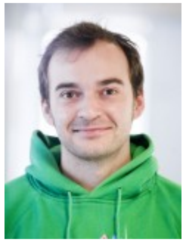
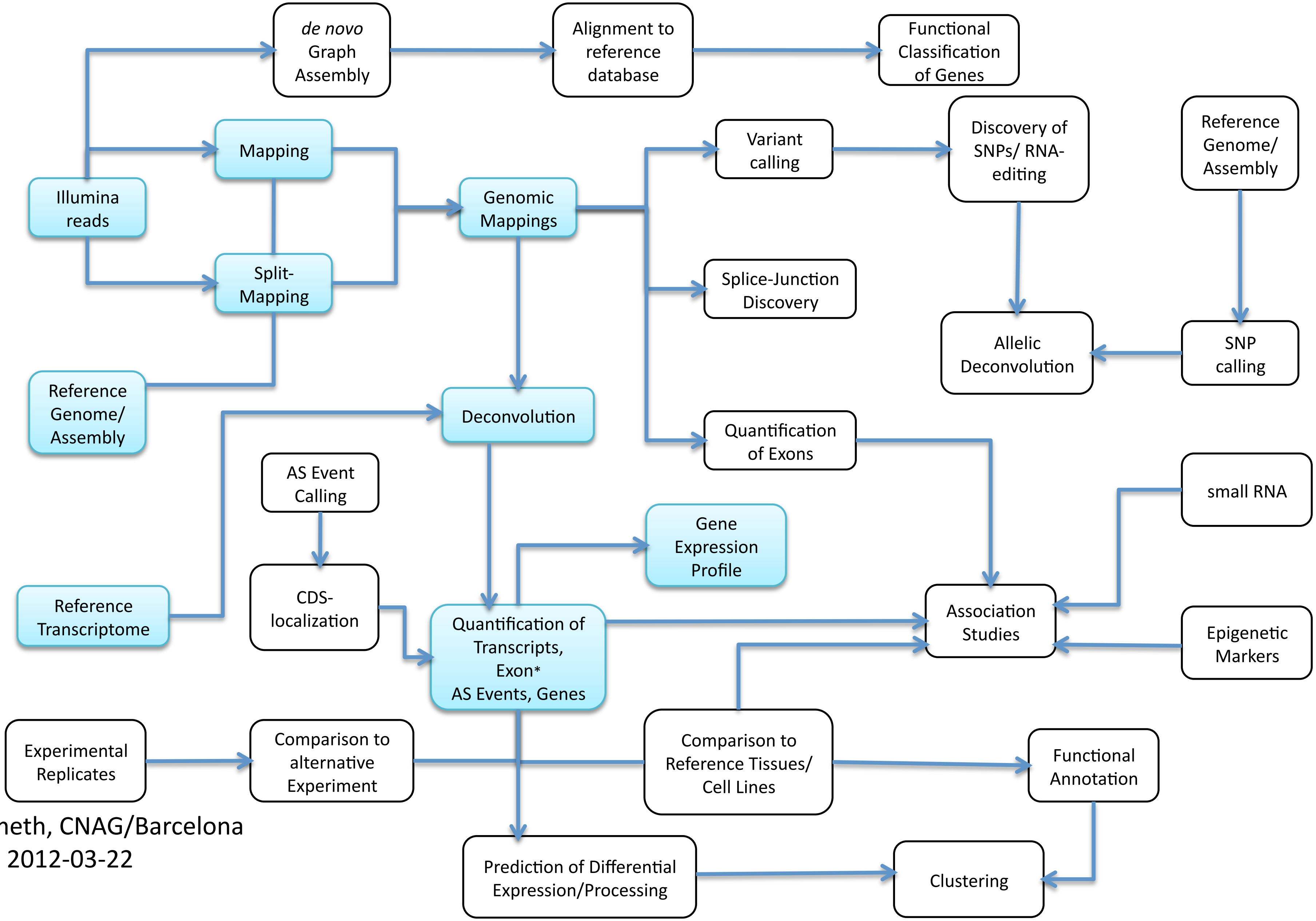
Pipeline: quantitative RNA-Seq

MVP:



Thasso Griebel

Jordi Camps



Micha Sammeth, CNAG/Barcelona
Geuvadis TC 2012-03-22

Mapping

- GEM (GEnome Multitool) split-/mapper (http://gemlibrary.sourceforge.net)

Santiago Marco

Leonor Frias

Paolo Ribeca

- exhaustive mapping up to the number of mismatches

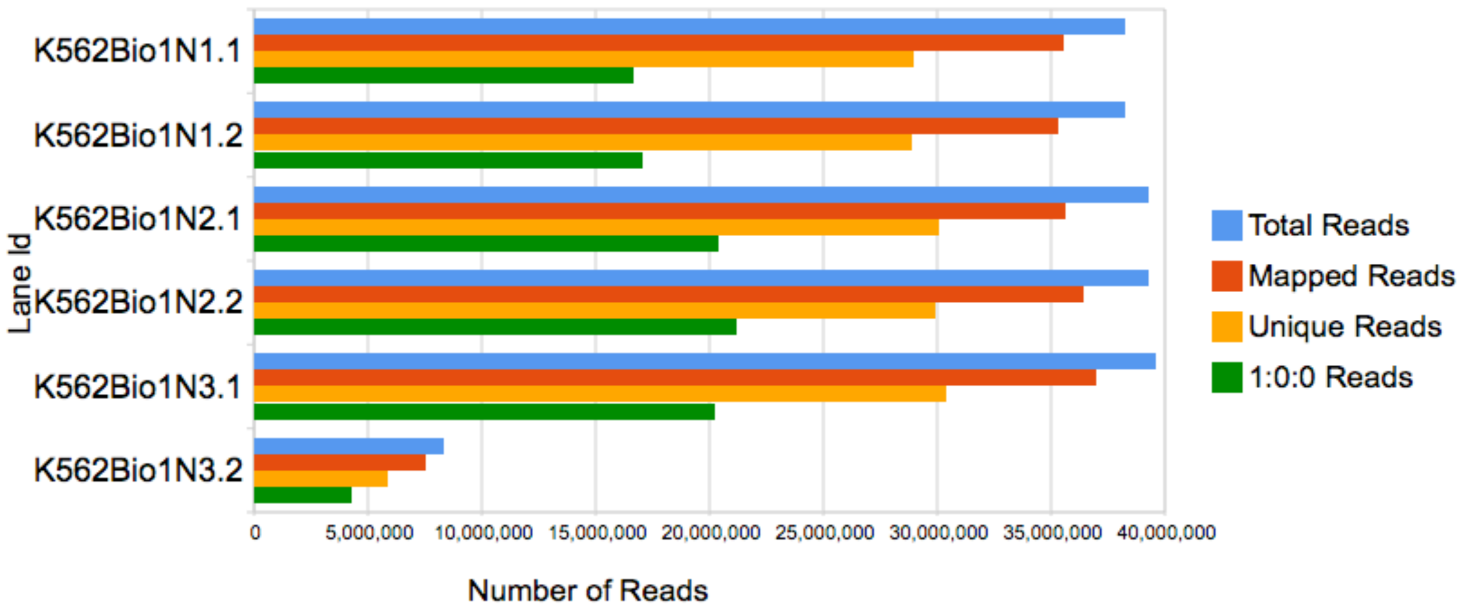
- quality mapping: downweight mismatches at positions with bad qualities (quality score)

