (2008) Reciprocally imprinted genes and the response to selection on one sex. *Genetics* **179:** 1389-94.

Pauler FM, Koerner MV, and Barlow DP (2007) Silencing by imprinted noncoding RNAs: is transcription the answer? *Trends Genet.* **23:** 284-92.

Paulsen M, El-Maarri O, Engemann S, Strödicke M, Franck O, Davies K, Reinhardt R, Reik W, and Walter J (2000) Sequence conservation and variability of imprinting in the Beckwith-Wiedemann syndrome gene cluster in human and mouse. *Hum Mol Genet.* **9:** 1829-1841.

Paulsen M, Takada S, Youngson NA, Benchaib M, Charlier C, Segers K, Georges M, and Ferguson-Smith AC (2001) Comparative sequence analysis of the imprinted *Dlk1-Gtl2* locus in three mammalian species reveals highly conserved genomic elements and refines comparison with the *Igf2-H19* region. *Genome Res.* **11:** 2085-2094.

Paulsen M and Ferguson-Smith AC (2001) DNA methylation in genomic imprinting, development, and disease. *J Pathol.* **195:** 97-110.

Paulsen M, Khare T, Burgard C, Tierling S, and Walter J (2005) Evolution of the Beckwith-Wiedemann syndrome region in vertebrates. *Genome Res.* **15:** 146-153.

Pearsall RS, Shibata H, Brozowska A, Yoshino K, Okuda K, deJong PJ, Plass C, Chapman VM, Hayhshizaki Y, and Held WA (1996) Absence of imprinting in *U2AFBPL*, a human homologue of the imprinted mouse gene *U2afbp-rs*. *Biochem Biophys Res Comm.* **222:** 171-177.

Pearsall RS, Plass C, Romano MA, Garrick MD, Shibata H, Hayashizaki Y, and Held WA (1999) A direct repeat sequence at the *Rasgrf1* locus and imprinted expression. *Genomics* **55:** 194-201.

Pennacchio LA, Loots GG, Nobrega MA, and Ovcharanko I (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.* **17:** 201-211.

Peters J and Robson JE (2008) Imprinted noncoding RNAs. *Mamm Genome* **19:** 493-502.

Ponger L, Duret L, and Mouchiroud D (2001) Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.* **11:** 1854-1860.

Ponger L and Mouchiroud D (2001) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18:** 631-633.

Quentin Y (1994) A master sequence related to a free left *Alu* monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucl Acids Res.* **22:** 2222-2227.

Rapkins RW, Hore T, Smithwick M, Ager E, Pask AJ, Renfree MB, Kohn M, Hameister H, Nicholls RD, Deakin JE, et al. (2006) Recent assembly of an imprinted domain from non-

