Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Gholson J. Lyon, M.D. Ph.D.
Results from Exome and WGS requires both Analytic and Clinical Validity

• Analytical Validity: the test is accurate with high sensitivity and specificity.

• Clinical Validity: Given an accurate test result, what impact and/or outcome does this have on the individual person?

Illusions of Certainty. Everything is Probabilistic.
CALCULATED RISKS
HOW TO KNOW WHEN NUMBERS DECEIVE YOU
GERD GIGERENZER
In the fields of science, engineering, industry, and statistics, the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. The precision of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results.

http://en.wikipedia.org/wiki/Accuracy_and_precision
<table>
<thead>
<tr>
<th></th>
<th>Accurate</th>
<th>Inaccurate (systematic error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precise</td>
<td><img src="image1.png" alt="Target" /></td>
<td><img src="image2.png" alt="Target" /></td>
</tr>
<tr>
<td>Imprecise</td>
<td><img src="image3.png" alt="Target" /></td>
<td><img src="image4.png" alt="Target" /></td>
</tr>
</tbody>
</table>
**Accuracy**

\[
\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}
\]

An accuracy of 100% means that the measured values are exactly the same as the given values.
“ground truth” Genome from blood of one person (of course, that is from millions of cells and only blood, not other tissues)

~3 billion nucleotides
Exon = set of contiguous nucleotides predicted to contribute toward a protein

“exon capture and sequencing”
“Exome”
Chose to sequence 15 “exomes”
2-3 rounds of sequencing at BGI to attain goal of >80% of target region at >20 reads per base pair

<table>
<thead>
<tr>
<th>Exome Capture Statistics</th>
<th>K24510-84060</th>
<th>K24510-92157-a</th>
<th>K24510-84615</th>
<th>K24510-88962</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target region (bp)</td>
<td>46,401,121</td>
<td>46,401,121</td>
<td>46,401,121</td>
<td>46,257,379</td>
</tr>
<tr>
<td>Raw data yield (Mb)</td>
<td>12,490</td>
<td>14,571</td>
<td>14,129</td>
<td>9,398</td>
</tr>
<tr>
<td>Reads mapped to genome</td>
<td>110,160,277</td>
<td>135,603,094</td>
<td>135,087,576</td>
<td>83,942,646</td>
</tr>
<tr>
<td>Reads mapped to target region</td>
<td>68,042,793</td>
<td>84,379,239</td>
<td>80,347,146</td>
<td>61,207,116</td>
</tr>
<tr>
<td>Data mapped to target region (Mb)</td>
<td>5,337.69</td>
<td>6,647.18</td>
<td>6,280.01</td>
<td>4,614.47</td>
</tr>
<tr>
<td>Mean depth of target region</td>
<td>115.03</td>
<td>143.25</td>
<td>135.34</td>
<td>99.76</td>
</tr>
<tr>
<td>Coverage of target region (%)</td>
<td>0.9948</td>
<td>0.9947</td>
<td>0.9954</td>
<td>0.9828</td>
</tr>
<tr>
<td>Average read length (bp)</td>
<td>89.91</td>
<td>89.92</td>
<td>89.95</td>
<td>89.75</td>
</tr>
<tr>
<td>Fraction of target covered &gt;=4X</td>
<td>98.17</td>
<td>98.38</td>
<td>98.47</td>
<td>94.25</td>
</tr>
<tr>
<td>Fraction of target covered &gt;=10X</td>
<td>95.18</td>
<td>95.90</td>
<td>95.97</td>
<td>87.90</td>
</tr>
<tr>
<td>Fraction of target covered &gt;=20X</td>
<td>90.12</td>
<td>91.62</td>
<td>91.75</td>
<td>80.70</td>
</tr>
<tr>
<td>Fraction of target covered &gt;=30X</td>
<td>84.98</td>
<td>87.42</td>
<td>87.67</td>
<td>74.69</td>
</tr>
<tr>
<td>Capture specificity (%)</td>
<td>61.52</td>
<td>62.12</td>
<td>59.25</td>
<td>73.16</td>
</tr>
<tr>
<td>Fraction of unique mapped bases on or near target</td>
<td>65.59</td>
<td>65.98</td>
<td>63.69</td>
<td>85.46</td>
</tr>
<tr>
<td>Gender test result</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>F</td>
</tr>
</tbody>
</table>
Depth of Coverage in 15 exomes > 20 reads per bp in target region

