Bioinformatics for everyone

Gentle introduction to RNA-seq

Konstantin Okonechnikov
Max Planck Institute For Infection Biology

Летняя школа биоинформатики
Москва, 2013

Lecture plan

- Biology primer
- RNA-seq technology
- Experiment types
- Gene expression studies
- Spliced alignment
- Quantification and normalization of gene counts
- Novel transcript inference
- Fusion genes discovery

Biology: central dogma

- DNA- basic inheritance material
- Central dogma:

  DNA → RNA → protein
- Components involved: DNA polymerase (replication), RNA polymerase (transcription), ribosome (translation)

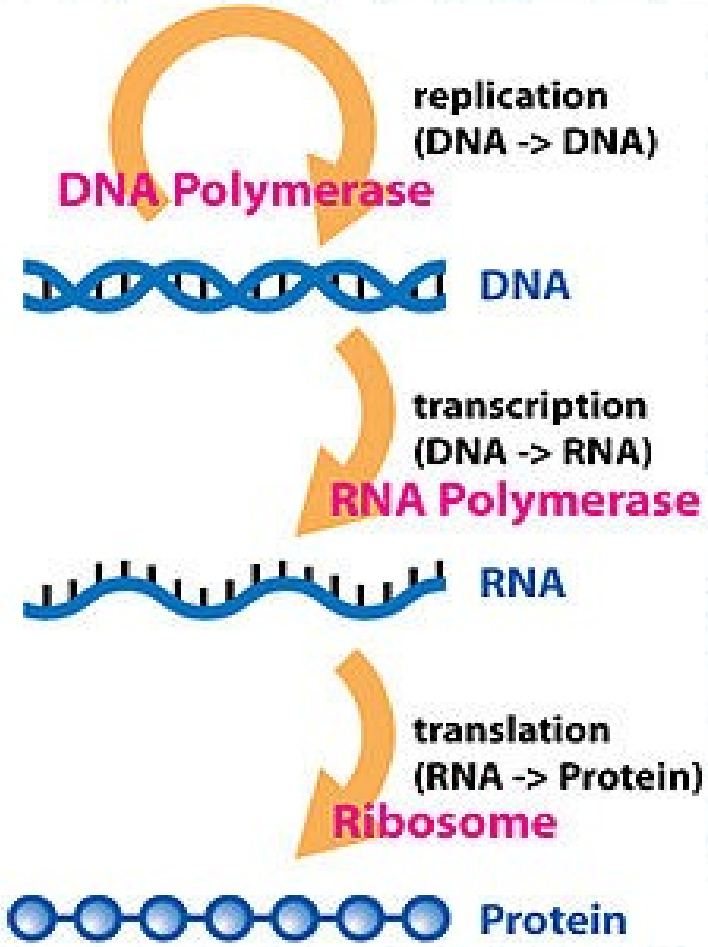
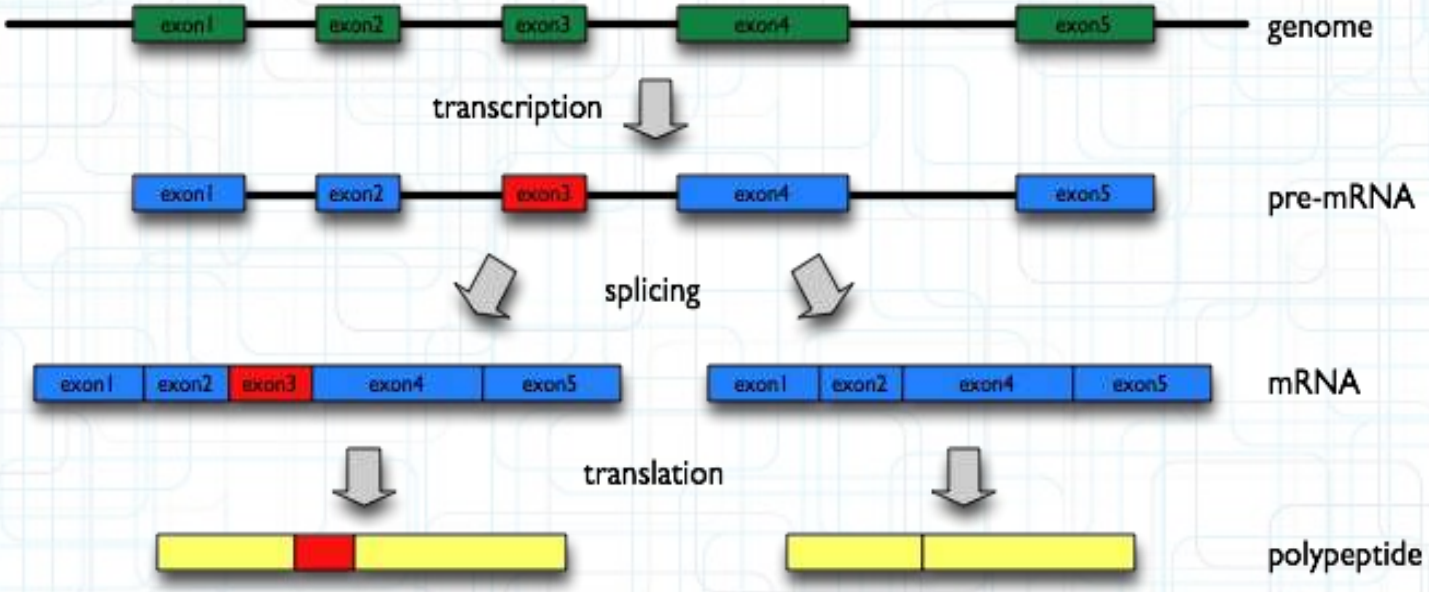


