Discovery and Quantification of RNA with RNASeq

Roderic Guigó Serra

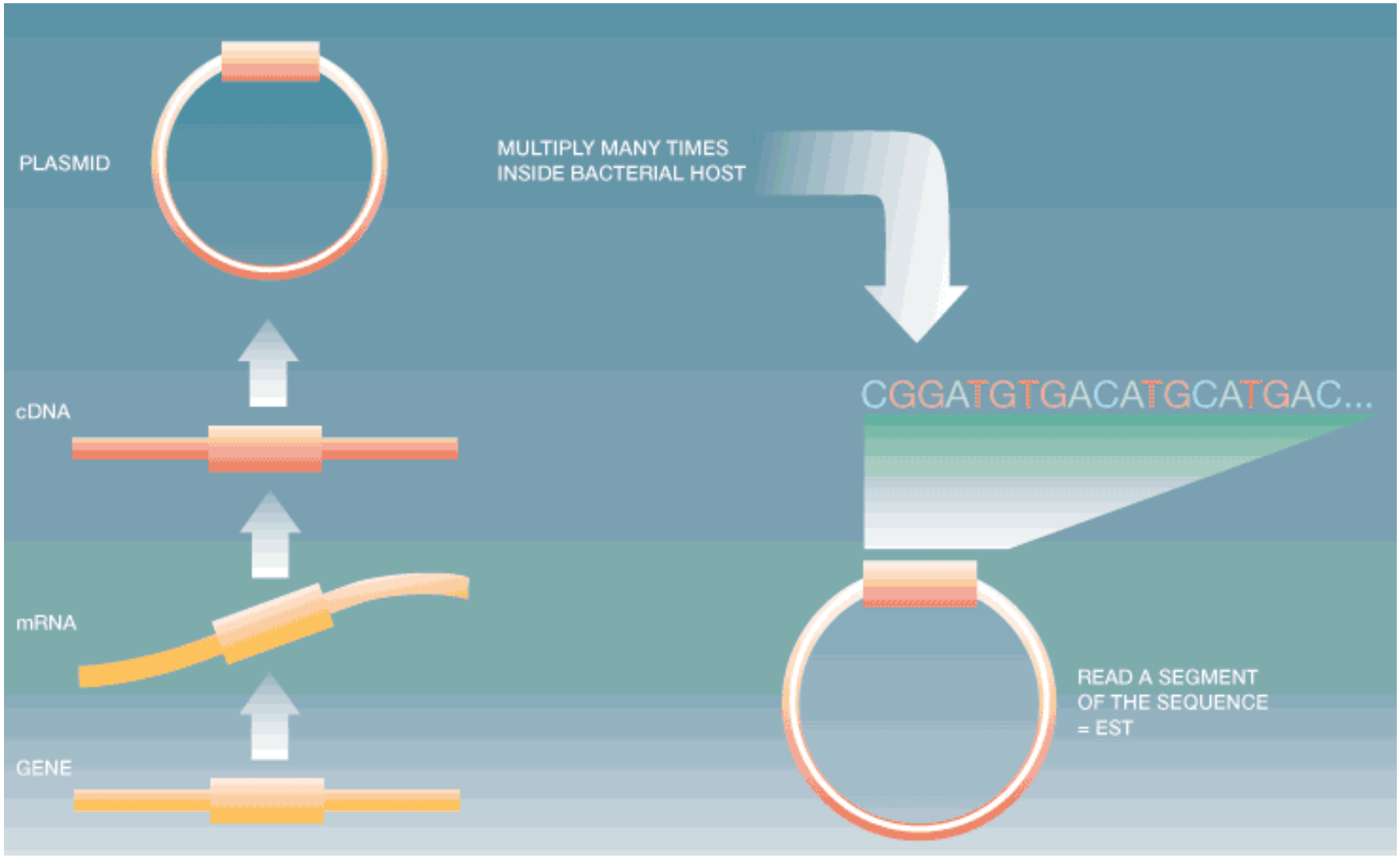
Centre de Regulació Genòmica (CRG)

roderic.guigo@crg.cat

RNA

- Transcription to RNA and subsequent processing is the first step in the unfolding of the intructions encoded in the DNA.
- There is a one-to-one correspondence between the RNA content of the cell and the cellular phenotype (which is obviolusly not true for the DNA)
- The phenotypic effects of the variations in the DNA are ultimately mediated by variations in the RNA content of cell.

cDNA cloning and sequencing

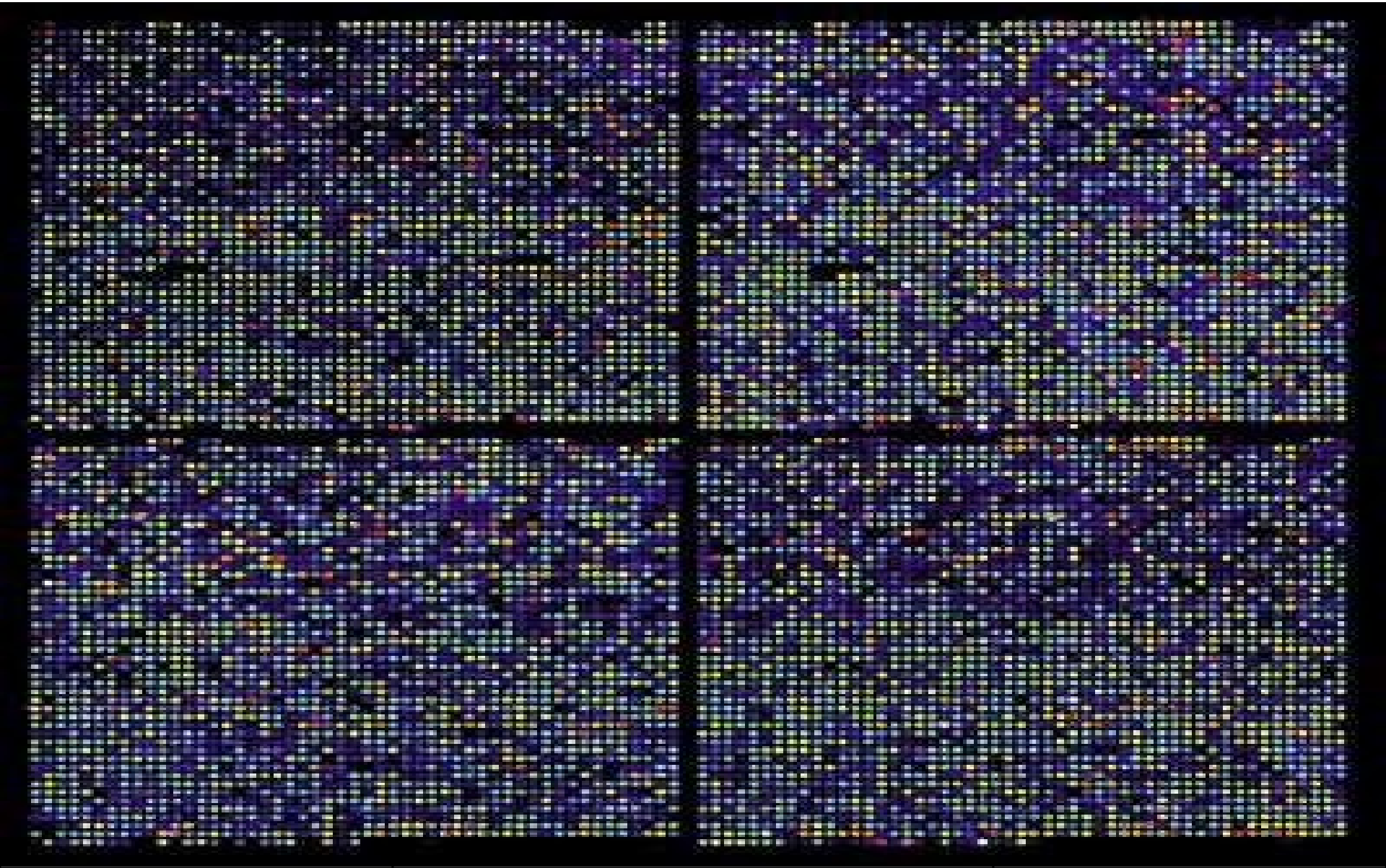


IBM Systems Journal, Inman et al (2001) , Deep Computing for life Sciences

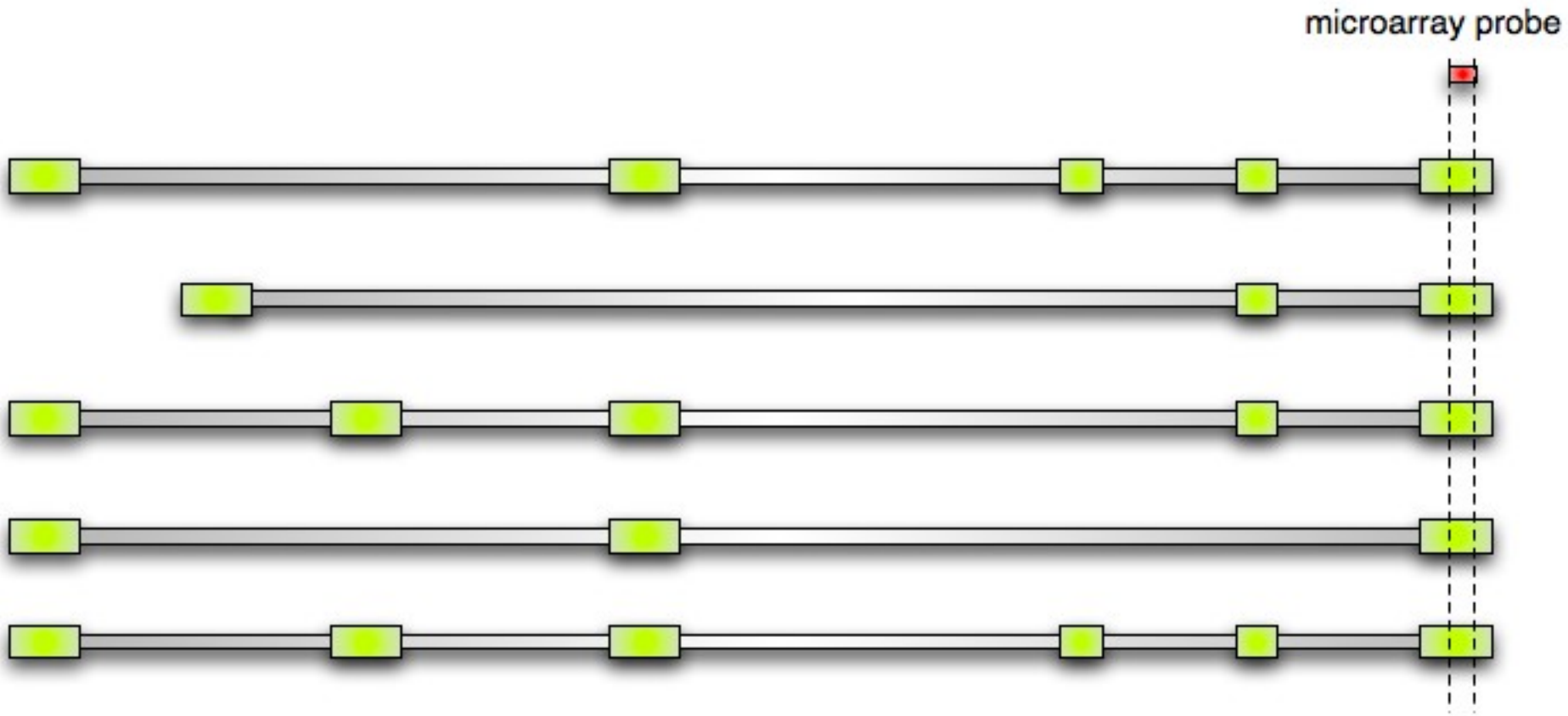
high dynamic range in RNA abundance

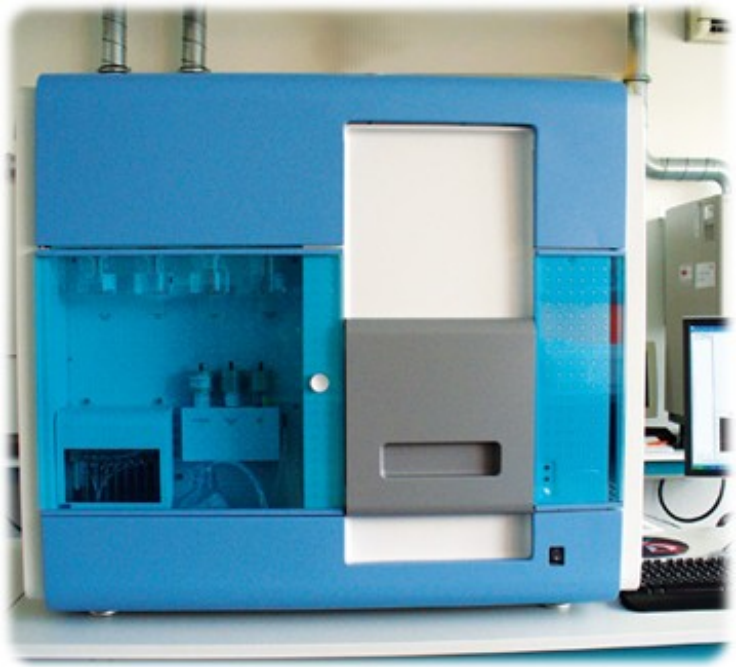
- RNA species vary greatly in abundance. Typically (Carninci et al., 2000)
  - 5-10 **highly expressed** species. thousands of copies ~ 20% of the mRNA mass
  - 500-2000 **intermediate** species. hundreds of copies 40%-60% of the mRNA mass
  - 10,000-20,000 **rare** messages. a few copies <20%-40% of the mRNA mass
- Random clone selection is ineffective in recovering low abundance transcripts.
- Normalization strategies are required, but the destroy transript abundances

DNA microarrays



Alternative splicing





Next generation sequencing technologies.

