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: 15.03.2018

Contents

1 Introduction

1.1 Regulation of cell functions and Epigenetics

- central dogma of biology
- rising importance of other factors atop DNA →epigenetics
- quick overview epigenetic marks
- chromatin states
- DNA methylation
- snp -¿ cpg -¿ expression pattern
- TF ¡-¿ cpg interaction

1.2 Human endogenous retroviruses

- first humane genome -¿ "junk DNA"
- ongoing discovery for non-coding regions
- still masking of difficult sequence for many analysis -¿ repeats
- repeat classes -¿ ... -¿ herv
- herv origin - ...-virus like
- herv structure: LTR - pol - env - ... - LTR
- discovered roles of hervs in general regulation/diseases
-

1.3 Effect network analysis

- many bioinformatics methods find correlations, but not direct cause
- attempt to discern direct connections from bigger data webs
- hope to find possible biological mechanisms of gene regulation = path in model
- used approach: Gaussian Graphical models

bin	swScore	milliDiv	milliDel	milliIns	genoName	genoStart	genoEnd	genoLeft	strand	repName	repClass	repFamily	repStart	repEnd	repLeft	id
585	1504	13	4	13	chr1	10000	10468	-249240153	+	(CCCTAA)n	Simple_repeat	Simple_repeat	1	463	0	1
585	3612	114	270	13	chr1	10468	11447	-249239174	-	TAR1	Satellite	telo	-399	1712	483	2
585	437	235	186	35	chr1	11503	11675	-249238946	-	L1MC	LINE	L1	-2236	5646	5449	3
585	239	294	19	10	chr1	11677	11780	-249238841	-	MER5B	DNA	hAT-Charlie	-74	104	1	4
585	318	230	38	0	chr1	15264	15355	-249235266	-	MIR3	SINE	MIR	-119	143	49	5
585	203	162	0	0	chr1	16712	16749	-249233872	+	(TGG)n	Simple_repeat	Simple_repeat	1	37	0	6
585	239	338	148	0	chr1	18906	19048	-249231573	+	L2a	LINE	L2	2942	3104	-322	7
585	652	346	85	42	chr1	19947	20405	-249230216	+	L3	LINE	CR1	3042	3519	-970	8
585	270	331	7	27	chr1	20530	20679	-249229942	+	Plat_L3	LINE	CR1	2802	2947	-639	9
585	254	279	47	39	chr1	21948	22075	-249228546	+	MLT1K	LTR	ERVL-MaLR	15	142	-453	1

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

2 Data

2.1 HERV annotation

HERV annotation was pertained from RepeatMasker[1] 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 elements from the RepeatMasker library for human genome hg19 was downloaded from the UCSC genome browser download section (`http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz`). It contains a total of 5298130 occurrences of repeats. Each entry consists of 17 values. Not in order 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 occurence. Furthermore it contains the position in the whole sequence, that is known for the identified repeat, that is covered by the occurrence. It also describes the quality of the alignment of the repeat sequence to the annotated position using the Smith Waterman alignment score[2] and the number of base mismatches, deletions and insertions per thousand base pairs. Finally there is an indexing field used to speed aup chromosme 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 **??**.

To extract HERV elements the annotation was filtered on different columns generating three sets variable size. Multiple HERV sets was created as an atempt to cover different possible definitions of what exactly a HERV element is.

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 35589 elements after merging. The elements have a mean width of 949 bp and cover a total of 33.8 MB, which is ca 1.04% of the human genome. The distribution of element lengths in HERV S1 is shown in Figure 2.1

Alternatively filtering the annotation for "ERV" in the superfamily column leads to 696689 annotations. HERV set 2 (HERV S2) contains all endogenous retroviral sequences found in the human genome. After merging overlapping and adjacent annotations this led to 633323 elements. Their mean width is 415 bp and they make up to 262.8 MB or ca 8.13% of the human genome.

A third set "HERV set 3" (JERV S3) was constructed by filtering the repeat name column for "HERV", which resulted in 21361 annotations. As this set contains only

