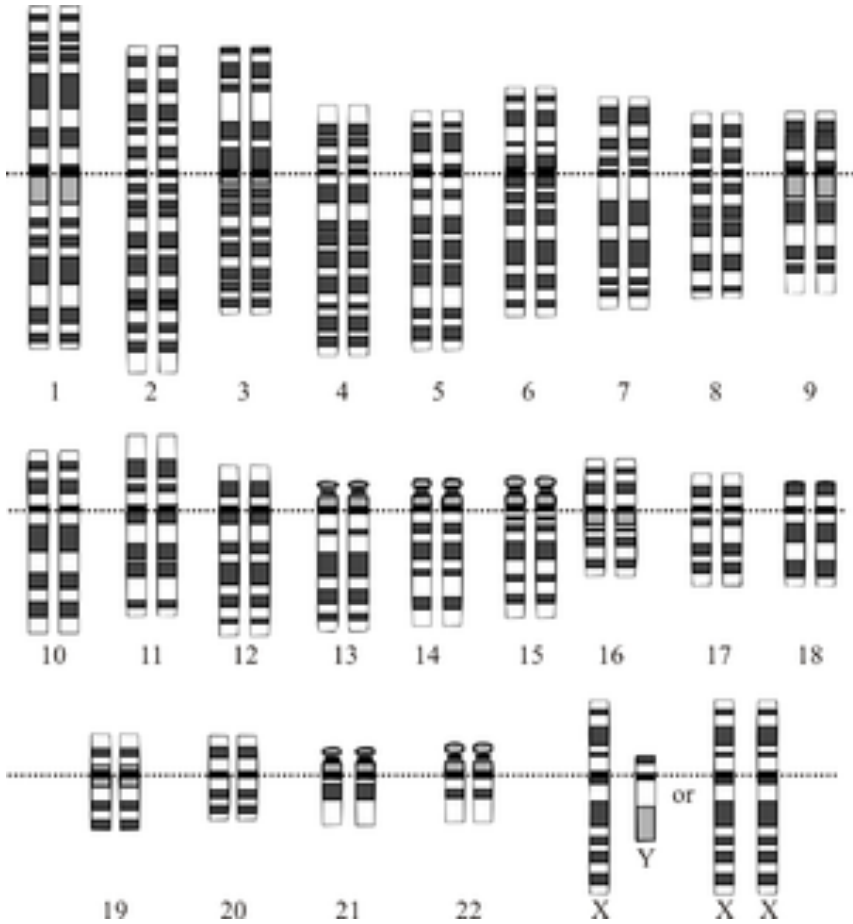
Genome Variant Calling: A statistical perspective

R. Gentleman

Genentech

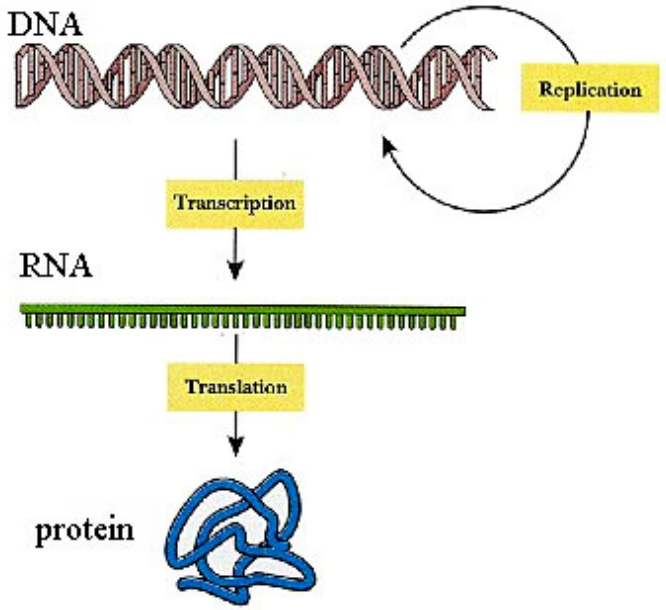
Notation

- the human genome is encoded on 23 (pairs) of chromosomes
- it is diploid (two copies of each); two copies of each gene
- the haploid version has ~3 billion nucleotides (nt), denoted ACGT
- at each locus you can be homozygote (the same on both chromosomes) or heterozygote (different)