<table>
<thead>
<tr>
<th>Coverage depth</th>
<th>&gt;=1</th>
<th>&gt;=10</th>
<th>&gt;=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction of target covered (%)</td>
<td>100</td>
<td>90</td>
<td>80</td>
</tr>
</tbody>
</table>

The diagram shows the fraction of the target covered for different coverage depths in 15 exomes.
Experimental Design

- Evaluate robustness of variant calling implemented by different bioinformatics analysts.

- Looking at False Positives and False Negatives.

- How reliable are variants that are uniquely called by individual pipelines?

- Are some pipelines better at detecting rare, or novel variants than others?
Promotion Details (valid for Americas and Europe customers NOW through MAY 31)

A. The 899 USD/sample package – 50X human exome sequencing
   
   Agilent SureSelect 50/51M Capture kit
   100 bp paired-end sequencing on HiSeq 2000
   5 Gb high quality* sequencing data
   50X average coverage for target regions guaranteed
   SNP & Indel calling and annotation included

B. The 1299 USD/sample package – 100X human exome sequencing
   
   Agilent SureSelect 50/51M Capture kit
   100 bp paired-end sequencing on HiSeq 2000
   10 Gb high quality* sequencing data
   100X average coverage for target regions guaranteed
   SNP & Indel calling and annotation included
Table 1. A descriptive summary of the variant calling pipelines included in the comparative analyses.

<table>
<thead>
<tr>
<th>Pipeline name</th>
<th>Alignment method</th>
<th>Variant-calling module</th>
<th>Description of variant-calling algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOAP</td>
<td>SOAPaligner</td>
<td>SOAPsnp version 1.03/ SOAPindel version 2.01</td>
<td>SOAP uses a method based on Bayes’ theorem to call consensus genotype by carefully considering the data quality, alignment, and recurring experimental errors [22].</td>
</tr>
<tr>
<td>GATK version 1.5</td>
<td>BWA version 0.5.9</td>
<td>UnifiedGenotyper version 1.5</td>
<td>GATK employs a general Bayesian framework to distinguish and call variants. Error correction models are guided by expected characteristics of human variation to further refine variant calls [19].</td>
</tr>
<tr>
<td>SNVer version 0.2.1</td>
<td>BWA version 0.5.9</td>
<td>SNVer version 0.2.1</td>
<td>SNVer uses a more general frequentist framework, and formulates variant calling as a hypothesis-testing problem [25].</td>
</tr>
<tr>
<td>GNUMAP version 3.1.0</td>
<td>GNUMAP version 3.1.0</td>
<td>GNUMAP version 3.1.0</td>
<td>GNUMAP incorporates the base uncertainty of the reads into mapping analysis using a probabilistic Needleman-Wunsch algorithm [24].</td>
</tr>
<tr>
<td>SAMtools version 0.1.18</td>
<td>BWA version 0.5.9</td>
<td>mpileup version 0.1.18</td>
<td>SAMtools [20] calls variants by generating a consensus sequence using the MAQ model framework, which uses a general Bayesian framework for picking the base that maximizes the posterior probability with the highest Phred quality score.</td>
</tr>
</tbody>
</table>
Accuracy \hspace{1cm} Precision

Error-Prone & Doubtful \hspace{1cm} Repeatable & Reproducible

B) Mean # of known SNVs (present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the Venn diagram is the percent of known SNVs called by all five pipelines.
C) Mean # of novel SNVs (not present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the Venn diagram is the percent of novel SNVs called by all five pipelines.
Indels called by GATK, SOAP and SAMtools
“ground truth” exons from blood of one person
Cross validation using orthogonal sequencing technology (Complete Genomics)
Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)
Accuracy of Complete Genomics Whole Human Genome Sequencing Data
Analysis Pipeline v2.0

Table 2. Concordance of Technical Replicates.

