# Personal genome outline

# 

Goals of the personal genome paper:

a) show the value of diverse data, what you can/can’t one get using the reference vs using the personal genome

b) define a new paradigm on how integrate data using personal genome sequencing.

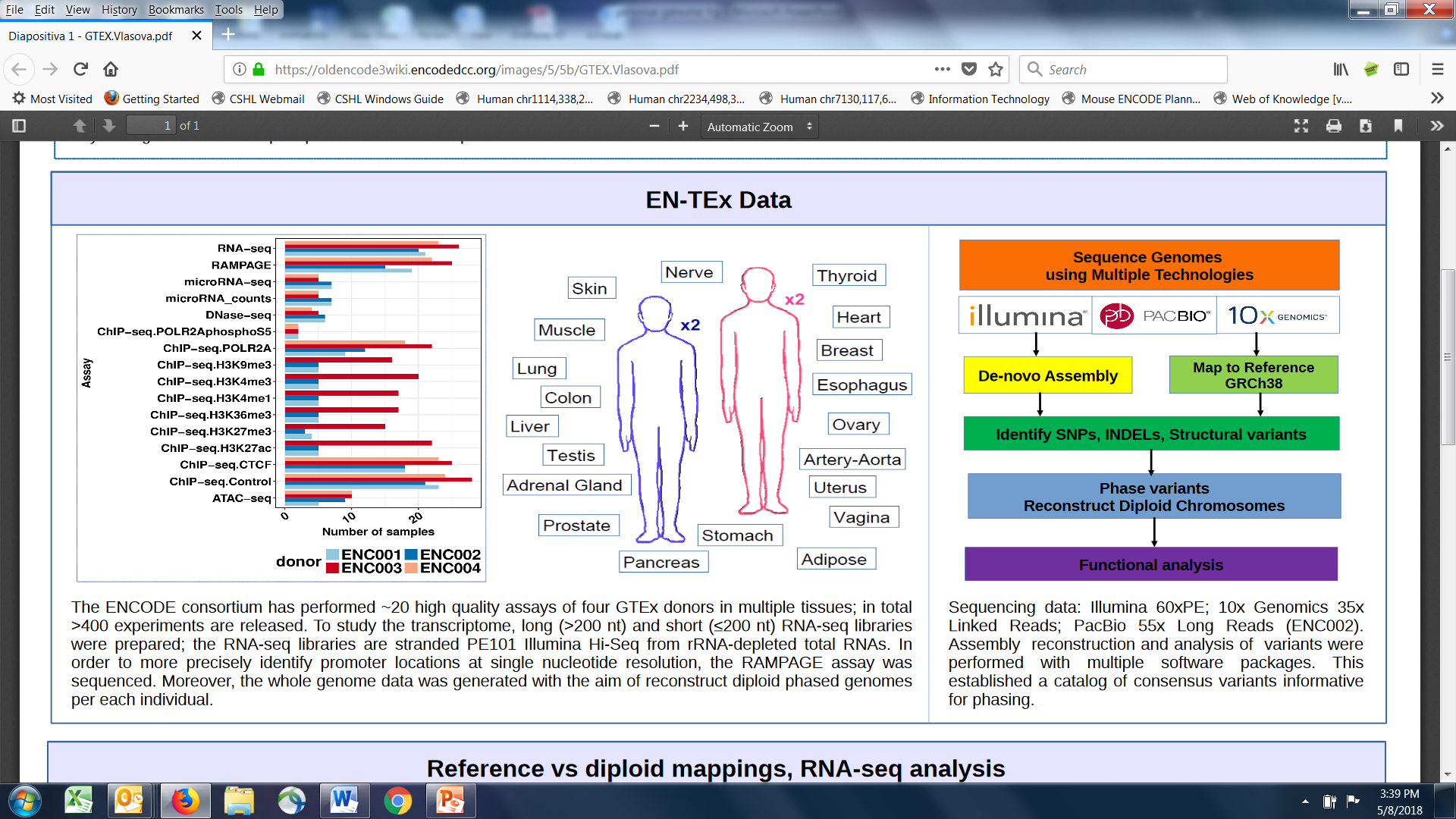
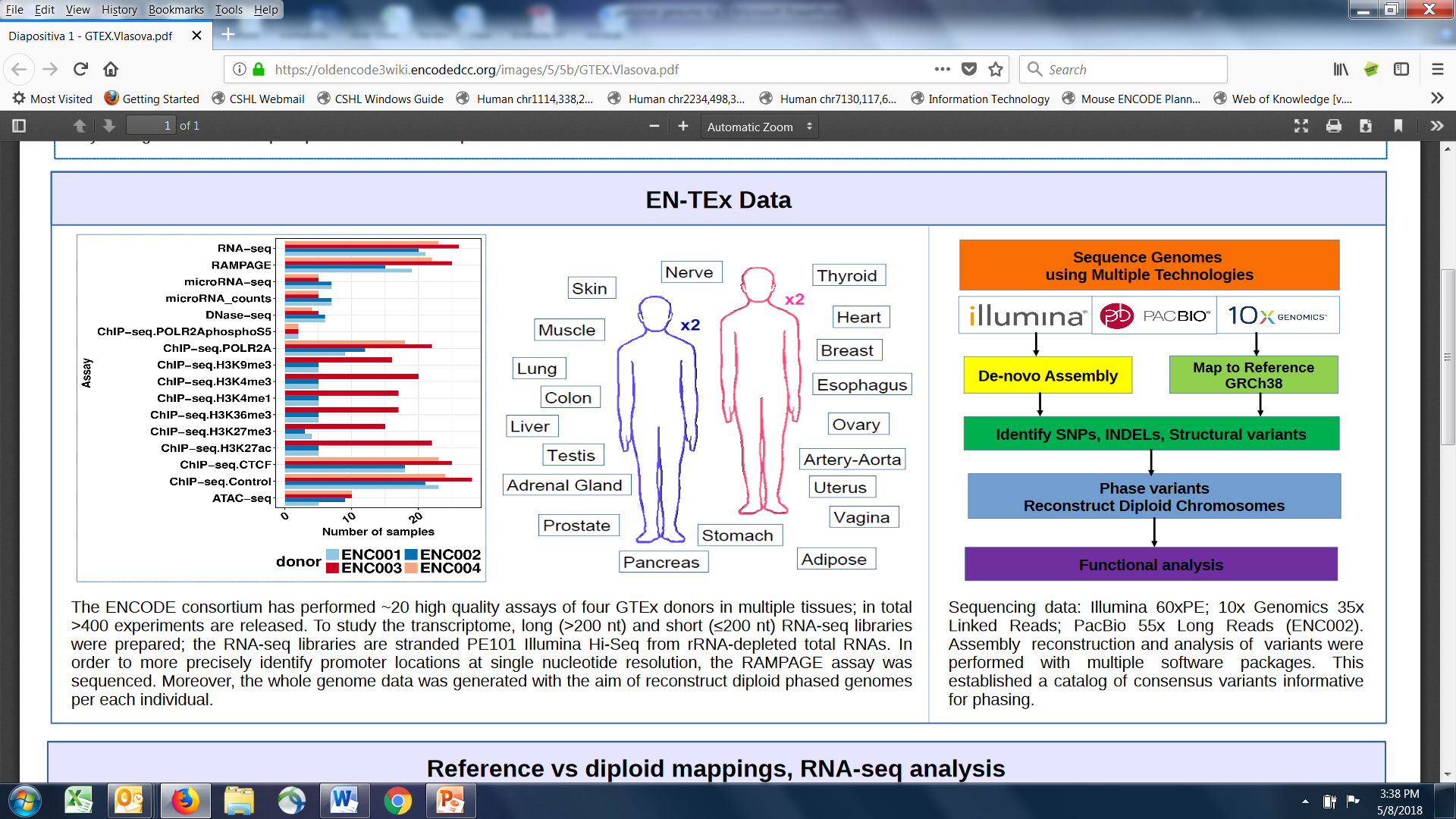
Values in Building Personal Phased Diploid Genomes Using Multiple Data Types  
Target: NBT Article: 4000 words, 6 display figures   
https://www.nature.com/nbt/about/content  
  