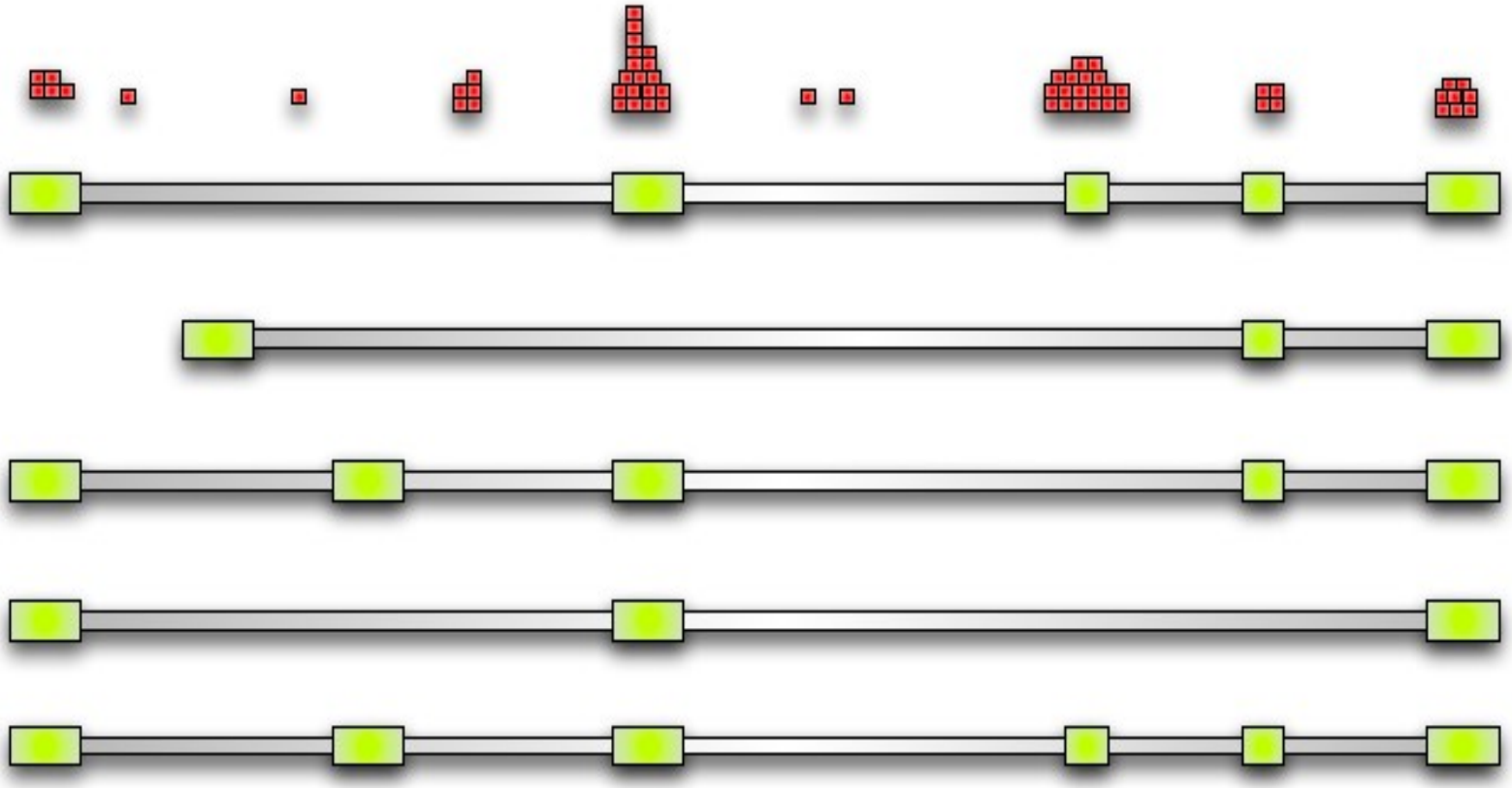
- They MAY provide enough sequence depth, so that the (unbiased) sequencing of the RNA complement of the cell MAY become feasible.

Deep sequencing of PTB knockdown

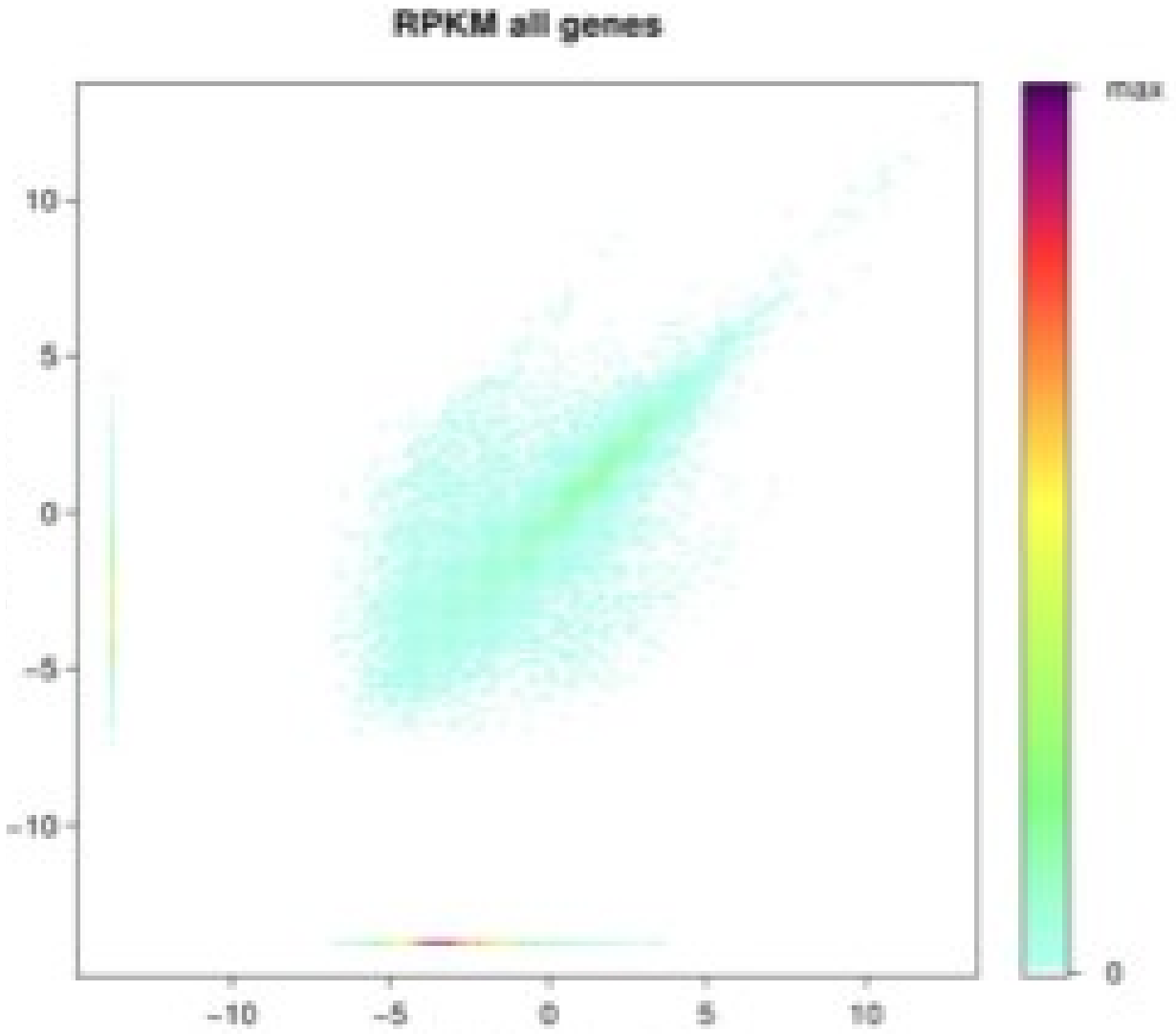
- The polypyrimidine tract binding protein (PTB) is an important regulator of alternative splicing
- Poly A+ RNA from Hela cells in two conditions
  - Control
  - Knock-down of PTB
- Single reads 11M +10.7M
- Paired-end reads
  - Short insert size: 200-250 nt 41M + 21.5M
  - Long insert size: 600-650 nt 32.6M + 17.7M
- Total 135M reads (85M control)

Vincent Lacroix
Heinz Himmelbauer
Juan Valcárcel
Chris Smith

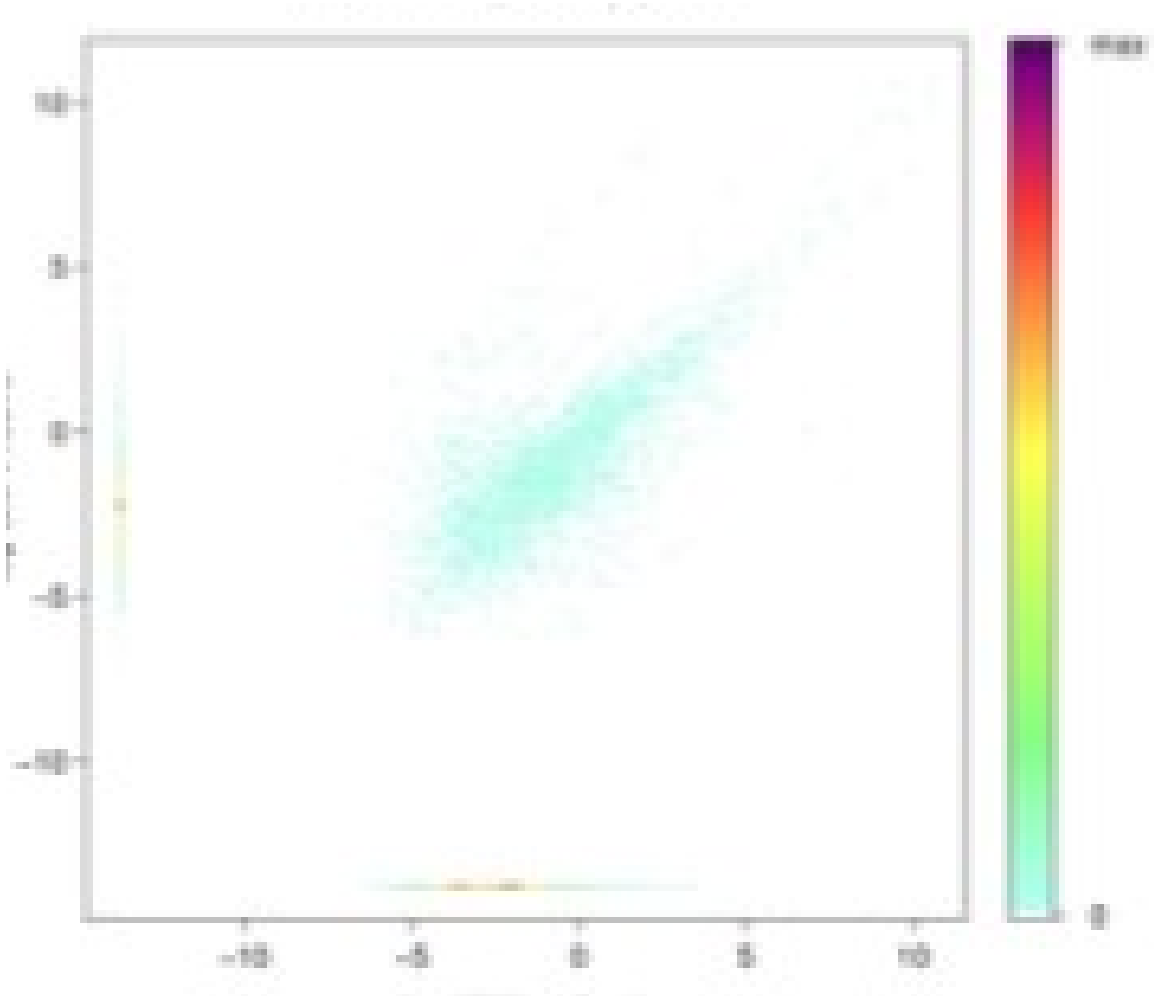
Mapping reads to the annotated genome/transcriptome



