Department of Bioinformatics and Computational Biology

Technische Universität München

Master's Thesis in Bioinformatics

Variation of HERV elements in the KORA cohort

Julian Schmidt

Department of Bioinformatics and Computational Biology

Technische Universität München

Master's Thesis in Bioinformatics

Variation of HERV elements in the KORA cohort

Variation von HERV elementen in der KORA Kohorte

Author: Julian Schmidt
Supervisor: Dr. Matthias Heinig
Advisor: Johann Hawe
Submitted: 15.04.2018

I confirm that this master's thesis is my own work and I have documented all sources and material used.

13.04.2018

Julian Schmidt

Contents

1 Introduction **1**
- 1.1 Multiomics . . . 1
- 1.2 Human endogenous retroviruses . . . 2
- 1.3 Effect network analysis . . . 3

2 Data **5**
- 2.1 HERV annotation . . . 5
- 2.2 KORA . . . 6
  - 2.2.1 Expression . . . 7
  - 2.2.2 Methylation . . . 8
  - 2.2.3 Genotypes . . . 9
  - 2.2.4 Covariates . . . 10
  - 2.2.5 Methylation quantitative trait loci . . . 10
- 2.3 Transcription factor binding . . . 10
- 2.4 Protein interaction network . . . 11
- 2.5 Chromatin states . . . 12

3 Methods **13**
- 3.1 Overlaps . . . 13
- 3.2 Data normalization . . . 13
- 3.3 eQTL/eQTM calculation . . . 14
- 3.4 Functional Analysis of Gene Sets . . . 14
- 3.5 Gaussian Graphical Models . . . 15
  - 3.5.1 Methodological Background . . . 15
  - 3.5.2 Data collection . . . 18
  - 3.5.3 BDgraph parameters . . . 19

4 Results **21**
- 4.1 Normalized Data . . . 21
- 4.2 HERV region features . . . 21
  - 4.2.1 Expression . . . 22
  - 4.2.2 Methylation . . . 23
  - 4.2.3 Genotypes . . . 24
  - 4.2.4 Chromatin states . . . 24
- 4.3 eQTLs . . . 24
- 4.4 eQTMs . . . 26
- 4.5 meQTLs . . . 27
- 4.6 HERV related regulatory networks . . . 28

4.6.1 Data collection . . . 28
4.6.2 General analysis . . . 28
4.6.3 Detailed analysis . . . 28

5 Discussion 31
5.1 HERV region features . . . 31
5.1.1 Expression . . . 31
5.1.2 Methylation . . . 31
5.1.3 Genotypes . . . 31
5.2 eQTLs . . . 31
5.3 eQTMs . . . 31
5.4 meQTLs . . . 31
5.5 Regulatory networks . . . 31

6 Conclusion and Outlook 33

Chapter 1

Introduction

1.1 Multiomics

The rise of high-throughput technologies allows researchers to paint a more complete picture of the mechanisms of cell processes than ever before. Multi-omics, the combination of the analysis of different types of cellular molecules, is made possible by that. The objective of multi-omics studies is to draw conclusions about cellular mechanisms and understand the flow of information by finding associations and patterns connecting them[1]. In this work we worked on data from the fields of transcriptomics, genomics and epigenomics to investigate the effects of changes in human endogenous retroviruses (HERVs), which are introduced in more detail in section 1.2.

Transcriptomics examines the occurrence and level of RNAs on a genome-wide scale. Following the central dogma of molecular biology, RNA is the intermediate between DNA and proteins and therefore of central importance[]. With the late progress in transcriptome studies, even more effects of RNA have been discovered in the form of several functional RNAs[2]. RNA levels can be measured using a technology called RNA-seq[3] or probe-based arrays[4], as was the case with the data used in this work.

The second field, genomics, describes variation between genomes. In this work we focused on single nucleotide polymorphisms (SNPs), which are single base changes. Genotypic information fills a special place in multi-omics analyses, as it not dependent on the environment and can be considered static in ones lifetime. Therefore, it is considered as the cause of changes instead of being a downstream effect of other changes[1]. This means they are an exceptional anchor for the start of multi-omics analyses. Genotypic changes are known to cause Mendelian diseases, when the change lies in the coding region of a protein. They also influence gene expression levels, which in turn is thought to be the cause of common diseases[1].

Epigenomics is the youngest field of 'omics' that is used in this work. It describes genome-wide reversible modifications of DNA and DNA-associated proteins. Among these are multiple different histone modifications and DNA methylation. Epigenetic modifications play a major role in the regulation of gene transcription[5, 6] and in general are a additional layer of information on top of the DNA sequence. Changes in epigenetic modifications are influenced by genetic and environmental factors[1].

DNA methylation, the modification that is the focus in this work, is the best understood and researched epigenetic mark[7]. The term refers to the addition of a methyl group to the fifth carbon atom of a cytosine ring, thus transforming the cytosine to 5-methylcytosine. In humans this happens almost exclusively in the context of a cytosine neighbored by a guanine - a CpG-site. As the complementary sequence of CG is also CG, methylation usually occurs on both strands.

There are about 28 million CpG sites in the human genome, of which 60-80% are generally methylated[7]. The distribution of CpGs can be divided into two modes: The majority of the genome is sparsely populated by CpGs, that are usually methylated. Second there are CG-dense regions, which are predominantly nonmethylated. These are called CpG islands and occur mainly around the transcription start sites of genes[6]. The methylation of these CpG islands is strongly connected to the silencing of adjacent genes, by preventing the transcription initiation[6]. Other mechanisms that are negatively affected by DNA methylation are general promoter, enhancer and insulator activity as well as transcription factor binding[7].

1.2 Human endogenous retroviruses

When the draft for the first human genome was published in 2001[8] it was expected, that it would allow to fully understand the human genome. But soon after genes were annotated less than 5% of the genome was found to be protein coding[9] and later on the amount of exonic protein coding DNA was estimated to be as low as 1.2%[10]. The remaining majority of the genome was often labeled as non-functional "junk DNA"[11].

Since then this notion has generally been revoked. Especially within the two phases of the Encyclopedia of DNA Elements (ENCODE) project[10], participation in some functional role has been assigned to up 80.4% of the human genome.

Still up to this day there are areas in the genome that are difficult to analyze. Especially repetitive sequences cause huge problems for any method that relies on sequencing[12]. In human repetitive DNA makes up about 50% of the whole Genome[12]. There are five major classes of repetitive elements: Satellites, short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), DNA transposons and long terminal repeat (LTR) retrotransposons.

In this work we will focus on a subclass of the latter repeat class. Human endogenous retroviruses (HERVs) belong to the class of LTR retrotransposons and make up about 8% of the human genome[13]. HERVs originate from retroviruses that infected the germ line of humans or human ancestors up to 30 million years ago[14]. Therefore, they became dormant in the genome and are inherited in a Mendelian fashion. Newly integrated endogenous retroviruses (ERVs) share the typical structure of the viruses they originate from. This means they contain at least the four main genes in the order $5' - gag - pro - pol - env - 3'$. Gag encodes the matrix and capsid proteins. Pro codes for a protease. Pol is the reverse transcriptase and integrase and env codes for the surface proteins. Furthermore, they have two flanking LTRs that originate from a duplication of the sequence, where the retrovirus was integrated[14]. An example of the structure of a HERV element is shown in figure 1.1.

However, endogenous retroviruses in human are generally not able to replicate. This is due to a high level of degeneration of the HERV gene sequences due to mutations and deletion introducing frame shifts and premature stop codons. Homologous recombination of the flanking LTRs can also lead to elements containing only a single LTR and losing the entire coding body[14]. Additionally, the promoter activity of LTRs is commonly silenced by hypermethylation[7].

Regardless of their loss of capability to replicate there still are multiple ways in which

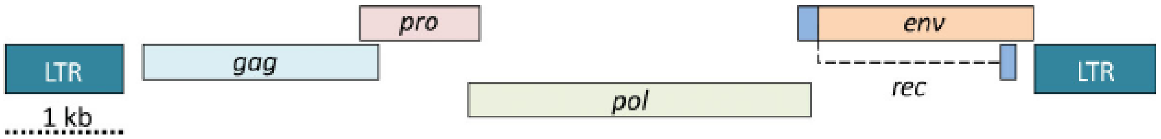

Figure 1.1: Exemplified structure of an intact HERV element. Based on a HERV-K provirus, which is considered the most recently introduced in humans and contains the most complete viral genes. Taken from Young et. al. [15]

HERVs can influence human cells. LTRs of HERVs can act as promoters or enhancers to regulate gene expression. This has been associated to various cancers[16]. While most HERV genes are degenerated, some are still expressed. A change in the transcription of several HERV subclasses was found in schizophrenia subjects, but the mechanisms of HERV involvement are not yet well understood[17]. Finally, non-coding HERV RNAs have been found to play a role in control of stem cell properties[13].

The multiomics data analyses performed in this work were either started from an anchor relating to HERV elements or performed genome wide and afterwards filtered to retain HERV specific results.

1.3 Effect network analysis

Many multiomics analyses are based on the calculation of associations within or between different data layers. Examples are co-expression of genes or the association of genotypic variants and transcription, protein or methylation levels. These approaches allow to identify associated entities by calculating correlations between single entities over available individuals or samples. They can't give any resolution on the direction of an effect. Furthermore, purely correlation based methods can not differentiate between direct and indirect effects[1].

Multomics data can be organized into networks, where each entity from the different layers is a node and the significant pairwise associations, that exceed a given threshold, are resembled by undirected edges. However when using Pearson correlation as measurement of association, the networks tend to be rather dense, as indirect associations are still contained[18].

An approach to remove these indirect associations is the use of partial correlation coefficients. The partial correlation coefficient can be seen as the pairwise correlation between two variables conditioned against their correlations with all other variables. A scheme of how partial correlations work is shown in figure **??**.

An approach that combines partial correlation coefficients with Bayesian structure learning is called "Gaussian Graphical Models"[18] (GGMs). We used GGMs to investigate the mechanisms of cellular control related to variation within HERV elements.

The Gaussian Graphical Models used in this work are described in more detail in section 3.5.

Chapter 2

Data

2.1 HERV annotation

As mentioned in section 1.2 HERV elements tend to be highly degenerated. Therefore, it is no trivial task to pertain a complete HERV annotation for the human genome. Furthermore, classification and naming of endogenous retroviruses is not performed in an unified manner but rather dependent on how and by whom a given element was identified. Instead of collecting a library of known HERV sequences and performing sequence search ourselves, HERV annotations were pertained from RepeatMasker[19] repeat library. RepeatMasker is a tool that screens DNA sequences against a library interspersed repeats and low complexity DNA sequences. It generates an annotation of identified repeats and masks them in the query sequence.

The track representing all identified repeats from the RepeatMasker library for human genome hg19 was downloaded from the UCSC genome browser download section (`http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz`). It contains a total of 5298130 occurrences of repeats. Each entry consists of 17 values. These are repeat name, the repeat class and family, as well as the chromosome, strand, and the genomic start and end position of the repeat occurrence. Furthermore it contains the area of the known repeat sequence, that is covered by the occurrence. It also contains the quality of the alignment of the repeat sequence to the annotated position using the Smith Waterman alignment score[20] and the number of base mismatches, deletions and insertions per thousand base pairs. Finally, there is an indexing field used to speed up chromosome range queries and the first digit of the id field in the RepeatMasker output file. The first 10 lines of the track are shown in table 2.1.