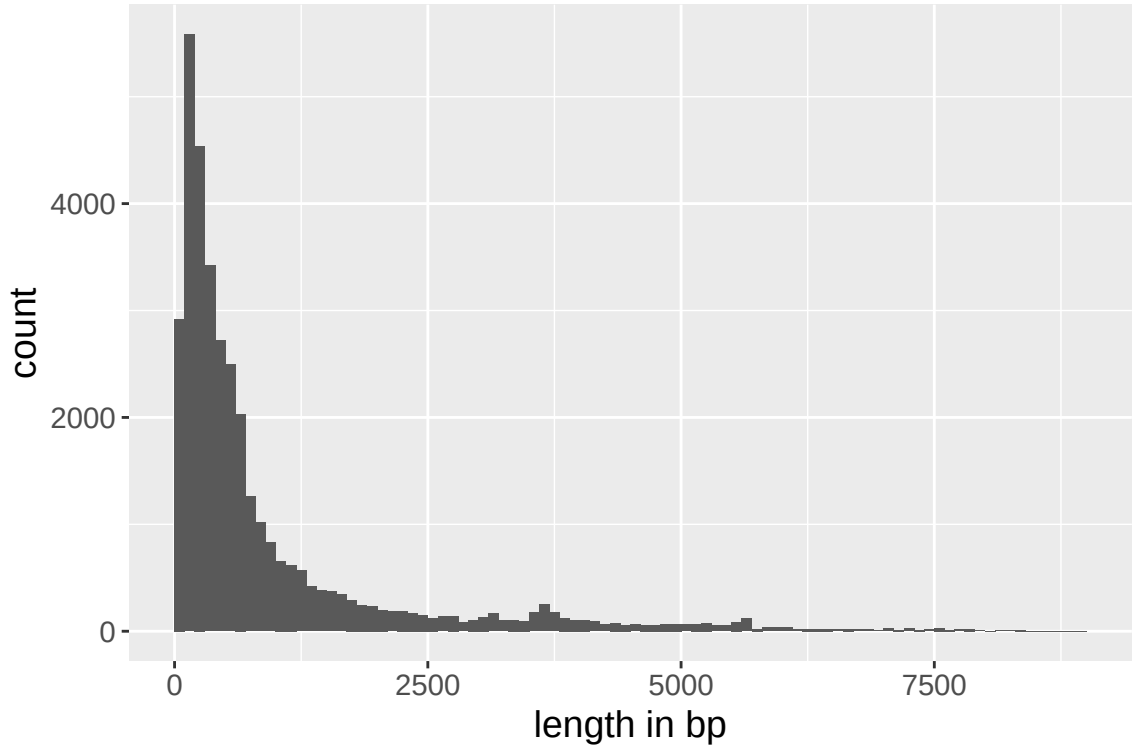

Figure 2.1: Length distribution of elements in HERV S1

elements that are explicitely named "HERV", it is my hope that might contain only endogenous retroviral elements, that were inserted directly into the human genome. After merging overlapping and adjacent annotations there were 15403 element with an average width of 1468 bp and a combined length of 22.6 MB.

2.2 KORA

The Expression, Methylation and Genotype data used in this work were generated by 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 chronical diseases.

The data comes from the Survey F4, which was conducted from 2006 to 2008 and comprised samples of 3080 individuals. F4 is a follow up study to the survey S4 performed from 1999 to 2001 and containing 4261 individuals.

All measurements were performed on whole blood samples. Houseman blood counts[] describing the composition of different cell types for each individual are available.

Not all essays are available for all samples. Therefore different analyses were performed on varying sets of indiviudals according to availability of the required data types.

