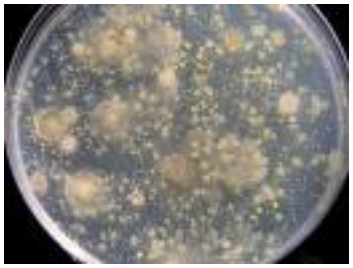
UPMC
PARIS UNIVERSITAS

Hugues Richard

Technology Transforming Biology

Until late 20th Century



Hypothesis Generation and Validation

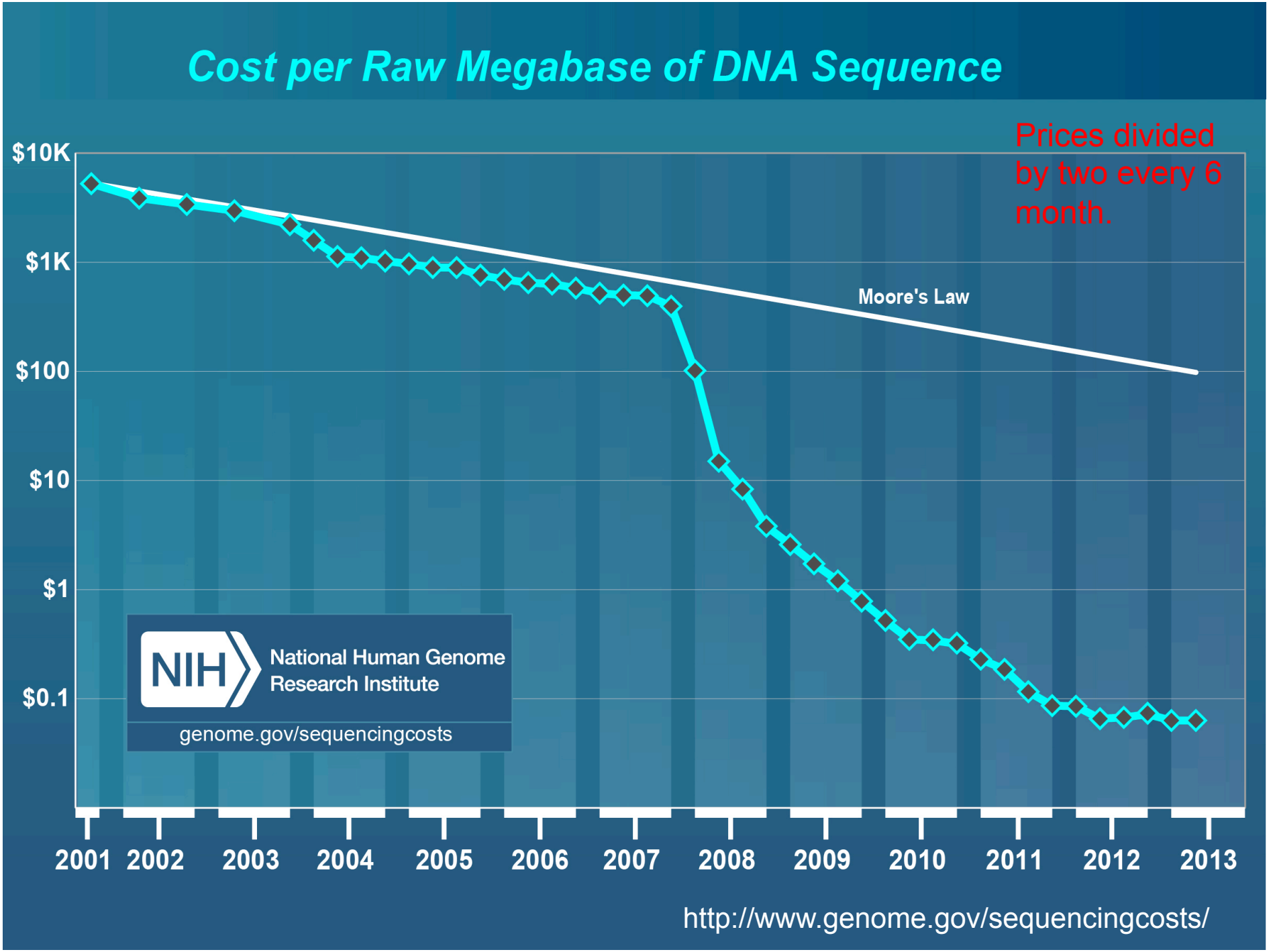
21th Century and Beyond

Mathematics & Computation

Hypothesis Generation and Validation

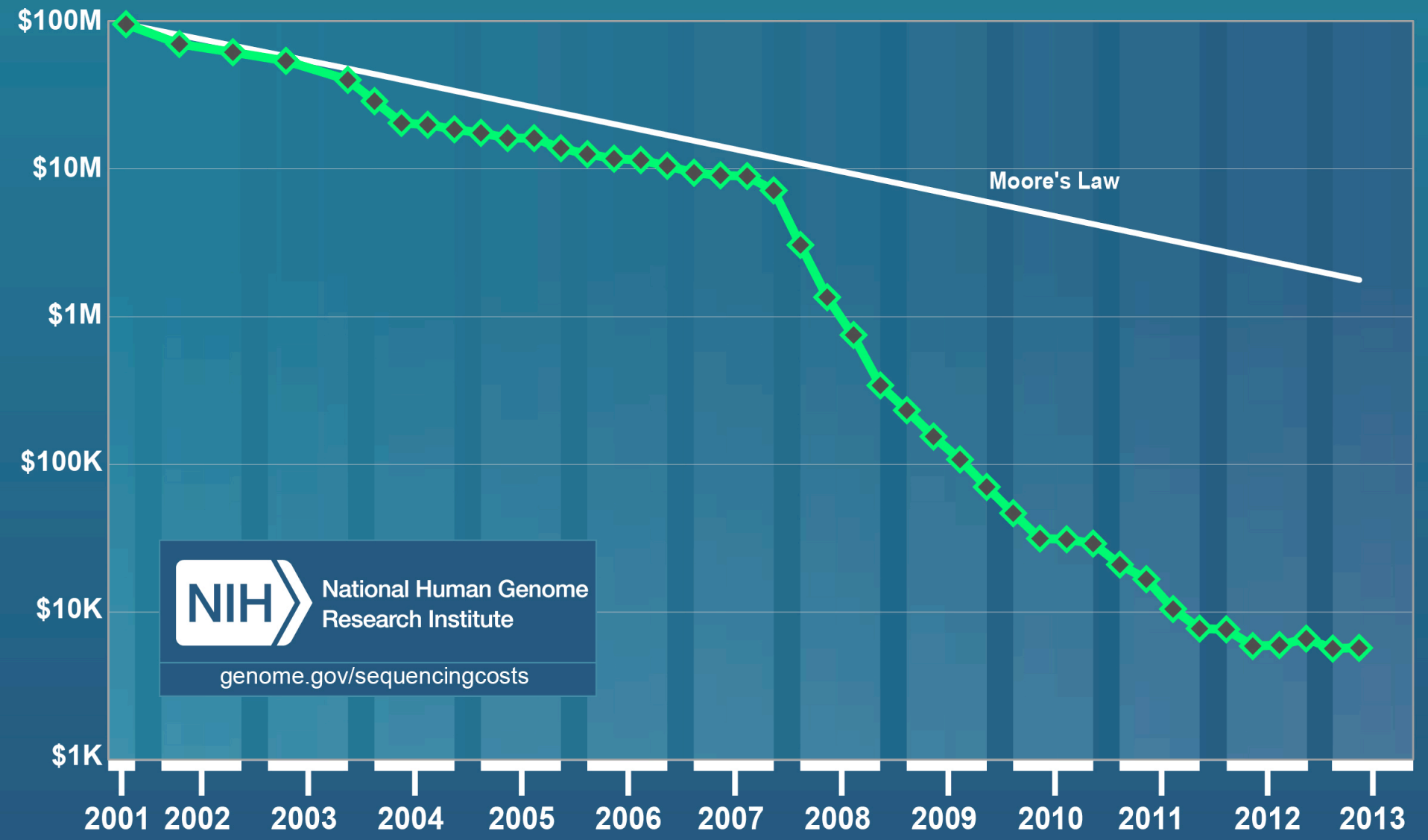
High throughput technologies

Cost per Raw Megabase of DNA Sequence



http://www.genome.gov/sequencingcosts/

Cost per Genome



Analysis of NGS data

- Lectures:
  - 12/01: High-throughput sequencing, sequence alignment and error correction
  - 19/01: Genome sequencing and resequencing
    - genome assembly
    - Annotation of sequence polymorphism (A. Gillet)
  - 26/01: Transcriptome analysis
- Tutorials:
  - 12/01, 26/01: Pan & Core Genome (A. Ugarte)
  - 19/01: detection of structural variants (A. Gillet)

1. Genome (Re)Sequencing

Sequence alignment / Analysis of variants

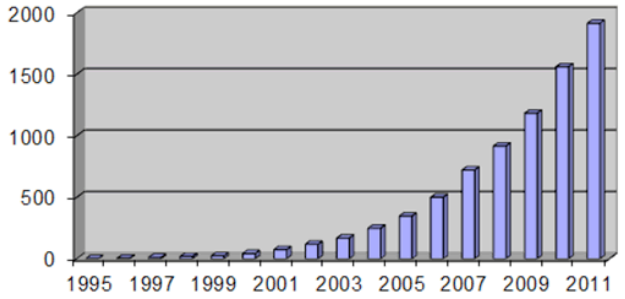
Outline

- Brief historical background on sequencing
  - the many applications of High throughput sequencing
- A typical resequencing workflow
  - Reads alignment
  - Basic statistics after alignment
  - Detection of sequence polymorphisms
  - Genome Assembly

Timeline of genomes

- 19 years of sequencing
  - 1995: first bacterial genome *H. Influenza*
  - 1996: first eukaryotic genome *S. Cerevisiae*
  - 2001: Completion of the Draft of the Human genome.
    - 10 Billion Dollar and 10 years to complete.
  - Now, almost 2k genomes

(free download)



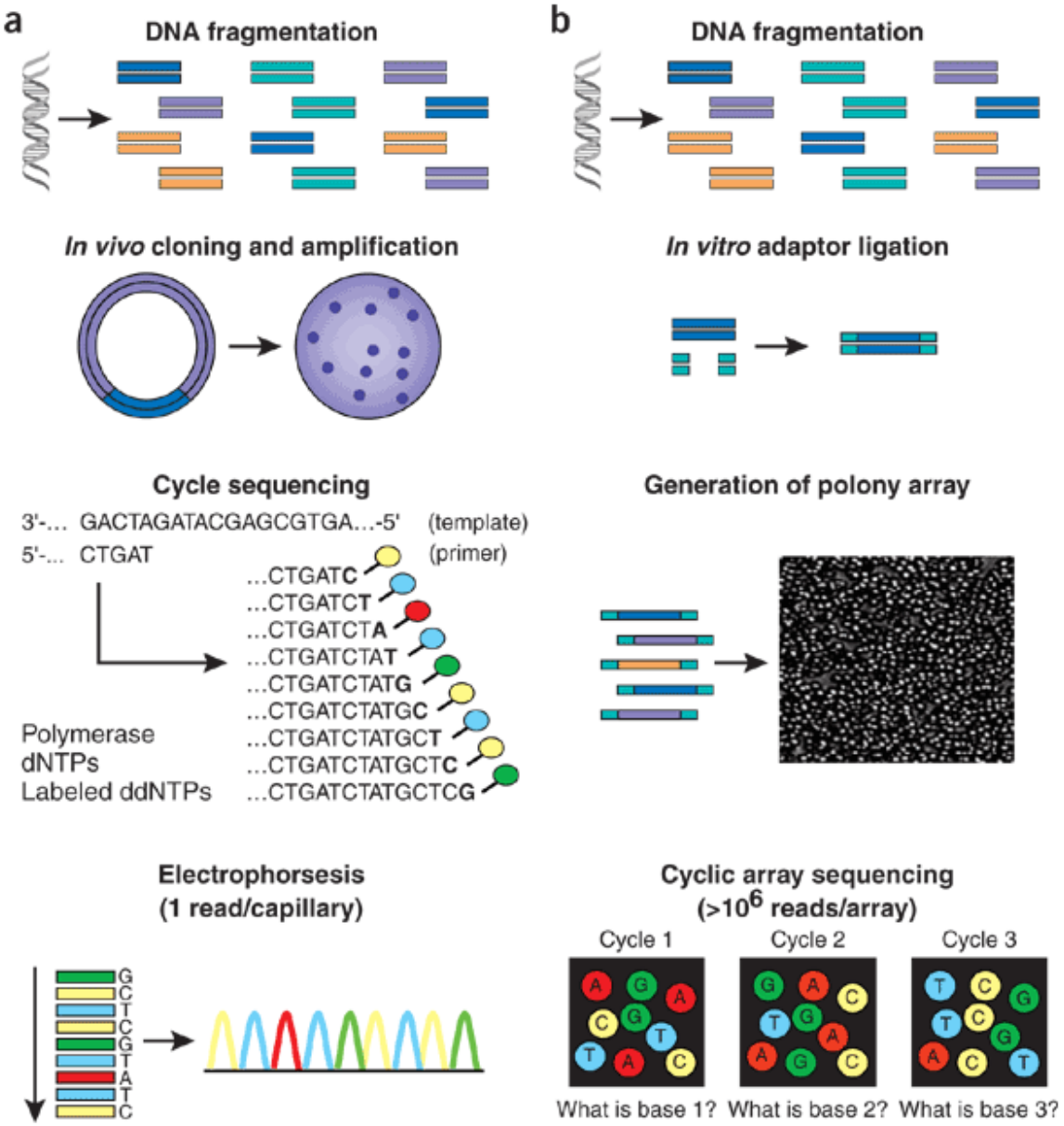
- 2005 on: Massive development of new technologies
  - One human genome < 5,000$, less than 1 week.

Changes in instrument capacity over the past decade, and the timing of major sequencing projects

nature

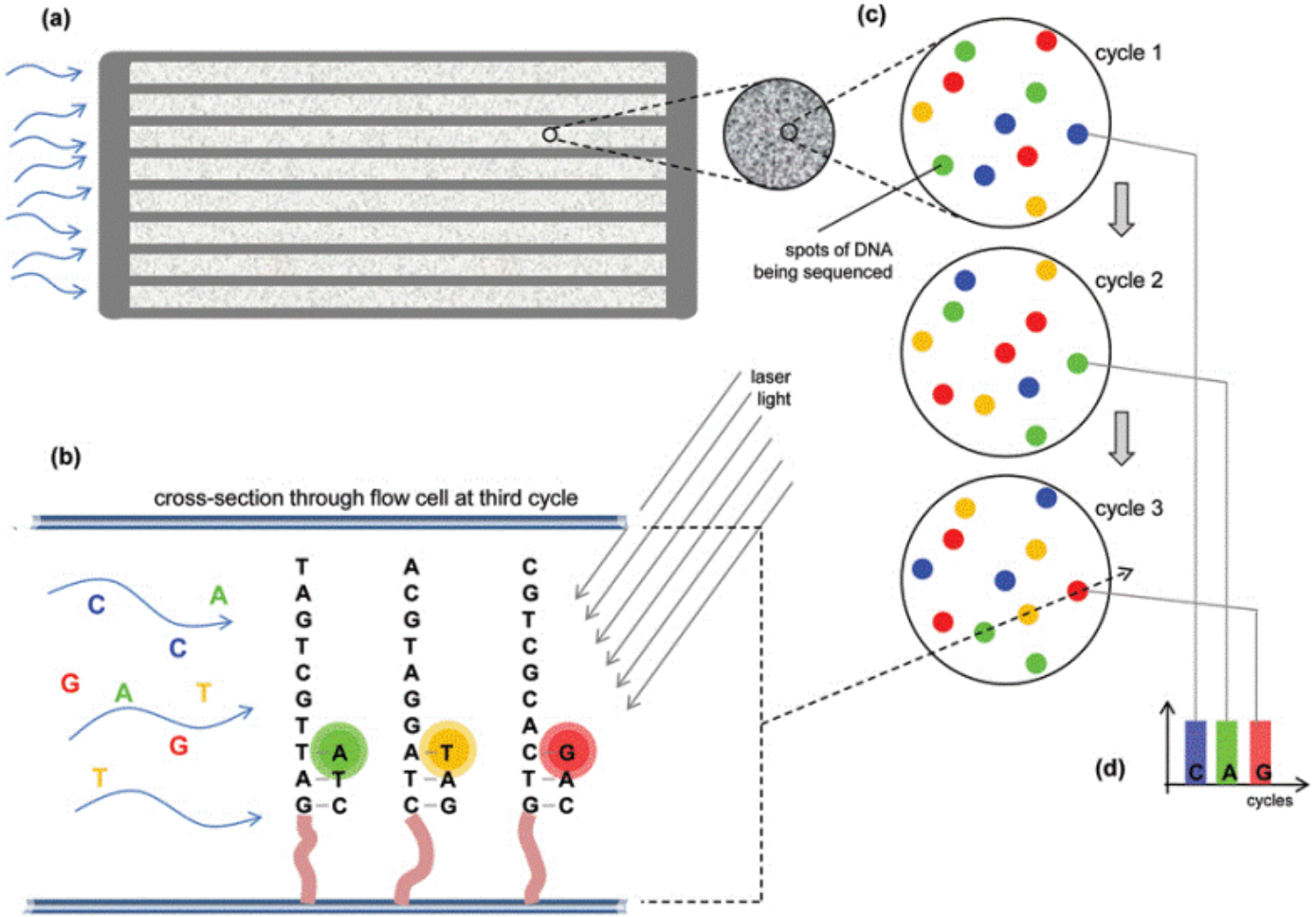
ER Mardis. *Nature* **470**, 198-203 (2011)

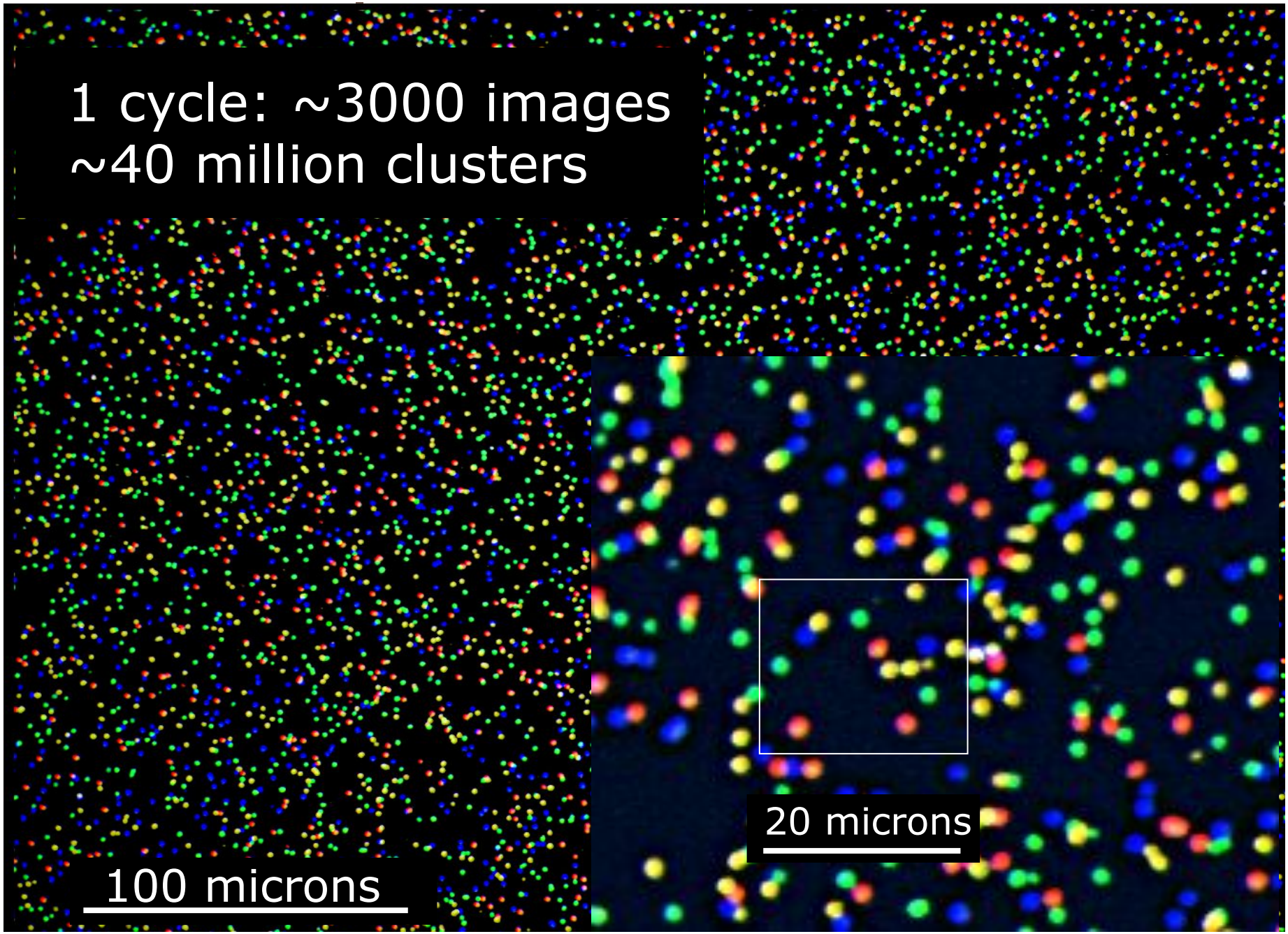
Specificity of 2nd generation technologies

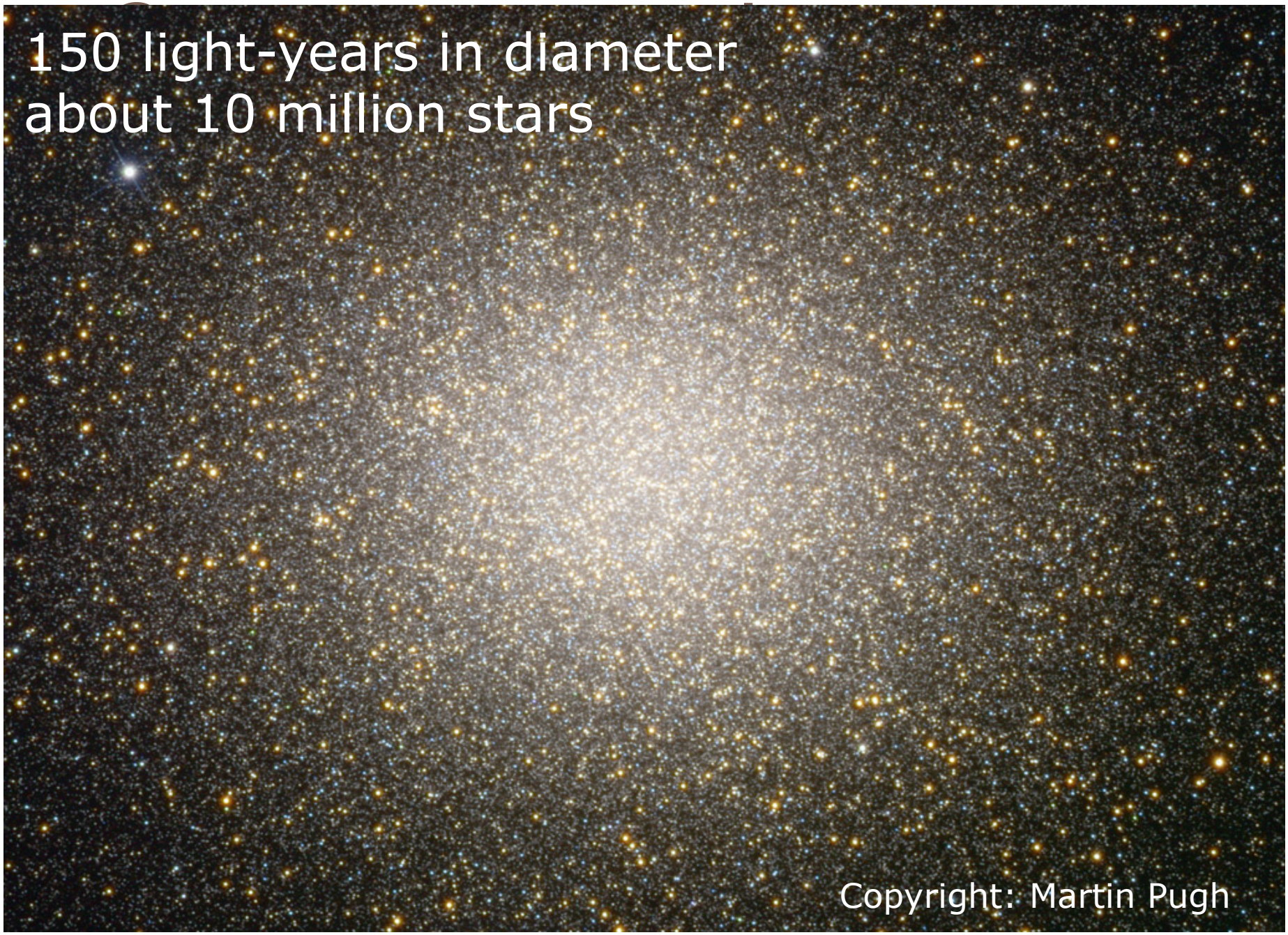


a. Sanger sequencing
b. Illumina sequencing

Illumina Sequencing

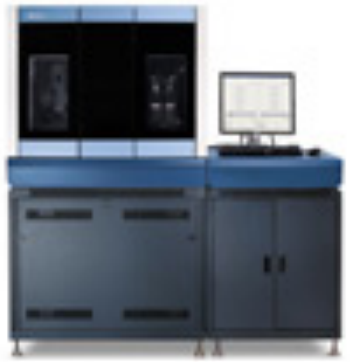



1 cycle: ~3000 images
~40 million clusters
100 microns
20 microns


150 light-years in diameter
about 10 million stars
Copyright: Martin Pugh

Next|2nd Generation Sequencing

Machine	454 FLX [Roche]	HiSeq 2000 [Illumina]	SOLiD 4 [ABI]
Sequencing Approach	Pyrophosphate Release	Bridge Amplification	Ligation
Read Lengths	≈ 400|2x140 bps	2x100 bp	50|50+35 bp
Paired Ends	Yes	Yes	Yes
Output per Run	≈ 400 Mb in 2-4 d	100|200 Gb in 4|8 d	100 Gb in 6 d
Accuracy depends on	Homopolymer Length (>6 problematic)	Position in the Read	Position in the Read
Costs per Run	$9,000	$10,000	$6,000



Numbers as of 2011.08

Next Generation Sequencing Platforms

Complete genomics

HiSeqXTen

Jusqu'à 1,8Tb

HiSeq

Jusqu'à 1000Gb

Jusqu'à 100 Gb

Solid5500

NextGen500

Jusqu'à 120 Gb

Ion Proton

Jusqu'à 10 Gb

MiSeq

Jusqu'à 15 Gb

PGM

De 10 Mb à 1,5 Gb

Jusqu'à 700 Mb

Jusqu'à 35 Mb

Recherche

Clinique, Diagnostic

courtesy C. Bole-Feysot

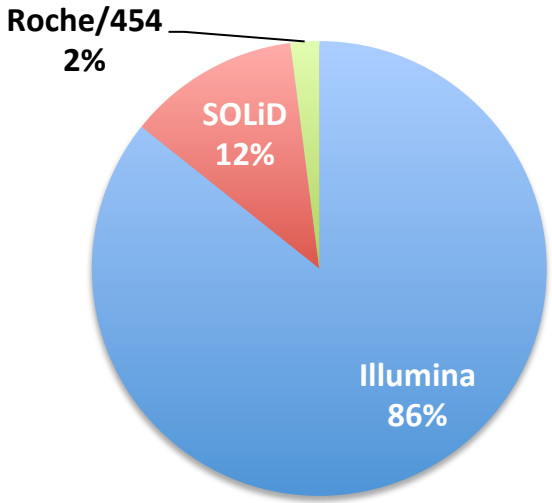
Single molecule sequencing

Oxford Nanopore Minion

High-Throughput Sequencing

technology	instrument	run time	yield [Mb/run]	read length [bp]	costs [$/Mb]	error rate [%]
Sanger	3730xl (capillary)	2 h	0.06	650	1500	0.1–1
Illumina	HiSeq 2000	8 d	200,000	2 × 100	0.10	≥ 0.1
SOLiD	SOLiD 4	12 d	71,400	50 + 35	0.11	> 0.06
Roche/454	FLX Titanium	10 h	500	400	12.4	1
SMRT™	PacBio RS	0.5–2 h	5–10	860–1100	11–180	16
HeliScope™	Helicos	N/A	28,000	35	N/A	N/A

(Glenn, 2011)

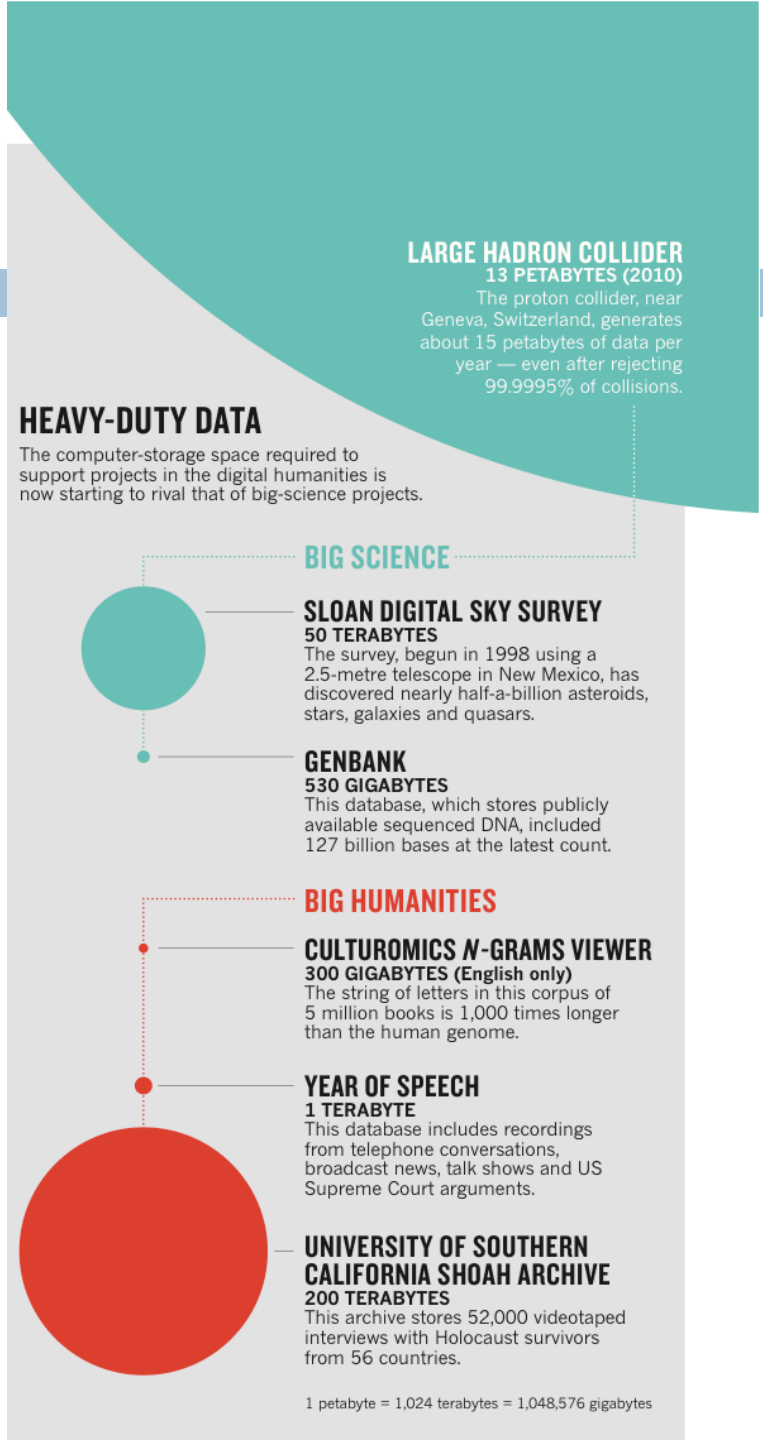


More than 100 Terabases publicly available in the Short Read Archive (Kodama et al., 2012)

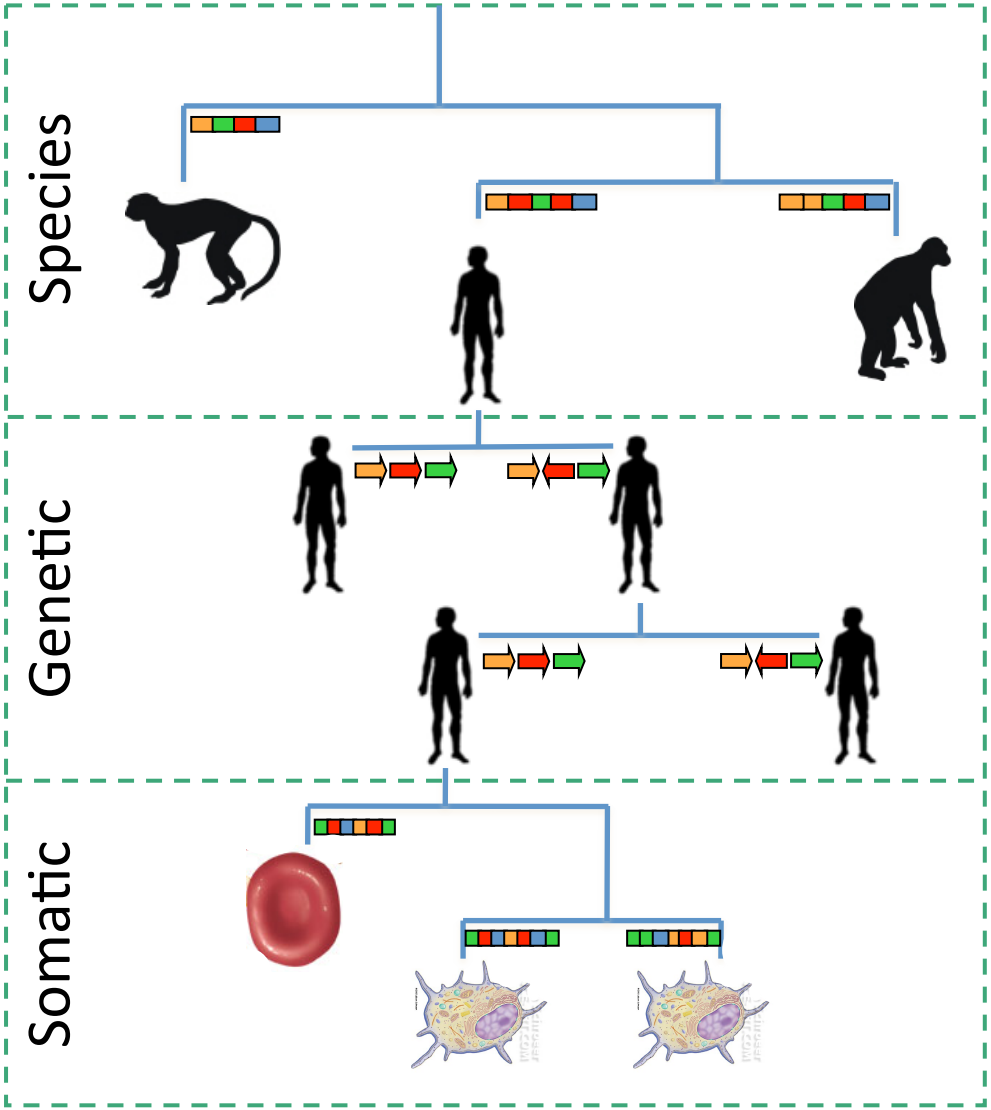
Big data

- One sequencing experiment:
  - > 100 Gb of sequence

    Human genome 3.3 Gb



Genome Sequencing and Comparison



Comparative genomics

Differences between species?

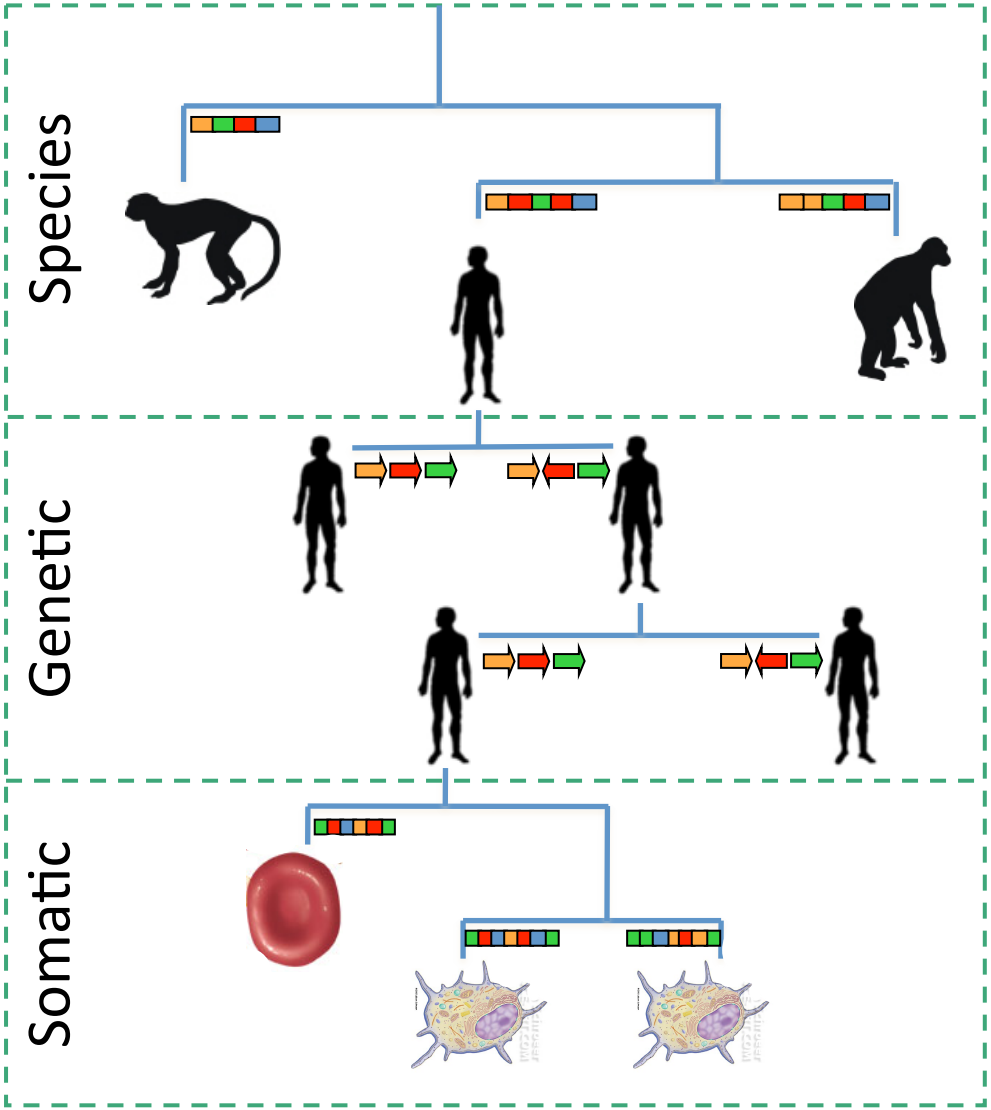
Personal genomics

Genetic basis for traits of individuals?

Cancer genomics

Which somatic mutations lead to cancer?

Petabytes of Genomes



GENOME 10K.

