

## 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

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Human endogenous retroviruses . . . . .	1
1.2	Multimomics . . . . .	2
1.3	Effect network analysis . . . . .	3
1.4	Objectives . . . . .	3
<b>2</b>	<b>Data</b>	<b>5</b>
2.1	HERV annotation . . . . .	5
2.2	KORA . . . . .	6
2.2.1	Expression . . . . .	7
2.2.2	Methylation . . . . .	7
2.2.3	Genotypes . . . . .	9
2.2.4	Covariates . . . . .	9
2.2.5	Methylation quantitative trait loci . . . . .	10
2.3	Transcription factor binding . . . . .	10
2.4	Protein interaction network . . . . .	11
<b>3</b>	<b>Methods</b>	<b>13</b>
3.1	Overlaps . . . . .	13
3.2	Data normalization . . . . .	13
3.3	eQTL/eQTM calculation . . . . .	13
3.4	Functional Analysis of Gene Sets . . . . .	14
3.5	Gaussian Graphical Models . . . . .	14
3.5.1	Methodological Background . . . . .	15
3.5.2	Data collection . . . . .	17
3.5.3	GGM calculation and filtering . . . . .	18
<b>4</b>	<b>Results</b>	<b>21</b>
4.1	Normalized Data . . . . .	21
4.2	HERV region features . . . . .	21
4.2.1	Expression . . . . .	21
4.2.2	Methylation . . . . .	22
4.2.3	Genotypes . . . . .	24
4.3	eQTLs . . . . .	25
4.4	eQTMs . . . . .	27
4.5	meQTLs . . . . .	28
4.6	HERV related regulatory networks . . . . .	29
4.6.1	Data collection . . . . .	30
4.6.2	General network overview . . . . .	31
4.6.3	Example networks . . . . .	32
<b>5</b>	<b>Discussion</b>	<b>37</b>
5.1	HERV region features . . . . .	37
5.2	Quantitative trait loci . . . . .	37
5.3	Regulatory networks . . . . .	38



# 1 Introduction

## 1.1 Human endogenous retroviruses

When the draft for the first human genome was published in 2001[1] 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[2] and later on the amount of exonic protein coding DNA was estimated to be as low as 1.2%[3]. The remaining majority of the genome was often labeled as non-functional "junk DNA"[4].

Since then this notion has generally been revoked. Especially within the two phases of the Encyclopedia of DNA Elements (ENCODE) project[3], 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[5]. In human repetitive DNA makes up about 50% of the whole Genome[5]. 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[6]. HERVs originate from retroviruses that infected the germ line of humans or human ancestors up to 30 million years ago[7]. 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[7]. 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 deletions 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[7]. Additionally, the promoter activity of LTRs is commonly silenced by hypermethylation[9].

Regardless of their loss of capability to replicate there still are multiple ways in which HERVs can influence human cells. LTRs of HERVs can act as promoters or enhancers

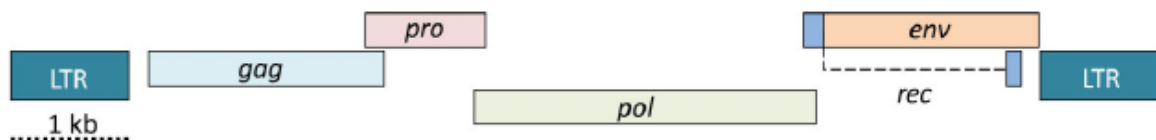


Figure 1.1: Exemplified structure of an intact HERV element. Based on a HERV-K provirus, which is considered the most recently introduced in humans and contains the most complete viral genes. Taken from Young et. al. [8]

to regulate gene expression. This can lead to disease and has been associated to various cancers[10]. However, in other known the promoter activity has been recruited and adds to transcription regulation in a positive way[11]. 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[12]. Finally, non-coding HERV RNAs have been found to play a role in control of stem cell properties[6].

## 1.2 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 integration of the analysis of different types of cellular molecules, is made possible by that. The objective of multiomics studies is to draw conclusions about cellular mechanisms and understand the flow of information by finding associations and patterns connecting them[13]. The objective of this work was to use data from the fields of transcriptomics, genomics and epigenomics to investigate the effects of variation in these three fields within human endogenous retroviruses.

Transcriptomics examines the expression 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[14]. With the late progress in transcriptome studies, even more effects of RNA have been discovered in the form of several functional RNAs[15]. RNA levels can be measured using a technology called RNA-seq[16] or probe-based arrays[17], as were used in this work.

The second field, genomics, describes variation between genomes. In this work I focused on single nucleotide polymorphisms (SNPs), which are single base changes. Such genotype information fills a special place in multiomics analyses, as it is not dependent on the environment and can be considered static in individuals. Therefore, it is considered as the cause of changes instead of being a downstream effect of other changes[13]. This means SNPs are well suited to be used as anchors 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 by, when occurring in enhancer or promoter regions of genes, which in turn is thought to be the cause of common diseases[13].

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[18, 19] 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[13].

DNA methylation, the epigenetic modification that is the focus in this work, is the best understood and researched epigenetic mark[9]. 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[9]. 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 (CGIs) and occur mainly around the transcription start sites of genes[19]. The methylation of these CpG islands is strongly connected to the silencing of adjacent genes, by preventing the transcription initiation[19]. The mechanisms this is realized by include the reduction of general promoter, enhancer and insulator activity as well as transcription factor binding[9].

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, called eQTL, pQTL and meQTL. These approaches allow to identify associated entities by calculating correlations between single entities over available individuals or samples. However, purely correlation based methods can not differentiate between direct and indirect effects[13].

Multomics data can used to construct 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[20].

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 after removing the common effects of all other variables. A scheme of how partial correlations work is shown in figure ??.

A method that uses partial correlation coefficients to construct networks is called "Gaussian Graphical Model"[20] (GGM).

The approach to Gaussian Graphical modeling used in this work utilizes Bayesian structure learning and is described in more detail in section 3.5.

### 1.4 Objectives

This work had three main objectives. The first objective was the generation of a comprehensive annotation of multiomics features in HERV elements. This allows to assess which expression, genotype and methylation data is available within HERVs, as well as identifying common patterns among these features, like the common association with certain pathways, or phenotypes. Furthermore, this annotation was used to filter later genome wide analyses.

The second objective was the integration of the multiomics data set by pairwise quantitative loci calculation. This was done by calculating eQTLs and eQTLs, and using an already existing dataset of meQTLs. The pairwise analysis allows the exploration of which cellular mechanisms are influenced by changes within HERV elements.



However, these associations can't differentiate between direct and indirect effects. Therefore, to elucidate the underlying mechanisms and pathways of the effects of changes within HERVs and integrate all omics, I generated networks representing direct interactions. This was done using Graphical Gaussian models, that are based on the concept of partial correlations.

## 2 Data

### 2.1 HERV annotation

As mentioned in section 1.1 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[21] 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<sup>a</sup>. 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[22] 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.

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

<sup>a</sup><http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz>

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 · 10 <sup>8</sup>	+	(CCCTAA)n	Simple_repeat	Simple_repeat	1	463	0	1
585	3,612	114	270	13	chr1	10,468	11,447	-2.49 · 10 <sup>8</sup>	-	TAR1	Satellite	telo	-399	1,712	483	2
585	437	235	186	35	chr1	11,503	11,675	-2.49 · 10 <sup>8</sup>	-	L1MC	LINE	L1	-2,236	5,646	5,449	3
585	239	294	19	10	chr1	11,677	11,780	-2.49 · 10 <sup>8</sup>	-	MER5B	DNA	hAT-Charlie	-74	104	1	4
585	318	230	38	0	chr1	15,264	15,355	-2.49 · 10 <sup>8</sup>	-	MIR3	SINE	MIR	-119	143	49	5
585	203	162	0	0	chr1	16,712	16,749	-2.49 · 10 <sup>8</sup>	+	(TGG)n	Simple_repeat	Simple_repeat	1	37	0	6
585	239	338	148	0	chr1	18,906	19,048	-2.49 · 10 <sup>8</sup>	+	L2a	LINE	L2	2,942	3,104	-322	7
585	652	346	85	42	chr1	19,947	20,405	-2.49 · 10 <sup>8</sup>	+	L3	LINE	CR1	3,042	3,519	-970	8
585	270	331	7	27	chr1	20,530	20,679	-2.49 · 10 <sup>8</sup>	+	Plat_L3	LINE	CR1	2,802	2,947	-639	9
585	254	279	47	39	chr1	21,948	22,075	-2.49 · 10 <sup>8</sup>	+	MLT1K	LTR	ERV1-MaLR	15	142	-453	1

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

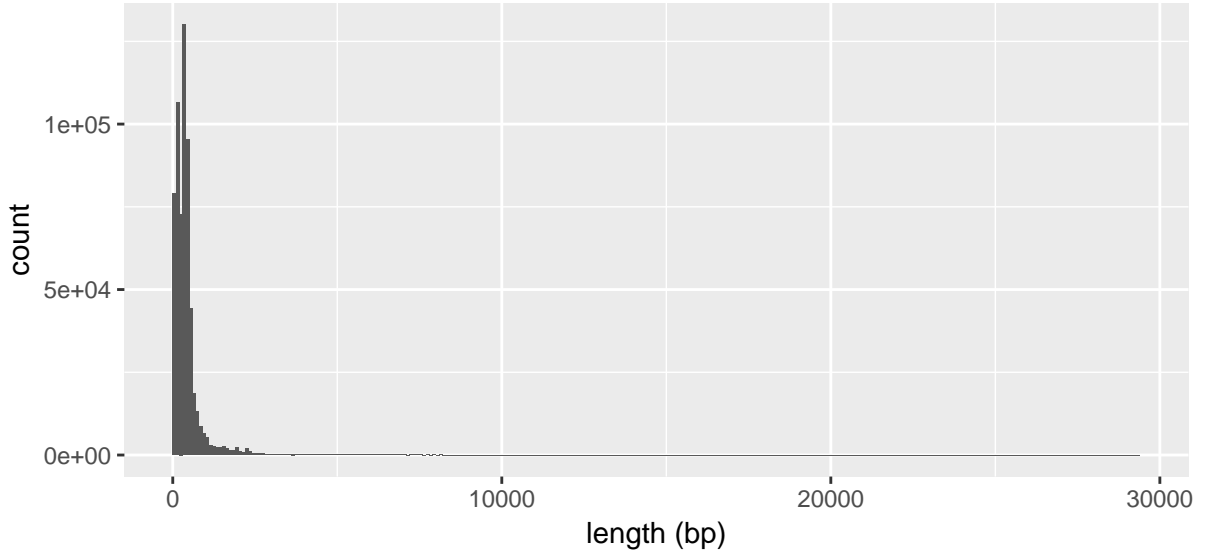


Figure 2.1: Length distribution of elements in HERV S2

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 over a long period in order to identify and examine the causes, effects and development of chronic 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[23] 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

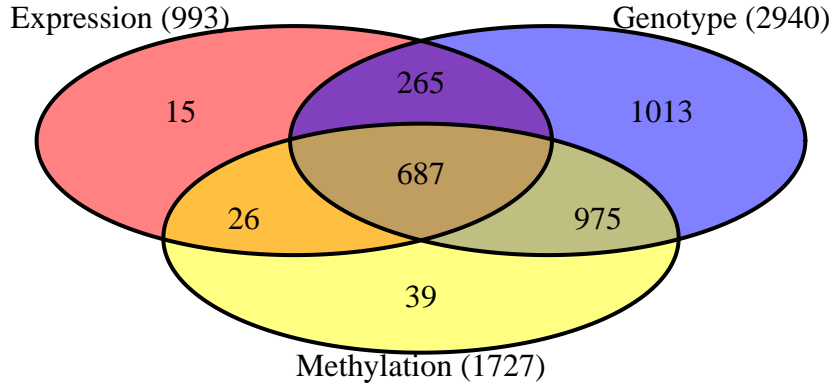


Figure 2.2: Number of samples with genotype, expression and methylation measurements in KORA F4 Survey

on varying sets of individuals according to availability of the required data types. A diagram of the samples available for each assay can be seen in figure 2.2. The joint analysis, using Gaussian Graphical Models, was performed on 687 individuals that had all three assays 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. They 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.

