Introduction to transcriptome analysis using high-throughput sequencing technologies

D. Puthier 2015

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

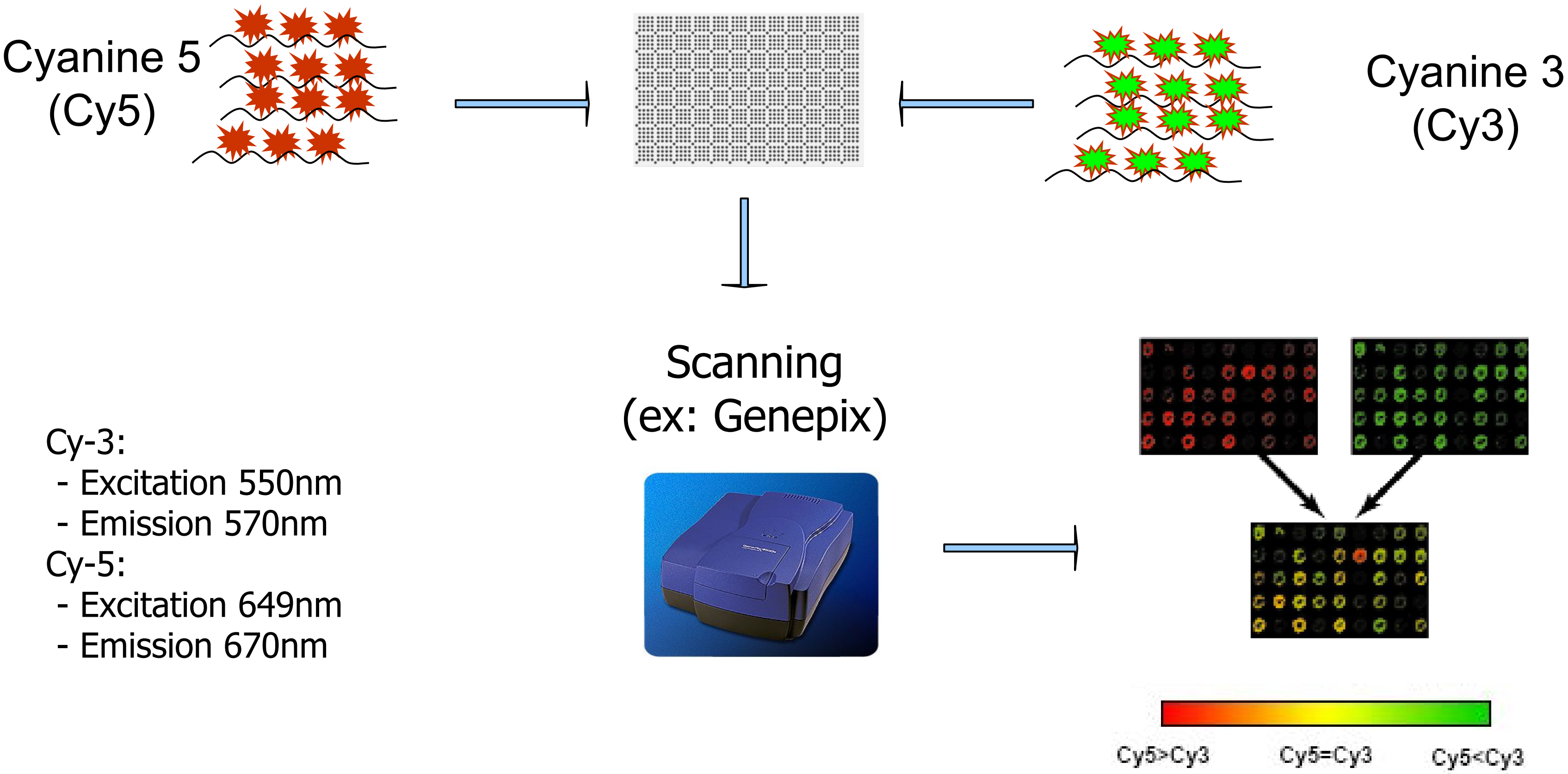
Some players of the RNA world

- 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...

Transcriptome: the old school



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

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

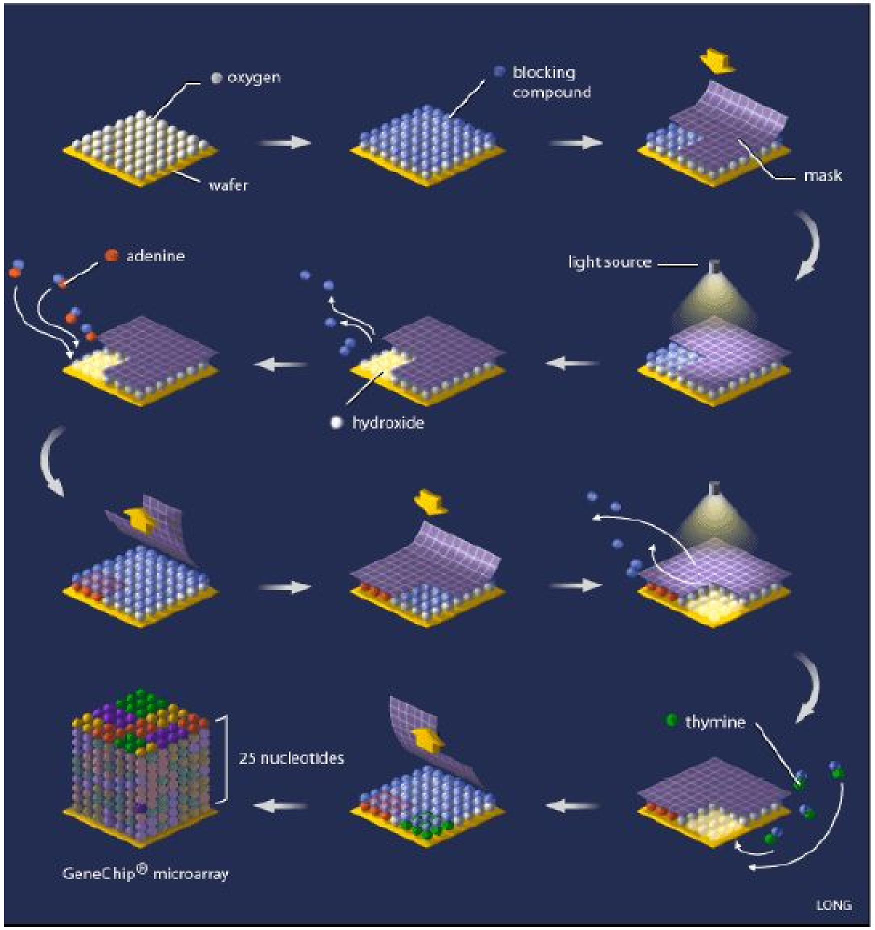
Science. 1995 Oct 20;270(5235):467-70.

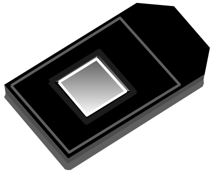
Quantitative monitoring of gene expression patterns with a complementary DNA microarray.

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

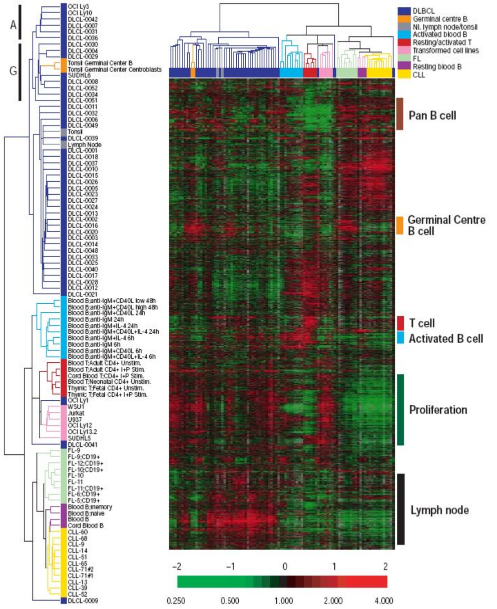
Transcriptome still the old school

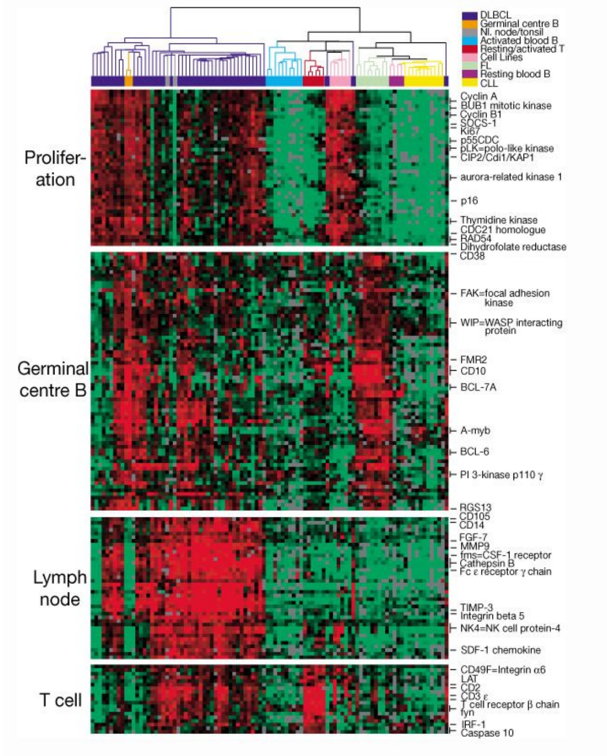
- Principle:
  - In situ synthesis of oligonucleotides
  - Features
    - Cells: 24µm x 24µm
    - ~10^7 oligos per cell
    - ~ 4.10^5-$1,5.10^6$ probes

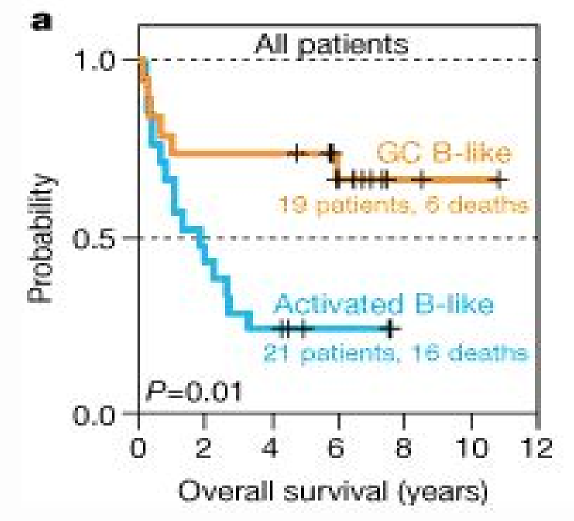




Some pioneering works: “Molecular portraits of tumors”





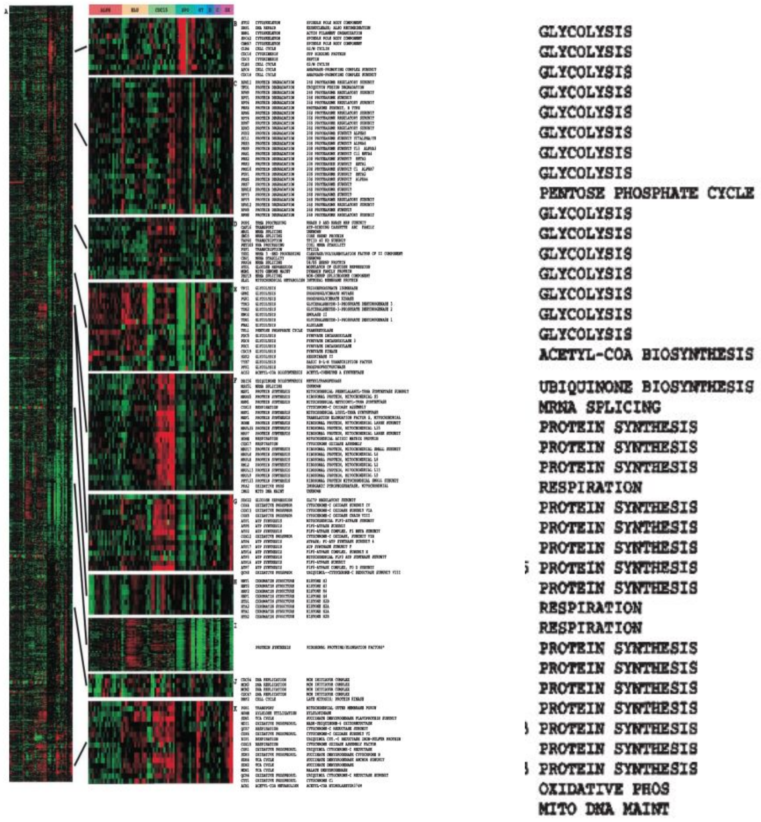


Nature. 2000 Feb 3;403(6769):503-11.