Example:

```
BILLIEHOLIDAY_0008:1:11:13870:15798#0/1
AAAACTAATAACTTAAAACTGCCACACNCAAAAAAGAAAACCAAAGTGGTCCACAAAACATTCTCCTTTCCTTCTG
hhhhhgghhhhhhdhhhhhhhhhhhffBfffffffhhhhhhhghhhdhhhhhhfhhhhhhhfhehhhhhhhhhhggh
0:4:4:3
chr6:F74227323G28@2/1,chr7:F22550021G28@2/1,chr9:R135896396G28@2/1,chr13:R96271784A28@2/1,
    chr5:F14651837T27G28@40/2,chr12:F8233706G28<
    +2>43@2/1,chr12:F19608973G26G28@40/2,chr10:R129322971T27G28@40/2
```

Mapping

Mapping Outcome:



Different Mapping Classes:

Unique: Reads that map uniquely (Strata 1:0:0, 0:1:0)

Multi: Reads that map multiple times in the reference

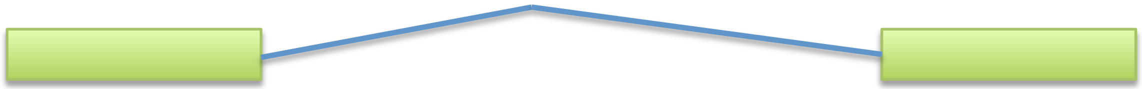
Ambiguous: Reads that map unique, but only in the most permissive Stratum (0:0:1)

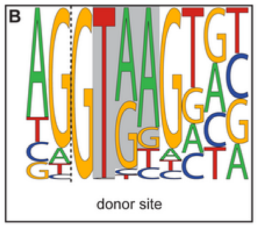
Redundant: Reads that have redundant hits in the reference, usually above the limit the output every hit's position (e.g., 14:23:58)

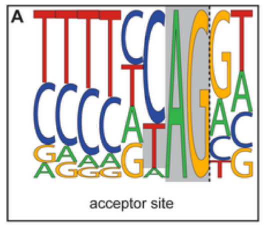
Unmapped: Reads that won't map to the given reference, with the given set of parameters

Split-Mapping

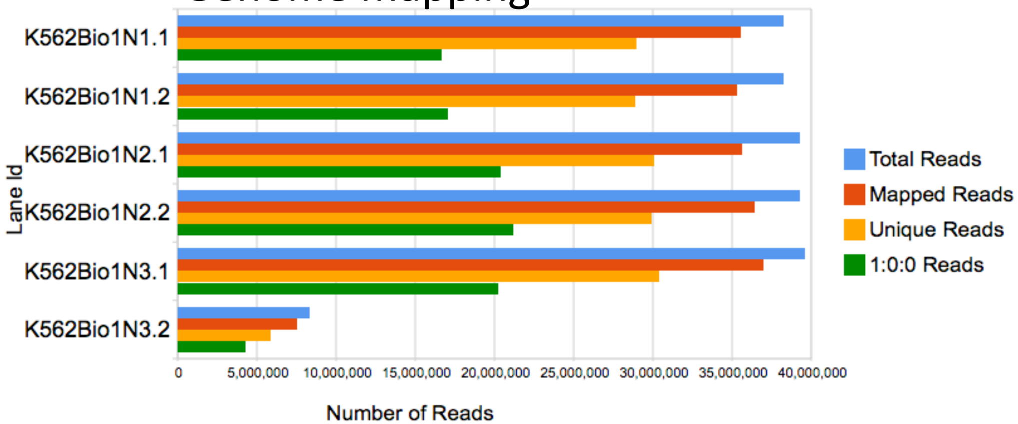
- match substrings of the read to the genomic sequence (expensive!)
- in RNA-Seq split-maps correspond (mainly) to the splice-junctions
- Splice Site consensus can be used to “guide” split-mapping



