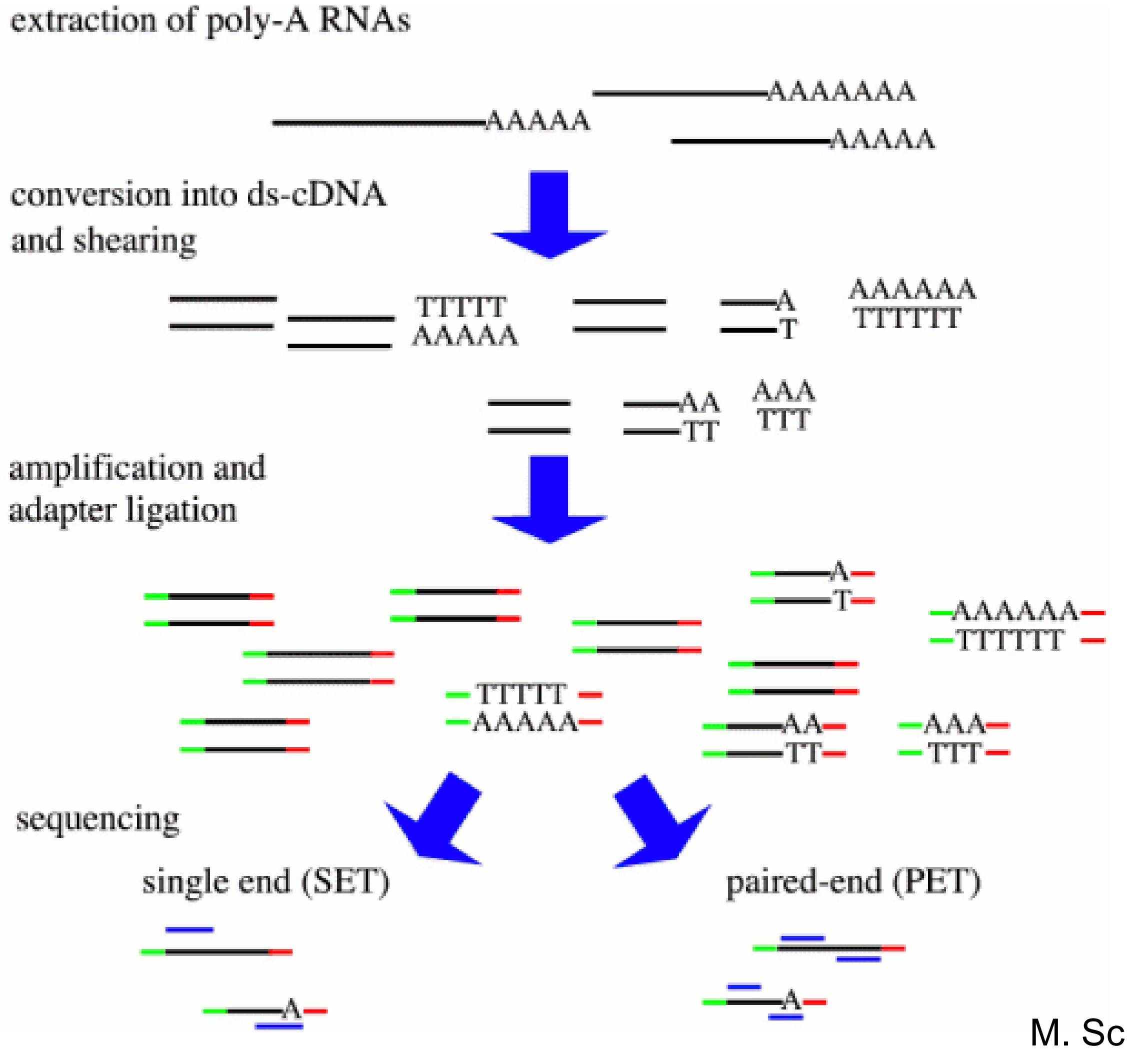
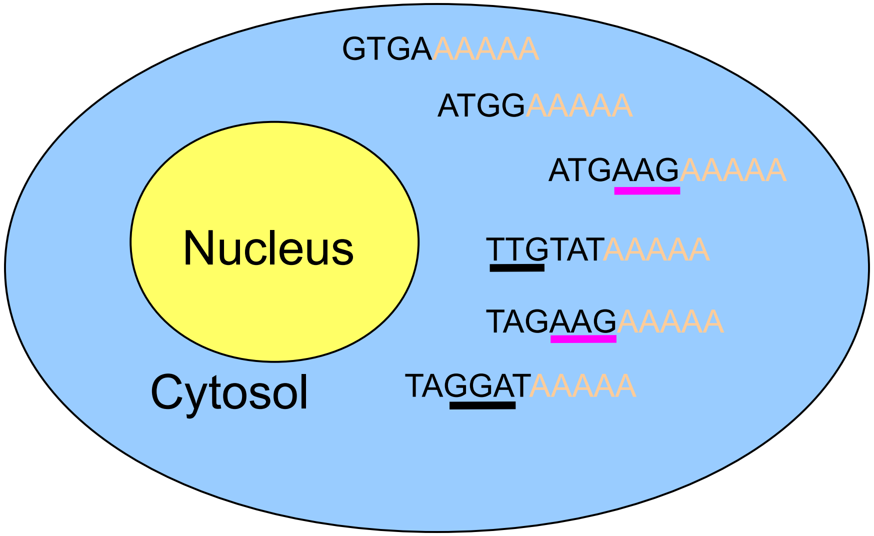
Methods for Analysis of RNA-Seq data.

Hugues Richard

RNA-Seq protocol (no strand information)



M. Schulz©



“bag of transcript positions”

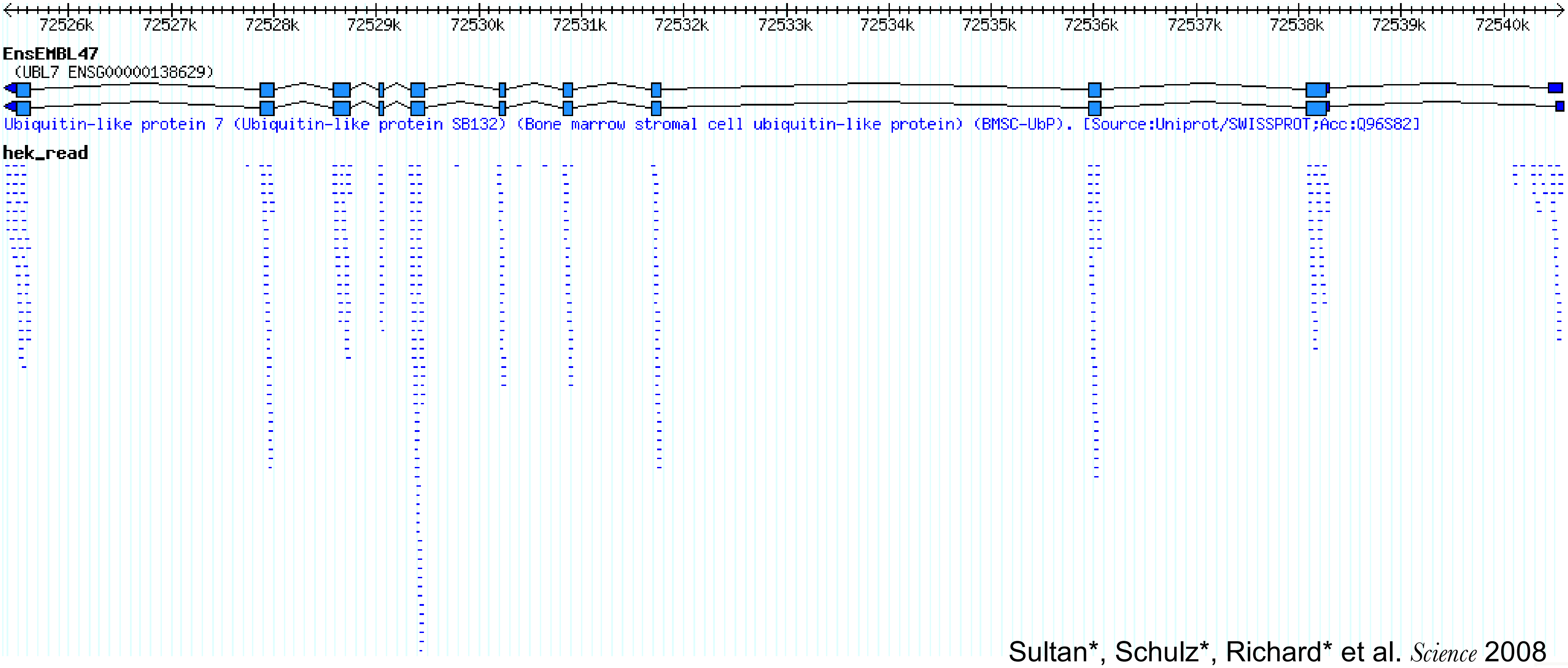
TTG
GGA
AAG

Variability of the counts (sampling) influenced by:

_ region length

_ copy number

Mapping the reads



72526k
72527k
72528k
72529k
72530k
72531k
72532k
72533k
72534k
72535k
72536k
72537k
72538k
72539k
72540k
EnsEMBL47
(UBL7 ENSG00000138629)
Ubiquitin-like protein 7 (Ubiquitin-like protein SB132) (Bone marrow stromal cell ubiquitin-like protein) (BMSC-UbP). [Source:Uniprot/SWISSPROT;Acc:Q96S82]
hek_read

Aggregate counts on exons/genes

Normalize by number of possible hits (RPKM)

Sultan*, Schulz*, Richard* et al. *Science* 2008

Mortazavi*, Williams* et al. *Nat Methods* 2008

Wang*, Sandberg* et al. *Nature* 2008

Cloonan*, Forest*, Kolle* et al. *Nat Methods* 2008

MAX-PLANCK-GESELLSCHAFT

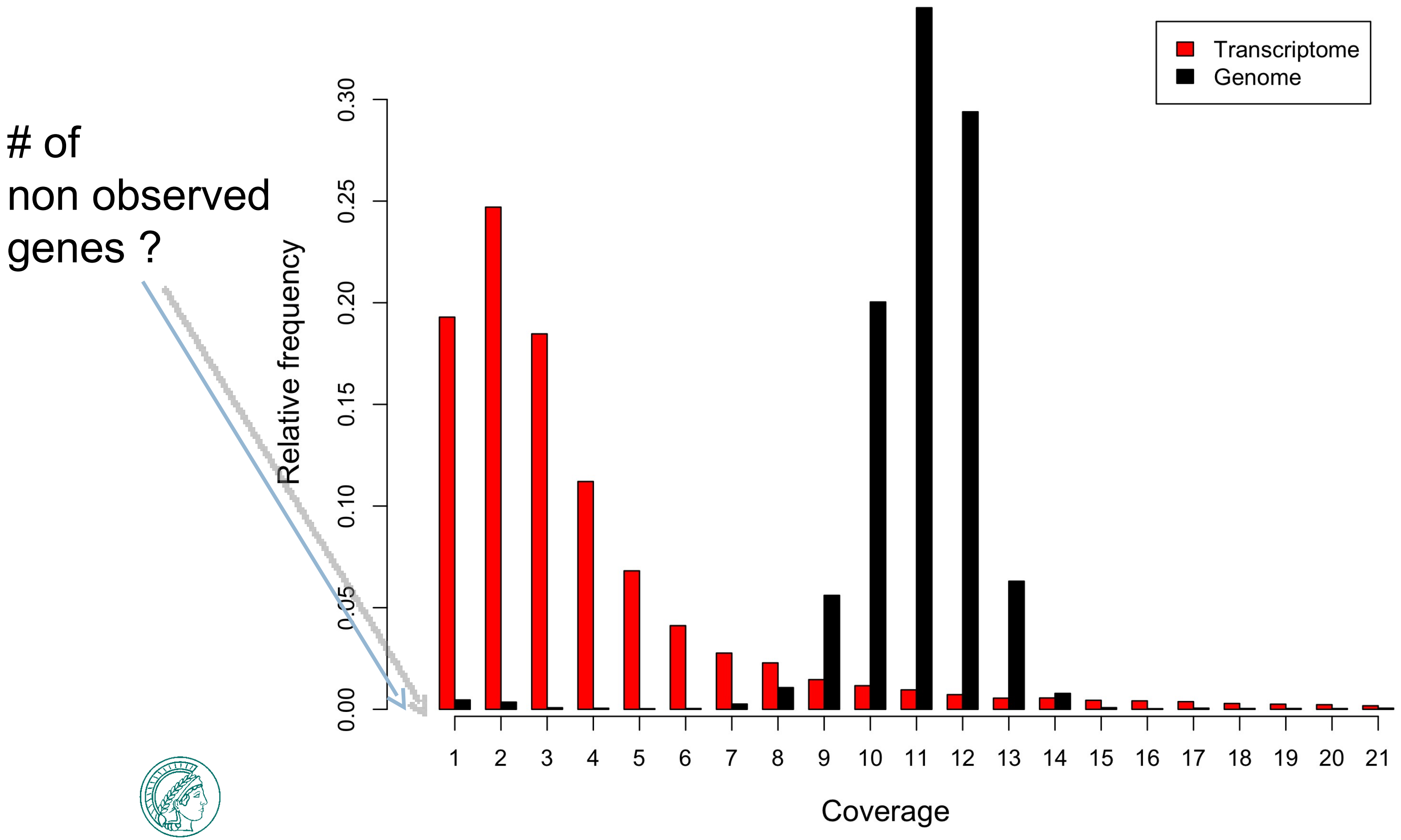
UPMC PARIS UNIVERSITAS

Outline

- Estimating depth of sequencing.
  - Do we see all the transcriptional units ?
- Infering new events:
  - Detection of new transcriptional units.
  - Detection of Alternative Splicing Events.
- RGASP competition:
  - Transcriptome assembly with Oases (Schulz/Zerbino)
  - Reads remapping with RazerS (Weese)

Transcriptome vs genome assembly

- RNA-Seq reads are distributed according to transcript expression levels.



