High throughput methods approches in genomics

D. Puthier

Genomics

“The science for the 21st century”

Ewan Birney(EMBL-EBI)

at GoogleTech talk

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 in the clinical field

- In the clinical field genomics is a tool of choice
  - Define Biomarkers
    - Diagnosis
      - E.g. Tumor class ?
    - prognosis
      - Patient outcome ?
    - Develop personalized medicine
      - Adapt treatment based on genetic background

Genomics an interdisciplinary science

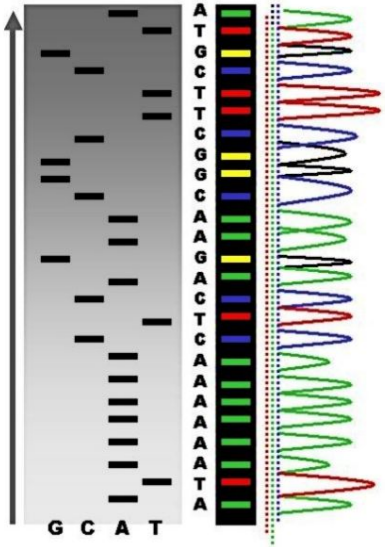
Analysing genomes requires teams/individuals with various skills

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

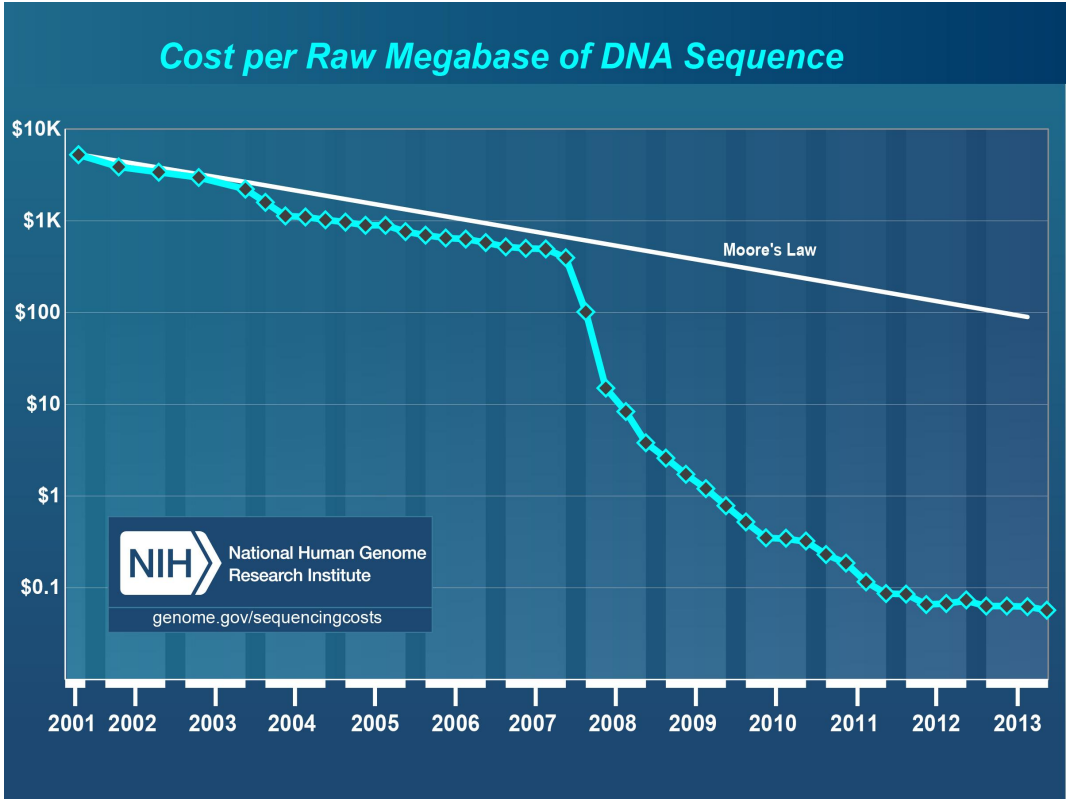
Breakthrough in DNA Sequencing

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

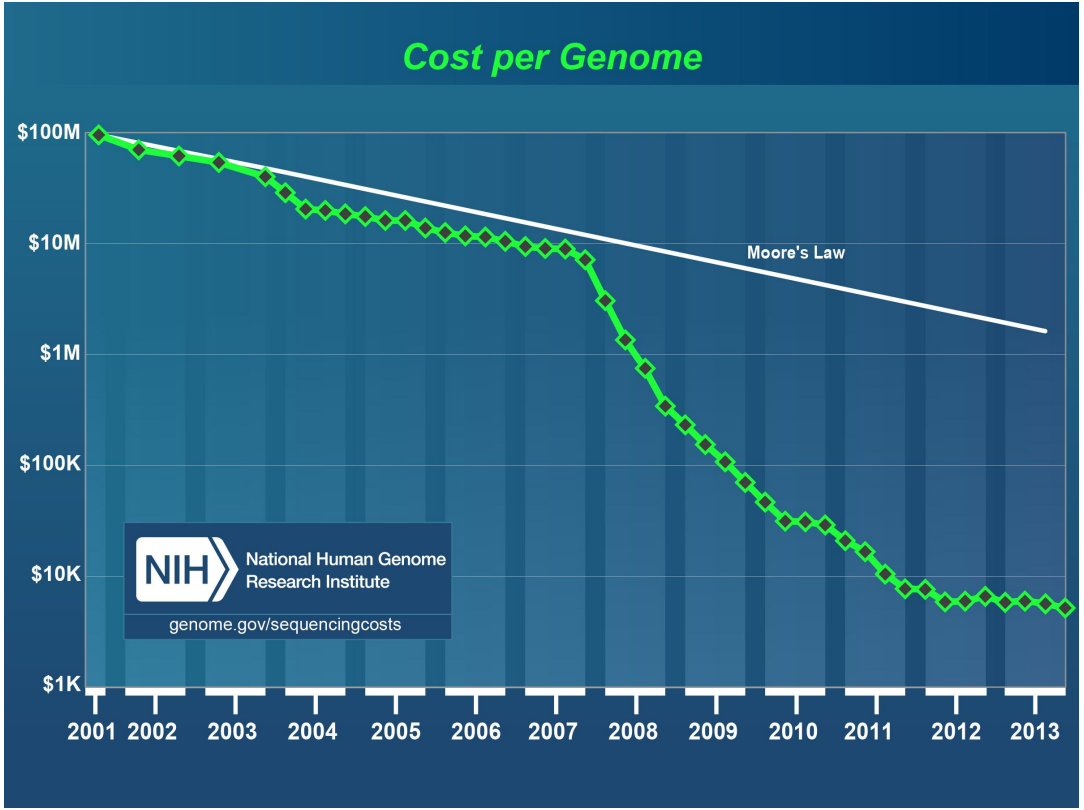
(2D sequencers “Next Generation Sequencing.”)



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

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.

Business Wire

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

- What about pricing for analysis ?

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

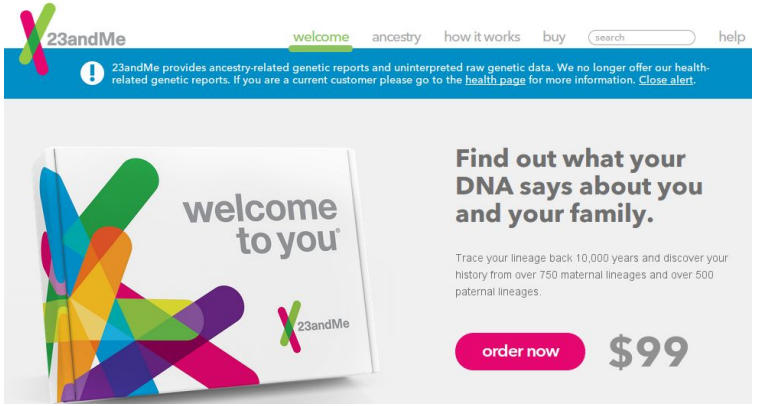
An infant delivered last week in California appears to be the first healthy person ever born in the U.S. with his entire genetic makeup deciphered in advance.

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," finishing a PhD in feline p University of California, D 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

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.

