Polymorphisms, SNPs and Alleles
What are SNPs?

• **Single Nucleotide Polymorphisms**
  – Differences between individuals (used in forensics)
  – EuPath: differences between strains / isolates
    • Kinetoplastida are diploid so will also have allelic SNPs within strain
  – Genes that are different due to SNPs are alleles.
  – Does not include indels currently

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>tgondii_gt1_chr</td>
<td>ATTCGATGCGCAGAGGAGGAACTACAGAGACGGAGCGGCACCTGAAGGCTTTTGCAAAAGAC</td>
</tr>
<tr>
<td>tgondii_veg_chr</td>
<td>ATTCGATGCGCAGAGGAGGAACTACAGAGACGGAGCGGCACCTGAAGGCTTTTGCAAAAGAC</td>
</tr>
<tr>
<td>tgondii_me49_chr</td>
<td>ATTCGATGCGCAGAGGAACTACAGAGACGGAGCGGCACCTGAAGGCTTTTGCAAAAGAC</td>
</tr>
<tr>
<td>neospora_chr</td>
<td>ATTCGCTGCGCAGAAAGAGAGCTCAGAAGACGCAGGACGCACGCACGCTCAGCAGAAGAC</td>
</tr>
<tr>
<td>tgondii_rh_chr</td>
<td>ATTCGCTGCGCAGAAAGAGAGCTCAGAAGACGCAGGACGCACGCACGCTCAGCAGAAGAC</td>
</tr>
<tr>
<td>tgondii_gt1_chr</td>
<td>TTAGTTTCTCTCTCCTGTCGAGGCTAGCCCTCTTCCTCGCTGCAGACAGCCTGGTAAGGCG</td>
</tr>
<tr>
<td>tgondii_veg_chr</td>
<td>TTGCTTTCTCTCTCCTGTCGAGGCTAGCCCTCTTCCTCGCTGCAGACAGCCTGGTAAGGCG</td>
</tr>
<tr>
<td>tgondii_me49_chr</td>
<td>TTGCTTTCTCTCTCCTGTCGAGGCTAGCCCTCTTCCTCGCTGCAGACAGCCTGGTAAGGCG</td>
</tr>
<tr>
<td>neospora_chr</td>
<td>CTTCTCCTCCTCCTGCTCGGAGAGCGTGCTCAGCTGCAACAGGCTGGTAAGGCG</td>
</tr>
<tr>
<td>tgondii_rh_chr</td>
<td>CTTCTCCTCCTCCTGCTCGGAGAGCGTGCTCAGCTGCAACAGGCTGGTAAGGCG</td>
</tr>
<tr>
<td>tgondii_gt1_chr</td>
<td>GCGGCGACGA----AGGTTGCTCTGAA------------------------GAGC</td>
</tr>
<tr>
<td>tgondii_veg_chr</td>
<td>GCGGCGACGA----AGGTTGCTCTGAA------------------------GAGC</td>
</tr>
<tr>
<td>tgondii_me49_chr</td>
<td>GCGGCGGCGAGCAAGGTTGCTCTGAA------------------------GAGC</td>
</tr>
<tr>
<td>neospora_chr</td>
<td>CCGGCCGGGAGGAGACGTGGCGCGCCACGGCAAGGCGAGAAAGGGGAAAGCCTGGTGAAGG</td>
</tr>
<tr>
<td>tgondii_rh_chr</td>
<td>CCGGCCGGGAGGAGACGTGGCGCGCCACGGCAAGGCGAGAAAGGGGAAAGCCTGGTGAAGG</td>
</tr>
</tbody>
</table>
What can we do with SNPs?

• 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 indicates whether genes are under purifying or diversifying (balancing) selection.

• SNPs are genetic markers
  – Distinguish specific strains / isolates.
  – Enable fine structure mapping of phenotypes in genetic crosses or association studies.

• Have sets of queries against SNPs to identify SNPs based on a variety of characteristics.
  – Location (on chromosome or within genes)
  – Allele frequency
  – Presence in isolate assays
  – Isolate characteristics
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*, best comparator is *P. reichenowi*.

• Diversifying 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.
  – Use comparators high on the list (have more sequence coverage and thus more SNPs). *Reichenowii* frequently is not a good comparator because these genes are changing so rapidly that they may not be conserved well enough to call SNPs.
Alleles in ToxoDB

• ToxoDB contains three *Toxoplasma* strains fully sequenced and annotated. Other species are not far behind (*E. Histolytica*, *T. brucei*, *P. falciparum* ...)
  – Ideally, there would be a 1-1-1 mapping for genes but this is not always the case.

• Results of gene queries are filtered by strain and species in the case of *Neospora*.

• Functional data (expression, proteomics) is all mapped to ME49 currently.

• This means that care must be taken when constructing *Toxoplasma* strategies as instances from different strains won’t intersect even though they may be from the same gene.

• We use the “Expand” function under add step to expand a result into all the instances (alleles) for that query set.