To extract HERV elements the annotation was filtered on different columns generating three sets of variable size. Multiple HERV sets were created as an attempt to cover different possible definitions of HERV elements.

bin	swScore	milliDiv	milliDel	milliIns	genoName	genoStart	genoEnd	genoLeft	strand	repName	repClass	repFamily	repStart	repEnd	repLeft	id
585	1,504	13	4	13	chr1	10,000	10,468	$-2.49 \cdot 10^8$	+	(CCCTAA)n	Simple_repeat	Simple_repeat	1	463	0	1
585	3,612	114	270	13	chr1	10,468	11,447	$-2.49 \cdot 10^8$	-	TAR1	Satellite	telo	−399	1,712	483	2
585	437	235	186	35	chr1	11,503	11,675	$-2.49 \cdot 10^8$	-	L1MC	LINE	L1	−2,236	5,646	5,449	3
585	239	294	19	10	chr1	11,677	11,780	$-2.49 \cdot 10^8$	-	MER5B	DNA	hAT-Charlie	−74	104	1	4
585	318	230	38	0	chr1	15,264	15,355	$-2.49 \cdot 10^8$	-	MIR3	SINE	MIR	−119	143	49	5
585	203	162	0	0	chr1	16,712	16,749	$-2.49 \cdot 10^8$	+	(TGG)n	Simple_repeat	Simple_repeat	1	37	0	6
585	239	338	148	0	chr1	18,906	19,048	$-2.49 \cdot 10^8$	+	L2a	LINE	L2	2,942	3,104	−322	7
585	652	346	85	42	chr1	19,947	20,405	$-2.49 \cdot 10^8$	+	L3	LINE	CR1	3,042	3,519	−970	8
585	270	331	7	27	chr1	20,530	20,679	$-2.49 \cdot 10^8$	+	Plat_L3	LINE	CR1	2,802	2,947	−639	9
585	254	279	47	39	chr1	21,948	22,075	$-2.49 \cdot 10^8$	+	MLT1K	LTR	ERVL-MaLR	15	142	−453	1

Table 2.1: First ten rows of the RepeatMasker annotation on hg19

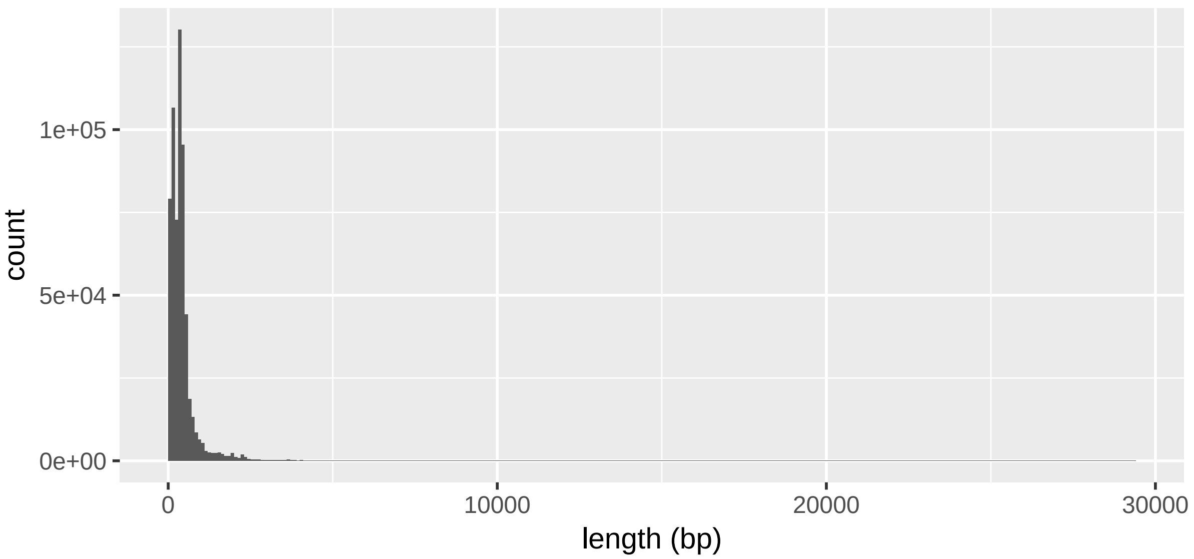

Figure 2.1: Length distribution of elements in HERV S2

This work only discerned between different HERV element types by defining different HERV sets. Therefore, within each set annotations, whose genomic positions overlap or are directly adjacent, were merged into one element.

A first set, HERV set 1 (HERV S1) was constructed by extracting all elements that contained "ERV" in the repeat name column. This set The resulting 42508 annotations condensed to 35358 elements after merging. The elements have a mean width of 956 bp and cover a total of 33.8 Mbp, which is about 1.04% of the human genome.

Alternatively, filtering the annotation for "ERV" in the super family column leads to 696689 annotations. HERV set 2 (HERV S2) contains all endogenous retroviral sequences found in the human genome. Therefore, it is the most comprehensive set and it is the focus of detailed results that are shown and discussed. After merging overlapping and adjacent annotations this led to 612594 elements. Their mean width is 430 bp and they make up to 263.4 Mbp or ca 8.13% of the human genome. The distribution of element lengths in HERV S2 is shown in Figure 2.1

A third set "HERV set 3" (HERV S3) was constructed by filtering the repeat name column for "HERV", which resulted in 21361 annotations. As this set contains only elements that are explicitly named "HERV", we expect it to contain elements, that were inserted directly into the human genome. Merging overlapping and adjacent annotations resulted in 15284 elements with an average width of 1480 bp and a combined length of 22.6 Mbp.

2.2 KORA

The functional data for our analysis of gene regulatory mechanisms are made up expression, methylation and genotype data. These were collected in the light of the platform for Cooperative Health Research in the Region of Augsburg, short KORA. It contains health surveys as well as examinations of individuals of German nationality living in the area of Augsburg, Bavaria. The objective of KORA is to track changes in health conditions

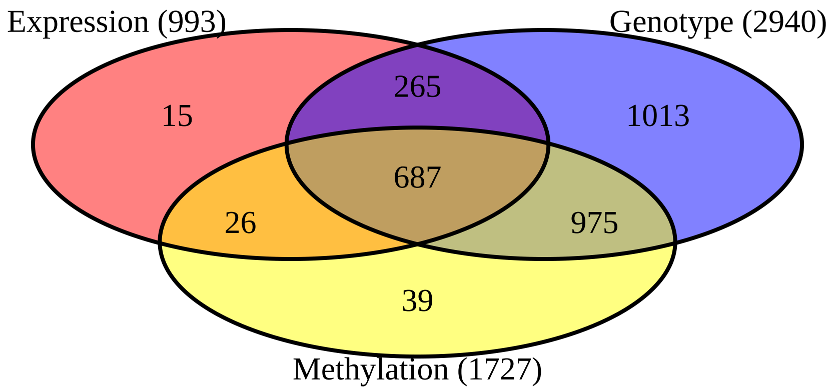

Figure 2.2: Number of samples with genotype, expression and methylation measurements in KORA F4 Survey

over a long period in order to identify and examine the causes, effects and development of chronical diseases.

The data used in this work originates from the KORA F4 Survey, which was conducted from 2006 to 2008 and comprised samples of 3080 individuals. Of these data was available for 3020 individuals. F4 is a follow up study to the KORA S4 Survey performed from 1999 to 2001. It contains 4261 individuals.

All measurements were performed on whole blood samples. Houseman blood counts[21] describing the composition of different cell types for each individual had previously been calculated from the methylation data. They were used to weight certain cell type specific data.

Not all essays are available for all samples. Therefore different analyses were performed on varying sets of individuals according to availability of the required data types. A diagram of the samples available for each essay can be seen in figure 2.2. The joint analysis, using Gaussian Graphical Models, was performed on 687 individuals that had all three essays available.

2.2.1 Expression

The expression data was generated using the HumanHT-12 v3.0 Gene Expression BeadChip. The chip can measure expression values for 49576 probes. However, only 47864 probes represent an actual genomic location. The remaining probes are control probes used to assess the quality of the measurements and infer the background for measurements.

Measurements for 993 individuals are available from the KORA F4 Survey. The¡ comprise values for a total of 48803 probes per sample. Probes that do not map to a genomic location were excluded in all analyses, leaving 47864 probes. Of these 29521 are annotated to a total of 19288 genes.

The available data had already been quantile normalized, corrected for the background and log-transformed.

Figure 2.3 shows the distribution and coefficient of variation over all probes and samples. Apart from a tail of some probes with very high expression values figure 2.3A roughly resembles a standard distribution.

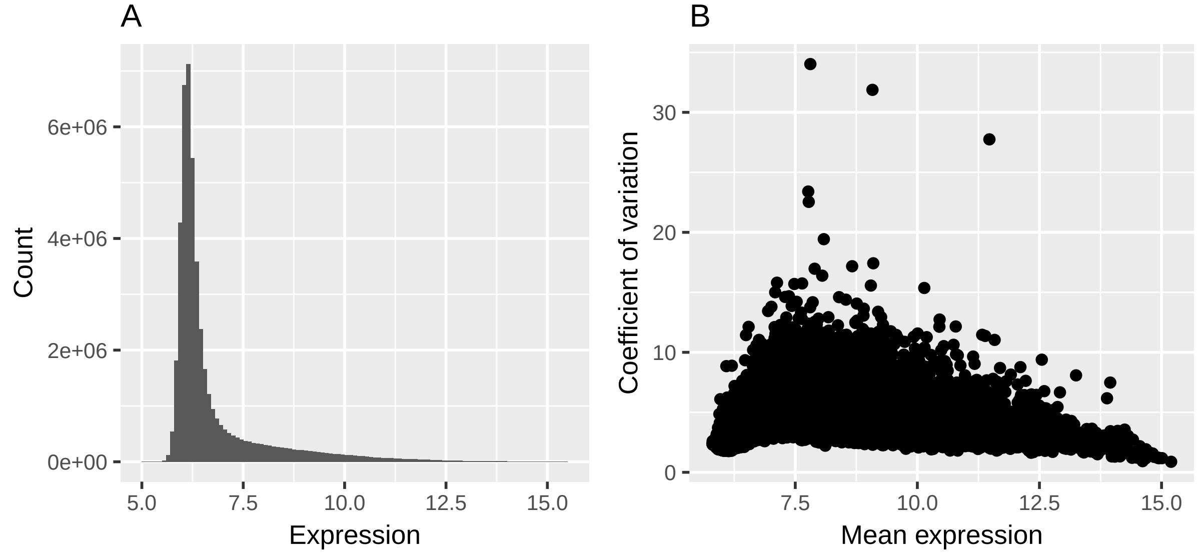

Figure 2.3: Distribution of expression values (A) and coefficient of expression variation (B) of 47864 probes in 993 individuals.

Probes that could not be mapped to genes were not discarded to avoid losing data for HERV regions, which are only sparsely annotated with genes and were the focus of this work. Therefore, most analyses were performed on probe level or a mix of probes and genes. Whenever there were multiple probes mapping to the same gene, the mean of the expression values of these probes was taken as estimation of expression level for the gene.

2.2.2 Methylation

DNA methylation was measured using the Infinium HumanMethylation450K BeadChip, which interrogates methylation levels at 485577 genomic locations. Methylation intensities had already been corrected and transformed to beta values. Beta values have a range from 0 to 1 and represent the fraction of copies of a CpG site that are methylated in a sample.

No imputation of missing measurements was performed for the methylation data. Therefore, only 44335 probes have no missing values at all and measurements in all samples available. Overall there are about 11.2 million missing values, which makes up 1.3% of all measurements and on average 23 samples missing per probe. The histogram of proportion of missing values per probe (A) and sample (B) can be seen in figure 2.4.

Methylation data was available for 1727 individuals and 485512 sites, which make up all 'cg' and 'ch' probe type probes. The distribution of beta values and the variances over all samples and probes is shown in figure 2.5. As seen in figure 2.5A, the interrogated CpG sites tend to be either entirely methylated or unmethylated within single samples. This means that all cells in a sample have the same, stable methylation pattern, which is expected as methylation is known to be stable within a certain condition[]. However, as subfigure B shows there is some variance between individuals in most CpG-sites. The low variance for very low and very high betas is explained by the fact that for a CpG site to have a mean close to 0 all samples have to have a value close to 0. Therefore, the variance also is very low.

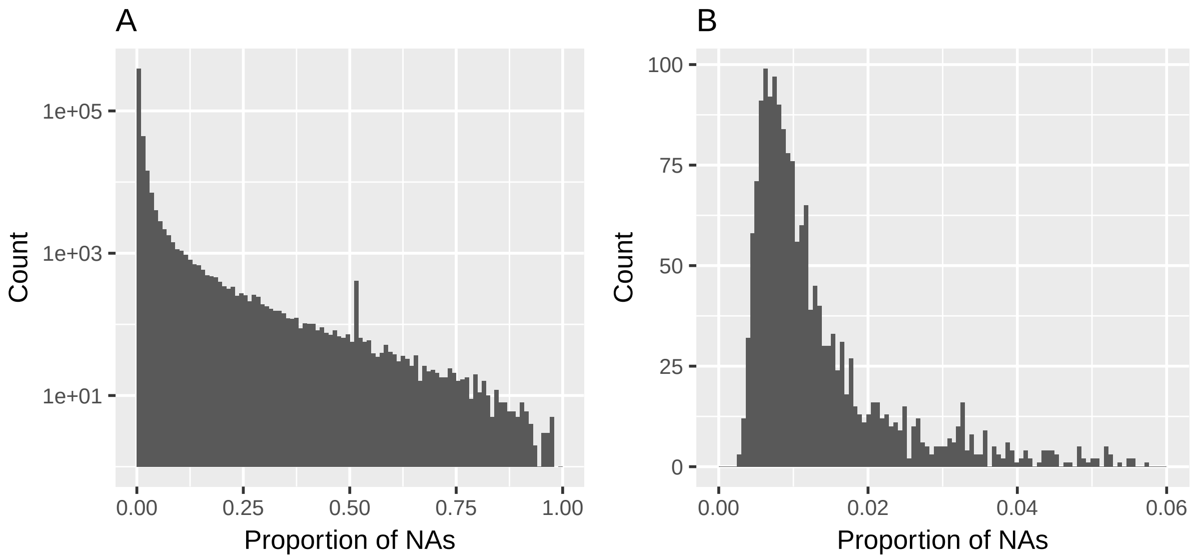

Figure 2.4: Histograms of proportions of missing values per CpG-site (A) and sample (B).

