High throughput sequencing methods in genomics. Focus on transcriptome analysis

D. Puthier

Tlemcen, may 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

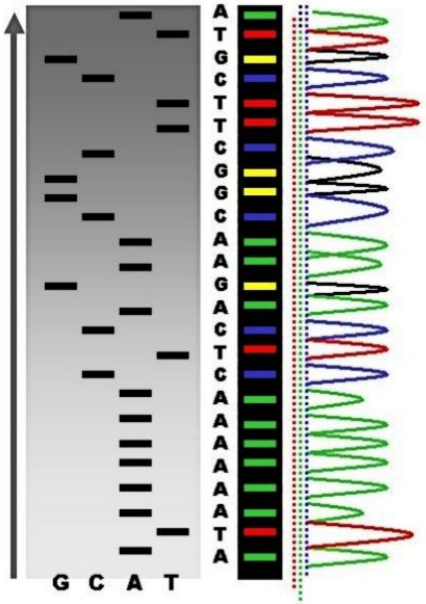
Analysing genomes requires teams/individuals with various skills

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

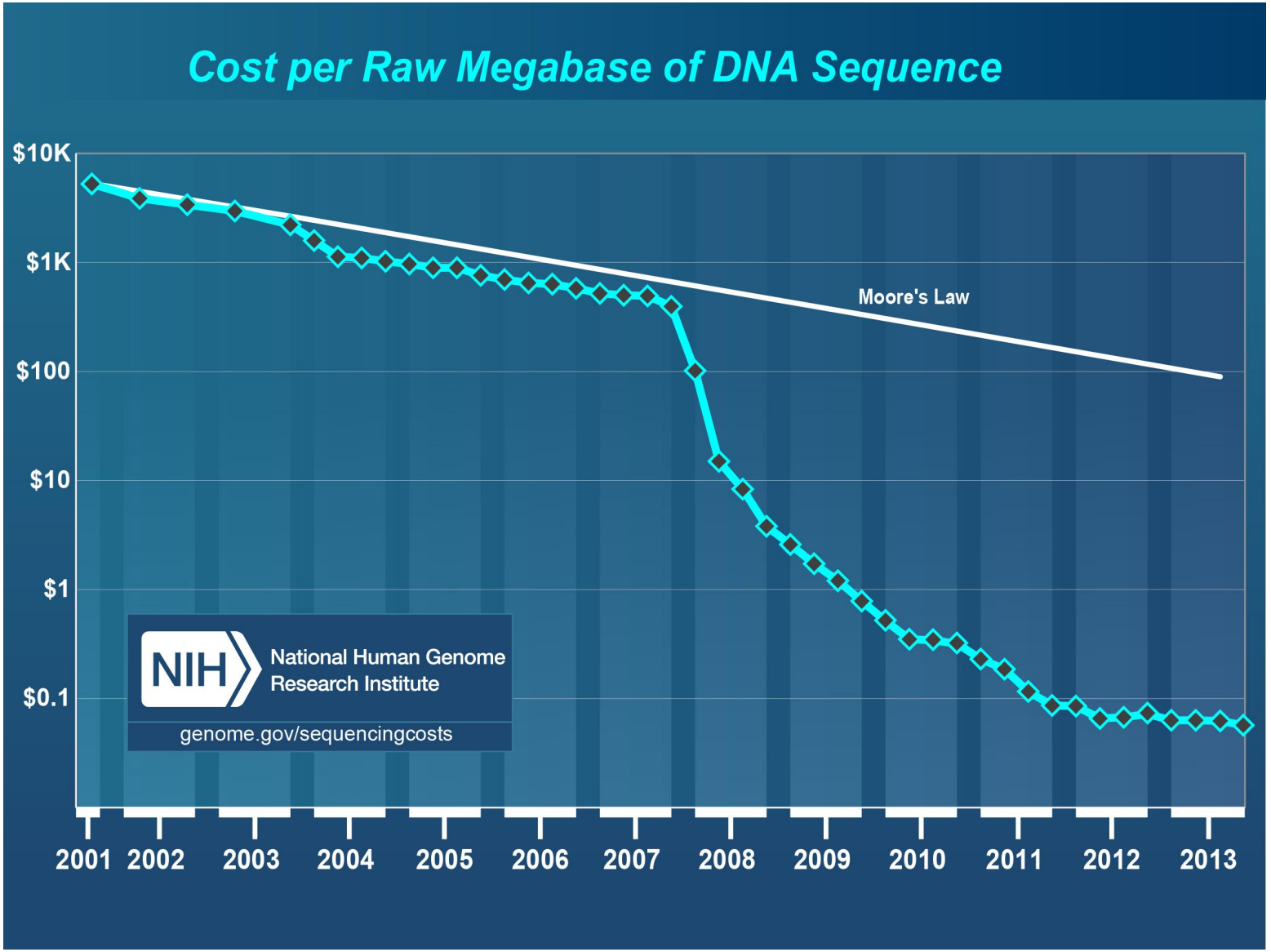
Breakthrough in DNA Sequencing

- 1977-1990, 500bp, manual analysis
- 1990-2000, 500Bp, computed assisted analysis (1D capillary sequencers)
- 2005-2014, 20-1000bp

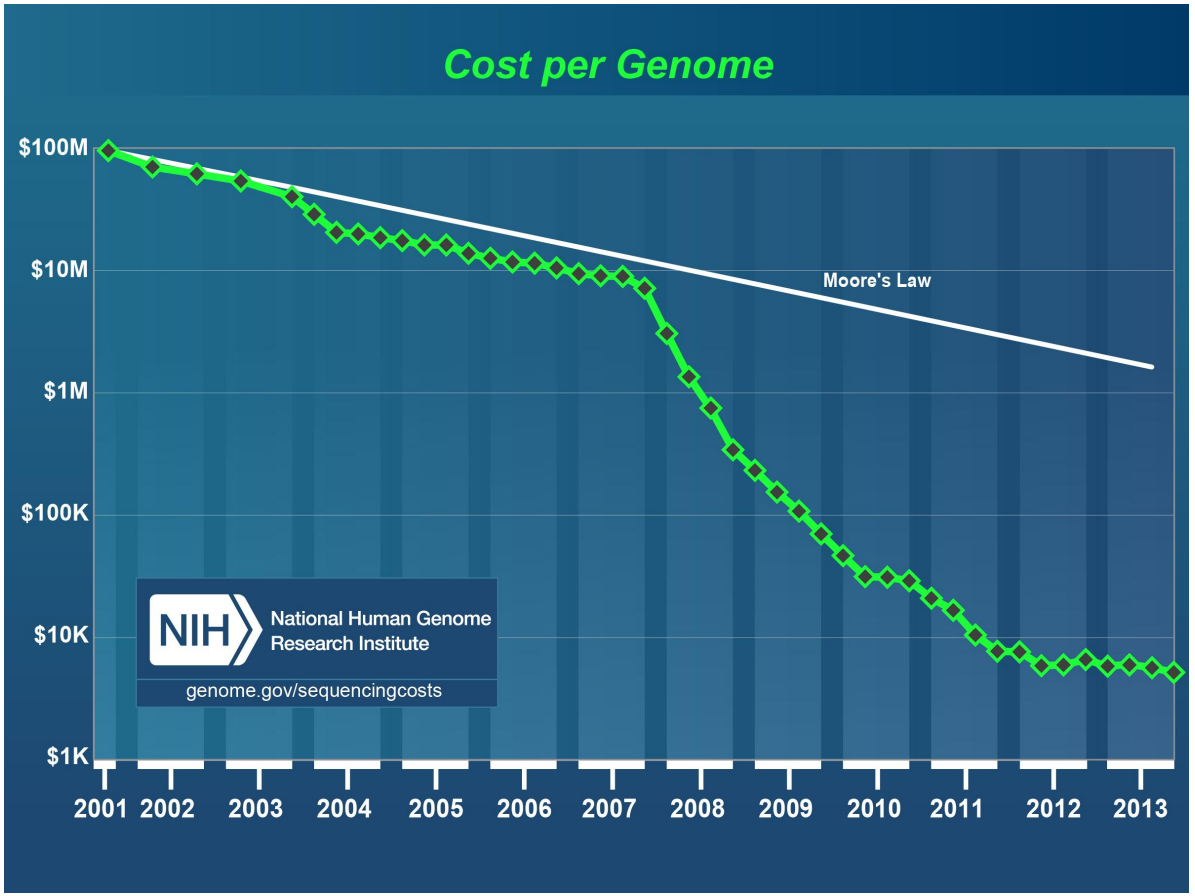
(2D sequencers “Next Generation Sequencing.”)



Cost per megabase (1 million base)



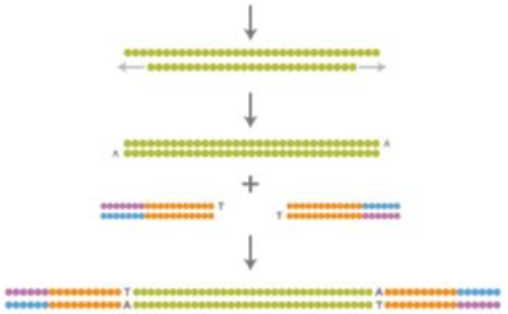
Cost per human genome



- Sanger-based sequencing (average read length=500-600 bases): 6-fold coverage
- 454 sequencing (average read length=300-400 bases): 10-fold coverage
- Illumina and SOLiD sequencing (average read length=50-100 bases): 30-fold coverage

NGS: a simplified view

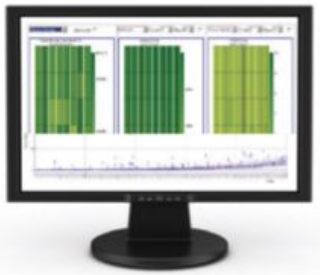
Figure 3: Next-Generation Sequencing Simplified



Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)

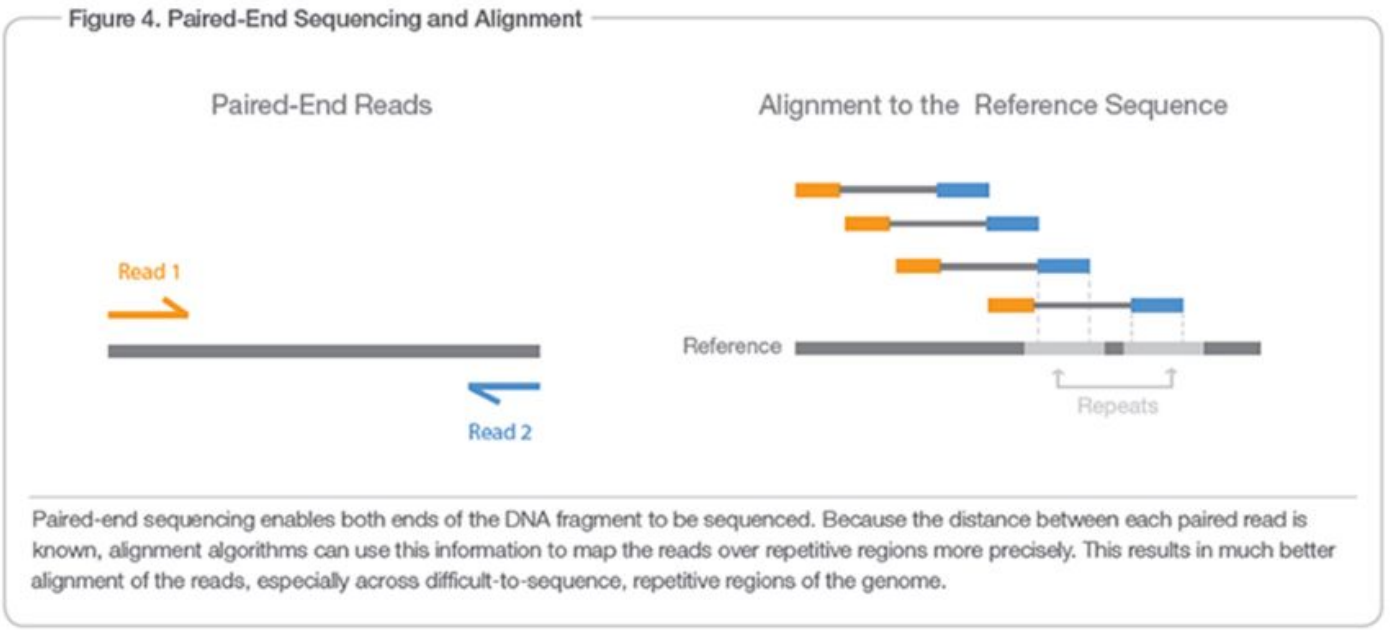
From simplified sample preparation kits and automated cluster generation, to streamlined sequencing by synthesis and complete data analysis, Illumina HiSeq sequencing systems offer the industry's simplest next-generation sequencing workflow.

Illumina sequencing principle

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

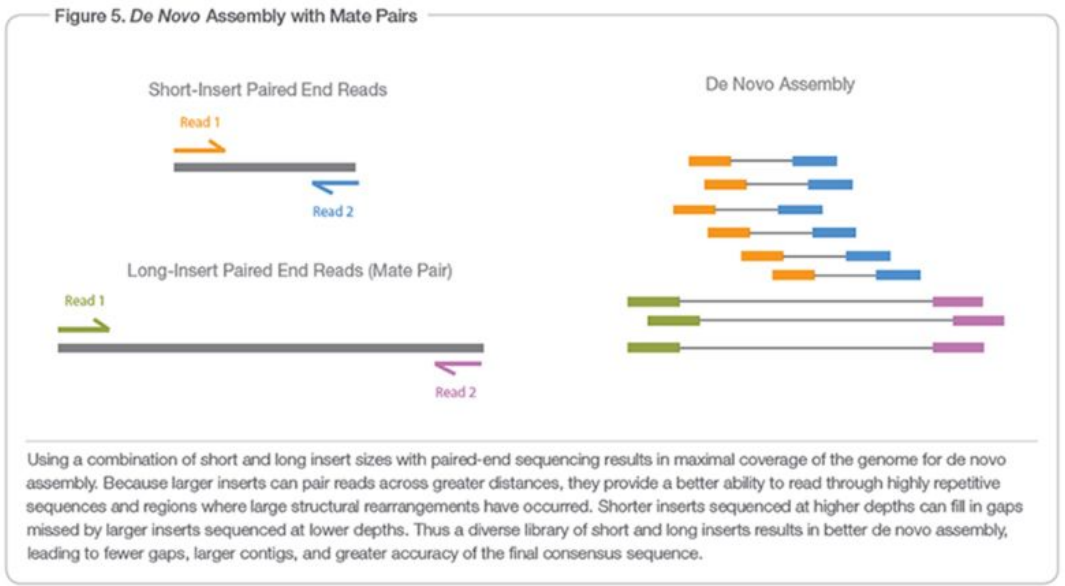
Single-end vs Paired

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

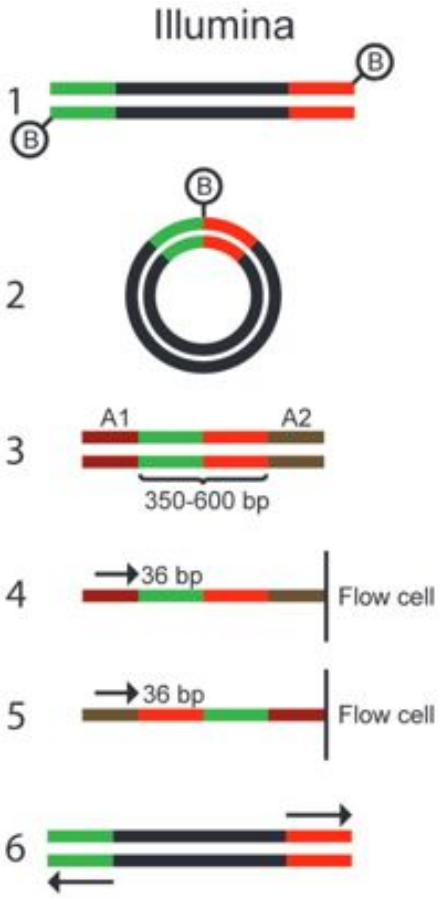


MATE-Pair sequencing ?

- For very long insert size preparation
  - Genome finishing
  - Structural variant detection
  - Identification of complex genomic rearrangements



MATE-Pair library preparation



- Fragments are end-repaired using biotinylated nucleotides (1). After circularization, the two fragment ends (green and red) become located adjacent to each other
- The circularized DNA is fragmented, and biotinylated fragments are purified by affinity capture. Sequencing adapters (A1 and A2) are ligated to the ends of the captured fragments (3).
- The fragments are hybridized to a flow cell, in which they are bridge amplified. (4,5,6).

Next-generation sequencing technologies and applications for human genetic history and forensics. Investigative Genetics, 2 (1), 1-15.

Applications of High throughput genomics

Is the 1000 $ genome for real ?

- The first sequenced human genome cost nearly $3 billion

The HiSeq X Ten probably will not be able to immediately sequence human genomes for under $1,000, but it will get close. Flatley's breakdown of projected HiSeq X Ten sequencing costs included the cost of reagents needed to run the machine ($797 per genome), the depreciated cost of the machine itself ($137 per genome), and the costs of paying technicians to run the machines and of preparing samples for sequencing ($55–65 per genome). But it left out the overhead costs that academic centers must pay, such as the costs of electricity needed to run the machines.

The HiSeq X Ten is composed of ten HiSeq X machines, and sells for at least $10 million.

- What about pricing for analysis ?

HiSeq X 10: a sequencer for factory-scale sequencing

- Illumina
- A set of 10 sequencers.
  - Each producing 1,8 Terabases / 3 day
- 18,000 genome / year
  - ”Factory-scale sequencing technology.
- 1000$ genome coming true….

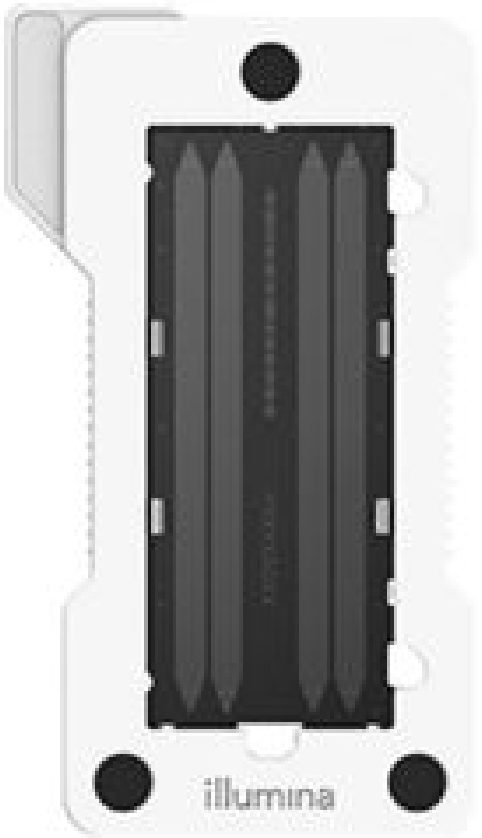
Some computing issues...

http://glennklockwood.blogspot.nl/

- 18,000 / year ~ 340/ week
- 30-50To storage / weak
  - Cost of long term storage ?
- 518 core hours / genome
- 175,000 core hours per week

Other Illumina sequencers

Key Methods	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.			
	NextSeq 500		HiSeq 2500		HiSeq 3000	HiSeq 4000
Run Mode	Mid-Output	High-Output	Rapid Run	High-Output	N/A	N/A
Flow Cells per Run	1	1	1 or 2	1 or 2	1	1 or 2
Output Range	20-39 Gb	30-120 Gb	10-300 Gb	50-1000 Gb	125-750 Gb	125-1500 Gb
Run Time	15-26 hours	12-30 hours	7-60 hours	<1-6 days	<1-3.5 days	<1-3.5 days
Reads per Flow Cell†	130 million	400 million	300 million	2 billion	2.5 billion	2.5 billion
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 250 bp	2 x 125 bp	2 x 150 bp	2 x 150 bp
System Overview	Speed and simplicity for everyday genomics.		Power and efficiency for large-scale genomics.		Maximum throughput and lowest cost for production-scale genomics.	Maximum throughput and lowest cost for production-scale genomics.



https://www.illumina.com/systems/sequencing.html

Sequencer comparison

Table 1 Characteristics of second-generation and third-generation sequencing instruments

Instrument	Read length (nucleotides)	No. of reads[a]	Output (Gb)[a]	No. of samples[a, b]	Runtime	Advantages	Disadvantages
Roche 454 GS FLX+	700[c]	1×10^6	0.7	192[d]	23 h	Long reads, short run time	Homopolymer errors, expensive
Illumina HiSeq2000	100[e]	3×10^9	600	384	11 days[f]	High yield	No. of index tags limiting
Life Technologies SOLiD 5500xl	75[g]	1.5×10^9	180	1,152	14 days[f]	Inherent error correction	Short reads[g]
Roche 454 GS Junior	400[c]	1×10^5	0.035	132	9 h	Long reads	Homopolymer errors, expensive
Illumina MiSeq	150	5×10^6	1.5	96	27 h	Short run time, ease of use	Expensive per base
Ion Torrent PGM Ion 316 chip	> 100[h]	1×10^6	0.1	16	2 h	Short run time, low reagent cost	Not well evaluated
Helicos BioSciences HeliScope	35[h]	1×10^9	35	4,800	8 days	SMS, sequences RNA	Short reads, high error rate
Pacific Biosciences PacBio RS	> 1,000[h]	1×10^5	0.1	1	90 min	SMS, long reads, short run time	High error rate, low yield

Most of this information is subject to rapid change, and the aim of this table is not to present absolute numbers but to provide a general comparison between different sequencing systems.

[a]Numbers calculated for two flow cells on HiSeq2000 and SOLiD 5500xl.

[b]Calculated as no. of index tags (provided by the sequencing company) × no. of divisions on solid support.

[c]Average for single-end sequencing, paired-end reads are shorter.

[d]No. of reads decreases when the PicoTiterPlate is divided.

[e]36 nucleotides for mate-pair reads.

[f]Run time depends on the read length, and on whether one or two flow cells are used.

[g]Second read in paired-end sequencing is limited to 35 nucleotides, and mate pair reads to 60 nucleotides.

[h]Average.

SMS = single molecule sequencing.

The MinION portable sequencer...

Long read lengths

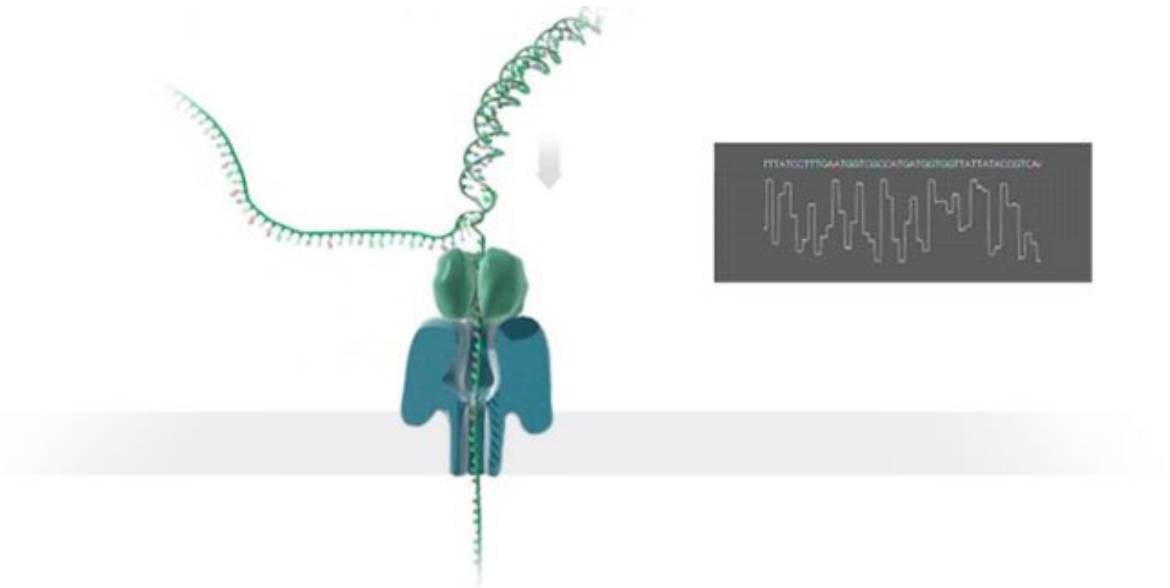
The Oxford Nanopore system processes the reads that are presented to it rather than generating specific read lengths. The longest read reported by a MinION user to date is more than 200Kb, but it can process the spectrum of read lengths.

Long read lengths

Real-time data

Direct molecular analysis

Portability



A nanopore is a nano-scale hole. In its devices, Oxford Nanopore passes an ionic current through nanopores and measures the changes in current as biological molecules pass through the nanopore or near it. The information about the change in current can be used to identify that molecule.

(DNA strand sequencing, illustrative data only)

"The Oxford Nanopore Technologies (ONT) MinION is a new sequencing technology that potentially offers read lengths of tens of kilobases (kb) limited only by the length of DNA molecules presented to it." ~1Gb to 2 Gb of sequence per minION. Detect DNA modifications.

https://nanoporetech.com/science-technology/how-it-works

Example application of MinION

Real-time, portable genome sequencing for Ebola surveillance.

Nature doi:10.1038/nature16996

View it >

Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Stephan Günther, Miles W. Carroll *et al*

Abstract

The Ebola virus disease epidemic in West Africa is the largest on record, responsible for over 28,599 cases and more than 11,299 deaths. Genome sequencing in viral outbreaks is desirable to characterize the infectious agent and determine its evolutionary rate. Genome sequencing also allows the identification of signatures of host adaptation, identification and monitoring of diagnostic targets, and characterization of responses to vaccines and treatments. The Ebola virus (EBOV) genome substitution rate in the Makona strain has been estimated at between 0.87 × 10−3 and 1.42 × 10−3 mutations per site per year. This is equivalent to 16–27 mutations in each genome, meaning that sequences diverge rapidly enough to identify distinct sub-lineages during a prolonged epidemic. Genome sequencing provides a high-resolution view of pathogen evolution and is increasingly sought after for outbreak surveillance. Sequence data may be used to guide control measures, but only if the results are generated quickly enough to inform interventions. Genomic surveillance during the epidemic has been sporadic owing to a lack of local sequencing capacity coupled with practical difficulties transporting samples to remote sequencing facilities. To address this problem, here we devise a genomic surveillance system that utilizes a novel nanopore DNA sequencing instrument. In April 2015 this system was transported in standard airline luggage to Guinea and used for real-time genomic surveillance of the ongoing epidemic. We present sequence data and analysis of 142 EBOV samples collected during the period March to October 2015. We were able to generate results less than 24 h after receiving an Ebola-positive sample, with the sequencing process taking as little as 15–60 min. We show that real-time genomic surveillance is possible in resource-limited settings and can be established rapidly to monitor outbreaks.

And now the Smidgion...

SmidgION: nanopore sensing for use with mobile devices

Using the same core technology as the handheld MinION device, we are now starting to develop an even smaller device.

In early development

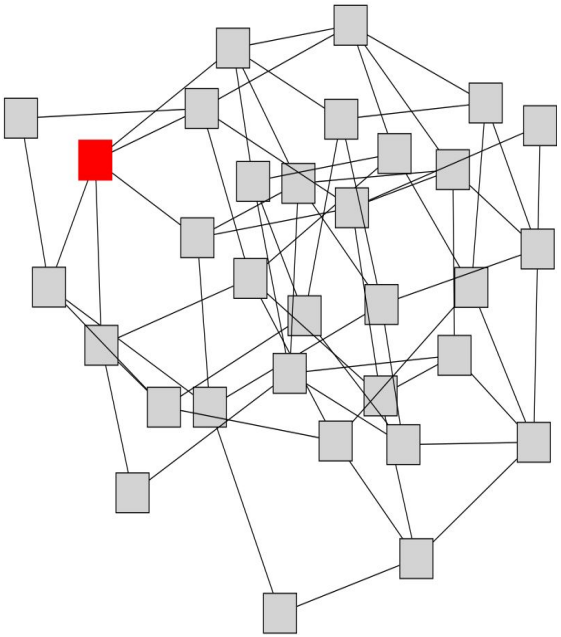
Some applications of DNA sequencing: genetic variation analysis

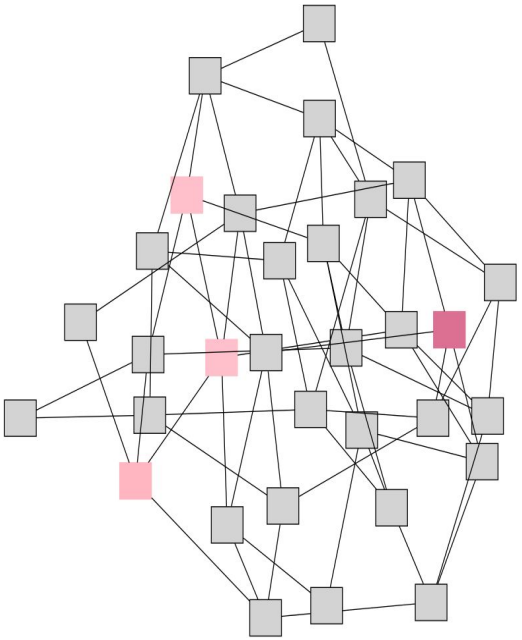
- Analysis of genome diversity
  - SNPs (Single Nucleotide Polymorphisms)
  - InDel (Insertion/Deletion)
  - CNV (Copy Number Variation)

- E.g The 1000 genome Project
- Goal: decifer the mechanisms driving complexe diseases

Monogenic vs complexe disease

- In complexe diseases, the phenotype is driven by a set of loci whose penetrance is low (polygenic)
- Complexe diseases are also viewed as multifactorial (i.e also influenced by environment)





Genetic variations in human

- 1000 genomes project

1,092 individuals from 14 populations, constructed using a combination of low-coverage **whole-genome** and **exome Sequencing**

- 38 millions SNPs, 1.4 million indels

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium

Affiliations | Contributions | Corresponding author

Nature **491**, 56–65 (01 November 2012) | doi:10.1038/nature11632

Received 04 July 2012 | Accepted 01 October 2012 | Published online 31 October 2012

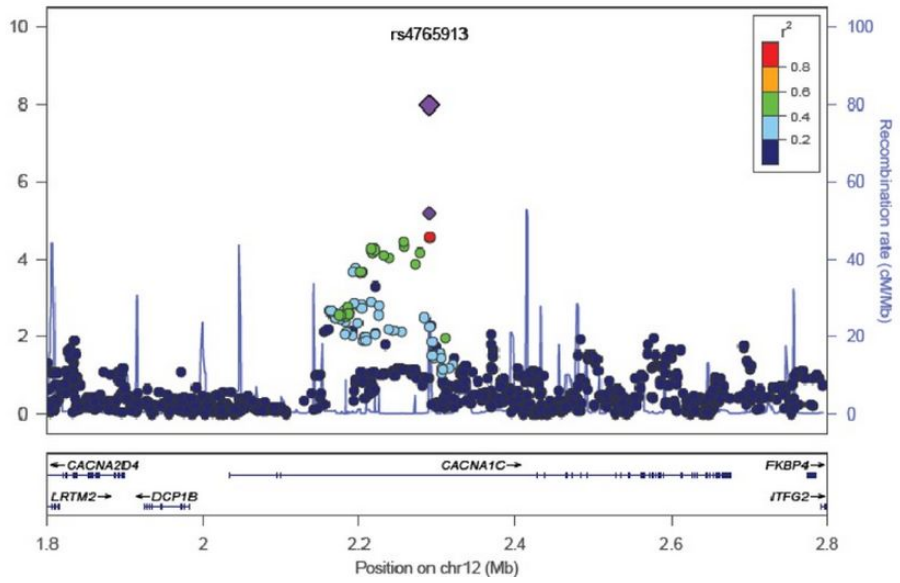
GWAS analysis

Nat Genet. Author manuscript; available in PMC May 1, 2013.
Published in final edited form as:
Nat Genet. Oct 2011; 43(10): 977–983.
Published online Sep 18, 2011. doi: 10.1038/ng.943
INSERM Subrepository
PMCID: PMC3637176
HALMS: HALMS634944

Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4

Bipolar disorder (BD) is a severe mood disorder affecting greater than 1% of the population[1]. Classical BD is characterized by recurrent manic episodes that often alternate with depression. Its onset is in late adolescence or early adulthood and results in chronic illness with moderate to severe impairments (...).

