Transcriptome and genome sequencing uncovers functional variation in humans

---

Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories

Journal Club - Stephan Pabinger - 2013-09-24

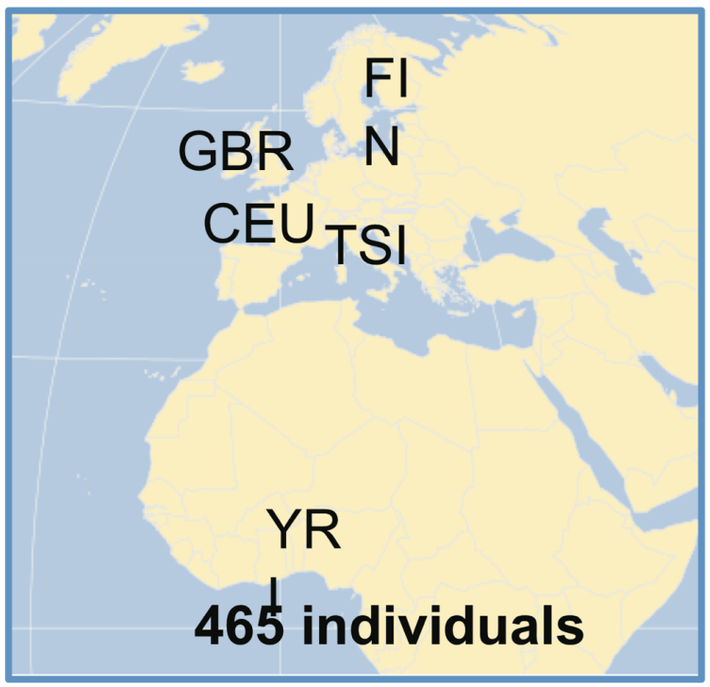
Aim of study

Characterize functional variation in human genomes
- catalogue novel **loci** with regulatory variation
- discover & characterize **molecular properties** of **causal** functional **variants**

What do GWAS variants actually do in the cell?

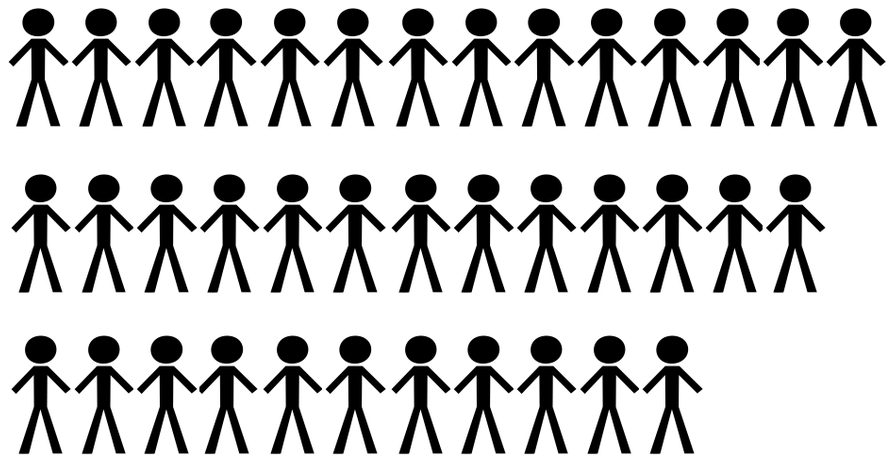
Setup

Lymphoblastoid cell lines
Samples from **1000 genomes** project
mRNA and sRNA sequencing



Populations
Europe, Finns, British, Toscani, Nigeria

Data (after control)