Abstract (150 words)  
Most current transcriptome and other functional genomics studies begin by mapping sequencing data to a standard haploid reference genome. While this approach has been very effective for measuring major trends to gene expression and regulation across individuals and tissue types, this approach does not account for the multiple and distinguishing genomic differences that defines each individual. Importantly the ability to phase the variants found in each genome is lost. Consequently, the community suffers from limited power to study the role of genomic variation on such effects as allele specific binding or allele specific expression modulated by distant cis-regulatory elements. To address this critical need, as part of the ENCODE project and in collaboration with the GTEx consortium we have deeply sequenced the genomes, analyzed the transcriptional and epigenetic profiles of 4 EN-TEx human samples using a combination of genomic technologies. Using combinations of these genomic data in the context of alignment-based and de novo assemblies, millions of single nucleotide and short indel variants per genome have been identified as well as thousands of larger structural variations establishing high quality phased personalized genomes, with average phase block sizes approaching 1 million bp. When combined with the varieties of RNA-seq, ChIP-Seq, and other functional genomics data also available for each of these samples, these data rich phased personal genomes provide insights not available in using consensus genome assemblies and a foundation for exploration into the interplay between variation, expression, and regulation with allele-specific resolution.  
  
  
Introduction (~1-2 pages)  
  
- Brief history of human genome reference  
- Needs for personalized genome  
- Introduce data, methods, results presented in paper  
  
**Results**  
  
**Section 1. Introduction of Entex datasets**  


Figure 1. RNAseq reads distribution by genomic elements. (DCC long RNAseq pipeline, GRCh38, gencode24)  
  
  
**Section 2. Personalized Genome Construction**  
  
- Properties of short, linked, and long reads  
 - De novo assembly versus mapping approaches for variant detection  
 - Assessment of different technologies for phasing   
- Discussion of how variants are detected, phased, and incorporated into personalized genome  
 - Accuracy / Completeness benchmark versus de novo approaches  
 - Transfer of annotation from reference to personalized genome  
- Variants in a typical human genome  
 - Breakdown of SNPs, indels, and SVs  
 - Size distribution, overlap & extension of 1000 genomes SVs  
 - Overlap of SVs in genes, regulatory features and other annotated sequences

|  |  |
| --- | --- |
| Figure 2a. Comparison of SV Calling Approaches | Table 2b. Summary of Variants per person |
| Figure 2c. Enrichment and Suppression of Variants in different annotated regions | |

