Chapter 7
SNP Quality Assessment

Shaolin Wang, Hong Liu, and Zhanjiang (John) Liu

Identification of single-nucleotide polymorphisms (SNPs) relies on the alignment of multiple sequences derived from different genomes that include genomes of different individuals or the two haploid sets of chromosomes of a single diploid individual. Traditionally, large-scale SNP identification depends on the availability of whole genome sequences. Genome sequencing projects using traditional Sanger sequencing usually involve a single individual with two sets of chromosomes from which SNPs can be identified. The situation is quite different, however, with aquaculture species where no genomes have been entirely sequenced using the traditional Sanger sequencing, with the exception of the Atlantic salmon genome that is being sequenced using Sanger sequencing for the first phase of the sequencing project. As a result, SNP identification in aquaculture depends on various resources including expressed sequence tag (EST) databases, genomic survey sequences, and very recently, sequences generated using next generation sequencing technologies.

Advances in next generation sequencing have allowed the generation of whole genome sequences in aquaculture species. Currently, whole genome sequences have been generated or are being generated from several aquaculture species including Atlantic cod, channel catfish, carp, rainbow trout, tilapia, Pacific oyster, and shrimp. In some of these cases, many SNPs will be generated through the whole genome sequencing project; but in other cases, no SNPs will be expected from the whole genome sequencing project because of the use of doubled haploid as sequencing templates. This is particularly true for several teleost fish species. Teleost fish are believed to have gone through one extra round of whole genome duplication and therefore are believed to harbor duplicated genes or duplicated genome regions. In order to reduce complications in whole genome sequencing assembly, doubled haploid is used in several cases for whole genome sequencing such as in channel catfish, carp, and rainbow trout. Therefore, SNPs are not expected from these species during whole genome sequencing. Instead, additional efforts will have to be devoted in these species for SNP discovery, and subsequent SNP quality assessment would become an issue of interest.

Recent use of next generation sequencing technologies allowed sequencing of the same genomic segments from multiple individuals by sequencing reduced representation libraries (RRLs; see Chapter 5), greatly enhancing the power of SNP discovery. SNPs can also be identified from genome survey sequences such as BAC-end

Next Generation Sequencing and Whole Genome Selection in Aquaculture Edited by Zhanjiang (John) Liu
© 2011 Blackwell Publishing Ltd. ISBN: 978-0-813-80637-2

sequences if multiple sequences exist for the same genomic location, or from EST databases when multiple sequences are available from the same transcripts. While a high level of sequence redundancy may be involved in whole genome sequencing situations because the number of genome coverage is usually high (at least 5–10 genome coverage) as required for whole genome assembly, sequence redundancy for genome survey sequences or ESTs varies depending on the depth of sequencing efforts. As such, putative SNPs are identified from sequence alignments involving a variable number of sequences, for example, as few as two sequences. When only two sequences are aligned, a discrepancy of sequences could mean either a true SNP or a sequence error. Such situations necessitate the need for quality assessment of SNPs. In addition, not all SNPs are equal in terms of their genomic location and distribution, if they are involved in the genes, or whether they are adaptable to efficient genotyping for follow-up genomic studies. In this chapter, we provide some general considerations for SNP quality assessment.

Quality Assessment Parameters for EST-derived SNPs

SNPs derived from ESTs are very important SNPs as they reside within genes and thus may represent differences in protein-coding capacities. Biologically, gene-associated SNPs are obviously more important because they could account for the causations of phenotype differences, whereas SNPs in intergenic regions are only associated in location with phenotypes. In addition, gene-associated SNPs offer two additional benefits: (1) gene sequences are complex sequences so genotyping should be more straightforward in general than SNPs from unknown genomic locations that often involve simple sequences or repetitive sequences; (2) as genes are distributed throughout the genome, SNPs derived from ESTs, collectively at the genome scale, could have a wide distribution in the genome, alleviating some problems of SNP clustering. However, as ESTs represent single-pass sequences of cDNAs, EST sequences could involve a significantly higher level of sequencing errors, leading to the identification of "pseudo-SNPs." Use of EST-derived SNPs requires extensive consideration of quality assessment.