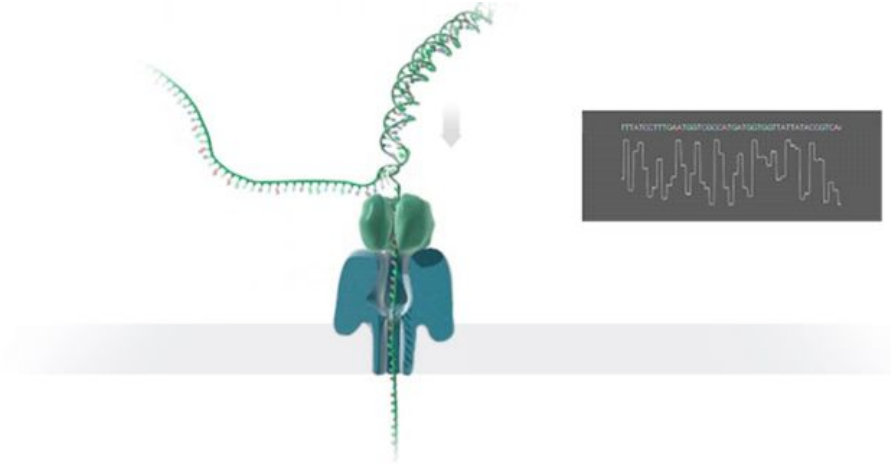
[h]Average.

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.



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)

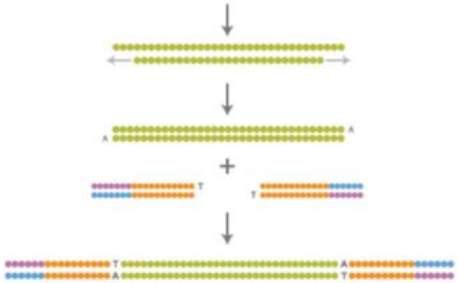
"The Oxford Nanopore Technologies (ONT) MinION is a new sequencing technology that potentially offers read lengths of tens of kilobases (kb) limited only by the length of DNA molecules presented to it."

~1Gb to 2 Gb of sequence per minION

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

NGS: a simplified view

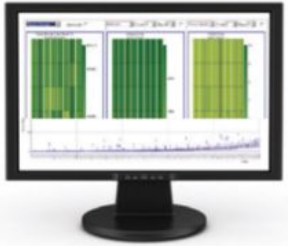
Figure 3: Next-Generation Sequencing Simplified



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



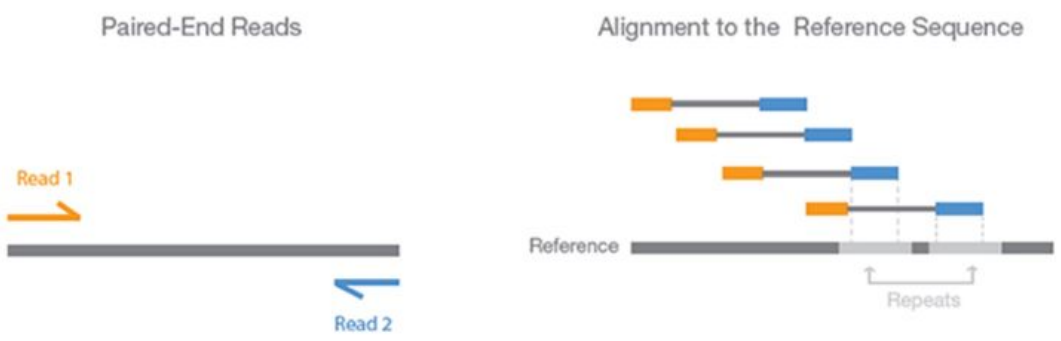
CASAVA
2 days (30 min hands-on)

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.

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

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

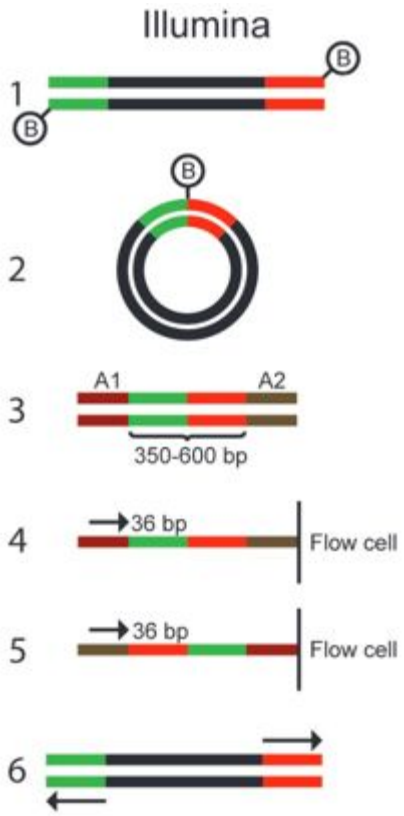
MATE-Pair sequencing ?

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

Figure 5. *De Novo* Assembly with Mate Pairs

Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

MATE-Pair library preparation



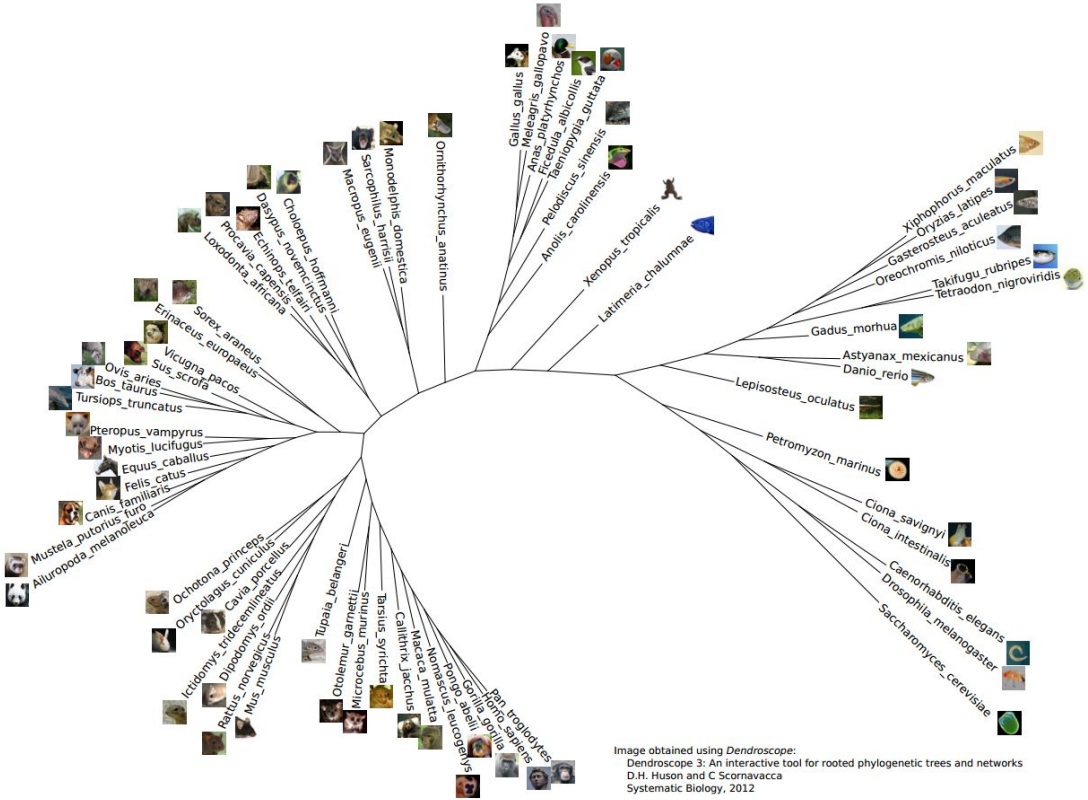
- Fragments are end-repaired using biotinylated nucleotides (1). After circularization, the two fragment ends (green and red) become located adjacent to each other
- The circularized DNA is fragmented, and biotinylated fragments are purified by affinity capture. Sequencing adapters (A1 and A2) are ligated 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.

Illumina sequencing principle

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

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

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

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

SNP or mutation ?

