

LSM2241

Personal genomics and the future of bioinformatics

Greg Tucker-Kellogg
dbsgtk@nus.edu.sg

4 November 2015

Outline

Trends driving bioinformatics into the future

Past

Present

Future

Topic

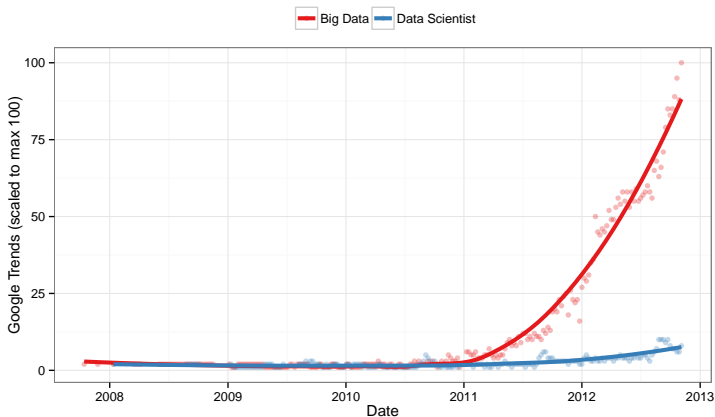
Trends driving bioinformatics into the future

Past

Present

Future

Bioinformatics, and modern biology, are becoming sciences of "big data analysis"



Epigenetics: the next layer of genomic information

- Drew Berry animation X inactivation and epigenetics
- Drew Berry animation Molecular visualizations of DNA machinery

The ENCODE project (1)

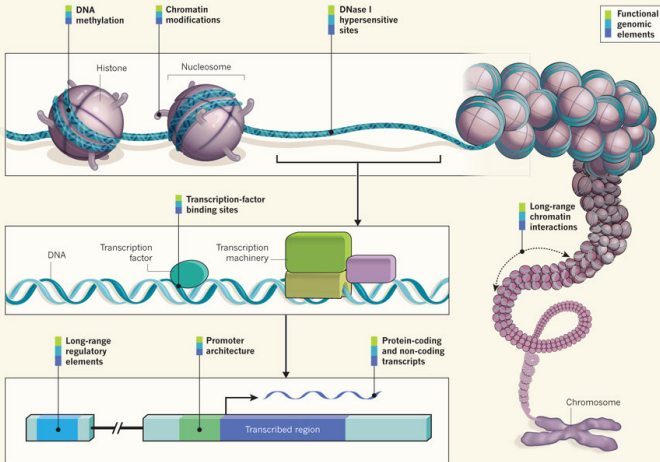
Definition (ENCODE)

- Encyclopedia of DNA Elements, to determine a map of all functional elements of the human genome
- 60 articles ([link](#)) published simultaneously in September 2012
- Integrated data available for the genome
<https://genome.ucsc.edu/ENCODE/>

modENCODE: ENCODE for model organisms ([link](#))

- *D. melanogaster*
- *C elegans*
- others

The ENCODE project (2)



from Ecker et al. 2012

Where are we headed?

Definition (Personal Genomics)

The sequencing, analysis and use of an *individual* genome for medical decisions, lifestyle choices, or self-knowledge

Timeline

Today (commercial)	personal genotyping
Today (available)	personal exome sequencing
Today (expensive)	Personal whole genome sequencing
Near future	Affordable whole genome sequencing

Why does it matter?

Today

- A variety of genetic tests are available for use with your physician to predict risk or manage care
- Generally each test is handled individually

By your graduation

- You will be able to have your *entire* genome sequenced for \ll \$1000
- The raw data will probably be discarded
- You may not want to know what the data means
- If you do want to know, you may not be able to get counselling

Will we just uncover the obvious?