Sequence Base Quality Scores

If sequences were 100% correct, SNPs identified by multiple sequence alignments then represent true SNPs. However, no matter what method was used in generating the sequences, 100% of sequence accuracy is not achievable. Practically, therefore, sequence base quality score derived from sequencing trace files is a primary source for SNP quality assessment. Base calling is one of the most important factors for the identification of true SNP instead of sequencing errors. During base calling, the program PHRED scans the trace files and generates sequence FASTA file along with base quality file. Q20 (99% accuracy) is the most common cutoff quality score. However, in some cases, in the effort to obtain longer and more sequences, the quality score could have been lowered to Q15 (97.5% accuracy) or Q13 (95% accuracy), which may then significantly increase the sequencing errors. In order to reduce

the chances of pseudo-SNPs, it is recommended that only high-quality sequences are used unless absolutely required otherwise.

Several applications have been developed based on the sequence trace file or quality score evaluation for the SNP identification, such as POLYBAYES and POLYPHRED. These methods have been proven to provide high validation rates of SNPs (Hayes et al., 2007). However, these applications require sequence quality scores or original sequence trace files for SNP identification, which may not be available for many researchers using GenBank sequences as resources.

Sequence Redundancy

To assess SNP qualities after base calling, a major criterion is the sequence redundancy at the SNP site, which is determined, for the most part, by the number of sequences in the contig; but we stress the redundancy at the SNP site, not just the number of sequences in the contigs.

The first factor influencing EST sequence redundancy is related to the level of gene expression of the transcripts. For instance, in the dbEST database, over 6 million ESTs have been generated for human transcriptomes, with an average of over 150 ESTs per transcript. However, some most highly expressed genes can be represented in huge numbers in tens of thousands of counts, whereas some transcripts are represented only few times. ESTs are sequenced from cDNA libraries made from various conditions. The abundance of ESTs is proportional to the expression level of the genes if the cDNA libraries were not normalized. Even with normalized libraries, highly expressed genes tend to be overrepresented in the EST databases.

Sequence redundancy at the SNP site can be significantly lower than the number of sequences in the contigs. Given that transcript sizes can vary in sizes from several hundred base pairs (bp) to multiple of kilobases (kb), sequence redundancy is usually high at the 5′ end and/or at the 3′ end. This is because ESTs were sequenced from either the 5′ end or the 3′ end of the transcripts. Single-pass sequencing usually generate 600–800 bp. Therefore, sequences at the middle of the transcripts tend to have a lower sequence redundancy (Figure 7.1).

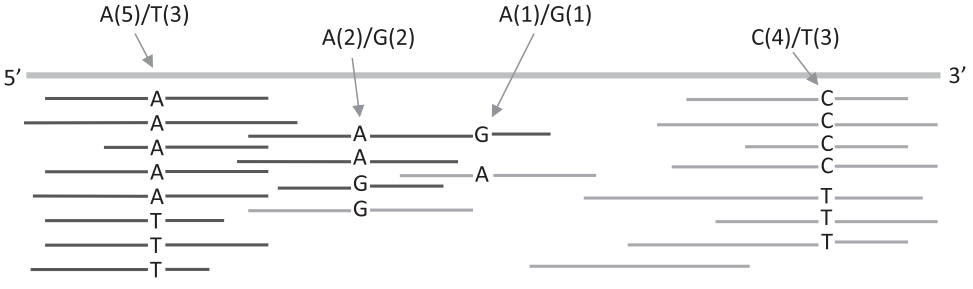

Figure 7.1 Importance of the minor sequence frequency and the number of sequences in the contig. Note that the number of sequences at the SNP site is the most important. For instance, more sequences are available at the 5′ and 3′ of a transcript, providing a greater level of reliability of sequences. In contrast, fewer sequences are available in the middle of transcripts. See color insert.

Large EST resources can be a tremendous source for SNP identification. For instance, the large number of human ESTs provided a tremendous resource for SNP identification from ESTs, which was proven to be sufficient for SNP identification for the HapMap project. Such a high level of sequence redundancy or coverage, however, may not be available for most species, especially not for the aquaculture species. The higher the sequence redundancy at the SNP sites, the more likely each of the alternative alleles can be observed more than one time, alleviating the problem of sequencing errors. For most aquaculture species, large sequence redundancy at the SNP sites may not be available. Therefore, quality assessment standards should be established for the SNPs identified through alignment of ESTs.

