Genome in a Bottle: You’ve sequenced. How well did you do?

April 2015

Justin Zook, Marc Salit, and the Genome in a Bottle Consortium

Presented by Steve Lincoln
Project leaders
at US National Institute of Standards and Technology

Justin Zook
Marc Salit
Sequencing technologies can disagree about 100,000s of variants per genome.

Platform #1:
- 230,311 (5.76%)
- 208,038 (5.21%)
- 125,574 (3.14%)

Platform #2:
- 121,440 (3.04%)
- 39,604 (0.99%)

Platform #3:
- 71,944 (1.80%)
- 39,604 (0.99%)

# SNPs (% of single nucleotide polymorphisms [SNPs] detected by any platform)
Bioinformatics pipelines also disagree on the same raw data

O’Rawe, et al., Genome Medicine 2013, 5:28
Bioinformatics pipelines also disagree on the same raw data

Who is right?

Is anyone right?
Genome in a Bottle Consortium (GIAB)

Hosted by US National Institute of Standards and Technology

Goal: Provide infrastructure for performance assessment of NGS.

- Appropriately consented, widely available **DNA samples**, distributed by the Coriell Institute
  - Also, QCed Reference Material (RM) versions from controlled lots will be available from NIST

- High-accuracy **reference data** for these samples

- **Tools** to facilitate their use
  - With the Global Alliance Data Working Group Benchmarking Team

[genomeinabo)le.org](http://genomeinabo)le.org)

[ga4gh.org](http://ga4gh.org)
GIAB selected samples

CEPH/Utah pedigree 1463

NA12889
NA12890
NA12891
NA12892
NA12877
NA12878

Ashkenazi Jewish trio

NA24149
NA24143
NA24385

Asian (Han Chinese) trio

NA24694
NA24695
NA24631

Note: Data from the complete families can be used to improve variant calls in the specific GIAB samples.
Pilot genome: NA12878
Goals for GIAB data

• Close to 0 false positives
• Close to 0 false negatives (missing calls)
• Identify confident regions that have this high accuracy.
  – Include as much of the genome as possible in the confident regions (i.e., don’t just use the intersection set).
• Avoid bias towards any particular platform or bioinformatics pipeline.
  – Leverage strengths of each individual platform.

Pilot genome (NA12878): integrated 14 datasets from five platforms

<table>
<thead>
<tr>
<th>Source*</th>
<th>Platform</th>
<th>Mapping algorithm</th>
<th>Coverage</th>
<th>Read length</th>
<th>Genome/exome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes</td>
<td>Illumina Gallx</td>
<td>Bwa</td>
<td>39</td>
<td>44</td>
<td>Genome</td>
</tr>
<tr>
<td>1000 Genomes</td>
<td>Illumina Gallx</td>
<td>Bwa</td>
<td>30</td>
<td>54</td>
<td>Exome</td>
</tr>
<tr>
<td>1000 Genomes</td>
<td>454</td>
<td>Ssaha2</td>
<td>16</td>
<td>239</td>
<td>Genome</td>
</tr>
<tr>
<td>X Prize</td>
<td>Illumina HiSeq</td>
<td>Novoalign</td>
<td>37</td>
<td>100</td>
<td>Genome</td>
</tr>
<tr>
<td>X Prize</td>
<td>SOLiD 4</td>
<td>Lifescope</td>
<td>24</td>
<td>40</td>
<td>Genome</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td>Complete Genomics</td>
<td>CGTools 2.0</td>
<td>73</td>
<td>33</td>
<td>Genome</td>
</tr>
<tr>
<td>Broad</td>
<td>Illumina HiSeq</td>
<td>Bwa</td>
<td>68</td>
<td>93</td>
<td>Genome</td>
</tr>
<tr>
<td>Broad</td>
<td>Illumina HiSeq</td>
<td>Bwa</td>
<td>66</td>
<td>66</td>
<td>Exome</td>
</tr>
<tr>
<td>Illumina</td>
<td>Illumina HiSeq</td>
<td>CASAVA</td>
<td>80</td>
<td>100</td>
<td>Genome</td>
</tr>
<tr>
<td>Illumina</td>
<td>HiSeq – PCR-free</td>
<td>Bwa</td>
<td>56</td>
<td>99</td>
<td>Genome</td>
</tr>
<tr>
<td>Illumina</td>
<td>HiSeq – PCR-free</td>
<td>Bwa</td>
<td>190</td>
<td>99</td>
<td>Genome</td>
</tr>
<tr>
<td>Life Technologies</td>
<td>Ion Torrent</td>
<td>tmap</td>
<td>80</td>
<td>237</td>
<td>Exome</td>
</tr>
</tbody>
</table>

