High throughput sequencing methods in genomics. Focus on transcriptome analysis

D. Puthier

2016

Genomics

- Genomics is the discipline which aims at studying genome (structure, function of DNA elements, variation, evolution) and genes (their functions, expression...).
- Genomics is mostly based on large-scale analysis
  - Microarrays
  - Sequencing
  - Yeast-two-hybrids,...

Genomics

“The science for the 21st century”

Ewan Birney(EMBL-EBI)

at GoogleTech talk

Genomics an interdisciplinary science

Analysing genomes requires teams/individuals with various skills

- Biology
- Informatics
- Bioinformatics
- Statistics
- Mathematics, Physics
- ...

Introduction to transcriptome analysis using high- throughput sequencing technologies

Transcriptome analysis

- Tentative definition
  - Transcriptome: the set of all RNA produced by a cell or population of cells at a given moment

Main objectives of transcriptome analysis

- Understand the molecular mechanisms underlying gene expression
  - Interplay between regulatory elements and expression
    - Create regulatory model
      - E.g; to assess the impact of altered variant or epigenetic landscape on gene expression
- Classification of samples (e.g tumors)
  - Class discovery
  - Class prediction

Relies on a holistic view of the system

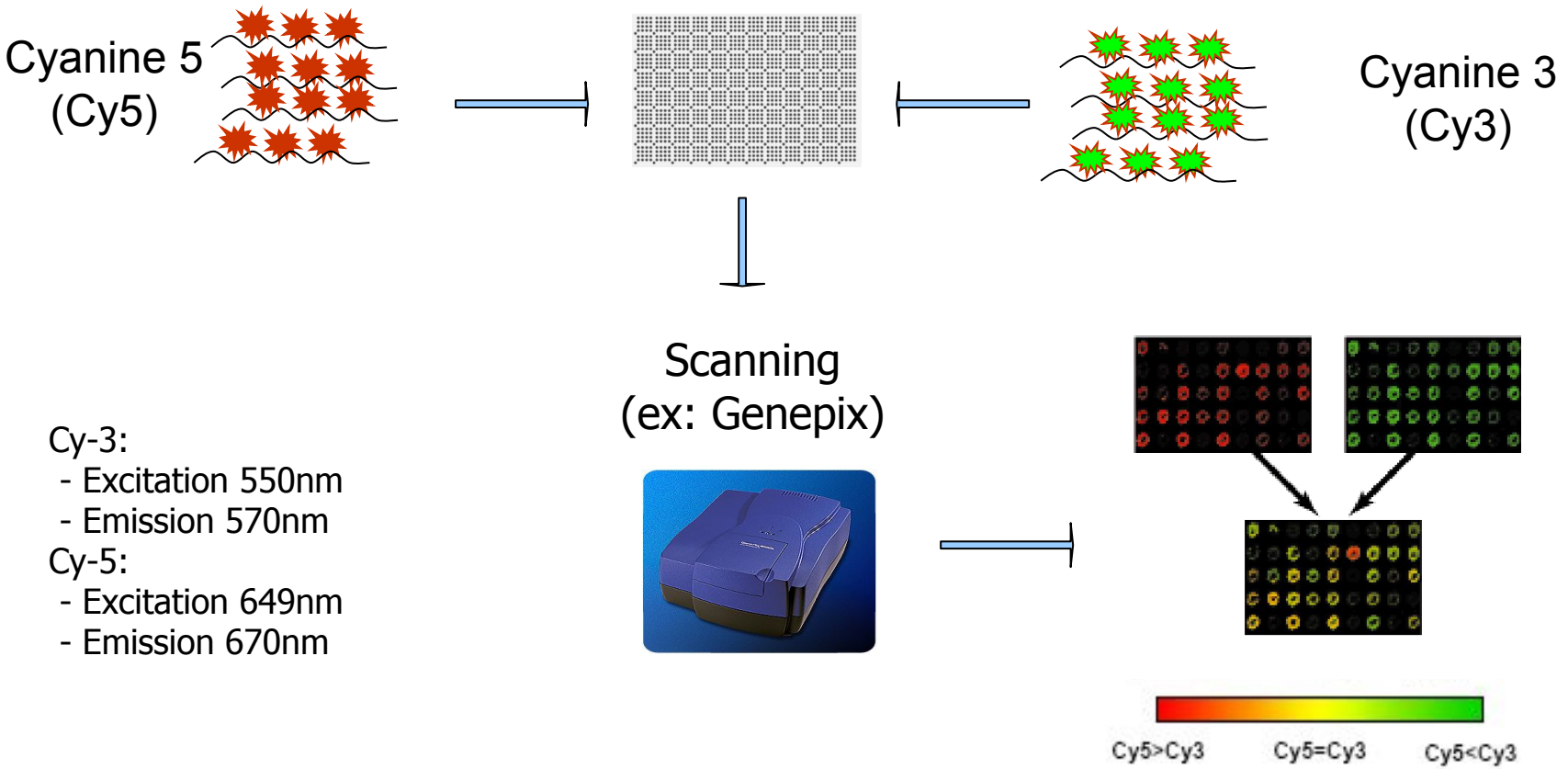
Some players of the RNA world

- Messenger RNA (mRNA)
  - Protein coding
  - Polyadenylated
  - 1-5% of total RNA
- Ribosomal RNA (rRNA)
  - 4 types in eukaryotes (18s, 28s, 5.8s, 5s)
  - 80-90% of total RNA
- Transfert RNA
  - 15% of total RNA

Some players of the RNA world

- miRNA
  - Regulatory RNA (mostly through binding of 3’UTR target genes)
- SnRNA
  - Uridine-rich
  - Several are related to splicing mechanism
  - Some are found in the nucleolus (snoRNA)
    - Related to rRNA biogenesis
- eRNA
  - Enhancer RNA
- And many others…(e.g LncRNA)

Transcriptome: the old school



Cy-3:
- Excitation 550nm
- Emission 570nm

Cy-5:
- Excitation 649nm
- Emission 670nm

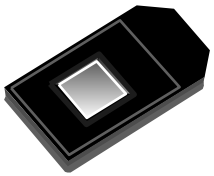
Science. 1995 Oct 20;270(5235):467-70.

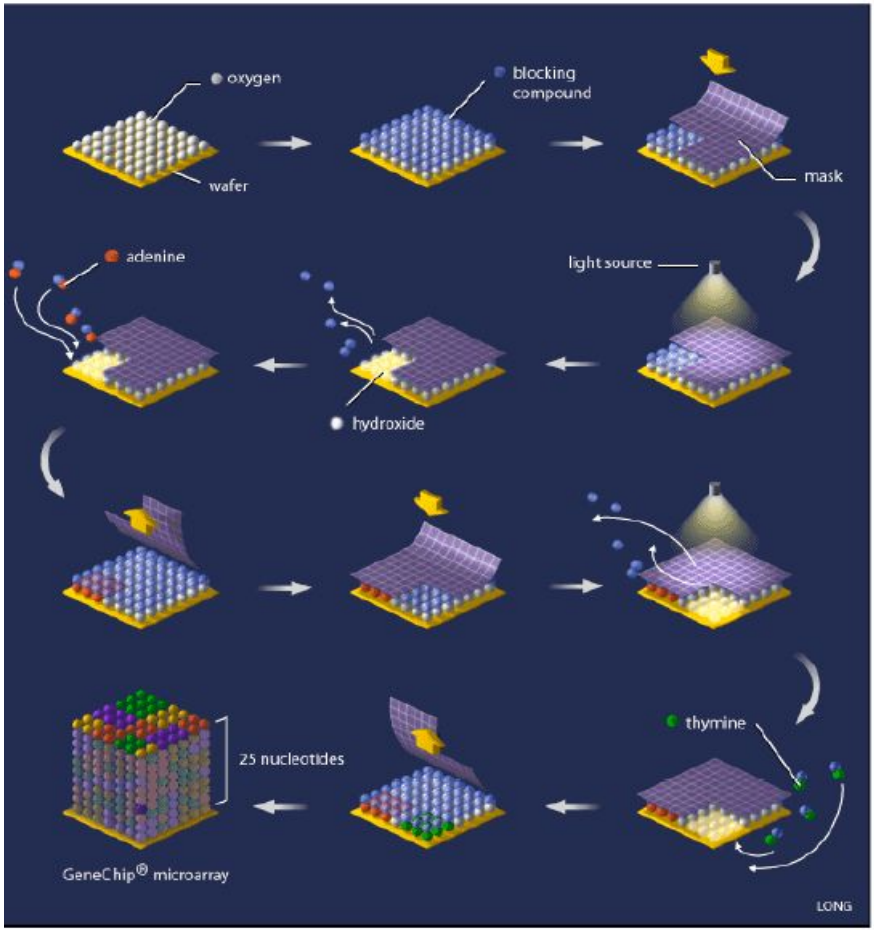
Quantitative monitoring of gene expression patterns with a complementary DNA microarray.

Schena M, Shalon D, Davis RW, Brown PO.

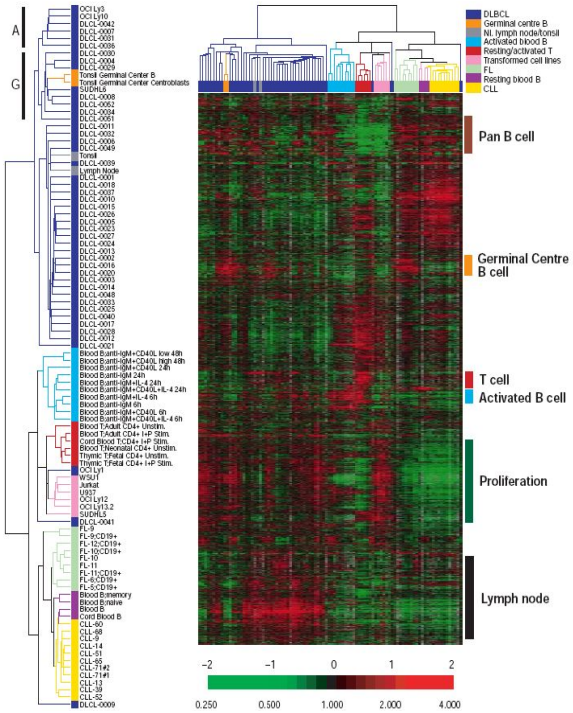
Transcriptome still the old school

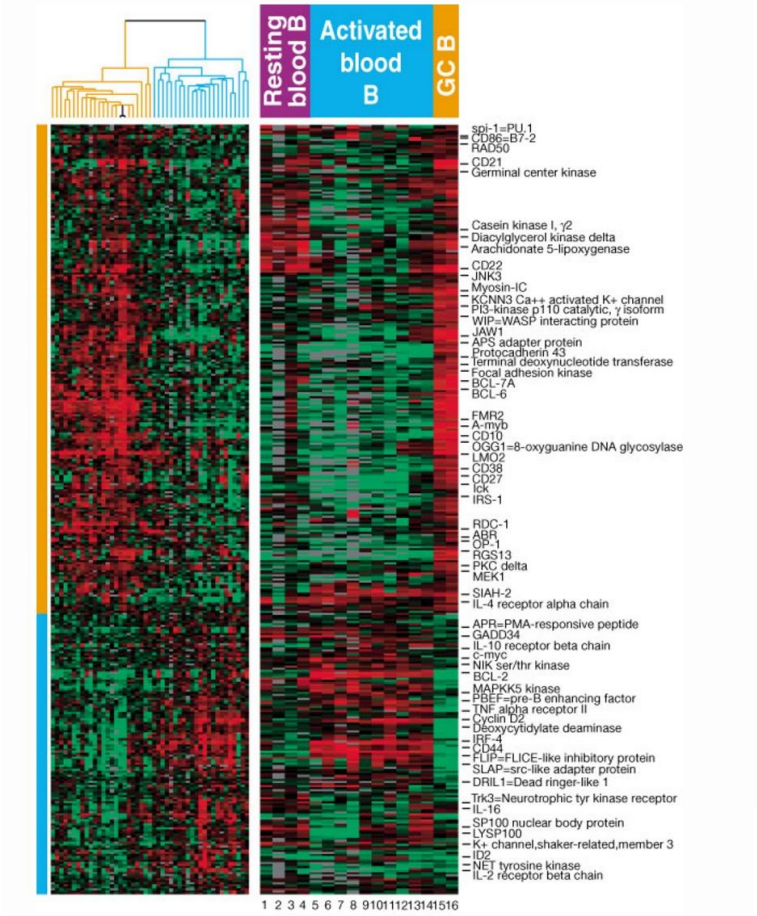
- Principle:
  - In situ synthesis of oligonucleotides
  - Features
    - Cells: 24µm x 24µm
    - ~10^7 oligos per cell
    - ~ 4.10^5-$1,5.10^6$ probes

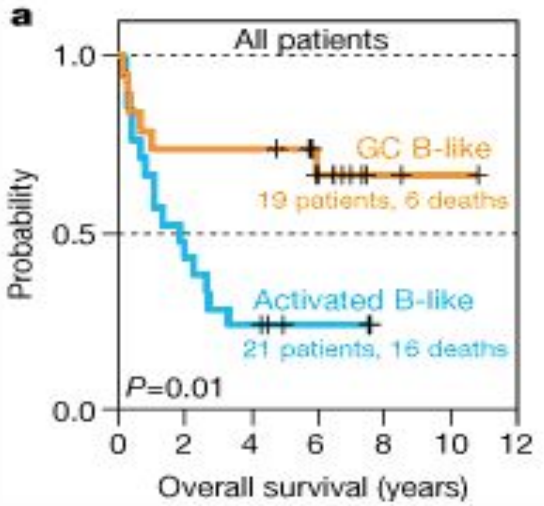




Some pioneering works







Nature. 2000 Feb 3;403(6769):503-11.

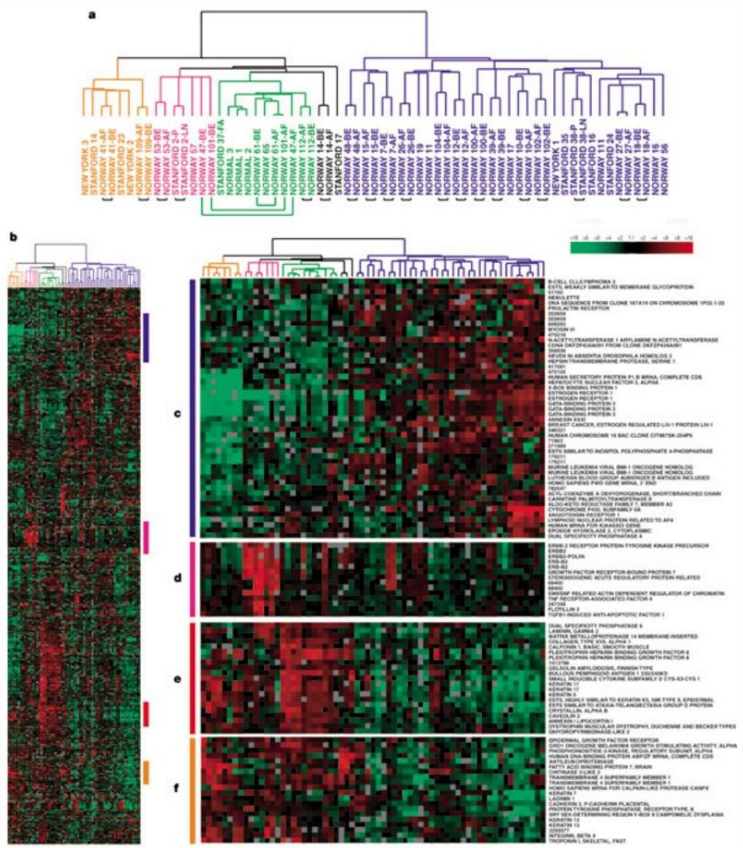
Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM.

Some pioneering works: “Molecular portrait of breast tumors”

Human breast tumours are diverse in their natural history and in their responsiveness to treatments[1]. Variation in transcriptional programs accounts for much of the biological diversity of human cells and tumours. In each cell, signal transduction and regulatory systems transduce information from the cell's identity to its environmental status, thereby controlling the level of expression of every gene in the genome. Here we have characterized variation in gene expression patterns in a set of 65 surgical specimens of human breast tumours from 42 different individuals, using complementary DNA microarrays representing 8,102 human genes. These patterns provided a distinctive molecular portrait of each tumour. Twenty of the tumours were sampled twice, before and after a 16-week course of doxorubicin chemotherapy, and two tumours were paired with a lymph node metastasis from the same patient. Gene expression patterns in two tumour samples from the same individual were almost always more similar to each other than either was to any other sample. Sets of co-expressed genes were identified for which variation in messenger RNA levels could be related to specific features of physiological variation. The tumours could be classified into subtypes distinguished by pervasive differences in their gene expression patterns.

- Top



Two large branches were apparent in the dendrogram, and within these large branches were smaller branches for which common biological themes could be inferred. Branches are coloured accordingly: basal-like, orange; *Erb-B2* +, pink; normal-breast-like, light green; and luminal epithelial/ER+, dark blue. **a**, Experimental sample associated cluster dendrogram. Small black bars beneath the dendrogram identify the 17 pairs that were matched by this hierarchical clustering; larger green bars identify the positions of the three pairs that were not matched by the clustering. **b**, Scaled-down representation of the intrinsic cluster diagram (see Supplementary Information Fig. 6). **c**, Luminal epithelial/ER gene cluster. **d**, *Erb-B2* overexpression cluster. **e**, Basal epithelial cell associated cluster containing keratins 5 and 17. **f**, A second basal epithelial-cell-enriched gene cluster.

Some pioneering works: Cluster analysis to infer gene function

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

Contributed by David Botstein, October 13, 1998

ABSTRACT **A system of cluster analysis for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes according to similarity in pattern of gene expression. The output is displayed graphically, conveying the clustering and the underlying expression data simultaneously in a form intuitive for biologists. We have found in the budding yeast *Saccharomyces cerevisiae* that clustering gene expression data groups together efficiently genes of known similar function, and we find a similar tendency in human data. Thus patterns seen in genome-wide expression experiments can be interpreted as indications of the status of cellular processes. Also, coexpression of genes of known function with poorly characterized or novel genes may provide a simple means of gaining leads to the functions of many genes for which information is not available currently.**

The rapid advance of genome-scale sequencing has driven the development of methods to exploit this information by characterizing biological processes in new ways. The knowledge of the coding sequences of virtually every gene in an organism, for instance, invites development of technology to study the expression of all of them at once, because the study of gene expression of genes one by one has already provided a wealth of biological insight. To this end, a variety of techniques has evolved to monitor, rapidly and efficiently, transcript abundance for all of an organism's genes (1–3). Within the mass of numbers produced by these techniques, which amount to hundreds of data points for thousands or tens of thousands of genes, is an immense amount of biological information. In this paper we address the problem of analyzing and presenting information on this genomic scale.

A natural first step in extracting this information is to examine the extremes, e.g., genes with significant differential expression in two individual samples or in a time series after a given treatment. This simple technique can be extremely efficient, for example, in screens for potential tumor markers or drug targets. However, such analyses do not address the full potential of genome-scale experiments to alter our understanding of cellular biology by providing, through an inclusive analysis of the entire repertoire of transcripts, a continuing comprehensive window into the state of a cell as it goes through a biological process. What is needed instead is a holistic approach to analysis of genomic data that focuses on illuminating order in the entire set of observations, allowing biologists to develop an integrated understanding of the process being studied.