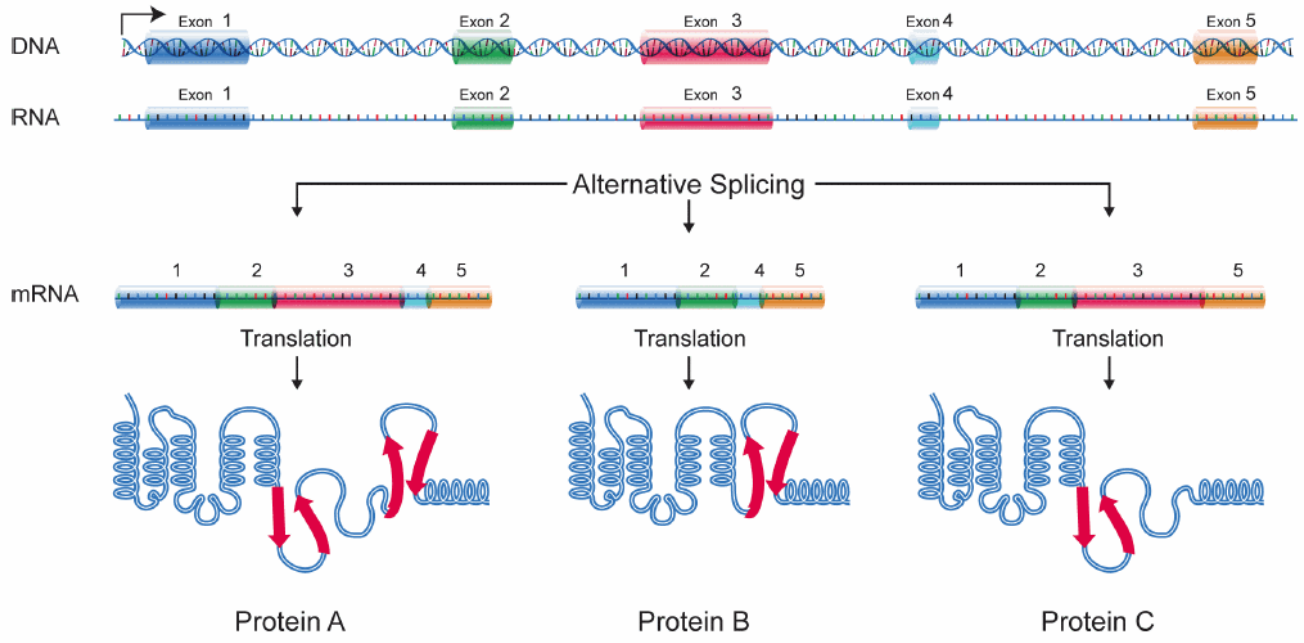
Central Dogma

- DNA -> RNA -> Protein
- DNA and RNA are relatively easy to sequence
- DNA: essentially two copies per cell
- RNA
  - not all genes expressed
  - some are at very high copy number
  - different lengths (capture probability is proportional to length and abundance)
  - transcription has higher error rates than DNA copying



Gene Structure

- genes are encoded in the DNA
  - variants are called alleles
- in higher organisms genes are organized with introns (spliced out) and exons (retained)



Sources of Variation

- Germline variation (SNPs or indels)
  - SNP: single nucleotide polymorphism
  - many known and reported in dbSNP (but there are lots of errors in dbSNP)
  - indel: insertion or deletion
  - copy# variation
- Germline or novel mutations
  - variation in normal tissue
- Somatic mutations (SNVs or indels)
  - variation in cancer
  - SNV: single nucleotide variation
- post-transcriptional modifications
  - RNA editing

Problem Specification

1. Variant calling:
   - what are the differences between the genome being sequenced and the/a reference
2. Genotyping:
   - what is the genotype of the genome being sequenced
3. Differences:
   - between two sequenced genomes
   - given data for two genomes (aligned to a reference) how do they differ

