universität wien

MASTERARBEIT

Titel der Masterarbeit

„Imprinted expression of known and novel transcripts in multiple tissues of the mouse“

verfasst von

Christoph Dotter, Bakk, BSc

angestrebter akademischer Grad

Master of Science (MSc)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 066 863
Studienrichtung lt. Studienblatt:	Masterstudium Biologische Chemie
Betreut von:	Prof. Denise Barlow

Contents

Table of contents ii

1 Introduction 1

1.1 Genomic imprinting 1

1.1.1 Identification of genes showing imprinted gene expression 2

1.1.2 The function of genomic imprinting in mammals 4

1.1.3 Organisation of genes showing imprinted expression 5

1.1.4 Tissue-specificity of genomic imprinting 7

1.1.5 Regulation of genomic imprinting in imprinted clusters 8

1.2 Non-coding RNAs 10

1.2.1 Non-coding RNAs in imprinted regions 12

1.2.2 Methods for the determination of non-coding potential 13

1.3 Aims of this project 15

2 Results 17

2.1 Preliminary tests of primer design - *Igf2r*, *Airn* and *Impact* 17

2.2 Preparation of PCR samples for known imprinted protein coding genes 21

2.2.1 Candidate selection and primer design 21

2.2.2 PCR for known imprinted targets 24

2.3 Preparation of PCR samples for non-coding candidates in imprinted regions. . 31

2.3.1 Validation of the non-coding pipeline 31

2.3.2 Candidate selection and primer design 34

2.3.3 PCR for non-coding targets in eight tissues 37

2.3.4 PCR for non-coding targets in testes/placenta 43

2.4 Massive parallel sequencing of PCR products from four tissues 44

2.4.1 Sonication and library preparation 46

2.4.2 Analysis of sequencing results 49

3 Discussion 57

3.1 Evaluation of the imprinting status of *Igf2r* and *Airn* agrees with previous findings 57

3.2 The targeted PCR approach 58

3.2.1 The targeted approach is feasible and cost efficient but more time consuming than Whole Transcriptome Sequencing methods 59

3.2.2 PCR results are highly reproducable across biological replicates 60

3.2.3 Visual quality checks by agarose gel electrophoresis give a good estimate on sequencing coverage 60

3.3 Coverage of PCR products 62

3.3.1 Known imprinted protein coding candidates were well covered 62

3.3.2 Non-coding candidates showed tissue specific differences in coverage . 63

3.3.3 Reactions with intronic primers yielded more product than their exonic counterparts 64

3.3.4 The majority of *de novo* assembled splice junctions could be validated . 65

3.3.5 Sequencing results revealed additional exons as well as possibly unspliced transcripts 67

3.4 Future perspectives 69

4 Methods 71

5 Materials 78

List of figures/tables 84

Bibliography 87

Abstract 97

Zusammenfassung 98

Curriculum vitae 98

Acknowledgements 100

1 Introduction

1.1 Genomic imprinting

Genomic imprinting is an example of an epigenetic phenomenon which is characterized by the parent-of-origin specific expression of affected genes where only one of the two alleles is expressed while the other one is silent [1]. This is in contrast to the rules of Mendelian expression in diploid organisms where both alleles are expressed equally [2]. Epigenetics is defined as heritable effects on gene expression not caused by changes in the DNA sequence [3]. Inbred mice have genetically identical chromosomes and the fact that genomic imprinting can be observed in inbred mice provided evidence that it is established and maintained by an epigenetic mechanism and that the imprint must be defined before the embryo becomes diploid, namely during gamete formation. Other features of genomic imprinting include that it is a *cis*-acting mechanism, meaning that the imprinting mechanism only influences one chromosome but not the other, and that genomic imprinting arises as a consequence of inheritance and not sex. This means that the expressed allele is determined by whether it was inherited from the father or mother and is not influenced by the sex of the offspring [4]. Due to the fact that both the active and the inactive parental allele are located within the same nucleus, genomic imprinting is a good model for epigenetic gene regulation in mammals since they are both located within the same transcriptional environment [5].

So far genomic imprinting has been described in mammals, sciarid flies [6], coccid insects [7] as well as the endosperm of some seed-bearing plants, e.g. *Arabidopsis* [8] or maize [9]. First evidence for this regulatory phenomenon was provided by nuclear transfer experiments in the 1980s which demonstrated that the parental genomes are not functionally equivalent [10] and that both the maternal and the paternal genome are required for embryogenesis in mice [11, 12]. This was later demonstrated by the identification of the first imprinted genes *Igf2r* [13], *Igf2* [14] and *H19* [15]. Genomic imprinting was also shown to be the main obstacle to parthenogenesis in mammals as generation of mice with two maternal genomes was made possible by the introduction of deletions in the *Igf2* and *Dlk1* imprinted clusters. These deletions included the differentially methylated regions (DMRs) regulating imprinted expression in these clusters and therefore led to re-expression of *Igf2* and *Dlk1* and downregulation of *Gtl2* [16].

1.1.1 Identification of genes showing imprinted gene expression

After the initial discoveries of genes showing imprinted expression the question arose how many of these genes there are in total. Imprinted expression can be detected by simply mapping parental allele specific expression without knowing any mechanistic details about how this expression pattern is established at these loci. These mapping efforts were carried out by employing methods such as single-nucleotide polymorphism (SNP) microarrays [17] or by analysing differential transcription between parthenogenote and androgenote mouse embryos, which contain only the maternal and paternal genome, respectively [18]. While these approaches lacked sensitivity and specificity, the advent of RNA sequencing brought about a new method of identifying imprinted genes which solved these problems. RNA sequencing employs massive parallel sequencing technologies to enable analysis of the transcriptome, which is defined as the sum of all transcripts in a cell. The usual workflow includes isolation of target RNA, transcription into cDNA and creating a library which can be used for massive parallel sequencing. Cellular RNA is often depleted of ribosomal RNA and can also be further enriched for RNA molecules with poly(A) tails, e.g. mRNAs. The result is a high number of short sequences called reads which can then be aligned to a reference genome [19]. The number of reads correlates with the level of expression of the transcript.

RNA sequencing as a method for the detection of imprinted expression has several advantages over previous techniques, for example that it provides an easily quantifiable measurement for expression or that this method does not need an exact annotation of the transcript beforehand. A prerequisite for the analysis of imprinted expression by RNA sequencing is the ability to distinguish between reads from the paternal and maternal allele based on their sequence. For this purpose SNPs are used, which are single bases that differ between the two alleles. Quantification of reads supporting either one or the other allele then enables the calculation of a ratio representing the relative abundance of one allele over the other. The method therefore requires the transcriptome analysis of F1 offspring from two genetically distinct mouse strains. For example, the *castaneus* mouse strain Cast/EiJ is genetically distinct from the reference strain C57BL/6, with around 21 million SNPs between them (source: Sanger institute[1]).

To this date RNA-sequencing has been employed by many studies to detect genes showing imprinted expression. This was done mostly for early developmental stages by investigating embryonic tissues like MEFs [20], brain [21, 22] or whole-embryo

---

[1] ftp://ftp-mouse.sanger.ac.uk/REL-1303-SNPs_Indels-GRCm38/

[23], extra-embryonic tissues like placenta [24, 25], but also neonatal tissues (e.g. brain [26]). While most of these studies mainly confirmed the already known genes and found only a few novel ones, one study in 2010 reported over 1300 genes demonstrating parental biases in gene expression in the brain (embryonic brain, medial prefrontal cortex and hypothalamus of adult mice), which was around 10-fold higher than all previous reports and also included 347 genes showing sex-specific imprinted expression [21]. However, attempts to replicate these findings by repeating the experiment and analysis failed to validate most of the novel genes reported before. It was concluded that the initial results contained a large number of false positives and that statistical methods are necessary to correct for this. Since this can only be done empirically so far, independent validation of the results is necessary to confidently declare genes as showing imprinted expression.[22]. Another caveat of this method is the requirement of SNPs that can be used to distinguish between the two strains. If the DNA sequences of the used strains are too similar to each other it might not be possible to evaluate certain genes because there simply is no SNP located within these genes that would allow the assignment of reads to a specific allele. Therefore the number of genes showing imprinted expression identified in an experiment might be greatly reduced if there are not enough SNPs between the two used strains [27]. For example the imprinting status of *Igf2r* could not be assessed in this study since there were no SNPs which were exclusive to this gene. Regarding the statistical method used to determine the significance of the parental bias most studies used an approach based on binomial distributions, although a recent study suggested a different model based on joint modeling of strain and parent of origin effects, stating that this method is less error-prone when detecting imprinted expression [28].

Several online resources summarise the current state of research regarding genes showing imprinted expression and report slightly different numbers of genes. For example the Web Atlas of Murine genomic Imprinting and Differential EXpression (WAMIDEX)[2] [29] declares 104 genes as showing imprinted expression, the Catalogue of Parent of Origin Effects from the university of Otago[3] [30] lists 122 entries and the MRC Harwell Imprinting Web pages[4] [31] report 150 genes. Since all of these resources compiled results from multiple sources the different numbers are mainly because of the use of different sources. A solution to come up with one final number would be to use one method to find all imprinted genes in one system. A cost-efficient method to study imprinted genes in multiple tissues is required to make such a project more feasible.

---

[2] https://atlas.genetics.kcl.ac.uk/

[3] http://igc.otago.ac.nz/home.html

[4] http://www.mousebook.org/catalog.php?catalog=imprinting

1.1.2 The function of genomic imprinting in mammals

To determine the function of a regulatory phenomenon like genomic imprinting it is useful to first look at the function of genes regulated in this fashion. The function of many of the known imprinted genes has been elucidated by investigating the effects of a knock-out of the gene. These experiments showed many imprinted genes are involved in the growth regulation of embryo, placenta and neonate. On closer inspection of these genes it was discovered that they are divided into paternally expressed growth promoters (*Igf2*, *Peg1*, *Peg3*, *Rasgrf1* and *Dlk1*) and maternally expressed growth repressors (*Gnas*, *Igf2r*, *Cdkn1c*, *H19* and *Grb10*). Another subset of genes showing imprinted expression regulates neurological processes which influence behaviour. This over-representation of growth and behaviour-associated genes especially supports two common theories on the function of genomic imprinting, the "parental conflict" and the "overian time-bomb" hypothesis. Further evidence for these two theories is provided by the fact that genomic imprinting in mammals is restricted to placental and marsupial mammals and that the paternal genome is necessary for fetal development [4].

The "parental conflict" hypothesis

This theory is based on a conflict of interest between the mother and father regarding the distribution of resources from the mother to the different offspring. While the former aims at a more equal distribution of maternal resources to multiple offspring and therefore increasing the number of offspring carrying the maternal genome, the latter strives to maximize the resources for the individual offspring carrying his paternal genome. This is also reflected in the functions of subsets of paternally and maternally expressed genes, which are responsible for enhancement and inhibition of growth, respectively [32]. Another subset of genes has been shown to be involved in regulating behaviour, in particular behaviour linked with demand for resources [33]. One example for this is the regulation of suckling behaviour which has been partly associated with *GnasXL* [34] on chromosome 2, as well as *Peg3* and *Dlk1* on chromosomes 7 and 12, respectively [35, 36], all of which are expressed exclusively from the paternal allele. This supports the theory that paternally expressed imprinted genes promote growth and survival of the individual offspring, ensuring the passing on of the paternal genome.

The "ovarian time-bomb" hypothesis

Since the female body is equipped for internal reproduction, a risk exists that spontaneous oocyte activation might lead to full embryonic development in the ovary (ovarian trophoblastic disease). This problem is not present in males since they lack the ability to reproduce internally. Therefore it was proposed that imprinting of genes required for placental development leads to the necessity of the paternal genome since these genes are not expressed from the maternal genome. On fertilization this genes would be expressed from the paternal genome and allow normal embryo development [37].

Many other non-conflict theories have been proposed, some of which predict sexually dimorphic imprinting, where expression from the paternal allele is favoured in male offspring and vice versa [38, 39], or evolution of imprinting as a mechanism for genomic defense against transposable elements [40]. Overall it appears that no theory can give a sufficient explanation for each case of imprinted expression. The ovarian time-bomb hypothesis, for example, fails to explain why genes not necessary for trophoblast development show imprinted expression or why imprinted expression is still maintained in later stages of development. Therefore it has been proposed that genomic imprinting came into being for different reasons and also that imprinted expression had been established by different mechanisms. This would also account for the different imprinting status of some genes across different species [2].

1.1.3 Organisation of genes showing imprinted expression

There are several key requirements for a mechanism to act as the basis for genomic imprinting. It has to be able to influence transcription, it must be stably inherited through mitosis and it has to remain on the same parental chromosome after fertilization to allow parental specific control of expression by a *cis*-acting mechanism. The imprint also has to be erasable, since the imprint in the gametes needs to be erased and re-established according to the sex of the embryo [4, 41].

The only epigenetic mechanism fulfilling these criteria is DNA methylation, which is defined as the methylation of cytidine residues in 5'CpG3' dinucleotides and is established and maintained by de novo and maintenance methyltransferases, respectively. Cytidine methylation fulfils the aforementioned criteria in that it can be established in a sex specific way [4, 42, 43], can be stably propagated through the use of maintenance methyltransferases and it can be erased by either replication without maintenance methylation (passive demethylation), or through enzymes with demethylating activity (active demethylation). Furthermore DNA methylation has been shown to be associated with transcriptional repression [44].

Name (location)	**mat. expr. genes**	**pat. expr. genes**	**Cluster size / kb**	**lncRNAs**
Gnas (chr2)	2 pc,	4 nc/1 pc	80	Nespas, Exon1A
Igf2 (chr7)	1 nc	2 pc	80	H19
Kcnq1 (chr7)	11 pc	1 nc	780	Kcnq1ot1
Pws (chr7)	2 pc	>7 nc/pc	3700	Ipw, Zfp127a, PEC2, PEC3, Pwcr1
Grb10 (chr11)	2 pc	2 pc	780	-
Dlk1 (chr12)	>1 nc	4 pc	830	Gtl2, Rian, Rtl1as, Mirg
Igf2r (chr17)	3 pc	1 nc	490	Airn

Table 1: Examples for murine imprinted gene clusters [4]. This table gives an overview over seven well studied clusters of imprinted genes in the mouse. The data shown includes the name of the cluster, which is equal also the name of the respective principle imprinted mRNA gene, the number of maternally and paternally expressed genes, the size of the cluster in kilobases and the long non-coding RNAs located within the cluster; M, maternal; P, paternal; pc, protein coding; nc: noncoding

Eighty percent of the imprinted genes identified so far in the mouse are organized in 16 clusters [45]. This also provided further possible evidence for *cis* regulation of imprinted genes by a shared DNA element [4]. However, other imprinted genes are located outside these clusters. *Mcts2*, for example, was created by retrotransposition and is located within the intron of H13, which also shows imprinted expression [46]. Seven of the known clusters, i.e. *Igf2r*, *Kcnq1*, *Pws*, *Gnas*, *Grb10*, *Igf2* and *Dlk1*-cluster, named after their respective principle imprinted mRNA gene, have been thoroughly investigated. These clusters have been shown to contain at least three imprinted genes (table 1) [4]. Regarding the control of imprinted expression in these clusters a common feature has been found in the existence of a DNA sequence carrying a gametic differentially methylated region (gametic DMR). The gametic DMR is established in one gamete and stays on this specific parental chromosome in the diploid state throughout all developmental stages. These DMRs have been shown to regulate the imprinted expression of genes within the respective cluster. An unmethylated gDMR has been associated with repression of expression of imprinted genes from this allele. A methylated gDMR on the other hand is associated with expression of these genes. This was demonstrated by experiments deleting the gDMR, which lead to a loss of imprinted expression, i.e. biallelic expression, when the gDMR was deleted on the allele where it is unmethylated. On the other hand, deletion of the methylated gDMR had no effect on the imprinted expression of genes in the cluster. Therefore these gametic DMRs were declared as imprinting control elements (ICE) [3].

1.1.4 Tissue-specificity of genomic imprinting

When comparing genes with imprinted expression between human and mouse it becomes apparent that the imprinting status is not always conserved, with some genes being imprinted in one organism while being biallelically expressed in the other [30]. An example of this is *Igf2r*, which is biallelically expressed in human gene while the murine gene shows a maternal expression pattern [47]. For other genes the imprinted expression is even reversed in other species, e.g. *Zim2* is expressed from the maternal allele in mouse but from the paternal allele in human [48]. Furthermore imprinted expression of a gene might not be equal among all tissues of one organism. A survey analysing the data from the WAMIDEX [29] showed that of 115 genes analysed 23 only showed imprinted expression in a single tissue type but were biallelically expressed in others, 59 showed imprinted expression in at least two tissues and 33 could not be assessed since their imprinting status had only been investigated in one tissue type at the time of the survey. Of the 23 tissue specific imprinted genes thirteen show imprinted expression in extra-embryonic tissues (eleven in placenta, two in yolk sac), nine in brain and one in heart [33]. An example for a gene showing cell-type specific expression in brain is *Ube3a* which is only imprinted in neurons but shows biallelic expression in glial cells [49]. Another example is Dopa decarboxylase (*Ddc*) the isoform *Ddc_exon1a* of which was shown to be imprinted in heart, while being biallelic in brain [50]. A single cluster of imprinted genes can contain genes with different imprinting expression in different tissues. For example, the *Igf2r* cluster contains *Igf2r* and *Airn*, both of which show imprinted expression in almost all tissues, as well as *Slc22a2* and *Slc22a3*, two genes which only show imprinted expression in extra-embryonic tissues [51]. Another subset of genes only shows biallelic expression in one tissue and imprinted expression in all others. This behaviour was first discovered for *Igf2*, which shows biallelic expression in the choroid plexus as well as the leptomeninges [14]. Other examples include the relaxation of imprinted expression of *Igf2r* in post-mitotic neurons but not in glial cells [52] and the relaxation of *Dlk1* in niche astrocytes and neural stem cells [53]. *Kcnq1* also reverts to biallelic expression during mid-gestation in brain, cardiac lineages and kidney [54] which has been shown to coincide with the formation of chromatin loops and interaction of the *Kcnq1* promoter with tissue-specific enhancers [55].

Grb10 displays an unusual pattern of imprinted expression since it is usually exclusively expressed from the maternal allele and functions to repress growth during development and regulates energy homoeostasis and glucose metabolism [56, 57, 58]. However, initial studies showed biallelic expression of *Grb10* in brain [59] which was resolved into the paternally imprinted expression of a brain specific *Grb10* isoform which is expressed together with the maternally imprinted isoform. The brain specific isoform has a separate transcription start site overlapping a different CpG island [60].

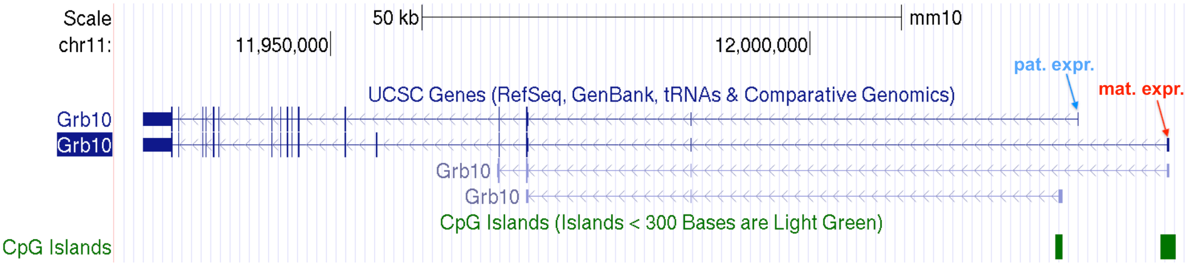

Scale
chr11:
50 kb
mm10
11,950,000
12,000,000
pat. expr.
mat. expr.
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Grb10
Grb10
Grb10
Grb10
CpG Islands (Islands < 300 Bases are Light Green)
CpG Islands

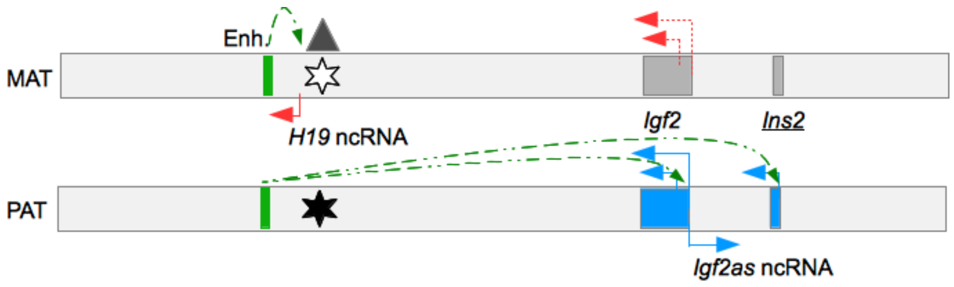
Figure 1: Different isoforms of Grb10 show different imprinting patterns. This figure shows a screenshot from the UCSC genome browser [61]. The data displayed is (from top to bottom): The scale in kilobases, the chromosomal location, gene annotations from the UCSC database [62] and a track displaying CpG island regions (green). Exons are depicted as filled rectangles while introns are shown as lines with arrows indicating the 5' to 3' direction. *Grb10* shows two isoforms, each originating at a different CpG island. One of these isoforms is mainly expressed in brain, where it shows paternal expression (blue arrow). The other isoform is the more common one expressed in most tissues and shows maternal expression (red arrow). The concurrent expression of both isoforms in brain has been interpreted as biallelic *Grb10* expression in brain, which could later be resolved by the use of isoform specific primers.

In summary tissue specificity of imprinted expression is one of the reasons why the analysis of different tissues enables a more precise evaluation of imprinted expression of a target gene since its expression might only be imprinted in very few tissues. Since a thorough analysis would require the study of many tissues, a cost-efficient method for the analysis of multiple tissues for specific candidates is required.

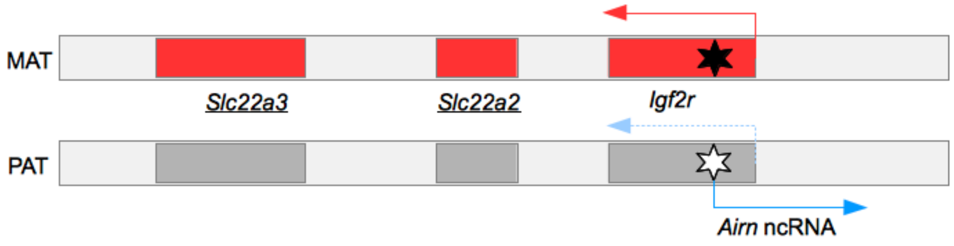
1.1.5 Regulation of genomic imprinting in imprinted clusters

Since imprinted expression shows developmental and/or tissue specific regulation it has been concluded that "printing" of the imprint at the ICE alone, i.e. methylation of CpG dinucleotides, is not enough to regulate transcription by itself but rather acts as a mark which is then recognized by other factors, called "readers" [63]. The "readers" then mediate the regulation of transcription. This system enables tissue and/or developmental specific control of imprinted expression by controlling the expression of the "readers" while the imprint is present in all tissues and at all stages in development. Different types of "readers" are used throughout the clusters thereby separating them into two main groups depending on the regulatory mechanism.

(a) *Igf2* cluster



(b) *Igf2r* cluster



(c) *Kcnq1* cluster

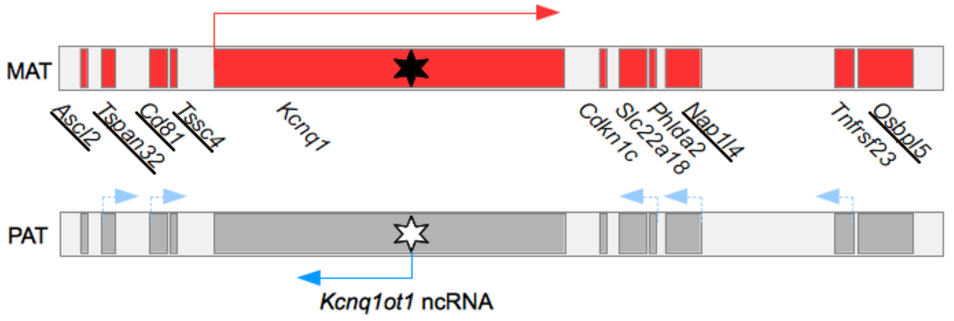


Figure 2: Schematic representation of imprinted clusters. Genes showing multi-lineage imprinted expression are shown in italics while genes with imprinted expression restricted to extra-embryonic tissues are underlined. The colors blue and red mark paternal and maternal expression, respectively, with dotted arrows in combination with grey boxes indicating weak expression from a repressed allele. The ICE of the cluster is depicted as a star and is black when methylated and white when unmethylated. In general protein coding genes have arrows above the scheme while expression of non-coding RNAs is illustrated below. (a) shows the *Igf2* cluster as an example for the insulator model. The enhancer is marked in green and the CTCF binding factor is shown as a grey triangle. (b) and (c) show the *Ig2fr* cluster and the *Kcnq1* cluster as examples for the lncRNA-mediated silencing model. Adapted from [3]

The insulator mechanism

This model was first described for the *Igf2* cluster [64, 65] and was discovered by deleting the ICE located within this cluster [66]. The "reader" in this model is CTCF, which binds to the ICE in a methylation-sensitive manner. The ICE is located between the *H19* ncRNA and *Igf2*. When bound to the unmethylated ICE, CTCF blocks the interaction between the *Igf2* promoter and downstream enhancers which are then only able to interact with the *H19* promoter, leading to the maternal repression of *Igf2* and expression of *H19* from this allele. On the paternal allele CTCF is unable to bind to the methylated ICE leading to interaction between the enhancer and the *Igf2* promoter while *H19* is repressed. Figure 2 gives an overview of this mechanism in the *Igf2* cluster. This model has since then also been shown for the *Grb10* cluster [67].

The lncRNA-mediated silencing mechanism

For completeness a brief description of this model is given here. More details on long non-coding RNAs (lncRNAs) will be provided in later chapters. This model employs transcription factors as the "readers" of the imprint. The ICE is located at the promoter region of a lncRNA and the lncRNA is only expressed if the ICE is unmethylated since methylation inhibits binding of transcription factors to this region. The long non-coding RNA then regulates transcription of the genes in the cluster. An example for this is the *Igf2r cluster*. This cluster includes three maternally expressed protein coding genes, i.e. *Igf2r*, *Slc22a2* and *Slc22a3*, as well as the paternally expressed lncRNA *Airn*. The promoter for *Airn* is located within an intron of *Igf2r* and *Airn* is transcribed in antisense to *Igf2r* (figure 2b. A truncation experiment reducing the lncRNA *Airn* from 108 to 3.7 kb was the first to show a role of lncRNAs in the regulation of imprinted expression. The truncated variant still showed paternal expression but silencing of *Igf2r*, *Slc22a2* and *Slc22a3* was abolished [51]. A silencing mechanism mediated by lncRNAs has also been demonstrated for the *Gnas* cluster [68], as well as the *Kcnq1* cluster, although other mechanisms might also play a role here [69, 70]. The *Kcnq1* cluster is illustrated in figure 2c. The figures for both the *Igf2r* and *Kcnq1* cluster show that the ICEs in these clusters overlap the promoter region of the respective lncRNA and that the lncRNA is only expressed from the allele where the ICE is unmethylated. The possible mechanisms by which these lncRNAs regulate the transcription of neighbouring genes include recruiting chromatin modifying enzymes and transcriptional interference and will be described in chapter 1.2.1.

1.2 Non-coding RNAs

The continuing improvement of methods for studying the transcriptome allowed the discovery that a large proportion of the human genome is transcribed into RNAs which are not coding for proteins, called non-coding RNAs (ncRNAs) [71]. In general these ncRNAs can be divided into two main groups. The first group, termed infrastructural ncRNAs, includes constitutively expressed RNAs such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs). These ncRNAs are mostly involved in RNA processing [72, 73]. The second group is named regulatory ncRNAs and the expression of ncRNAs in this group is often tissue specific and dependent on the developmental stage or is triggered as a response to environmental conditions [74]. This group includes short ncRNAs, such as Piwi-interacting RNAs (piRNAs) and micro RNAs (miRNAs), which are involved in the RNA interference (RNAi) mechanism [75], as well as long non coding RNAs (lncRNAs), which are usually defined as being larger than 200 nucleotides. Since the discovery that there are at least as many lncRNAs as protein coding genes, the question whether these transcripts are functional or not has been subject to debate [76]. Although the function of most lncRNAs is unknown there are some examples of functional lncRNAs. In general lncRNAs have been associated with gene regulatory functions through the formation of ribonucleic-protein complexes. By binding to regulatory proteins these lncRNAs are able to enhance or restrict the recruitment of these regulatory factors to certain chromatin regions [77]. One example for a regulatory lncRNA interacting with a protein is *SRA*, which enhances the insulator function of the CTCF protein involved in the regulation of imprinted expression at the *Igf2* cluster [78].

Functional lncRNAs can be divided into two groups, depending whether they act in *cis* or in *trans*, meaning on other chromosomes as well. One way that lncRNAs can work in *trans* is by influencing RNA polymerase II (RNAPII) activity, for example the inhibition of RNAPII phosphorylisation by the heat-shock induced *B2* lncRNA in mouse [79]. The *trans*-acting lncRNA *HOTAIR* in human shows a different mode of action as it is interacting with the Polycomb repressive complex 2 (PRC2) to aid in the establishment of the repressive H3 lysine 27 trimethylation (H3K27me3) mark at the *HOXD* cluster [80]. *Cis* acting lncRNAs, on the other hand, can be divided into two classes by their mode of action. The first mechanism suggests that the lncRNA product mediates the regulatory function. One well examined example of a lncRNA employing this mechanism is *Xist*, which is required for X chromosome dosage compensation in mammals [81]. The *Xist* lncRNA is necessary for the silencing of genes on the inactive X chromosome. *Xist* is also involved in the recruitment of polycomb complex 2 to the inactive X chromosome, although the role of PRC2 in X chromosome inactivation is still unclear [82]. Another example for a lncRNA product mediating a regulator function is the *HOTTIP* lncRNA which is expressed from a locus within the *HOXA* cluster and positively regulates neighbouring genes by the recruitment of enzymes responsible for trimethylation of H3 lysine 4 (H3K4me3) at their promoter regions [83]. A general explanation for the *cis*-acting nature of lncRNA products is that they recruit the respective regulatory factor while they are still attached to the elongating RNAPII, a mode of action named "tethering", which restricts the effect to the same chromosome [84, 85]. The second *cis*-mechanism, termed transcriptional interference (TI), proposes regulatory function through the act of transcription itself. TI was shown for lncRNAs overlapping promoters of their regulatory targets, although in general TI could work by transcription of the lncRNA across any regulatory sequence [84]. Possible ways in which TI could influence these regions are by repositioning of promoter nucleosomes, the establishment of histone modifications at promoter regions, both of which have been demonstrated in yeast [86, 87]. Another mechanism has been recently demonstrated for the imprinted ncRNA *Airn*, which is responsible for silencing the paternal allele of the maternally expressed *Igf2r* gene among other genes in this cluster. By creating truncations of *Airn* it has been shown that this silencing occurs in the absence of repressive chromatin and independent of DNA methylation but dependent on *Airn* transcription through the promoter region of *Igf2r* [88, 89].

1.2.1 Non-coding RNAs in imprinted regions

The regulation of imprinted expression in imprinted clusters can be mediated by either inhibiting the interaction of regulatory regions with the promoter of the target gene ("insulator model"), or via the action of a macro non-coding RNAs, which are defined as mainly unspliced long ncRNAs [90]. Four of the aforementioned clusters have been well studied, i.e the *Igf2r*-, *Kcnq1*-, *Pws*- and the *Gnas*-cluster, and contain at least one lncRNA, i.e. *Airn*, *Kcnq1ot1*, *Ube3a-ats* and *Nespas*, respectively, which overlap the respective protein coding gene in antisense. In contrast to this the lncRNA *H19* in the *Igf2* cluster does not overlap *Igf2* and it has been shown that *H19* is not necessary for imprinted expression of *Igf2* [91] as can be explained by the aforementioned insulator model.

Both *Airn* and *Kcqn1ot1* have been well studied and it has been established that expression of these lncRNAs is necessary for the repression of *Igf2r* and *Kcnq1*, respectively [69, 70, 51]. While it has been shown recently that *Airn* mediates silencing of *Igf2r* by transcriptional interference [88] the exact way in which *Kcnq1ot1* silences *Kcqn1* and other genes in this cluster has yet to be determined. Earlier studies suggest that *Kcnq1ot1* regulates gene expression in the cluster by interacting with Polycomb proteins to create a repressive chromatin environment [92]. In contrast to this, a recent study proposed a mode of action where *Kcnq1ot1* acts as a sort of scaffold which mediates the formation of an intrachromosomal loop involved in the regulation of imprinted expression of *Kcnq1* [93]. Although the way in which *Airn* acts in the regulation of *Igf2r* has been elucidated, the regulatory mechanism for the two remaining imprinted protein coding genes in the *Igf2r*-cluster, i.e. *Slc22a2* and *Slc22a3*, remains to be determined. Both of this genes only show imprinted expression in extra-embryonic tissues and are therefore termed extra-embryonic lineage-specific (EXEL) genes [94]. It has been shown that G9a, a histone lysine 9 methyltransferase is required for imprinted expression of *Slc22a3* and that *Airn* interacts with both G9a and the promoter region of *Slc22a3* suggesting a model where *Airn* recruits G9a towards the promoter region [95]. A different model which has been suggested but not shown so far is that *Airn* inhibits the formation of a chromosomal loop which would be necessary for activation of *Slc22a2* and *Slc22a3* expression [90]. The situation within the large *Pws* cluster is not yet fully understood although it has been indicated that the lncRNA *Ube3a-ats* which is probably a part of larger transcript together with *Ipw* [96], is involved in the regulation of imprinted expression of *Ube3a* [97]. Further evidence was provided by a recent study showing that a truncation of *Ube3a-ats* leads to reactivation of *Ube3a* and even improvement of Angelman syndrome related symptoms in the mouse model. The proposed mechanism was termed "transcriptional collision" and is based on the collision of the two polymerases transcribing *Ube3a* and *Ube3a-ats* followed by the stalling of transcription and dissociation of both polymerases [98].

Besides the main regulator lncRNAs in these clusters, e.g. *Airn*, only little is known about imprinted lncRNAs. Many lncRNA databases, e.g. NONCODE [99] do not offer a comprehensive annotation of the imprinting status. lncRNA databases that include a search for imprinted genes either only provide a limited amount of imprinted lncRNAs, e.g. 12 imprinted lncRNAs in the lncRNAdb [100], or make it difficult to separate coding from non-coding candidates, e.g. the ncRNA Expression Database (NRED) [101]. Another publication mentioned the setup of a specific database for imprinted ncRNAs but the website was unavailable at the time [102]. A more comprehensive description of this data in a easily browsable form is therefore still required to give a better overview on the field of imprinted ncRNAs. Another remaining question is whether there are still imprinted lncRNAs which have not been found yet and what the function of these lncRNAs is. One of the challenges to answer this question is to provide a thorough description of all ncRNAs which could then be analysed for imprinted expression.

1.2.2 Methods for the determination of non-coding potential

So far there have been many studies in different organisms with the aim of creating a comprehensive catalogue of non-coding RNAs in different species and reporting varying numbers of ncRNAs. One thing most of these studies have in common is the focus on long intervening non-coding RNAs (lincRNAs) which are defined as transcripts not overlapping the exons of other non-lincRNA or protein-coding genes, since these overlaps may give rise to complications in the analysis [103, 104, 105, 106]. Different techniques were used for identifying lncRNAs over the years, such as cDNA cloning [107], tiling microarrays [108] or RNA sequencing [103, 104, 105, 106]. These datasets can be combined with others, for example chromatin data to further improve the derived transcript models [108, 109]. One thing that distinguishes the classification of lncRNAs from classifying other types of RNAs is that they are mostly defined by negative descriptors, i.e. not fulfilling certain criteria. One positive descriptor is being a product of RNA polymerase II while an example for a negative descriptor is that they must not overlap other transcripts or code for proteins themselves [110].

Deciding whether a transcript is coding for a protein or not is not a trivial task. The simple presence of an open reading frame (ORF) that could be translated does not necessarily exclude a non-coding function, since a long transcript might by chance contain a putative ORF of a certain length. A classical minimum ORF cutoff was set to 300 nucleotides (or 100 codons) by the FANTOM consortium [107] based on the observation that most known proteins showed a length greater than 100 amino acids [111]. This arbitrary cutoff is however prone to misclassification since on the one hand very long non-coding RNAs can contain an ORF of that size by chance [112] and, on the other hand, the amount of proteins smaller than 100 amino acids has been estimated to be around 3700 in mouse [111]. It is therefore apparent that ORF size alone cannot be a sole criteria for distinguishing non-coding RNAs and mRNAs, so additional determinants have to be employed.

One widely used approach is to evaluate the conservation of the occurring ORFs focusing on similarity to related DNA sequences from other species (Coding Region Identification Tool Invoking Comparative Analysis, CRITICA [113]), the tendency that for functional protein coding regions the probability for synonymous base changes is higher than for non-synonymous changes (Codon Substitution Frequencies, CSF [114], PhyloCSF [115]) or the nucleotide composition/codon usage bias of the mRNA sequence (as applied in the Coding-Potential Assessment Tool, CPAT [116], based on the 'Fickett score" [117]). Additional information can be gained by analysing the in silico translated sequence of predicted ORFs and querying this sequence against a protein database to find sequence similarities with known proteins or to investigate if this sequence contains any known protein motifs. Tools for this task include BLASTX (as used in [118]), the Pfam protein families database [119] and the HMMER algorithms [120]. A summary of these criteria is given in figure 3 in the form of a possible pipeline employing the mentioned filtering steps.