A natural basis for organizing gene expression data is to group together genes with similar patterns of expression. The first step to this end is to adopt a mathematical description of similarity. For any series of measurements, a number of sensible measures of similarity in the behavior of two genes can be used, such as the Euclidean distance, angle, or dot products of the two n-dimensional vectors representing a series of n measurements. We have found that the standard correlation coefficient (i.e., the dot product of two normalized vectors) conforms well to the intuitive biological notion of what it means for two genes to be "coexpressed;" this may be because this statistic captures similarity in "shape" but places no emphasis on the magnitude of the two series of measurements.

It is not the purpose of this paper to survey the various methods available to cluster genes on the basis of their expression patterns, but rather to illustrate how such methods can be useful to biologists in the analysis of gene expression data. We aim to use these methods to organize, but not to alter, tables containing primary data; we have thus used methods that can be reduced, in the end, to a reordering of lists of genes. Clustering methods can be divided into two general classes, designated supervised and unsupervised clustering (4). In supervised clustering, vectors are classified with respect to known reference vectors. In unsupervised clustering, no predefined reference vectors are used. As we have little *a priori* knowledge of the complete repertoire of expected gene expression patterns for any condition, we have favored unsupervised methods or hybrid (unsupervised followed by supervised) approaches.

Although various clustering methods can usefully organize tables of gene expression measurements, the resulting ordered but still massive collection of numbers remains difficult to assimilate. Therefore, we always combine clustering methods with a graphical representation of the primary data by representing each data point with a color that quantitatively and qualitatively reflects the original experimental observations. The end product is a representation of complex gene expression data that, through statistical organization and graphical display, allows biologists to assimilate and explore the data in a natural intuitive manner.

To illustrate this approach, we have applied pairwise average-linkage cluster analysis (5) to gene expression data collected in our laboratories. This method is a form of hierarchical clustering, familiar to most biologists through its application in sequence and phylogenetic analysis. Relationships among objects (genes) are represented by a tree whose branch lengths reflect the degree of similarity between the objects, as assessed by a pairwise similarity function such as that described above. In sequence comparison, these methods are used to infer the evolutionary history of sequences being compared. Whereas no such underlying tree exists for expression patterns of genes, such methods are useful in their ability to represent varying degrees of similarity and more distant relationships among groups of closely related genes, as well as in requiring few assumptions about the nature of the data. The computed trees can be used to order genes in the original data table, so that genes or groups of genes with similar expression patterns are adjacent. The ordered table can then be displayed graphically, as above, with a representation of the tree to indicate the relationships among genes.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡To whom reprint requests should be addressed. e-mail: botstein@genome.stanford.edu.

Some pioneering works: tumor class prediction

Science. 1999 Oct 15;286(5439):531-7.

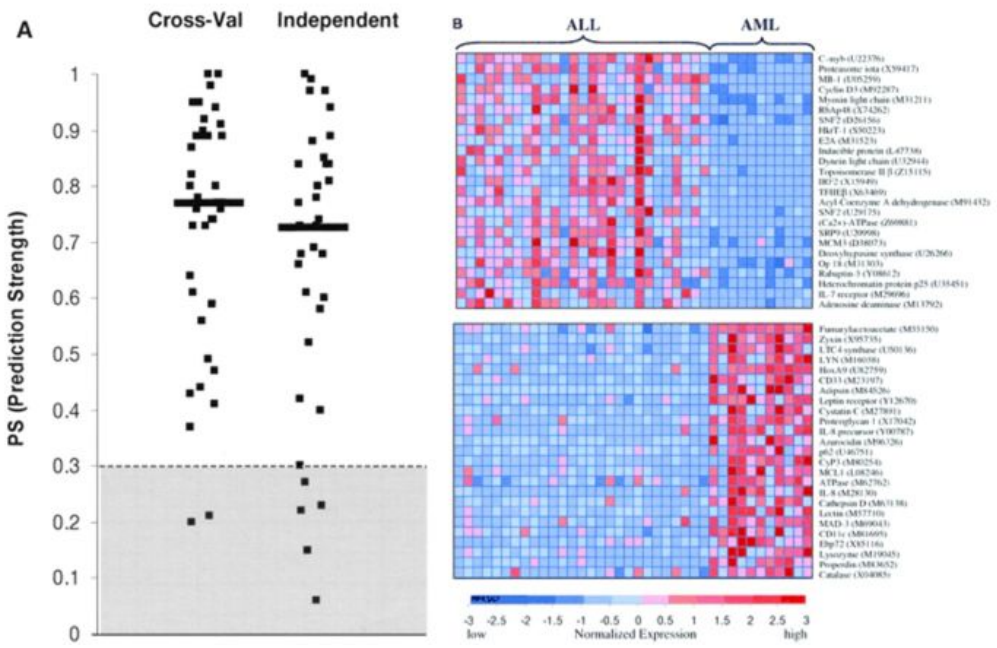
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.

Golub TR[1], Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES.

Author information

Abstract

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.



Microarrays drawbacks

- Cross-hybridization
  - Probe design issues
- Content limited
  - Can only show you what you're already looking for
- Indirect record of expression level
  - Complementary probes
  - Relative abundance

Even more powerful technology: RNA-Seq

Nature Methods - **5**, 585 - 587 (2008)
doi:10.1038/nmeth0708-585

The beginning of the end for microarrays?

Jay Shendure

Jay Shendure is in the Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. shendure@u.washington.edu

Two complementary appr
successfully tackled the s
once revealing unprecede

Published online 15 October 2008 | *Nature* **455**, 847 (2008) |
doi:10.1038/455847a

News

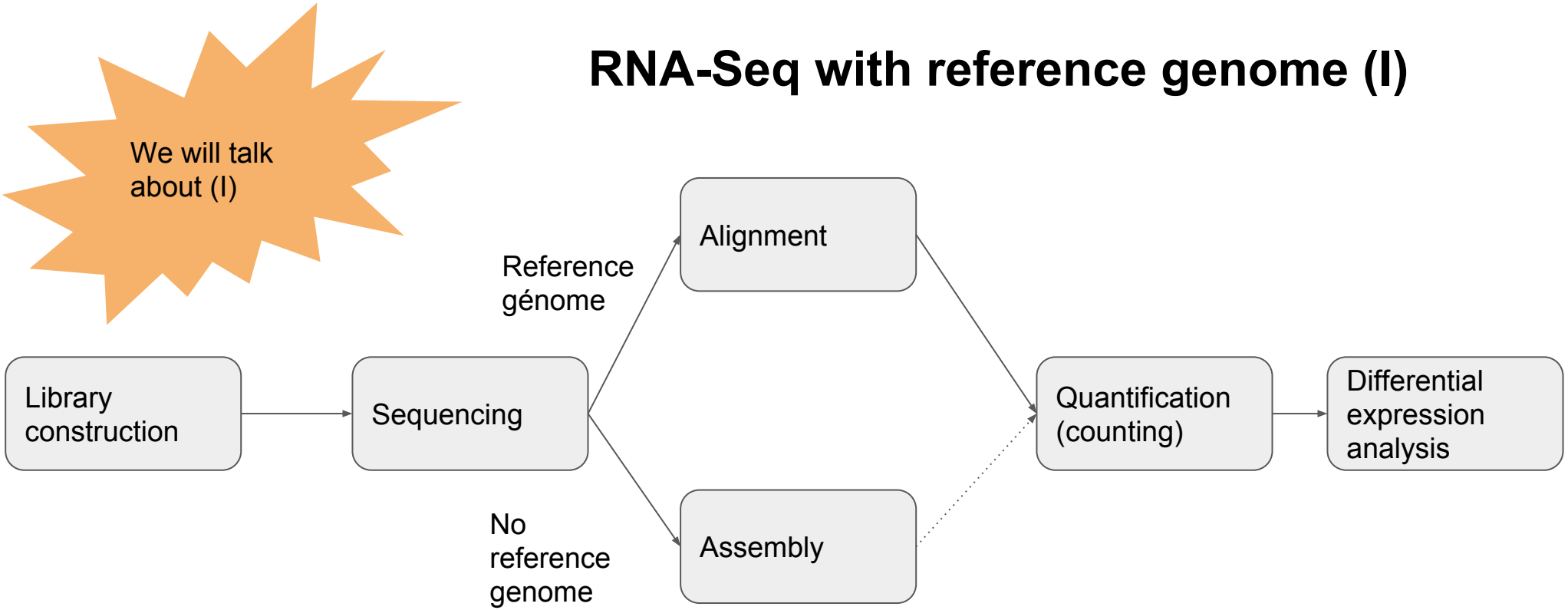
The death of microarrays?

High-throughput gene sequencing seems to be stealing a march on microarrays. Heidi Ledford looks at a genome technology facing intense competition.

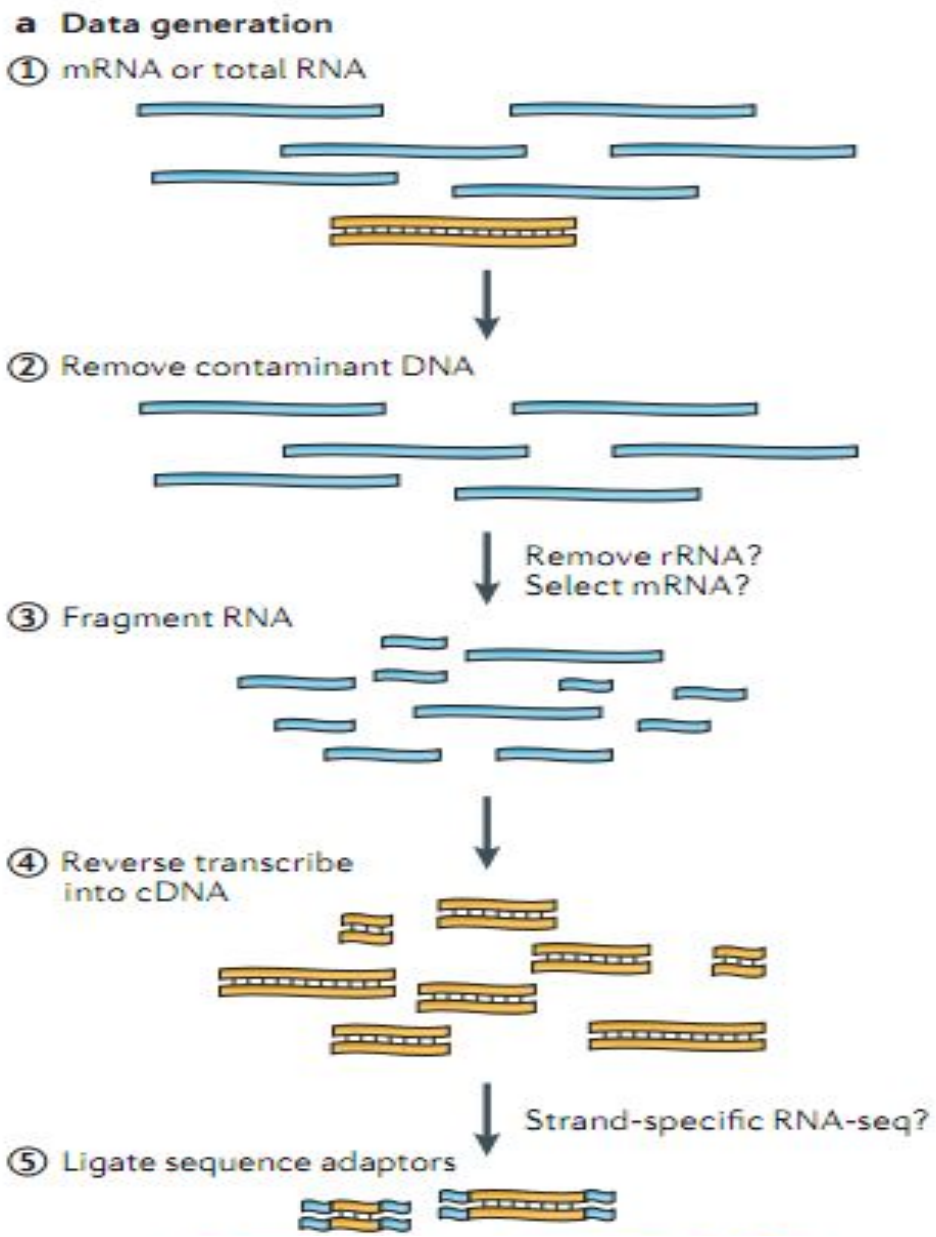
Heidi Ledford

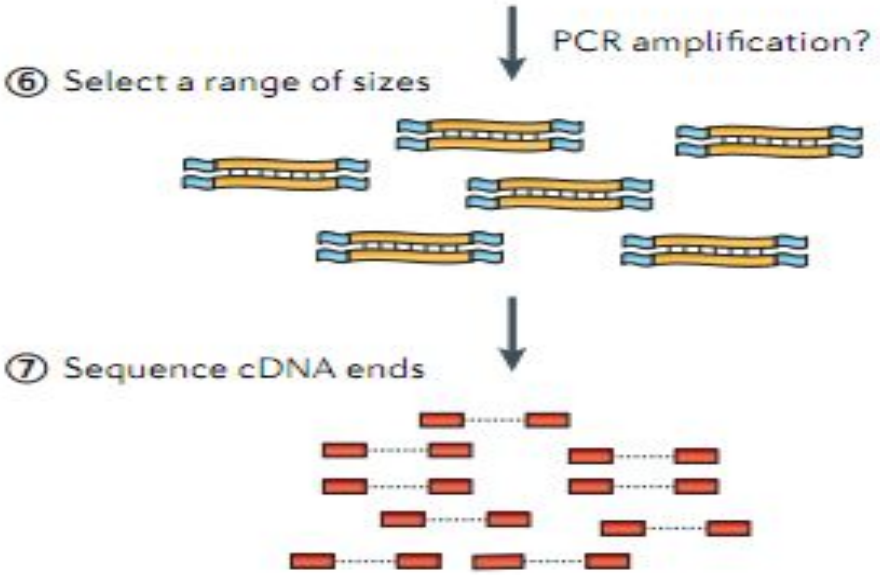
RNA-Seq global overview

- Objectives: sequencing of DNA fragments derived from transcripts



RNA-Seq: library construction simplified





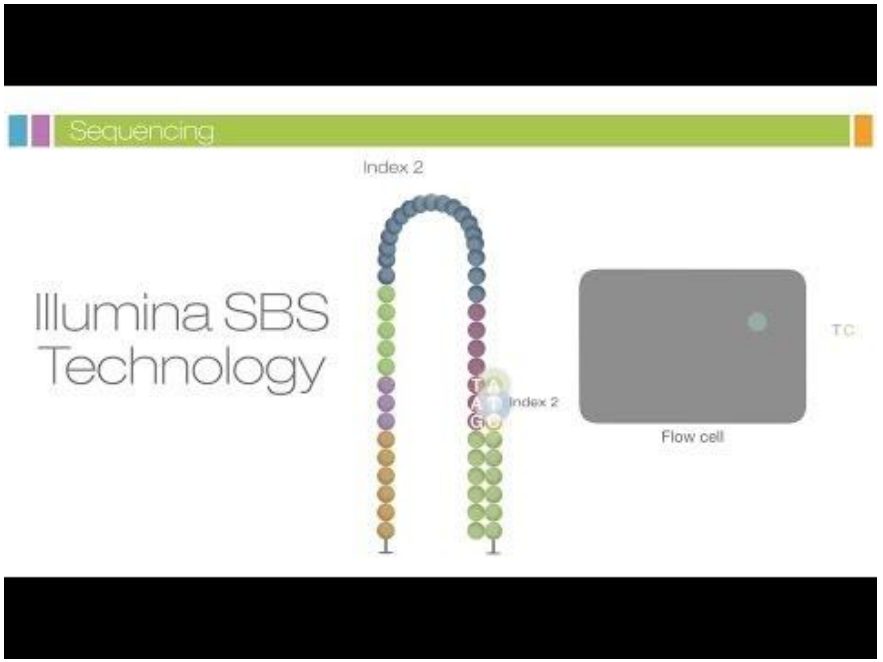
Nature Reviews Genetics **12**, 671-682 (October 2011) | doi:10.1038/nrg3068

ARTICLE SERIES: Study designs

Next-generation transcriptome assembly

Jeffrey A. Martin[1] & Zhong Wang[1] About the authors

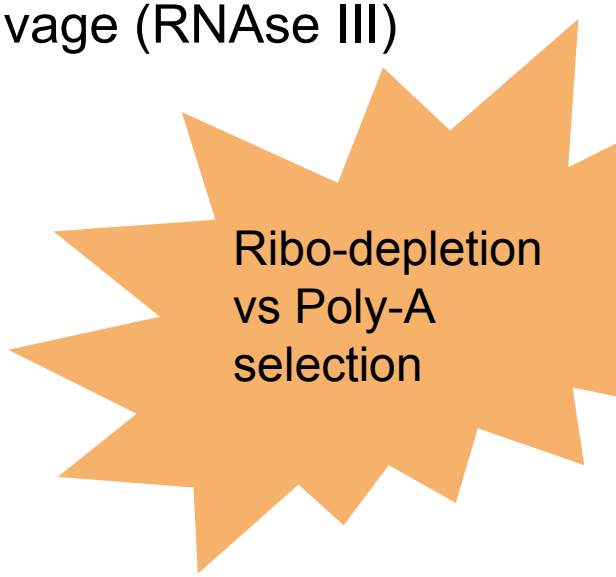
Illumina sequencing general principle



http://www.illumina.com/company/video-hub/HMyCqWhwB8E.html

RNA-Seq library construction: protocol variations

- **Fragmentation methods**
  - RNA: magnesium-catalyzed hydrolysis, enzymatic clivage (RNAse III)
  - cDNA: sonication, Dnase I treatment
- **Targeted RNA populations**
  - **Poly(A) RNA-Seq**:
    - **Positive selection** of mRNA . Poly(A) selection.
  - **Total RNA-Seq** :All transcript excluding ribosomal RNA (rRNA)
    - 'Ribo depletion'. **Negative selection**. (RiboMinus™)
    - Select also **pre-messenger**
  - **Small RNA-seq**
    - Size selection (e.g between 17nt and 35nt). E.g for miRNA profiling

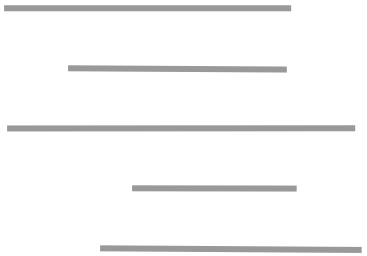


RNA-Seq library construction: protocol variations

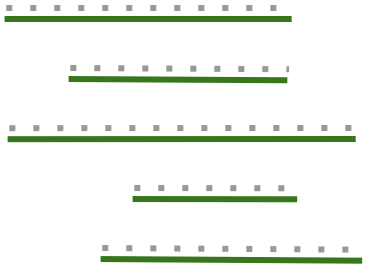
- **Stranded vs unstranded RNA-seq**
  - **Unstranded**
    - No information regarding the strand of the gene producing the fragment. Ambiguous reads should be discarded
  - **Stranded**
    - The strand of the gene producing the fragment can be inferred from alignment
    - No ambiguity. Better estimation of gene expression level.
    - Better reconstruction of transcript model.