What is current coverage ? (Fisher & Corbett 43)

- Total number of transcripts :
  - Count of a transcriptional unit i:

$$X_i \sim \mathcal{P}(\lambda_i)$$

  - Unspecified counts:

$$f(c;\mathbf{g}) = \int_0^\infty \exp^{-\lambda} \frac{\lambda^c}{c!} \, \mathrm{d}\mathbf{g}(\lambda)$$

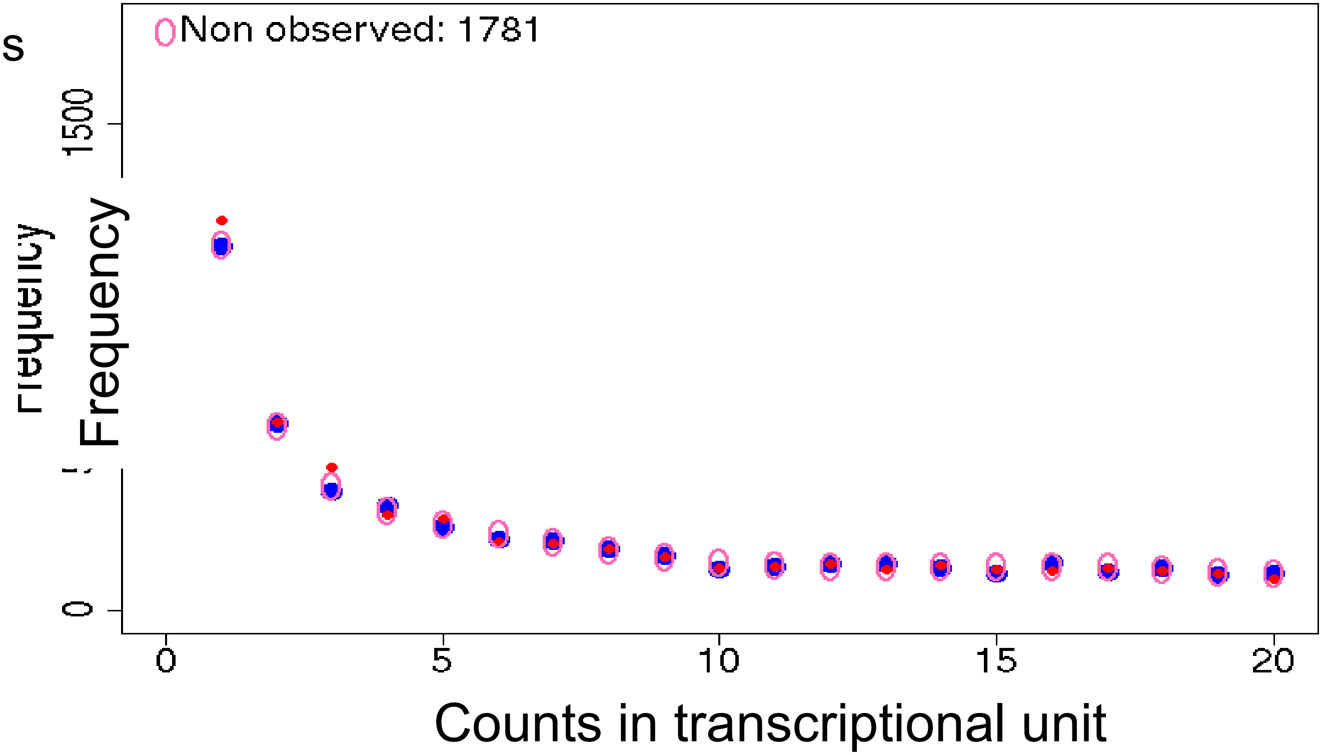
  - If we estimate $\mathfrak{g}$, then:

$$\hat{N} = \left\langle \frac{\#\{\text{observed}\}}{1 - f(0,\hat{g})} \right\rangle$$

- Expected number of new discoveries with more lanes

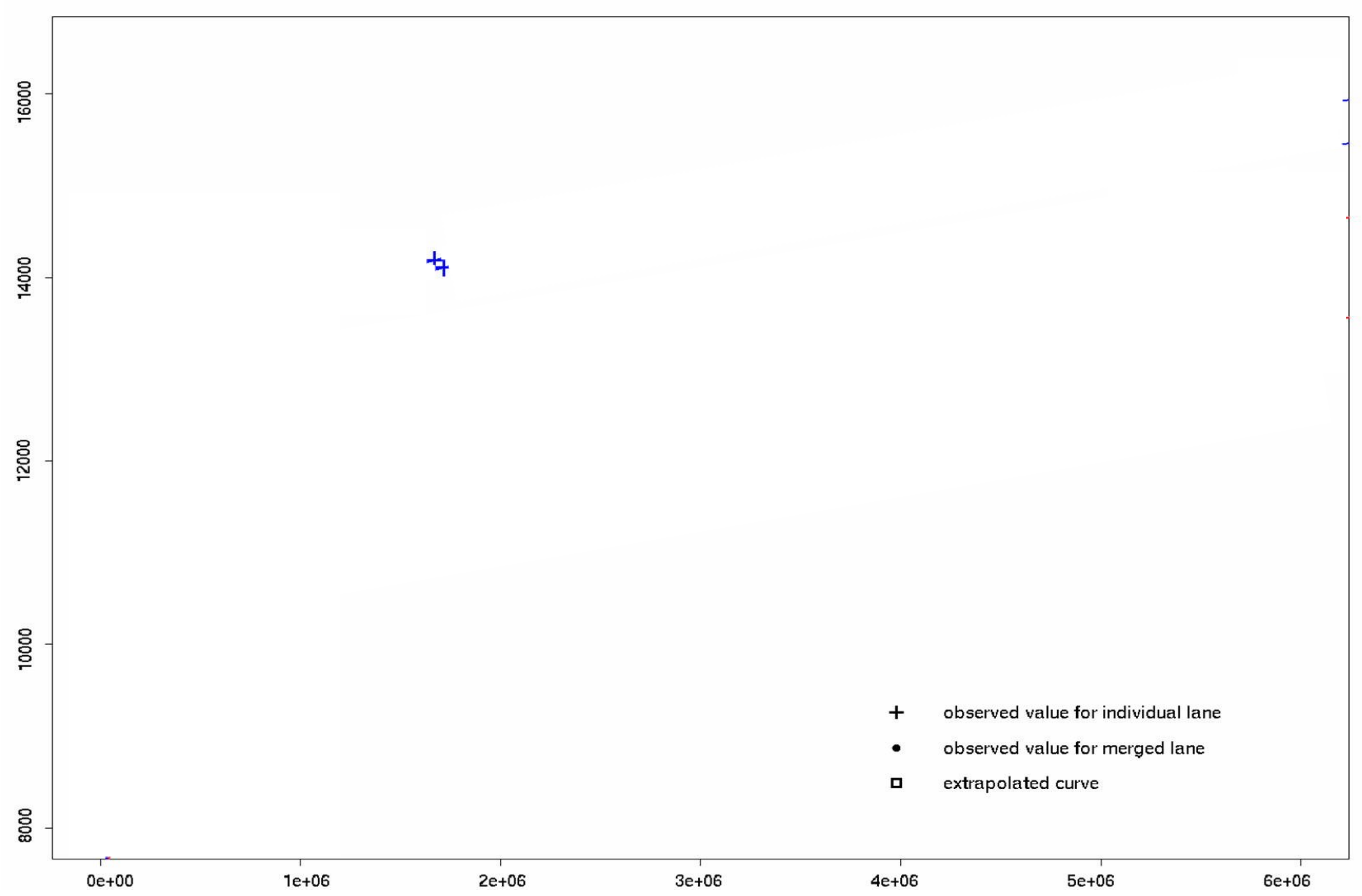
Estimating count frequency law

2 lanes



Penalized Non Parametric Maximum Likelihood method (Wang & Lindsay 05)

Dynamic range



Number of transcriptional units

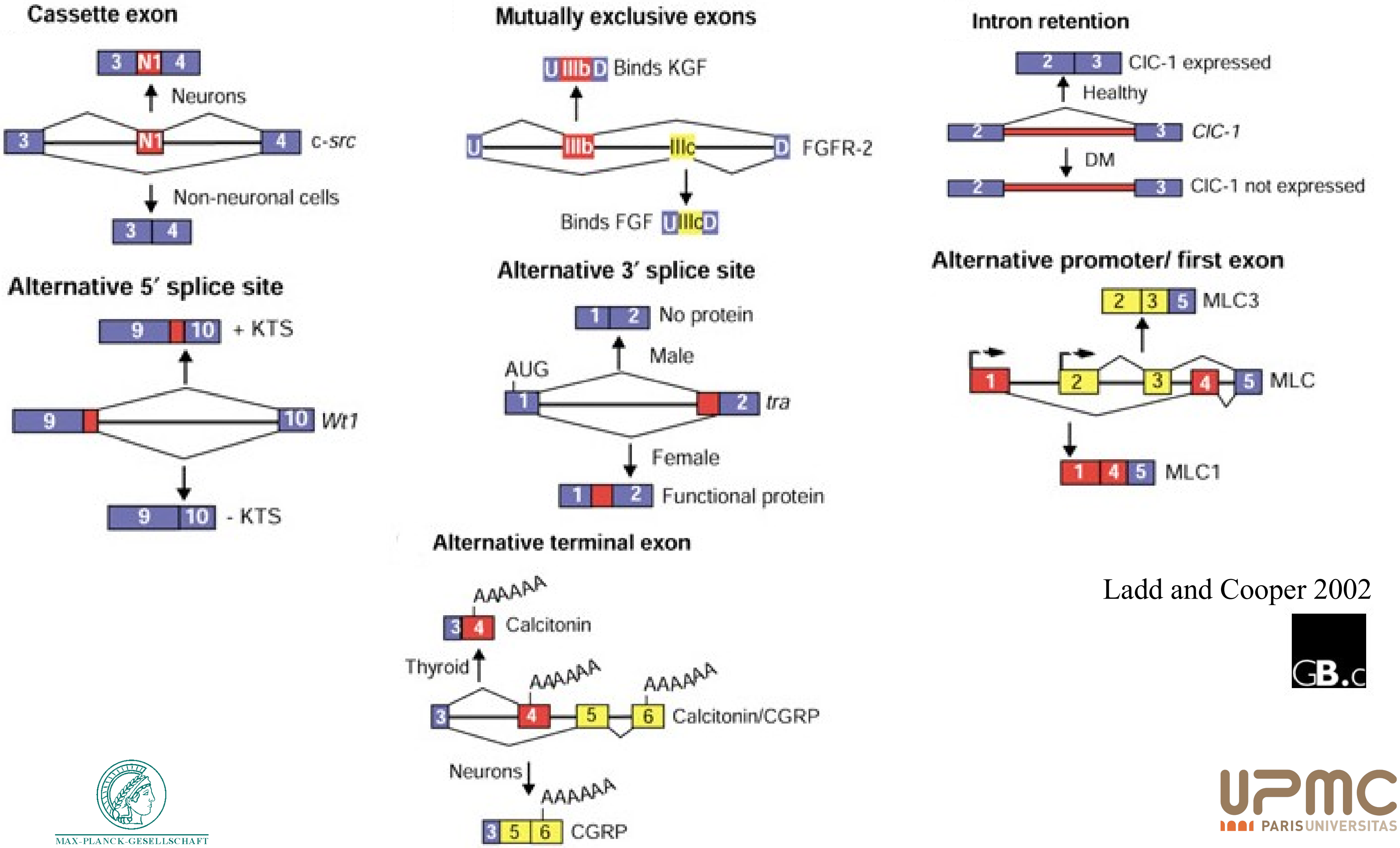
Number of mapped reads

~ 4 Mio reads/replicate

Outline

- Estimating depth of sequencing.
  - do we see all the genes ?
- Infering new events:
  - Detection of new transcriptional units.
  - Alternative Splicing events detection.
- RGASP competition:
  - Transcriptome assembly with Oases (Schulz/Zerbino)
  - Mapping with RazerS (Weese)

Alternative Exon Events (AEEs)