imprinted components. *PLoS Genet.* **2:** e182.
Reed M, Riggs AD, and Mann JR (2001) Deletion of a direct repeat element has no effect on *Igf2* and *H19* imprinting. *Mamm Genome* **12:** 873-876.
Reik W and Walter J (2001) Genomic imprinting: parental influence on the genome. *Nat Rev Genet.* **2:** 21-32.
Reik W and Lewis A (2005) Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nat Rev Genet.* **6:** 403-410.
Reinhart B, Eljanne M, and Chaillet JR (2002) Shared role for differentially methylated domains of imprinted genes. *Mol Cell Biol.* **22:** 2089-2098.
Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, and Koonin EV (2007) Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol.* **24:** 1821-31.
Rice P, Longden I, and Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16:** 276-277.
Riesewijk AM, Schepens MT, Welch TR, van den Berg-Loonen EM, Mariman EM, Ropers H-H, and Kalscheuer VM (1996) Maternal-specific methylation of the human *IGF2R* gene is not accompanied by allele-specific transcription. *Genomics* **31:** 158-166.
Robertson KD and Wolffe AP (2000) DNA methylation in health and disease. *Nat Rev Genet.* **1:** 11-19.
Rodriguez-Jato S, Nicholls RD, Driscoll DJ, and Yang TP (2005) Characterization of cis- and trans-acting elements in the imprinted human *SNURF-SNPRN* locus. *Nucl Acids Res.* **33:** 4740-4753.
Royo H and Cavaillé J (2008) Non-coding RNAs in imprinted gene clusters. *Biol Cell* **100:** 149-66.
Rubin CM, VandeVoort CA, Teplitz RL, and Schmid CW (1994) *Alu* repeated DNAs are differentially methylated in primate germ cells. *Nucl Acids Res.* **22:** 5121-5127.
Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, Aebersold R, Ranish JA, and Krumm A (2008) CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci.* **105:** 8309-14.
Ruf N, Bahring S, Galetzka D, Pliushch G, Luft FC, Nurnberg P, Haaf T, Kelsey G, and Zechner U (2007) Sequence-based bioinformatic prediction and QUASEP identify genomic imprinting of the *KCNK9* potassium channel gene in mouse and human. *Hum Mol Genet.* **16:** 2591-9.
Sandovici I, Kassovska-Bratinove S, Vaughan JE, Steward R, Leppert M, and Sapienza C (2006) Human imprinted chromosomal regions are historical hot-spots of recombination. *PLoS Genetics* **2:** e101.
Sanz LA, Chamberlain S, Sabourin JC, Henckel A, Magnuson T, Hugnot JP, Feil R, and Arnaud P (2008) A mono-allelic bivalent chromatin domain controls tissue-specific imprinting at *Grb10*. *Embo J.* **27:** 2523-32.
Sasaki H, Shimozaki K, Zubair M, Aoki N, Ohta K, Hatano N, Moore T, Feil R, Constância M, Reik W, et al. (1996) Nucleotide sequence of a 28-kb mouse genomic region comprising the imprinted *Igf2* gene. *DNA Research* **3:** 331-335.
Saveliev A, Everett C, Sharpe T, Webster Z, and Festenstein R (2003) DNA triplet repeats mediate heterochromatin-protein-1-sensitive variegated gene silencing. *Nature* **422:** 909-913.
Saxonov S, Berg P, and Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci.* **103:** 1412-1417.
Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucl Acids Res.* **18:** 6097-100.
Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, and Miller W (2000) PipMaker - a web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577-586.
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, and Miller W (2003) Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103-107.
Shaner MC, Blair IM, and Schneider TD (1993) Sequence logos: a powerful, yet simple, tool. *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*,

Los Alamitos, 813-821.
Shannon CE (1948) A mathematical theory of communication. *Bell Sys Tech J.* **27:** 379-423, 623-656.
Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucl Acids Res.* **29:** 308-11.
Shiu SH, Byrnes JK, Pan R, Zhang P, and Li WH (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc Natl Acad Sci.* **103:** 2232-6.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034-1050.
Sigurdsson MI, Smith AV, Bjornsson J, and Jonsson J(2009) HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Res.* **19:** 581-589.
Sinha S and Tompa M (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucl Acids Res.* **30:** 5549-5560.
Sinha S, Blanchette M, and Tompa M (2004) PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5:** 170-186.
Slotkin RK and Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet.* **8:** 272-85.
Smith NG and Hurst LD (1998) Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. *Genetics* **150:** 823-33.
Smith NG and Hurst LD (1999) The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics* **152:** 661-73.
Smith RJ, Dean W, Konfortova G, and Kelsey G (2003) Identification of novel imprinted genes in a genome-wide screen for maternal methylation. *Genome Res.* **13:** 558-69.
Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D, Matthews L, Rogers J, Pask AJ, Shaw G, VandeBerg JL, et al. (2008) Conservation of the *H19* noncoding RNA and *H19-IGF2* imprinting mechanism in therians. *Nature Genet.* **40:** 971-6.
Smrzka OW, Faé I, Stöger R, Kurzbauer R, Fischer GF, Henn T, Weith A, and Barlow DP (1995) Conservation of a maternal-specific methylation signal at the human *IGF2R* locus. *Hum Mol Genet.* **4:** 1945-1952.
Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, and Held WA (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc Natl Acad Sci.* **102:** 3336-3341.
Spillane C, Schmid KJ, Laoueille-Duprat S, Pien S, Escobar-Restrepo JM, Baroux C, Gagliardini V, Page DR, Wolfe KH, and Grossniklaus U (2007) Positive darwinian selection at the imprinted *MEDEA* locus in plants. *Nature* **448:** 349-352.
Steinhoff C, Paulsen M, Kielbasa S, Walter J, and Vingron M (2009) Expression profile and transcription factor binding site exploration of imprinted genes in human and mouse. *BMC Genomics* **10:** 144.
Strichman-Almashanu LZ, Lee RS, Onyango PO, Perlman E, Flam F, Frieman MB, and Feinberg AP (2002) A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res.* **12:** 543-554.
Studer RA and Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* **25:** 210-6.
Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci.* **101:** 6062-6067.
Surani MA, Barton SC, and Norris ML (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature* **308:** 548-50.
Suzuki MM and Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet.* **9:** 465-76.
Suzuki S, Renfree MB, Pask AJ, Shaw G, Kobayashi S, Kohda T, Kaneko-Ishino T, and Ishino F