<table>
<thead>
<tr>
<th>COMPLETE GENOMICS CALL</th>
<th>OTHER PLATFORM</th>
<th>PLATFORM-SPECIFIC SNVs</th>
<th>VALIDATION RATE</th>
<th>EST FPs</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Het or Hom SNV</td>
<td>No SNV Reported</td>
<td>99K</td>
<td>17/18 = 94.4%</td>
<td>5,577</td>
<td>0.16%</td>
</tr>
<tr>
<td>No-call or Hom-Ref</td>
<td>SNV Reported</td>
<td>345K</td>
<td>2/15 = 13.3%</td>
<td>299,115</td>
<td>8.2%</td>
</tr>
</tbody>
</table>

Table 3. False Positive Rate.

Performance comparison of whole-genome sequencing platforms

Hugo Y K Lam, Michael J Clark, Rui Chen, Rong Chen, Georges Natsoulis, Maeve O’Huallachain, Frederick E Dewey, Lukas Habegger, Euan A Ashley, Mark B Gerstein, Atul J Butte, Hanlee P Ji & Michael Snyder
Union of all SNVs from 5 pipelines

Union of all indels from 3 pipelines
What is the "True" Personal Genome?

Illumina SNVs

- 35653
- 18331 (48.6%)

CG SNVs

- 19407
- 2085 (5.5%)

Illumina indels

- 4364
- 2666 (50.5%)

CG Indels

- 2613
- 915 (17.3%)

- 1698 (32.2%)
Genomic Dark Matter: The reliability of short read mapping illustrated by the Genome Mappability Score

Hayan Lee¹,²* and Michael C. Schatz¹,²

¹Department of Computer Science, Stony Brook University, Stony Brook, NY
²Simons Center for Quantitive Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Bioinformatics Advance Access published June 4, 2012

• Genome Mappability Score (GMS) -- measure of the complexity of resequencing a genome = a weighted probability that any read could be unambiguously mapped to a given position, and thus measures the overall composition of the genome itself.
Known Novel
Percent of Illumina SNVs validated by CG data

Personal genome “K8101”

All
- GATK
- GNUMAP
- SAMTools
- SNVer
- SOAPsnp

0 10 20 30 40 50 60 70 80 90 100

Percent of Illumina SNVs validated by CG data
Higher Validation by CG of SNVs with the BWA-GATK(v1.5) pipeline

- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK(v1.5), in comparison to the other 4 pipelines (including SOAP).
Validating Indels with Complete Genomics Data for the 3 pipelines
Comparing to New Versions of GATK
Validation of SNVs and Indels called by GATK, SOAP and both, with another platform
Validation with PCR amplicons and MiSeq 150 bp reads at ~5000x coverage

1,140 SNVs, with random sampling of 380 from the set of unique-to-GATK SNVs, 380 from the set of unique-to-SOAPsnp SNVs, and 380 from the set that were overlapping between these two pipelines.

960 indels, with random sampling of 386 from the unique-to-GATK indel set, 387 from the unique-to-SOAPindel set, and 187 from the set of indels overlapping between the two (SOAPindel and GATK).
Validation of ~2000 PCR amplicons with PacBio reads from two SMRT cells (~50,000 useable reads per cell)

- **GATK v1.5 indel validation**
  - Validated: 216 / 369
  - Not validated: 153 / 369

- **SOAPindel v2.01 indel validation**
  - Validated: 220 / 365
  - Not validated: 145 / 365

- **Validation of overlapping indels (GATK and SOAPindel)**
  - Validated: 74 / 183
  - Not validated: 109 / 183

- **GATK v1.5 SNV validation**
  - Validated: 289 / 357
  - Not validated: 68 / 357

- **SOAPsnp v1.03 SNV validation**
  - Validated: 206 / 339
  - Not validated: 133 / 339