1000 Genomes
A Deep Catalog of Human Genetic Variation

Autism Genome 10K

THE CANCER GENOME ATLAS

International Cancer Genome Consortium

Genome (re)-sequencing

- Evolutionary studies:
  - 10k genomes on vertebrate species

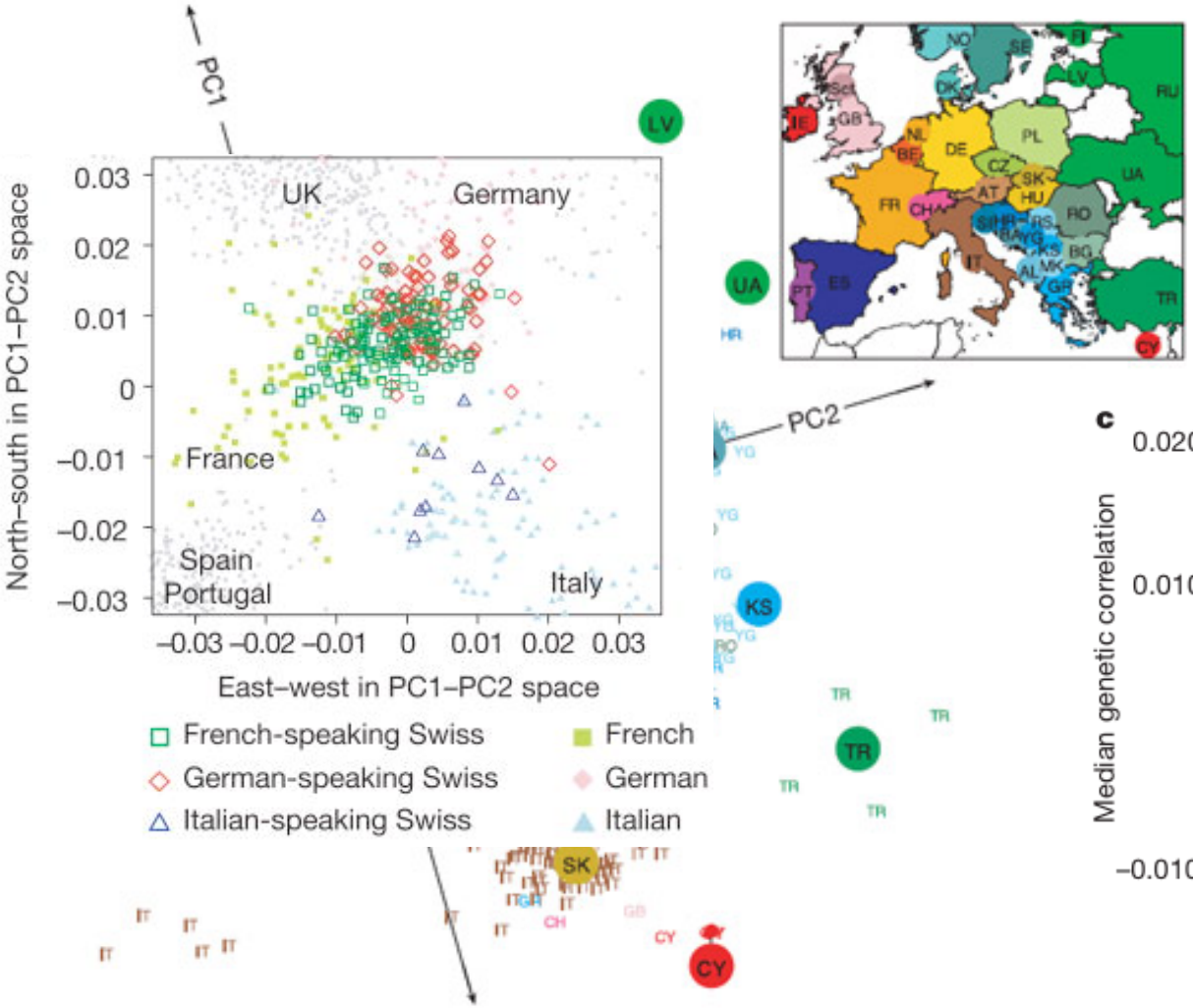
- Gene Wide association studies
  - Autism genome 10k
- Population genomics:
  - 1001 genomes (A. Thaliana)

- Personal genomics
- Tumor cell profiling (cancer genomics)

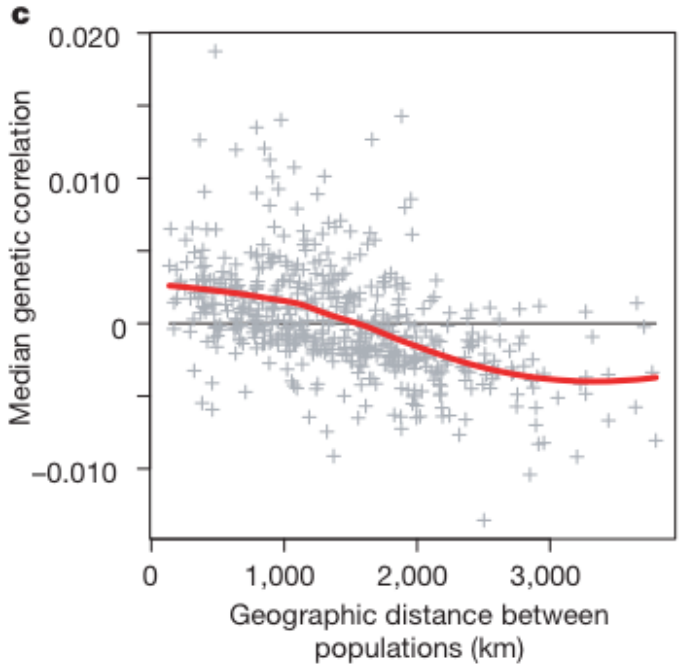
The Cancer Genome Atlas

Understanding genomics to improve cancer care

Genotype variation



PCA on 3,192 individual from all Europe. Data are alleles at a subset of genomic loci (~ 200k loci)



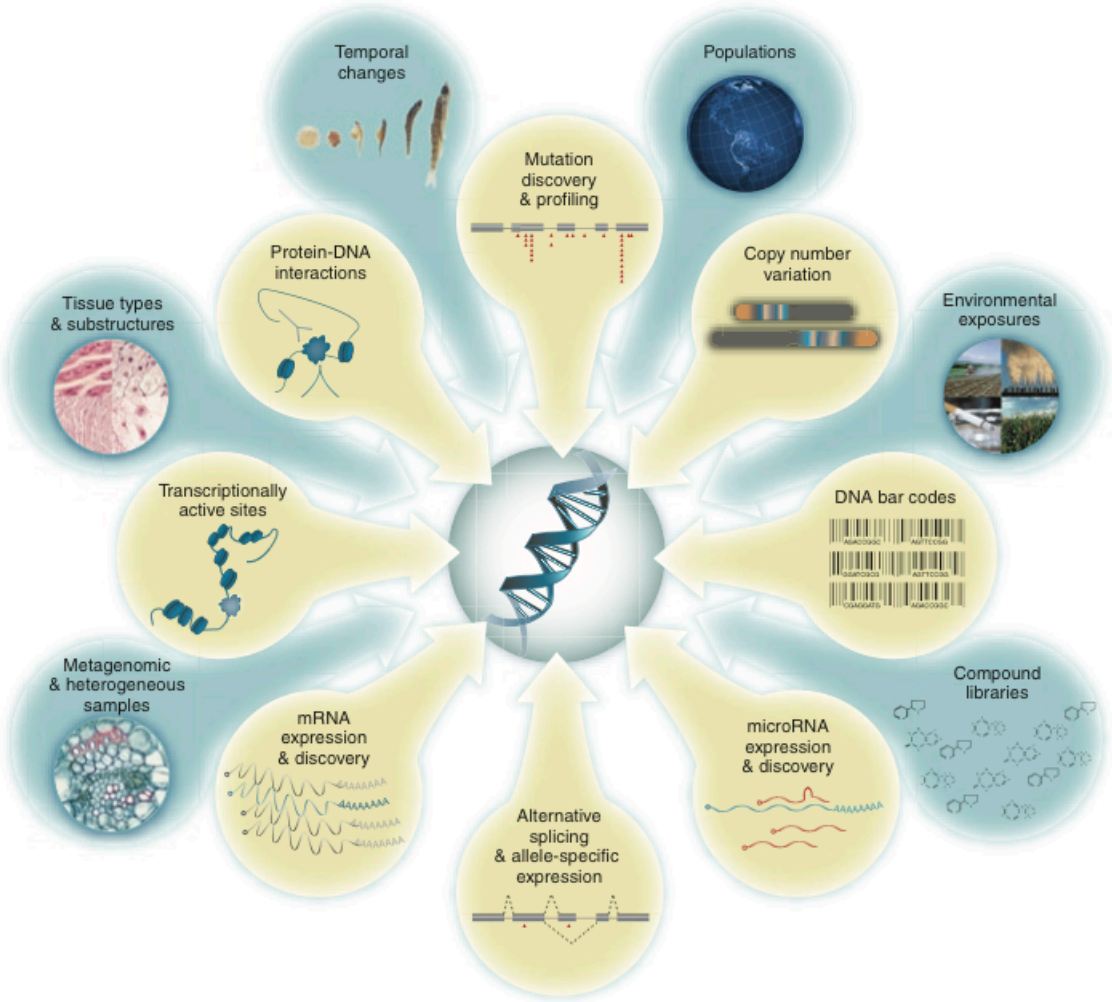
J. Novembre, Nature 2008

Sum-up

- Increasing availability of genomes will have a profound impact.
  - Personal Computers in the 80s, now Personal Genomes
- Large amounts of data:
  - Need efficient algorithms
  - Many (Mio) sequences + errors → Statistical reasoning
- Main questions for primary analysis in genomics:
  - Assign each sequence to its originating position
  - Annotate sequence variation w.r. to the reference

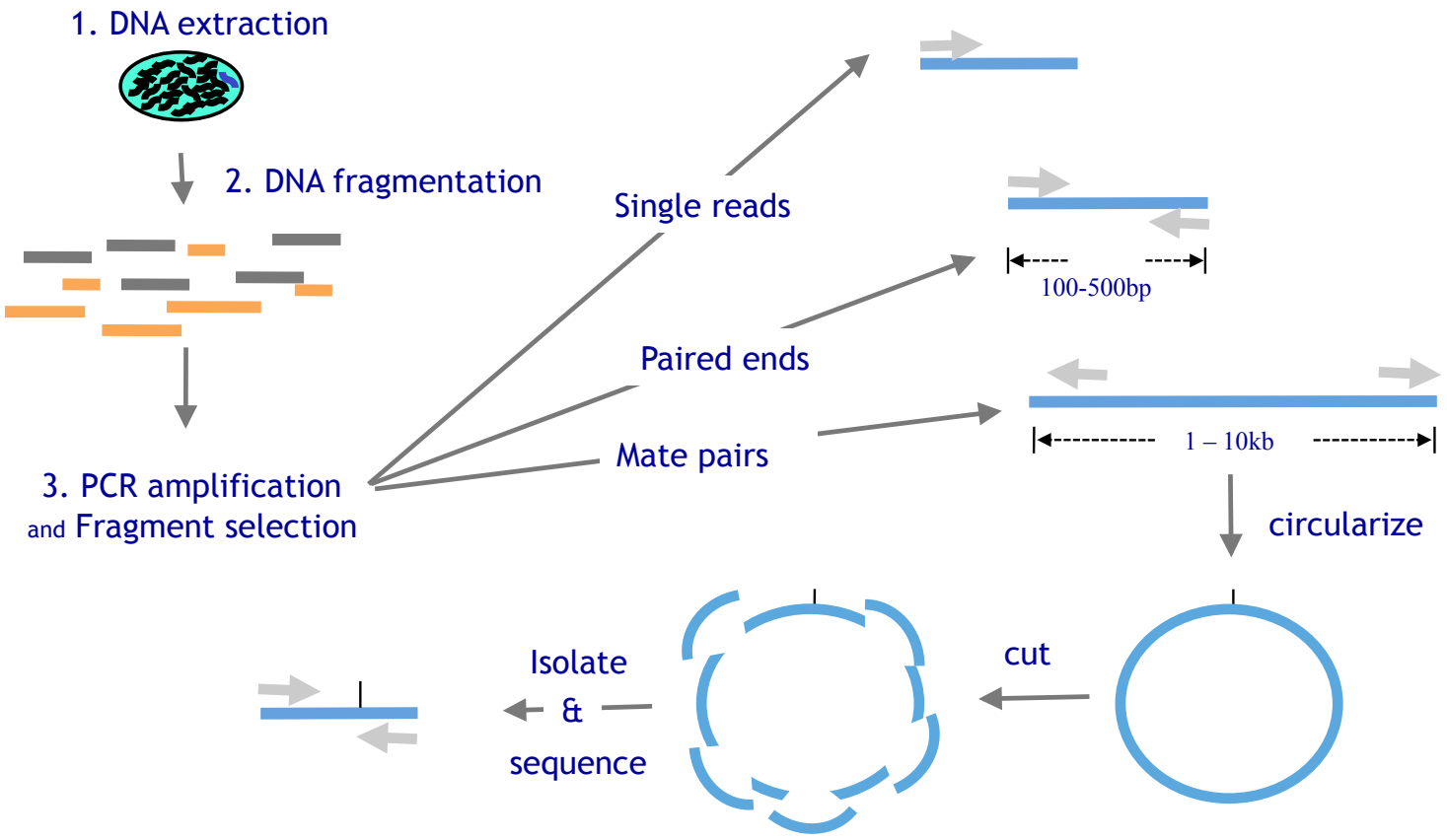
The many applications of NGS

- Genomic
  - **(re)sequencing**
  - metagenomic (tutorial)
  - targeted sequencing
- Functional
  - Protein-DNA (chIP-Seq)
  - **mRNA (RNA-seq, CAGE)**
  - small RNA (µRNA-Seq)
  - methylome (meDIP-Seq)
  - Chromosomal structure (Hi-Seq, 3C)

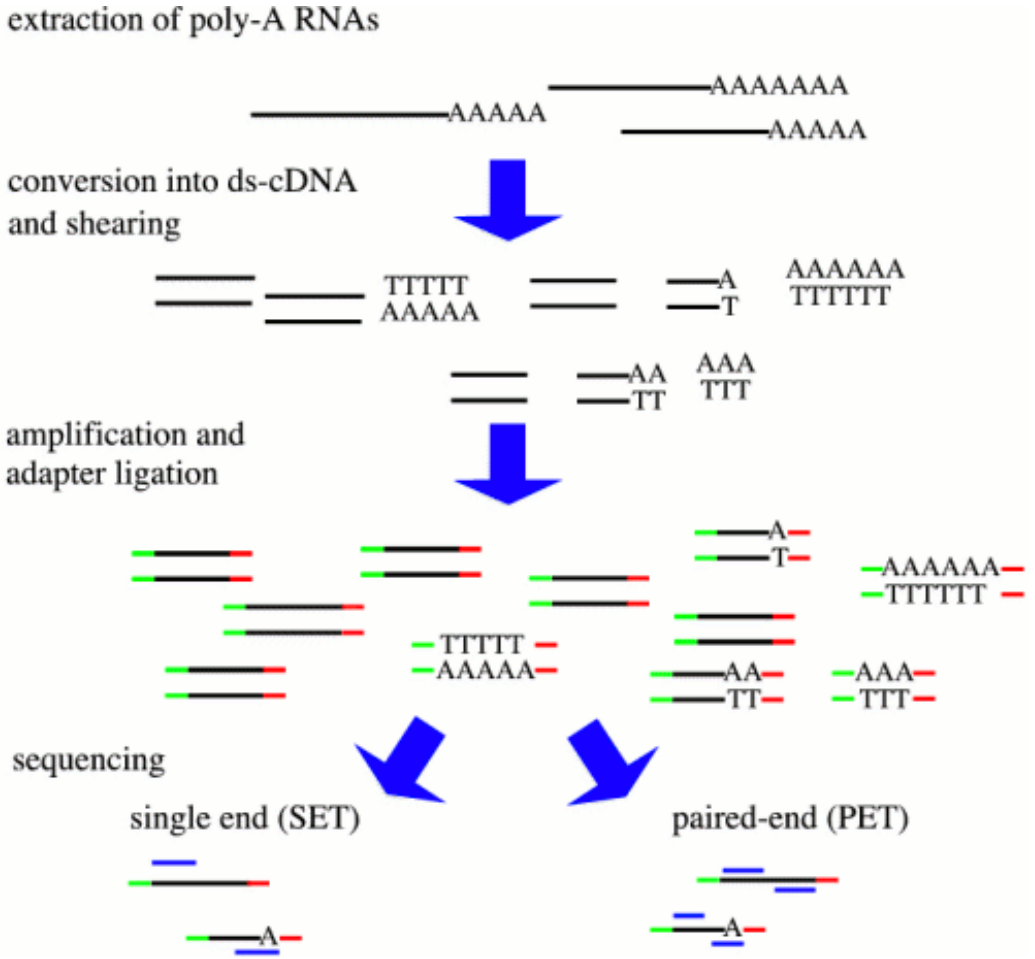


The sequences have to be aligned or assembled

Genome sequencing



RNA-Seq protocol

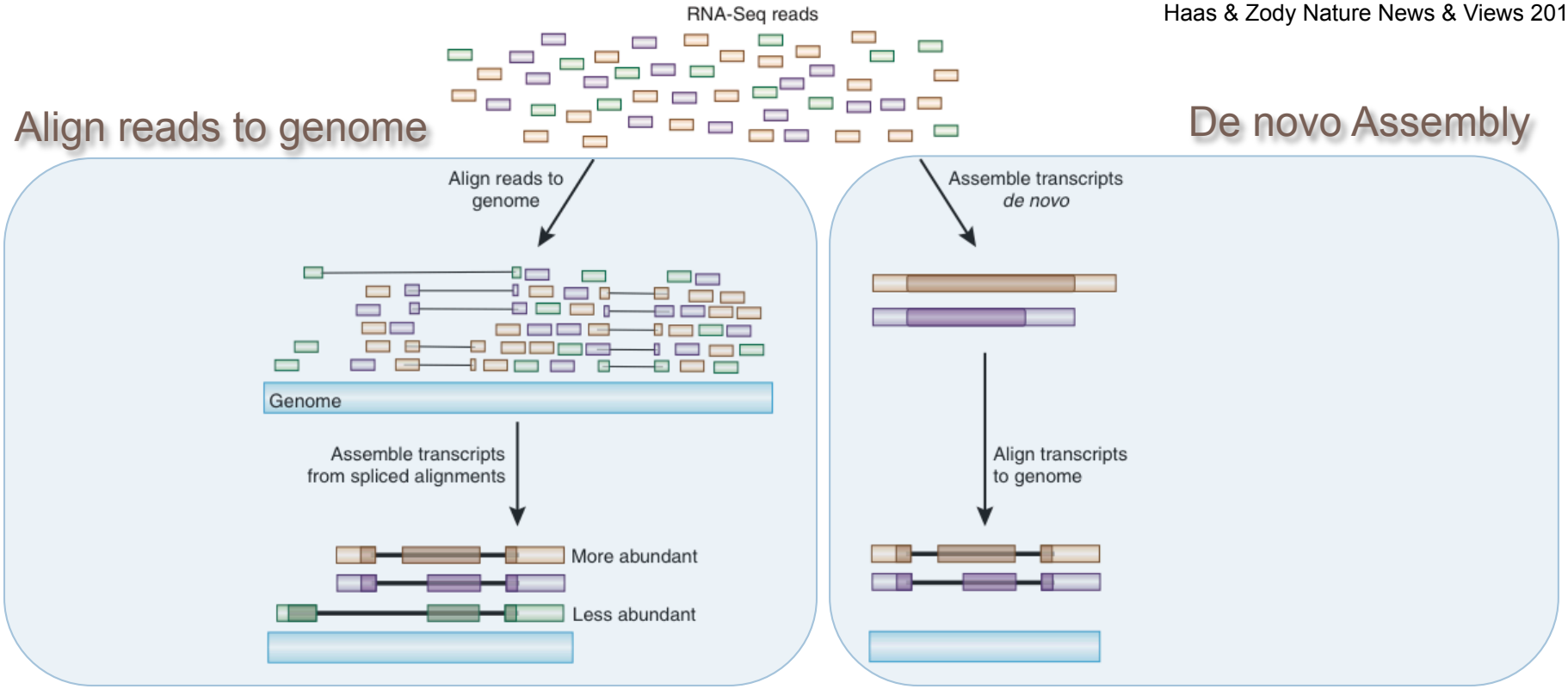


Whole Transcriptome Shotgun Sequencing

M. Schulz©

From reads to transcripts

Haas & Zody Nature News & Views 2010



- (Re)-annotation of gene models (transcripts)
- Transcripts quantification
- Transcriptional profiles

3C-based technologies

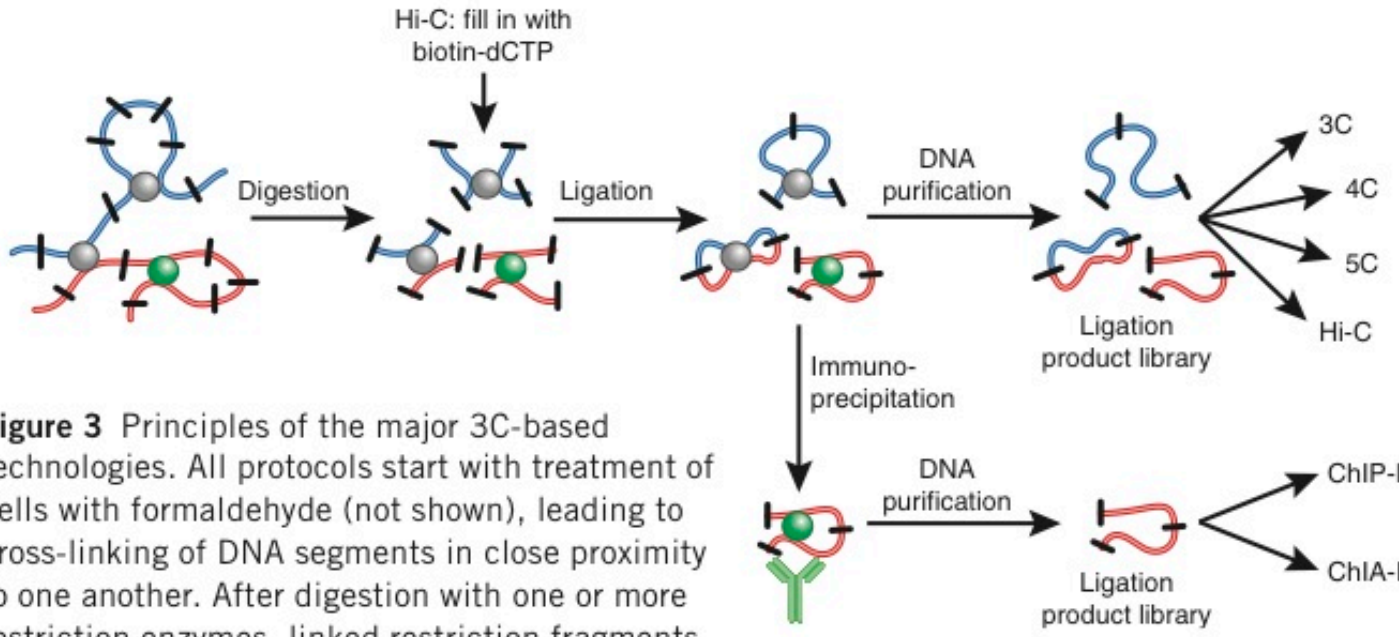


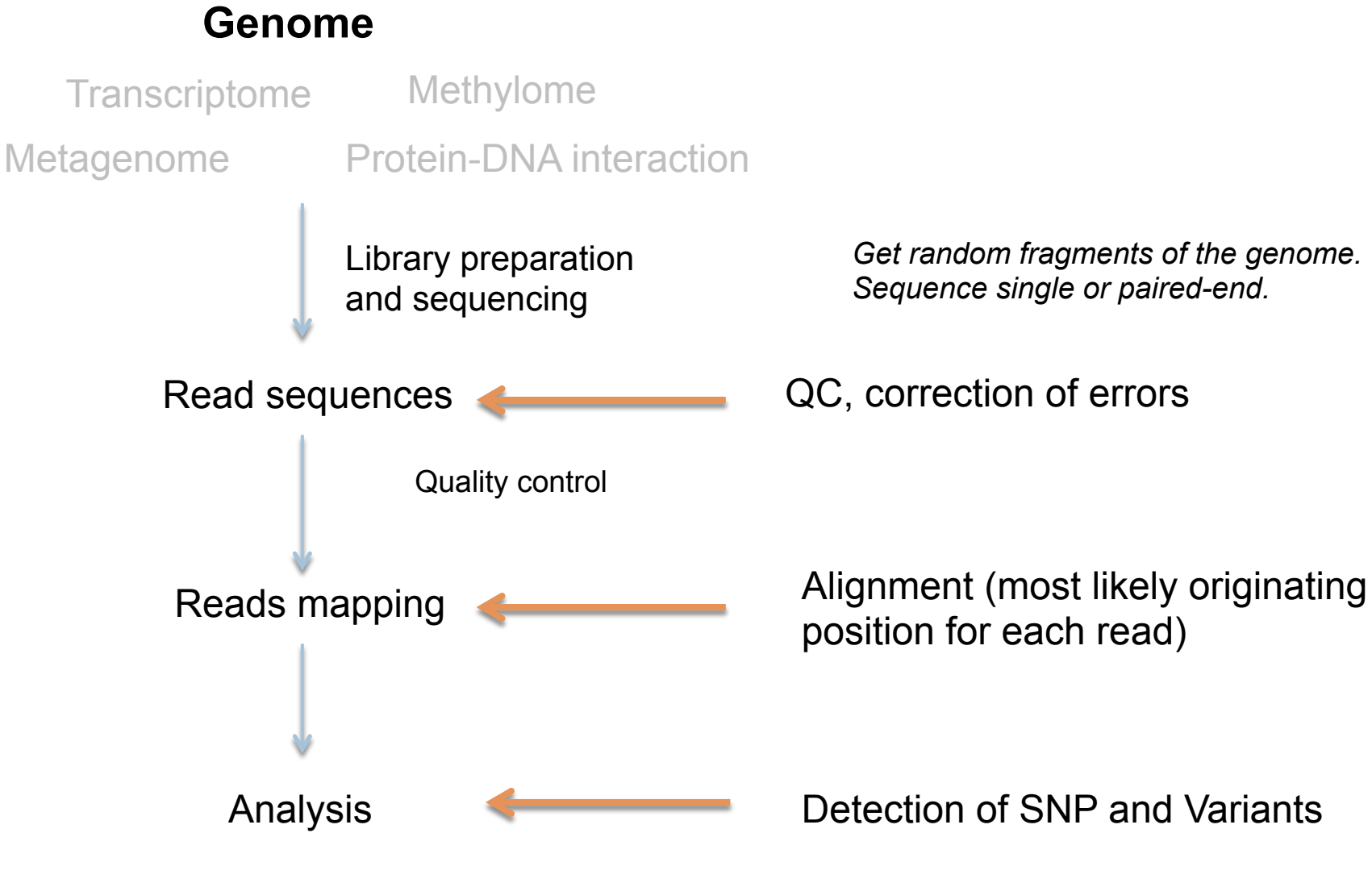
Figure 3 Principles of the major 3C-based technologies. All protocols start with treatment of cells with formaldehyde (not shown), leading to cross-linking of DNA segments in close proximity to one another. After digestion with one or more restriction enzymes, linked restriction fragments are intramolecularly ligated. In the case of Hi-C, the ends of the restriction fragments are first filled in with biotinylated dNTPs before ligation to facilitate purification of ligation junctions using streptavidin-coated beads. Single or multiple ligation events are detected directly (using 3C, 4C, 5C and Hi-C), or immunoprecipitation is first used to enrich for DNA associated with a protein of interest (using ChIP-loop and ChIA-PET). See **Table 2** for overview of different detection strategies and their scope.

Outline

- Brief historical background on sequencing

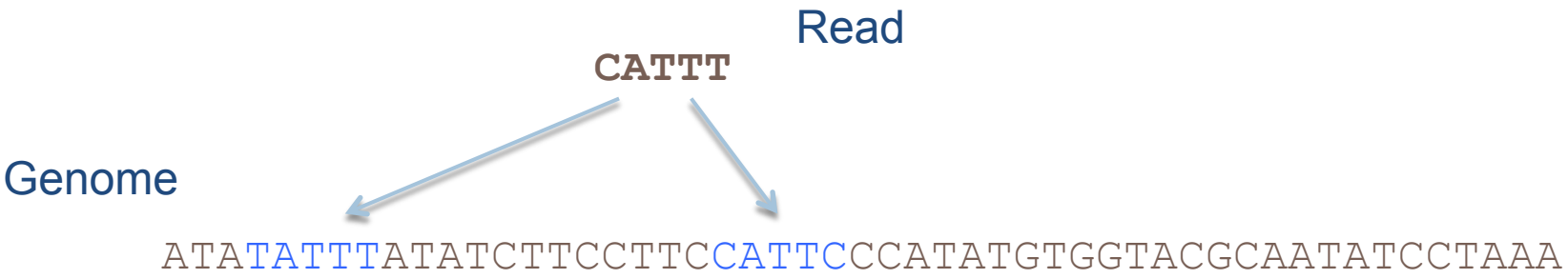
- **A typical resequencing workflow**
  - Reads alignment
  - Basic statistics after alignment
  - Automatic correction of sequencing errors
  - Detection of sequence polymorphisms
  - Genome Assembly

NGS workflow



Question 1: read alignment

- We sequenced a read. Where should we place it on the genome ?



- Genome is available
  - Recall pairwise alignment ?
  - Has to be done in short time.

Time is money

- Evaluating by dynamic programming would cost k.n operations for each read.
  - Nowadays need to align ~500Mio reads / exp.
- Need to do cost saving cuts by either:
  - Index the genome with an appropriate data structure
    - Hash table of length k
    - Suffix tree of the genome
  - For each read, filter the set of possible starting positions with the data structure.

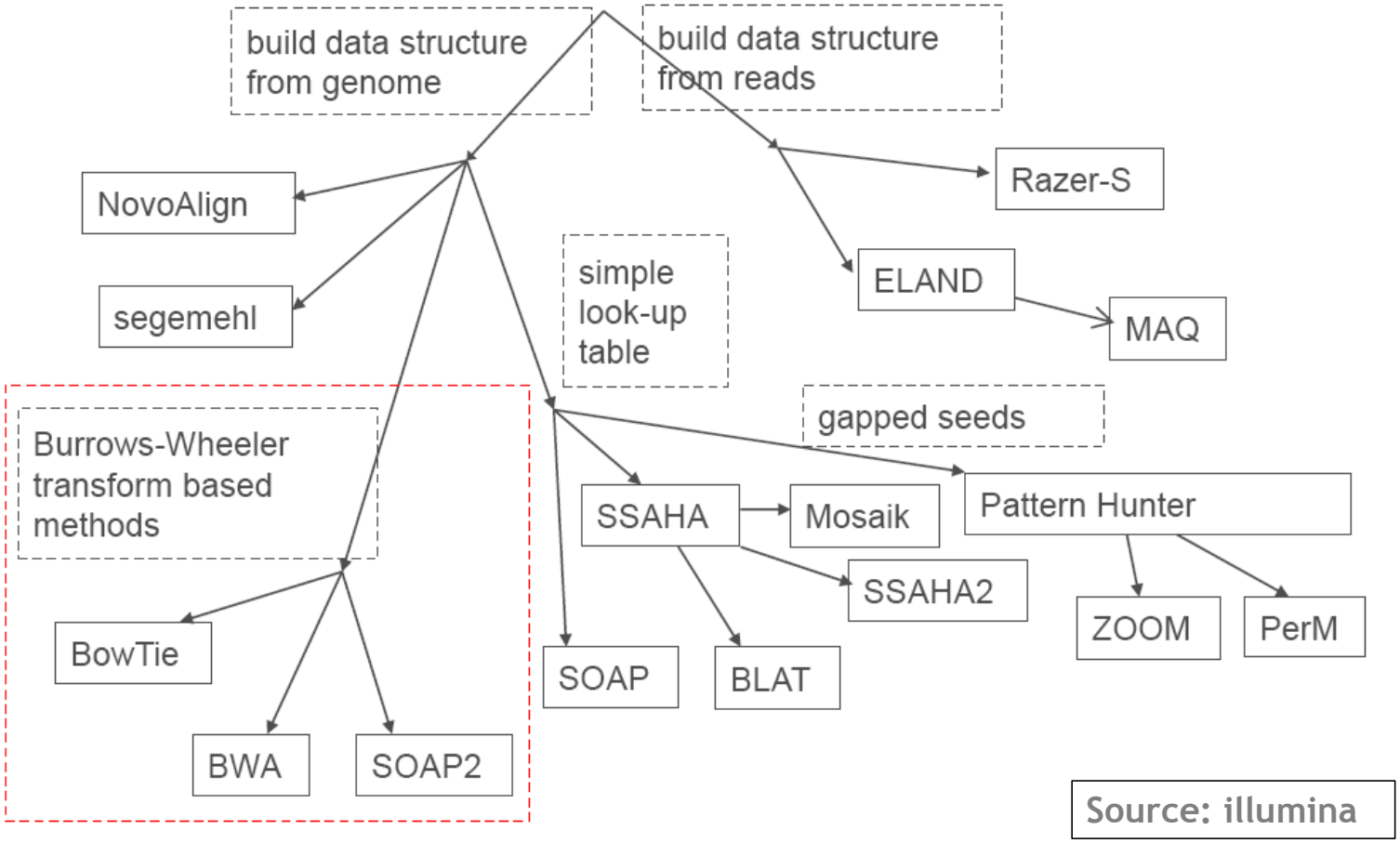
High throughput read alignment

- Usually semi-global alignment
  - Sometimes allow soft trimming at one/both end
- Many aligners exist now.

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

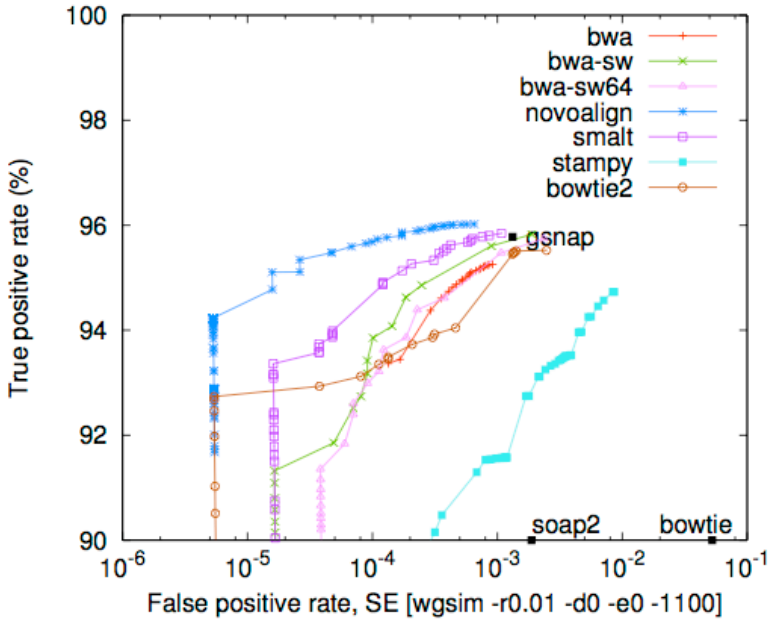
[a]Work well for Sanger and 454 reads, allowing gaps and clipping. [b]Paired end mapping. [c]Make use of base quality in alignment. [d]BWA trims the primer base and the first color for a color read. [e]Long-read alignment implemented in the BWA-SW module. [f]MAQ only does gapped alignment for Illumina paired-end reads. [g]Free executable for non-profit projects only.

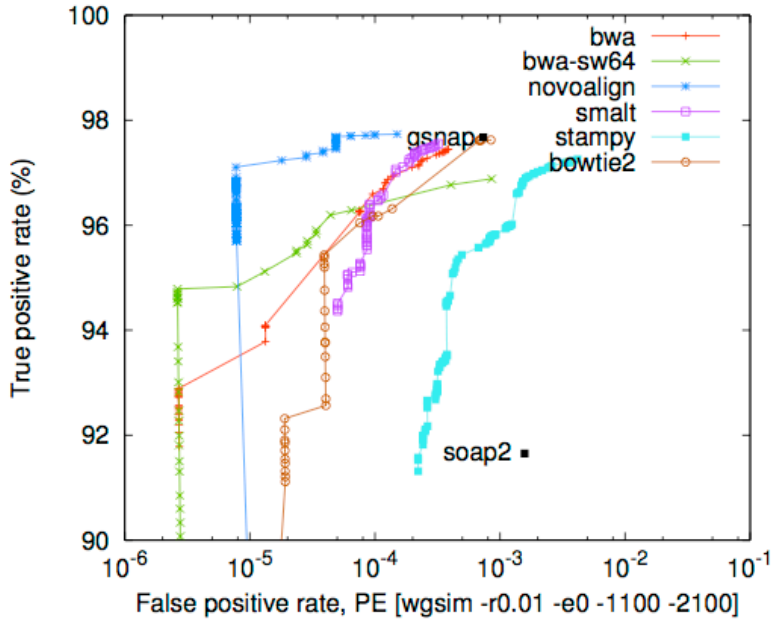
Read mapping tools



Last generation of mappers combine genome and read indexing (Massai, OLego)

Comparison read mappers

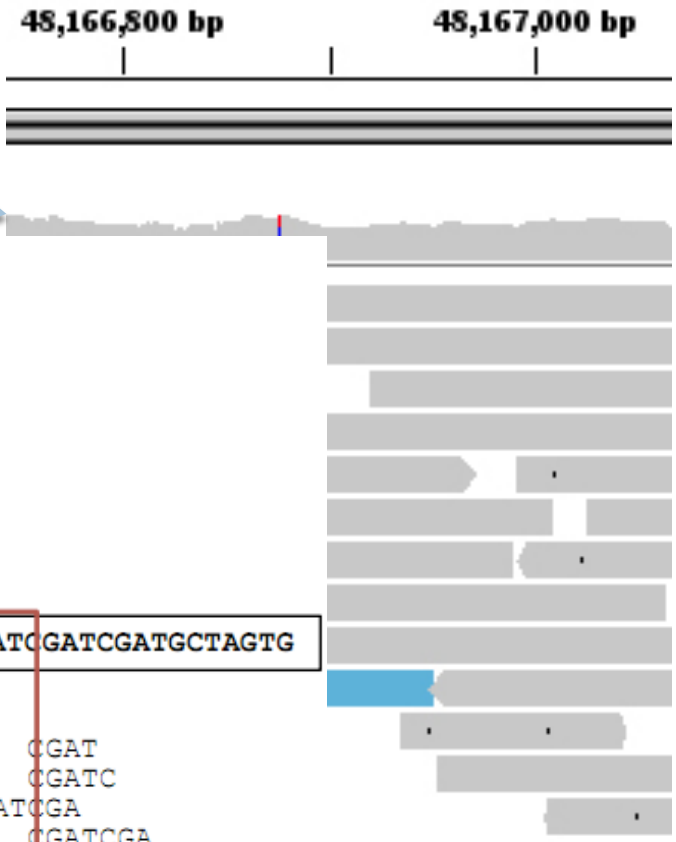




from Heng Li (BWA) benchmark page: http://lh3lh3.users.sourceforge.net/alnROC.shtml
100k reads simulated with wgsim **without** sequencing errors

An alignment result

Sequence coverage
(# of reads covering a position)



SNPs

INDELs

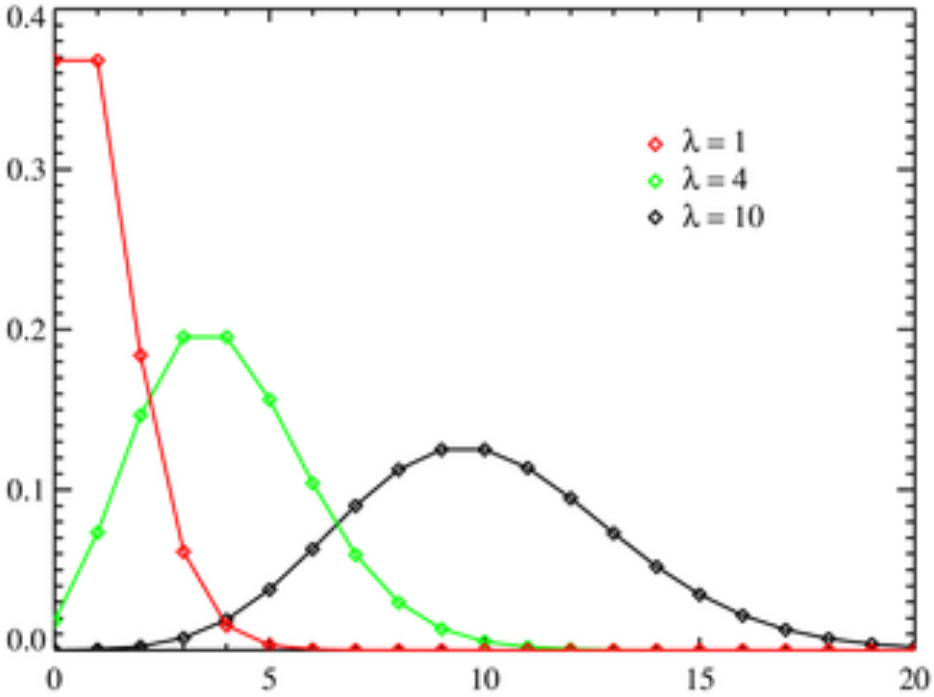
Question 2: Sequence Coverage ?

