Variant and Genotype Calling in Polyploids

Lindsay Clark, University of Illinois, Urbana-Champaign

Excellence in Breeding meeting, CIP, 8 May 2019

See https://lvclark.github.io for copies of my presentation materials

Terminology

- **Variant calling:** Identifying SNPs and other variants (and their genomic locations, if there is a reference)
  - Is this a true SNP, a sequencing error, or a difference between paralogs?

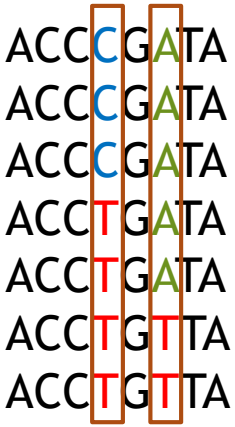
A → G

- **Genotype calling:** For all identified SNPs, determining the genotype for every individual in a population.
  - Is this a homozygote or heterozygote? What allele dosage in a heterozygote?

Sam1 AAAG
Sam2 AAAA
Sam3 AAAA
Sam4 AAGG

Issues with variant calling in polyploids

- **Allopolyploids** – Two or more subgenomes from different species, typically not recombining with each other
- **Isoloci** – paralogous loci originating from different subgenomes
- Fixed differences between subgenomes are not informative and should not be called as variants
- Ideally, every read is aligned/assigned to the correct isolocus
- For autopolyploids, the software should be aware that read depth in heterozygotes might not be a 1:1 ratio



Example:
C/T distinguishing isoloci
A/T variable in one isolocus

Variant calling software: UNEAK

- Non-reference pipeline
- Part of TASSEL3
- Keeps pairs of sequence tags that differ by one nucleotide
- Groups of more than two similar sequence tags get discarded
- This eliminates most paralogs, but many good markers as well
- Can run on a laptop
- Read depth higher than 127 not reported
- Software not updated or maintained

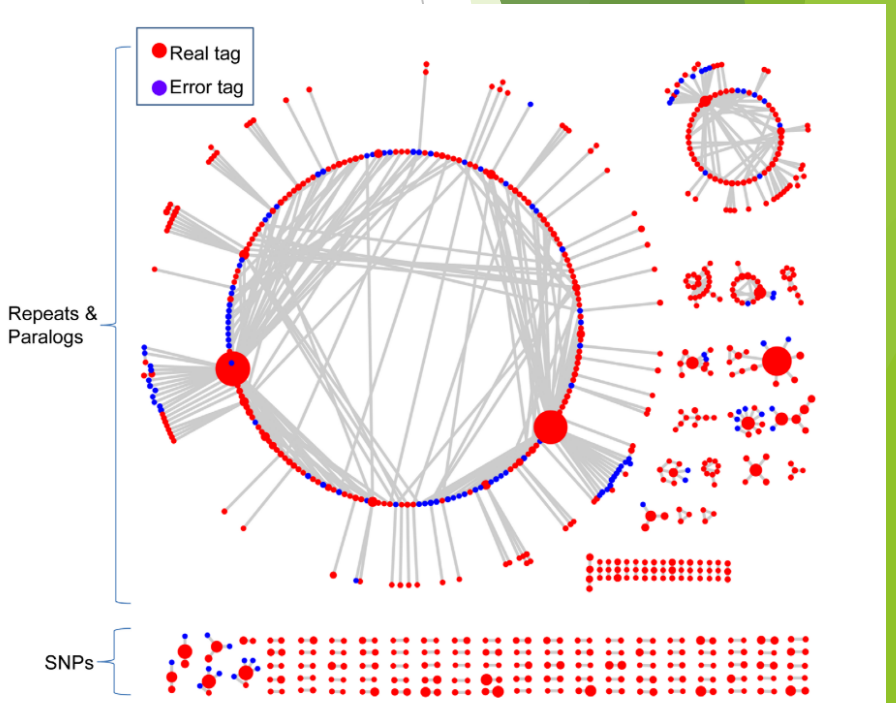


Figure 2. The networks of 802 representative tags from actual switchgrass data. The red circles are putative "real" tags. The blue circles are low frequency, putative error tags (see Methods). The size of each circle denotes the count of a tag. Lines connecting the circles ("edges") join tags that differ by a single bp mismatch. Of the 802 tags, 192 (24%) formed reciprocal tag pairs and thus, were identified as SNPs by the network filter.
doi:10.1371/journal.pgen.1003215.g002

Variant calling software: GBS-SNP-CROP

- Works with or without reference
- Set of Perl scripts utilizing existing tools such as BWA, Samtools, and Vsearch
- Without a reference, Vsearch is used to cluster reads to make a mock reference
- Ratio of read depth within individuals is used to help filter paralogs (mnAlleleRatio parameter)
- Allows use of paired-end reads