(2005) Genomic imprinting of *IGF2*, *p57^{KIP2}* and *PEG1/MEST* in a marsupial, the tammar wallaby. *Mech Dev.* **122:** 213-222.

Suzuki S, Ono R, Narita T, Pask AJ, Shaw G, Wang C, Kohda T, Alsop AE, Marshall Graves JA, Kohara Y, et al. (2007) Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genet.* **3:** e55.

Szabó PE, Tang SH, Silva FJ, Tsark WM, and Mann JR (2004) Role of CTCF binding sites in the *Igf2/H19* imprinting control region. *Mol Cell Biol.* **24:** 4791-800.

Takada S, Paulsen M, Tevendale M, Tsai C-E, Kelsey G, Cattanach BM, and Ferguson-Smith AC (2002) Epigenetic analysis of the *Dlk1-Gtl2* imprinted domain on mouse chromosome 12: implications for imprinting control from comparison with *Igf2-H19*. *Hum Mol Genet.* **11:** 77-86.

Takai D and Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci.* **99:** 3740-3745.

Tazi J and Bird AP (1990) Alternative chromatin structure at CpG islands. *Cell* **60:** 909-920.

Thompson JD, Gibson TJ, and Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics.* **Chapter 2:** Unit 2.3.

Tierling S, Dalbert S, Schoppenhorst S, Tsai C-E, Oliger S, Ferguson-Smith AC, Paulsen M, and Walter J (2006) High-resolution map and imprinting analysis of the *Gtl2-Dnchc1* domain on mouse chromosome 12. *Genomics* **87:** 225-235.

Tierling S, Gasparoni G, Youngson N, and Paulsen M (2009) The *Begain* gene marks the centromeric boundary of the imprinted region on mouse chromosome 12. *Mamm Genome*, in press.

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotech.* **23:** 137-144.

Tucker KL, Beard C, Dausmann J, Jackson-Grusby L, Laird PW, Lei H, Li E, and Jaenisch R (1996) Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. *Genes Dev.* **10:** 1008-20.

Tuiskula-Haavisto M and Vilkki J (2007) Parent-of-origin specific QTL - a possibility towards understanding reciprocal effects in chicken and the origin of imprinting. *Cytogenet Genome Res.* **117:** 305-12.

Umlauf D, Goto Y, Cao R, Cerqueira F, Wagschal A, Zhang Y, and Feil R (2004) Imprinting along the *Kcnq1* domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. *Nature Genet* **36:** 1296-300.

van Helden J, André B, and Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol.* **281:** 827-842.

van Helden J, Rios AF, and Collado-Vides J (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl Acids Res.* **28:** 1808-1818.

Varrault A, Gueydan C, Delalbre A, Bellmann A, Houssami S, Aknin C, Severac D, Chotard L, Kahli M, Le Digarcher A, et al. (2006) Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth. *Dev Cell* **11:** 711-22.

Vicoso B and Charlesworth B (2006) Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet.* **7:** 645-53.

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, and Martienssen RA (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297:** 1833-1837.

