Chapter 3
WHAT ARE SNPs?

David Edwards[1], John W. Forster[1], David Chagné[2], and Jacqueline Batley[1]

3.1 SNP DEFINITION

In the simplest form, a single nucleotide polymorphism (SNP) is an individual nucleotide base difference between two DNA sequences. SNPs can be categorised according to nucleotide substitution as either transitions (C/T or G/A) or transversions (C/G, A/T, C/A, or T/G). As a nucleotide base is the smallest unit of inheritance, SNPs provide the ultimate form of molecular genetic marker. They also represent the most frequent type of genetic polymorphism, and the potential number of such markers is enormous in comparison with any but the most closely related genotypes within a given species (Rafalski 2002a,b). Sequence variation can have a major impact on how the organism develops and responds to the environment. Furthermore they are evolutionarily stable, not changing significantly from generation to generation (Lopez *et al*. 2005). SNPs provide an important source of molecular markers that are useful in genetic mapping, map-based positional cloning, detection of marker-trait gene associations through linkage and linkage disequilibrium (LD) mapping and the assessment of genetic relationships between individuals. The low mutation rate of SNPs makes them excellent markers for studying complex genetic traits and as a tool for the understanding of genome evolution (Syvanen 2001).

SNPs, at any particular site, could in principle involve four different nucleotide variants, but in practice they are generally biallelic. However, this disadvantage, when compared to multiallelic markers such as SSRs, is compensated by the relative abundance of SNPs. In humans, for a variation to be considered a true SNP, it must occur in at least 1% of the population. SNPs are suitable for automated discovery and detection, and can be applied to a wide range of purposes, including rapid identification of crop cultivars,

[1] Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia
[2] HortResearch, Plant Gene Mapping group, Private Bag 11030, Palmerston North, New Zealand

construction of ultra-high density genetic maps and association with genetic disorders (in humans and livestock) and agronomic traits (in livestock and crop plants).

SNPs provide an abundant source of DNA polymorphism in a number of eukaryote species. Information on the frequency, nature and distribution of SNPs in the majority of plant genomes is limited. However, the level of development and application of SNPs in higher plants, including some crop and tree species, is increasing, and consequently they provide an attractive marker system to plant breeders and geneticists. With the increasing availability of public sequence data and the rapid discovery of SNPs in plants, the development and application of SNP markers will continue to accelerate.

3.2 SNP FREQUENCY

SNPs can differentiate between related sequences both within an individual and between individuals in a population. In diploid species, in which an individual is heterozygous at a genetic locus, there are two homologous gene copies that may be differentiated by SNPs. The inheritance of each variant may be directly measured in the progeny. Detection of SNPs in individuals becomes complicated in the presence of gene or genome duplication. In these instances, it is often difficult to differentiate between homoeologous (between genome) and paralogous (within genome) duplication of genetic loci without detailed genetic inheritance studies. Because the majority of DNA in individuals within a related population is the same, genetic differences between individuals can be defined by SNPs. The frequency of SNPs (nucleotide diversity) and the haplotypic diversity (heterozygosity) between two individuals or within a population are direct measures of genetic diversity. Under conditions of forced inbreeding, such as recurrent backcrossing to parental individuals, sib-mating or mating between individuals with lower-degree relatedness, reduced genetic diversity and SNP frequency is observed. Such conditions may have arisen due to population reduction or isolation in natural populations (the so-called 'founder effect'). For domesticated crop plants, narrow genetic bases have contributed to corresponding reduced genetic diversity at the nucleotide level.

The frequency and nature of SNPs in plants is beginning to receive considerable attention. A number of reports in *Arabidopsis thaliana*, rice and maize have provided estimates of sequence diversity in these species. In many species, the analysis of DNA sequence variation has been confined to single genes or DNA fragments with the goal of defining gene structure, function or evolutionary relationships. It is known that SNPs are widely distributed throughout genomes, although various studies show that the occurrence and distribution of SNPs differs between species, in particular between inbreeding and outbreeding species, or in those species with a narrow genetic base. It is generally well accepted that some species, for example maize, are highly polymorphic, whilst others, such as soybean and melon, are less polymorphic. Detailed studies of sequence diversity have now been performed at selected loci for a range of plant species and in plants, the typical frequencies are in the range of 1 SNP every 100–300 bp.