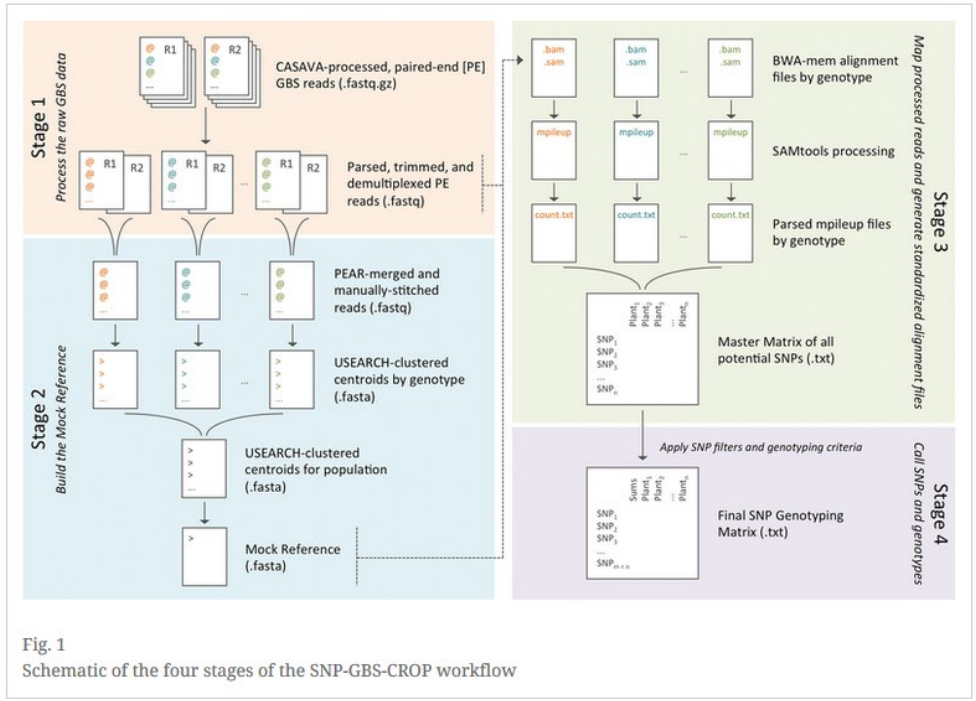
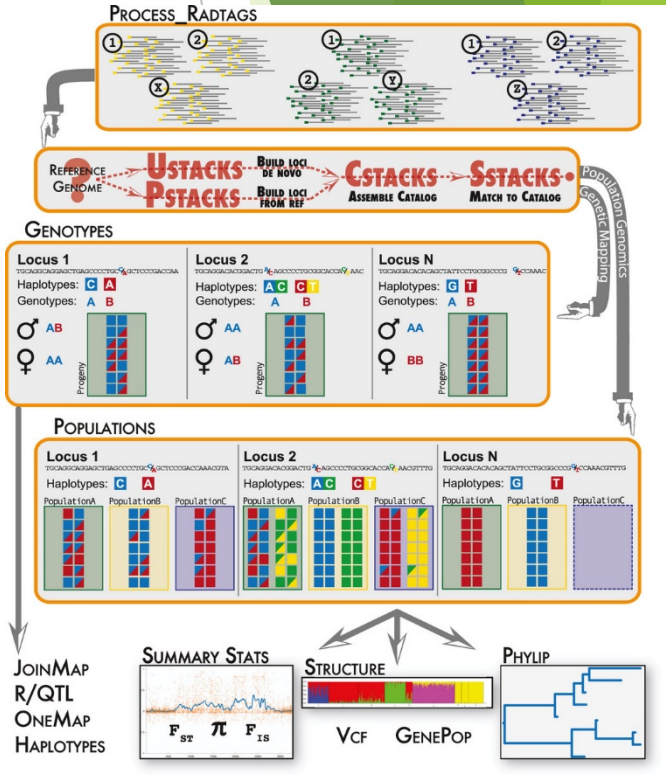


