Applied Bioinformatics

Variant Calling
October 2015, Belgrade

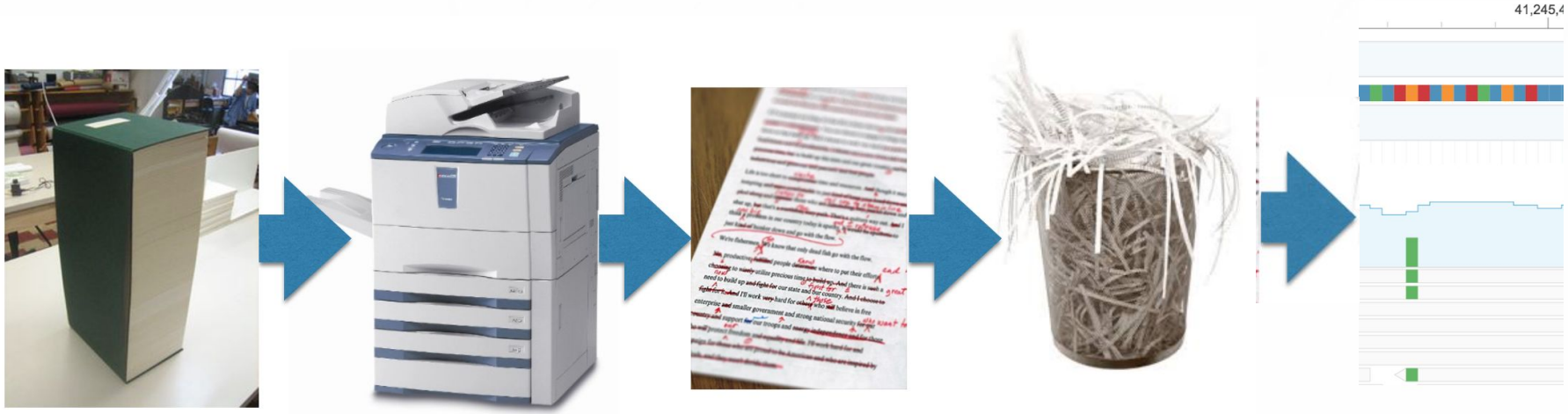
Goran Rakočević, PhD
goran.rakocevic@sbgenomics.com

Today's agenda

1. Overview of variant calling strategies
    a. Simple binomial model
    b. Additional information available
    c. Bayesian approaches
    d. Haplotype-based approaches
    e. Multisample calling
2. Using GATK Variant callers
    a. Exercise with the UnifedGenotyper
3. Variant filtration
    a. Hard filtering
    b. Variant Quality Score Recalibration