The most advanced SNP studies in plants have been performed on model species where a large quantity of genomic or EST sequence is available. SNPs have been detected using high-throughput analysis in *A. thaliana* (Cho *et al*. 1999). ESTs are a good resource for SNP discovery and they have been used for SNP discovery in sugarbeet (Schneider *et al.* 2001), maize (Ching *et al.* 2002; Batley *et al.* 2003), rice (Nasu *et al.* 2002), soybean (Zhu *et al*. 2003) and sugarcane (Grivet *et al.* 2003). In soybean 280

SNPs from 143 amplicons (76.3 kb) have been identified (Zhu *et al.* 2003). In maize, one non-coding SNP/31 bp and one coding SNP/124 bp has been reported for 18 maize genes in 36 inbred lines.

A genome-wide polymorphism database of rice has been constructed defining polymorphisms between the cultivars Nipponbare (from sub-species *japonica*) and 93-11 (from sub-species *indica*) (Shen *et al.* 2004). The database contains 1,703,176 SNPs and 479,406 insertions/deletions (indels) (see Section 3.6 for further discussion on indels). This equates to approximately 1 SNP/268 bp in the rice genome. A similar study was also performed by Feltus *et al.* (2004). After aligning drafts of rice *indica* and *japonica* sequence and filtering to remove multiple copy and low-quality sequence, 384,341 candidate interspecific SNPs were identified, at a frequency of approximately 1.7 SNPs/kb. Due to the stringent filtering process, this is probably an underestimate of the real SNP frequency in rice. This work was performed again in 2005 (Yu *et al*. 2005) using alignments of the improved whole-genome shotgun sequences for *japonica* and *indica* rice. SNP frequencies varied from 3 SNPs/kb in coding sequence to 27.6 SNPs/kb in the transposable elements, with a genome wide measure of 15.13 SNPs/kb, or 1 SNP per 66 bp.

Further studies in rice have involved SNP discovery and characterisation in the *Piz* and *Piz-t* regions (Hayashi *et al.* 2004). The frequency was found to be similar to the previous studies, with an SNP found every 248 bp (Hayashi *et al.* 2004). The SNP frequency varied slightly depending on the cultivars being assessed. On average, 1 SNP was detected every 390 bp between cultivars Nipponbare and Zenith and 1 SNP per 173 bp between cultivars Nipponbare and Toride 1. The SNP frequency was higher between Zenith and Toride 1, with an SNP on average every 140 bp. In earlier studies, Yu *et al*. (2002) compared sequences from *japonica* and *indica* cultivars and found an average of 1 SNP every 170 bp, while Nasu (2002) reported a similar frequency for rice SNPs.

Extensive research has been performed on SNP frequency in barley. Russell *et al.* (2004) examined the frequency and distribution of SNPs within 23 genes associated with grain germination in barley in a range of accessions including European cultivars, landraces and wild barley. The frequency of SNPs was found to be 1 SNP every 78 bp. In a further study, the *Isa* (inhibitor of α-amylase) gene was sequenced in 16 barley genotypes to detect sequence polymorphisms (Bundock and Henry 2004). A total of 80 SNPs were identified in the 2,164 bp sequence, containing the *Isa* promoter, transcript and 3′-untranslated region (UTR), giving a high frequency of 1 SNP/27 bp. Kota *et al.* (2001) identified 72 polymorphic SNPs in seven genotypes of barley. The frequency of SNPs was estimated to be 1 every 240 bp. This was calculated from 52,140 bp of sequence from each genotype analysed. Similar studies have been performed in other crop species such as *Beta vulgaris* and *Zea mays*, for which the relevant frequencies were 1 every 60–130 bp and 104 bp, respectively (Schneider *et al.* 2001; Ching *et al.* 2002; Tenaillon *et al.* 2001). As expected, the frequency of SNPs in inbreeding species such as barley is lower than that observed in outbreeding species. This is further demonstrated in poplar, an out-breeding tree species, which exhibits a high level of variation. Cronk (2005) determined the presence of an SNP every 100 bp in poplar, increasing to 1 every 50 bp when geographically diverse species were included in the study.

In a study of 25 diverse genotypes of soybean (Zhu *et al.* 2003), a total of 280 SNPs were identified in 143 amplicons, totalling 76.3 kb sequence, providing 1 SNP per 273 bp. It was found that nucleotide diversity was lower in soybean than maize or *A. thaliana*, and this may be due to inbreeding. However, as *A. thaliana* is also self-pollinating, this does not explain all the findings. These results may also be due to

the narrow genetic base of soybean. SNP discovery has also been performed in lesser-known crops. Based on EST sequence information, fragments of 34 genes were amplified from five diverse quinoa (*Chenopodium quinoa* Willd.) accessions and the related weed species *C. berlandieri* and sequenced (Coles *et al.* 2005). Analysis of the quinoa EST sequences revealed a total of 51 polymorphisms in 20 EST sequences, including 38 SNPs and 13 indels. This was an average of 1 SNP every 462 bp, which increased to 1 SNP every 179 bp when *C. berlandieri* was included in the analysis. This SNP frequency is lower than that observed in barley (1/189 bp), maize (1/104 bp) and sugarbeet (1/130 bp), but similar to levels observed in soybean (1/503 bp) and *A. thaliana* (1/336 bp). Although the sample size was small, the SNP frequency reflects the narrow genetic base for cultivated quinoa.

