Chapter 6
SNP Discovery through EST Data Mining

Shaolin Wang and Zhanjiang (John) Liu

The key issue of single-nucleotide polymorphism (SNP) marker applications in aquaculture species is the availability of SNPs. Identification of large numbers of SNPs requires massive sequencing efforts and resources (Picoult-Newberg et al., 1999). Prior to the application of next generation sequencing, large-scale genome sequencing was not possible with aquaculture species. With recent adoption of next generation sequencing, it is obvious that SNP discovery has been made possible through both whole genome sequencing or through sequencing of reduced representation libraries (RRLs; for details, see Chapter 5). In spite of such efforts and possibilities, it is expected that in the near future, whole genome sequences will still not become available for the vast majority of aquaculture species, especially not for the minor aquaculture species. Therefore, SNP identification in aquaculture species will still likely be using various alternative available resources such as expressed sequence tags (ESTs). In this chapter, we will focus on SNP discovery through mining of EST databases.

Advantages and Disadvantages of SNP Discovery through EST Data Mining

ESTs are already available in the databases, so additional sequencing efforts are not essential. ESTs are single-pass sequence reads generated by direct sequencing of cDNA clones. They have been generated in the course of expression studies. In recent years, EST resources have become available for many aquaculture species, and a summary is provided in Table 6.1 for major aquaculture species.

ESTs-derived SNPs are associated with genes, therefore they are type I markers. Gene-associated SNPs can account for genomic causes of phenotypes. In this regard, gene-associated markers are superior to markers identified from anonymous genomic regions.

EST-derived SNPs should correlate genes in terms of genomic locations. While the numbers of markers available is very important, it is even more important to have markers that are evenly distributed in the genome. In genomic scale, genes are distributed in all chromosomes and chromosome segments, allowing EST-derived SNPs to have the potential of the same distribution in the genome, thereby reducing the levels of marker clustering.

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

Table 6.1 EST resources available from the major aquaculture species as of April 2, 2010 (dbEST release 040210).

Species	Number of ESTs
Sanger EST sequences	
Salmo salar (Atlantic salmon)	494,392
Ictalurus punctatus (channel catfish)	354,434
Oncorhynchus mykiss (rainbow trout)	287,928
Gadus morhua (Atlantic cod)	206,649
Litopenaeus vannamei (white shrimp)	161,075
Ictalurus furcatus (blue catfish)	139,475
Oreochromis niloticus (tilapia)	117,193
Crassostrea gigas (Pacific oyster)	57,139
Dicentrarchus labrax (European sea bass)	54,200
Penaeus monodon (giant tiger prawn)	35,180
Cyprinus carpio (common carp)	34,056
Eriocheir sinensis (Chinese river crab)	16,882
Mytilus galloprovincialis (Mediterranean mussel)	15,408
Crassostrea virginica (Eastern oyster)	14,560
Fenneropenaeus chinensis (fleshy prawn)	10,446
454 EST sequences	
O. mykiss (rainbow trout)	1,298,911
C. carpio (common carp)	242,263

However, ESTs are single-pass sequences, and the sequence qualities are relative low. Therefore, identification of SNPs from ESTs may lead to pseudo-SNPs because of sequencing errors. In order to overcome the sequencing errors, deep sequencing and higher coverage are required to achieve higher sequencing accuracy, which may not be available for most aquaculture species. Without the whole genome sequences, the intron information is unknown, which also could lead to genotyping failures because of failed primer amplification caused by intron involvement. During the assembly of EST data set, the low stringent assembly may bring the related gene family members into the same contigs, which also leads to identification of pseudo-SNPs.

Efforts have been made to develop SNP markers in aquaculture species (He et al., 2003). Recently, efforts for application of EST-derived SNPs are increasing (Hayes et al., 2007; Moen et al., 2008; Wang et al., 2008). High-throughput SNP genotyping chips have been developed for salmon and catfish for whole genome association studies. Identification of SNPs from ESTs requires extensive bioinformatics support. Here we introduce several software packages commonly used for SNP discovery.