- **Validation of overlapping SNVs (GATK and SOAPsnp)**
  - Validated: 336 / 375
  - Not validated: 39 / 375
Validation of overlapping indels (GATK and SOAPindel)

GATK v1.5 indel validation

SOAPindel v2.01 indel validation

Validate
Not validated

Validation of overlapping SNVs (GATK and SOAPsnp)

GATK v1.5 SNV validation

SOAPsnp v1.03 SNV validation

Validate
Not validated

MiSeq

GATK v1.5 indel validation

SOAPindel v2.01 indel validation

Validation of overlapping indels (GATK and SOAPindel)

GATK v1.5 SNV validation

SOAPsnp v1.03 SNV validation

Validation of overlapping SNVs (GATK and SOAPsnp)

PacBio
What is the “True” Personal Genome?

Illumina SNVs
- 35653
- 18331 (48.6%)
- 17322 (45.9%)
- 2085 (5.5%)

Illumina indels
- 4364
- 2666 (50.5%)
- 1698 (32.2%)
- 915 (17.3%)

CG SNVs
- 19407

CG Indels
- 2613
We looked for SNVs that were detected in children but not in parents using 3 different strategies:

1. We used all of the SNVs that were detected by all 5 pipelines for both parents and children.
2. We used all of the detected SNVs for parents, but only the concordant SNVs between the 5 different pipelines for children.
3. We used SNVs concordant between the 5 different pipelines for children and parents.
Age 49
Possible Motor Tic, but no diagnosis
YGTSS 6
OCD w/ YBOCS 25

Age 22
OCD-mild
YBOCS 18

Age 19
OCD-mild
YBOCS 14

Age 24
TS- definite
YGTSS 47
OCD? ADHD?

TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

No Tics
OCD-mild
ADHD
Age 22
YBOCS 14

No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 14

No tics yet
Subclinical OCD
Age 14
YBOCS 12

Mild OCD w YBOCS 14
Possible ADHD

-01 88458
Age 51
NO TICS

-03 88460
TS
ADHD, definite

-02 88459
Age 49
Possible Motor Tic, but no diagnosis
YGTSS 6
OCD w/ YBOCS 25

-04 88461
No tics yet
Subclinical OCD
Age 14
YBOCS 12

-05 89587
No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 14

-06 89588
No tics
OCD-mild
ADHD
Age 22
YBOCS 18

-07 91583
Age 79, TS- definite,
YGTSS 47
OCD? ADHD?
-01 88458
Age 51
NO TICS
Mild OCD w YBOCS 14
Possible ADHD

-02 88459
Age 49
Possible Motor Tic, but no diagnosis
YGTSS 6
OCD w/ YBOCS 25

-03 88460
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

-06 89588
No Tics
OCD-mild
ADHD
Age 22
YBOCS 18

-05 89587
No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 18

-04 88461
No tics yet
Subclinical OCD
Age 14
YBOCS 12

-07 91583
Age 79, TS- definite,
YGTSS 47
OCD? ADHD?
88458
Age 51
NO TICS
Mild OCD w YBOCS 14
Possible ADHD

88460
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

89588
No Tics
OCD-mild
ADHD
Age 22
YBOCS 18

89587
No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 18

88461
No tics yet
Subclinical OCD
Age 14
YBOCS 12

91583
Age 79, TS-
definite,
YGTSS 47
OCD? ADHD?

??

??

??

??

TDT- 09 -1018
K26679

“Parents”
Analysis based on various pipelines

• “Parents” in this case means the mother, father AND grandmother.

• Taking the Union of SNVs from all 5 pipelines from “Parents”, and subtract that from the Union of all SNVs in each child.

• Or Subtract the Union of these “Parents” from the SNVs in the child concordant between 5 pipelines.

• Or, subtract the concordant variants from 5 pipelines in “Parents” from the concordant variants for 5 pipelines in each child.
Table 3. *De novo* single-nucleotide variants (SNVs) were detected in two families contained within the 15 study exomes.