Lopez *et al.* (2005) exploited a recently developed EST collection to identify SNPs in five cultivars of cassava (*Manihot esculenta* Crantz). One SNP per 905 bp was detected in intra-cultivar comparisons and 1 SNP per 1,032 bp was detected in inter-cultivar comparisons, based on data from 111 contigs, with an overall value of 1 SNP every 509 bp. This study also obtained further information on SNP frequency in six cultivars from 33 amplicons from 3′-EST and BAC end sequences. A total of 11 kb of sequence was obtained for each cultivar, with 186 SNPs being identified. Of these, 146 were observed within cultivars and 80 were observed between cultivars. The total frequency of SNPs was found to be one per 62 bp, a value similar to that observed for other crops. The intra-cultivar variation may be due to the presence of background heterozygosity and inbreeding depression within the lines. Cassava is also an ancient polyploid and predicted SNPs may be due to the presence of paralogous comparisons between members of multi-gene families.

In potato, 277 SNPs were identified between two alleles of the urease gene, with an average of 2.5 SNPs per 100 bp (Wittle *et al.* 2005). This average frequency of 1 SNP per 40 bp is relatively high for comparison between two alleles of a single copy gene. This is also reflected by studies of SNP variation in resistance gene analogues (RGAs) of cultivated potato, as described in Chapter 4.

3.3 SNP DISTRIBUTION

DNA is inherited in long stretches or blocks that are only separated by recombination events at meiosis. Because of this, groups of SNPs that are located in physical proximity to each other on the same stretch of DNA tend to be inherited together as a single linked group. A haplotype can be defined as a contiguous DNA sequence of an organism and may extend over physical distances characteristic of genes, gene clusters, chromosome segments, whole chromosomes or, in the case of asexual lineages, whole genomes. SNPs may be considered to define a haplotype, in that they are a series of DNA polymorphisms that differentiate between DNA sequences. As groups of SNPs that are in physical proximity tend to be inherited together (due to reduced capacity for genetic recombination and defining the extent of LD), haplotypes segregating in populations may be identified through the interrogation of one or a small number of diagnostic SNP loci.

The frequency of SNPs varies within each genome. Currently available data shows that the distribution of polymorphic sites is not random across the nuclear genome, or within a gene. SNPs can occur in coding and non-coding regions of the genome and at

different frequencies in different genomic regions. This uneven distribution may be due to differences in recombination rate, gene density, transmission pattern, selection strength and compositional pressure. Genomic regions with low recombination rates generally have reduced levels of polymorphisms (Rafalski and Morgante 2004). Regions subject to strong balancing selection (i.e. two or more alleles or haplotypes are maintained), such as those containing disease resistance genes, show the greatest diversity (Kuang *et al.* 2004).

The local abundance of SNPs within the genome varies due to a combination of the mutation rate that generates new polymorphisms and any positive or negative selection for regions linked to these mutations. SNP generation *de novo* may be more frequent outside of transcribed genic regions as these regions tend to exhibit greater levels of 5-methylcytosine (^{5me}C) abundance, an important factor in the generation of the most abundant C to T mutation due to deamination of ^{5me}C (which is aminothymidine) to T over evolutionary time. The majority of SNPs would be expected to be evolutionary neutral, that is, they would be neither selected for nor against, and their abundance in a population would vary due to random genetic drift. Rare deleterious mutations are counter-selected at a rate characteristic of the specific fitness penalty. For example, SNPs or Indels in transcribed sequences that lead to the production of altered proteins are relatively infrequent in populations when compared to similar polymorphisms within intron or untranscribed sequence. Selection, either natural or through breeding would lead to the removal of deleterious sequences from the population and increase the abundance of beneficial sequences. Selective pressure would apply to sequences in proximity to the selected sequence (the so-called 'hitch-hiking' phenomenon) unless they are separated by recombination during meiosis. Thus, strong selective pressure is likely to lead to genomic regions with reduced genetic diversity and fewer SNPs. This hypothesis is supported by the observation that in most organisms studied to date, SNPs are more prevalent in the non-coding regions of the genome. These mutations should theoretically only affect the phenotype if they cause a change in the regulation of gene expression, changing the expression pattern of surrounding transcribed regions. Within the coding regions, an SNP is either non-synonymous and results in an amino acid change, or is synonymous and does not alter the amino acid sequence and therefore is neutral. Non-synonymous SNPs may also be radical or conservative in nature, depending on transitions between positively charged, uncharged and negatively charged amino acid side-groups. Synonymous change may, however, potentially modify an RNA splice processing site resulting in phenotypic changes. SNPs have become popular tools for identifying genetic loci that contribute to phenotypic variation based on LD (see Chapters 2 and 7 for further discussion on the principles of LD).