- Every read sample a position along the genome.
  - Coverage: Number of reads overlapping a position
  - What is the expected distribution ?

- Suppose uniform sampling.

Poisson Distribution

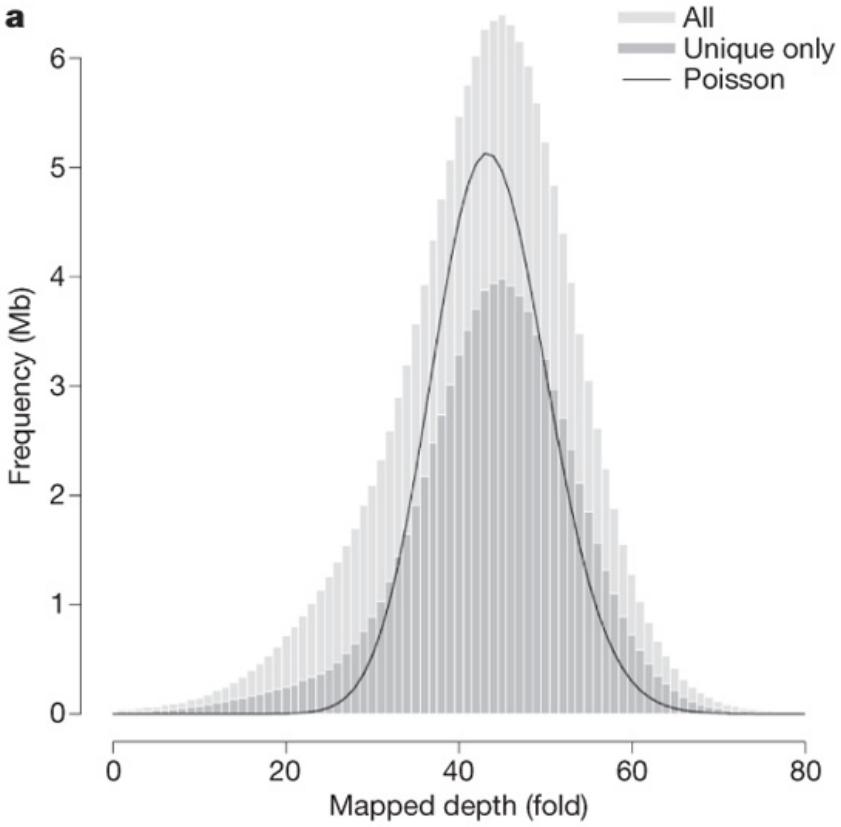
- Only one parameter, the expected coverage λ
- Variance and expectation are equal

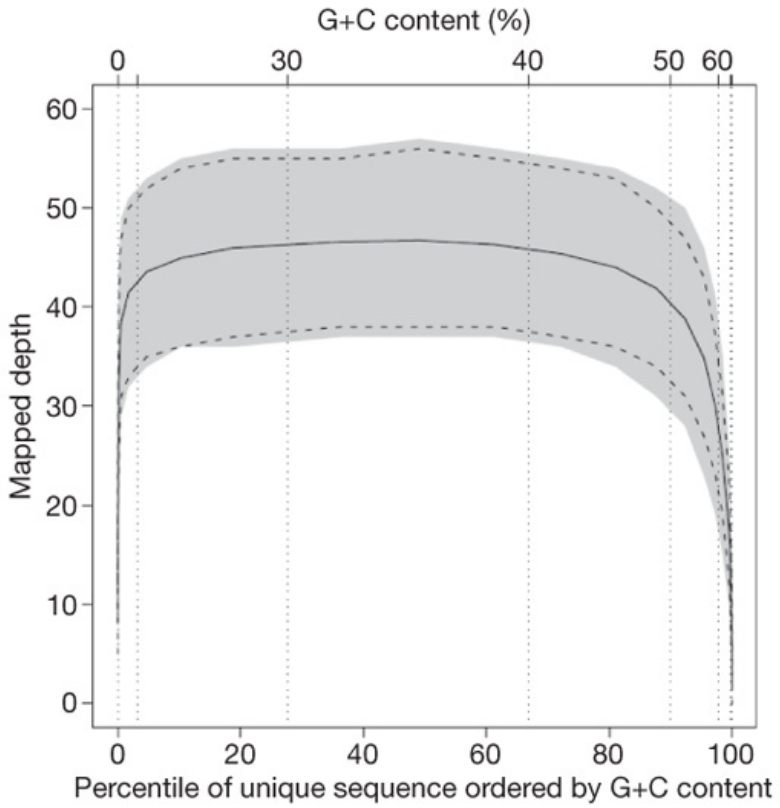


Probability distribution function for various parameters of λ

In practice

- Poisson assumes uniform sampling, but multiple artifacts in the data (also: PCR duplicates)





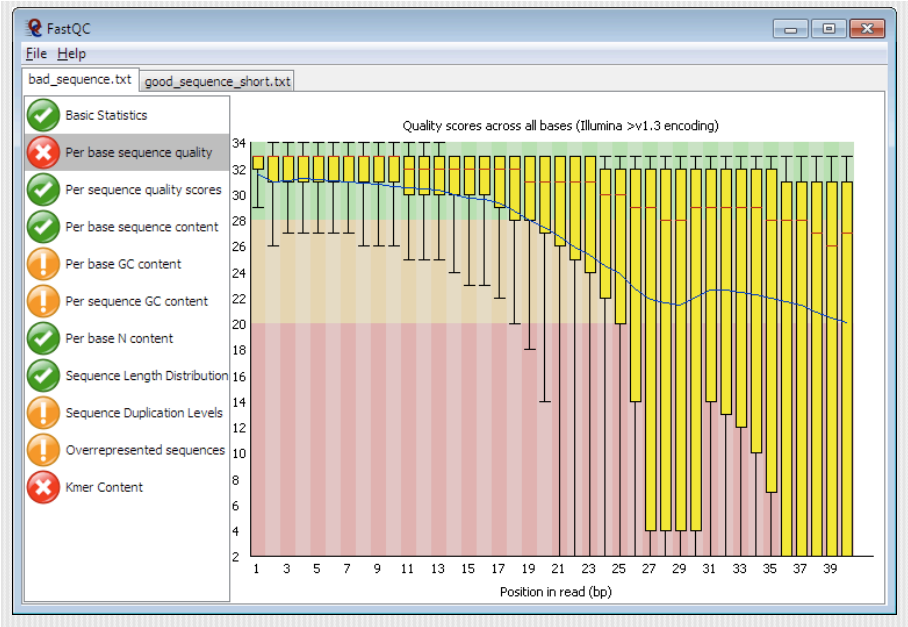
Bentley et al. Nature 2008

Quality evaluation

- Sequence Quality score
  - reported by the sequencer

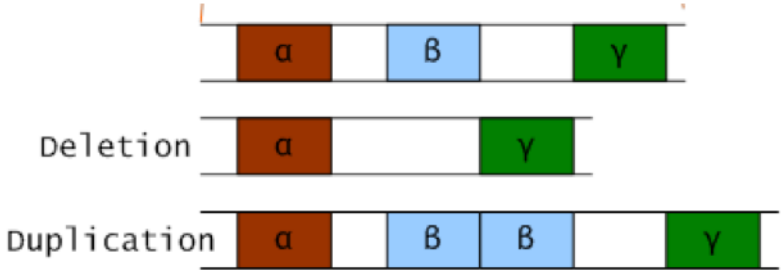
Phred score: $-10.\log_{10}(P_{err})$

- % mapped reads
- coverage distribution



Variant Calling with a reference

- Single Nucleotide Variants (SNV)
  - 1bp difference to the reference
- Short insertions/deletions (indels)
  - <50 bp variation to the reference
- Structural variants



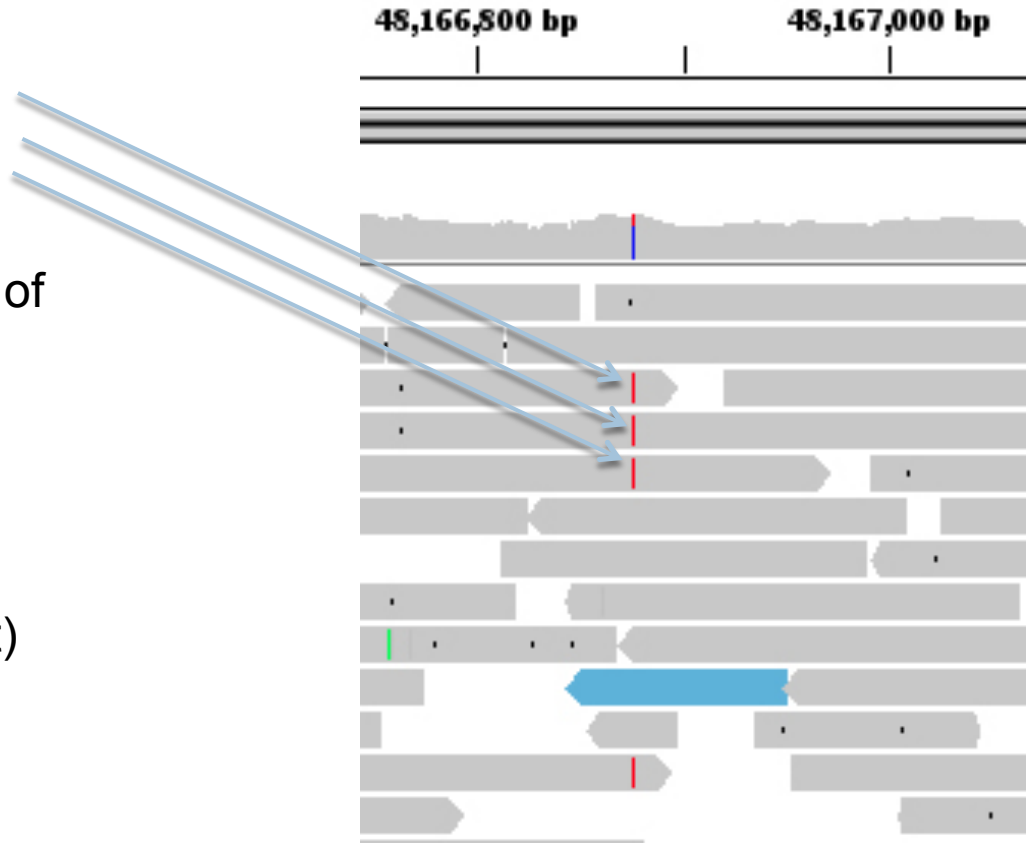
Detection of SNV

Are these genuine or due to sequencing errors ?

Hypothesis testing on the number of reads supporting the SNV vs expected error rate.

Tools:

- GATK (Genome analysis toolkit)
- SAMtools mpileup
- VarScan

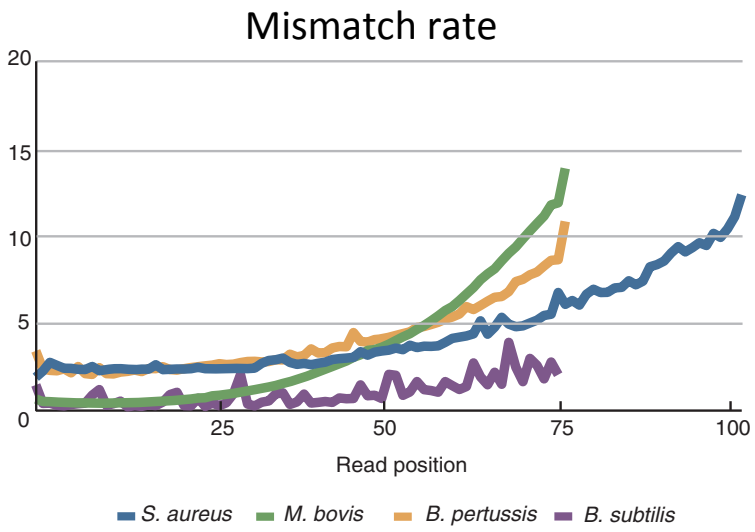


SNV annotation

- Need to be annotated w.r. to genomic information
  - Homo/heterozygous
  - coding/non-coding SNVs
  - Known/unknown (dbSNP, 1k genome)
- Functional implication of a SNV
  - Silent/Non-silent mutation (within CDS)
  - Predicting possible impact of an amino acid substitution on the protein stability (Sift, Polyphen2…)

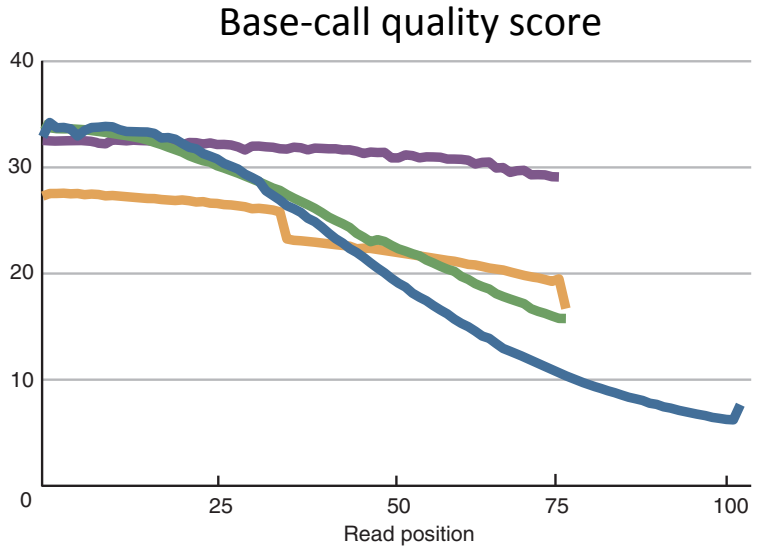
Illumina error profile

- Biases found by Dohm *et al.* (2008):
  - **Mismatches** are dominant errors
  - Positional error error profile with an increase towards the 3' end
  - non-uniform substitution matrix
    - 8 times more A→C errors compared to A→G or A→T errors



- Base-call qualities give a hint about errors (Nakamura et al., 2011)

$$\text{Phred score: } -10.\log_{10}(P_{err})$$



Errors in other Technologies

- Other platforms are dominated by indel errors due to:
  - Indistinguishable signals for homopolymers (>6bp)
  - Undetected low-intensity signals (single molecule sequencing)

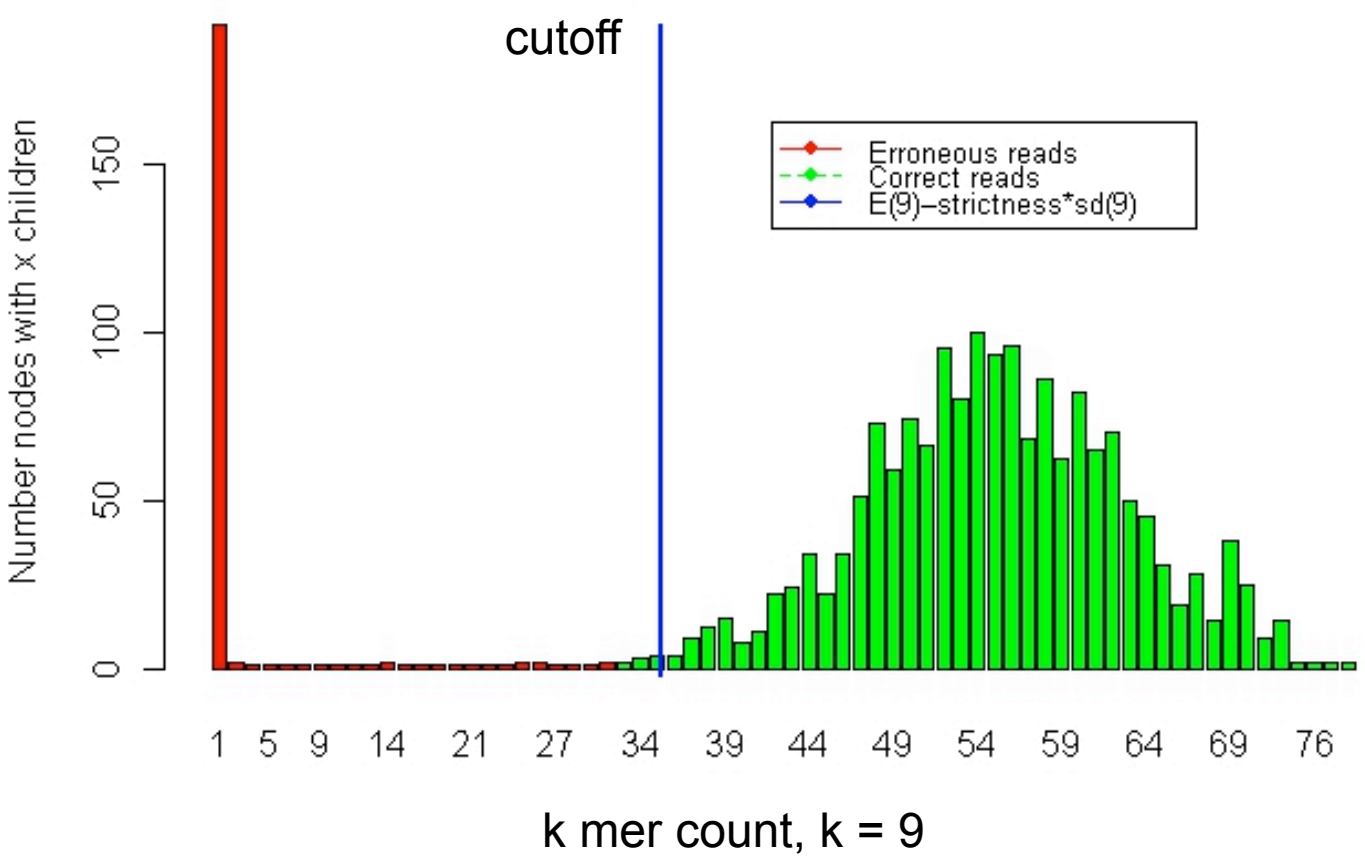
Table 1: Characteristics of major NGS platforms as of February 2012

Company	Platform	Read length (bp)	Throughput & time per run	Technique	Dominant error type
Illumina	HiSeq 2000[a]	36 50 100	105–600 Gb 2–11 days	Reversible terminator	Substitution
Applied Biosystems	5500 SOLiD™ System[b]	356 075	7–9 Gb/day	Sequencing by ligation	–
Complete Genomics		35	–	Ligation based	–
Helicos BioSciences	HeliScope SMS[c]	25–55	21–35 Gb	**Single molecule sequencing**	Insertion Deletion
454 Life Sciences	GS FLX Titanium XL+[d]	≤1000	700 Mb 23 h	Sequencing by **synthesis**	Insertion Deletion
Ion Torrent	Ion PGM Sequencer 318[e]	>200	>1 Gb 2 h	Ion semiconductor sequencing	Insertion Deletion
Pacific Biosciences	PacBio RS	1 k–10 k	–	**Single molecule sequencing**	Insertion Deletion

(Yang *et al.*, 2012)

Detecting Potential Errors

- reads with errors will have a low coverage.



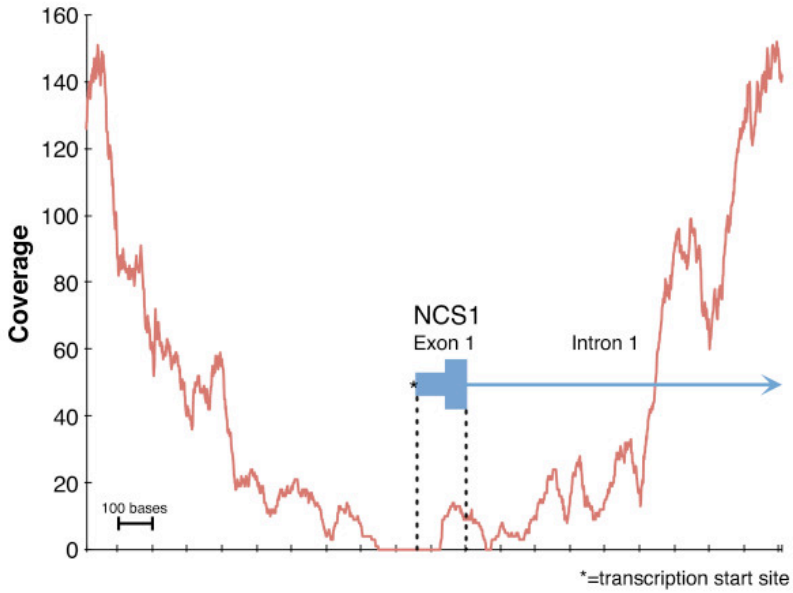
- Automatic correction of sequencing errors without alignment to the genome

  → cf other slides

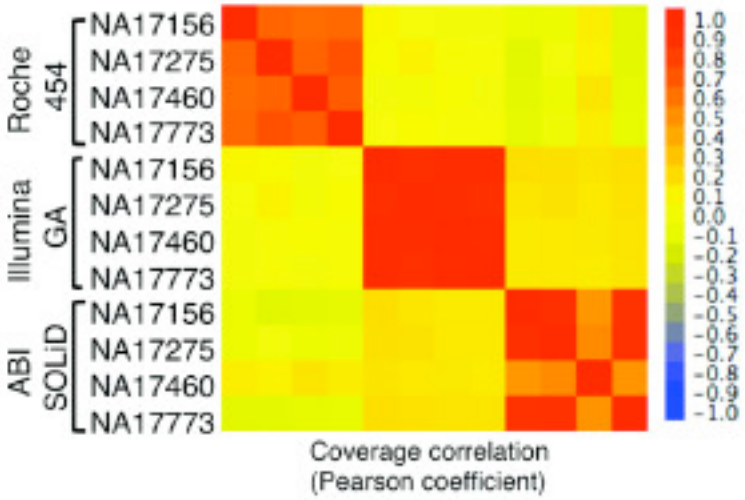
Performance of Genome resequencing

- Depends on artifacts/biases:
  - Each technology/preparation has its own
- Two main artifacts:
  - Sequencing bias (quality score, error correction)
  - Coverage bias
    - Can be related to GC composition
    - Some regions with consistent lower coverage cannot be genotyped
- Assess by comparing multiple technologies with Sanger sequencing

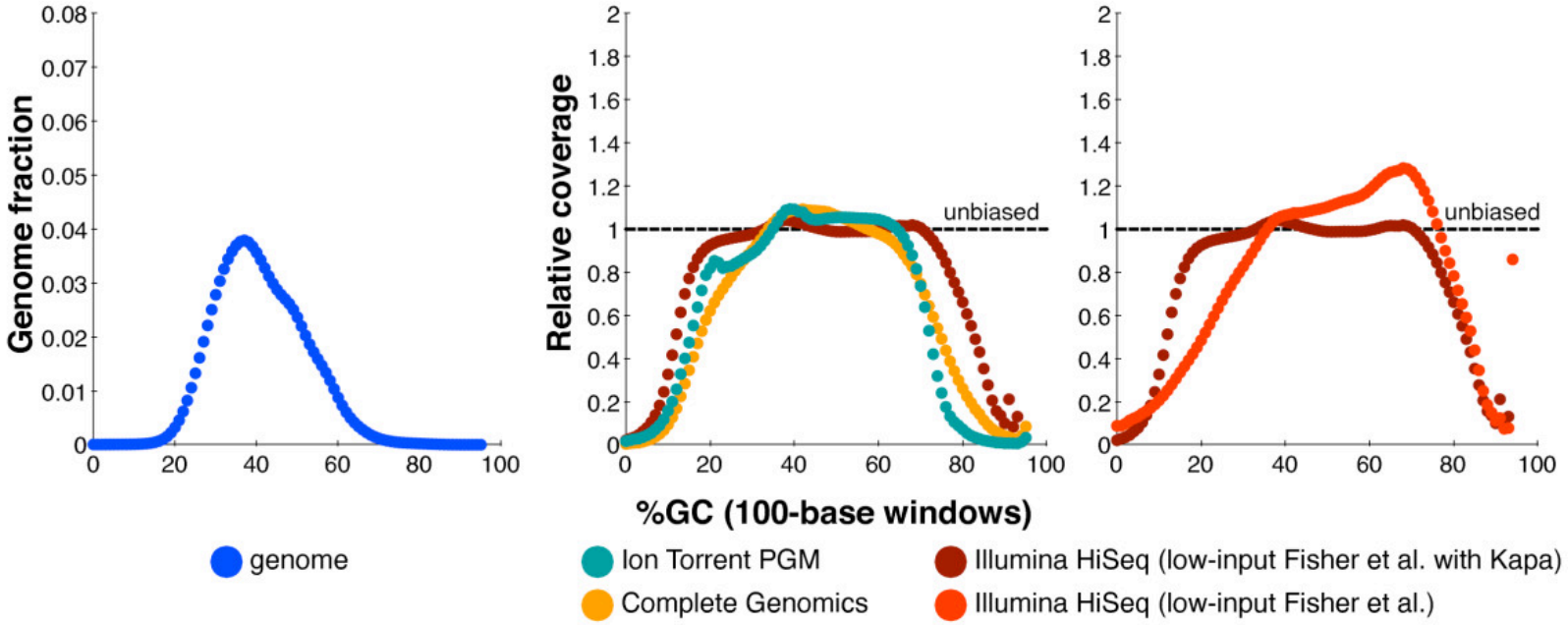
Ross et al. Genome biology 2013



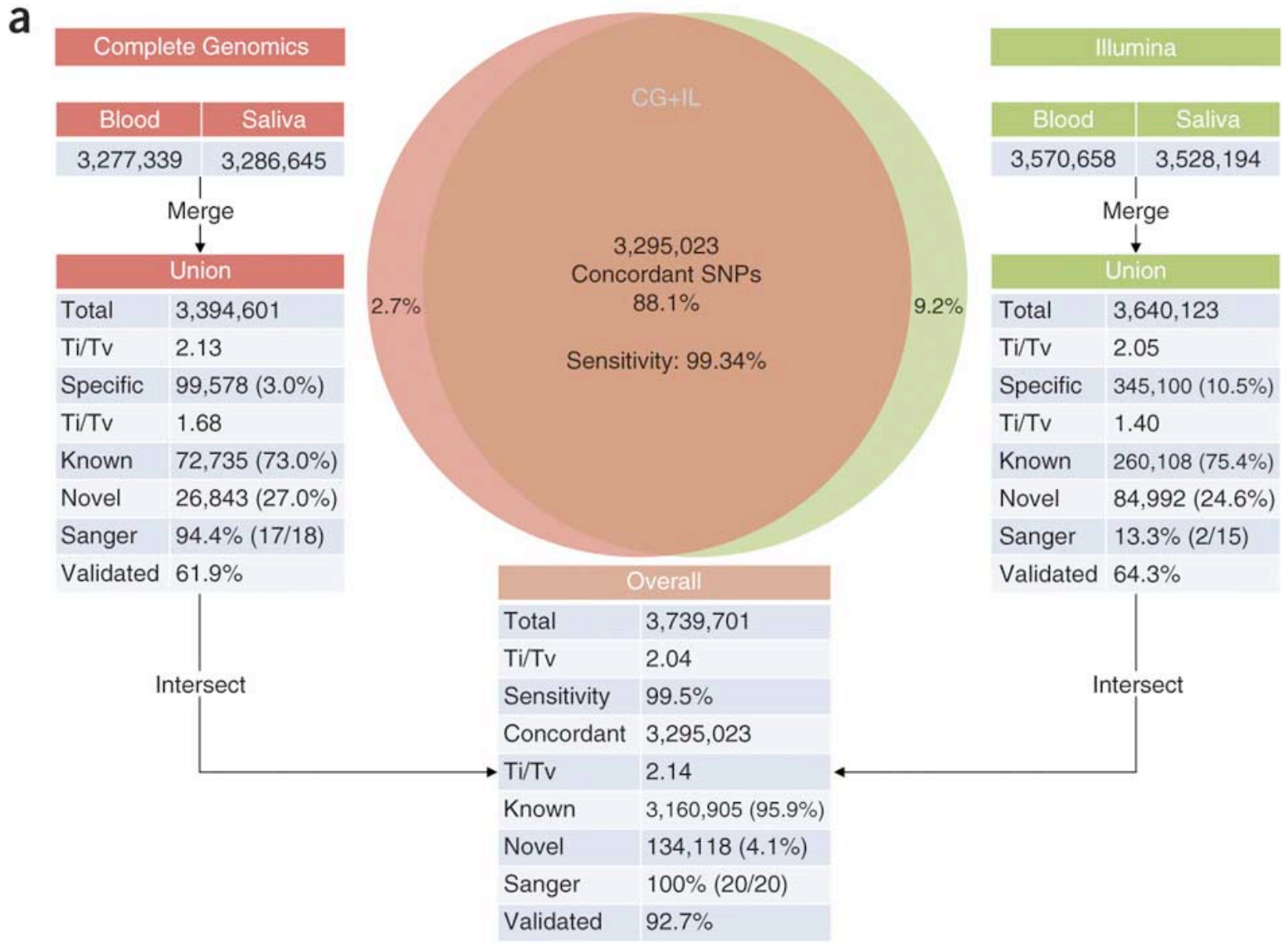
Harismendy et al. Genome biology 2009



- Some region will never be covered enough for SNV calling
  - Frequently linked to high/low GC
- Coverage tend to be correlated with technology (here targeted resequencing)

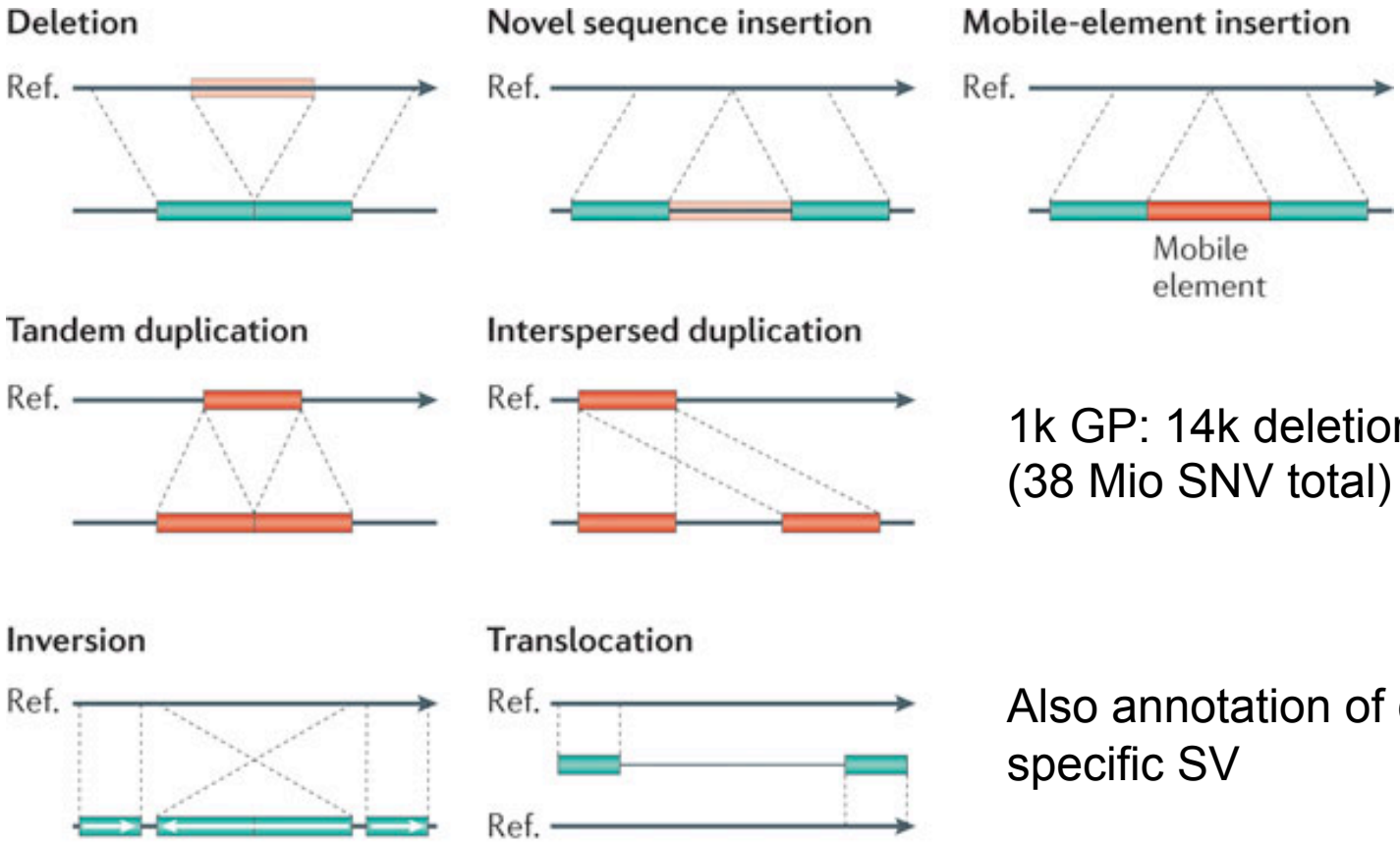


Ross et al. Genome biology 2013



Lam et al. Nature Biotechnology 2012

Annotation of structural variants (SV)



1k GP: 14k deletions
(38 Mio SNV total)

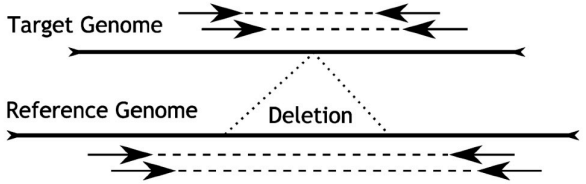
Also annotation of cancer specific SV

Nature Reviews | Genetics

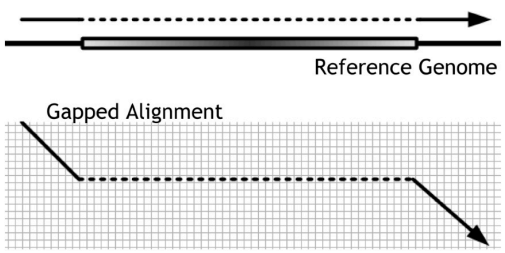
Alkan et al 2011

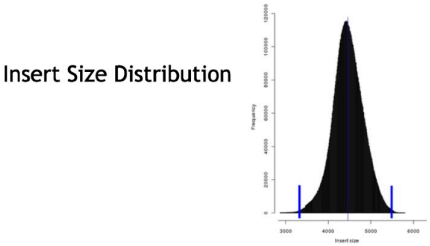
SV detection methods

- Different information can be used (or merged)
  - Paired-end Mapping

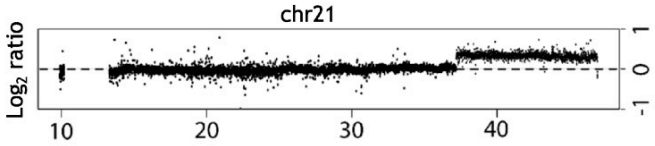


  - Split-read alignment





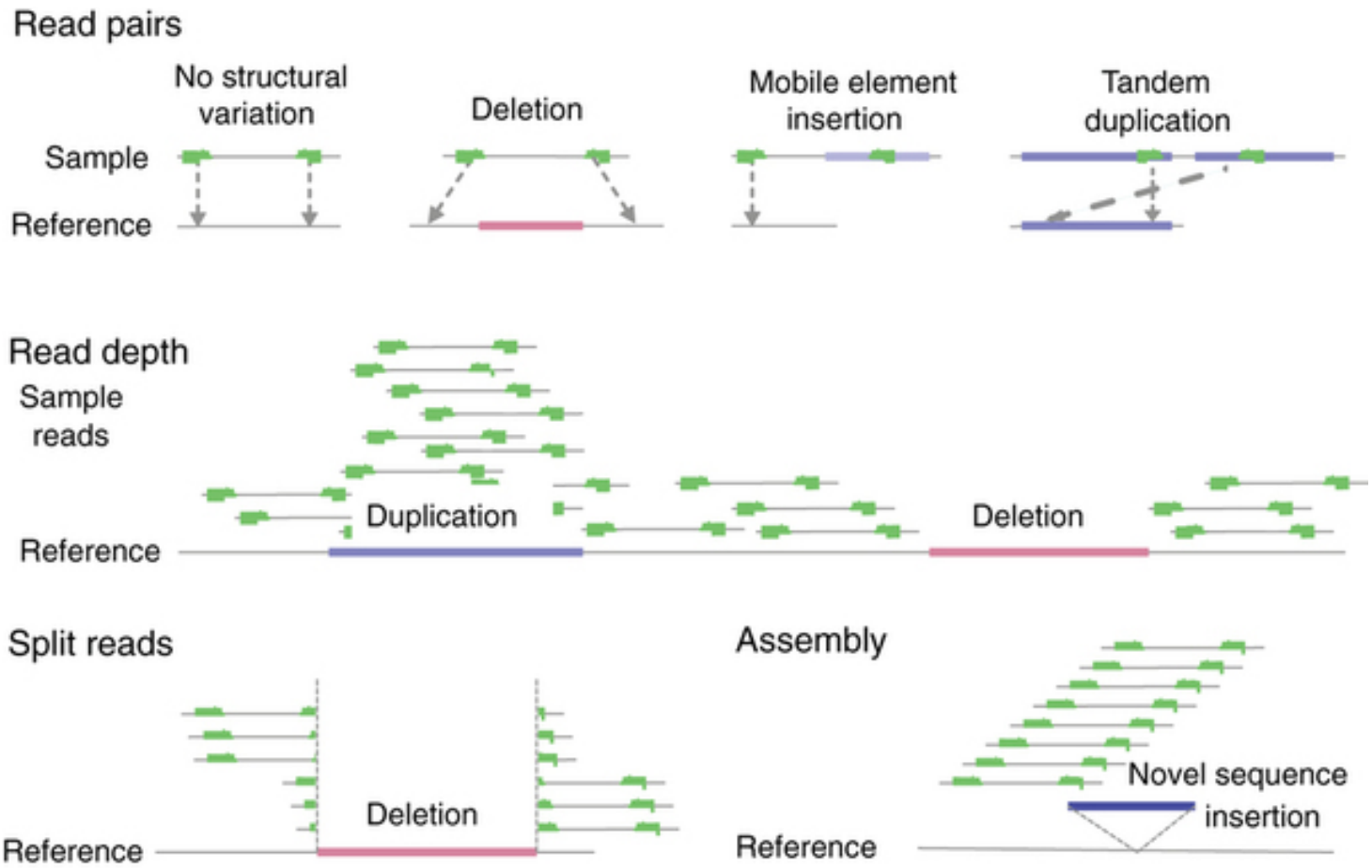
  - Read depth analysis

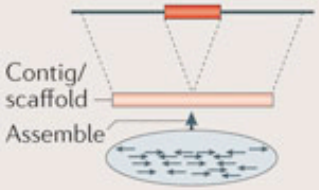
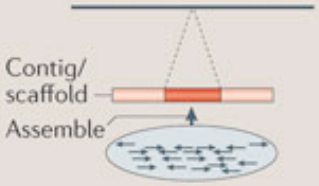
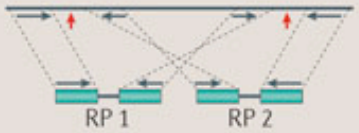
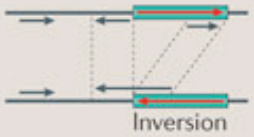
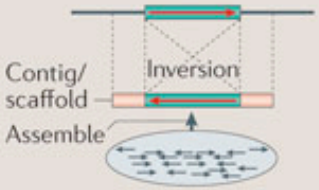
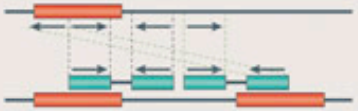
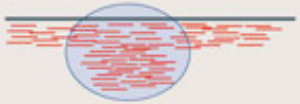
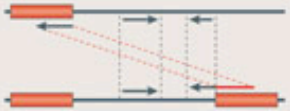
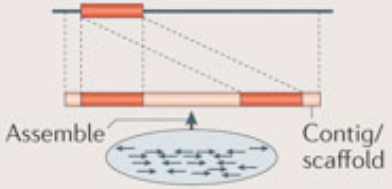
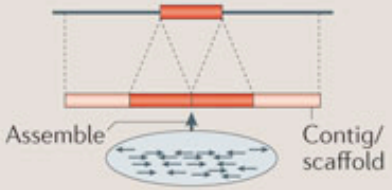


$$\log_2 \frac{\#Reads_{Diseased}}{\#Reads_{Normal}}$$

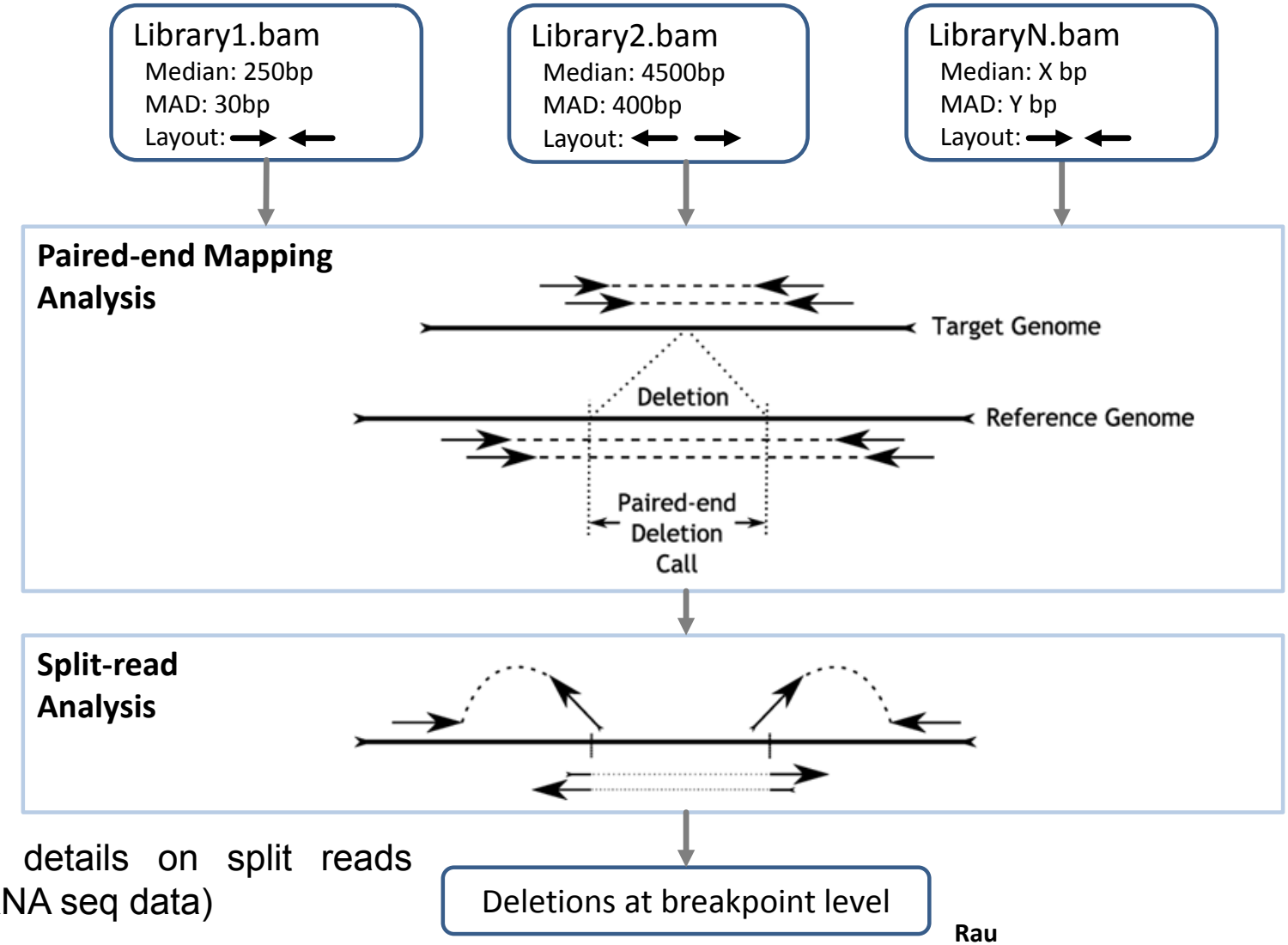
  - Assembly and mapping

Detection of variants (with Paired end)



SV classes	Read pair	Read depth	Split read	Assembly
Deletion				
Novel sequence insertion		Not applicable		
Inversion		Not applicable		
Interspersed duplication				
Tandem duplication				

Delly: Combine PE and split read analysis



(More details on split reads with RNA seq data)

Rau

Rausch et al. Bioinformatics 2012.

