An introduction to (small) variant detection

Erik Garrison
Advanced Sequencing 13
Cold Spring Harbor Laboratory
November 21, 2013
Resequencing

We've sequenced some DNA.

Maybe we have one sample, maybe many. Maybe we have a pool.

We want to understand phylogeny, function, or population.

(Luckily) we can use a reference to guide our analysis.

Functional Analysis of an Arabidopsis T-DNA “Knockout”...
doi: http://dx.doi.org/10.1104/pp.102.010843
Understanding differences

We can represent relative differences using literal, whole-genome sequences

- But, interpretation is often relative in nature
- We don't need to assemble a genome *de novo* in all cases (although this helps observe structural variation, e.g.):

Two strains of *H. pylori*, showing an inversion

Understanding differences

A simpler approach (especially for short reads) is sets of differences observed when aligning our reads back to a reference:

- 1bp substitutions (SNPs)
  - AATC -> AGTC
- insertions and deletions of sequence (indels)
  - AATC -> ATC or AATC -> AATTTC
- block substitutions, (MNPs, or complex)
  - AATC -> AGCC

These are "small" variants.
Resequencing methodology

Typically "short", \(~100\)bp reads of DNA are generated, then aligned to a reference genome.

Alignment generates candidate short variants, dependent on read length and method.

We can then read off variants from the alignments.
Alignment

Alignment: Matching one sequence to another, finding the closest-matching sequence, and describing the mapping between the two in terms of mismatches, insertions, and deletions.

Schematic of Needleman and Wunsch algorithm. Smith and Waterman added the 0.  [http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/2006_7/ABECASIS/abecasis.html](http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/2006_7/ABECASIS/abecasis.html)
(Global) alignment

To make alignment feasible, we have to employ techniques that allow rapid matching of short sequences (reads) to large ones (references). For instance, a suffix tree:

In contemporary practice, compressed suffix arrays (via the Burrows-Wheeler transform) and high-performance hash systems are used for global localization of reads.
Short reads are noisy

In 2nd and 3rd-generation sequencing, most putative variants are typically artifacts (>99%).

We must filter putative variants to remove artifacts or our analyses will be overwhelmed by noise.
Variant detection

Many current methods use a Bayesian model which combines several sources of information:

- **Sequencing provides observation quality estimates:**
  - base quality = 1 - prob(observation|sequence)

- **Biology provides prior expectations about polymorphism:**
  - $\Theta = p(\text{site polymorphic}) \sim 1e^{-3}$ (for humans)

- **And also population structure,**
  - which matters if we call several samples together.
Bayesian (visual) intuition

We have a universe of individuals.

A = samples with a variant at some locus

B = putative observations of variant at some locus

Figures from http://oscarbonilla.com/2009/05/visualizing-bayes-theorem/
probability(A|B)

We want to estimate the probability that we have a real polymorphism "A" given "|" that we observed variants in our alignments "B".

\[
P(A|B) = \frac{|AB|}{|B|}
\]

\[
P(A|B) = \frac{P(AB)}{P(B)}
\]

\[
P(B|A) = \frac{P(AB)}{P(A)}
\]

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]
In our case it's a bit more like this...

Observations (B) provide pretty good sensitivity, but poor specificity.
Working with sequences

Variant detection is a composite process.

First, we process our sequencing data to find patterns indicative of variation.

Only then can we apply models (e.g. Bayesian) to filter, interpret, and describe variation.
Working with sequences, finding variations

Genome (FASTA)

Variation (VCF)

alignment and variant calling

Reads (FASTQ)
One position at a time

Reference

Reads

Variant observations
One position at a time

<table>
<thead>
<tr>
<th>Reference</th>
</tr>
</thead>
</table>

Haplotype information is lost.
Direct detection of haplotypes (FreeBayes)

Direct detection of haplotypes from reads resolves differentially-represented alleles (as the sequence is compared, not the alignment).

Allele detection is still alignment-based.
Why haplotypes?

- Variants cluster.
- This has functional significance.
- Observing haplotypes lets us be more certain of the local structure of the genome.
- We can improve the detection process itself by using haplotypes rather than point mutations.
Sequence variants cluster

In ~1000 individuals, \( \frac{1}{2} \) of variants are within ~22bp of another variant.

Variance to mean ratio (VMR) = 1.4.
The functional effect of variants depends on other nearby variants on the same haplotype.