Fig. 1
Schematic of the four stages of the SNP-GBS-CROP workflow

https://doi.org/10.1186/s12859-016-0879-y

Variant calling software: Stacks

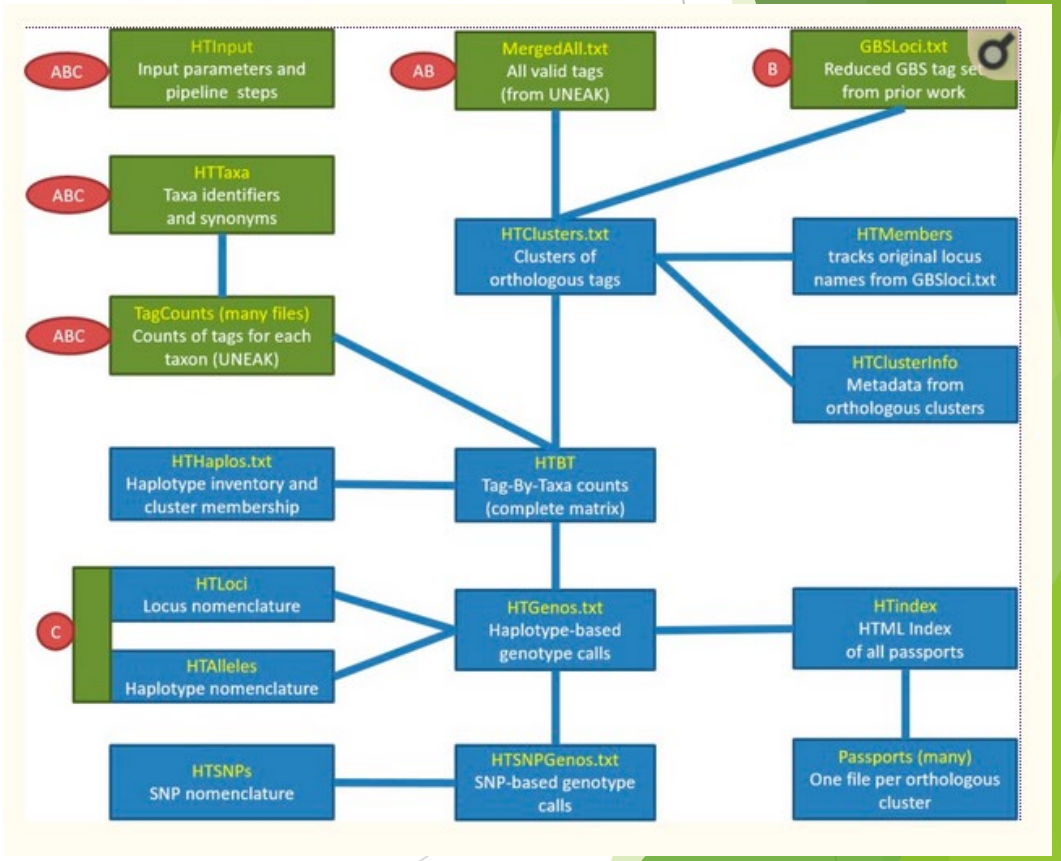
- Works with or without a reference
- Variant and genotype calling integrated with software for population genetics
- Assumes diploidy
- For polyploids, it is recommended to lower the “M” parameter to help filter paralogs (http://doi.org/10.1111/2041-210X.12775)
- Outputs VCF, but intermediate files are tab-delimited text and can be processed with custom software



https://doi.org/10.1111/mec.12354

Variant calling software: HaploTag

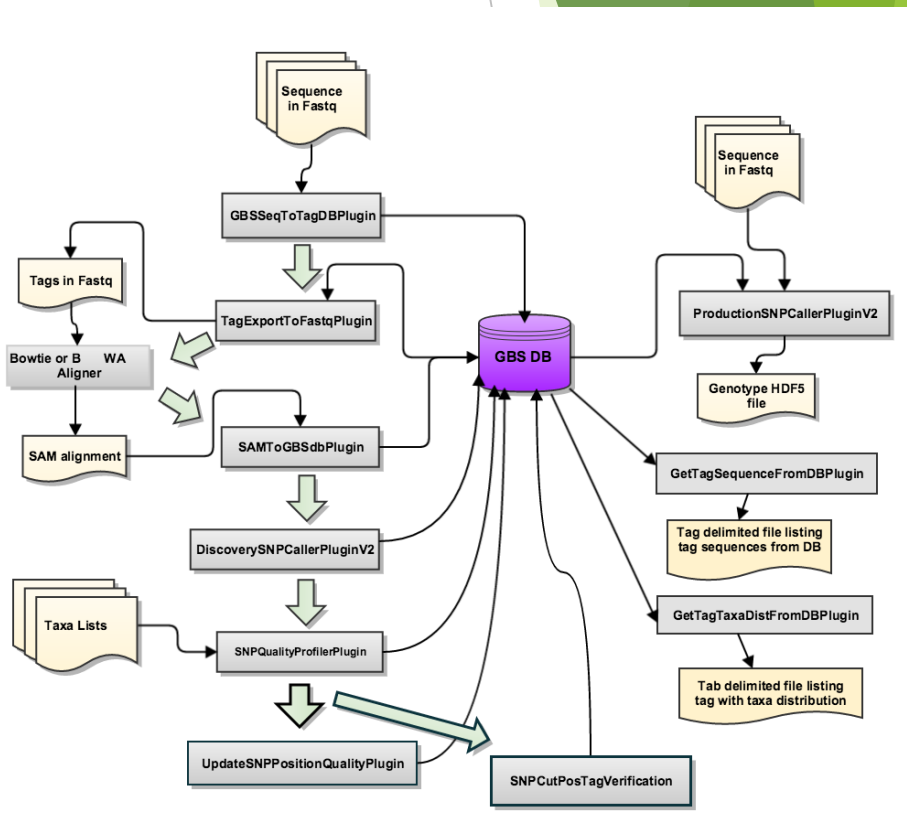
- Does not require reference genome
- Optimized for self-pollinating polyploid species
- Can output SNPs or haplotype-based genotypes



https://doi.org/10.1534%2Fg3.115.024596

Variant calling software: TASSEL-GBS

- Requires a reference genome
- Can run on a laptop
- Use TASSEL5 for most current version
- Assumes diploidy, but does output read depth in VCF
- Can always use “GetTagTaxaDistFromDBPlugin” to export raw table of read depth for each unique tag, and do your own processing from there



