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

2 Data

2.1 hERV annotation

hERV annotation was pertained from RepeatMasker human annotation downloaded from the UCSC genome browser download section (`http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz`).

The annotation then was filtered for elements containing "ERV" in the repeat name column. This lead to a total of 42508 elements with a mean width of 796 bp covering a total of 33.9 MB. This set will be called "hERV set 1", or short hERV s1.

Alternatively filtering the annotation for "ERV" in the superfamily column leads to 696689 elements. They have a mean width of 379 bp and cover 263.7 MB. (hERV set 2/hERV s2)

A third set "hERV set 3" (hERV s3) was constructed by filtering the repeat name column for "hERV", which resulted in 21361 elements.

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.

2.2.1 Expression

The expression data was generated using the HumanHT-12 v3.0 Gene Expression Bead-Chip. 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.
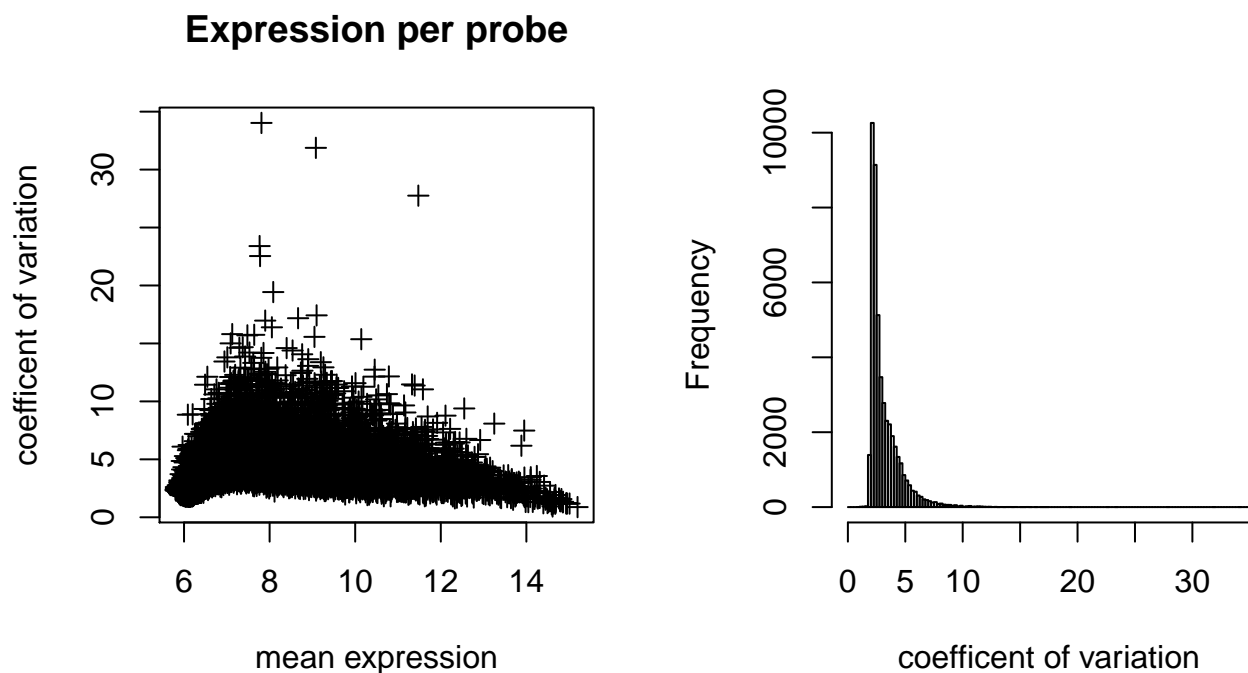
Expression per probe

Figure 2.1: Coefficient of expression variance over 993 individuals

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[1]. 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"[2] 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[2]. The pipeline uses the tools MEME-ChIP[3] and FIMO[4] from the MEME software suite and merges discovered motifs with known motifs from Jaspar[5] and TransFac[6] 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[7]. 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[8] 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

Overlaps between HERV elements and annotated elements like expression probes, cpg sites and transcription factor binding sites (tfbs) were calculated using the Bioconductor package "GenomicRanges"[9]. Two entities were considedered overlapping, if they share at least one base pair.

All overlaps were also generated for the herv elements plus 1kb and 2kb flanking regions.

3.2   Data normalization

- calculate residuals with linear regression to select covariates

- expression:

- methylation:

3.3   eQTL/eQTM calculation

- R package MatrixEQTL

- determines correlation between snp and expression values by linear regression model

- reduction to simple matrix multiplications
    - mean -¿ 0
    - variance -¿ 0-1
    - allows fast calculation

- paramters:
    - model: linear
    - thresholds: cis = 10e-6, trans 10e-8
    - cis-dist: 5e5

3.4   Functional Analysis of Gene Sets

- packages GSEABase, GOstats

- generate GeneSetCollection for Gene Symbol -¿ GO term, for all pairs with evidence

- test vor overrepresentation of terms in set of genes against background set using Hypergeometric test (hyperGTest)

3.5   Gaussian Graphical Models

- I'm fucking afraid of writing this part...

4 Results

4.1 Normalized Data

- 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 sample, are available.

4.2.4   Chromatin states

4.3   eQTLs

4.4   eQTMs

4.5   hERV realated regulatory networks

5 Discussion

6   Conclusion

References

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

[2] 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.

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

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

[5] 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.

[6] 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].

[7] 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.

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

[9] 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.