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: 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.

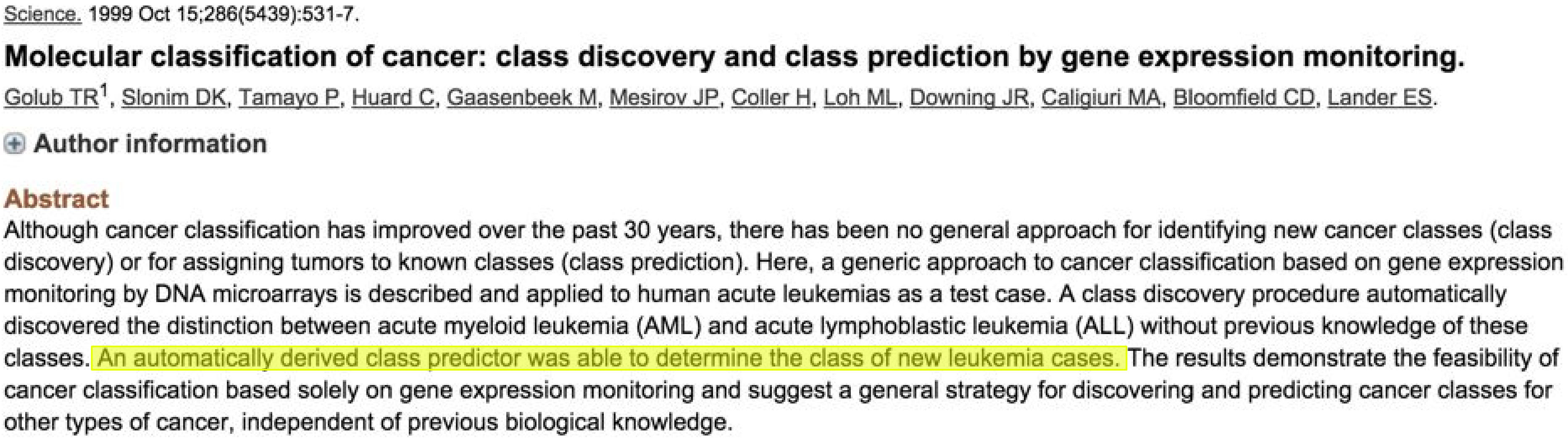
Although various clustering methods can usefully organize tables of gene expression measurements, the resulting ordered but still massive collection of numbers remains difficult to assimilate. Therefore, we always combine clustering methods with a graphical representation of the primary data by representing 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 work: tumor class prediction



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

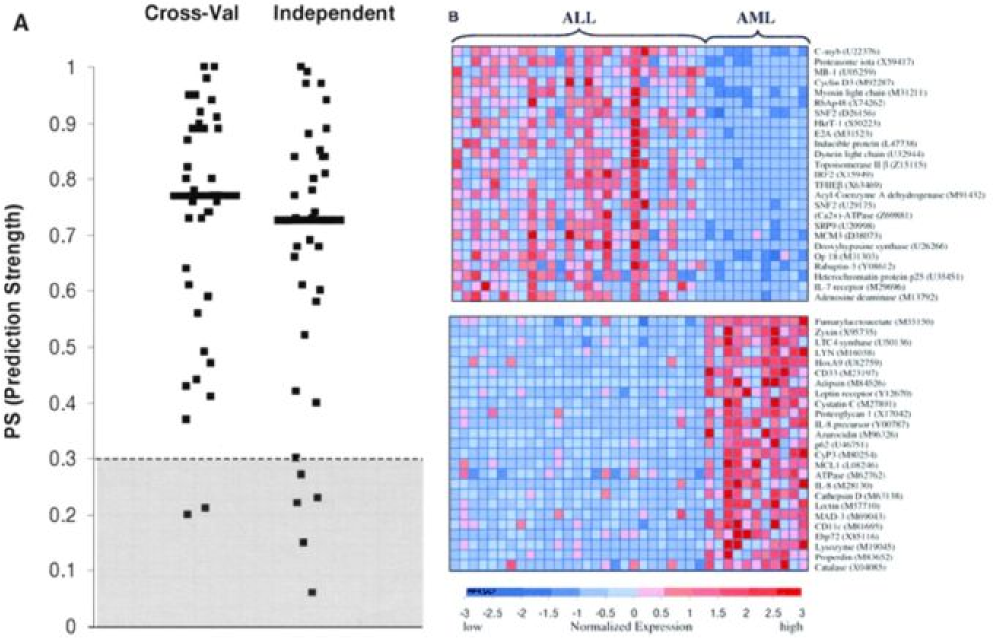
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.

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.

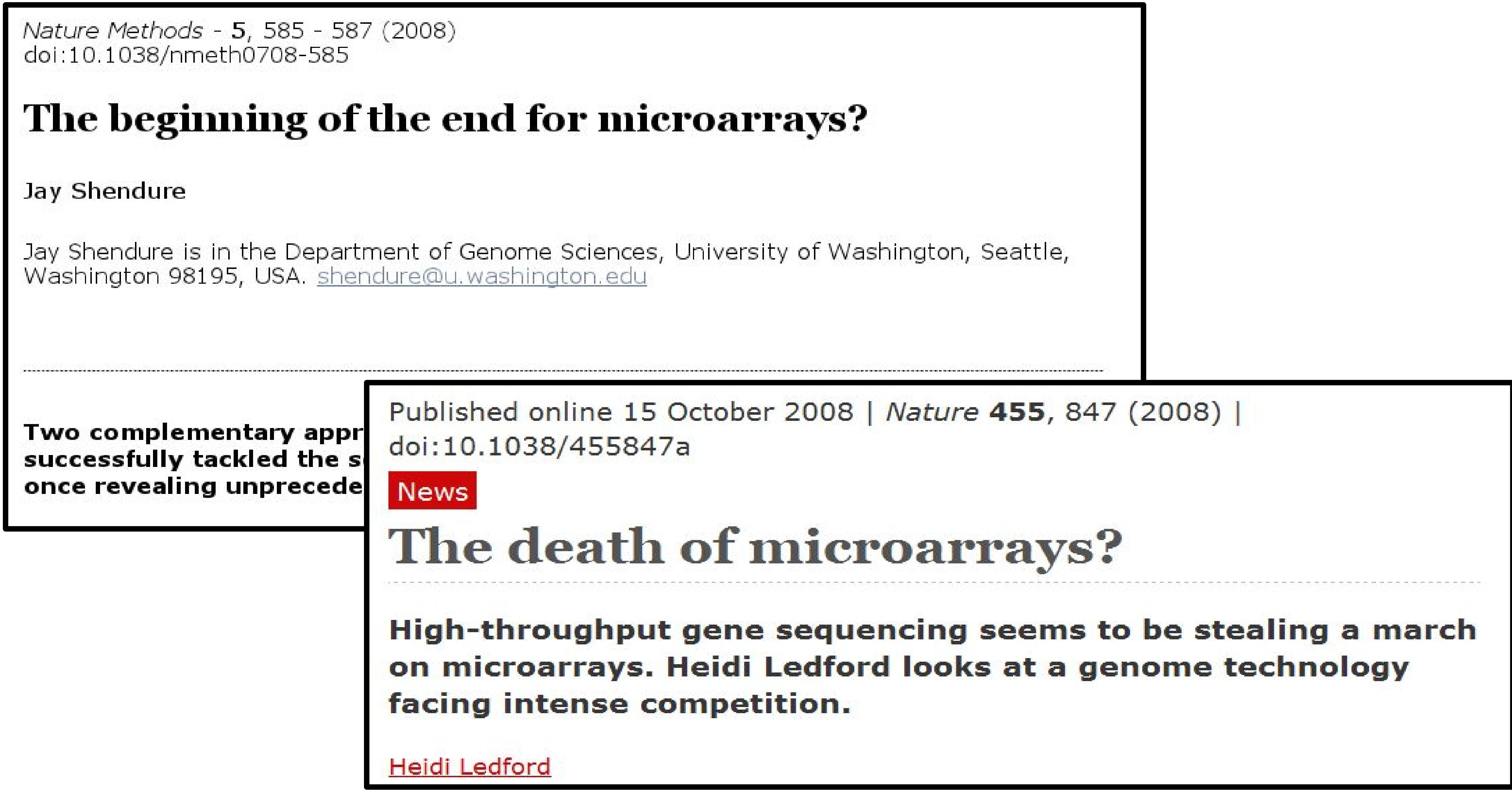
⊕ Author information

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.



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

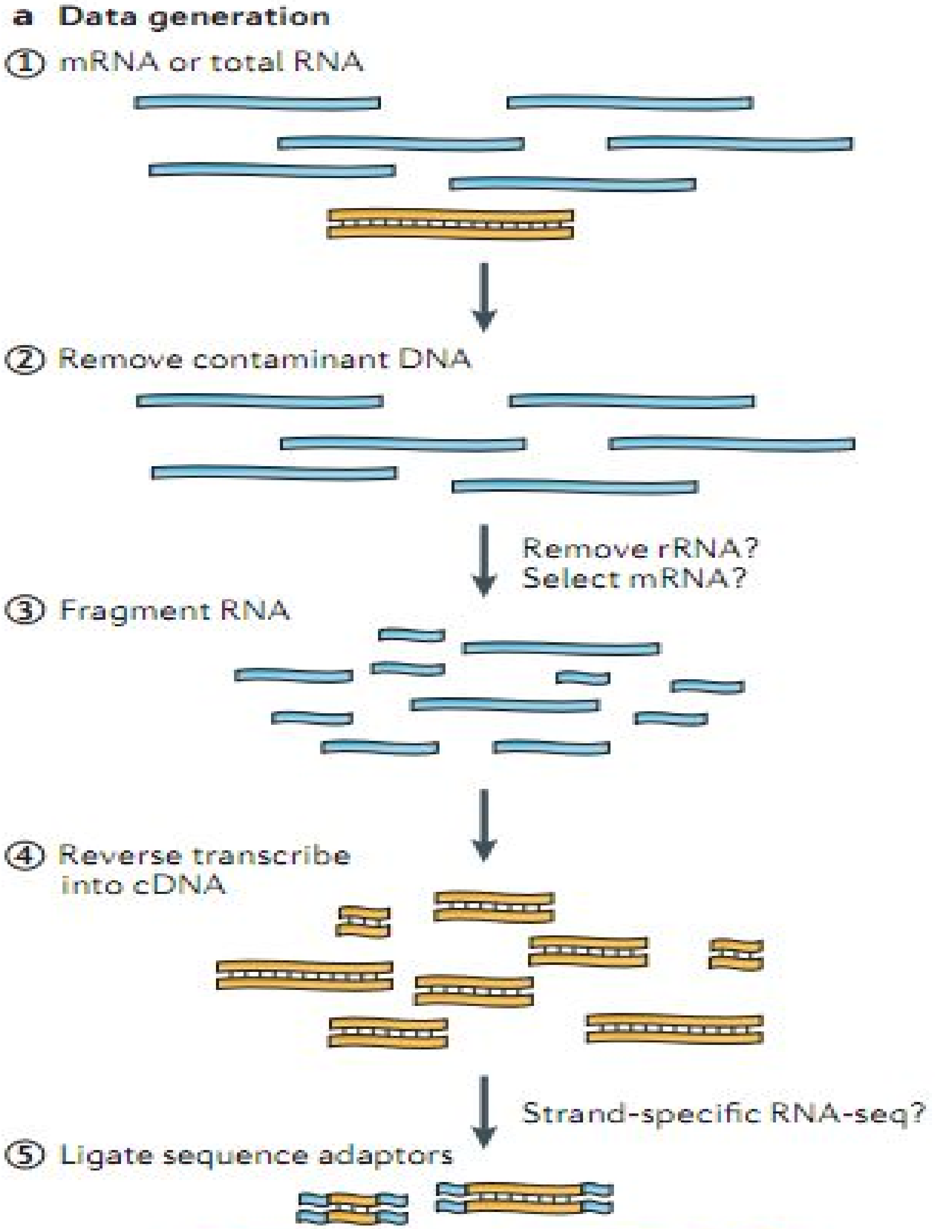
News

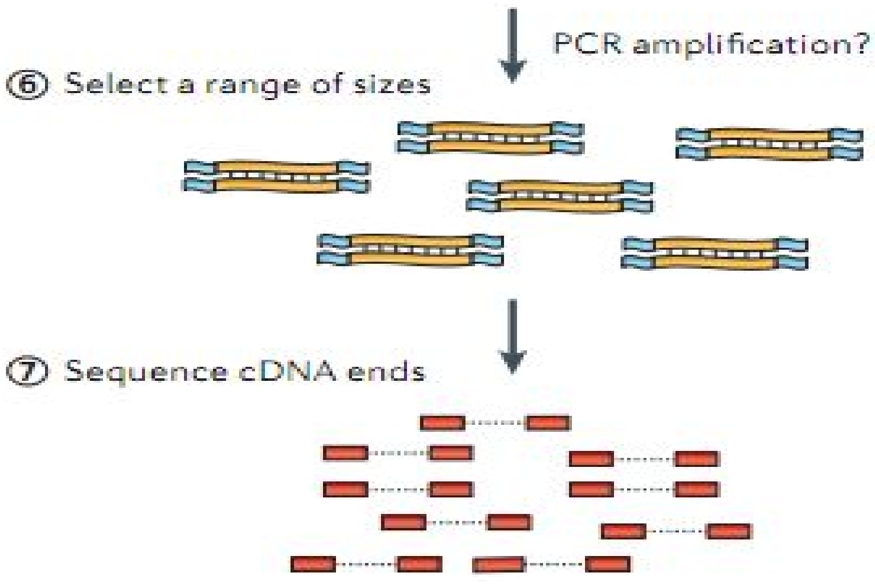
The death of microarrays?