<table>
<thead>
<tr>
<th></th>
<th>Number of putative <em>de novo</em> coding non-synonymous or nonsense SNVs detected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without using the grandparents as a filter</td>
</tr>
<tr>
<td><strong>Family 1</strong></td>
<td></td>
</tr>
<tr>
<td>Child A</td>
<td>241</td>
</tr>
<tr>
<td>Child B</td>
<td>211</td>
</tr>
<tr>
<td>Child C</td>
<td>102</td>
</tr>
<tr>
<td>Child D</td>
<td>242</td>
</tr>
<tr>
<td><strong>Family 2</strong></td>
<td></td>
</tr>
<tr>
<td>Child A</td>
<td>49</td>
</tr>
<tr>
<td>Child B</td>
<td>41</td>
</tr>
</tbody>
</table>

<sup>a</sup>N/A, no grandparent available.

Family 1 had a grandparent available for filtering purposes, whereas family 2 did not. To minimize false positives in the pool of SNVs associated with each child, only highly concordant SNVs were used (SNVs detected by all five pipelines). To construct a comprehensive set of SNVs for each parent, and hence increase filtering accuracy, false negatives for parent SNVs were reduced by taking the union of all SNV calls from all five pipelines.
In the fields of science, engineering, industry, and statistics, the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. The precision of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results.

http://en.wikipedia.org/wiki/Accuracy_and_precision
Conclusions

• Sequencing a grandparent seems to help eliminate errors derived from the current depth of sequencing coverage in the mother and father.

• For now, we advocate using more than one pipeline on one set of sequencing data, but we expect the field to move toward >2 sequencing platforms per sample.

• Still need substantial work on indel-calling and validation.
Acknowledgements

Cold Spring Harbor Laboratory
STANLEY INSTITUTE FOR COGNITIVE GENOMICS

Jason O’Rawe

Kai Wang

Michael Schatz
Giuseppe Narzisi
Eric Antoniou
Dick McCombie
Sequencing core facility

Yiyang Wu

Tao Jiang
Guangqing Sun
EXTRA SLIDES – Not Shown
Indels called by GATK, SOAP and SAMtools

![Venn Diagram showing the overlap of Indels called by GATK, SOAP, and SAMtools.]

- GATK: 1114 (28.5%)
- SOAP: 278 (7.1%)
- SAMTools: 305 (7.8%)
- Overlapping with all three: 20 (0.5%)
- Unique to each tool:
Validation of SNVs and Indels with an additional platform
Additional file 2, Figure S3. Histograms of Illumina read depth taken from each pipeline’s independently aligned BAM file at genomic coordinates of SNVs called by each of the 5 alignment and variant calling pipelines. A) SOAPsnp, B) SNVer, C) SAMTools, D) GNUMAP and E) GATK, respectively. Frequency of read depths for all SNVs (A, B, C, D, and E) as well as for SNVs having depths between 0 and 50 (a, b, c, d, and e) were plotted.
Additional file 2, Figure S4. SNV concordance for a single exome, “k8101-49685”, between five alignment and variant detection pipelines: GATK, SOAPsnp, SNVer, SAMTools, and GNUMAP. Concordance between each pipeline was determined by matching the genomic coordinate as well as the base pair change and zygosity for each detected SNV. Concordance was measured at varying Illumina read depth threshold values in each independently aligned BAM file, ranging from >0 (no threshold) to >30 reads.
Additional file 2, Figure S5. Histograms of read depth taken from each of the five Illumina pipeline’s independently aligned BAM file at genomic coordinates of SNVs that were found by Complete Genomics but not by any of the 5 Illumina pipelines: GATK, GNUMAP, SNVer, SAMTools and SOAPsnp, A, B, C, D and E respectively. All coordinates fell within the range of the Agilent SureSelect v.2 exons.
Additional file 2, Figure S8. SNVs called by each Illumina-data pipeline were cross-validated using SNVs called by Complete Genomics, an orthogonal sequencing technology, in sample “k8101-49685”. The percentage of Illumina SNVs that were validated by CG sequencing was measured for variants having varying degrees of Illumina-data pipeline concordance. The same analysis was performed for variants that were considered novel (absent in dbSNP135).
Additional file 2, Figure S9. Indels called by each Illumina-data pipeline were cross-validated using indels called by Complete Genomics for sample “k8101-49685”. The percentage of Illumina indels that were validated by CG sequencing was measured across varying degrees of Illumina pipeline concordance. The same analysis was done for novel indels (indels not found in dbSNP 135).
Comparing the concordance among the 5 pipelines used to analyze Illumina data, also stratified by read depth from >0 to >30 reads.
<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean*</td>
<td>SD</td>
</tr>
<tr>
<td>SOAPsnp</td>
<td>94.68</td>
<td>2.26</td>
</tr>
<tr>
<td>GATK1.5</td>
<td>95.34</td>
<td>1.16</td>
</tr>
<tr>
<td>SNVer</td>
<td>92.33</td>
<td>4.40</td>
</tr>
<tr>
<td>GNUmap</td>
<td>86.60</td>
<td>3.23</td>
</tr>
<tr>
<td>SAMtools</td>
<td>94.47</td>
<td>4.22</td>
</tr>
<tr>
<td>Any pipeline</td>
<td>97.67</td>
<td>1.20</td>
</tr>
<tr>
<td>≥2 pipelines*</td>
<td>96.64</td>
<td>2.28</td>
</tr>
<tr>
<td>≥3 pipelines*</td>
<td>95.62</td>
<td>3.13</td>
</tr>
<tr>
<td>≥4 pipelines*</td>
<td>92.60</td>
<td>3.40</td>
</tr>
<tr>
<td>5 pipelines*</td>
<td>80.58</td>
<td>5.26</td>
</tr>
</tbody>
</table>