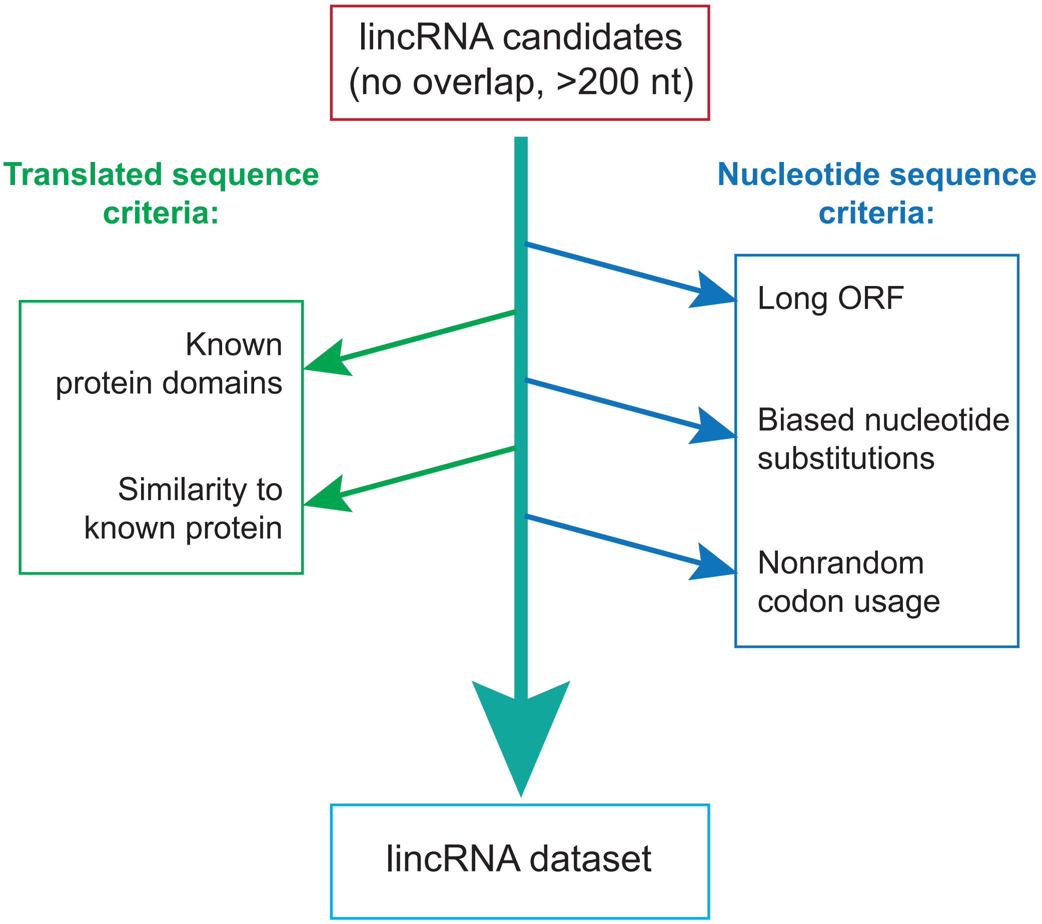

lincRNA candidates (no overlap, >200 nt)
Translated sequence criteria:
Nucleotide sequence criteria:
Known protein domains
Similarity to known protein
Long ORF
Biased nucleotide substitutions
Nonrandom codon usage
lincRNA dataset

Figure 3: A model pipeline illustrating common filtering steps involved in the selection of non-coding RNAs. The initial set of candidates is filtered using both nucleotide sequence criteria and criteria defined by the amino acid sequence of a hypothetical protein. Adapted from [110].

Since each of these criteria has its caveats, a possible solution could be the combination of multiple filters, one example being the Coding Potential Calculator (CPC) [121], which in turn are difficult to calibrate due to the lack of standards [110]. Another example, which was also the template for the pipeline used in this study, is the combination of a conservation approach, in this case PhyloCSF, with a query for known protein motifs in the Pfam database [103].

1.3 Aims of this project

The goal of this project is to establish a targeted PCR approach for the subsequent analysis of allelic expression in eight adult tissues of the mouse. Since imprinted expression can be regulated in a developmentally and/or tissue specific way [63], the analysis of eight different tissues enables the investigation of tissue specific imprinting in adult tissues. Two sets of transcripts were investigated. The first set consisted of known imprinted genes and served mainly as a control set to show that this approach works. In addition to that the analysis of this set will lead to a more comprehensive description of imprinted expression of these candidates since most of them have not been investigated in multiple adult tissues yet.

The second set was designed as a set of non-coding RNA (ncRNA) candidates gathered from a *de novo* assembly based on RNA sequencing data. This set contained both ncRNAs which have been demonstrated to show imprinted expression and novel candidates for which either no information about imprinted expression was known or which were not priorly annotated at all. The analysis of this set had two main goals. The first one was to provide a dataset for the subsequent analysis of imprinted expression of these candidates, which could lead to the discovery of novel imprinted ncRNAs. The second goal was to use this approach as a method to validate the *de novo* assembled exon models. To increase the number of candidates for this part of the analysis a set of candidates expressed in testes was also selected.

2 Results

The main goal of the work done in this thesis was to provide sequencing data for the subsequent analysis of imprintd expression. This was done using a targeted PCR based approach. Two sets of primers were designed for known imprinted targets and known and novel non-coding RNAs in imprinted regions. Eight adult mouse tissues were selected for this analysis to enable a more comprehensive description of imprinted expression in adult tissues. These tissues were brain, heart, kidney, leg muscle, liver, lung, spleen and thymus. Since the analysis of imprinted expression relied on SNPs to distinguish between the paternal and maternal allele F1 progeny of the strains Cast/EiJ and FVB/NJ were chosen as they have 20 million SNPs between them. The experimental setting was chosen so that both CxF and FxC progeny were used, the first letter indicating the strain of the mother and the second the strain of the father. These reciprocal crosses were necessary to correct for expression biases introduced by the sequence difference between the two alleles. In a single cross setting a bias like this would already be called imprinted while in the reciprocal setting a gene with this kind of bias would not as it would have a paternal bias in one cross and a maternal bias in the reciprocal cross. Two female replicates were used for each of the two crosses. These replicates were termed FxC f3/f4 and CxF f2/f3.

2.1 Preliminary tests of primer design - *Igf2r*, *Airn* and *Impact*

After implementing the primer design script preliminary tests were performed to validate that the designed primers work, i.e. yield the expected product, and that allele specific expression could be detected by evaluating the sequence information at the SNP positions. The selected candidates for these tests were *Impact*, *Igf2r* and *Airn*. The results of the primer design are shown for *Igf2r* as a representative example in figure 4. The image shows that three SNPs (orange) are located within the PCR product (highlighted in green) and can therefore be used to analyse imprinted expression by sequencing. Heterozygosity at the SNPs located in the PCR products of *Igf2r* and *Airn* was validated by PCR on genomic DNA. The primers used for *Igf2r* in this control are shown in figure 4, the primers for *Airn* were the same as used on cDNA as they already produced a PCR product coaligning with genomic DNA. Since there was no genomic DNA available for the samples used for the cDNA preparation, genomic DNA from a brain sample of a different FxC specimen was used as template. After preparing the cDNA for the tissues of FxC f4 and CxF f2, the PCR was done for the three candidates in all these samples (8 tissues, 2 reciprocal samples).

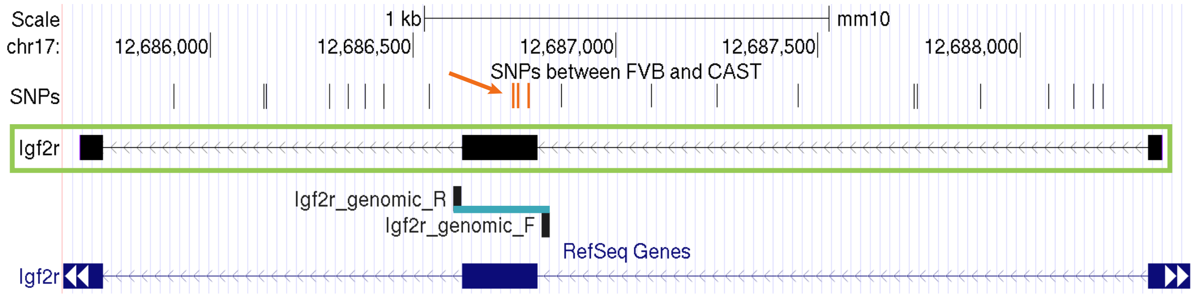


Figure 4: Illustration of automatically designed primers for Igf2r. This figure shows a UCSC screenshot as described in figure 1. Depicted are the SNPs between the two strains investigated (top), the PCR product for the primers of the initial test (green box) which spans two junctions, the primer pair designed to yield a product on genomic DNA (teal),and the RefSeq annotation of *Igf2r* in this area (bottom, blue). The track for the PCR product shows how this product aligns to genomic DNA, the black rectangles represent the cDNA sequence. SNPs covered by the initial PCR product (orange arrow) are also covered by the PCR product of the primer pair for the genomic DNA control.

All *Igf2r* reactions showed similar amounts of PCR product for all 16 samples (figure 5) after adjusting the amount of template to 3-9 µl. A second batch of cDNA was used for the *Airn* PCR since it did not work as well for the first batch.

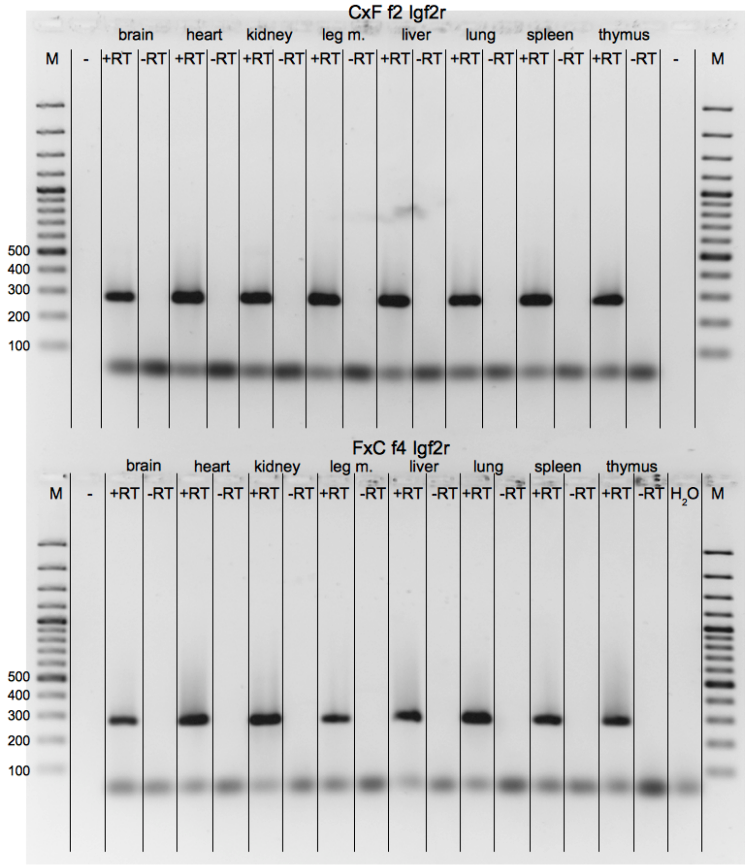


Figure 5: PCR results for *Igf2r* in eight tissues each of two mice. The designed *Igf2r* primers for the theoretical PCR product shown in figure 4. The gel shows that the PCR worked equally well for all samples and products were of the right size (279 bp). PCR was done with both normally prepared cDNA (+RT) and minus RT control (-RT) as well as a control reaction without template (H_2O). Sizes of marker fragments are indicated on the left hand side (in bp). Marker: GeneRuler 100 bp Plus DNA Ladder (Thermo Scientific); leg m., leg muscle

All 16 samples for *Igf2r* and *Airn*, as well as the genomic DNA control samples and 14 samples for *Impact* were sequenced by Sanger sequencing. The two samples for *Impact* in leg muscle were not sequenced since the reactions showed no PCR product on the gel. Figure 6 shows representative results of this sequencing. Two SNPs are shown for each candidate. The SNP variants for the two SNPs for *Airn* were A/C and C/T, respectively, the first one being the Cast/EiJ variant while the second represents the FVB/NJ variant. The results for *Airn* demonstrated that the crosses showed one SNP variant at the marked position and that this variant was different between the two crosses. The variants were validated by the genomic DNA control (figure 6, bottom left) and the fact that both variants were observed in the two reciprocal crosses. Both SNPs showed the paternal variant in both crosses (C+T in CxF, A+C in FxC), which is in agreement with the known paternal specific expression of *Airn* in mouse.

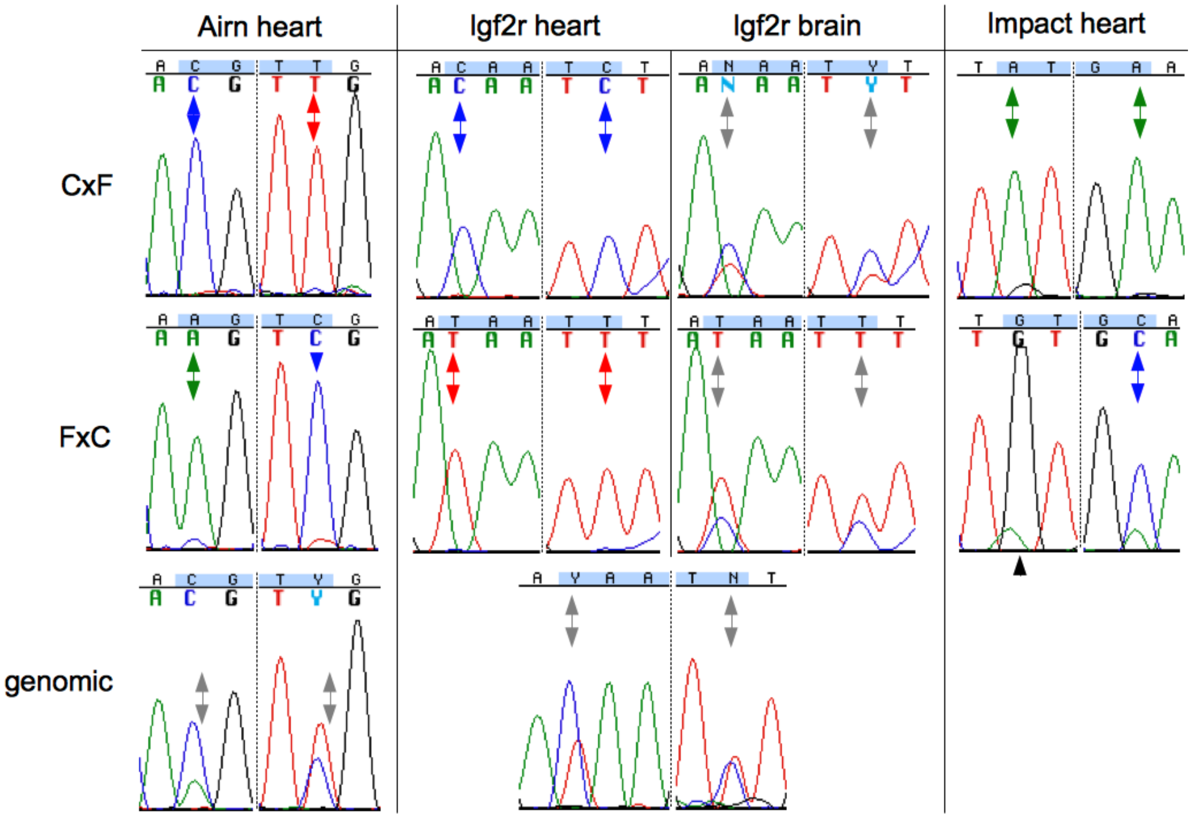


Figure 6: Representative Sanger sequencing results for *Airn*, *Igf2r* and *Impact*. Shown are the results in heart for *Airn*, heart and brain for *Igf2r* and heart in *Impact* (left to right) for both the CxF and the FxC cross as well as the genomic DNA sample (top to bottom). Two SNPs are shown for all three candidates as indicated by the arrows. The arrows are color coded according to the respective base, grey meaning that two variants are present at this position. SNP-variants: *Airn*: A/C + C/T, *Igf2r*: C/T + C/T, *Impact*: G/A + C/A for Cast/EiJ / FVB/NJ, respectively. All candidates show the expected parent-of-origin specific expression in the crosses as *Airn* showed FVB variants C+T in the CxF sample and Cast variants A+C in the FxC sample, corresponding to paternal expression. *Impact* also showed a pattern of paternal expression (A+A in CxF, G+C in FxC), while *Igf2r* is maternally expressed (C+C in CxF, T+T in FxC) in heart. The results also indicate relaxation of imprinted *Igf2r* expression in brain.

The annotated SNP variants for the two SNPs in *Igf2r* were C (Cast)/T (FVB) for both SNPs, which could be again validated in the genomic DNA control (figure 6, bottom middle). The results for *Igf2r* in heart (figure 6, column 2) showed different single SNP variants at the indicated positions in each of the crosses. Both crosses only showed the maternal variant (C+C in CxF and T+T in FxC) at both SNP positions, which confirmed the known maternal expression of murine *Igf2r*. The results in brain (figure 6, column 3) showed a different picture. While a bias towards the maternal allele can still be observed, a clear relaxation of imprinted expression was indicated by these results which corresponds to the relaxation of imprinted *Igf2r* expression in post-mitotic neurons, not glial cells [52]. The SNP variants for *Impact* were G (Cast)/A (FVB) and C (Cast)/A (FVB). Sequencing results for showed monoallelic expression at the SNP positions. Both crosses only showed the paternal variant (A+A in CxF, G+C in FxC) for both crosses. These findings showed that *Impact* was also correctly identified as being paternally expressed.

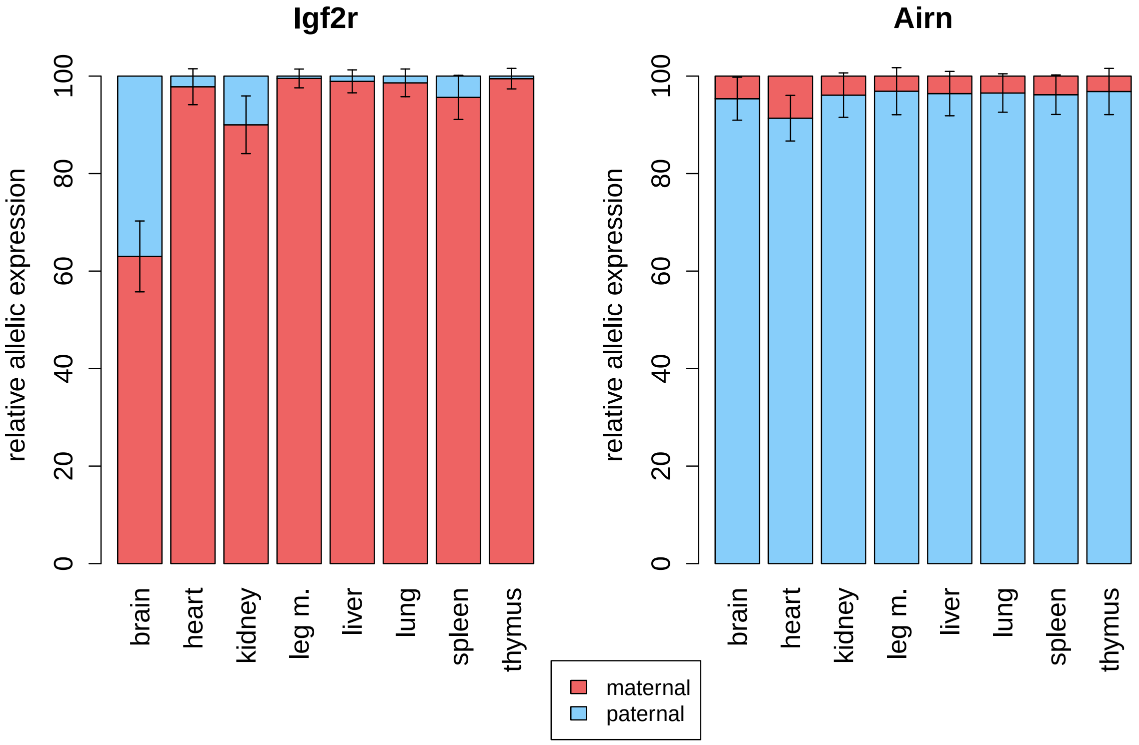


Figure 7: Quantified Sanger sequencing results for *Igf2r* and *Airn* in eight tissues of the adult mouse The graphs show the relative abundance of the transcript from the maternal (red) or paternal (blue) allele as observed in the Sanger sequencing data averaged over all four biological replicates (two for each cross). Error bars depict standard deviations across these four replicates. Overall the results are in agreement with previous data; leg m., leg muscle

After the initial results for the samples of CxF f2 and FxC f4 showed the experiment worked in all tissues the PCR for *Igf2r* and *Airn* was repeated in one biological replicate for each cross, i.e. CxF f3 and FxC f3. Sequencing the resulting products the data of all samples was then used to quantify the ratio between the two alleles. The results are shown in figure 7. The quantified data confirmed the initial results and showed between 91 and 96% paternal expression of *Airn* in all eight examined tissues, as well as 90 to 99% maternal expression of *Igf2r* in all tissues but brain, where the relaxation of imprinted expression resulted in a maternal/paternal ratio of 63/37%.

Overall these preliminary tests demonstrated that the primer design script yielded working PCR primers and evaluation of the SNPs contained in the PCR products by sequencing was shown to be a feasible method to detect maternal or paternal expression, as imprinted expression of all three tested known candidates, *Igf2r*, *Airn* and *Impact* could be validated.

2.2 Preparation of PCR samples for known imprinted protein coding genes

The first part of this project was the investigation of the imprinting status of protein coding genes which have been previously demonstrated to show imprinted expression. This was planned to be done in eight different tissues of the adult mouse for two crosses between genetically distinct mouse strains, which are reciprocal to each other and two biological replicates for each cross. The need for reciprocal crosses lies in the subsequent analysis of imprinted expression as explained at the beginning of the results section. Since this project uses a targeted approach instead of a genome-wide one the first step was to select the targets and design primers for the subsequent PCR based sequencing analysis.

2.2.1 Candidate selection and primer design

Starting point of this selection process was a list curated by Prof. Denise Barlow based on previously published data (see methods) yielding 167 candidates. The first step in the selection process depicted in figure 8 was to obtain a RefSeq annotation for these candidates, which worked for 99 candidates. Since some of these candidates had more than one RefSeq annotated isoform these 99 candidates corresponded to 165 isoforms which were selected for primer design. Isoforms were grouped together into one locus if they had the similar transcription start sites (see methods). This yielded 113 loci for the 99 RefSeq annotated candidates. In the initial run suitable primer pairs were designed for 141/165 transcripts corresponding to 95 loci and 83 candidates. Primers were designed so that the PCR product had a size of 100-500 bp and included at least two SNPs to allow a more reliable analysis regarding imprinted expression. 16 candidates had to be excluded at this stage because the SNPs located within the exons of the gene didn't allow the design of primers fitting these criteria. On the one hand there were candidates which only contained one SNP or no SNP at all, on the other hand some candidates had two or more SNPs but each of them was located so far away from the others that it was not possible to include them in the same PCR product due to the size limitation.

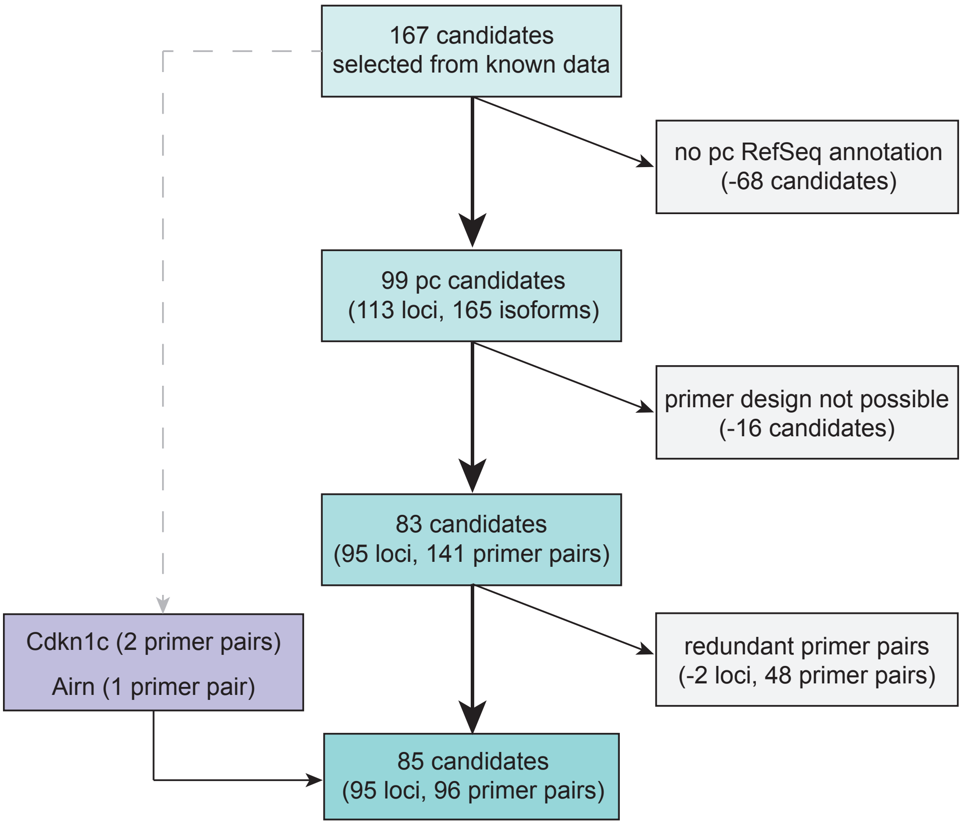


Figure 8: Selection of suitable known imprinted protein coding candidates. The chart gives an overview on the filtering process starting at the initial list and removing candidates with no RefSeq protein coding gene annotation (NM), candidates for which no suitable primers could be designed and lastly reducing the set to one primer pair per locus. A locus in this analysis is defined as one or more transcripts with the same transcriptional start site (see methods). *Cdkn1c* was added (although filtered out) to be included it in the analysis. Airn was chosen to complete the set of 96 primers. In total 96 primer pairs were used covering 95 loci and 85 candidates; pc, protein coding.

In the next step 48 primer pairs were excluded to obtain a 1:1 relationship between primer pairs and loci. Two loci also were removed in this step because the primer pairs designed were equal to the primer pairs of another locus and no primer pair for a product covering SNPs specific for isoforms of this locus could be designed. Overall this resulted in 93 primer pairs selected for 93 loci and 83 candidates. A special case was made for *Cdkn1c*, a candidate which was excluded at first because the SNPs were located too far apart from each other to be included in one PCR product. Therefore two primer pairs were designed covering one SNP each. Since this brought the total number of primer pairs to 95 *Airn* was included to complete the set of 96 primer pairs. This brought the total to 85 candidates and 95 loci. The list of selected candidates is given in table 2.

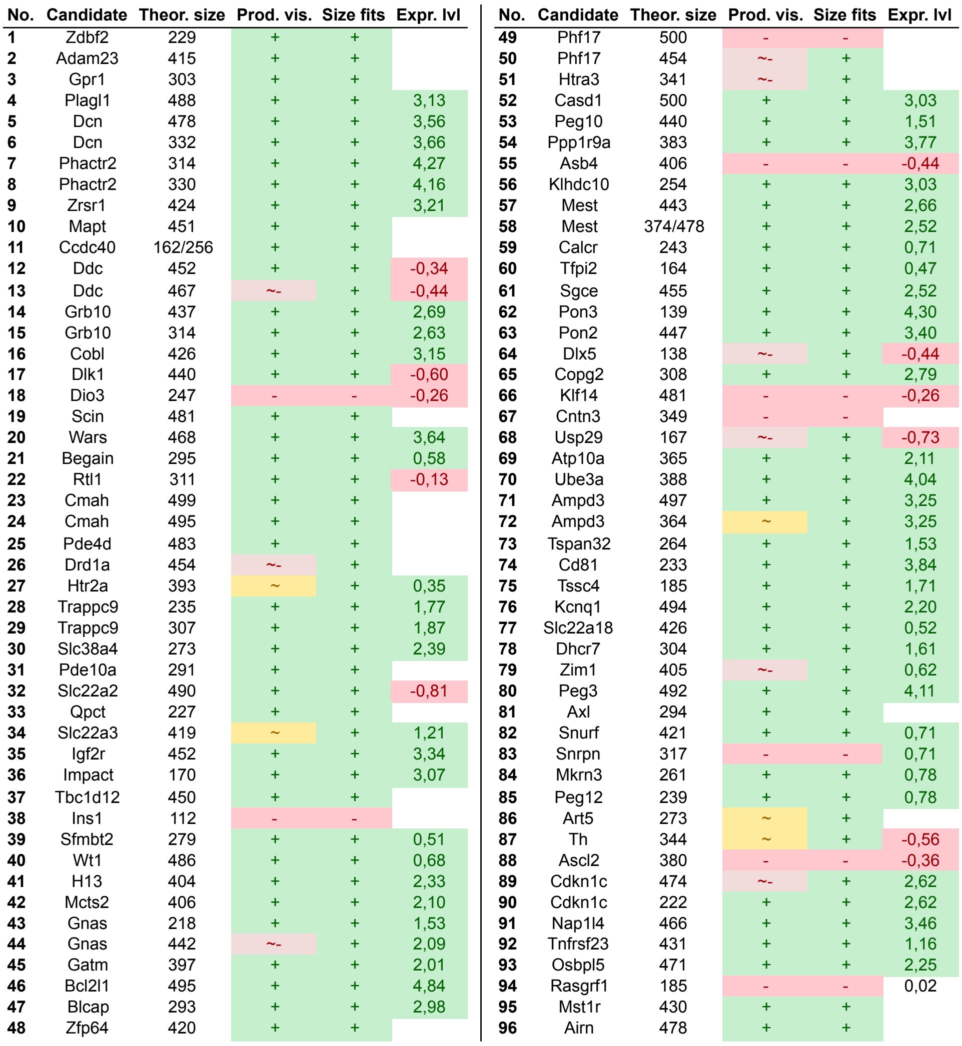
No.	Name	Chr.	No.	Name	Chr.	No.	Name	Chr.
1	Zdbf2	1	35	Igf2r	17	66	Klf14	6
2	Adam23	1	36	Impact	18	67	Cntn3	6
3	Gpr1	1	37	Tbc1d12	19	68	Usp29	7
4	Plagl1	10	38	Ins1	19	69	Atp10a	7
5+6	*Dcn*	*10*	39	Sfmbt2	2	70	Ube3a	7
7+8	*Phactr2*	*10*	40	Wt1	2	*71+72*	*Ampd3*	*7*
9	Zrsr1	11	41	H13	2	73	Tspan32	7
10	Mapt	11	42	Mcts2	2	74	Cd81	7
11	Ccdc40	11	*43+44*	*Gnas*	*2*	75	Tssc4	7
12+13	*Ddc*	*11*	45	Gatm	2	76	Kcnq1	7
14+15	*Grb10*	*11*	46	Bcl2l1	2	77	Slc22a18	7
16	Cobl	11	47	Blcap	2	78	Dhcr7	7
17	Dlk1	12	48	Zfp64	2	79	Zim1	7
18	Dio3	12	*49+50*	*Phf17*	*3*	80	Peg3	7
19	Scin	12	51	Htra3	5	81	Axl	7
20	Wars	12	52	Casd1	6	82	Snurf	7
21	Begain	12	53	Peg10	6	83	Snrpn	7
22	Rtl1	12	54	Ppp1r9a	6	84	Mkrn3	7
23+24	*Cmah*	*13*	55	Asb4	6	85	Peg12	7
25	Pde4d	13	56	Klhdc10	6	86	Art5	7
26	Drd1a	13	*57+58*	*Mest*	*6*	87	Th	7
27	Htr2a	14	59	Calcr	6	88	Ascl2	7
28+29	*Trappc9*	*15*	60	Tfpi2	6	*89+90*	*Cdkn1c*	*7*
30	Slc38a4	15	61	Sgce	6	91	Nap1l4	7
31	Pde10a	17	62	Pon3	6	92	Tnfrsf23	7
32	Slc22a2	17	63	Pon2	6	93	Osbpl5	7
33	Qpct	17	64	Dlx5	6	94	Rasgrf1	9
34	Slc22a3	17	65	Copg2	6	95	Mst1r	9
						96	Airn	17

Table 2: Known imprinted protein coding candidates selected for this study. Listed are the reaction numbers per candidate, the candidate name and the chromosome the candidate gene is located on. Candidates for which more than one primer pair was used, e.g. to cover isoforms of different loci are hightlighted in italics; Chr, chromosome; No., reaction number.

2.2.2 PCR for known imprinted targets

The PCR was done for all eight tissues in four replicates, two for each cross, resulting in a total of thirty-two 96-well plates. A list of the theoretical PCR product sizes compiled by the primer design script was used to validate the PCR reactions based on the location of the bands on the agarose gel. Figure 10 shows examples for such gel images, depicting the results for all four lung samples. The lanes were numbered according to the corresponding PCR reaction as listed in table 2. On first glance the images showed that a large proportion of the reactions worked and yielded a product which could be visualised using the gel. Furthermore it could be observed that some reactions produced more than one product. Since massive parallel sequencing will be used for the analysis of the PCR products the sequences for both the primary product and potential secondary products will be obtained. Therefore secondary products do not pose a problem for the downstream analysis as they do not interfere with the sequencing of the primary product. For further analysis visual inspection of the band intensity was used to grade the DNA yield of the PCR reaction using a subjective scale: -, no product visible; ~-, barely visible product; ~, product visible but weak intensity; ~+, clearly visible band but less intensity than +, which stands for a strong band. An example of this visual investigation for one of the gels for a lung sample (CxF f2 lung, figure 10 top left) is depicted in figure 9. The results of the inspection indicated that most reactions worked and showed at least a slightly visible product. All reactions yielding visible products showed products of the right size. In summary 9/96 reactions did not show a product ("-") and 73/96 showed a strong product ("+") in this sample. The remaining reactions were graded as having intermediate results.

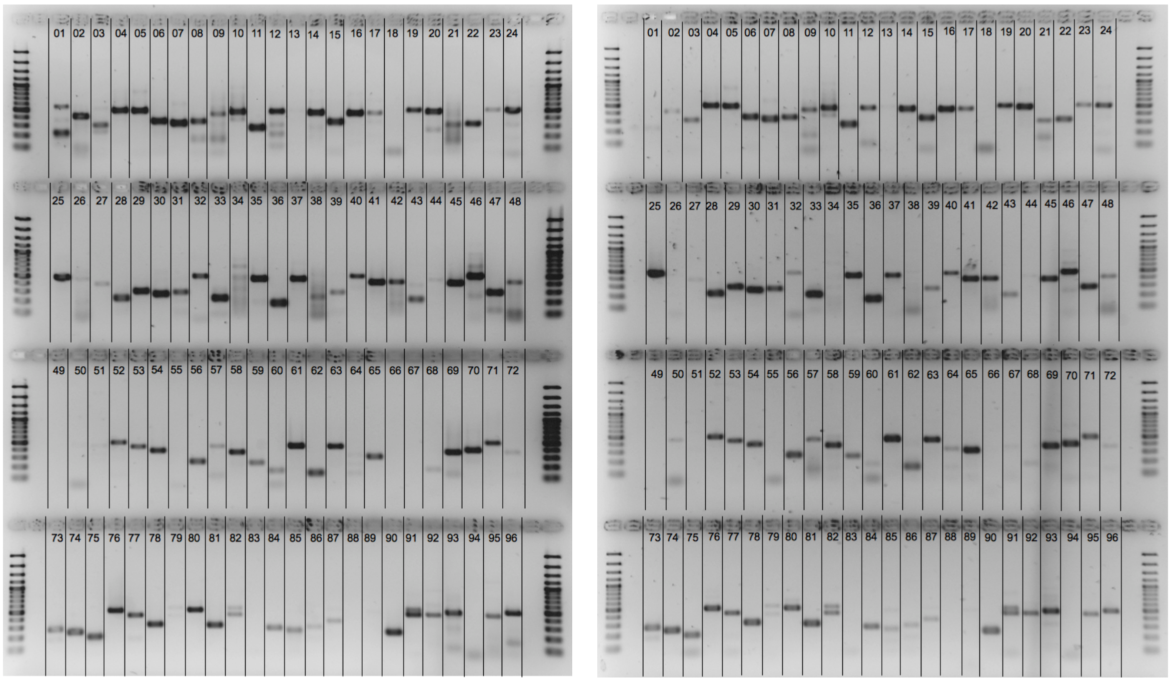
Since some of the candidate genes might not be expressed in the tissues investigated an additional validation step was introduced correlating expression levels with the PCR results. For this purpose data from the Mouse Imprinted Region Tiling Arrays (MIRTAs) was used. These arrays had been used to map total RNA to known imprinted regions of the mouse genome and can therefore be employed to estimate gene expression based on the levels of mRNA mapped [122]. The average hybridisation signal over the exons of the gene was used and values for one example are also included in figure 9. This method of estimating expression by the average hybridisation signal was also employed in a previous study [123]. Some candidates were located outside the regions covered by the MIRTAs, therefore the expression values could only be obtained for 73 reactions (72 loci, 64 candidates) using this data. The comparison with the expression levels showed that 59 of these 73 reactions showed a PCR product for a candidate with positive expression levels, 4 showed no product confirmed by negative expression levels, 8 showed a PCR product despite having negative expression levels and 2 showed no product although expression levels suggested that the reactions should have worked.