The distribution of SNPs across the genome has been studied in a variety of plant species. Perhaps the most comprehensive study is in *A. thaliana*, where over 37,000 SNPs were identified by comparing partial genome sequence from the *Ler* accession with the near complete sequence of Col-0 (Schmid *et al.* 2003). The distribution of SNPs was found to be even across the five chromosomes, with the exception of centromeric regions, which contain few transcribed genes. In the ESTs studied, a total of 4,327 SNPs were identified. Analysis of amplicons derived from sequence tagged sites (STSs), corresponding to 4,955 consensus sequences revealed 3,773 SNPs. Of these, 2,922 (77%) were in non-coding regions of the genome. In the EST-derived SNPs, there was an average of 1 SNP per 336 bp. There was a higher ratio of synonymous to non-synonymous polymorphisms in EST compared to STS data, supporting the concept that expressed genes are more constrained by sequence evolution than randomly selected genomic loci.

A decreased frequency of SNPs in coding regions was also observed in quinoa (Coles *et al.* 2005). One SNP per 2,614 bp was observed in coding sequence, which increased to 1 SNP per 697 bp if the closely related weed species *C. berlandieri* was included in the analysis. The frequency of SNPs was much higher in the non-coding sequence, with an SNP per 385 bp, increasing to 1 every 144 bp in comparison to *C. berlandieri*. Of the SNPs in coding sequence, one was synonymous and three were non-synonymous. A detailed sequence analysis of four SSCP-SNP loci, over a panel of eight inbred pearl millet genotypes, revealed one SNP every 59 bp in introns, but considerably fewer in exons (Bertin *et al.* 2005). An elevated SNP frequency in non-coding sequence was also observed in maize, with 1 per 31 bp in non-coding regions and 1 per 124 bp in coding sequence (Ching *et al.* 2002). Five of the 18 SNPs in coding sequence were non-synonymous. In a study of SNP distribution in melon, 75% of the polymorphisms were located in introns and 3′-UTRs (Morales *et al.* 2004). Eleven SNPs (32%) were found in coding regions and the remaining 23 (68%) were found in 3′-UTR or intronic sequence. Seven of the eleven SNPs in coding sequence gave rise to synonymous changes. The proportion of synonymous compared to non-synonymous SNPs was also comparable with observations in maize, where 72% of SNPs in coding regions were synonymous (Ching *et al.* 2002). The higher presence of SNPs in non-coding regions has also been demonstrated in soybean (Zhu *et al.* 2003). These results suggest that UTRs and introns should be preferentially targeted for SNP discovery in candidate genes.

In a study of 25 diverse genotypes of soybean, 51 SNPs were identified in 28.7 kb coding sequence. Of these, 25 were synonymous and 26 were non-synonymous (Zhu *et al.* 2003). The rate of synonymous to non-synonymous base changes was lower in soybean than in maize, although similar to that seen in *A. thaliana*. Low diversity at non-synonymous sites is the result of selection against deleterious mutations. Out-crossing species are generally more effective at removing deleterious mutations as a consequence of large effective population size. Soybean and *A. thaliana* both exhibit low ratios of synonymous to non-synonymous mutation, suggesting the presence of a relatively high level of slightly deleterious mutations. The SNP distribution also varies in rice. Yu *et al*. (2005) aligned the improved whole-genome shotgun sequences for *japonica* and *indica* and found that SNP rates varied from 3 SNPs per kb in coding sequence to 27.6 SNPs per kb in the transposable elements. Furthermore, there were 4.72 SNPs per kb in the 5′UTR, 6.07 SNPs per kb in introns and 4.5 per kb in the 3′-UTR.