*Intersection of variants contained in the number of pipelines specified.

Sensitivity and specificity was calculated for each pipeline by comparing Illumina Human610-Quad version 1 SNP arrays with exome-capture sequencing results, based on the four samples whose genotyping data was available.
Table S1. Concordance rates with common SNPs genotyped on Illumina 610K genotyping chips.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Software</th>
<th>Compared Sites</th>
<th>Concordance Sites</th>
<th>Concordance rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother-1</td>
<td>SOAPsnp</td>
<td>6088</td>
<td>6074</td>
<td>99.77%</td>
</tr>
<tr>
<td></td>
<td>GATK 1.5</td>
<td>6249</td>
<td>6224</td>
<td>99.60%</td>
</tr>
<tr>
<td></td>
<td>SNVer</td>
<td>5723</td>
<td>5708</td>
<td>99.74%</td>
</tr>
<tr>
<td></td>
<td>GNUMAP</td>
<td>5458</td>
<td>5434</td>
<td>99.56%</td>
</tr>
<tr>
<td></td>
<td>SAMTools</td>
<td>5885</td>
<td>5848</td>
<td>99.37%</td>
</tr>
<tr>
<td>Son-1</td>
<td>SOAPsnp</td>
<td>6366</td>
<td>6353</td>
<td>99.80%</td>
</tr>
<tr>
<td></td>
<td>GATK 1.5</td>
<td>6341</td>
<td>6323</td>
<td>99.72%</td>
</tr>
<tr>
<td></td>
<td>SNVer</td>
<td>6255</td>
<td>6239</td>
<td>99.74%</td>
</tr>
<tr>
<td></td>
<td>GNUMAP</td>
<td>5850</td>
<td>5828</td>
<td>99.62%</td>
</tr>
<tr>
<td></td>
<td>SAMTools</td>
<td>6383</td>
<td>6362</td>
<td>99.67%</td>
</tr>
<tr>
<td>Son-2</td>
<td>SOAPsnp</td>
<td>6412</td>
<td>6401</td>
<td>99.83%</td>
</tr>
<tr>
<td></td>
<td>GATK 1.5</td>
<td>6426</td>
<td>6413</td>
<td>99.80%</td>
</tr>
<tr>
<td></td>
<td>SNVer</td>
<td>6336</td>
<td>6325</td>
<td>99.83%</td>
</tr>
<tr>
<td></td>
<td>GNUMAP</td>
<td>5906</td>
<td>5889</td>
<td>99.71%</td>
</tr>
<tr>
<td></td>
<td>SAMTools</td>
<td>6477</td>
<td>6450</td>
<td>99.58%</td>
</tr>
<tr>
<td>Father-1</td>
<td>SOAPsnp</td>
<td>6247</td>
<td>6238</td>
<td>99.86%</td>
</tr>
<tr>
<td></td>
<td>GATK 1.5</td>
<td>6304</td>
<td>6288</td>
<td>99.75%</td>
</tr>
<tr>
<td></td>
<td>SNVer</td>
<td>6205</td>
<td>6192</td>
<td>99.79%</td>
</tr>
<tr>
<td></td>
<td>GNUMAP</td>
<td>5805</td>
<td>5786</td>
<td>99.67%</td>
</tr>
<tr>
<td></td>
<td>SAMTools</td>
<td>6344</td>
<td>6327</td>
<td>99.73%</td>
</tr>
</tbody>
</table>