Data Sources

- DNA: normal cells
  - this is the “easiest” case
  - cells have known ploidy (diploid for humans)
  - the variations occur at rates that are known (or knowable)
  - cells are presumed clonal at the DNA level
- DNA tumor cells
  - harder because the ploidy is unkown
  - the cause and rates of mutation are unknown
  - the tumor is likely to be heterogenous
  - tumor has normal cells mixed in with it in almost all cases

Data Sources

- RNA: germline cells
  - harder than DNA because of variation in the rate of expression of different genes
  - post transcriptional modifications can occur
  - transcriptional fidelity is not that high
  - allele specific expression (it seems unlikely that alleles are expressed at equal rates)
- RNA: tumor cells (hardest)
  - all the problems with DNA + the problems listed above re RNA

DNA Variants

- identifying variants at particular genomic locations is straightforward
- translating that information into whether the variant is in a coding region, if so is it synonymous, non-synonymous (nonsense) etc depends on the **gene models** being used
- the VariantAnnotation package helps with these questions

RNA Variants

- alignment to the genome
  - likely more bias in this due to both differences between the RNA and the DNA plus splicing issues
- FIXME: more detail pls

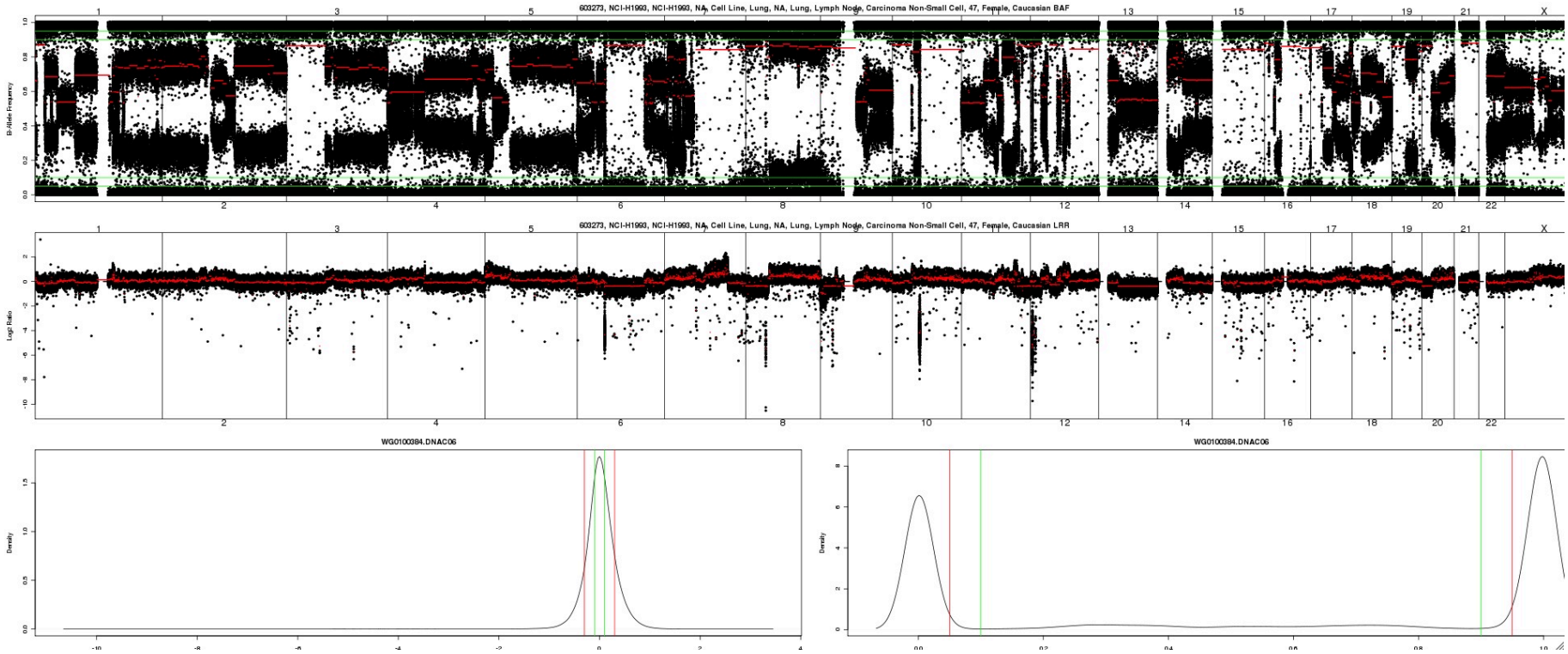
Software

- Reference genomes are distributed using the BSGenome class
  - eg. BSgenome.Hsapiens.UCSC.hg19
  - gives sequence level data
- Transcripts are distributed using the TranscriptDB class
  - eg. TxDb.Hsapiens.UCSC.hg19.knownGene
  - you can have multiple versions
  - provides a way to specify a set of transcripts for downstream processing

