Variant Calling

Gonçalo Abecasis
Hyun Min Kang

Calling Consensus Genotype - Details

- Each aligned read provides a small amount of evidence about the underlying genotype
  - Read may be consistent with a particular genotype …
  - Read may be less consistent with other genotypes …
  - A single read is never definitive
- This evidence is cumulated gradually, until we reach a point where the genotype can be called confidently
- Let’s outline a simple approach …

Shotgun Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

A/C

Predicted Genotype

Shotgun Sequence Data

Sequence Reads

Reference Genome

P(reads|A/A, read mapped)= 1.0

P(reads|A/C, read mapped)= 1.0

P(reads|C/C, read mapped)= 1.0

Possible Genotypes

Shotgun Sequence Data

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A, read mapped)= P(C observed|A/A, read mapped)

P(reads|A/C, read mapped)= P(C observed|A/C, read mapped)

P(reads|C/C, read mapped)= P(C observed|C/C, read mapped)

Possible Genotypes

Shotgun Sequence Data

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A, read mapped)= 0.01

P(reads|A/C, read mapped)= 0.50

P(reads|C/C, read mapped)= 0.99

Possible Genotypes

Shotgun Sequence Data

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A, read mapped)= 0.0001

P(reads|A/C , read mapped)= 0.25

P(reads|C/C , read mapped)= 0.98

Possible Genotypes

Shotgun Sequence Data

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A , read mapped)= 0.000001

P(reads|A/C , read mapped)= 0.125

P(reads|C/C , read mapped)= 0.97

Possible Genotypes

Shotgun Sequence Data

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A , read mapped)= 0.00000099

P(reads|A/C , read mapped)= 0.0625

P(reads|C/C , read mapped)= 0.0097

Possible Genotypes

Shotgun Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A , read mapped)= 0.00000098

P(reads|A/C , read mapped)= 0.03125

P(reads|C/C , read mapped)= 0.000097

Possible Genotypes

Shotgun Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A, read mapped)= 0.00000098

P(reads|A/C, read mapped)= 0.03125

P(reads|C/C, read mapped)= 0.000097

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

Shotgun Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$P(Genotype|reads) = \frac{P(reads|Genotype)Prior(Genotype)}{\sum_G P(reads|G)Prior(G)}$$

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

Ingredients That Go Into Prior

- Most sites don't vary
  - P(non-reference base) ~ 0.001
- When a site does vary, it is usually heterozygous
  - P(non-reference heterozygote) ~ 0.001 * 2/3
  - P(non-reference homozygote) ~ 0.001 * 1/3
- Mutation model
  - Transitions account for most variants (C↔T or A↔G)
  - Transversions account for minority of variants

From Sequence to Genotype: Individual Based Prior

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098	**Prior(A/A) =** 0.00034	**Posterior(A/A) =** <.001
P(reads|A/C)= 0.03125	**Prior(A/C) =** 0.00066	**Posterior(A/C) =** 0.175
P(reads|C/C)= 0.000097	**Prior(C/C) =** 0.99900	**Posterior(C/C) =** 0.825

Individual Based Prior: Every site has 1/1000 probability of varying.

From Sequence to Genotype: Individual Based Prior

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098 **Prior(A/A) =** 0.00034 **Posterior(A/A) =** <.001

P(reads|A/C)= 0.03125 **Prior(A/C) =** 0.00066 **Posterior(A/C) =** 0.175

P(reads|C/C)= 0.000097 **Prior(C/C) =** 0.99900 **Posterior(C/C) =** 0.825

Individual Based Prior: Every site has 1/1000 probability of varying.

Sequence Based Genotype Calls

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson …
  - … 14x coverage would allow for 99.8% genotype accuracy
  - … 30x coverage of the genome needed to allow for errors and clustering

From Sequence to Genotype: Population Based Prior

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098	**Prior(A/A) =** 0.04	**Posterior(A/A) =** <.001
P(reads|A/C)= 0.03125	**Prior(A/C) =** 0.32	**Posterior(A/C) =** 0.999
P(reads|C/C)= 0.000097	**Prior(C/C) =** 0.64	**Posterior(C/C) =** <.001

Population Based Prior: Use frequency information from examining others at the same site.