Reference: AGG   GAG   CTG
           Arg   Glu   Leu

Apparent:  AGG   TAG   CTG
           Arg   Ter   ---

Actual:   AGG   TTTG  CTG
           Arg   Leu   Leu

*OTOF* gene – mutations cause profound recessive deafness.

Apparent nonsense variant, one YRI homozygote.

Actually a block substitution that results in a missense substitution.

(Daniel MacArthur)
Importance of haplotype effects: frame-restoring indels

- Two apparent frameshift deletions in the \textit{CASP8AP2} gene (one 17 bp, one 1 bp) on the same haplotype
- Overall effect is in-frame deletion of six amino acids

(Daniel MacArthur)
Impact on genotyping chip design

- Biallelic SNPs detected during the 1000 Genomes Pilot project were used to design a genotyping microarray (Omni 2.5).

- When the 1000 Genomes samples were genotyped using the chip, 100k of the 2.5 million loci showed no polymorphism (monomorphs).
Measuring haplotypes improves specificity

Indels from AFR191 sample set, 1000G phase2 testing.
2bp MNPs and dinucleotide intermediates

<table>
<thead>
<tr>
<th></th>
<th>TT</th>
<th>TG</th>
<th>TC</th>
<th>TA</th>
<th>GT</th>
<th>GG</th>
<th>GC</th>
<th>GA</th>
<th>CT</th>
<th>AG</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alternate and reference states are shown in a grid format.
A distinct mutational mechanism is responsible for the most frequent 2bp MNP

Deamination of methyl-C to T

CA/TG

CA
GT
CG
GC
TG
AC

A→G transition to CpG intermediate

TG/CA

TG
AC
CG
GC
CA
GT

Same process on opposite strand
Comparative results

These improvements are not only theoretical.

They yield a high-quality detector!

(simulation of 100 10x samples.)
Variant detector lineage

**PolyBayes**— original Bayesian variant detector (Gabor Marth, 1999); written in perl

**GigaBayes**— ported to C++

**BamBayes**— “modern” formats (BAM)

**FreeBayes**— 2010-present
FreeBayes-specific developments

FreeBayes models (~in order of introduction):

- Multiple alleles
- Indels, SNPs, MNPs, complex alleles
- Local copy number variation (e.g. sex chromosomes)
- Global copy-number variation (e.g. species-level, genome ploidy)
- Pooled detection, both discrete and continuous
- Many, many samples (>30k exome-depth samples)
- Genotyping using known alleles (hints, haplotypes, or alleles)
- Genotyping using a reference panel of genotype likelihoods
- Direct detection of haplotypes from short-read sequencing
- Haplotype-based consensus generation (clumping)
- Allele-length-specific mapping bias
- Contamination-aware genotype likelihoods
Design/development methodology

- Portable C++
- MIT license
- Modular codebase (multiple git submodules)
  - BAM library (bamtools)
  - VCF library (vcflib)
  - FASTA library (fastahack)
- Streams > rewriting files
- A priori models > technology-specific recalibration
- Probabilities > hard filters
- Simulations > ts/tv, bulk metrics, Omni concordance
- Annotations for classification and filtering of variants (vcfsom)
- Produce internally-consistent VCF output
- Run fast
The model

- Bayesian model estimates the probability of polymorphism at a locus given input data and the population mutation rate (∼pairwise heterozygosity) and assumption of “neutrality” (random mating).
- Following Bayes theorem, the probability of a specific set of genotypes over some number of samples is:
  - \( P(G|R) = \frac{P(R|G) P(G)}{P(R)} \)
- Which in FreeBayes we extend to:
  - \( P(G,S|R) = \frac{P(R|G,S) P(G) P(S)}{P(R)} \)
- \( G = \) genotypes, \( R = \) reads, \( S = \) locus is well-characterized/mapped
- \( P(R|G,S) \) is our data likelihood, \( P(G) \) is our prior estimate of the genotypes, \( P(S) \) is our prior estimate of the mappability of the locus, \( P(R) \) is a normalizer.
The process

● Parse alleles (small haplotypes) from alignments using CIGAR strings.
● Pick suitable alleles (very weak input filters to improve runtime).
● Build haplotypes across target locus.
● Generate genotype likelihoods.
● Sample a posterior space around the data-likelihood maximum
  ○ update genotype estimates and iterate (hill-climbing posterior search) until convergence on maximum a posteriori genotyping over all samples
● Output a record, and do it again
Working with data

We’ll use NA12878 to experiment with variant calling.

tutorial @ http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html
Results on 5x NA12878 vs. Broad truth set.