Cassette exon

Alternative 5′ splice site

Mutually exclusive exons

Alternative 3′ splice site

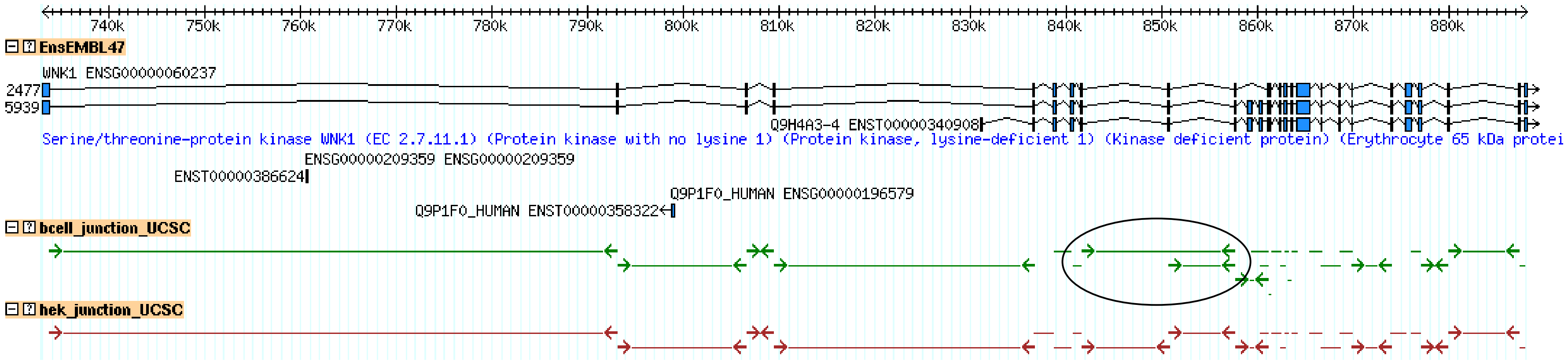
Alternative terminal exon

Intron retention

Alternative promoter/ first exon

Ladd and Cooper 2002

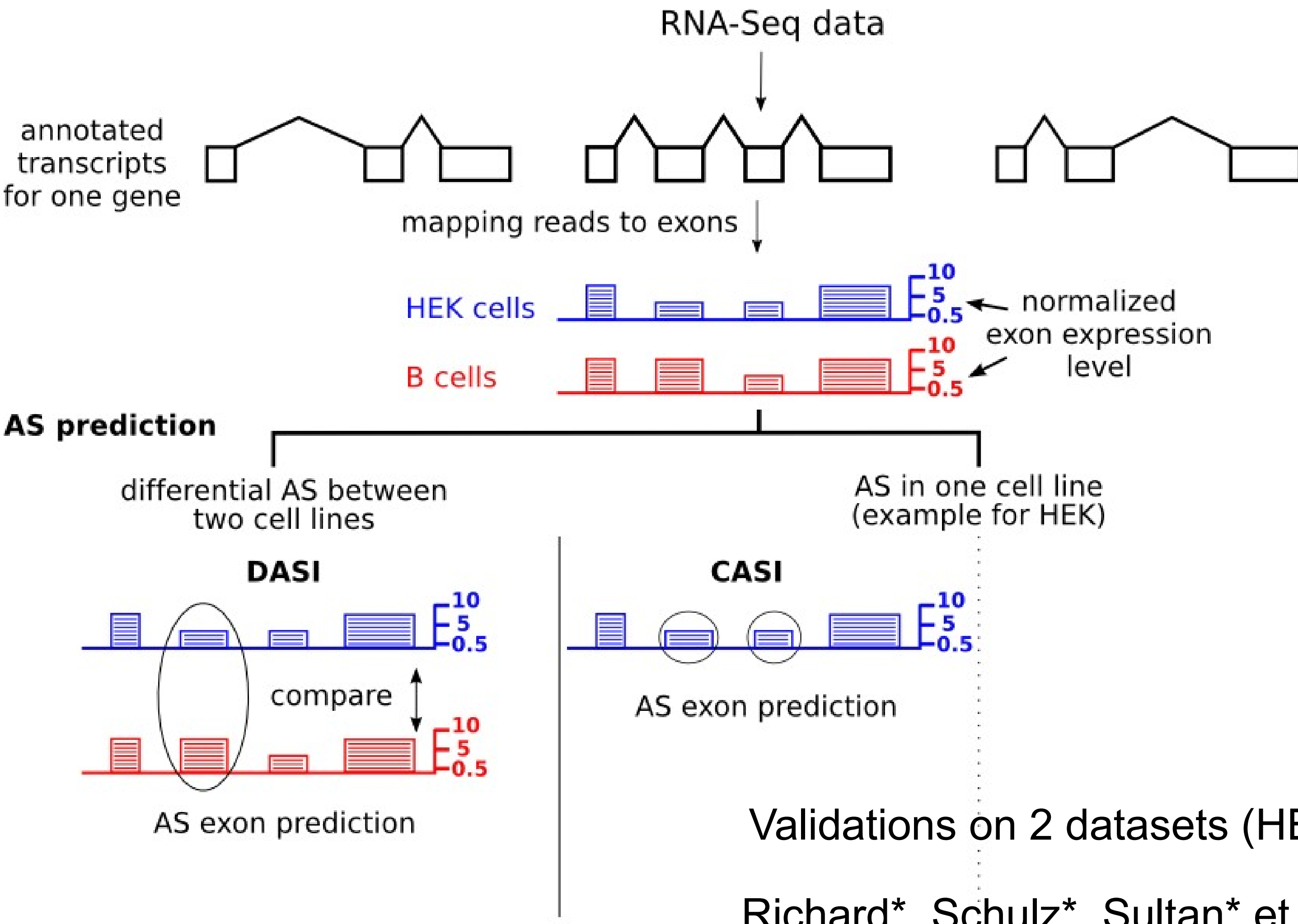
Splice junctions


740k 750k 760k 770k 780k 790k 800k 810k 820k 830k 840k 850k 860k 870k 880k
EnsEMBL47
WNK1 ENSG00000060237
2477
5939
Q9H4A3-4 ENST00000340908
Serine/threonine-protein kinase WNK1 (EC 2.7.11.1) (Protein kinase with no lysine 1) (Protein kinase, lysine-deficient 1) (Kinase deficient protein) (Erythrocyte 65 kDa protei
ENSG00000209359 ENSG00000209359
ENST00000386624
Q9P1F0_HUMAN ENSG00000196579
Q9P1F0_HUMAN ENST00000358322
bcell_junction_UCSC
hek_junction_UCSC

Align unmatched reads to a set of artificially created junctions

Use a splice junction aligner (Tophat, QPalma, GEMM)

Detecting AEE from Exons Expression

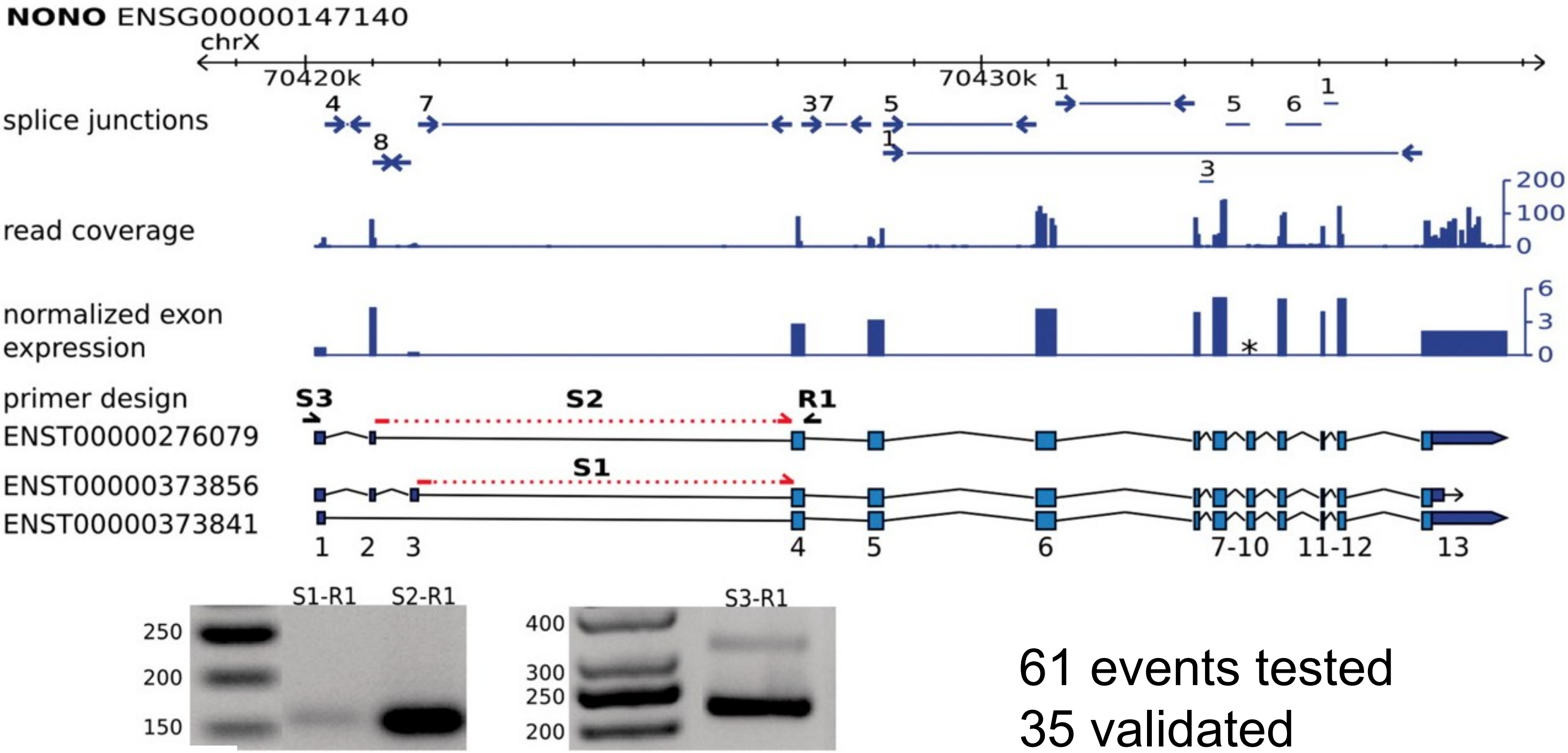

RNA-Seq data
annotated transcripts for one gene
mapping reads to exons
HEK cells
B cells
10
5
0.5
normalized exon expression level
AS prediction
differential AS between two cell lines
AS in one cell line (example for HEK)
DASI
compare
AS exon prediction
CASI
AS exon prediction

Solas

Validations on 2 datasets (HEK and B cells)

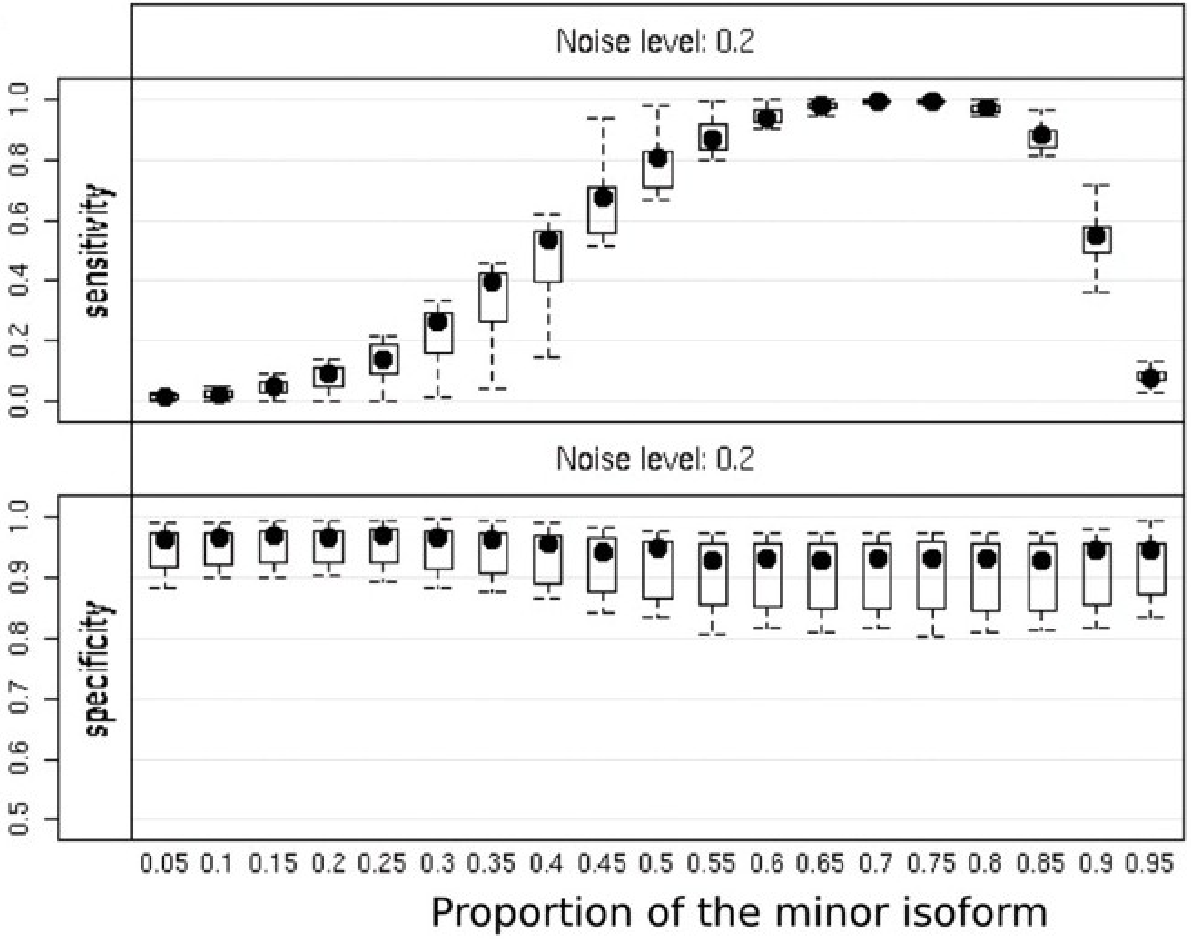
Richard*, Schulz*, Sultan* et al. *NAR* 2010