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

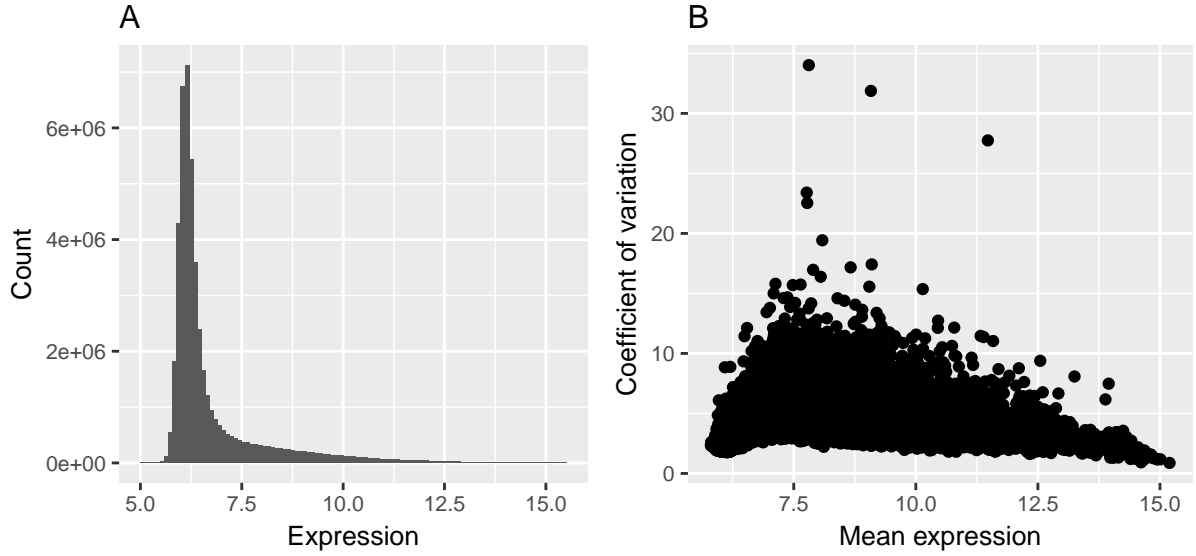


Figure 2.3: Distribution of expression values (A) and coefficient of expression variation (B) of 47864 probes in 993 individuals.

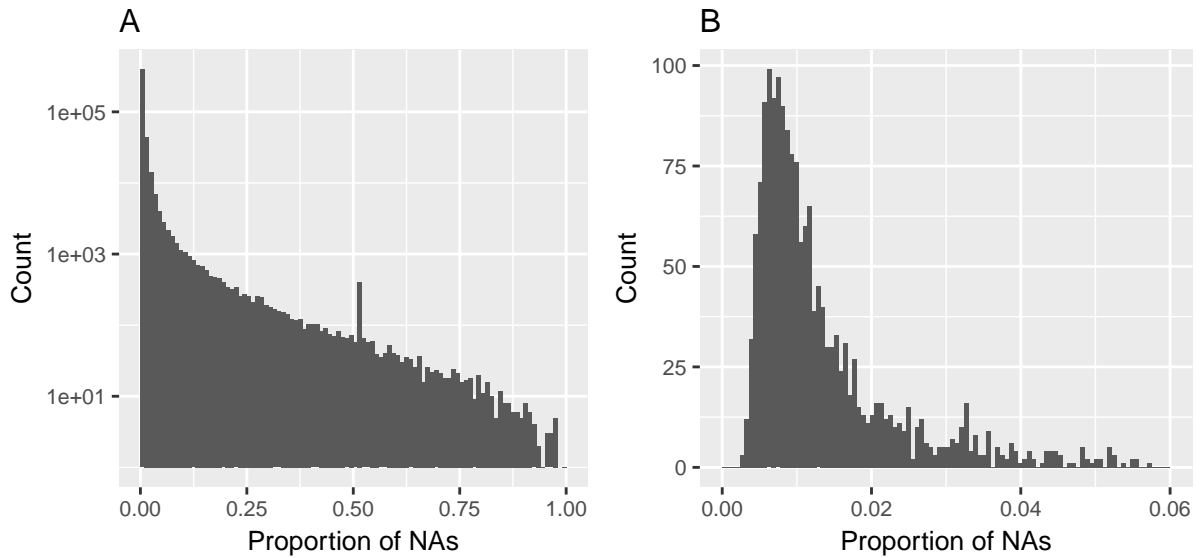


Figure 2.4: Histograms of proportions of missing values per CpG-site (A) and sample (B).

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

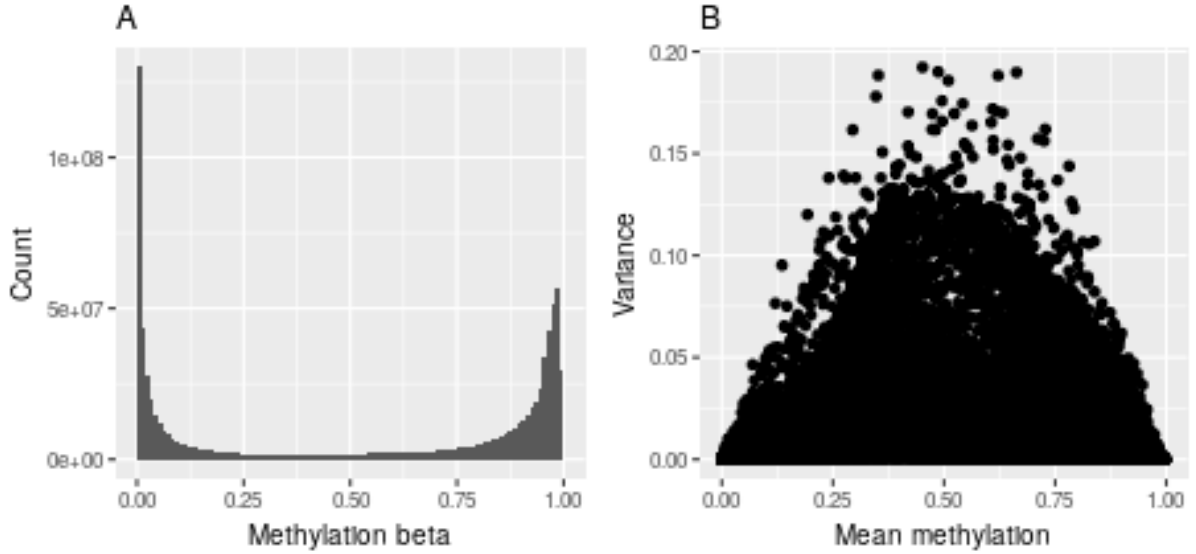


Figure 2.5: Distribution of all methylation beta values (A) and variance (B) of 485577 CpGs between 1727 individuals.

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.

### 2.2.3 Genotypes

Genotyping was performed with the Affymetrix Axiom 6.0 array. On the data set used in this work genotypes had already been called using the Illuminus calling algorithm and missing values had been imputed using the IMPUTE2 software[24]. Furthermore, imputation results had already been filtered at IMPUTE value of 0.4.

Additionally SNPs with a minor allele frequency (MAF) of less than one percent had been removed. This allows association analyses to 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	Covariate	Minimum	Maximum	Mean
Age	31	82	56.37	Age	61	81	69.04
BMI	16.05	55.99	27.65	BMI	18.99	47.58	28.87
WBC	2.5	40.9	5.98	WBC	2.7	12.6	5.9

Table 2.2: Covariate overview of all 3020 individuals in the KORA S4 survey (left) and the 687 individuals with available genotype, expression and methylation data (right).

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[25], and amplification plate, were known. These were used to correct expression values for batch effects.

### 2.2.5 Methylation quantitative trait loci

Genotypic variants are known to effect methylation patterns in humans [26]. 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 as described in section 3.5.2.

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 11,165,559 significantly associated SNP-CpG pairs. 70,709 distinct CpG-sites and 2,709,428 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 467,915 pairs consisting of 3,592 CpG-sites and 200,761 SNPs were between different chromosomes.

## 2.3 Transcription factor binding

DNA methylation plays a major role in the specifics of transcription factor binding sites[27]. 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"[28] downloaded from the UCSC genome browser download section <sup>b</sup>.

It combines 690 high quality ENCODE ChIP-seq data sets, which were processed with the Factorbook motif discovery and annotation pipeline[28]. The pipeline uses the

<sup>b</sup><http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz>

tools MEME-ChIP[29] and FIMO[30] from the MEME software suite and merges discovered motifs with known motifs from Jaspar[31] and TransFac[32] using machine learning methods and manual curation.

The track contains a total of 4,380,444 distinct peaks, representing binding sites for 161 transcription factors in 91 cell lines. For our analyses we filtered and combined the peaks for 23 blood cell lines, consisting of k562 and 22 B lymphocyte derived cell lines. This leaves a total of 2,475,316 peaks for 125 transcription factors.

The second source was the ReMap project[33]. It combines 395 publicly available ChIP-seq data sets covering 132 different transcription factors across 83 cell lines and 264 conditions. ReMap uses Bowtie2[34] to map reads to the human genome and the tool MACS[35] for peak calling. The final data set was downloaded from the ReMap website<sup>c</sup>.

There were a total of 8,905,156 peaks for 131 transcription factors in the data set. After filtering for 54 conditions, made up by 19 blood related cell lines, 1,372,245 peaks for 35 different transcription factors remained.

Combining both filtered data sets lead to a total of 3,847,561 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[36]. 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[37].

The data set contained 3,019,612 pairwise protein interactions between 16,590 different proteins. For each pair associations scores for each source were available. Protein interactions were filtered for the availability of experimental or database evidence by excluding pairs with an association score of 0 in both categories, which left 375,702 interactions between 12,769 different proteins.

Further filtering was performed by excluding proteins, whose genes were not found expressed in whole blood. For this, expression values from RNA-seq experiments for 55 tissues in human were downloaded from the Genotype-Tissue Expression (GTEx) Project[38]. Transcripts with an RPKM (Reads per kilobase per million mapped reads) of more than 0.1 in GTEx blood samples were considered expressed. Interactions where one or both members were not represented by these 17,343 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.

---

<sup>c</sup>[http://tagc.univ-mrs.fr/remap/download/All/filPeaks\\_public.bed.gz](http://tagc.univ-mrs.fr/remap/download/All/filPeaks_public.bed.gz)





## 3 Methods

Most computations were done using the statistical computing environment and programming language R, version 3.4.1[39]. 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"[40]. 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"[41] and "FDb.InfiniumMethylation.hg19"[42] 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 1 and considers the covariates age, sex, RNA integrity number (RIN), plate and storage time. The model for methylation values, shown in equation 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.

$$Expr \sim 1 + age + sex + RIN + plate + storage.time \quad (1)$$

$$Meth \sim 1 + CD4T + CD8T + NK + Bcell + Mono + PC_1 + \dots + PC_{20} \quad (2)$$

### 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 eQTLs was performed analogously, the only difference being that methylation residuals were supplied instead of SNP dosages.

Calculations were performed using the Bioconductor package MatrixEQTL[43]. 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[44], adapted for not recording all p-values.

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[45].

Then a hypergeometric test[46] for overrepresentation 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[47]. 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[48]. First, I will shortly go over the theoretical background and the method, following Mohammadi and Wit[49, 48]. All formulas are taken from the papers. For the complete theory I suggest reading the original paper[49].

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  $V = \{1, 2, \dots, p\}$  is the set of nodes and  $E \subset V \times V$  the set of edges. A zero mean Gaussian Graphical model with respect to the graph  $G$  is defined as

$$\mathcal{M}_G = \{\mathcal{N}_p(0, \Sigma | K = \Sigma^{-1} \in \mathbb{P}_G)\} \quad (3)$$

where  $\Sigma$  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  $\mu$  is zero.

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

$$Pr(Z|K, G) \propto |K|^{n/2} \exp \left\{ -\frac{1}{2} \text{tr}(KU) \right\}, \quad (4)$$

where  $U = Z^T Z$ .

In section 1.3 I introduced GGMs as being based on partial correlations. For multivariate Gaussian distributions conditional dependence is a direct representation of partial correlation[50]. Therefore, to stay closer to the paper we will only use the term conditional dependence 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  $G$ .

The space of possible graphs is too large 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 this 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 specifically, 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  $K$  given as  $Pr(K, G|Z)$ .

The birth-rates and death-rates are defined as

$$\beta_e(K) = \min \left\{ \frac{Pr(G^{+e}, K^{+e}|Z)}{Pr(G, K|Z)}, 1 \right\}, \text{ for each } e \notin E, \quad (5)$$

$$\delta_e(K) = \min \left\{ \frac{Pr(G^{-e}, K^{-e}|Z)}{Pr(G, K|Z)}, 1 \right\}, \text{ for each } e \in E, \quad (6)$$

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

$$\beta(K) = \sum_{e \notin E} \beta_e(K), \quad (7)$$

$$\delta(K) = \sum_{e \in E} \delta_e(K). \quad (8)$$

The waiting time between successive events is exponentially distributed and has the mean  $1/(\beta(K) + \delta(K))$ . This means the length of the waiting time is dependent on the how likely a graph is, compared to the graphs that could be reached by a birth or death event. The probability of a birth or death event is given by

$$Pr(\text{birth of link } e) = \frac{\beta_e(K)}{\beta(K) + \delta(K)}, \text{ for each } e \notin E, \quad (9)$$

$$Pr(\text{death of link } e) = \frac{\delta_e(K)}{\beta(K) + \delta(K)}, \text{ for each } e \in E. \quad (10)$$

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

$$Pr(K, G|Z) \propto Pr(Z|K)Pr(K|G)Pr(G) \quad (11)$$

$Pr(Z|K)$  is the likelihood and was defined previously. The default for the prior of the graph is the uniform distribution  $Pr(G) = \frac{1}{|\mathcal{G}|}$ , where  $\mathcal{G}$  is the space of possible graphs. 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:

$$Pr(K|G) = \frac{1}{I_G(b, D)} |K|^{\frac{b-2}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(DK) \right\} 1(K \in \mathbb{P}_G), \quad (12)$$

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.