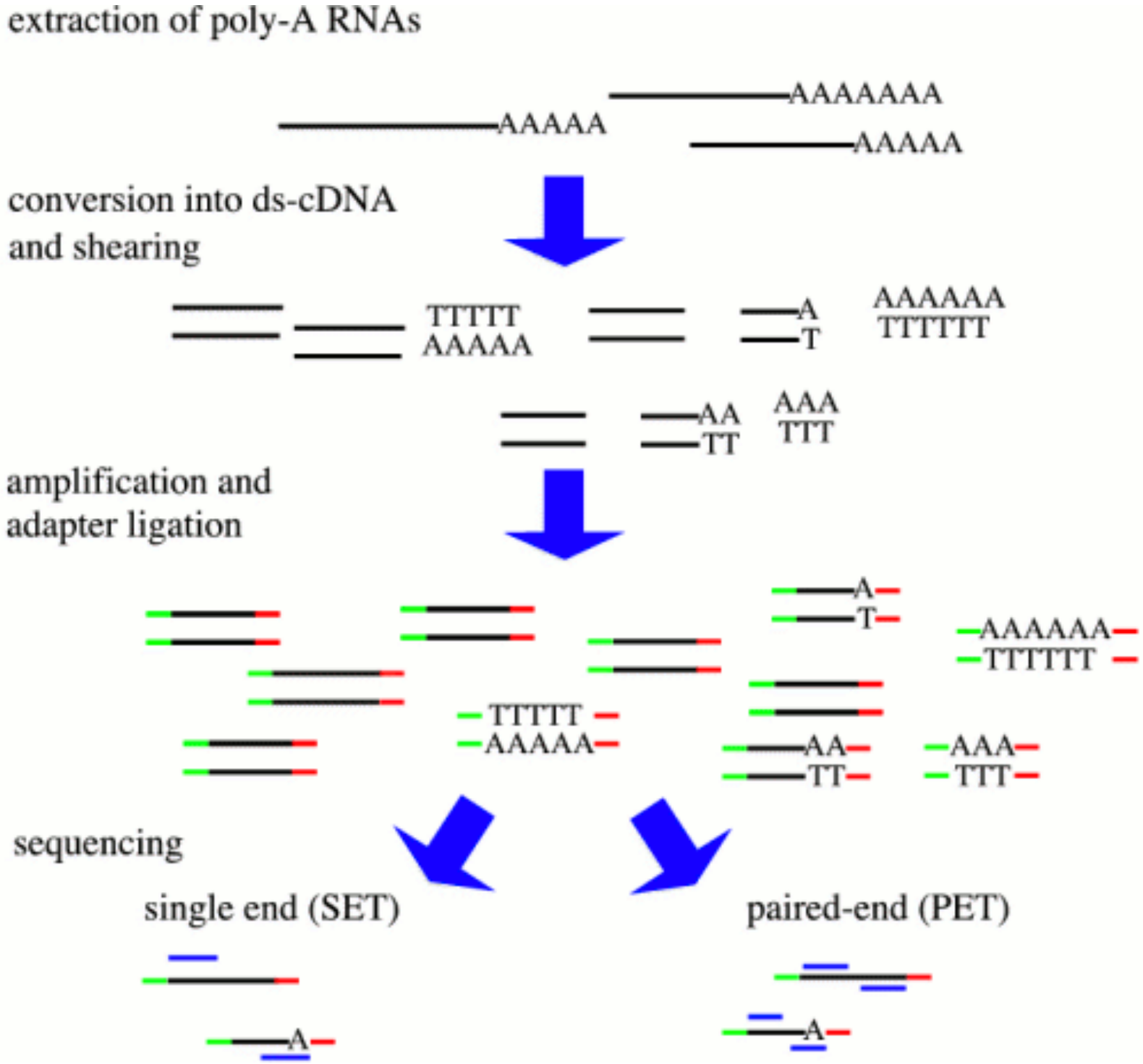


462 - mRNA

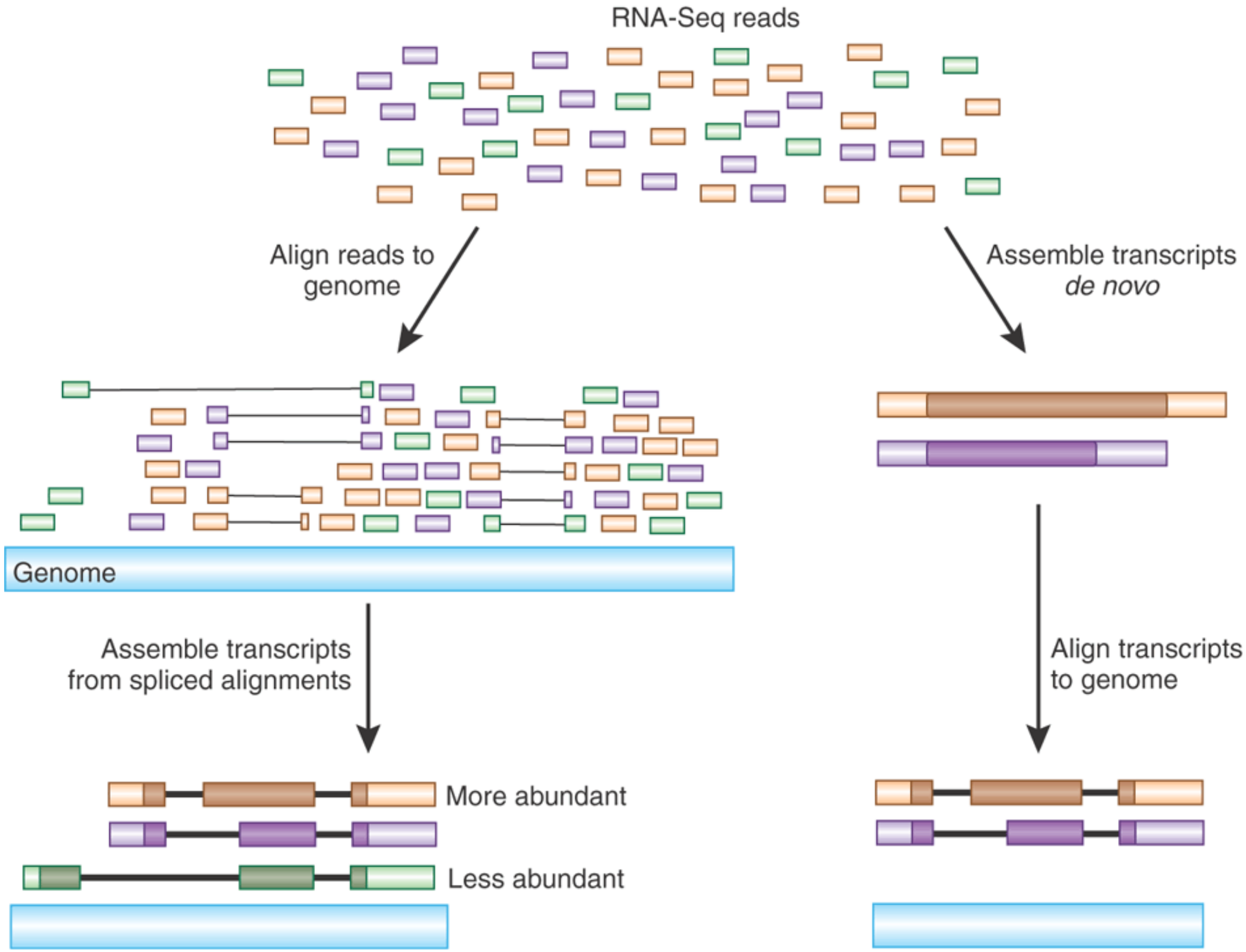
452 - sRNA

421 - in 1000 Genome Phase 1 (rest SNP array)

RNA Seq


extraction of poly-A RNAs
AAAAAAA
AAAAA
AAAAA
conversion into ds-cDNA
and shearing
TTTTT
AAAAA
A
T
AAAAAA
TTTTTT
AA
TT
AAA
TTT
amplification and
adapter ligation
A
T
AAAAAA
TTTTTT
TTTTT
AAAAA
AA
TT
AAA
TTT
sequencing
single end (SET)
paired-end (PET)
A
A

RNA Seq



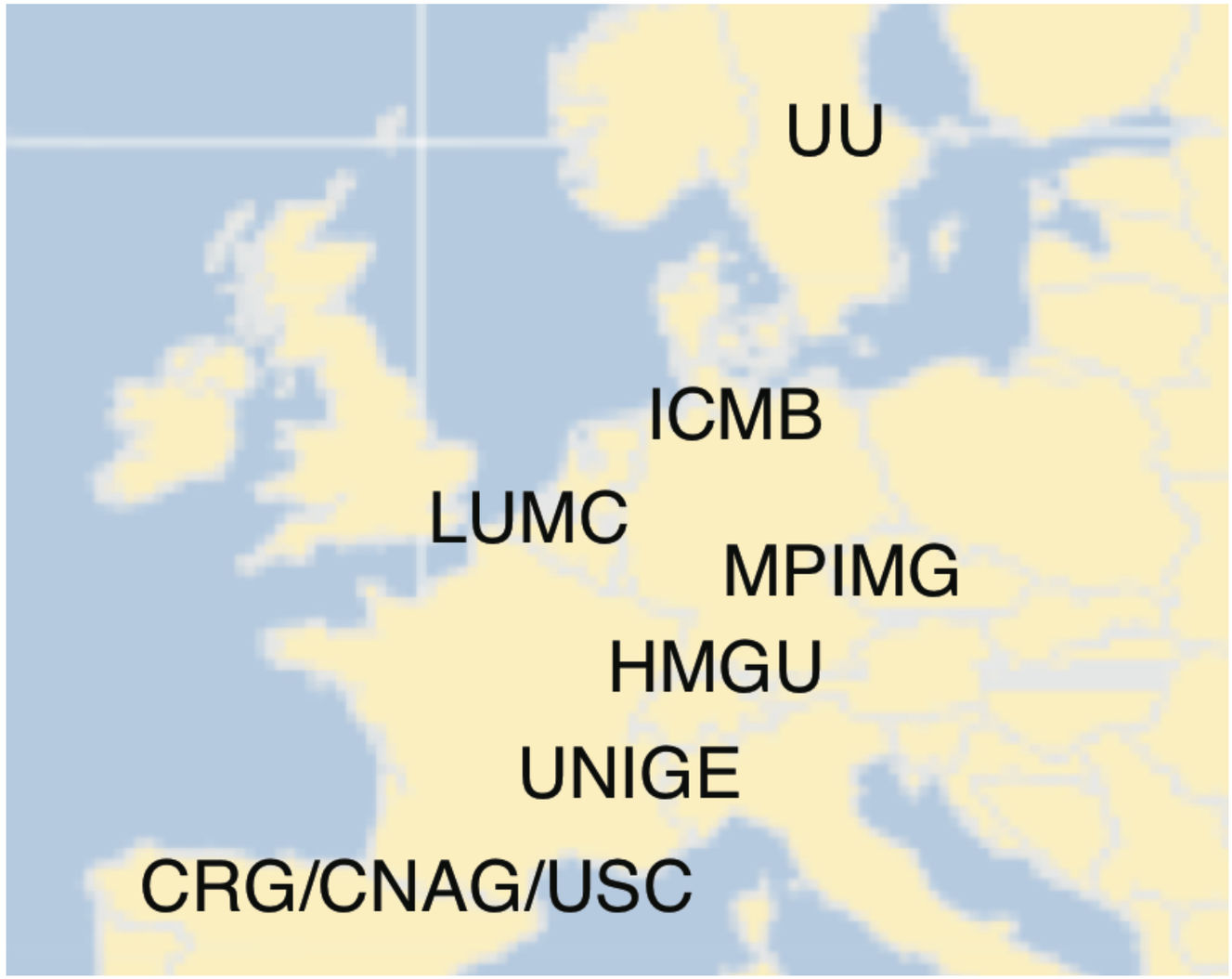
nature.com

Sequencing labs

Transcriptome sequencing

SEUVADIS CONSORTIUM

7 laboratories
48-116 samples each



Quantitative trait locus (QTL) analysis

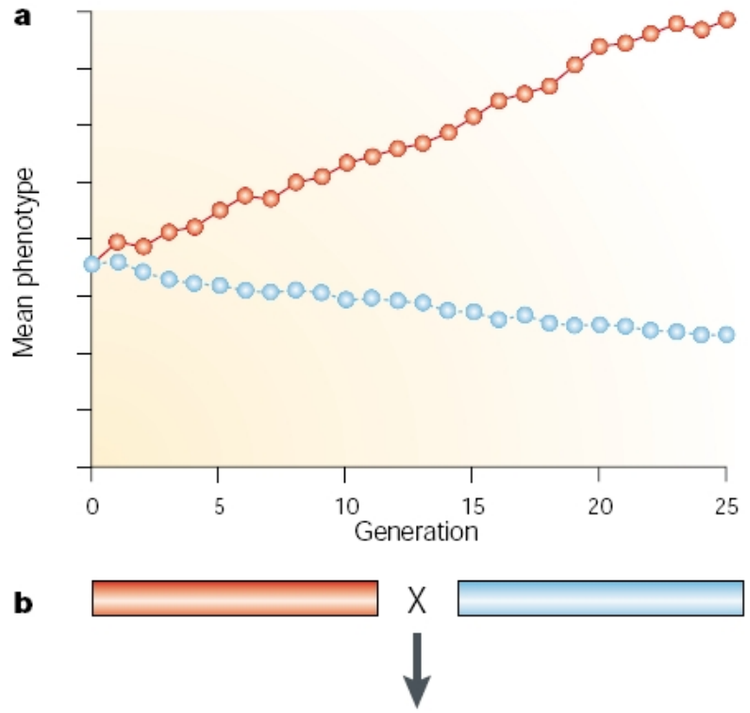
Links **phenotypic** and **genotypic** data

Stretches of DNA containing or linked to the **genes** that underlie a **quantitative trait** (phenotypes; can be attributed to polygenic effects)

Expression QTLs (**eQTL**):

cis- and *trans*-controlling elements for the **expression** of **genes**

Quantitative trait locus (QTL) analysis



Traits

Parental lines

Analysis

Mapped *cis*- QTLs to transcriptome traits of protein-coding and miRNA genes

Separately for European and Yoruba (Nigeria)

***cis*-regulatory element**

region of DNA or RNA that regulates the expression of genes located on that same molecule of DNA

***trans*-acting element**

is usually a DNA sequence containing a gene; resulting protein (or microRNA) is used in the regulation of another target gene