Genome-wide significant evidence for association was confirmed for *CACNA1C* and found for a novel gene *ODZ4* (...).
Pathway analysis identified a pathway comprised of subunits of calcium channels enriched in the bipolar disorder association intervals.



Genetic variation ongoing project: BGI

Human

The Million Human Genomes Project was launched by BGI to decode the genome of over 1 million people in November 2011. This project concludes five essential parts: Ancient genomes, Population genomes, Medical genomes, Cell genomes and Personal genomes.

The aim of this project is to establish the research baseline and reference standard for specific populations, as well as to connect the phenotypes of diseases and traits with the genetic variations to understand the disease mechanism.

The integrative genome message and scientific discoveries obtaining from the project will lay the foundation for guiding the innovative clinical diagnosis and treatment, and ultimately advancing personalized healthcare and improving human health.

华大基因 BGI

U.S. proposes effort to analyze DNA from 1 million people

WASHINGTON | BY TONI CLARKE AND SHARON BEGLEY

The Obama Administration has just announced a Million Genomes Project – and it's not even the first.

Now both Craig Venter and Francis Collins, leads of the private and public versions of the Human Genome Project, are working on their million-omes.

The company 23andMe might be the first 'million-ome-aire'. By 2014, the company founded by Ann Wojcicki processed upwards of 800,000 customer samples. Pundit Eric Topol suggests in his article "Who Owns Your DNA" that without the skirmish with the FDA, 23andMe would already have millions.

In 2011, China's BGI, the world's largest genomics research company, boldly announced a million human genomes project. Building on projects like the panda genome and the 3000 Rice Genomes project, the BGI is building new next-generation sequencing technologies to support its flagship project.

Also in 2011, the United States Veterans Affairs (VA) Research and Development program launched its Million Veteran Program (MVP) aiming to build the world's largest database of genetic, military exposure, lifestyle, and health information. The "large, diverse, and altruistic patient population" of the VA puts it ahead of the others in collecting samples.

http://blog.oup.com/2015/02/millions-genomes-project/

Yet another ongoing project: Calico

Larry Page at Google's headquarters

being

MOUNTAIN VIEW, CA – September 18, 2013 – Google today announced Calico, a new company that will focus on health and well-being, in particular the challenge of aging and associated diseases. Arthur D. Levinson, Chairman and former CEO of Genentech and Chairman of Apple, will be Chief Executive Officer and a founding investor.

Announcing this new investment, Larry Page, Google CEO said: "Illness and aging affect all our families. With some longer term, moonshot thinking around healthcare and biotechnology, I believe we can improve millions of lives. It's impossible to imagine anyone better than Art—one of the leading scientists, entrepreneurs and CEOs of our generation—to take this new venture forward." Art said: "I've devoted much of my life to science and technology, with the goal of improving human health. Larry's focus on outsized improvements has inspired me, and I'm tremendously excited about what's next."

Art Levinson will remain Chairman of Genentech and a director of Hoffmann-La Roche, as well as Chairman of Apple.

Commenting on Art's new role, Franz Humer, Chairman of Hoffmann-La Roche, said: "Art's track record at Genentech has been exemplary, and we see an interesting potential for our companies to work together going forward. We're delighted he'll stay on our board."

Tim Cook, Chief Executive Officer of Apple, said: "For too many of our friends and family, life has been cut short or the quality of their life is too often lacking. Art is one of the crazy ones who thinks it doesn't have to be this way. There is no one better suited to lead this mission and I am excited to see the results."

Yet another ongoing project : HLI

Human Longevity Inc. (HLI) Launched to Promote Healthy Aging Using Advances in Genomics and Stem Cell Therapies

HLI is Building World's Largest Genotype/Phenotype Database by Sequencing up to 40,000 Human Genomes/Year Combined with Microbiome, Metabolome and Clinical Data to Develop Life Enhancing Therapies

HLI has Purchased Two Illumina HiSeq X Ten Sequencing Systems

SAN DIEGO, CA (March 4, 2014)—Human Longevity Inc. (HLI), a genomics an cell therapy-based diagnostic and therapeutic company focused on extendin the healthy, high performance human life span, was announced today by co-founders J. Craig Venter, Ph.D., Robert Hariri, M.D., Ph.D., and Peter H. Diamandis, M.D.

The company, headquartered in San Diego, California, is being capitalized wi an initial $70 million in investor funding.

HLI's funding is being used to build the largest human sequencing operation in the world to compile the most comprehensive and complete human genotype, microbiome, and phenotype database available to tackle the diseases associated with aging-related human biological decline. HLI is also leading the development of cell-based therapeutics to address age-related decline in endogenous stem cell function. Revenue streams will be derived

HLI has initially purchased two Illumina HiSeq X Ten Sequencing Systems (with the option to acquire three additional systems) to sequence up to 40,000 human genomes per year, with plans to rapidly scale to 100,000 human genomes per year. HLI will sequence a variety of humans—children, adults and super centenarians and those with disease and those that are healthy.

HLI is uniquely positioned to identify therapeutic solutions to preserve the healthy, high performing body by focusing on some of the most prevalent and actionable areas. HLI is concentrating on cancer, diabetes and obesity, heart and liver diseases, and dementia with its team of expert scientists and clinicians. The company has established strategic collaborations with Metabolon Inc., University of California, San Diego, and the J. Craig Venter Institute (JCVI).

Whole-Genome Sequencing of the World's Oldest People

Hinco J. Gierman, Kristen Fortney, Jared C. Roach, Natalie S. Coles, Hong Li, Gustavo Glusman, Glenn J. Markov, Justin D. Smith, Leroy Hood, L. Stephen Coles, Stuart K. Kim

Published: November 12, 2014 • DOI: 10.1371/journal.pone.0112430

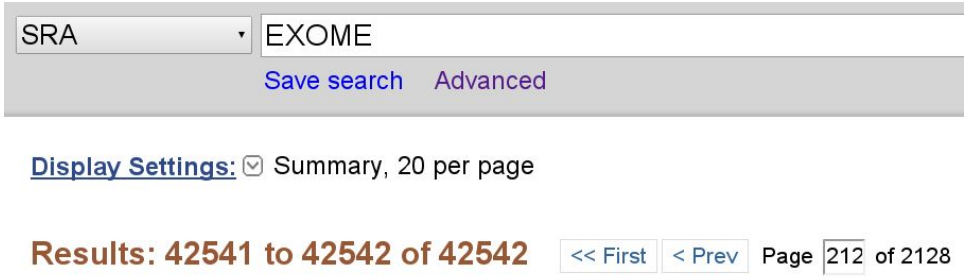
Affiliation: Depts. of Developmental Biology and Genetics, Stanford University, Stanford, CA, United States of America

Abstract

Supercentenarians (110 years or older) are the world's oldest people. Seventy four are alive worldwide, with twenty two in the United States. We performed whole-genome sequencing on 17 supercentenarians to explore the genetic basis underlying extreme human longevity. We found no significant evidence of enrichment for a single rare protein-altering variant or for a gene harboring different rare protein altering variants in supercentenarian compared to control genomes. We followed up on the gene most enriched for rare protein-altering variants in our cohort of supercentenarians, TSHZ3, by sequencing it in a second cohort of 99 long-lived individuals but did not find a significant enrichment. The genome of one supercentenarian had a pathogenic mutation in DSC2, known to predispose to arrhythmogenic right ventricular cardiomyopathy, which is recommended to be reported to this individual as an incidental finding according to a recent position statement by the American College of Medical Genetics and Genomics. Even with this pathogenic mutation, the proband lived to over 110 years. The entire list of rare protein-altering variants and DNA sequence of all 17 supercentenarian genomes is available as a resource to assist the discovery of the genetic basis of extreme longevity in future studies.

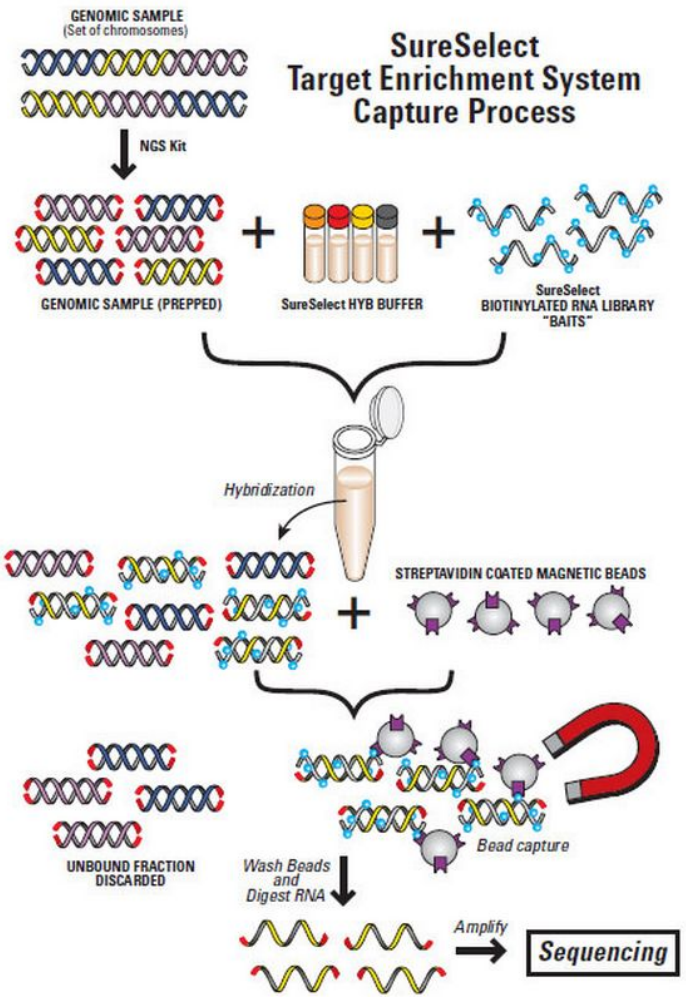
Reducing costs: exome sequencing

- Exome sequencing
  - Sequencing large dataset is expensive
    - Focus on exons (using beads or microarrays to capture genomic regions)
  - Application examples
    - Tumor genome Sequencing
    - Monogenic disease
    - Complexe disease



Targeted sequencing (E.g Exome)

- Agilent
  - SureSelect
- Roche NimbleGen
  - SeqCap EZ library
- Illumina
  - Nextera



Exome Sequencing : Miller Syndrome

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure & Michael J Bamshad

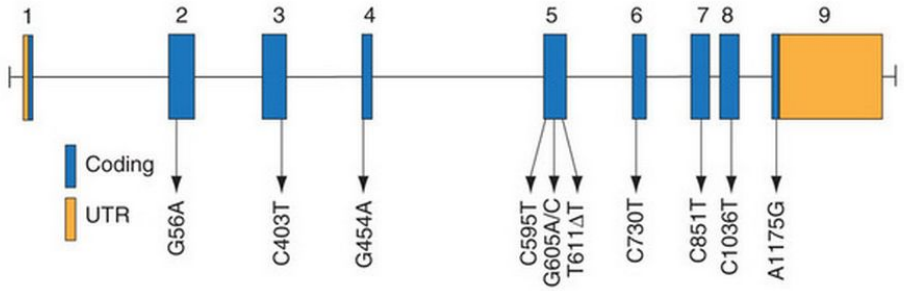
Affiliations | Contributions | Corresponding authors

Nature Genetics **42**, 30–35 (2010) | doi:10.1038/ng.499

Received 02 October 2009 | Accepted 09 November 2009 | Published online 13 November 2009

We demonstrate the first successful application of exome sequencing to discover the gene for a rare mendelian disorder of unknown cause, Miller syndrome (MIM%263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40× and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and eight HapMap exomes for genes with two previously unknown variants in each of the four individuals identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes underlying rare mendelian disorders and will likely transform the genetic analysis of monogenic traits.

Figure 2: Genomic structure of the exons encoding the open reading frame of *DHODH*.



DHODH is composed of nine exons that encode untranslated regions (UTR) (orange) and protein coding sequence (blue). Arrows indicate the locations of 11 different mutations found in 6 families with Miller syndrome.

Studying tumors

- Mutations / Indel
  - Exome seq
  - Whole genome sequencing
- Genomic rearrangements analysis
  - E.g Mate-pair approach (translocation,...)
- Gene expression deregulation
  - Transcriptome analysis (RNA-Seq)
  - Regulatory region analysis (ChIP-Seq)
  - Fusion transcripts

Exome sequencing of renal cell carcinoma

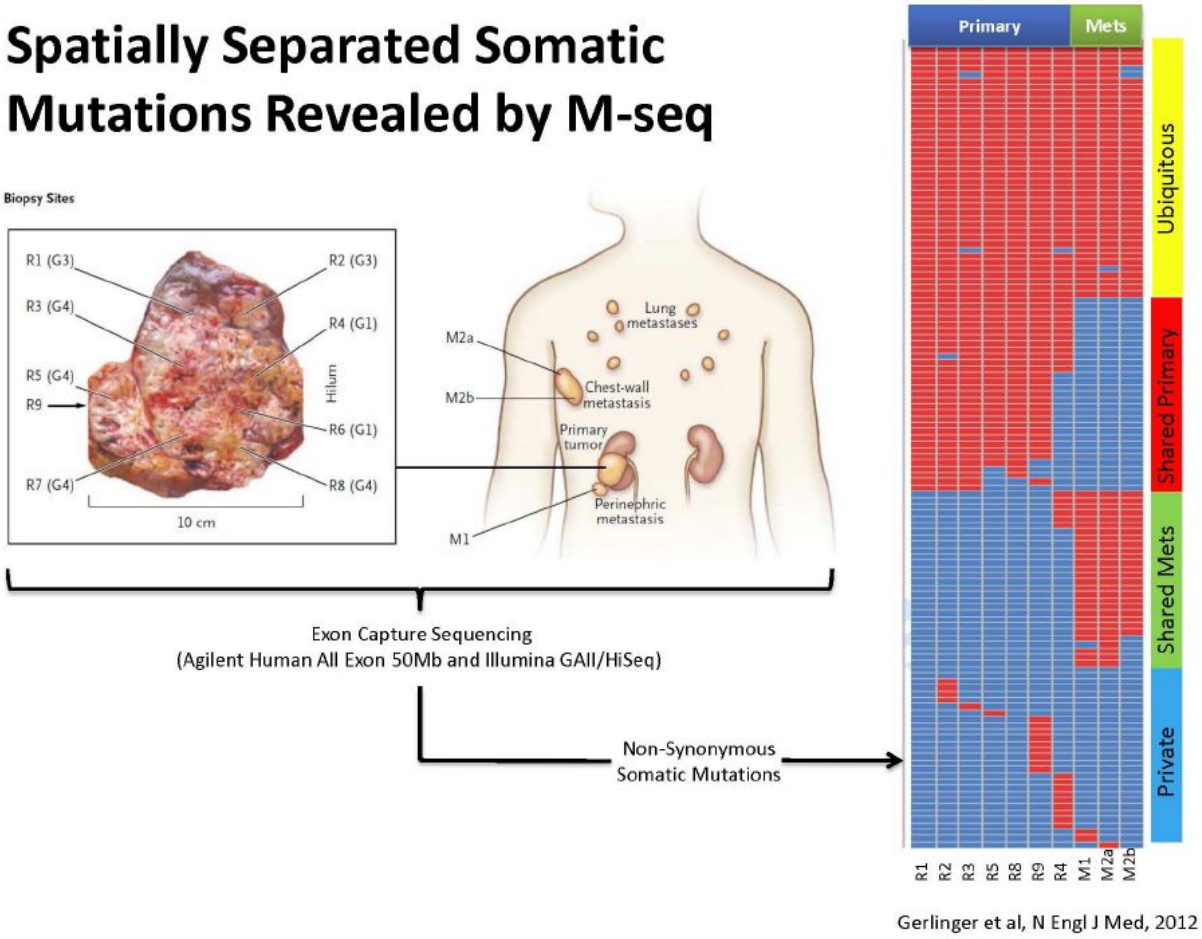
Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing

Marco Gerlinger, M.D., Andrew J. Rowan, B.Sc., Stuart Horswell, M.Math., James Larkin, M.D., Ph.D., David Endesfelder, Dip.Math., Eva Gronroos, Ph.D., Pierre Martinez, Ph.D., Nicholas Matthews, B.Sc., Aengus Stewart, M.Sc., Patrick Tarpey, Ph.D., Ignacio Varela, Ph.D., Benjamin Phillimore, B.Sc., Sharmin Begum, M.Sc., Neil Q. McDonald, Ph.D., Adam Butler, B.Sc., David Jones, M.Sc., Keiran Raine, M.Sc., Calli Latimer, B.Sc., Claudio R. Santos, Ph.D., Mahrokh Nohadani, H.N.C., Aron C. Eklund, Ph.D., Bradley Spencer-Dene, Ph.D., Graham Clark, B.Sc., Lisa Pickering, M.D., Ph.D., Gordon Stamp, M.D., Martin Gore, M.D., Ph.D., Zoltan Szallasi, M.D., Julian Downward, Ph.D., P. Andrew Futreal, Ph.D., and Charles Swanton, M.D., Ph.D.

N Engl J Med 2012; 366:883-892 | March 8, 2012 | DOI: 10.1056/NEJMoa1113205

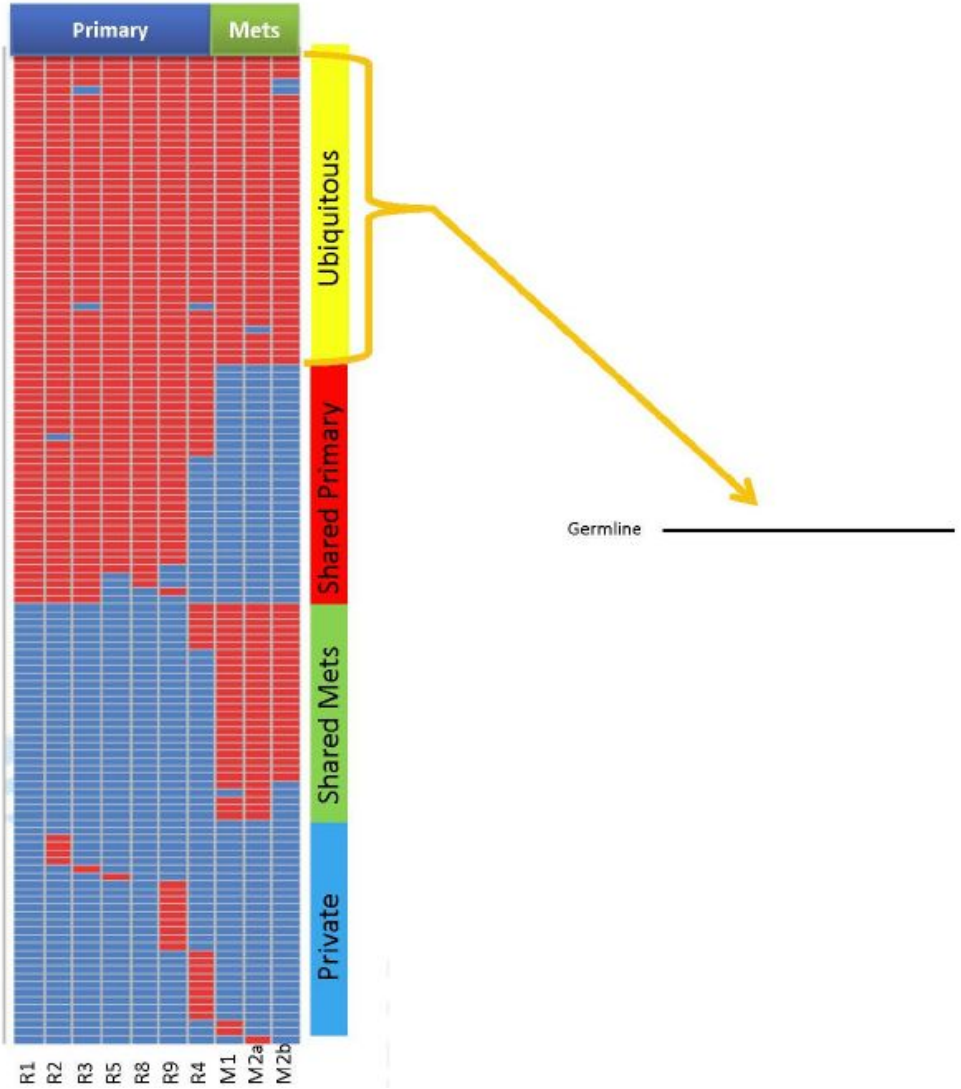
Cancer a clonal disease evolving in a linear fashion ?
What about tumor heterogeneity ?
Can we re-constitute the evolution of the tumor ?

Exome-Seq of Renal cell carcinoma

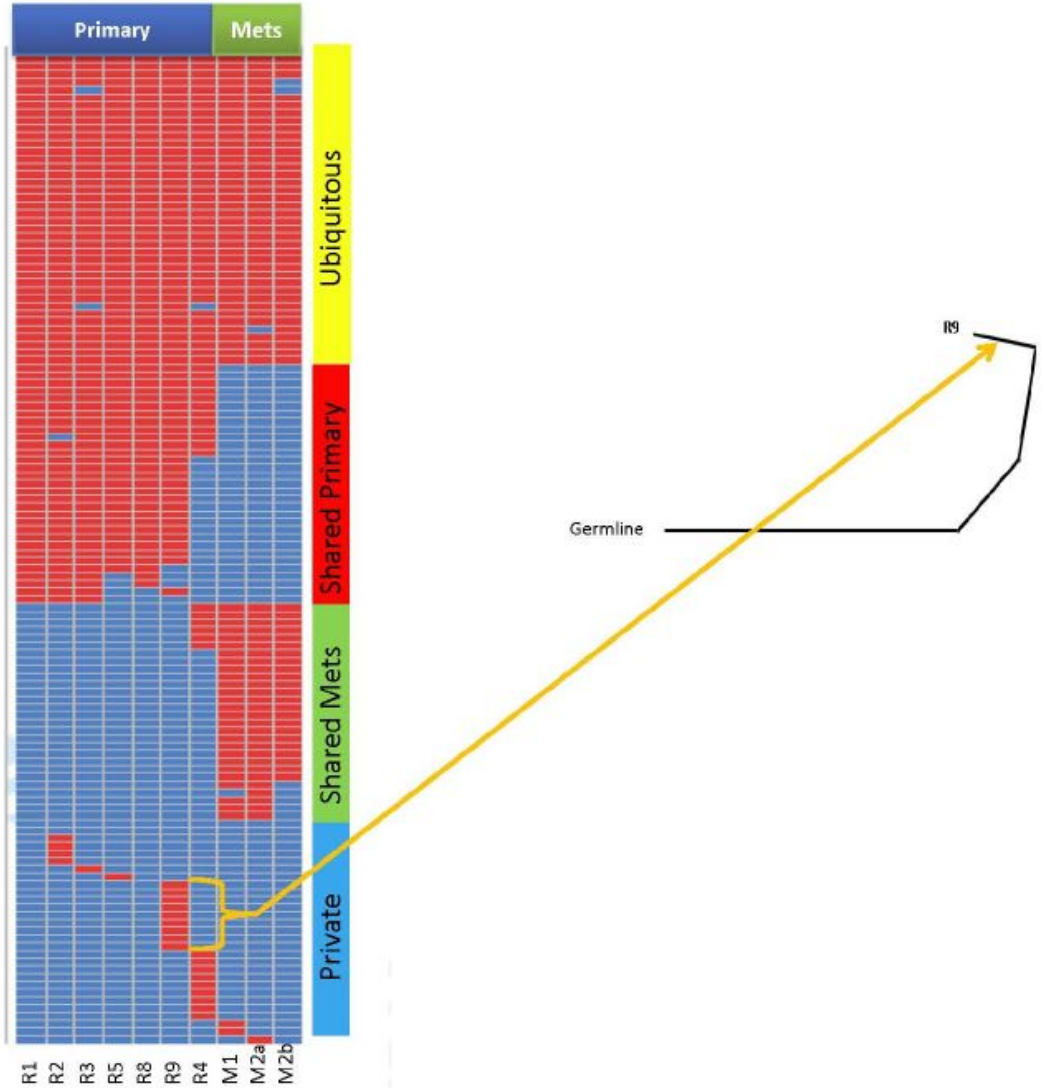


Gerlinger et al, N Engl J Med, 2012

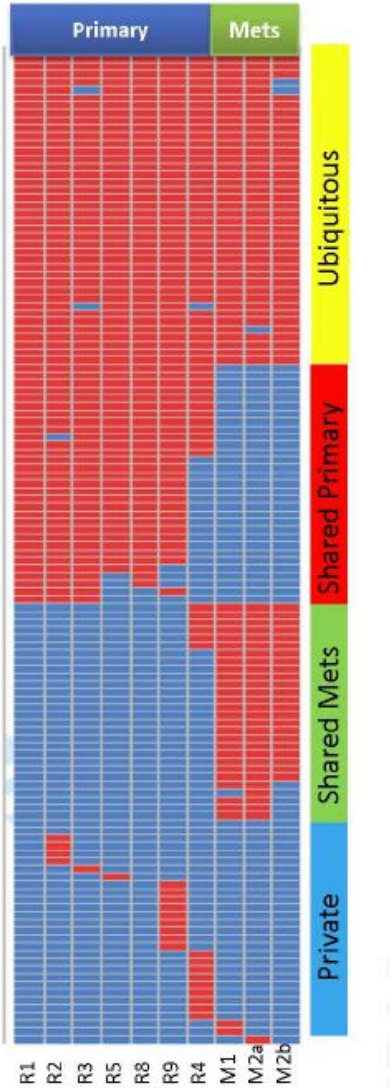
Phylogenetic reconstruction by clonal ordering

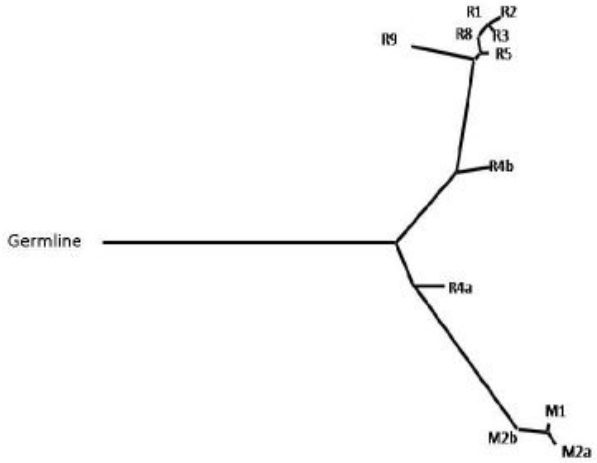


Phylogenetic reconstruction by clonal ordering

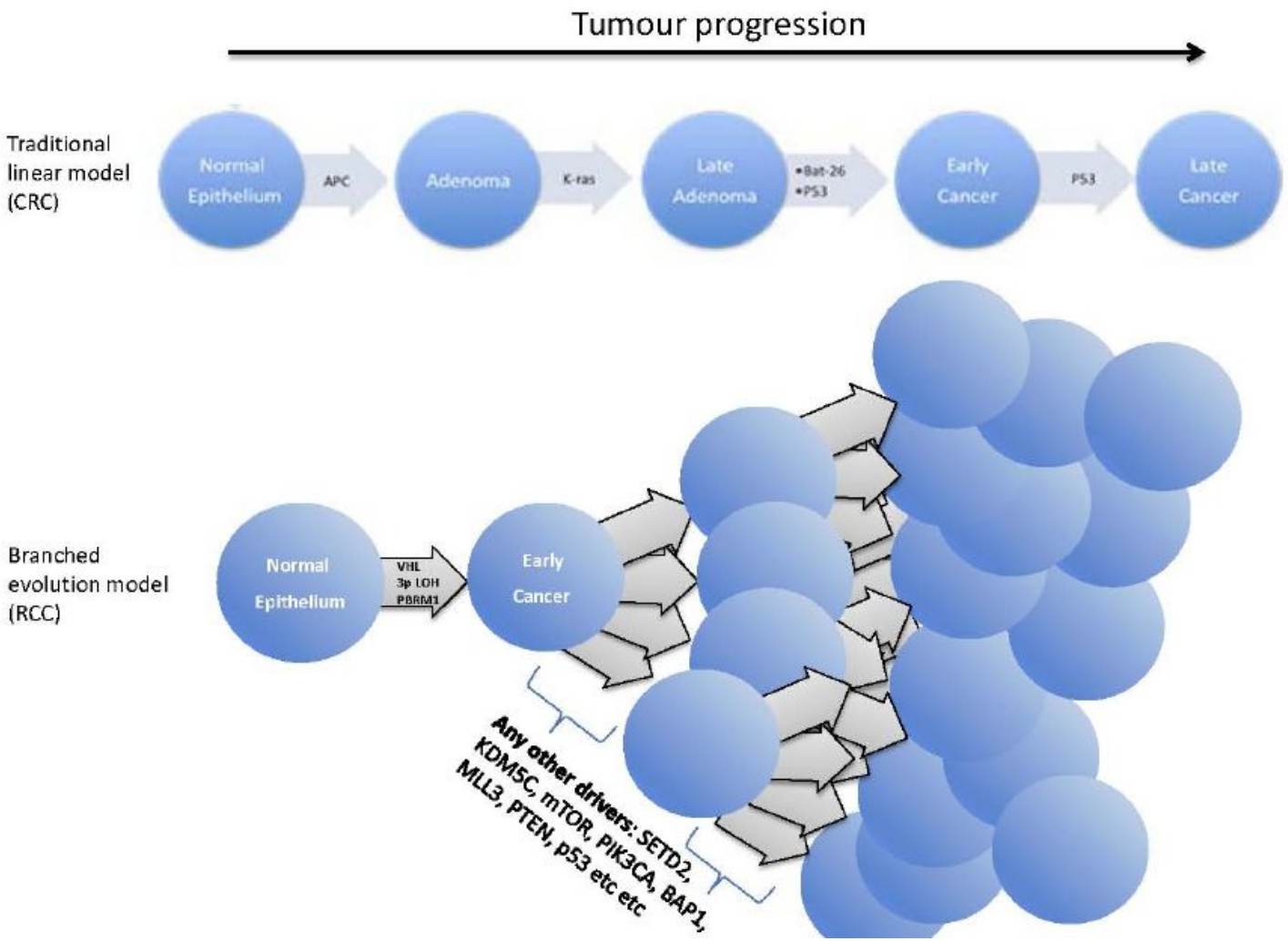


Phylogenetic reconstruction by clonal ordering

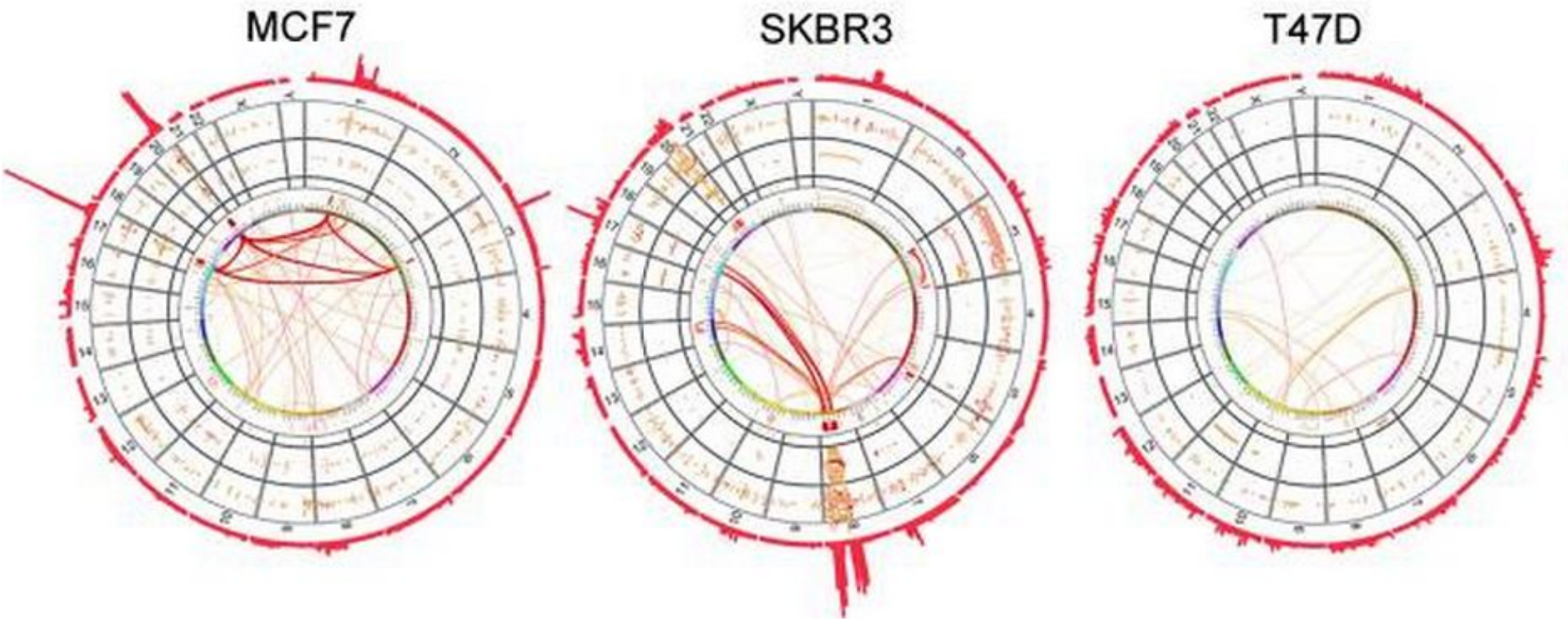


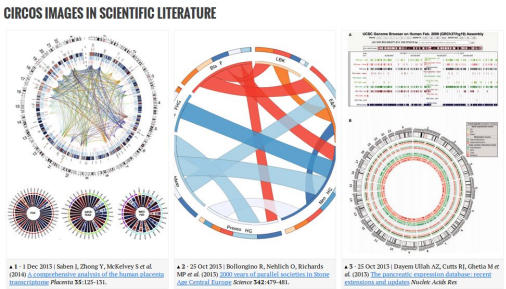


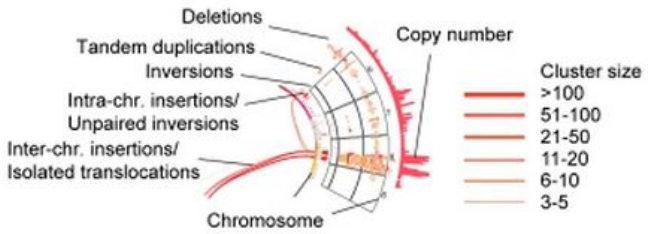
Cancer: A clonal disease evolving in a linear fashion?



