Small variant detection with Galaxy

Erik Garrison

GCC 2012, July 25, 2012

Sequencing-based variant detection

- Assembly
  - expensive, both in sequencing costs and compute
  - (but potentially unbiased)
  - difficulty dealing with errors
- Reference-guided assembly (alignment)
  - use population information to improve guided assembly of individual genomes
  - more biased, but more tractable
    - difficulty with complex patterns of variation, e.g. clusters of variants and larger indels
  - use of mutual information in detection (population calling)

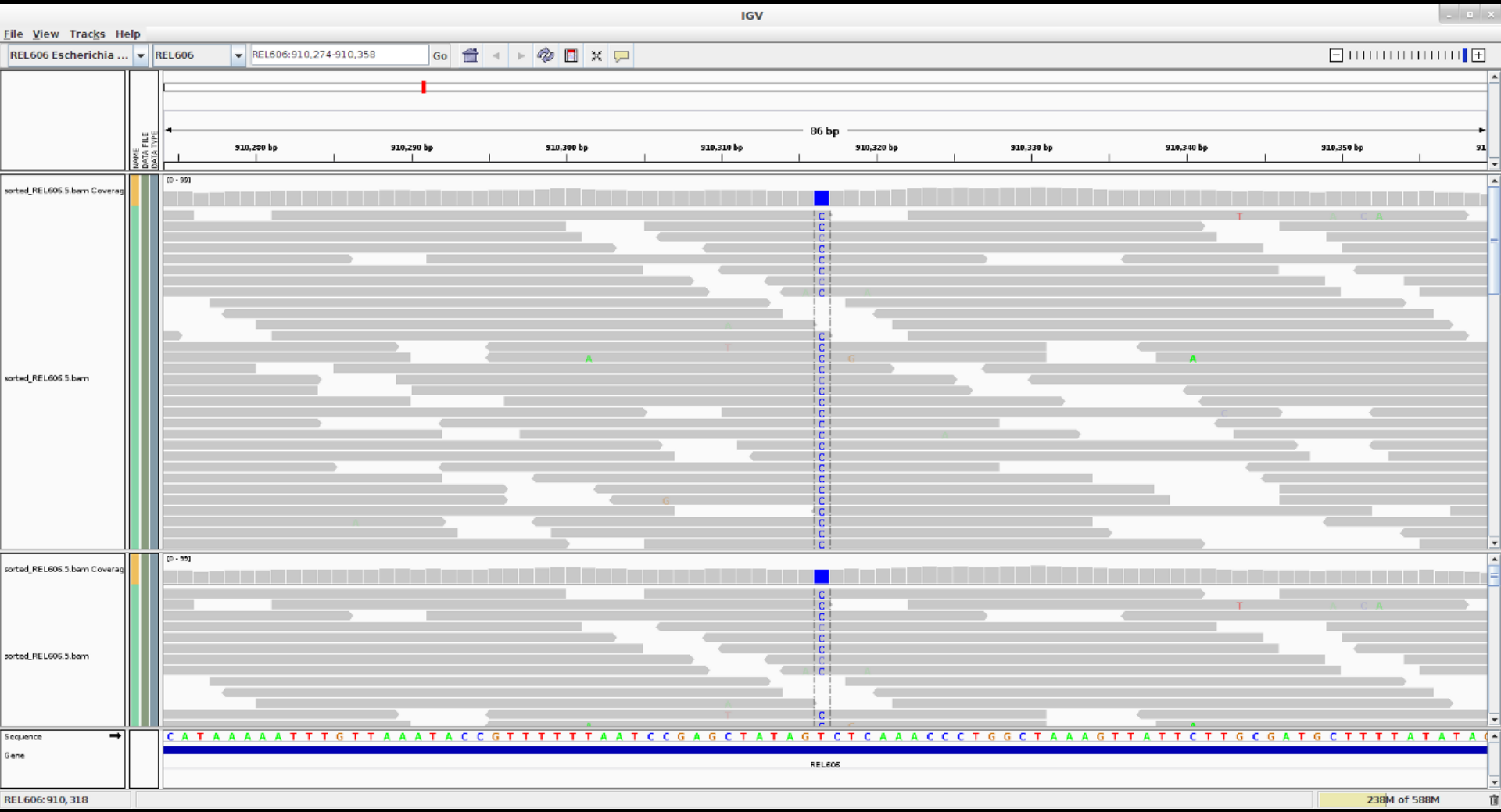
NGS variant detection

- short reads
  - 100bp is the upper limit
- high error rates
  - 1-2% at best
- error correction is required
  - rates of variation are far below the error rate (10^-3 versus 10^-2)

Variant detection pipeline

- Sequence
- Align (e.g. bwa, mosaik, bfast)
  - General approach is to index the genome, then find matches in the reference to pieces of the reads
  - Pairwise alignment allows the description of
- Process alignments
  - Recalibrate quality
  - Homogenize indels
  - Assemble locally?
- Filter variants
  - Simplistic (e.g. counting)
  - complex (e.g. Bayesian, using population information)

Alignments...


IGV
File View Tracks Help
REL606 Escherichia ...
REL606
REL606:910,274-910,358
Go
86 bp
910,280 bp
910,290 bp
910,300 bp
910,310 bp
910,320 bp
910,330 bp
910,340 bp
910,350 bp
sorted_REL606.5.bam Coverag
sorted_REL606.5.bam
sorted_REL606.5.bam Coverag
sorted_REL606.5.bam
Sequence
Gene
REL606
REL606:910,318
238M of 588M

Bayesian approach

Idea: Use a population model to combine information from all available samples to improve detection power.

Effects:

- increased sensitivity
  - low-frequency events may be hard to find in only one sample
- improved discrimination against sequencing artifacts
  - artifacts don't abide by population genetic principles

What do we do?

Assuming the alignment and recalibration is done...

We take alignments and run a variant detector.

What do we get?

A typical variant detector generates a variant report with some description of confidence.

We'll be using freebayes:

- https://github.com/ekg/freebayes
- http://arxiv.org/abs/1207.3907v2

It outputs a QUAL score which is the probability of polymorphism under its model, provided all available sequencing information.

Variant Call Format output

How do we know how to use this?

We can look at biological markers (e.g. transitions vs. transversions).

We can look at recall rates on known markers.

We can validate experimentally, then repeat.

We can use simulations to get some sense of the tradeoffs of various alignment, variant calling, and filtering methods. (Today...)

Simulated data

simulate one sample, or a set of samples
https://github.com/ekg/mutatrix

simulate reads, dwgsim (available via toolshed) or wgsim

align using MOSAIK (although, you don't have to...)

Simulated data

http://clavius.bc.edu/~erik/galaxy_workshop_2012/
reference: chr20_bit.fa (a piece of human chr20)

Subdirectories haploid, diploid, and tetraploid contain simulated individuals (1_sample) and samples of populations (10_samples).

File lists:

http://clavius.bc.edu/~erik/galaxy_workshop_2012/bam_urls/
http://clavius.bc.edu/~erik/galaxy_workshop_2012/fq_urls/
http://clavius.bc.edu/~erik/galaxy_workshop_2012/fa_urls/