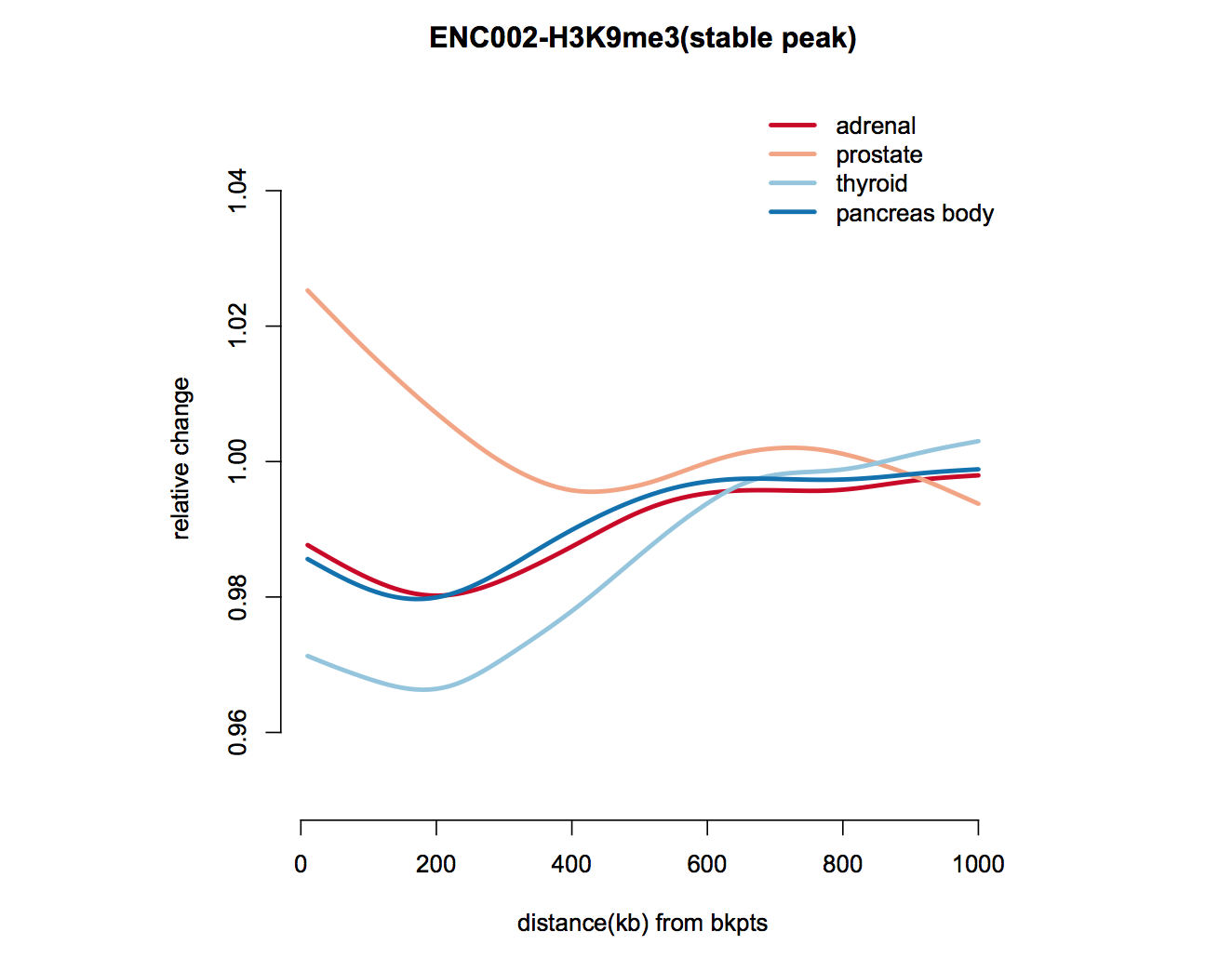
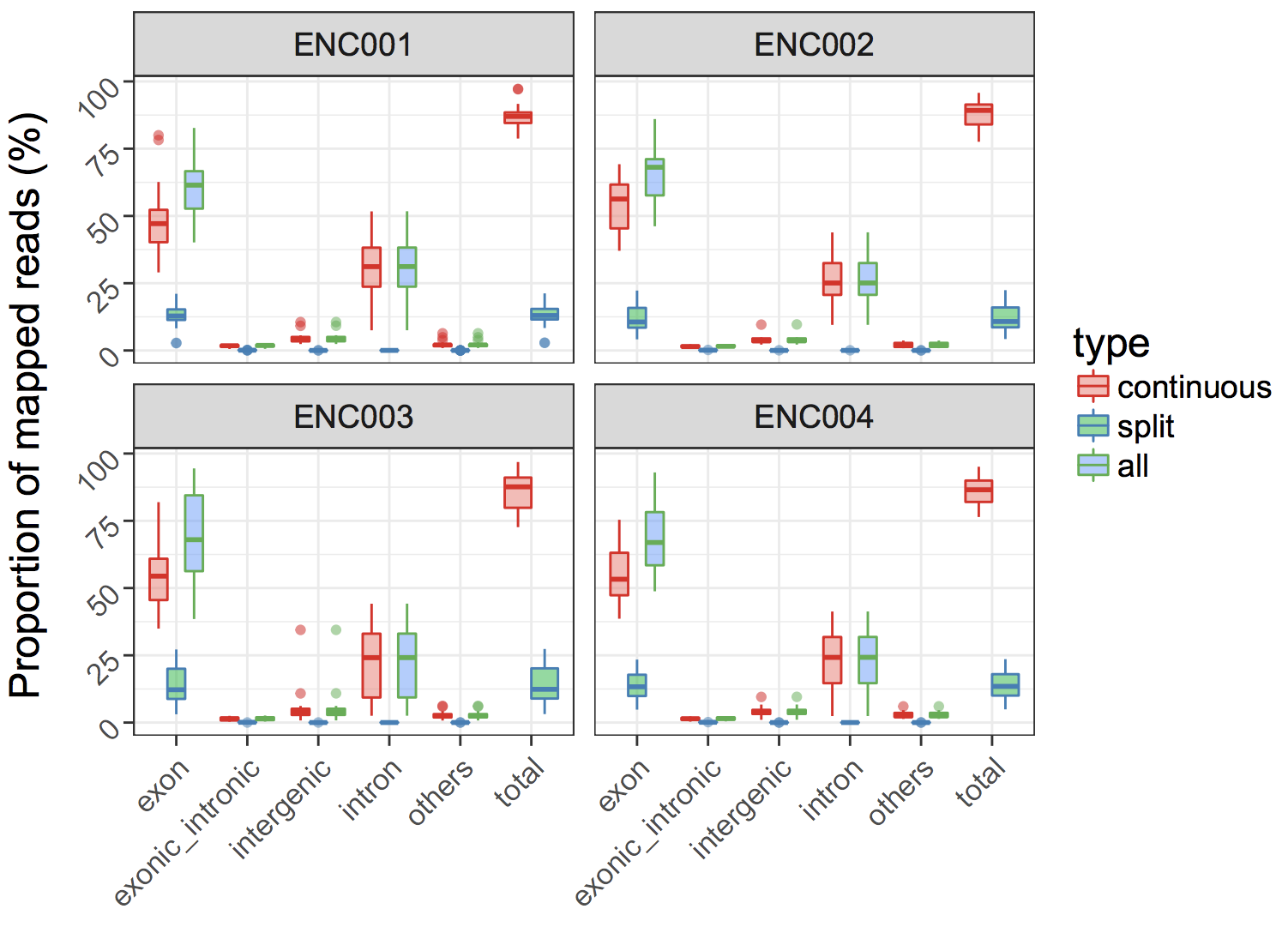