The *Isa* gene, which is significant for control of α-amylase activity during germination, was sequenced in 16 barley genotypes to detect sequence polymorphisms (Bundock and Henry 2004). A total of 80 SNPs and 23 indels were identified in 2,164 bp of sequence containing the *Isa* promoter, transcript and 3′-UTR. The frequency of SNPs was greatest in the 3′-non-translated region, downstream of the gene (1 SNP/16 bp), due to the contribution of comparison with sequences derived from wild barley (*Hordeum spontaneum*). One SNP per 75 bp was observed in the transcribed region, with 10 SNPs in the coding sequence, none in the 5′-UTR and 1 in the 3′-UTR. The region flanking the SSR in the promoter was highly polymorphic, with twice the number of SNPs expected given the overall frequency observed. This high frequency of SNPs surrounding SSRs has also been observed in maize (Mogg *et al.* 2002).

Two cultivars of soybean were distinguished by a non-synonymous transversion within the *GmNARK* (*Glycine max* nodule autoregulation receptor kinase) gene. Further sequence variants, including an indel and 5 SNPs, were detected in the intron and 5′-UTR respectively. There were a further 6 SNPs in the exons, all of which were synonymous (Kim *et al.* 2005). In the *Piz* and *Piz-t* regions of rice associated with rice blast resistance (Hayashi *et al.* 2004), SNPs were found every 248 bp (Hayashi *et al*. 2004).

Nucleotide polymorphism in the gene encoding phenylalanine ammonia-lyase (*Pal1*) of Scots pine (*Pinus sylvestris*) was studied by Dvornyk *et al.* (2002). A 2,045 bp exonic fragment of *Pal1* was sequenced in five megagametophytes from different individuals belonging to four populations, from Finland, Russia and Spain. Twelve polymorphisms were identified, and two alleles from a further 11 loci were studied (4,606 bp). Nine of the polymorphisms were synonymous and there were no introns in the sequence studied.

3.4 TRANSITIONS OR TRANSVERSIONS

SNPs are produced by mutations. The mutation frequency between any two nucleotides is not random but is dependent on the nucleotide base, the base sequence in its immediate proximity and the methylation status of the DNA. A major mechanism of spontaneous mutation is due to errors in DNA replication. Nucleotide bases in DNA can exist in two different structural forms (tautomers) called KETO and ENOL forms, but are predominantly found in the KETO form. Shifts to the ENOL form (tautomerisation) can alter pairing preferences, such that A may pair with C rather than T. Reversion of the tautomeric shift following DNA replication leads to fixation of a base mutation. The predicted average frequency of such processes is c. 1 per 10^4 bp copied, but the influence of fidelity maintenance systems such as polymerase proof-reading and post-replication mismatch repair results in observed frequencies of c. 1 in 10^{10} bp copied, corresponding to c. 1 in 10^6 per gene across a broad range of organisms.

Transitions are the most common form of SNP (Garg *et al.* 1999; Picoult-Newberg *et al*. 1999; Deutsch *et al.* 2001; Batley *et al.* 2003) reflecting the high frequency of the C to T mutation following deamination of methylated cytosine residues (Coulondre *et al.* 1978). C/T transitions constitute 67% of the SNPs observed in humans. Other variations in base substitution abundance are observed, but the underlying mechanisms for these differences remain to be explained (Batley *et al*. 2003).

Lopez *et al.* (2005) observed a significantly higher number of transitions than transversions in intra-cultivar (64% transitions) and inter-cultivar (65% transitions) comparisons in cassava. However, Coles *et al.* (2005) found an approximate 1:1 transition:transversion ratio in quinoa. A total of 20 transitions and 18 transversions were identified, increasing to 61 and 45, respectively, if the closely related weed species, *C. berlandieri,* was included in the analysis. This ratio was similar to those observed in maize, soybean (Zhu *et al.* 2003) and *A. thaliana*, but lower than the 2:1 ratios observed in sugarbeet, melon (Morales *et al.* 2004) and barley (Soleimani *et al.* 2003). The higher-than-expected C/T transition rate is likely to be due to the methylation effects described previously. Hayashi *et al.* (2004) found that 72–75% of SNPs between *indica* and *japonica* rice cultivars were transitions. This finding was supported by Feltus *et al.* (2004) who aligned drafts of the rice subspecies *japonica* and *indica* sequence and found that 65.8% SNPs were transitions and 34.2% were transversions. The high frequency of

transitions in this study is also compatible with the consequences of epigenetic modification of CG nucleotide motifs by DNA methylation in rice.

3.5 SNPs ASSOCIATED WITH ECONOMICALLY IMPORTANT GENES

As more genomes are being completely sequenced, interest is re-focusing on the discovery and analysis of intra-specific differences. SNPs can be used as simple genetic markers which may be identified in the vicinity of virtually every gene. There is potential for the use of SNPs to detect associations between the allelic forms of a gene and phenotypes, especially for common diseases in humans (see Chapter 2). SNPs have been identified in a number of plant genes of economic value. For example, SNPs were identified discriminating allelic variants of the potato urease gene in cultivar Desiree (Wittle *et al.* 2005).