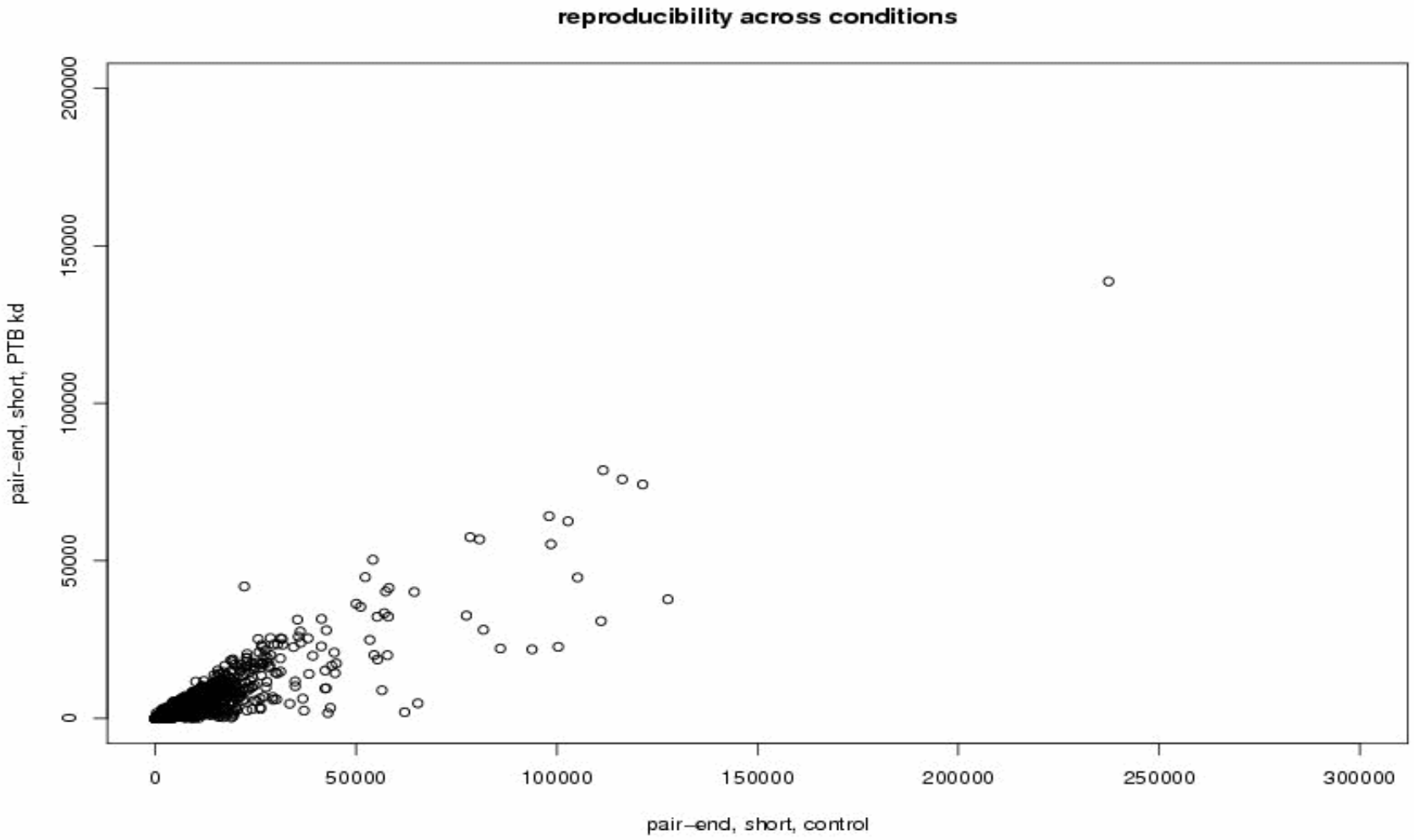
Differential gene expression



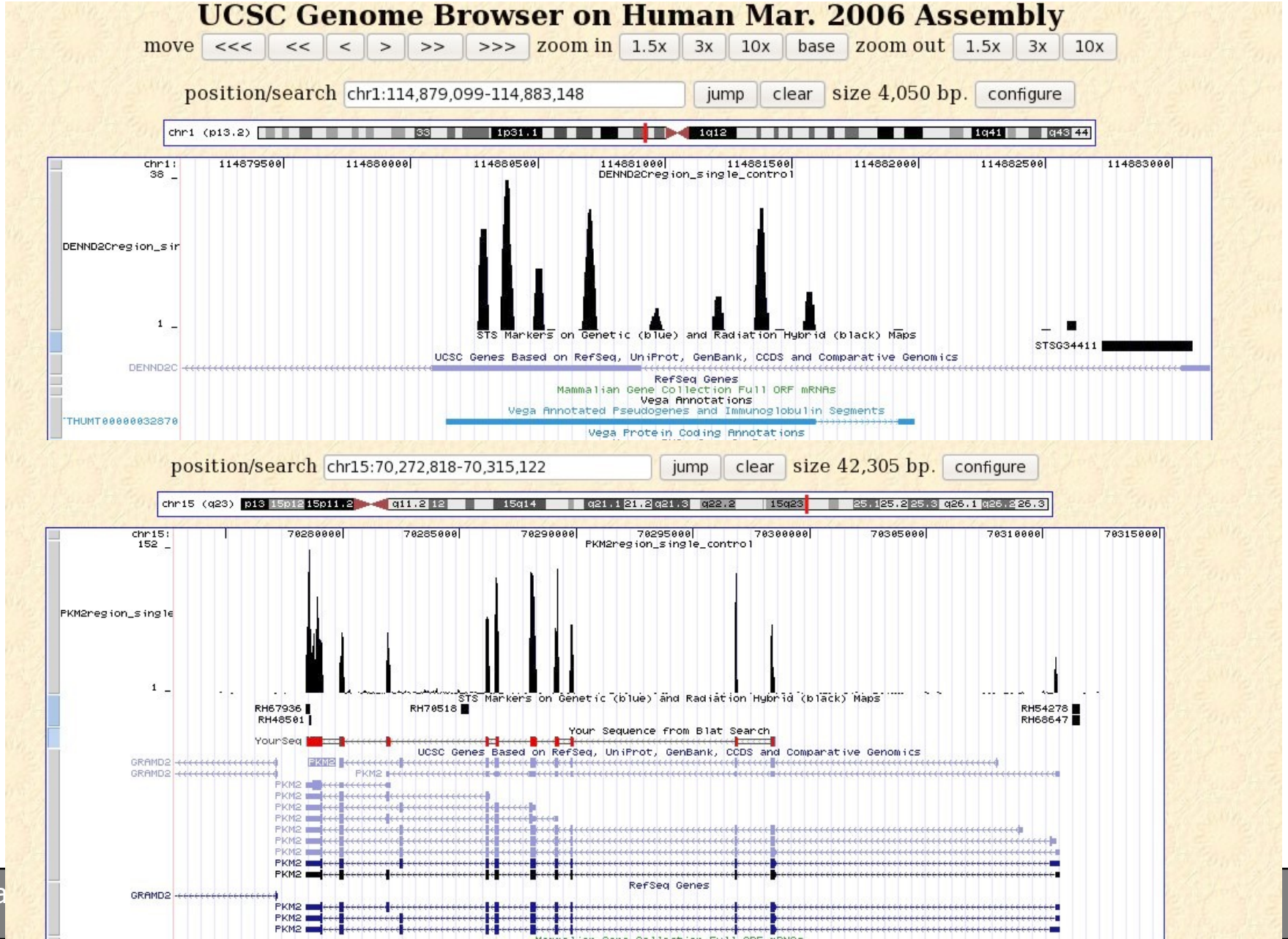
Differential gene expression in non coding genes



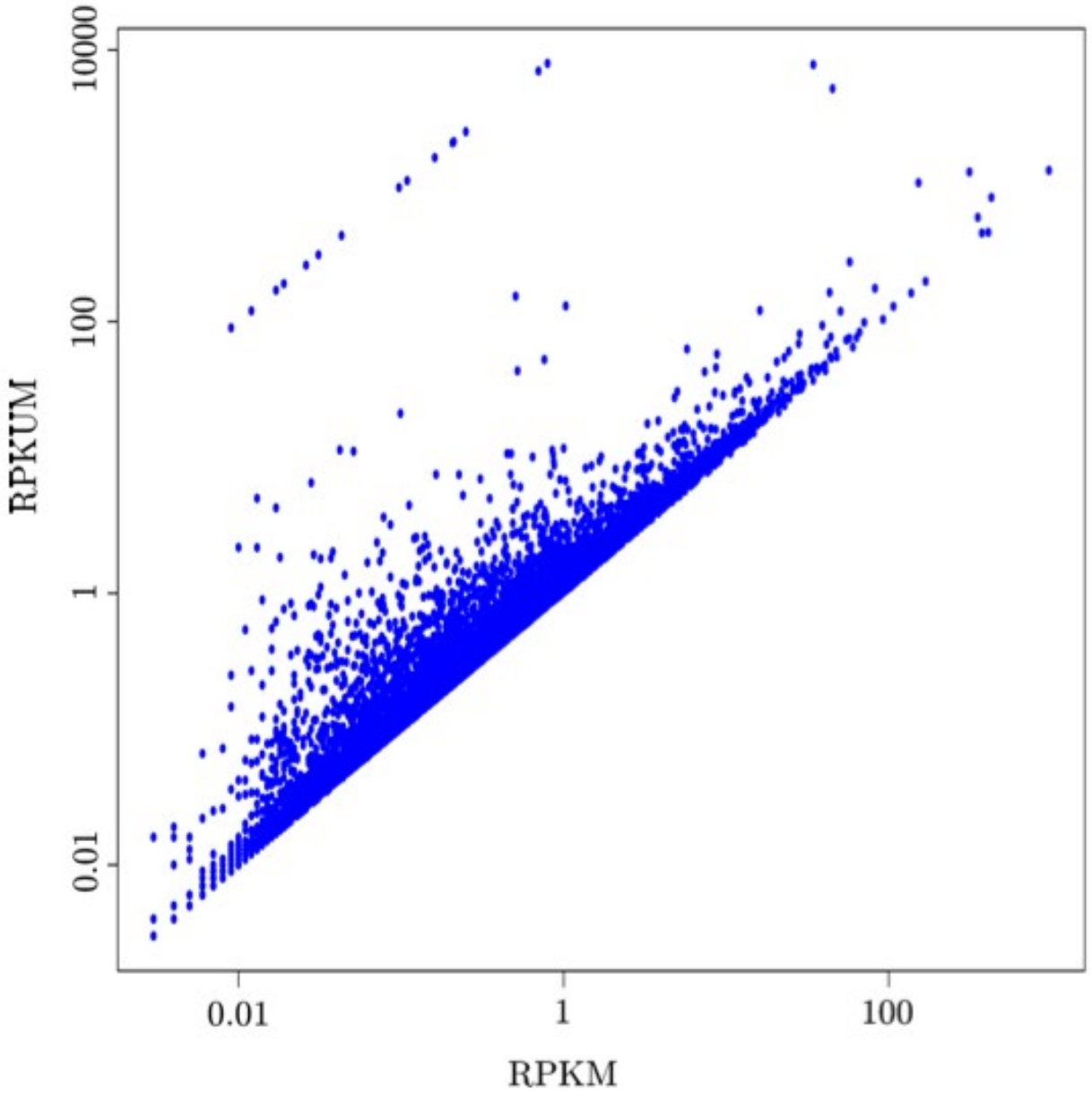
Differential gene expression between control and knockdown



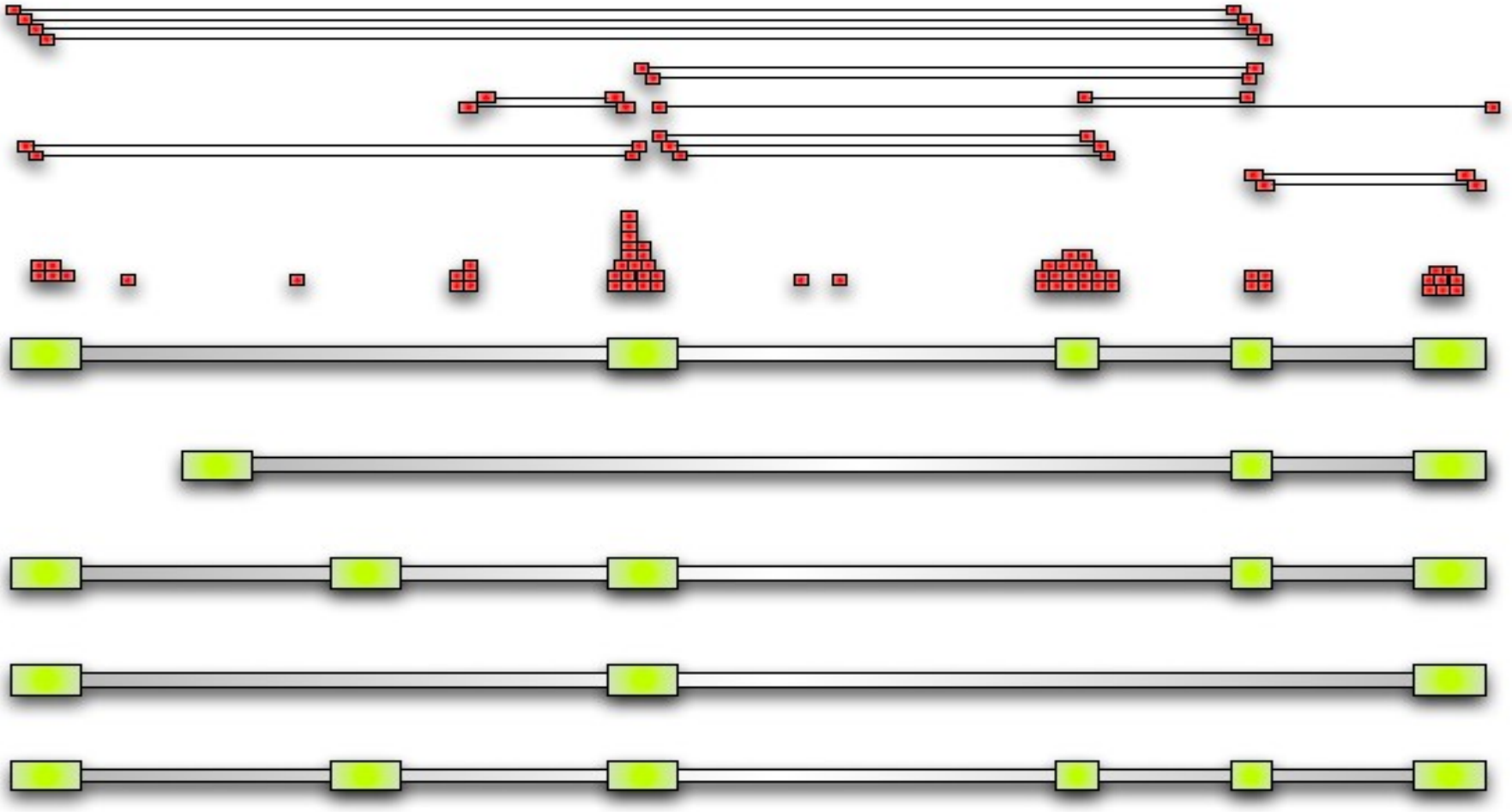
Mapping of RNAseq reads is not trivial



Mapping of RNAseq reads is not trivial



Reads mapping to the splice junctions

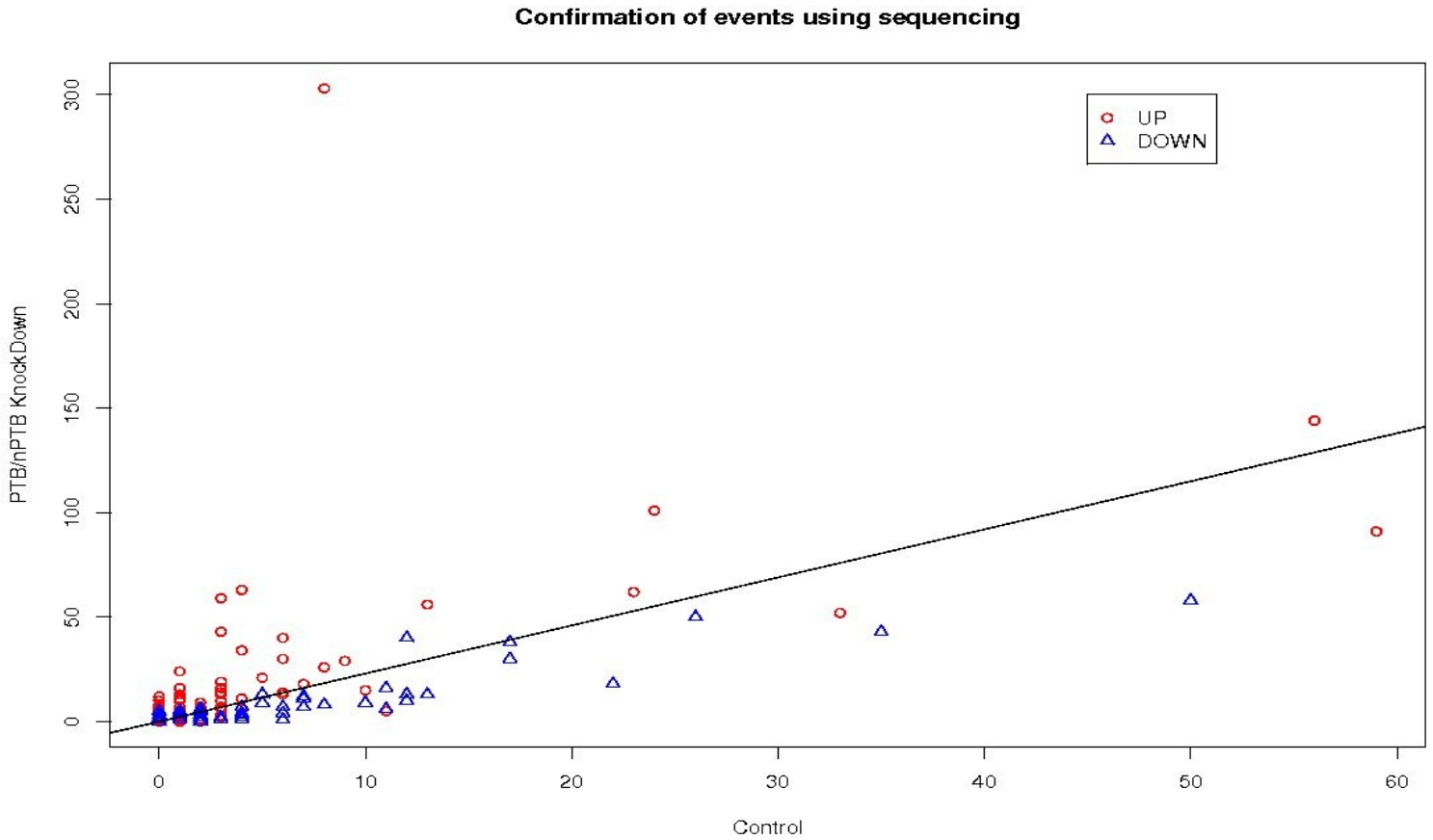


RNAseq vs splicing arrays

EURASNET CONSORTIUM.

