What is Big Data?

• Information Retrieval *not* Big Data
• Computing across TB’s of data (Higgs boson) *is* Big Data
• Computing across credit card data to detect possible fraud *is* Big Data

& Diapers == *Predictive Analytics*

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?smid=pl-share
Daily Data Processing at NCBI

126 Petabytes/year

4 Terabytes (TB)/day
365TB/day

3TB/day

26TB/day

Interactive Web
Data Downloads

2012 data...
Stationary night blindness due to premature termination in TRPM1

Data for NA11918 placed by two different aligners (mosaik & bwa)

All individual genotypes for rs3784589

Nonsense C→A in TRPM1
What is the Big Data Problem in Biology?
Example: Reducing the 1000 Genome Dataset

- Computing on reads is a short term technical challenge
- Computing on growing amounts of derived data is a long term challenge & opportunity
NextGen Churn

Basic Data Processing
- Mapping reads, calling SNPs, splices, peaks, etc.

Derived Data
- Genotypes, Genes, expression levels, motifs, etc.

Etc....

Biological Questions
As cost of computers decreased, overall investments increased...

Same thing with DNA sequencing and other high-throughput biomed tools
Economics of NextGen

Grantees want to maximize impact of budget

Sequencing will continue to increase if total cost continues to drop:

sample prep + raw sequencing + IT
Big Data:

What mess?

When

BIG DATA meets Small Signal

the signal has to be teased out of the data

- There are many ways to do this: a small change in the analysis details can cause a large change in the results.
- It is too easy to distort your findings, either by fooling yourself or on purpose.
Systematic errors are routinely observed and have been reported in many papers

- Identification and correction of systematic error in high-throughput sequence data by Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David IK Martin, Meromit Singer and Lior Pachter

This paper shows the existence of systematic errors, even at high coverage, often strand dependent but not always
Mismatch profiles are dominated by ‘spikes’ occurring at particular sequencing cycles (sequencing batch effect). Furthermore, each problematic cycle has a specific limited profile of mismatch types, adding to the bias.

This effect on Illumina HiSeq is sequencing lane dependent.

SEQC neuroblastoma study, L Shi, Fischer et al
Systematic errors generate noise in the low to intermediate allele fraction (1 to 30%), making identification of true SNPs hard in that area, even at high coverage.

Histogram of candidate variant allele fraction before filtering: massive presence of SNP candidates with low to intermediate variant allele fraction values

Pacific biosciences sequencing technology for genotyping and variation discovery in human data
Mauricio O Carneiro, Carsten Russ, Michael G Ross, Stacey B Gabriel, Chad Nusbaum and Mark A DePristo
BMC Genomics 2012, 13:375
RNA-Seq often differs from ‘true’ concentration by a factor 2 or more. This affects all platforms, in particular Illumina and PGM/Proton/Solid.

The two red and green runs should be superimposed on the blue nominal concentration. Yet some specific ERCC molecules are vastly different, e.g. ERCC116 is measured 16 fold below nominal.

Data by Setterquist, LifeTech, QC for PGM
“That partly explains why observational studies in general can be replicated only 20% of the time, versus 80% for large, well-designed randomly controlled trials, says Dr. Ioannidis. Dr. Young, meanwhile, pegs the replication rate for observational data at an even lower 5% to 10%.”
Correlation between impact factor and retraction index.

Repeatability of published microarray gene expression analyses
"Preclinical research generates many secondary publications, even when results cannot be reproduced."

Table 1: Reproducibility of research findings

<table>
<thead>
<tr>
<th>Journal impact factor</th>
<th>Number of articles</th>
<th>Mean number of citations of non-reproduced articles*</th>
<th>Mean number of citations of reproduced articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;20</td>
<td>21</td>
<td>248 (range 3–800)</td>
<td>231 (range 82–519)</td>
</tr>
<tr>
<td>5–19</td>
<td>32</td>
<td>169 (range 6–1,909)</td>
<td>13 (range 3–24)</td>
</tr>
</tbody>
</table>

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.

