Polymorphisms, SNPs and Alleles

EuPathDB Workshop June 2009
What are SNPs?

- **Single Nucleotide Polymorphisms**
  - Differences between individuals (used in forensics)
  - EuPath: differences between strains / isolates
  - Genes that are different due to SNPs are alleles.
  - Does not include indels currently
What can we do with SNPs?

• Genes
  – Identify genes that 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 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.
Purifying vs. Diversifying selection

- **Purifying selection** (gene is evolutionarily constrained to maintain the primary amino acid sequence)
  - Genes that have a very 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 uses to escape immune detection.
  - Use comparators high on the list (have more sequence coverage and thus more SNPs). *Reichenowi* 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.
  - 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 the different alleles won’t intersect even though they may be from the same gene.
- We use the “Expand” function under the transform pull down to expand a result into all the instances (alleles) for that query set.