Assessment of variant calling pipelines for clinical diagnosis

Vandhana Krishnan
Stanford Center for Genomics and Personalized Medicine
Stanford Clinical Genomics Service
Overview

- Advancements in research towards patient care
- Challenges in translating genomics research
- Clinical Genomics Service pipeline (@Stanford Health Care)
- Accuracy benchmarking
Advancements in research towards patient care

- Rare variants causing disease
- Pharmacogenomics
- Clinical research informatics (health records)
- Computational modelling
- Gene interaction networks
Translating research for patient care

State of the art facilities doing groundbreaking research @ Stanford University

Leading clinical services @ Stanford Health Care, Lucile Packard Children’s Hospital

Provide benefits to patient care
Translating genomics research for patient care

1. Physician requests patient sequencing
2. Sequenced patient data
3. Bioinformatics analyses
4. Clinical data interpretation
5. Patient results delivered to physician
Translating genomics research for patient care: Challenges

Raw data to interpretation

- Several sequencing platforms constantly evolving
  - Long reads versus short reads
  - Use one platform or a combination?

- Identification of variants
  - Choosing among multiple sequence aligners
  - Choosing among multiple variant callers
  - Testing variant calling pipelines (aligner + variant caller combos)
Challenges in translating genomics research

Raw data to interpretation

- Evaluation of variant calling pipelines (existing publications)
  - BWA + GATK-HC
  - Bowtie2 + GATK-UG / Bowtie2 + GATK-HC
  - NovoAlign + GATK-HC
  - Isaac + Isaac
  - BWA + FreeBayes
  - More combinations…
  - Involves testing with truth sets
Challenges in translating genomics research

Raw data to interpretation

- Performance is different for detection of SNPs and INDELs
  - Different biases on variant calling (ignoring or adding REF allele)
  - Sequencing platform used and sequence coverage
  - Specificity and sensitivity based on truth set calls

Benchmarking is required
Challenges in translating genomics research

Raw data to interpretation

- What is true can be challenging
  - Genome-in-a-Bottle (GIAB) truth sets
- Databases: ExAC, dbSNP, DGV, ClinVar, dbGaP, ...
Truth/Gold standards available - GIAB/NIST

<table>
<thead>
<tr>
<th>GIAB (NA12878)</th>
<th>Number of bases in the high confidence region ($10^9$)</th>
<th># Truth SNPs + Indels in GIAB high confidence bed ($10^6$)</th>
<th>#Truth SNPs in GIAB high confidence bed ($10^6$)</th>
<th># Truth Indels in GIAB high confidence bed</th>
</tr>
</thead>
<tbody>
<tr>
<td>NISTv2.19</td>
<td>~ 2.22</td>
<td>~ 3.15</td>
<td>~ 2.79</td>
<td>365,459</td>
</tr>
<tr>
<td>NISTv3.2.2</td>
<td>~ 2.53</td>
<td>~ 3.51</td>
<td>~ 3.15</td>
<td>358,207</td>
</tr>
<tr>
<td>NISTv3.3</td>
<td>~ 2.57</td>
<td>~ 3.56</td>
<td>~ 3.19</td>
<td>371,889</td>
</tr>
</tbody>
</table>
## Integrated data to generate GIAB truth sets

<table>
<thead>
<tr>
<th>GIAB (NA12878)</th>
<th>Sequencing platforms used to generate the integrated data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>NISTv2.19</td>
<td>Illumina Gallx, Illumina HiSeq, 454, Complete Genomics, SOLiD, Ion Torrent</td>
</tr>
<tr>
<td>NISTv3.2.2</td>
<td>Illumina, SOLiD, Complete Genomics, Ion Torrent</td>
</tr>
<tr>
<td>NISTv3.3</td>
<td>Illumina, BioNano Genomics, Nabsys, Complete Genomics, 10X Genomics, Ion Proton, Oxford Nanopore, Pacific Biosciences, SOLiD</td>
</tr>
</tbody>
</table>
GIAB updates (Sep 2016)

- NISTv3.3 - phased calls and phased ID’s from GATK-HC
- New way to define callable sets that
  - excludes decoy and certain segmental duplications
  - includes some variant and homozygous reference calls found in >10bp homopolymers
- Includes data from Ashkenazim Jew Trio and Chinese son (Asian Trio)
Choosing variant calling pipeline(s)

- **MedGAP pipeline (in-house)**
  - BWA for alignment
  - GATK-HC for variant calling

- **Speedseq pipeline**
  - BWA for alignment
  - FreeBayes for variant calling

- **Future directions:** Novoalign or another aligner combined with variant callers
Choosing variant calling pipeline(s)

Currently WES data used
Possible data sources:
Personalis, Nextera

*Medical Genome Analysis Pipeline
Benchmarking variant calling pipeline(s)

Reference cell line (e.g., NA12878)

NIST/GIAB data integration

Truth set (GIAB)

Variant calling pipeline

Query data

Benchmarking (vcf comparison):
(1) Illumina’s hap.py  (2) Real Time Genomics’ vcfeval
What is unique about the pipeline?