SNPs associated with functional genes are candidate qualitative or quantitative trait nucleotides (QTNs) that are causally associated with the phenotypic effects of different alleles. However, the determination of QTNs is an intensive process which involves the use of data from methods such as induced mutagenesis, protein modelling and *in vitro* RNA and protein synthesis studies, as well as genetic analysis. For species without extensive LD, association studies can potentially be used to obtain very high map resolution, to the level of the QTN. Tree species such as Norway spruce (*Picea abies*), for which LD declines to minimal levels over very short distances (c. 50–100 bp) within genes (M. Morgante, pers. commun.), or the pine species *Pinus taeda*, for which the equivalent value is c. 1,500 bp, may be amenable to this form of analysis (Neale and Savolainen 2004). Species with moderate or high LD do not offer these advantages, and hence SNP haplotypes over the length of genes or gene clusters must be used to provide diagnostic tests for superior allele content.

3.6 INDELS

Small insertion or deletion events (indel for insertion/deletion) are another common form of genetic mutation. These mutations may be detected as SNPs as the insertion or deletion of nucleotides changes the sequence read. Indels may be produced by errors in DNA synthesis, repair or recombination, or may be due to the insertion and excision of transposable elements that often leave a characteristic DNA footprint of several nucleotide bases. For example, the relative abundance of eight base indels observed in maize by Bhattramakki *et al.* (2002) may be due to sequence duplication during insertion and excision of *Ac/Ds* transposable elements (Sutton *et al.* 1984).

Tenaillon *et al.* (2002) studied SNPs and indels located in previously published sequences from 21 loci on maize chromosome 1. Small indels (1–5 bp) were frequent, 56% of the indels being 1–2 bp in length and 92% were less than 20 bp in length. Furthermore, 5 of the 21 indels longer than 20 bp were found to be previously characterised Miniature Inverted-repeat Transposable Elements (MITEs). A total of 263 indels were observed in 17/21 loci. Indel size ranged from 1 to 640 bp, and the number per locus ranged from 2 to 59. This frequency of small indels was also observed in the *Piz* and *Piz-t* regions of rice. Of the 52 indels identified, 42 (81%) were 1–5 bp in length and only 4 were longer than 40 bp (Hayashi *et al.* 2004).

In a study of the urease gene in potato, 40 indels were observed within non-coding regions, of which 70% were 1–4 bp in length, 20% 5–10 bp and 10% (4 indels) were greater than 10 bp. The instances of these long indels may be explained by the relevant sequence features. One insertion was found to be due to a retrotransposon. A 30 bp indel is found in an array of 30 bp repeats within an intron and may have been caused by unequal cross-over, while a 34 bp indel is present in an SSR-containing region, which are known to undergo expansion and contraction (Wittle *et al.* 2005).

Morales *et al.* (2004) searched for indels in 34 ESTs between two distantly related melon genotypes. On average 1 indel was found per 1,666 bp. No indel was found inside the coding region. The indel length ranged from 1 to 13 bp, with single bp indels being the most frequent. This indel frequency was higher than in the total *A. thaliana* genome, in which one indel per 6.6 kb was observed (Jander *et al.* 2002). However, these data are not directly comparable to the melon study as both coding and non-coding regions were used in the *A. thaliana* study. Ching *et al.* (2002) examined the frequency and distributions of polymorphisms at 18 maize genes in 36 maize inbreds. Indels were found to be frequent in non-coding regions (1/85 bp) but rare in coding sequences.

In the genome wide polymorphism database of rice, using cultivars Nipponbare (*japonica*) and 93-11 (*indica*) (Shen *et al.* 2004), 479,406 indels were detected. This corresponds to approximately 1 indel per 953 bp in the rice genome. This indel frequency is higher than that observed in a similar study of the rice subspecies *indica* and *japonica* sequence by Feltus *et al.* (2004), who found approx. 0.11 indels/kb. However, due to the stringent sequence filtering performed in this later study, the result probably underestimates indel frequency in rice.

A total of 23 indels were identified between 16 barley genotypes in the 2,164 bp of *Isa* gene sequence (Bundock and Henry 2004), a measure of 1 indel per 94 bp. Four of these indels were within a microsatellite region and were excluded. Of the remaining 19 indels, 9 were 1 bp in length and the others ranged from 4 to 306 bp, giving an average frequency of 1 indel per 114 bp.