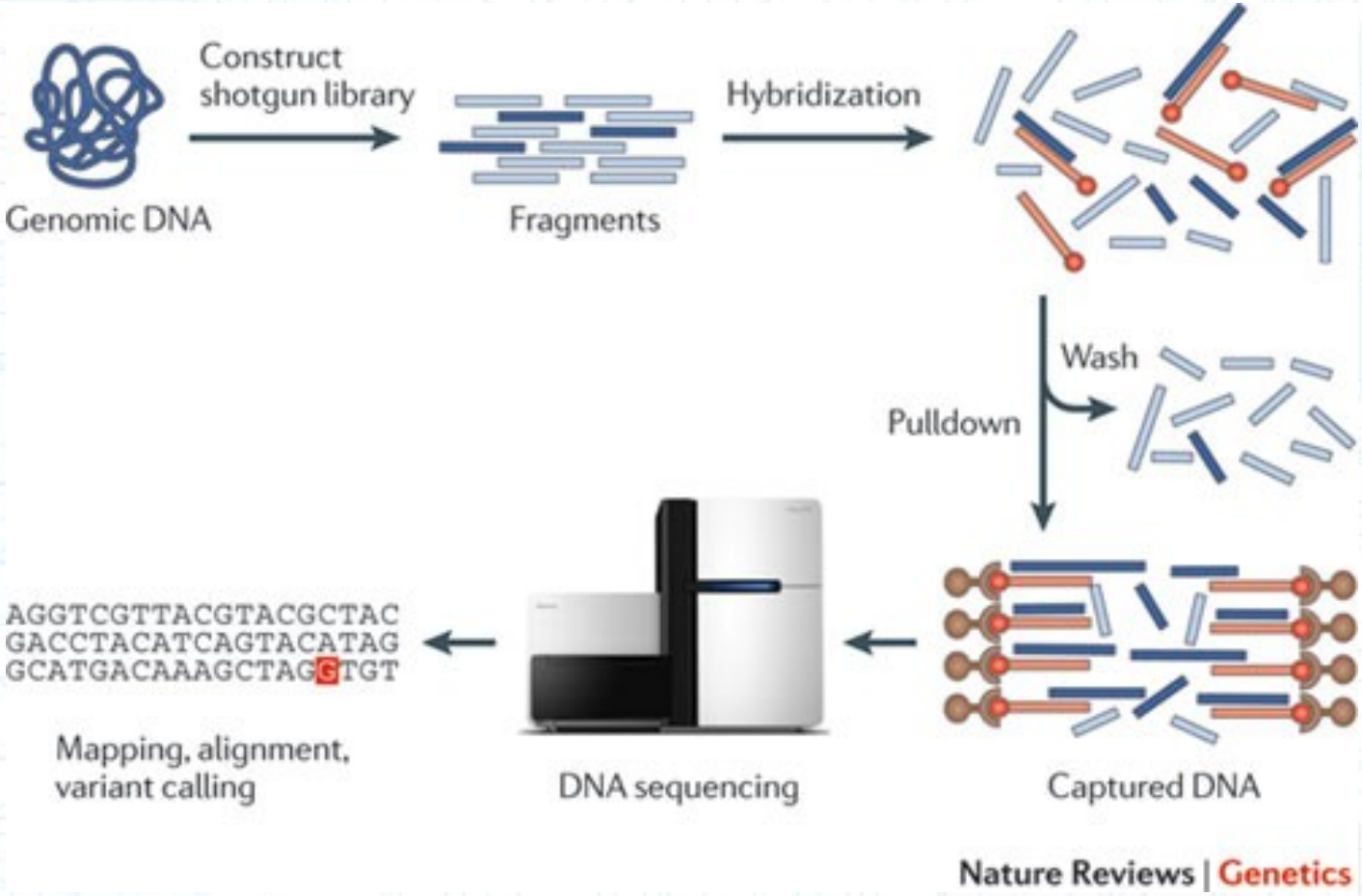
Image source: Wikipedia

Biology: gene structure

- Exons and introns
- Junctions
- Alternative splicing

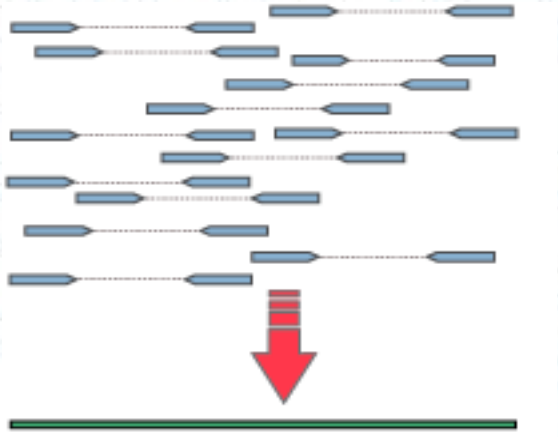


Next Generation Sequencing



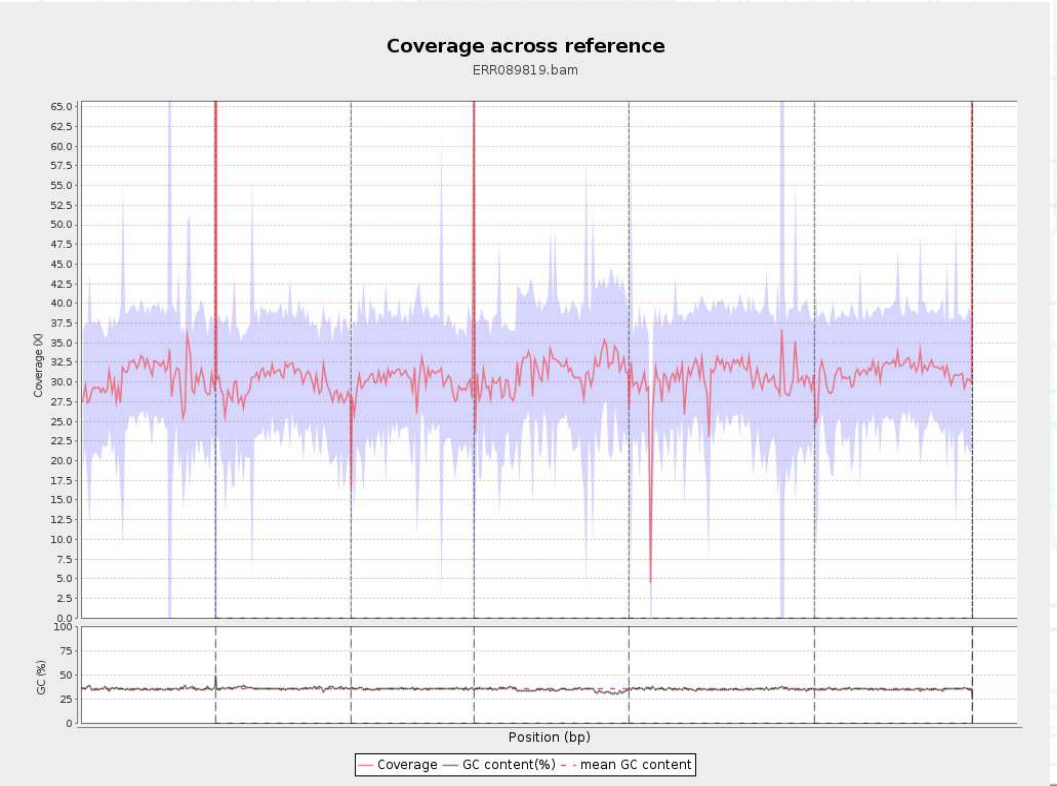
NGS: coverage

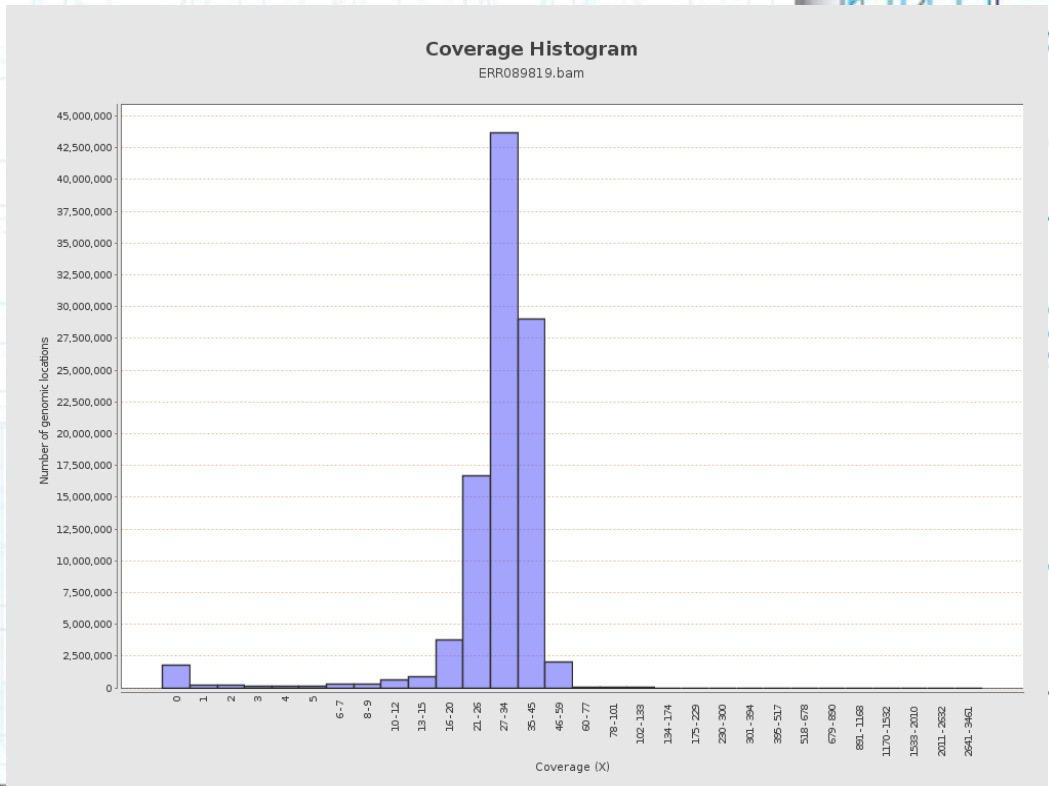
- Coverage: average number of times a genomic base is represented
- In case of paired reads represented as whole fragment (i.e. also counting unsequenced bases)