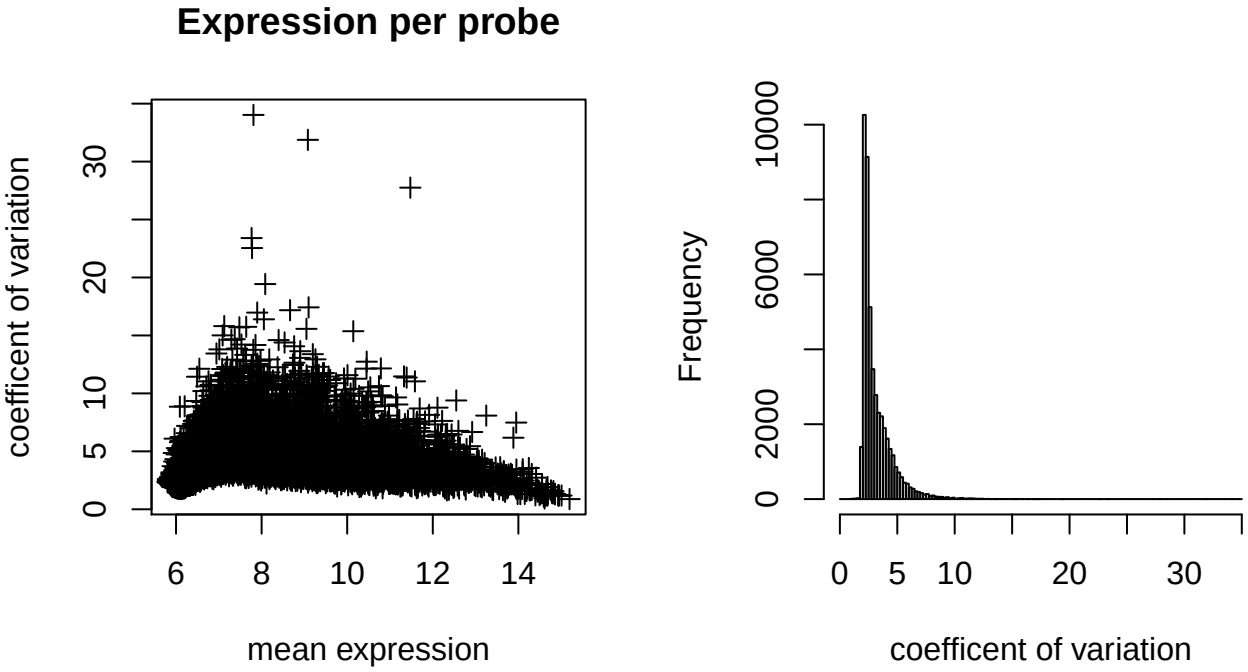

Figure 2.2: Coefficient of expression variance over 993 individuals

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.

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

To not lose information, especially in hERV regions that are usuallly sparesly annotated with genes, probes without genes were not discarded and most analyses were performed on probe level or only partially abstracting to gene level.

2.2.2 Methylation

DNA methylation was measured using the Infinium HumanMethylation450K BeadChip, which interrogates methylation levels at 485577 genomic locations.

Methylation data was available for 1727 individuals and 485512 sites, which make up all 'cg' and 'ch' probe type probes.

2.2.3 Genotypes

Genotyping was performed with the Affymetrix Axiom array. The Iluminus calling algorithm was used for genotype calling and missing values were imputed using the IMPUTE2

software[3]. SNPs were filtered at IMPUTE value of 0.4.

After excluding all SNPs with a minor allele frequency of less than one percent measurements of 9533127 SNPs for 3788 individuals were available.

2.2.4 Covariates

Several covariates were known for each sample. These were age, sex, body mass index (BMI) and wide blood cell count, as well as experimental factors like storage time and RNA integrity number (RIN).

2.2.5 Methylation quantitative trait loci

Previously process methylation quantitative trait loci (meQTL) data was used. The data set contained a total of xxxxxx significantly associated cpg-snp pairs. xxxx distinct cpg-sites and xxxxxx snps were part of at least one meQTL. xxxx pairs consisted of cpgs and snps on the same chromosome, while xxxxx association were between different chromosomes.

2.3 Transcription factor binding

Transcription factor binding sites were obtained from two publicy available sources:

First was the third version of the track "Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs"[4] downloaded from the UCSC genome browser download section (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz`).

It combines 690 high quality ENCODE ChIP-seq data sets, which were processed with the Factorbook motif discovery and annotation pipeline[4]. The pipeline uses the tools MEME-ChIP[5] and FIMO[6] from the MEME software suite and merges discovered motifs with known motifs from Jaspar[7] and TransFac[8] using machine learning methods and manual curation.

The track contains a total of 438044 distinct peaks for 161 transcription factors in 91 cell types. For our analyses we filtered and combined the peaks for 23 blood related cell types. This leaves a total of 2173371 peaks for 125 transcription factors.

The second source was the ReMap project[9]. It combines 395 publicy available ChIP-seq data sets covering 132 different transcription factors across 83 cell lines. ReMap uses Bowtie2[] to map reads to the human genome and the tool MACS[] for peak calling. The finished data set was downloaded from the ReMap website(`http://tagc.univ-mrs.fr/remap/download/All/filPeaks_public.bed.gz`).

It contains xxxxx peaks. After filtering for 19 blood related cell types a of xxxx peaks of xxx different transcription factors remained.

Combining both filtered data sets lead to a total of xxxxx peaks of xxx transcription factors.

2.4 Chromatin states

Chromatin state annotations were downloaded from Roadmap Epgenomics Core 15-state model`http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/`

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

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

In this work the annotations for 27 blood related cell lines were used. To get the distribution of states for a single feature or a set of features the different annotations for the different cell lines were weighted according to the houseman counts and summed up.

3 Methods

3.1 Overlaps

The task of whether HERV elements contained any annotated elements like expression probes, genes, cpg sites and transcription factor binding sites (tfbs) was performed using funtion "findOverlaps" from the Bioconductor package "GenomicRanges"[11].

