Integrated variant detection

Erik Garrison, Boston College

Overview

- Single-sample variant detection
- Population-based variant detection
- Our implementation (freeBayes)
- Challenges for population-based variant detection.

What varies?

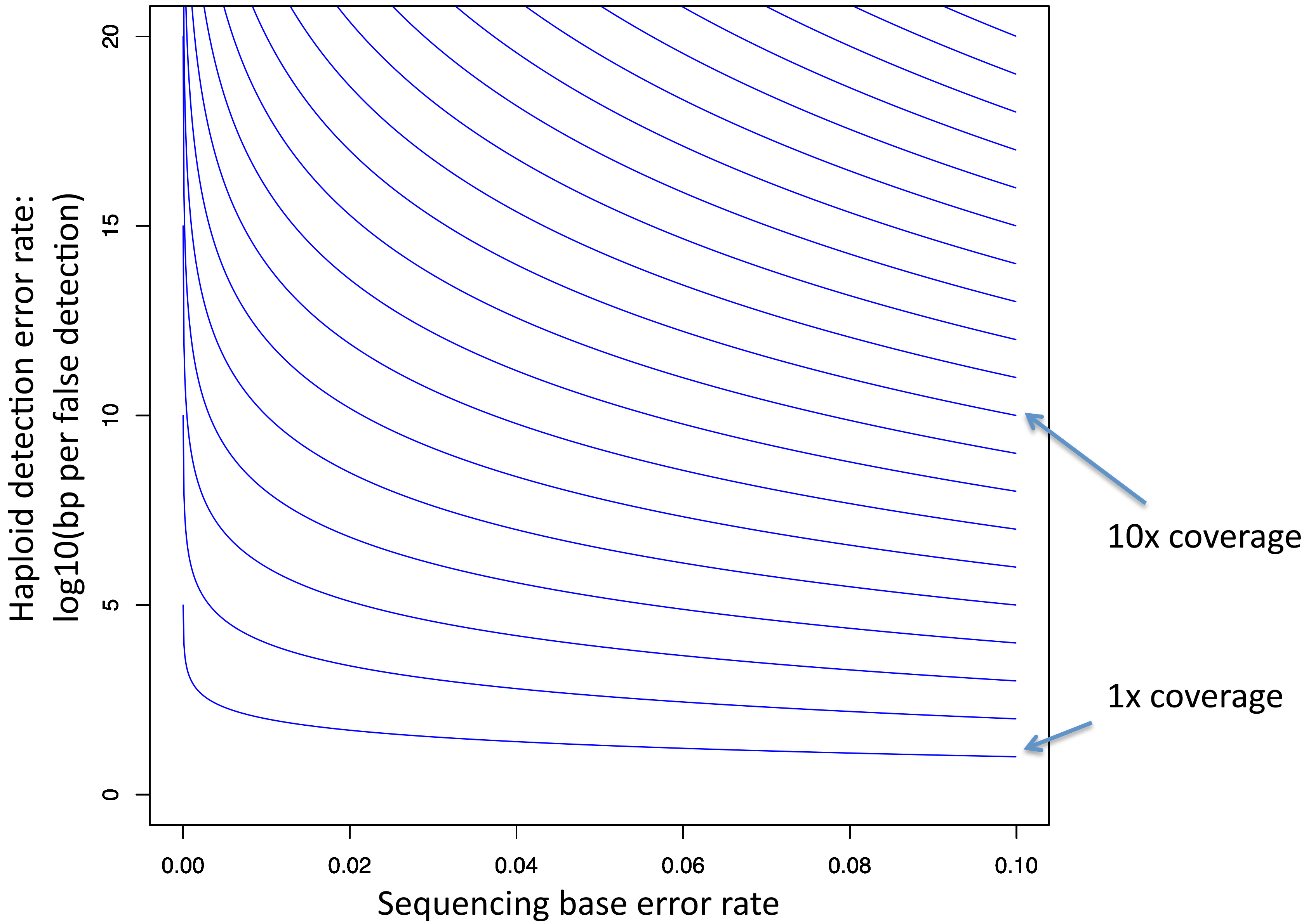
- Given short read data from some individual, how do we determine what true polymorphisms they have relative to another?

- A few approaches come to mind:
  - Count alternate (non-reference) alleles
  - Use a binomial test
  - Integrate quality scores from reads

Maximum likelihood variant detection

- Short reads are noisy
- Alignments are noisy
- Even with a relatively low base error rate for short read sequence data, we need coverage to ensure that we have sufficient power.

~Error rate versus coverage, 1-20x



Maximum likelihood variant detection

- Looking at one sample is informative, but limited by per-sample coverage.
- Using a single-sample model is difficult because we lack power to filter out artifacts which result from errors within our sequencing and alignment system:
  - paralogs
  - spurious mismatch agreement
  - systematic misalignment

Population calling

- 1 sample is noisy
- Your study may obtain data from many. Why not use them together to improve the power of your variant detection?