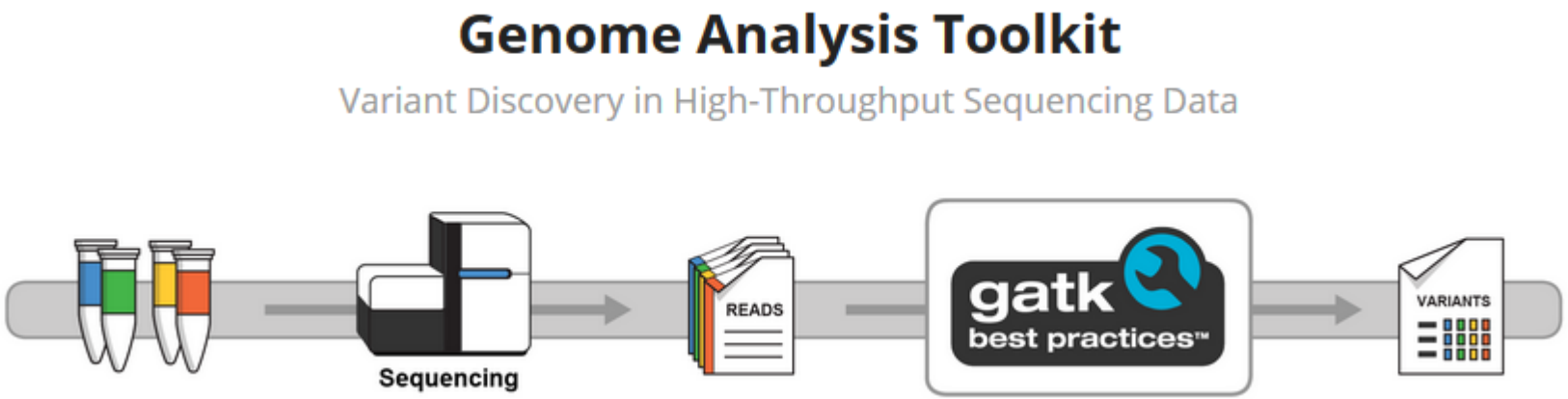
https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline

Variant calling software: TASSEL4-poly

- Requires reference genome
- Custom modified version of TASSEL4 that is not capped at read depth of 127
- Integrates with VCF2SM software for performing genotype calls with SUPERMASSA

Variant calling software: GATK

- Requires reference genome
- Designed for whole genome resequencing data, but can also work with GBS
- Can output polyploid genotypes, but uses a naïve model for genotype calling



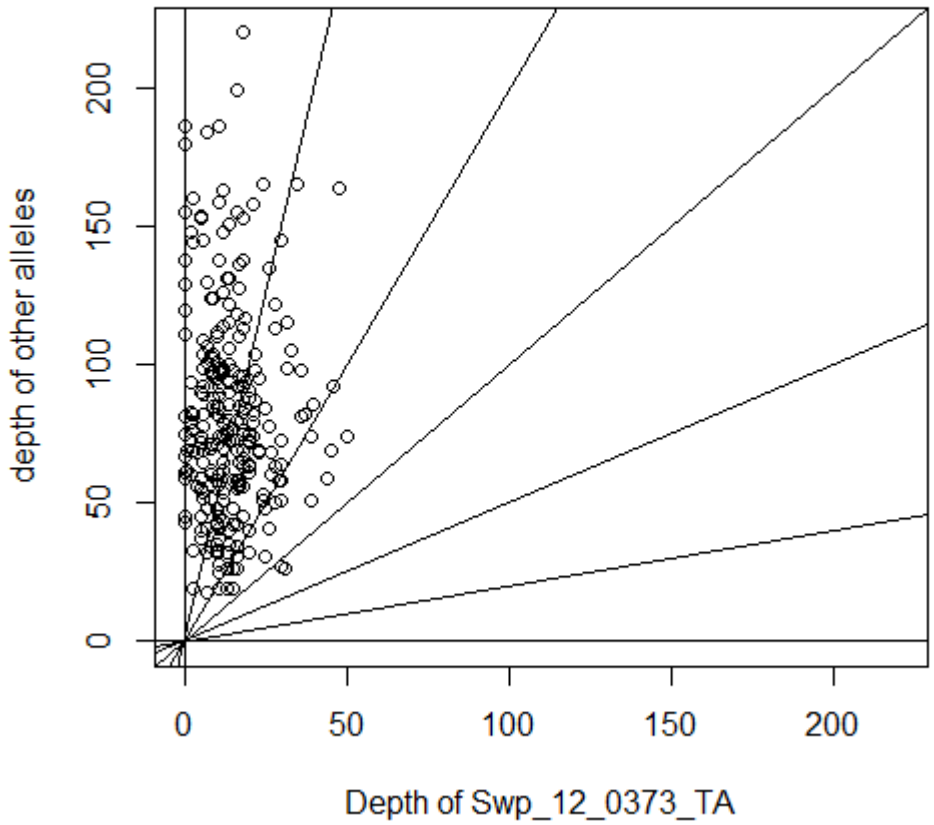
Variant calling software: FreeBayes

- Requires reference genome
- Designed for resequencing but works for GBS
- Uses sequence reads rather than alignments for calling variants (since one read can have multiple alignments)
- For polyploid variant discovery, lower the “min-alternate-fraction” argument below the default of 0.2
- Can perform polyploid genotype calling
- Preprint published 2012, hasn’t been through peer-review

Genotype calling issues in polyploids

- Biggest challenge: inferring allele dosage
- High genotype certainty requires very high read depth, which can be cost-prohibitive
- Heterozygote undercalling (allelic dropout) becomes a bigger issue when allele copy ratio is not 1:1
- Technical issues can cause read depth ratios to deviate from allele copy ratios more than we would expect
- How can we make the best genotype estimations for the amount of read depth that we can afford?