The Number of Sequences in the Contigs

Mining SNPs out of EST data with much smaller EST resources in aquaculture species (usually fewer than 500,000 ESTs per species) is still a highly productive approach. However, quality of the SNPs can become a greater challenge because of much smaller number of sequences in the contigs. Researchers are facing a dilemma of sacrificing for the number of SNPs or sacrificing for SNP quality. A wise resolution of such a dilemma can prove to be difficult but extremely useful. Keeping this dilemma in mind, we have conducted a pilot project (Wang et al., 2008). The objective of the pilot project was to develop a strategy for rapid and reliable identification and evaluation of qualities of EST-derived SNPs to reduce the rate of pseudo-SNPs resulted from sequence errors typically found in single-pass EST data sets, especially those deposited in the National Center for Biotechnology Information (NCBI) where sequence trace files or quality files may or may not be available. In this pilot project, about 55,000 catfish ESTs were downloaded from the NCBI dbEST database. The ESTs were assembled using CAP3, and putative SNPs were identified by using AUTOSNP.

Obviously, the number of sequences representing a specific transcript is the primary source for SNPs to be discovered. As such, the more sequences in the contigs, the more likely for one to detect SNPs from the contigs. Think of a contig with two ESTs; any sequence discrepancy would be identified as a putative SNP, providing no clue as to if the putative SNP is a true SNP or a sequencing error. Now, think of a contig with three sequences; a potential SNP would be presented by the sequence alignment as 1:2 at the SNP site. While a repeated calling of a base twice increased the chances of correctness for that allele, the base calling just once for the alternative allele provided no assurances if the base represent a SNP or a sequence error. Similarly, when four sequences are involved in a contig, one would have two possibilities: a 2:2 situation or a 1:3 situation. Obviously, one would feel more comfortable for the 2:2 situation to call the SNP than for the 1:3 situation as the one sequence could still be a sequence error. By the same token, when 10 sequences are involved in a contig, one can have allele ratios of 1:9, 2:8, 3:7, 4:6, or 5:5, and obviously, the confidence for a true SNP increases with the same order, with the 5:5 being the most likely to represent a true SNP. Technically, this involves both the chances of sequencing errors and the possibility of having both alleles being sequenced, given a fixed SNP allele frequency in the population. Clearly, it is not the contig sizes primarily (the number of sequences in the contigs), but the allele frequency that is most

important for the assessment of SNP quality; however, it is only when the number of sequences in the contig is large that it becomes possible to detect sequence allele distributions, thereby making some assessment of SNP quality.

Minor Allele Frequency

The sequence allele distribution is the most important indicator of SNP quality. The higher the minor allele frequency, the higher chance of a putative SNP to be a real SNP can be. Given a fixed number of sequences in a contig, the more equal the minor and major allele frequencies, the more likely the putative SNP can be a true SNP (Figure 7.2). Theoretically, a compound of sequencing quality and real SNP allele distribution determines the chances for the detected SNPs. For sequencing errors, the chances of getting the same sequencing error at the same sequence location multiple times becomes smaller and smaller as the number of sequences increases. For allele frequencies, the greater the minor allele frequency in the population, the

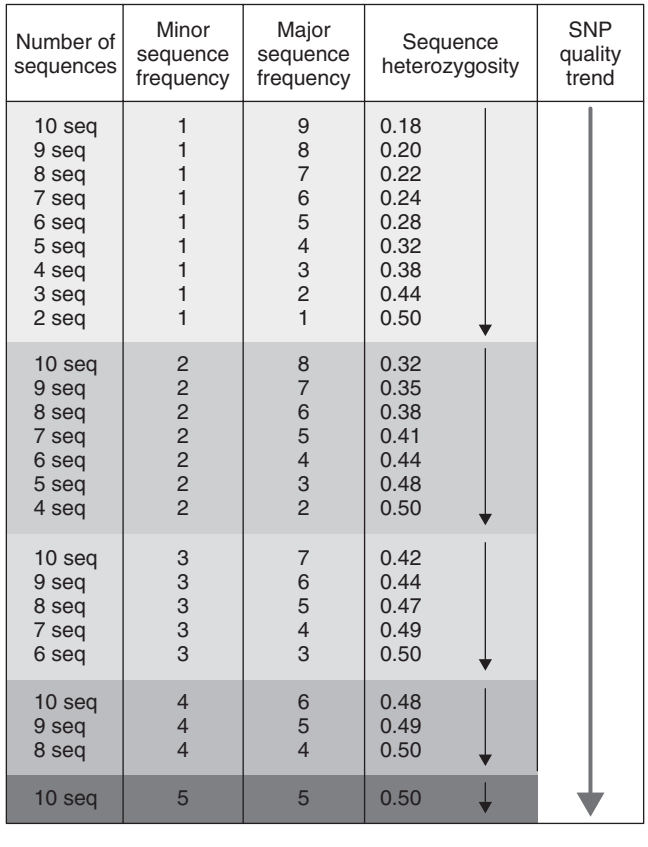

Figure 7.2 SNP quality assessment based on EST contig size and sequence frequency of the alleles. Arrows indicate the trend of increases of heterozygosity and the trend of increases in SNP quality. See color insert.

more likely one can detect the SNP. Because SNPs are regarded as biallelic markers (even though there are four alleles, theoretically, at any given SNP site), the largest chance for the detection of the SNP is when the two alleles share equal allele frequency of 50% each (Figure 7.2). These theoretical predictions were validated with our pilot studies (Wang et al., 2008).