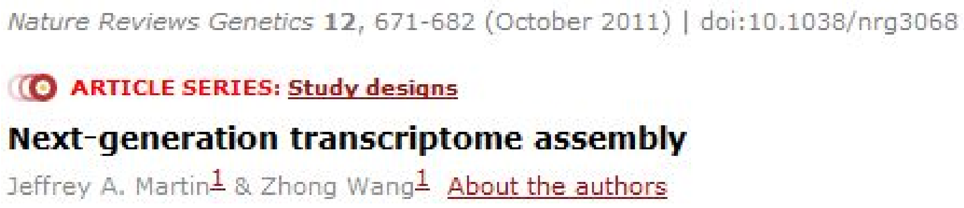
High-throughput gene sequencing seems to be stealing a march on microarrays. Heidi Ledford looks at a genome technology facing intense competition.

Heidi Ledford

RNA-Seq: library construction





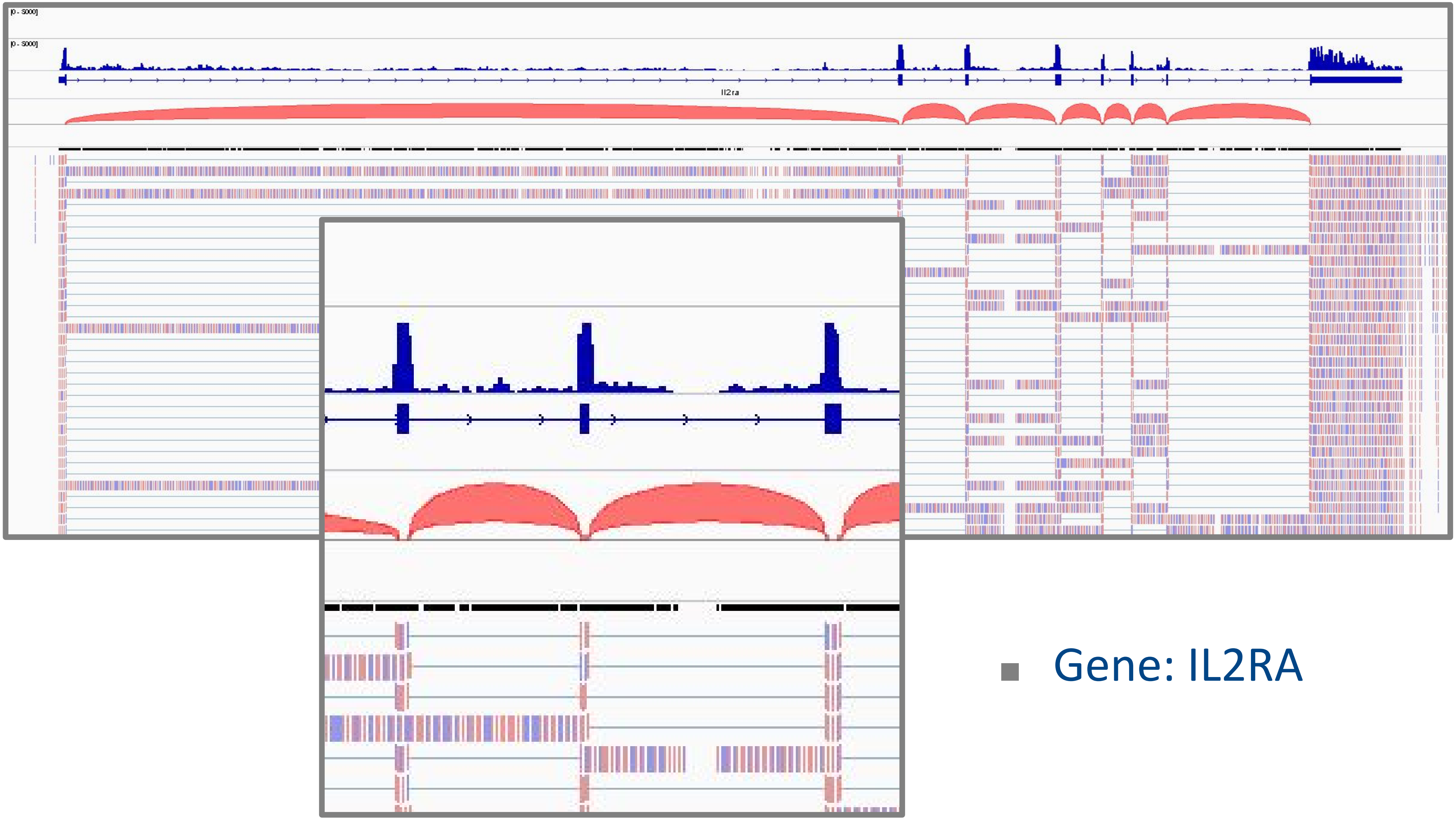

Nature Reviews Genetics **12**, 671-682 (October 2011) | doi:10.1038/nrg3068

ARTICLE SERIES: Study designs

Next-generation transcriptome assembly

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

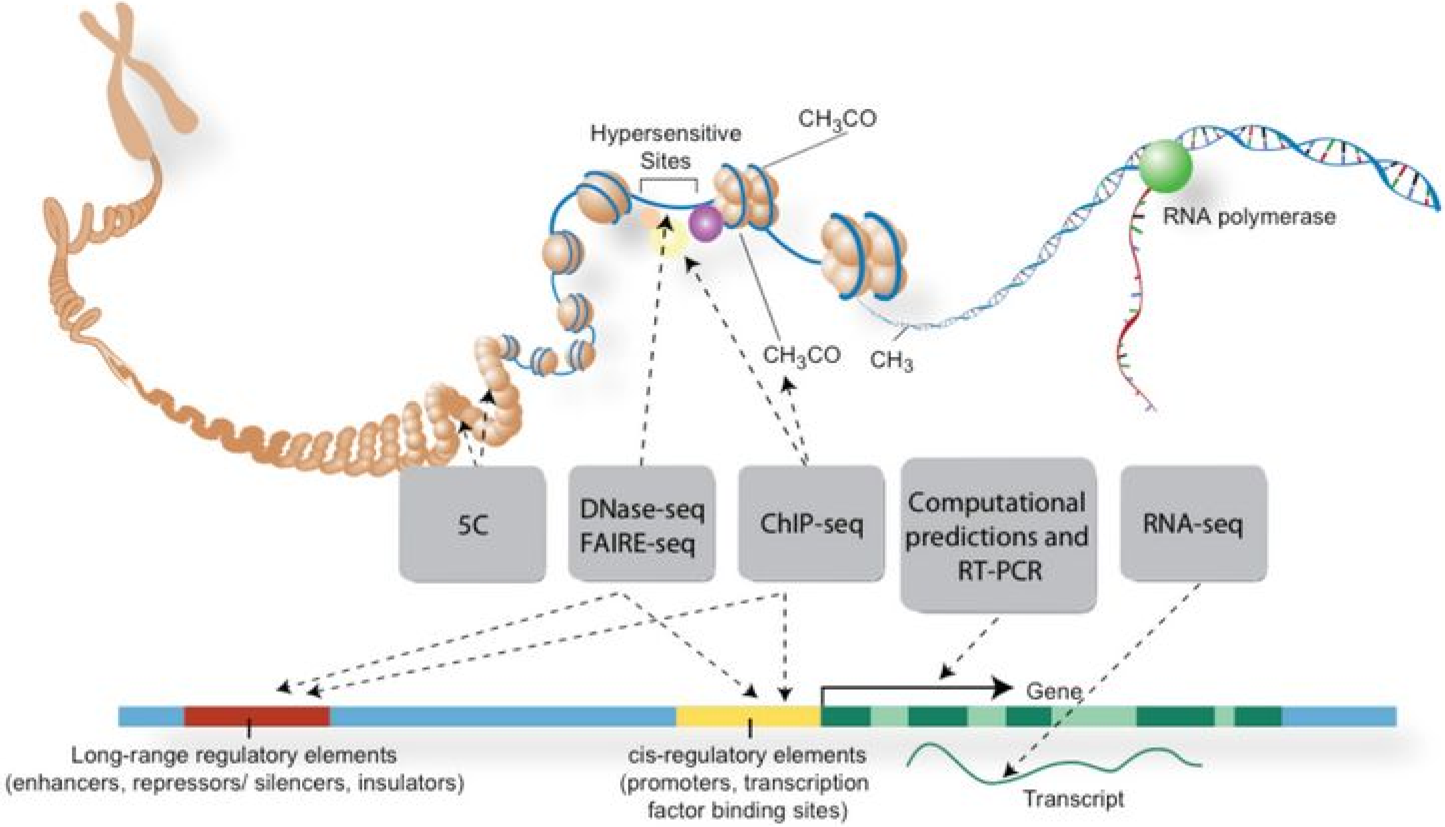
RNA-Seq: aligned reads (Paired-end sequencing on Total RNA)



- Gene: IL2RA

What can we learn from RNA-Seq ?

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


Hypersensitive Sites
CH$_3$CO
RNA polymerase
CH$_3$CO
CH$_3$
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

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

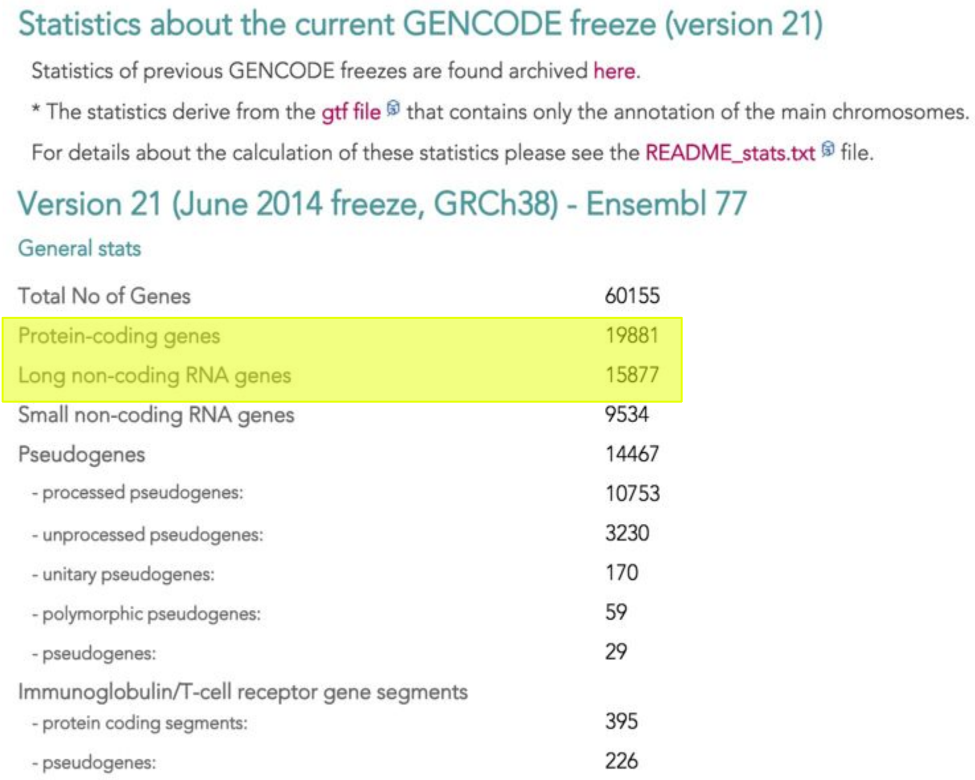
Nature. 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233.