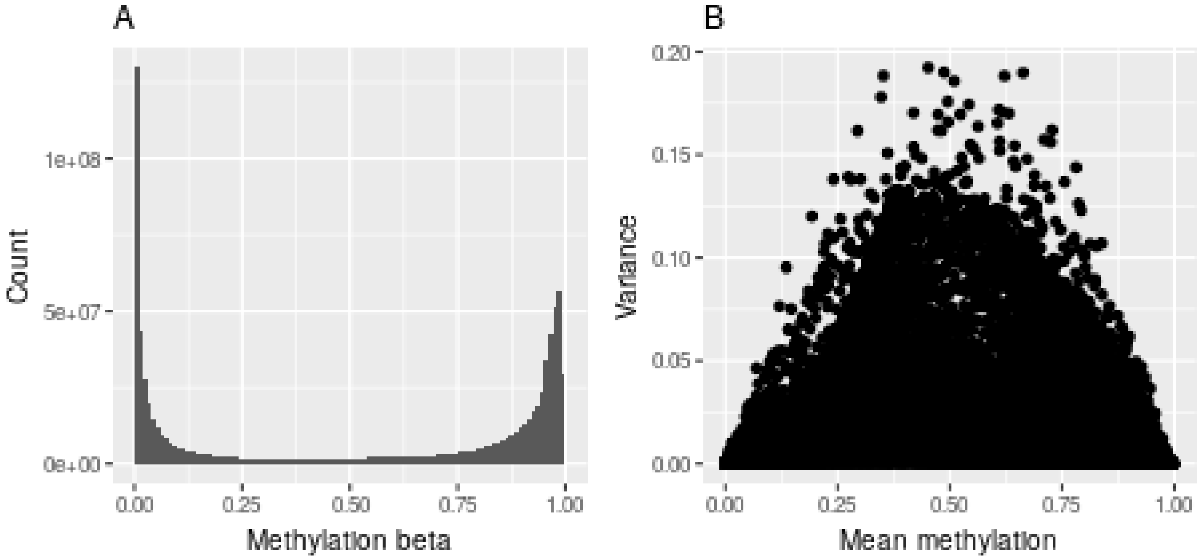

Figure 2.5: Distribution of all methylation beta values (A) and variance (B) of 485577 CpGs between 1727 individuals.

2.2.3 Genotypes

Genotyping was performed with the Affymetrix Axiom array. On the data set used in this work genotypes had already been called using the Iluminus calling algorithm and missing values had been imputed using the IMPUTE2 software[22]. Furthermore, imputation results had already been filtered at IMPUTE value of 0.4.

Additionally SNPs with a minor allele frequency of less than one percent had been removed, which allows used association analyses to be have more powerful test statistics.

In total measurements of 9533127 SNPs for 2940 individuals were available in the form of continuous dosages from 0 to 2 with 0 representing no occurrence of the SNP and 2 meaning the SNP is present in both chromosomes. Non integer values mean that different genotypes were measured in different cells in the sample and the resulting value describes the fraction of the occurrence of the variant.

2.2.4 Covariates

Several covariates were known for each sample. They were used to correct expression and methylation values for shared effects.

The sex of all participants was known. Of the 3020 participants with any available data 1458 were SEX1 and the remaining 1562 were SEX2. The 697 individuals that had all three essays available contained 348 men/women and 339 women/men. Further covariates specific to participating individuals are sex, age, body mass index (BMI) and white blood cell count. An overview of these covariate values for all 3020 individuals and the individuals with all three essays available is shown in table 2.2.

Covariate	Minimum	Maximum	Mean
Age	31	82	56.37
BMI	16.05	55.99	27.65
WBC	2.5	40.9	5.98

Covariate	Minimum	Maximum	Mean
Age	61	81	69.04
BMI	18.99	47.58	28.87
WBC	2.7	12.6	5.9

Table 2.2: Covariate overview

Finally, three experimental factors for the expression measurements, storage time, RNA integrity number (RIN), a measure that describes the degree of degradation of RNA molecules[23], and plate, were known.

2.2.5 Methylation quantitative trait loci

Genotypic variants are known to effect methylation patterns in humans []. Therefore, we used previously processed methylation quantitative trait loci (meQTL) data to explore the mechanism of trans acting SNP-CpG associations. Associated pairs were used as a seed for trans-acting interaction networks.

For meQTLs a stringent definition of trans effects was used: A SNP-CpG pair was called a trans-meQTL only if they are located on different chromosomes. This definition differs from the one used for eQTM and eQTL calculation as described later.

The data set contained a total of 11165559 significantly associated SNP-CpG pairs. 70709 distinct CpG-sites and 2709428 SNPs were part of at least one meQTL. On average each SNP was associated with 4.12 CpGs and each CpG was part of on average 157.91 meQTLs.

Of these 467915 pairs consisting of 3592 CpG-sites and 200761 SNPs were between different chromosomes.

2.3 Transcription factor binding

DNA methylation plays a major role in the specifics of transcription factor binding sites[24]. We used transcription factor binding sites obtained from two publicly available sources to enrich information about CpG-sites of interest with transcription factors that bind nearby:

The first source was the third version of the track "Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs"[25] downloaded from the UCSC genome browser download section (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz`).

It combines 690 high quality ENCODE ChIP-seq data sets, which were processed with the Factorbook motif discovery and annotation pipeline[25]. The pipeline uses the tools MEME-ChIP[26] and FIMO[27] from the MEME software suite and merges discovered motifs with known motifs from Jaspar[28] and TransFac[29] using machine learning methods and manual curation.

The track contains a total of 438044 distinct peaks for 161 transcription factors in 91 cell types. For our analyses we filtered and combined the peaks for 23 blood related cell types. This leaves a total of 2475316 peaks for 125 transcription factors.

The second source was the ReMap project[30]. It combines 395 publicly available ChIP-seq data sets covering 132 different transcription factors across 83 cell lines. ReMap uses Bowtie2[31] to map reads to the human genome and the tool MACS[32] for peak calling. The final data set was downloaded from the ReMap website(`http://tagc.univ-mrs.fr/remap/download/All/filPeaks_public.bed.gz`).

There were a total of 8905156 peaks for 131 transcription factors in the data set. After filtering for 19 blood related cell types 1372245 peaks for 35 different transcription factors remained.

Combining both filtered data sets lead to a total of 3847561 peaks of 145 different transcription factors.

2.4 Protein interaction network

Protein interaction data were used to include potential interaction partners in the generation of interaction networks.

Human protein interaction data were downloaded from the STRING database version 9[33]. STRING is a database for direct and functional protein associations that is compiled by combining multiple sources: Experimental evidence of interaction, known pathways and protein complexes from curated databases, co-expression analysis, knowledge transfer from other species based on gene orthology, and automated text-mining of scientific literature[34].

The data set contained 3019612 pairwise protein interactions between 16590 different proteins. For each pair associations scores for the described categories were available. Protein interactions were filtered for the availability of experimental and/or database evidence by excluding pairs with and association score of 0 in both categories, which left 375702 interactions between 12769 different proteins.

Further filtering was performed, by excluding proteins, whose genes were not found expressed in whole blood. Therefore, expression values from RNA-seq experiments for 55 tissues in human were downloaded from the Genotype-Tissue Expression (GTEx) Project[35]. Transcripts with an RPKM (Reads per kilobase per million mapped reads) of more than 0.1 in blood were considered expressed. Interactions where one or both members were not represented by these 17343 transcripts were removed.

Finally, connected components were calculated on the network represented by the remaining pairwise interactions and all genes not within the biggest connected component were removed. The final network contained 8546 proteins and 97447 interactions.

2.5 Chromatin states

Chromatin state annotations were downloaded from Roadmap Epgenomics Core 15-state model (`http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final`). The Model provides a whole genome chromatin state annotation of 200 bp wide windows to the following 15 states: Active Transcription Start Site (TSS), Flanking Active TSS, Transcription at gene 5' and 4', Strong transcription, Weak transcription, Genic enhancers, Enhancers, ZNF genes & repeats, Heterochromatin, Bivalent/Poised TSS, Flanking Bivalent TSS/Enhancer, Bivalent Enhancer, Repressed PolyComb, Weak Repressed PolyComb and Quiescent/Low. The model is available for 127 diverse cell lines.

It was generated using ChromHMM v1.10[36] on the chromatin marks H3K4me1, H3K4me3, H3K27me3, H3K9me3, and H3K36me3. ChromHMM is based on a multivarate Hidden Markov Model.

In this work the annotations for 27 blood related cell lines were used. When combining the annotation for a genomic location of interest like a HERV element or a SNP, houseman blood counts were used to account of the cell composition of the whole blood samples.

Chapter 3

Methods

Most computations were done using the statistical computing environment and programing language R, version 3.4.1[37]. Pre filtering steps on large files, that could not be loaded into random access memory, were performed using the bash command ”awk”, which runs scripts written in the ”AWK” programming language.

3.1 Overlaps

The main objective of this thesis was to analyze the effects of HERV elements on the regulation of cell activity.

To identify functional features like expression probes, genes, SNPs and CpG-sites that were associated with HERV elements, overlaps between these features and elements were calculated. Features were considered to be of interest for the analysis of an HERV element, if they overlapped by at least one base pair.

The calculation was performed using the function ”findOverlaps” from the Bioconductor package ”GenomicRanges”[38]. The function allows to identify overlaps between ”GRanges” objects. These were constructed from BED files for all HERV sets as well as SNPs and retrieved from the Bioconductor packages available for analyses on the expression and methylation essays, ”illuminaHumanv3.db”[39] and ”FDb.InfiniumMethylation.hg19”[40] respectively.

The search for overlapping features was also performed on ”GRanges” objects including all HERV elements as well as the 1kb or 2kb flanking regions of all elements. This was done to include functional features, that while not directly within HERV regions, are still likely to be influenced by them. Furthermore, this analysis was performed on all three HERV sets.

3.2 Data normalization

Before expression and methylation values were used for analyses, they were corrected for batch effects. For this we used the available covariates to calculate linear models and regress out the respective residuals. The model used for expression values is shown in equation 3.1 and considers the covariates age, sex, RNA integrity number (RIN), plate and storage time. The model for methylation values, shown in equation 3.2, includes the cell composition for the five blood cell types known from the houseman blood counts, as well as the first 20 principal components of the Illumina 450k array control-probes.

$$Expr \sim 1 + age + sex + RIN + plate + storage.time \tag{3.1}$$

$$Meth \sim 1 + CD4T + CD8T + NK + Bcell + Mono + PC_1 + ... + PC_{20} \tag{3.2}$$

3.3 eQTL/eQTM calculation

As a preliminary analysis to interrogate the effect of genotype and methylation variants on expression, expression quantitative trait loci (eQTL) and expression quantitative trait methylation (eQTM) were calculated.

In the following paragraphs the calculation of eQTLs will be described. The calculation of eQTMs was performed analogously, the only difference being that methylation residuals were supplied instead of SNP dosages.

Calculations were performed using the Bioconductor package MatrixEQTL[41]. MatrixEQTL tests for association of SNP-transcript pairs.

We set the parameter `useModel = modelLINEAR`, which leads to the use of an additive linear model. The association is modeled as simple linear regression and the absolute value of the sample correlation is used as test statistic.

After calculating the test statistics the p-values for all pairs that pass a defined significance threshold are calculated. These are corrected multiple testing using a Benjamini-Hochberg procedure[42], adapted for not recording all p-values, as shown in equation **??**

MatrixEQTL is rather efficient because it manages to reduce the calculation of the sample correlation over all SNPs and transcripts to one single large matrix multiplication. This is achieved by transforming the genotype and transcription values.

MatrixEQTL also allows to include covariates in the QTL calculation. As the expression and methylation values used are residuals and therefore already corrected this option is not used. Furthermore, MatrixEQTL can differentiate between cis- and trans-interactions based on distance. The maximal distance to consider a pair on the same chromosome as cis was set to 500 kb.