©Cartoonbank.com



"We think it has something to do with your genome"

Topic

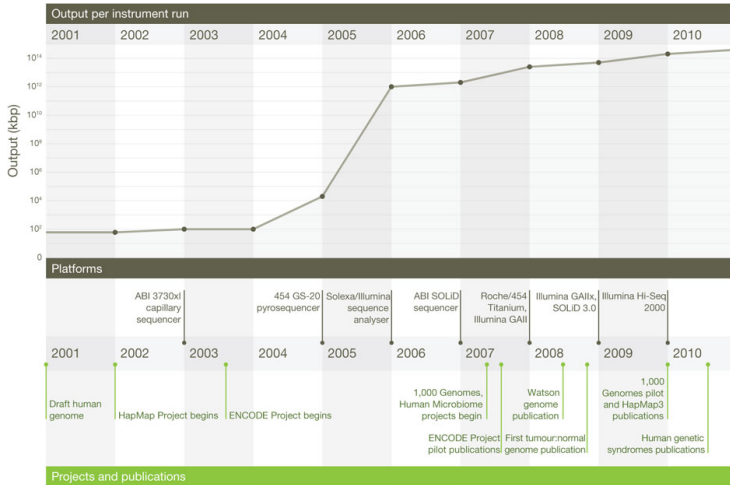
Trends driving bioinformatics into the future

Past

Present

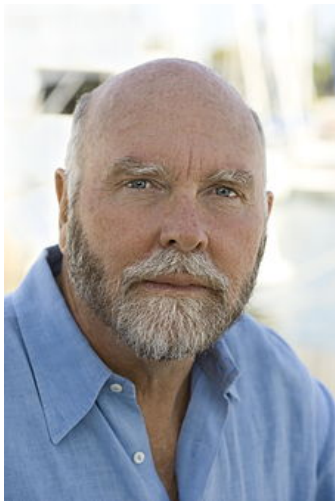
Future

Towards personal genomics



Mardis 2011

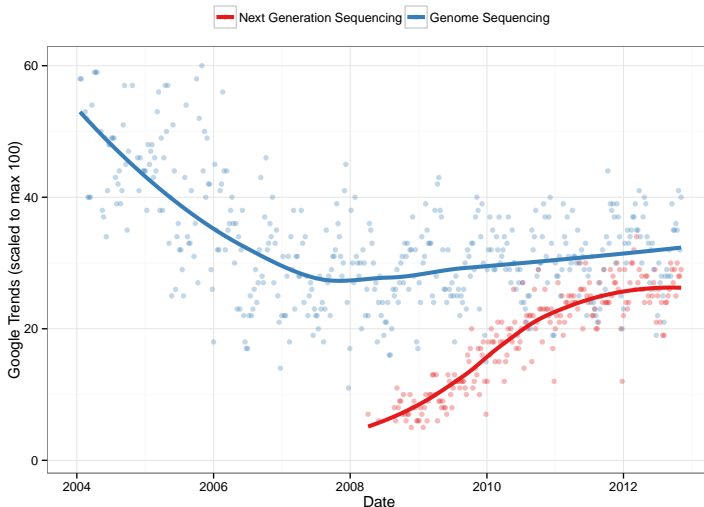
The first individual human genome



J. Craig Venter

- Founder, Celera Genomics (private human genome project)
- Proponent, whole genome shotgun sequencing
- Whole genome shotgun sequencing used for human genome assembly by Celera
- Personal genome sequenced, 2007, using Sanger sequencing

The emergence of Next Generation Sequencing (NGS)



Next generation sequencing makes things different

Pyrosequencing (454) emulsion-based PCR, picolitre wells, and luciferase sequence detection

Illumina (formerly Solexa) Reversible dye termination. DNA is attached to a glass slide, amplified to form a local colony, and image-based readouts of the slide

SOLiD Sequencing by ligation. The conceptual strategy here is a little different, because it is using ligation instead of extension.