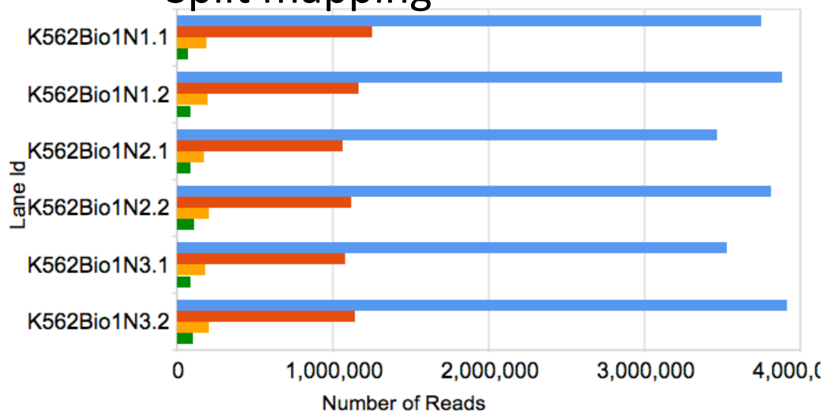




Genome mapping

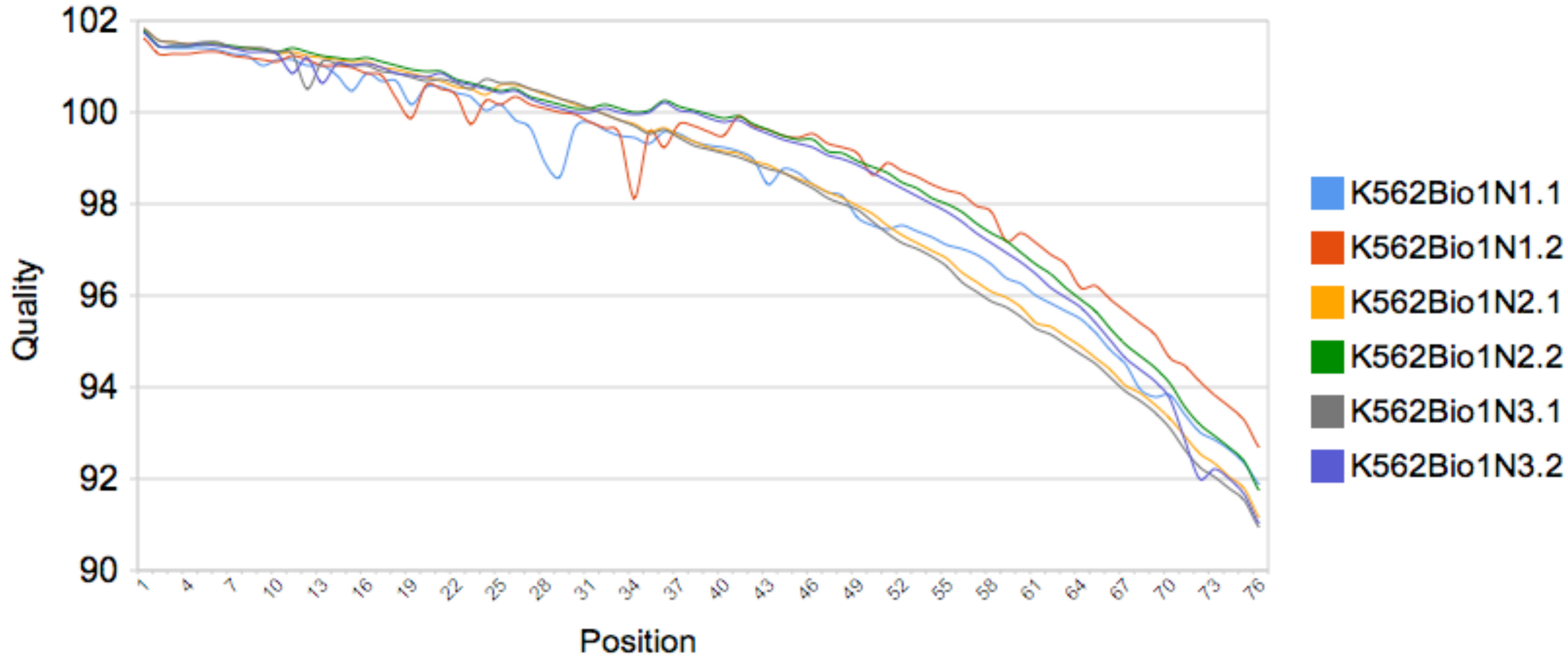


Split mapping



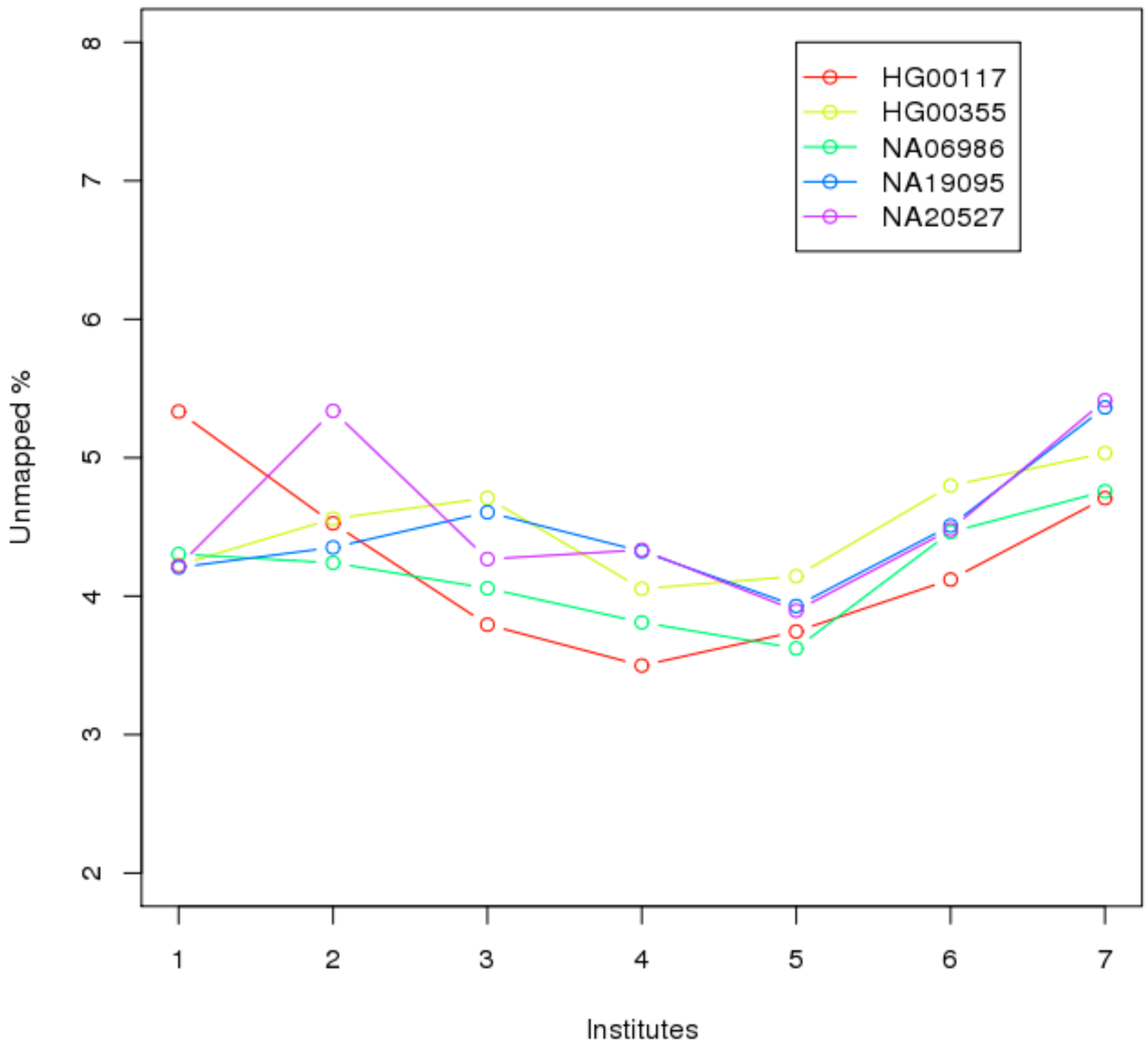
Usually only ~10% of the mapped reads are split-maps,
but for some appications they carry ~90% of information!

Trimming



- nucleotides at the end tend to accumulate more mismatches
- multiple rounds of split-/mapping with increasing trimming steps
- for Geuvadis: entire reads (76nt), quality trimming, trim-to-50nt (trim-to-30nt)
- BAM files contain additionally the information about (genomic) pairing