Lines = expected depth ratios
Points = actual depth ratios



Bayesian genotype calling

- $L(D|G)$: Likelihood of the observed distribution of allelic read depth (D) if a given genotype (G) were the true genotype
  - If the genotype is AAAB, what is the probability of getting 7 reads of A and 4 reads of B?
- $P(G)$: Prior probability of the genotype
  - How frequently to we expect to find AAAB in the population?
- $P(G|D)$: Posterior probability of the genotype
  - Given that we have 7 reads of A and 4 reads of B, what is the probability that AAAB is the true genotype?

$$P(G|D) = \frac{L(D|G)P(G)}{\sum_{i=1}^{k} L(D|G_i)P(G_i)}$$

For k possible genotypes

Practical effects of Bayesian genotype calling

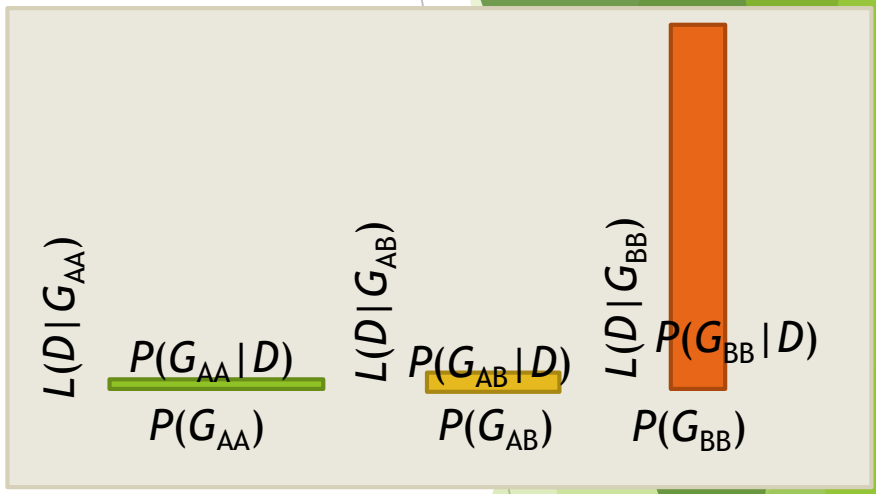
- High read depth → $P(G|D)$ is more influenced by $L(D|G)$
  - i.e. the observed allelic read depth ratio
- Low read depth → $P(G|D)$ is more influenced by $P(G)$
  - i.e. population parameters
- Read depth of zero → $P(G|D) = P(G)$

- At low read depth a genotype might appear homozygous, but if that allele is rare in the population, the homozygous genotype will have a low $P(G)$, and a heterozygous genotype might have the highest $P(G|D)$

$$P(G|D) = \frac{L(D|G)P(G)}{\sum_{i=1}^{k} L(D|G_i)P(G_i)}$$

Practical effects of Bayesian genotype calling

- High read depth → $P(G|D)$ is more influenced by $L(D|G)$
  - i.e. the observed allelic read depth ratio
- Low read depth → $P(G|D)$ is more influenced by $P(G)$
  - i.e. population parameters
- Read depth of zero → $P(G|D) = P(G)$

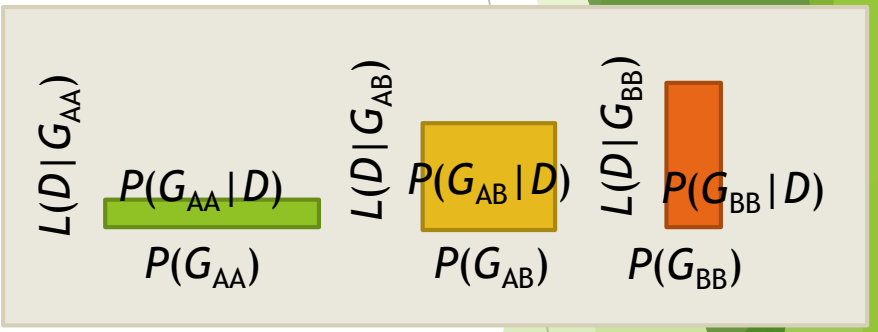


- At low read depth a genotype might appear homozygous, but if that allele is rare in the population, the homozygous genotype will have a low $P(G)$, and a heterozygous genotype might have the highest $P(G|D)$

$$P(G|D) = \frac{L(D|G)P(G)}{\sum_{i=1}^{k} L(D|G_i)P(G_i)}$$

Practical effects of Bayesian genotype calling

- High read depth → $P(G|D)$ is more influenced by $L(D|G)$
  - i.e. the observed allelic read depth ratio
- Low read depth → $P(G|D)$ is more influenced by $P(G)$
  - i.e. population parameters
- Read depth of zero → $P(G|D) = P(G)$