Within cell AS : CASI



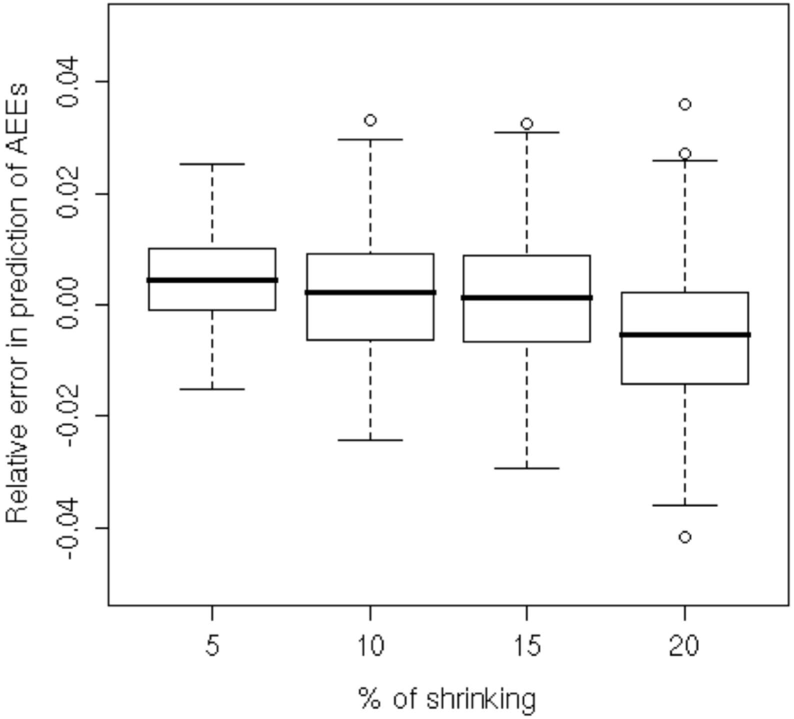
61 events tested
35 validated

Robustness (simulation)



1000 simulations of alternative exon events
of one gene with 6 exons and 300 reads in total

Robustness (bootstrap)

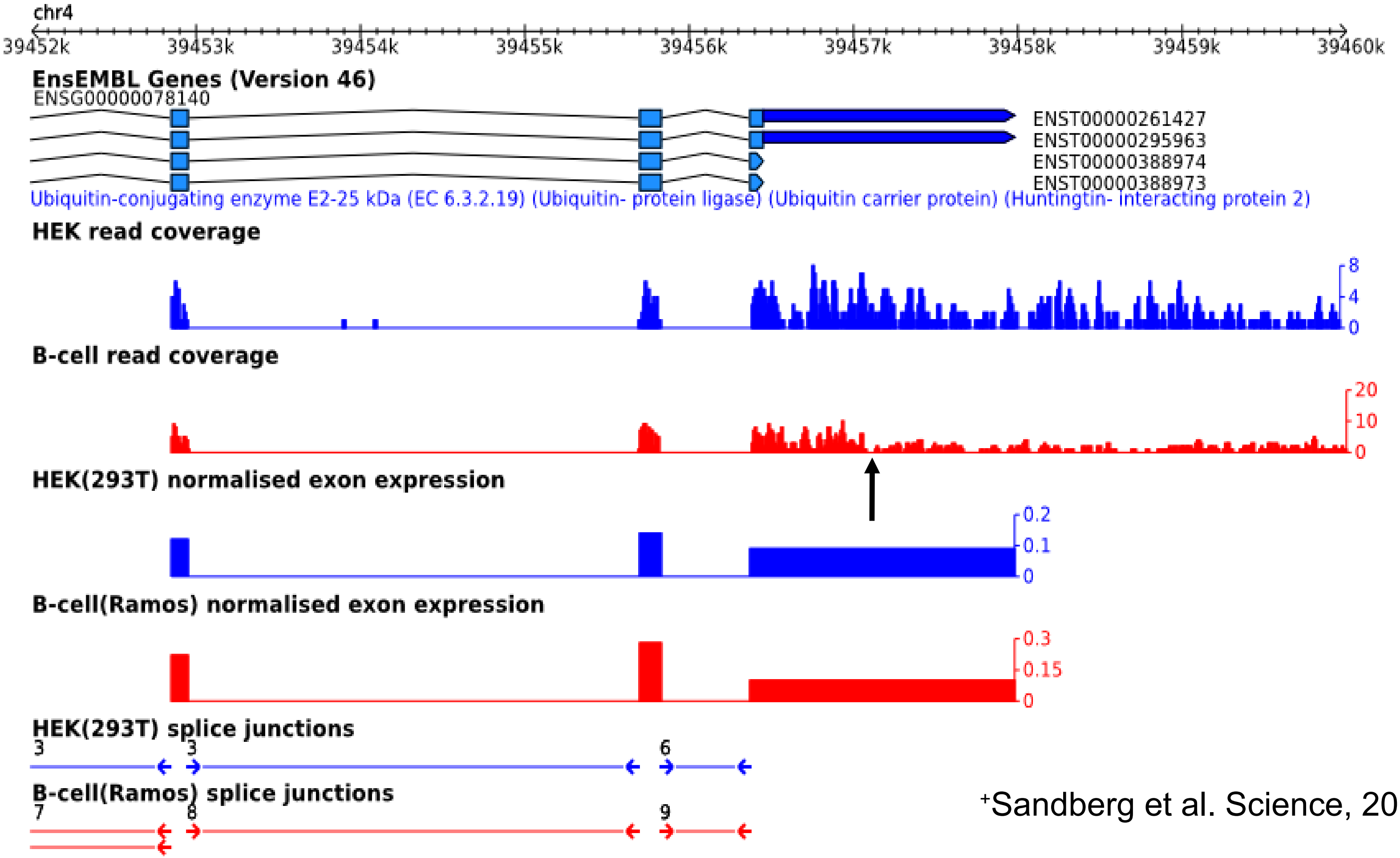


(500 repetitions)

Bootstrap:

_ Randomly alter exon boundaries
_ Monitor changes in prediction

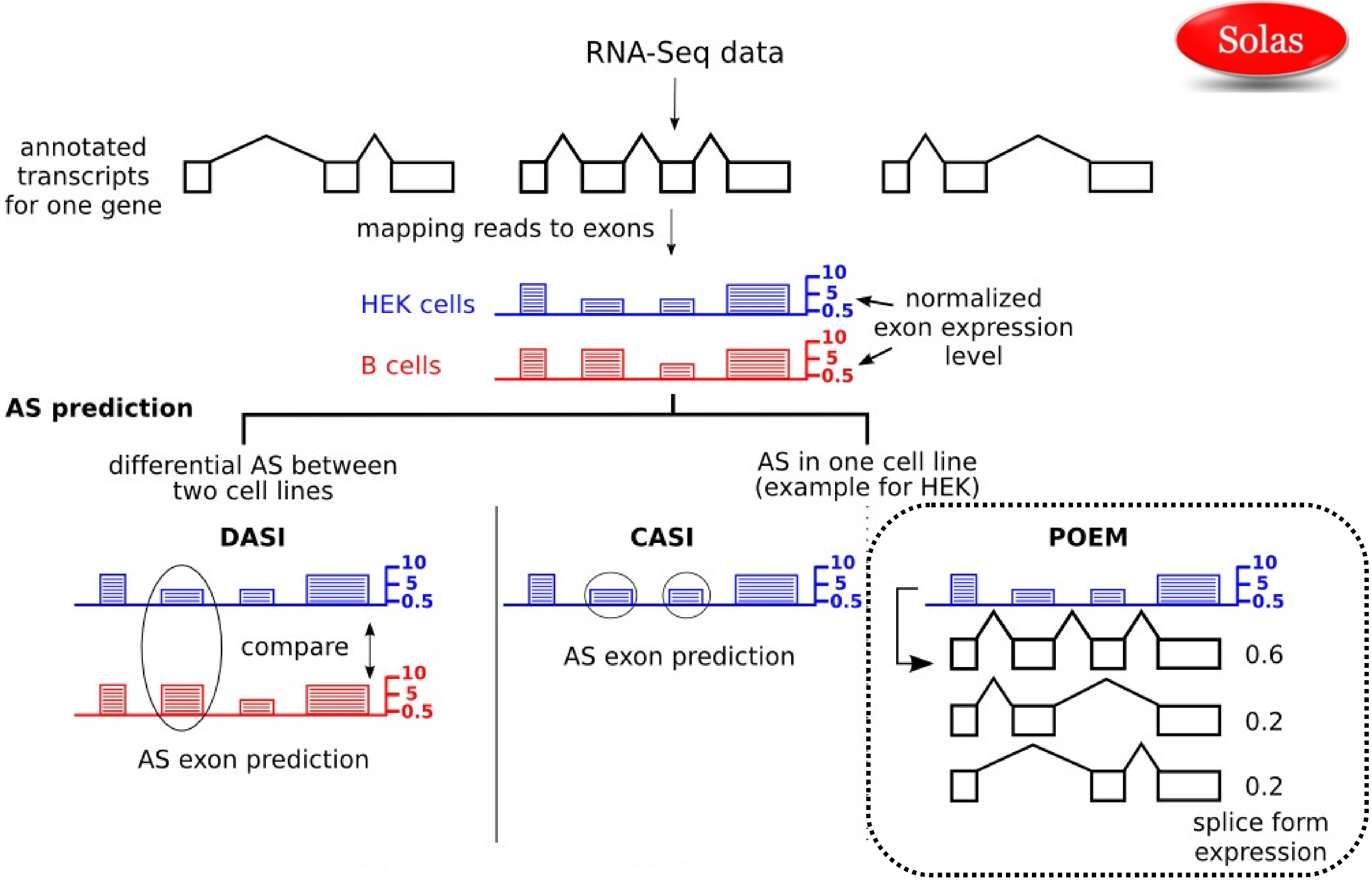
Alternative Polyadenylation in HIP2[+]



[+]Sandberg et al. Science, 2008

Detecting AEE from Exons Expression

Solas



RNA-Seq data

annotated transcripts for one gene

mapping reads to exons

HEK cells

B cells

10
5
0.5

normalized exon expression level

AS prediction

differential AS between two cell lines

AS in one cell line (example for HEK)