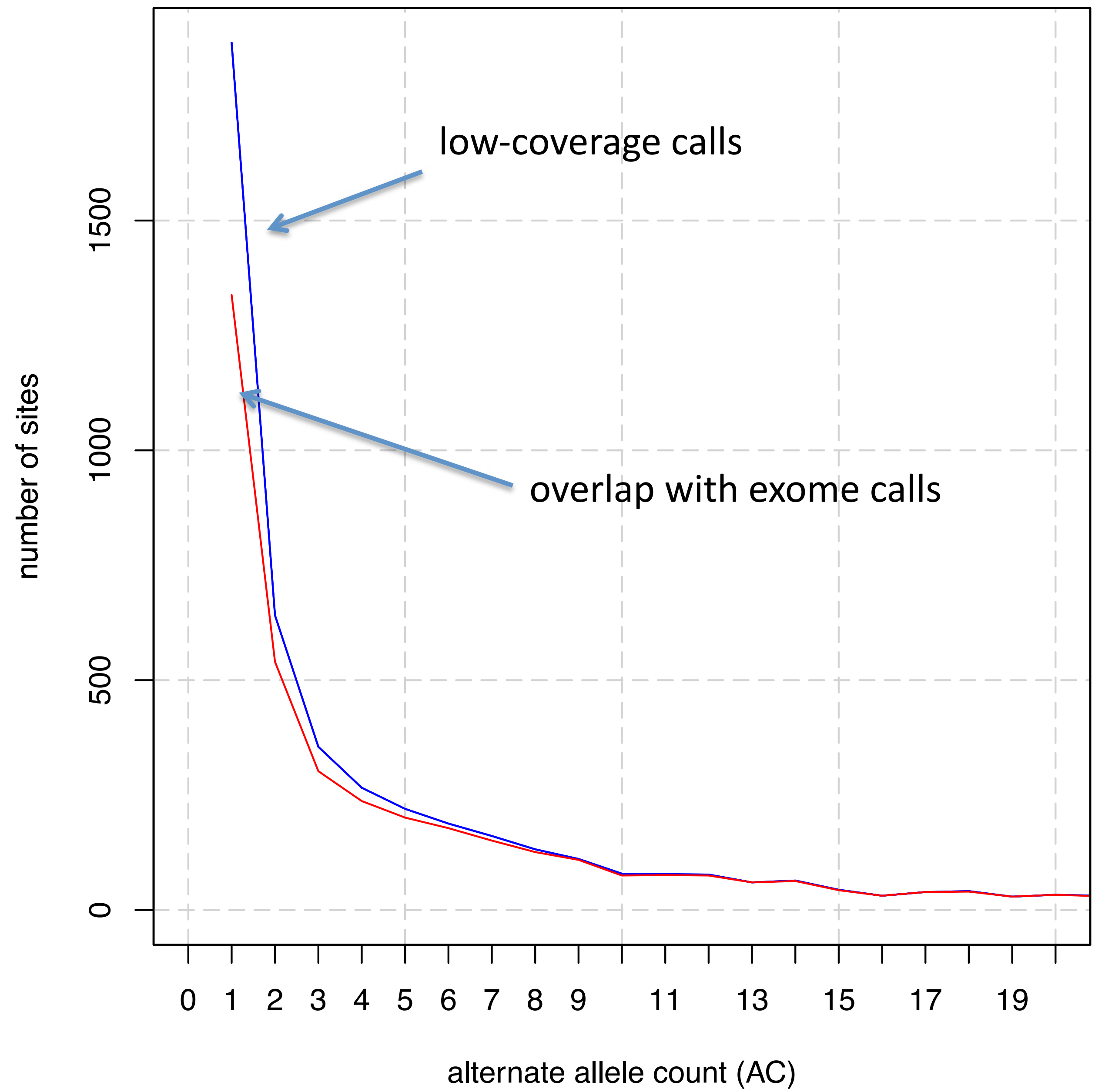
Bayesian population-level variant detectors and genotypers

- freeBayes
  - Marth Lab, Boston College: http://bioinformatics.bc.edu/marthlab/
- GATK
  - http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper
- glfMultiples
  - http://genome.sph.umich.edu/wiki/GlfMultiples
- others …