Rates of Variation (DNA)

- SNPs should be found at either 50% frequency or fixed
- Germline variants that are novel should be found at 50% frequency in the offspring
- Somatic mutations will be found at a frequency that is dependant on the age of the mutation and/or the fitness of the mutation (generally <50% frequency, however, allelic imbalance can also lead to higher frequency)

SNP Arrays


603273, NCI-H1993, NCI-H1993, NA, Cell Line, Lung, NA, Lung, Lymph Node, Carcinoma Non-Small Cell, 47, Female, Caucasian BAF
B-Allele Frequency
603273, NCI-H1993, NCI-H1993, NA, Cell Line, Lung, NA, Lung, Lymph Node, Carcinoma Non-Small Cell, 47, Female, Caucasian LRR
Log2 Ratio
WG0100384.DNAC06
Density
WG0100384.DNAC06
Density

Tumors/Cancer

- tumors arise from normal tissue
  - genome is very similar to the normal
- variants
  - point mutations: was C becomes A
  - insertions or deletions: a (small) amount of DNA is gained or lost
  - loss of heterozygosity (LOH): either lose a (part of) chromosome or select two copies of the same chromosome (now homozygous over that region)
- tumor samples tend to have some normal contamination
  - immune cells, blood, other tissue
  - attenuates our estimates of tumor specific variants towards zero

Sequencing

- whole genome sequencing (WGS)
  - all DNA is used
- exome sequencing
  - sequence only the exons
    - misses much of the regulatory genome
  - tends to be cheaper and gives higher coverage
  - only a small part of the genome is sequenced (3%)
- coverage:
  - number of reads that align over a locus
  - varies substantially (0 – 100's or 1000's)
  - determines your power and ability to detect variation