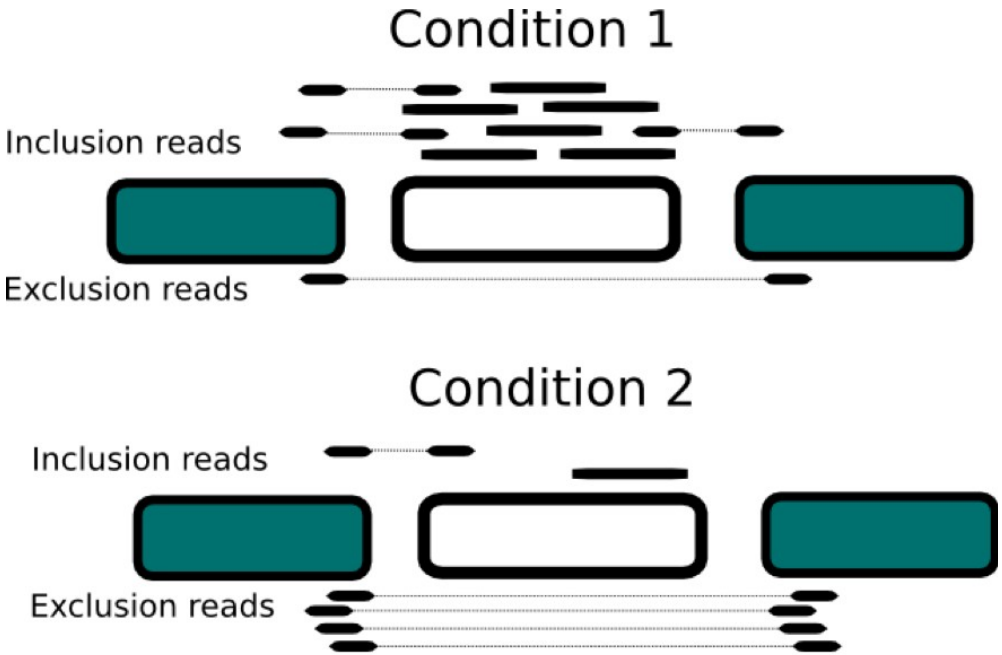
255 known targets

estalished using splicing arrays, and validated through Q-PCR and proteomics



- 75% of the cases detected by arrays are confirmed by RNAseq
  - 81% for highly expressed events (>5 reads)

Detection of new PTB targets



	Condition 1	Condition 2
Inclusion	9	2
Exclusion	1	4

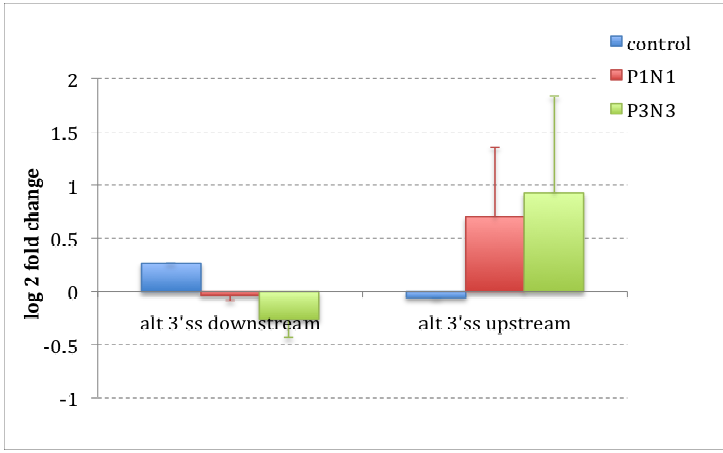
p=0.036

Fisher's exact test

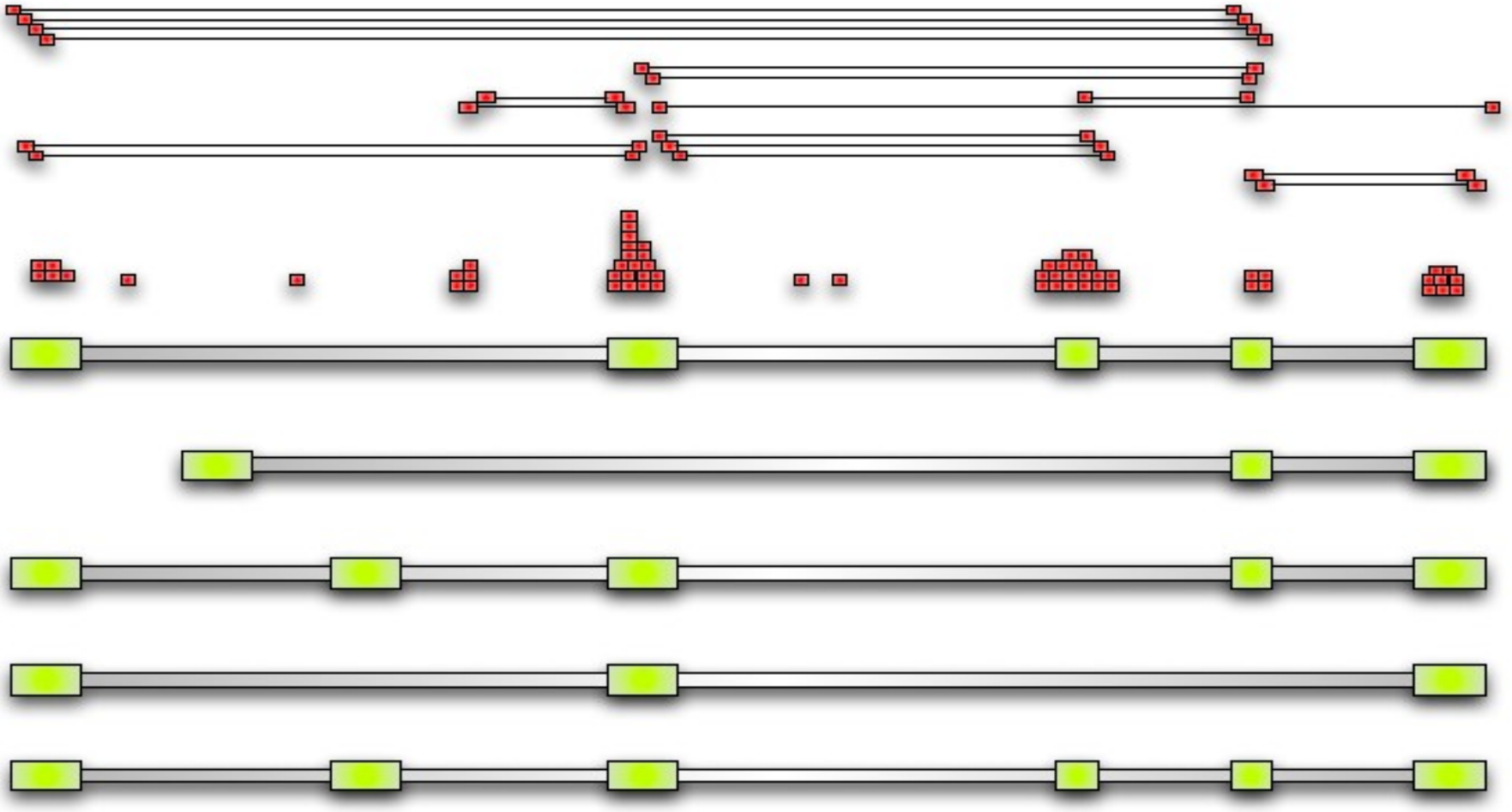
Detection of new PTB targets

31 candidates identified

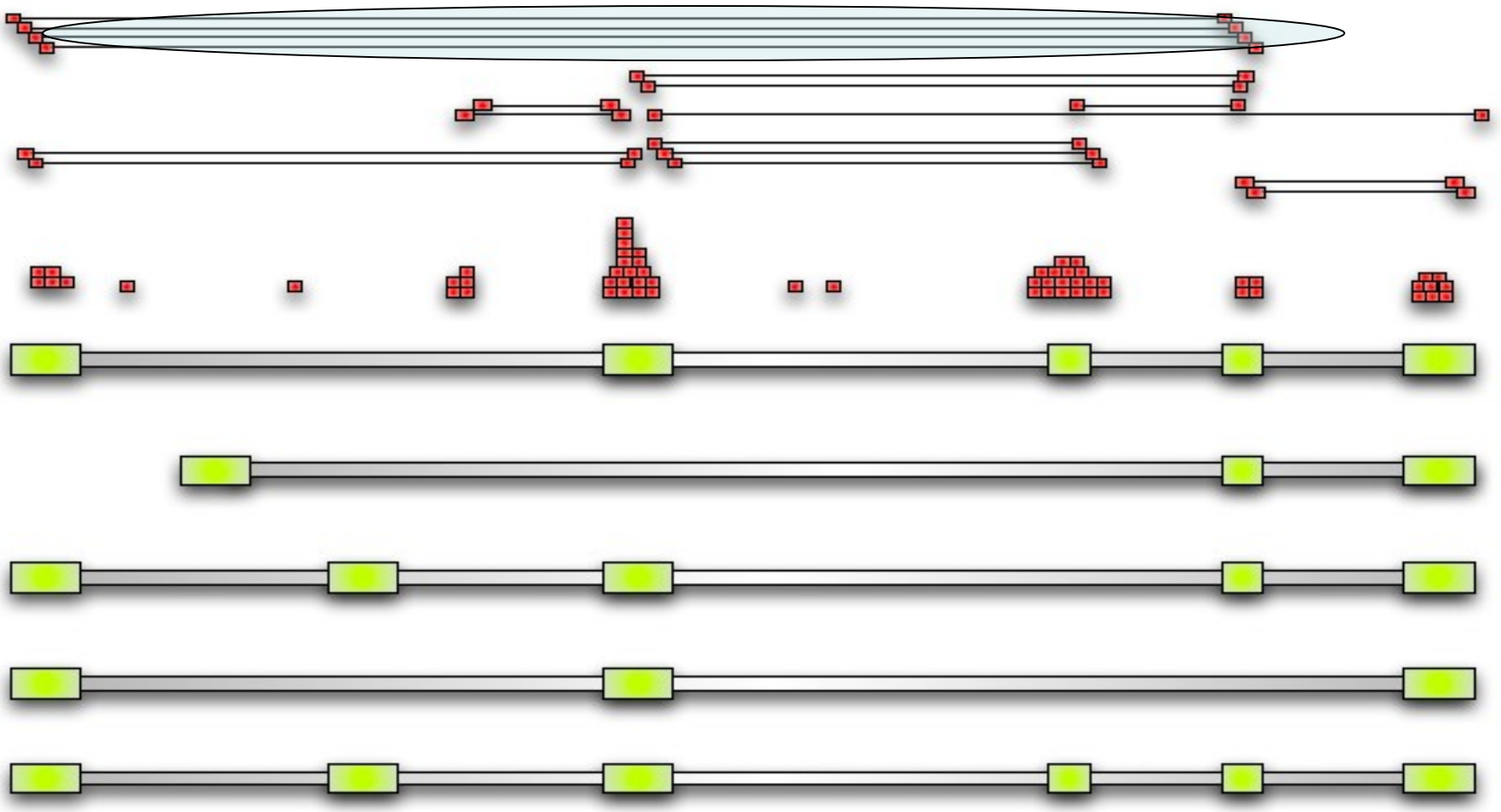
- 30 with evidence of alternative splicing selected for experimental verification by Q-PCR
- **20 confirmed (66.7%).** Identification of previously unknown PTB targets



Novel splice junctions



Novel splice junctions

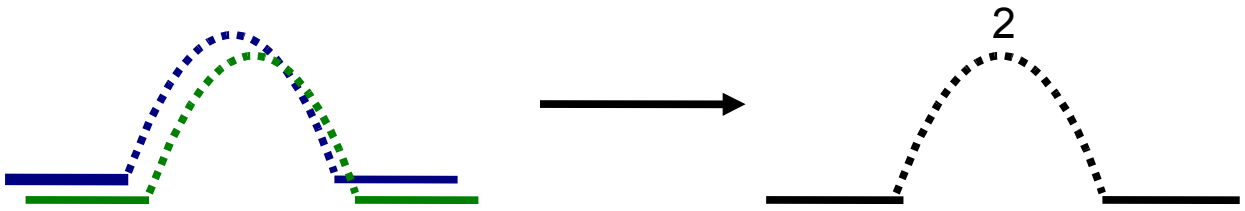


Interchromosomal interactions

- **Paired-end reads**: the two reads are read from the same molecule

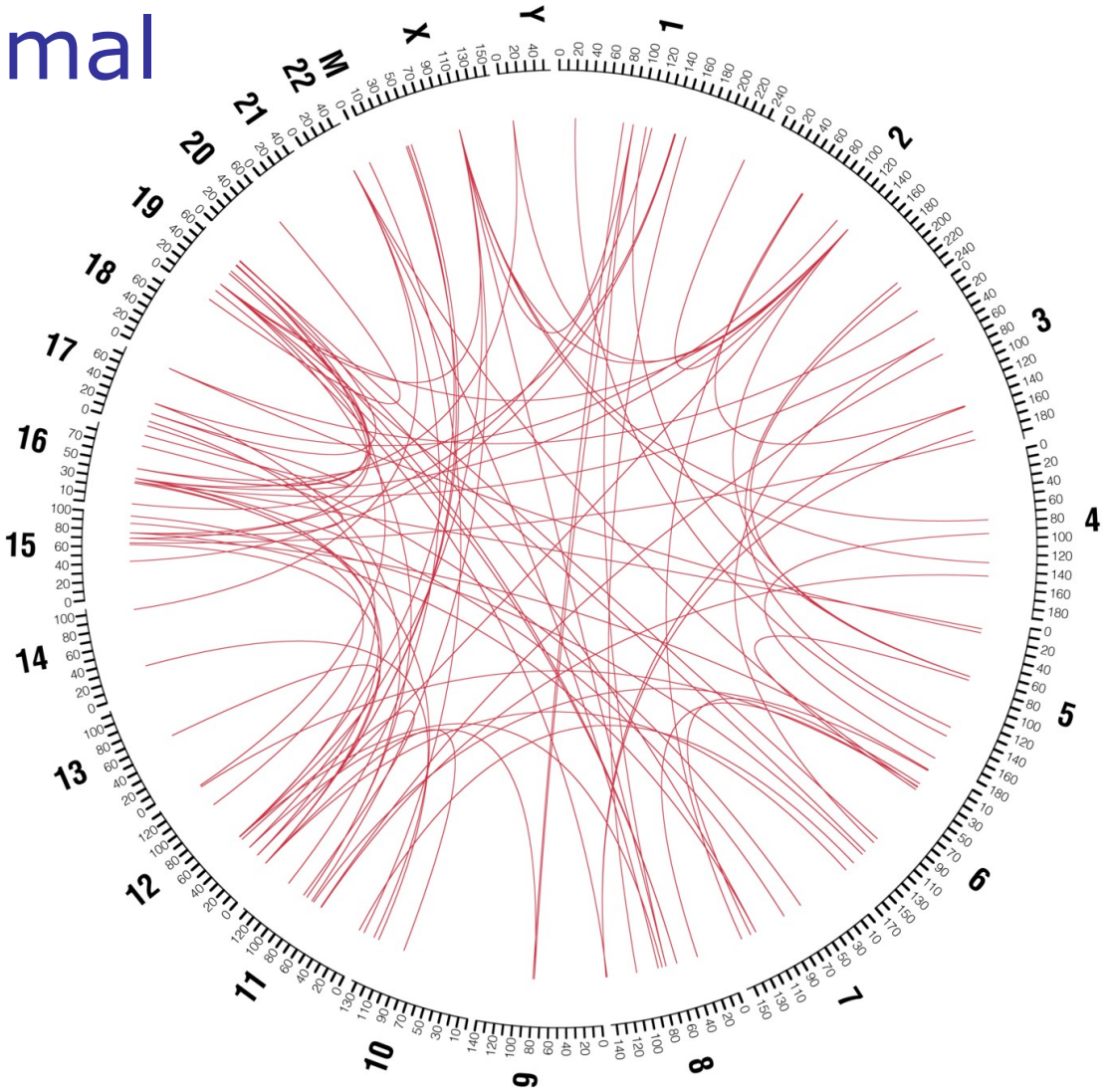
- **Mapping**: The two reads from a pair are mapped independently, and uniquely
- **Interchromosomal interaction:** There are cases where the two reads map to different chromosomes (1.6 % of short inserts, 2.8% of long inserts)
  - 382,639 interchromosomal interactions
    - rRNA 14%
    - chrM 33%