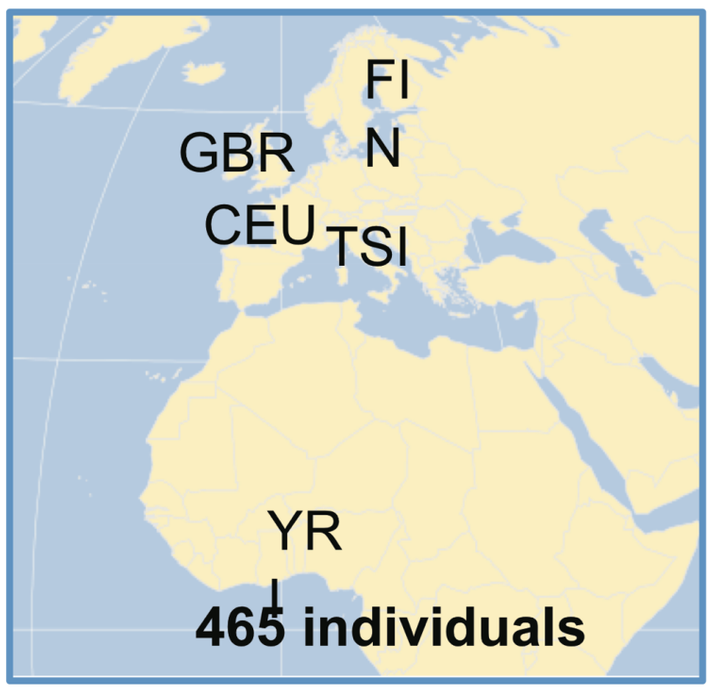
Transcriptome variation in populations

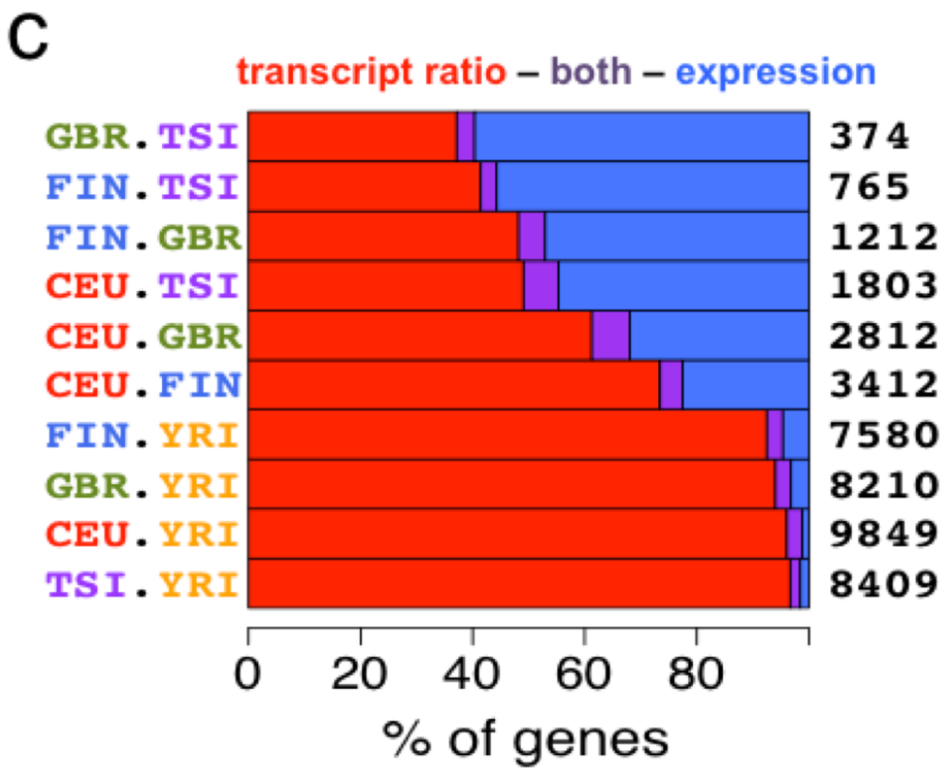
Shown in
- overall expression levels
- relative abundance of transcripts from the same gene

263 - 4,379 genes with **differential expression** or **transcript ratios** between population pairs

---

Continental tr differences higher than within Europe



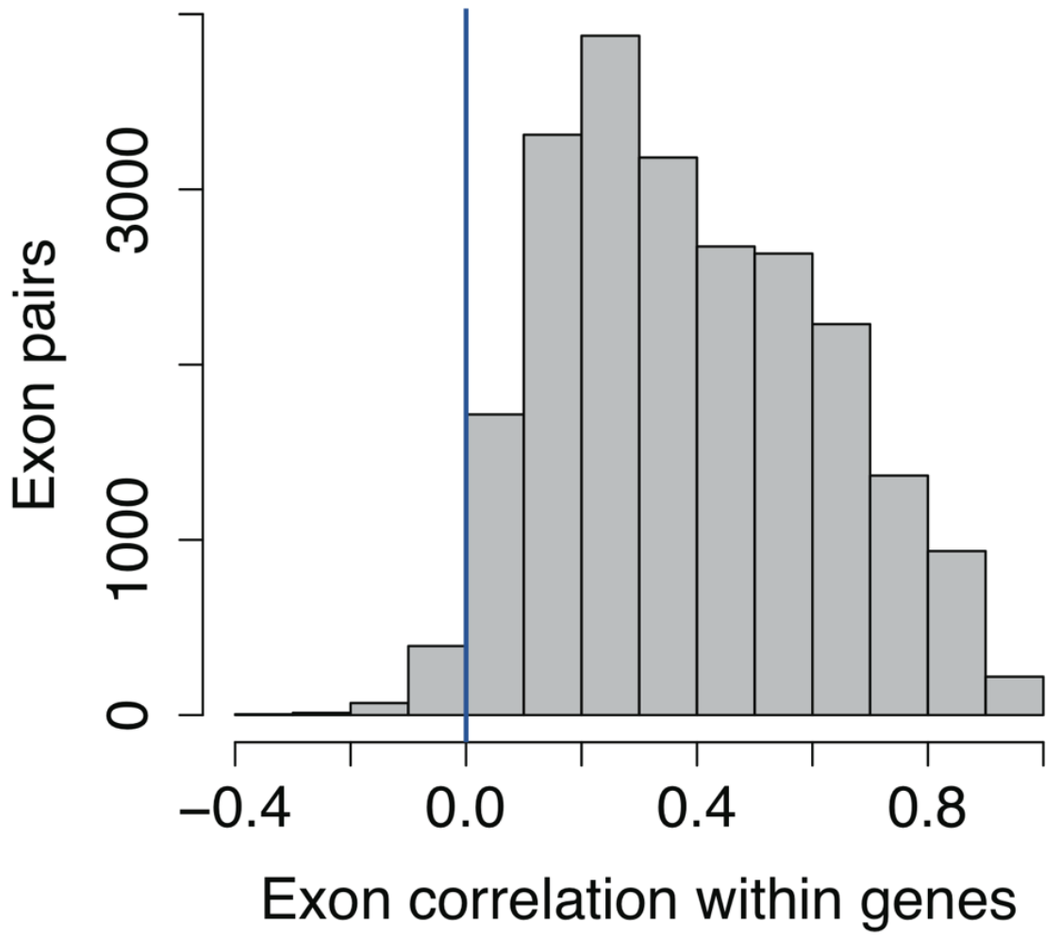


The proportion of differentially expressed and differentially spliced genes between population pairs using the DEXSeq method.[10]

eQTL analysis

eQTL analysis of **protein-coding** and **lincRNA** genes
-> **3,773** genes having “classical” eQTL