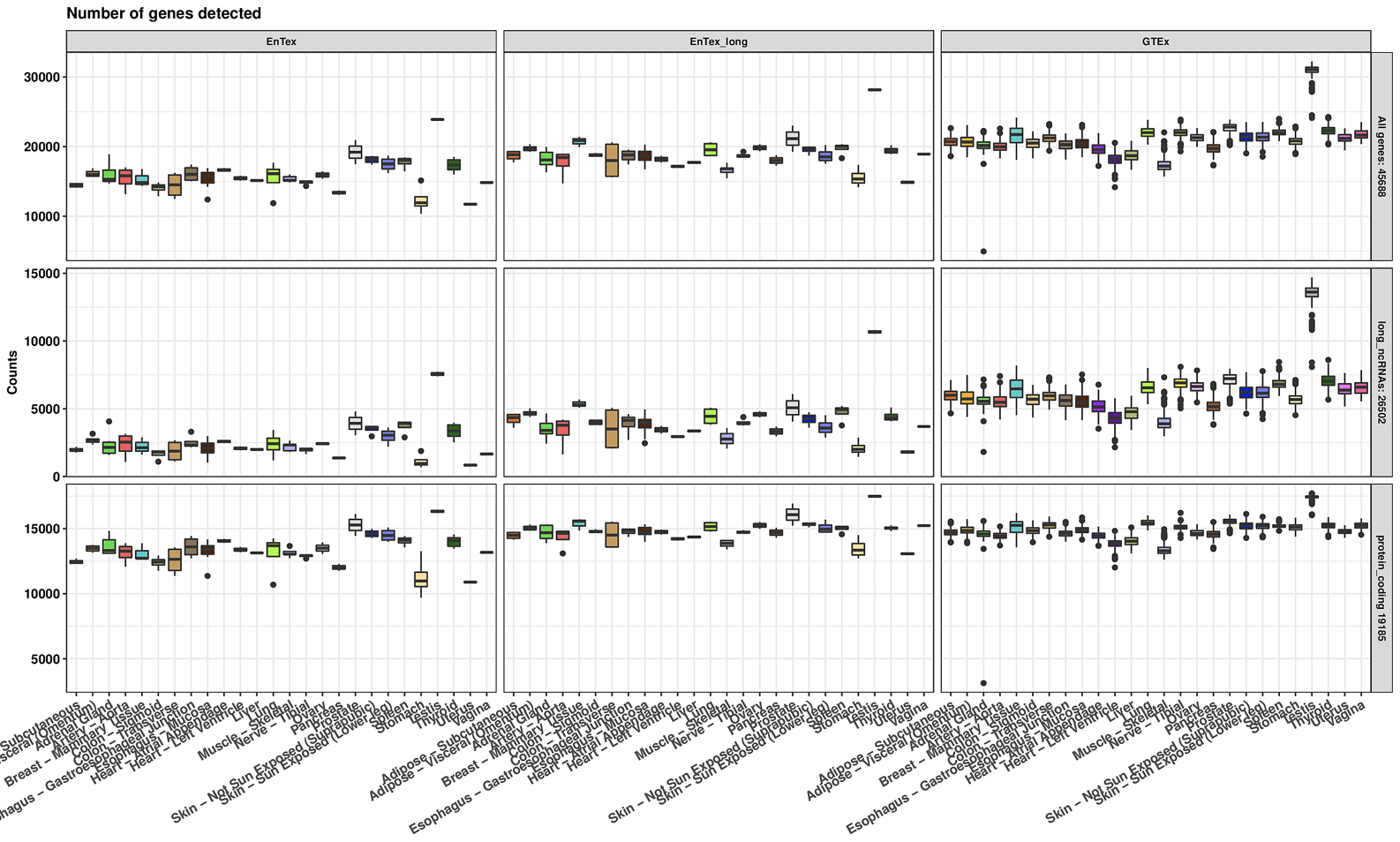