Geuvadis: Mapping Success of the Sandbox Data across the 7 Institutes



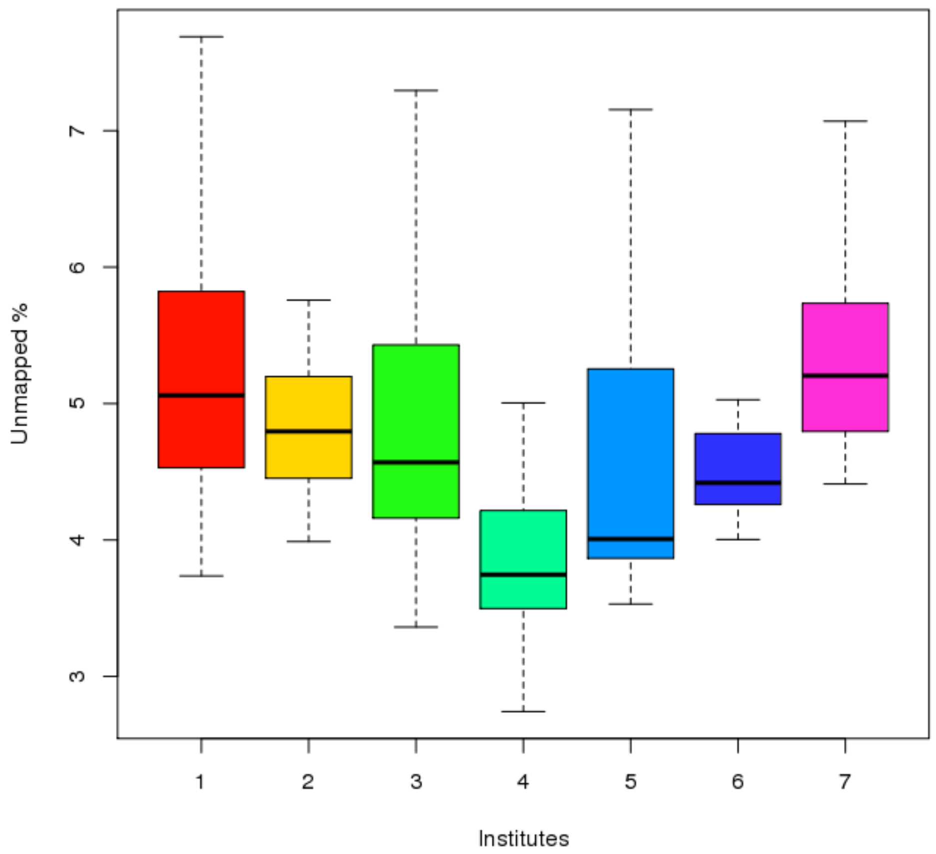
- Approach maps >95% of the data in most datasets

- Marginal variations between samples/institutes

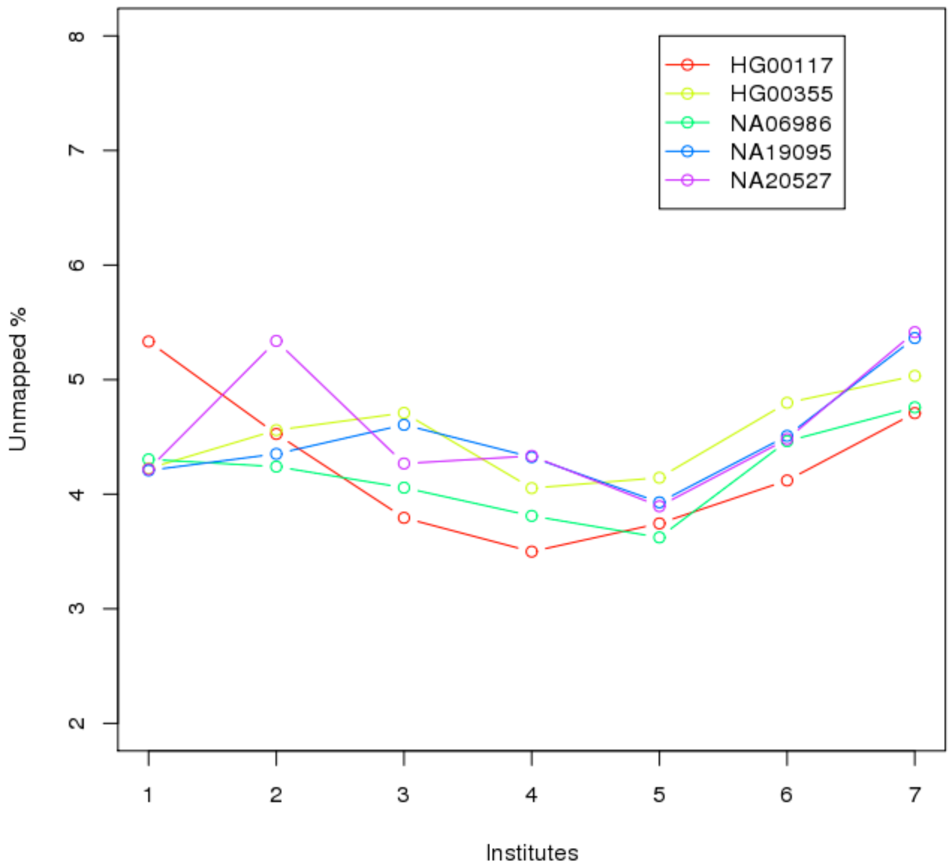
- No dominant effect of the sample in the Sandbox Data, some samples support an institute-trend

Geuvadis: Mapping Success of All Datasets by Institution

All Datasets of every Institute (unequal sets)



Sandbox Data (distributed to all)



Differences: sample, number of samples (> 2-fold), experimental influences, ...

Contamination by EBV

- Virus used to transfect samples, virus load can be differentially high in the cells at the time point of RNA extraction

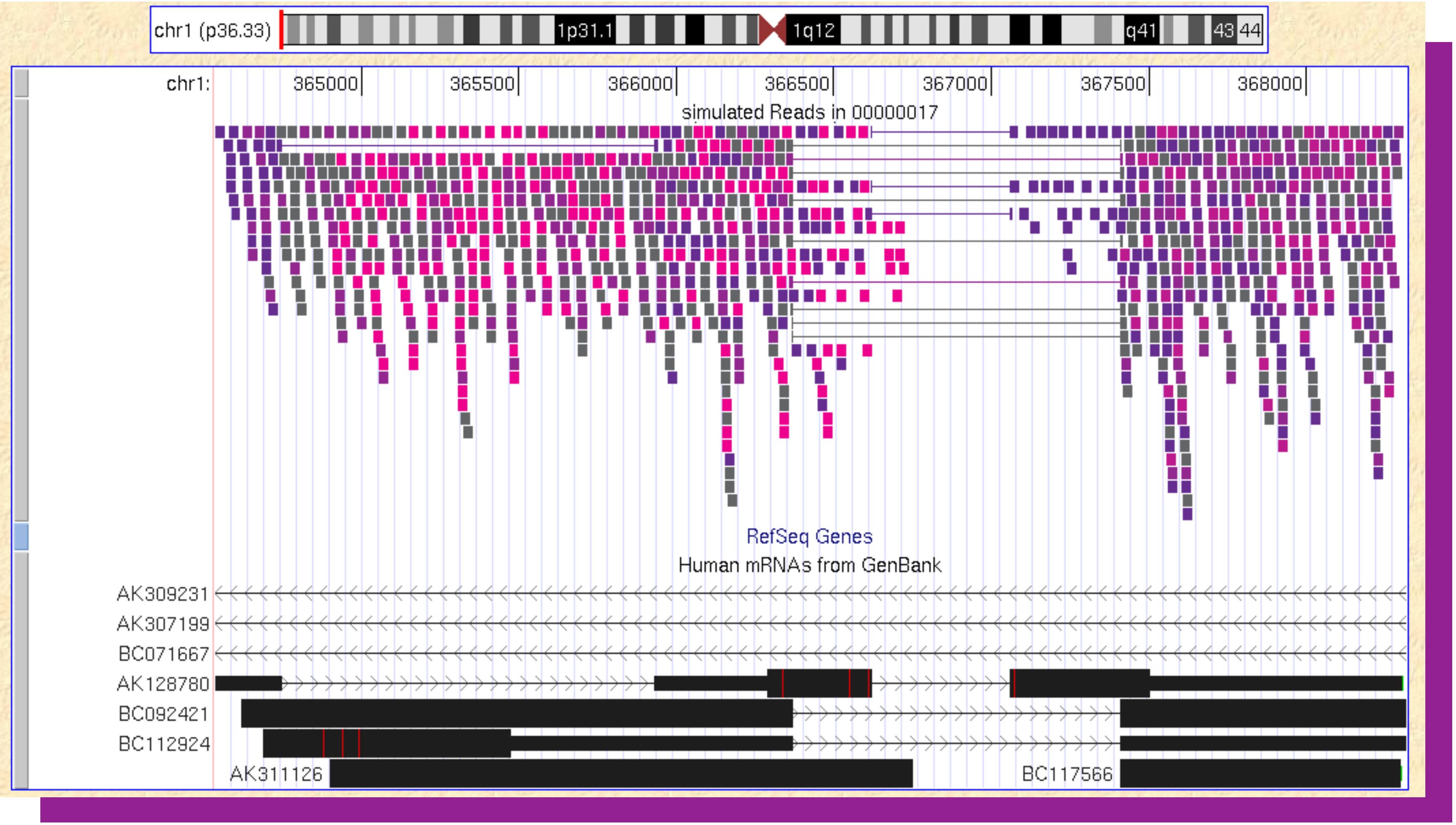
- Does the reads that origin from the virus falsify the mappings?