All pipelines are very good with identifying already known, common SNPs.
Taking SNVs concordant in 5 Illumina pipelines, and comparing to SNVs in Complete Genomics Data from same sample
Taking SNVs concordant in 5 Illumina pipelines as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample
Taking SNVs found by ALL 5 Illumina pipelines (Union), and comparing to SNVs in Complete Genomics Data from same sample
Taking the UNION of all SNVs called by Illumina pipelines, as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample.
Comparing the UNION versus the CONCORDANCE of 5 pipelines to the Complete Genomics Data

Union of Illumina variants

Concordant Illumina variants
Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GATK or SOAP pipelines

- Read depth of SNVs called by CG and not GATK
- Read depth of SNVs called by CG and not SOAPsnp

Read depth taken from GATK bam file

Read depth taken from SOAPsnp bam file
Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GNUMAP, SNVer or SamTools pipelines

Read depth of SNVs called by CG and not GNUMAP
Read depth taken from GNUMAP bam file

Read depth of SNVs called by CG and not SNVer
Read depth taken from SNVer bam file

Read depth of SNVs called by CG and not SAMTools
Read depth taken from SAMTools bam file
Genomic Dark Matter, cont....

• That means that unlike typical false negatives, increasing coverage will not help identify mutations in low GMS regions, even with 0% sequencing error.

• Instead this is because the SNP-calling algorithms use the mapping quality scores to filter out unreliable mapping assignments, and low GMS regions have low mapping quality score (by definition). Thus even though many reads may sample these variations, the mapping algorithms cannot ever reliably map to them.

• Since about 14% of the genome has low GMS value with typical sequencing parameters, it is expected that about 14% of all variations of all resequencing studies will not be detected.

• To demonstrate this effect, we characterised the SNP variants identified by the 1000 genomes pilot project, and found that 99.99% of the SNPs reported were in high GMS regions of the genome, and in fact 99.95% had GMS over 90.
Figure 1. Mean single-nucleotide variants (SNV) concordance over 15 exomes between five alignment and variant-calling pipelines. The alignment method used, followed by the SNV variant calling algorithm is annotated here in shorthand: BWA-GATK, SOAP-Align-SOAPsnp, BWA-SNVer, BWA-SAMtools, and GNUMAP-GNUMAP. (A) Mean SNV concordance between each pipeline was determined by matching the genomic coordinate as well as the base-pair change and zygosity for each detected SNV. (B) The same analysis as in (A) but filtered to include only SNVs already found in dbSNP135. (C) The same analysis as in (A), but filtered to include novel SNVs (that is, SNVs not found in dbSNP135).
Pipelines Used on Same Set of Seq Data by Different Analysts, using Hg19 Reference Genome

1) BWA - Sam format to Bam format - Picard to remove duplicates - GATK (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper.

2) BWA - Sam format to Bam format-Picard to remove duplicates - SamTools version 0.1.18 to generate genotype calls -- The “mpileup” command in SamTools were used for identify SNPs and indels.

