Variation

- Types of Variation
- SNPs

Types of Variation

- Indels – Insertions or deletions
- CNV – Copy Number Variation
- SNPs – Single Nucleotide Polymorphisms

Slides adapted from Irene Paratheodorou

What are SNPs?

- DNA sequence variations occurring when a single nucleotide in the genome is altered
- Frequency of 1% or more
- Occur in both coding and non-coding regions
- Occur every 100-300 bases
- ~15 million in human genome

Figure from Alexander Kozik, Compositae Genome Project, UCDA

Categories of SNPs

- Missense/Non-synonymous
  - Changes AA
  - May alter function / structure of protein
  - Cause of some monogenetic diseases e.g. cystic fibrosis
- Nonsense
  - Introduces a stop codon
  - Similar consequences to missense SNPs

Categories of SNPs

- Synonymous
  - Does not change coding sequence
  - May alter splicing
- Non-coding
  - May be promoter or regulatory sequences
  - Might affect gene expression

SNP Discovery

- Usually from sequencing
- Separate errors from 'real' differences in sequence and assessing frequency in population

Figure from Alexander Kozik, Compositae Genome Project, UCDA

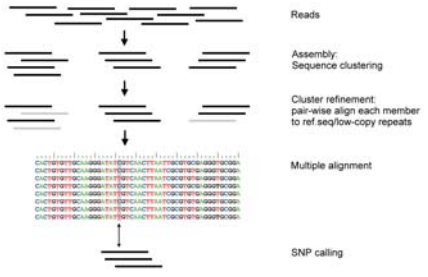
SNP Discovery: Experimental

- Re-sequencing alleles from different haplotypes
- Targeted re-sequencing of certain regions
- SSCP: Single Strand Conformation Polymorphism analysis

SNP Discovery: Computational Discovery Systems

Pipelines/Systems	Data types	Source
ssahaSNP/PolyBayes	EST data; paralogue identification; genome reference requirement	Ning *et al.* 2001/Marth et al. 1999
PolyPhred/SNPdetector/novoSNP	PCR re-sequencing from diploid samples	Stephens *et al.* 2006/Zhang *et al.* 2005/Weckx *et al.* 2005
QualitySNP	EST data; paralogue identification	Tang J *et al.* BMC Bioinformatics 2006
PanGEA	Reads from 454 pyrosequencing	Kofler R *et al.* BMC Bioinformatics 2009
MAQ	Reads from NGS; genome reference requirement	Li H *et al.* Genome Research 2008

Pipeline of SNP discovery



SNP resources

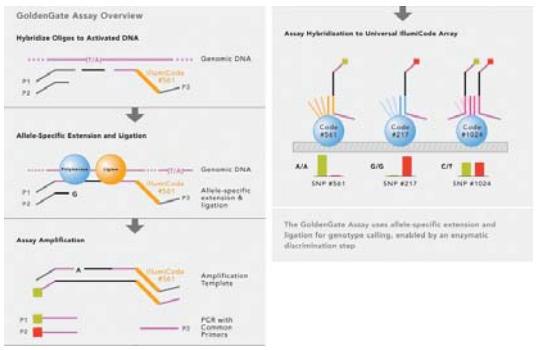
- dbSNP
  - Central repository for SNPs
  - Initial SNPs identified with PolyBayes
  - dbSNP build 138, human genome build 37.5 – 233M submissions at 63M loci
  - High false positive rate?
- HapMap
  - Database of haplotypes and 'tag' SNPs which identify them
  - Samples from 270 people from Nigeria, Japan, China, USA (of North and West European decent)

What are SNPs used for?

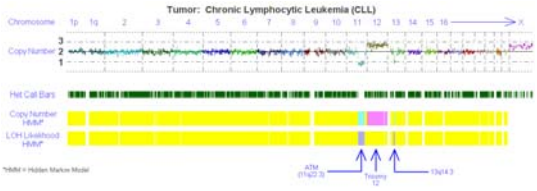
- Association studies: SNPs as markers to identify regions associated with a phenotype
- Study variation in human populations
- Evolutionary analyses
- To infer disease susceptibility
- To infer drug resistance

SNP Genotyping

- PCR - Taqman assays
- Bead-based - Illumina
- Arrays – Affymetrix SNP chips
- Mass Spectrometry - Sequenom



Virtual karyotype - SNP arrays



References

- Kwok *et al*. - Detection of single nucleotide polymorphisms – Curr. Issues Mol. Bio. 2003
- Mitchell *et al*. - Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. – Bioinformatics 2004

Practical 5:

- Categorise a number of SNPs from the bacterium *Streptococcus pneumoniae* TIGR4.
- SNPs produce by alignment of reads to reference genome using MAQ.