- Mutation : any change in a DNA sequence away from normal (this implies a normal allele which is prevalent in the population)
- Polymorphism : a DNA sequence variation that is common in the population (an alternative).
  - The arbitrary cut-off point between a mutation and a polymorphism is generally 1 per cent (0.5 for the 1000 genome project)

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

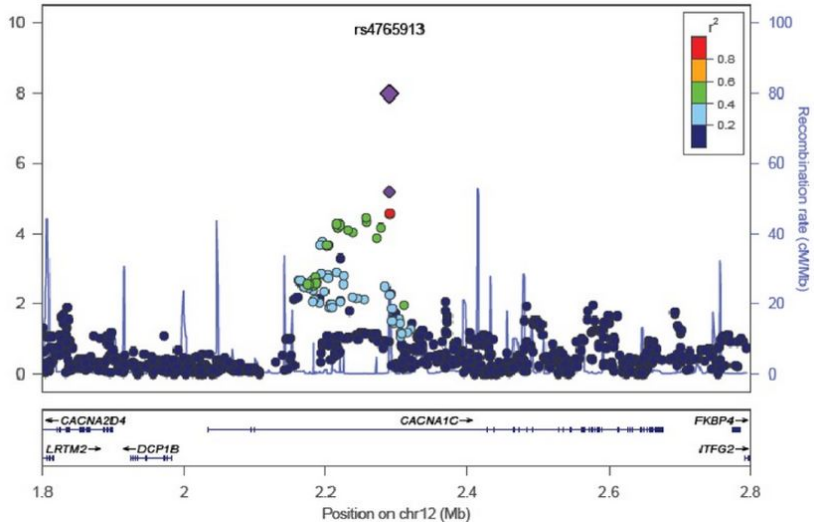
Bipolar disorder (BD) is a severe mood disorder affecting greater than 1% of the population[1]. Classical BD is characterized by recurrent manic episodes that often alternate with depression. Its onset is in late adolescence or early adulthood and results in chronic illness with moderate to severe impairments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

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.

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

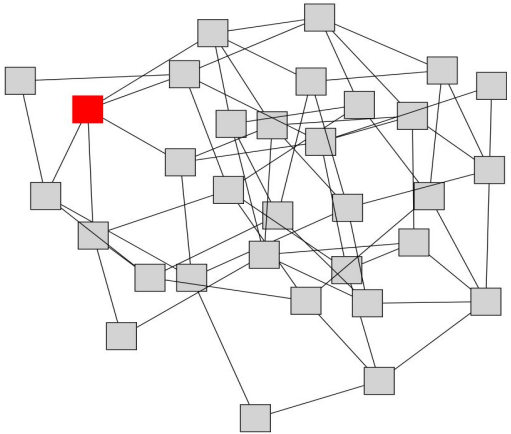
PMCID: PMC3637176
HALMS: HALMS634944

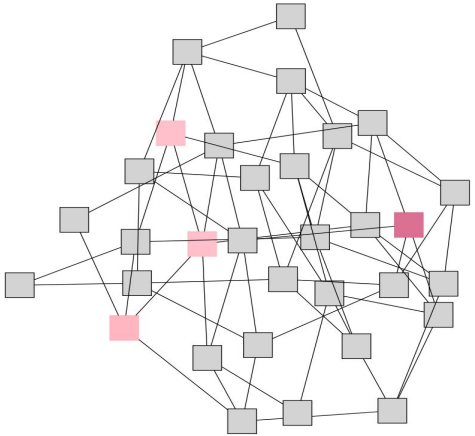
Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4



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 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 an 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

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

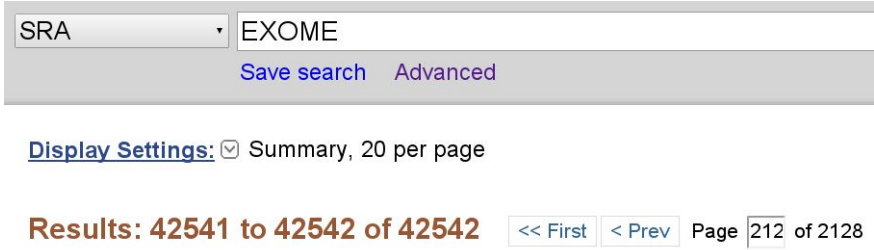
Published: November 12, 2014 • DOI: 10.1371/journal.pone.0112430

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.

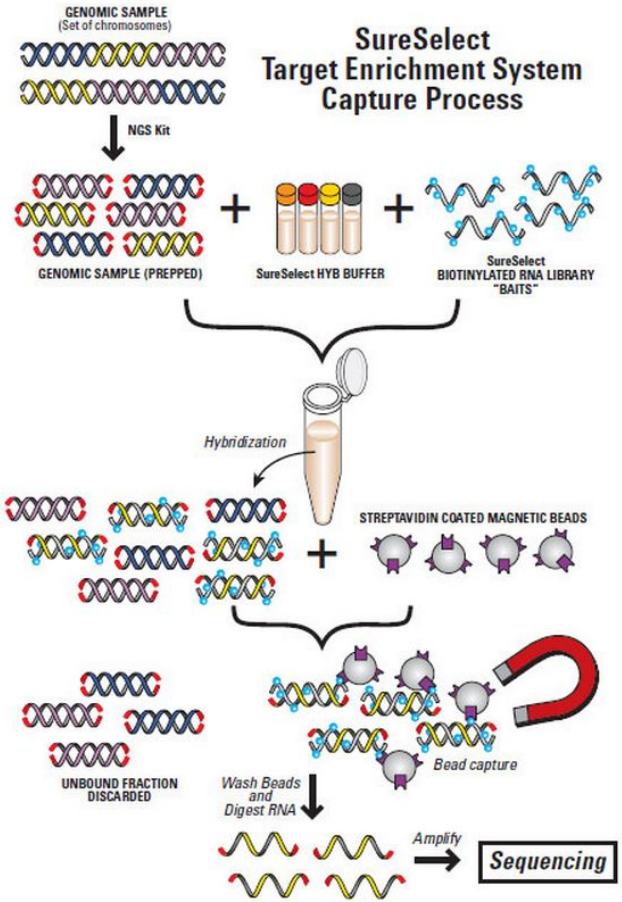
Analysing variations in exome

- 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

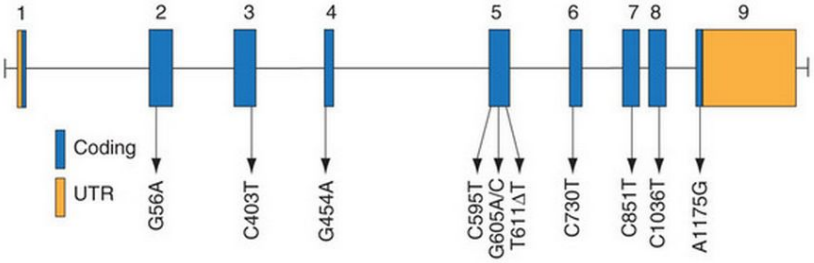
Affiliations | Contributions | Corresponding authors

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

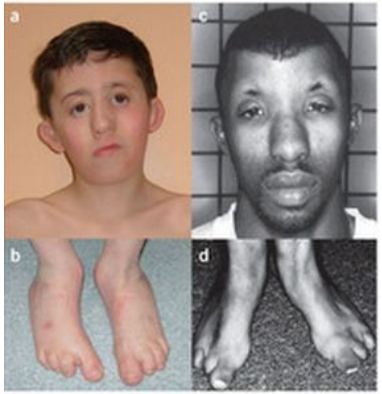
Received 02 October 2009 | Accepted 09 November 2009 | Published online 13 November 2009

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)