Ion semiconductor (Ion Torrent) Very novel detection system, no optics. Benchtop sequencers

DNA Nanoballs (Complete Genomics) Rolling circle amplification, followed by sequencing by ligation

Single molecule methods No time for details

Second individual human genome



James D. Watson

- Co-discoverer of the double helix, 1953
- Nobel Prize in Physiology and Medicine, 1962
- Professor at Harvard, 1956–1976
- Head, Human Genome Project, 1990–1992
- President, Cold Spring Harbor Laboratories, 1968–2007
- Personal genome sequenced, April 2008

Watson's rationale

"I am putting my genome sequence on line to encourage the development of an era of personalized medicine, in which information contained in our genomes can be used to identify and prevent disease and to create individualized medical therapies"

3rd, 4th, 5th, and 6th

- Han Chinese individual (2008, > 30 fold coverage,)
- Male Yoruban, Nigeria (2008, 30 fold coverage, 35 bp reads)
- Acute Myeloid Leukemia patient (2008, cytogenetically normal cancer)
- Korean (2009, Genome Research)

Topic

Trends driving bioinformatics into the future

Past

Present

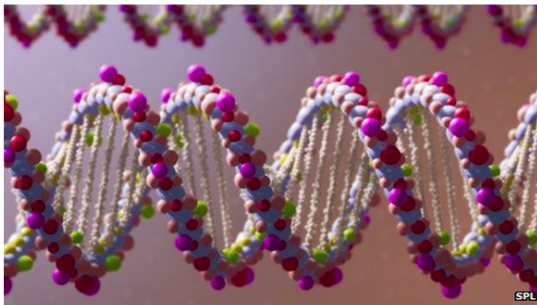
Future

Looking for volunteers

Massive DNA volunteer hunt begins

By James Gallagher

Health and science reporter, BBC News



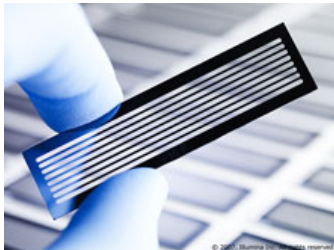
Scientists are looking for 100,000 volunteers prepared to have their DNA sequenced and published online for anyone to look at.

Related Stories

7 November 2013

<http://www.bbc.co.uk/news/health-24834375>

Illumina flow cells

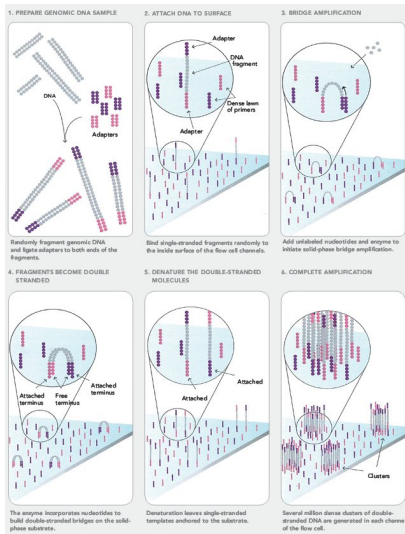


Why Illumina

- Is the most widely used NGS technology (>80% of all DNA sequence data ever generated)
- Size of a microscope slide
- Millions of sequences generated in a single run

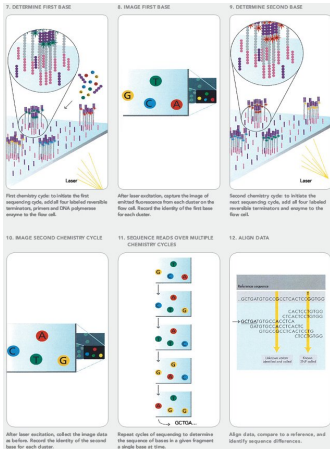
How does it work?