In the calculation of GGMs with the package BDgraph the following steps are repeated for a set 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 9 and 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 [49], based on work by Lenkoski[51].

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 the waiting times were long due to low birth and death rates. The approach takes some time

to converge close to the posterior distribution. Therefore, BDgraph offers 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.

### 3.5.2 Data collection

The stability of Gaussian Graphical models is limited by the property, that sample covariance and correlation matrices are well conditioned only when the number of variables is smaller than the number of samples[52]. Therefore, it is absolutely necessary to perform a pre-selection of included entities when calculating GGMs on our multiomics data set instead of calculating genome wide networks.

We performed two approaches of selecting potential node sets. Both are anchored in trans-acting SNP-CpG interactions as well as a relationship to HERV elements. Trans-meQTLs provide interesting boundaries for the analysis of regulatory mechanisms, as the effect is guaranteed to be indirect. Furthermore, the effect size of a trans-meQTL has to be big to reach significance.

The first approach starts 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 nodes. To include genes, that are potentially associated with this SNP, all expression probes within a 250 kb flanking region up- and downstream of the SNP (SNP genes/probes) are included. Analogously, to find genes that are likely to be affected by methylation changes, for each associated CpG all overlapping probes and the

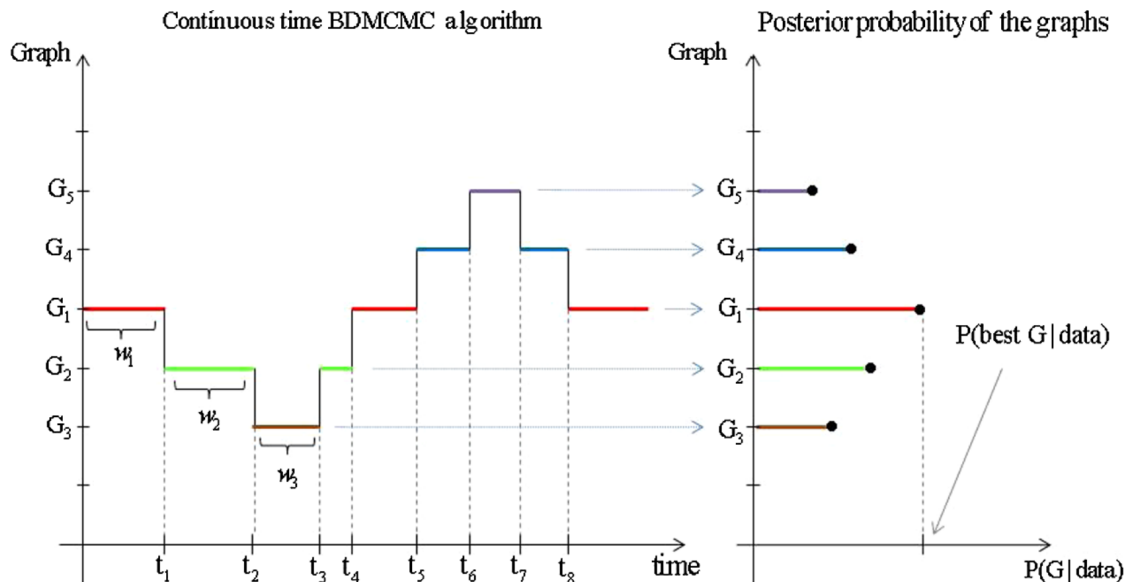


Figure 3.1: (Left) Continuous time BDMCMC algorithm where  $\{t_1, t_2, t_3, \dots\}$  are jumping times and  $\{w_1, w_2, w_3, \dots\}$  are waiting times. (Right) Posterior probability estimation of the graphs based on the proportions of their waiting times.

first up- and downstream probe each within a distance of less than 250 kb are added (CpG genes). To include transcription factors (TFs), 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, to connect genes around the sentinel SNP with genes around the methylation sites or transcription factors, we used the protein-protein interaction network described in section 2.4. 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 3.2A.

The second approach of data collection, shown in figure 3.2B, 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 and are part of a trans-meQTL, they are added to the analysis. Then for each CpG of interest the most strongly associated SNP that lies on another chromosome is added. Other CpGs that are in a trans-meQTL with these SNPs are included as well. Finally neighboring genes, transcription factors and shortest path genes are included according to the same criteria as in the first approach.

For both approaches data sets that contained too few members were excluded as follow. Sets with less than one gene or probe neighboring a SNP and one neighboring a CpG-site were removed. Furthermore, if no transcription factor and no shortest path gene was found, the set was excluded. Therefore, each GGM calculated contained at least 5 entities.

### 3.5.3 GGM calculation and filtering

The BDgraph main method "bdgraph" was used to calculate the GGMs on the data sets that were just defined. I used the parameters, 'method = "gcmg"', which leads to the

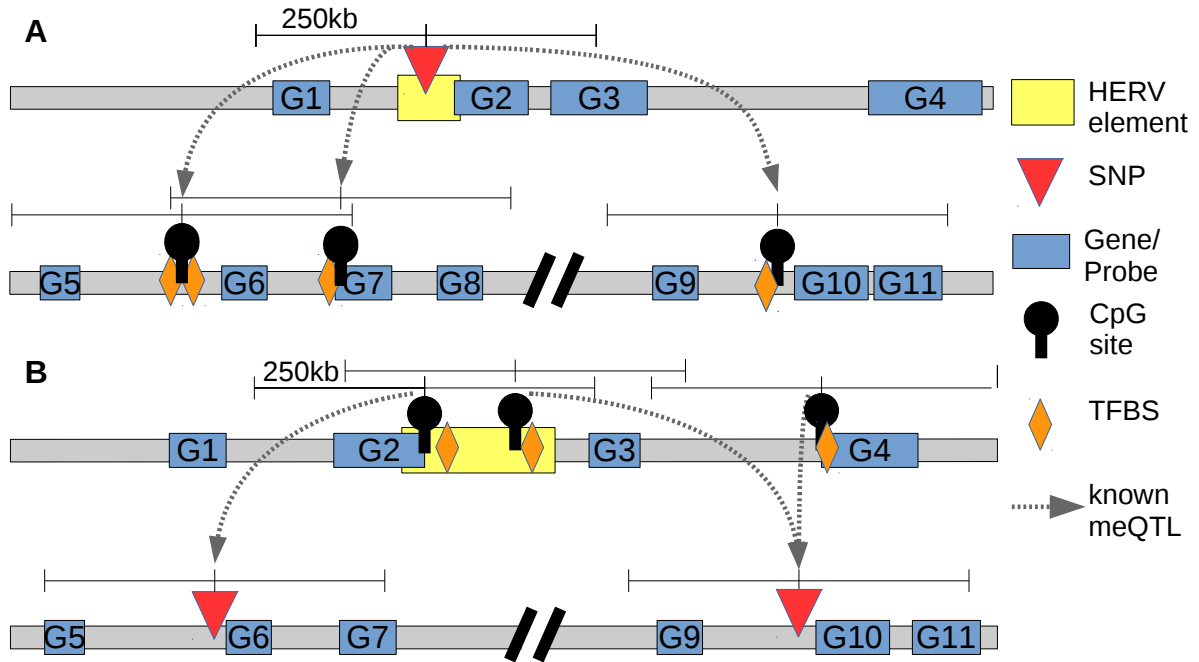


Figure 3.2: Scheme for data collection starting from a SNP (A) and a CpG (B) within a HERV element.

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'.

Due to large number and size of calculated GGMs no manual analyses of all networks was possible. Therefore, multiple criteria were defined to identify potentially interesting GGMs.

First the edges in the network were filtered for an edge inclusion probability of more than 0.9. Then the connected components were determined. Networks were restricted to the connected component, which contained the seed SNP or seed CpG. If several seed CpGs were in different connected components, the bigger one was chosen. Finally, restricted networks with less than one SNP/CpG gene or probe and no shortest path gene or transcription factor were discarded.

Figure 3.3 shows the structure we expected to find in the networks. A SNP is associated with a neighboring gene. This gene connects to a CpG site through a route of transcription factor and/or shortest path genes. The CpG in turn has an edge to a neighboring CpG gene.

Therefore, the number edges between SNPs/CpGs and neighboring genes were counted, to identify restricted networks, that potentially follow this pattern. Finally, the eQTL and eQTM results described in sections 4.3 and 4.4 were used to assess how many of the identified associations were found in the network, either through a path over other nodes or as a direct edge.

A manual inspection of these criteria was performed in order to select which GGMs would be the target of detailed manual analysis.

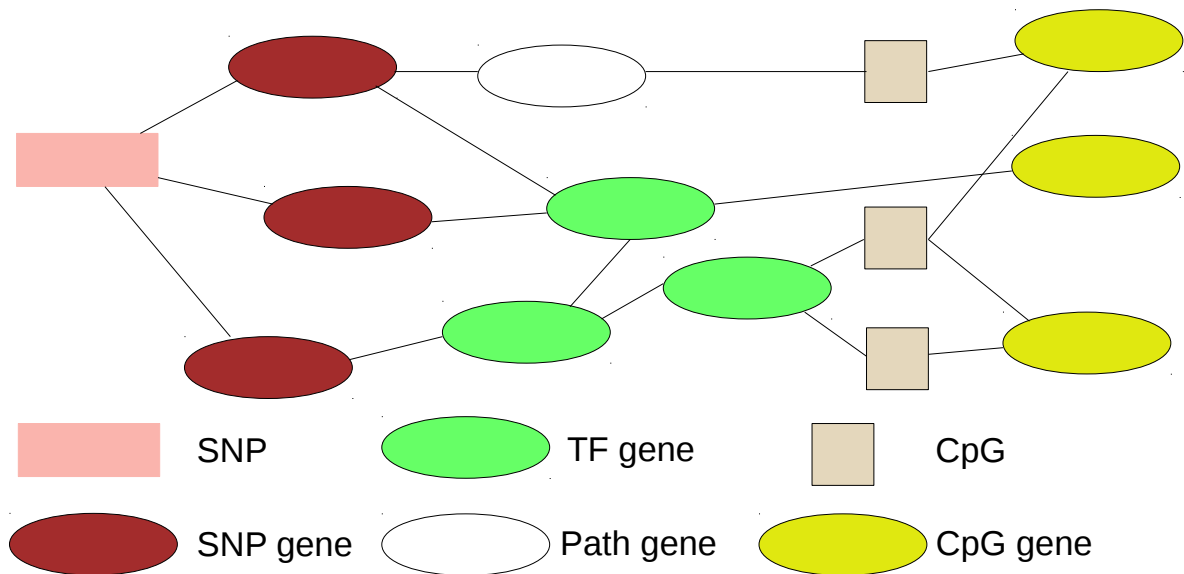


Figure 3.3: Expected structure of regulatory networks.





## 4 Results

### 4.1 Normalized Data

The distributions of the expression and methylation residuals, calculated as described in section 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[49].

### 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 section 2.1 and including flanking regions will be shown in tables or in the supplementary data.

