Bioinformatics for SNP Discovery

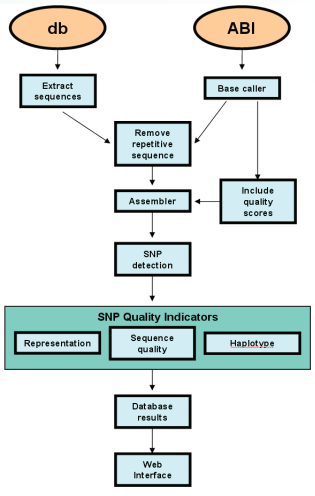
Linda Milne
SCRI, Invergowrie, Dundee DD2 5DA.

Single nucleotide polymorphisms are an important tool in the development of molecular markers for important genes, traits and biodiversity in crop plants.

We have designed a SNP discovery pipeline which makes use of sequence data from the major DNA repositories and from the re-sequencing of target genes.

The starting point for data-mining SNPs is an alignment of homologous sequences from a variety of cultivars. To produce alignments we can use assembly software to produce a large-scale sequence assembly. Each unigene in the assembly can be then screened for potential SNPs.

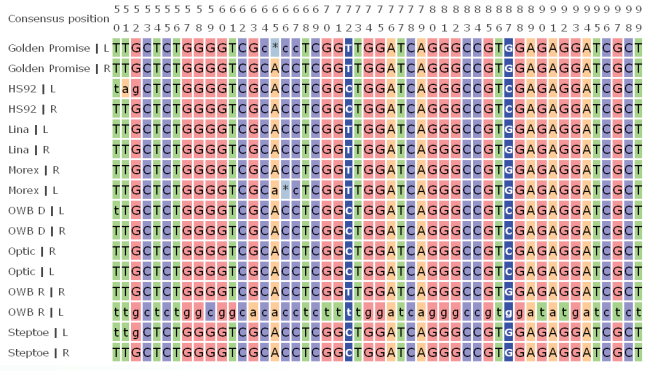
With the introduction of Next Generation sequencing we are currently reviewing our SNP pipeline and evaluating new de-novo and template-based assembly software, such as Velvet, MAQ and MOSAIC.



Right: A typical workflow for SNP discovery.

SNP Discovery

From a sequence alignment we can scan along seeking locations where there is a base that does not match the consensus of the alignment, which gives us a list of putative SNPs. From there, various approaches can be used to help verify that a putative SNP is not a read-error. To help avoid SNPs that are false positives there should be more than one example of the alternative base at that location, especially if there is no quality score data available. Another clue that helps verify a true SNP from background noise is if the SNP in question appears to occur in concert with a SNP at another location, giving a consistently different haplotype. We are currently evaluating the read-error bias of new sequencing technologies and our verification process is also being adapted to incorporate these new features.



Aligned re-sequenced alleles of a barley gene, showing SNPs changing in concert at two sites.

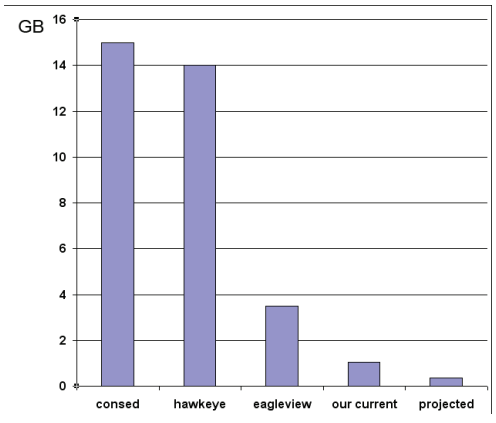
An example of a putative SNP in a barley gene, showing how poorer quality base calls (greyscale) can be used to avoid false positives.

Next Generation Sequencing

The advent of next generation sequencing technologies has presented bioinformatics with a series of challenges. Not only does Next Generation sequencing require novel assembly methods and consideration of different read-error bias, the sheer volume of data means we need new tools for visualizing assembly data. We are addressing the difficulties of displaying millions of reads and manipulating it in real-time.

We are in the process of piloting new methods of storing and manipulating large assembly datasets, and already we have substantially reduced the memory requirement compared to other assembly viewers.

For a dataset of 7 million 32bp reads (a 1.4GB file), we have reduced the memory footprint from over 3.5GB in Eagleview (the most efficient viewer) to just 1GB, and we project we can reduce this even further to 350MB.



Assemblies can contain millions of reads requiring gigabytes of RAM which means computational optimisation of such data visualizations are paramount.