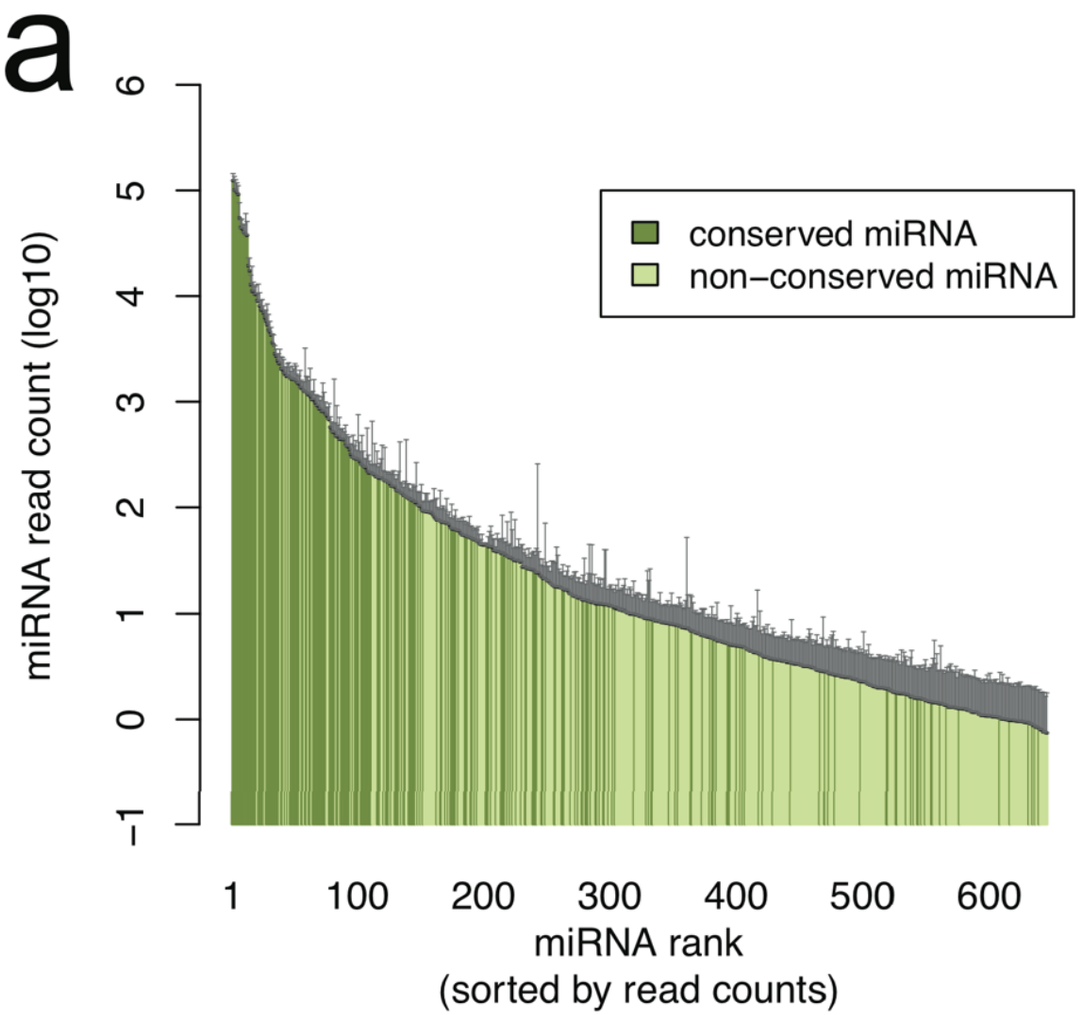
mapped eQTLs for exon quantifications that can capture **gene expression** & **splicing variations** -> **7,825** genes



Correlation between quantifications of exons from the same gene for chr20 in EU dataset.

Often correlation is not very high
-> frequent splicing variation

miRNA expression study



Expression of 644 autosomal miRNAs detected in 452 individuals

60 (out of 644) have significant *cis*-eQTLs

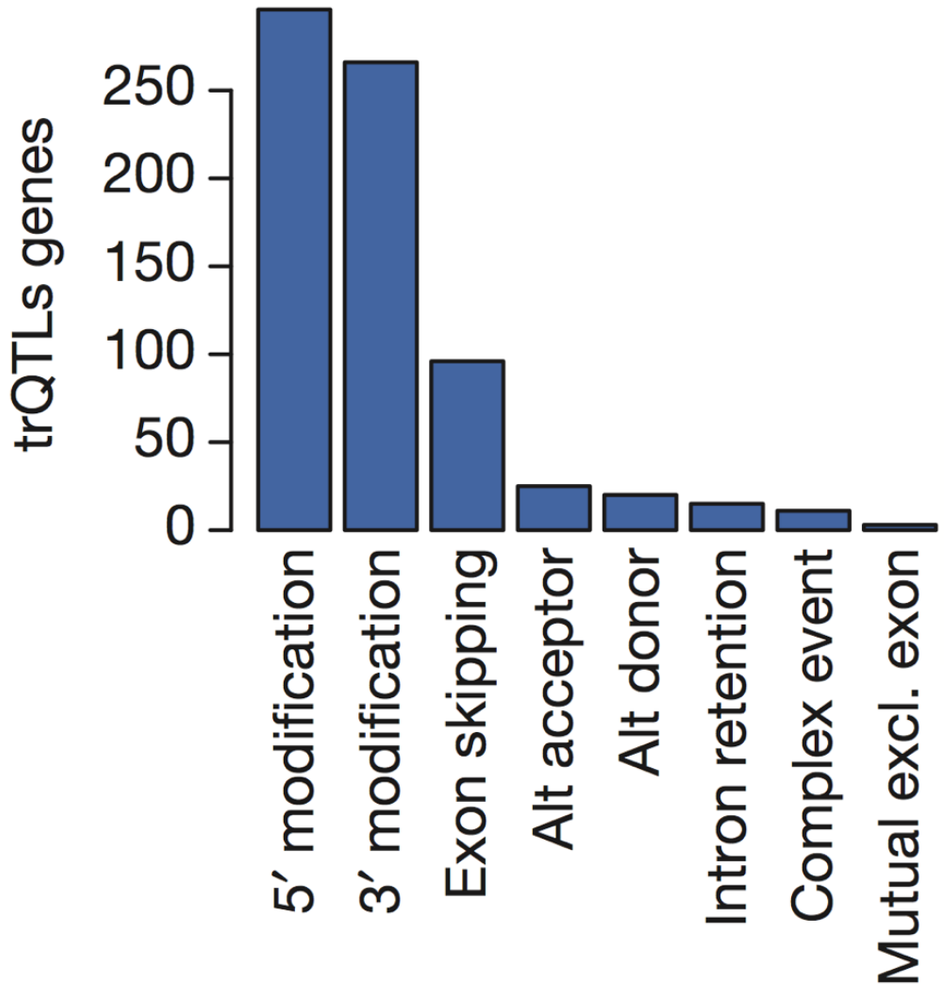
miRNA - mRNA effect

32 (out of 100): expression correlated with target exons -> miRNA downregulate genes

transcript ratio QTLs (trQLTs)

Investigate genetic effects on **splicing**

Discovered **639** genes with transcript ratio QTLs (**trQTLs**) affecting the ratio of each **transcript** to the **gene total**



Classification of changes caused by trQTLS

transcript ratio QTLs (trQLTs)

Investigate genetic effects on **splicing**

Discovered **639** genes with transcript ratio QTLs (**trQTLs**) affecting the ratio of **transcript** to the **gene total**

Identified **279** genes where **eQTL and trQTL** causal variants are **independent** (in most genes) -> transcriptional activity and transcript usage are controlled by different regulatory elements

Properties of regulatory variants

Compared properties of **top** (most significant) **eQTL** variant per gene to a **null of non-eQTL** variants

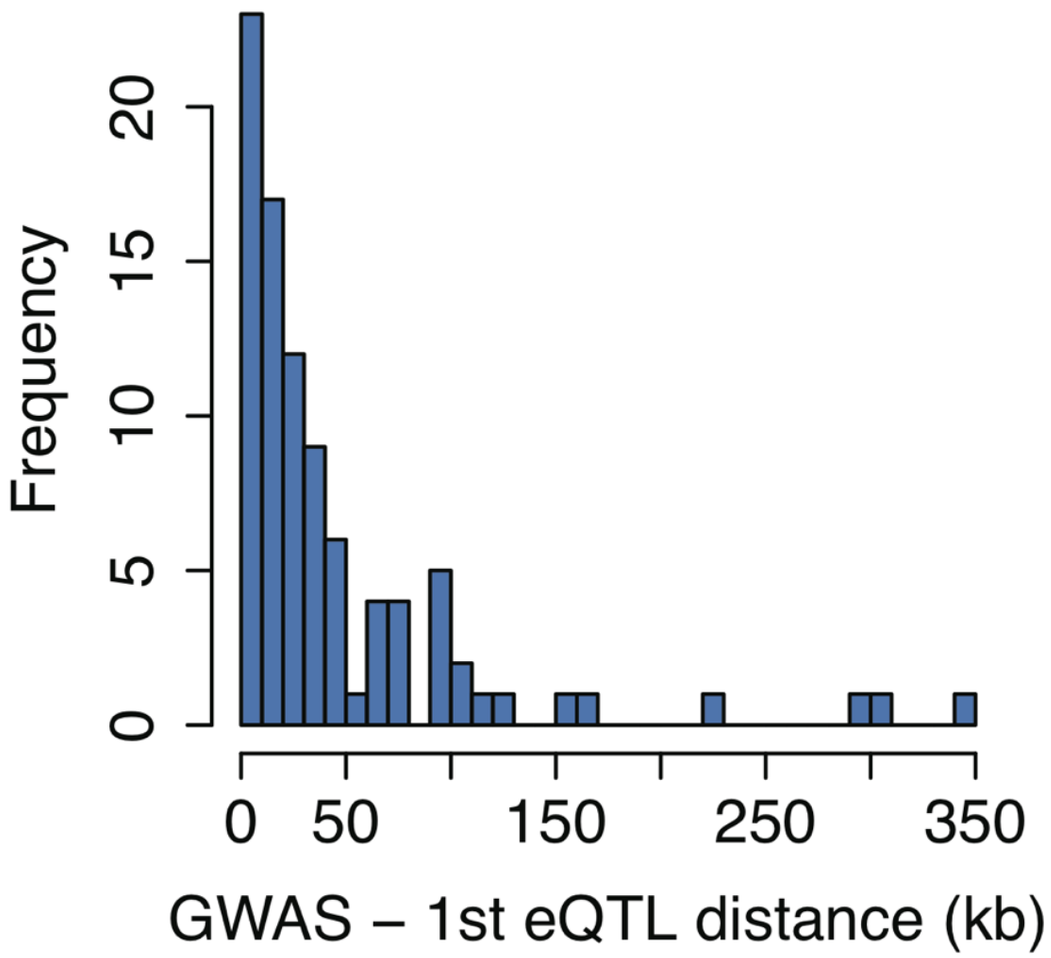
InDels enriched -> more likely to have effect (vs SNPs)

eQTLs
- non-coding elements | transcription factor peaks
- Dnase1 hypersensitive sites
- chromatin states of active promoters & enhancers
- splice sites
- non-synonymous enrichments