Clustering of read pairs



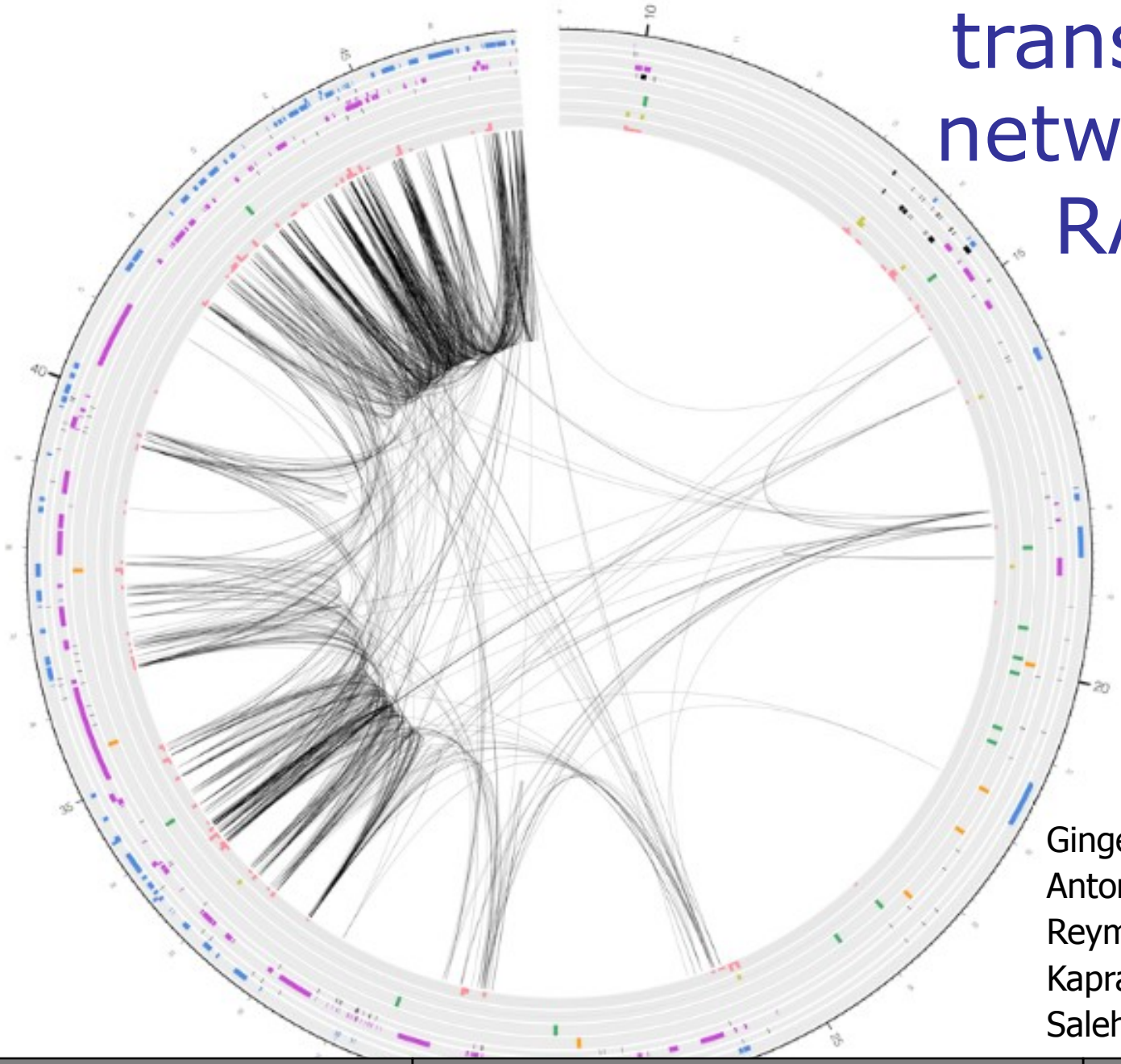
- Read pairs are clustered when both read mappings overlap

Interchromosomal transcripts



Interchromosomal interactions as detected by paired-end reads, long inserts, control
Only interactions supported by more than 100 pairs are shown

transcriptional network. Ch21 RACEarrays

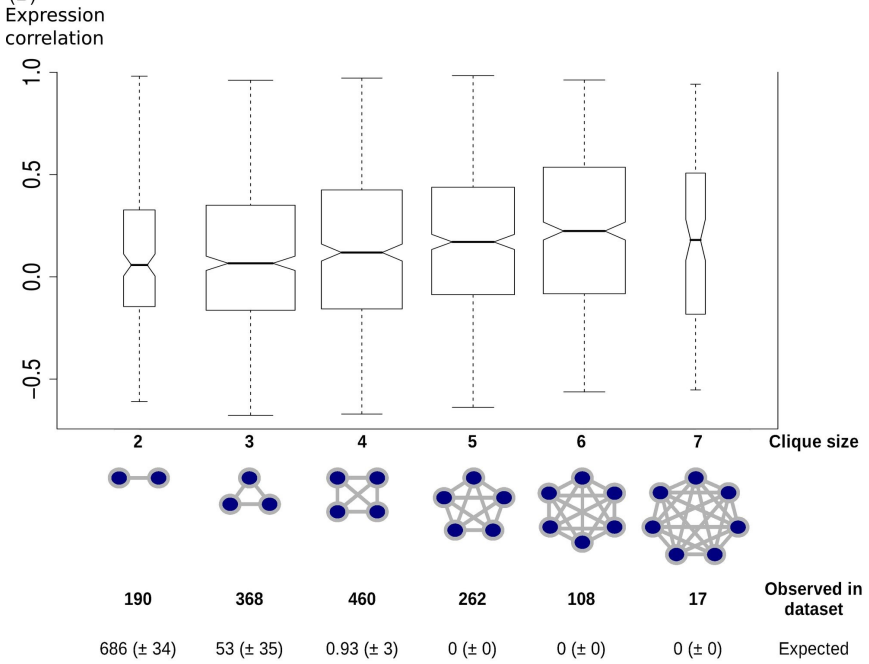


Gingeras,
Antonarakis,
Reymond,
Kapranov,
Salehi-Ashtiani,…

The higher the connectivity of the transcripts, the higher the coexpression

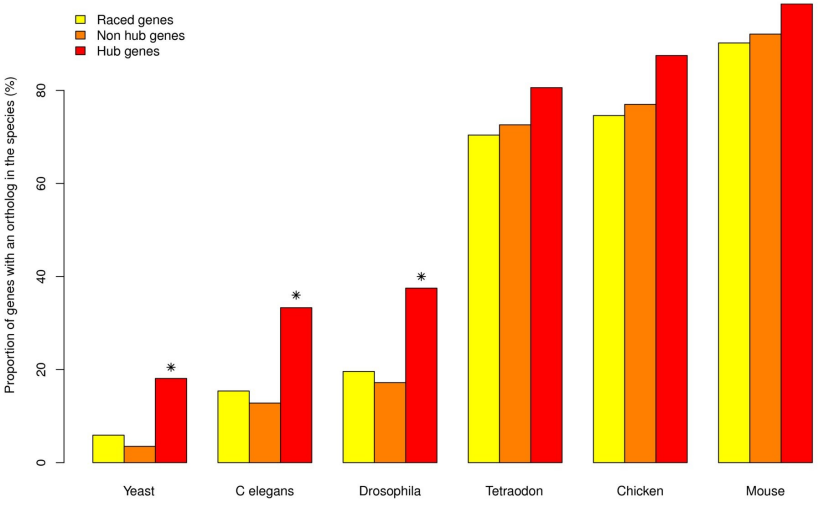
Sarah Djebali
Julien Lagarde
Vincent Lacroix
Sylvain Foissac

(B)



(A)

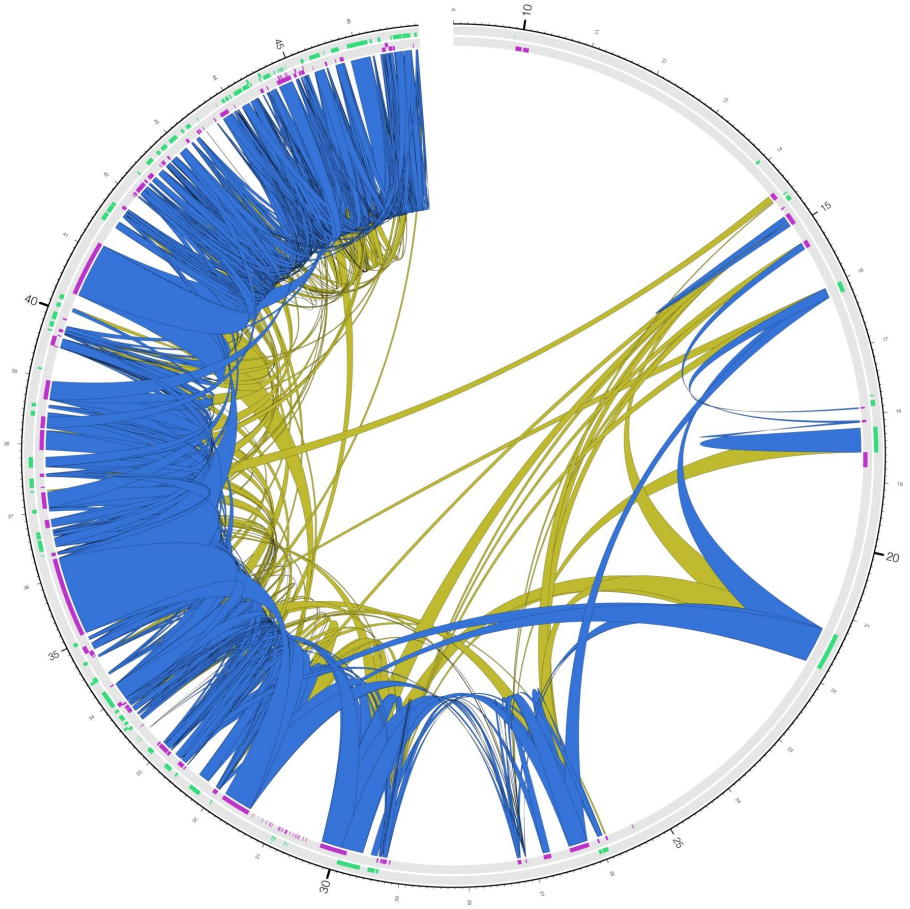
Hubs have higher phylogenetic depth than no hubs



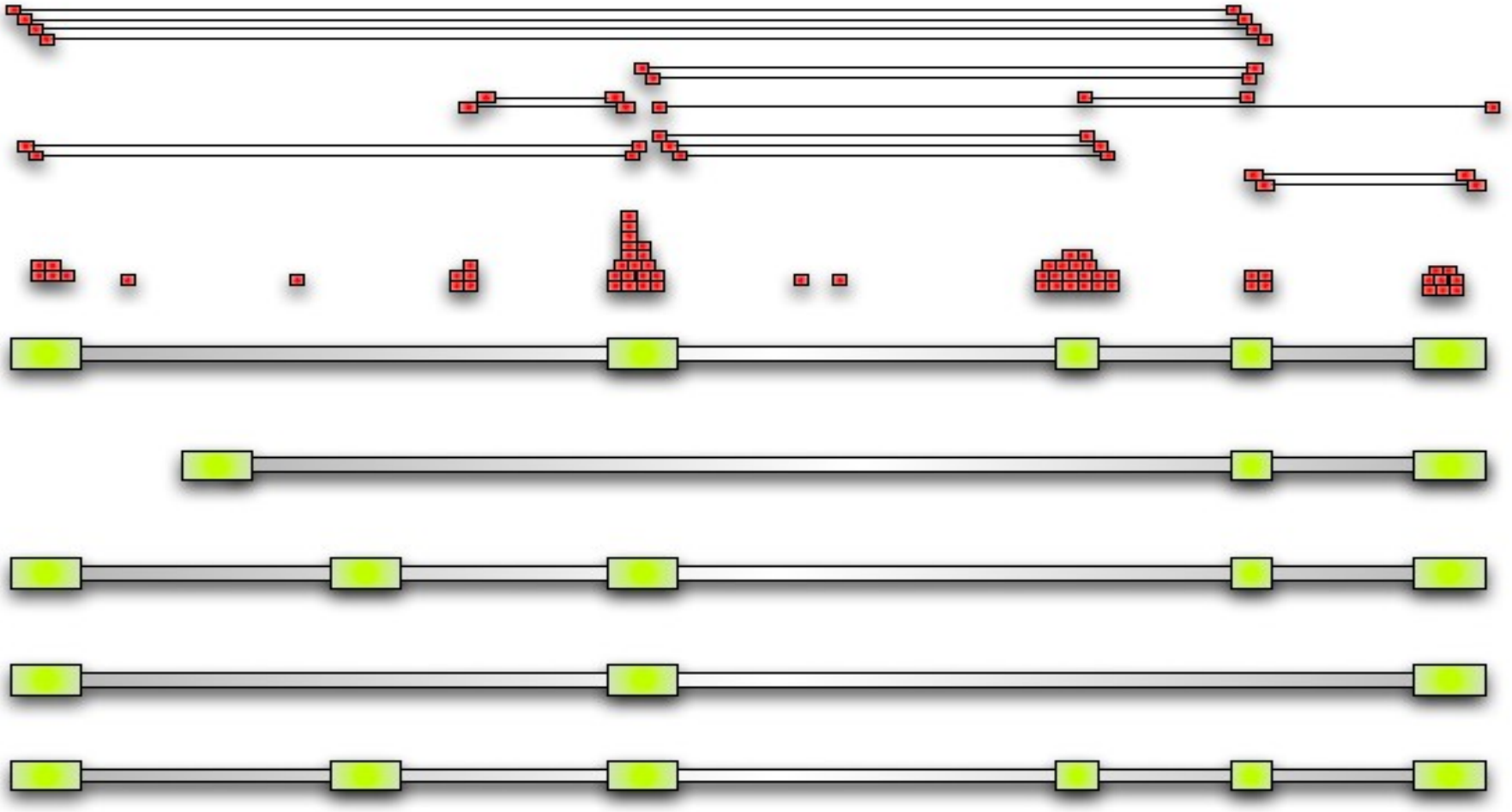
(C)

