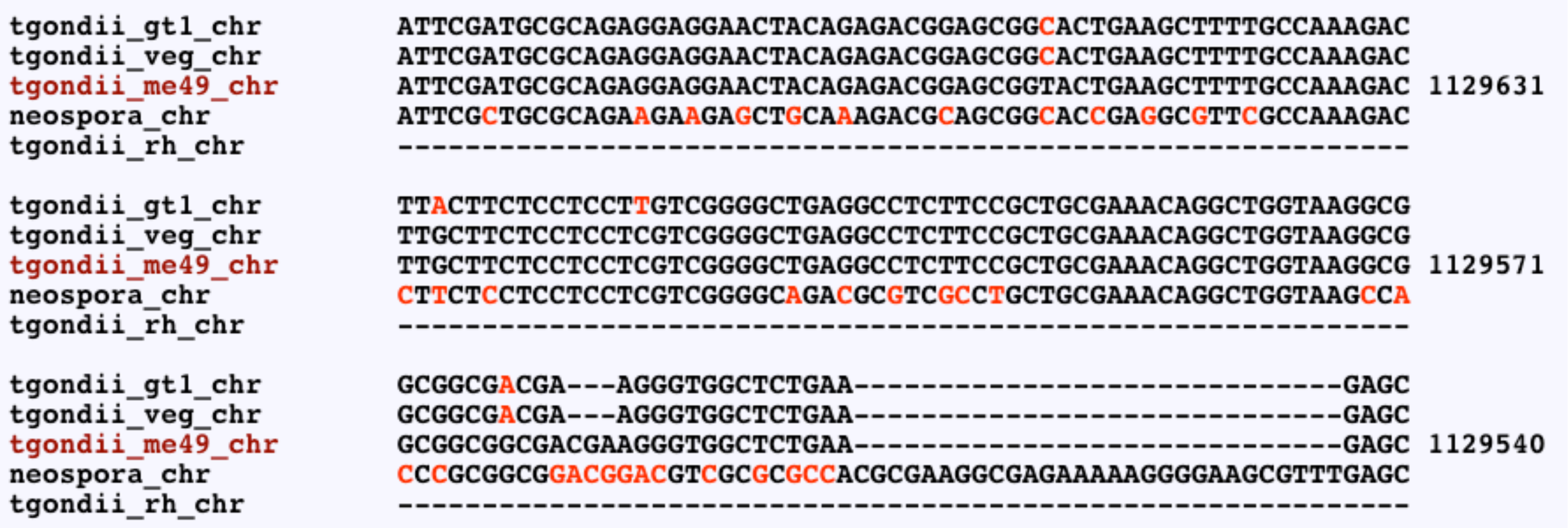
What are SNPs?

- Single Nucleotide Polymorphisms
  - DNA sequence differences distinguishing individuals within a species (allelic polymorphisms)
  - Diploid (aneuploid / polyploid) species may have allelic SNPs within an individual isolate (heterozygotes)
  - Note that SNPs are not the only form of allelic polymorphism, however … but EuPathDB does not currently support insertions & deletions (indels)



```
tgondii_gt1_chr         ATTCGATGCGCAGAGGAGGAACTACAGAGACGGAGCGGCACTGAAGCTTTTGCCAAAGAC
tgondii_veg_chr         ATTCGATGCGCAGAGGAGGAACTACAGAGACGGAGCGGCACTGAAGCTTTTGCCAAAGAC
tgondii_me49_chr        ATTCGATGCGCAGAGGAGGAACTACAGAGACGGAGCGGTACTGAAGCTTTTGCCAAAGAC 1129631
neospora_chr            ATTCGCTGCGCAGAAGAAGAGCTGCAAAGACGCAGCGGCACCGAGGCGTTCGCCAAAGAC
tgondii_rh_chr          ------------------------------------------------------------

tgondii_gt1_chr         TTACTTCTCCTCCTTGTCGGGGCTGAGGCCTCTTCCGCTGCGAAACAGGCTGGTAAGGCG
tgondii_veg_chr         TTGCTTCTCCTCCTCGTCGGGGCTGAGGCCTCTTCCGCTGCGAAACAGGCTGGTAAGGCG
tgondii_me49_chr        TTGCTTCTCCTCCTCGTCGGGGCTGAGGCCTCTTCCGCTGCGAAACAGGCTGGTAAGGCG 1129571
neospora_chr            CTTCTCCTCCTCCTCGTCGGGGCAGACGCGTCGCCTGCTGCGAAACAGGCTGGTAAGCCA
tgondii_rh_chr          ------------------------------------------------------------

tgondii_gt1_chr         GCGGCGACGA---AGGGTGGCTCTGAA-----------------------------GAGC
tgondii_veg_chr         GCGGCGACGA---AGGGTGGCTCTGAA-----------------------------GAGC
tgondii_me49_chr        GCGGCGGCGACGAAGGGTGGCTCTGAA-----------------------------GAGC 1129540
neospora_chr            CCCGCGGCGGACGGACGTCGCGCGCCACGCGAAGGCGAGAAAAAGGGGAAGCGTTTGAGC
tgondii_rh_chr          ------------------------------------------------------------
```