The p-value threshold for significant cis-QTLs during calculation was set to 10^{-6} and 10^{-8} for trans. As this cutoff is used for uncorrected p-values, a more stringent threshold is set for trans QTLs to account for the bigger number of possible pairs.

3.4 Functional Analysis of Gene Sets

In multiple analyses functional Gene Ontology enrichments were performed.

First a set of all GO annotations with any evidence code for gene symbols was retrieved from the Bioconductor package AnnotationDbi[43].

Then a hypergeometric test[44] for overreprensentation was performed on a set of genes of interest. A custom background set of genes specific to the analysis was given for each enrichment. Finally, the p-values for overrepresented GO terms were adjusted for multiple testing using the Holm method[45]. Terms with an adjusted p-value of less than 0.05 were considered significantly overrepresented.

3.5 Gaussian Graphical Models

In this work Gaussian Graphical Models (GGMs) on multiomics data with a connection to HERV elements were used to investigate the effects and effect pathways of changes in HERV elements. The calculation of GGMs was performed using the R package "BDgraph" version 2.44[46]. First I will shortly go over the theoretical background and the method, following Mohammadi and Wit[47, 46]. Then I will elaborate on how I defined the data sets to generate GGMs on.

3.5.1 Methodological Background

In Gaussian Graphical Models, random variables are represented by nodes. Conditional dependence relationships, or partial correlations, between these variables are represented as undirected edges in a graph $G = (V, E)$, where $V = \{1, 2, ..., p\}$ is a set of nodes and $E \subset V \times V$ the set of edges. A zero mean Gaussian Graphical model with respect to graph G is defined as

$$\mathcal{M}_G = \{\mathcal{N}_p(0, \Sigma | K\Sigma^{-1} \in \mathbb{P}_G\} \tag{3.3}$$

where Σ is the covariance matrix, its inverse K is the precision matrix and $\mathbb{P}_G$ is the cone of symmetric positive definite matrices with elements K_{ij} equal to zero for all $(i, j) \notin E$. This means the random variables representing our data are assumed to follow the multivariate Gaussian distribution $\mathcal{N}(\mu, K^{-1})$, where the mean μ is zero.

With $Z = (Z^{(1)}, ..., Z^{(n)})$ being the observed data of p independent samples, the likelihood function is

$$Pr(Z|K, G) \propto |K|^{n/2} exp\left\{-\frac{1}{2}tr(KU)\right\}, \tag{3.4}$$

where $U = Z^T Z$.

As for multivariate Gaussian distributions conditional dependence is a direct representation of partial correlation[48], we will only use the former going forward.

Conditional independence in the model is represented by the precision matrix in the form that two variables i and j are conditionally independent given the remaining variables, if and only if $K_{ij} = 0$. Therefore, the nonzero entries of K correspond to links in the graph G.

As mentioned in section 1.3, the space of possible graphs is too big to explore exhaustively. When considering a case with as low as $p = 50$ variables, there are $2^{p(p-1)/2} > 10^{300}$ possible graphs. Therefore a heuristic approach is necessary.

In the used implementation of GGM calculation the space of possible graphs and precision matrices is explored by a birth-death Markov chain Monte Carlo (BDMCMC) method. It is based on a continuous time Markov process, where jumps to a larger or a smaller dimension occur based on birth and death rates, that are chosen so that the stationary distribution of the Markov chain represents the chosen posteriori distribution.

More specific in the case of Graphical Models a birth event represents the addition of an edge $e = (i, j) \notin E$ and leads the process to a new state $(G^{+e}, K^{+e}$, where $G^{+e} = (V, E \cup \{e\})$, and K^{+e} is equal to K except for the positions $\{(i, j), (j, i), (j, j)\}$.

Analogously a death event represents the deletion of an edge $e = (i, j) \in E$ and leads the process to a new state $(G^{-e}, K^{-e}$, where $G^{-e} = (V, E \setminus \{e\})$, and K^{-e} is equal to K except for the positions $\{(i, j), (j, i), (j, j)\}$.

These events are considered to be independent Poisson processes. Birth and death rates are conditional to the joint posterior distribution of a graph G and precision matrix K, $Pr(K, G|Z)$.

The birth-rates and death-rates are defined as

$$\beta_e(K) = min\left\{\frac{Pr(G^{+e}, K^{+e}|Z)}{Pr(G, K|Z)}, 1\right\}, \text{ for each } e \notin E, \tag{3.5}$$

$$\delta_e(K) = min\left\{\frac{Pr(G^{-e}, K^{-e}|Z)}{Pr(G, K|Z)}, 1\right\}, \text{ for each } e \in E, \tag{3.6}$$

As birth and death rates are independent the total birth and death rates are

$$\beta(K) = \sum_{e \notin E} \beta_e(K), \tag{3.7}$$

$$\delta(K) = \sum_{e \in E} \delta_e(K). \tag{3.8}$$

The time between successive events is exponentially distributed and has the mean $1/(\beta(K) + \delta(K))$. Therefore, the probability of a birth or death event is given by

$$Pr(\text{birth of link e}) = \frac{\beta_e(K)}{\beta(K) + \delta(K)}, \text{ for each } e \notin E, \tag{3.9}$$

$$Pr(\text{death of link e}) = \frac{\delta_e(K)}{\beta(K) + \delta(K)}, \text{ for each } e \in E. \tag{3.10}$$

The joint posterior distribution of a graph G with the precision matrix K is

$$Pr(K, G|Z) \propto Pr(Z|K)Pr(K|G)Pr(G) \tag{3.11}$$

$Pr(Z|K)$ is the likelihood and already defined. The default for the prior of the graph is $Pr(G) = \frac{1}{|\mathcal{G}|}$, where $\mathcal{G}$ is the graph space. Alternatively $Pr(G)$ can be supplanted by a distribution representing prior information of relationships between variables, if available. Finally, the prior distribution for the precision matrix is defined to be the G-Wishart distribution $W_G(b, D)$ with the following density:

$$Pr(K|G) = \frac{1}{I_G(b, D)}|K|^{\frac{b-2}{2}} exp\left\{-\frac{1}{2}tr(DK)\right\} 1(K \in \mathbb{P}_G), \tag{3.12}$$

where $b > 2$ is the degrees of freedom, D is a symmetric positive definitive matrix, $I_G(b, D)$ is the normalizing constant with respect to the graph G and $1(x)$ is 1 if x is true, and 0 otherwise.

The used implementation of GGMs considers the G-Wishart distribution $W_G(b, D)$ to be a prior distribution for the precision matrix K with the following density:

$$Pr(K|G) = \frac{1}{I_G(b, D)} |K|^{\frac{b-2}{2}} exp \left\{ -\frac{1}{2} tr(DK) \right\} 1(K \in \mathbb{P}_G), \tag{3.13}$$

where $b > 2$ is the degrees of freedom, D is a symmetric positive definitive matrix, $I_G(b, D)$ is the normalizing constant with respect to the graph G and $1(x)$ is 1 if x is true, and 0 otherwise.

In the calculation of GGMs with the package BDgraph the following steps are repeated for a given number of iterations:

First The birth and death rates as well as the waiting time is calculated. Then a jump is simulated according to the probabilities in equations 3.9 and 3.10 and an edge is added or removed. Finally, the precision matrix for the new graph is sampled. This is done using a sampler defined in [47], that I will not elaborate on.

The output of the calculation consists of the graphs, precision matrices and waiting times for each iteration. The posterior probabilities of all visited graphs are estimated by the total waiting time on each graph during the iterations as shown in figure 3.1. This means a graph is considered more likely when it was visited multiple times and/or the total birth and death rates when at this graph were low leading to long waiting times. The approach usually takes some time to converge close to the posterior distribution. Therefore, BDgraph offers the the function to define a number of iterations, that are used as "burn-in" period, which are not considered for the calculation of graph probabilities.

An extension to GGMs that allows the method to be used on both discrete and continuous variables is called Gaussian copula graphical modeling. This is done by introducing a multivariate Gaussian latent variable that transform the raw data in way that makes it usable in the calculation of GGMs.

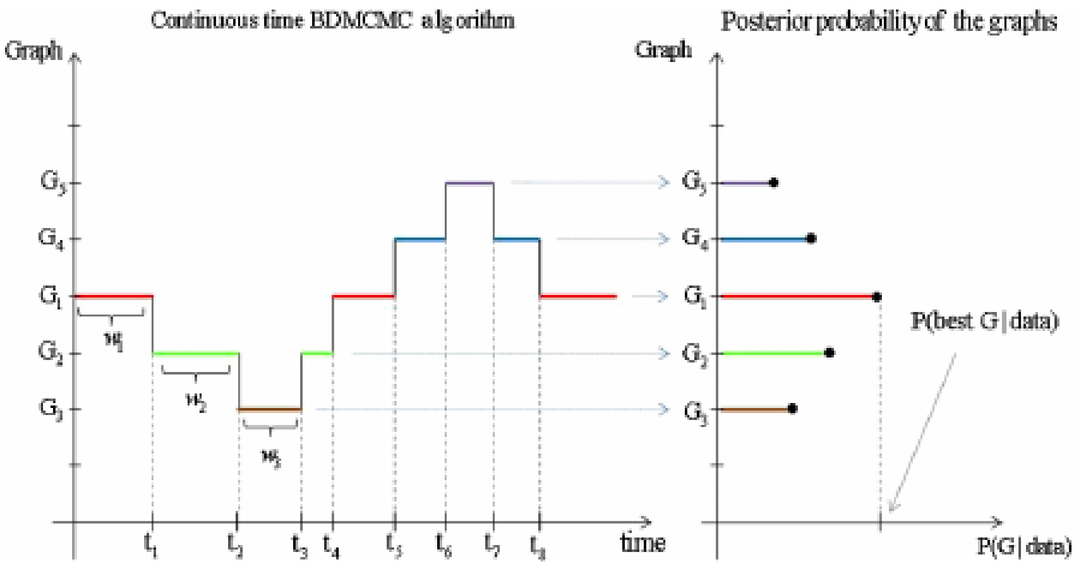

Figure 3.1:

3.5.2 Data collection

Gaussian Graphical models have the severe limitation, that sample covariance and correlation matrices are well conditioned only when the number of variables is smaller than the number of samples[49]. Therefore, it is absolutely necessary to perform a pre-selection of included entities when calculating GGMs on our multiomics data set.

We performed two mechanisms of selecting node sets. Both are anchored in trans-acting SNP-CpG interactions as well as a relationship to HERV elements.

The first way starts of with a SNP that lies within a HERV element and has significant associations with at least 5 CpG sites located on another chromosome. These CpGs are added to the analysis. To include genes, that are potentially associated with this SNP, all expression probes within a 250 kb flanking region up- and downstream of the sentinel SNP are included. Analogously to find genes that are likely to be affected by methylation changes, for each associated CpG all overlapping probes and the first up- and downstream probe each within a distance of less than 250 kb are added. To include transcription factors, whose binding might be affected by DNA methylation changes, all TFs that have a binding site within 100 bp of a CpG of interest are added. Finally, in an attempt to connect genes around the sentinel SNP with genes around the methylation sites or transcription factors, we used the String network described in section **??**. We calculated the shortest path from a representative SNP gene to any TF or CpG gene and included all genes on this path.