first → second

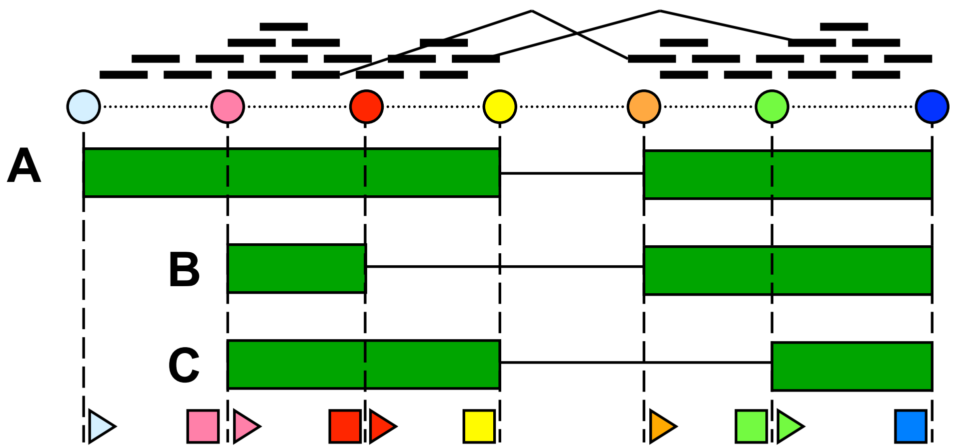
Sample	hu.uniq-76	eb.uniq-76	hu.uniq-76	eb.uniq-76
HG00355	60,803,244	320,980	60,803,217	320,304
NA06986	52,080,577	125,951	52,080,552	125,378

~0.00004% of the human unique mappings (27, respectively 5) are not unique anymore

Deconvolution



Splice Graph + Reads = Flow Network



Annotation mapping

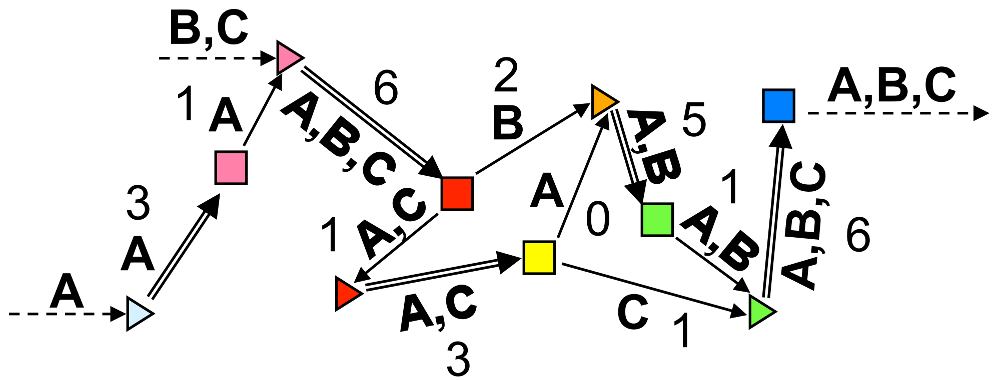
Superimpose reference annotation to genomic mappings

Flow networks

transportation problem
bipartite matching
assignment problem
transportation problem
genomic assembly
(repeats)