Integration methods to establish reference variant calls for NA12878

Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook¹, Brad Chapman², Jason Wang³, David Mittelman³,⁴, Oliver Hofmann², Winston Hide² & Marc Salit¹
Integration methods to establish reference variant calls for NA12878

1. Candidate variants from each platform
2. Identify concordant/discordant variants
3. Identify characteristics of systematic error
4. Arbitrate using evidence of systematic error
5. Determine confidence levels by region

Assigning confidence to genomic regions for NA12878

High confidence (77%)

• Agreement across platforms or an understanding of the systematic biases causing disagreement
• At least some methods with no evidence of systematic errors
• Consistent Mendelian inheritance

Lower confidence (23%)

• In a region known to be difficult for current technologies
  – Segmental duplications
  – Repeats, low complexity
  – High/low GC
• Evidence of systematic error across many platforms
• Inconsistent inheritance

Assigning confidence to genomic regions for NA12878

High confidence (77%)
- Agreement across platforms or agreement with existing systems
- Disagreement with existing systems
- At least some methods with no evidence of systematic errors
- Consistent Mendelian inheritance

Lower confidence (23%)
- In a region known to be difficult for current technologies – Segmental duplications – Repeats, low complexity – High/low GC
- Evidence of systematic error across many platforms
- Inconsistent inheritance

A BED file of high confidence regions is provided.
You should use it!
Really!

Quality assessment of NA12878 calls in high-confidence regions

• 100% sensitivity/specificity vs. GeT-RM data
  – 427 SNPs and 42 indels, 366kb negative

• ≈99.8% initial sensitivity vs. Xprize Fosmid Sequences
  – 135,652 SNVs and 10,942 insertions/deletions
  – 124 SNPs and 37 indels (Sanger confirmed)
  – After manual data review of discrepancies: >99.9%
    • Most were errors were in the Fosmid data.

• Other QC
  – Comparison vs SNP microarrays
  – Ti/Tv ratio
  – Et cetera
Quality assessment of NA12878 calls in high-confidence regions

- 100% sensitivity/specificity vs. Get-RM data
  - 427 SNPs and 42 indels, 366 kb negative
- ≈99.8% identity sensitivity vs. Xprize Fosmid Sequences
  - 135,652 SNVs and 10,942 insertions/deletions
  - 124 SNPs and 37 indels (Sanger confirmed)
- Aser manual data review of discrepancies: >99.9%
  - Most were errors in the Fosmid data.
- Other QC - Comparison vs. SNP Microarrays - Ti/Tv tandem repeats
- ≈94,500 true+ SNVs
- ≈1400 true+ indels
- 0 or 1 false+ or false- (SNVs and indels together)
- But ≈3 partial calls of complex small variants

Per 120 MB*...

GeT-RM browser from NCBI and CDC

NIST human genome reference materials (RMṣ)

• NA12878 is available from Coriell today.

• NIST RM version will available later in 2015.
  – DNA isolated from large-growth cell cultures
  – Stable, homogeneous
  – Best for regulated uses

• There are new Ashkenazi Jewish and Asian samples.
  – Available for Coriell now
  – NIST RM available in 2016
Comparing your data to the reference data
NGS validation process using genomes in bottles

Pre-analytical process
- Sample
- gDNA isolation

Analytical process
- Library prep
- Sequencing
- Alignment/mapping
- Variant calling
- Confidence estimates
- Downstream analysis

Clinical interpretation

Genome in a Bottle scope

GIAB Data

genomeinabottle.org
Uses of GIAB NA12878

**RESEARCH ARTICLE**

Analytical validation of whole exome and genome sequencing for clinical applications

Michael D Linderman, Tracy Brandt, Lisa Edelmann, Omar Jabado, Yu Wang, Milind Mahajan, Hardik Shah, Andrew Kasarskis and Eric E Schadt

Xu et al. BMC Genomics 2014, 15:244

Comparison of somatic mutation calling in amplicon and whole exome sequencing

Huilei Xu, John DiCarlo, Ravi Vijaya Satya, Quan Peng and Yexun Wang

Oncology – molecular and cellular tumor markers...