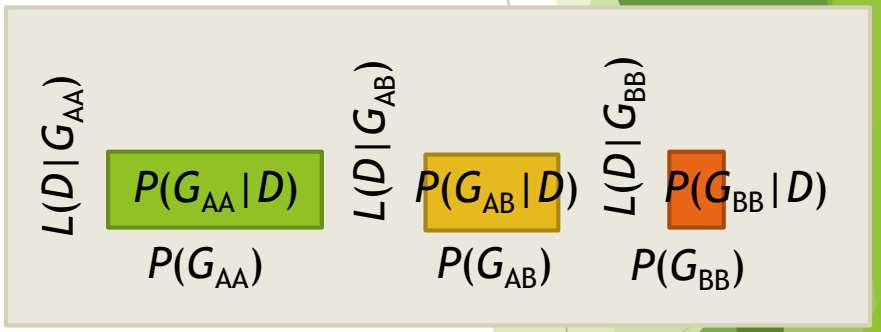


- At low read depth a genotype might appear homozygous, but if that allele is rare in the population, the homozygous genotype will have a low $P(G)$, and a heterozygous genotype might have the highest $P(G|D)$

$$P(G|D) = \frac{L(D|G)P(G)}{\sum_{i=1}^{k} L(D|G_i)P(G_i)}$$

Practical effects of Bayesian genotype calling

- High read depth → $P(G|D)$ is more influenced by $L(D|G)$
  - i.e. the observed allelic read depth ratio
- Low read depth → $P(G|D)$ is more influenced by $P(G)$
  - i.e. population parameters
- Read depth of zero → $P(G|D) = P(G)$

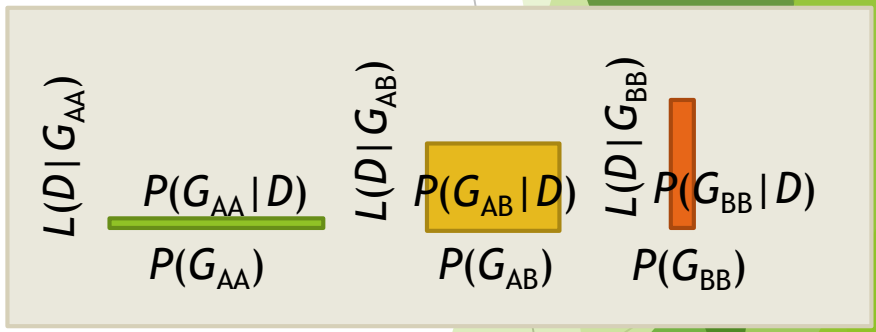


- At low read depth a genotype might appear homozygous, but if that allele is rare in the population, the homozygous genotype will have a low $P(G)$, and a heterozygous genotype might have the highest $P(G|D)$

$$P(G|D) = \frac{L(D|G)P(G)}{\sum_{i=1}^{k} L(D|G_i)P(G_i)}$$

Practical effects of Bayesian genotype calling

- High read depth → $P(G|D)$ is more influenced by $L(D|G)$
  - i.e. the observed allelic read depth ratio
- Low read depth → $P(G|D)$ is more influenced by $P(G)$
  - i.e. population parameters
- Read depth of zero → $P(G|D) = P(G)$



- At low read depth a genotype might appear homozygous, but if that allele is rare in the population, the homozygous genotype will have a low $P(G)$, and a heterozygous genotype might have the highest $P(G|D)$

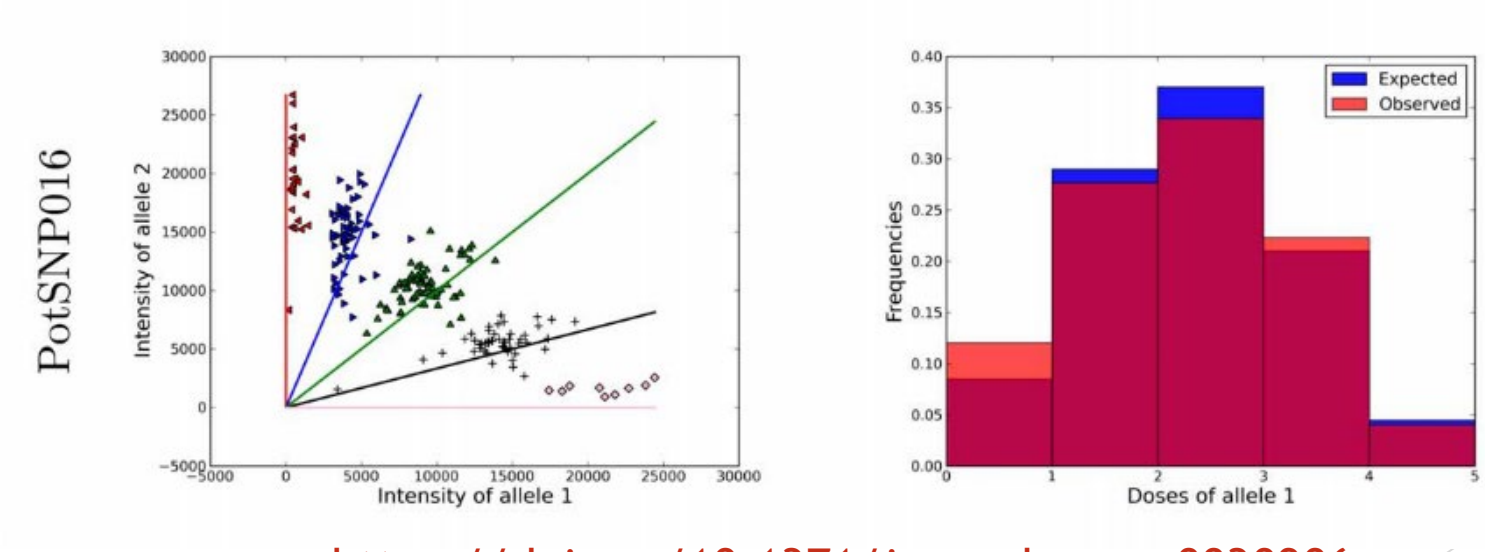
$$P(G|D) = \frac{L(D|G)P(G)}{\sum_{i=1}^{k} L(D|G_i)P(G_i)}$$