SNP Discovery Using ESTs Generated by Sanger Sequencing

SNP discovery using EST requires the alignment of multiple EST sequences generated from a single transcript. Its basic procedures involve (1) retrieving EST resources from the National Center for Biotechnology Information (NCBI) dbEST database;

(2) cluster analysis of ESTs by sequence alignments; and (3) identification of mismatch (SNP) based on the alignment of multiple sequences.

Retrieving EST Sequences

Retrieving ESTs from existing databases is the first step for researchers who wish to use various EST sources generated from different laboratories. This step is not required if one already has the ESTs with the original trace files or stored as FASTA format.

NCBI dbEST is the most useful and powerful database, including over 65 millions (as of April 2010) of ESTs from various species. To retrieve all the EST resource from the related species, go to the NCBI Web site (www.ncbi.nlm.nih.gov): Click the "Search" drop-down menu and select EST, and then input the species name, such as *Ictalurus punctatus* (scientific names can provide accurate searches); choose the "FASTA" format from the "Display" drop-down menu; go to the "Send" drop-down menu and select file, and you should be ready to download the EST data set. The EST sequences retrieved using the above methods is in the FASTA file format. If you need to retrieve the trace files for some special SNP discovery program, such as POLYBAYES and POLYPHRED (Nickerson et al., 1997; Marth et al., 1999), you need to go to the trace repository at NCBI that is under construction at the NCBI. Currently, EST sequence traces can be downloaded from the Washington University FTP site: (ftp://genome.wustl.edu/pub/ gsc1/est) for ESTs produced there.

Multiple Sequence Alignments

A number of methods are available for multiple sequence alignments, and here we will just discuss a few. If a small number of genes are involved, some of the easy portals, such as the Basic Local Alignment Search Tool (BLAST) and ClustalW probably would provide a quick and convenient way for SNP identification. If large numbers of ESTs are involved, EST assembly would be needed.

NCBI BLAST

Computational SNP discovery, in a general sense, refers to the process of compiling and organizing DNA sequences that represent orthologous regions in samples of multiple individuals, followed by the identification of polymorphic sequence locations. The first step typically involves a similarity search with BLAST to compile groups of sequences that originate from the region under examination (Altschul et al., 1990). This is followed by the construction of a base-wise multiple alignment to determine the precise, base-to-base correspondence of residues present in each of the samples in a group. Finally, each position of the multiple alignments is scanned for nucleotide mismatches. The following is a step-by-step example of how to use BLAST

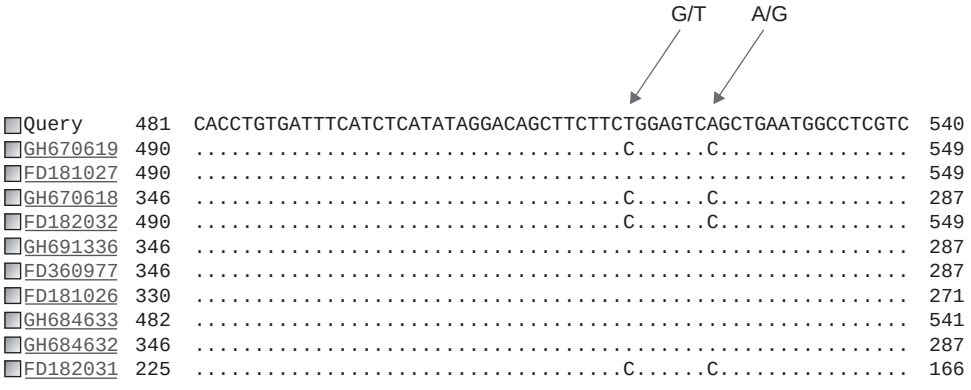

Figure 6.1 SNP visualization from BLAST results with the option of "Query-anchored with dots for identities." The dots represent identical nucleotides at the locus. The positions indicated by arrows are the SNP site with C/T and A/C SNPs.