“next generation” sequencing (NGS) guidelines for somatic genetic variant detection
Contribution to Invitae hereditary cancer validation study (29 genes)

- Seven whole genomes contributed 310 of 750 selected variants.
  - These genomes included GIAB NA12878 and six others with similarly cleaned data.
  - 41% of the total set of variants came from 0.6% of the samples.

- In 15 of 29 genes, the seven samples at least doubled the selected variant count.

- WCGs added variants in one gene (PTCH1) that otherwise had none selected from clinical data.

- WCGs saved 310 Sanger confirmations.
  - Unlike confirmation, WCGs contribute both to sensitivity and specificity measurements in a strong way.
Limitations of GIAB

- There are no coding variants in five of 29 genes.
  - CDKN2A, PALB2, RAD51C, SMAD4
  - CHEK2 (a special case)

- There is only one coding variant in two other genes.
  - PTEN, TP53

- The only errors in any reference data we saw were in WGS data (but not GIAB!).
  - Two in NA19240; one in NA12892
  - All errors in low-complexity sequence

- Many of the variants are repeated.
  - Partly due to using related individuals
  - Partly because most are common polymorphisms
Limitations of GIAB

- There are no coding variants in five of 29 genes.
  - CDKN2A, PALB2, RAD51C, SMAD4
  - CHEK2 (a special case)

- There is only one coding variant in two other genes.
  - PTEN, TP53

- The only errors in any reference data we saw were in WGS data (but not GIAB).
  - Two in NA19240; one in NA12892
  - All errors in low-complexity sequence

- Many of the variants are repeated.
  - Partly due to using related individuals
  - Partly because most are common polymorphisms

Genome in a Bottle is a valuable—but not sufficient—part of a validation study.
The challenge in variant comparison: Complex small variants have multiple representations.

Ref: AAA CAGTGA GAA
Alt: AAA TCTCTCT GAA
The challenge in variant comparison: Complex small variants have multiple representations.

Ref: AAA-CAGTGTGAGAA
Alt: AAATC--TCTGAA

Ref: AAACAGGTGAGAA
Alt: AAA--TCTCTGAA

Ref: AAACAGGTGAGAA
Alt: AAATC-CTCTGAA

Ref: AAACAGGTGAGAA
Alt: AAATCTCTGAA

Ref: AAACAGGTGAGAA
Alt: AAA-TCTCTGAA

Ref: AAACAGTGA------GAA
Alt: AAA------TCTCTGAA

hg19  NC_000001.10:114841792–114841797
The challenge in variant comparison: Complex small variants have multiple representations.

```
123-456789012
Ref: AAA-CAGTGAGAA 3_4insT
     |||--|--|:::||| 5_6delAG
Alt: AAATC--TCTGAA 8_9delGAinsCT

123456789012
Ref: AAACAGTGAGAA 4_6delCAGinsTC 5_6delAGinsTC
     |||--|--|:::||| 8_9delGAinsCT 8_9delGAinsCT
Alt: AAA-TCTCTGAA

123456789012
Ref: AAACAGTGAGAA 4_9delCAGTGAinsTCTCT 5_6delAGinsTC
     |||--|--|:::||| 8_9delGAinsCT
Alt: AAAT-CCTCTGAA

123456789012
Ref: AAACAGTGAGAA 4_9delCAGTGAinsTCTCT 5_6delAGinsTC
     |||--|--|:::||| 8_9delGAinsCT
Alt: AAAT-CCTCTGAA

CG requires two reference bp to separate variants.
```

4C>T
5delA
6G>C
8_9delGAinsCT
The challenge in variant comparison: Complex small variants have multiple representations.