Vu TH, Li T, and Hoffman AR (2004) Promoter-restricted histone code, not the differentially methylated DNA regions or antisense transcripts, marks the imprinting status of *IGF2R* in human and mouse. *Hum Mol Genet.* **13:** 2233-2245.

Walter J and Paulsen M (2003) The potential role of gene duplications in the evolution of imprinting mechanisms. *Hum Mol Genet.* **12:** R215-R220.

Walter J, Hutter B, Khare T, and Paulsen M (2006) Repetitive elements in imprinted genes. *Cytogenet Genome Res.* **113:** 109-115.

Wan LB, Pan H, Hannenhalli S, Cheng Y, Ma J, Fedoriw A, Lobanenkov V, Latham KE, Schultz

RM, and Bartolomei MS (2008) Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development* **135:** 2729-38.

Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, and Clark AG (2008) Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS ONE* **3:** e3839.

Wang Y, Joh K, Masuko S, Yatsuki H, Soejima H, Nabetani A, Beechey CV, Okimani S, and Mukai T (2004) The mouse *Murr1* gene is imprinted in the adult brain, presumably due to transcriptional interference by the antisense-oriented *U2af1-rs1* gene. *Mol Cell Biol.* **24:** 270-279.

Wang Z, Fan H, Yang HH, Hu Y, Buetow KH, and Lee MP (2004) Comparative sequence analysis of imprinted genes between human and mouse to reveal imprinting signatures. *Genomics* **83:** 395-401.

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT, et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453:** 175-83.

Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet.* **5:** 276-287.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Weidman JR, Murphy SK, Nolan CM, Dietrich FS, and Jirtle RL (2004) Phylogenetic footprint analysis of *IGF2* in extant mammals. *Genome Res.* **14:** 1726-1732.

Weidman JR, Dolinoy DC, Maloney KA, Cheng JF, and Jirtle RL (2006) Imprinting of opossum *Igf2r* in the absence of differential methylation and air. *Epigenetics* **1:** 49-54.

Wen B, Wu H, Bjornsson H, Green RD, Irizarry R, and Feinberg AP (2008) Overlapping euchromatin/heterochromatin-associated marks are enriched in imprinted gene regions and predict allele-specific modification. *Genome Res.* **18:** 1806-13.

Wilkins JF and Haig D (2001) Genomic imprinting of two antagonistic loci. *Proc Biol Sci.* **268:** 1861-7.

Wilkins JF and Haig D (2002) Parental modifiers, antisense transcripts and loss of imprinting. *Proc Biol Sci.* **269:** 1841-6.

Wilkins JF and Haig D (2003) What good is genomic imprinting: the function of parent-specific gene expression. *Nat Rev Genet.* **4:** 359-68.

Wood AJ and Oakey RJ (2006) Genomic imprinting in mammals: emerging themes and established theories. *PLoS Genetics* **2:** e147.

Wood AJ, Roberts RG, Monk D, Moore GE, Schulz R, and Oakey RJ (2007) A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PLoS Genetics* **3:** e20.

Wutz A and Barlow DP (1998) Imprinting of the mouse *Igf2r* gene depends on an intronic CpG island. *Mol Cell Endocrinol.* **140:** 9-14.

Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, and Lander ES (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci.* **104:** 7145-50.

Xing Y and Lee C (2006) Can RNA selection pressure distort the measurement of Ka/Ks? *Gene* **370:** 1-5.

Xu N, Donohoe ME, Silva SS, and Lee JT (2007) Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein. *Nature Genet.* **39:** 1390-6.

Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, Iwasaka T, Mukai T, Sakaki Y, and Ito T (2004) A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Res.* **14:** 274-266.

Yamashita R, Suzuki Y, Sugano S, and Nakai K (2005) Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350:** 129-136.

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* **15:** 568-73.

Yang Z and Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol.* **15:** 496-503.

Yang Z and Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* **17:** 32-43.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* **24:** 1586-1591.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, et al. (2003) Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnol.* **21:** 379-86.

Yoder JA, Walsh CP, and Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13:** 335-340.

Yoon B, Herman H, Hu B, Park YJ, Lindroth A, Bell A, West AG, Chang Y, Stablewski A, Piel JC, et al. (2005) *Rasgrf1* imprinting is regulated by a CTCF-dependent methylation-sensitive enhancer blocker. *Mol Cell Biol.* **25:** 11184-90.