Outline

- Brief historical background on sequencing

- A typical resequencing workflow
  - Reads alignment
  - Basic statistics after alignment
  - Automatic correction of sequencing errors
  - Detection of sequence polymorphisms (A. Gillet)
  - **Genome Assembly**

Sequence Assembly

- Assemble:
  - reconstruct the large sequence (the genome) by merging and ordering the reads from the sequencing experiments
  - need to summarize overlap between reads in a data structure to reconstruct contiguous sequences (*contigs*)
- Assembly is:
  - trivial if repeats are shorter than read length
  - computationally intractable with too short reads

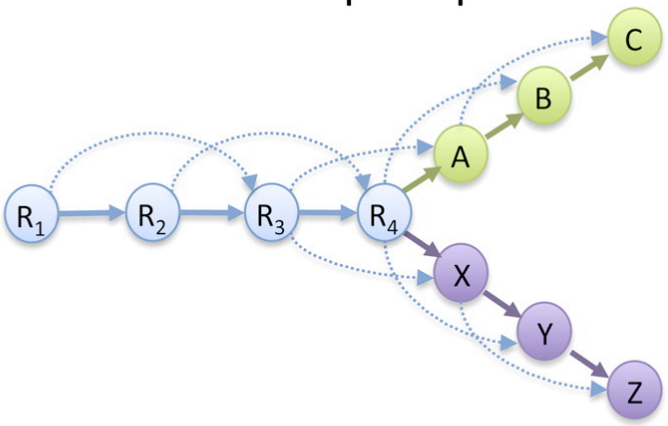
(sequencing errors play a role)

Assembly paradigms

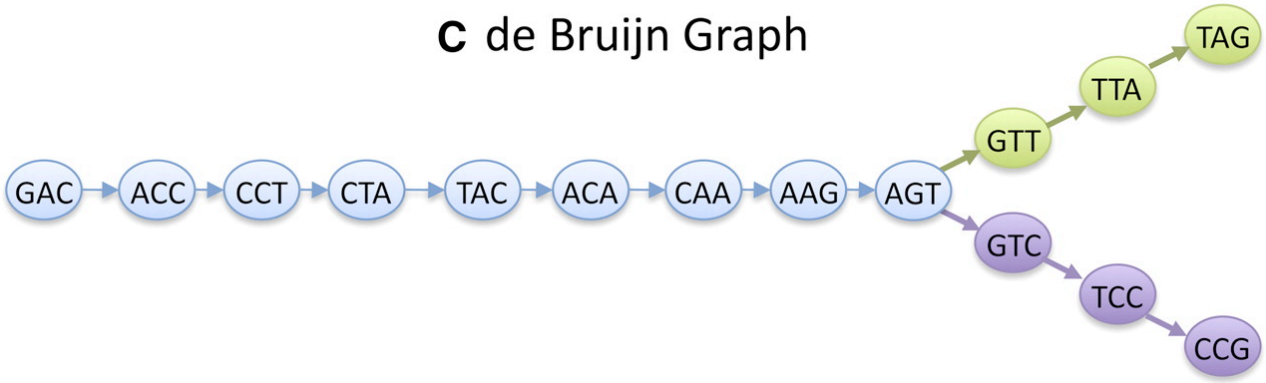
A Read Layout

```
R1: GACCTACA
R2:   ACCTACAA
R3:    CCTACAAG
R4:     CTACAAGT
A:       TACAAGTT
B:        ACAAGTTA
C:         CAAGTTAG
X:       TACAAGTC
Y:        ACAAGTCC
Z:         CAAGTCCG
```

B Overlap Graph



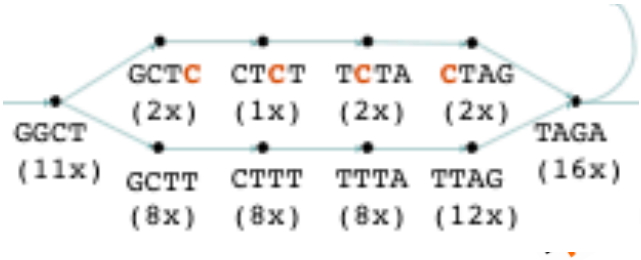
C de Bruijn Graph



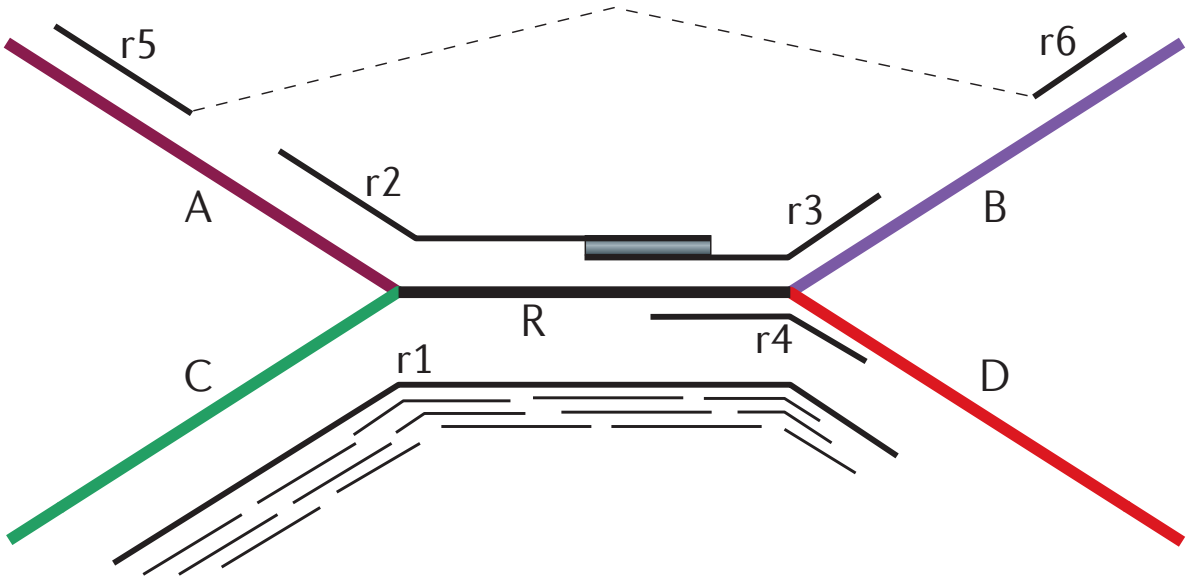
Genome assembly paradigms

- Overlap Layout consensus and string graph
  - repeats are “perfectly” summarized
  - Computationally costly to do all pairwise alignments

    (optimized with FM-index/burrows wheeler, SGA assembler recently)
- De Bruijn Graph on the set of sequences
  - easy to do a hash map of all reads
  - repeats needs to be transitively reduced
  - errors impact the graph structure



Effect of sequencing parameters

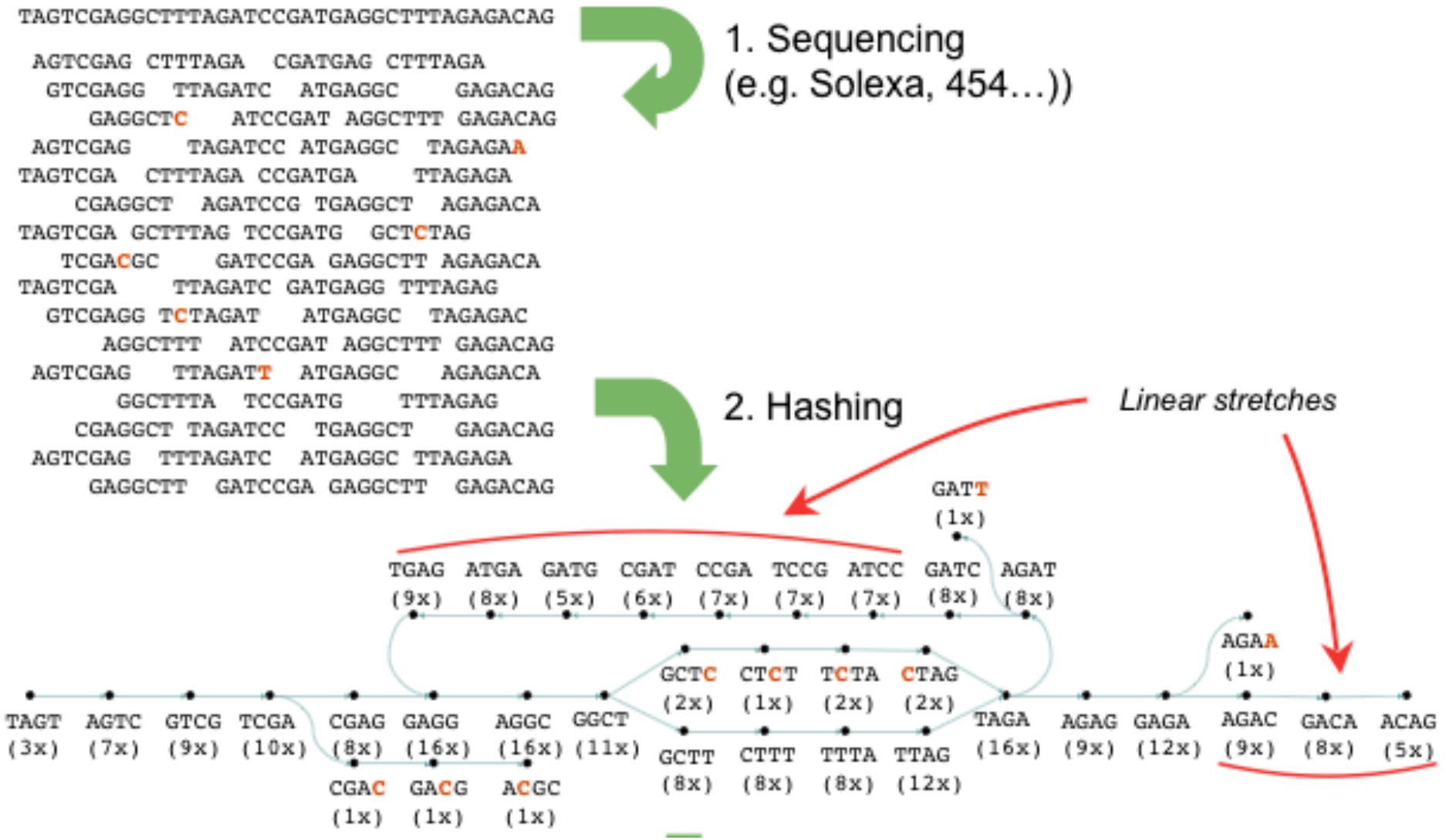


- Possible reconstructions: ARB+CRD or ARD+CRB
- some reads are longer than the repeat region R

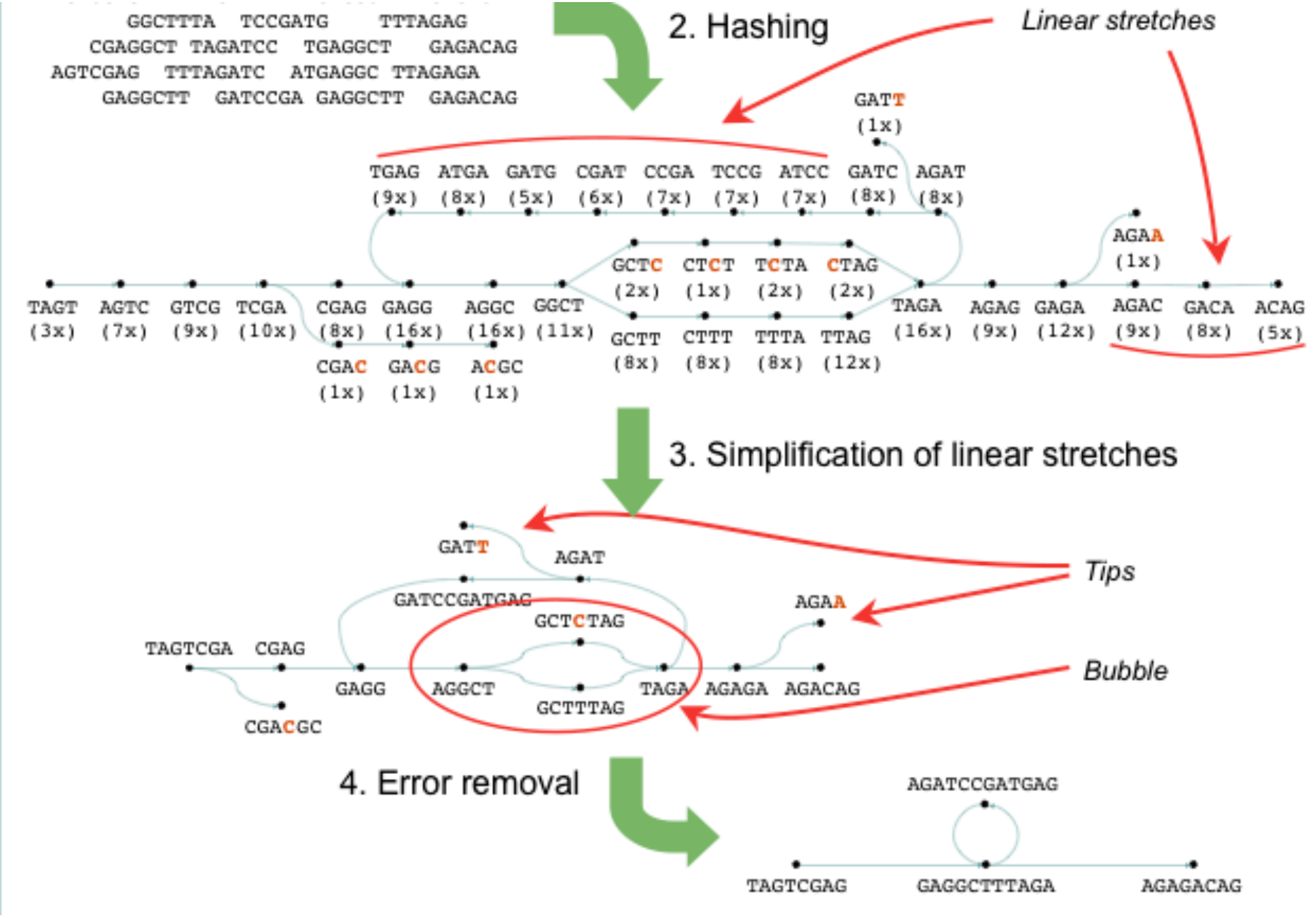
Nagarajan & Pop, Nature Review Genetics, 2013

Difficulties with assembly

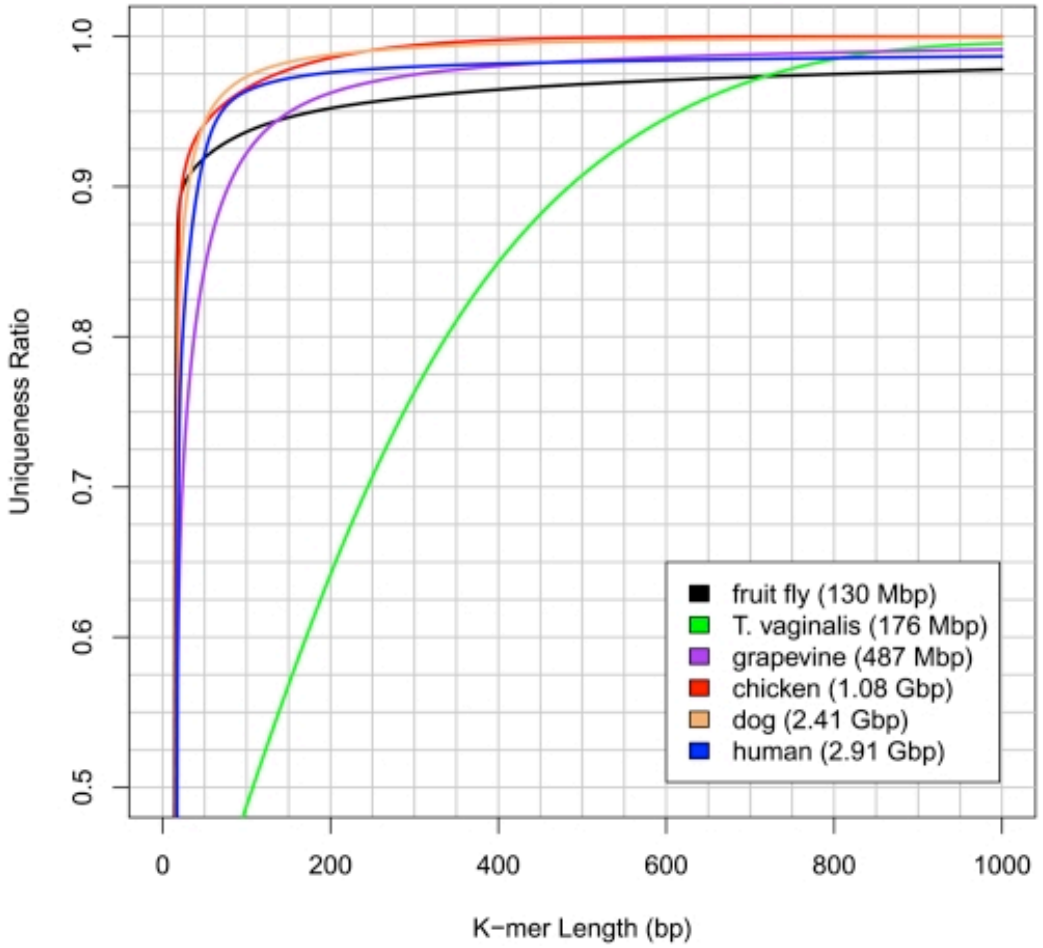
- **Repeats** create cycles/branching
  - need longer reads
  - Use paired-end and mate pair information
- **Sequencing errors**
  - increase the size of the DB graph (Tips and bubbles)
  - need to be lower than intra genomic variations
- Need to **experimentally design** the sequencing experiment (sequence recipes from ALLPATH-LG) according to the genome characteristics and sequencing experiment.



Velvet strategy (Zerbino et al. 2008)


GGCTTTA TCCGATG TTTAGAG
CGAGGCT TAGATCC TGAGGCT GAGACAG
AGTCGAG TTTAGATC ATGAGGC TTAGAGA
GAGGCTT GATCCGA GAGGCTT GAGACAG
2. Hashing
Linear stretches
GATT (1x)
TGAG (9x) ATGA (8x) GATG (5x) CGAT (6x) CCGA (7x) TCCG (7x) ATCC (7x) GATC (8x) AGAT (8x)
AGAA (1x)
GCTC (2x) CTCT (1x) TCTA (2x) CTAG (2x)
GCTT (8x) CTTT (8x) TTTA (8x) TTAG (12x)
TAGT (3x) AGTC (7x) GTCG (9x) TCGA (10x) CGAG (8x) GAGG (16x) AGGC (16x) GGCT (11x)
CGAC (1x) GACG (1x) ACGC (1x)
TAGA (16x) AGAG (9x) GAGA (12x) AGAC (9x) GACA (8x) ACAG (5x)
3. Simplification of linear stretches
GATT AGAT
GATCCGATGAG
GCTCTAG
AGAA
Tips
TAGTCGA CGAG
GAGG AGGCT
GCTTTAG
TAGA AGAGA AGACAG
Bubble
CGACGC
4. Error removal
AGATCCGATGAG
TAGTCGAG GAGGCTTTAGA AGAGACAG

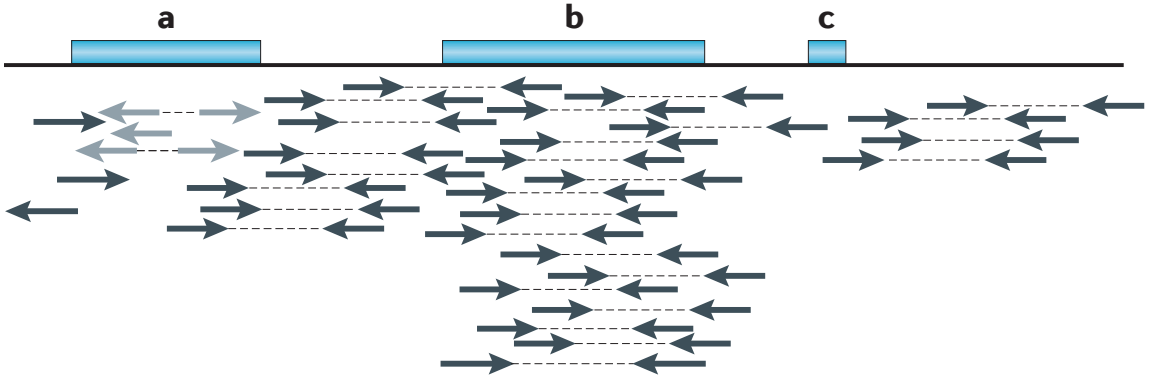
Repeat content

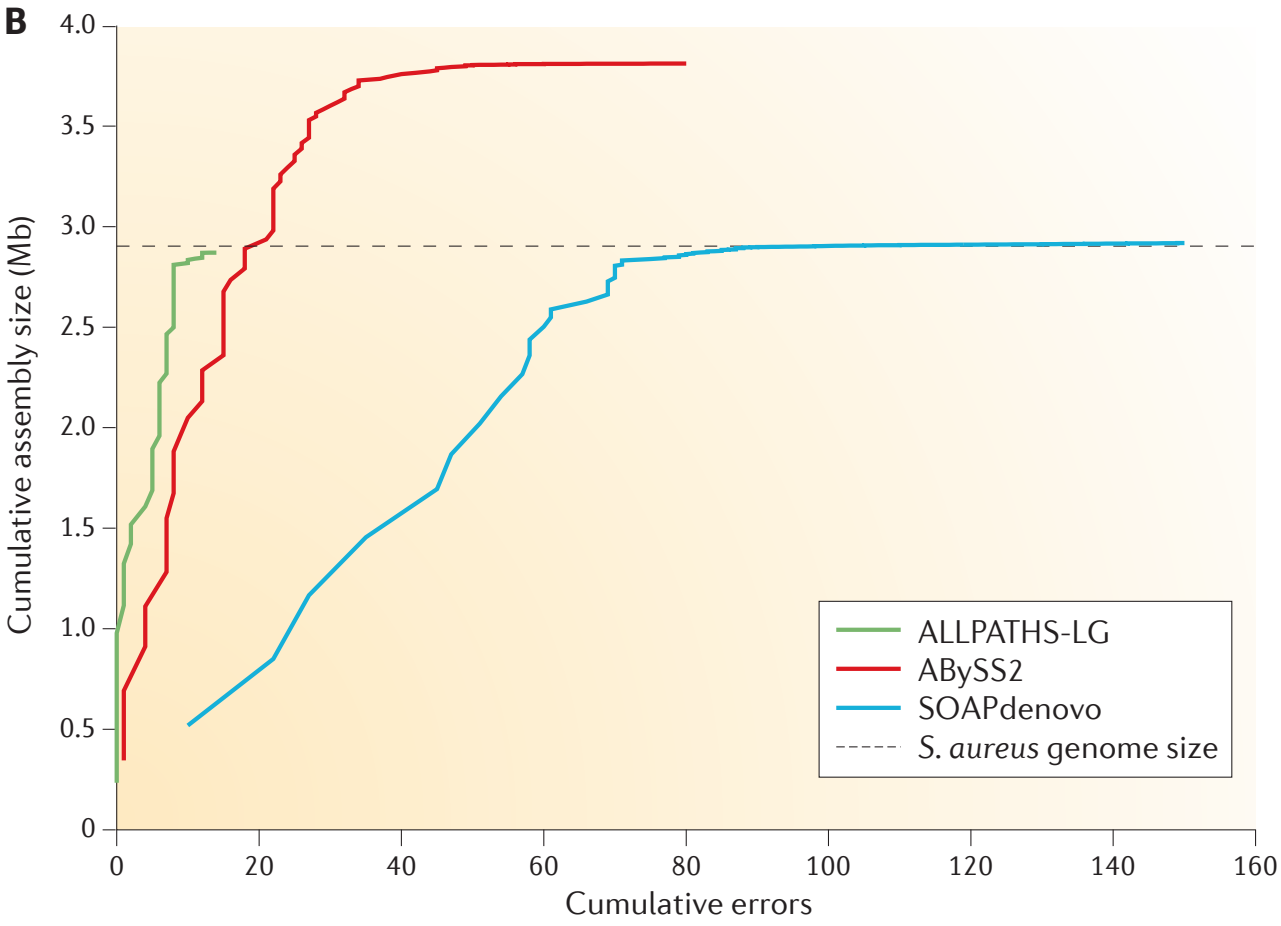


Uniqueness ratio: percentage of the genome that is covered by unique sequences of length *k* or longer

Assembly validation

- Measure contiguity of assembled sequences
  - total size and number of contigs
  - N50: weigthed median contig size
- Evaluate assembly errors
  - compare to BAC sequences, transcriptome, closely related organism
  - consistency after aligning the reads




B
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
Cumulative assembly size (Mb)
0
20
40
60
80
100
120
140
160
Cumulative errors
ALLPATHS-LG
ABySS2
SOAPdenovo
S. aureus genome size

THE ASSEMBLATHON

http://assemblathon.org/

- Competitive assessment of *de novo* assembly
  - Assemblathon3 plans to work with limited ressource
    - 20k$ to each team
    - Use money to buy sequencing
- Assemblathon1:
  - Evaluate on a simulated diploid genome (from chr13)
  - Setting up metrics is not an easy task:
    N50-NG50-CPNG50-CC50
- Evaluation of classical tools:

  Allpath-LG, Velvet, Euler-SR, Ray, SOAPdenovo, Abyss…

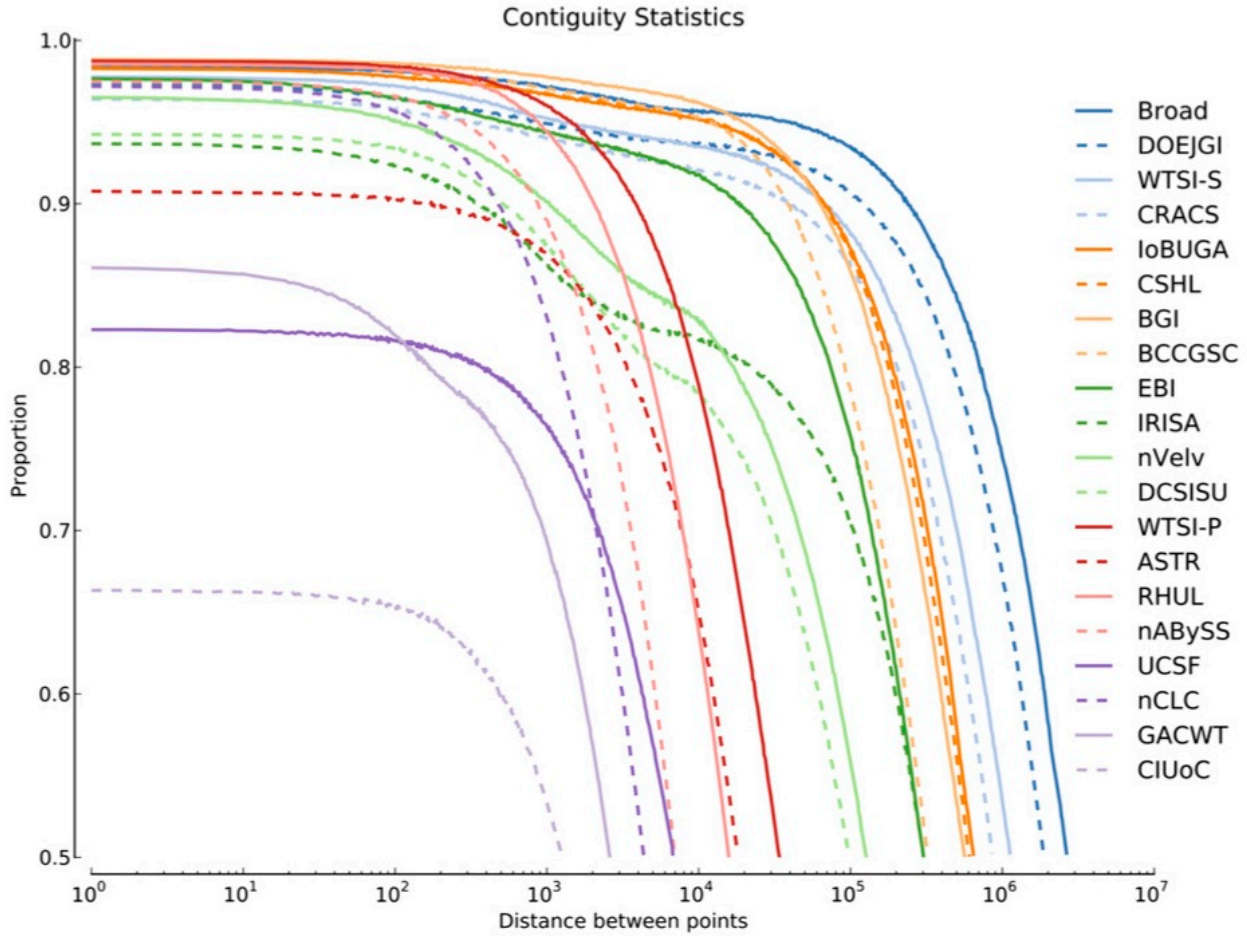


Figure 5. The proportion of correctly contiguous pairs as a function of their separation distance. Each line represents the top assembly from each team. Correctly contiguous 50 (CC50) values are the lowest point of each line. The legend is ordered *top* to *bottom* in descending order of CC50. Proportions were calculated by taking 100,000,000 random samples and binning them into 2000 bins, equally spaced along a $\log_{10}$ scale, so that an approximately equal number of samples fell in each bin.

Other Applications of Assembly

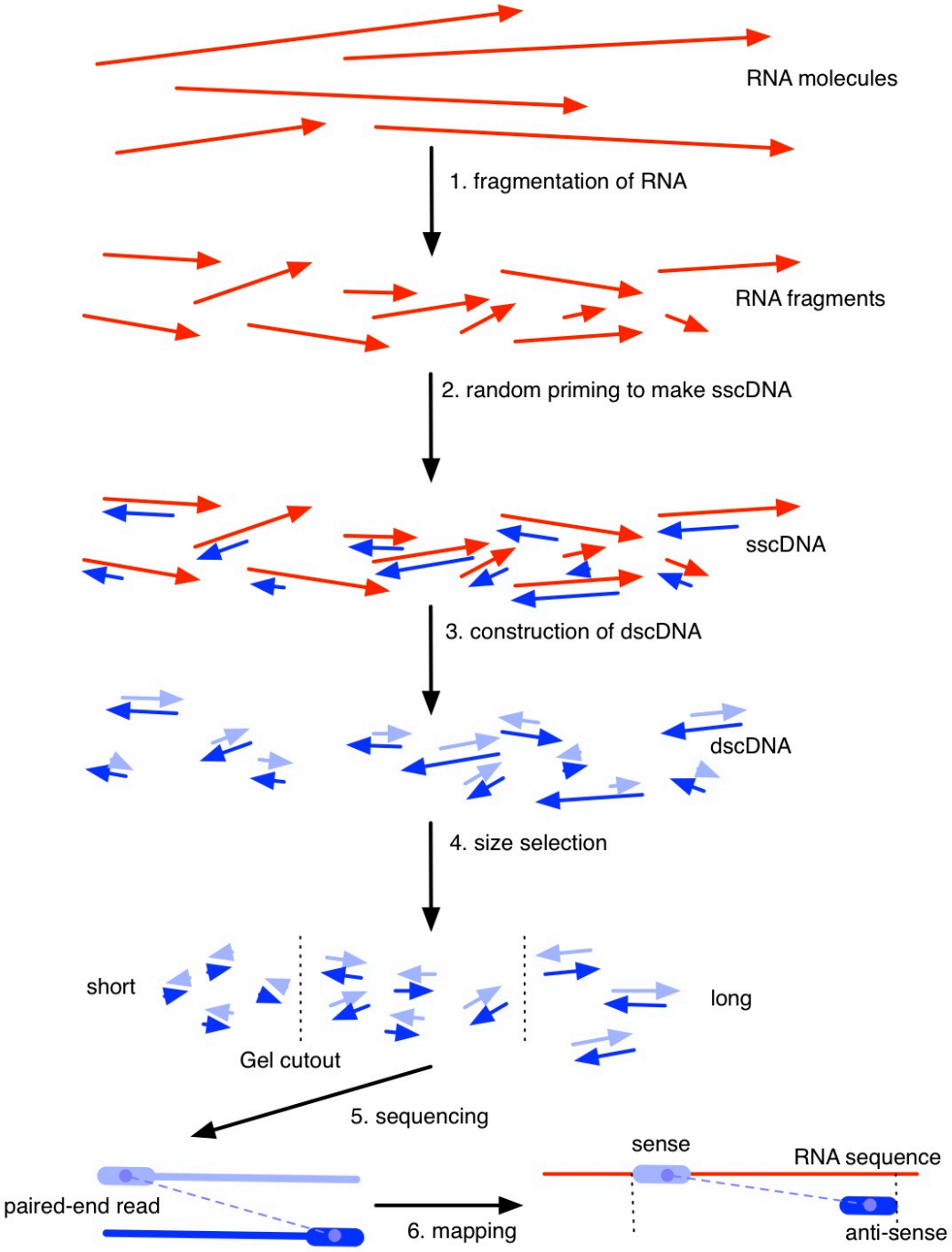
- Whole genome sequencing of isolate genome
  - plenty of material, uniform coverage
- Single cell whole genome sequencing
  - uneven coverage due to amplification
- Transcriptome assembly
  - many transcript contigs with uneven abundances
  - sequence similarity between isoforms
- Metagenomics

Assemblers	Technology	Availability	Notes
Genome assemblers			
ALLPATHS-LG	Illumina, Pacific Biosciences	ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG	Requires a specific sequencing recipe (BOX 3)
SOAPdenovo	Illumina	http://soap.genomics.org.cn/soapdenovo.html	Also used for transcriptome and metagenome assembly
Velvet	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/velvet	May have substantial memory requirements for large genomes
ABySS	Illumina, SOLiD, 454, Sanger	http://www.bcgsc.ca/platform/bioinfo/software/abyss	Also used for transcriptome assembly
Metagenome assemblers			
Genovo	454	http://cs.stanford.edu/group/genovo	Uses a probabilistic model for assembly
MetaVelvet	Illumina, SOLiD, 454, Sanger	http://metavelvet.dna.bio.keio.ac.jp	Based on Velvet
Meta-IDBA	Illumina	http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba	Based on IDBA
Transcriptome assemblers			
Trinity	Illumina, 454	http://trinityrnaseq.sourceforge.net	Tailored to reconstruct full-length transcripts; may require substantial computational time
Oases	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/oases	Based on Velvet
Single-cell assemblers			
SPAdes	Illumina	http://bioinf.spbau.ru/en/spades	
IDBA-UD	Illumina	http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud	Based on IDBA

Nagarajan & Pop, Nature Review Genetics, 2013

2. Analysis of the Transcriptome

RNA-Seq data analysis


RNA molecules
1. fragmentation of RNA
RNA fragments
2. random priming to make sscDNA
sscDNA
3. construction of dscDNA
dscDNA
4. size selection
short
long
Gel cutout
5. sequencing
paired-end read
6. mapping
sense
RNA sequence
anti-sense