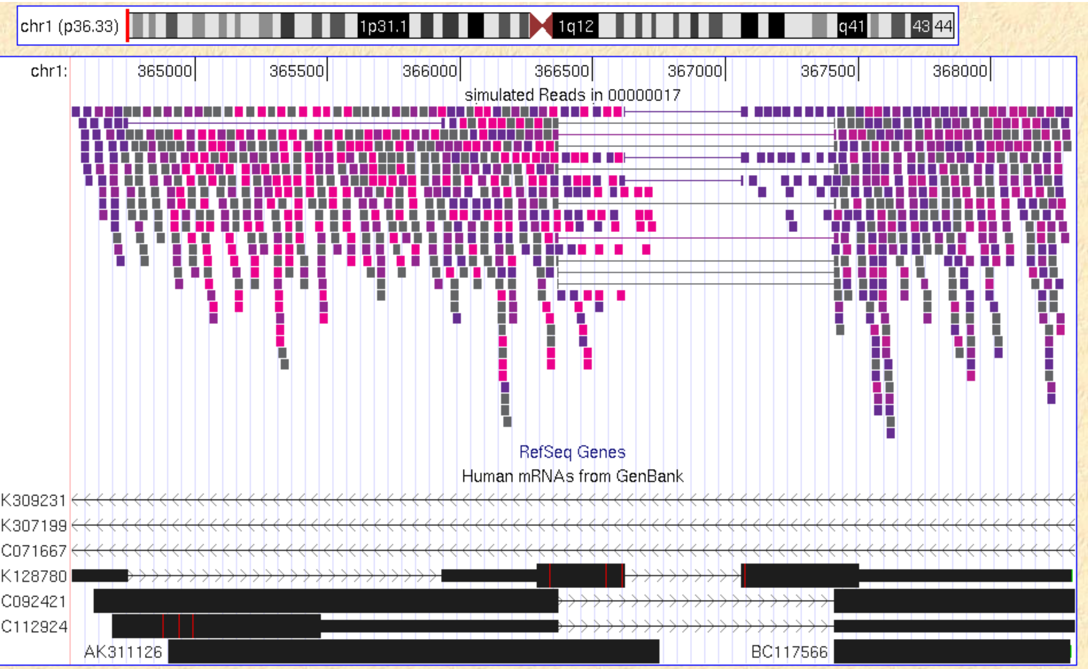
Flow vs. Flux

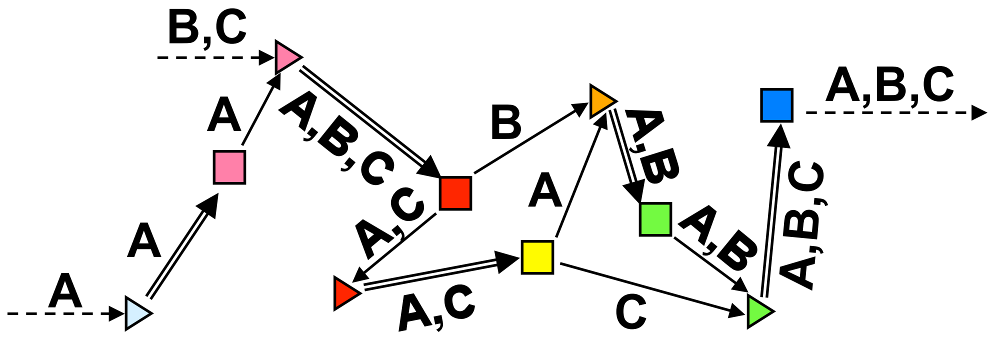
readcount along exon length

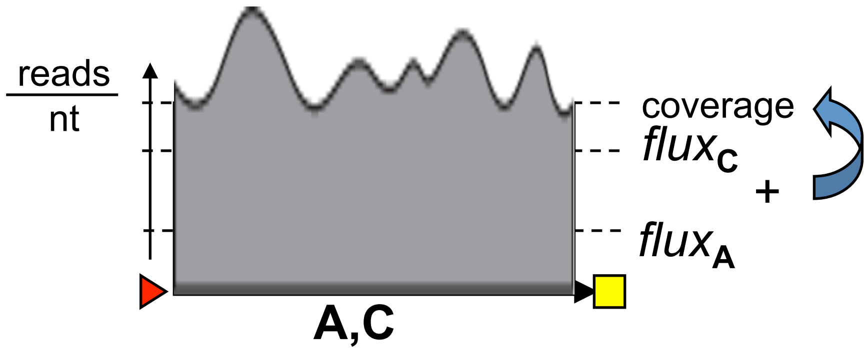


Inverse Transportation Problem, Flow Network Stabilizes Noise

Flux Capacitor: Algorithm Outline







edge ▶ ⟶ ▢ poses the constraint:

$flux_A + flux_C$ +/- $error_{cov}$ = coverage

respectively

$flow_A + flow_C$ +/- $error_{reads}$ = readcount

⟶ set of constraints across network

⟶ solve as a linear program, OF: *minimize error*

⟶ output the predicted expressions $flux_X$ resp. $flow_X$

$flux_X$:= coverage [reads/nt] across whole transcript **X**

$flow_X$:= expected number of reads sampled from **X** between ▶ and ▢

$= \int p_X(x)\, dx$

Normalization: Straightforward

The **RPKM** value [Ali & Co 2008]