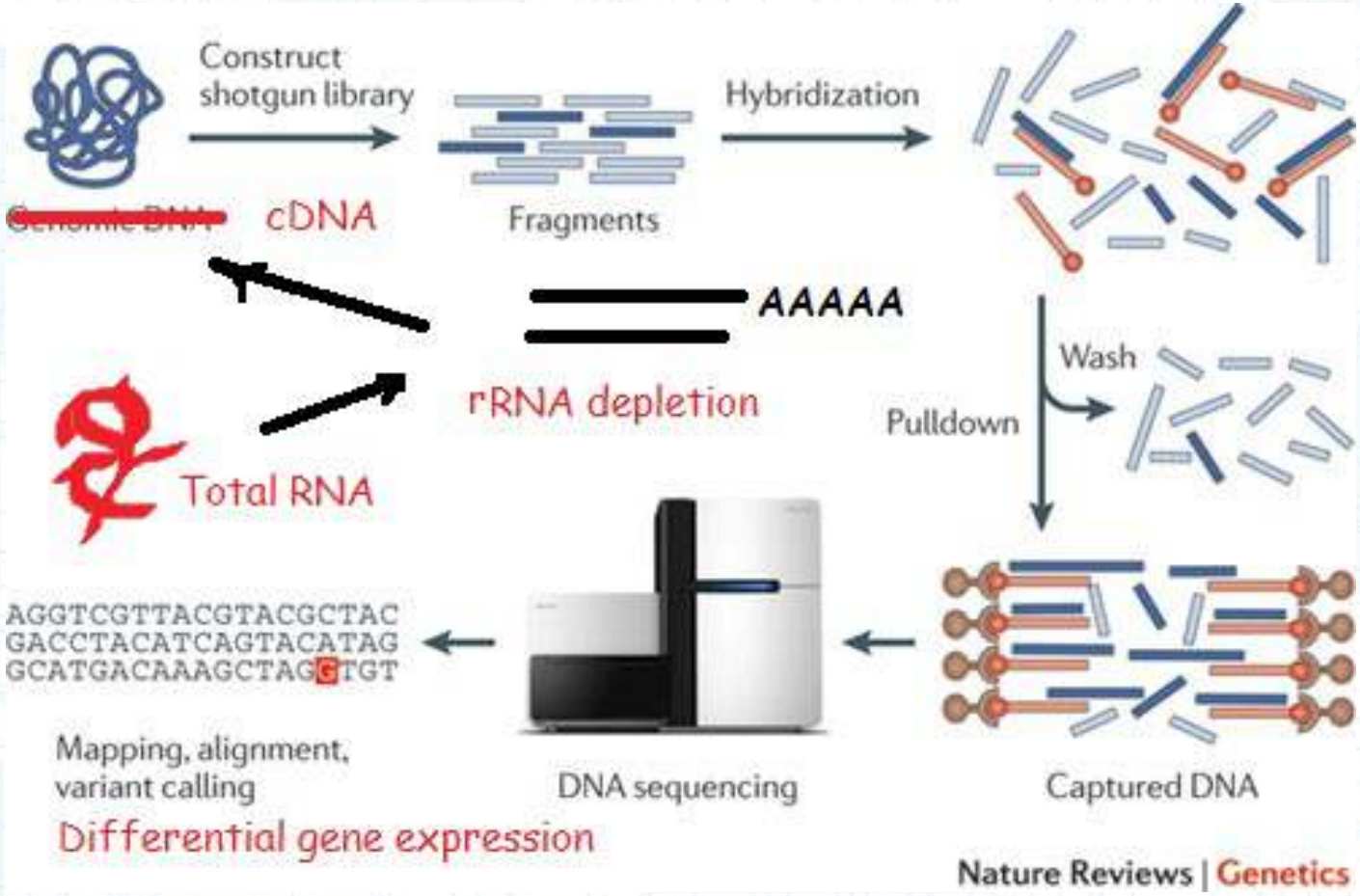
NGS: coverage variation

- Statistical (Poisson distribution)
- Polymorphisms and structural variation (seg dups, karyotype abnormalities)
- Reference issues (e.g. centromere, telomere)
- PCR biases (extremes of GC content usually under-represented)





Whole transcriptome sequencing



RNA-seq is awesome

- More sensitive than microarrays, almost as specific as qPCR
- Fast speed
- High-throughput
- Can also learn about small RNAs, expression outside of annotated exons, alternative splicing, etc

RNA-seq is not so **awesome**

- All problems specific to NGS
  - High price
  - High error rate
  - Computational requirements
- RNA-seq specific problems:
  - New analysis methods required
  - Protocol biases

RNA-seq specific biases

- PCR artifacts: results in identical reads
- 5' and 3' biases: uneven coverage in fragment
- Random hexamer priming is not random
- Strand-specificity
  - Allows infer the strand of the transcript, but has problems in construction

Rna-Seq specific biases

How to solve:

- Perform quality controls. Tools:

  FastQC, Qualimap, RNASeq QC
- Use better algorithms
- Use replicates and statistics

Experiment types

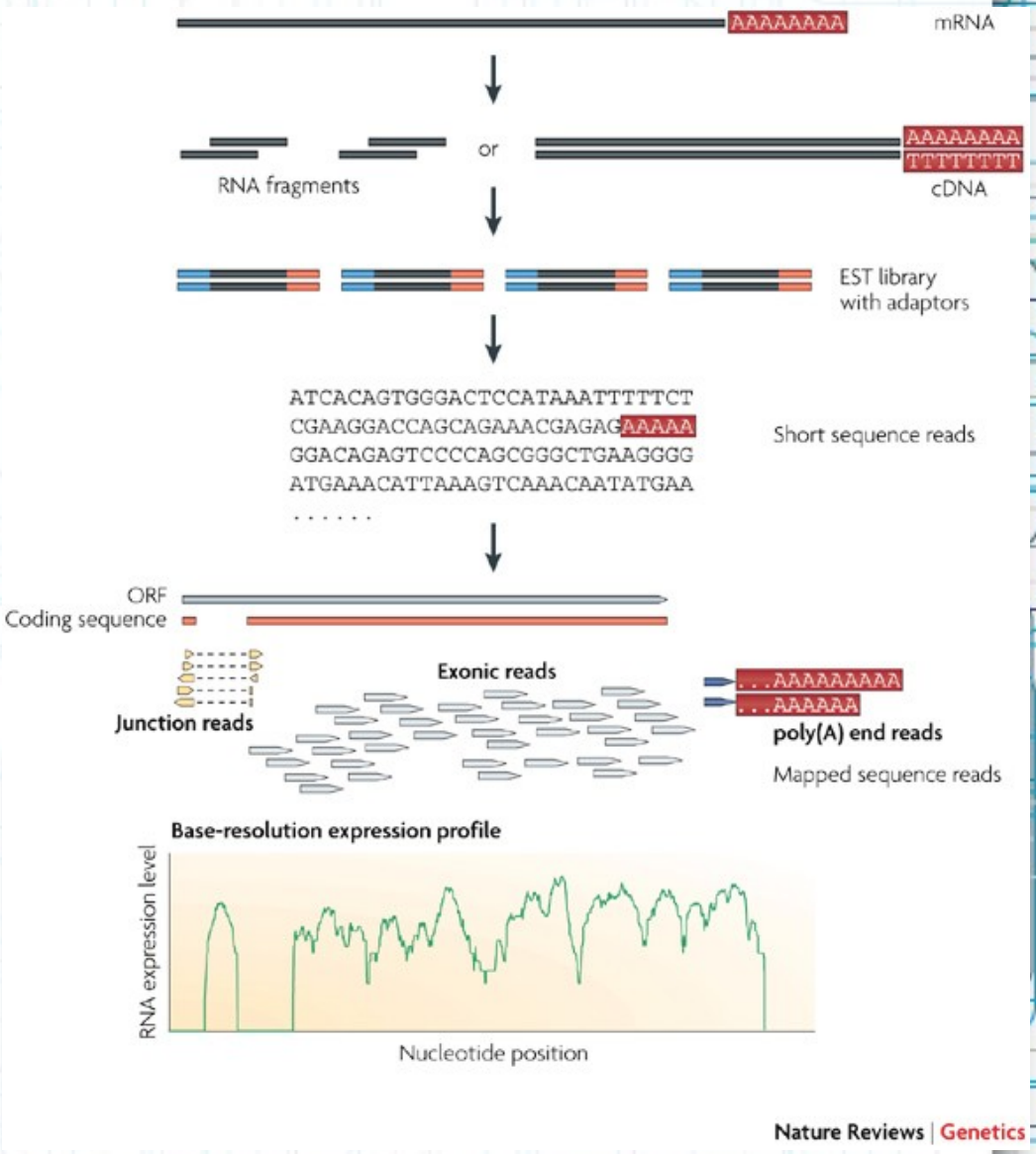
- Evaluation of a tissue's transcriptome
  - What is the composition of the transcriptome?
- **Differential gene expression (DE)**
- Novel genes discovery
- Alternative splicing
- Small RNA studies
- Other studies