#### 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).

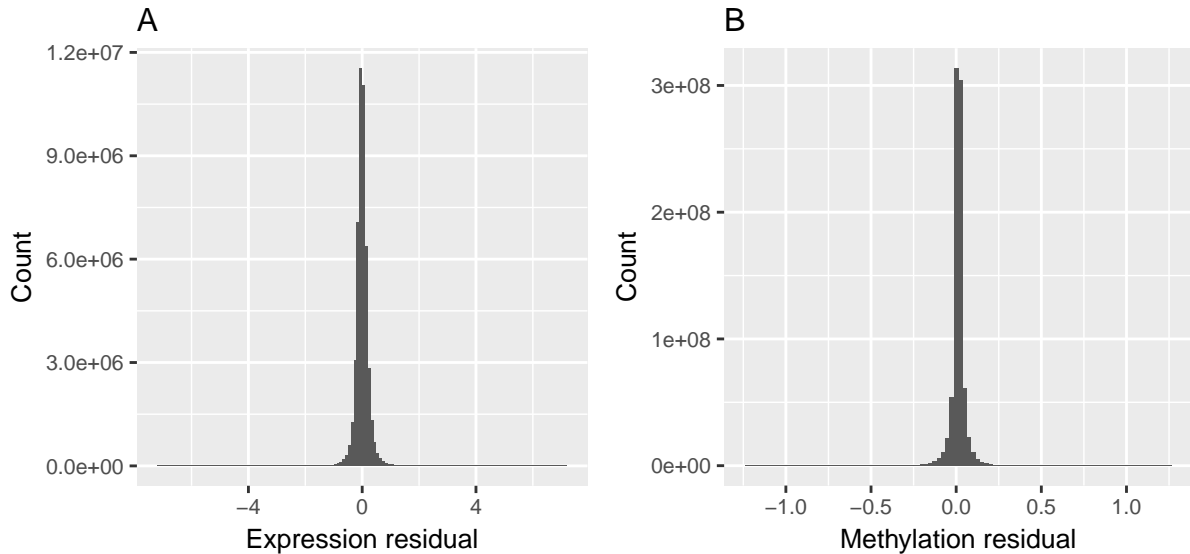


Figure 4.1: Distribution of expression (A) and methylation (B) residuals over all samples and probes

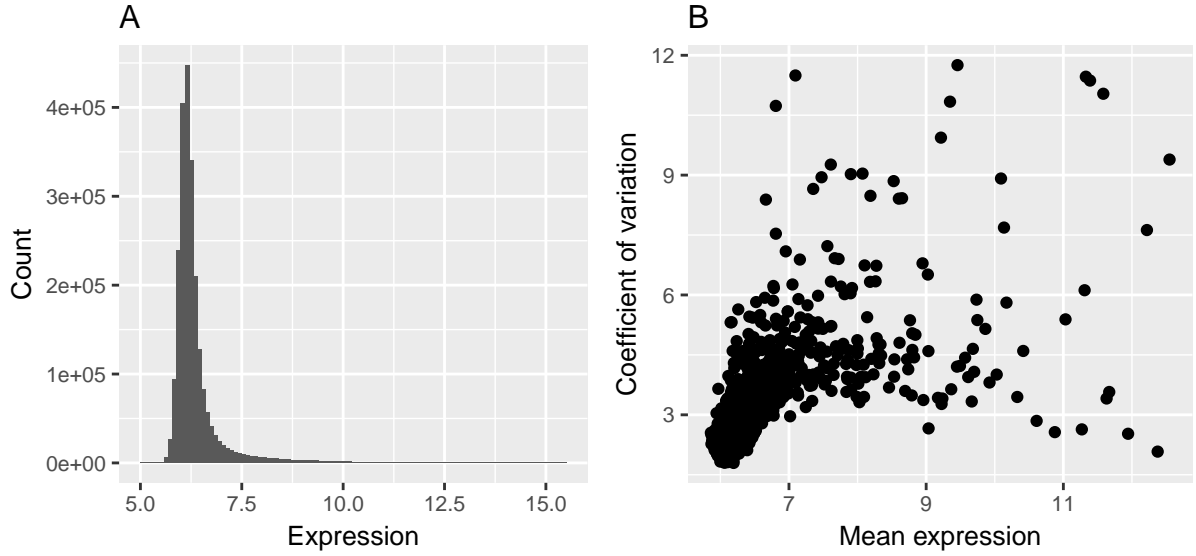


Figure 4.2: Distribution of expression values (A) and coefficient of expression variation (B) of 2343 expression probes overlapping with HERV S2 in 993 individuals.

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,  $p\text{-value} = 1.37 \cdot 10^{-5}$ ,  $fdr = 0.047$ ) 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 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.

## 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.

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.

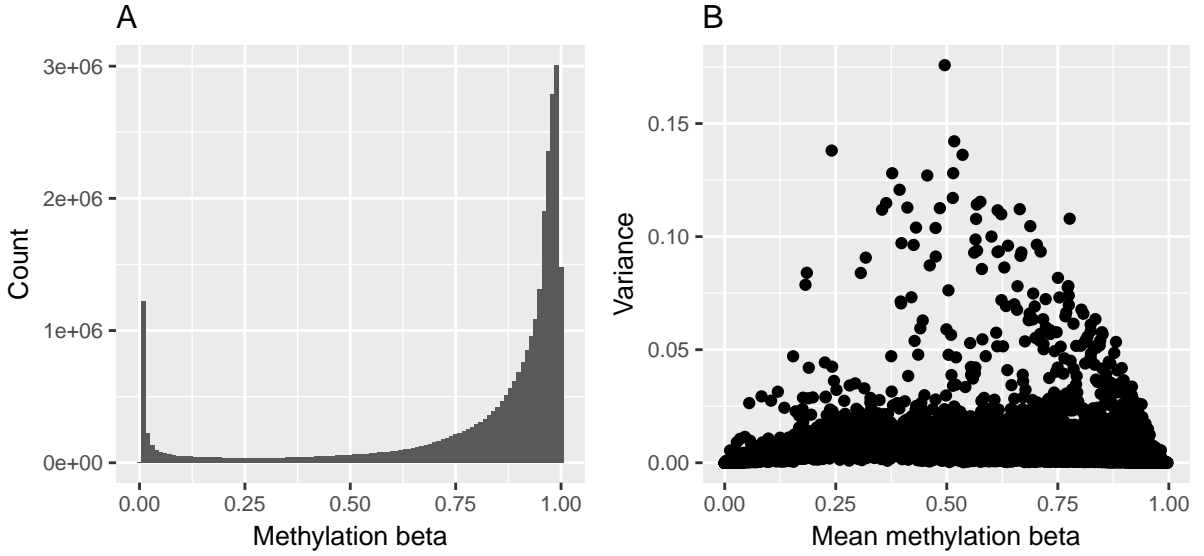


Figure 4.3: Distribution of methylation beta values (A) and coefficient of methylation variation (B) of 17077 GpG-sites located within HERV S2 in 1727 individuals.

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.

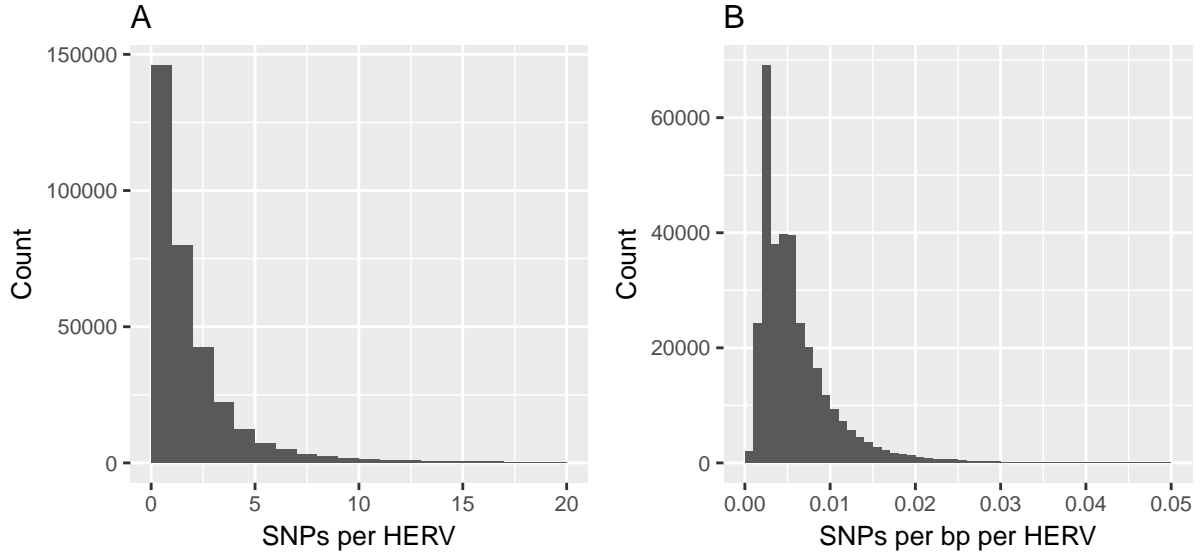


Figure 4.4:

### 4.2.3 Genotypes

A total of 890780 of 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. When considering a uniform distribution of the measured SNPs over the whole genome, a concentration of 3.18 SNPs per kilobase would be expected. The majority of HERV elements have a similar concentration, however, there is a sizable subset, that is more densely populated by measured SNPs.

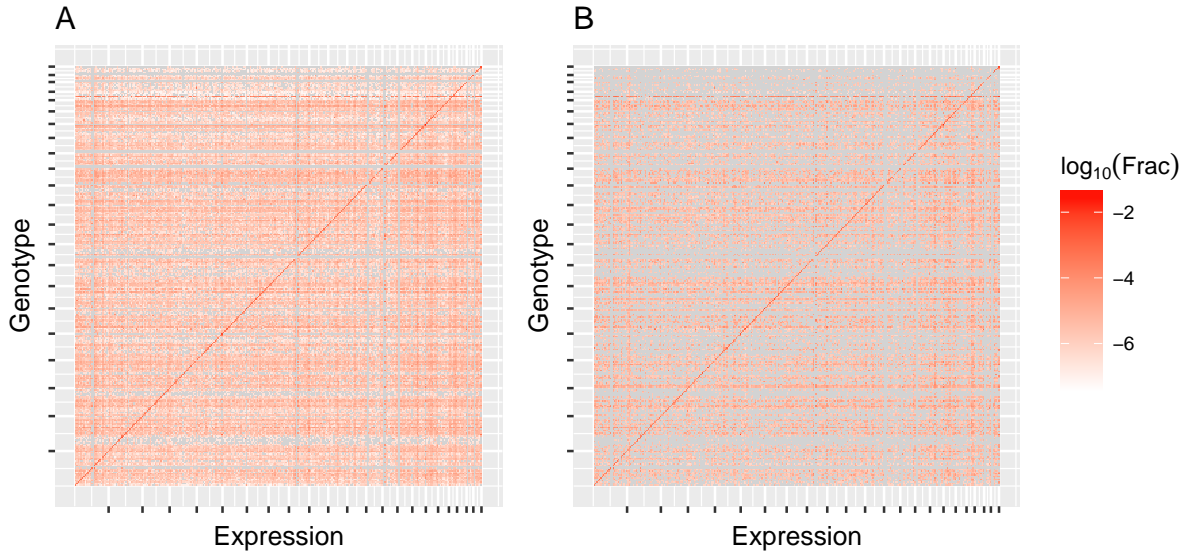


Figure 4.5: Fraction of SNP-expression probe pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTLs in the given pair of bins. A considers all eQTLs and pairs, while B is limited on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

### 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 these cis-pairs were significantly associated with p-values of  $10^{-6}$  or less. These significant eQTLs have a false discovery rate of less than  $1.93 \times 10^{-4}$ . 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.

There were a total of 1511235 significant trans-eQTLs with an p-value of less than  $10^{-8}$  and a FDR of less than  $3.02 \times 10^{-4}$ . 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. [53] calculated eQTLs on expression and genotype data from 2,708 subjects, including 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 genomic windows. As expected the values at the diagonal are comparatively higher, containing all cis-eQTLs.

The expression residual signatures of homozygous reference allele, heterozygous and homozygous alternative allele for the two eQTLs with the most and least significant 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.

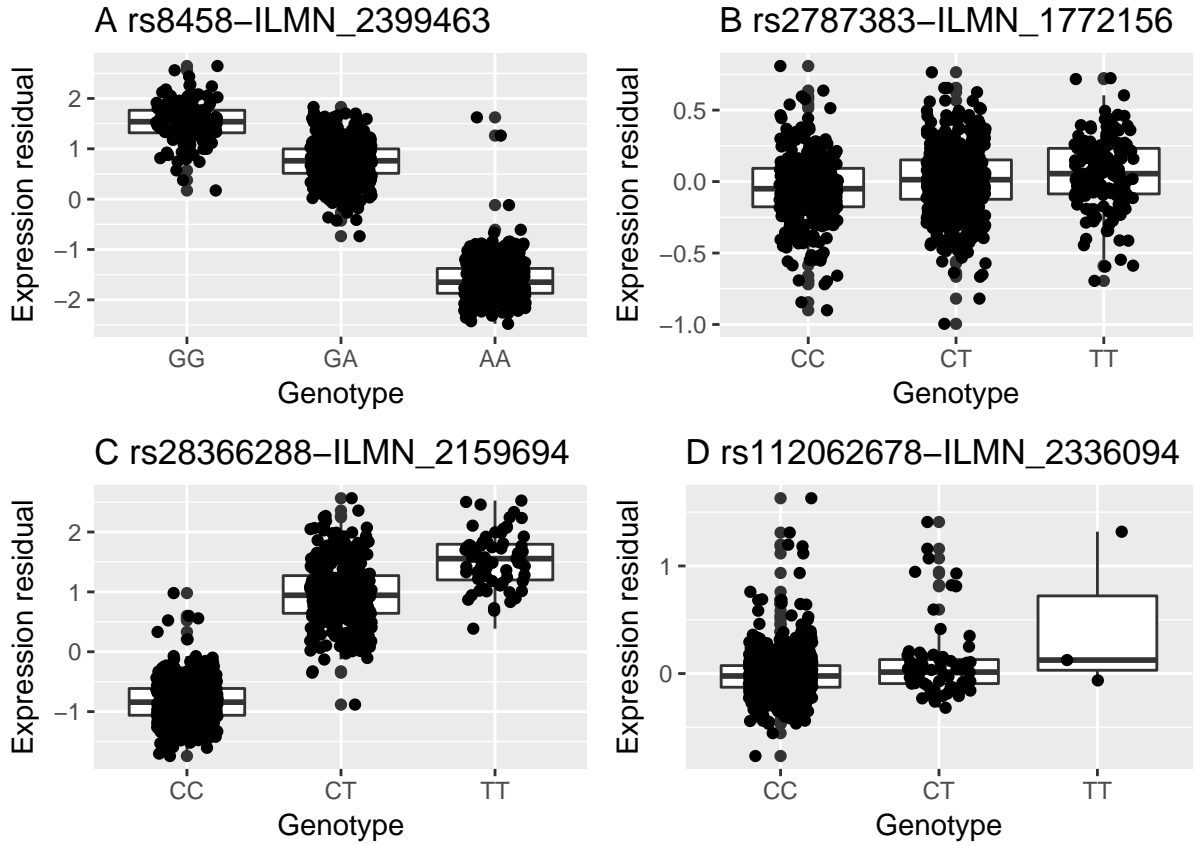


Figure 4.6: Distribution of expression residuals with respect to genotypic variant in most significant (left column) and least significant (right column) eQTLs for cis (upper row) and trans(lower row).