DASI

compare

AS exon prediction

CASI

AS exon prediction

POEM

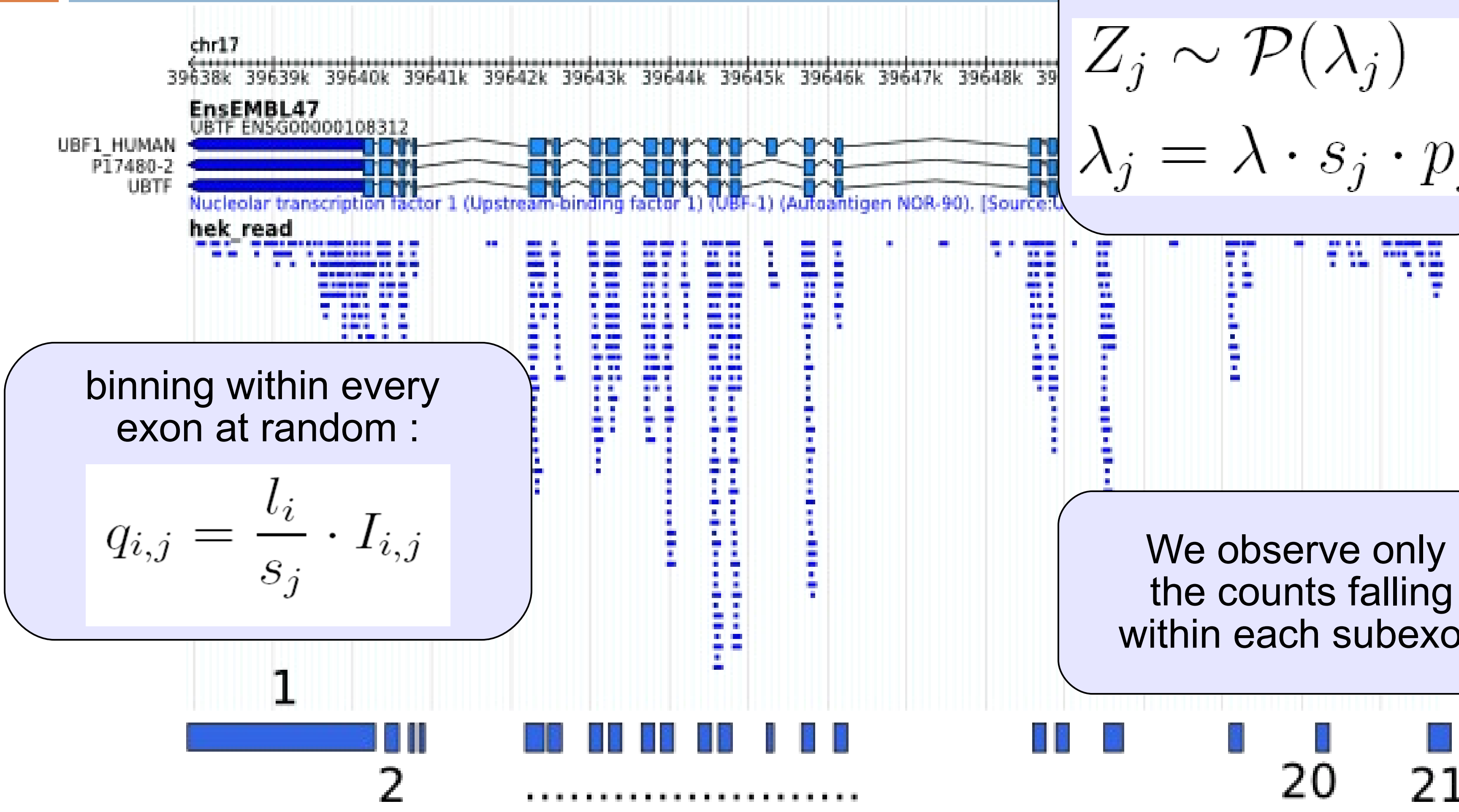
0.6

0.2

0.2

splice form expression

Isoform quantification: POEM


chr17
39638k 39639k 39640k 39641k 39642k 39643k 39644k 39645k 39646k 39647k 39648k 39
EnsEMBL47
UBTF ENSG00000108312
UBF1_HUMAN
P17480-2
UBTF
Nucleolar transcription factor 1 (Upstream-binding factor 1) (UBF-1) (Autoantigen NOR-90). [Source:U
hek_read
1
2
....
20
21

Count for each isoform j

$$Z_j \sim \mathcal{P}(\lambda_j)$$

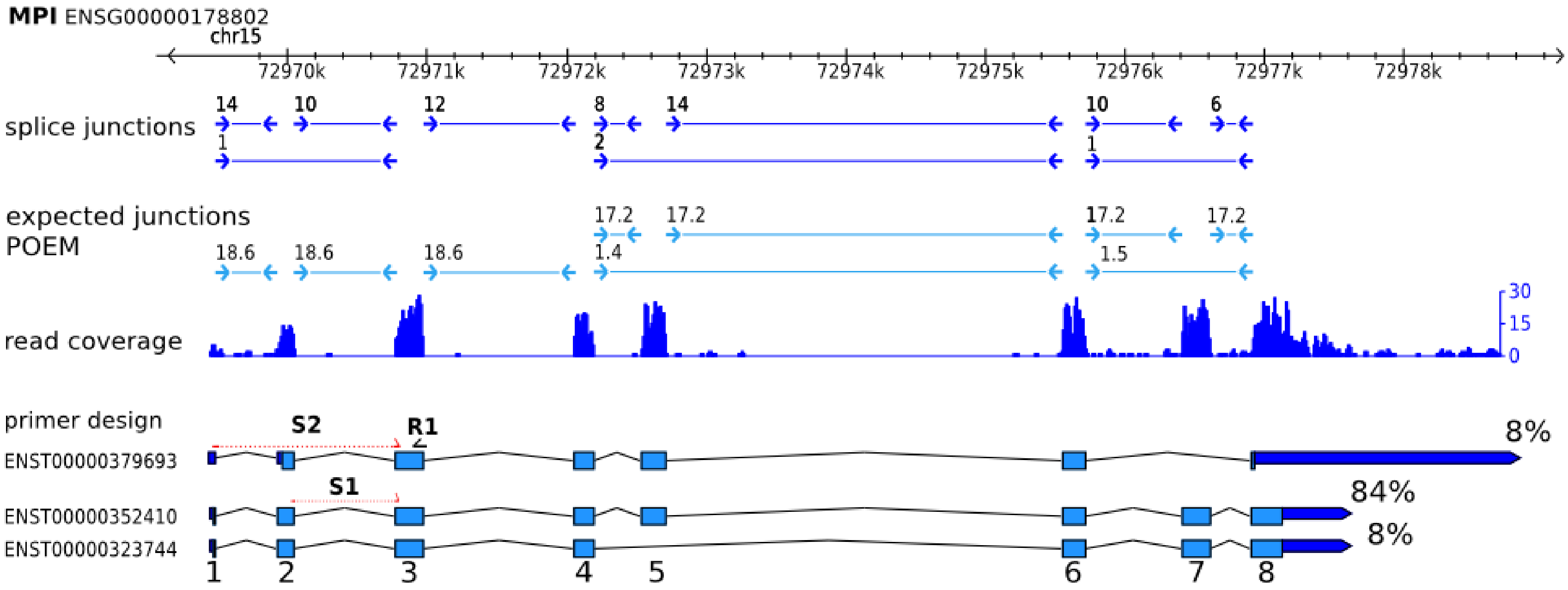
$$\lambda_j = \lambda \cdot s_j \cdot p_j$$

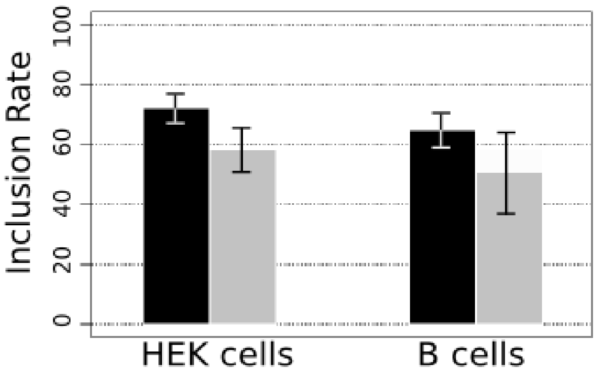
binning within every exon at random :

$$q_{i,j} = \frac{l_i}{s_j} \cdot I_{i,j}$$

We observe only the counts falling within each subexon

Isoform quantification: POEM





Comparison to:

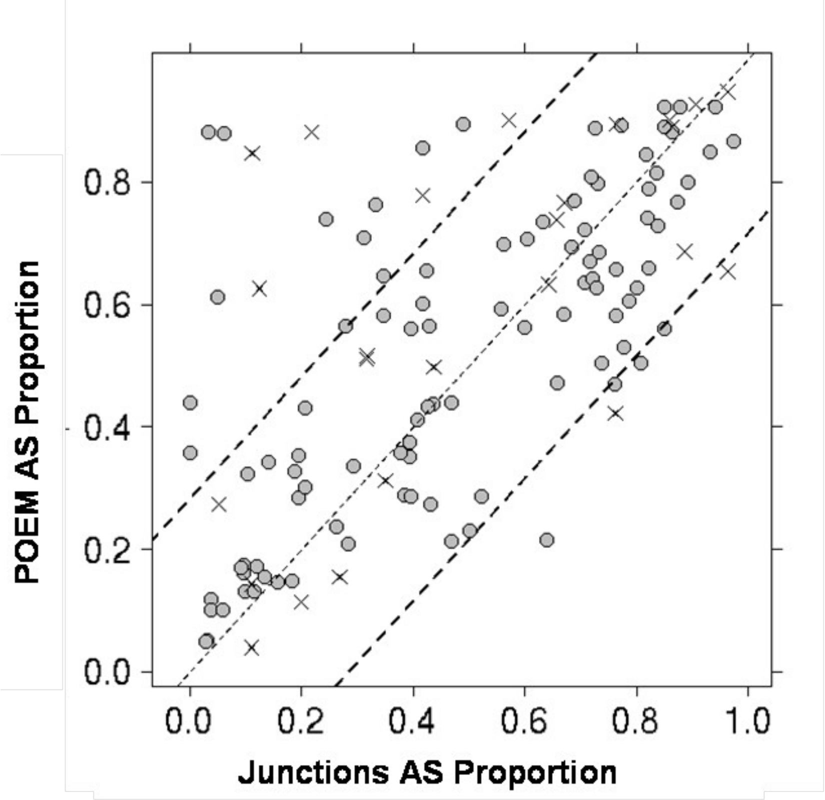
_ qPCR (47 events)

_ Estimate from junction counts (267 events)

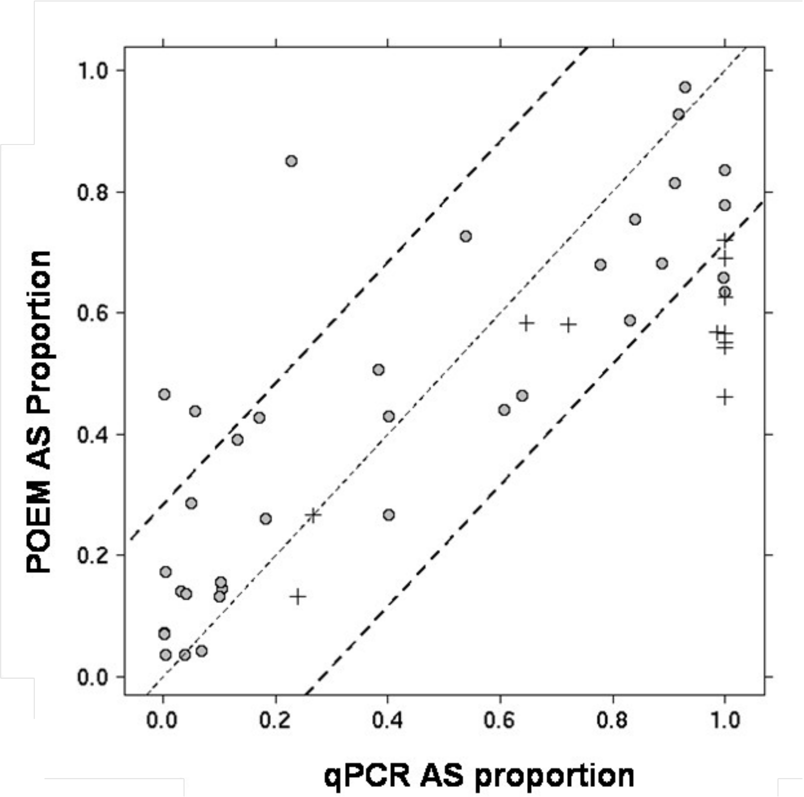
grey - POEM
black - qPCR

Validation

Junction reads



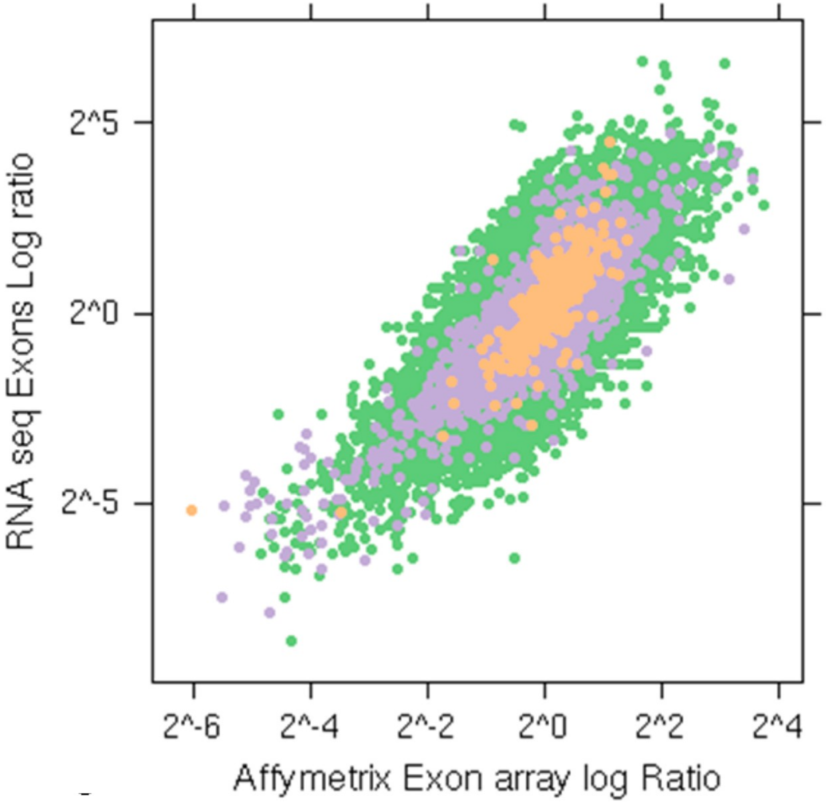
qRT-PCR

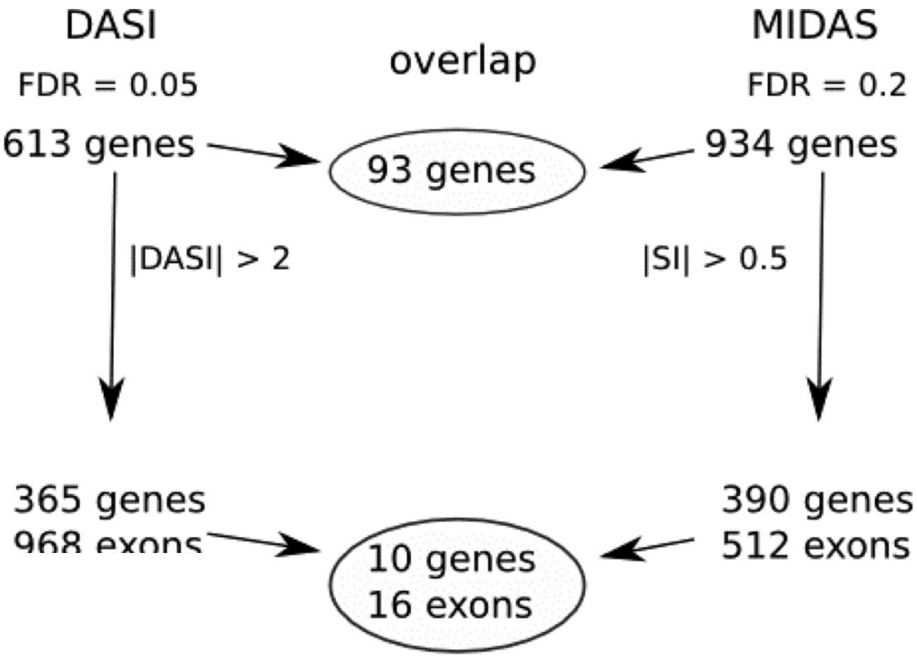


PCC: 0.65 0.81

Comparison to Exon arrays

_ produced 4 replicates Exon Array hybridizations for each cell line
_ based on ENSEMBL 25% more exona are detected by RNA-Seq