The schematic representation of the selection starting from an example SNP is shown in figure 3.2

The second approach of data collection, shown in figure 3.3, starts with an CpG-site within a HERV element, that is part of a trans-meQTL. First, if any other measured CpGs lie within the same HERV, they are added to the analysis. Then for each CpG of interest the most strongly associated SNP that lies on another chromosome is added. Finally

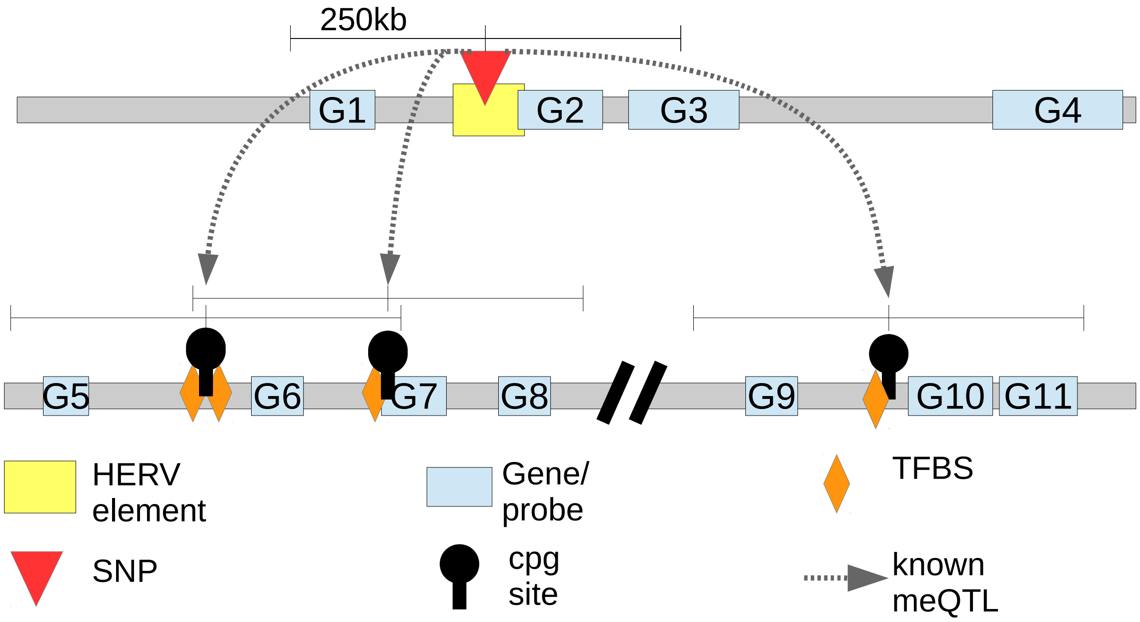

Figure 3.2:

neighboring genes, transcription factors and shortest path genes are included according to the same criteria as in the first approach.

3.5.3 BDgraph parameters

The BDgraph main method "bdgraph" was used to calculated the GGMs on the data sets that were just defined. I used the parameters, `'method = "gcgm"`, which leads to the use of the extension for mixed data. As algorithm I used the package default 'algorithm = "bdmcmc". The number of iterations was set to `'iter = 50000'`, while the burn-in period was `'burnin = 25000'`. "bdgraph" was run with a prior edge distribution of 'g.prior = 0.5'. The starting graph was set to `'g.start = "empty"'`, which means the calculation is started in an unconnected graph. Finally, the degrees of freedom for the G-Wishart distribution used to calculate the prior of the precision matrix was chosen as `'prior.df = 3'`.

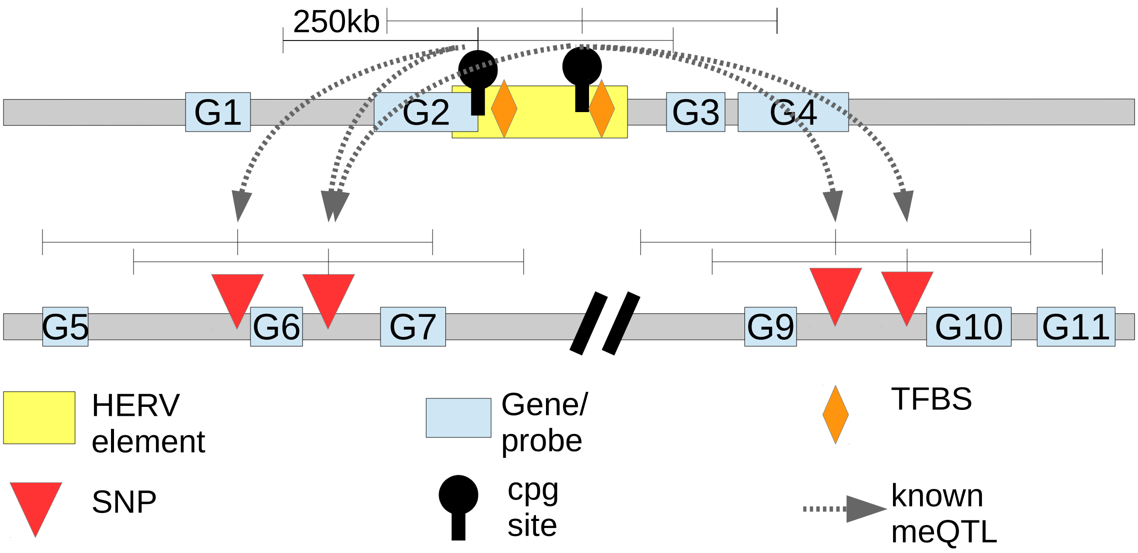

Figure 3.3:

Chapter 4

Results

4.1 Normalized Data

The distributions of the expression and methylation residuals, calculated as described in chapter 3.2, over all probes and samples can be seen in figure 4.1. As expected the residuals follow a normal distribution. This is important for the calculation of the Gaussian graphical models, as the normal distribution of the data is one of base assumptions for Gaussian graphical models[].

4.2 HERV region features

In this section I will describe the expression probes that overlap with and the cpgs and SNPs that lie within any HERV element and/or their flanking regions. The results for the set of all endogenous retroviral elements, HERV set 2, without flanking regions are described in detail. The results for the other sets defined in chapter 2.1 and including flanking regions will be shown in tables or in the supplementary data.

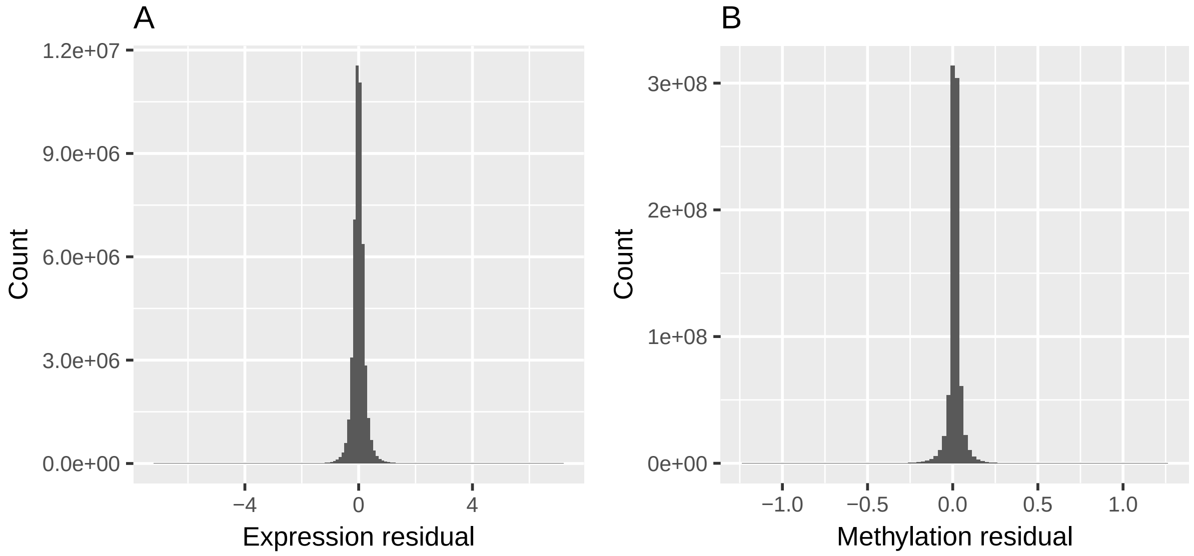

Figure 4.1: Distribution of expression (A) and methylation (B) residuals over all samples and probes

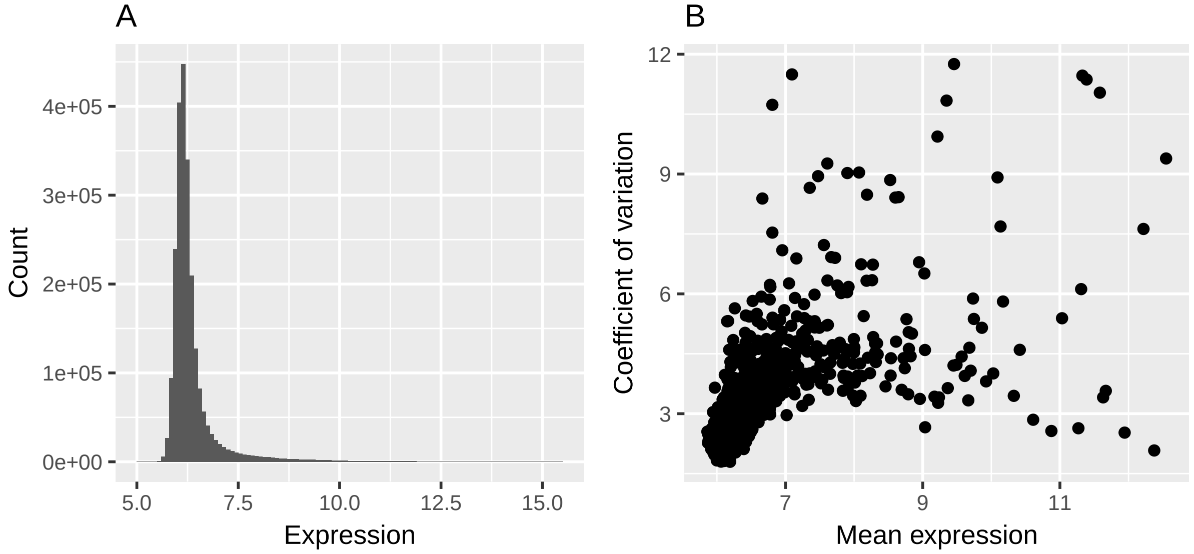

Figure 4.2: Distribution of expression values (A) and coefficient of expression variation (B) of 2343 expression probes overlapping with HERV S2 in 993 individuals.

4.2.1 Expression

A total of 2343 expression probes overlap directly with at least one of 2271 HERV elements from HERV S2. This means ca 4.50% of all expression probes overlap with HERV elements and expression measurements are known for parts of ca 0.37% of all HERV elements.

The distribution of expression values of the probes overlapping with HERV elements, as shown in figure 4.2A is almost identical to the distribution of all probes (figure 2.3A). The coefficient of variation (figure 4.2B), however, is lower on average for the considered subset compared to the whole set of probes (figure 2.3B).

Of these 2343 probes 510 were annotated to one of 449 different genes. I performed a GO enrichment for the biological process ontology on these genes with the set of all genes with available expression data as background. After correcting for multiple testing only the term defense response (GO:0006952, p-value=1.37e-5, fdr=0.047) was significantly enriched.

Term ID	Term	p	fdr
GO:0006952	defense response	$4.82 \cdot 10^{-11}$	$5.2 \cdot 10^{-7}$
GO:0045087	innate immune response	$5.5 \cdot 10^{-11}$	$5.94 \cdot 10^{-7}$
GO:0006955	immune response	$2.8 \cdot 10^{-9}$	$3.02 \cdot 10^{-5}$
GO:0098542	defense response to other organism	$1 \cdot 10^{-7}$	$1.08 \cdot 10^{-3}$
GO:0009615	response to virus	$3.2 \cdot 10^{-7}$	$3.45 \cdot 10^{-3}$
GO:0051607	defense response to virus	$3.45 \cdot 10^{-6}$	$3.73 \cdot 10^{-2}$

Table 4.1: Significantly enriched GO biological process terms among genes overlapping with HERV S2.

However, when including the 2kb flanking regions of HERV S2 and performing the GO enrichment on the 5518 identified genes, the six GO terms shown in table 4.1 are

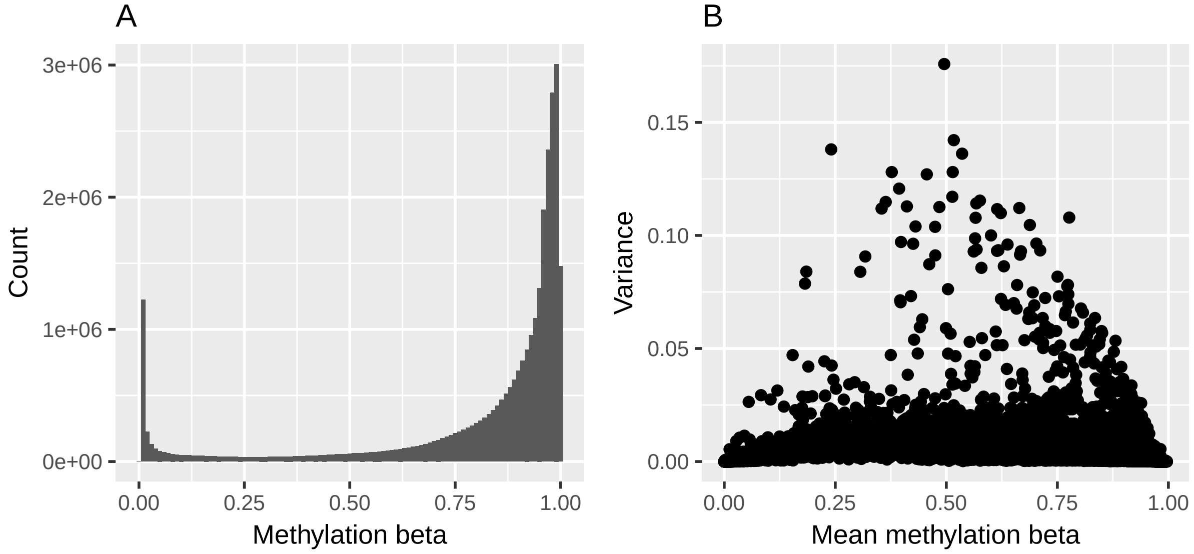

Figure 4.3: Distribution of methylation beta values (A) and coefficient of methylation variation (B) of 17077 GpG-sites located within HERV S2 in 1727 individuals.

significantly overrepresented. They are all connected to defense or immune response.