Distal sites of transcription associate to sites of chromosome interaction

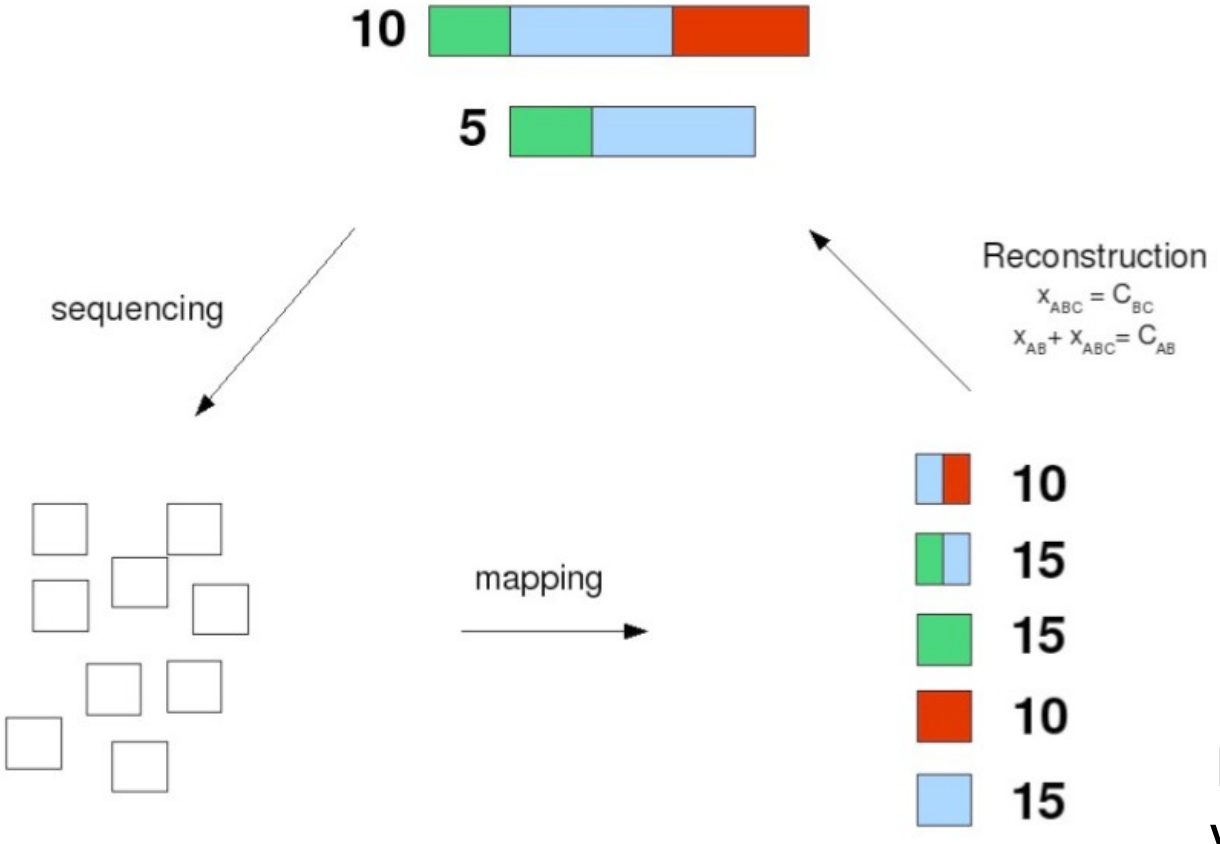
(Job Dekker, John Stamm)



Transcript quantification



Reconstructing transcript abundances from short sequence reads

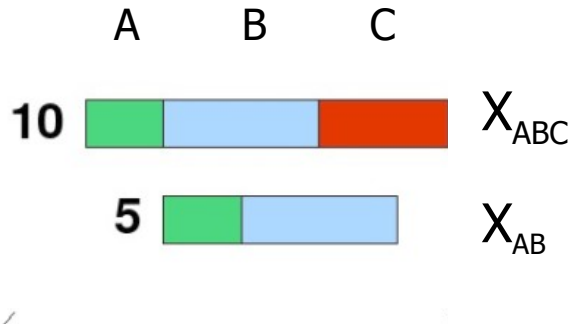


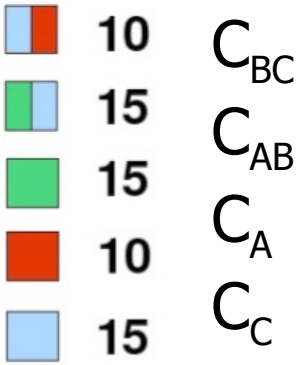
Micha Sammet

Vincent Lacroix

Paolo Ribeca

the problem can be posed as a system of linear equations

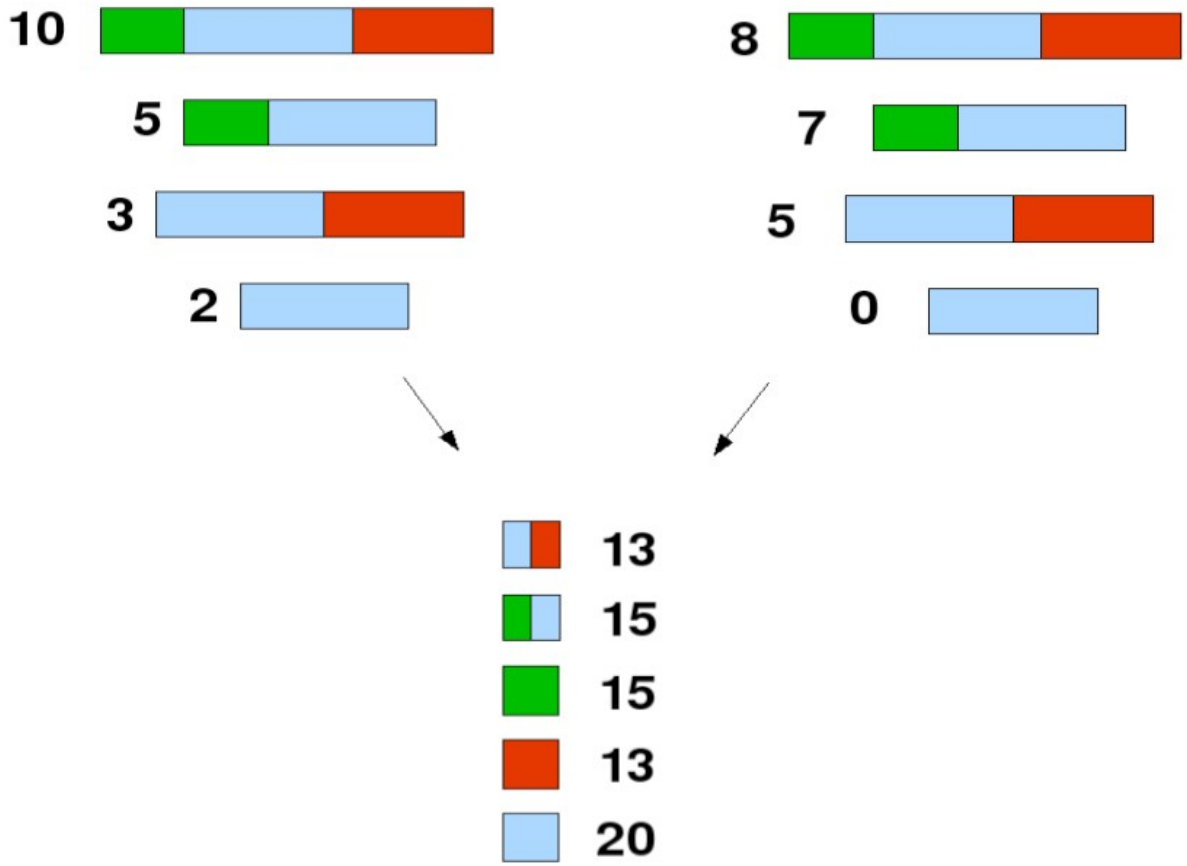




$$C_{BC} = X_{ABC}$$
$$C_{AB} = X_{ABC} + X_{AB}$$
$$C_A = X_{ABC} + X_{AB}$$
$$C_C = X_{ABC}$$
$$C_B = X_{ABC} + X_{AB}$$

V. Lacroix, M. Sammeth, R. Guigo, A. Bergeron. Exact transcriptome reconstruction from short sequence reads. WABI (2008)

The system of equations is often underdetermined: different transcriptomes result in exactly the same set of sequenced reads

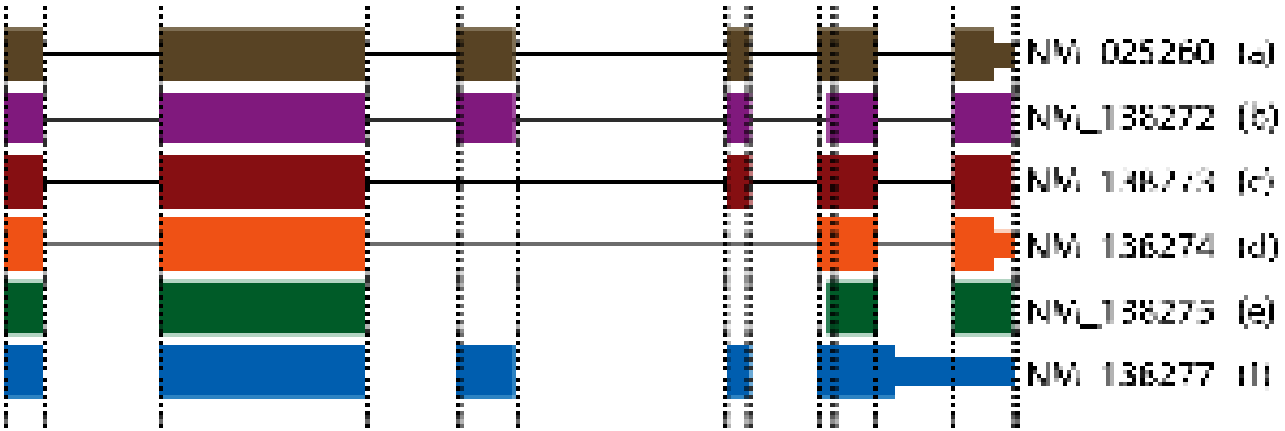


How bad is this in the practice?

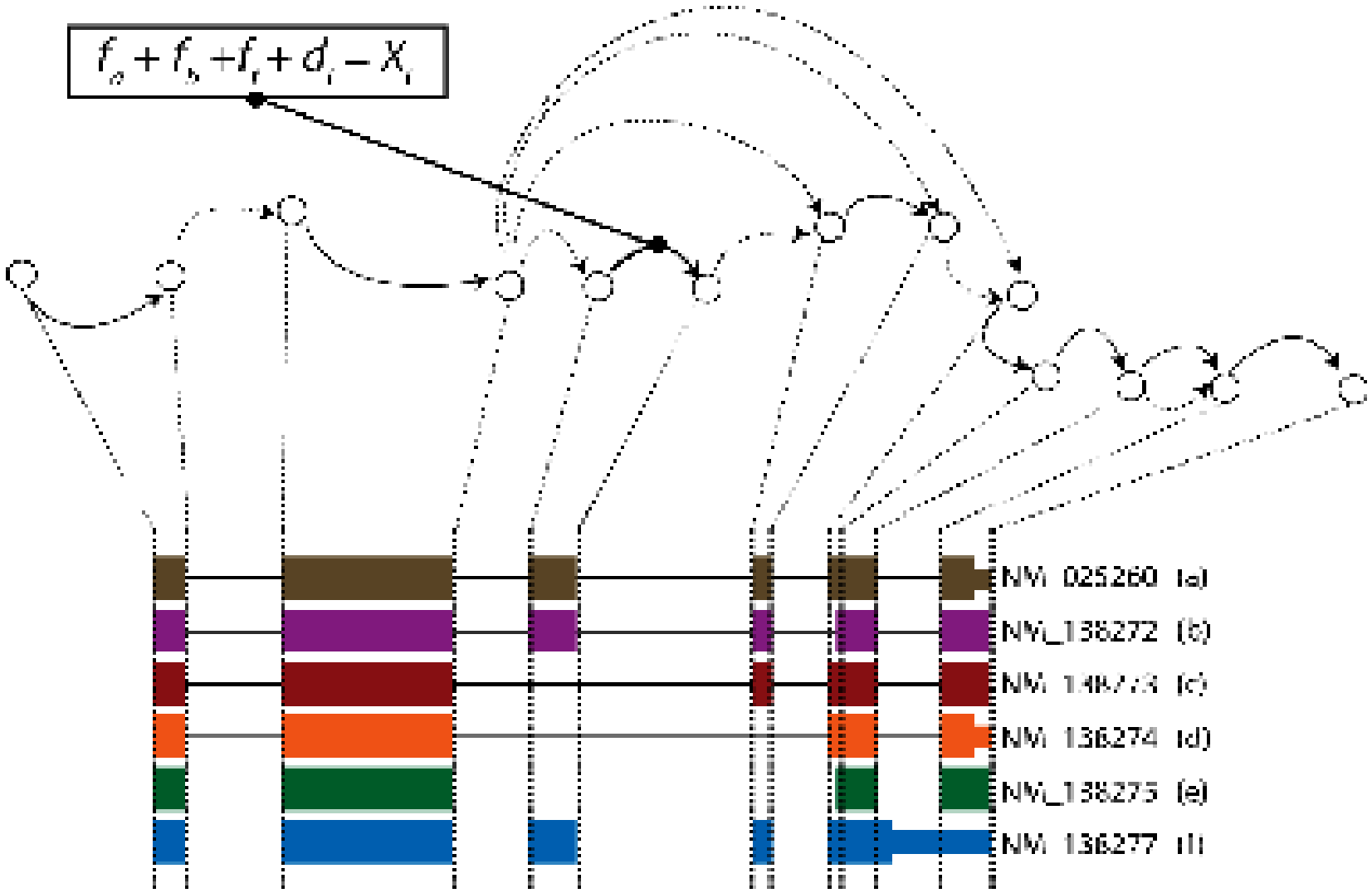
- GENCODE annotation of the ENCODE pilot regions (44 regions, 1% of the human genome)
  - 681 genes and **2981** variants
- Assume "unlimited" number of **single** reads equally distributed across the transcript length ("infinite" coverage)
- If we **do not assume** the underlying **transcriptome**--that is, all combinations of annotated exons are possible--we can only unequivocally determine the abundance of **30** variants.
- **Assuming that the underlying transcriptome**, however we can determine the abundances **in all cases**, but one.

The Flux Capacitor.