3.7 CONCLUDING REMARKS

SNPs are individual nucleotide base differences between DNA sequences and can represent differences between individuals or within populations. The specific base difference is determined by the cause of mutation and is non-random, with C to T transitions being the most frequent form. Insertion/deletion events (indels) are a special form of SNP caused by the addition or removal of DNA sequence, resulting in both length and sequence polymorphisms. The frequency of SNPs is dependent on both their generation and selection in populations. SNPs are generally evolutionally neutral, with frequencies varying due to random genetic drift. Some SNPs, particularly those associated with expressed genes, may be under positive or negative evolutionary selection pressure and will be maintained or rapidly removed from populations (Przeworski 2002; Bamshad and Wooding 2003). SNPs not separated by recombination at meiosis and thus in LD with other SNPs will be inherited as a linkage block and thus maintained at a frequency determined by the cumulative selection pressure of the haplotypic group. SNPs and indels are valuable molecular genetic markers due to both their abundance and relative stability in the genome, and can be applied as perfect molecular markers when identified within genes underlying observed traits.

3.8 REFERENCES

Bamshad, M., Wooding, S.P., 2003, Signatures of natural selection in the human genome. Nat. Rev. Genet. 4: 99–111.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., Edwards, D., 2003, Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol. 132: 84–91.

Bertin, I., Zhu, J.H., Gale, M.D., 2005, SSCP-SNP in pearl millet – a new marker system for comparative genetics. Theor. Appl. Genet. 110: 1467–1472.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J.C. III, Tingey, S.V., Rafalski, A., 2002, Insertion–deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol. Biol. 48: 539–547.

Bundock, P.C., Henry, R.J., 2004, Single nucleotide polymorphism, haplotype diversity and recombination in the *Isa* gene of barley. Theor. Appl. Genet. 109: 543–551.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M., Rafalski, A.J., 2002, SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3: 1–14.

Cho, R.J., Mindrinos, M., Richards, D.R., Sapolsky, R.J., Anderson, M., Drenkard, E., Dewdney, J., Reuber, T.L., Stammers, M., Federspiel, N., Theologis, A., Yang, W.H., Hubbell, E., Au, M., Chung, E.Y., Lashkari, D., Lemieux, B., Dean, C., Lipshutz, R.J., Ausubel, F.M., Davis, R.W., Oefner, P.J., 1999, Genome wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat. Genet. 23: 203–207.

Coles, N.D., Coleman, C.E., Christensen, S.A., Jellen, E.N., Stevens, M.R., Bonifacio, A., Rojas-Beltran, J.A., Fairbanks, D.J., Maughan, P.J., 2005, Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Sci. 168: 439–447.

Coulondre, C., Miller, J.H., Farabaugh, P.J., Gilbert, W., 1978, Molecular basis of base substitution hot spots in *Escherichia coli*. Nature 274**:** 775–780.

Cronk, Q.C.B., 2005, Plant eco-devo: the potential of poplar as a model organism. New Phytol. 166: 39–48.

Deutsch, S., Iseli, C., Bucher, P., Antonarakis, S.E., Scott, H.S., 2001, A cSNP map and database for human chromosome 21. Genome Res. 11**:** 300–307.

Dvornyk, V., Sirviö, A., Mikkonen, M., Savolainen, O., 2002, Low nucleotide diversity at the *pal*1 locus in the widely distributed *Pinus sylvestris*. Mol. Biol. Evol. 19: 179–188.

Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N., Paterson, A.H., 2004, An SNP resource for rice genetics and breeding based on subspecies *Indica* and *Japonica* genome alignments. Genome Res. 14: 1812–1819.

Garg, K., Green, P., Nickerson, D.A., 1999, Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res. 9**:** 1087–1092.

Grivet, L., Glaszmann, J.-C., Vincentz, M., da Silva, F., Arruda, P., 2003, ESTs as a source for sequence polymorphism discovery in sugarcane: example of *Adh* genes. Theor. Appl. Genet. 106: 190–197.

Hayashi, K., Hashimoto, N., Daigen, M., Ashikawa, I., 2004, Development of PCR-based SNP markers for rice blast resistance genes at the *Piz* locus. Theor. Appl. Genet. 108: 1212–1220.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M., Last, R.L., 2002, *Arabidopsis* map based cloning in the post genome era. Plant Physiol. 129: 440–450.

Kim, M.Y., Van, K., Lestari, P., Moon, J.-K., Lee, S.-H., 2005, SNP identification and SNAP marker development for a GmNARK gene controlling supernodulation in soybean. Theor. Appl. Genet. 110: 1003–1010.

Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J., Graner, A., 2001, Generation and compairson of EST derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas 135: 145–151.