The presence of minor allele sequence in relation to the contig size is important. For instance, if the minor allele sequence was present only once, then the smaller the contig size, the more likely the SNP could be real. This is because the contig size of ESTs is simply a reflection of expression abundance. If a rarely expressed gene was sequenced twice, with the alternative allele being present once each, one can still expect that the allele frequency could be equal or close to be equal when the transcript is sequenced 10 times. However, if the transcript was already sequenced 10 times, with the minor allele sequence being present only once, it is more likely that the minor allele could have been derived from sequencing errors (Figure 7.2). This relation is obvious when sequence heterozygosity is considered as shown in Figure 7.2. A contig of two sequences with one each of the alternative alleles would have a sequence heterozygosity of 0.5, while a contig with 10 sequences of major allele : minor allele = 9 : 1 would have a sequence heterozygosity of only 0.18.

Although the above discussion is correct, practically however, one does not have to increase SNP quality at the extreme expense of sacrificing the number of SNPs. This is to say that once the quality is high enough, one does not need to push for the highest SNP quality. In our pilot study, we attempted to determine the optimal level of quality in terms of minimal number of sequences in the contig and the minimal allele frequency, and the SNP validation rate. Our results indicated that a minimum of four sequences in the contig, with minor sequence allele to be represented at least twice, provided a relatively high level of SNP validation rate as tested in a single resource family (Table 7.1).

Table 7.1 SNP polymorphic rates as a function of contig size and minor sequence allele frequency as determined from a pilot study in catfish (Wang et al., 2008).

Number of sequences in the contig	Number of successful loci	Sequence ratio	Minimal minor sequence frequency	Polymorphic rate (%)
2	24	1:1	50%	33.3
3	37	1:2	33.3%	45.9
4	26	1:3	25%	15.4
Subtotal	87			33.3
4	44	2:2	50%	70.5
5–6	60	2:3 & 2:4 & 3:3	33.%	60.0
7–8	17	3:4 & 3:5 & 4:4	37.5%	64.7
9–12	21	4:5 & 4:6 & 4:7 & 4:8 & 5:5 & 5:6 & 5:7 & 6:6	33.3%	76.2
>12	37	5:7 & 6:6 & 5:8 & 6:7 … & 12:57	17.4%	89.2
Subtotal	179			70.9
Total	266			58.6