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



Statistics about the current GENCODE freeze (version 21)

Statistics of previous GENCODE freezes are found archived here.

* The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

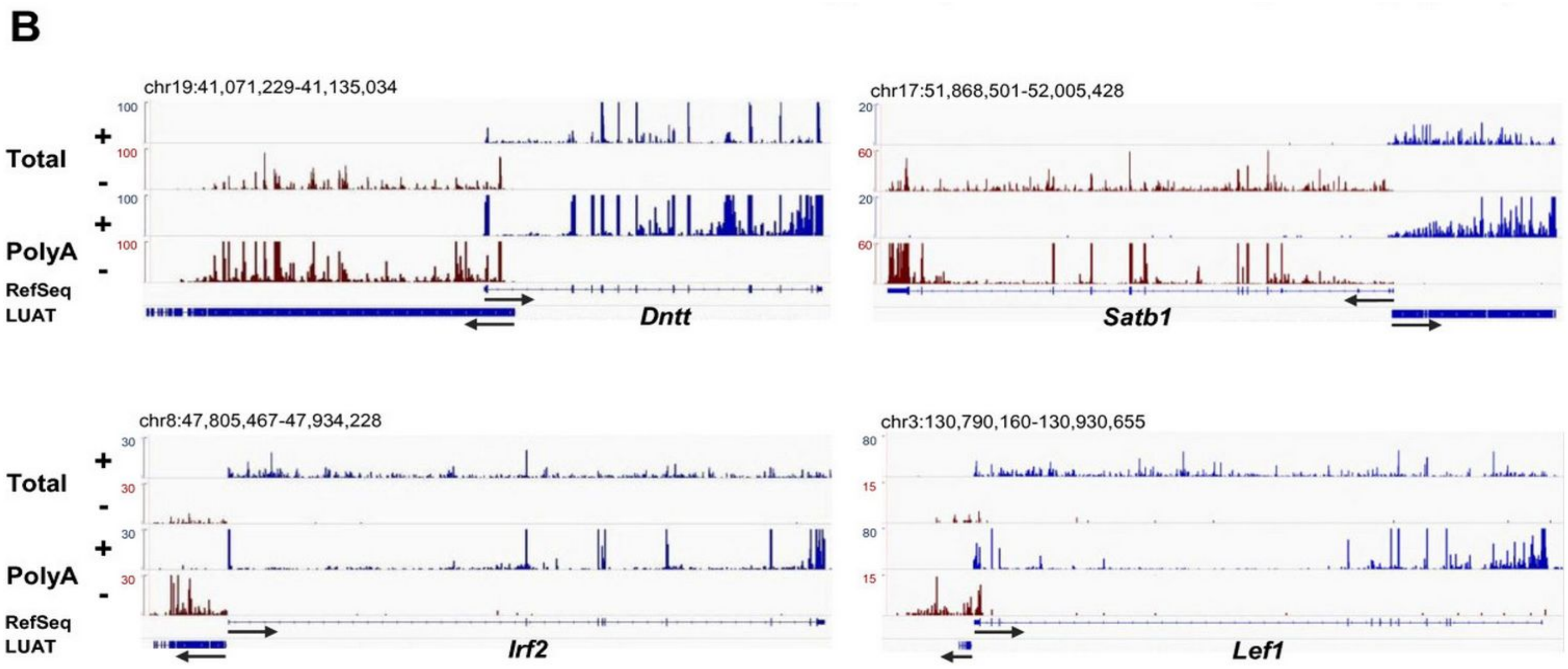
Version 21 (June 2014 freeze, GRCh38) - Ensembl 77

General stats

Total No of Genes	60155
Protein-coding genes	19881
Long non-coding RNA genes	15877
Small non-coding RNA genes	9534
Pseudogenes	14467
- processed pseudogenes:	10753
- unprocessed pseudogenes:	3230
- unitary pseudogenes:	170
- polymorphic pseudogenes:	59
- pseudogenes:	29
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments:	395
- pseudogenes:	226

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



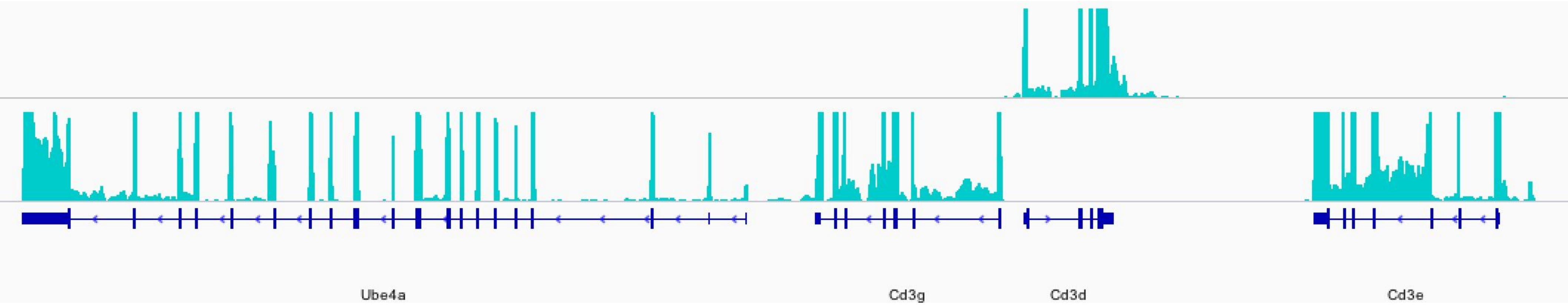
RNA-Seq: protocol variations

- Fragmentation methods
  - RNA: nebulization, magnesium-catalyzed hydrolysis, enzymatic clivage (RNAse III)
  - cDNA: sonication, Dnase I treatment
- Depletion of highly abundant transcripts
  - Ribosomal RNA (rRNA)
    - Positive selection of mRNA . Poly(A) selection.
    - Negative selection. (RiboMinus^{TM})
      - Select also pre-messenger
- Strand specificity
- Single-end or Paired-end sequencing

http://www.bioconductor.org/help/course-materials/2009/EMBLJune09/Talks/RNAseq-Paul.pdf

Strand specific RNA-Seq

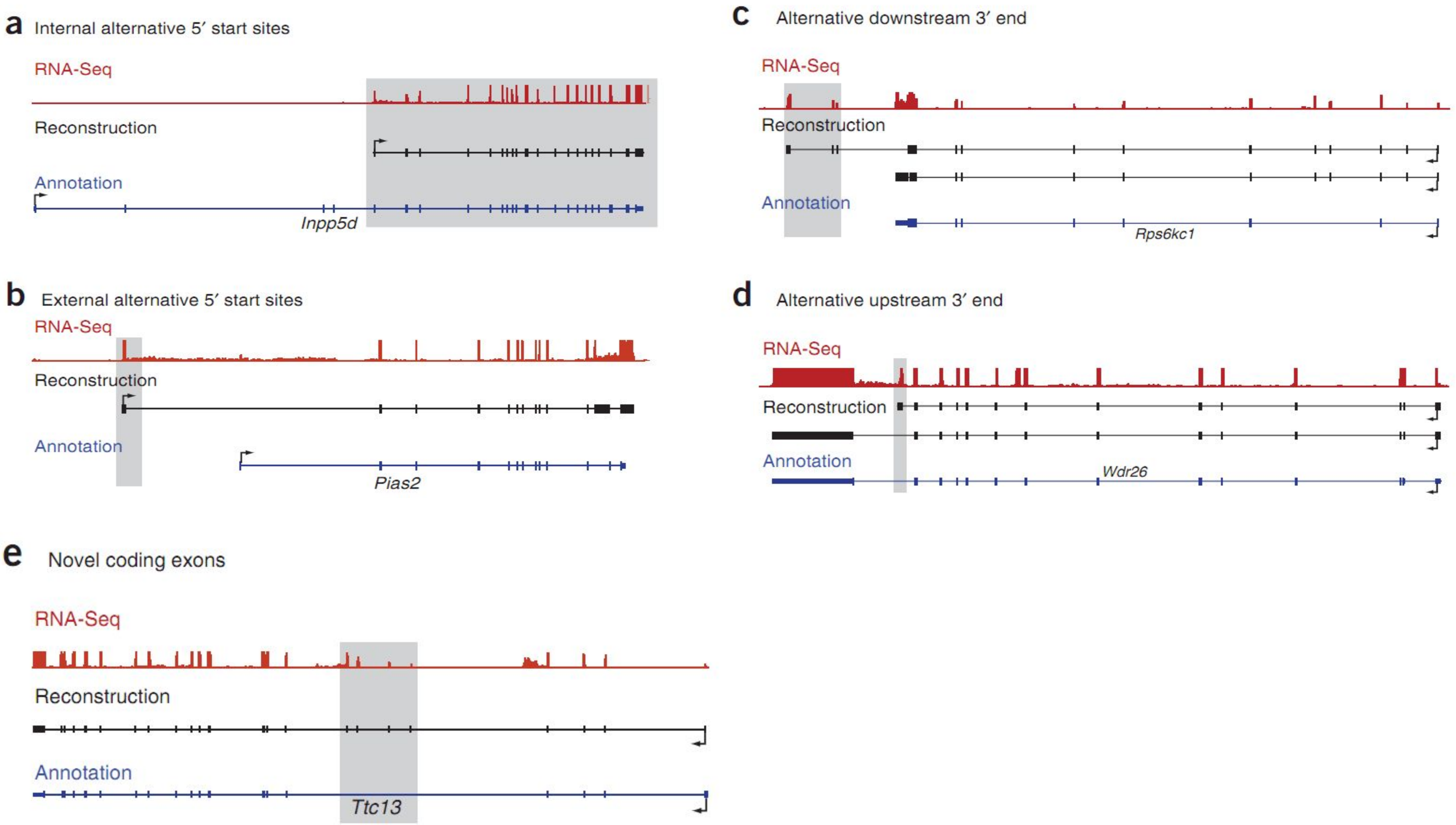
- Most kits are now strand-specific
  - Better estimation of gene expression level.
  - Better reconstruction of transcript model.



Microarrays vs RNA-Seq

- RNA-seq
  - Counting
  - Absolute abundance of transcripts
  - All transcripts are present and can be analyzed
    - mRNA / ncRNA (snoRNA, linc/lncRNA, eRNA, miRNA,...)
  - Several types of analyses
    - Gene discovery
    - Gene structure (new transcript models)
    - Differential expression
    - Allele specific gene expression
    - Detection of fusions and other structural variations

Microarrays vs RNA-Seq



Nat Biotechnol. 2010 May;28(5):503-10. Epub 2010 May 2.

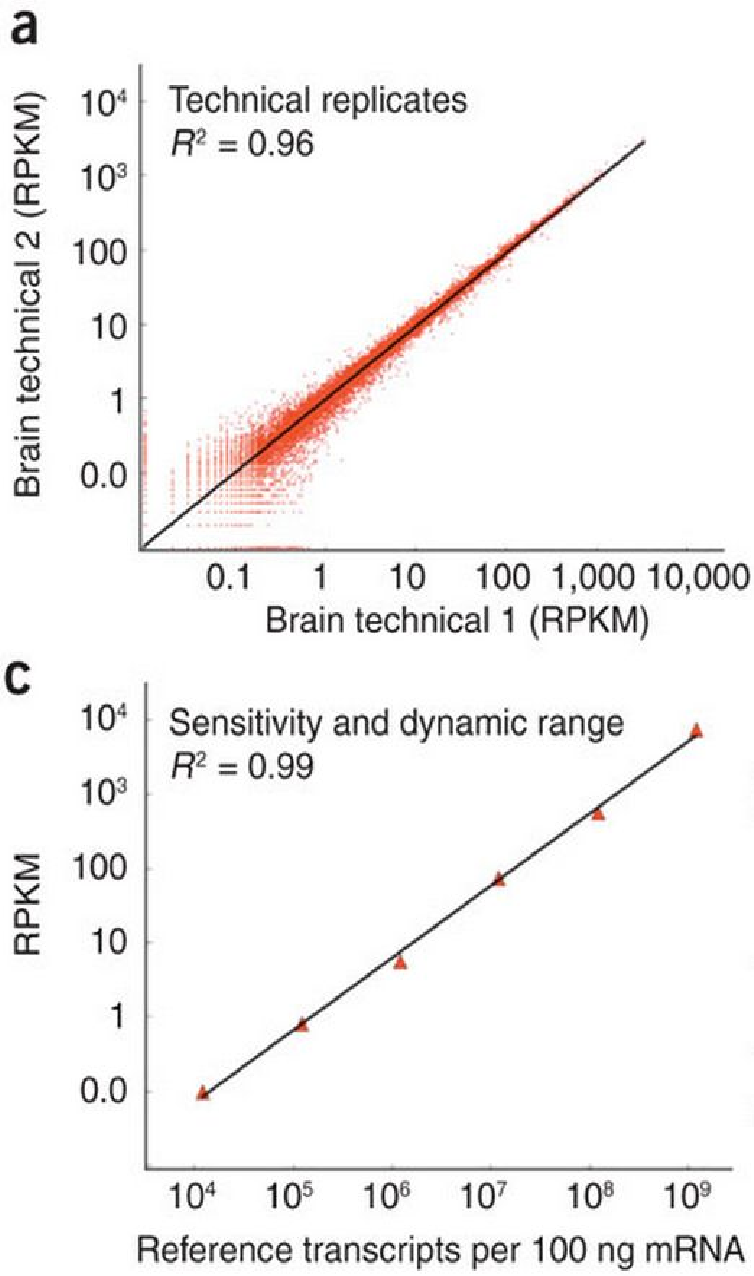
Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A.