Exome sequencing of renal cell carcinoma

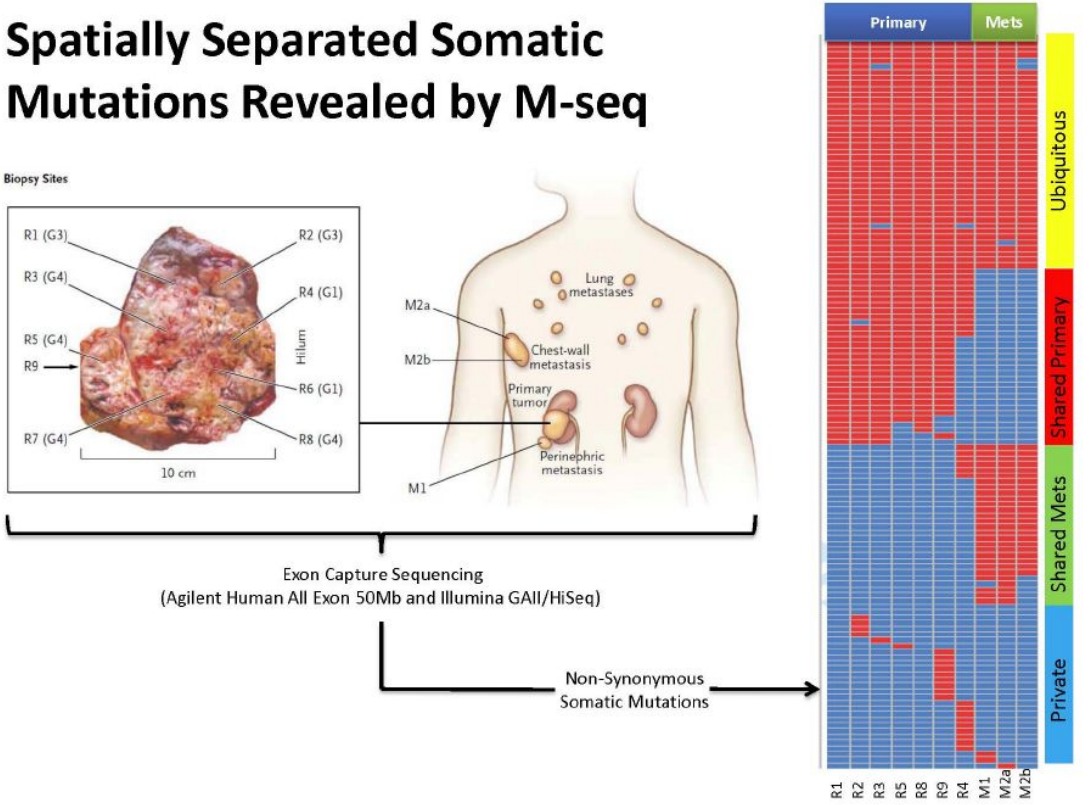
Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing

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

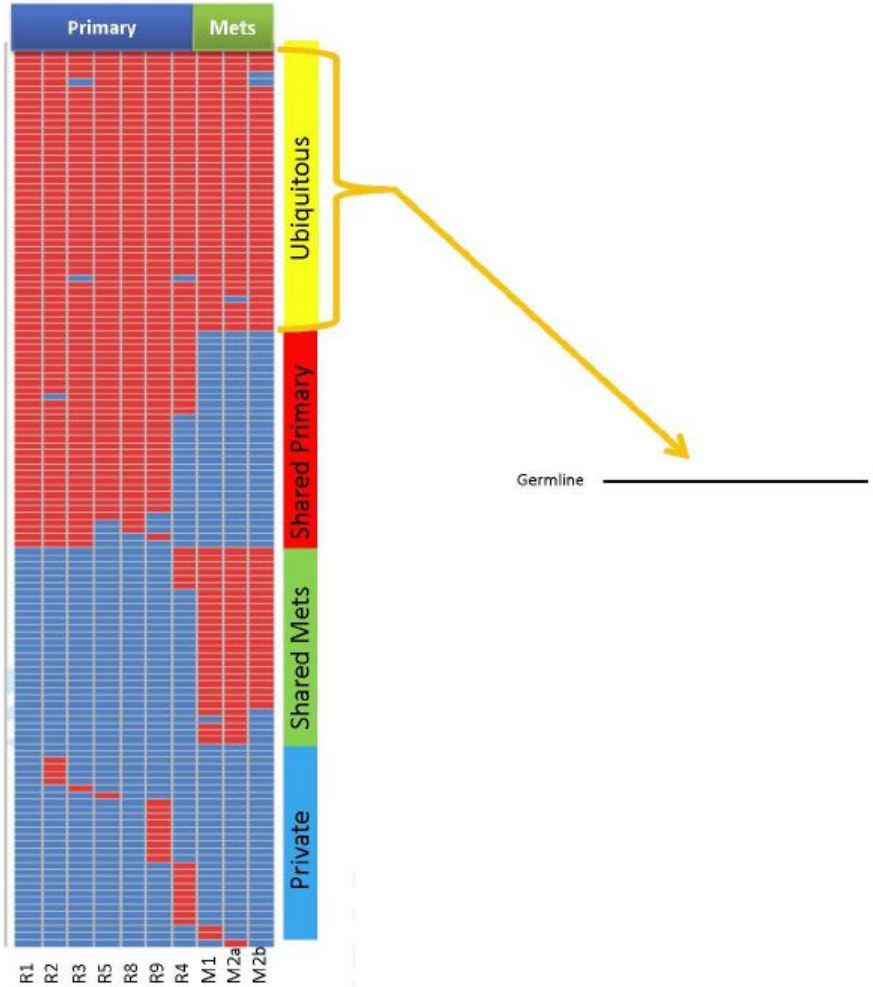
Cancer a clonal disease evolving in a linear fashion ?
What about tumor heterogeneity ?
Can we re-constitute the evolution of the tumor ?

Exome-Seq of Renal cell carcinoma

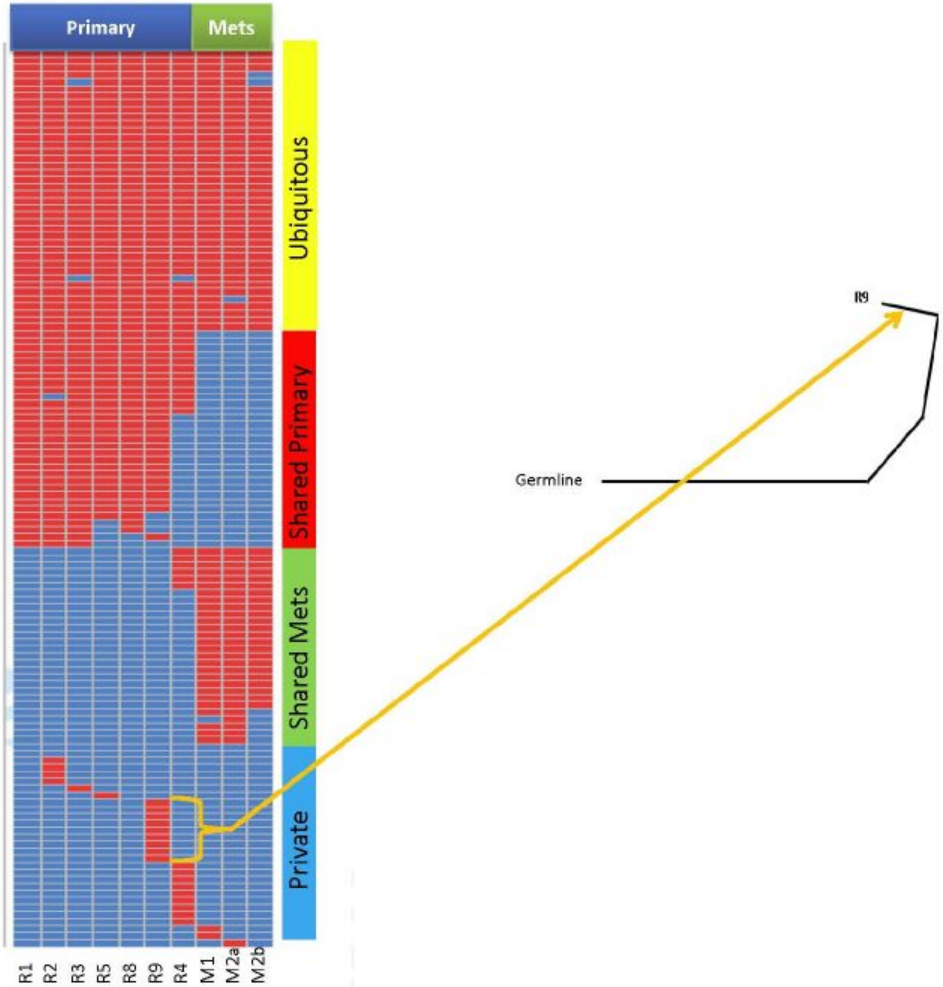


Gerlinger et al, N Engl J Med, 2012

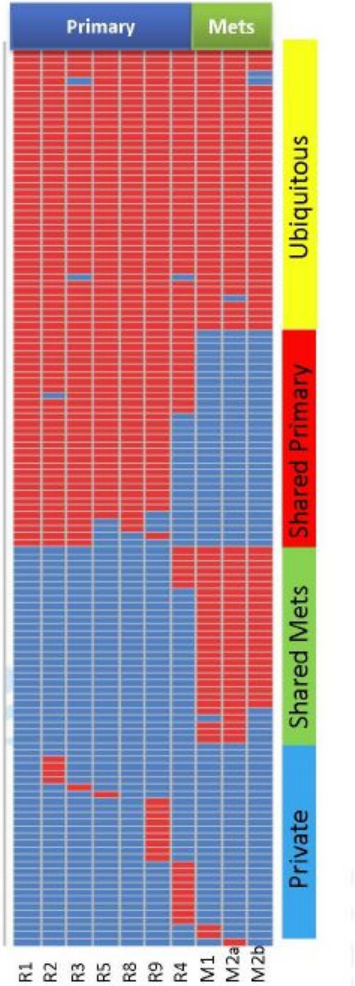
Phylogenetic reconstruction by clonal ordering