45862 of the SNPs and 166 of the expression probes found in at least one cis-eQTL were 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 remained. They were made up by 3551 different expression probes, annotated to 2638 genes, and 80573 SNPs.

There were 5855 cis-eQTLs between one of 4748 SNPs within a HERV element and one of 128 expression probes, that overlap with a HERV element.

When limiting trans-eQTLs to the ones that contain a HERV S2 related SNP (23396) and/or expression probe (1227), 240301 remained, comprising 14573 unique expression probes (6357 genes) and 56672 SNPs. In 9177 of these pairs both members were HERV related.

The fraction of HERV S2 related SNP-expression probe pairs that proved to be significantly associated is shown in figure 4.5B. Due the smaller number of pairs, there were more cells having no data available. The general fraction of significant eQTLs was lower than when considering all eQTLs. However, the positional distribution was rather similar, with the strong diagonal containing all cis-eQTLs being clearly visible.

A GO enrichment was performed on all 8211 genes associated to HERV related eQTLs using the set of all genes found in eQTLs as background. However, there were no significantly enriched genes.

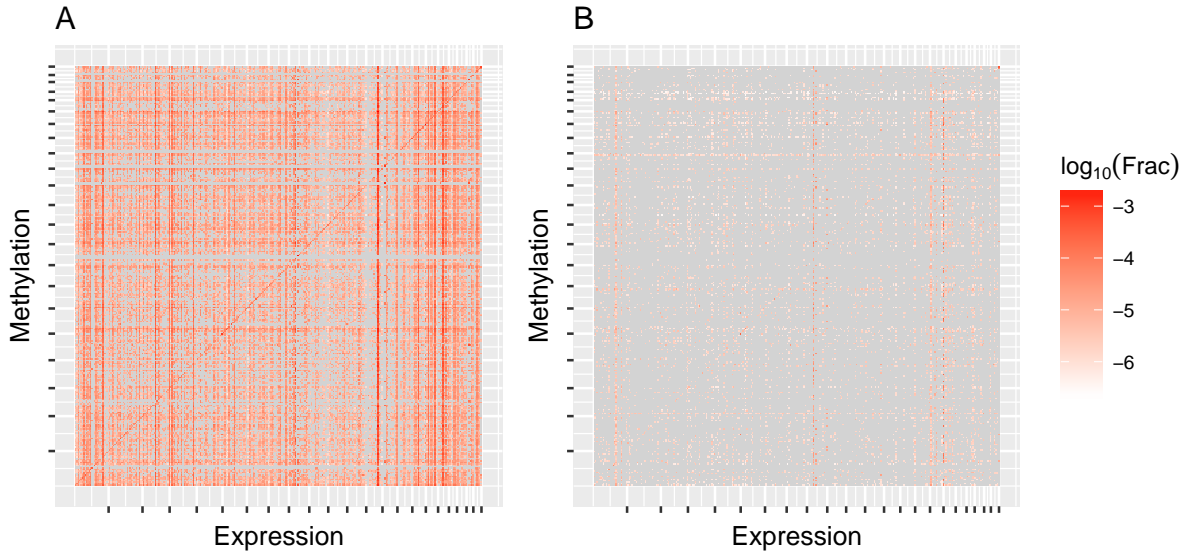


Figure 4.7: Fraction of CpG-expression probe pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTM in the given pair of bins. A considers all eQTMs and pairs, while B is limited on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

## 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  $10^{-6}$  for cis resulted in 8187 significant associations ( $fdr < 1.7 * 10^{-4}$ ) 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  $10^{-8}$  ( $fdr < 6.5 * 10^{-4}$ ). 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.

The distribution of methylation betas for all CpGs that are part of at least one eQTM are shown in figure 4.9A. It has a higher proportion of beta values close to zero than the distribution for all CpGs (figure 2.5).

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 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



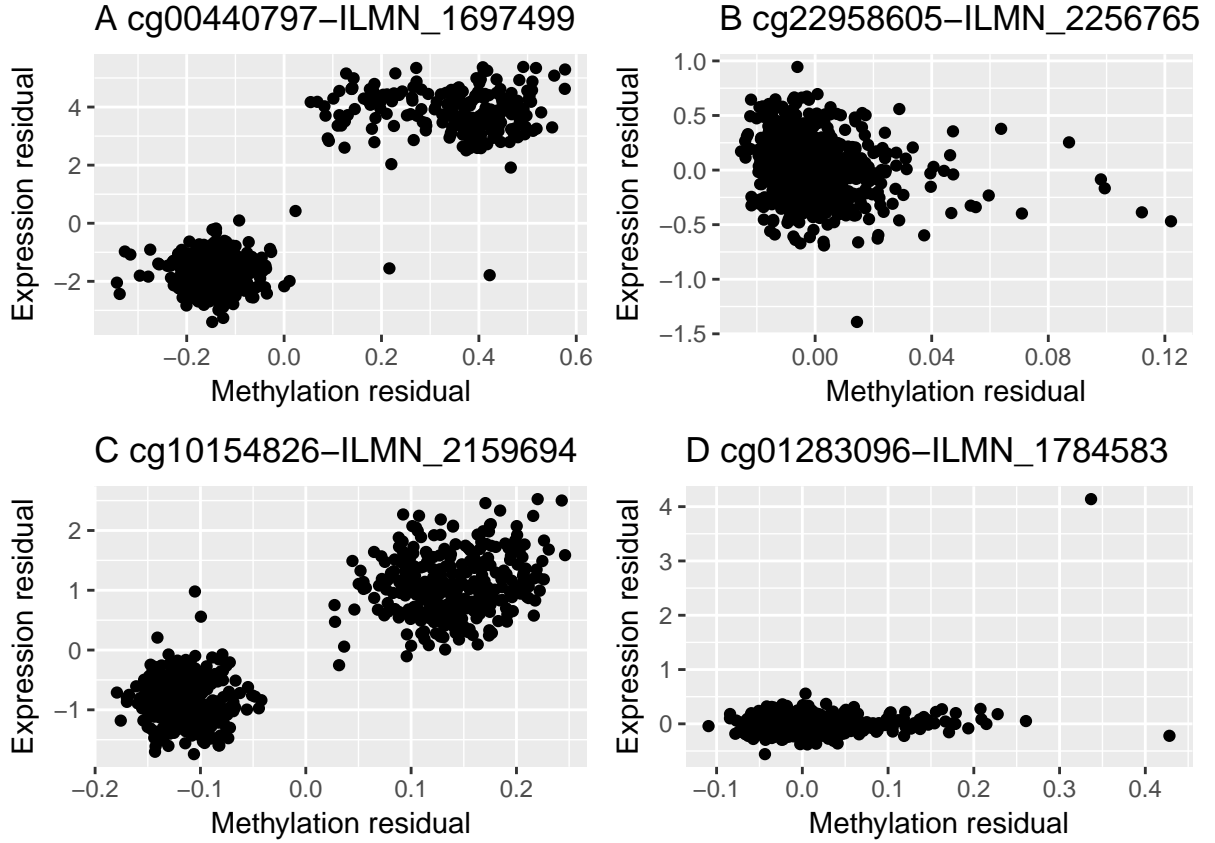


Figure 4.8:

15085 trans-eQTM are related to HERV S2.

The fraction of potential HERV related eQTMs that were significant is shown in figure 4.7B. Due to low number of significant pairs compared to the 90000 cells in the graph it was only sparsely filled. Only a weak preference for cis-eQTM could be seen.

Figure 4.9B shows the distribution of methylation beta values for all CpGs that are part of at least one eQTM and lie within HERV S2. It is very similar to the distribution for all eQTM-CpGs. However, comparing the distribution to the one for all CpGs within HERV elements (figure 4.3A), shows that the proportion of beta values close to zero is higher.

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 section 2.2.5. Therefore, in this section 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 1915 unique CpGs and 286289 SNPs. This means each of these CpG-

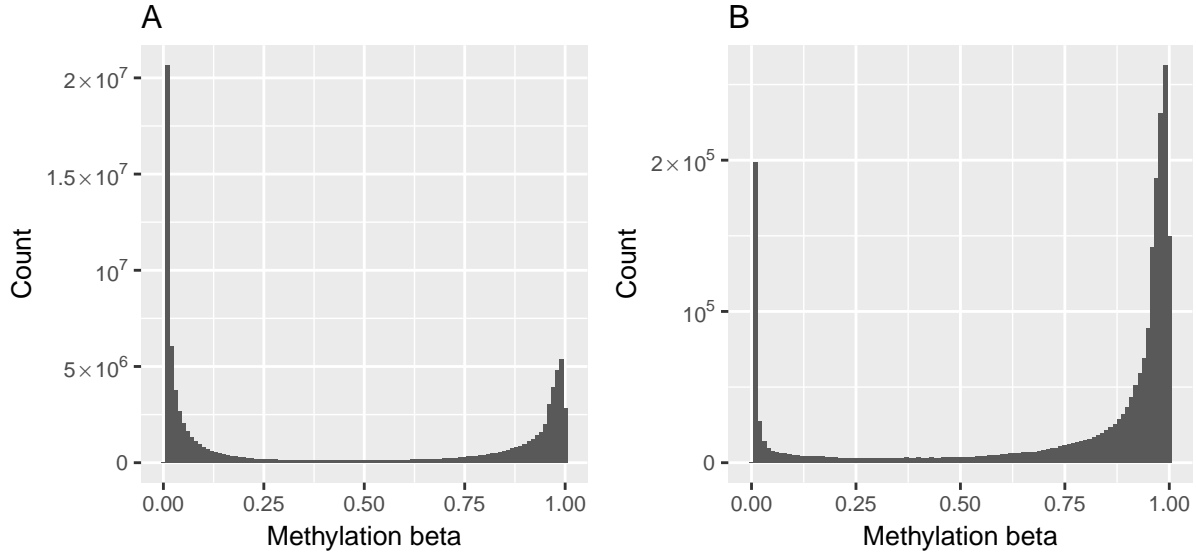


Figure 4.9:

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 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 with respect to their genomic locations are shown in figure 4.10A. 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, what we defined 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.

A GO biological process enrichment was performed on the 5128 genes, that contained at least one of the 53167 CpGs that were part of an HERV related meQTL. The significantly enriched terms are shown in table 4.5

## 4.6 HERV related regulatory networks

In the following sections I will present the results of the GGM analysis on HERV related multiomics data. First, I will describe the data sets that were generated according to the two methods described in section 3.5.2. Next I will give a general overview over the GGMs that were calculated using BDgraph and describe several measures that I used to

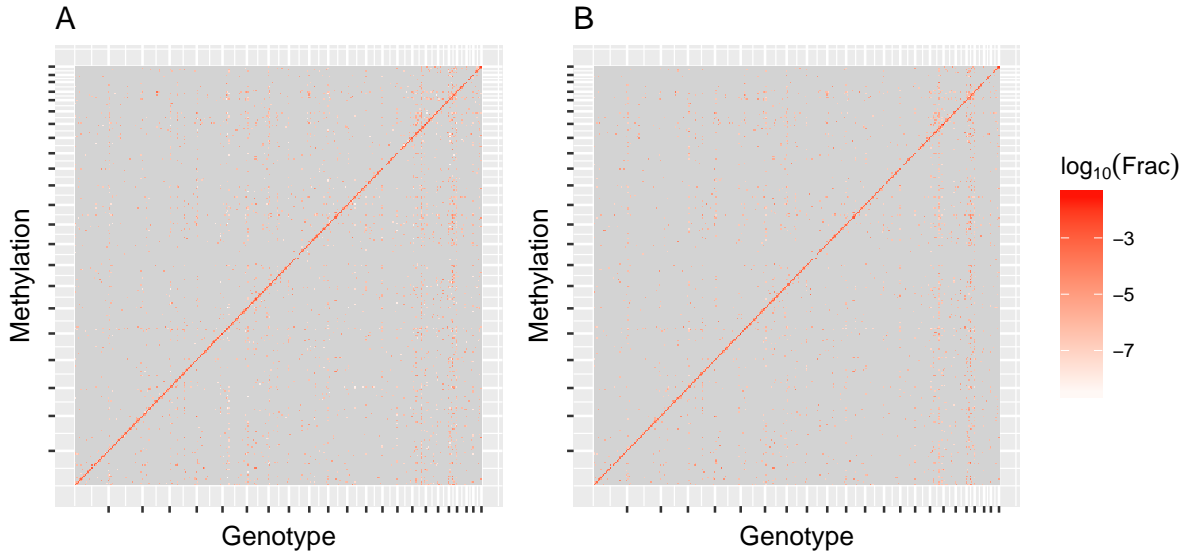


Figure 4.10: Fraction of SNP-CpG pairs that were significantly associated sorted into bins according to their genomic locations. Ticks on the axis denote chromosome boundaries. Grey means there were no eQTLs in the given pair of bins. A considers all meQTLs and pairs, while B is limited on the ones whose CpGs and/or SNPs are related to HERV set 2.