DASI

FDR = 0.05

613 genes

|DASI| > 2

365 genes
968 exons

overlap

93 genes

10 genes
16 exons

MIDAS

FDR = 0.2

934 genes

|SI| > 0.5

390 genes
512 exons

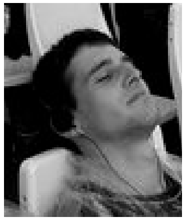
Outline

- Coverage for count based problems.
  - do we see all the genes ?
- Infering new events:
  - Detection of new transcriptional units.
  - Alternative Splicing events detection.
- RGASP competition:
  - Transcriptome assembly with Oases (Schulz/Zerbino)
  - Mapping with RazerS (Weese)

MAX-PLANCK-GESELLSCHAFT

UPMC PARIS UNIVERSITAS

Transcripts assembly



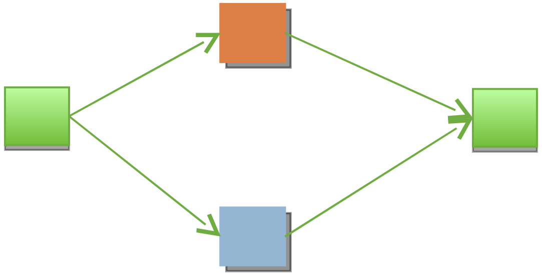
Imagine two transcripts:

TAGTCGAG GCTT TAGAGACAG

TAGTCGAG TCCGA TAGAGACAG

```
 AGTCGAG CTTTAGA  CGATGAG CTTTAGA
  GTCGAGG  TTAGATC  ATGAGGC    GAGACAG
     GAGGCTC   GTCCGAT AGGCTTT GAGACAG
 AGTCGAG    TAGATCC ATGAGGC  TAGAGAA
TAGTCGA  CTTTAGA CCGATGA    TTAGAGA
    CGAGGCT  AGATCCG TGAGGCT  AGAGACA
TAGTCGA GCTTTAG TCCGATG  GCTTTAG
   TCGATTGC    GATCCGA GAGGCTT AGAGACA
TAGTCGA    TTAGATC GATGAGG TTTAGAG
  GTCGAGG TCTAGAT   ATGAGGC  TAGAGAC
      AGGCTTT  GTCCGAT AGGCTTT GAGACAG
 AGTCGAG   TTAGATA  ATGAGGC   AGAGACA
       GGCTTTA  TCCGATG    TTTAGAG
    CGAGGCT TAGATCC  TGAGGCT   GAGACAG
 AGTCGAG  TTTAGATC  ATGAGGC TTAGAGA
```

Assemble reads into contigs



Oases:

_De Bruijn graph

_Velvet framework

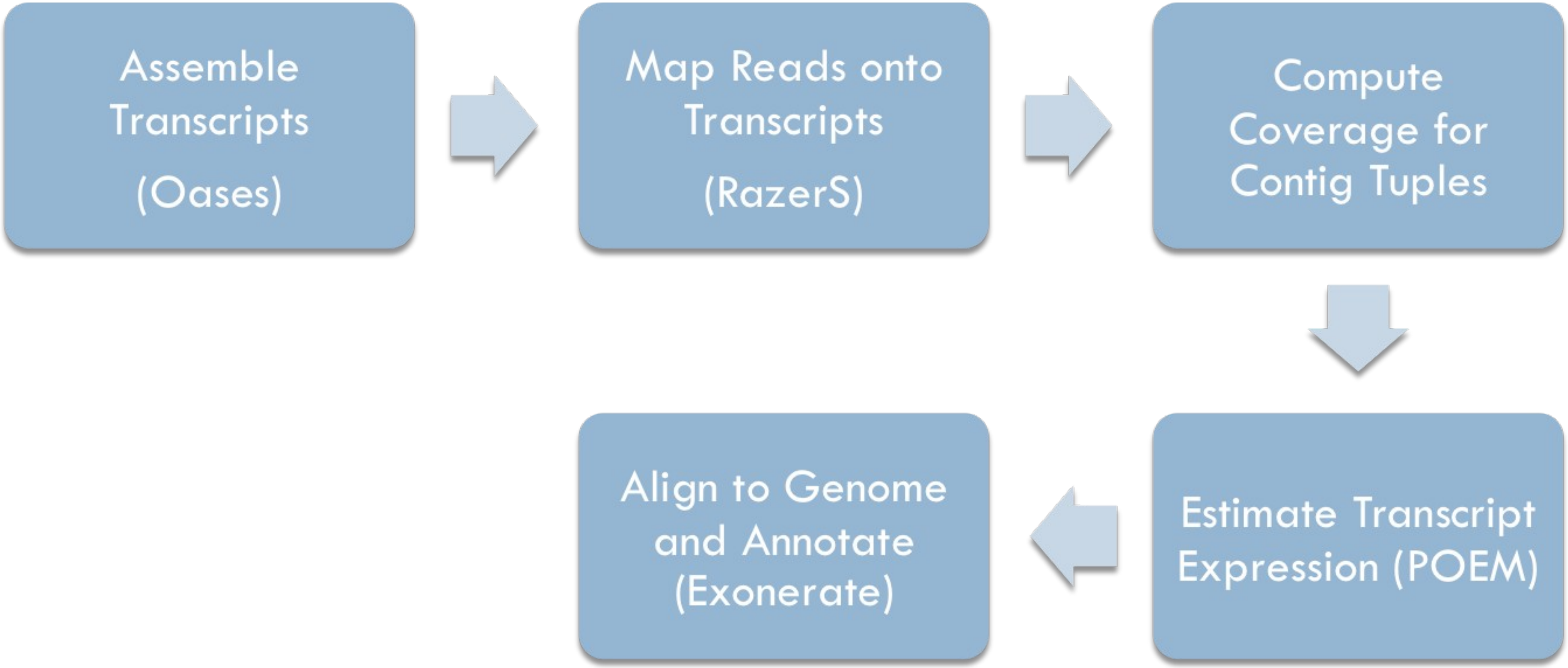
Schulz, Zerbino et al (in preparation)

De Novo transcripts assembly

- Advantages:
  - No reference or bad quality genome
  - Cancer transcriptomics (genes fusion)
  - Micro exons

- Assembly specific challenges:
  - sequencing errors are hard to rescue
  - differentiation of (post-)transcriptional modifications like alt. splicing, alt. polyadenylation, alt. first exon or trans-splicing
  - judgement of assembly quality without reference
- paralogous domain genes

Workflow for RGASP

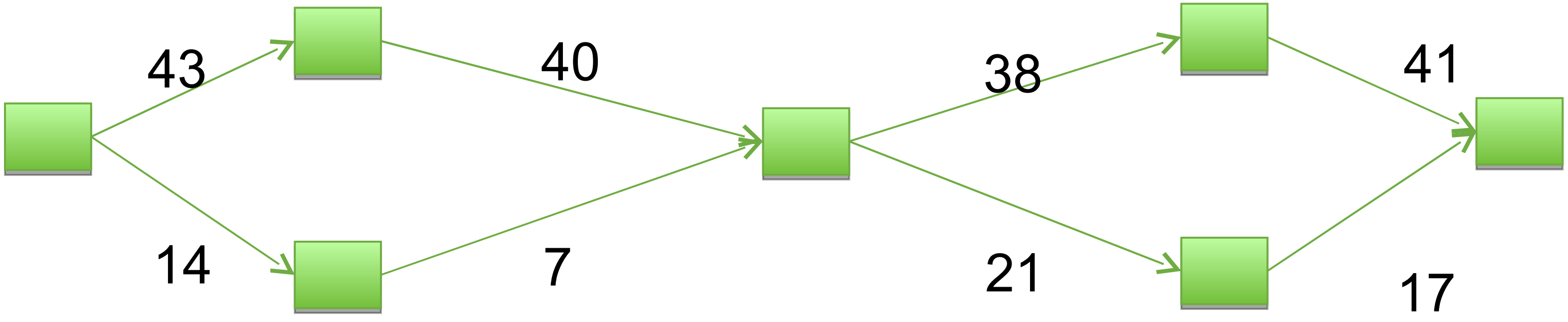

Assemble Transcripts (Oases)
Map Reads onto Transcripts (RazerS)
Compute Coverage for Contig Tuples
Estimate Transcript Expression (POEM)
Align to Genome and Annotate (Exonerate)

MAX-PLANCK-GESELLSCHAFT

UPMC PARIS UNIVERSITAS

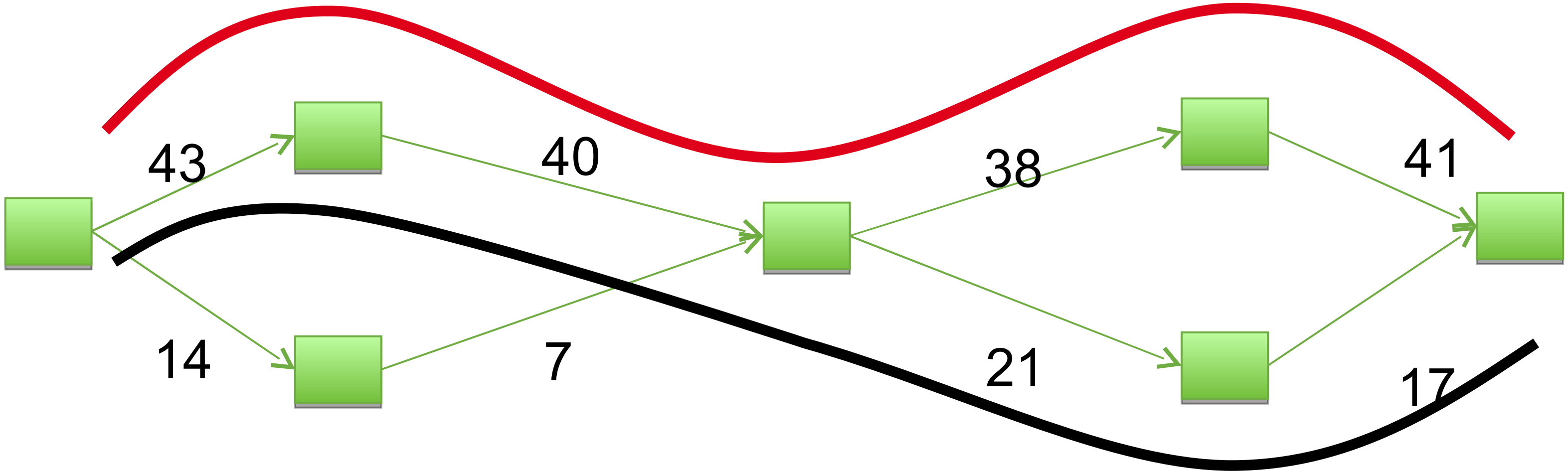
From contigs to transcripts

- weighted contig graph
  - reconstruction hard (Lacroix et al. WABI 2008)
- iterative maximum likelihood reconstruction method (heuristic) (Lee 2003)