<table>
<thead>
<tr>
<th></th>
<th>FP SNPs</th>
<th>FP MNPs</th>
<th>FP indels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional comparison</td>
<td>0.38%</td>
<td>100%</td>
<td>6.5%</td>
</tr>
<tr>
<td></td>
<td>(610)</td>
<td>(915)</td>
<td>(733)</td>
</tr>
<tr>
<td>Comparison with allele realignment</td>
<td>0.15%</td>
<td>4.2%</td>
<td>2.6%</td>
</tr>
<tr>
<td></td>
<td>(249)</td>
<td>(38)</td>
<td>(298)</td>
</tr>
</tbody>
</table>

Improved comparison tool in development by Kevin Jacobs of 23andme
Global Alliance Benchmarking Team
New GIAB samples
New data for Ashkenazi Jewish and Asian samples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Characteristics</th>
<th>Coverage</th>
<th>Availability</th>
<th>Most useful for...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina Paired-end</td>
<td>Short read 150x150bp</td>
<td>~300x</td>
<td>On FTP now</td>
<td>SNPs/indels/some SVs</td>
</tr>
<tr>
<td>Complete Genomics Standard</td>
<td>Short read 26x26</td>
<td>~100x</td>
<td>On FTP now</td>
<td>SNPs/indels/some SVs</td>
</tr>
<tr>
<td>Ion Proton Exome</td>
<td>Short read 190bp average</td>
<td>~1000x</td>
<td>On FTP now</td>
<td>SNPs/indels in exome</td>
</tr>
<tr>
<td>Illumina Mate- pair</td>
<td>~6kb insert size 150x150bp</td>
<td>~30x/individual</td>
<td>Finished; on FTP Apr 2015</td>
<td>SVs</td>
</tr>
<tr>
<td>Illumina “Molecule”</td>
<td>Synthetic long reads</td>
<td>~30x long contigs</td>
<td>Finished; on FTP Apr-May 2015</td>
<td>SVs/phasing/assembly</td>
</tr>
<tr>
<td>Complete Genomics LFR</td>
<td>Synthetic long reads</td>
<td>~100x</td>
<td>Finished; on FTP Apr-May 2015</td>
<td>SNPs/indels/phasing</td>
</tr>
<tr>
<td>10X</td>
<td>Synthetic long reads</td>
<td>30-45x sequence</td>
<td>Target May 2015</td>
<td>SVs/phasing/assembly</td>
</tr>
<tr>
<td>BioNano Genomics</td>
<td>200-250kbp optical map fragments</td>
<td>~100x AJ individuals ~57x on Asian son</td>
<td>On FTP now</td>
<td>SVs/assembly</td>
</tr>
<tr>
<td>PacBio</td>
<td>~10kb reads</td>
<td>~69x each child ~30x each parent</td>
<td>On FTP now</td>
<td>SVs/phasing/assembly/STRs</td>
</tr>
</tbody>
</table>
Data analysis group–new data sets

Leaders
• Francisco de la Vega
  – Annai Systems
• Chris Mason
  – Weill Cornell Medical Center
• Tina Graves
  – Washington University
• Valerie Schneider
  – NCBI

Status
• Collecting data into a central FTP site
• Developing an analysis plan
• Developing a validation plan and metrics
• Recruiting people to help with the work

This could be you. We need volunteers!
Data analysis group—new data sets

Leaders
• Francisco de la Vega
  — Annai Systems
• Chris Mason
  — Weill Cornell Center
• Valerie Schneider
  — NCBI

Status
• Collecting data into a central FTP site
• Developing an analysis plan
• Developing a validation plan and metrics
• Recruiting people to help with the work

This could be you.
We need volunteers!

Fame!
Glory!
Publications!
Free Data!
High-level analysis plan

Assembly
- Long reads
- Synthetic long reads
- Mapping
- Integration

Small variants
- Long reads
- Synthetic long reads
- Short paired-end
- Mate-pair
- Exome
- Integration

SVs/VNTRs
- Long reads
- Synthetic long reads
- Short paired-end
- Mate-pair
- Mapping
- Integration

Phasing
- Long reads
- Synthetic long reads
- Mapping
- Integration

Phased, integrated, high-confidence variant calls

Personal genome reference for mapping

Francisco de la Vega
Detailed draft analysis plan
Acknowledgments

• FDA–Elizabeth Mansfield, computing staff

• Harvard School of Public Health

• GCAT website–David Mittelman, Jason Wang

• Members of Genome in a Bottle
  – New members welcome!
  – Sign up on website for email newsletters.