RNA-Seq protocol (2)

sample of interest

Extraction of RNAs, RT,
random shearing

Fragments

amplification
sequencing

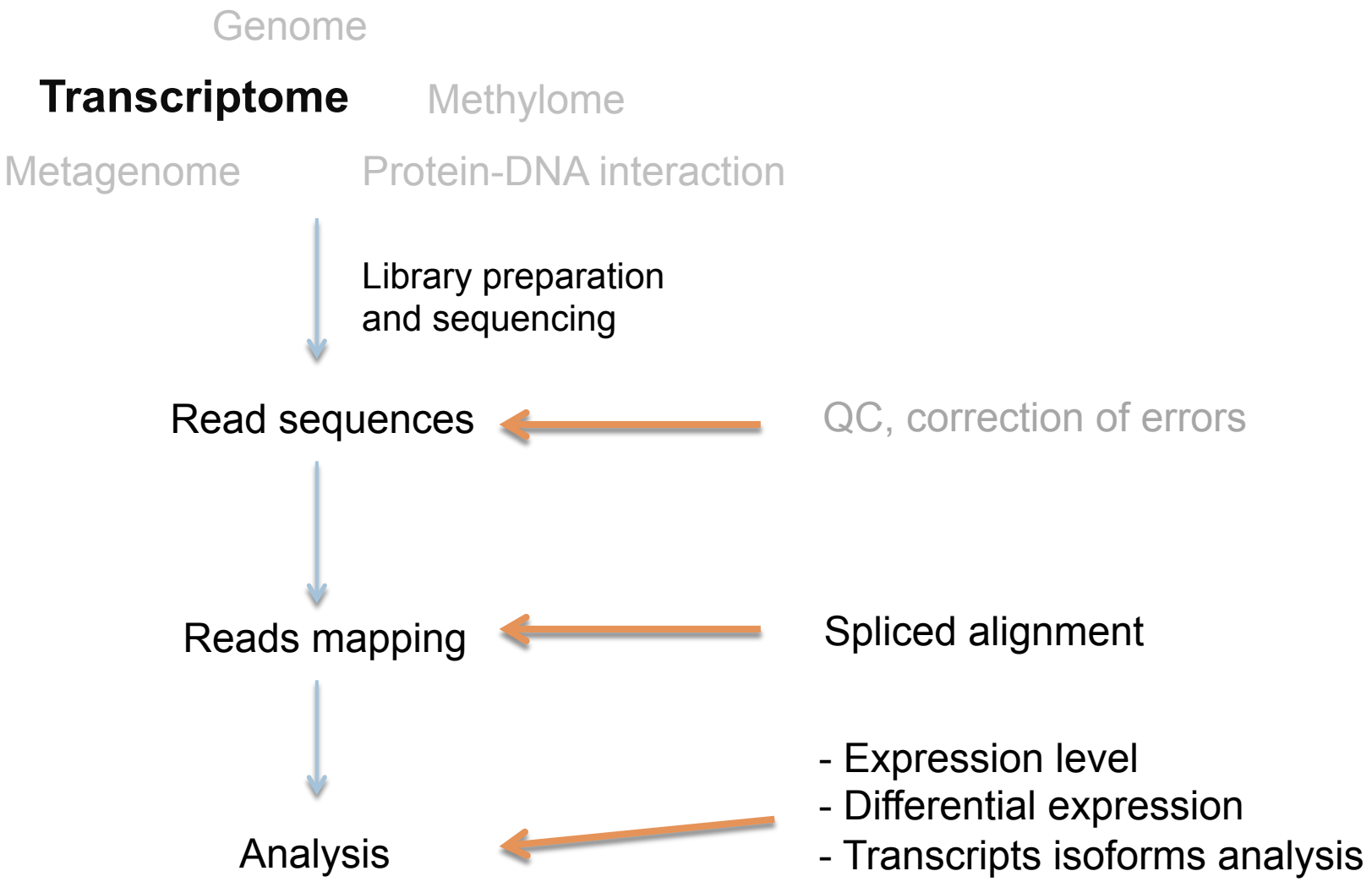
Sequenced reads

alignment

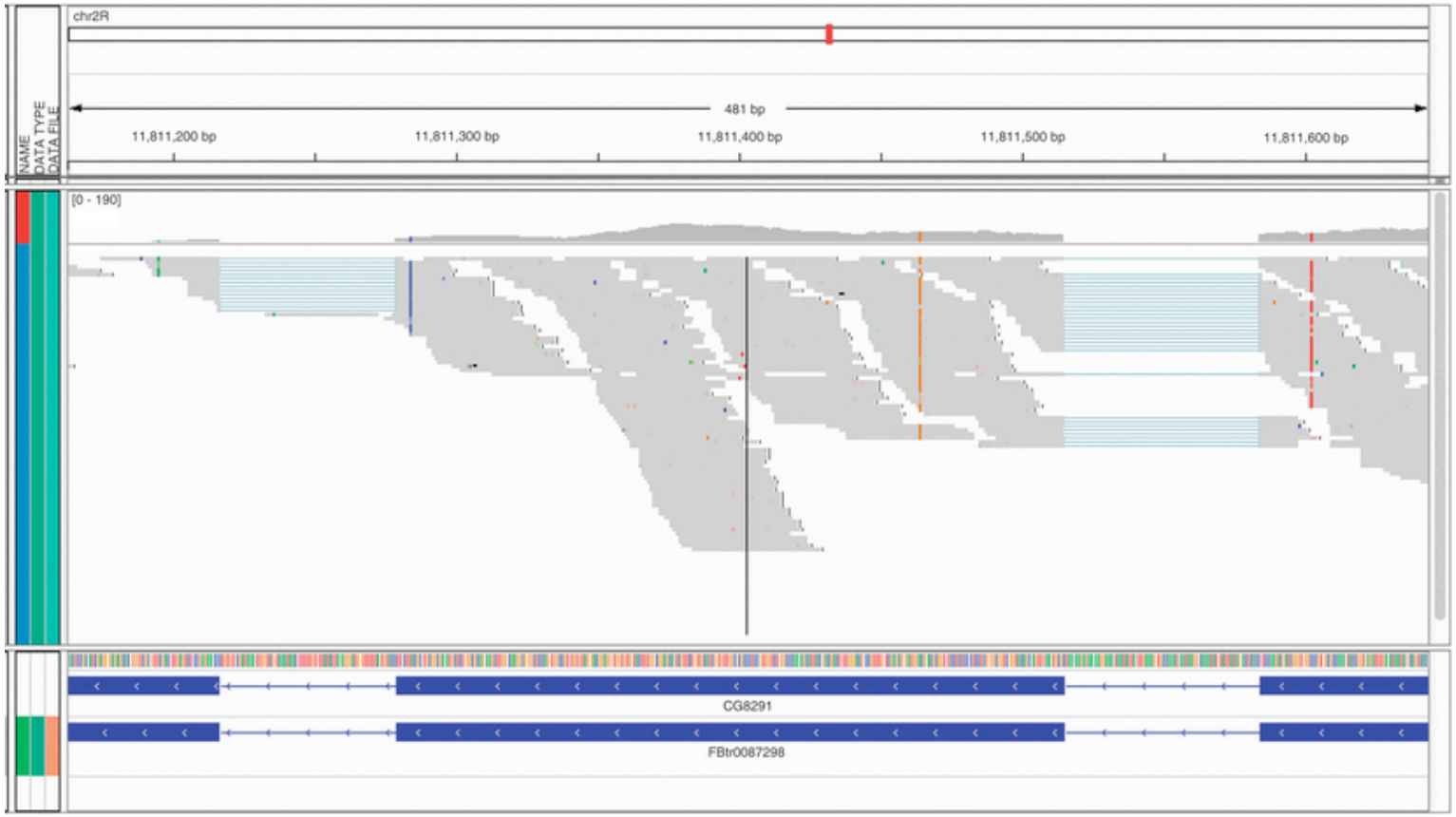
Some variants:

Small RNA (<200nt)
Transcripts ends (CAGE, poly-dT)
Ribosomal profiling

RNA-Seq workflow



RNA-Seq data



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

RNA-Seq

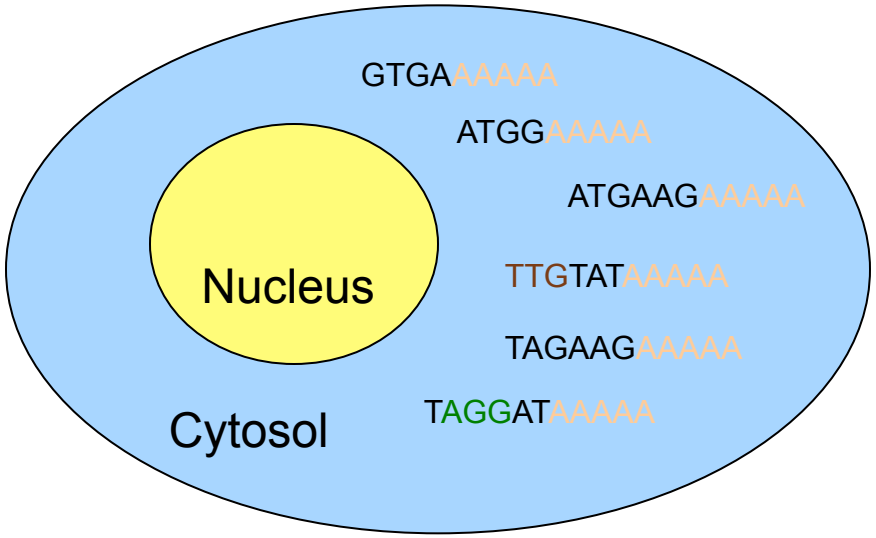
- Snapshot of transcripts activity
  - Direct identification (sequencing cDNAs)
  - Digital counting (large dynamic range)
- Questions
  - Determine transcript structure (boundaries, splicing…)
  - post-transcriptional modifications
  - Quantify expression levels/differential expression
- Challenges
  - large quantities of data (1 Gb/Mio reads)
  - still many biases/artifacts
  - transcript reconstruction is non trivial

Outline

- Gene/Exon Expression level quantification
  - Sampling model and setup
  - Testing for differential expression

- Analysis of Alternative Splicing/Transcripts isoforms
  - Spliced alignment
  - Transcripts isoforms analysis
    - Transcripts isoforms quantification
    - Transcriptome reconstruction/*de novo* assembly

Counting reads



"bag of transcripts positions"

TTG

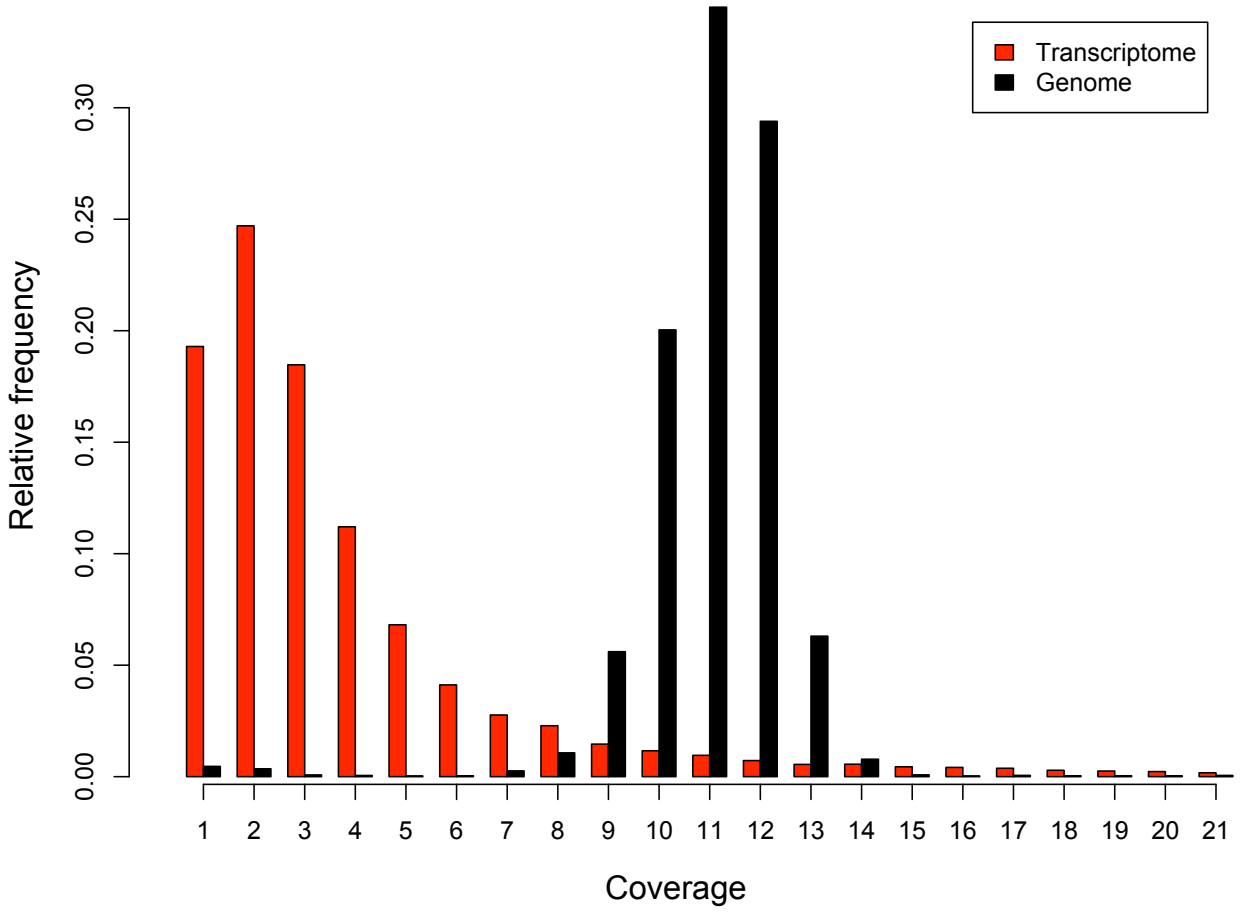
AGG

AAG

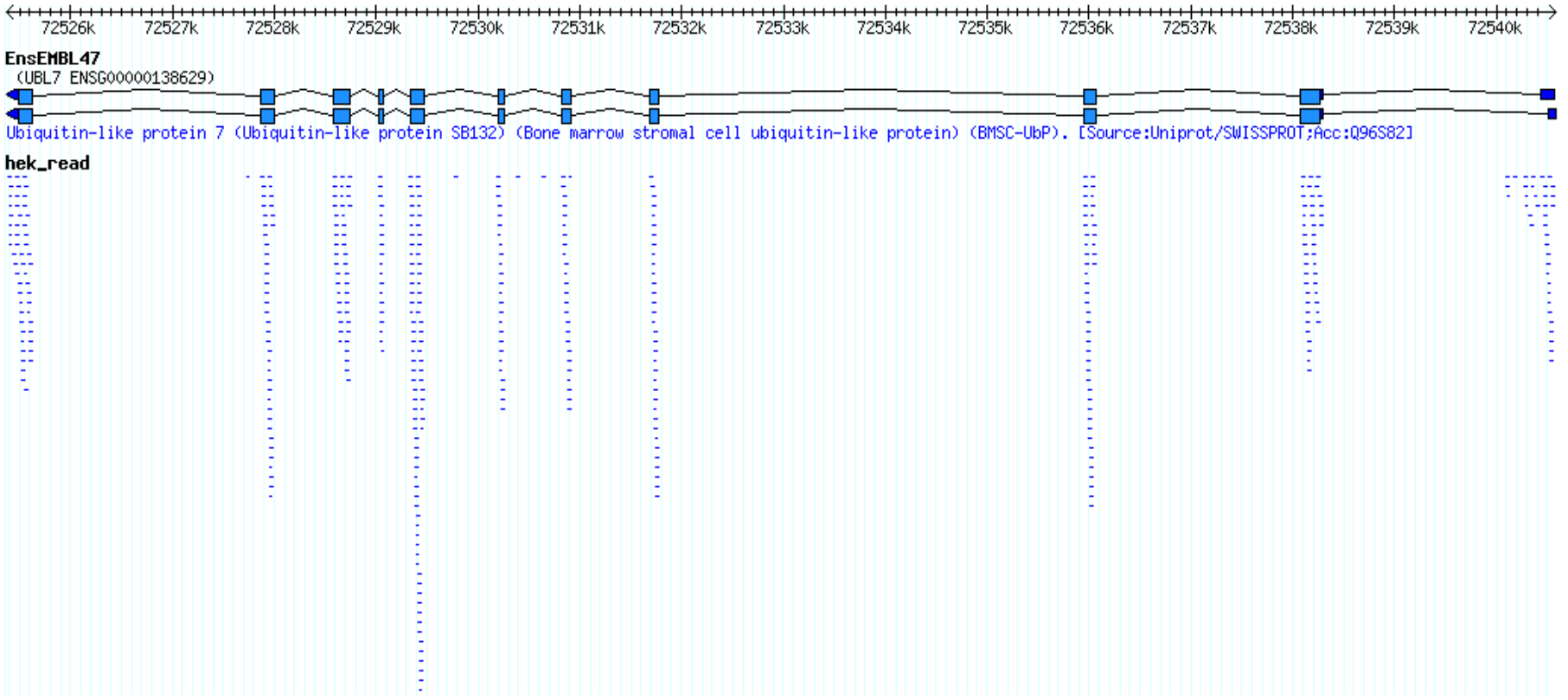
Counts ideally follow a Poisson distribution, with an influence for transcript length

Transcriptome vs genome coverage

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



Mapping to annotated exons

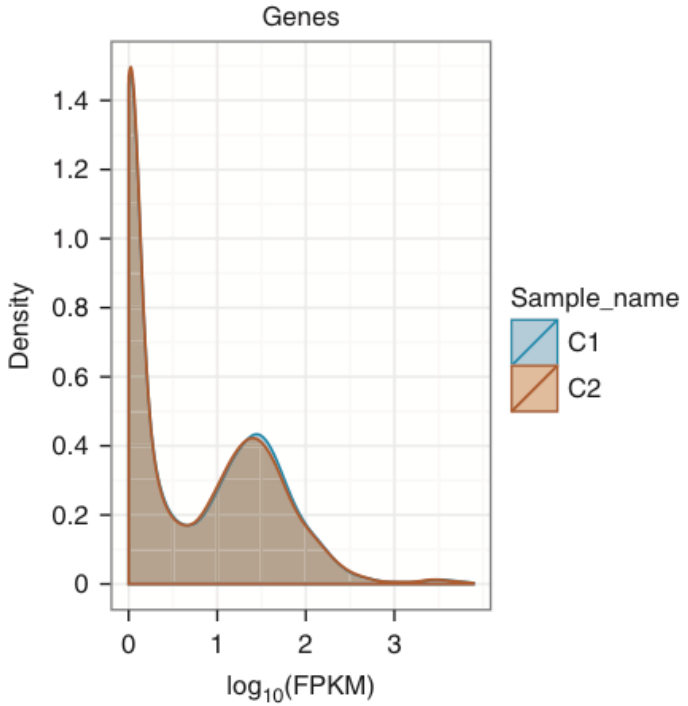


Questions:

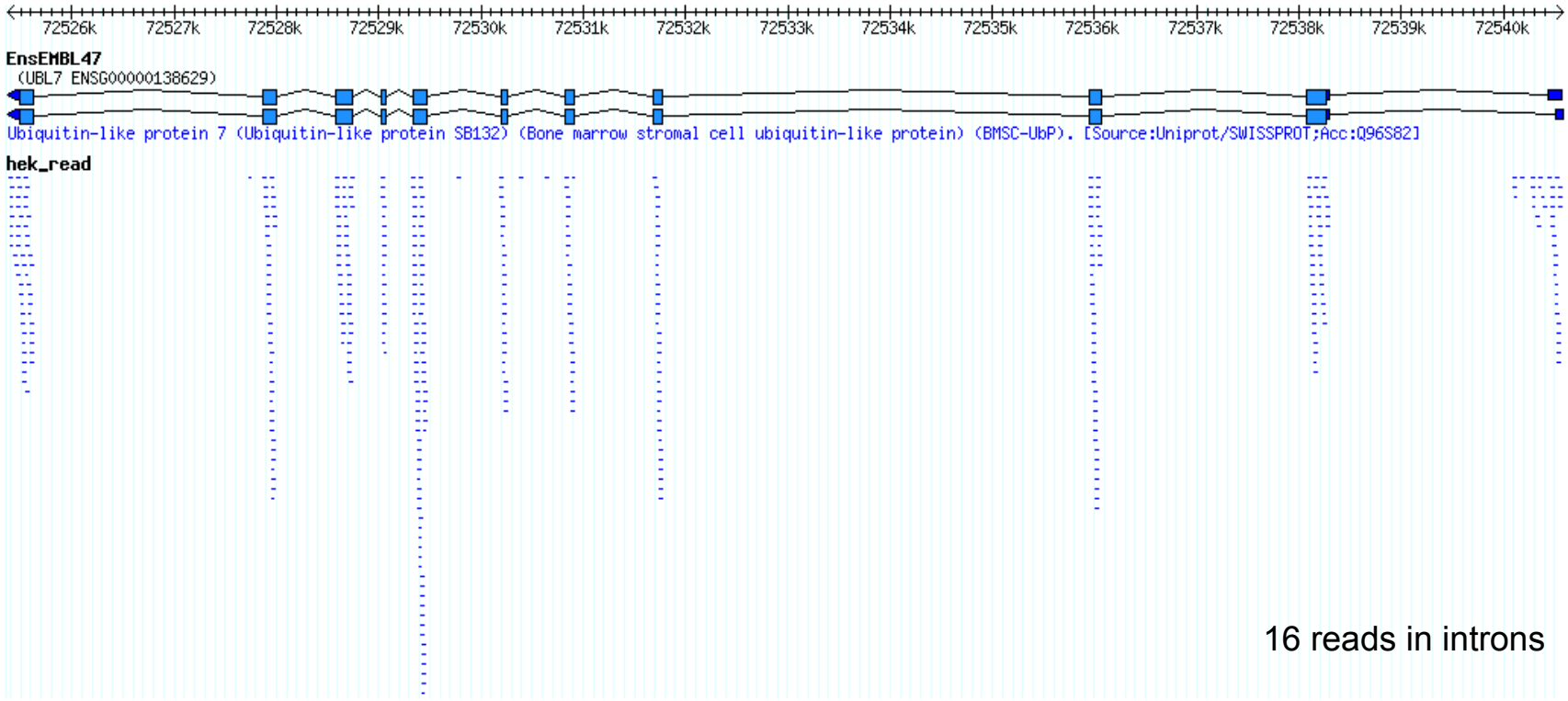
- Is a gene/region expressed?
- How to quantify gene expression level ?

Normalizing expression

- Count mapped reads on regions (exon, gene…)
  - proportional to sequencing depth and region length.
- RPKM: Reads Per Kilobase per Million (reads mapped) (Mortazavi et al 2008)
  - FragmentPKM for mate pairs.
  - Expressed region >1-10 RPKM (depending on complexity).
  - ! Normalization not robust (a few genes account for most of the reads)

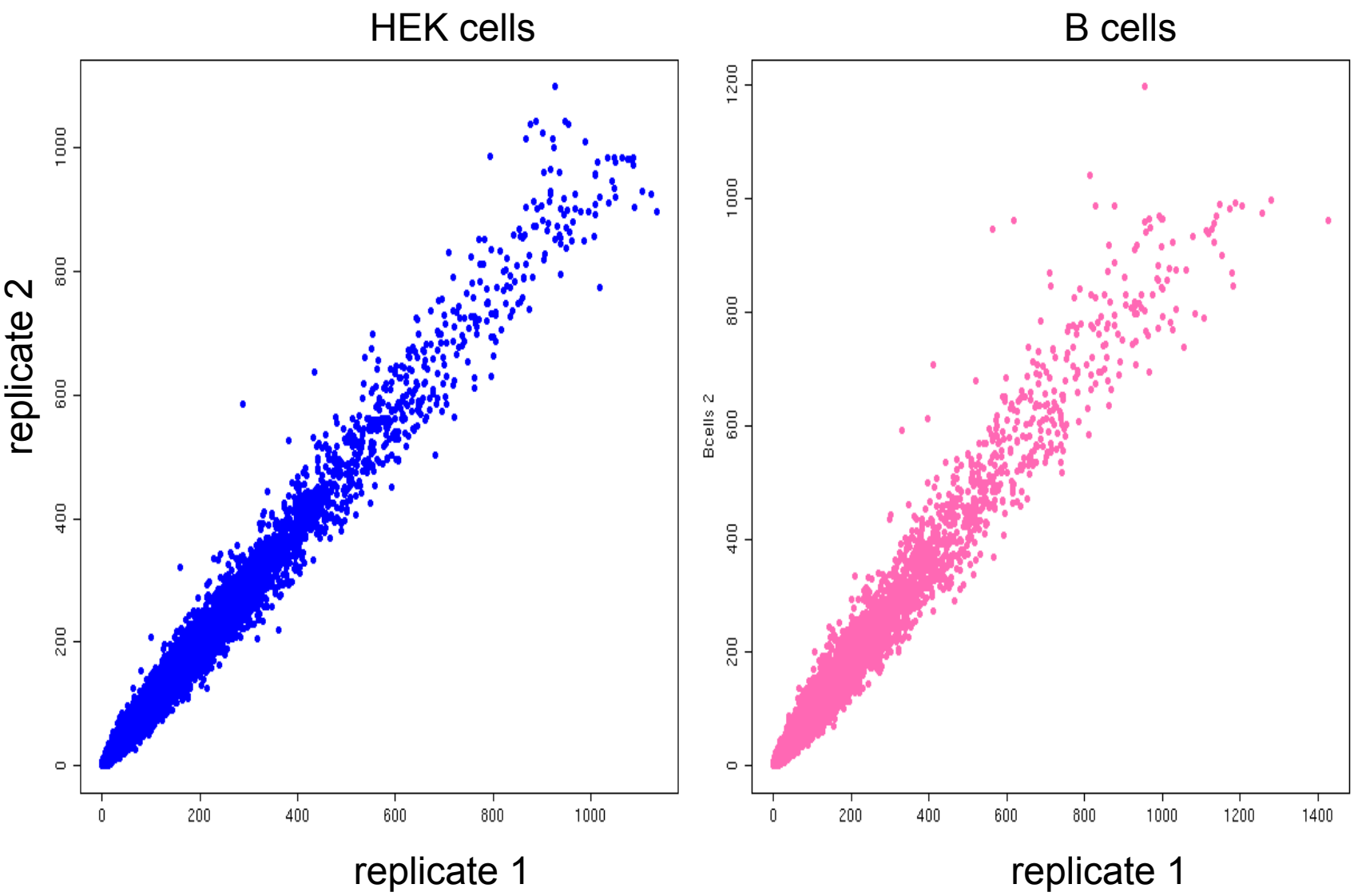


Variability on 2 technical replicates



HEK 1	24		16	17	6	42	8	15	16		20	20	7
HEK 2	25		19	23	8	34	10	8	15		20	21	6

reads count are variable
(sampling by sequencing)


HEK cells
B cells
replicate 2
replicate 1
Bcells 2
replicate 1
0
200
400
600
800
1000
1200
1400

Count Model For Expression

- Read counts summarized over a region are variable:
  - sampling of sequence fragments.
  - (library preparation effect)
  - (biological variability)
- Observed counts $Y_{i,j}$ (only with sampling variability):

$$Y_{i,j} \sim \quad \mathcal{P}(\mu_{i,j}) \quad E(Y_{i,j}) = Var(Y_{i,j}) = \mu_{i,j}$$

- Discrete distribution (different from Gaussian)
- Compare counts from same region
  - no length normalization, only sample size

Testing differential expression

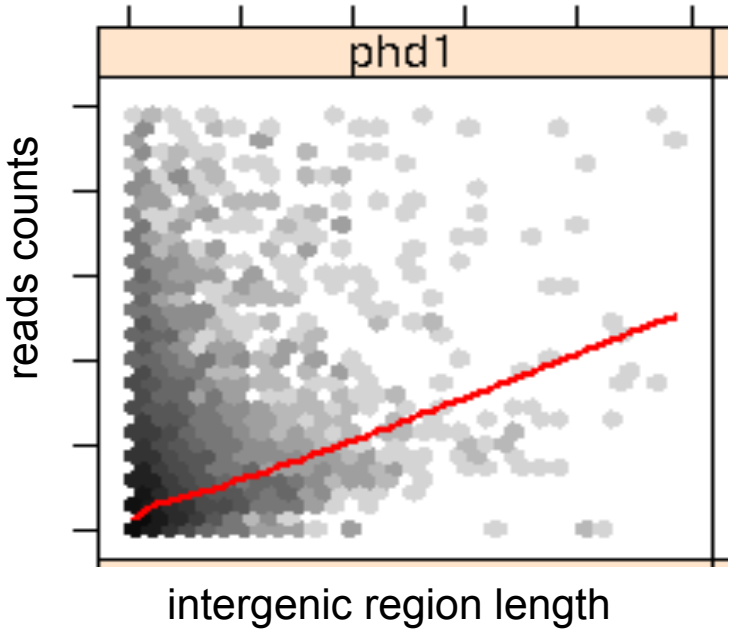
- Are differences in read counts higher than the variability expected from the experiment ?
  - Is the count on a region higher than the noise from experiment?
  - Is the expression of a gene/exon different between two experimental conditions?
- Do statistical hypothesis testing:
  - How unlikely is the observed difference in counts? (how significant)
  - Need a null model and a test statistics…

Test expression above noise

- Assume a proportion $p_{0,j}$ for noise.

  H_0 = { region present in proportion $p_{0,j}$ }

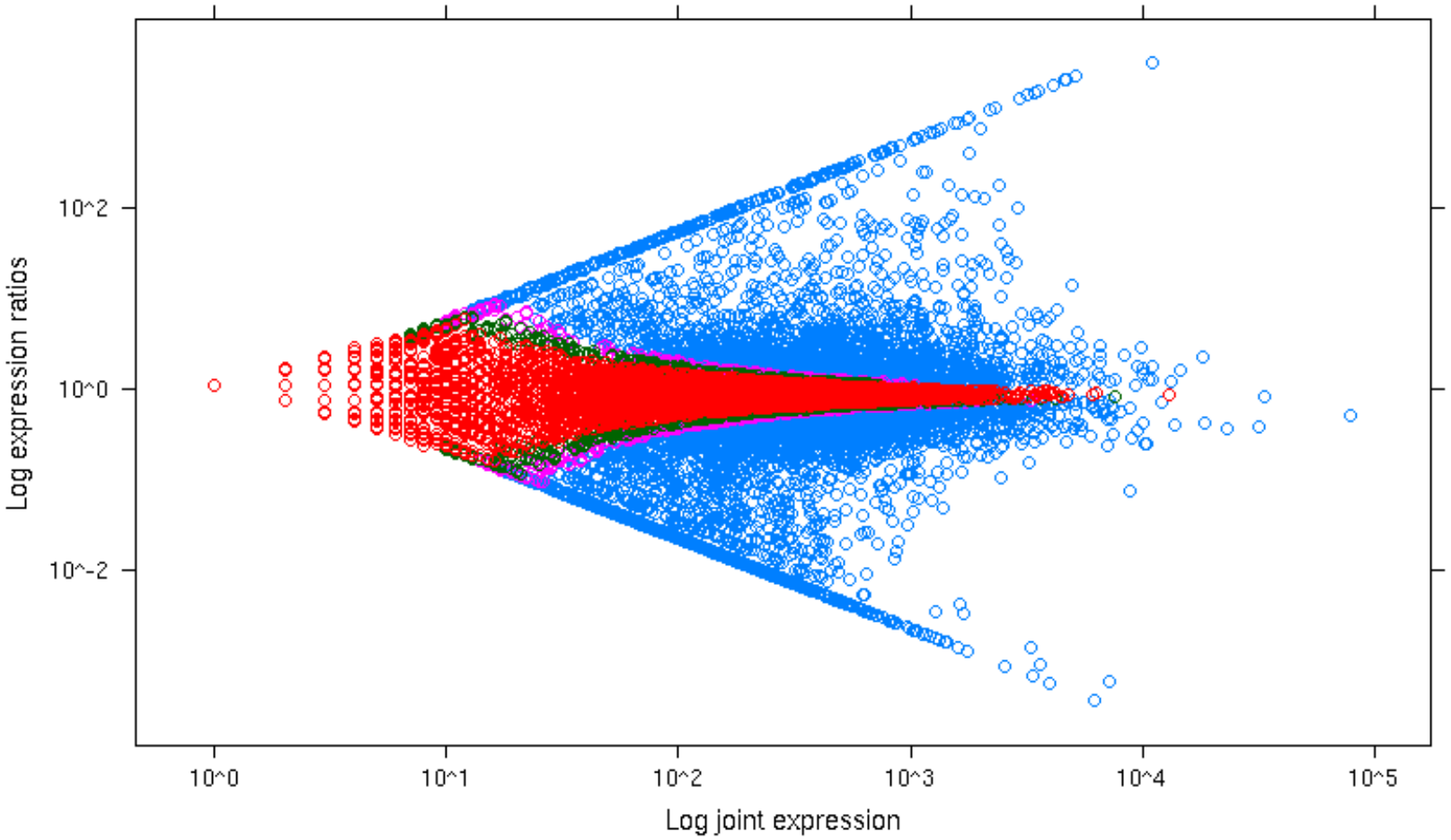
  $p_{0,j}$ estimated on the set of intergenic regions



Wang ET et al, *Nature* (2008)

Noise ~ 0.4 RPKM (Zebrafinch forebrain RNA-Seq)

Higher variability with lower counts



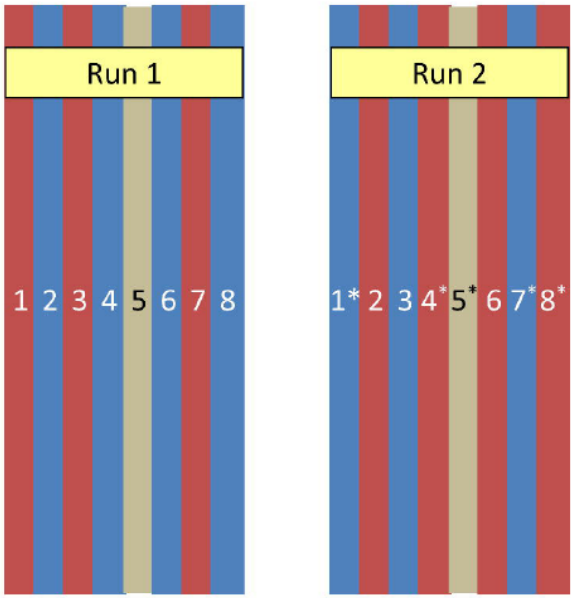
Colors refer to the probability of differential expression

Verifying Poisson assumption

(on technical replicates)

- Assess technical variability in 2 conditions.
  - should obey Poisson distribution
- Only 0.5% of genes show departure from the Poisson distribution.

Illumina study design



Kidney
Liver
* Sequenced at a concentration of 1.5 pM

Marioni et al, *Genome Research,* 2008

Extra variability

- But there are other terms than sampling variability:
  - library preparation
  - biological variability (at gene and sample level)

Var(Expr) = Across Group Variability + Measurement Error + Biological Variability

KD Hansen et al. 2011

- Negative Binomial is usually used to account for overdispersion on counts

Poisson	$v = \mu$
Poisson + constant CV	$v = \mu + \alpha \mu^2$

(knowing mean and variance is sufficient)

Estimate variance

- Estimate on each gene:
  - originally developed for SAGE data (Baggerley et al 2005)
  - but limited power due to small number of replicates

- Estimate a term common to most genes using mean variance relationship:
  - EdgeR (Robinson et al 2010, Mc Carthy et al 2012)
  - DESeq (Anders et al 2010, 2012)

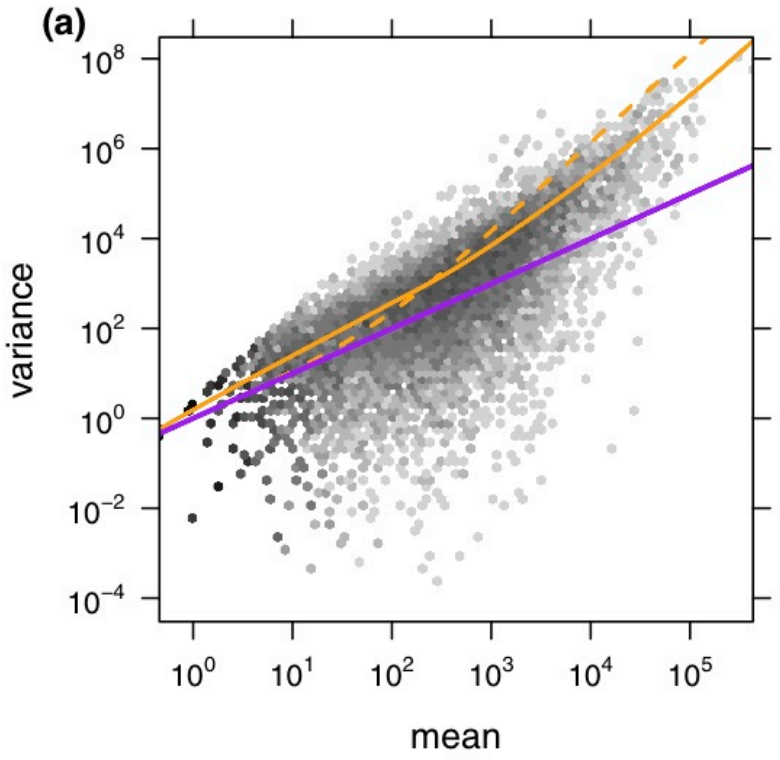
DESeq

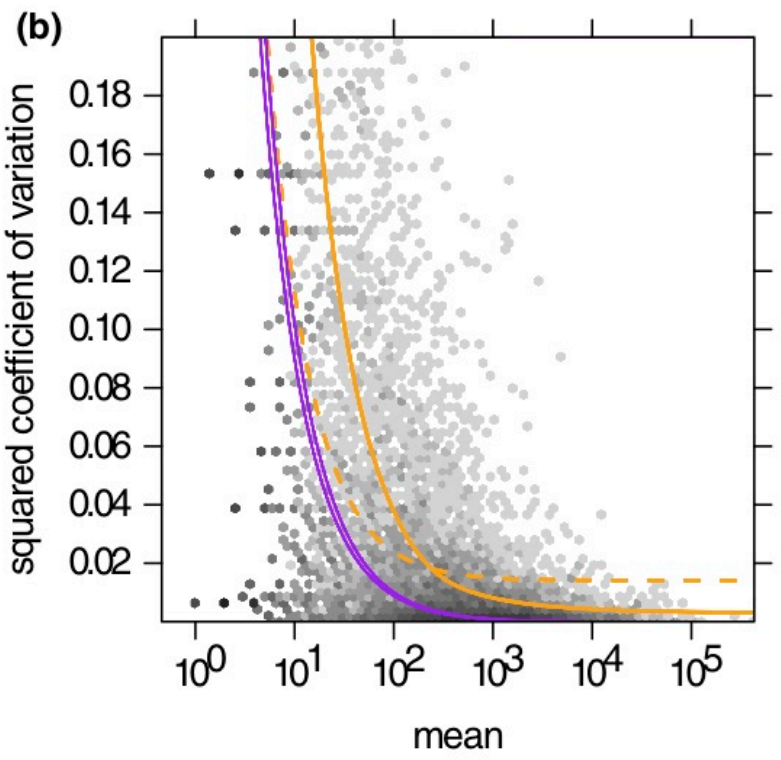
- Estimates the mean vs variance relationship with a local regression on replicates

$$v = \mu + f(\mu^2)$$

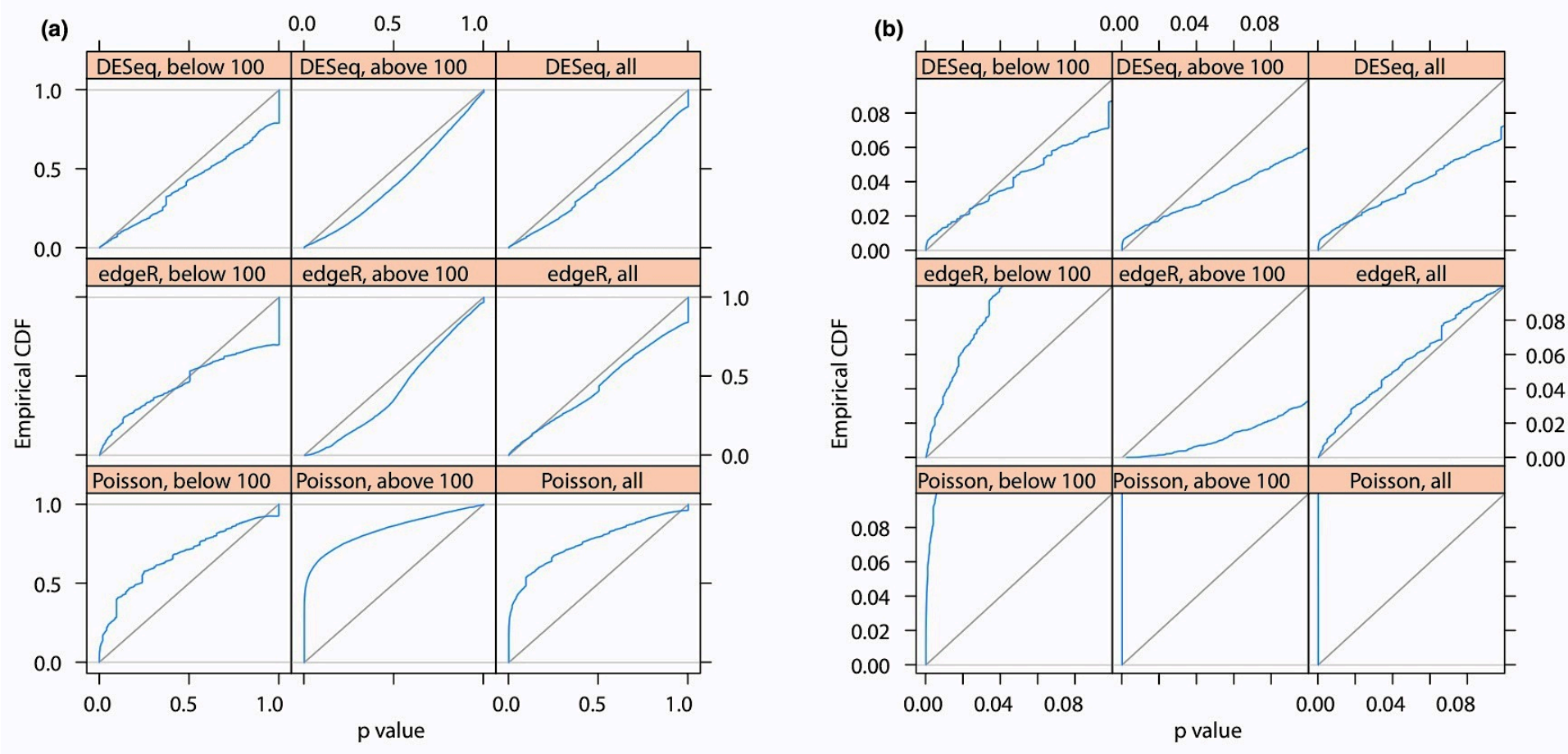
(also scaling factor from sequencing depth)

- Test for differential expression for the pooled counts of all replicates in each conditions c_A and c_B.
  - probability of observing c_A+c_B or more extreme cases, given the sum is c_A+c_B
  - c_A and c_B distribution are sums of NB and approximated as NB.