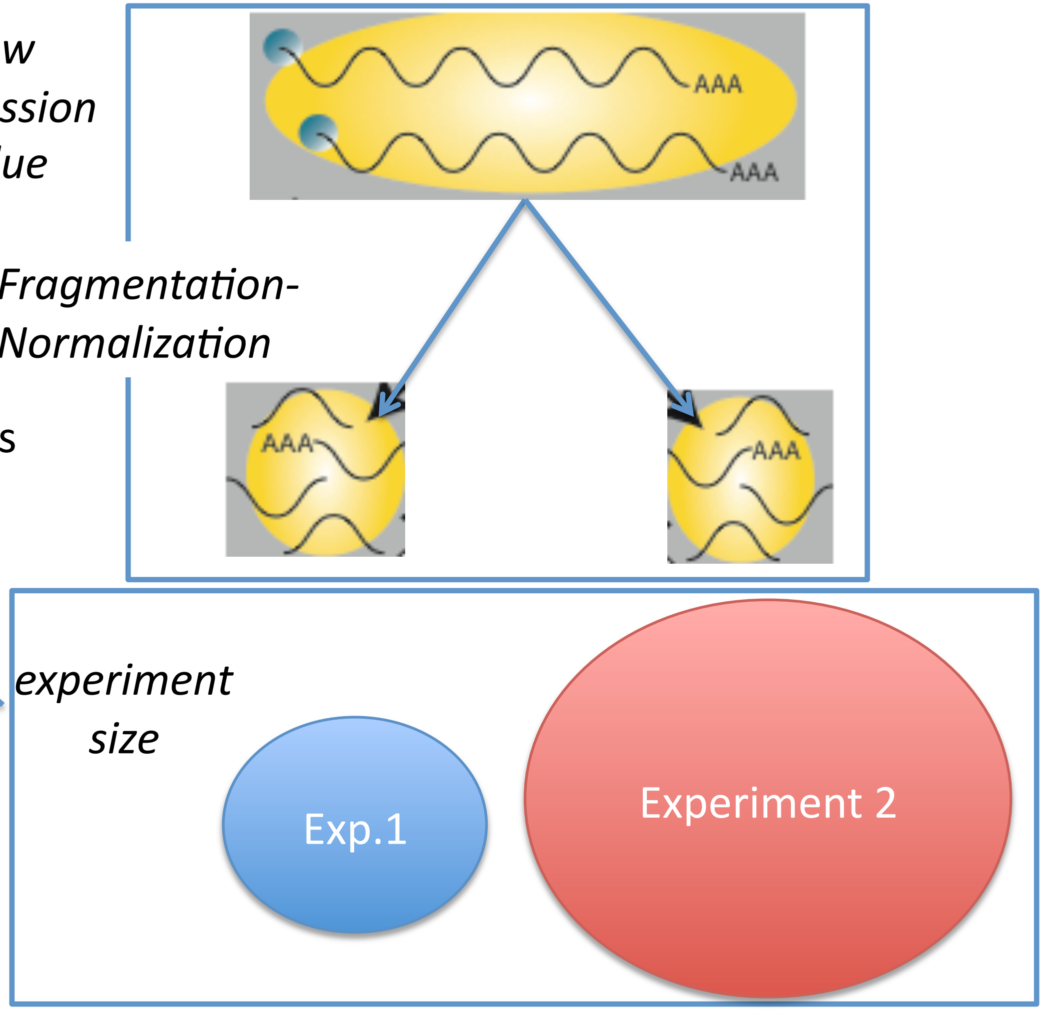
raw expression value

Reads

Fragmentation-Normalization

Per **K**ilo-Base

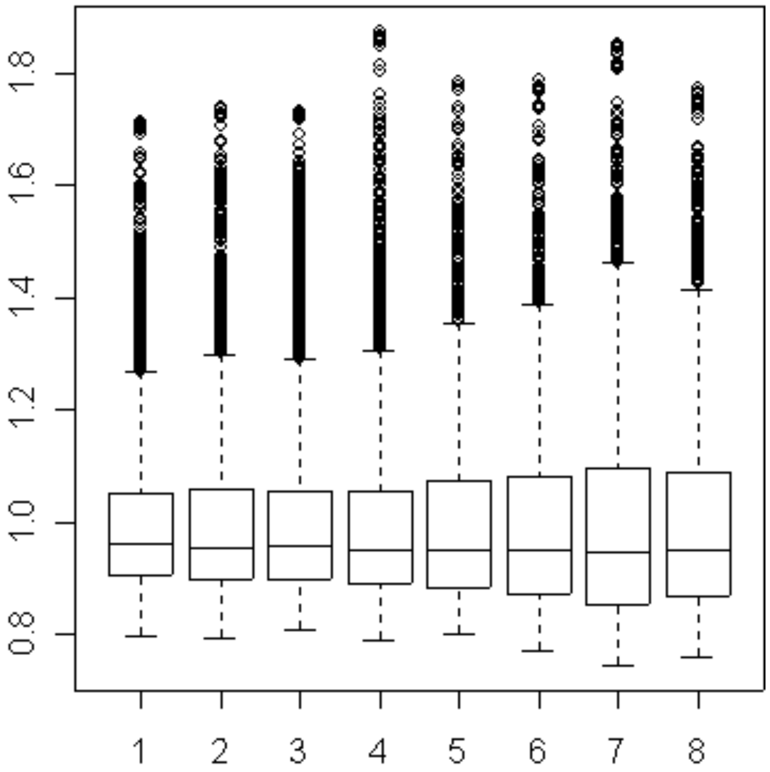
per **M**illion mapped Reads

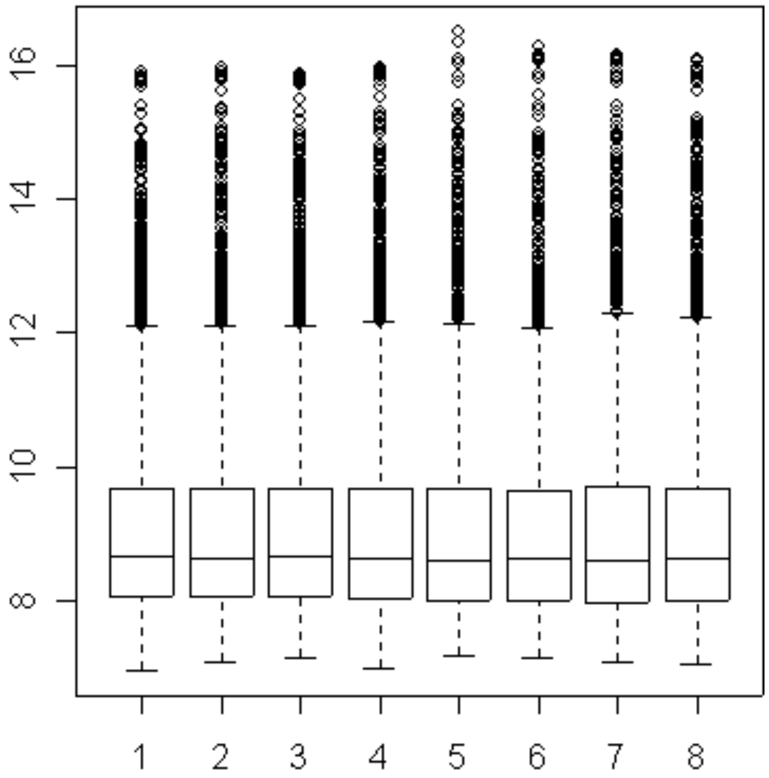


Variations: RPK, FPKM, …

Normalization of Distribution

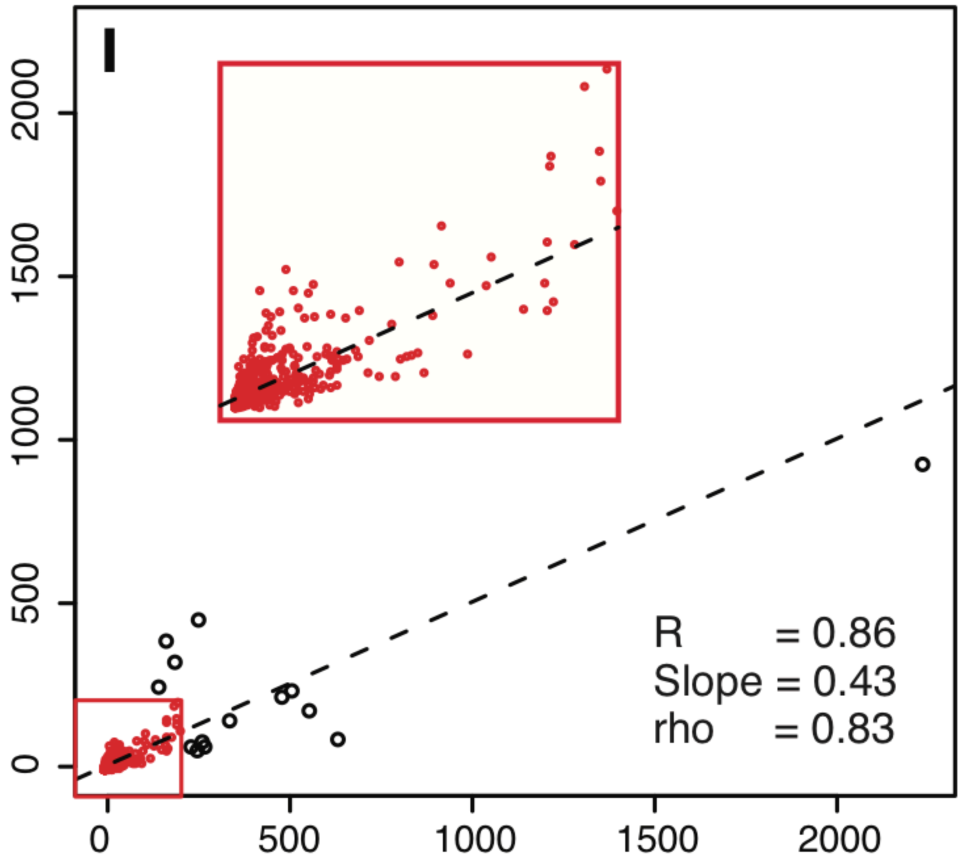
Normalization of the Distribution: e.g., Quantile Normalization, etc.





(this is NO Geuvadis Data)

Normalization to compare Distribution



- Gene expression data has outliers (Zipf's Law)

- Outliers bias statistical indicators based on data point values (e.g., Pearson's product-moment coefficient, LSLR, dispersion, etc.)

