Challenges in precision medicine: From sequencing to big data processing

Mikel Hernaez

CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY
Where Science Meets Society

Torino, July 19th 2017
“Tonight I’m launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes”

President Obama, 2015
Precision Medicine

• This growing field will revolutionize how we treat disease, taking into account the individual’s unique genetic makeup, environment, and lifestyle.

• Hospitals around the world are working hard to embrace the personalized medicine paradigm.
Precision Medicine: Technology-enabled medicine

Today

Disease Progression

EXAMINE, DIAGNOSE, TREAT

Future

Health & Wellness, Disease Etiology

SYSTEMS OR PRECISION MEDICINE

The Human System – A Grand Challenge

How does the system work

Medical Imaging
Advanced Microscopy

Sensing & Imaging
Personalized Diagnostics
Bionanotechnology

Computational Medicine
Data Integration

Computational Modeling
Systems Biology
Omic Networks

How to model the system

How to improve the system

Regenerative Medicine
Therapeutic Delivery

Cell and Molecular
Reengineering

Synthetic Bioengineering
Drug Discovery

CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY
Where Science Meets Society
The Mayo-Illinois Alliance

Top-Ranked Medical Center, World Renowned Tradition in Quality Health Care Delivery

Top-Ranked Programs in Engineering, Computation, Bioinformatics, Genomics and Nanotechnology

INFORMATION-BASED MEDICINE

GENOMICS

POINT-OF-CARE DIAGNOSTICS
The Mayo-Illinois Alliance

• The goal is to advance clinical and translational research and to harness the power of big data to transform precision medicine for the 21st century.

• The Alliance’s interests extend to numerous areas, including:
  – High-performance computing
  – Machine learning
  – Signal processing
  – Nanotechnology
The Mayo Grand Challenge: The 10,000 Genome Project

• The aim is to develop a pipeline capable of sequencing, analyzing, visualizing, and interpreting genomes of at least 10,000 patients per year, each within 48 hours.

• Some challenges:
  – Data Volume: ~1PB a year
  – Data Processing: GWAS analysis in less than 48 h (per patient)
Problems towards personalized medicine

• Data size:
  – New sequencing machines can generate 1 TB of data per day.
  – Institutions are shipping HDs through FedEx instead of transmitting data through the internet!

• Data analysis:
  – Interoperability across computational biology methods is not fully guaranteed, causing huge headaches to researchers and data scientists from hospitals and institutions.
## Evolution of Genome Sequencing

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost/Genome</strong></td>
<td>$100K</td>
<td>$1K</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>30x</td>
<td>&gt; 200x</td>
</tr>
<tr>
<td><strong>Number of reads</strong></td>
<td>1 Billion</td>
<td>&gt; 6 Billion</td>
</tr>
<tr>
<td><strong>Size of raw sequencing files</strong></td>
<td>0.25 TB</td>
<td>&gt; 1.5 TB</td>
</tr>
</tbody>
</table>
## Evolution of Genome Sequencing

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost/TFB</strong></td>
<td>$100/TFB</td>
<td>$50/TFB</td>
</tr>
<tr>
<td><strong>Download Speed</strong></td>
<td>10Mbps</td>
<td>100Mbps</td>
</tr>
<tr>
<td><strong>Cost/Genome</strong></td>
<td>$100K</td>
<td>$1K</td>
</tr>
</tbody>
</table>

No technology is keeping with the pace of genome sequencing!
Some proposed solutions

Contributors:
Stanford University, University of Washington, Carnegie Melon, Simon Fraser University, Cornell University, University of Hannover, European Bioinformatics Institute, Silesian University of Technology, University of Illinois, ...
Weight of the lossless compressed data

CRAM

Quantization

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiers</td>
<td>13.20%</td>
</tr>
<tr>
<td>Quality Scores</td>
<td>76.50%</td>
</tr>
<tr>
<td>Sequence</td>
<td>5.30%</td>
</tr>
<tr>
<td>Aux</td>
<td>1.97%</td>
</tr>
<tr>
<td>Other (pairing, flags, mqual, format overheads, etc)</td>
<td>3.02%</td>
</tr>
</tbody>
</table>
Quantization

350KB - 2:1 COMPRESSION

49KB - 30:1 COMPRESSION

25KB - 60:1 COMPRESSION

12KB - 100:1 COMPRESSION
Quantization of quality scores

Contributors:
Stanford University, MIT, Simon Fraser University, University of Hannover, EBI, University of Illinois, Illumina Inc., University of Melbourne, …
Quantization of quality scores

- Order of magnitude improvements in compression.

- Several extensive analyses of its effect on variant calling and RNA-Seq gene expression:
  - Only negligible variation of results
  - Consistent improvements in variant calling also shown!
## Benefits of Quantization

<table>
<thead>
<tr>
<th>Human data</th>
<th># Reads</th>
<th>SAM</th>
<th>BAM</th>
<th>CRAM</th>
<th>Quantize QS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr20 (5X)</td>
<td>3.4 M</td>
<td>800 MB</td>
<td>172 MB</td>
<td>89 MB</td>
<td>27 MB</td>
</tr>
<tr>
<td>Chr20 (85X)</td>
<td>52 M</td>
<td>13 GB</td>
<td>3.5 GB</td>
<td>2.4 GB</td>
<td>0.7 GB</td>
</tr>
</tbody>
</table>

- **75%** reduction in storage space
- **95%** reduction in storage space
- **70%** reduction in storage space
Current *de facto* solutions

- **Raw data: GZIP format**
  - GZIP is a generic file compressor

- **Aligned data: BAM format**
  - BAM is a binarized and GZIP-ed raw aligned data

All the solutions use off-the-self general compressors!
Why are not adopted?

- Most solutions implemented by “students”
- Usually not long-term maintained
- Buggy

There is a need for a well maintained standard format!
Requirements for a good genomic data format

Reliable

Efficient

Maintained

Usable
Requirements for a good genomic data format

• It is not only about compression:
  – Random Access over the compressed domain
  – Indexing capabilities
  – Interoperability among systems
Solutions needed by Hospitals and Institutions

• Mayo Clinic’s personalized medicine initiative is expected to generate ~ 1 PB of data per year.

• These data needs to be store, transmitted and analyzed:
  – Extremely cumbersome with current *de facto* formats!

• A well maintained data standard format is urgently needed to fully enable the personalized medicine paradigm.
Thanks!

CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY
Where Science Meets Society

Torino, July 19th 2017