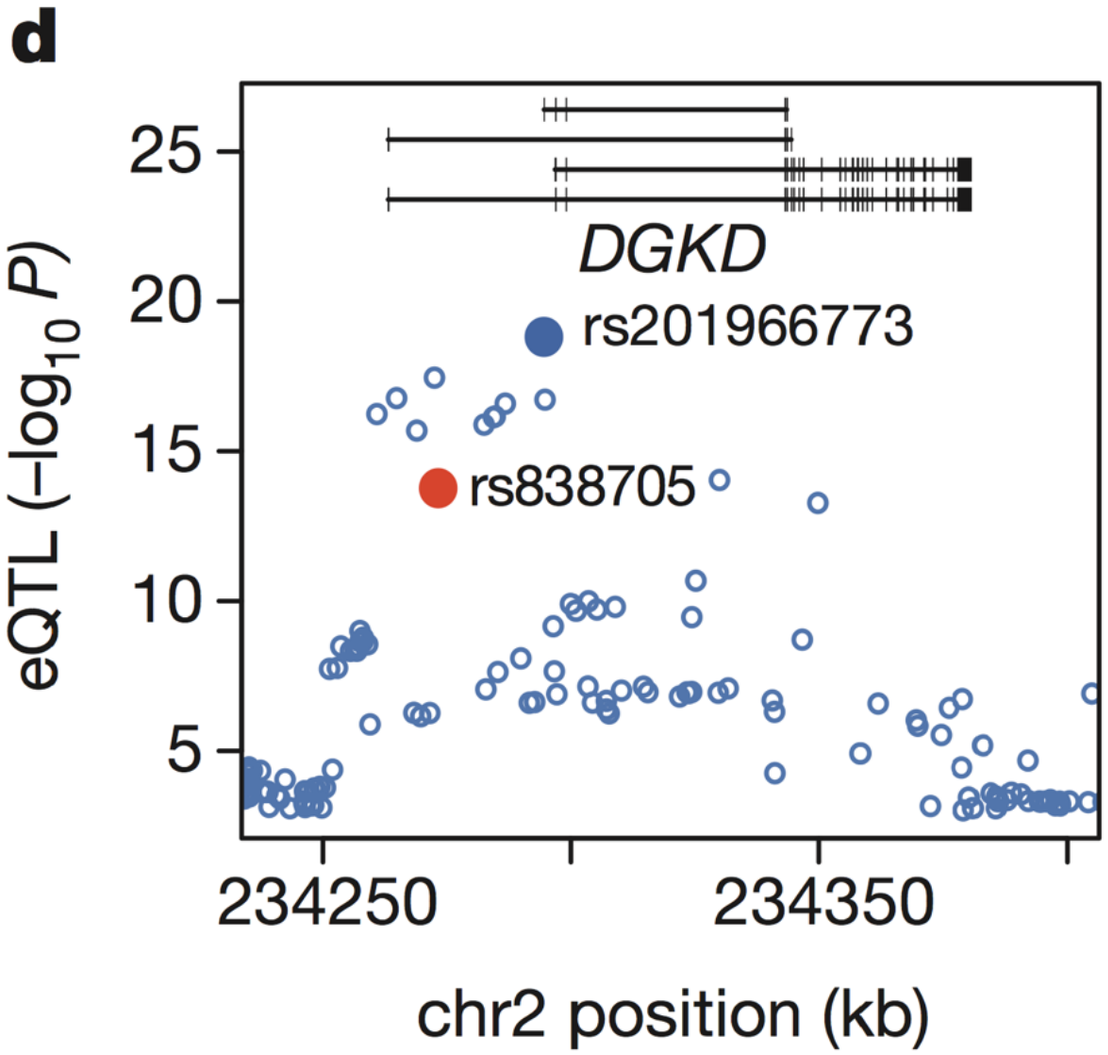
trQTLs
- splices sites | 3’ UTRs | promoters

GWAS SNPs - eQTL

78 eQTL regions that are likely causal signals for 91 GWAS SNPs



How close the GWAS variant is to best EUR eQTL



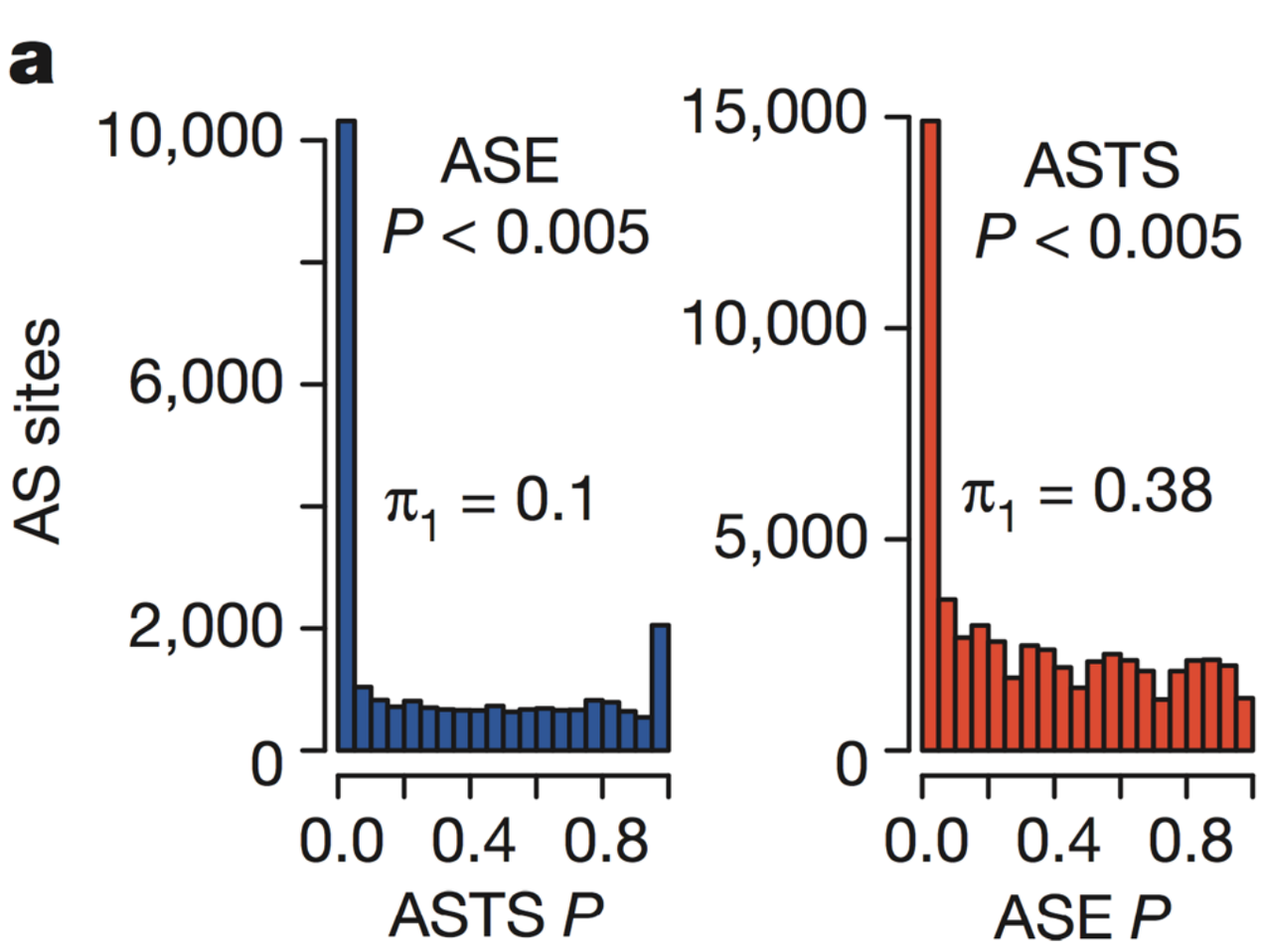
Intronic SNP is associated with Ca levels (red), and the top eQTL variant (blue) is a very likely causal variant

Allelic effects

ASE equally common as **ASTS**

ASE seems to be genetically driven rather than epigenetic effects

ASE may be **driven by ASTS**



Distribution of ASTS p-value of sites with significant ASE and vice verse

Summary

Genetic regulation is very common
- over **one half of genes** affected by **eQTLs**
- major determinant of allelic expression

16% of known **GWAS** variants are **eQTLs**
- but by chance overlap is 11%

However, with **statistical methods eQTL** can be used to **identify causal GWAS** variant

eQTLs and **ASE** appear common but independent

UU

ICMB

LUMC

MPIMG

HMGU

UNIGE

CRG/CNAG/USC

Reproducible?