No.	Candidate	Theor. size	Prod. vis.	Size fits	Expr. lvl
1	Zdbf2	229	+	+	
2	Adam23	415	+	+	
3	Gpr1	303	+	+	
4	Plagl1	488	+	+	3,13
5	Dcn	478	+	+	3,56
6	Dcn	332	+	+	3,66
7	Phactr2	314	+	+	4,27
8	Phactr2	330	+	+	4,16
9	Zrsr1	424	+	+	3,21
10	Mapt	451	+	+	
11	Ccdc40	162/256	+	+	
12	Ddc	452	+	+	-0,34
13	Ddc	467	~-	+	-0,44
14	Grb10	437	+	+	2,69
15	Grb10	314	+	+	2,63
16	Cobl	426	+	+	3,15
17	Dlk1	440	+	+	-0,60
18	Dio3	247	-	-	-0,26
19	Scin	481	+	+	
20	Wars	468	+	+	3,64
21	Begain	295	+	+	0,58
22	Rtl1	311	+	+	-0,13
23	Cmah	499	+	+	
24	Cmah	495	+	+	
25	Pde4d	483	+	+	
26	Drd1a	454	~-	+	
27	Htr2a	393	~	+	0,35
28	Trappc9	235	+	+	1,77
29	Trappc9	307	+	+	1,87
30	Slc38a4	273	+	+	2,39
31	Pde10a	291	+	+	
32	Slc22a2	490	+	+	-0,81
33	Qpct	227	+	+	
34	Slc22a3	419	~	+	1,21
35	Igf2r	452	+	+	3,34
36	Impact	170	+	+	3,07
37	Tbc1d12	450	+	+	
38	Ins1	112	-	-	
39	Sfmbt2	279	+	+	0,51
40	Wt1	486	+	+	0,68
41	H13	404	+	+	2,33
42	Mcts2	406	+	+	2,10
43	Gnas	218	+	+	1,53
44	Gnas	442	~-	+	2,09
45	Gatm	397	+	+	2,01
46	Bcl2l1	495	+	+	4,84
47	Blcap	293	+	+	2,98
48	Zfp64	420	+	+	
49	Phf17	500	-	-	
50	Phf17	454	~-	+	
51	Htra3	341	~-	+	
52	Casd1	500	+	+	3,03
53	Peg10	440	+	+	1,51
54	Ppp1r9a	383	+	+	3,77
55	Asb4	406	-	-	-0,44
56	Klhdc10	254	+	+	3,03
57	Mest	443	+	+	2,66
58	Mest	374/478	+	+	2,52
59	Calcr	243	+	+	0,71
60	Tfpi2	164	+	+	0,47
61	Sgce	455	+	+	2,52
62	Pon3	139	+	+	4,30
63	Pon2	447	+	+	3,40
64	Dlx5	138	~-	+	-0,44
65	Copg2	308	+	+	2,79
66	Klf14	481	-	-	-0,26
67	Cntn3	349	-	-	
68	Usp29	167	~-	+	-0,73
69	Atp10a	365	+	+	2,11
70	Ube3a	388	+	+	4,04
71	Ampd3	497	+	+	3,25
72	Ampd3	364	~	+	3,25
73	Tspan32	264	+	+	1,53
74	Cd81	233	+	+	3,84
75	Tssc4	185	+	+	1,71
76	Kcnq1	494	+	+	2,20
77	Slc22a18	426	+	+	0,52
78	Dhcr7	304	+	+	1,61
79	Zim1	405	~-	+	0,62
80	Peg3	492	+	+	4,11
81	Axl	294	+	+	
82	Snurf	421	+	+	0,71
83	Snrpn	317	-	-	0,71
84	Mkrn3	261	+	+	0,78
85	Peg12	239	+	+	0,78
86	Art5	273	~	+	
87	Th	344	~	+	-0,56
88	Ascl2	380	-	-	-0,36
89	Cdkn1c	474	~-	+	2,62
90	Cdkn1c	222	+	+	2,62
91	Nap1l4	466	+	+	3,46
92	Tnfrsf23	431	+	+	1,16
93	Osbpl5	471	+	+	2,25
94	Rasgrf1	185	-	-	0,02
95	Mst1r	430	+	+	
96	Airn	478	+	+	

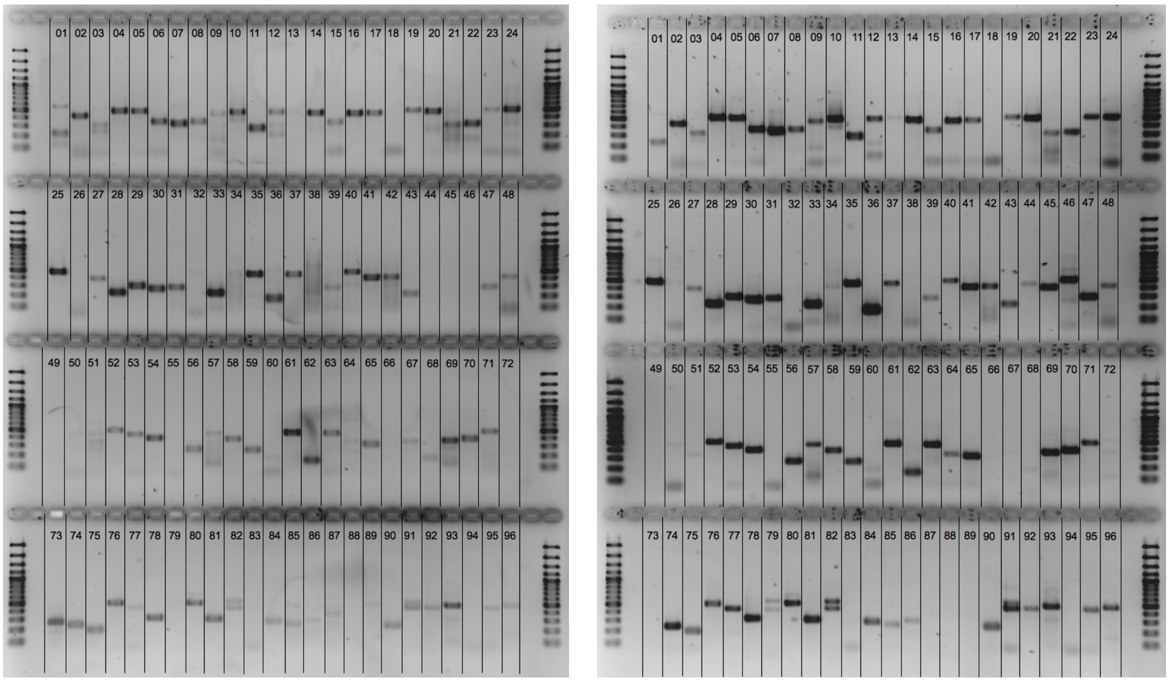
Figure 9: Visual inspection of PCR results in CxF f2 lung. The PCR results were evaluated using two criteria: (1) if there was a product visible at all (fourth column); (2) if the observed product is of the expected size as predicted by the primer design script (fifth column). A five level subjective scale was used ranging from - (no product, color code: red) to + (strong product, green) with intermediate steps: ~- (barely visible product, light red), ~(weak product, yellow) and ~+ (cleary visible but less than +, olive green, not used here) in between. Expression levels were estimated from the MIRTA data as average hybridisation signal over the exons of the gene and color coded as "not expressed" (red , average < 0), "weakly expressed" (white , average 0 - 0.3) and "expressed" (green, average > 0.3). Some candidates were not covered by the MIRTA data (empty cells).

The positive expression signal of one of the latter two candidates, *Snrpn*, can be explained by the shared exons with *Snurf* near the 3' end as well as an overlapping signal of a retroposed gene. This is depicted in figure 11. The green box indicates the positive MIRTA signal (bottom track) at the 3' end of *Snrpn* including the exons shared with *Snurf*. The positive signal of the retroposed gene (orange arrow) also overlaps an exon of Snrpn. In contrast to this the first exon of the long *Snrpn* isoform shows a negative MIRTA signal (blue arrow). Overall the overlap with positive signals most



(a) CxF f2

(b) CxF f3



(c) FxC f4

(d) FxC f3

Figure 10: PCR results for known imprinted protein coding candidates in lung samples. The agarose gels used to evaluate the success of the PCR reactions in all four biological replicates of adult lung are shown in figures (a) to (d), named according to the cross. Lane numbers indicate the corresponding PCR reaction as listed in table 2. Overall the pattern of working/not working reactions are largely reproducible in all four replicates. A detailed inspection of gel (a) is given in figure 9. Marker: GeneRuler 100 bp Plus DNA Ladder (Thermo Scientific).

likely not originating from the *Snrpn* transcript is sufficiently high to yield a positive average signal.

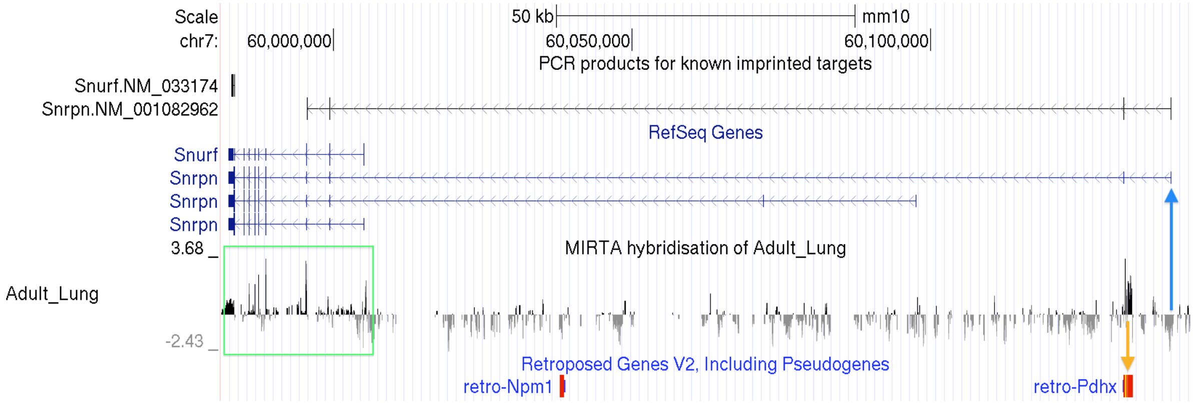
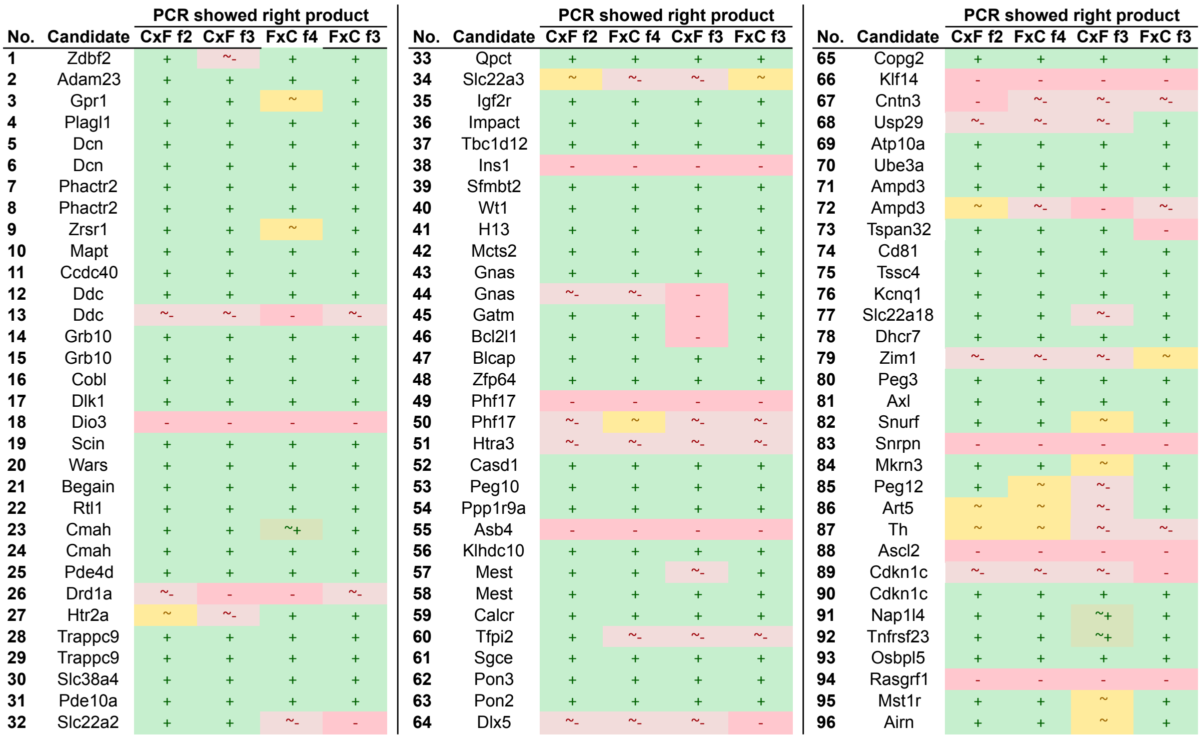


Figure 11: Overlapping transcripts are a caveat of using the average MIRTA expression value as an estimator of PCR results. This UCSC genome browser figure (described before) displays the PCR products designed for *Snrpn* and *Snurf* (top track), the RefSeq annotaton in this region and the MIRTA data for adult lung. The MIRTA data is displayed as the log2 of the cDNA/genomic DNA ratio. The bottom track shows retroposed genes. This example demonstrates that overlap with other transcripts which show a positive MIRTA signal (bottom track), in this case a retroposed gene (orange arrow) and *Snurf* (green box), might lead to an overall positive average even though the target transcript, here the long *Snrpn* isoform, might not be expressed at all, as indicated by the negative signal at the first exon (blue arrow). The top track shows the location of the two PCR products, with the product for *Snurf* located within the last exon and the product for *Snrpn* spanning multiple exons starting at the first exon of *Snrpn*.

The other discrepancy between PCR product and expression data of this kind was found for *Rasgrf1*, a candidate which only worked well in brain. The average expression signal in lung was about 7 fold reduced compared to the average signal in brain. This low positive expression signal might be the reason why the PCR did not work as there was not enough template. In conclusion the use of the average expression based on array data was a quick way to assess the validity of the PCR results with a high percentage of reactions showing the suggested results. The few examples where the PCR result did not match the expression data could be explained by overlapping transcripts and very low expression.

To assess the reproducibility of the PCR reactions the visual inspection results were compared across all four replicates, as visualised for lung samples (gel picture: figure 10) in figure 12. All reactions showed products of the correct size and the figure shows that most reactions worked in either all samples or in none. Table 3 summarises this finding, showing that in total 85/96 (89%) reactions showed consistent results across all four replicates. Of the remaining 11, 10 worked in all samples but one and it was further evaluated if the failed reactions should be repeated or not. The decision to repeat six of these reactions was mainly based on expression data suggesting that the reaction should theoretically work in lung. Five of the repeated reactions showed



No.	Candidate	PCR showed right product CxF f2	 CxF f3	 FxC f4	 FxC f3
1	Zdbf2	+	~-	+	+
2	Adam23	+	+	+	+
3	Gpr1	+	+	~	+
4	Plagl1	+	+	+	+
5	Dcn	+	+	+	+
6	Dcn	+	+	+	+
7	Phactr2	+	+	+	+
8	Phactr2	+	+	+	+
9	Zrsr1	+	+	~	+
10	Mapt	+	+	+	+
11	Ccdc40	+	+	+	+
12	Ddc	+	+	+	+
13	Ddc	~-	~-	-	~-
14	Grb10	+	+	+	+
15	Grb10	+	+	+	+
16	Cobl	+	+	+	+
17	Dlk1	+	+	+	+
18	Dio3	-	-	-	-
19	Scin	+	+	+	+
20	Wars	+	+	+	+
21	Begain	+	+	+	+
22	Rtl1	+	+	+	+
23	Cmah	+	+	~+	+
24	Cmah	+	+	+	+
25	Pde4d	+	+	+	+
26	Drd1a	~-	-	-	~-
27	Htr2a	~	~-	+	+
28	Trappc9	+	+	+	+
29	Trappc9	+	+	+	+
30	Slc38a4	+	+	+	+
31	Pde10a	+	+	+	+
32	Slc22a2	+	+	~-	-

No.	Candidate	PCR showed right product CxF f2	 FxC f4	 CxF f3	 FxC f3
33	Qpct	+	+	+	+
34	Slc22a3	~	~-	~-	~
35	Igf2r	+	+	+	+
36	Impact	+	+	+	+
37	Tbc1d12	+	+	+	+
38	Ins1	-	-	-	-
39	Sfmbt2	+	+	+	+
40	Wt1	+	+	+	+
41	H13	+	+	+	+
42	Mcts2	+	+	+	+
43	Gnas	+	+	+	+
44	Gnas	~-	~-	-	+
45	Gatm	+	+	-	+
46	Bcl2l1	+	+	-	+
47	Blcap	+	+	+	+
48	Zfp64	+	+	+	+
49	Phf17	-	-	-	-
50	Phf17	~-	~	~-	~-
51	Htra3	~-	~-	~-	~-
52	Casd1	+	+	+	+
53	Peg10	+	+	+	+
54	Ppp1r9a	+	+	+	+
55	Asb4	-	-	-	-
56	Klhdc10	+	+	+	+
57	Mest	+	+	~-	+
58	Mest	+	+	+	+
59	Calcr	+	+	+	+
60	Tfpi2	+	~-	~-	~-
61	Sgce	+	+	+	+
62	Pon3	+	+	+	+
63	Pon2	+	+	+	+
64	Dlx5	~-	~-	~-	-

No.	Candidate	PCR showed right product CxF f2	 FxC f4	 CxF f3	 FxC f3
65	Copg2	+	+	+	+
66	Klf14	-	-	-	-
67	Cntn3	-	~-	~-	~-
68	Usp29	~-	~-	~-	+
69	Atp10a	+	+	+	+
70	Ube3a	+	+	+	+
71	Ampd3	+	+	+	+
72	Ampd3	~	~-	-	~-
73	Tspan32	+	+	+	-
74	Cd81	+	+	+	+
75	Tssc4	+	+	+	+
76	Kcnq1	+	+	+	+
77	Slc22a18	+	+	~-	+
78	Dhcr7	+	+	+	+
79	Zim1	~-	~-	~-	~
80	Peg3	+	+	+	+
81	Axl	+	+	+	+
82	Snurf	+	+	~	+
83	Snrpn	-	-	-	-
84	Mkrn3	+	+	~	+
85	Peg12	+	~	~-	+
86	Art5	~	~	~-	+
87	Th	~	~	~-	~-
88	Ascl2	-	-	-	-
89	Cdkn1c	~-	~-	~-	-
90	Cdkn1c	+	+	+	+
91	Nap1l4	+	+	~+	+
92	Tnfrsf23	+	+	~+	+
93	Osbpl5	+	+	+	+
94	Rasgrf1	-	-	-	-
95	Mst1r	+	+	~	+
96	Airn	+	+	~	+

Figure 12: Visual inspection results for all four replicates of lung samples. The four columns show the results of the "Product visible" criterion (see figure 9) only if the product was of the correct size, otherwise "-" is displayed, regardless of intensity of the wrongly sized product. Most reactions worked for either all or none of the samples, suggesting high reproducibility. This is also summarised in table 3.

a product and were pooled with the PCR reactions of the first run. The corrected numbers are given in the last row of table 3, now showing that 90/96 (94%) reactions worked consistently across all replicates.

	worked:failed **4:0**	 **3:1**	 **2:2**	 **1:3**	 **0:4**
initial	77	10	1	0	8
after repetition	82	5	1	0	8

Table 3: Summary of visual inspection results for all four lung samples. Reactions are grouped according to the ratio between samples in which the reactions worked over samples in which they did not. The first row gives the results after the first PCRs. Six of the reactions with a 3:1 ratio were repeated for the respective samples where the reaction did not work as MIRTA data suggested expression. Five of these reactions yielded a product in the repetition. The last row shows the results after inclusion of the results from the repetitions.

The analysis described here for lung was then applied to all of the remaining seven tissues. The results of the visual inspection for all 32 samples are summarised in figure 13. These results all include repeated reactions. Reactions were grouped into four categories as indicated by the color code: Green highlights reactions that worked in all 32 samples while red shows reactions that worked in less than four samples.

Light green and yellow indicate reactions that worked in 24-31 and 4-23 samples, respectively. The category boundaries were derived from the fact that one tissue is represented by four samples so the red category contains reactions that did not even work for all replicates of one tissue.

No.	Candidate	worked in	No.	Candidate	worked in	No.	Candidate	worked in	No.	Candidate	worked in
1	Zdbf2	29	**25**	Pde4d	32	**49**	Phf17	0	**73**	Tspan32	31
2	Adam23	32	**26**	Drd1a	10	**50**	Phf17	26	**74**	Cd81	32
3	Gpr1	31	**27**	Htr2a	26	**51**	Htra3	20	**75**	Tssc4	31
4	Plagl1	32	**28**	Trappc9	32	**52**	Casd1	32	**76**	Kcnq1	32
5	Dcn	32	**29**	Trappc9	32	**53**	Peg10	32	**77**	Slc22a18	30
6	Dcn	31	**30**	Slc38a4	32	**54**	Ppp1r9a	32	**78**	Dhcr7	32
7	Phactr2	32	**31**	Pde10a	32	**55**	Asb4	10	**79**	Zim1	22
8	Phactr2	32	**32**	Slc22a2	17	**56**	Klhdc10	32	**80**	Peg3	32
9	Zrsr1	32	**33**	Qpct	32	**57**	Mest	32	**81**	Axl	32
10	Mapt	32	**34**	Slc22a3	31	**58**	Mest	32	**82**	Snurf	32
11	Ccdc40	25	**35**	Igf2r	32	**59**	Calcr	20	**83**	Snrpn	2
12	Ddc	32	**36**	Impact	32	**60**	Tfpi2	21	**84**	Mkrn3	32
13	Ddc	13	**37**	Tbc1d12	32	**61**	Sgce	32	**85**	Peg12	32
14	Grb10	32	**38**	Ins1	4	**62**	Pon3	32	**86**	Art5	29
15	Grb10	32	**39**	Sfmbt2	31	**63**	Pon2	32	**87**	Th	21
16	Cobl	32	**40**	Wt1	23	**64**	Dlx5	18	**88**	Ascl2	1
17	Dlk1	30	**41**	H13	32	**65**	Copg2	32	**89**	Cdkn1c	9
18	Dio3	0	**42**	Mcts2	32	**66**	Klf14	0	**90**	Cdkn1c	32
19	Scin	29	**43**	Gnas	32	**67**	Cntn3	23	**91**	Nap1l4	32
20	Wars	32	**44**	Gnas	29	**68**	Usp29	23	**92**	Tnfrsf23	32
21	Begain	30	**45**	Gatm	32	**69**	Atp10a	32	**93**	Osbpl5	32
22	Rtl1	32	**46**	Bcl2l1	32	**70**	Ube3a	32	**94**	Rasgrf1	6
23	Cmah	26	**47**	Blcap	32	**71**	Ampd3	32	**95**	Mst1r	32
24	Cmah	32	**48**	Zfp64	32	**72**	Ampd3	23	**96**	Airn	32

Figure 13: Summary of PCR results for known imprinted protein coding candidates. The figure lists for how many of the 32 samples the reaction yielded a product of the correct size. The color code is: green = worked in 32 samples, light green = worked in 24 - 31 samples, yellow = worked in 4 - 23 samples red = worked in less than 4 samples; Four reactions worked only in 0 - 1 sample, possible reasons are given in the text.

Grouped into these four categories 57/96 (59%) reactions worked in all 32 samples and 17/96 (18%) reactions worked in more than 24 samples but less than 32. 17/96 (18%) reactions showed a product of the correct size in at least four but less than 24 samples and 5/96 (5%) reactions worked in less than four reactions. Four of these five basically did not work in any sample. Reasons for this include expression exclusive to other developmental stages than adult for *Dio3* (reaction 18) [124] and *Klf14* (react. 66) [125] or tissue specific expression, either of just a specific isoform (*Phf17*, react. 49) or the candidate in general (*Ascl2*, react. 88) based on MIRTA data. The reaction for *Snrpn* only showed a product in brain, two times of the right size and two times of a different size. Reproducibility was assessed for all tissues as before (see table 3). The results are summarised in table 4 and showed that 82-93 reactions worked in either all or none of the samples of one tissue.

	brain	**heart**	**kidney**	**leg m.**	**liver**	**lung**	**spleen**	**thymus**
worked in all	79	79	82	74	69	82	74	72
worked in none	7	7	11	9	15	8	8	16
reproducible in all	86 (90%)	86 (90%)	93 (97%)	83 (86%)	84 (88%)	90 (94%)	82 (85%)	88 (92%)

Table 4: Reproducibility of PCR reactions for known imprinted protein coding candidates. The table summarizes the number of reactions that worked for either all (first row) or none (second row) of the four replicates per examined tissue. The third row gives a sum of the two to indicate how many reactions show reproducible results across all replicates of a tissue.

Finally two control 96-well PCRs using the whole primer set were performed using cDNA from inbred FVB/NJ and Cast/EiJ mouse strains, respectively. The aim of these controls was to show that the primers work equally on both alleles so that no bias from the PCR itself could influence the evaluation of imprinted expression in the downstream analysis. Another reason for this control was the ability to validate the SNPs located within the PCR products by showing that both variants can be found in the two strains. The template tissues selected for this test were based on the results from the previous PCRs choosing the respective tissue in which the reaction worked best before. The results show that 7/96 reactions did not work in both control reactions and 88 worked in both yielding a product of the correct size. *Snrpn* was an interesting case which only showed a product of the right size for CxF samples before, but not for the FxC samples. This could be reproduced in the FVB/Cast controls as only the FVB sample gave a product of the right size while the Cast sample gave a product that was around 100 bp larger than expected. This is in line with the paternal expression of *Snrpn*. Insertions/deletions between the two strains were investigated as a possible cause for this discrepancy but no insertion/deletion was found in the region of the PCR product. In the end it was decided to further investigate this once the sequencing of the control samples is finished. Overall the results indicate that most of the selected candidates were sufficiently well expressed in all examined tissues and the reproducibility of the PCR reactions for these candidates was high. With the possible exception of *Snrpn* there appeared to be no technical strain bias.

2.3 Preparation of PCR samples for non-coding candidates in imprinted regions.

The second part of the project encompasses the selection of non-coding candidates from a de novo assembly of transcripts and subsequent primer design and PCR for this selection. The focus here lies on ncRNAs in imprinted regions, as covered by the MIRTA data, as this part of the project is aimed at the validation of the imprinting status of known imprinted non-coding RNAs in these regions as well as the potential discovery novel imprinted non-coding RNAs. The target adult mouse tissues were the same as for the analysis of known imprinted protein coding genes. The selection of non-coding candidates required a method for separating protein coding and non-coding transcripts. A pipeline was implemented, validated and employed for this purpose.

2.3.1 Validation of the non-coding pipeline

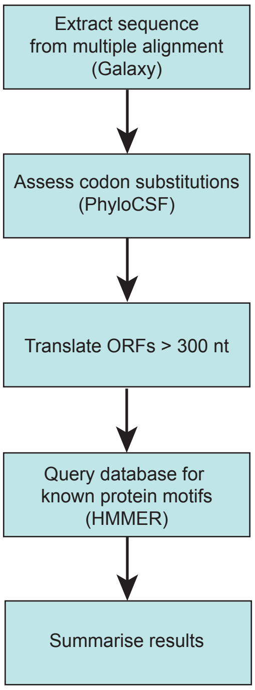

Extract sequence from multiple alignment (Galaxy)
Assess codon substitutions (PhyloCSF)
Translate ORFs > 300 nt
Query database for known protein motifs (HMMER)
Summarise results

Figure 14: The main steps of the non-coding pipeline employed to assess the coding potential of transcripts. After the extraction of sequence data from the multiple alignment files the coding potential was assessed by examining the codon substitutions that occurred between species using PhyloCSF for each exon separately. Afterwards the sequence was scanned for open reading frames of a minimum size (300 nucleotides for this analysis) and the translated sequences of these ORFs were queried against a database of known protein motifs using HMMER. The summarised results included the maximum PhyloCSF score as well as a boolean variable indicating if hmmer returned a significant hit or not.

The main steps of this pipeline are demonstrated in figure 14, a more in-depth description of the implementation is included in the methods section. The two criteria used to classify transcripts as coding were a PhyloCSF score of greater than 100 or if a significant hit in the protein motif database was found. The validity of using these two criteria was then assessed by testing the pipeline on RefSeq annotated transcripts. This was done for both RefSeq annotations having the NM RefSeq accession number prefix as a test set of protein coding genes and annotations with the NR RefSeq accession number prefix as a set of non-coding candidates. The overall results of these validation runs are shown in figure 15.

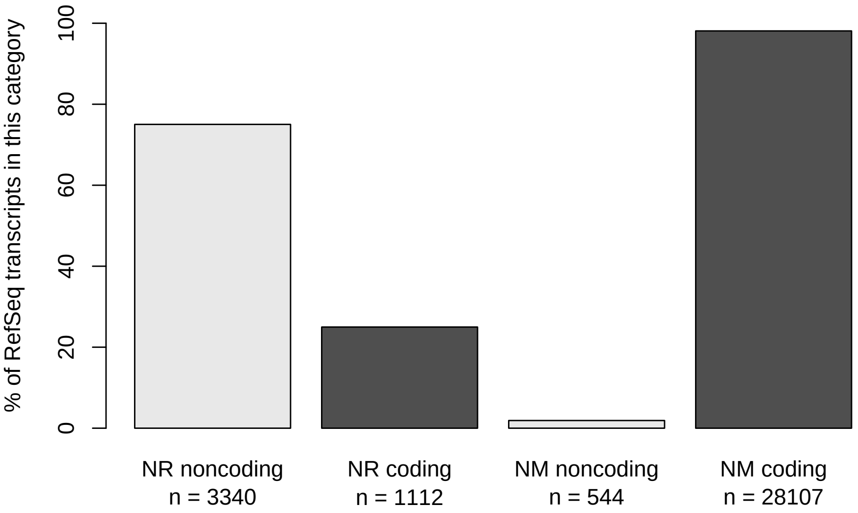
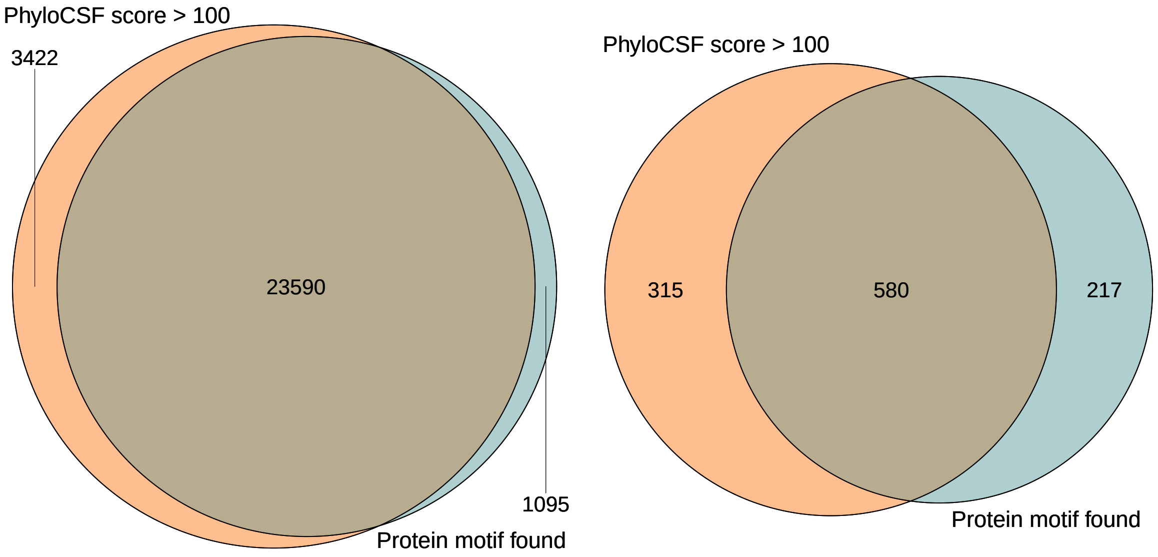


Figure 15: Results for the validation of the non-coding pipeline using RefSeq candidates. The graphs display the relative amount of non-coding (light grey) and protein coding transcripts (black) within each set. Total numbers of transcripts are given below each bar.

Concerning the set of protein coding candidates almost all, 98%, were classified correctly as protein coding suggesting a low rate of false-negatives using these criteria. Some of the false-negatives might encode for proteins smaller than 100 amino acids, which would lead to a missing hmmer result. One example for this is the gene *Vmac*, which encodes for a protein of 82 amino acids. In contrast to this, the results for the NR RefSeq set show that around one fourth of all NR transcripts were classified as protein coding. This suggests that by using these criteria the pipeline tends to overestimate the coding potential of transcripts and might classify truly non-coding transcripts as protein coding. However, these results also suggest that the criteria used allow a strict filtering so that candidates declared as non-coding by this pipeline comprise a reliable set of non-coding RNAs.

In addition to evaluating the overall classification a closer look was taken at the agreement between the two used criteria, as illustrated in figure 16.



(a) NM candidates **(b)** NR candidates

Figure 16: Comparison of the two criteria used to classify candidates as protein coding. The two diagrams show all candidates classified as protein coding for (a) NM RefSeq candidates and (b) NR RefSeq candidates. The areas indicate: orange - only supported by a PhyloCSF score > 100 but no significant hmmer result; blue - only supported by a significant hit in the protein motif database but the PhyloCSF score is lower than the threshold: brown - supported by both; The agreement between both classifiers is high for the NM candidates and about half of the NR candidates classified as protein coding are supported by both classifiers.

The results indicated a high agreement of the two classifiers regarding the classification of NM RefSeq candidates, with 98% of all classifications having both a PhyloCSF score greater than 100 and a protein motif could be found by hmmer. In contrast the classification of NR candidates showed a higher dissonance between these two classifiers as only around 50% of all classifications were supported by both of them. 28% showed a PhyloCSF score above the threshold but no hit could be found in the protein motif database. On the other hand 19% of all classifications had a PhyloCSF score lower than the threshold but the hmmer query yielded a significant hit in the database. These results suggested that most NM candidates could be confidently classified as protein coding, while the NR candidates classified as protein coding should be treated with more caution.

In summary both the results of the overall classification as well as the more detailed look into the agreement between the two classifiers suggested that candidates classified as protein coding by this pipeline should be treated with caution, as they might include a significant amount of false positives. On the other hand, and more relevant for this study, using the pipeline to filter for non-coding transcripts yielded a conservative but reliable set of non-coding RNAs.

2.3.2 Candidate selection and primer design

Candidates for this analysis were selected based on an annotation by Florian Pauler (see methods). The provided annotation consisted of 87 regions with varying numbers of exon models at these loci, ranging from zero for three regions with good MIRTA signal coverage but no exon model to >50 for the Meg3 locus. Additionally three known imprinted non-coding RNA candidates, *H19*, *Nctc1* and *Dio3os* were added manually since they were removed by prior filtering steps. This lead to a total of 90 regions at the start of the selection.

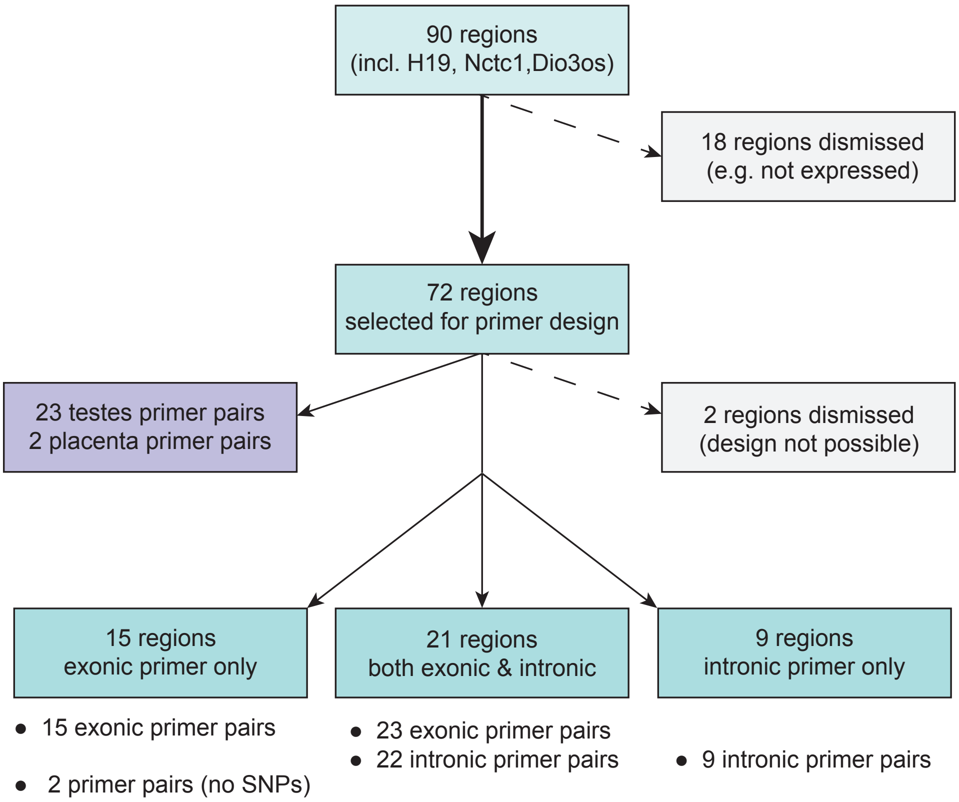


Figure 17: Selection of suitable non-coding candidates. Depicted is an overview of the steps during the selection of non-coding candidates starting at the provided annotation of 90 regions and first filtering for regions with transcripts expressed in the eight tissues of interest and deciding whether to also design intronic primers for the selected candidates (details see text). The number of selected primer pairs is given at the bottom of the figure. In addition to this, candidates with expression in testes or placenta were selected to validate the assembled splice variants (purple box).

An overview of the following selection process is given in figure 17. Candidates were selected based on expression data from the MIRTAs and RNA sequencing forming three sets of candidates for analysis in a) the eight adult tissues of this study, b) testes or c) placenta. This resulted in the dismissal of 16/90 regions. Two additional candidates, *Airn* and *AK041647*, were excluded because the former was already included in the first part of the study and the latter might encode for a protein of 1382 amino acids according to the UCSC database. In total 18 regions were dismissed during the first step leaving 72 regions comprised of 47 regions for candidates in the eight adult tissues, 23 regions for candidates in testes and two regions for placenta candidates (figure 17 first step).

The following step was the design of primers for products covering at least 2 SNPs and if possible spanning at least one splice junction. Primer design worked for all regions except six. Furthermore two regions were included just to validate the existence of the isoforms so the products did not have to fulfil the criterion of covering two SNPs. Special focus was put on *Kcnq1ot1* as its transcript was previously found to be unspliced [126] but a spliced isoform was found in the *de novo* assembly.

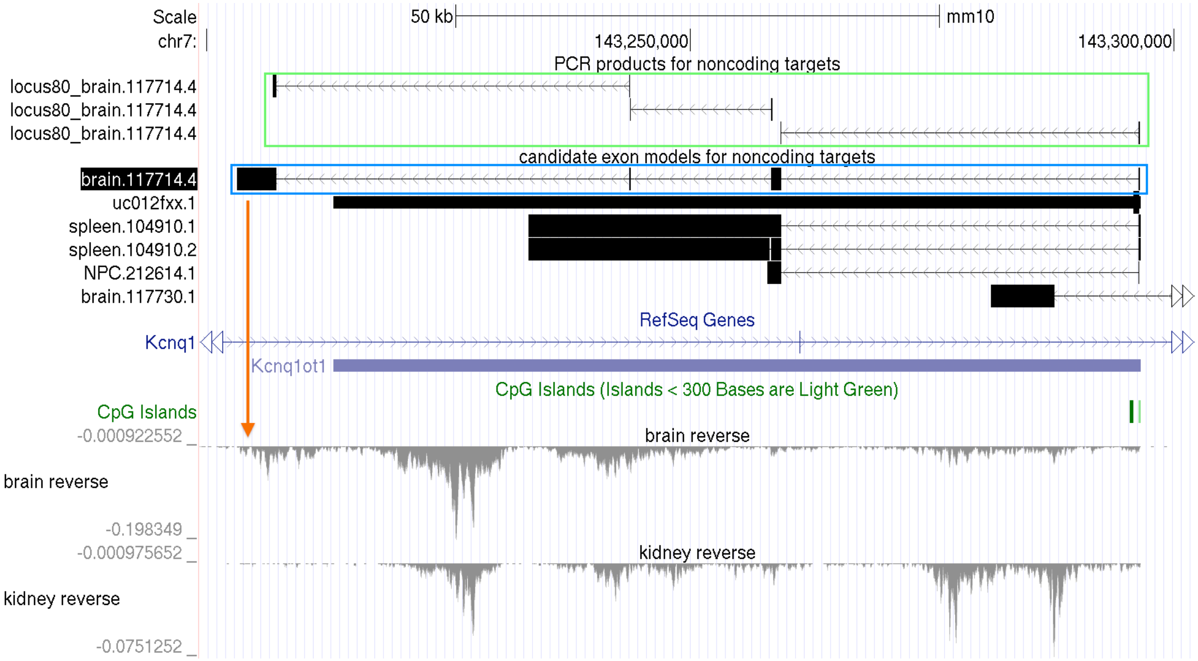


Figure 18: Junction spanning primer design for *Kcnq1ot1*. This UCSC genome browser screenshot (described before) shows the three primer pairs (green) designed for each junction of the long spliced isoform brain.117714.4 (blue box). Other isoforms assembled in this region are also shown in this track. The RefSeq annotation of *Kcnq1* and *Kcnq1ot1* as well as a track showing CpG islands are given below. The isoform appeared to be supported by RNA sequencing data (bottom two tracks). These tracks depict the amount of sequences (reads) gained by RNA sequencing that aligned to these regions. The last exon appeared to be supported only in brain (orange arrow) as the other examined tissues showed low amounts of sequencing reads in this region (example: kidney).

