

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)

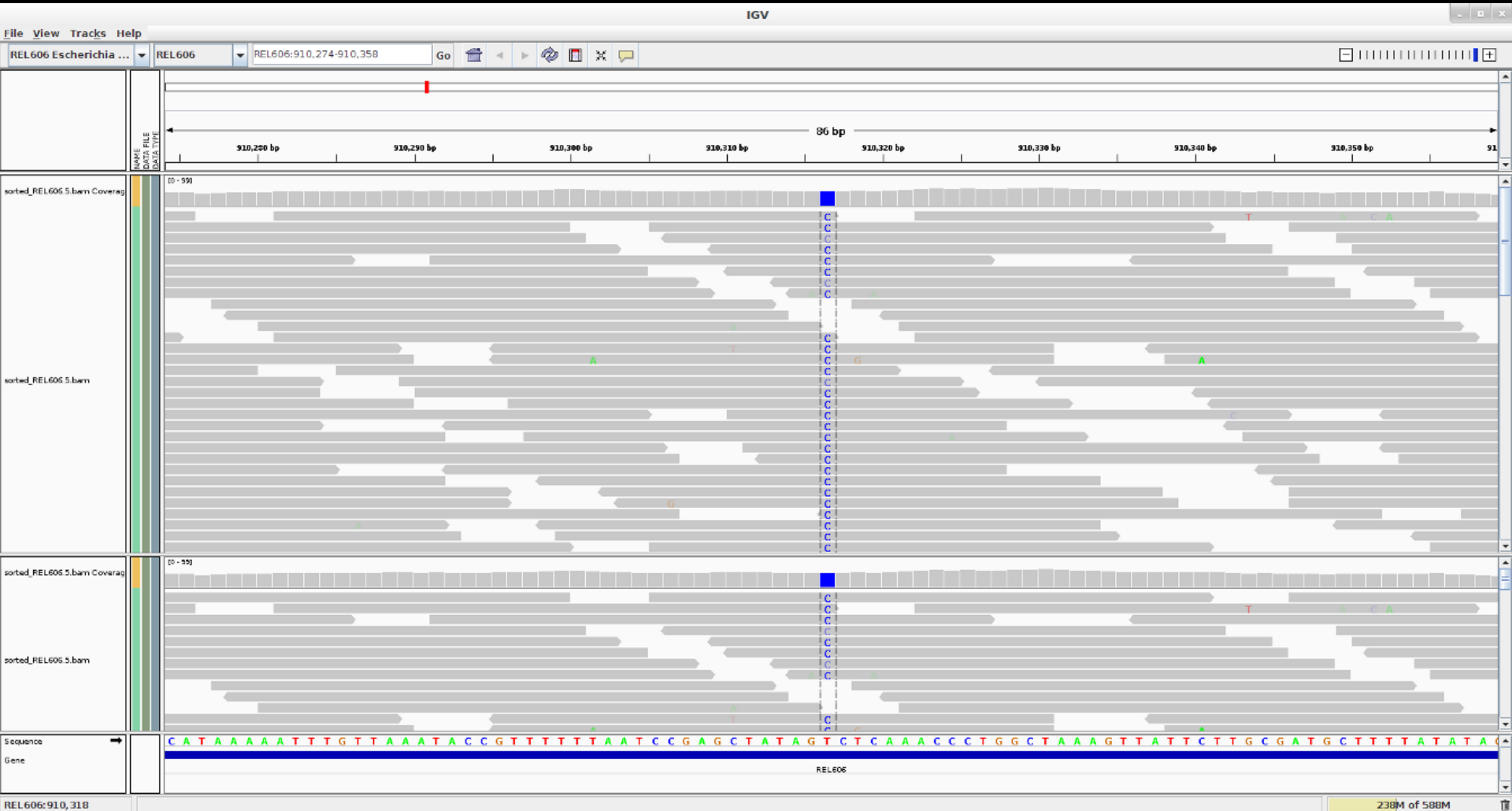
NGS variant detection

- 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...