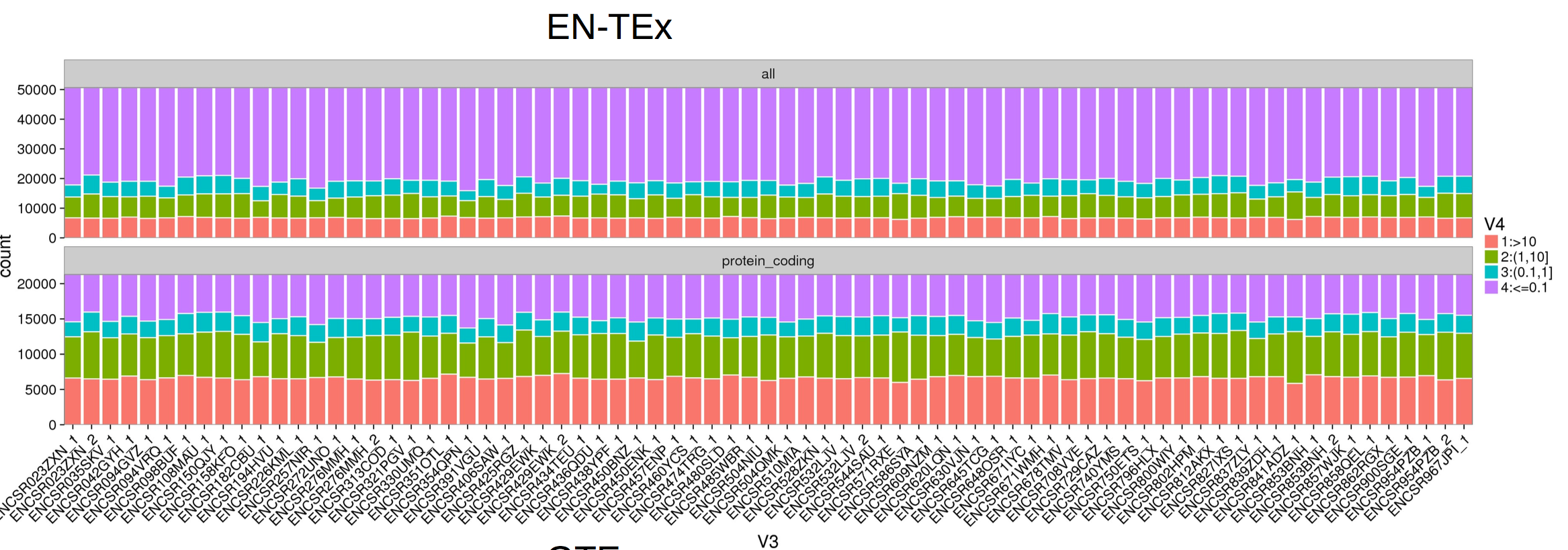
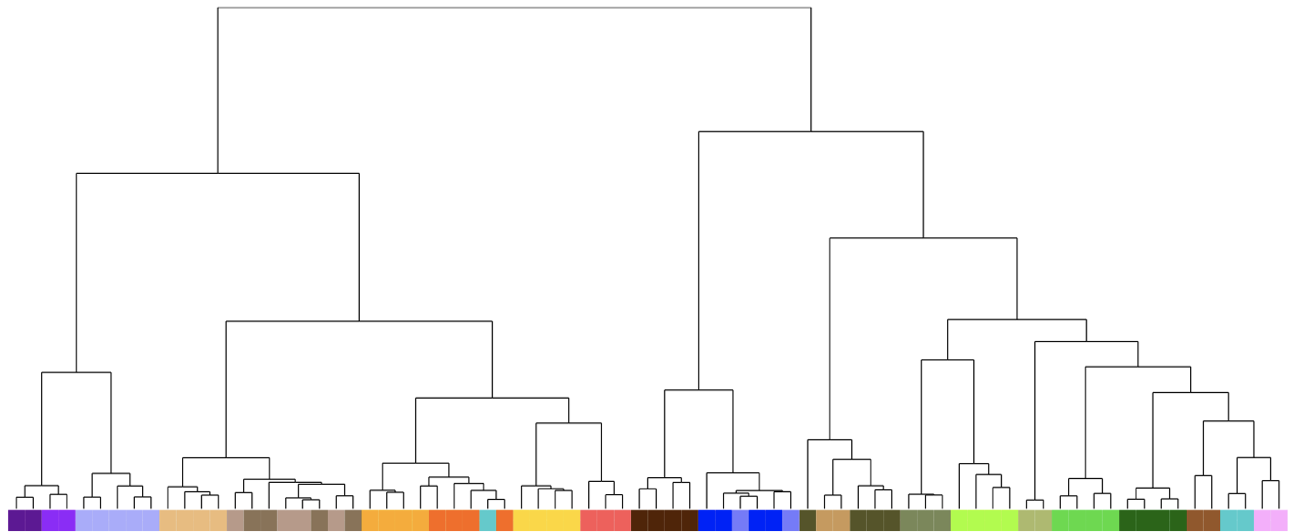