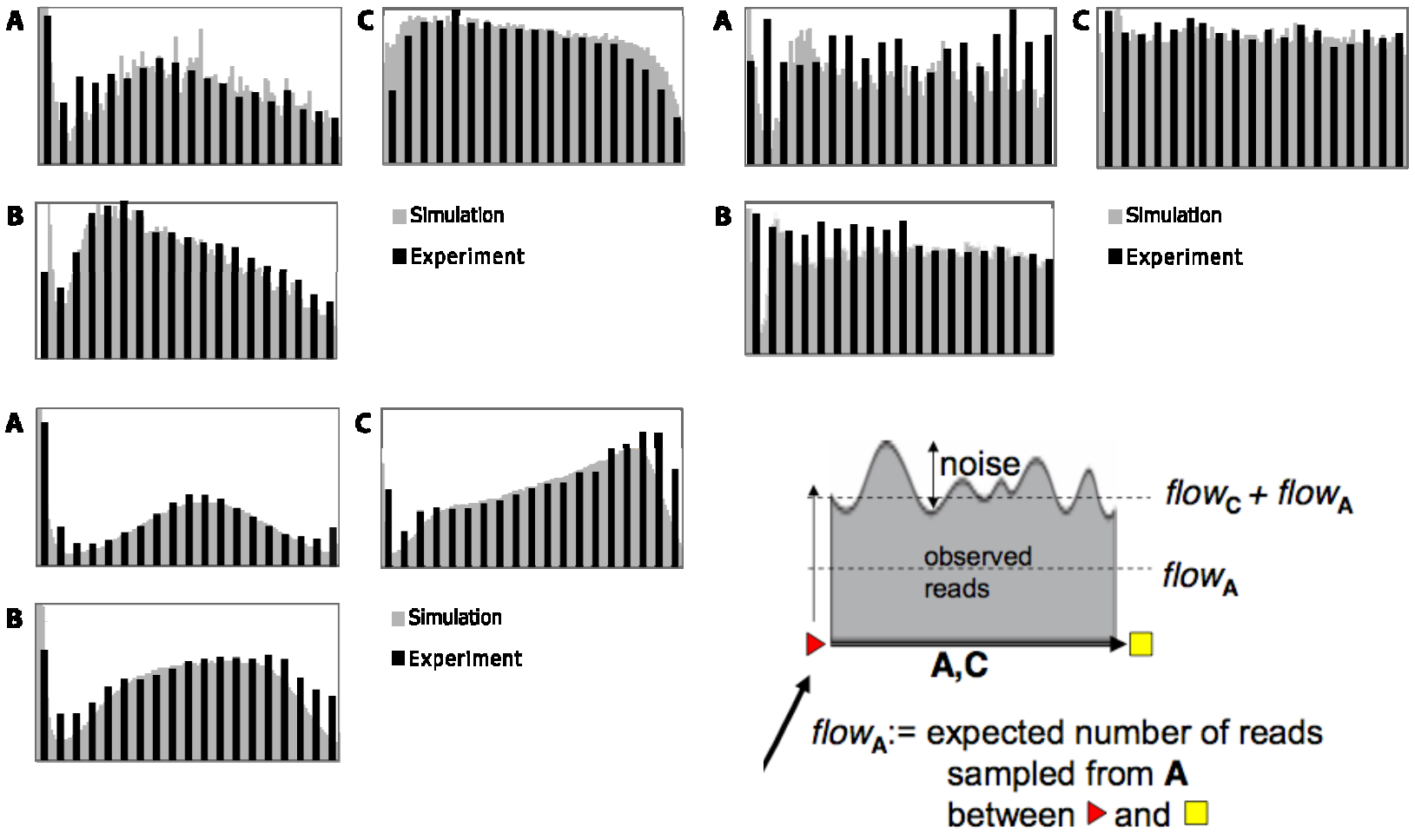
Micha Sammeth



The Flux Capacitor



Read density bias along transcripts



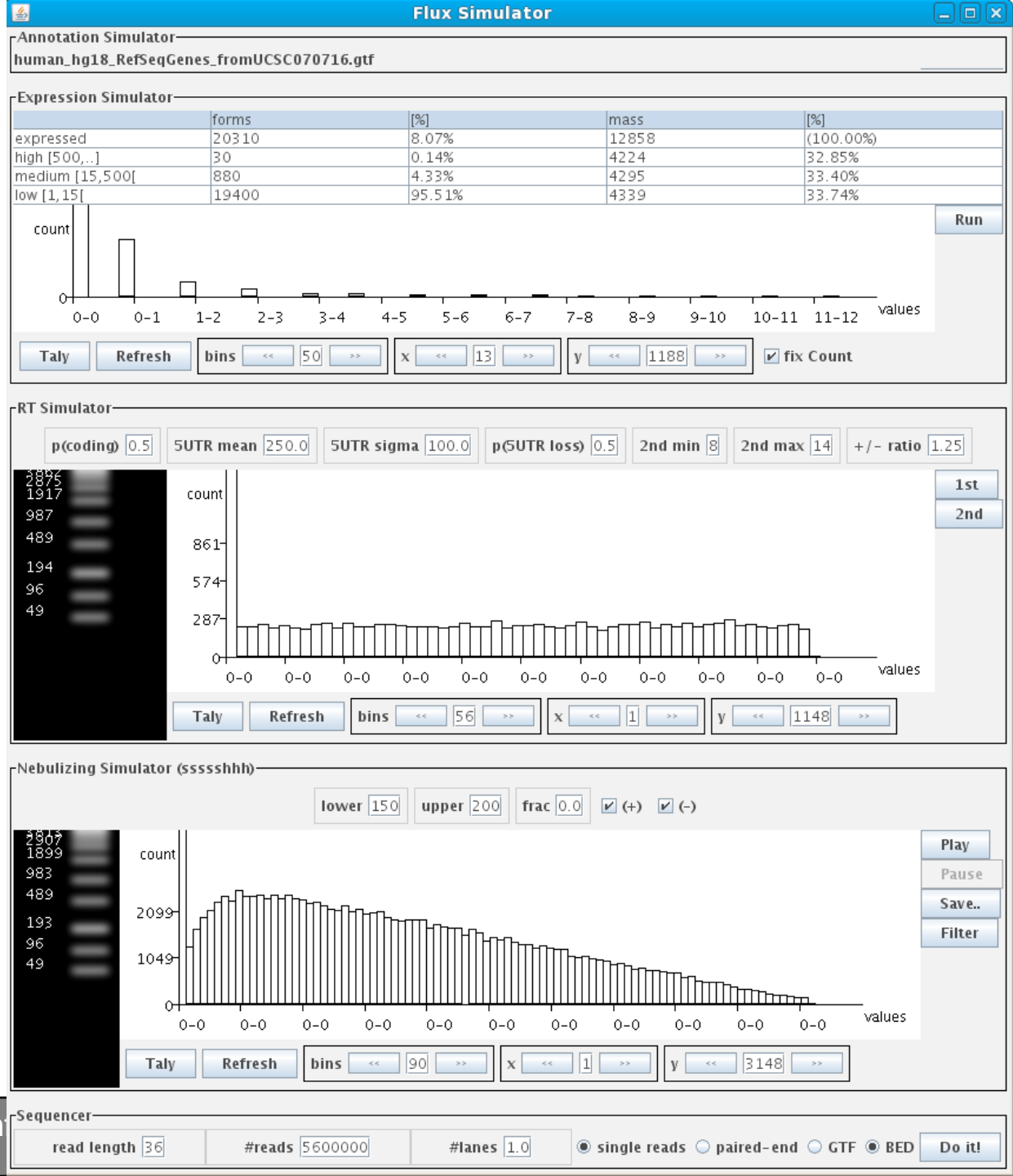
RNAseq simulator

First, given a transcript annotation, and a model of transcript abundance

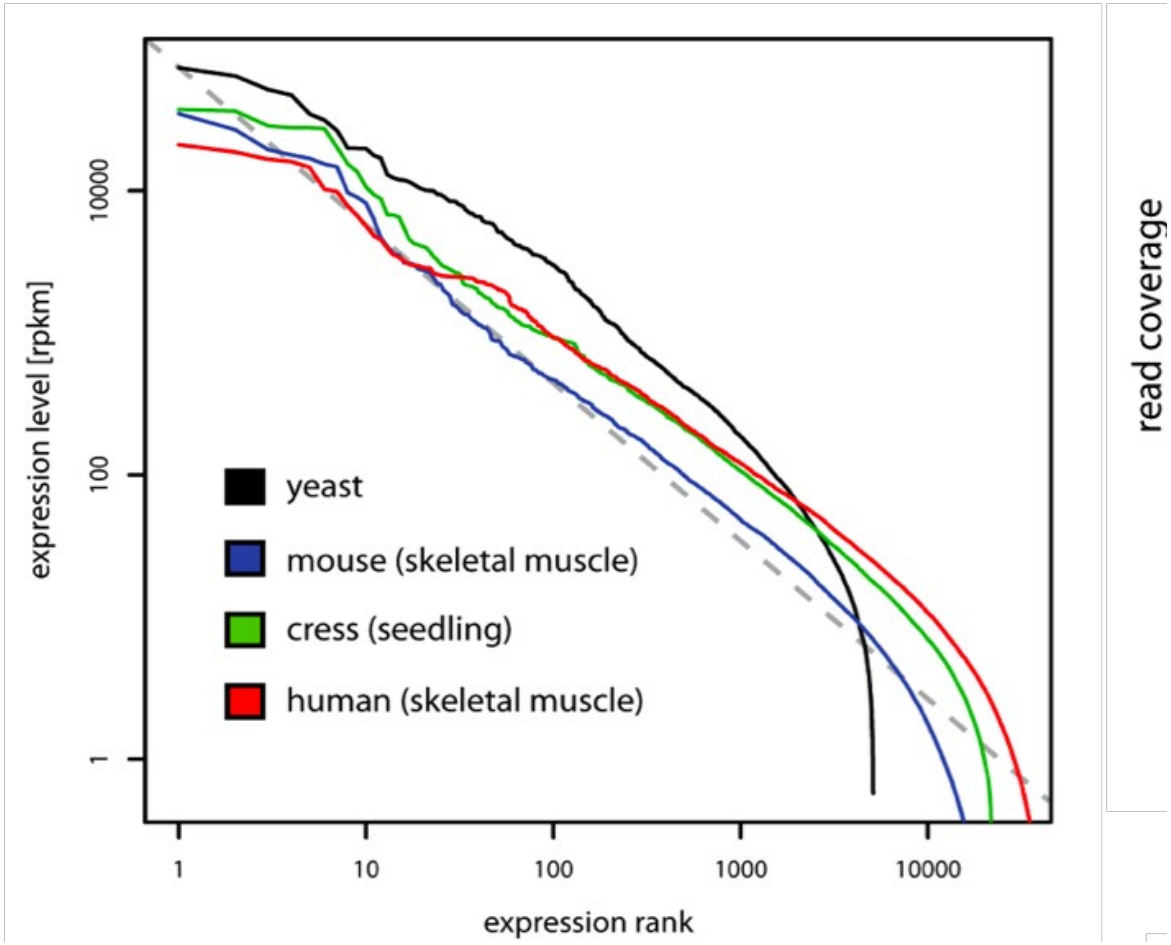
Then, the simulator produces a transcriptome (a set of transcripts with abundances attached)

Second, it simulates the different steps of the sequencing experiment (RT, nebulization, etc)

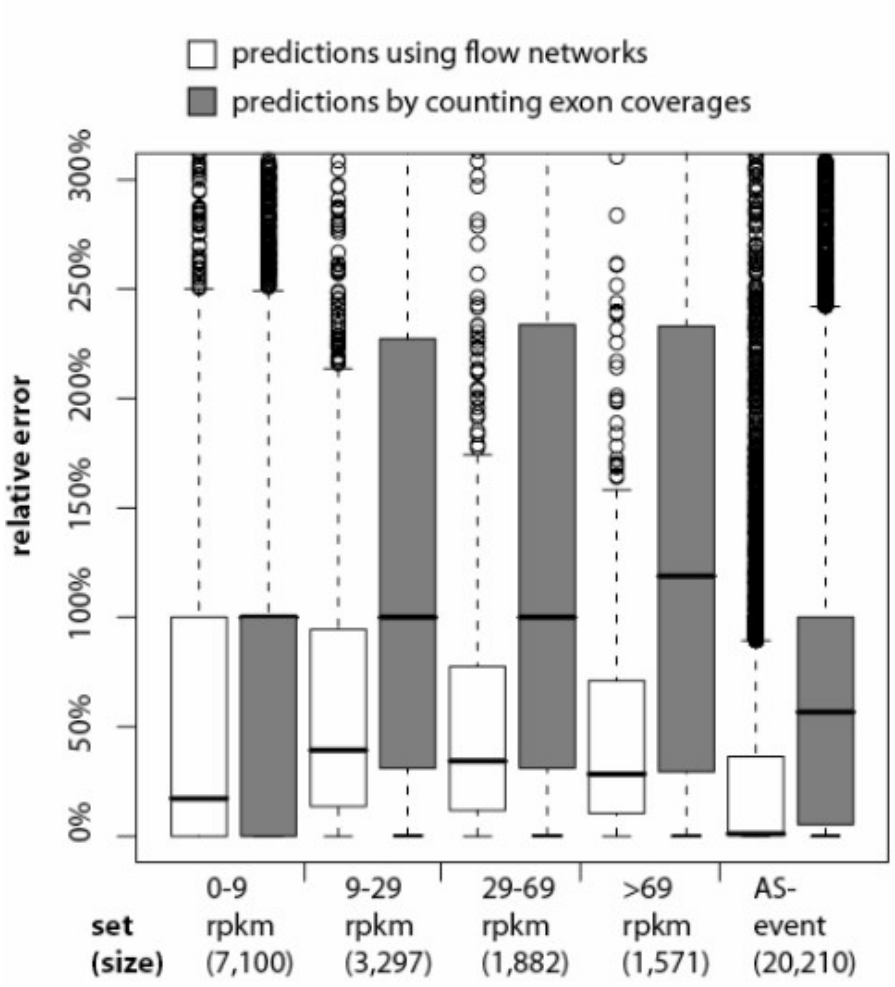
To **produce** a set of sequence reads as realistic as possible



RNAseq simulator; simulating the transcriptome



Quantitative assessment



relative error :=

$$\frac{|\, pred\_cov - ref\_cov \,|}{ref\_cov}$$

resp. 100% for false positives (i.e., *ref_cov* = 0).

pred_cov = predicted coverage

ref_cov = reference coverage

rpkm = reads per kilobase per million mapped reads (coverage measure)

Functional Genomics
Human (HGP)
Pathogens
Blast

ENCODE
Home
Data Management
Participants
News
RGASP
Pilot Phase

Website Search
People Search
Library Services
Site Map
Feedback / Help

Hide Navigation

RGASP - The RNAseq Genome Annotation Assessment Project

Introduction

Following the successful format of the EGASP workshop in 2005, the RNAseq genome annotation assessment project is being launched to assess the current progress of automatic gene building using RNAseq as its primary dataset.