Structural variations analysis







GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. May 2011; 21(5): 665–675. PMCID: PMC3083083
doi: 10.1101/gr.113555.110

Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes

Ongoing Project...

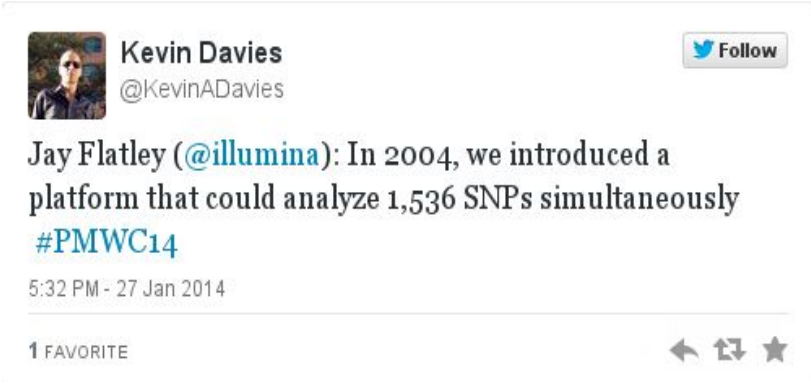
Illumina's Jay Flatley at #PMWC14: Get Sequence of 1 million cancer patients in next 5 years

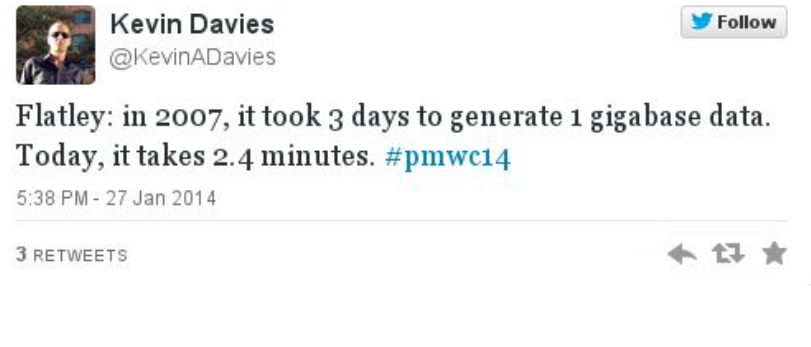
January 27, 2014 by nextgenseek · 1 Comment

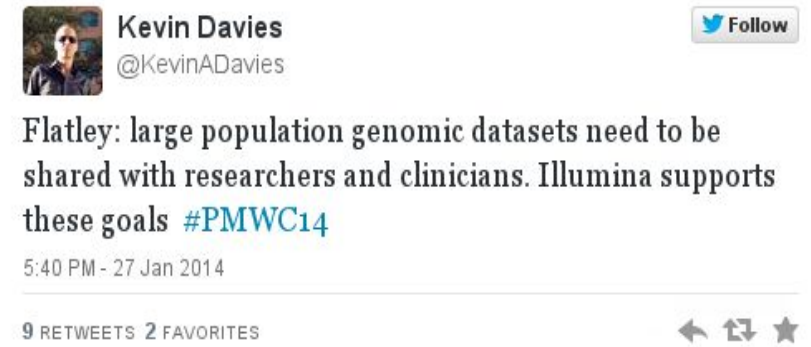
Like 3 Tweet 23 Pin it Share 11

Illumina's Jay Flately said at #PMWC14 that Illumina wants to have the sequence of 1 million cancer patients in a database in the next five years. And one of his personal goal is to make cancer a "chronic" disease within 10 years. Jay Flatley said Illumina support the goals of sharing large population genomic datasets with researchers and clinicians. This is the gist of Jay Flately's talk at #PMWC14 happening right now at Mountain View, CA.

Thanks to awesome live tweets by Kevin Davies, @DivaBioTech, and Theral Timpson. Here are the links to the original tweets.







Genome for everyone...

For One Baby, Life Begins with Genome Revealed

How a California father made an end run around medicine to decode his son's DNA.

By Antonio Regalado on June 13, 2014

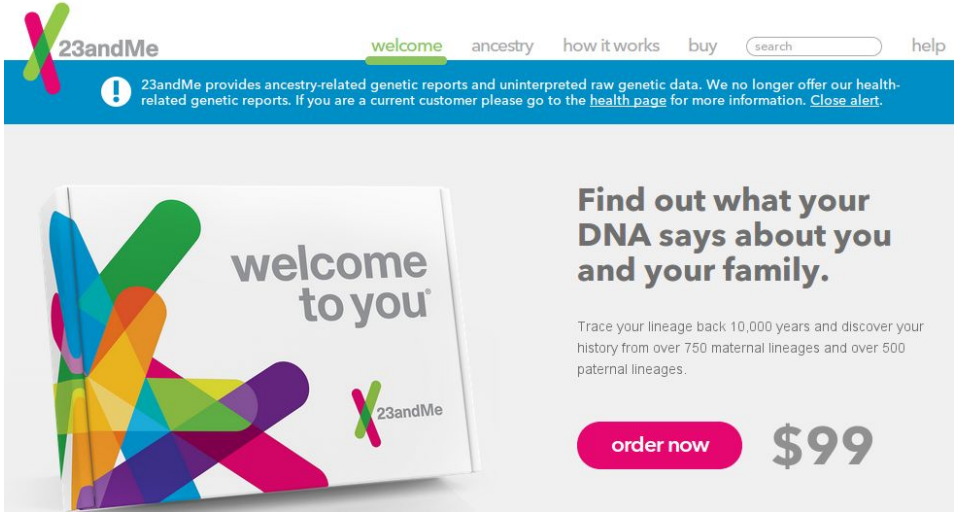
An infant delivered last week in California appears to be the first healthy person ever born in the U.S. with his entire genetic makeup deciphered in advance.

WHY IT MATTERS

Medical ethics is colliding with parents' desire for DNA data during pregnancy.

His father, Razib Khan, is a graduate student and professional blogger on genetics who says he worked out a rough draft of his son's genome early this year in a do-it-yourself fashion after managing to obtain a tissue sample from the placenta of the unborn baby during the second trimester.

"We did a work-around," says Khan, 37, who is now finishing a PhD in feline population genetics at the University of California, Davis. "There is no map for doing this, and there's no checklist."

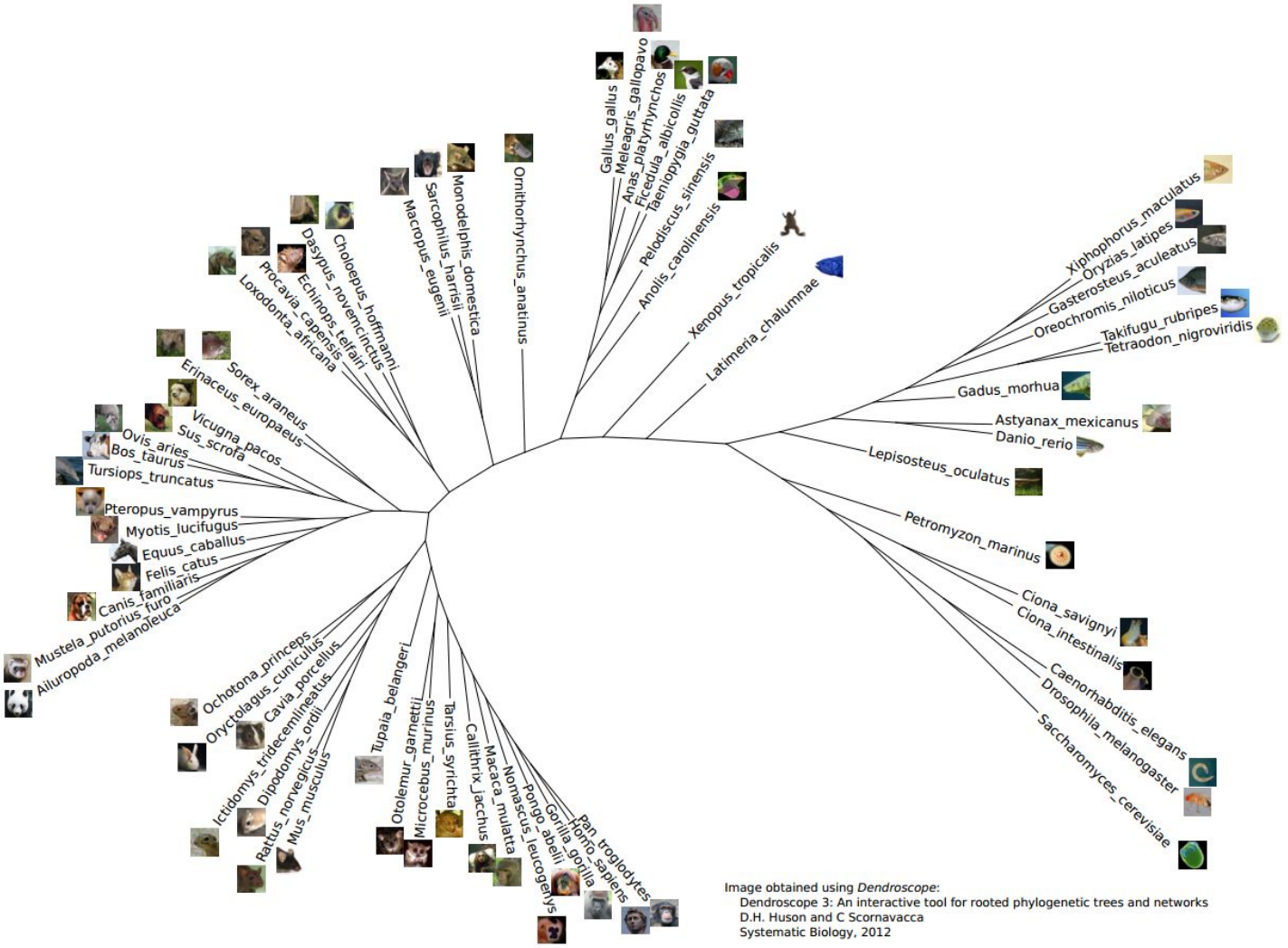


Vox GENETICS

With genetic testing, I gave my parents the gift of divorce

Updated by George Doe on September 9, 2014, 7:50 a.m. ET

Some examples of sequenced organims



Applications: analysing genome diversity across species

Plant & Animal

The Million Plant & Animal Genomes Project aims to generate reference genomes for thousands of economically and scientifically important plant/animal species and resequence millions of plant/animal specimens. This enormous project, to be carried out in collaboration with scientists worldwide, will ultimately generate a huge database of genetic information, allow dramatic improvement in the research of biodiversity conservation, evolutionary mechanism studies, gene function analyses, and help to build animal models for diseases, accelerate molecular breeding, etc. The primary goal for this project is to use genome sequencing and bioinformatics technologies to accelerate the development of practical mechanisms to ensure food security, promote medical applications, improve ecological conservation, and develop new energy sources.

GENOME 10K

Genome 10K Project

The Genome 10K project aims to establish a genomic 'zoo' — a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. Capturing the genetic diversity of vertebrate species will create an unprecedented resource for the life sciences and for worldwide conservation efforts.

i5k

i5k Initiative

The i5k initiative plans to sequence the genomes of 5,000 insect and related arthropod species over the next 5 years. It aims to sequence the genomes of all insect species known to have worldwide importance in agriculture, food safety, medicine, and energy production, and those with important scientific value in evolution and phylogeny research.

Million plant and animal genomes project

华大基因 BGI

Sequencing as a strategy to improve quality of crops (agrigenomics)

Data Note

Highly accessed Open Access

The 3,000 rice genomes project

The 3,000 rice genomes project†

Correspondence: The 3,000 rice genomes project

† Equal contributors

▾ Author Affiliations

Institute of Crop Sciences/National Key Facilities for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, 12 S. Zhong-Guan-Cun St, Beijing 100081, China

BGI, Bei Shan Industrial Zone, Yantian District, Shenzhen 518083, China

International Rice Research Institute, DAPO 7777, Metro Manila 1301, Philippines

GigaScience 2014, **3**:7 doi:10.1186/2047-217X-3-7

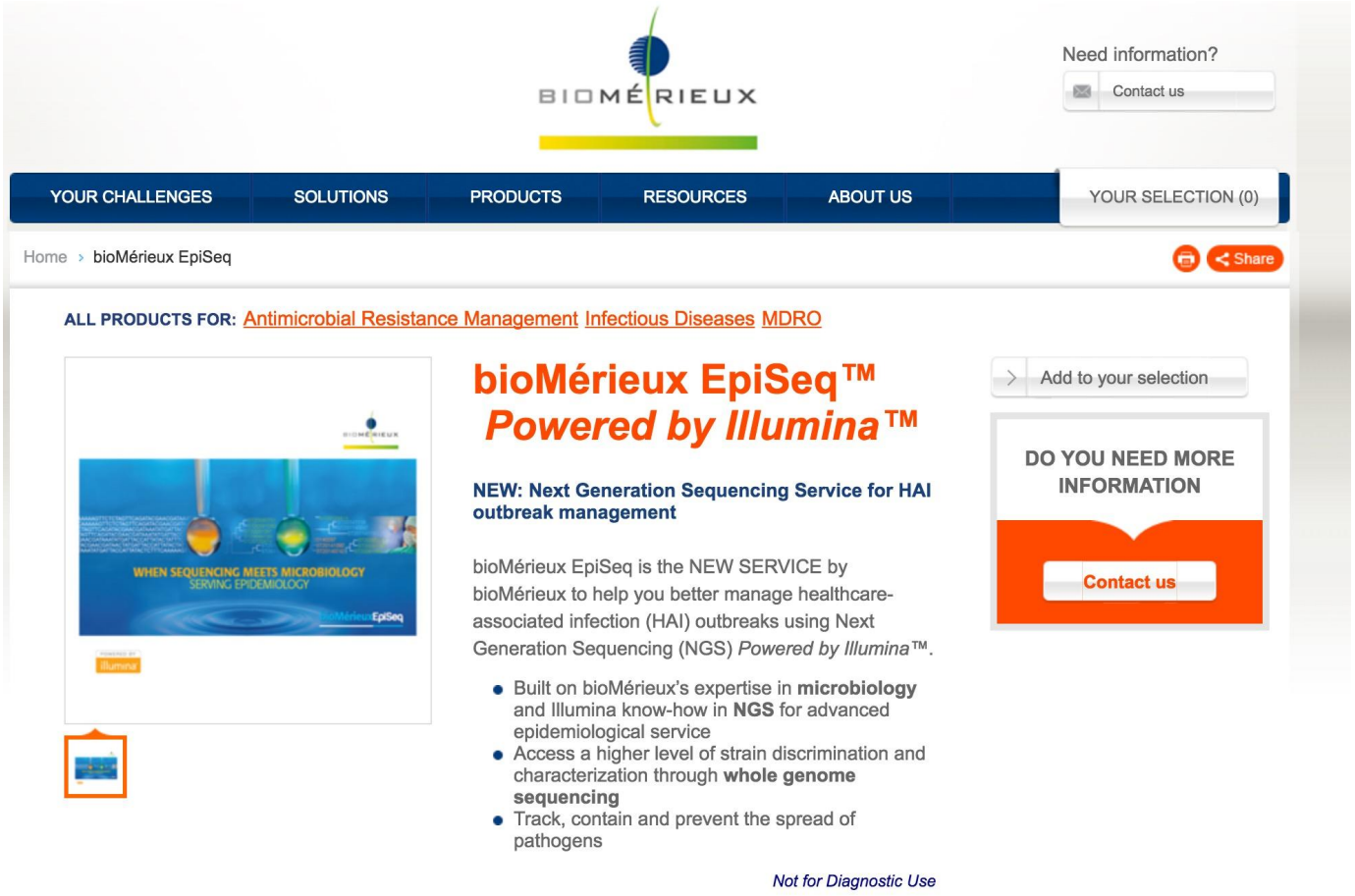
Background

Rice, *Oryza sativa* L., is the staple food for half the world's population. By 2030, the production of rice must increase by at least 25% in order to keep up with global population growth and demand. Accelerated genetic gains in rice improvement are needed to mitigate the effects of climate change and loss of arable land, as well as to ensure a stable global food supply.

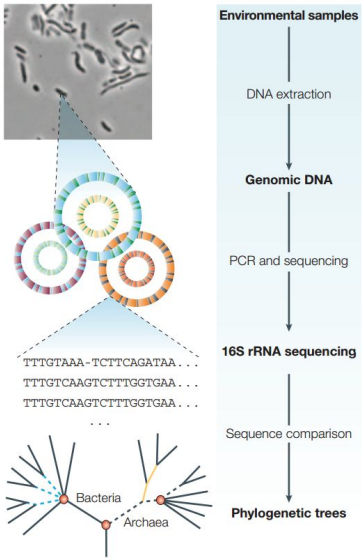
NB: rice genome size 430Mb

Microbial genomics

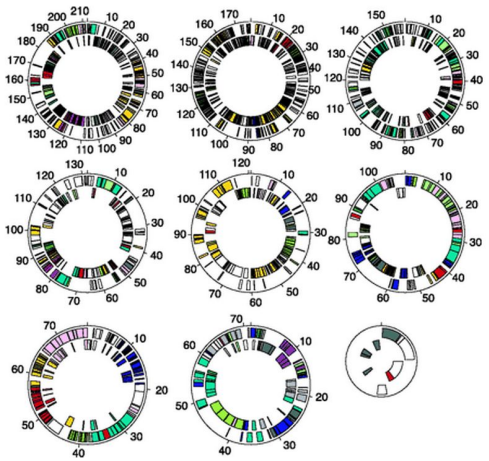
- Human microbiome
- Healthcare associated infections



Some application of DNA Sequencing: Metagenomics



Circular diagrams of nine complete megaplasmids. Genes encoded in the forward direction are shown in the outer concentric circle; reverse coding genes are shown in the inner concentric circle. The genes have been given role category assignment and colored accordingly: amino acid biosynthesis, violet; biosynthesis of cofactors, prosthetic groups, and carriers, light blue; cell envelope, light green; cellular processes, red; central intermediary metabolism, brown; DNA metabolism, gold; energy metabolism, light gray; fatty acid and phospholipid metabolism, magenta; protein fate and protein synthesis, pink; purines, pyrimidines, nucleosides, and nucleotides, orange; regulatory functions and signal transduction, olive; transcription, dark green; transport and binding proteins, blue-green; genes with no known homology to other proteins and



Soil and Agricultural Metagenomics

Sequencing has transformed environmental metagenomics, enabling the study of large microbial communities directly in their natural environment without prior culturing. These studies can yield important information about diverse microbial populations associated with animal and plant development, from rumen flora that enhance animal digestion to root-associated bacteria involved in nitrogen fixation.

NGS has been instrumental in advancing microbiology research. With NGS, you can measure changes anywhere in the genome without prior knowledge, which is critical for unculturable organisms. Single-base resolution allows tracking of microbial adaptation over short periods of time, both in the laboratory and in the environment.

Metagenomics: DNA sequencing of environmental samples

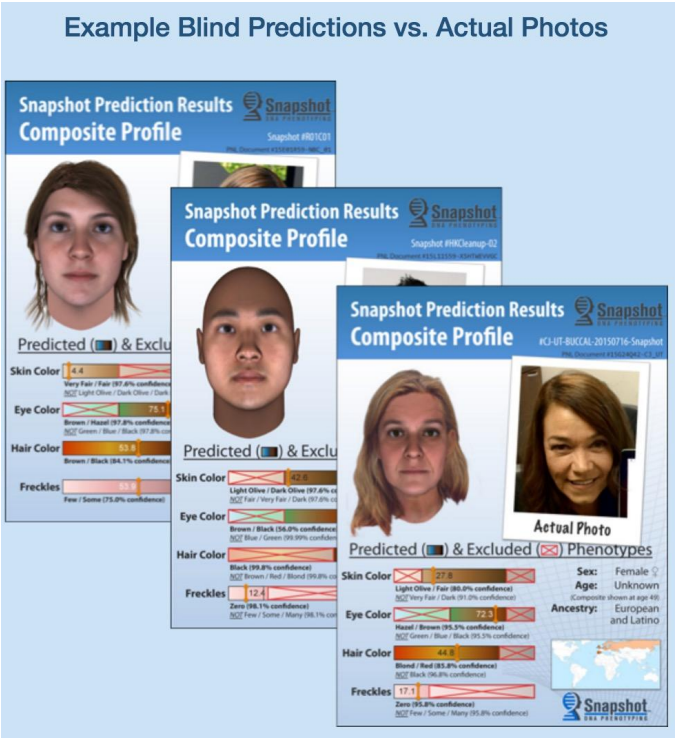
Susannah Green Tringe[1] & Edward M. Rubin[1] About the authors

Science. 2004 Apr 2;304(5667):66-74. Epub 2004 Mar 4.

Environmental genome shotgun sequencing of the Sargasso Sea.

Venter JC[1], Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO.

Forensic genomics



Workflow for Criminal Casework

Illumina sequencing by synthesis (SBS) technology on the MiSeq FGx Forensic Genomics System performs multiple tests at the same time with 1 nanogram or less of forensic casework samples. Using this sequencing system, crime laboratories can simultaneously analyze every locus now in use, plus hundreds more. This forensic next-generation sequencing (NGS) approach generates the maximum information possible from an evidence sample, or known reference sample, in a single MiSeq FGx run.

Workflow for Disaster Victim Identification

For disaster victim identification, Illumina offers a complete, fully validated DNA-to-data solution designed for forensic genomics. Our recommended workflow on the MiSeq FGx System utilizes the ForenSeq DNA Signature Prep Kit or the Nextera XT DNA Library Prep Kit.

Achieve high resolution and exceptional accuracy from as little as 1 ng of DNA—Even with complex mixtures or degraded DNA. The inherent sensitivity of Illumina chemistry helps detect minor components that might go undetected by conventional STR and CE analysis.

Sequencing and regulatory elements analysis

See S. Spiguglia talk.

Analysing chromosome cross-talks in three dimensions

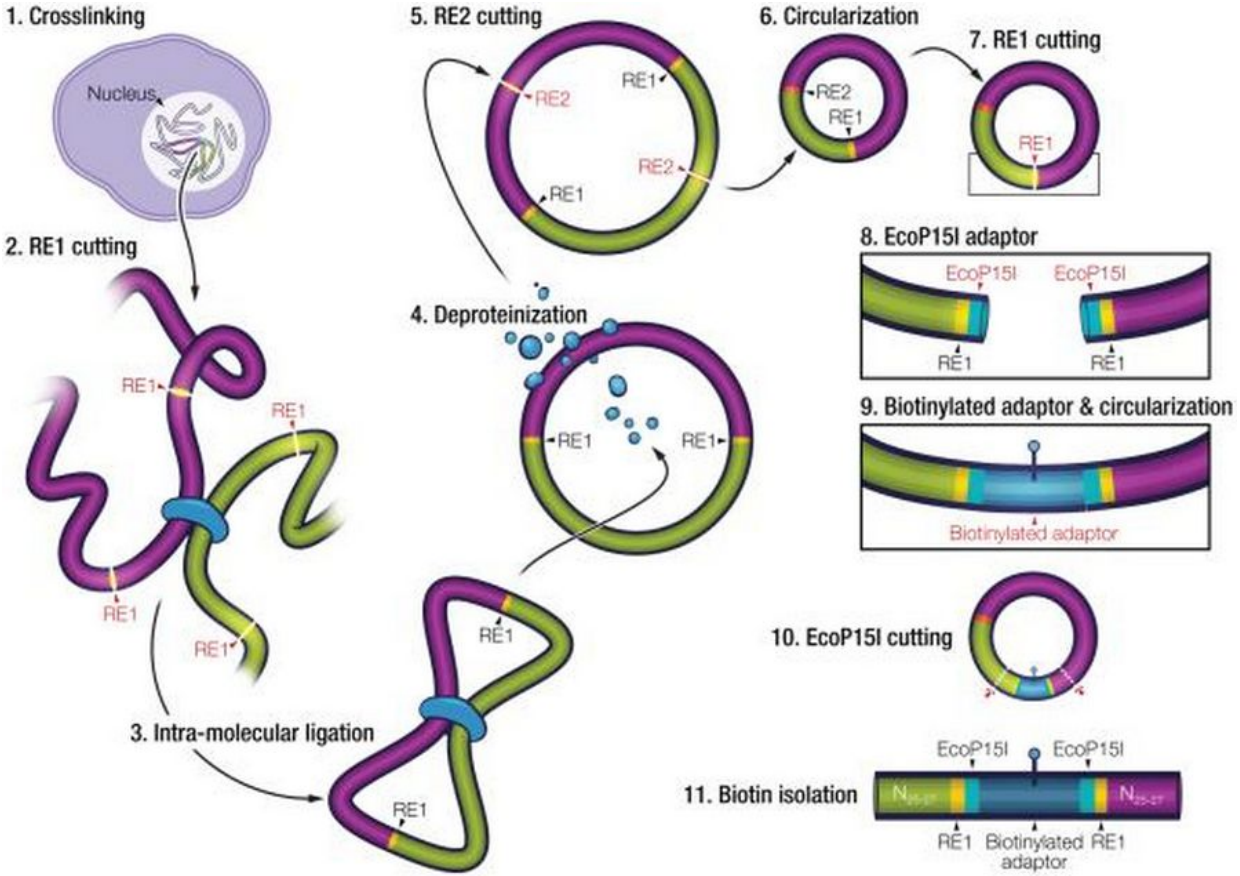
Box 1 | **3C-based methods**

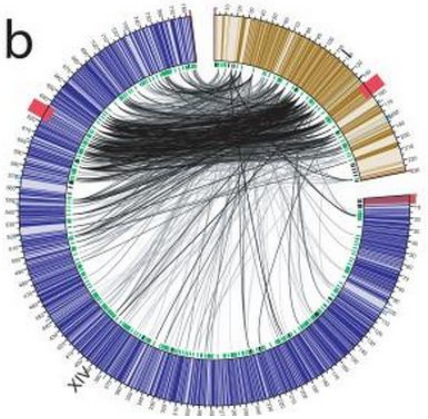
a **3C: converting chromatin interactions into ligation products**

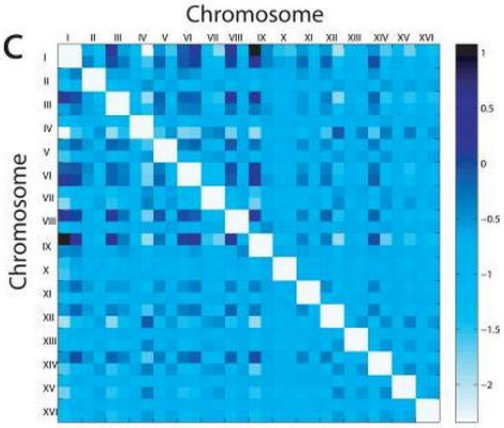
b **Ligation product detection methods**