Gene expression studies

- Without novel transcripts
  - Quick identification and analysis of differentially expressed genes
  - Required annotated reference genome, such as human and mouse
- With discovery of novel transcripts
  - Not limited by previous knowledge
  - Extends current knowledge banks
  - More complicated analysis

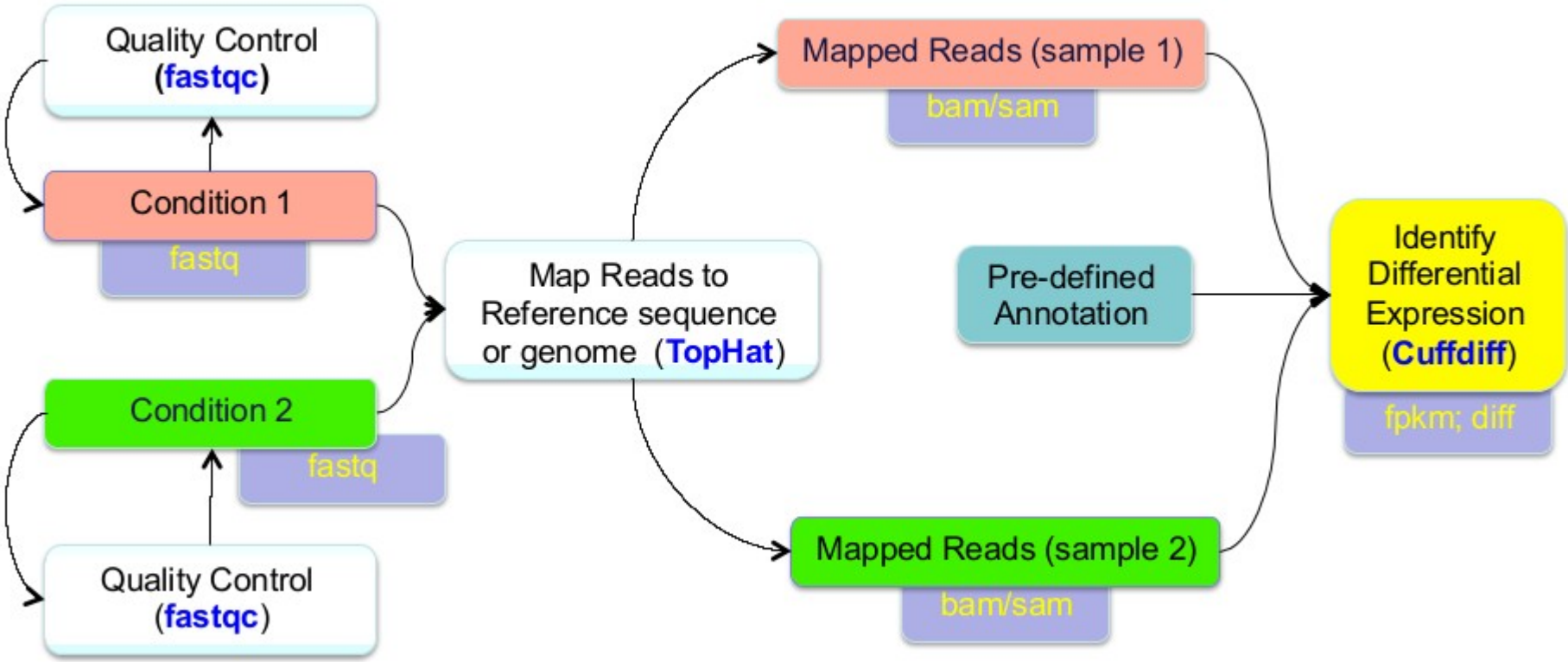
Differential gene expression: general steps

- Quality control
- Reads alignment
- Quantification
- Normalization
- Comparison
- Biological inference



Differential gene expression: pipeline example

- 2 or more conditions are compared

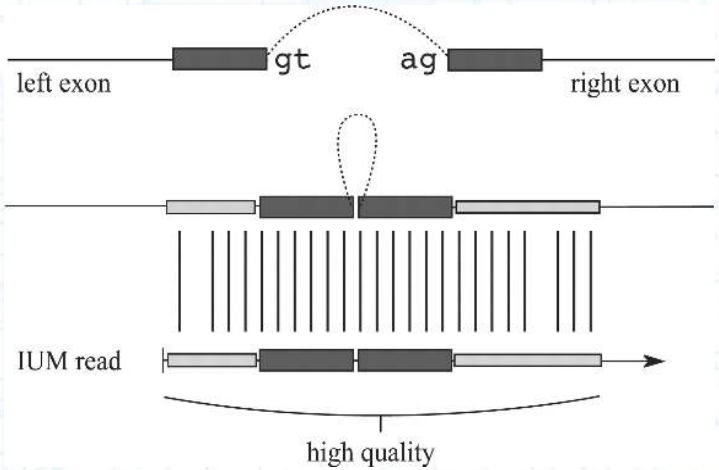


Spliced alignment

- Problem: the genes contain introns. How to infer reads covering the junction?
- Naive solution: align to transcriptome
  - Can use any existing aligner for short reads: **bwa, bowtie,** etc.
  - Problem: we are loosing novel transcripts and other information
- Better solution: spliced alignment