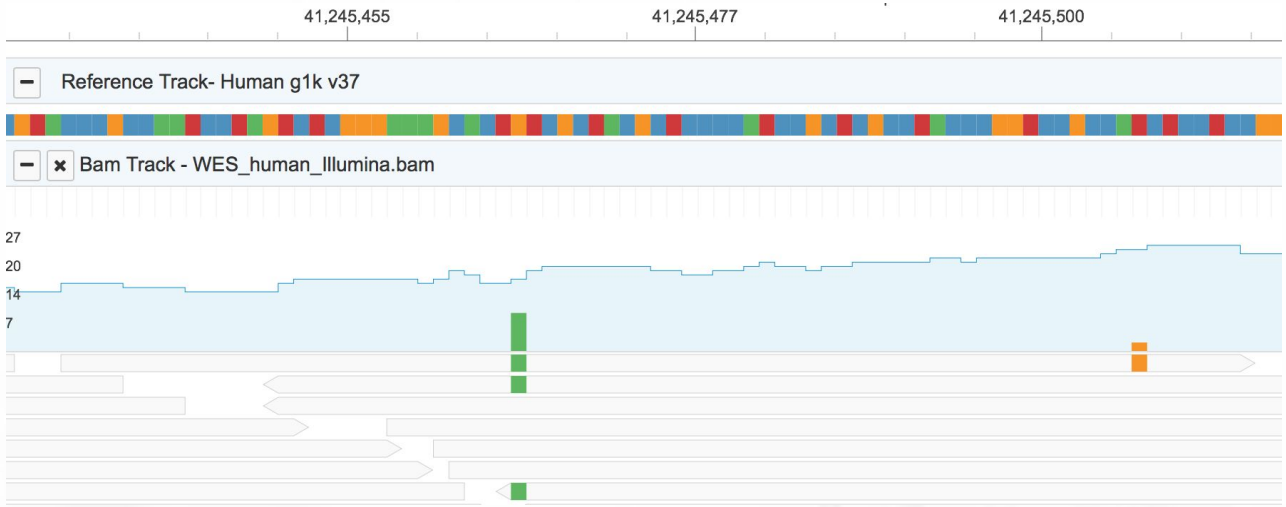
DNA Sequencing - Reminder

- We got a FASTQ file with the “reads” - little pieces of the genome



Looking at a pileup

- Pileup is the set of bases aligned to a single position on the genome



SevenBridges genomics

Examining a pileup

- Two possible cases:
  - All of the bases are the same nucleotide [A,T,C,G]
  - Different nucleotides exist in the pileup
- In the simplest case, assume diploidy
  - There can be only two alleles at a site
  - If there are more than two different letters in the pileup we will only consider the most common two (assume others are errors and discard them)

Case 1: All bases are the same

- Once again, two options:
  - All bases are the same and match the reference
    - Consider the site to be homozygous reference
  - All bases are the same and do not match the reference
    - Consider the site to be homozygous variant
    - But what if the pileup contains only one or two bases?
    - Probably an error, but still make the call and leave it to filtering
- Making the call looks fairly simple

Case 2: Two “letters” in the pileup

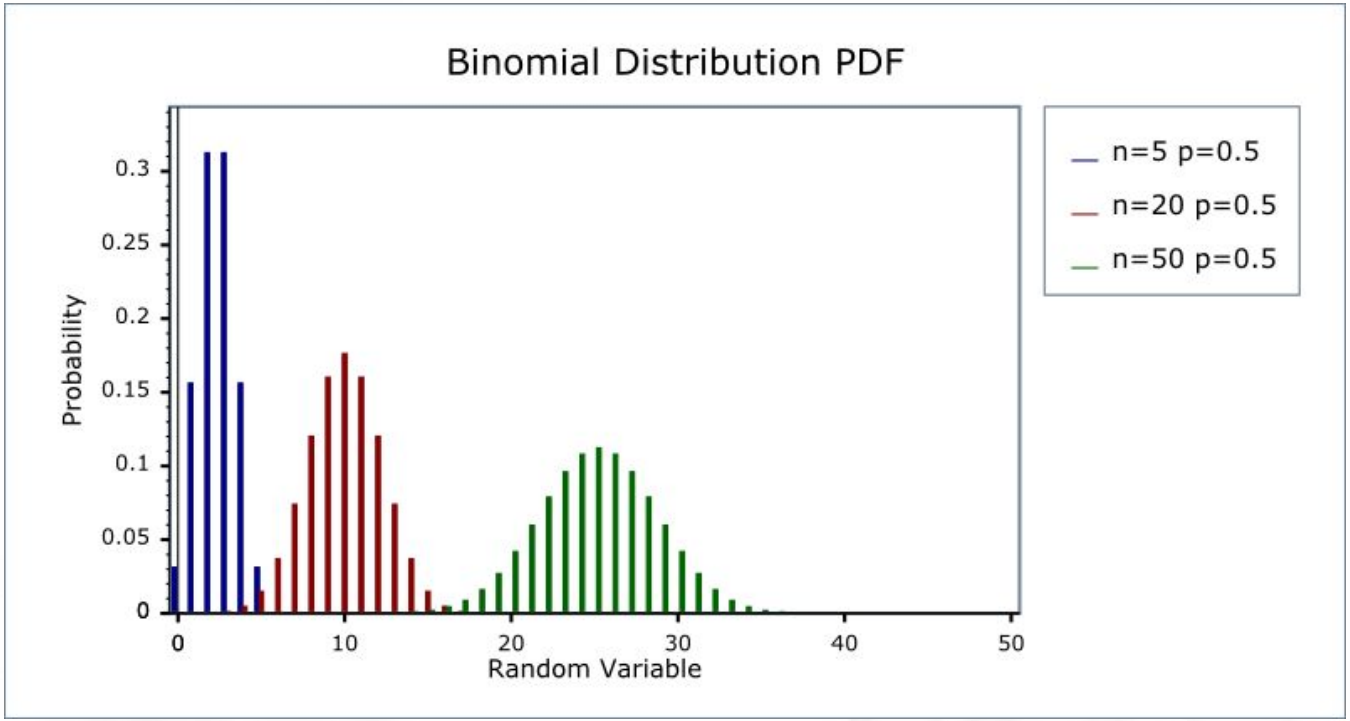
- If we have 15 As and 15 Ts, it’s a heterozygote!
- If we have 29 As and 1 T, the T is an error, previous slide!
- What about 5 Ts? Or 7?
  - Where is the threshold?
  - What happens with more or less than 30 bases?
- For this case we will need to use statistics
- We will look at the simplest model that was actually used