Microarrays vs RNA-Seq

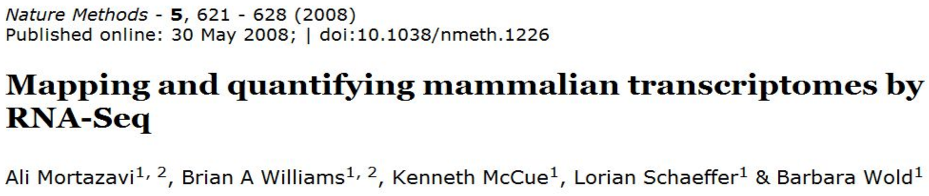
- Microarrays
  - Indirect record of expression level (complementary probes)
  - Relative abundance
  - Cross-hybridization
  - Content limited (can only show you what you're already looking for)

High reproducibility and dynamic range



(a) Comparison of two brain technical replicate RNA-Seq determinations for all mouse gene models (from the UCSC genome database), measured in reads per kilobase of exon per million mapped sequence reads (RPKM), which is a normalized measure of exonic read density; $R^2 = 0.96$.

(c) Six in vitro–synthesized reference transcripts of lengths 0.3–10 kb were added to the liver RNA sample (1.2 104 to 1.2 109 transcripts per sample; R2 > 0.99).

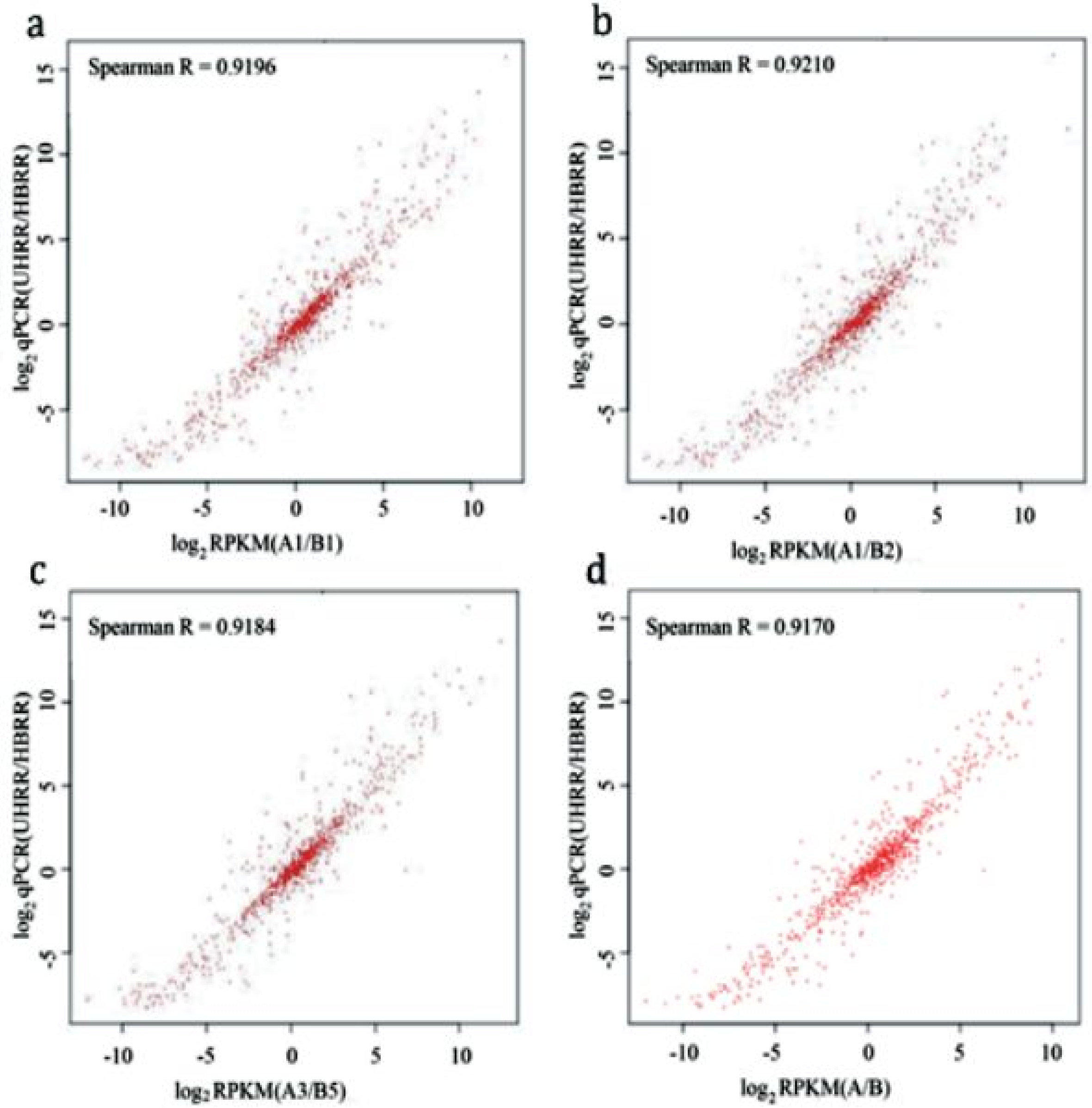


Nature Methods - **5**, 621 - 628 (2008)
Published online: 30 May 2008; | doi:10.1038/nmeth.1226

Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1, 2], Brian A Williams[1, 2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

RNA-seq vs QPCR



http://bgiamericas.com/wp-content/uploads/2011/12/RNA-Aeq-100-ng-20111209.pdf

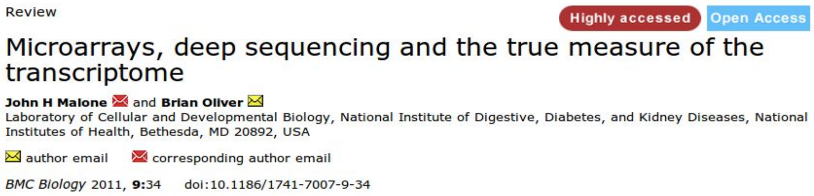
Some RNA-Seq drawbacks

- Current disadvantages
    - More time consuming than any microarray technology
    - Some (lots of) data analysis issues
        - Mapping reads to splice junctions
        - Computing accurate transcript models
        - Contribution of high-abundance RNAs (eg ribosomal) could dilute the remaining transcript population; sequencing depth is important

http://www.bioconductor.org/help/course-materials/2009/EMBLJune09/Talks/RNAseq-Paul.

Do arrays and RNA-Seq tell a consistent story?

- Do arrays and RNA-Seq tell a consistent story?
  - ”The relationship is not quite linear … but the vast majority of the expression values are similar between the methods. Scatter increases at low expression … as background correction methods for arrays are complicated when signal levels approach noise levels. Similarly, RNA-Seq is a sampling method and stochastic events become a source of error in the quantification of rare transcripts ”
  - ”Given the substantial agreement between the two methods, the array data in the literature should be durable”



Review

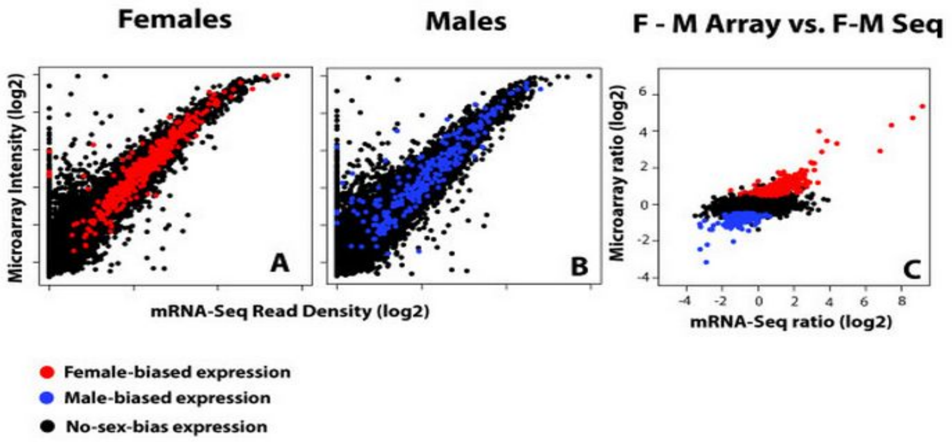
Highly accessed Open Access

Microarrays, deep sequencing and the true measure of the transcriptome

John H Malone ✉ and **Brian Oliver** ✉
Laboratory of Cellular and Developmental Biology, National Institute of Digestive, Diabetes, and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

✉ author email ✉ corresponding author email

BMC Biology 2011, **9**:34 doi:10.1186/1741-7007-9-34



Comparison of array and RNA-Seq data for measuring differential gene expression in the heads of male and female D. pseudoobscura

Raw data: the fastq file format

- Header
- Sequence
- + (optional header)
- Quality (default Sanger-style)

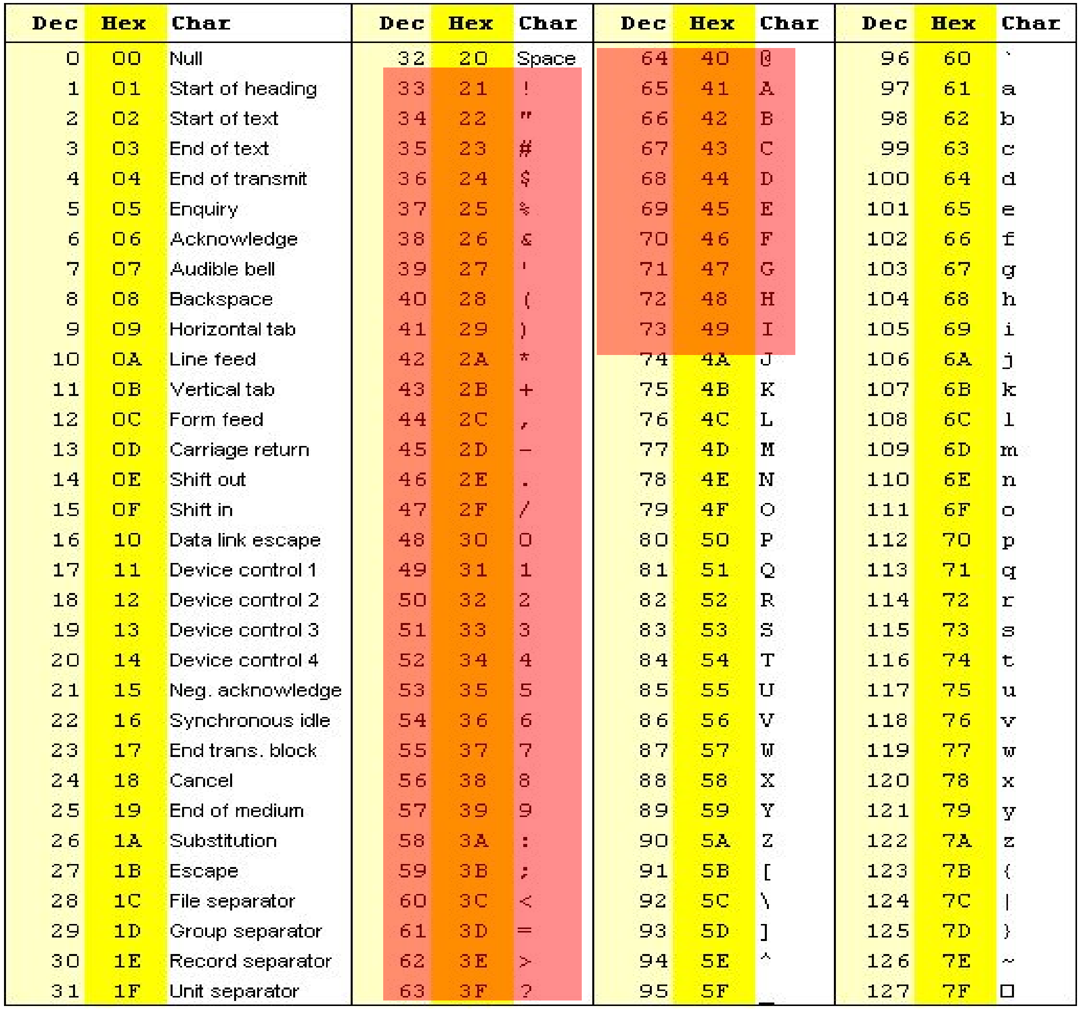
```
@QSEQ32.249996 HWUSI-EAS1691:3:1:17036:13000#0/1 PF=0 length=36
GGGGGTCATCATCATTTGATCTGGGAAAGGCTACTG
+
=.+5:<<<<>AA?0A>;A*A################
@QSEQ32.249997 HWUSI-EAS1691:3:1:17257:12994#0/1 PF=1 length=36
TGTACAACAACAACCTGAATGGCATACTGGTTGCTG
+
DDDD<BDBDB??BB*DD:D#################
```