Steering Committee
  – Marc Salit
  – Justin Zook
  – David Mittelman
  – Andrew Grupe
  – Michael Eberle
  – Steve Sherry
  – Deanna Church
  – Francisco De La Vega
  – Christian Olsen
  – Monica Basehore
  – Lisa Kalman
  – Christopher Mason
  – Elizabeth Mansfield
  – Liz Kerrigan
  – Leming Shi
  – Melvin Limson
  – Alexander Wait Zaranek
  – Nils Homer
  – Fiona Hyland
  – Steve Lincoln
  – Don Baldwin
  – Robyn Temple-Smolkin
  – Chunlin Xiao
  – Kara Norman
  – Luke Hickey
For more information

www.genomeinabottle.org

www.bioplanet.com/gcat—(old) comparison tool


Global Alliance Benchmarking work group
  – ga4gh.org/#/benchmarking-team

Twice-yearly workshop
  – Winter: Stanford University, California, USA
  – Summer: August 27-28, 2015 at NIST, Maryland, USA
    • Near Washington, DC

Open Meeting!

Justin Zook: jzook@nist.gov
Marc Salit: salit@nist.gov
Francisco de la Vega: francesco.dlv@gmail.com
Steve Lincoln: steve.lincoln@me.com
Genome in a Bottle consortium development

• NIST meeting with sequencing technology developers to assess standards needs
  – Stanford, June 2011
• Open, exploratory workshop
  – ASHG, Montreal, Canada
  – October 2011
• Small, invitational workshop at NIST to develop consortium for human genome reference materials
  – FDA, NCBI, NHGRI, NCI, CDC, Washington University, Broad, technology developers, clinical labs, CAP, PGP, Partners, ABRF, others
  – Developed draft work plan
  – April 2012
• Open, public meetings of GIAB
  – August 2012 at NIST
  – March 2013 at Xgen
  – August 2013 at NIST
  – January 2014 at Stanford
  – August 2014 at NIST
  – January 2015 at Stanford
• Website
  – www.genomeinabottle.org
Others working in this space...

Well-characterized genomes
- Illumina Platinum Genomes
- CDC GeT-RM
- Korean Genome Project
- Human Longevity, Inc.
- Hyditaform mole haploid cell line
- Genome Reference Consortium

Performance metrics
- Global Alliance for Genomics and Health Benchmarking Team
- NCBI/CDC GeT-RM Browser
- GCAT website
Perspective

First FDA Authorization for Next-Generation Sequencer
Francis S. Collins, M.D., Ph.D., and Margaret A. Hamburg, M.D.

This year marks 60 years since James Watson and Francis Crick described the structure of DNA and 10 years since the complete sequencing of the human genome. Fittingly, today the Food and Drug Administration (FDA) has granted marketing authorization for the first high-throughput (next-generation) genomic sequencer, Illumina’s MiSeqDx, which will allow the development and use of innumerable new genome-based tests.

When a global team of researchers sequenced that first human genome, it took more than a decade and cost hundreds of millions of dollars. Today, because of federal and private investment, sequencing technologies have advanced dramatically, and a human genome can be sequenced in about 24 hours for what is now less than $5,000 (see graph). This is a rare example of technology development in which faster, cheaper, and better have coincided: as costs have plummeted and capacity has increased, the accuracy of sequencing has substantially improved. With the FDA’s announcement, a platform that took nearly a decade to develop from an initial research project funded by the National Institutes of Health will be brought into use for clinical care. Clinicians can selectively look for an almost unlimited number of genetic changes that may be of medical significance. Access to these data opens the door for the transformation of research, clinical care, and patient engagement.

To see how this technology could be used, consider cancer. Comprehensive analysis of the genome sequence of individual cancers has helped uncover the specific mutations that contribute to the malignant phenotype, identify new targets for therapy, and increase the opportunities for choosing the optimal treatment for each patient. For instance, lung adenocarcinoma can now be divided into subtypes with unique genomic fingerprints associated with different outcomes and different responses to particular therapies. More broadly, recent work from the Cancer Genome Atlas demonstrates that the tissue of origin of a particular cancer may be much less relevant to prognosis and response to therapy than the array of causative mutations. As a result, patients diagnosed with a cancer for which there are few therapeutic options may increasingly benefit from drug therapies originally aimed.