Sequencing: Error Rates

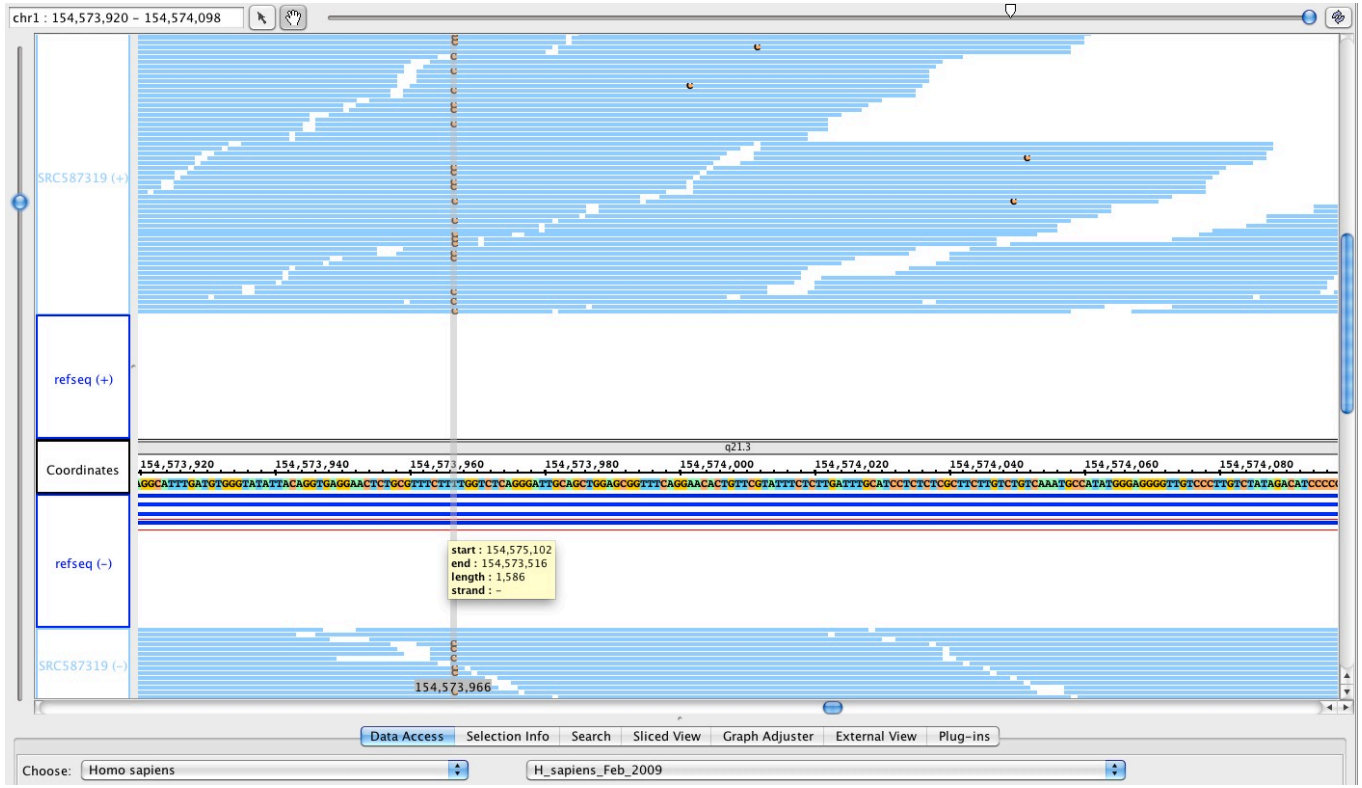
- DNA copying fidelity is about one error in 10^{-8}
  - each cell will have private mutations
- RNA transcription fidelity is one error in 10^{-4}
  - post-transcriptional modifications add complexity
- sequencing error rates vary but tend to be around one error in 10^{-3} (some reports of 1/300)
  - but there are location, sequence, biochemical reasons
- suggests the bulk of the observed differences are sequencing errors

Alignment

- we align reads to the reference genome
- we will do worse (not align or align fewer) where the genome is different from reference
  - this gives rise to reference bias
  - some groups perform a local de novo alignment to alleviate some of this
  - tumor genomes differ more – so we align worse and hence likely under-report
- it is difficult to align to regions that are duplicated or nearly duplicated in the genome
  - increases errors and can result in increased variant calling
  - UCSC provides self-chain data (you could also look at mappability)

How do we discover variation?

In a perfect world, after aligning these reads to the genome, variant calling would become a simple counting exercise…



Statistical Challenges

- multiple testing
  - many millions of tests (discrete probability distribution)
- varying power
  - coverage determines power, coverage varies
- varying size
  - also determined by coverage and since we have discrete distributions it varies
- bias
  - potential to under-call
  - we align to the reference genome (reference bias)

Preprocessing

- each variant must be supported by a minimum of two reads
- one must have a quality value greater than Q22
- variant must occur at different positions within the read
  - variants supported by only one cycle are removed
- one or more of the supporting cycles must occur outside the first and last 10% of the read
- remove variants with a more significant strand bias than the reference allele
  - default p-value cutoff is set to 0.001
  - for some capture methods there is significant strand bias at the extremes of the capture region

Variant Calling

- where are there differences between the genome sequence data and the reference?
- our reference genome is haploid
  - we assume homozygous at every locus
- H_0: the genome (G) and ref (R) are the same (G is homozygous identical to the reference)
- under H_0 all reads should be the reference allele
  - errors are due to sequencing errors
- every heterozygous locus is a variant (in this case), some homozygous loci are too