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

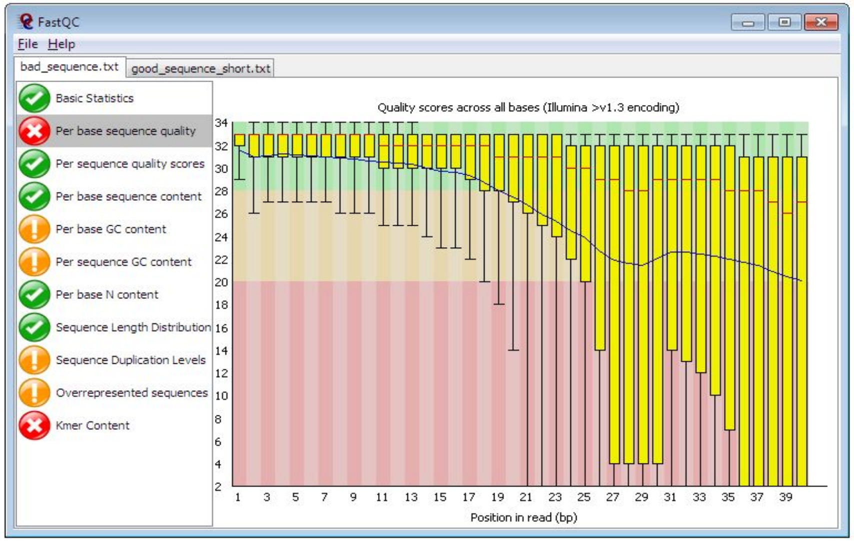
- Storing PHRED scores as single characters gave a simple and space efficient encoding:
- Character ”!” means a quality of 0
- Range 0-40



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	□

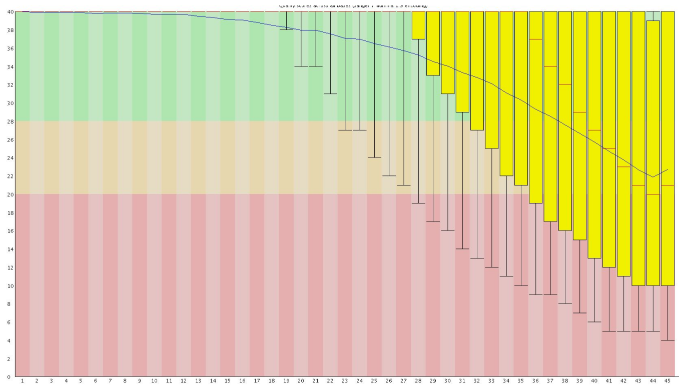
Quality control for high throughput sequence data

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

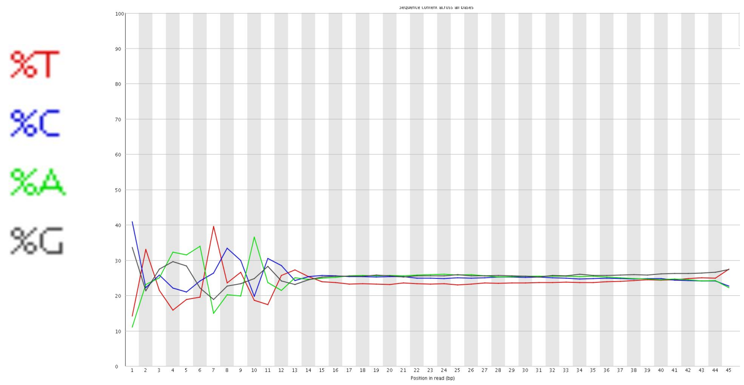

FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

Quality control with FastQC

Quality

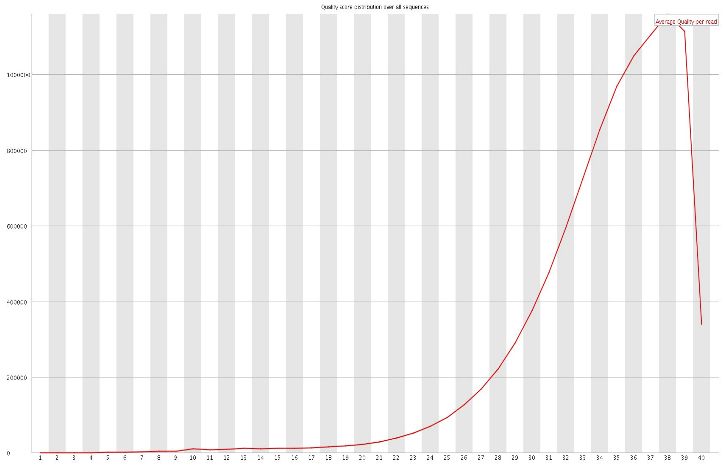


Position in read



Position in read

Nb Reads



Mean Phred Score

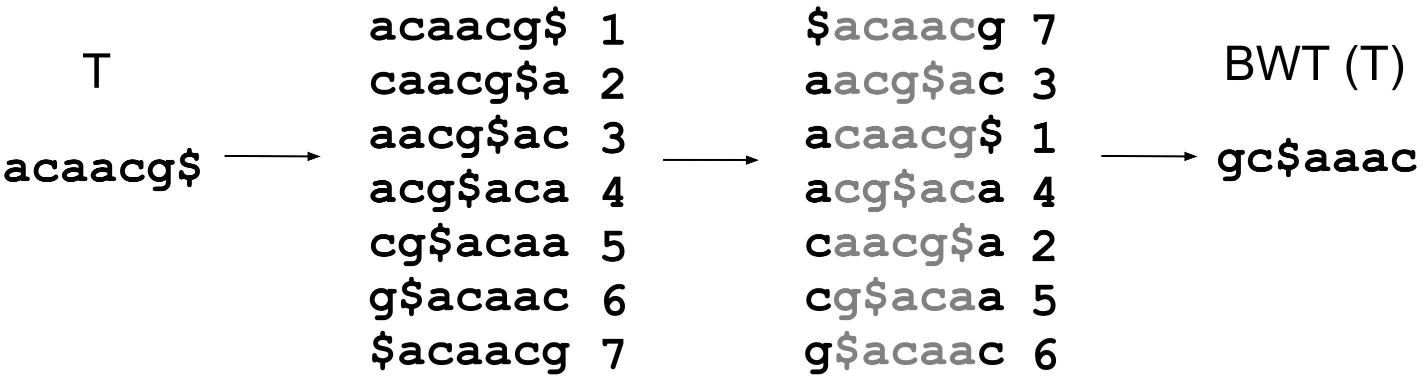
Look also at over-represented sequences

Reference mapping and de novo assembly

- Downstream approaches depend on the availability of a reference genome
  - If reference :
    - Align the read to that reference
      - Rather straightforward
  - If no reference
    - Perform read assembly (contigs) and compare them to known RNA sequences (e.g blast).
      - More complex approaches.

Bowtie a very popular aligner

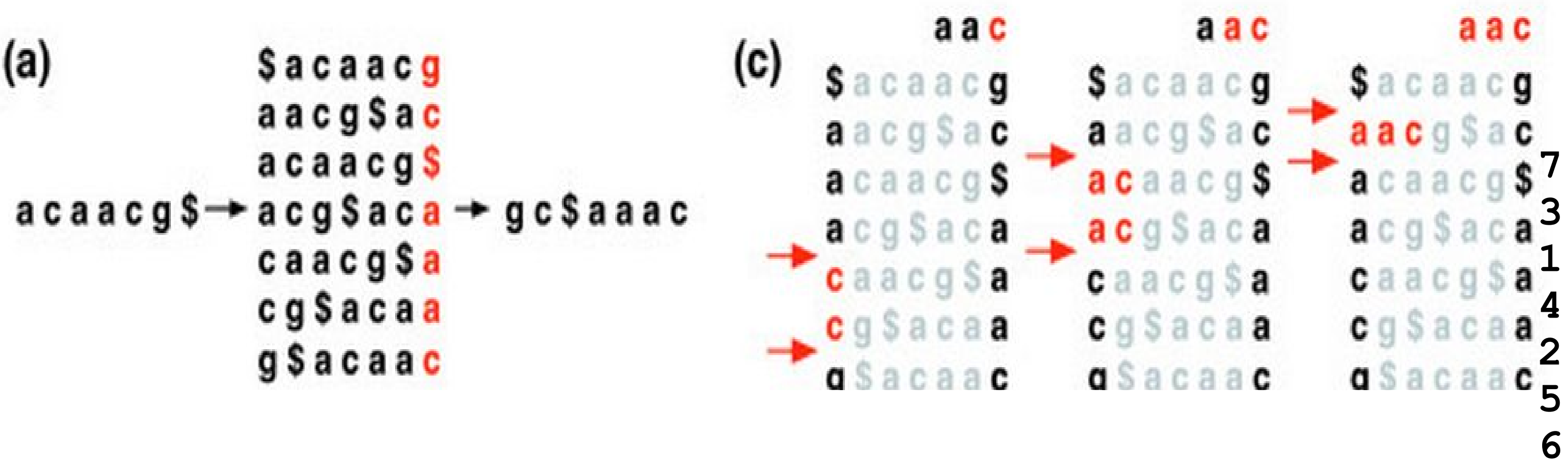
- Burrows Wheeler Transform-based algorithm
- Two phases: “seed and extend”.
- The Burrows-Wheeler Transform of a text T, BWT(T), can be constructed as follows.
  - The character $ is appended to T, where $ is a character not in T that is lexicographically less than all characters in T.
  - The Burrows-Wheeler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



```
              acaacg$ 1        $acaacg 7
   T          caacg$a 2        aacg$ac 3         BWT (T)
              aacg$ac 3        acaacg$ 1
acaacg$  →    acg$aca 4   →    acg$aca 4   →     gc$aaac
              cg$acaa 5        caacg$a 2
              g$acaac 6        cg$acaa 5
              $acaacg 7        g$acaac 6
```