NIST plays a role in the first FDA authorization for next-generation sequencer
November 20, 2013
## Candidate NIST reference materials

<table>
<thead>
<tr>
<th>Genome</th>
<th>PGP ID</th>
<th>Coriell ID</th>
<th>NIST ID</th>
<th>NIST RM #</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEPH Mother/ Daughter</td>
<td>N/A</td>
<td>GM12878</td>
<td>HG001</td>
<td>RM8398</td>
</tr>
<tr>
<td>AJ Son</td>
<td>huAA53E0</td>
<td>GM24385</td>
<td>HG002</td>
<td>RM8391 (son)/RM8392 (trio)</td>
</tr>
<tr>
<td>AJ Father</td>
<td>hu6E4515</td>
<td>GM24149</td>
<td>HG003</td>
<td>RM8392 (trio)</td>
</tr>
<tr>
<td>AJ Mother</td>
<td>hu8E87A9</td>
<td>GM24143</td>
<td>HG004</td>
<td>RM8392 (trio)</td>
</tr>
<tr>
<td>Asian Son</td>
<td>hu91BD69</td>
<td>GM24631</td>
<td>HG005</td>
<td>RM8393</td>
</tr>
<tr>
<td>Asian Father</td>
<td>huCA017E</td>
<td>GM24694</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Asian Mother</td>
<td>hu38168C</td>
<td>GM24695</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Putting “genomes” in bottles

• NIST worked with GIAB to select genomes.
• Current genomes
  – NA12878 HapMap sample as pilot sample (part of 17-member pedigree)
  – 2 trios from PGP
    • Ashkenazim
    • Asian
Integration methods to establish benchmark variant calls

- Candidate variants
- Concordance
- Find systematic errors
- Arbitrate
- Determine confidence

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Site A</th>
<th>Site B</th>
<th>Site C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform #1</td>
<td>0/0</td>
<td>0/0</td>
<td>1/1</td>
</tr>
<tr>
<td>Platform #2</td>
<td>0/1</td>
<td>0/1</td>
<td>1/1</td>
</tr>
<tr>
<td>Platform #3</td>
<td>0/0</td>
<td>0/1</td>
<td>1/1</td>
</tr>
<tr>
<td>Integrated data</td>
<td>0/0</td>
<td>0/1</td>
<td>No-call</td>
</tr>
</tbody>
</table>

• Mapping bias
• Systematic sequencing error
• Abnormal allele balance
• Local alignment bias
• < 2 datasets
• Conflicting genotypes
• Low coverage/low mapping quality
• Reference in HaplotypeCaller
• Segmental duplications
• 1000 Genomes decoy
• Simple repeats
• dbVar structural variants
NCBI/CDC GeT-RM browser

- Allows visualization of questionable calls

All data is available on our FTP site.

**Download bulk data**

Download data for a subset of the genome. To get all data use FTP.

- **Download pre-defined datasets**
  - TSV = Tab-separated value file, suitable for opening in Excel
  - VCF = Variant call format
  - BED = Tab-separated value file, suitable for uploading to most genome browsers
  - Variant calls validated with an orthogonal technology

**High-quality variants**

- NA12878 TSV VCF
- NA19240 TSV VCF

High-quality regions where no-variant call suggests the sample is homozygous reference

**Sanger regions**

- NA12878 BED

Download data for just your regions of interest

To bulk download variant data, upload a file with a list of genes or positions.

**Sample** NA12878

**Upload options**

- **Upload your gene list or bed file**
  - Choose File: No file chosen
Challenges with assessing performance

- All variant types are not equal.
- All regions of the genome are not equal.
- Labeling difficult variants as *uncertain* leads to higher apparent accuracy when assessing performance.
- Genotypes fall in 3+ categories (not just positive and negative).
  - Standard diagnostic accuracy measures are not well-posed.
Global Alliance for Genomics and Health
Benchmarking Task Team

- Formed in June 2014 to develop methods and tools for comparing variant calls to a benchmark
- Developed standardized definitions for performance metrics like TP, FP, and FN
- Initial focus on germline SNPs/indels
- Developing benchmarking tools
  - Comparison engine
  - Pluggable web interface with modules for:
    - Reporting/calculcation of metrics
    - Visualization/user interface
- Working with Genome in a Bottle Consortium to host data and calls from their well-characterized genomes

User interface example

www.bioplanet.com/gcat
Stratifying performance

• Measure performance for different types of variants in different sequence contexts
  – Types of variants
    • SNPs
    • Indels of different sizes
    • Complex variants
    • Structural variants
  – Sequence contexts
    • Homopolymers,
    • STRs
    • Duplications
  – Functional context
    • Exome vs genome, etc.
  – Data characteristics
    • Coverage
    • Mapping quality

• Challenge of smaller gene panels vs. genome sequencing
  – One piece of reference material may not have a sufficient number of examples of different classes of variants or sequence contexts.
  – More samples with specific types of variants will likely be needed.