Term ID	Term	p	fdr
GO:0007399	nervous system development	$3.18 \cdot 10^{-8}$	$4.01 \cdot 10^{-4}$
GO:0048699	generation of neurons	$2.94 \cdot 10^{-7}$	$3.71 \cdot 10^{-3}$
GO:0022008	neurogenesis	$5.95 \cdot 10^{-7}$	$7.49 \cdot 10^{-3}$
GO:0030182	neuron differentiation	$7.21 \cdot 10^{-7}$	$9.08 \cdot 10^{-3}$
GO:0023051	regulation of signaling	$1.13 \cdot 10^{-6}$	$1.42 \cdot 10^{-2}$
GO:0023052	signaling	$2.08 \cdot 10^{-6}$	$2.62 \cdot 10^{-2}$
GO:0048585	negative regulation of response to stimulus	$2.18 \cdot 10^{-6}$	$2.75 \cdot 10^{-2}$
GO:0044700	single organism signaling	$2.43 \cdot 10^{-6}$	$3.05 \cdot 10^{-2}$
GO:0010646	regulation of cell communication	$2.65 \cdot 10^{-6}$	$3.34 \cdot 10^{-2}$
GO:0040011	locomotion	$3.57 \cdot 10^{-6}$	$4.49 \cdot 10^{-2}$

Table 4.5: Significantly enriched GO biological process terms among genes containing CpG-sites participating in HERV S2 related meQTLs.

identify interesting networks. Finally, I will show some of these interesting networks and analyze some of their structures in detail.

## 4.6.1 Data collection

In the following "entity" describes one of a CpG site, a SNP, a gene or an expression probe.

Filtering the SNPs that were part of at least five trans-meQTLs for relation to elements in HERV S2 resulted in 1885 seed-SNPs. These were consequently used to generate 1882

data sets for GGM calculation accordingly to the criteria defined in 3.5.2. These sets contained between 13 and 428 entities. The exact size and composition of the biggest (A) and smallest (B) sets is shown in figure 4.11.

In general the biggest groups of entities were CpGs, genes and probes neighboring CpGs and transcription factors. The number genes and probes around the SNP was rather consistent. 1103 of the data sets were enriched with shortest path genes derived from protein-protein interactions. The length of these paths was between 1 and 8.

For the selection of the CpG based approach, there were 148 CpGs that were part of an trans-meQTL and laid within HERV S2. 103 of these CpGs produced data sets, that contained all demanded types of entities. However, the inclusion criterion of all other CpGs associated to the best associated SNP to any seed CpG lead to some duplicated sets, when the additional CpGs were seed-CpGs themselves. These sets were merged, as their data matrices were identical, which left 100 data sets.

93 of these started from a single seed CpG, while 4 sets contained two seed CpGs, two sets three and one set four.

In total there were between 5 and 914 entities per data set. The number and composition of entities in the ten data sets with most (C) and least (D) entities is shown in figure 4.11.

The distribution of different types of entities was similar to the data sets constructed starting from HERV related SNPs. A total of 34 data sets contained between 1 and 10 shortest path genes.

## 4.6.2 General network overview

The following section will give a short overview over GGMs calculated as described in section 3.5.3.

937 of the 1882 data sets generated from HERV related SNPs remained after filtering. The restricted networks consisted of between 16 and 425 nodes. The number and composition of nodes for the biggest (A) and smallest networks can be seen in figure 4.12.

The networks contained between xx and xxxx edges.

The networks were investigated for the presence of SNP-expression probe pairs as well as CpG-expression probe pairs known to be associated from the eQTL and eQTM analysis.

484 networks contained both members of up to three eQTLs. In 448 of these the eQTLs were representend as an edge. Analogously 211 networks contained eQTMs, but only in three of the networks one eQTM was present as an edge.

After filtering the 100 networks seeded in HERV-CpGs, 37 networks remained. The restricted networks contained between 23 and 914 nodes. Their number and composition of nodes are shown in figure 4.12C.

None of these networks contained any eQTLs. 18 networks contained eQTMs, of which 3 were present as edge.

As displayed in figure 4.13B almost all networks had one connected component which contained almost all nodes. Furthermore, GGMs with more entities were more likely to be retained after the filtering.

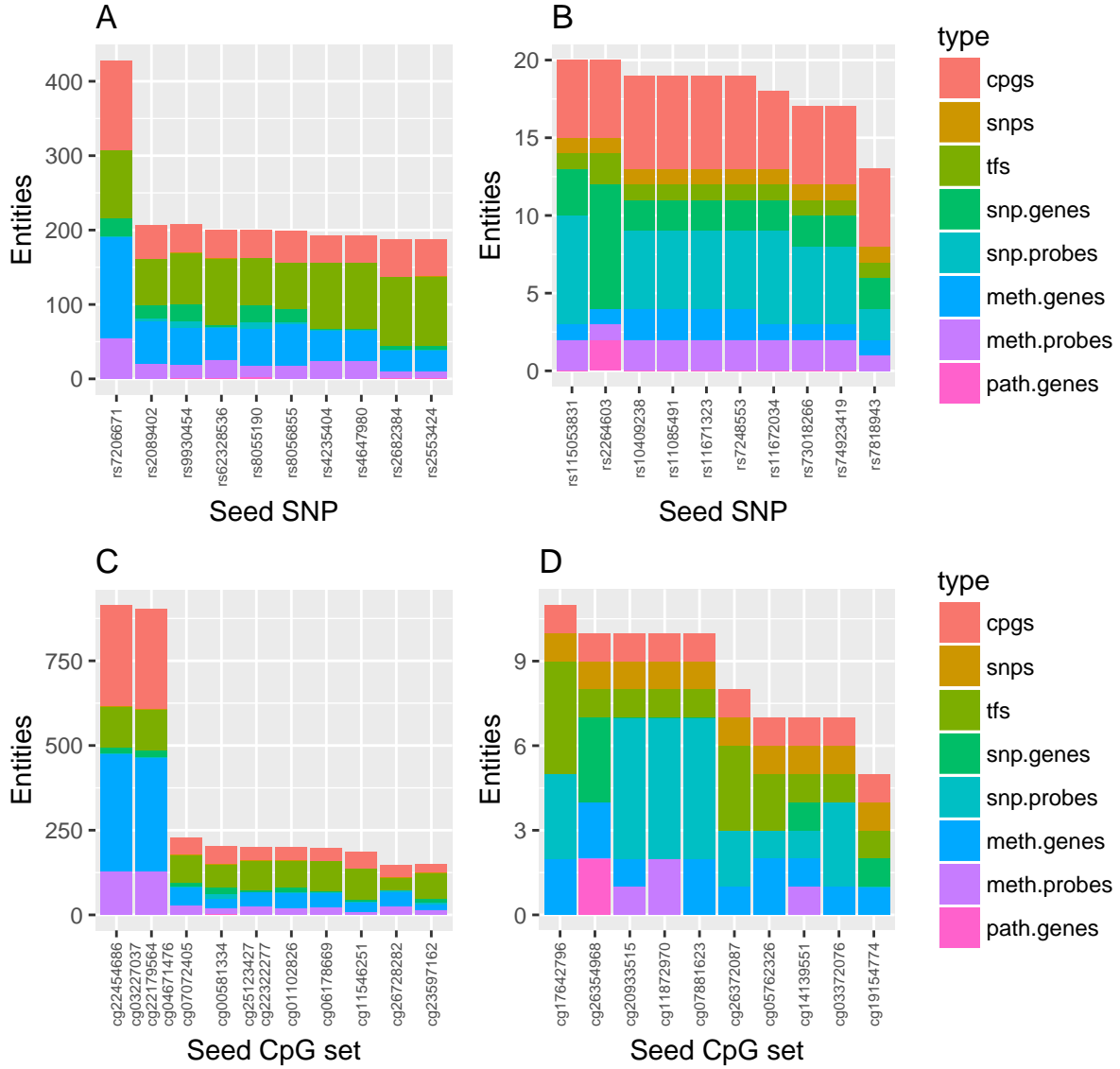


Figure 4.11: Size and composition of biggest (left) and smallest (right) GGM data sets. The upper row are sets seeded in HERV-SNPs, while the lower row is seeded in HERV-CpGs. In few cases genes were included as SNP or CpG neighboring gene as well as transcription factor. These genes are displayed as the former.

### 4.6.3 Example networks

In the following section I will show some interesting restricted networks.

Figure 4.14 shows the network generated from the dataset seeded in rs11079148 after edges were filtered at an edge inclusion probability of 0.9 or more. Only the biggest connected component, containing the SNP, is shown.

The dataset contained 95 entities, made up of one SNP, 29 CpGs that were significantly associated with it, 4 genes or unmapped probes around the SNP, 29 genes or unmapped probes neighboring the CpGs, 27 transcription factors and 8 shortest path genes. Of these 36 remained connected to the SNP. Apart from the SNP itself, they are 19 CpGs, one SNP gene, 11 CpG genes and probes, 10 transcription factors and 2 shortest path genes. The

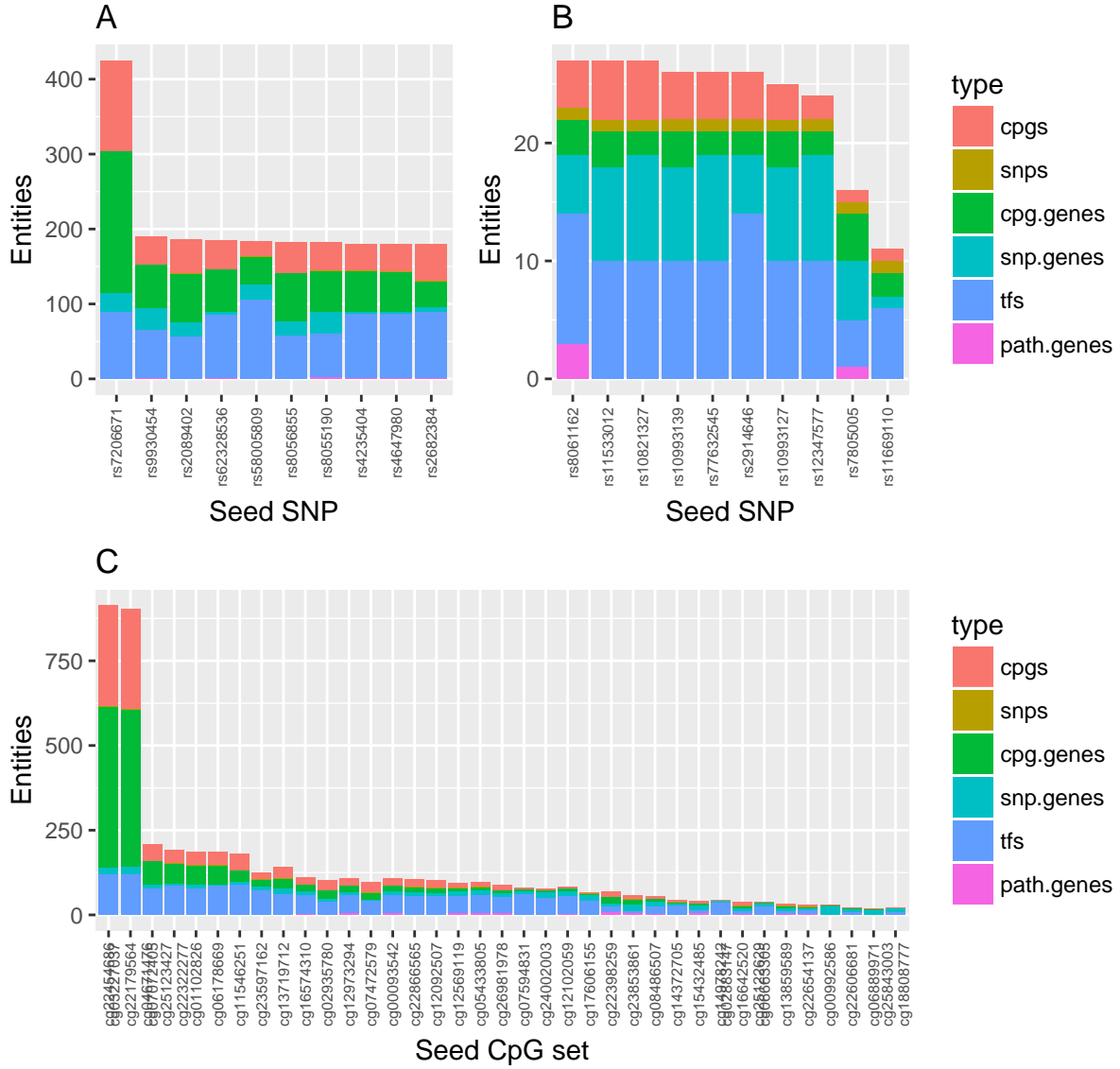


Figure 4.12: Number and composition of nodes in restricted GGM networks. A and B are the ten biggest networks for GGMs seeded in SNPs, C are all networks for GGMs seeded HERV-CpGs.

restricted network contains 35 edges, of which 3 are of special interest, as they connect CpG sites and their neighboring genes. There is no edge between the SNP and its neighboring gene. Instead the SNP is directly connected to three CpGs and one probe neighboring a CpG.

