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 sparsely annotated with genes, probes without genes were not discarded and most analyses were performed on probe level or only partially abstracting to gene level.
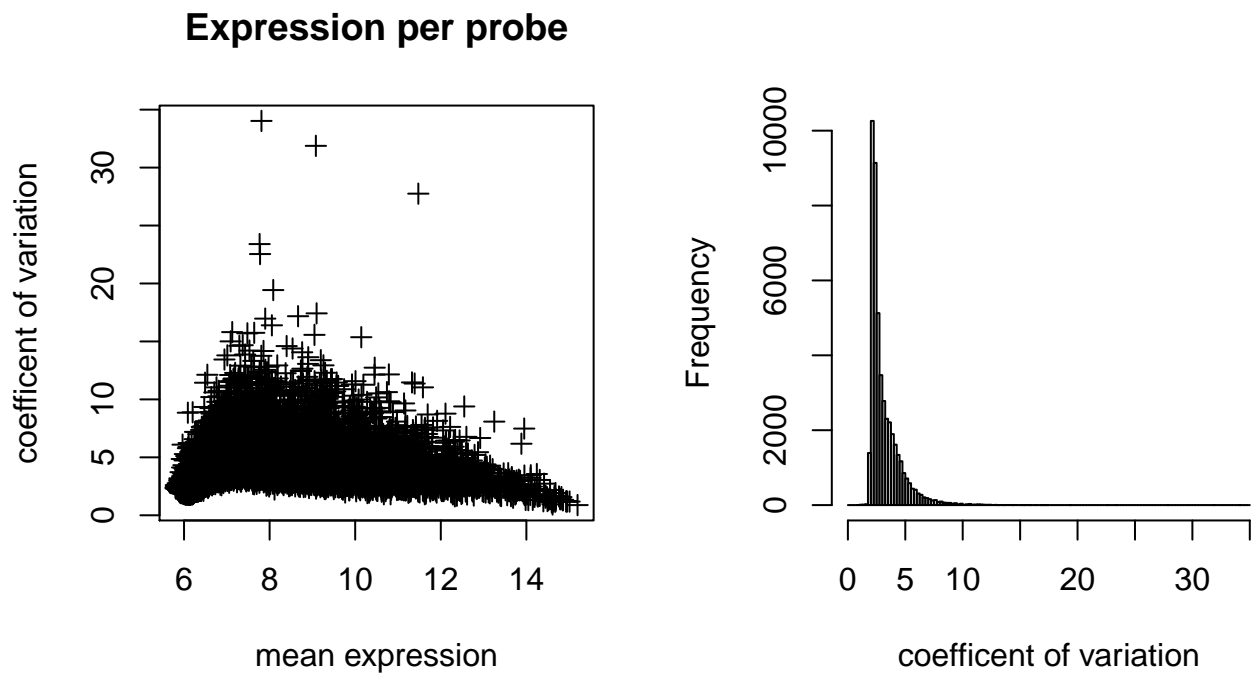
Expression per probe

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 imputed using the IMPUTE2 software. 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 xxxx 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 data was used. There were a total of xxxxxx significantly associated cpg-snp pairs. Of these xxxx were cis-interactions and

xxxxx were trans.

2.3   Transcription factor binding

2.4   chromHMM

- chromatin modification as interesting/highly invested epigenetics mark

- HMM used to predict 15 state model

- use data for 27 blood cell types

- sum up with houseman counts as weights for joint analysis

3   Methods

3.1   Overlaps

- R GenomicRanges package

- overlap = at least 1 bp of either element overlapping

- direct and +1/2kb flanking region

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.

5 Discussion

6 Conclusion