3C	4C	5C	ChIA–PET	Hi-C
One-by-one All-by-all	One-by-all	Many-by-many	Many-by-many	All-by-all
			• DNA shearing • Immunoprecipitation 	• Biotin labelling of ends • DNA shearing 
PCR or sequencing	Inverse PCR sequencing	Multiplexed LMA sequencing	Sequencing	Sequencing

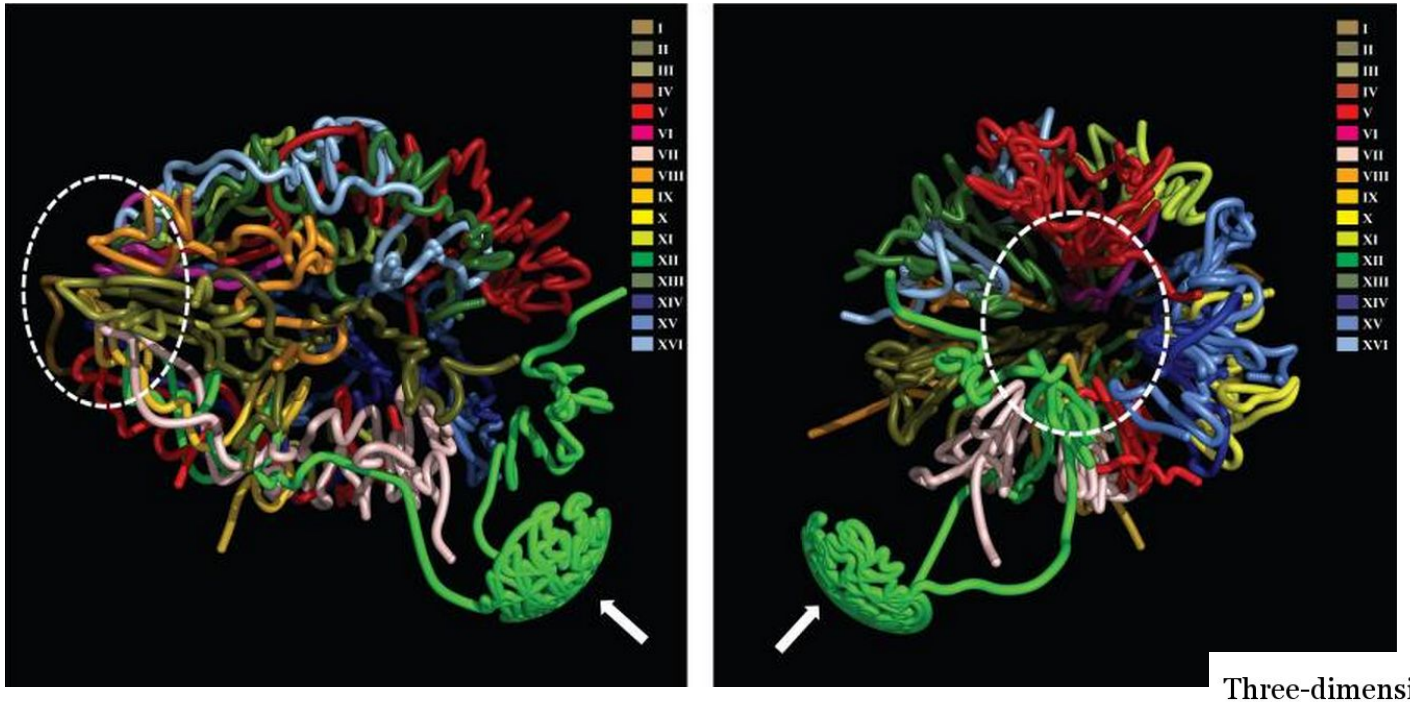
Some application: 3D architecture of the genome (yeast)







Some application: 3D architecture of the genome (yeast)



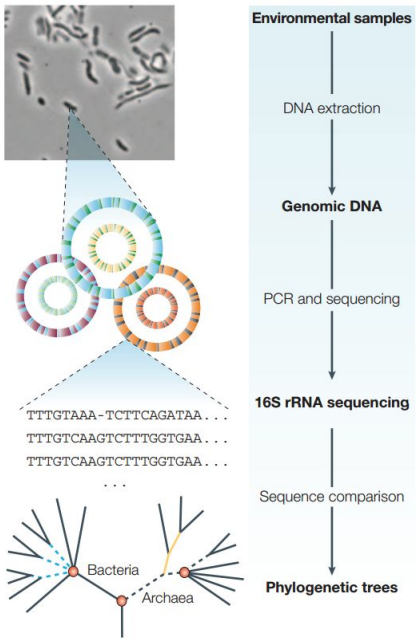
A Three-Dimensional Model of the Yeast Genome

Zhijun Duan,[1,2,*] Mirela Andronescu,[3,*] Kevin Schutz,[4] Sean McIlwain,[3] Yoo Jung Kim,[1,2] Choli Lee,[3] Jay Shendure,[3] Stanley Fields,[2,3,5] C. Anthony Blau,[1,2,3,#] and William S. Noble[3,#]

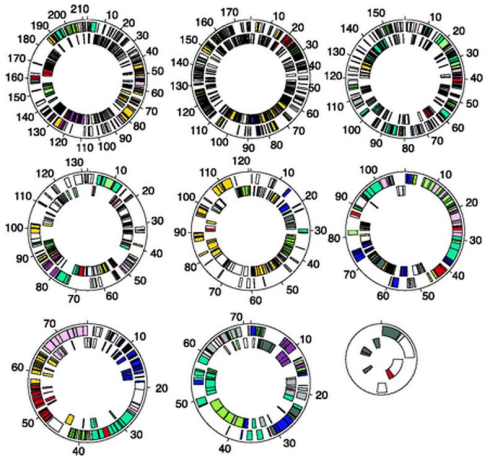
[1]Institute for Stem Cell and Regenerative Medicine, University of Washington
[2]Department of Medicine, University of Washington
[3]Department of Genome Sciences, University of Washington
[4]Graduate Program in Molecular and Cellular Biology, University of Washington
[5]Howard Hughes Medical Institute
*

Three-dimensional model of the yeast genome. Two views representing two different angles are provided. Chromosomes are colored as in Figure 4a (also indicated in the upper right). All chromosomes cluster via centromeres at one pole of the nucleus (the area within the dashed oval), while chromosome XII extends outward toward the nucleolus, which is occupied by rDNA repeats (indicated by the white arrow). After exiting the nucleolus, the remainder of chromosome XII interacts with the long arm of chromosome IV.

Some application of DNA Sequencing: Metagenomics



Circular diagrams of nine complete megaplasmids. Genes encoded in the forward direction are shown in the outer concentric circle; reverse coding genes are shown in the inner concentric circle. The genes have been given role category assignment and colored accordingly: amino acid biosynthesis, violet; biosynthesis of cofactors, prosthetic groups, and carriers, light blue; cell envelope, light green; cellular processes, red; central intermediary metabolism, brown; DNA metabolism, gold; energy metabolism, light gray; fatty acid and phospholipid metabolism, magenta; protein fate and protein synthesis, pink; purines, pyrimidines, nucleosides, and nucleotides, orange; regulatory functions and signal transduction, olive; transcription, dark green; transport and binding proteins, blue-green; genes with no known homology to other proteins and



Metagenomics: DNA sequencing of environmental samples

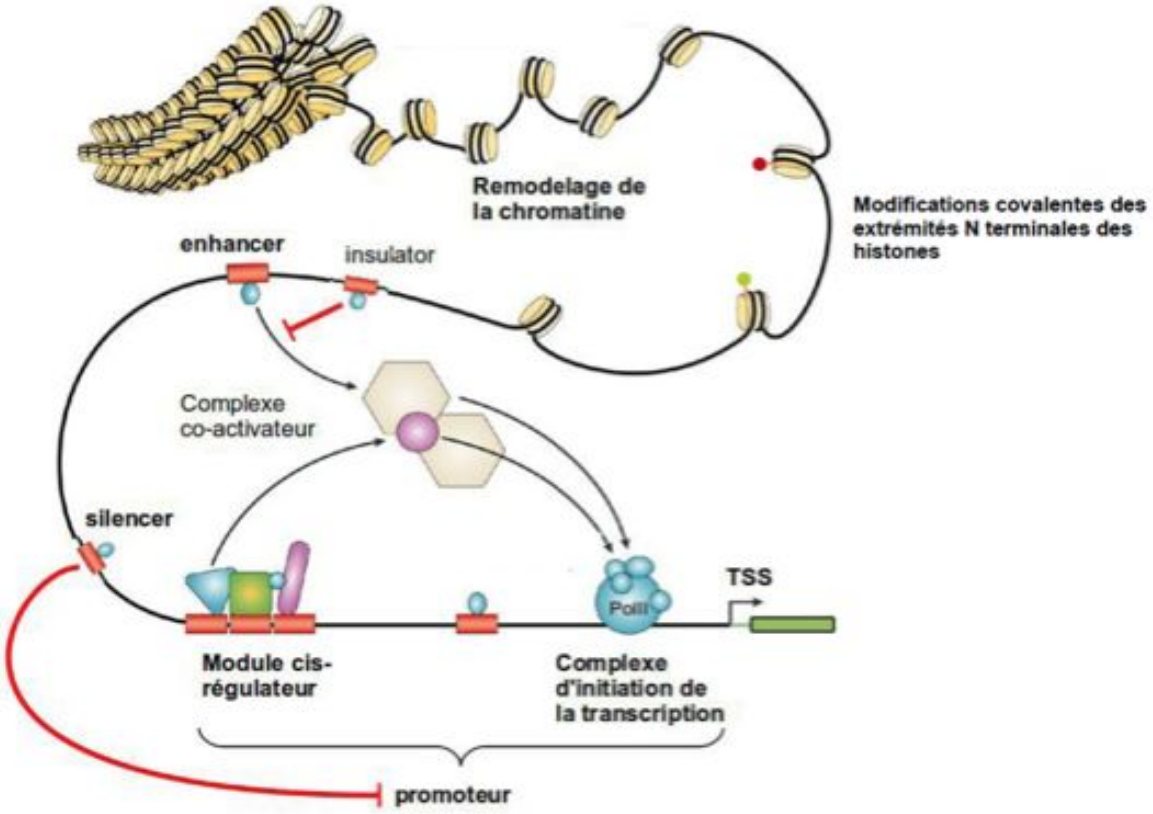
Susannah Green Tringe[1] & Edward M. Rubin[1] About the authors

Science. 2004 Apr 2;304(5667):66-74. Epub 2004 Mar 4.

Environmental genome shotgun sequencing of the Sargasso Sea.

Venter JC[1], Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO.

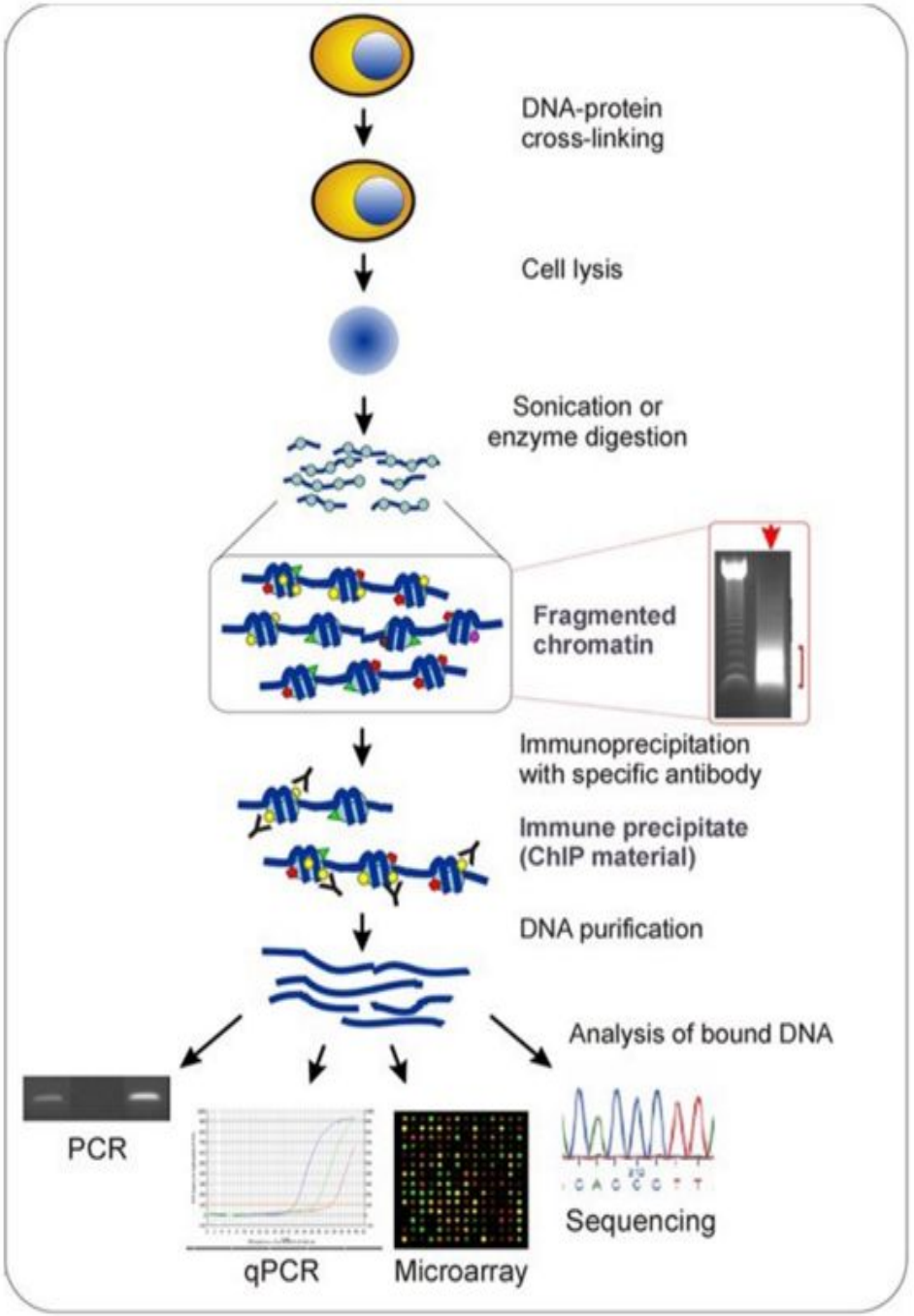
Sequencing to detect regulatory elements



The ENCODE project

- The National **Human Genome Research Institute** (NHGRI) launched a public research consortium in 2003

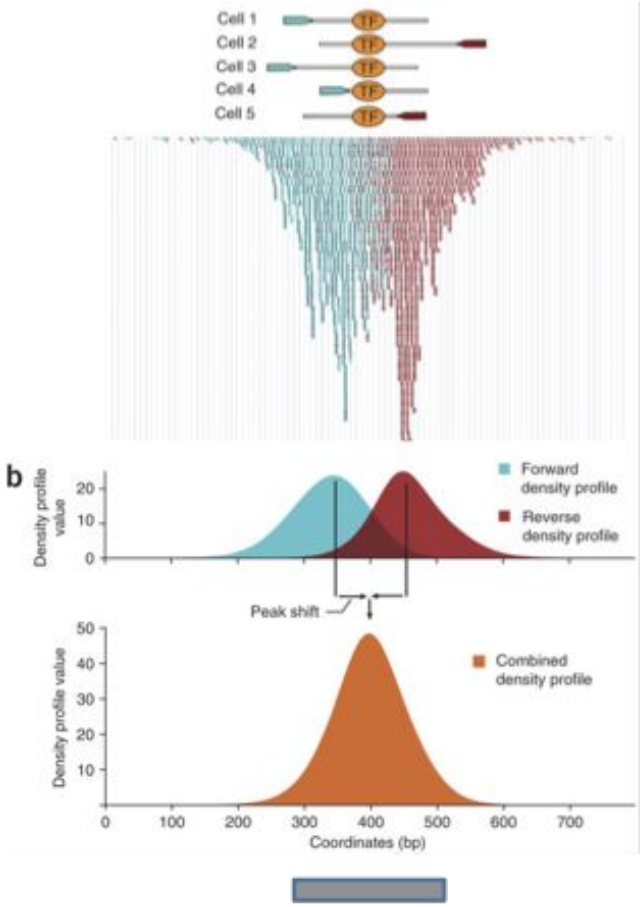
    - **ENCODE**, the **Encyclopedia Of DNA Elements**
        - objective: carry out a project to identify **all functional elements** in the human genome sequence.
        - Lots of experiments rely on ChIP-Seq and RNA-Seq.



ChIP-Seq principle

- Use to analyze
  - Transcription factor location
  - Histone modification across genome

ChIP-Seq analysis (in brief…)

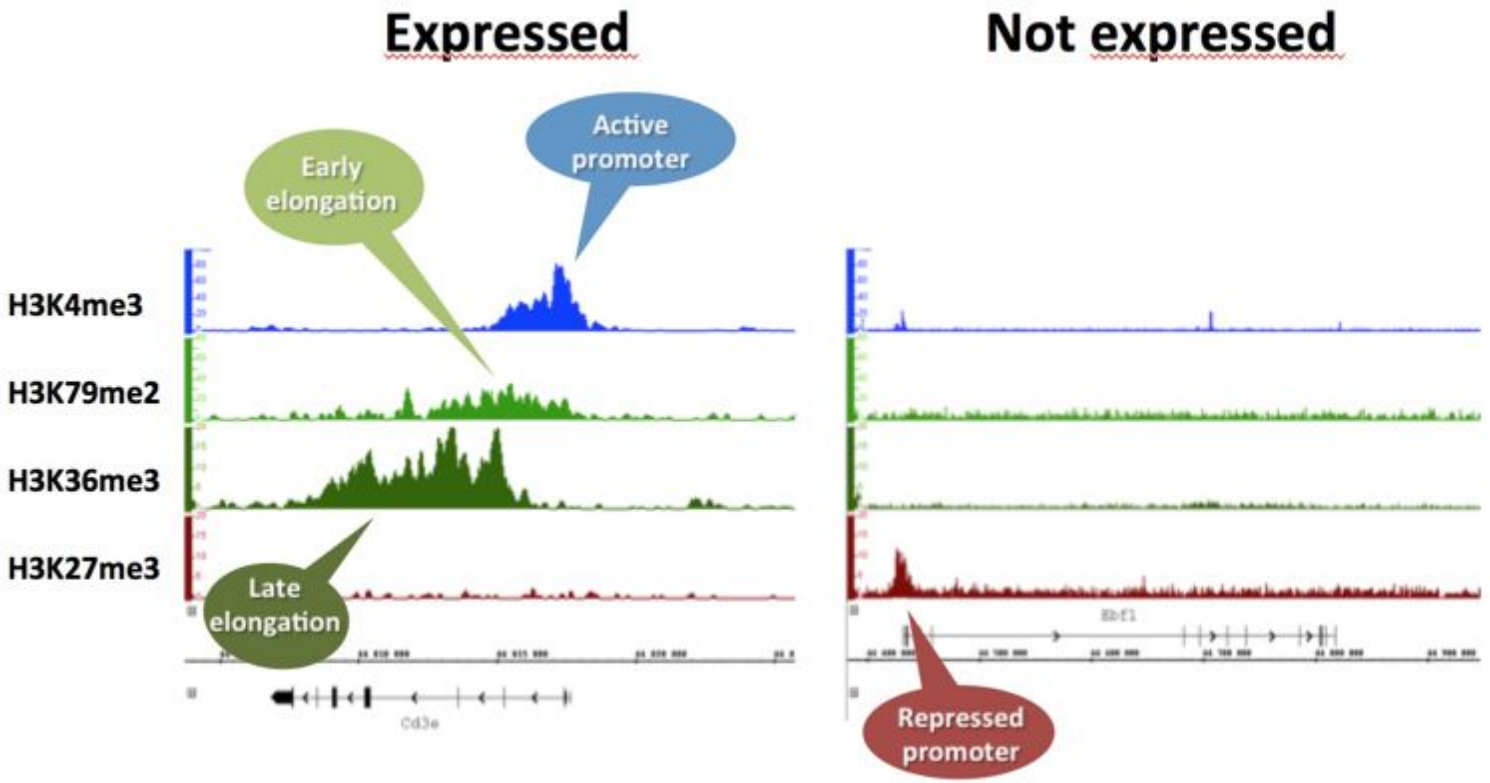


Aligned reads

Binding profile

Binding Peak

Epigenetic modification on histones

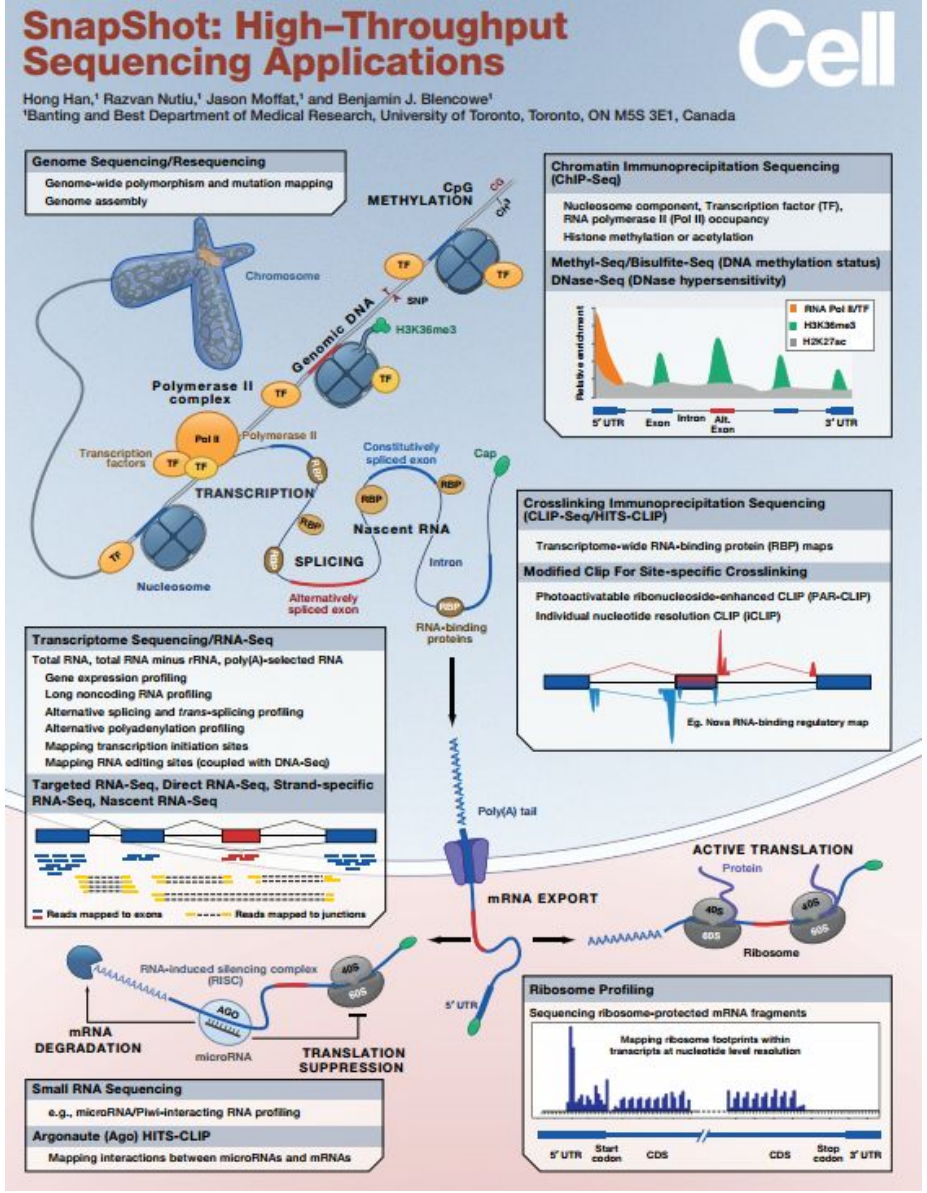


Application of ChIP-Seq

- Defining transcription factor location
  - Define precise motif
    - peak sequence analysis
    - Define co-factor through motif analysis
  - Differential analysis : e.g normal vs tumor
    - lost/acquired regulatory site in tumors
  - Impact of mutation on binding sites
  - ...

Application of ChIP-Seq

- Define epigenetic landscape
  - Active / inactive regions
    - Differential expression
      - Impact of mutation on transcriptional status
  - Essential to detect proximal or distal regulatory regions
    - Help to define promoter regions (H3K4me3)
    - Help to define enhancer regions (e.g H3K27ac)
    - Super-enhancer (large regions with H3K27ac)
      - Frequently associated with cell identity
      - SNP falling in these regions are more likely to be



Nucleosome-positioning, Ribosome profiling, ...

Transcriptome analysis

- Tentative definition
  - 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

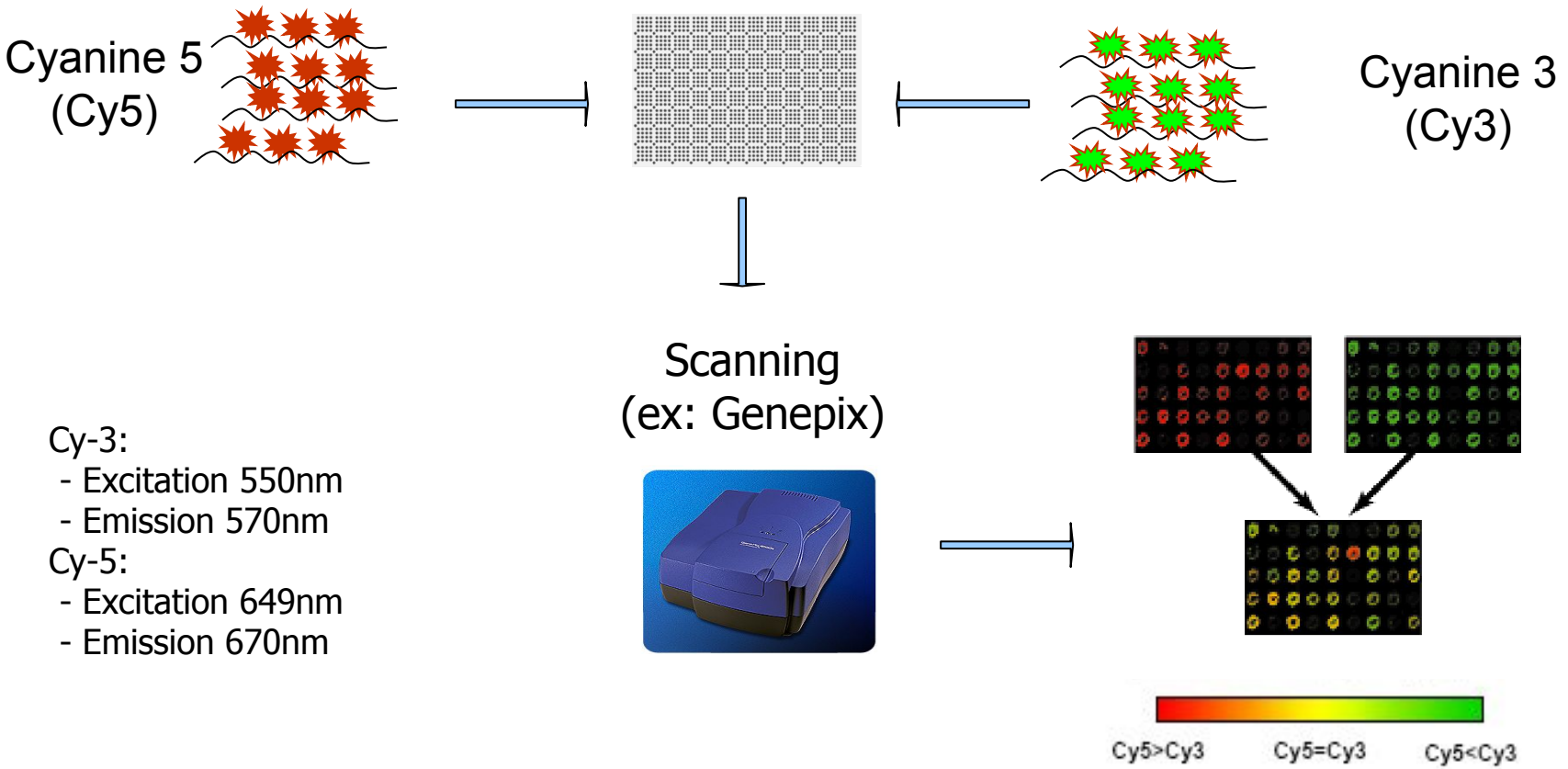
Some players of the RNA world

- 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

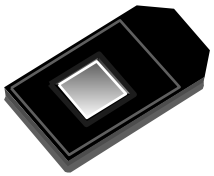
Science. 1995 Oct 20;270(5235):467-70.

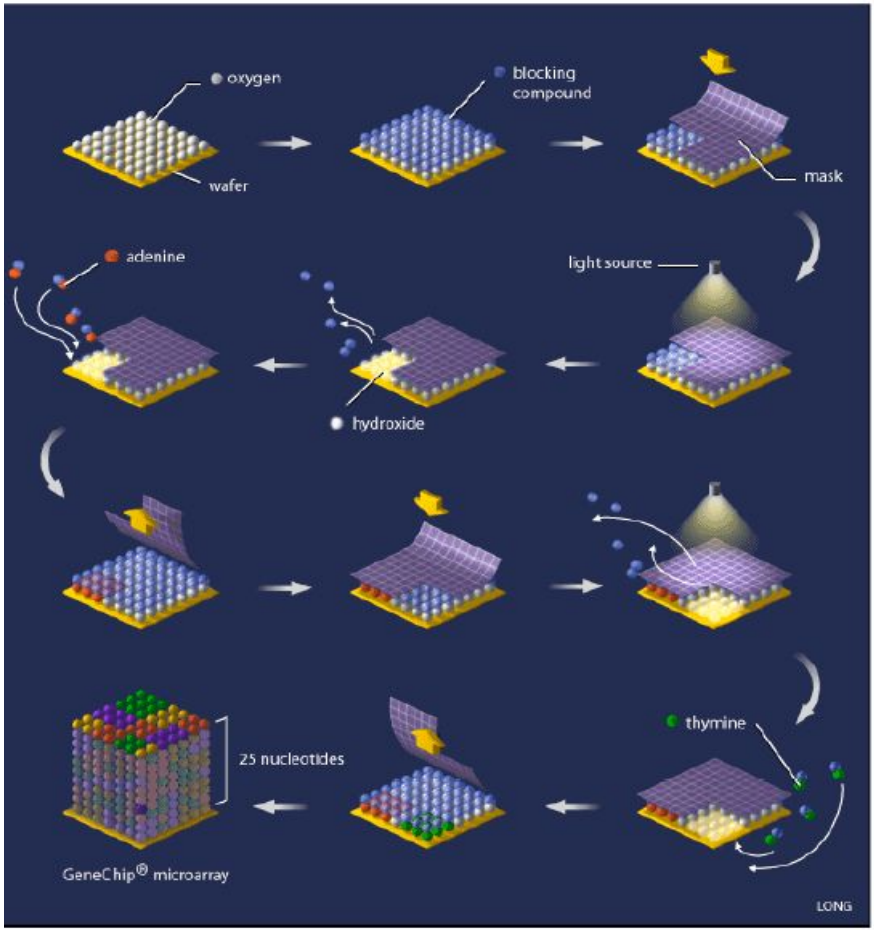
Quantitative monitoring of gene expression patterns with a complementary DNA microarray.