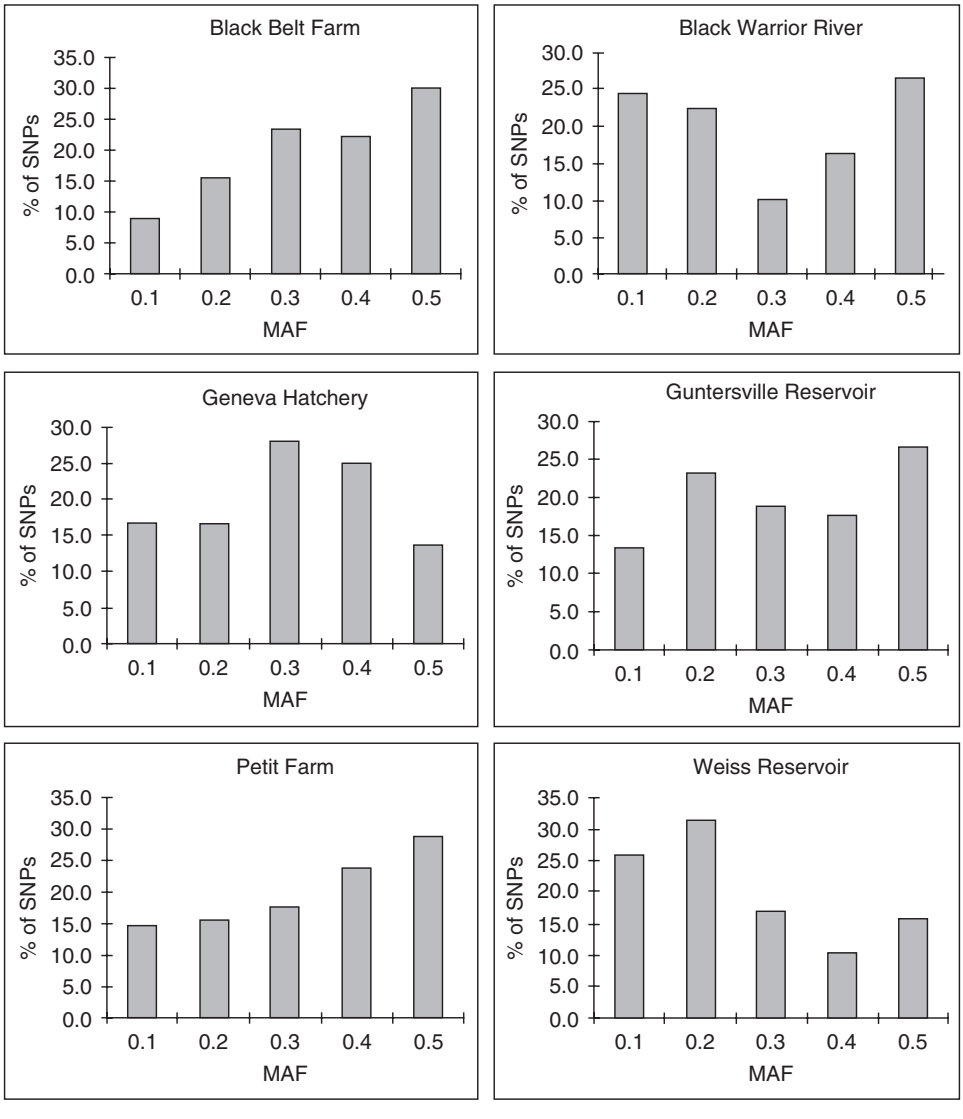

Figure 7.3 Distribution of minor allele frequency (MAF) in domestic and wild channel catfish strains. The label at the top of each panel refers to the name of the populations.

It is noteworthy to point out that the sequence allele frequencies generated from EST studies by no means reflect the real allele distribution in various populations as revealed by SNP analysis with various domestic and wild populations of catfish (Mickett et al., 2003; Simmons et al., 2006; Figure 7.3).

Sequences Flanking SNPs and Their Sequence Quality

In addition to sequence redundancy at the SNP site and minor allele frequency, several other factors were also important for SNP genotyping and validation. As

Illumina SNP genotyping technology is one of the most popular high-throughput SNP genotyping platforms; here we will introduce several important factors that can affect the success rate of EST-derived SNP genotyping. Illumina genotyping technology, and perhaps several other genotyping technologies as well, requires highly reliable SNP flanking sequences for efficient base extension, polymerase chain reaction (PCR), and genotyping. This would be particularly true for Affymetrix MyGeneChip Custom Arrays as this system is a hybridization-based system relying on sequences surrounding SNP sites for probes.

With EST-derived SNPs, sequence quality flanking the SNP sites was found to be important for successful SNP genotyping using Illumina's BeadArray technology, but the flanking sequence context was less important, when the Illumina quality score was above 0.5. It is probably true that SNP genotyping primers would have worked properly for the most part even if the sequence context was somewhat simple or A/T-rich, or G/C-rich. However, sequence errors in the SNP region could directly affect the base pairing of the SNP genotyping primers. Low-quality sequences could easily generate false SNPs, especially at the beginning or at the end of the sequence. Therefore, sequence quality surrounding the SNP site should be used as one parameter to identify reliable SNPs. However, many EST sequences retrieved from the NCBI do not have quality scores or trace files. In such cases, greater caution should be exercised. In particular, hot spot of SNP occurrence should be avoided if possible (Wang et al., 2008). It is worthwhile noting that the quality scores could become more important when genomic sequences are involved, which often involves repetitive sequences (Lepoittevin et al., 2010).

Flanking sequence quality greatly affected the SNP success rate. In the pilot study (Wang et al., 2008), we identified 28 contigs with hot spots of SNP occurrence where a region of sequence was highly variable with many "SNPs" detected. Sequence quality examination suggested low quality scores in the sequencing reactions. We intentionally included these SNPs in the pilot project to determine the effect of quality of sequences flanking SNPs. Of the 28 SNPs tested, 14 (50%) failed in genotyping, suggesting that high sequence quality is required in the SNP region as they are involved in genotyping primer binding regions.