Illumina: from DNA to prepared slides



- DNA is fragmented and ligated to adapters
- PCR amplify in a lawn of adapter sequences
- The lawn creates “bridges” so a cluster of identical sequences appear where each molecule of DNA started

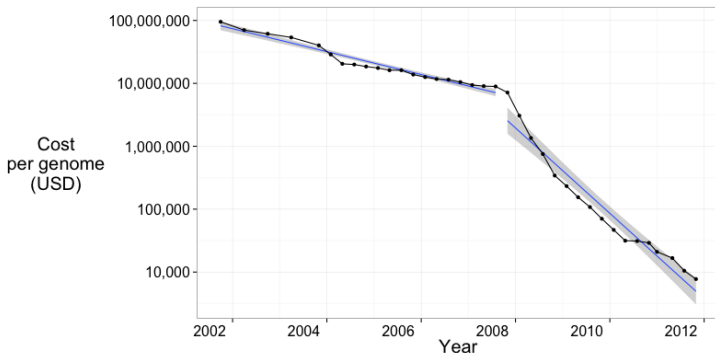
Illumina: from slide to sequence



- Each cluster yields its sequence
- Every time a new base is added, it is color coded
- A single image of the slide gives one base addition
- A single run of a single machine can yield tens of millions of sequence reads

<http://www.youtube.com/watch?v=womKfikWlxM>

The technology provided by NGS is an *unprecedented trend*



Data from US National Human Genome Research Institute

What are some consequences?

Science

- In 2008, an individual genome could get a paper in Science or Nature
- In 2010, a tumour/normal genome pair could do the same
- In 2011, papers with >50 individual genomes are being published
- by end of 2012, 1,000–10,000 individual genomes will have been sequenced!

Data

- The “1000 Genomes Project” deposited more data into GenBank in 6 months than total of past 30 years
- It is faster to send NGS data by mail than by using the internet
- GenBank has stopped archiving NGS short sequence reads
- It will soon be cheaper to regenerate raw sequence data than to store it on hard disk

Personal genomics genotyping services

Pathway Genomics Tests hundreds of genetic markers for disease risk, carrier status, and drug response

23andMe SNP Array of about 1,000,000 SNPs

- As cheap as US\$399 flat price, or \$99 plus 1 year subscription.
- Ancestry painting, relative finder

NaviGenics Health-focused. No relative finding, but focused on prediction and risk

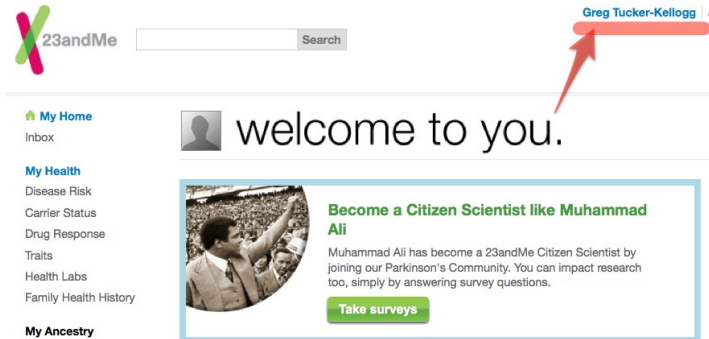
Promethease Free program (\$2 if you want it to run fast!) to analyze SNP data from genotyping sources

Personal whole genome sequencing

A variety of companies are gearing up for the expected rush for personal genomics of entire genomes

- Knome offers whole genome sequencing for \$39,500 (USD)
- BGI has provided whole exome sequencing for one bioinformaticist¹ for \$999 (USD)
- Illumina, Complete Genomics, and others are offering whole genome sequencing with prescription and physician involvement.

Who does this?



The image is a screenshot of a 23andMe user profile page. At the top left is the 23andMe logo. To its right is a search bar with the word "Search" inside. In the top right corner, the user's name "Greg Tucker-Kellogg" is displayed in blue text, with a red arrow pointing to it from below. Below the name is a red horizontal bar. On the left side of the page, there is a sidebar with several menu items: "My Home" (with a house icon), "Inbox", "My Health" (with a heart icon), "Disease Risk", "Carrier Status", "Drug Response", "Traits", "Health Labs", "Family Health History", and "My Ancestry" (with a DNA helix icon). The main content area features a large "welcome to you." message next to a placeholder profile picture. Below this, there is a promotional banner for "Become a Citizen Scientist like Muhammad Ali". The banner includes a circular image of Muhammad Ali, text stating that he has become a 23andMe Citizen Scientist by joining the Parkinson's Community, and a green button labeled "Take surveys".

