What are SNPs?

- Single Nucleotide Polymorphisms
  - Differences between individuals (isolates) of a species
  - EuPathDB: differences between strains / isolates (if clonal)
    - Some organisms are diploid so will also have allelic SNPs within strain
  - Genes that are different due to SNPs are alleles.
  - Our model does not include insertions/deletions (indels) currently

```
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 assays
  - Arrays are designed that allow identification of SNP alleles given a DNA sample. There are multiple different arrays in PlasmoDB. Also Barcode assays (24 SNPs) assayed by PCR
  - Isolate DNA is then assayed on these arrays.
- Direct deep sequencing of DNA from isolates.
  - Reads are aligned to a reference genome and SNPs called.
- What are Isolates – Session this afternoon where we look at isolates.

How are resequencing SNPs called in EuPathDB

- Reads (hopefully paired end) are received from provider (best if via SRA)
- Reads are aligned to the reference using Bowtie2 (end-to-end)
- Reads realigned around indels using GATK
- SNPs, indels and consensus sequence generated using VarScan (min read depth 5, min read frequency 20%)
- SNPs based on this reference comparison are stored in the DB
- Every isolate alignment is checked for every SNP position and if sufficient evidence to make “like reference” call then this is indicated in the DB along with evidence for call.
- In the end, have allele(s) for every isolate at every position if the evidence from the alignment warrants it.
- http://tinyurl.com/pebcdlz (SNP record page - link to alignment)

Homozygous / Heterozygous SNPs

- Ploidy of organism is critical
  - Apicomplexans are haploid for majority of cycle
  - Trypanosomatids, Amoebas, etc are diploid (or worse)
- Why does this matter for SNP calling/queries?
  - Read frequency is the defining parameter
- What does a heterzygous SNP look like?
  - http://tinyurl.com/o23wndt – record page
  - http://tinyurl.com/nh5jbxj

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.
  - Enable population studies etc.
- Identify SNPs based on a variety of characteristics.
  - Within a group of isolates (includes allele frequency and confidence parameters)
    - Restrict to location (on chromosome or within genes)
  - Compare two groups of isolates to identify SNPs that distinguish the two groups.
- 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 provides an indication of whether genes are under purifying or diversifying (balancing) selection.

Purifying vs. Diversifying selection

- Purifying selection (gene is evolutionarily constrained to maintain the primary amino acid sequence)
  - Genes that have a low Non-synonymous / Synonymous ratio
  - Tend to be genes critical for basic metabolic processes such as enzymes, cell cycle related etc.
  - Due to very high A/T bias in *falciparum*, the ratio of non-synonymous/synonymous can be skewed due to severe codon bias.
    - *P. reichenowi* (closely related species infecting chimps) is less A/T rich and makes a good "strain" for identifying genes under purifying selection.
- Diversifying (balancing) selection (it is evolutionarily advantageous to quickly change the amino acid sequence)
  - Genes that have a high Non-synonymous / Synonymous ratio.
  - Tend to be things like surface antigens that the organisms use to escape immune detection.

Note Parameter Help

- Parameters all have help.
- Description of search contains detailed help with figures for some of the parameters…