Genotype calling software: GATK and FreeBayes

- GATK uses uniform priors, and therefore can have high error rate at low read depth
- FreeBayes estimates priors from allele frequencies under Hardy-Weinberg Equilibrium

Genotype calling software: SUPERMASSA

- Originally designed for SNP array data, but works with read depth
- *P*(*G*) (genotype priors) can be based on Hardy-Weinberg Equilibrium or an F1 mapping population design
- *L*(*D*|*G*) (genotype likelihoods) are estimated a normal distribution of signal ratio, centered on the expected value (e.g. 0.33, or 1:3, for ABBB)
- Can also estimate ploidy if unknown



https://doi.org/10.1371/journal.pone.0030906

Genotype calling software: fitPoly

- Originally designed for SNP array data
- Priors can be based on HWE, F1, no constraints, or user-specified
- Tries different likelihood estimation based on bias of signal towards one allele or the other, and linear or quadratic relationship between signal and dosage

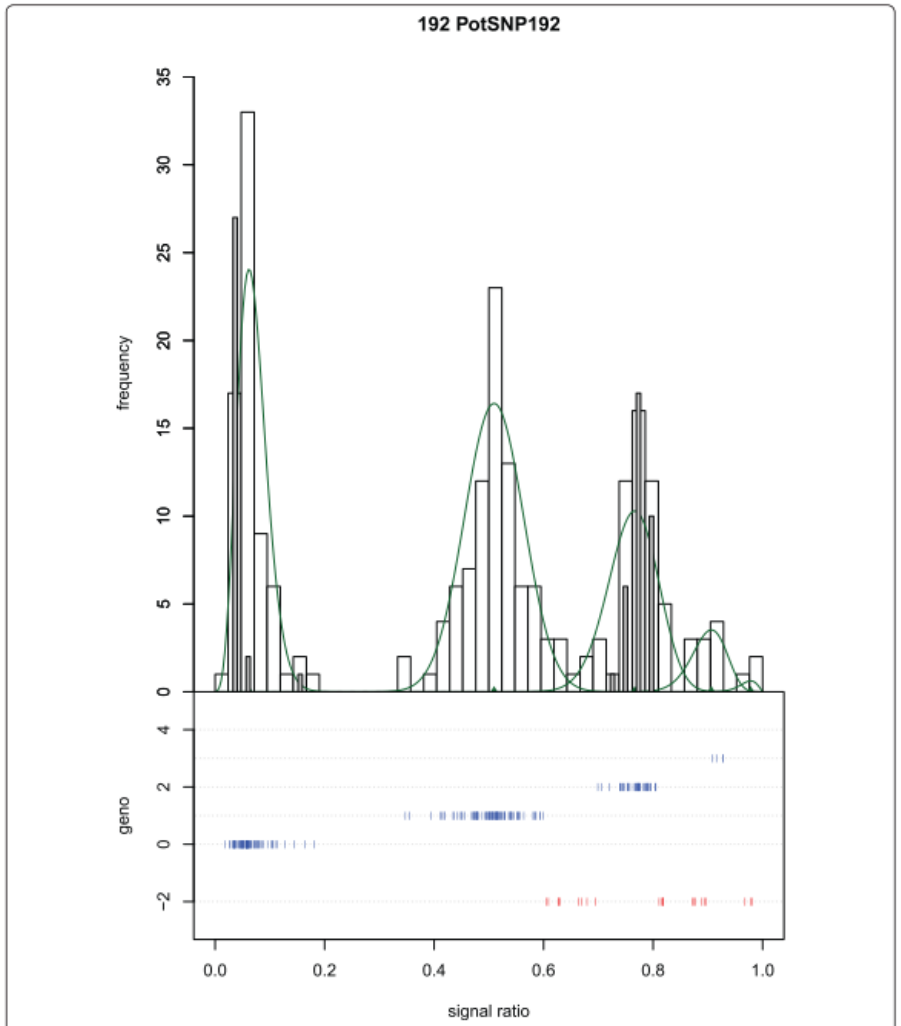


Figure 2 SNP showing a large gap. For this SNP (PotSNP192) the signal strength of the b allele is smaller than that of the a allele, resulting in a non-central position of the duplex peak and a wide gap between the nulliplex and simplex peak. For a general explanation see Figure 1.