Unstranded

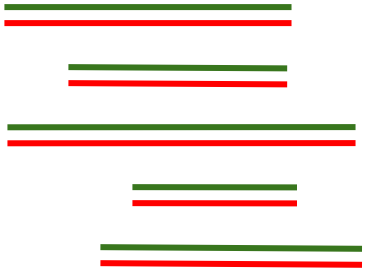
(1) - RNA fragments



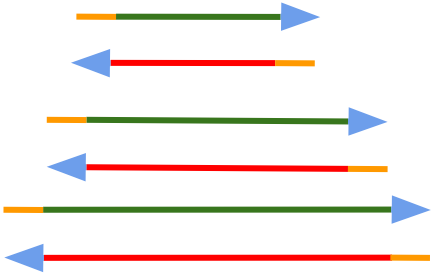
(2) - Reverse transcription and RNA degradation



(3) - dsDNA



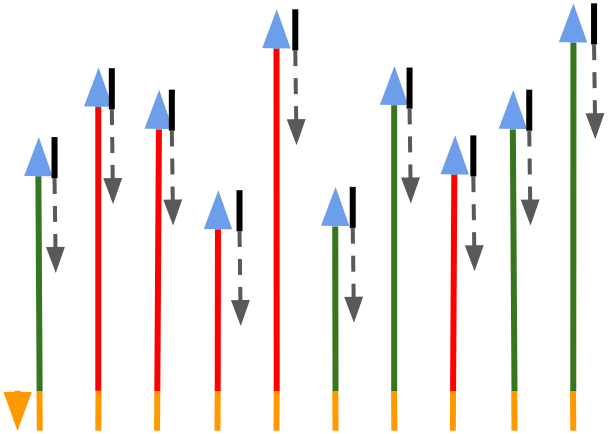
(4) - ligation of adapters



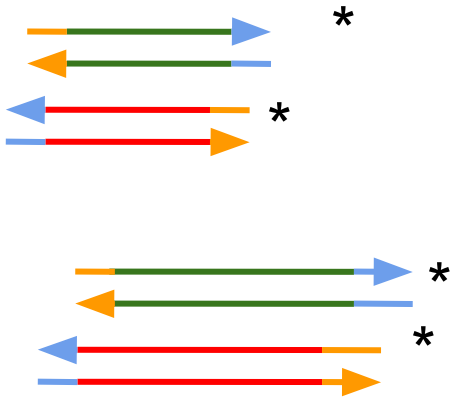
(6) - Results

Each colony may produce two types of sequences corresponding to both ends of the fragment.

(5) - Sequencing : bridge amplification (not shown) and sequencing of each fragment

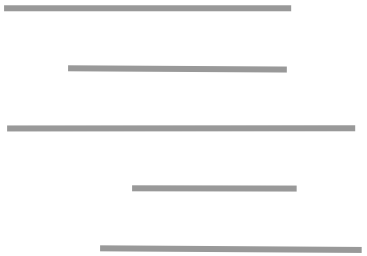


(4) - amplification

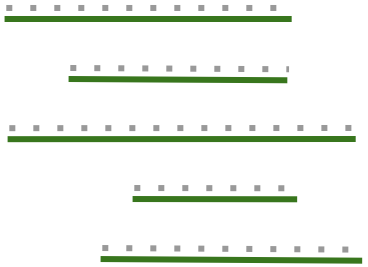


Stranded

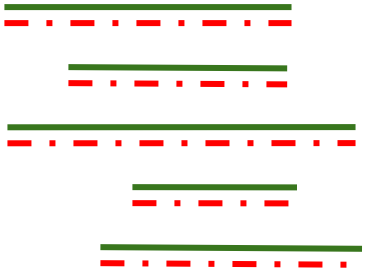
(1) - RNA fragments



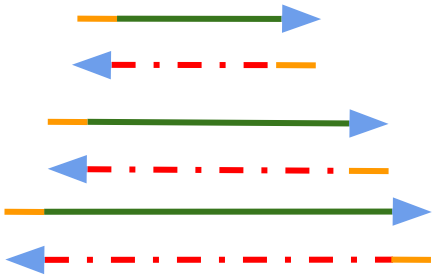
(2) - Reverse transcription and RNA degradation



(3) - second strand synthesis with dUTP



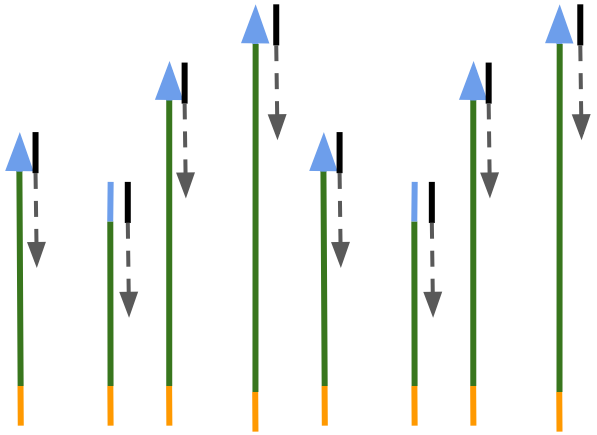
(4) - Ligation of adapters



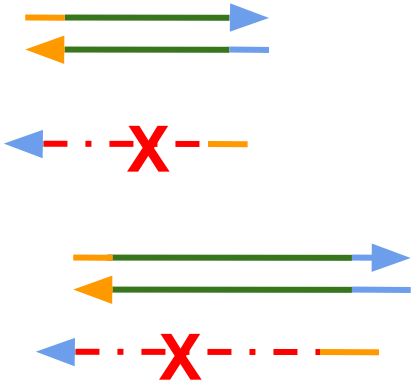
(6) - Results

Each colony may produce only one type of sequences corresponding to the 5’ or 3’ end depending on the kit.

(5) - Sequencing : bridge amplification (not shown) and sequencing of each fragment



(4) - amplification



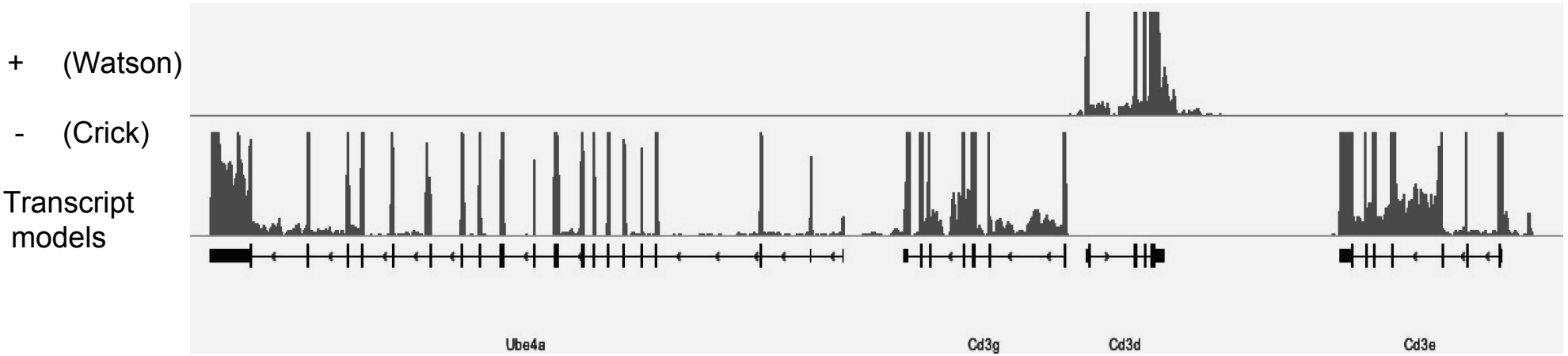
Example of stranded single-end RNA-Seq alignment

Forward (Red)
Reverse (Blue)

Cd3g (brin -)

Cd3d (brin +)

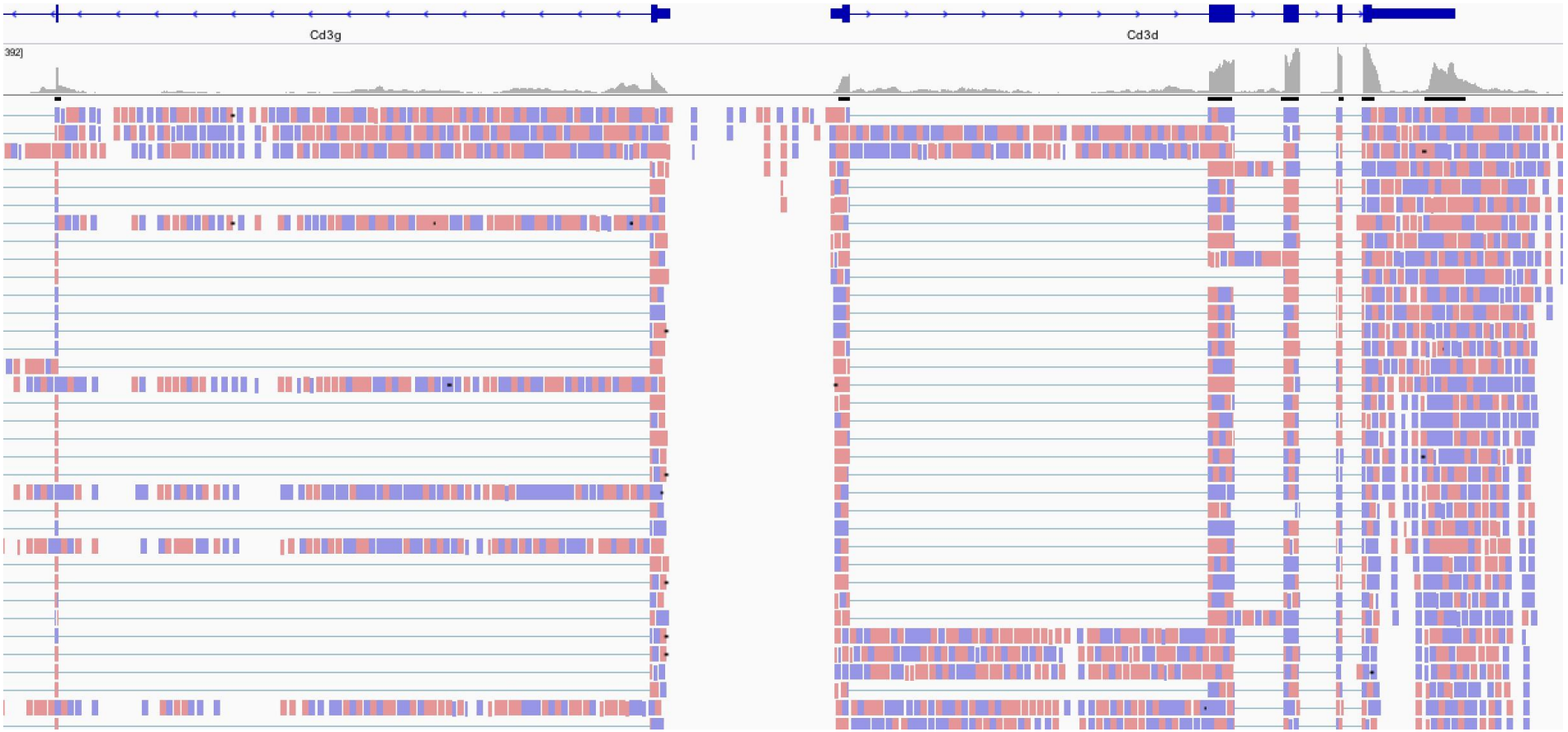
Stranded RNA-Seq result



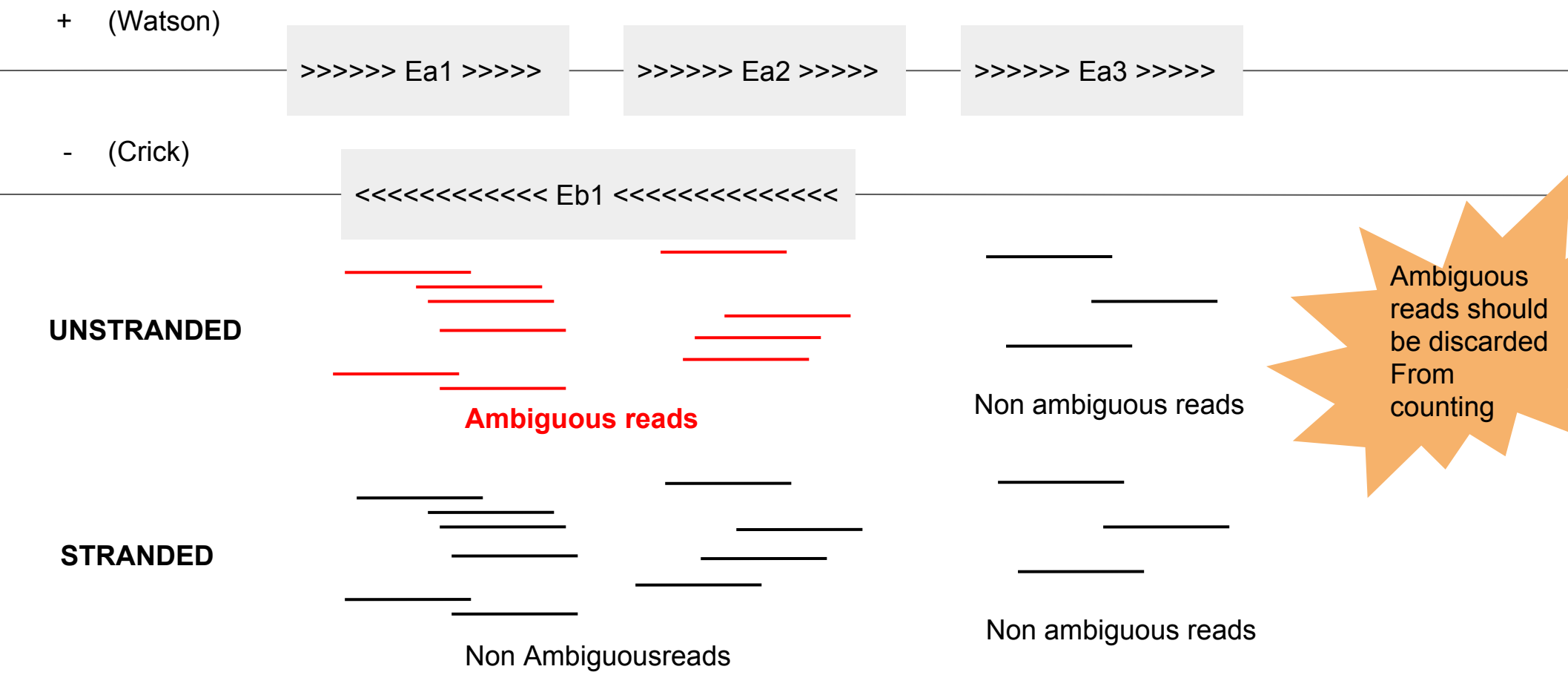
Stranded RNA-Seq allows one to extract signal produced from both strands

Example of unstranded single-end RNA-Seq alignment

Forward (Red)
Reverse (Blue)

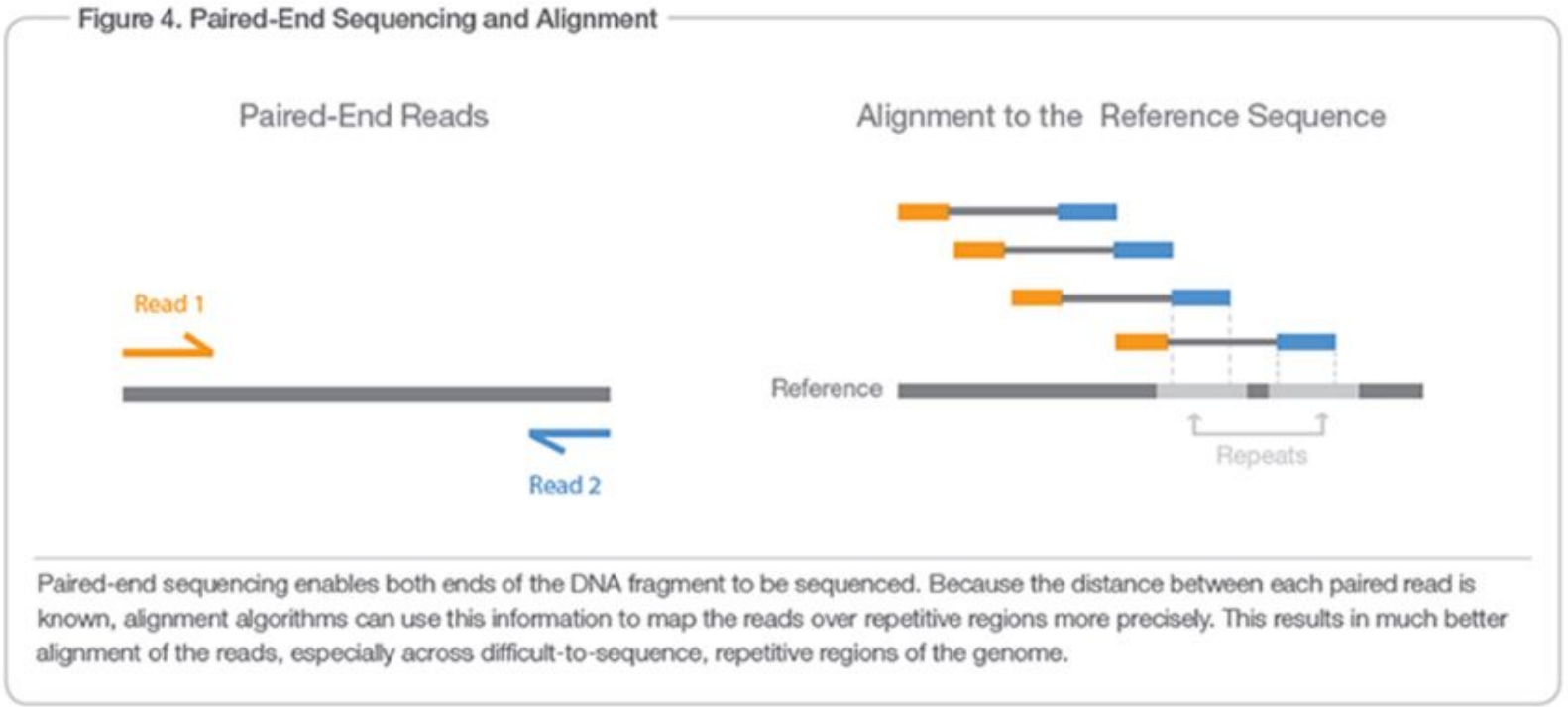


Unstranded RNA-Seq library limitations



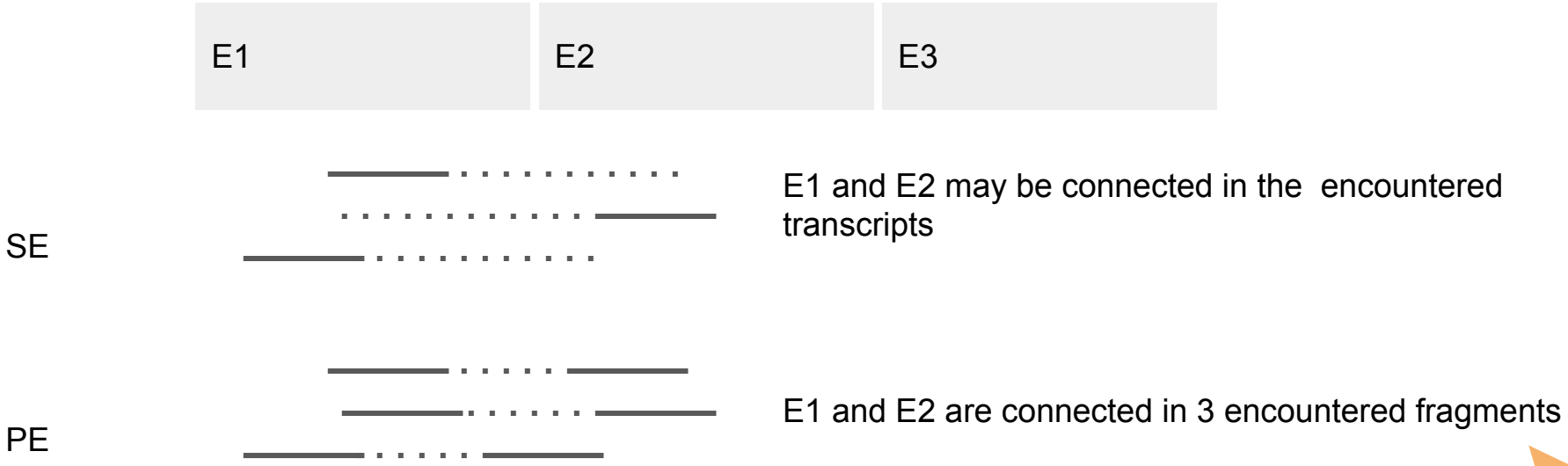
Sequencing variation: single-end vs Paired