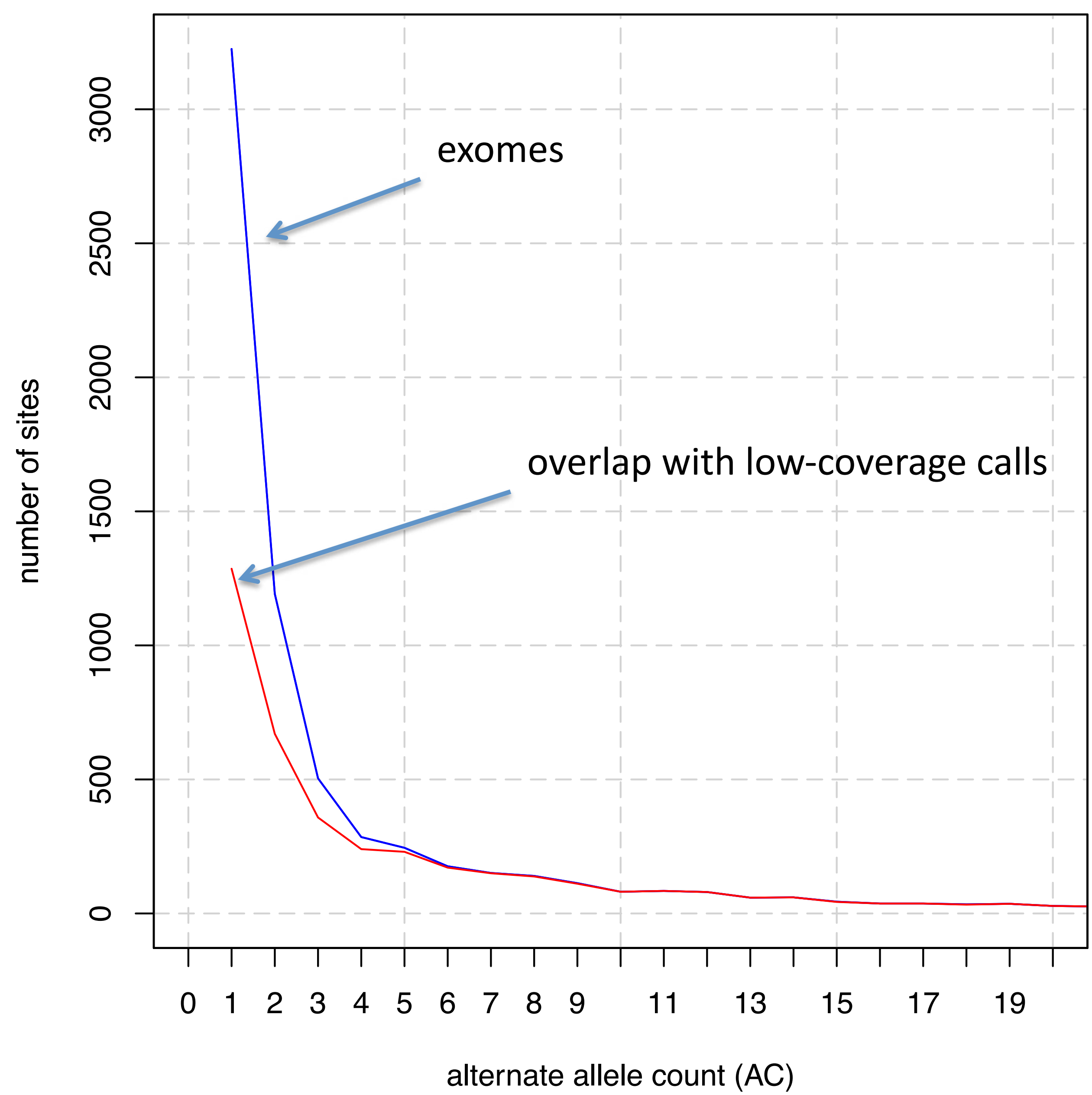
How does a population model cope with errors?

- Directly, via incorporation of information from multiple samples.
- It’s much less likely to miss or miss-call variants with even low frequency in the population.
  - In the 1000 Genomes project, we see error rates (both FP and FN) drop very low at alternate allele count >10, ~ 1% allele frequency.

Sites found in the 1000G working low-coverage consensus against sites in 688 exomes



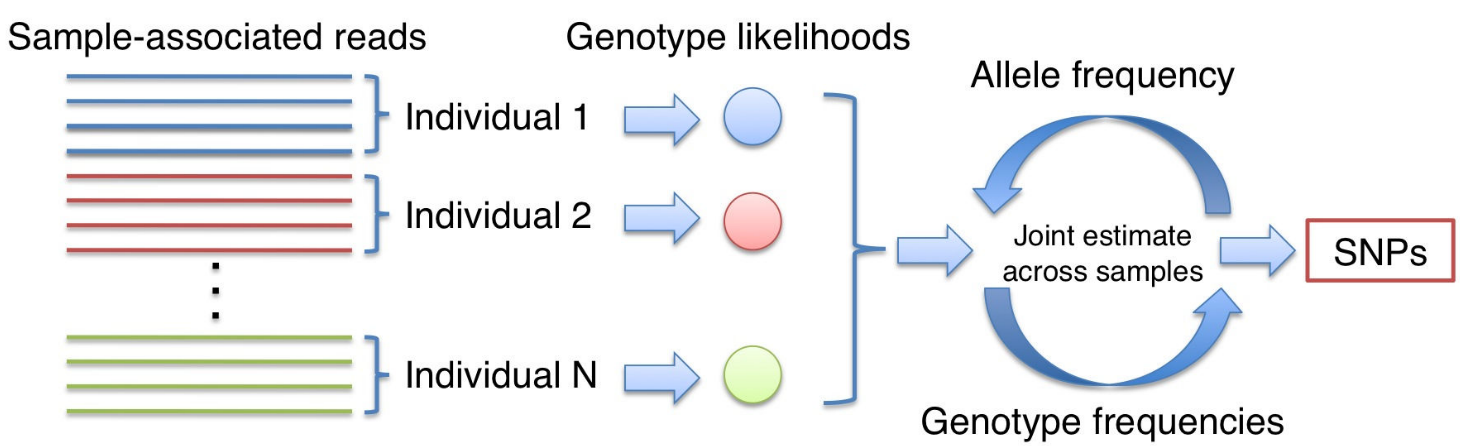
Sites found in 688 exomes against sites in the 1000G low-coverage consensus



Bayesian detection, multiple samples

- We can improve power by collecting our samples together in a Bayesian framework.
- Because population-based variation looks very different than sequencing error and alignment artifact, we can compare what we observe against prior expectations about the way that alleles are distributed in a population.
- The natural way to do this is in a Bayesian setting.