Figure 2d. Cumulative analysis of histone modifications flaking SVs  
  
**Section 3. Personal Genome Expression Analysis**- Improvement to RNA-seq mapping, “differential expression” compared to reference

- Expression of genes spanning and near SVs







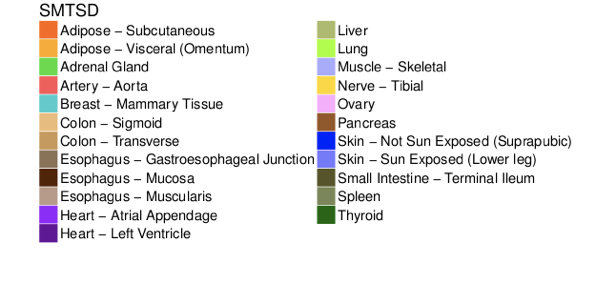


Figure5. Hierarchical clustering EnTex samples, long genes

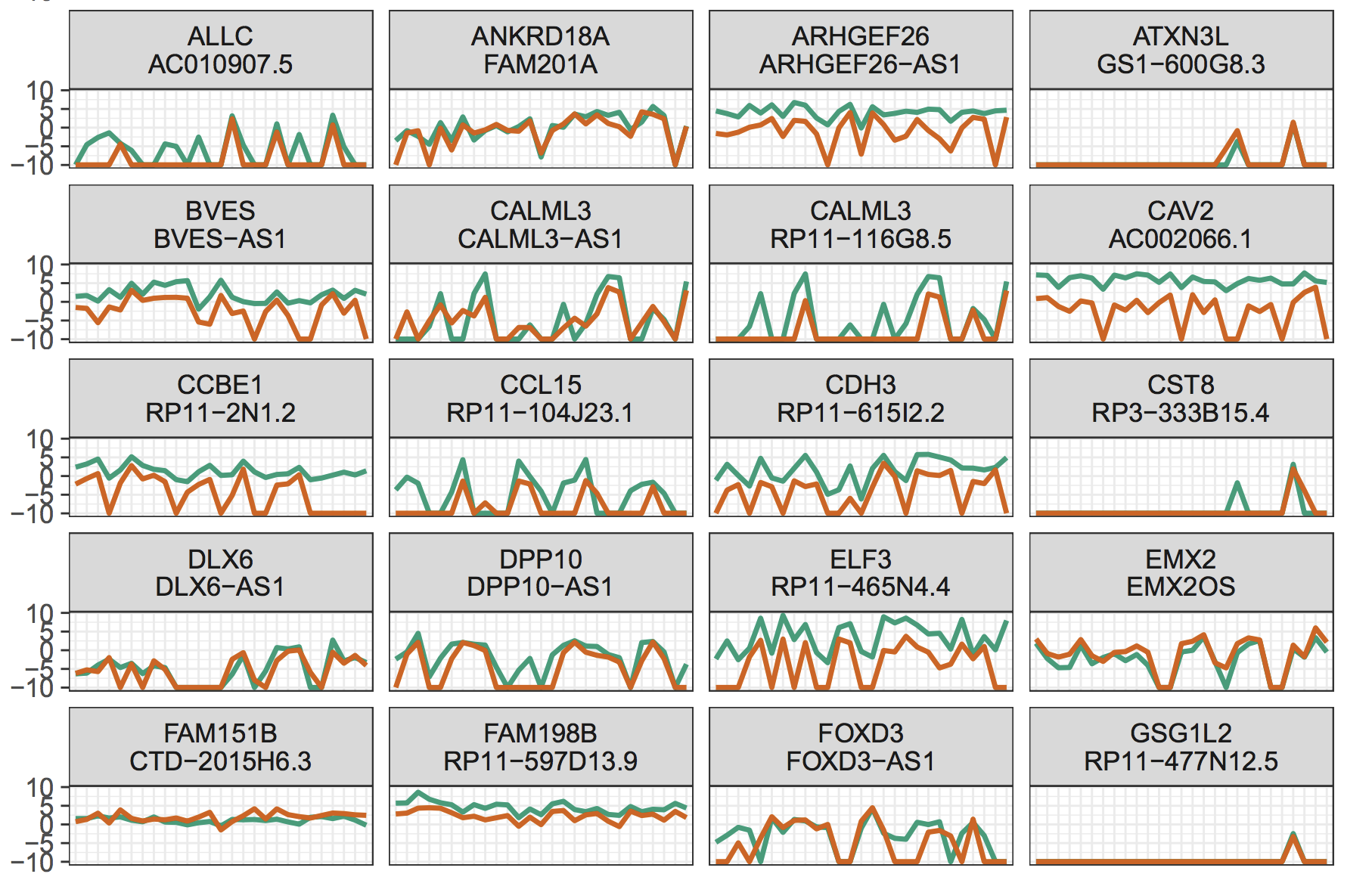


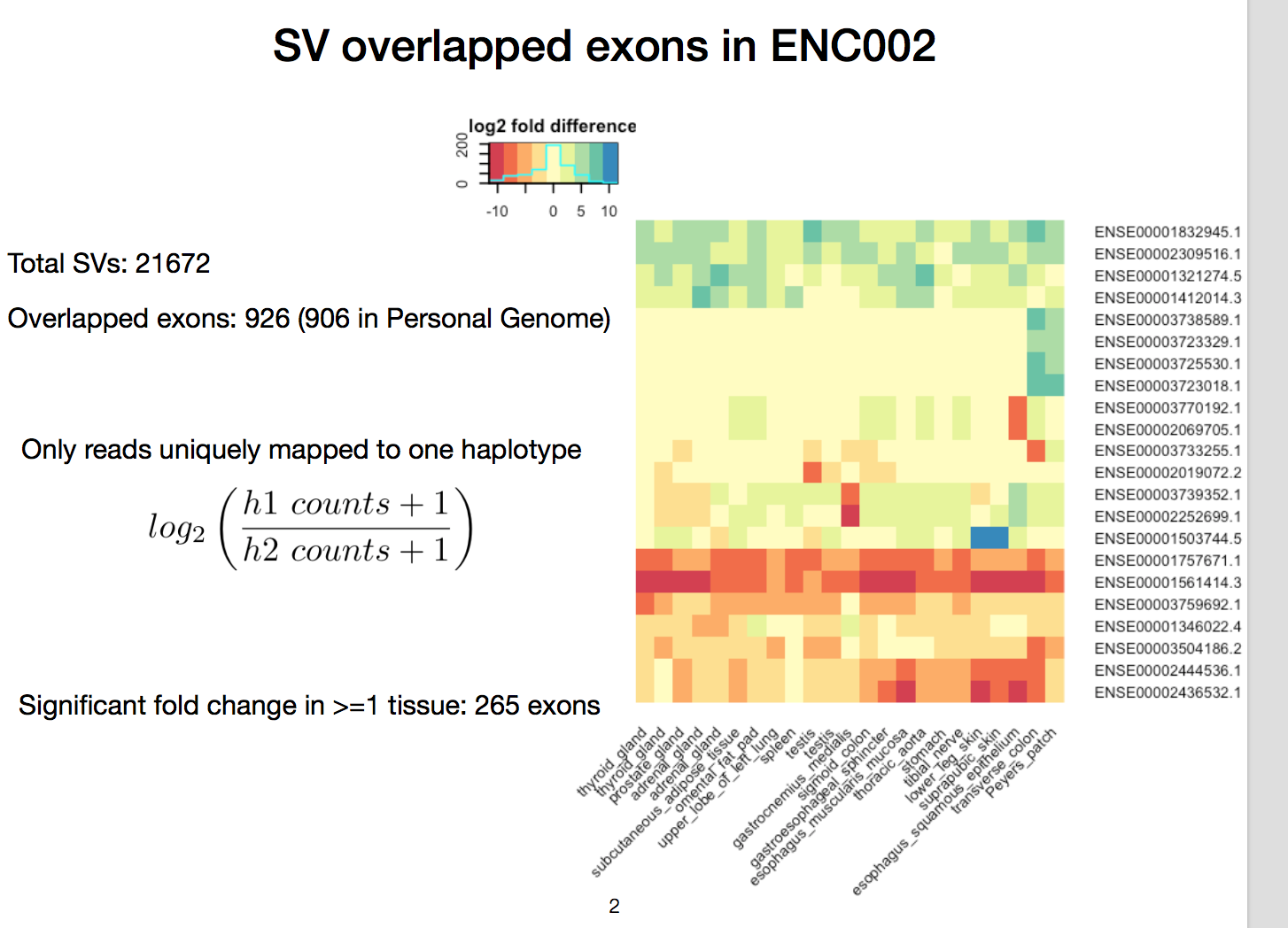
Figure 5. Example of antisense expression between protein\_coding genes and lncRNAs

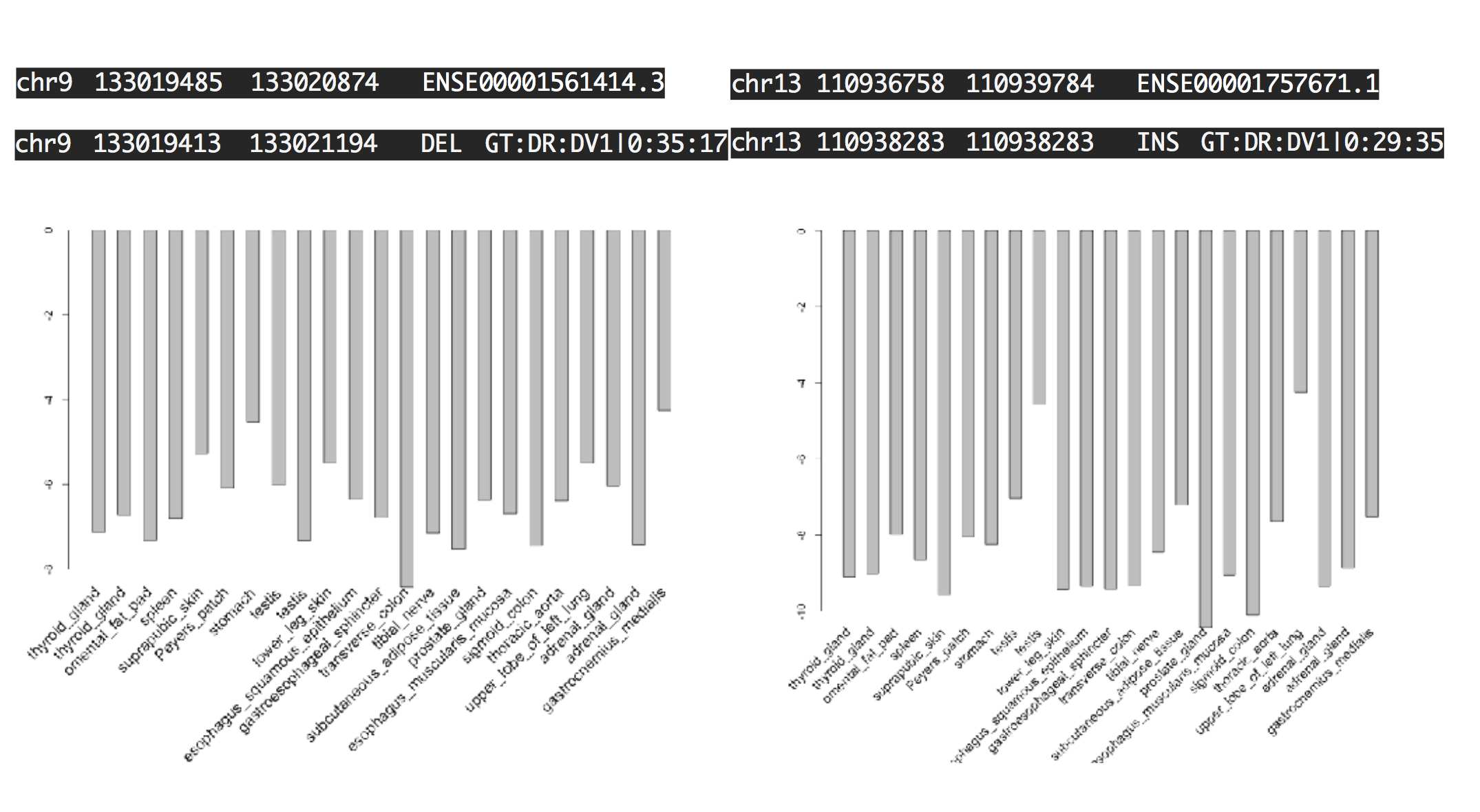
[1] "Number of pairs with positive correlation (PCC>= 0.6 ) 91"