The results for all HERV sets and including flanking regions is shown in table 4.2.

Set	S1	S1.1kb	S1.2kb	S2	S2.1kb	S2.2kb	S3	S3.1kb	S3.2kb
Pairs	239	766	1,496	2,368	13,601	27,476	165	390	703
HERVs	219	706	1,336	2,271	12,551	24,275	146	350	615
Probes	239	563	970	2,343	9,364	15,466	165	302	487
Genes	21	131	293	449	2,968	5,517	15	69	138

Table 4.2: Overview of expression probes overlapping with different HERV sets and flanking regions. "Pairs" describes the total number of overlaps occurring, "HERVs" and "Probes" are the number of distinct HERV elements/expression probes that are part of at least one overlap.

4.2.2 Methylation

17077 CpG sites, equaling 3.52% of all measured CpGs, were found within HERV S2 and 12871 distinct HERV elements (2.10%) contain at least one interrogated CpG site.

CpGs associated with HERV S2, as shown in figure 4.3A, tend to be highly methylated. The variance also is (figure 4.3B) is on average lower than the one for all CpG-sites (figure 2.5B).

The number of CpG sites within all HERV sets and including flanking regions is shown in table 4.3

Set	S1	S1.1kb	S1.2kb	S2	S2.1kb	S2.2kb	S3	S3.1kb	S3.2kb
Pairs	1,587	6,105	12,005	17,077	$1.11 \cdot 10^5$	$2.39 \cdot 10^5$	1,152	3,465	6,408
HERVs	973	3,091	4,885	12,871	60,727	$1.01 \cdot 10^5$	614	1,561	2,425
CpGs	1,587	4,497	7,790	17,077	78,466	$1.39 \cdot 10^5$	1,152	2,671	4,404

Table 4.3: Overview of CpGs overlapping with different HERV sets and flanking regions. ”Pairs” describes the total number of overlaps occurring, ”HERVs” and ”CpGs” are the number of distinct HERV elements/GpG sites that are part of at least one overlap.

4.2.3 Genotypes

A total of 890780 the considered SNPs are located within elements of HERV S2. This constitutes 9.34% of all SNPs. These SNPs are found in 330744 distinct HERV elements. Therefore, 53.99% contain at least one SNP.

The results for all sets and flanking regions are shown in table 4.4

Set	S1	S1.1kb	S1.2kb	S2	S2.1kb	S2.2kb	S3	S3.1kb	S3.2kb
Pairs	$1.25 \cdot 10^5$	$3.84 \cdot 10^5$	$6.37 \cdot 10^5$	$8.91 \cdot 10^5$	$5.15 \cdot 10^6$	$9.35 \cdot 10^6$	89,066	$2.08 \cdot 10^5$	$3.21 \cdot 10^5$
HERVs	21,805	31,412	31,601	$3.31 \cdot 10^5$	$5.55 \cdot 10^5$	$5.59 \cdot 10^5$	10,139	13,125	13,189
SNPs	$1.25 \cdot 10^5$	$2.64 \cdot 10^5$	$3.72 \cdot 10^5$	$8.91 \cdot 10^5$	$3.31 \cdot 10^6$	$4.79 \cdot 10^6$	89,066	$1.54 \cdot 10^5$	$2.05 \cdot 10^5$

Table 4.4: Overview of SNPs overlapping with different HERV sets and flanking regions. ”Pairs” describes the total number of overlaps occurring, ”HERVs” and ”SNPs” are the number of distinct HERV elements/SNPs that are part of at least one overlap.

The distribution of number of SNPs found in each HERV element are shown in figure 4.4A. A total of 2423 HERV elements had more than 20 within its boundaries and were not shown in the figure. The maximum number of SNPs found in a single HERV element was 368. This HERV element, however, is unusually long at 7003 bp.

In figure 4.4B the density of SNPs per base pair of its HERV element are shown. 356 HERV elements that contained more than 0.05 SNPs per base pair were excluded from this figure.

4.2.4 Chromatin states

Data in /storage/groups/groups_epigenreg/users/julian.schmidt ...

4.3 eQTLs

Associations were calculated for 156.4 millions cis acting SNP-expression probe pairs and 456.1 billion possible pairs in trans.

812147 of cis-pairs were significantly associated with p-values of $1e-6$ or less. These significant eQTLs have a false discovery rate of less than $1.93e-4$. A total of 551728 distinct SNPs and 4903 expression probes were part of at least one cis-eQTL. 4145 of these probes are annotated to 3552 different genes. The remaining 758 could not be assigned to a specific gene.

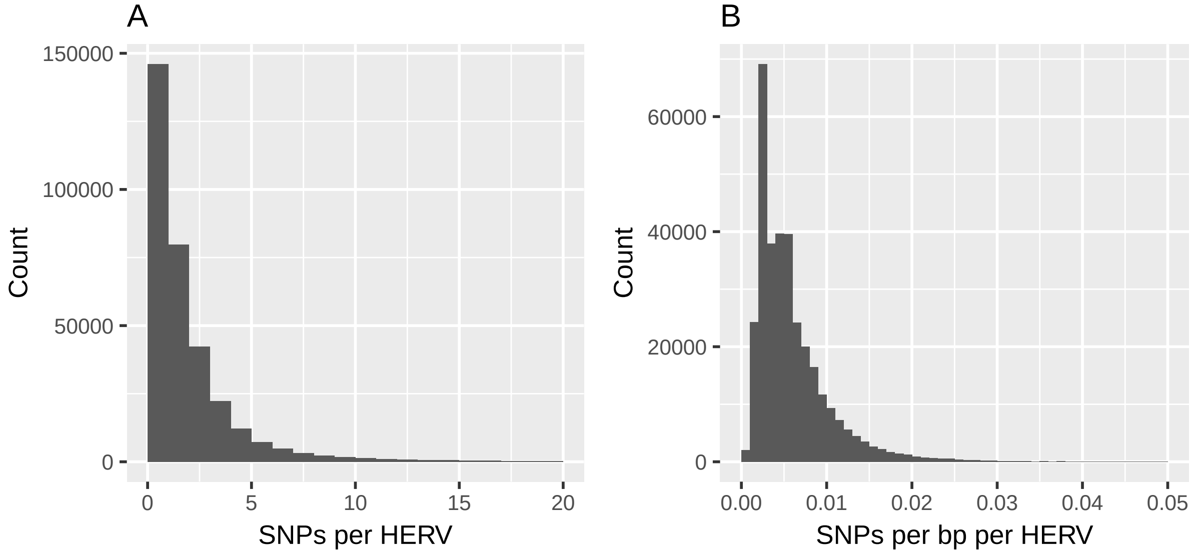

Figure 4.4:

There were a total of 1511235 significant trans-eQTL with an p-value of less than 1e-8 and a FDR of less than $3.02e-4$. These are made up by 229332 distinct SNPs and 21338 expression probes. 11505 of the probes found in at least one trans-eQTL are annotated to 9389 different genes, while 9767 probes are not assigned to a gene.

In a previous work Schramm et al. [50] calculated eQTLs on expression and genotype data from 890 KORA F4 samples. They found significant cis-eQTLs for 4116 probes and considered only the pair of the most strongly associated SNPs with the probe. Of these pairs 2680 were also found in my analysis.

Figure 4.5A shows the fraction of significant to all possible SNP-expression probe pairs, mapped to 10 Mpb windows. As expected the values at the diagonal are comparatively higher, containing all cis-eQTLs.

The expression residual signatures of wildtype, heterogenous and homogenous mutated SNP site for the two eQTLs with the best and worst p-values in cis and trans are shown in figure 4.6. The eQTLs with the best p-values in subfigures A and C show clear distinctions between the different SNP variants. While there is a small trend between variants in subfigures B and D, the distributions of expression residuals are rather similar.

45862 of the SNPs and 166 of the expression probes found in at least one cis-eQTL are located within HERV S2. When considering all significant associations, where either the SNP or the expression probe lie within a HERV element, a total of 112604 pairs remain. There are 5855 cis-eQTLs constituting one of 4748 SNPs within a HERV element controlling one of 128 expression probes, that overlap with a HERV element.

When limiting trans-eQTLs to the ones that contain a HERV S2 SNP (23396) and/or expression probe (1227), 240301 remain. In 9177 of these pairs both are HERV related.

The fraction of HERV S2 related SNP-expression probe pairs that proved to be significantly associated is shown in figure 4.5B. The distribution of this subset of eQTLs is very similar to the whole set. At least at this resolution there does not seem to be a difference between HERV regions and the whole genome when it comes to the probability of a SNP or expression probe to be part of an eQTL.

A GO enrichment was performed on all 8211 genes associated to HERV related eQTLs

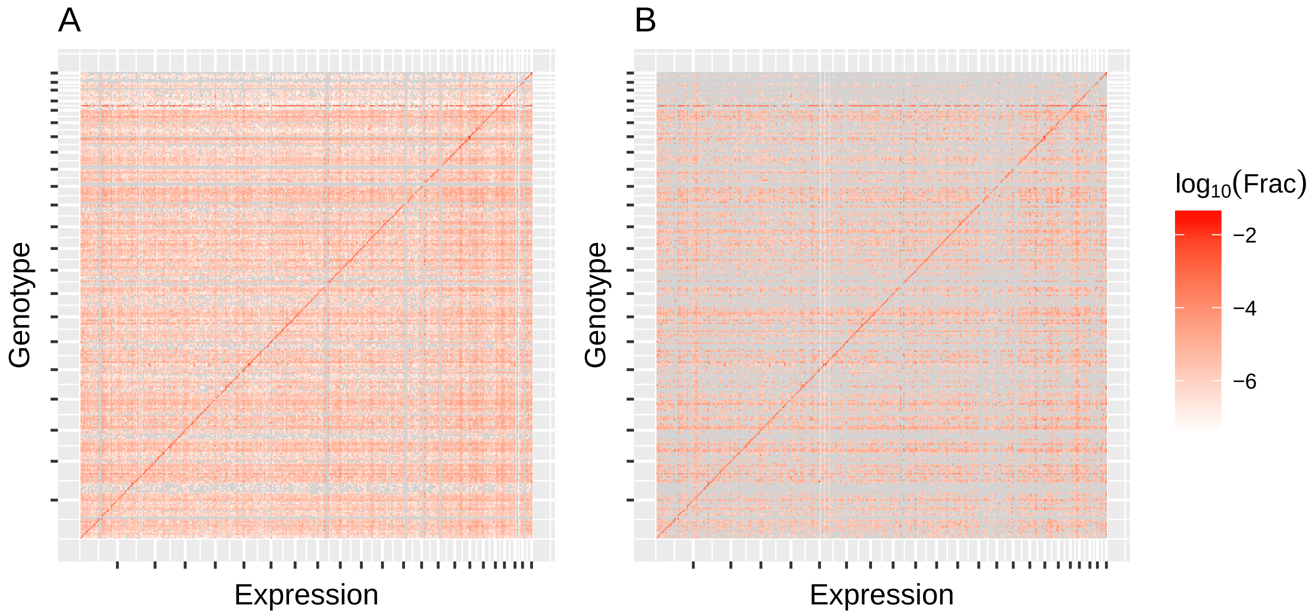

Figure 4.5: Fraction of SNP-expression probe pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTMs in the given pair of bins. A considers all eQTLs an pairs, while B is limited on the ones whose SNP and/or expression probes are connected to HERV set 2.

using the set of all genes found in eQTLs as background. However, there were no significantly enriched genes.

4.4 eQTMs

When calculating expression quantitative trait methylation there were a total of 13.88 millions CpG-expression probe pairs within a distance of 50 Kpb or less of each other. The number of potential trans-acting pairs equaled around 23.22 billions.