Variant Calling

- often used algorithm: if #Variants > L, and coverage > K, call a variant
  - K is artificial, the requirement should be based on evidence against H_0, not on coverage
  - size and power changes with coverage
- Pr(2 or more non-reference alleles| H_0) is a Binomial calculation, p=10^{-3}, n=coverage
  - n=10, 10^{-5}
  - n=50, 10^{-3}

Variant Calling

- SNVmix (Goya *et al*, Bioinformatics, 2010) had two additional criteria
  - quality of the nt sequenced
  - quality of the alignment of the read
- suggest we should discount evidence from
  - low quality nts
  - low quality alignments
- propose a complicated estimation procedure

Variant Calling: p-value adj

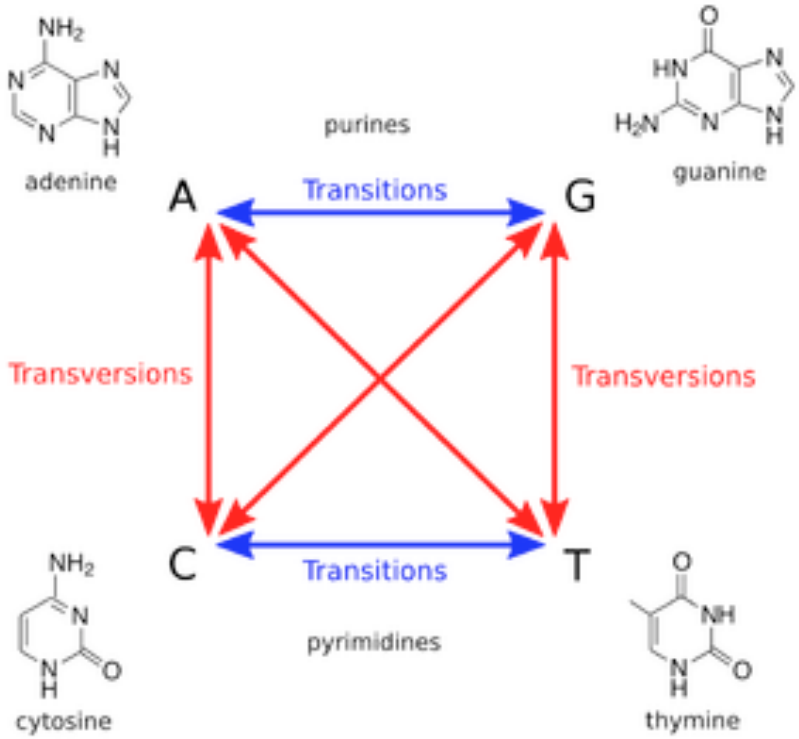
- the distributions of the test statistic is discrete
- the distributions of the p-values are too
- as coverage increases, for a fixed cut-off, the size of the test decreases
- our p-values, if aggregated and sorted, would come in runs according to coverage and observed count
- a stratified approach would be useful
  - divide the genome into coverage regions
  - compute FDR or other within coverage regions

Genotyping

- call the actual genotype at a locus
- typically done using a Bayesian approach
  - we can compute P(D|G)

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)}$$

  - use prior information on P(G)