Kuang, H., Woo, S.-S., Meyers, B.C., Nevo, E., Michelmore, R.W., 2004, Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell 16: 2870–2894.

Lopez, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J., Verdier, V., 2005, Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). Theor. Appl. Genet. 110: 425–431.

Mogg, R., Batley, J., Hanley, S., Edwards, D., O'Sullivan, H., Edwards, K.J., 2002, Characterisation of the flanking regions of *Zea Mays* microsatellites reveals a large number of useful sequence polymorphisms. Theor. Appl. Genet. 105: 532–543.

Morales, M., Roig, E., Monforte, A.J., Arús, P., Garcia-Mas, J., 2004, Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). Genome 47: 352–360.

Nasu, S., Suzuki, J., Ohta, R., Hasegawa, K., Yui, R., Kitazawa, N., Monna, L., Minobe, Y., 2002, Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa, Oryza rufipogon*) and establishment of SNP markers. DNA Research 9: 163–171.

Neale, D.B., Savolainen, O., 2004, Association genetics of complex traits in conifers. Trends Plant Sci. 9: 325–330.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M., 1999, Mining SNPs from EST databases. Genome Res. 9: 167–174.

Przeworski, M., 2002, The signature of positive selection at randomly chosen loci. Genetics 160: 1179–1189.

Rafalski, J.A., 2002a, Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci. 162: 329–333.

Rafalski, J.A., 2002b, Applications of single nucleotide polymorphisms in crop genetics. Current Opin. Plant Biol. 5: 94–100.

Rafalski, A., Morgante, M., 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20: 103–111.

Russell, J., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G., Powell, W., 2004, A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome 47: 389–398.

Schmid, K.J., Rosleff Sörensen, T., Stracke, R., Törjék, O., Altmann, T., Mitchell-Olds, T., Weisshaar, B., 2003, Large-scale identification and analysis of genome wide single nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res. 13: 1250–1257.

Schneider, K., Weisshaar, B., Borchardt, D.C., Salamini, F., 2001, SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. Mol. Breed. 8: 63–74.

Shen, Y.-J., Jiang, H., Jin, J.-P., Zhang, Z.-B., Xi, B., He, Y.-Y., Wang, G., Wang, C., Qian, L., Li, X., Yu, Q.-B., Liu, H.-J., Chen, D.-H., Gao, J.-H., Huang, H., Shi, T.-L., Yang, Z.-N., 2004, Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. 135: 1198–1205.

Soleimani, V.D., Baum, B.R., Johnson, D.A., 2003, Efficient validation of single nucleotide polymorphisms in plants by allele specific PCR, with an example from barley. Plant Mol. Biol. Rep. 21: 281–288.

Sutton, W.D., Gerlach, W.L., Schwartz, D., Peacock, W.J., 1984, Molecular analysis of *Ds* controlling element mutations at the Adh1 locus of maize. Science 223: 1265–1268.

Syvanen, A.C., 2001, Genotyping single nucleotide polymorphisms. Nat. Rev. Genet. 2: 930–942.

Tenaillon, M.I., Sawkins, M.C., Anderson, L.K., Stack, S.M., Doebley, J., Gaut, B.S., 2002, Patterns of diversity and recombination along Chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics 162: 1401–1413.

Wittle, C.-P., Tiller, S., Isidore, E., Davies, H.V., Taylor, M.A., 2005, Analysis of two alleles of the urease gene from potato: polymorphisms, expression and extensive alternative splicing of the corresponding mRNA. J. Exp. Botany 56: 91–99.

Yu, J., Hu, S., Wang, J., Wong, G.K., Songgang Li, S., Liu, B., Deng, Y., Dai, L., Zhou,Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Xiangang Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L. Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Wei Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., Yang, H., 2002, A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296: 79–92.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Wang, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Wang, J., Wang, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Liu, J., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Zhang, J., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Zhang, Z., Zhang, Y., Huang, X., Su, Z., Tong, W., Li, J., Tong, Z., Li, S., Ye, J., Wang, L., Fang, L., Lei, T., Chen, C., Chen, H., Xu, Z., Li, H., Huang, H., Zhang, F., Xu, H., Li, N., Zhao, C., Li, S., Dong, L., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Zhang, B., Hu, W., Zhang, Y., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wang, J., Wong, G.K.-S., Yang, H., 2005, The genomes of *Oryza sativa*: A history of duplications. PLoS Biology 3: 0266-0281.

Zhu, Y.L., Song, Q.J., Hyten, D.L., van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., Cregan, P.B., 2003, Single nucleotide polymorphisms in soybean. Genetics 163: 1123–1134.