- Paired-end sequencing: sequence both ends of a fragment
  - Facilitate alignment
  - Facilitate gene fusion detection
  - Better to reconstruct transcript model from RNA-Seq



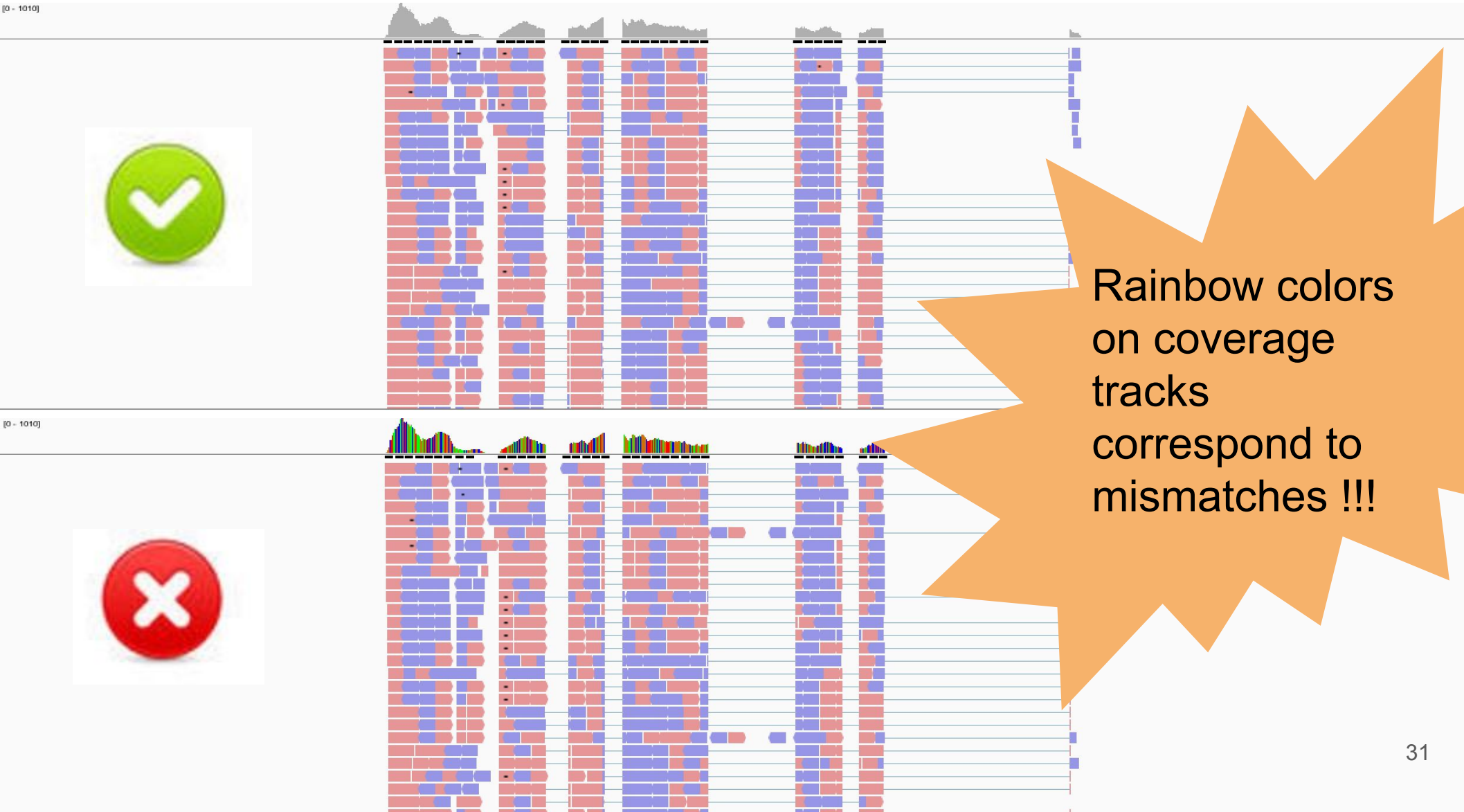
RNA-Seq library preparation: PE vs SE

- Paired-end vs Single-end
  - Better reconstruction of transcripts with Paired-end
  - Paired-end : more expensive

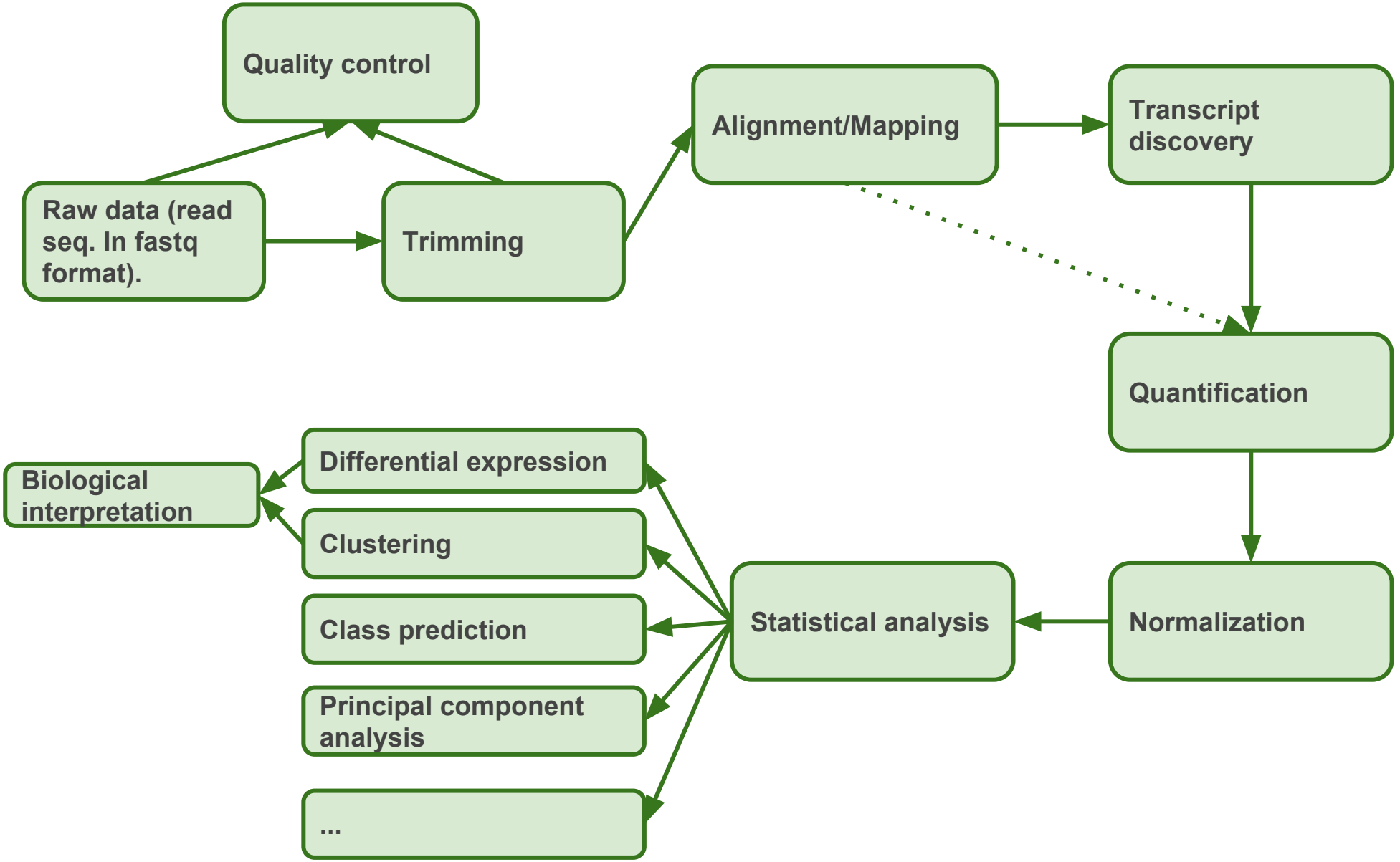


Take care to genome version

- ACTB (chr5) mm9 vs mm10 in **IGV (integrated Genome Viewer)**



Bioinformatic processing of sequencing data



The raw data are provided in fastq format

- Header
- Sequence
- + (optional header)
- Quality (Sanger quality score or other format)

```
@QSEQ32.249996 HWUSI-EAS1691:3:1:17036:13000#0/1 PF=0 length=36
GGGGGTCATCATCATTTGATCTGGGAAAGGCTACTG
+
=.+5:<<<<>AA?0A>;A*A##################
@QSEQ32.249997 HWUSI-EAS1691:3:1:17257:12994#0/1 PF=1 length=36
TGTACAACAACAACCTGAATGGCATACTGGTTGCTG
+
DDDD<BDBDB??BB*DD:D##################
```

The Sanger quality score

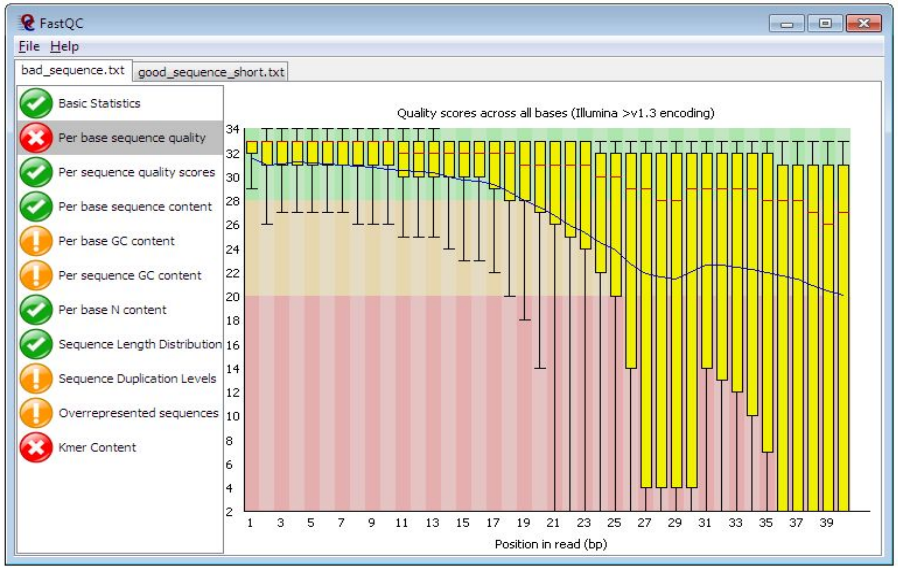
- Sanger quality score (Phred quality score): Measure the quality of each base call
  - Based on p, the probality of error (the probability that the corresponding base call is incorrect)
  - Qsanger= -10*log10(p)
  - p = 0.01 <=> Qsanger 20
- Quality score are in ASCII 33
- Note that SRA has adopted Sanger quality score although original fastq files may use different quality score (see: http://en.wikipedia.org/wiki/FASTQ_format)

ASCII 33

- Storing PHRED scores as single characters gave a simple and space efficient encoding:
- Character ”!” means a quality of 0
- Range 0-40

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

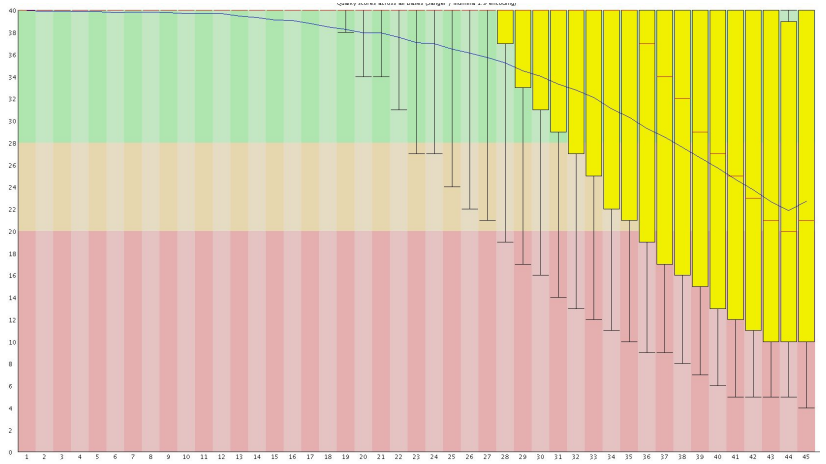
Quality control for high throughput sequence data



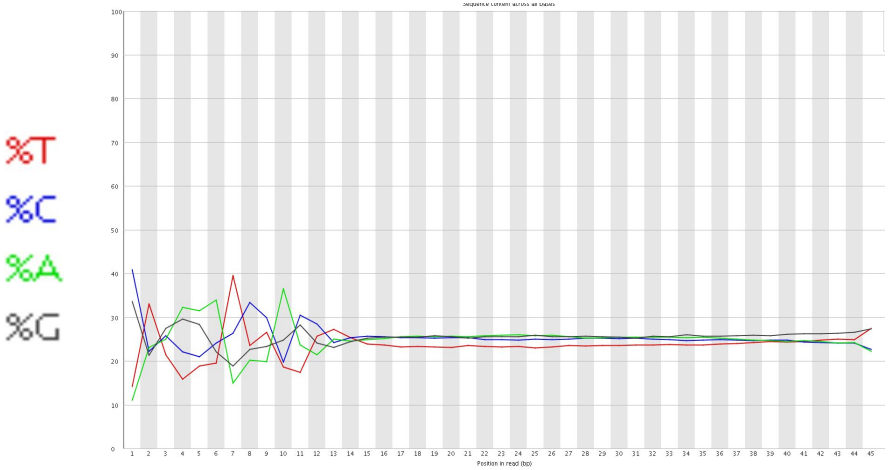
- First step of analysis
  - Quality control
  - Ensure proper quality of selected reads.
    - The importance of this step depends on the aligner used in downstream analysis

Quality control with FastQC program

Quality

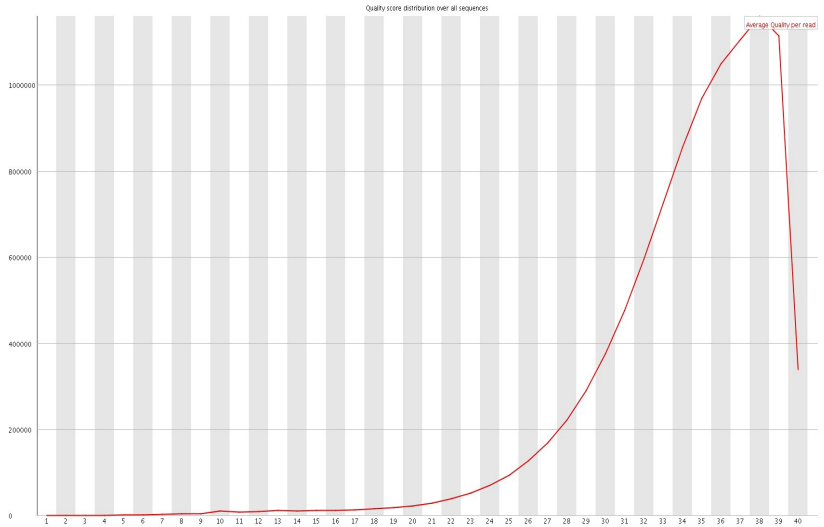


Position in read



Position in read

Nb Reads

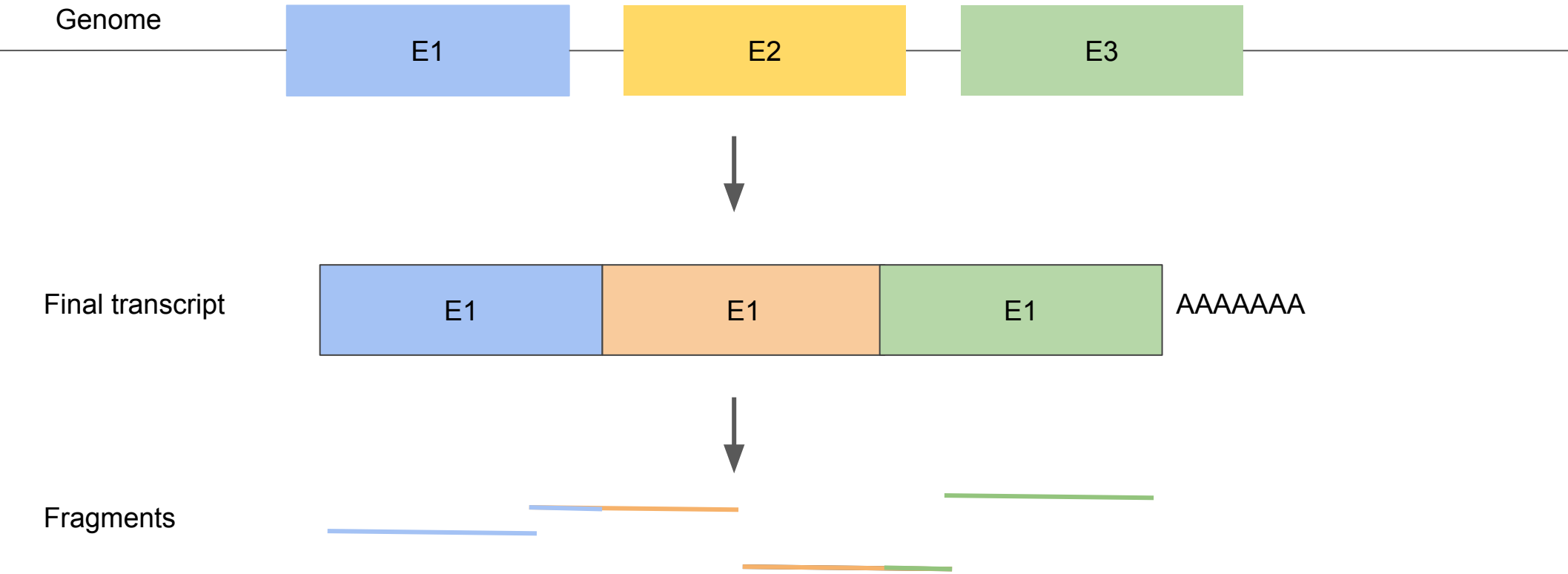


Mean Phred Score

Look also at over-represented sequences

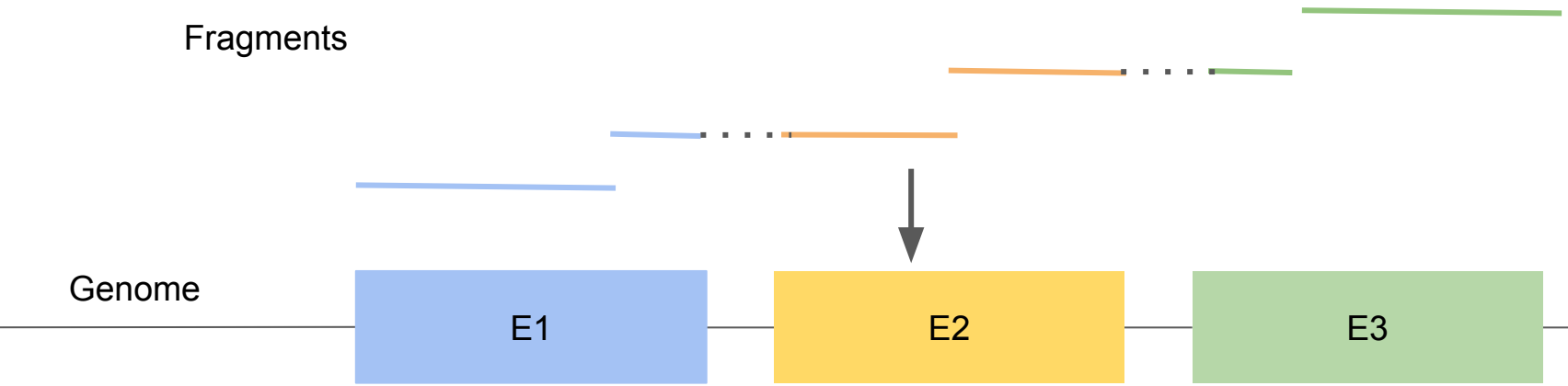
Alignement: splice-aware aligners

- Reads that overlaps several exons may not be mapped properly by splice-unaware aligners (e.g bowtie)