The GATK pipeline

GATK uses a Bayesian model to reduce false positives

Use assumptions about heterozygosity, and platform-specific error probabilities

Assumes data are generated according to a Binomial distribution

GATK single sample genotype likelihoods

Bayesian model

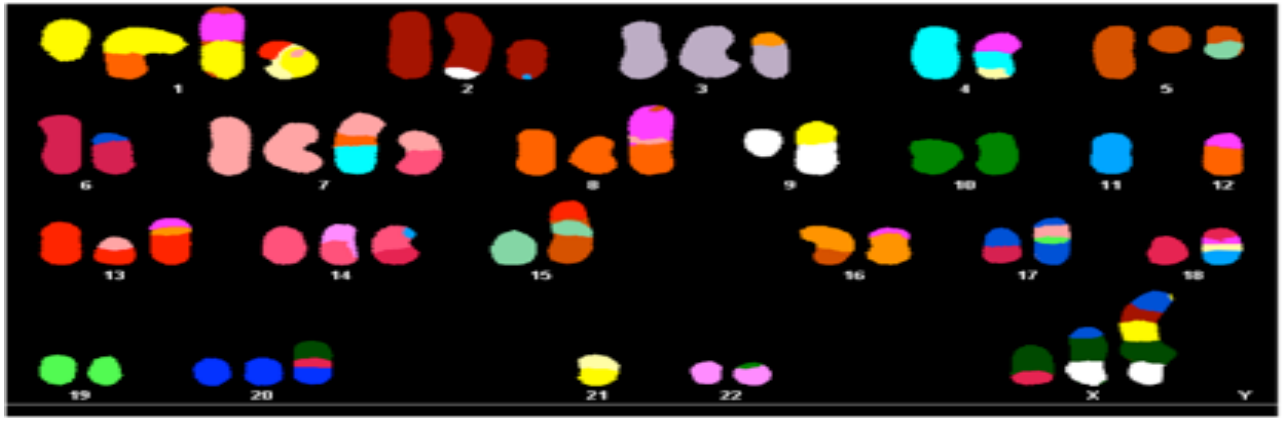
Likelihood for the genotype | Prior for the genotype | Likelihood of the data given the genotype | Independent base model

$$L(G \mid D) = P(G)P(D \mid G) = \prod_{b \in \{good\_bases\}} P(b \mid G)$$

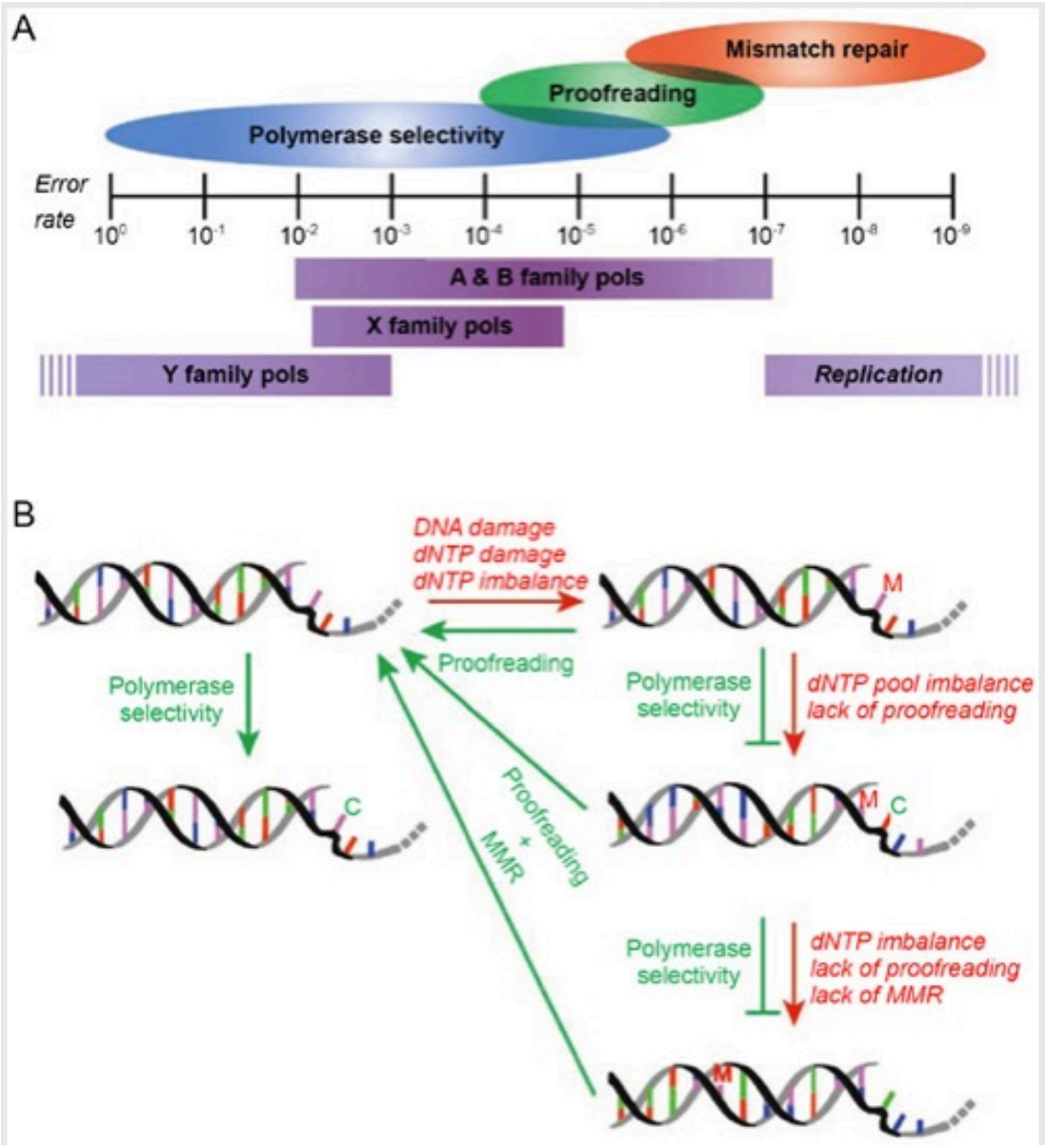
- Priors applied during multi-sample calculation; P(G) = 1
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only "good bases" are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS
- P(b | G) uses a platform-specific confusion matrix
- L(G|D) computed for all 10 genotypes