A population of samples



GATK documentation

http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper

A population of samples

Genotyping across samples

Prior probability of the genotyping

$$P(G_1,\ldots,G_n|B_1,\ldots,B_M) = \frac{P(G_1,\ldots,G_n)P(B_1,\ldots,B_M|G_1,\ldots,G_n)}{P(B_1,\ldots,B_M)}$$

$$P(G_1,\ldots,G_n|B_1,\ldots,B_M) = \frac{P(G_1,\ldots,G_n)\prod_{i=1}^{n} P(B_i|G_i)}{\sum_{\forall G_1,\ldots,G_n}(P(G_1,\ldots,G_n)\prod_{i=1}^{n} P(B_i|G_i))}$$

Sequencing observations from the entire population

Data likelihoods

G_i = genotype for a sample, B_i = observations for a sample

The Neutral Model

- Most variation in populations is relatively neutral with reflect to base context.
- Assuming neutrality, we can build some simple mathematical descriptions of the probability of observing a given set of alleles and genotypes at a given locus.
- We can use this model to integrate data likelihood estimates from a population of samples.

Genotype sampling probability

- ~ Hardy-Weinberg Equilibrium (as used in other callers).
- Genotypings like this: AB, AB, AB, AB, AB, AB have much lower probability than AA, AA, AB, BB, AA, AB, AA.
- (Technical: discrete scaling allows us to use numerical integration methods....)

Allele frequency prior probability: Ewens' sampling formula

- Provides the probability of a given set of allele frequencies at a locus given an expected diversity rate (we use estimated pairwise diversity ~0.001).
- Seamlessly incorporates multiple alleles.

$$P(f_1, \ldots, f_k) = P(a_1, \ldots, a_n) = \frac{M!}{\theta \prod_{z=1}^{M-1} (\theta + z)} \prod_{j=1}^{M} \frac{\theta^{a_j}}{j^{a_j} a_j!}$$

The probability of a given set of allele frequencies…

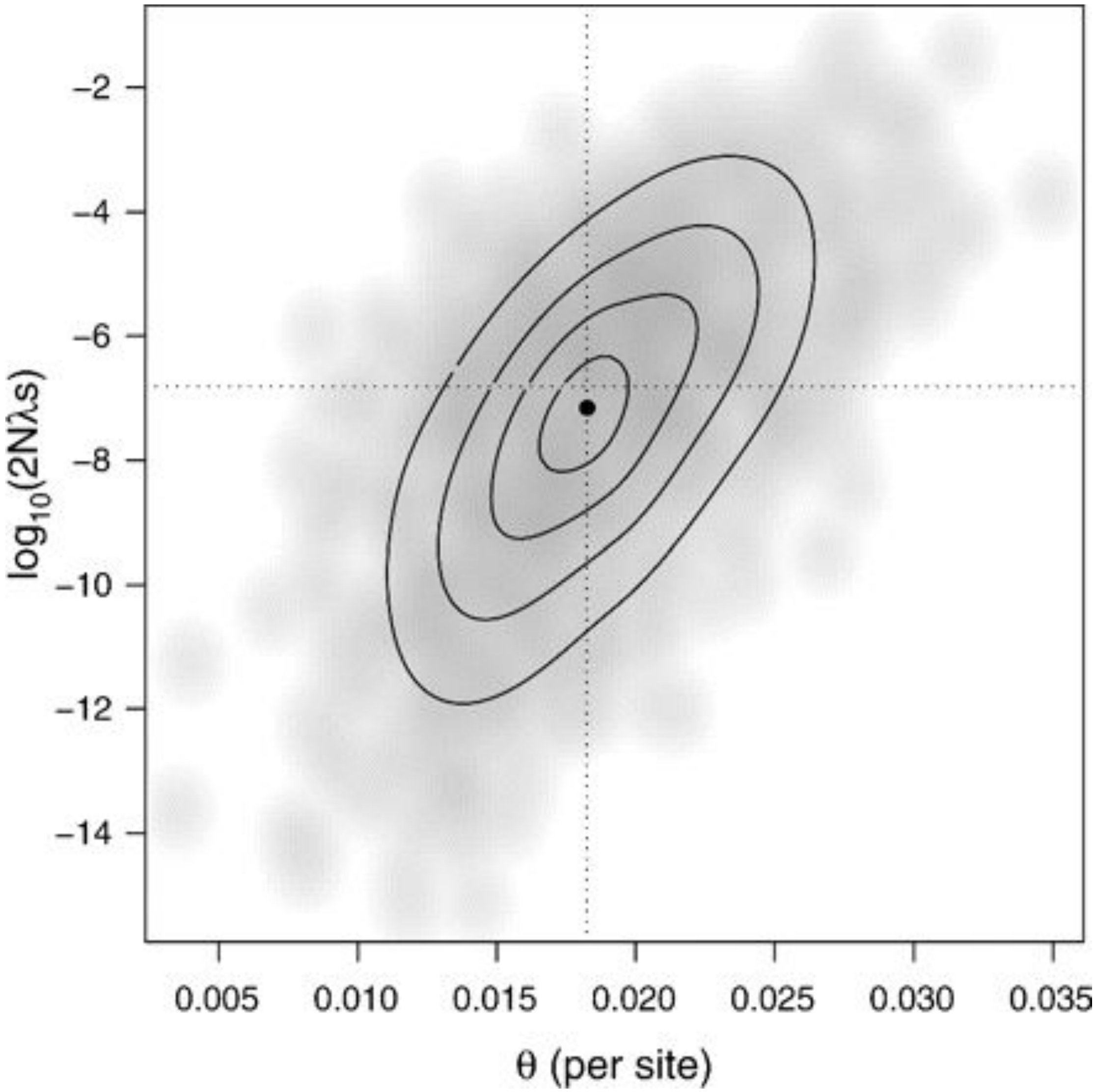
… can be expressed as allele frequency counts …