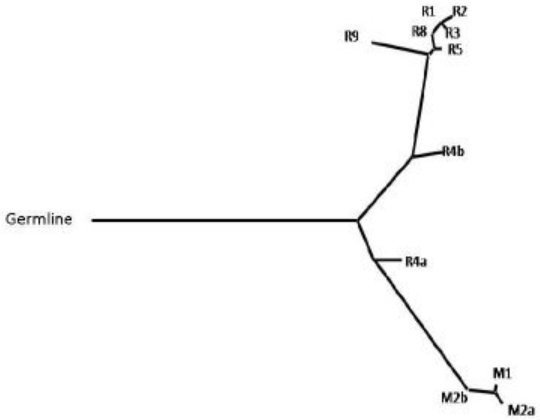


Phylogenetic reconstruction by clonal ordering

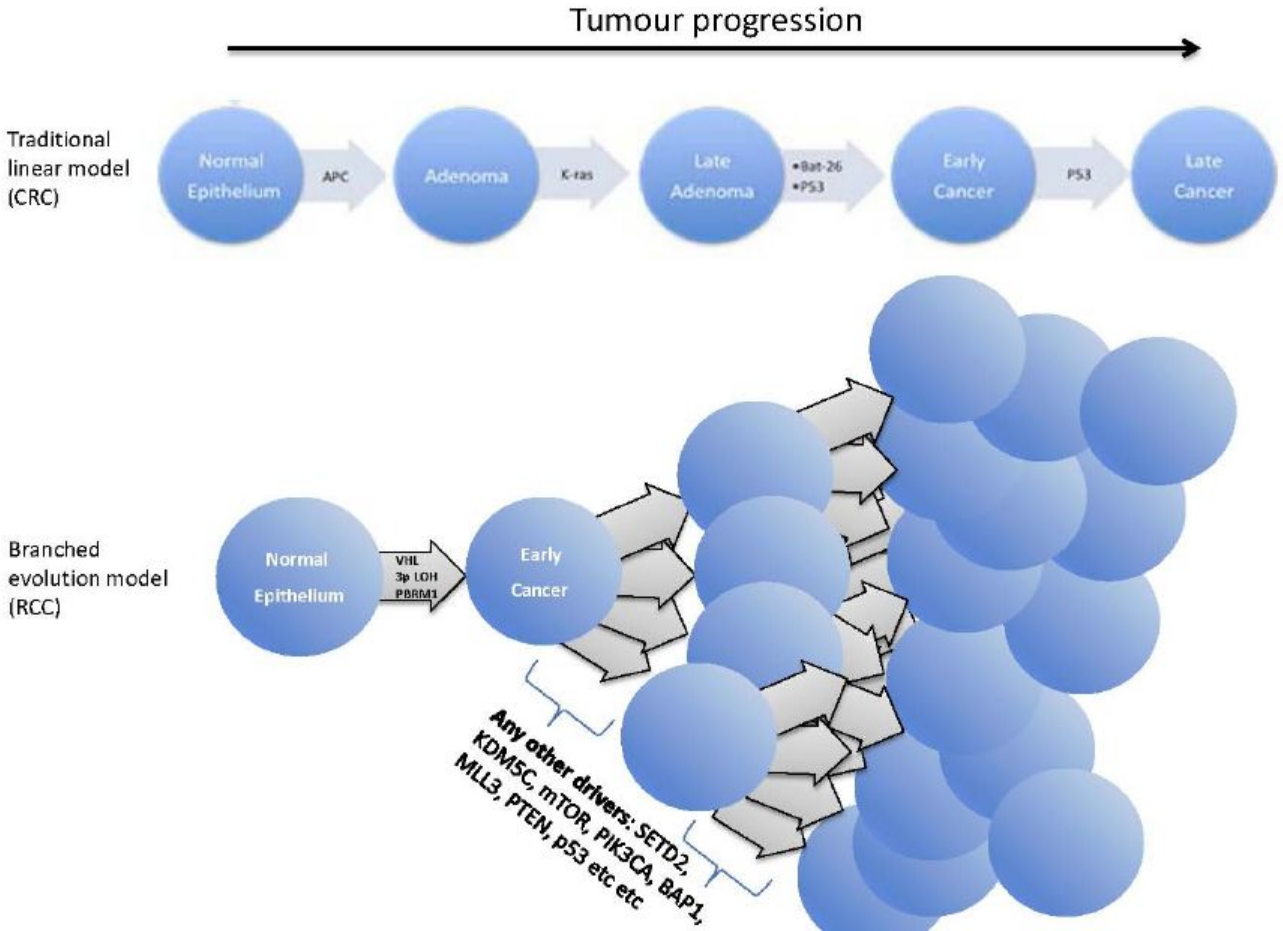


Phylogenetic reconstruction by clonal ordering

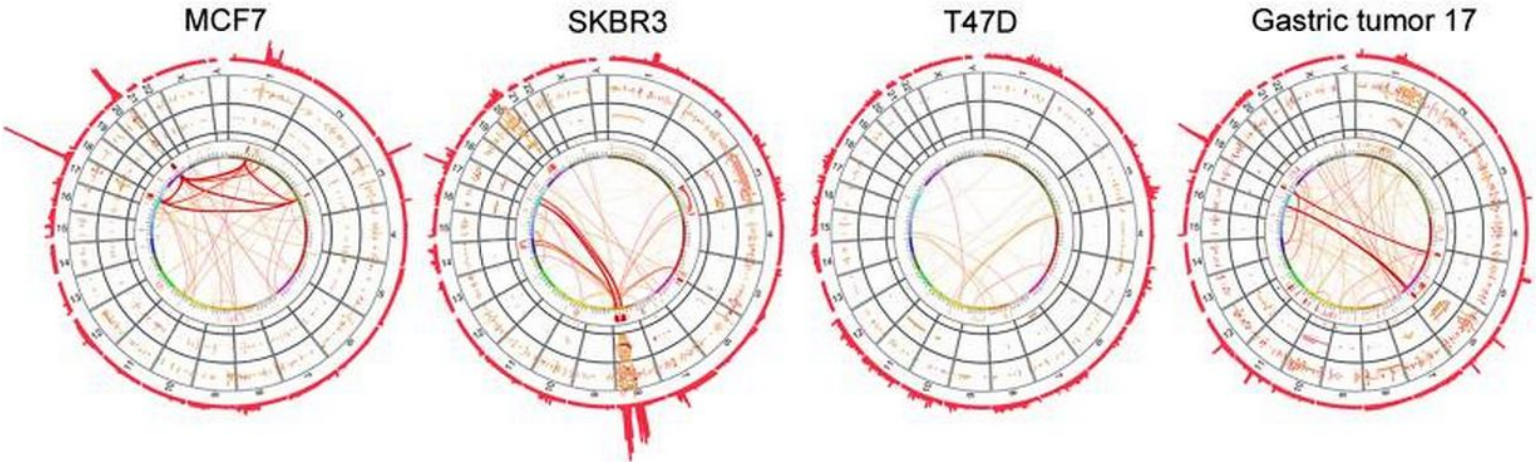


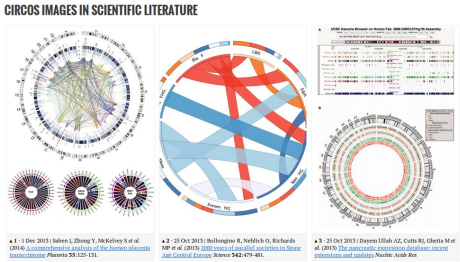


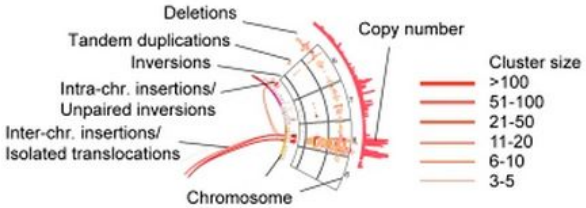
Cancer: A clonal disease evolving in a linear fashion?