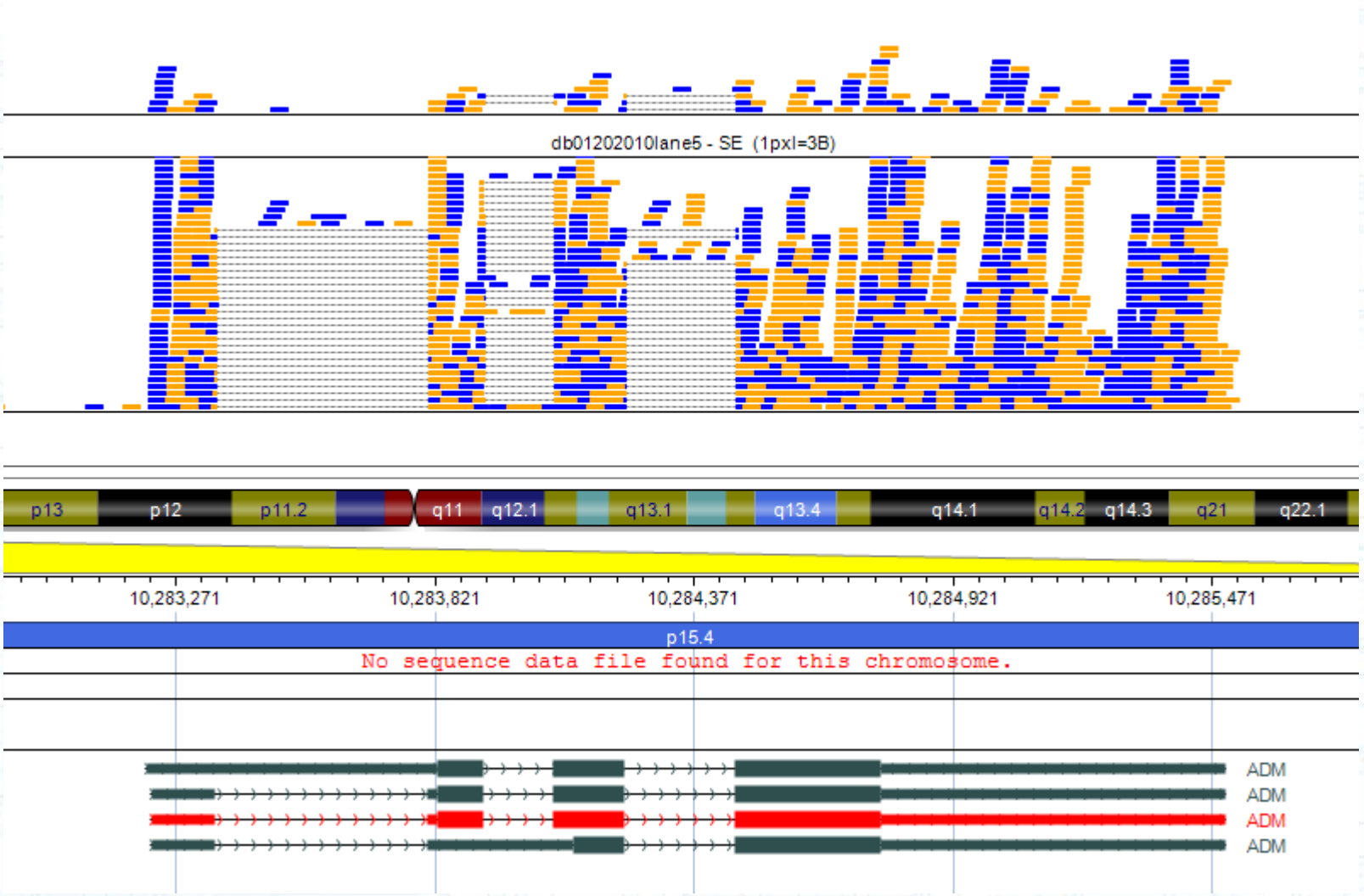
Spliced alignment

- Idea: split read into parts and align them independently



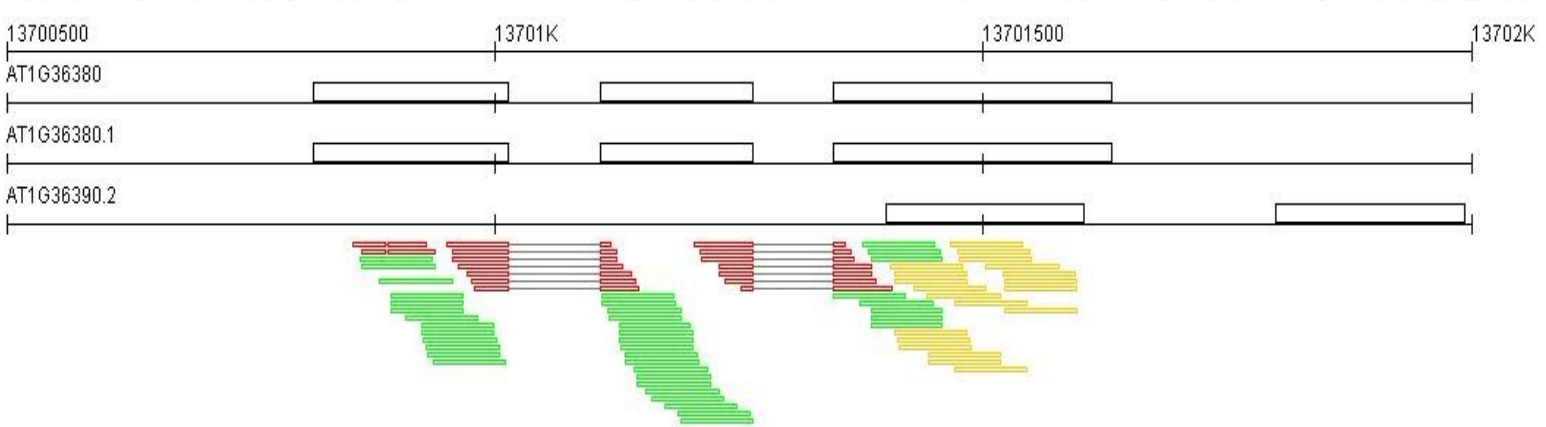
- Some tools: **tophat, splicemap, mapsplice, rum, star...**
- Problems: psuedogenes, repetitive regions

Spliced alignment



Quantification

- Quantification: counting how many reads are mapped to genes



  - Multimapped reads?
  - Reads that map to introns or outside exon boundaries?
  - Gene or transcript level? (will return later)
  - What about overlapping genes?

Normalization

- We compare 2 or more RNA-seq samples. Coverage is not the same for each sample.
- Problems: Need to scale RNA counts per gene to total sample coverage. Longer genes have more reads, gives better chance to detect DE
- Simple solution – divide counts by gene length

  RPKM (Reads Per KB per Million)

Normalization

- Similar metrics: **FPKM** (fragments per kilobase per million of reads)