Another restricted network generated from a dataset based on a seed SNP, rs1833977, is shown in figure 4.15.

It contains 28 of the 42 entities in the dataset. Apart from the seed SNP these are 5 of 5 CpGs, 3 of 10 CpG genes and probes, 5 of 9 SNP genes and probes, 11 of 14 transcription factors and all 3 shortest path genes in the dataset. The network has 49 edges, among which there is one edge between the SNP and one of its neighboring genes, as well as one edge between a CpG and a CpG gene. The network actually allows to trace the following path from the SNP a CpG gene as described in figure 3.3: rs1833977 -

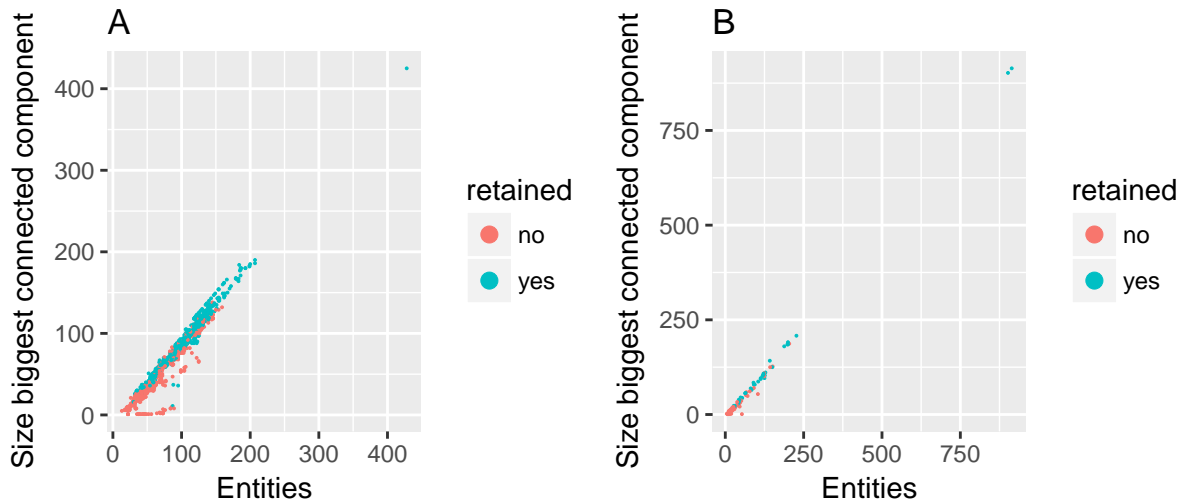


Figure 4.13: Number of entities and size of the biggest connected component of GGMs based on SNP (A) and CpG (B) seeds.

LDHD - CEBPA - cg25930786 - OSCAR. However, this path is not the only and not the shortest path between these nodes, as the CpG also is directly connected to the SNP.

The network shown in figure 4.16 is generated from the dataset collected starting with the CpG-site cg06889971.

19 of the 29 entities in the dataset are in the same connected component as the seed CpG. These are one of the two CpG genes, 15 of the 24 SNP genes and probes, two of which () lie within a HERV element themselves, and the only transcription factor in the data set. The restricted network contains 28 edges. The CpG has only one edge to the SNP, which shares an edge with the SNP gene ZNF415.

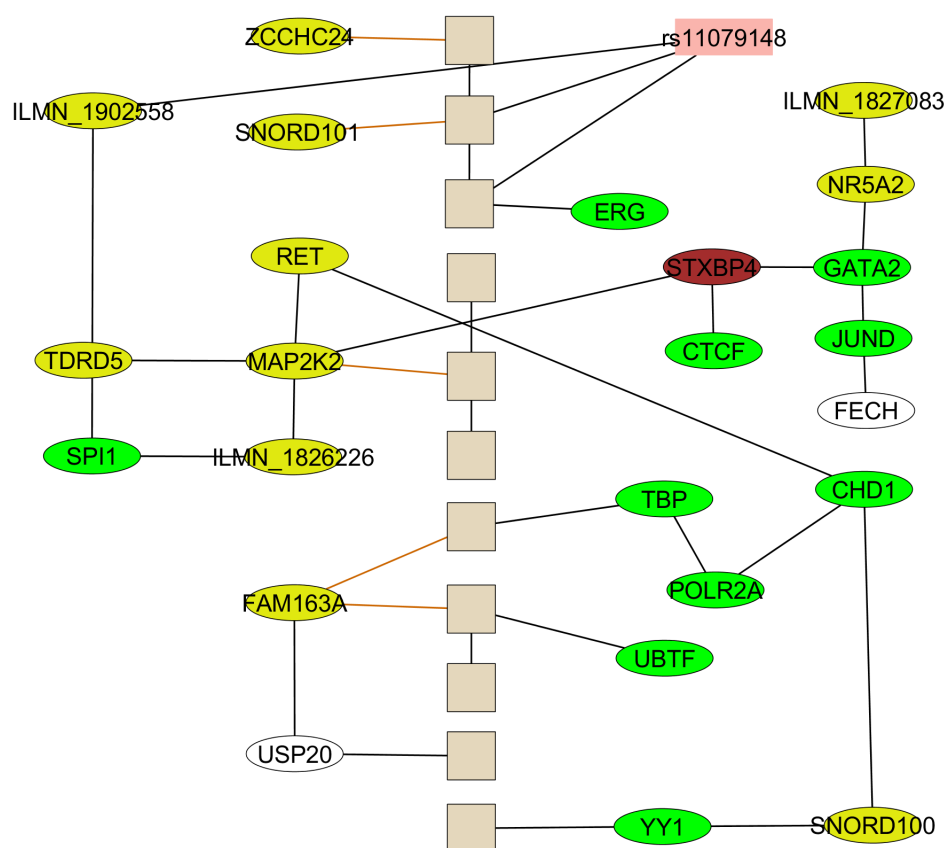


Figure 4.14: Restricted network seeded in rs11079148.

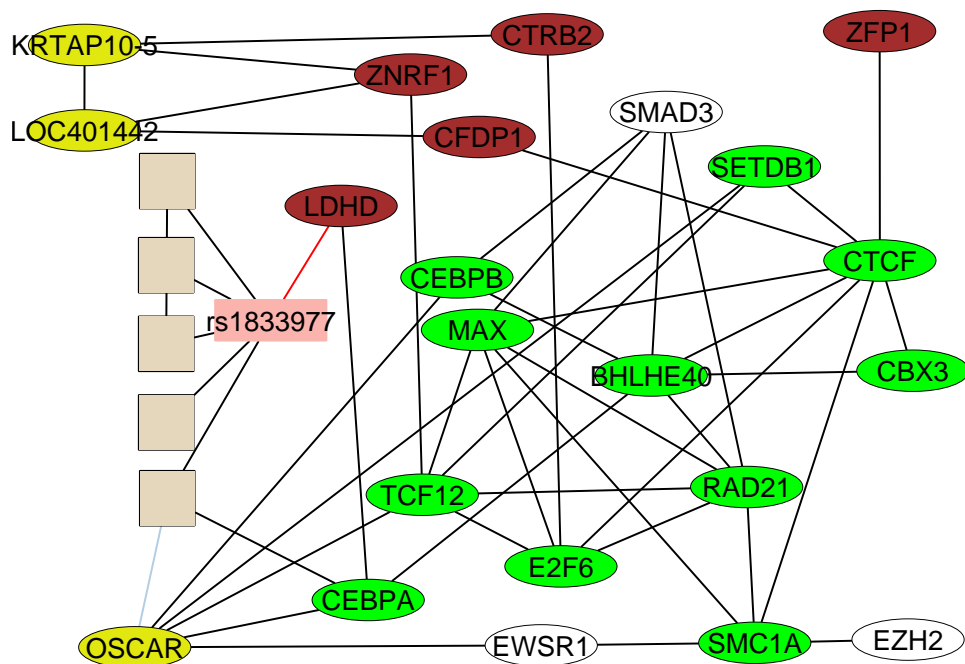


Figure 4.15: Restricted network seeded in rs1833977.



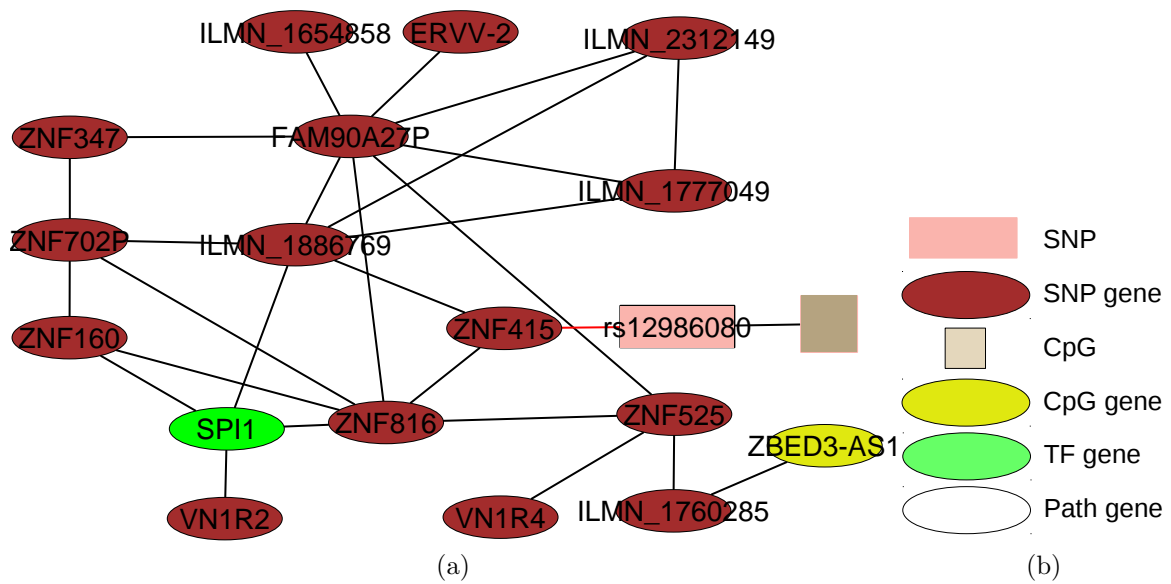


Figure 4.16: Restricted network seeded in cg06889971.

## 5 Discussion

### 5.1 HERV region features

Genotypes, expression levels and methylation levels were measured with microarray based methods. This means no global, unbiased analysis is possible. Therefore, the absence of measurements for many HERV elements is not sufficient to infer e.g. the lack of expression activity. The design of these arrays is not only based on prior knowledge and an assessment of the interest of each probe, but also sequence properties. For example the data sheet for the HumanHT-12 v3 Expression BeadChip[54] mentions as parameters for the probe design among others the lack of similarity to other genes and the absence of highly repeated sequence in the genome. As a result it is to be expected, that HERVs, which are considered repeats after all, are covered sparsely.

The number of available expression measurements within HERV elements was lower than expected, if the probes were randomly distributed. This is explained by the mentioned limitations in probe design, as well as the focus on annotated genes. This was supported by the result, that while over 60% of all probes on the HumanHT-12 v3 Expression BeadChip are annotated to a gene, less than one fourth of the probes overlapping with HERVs belong to a gene.

The enrichment of genes associated with the GO term "defense response" among them might conform with prior findings. Several env genes of endogenous retroviruses have been found to play a role in host protection against viral infection in different vertebrates[55]. The mechanism of this defense, while not yet exhaustively understood, is suspected to be facilitated the binding of the env surface subunit to the host receptor ASCT2, which blocks the binding of exogenous retroviruses[55].

The distribution of expression values of HERV associated probes shows, that there certainly is expression activity within HERV elements. While the variance of expression values tends to be lower than on the entire data set, it is still enough for differential analyses.

The amount of probes measuring CpG-sites with HERV elements also was lower than would be expected, if probes were randomly distributed. This is expected to be due to probe design instead actual CpG occurrence. However an actual calculation of CpG density in HERV elements was not performed.