Calculating eQTMs with a significance threshold of 10e-6 for cis resulted in 8187 significant associations ($fdr < 1.7e-3$) consisting of 5957 distinct CpG sites and 1959 different expression probes. 1658 of these probes are annotated to 1461 genes, while there are no gene annotations for the remainder.

361485 of the potential trans-action CpG-expression probe pairs proved to be significant at a p-value threshold of 10e-8 ($fdr < 6.5e-4$). These trans-eQTLs are made up by 48206 CpGs and 11673 expression probes, for which 5738 gene annotations are available.

The fractions of potential CpG-expression probe pairs that were significantly associated between their genomic locations are shown in figure 4.7A. In contrast to the eQTL results there is only a weaker preference for cis interactions.

The pairs of expression and methylation residual values for the eQTMs with the best and worst p-values for cis and trans each, are shown in figure 4.8.

HERV S2 contains 311 CpG sites and 420 expression probes, that were present in cis-eQTMs. A total of 738 cis-eQTMs are related to the set. 33 of these are associations between one of 33 CpG sites and one of 14 expression probes that are both within a HERV

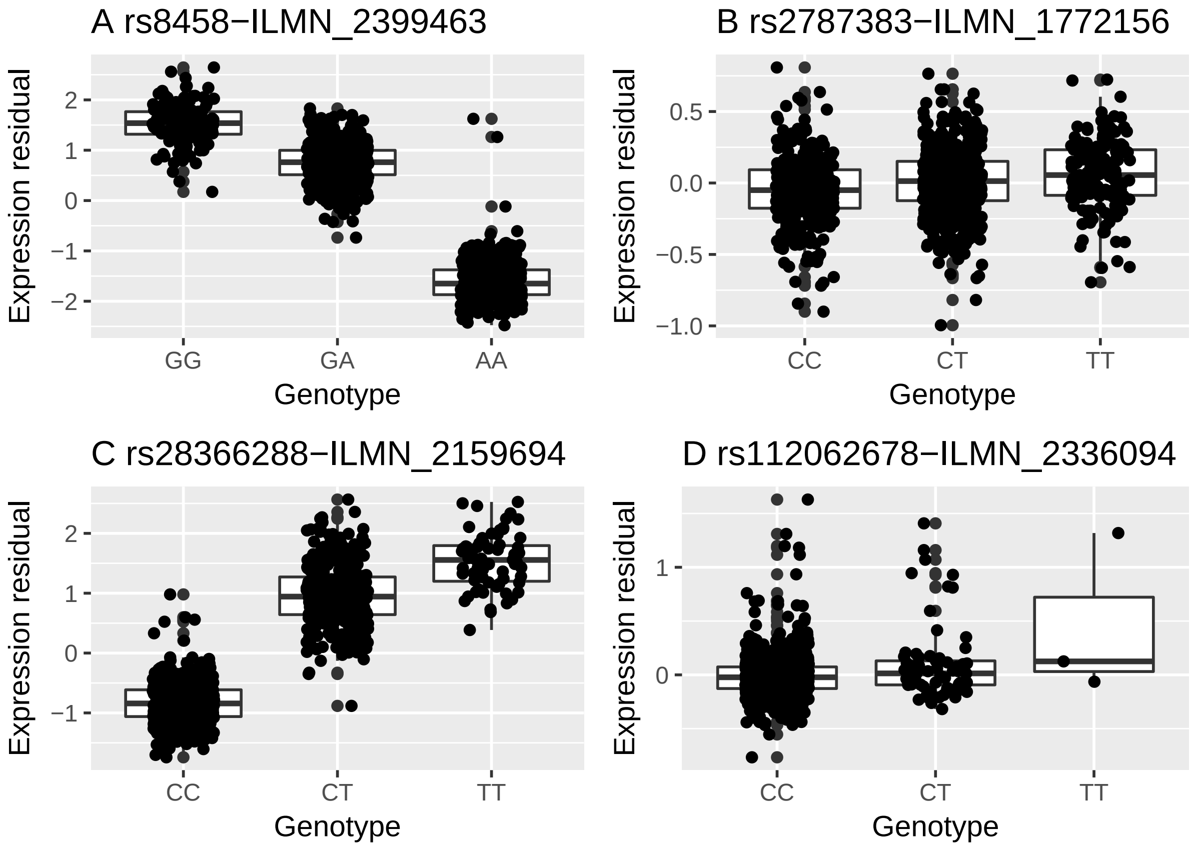

Figure 4.6:

element.

Considering trans-eQTMs, 1109 significantly associated CpGs lie within HERV S2 and 5422 expression probes overlapping a HERV element are part of a trans-eQTM. In total 15085 trans-eQTMs are related to HERV S2.

A GO biological process enrichment was performed on the 1316 genes that the expression probes found in HERV S2 related eQTMs are annotated to. All gene that were part of an eQTM were used as background. After correction for multiple testing no GO terms were found to be significantly enriched.

4.5 meQTLs

The number of cis- and trans-meQTLs in the obtained data set was already elaborated on in chapter 2.2.5. Therefore, in this chapter the focus is laid on the meQTLs that are related to HERV S2.

There is a total of 360048 meQTLs, whose CpG-site lies within a HERV element. They are made up by different 1915 CpGs and 286289 SNPs. This means each of these CpG-sites is associated with 188 SNPs on average, which is slightly higher than the average of 157.91 for CpGs in the total set of meQTLs. The CpG, that is part of most meQTLs, cg07143125, is associated with 579 SNPs.

When considering meQTLs, whose SNP lies within HERV S2, there are 983664 significant CpG-SNP pairs. These consist of 229819 different SNPs and 52951 CpGs. On

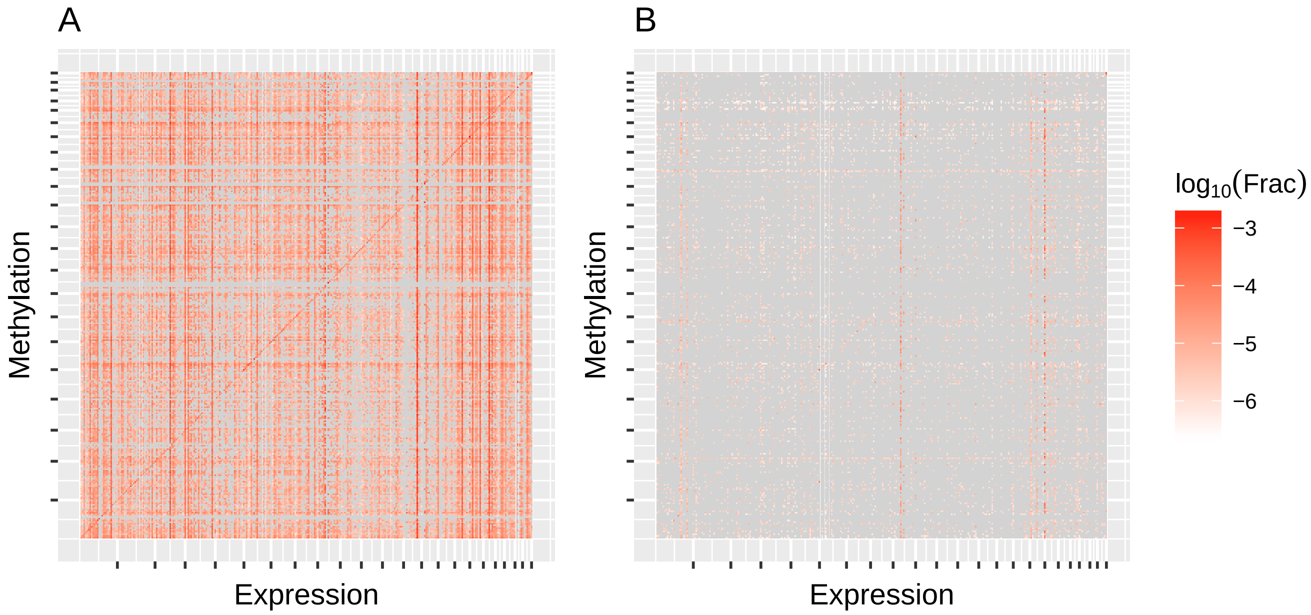

Figure 4.7: Fraction of CpG-expression probe pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTMs in the given pair of bins. A considers all eQTMs and pairs, while B is limited on the ones whose CpGs and/or expression probes are connected to HERV set 2.

average each of these SNPs is associated with 4.28 CpG-sites.

In total there are 1301127 meQTLs, that are related to HERV S2. They are made up of 53167 different CpGs and 483019 SNPs

The fractions of potential SNP-CpG pairs that were significantly associated between their genomic locations are shown in figure 4.9A. Subfigure B shows the same analysis for the subset of meQTLs that were just described. The Fractions of significantly enriched pairs are very similar. The high values along the diagonal show, that there is a strong preference for cis-interactions in both sets.

When limiting the analysis to meQTLs, that represent associations between different chromosomes, which are called trans-meQTLs, a total of 64852 significant pairs between 2681 CpGs and 34957 SNPs remain. Of these meQTLs, 23463 have their CpGs within HERV S2. They consists of 155 different CpGs and 18825 SNPs. Analogously for 45175 pairs their 19192 SNPs are located within HERV S2 and associated with 2655 different CpGs.

4.6 HERV related regulatory networks

4.6.1 Data collection

4.6.2 General analysis

4.6.3 Detailed analysis

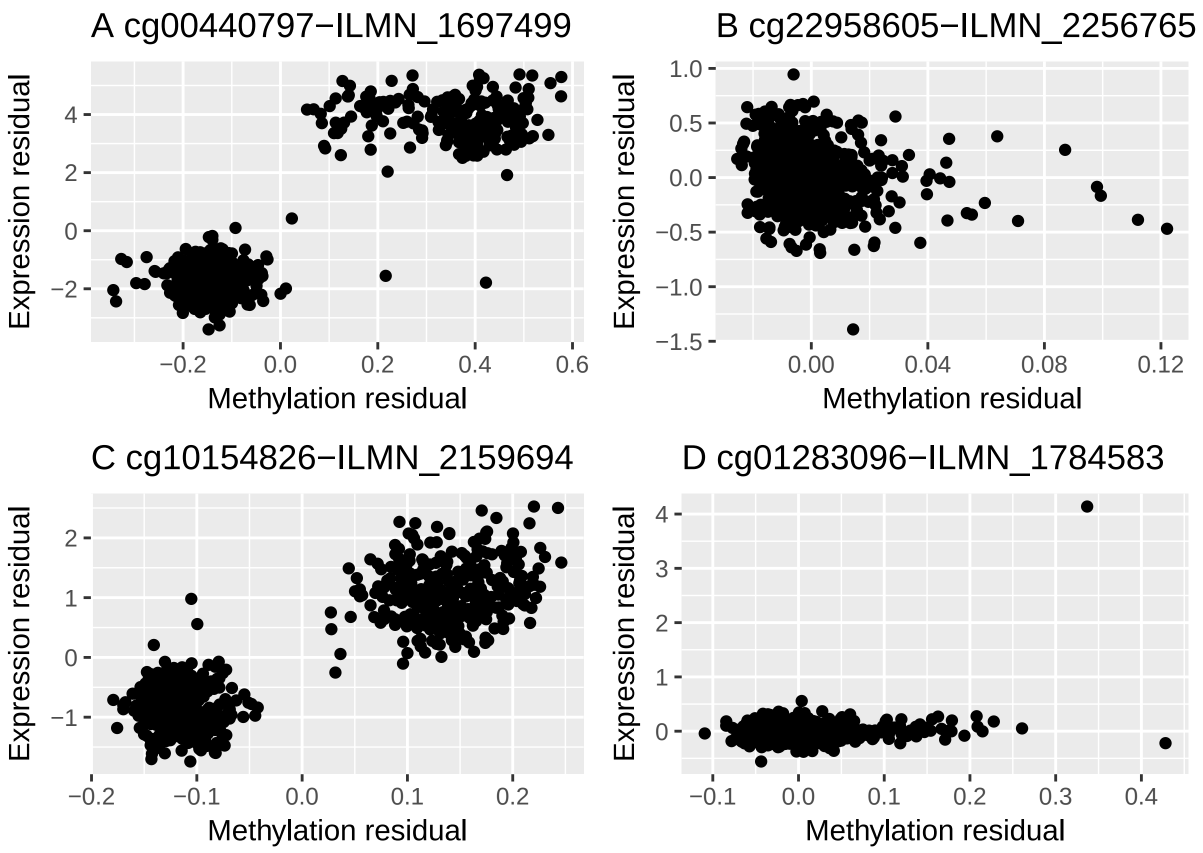

Figure 4.8:

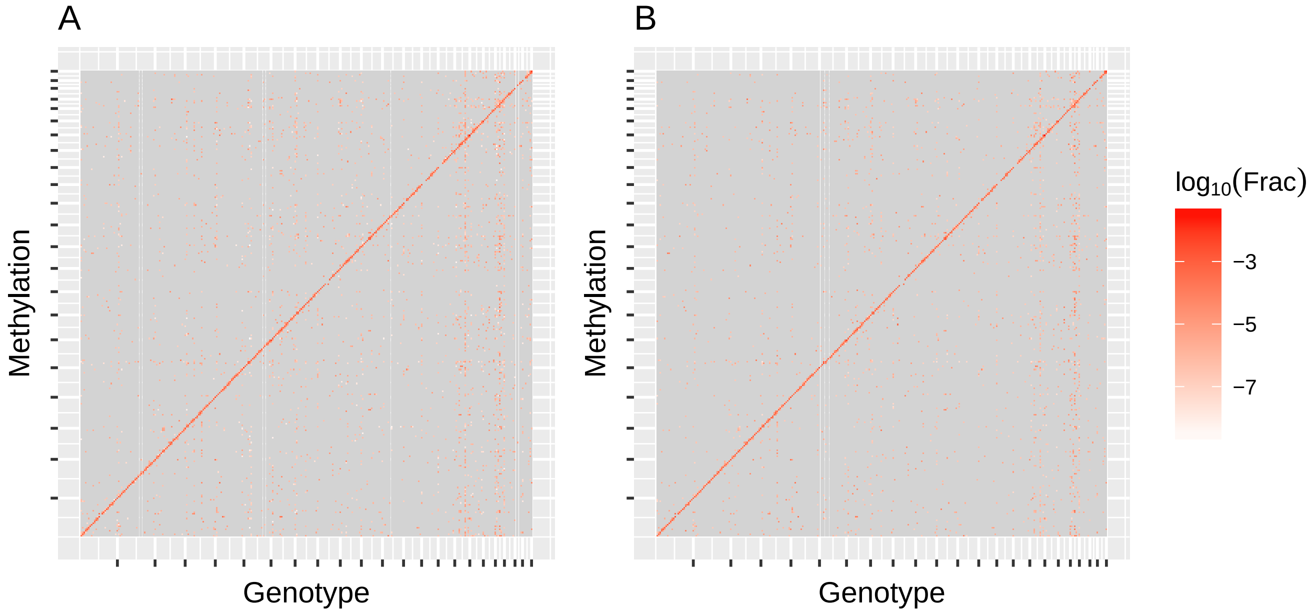

Figure 4.9: Fraction of SNP-CpG pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTMs in the given pair of bins. A considers all meQTLs and pairs, while B is limited on the ones whose CpGs and/or SNPs are related to HERV set 2.

Chapter 5

Discussion

5.1 HERV region features

5.1.1 Expression

- less probes than expected $->$ chip design

5.1.2 Methylation

- less data available than expected? chip design or less CpG-sites?
- mostly methylated $->$ inactivated LTRs

5.1.3 Genotypes

5.2 eQTLs

5.3 eQTMs

5.4 meQTLs

5.5 Regulatory networks

Chapter 6

Conclusion and Outlook

List of Figures

1.1 Exemplified structure of an intact HERV element. Based on a HERV-K provirus, which is considered the most recently introduced in humans and contains the most complete viral genes. Taken from Young et. al. [15] . . . 3

2.1 Length distribution of elements in HERV S2 6
2.2 Number of samples with genotype, expression and methylation measurements in KORA F4 Survey 7
2.3 Distribution of expression values (A) and coefficient of expression variation (B) of 47864 probes in 993 individuals. 8
2.4 Histograms of proportions of missing values per CpG-site (A) and sample (B). 9
2.5 Distribution of all methylation beta values (A) and variance (B) of 485577 CpGs between 1727 individuals. 9

3.1 17
3.2 18
3.3 19

4.1 Distribution of expression (A) and methylation (B) residuals over all samples and probes 21
4.2 Distribution of expression values (A) and coefficient of expression variation (B) of 2343 expression probes overlapping with HERV S2 in 993 individuals. 22
4.3 Distribution of methylation beta values (A) and coefficient of methylation variation (B) of 17077 GpG-sites located within HERV S2 in 1727 individuals. 23
4.4 25
4.5 Fraction of SNP-expression probe pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTMs in the given pair of bins. A considers all eQTLs an pairs, while B is limited on the ones whose SNP and/or expression probes are connected to HERV set 2. 26
4.6 27
4.7 Fraction of CpG-expression probe pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTMs in the given pair of bins. A considers all eQTMs and pairs, while B is limited on the ones whose CpGs and/or expression probes are connected to HERV set 2. 28
4.8 29

4.9 Fraction of SNP-CpG pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTMs in the given pair of bins. A considers all meQTLs and pairs, while B is limited on the ones whose CpGs and/or SNPs are related to HERV set 2. 29

List of Tables

2.1 First ten rows of the RepeatMasker annotation on hg19 5
2.2 Covariate overview 10

4.1 Significantly enriched GO biological process terms among genes overlapping with HERV S2. 22
4.2 Overview of expression probes overlapping with different HERV sets and flanking regions. "Pairs" describes the total number of overlaps occurring, "HERVs" and "Probes" are the number of distinct HERV elements/expression probes that are part of at least one overlap. 23
4.3 Overview of CpGs overlapping with different HERV sets and flanking regions. "Pairs" describes the total number of overlaps occurring, "HERVs" and "CpGs" are the number of distinct HERV elements/GpG sites that are part of at least one overlap. 24
4.4 Overview of SNPs overlapping with different HERV sets and flanking regions. "Pairs" describes the total number of overlaps occurring, "HERVs" and "SNPs" are the number of distinct HERV elements/SNPs that are part of at least one overlap. 24

Bibliography

[1] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, May 2017.

[2] Alexander F. Palazzo and Eliza S. Lee. Non-coding rna: what is functional and what is junk? *Frontiers in Genetics*, 6:2, 2015.

[3] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009. 19015660[pmid].

[4] Almut Schulze and Julian Downward. Navigating gene expression using microarrays – a technology review. *Nature Cell Biology*, 3:E190 EP –, Aug 2001.

[5] Andrea Piunti and Ali Shilatifard. Epigenetic balance of gene expression by polycomb and compass families. *Science*, 352(6290), 2016.

[6] Aimée M. Deaton and Adrian Bird. Cpg islands and the regulation of transcription. *Genes Dev*, 25(10):1010–1022, May 2011. 21576262[pmid].

[7] Zachary D. Smith and Alexander Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14:204 EP –, Feb 2013. Review Article.

[8] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui

Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigó, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[9] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860 EP –, Feb 2001.

[10] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57 EP –, Sep 2012. Article.

[11] Elizabeth Pennisi. Encode project writes eulogy for junk dna. *Science*, 337(6099):1159–1161, 2012.

[12] Todd J. Treangen and Steven L. Salzberg. Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13:36 EP –, Nov 2011. Review Article.

[13] Robin A. Weiss. Human endogenous retroviruses: friend or foe? *APMIS*, 124(1-2):4–10, 2016.

[14] Norbert Bannert and Reinhard Kurth. The evolutionary dynamics of human endogenous retroviral families. *Annual Review of Genomics and Human Genetics*, 7(1):149–173, 2006. PMID: 16722807.

[15] George R. Young, Jonathan P. Stoye, and George Kassiotis. Are human endogenous retroviruses pathogenic? an approach to testing the hypothesis. *BioEssays*, 35(9):794–803, 2013.

[16] H. Yu, Z. Zhao, and F. Zhu. The role of human endogenous retroviral long terminal repeat sequences in human cancer (review). *International Journal of Molecular Medicine*, 32(4):755–762, 2013.

[17] Gorjan Slokar and Gregor Hasler. Human endogenous retroviruses as pathogenic factors in the development of schizophrenia. 6, 01 2016.

[18] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5:21–21, Jan 2011. 1752-0509-5-21[PII].

[19] R Smit AFA, Hubley and Green P. Repeatmasker open-4.0. 2013-2015.

[20] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.

[21] Lindsay L. Waite, Benjamin Weaver, Kenneth Day, Xinrui Li, Kevin Roberts, Andrew W. Gibson, Jeffrey C. Edberg, Robert P. Kimberly, Devin M. Absher, and Hemant K. Tiwari. Estimation of cell-type composition including t and b cell subtypes for whole blood methylation microarray data. *Frontiers in Genetics*, 7:23, 2016.

[22] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, 5(6):1–15, 06 2009.

[23] Andreas Schroeder, Odilo Mueller, Susanne Stocker, Ruediger Salowsky, Michael Leiber, Marcus Gassmann, Samar Lightfoot, Wolfram Menzel, Martin Granzow, and Thomas Ragg. The rin: an rna integrity number for assigning integrity values to rna measurements. *BMC Molecular Biology*, 7(1):3, Jan 2006.

[24] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K. Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, Kazuhiro R. Nitta, Minna Taipale, Alexander Popov, Paul A. Ginno, Silvia Domcke, Jian Yan, Dirk Schübeler, Charles Vinson, and Jussi Taipale. Impact of cytosine methylation on dna binding specificities of human transcription factors. *Science*, 356(6337), 2017.

[25] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J.

Rando, Ewan Birney, Richard M. Myers, William S. Noble, Michael Snyder, and Zhiping Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*, 22(9):1798–1812, Sep 2012.

[26] Philip Machanick and Timothy L. Bailey. Meme-chip: motif analysis of large dna datasets. *Bioinformatics*, 27(12):1696–1697, 2011.

[27] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[28] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne ChÃ¨neby, Shubhada R Kulkarni, Ge Tan, Damir Baranasic, David J Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, BenoÃ®t Ballester, Wyeth W Wasserman, FranÃ§ois Parcy, and Anthony Mathelier. Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D260–D266, 2018.

[29] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. Transfac(?) and its module transcompel(?): transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, Jan 2006. 16381825[pmid].

[30] Aurelien Griffon, Quentin Barbier, Jordi Dalino, Jacques van Helden, Salvatore Spicuglia, and Benoit Ballester. Integrative analysis of public chip-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Research*, 43(4):e27, 2015.

[31] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357 EP –, Mar 2012.

[32] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137–R137, Sep 2008. gb-2008-9-9-r137[PII].

[33] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J. Jensen. String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, 41(Database issue):D808–D815, Jan 2013. gks1094[PII].

[34] Damian Szklarczyk, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von?Mering. The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*, 45(Database issue):D362–D368, Jan 2017. 27924014[pmid].

[35] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45:580 EP –, May 2013.

[36] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature Methods*, 9:215 EP –, Feb 2012. Correspondence.

[37] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017.

[38] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013.

[39] Mark Dunning, Andy Lynch, and Matthew Eldridge. *illuminaHumanv3.db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3)*, 2015. R package version 1.26.0.

[40] Tim Triche, Jr. *FDb.InfiniumMethylation.hg19: Annotation package for Illumina Infinium DNA methylation probes*, 2014. R package version 2.2.0.

[41] Andrey A. Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[42] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[43] HervÃ© PagÃ¨s, Marc Carlson, Seth Falcon, and Nianhua Li. *AnnotationDbi: Annotation Database Interface*, 2017. R package version 1.38.1.

[44] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

[45] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[46] Abdolreza Mohammadi and Ernst C Wit. Bdgraph: An r package for bayesian structure learning in graphical models. 2015.

[47] Abdolreza Mohammadi and Ernst C. Wit. Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

[48] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664.

[49] Juliane SchÃ¤fer and Korbinian Strimmer. Learning large-scale graphical gaussian models from genomic data. In *In Science of Complex Networks: From Biology to the Internet and WWW*, 2005.

[50] Katharina Schramm, Carola Marzi, Claudia Schurmann, Maren Carstensen, Eva Reinmaa, Reiner Biffar, Gertrud Eckstein, Christian Gieger, Hans-JÃ¶rgen Grabe, Georg Homuth, Gabriele KastenmÃ¼ller, Reedik MÃ¤gi, Andres Metspalu, Evelin Mihailov, Annette Peters, Astrid Petersmann, Michael Roden, Konstantin Strauch, Karsten Suhre, Alexander Teumer, Uwe VÃ¶lker, Henry VÃ¶lzke, Rui Wang-Sattler, Melanie Waldenberger, Thomas Meitinger, Thomas Illig, Christian Herder, Harald Grallert, and Holger Prokisch. Mapping the genetic architecture of gene regulation in whole blood. *PLOS ONE*, 9(4):1–13, 04 2014.