… over which the Ewens' sampling formula is defined, given some theta.

Technical issues

- Posterior integration… $\sum_{\forall G_1,\ldots,G_n}(P(G_1,\ldots,G_n)\prod_{i=1}^{n}P(B_i|G_i))$
  - freeBayes uses a greedy method centered on the data likelihood maximum (OK in most cases due to extreme "spikiness" of the distribution)
- Maximum a posteriori estimation
  - Convergent, greedy method: local search followed by gradient ascent.
  - Provides a decent balance of speed and sensitivity relative to MCMC approaches often used.
  - Deterministic

Finding the posterior maximum



(From: Ingvarsson, "Natural Selection on Synonymous and Nonsynonymous Mutations Shapes Patterns of Polymorphism in Populus tremula.")

All together now…

SNPs, INDELs, and MNPs

- Abstract representation of alleles allows freeBayes to simultaneously call all these classes.
- Piping BAM input allows for base quality recalibration methods, INDEL realignment, gap opening realignment, and other approaches on the fly, without rewriting BAM files.
- Call all small variants in one pass over the data.
- Full support for poly-allelic sites (>2 present alleles).

Polyploidy, variable copy number, and pooled sequencing analysis

- We use a fully generalizeable mathematical model, allowing for per-sample, per-region specification of ploidy.
- Pooled sequencing is a special case of variable ploidy, and is enabled via a flag to freeBayes and the specification of ploidy == the number of genomic copies in the pooled sample.

Combined variant output

(VCF 4.1)

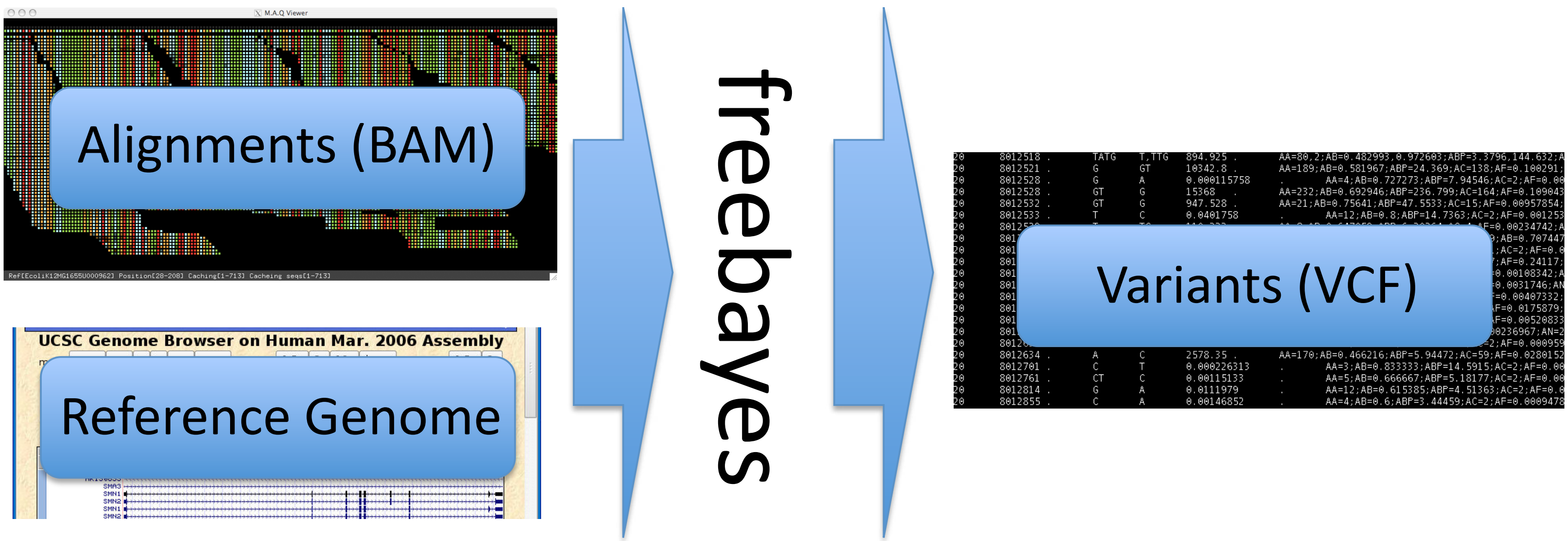
20	8012518	.	TATG	T,TTG	894.925	.	AA=80,2;AB=0.482993,0.972603;ABP=3.3796,144.632;A
20	8012521	.	G	GT	10342.8	.	AA=189;AB=0.581967;ABP=24.369;AC=138;AF=0.100291;
20	8012528	.	G	A	0.000115758	.	AA=4;AB=0.727273;ABP=7.94546;AC=2;AF=0.00
20	8012528	.	GT	G	15368	.	AA=232;AB=0.692946;ABP=236.799;AC=164;AF=0.109043
20	8012532	.	GT	G	947.528	.	AA=21;AB=0.75641;ABP=47.5533;AC=15;AF=0.00957854;
20	8012533	.	T	C	0.0401758	.	AA=12;AB=0.8;ABP=14.7363;AC=2;AF=0.001253
20	8012539	.	T	TG	110.233	.	AA=8;AB=0.647059;ABP=6.20364;AC=4;AF=0.00234742;A
20	8012540	.	GATGT	G,GTATGT,GATGTATGT,GCATGT	3446.45	.	AA=68,41,24,9;AB=0.707447
20	8012543	.	G	T	0.00155097	.	AA=10;AB=0.571429;ABP=3.32051;AC=2;AF=0.0
20	8012546	.	T	C	11845.4	.	AA=717;AB=0.501852;ABP=3.04247;AC=437;AF=0.24117;
20	8012549	.	A	G	13.9665	.	AA=8;AB=0.636364;ABP=4.78696;AC=2;AF=0.00108342;A
20	8012552	.	T	C	109.051	.	AA=27;AB=0.692308;ABP=15.538;AC=6;AF=0.0031746;AN
20	8012559	.	GT	G	315.757	.	AA=10;AB=0.756757;ABP=24.1968;AC=8;AF=0.00407332;
20	8012563	.	GC	G	1878	.	AA=41;AB=0.789773;ABP=131.374;AC=35;AF=0.0175879;
20	8012624	.	C	T	608.011	.	AA=41;AB=0.507042;ABP=3.04088;AC=11;AF=0.00520833
20	8012625	.	G	T	219.228	.	AA=27;AB=0.36;ABP=7.26639;AC=5;AF=0.00236967;AN=2
20	8012633	.	GA	G	0.000147503	.	AA=2;AB=0.75;ABP=7.35324;AC=2;AF=0.000959
20	8012634	.	A	C	2578.35	.	AA=170;AB=0.466216;ABP=5.94472;AC=59;AF=0.0280152
20	8012701	.	C	T	0.000226313	.	AA=3;AB=0.833333;ABP=14.5915;AC=2;AF=0.00
20	8012761	.	CT	C	0.00115133	.	AA=5;AB=0.666667;ABP=5.18177;AC=2;AF=0.00
20	8012814	.	G	A	0.0111979	.	AA=12;AB=0.615385;ABP=4.51363;AC=2;AF=0.0
20	8012855	.	C	A	0.00146852	.	AA=4;AB=0.6;ABP=3.44459;AC=2;AF=0.0009478