Setup

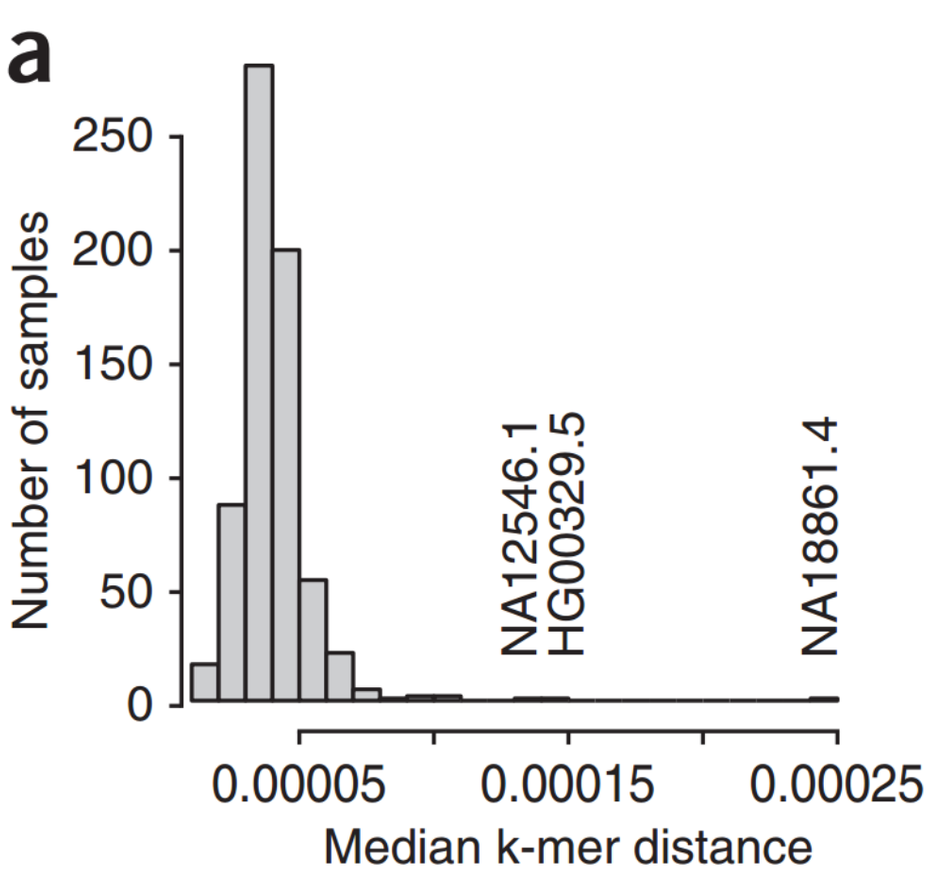
Illumina **HiSeq 2000**
Illumina's TruSeq kits for preparation
mRNA - 75bp PE
sRNA - 36bp SE

5 RNA samples sequenced **at all sites**
→ for comparison

Software
GEM (mRNA)
miraligner (sRNA)

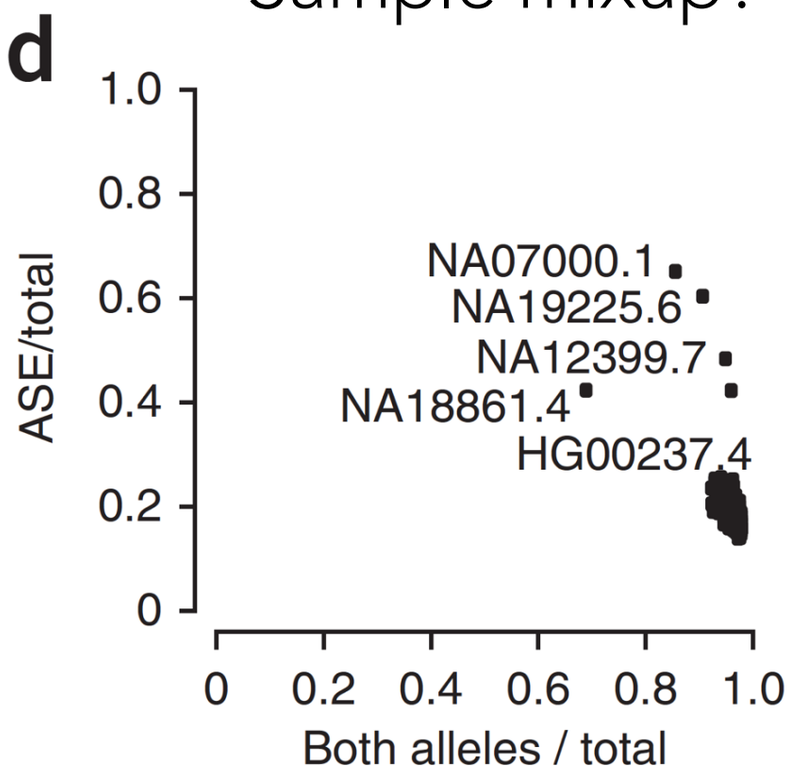
Quality check

k-mer distance using abundance in samples
promising & does not require mapping
Exon/transcript expression correlation between samples
→ low correlation = outlier
Allele-specific expression



with k=9; pair wise distance

Sample mixup?

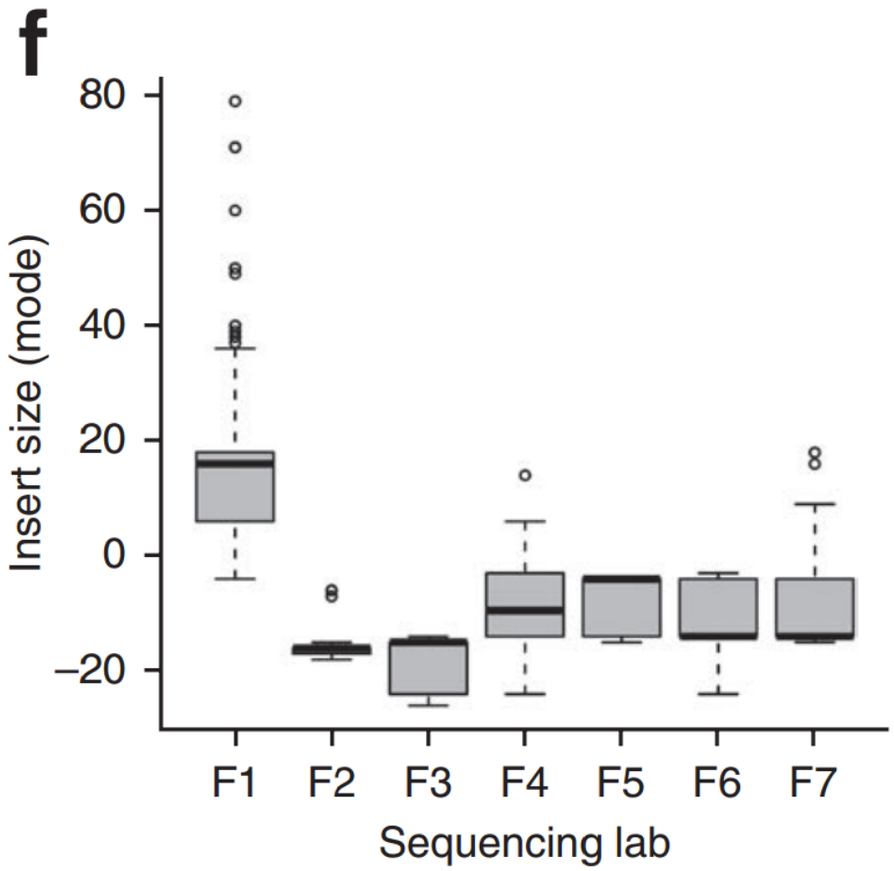


For all heterozygous sites: proportion of het SNPs where both alleles were observed vs proportion of ASE