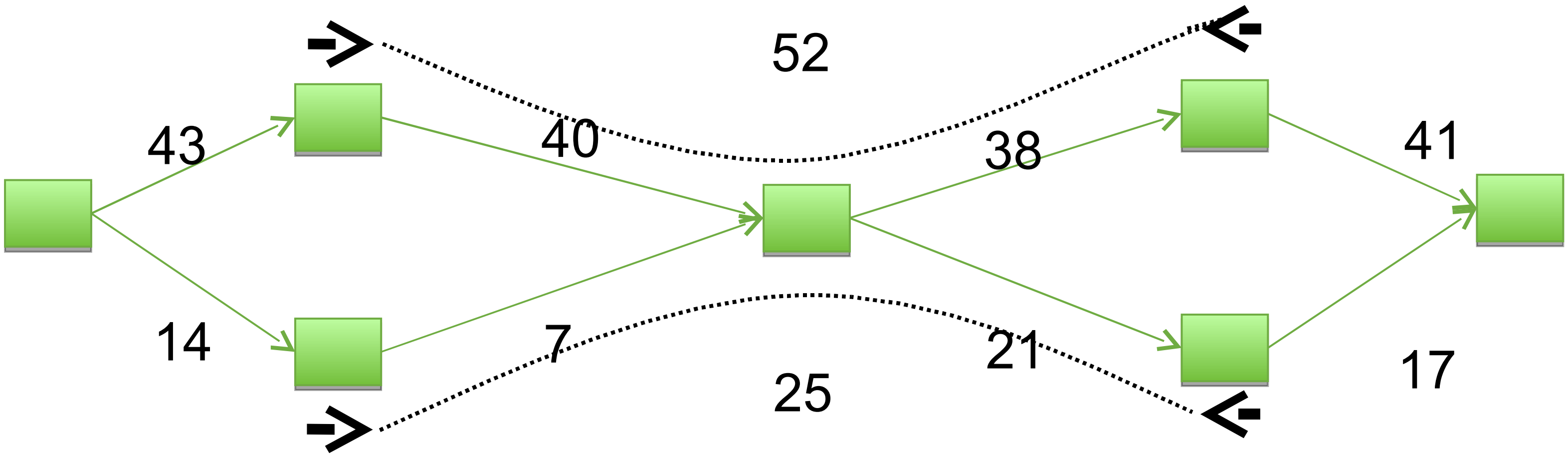
From contigs to transcripts

- weighted contig graph
  - reconstruction hard (Lacroix et al. WABI 2008)
- iterative maximum likelihood reconstruction method (heuristic) (Lee 2003)



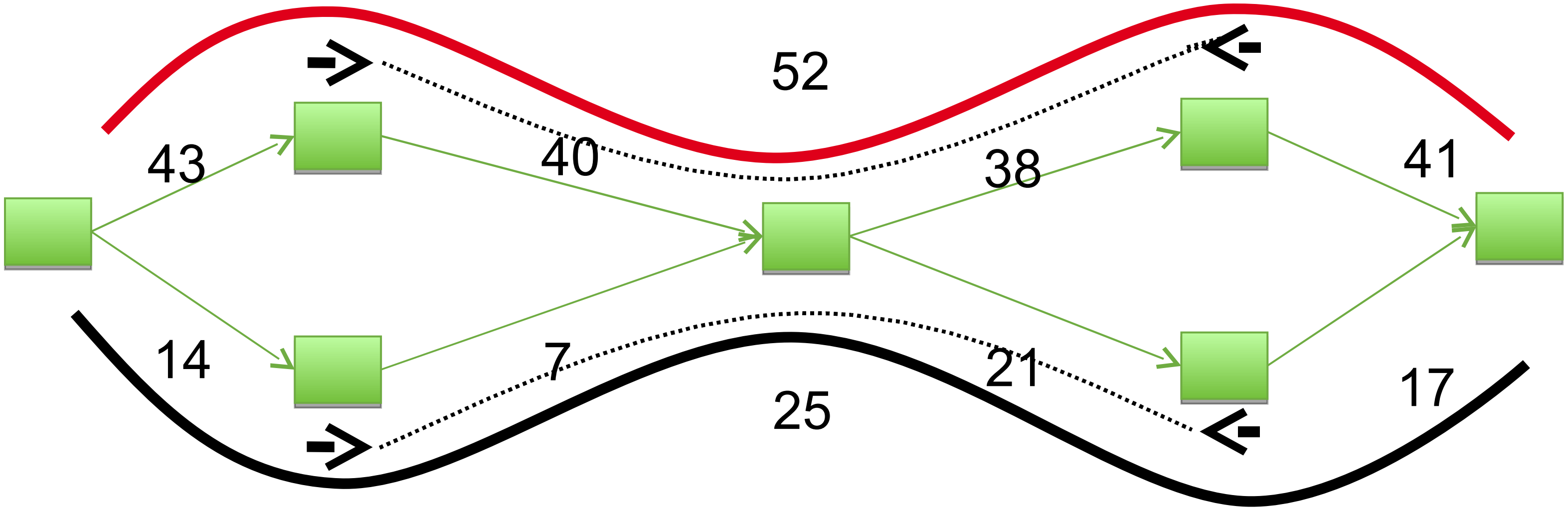
From contigs to transcripts

- weighted contig graph
  - reconstruction hard (Lacroix et al. WABI 2008)
- incorporation of paired-end reads (transitive reduction)



From contigs to transcripts

- weighted contig graph
  - reconstruction hard (Lacroix et al. WABI 2008)
- incorporation of paired-end reads (transitive reduction)



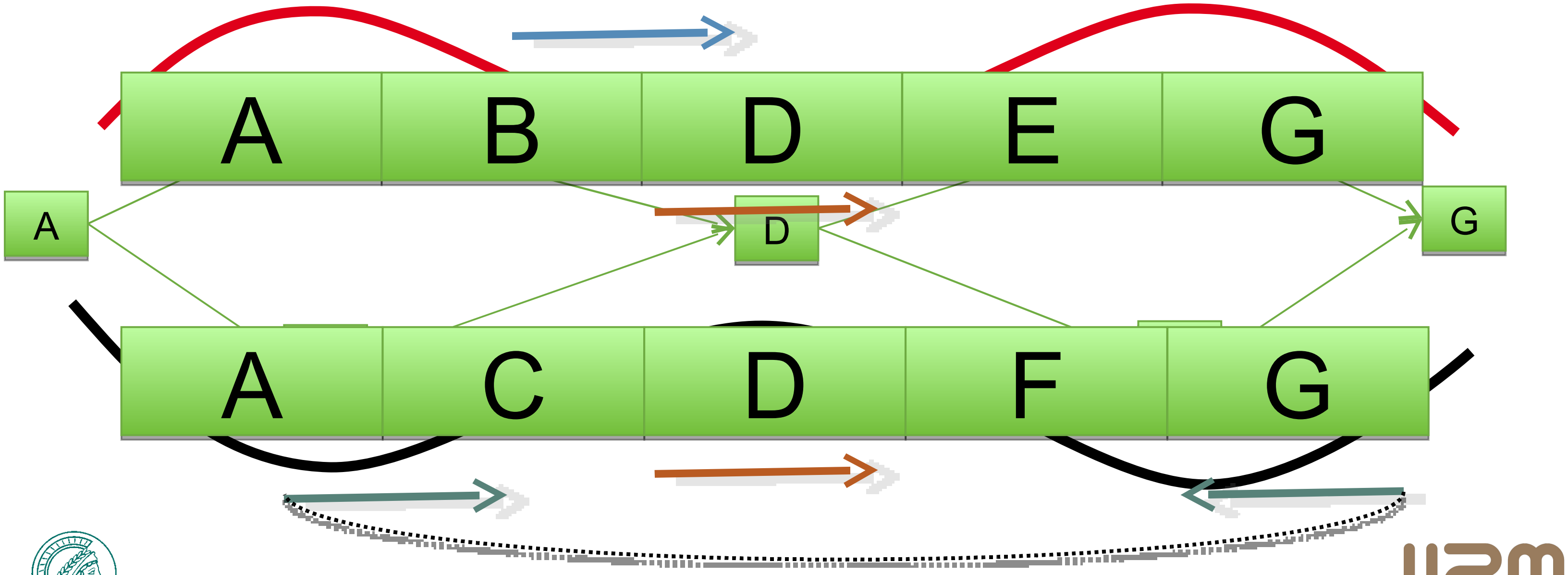
Map Reads onto Transcripts

- map all reads onto assembled transcripts
- allow for mismatches, indels
- record all (multiple) matches

- RazerS - Fast Read Mapping with Sensitivity Control (Weese et al. 2009)

Compute Coverage for Contig Tuples

- for every read match find tuples of covered contigs
- count for every tuple the number of covering reads
  - reweight multiple and paired-end matches



Perspectives

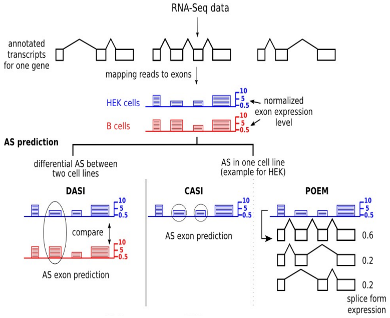
- Estimating the depth of sequencing
  - Estimates for the total number of *transcripts*
- Infering new events
  - Automatic correction of experimental biases
- Stay tuned for RGASP results

Acknowledgements

MPIMG (Berlin)

Marcel Schulz

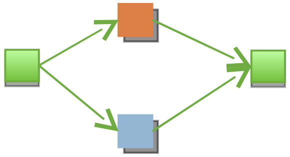
Marc Sultan



Alon Magen
Tatjana Borodina
Aleksey Soldatov
Dmitri Parkhomchuk
Stefan Haas
Martin Vingron
Hans Lehrach
Marie-Laure Yaspo

MAX-PLANCK-GESELLSCHAFT

RGASP pipeline:



Marcel Schulz
Daniel Zerbino (UCSC, former EBI)
David Weese (FU Berlin)

Ewan Birney (EBI)
Knut Reinert (FU Berlin)
Martin Vingron (MPIMG)

Now at Cordeliers
Come and visit !

UPMC
PARIS UNIVERSITAS

Results on Drosophila Paired-end reads

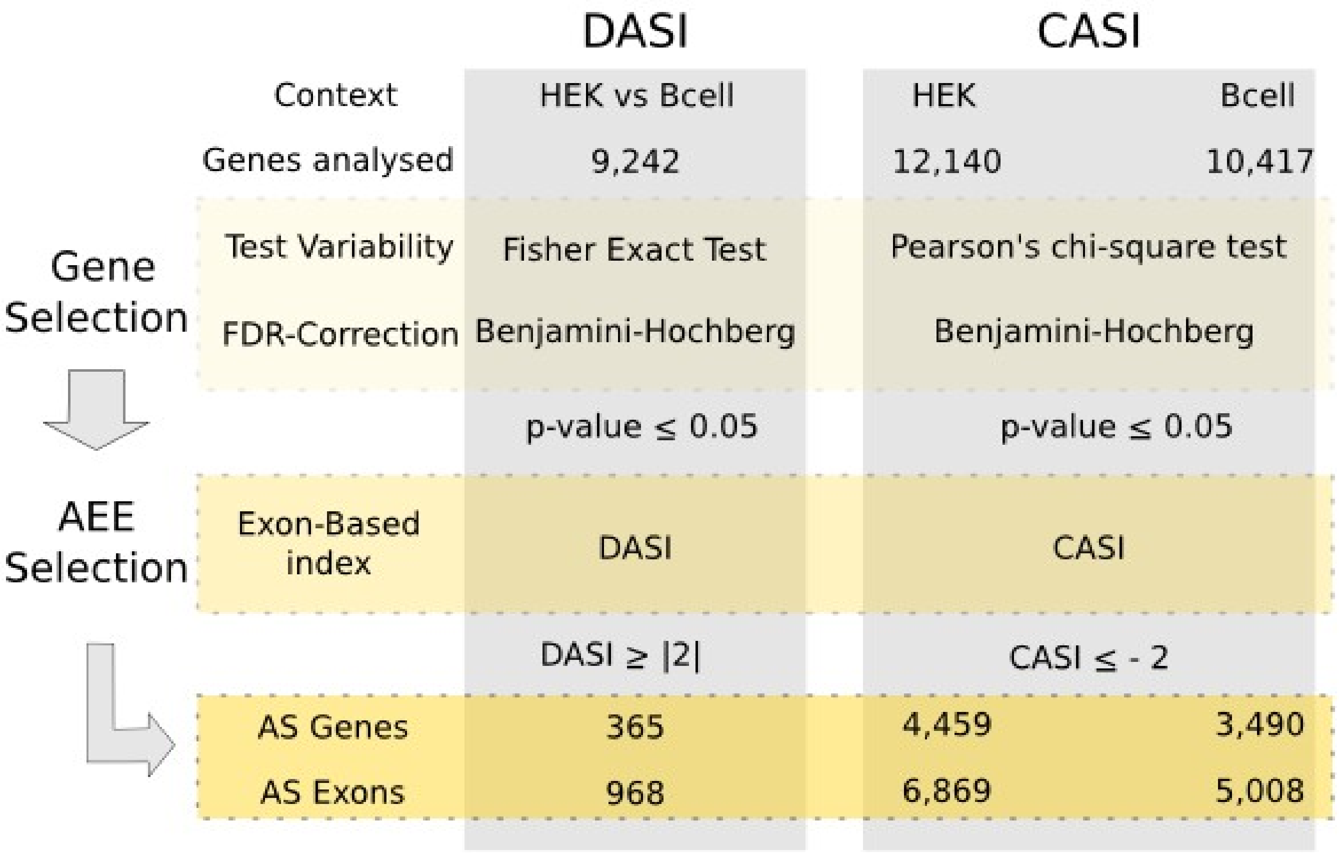
read length=36
k=21
fragment length=200

#S2-DRSC Paired-reads	# loci	# transcripts	median length	mapped to genome	Overlapping ENSEMBL transcripts	#identified exons
43,836,085*	32,905	53,299	214	89%	61%	72,892