Bowtie principle

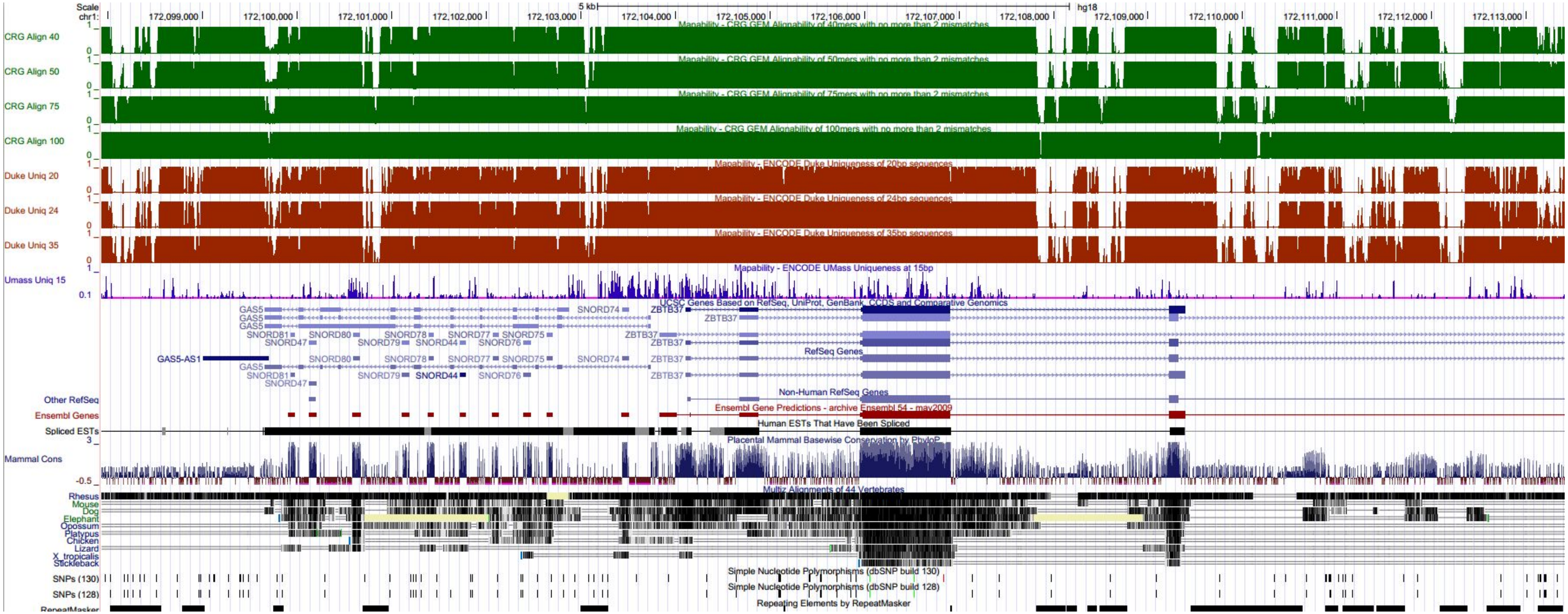
- Burrows-Wheeler Matrices have a property called the Last First (LF) Mapping.
  - The ith occurrence of character c in the last column corresponds to the same text character as the ith occurrence of c in the first column
  - Example: searching ”AAC” in ACAACG



- Second phase is “extension”

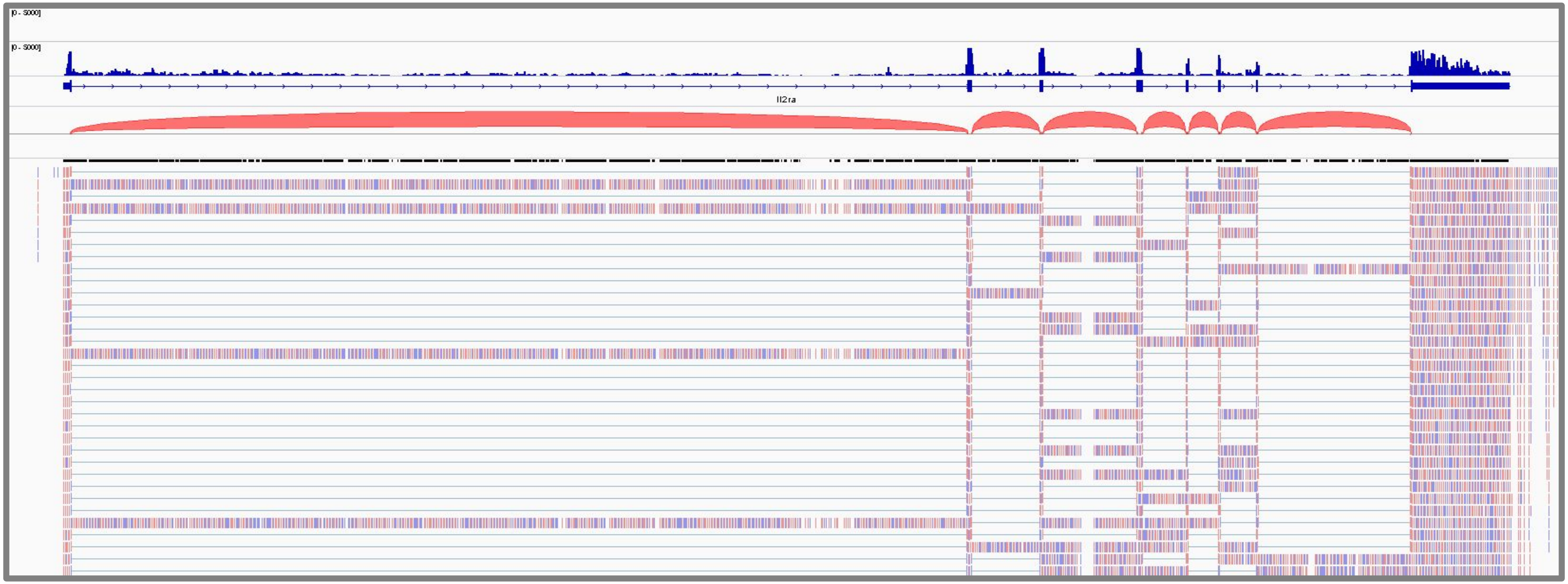
Mappability issues

- Mappability: sequence uniqueness of the reference
- These tracks display the level of sequence uniqueness of the reference NCBI36/hg18 genome assembly. They were generated using different window sizes, and high signal will be found in areas where the sequence is unique.

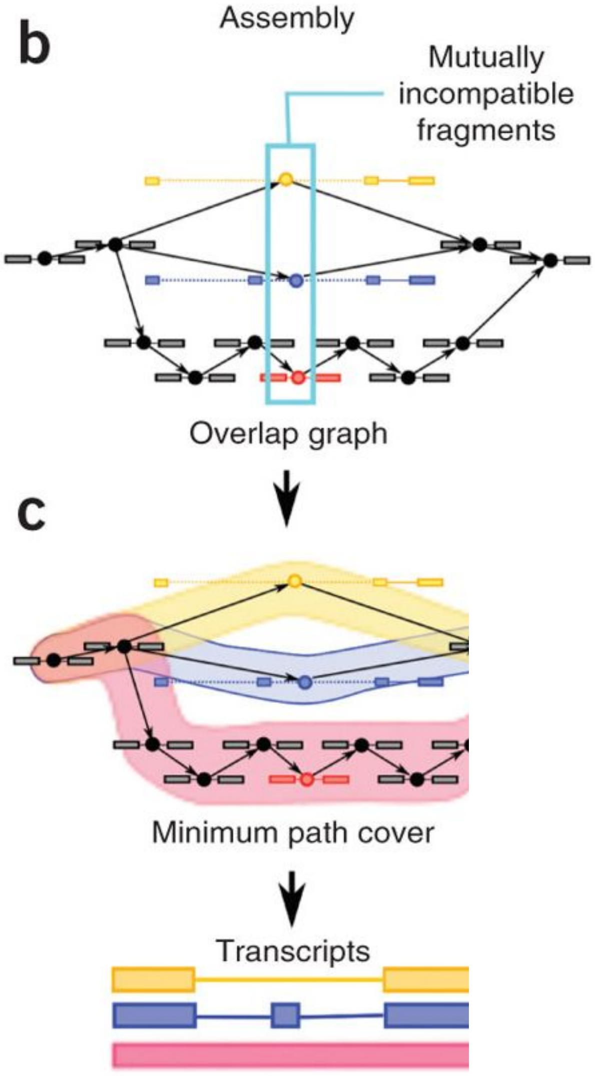

Scale
chr1:
172,099,000
172,100,000
172,101,000
172,102,000
172,103,000
5 kb
hg18
172,104,000
172,105,000
172,106,000
172,107,000
172,108,000
172,109,000
172,110,000
172,111,000
172,112,000
172,113,000
CRG Align 40
Mapability - CRG GEM Alignability of 40mers with no more than 2 mismatches
CRG Align 50
Mapability - CRG GEM Alignability of 50mers with no more than 2 mismatches
CRG Align 75
Mapability - CRG GEM Alignability of 75mers with no more than 2 mismatches
CRG Align 100
Mapability - CRG GEM Alignability of 100mers with no more than 2 mismatches
Duke Uniq 20
Mapability - ENCODE Duke Uniqueness of 20bp sequences
Duke Uniq 24
Mapability - ENCODE Duke Uniqueness of 24bp sequences
Duke Uniq 35
Mapability - ENCODE Duke Uniqueness of 35bp sequences
Umass Uniq 15
Mapability - ENCODE UMass Uniqueness at 15bp
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
GAS5
SNORD81
SNORD80
SNORD47
SNORD79
SNORD78
SNORD44
SNORD77
SNORD76
SNORD75
SNORD74
ZBTB37
GAS5-AS1
RefSeq Genes
Other RefSeq
Non-Human RefSeq Genes
Ensembl Genes
Ensembl Gene Predictions - archive Ensembl 54 - may2009
Spliced ESTs
Human ESTs That Have Been Spliced
Mammal Cons
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 44 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
SNPs (130)
Simple Nucleotide Polymorphisms (dbSNP build 130)
SNPs (128)
Simple Nucleotide Polymorphisms (dbSNP build 128)
RepeatMasker
Repeating Elements by RepeatMasker

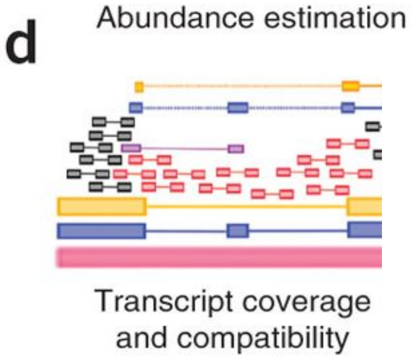
Mapping read spanning exons

- One limit of bowtie
  - mapping reads spanning exons
- Solution: splice-aware short-read aligners
  - E.g: tophat



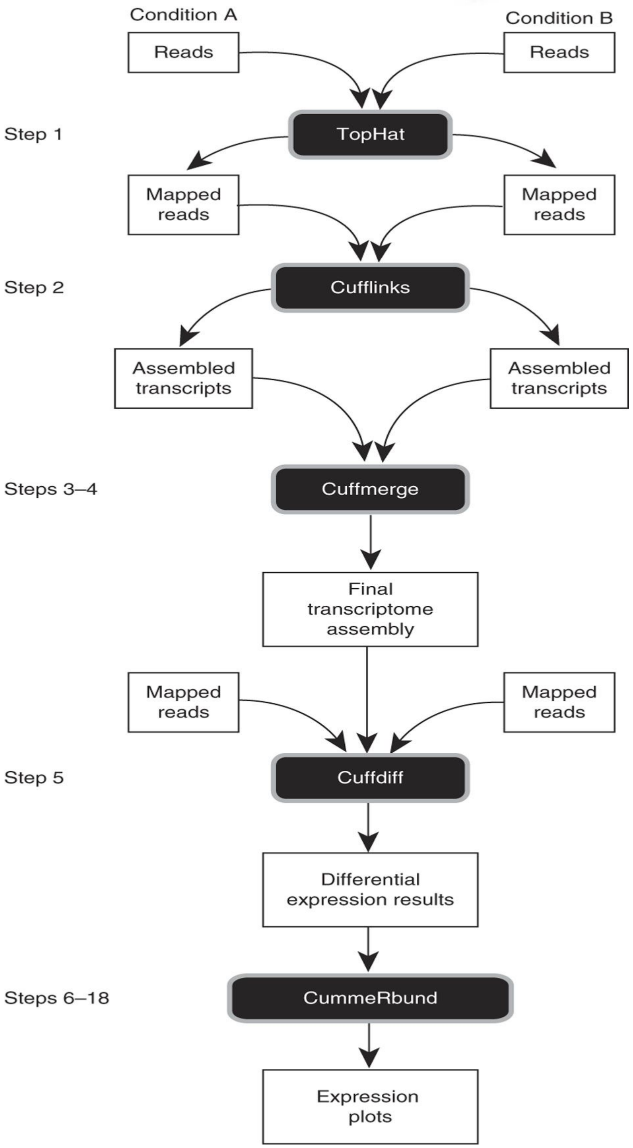
Searching for novel transcript model: cufflinks