The main aim of the workshop is to **assess the status of computational methods to map human RNAseq data, assemble them into transcripts and quantify the abundance of that transcript in particular datasets**.

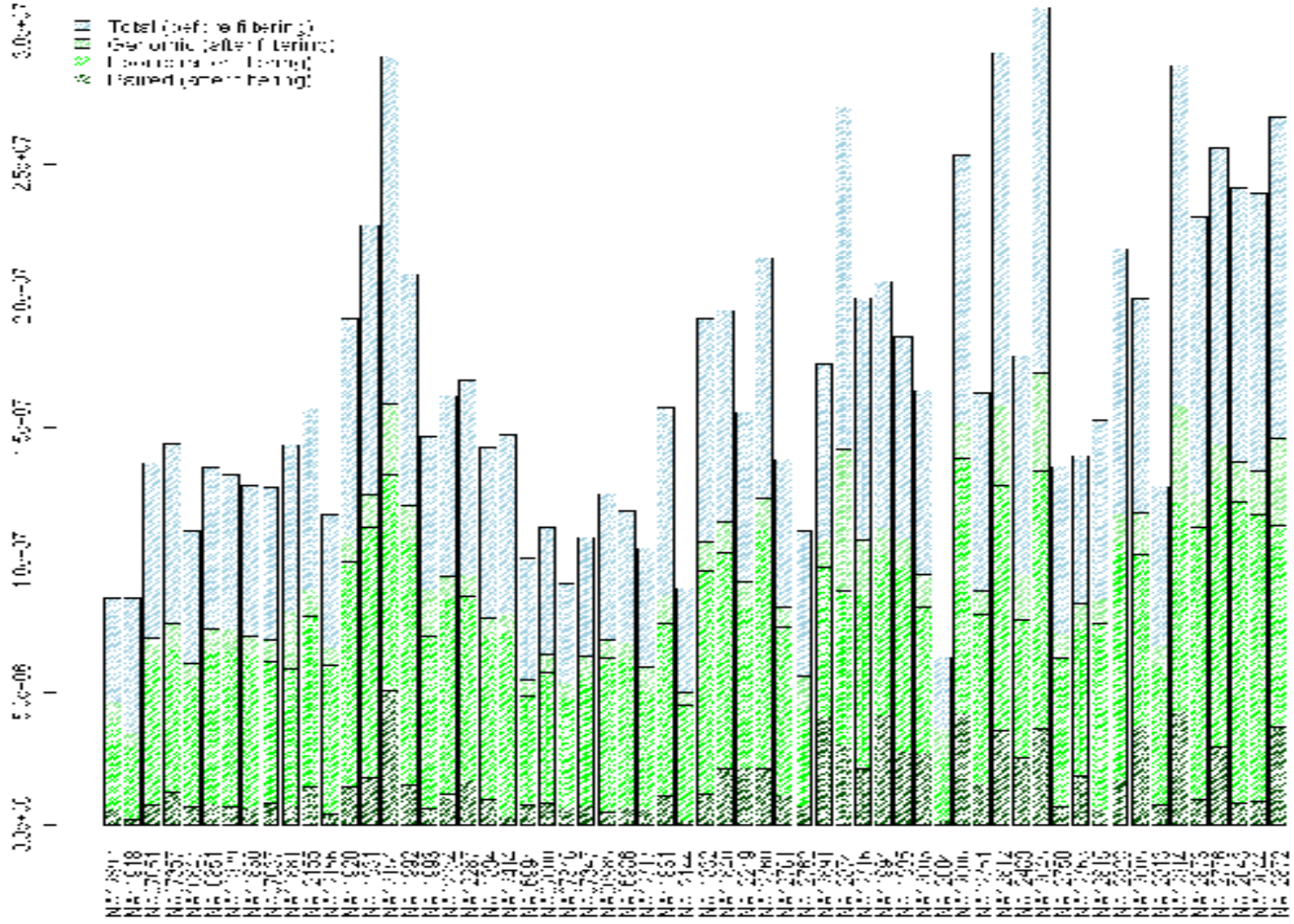
Transcript predictions will be evaluated against the GENCODE annotation produced as part of the ENCODE project. Special attention/assessment will be given to newly manually annotated chromosomes not previously publicly released before the workshop release. Promising transcript predictions not covered by Gencode annotation will be validated by experimental methods

The project is open to all researchers, but places are limited.
We will also include modEncode drosophila data and would like to encourage participants working on model organisms to participate.
More details will be posted on this website.

Dates

Registration deadline:	03.07.2009
Submission of predictions:	21.09.2009 (Extended)
Workshop days:	10. & 11.11.2009

Sequence of the transcriptome of 60 CEU individuals for the HapMap project.

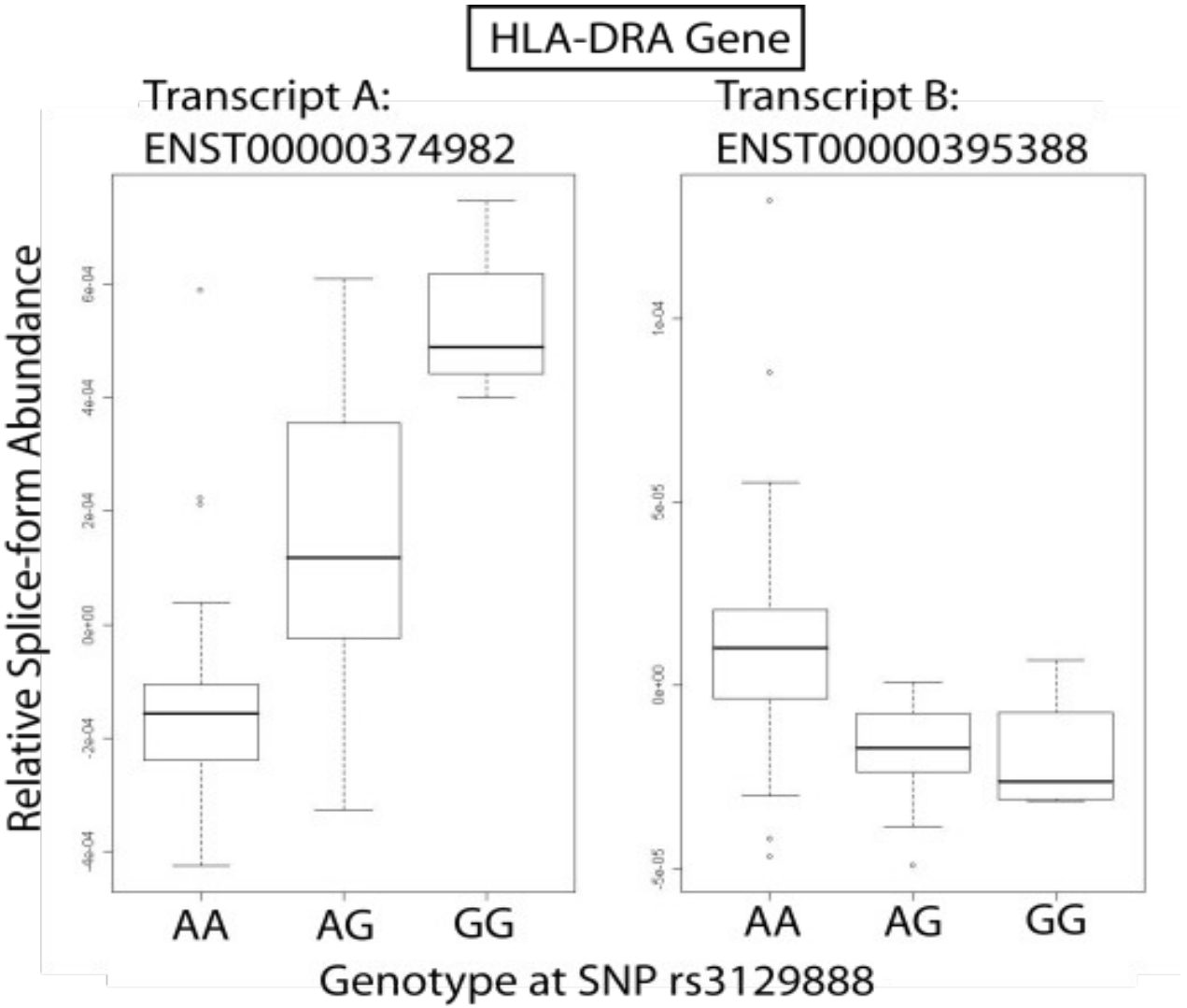


Transcriptome genetics using second generation sequencing in a Caucasian population

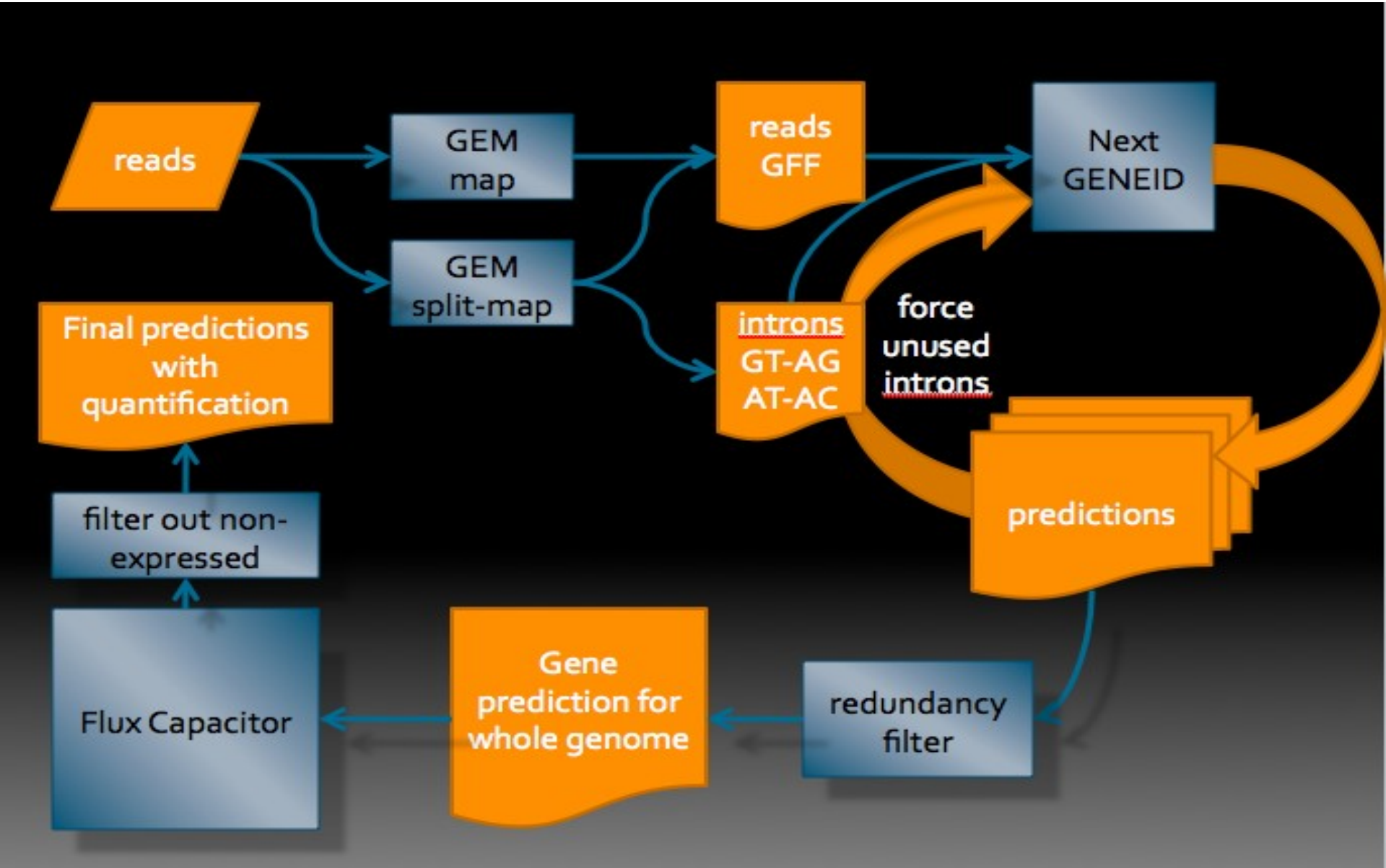
Stephen B. Montgomery[1,2*], Micha Sammeth[3], Maria Gutierrez-Arcelus[1], Radoslaw P. Lach[2], Catherine Ingle[2], James Nisbett[2], Roderic Guigo[3] & Emmanouil T. Dermitzakis[1,2*]

- ~10 million reads provide access to the same dynamic range as arrays with better quantification of alternative and highly abundant transcripts.
- Correlation with SNPs leads to a larger discovery of eQTLs than with arrays.
  - eQTLs are discovered with similar frequencies in protein coding and long non coding RNAs
- a substantial number of variants influence the structure of mature transcripts indicating variants responsible for alternative splicing.
- Measures of allele- specific expression allow for the identification of rare eQTLs and allelic differences in transcript structure

Significant eQTL with opposite effect in two different isoforms of the gen HLA-DRA



NextGeneid: transcript assembly and discovery



http://gemlibrary.sourceforge.net

navigation

- The GEM library
- News
- Browse code
- Downloads
- Documentation
- Publications
- Bug reports
- Suggestions
- Forums
- Mailing lists
- Recent changes

search

Go Search

toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link

The GEM library

(Also home to: *The GEM mapper*, *The GEM split-mapper*, and others)

Next-generation sequencing platforms (Illumina/Solexa, ABI/SOLiD, etc.) call for powerful and very optimized tools to index/analyze huge genomes. The GEM library strives to be a true "next-generation" tool for handling any kind of sequence data, offering state-of-the-art algorithms and data structures specifically tailored to this demanding task. At the moment, efficient indexing and searching algorithms based on the Burrows-Wheeler transform (BWT) have been implemented. The library core is written in C for maximum speed, with concise interfaces to higher-level programming languages like OCaml and Python. Many high-performance standalone programs (mapper, split-mapper, etc.) are also provided along with the library; in general, new algorithms and tools can be easily implemented on the top of it.

The GEM project started at the CRG in June 2008. Since fall 2008, early versions of the GEM tools have been in everyday use to help the development of many different scientific projects involving mapping of DNA/RNA data, reconstruction of RNA splice-form abundances, SNP calling, microRNA analysis, ChIP-seq experiments, metagenomics studies, and other tasks related to next-generation sequencing.

If you are happy with GEM, you might also like its friend projects:

- MIRO, a pipeline to analyze microRNAs using next-generation sequencing data
- The Flux Capacitor, a set of tools to predict the abundance of splice-forms from next-generation sequencing data.

All these projects provide full integration for the `gem-mapper` to be used as the engine of their mapping stage.

News

24/11/2009 Added `gem-mappability` (experimental)

10/09/2009 Started optimization round. Stay tuned!

20/07/2009 Added `gem-do-index` man page

14/07/2009 Added `gem-mapper` man page

11/07/2009 Added `gem-2-sam` converter (experimental)

30/06/2009 Added Perl converters

23/06/2009 First official binary pre-release

Paolo Ribeca

http://flux.sammeth.net/

Micha Sammet