Binomial distribution

- Models the number of successes in a sequence of yes/no experiments
- Parameters:
  - n - number of trials
  - p - probability of a success in a single trial

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial distribution



SevenBridges genomics

Back to “two-letter” pileup

- Let’s call the two “letters” **b** and **b’** *(b, b’ ∈ [A, C, T, G])*
- Let **n** be the total number of bases, and **k** number of b’ bases
- Three possible explanations for the pileup:
  - Genotype is bb; k bases are errors, n-k are correct
  - Genotype is b’b’; n-k bases are errors, k are correct
  - Genotype is bb’; all n bases are correct
- Now we need to find the probabilities of these three cases
  - Will pick the most probable one!

Probabilities for different options

- Genotype is bb; k bases are errors, n-k are correct
  - Let ε be the probability of a sequencing error
  - What is the probability we draw n bases, and get k errors?

$$P(D|bb) = \binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k}$$

Probabilities for different options (2)

- Genotype is b'b'; n-k bases are errors, k are correct
  - Let ε be the probability of a sequencing error
  - What is the probability we draw n bases, get n-k errors?
    - It's binomial!

$$P(D|b'b') = \binom{n}{n-k}(1-\varepsilon)^k \varepsilon^{n-k}$$

Probabilities for different options (3)

- Genotype is bb’; n-k bases are errors, k are correct
  - Probability of drawing b or b’ is equal, and 0.5
  - What is the probability we draw n bases, k bs (or b’s)?
    - It’s binomial!

$$P(D|bb') = \binom{n}{k}\frac{1}{2^n}$$

Putting it together

- We now have P(D|bb), P(D|b'b'), and P(D|b'b)
- What about P(bb|D), P(b'b'|D), and P(b'b|D)?
- Bayes' theorem:

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

- P(D) - is always the same
- P(b'b) - fixed to 0.001 (or 0.2 for known variants)
- P(bb) = P(bb) = (1-r)/2
- Now we can pick max(P(bb|D), P(b'b'|D), and P(b'b|D))