https://doi.org/10.1186/1471-2105-12-172

Genotype calling software: EBG

- Designed for sequencing data
- $L(D|G)$ (genotype likelihoods) are estimated under a binomial distribution
  - E.g. 7 reads of A and 4 reads of B from AAAB =
  - $11!/(7! * 4!) * 0.75^7 * 0.25^4 = 0.172$
- $P(G)$ (genotype priors) based on HWE or inbreeding in autopolyploids.
- $P(G)$ can be estimated for allopolyploids if allele frequency in a parental species is known

https://doi.org/10.1093/bioinformatics/btx587

Genotype calling software: updog

- Designed for sequencing data
- Models technical issues with the data
  - Bias: some alleles get proportionately more sequencing reads than others
  - Overdispersion: allele depth ratios vary more than expected from the expected ratio
- $L(D|G)$ (genotype likelihoods) are estimated under a beta-binomial distribution
  - The probability of sampling a given allele from a given genotype is assumed to vary
- $P(G)$ (genotype priors) based on HWE, F1, or statistical distributions
- Outputs posterior mean genotypes
- Runs slowly due to estimation of many parameters

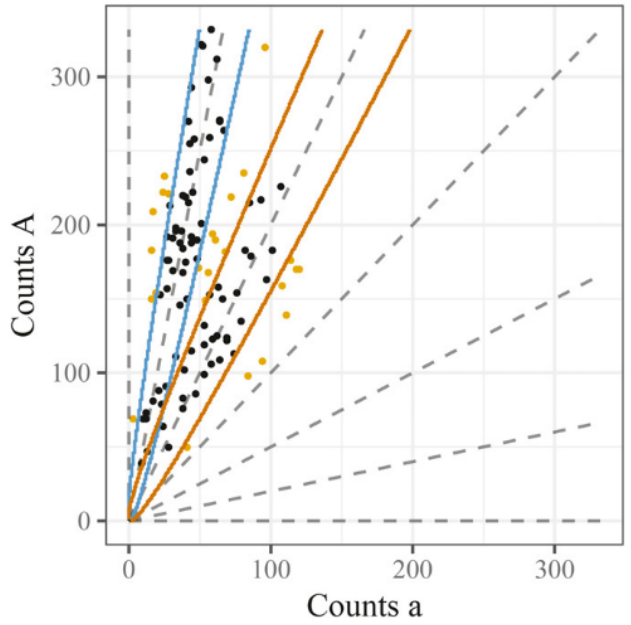
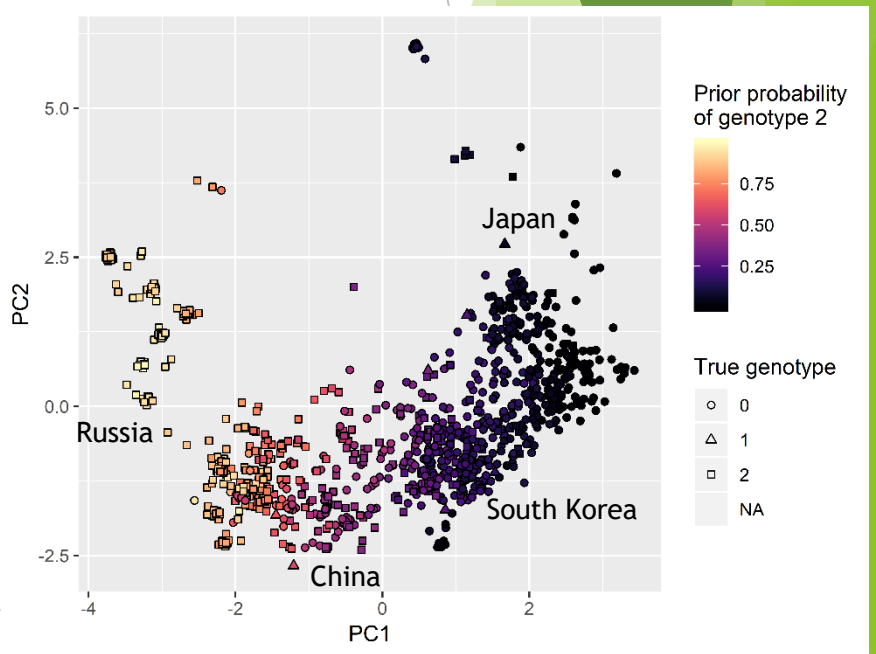


Figure 5 A genotype plot illustrating overdispersion compared with the simple binomial model. This figure shows the same SNP as the left panel of Figure 3 but with the points with >95% A reads removed. Solid lines indicate the 0.025 and 0.975 quantiles for the Binomial distribution with probabilities 4/6 (red) and 5/6 (blue). Points that lie within these lines are colored black; outside are colored orange. Under the binomial model only, 5% of the points should be orange, but, in fact, a significantly higher proportion (23.6%; *P*-value 1.2×10^{-11}) are orange.

https://doi.org/10.1534/genetics.118.301468

Genotype calling software: polyRAD

- Designed for sequencing data, especially GBS data
- $L(D|G)$ (genotype likelihoods) are estimated under a beta-binomial distribution
  - Overdispersion parameter only estimated once, reducing computation time with respect to updog
- $P(G)$ (genotype priors) are highly informed by biology, improving accuracy at low read depth, including zero depth
  - Any biparental mapping population design
  - Priors updated per-individual based on pop. structure and linkage disequilibrium
  - Allopolyploid and autopolyploid inheritance modes
- Outputs posterior mean genotypes



Quick demo of polyRAD with remaining time…