human, fly, and worm predictions submitted to RGASP competition

*high quality reads

Statistical tests and assumptions



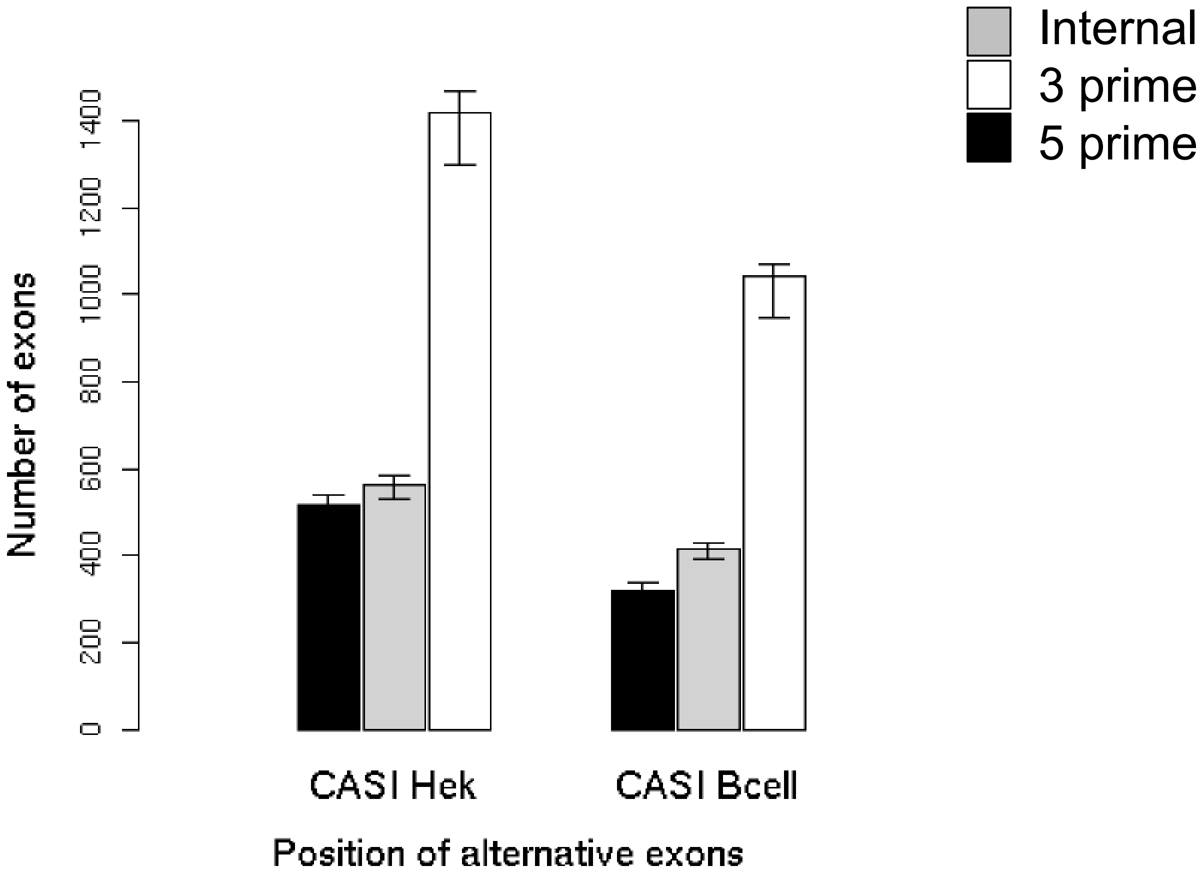
		DASI	CASI	
	Context	HEK vs Bcell	HEK	Bcell
	Genes analysed	9,242	12,140	10,417
Gene Selection	Test Variability	Fisher Exact Test	Pearson's chi-square test	
	FDR-Correction	Benjamini-Hochberg	Benjamini-Hochberg	
		p-value ≤ 0.05	p-value ≤ 0.05	
AEE Selection	Exon-Based index	DASI	CASI	
		DASI ≥ |2|	CASI ≤ - 2	
	AS Genes	365	4,459	3,490
	AS Exons	968	6,869	5,008

exon based index is a robust z-score estimated with median and median absolut deviation

MAX-PLANCK-GESELLSCHAFT

UPMC PARIS UNIVERSITAS

CASI complements splice junctions



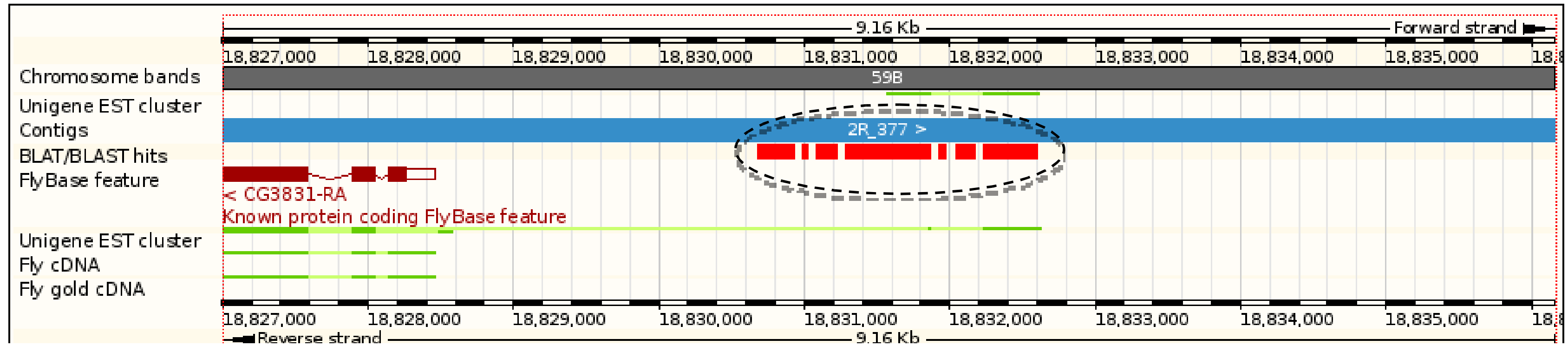
MAX-PLANCK-GESELLSCHAFT

UPMC PARIS UNIVERSITAS

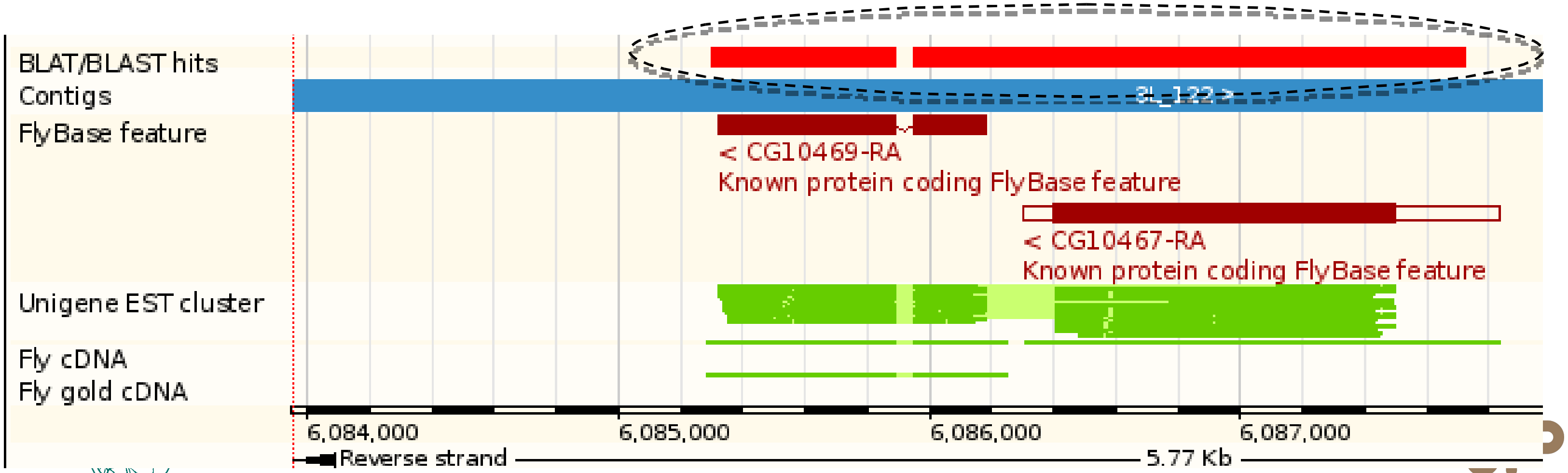
Examples

unannotated gene with 7 exons (locus 136)

assembled transcript



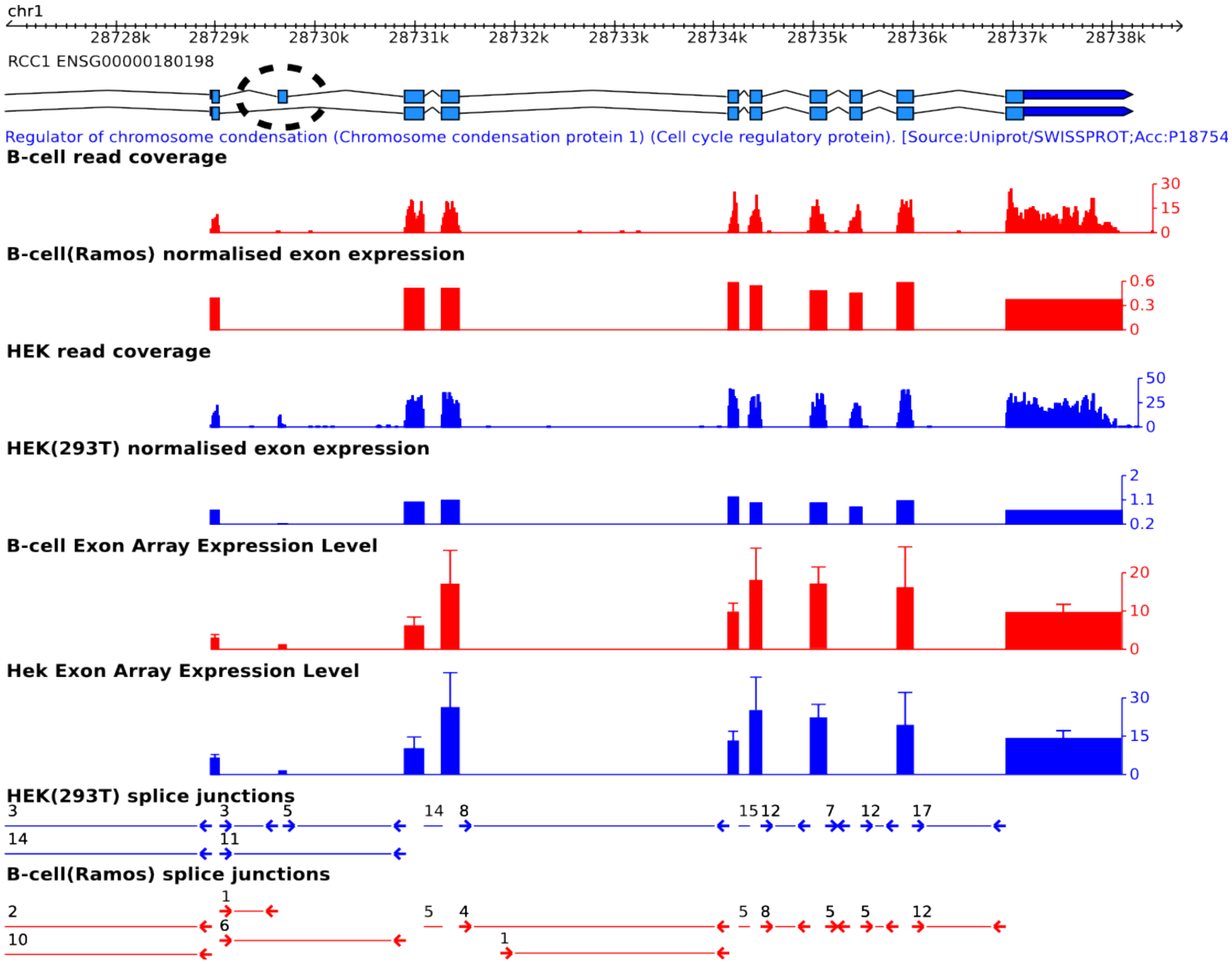
transcript connecting 2 genes (locus 2589)



MAX-PLANCK-GESELLSCHAFT

UPMC PARIS UNIVERSITAS

Exons Arrays vs RNA-Seq



MAX-PLANCK-GESELLSCHAFT

UPMC PARIS UNIVERSITAS