As the existence of this isoform would disagree with published data it was investigated in more detail. Three primer pairs were designed to cover all junctions of a long spliced isoform. This is demonstrated in figure 18. The isoform appeared to be supported by RNA sequencing data (bottom tracks). The last exon appeared to be used only in brain, as compared to other tissues, in this example kidney.

In addition to junction spanning primers "intronic" primers (see methods) were designed for some candidates selected for analysis in the eight adult tissues mentioned before. The criterion for this was that they did not overlap any other genes. Furthermore primers were designed for three regions without an assembled exon-model. In summary two regions were dismissed during the primer design step since neither a suitable exonic nor intronic primer pair could be designed.

Region	Exonic primer pairs	Junctions covered	Intronic primer pairs
locus3	1	1	1
locus5	1	0	1
locus6	1	4	1
locus7	1	1	1
locus8	-	-	1
locus9	1	-	1
locus12	1	4	-
Meg3	-	-	1
locus17	1	2	1
locus18	1	3	1
locus19	1	2	1
locus23	-	-	1
locus32	1	2	1
locus38	1	1	-
locus40	1	1	1
Nespas	2	1+1	1
locus45	1	1	-
locus46	1	1	-
locus48	1	3	1
locus51	1	3	-
locus53	1	1	-
locus56	1	2	1
locus58	1	2	-
Ipw	1	2	1
locus64	1	5	1
locus65	-	-	1
locus66	2	4+2	1
locus67	1	4	1
locus68	1	6	2
locus74	1	2	1
locus75	-	-	1
locus78	1	3	-
locus79	1	4	1
Kcnq1ot1	3	1+1+1	-
locus81	1	3	-
locus83	1	2	-
locus84	1	1	-
locus85	-	-	1
locus86	-	-	1
locus87	-	-	1
Nctc1	1	1	1
H19	-	-	1
Dio3os	1	1	-
locus36	1	1	-
locus61	1	2	-

Table 5: Summary of non-coding candidates selected for this analysis. This table lists the 45 regions for which primers were designed. The data provided includes the number of exonic primer pairs designed for the region together with how many junctions were covered by the product (if any). In addition the number of intronic primers designed is given. The last two regions listed were the ones selected for junction validation only.

Table 5 In total the selection yielded 45 regions to investigate in all eight tissues, 43 of which could be used for an analysis of imprinted expression due to inclusion of SNPs in the PCR primer design. 38 junction spanning primer pairs as well as 31 intronic

primer pairs were designed for these 43 regions. Combining this set with the aforementioned two primer pairs for products without SNPs gives a total of 71 primer pairs to be examined in all eight tissues. In addition 23 and two junction spanning primer pairs were designed for candidates expressed in testes and placenta, respectively, completing the set of 96 primer pairs for the second part of this study.

2.3.3 PCR for non-coding targets in eight tissues

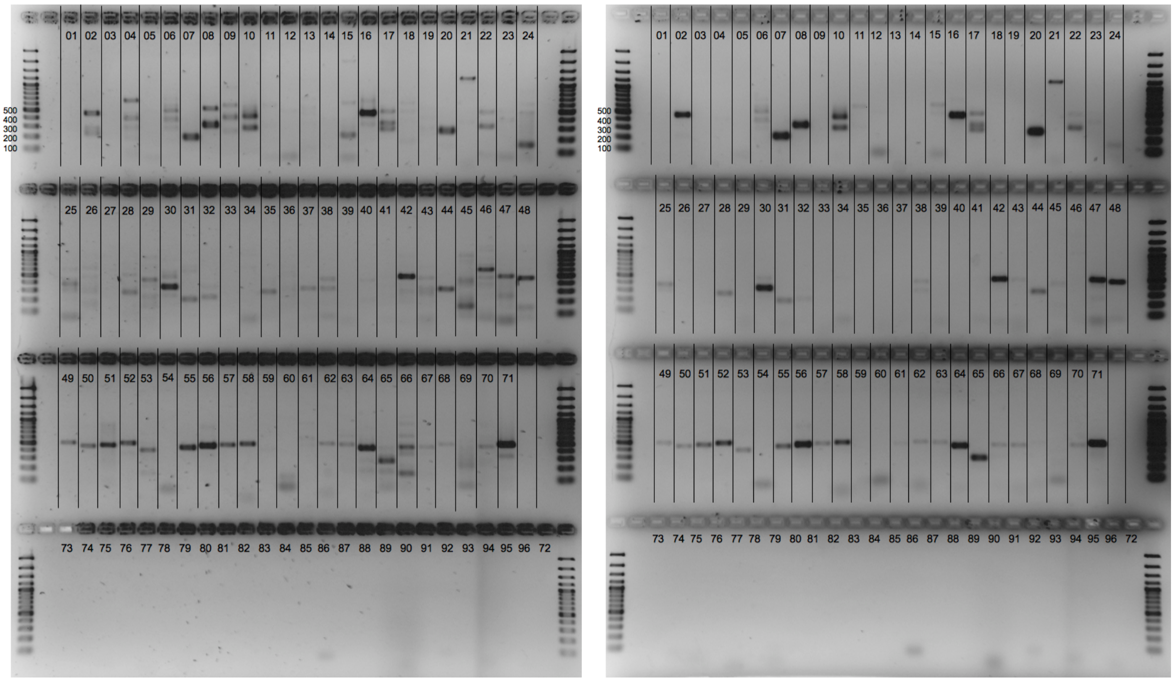
PCR was performed for all eight tissues and all four replicates per tissue as in section 2.2.2. At the start a pilot experiment was done using the FxC f4 brain sample to check if the primers were working. Visual inspection of the results was done as described on page 24 and correlation with expression was analysed using both the average MIRTA signal over the whole transcript and a visual evaluation of the MIRTA signal at the specific regions of the PCR product. The results are listed in table 6.

	pos. match	**neg. match**	**no match** (negative signal)	**no match** (positive signal)
average signal	37	7	24	3
visual evaluation	46	5	15	5

Table 6: Evaluation of the correlation between MIRTA signal and PCR results Results are given for both the average MIRTA signal over the whole transcript and the MIRTA signal in the region of the PCR product based on a visual inspection. Categories: pos. match, positive MIRTA signal and PCR product of the right size; negative match, negative MIRTA signal and no PCR product; no match (negative signal), PCR showed a product but MIRTA signal was negative; no match (positive signal), PCR did not work although MIRTA signal was positive.

Overall the agreement between MIRTA expression data and PCR results was better when using the visual evaluation of the specific region and not the average signal over the whole gene, giving 51 and 44 matches, respectively. Compared with the correlation results of the known imprinted protein coding candidates these results showed that the MIRTA expression data was a less reliable tool for predicting the success of PCR reactions for de novo assembled transcripts.

Five primers were redesigned after this pilot run because they appeared not to work properly or they overlapped an insertion or deletion in one of the two crosses (see methods), a fact that was overlooked in the initial design process. The redesigned primers were tested successfully and replaced the initial primers for the PCRs in this analysis. PCR and subsequent agarose gel analysis was performed for all eight tissues in four replicates. An example for gel pictures of such PCR reaction results is given in figure 19 showing all four analysed heart samples.



(a) CxF f2 **(b)** CxF f3

(c) FxC f4 **(d)** FxC f3

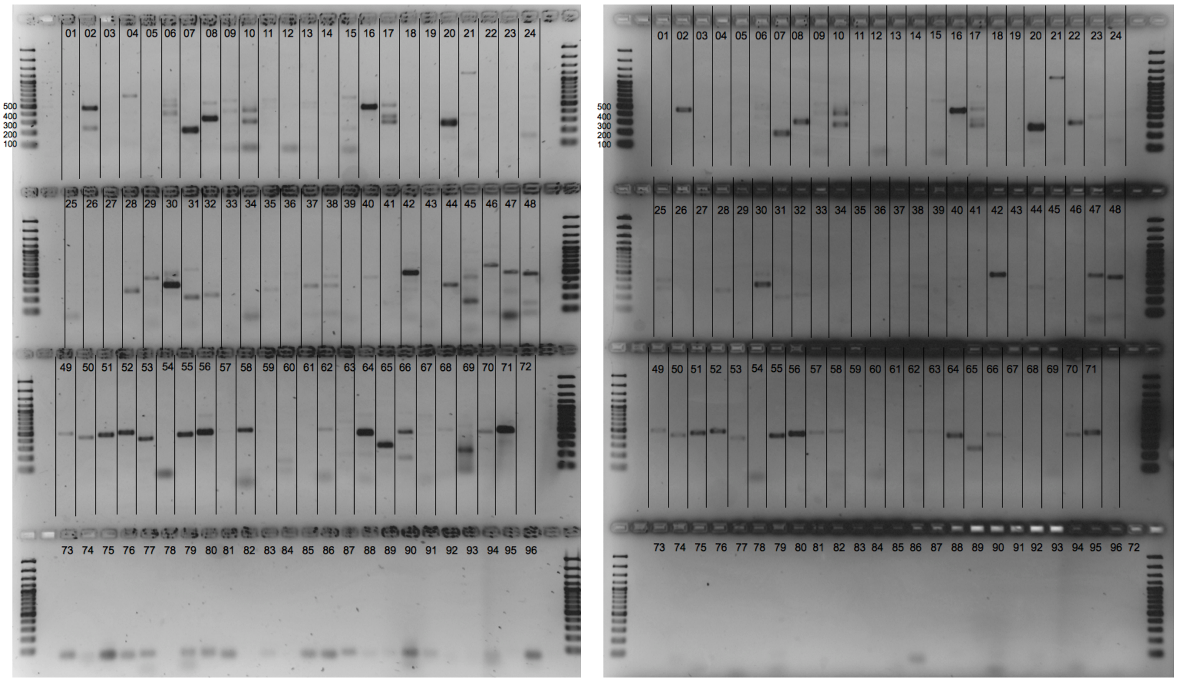
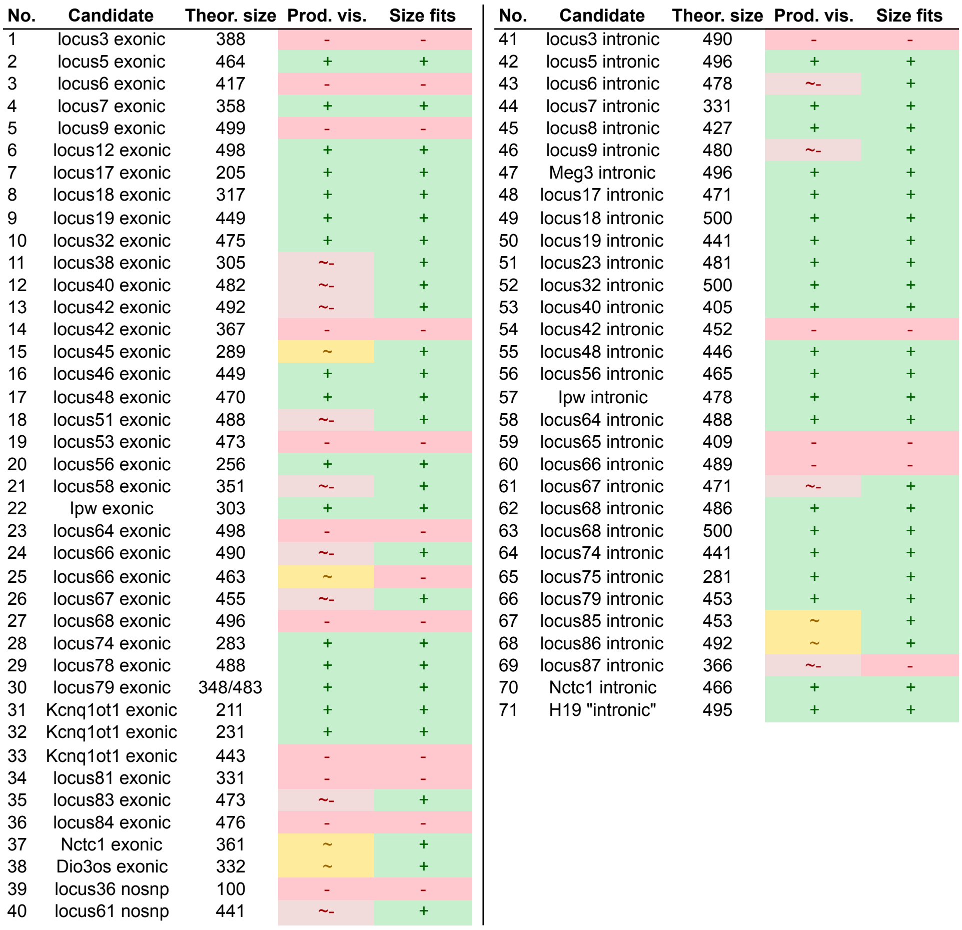


Figure 19: PCR results for non-coding candidates in heart samples. The figures show the agarose gels visualizing the results of the PCR reactions in all four biological replicates of adult heart. The four replicates are shown in (a) to (d). The lane numbers indicate the corresponding PCR reaction. Reactions 73-96 and 72 were negative controls with -RT samples and H_2O as template, respectively. The reactions 1-38 were for junction-spanning PCR products covering SNPs, 39-40 for junction-spanning products without SNPs and 41-71 for intronic PCR products. A detailed inspection of gel (a) is given in figure 20. Marker: GeneRuler 100 bp Plus DNA Ladder (Thermo Scientific).

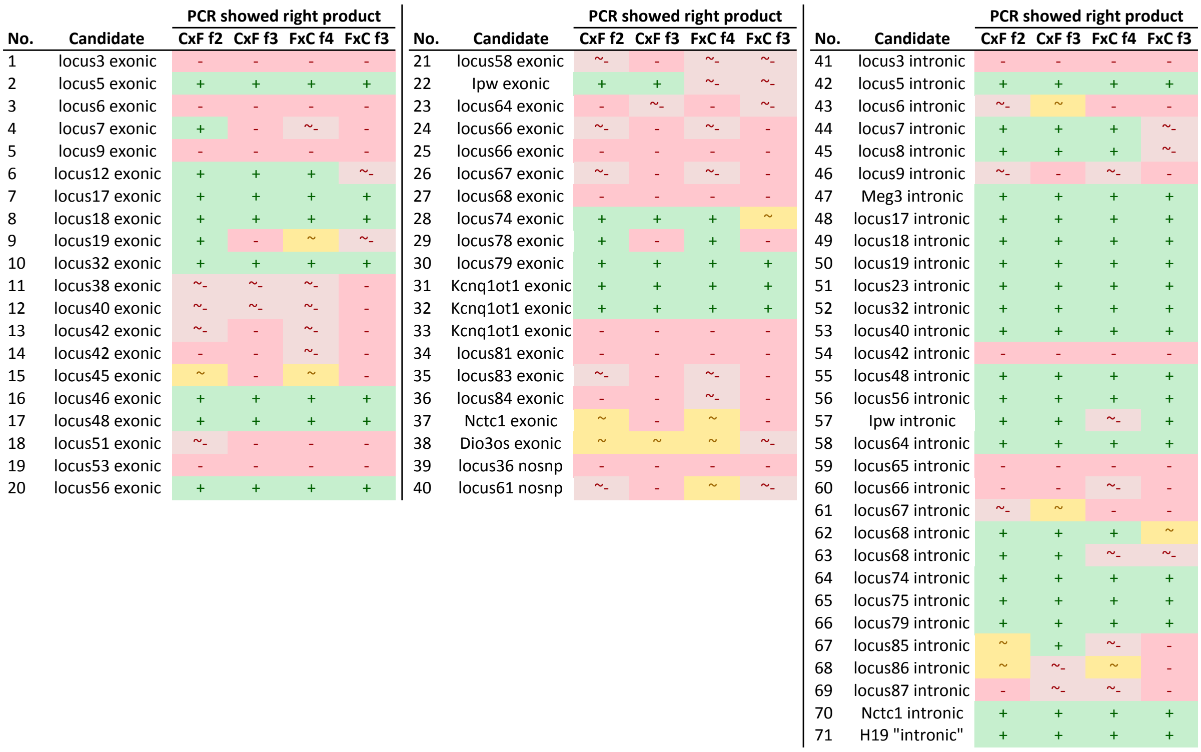
Since only 71 reactions were needed this time the remaining 25 were used for negative controls using -RT samples and H_2O as templates, numbered 73-96 and 72, respectively. Intronic primers 41-64 were used for the -RT controls and intronic primer 41 for the water control. In general only negative control samples were loaded on the last row of each gel and it could be observed that no product was visible for any of them. The numbers 1-71 correspond to the reaction number, reactions 1-38 being junction spanning PCR products covering SNPs, 39-40 being junction-spanning products without SNPs and 41-71 being intronic PCR products. In summary the gel images were similar across the four replicates. Overall less reactions appeared to have worked for this set of primers and more secondary products were observed than for the known imprinted candidates (figure 10).



No.	Candidate	Theor. size	Prod. vis.	Size fits
1	locus3 exonic	388	-	-
2	locus5 exonic	464	+	+
3	locus6 exonic	417	-	-
4	locus7 exonic	358	+	+
5	locus9 exonic	499	-	-
6	locus12 exonic	498	+	+
7	locus17 exonic	205	+	+
8	locus18 exonic	317	+	+
9	locus19 exonic	449	+	+
10	locus32 exonic	475	+	+
11	locus38 exonic	305	~-	+
12	locus40 exonic	482	~-	+
13	locus42 exonic	492	~-	+
14	locus42 exonic	367	-	-
15	locus45 exonic	289	~	+
16	locus46 exonic	449	+	+
17	locus48 exonic	470	+	+
18	locus51 exonic	488	~-	+
19	locus53 exonic	473	-	-
20	locus56 exonic	256	+	+
21	locus58 exonic	351	~-	+
22	Ipw exonic	303	+	+
23	locus64 exonic	498	-	-
24	locus66 exonic	490	~-	+
25	locus66 exonic	463	~	-
26	locus67 exonic	455	~-	+
27	locus68 exonic	496	-	-
28	locus74 exonic	283	+	+
29	locus78 exonic	488	+	+
30	locus79 exonic	348/483	+	+
31	Kcnq1ot1 exonic	211	+	+
32	Kcnq1ot1 exonic	231	+	+
33	Kcnq1ot1 exonic	443	-	-
34	locus81 exonic	331	-	-
35	locus83 exonic	473	~-	+
36	locus84 exonic	476	-	-
37	Nctc1 exonic	361	~	+
38	Dio3os exonic	332	~	+
39	locus36 nosnp	100	-	-
40	locus61 nosnp	441	~-	+
41	locus3 intronic	490	-	-
42	locus5 intronic	496	+	+
43	locus6 intronic	478	~-	+
44	locus7 intronic	331	+	+
45	locus8 intronic	427	+	+
46	locus9 intronic	480	~-	+
47	Meg3 intronic	496	+	+
48	locus17 intronic	471	+	+
49	locus18 intronic	500	+	+
50	locus19 intronic	441	+	+
51	locus23 intronic	481	+	+
52	locus32 intronic	500	+	+
53	locus40 intronic	405	+	+
54	locus42 intronic	452	-	-
55	locus48 intronic	446	+	+
56	locus56 intronic	465	+	+
57	Ipw intronic	478	+	+
58	locus64 intronic	488	+	+
59	locus65 intronic	409	-	-
60	locus66 intronic	489	-	-
61	locus67 intronic	471	~-	+
62	locus68 intronic	486	+	+
63	locus68 intronic	500	+	+
64	locus74 intronic	441	+	+
65	locus75 intronic	281	+	+
66	locus79 intronic	453	+	+
67	locus85 intronic	453	~	+
68	locus86 intronic	492	~	+
69	locus87 intronic	366	~-	-
70	Nctc1 intronic	466	+	+
71	H19 "intronic"	495	+	+

Figure 20: Visual inspection of PCR results in CxF f2 heart. Junction spanning primers are listed on the left side, while intronic primers are listed on the right. Two characteristics of the results were assessed. First, if there was a product visible (fourth column) and second if the observed product is of the expected size as predicted by the primer design script (fifth column). The same scale as in figure 9 was used ranging from - (no product, color code: red) to + (strong product, green) with intermediate steps ~- (barely visible product, light red), ~(weak product, yellow) and ~+ (cleary visible but less than +, olive green, not used here) in between.

A more detailed visual inspection of the gel for CxF f2 is summarised in figure 20. The same scoring system as before was used to assess both the amount and the size of the PCR products. 15/71 (21%) reactions showed no product at all ("-") while 37/71 (52%) showed a strong band on the gel ("+"). The remaining 19 reactions showed intermediate results. These results indicate less ubiquitous expression of the non-coding candidates compared to the protein coding targets. Two reactions showed a PCR product but not of the right size. This could also be observed before for different reactions in the pilot PCR in brain and was probably an indication that the target was not expressed in this tissue, as subsequent tests after the pilot run showed that the reactions that did not give a product of the right size in brain did so when using a different tissue as template in which the target should be expressed according to the MIRTA signal.



No.	Candidate	PCR showed right product			
		CxF f2	CxF f3	FxC f4	FxC f3
1	locus3 exonic	-	-	-	-
2	locus5 exonic	+	+	+	+
3	locus6 exonic	-	-	-	-
4	locus7 exonic	+	-	~-	-
5	locus9 exonic	-	-	-	-
6	locus12 exonic	+	+	+	~-
7	locus17 exonic	+	+	+	+
8	locus18 exonic	+	+	+	+
9	locus19 exonic	+	-	~	~-
10	locus32 exonic	+	+	+	+
11	locus38 exonic	~-	~-	~-	-
12	locus40 exonic	~-	~-	~-	-
13	locus42 exonic	~-	-	~-	-
14	locus42 exonic	-	-	~-	-
15	locus45 exonic	~	-	~	-
16	locus46 exonic	+	+	+	+
17	locus48 exonic	+	+	+	+
18	locus51 exonic	~-	-	-	-
19	locus53 exonic	-	-	-	-
20	locus56 exonic	+	+	+	+
21	locus58 exonic	~-	-	~-	~-
22	Ipw exonic	+	+	~-	~-
23	locus64 exonic	-	~-	-	~-
24	locus66 exonic	~-	-	~-	-
25	locus66 exonic	-	-	-	-
26	locus67 exonic	~-	-	~-	-
27	locus68 exonic	-	-	-	-
28	locus74 exonic	+	+	+	~
29	locus78 exonic	+	-	+	-
30	locus79 exonic	+	+	+	+
31	Kcnq1ot1 exonic	+	+	+	+
32	Kcnq1ot1 exonic	+	+	+	+
33	Kcnq1ot1 exonic	-	-	-	-
34	locus81 exonic	-	-	-	-
35	locus83 exonic	~-	-	~-	-
36	locus84 exonic	-	-	~-	-
37	Nctc1 exonic	~	-	~	-
38	Dio3os exonic	~	~	~	~-
39	locus36 nosnp	-	-	-	-
40	locus61 nosnp	~-	-	~	~-
41	locus3 intronic	-	-	-	-
42	locus5 intronic	+	+	+	+
43	locus6 intronic	~-	~	-	-
44	locus7 intronic	+	+	+	~-
45	locus8 intronic	+	+	+	~-
46	locus9 intronic	~-	-	~-	-
47	Meg3 intronic	+	+	+	+
48	locus17 intronic	+	+	+	+
49	locus18 intronic	+	+	+	+
50	locus19 intronic	+	+	+	+
51	locus23 intronic	+	+	+	+
52	locus32 intronic	+	+	+	+
53	locus40 intronic	+	+	+	+
54	locus42 intronic	-	-	-	-
55	locus48 intronic	+	+	+	+
56	locus56 intronic	+	+	+	+
57	Ipw intronic	+	+	~-	+
58	locus64 intronic	+	+	+	+
59	locus65 intronic	-	-	-	-
60	locus66 intronic	-	-	~-	-
61	locus67 intronic	~-	~	-	-
62	locus68 intronic	+	+	+	~
63	locus68 intronic	+	+	~-	~-
64	locus74 intronic	+	+	+	+
65	locus75 intronic	+	+	+	+
66	locus79 intronic	+	+	+	+
67	locus85 intronic	~	+	~-	-
68	locus86 intronic	~	~-	~	-
69	locus87 intronic	-	~-	~-	-
70	Nctc1 intronic	+	+	+	+
71	H19 "intronic"	+	+	+	+

Figure 21: Visual inspection results for all four replicates of heart samples. The four columns give the results for the criteria used before, showing the results of the "PCR worked" criterion only if the product had the right size. The first tables show the results for the junction spanning primers, while the thirdshows the results for the intronic primers. A summary of these results is given in table 7.

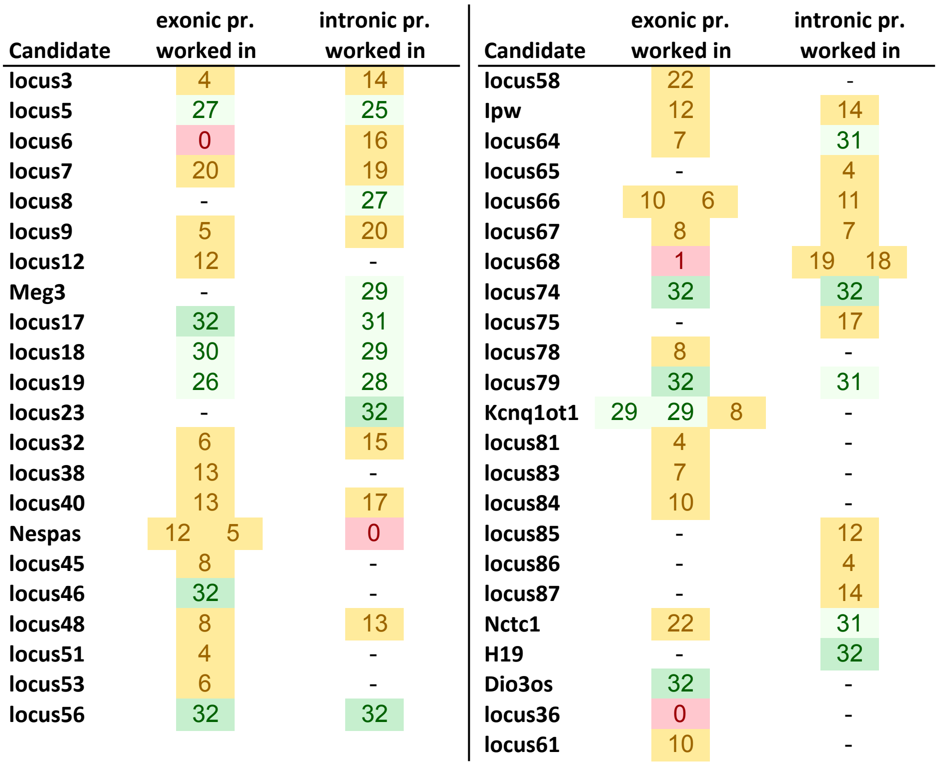
The reproducibility of the PCR results was assessed by comparing the results of the visual inspection between all four replicates of one tissue. This comparison is shown for heart in figure 21 and summarised in table 7. Overall 47/71 (66%) reactions worked for either none or all of the four replicates, indicating lower reproducibility of the PCR results than for the protein coding candidates. On separation of junction spanning from intronic primers it became apparent that the intronic primers worked more often than the junction spanning primers indicating that some of the assembled exon models used for primer design might not be valid or indicate tissue specific splice variants. The reproducibility of the PCR results was also different between the two sets with

23/40 (58%) reactions with junction spanning primers and 24/31 (77%) reactions with intronic primers working for either all or none of the four replicates.

	worked:failed				
	4:0	**3:1**	**2:2**	**1:3**	**0:4**
exonic	14	5	9	3	9
intronic	21	2	4	1	3
total	35	7	13	4	12

Table 7: Summary of visual inspection results for all four heart samples. Reactions are grouped according to the ratio between samples in which the reactions worked over samples in which they did not. Results are given for intronic and junction spanning primer pairs separately and for both combined.

The results for all eight tissues are given in figure 22. The numbers indicate how many of the 32 reactions per primer pair in all samples worked. The colour code is the same as before, green indicating reactions that worked in all 32 samples, light green in 24-31 samples, yellow in 4-23 samples and red in less than 4 samples.



Candidate	exonic pr. worked in	intronic pr. worked in	Candidate	exonic pr. worked in	intronic pr. worked in
locus3	4	14	locus58	22	-
locus5	27	25	Ipw	12	14
locus6	0	16	locus64	7	31
locus7	20	19	locus65	-	4
locus8	-	27	locus66	10 6	11
locus9	5	20	locus67	8	7
locus12	12	-	locus68	1	19 18
Meg3	-	29	locus74	32	32
locus17	32	31	locus75	-	17
locus18	30	29	locus78	8	-
locus19	26	28	locus79	32	31
locus23	-	32	Kcnq1ot1	29 29 8	-
locus32	6	15	locus81	4	-
locus38	13	-	locus83	7	-
locus40	13	17	locus84	10	-
Nespas	12 5	0	locus85	-	12
locus45	8	-	locus86	-	4
locus46	32	-	locus87	-	14
locus48	8	13	Nctc1	22	31
locus51	4	-	H19	-	32
locus53	6	-	Dio3os	32	-
locus56	32	32	locus36	0	-
			locus61	10	-

Figure 22: Summary of PCR results for non-coding candidates. The numbers shown indicate how many of the 32 reactions in all samples showed a product of the right size. Both reactions for junction spanning (exonic) primer pairs and intronic primer pairs are shown per candidate. In some cases more than one exonic or intronic primer pair was designed per candidate with a maximum of three primer pairs for *Kcnq1ot1*. The colour code is the same as before. green: 32; light green: 24-31; yellow: 4-23; red: <4;

Overall 10/71 (14%) reactions worked in all 32 samples, 14/71 (20%) reactions worked in more than 24 but less than 32 samples, 43/71 (61%) reactions worked in 4-23 samples and 4/71 (6%) reactions worked in less than 4 samples. These results indicate less ubiquitous expression for the non-coding candidates than for the protein coding candidates with 18/71 (25%) reactions working in 4-8 samples. Four reactions worked 0-1 times according to the visual inspection. Reactions are furthermore listed by locus to enable a comparison between intronic and exonic primers at the same locus. There appears to be a general trend of reactions with intronic primers working in more samples than their junction spanning counterparts. 11/21 loci showed similar (+/- 3) results for exonic and intronic primer. The remaining 10 loci showed discrepancies between the number of times the exonic and intronic primers worked which were larger than the set margin of +/- 3. 9/21 loci had intronic primers working in more samples than the exonic primers while 1 locus had an exonic primer showing more results than the intronic one.

Reproducibility was assessed for the remaining tissues in the way described for heart. The results for reproducibility of PCR results for all tissues are summarised in table 8. 41-60 (58-85%) reactions worked in a reproducible way in all tissues, supporting the statement that reproducibility is lower than for the first set of candidates in the remaining tissues as well. Again these results were compiled after some reactions were repeated.

	brain	**heart**	**kidney**	**leg m.**	**liver**	**lung**	**spleen**	**thymus**
worked in all	36	35	21	21	14	35	33	27
worked in none	14	12	20	29	25	14	17	33
reproducible in all	50	47	41	50	39	49	50	60
	(70%)	(66%)	(58%)	(70%)	(55%)	(69%)	(70%)	(85%)

Table 8: Reproducibility of PCR reactions for non-coding candidates. An overview on how many reactions worked for either all (first row) or none (second row) of the four replicates per tissue. The last row gives the total number of reactions with reproducible results in all replicates.

As before the PCRs for this part were finalised with two PCRs for pure FVB and Cast strain samples. Suitable templates were selected in the same way and results showed that 52/71 reactions worked in both samples and 10 worked in none. Nine reactions worked inconsistently yielding a product in one of the two samples while failing in the other.

In summary the reactions for the non-coding candidates worked sufficiently well to continue with sequencing. The control reactions in FVB and Cast showed some discrepancies which will have to be considered when analysing imprinted expression.

2.3.4 PCR for non-coding targets in testes/placenta

Employing the same PCR and agarose gel electrophoresis analysis as in sections 2.2.2 and 2.3.3 PCR products for 23 candidates with suggested expression in testes and 2 candidates with suggested expression in placenta were obtained and analysed. This was done in two biological replicates, named FxC m2 and CxF m1, for testes and one sample of placental RNA provided by Quanah Hudson. Figure 23 shows the gel for these 48 PCR reactions and matching -RT controls. The negative controls for testes were made by pooling -RT reactions of both biological replicates.

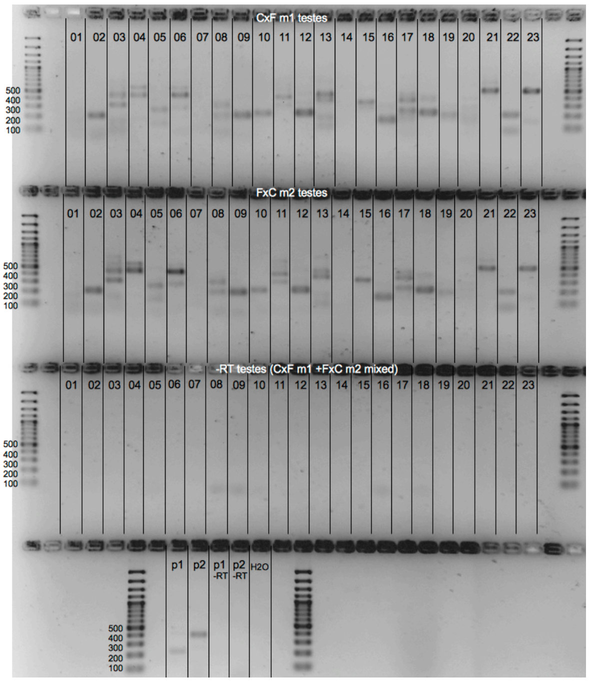
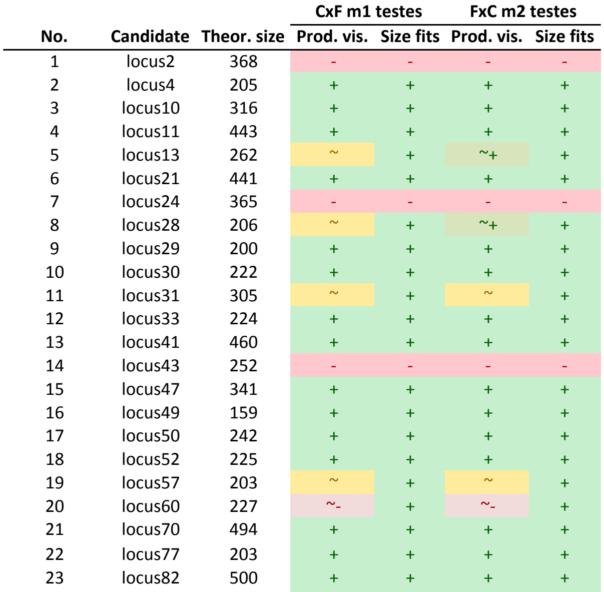


Figure 23: PCR results for non-coding candidates in testes/placenta samples visualised on an agarose gel. The first two rows show the results for the reactions in testes, the third row gives the corresponding -RT negative controls (mixed -RT samples of the two replicates) and in the last row the results for the two primer pairs used in a placenta sample can be observed. Overall 3 of 23 reactions showed no product in both replicates while the remaining 20 reactions all showed a product in both replicates. This indicated high reproducibility of the PCR results for this primer set. A detailed inspection of the results for testes is given in figure 24. Marker: GeneRuler 100 bp Plus DNA Ladder (Thermo Scientific).

Results show that most of the reactions worked in both replicates with just three not showing a product in both. This indicated that the results were highly reproducible. All products for reactions in placenta showed the right size. The gel image was visually inspected as before and results for the reactions in testes samples are given in figure 24. A more detailed analysis by visual inspection provided further evidence for the high reproducibility as even the intensity of the bands for each reaction was very similar between the two replicates. Testes samples were also sequenced and analysed regarding coverage and support of the junctions annotated by the de novo assembly.



			CxF m1 testes		FxC m2 testes	
No.	**Candidate**	**Theor. size**	**Prod. vis.**	**Size fits**	**Prod. vis.**	**Size fits**
1	locus2	368	-	-	-	-
2	locus4	205	+	+	+	+
3	locus10	316	+	+	+	+
4	locus11	443	+	+	+	+
5	locus13	262	~	+	~+	+
6	locus21	441	+	+	+	+
7	locus24	365	-	-	-	-
8	locus28	206	~	+	~+	+
9	locus29	200	+	+	+	+
10	locus30	222	+	+	+	+
11	locus31	305	~	+	~	+
12	locus33	224	+	+	+	+
13	locus41	460	+	+	+	+
14	locus43	252	-	-	-	-
15	locus47	341	+	+	+	+
16	locus49	159	+	+	+	+
17	locus50	242	+	+	+	+
18	locus52	225	+	+	+	+
19	locus57	203	~	+	~	+
20	locus60	227	~-	+	~-	+
21	locus70	494	+	+	+	+
22	locus77	203	+	+	+	+
23	locus82	500	+	+	+	+

Figure 24: Visual inspection of PCR results for non-coding candidates in testes/placenta samples. The same criteria as before, i.e. if there is a visible product (Prod. vis.) and if the product is of the right size (Size fits), were rated using the same scoring system. Scores are ranging from - (no product, color code: red) to + (strong product, green) with ~- (barely visible product, light red), ~(weak product, yellow) and ~+ (cleary visible but less than +, olive green, not used here) in between. Results appeared to be very reproducible as all samples worked in all or none of the two replicates and even the intensity of the bands was ranked in a similar fashion.