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

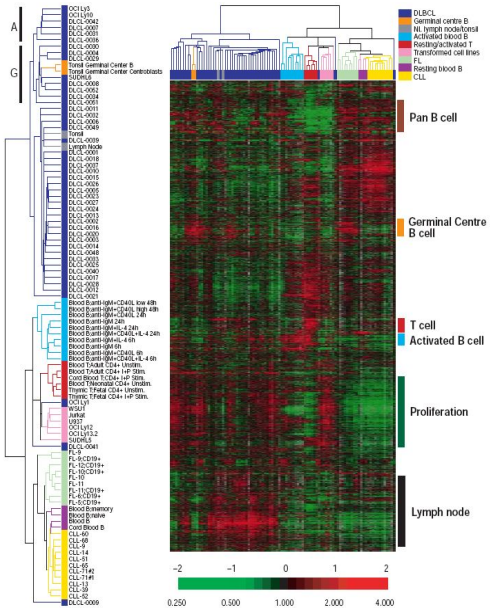
Transcriptome still the old school

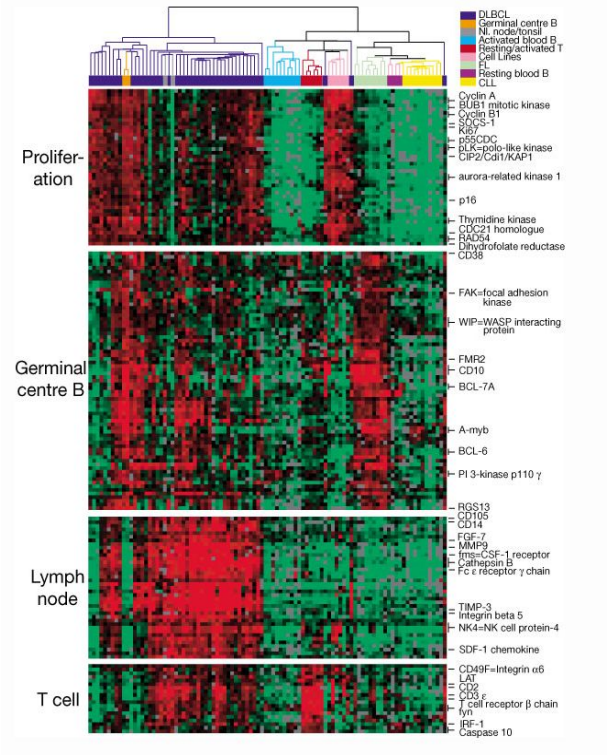
- Affymetrix biochip
- Principle:
  - In situ synthesis of oligonucleotides
  - Features
    - Cells: 24µm x 24µm
    - ~10^7 oligos per cell
    - ~ 4.10^5-$1,5.10^6$ probes

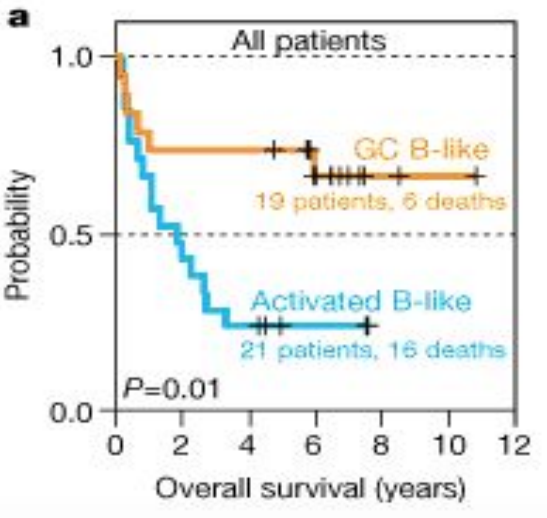




Some pioneering works: "Molecular portraits of tumors"







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

GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
PENTOSE PHOSPHATE CYCLE
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
GLYCOLYSIS
ACETYL-COA BIOSYNTHESIS

UBIQUINONE BIOSYNTHESIS
MRNA SPLICING
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
RESPIRATION
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
RESPIRATION
RESPIRATION
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
PROTEIN SYNTHESIS
OXIDATIVE PHOS
MITO DNA MAINT

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.

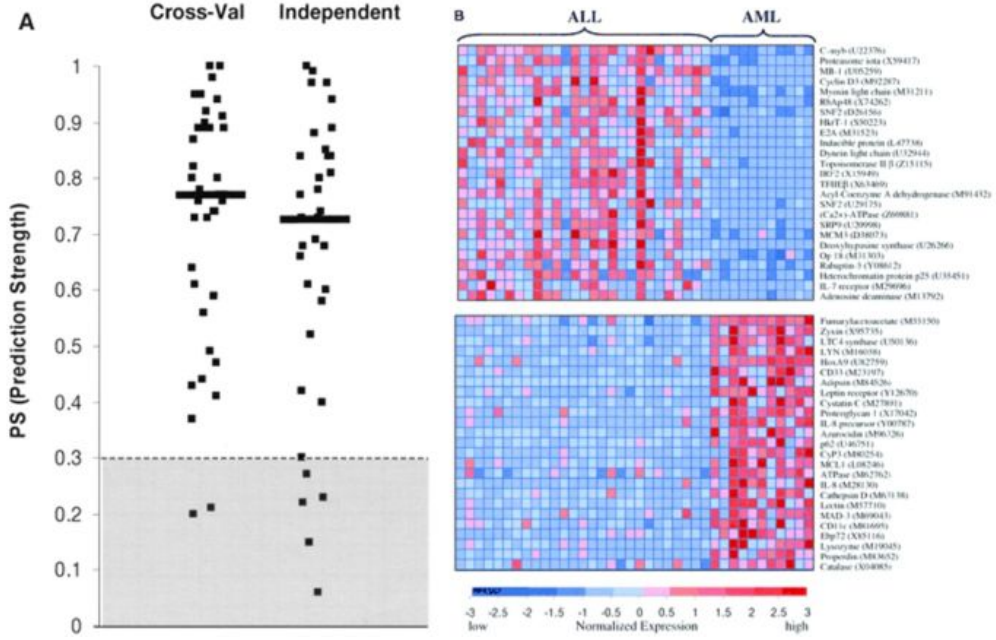
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.

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

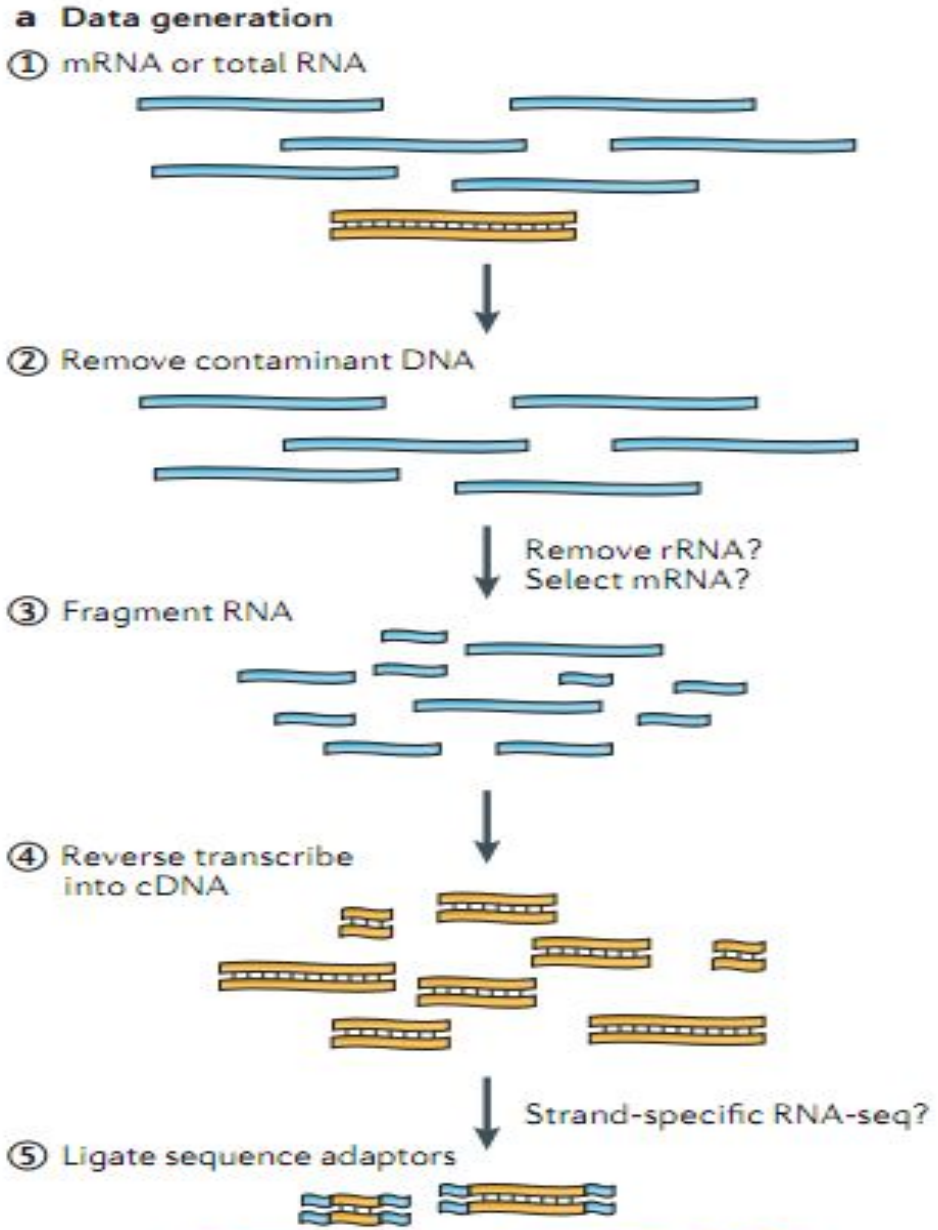
News

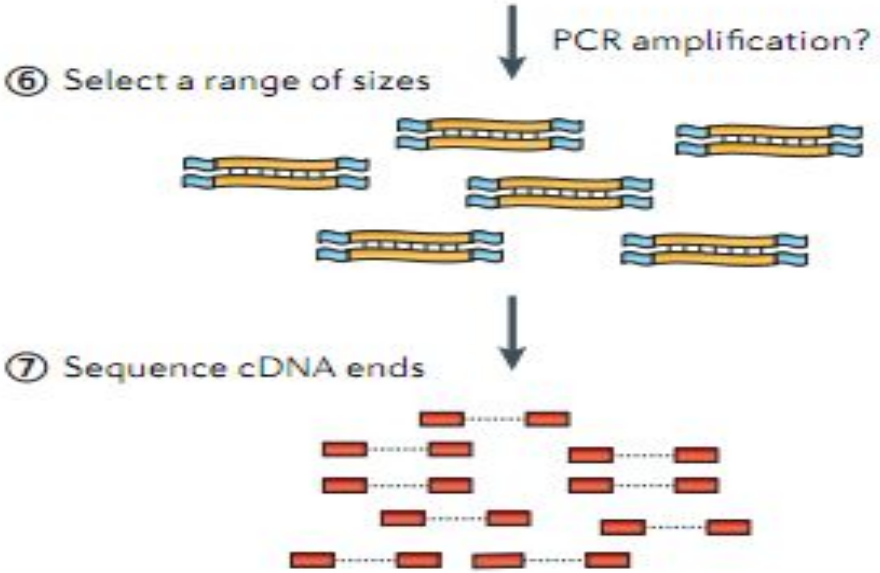
The death of microarrays?

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

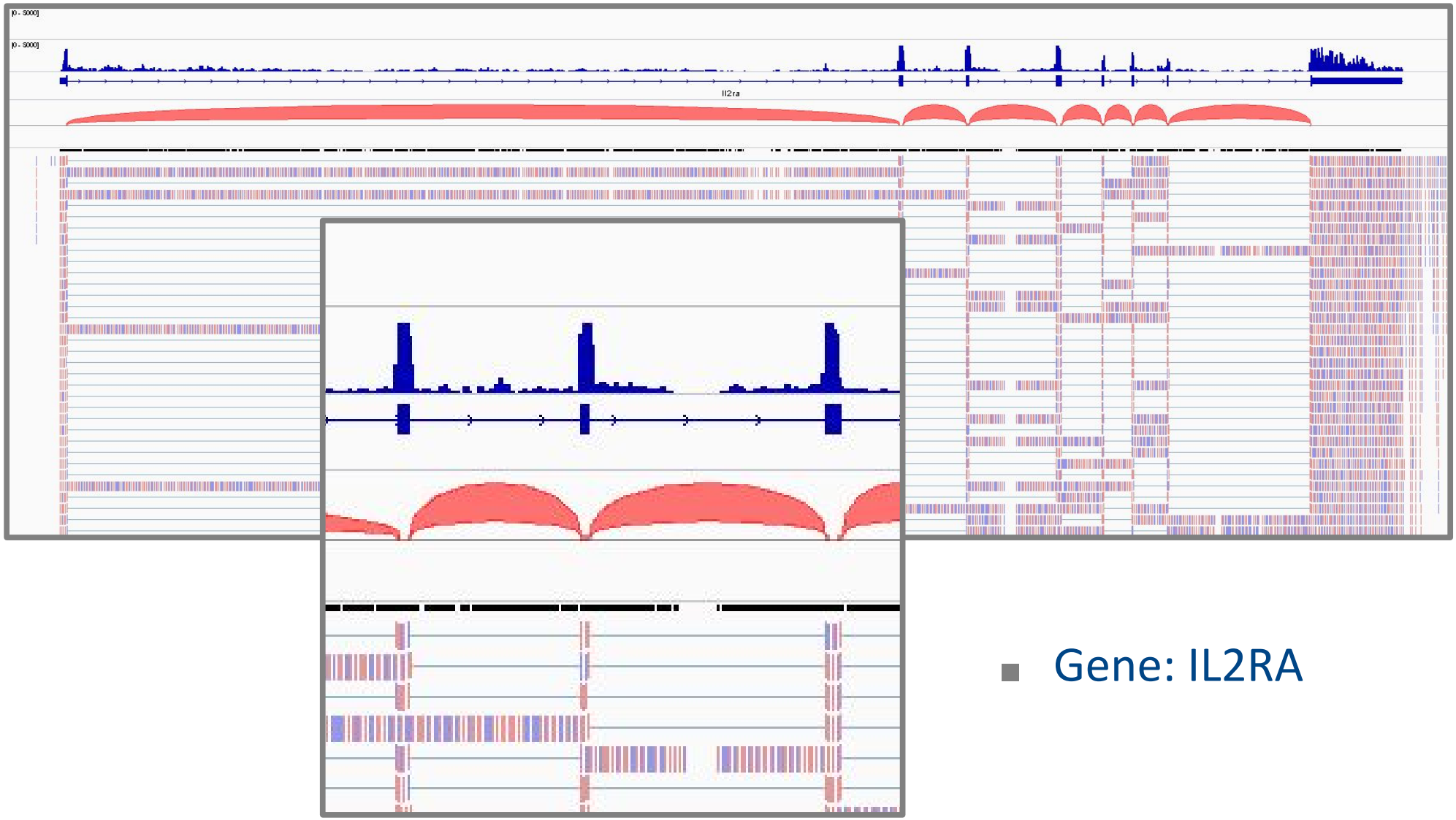
ARTICLE SERIES: Study designs

Next-generation transcriptome assembly

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

Aligned reads on mouse genome (mm9 version)

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



- Gene: IL2RA

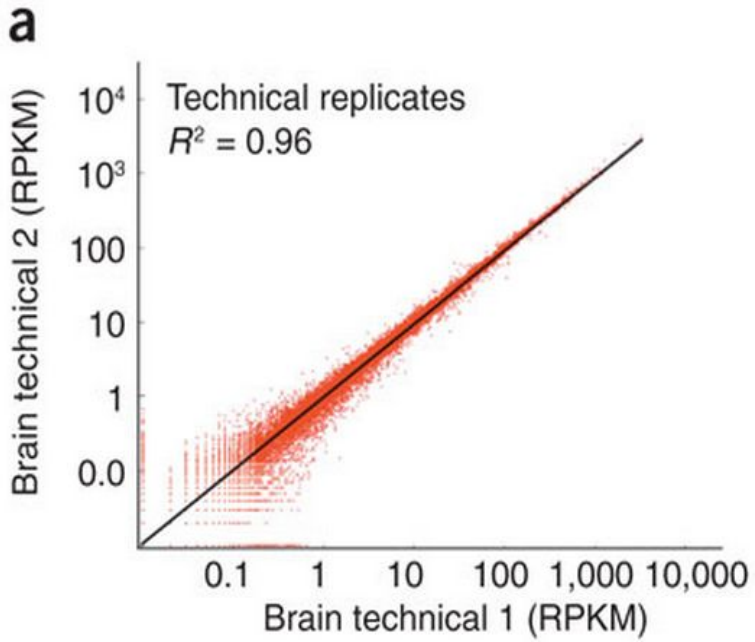
Microarrays vs RNA-Seq

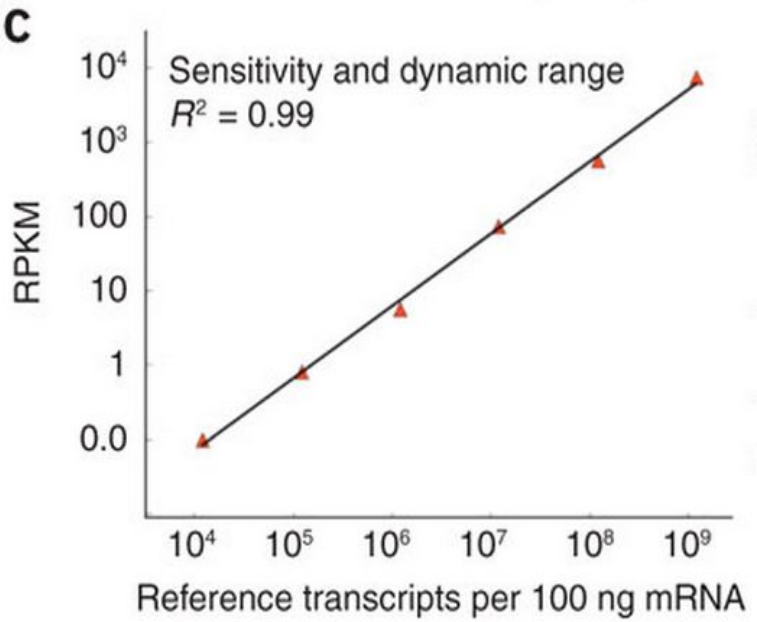
- RNA-seq
  - Counting
  - Absolute abundance of transcripts
  - All transcripts are present and can be analyzed
    - mRNA / ncRNA (snoRNA, linc/lncRNA, eRNA, miRNA,...)
  - Several types of analyses
    - Gene discovery
    - Gene structure (new transcript models)
    - Differential expression
   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

Microarrays vs RNA-Seq

- 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; $R^2 = 0.96$.

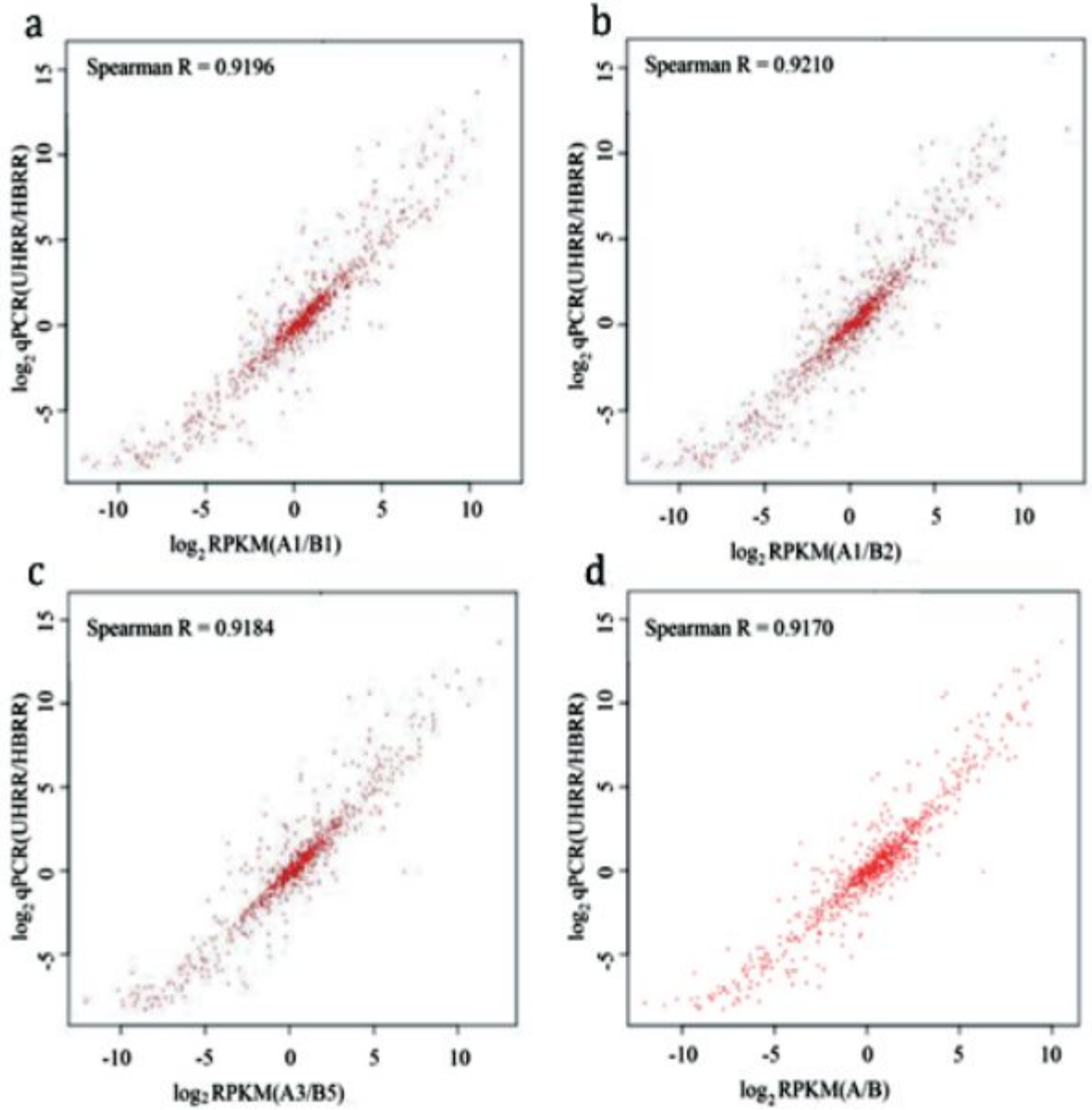
(c) Six in vitro–synthesized reference transcripts of lengths 0.3–10 kb were added to the liver RNA sample (1.2 104 to 1.2 109 transcripts per sample; R2 > 0.99).

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.pdf

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

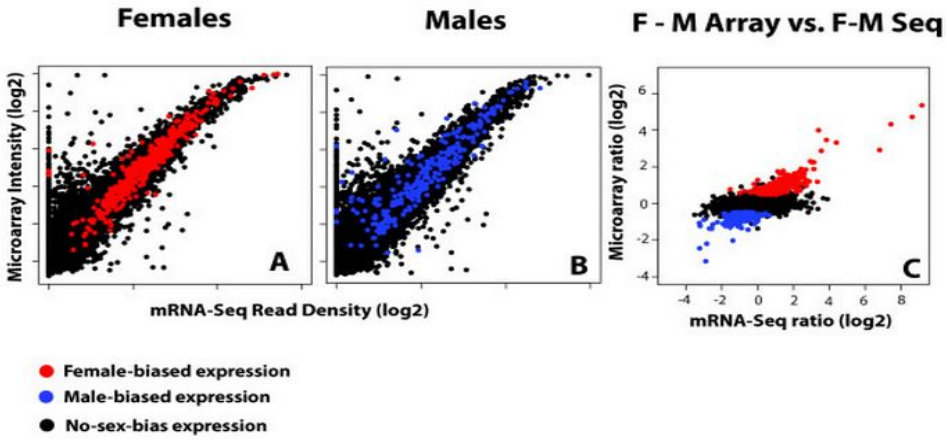
Highly accessed Open Access

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

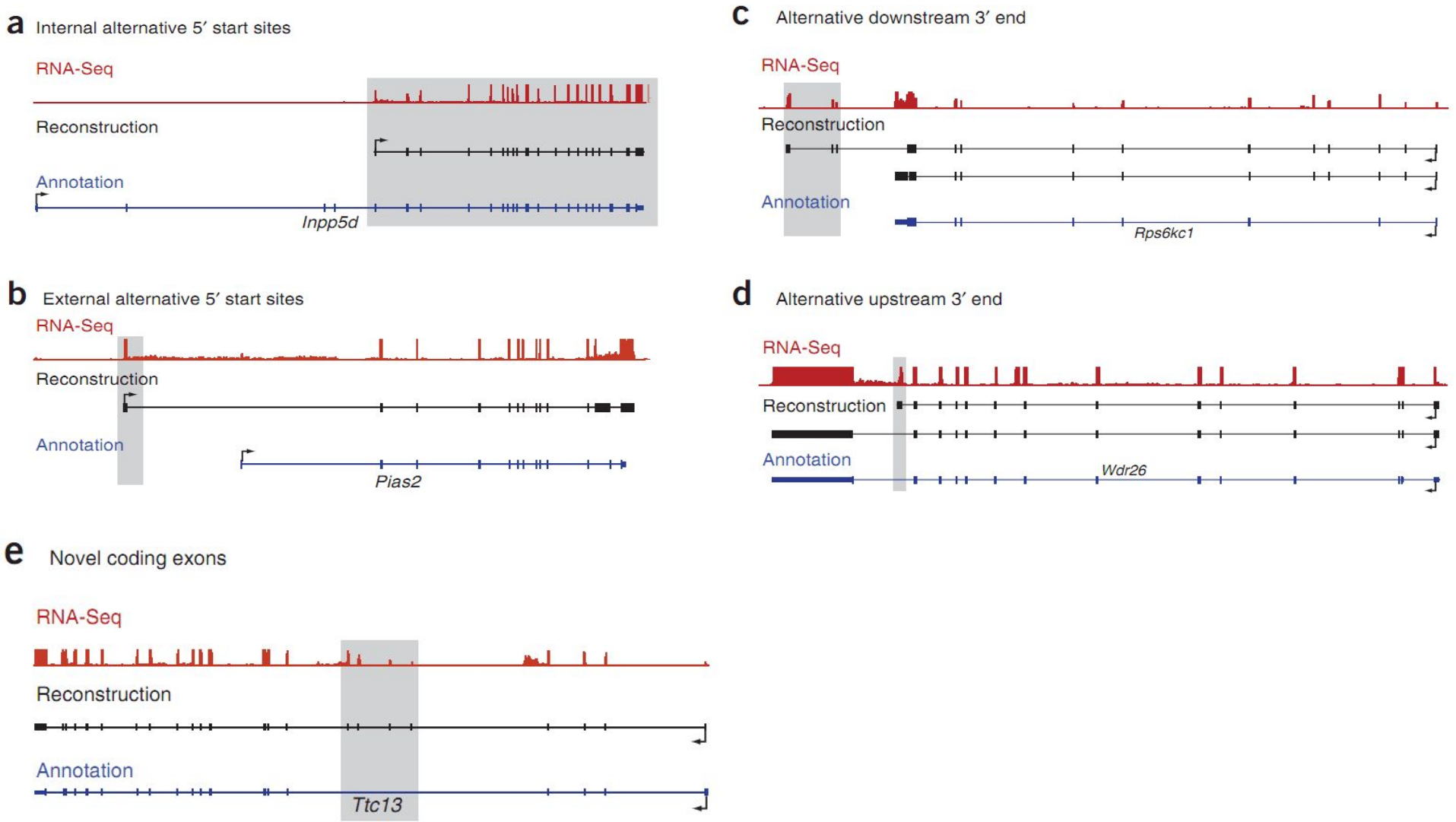
author email corresponding author email

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

Microarrays vs RNA-Seq



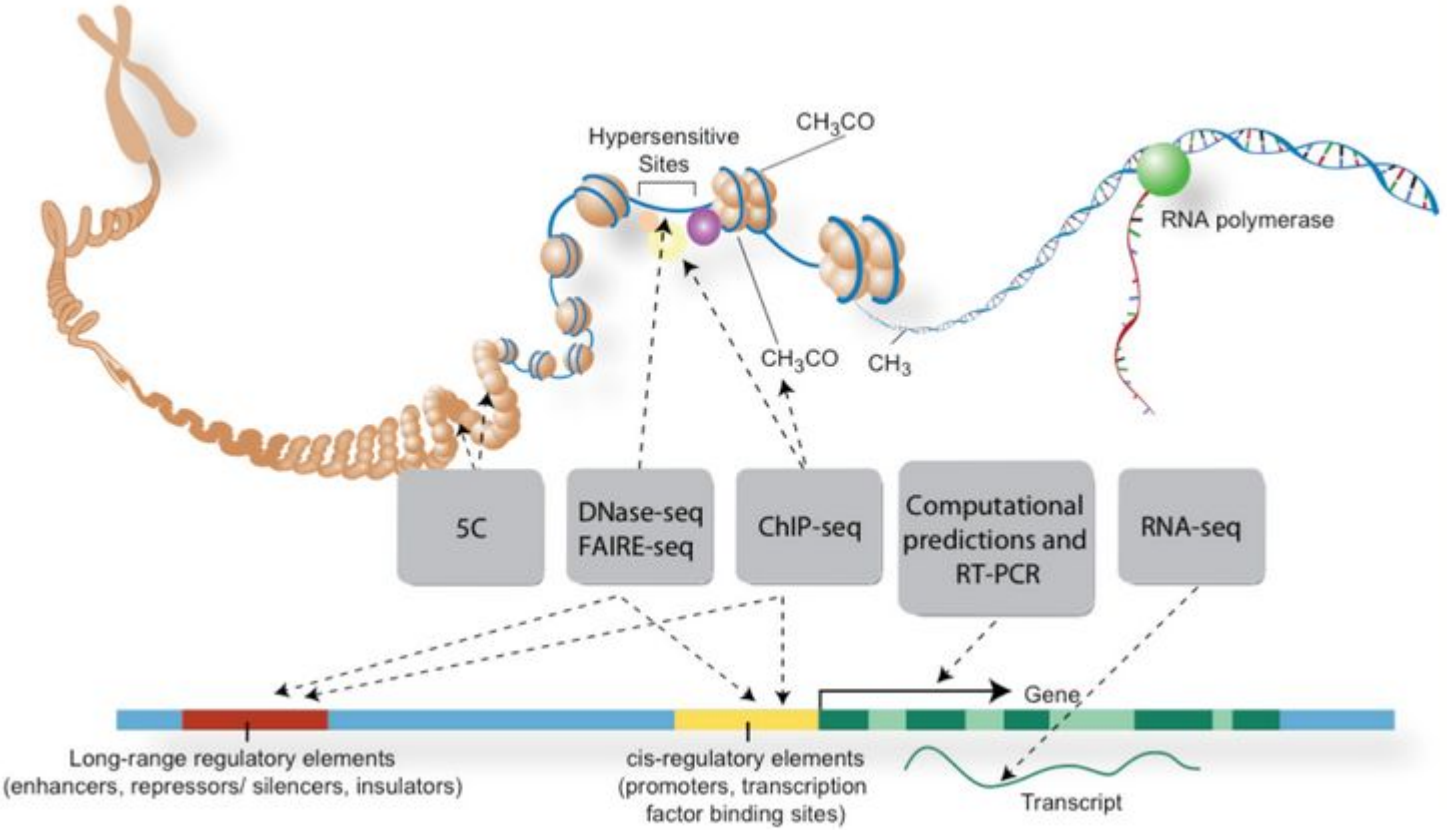
Nat Biotechnol. 2010 May;28(5):503-10. Epub 2010 May 2.