The high average of methylation level of HERV CpGs is in line with observations, that HERV LTRs tend to be methylated in order to silence their promoter activity[9].

The availability of genotype measurements was an exception, as the fraction of SNPs within HERV elements was slightly higher than the fraction of the genome covered by HERVs. This might be represent an actual higher density of common variants in HERV elements, as they are known to be highly degenerated and experience almost no evolutionary pressure[7].

### 5.2 Quantitative trait loci

We identified a high number of significant cis- as well as trans-eQTLs in the data set. Nevertheless, pairs of SNPs and expression probes are a lot more likely to be associated when they are in cis. This agrees with previous findings[56].

Interestingly a far higher percentage of probes taking part in cis-eQTLs at about 85% is annotated to a gene, compared to probes in trans-eQTLs at 44%.

The distributions of expression residuals for the significant eQTLs with highest p-values, shown in figure 4.6B and D, might hint towards setting a stricter cutoff. However, matrixEQTL works with continuous dosages instead of the discrete variants shown in the figure. This might disguise a more palpable association of the variant on the expression levels.

About 65% of the cis-eQTLs identified by Schramm et al[53] were reproduced in this analysis. Considering that only a subset of the data was used and the methods as well as thresholds for the calculation of significant eQTLs were different, this is a good overlap.

The number of HERV related eQTLs shows that there certainly is some variation within HERV elements. Especially the high number of expression probes significantly associated with SNPs that lie within HERV elements hints at wide reaching effects. However, eQTLs do not grant insight into the underlying mechanisms.

This high number of probes and subsequently also genes might be the reason, that the GO enrichment on genes being part of HERV related eQTLs produced no significant results. 78% of the genes that were part of any eQTL were also found in the filtered set.

The number of significant eQTLs was rather small. Just as observed in the eQTL results, there is a bias towards cis-acting associations, even though its not quite as strong.

Performing an GO biological process enrichment on the genes that were parts of HERV S2 related eQTLs, didn't result in any significantly enriched terms. This might be caused by the low number of genes among eQTL expression probes in general.

The increased proportion of low methylation beta values among CpGs that are part of eQTLs and are located within HERV S2 is very interesting. A possible explanation is, that a lower level methylation within the HERV increases the regulatory function of the HERV LTRs. The expression probes associated to these CpGs, therefore, have a changed expression.

The GO enrichment performed on HERV related meQTLs showed an enrichment of terms regarding the development nervous system and especially neurons. This is interesting as HERVs have previously been associated with neurodegenerative diseases [57], like schizophrenia[12].

## 5.3 Regulatory networks

I expected to find paths matching the pattern shown in figure 3.3 in the calculated networks. A SNP influences the expression of a neighboring gene. This gene connects to a CpG site through a route made up by transcription factors and shortest path genes. The methylation level of the CpG in turn controls the expression of a neighboring gene. Nodes not taking part the same in this functional module would be expected to not be strongly connected to any of the visited nodes. As a result there would be more than one connected component in the graph and some nodes would be totally isolated.

To be able to correctly identify indirect associations, when using partial correlations, it is necessary that all variables that mediate the effect are considered. In this case this means that a strong indirect association can only be identified as such, when all mediating entities are in the collected data set.

This, however, didn't seem to be the case in most calculated networks. The most obvious consequence could be seen for the meQTLs that were used as cornerstones of the

data collection. They were present as direct edges in almost all networks. However, it is highly unlikely that these associations represent a direct effect due to being located on different chromosomes. Therefore, it was very likely that not all mediators were in the considered data sets. Another contributing reason might be, that if the actual path the effect follows is longer than a few elements, even small errors in the data could lead to the failure to explain the indirect effect.

In many cases the SNP and CpGs were not only directly connected, but this edge was the only edge for one of them. This was the case in the network for cg06889971 (Figure 4.16).

For once the networks tended to be highly connected even after filtering edges at an edge inclusion probability cutoff as high as 0.9. This meant that in almost all cases there was one big connected component, that contained almost all nodes. Therefore, analyses on whether genes are included in the restricted graph, like GO enrichments, were not meaningful.

A potential solution to these problems could be the introduction of priors to the calculation. This might guide the iterative GGM calculation to be quicker to converge to the correct underlying interactions. I was not able to include priors due to time constraints and the troubles at finding publicly available datasets that contain expression values not only on gene but on probe level. As expression probes without a known gene made up a big part of the networks, the information to define complete priors was not available.

Their sheer size posed a major problems to the manual analysis of most networks. More sophisticated automated analyses were sorely needed, but could not be implemented due to the lack of time.



## 6 Conclusion and Outlook

The analysis of multiomics data from the KORA S4 Survey with a focus on human endogenous retroviruses showed that HERV elements have wide reaching effects in the human genome. While the array based measurements did not allow a truly genome wide analysis of the expression and methylation of HERV elements,

# 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. [8]	1
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).	8
2.5	Distribution of all methylation beta values (A) and variance (B) of 485577 CpGs between 1727 individuals.	9
3.1	(Left) Continuous time BDMCMC algorithm where $\{t_1, t_2, t_3, \dots\}$ are jumping times and $\{w_1, w_2, w_3, \dots\}$ are waiting times. (Right) Posterior probability estimation of the graphs based on the proportions of their waiting times.	17
3.2	Scheme for data collection starting from a SNP (A) and a CpG (B) within a HERV element.	18
3.3	Expected structure of regulatory networks.	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		24
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 eQTLs in the given pair of bins. A considers all eQTLs and pairs, while B is limited on the ones whose SNP and/or expression probe are connected to HERV set 2.	25
4.6	Distribution of expression residuals with respect to genotypic variant in most significant (left column) and least significant (right column) eQTLs for cis (upper row) and trans(lower row).	26
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 eQTM in the given pair of bins. A considers all eQTM and pairs, while B is limited on the ones whose CpGs and/or expression probes are connected to HERV set 2.	27
4.8		28
4.9		29

4.10	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 eQTLs 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. . . . .	30
4.11	Size and composition of biggest (left) and smallest (right) GGM data sets. The upper row are sets seeded in HERV-SNPs, while the lower row is seeded in HERV-CpGs. In few cases genes were included as SNP or CpG neighboring gene as well as transcription factor. These genes are displayed as the former. . . . .	32
4.12	Number and composition of nodes in restricted GGM networks. A and B are the ten biggest networks for GGMs seeded in SNPs, C are all networks for GGMs seeded HERV-CpGs. . . . .	33
4.13	Number of entities and size of the biggest connected component of GGMs based on SNP (A) and CpG (B) seeds. . . . .	34
4.14	Restricted network seeded in rs11079148. . . . .	35
4.15	Restricted network seeded in rs1833977. . . . .	35
4.16	Restricted network seeded in cg06889971. . . . .	36

## List of Tables

2.1	First ten rows of the RepeatMasker annotation on hg19 . . . . .	5
2.2	Covariate overview of all 3020 individuals in the KORA S4 survey (left) and the 687 individuals with available genotype, expression and methylation data (right). . . . .	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. . . . .	23
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
4.5	Significantly enriched GO biological process terms among genes containing CpG-sites participating in HERV S2 related meQTLs. . . . .	30





# References

- [1] J. Craig Venter, Mark D. Adams, Eugene W. Myers, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.
- [2] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860 EP –, Feb 2001.
- [3] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57 EP –, Sep 2012. Article.
- [4] Elizabeth Pennisi. Encode project writes eulogy for junk dna. *Science*, 337(6099):1159–1161, 2012.
- [5] 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.
- [6] Robin A. Weiss. Human endogenous retroviruses: friend or foe? *APMIS*, 124(1-2):4–10, 2016.
- [7] 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.
- [8] 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.
- [9] Zachary D. Smith and Alexander Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14:204 EP –, Feb 2013. Review Article.
- [10] 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.
- [11] Cédric Feschotte and Clément Gilbert. Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics*, 13:283 EP –, Mar 2012. Review Article.
- [12] Gorjan Slokar and Gregor Hasler. Human endogenous retroviruses as pathogenic factors in the development of schizophrenia. 6, 01 2016.
- [13] Yehudit Hasin, Marcus Seldin, and Aldons Lusi. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, May 2017.
- [14] FRANCIS CRICK. Central dogma of molecular biology. *Nature*, 227:561 EP –, Aug 1970.
- [15] Alexander F. Palazzo and Eliza S. Lee. Non-coding rna: what is functional and what is junk? *Frontiers in Genetics*, 6:2, 2015.

- [16] 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].
- [17] Almut Schulze and Julian Downward. Navigating gene expression using microarrays – a technology review. *Nature Cell Biology*, 3:E190 EP –, Aug 2001.
- [18] Andrea Piunti and Ali Shilatifard. Epigenetic balance of gene expression by polycomb and compass families. *Science*, 352(6290), 2016.
- [19] Aimée M. Deaton and Adrian Bird. CpG islands and the regulation of transcription. *Genes Dev*, 25(10):1010–1022, May 2011. 21576262[pmid].
- [20] 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].
- [21] R Smit AFA, Hubley and Green P. Repeatmasker open-4.0. 2013-2015.
- [22] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.
- [23] Lindsay L. Waite, Benjamin Weaver, Kenneth Day, et al. Estimation of cell-type composition including t and b cell subtypes for whole blood methylation microarray data. *Frontiers in Genetics*, 7:23, 2016.
- [24] 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.
- [25] Andreas Schroeder, Odilo Mueller, Susanne Stocker, et al. The rin: an rna integrity number for assigning integrity values to rna measurements. *BMC Molecular Biology*, 7(1):3, Jan 2006.
- [26] Meaghan J. Jones, Anthony P. Fejes, and Michael S. Kobor. Dna methylation, genotype and gene expression: who is driving and who is along for the ride? *Genome Biology*, 14(7):126, Jul 2013.
- [27] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, et al. Impact of cytosine methylation on dna binding specificities of human transcription factors. *Science*, 356(6337), 2017.
- [28] Jie Wang, Jiali Zhuang, Sowmya Iyer, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*, 22(9):1798–1812, Sep 2012.
- [29] Philip Machanick and Timothy L. Bailey. Meme-chip: motif analysis of large dna datasets. *Bioinformatics*, 27(12):1696–1697, 2011.
- [30] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

- [31] Aziz Khan, Oriol Fornes, Arnaud Stigliani, et al. 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.
- [32] V. Matys, O. V. Kel-Margoulis, E. Fricke, et al. Transfac(?) and its module transcompel(?): transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, Jan 2006. 16381825[pmid].
- [33] Aurelien Griffon, Quentin Barbier, Jordi Dalino, et al. Integrative analysis of public chip-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Research*, 43(4):e27, 2015.
- [34] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357 EP –, Mar 2012.
- [35] Yong Zhang, Tao Liu, Clifford A. Meyer, et al. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137–R137, Sep 2008. gb-2008-9-9-r137[PII].
- [36] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, et al. String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, 41(Database issue):D808–D815, Jan 2013. gks1094[PII].
- [37] Damian Szklarczyk, John H. Morris, Helen Cook, et al. 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].
- [38] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45:580 EP –, May 2013.
- [39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
- [40] Michael Lawrence, Wolfgang Huber, Hervé Pagès, et al. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013.
- [41] Mark Dunning, Andy Lynch, and Matthew Eldridge. *illuminaHumanv3.db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3)*, 2015. R package version 1.26.0.
- [42] Tim Triche, Jr. *FDb.InfiniumMethylation.hg19: Annotation package for Illumina Infinium DNA methylation probes*, 2014. R package version 2.2.0.
- [43] Andrey A. Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.
- [44] 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.
- [45] Hervé Pagès, Marc Carlson, Seth Falcon, and Nianhua Li. *AnnotationDbi: Annotation Database Interface*, 2017. R package version 1.38.1.

- [46] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.
- [47] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [48] Abdolreza Mohammadi and Ernst C. Wit. Bdgraph: An r package for bayesian structure learning in graphical models. 2015.
- [49] Abdolreza Mohammadi and Ernst C. Wit. Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.
- [50] 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.
- [51] Alex Lenkoski. A direct sampler for g-wishart variates. *Stat*, 2(1):119–128.
- [52] Juliane Schäfer 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.
- [53] Katharina Schramm, Carola Marzi, Claudia Schurmann, et al. Mapping the genetic architecture of gene regulation in whole blood. *PLOS ONE*, 9(4):1–13, 04 2014.
- [54] Inc Illumina. Data sheet: Humanht-12 v3 expression beadchip, 2010.
- [55] Ray Malfavon-Borja and Cédric Feschotte. Fighting fire with fire: Endogenous retrovirus envelopes as restriction factors. *J Virol*, 89(8):4047–4050, Apr 2015. 03653-14[PII].
- [56] Roby Joehanes, Xiaoling Zhang, Tianxiao Huan, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biology*, 18(1):16, Jan 2017.
- [57] Kristien Mortelmans, Feng Wang-Johanning, and Gary L. Johanning. The role of human endogenous retroviruses in brain development and function. *APMIS*, 124(1-2):105–115.