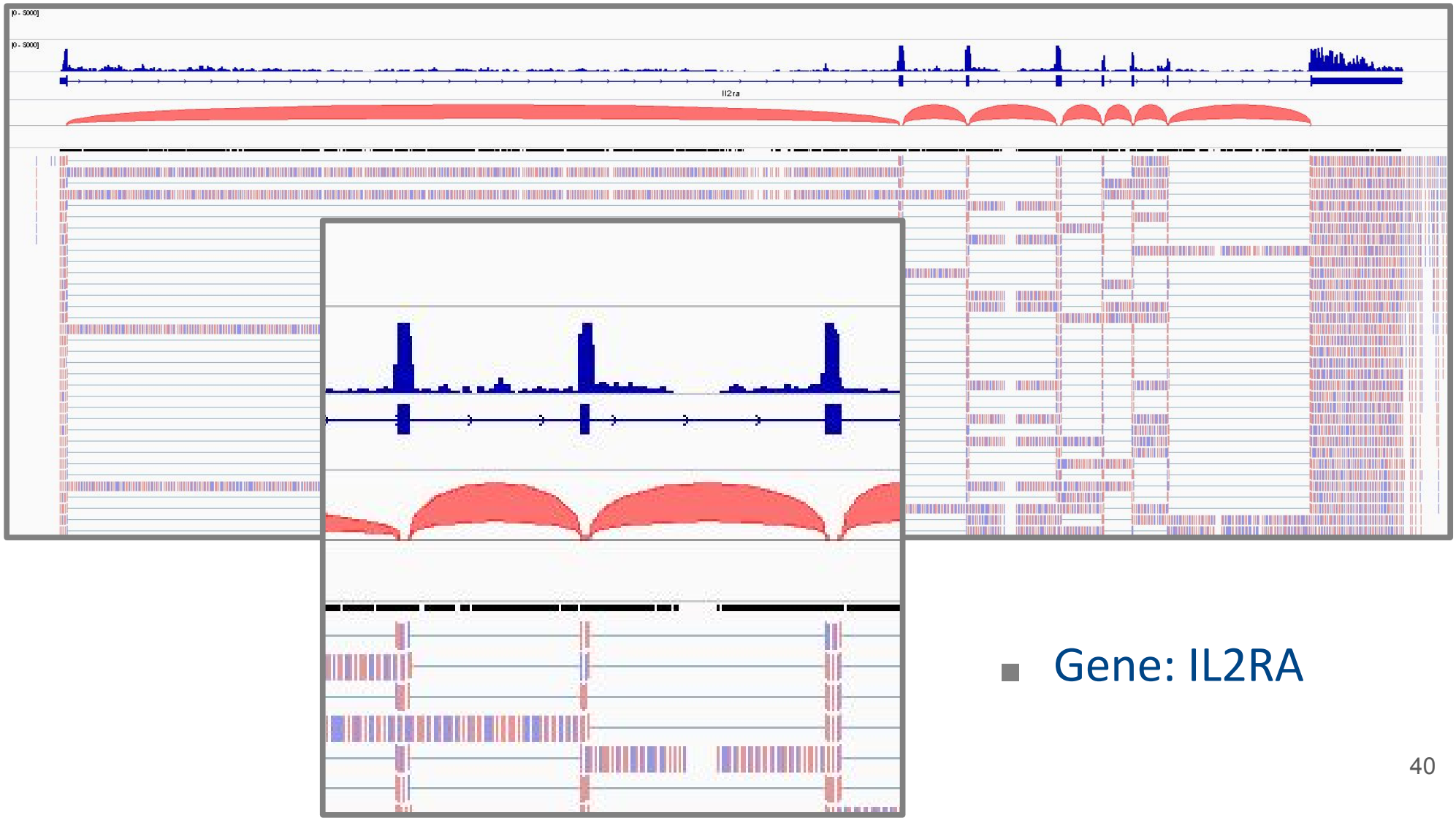
Splice-aware aligners ?

- Reads that overlaps several exons may not be mapped properly by splice-unaware aligners (e.g bowtie)



We will obtained spliced reads (gapped alignments)

RNA-Seq: aligned reads (Stranded paired-end sequencing on Total RNA)



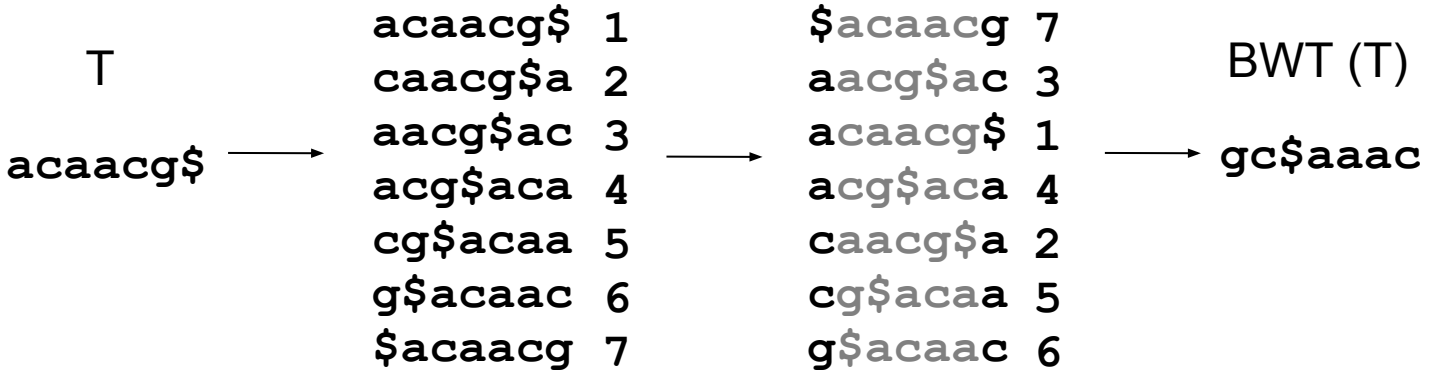
- Gene: IL2RA

Example of splice aware aligners

- Tophat
  - Part of a complete pipeline (the tuxedo pipeline)
  - Make call to **bowtie** to perform **initial, unspliced-alignments**
- STAR
  - Developed in the context of ENCODE project
  - Very fast (>> compared to tophat)
  - Need ~30Go of memory for human/mouse genome
    - Based on a associative table (hash).
  - Usage is painful
  - Compatible with the tuxedo pipeline

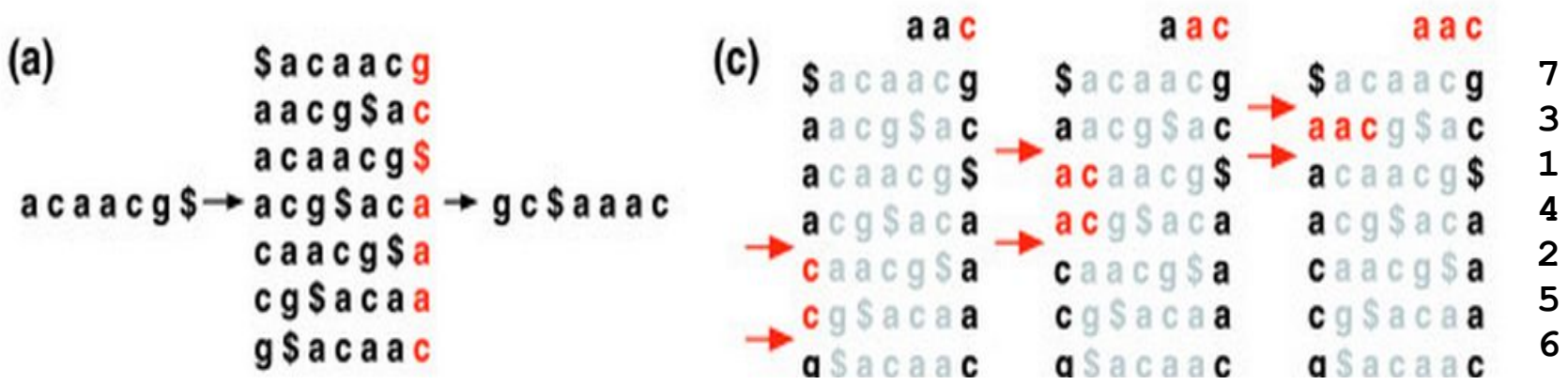
Behind tophat: Bowtie a very popular aligner (for unspliced alignments)

- Burrows Wheeler Transform-based algorithm
- Two phases: "**seed and extend**".
- The Burrows-Wheeler Transform of a text T, BWT(T), can be constructed as follows.
  - The character $ is appended to T, where $ is a character not in T that is lexicographically less than all characters in T.
  - The Burrows-Wheeler Matrix of T, BWM(T), is obtained by computing the matrix whose rows comprise all cyclic rotations of T sorted lexicographically.



Bowtie principle

- Burrows-Wheeler Matrices have a property called the Last First (LF) Mapping.
  - The ith occurrence of character c in the last column corresponds to the same text character as the ith occurrence of c in the first column



TopHat pipeline

- RNA-Seq **reads are mapped** against the whole reference genome (**bowtie**).
- TopHat allows Bowtie to report more than one alignment for a read (default=10), and suppresses all alignments for reads that have more than this number
- Reads that do not map are set aside (**initially unmapped reads**, or IUM reads)
- TopHat then **assembles** the mapped reads using the assembly module in Maq. An initial consensus of mapped regions is computed.
- The ends of exons in the pseudoconsensus will initially be covered by few reads (most reads covering the ends of exons will also span splice junctions)
    - Tophat **add a small amount of flanking sequence** of each island (default=45 bp).

Bioinformatics. 2009 May 1;25(9):1105-11. Epub 2009 Mar 16.

TopHat: discovering splice junctions with RNA-Seq.

Trapnell C, Pachter L, Salzberg SL.

TopHat pipeline

- Weakly expressed genes should be poorly covered
  - Exons may have gaps
- To map reads to splice junctions, TopHat first **enumerates all canonical donor and acceptor sites** within the island sequences (as well as their reverse complements)
- Next, tophat **considers all pairings** of these sites that could form canonical (GT–AG) introns between neighboring (but not necessarily adjacent) islands.
  - By default, TopHat examines potential introns longer than 70 bp and shorter than 20 000 bp (more than 93% of mouse introns in the UCSC known gene set fall within this range)
- **Sequences** flanking potential donor/acceptor splice sites within neighboring regions **are joined to form potential splice junctions**.
- Read are **mapped onto these junction library**

Mapping read spanning exons

Aligner output: SAM/BAM files

- SAM = ‘Sequence Alignment/MAP’
- BAM: binary/compressed version of SAM
- Store information related to alignments
  - Read ID
  - Alignment position
  - Mapping quality
  - CIGAR String
  - Bitwise FLAG
    - read paired, read mapped in proper pair, read unmapped, ...
  - ...

Sequence Alignment/Map Format Specification

The SAM/BAM Format Specification Working Group

Bitwise flag

- read paired
- read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate

Bitwise flag

- 00000000001 → 2^0 = 1 (read paired)
- 00000000010 → 2^1 = 2 (read mapped in proper pair)
- 00000000100 → 2^2 = 4 (read unmapped)
- 00000001000 → 2^3 = 8 (mate unmapped) …
- 00000010000 → 2^4 = 16 (read reverse strand)
- 00000001001 → 2^0+ 2^3 = 9 → (read paired, mate unmapped)
- 00000001101 → 2^0+2^2+2^3 =13 ...
- ...

http://picard.sourceforge.net/explain-flags.html

The extended CIGAR string

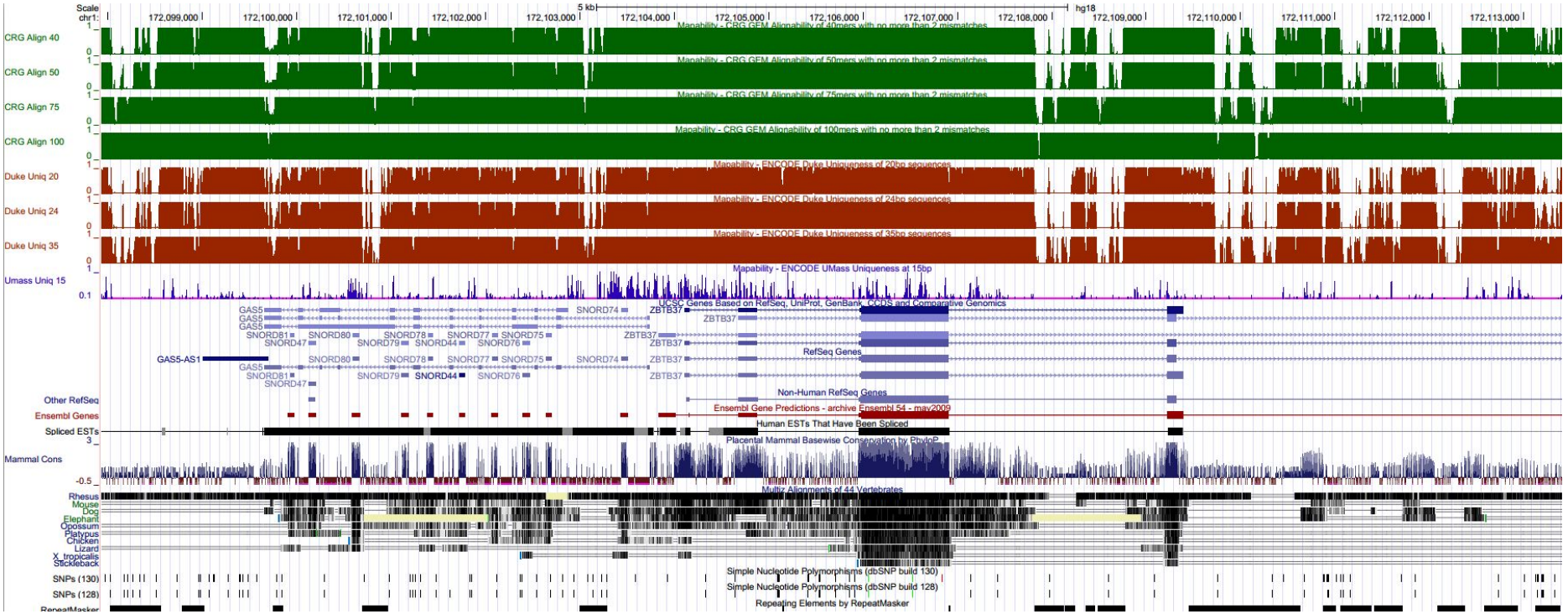
- Exemple flags:
  - M alignment match (can be a sequence match or mismatch !)
  - I insertion to the reference
  - D deletion from the reference
  - http://samtools.sourceforge.net/SAM1.pdf

ATTCAGATGCAGTA
ATTCA--TGCAGTA

5M2D7M

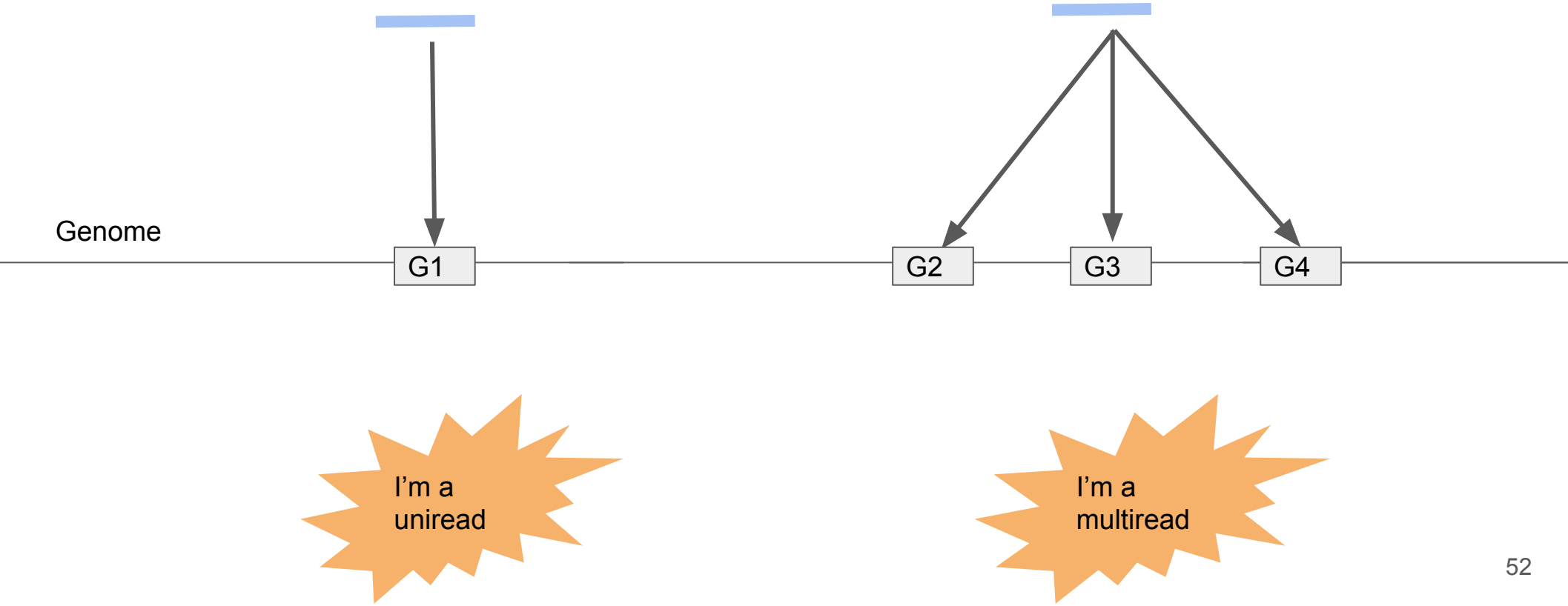
Mappability issues

- Mappability: sequence uniqueness of the reference
- Mappability = 1/(#genomic position for a given word)
- Mappability of 1 for a unique k-mer
- Mappability < 1 for a non unique k-mer



Uniread ? Multireads ?