Advanced stuff

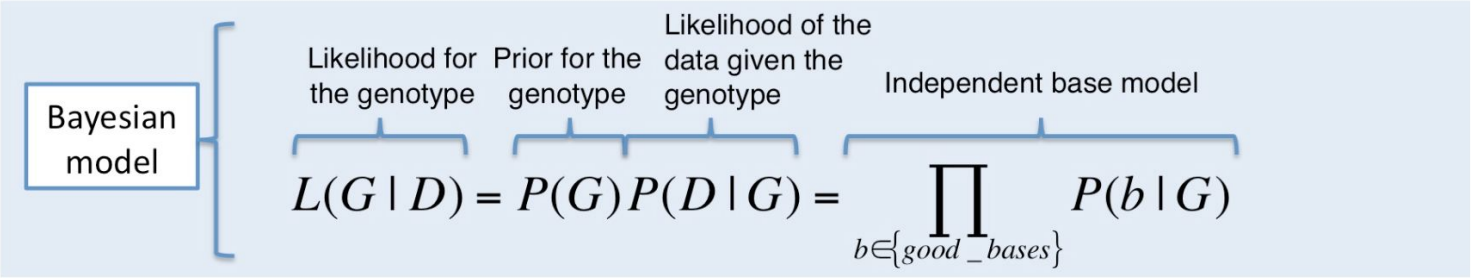
- We assumed a flat error rate
  - But we have Base qualities from the sequencer
  - Machine-specific error profiles
- We can look at mapping qualities
  - Mapping errors are a big source of errors
- We can look at haplotypes
  - Errors don’t segregate nicely
- Population-based methods
  - Separate variant calling from genotyping

Genome Analysis Toolkit (GATK)

- A collection of tools for NGS analyses
- Two variant Caller (de-facto standard)
  - UnifedGenotyper - a Bayesian model
  - HaplotypeCaller
- GATK Also includes tool for filtering variants
- …as well as many other things
- Written in Java
- https://www.broadinstitute.org/gatk/

GATK UnifiedGenotyper

- A Bayesian genotype likelihood model model



- Uses a platform-specific confusion matrix
- Can do joint calling on multiple samples
- Can call both SNP and Indels

Exercise 1: Calling variants

- Use UnifiedGenotyper to call variants on the small example BAM
- GATK jar file is located in the /opt folder
- UnifiedGenotyper command line:
  java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R reference.fasta
  -I reads.bam -o variants.vcf [-gim BOTH]
  [--dbsnp dbsnp.vcf]
- All of the input files are located in the ../data folder
- Examine the produced variants in the VCF file
  - VCF files are in a plain text format

Filtering variants

- Many Variant callers are designed for *sensitivity*
  - Call everything that looks plausible
- High sensitivity comes at expense of specificity
  - Some of the called stuff are false positives
- Filtering steps are used to reduce the false positives
  - Hard filtering (GATK VariantFiltration)
  - Machine learning (VQSR)
- http://gatkforums.broadinstitute.org/discussion/2806/howto-apply-hard-filters-to-a-call-set
- http://gatkforums.broadinstitute.org/discussion/39/variant-quality-score-recalibration-vqsr

VCF file format

- A plain text file format for storing variant data
- A number of line starting with ## -the header
- Main header line:
  #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1
- This is followed by the actual variant data, one entry per line
  22 10001 . A C 40 PASS DP=14 GT 0/1
- More than one sample can be in one line
- For details: http://samtools.github.io/hts-specs/VCFv4.2.pdf

Bgzip and Tabix

- Bgzip - block zip format
  - Break the VCF into blocks of several lines
  - Compress each block separately
- Tabix indexing tools
  - Makes and index on a bgzipped files
  - Allows a genomic range to fetched
- Not only VCF files can be indexed!
  - As long as there are columns with coordinates

Pysam - Python interface for VCFs

- Pysam can be used to process VCF files
- pysam.VariantFile
  - VariantFile(path_to_file)
  - for read in VariantFile(path_to_file):
- Reads are wrapped in VariantRecord objects
  - VariantRecord gives access to all of the data
- pysam.VariantFile supports fetching regions
  - The VCF file needs to be bgziped and tabix indexed!

Pysam VCF -exercise 2

- Create an VariantFile object
  - Use SRS000638.vcf.gz from the *data* folder
  - How many samples are there
- Take the first record from the VariantFile
  - What is the variant quality? Is the read filtered?
  - What INFO fields are present? What are the values?
  - What is the genotype of the first variant?
- How many unmapped reads are there in the file?
- Create a BedTool with the exome.bed from the ../data folder
- Get the regions from the BedTool that cover BRCA1 gene
- Fetch all of the variants from the BAM file mapped to BRCA1
  - Are any known? Check OMIM and DbSnp for associations

Questions?