2.4 Massive parallel sequencing of PCR products from four tissues

The final part of the project covered in this thesis was the sequencing of the prepared PCR products. For this purpose purified PCR products from both sets were pooled together resulting in one pool per tissue per replicate. This was also done for the Cast/FVB controls, adding two additional pools. The last two samples to be sequenced were the two pools of PCR products from testes.

	tissues (4 repl.)	**additional sample**	**status**
set 1	brain, heart	FxC m2 testes	sequenced
set 2	kidney, leg muscle	CxF m1 testes	sequenced
set 3	liver, lung	Cast ctrl	libraries prepared
set 4	spleen, thymus	FVB ctrl	libraries prepared

Table 9: Sets of samples where libraries were prepared together and later multiplexed and sequenced on one lane. One set consisted of four replicates each of two tissues and one additional sample, either a testes sample, the Cast control sample or the FVB control sample. For the scope of this thesis two of these sets were sequenced and libraries were prepared for the other two.

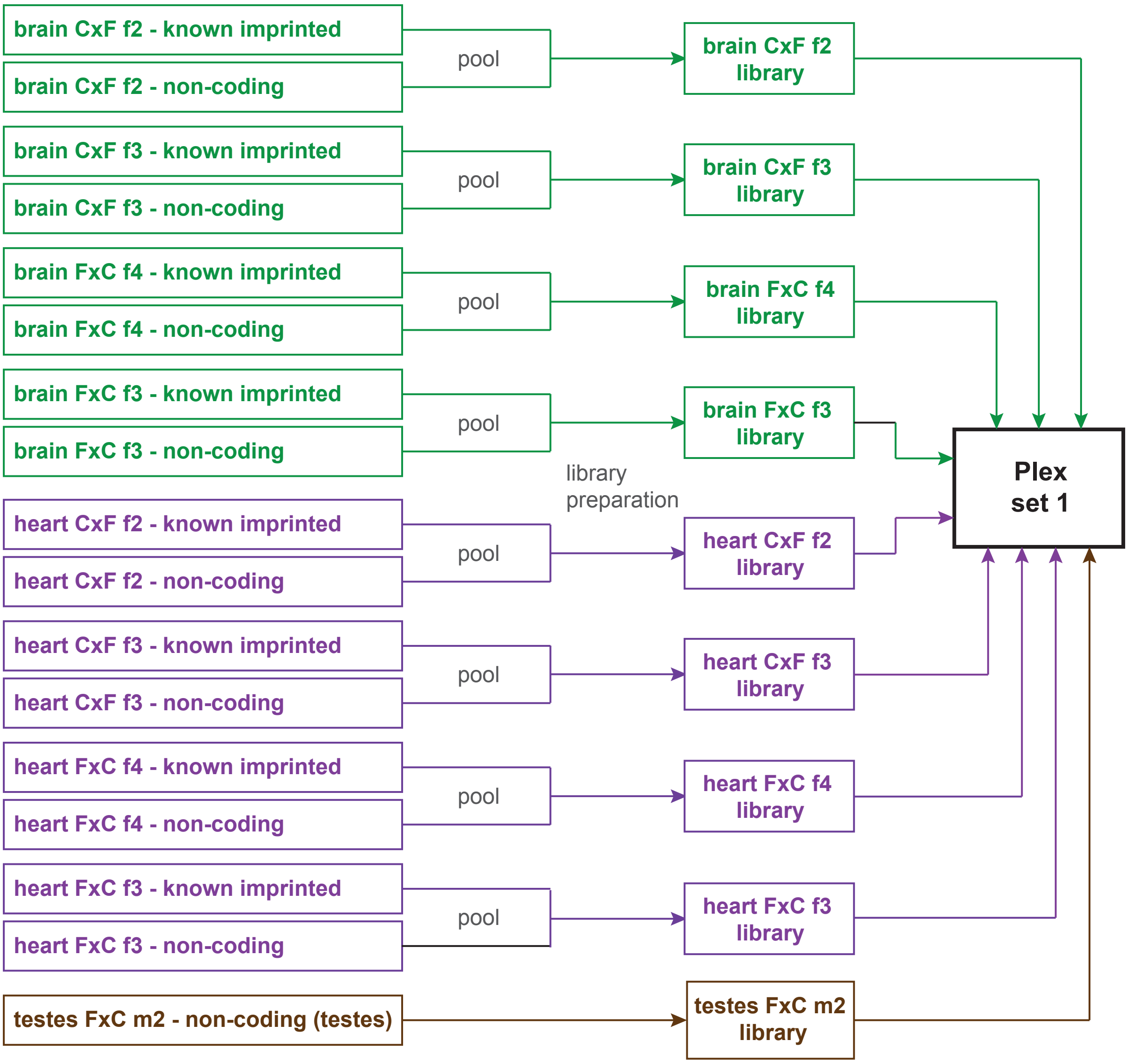

brain CxF f2 - known imprinted
brain CxF f2 - non-coding
pool
brain CxF f2 library
brain CxF f3 - known imprinted
brain CxF f3 - non-coding
pool
brain CxF f3 library
brain FxC f4 - known imprinted
brain FxC f4 - non-coding
pool
brain FxC f4 library
brain FxC f3 - known imprinted
brain FxC f3 - non-coding
pool
brain FxC f3 library
library preparation
Plex set 1
heart CxF f2 - known imprinted
heart CxF f2 - non-coding
pool
heart CxF f2 library
heart CxF f3 - known imprinted
heart CxF f3 - non-coding
pool
heart CxF f3 library
heart FxC f4 - known imprinted
heart FxC f4 - non-coding
pool
heart FxC f4 library
heart FxC f3 - known imprinted
heart FxC f3 - non-coding
pool
heart FxC f3 library
testes FxC m2 - non-coding (testes)
testes FxC m2 library

Figure 25: Workflow for the creation of a multiplexed sample for sequencing. Shown are the pooled PCR products used for the multiplexed set 1 sample. Both products for known imprinted and non-coding candidates were pooled together per replicate and tissue, libraries were prepared and plexed together in equal amounts. PCR products for non-coding candidates in testes were obtained using a different primer set than for the other tissues, which is indicated as "non-coding (testes). Samples are shown in green for brain, purple for heart and brown for testes.

This resulted in a total of 36 samples for sequencing and it was decided to divide these into four sets of nine samples each. The samples of one set were then multiplexed after library preparation and sequenced together on one lane on an Illumina HiSeq 2500 (see methods). The process of preparing these multiplexed samples is illustrated for set 1 in figure 25. This set contained brain and heart sample as well as one testes sample. The composition of all sets prepared in the way depicted in figure 25 is shown in table 9.

2.4.1 Sonication and library preparation

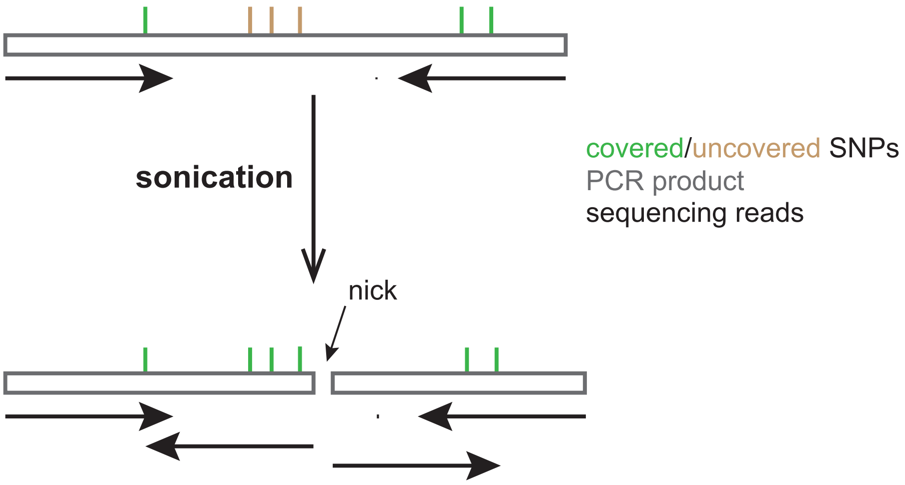
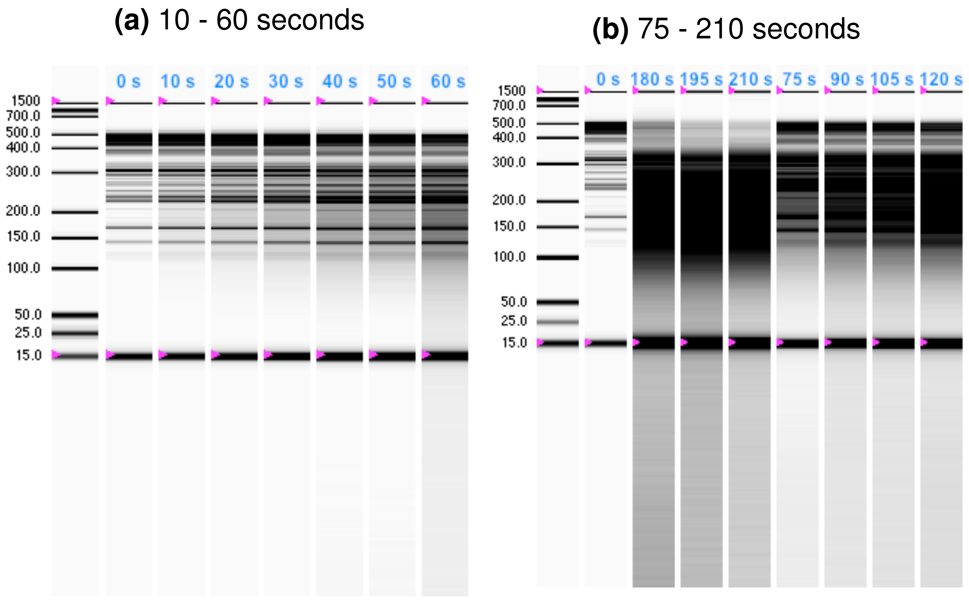


Figure 26: The reason why sonication is necessary is because massive parallel sequencing can only sequence from the two ends of DNA fragments. That is why SNPs located further within the product would not be sequenced this way (orange,top). Sonication introduces nicks and by sequencing those fragments SNPs that initially couldn't be sequenced are now accessible (blue, bottom). Shown are both SNPs accessible (blue) and inaccessible (orange) for sequencing in the respective situtation, an example PCR product of 500 bp and sequencing reads of 150 bp.

Before the PCR products could be sequenced they had to be sonicated to allow the full coverage of products larger than twice the length of a sequencing read. The reason for this is that 150 bp massive parallel sequencing is only able to sequence 150 bp from each end of the PCR product. Because of this, SNPs located in the middle of products larger than 300 bp would not be covered by sequencing reads. The sonication process introduces nicks into the PCR products and enables the sequencing of such SNPs. This is demonstrated in figure 26.

Since there were to my knowledge no references for sonicating PCR products of 100-500 bp a trial run had to be performed to determine the optimal duration of this process. This was done for one pool of PCR products for known imprinted protein coding genes of CxF f2 heart. Initially sonication times between 10 and 60 seconds were tested and the result was analysed using the Bio-Rad Experion^{TM} DNA 1K Analysis Kit. The resulting gel image produced by the kit is shown in figure 27a. When comparing the results of the sonicated samples with the untreated control (0 s) it became apparent that the general pattern of bands was still clearly visible even after 60 seconds of sonication. This was interpreted as most of the PCR products still being intact and therefore it was decided to try longer sonication times up to 210 seconds. Results for 75-105 and 180-210 seconds are displayed in figure 27b.

A difference to the shorter times in figure 27a could already be observed starting at 90 second with the continuous smear of the sonicated fragments starting to overshadow



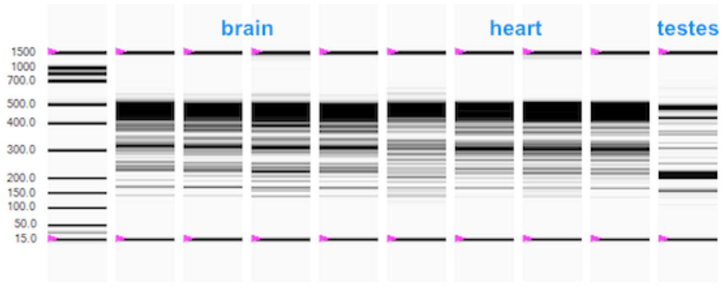
(a) 10 - 60 seconds

(b) 75 - 210 seconds

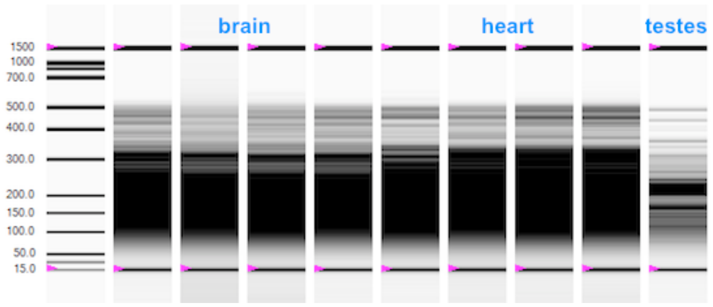
Figure 27: Optimisation of sonication times for the fragmentation of PCR products. (a) Results of the initial test run using sonication times between 10 and 60 seconds. The banding pattern visible in the original sample (0 s) was still visible after sonication. (b) Results from using longer sonication times between 75 and 210 s. The band pattern began to change into a continuous smear at a sonication time of 90 seconds but the largest band around 500 bp was still very prominent. Longer sonication times above 180 seconds showed a clear reduction in the intensity of the top band indicating sonication of the largest products.

the band pattern of the original sample. This smear is indicative of successful sonication as the random introduction of nicks produces a continuous spectrum of fragment sizes. Still, the largest products around 500 bp remained unchanged. When taking a look at higher sonication times between 180 and 210 seconds it could be observed that even the highest band had a clearly reduced intensity suggesting fragmentation of the largest products. In addition the results showed that most fragments have a maximum size of around 300 bp and a minimum size of 100 bp, suggesting that fragmentation below 100 bp is happening rarely. I therefore concluded that sonication for 210 seconds resulted in good overall fragmentation of PCR products without overfragmentating them indicating that the majority of fragments could still be sequenced.

Sonication of samples for library preparation of sets 1 and 2 was done using a sonication time of 210 seconds. An example comparison of unsonicated and sonicated samples is given for set 1 in figure 28. The results show that the pattern visible before sonication (figure 28a) is less visible afterwards and also the intensity of the strong band around 500 bp is greatly reduced in all eight samples of brain and heart. The testes sample had a different pattern from the beginning since only 23 reactions are pooled here compared with 167 reactions for the other samples. Still, the testes sample shows similar results with a general reduction of the average fragment size,



(a) Set 1 before sonication



(b) Set 1 after sonication

Figure 28: Sonication of set 1. Samples from left to right are: brain CxF f2, CxF f3, FxC f4, FxC f3, heart CxF f2, CxF f3, FxC f4, FxC f3 and testes FxC m2. The lane on the far left contained the ladder used by the kit. The upper image (a) shows the pooled samples of set 1 before sonication while the lower image (b) depicts the results of the sonication. The results are comparable with the results seen in the preliminary test run (figure 27b. Sonication time was 210 seconds.

especially for the band around 500 bp. Results for the other sets were comparable to set 1. Following sonication, libraries were prepared for one set at a time using the Illumina® TruSeq® ChIP Sample Preparation Kit (see methods) using 5 ng of sonicated sample. The quality of the libraries was assessed using the Experion kit (see methods). The results for set 1 are given in figure 29.

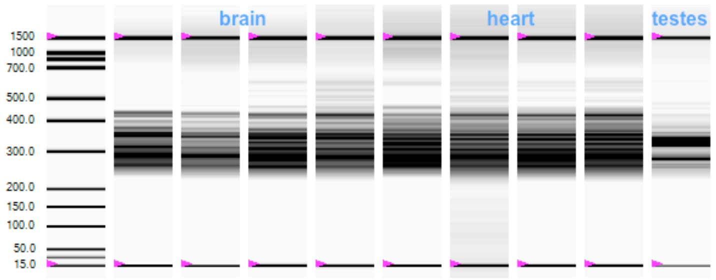


Figure 29: Quality check of the finished library for set 1. Samples from left to right are the four replicates for brain, heart as well as one testes sample in the same order as in figure 28. The first lane on the left contained the ladder provided by the kit. The results provided evidence that the quality of the libraries was fine.

These results showed that library preparation worked well, yielding fragments between 250 and 450 bp (including adapters). The results for the other sets were comparable. Samples of one set were multiplexed and both multiplexed sets were sent for 150 bp single end sequencing by the CSF at the Vienna Biocenter.

2.4.2 Analysis of sequencing results

This analysis focused on the general coverage of the PCR products, going into more detail on coverage of SNPs which could be used for downstream applications. Furthermore *de novo* assembled exon models were verified by checking the coverage of the non-coding candidate PCR products in more detail. General statistics about the sequencing runs and the following alignment (see methods) are given in table 10. The total number of reads given is the input used for the alignment and therefore after trimming of adapter sequences. Overall the results looked promising with around 16-17 million input reads per sample and at least 96.6% of reads uniquely aligning to the genome.

Set 1			**Set 2**		
Sample	**Input reads**	**Uniquely mapped**	**Sample**	**Input reads**	**Uniquely mapped**
brain CxF f2	16397466	98.27%	kidney CxF f2	15466081	98.74%
brain CxF f3	15979401	98.48%	kidney CxF f3	15213094	98.68%
brain FxC f4	17995745	97.34%	kidney FxC f4	14933909	98.50%
brain FxC f3	16517000	98.35%	kidney FxC f3	16899625	98.71%
heart CxF f2	17628788	96.05%	leg m. CxF f2	15242369	98.56%
heart CxF f3	15750082	98.32%	leg m. CxF f3	14430577	98.82%
heart FxC f4	16440141	97.66%	leg m. FxC f4	15604974	97.43%
heart FxC f3	16437369	98.53%	leg m. FxC f3	16223829	98.71%
testes FxC m2	16759718	98.05%	testes CxF m1	14630728	96.62%

Table 10: General information regarding the sequencing results. Shown are the number of input reads left after the removal of adapter sequences. Furthermore the percentage of uniquely aligned reads (Uniq. mapped). for these remaining reads is given.

The quality of the sequencing data was checked using FastQC[5]. The results are shown in figure 30 for CxF f2 kidney, which was chosen as a representative sample. Base quality was good across the whole read with decreased but still good quality at the beginning and end of the reads. These results, together with the overall read counts and mapping percentage given in table 10 showed that the chosen sequencing strategy was a suitable choice for this experiment.

---

[5] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

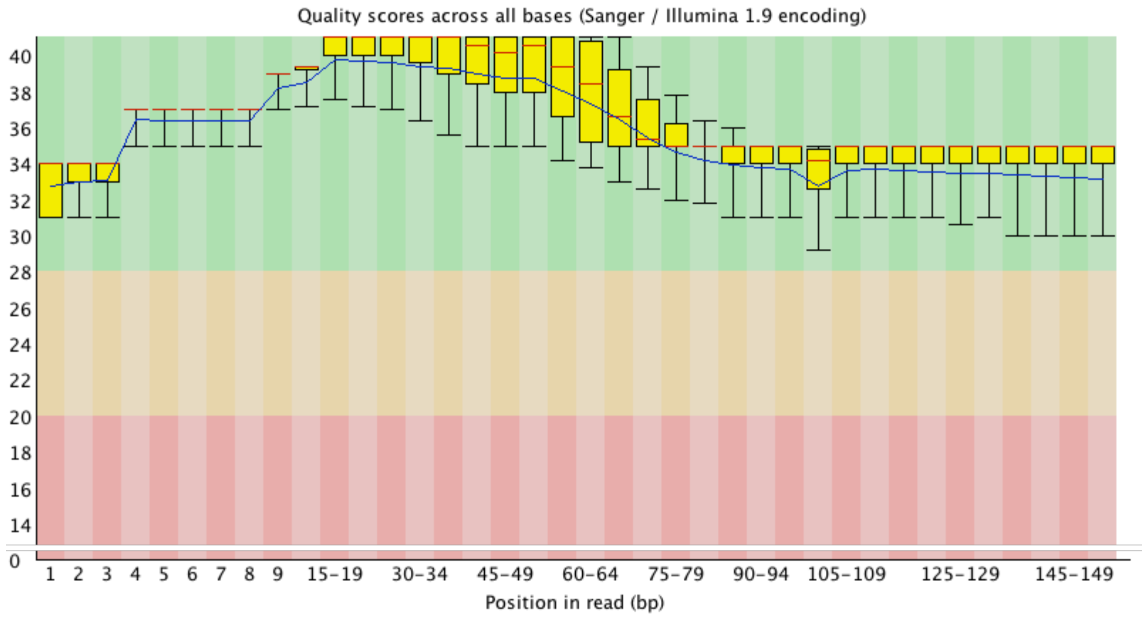


Figure 30: Quality assessment of sequencing data for CxF f2 kidney. The graph shows the average base quality scores across the 150 bp reads. Green/Yellow/Red indicate good, medium and bad scores according to the tool.

Coverage of protein coding targets

The coverage was assessed to calculate the fraction of the PCR product covered by sequencing reads. Subsequent data analysis using R was performed to assess the coverage of the whole PCR product for the known imprinted protein coding genes and the coverage of the assembled splice junctions for the non-coding candidates (see methods). The results for the protein coding candidates showed that 88-93/96 PCR products were covered in all four replicate samples of one tissue. A PCR product was rated as covered when 80% of its sequence were supported by at least one read allowing the inclusion of lowly covered PCR products. Two candidates which were covered in very few replicates were *Ins1*, which was only covered in one replicate each of kidney and brain and *Dio3*, which was covered in all brain replicates but besides that only two times each in kidney and legmuscle and once in heart. Other PCR products not covered in all replicates of at least three tissues were *Ascl2*, *Ddc*, *Klf14*, *Mest*, *Snrpn* and *Th*. Overall the results showed that 85/96 (89%) PCR products were covered in all 16 replicates examined and that all products were covered in at least two samples. The results are visualised in figure 31.

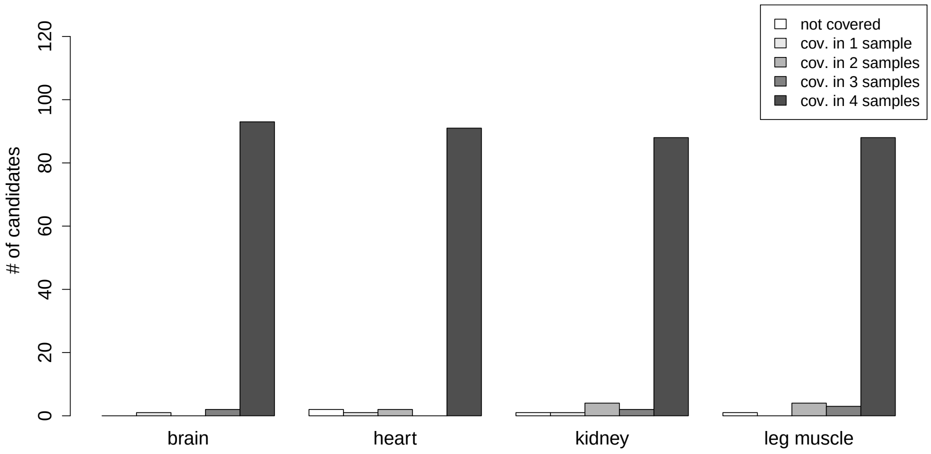


Figure 31: Coverage of PCR products for known imprinted protein coding candidates. PCR products were grouped per tissue based on how many replicates they were covered in. The y-axis represents the number of products in each group. The results show a high percentage of PCR products covered in all four replicates within a tissue.

Coverage of non-coding targets in eight adult tissues

The same analysis was done for intronic PCR products of non-coding candidates. Junction spanning PCR products were used for the validation of splice junctions and were therefore described per junction instead of per product. The same coverage cutoff as before was used, requiring 80% coverage by at least one read.

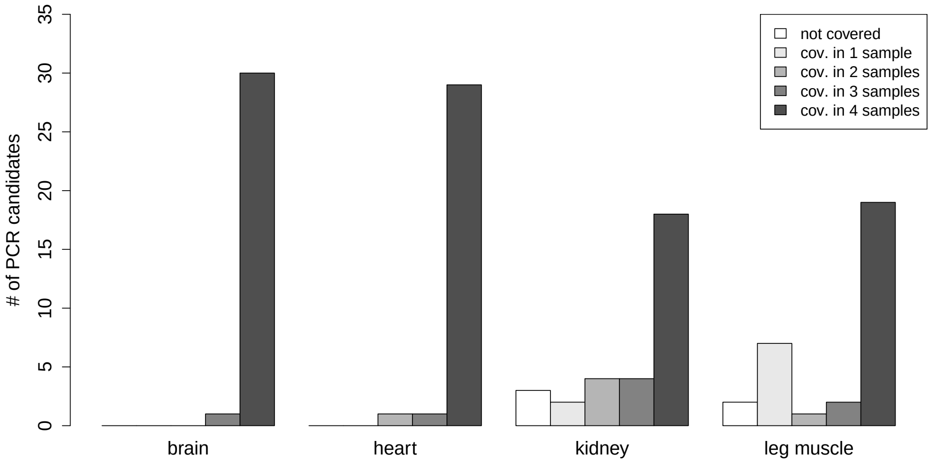


Figure 32: Coverage of intronic PCR products for non-coding candidates. The graphs display the amount of PCR products covered in the indicated number of biological replicates in one tissue for all 31 intronic primer pairs. At least 80 % coverage by at least one sequencing read was required to call a product covered. The results show a high percentage of PCR products covered in all four replicates within brain and heart but a reduced coverage in kidney and leg muscle.

Results of the coverage assessment for 31 intronic PCR products are given in figure 32. Almost all products are covered in all four replicates of brain and heart with 30 and 29 products covered, respectively. Kidney and leg muscle results show only 18 and 19 products covered in all replicates. The set of junction spanning PCR products encompassed 40 products for all eight tissues and 23 products only investigated in testes. These products covered 83 junctions in candidates for all tissues and 33 junctions in candidates for testes. The coverage of a junction was evaluated by checking the coverage of the two exons between which the junction had been annotated. The coverage was determined the same way as before, classifying exons as covered if 80% are supported by at least one read. The cutoff of 80% was chosen to account for small differences between the annotated exon structure and the experimentally observed one. Summary statistics were created for each tissue separately to enable correlation between coverage and expression data from the MIRTAs. Candidates were grouped into two bins, one bin showing a negative average MIRTA signal indicating no expression, while the other had a positive signal average, indicating expression. Results for the four tissues investigated are given in figures 33-36.

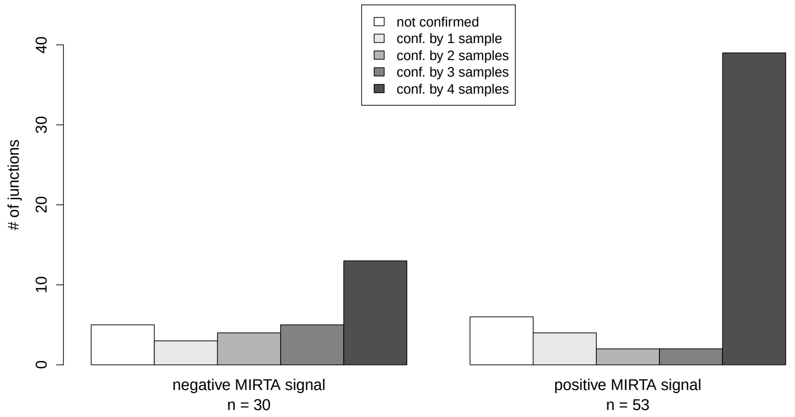


Figure 33: Validation of *de novo* assembled splice junctions in brain. The validation of junctions was summarised showing how many of the four reactions in four biological replicates showed a PCR product supporting the junction, according to coverage of the exons forming the junction. In addition junctions were grouped into two bins according to whether the candidate containing the junction is expressed in this tissue or not. A negative MIRTA signal is an indicator of no expression while a positive signal suggests expression of the target.

The results for brain in figure 33 showed that 53 junctions belong to candidates with MIRTA data suggesting expression in this tissue. The junctions in this "expressed" group showed a higher number of replicates supporting them with 39/53 junctions (74%) in this bin being supported in all four replicates compared to 13/30 (43%) in the "not expressed" bin. Both bins contain junctions not supported within any of the replicates suggesting that the transcript annotated in the *de novo* assembly might not be expressed in brain. Overall 11 junctions (13%) could not be validated in brain and

52 (63%) were confirmed in all four replicates.

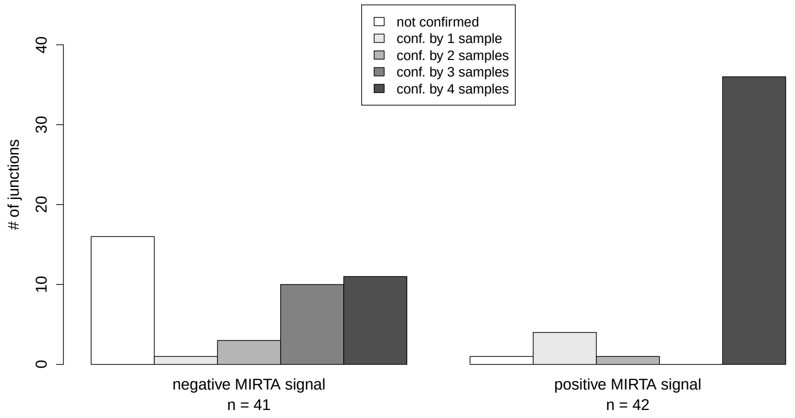


Figure 34: Validation of *de novo* assembled splice junctions in heart. Junctions were validated and grouped into bins as described in figure 33.

Results in heart showed a similar picture to brain, although less junctions were expressed according to the MIRTA data. As before the "expressed" bin contained a high number of junctions, 36/42 (86%), confirmed in all replicates. The other bin, in comparison, only contained 11/41 supported junctions (27%). In summary 16 junctions (19%) were not confirmed in any of the four replicates while 49 junctions (59 %) could be validated in all of them.

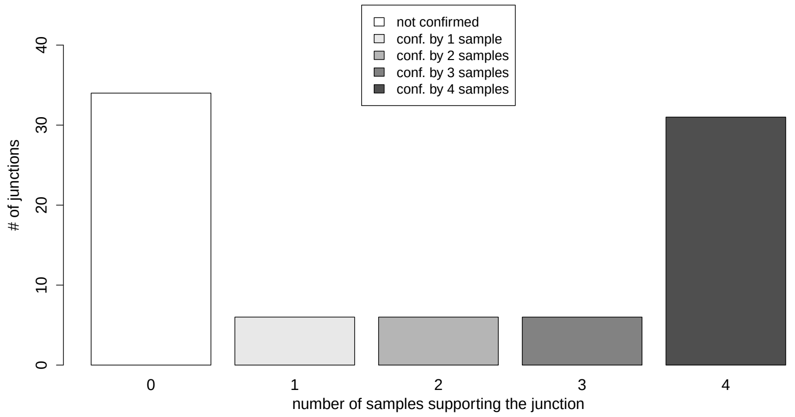


Figure 35: Validation of *de novo* assembled splice junctions in kidney. Junction coverage was assessed as described for figure 33. No grouping according to expression level was performed since no MIRTA data was available for this tissue.

Figure 35 shows the results for kidney. Since the MIRTA data did not include kidney samples no expression data was available for this analysis. Therefore no grouping according to expression level was performed and all 83 junctions were investigated together. Compared to the last two tissues the amount of junctions validated in all four replicates was lower with only 31 junctions (37%) being confirmed. On the other hand, the number of junctions which could not be verified showed a two-fold increase compared to the two tissues before with 34 (41%) junctions not confirmed in kidney. This suggested that less of the candidates are actually expressed in kidney, relative to brain and heart.

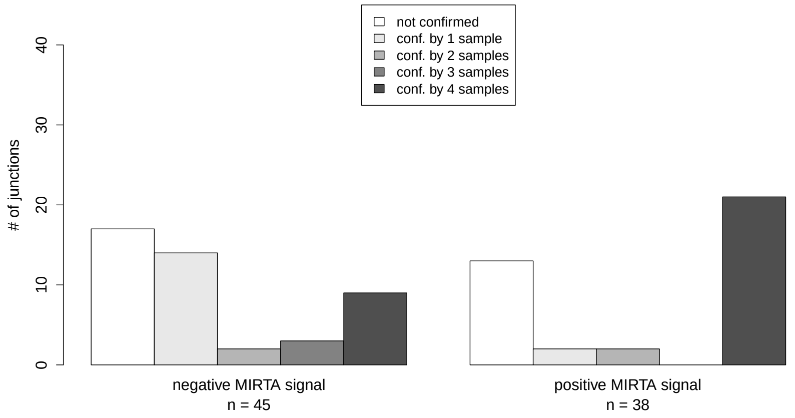


Figure 36: Validation of *de novo* assembled splice junctions in leg muscle. Junctions were validated and grouped into bins as described in figure 33.

The results for leg muscle are shown in figure 36. 13/38 (34%) junctions in the "expressed" bin and 16/45 (36%) junctions in the "not expressed" bin could not be confirmed in any of the four replicates. In addition the amount of junctions only confirmed in just one replicate, 16/83 (19%), is at least two fold higher than for the other tissues. Overall 29 junctions (35%) could not be confirmed in leg muscle and 30 (36%) were validated in all four replicates. Both of these numbers are similar to what was observed for kidney.

As a summary of the junction validation the combined results for brain, heart, kidney and leg muscle are shown in figure 37. Junctions were grouped into four bins according to whether they were not confirmed at all or confirmed in 1-3, 4-7, 8-11 or 12-16 samples. Of the 35 junctions confirmed in 12-16 samples 21 could be validated in all 16 replicates. Overall 68/83 junctions could be confirmed in at least four replicates and 10 junctions could not be validated in any of the investigated samples.

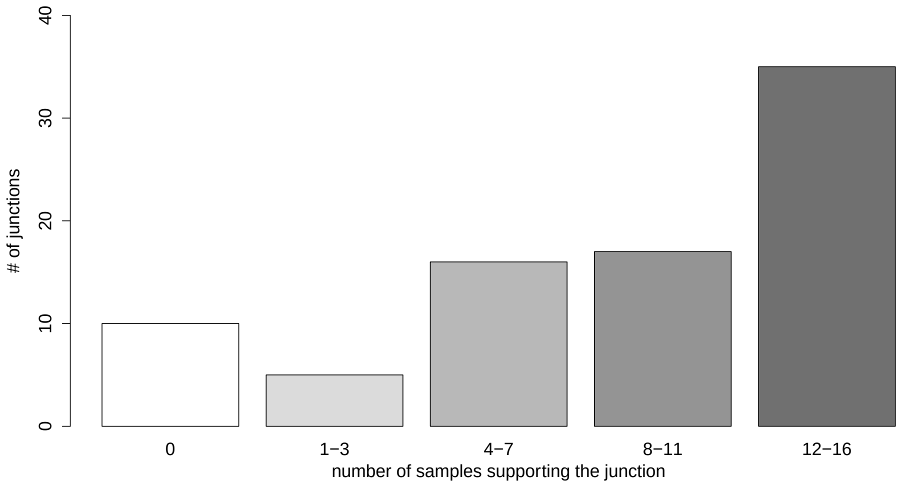


Figure 37: Summary of junction validation in all four tissues combined. Results were integrated by summing up the number of replicates a junction was validated in across all four tissues. Five bins were created representing junctions that were not confirmed in any sample or junctions that could be confirmed in 1-3, 4-7, 8-11 and 12-16 samples, respectively.

Coverage of non-coding targets in testes

Coverage for the 33 junctions covered by PCR products in testes was assessed by checking whether the exons taking part in the junction were at least 80% covered. The results of the analysis are given in figure 38. A high percentage of junctions, 31/33 (94%), could be confirmed in both replicates while 2/33 (6%) could not be confirmed in either of them.

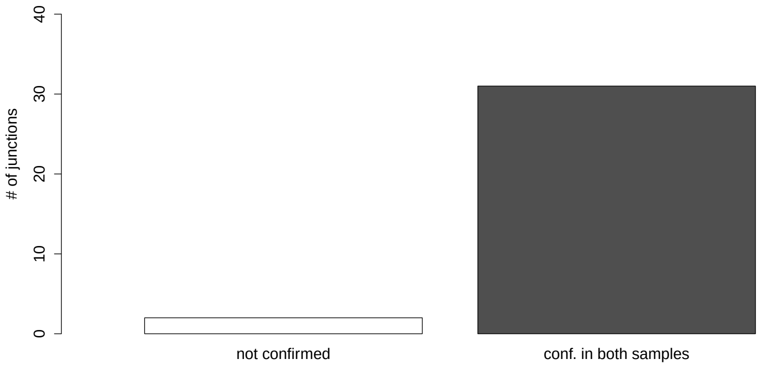


Figure 38: Validation of *de novo* assembled splice junctions in testes. 33 junctions were assessed for PCR products in testes and grouped according to whether they were supported in none or both replicates. No junctions were validated in just one replicate. Junctions were validated as described for figure 33.

The results also demonstrate that junctions in testes were either confirmed in both replicates or not confirmed at all, showing high reproducibility of the PCR results between these two replicates (see figure 24).

Coverage of SNPs located within the PCR products

Since this sequencing data was primarily produced to be used for the analysis of imprinted expression the coverage of the SNPs located within the PCR products was assessed. SNP coverage was extracted from the pileup of sequencing reads using samtools (see methods). A total of 727 SNPs were located within the PCR products for both known imprinted protein coding and non-coding candidates combined. The computed coverages for these SNPs is plotted in figure 39.