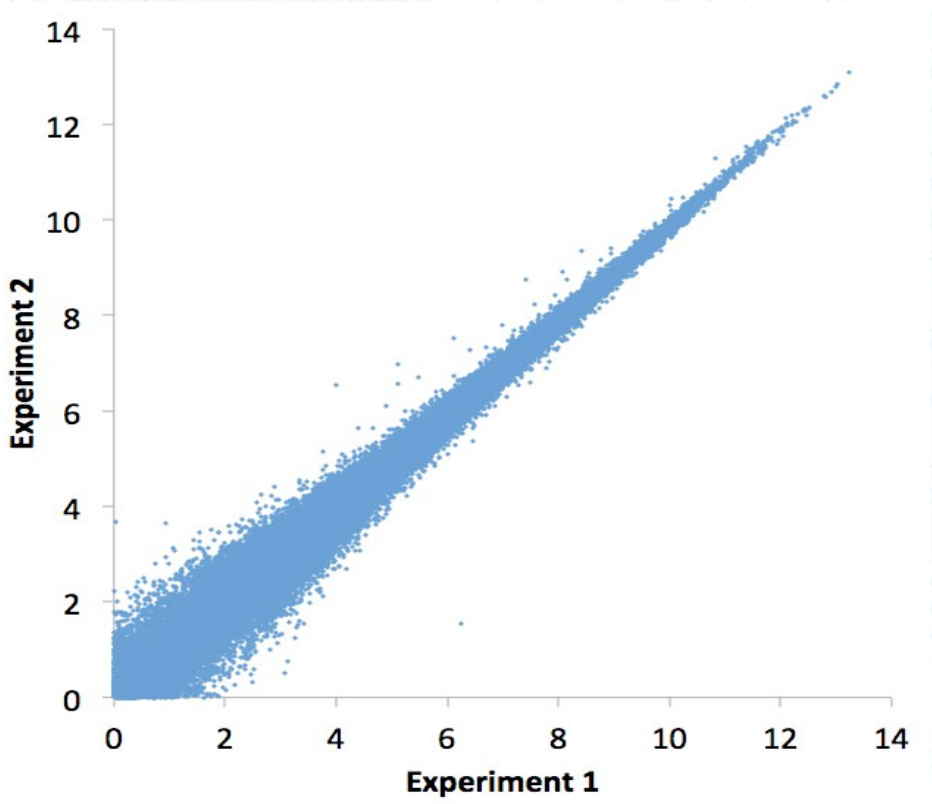
transcript	Sample A	Sample V	Sample O	Sample E	Sample I	Sample U
gene1	6.18	6.64	6.46	6.30	6.58	6.54
gene2	5.48	0.11	1.00	0.24	0.02	0.68
gene3	20.53	18.93	18.79	18.51	18.00	18.26
gene4	55.47	52.71	50.39	54.66	49.15	44.68
gene5	7.28	8.09	8.57	7.82	8.29	9.38
gene6	14.65	13.88	13.48	13.98	14.72	12.47
gene7	16.41	13.80	14.99	17.20	14.39	13.50
gene8	6.17	6.79	7.20	6.70	8.42	7.26
gene9	25.83	24.24	25.63	27.09	22.18	23.09
gene10	38.04	30.39	35.53	37.42	28.72	27.28
gene11	195.06	179.88	178.18	208.25	179.01	155.15
gene12	32.82	32.04	31.84	33.62	31.06	29.46
gene13	18.41	16.75	16.72	17.33	16.32	16.87
gene14	24.00	21.05	22.68	22.72	22.08	22.45
			.....			

Group 1 (A,V,O)

Group 2 (E,I,U)

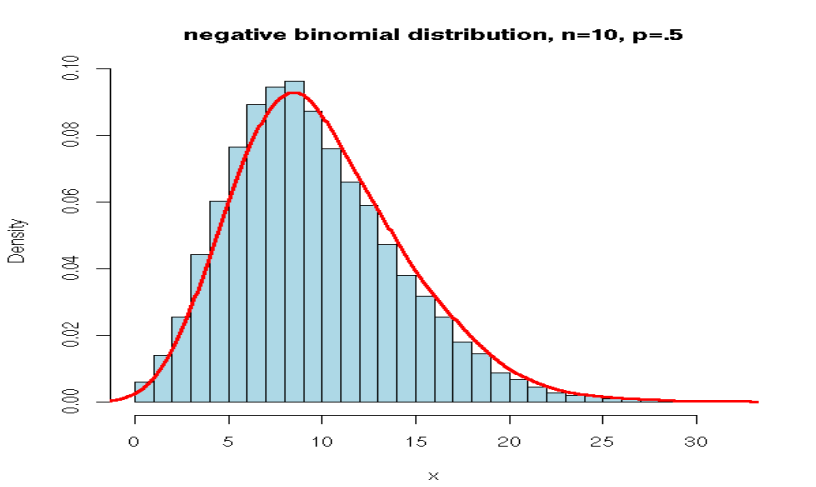
DE Statistics

- Problem: random technical noise vs biological variation



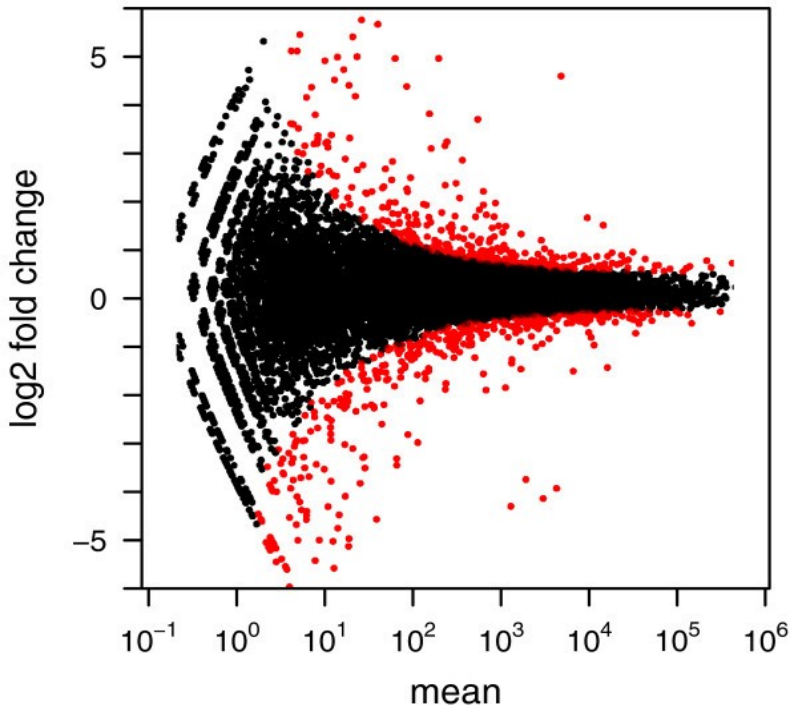
DE Statistics

- Statistical testing: for a given gene, an observed difference in read counts is significant (is it greater than what would be expected just due to natural random variation)
- Use probabilistic distribution to model number of reads to assigned gene: **negative binomial**



DE Statistics

- Additional model changes are used to improve the mean~variance relationship
- Popular algorithms: **DESeq, edgeR, cuffmerge-cuffcompare**