- First aligners defined the notions of uni-reads and multireads
- An uniread is thought to map to a single position on the genome
- A multiread is thought to map to several position on the genome
  - Which position/gene produced the signal ?

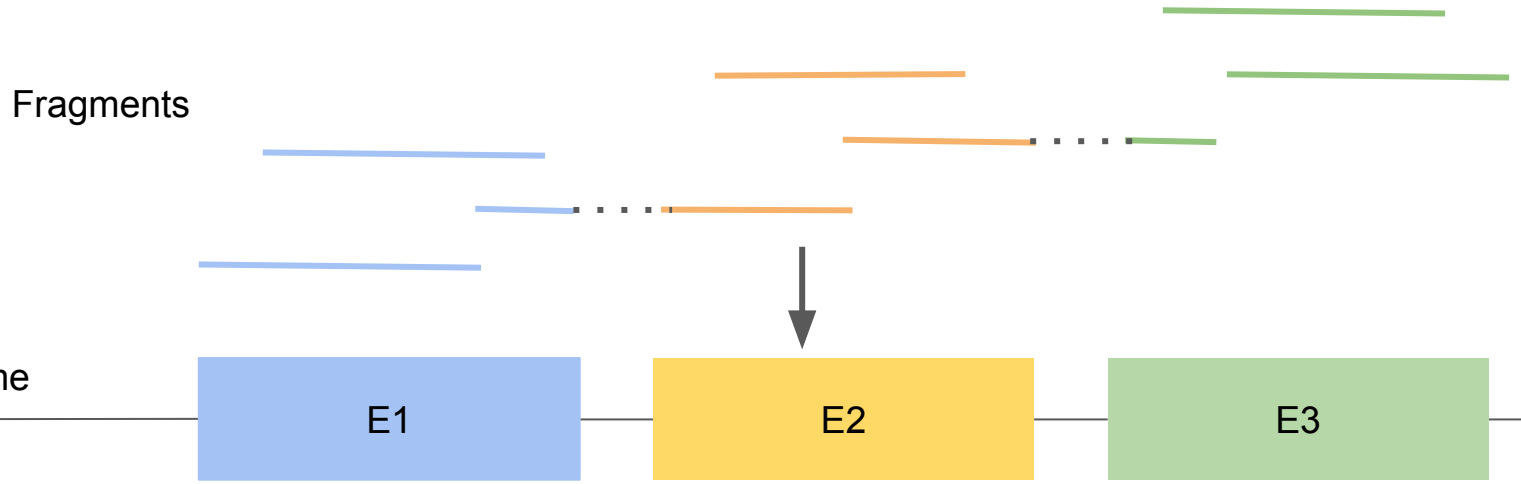


Uniread ? Multireads ?

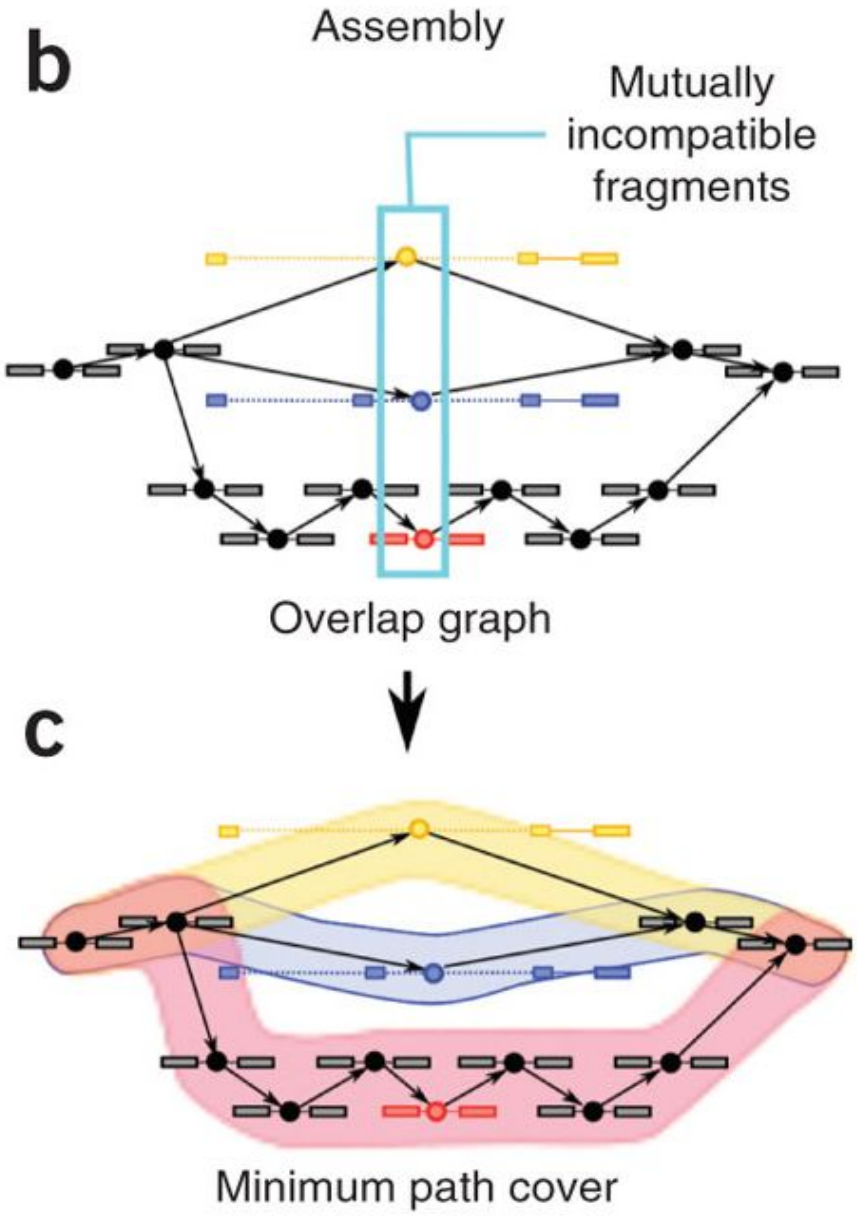
- Several aligners still use this notion
  - E.g tophat(2)
  - See -x -g arguments
- The notion has been superseded by the mapping quality score.
  - Mapping quality score indicates is computed from the probability that alignement is wrong
  - -log10(prob. alignment is wrong)
- It is particularly advised to take into account mapping quality (e.g by selecting high quality alignments from the BAM file)
  - Samtools view -q 30 file.bam

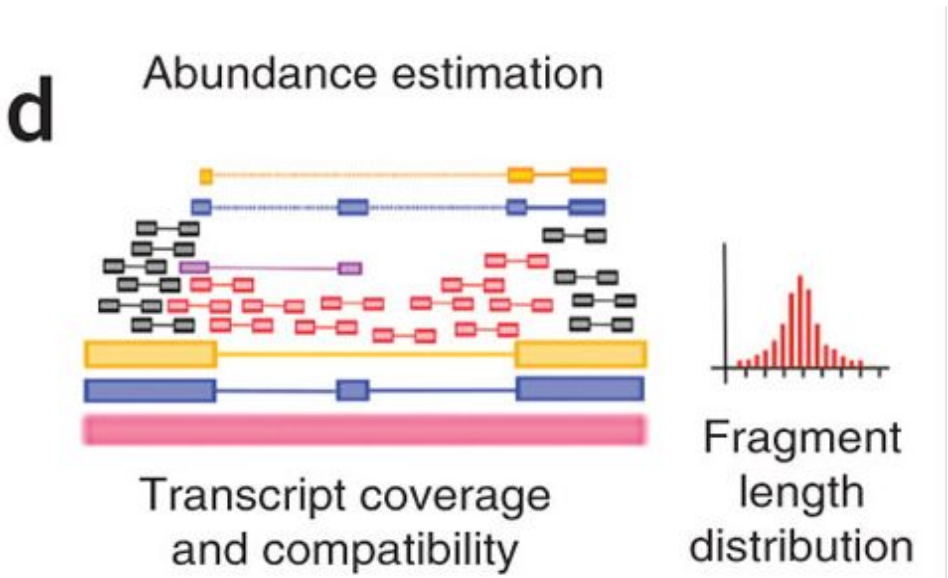
Searching for novel transcript models

- RNA-Seq may be used to discover novel transcripts inside the dataset
- Several software:
  - Cufflinks, MATS, MISO…
- Cufflinks is the most popular
  - Performs much better with stranded RNA-Seq
  - Analyse read overlap to infer transcript structure



Searching for novel transcript model: cufflinks



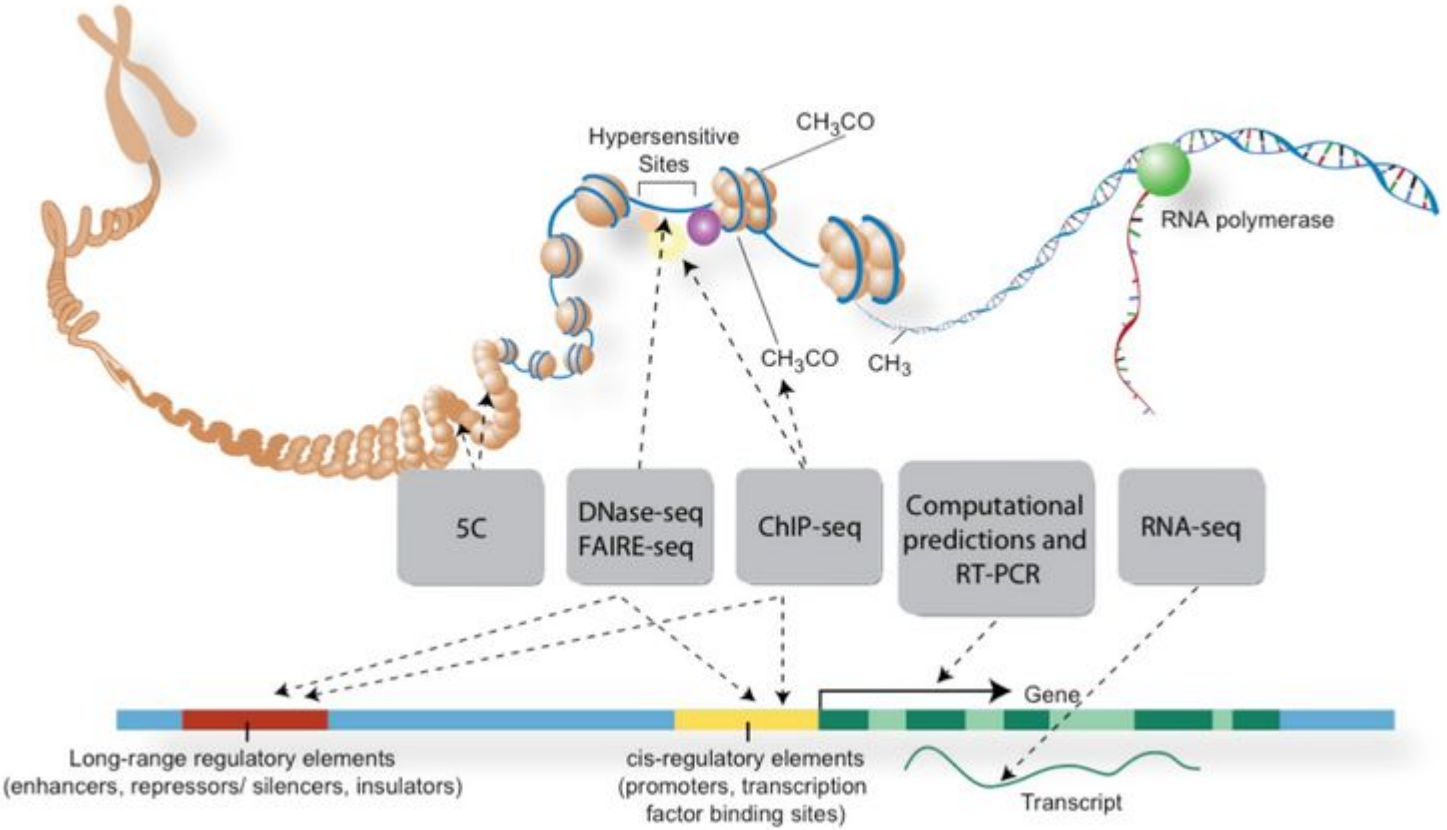


Read pair

Gapped alignment

Transcript discovery in the context of the ENCODE project

- E.g ENCODE (Encyclopedia Of DNA Elements)
  - A catalog of express transcripts



Some key results of ENCODE analysis

- 15 cell lines studied
  - RNA-Seq, CAGE-Seq, RNA-PET
  - Long RNA-Seq (76) vs short (36)
  - Subnuclear compartments
    - chromatin, nucleoplasm and nucleoli

- Human genome coverage by transcripts
  - 62.1% covered by processed transcripts
  - 74.7 % covered by primary transcripts,
  - Significant reduction of ”intergenic regions”
  - 10–12 expressed isoforms per gene per cell line

Nature. 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233.

Landscape of transcription in human cells.

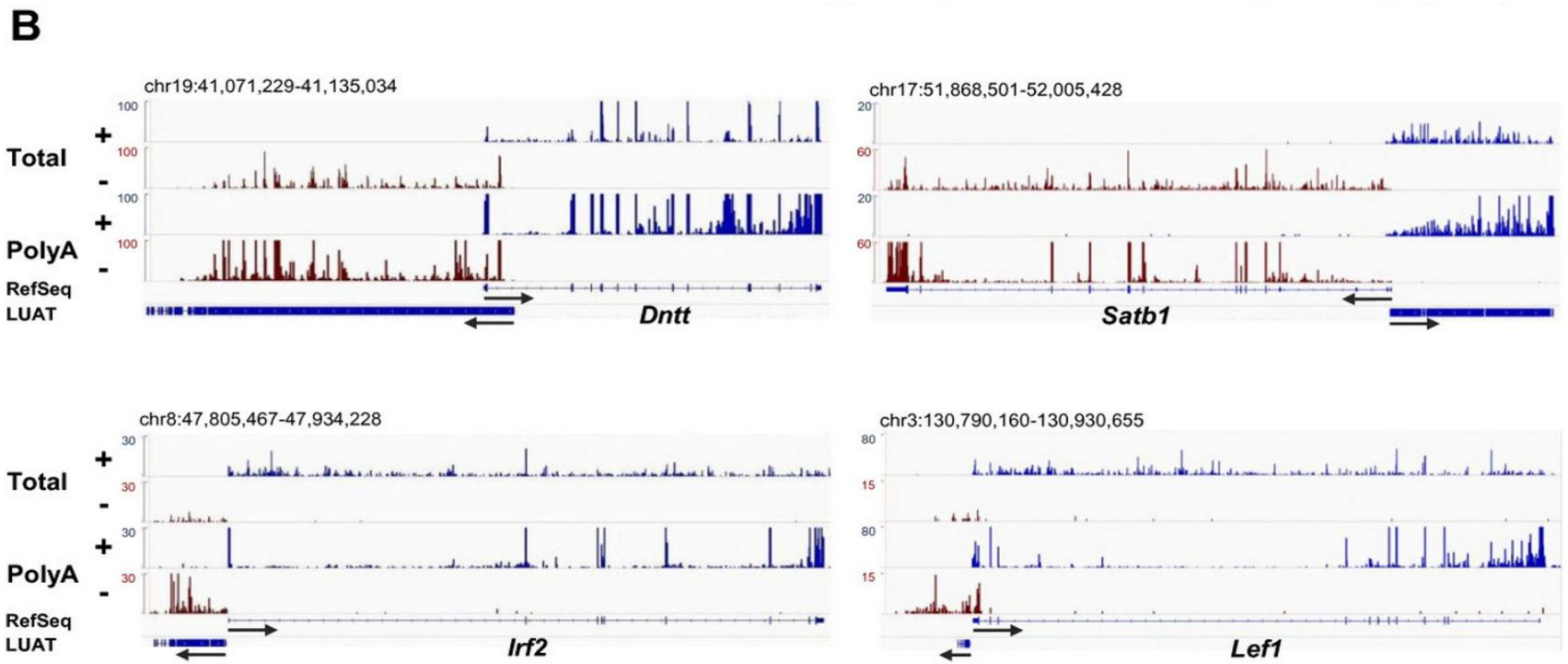
Djebali S[1], Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR.

The world of long non-coding RNA (LncRNA)

- Long: i.e cDNA of at least 200bp
- A considerable fraction (29%) of lncRNAs are detected in only one of the cell lines tested (vs 7% of protein coding)
- 10% expressed in all cell lines (vs 53% of protein-coding genes)
- More weakly expressed than coding genes
- The nucleus is the center of accumulation of ncRNAs

Some LncRNA are functional

- Some results regarding their implication in cancer
- May help recruitment of chromatine modifiers
- May also reveal the underlying activity of enhancers
- A large fraction are divergent transcripts



The Gencode database (hs/mm)

Version 25 (March 2016 freeze, GRCh38) - Ensembl 85

General stats

Total No of Genes	58037
Protein-coding genes	19950
Long non-coding RNA genes	15767
Small non-coding RNA genes	7258
Pseudogenes	14650
- processed pseudogenes:	10725
- unprocessed pseudogenes:	3400
- unitary pseudogenes:	214
- polymorphic pseudogenes:	51
- pseudogenes:	21
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments:	411
- pseudogenes:	239

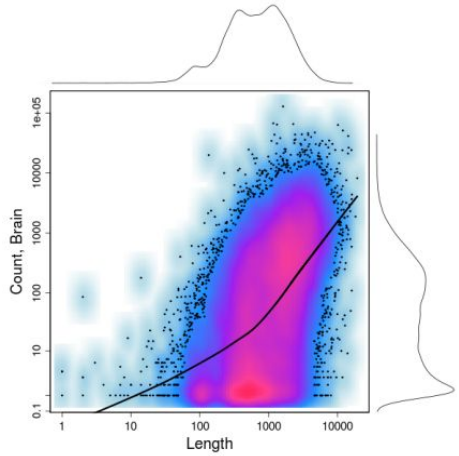
Total No of Transcripts	198093
Protein-coding transcripts	80087
- full length protein-coding:	54755
- partial length protein-coding:	25332
Nonsense mediated decay transcripts	13769
Long non-coding RNA loci transcripts	27692
Total No of distinct translations	60033
Genes that have more than one distinct translations	13536

Quantification

- Objective
  - Count the number of reads or fragments (PE) that fall in each gene
    - featureCounts, HTSeq-count,...
    - The output is a **count matrix (or expression matrix)**