SNPs in EuPathDB are derived from two sources

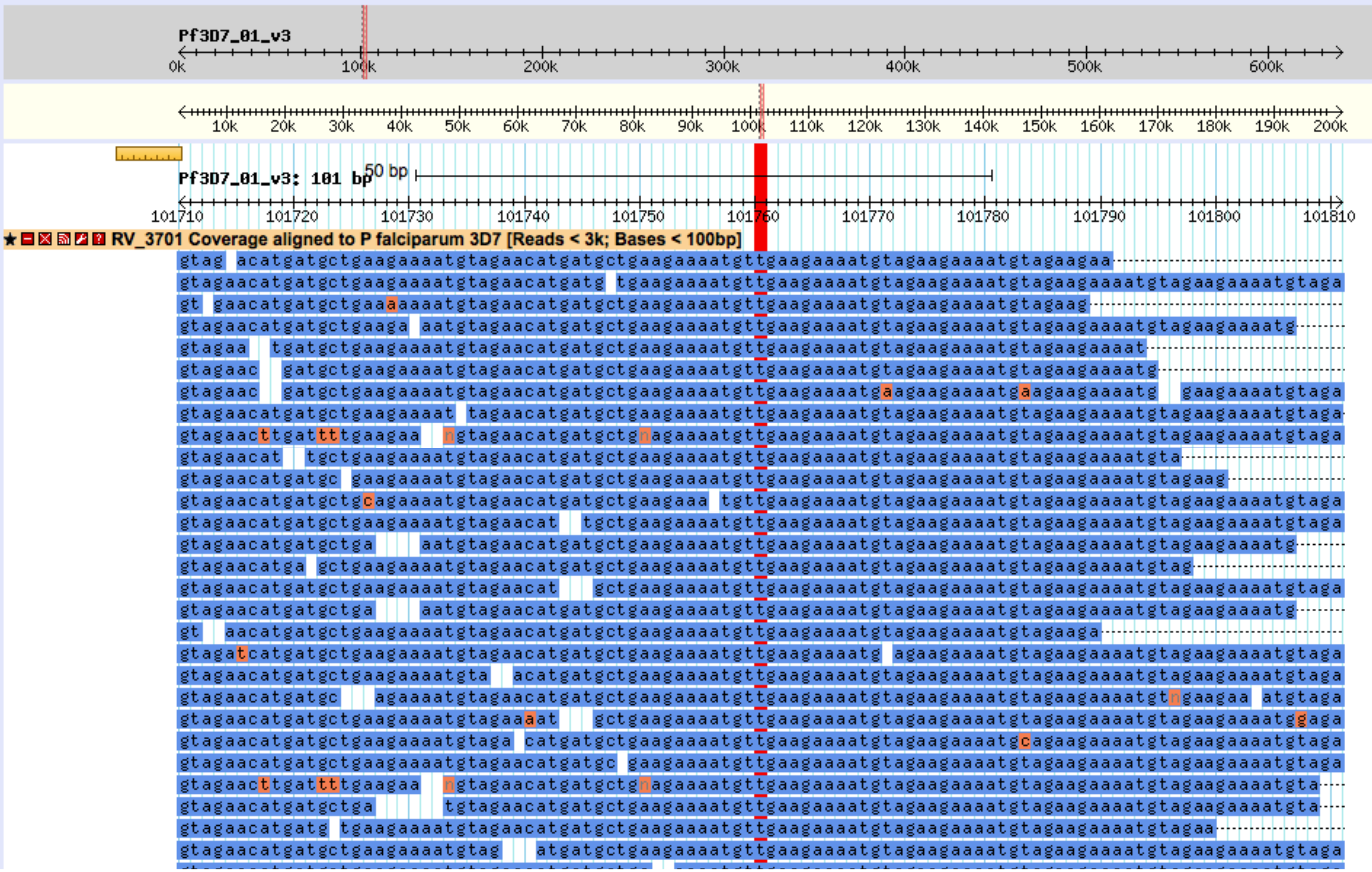
- Chip-based hybridization assays
  - Arrays can bedesigned toallow identification of SNP alleles in a given DNA sample.
  - PlasmoDB supports several such arrays, including 'barcode' arrays designed rapidly and inexpensively type field isolates
- Direct (deep) sequencing of isolate DNA
  - Reads are aligned to reference genome(s), and SNPs called based on differences
- What are isolates? *– explore on EuPathDB!*