In the example above, we estimated P(A) = 0.20

From Sequence To Genotype: Population Based Prior

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098 **Prior(A/A) =** 0.04 **Posterior(A/A) =** <.001

P(reads|A/C)= 0.03125 **Prior(A/C) =** 0.32 **Posterior(A/C) =** 0.999

P(reads|C/C)= 0.000097 **Prior(C/C) =** 0.64 **Posterior(C/C) =** <.001

Population Based Prior: Use frequency information from examining others at the same site.

In the example above, we estimated P(A) = 0.20

Sequence Based Genotype Calls

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson …
  - … 14x coverage would allow for 99.8% genotype accuracy
  - … 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data

Shotgun Sequence Data

Haplotype Based Prior

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098	**Prior(A/A) =** 0.81	**Posterior(A/A) =** <.001
P(reads|A/C)= 0.03125	**Prior(A/C) =** 0.18	**Posterior(A/C) =** 0.999
P(reads|C/C)= 0.000097	**Prior(C/C) =** 0.01	**Posterior(C/C) =** <.001

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.

In the example above, we estimated that 90% of similar chromosomes carry allele A.

Shotgun Sequence Data

Haplotype Based Prior

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098 **Prior(A/A) =** 0.81 **Posterior(A/A) =** <.001

P(reads|A/C)= 0.03125 **Prior(A/C) =** 0.18 **Posterior(A/C) =** 0.999

P(reads|C/C)= 0.000097 **Prior(C/C) =** 0.01 **Posterior(C/C) =** <.001

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.

In the example above, we estimated that 90% of similar chromosomes carry allele A.

Sequence Based Genotype Calls

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson …
  - … 14x coverage would allow for 99.8% genotype accuracy
  - … 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data

- **Haplotype Based Prior or Imputation Based Analysis**
  - Compares individuals with similar flanking haplotypes
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Can make accurate genotype calls with 2-4x coverage of the genome
  - Accuracy improves as more individuals are sequenced

Why Like Low Pass Sequencing? Whole Genome Sequencing Studies

	.5 – 1%	1 – 2%	2-5%
400 Deep Genomes (30x)			
Discovery Rate	100%	100%	100%
Het. Accuracy	100%	100%	100%
Effective N	400	400	400
3000 Shallow Genomes (4x)			
Discovery Rate	100%	100%	100%
Het. Accuracy	90.4%	97.3%	98.8%
Effective N	2406	2758	2873

Variant Filtering

- Typically, an initial set of variant calls might include many artifacts

- Artifacts can be identified through a series of unusual features:
  - Clusters of SNPs
  - Unusual flanking sequences, particularly homopolymer runs
  - Unusual balance of reads with forward / reverse orientation
  - Unusual fraction of reads with reference base in heterozygotes

- Typically, one compares features like allele frequency or transition / transversion ratio for different sets of potential filters
  - For example, if we find that the transition/transversion ratio is unusually low when there is another SNP within the flanking 5-bp, we might decide to filter those.

Example of Raw and Filtered Callsets

CATEGORY	# SNPs	#dbSNP (b129)	%dbSNP (b129)	Known Ts/Tv	Novel Ts/Tv	Overall Ts/Tv	%HM3 Redis-covery
PASS	**17,490,556**	**6,768,218**	**38.7**	**2.12**	**2.07**	**2.09**	**90.164**
FAIL	3,138,231	316,178	10.0	1.34	0.81	0.85	0.410
TOTAL	20,628,787	7,084,396	34.3	2.07	1.68	1.80	90.574

- Notice that the “failed” SNPs have:
  - lower fraction of previously known SNPs
  - Lower transition transversion ratio

What is the right amount of filtering?

- For a population genetics study?
- For a disease association study?

- What might be pitfalls of too much filtering?
- What might be pitfalls of too little filtering?

Rare variant test for high LDL vs low LDL