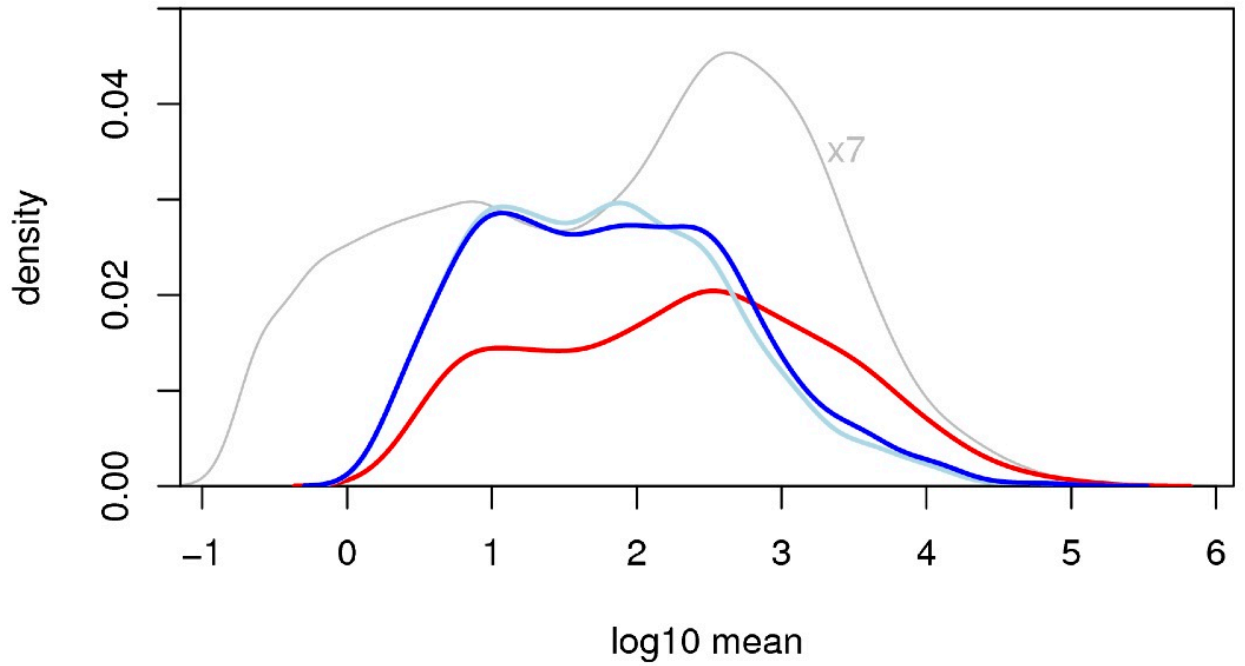




Poisson	$v = \mu$	———
Poisson + constant CV	$v = \mu + \alpha\,\mu^2$	- - - -
Poisson + local regression	$v = \mu + f(\mu^2)$	———



evaluating pvalues distribution on one replicate after estimating variance.
edgeR more sensitive on low counts, but more conservative for higher counts



all genes
differentially expressed according to DESeq
differentially expressed according to edgeR

Alternative to count model

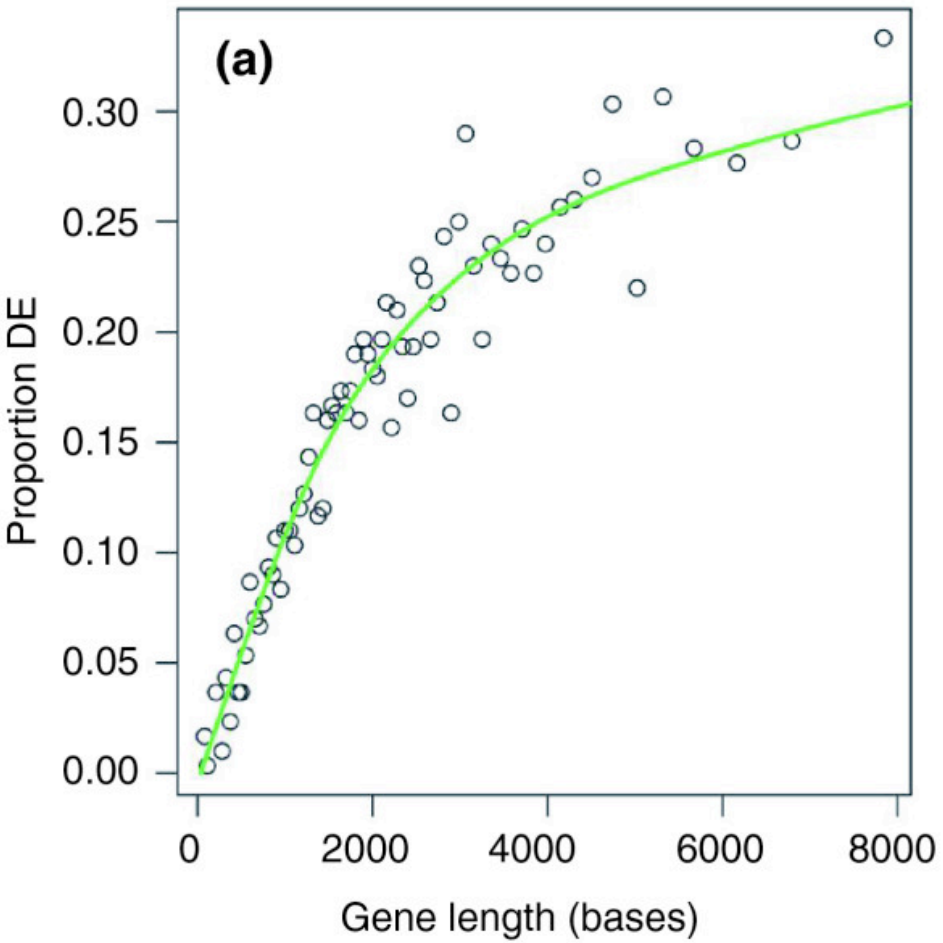
- Transform data to approximate normality
  - Usually not accurate for low counts
  - can use log, square root, arcsin…
- Limma-VOOM: log transformation to counts per million (log-cpm):
  - Can leverage linear model techniques after capturing the mean vs variance relationship on log-cpm
  - Computation of significance levels is faster and more flexible (normal distribution)

Law et al 2013

Summary DE

- 2 main terms of variability
  - Variability from sampling
    - mainly affects low counts (<100)
    - Increase power with more sequences
  - Group/biological Variability
    - more present on high counts
    - Increase power with more replicates
- Usually limited number of replicates
  - Distributional assumptions permit a tradeoff by pooling variance estimates on regions with same expression level.

Influence of the region length



- longer/highly expressed genes more likely to be DE
- biases downstream GO analysis
- reweight genes' pvalues based on their length

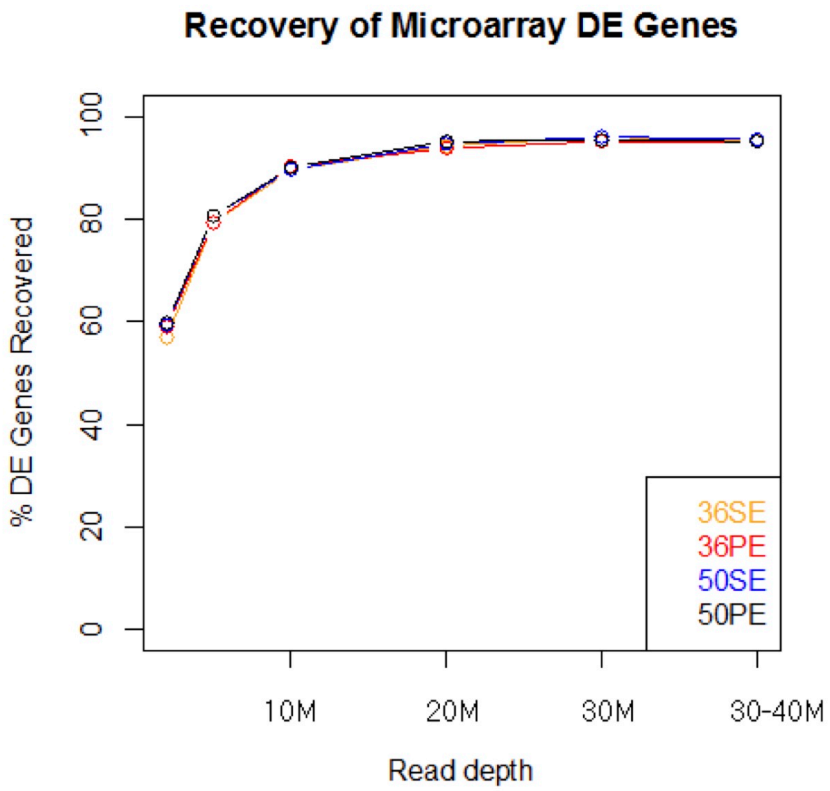
Young et al. 2010

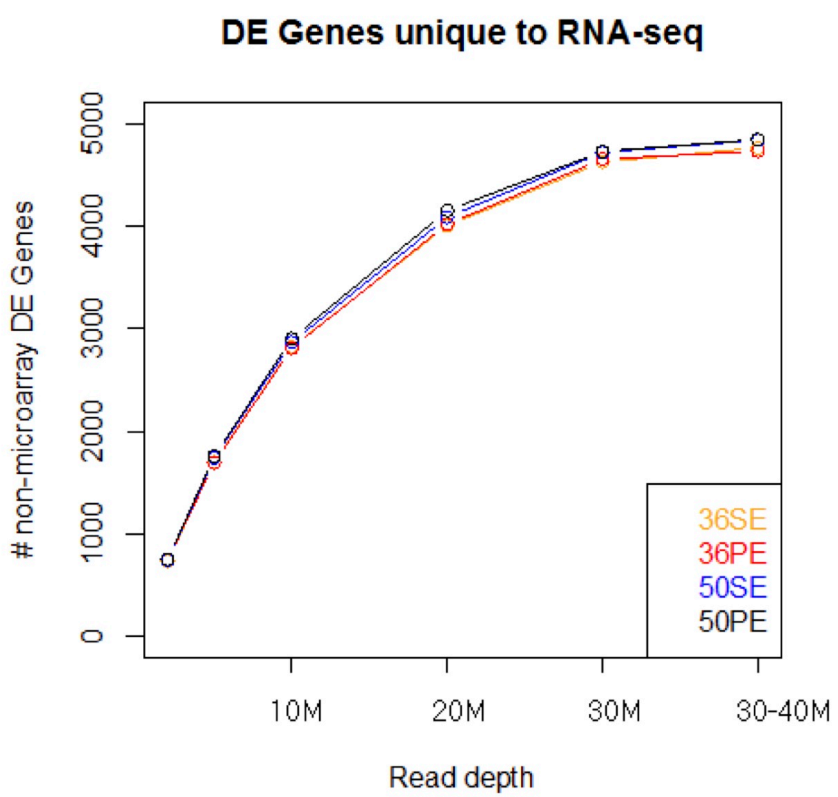
Experimental design

- Need to optimize parameters of the experiment according to budget.

- Sequencing parameters
  - read length (36 to 120 bp)
  - sequencing depth (2Mio to 100 Mio)
  - Single-end or paired-end

- Number of biological replicates

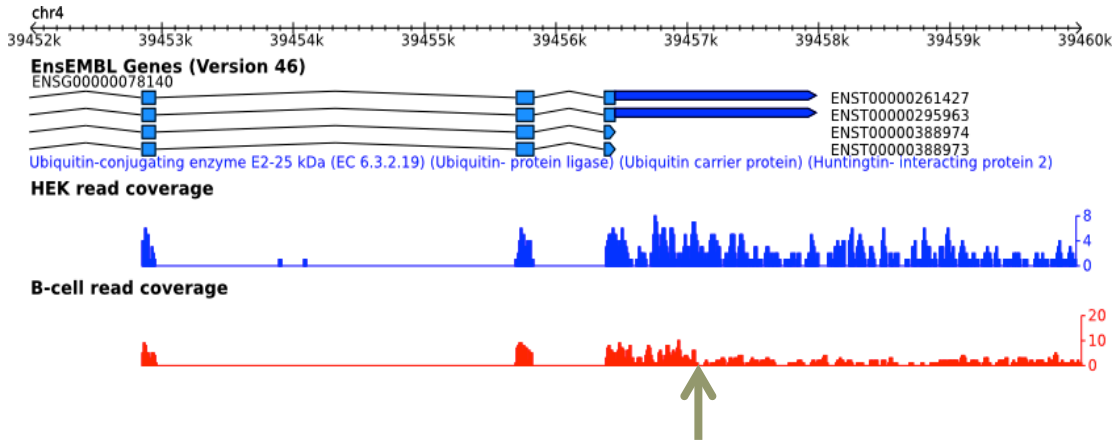




From Suraj Menon, comparison microarray vs RNA-Seq
6 replicates, tamoxifen treatment on MCF7 cell line

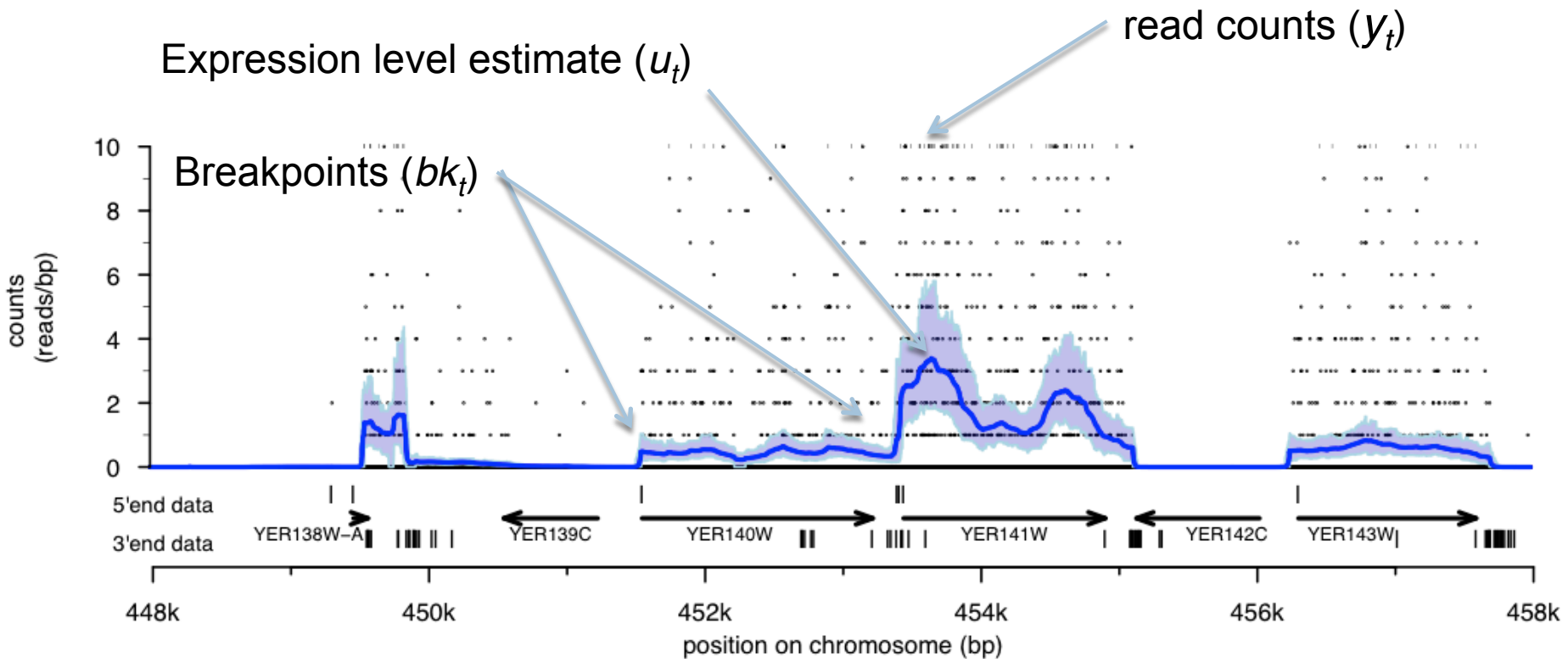
Read counts along the sequence

- Detect shift in read density without prior information



- Estimate expression level with confidence intervals

Read counts


Expression level estimate (u_t)
read counts (y_t)
Breakpoints (bk_t)
counts (reads/bp)
10
8
6
4
2
0
5'end data
3'end data
YER138W–A
YER139C
YER140W
YER141W
YER142C
YER143W
448k
450k
452k
454k
456k
458k
position on chromosome (bp)

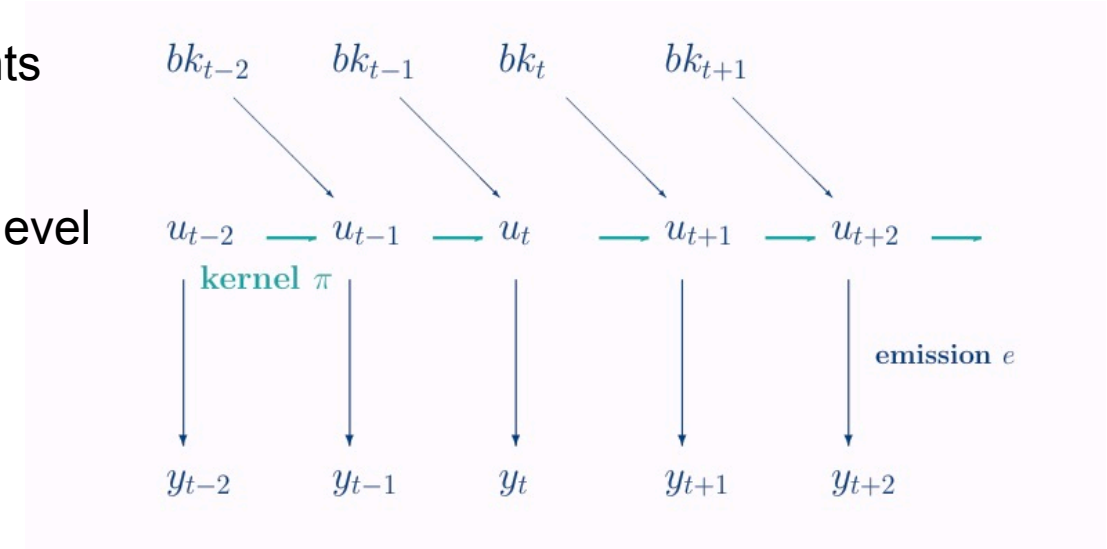
Parseq

Joint work with B. Mirauta and P. Nicolas.

- Reconstruct u_t trajectory at a bp resolution:
  - Without prior annotation
  - Not necessarily homogeneous

stranded RNA-Seq

- Segmentation with Hidden Markov models:

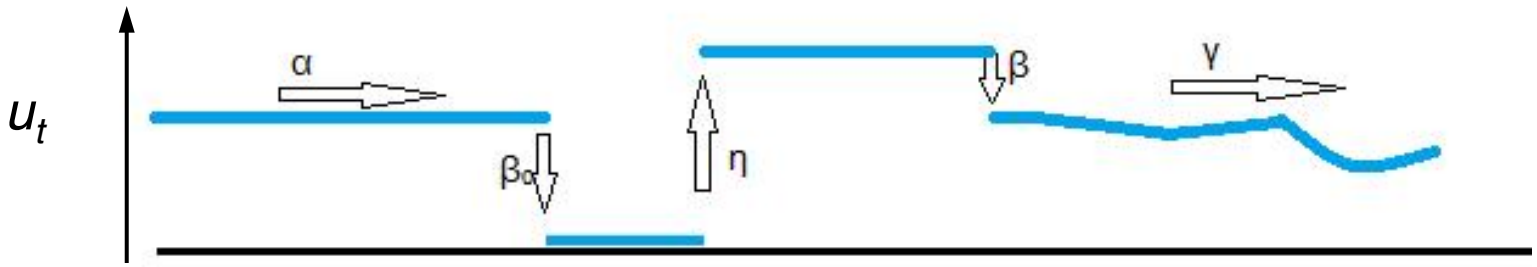


Ideally:

$$y_t \sim \mathcal{P}(u_t)$$

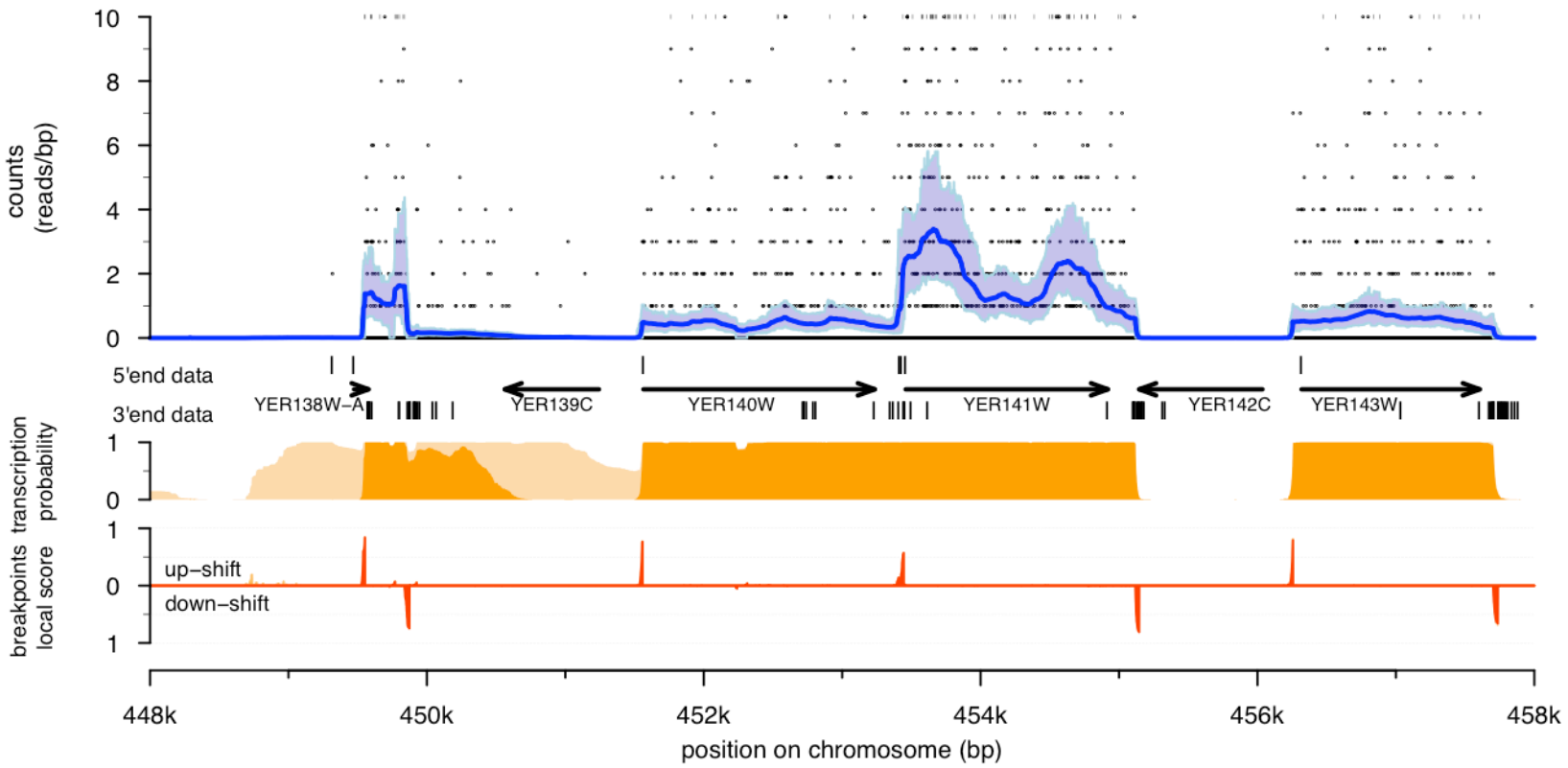
Expression level changes

- Various types of change on transition kernel



- Read counts distribution accounts for two types of variability: local and correlated
- Estimation not possible with exact methods, use *Sequential Monte Carlo* methods.

A Parseq output



http://www.lcqb.upmc.fr/parseq/

Mirauta, Nicolas*, Richard* Bioinformatics 2014

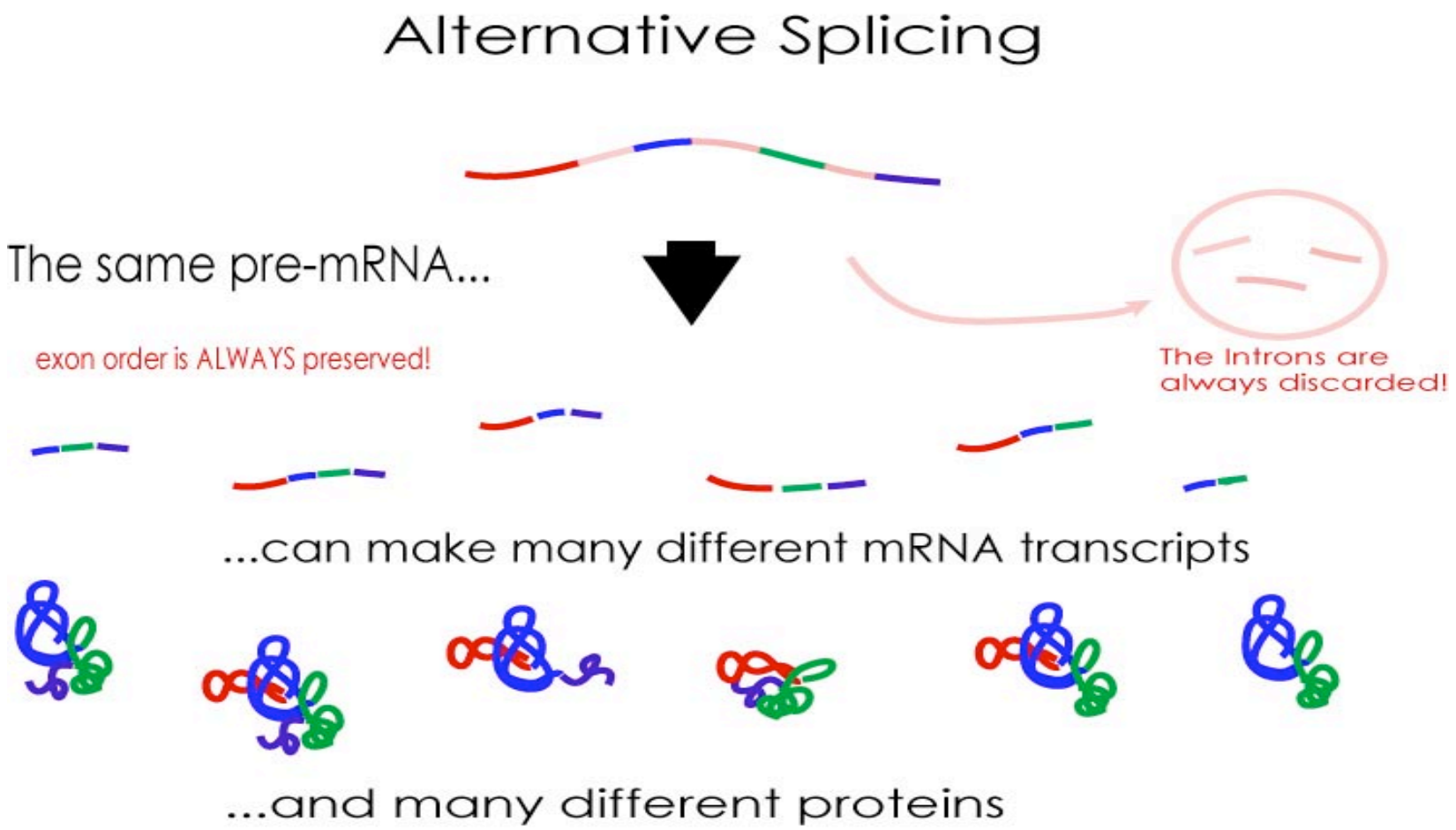
Outline

- Gene/Exon Expression level quantification
  - Sampling model and setup
  - Testing for differential expression

- Analysis of Alternative Splicing/Transcripts isoforms
  - Spliced alignment
  - Transcripts isoforms analysis
    - Transcripts isoforms quantification
    - Transcriptome reconstruction/*de novo* assembly

Introducing Alternative splicing

1 pre mRNA → multiple transcripts (also called isoforms)

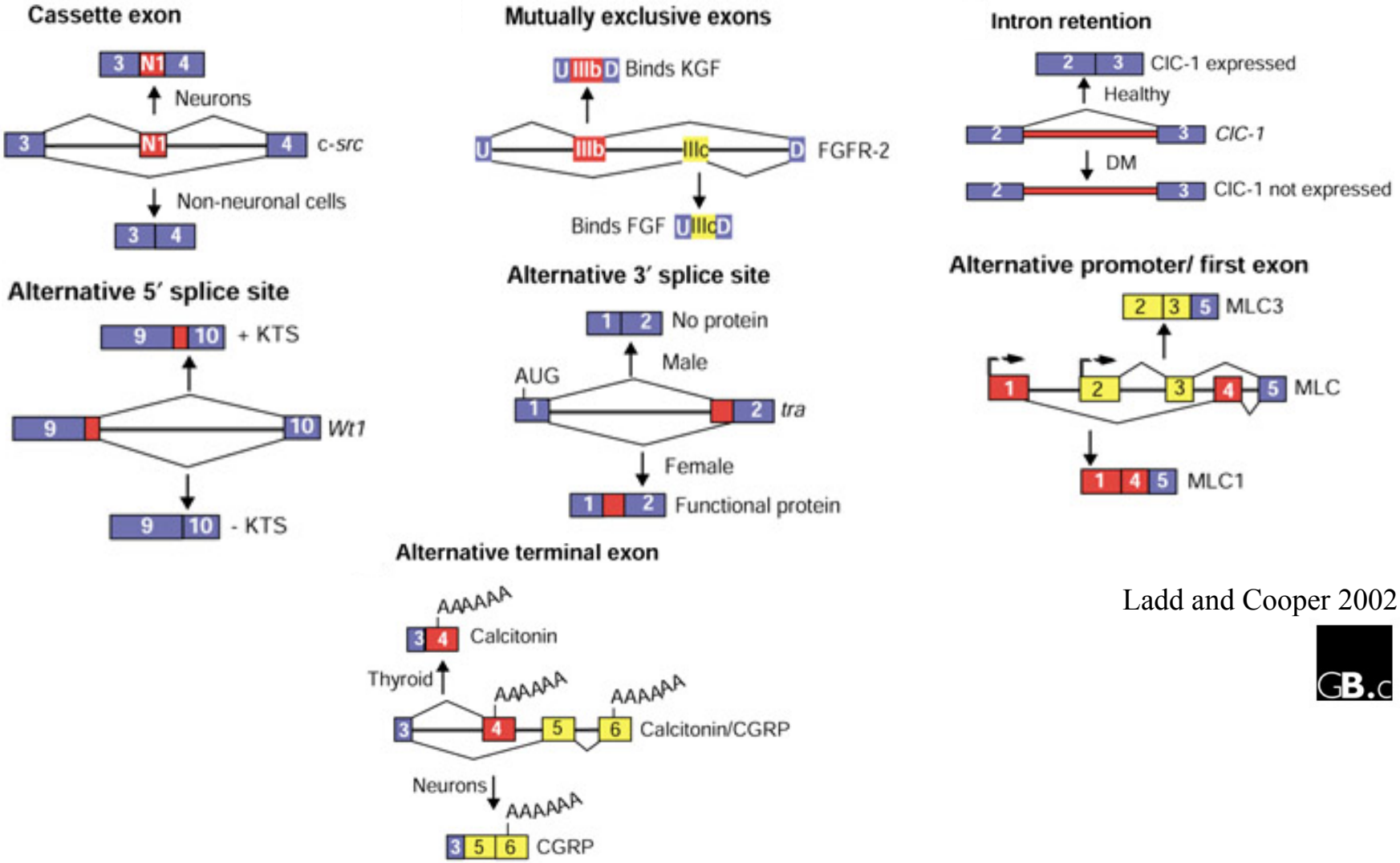


Importance of AS

- >90% of human multi-exon genes undergo alternative splicing.
- Important in regulation of gene function.
- Aberrant splicing is a major cause of human diseases.
- An important mechanism for acquisition of evolutionary novelties (e.g. between human and other primates)

1. Xing and Lee, *Nature Reviews Genetics*, 2006, 7: 499-510.
2. Xing and Lee, *PNAS*, 2005, 102(38): 13526 - 13531.
3. Calarco*, Xing*, Caceres*, et al, *Genes & Dev*, 2007, 21:2963-2975.

