The next 10 years of quantitative biology

Michael Schatz

March 25, 2014
Keystone Meeting on Big Data in Biology

@mike_schatz / #KSBigData
Unsolved Questions in Biology

- What is your genome sequence?
- How does your genome compare to my genome?
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?
- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- Where do proteins bind and regulate genes?
- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- What drugs should we give you?
- Plus hundreds and hundreds more
Quantitative Biology Technologies

Results
Domain Knowledge

Machine Learning
classification, modeling, visualization & data Integration

Scalable Algorithms
Streaming, Sampling, Indexing, Parallel

Compute Systems
CPU, GPU, Distributed, Clouds, Workflows

IO Systems
Hardrives, Networking, Databases, Compression, LIMS

Sensors & Metadata
Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies
Quantitative Biology Technologies

- Sensors & Metadata
  - Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

- IO Systems
  - Hardrives, Networking, Databases, Compression, LIMS

- Compute Systems
  - CPU, GPU, Distributed, Clouds, Workflows

- Scalable Algorithms
  - Streaming, Sampling, Indexing, Parallel

- Machine Learning
  - classification, modeling, visualization & data Integration

- Results
  - Domain Knowledge

- Domain Knowledge

- Visualizations & Data Integration

- Scalable Algorithms
  - Streaming, Sampling, Indexing, Parallel

- Compute Systems
  - CPU, GPU, Distributed, Clouds, Workflows

- IO Systems
  - Hardrives, Networking, Databases, Compression, LIMS

- Sensors & Metadata
  - Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies
Cost per Genome

http://www.genome.gov/sequencingcosts/
Worldwide capacity exceeds 15 Pbp/year
25 Pbp/year as of Jan 15

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com
How much is a petabyte?

<table>
<thead>
<tr>
<th>Unit</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte</td>
<td>1</td>
</tr>
<tr>
<td>Kilobyte</td>
<td>1,000</td>
</tr>
<tr>
<td>Megabyte</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Gigabyte</td>
<td>1,000,000,000</td>
</tr>
<tr>
<td>Terabyte</td>
<td>1,000,000,000,000</td>
</tr>
<tr>
<td>Petabyte</td>
<td>1,000,000,000,000,000,000</td>
</tr>
</tbody>
</table>

*Technically a kilobyte is $2^{10}$ and a petabyte is $2^{50}$*
How much is a petabyte?

100 GB / Genome
4.7 GB / DVD
~20 DVDs / Genome

X

10,000 Genomes

= 

1 PB Data
200,000 DVDs

787 feet of DVDs
~1/6 of a mile tall

500 2 TB drives
$500k
DNA Data Tsunami

*Current world-wide sequencing capacity is growing at ~3x per year!*
DNA Data Tsunami

Current world-wide sequencing capacity is growing at ~3x per year!

~1 exabyte by 2018

~1 zettabyte by 2024
## How much is a zettabyte?

<table>
<thead>
<tr>
<th>Unit</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte</td>
<td>1</td>
</tr>
<tr>
<td>Kilobyte</td>
<td>1,000</td>
</tr>
<tr>
<td>Megabyte</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Gigabyte</td>
<td>1,000,000,000</td>
</tr>
<tr>
<td>Terabyte</td>
<td>1,000,000,000,000</td>
</tr>
<tr>
<td>Petabyte</td>
<td>1,000,000,000,000,000</td>
</tr>
<tr>
<td>Exabyte</td>
<td>1,000,000,000,000,000,000,000</td>
</tr>
<tr>
<td>Zettabyte</td>
<td>1,000,000,000,000,000,000,000,000</td>
</tr>
</tbody>
</table>
How much is a zettabyte?

100 GB / Genome
4.7 GB / DVD
~20 DVDs / Genome

X

10,000,000,000 Genomes

= 

1 ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

Both currently ~100 Pb
But growing exponentially
Sequencing Centers

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com
Sequencing Centers

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com
Biological Sensor Network

The rise of a digital immune system

(@ewanbirney)  (@latimes)
Data Production & Collection

Expect massive growth to sequencing and other biological sensor data over the next 10 years
- Exascale biology is certain, zettascale on the horizon
- Compression helps, but need to aggressively throw out data
- Requires careful consideration of the “preciousness” of the sample

Major data producers concentrated in hospitals, universities, agricultural companies, research institutes
- Major efforts in human health and disease, agriculture, bioenergy

But also widely distributed mobile sensors
- Schools, offices, sports arenas, transportations centers, farms & food distribution centers
- Monitoring and surveillance, as ubiquitous as weather stations
- The rise of a digital immune system?
Quantitative Biology Technologies