Homozygous / Heterozygous SNPs

- Ploidy of organism is critical
  - Replicative forms of Apicomplexans are haploid
  - Amoebae are diploid
  - *Giardia* is (approximately) tetraploid
  - African trypanosomes are diploid
  - Ploidy in *T. cruzi* is not entirely clear
  - *Leishmania* is (sometimes, partially) aneuploid
- Why does this matter for SNP calling/queries?
  - Read frequency is the defining parameter
- What does a heterzygous SNP look like?
  - http://tinyurl.com/o3tr9ly *record page*
  - http://tinyurl.com/ppcxmqo *GBrowse view*

In a haploid isolate -- all reads should be identical

- *What might account for variation?*

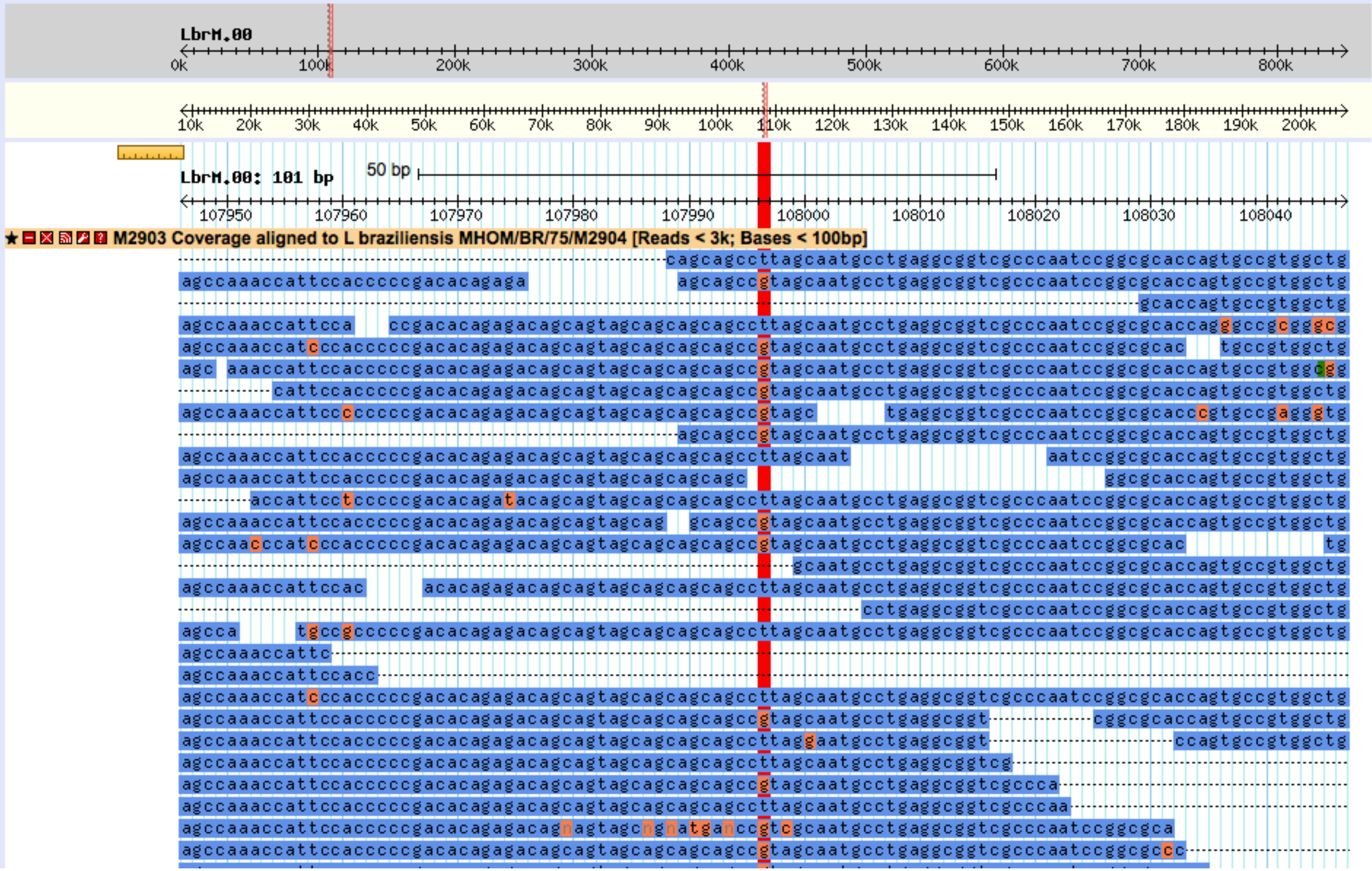


RV_3701 Coverage aligned to P falciparum 3D7 [Reads < 3k; Bases < 100bp]

```
gtag acatgatgctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaa
gtagaacatgatgctgaagaaaatgtagaacatgatg tgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gt gaacatgatgctgaaaaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaag
gtagaacatgatgctgaaga aatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatg
gtagaa tgatgctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaat
gtagaac gatgctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatg
gtagaac gatgctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgaagaagaaaatgaagaagaaaatg gaagaaaatgtaga
gtagaacatgatgctgaagaaaat tagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacttgatttgaagaa ngtagaacatgatgctgnagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacat tgctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgta
gtagaacatgatgc gaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaag
gtagaacatgatgctgcagaaaatgtagaacatgatgctgaagaaa tgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacatgatgctgaagaaaatgtagaacat tgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacatgatgctga aatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatg
gtagaacatga gctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtag
gtagaacatgatgctgaagaaaatgtagaacat gctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacatgatgctga aatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatg
gt aacatgatgctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaaga
gtagatcatgatgctgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatg agaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacatgatgctgaagaaaatgta acatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacatgatgc agaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtngaagaa atgtaga
gtagaacatgatgctgaagaaaatgtagaaaat gctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatggaga
gtagaacatgatgctgaagaaaatgtaga catgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgcagaagaaaatgtagaagaaaatgtaga
gtagaacatgatgctgaagaaaatgtagaacatgatgc gaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
gtagaacttgatttgaagaa ngtagaacatgatgctgnagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgta
gtagaacatgatgctga tgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgta
gtagaacatgatg tgaagaaaatgtagaacatgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaa
gtagaacatgatgctgaagaaaatgtag atgatgctgaagaaaatgttgaagaaaatgtagaagaaaatgtagaagaaaatgtagaagaaaatgtaga
```

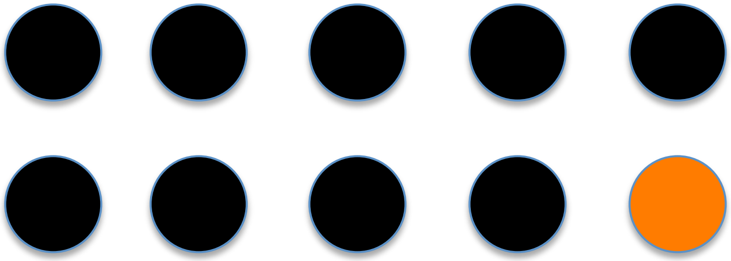
In a diploid isolate – expect a 50:50 ratio

- *What might account for variation?*



Minor allele frequency

- *10 isolates in this example … 9 display one allele and one has a different allele. For example, at position 237,748 on P. falciparum chromosome 2, 9 isolates contain a T and one contains G.*



= 10% minor allele frequency

Calling eukaryotic SNPs (in EuPathDB)

- Retrieve reads (pref. paired end), ideally from SRA).
- Align to reference using Bowtie2 (end-to-end).
- Realign around indels using GATK.
- Identify SNPs, indels and consensus sequence using VarScan (min depth 5, min frequency 20%).
- Every isolate alignment checked for every SNP … infer comparisons between non-reference isolates if sufficient evidence to make statistically valid call (like reference, or not like reference).
- SNPs stored in database … based on this reference.
- http://tinyurl.com/pebcdlz

  (SNP record page & link to alignment)

Why do we care? What can we do with SNPs?

- SNPs are genetic markers
  - Distinguish specific strains / isolates.
  - Enable fine structure mapping of phenotypes in genetic crosses or association studies.
- Identify SNPs based on a useful characteristics.
  - Within a group of isolates, based on:
    - allele frequency
    - chromosomal position (or position within genes)
    - other parameters
  - Compare two groups of isolates to identify distinctive SNPs.
- Identify Genes
  - Identify genes that are appear to be under selection based on SNP characteristics:
    - Number of SNPs (coding, non-coding, synonymous *etc*)
    - Ratio of non-synonymous / synonymous SNPs … identifying genes under purifying or diversifying (balancing) selection.

Purifying vs Diversifying Selection

- Purifying selection: evolutionarily constraints serve to maintain primary amino acid sequence
  - Low ratio of non-synonymous / synonymous codons
  - Tend to be genes critical for basic metabolic processes (encoding enzymes, cell cycle related proteins, *etc*).
  - *Note: very high A+T content in P. falciparum yields biased codon usage, skewing NS/S ratio.*
- Diversifying selection: evolutionarily advantageous to change amino acid sequence rapidly
  - High non-synonymous / synonymous codon ratio
  - Tend to be genes encoding proteins recognized by the host immune response: surface antigens, *etc*
- Assessing NS/S requires good coverage, maximizing reliable SNP calls
- *P. reichenowi* a recent outgroup … useful for highly conserved genes, but not those changing rapidly