Features were considered to be of interest for the analysis of an HERV element, if they overlapped by at least one base pair.

Features were also filtered for the HERV elements and their 1kb or 2kb up- and downstream regions.

3.2 Data normalization

Expression and methylation values were corrected for available covariates by calculating the residual matrix. A linear model containing the cell compositions and the first 20 principal components was used for methylation.

Expression residuals were calculated using a linear model of age, sex, RNA integrity number (RIN), plate and storage time.

3.3 eQTL/eQTM calculation

Expression quantitative trait loci (eQTL) and expression quantitative trait methylation were calculated using the Bioconductor package MatrixEQTL[12]. MatrixEQTL tests for association of SNP-transcript pairs. It offers two modes of modeling the effect of the genotype on transcription levels:

When setting the parameter "$useModel = modelLINEAR$", as was done in this work, an additive linear model is used. The association ism modeled as simple linear regression and the absolute value of the sample correlation is used as test statistic.

Alternatively when choosing "$modelAnova$" for the paramter, the effect is modeled with ANOVA model. In this case the test statistic is the squared sample correlation.

After calculating the test statistics the p-values for the all pairs that pass a defined significance threshold are calculated. These are corrected multiple testing using a Benjamini-Hochberg procedure, adapted for not recording all p-values.

MatrixEQTL is very performant because it manages to reduce the calculation of the test statistic to one single large matrix multiplication by cleverly transforming the genotype and transcription variables.

MatrixEQTL also allows to include covariates in the QTL calculation. As the expression and methylation values used are residuals and therefore already consider covariates this option is not used. Furthermore MatrixEQTL can differentiate between cis- and trans-interactions. The maximal distance to consider a pair on the same chromosome as cis was set to 50kb.

The threshold for significant cis-QTLs during calculation was set to 10e-6 and 10e-8 for trans.

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[13].

Then a hypergeometric test[14] for overreprensentation is performed on a set of genes of interest. For most enrichments a custom backround set of genes specific to the analysis is given. Finally the p-values for overrepresented GO terms were adjusted for multiple testing using the Holm method[15].

3.5 Gaussian Graphical Models

4 Results

4.1 Normalized Data

The distribution of the quantile normalized expression values and the expression residuals can be seen in figure **??**. As expected it follows a normal distribution. This is important for the calculation of the Gaussian graphical models, as it's one of base assumptions.

- raw var vs residual var
-

4.2 HERV region features

4.2.1 Expression

Using hERV set 1 there are a total of 191 overlaps of at least 1 bp between a hERV element and a region measured in the expression array. 174 hERV elements overlap with 188 different expression probes.

The expression probes have 1338 overlaps with the annotated elements in hERV set 2. 1274 hERV elements overlap with 1317 different expression probes.

The coefficents of variance for the expression probes that overlap with the hERV sets are shown in Figure **??**

When inspecting not only direct overlaps, but the region of +/- 1kb around the hERV elements, there are 517 overlaps (476 hERV elements, 349 probes) for set 1 and 6812 overlaps (6336 hERV elements, 4712 probes) for set 2.

Enlarging the flanking regions to 2kb leads to 973 (870 hERV elements, 548 probes) and 13398 (12201 hERV elements, 8044 probes) overlaps for set 1/2 respectively.

4.2.2 Methylation

Using hERV set 1 there are a total of 1602 overlaps of hERV elements and measured methylation sites. 1021 hERV elements overlap with 1595 different methylation sites.

hERV set 2 has 17162 overlaps. These are constituted by 13141 hERV elements and 17137 methylation sites.

Including the 1kb flanking regions of the hERV elements leads to 6785 overlaps (3470 hERV elements, 4497 methylation probes) for set 1 and 119763 overlaps (66249 hERV elements, 78501 methylation probes) for set 2.

With a flanking region of 2k this increases to 13559 overlaps (5645 hERV elements, 7792 methylation probes) for set 1 and 259739 overlaps (110524 hERV elements, 139036 methylation probes) for set 2.

4.2.3 Genotypes

Measurements for a total of 80754 SNPs within hERV set 1, that occurr in at least one sample, are available.

4.2.4 Chromatin states

4.3 eQTLs

4.4 eQTMs

4.5 hERV realated regulatory networks

4.5.1 Data collection

5 Discussion

6 Conclusion

References

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

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

[3] 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.

[4] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J. Rando, Ewan Birney, Richard M. Myers, William S. Noble, Michael Snyder, and Zhiping Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*, 22(9):1798–1812, Sep 2012.

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

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

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

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

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

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

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

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

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

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

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