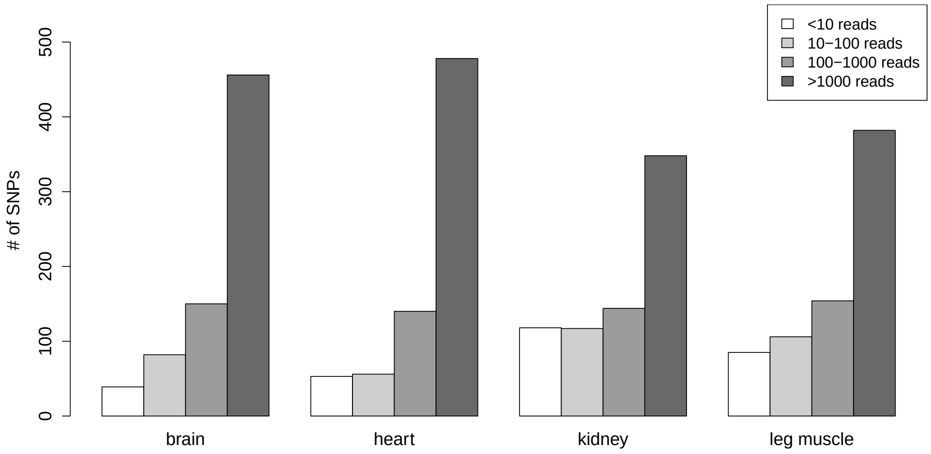


Figure 39: Summary of SNP coverage in all PCR products and four tissues SNP coverage, as calculated from sequencing pileups, was plotted for each tissue and all PCR products from both the known imprinted protein coding and the non-coding set. SNPs were grouped according to their coverage into four bins.

SNPs were grouped into bins according to their coverage, setting the first cutoff to 10 reads as a minimum coverage required to detect imprinted expression. The other categories were defined as SNPs covered by 10-100, 100-1000 and more than 1000 reads, respectively. Overall, the majority of SNPs was well covered by more than 1000 reads allowing a good statistical analysis of imprinted expression. However, there were differences in overall SNP coverage between brain/heart on one hand and kidney/legmuscle on the other hand with less SNPs covered by more than 1000 reads, i.e. (~460 vs ~370), and around twice as many SNPs covered by less than 10 reads.

3 Discussion

The scope of this thesis included implementing a targeted PCR based approach employing automated primer design to allow the subsequent investigation of the imprinting status of both known imprinted protein coding candidates and known and novel non-coding candidates. PCR was performed for these two sets in eight adult mouse tissues, i.e. brain, heart, kidney, leg muscle, liver, lung, spleen and thymus, with four replicates for each tissue. The sets were chosen to allow for both an improvement of the description of imprinted expression in adult mouse tissues on the one hand and, on the other hand, a validation of the annotation of imprinted non-coding RNAs by evaluating their imprinting status in adult tissues and possibly discovering novel imprinted lncRNAs. Library preparation and sequencing was done for four tissues for the data presented in this thesis. The analysis in this thesis focused on the feasibility of the targeted approach by assessing the coverage of PCR products by sequencing reads and how well this data could be used for a subsequent analysis of imprinted expression. Another part of the analysis was the validation of *de novo* assembled spliced transcripts by investigating the coverage of junction spanning PCR products.

3.1 Evaluation of the imprinting status of *Igf2r* and *Airn* agrees with previous findings

As a validation for this method a pilot experiment was performed assessing the imprinting status of *Igf2r* and *Airn* in the eight tissues mentioned above. Previous studies showed that *Igf2r* is exclusively expressed from the maternal allele in both post-implantation embryonic stages as well as adult stages [13, 127, 128]. The only adult tissue with biallelic expression of *Igf2r* is brain [128]. This was later shown to be restricted to post-mitotic neurons, whereas glial cells show only maternal expression of *Igf2r* [52]. *Airn* is a paternally expressed non-coding RNA which partially overlaps *Igf2r* in antisense and has been shown to be the main regulator of imprinted expression in the *Igf2r* cluster [51]. Imprinted expression in adult tissues was determined by designing SNP covering PCR products and evaluating the relative abundance of maternal/paternal variance by quantitation of Sanger sequencing results (figure 7). The results showed 91-96% paternal expression for *Airn* across all tissues which agrees that expression of *Airn* is imprinted in all of the examined tissues. *Igf2r* showed 90-99% maternal expression in all tissues except brain. Expression of *Igf2r* in the brain was almost biallelic with a slight bias towards the maternal allele resulting in a 63:37 maternal:paternal ratio. Since the brain consists of both neurons and glial cells these results can be explained by the relaxation of imprinted expression of *Igf2r* in neurons on the one hand and the exclusively maternal expression in glial cells on the other hand. Overall these results were a validation of the targeted method in showing that imprinted expression can be detected correctly by sequencing of SNP covering PCR products. For the future it would be interesting to examine DNA methylation of the promoter regions in an allele specific manner and see how these findings match the results for imprinted expression presented here.

Previously published expression data on *Airn* and *Igf2r* grouped adult tissues into three categories, i.e. tissues with high, medium and low steady state levels of *Airn* and *Igf2r*. [129] According to this data, heart is the only tissue in the set of eight tissues showing high expression, lung and thymus belong in the medium group and brain, kidney, liver and spleen show low levels of *Airn* and *Igf2r*. When comparing this expression data with the imprinting results, no correlation between steady state level and maternal/paternal bias could be observed for neither *Igf2r* or *Airn*. This indicates that the ability of *Airn* to silence *Igf2r* is independent of steady state levels of *Airn* which is in agreement with findings that *Airn* is necessary for initiating imprinted *Igf2r* silencing but is dispensable after DNA methylation of the *Igf2r* promoer has been established [89].

3.2 The targeted PCR approach

PCR based targeted approaches are a known method for target enrichment in massive parallel sequencing. Compared to enrichment by hybridisation, PCR based approaches excel at sensitivity and specificity but do not scale easily and are therefore not suited for analysing large regions of the genome [130, 131]. Target enrichment strategies are widely used in the discovery of DNA mutations in cancer [132] or other genetic disorders [133, 134, 135]. The approach used in this thesis differs from these experiments in that it is an analysis of the RNA to investigate the imprinting status of target genes based on already known SNPs rather than testing genomic DNA for priorly unknown variants. The motivation behind choosing a targeted approach for this project was to increase sensitivity for the detection of imprinted expression of lowly expressed candidates. Furthermore the design of junction spanning PCR products allowed the specific investigation of overlapping isoforms of a gene as long as the exon structure is not completely identical. Another factor was the cost of the analysis, which enabled the investigation of a larger number of tissues.

3.2.1 The targeted approach is feasible and cost efficient but more time consuming than Whole Transcriptome Sequencing methods

A custom script was implemented for the design of suitable primers, allowing for a faster and more streamlined process of designing the primers needed for this analysis. The main criterion was the inclusion of SNPs in the PCR product, which could subsequently be used to investigate imprinted expression. Two sets of 96 primers each were prepared for both the set of known imprinted protein coding targets (set 1) and the set of non-coding targets (set 2). The resulting primers were tested and redesigned if needed. Overall most of the designed primers worked as expected with only 6 and 5 primer pairs selected for redesign in sets 1 and 2, respectively. Isoform specific primer pairs were designed in 12 cases, 10 known imprinted and 2 ncRNAs, where isoforms of one candidate had different transcriptional start sites, the most notable being *Grb10* since it has been reported that the two isoforms expressed in brain show differentially imprinted expression, i.e. one being paternally expressed while the other one is maternally expressed [60].

Besides practicality another factor determining whether this approach was a suitable choice or not was the cost efficiency. This factor was assessed using a comparison with a whole transcriptome sequencing (WTS) approach. It has been shown in preliminary studies in the lab that it is possible to study imprinted expression using WTS data (D. Andergassen, unpublished data). However this requires a more expensive library preparation than for the approach used in this thesis. Furthermore less libraries can be multiplexed in a single sequencing run making the sequencing more expensive as more sequencing runs are needed. Lastly 100 bp paired end sequencing with HiSeq chemistry is required for WTS which is more expensive and time consuming than 150 bp single end sequencing with MiSeq chemistry. In total the WTS approach would need nine lanes on a flow cell compared to four lanes for the targeted approach. Considering only library preparation and sequencing costs this would come to a total of around € 20000 for the WTS approach. The expenses for the targeted approach, on the other hand, consisted of fees for the around 6500 PCR reactions (€ 1800), the primers (€ 1200), library preparation for 36 half reactions (€ 800) and the cost for 4 lanes of 150 bp single end sequencing using MiSeq chemistry (€ 4000). This brought the total to around € 7800, equal to around 39% of the cost of the whole transcriptome sequencing approach. While this reduction in cost is certainly an advantage of the targeted approach, one should also consider that this approach required a considerably larger amount of work, i.e primer design, PCRs and agarose gels. Overall the work could be done in a quick and straight forward way once the primer design script was implemented, the targets were selected and primer design and selection were completed. The bulk of PCR reactions and corresponding gels could be completed within three weeks. Overall this lead to the conclusion that the reduction of the costs by more than half justified the increase in workload and time it took to complete the analysis.

3.2.2 PCR results are highly reproducable across biological replicates

Investigation of the reproducibility of PCR results (see chapters 2.2.2 and 2.3.3) already showed high reproducibility of PCR results which was then further improved by repeating some reactions. Overall 3072 (96 x 32) PCR reactions were performed for the known imprinted candidates in set 1 and 2272 (71 x 32) for the ncRNAs in set 2. 52 and 32 were selected for repetition for set 1 and 2, respectively. After improving the reproducibility by these repetitions, set 1 showed highly reproducible results with around 86-97% of reactions working either in all or none of the four replicates of a tissue. The reactions of the second set worked less reproducible with 55-85% working in all or none of the replicates, but the reproducibility was deemed high enough to refrain from doing further repetitions. One of the reasons behind this was that these candidates are most likely not as highly expressed as the candidates of the first set and therefore the amount of PCR product in the volume loaded on the gel might not have been enough to give a visible band. However even though the product was not visible on the gel did not necessarily mean that sequencing would not work as well, as described in the next section.

3.2.3 Visual quality checks by agarose gel electrophoresis give a good estimate on sequencing coverage

Agarose gel electrophoresis was used as a tool for immediate evaluation of the success of the PCR reactions. 5 µl of 25 µl total reaction volume were loaded on the gel and the results were checked using a list of theoretical PCR product sizes compiled by the script (see figure 9 for an example). The hypothesis was that this was a good estimator of PCR success and therefore will most likely be an estimator whether the massive parallel sequencing of the PCR products in the next step will work or not. This was verified by correlating the sequencing results with the results from the visual inspection. The results are depicted in figure 40. PCR reactions were grouped according to tissue and primer set, 384 reactions per known imprinted bin, 284 per non-coding bin, and classified into three categories. The first category, displayed in dark grey and comprising the majority of PCR reactions, is defined as reactions showing matching results between visual inspection and sequencing data. This could be either a reaction with a visible product which was also covered or a reaction with no visible product which was not covered. In either case the gel analysis predicted the outcome of the sequencing in a correct way. The other two categories contained reactions where this prediction was false, either by showing coverage although no product was visible (grey, middle) or by showing no coverage although a product was visible on the gel (light grey, top).

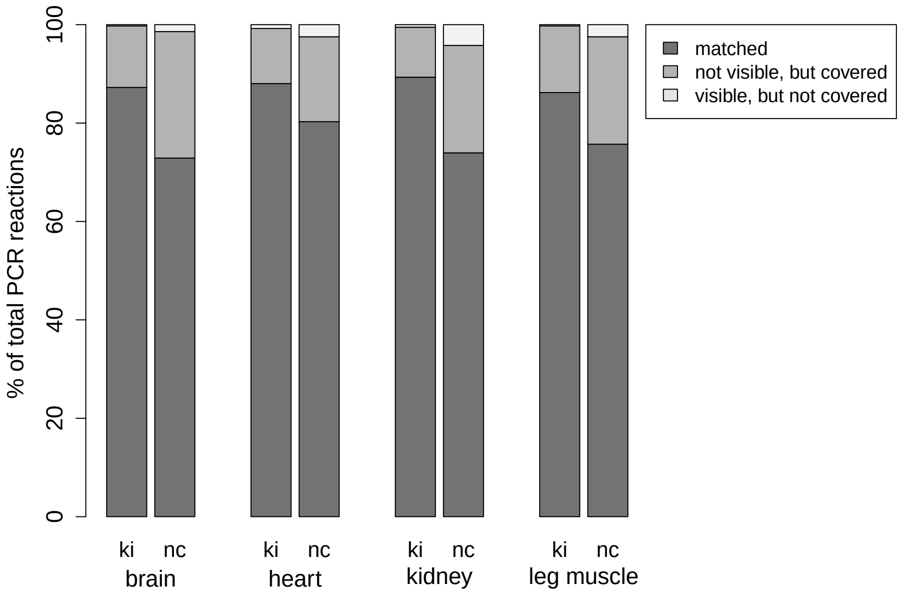


Figure 40: Visual inspection as a qualitative predictor of sequencing success. The results of the correlation between visual inspection and sequencing coverage are plotted as the relative abundance of PCR reactions that showed agreement between visual inspection and sequencing coverage (dark grey), showed no visible product on the gel but were still covered by sequencing reads (grey) or showed a visible product but were not covered by sequencing reads (light grey). PCR reactions are divided into two bins corresponding to the two primer sets of known imprinted (ki) and non-coding (nc) targets. Total numbers of reactions per tissue were 384 (4x96) for known imprinted candidates and 284 (4x71) for the non-coding candidates.

Overall a few trends were visible when analysing these results. First, it appeared that the estimation worked better for the known imprinted candidates with 87% of all reactions showing matching results compared to 75% for the non-coding candidates. Second, there were only little differences across tissues concerning the categorisation. The majority of the wrongly predicted reactions were classified as being covered by sequencing reads despite showing no product on the gel. These findings provided evidence that the detection threshold of the visual analysis is higher than for sequencing. Therefore very little amounts of product, which could nevertheless still be sequenced, do not get detected by this method. The reactions for the protein coding set assigned to the third category all belonged to two candidates, i.e. *Ins1* and *Mest*. *Ins1* is only expressed in pancreatic β-cells in adult tissues [136] which explained why the reaction barely worked at all. The classification of some reactions into the third category might be the result of secondary products of similar size which were interpreted as a product of the correct size. *Mest* is an example for a caveat of the method used to determine the coverage and will be described in more detail in the next section.

3.3 Coverage of PCR products

Sequencing data was analysed for four of the eight prepared PCR product pools. In addition to that the two pools of PCR products for testes candidates were also sequenced and analysed.

3.3.1 Known imprinted protein coding candidates were well covered

The results for the known imprinted candidates of the first set showed high coverage of the PCR products by sequencing reads. Varying between tissues 88-93/96 PCR products were covered in all four replicates of one tissue with 85 products covered in all 16 replicates. One candidate, *Ins1*, barely worked in any of the 16 replicates and was covered only in one replicate each of kidney and brain. As mentioned before this could be explained by the tissue-specific expression of this candidate. *Dio3* was only covered in the four brain replicates and five replicates of other tissues, which is also in line with the finding that this gene is primarily expressed in the central nervous system and is down regulated in later stages of development [137]. Among the other candidates showing coverage in less than 12 samples, *Mest* was a special case. Since MIRTA data suggested expression of this candidate in the investigated tissues a closer look was taken at the sequencing results at this locus. This revealed that the lack of full coverage came from the fact that the PCR product was designed for an isoform with an alternative second exon. This exon appeared to be barely used at all, as supported by the very low RNA sequencing signal in this region, resulting in reduced coverage of this exon. This lead to a reduction of the overall coverage below the threshold of 80% and the subsequent classification of this reactions as "uncovered". The structure of *Mest* is illustrated in figure 41. In summary coverage of PCR products for this primer set was high and should provide suitable data for a downstream analysis (see also 3.4).

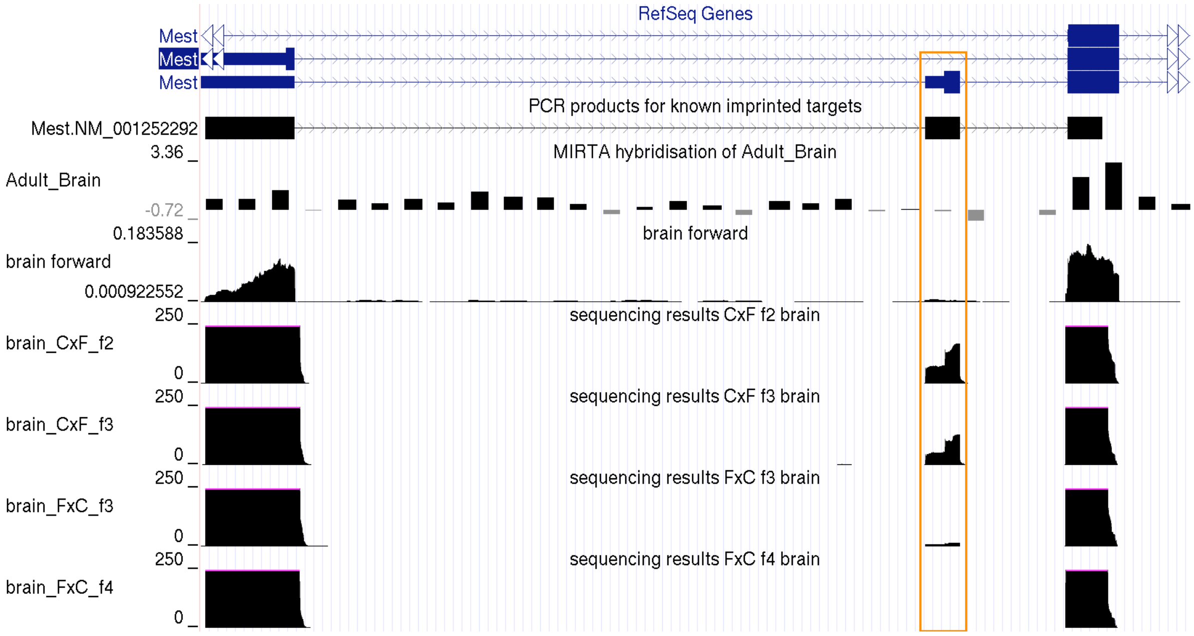

RefSeq Genes
Mest
Mest
Mest
PCR products for known imprinted targets
Mest.NM_001252292
3.36
MIRTA hybridisation of Adult_Brain
Adult_Brain
-0.72
0.183588
brain forward
brain forward
0.000922552
250
sequencing results CxF f2 brain
brain_CxF_f2
0
250
sequencing results CxF f3 brain
brain_CxF_f3
0
250
sequencing results FxC f3 brain
brain_FxC_f3
0
250
sequencing results FxC f4 brain
brain_FxC_f4
0

Figure 41: Susceptibility of coverage analysis to alternative exons - Example: *Mest*. This figure shows a UCSC genome browser screenshot as described in more detail in figure 1. The tracks shown are (from top to bottom): RefSeq annotation of three different *Mest* isoforms, the alignment of the junction spanning PCR product for this target, MIRTA data from adult brain, RNA sequencing data from the forward strand in adult brain and sequencing data for all four brain replicates. Pink lines on top of the signal indicate truncated peaks, since the size was set to a fixed value of 250 to enable viewing the alternative exon. The figure demonstrates a situation where the presence of an alternative exon in the isoform used for the design of the PCR product might lead to a misclassification of this candidate as uncovered. The alternative exon is hightlighted in orange and is supported by only a low number of RNA sequencing reads. Sequencing results show coverage in two replicates, very low coverage in one and no coverage in the fourth replicate.

3.3.2 Non-coding candidates showed tissue specific differences in coverage

The second set was divided into intronic and junction spanning PCR products for the coverage analysis. The results of this analysis for the intronic candidates showed results similar to the known imprinted targets for brain and heart. 30 and 29/31 products were covered in all four replicates of these tissues. In kidney and leg muscle the number of PCR products covered in all replicates is reduced by one third with only 18 and 19 products covered in all, respectively. When looking at the MIRTA data for leg muscle it could be observed that the average signal of "expressed" candidates is lower than for brain or heart. This difference to the results for the first set can be explained by the target selection process. The selection of the protein coding targets for the first set was based on their known imprinting status and most genes selected are expressed in many of the examined tissues. In contrast to this the selection of non-coding targets was based on expression in at least one of the examined tissues. This lead to the selection of candidates with highly tissue specific expression. Tissue specific candidates were mostly found for brain, heart and spleen. Furthermore, since the selection was in part based on the MIRTA data, there was just RNA sequencing data for kidney as this tissue was not included in the MIRTA dataset. Taken together these arguments lead to the conclusion that the difference in coverage between brain/heart and kidney/leg muscle can be explained by tissue specific expression.

3.3.3 Reactions with intronic primers yielded more product than their exonic counterparts

Results for non-coding candidates for which both exonic and intronic primers have been designed showed a trend that reactions with intronic primers tended to work more often than the junction spanning primers for the same candidate (see figure 22). An example for this is given in figure 42. This candidate on chromosome 7 showed a drastic reduction in the number of samples the exonic primer pairs worked in, 7, compared to the intronic primers which worked in 31 samples. The candidate exon model overlapped the UCSC annotation of a long *D7Ertd715e* isoform. The exonic PCR product is highlighted by a green arrow and spans across six exons. The intronic PCR product is located within the sixth intron of the transcript as indicated by the orange arrow. Expression data is given for both adult brain and heart in the form of MIRTA data and RNA sequencing data. Comparison of the expression data with the location of the exons showed that while brain showed signals corresponding to the exons covered by the PCR product in both MIRTA and RNA sequencing data (green dotted lines), heart shows a negative signal in the MIRTA data and no coverage in the RNA sequencing data. The situation in heart could also be observed for all other tissues examined (data not shown), suggesting brain specific expression of the long spliced isoform. In contrast the location of the intronic PCR product showed a MIRTA signal in both brain and heart (orange dotted line), as well as most other tissues. This explained why the exonic primer pair only showed a visible product in just the four brain samples, two heart samples and one lung sample, while the intronic primer pair worked in almost all samples. Results from the coverage analysis also confirmed this trend as the intronic product is covered in all 16 samples while the junction spanning product shows far less coverage in heart, kidney and leg muscle compared to brain and incomplete coverage of the product in six replicates.

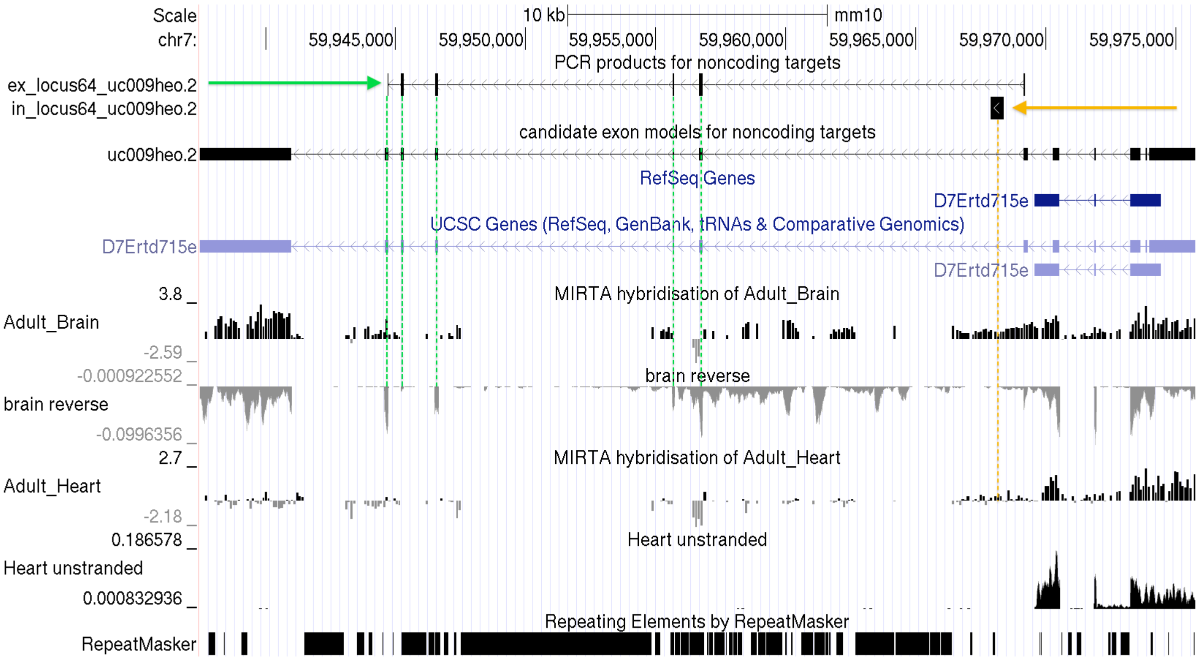


Figure 42: Example of a tissue specific splice variant. This UCSC genome browser screenshot (described in more detail in figure 1) shows the following tracks (from top to bottom): the alignment of both the junction spanning PCR product (green arrow) and the intronic PCR product (orange arrow) for this target, the exon model used to design these products, RefSeq and UCSC annotations of *D7Ertd715e* , MIRTA data from adult brain, RNA sequencing data from the reverse strand in adult brain, MIRTA data from adult heart, RNA sequencing data from both strands in adult heart and repeating elements. MIRTA and RNA sequencing data were used to estimate whether the candidate is expressed in these tissues. The candidate exon model in this case is equal to an UCSC annotated long isoform of *D7Ertd715e*. The positions of the exons covered by the PCR product only show a MIRTA and RNA sequencing signal in adult brain (green dotted lines) but not in adult heart. On the other hand, the location of the intronic primer is supported by MIRTA data in both tissues. Large MIRTA regions without any signal can be explained by their overlapping repeat regions (bottom track).

3.3.4 The majority of *de novo* assembled splice junctions could be validated

The junction spanning PCR products were used to validate the splice junctions assembled by Cufflinks. This was done by analysing the coverage of the exons flanking the junction. The coverage was investigated per tissue and correlated to expression data from the MIRTAs. Overall the same trend of tissue specific coverage differences between brain/heart and kidney/leg muscle could be observed. 52 and 47 junctions were covered in all four replicates of brain and heart, respectively, compared to 31 and 30 in kidney and leg muscle. Separating expressed and not expressed candidates based on MIRTA data showed that junctions in the expressed bin were on average covered in more replicates. Combining the results of all four tissues showed that a high number of junctions, 73/83, could be validated in at least one sample and 35 were confirmed in 12-16 samples. Special focus of this coverage analysis was on *Kcnq1ot1*, since expression data suggested that the last exon of the assembled spliced isoform is only used in brain (see figure 18). The junction coverage analysis showed that the junction using this exon could only be confirmed in brain and heart, with the exception of one kidney sample. On closer inspection it could be observed that the coverage of this exon in heart is much lower than for the other exons indicating that the exon is used rarely. This is also illustrated in figure 43. The exon marked by the arrow showed high coverage in brain, which was comparable to the coverage of the other exons. Overall this data suggested tissue specific expression of the longer isoform.

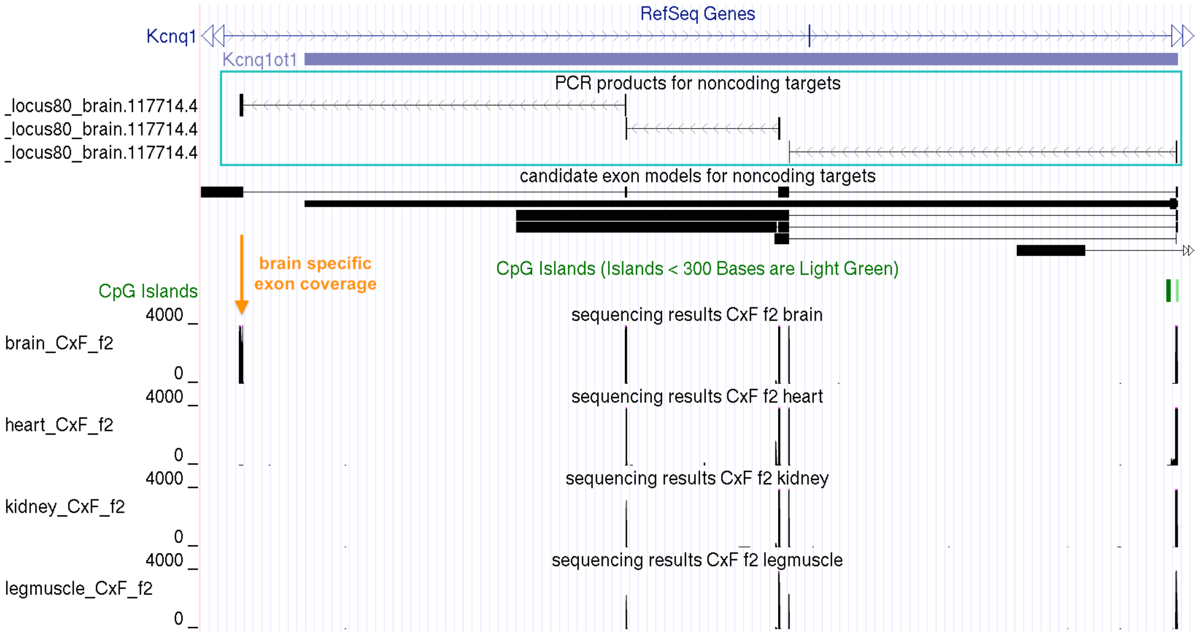

RefSeq Genes
Kcnq1
Kcnq1ot1
PCR products for noncoding targets
_locus80_brain.117714.4
_locus80_brain.117714.4
_locus80_brain.117714.4
candidate exon models for noncoding targets
brain specific exon coverage
CpG Islands
CpG Islands (Islands < 300 Bases are Light Green)
4000
brain_CxF_f2
0
sequencing results CxF f2 brain
4000
heart_CxF_f2
0
sequencing results CxF f2 heart
4000
kidney_CxF_f2
0
sequencing results CxF f2 kidney
4000
legmuscle_CxF_f2
0
sequencing results CxF f2 legmuscle

Figure 43: Brain specific coverage of an exon of a *Kcnq1ot1* splice variant. The tracks displayed in this UCSC genome browser screenshot (see figure 1) are (from top to bottom): RefSeq annotation of *Kcnq1* and *Kcnq1ot1* in this region, the three PCR products (teal box) designed for the spliced isoform of *Kcnq1ot1*, exon models assembled in this region with the topmost being the one used for the design of the primers, a track showing CpG islands and four tracks showing sequencing data for one representative replicate each of brain, heart, kidney and leg muscle. As previously indicated three primer pairs (teal box) were designed for a spliced isoform of *Kcnq1ot1*. The exon indicated by the orange arrow shows high coverage in brain but little to no coverage in other tissues. Axis scales were chosen equal for all four sequencing tracks.

Validation of the 33 junctions covered by the 31 PCR products in testes also showed that the large majority (31/33) could be validated in both replicates (figure 38). Overall the results for these PCR products were highly reproducible between these two replicates as products were either covered in both or in none. The high amount of confirmed junctions, as well as the high reproducibility might be explained by the fact that these candidates were selected exclusively for testes compared to the other candidates which were selected if they were expressed in at least one of the eight tissues included.

3.3.5 Sequencing results revealed additional exons as well as possibly unspliced transcripts

In addition to validating the junctions that were annotated in the *de novo* assembly the data was checked for additional exons which were not part of the assembly. This was done by computationally testing for reads that aligned within the PCR product but not in the annotated regions. Visual inspections of such regions revealed two possible deviations from the assembly. First, a candidate was discovered that showed multiple additional exons. Although these exons were covered weakly in respect to the annotated exons the finding still provided evidence that this locus might harbor an isoform with more exons than annotated by the assembly. This is shown in figure 44.

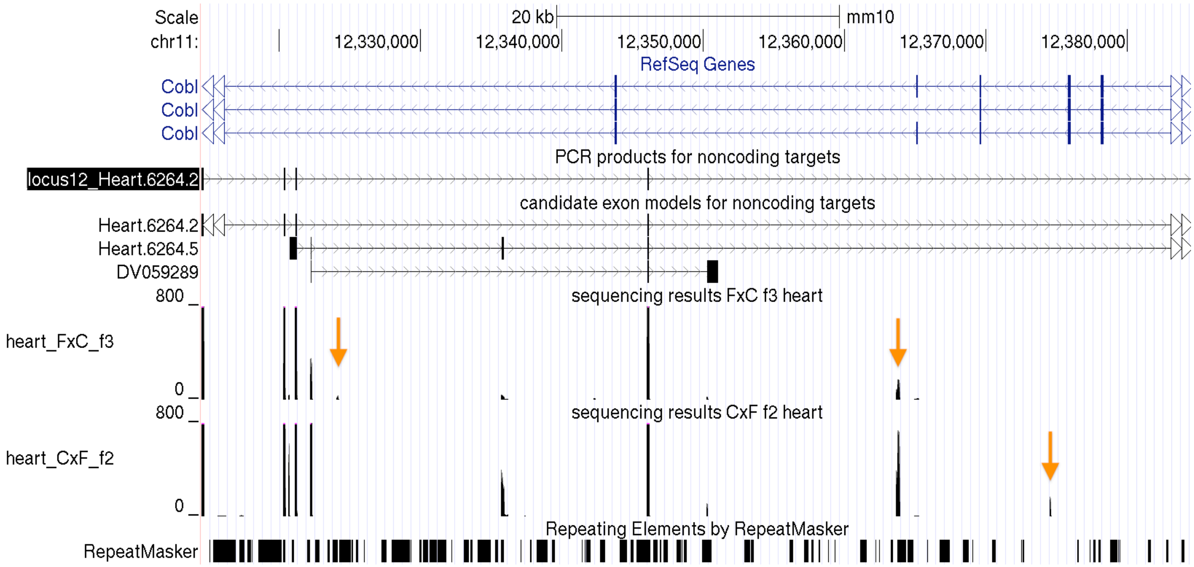


Figure 44: A candidate showing additional priorly not annotated exons. This UCSC genome browser screenshot (see figure 1) shows the following tracks (from top to bottom): RefSeq annotation of *Cobl* isoforms in this region, the annotation of the PCR product designed for the non-coding target, exon models provided by the assembly and sequencing results for two heart replicates where expression of this candidate was highest. A track depicting repeat regions is shown at the bottom. It should be emphasized that the exon models for the non coding candidate are all in antisense orientation to the protein coding gene *Cobl*. Orange arrows indicate reads aligning to regions outside exons annotated by either an exon model or the RefSeq annotation of *Cobl*. Overlap with repeat regions was also checked and only partial overlap was found suggesting that these reads aligned uniquely to these positions.

The second phenomenon discovered by this part of the analysis was that some transcripts appeared to be only partly spliced, at least at the junctions covered by the PCR product. The first example for this was a transcript consisting of two large exons with a small intron between, as shown in figure 45. The coverage pattern only partly supported the spliced variant of this candidate since the high coverage at both ends was around 150 bp long and overlapping the exon/intron junction. These pileups of high coverage at the ends of a PCR product were a pattern also found in pilot sequencing experiments with 50 bp reads, where they were 50 bp in size (data not shown).

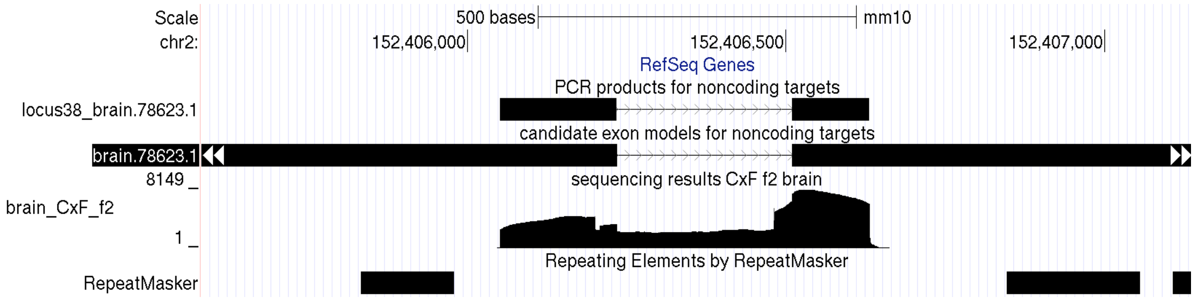

Scale
500 bases
mm10
chr2:
152,406,000
152,406,500
152,407,000
RefSeq Genes
PCR products for noncoding targets
locus38_brain.78623.1
candidate exon models for noncoding targets
brain.78623.1
8149 _
sequencing results CxF f2 brain
brain_CxF_f2
1 _
Repeating Elements by RepeatMasker
RepeatMasker

Figure 45: A possible novel single exon transcript priorly annotated as spliced. This figure shows an UCSC genome browser screenshot which shows the following tracks (from top to bottom): the RefSeq genes track showing that there is no RefSeq annotated gene in this region, the annotation of the PCR product for this candidate, the corresponding exon model showing two large exons and a short intron and the sequencing track for one brain replicate. Sequencing data showed a pattern which would also support an unspliced variant since the high coverage regions at the ends of the PCR product do not perfectly align to the annotated exons (see text).

The pattern resulted from the fact that these regions could be sequenced in unsonicated PCR products as well and were therefore overrepresented compared to the middle part of the PCR product which needed sonication to be accessible for sequencing. The inclusion of the intron in the PCR product was also visible on the gel as a larger secondary product (not shown as it is not visible in the gels for heart). Overall these results suggested that the annotated transcript is at least partly unspliced.

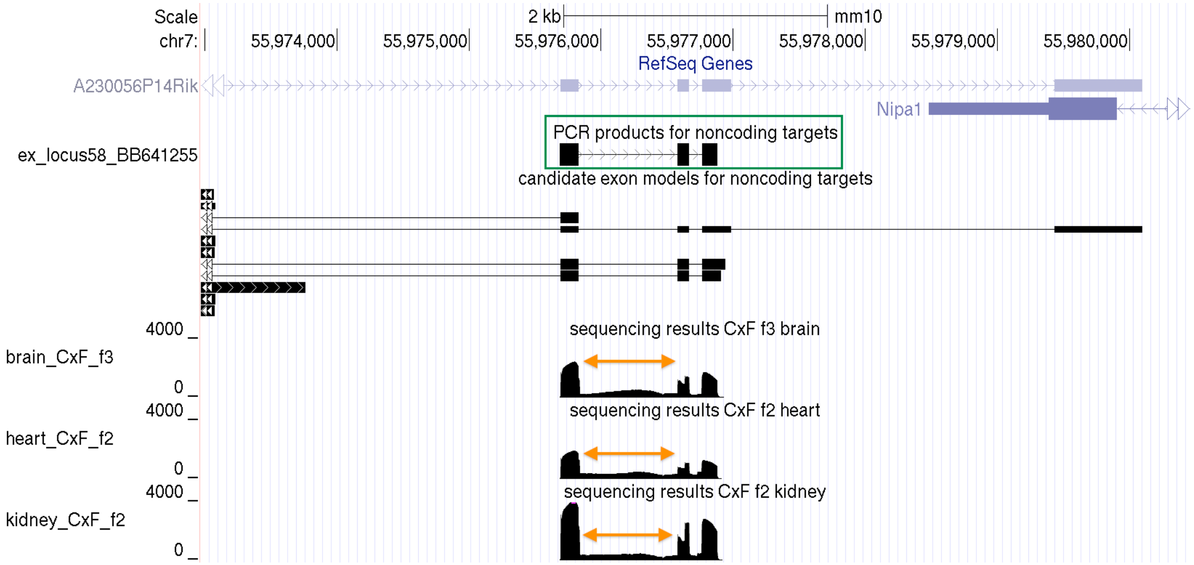


Figure 46: Continuous coverage across two junctions suggest an inefficiently spliced transcript. The tracks shown in this UCSC genome browser screenshot (see figure 1) are (from top to bottom): RefSeq annotation of genes in the region, the PCR product for the candidate (green box), exon models assembled in this region and three tracks showing sequencing results for one replicate each of brain, heart and kidney. The exons of the assembled exon model used for the primer design overlapped the exons of *A230046P14Rik*. Orange arrows indicate coverage of the larger of the two introns indicative of inefficient splicing.

