Variant Detection in Next Generation Sequencing Data

John Osborne
Sept 14, 2012
Overview

- **My Bias**
  - Talk slanted towards analyzing whole genomes using Illumina paired end reads with open source tools

- **Background**

- **Alignment Software**

- **Detecting Variation**
  - Nucleotide
  - Structural

- **Analyzing and Interpreting Variation**
  - *Best practices change extremely rapidly*
Next Generation Sequencing

Figure 1: Conceptual Overview of Whole-Genome Resequencing

A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

Taken from Illumina website
Short Read Alignment

- Making comparisons is very difficult!

- Test Parameters
  - Read length size
  - Introduced errors
  - Paired versus single end reads

- Metrics
  - Discovery? Accuracy? Area under curve?
  - What is correct?

- Downstream analysis

- Comparisons are time consuming to do and are typically only done when somebody releases a new aligner
From homepage of Heng Li:  http://lh3lh3.users.sourceforge.net/alnROC.shtml
Short Read Aligner Conclusions

- The differences between aligners are not that large anymore
  - BWA, Bowtie2 are all available on cheaha

- I currently recommend BWA, but I suspect it will be supplanted by something else
  - Bowtie2
  - Novoalign
  - SeqAlto or something newer

- For longer reads (>=200bp) I would recommend BWA-SW, Bowtie2 (long read version) or CUSHAW2 (new)

- Select your aligner based on your downstream workflow, for example use of BWA is recommended by GATK
HugeSeq Workflow

1. Mapping
   - Reads
   - Dividing reads
     - Set 1
     - ... Set n
   - Gapped alignment
     - BWA mapping
     - Bam generation
   - Aligned BAM 1
   - ... Aligned BAM n

2. Sorting
   - Sorting by chromosomes
     - chr1 BAM
     - ... chrM BAM
   - Cleanup
     - Duplicating removal
     - Local realignment
     - Base recalibration
   - Cleaned chr1 BAM
   - ... Cleaned chrM BAM

3. Reduction
   - Detecting variants
     - SNP/indel
     - GATK
     - SAMtools
   - SV/CNV
     - Pindel (SR)
     - CNVnator (RD)
     - BreakDancer (RP)
     - BreakSeq (JM)
   - Final
     - Combining and merging
     - Functional annotation
   - SNP/indel (VCF)
   - SV/CNV (GFF)

From “Detecting and annotating genetic variations using the HugeSeq pipeline” Lam et al., 2012
Variation Detection

- Nucleotide Polymorphisms
- “Structural Variants” / Rearrangements

```plaintext
#fileformat=VCFv4.1
#fileDate=20090805
##source=myImputationProgramV3.1
##reference=file://seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66b6b2da,species="Homo sapiens"
#phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description=" Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS  ID  REF  ALT  QUAL  FILTER  INFO FORMAT  NA0000
20   14370  rs6054257 G   A   29  PASS  NS=3;DP=14;AF=0.5;DB;H2  GT:GQ:DP:HQ 0:0:4
20   17330  .     T   A   3   q10  NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0:0:4
20   1110696 rs6040355 A   G,T  67  PASS  NS=2;DP=10;AF=0.333,0.667;AA=T,DB GT:GQ:DP:HQ 1:2:2
20   1230237  .     T   .   47  PASS  NS=3;DP=13;AA=T GT:GQ:DP:HQ 0:0:5
20   1234567 microsat1 GTC G,GTCT 50  PASS  NS=3;DP=9;AA=G GT:GQ:DP 0/1:3
```
SNP Detection

- Most advanced and reliable variant detection
  - New version of GATK can detect MNPs as well
- Coverage and Toolkit matter
- Problem isn’t finding SNPs, it is finding the right SNPs
Coverage versus Sensitivity

From “Detecting and annotating genetic variations using the HugeSeq pipeline” Lam et al., 2012
GATK and SAMTools Variant Calling

- From “Detecting and annotating genetic variations using the HugeSeq pipeline” Lam et al., 2012
False Positives

- Data from human monozygotic twins
- Artifacts from borderline low coverage, top twin has 17 high quality reads (7 A) and the bottom has 23 high quality reads (2 A)
Structural Variation

Methods

- Small indels within single reads (GATK)
- Discordant paired-end reads (Breakdancer, VariationHunter)
- Depth of coverage (CNVnator, SegSeq)
- Split reads (Pindel, ClipCrop)

Very active area of research

- Combined approaches becoming more common
Neither deletion was detected by Pindel..
Pitfalls of Structural Variant Detection with NGS

Tips

- Get as much coverage as possible
  - Not possible to find breakpoints with 5 fold coverage
- Use multiple approaches
- Remove duplicates
- If it is important and you have time... look
  - In twin study, only 2 out of 12 SVs found by Pindel

Personal Bias

- GATK (small indels), Breakdancer (rearrangements), Pindel (split reads) and CNVator (repeat size estimation)
Interpreting Variation

- Getting some variants is easy, analyzing them is hard
- Commonly used tools in CCTS
  - IGV, BedTools, VCFTools, SNPEff
- Pipelines are becoming more popular
  - Annovar (Sift, Polyphen2)
- Online Resources
Questions?
“Fast gapped-read alignment with Bowtie 2”, Langmead and Saltzburg (2012)
### (a) Samtools SNPs Called