In contrast to the quality of sequences flanking SNPs, the sequences themselves were not crucially important once the quality score is above 0.5. Illumina actually assigns a quality score as a reflection of the flanking sequence complexity and sequence context flanking SNPs. In our pilot study, we selected 384 SNPs with quality scores ranging from 0.5 to 1.0 to determine how Illumina quality scores affect the success rate of genotyping. Quality scores were not associated with the failures of SNP genotyping when the quality scores were above 0.5 (Wang et al., 2008).

The Presence of Introns Can Significantly Reduce the Success Rate of SNPs

With EST-derived SNPs, a major challenge is the potential presence of introns near the SNP sites. The presence of introns greatly reduced the SNP genotyping success rate. In our pilot study, among the contigs containing SNPs, four had genomic DNA

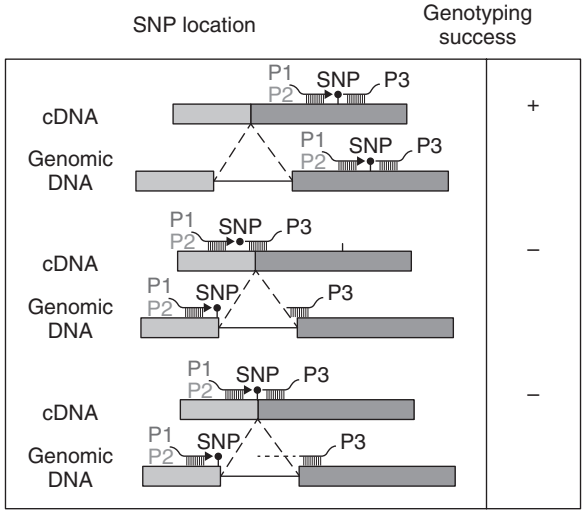

Figure 7.4 Schematic illustration of the effect of introns involved in SNP genotyping. In the first case, all the genotyping primers are located in the same exon nearby, leading to successful genotyping (+); in the second case (middle), one of the genotyping primers (P3 as shown) was located at the exon–intron border, causing nonbase pairing that leads to failure of genotyping (–); and in the third case, all primers were located in exon regions, but an intron was involved that demands PCR extension to across the intron. Apparently, the BeadArray technology provides very limited extension capability, leading to genotyping failure (–) as well. See color insert.

information that allowed us to test if the involvement of introns has a major effect on SNP genotyping and validation rates. All four SNPs failed to provide genotypes.

The reasons for the inability of successful genotyping when introns are present near the SNP sites may include the following. (1) The genotyping primers are located at the exon–intron boundary, leading to nonbase pairing of the primers with DNA amplified from genomic DNA (Figure 7.4). This is easy to understand as the genotyping primers are designed using cDNA sequences, while the real template used in genotyping is from genomic DNA. One possibility would be to genotype using RNA samples converted to cDNA, but this has not been tested. (2) The BeadArray technology depends on very short extension and subsequent ligation for success. The presence of introns requires long extension followed by ligation, which reduces the success rate; this is somewhat unexpected as one would expect that DNA polymerase should be able to extend easily a few hundred bases. Nonetheless, with experimental science, the guiding principle is that we should do whatever works.

Selection of SNPs to allow both allele-specific and locus-specific primers to be located in a single exon is the key to achieving high success rate of SNP genotyping. In absence of a whole genomic sequence or gene structural information for the involved SNPs, comparative gene organization analysis is a useful and productive approach to resolve the problems caused by the presence of introns near SNP sites. Bioinformatics analysis using *in silico* comparative sequence and gene structural analysis is important when dealing with EST-derived SNPs.

We conducted comparative sequence analysis of catfish ESTs with corresponding zebrafish genes as references. The rationale is that if the gene organization is similar in catfish and zebrafish, then sequence similarity comparison would allow the location of SNP sites to be aligned to the zebrafish genome. If the SNP sites are close to the exon–intron junction, then that could have caused the genotyping failures, assuming conservation of gene structure and organization between catfish and zebrafish. In our pilot study, 92 of the 99 catfish SNP loci had significant BLAST hits with the zebrafish genome, but of these, only 50 allowed sequence alignment in the region containing the involved SNPs. Sequence alignment and gene structure in zebrafish indicated that 32 (64%) of the 50 SNPs were located at the exon–intron border, suggesting that the presence of the presumed introns was the major cause for the failures of the SNP genotyping.