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.

What can we learn from RNA-Seq ?

- 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

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.

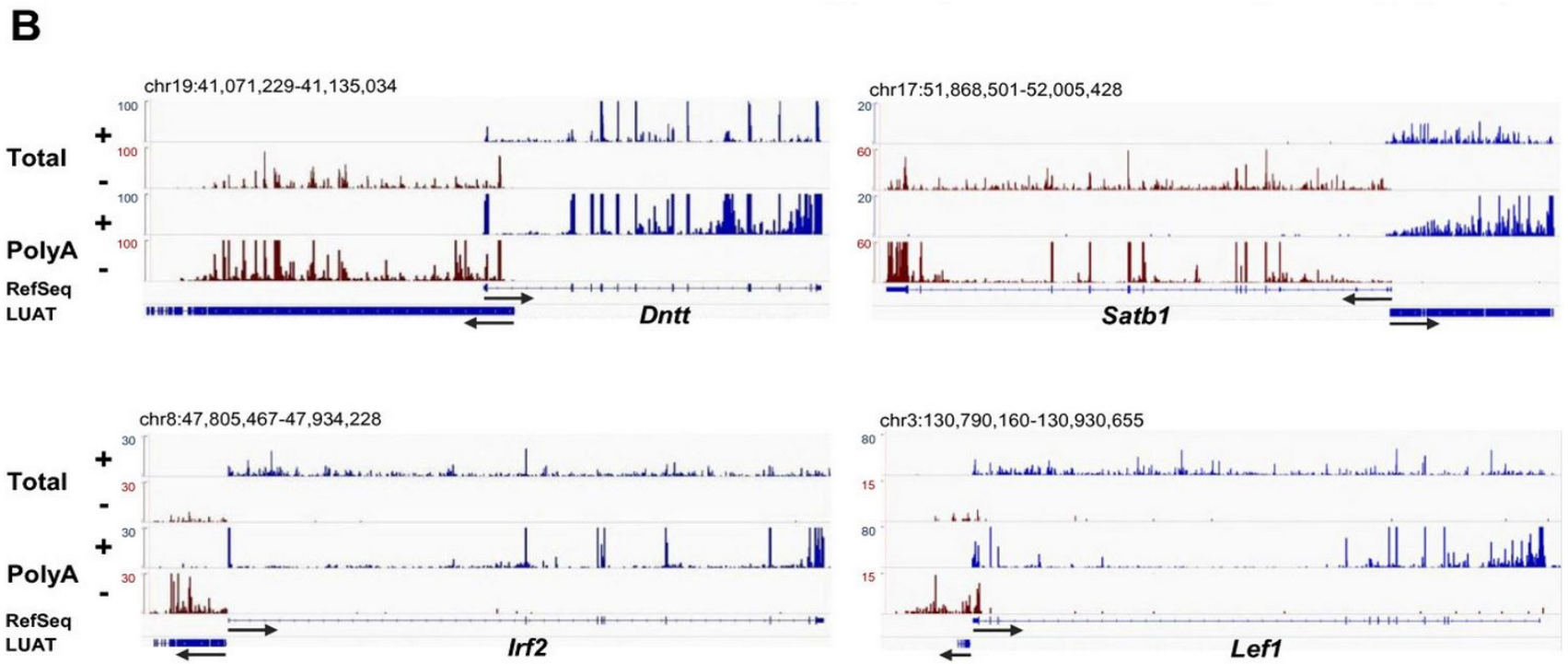
Version 21 (June 2014 freeze, GRCh38) - Ensembl 77

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



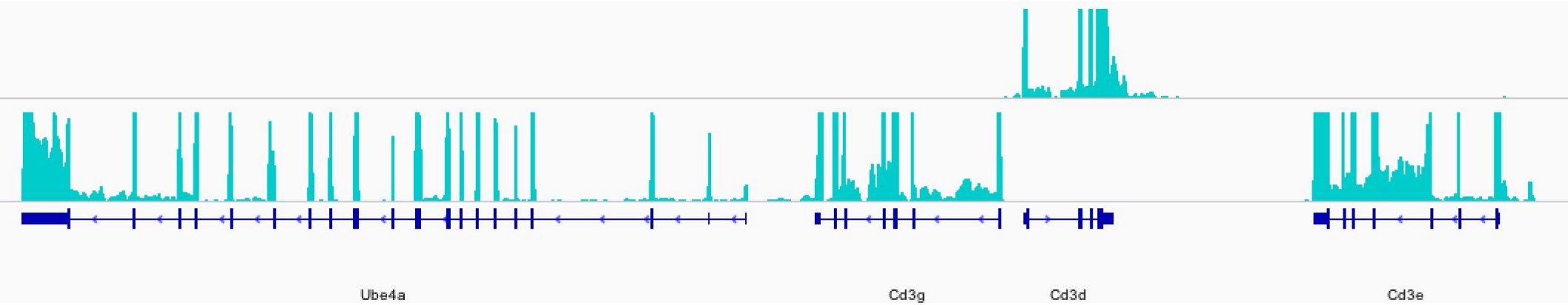
RNA-Seq: protocol variations

- 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

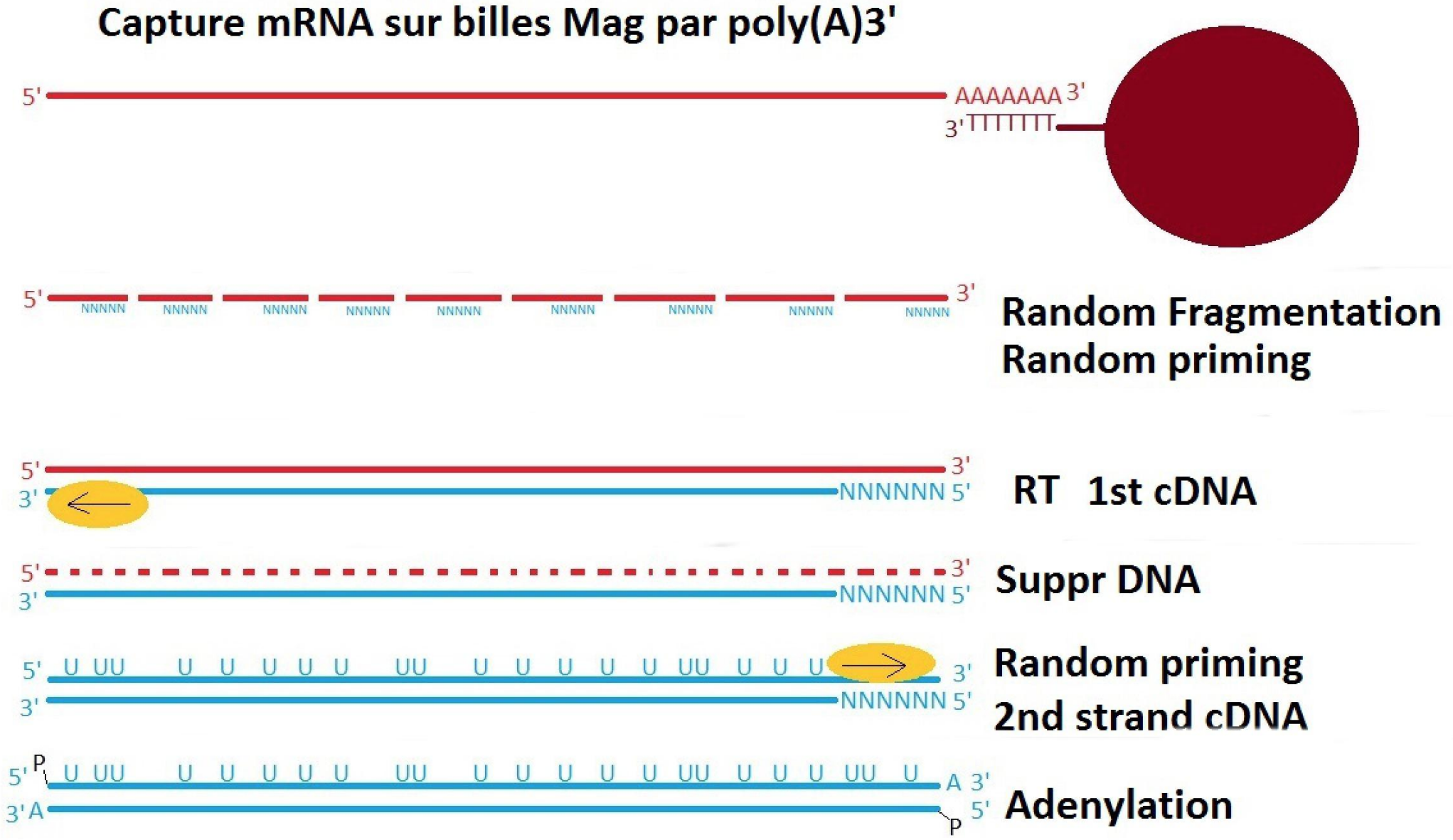
http://www.bioconductor.org/help/course-materials/2009/EMBLJune09/Talks/RNAseq-Paul.pdf

Strand specific RNA-Seq

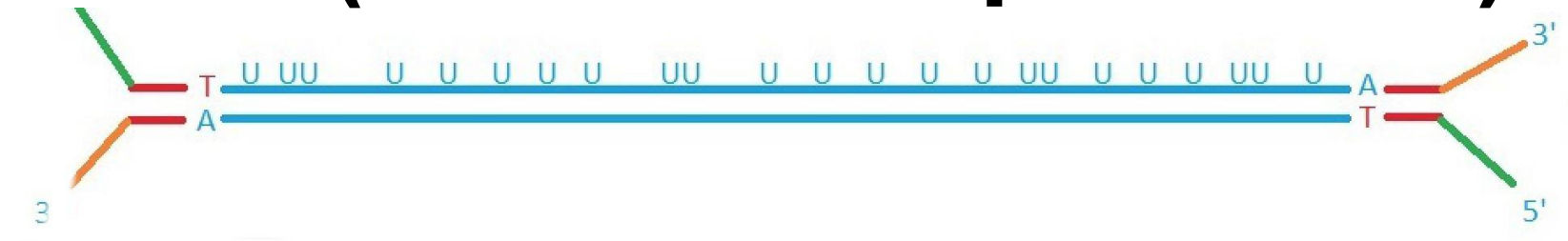
- Most kits are now strand-specific
  - Better estimation of gene expression level.
  - Better reconstruction of transcript model.



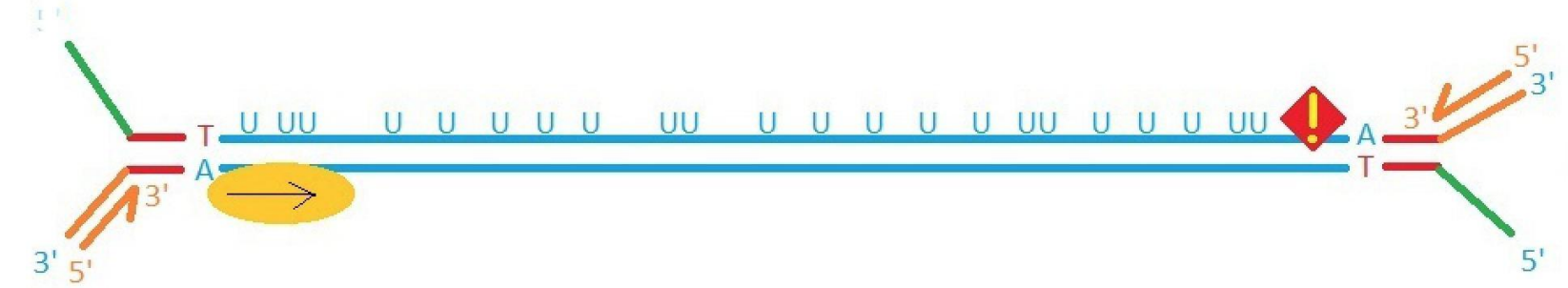
Strand specifity protocol (tentative explanation)



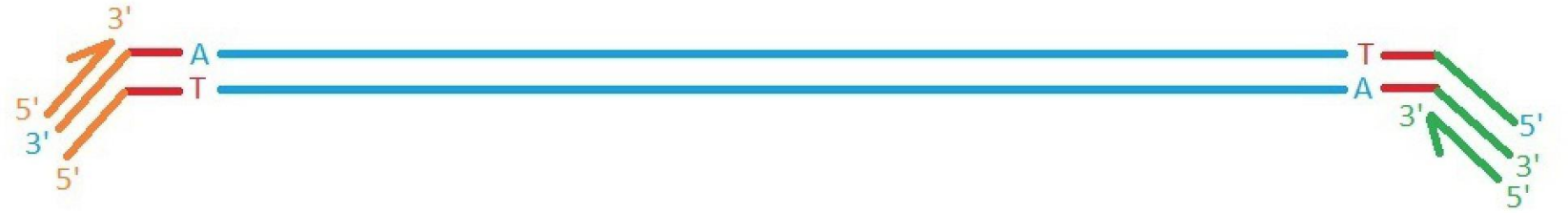
Strand specifity protocol (tentative explanation)



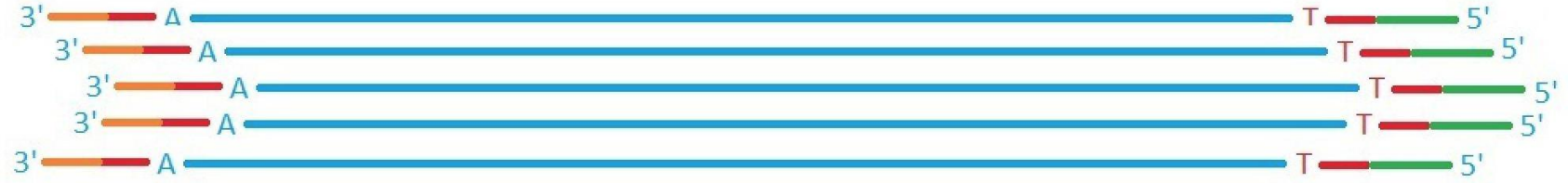
Ligation Adapt.



PCR



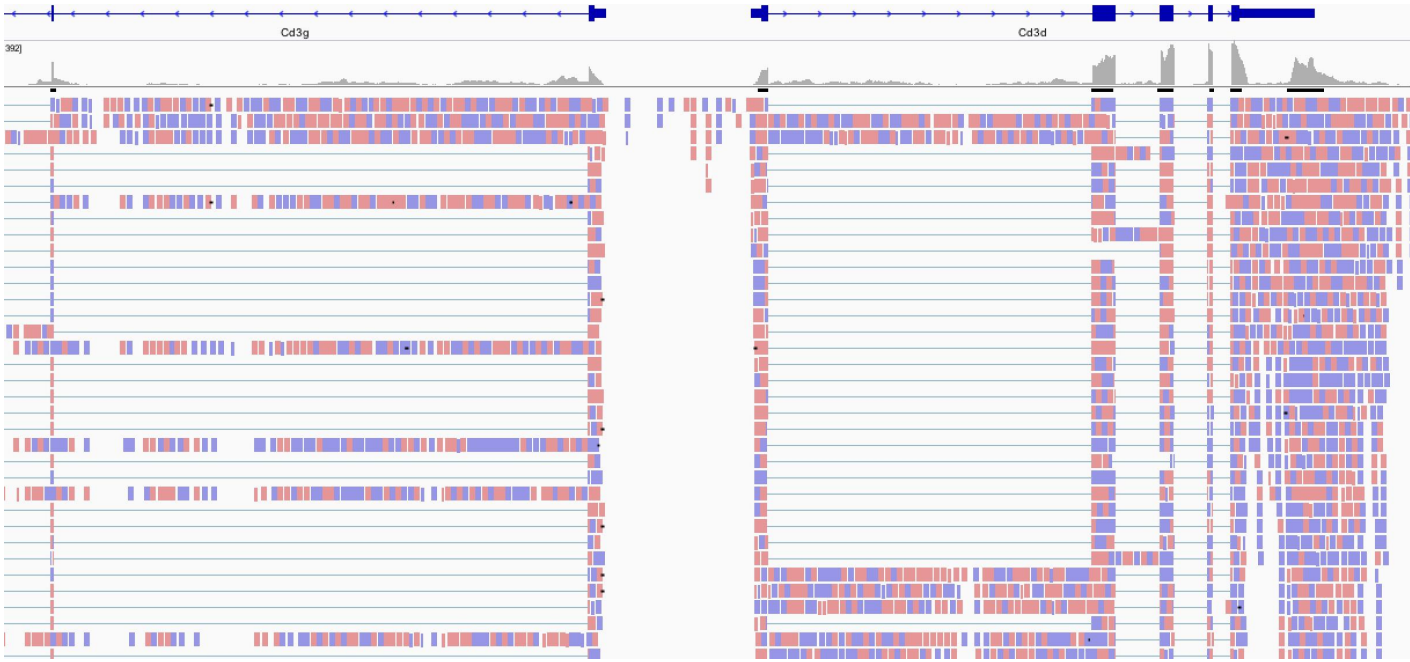
1er produit



Produits finaux "Utile"

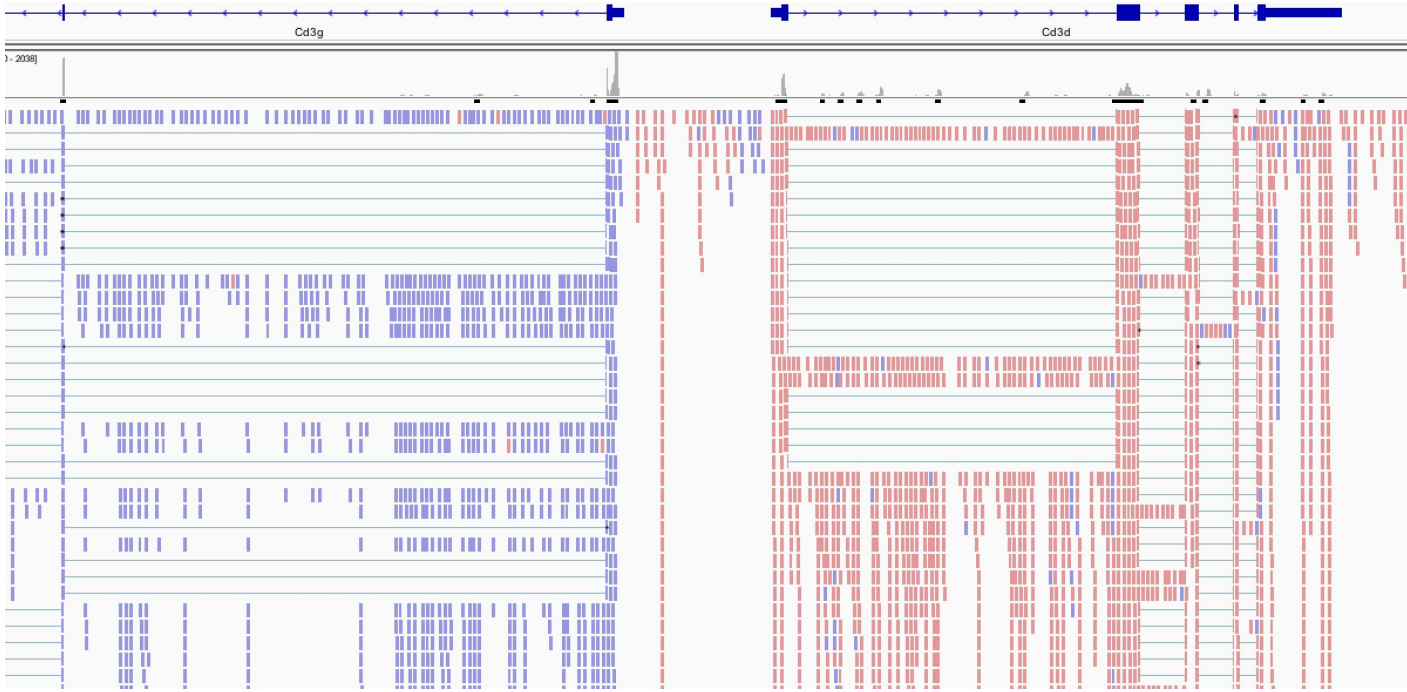
"Inutile"

Unstranded (single-end)

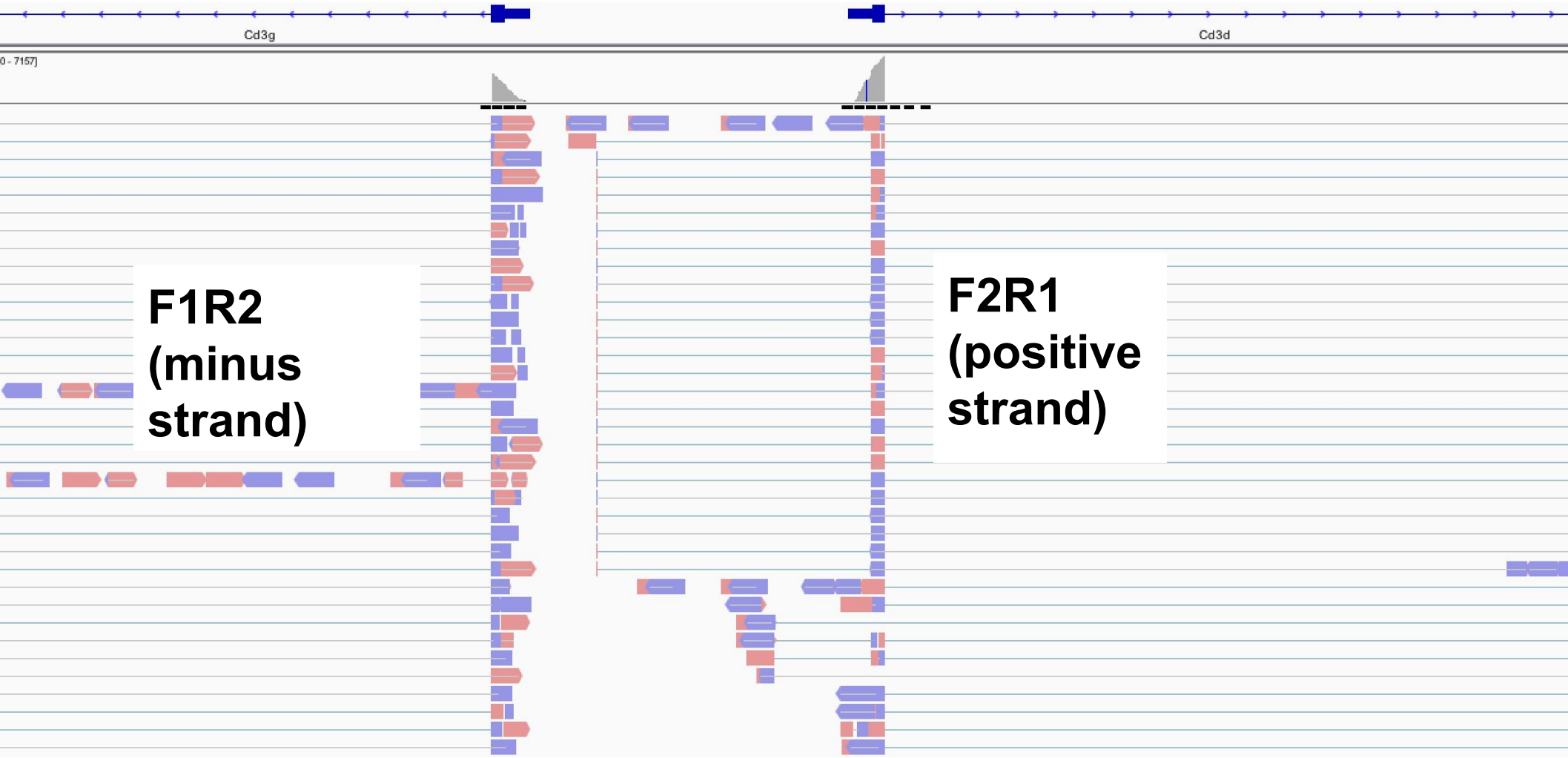


Forward (Red)
Reverse (Blue)

Stranded (single-end)



Stranded (paired, forward-reverse)



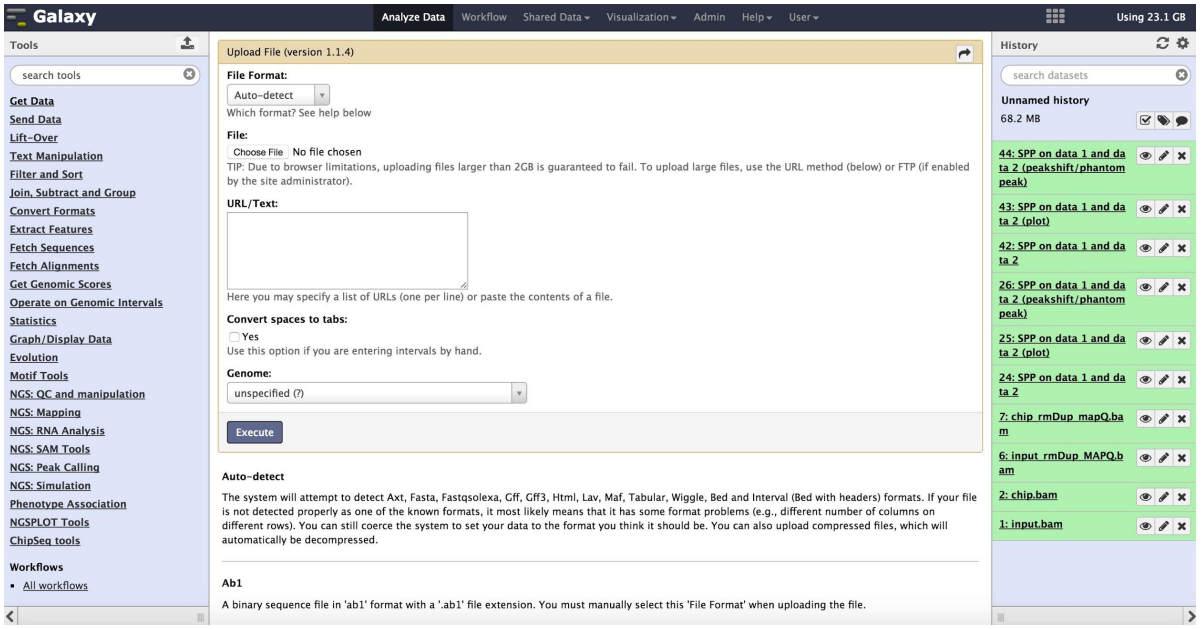
Our dataset

- P5424 Thymocytes
  - Control condition (DMSO, “DM”)
  - Treated (PMA/Ionomycin, “PI”)
- Paired-end strand-specific RNA-Seq
  - NextSeq 500
  - Full dataset contains 50.106 sequenced **fragments**
  - We will work with a subset of reads (that sould mostly map to the chr18)

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

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



Raw data: the fastq file format

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

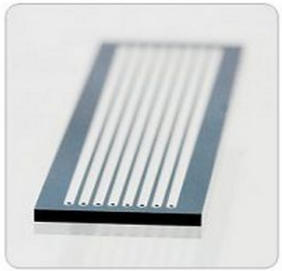
```
@QSEQ32.249996 HWUSI-EAS1691:3:1:17036:13000#0/1 PF=0 length=36
GGGGGTCATCATCATTTGATCTGGGAAAGGCTACTG
+
=.+5:<<<<>AA?0A>;A*A##################
@QSEQ32.249997 HWUSI-EAS1691:3:1:17257:12994#0/1 PF=1 length=36
TGTACAACAACAACCTGAATGGCATACTGGTTGCTG
+
DDDD<BDBDB??BB*DD:D##################
```

Illumina sequence identifiers

- Sequences from the Illumina software use a systematic identifier:

```
@SRR038538.sra.2 HWI-EAS434:4:1:1:1701 length=36
NAATCGGAAATTTTATTTGTTCAGTACACCAAATAG
+SRR038538.sra.2 HWI-EAS434:4:1:1:1701 length=36
!0<<;:::<<<<<<<<<<<<<<;;;<<<<<<<<<;76
```

HWI-EAS434	Unique instrument name
4	Flowcell lane
1	Tile number within the flow cell
1	'x'-coordinate of the cluster within the tile
1701	'y'-coordinate
#0	Index number for a multiplexed sample (opt.)
/1	/1 or /2 for paired-end and mate-pair sequencing (opt.)