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

```
20 138178 . GC G 49314.7 . AB=0.459264;ABP=1412.47;AC=1334;AF=0.427564;AN=3120;AO=87369;CIGAR=1M1
20 138194 . C T 3448.93 . AB=0.397101;ABP=34.739;AC=4;AF=0.00128205;AN=3120;AO=441;CIGAR=1X;DP=2
20 138200 . G A 1814.05 . AB=0.306122;ABP=83.0003;AC=1;AF=0.000320513;AN=3120;AO=200;CIGAR=1X;DP=
20 138297 . G T 876.352 . AB=0.493506;ABP=3.0385;AC=1;AF=0.000320513;AN=3120;AO=266;CIGAR=1X;DP=
20 138370 . AT A 2385.07 . AB=0.48731;ABP=3.28587;AC=7;AF=0.00224647;AN=3116;AO=103;CIGAR=1M1D;DP=
20 138460 . A G 49314.7 . AB=0.495512;ABP=4.04327;AC=878;AF=0.306778;AN=2862;AO=6126;CIGAR=1X;DP=
20 138555 . T C 28542.3 . AB=0.457986;ABP=23.2637;AC=822;AF=0.352185;AN=2334;AO=1425;CIGAR=1X;DP=
20 138563 . AATAT AAT,A 50000 . AB=0.563883,0.463453;ABP=73.4805,17.9303;AC=932,586;AF=0.446788,0.2809
20 138599 . TGTGT TGCGC,CGCGC 3498 . AB=0.512461,0.497354;ABP=3.87618,3.02179;AC=935,243;AF=0.50160
20 138606 . A G 4738.5 . AB=0.468484;ABP=8.0746;AC=481;AF=0.252361;AN=1906;AO=743;CIGAR=1X;DP=2
20 138615 . G A 10894.8 . AB=0.451664;ABP=6.82526;AC=527;AF=0.266971;AN=1974;AO=699;CIGAR=1X;DP=
20 138620 . A C 11168.7 . AB=0.48111;ABP=3.86297;AC=531;AF=0.266834;AN=1990;AO=700;CIGAR=1X;DP=
20 139154 . A G 30605.2 . AB=0.47622;ABP=9.41148;AC=752;AF=0.303716;AN=2476;AO=1609;CIGAR=1X;DP=
20 139173 . T C 49346.1 . AB=0.470981;ABP=17.7515;AC=914;AF=0.343609;AN=2660;AO=2136;CIGAR=1X;DP=
20 139268 . C T 155.157 . AB=0.571429;ABP=8.6377;AC=1;AF=0.000324675;AN=3080;AO=21;CIGAR=1X;DP=
20 139269 . G A 118.296 . AB=0.368421;ABP=5.8655;AC=1;AF=0.000324465;AN=3082;AO=23;CIGAR=1X;DP=2
20 139312 . C A 200.62 . AB=0.178571;ABP=1158.85;AC=58;AF=0.0186017;AN=3118;AO=1146;CIGAR=1X;DP=
20 139358 . CA C,TA 1758.54 . AB=0.135576,0.364865;ABP=1134.62,26.4857;AC=41,4;AF=0.013141,0.0012820
20 139361 . C T 808.626 . AB=0.474359;ABP=3.45573;AC=2;AF=0.000641026;AN=3120;AO=52;CIGAR=1X;DP=
20 139362 . G A 49314.7 . AB=0.492805;ABP=15.5089;AC=1767;AF=0.566346;AN=3120;AO=33735;CIGAR=1X;
20 139363 . TG T 570.068 . AB=0.175725;ABP=1515.54;AC=77;AF=0.0246795;AN=3120;AO=539;CIGAR=1M1D;D
20 139405 . CTG C 0.442479 . AB=0.145833;ABP=55.3066;AC=2;AF=0.000641026;AN=3120;AO=250;CIG
20 139409 . G A 49314.7 . AB=0.438562;ABP=1115.92;AC=1028;AF=0.329487;AN=3120;AO=29599;CIGAR=1X;
20 139419 . C T 50000 . AB=0.471803;ABP=32.7674;AC=66;AF=0.0211538;AN=3120;AO=2148;CIGAR=1X;DP=
20 139455 . AG AA,A 49314.7 . AB=0.438793,0.0182558;ABP=1594.73,50023;AC=1086,1;AF=0.348077,0.000320
20 139504 . T C 331.268 . AB=0.486486;ABP=3.06899;AC=1;AF=0.000320513;AN=3120;AO=105;CIGAR=1X;DP=
20 139510 . A C 0.0130107 . AB=0.222222;ABP=21.1059;AC=1;AF=0.000320513;AN=3120;AO=1271;CI
20 139521 . C G 465.989 . AB=0.425532;ABP=5.27418;AC=1;AF=0.000320513;AN=3120;AO=84;CIGAR=1X;DP=
20 139538 . A G 3938 . AB=0.459658;ABP=8.79204;AC=5;AF=0.00160256;AN=3120;AO=338;CIGAR=1X;DP=
20 139540 . G A 3621.2 . AB=0.520161;ABP=3.88589;AC=2;AF=0.000641026;AN=3120;AO=327;CIGAR=1X;DP=
20 139576 . C A 49314.7 . AB=0.423178;ABP=3434.1;AC=1071;AF=0.343269;AN=3120;AO=54843;CIGAR=1X;D
20 139681 . G A 708.799 . AB=0.456897;ABP=4.88226;AC=2;AF=0.000641026;AN=3120;AO=124;CIGAR=1X;DP=
20 139701 . C A 820.398 . AB=0.472222;ABP=3.49285;AC=1;AF=0.000320513;AN=3120;AO=68;CIGAR=1X;DP=
20 139741 . CCTTT CCTTC,TCTTC 49314.7 . AB=0.491217,0.444444;ABP=5.07016,15.074;AC=2643,14;AF=0.847659
20 139773 . C T 2758.73 . AB=0.471311;ABP=4.7546;AC=9;AF=0.00289203;AN=3112;AO=136;CIGAR=1X;DP=
```



QUAL

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/