• Biomarkers – most highly cited studies overestimated effect sizes (*JAMA* 2011; 305(21): 2200-2210)

• Faculty & Trainee survey at MD Anderson – 50% had experienced at least one case of irreproducibility (*PLOS One* 2013 May 15; 8(5))

• More first-in-class small molecule drugs approved between 1999-2008 identified by “classical” methods than genomics approaches (*J. Biomol Screen* 2013 Dec; 18(10): 1143-55)

• “Koch’s Postulates” for assigning causality between genetic variants & disease phenotypes (*Cell* 2013 Sep 26; 155(1) : 21-6)
“Cancer experts seeking to solve the problem have found that a fifth to a third or more of cancer cell lines tested were mistakenly identified— with researchers unwittingly studying the wrong cancers, slowing progress toward new treatments and wasting precious time and money.”

“...Dr. Masters, in a study of scientific papers published between 2000 and 2004, found nearly a 1,000 citations of the same contaminated cancer lines revealed in Dr. Gartler's 1966 findings, which have since been replicated many times using more advanced techniques.”
Dr. John Snow - location of pumps and cholera deaths, London, England, 1854
The Well-Traveled Salad.
Do You Know Where Your Food Has Been?

As consumers, many of us fail to recognize that even our domestic and local food supplies are part of a global network. The daily activity of consuming food directly links our health as humans to the health of crops and produce, food animals, and the environments in which they are produced.

A “One Health” approach to food safety—bringing together expertise and resources from the clinical, veterinary, wildlife health, and ecology communities—has the potential to reveal the sources, pathways, and factors driving the outbreaks of foodborne illness and possibly prevent them from occurring in the first place.

NOTE: Countries are listed in alphabetical order and not by volume of export.
Foodborne Illness

- ~48 million US cases annually (CDC)
  - 128,000 hospitalized
  - 3,000 deaths
- Trends show little evidence of progress
DNA Forensics...

A Pathogen Genome Is The Fingerprint

Functional prediction can be developed and refined more slowly from this base.
Search for phylogenetic signal at the level of SNPs
Why do we need WGS? To Shift the Paradigm from a “low resolution” Public Health Approach to A Real-Time “high resolution” Approach

Clinical ID and fingerprint

Identify Food and confirm PFGE Fingerprint

CDC

FDA/FSIS

Product enters commerce

Source of contamination identified too late
Current Genometrakr network
7 state Laboratories + 11 FDA-ORA

Network of Sequencers
First use of Genometrakr network and WGS approach to support regulatory action and positive public health outcome in real-time

Listeria monocytogenes project with CDC, FDA, NCBI, USDA March 2014

> 6 SNPs

Isolates from cheese facility, distributed product, and patients who consumed product
An increased degree of certainty that comes with matching strains of pathogens through whole genome sequencing allowed for detection of this Salmonella contamination event in nut butter across several states with low level contamination and a widely distributed product. In this case, WGS identifies the link and preempts an outbreak even w/o availability of food - it informs the epidemiology and our inspectors.

Outbreak/Pre-outbreak summer of 2014
Health and Economic Impact of Active WGS-based Surveillance

Comparison of Nspired and Sunland contamination events.

- Similar facilities – broad domestic distribution.
- Sunland 42 cases and 10 hospitalizations with as many as 1,260 illnesses unreported (Fall 2012)
- Nspired – 4 confirmed cases, 1 hospitalization (Summer 2014)
- WGS informed investigation prevented significant illness and hospitalizations
  - lower illness rate and treatment cost ($3000-$9000) + fraction of longterm and chronic onset complications associated with Salmonella infection (ie, Reiter’s syndrome, GBS)
Clinical ID WGS in real-time and in parallel food and environmental WGS

FDA, CDC, FSIS, States

Source of contamination identified through WGS combined database queries

Food enters commerce
National Digital Immune system: Big Data?

• WGS + metadata

• PulseNet - ~50,000 isolates per year + Hospital based infections (AMR strains) 50-500K per year (??) + Environmental samples

• Ecology & population genetics of pathogens...