Assessment of SNP Distribution in the Genome

Even if the identified SNPs are real and can support successful genotyping, their utilities for genetic studies are dependent on their genomic location and distribution. It is not only the total number of SNPs that makes a difference in the power for genetic analysis but also their genomic distribution. Obviously, the more even distribution the SNPs have, the greater their power for genetic analysis. Highly clustered SNP will not provide additional power as they may not represent any haplotypes or there is no recombination among them.

The best scenario is to achieve as even a distribution as possible. Theoretically, for a genome with a size of 1×10^9 bp such as that of the catfish, if 20,000 SNPs are used, one would be able to have one SNP per 50,000 bp, assuming all SNPs are distributed at equal space throughout the genome. If this level of coverage can be achieved, the power for genetic analysis would be great as one can analyze association of SNPs with traits within a 50-kb region, which is highly manageable in most laboratories. The problem is the inability to achieve this level of distribution, leading to the presence of large gaps in the genome without any SNP coverage (Figure 7.5).

The most effective way for the assessment of SNP genomic location and distribution is through *in silico* mapping of SNPs to whole genome sequence assembly. This will soon be possible for several species of aquaculture species. One can select SNPs with roughly equal spaces in large scaffolds, and select at least one SNP in small scaffolds.

In the absence of whole genome sequences, comparative genome analysis may lend some insight into genomic distribution of SNPs, assuming a colinearity of gene

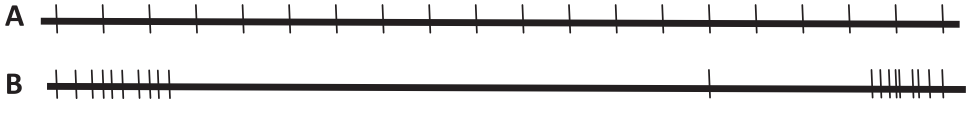

Figure 7.5 Relationship of SNP distribution in the genome and their abilities to provide good genome coverage. The best scenario is the even genomic distribution of SNP markers as shown in A; SNP clusters can significantly reduce the power of genome coverage as shown in B.

sequence arrangement at the genome scale. Whole genome sequence assembly of closely related species can be used as a resource sequence for the analysis of SNP distribution. Complexities can be caused by inabilities to locate genomic sequences to a related genome because of the lack of sequence similarities. Such an approach should, in theory, work well for gene-associated SNPs.

Quality Issues of SNPs Generated from Sequencing RRLs

As detailed in Chapter 5, large numbers of SNPs can be rapidly identified through sequencing of RRLs by using next generation sequencing. SNPs so identified should not be much different from those identified through whole genome sequencing projects except that (1) the contig size (total length in base pairs) may be short, thereby limiting the utility of the identified SNPs; (2) the sequence context could be quite simple or repetitive in nature, which limits the utility of the identified SNPs; and (3) the contig assembly provide no information as to the genomic location and distribution of the identified SNPs.

Small contig length can significantly reduce the utility of the identified SNPs. For instance, if SNPs are located at or close to the beginning and end of contigs, there would be insufficient flanking sequences for the design of genotyping primers.

The sequence context can have a significant effect on the utility of the SNPs. Given the short length of the contigs constructed using the next generation sequencing technology, the flanking sequences near the SNP sites can fall under either the very simple sequences or repetitive sequences. Teleost fish genomes are high in repetitive sequences such as Tc-1/mariner transposons. SNPs involved in such sequences may appear to be useful, the genotyping of which may prove to be difficult. For instance, Tc-1/mariner repetitive elements represent 4.6% of the catfish genome (Xu et al., 2006; Nandi et al., 2007; Liu et al., 2009). Genomic distribution of SNPs from sequencing the RRLs should be random as the genomic segment is randomly chosen, assuming avoidance of repetitive elements selected for the RRLs.

Conclusions

As compared with SNPs identified from genomic sequences, EST-derived SNPs have several advantages. Since ESTs are transcribed sequences, EST-derived SNPs are associated with actual genes, allowing use of gene-associated SNPs for mapping and subsequent use in comparative genome studies (Sarropoulou et al., 2008). This is particularly important for species without a genome sequence such as aquaculture species. In addition to be used as markers for mapping, SNPs are also considered a rich source of candidate polymorphisms underlying important traits, leading to the identification of causative genes or quantitative trait nucleotide (QTN) (Jalving et al., 2004). However, there are several important factors to be considered when using EST-derived SNPs. The major issue for development of SNPs from EST resources is not whether SNPs can readily be identified, but to what degree these SNPs would be reliable because parameters for quality assessment of EST-derived

SNPs simply do not exist. This reliability issue was mostly due to sequence errors; assembled contigs with sequence variation could simply be sequence errors. Additionally, since SNPs derived from ESTs can only be identified from EST contigs where the same gene transcripts were sequenced at least twice and sequencing frequency of ESTs is not random, large-scale sequencing is required to identify SNPs from rarely expressed genes. Moreover, SNP rates could be lower in coding regions because of evolutionary restraints of selection pressure.