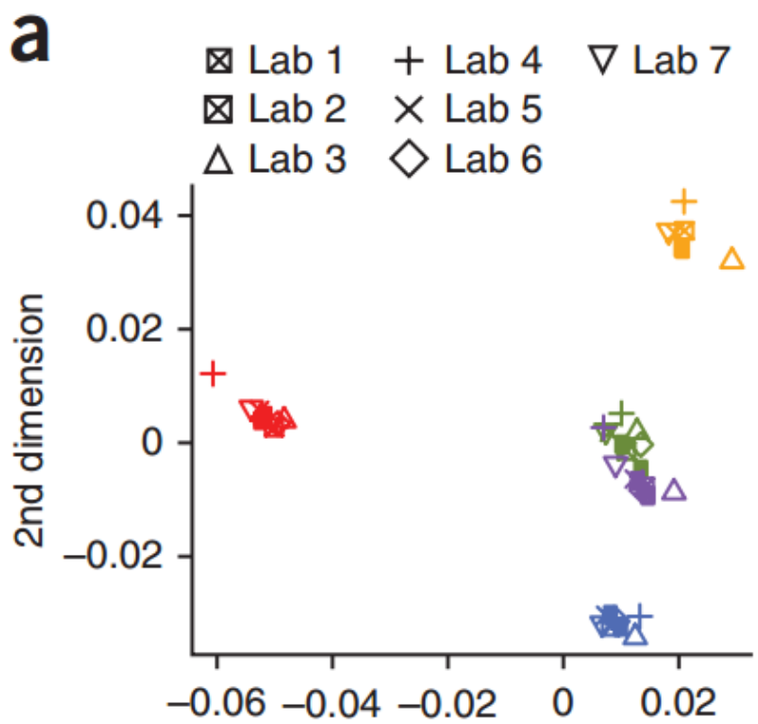
Sources of variation

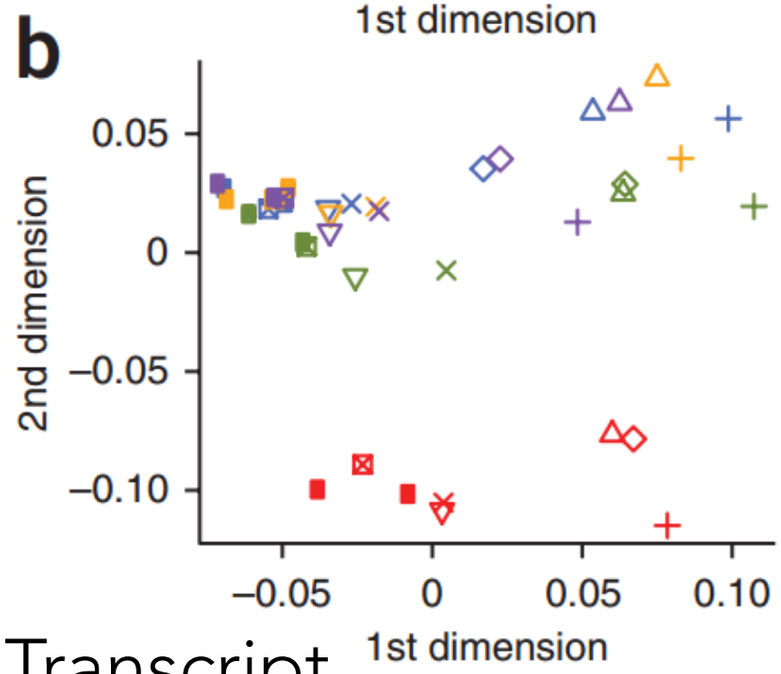
Samples **clustered by sample** and not lab

Clustering for **exon** quantification
stronger than for **transcript**
Lib prep difference in **GC percentage**
Smaller inserts than targeted



Exon





Transcript

Correction for variation in mRNA

Used **PEER**

takes quantification of genes (or transcripts) and uses factor analysis-based methods to infer factors explaining variance.

After PEER

clustering by lab **less pronounced**

technical variation can be properly accounted for

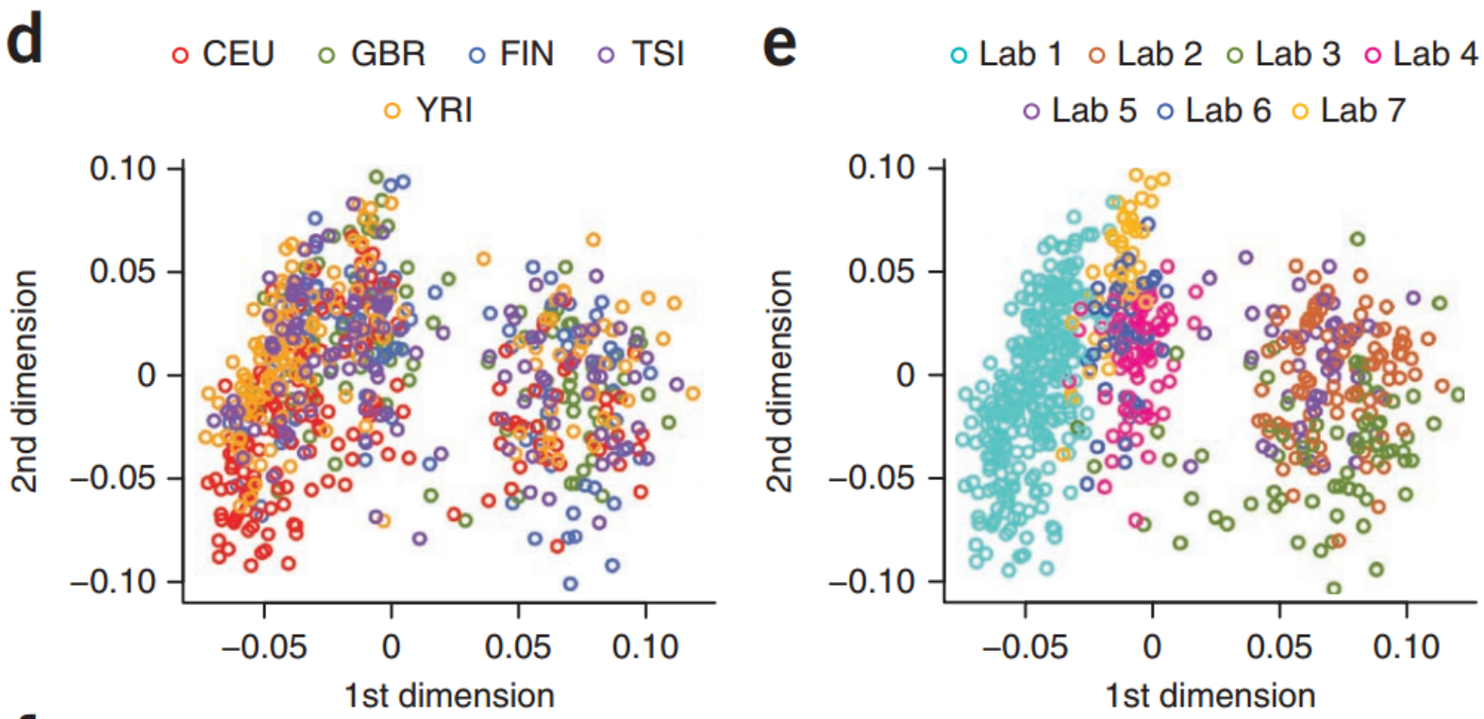
Stegele et al. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability ofgene expression analyses. Nat Protoc 2012