Quantification

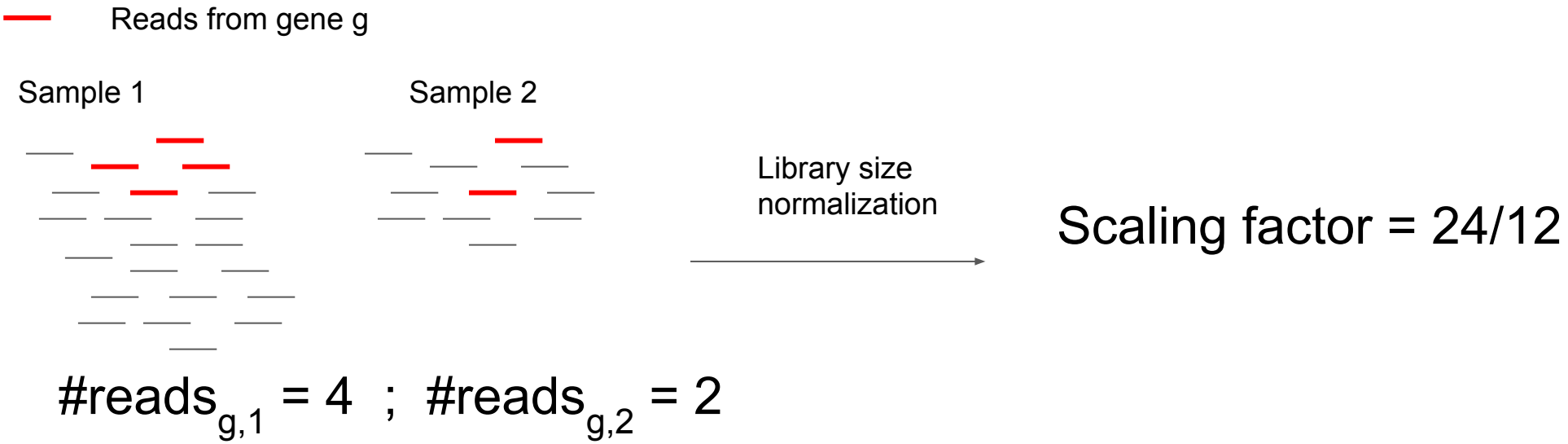
- Quantification is most generally performed **at the gene level**
  - Some specialized software may provide you with transcript abundance **estimations**
    - **Cufflinks (tuxedo pipeline)**
    - **Kallisto**
- Known issues
  - Positive association between gene counts and length
    - May be problematic for gene-wise comparisons
    - Suggests higher expression among longer genes
  - Unstranded data may lead to ambiguous reads that should be discarded



(a) Count vs. length

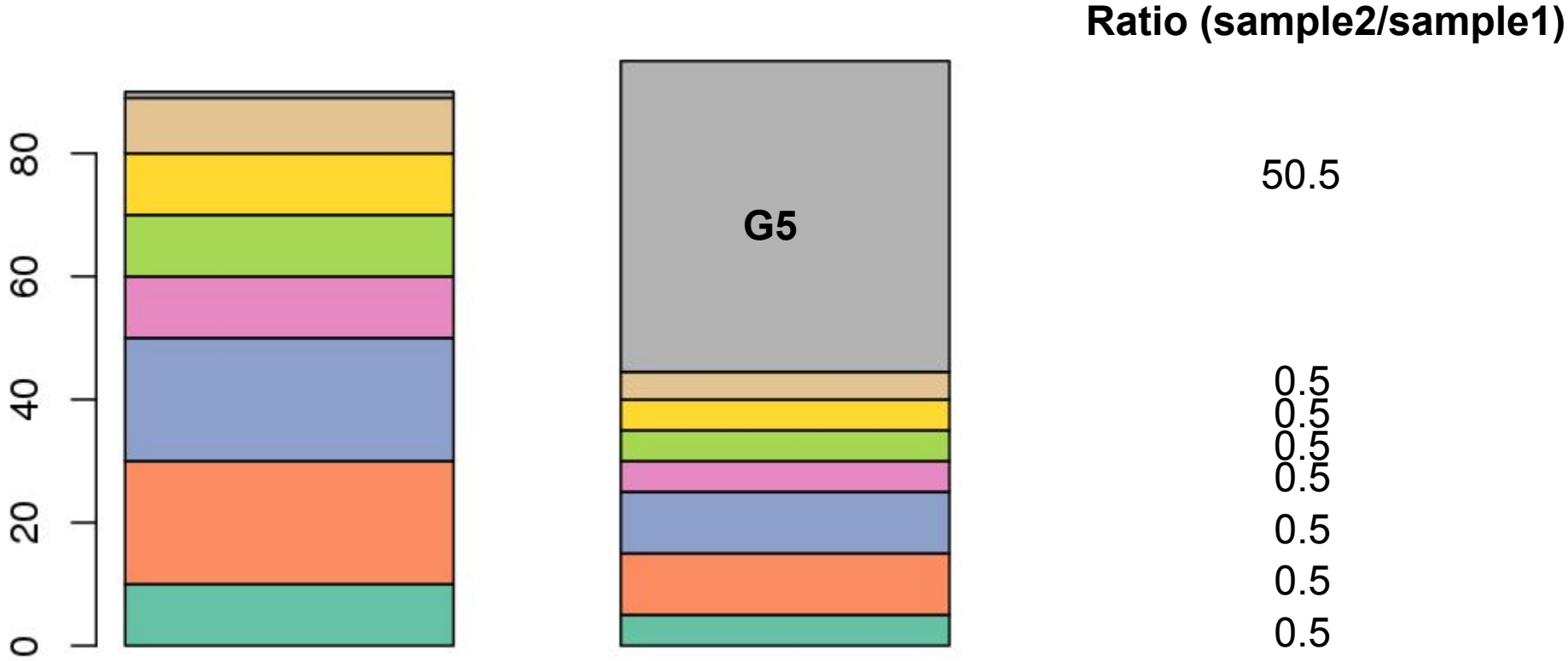
Intersample normalization: library size

- Inter-sample normalization is a prerequesite for differential expression analysis
- This normalization is mostly applied because of some imbalance in read counts between
  - Here sample 1 has 2 times more reads (24 vs 12)
  - Gene g expression will be overestimated in sample 1 although is expression is unchanged
  - A basic normalisation factor could be the library size (total number of reads)

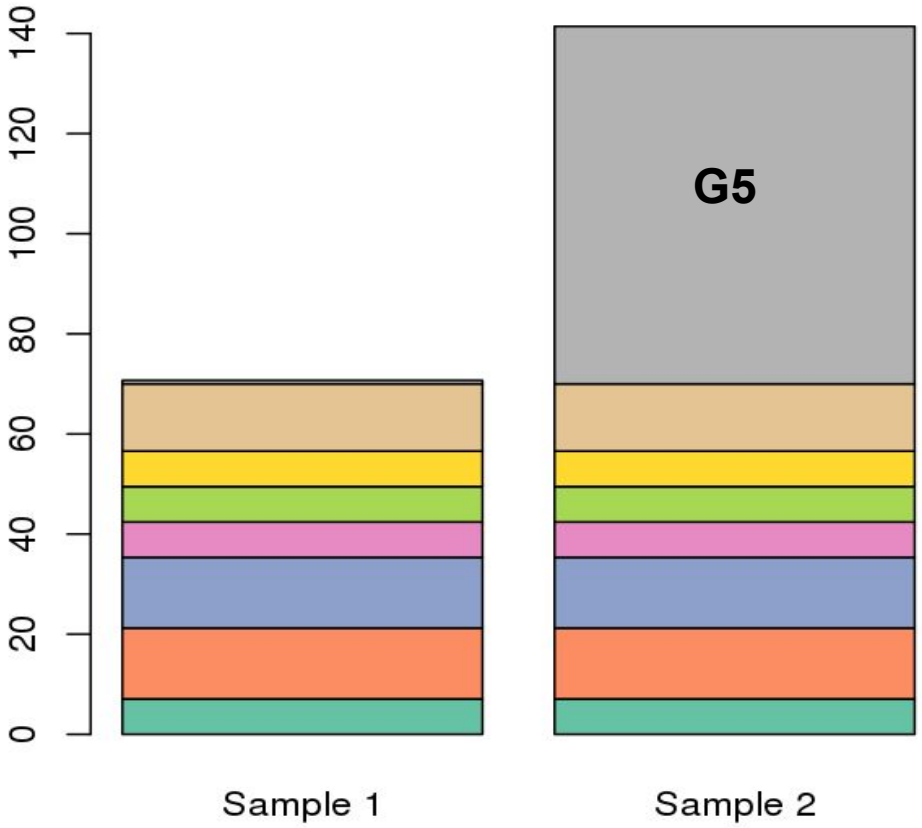


Inter-sample normalization: limits of library size

- If a large number of genes are highly expressed in, one experimental condition, the expression of the remaining genes will artefactually appear as decreased.
  - Can force the differential expression analysis to be skewed towards one experimental condition.



TMM Normalization (Robinson and Oshlack, 2010)



- Trimmed Mean of M values
- Outline
  - Compute the M values (log ratio).
    - Take the trimmed mean of the M value as scaling factor.
    - Multiply read counts by scaling factor (they multiply to one)
    - If more than two columns
      - The library whose 3rd quartile is closest to the mean of 3rd quartile is used.
    - **Very similar to RLE**

Genome Biol. 2010;11(3):R25. Epub 2010 Mar 2.

A scaling normalization method for differential expression analysis of RNA-seq data

Robinson MD, Oshlack A.

Intra-sample normalization

- Here the objective is to compare the expression level of genes in the same sample
  - Counts ?
    - Problem with long transcripts
      - Produce lots of fragments
      - Will appear artefactually highly expressed compare to other…
- Proposed method
  - RPKM
    - Read per kilobase per million mapped reads (SE)
  - FPKM
    - Fragment per kilobase per million mapped reads (PE)

RPKM/FPKM normalization

- 2kb transcript with 3000 alignments in a sample of 10 millions of mappable reads
  - RPKM = 3000/(2 * 10) = 150

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber[1], Manfred G Grabherr[1], Mitchell Guttman[1,2] & Cole Trapnell[1,3]

Accurate quantification of transcriptome from RNA-Seq data by effective length normalization

Soohyun Lee[1], Chae Hwa Seo[1], Byungho Lim[2], Jin Ok Yang[1], Jeongsu Oh[1], Minjin Kim[2], Sooncheol Lee[2], Byungwook Lee[1], Changwon Kang[2] and Sanghyuk Lee[1,3,*]

Differential expression analysis

- Use **statistical tests** (e.g based on **negative binomial model**) to find **differentially expressed genes**
  - **Biological replicates** prefered/needed (not technical replicates)
  - Tools:
    - EdgeR, DESeq2…
- The **list of differentially expressed genes** may be used for subsequent **analysis.**


An example list.
Pubmed query for all of them ?

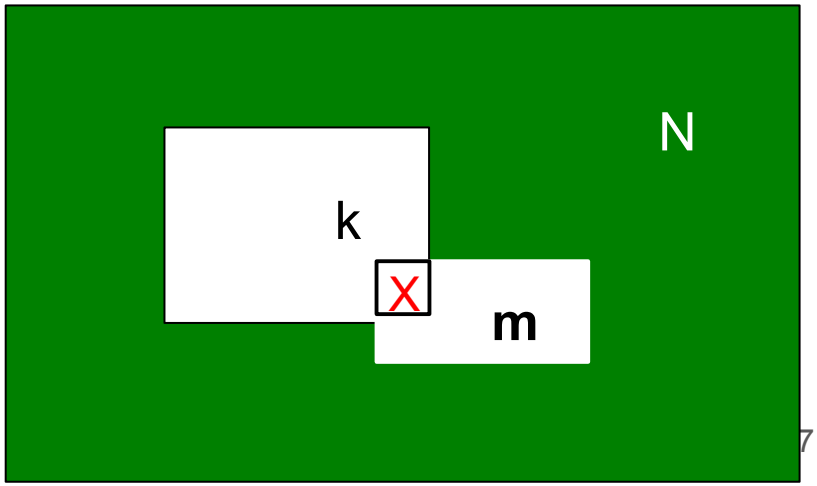
What is the biological meaning of a gene lists ?

- Example: the list of gene upregulated in tumors compared to normal counterpart.
- Is there any hidden biological meaning ?
- Solution: compare this list to known lists. Eg:
  - Gene involved in cell cycle, apoptosis, T-cell activation
  - Gene involved in chimiotactism
  - Gene whose products are located in mitochondria
  - Gene involved in a given pathway
  - Predicted targets of miRNA, transcription factors….
  - Gene located in a given chromosome
  - Genes known to be associated with mutations in a given tumor type….
  - Genes known or predicted as being regulated by a given transcription factor

Is my list enriched in gene whose function is known ?

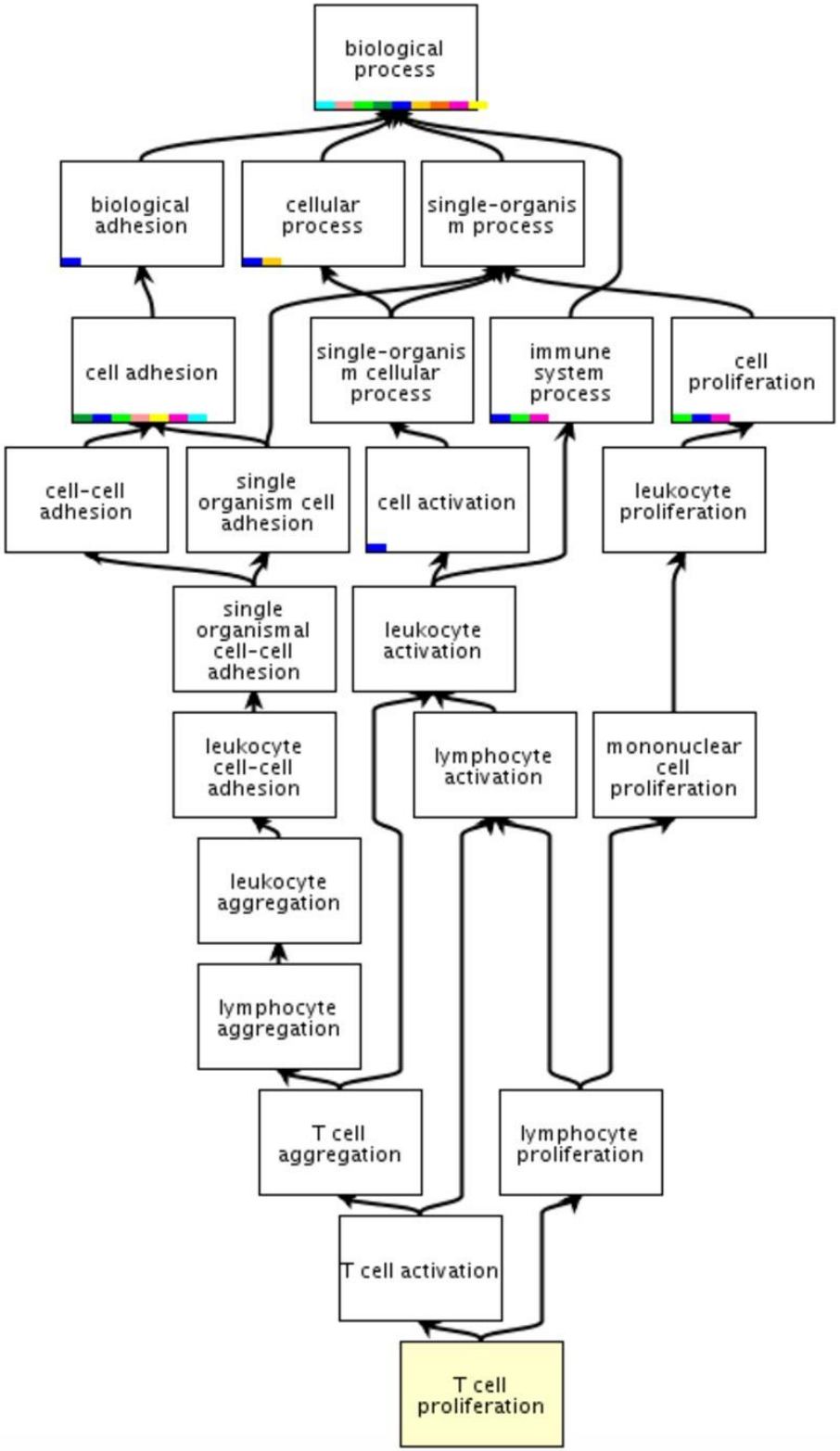
- N genes
- m genes known to be associated to a term/function T.
- n genes not associated to the term/function T.
- k selected genes (upregulated in the tumor compared to normal counterpart)
- x genes associated to term/function T in k.
- What is probability to observe x associated with term/function T in k ?
  - X follows a hypergeometric distribution
  - Hypergeometric test / Fisher exact test

	Term	!Term	
List	x	k-x	k
!List	m-x	n-(k-x)	N-k
	m (white)	n (black)	N



Where are these lists coming from ?

- Pathways: KEGG pathways, Reactome, Biocarta, GenMapp...
- Gene Ontology
  - Ontology: definition of types, properties and relationships between entities using a control vocabulary
  - The GO (http://geneontology.org/) defines concepts/classes used to **describe gene/product function**, and relationships between these concepts. It classifies functions along three aspects:
    - molecular function
      - molecular activities of gene products
    - cellular component
      - where gene products are active
    - biological process



Example GO term: T cell activation (GO:0042098)

- 225 genes in human are annotated with GO term GO/0042098:
  - E.g:
    - IL27, IRF1, CD28, CD1D, CD5, CD6, CD4, CD8, LCK, ZAP70...