3) SOAP-Align – SOAPsnp – then BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph )

4) GNUMAP-SNP (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion)

5) BWA - Sam format to Bam format - Picard to remove duplicates - SNVer

6) BWA - Sam format to Bam format - Picard to remove duplicates - SCALPEL
Pipelines Used on Same Set of Seq Data by Different Analysts, using Hg19 Reference Genome

1) **BWA - GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper. For SNVs and indels.

2) **BWA - SamTools** version 0.1.18 to generate genotype calls -- The “mpileup” command in SamTools was used for identify SNVs and indels.

3) **SOAP-Align** – SOAPsnp for SNVs– and BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph) for indels.

4) **GNUMAP-SNP** (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion), for SNVs only.

5) **BWA - Sam format to Bam format - Picard to remove duplicates – SNVer**, for SNVs only
<table>
<thead>
<tr>
<th></th>
<th>All SNVs, both for parents and children, were considered</th>
<th>All parental SNVs that were detected were considered. Only SNVs concordant between the 5 pipelines were considered for children</th>
<th>SNVs concordant between 5 pipelines for children and parents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SNVs found in child A but not in parents</td>
<td>1057</td>
<td>2</td>
<td>637</td>
</tr>
<tr>
<td>Number of SNVs found in child B but not in parents</td>
<td>1084</td>
<td>1</td>
<td>672</td>
</tr>
<tr>
<td>Number of SNVs found in child C but not in parents</td>
<td>2363</td>
<td>20</td>
<td>1703</td>
</tr>
<tr>
<td>Number of SNVs found in child D but not in parents</td>
<td>1518</td>
<td>5</td>
<td>876</td>
</tr>
<tr>
<td>Number of nonsyn SNVs in child A but not in parents</td>
<td>411</td>
<td>1</td>
<td>150</td>
</tr>
<tr>
<td>Number of nonsyn SNVs in child B but not in parents</td>
<td>396</td>
<td>0</td>
<td>135</td>
</tr>
<tr>
<td>Number of nonsyn SNVs in child C but not in parents</td>
<td>911</td>
<td>6</td>
<td>459</td>
</tr>
<tr>
<td>Number of nonsyn SNVs in child D but not in parents</td>
<td>619</td>
<td>3</td>
<td>225</td>
</tr>
<tr>
<td>Number of shared nonsyn SNVs in the children, but not in parents</td>
<td>8</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>
Optimizing pipeline based on literature value of \( \sim 1 \) true de novo protein-altering mutation per exome

<table>
<thead>
<tr>
<th></th>
<th>All SNVs, both for parents and children, were considered</th>
<th>All parental SNVs that were detected were considered. Only SNVs concordant between the 5 pipelines were considered for children</th>
<th>SNVs concordant between 5 pipelines for children and parents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SNVs found in child A but not in parents</td>
<td>1308</td>
<td>186</td>
<td>1795</td>
</tr>
<tr>
<td>Number of SNVs found in child B but not in parents</td>
<td>1332</td>
<td>161</td>
<td>1762</td>
</tr>
<tr>
<td>Number of nonsyn SNVs in child A but not in parents</td>
<td>381</td>
<td>52</td>
<td>420</td>
</tr>
<tr>
<td>Number of nonsyn SNVs in child B but not in parents</td>
<td>392</td>
<td>42</td>
<td>394</td>
</tr>
<tr>
<td>Number of shared nonsyn SNVs in the children, but not in parents</td>
<td>98</td>
<td>14</td>
<td>171</td>
</tr>
</tbody>
</table>

The result is that using all of the detected SNVs for both parents and children should minimize the false negative rate but similarly show a relatively high false positive rate. Using all of the SNVs detected for parents but only the SNVs concordant among the five pipelines shows mutation rates similar to those reported by the literature and is expected to have moderate false positive rates and moderate false negative rates. Using only the SNVs concordant among the 5 different pipelines for both parents and children should minimize the false positive rate but similarly show a relatively high false negative rate.
Much Higher Validation of the Concordantly Called SNVs (by the CG data)