[1] "Number of pairs with negative correlation (PCC<= - 0.6 ) 0"

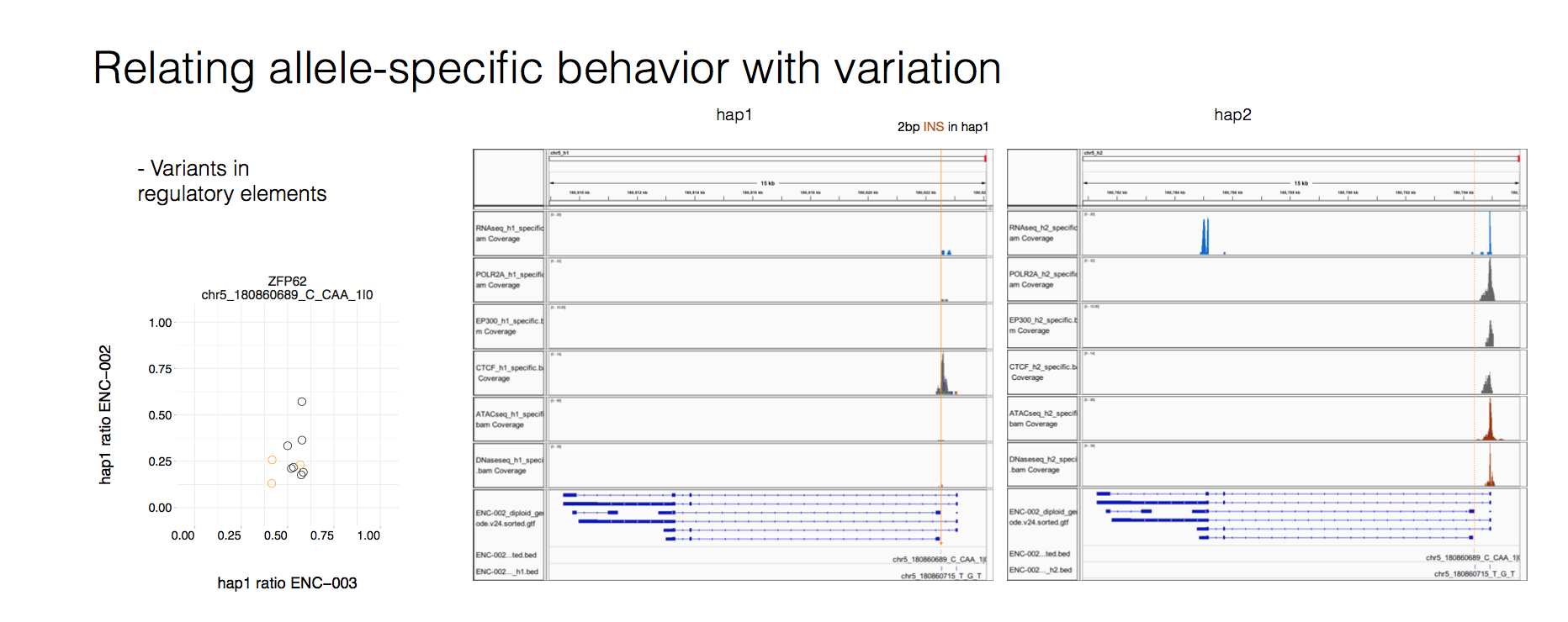
Negtaive correlation minimum value is -0.45

-Differential gene expression

* SVs affecting exon/gene expression
* 



- Allele-specific analysis  
  
  
RNAseq analysis on reference genome (example of indel causing allele specific expression)



Summary statistics plots

Figure5. Hierarchical clustering EnTex samples, long genes  
  
Figure 5. Example of antisense expression between protein\_coding genes and lncRNAs  
[1] "Number of pairs with positive correlation (PCC>= 0.6 ) 91"  
[1] "Number of pairs with negative correlation (PCC<= - 0.6 ) 0"  
Negtaive correlation minimum value is -0.45  
  
-Differential gene expression  
  
  
RNAseq analysis on diploid genome  
Some text here

|  |  |  |
| --- | --- | --- |
| Figure 3a. Observed Differential Expression mapping to reference or personal genome | Figure 3b. Deletions cause decrease in expression | Figure 3c. Impact of promoter deletions |

**Section 4. Integrative analysis**- SVs overlapping eQTLs  
- Analysis of ChipSeq, Methylation, other data types?

|  |
| --- |
| Figure 4. Example of SV overlapping eQTL. |

**Discussion**- Summarize findings  
- Need to extend to more genomes  
  
**Online Methods**- Details on data generation  
 - Sequencing, variant calling, phasing  
- Details on reference-based analysis  
- Details on CrossStitch pipeline: personal genome construction  
- Details on personal genome-based analysis