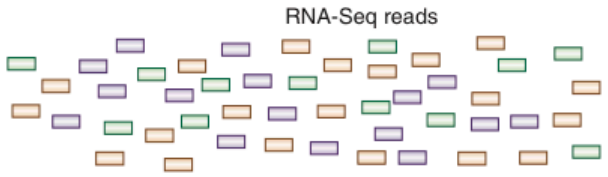
Types of splicing events



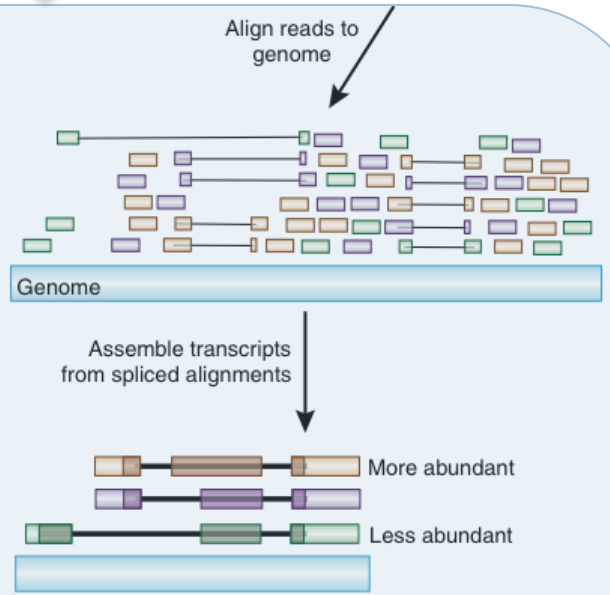
Ladd and Cooper 2002

From reads to transcripts

Haas & Zody Nature News & Views 2010

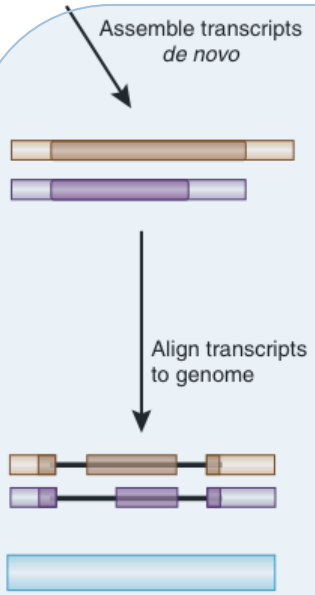


Align reads to genome



- Isoforms quantification
- Isoforms reconstruction

De novo Assembly



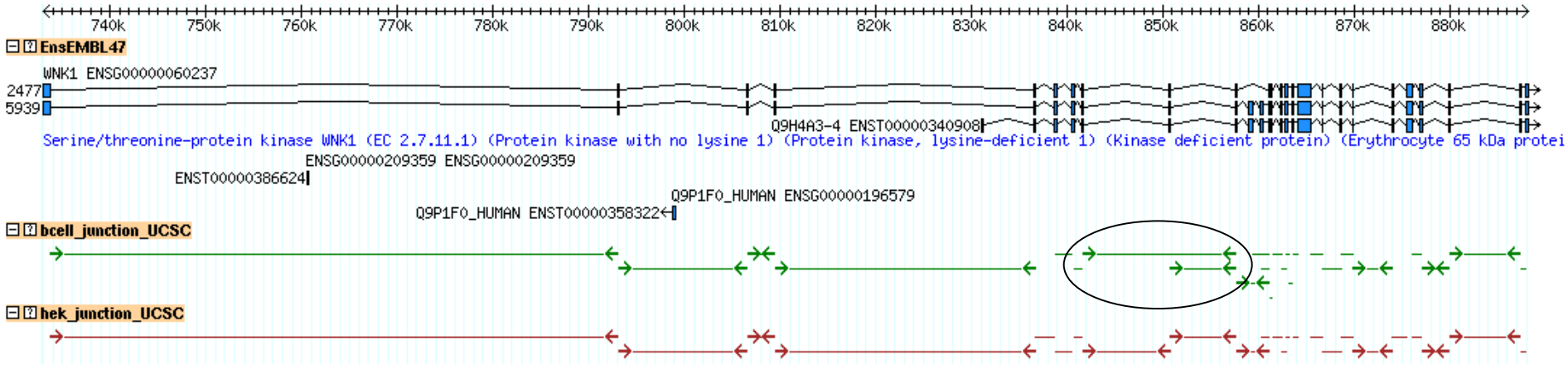
- Isoforms reconstruction

Outline

- Gene/Exon Expression level quantification
  - Sampling model and setup
  - Testing for differential expression

- Analysis of Alternative Splicing/Transcripts isoforms
  - **Spliced alignment**
  - Transcripts isoforms analysis
    - Transcripts isoforms quantification
    - Transcriptome reconstruction/*de novo* assembly

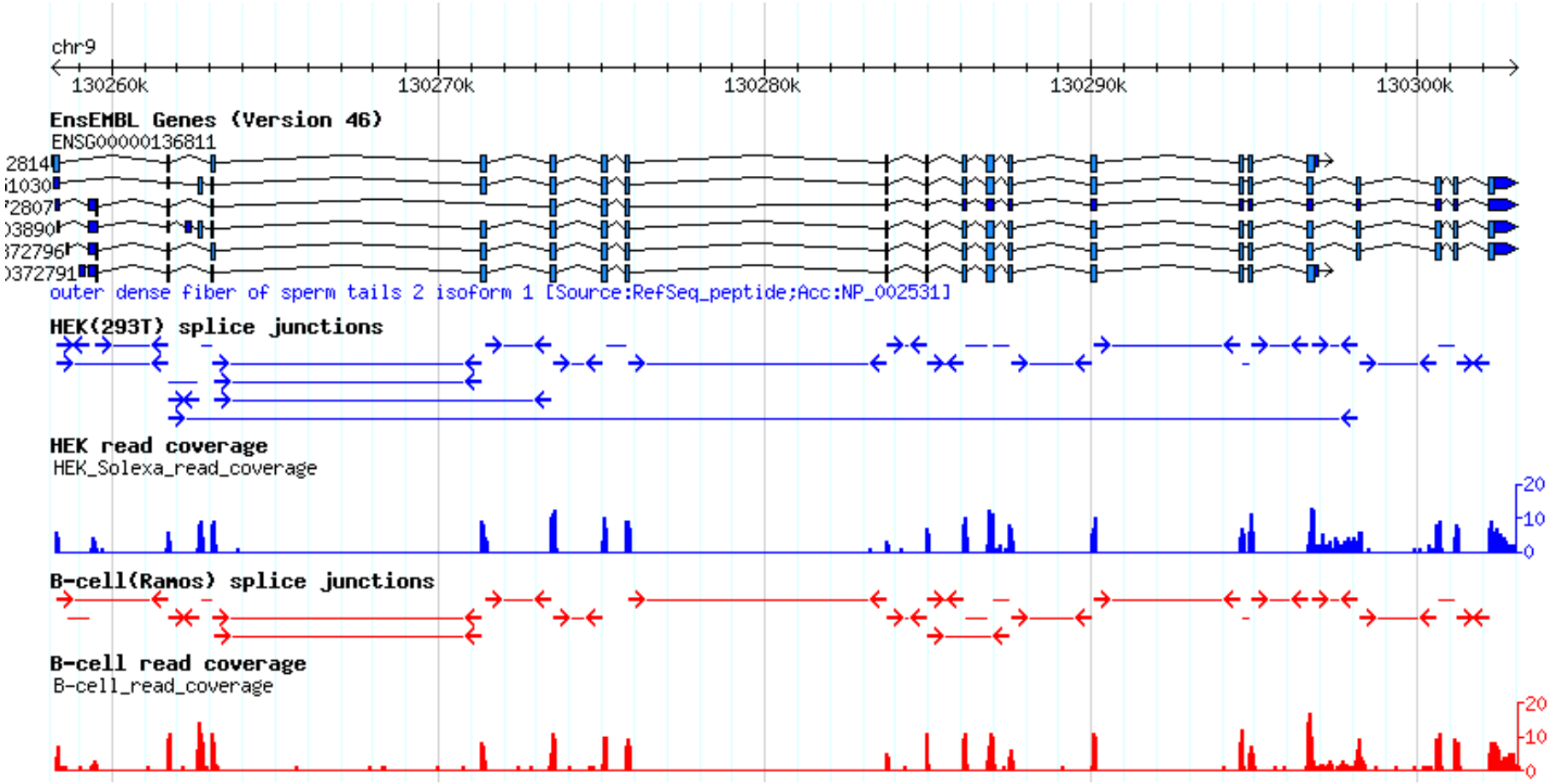
Identify from Splice junctions



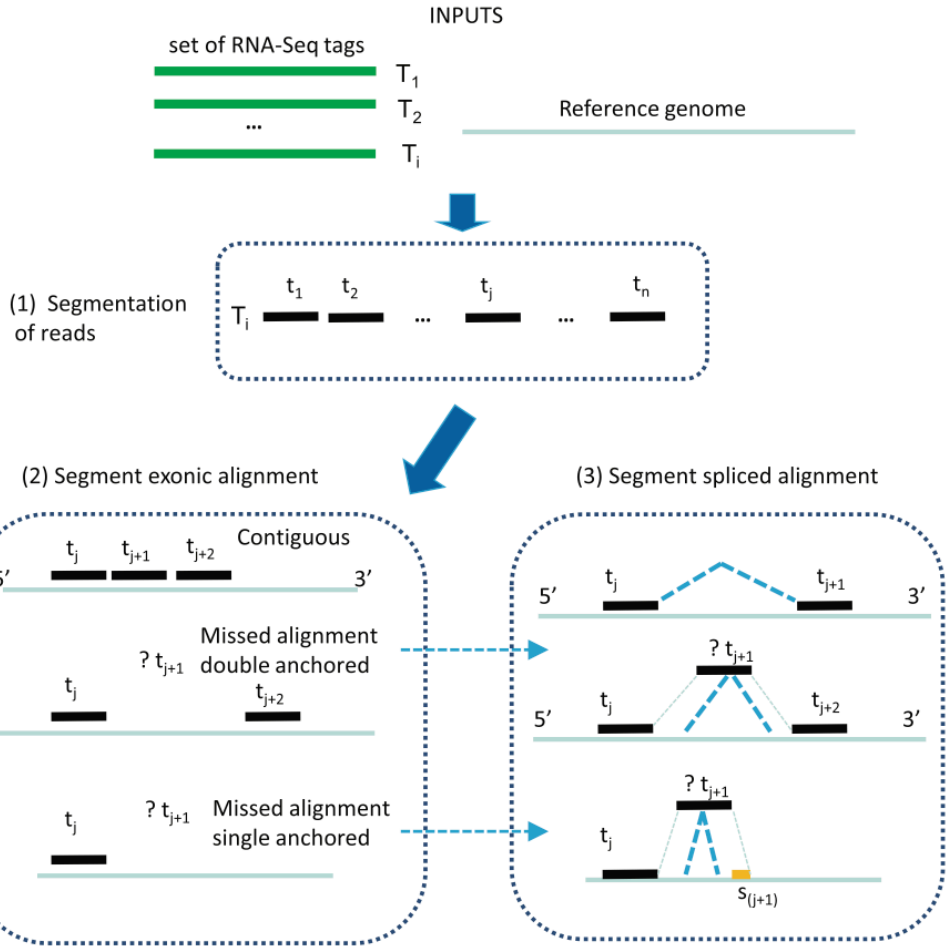
Map to a base of exons-exons junctions:

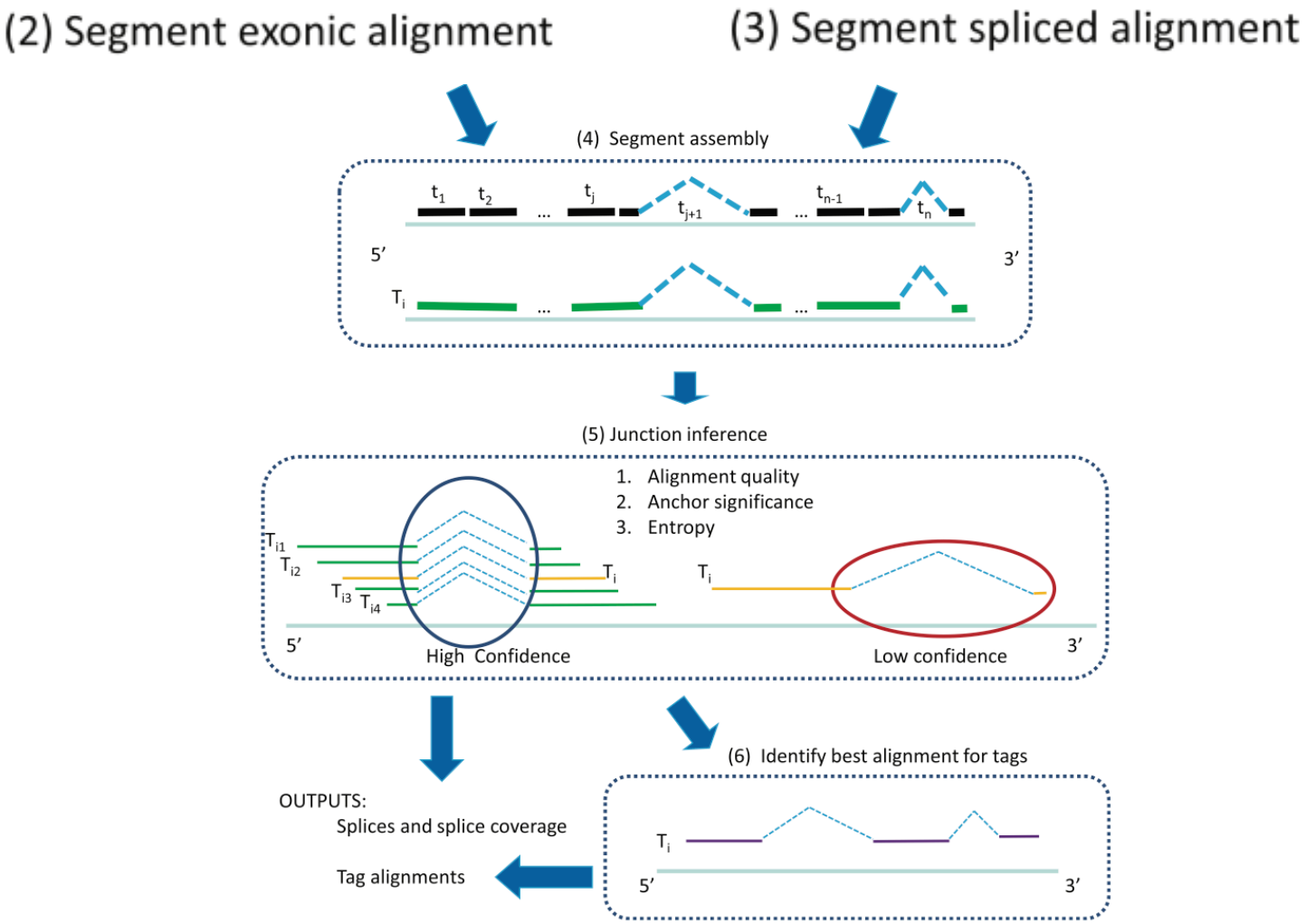
- Derived from genome annotation.
- Using the reads on exons as a guideline (TopHat).
- Perform alignment with arbitrary gap length, using pair end information, reads cut in half (SpliceMap, MapSplice).
- Use information from splice site in score function

Example junction



Aligning splice junctions (splicemap)



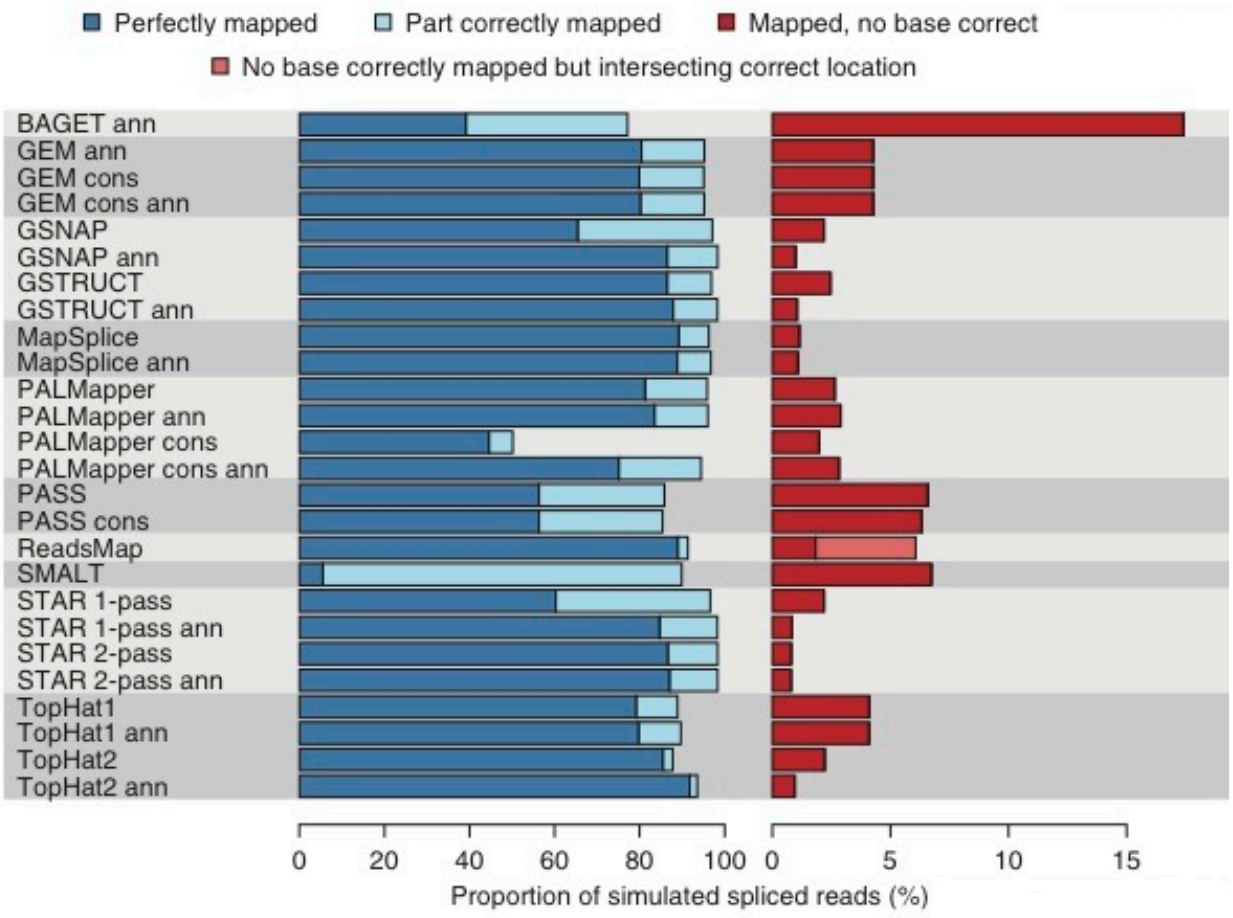

(2) Segment exonic alignment
(3) Segment spliced alignment
(4) Segment assembly
t1 t2 ... tj tj+1 ... tn-1 tn
5′
3′
Ti
(5) Junction inference
1. Alignment quality
2. Anchor significance
3. Entropy
Ti1
Ti2
Ti3 Ti4
Ti
Ti
5′
High Confidence
Low confidence
3′
(6) Identify best alignment for tags
Ti
5′
3′
OUTPUTS:
Splices and splice coverage
Tag alignments

RGASP-Benchmark splice aligners

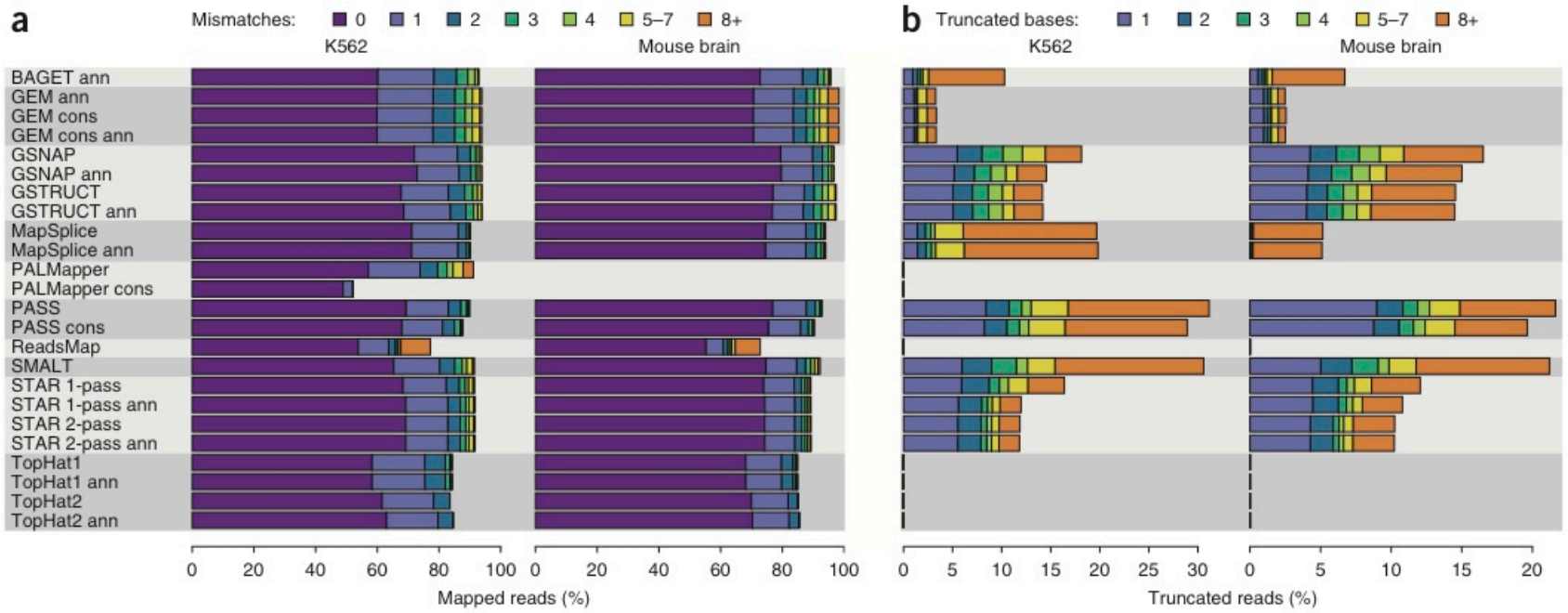
- RNA-Seq Genome Annotation Assessment Project
  - 9 teams submitted 26 alignment protocols
    - 2 real (human cell line, mouse)
    - Two simulated datasets (1 easy and 1 complex)
- Softwares:
    - GSNAP, MapSplice, PALMapper, ReadsMap, STAR, TopHat.
    - GEM, PASS, GSTRUCT, BAGET.
    - SMALT (contiguous aligner not tuned for splice junctions)

Engström et al 2013

Read placement



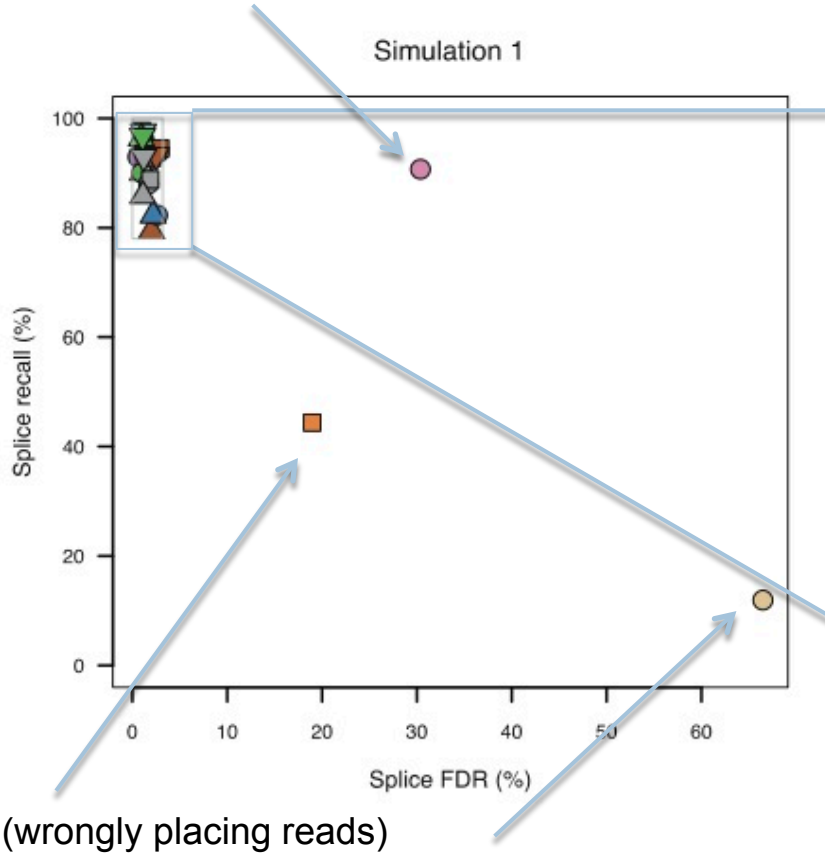
Mismatches


a
Mismatches: 0 1 2 3 4 5–7 8+
K562
Mouse brain
BAGET ann
GEM ann
GEM cons
GEM cons ann
GSNAP
GSNAP ann
GSTRUCT
GSTRUCT ann
MapSplice
MapSplice ann
PALMapper
PALMapper cons
PASS
PASS cons
ReadsMap
SMALT
STAR 1-pass
STAR 1-pass ann
STAR 2-pass
STAR 2-pass ann
TopHat1
TopHat1 ann
TopHat2
TopHat2 ann
0 20 40 60 80 100
0 20 40 60 80 100
Mapped reads (%)
b
Truncated bases: 1 2 3 4 5–7 8+
K562
Mouse brain
0 5 10 15 20 25 30
0 5 10 15 20
Truncated reads (%)

Evaluation

READSmap (overestimating indels)

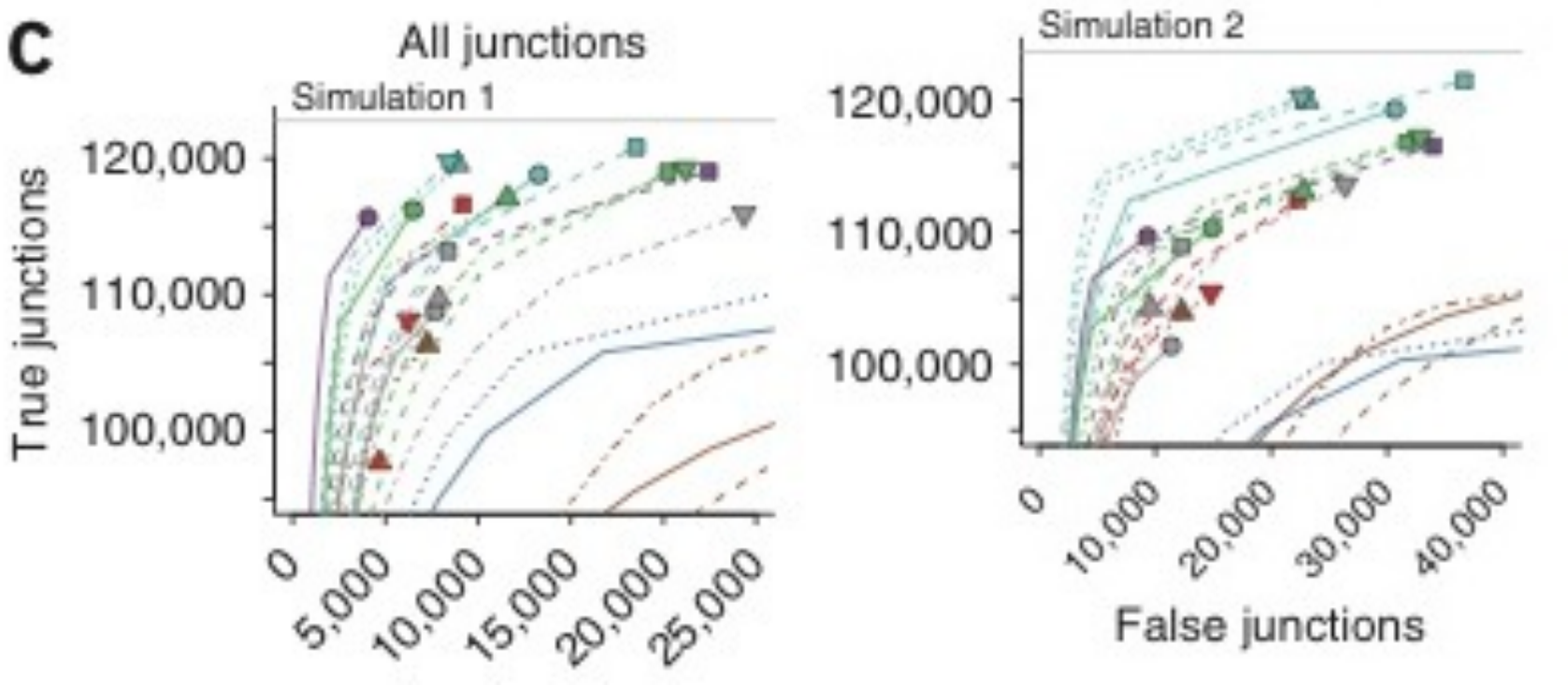
GSNAP, GSTRUCT



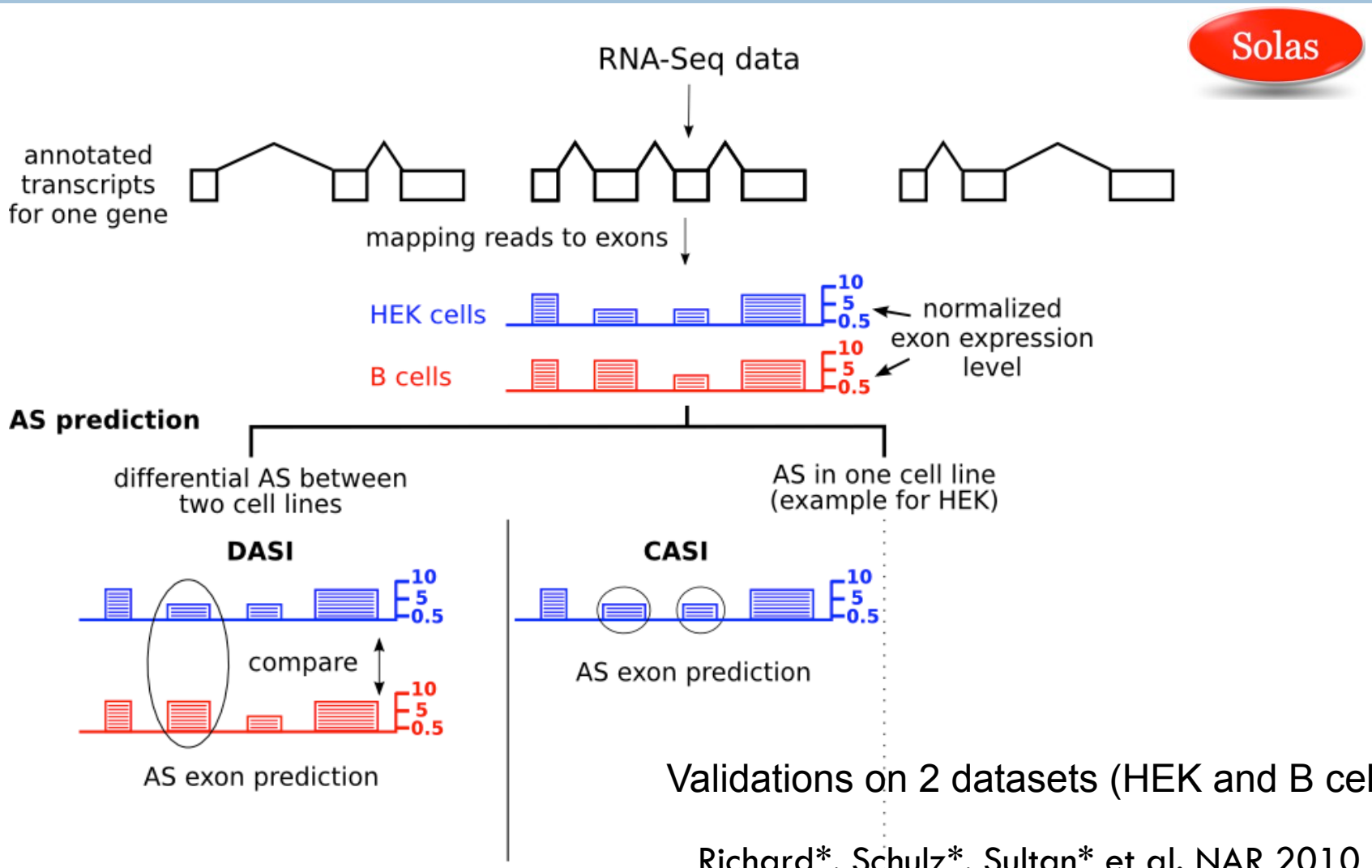
Baget (wrongly placing reads)

SMALT (missing annotated junc., wrongly placing reads)

Tophat


C
All junctions
Simulation 1
True junctions
120,000
110,000
100,000
0
5,000
10,000
15,000
20,000
25,000
Simulation 2
120,000
110,000
100,000
0
10,000
20,000
30,000
40,000
False junctions

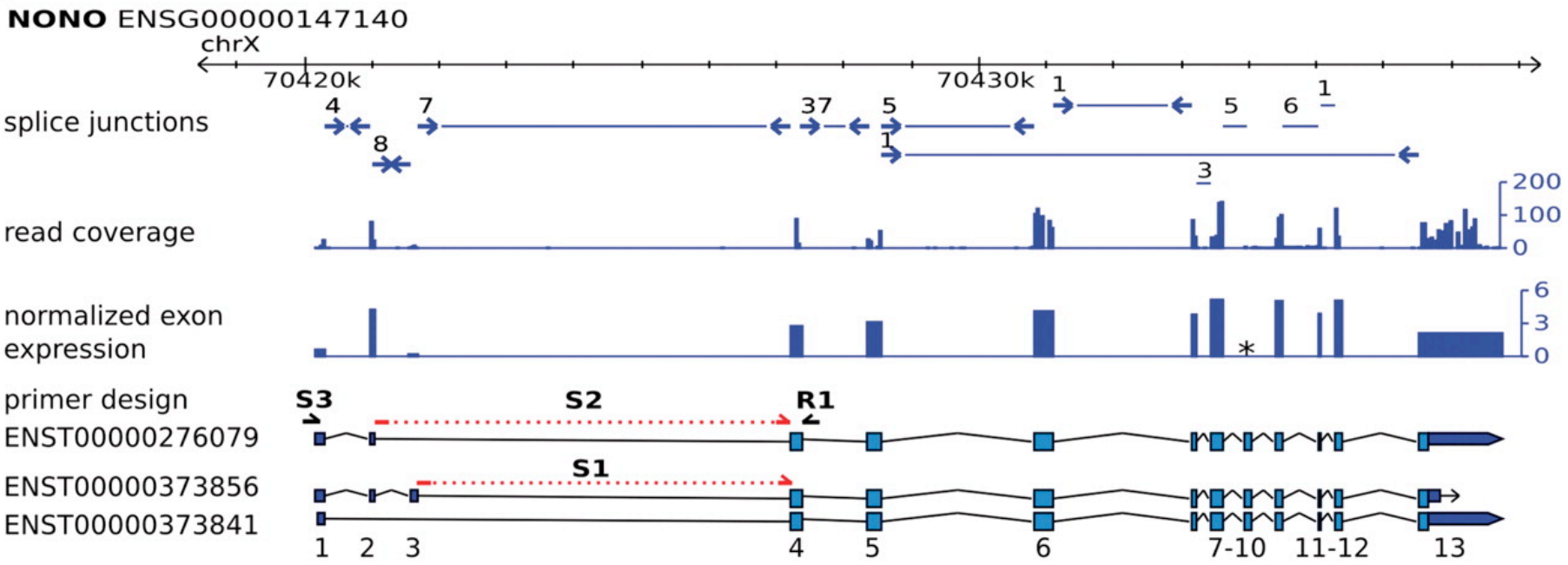
Detecting AEE from Exons Expression

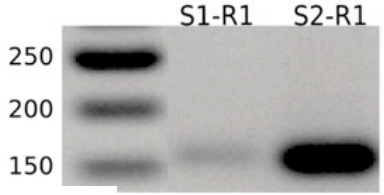


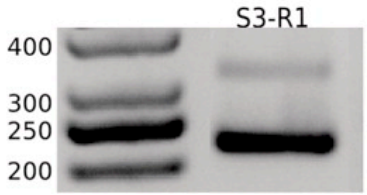
Validations on 2 datasets (HEK and B cells)

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

Within cell AS : CASI

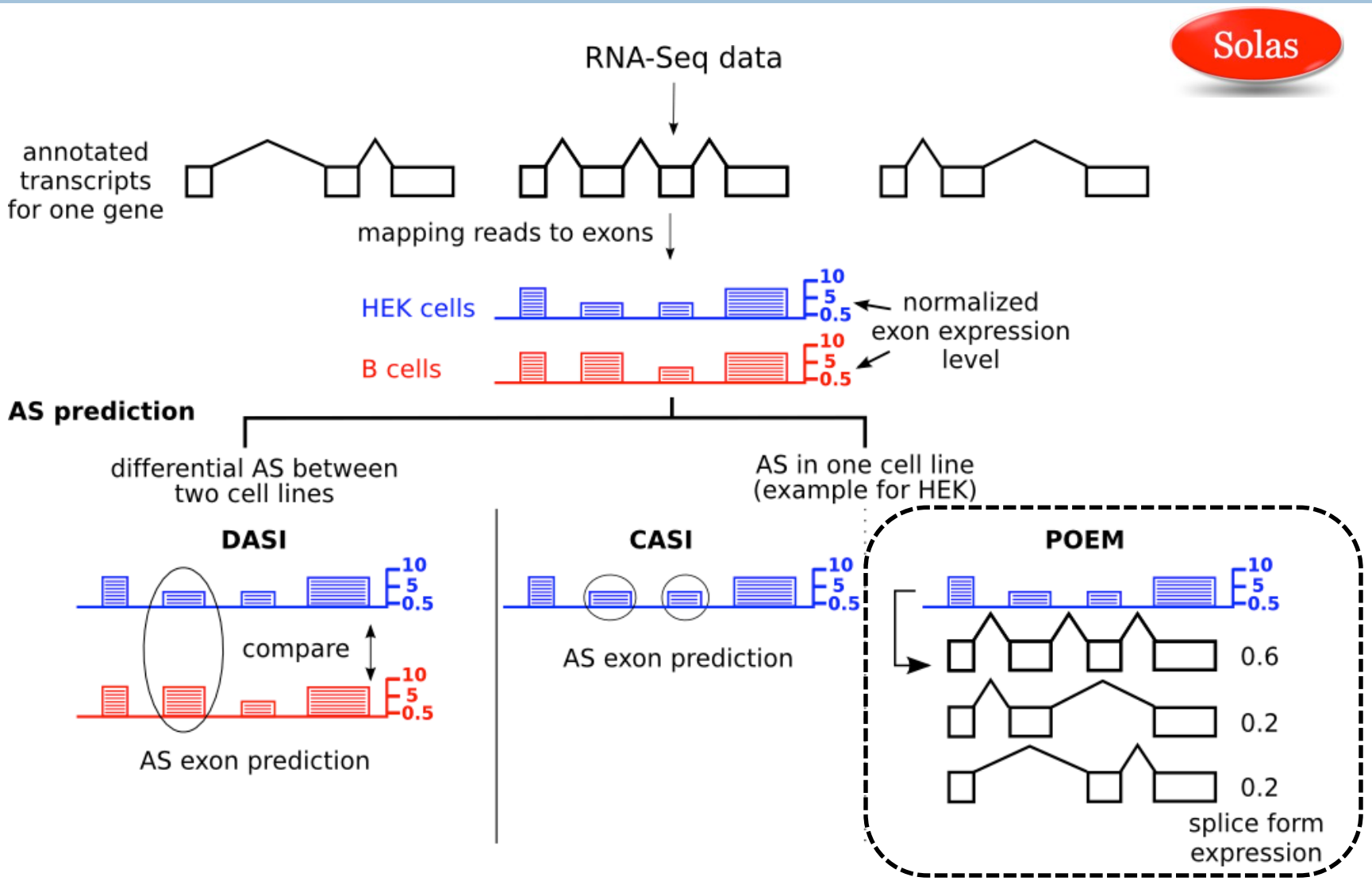






61 events tested
35 validated

Detecting AEE from Exons Expression



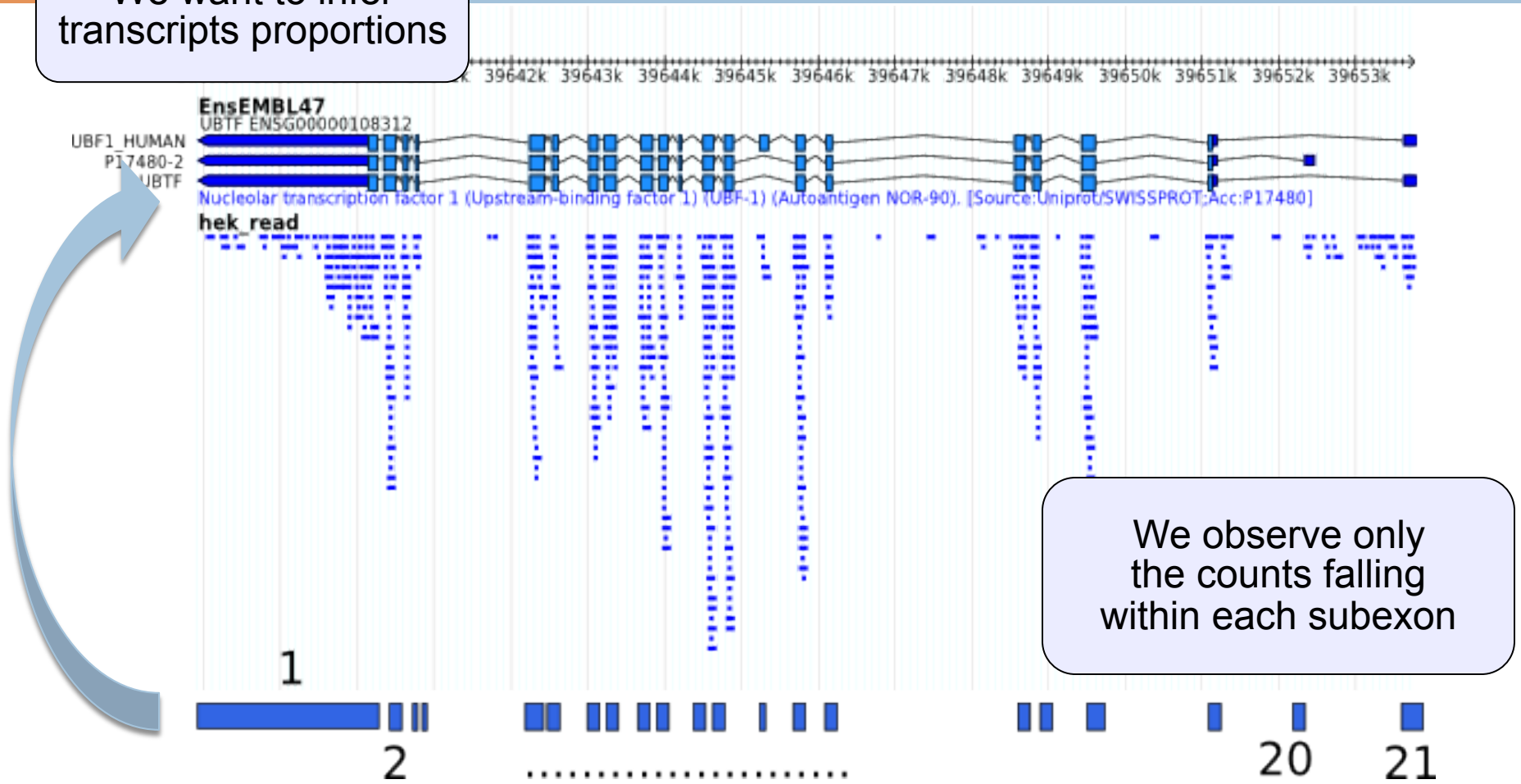
Outline

- Gene/Exon Expression level quantification
  - Sampling model and setup
  - Testing for differential expression

- Analysis of Alternative Splicing/Transcripts isoforms
  - Spliced alignment
  - **Transcript isoforms analysis**
    - Transcripts isoforms quantification
    - Transcriptome reconstruction/*de novo* assembly

Transcripts quantification

We want to infer transcripts proportions

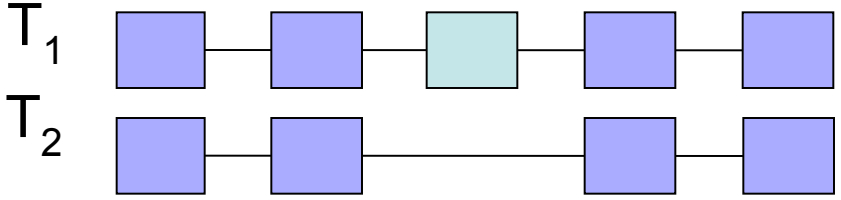


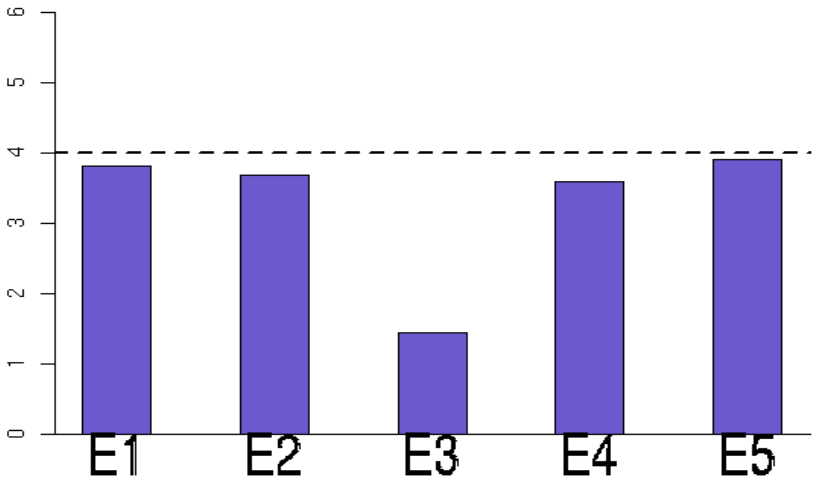
We observe only the counts falling within each subexon