Yusufzai TM, Tagami H, Nakatani Y, and Felsenfeld G (2004) CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell* **13:** 291-8.

Zhang J (2003) Evolution by gene duplication: an update. *Trends Ecol Evol.* **18:** 292-298.

Zhang Y, Liu XS, Liu QR, and Wei L (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucl Acids Res.* **34:** 3465-75. Print 2006.

Zhao Z and Zhang F (2006a) Sequence context analysis in the mouse genome: Single nucleotide polymorphisms and CpG island sequences. *Genomics* **87:** 68-74.

Zhao Z and Zhang F (2006b) Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. *Gene* **366:** 316-324.

Zilberman D, Gehring M, Tran RK, Ballinger T, and Henikoff S (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet.* **39:** 61-69.

Web references

Genome Information

dbSNP http://www.ncbi.nlm.nih.gov/SNP/

Ensembl http://www.ensembl.org

Epigenetic scores http://neighborhood.bioinf.mpi-inf.mpg.de/CpG_islands_revisited

European Bioinformatics Institute http://www.ebi.ac.uk

GenBank http://www.ncbi.nlm.nih.gov/Genbank/index.html

Genetic Information Research Institute http://www.girinst.org

HomoloGene http://www.ncbi.nlm.nih.gov/homologene

MapViewer http://www.ncbi.nlm.nih.gov/mapview/

Mouse Genome Informatics Nomenclature Committe
http://www.informatics.jax.org/mgihome/nomen/gene.shtml

NCBI http://www.ncbi.nlm.nih.gov

ORegAnno http://www.oreganno.org/oregano/Index.jsp

Sanger Institute http://www.sanger.ac.uk

TransFac at Biobase http://www.gene-regulation.com/pub/databases.html

UCSC Genome Browser http://genome.ucsc.edu

UCSC Wiki http://genomewiki.cse.ucsc.edu/

Imprinting

Imprinted Gene Catalogue at the University of Otago http://igc.otago.ac.nz/home.html
Mouse Imprinting at MRC Harwell http://www.har.mrc.ac.uk/research/genomic_imprinting/

Programs

Blat http://genome.ucsc.edu/cgi-bin/hgBlat
Blast http://blast.ncbi.nlm.nih.gov/Blast.cgi
Blastz http://www.bx.psu.edu/miller_lab/
ClustalW http://www.ebi.ac.uk/Tools/clustalw2/index.html
cpg http://www.nslij-genetics.org/wli/dnaseg/
CpG cluster http://bioinfo2.ugr.es/CpGcluster/
CpG Island Searcher http://cpgislands.usc.edu
cross_match alignment tool http://www.phrap.org
DCODE http://www.dcode.org
Entrez Programming Utilities http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
MEME http://meme.sdsc.edu/meme4/
PAML http://abacus.gene.ucl.ac.uk/software/paml.html
PipTools package http://www.bx.psu.edu/miller_lab/
PipMaker http://bio.cse.psu.edu/pipmaker (http://pipmaker.bx.psu.edu/pipmaker/)
Regulatory sequence analysis tools http://rsat.ulb.ac.be/rsat/
RepeatMasker http://www.repeatmasker.org
rVista http://www-gsd.lbl.gov/vista/
transAlign http://www.personal.uni-jena.de/~b6biol2/ProgramsMain.html#Sequences
Tandem Repeats Finder http://tandem.bu.edu/trf/trf.html
Washington University Blast (*WuBlast*) http://blast.wustl.edu
WebLogo http://weblogo.berkeley.edu

Other tools

BioMart http://biomart.org
EpiGraph http://epigraph.mpi-inf.mpg.de/WebGRAPH/
Perl htpp://www.perl.org
R htpp://www.r-project.org
Random number generator http://www.random.org
Spearman's correlation http://www.sics.se/~jussi/Vertyg/spearman.html
Wilcoxon test http://www.fon.hum.uva.nl/rob/SignedRank/WlcxTest.pl