Big data, Genomics and Public Health:
Big Data meets DNA

Winston Hide, Harvard School of Public Health and Harvard Stem Cell Institute

Critical Data - Secondary use of Big Data from Critical Care - January 7 2014
Bioinformatics

“We believe the field of bioinformatics for genetic analysis will be one of the biggest areas of disruptive innovation in life science tools over the next few years”

Goldman Sachs
Big Data - Quantification

Science, sports, advertising and **public health**

— a drift toward data-driven discovery and decision-making.

“**It’s a revolution. We’re really just getting under way. But the march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government. There is no area that is going to be untouched.**”

- **Gary King**, director of Harvard’s Institute for Quantitative Social Science
Scaling genomes
Yikes: Peter Tonellato (BIDMC): The sequenced genome of every patient in the world = 1 Yottabyte (10 to 24 power). Big data indeed. #Tricon

Data management
Heavy lifting

Next-gen sequencing/pipelining

Analytics for functional significance

Research computing
Rapid adoption of NGS brings new challenges
"And that’s why we need a computer."
The new shared knowledge that doesn’t fit into one domain

How to recognise expertise?
Can you train a post doc in my group?
Why can’t you do this faster than the Broad?
Why does this cost so much?
Public discourse
Will I get cancer?
Informed consent
How to understand my genome data?
Angelina Jolie's BRCA1 gene and her decision to have a double mastectomy. It's hard not to be a bit unnerved. I've always thought it better to have information than not have it, and that, anyway, you largely know from your family history what nasty diseases are floating around your gene pool.

What happened when I had my genome sequenced

It is 10 years since the human genome was first sequenced. In that time the cost per person has fallen from $2.7bn to just $5,000. Revealing our full DNA will revolutionise medicine – but it will also raise huge ethical questions about what we do with the information ...
Genome variants
OTC genome variation analysis
Personal Incentive to genome variants
The Case of Kira Peikoff
Do OTC test work? (NYT)

- 240 health conditions and traits - 23&me
- 25 disease risks - Genetic Testing Laboratories
- 24 disease risks - Pathway Genomics
  - Psoriasis *Elevated* 20%
  - Psoriasis *Decreased* 2%
Predicting health

- source of variation in the *estimates of average risks*

- **Sampling Error**
  - SNPs vary between each platform
  - Tiny fraction of all variants
  - Role of most genes currently unknown in disease
Scalable Tools for Open-source Read Mapping

Upload your data
Create an S3 bucket as described on the STHRMSeq page and upload your files.

Enter your AWS Account Number:

Enter your Access Key ID:

Enter your Secret Access Key:

Enter the Amazon S3 bucket where your data are stored:

Amazon EC2 Advanced Options

Set your parameters

Name your sample:

Genome version: hg19

dbSNP version: dbSNP 135

Alignment: BWA

Cleaning Pipeline: GATK

SNP Calling: GATK

Indel Calling

SV Calling

Advanced settings

Start processing

GO!

Note that by clicking "GO!", you are starting up Amazon instances and volumes that will be charged to your account. The mapping and variant calling for a whole genome (30X coverage) is estimated at $X for BWA and $X for SNAP. For a full exome (80X coverage), the cost is estimated at $X for BWA $X for SNAP. However, note that processing times (and thus costs) may vary.
Look It Up

> Individual GWA studies
> A handful of genes
> from 1 genome to thousands, hundreds of thousands, millions of genomes
> Power to detect associated genes increases
> Meaningful sets of genes
> the complete catalogue of disease genes within 6 years

“This is the first decade when we can actually look across diseases in this systematic way”
Examples

- Schizophrenia - nothing with 3000 cases
- 2 genes 10 000 cases
- 25 000 cases yield 62 - 93 significance loci - association with post synaptic density
Realizing the complexity

- can we define level of cellular organization above the genome?

Decompose data to identify **finite number of cellular states and response modules**

What data is needed?

*Gene expression?*

*Protein levels?*

*Post translational modifications?*

*Single cells?*

How to assess noise **accurately and to completeness**?

- Towards a catalog of functions?
Interpreting complexity

What gene is responsible?
Ok, Um, What key genes and their pathways are responsible?
Which are biomarkers?
Targets?
What drugs will be specific for targets?
Cohort/Samples for understanding disease

- “Classic” Phenotype
- Clinical/Physiological/Physical/Gene markers
- Molecular
FUNCTIONAL ASSAY

TRANSCRIPTOMICS
CHIP-SEQ
EPIGENETICS

PHENOTYPE
PHENOMICS
ASSOCIATION

TRANSCRIPTION

DISEASE MODEL

CAUSAL

INFEERENCE

PERTURBATION

VALIDATION
Practitioner-scientists?

- Learn online tools for assessment of genome variation
- Learn appropriate strategies for cohort data interpretation
- Systematically compare signals against drug association databases
- Test existing repurposed drugs against patients
Public Information

medical practitioners

professionals working in the biomedical sciences

motivated lay individuals interested in exploring their personal genetic data
Challenges
Reproducibility?

- replication does not lead to the same results
- Tests at Bayer reproduced published results in a quarter of 67 seminal studies


Florian Prinz Bayer Health Care Nature Drugs Discovery
Reproducible Genome Signals?

- SNPs
- Genes
- Networks
- Processes
- Systems
- Phenotype
Addressing the future

Data Science
Research computing
Data sharing/Publication paradigms
Sharing for Personal Genome Data

Information commons for precision medicine
Patients
Physicians
Clinics
Payers

Biomedical researchers
Computer scientists
Social scientists
Feedback to the scientists from the sample providers?

A Million Cancer Genome Warehouse
The million cancer genome warehouse

http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html
Strategies

- Vast majority of the biomedical corpus is ignored
- ‘Raw’ omics data has associated **phenotype**
  - Metadata
- Molecular profiles can be derived from raw molecular data and clustered
- Metadata can be clustered

Grouping Metadata
Make data that is similar ‘findable’ Google style.
Is peer review the answer?
Digital Object Identifier

DOI:.....

Not just for Publications any more...
Provenance
Citability

3A Commons: facilitates collection, curation and sharing of experiments in life science domains, using a common, structured representation of the experiments.
Towards interoperable bioscience data

Nature Journals
“Science Data”

Common metadata format
Searchable
Peer reviewed
Curated/Clean
Freely accessible
Computable

Look It Up
Catalogues of variants gain value as they are assimilated and constantly analyzed.
- development and suites of relevant variant tools.

Towards precision medicine: advances in computational approaches for analysis of human variants J. ol biol 2013

Maricella Kann lab
Cancer Genome Atlas
PharmGKB
LINCS/Connectivity Map

Open sharing and access
Publication paradigm
Manner of the pursuit of research
Measures of impact (resetting tenure review)
Actual innovation and progress
A focus on **functional discovery** vs **genomics observation**

Public Health Genomics
- Shareable resources for genome understanding
- LookUp systems for disease - gene - drug association
- Grouping associations together
- Networks of relationships between treatments, genes, pathways, genome variants, tissue imaging, cohort features

- **The informed physician will become a data centric specialist.**