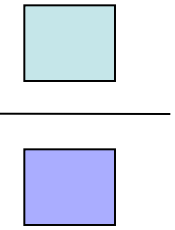
Transcripts quantification

Add the transcript structure:

- Deduce directly transcript proportions
- Like solving linear equations (but with variability)





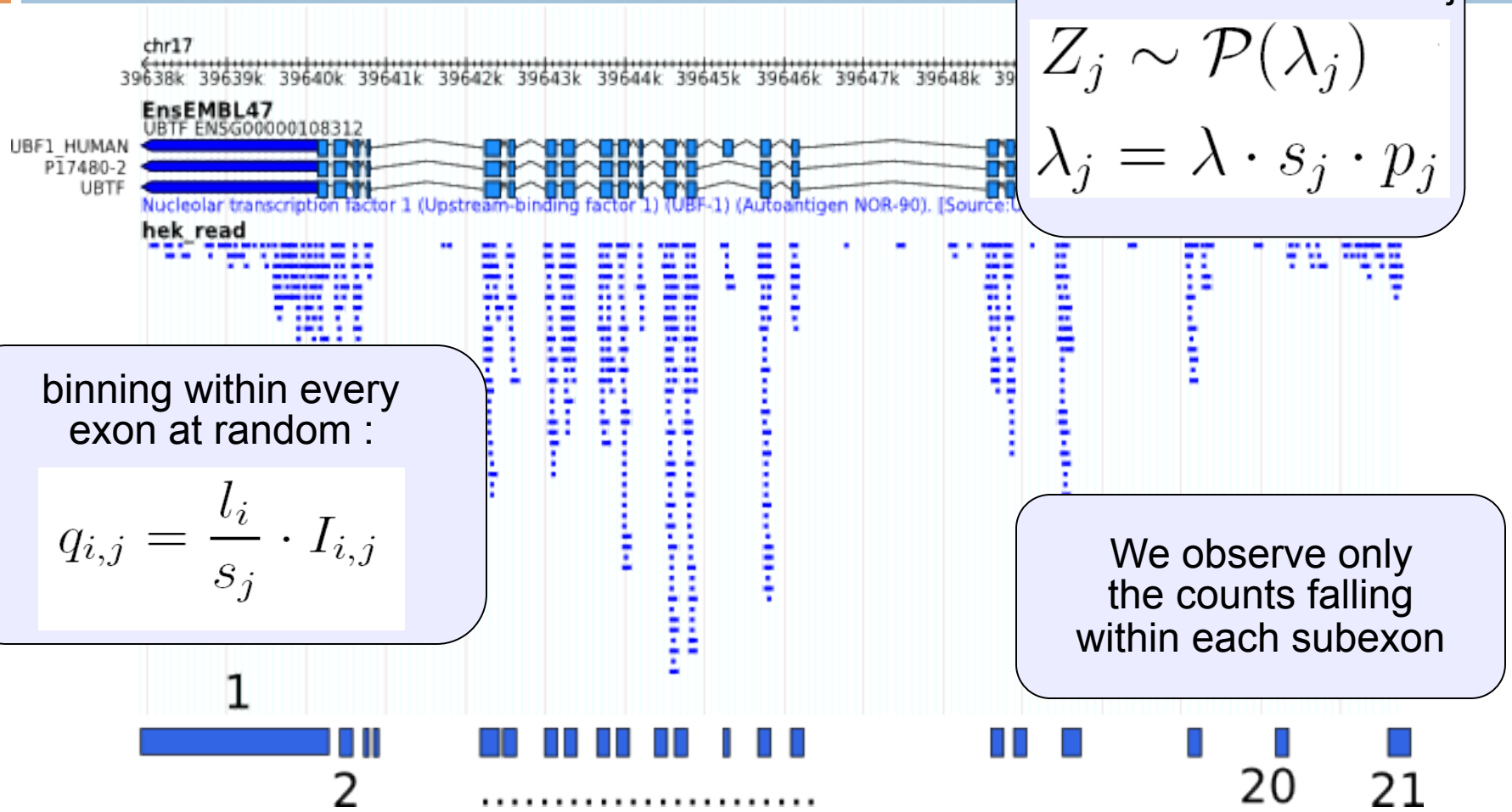
$T_1\% \sim$ 

Transcripts quantification: EM

Count for each isoform j

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

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

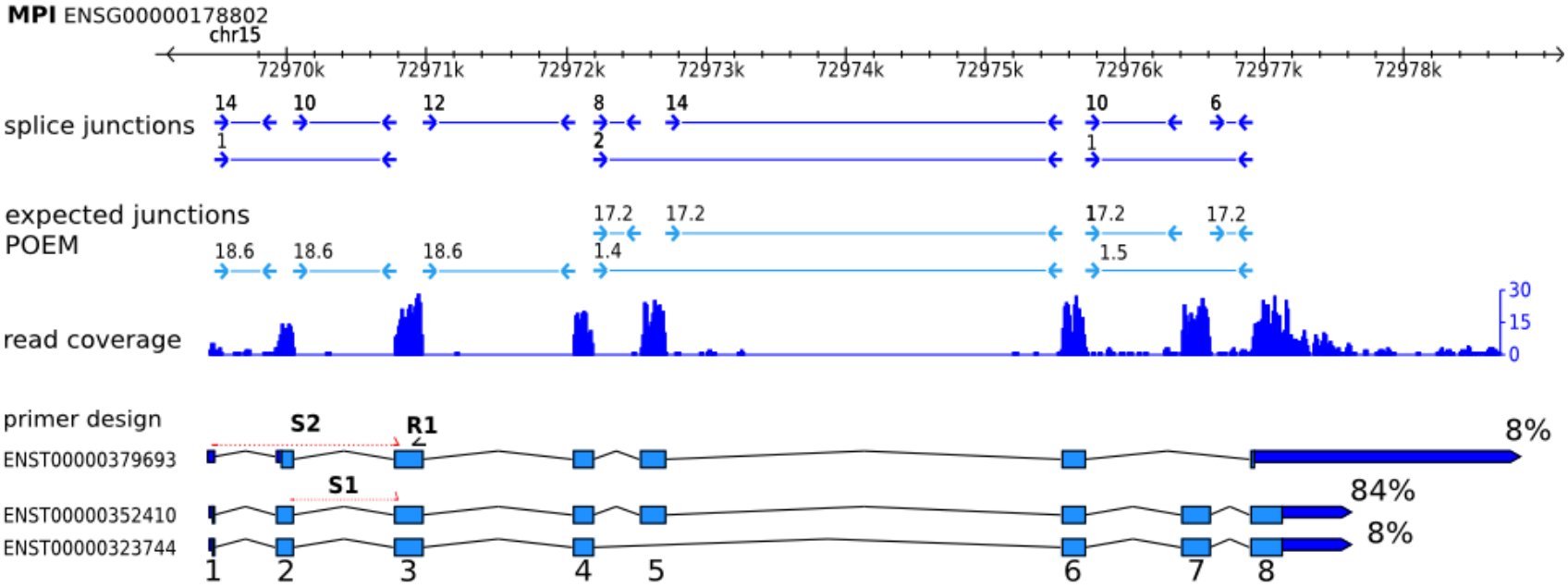


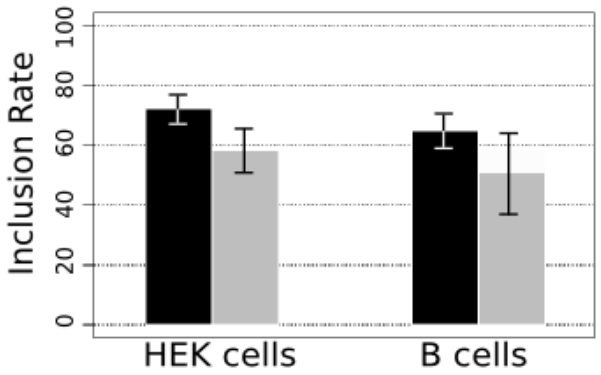
binning within every exon at random :

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

We observe only the counts falling within each subexon

Isoform quantification: POEM





Comparison to:

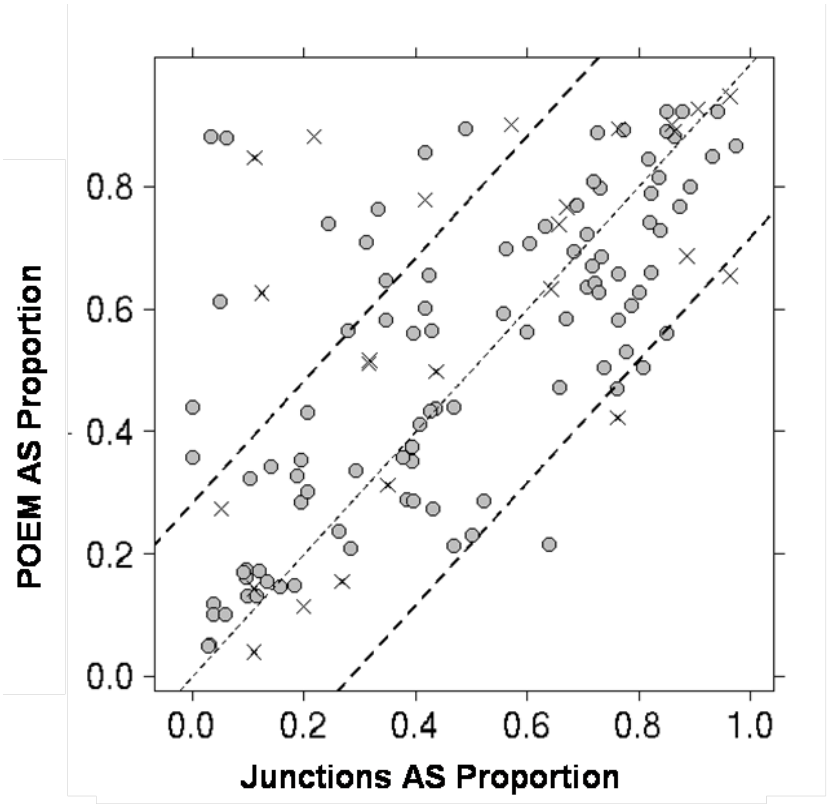
_ qPCR (47 events)

_ Estimate from junction counts (267 events)

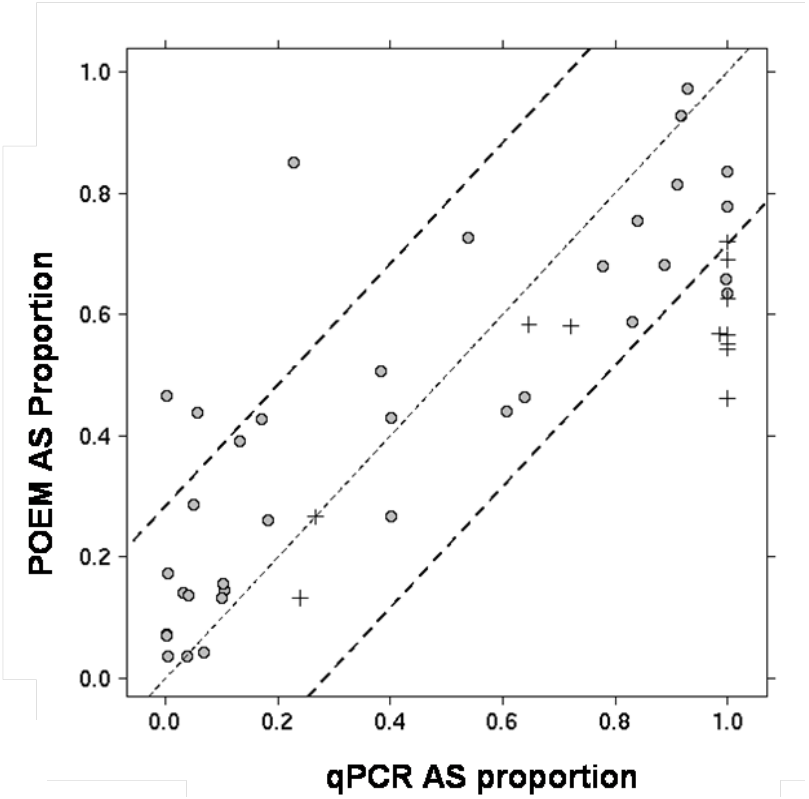
grey - POEM
black - qPCR

Validation

Junction reads



qRT-PCR



PCC: 0.65 0.81

Transcript Quantification now

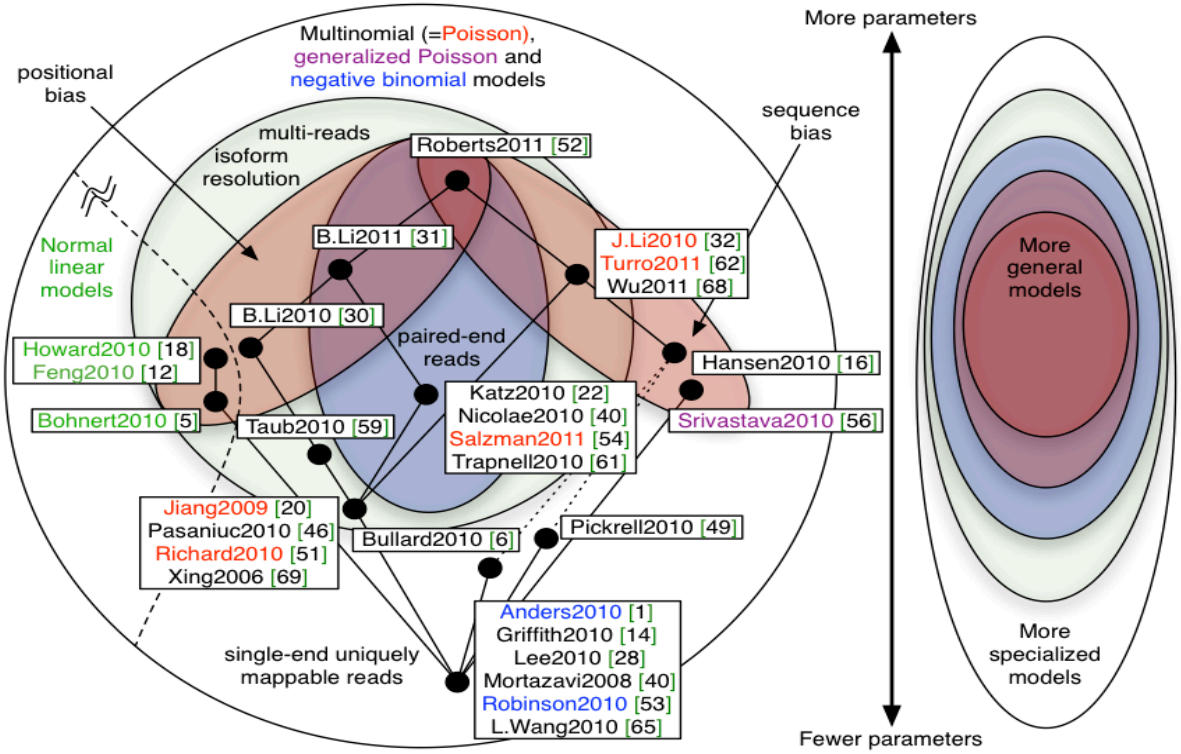


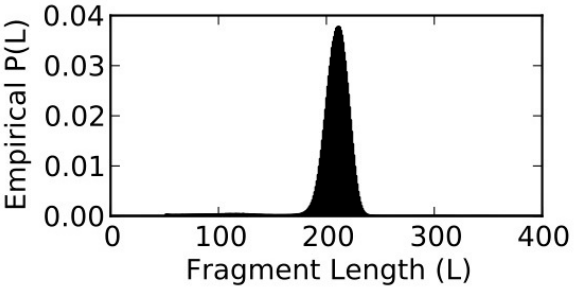
FIGURE 1. Models for RNA-Seq. The figure shows a Venn diagram (and the partially ordered set it induces) representing relationships among models. More general models are nested inside simpler models.

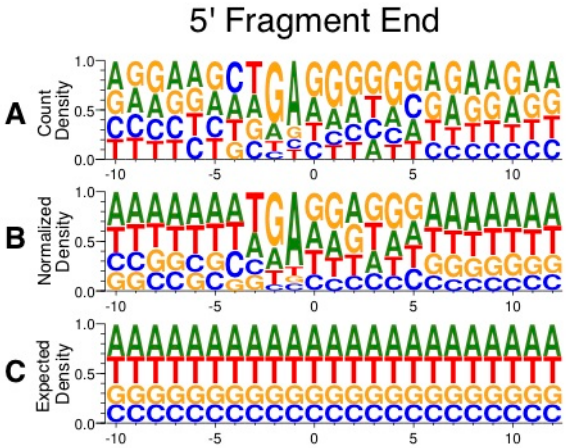
Pachter, ArXiv 2011

Multiple improvements lately :

- Multi-reads
- Sequence/longitudinal artifacts
- Paired-end reads

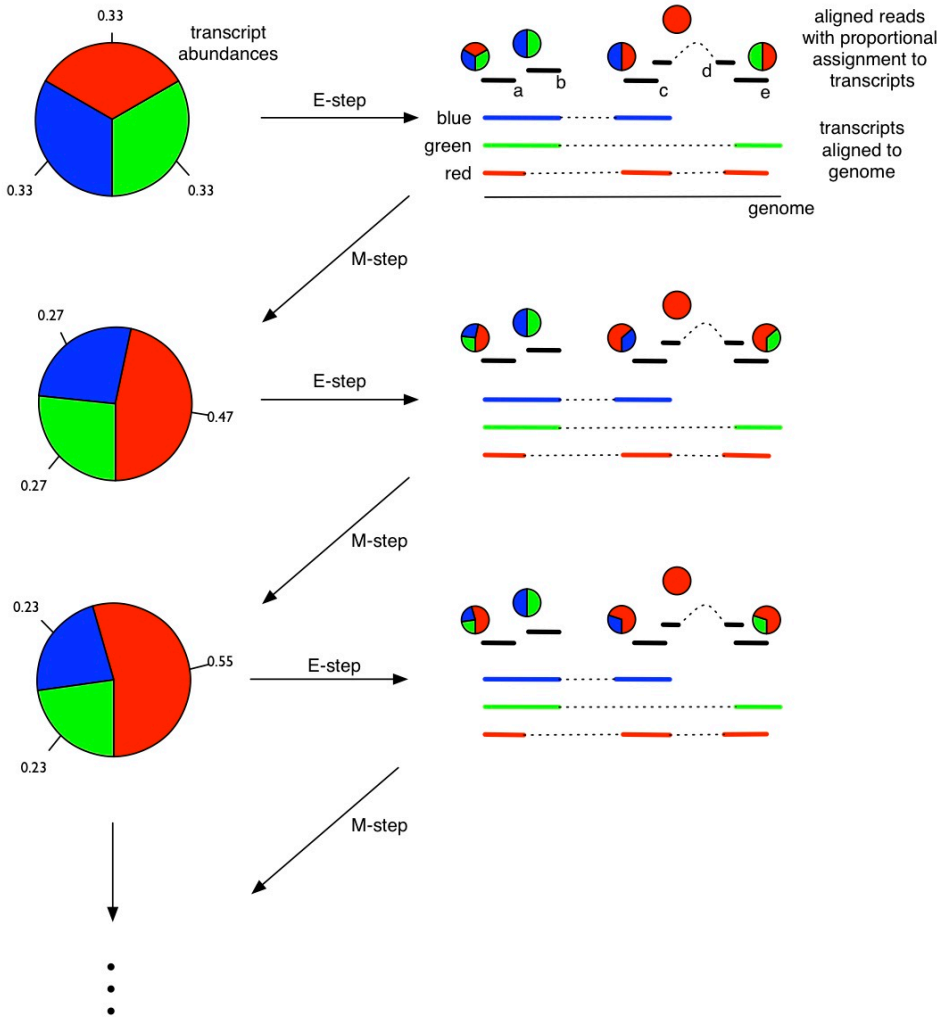
RSEM and other





- Incorporate library artifacts
  - fragment length distribution
  - positional biases
  - context information
- Model fragments to transcripts assignations
  - RSEM, Cufflinks, Express, IsoEM, Ireckon…
  - EM to estimate parameters and transcripts abundances

Batch EM for transcript proportions

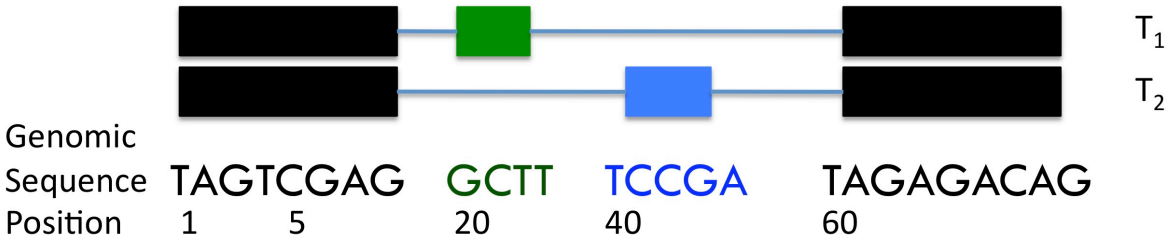


Pachter, ArXiv 2011

Transcripts reconstruction

- The reconstruction problem is non trivial
  - Combination of AS events can be **non identifiable** given the limited read length.
    - Paired end information is critical
    - Enumerating all possible isoforms is exponential in the number of AS events
  - Need a regularization strategy (Parcimony, Lasso)
- Can be formulated:
  - *ab initio* with a genome (Cufflinks, Ireckon, Slider, IsoLasso…)
  - *de novo* assembly (Trinity, Oases, TransAbyss…)

Splicing graph



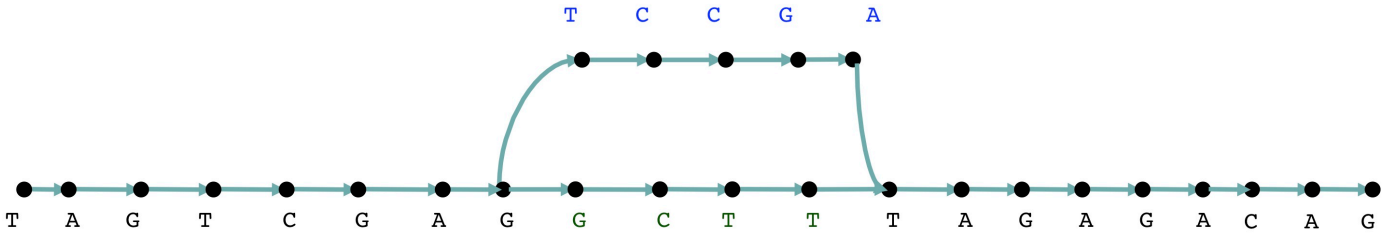
T_1= { 1,2,3,4,5,6,7,8,20,21,22,23,60,61,62,63,64,65,66,67,68 }

T_2= { 1,2,3,4,5,6,7,8,40,41,42,43,60,61,62,63,64,65,66,67,68 }, T_1 and T_2 on the forward strand!

Definition: Splicing graph G=(V,E) (G is a Directed Acyclic Graph)

$V = T_1 \cup T_2$

edge (v,w) in E, if v and w are consecutive positions in T_1 or T_2



Splicing graph over genomic positions of T_1 and T_2
with representative letter at that position as node label

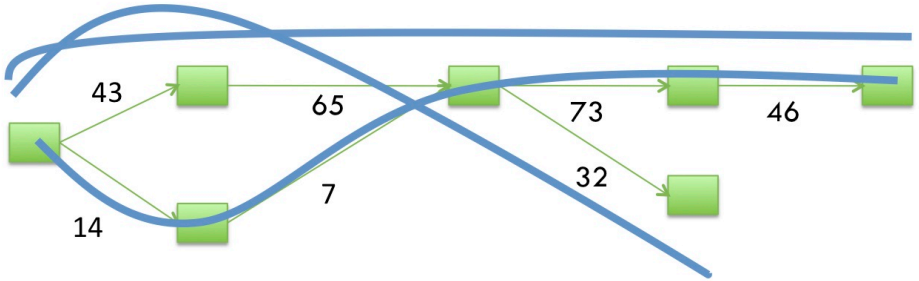
Heber et al. Bioinformatics, 2001

Courtesy M.Schulz

Reconstruct transcripts

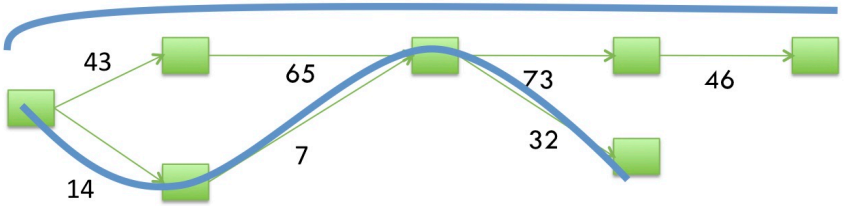
- Maximum Likelihood approach:

  assume independent AS event and a majority of constitutive exons



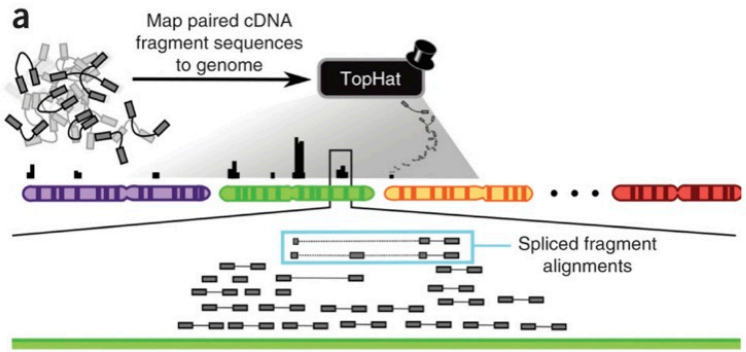
- Maximum parsimony

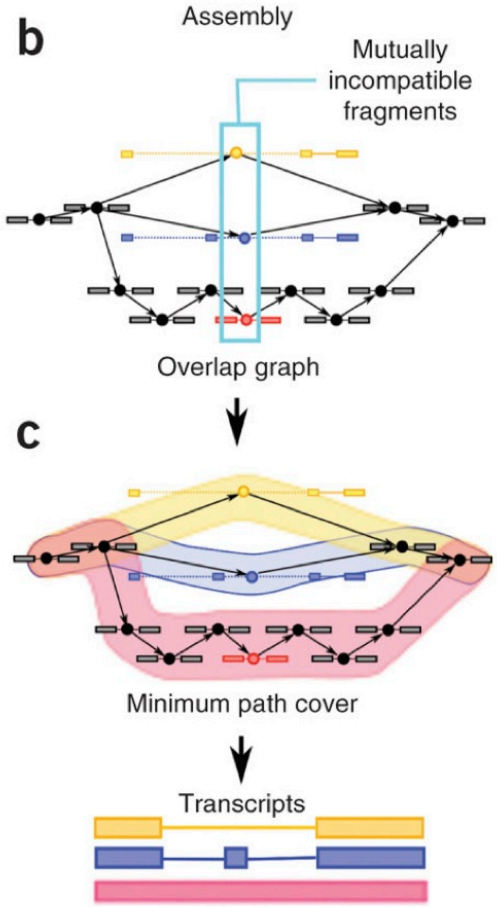
  constrain the number of transcripts



Cufflinks

spliced alignment of paired-end reads to genome





modified Trapnell et al. 2010

Comparing each approach

Approach	Advantages	Disadvantages
ab initio Reference based	- alignment tolerate seq. errors - repeat detected through alignment - grouping by genomic proximity	- need the reference seq. - assumes collinearity of the transcripts with the genome
de novo	- no reference needed - detection of non-collinear transcripts - handles micro-exons	- lowly expressed genes are indistinguishable from seq. errors - missassemblies due to repeats

Courtesy M.Schulz

Comparing approaches on mouse RNA-Seq

Method	transfrag > 100 bps	N50	total in mb	Nucleotide Sensitivity			Nucleotide Specificity
				all genes	lowly exp. genes RPKM < 1	highly exp. genes RPKM > 20	
Cufflinks*	72,745	2,613	~ 73	45.3	24.1	71.4	67
Oases[$]	73,357	1,287	~ 64	28.3	0.9	64.8	85.3

[k]Trapnell *et al. Nat. Biotech.* 2010
[$]Schulz, Zerbino et. al, submitted

Results for Ensembl 57 annotation

Courtesy M.Schulz

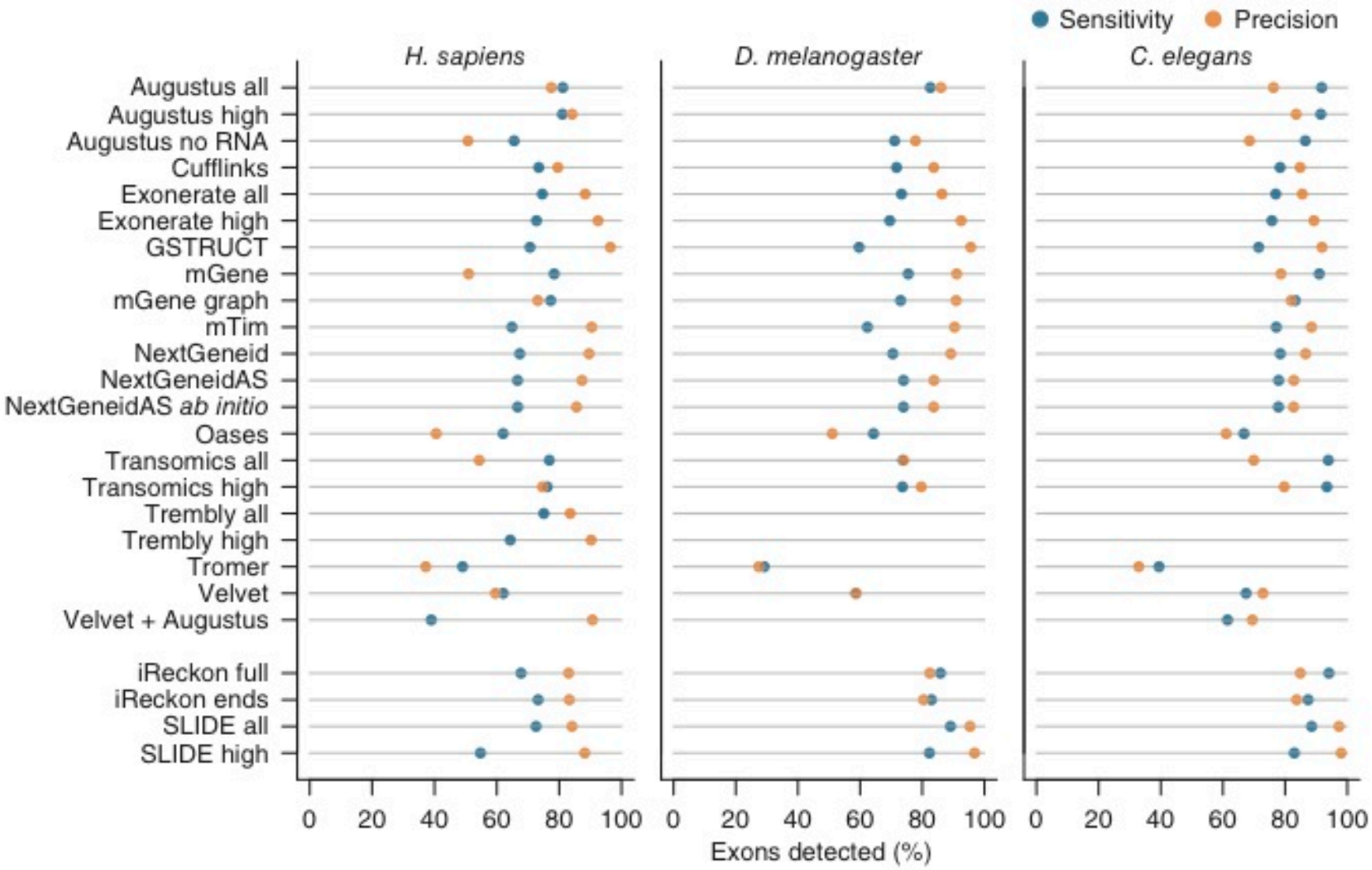
RGASP – transcript reconstruction

- Evaluate 25 transcripts reconstruction methods:
  - Ab initio methods:

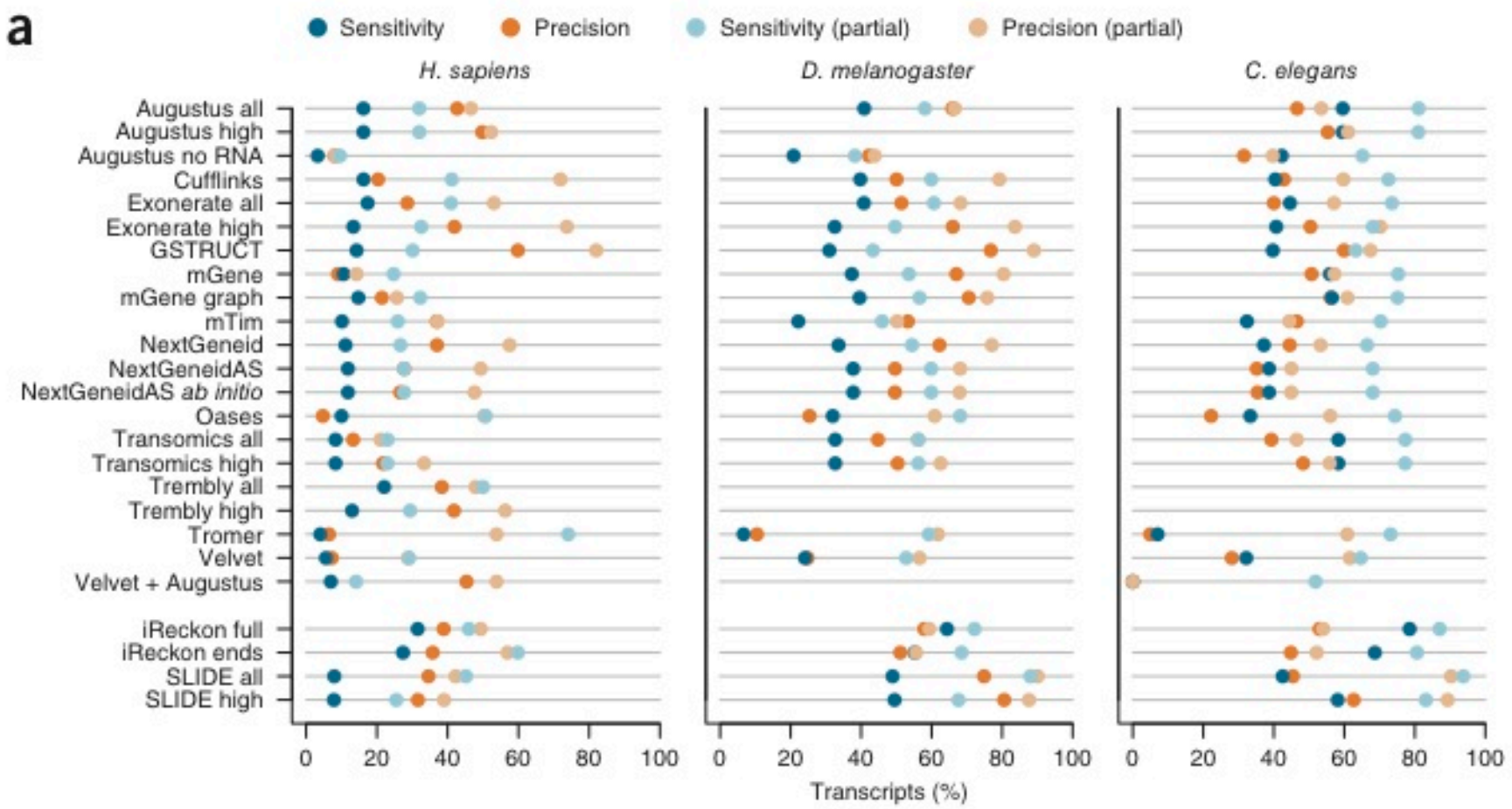
    Cufflinks, iReckon, SLIDE,Augustus, mGene, Transomics
  - De novo:

    Oases, Velvet + Augustus.

results- exon level

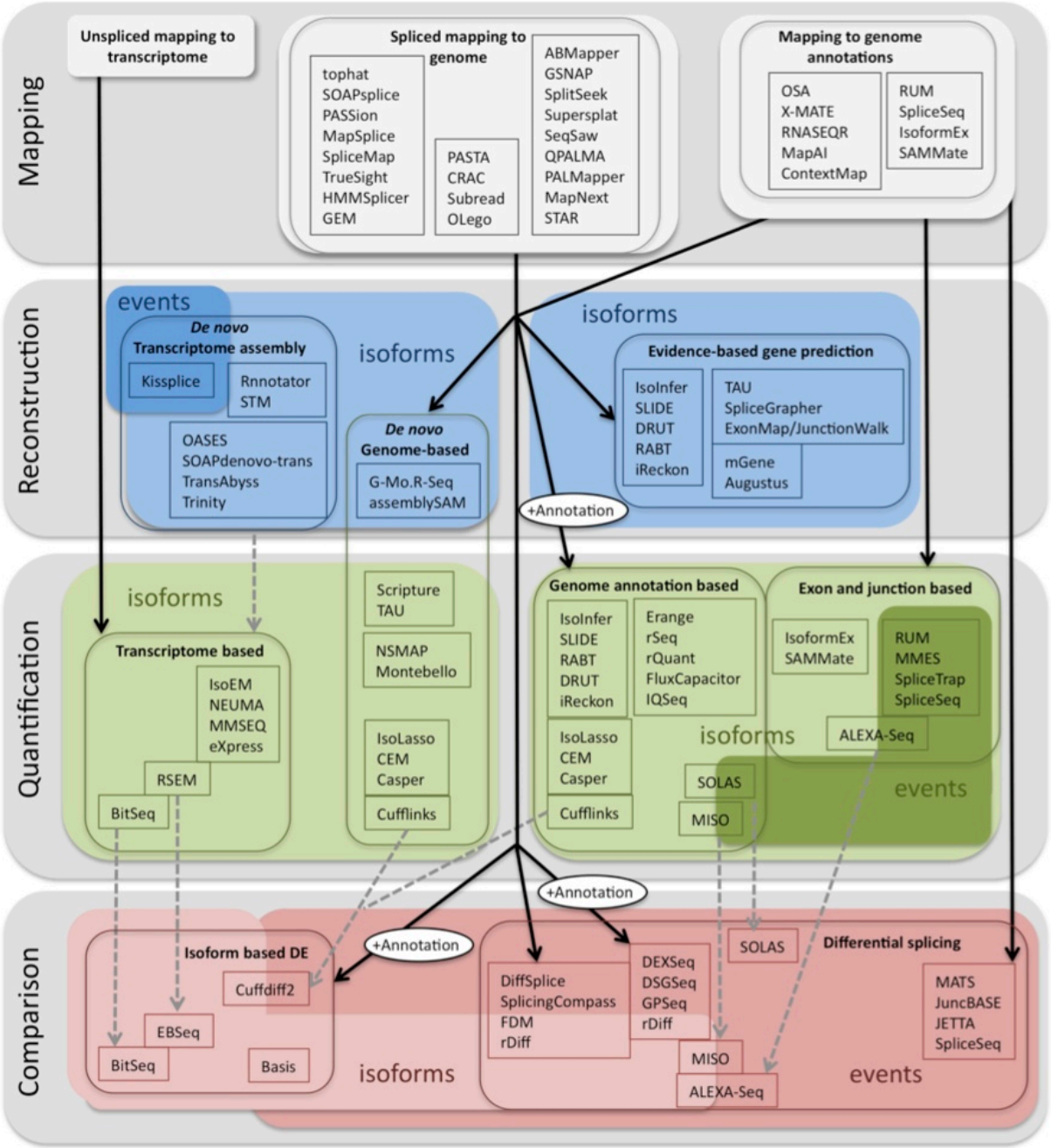


results- transcript level



Summary transcriptome

- Transcriptome is more complicated than the genome
  - Often still use a genome proxy
  - Quantification from digital data (counts)
  - Identification of exons boundary is a special alignment problem with large gaps
- Transcripts quantification and reconstruction:
  - Inferring transcript proportions: linear system or models with Hidden variables
  - When transcripts are unknown, reconstructing transcript structure is an additional challenge.



Alamancos et al. 2013