The Database for Annotation, Visualization and Integrated Discovery (DAVID)

Current Gene List: demolist2
Current Background: Homo sapiens
379 DAVID IDs

⊞ **Options**

Rerun Using Options | Create Sublist

468 chart records Download File

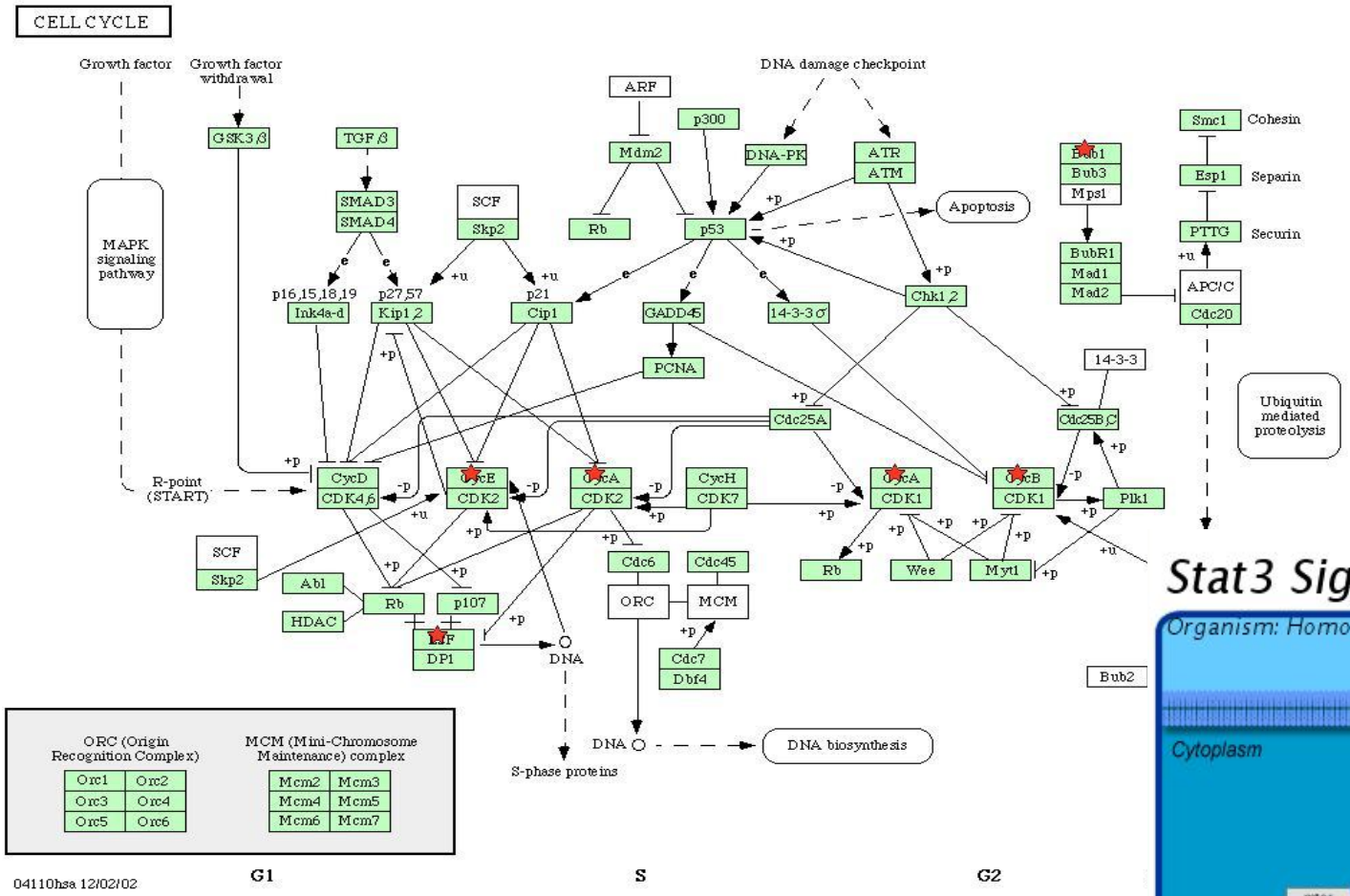
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_FAT	plasma membrane part	RT		83	21.9	2.8E-7	9.3E-5
	GOTERM_BP_FAT	response to wounding	RT		34	9.0	9.2E-7	2.2E-3
	GOTERM_BP_FAT	regulation of apoptosis	RT		42	11.1	7.3E-6	8.9E-3
	GOTERM_BP_FAT	response to organic substance	RT		39	10.3	7.6E-6	6.2E-3
	GOTERM_BP_FAT	regulation of programmed cell death	RT		42	11.1	9.2E-6	5.6E-3
	GOTERM_BP_FAT	regulation of cell death	RT		42	11.1	9.9E-6	4.8E-3
	GOTERM_BP_FAT	regulation of cell proliferation	RT		41	10.8	1.0E-5	4.1E-3
	GOTERM_MF_FAT	transcription factor activity	RT		45	11.9	1.1E-5	6.3E-3
	GOTERM_BP_FAT	inflammatory response	RT		23	6.1	1.8E-5	6.4E-3
	GOTERM_MF_FAT	sequence-specific DNA binding	RT		32	8.4	2.6E-5	7.3E-3
	GOTERM_CC_FAT	cell fraction	RT		45	11.9	4.1E-5	6.7E-3

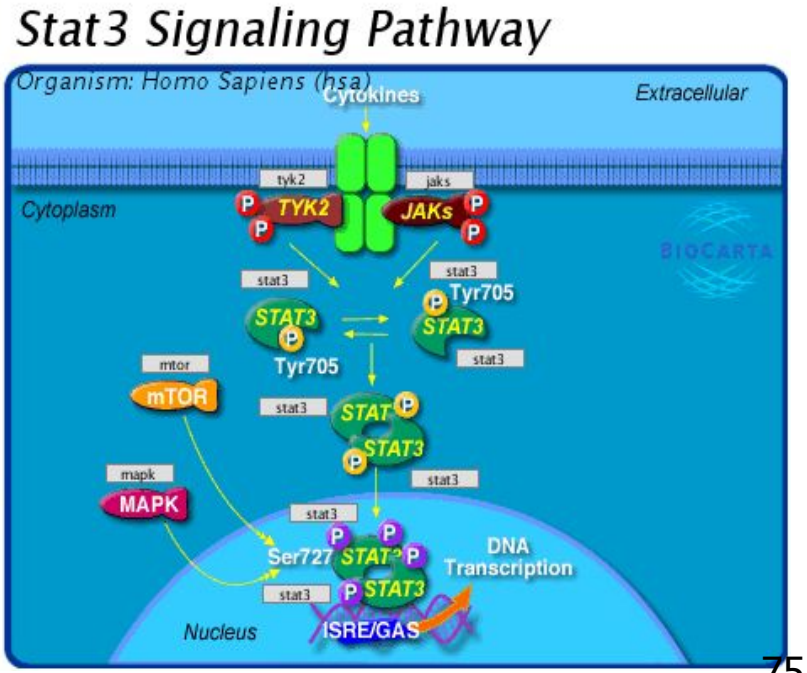
Database **Open Access**

DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis

Brad T Sherman[†1], Da Wei Huang[†1], Qina Tan[1], Yongjian Guo[4], Stephan Bour[4], David Liu[3], Robert Stephens[3], Michael W Baseler[5], H Clifford Lane[2] and Richard A Lempicki[*1]

The Database for Annotation, Visualization and Integrated Discovery (DAVID)



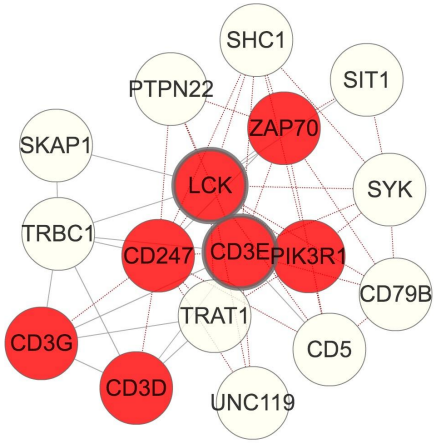


Ontologies for almost everything !

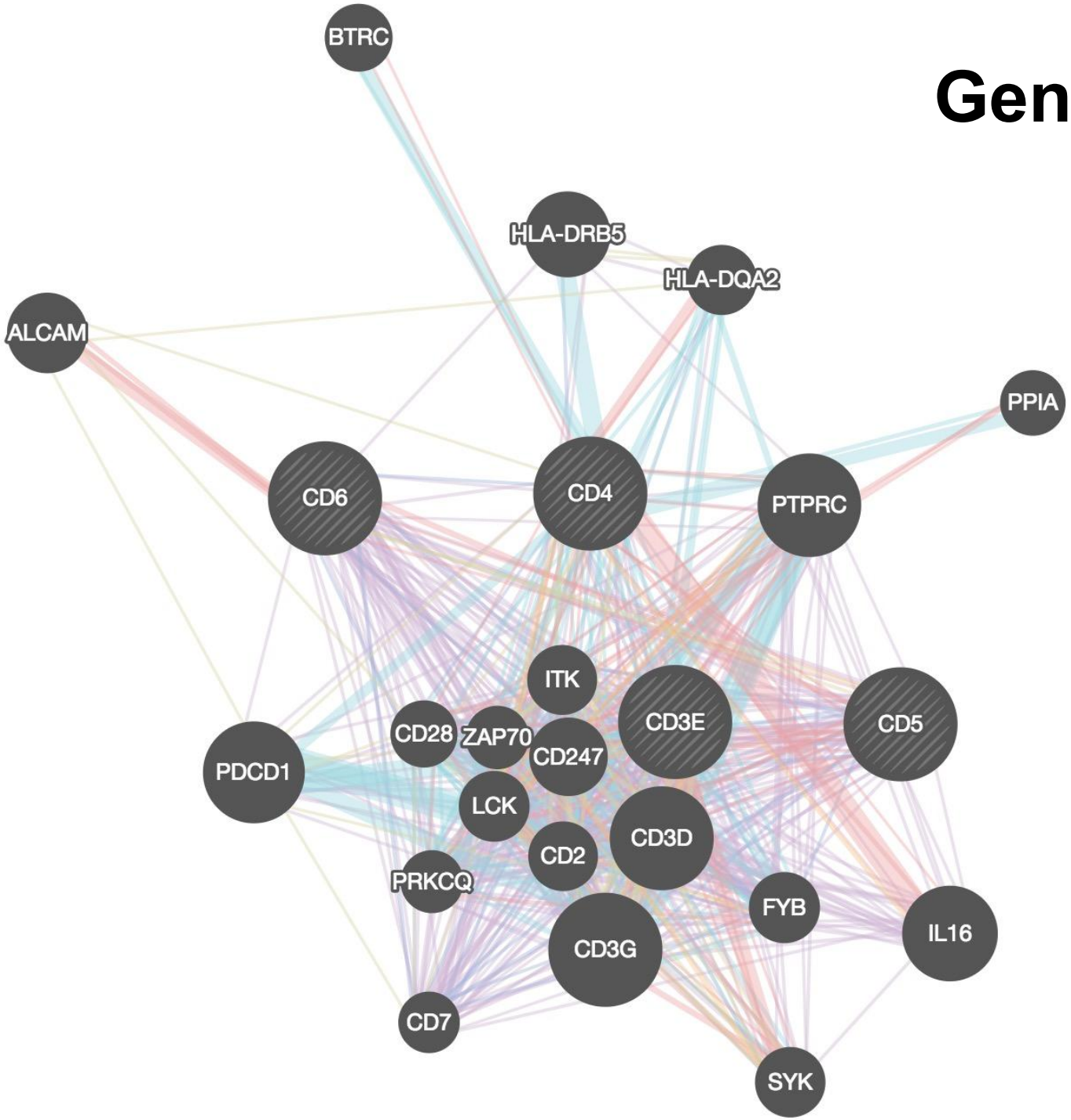
- https://www.bioontology.org/
- Bioportal at http://bioportal.bioontology.org/

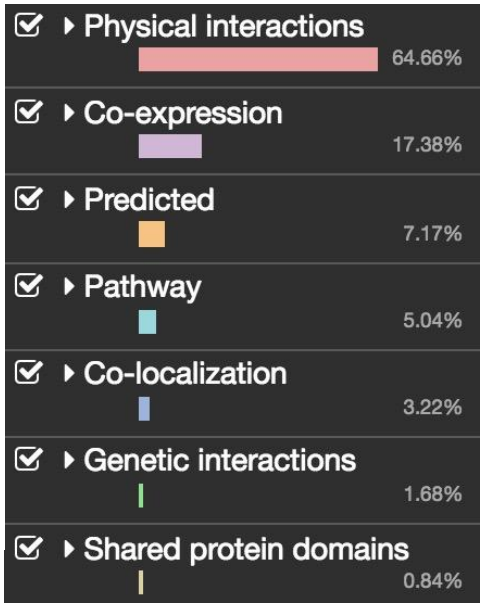
Network analysis through datamining

- Mine various databases in search for meaningful connections between gene/products
  - Interactome analysis
    - Known or predicted Protein-Protein Interactions
    - Several databases : IntAct, BioGrid, mint
      - Yeast-two-hybrid
      - Litterature…
  - Co-expression analysis
    - E.g microarray data or RNA-Seq data
      - http://coxpresdb.jp/
  - Text-mining
    - ??
  - Combined analysis
    - String
    - Reactome
    - GeneMania



GeneMania

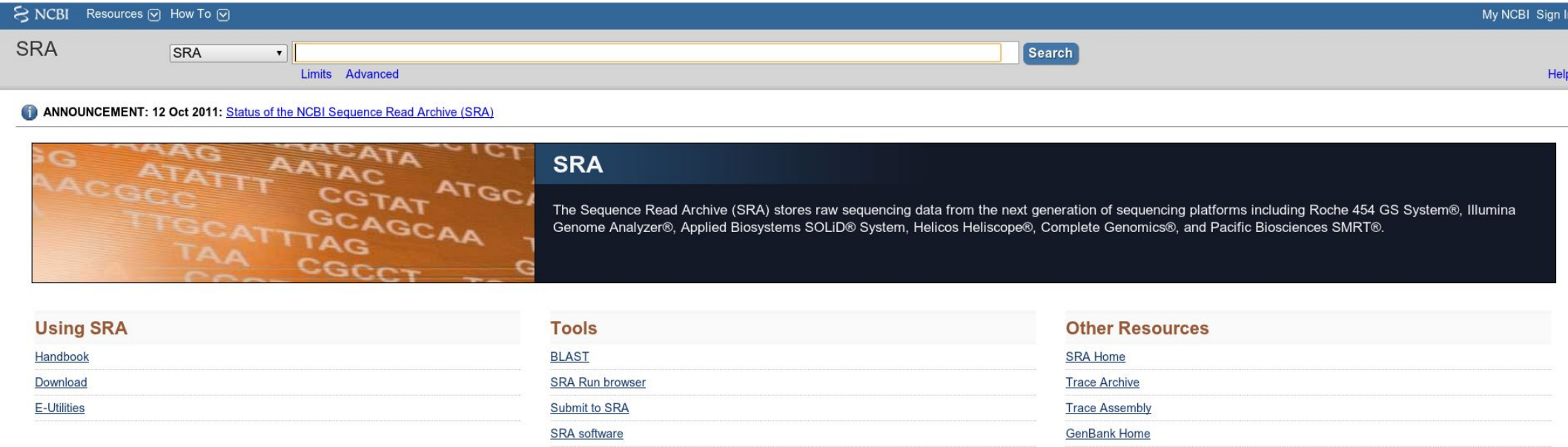

BTRC
HLA-DRB5
HLA-DQA2
ALCAM
PPIA
CD6
CD4
PTPRC
ITK
CD28
ZAP70
CD247
CD3E
CD5
PDCD1
LCK
CD3D
CD2
PRKCQ
FYB
IL16
CD3G
CD7
SYK


Physical interactions
64.66%
Co-expression
17.38%
Predicted
7.17%
Pathway
5.04%
Co-localization
3.22%
Genetic interactions
1.68%
Shared protein domains
0.84%

Yet other applications of RNA-Seq

- **Fusion transcript** analysis
  - Are there any fusion transcript specific of my tumors ?
- **Isoforms or exons-level differential analysis**
- **Allele-specific expression**
  - Preferential expression of one of the two alleles in a diploid genome
  - The allele-specific expression of a gene is attributed to a distinct epigenetic status of its two parental alleles
- Short RNA-Seq (miRNA)
- Single cell analysis
  - C1 (Fluidigm)
  - 10X Genomics

Sequence read Archive (SRA)



- The SRA archives high-throughput sequencing data that are associated with:
- RNA-Seq, ChIP-Seq, and epigenomic data that are submitted to GEO

SRA growth

Display Settings: Abstract

Send to:

Nucleic Acids Res. 2011 Oct 18. [Epub ahead of print]

The sequence read archive: explosive growth of sequencing data.

Kodama Y, Shumway M, Leinonen R; on behalf of the International Nucleotide Sequence Database Collaboration.

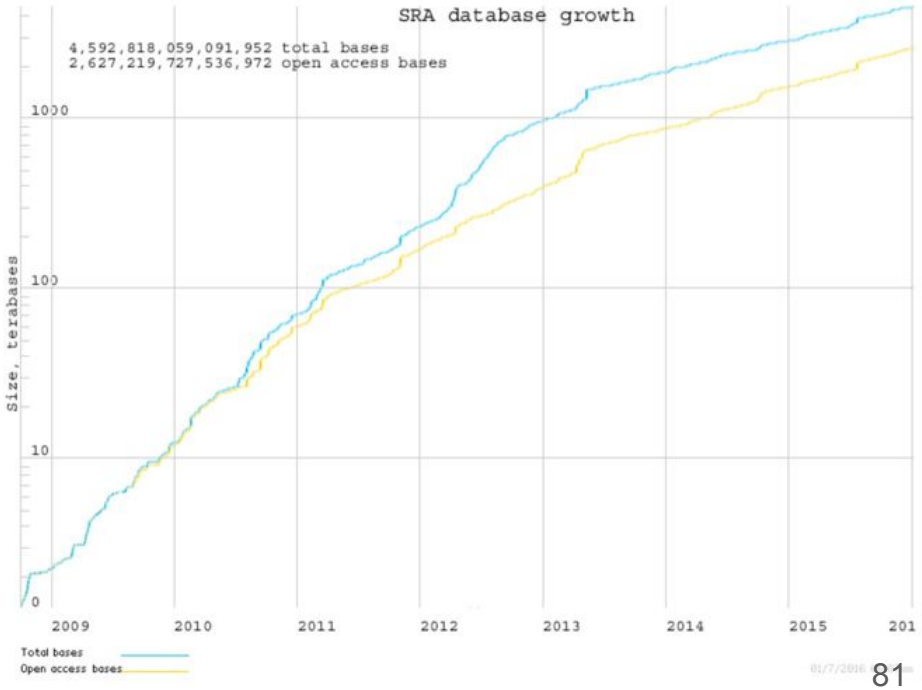
Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Abstract

New generation sequencing platforms are producing data with significantly higher throughput and lower cost. A portion of this capacity is devoted to individual and community scientific projects. As these projects reach publication, raw sequencing datasets are submitted into the primary next-generation sequence data archive, the Sequence Read Archive (SRA). Archiving experimental data is the key to the progress of reproducible science. The SRA was established as a public repository for next-generation sequence data as a part of the International Nucleotide Sequence Database Collaboration (INSDC). INSDC is composed of the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). The SRA is accessible at www.ncbi.nlm.nih.gov/sra from NCBI, at www.ebi.ac.uk/ena from EBI and at trace.ddbj.nig.ac.jp from DDBJ. In this article, we present the content and structure of the SRA and report on updated metadata structures, submission file formats and supported sequencing platforms. We also briefly outline our various responses to the challenge of explosive data growth.

PMID: 22009675 [PubMed - as supplied by publisher] **Free full text**

In 2011 the SRA surpassed 100 Terabases of open-access genetic sequence reads from next generation sequencing technologies. The Illumina™ platform comprises 84% of sequenced bases, with SOLiD™ and Roche/454™ platforms accounting for 12% and 2%, respectively. The most active SRA submitters in terms of submitted bases are the Broad Institute, the Wellcome Trust Sanger Institute and Baylor College of Medicine with 31, 13 and 11%, respectively. The largest individual global project generating next-generation sequence is the 1000 Genomes project which has contributed nearly one third of all bases. The most sequenced organisms are *Homo sapiens* with 61%, human metagenome with 6% and *Mus musculus* with 5% share of all bases. The common

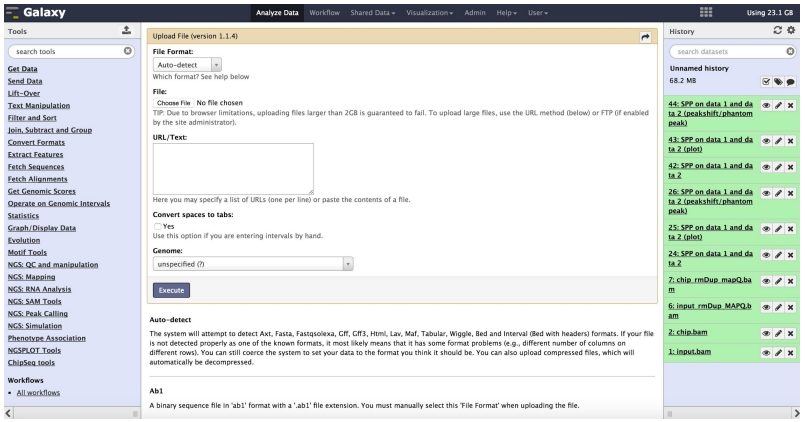


Tools to create reproducible workflows

- https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems
- E.g make, snakemake, galaxy, taverna...

Galaxy server (https://usegalaxy.org/)

- Interface to a computing cluster
- Highly flexible
  - Large palette of bioinformatic programs
  - ‘Easy’ to add your own
- Fully reproducible workflows



Snakemake

- A make-like solution

```
rule targets:
    input:
        "plots/dataset1.pdf",
        "plots/dataset2.pdf"

rule plot:
    input:
        "raw/{dataset}.csv"
    output:
        "plots/{dataset}.pdf"
    shell:
        "somecommand {input} {output}"
```

Merci