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)

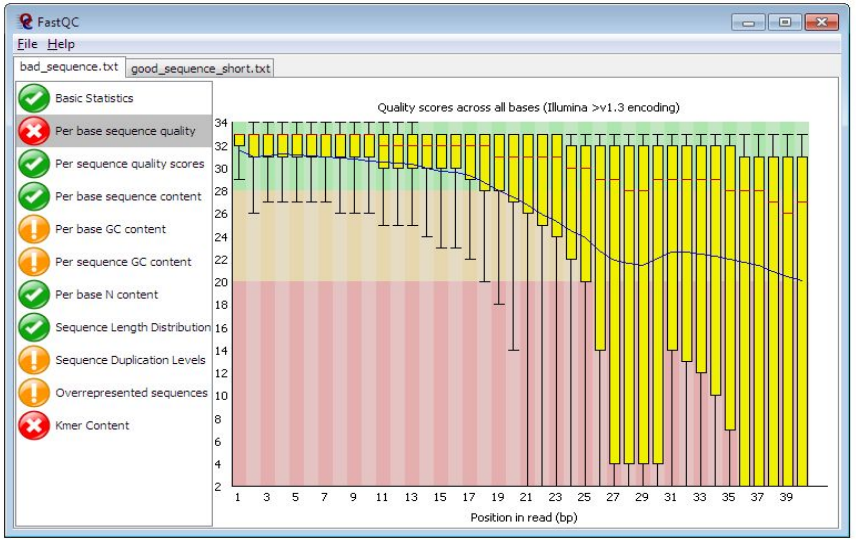
ASCII 33

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

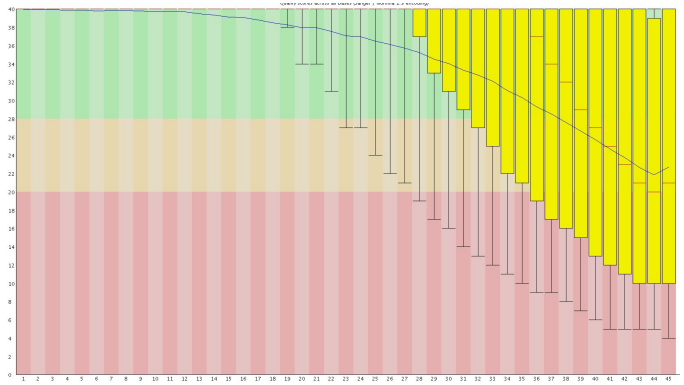
Quality control for high throughput sequence data

- 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

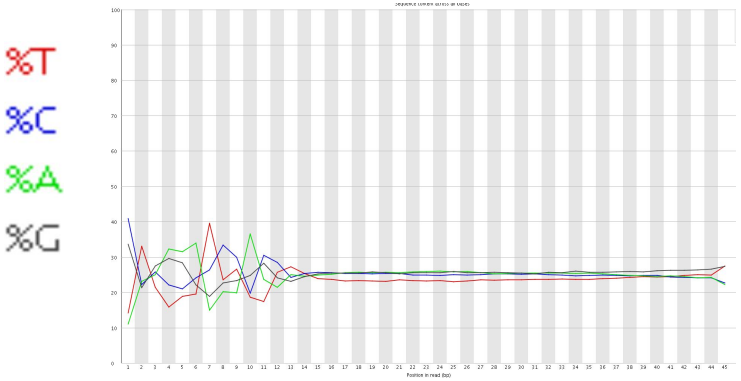


Quality control with FastQC

Quality

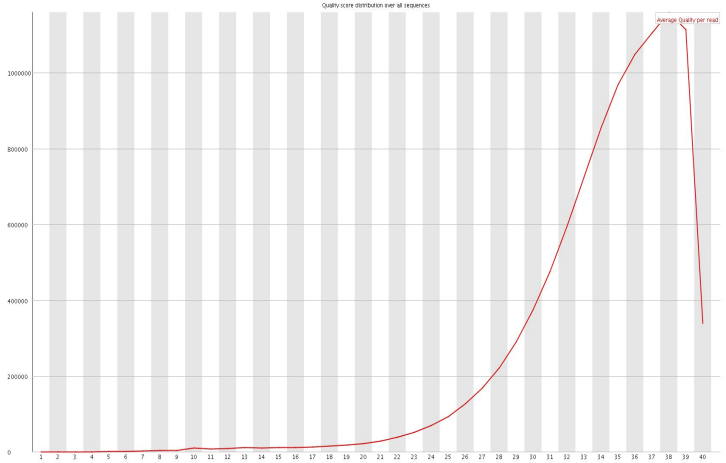


Position in read



Position in read

Nb Reads



Mean Phred Score

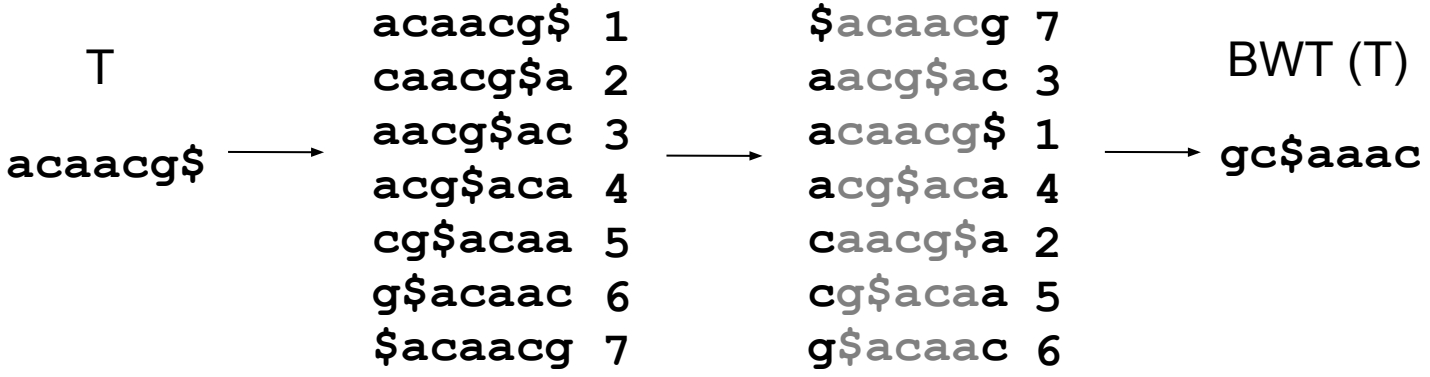
Look also at over-represented sequences

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.

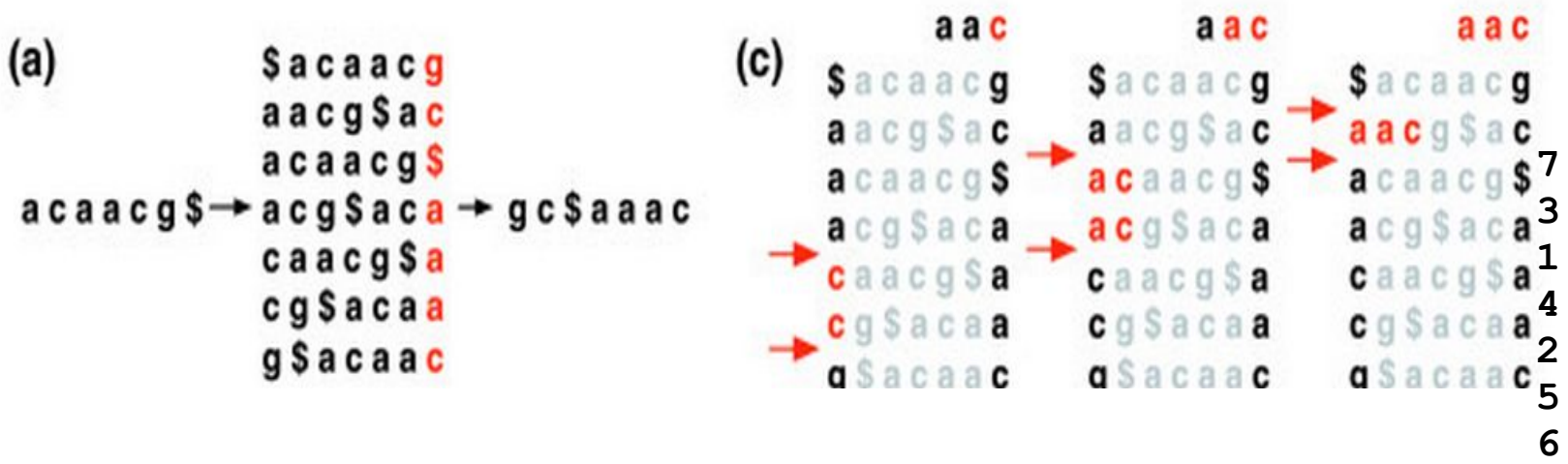
Bowtie a very popular aligner

- 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.



Bowtie principle

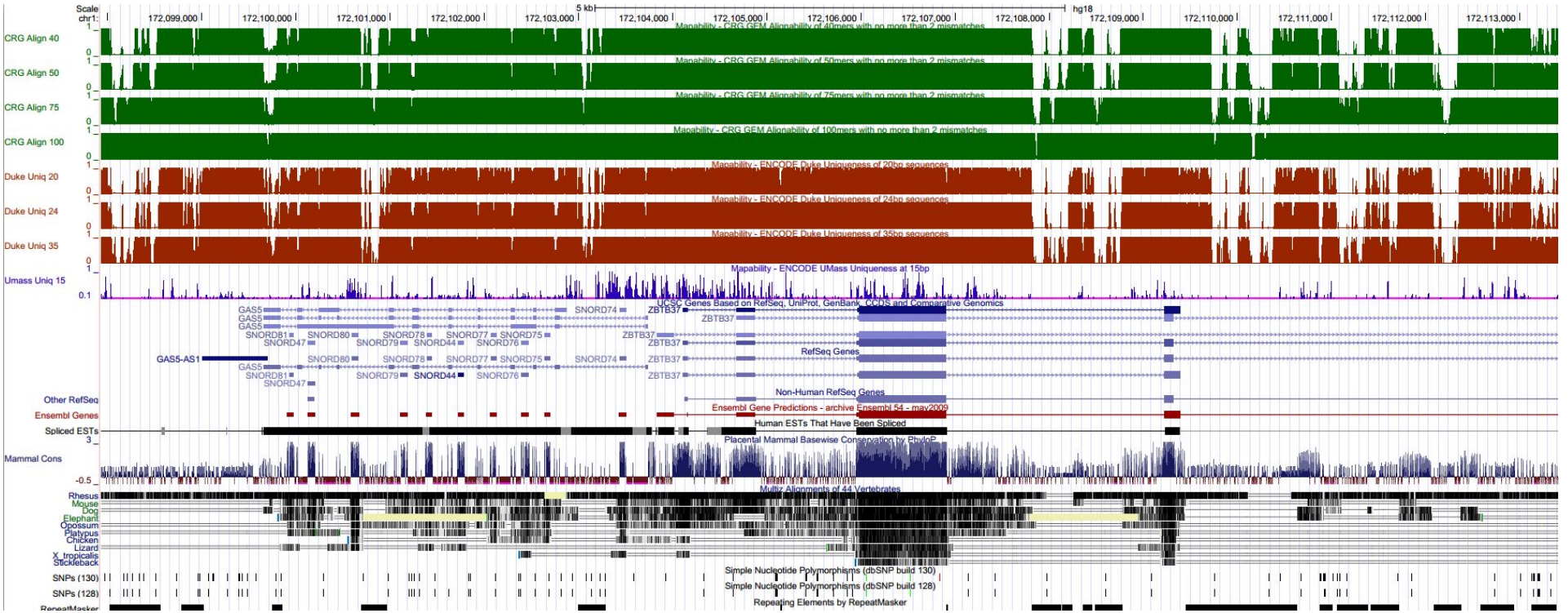
- Burrows-Wheeler Matrices have a property called the Last First (LF) Mapping.
  - The ith occurrence of character c in the last column corresponds to the same text character as the ith occurrence of c in the first column
  - Example: searching ”AAC” in ACAACG



- 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.



Mappability issues

- If one discards “multi-reads” (read with poor mapping qualities)
  - One may miss some true novel transcripts
  - One may discard the signal corresponding to some gene families
- If one keeps “multi-read”
  - Transcript discovery may lead to false-positives
  - May ends with biased signal for genes related to a gene family
- Important to check software policies
  - If multi-reads are accepted upon mapping
    - What is the rule (*e.g.* random choice)?
  - E.g Upon quantification featureCounts don’t take multi-reads into account (default parameters)

Mapping read spanning exons

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

Storing alignments: the SAM/BAM format

- SAM file stores information related to alignment
  - E.g read was aligned at position … on chromosome ...
- The SAM files contains:
  - Read ID
  - Chomosome and position
  - CIGAR String
  - Bitwise FLAG
  - Alignment position
  - Mapping quality
  - ...

Bitwise flag is a single column

- But it contains various informations
  - 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 explained

- 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 ...
- ...

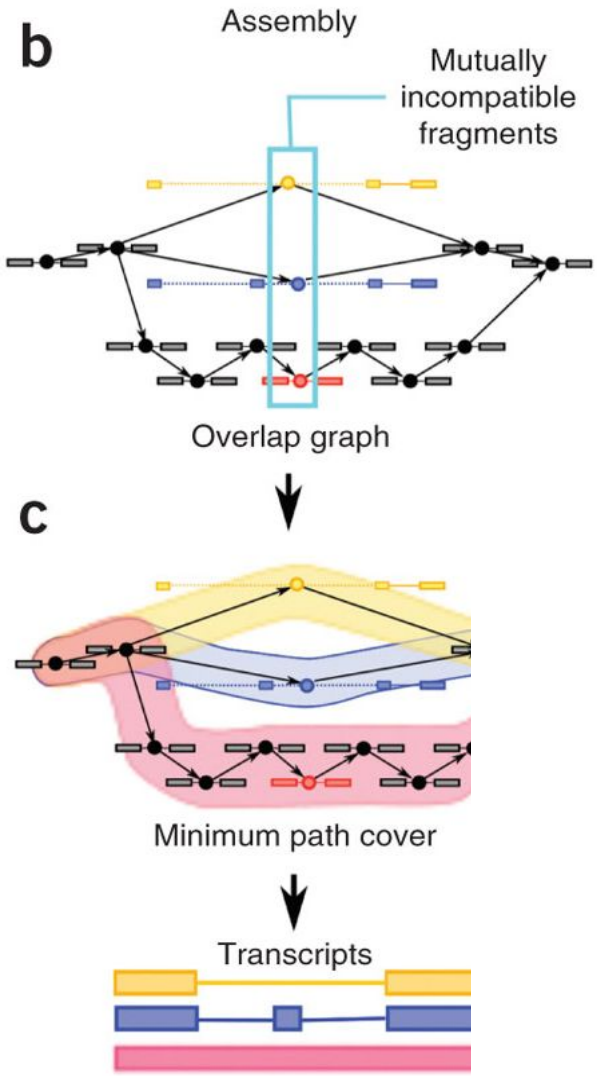
The extended CIGAR string

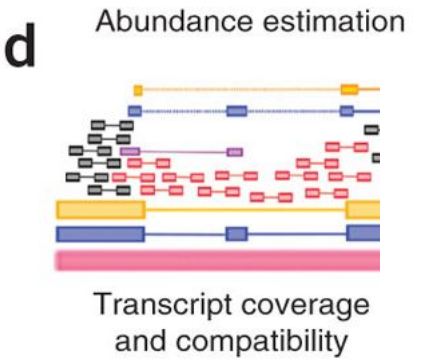
- Exemple flags:
  - M alignment match (can be a sequence match or mismatch !)
  - I insertion to the reference
  - D deletion from the reference
  - http://samtools.sourceforge.net/SAM1.pdf

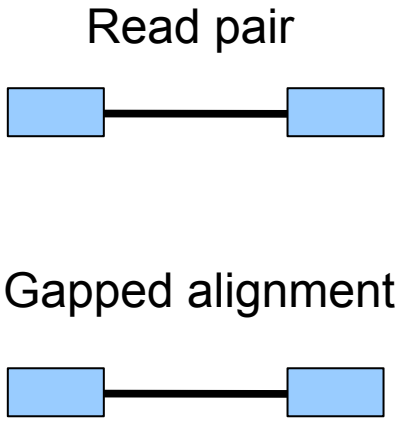
```
ATTCAGATGCAGTA
ATTCA--TGCAGTA
```

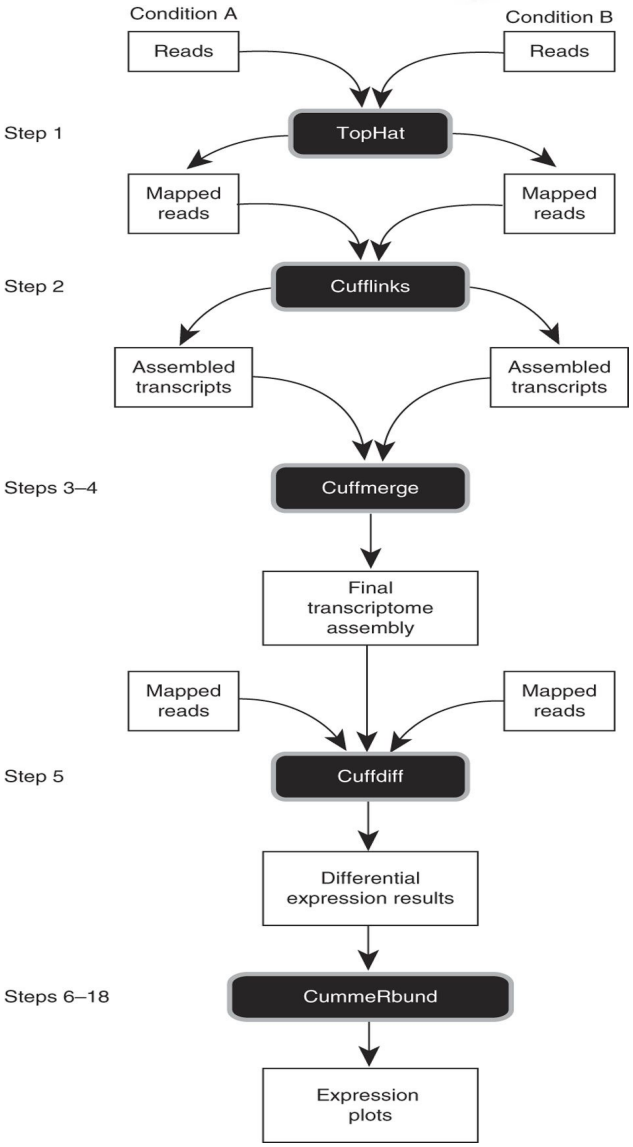
5M2D7M

Searching for novel transcript model: cufflinks

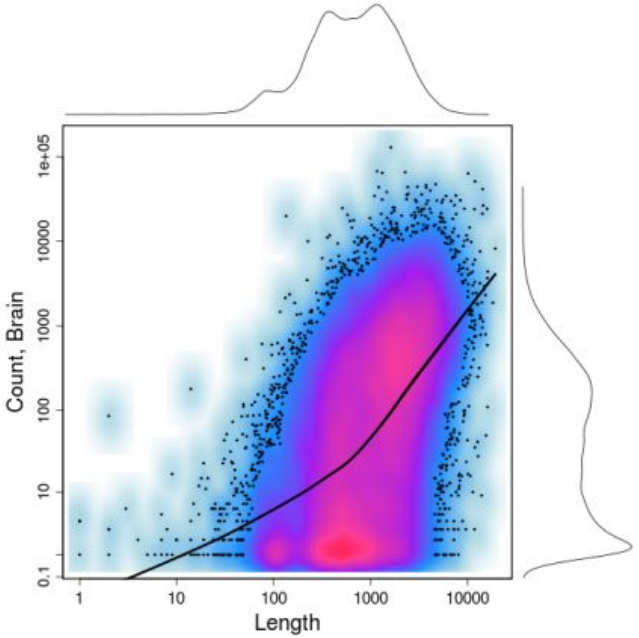








Quantification



(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

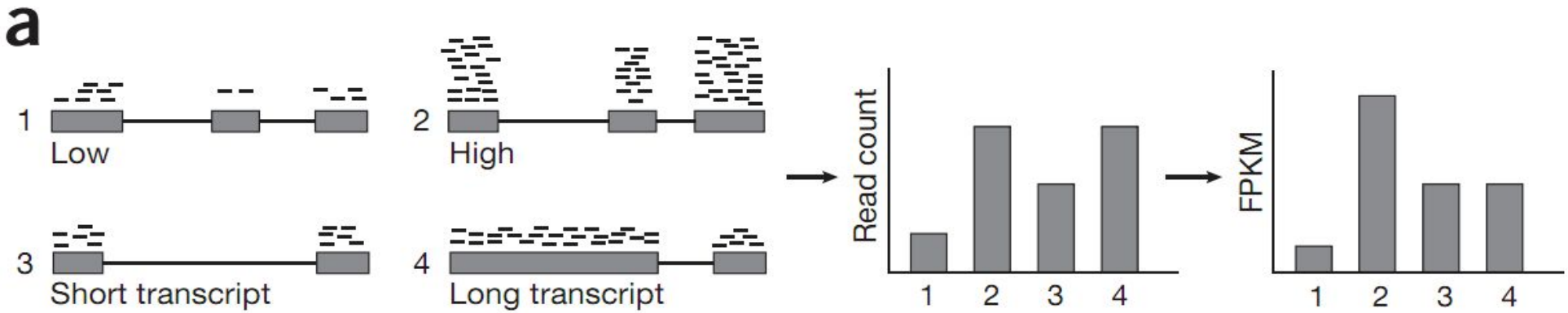
BMC Bioinformatics. 2010 Feb 18;11:94.

Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.

Bullard JH, Purdom E, Hansen KD, Dudoit S.

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,*]

RPKM/FPKM normalization

- RPKM: Tag count is normalized for transcript length and total read number in the measurement (RPKM, Reads Per Kilobase of exon model per Million mapped reads)
  - 2kb transcript with 3000 alignments in a sample of 10 millions of mapped reads
  - RPKM = 3000/(2 * 10) = 150
- 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,*]

Other issues in normalization

- Normalizing based on read counts alone is not sufficient
  - Problem with very abundant tissue specific transcripts.
    - Since a large amount of sequencing is dedicated to these transcript, there is less sequencing available for the remaining genes. The remaining genes may thus appear as down-regulated

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.

Some proposed normalization methods

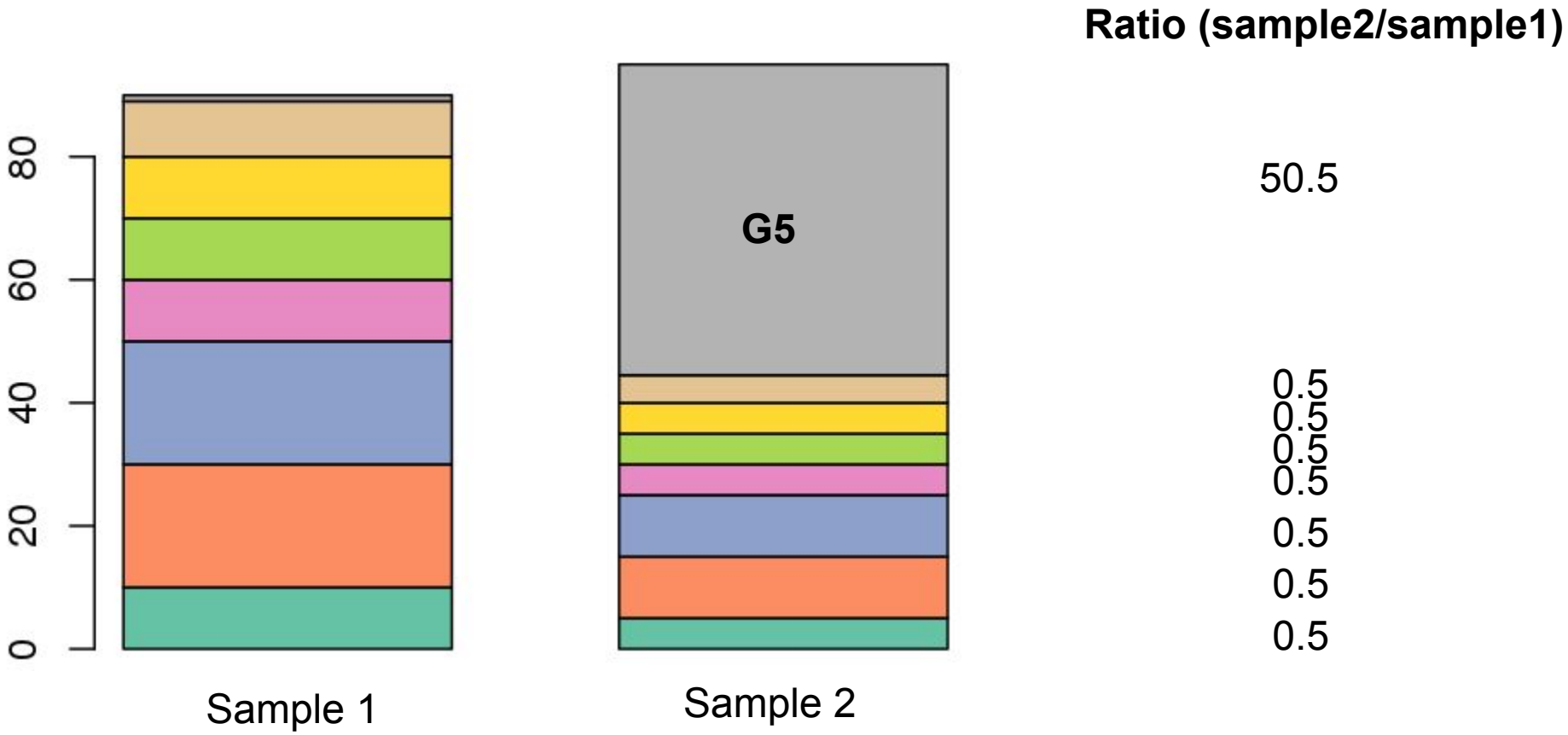
- Reads Per Kilobase per Million mapped reads (RPKM): This approach was initially introduced to facilitate comparisons between genes within a sample.
  - Not sufficient

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

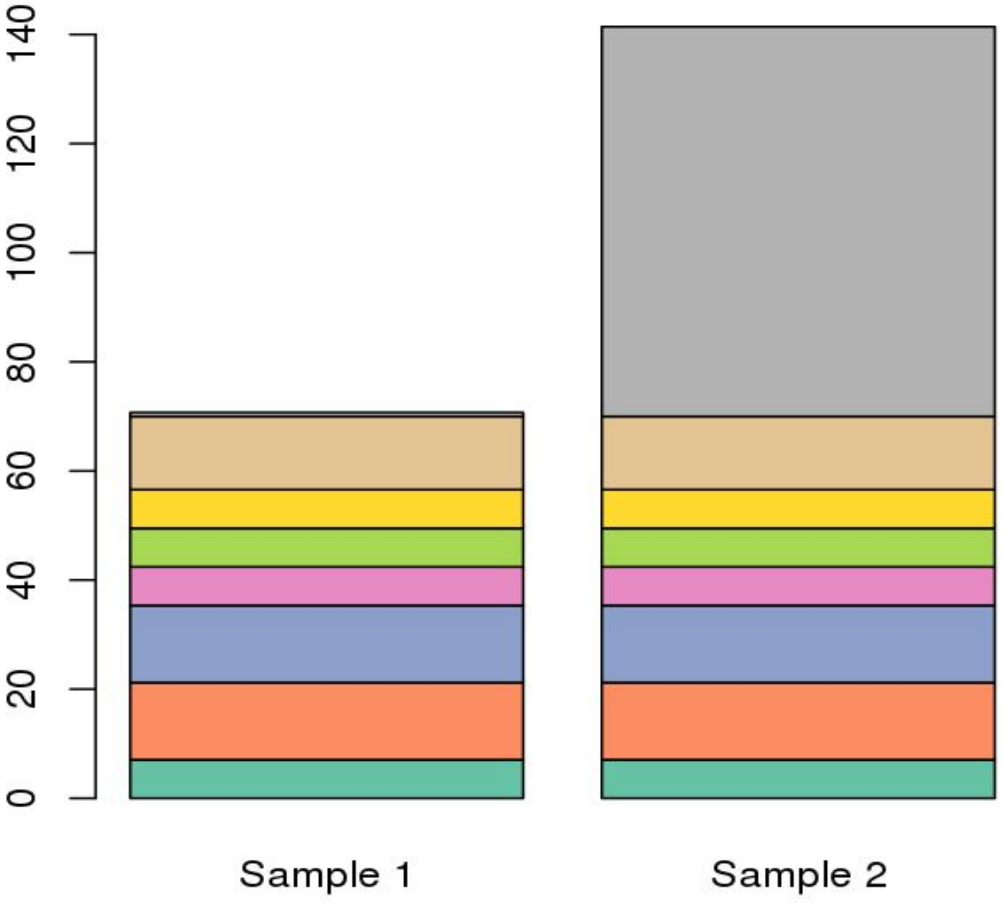
- Trimmed Mean of M-values (TMM): This normalization method is implemented in the edgeR Bioconductor package (version 2.4.0). Scaling is based on a subset of M values
  - TMM seems to provide a robust scaling factor.

Main issues in RNA-Seq normalization

- Highly abundant genes:
  - E.g; All genes unchanged but G5
    - Total count → repression of all other genes by a factor 2 !



TMM Normalization (Robinson and Oshlack, 2010)



- Outline
  - Compute the M values (log ratio).
    - Take the trimmed mean of the M value as scaling factor.
    - Divide 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**

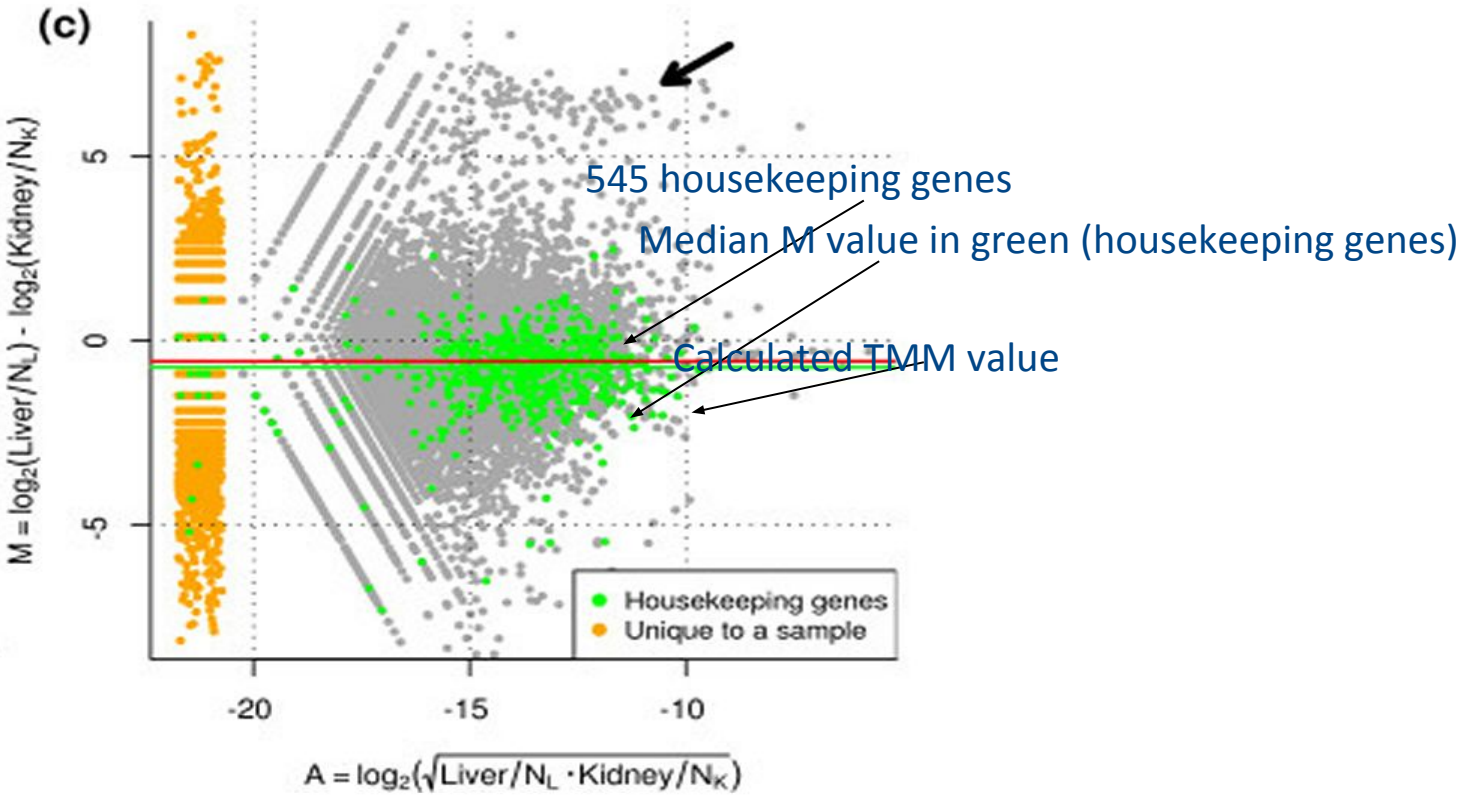
Genome Biol. 2010;11(3):R25. Epub 2010 Mar 2.