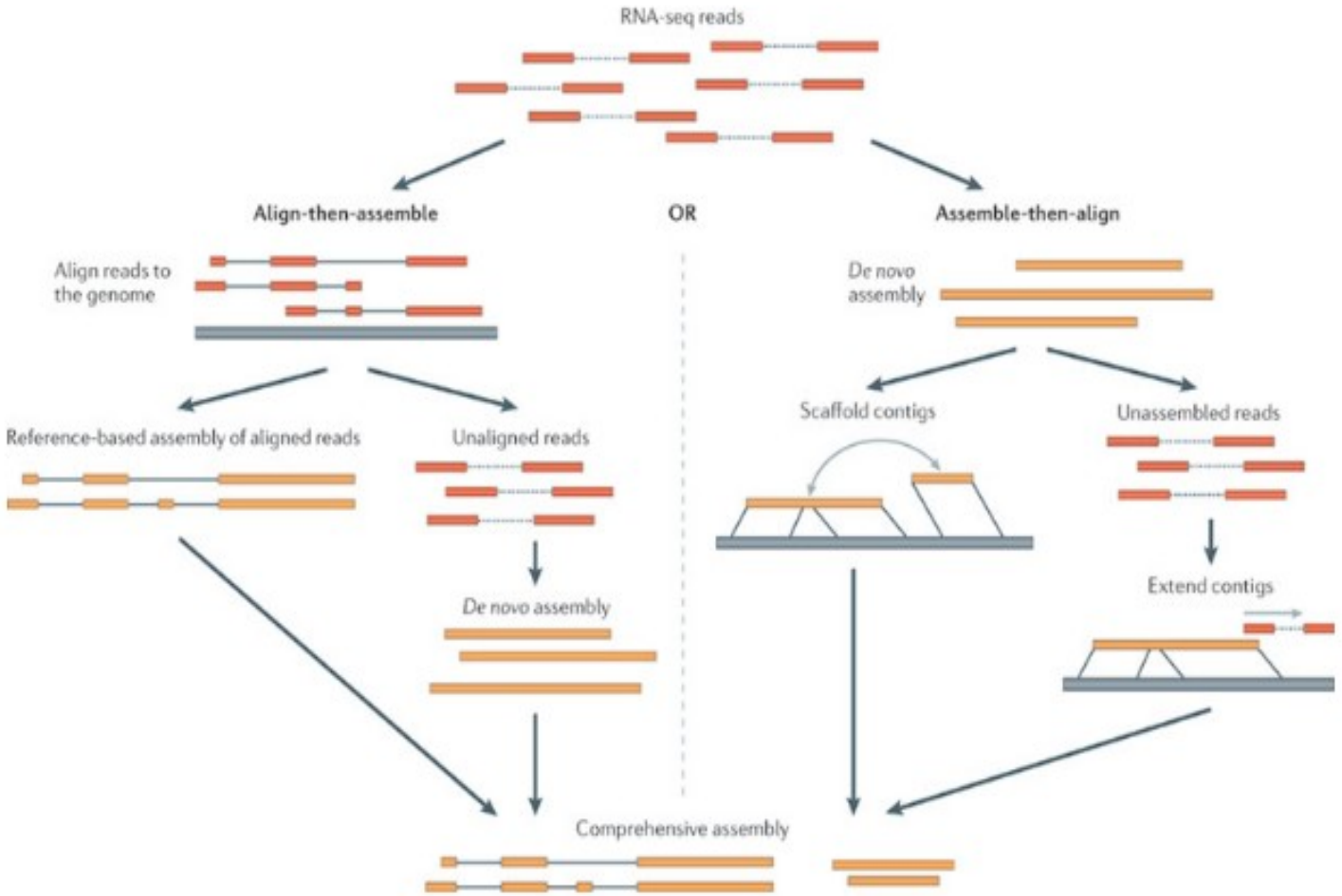


Biological inference

- Use data analysis to find your genes (clustering, principal component analysis, etc.)
  - Example tool: **R, Matlab**
- Detect pathways
  - Example tool: **IPA**
- Analyze Gene Ontologies
  - Example: **DAVID, Blast2GO**

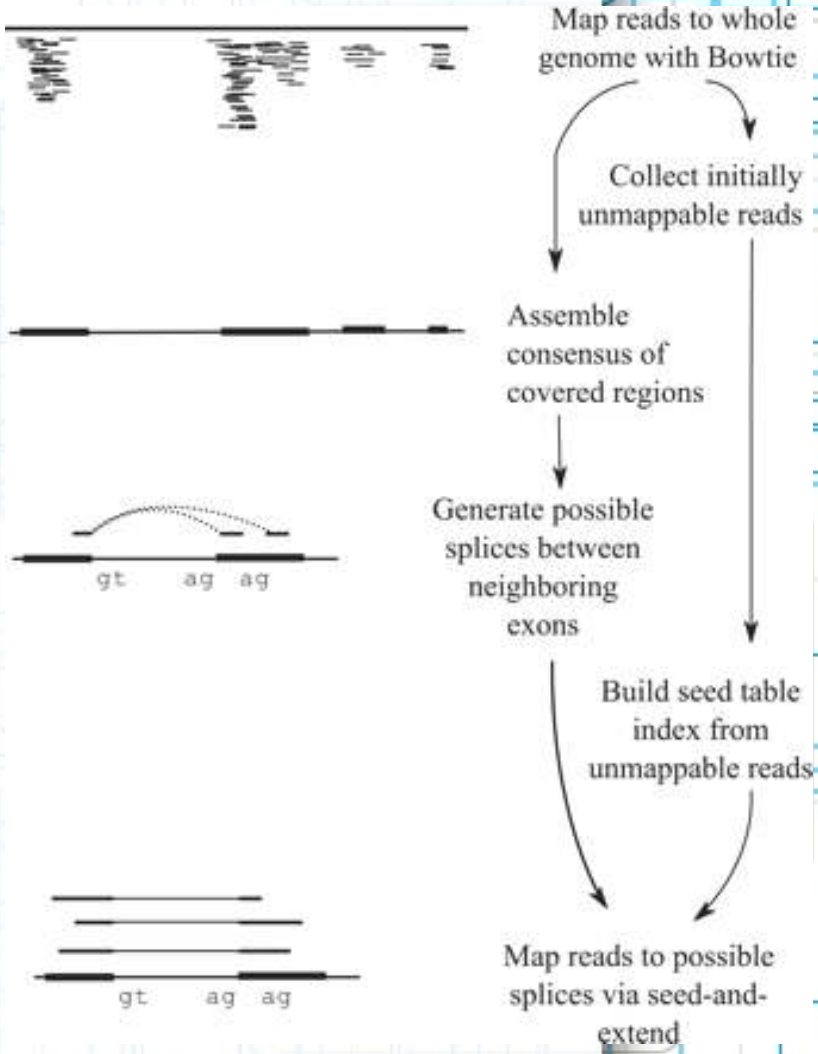
Novel transcript detection

- Discovery mode: on



Alternative splicing and reference based assembly

- Use exon junctions
- Create splicing graph
- Assign multimapped reads
- Infer transcripts using a probailistic model
- Popular tools:
  - Cufflinks
  - Splicing Compass

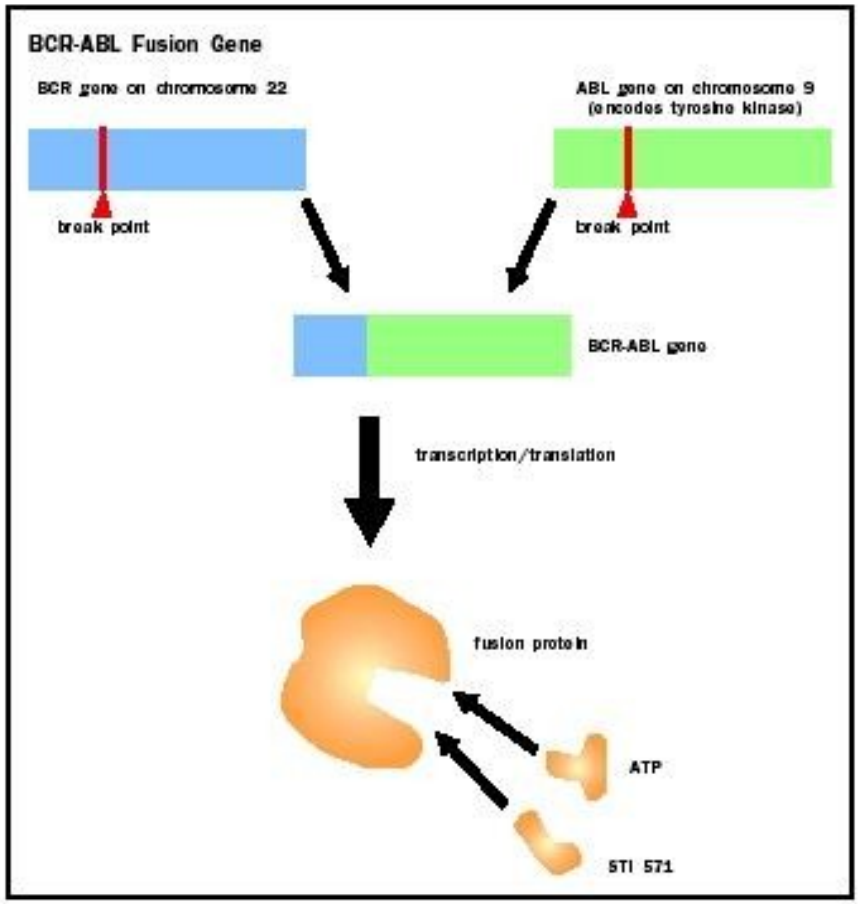


Transcriptome assembly

- Popular tools: **Trinity, TransAbyss** (based on de Brujn graphs)
  - Find all non-overlapping k-mers and build graphs
  - Extend using paired information
  - Create transcripts and align to genome
- Computational problem: all-against-all similarity searching and multiple overlapping transcripts