Another example for inefficient splicing is shown in figure 46. In this case the exon structure is still clearly visible in the coverage pattern but coverage between the exons (orang arrows) suggested the presence of an unspliced variant as well. Since the inclusion of the introns increased the size of the PCR product to around 1100 bp this difference in coverage could also have been caused by a bias introduced by the PCR and subsequent size exclusion steps during the purification. Again the large PCR product was visible on the agarose gel images as a secondary product (see gels in figure 19, lane 21). Since the investigated exon model is identical to *A230056P14Rik* this would make an interesting candidate to make a detailed analysis of splicing efficiency.

3.4 Future perspectives

The results presented in this thesis showed that the dataset is suitable for further downstream analyses in terms of coverage and reproducibility across replicates. Since the main focus of the overall project is genomic imprinting the dataset was checked if it could be used for this purpose. The pipeline that will be used for the analysis of imprinted expression is centred around SNP variant quantification and requires good coverage of the SNP positions to increase the statistical power. Of the 727 SNPs covered by all PCR products in both sets at least 50 and 63% were covered by more than 1000 reads in kidney/leg muscle and brain/heart, respectively. SNPs with low coverage below 10 reads only made up a small portion with 5-7% in brain/heart and 11-16% in kidney/leg muscle belonging to that category. Differences between these two tissue pairs can be explained by tissue specific coverage differences as mentioned above. In summary these results show that this dataset is suited for the analysis of imprinted expression based on SNP variant quantification. For a more comprehensive view on genomic imprinting in adult tissues PCR products from samples of the remaining tissues will also be sequenced and samples were already submitted to the facility at the time this thesis was written. The remaining sets also contain the two control samples with pure Cast and FVB samples which should aid in the correct identification of imprinted genes by enabling to correct for strain biases introduced by PCR.

This dataset was also shown to be a good tool to validate *de novo* assembled splice variants. Especially candidates where the sequencing data showed differences to the assembled annotation could be further investigated regarding their true exon structure or their splicing efficiency. Since no spliced variant of *Kcnq1ot1* has been annotated yet further investigations of this transcript could lead to a characterisation of this splice variant and shed more light on the nature of the brain specific usage of the last exon.

In summary the data presented in this thesis showed that the employed targeted PCR approach produced a dataset which can be be easily used for an analysis of imprinted expression of the selected candidates. Furthermore the data showed that this approach is feasible and provides a less expensive method for the analysis of imprinted expression than whole transcriptome sequencing.

4 Methods

Non-coding pipeline

This pipeline was based on the one used by Cabili et al. [103] and used publicly available multiple alignment data (reference: mouse) between 60 species from UCSC[6] in the form of .maf files. The species used in the alignment are listed at the UCSC genomewiki[7]. Before the .maf files could be used they had to be indexed using the `maf_build_index.py` script, which was part of the bx-python project[8]. Based on the coordinates of the transcript the sequence-data of the corresponding multiple alignment was extracted using the `interval_maf_to_merged_fasta.py` provided by Galaxy tools (local instance[9]) [138, 139, 140] with standard parameters for the use of a BED file as input. Additionally a list of species was given for the `-p` parameter to only extract data for the species present in the phylogenetic data of PhyloCSF, since otherwise this would have produced an error.

The extracted multiple alignment data, now in Fasta format, was then subjected to analysis by PhyloCSF[10], a method which used phylogenetic codon models to distinguish between protein-coding and non-coding sequences [115]. Before using PhyloCSF with the UCSC .maf files the identifiers in the phylogenetic tree file (`.nh`) had to be changed to comply with the identifiers in the maf files (for example mm10 instead of Mouse). The options used besides the required phylogenetic tree and input fasta file were `-frames=6 -removeRefGaps`, the former telling PhyloCSF to look in all 6 possible reading frames and report the highest score while the latter is set to handle gaps in the references sequence which would otherwise cause an error. I discovered that PhyloCSF results are influenced by UTRs which are relatively long compared to the rest of the transcript. This lead to an underestimation of the coding potential. To tackle this issue the PhyloCSF analysis was done for each exon of the transcript separately and the maximum score of a single exon is reported.

In addition to this the pipeline then searched for an open reading frame of at least 300 nt length (equal to 100 amino acids) using the EMBOSS getorf tool which was then translated using the transeq tool from EMBOSS with the option `-frame 6` to translate all possible reading frames. If there was no start codon present, which might be the case for incompletely assembled transcripts, the region from start until the

---

[6] http://hgdownload.cse.ucsc.edu/goldenPath/mm10/multiz60way/maf/

[7] http://genomewiki.ucsc.edu/index.php/Mm10_conservation_alignment

[8] https://bitbucket.org/james_taylor/bx-python/wiki/Home

[9] https://wiki.galaxyproject.org/Admin/GetGalaxy

[10] https://github.com/mlin/PhyloCSF/wiki

first stop codon was extracted by the `fasta_formatter` tool, part of the fastx-toolkit[11]. The translated amino acid sequences were then used by the phmmer tool of a local instance of HMMER[12] [120] which queried the sequences against a local copy of the UniProt UniProtKB/Swiss-Prot database (obtained from the uniprot site (release 2012_01) [13]) using a Hidden Markov Model created from the query sequence. The non-standard parameter used for this tool was `-E 0.01` which specified the limit for the E-value below which hits were reported as significant. The results were then summarized reporting the phyloCSF score together with a boolean value indicating whether the phmmer query reported a significant hit or not.

Automated primer design

Primer design was done by designing a custom script which employed Primer3 2.3.4 [141, 142] to design primers for transcripts defined in a BED file. For each line of the BED file the script extracted the cDNA sequence of the transcript from whole chromesome fasta files using bedtools. Positions of SNPs within the transcript were determined by scanning through a list of annotated SNPs obtained from the Sanger institute[14]. SNPs then got grouped together if the distance between them is beneath a threshold (default: 200 bases). Grouped SNP regions then formed the targets for the primer design using Primer3. Single SNP positions were used as excluded regions to prevent primers from overlapping SNPs. Primer design was implemented so that the script first tried to design primers in the first 5th of the transcript, starting at the 5' end, then the first two fifths and so on. Large SNP groups were prioritised by first using only the groups including the most SNPs as targets then the groups including the second most SNPs and so on. In summary primer design prioritised primers designed near the 5' end and including as many SNPs as possible. This was done because some candidates had isoforms with different 5' ends which only had isoform-exclusive SNPs, i.e. SNPs in a region not overlapped by another transcript, near the 5' end. A similar script was implemented to generate junction spanning primers for candidates defined in a BED file.

---

[11] http://hannonlab.cshl.edu/fastx_toolkit/download.html

[12] http://hmmer.janelia.org/software

[13] ftp://ftp.uniprot.org/pub/databases/uniprot/previous_major_releases/release-2012_01/

[14] ftp://ftp-mouse.sanger.ac.uk/REL-1303-SNPs_Indels-GRCm38/

Selection of known imprinted protein coding genes

The initial list for this selection process was curated by Prof. Denise Barlow based on published data from both the MRC Harwell Imprinting Web pages [31] as well as a list published in 2012 by Okae et al. [24]. Only candidates with an available RefSeq annotation (BED file) were used. For this analysis windows of +/- 1 kb around the annotated transcription start site (TSS) were used to assign transcripts to loci, i.e. if the TSS-windows of two transcripts overlapped they were assigned to the same locus. The whole list of transcripts with RefSeq annotation, including multiple transcripts per locus, was used as input for the previously validated primer design script. In general a maximum of one primer pair per locus was selected for the following PCR analysis.

Selection of non-coding candidates for primer design

The selection was based on an annotation created by Florian Pauler. Transcripts were assembled with Cufflinks from RNA sequencing data aligned with STAR [143]. These transcripts were then filtered to select only those located within imprinted regions specified by the MIRTA data. The remaining transcripts were then assessed by the non-coding pipeline and transcripts with coding potential were filtered out. In addition to this a script was used to select for transcript overlapping regions with a continuous MIRTA signal (Florian Pauler, unpublished data). Based on this provided annotation I manually selected candidates if they showed expression in at least one of the tissues of interest. Expression was derived from both RNA sequencing data as well as the MIRTA data. An additional criterion used to aid in the selection process was the presence of a H3K4me3 peak near the assembled transcription start site, which was used when deciding between different transcripts at a single locus. If the transcript did not overlap any other genes an additional "intronic" primer pair was designed using the genomic DNA of the whole region as a template for primer design. Most of these primers were located within introns and some were overlapping an intron/exon junction. Intronic primers were also used for regions without an exon model or regions with an excessive amount of exon models. In addition, candidates with expression in testes and/or placenta but none of the eight tissues checked before were selected and junction spanning primer pairs were designed using the script described above. Coverage of SNPs was of no concern here since testes was not a tissue to be included in the analysis of imprinted expression. The SNP annotation was only used to check if the primer themselves did not overlap any SNP so that the primers would work on both alleles.

Checking for overlap with insertions/deletions

Overlap of primers with insertions/deletions in either one of the two strains was analysed using the intersectBed tool provided by bedtools [144]. For this primer sequences were aligned via bowtie using the parameters -a -f -v0 –un. An annotation of insertions/deletions was acquired from the Sanger institute (same source as the SNPs, see footnote). Insertions/deletions data of the strains was extracted using a custom script and overlaps were determined using intersectBed.

Harvesting of samples

Samples were harvested from F1 hybrid mice between strains Cast/EiJ and FVB/NJ, henceforth termed CxF and FxC, the first letter indicating the mother, the latter the father strain. Two replicates were prepared from each cross termed FxC f4/f3 and CxF f2/f3. No animal experiments according to the Austrian Laboratory Animal Act were performed in this study, because humane killing of laboratory animals is not defined as animal experimentation under the Austrian Laboratory Animal Act (Animal Experiments Act, Federal Law Gazette No. 501/1989). For this reason approval of the study by an institutional ethics committee was not required. The mice were sacrificed at ages 9 weeks and 11 weeks for crosses CxF and FxC, respectively, and brain, heart, kidney, leg muscle, liver, lung, spleen and thymus were harvested by Philipp Günzl and Quanah Hudson. The samples were immediately frozen after dissection by using liquid nitrogen and stored at -80 °C until further use. Mice were bred and housed at the Forschungsinstitut für Molekulare Pathologie GmbH, Dr. Bohr-Gasse 7, 1030 Vienna, Austria in strict accordance with national recommendations described in the "IMP/IMBA Common Institutional policy concerning the care and use of live animals" with the permission of the national authorities (Laboratory Animal Facility Permit MA58-0375/2007/4).

RNA isolation

Organ samples were first homogenised in TRIzol® Reagent using the Polytron® PT 2100 homogeniser. The homogeniser was prepared by soaking the parts in freshly prepared DEPC water for at least 30 min. 4 ml of TRIzol® Reagent were used for the brain samples and 3 ml for the other organ samples. Homogenisation was done in 14 ml Falcon™ tubes until the organ was completely homogenised, followed by washing the homogeniser 3 times with autoclaved DEPC water after each sample.

The homogenised samples were either stored on ice and isolated immediately afterwards or stored at -80 °C until further processing. Samples were divided into 1 ml aliquots in 1.5 ml RNAse-free tubes. RNA was then isolated according to the standard TRIzol® Reagent protocol but substituting 200 µl chloroform with 100 µl BCP. RNA was resuspended in 20-100 µl RNA Storage Solution corresponding to pellet size. Concentration of isolated RNA was assessed by NanoDrop® 1000.

cDNA preparation

Isolated RNA was first DNase I treated by applying the DNA-*free*™ Kit (Ambion®) according to protocol using 1 µl of rDNAse I and 5 µl of 10X DNase I Buffer in a total volume of 50 µl per 10 µg RNA and digesting for 30 min. DNAse treated RNA was precipitated by adding 2.5 vol. EtOH (96%) and 0.1 vol. sodium acetate (3 M) at -20 °C overnight. cDNA was prepared using the RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific™) according to protocol. The minus RT control was prepared by dividing the reaction mix before adding the reverse transcriptase. Afterwards the samples were diluted using 100 µl ultrapure H_2O per 20 µl cDNA reaction.

PCR of Igf2r/Airn/Impact for Sanger sequencing

PCR was performed using a reaction-mix containing 0.8 M beatine, 2.5 mM $MgCl_2$, 0.2 mM dNTP mix, 0.5 µM of both forward and reverse primer and 0.02 U/µl goTaq® DNA polymerase for 1 µl cDNA template in 25 µl total reaction volume. PCR program used: 94 °C for 3 min; 35 cycles of 96 °C for 10 sec, 94 °C for 30 sec, 58 °C for 30 sec and 72 °C for 30 sec; 72 °C for 5 min - keep at 4 °C. The PCR was checked by 2 % agarose gel electrophoresis (EtBr staining), loading the whole reactions. Reversed images were captured using the Biometra UVsolo TS Imaging System. Purification of the PCR products was done by cutting the corresponding bands from the gel and purifying them using the Promega Wizard® SV Gel and PCR Clean-Up System by centrifugation according to the manufacturer's protocol or by sending the unpurified PCR reactions to the Microsynth company for purification prior to sequencing. Sanger sequencing was done by Microsynth using their Barcode Economy Run Service. Samples were prepared according to Microsynth's specifications and the sequencing results were analysed using Sequencher 4.7 (Gene Codes Corporation). Quantification of the ratio between the two alleles was done using the Softgenetics® Mutation Surveyor® software and visualized using R.

Preparation of PCR products for libraries

The same reaction mix and PCR settings were used as for the PCR described above. Afterwards 5 µl of the PCR reactions were analysed using 2 % agarose gel electrophoresis as before. The remaining 20 µl reactions were pooled together per tissue sample, yielding 8 tissues x 4 replicates = 32 pools, precipitated with 2.5 vol. EtOH (96 %) and 0.1 vol. sodium acetate (3 M) at -20 °C overnight and resuspended in H_2O. Afterwards the PCR products were purified using Agencourt® AMPure® XP magnetic beads. To attain the right size selection the purification was done in two steps. In the first step 0.5 vol beads were added to the pooled PCR products and the beads were discarded afterwards. In the second step 1.0 vol of beads were added to the supernatant of the first purification and after discarding the supernatant the purified products were eluted from the beads using 100 µl H_2O. Purified PCR products were stored at -20 °C until further use. Purified PCR product samples were sonicated using a Covaris S2X sonicator using Snap-Cap microtubes. Settings used were Duty cycle 10%, Intensity 5.0 and Cycles/burst 200 for a volume of 50 µl at temperatures around 5-7 °C. The results of the sonication were analysed using the Bio-Rad Experion™ DNA 1K Analysis Kit according to the manufacturer's protocol but priming twice, premixing samples with loading dye before loading and omitting the last vortexing step.

Library preparation and Next generation sequencing

Sonicated samples were diluted (4 µl sample + 26 µl H_2O) and the concentration was measured by Qubit® fluorometer according to the manufacturer's protocol. Library preparation was done for 5 µg sample in 25 µl total volume using the Illumina® TruSeq® ChIP Sample Preparation Kit according to protocol but omitting the purification step by agarose gel electrophoresis and using half volumes for everything except the last elution step as well as the washing steps. The success of the library preparation was assessed by measuring the concentration with a Qubit fluorometer and checking the size distribution using the Bio-Rad Experion™ DNA 1K Analysis Kit as before. Massive parallel sequencing was done by the CSF at the Vienna Biocenter using an Illumina HiSeq 2500 with the MiSeq chemistry and 150 bp single end sequencing multiplexing nine samples per lane.

Bioinformatic analysis of sequencing results

Sequencing results were aligned using the STAR aligner [143] and further processed using samtools [145]. For analysing the achieved coverage of the PCR products a BED file of the PCR products was created by first employing a custom script implementing in silico PCR on custom cDNA sequences to get the sequence of the PCR products and aligning these products using STAR. Correctness of the alignment was verified by visual inspection of some aligned PCR products using the UCSC Genome Browser. Coverage of PCR products and SNPs in particular was evaluated using the coverageBed tool, part of the bedtools toolset [144], and subsequent analysis of the data using a custom R script.

5 Materials

Chemicals and consumables

	Product	Supplier
Chemicals/Enzymes	AMPure XP Beads	Biozym
	Bromo-chloropropane (BCP)	MRC
	Diethylpyrocarbonate	Sigma
	Ethanol 96 %	Merck
	Ethidium bromide (EtBr)	Applichem
	GeneRuler 100 bp Plus DNA Ladder	Thermo Scientific
	GoTaq® Flexi DNA Polymerase	Promega
	Isopropanol	Merck
	PCR nucleotide mix	Promega
	RNA storage solution	Ambion
	Sodium acetate	Applichem
	TRIzol Reagent	Sigma
	Tween 20	Sigma
Kits	DNA-free Kit	Ambion
	Experion DNA 1K Analysis Kit	Bio-Rad
	Qubit dsDNA HS Assay Kit	Invitrogen
	RevertAid First Strand cDNA Synthesis Kit	Thermo Scientific
	TruSeq ChIP Sample Preparation Kit	Illumina
	Wizard SV Gel and PCR Clean-Up Kit	Promega

Primers

Preliminary study for *Airn* and *Igf2r*

Target	Forward primer	Reverse primer
Airn	TCAATGTTAGCAACTTTGGGGG	CAGTCCAAGGTCACCGTAACA
Igf2r	GGCACCTCTGACATGACCAA	ACTCCGCTCTGAGAGTCCTT
Igf2r genomic	TGCTTGCTGTCTCCTTTCCT	GGACATGGGCATCACACTCA

Table 11: Primers for the preliminary study on Airn and Igf2r.

Known imprinted targets - set 1

Target	Forward primer	Reverse primer
001.Zdbf2.NM_001267872	AGGCAGGTACAGCAGGAAAC	GCTTGGGAAGGACCAGGTAC
002.Adam23.NM_001177600	ACAAAGGCCAGACACCAACA	ACTGGTCACTGCCATCTGTG
003.Gpr1.NM_146250	TCCTGTCTGTGGTCATTGCC	ACAGGCTCTTGGTTTCAGCA
004.Plagl1.NM_009538	GCTGGACCACCTCAAGTCTC	AAGACTGCATCGGCTCCAAA
005.Dcn.NM_001190451	AGGGCTCCTGTGGCAAATAC	TCAGGCAGTTCCTTTAGTTGGT
006.Dcn.NM_007833	GGGGTAAACACAGAAAGCCC	TCAGGGGATTGTCAGGGTCA
007.Phactr2.NM_001033257	CCTTCTTCTGCCCCCTGAAG	AGTCAGAGTCGTTGGCTTGG
009.Phactr2_intron.NM_001195065	AGAAGCAGCTCATGGACACC	GGCAGAGAAATCTGGGAAGGA
010.Zrsr1.NM_011663	GATACCTGGATGACTGCGCA	TTATCCGTGGTATGGCCTGC
011.Mapt.NM_010838	AGCCCTAAGACTCCTCCAGG	TGTTCCCTAACGAGCCACAC
012.Ccdc40_thick.NM_175430	CCCAGCGCATCCCTATCAAA	CTCCTCTGCTTCATCACTCCC
013.Ddc.NM_016672	CACGGCTAGCTCATACCCAG	CTGCCTTGTCCCGCTCCAG
014.Ddc_intron.NM_001190448	CACTAGTTGCTGAAAAGGCACC	TCCAGGCAAGGGTCCTTCTA
015.Grb10.NM_010345	AGATGGGACCAGCAAAGTGG	AATAGAGGCCAGATCTGCGC
016.Grb10_intron.NM_010345	GCAGCTCTGTGTCTCCAGTT	ATGGACAATTGCCCCCAGAG
017.Cobl.NM_172496	AGATGCCATCTCCCTGGACT	AGCCCTGAGTCAAAAGGCTC
019.Dlk1.NM_010052	GTGCAACCCTGGCTTTCTTC	CAAGTTCCATTGTTGGCGCA
020.Dio3.NM_172119	TCGAGACCAAAGGAAGTGGC	AGAGCAACTTCCTTCAAGTCTT
021.Scin.NM_001146196	CTCAGGGCACGGATCAAGTC	CCCTTCACGTGCAGAAGTCT
022.Wars.NM_001164314	ATCGGCCATCCTAAACCTGC	TACAGACAGGCTTGCCACAG
023.Begain.NM_001163175	CACAAGAGGGGAGAAGCCAG	AAGACAGGTTCACAGGGGTG
024.Rtl1.NM_184109	TCCCTCAGCCGAATTCCCTA	CTGGGCAAACCTCTCATCCA
025.Cmah_intron.NM_001111110	CCCTGCCCCTTGCCATTTAT	CAAAACATGCAGAAGACCAGGA
026.Cmah.NM_007717	AAGCTCAGCTGGTGAAGGAC	GGCTGGGTCTTTCTTTCGGA
027.Pde4d.NM_011056	AAAGGATGCTGAACCGGGAG	CGTCTGCAGCATGGATGTTG
028.Drd1a.NM_010076	GAGCGGCACAGGAGAGGG	GGCTTAGCCCTCACGTTCTT
029.Htr2a.NM_172812	GCAACCAGGAGGGGCTTATT	ACAGCAGCCGAGGAACTTAC
031.Trappc9.NM_029640	GCTGGACTTCCTGTCTGACC	GCTCTTGGTGGACATGCTCT
032.Trappc9.NM_180662	GTCCCCGACTACATGCAGTG	TCTCCTTCTGCACGTGGAAC
033.Slc38a4.NM_027052	GAGCTCGGGTGTGAACTTCA	ATGTTGGACACCGTCTGCAT
035.Pde10a.NM_011866	CGATTCAAGGCTTGCTGCTC	TGAGGAAACACTCGGTCAGC

036.Slc22a2.NM_013667	CGGAAGTTCTGCCTCTTGGT	CCAAGTCCAGGAACGAAGGG
037.Qpct.NM_027455	ATCTGAGTCTGGTCTCCGCT	GACCCACTCAGCCTGAAGTC
038.Slc22a3.NM_011395	CCTTTAGGGCAGGCTACAGG	CCAATAGTAGTCGGGCTGGC
039.Igf2r.NM_010515	CCCTTGGCCCAATATGGAGG	GAGTGACGAGCCAACACAGA
040.Impact.NM_008378	CTATTCTAGAGCACCCGCCG	CTGCTCAGGACAGACGACAG
041.Tbc1d12.NM_145952	AACGGAGACTCGGGCTTTTT	TGGGAAAAGGTTTCTGGCGA
042.Ins1_thick.NM_008386	CTTCCTACCCCTGCTGGC	ACACACCAGGTAGAGAGCCT
043.Sfmbt2.NM_001198808	TCATGAGCTTTGCTTGCTGC	TCCTCCTGGGAGTTCCACTC
044.Wt1.NM_144783	TTCACCTTGCACTTCTCGGG	CCCAGCAGCCATTCCCTTTA
045.H13.NM_001159552	TTAGGGGAACGTGGCTTTCC	AGGTTGATGTACTCCTGGGAGA
046.Mcts2.NM_025543	TCAGGCAGAGAAAAGGACCT	TGTCACATACAGACACACACAGA
048.Gnas.NM_019690	AACCAGTCACTCACTCAGCG	GGCCTCCTGGTCTTGCAG
049.Gnas.NM_010309	GCAGTGAGATCAGTGGACCC	CGGAAGCCAGCGTTTTCAAA
051.Gatm.NM_025961	GACCTGGTCTTGTGCTCTCC	AGGGCCTTGCTGCTTCTTAG
052.Bcl2l1.NM_009743	CCCCTAAACCAGCTCCTTGG	GTGGACAAGGATCTTGGGGG
053.Blcap.NM_016916	GTTCAGACAAGACCCAGGGG	TCCAGCTCTGTCCTATGCCT
054.Zfp64.NM_009564	AGACCACCACAACGACCATC	GCTTGTTAAGGCTGCTGCTG
055.Phf17_intron.NM_001130184	CGTTTGACAGACTGCAACCA	TCCCATGTGACAATTACAATGC
056.Phf17.NM_172303	ACTTACATGGTGACCCGCAG	CATGGTCTGGGGTCACCAC
057.Htra3.NM_030127	AGGGGTTCCTCTGTGAAGGA	CATGGTGTGGGGACTGACC
058.Casd1.NM_145398	TCATAATGGCAGCGAGGAGG	TCCGGTGAGCATTACGATGA
059.Peg10.NM_130877	GAGATGATTCCTGGAGCGCA	GGCTGGCGGTTCGTATTTTG
060.Ppp1r9a.NM_181595	GCCGCTCGATGAACCTCTAG	GGGAGCCCTCACCTTCTTTC
061.Asb4.NM_023048	AAGCTGAAGTCTTCCTGGGC	TCATGCTATCCACGATGGCC
062.Klhdc10.NM_029742	GCTGTGGTTGCCTAGTGACT	GACTGCACCTATGGACACCC
063.Mest.NM_001252293	AGAGCTGCTGTTGTTTTGTGT	TTTAGGTTCGATGGGCTGGG
064.Mest.NM_001252292	CACATCCCGGTGCTTCTTCT	TTCCATGAGTGCAGAGCAGG
066.Calcr.NM_001042725	CTGGAGCCACAGCCTATCAG	GCATGGAAGCAACCAAAGCA
067.Tfpi2.NM_009364	CCATGACATGTGAGCCCCTT	AAAGCGTGTCACATTGGCAG
068.Sgce.NM_011360	GGAAGAAATGTTGGCCAGCG	CAGCCAGGGTAACGAGGAAA
069.Pon3.NM_173006	ACCACAAAACTGCCACCTGA	TCTTCCTGGTTTGTCCGGTG
070.Pon2.NM_183308	ACGAGCTCCTTCCAAGTGTG	TGTTCTGAATGCGGAGGACC
071.Dlx5.NM_198854	GTAACTCGCCACAGTCACCA	GGTACCAGGAAGCCGAGTTC
072.Copg2.NM_017478	GCACCTGCTGTCTCAGTTCT	TGACAGAGAGCACTGATGGC
073.Klf14.NM_001135093	GAGGATGAGCTCTCTGACGC	GGGGAGTGCGACGACGAC
075.Cntn3.NM_008779	CCAGCACGGTGACAAATACC	TAGTAAGTGAGCCGCCCTCT
076.Usp29.NM_021323	ACGTTGAGGTCTAACCACGA	CCTGTGACTTGCGGTCTCTT
077.Atp10a.NM_009728	GGAGGAAGTGGTGTCCAAGG	GTGGTTGGTCTGGAGAGGTG
078.Ube3a.NM_011668	ACCTGTTGAGCTTGTGCCTT	CACAGTGGACAAGATGCCCA
080.Ampd3_intron.NM_001276301	AGGTACAAAGAGTCAATGCCA	ACAGCACATGGAGGAAATAGGG
081.Ampd3.NM_009667	AGGTGACTCAGGTCCAAGA	AAACACCTTCTCCGCCAACA
082.Tspan32.NM_001128080	TGAGCACCATAGCCACTGTG	CCAGAAGCCCAAAAGGAGAT
083.Cd81.NM_133655	GAGGTCTATAAAGAGTGAGGAGC	GGAGCTTGGATTGGTCCTGG
084.Tssc4_thick3UTR.NM_138631	GAGACCACTTCCGGAACAGG	CCTGACCCACAATTCCCACA
085.Kcnq1.NM_008434	ATCAGGGGTATCCGCTTCCT	GTGGCTGAGTCAGGGTTCTC

086.Slc22a18.NM_008767	ATCCAAGGCCTGGTCATTGG	AGTCACTGGGCTTTGTGGTC
087.Dhcr7.NM_007856	ATGGCTTCGAAATCCCAGCA	AGCTGACCCACAAGGCATAC
088.Zim1.NM_011769	CGGACGGACGGATAACCAAG	AGATCACTGGTTCCTTGTACC
089.Peg3.NM_008817	GAGTCCAGCTTGCCGAAGAT	CAGCACCACACTCAAAAGGC
090.Axl.NM_009465	TCCCGTACTTCCTGGAGGAG	TGGAAACCACGTGGAGATGG
091.Snurf.NM_033174	GGGACACCTATAGGCATGCC	ACCAATGCTTGAAGTGAATGTCA
093.Snrpn.NM_001082962	TGCCTCTCACATCCACCCTA	CGTCGTGGGTACAAGTGACA
094.Mkrn3.NM_011746	CCGAGATTGACAATGCAAGCC	CAGCATAGCGGCAAAGCG
095.Peg12.NM_013788	ACGACAACAGCTTCCTCCTG	CTCCGCTGATGCTGCTCTC
096.Art5.NM_007491	CCACCGGAGAAGAGACCAAG	ACTGGAGCTGAAGTGCCG
100.Th.NM_009377	AAGGGCCTCTATGCTACCCA	GCAAGTCCAATGTCCTGGGA
101.Ascl2.NM_008554	AAGTGCTGACTGACCTCTGC	CTTTGCAACAGCAGGGTTCC
102.Cdkn1c.NM_009876	CAGGATGTGCCTCTTCGAGG	CTCAGAGACCGGCTCAGTTC
102.Cdkn1c.NM_009876_second	CAGCGGACGATGGAAGAACT	TGCACTGAGAGCGAGTAGAG
104.Nap1l4.NM_008672	CCGAGTTCACCTTAGCCTCTG	CTGAACTGCAGCAGTCTCCA
105.Tnfrsf23.NM_024290	AGCCATGGTTACCTTCAGCC	CAGGGATTCCTTGGGGACAC
106.Osbpl5.NM_024289	GGACGAAGCTGTGGTGTGTA	ATGCATCGTTCTCCAAGGCA
107.Rasgrf1.NM_011245	GTAGATCTTGAAGGGGGGCGG	CCTGAGAGACTGCTACGCAC
108.Mst1r.NM_009074	CTCACCCTTGAAGGCCAGAG	TCAGATTCCCTGTTGCCCAC
109.Airn_nooverlap.NR_027772.1	TTGTCCCTTGCCTTCAGAGA	TCTGCTTTCTGTCTGTTTCCCA

Table 12: Primers designed for known imprinted targets for eight tissues. The names consist of the locus number, the name of the target and the RefSeq ID of the isoform the primers were designed for.

Non-coding targets for eight tissues - set 2

Target	**Forward primer**	**Reverse primer**
ex_locus3_spleen.9702.4	GCTGAAGTGGCTGATCGAGT	AAGAGCCGTGATTGGTGCAT
ex_locus5_Heart.6056.2	AGCACAGGGGAGTTTGTTGA	TCCATGTGTACTCCTTTGGGC
ex_locus6_brain.15210.1	GGCTGTTACAGTGTGAGGCT	GATTCTTTTCTTCATTTGACTTTTCCA
ex_locus7_BB612635	ATGCTTGGGCTGTGTTCTGA	CATGAACCCCTCCCTTCCAC
ex_locus9_spleen.11110.9	CGGGGAACTTAAATCCTCCCA	GCATGTAAGGGCTTAGGTTGC
ex_locus12_Heart.6264.2	TCCAGAGTGACAGGAGAGCT	TCCACTTCCTTGTTATTAAGCCT
ex_locus17_uc007paz.2	ATGGAGACTAGCCTGCTGGT	TTGGCACGTCAGGATGAGTC
ex_locus18_BQ830741	TGTCACGGTCAGCTCTGTTC	TGCTTACCCGAAGAAAACAGAAT
ex_locus19_uc011ytv.1	CCAGCTGTTTGGGATTCTGG	AAGGGCACAGGCATTACCAA
ex_locus32_Heart.14437.2	TGGAGTGTTTGCCATGTCCA	ATTGCCTCCATTCCTGGTCG
ex_locus38_brain.78623.1_trunc2	GAGACCGAAGAGCACTCAGG	CATAACCCTGGCGCTAGGAG
ex_locus40_uc008lkn.1	GCTGCCATCTCAGAATCCCA	CCTTCTCCCTCTTTGGTGACA
ex_locus42_testes.118994.1	GCTTCAGCCCAAGGAGGAAT	GCTCAGAGGTGTGTGGTCTC
ex_locus42_testes.118994.6	CGTAGGAGCTCCAGCATCAG	TTCGCCAGGCATCCTTGATT
ex_locus45_testes.138984.2_trunc	ACTGGCAGGAATCGGAACAG	GCGGTTAGTTCATCCAGCCT
ex_locus46_brain.94169.2	TGCGAAGGCCGGCATCTT	GAGTCGCCCCTGAATCACTA
ex_locus48_Heart.32260.10	AGCCCTGAATGTGAGCAGTC	CACCTGGAGGATCTGGGGTA

ex_locus51_ESC.36014.1	CCTGCTGCCTACGGATCAAT	CTGTGTCAGAAGTGGCCAGT
ex_locus53_lung.80593.1_trunc	GGCGGAGAGACAGATCACAG	TGACCTTATAAAAAGAAAGGACACAGT
ex_locus56_DV071898	TTGCCTCACTTTCGGGAGTC	GAGAACTGCACGTGAGCTCT
ex_locus58_BB641255_trunc	TGTGGAGGCCGTCTGTTTTT	TCACGTGTAAAGCGAGATTGT
ex_locus63_ipw	GCTGTTATTGCCTTGCCTGA	CTTCCTTTCCAAAATTGCTTCAGAG
ex_locus64_uc009heo.2	GAGACTAGATTGCTTTAAATTGAGTCT	TCCTGGATCAGAGAAAAACCACA
ex_locus66_brain.116263.197	ACAACAGGAGCAGCAACAGA	GGTTTCCCAAGCAATCCTCC
ex_locus66_uc009hey.1	AAAACGTCACCATGGCCAGA	TCCTCCTGTTATTTCTCCTCCTAC
ex_locus67_brain.112173.11	TGCAGATGAAAAGGCAAATAGAAAG	GGAAGCATGATCCCAGAAGCT
ex_locus68_brain.112299.6	GCAACATGGCTACTTCTGCG	TCTCCCTCCTGACCAGATTGT
ex_locus74_BM934118	CTCGCTGAAGGCTTCTGGTT	CTTTTCCTCGCCTCTCCTCC
ex_locus78_uc009kok.1	TTCCTGAGGGCTCCTGGATC	ATGGGAGGTCCTGTGTCCTT
ex_locus79_uc012fxr.1	CTGCTCACCCAGAAGACGAA	GCTGGAAAGTGCTCACTCCT
ex_locus80_Kcnq1ot1_ex1-2	GTGTGGTCGGCCTCATTTTG	GACTGCTGTAGTAGTTGTGGGA
ex_locus80_Kcnq1ot1_ex2-3	CCTCTAGTTGCACCAAACAAGT	GGTGCTTTCTGTTTAGGTTGCC
ex_locus80_Kcnq1ot1_ex3-4	GCACCATGCTTGGATGGATT	ACCAGTTGTATGCCATGTCGT
ex_locus81_brain.117732.1	AGTCCCCCTATGATCCAGAG	GGCTGTCTTCTGGCTGTGAT
ex_locus83_liver.63435.2	TCCAAAGGAGTGATTGGCAT	TGTAAAGGGGAAAGCAGAAGGA
ex_locus84_spleen.116248.2	CTGACTCTGTGCTGCCTCAG	TTCCTTGGAGCTCTGTCTGT
ex_locus89_Nctc1	TGGGTCCCCAGGTCTTAGAG	GCCGGGAGTCTCTTGTTCAA
ex_locus91_Dio3os	CGCACTCCCTAGAATGCTCC	CAGGTGGGAAGTGCTGATGT
ns_locus36_uc012bqn.2_1F	GAGTATACATATACATACGCACATACC	ATATTCATATACACATGTTCATGCACA
ns_locus61_uc009kpr.1_1F	AGGCACTTCAGCGTCAAGAA	GTGTCCCCAAGGAATCCCTG
in_locus3_spleen.9702.4	AGTCAGAATAGTATTACAGTGCATCCT	CGGTGGAAGCTGACCATTGA
in_locus5_Heart.6056.2	TGCCATCTTATTGCCGTGGA	GCGTAAAAGGCTTGCAGAACA
in_locus6_brain.15210.1	CAGGCTCCATGTTGGTGTGA	GGCGTCAGTTCCATTTGCTG
in_locus7_BB612635	CCATGATCCCAAGCCCTTGT	GCAGTGAGCGAATTAGCAAGC
in_locus8_spleen.9955.1	GCCTATGAGAAGCCAAGGCA	GTTTCACGGAAGAGCCTCCA
in_locus9_spleen.11110.6	GGCCAACCAGAAGCTTTGTG	TGGGATAGACCTGGTGCTGA
in_locus16_Meg3	CTTCACTGTCTGCAGGGTCC	AGTGCTTTTCCTGCCTCCTC
in_locus17_brain.2332.1	AAACCACCCAAACGCCAAAG	CCCTCAACCAGACACACCTG
in_locus18_lung.16003.1	GAGGAGGACTGATGTTGGATCA	AGGGCCATCTGTAAGTACCA
in_locus19_placenta.12560.10	GGGCAGCCAATCTCACAGAT	TGTTCTCCAGACCAAGCACA
in_locus23_introncovered	GGAGCCAACTGGGACAAAGA	TGATCAGCAGCAGTGAGTGA
in_locus32_Heart.14437.1	GGGATGAACGGGTGGGTTTT	GGTGGGTGTGGGTATGAAACT
in_locus40_spleen.68728.3	AACTGGGCCAGAAACTGACC	CTCTGGCTGCTGAAGGGATT
in_locus42_AV506483_trunc	CGGCTGCTCTAGTCAGTTGG	CAACAGACGACCACCATCCA
in_locus48_Heart.32260.9	CTTGGCAAACACCCACCTTG	TGTTTGATCAGAGAGATGTGATTTGA
in_locus56_liver.51114.2	CCATGGCTTGATGGTTCCCT	ATCCCCACACATCCCAGAGT
in_locus63_brain.112117.2_trunc	ACTGCTTGTGTGTTGAATAAATAGT	AGTCTCCTTATTTCCCTGTAAGGT
in_locus64_uc009heo.2	GGGTGTAATGGTATAAACTTACAACCT	AGGCTTTCATGTTAGGAGAGCA
in_locus65_BY123339_trunc	GGGGGAGCTGAAAAGGAAGA	ATGAGTTTATTGAGGAAACTGACAG
in_locus66_uc009hey.1	TCTCAAAAGGAACTCCAACCTGT	GAACAAATGAGGCTCTTGGGC
in_locus67_brain.112173.11_trunc	GCAGGAAATTAGATGCCAGGC	TGCATTTTTGTCTTCAAGATAAACCA
in_locus68_brain.112299.50	CCCACACTGATTGGAATCATGG	GGAGGTTCAAGCATGACTTTGT

in_locus68_testes.165186.1	TGTAGCAGGACCACAAAAGACA	AGCATATACTTGTTAAAAGATCCCACA
in_locus74_intronic	ACCTTGAGCTGGTTCCAGTC	AGTGAGAACATTTAAAAAGTTTGCCT
in_locus75_Heart.36806.1	TGTGTCACTCCACAGAGCTAG	AGATTCAGCTATCCCTCCTGC
in_locus79_intron_longiso	ATAGGAGACCCTGTGCCAGT	GATTCACCTGCATTCGTGGG
in_locus85	CCAACCCATGGCTTGTTCTA	CCTACTCAATTGTGGCAGGC
in_locus86	ACTGGTATGAATTCATGCTATCACA	AAGGTGAAATCTCCAAAGTTATGTT
in_locus87_rev	TGTGAGATGATCTATGGGTGGT	ATACTGGGATTGTGGCTGGC
in_locus89_Nctc1_trunc	TTTGTTGTCCACCCCACCAA	GACCCCAGGATGCAGGAAAA
in_locus90_H19	TCAGGCAGAGCAAAGGCATC	CAGCGGCCTCAGTCTTTACT

Table 13: Primers designed for non-coding targets for eight tissues. The names consist of an indicator whether they were designed as intronic (in_) or exonic (ex_) primers, the locus number and the name of the assembled exon models these primers were designed for.