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.

Kevin Davies
@KevinADavies
Follow

Jay Flatley (@illumina): In 2004, we introduced a platform that could analyze 1,536 SNPs simultaneously #PMWC14

5:32 PM - 27 Jan 2014

1 FAVORITE

Kevin Davies
@KevinADavies
Follow

Flatley: The first NGS platform, 454, was bought by Roche in 2007 and closed down 6 years later. #PMWC14

5:34 PM - 27 Jan 2014

Kevin Davies
@KevinADavies
Follow

Flatley: in 2007, it took 3 days to generate 1 gigabase data. Today, it takes 2.4 minutes. #pmwc14

5:38 PM - 27 Jan 2014

3 RETWEETS

Kevin Davies
@KevinADavies
Follow

Flatley: large population genomic datasets need to be shared with researchers and clinicians. Illumina supports these goals #PMWC14

5:40 PM - 27 Jan 2014

9 RETWEETS 2 FAVORITES

Analysing chromosome cross-talks in three dimensions

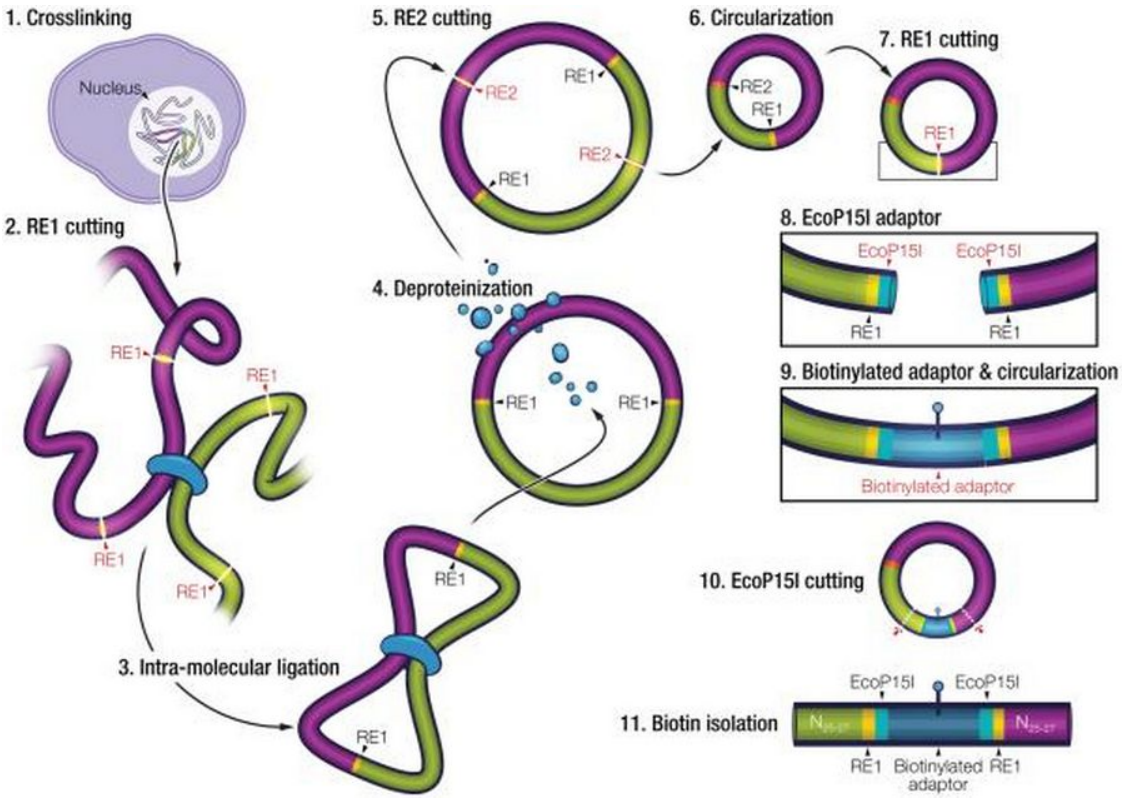
Box 1 | **3C-based methods**

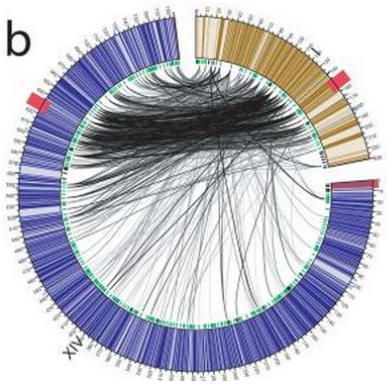
a **3C: converting chromatin interactions into ligation products**

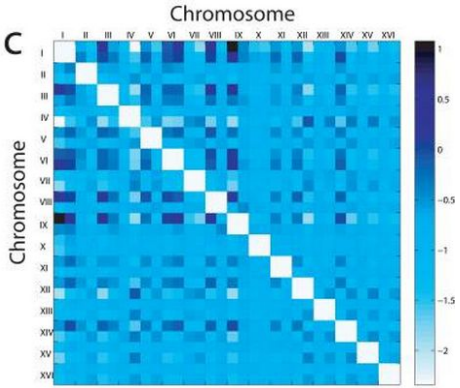
b **Ligation product detection methods**

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

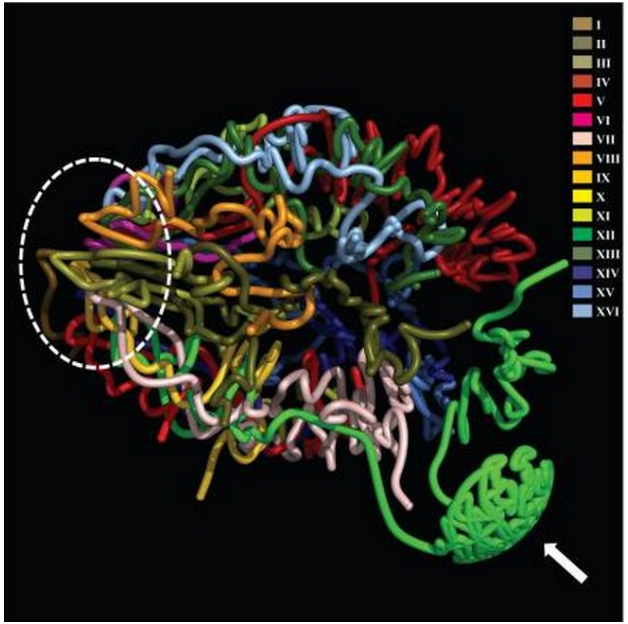
Some application: 3D architecture of the genome (yeast)

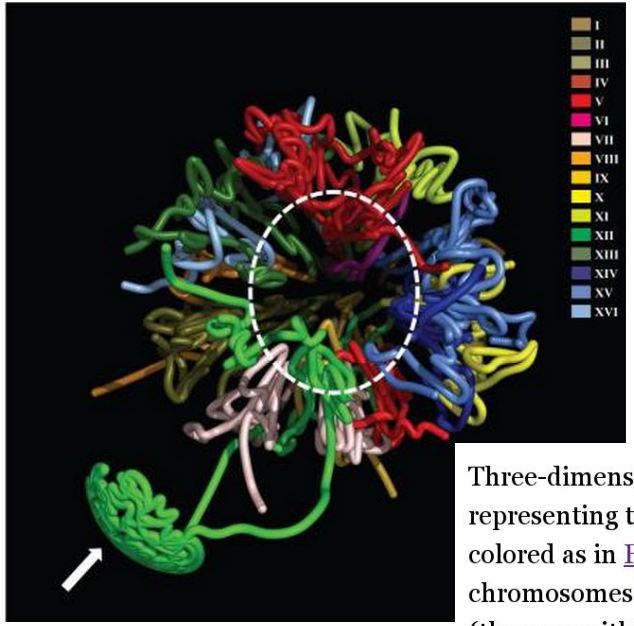






Some application: 3D architecture of the genome (yeast)





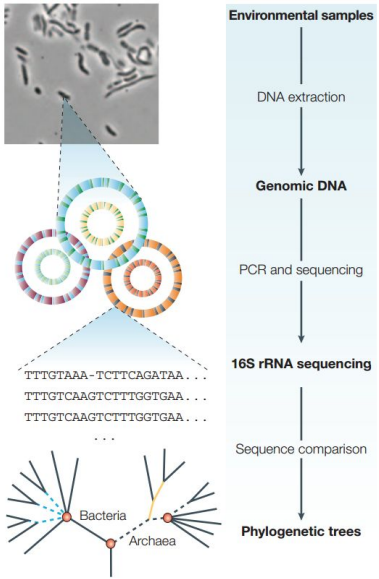
A Three-Dimensional Model of the Yeast Genome

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.