Results
Domain Knowledge

Machine Learning
classification, modeling, visualization & data Integration

Scalable Algorithms
Streaming, Sampling, Indexing, Parallel

Compute Systems
CPU, GPU, Distributed, Clouds, Workflows

IO Systems
Hardrives, Networking, Databases, Compression, LIMS

Sensors & Metadata
Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies
Sequencing Centers
The DNA Data Deluge
Informatics Centers

The DNA Data Deluge
Parallel Algorithm Spectrum

<table>
<thead>
<tr>
<th>Embarrassingly Parallel</th>
<th>Loosely Coupled</th>
<th>Tightly Coupled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Computing</td>
<td>MapReduce</td>
<td>Graphs &amp; MD simulations</td>
</tr>
<tr>
<td>Each item is Independent</td>
<td>Independent-Sync-Independent</td>
<td>Constant Sync</td>
</tr>
</tbody>
</table>
MUMmerGPU
http://mummergpu.sourceforge.net

- Index reference using a suffix tree
  - Each suffix represented by path from root
  - Reorder tree along space filling curve

- Map many reads simultaneously on GPU
  - Find matches by walking the tree
  - Find coordinates with depth first search

- Performance on nVidia GTX 8800
  - Match kernel was ~10x faster than CPU
  - Search kernel was ~4x faster than CPU
  - End-to-end runtime ~4x faster than CPU

- Cores are only part of the solution.
- Need storage, fast IO
- Locality is king

High-throughput sequence alignment using Graphics Processing Units.
Crossbow
http://bowtie-bio.sourceforge.net/crossbow

• Align billions of reads and find SNPs
  – Reuse software components: Hadoop Streaming
  – Mapping with Bowtie, SNP calling with SOAPsnp

• 4 hour end-to-end runtime including upload
  – Costs $85; Today's costs <$30

• Very compelling example of cloud computing in genomics
• Transfer takes time, but totally depends on institution
• Need more applications!

Searching for SNPs with Cloud Computing.
Pan-Genome Alignment & Assembly

Time to start considering problems for which N complete genomes is the input to study the “pan-genome”
- Available today for many microbial species, near future for higher eukaryotes

Pan-genome colored de Bruijn graph
- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?
- What are the pan-genome network properties?

Rapid pan genome analysis with augmented suffix trees
Marcus, S, Schatz, MC (2014) *In preparation*
Compute & Algorithmic Challenges

Expect to see many dozens of major informatics centers that consolidate regional / topical information
• Clouds for Cancer, Autism, Heart Disease, etc
• Plus many smaller warehouses down to individuals
• Move the code to the data

Parallel hardware and algorithms are required
• Expect to see >1000 cores in a single computer
• Compute & IO needs to be considered together
• Rewriting efficient parallel software is complex and expensive

Applications will shift from individuals to populations
• Read mapping & assembly fade out
• Population analysis and time series analysis fade in
• Need for network analysis, probabilistic techniques
Quantitative Biology Technologies

1. Sensors & Metadata
   - Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies
2. IO Systems
   - Hardrives, Networking, Databases, Compression, LIMS
3. Compute Systems
   - CPU, GPU, Distributed, Clouds, Workflows
4. Scalable Algorithms
   - Streaming, Sampling, Indexing, Parallel
5. Machine Learning
   - classification, modeling, visualization & data Integration
6. Results
   - Domain Knowledge

---

Domain Knowledge

---

Quantitative Biology Technologies
Exome sequencing of the SSC

Last year saw 3 reports of >593 families from the Simons Simplex Collection

- Parents plus one child with autism and one non-autistic sibling
- All attempted to find “gene killing mutations” specific to the autistic children to find genes associated with the disease
- Iossifov (343) and O’Roak (50) used GATK, Sanders (200) didn’t attempt to identify indels

De novo gene disruptions in children on the autism spectrum

De novo mutations revealed by whole-exome sequencing are strongly associated with autism

Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations
Scalpel: Haplotype Microassembly

DNA sequence micro-assembly pipeline for accurate detection and validation of de novo mutations (SNPs, indels) within exome-capture data.

Features

1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations

NRXN1 de novo SNP
(auSSC12501 chr2:50724605)

Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly
Scalpel Pipeline

1. Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs.