Non-coding targets for testes and placenta

	Target	Forward primer	Reverse primer
testes	locus2_testes.16526.2	CTGGGGCCTTGATTCTCTGG	TTAGCTGGACTGAGGAGGCA
	locus10_placenta.29429.1	GGTTTCTCAGTGTCTGCCCA	GTCCCCAGGTCCTTTTGCTT
	locus11_placenta.29429.6	TGACTCCTGTGCCTAAGAAGC	ACCCTGGCAGAGGAACAAAG
	locus28_testes.54891.1	GGTACAGGATGCTGGTCTGG	TCTGTCCGCCTAACTCCTGA
	locus41_testes.115840.6	CTATAAGAGGCAAGGGGCGG	GAGGGAGGGGTGTCATGTTG
	locus43_CB231709	CACAGCCACACACAGTAGGT	ACGTTGCCCATGTGGTAGAG
	locus52_testes.152076.1	GGATCCTTCGTCCGATGCTT	CACCAGATCCAAGCAAACACA
	locus50_testes.153466.6	CTCTCCACTCTGCTCCCTCT	CAGCAACTGCAGCAGGTCA
	locus70_testes.165255.6	AGTCCCAGGTGCCTTCAAAC	TCTGAAGCCATTGTCAAGCA
	locus60_testes.170669.4	CTGAGACTACGCCCAGTGAC	GTAGTGCAGAGTGGCTCCTT
	locus77_testes.171865.2	GCCAAAGCTCAGGTCAAGTC	GAGAGTCTCCCCTGGGATCC
	locus82_testes.189458.2	TGTTTCCATAGTTCCTGGAGGC	GTCAACCCCAAACCACTCCA
	locus13_testes.20251.1	TCCCCCTTCGTAGAAAAATG	ATGGTACAGCCAATCATCCA
	locus21_testes.7040.1	CCCTGGGGACTTTGAATCCC	ACAATTAGAATTCCGGGCCCT
	locus33_ESC.14534.1	AGTAGGTGGCTCCTTCTGGT	GGTGTCTGCCTGACCTGTAG
	locus29_testes.61595.1	TGAGAGAGGTTCTGGTTCACC	AACCAGGCAGTGTAGAAGTGG
	locus31_testes.61934.2	AGGCCTGTGGATTTGAAGCA	GCCTCTCAGGAGACCTCTCA
	locus4_testes.16472.1	CCTCCTCATGTTCCTCCTCC	GCCCCCTGAGATCATACGTG
	locus47_testes.152112.1	TCCATGGAGGCAGGAAGAGA	AGAGGGCAGTCGTACCTCAT
	locus49_testes.152378.8	TGCTTTGTTTTATGGCCTCTTGA	GGACACATTTCTGGCGCATC
	locus57_testes.156256.2	TTTGCGGGTTCTTTGTCCCT	TCCCCCTCACATCCACGAAT
	locus24_testes.38888.1	TGAGCACCACAGCAATCGAT	GTATGAAGCGCAGGGTCACT
	locus30_ESC.14066.1	ACTTCCTCTTGCCACCCATG	TCACATGGCCACCTTCACTC
placenta	locus22_BY718104	CTTACCACACTAGCCTGGGC	TCTCTGGGCCAACAACAGTT
	locus34_placenta.51023.7	GAGTTGGGATCTGGACGGTG	CCTTCGCTCTCCTTCCTTCC

Table 14: Primers designed for non-coding targets in testes and placenta. Same naming format as for the other non-coding primers.

List of Figures

1 Grb10 - different isoforms 8

2 Examples of imprinted clusters 9

3 Filtering pipeline for ncRNAs 15

4 Automatically designed primers for Igf2r 18

5 PCR results for CxF f2 and FxC f4 - Igf2r 18

6 Sanger sequencing results - Airn, Igf2r, Impact 19

7 Quantified Sanger sequencing - Igf2r, Airn 20

8 Selection of known imprinted protein coding candidates 22

9 Known imprinted candidates: Visual inspection of PCR results in CxF f2 lung . 25

10 Known imprinted candidates: PCR results for known imprinted protein coding candidates in lung 26

11 Known imprinted candidates: Snrpn, an example for a caveat of validating PCR results using MIRTA expression data 27

12 Known imprinted candidates: Comparison of visual inspection results in all four lung samples 28

13 Overview of PCR results for known imprinted protein coding genes 29

14 Main steps of the non-coding pipeline 31

15 Non-coding pipeline: Validation results 32

16 Non-coding pipeline: Comparison of criteria for coding potential 33

17 Selection of non-coding candidates 34

18 Non-coding candidates: Kcnq1ot1 - an example for primer design 35

19 Non-coding candidates: PCR results for non-coding candidates in heart 38

20 Non-coding candidates: Visual inspection of PCR results in CxF f2 heart 39

21 Non-coding candidates: Comparison of visual inspection results in all four heart samples 40

22 Overview of PCR results for non-coding genes 41

23 Non-coding candidates: PCR results for non-coding candidates in testes/placenta 43

24 Non-coding candidates: Visual inspection of PCR results in testes/placenta . . 44

25 Workflow for the creation of a multiplexed sample for sequencing. 45

26 The sonication principle 46

27 Library preparation - Sonication test 47

28 Library preparation - Sonication of set 1 48

29 Library preparation - Finished library of set 1 48

30 Sequencing results: Quality assessment of sequencing data 50

31 Sequencing results: Coverage of PCR products for known imprinted protein coding candidates 51

32 Sequencing results: Coverage of intronic PCR products for non-coding candidates 51

33 Sequencing results: Junction validation in brain 52

34 Sequencing results: Junction validation in heart 53

35 Sequencing results: Junction validation in kidney 53

36 Sequencing results: Junction validation in leg muscle 54

37 Sequencing results: Junction validation in all four tissues 55

38 Sequencing results: Junction validation in testes 55

39 Sequencing results: SNP coverage in all PCR products and four tissues 56

40 Comparison of visual inspection results with sequencing coverage 61

41 Susceptibility of coverage analysis to alternative exons - Mest 63

42 Example of a tissue specific splice variant 65

43 Brain specific coverage of a Kcqn1ot1 exon 66

44 A candidate showing coverage of additonal exons 67

45 A possible novel single exon transcript priorly annotated as spliced 68

46 A junction showing continuous coverage, indicating an unspliced transcript . . 68

List of Tables

1 Examples for murine imprinted gene clusters 6

2 Selected known imprinted protein coding candidates 23

3 Known imprinted candidates: Summary of visual inspection results in all four lung samples 28

4 Reproducibility of PCR reactions for known imprinted protein coding candidates 30

5 Summary of selected non-coding candidates 36

6 Non-coding candidates: Evaluation of correlation between MIRTA signal and PCR results. 37

7 Non-coding candidates: Summary of visual inspection results in all four lung samples 41

8 Reproducibility of PCR reactions for non-coding candidates 42

9 Library preparation - Sets of samples prepared together 44

10 Sequencing - General data 49

11 Primers for the preliminary study on Airn and Igf2r 79

12 Primers - Known imprinted targets for eight tissues 81

13 Primers - Non-coding targets for eight tissues 83

14 Primers - Non-coding targets for testes and placenta 83

Bibliography

[1] Bartolomei MS, Ferguson-Smith AC. Mammalian genomic imprinting. Cold Spring Harb Perspect Biol. 2011 Jul;3(7).

[2] Spencer HG, Clark AG. Non-conflict theories for the evolution of genomic imprinting. Heredity (Edinb). 2014 Jan;Epub ahead of print.

[3] Barlow DP. Genomic imprinting: a mammalian epigenetic discovery model. Annu Rev Genet. 2011;45:379–403.

[4] Barlow DP, Bartolomei MS. Genomic imprinting in mammals. Cold Spring Harb Perspect Biol. 2014;6(2).

[5] Bartolomei MS. Genomic imprinting: employing and avoiding epigenetic processes. Genes Dev. 2009 Sep;23(18):2124–2133.

[6] Crouse HV. The Controlling Element in Sex Chromosome Behavior in Sciara. Genetics. 1960 Oct;45(10):1429–1443.

[7] Khosla S, Mendiratta G, Brahmachari V. Genomic imprinting in the mealybugs. Cytogenet Genome Res. 2006;113(1-4):41–52.

[8] Kinoshita T, Ikeda Y, Ishikawa R. Genomic imprinting: a balance between antagonistic roles of parental chromosomes. Semin Cell Dev Biol. 2008 Dec;19(6):574–579.

[9] Waters AJ, Bilinski P, Eichten SR, Vaughn MW, Ross-Ibarra J, Gehring M, et al. Comprehensive analysis of imprinted genes in maize reveals allelic variation for imprinting and limited conservation with other species. Proc Natl Acad Sci U S A. 2013 Nov;110(48):19639–19644.

[10] McGrath J, Solter D. Maternal Thp lethality in the mouse is a nuclear, not cytoplasmic, defect. Nature. 1984;308(5959):550–551.

[11] McGrath J, Solter D. Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell. 1984 May;37(1):179–183.

[12] Surani MA, Barton SC, Norris ML. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature. 1984;308(5959):548–550.

[13] Barlow DP, Stoeger R, Herrmann BG, Saito K, Schweifer N. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature. 1991 Jan;349(6304):84–87.

[14] DeChiara TM, Robertson EJ, Efstratiadis A. Parental imprinting of the mouse insulin-like growth factor II gene. Cell. 1991 Feb;64(4):849–859.

[15] Bartolomei MS, Zemel S, Tilghman SM. Parental imprinting of the mouse H19 gene. Nature. 1991 May;351(6322):153–155.

[16] Kawahara M, Wu Q, Takahashi N, Morita S, Yamada K, Ito M, et al. High-frequency generation of viable mice from engineered bi-maternal embryos. Nat Biotechnol. 2007 Sep;25(9):1045–1050.

[17] Pant PVK, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. Analysis of allelic differential expression in human white blood cells. Genome Res. 2006 Mar;16(3):331–339.

[18] Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M, et al. Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. Genome Res. 2003 Jun;13(6B):1402–1409.

[19] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63.

[20] Tran DA, Bai AY, Singh P, Wu X, Szabo PE. Characterization of the imprinting signature of mouse embryo fibroblasts by RNA deep sequencing. Nucleic Acids Res. 2014 Feb;42(3):1772–1783.

[21] Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, et al. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science. 2010 Aug;329(5992):643–648.

[22] DeVeale B, van der Kooy D, Babak T. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. PLoS Genet. 2012;8(3):e1002600.

[23] Babak T, Deveale B, Armour C, Raymond C, Cleary MA, van der Kooy D, et al. Global survey of genomic imprinting by transcriptome sequencing. Curr Biol. 2008 Nov;18(22):1735–1741.

[24] Okae H, Hiura H, Nishida Y, Funayama R, Tanaka S, Chiba H, et al. Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. Hum Mol Genet. 2012 Feb;21(3):548–558.

[25] Wang X, Soloway PD, Clark AG. A survey for novel imprinted genes in the mouse placenta by mRNA-seq. Genetics. 2011 Sep;189(1):109–122.

[26] Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. PLoS One. 2008;3(12):e3839.

[27] Lagarrigue S, Martin L, Hormozdiari F, Roux PF, Pan C, van Nas A, et al. Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. Genetics. 2013 Nov;195(3):1157–1166.

[28] Zou F, Sun W, Crowley JJ, Zhabotynsky V, Sullivan PF, Pardo-Manuel de Villena FF. A Novel Statistical Approach for Jointly Analyzing RNA-seq Data from F1 Reciprocal Crosses and Inbred Lines. Genetics. 2014 Feb;Epub ahead of print.

[29] Schulz R, Woodfine K, Menheniott TR, Bourc'his D, Bestor T, Oakey RJ. WAMIDEX: a web atlas of murine genomic imprinting and differential expression. Epigenetics. 2008;3(2):89–96.

[30] Morison IM, Ramsay JP, Spencer HG. A census of mammalian imprinting. Trends Genet. 2005 Aug;21(8):457–465.

[31] Williamson CM, Blake A, Thomas S, Beechey CV, Hancock J, Cattanach BM, et al.. World Wide Web Site - Mouse Imprinting Data and References. Oxfordshire; 2014. Available from: `http://www.har.mrc.ac.uk/research/genomic_imprinting/`.

[32] Moore T, Haig D. Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet. 1991 Feb;7(2):45–49.

[33] Prickett AR, Oakey RJ. A survey of tissue-specific genomic imprinting in mammals. Mol Genet Genomics. 2012 Aug;287(8):621–630.

[34] Plagge A, Gordon E, Dean W, Boiani R, Cinti S, Peters J, et al. The imprinted signaling protein XL alpha s is required for postnatal adaptation to feeding. Nat Genet. 2004 Aug;36(8):818–826.

[35] Curley JP, Barton S, Surani A, Keverne EB. Coadaptation in mother and infant regulated by a paternally expressed imprinted gene. Proc Biol Sci. 2004 Jun;271(1545):1303–1309.

[36] da Rocha ST, Charalambous M, Lin SP, Gutteridge I, Ito Y, Gray D, et al. Gene dosage effects of the imprinted delta-like homologue 1 (dlk1/pref1) in development: implications for the evolution of imprinting. PLoS Genet. 2009 Feb;5(2):e1000392.

[37] Varmuza S, Mann M. Genomic imprinting–defusing the ovarian time bomb. Trends Genet. 1994 Apr;10(4):118–123.

[38] Day T, Bonduriansky R. Intralocus sexual conflict can drive the evolution of genomic imprinting. Genetics. 2004 Aug;167(4):1537–1546.

[39] Spencer HG, Clark AG. A chip off the old block: a model for the evolution of genomic imprinting via selection for parental similarity. Genetics. 2006 Oct;174(2):931–935.

[40] McDonald JF, Matzke MA, Matzke AJ. Host defenses to transposable elements and the evolution of genomic imprinting. Cytogenet Genome Res. 2005;110(1-4):242–249.

[41] Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet. 2011 Aug;12(8):565–575.

[42] Sanford JP, Clark HJ, Chapman VM, Rossant J. Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse. Genes Dev. 1987 Dec;1(10):1039–1046.

[43] Mertineit C, Yoder JA, Taketo T, Laird DW, Trasler JM, Bestor TH. Sex-specific exons control DNA methyltransferase in mammalian germ cells. Development. 1998 Mar;125(5):889–897.

[44] Li E, Bird A. DNA Methylation in Mammals. In: Allis CD, Jenuwein T, Reinberg D, editors. Epigenetics. Cold Spring Harbor Press; 2007. .

[45] Wan LB, Bartolomei MS. Regulation of imprinting in clusters: noncoding RNAs versus insulators. Adv Genet. 2008;61:207–223.

[46] Cowley M, Oakey RJ. Retrotransposition and genomic imprinting. Brief Funct Genomics. 2010 Jul;9(4):340–346.

[47] Wutz A, Smrzka OW, Barlow DP. Making sense of imprinting the mouse and human IGF2R loci. Novartis Found Symp. 1998;214:251–9; discussion 260–3.

[48] Kim J, Bergmann A, Lucas S, Stone R, Stubbs L. Lineage-specific imprinting and evolution of the zinc-finger gene ZIM2. Genomics. 2004 Jul;84(1):47–58.

[49] Yamasaki K, Joh K, Ohta T, Masuzaki H, Ishimaru T, Mukai T, et al. Neurons but not glial cells show reciprocal imprinting of sense and antisense transcripts of Ube3a. Hum Mol Genet. 2003 Apr;12(8):837–847.

[50] Menheniott TR, Woodfine K, Schulz R, Wood AJ, Monk D, Giraud AS, et al. Genomic imprinting of Dopa decarboxylase in heart and reciprocal allelic expression with neighboring Grb10. Mol Cell Biol. 2008 Jan;28(1):386–396.

[51] Sleutels F, Zwart R, Barlow DP. The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature. 2002 Feb;415(6873):810–813.

[52] Yamasaki Y, Kayashima T, Soejima H, Kinoshita A, Yoshiura KI, Matsumoto N, et al. Neuron-specific relaxation of Igf2r imprinting is associated with neuron-specific histone modifications and lack of its antisense transcript Air. Hum Mol Genet. 2005 Sep;14(17):2511–2520.

[53] Ferron SR, Charalambous M, Radford E, McEwen K, Wildner H, Hind E, et al. Postnatal loss of Dlk1 imprinting in stem cells and niche astrocytes regulates neurogenesis. Nature. 2011 Jul;475(7356):381–385.

[54] Gould TD, Pfeifer K. Imprinting of mouse Kvlqt1 is developmentally regulated. Hum Mol Genet. 1998 Mar;7(3):483–487.

[55] Korostowski L, Raval A, Breuer G, Engel N. Enhancer-driven chromatin interactions during development promote escape from silencing by a long non-coding RNA. Epigenetics Chromatin. 2011;4:21.

[56] Miyoshi N, Kuroiwa Y, Kohda T, Shitara H, Yonekawa H, Kawabe T, et al. Identification of the Meg1/Grb10 imprinted gene on mouse proximal chromosome 11, a candidate for the Silver-Russell syndrome gene. Proc Natl Acad Sci U S A. 1998 Feb;95(3):1102–1107.

[57] Smith FM, Holt LJ, Garfield AS, Charalambous M, Koumanov F, Perry M, et al. Mice with a disruption of the imprinted Grb10 gene exhibit altered body composition, glucose homeostasis, and insulin signaling during postnatal life. Mol Cell Biol. 2007 Aug;27(16):5871–5886.

[58] Charalambous M, Cowley M, Geoghegan F, Smith FM, Radford EJ, Marlow BP, et al. Maternally-inherited Grb10 reduces placental size and efficiency. Dev Biol. 2010 Jan;337(1):1–8.

[59] Hitchins MP, Bentley L, Monk D, Beechey C, Peters J, Kelsey G, et al. DDC and COBL, flanking the imprinted GRB10 gene on 7p12, are biallelically expressed. Mamm Genome. 2002 Dec;13(12):686–691.

[60] Arnaud P, Monk D, Hitchins M, Gordon E, Dean W, Beechey CV, et al. Conserved methylation imprints in the human and mouse GRB10 genes with divergent allelic expression suggests differential reading of the same mark. Hum Mol Genet. 2003 May;12(9):1005–1019.

[61] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996–1006.

[62] Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014 Jan;42(Database issue):D764–D770.

[63] Santoro F, Barlow DP. Developmental control of imprinted expression by macro non-coding RNAs. Semin Cell Dev Biol. 2011 Jun;22(4):328–335.

[64] Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature. 2000 May;405(6785):486–489.

[65] Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature. 2000 May;405(6785):482–485.

[66] Thorvaldsen JL, Duran KL, Bartolomei MS. Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. Genes Dev. 1998 Dec;12(23):3693–3702.

[67] Hikichi T, Kohda T, Kaneko-Ishino T, Ishino F. Imprinting regulation of the murine Meg1/Grb10 and human GRB10 genes; roles of brain-specific promoters and mouse-specific CTCF-binding sites. Nucleic Acids Res. 2003 Mar;31(5):1398–1406.

[68] Williamson CM, Ball ST, Dawson C, Mehta S, Beechey CV, Fray M, et al. Uncoupling antisense-mediated silencing and DNA methylation in the imprinted Gnas cluster. PLoS Genet. 2011 Mar;7(3):e1001347.

[69] Mancini-Dinardo D, Steele SJS, Levorse JM, Ingram RS, Tilghman SM. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev. 2006 May;20(10):1268–1282.

[70] Shin JY, Fitzpatrick GV, Higgins MJ. Two distinct mechanisms of silencing by the KvDMR1 imprinting control region. EMBO J. 2008 Jan;27(1):168–178.

[71] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012 Sep;489(7414):101–108.

[72] Kowalczyk MS, Higgs DR, Gingeras TR. Molecular biology: RNA discrimination. Nature. 2012 Feb;482(7385):310–311.

[73] Szymanski M, Erdmann VA, Barciszewski J. Noncoding RNAs database (ncRNAdb). Nucleic Acids Res. 2007 Jan;35(Database issue):D162–D164.

[74] Storz G. An expanding universe of noncoding RNAs. Science. 2002 May;296(5571):1260–1263.

[75] Collins LJ. The RNA infrastructure: an introduction to ncRNA networks. Adv Exp Med Biol. 2011;722:1–19.

[76] Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 2007 May;17(5):556–565.

[77] Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–166.

[78] Yao H, Brick K, Evrard Y, Xiao T, Camerini-Otero RD, Felsenfeld G. Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. Genes Dev. 2010 Nov;24(22):2543–2555.

[79] Espinoza CA, Allen TA, Hieb AR, Kugel JF, Goodrich JA. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nat Struct Mol Biol. 2004 Sep;11(9):822–829.

[80] Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007 Jun;129(7):1311–1323.

[81] Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N. Requirement for Xist in X chromosome inactivation. Nature. 1996 Jan;379(6561):131–137.

[82] Wutz A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat Rev Genet. 2011 Aug;12(8):542–553.

[83] Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011 Apr;472(7341):120–124.

[84] Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. Gene regulation by the act of long non-coding RNA transcription. BMC Biol. 2013;11:59.

[85] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012 Feb;482(7385):339–346.

[86] Martens JA, Laprade L, Winston F. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature. 2004 Jun;429(6991):571–574.

[87] van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, van Oudenaarden A, et al. Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. Cell. 2012 Sep;150(6):1170–1181.

[88] Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, et al. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science. 2012 Dec;338(6113):1469–1472.

[89] Santoro F, Mayer D, Klement RM, Warczok KE, Stukalov A, Barlow DP, et al. Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window. Development. 2013 Mar;140(6):1184–1195.

[90] Pauler FM, Barlow DP, Hudson QJ. Mechanisms of long range silencing by imprinted macro non-coding RNAs. Curr Opin Genet Dev. 2012 Jun;22(3):283–289.

[91] Schmidt JV, Levorse JM, Tilghman SM. Enhancer competition between H19 and Igf2 does not mediate their imprinting. Proc Natl Acad Sci U S A. 1999 Aug;96(17):9733–9738.

[92] Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, et al. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell. 2008 Oct;32(2):232–246.

[93] Zhang H, Zeitz MJ, Wang H, Niu B, Ge S, Li W, et al. Long noncoding RNA-mediated intrachromosomal interactions promote imprinting at the Kcnq1 locus. J Cell Biol. 2014 Jan;204(1):61–75.

[94] Hudson QJ, Seidl CIM, Kulinski TM, Huang R, Warczok KE, Bittner R, et al. Extra-embryonic-specific imprinted expression is restricted to defined lineages in the post-implantation embryo. Dev Biol. 2011 May;353(2):420–431.

[95] Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, et al. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science. 2008 Dec;322(5908):1717–1720.

[96] Landers M, Bancescu DL, Le Meur E, Rougeulle C, Glatt-Deeley H, Brannan C, et al. Regulation of the large (approximately 1000 kb) imprinted murine Ube3a antisense transcript by alternative exons upstream of Snurf/Snrpn. Nucleic Acids Res. 2004;32(11):3480–3492.

[97] Chamberlain SJ, Brannan CI. The Prader-Willi syndrome imprinting center activates the paternally expressed murine Ube3a antisense transcript but represses paternal Ube3a. Genomics. 2001 May;73(3):316–322.

[98] Meng L, Person RE, Huang W, Zhu PJ, Costa-Mattioli M, Beaudet AL. Truncation of Ube3a-ATS unsilences paternal Ube3a and ameliorates behavioral defects in the Angelman syndrome mouse model. PLoS Genet. 2013 Dec;9(12):e1004039.

[99] Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, et al. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 2014 Jan;42(Database issue):D98–103.

[100] Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 2011 Jan;39(Database issue):D146–D151.

[101] Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS. NRED: a database of long noncoding RNA expression. Nucleic Acids Res. 2009 Jan;37(Database issue):D122–D126.

[102] Zhang Y, Guan DG, Yang JH, Shao P, Zhou H, Qu LH. ncRNAimprint: a comprehensive database of mammalian imprinted noncoding RNAs. RNA. 2010 Oct;16(10):1889–1901.

[103] Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011 Sep;25(18):1915–1927.

[104] Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A. 2013 Feb;110(8):2876–2881.

[105] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010 May;28(5):503–510.

[106] Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012 Sep;22(9):1775–1789.

[107] Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002 Dec;420(6915):563–573.

[108] Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009 Mar;458(7235):223–227.

[109] Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009 Jul;106(28):11667–11672.

[110] Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013 Jul;154(1):26–46.

[111] Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, et al. The abundance of short proteins in the mammalian proteome. PLoS Genet. 2006 Apr;2(4):e52.

[112] Prasanth KV, Spector DL. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes Dev. 2007 Jan;21(1):11–42.

[113] Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol. 1999 Apr;16(4):512–524.

[114] Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, et al. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res. 2007 Dec;17(12):1823–1836.

[115] Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011 Jul;27(13):i275–i282.

[116] Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013 Apr;41(6):e74.

[117] Fickett JW. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 1982 Sep;10(17):5303–5318.

[118] Gish W, States DJ. Identification of protein coding regions by database similarity search. Nat Genet. 1993 Mar;3(3):266–272.

[119] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014 Jan;42(1):D222–D230.

[120] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011 Jul;39(suppl 2):W29–W37.

[121] Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007 Jul;35(suppl 2):W345–W349.

[122] Huang R. Mapping and regulation of imprinted macro non-coding RNAs in the mouse genome [Ph.D. dissertation]. University of Vienna; 2010.

[123] Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, et al. H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome Res. 2009 Feb;19(2):221–233.

[124] Capelo LP, Beber EH, Huang SA, Zorn TMT, Bianco AC, Gouveia CHA. Deiodinase-mediated thyroid hormone inactivation minimizes thyroid hormone signaling in the early development of fetal skeleton. Bone. 2008 Nov;43(5):921–930.

[125] Parker-Katiraee L, Carson AR, Yamada T, Arnaud P, Feil R, Abu-Amero SN, et al. Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. PLoS Genet. 2007 May;3(5):e65.

[126] Redrup L, Branco MR, Perdeaux ER, Krueger C, Lewis A, Santos F, et al. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Development. 2009 Feb;136(4):525–530.

[127] Szabo PE P E, Mann JR. Allele-specific expression and total expression levels of imprinted genes during early mouse development: implications for imprinting mechanisms. Genes Dev. 1995 Dec;9(24):3097–3108.

[128] Hu JF, Balaguru KA, Ivaturi RD, Oruganti H, Li T, Nguyen BT, et al. Lack of reciprocal genomic imprinting of sense and antisense RNA of mouse insulin-like growth factor II receptor in the central nervous system. Biochem Biophys Res Commun. 1999 Apr;257(2):604–608.

[129] Pauler FM, Stricker SH, Warczok KE, Barlow DP. Long-range DNase I hypersensitivity mapping reveals the imprinted Igf2r and Air promoters share cis-regulatory elements. Genome Res. 2005 Oct;15(10):1379–1387.

[130] Mertes F, Elsharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics. 2011 Nov;10(6):374–386.

[131] De Wilde B, Lefever S, Dong W, Dunne J, Husain S, Derveaux S, et al. Target enrichment using parallel nanoliter quantitative PCR amplification. BMC Genomics. 2014;15(1):184.

[132] Balabanski L, Antov G, Dimova I, Ivanov S, Nacheva M, Gavrilov I, et al. Next-generation sequencing of BRCA1 and BRCA2 in breast cancer patients and control subjects. Mol Clin Oncol. 2014 May;2(3):435–439.

[133] Bowne SJ, Sullivan LS, Koboldt DC, Ding L, Fulton R, Abbott RM, et al. Identification of disease-causing mutations in autosomal dominant retinitis pigmentosa (adRP) using next-generation DNA sequencing. Invest Ophthalmol Vis Sci. 2011 Jan;52(1):494–503.

[134] Jones MA, Bhide S, Chin E, Ng BG, Rhodenizer D, Zhang VW, et al. Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. Genet Med. 2011 Nov;13(11):921–932.

[135] Halbritter J, Diaz K, Chaki M, Porath JD, Tarrier B, Fu C, et al. High-throughput mutation analysis in patients with a nephronophthisis-associated ciliopathy applying multiplexed barcoded array-based PCR amplification and next-generation sequencing. J Med Genet. 2012 Dec;49(12):756–767.

[136] Murakami-Kawaguchi S, Takasawa S, Onogawa T, Nata K, Itaya-Hironaka A, Sakuramoto-Tsuchida S, et al. Expression of Ins1 and Ins2 genes in mouse fetal liver. Cell Tissue Res. 2014 Feb;355(2):303–314.

[137] Hernandez A. Structure and function of the type 3 deiodinase gene. Thyroid. 2005 Aug;15(8):865–874.

[138] Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, et al. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. Current protocols in molecular biology. 2010;p. 19–10.

[139] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome research. 2005;15(10):1451–1455.

[140] Goecks J, Nekrutenko A, Taylor J, Team TG. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.

[141] Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007 May;23(10):1289–1291.

[142] Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3–new capabilities and interfaces. Nucleic Acids Res. 2012 Aug;40(15):e115.

[143] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.

[144] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar;26(6):841–842.

[145] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–2079.

Abstract

Genomic imprinting is an epigenetic phenomenon characterized by monoallelic epression of genes in a parent-of-origin specific fashion. Imprinted genes are mostly organised in clusters and imprinted expression in these clusters is often regulated by an imprinted non-coding RNA (ncRNA). Imprinted expression is often tissue specific and dependent on the developmental stage. This has been shown for example in the biallelic expression of *Igf2r* in neurons or the relaxation of imprinted expression of *Kcnq1* in adult tissues. The development of next generation RNA sequencing brought about a feasible method to study the whole transcriptome. This technique was applied by a number of studies to investigate imprinted genes. Since most studies focused mainly on protein coding genes and early developmental stages little is known about imprinted expression in adult tissues and imprinted non-coding RNAs, with the exception of the main regulators in imprinted clusters.

The work presented in this thesis is part of a larger effort to evaluate the imprinting status of both known imprinted protein coding genes and novel long non-coding RNAs (lncRNAs) in eight adult tissues of the mouse. This was done in reciprocal crosses of the two mouse strains FVB/NJ and Cast/EiJ. As this would be quite cost intensive to do by whole transcriptome sequencing I implemented a targeted PCR based approach for the analysis of 96 amplicons for known imprinted genes and 71 amplicons for non-coding RNAs. I implemented a script to allow time efficient design of primers fulfilling the condition that the amplicons included at least two SNPs for the analysis of imprinted expression. Amplicons were plexed together and sequenced by massive parallel sequencing. The results showed that the amplicons were well covered with high reproducibility among biological replicates. The data was also used for a validation of the predicted *de novo* assembled lncRNA exon models. Almost all of the junctions covered by amplicons could be validated in at least one sample and some cases showed evidence for inefficient splicing. Overall these results indicated that the exon models are reliable although further investigation into additional exons and splicing efficiency is advisable. Coverage of SNPs was high on average so the results produced in this study will be a good dataset to continue with the analysis of imprinted expression in adult tissues.

Zusammenfassung

Genomische Prägung ist ein regulatorisches Phänomen, welches durch die monoallelische Genexpression von entweder dem mütterlichen oder dem väterlichen Allel charakterisiert ist. Dies wird durch einen epigenetischen Mechanismus etabliert. Geprägte Gene sind in Clustern organisiert und geprägte Genexpression in diesen Clustern ist oft von einer geprägten nicht kodierenden RNS (nkRNS) reguliert. Ge-prägte Genexpression ist des Weiteren oft gewebespezifisch und abhängig vom Entwicklungsstadium. Dies wurde zum Beispiel anhand der biallelischen Expression von *Igf2r* in post-mitotischen Neuronen und der Relaxation der geprägten Expression von *Kcnq1* in späteren Stadien der Entwicklung demonstriert. Durch die Entwicklung von RNS Sequenzierung der nächsten Generation entstand eine neue Methode zur Analyse des Transkriptoms welche in mehreren Studien zur Detektion von geprägten Genen verwendet wurde. Da die meisten dieser Studien auf proteinkodierende Gene und frühe Entwicklungsstadien beschränkt sind, ist wenig über geprägte Genexpression in adulten Geweben oder geprägte nkRNS bekannt. Die Ausnahme bilden hier die bekannten Hauptregulator-nkRNS der zuvor genannten Cluster.

Die Experimente die in dieser Arbeit präsentiert werden sind Teil eines größeren Projekts, welches zum Ziel hat geprägte Genexpression von proteinkodierenden Genen und nicht kodierenden RNS in acht Geweben von ausgewachsenen Mäusen zu untersuchen. Dies wurde in reziproken Kreuzungen zwischen den Mausstämmen Cast/EiJ und FVB/NJ durchgeführt. Da es relativ teuer wäre diese Analyse durch Sequenzieren des ganzen Transkriptoms durchzuführen wurde eine auf PCR basierende zielgerichtete Methode entwickelt um 96 Amplikons für bekannte geprägte protein-kodierende Gene und 71 Amplikons für *de novo* assemblierte nkRNS zu untersuchen. Zu diesem Zweck wurde ein Skript implementiert welches das zeiteffiziente Design von Primern ermöglicht, mit deren Hilfe sich Amplikons generieren lassen die mindestens zwei SNPs (Einzelnukleotid-Polymorphismen) inkludieren. Die Amplikons wurden mittels RNS Sequenzierung der nächsten Generation sequenziert. Die Resultate zeigten, dass die Sequenzierung der Amplikons gut funktionierte und die Reproduzierbarkeit der Resultate zwischen den biologischen Replikaten hoch war. Die Daten wurden außerdem für eine Validierung der *de novo* assemblierten Exonmodelle verwendet. Beinahe alle splice junctions konnten auf diese Weise validiert werden wobei teilweise zusätzliche, nicht im assemblierten Modell enthaltene Exons oder Hinweise auf ineffizientes Splicing erkennbar waren. Die Mehrheit der SNP Positionen war tief genug sequenziert, so dass sich diese Daten gut für die Analyse von geprägter Genexpression eignen werden.

Curriculum Vitae

Christoph Dotter

Higher Education:

2011 - 2014	Master’s program Biological Chemistry, University of Vienna
2008 - 2011	Bachelor’s program Chemistry, University of Innsbruck
2006 - 2009	Bachelor’s program Biomedical Informatics, UMIT

Research experience:

04/2013 - 04/2014	Master’s project on "Imprinted expression of known and novel transcripts in multiple tissues of the mouse", group of Denise Barlow, CeMM Research Center for Molecular Medicine, Vienna
02-03/2013	Research project on "Recombinant expression of polyphenoloxidase 4 from A. bisporus in E. coli" Institute for Biophysical Chemistry, University of Vienna
01-02/2013	Research project on "Synthesis of Elastin-like peptides using SPPS" Institute for Biochemistry, University of Vienna
08-09/2012	Internship at the group of Denise Barlow, CeMM Research Center for Molecular Medicine, Vienna
02/2012	Internship as a Bioinformatician at the group of Denise Barlow, CeMM Research Center for Molecular Medicine, Vienna
03-07/2011	Bachelor’s project on "Functional characterization of the arginine methylation site in the p65 subunit of NF-κB", Institute of Biochemistry, University of Innsbruck
03-06/2009	Bachelor’s project on "Detection of Gene Expression Patterns in different Gleason Scores of Prostate Cancer exploiting Meta Analysis", Institute for Bioinformatics and Translational Research, UMIT

Acknowledgements

I want to thank Prof. Denise Barlow for giving me the opportunity to work on this thesis in her lab. It was a real pleasure to work in a lab as well organized as hers. Thank you also for your work on providing the list of imprinted genes I used for the candidate selection and for your help on correcting this thesis.

Big thanks also go to Dr. Florian Pauler for being a great supervisor. He was always there when I had even the tiniest question and also considered my own ideas during this project. Thank you also for your patience at times, especially during the time I was writing this thesis, and for the great job you did on correcting it. It was a real pleasure to work with you.

I also want to specially mention two members in the lab who made this year as great as it was: Philipp Bammer, who became a great friend and even though I may not have acted like it all the time I really enjoyed helping you out. I can only return the thanks you gave me for being a great climbing partner and I'll always remember our music filled evenings in the lab. Alexandra Kornienko is the second person I want to mention here, since she rarely failed to brighten up my day with her humour and all sorts of sweets. I really liked having someone around I could share those family pizzas during the late hours in the lab.

Special thanks to Justyna for always supplying me with the huge load of pipette tips and agarose gels I needed, for keeping the lab as organized as it is and of course for sharing my humour and a small section of my music taste.

Thanks to Philipp Günzl for showing me how to prepare libraries, Matthias Farlik for helping with the sonication, Thomas Penz and Klaudia Bagienski for helping with the Experion chips and Andreas Sommer and Ru Huang at the CSF for doing the sequencing. Also thanks to all the lab members not mentioned so far, Daniel, Daniela, Markus, Rita, Sabine, Sarah and Tomasz for the many coffee breaks and for providing the work atmosphere I enjoyed so much.

Huge thanks go to my parents who always supported me and enabled me to study what really interests me, even if I didn't find it at first. Your patience and support is unmatched and I don't know where I would be without it.

Last but not least I want to thank everyone not mentioned so far who helped keeping me motivated, especially during the months I spent writing.