The contig size (number of sequences in the contig) and minor sequence allele frequency were the two major factors affecting the validation rates of EST-derived SNPs. Small contigs had much lower SNP validation rates. Obviously, in small contigs with two or three sequences, the alternative base is represented only once, and this could be due to sequencing errors. Similarly, in contigs with four sequences when the minor sequence allele is represented only once, it is highly likely that the minor allele is due to sequencing errors. Contigs of four or more sequences, with the minor sequence allele frequency being present at least twice in the contig, provided high levels of SNP validation rates (averaging 70.9% up to 89.2%). This makes good sense because it is highly unlikely for sequencing errors of two independently sequenced ESTs to occur at the same base location. When at least two ESTs exhibit an alternative base at the putative SNP sites, it is highly likely that such sequence variations are real.

Even with true SNPs, a key issue to the success of SNP genotyping using EST-derived SNPs is the avoidance of introns. Genotyping using cDNAs as templates could, in theory, reduce genotyping complications due to the presence of introns, but such an approach yet needs to be tested.

SNP locations and genome distributions are equally important for genetic analysis powers. The best scenario is even distribution of all SNPs, avoiding SNP clustering. If even genomic distribution can be achieved, a large genome can be effectively covered with a reasonable number of total SNPs. The best approach for the analysis of genome distribution of SNPs is *in silico* mapping if the whole genome sequence is available. If the whole genome sequence assembly is not available, cross-species *in silico* comparative analysis is highly useful.

In spite of the power of SNP identification using next generation sequencing with RRLs, SNPs so identified are associated with potential complications of being located at the ends of sequences, flanked with simple sequences or repetitive elements, and without information on genomic location and distribution.

References

Hayes BJ, Nilsen K, Berg PR, Grindflek E, and Lien S. 2007. SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. *Bioinformatics*, 23:1692–1693.

Jalving R, van't Slot R, and van Oost BA. 2004. Chicken single nucleotide polymorphism identification and selection for genetic mapping. *Poult Sci*, 83:1925–1931.

Lepoittevin C, Frigerio J-M, Garnier P, Salin F, Cervera T, Vornam B, Harvengt L, and Plomion C. 2010. *In vitro* vs *in silico* detected SNPs for the development of a genotyping array: What can we learn from a non-model species? *PLoS ONE*, 5:e11034.

Liu H, Jiang Y, Wang S, Ninwichian P, Somridhivej B, Xu P, Abernathy J, Kucuktas H, and Liu Z. 2009. Comparative analysis of catfish BAC end sequences with the zebrafish genome. *BMC Genomics*, 10:592.

Mickett K, Morton C, Feng J, Li P, Simmons M, Cao D, Dunham RA, and Liu Z. 2003. Assessing genetic diversity of domestic populations of channel catfish (*Ictalurus punctatus*) in Alabama using AFLP markers. *Aquaculture*, 228:91–105.

Nandi S, Peatman E, Xu P, Wang S, Li P, and Liu Z. 2007. Repeat structure of the catfish genome: A genomic and transcriptomic assessment of Tc1-like transposon elements in channel catfish (Ictalurus punctatus). *Genetica*, 131:81–90.

Sarropoulou E, Nousdili D, and Magoulas AGK. 2008. Linking the genomes of nonmodel teleosts through comparative genomics. *Mar Biotechnol (NY)*, 10:227–233.

Simmons M, Mickett K, Kucuktas H, Li P, Dunham R, and Liu ZJ. 2006. Comparison of domestic and wild channel catfish (*Ictalurus punctatus*) populations provides no evidence for genetic impact. *Aquaculture*, 252:133–146.

Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, Peatman E, Kucuktas H, and Liu Z. 2008. Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics*, 9:450.

Xu P, Wang S, Liu L, Peatman E, Somridhivej B, Thimmapuram J, Gong G, and Liu Z. 2006. Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. *Anim Genet*, 37:321–326.