Poly-allelic INDEL

SNP

+ Sample-specific genotyping information (not shown)

Variant detection pipeline



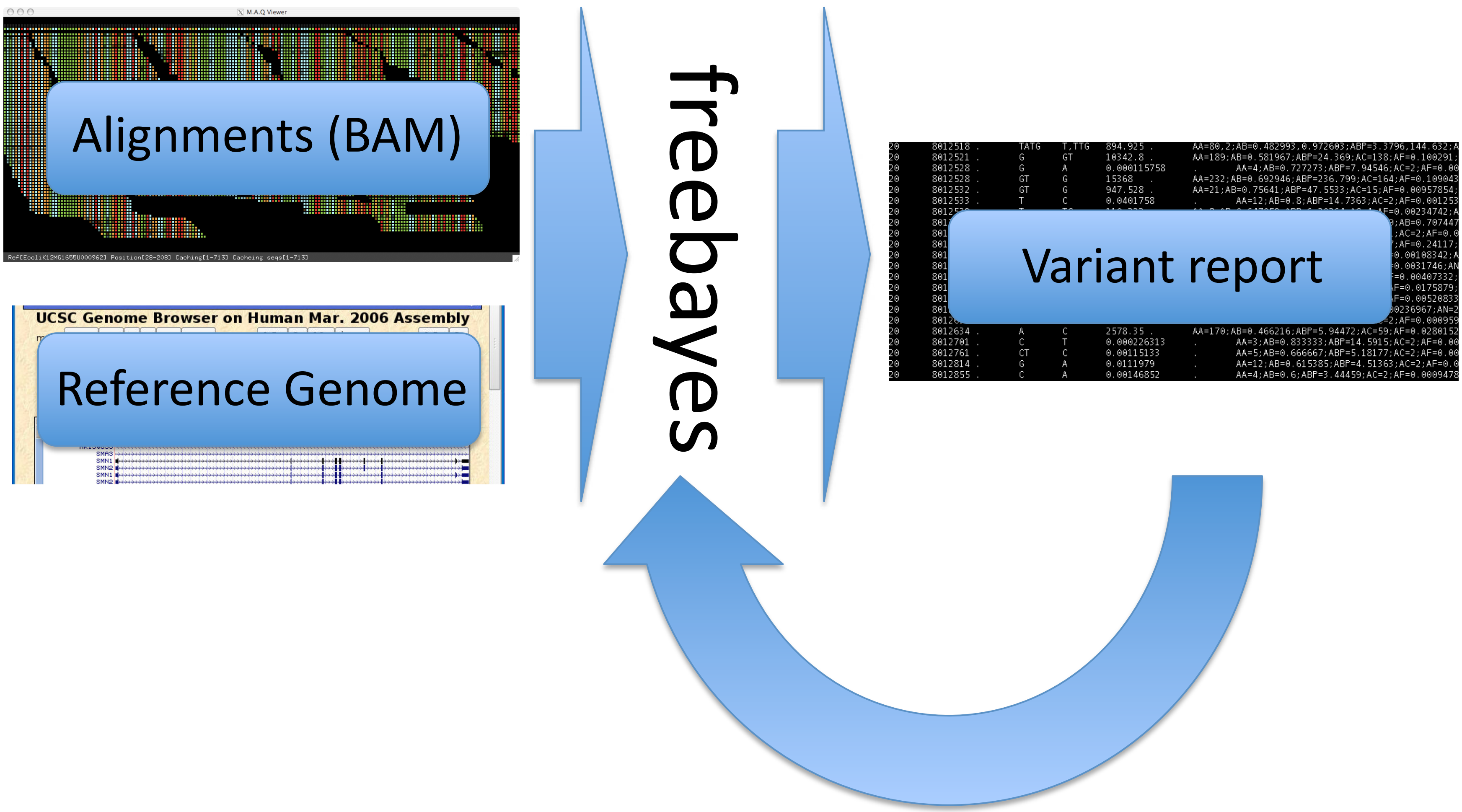
Problems with population-level variant detection

