Genomic Analysis Pipeline: Overview, Challenges, and Proposed Solutions

Idoia Ochoa
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Evolution of Genome Sequencing

Evolution of Genome Sequencing

Cost per Genome

Moore's Law

National Human Genome Research Institute

genome.gov/sequencingcosts
Evolution of Genome Sequencing
Evolution of Genome Sequencing

- Increased throughput of sequencing machines
  - Illumina’s HiSeq X Ten System can sequence 18,000 genomes per year
  - 3.5 PB per machine per year
  - NovaSeq 6000 System: 3 TB per run
Evolution of Genome Sequencing

▪ Advances in knowledge
  – Rich information encoded in the genome
  – Genome sequencing advances our understanding of biology and human genetic disease
# Evolution of Genome Sequencing

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost/Genome</td>
<td>$100K</td>
<td>&lt; $1K</td>
</tr>
<tr>
<td>Number of reads</td>
<td>1 Billion</td>
<td>6 Billion</td>
</tr>
<tr>
<td>Size of raw files</td>
<td>0.25 TB</td>
<td>&gt; 1.5 TB</td>
</tr>
<tr>
<td>Cost/TB</td>
<td>$100/TB</td>
<td>$50/TB</td>
</tr>
<tr>
<td>Download Speed</td>
<td>10Mbps</td>
<td>100Mbps</td>
</tr>
</tbody>
</table>

No other technology is keeping up with the pace of genome sequencing!
The Human Genome

- 4 possible symbols (base pairs)

- Human genome
  - 23 chromosomes
  - 3 billion base pairs

- Stored in a FASTA file

3 GB
Genome Sequencing

Genome Reads Coverage
Current Genomic Data Formats and Analysis Pipeline

FASTQ File
@srr0626347 13976/1
TGGGAATCAGATGGAATCATCGAATGGTC
+GGGGGFDGGGGGFGGGGFGGGGGGGGGGF

Output: FASTQ file

~1.5 TB
200x Clinical usage
Current Genomic Data Formats and Analysis Pipeline

Output: FASTQ file

SAM file

Reference genome

Reference Genome (template)

SAM File

@SRR0626347 13976/1  chr1 50 100M TGGAAATCAGATGGAATCATCGAATGGTC
GGFGGFDGGGFGGGGGGGGGGGGGGGF? <Extra fields> <Optional fields>

~3 TB
Current Genomic Data Formats and Analysis Pipeline

FASTQ file → Alignment → SAM file → Variant caller → VCF file

VCF File

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>REF</th>
<th>ALT</th>
<th>&lt;Annotations&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>150</td>
<td>T</td>
<td>C</td>
<td>---</td>
</tr>
<tr>
<td>20</td>
<td>175</td>
<td>A</td>
<td>T</td>
<td>---</td>
</tr>
</tbody>
</table>

CHALLENGES:
- SEQUENCING ERROR
- LOW COVERAGE AREAS
- SUB-OPTIMALITY OF PRECEDING ANALYSIS TOOLS
- ETC.

A
A
A
A
C
C
C
C
C
A
A

~1GB
Outline

- Bringing Digital-Era formats to Genomic Information Representation
- Improving Precision of Variant Identification
- Some Thoughts on the Future
Bringing Digital-Era Formats To Genomic Information Representation
Human Data Volumes

Genomic Information Representation

- FASTQ file
- Alignment
- SAM file
- Variant caller
- Reference genome
- Reference genome
- 3 GB
- 3 GB
- ~1.5 TB
- ~3 TB
- ~1 GB

~14

~3 TB
Future of Data Storage

Current Research studies:
1108 SAM files (277 samples)
35 TB
$4,200/year in AWS

Future Clinical Data:
130,000 SAM files/year
9.75 PB/year
~$1M/year in AWS
Proposed Solutions

Genomic Information Representation
# Compression Results for SPRING (lossless)

<table>
<thead>
<tr>
<th>WGS @50x</th>
<th>(GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM</td>
<td>732</td>
</tr>
<tr>
<td>Gzipped SAM (a.k.a. BAM)</td>
<td>144</td>
</tr>
<tr>
<td>CRAM</td>
<td>86</td>
</tr>
<tr>
<td>FASTQ</td>
<td>392</td>
</tr>
<tr>
<td>Gzipped FASTQ</td>
<td>139</td>
</tr>
<tr>
<td>SPRING</td>
<td>49</td>
</tr>
</tbody>
</table>
Still, the community uses inadequate solutions...

Shipping hard drives

Universal lossless coding not particularized to applications
Requirements for a good compression format

- Efficient
- and scalable
- Reliable
- Maintained
- Usable

Compression

Computing and Learning

- Statistical Analysis
- Information Fusion
- Dimensionality reduction
- Machine learning
- Selective Access

Genomic Information Representation
MPEG-G: An international standard for genomic data

- ISO started a standardization activity in 2015

- First attempt to design an industry grade solution for genomic data compression

- Bringing the Digital Era to genomic data formats!

- U of I members are among the core developers of MPEG-G
An MPEG-G standard

- Standard = normative text:
  - It is a book: a set of instructions on how to retrieve genomic data from the compress domain

- Why is this a book?
  - It is generic and not tied to particular implementations

- E.g., MP3 (MPEG-1)
  - Everybody was able to listen to digital music. Why?
  - Companies relied on the MP3 standard book to develop their products
  - An ecosystem thrived
MPEG-G file format

SAM File

@SRR0626347 13976/1 chr1 50 100M TGGAATCAGATGGAATCATCGAATGGTC GGGFFGDGGGGGGFGGGGGGGGGGF? <Extra fields> <Optional fields>

For example,
- **MPEG-G file**: sequencing data of a trio:
- **File Header**: metadata related to the study.
- **Dataset Group**: one per individual + metadata from the individual.
- **Dataset**: sequencing data + metadata from one experiment.
- **The colored structures**: This is how genomic data is represented in MPEG-G.

The MPEG-G file can encapsulate the entire genomic history of one or more individuals in a unique file including the metadata describing the study, samples, etc.
MPEG-G Transport format (streaming)

- Packet size adaptation to the channel characteristics/state
- Error detection and support of re-transmission of erroneous/incomplete data for error-free delivery
- Support of out-of-order delivery
- Packet-based filtering of genomic data
Compression achievable by MPEG-G

NA12878 S1 High Coverage

- Sequences
- QV
- AUX

Compression Factor (CF) ~ 50

- SAM: 377 GB
- BAM: 150 GB
- MPEG-G: 53 GB
- MPEG-G: 8.7 GB

Genomic Information Representation
Benefits provided by MPEG-G

- Selective access to compressed data
- Data streaming
- Genomic studies aggregation
- Enforcement of privacy rules with encryption
- Annotation and linkage of genomic segments in the compressed domain
- Interoperability with main existing technologies and legacy formats

It’s not just about compression efficiency!
A genomic ecosystem fueled by MPEG-G
MPEG-G is an international collaborative effort.
Thanks!

idoia@Illinois.edu
http://idoia.web.engr.illinois.edu/