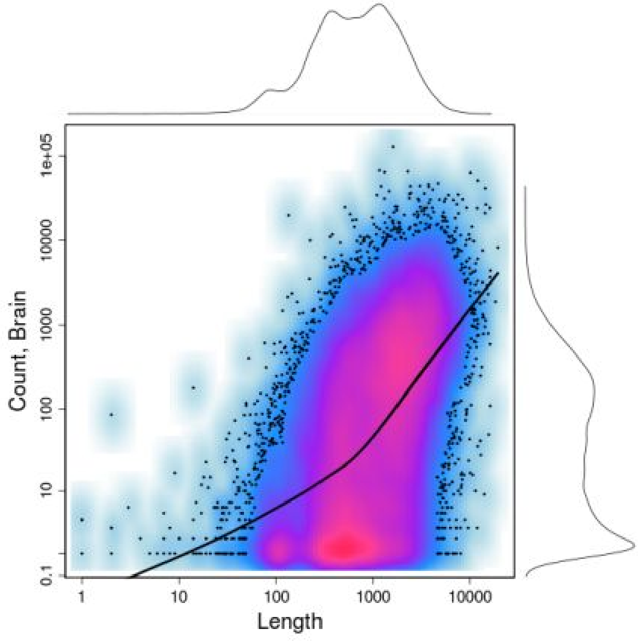


Read pair

Gapped alignment



Quantification

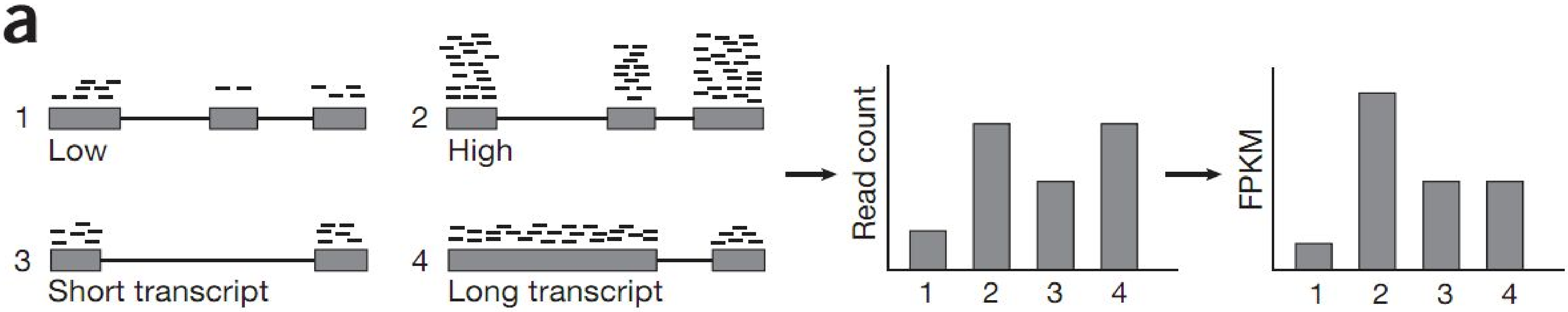


(a) Count vs. length

- Objective
  - Count the number of reads that fall in each gene
    - HTSeq-count, featureCounts,...
- Known issue
  - Positive association between gene counts and length
    - suggests higher expression among longer genes

RPKM / FPKM

- Transcrits of different length have different read count



- Tag count is normalized for transcrit length and total read number in the measurement (RPKM, Reads Per Kilobase of exon model per Million mapped reads)
- 1 RPKM corresponds to approximately one transcript per cell
- FPKM, Fragments Per Kilobase of exon model per Million mapped reads (paired-end sequencing)

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,*]

Some proposed normalization methods

- Reads Per Kilobase per Million mapped reads (RPKM): This approach was initially introduced to facilitate comparisons between genes within a sample.
  - Not sufficient

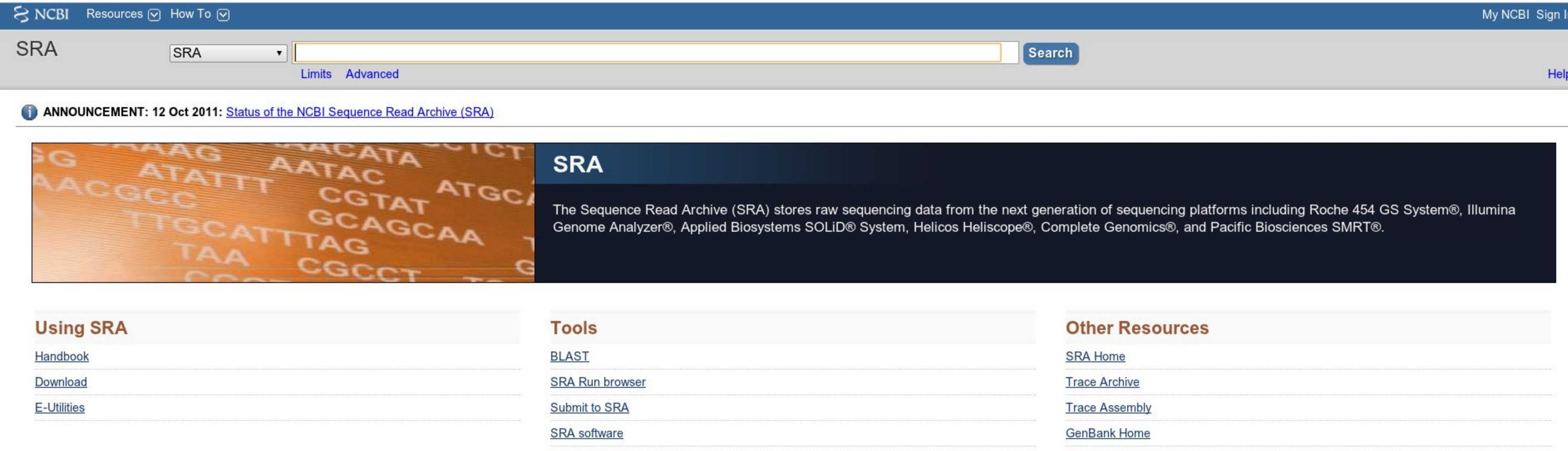
- Upper Quartile (UQ): the total counts are replaced by the upper quartile of counts different from 0 in the computation of the normalization factors.

- Trimmed Mean of M-values (TMM): This normalization method is implemented in the edgeR Bioconductor package (version 2.4.0). Scaling is based on a subset of M values
  - TMM seems to provide a robust scaling factor.

Next step ?

- Compare various samples
    - Eg.
        - control vs treated
        - Normal vs tumor
        - Poor/bad prognosis
        - …
    - Compare expression level, isoforms, fusions,...
- Perform classification
- Compare RNA-Seq data to regulatory data (ChIP-Seq,...)

Sequence read Archive (SRA)

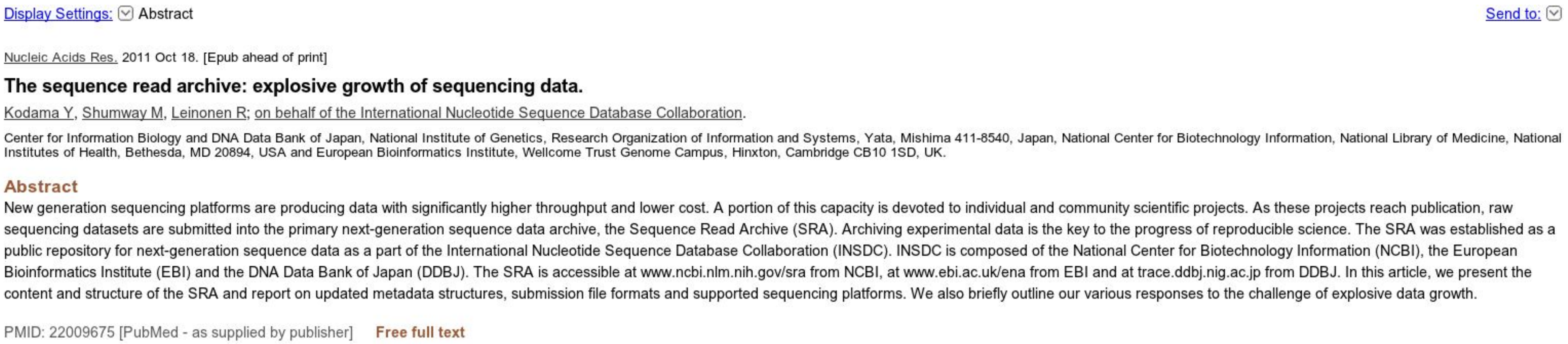


ANNOUNCEMENT: 12 Oct 2011: Status of the NCBI Sequence Read Archive (SRA)

The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

- 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.

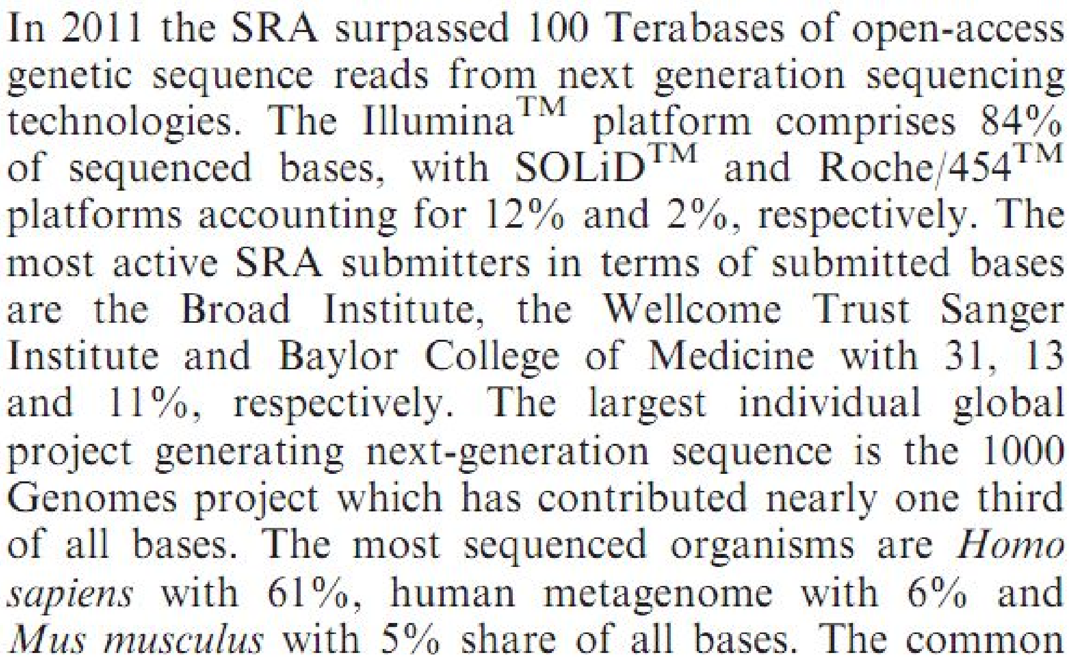
Kodama Y, Shumway M, Leinonen R; on behalf of the International Nucleotide Sequence Database Collaboration.

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.

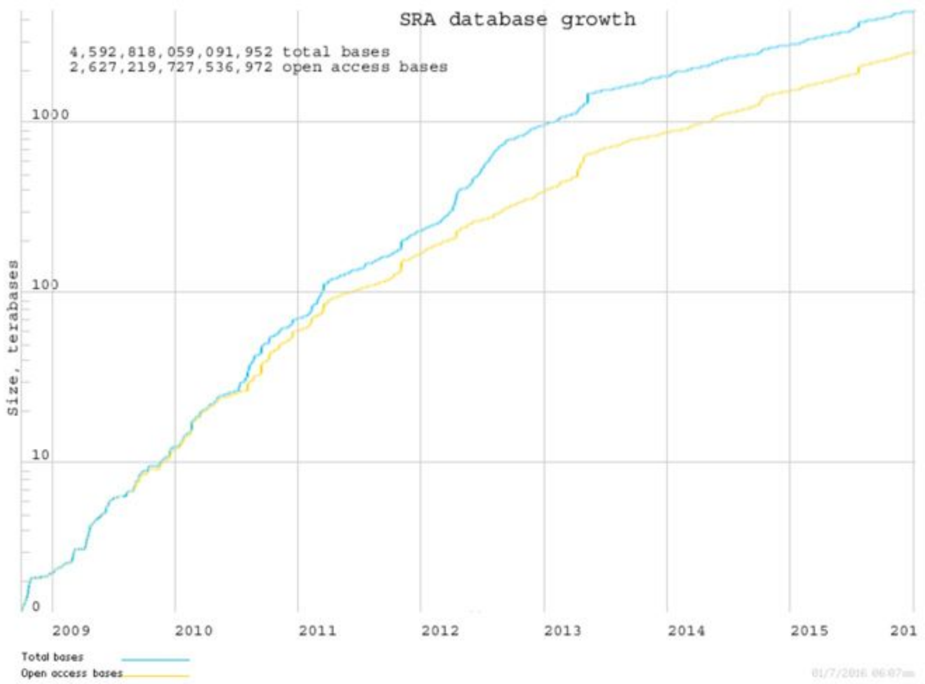
Abstract

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



Merci