• Bed files for sequence context stratification
# Overview of NIST RM development

<table>
<thead>
<tr>
<th>Genome(s)</th>
<th>Q4 2014</th>
<th>Q1 2015</th>
<th>Q2 2015</th>
<th>Q3 2015</th>
<th>Q4 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG-001/NA12878 (&quot;pilot&quot; genome)</td>
<td></td>
<td></td>
<td>Release NIST RM8398; preliminary large deletions</td>
<td>Refined structural variants</td>
<td></td>
</tr>
<tr>
<td>HG-002 to HG-004 (Ashkenazim trio)</td>
<td>Illumina, Complete Genomics, Ion, BioNano, homogeneity/stability</td>
<td>Preliminary SNPs/indels; ~130x PacBio; &quot;moleculo&quot;; mate-pair; CG-LFR</td>
<td>Refined SNPs/indels; preliminary structural variants</td>
<td>Refined structural variants</td>
<td>NIST RM8391/8392 release</td>
</tr>
<tr>
<td>HG-005 (son in Asian trio)</td>
<td>Illumina, Complete Genomics, Ion, BioNano, homogeneity/stability</td>
<td>&quot;moleculo&quot;; mate-pair; CG-LFR</td>
<td>Preliminary SNPs/indels</td>
<td>Refined SNPs/indels; refined structural variants</td>
<td>NIST RM8393 release</td>
</tr>
</tbody>
</table>
Initial uses of high-confidence NIST-GIAB genotypes for NA12878

- NIST have released several versions of high-confidence genotypes for its pilot research material (RM).
- These data are presently being used for benchmarking/
  - Prior to release of RMs
  - SNPs and indels
    - ~77% of the genome
An analytical framework for optimizing variant discovery from personal genomes

Gareth Highnam¹, Jason J. Wang¹, Dean Kusler¹, Justin Zook², Vinaya Vijayan³, Nir Leibovich¹
& David Mittelman¹³

www.bioplanet.com/gcat
Implications of technical accuracy in medical genome sequencing

• Collaboration with Euan Ashley group at Stanford
• What is accuracy for functional variants?
• How much of the exome falls in high confidence regions?
• “Black list” in databases

• Sensitivity
  – WExS (95%) < WGS (98%)
    • Especially splicing
  – Genome < nonsyn < syn
  – Most exome false negatives caused by low coverage
  – Most WGS false negatives caused by filtering

• Only 81% of ClinVar pathogenic or likely pathogenic SNPs fall in high-confidence regions.
  – Lots of work to do!
Future directions

**Germline mutations**
- Difficult regions/variants
  - Long-read technologies
  - Forming an analysis group
- Tools for assessing performance
  - How to stratify performance and understand biases?

**Somatic mutations**
- Pilot inter-laboratory study to assess comparability of spike-ins
- Commercial members developing FFPE cell lines
- Participants interested in mixing different research materials
How to get involved

• Use our integrated SNP/indel genotypes for NA12878 and give us feedback.
  – Cells and DNA currently available from Coriell
  – NIST RM available April 2015

• Join our new Analysis group.
  – Use long-read technologies
  – Structural variant calls
  – De novo assembly
  – Help create the best-ever characterized trio

• Attend our biannual workshops (January in California; August in Maryland)

• Develop tools and metrics with Global Alliance for Genomics and Health Benchmarking Team
Acknowledgments

- FDA—Elizabeth Mansfield, HPC staff
- HSPH spell
- GCAT—David Mittelman, Jason Wang
- Francisco De La Vega
- Illumina—Mike Eberle
- Personalis—Deanna Church
- NCBI—Chunlin Xiao
- Celera—Andrew Grupe

- Genome in a Bottle
  - www.genomeinabottle.org
  - New members welcome!
  - Sign up for email newsletters
  - jzook@nist.gov
Integration methods to establish benchmark variant calls

Dataset | Site A | Site B | Site C
---|---|---|---
Platform #1 | 0/0 | 0/0 | 1/1
Platform #2 | 0/1 | 0/1 | 1/1
Platform #3 | 0/0 | 0/1 | 1/1
Integrated data | 0/0 | 0/1 | No-call