<table>
<thead>
<tr>
<th>Aligner</th>
<th>Called</th>
<th>%Correct</th>
<th>%Discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqAlto</td>
<td>412547</td>
<td>97.469</td>
<td>95.259</td>
</tr>
<tr>
<td>Snap</td>
<td>416288</td>
<td>96.672</td>
<td>95.336</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>399420</td>
<td>98.521</td>
<td>93.223</td>
</tr>
<tr>
<td>BWA</td>
<td>410085</td>
<td>97.924</td>
<td>95.132</td>
</tr>
<tr>
<td>Stampy</td>
<td>410682</td>
<td>97.859</td>
<td>95.207</td>
</tr>
<tr>
<td>Novoalign</td>
<td>415850</td>
<td>97.088</td>
<td>95.646</td>
</tr>
</tbody>
</table>

### (b) Samtools Indels Called

<table>
<thead>
<tr>
<th>Aligner</th>
<th>Called</th>
<th>%Correct</th>
<th>%Discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqAlto</td>
<td>21949</td>
<td>99.932</td>
<td>98.329</td>
</tr>
<tr>
<td>Snap</td>
<td>18646</td>
<td>98.970</td>
<td>82.907</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>20486</td>
<td>99.981</td>
<td>91.925</td>
</tr>
<tr>
<td>BWA</td>
<td>21181</td>
<td>99.971</td>
<td>95.246</td>
</tr>
<tr>
<td>Stampy</td>
<td>21824</td>
<td>99.936</td>
<td>97.929</td>
</tr>
<tr>
<td>Novoalign</td>
<td>17978</td>
<td>99.967</td>
<td>94.397</td>
</tr>
</tbody>
</table>

Fast and Accurate Read Alignment for Resequencing, Mu et al, (2012)
### GATK SNPs Called

<table>
<thead>
<tr>
<th>Aligner</th>
<th>Called</th>
<th>%Correct</th>
<th>%Discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqAlto</td>
<td>429949</td>
<td>96.688</td>
<td>98.481</td>
</tr>
<tr>
<td>Snap</td>
<td>432023</td>
<td>96.078</td>
<td>98.332</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>412753</td>
<td>98.250</td>
<td>96.070</td>
</tr>
<tr>
<td>BWA</td>
<td>426207</td>
<td>97.466</td>
<td>98.409</td>
</tr>
<tr>
<td>Stampy</td>
<td>427137</td>
<td>97.290</td>
<td>98.446</td>
</tr>
<tr>
<td>Novoalign</td>
<td>430906</td>
<td>96.674</td>
<td>98.686</td>
</tr>
</tbody>
</table>

### GATK Indels Called

<table>
<thead>
<tr>
<th>Aligner</th>
<th>Called</th>
<th>%Correct</th>
<th>%Discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqAlto</td>
<td>22057</td>
<td>99.941</td>
<td>98.477</td>
</tr>
<tr>
<td>Snap</td>
<td>25563</td>
<td>93.319</td>
<td>90.303</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>20750</td>
<td>99.918</td>
<td>91.809</td>
</tr>
<tr>
<td>BWA</td>
<td>21228</td>
<td>99.915</td>
<td>95.174</td>
</tr>
<tr>
<td>Stampy</td>
<td>22696</td>
<td>99.277</td>
<td>98.288</td>
</tr>
<tr>
<td>Novoalign</td>
<td>20899</td>
<td>99.947</td>
<td>93.610</td>
</tr>
</tbody>
</table>
Key Points

- Best practices change extremely rapidly
  - We don’t know what the single best workflow is today

- Core variant toolset used by UAB CCTS
  - BWA for reference based alignment
  - Picard (duplicate removal)
  - GATK for SNP calling, realignment and recalibration
  - Breakdancer, Pindel for Structural Variant Detection
  - BedTools, VCFtools, IGV for interpretation
### Actual GATK Data

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>Ref</th>
<th>Alt</th>
<th>Quality</th>
<th>Genotype</th>
<th>Average Depth</th>
<th>Depth Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>802093</td>
<td>G</td>
<td>A</td>
<td>521.67</td>
<td>1/1:1:23:24:48.11:555,48,0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>802191</td>
<td>G</td>
<td>A</td>
<td>54.33</td>
<td>0/1:31,12:43:84.36:84,0,458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>802320</td>
<td>G</td>
<td>A</td>
<td>349.65</td>
<td>0/1:9,15:27:10.30:379,0,10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 3 genotypes (0/0, 0/1, 1/1)
- GQ:PL
  - Genotype Quality
- AD:DP
  - Average Depth : Depth Quality
Workflow Overview

Workflow from “Consensus Rules in Variant Detection from Next-Generation Sequencing Data”, Jia et al. (2012)
Variation Detection

- Nucleotide Polymorphisms
  - SNPs
  - MNPs

- “Structural Variants” / Rearrangements
  - Insertions/Deletions (small and large)
  - Inversions
  - Tandem Duplications
  - Translocations