- With present callers, you'll need ALL the alignments from the samples you want to include in the population.
  - This is good if you want to use complex priors involving read positional information, allele balance across all heterozygotes,
  - But this is bad if you don't have 50TB of storage space available!

Solution: use a VCF file to describe the population allele frequencies and sites

- Read sites and allele frequencies from a VCF file, such as that produced by the 1000 Genomes project.
- Report results for input samples at all sites, conditioned on allele frequencies provided by the input.
- Implementation in freeBayes ongoing
  - tabix indexing system for VCF files (allows data parallelization via analysis targeting).

Adding prior variants



Benefits of VCF-derived variant priors

- Genotype your samples at known sites.
- Variations with low supporting information can still be called.
- No need to shuffle around dozens of terabytes of BAM alignments, or process them!
- Priors are unlikely to overwhelm true variant sites (testing is underway to balance this).

FreeBayes and Galaxy

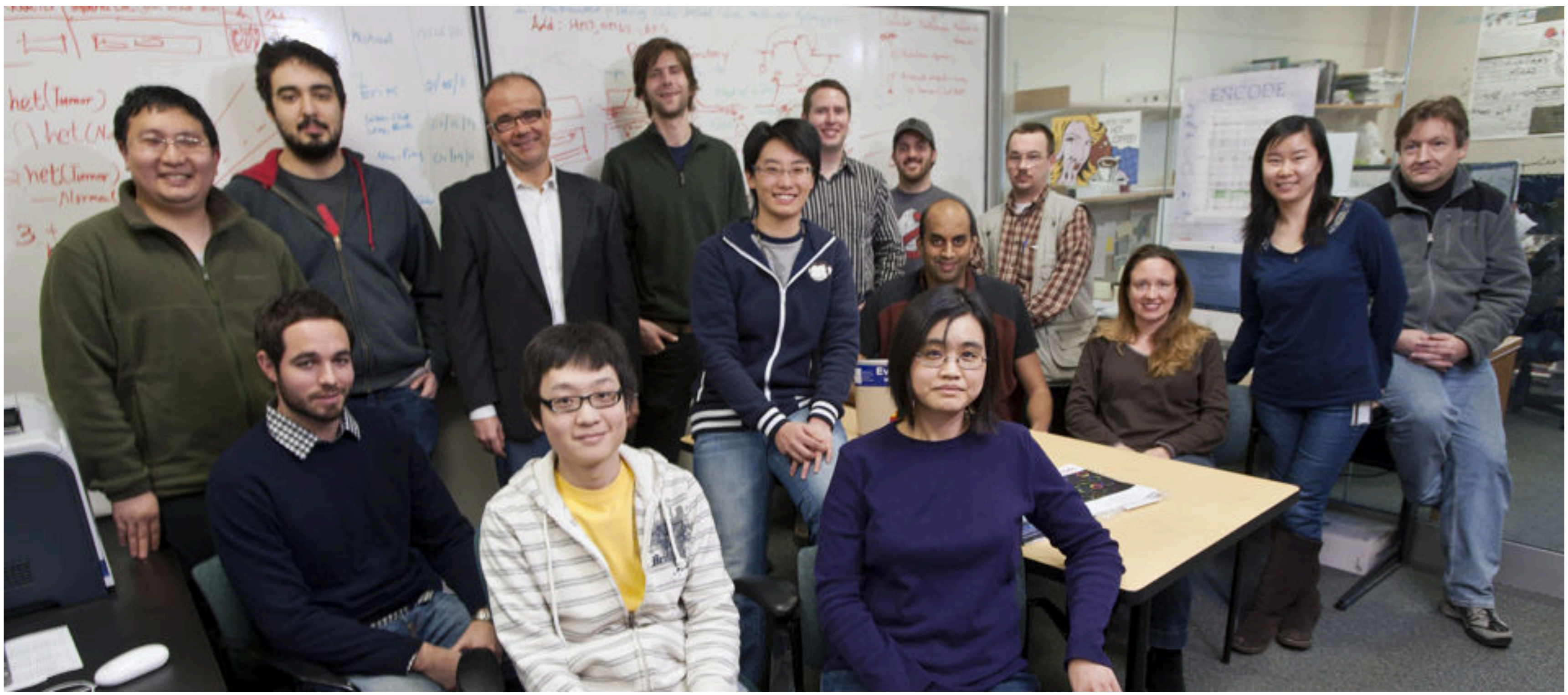
- We've done preliminary work integrating freeBayes into the Galaxy framework:
  - https://bitbucket.org/galaxy/galaxy-central/src/2f84c42a548a/tools/human_genome_variation/freebayes.xml
  - But we want to know more about how Galaxy users envision using freeBayes!