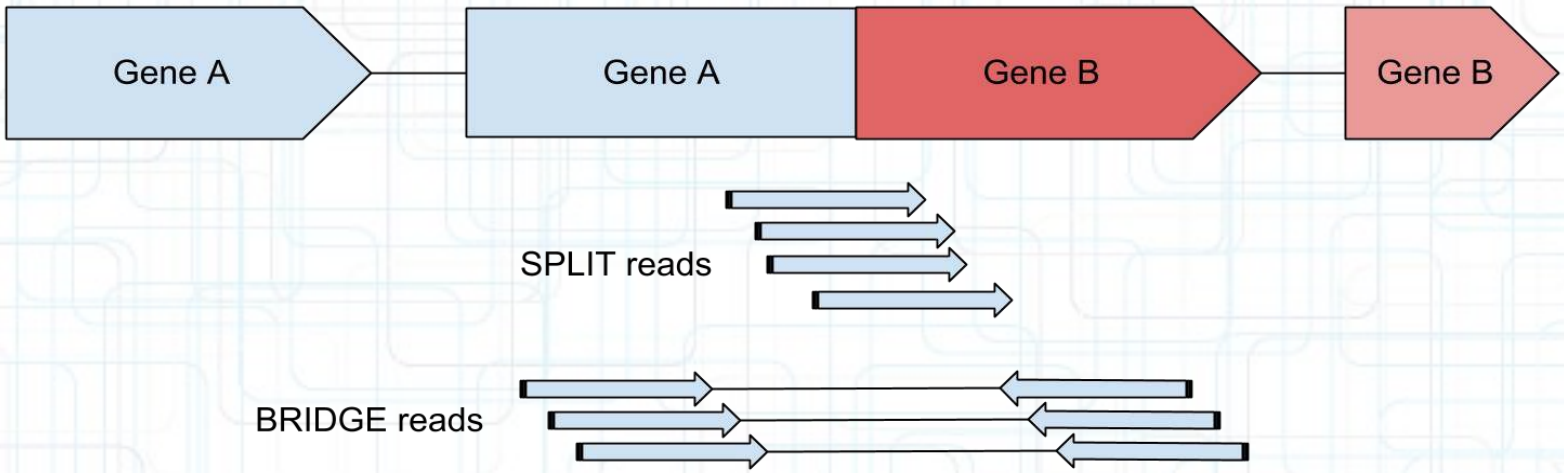
Advanced analysis: fusion genes

- Relevant for several types of cancers, but can be found in normal tissues too
- Can occur to genome breaks (fusions) or transcription process errors (chimeric transcripts)

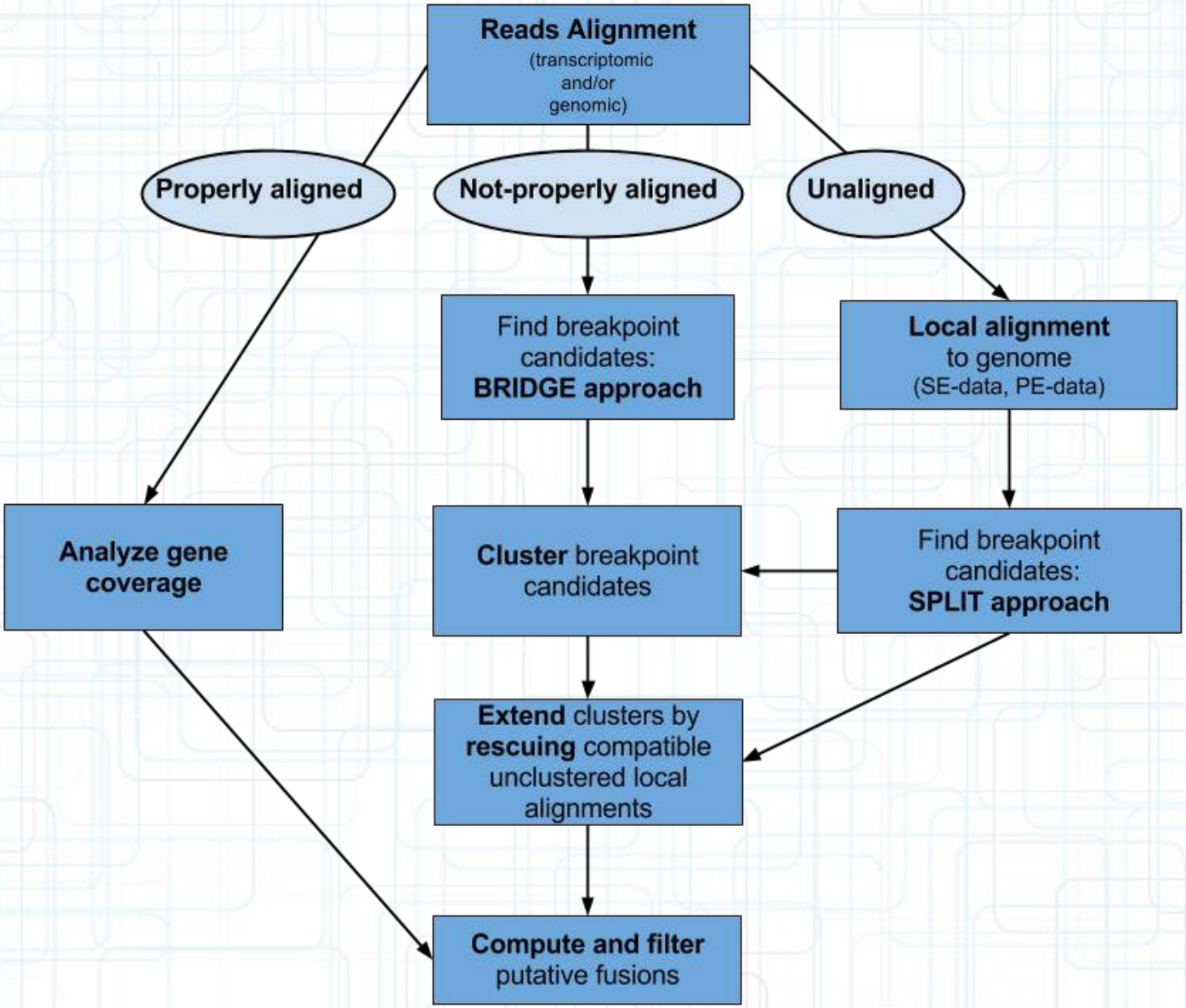


Fusion Gene Discovery

- Basic idea: find evidence from short reads



InFusion pipeline



Fusion genes filtering

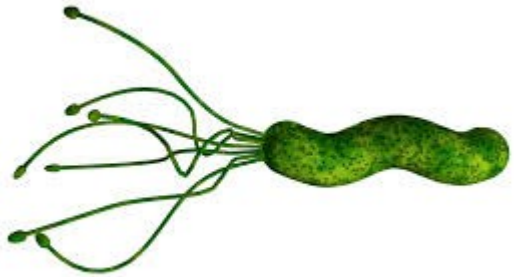
- 1000 of candidates. How to filter false positives?
  - Supporting reads
  - Homology of genes
  - Insert size distribution
  - Coverage pattern
  - Prediction via machine learning
- Fusion properties: type, ORF, isoforms
- Expression of fusion genes

References

- **Ying Zhang, John Garbe** RNA-seq tutorial
- **Stuart M. Brown, Zuojian** Tang Introduction to RNA-seq

For more references google example tools.

Final remarks



Спасибо за внимание!