Current clinical standards use Coding Exons, @ Stanford Health Care we also use Exons +/- 2 bases to capture rare variants in splice sites

<table>
<thead>
<tr>
<th></th>
<th>GIAB (NA12878)</th>
<th>Number of bases in the region of interest (10⁶)</th>
<th># Truth SNPs + Indels in GIAB high confidence bed</th>
<th>#Truth SNPs in GIAB high confidence bed</th>
<th># Truth Indels in GIAB high confidence bed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding Exons</td>
<td>NISTv3.3</td>
<td>~ 30.7</td>
<td>19,409</td>
<td>18,613</td>
<td>436</td>
</tr>
<tr>
<td>Exons +/-2 bases</td>
<td>NISTv3.3</td>
<td>~ 69.9</td>
<td>62,938</td>
<td>57,308</td>
<td>5,630</td>
</tr>
</tbody>
</table>
What is unique about the pipeline?

- Best practices in GATK and GA4GH
  - PrecisionFDA standards - rtg and hap.py tools
- Enabling reproducibility, traceability and scalability of the computational pipeline
  - Loom (workflow engine that ensures reproducibility)
  - Docker images for each step in a loom workflow
- Optimizing specificity and sensitivity – assessing against truth set by exploring tools’ parameter space
Clinical Genomics Service pipeline: Addressing Challenges

- Upgrading pipeline as versions of software/tools change
  - Variant callers
  - Benchmarking tools
  Docker helps isolate dependencies

- GIAB truth sets keep evolving
  - Frequent updates to exome pipeline (benchmarking)

- Improving specificity and sensitivity
  - Identifying variants in growing clinical databases (e.g., ExAC) to add to the GIAB truth set
ACCURACY BENCHMARKING
TESTING THE PERFORMANCE OF VARIANT CALLING PIPELINES
Benchmarking results for NA12878 Coding Exons using hap.py

Truth set used: GIAB - NISTv3.3, Query data source: Personalis

<table>
<thead>
<tr>
<th></th>
<th>Variant calling pipeline</th>
<th>#Truth total/#Query total</th>
<th># True positives (% of truth covered)</th>
<th>#False negatives</th>
<th>#False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indels</strong></td>
<td>MedGAP</td>
<td>428 / 366</td>
<td>346 (80.84%)</td>
<td>84</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Speedseq</td>
<td>428 / 389</td>
<td>364 (85.05%)</td>
<td>64</td>
<td>24</td>
</tr>
<tr>
<td><strong>SNPs</strong></td>
<td>MedGAP</td>
<td>18,304 / 16,945</td>
<td>16,911 (92.39%)</td>
<td>1,390</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Speedseq</td>
<td>18,304 / 18,233</td>
<td>18,058 (98.66%)</td>
<td>246</td>
<td>173</td>
</tr>
</tbody>
</table>
Benchmarking results for AJ Trio son Coding Exons using hap.py

Truth set used: GIAB (NA24385) - NISTv3.3, Query data source: Personalis

<table>
<thead>
<tr>
<th>Variant calling pipeline</th>
<th>#Truth total/ #Query total</th>
<th># True positives (% of truth covered)</th>
<th>#False negatives</th>
<th>#False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MedGAP</td>
<td>419 / 320</td>
<td>311 (74.22%)</td>
<td>111</td>
<td>9</td>
</tr>
<tr>
<td>Speedseq</td>
<td>419 / 385</td>
<td>377 (89.98%)</td>
<td>42</td>
<td>8</td>
</tr>
<tr>
<td>SNPs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MedGAP</td>
<td>17,667 / 16,198</td>
<td>16,180 (91.58%)</td>
<td>1,483</td>
<td>18</td>
</tr>
<tr>
<td>Speedseq</td>
<td>17,667 / 17,625</td>
<td>17,505 (99.08%)</td>
<td>162</td>
<td>120</td>
</tr>
</tbody>
</table>
Enabling Precision Medicine: Translational research

Clinical Genomics Service pipeline

- Sequencing platform(s)
- Identification of variants
- Evaluation of variant calling pipelines
- Assessing against truth sets – exploring tools’ parameter space

Continuous Benchmarking

Clinical data interpretation

Research institution

Medical institution

Patient care
Acknowledgements

**Stanford Center for Genomics & Personalized Medicine:**
Mike Snyder (Director), Somalee Datta (Director of Bioinformatics), Isaac Liao (Software Engineer, Loom), Amin Zia (Sr. Bioinformatics Scientist)

**Clinical Genomics Service @ Stanford Health Care:**
Euan Ashley & Jason Merker (Directors), Sowmi Utiramerur (Director of Bioinformatics), Nathan Hammond (Sr. Scientist, Loom), Lalitha Viswanathan (Sr. Pipeline Engineer, Workflows)
Thank you!