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

Robinson MD, Oshlack A.

TMM normalization

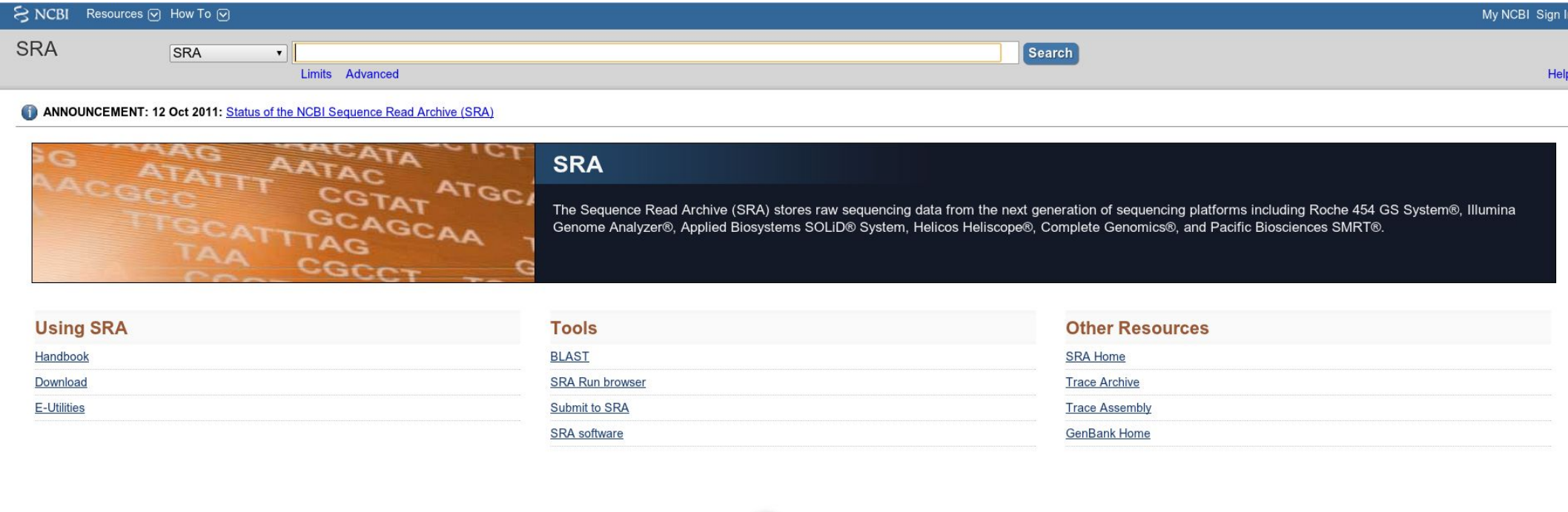
- Example data set:
  - Liver vs kidney analysis
- Legend
  - Green points indicate 545 housekeeping genes
  - the green line signifies the median log-ratio of the housekeeping genes.
  - The red line shows the estimated TMM normalization factor.



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)



- 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.

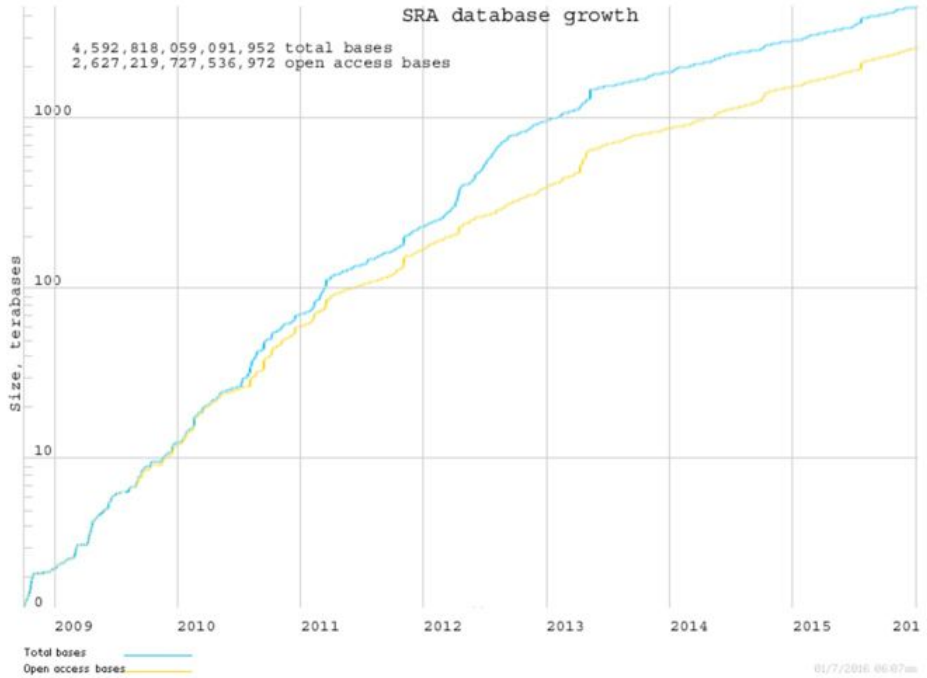
Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA 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

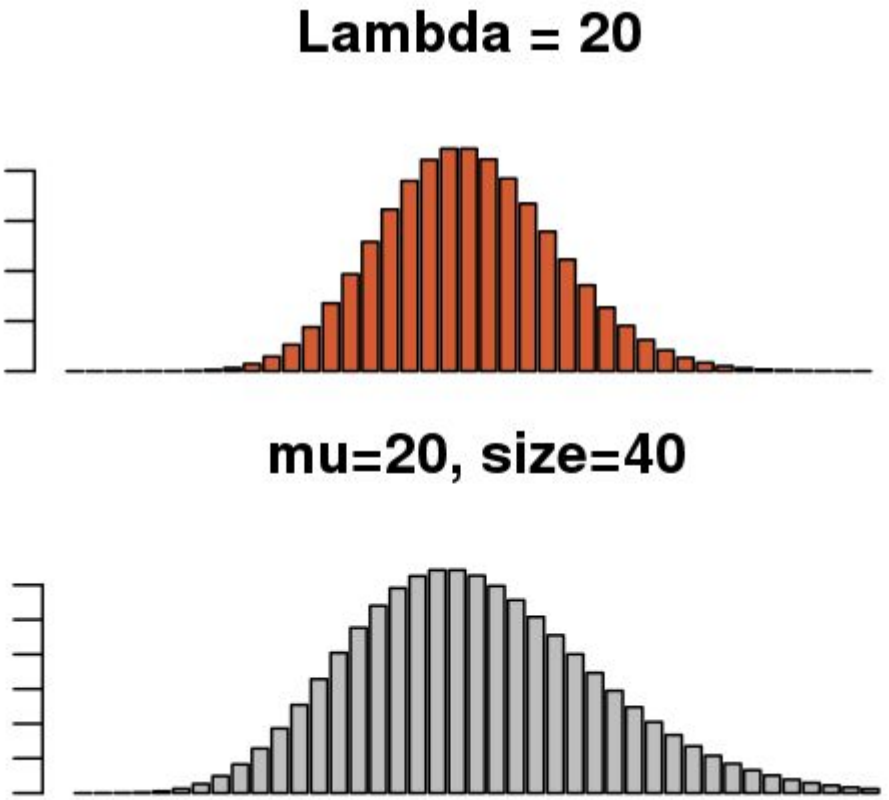


Merci

Two-class Differential expression analysis

- Problematic
  - What is the underlying distribution of read counts
    - If reads for gene *g* were obtained from a population of samples with equal expression level one could model read counts of *g* as a **poisson distribution**
    - However, depending on samples, expression level may vary in each class according to :
      - Genes type (*e.g, stress-responsive genes*)
      - Biological samples (*e.g*, purity)
      - → Overdispersion
    - Poisson distribution predict smaller dispersion than observed in the data
      - incorrectly optimistic p values

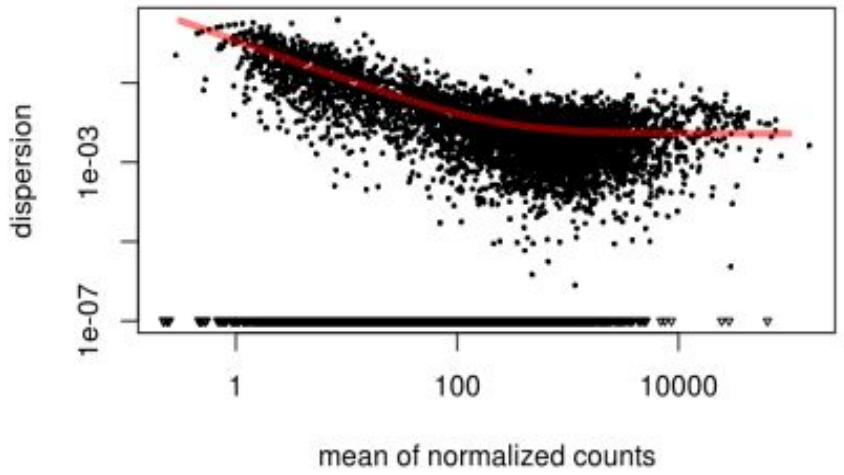
Negative binomial



- Poisson
  - One parameter, λ
  - Variance is equal to λ
- Negative binomial
  - Has two parameters mean (μ) and variance (σ^2).
  - Can be used as an alternative model to the Poisson distribution when sample variance exceeds the sample mean.

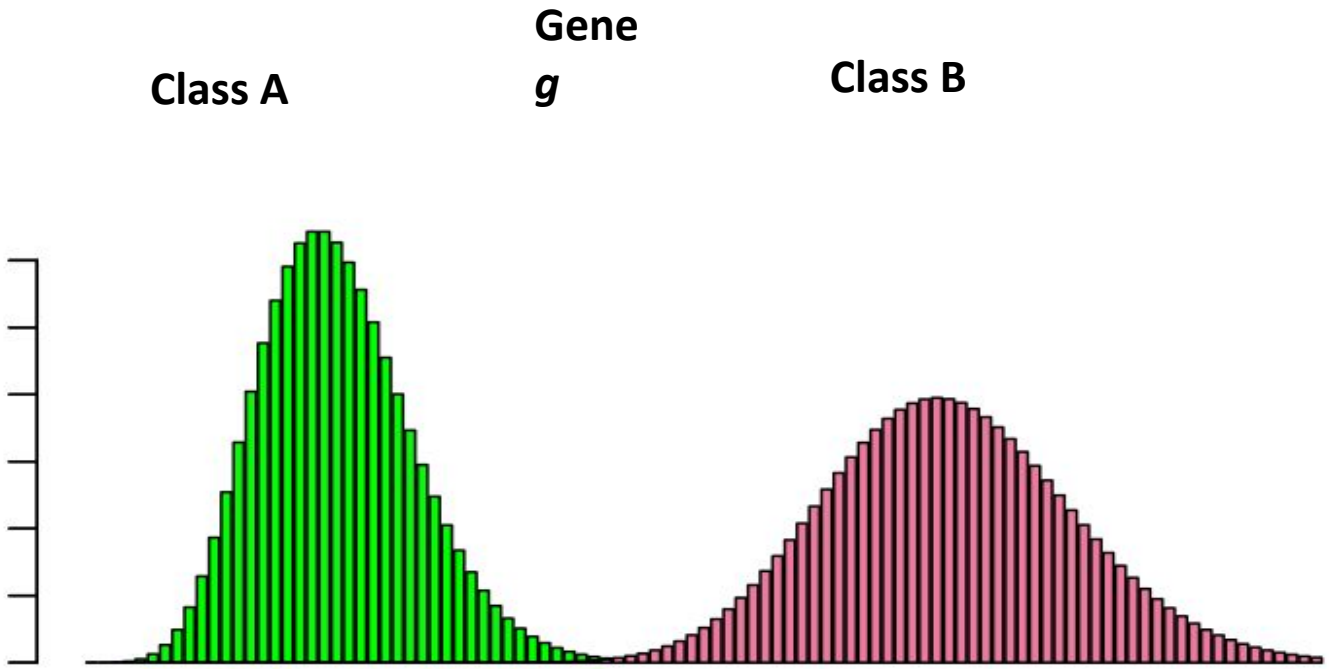
Estimating dispersion (DESeq)

- Variance observed between counts is the sum of two components
  - Sample-to-sample variation, biological variation (**dispersion**, dominate in highly expressed genes)
  - Uncertainty in measure (**shot noise**, dominate in weakly expressed genes)
- Variance is estimated in each class by using a shrinkage method



Test for differential expression (DESeq)

- Intuition

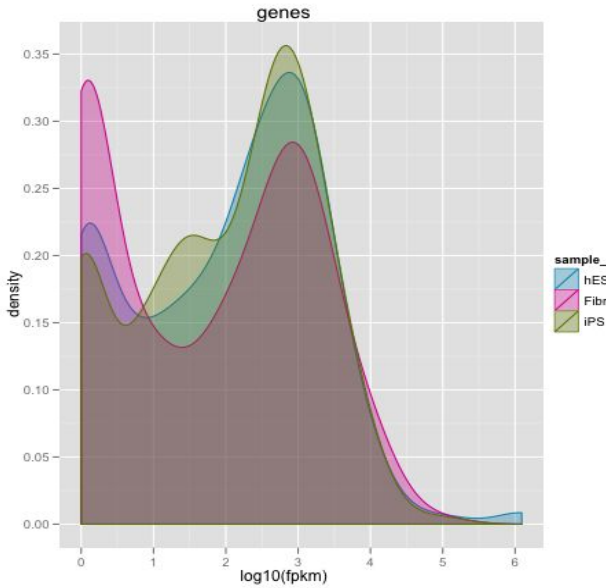


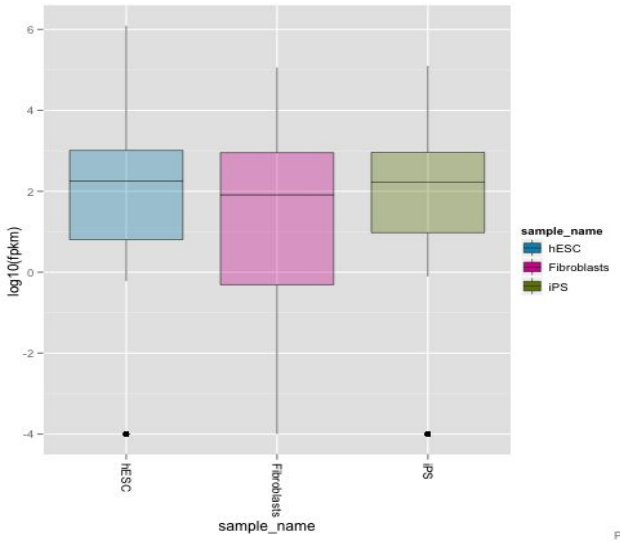
- The test implemented in DESeq is based on the sums of counts in class A and B (that are NB distributed variables)

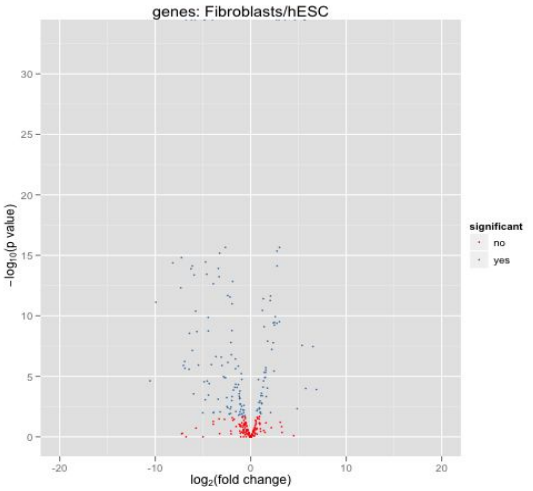
Cuffdiff

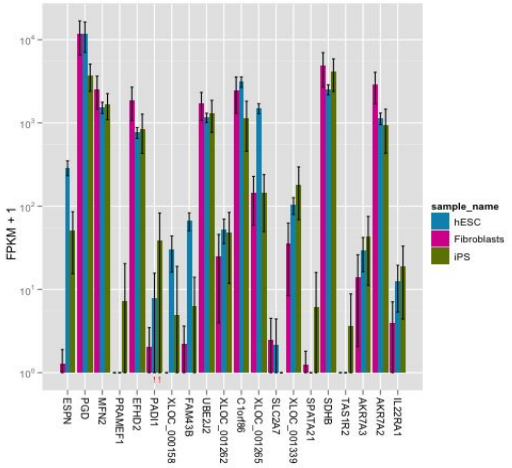
- Differential expression
  - Gene
  - Alternative transcripts
  - Alternative 5' UTRs
  - ...

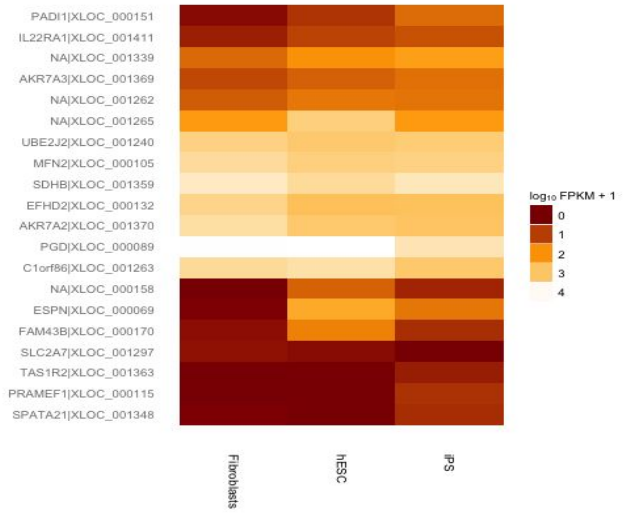
CummeRbund











- cummeRbund is a visualization package for Cufflinks high-throughput sequencing data.

Limits of RPKM/FPKM

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

Other normalization methods

- **Several methods proposed**

- Total count (TC): Gene counts are divided by the **total number of mapped reads** (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset.
  - First proposed
- Median (Med): Also similar to TC, the total counts are replaced by the **median counts different from 0** in the computation of the normalization factors.
  - Warning : if lots of weakly expressed values
- Upper Quartile (UQ): the total counts are replaced by the **upper quartile** of counts different from 0 in the computation of the normalization factors.
  - Very similar in principle to TC (but really more powerful).
- Trimmed Mean of M-values (TMM): This normalization method is implemented in the **edgeR** package and the scaling is based on a subset of M values.

Brief Bioinform. 2012 Sep 17. [Epub ahead of print]

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; on behalf of The French StatOmique Consortium.

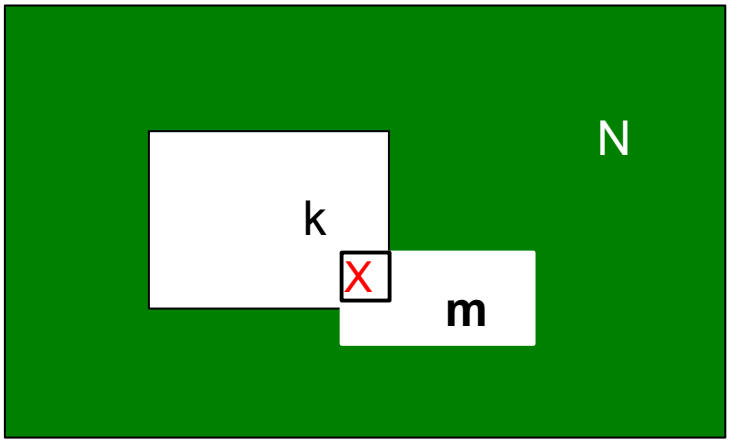
Differential Expression

- Several methods proposed
  - Fisher, EdgeR, DESeq, NOISeq, Cuffdiff...

Fisher's exact test

- If reads count for a gene G are balanced we should expect ~ same number of read in both conditions.
- Can be viewed as the number of white balls drawn without replacement from an urn which contains both black and white balls.

	Cont	Treated	
Reads from gene G	x	m-x	m (white)
Remaining reads	k-x	n-(k-x)	n (black)
	k	N-k	N

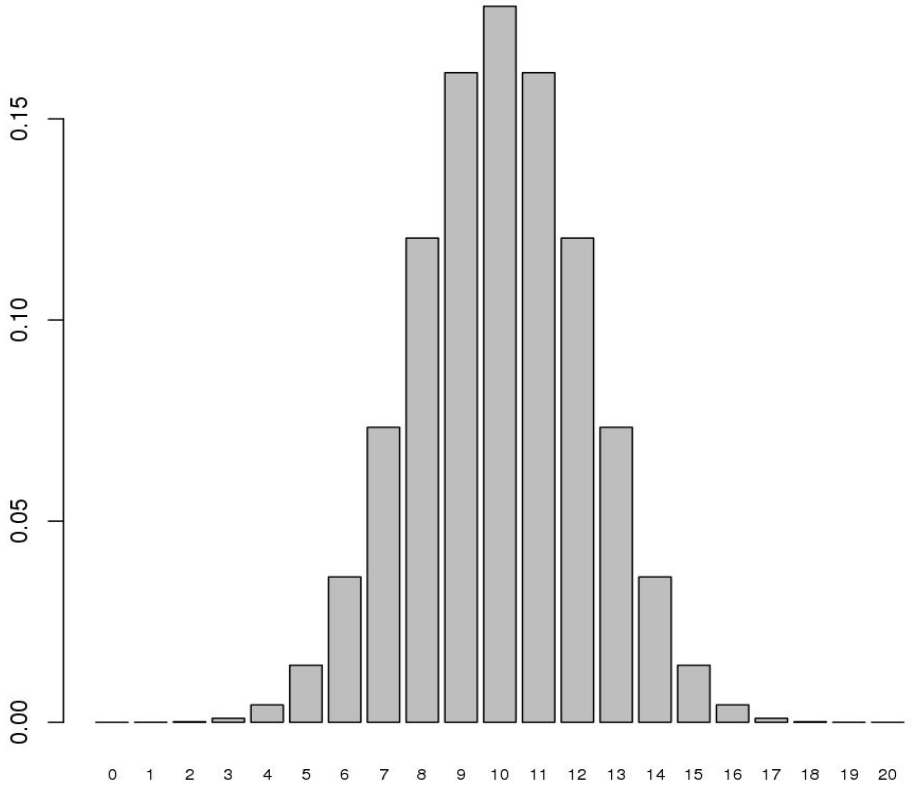


$$P(X = x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

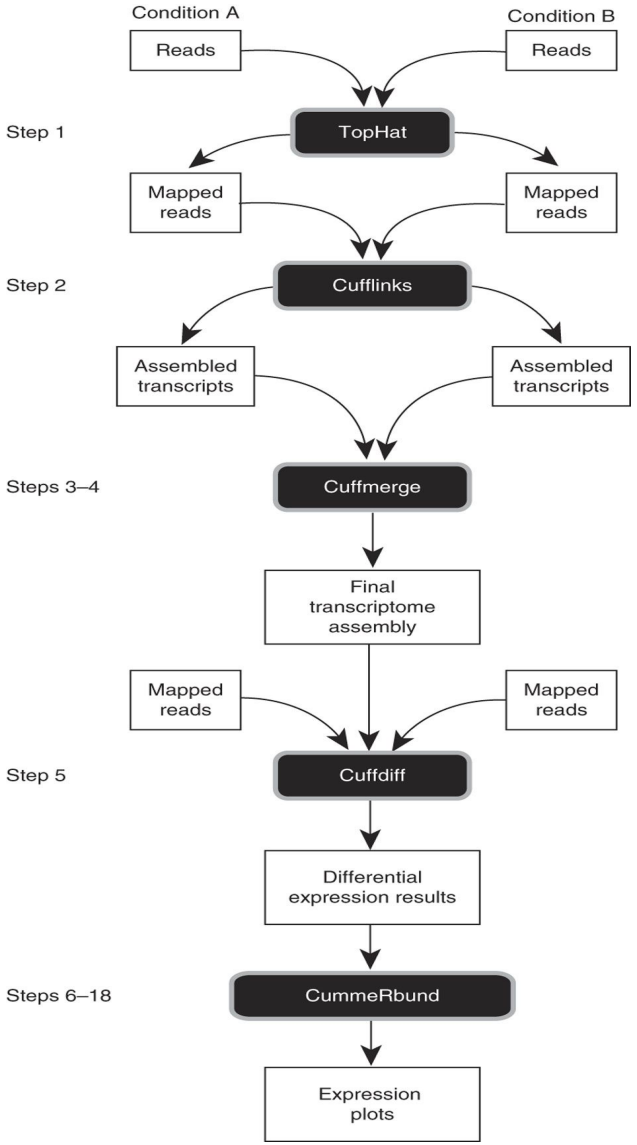
- x ~ Hypergeometric (N, K, n)

Example

- 1000 reads
- Reads for gene G = m =20
- Remaining reads = n = 980
- Reads in control sample = k = 500



The Tuxedo pipeline



Nat Protoc. 2012 Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016.

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L.

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 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.

Mapping results

- IL2RA
- Output : BAM file (compressed version of SAM)

TopHat pipeline

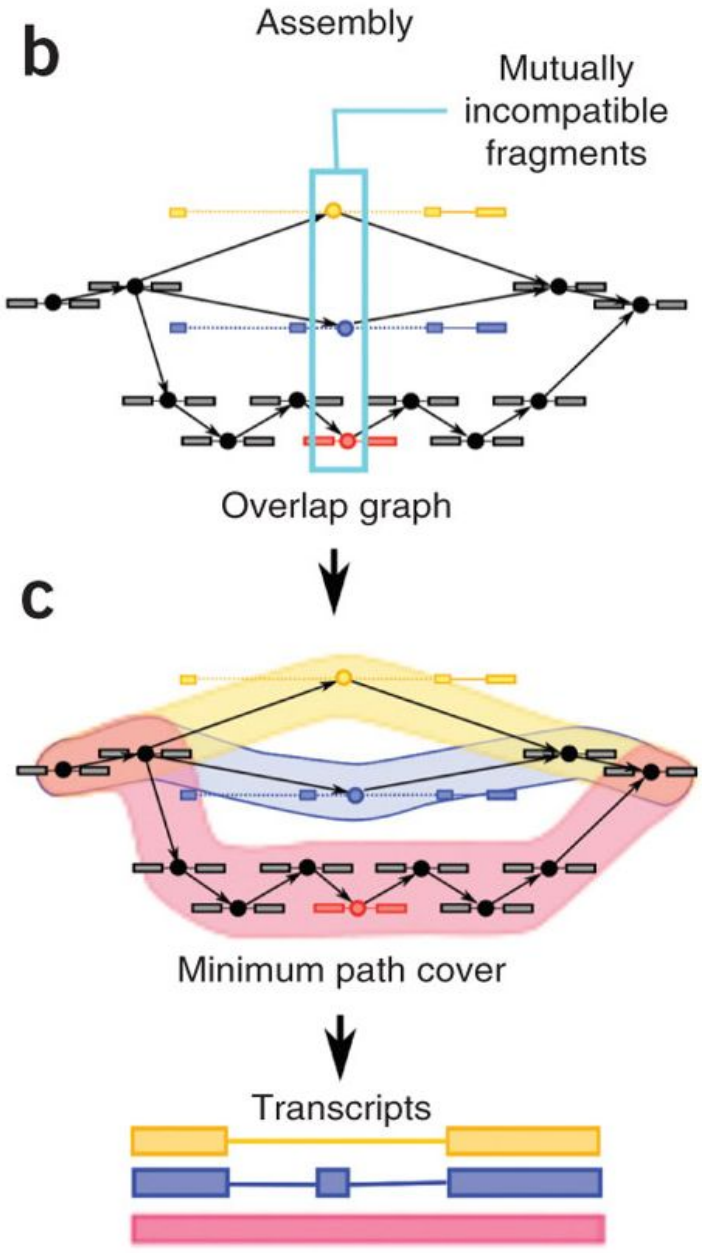
- Weakly expressed genes should be poorly covered
  - Exons may have gaps
- To map reads to splice junctions, TopHat first enumerates all canonical donor and acceptor sites within the island sequences (as well as their reverse complements)

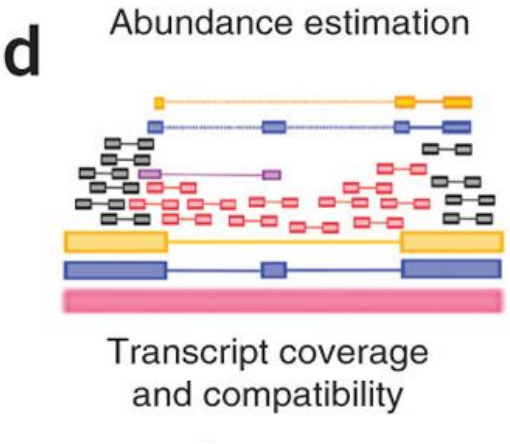
- Next, tophat considers all pairings of these sites that could form canonical (GT–AG) introns between neighboring (but not necessarily adjacent) islands.
  - By default, TopHat examines potential introns longer than 70 bp and shorter than 20 000 bp (more than 93% of mouse introns in the UCSC known gene 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 reads

- Main Issues:
  - Number of allowed mismatches
    - Depend on sequence size (sometimes heterogeneous length)
    - Depend of the aligner
  - Number of multi-hits
    - Issue with short reads
  - PCR duplicates
    - Accepted with RNA-Seq
    - Warning with ChIP-Seq (library complexity)
  - Mates expected distance (mate/paired-sequencing)

Cufflinks: transcript assembly and quantification





Read pair

Gapped alignment

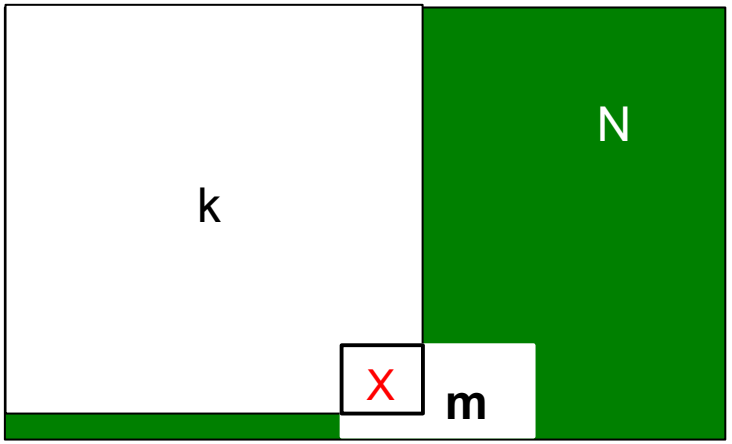
Differential Expression

- Several methods proposed
  - Fisher, EdgeR, DESeq, NOISeq, Cuffdiff...

Fisher's exact test
(or hypergeometric test)

- Simple two-library comparison
- If read counts for a gene g are balanced we should expect ~ same number of read in both conditions.

	Cont	Treated	
Reads from gene G	x	m-x	m (white)
Remaining reads	k-x	n-(k-x)	n (black)
	k	N-k	N



- x follows a hypergeometric distribution with parameter N, K, n