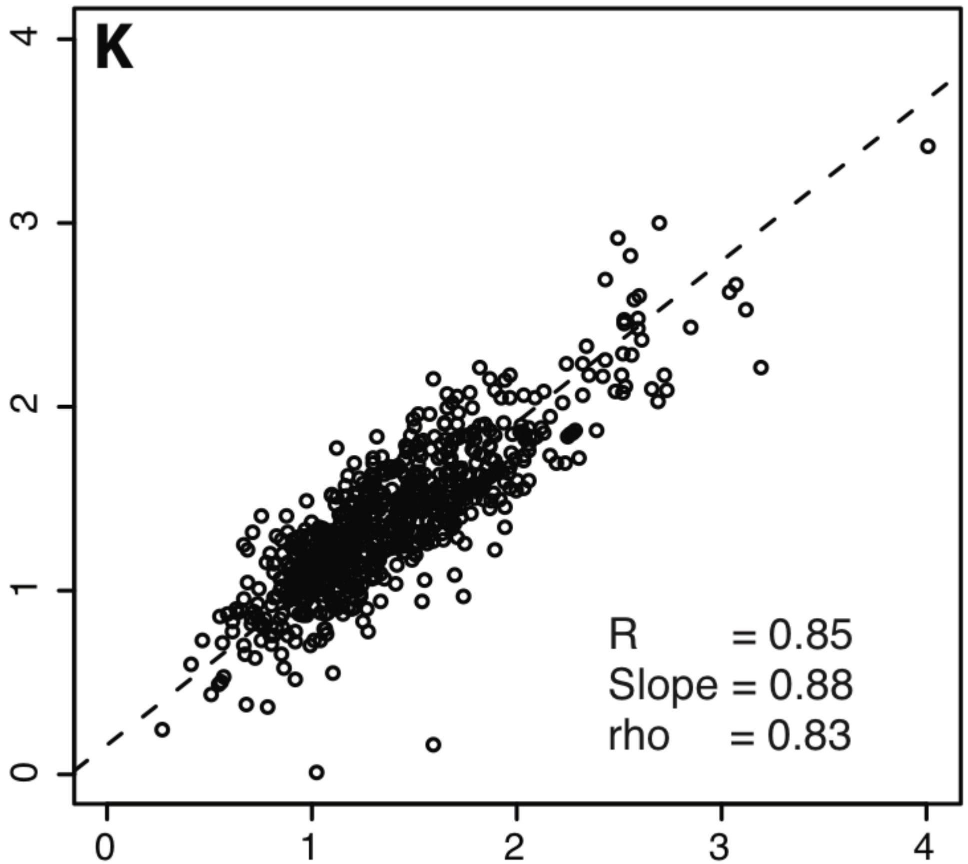
- Some indicators are robust (e.g., Spearman's rank corr.)
  BUT do not assess gene expression similarity

- Alternative: Normalization before stat. assessment

	Number of removed outliers									
	0	1	2	3	4	5	6	7	8	9
R	0.86	0.72	0.76	0.76	0.75	0.75	0.75	0.76	0.77	0.77
Slope	0.43	0.47	0.54	0.59	0.62	0.65	0.68	0.72	0.75	0.67

(this is also NO Geuvadis Data)

Normalization to compare Distribution



- Gene expression data has outliers (Zipf's Law)

- Outliers bias statistical indicators based on data point values (e.g., Pearson's product-moment coefficient, LSLR, dispersion, etc.)

- Some indicators are robust (e.g., Spearman's rank corr.) BUT do not assess gene expression similarity

- Alternative: **Normalization before** stat. assessment

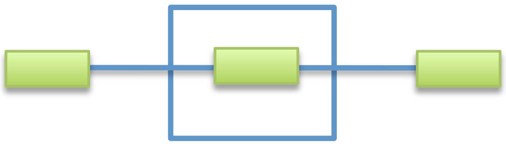
	Number of removed outliers									
	0	1	2	3	4	5	6	7	8	9
R	0.85	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.83	0.83
Slope	0.88	0.88	0.89	0.89	0.90	0.90	0.90	0.91	0.91	0.90

(this is also NO Geuvadis Data)

Quantification of other Elements

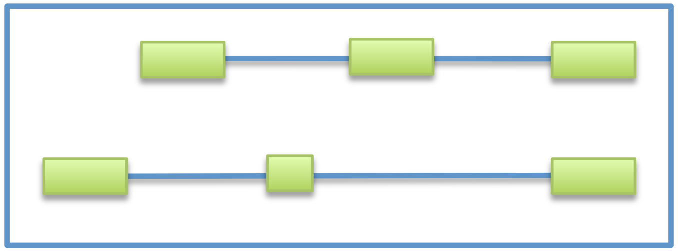
Extrapolation of Transcript RPKM alternatively to
Re-quantification by complementary methods

(A) Exons



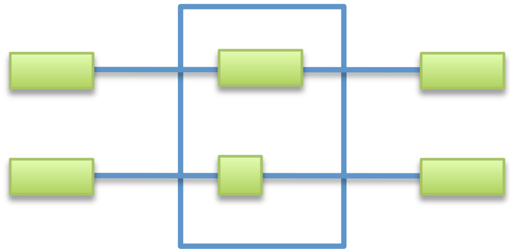
unique boundaries
vs.
genomic overlap

(B) Genes



protein-coding units
vs.
genomic loci
(hybrid transcripts,
nc transcripts)

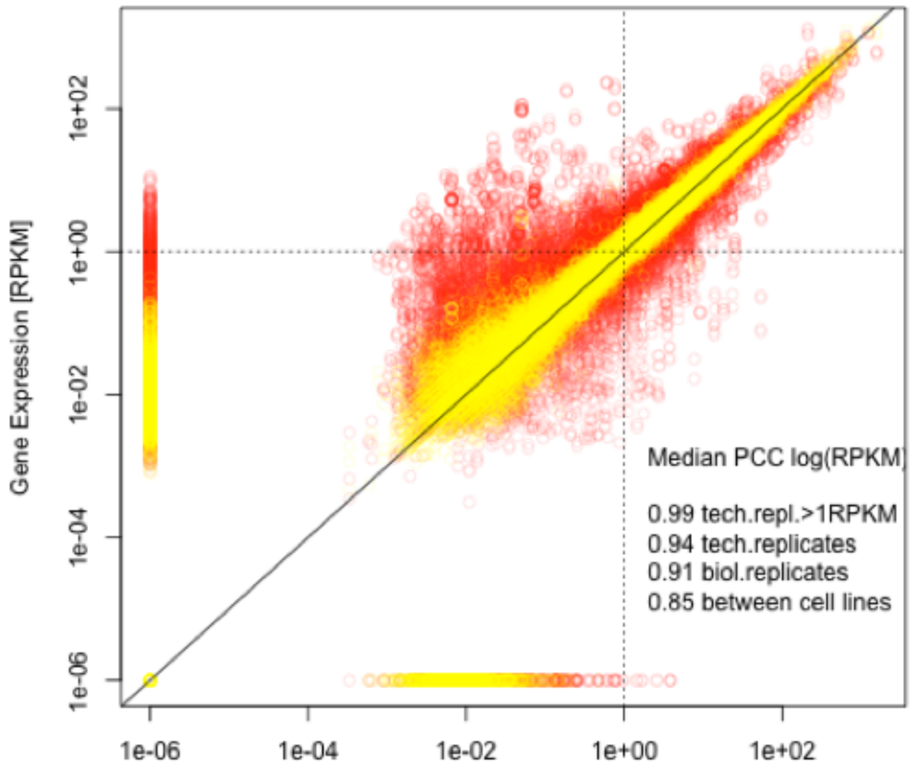
(C) AS Events



AStalavista
[Sylvain & Micha,
2007-2009]

Sylvain
Foissac

Comparison of Expression Values



-Different approaches depending on question / element that is compared

- Some Statistics do not require (much) a priori normalization (e.g., comparison of same element in different states)
- Here, comparison of Gene Expression Landscape by Pearson coefficients

(again, this is NO Geuvadis Data)

Acknowledgements