Genotyping: Tumors

- for tumors copy number varies and the variation in the genome tends to be a function of the type of cancer (or lifestyle: smoking induces G->T transversions) so reasonable priors are harder to obtain
- the genome is not diploid!
- tumor may not be clonal (so this is not a well posed problem)
- different DNA repair mechanisms fail in cancer increasing the rate of specific variations



Repair mechanisms



- Different Pol molecules have different replication fidelity
- Errors in replication are normally corrected MMR process

Calling Differences

- we focus on differences at the single nucleotide level
  - structural variation and indels are not considered just yet
- we now ignore the reference (sort of) and just want to compare two genomes
  - common comparison tumor (T) and normal (N)
- comparison is asymmetric
  - we want to discover gains in tumor (mutations)
  - losses are less interesting (capture with LOH, in/del)
  - losses tend to be due to structural changes not single nucleotide events
- we cannot call tumor specific variants at loci where we have insufficient coverage in N to make a call

Differences: Algorithm

- Case I: identify all loci where we call a variant in **T**umor and not in **N**ormal
- our concern is that the variant is present in **N** we just did not detect it
- assume **N** is heterzygous for the **T** allele and one other, with prob determined by the proportions observed in **T**
- test: Pr(as extreme or more extreme in the **N** | **T** frequencies)

Example

- **T** has 10 A's and 2 G's at locus L:
  - called variants: A and G
  - p(A) = 10/12, p(G) = 2/12
- **N** has 22 A's and 1 G at locus L:
  - called variants: A
- test: what is the probability that we see 0 or 1 G in **N**, when p(A) = 10/12 and p(G)=2/12, and we had 23 "tries"
  - P(X<=1 | p=2/12, n=23) = 0.084
  - so we **would not** call this a mutation
  - if the coverage was 33, with one G, then p=0.01 and we **would** call this a mutation

Example

- Criticisms
  - we have treated the Tumor data as special and used the observed proportions as if they were known values
  - for low coverage this is somewhat more problematic than for high coverage
  - copy number might change between T and N
- you could try other approaches, including a variety of two sample tests
  - but you would need to be careful that you are testing the hypothesis you intend
  - Fisher's exact test (FET) is not appropriate for example as we are not interested in whether the frequencies differ (which is what it tests)

Algorithm

- Case II: No variant in T (same as ref) but N is not ref.
  - essentially the same approach as before

Next Steps

- what is the effect of my variant?
- this depends very much on the set of gene models you want to use
- VariantAnnotation package provides tools to start to investigate this question
- locateVariants function
- predictCoding function

Thanks

- C. Barr
- R. Bourgon
- **J. Degenhardt**
- M. Huntley
- M. Lawrence
- T. Wu
- Z. Zhang
- W. Huber
- S. Dudoit
- V. Obenchain