Plots of transcript quantification

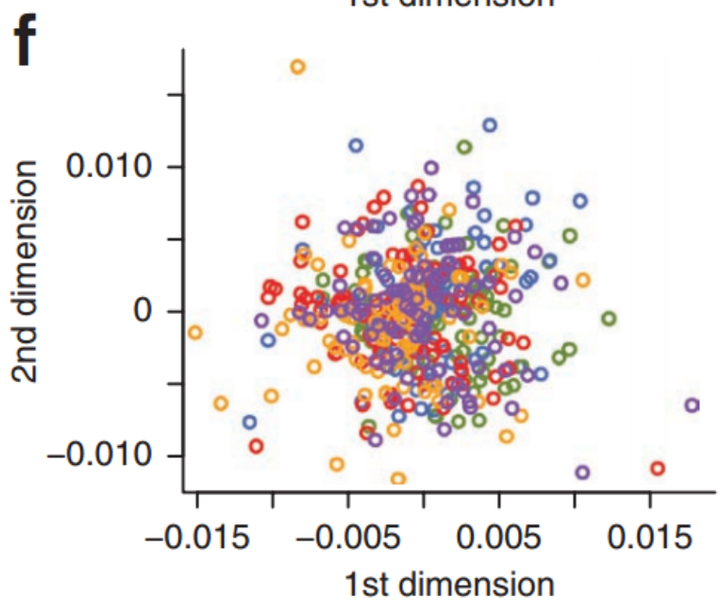
By population

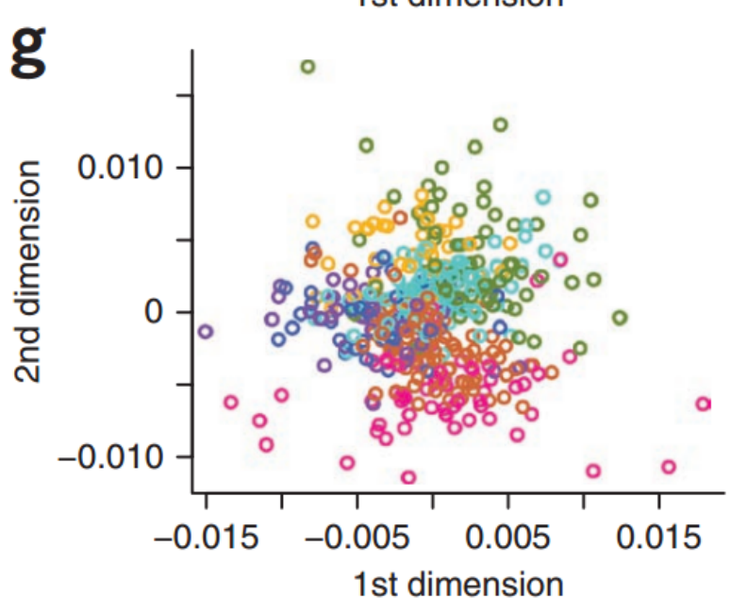
By laboratory

Before PEER



After PEER

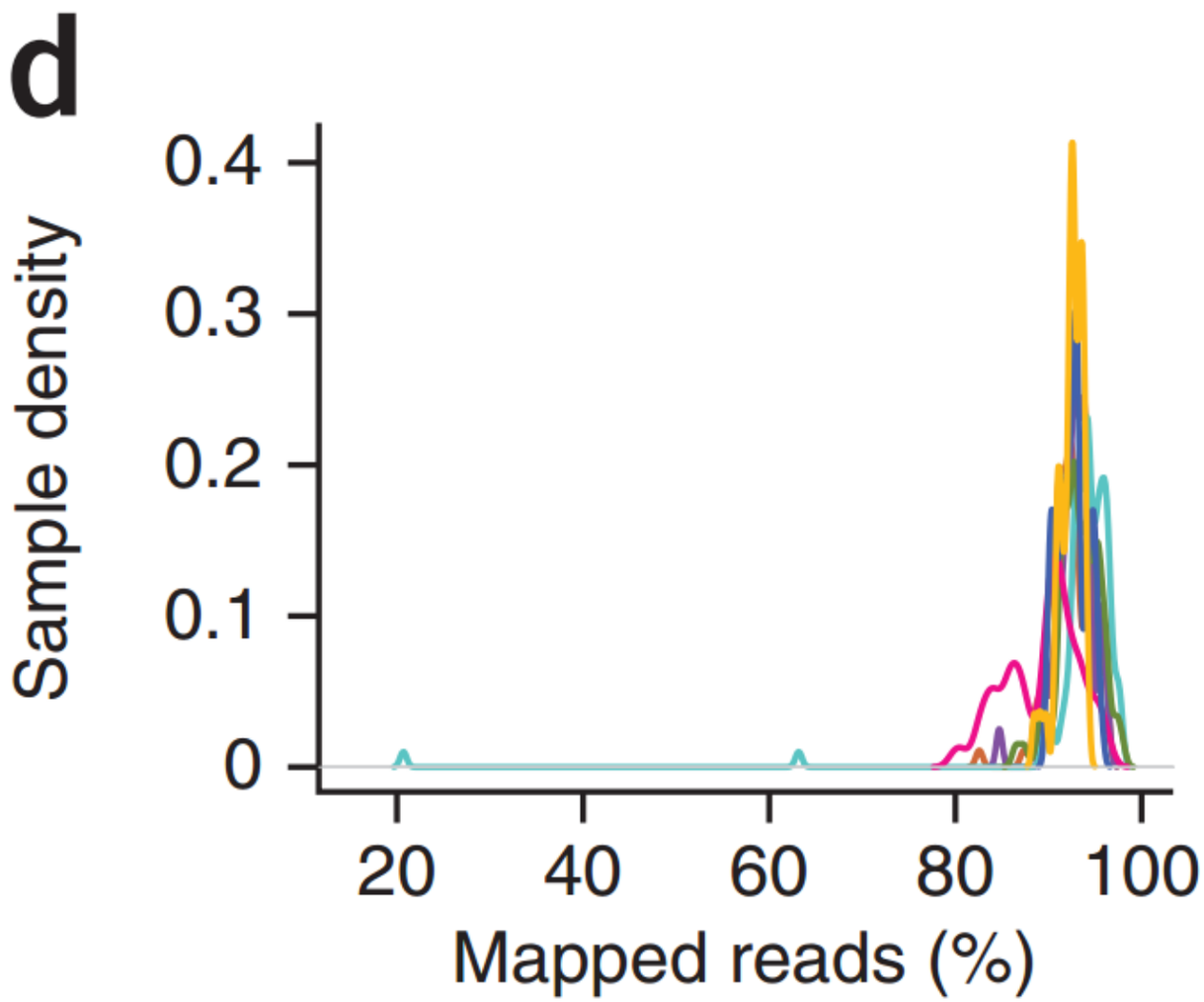




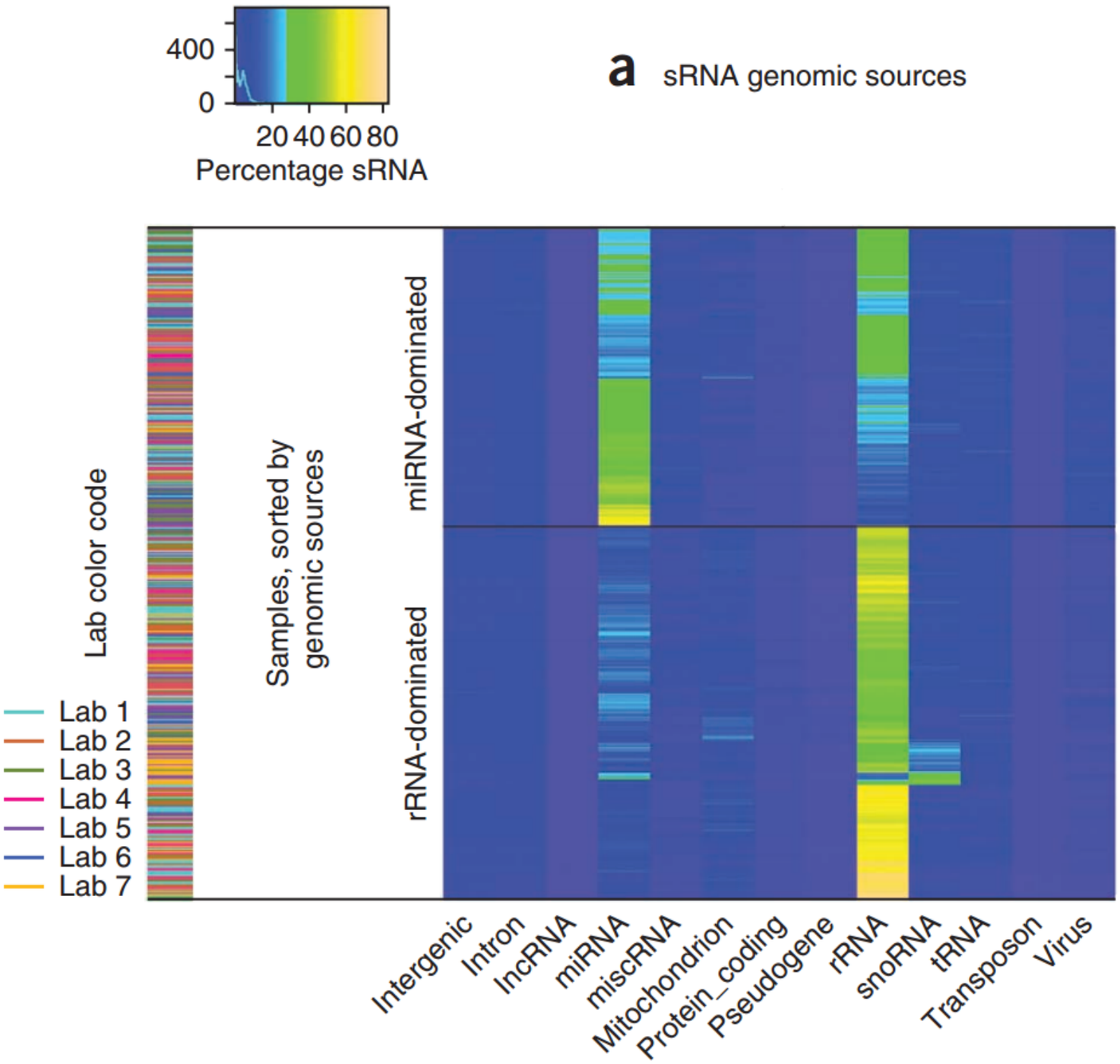
sRNA sequencing

Removed reads <18bp

Sequencing **quality & mapping quality high**



Heatmap clustered by expression of sRNA sources



sRNA sequencing

Removed reads **<18bp**

Sequencing **quality & mapping quality high**

Origins from miRNA and rRNA

differences in sRNA likely due to RNA extraction (not labs)

Corrected by **PEER**

GC% biggest source of variation

Summary

Technical **variation** in RNA-seq is **small** (when)

- using exact **same protocols** for sample prep and sequencing kits

Slight variations in **GC%** and **insert size**

PEER algorithm

account for & reduce technical factors (large studies needed)

Quality checks

Proposed quality checks

mRNA & sRNA sequencing

- Distribution of **base quality** scores
- Average and width of the distribution of **GC content**
- Percentage of **reads mapping** to the genome
- Checks for **sample swaps** and **contaminations**
- Outlier detection: **pairwise correlations** in **expression** quantification between samples

Proposed quality checks

mRNA

- Mean & s.d. of **insert size**
- Percentage of **reads mapping** to annotated **exons** (at least 60 %)
- 5′–3′ trends in **coverage across transcripts**

sRNA

- **Length distribution** after adaptor clipping
- Percentage of reads **mapping** to **known sRNA** genes

Transcriptome and genome sequencing uncovers functional variation in humans

---

Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories

Journal Club - Stephan Pabinger - 2013-09-24