FreeBayes and Galaxy, plans

- Incorporation of data parallelization framework (aka map+reduce) for freeBayes.
- Integration of of described VCF input system into Galaxy.
- VCF filtering systems for post-processing (vcflib).

Acknowledgments

Gabor Marth, Alistair Ward, Chase Miller (galaxy integration), Amit Indap, Wen Fung Leong, & the rest of the Marth Lab



Single-sample maximum-likelihood Bayesian model, no errors

true alleles

observations

multinomial sampling probability

$$P(B_i|G_i) \approx P(B'_i|G_i) = \left\{ \frac{s_i!}{\prod_{j=1}^{k_i} o'_j!} \prod_{j=1}^{k_i} \left(\frac{f_{i_j}}{m_i} \right)^{o'_j} \right\}$$

genotype

s_i = number of observations
k_i = number of observed alleles
o'_j = number of observations of allele
f_i_j = frequency of allele in genotype
m_i = sample ploidy

Single-sample maximum-likelihood Bayesian model, incorporating errors

$$P(B_i|G_i) = \sum_{\forall(B_i \in G_i)} \left(\frac{s_i!}{\prod_{j=1}^{k_i} o'_j!} \underbrace{\prod_{j=1}^{k_i} \left(\frac{f_{i_j}}{m_i}\right)^{o'_j}}_{\text{sampling probability}} \underbrace{\prod_{l=1}^{s_i} P(b'_l|b_l)}_{\text{error estimate via base qualities}} \right)$$

s_i = number of observations
k_i = number of observed alleles
o'_j = number of observations of allele
f_i_j = frequency of allele in genotype
m_i = sample ploidy

A population of samples

Genotyping across samples

Prior probability of the genotyping

$$P(G_1,\ldots,G_n|B_1,\ldots,B_M) = \frac{P(G_1,\ldots,G_n)P(B_1,\ldots,B_M|G_1,\ldots,G_n)}{P(B_1,\ldots,B_M)}$$

$$P(G_1,\ldots,G_n|B_1,\ldots,B_M) = \frac{P(G_1,\ldots,G_n)\prod_{i=1}^{n} P(B_i|G_i)}{\sum_{\forall G_1,\ldots,G_n}\left(P(G_1,\ldots,G_n)\prod_{i=1}^{n} P(B_i|G_i)\right)}$$

Sequencing observations from the entire population

Data likelihoods

Genotype priors

The integration of allele frequency and genotype frequency information is a common theme among callers using this approach. We break our genotype prior term into its subcomponents like this:

$$P(G_1, \ldots, G_n) = P(G_1, \ldots, G_n \cap f_1, \ldots, f_k)$$

Via Bayes's rule:

$$P(G_1, \ldots, G_n \cap f_1, \ldots, f_k) = P(G_1, \ldots, G_n | f_1, \ldots, f_k) P(f_1, \ldots, f_k)$$

Now we have two prior components which are very straightforward to model.

Genotype sampling probability

The probability of sampling a given genotyping across all samples, a-priori, given a specific allele frequency distribution

(Multiset permutations of alleles in genotypes * multinomial sampling probability)

$$P(G_1, \ldots, G_n | f_1, \ldots, f_k)$$

$$= \binom{M}{f_1, \ldots, f_k}^{-1} \prod_{i=1}^{n} \binom{m_i}{f_{i_1}, \ldots, f_{i_{k_i}}}$$

$$= \frac{1}{M!} \prod_{l=1}^{k} f_l! \prod_{i=1}^{n} \frac{m_i!}{\prod_{j=1}^{k_i} f_{i_j}!}$$

Allele frequency prior probability: Ewens' sampling formula

- Provides the probability of a given set of allele frequencies at a locus given an expected diversity rate (we use estimated pairwise diversity ~0.001).
- Seamlessly incorporates multiple alleles.

$$P(f_1, \ldots, f_k) = P(a_1, \ldots, a_n) = \frac{M!}{\theta \prod_{z=1}^{M-1}(\theta + z)} \prod_{j=1}^{M} \frac{\theta^{a_j}}{j^{a_j} a_j!}$$

The probability of a given set of allele frequencies…

… can be expressed as allele frequency counts …

… over which the Ewens' sampling formula is defined, given some theta.