23andMe

Search

Greg Tucker-Kellogg

My Home

Inbox

My Health

Disease Risk

Carrier Status

Drug Response

Traits

Health Labs

Family Health History

My Ancestry

welcome to you.

Become a Citizen Scientist like Muhammad Ali

Muhammad Ali has become a 23andMe Citizen Scientist by joining our Parkinson's Community. You can impact research too, simply by answering survey questions.

Take surveys

A lot of people!

Topic

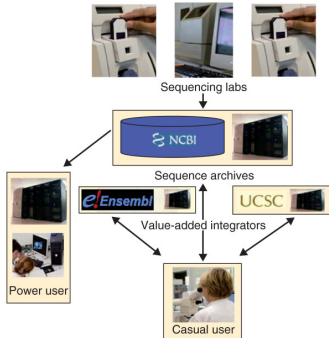
Trends driving bioinformatics into the future

Past

Present

Future

Can we keep doing bioinformatics the same way?

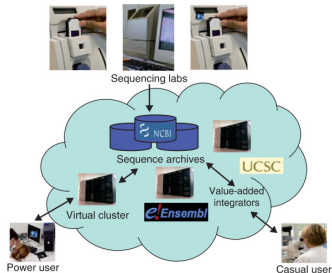


Stein, L.D., 2010. The case for cloud computing in genome informatics. *Genome biology*, 11(5), p.207

The traditional model

- Sequencing labs generate data, send to collaborators and customers
- Data sets uploaded to central databases at NCBI, EMBL, etc.
- Other databases (UCSC, Ensembl) integrate disparate data
- Users access the data from distant locations

Moving it to the cloud



Stein, L.D., 2010. The case for cloud computing in genome informatics. *Genome biology*, 11(5), p.207

A cloud-based model

- It is too costly, and too slow, to download data
- Huge amounts of data are available "on the cloud", like you have with Dropbox, Evernote, and Amazon
- Most data will never be downloaded, but will be analyzed "up there"

Personal genomics will change your health care

- Personal genomics will encourage behaviour modification for improved health
- The Internet of things will allow you to passively monitor your behaviour
- Information about your behaviour will be shared with you and with your doctor
- Social media will enable communities of tools to accelerate development
- Are you the customer or the product of these developments?

An example of the Internet of Things



GlowCaps

- Internet-aware medicine bottle caps
- Fits common prescription bottles in the US
- Promotes "compliance" with medication

How does this work?



Internet-aware bottle caps

- Flash when it is time for your meds
 - ▶ play a ring tone
 - ▶ telephone you
- Weekly email to friend or family member
- Monthly report to your doctor and to you
 - ▶ Medicine becomes cheaper if you are more compliant
- Automatically orders refills from the pharmacy

The crowdsourcing of personal genomics

The prospect of personal genomics has led to “open source” projects driven by people making their own genetic information public

- The **Personal Genome Project** Led by George Church at Harvard, aiming to share 100,000 individual genomes for research
- The Quantified Self movement
- **CureTogether**. Aligned with the Quantified Self movement, self-reporting conditions and tracking. Not currently genomic-centred
- Many, many other projects

Is this my (our) coming life?



**why aren't you
taking your medicine?**

References



Ecker, Joseph R et al. (2012). “Genomics: ENCODE explained.” In: *Nature* 489.7414, pp. 52–5. DOI: [10.1038/489052a](https://doi.org/10.1038/489052a) (cit. on p. 7).



Mardis, Elaine R (2011). “A decade’s perspective on DNA sequencing technology.” In: *Nature* 470.7333, pp. 198–203. DOI: [10.1038/nature09796](https://doi.org/10.1038/nature09796) (cit. on p. 12).