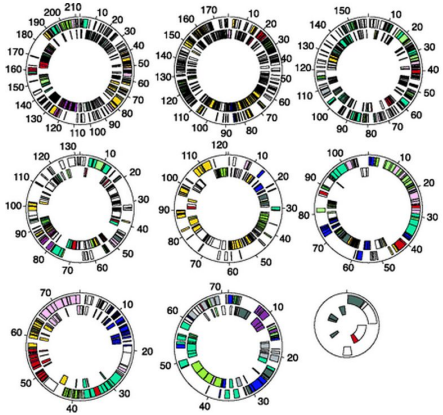
Some application of DNA Sequencing: Metagenomics



Metagenomics: DNA sequencing of environmental samples

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

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

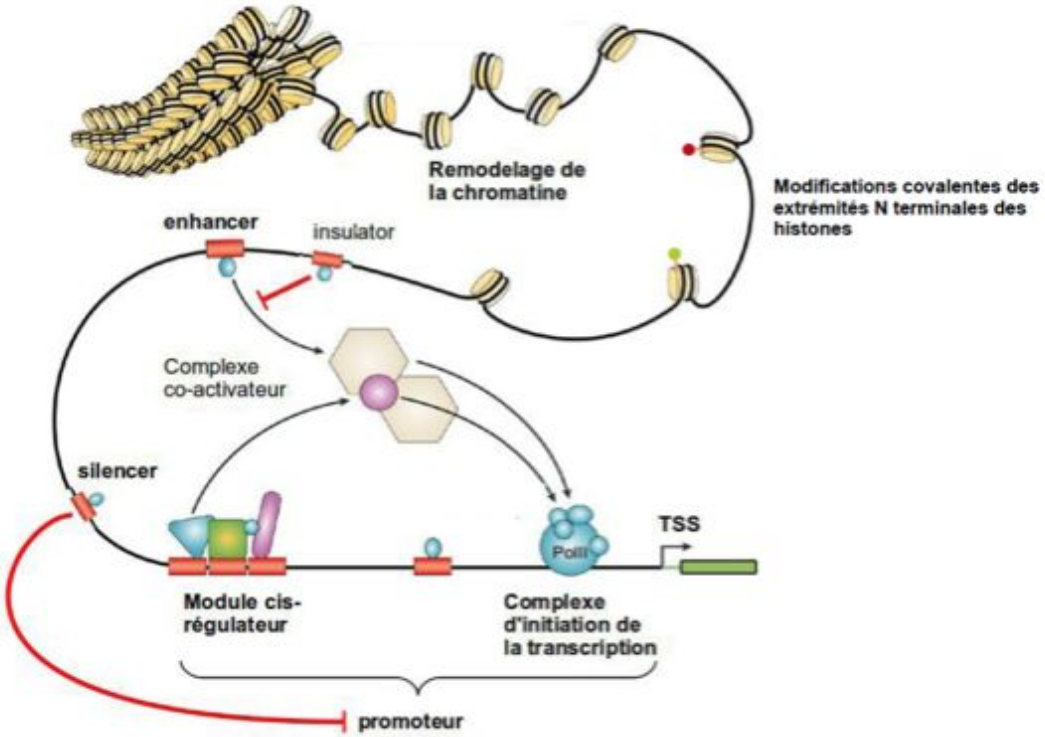


Science. 2004 Apr 2;304(5667):66-74. Epub 2004 Mar 4.

Environmental genome shotgun sequencing of the Sargasso Sea.

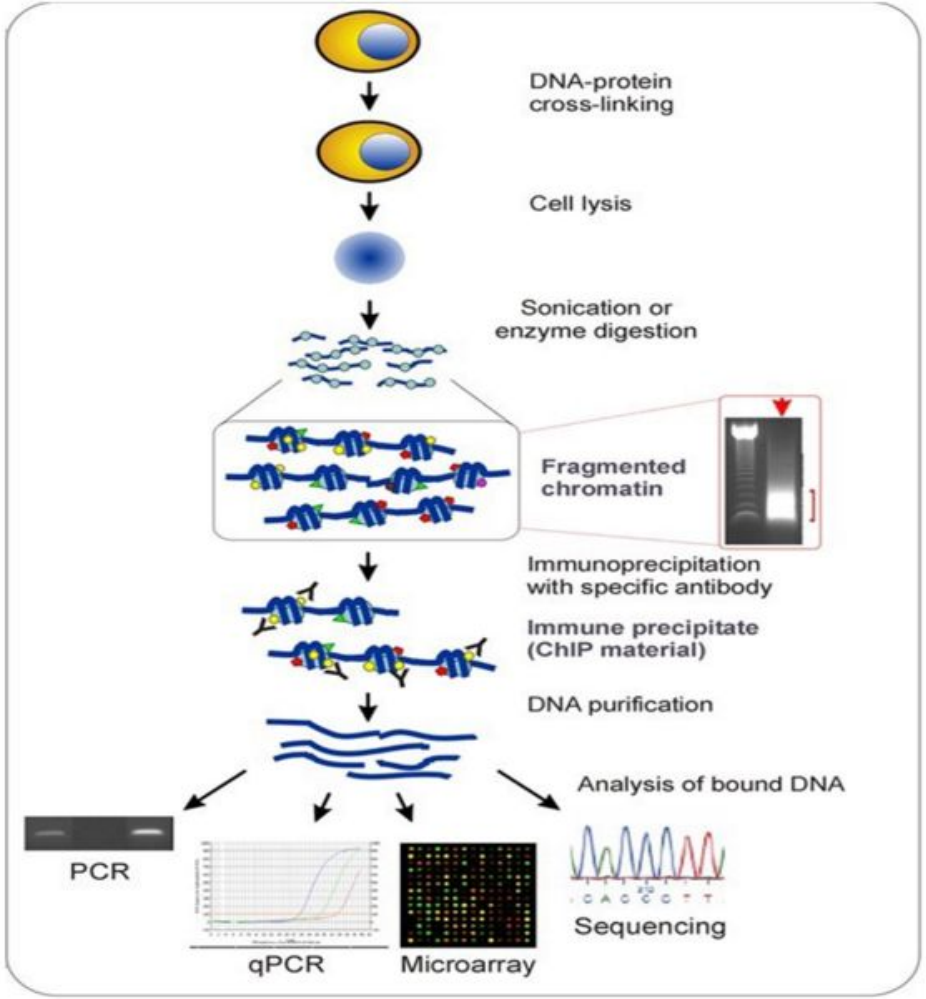
Venter JC[1], Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO.

Sequencing to detect regulatory elements



The ENCODE project

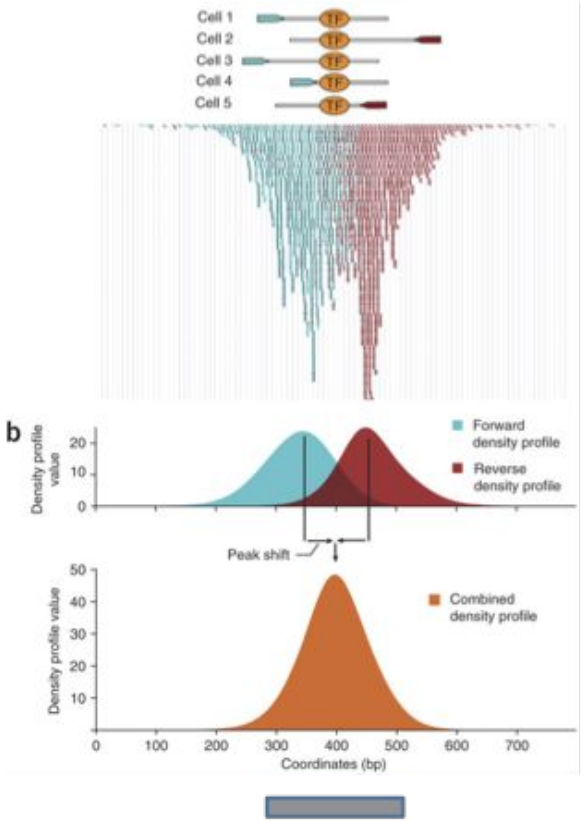
- The National **Human Genome Research Institute** (NHGRI) launched a public research consortium in 2003
  - **ENCODE**, the **Encyclopedia Of DNA Elements**
    - objective: carry out a project to identify **all functional elements** in the human genome sequence.
    - Lots of experiments rely on ChIP-Seq and RNA-Seq.