2. Decompose reads into overlapping \( k \)-mers and construct de Bruijn graph from the reads.

3. Find end-to-end haplotype paths spanning the region.

4. Align assembled sequences to reference to detect mutations (deletion and insertion).
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis
- Individual was diagnosed with ADHD
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis
- Individual was diagnosed with ADHD (See Gholson for details)
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)

![Venn diagram showing the overlap of Scalpel, SOAPindel, and GATK Haplotype Caller results.]

- 77% PPV
- 99% PPV
- 50% PPV
- 23% PPV
Revised Analysis of the SSC

Constructed database of >1M transmitted and de novo indels
Many new gene candidates identified, population analysis underway
De novo mutation discovery and validation

**Concept**: Identify mutations not present in parents.

**Challenge**: Sequencing errors in the child or low coverage in parents lead to false positive de novos

**Reference**: ...TCAAATCCTTTTAATAAAAGAAGAGCTGACA...

**Father**: ...TCAAATCCTTTTAATAAAAGAAGAGCTGACA...

**Mother**: ...TCAAATCCTTTTAATAAAAGAAGAGCTGACA...

**Sibling**: ...TCAAATCCTTTTAATAAAAGAAGAGCTGACA...

**Proband (1)**: ...TCAAATCCTTTTAATAAAAGAAGAGCTGACA...

**Proband (2)**: ...TCAAATCCTTTTAAT****AAGAGCTGACA...

4bp heterozygous deletion at chr15:93524061 CHD2
De novo Genetics of Autism

- In 593 family quads so far, we see significant enrichment in de novo *likely gene killers* in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMRP
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

*Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly*
The potential for big data?

Seasonal influenza epidemics causing tens of millions of cases and 500,000 deaths worldwide each year, a new strain of influenza, if immunity exists and that dissemination could result in a pandemic. Early detection of disease outbreaks, and hence a rapid response, can reduce the impact of influenza. One way to identify people seeking health care is through search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness.

Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.
The fallacy of big data?

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (9–11). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data. The assumption that big data is a substitute for traditional methods and theories is false. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories. We explore two common examples of the fallacy of big data hubris and algorithm dynamics—and other lessons for moving forward in the use of data.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (9–11). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data. The assumption that big data is a substitute for traditional methods and theories is false. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories. We explore two common examples of the fallacy of big data hubris and algorithm dynamics—and other lessons for moving forward in the use of data.
The risks of big data?

Predicting Social Security numbers from public data

Alessandro Acquisti\textsuperscript{1} and Ralph Gross
Carnegie Mellon University, Pittsburgh, PA 15213
Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of personal multiple sources, such as data brokers or proworking sites. Our results highlight the unexplored sequences of the complex interactions among sources in modern information economies and the risks associated with information revelation in these emerging environments.

In modern information economies, sensitive personal data is at risk of being misused or exposed. One example is the Social Security number (SSN). The SSN is a unique identifier assigned to each U.S. citizen and is used for a variety of purposes, including employment, government benefits, and financial transactions. The SSN is often used in conjunction with personal information such as date of birth, address, and other identifiers to verify an individual's identity.

In recent years, there has been an increase in the number of instances where SSNs have been exposed through breaches of security. This has led to concerns about the safety of SSNs and the risk of identity theft. To address these concerns, researchers have developed methods to predict SSNs from publicly available data.

In this study, the authors used publicly available data to predict SSNs. They observed a correlation between SSNs and birth data, which allowed them to make statistical inferences about private SSNs. The inferences were made possible by the availability of the Social Security Administration's Death Master File and the widespread accessibility of personal data sources, such as data brokers or proworking sites.

These findings highlight the need for increased security measures to protect SSNs and other sensitive personal data. The implications of this research are significant, as it demonstrates the potential for statistical inference to be used in the identification of personal information.
Learning and Translation

**Tremendous power from data aggregation**
- Observe the dynamics of biological systems
- Breakthroughs in medicine and biology of profound significance

**Be mindful of the risks**
- The potential for over-fitting grows with the complexity of the data, statistical significance is a statement about the sample size
- Reproducible workflows, APIs are a must
- Caution is prudent for personal data

**The foundations of biology will continue to be observation, experimentation, and interpretation**
- Technology will continue to push the frontier
- Feedback loop from the results of one project into experimental design for the next
Who is a Data Scientist?

Acknowledgements

Schatz Lab
Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Srividya
Ramakrishnan
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Alejandro Wences
Greg Vurture
Eric Biggers
Aspyn Palatnick
CSHL
Hannon Lab
Gingeras Lab
Jackson Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab
IT Department
Thank you

http://schatzlab.cshl.edu
@mike_schatz / #KSBigData