ChIP-Seq principle

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

ChIP-Seq analysis (in brief…)

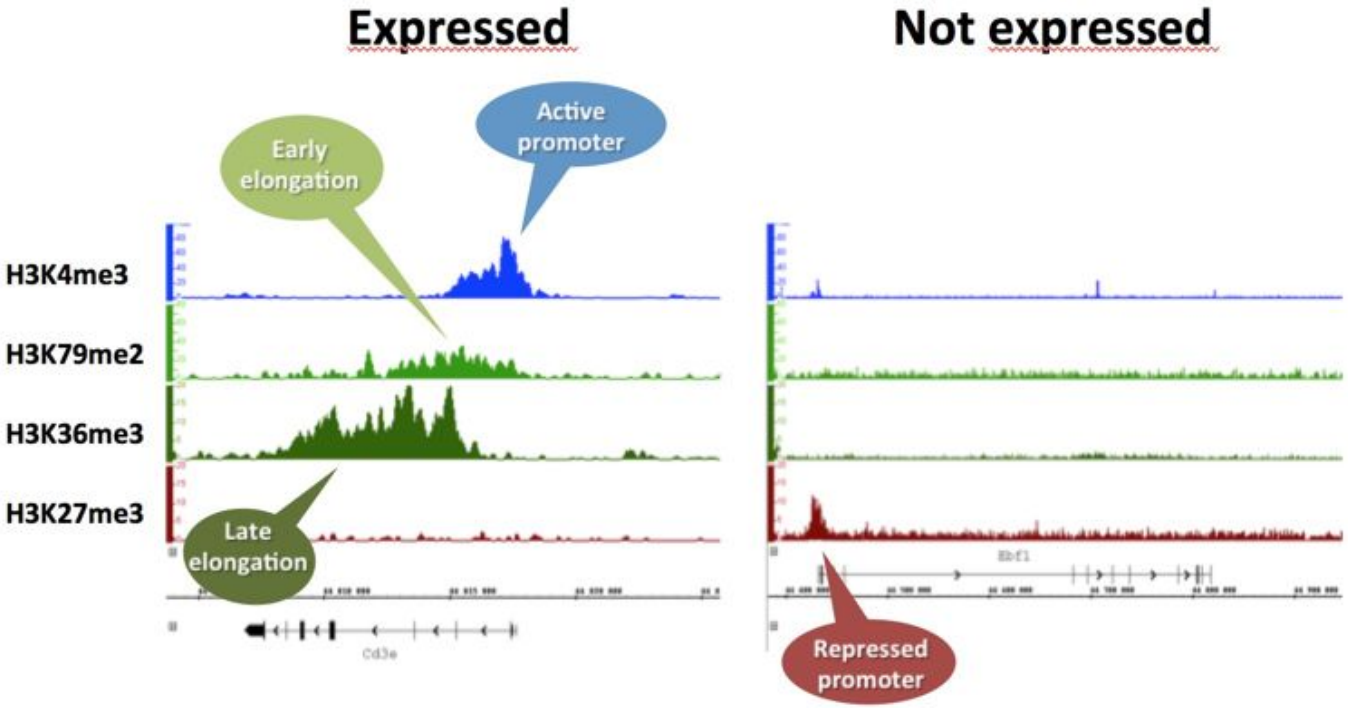


Aligned reads

Binding profile

Binding Peak

Epigenetic modification on histones

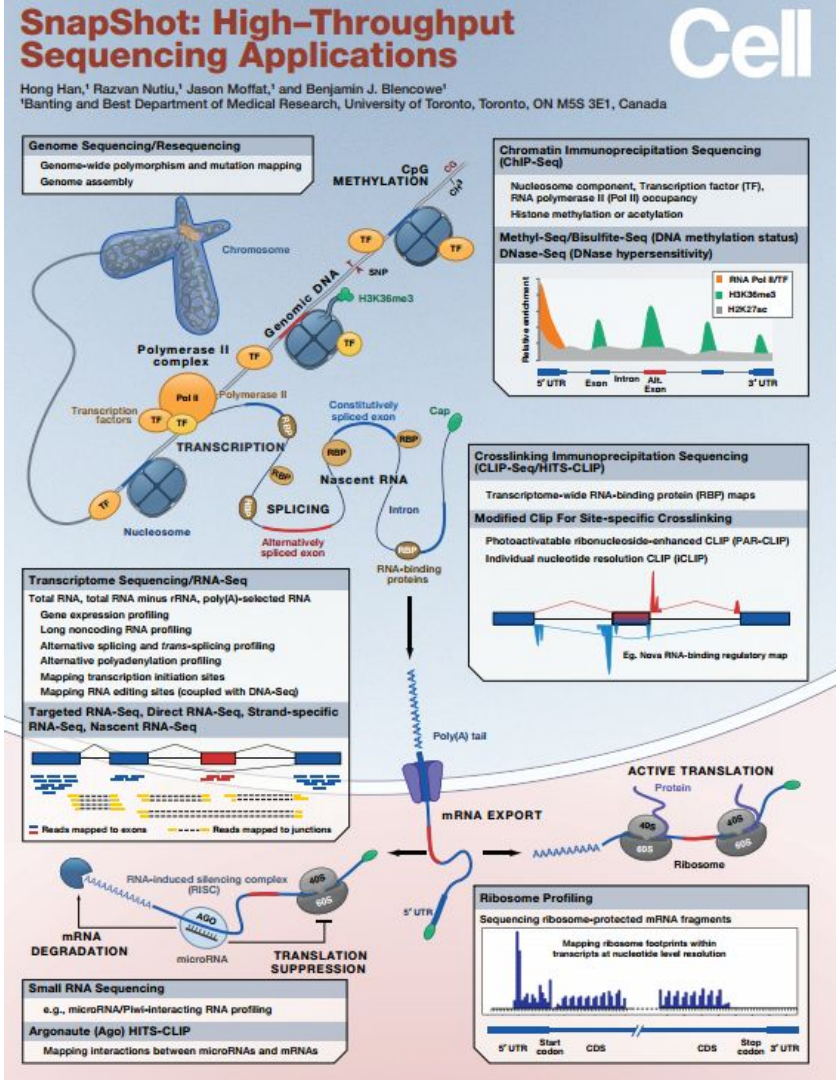


Application of ChIP-Seq

- 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 associated



Nucleosome-positioning, Ribosome profiling, ...

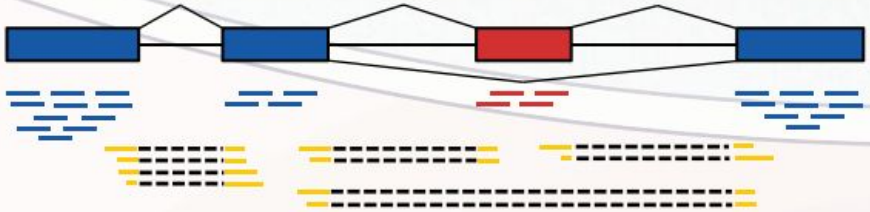
Transcriptome analysis

Transcriptome Sequencing/RNA-Seq

Total RNA, total RNA minus rRNA, poly(A)-selected RNA

- Gene expression profiling
- Long noncoding RNA profiling
- Alternative splicing and *trans*-splicing profiling
- Alternative polyadenylation profiling
- Mapping transcription initiation sites
- Mapping RNA editing sites (coupled with DNA-Seq)

Targeted RNA-Seq, Direct RNA-Seq, Strand-specific RNA-Seq, Nascent RNA-Seq



Small RNA Sequencing

e.g., microRNA/Piwi-interacting RNA profiling

Argonaute (Ago) HITS-CLIP

Mapping interactions between microRNAs and mRNAs

And many others...

Merci