to discover SNP from EST sequences: (1) If you already have a target gene or EST sequence for SNP discovery, (2) go to NCBI BLAST (blast.ncbi.nlm.nih.gov/Blast.cgi); under the Basic BLAST category select nucleotide BLAST, (3) paste or input the query sequence, select EST database for the species you are looking, such as "est_others" for all aquaculture species, and specify the species for the accurate search, such as *I. punctatus*, then click BLAST button; (4) once the blast results pop up, go to "format option" on the top of the BLAST result page, select "Query-anchored with dots for identities" under alignment menu, then click "Reformat" menu; the alternative alleles for the putative SNPs will show with the letters in the multiple alignment view results; all identical nucleotides will turn into dots in the results (Figure 6.1).

The NCBI BLAST also has a stand-alone version, which could be run on all platforms including Windows, Unix/Linux, and Mac OS. The BLAST programs could be found on the NCBI FTP site (blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). The command for running the BLAST program is identical on all platforms. In order to run BLAST searches on a local computer, the sequences need to be downloaded and the databases need to be prepared first. NCBI BLAST program provides a command "formatdb," which could be used to set up the database. The EST sequences file (FASTA format) can be directly used to set up the database with command "formatdb -i input_file (seqeunces) -p F (nucleotide)." Once the database has been set up, the BLAST searches could be performed by using "blastall -p blastn -i input_file (target gene or sequences for SNP identification) -d database -o output_file -m 2 (formatting the results to Query-anchored no identities)."

Multiple Alignment Tool

The ClustalX/ClustalW program has been widely used for both protein and nucleic acid multiple sequence alignments and construction of phylogenetic trees (Higgins et al., 1996; Jeanmougin et al., 1998). The program has undergone many improve-

ments since Clustal was first described in 1988 (Higgins and Sharp, 1988) and is available for different platforms including, most recently, ClustalW. It has an interface to Unix/Linux, the Macintosh Mac OS system, and MS Windows systems (Thompson et al., 1994, 1997). If you have multiple genes and ESTs, ClustalW can be used to construct multiple alignments instead of doing the multiple alignments using BLAST one by one. The program ClustalW has a stand-alone version for all platforms including Windows, Unix/Linux, and Mac OS. The European Molecular Biology Laboratory (EMBL) also has developed ClustalW for Web site access through (www.ebi.ac.uk/Tools/clustalw2/index.html), which is very convenient. ClustalW2 currently supports seven multiple sequence formats. These are

- NBRF/PIR
- EMBL/UniProtKB/Swiss-Prot
- Pearson (FASTA)
- GDE
- ALN/ClustalW
- GCG/MSF
- RSF

Identification of SNP using ClustalW is similar to the NCBI BLAST. First, the sequences (FASTA format) can be pasted into the input BOX of ClustalW (Web site). If you would like to use the default parameters, just click the "RUN" button. If you just have several sequences, you can wait for a while until the results is generated. If you have a large number of sequences, you can upload your sequence file through the Web site and leave your email address, and the notice for retrieving the results will be sent to your mailbox once the multiple alignments are finished.

The alignment results can be directly used to find the SNP information (Figure 6.2). The stand-alone ClustalW program has both Windows and Linux versions. Under Windows ClustalX, sequences and profiles (a term for preexisting alignments) are input using the File menu "Load Sequences." ClustalX has two modes that can be selected using the switch directly above the sequence display: Multiple Alignment Mode and Profile Alignment Mode. To do a multiple alignment on a set of sequences, make sure Multiple Alignment Mode is selected. A single sequence data area is then displayed. The alignment menu then allows you to either produce a guide tree for the alignment, or to do a multiple alignment following the guide tree, or to do a full multiple alignment. In Profile Alignment Mode, two sequence data areas are displayed, allowing you to align two alignments (termed profiles). Profiles are also used

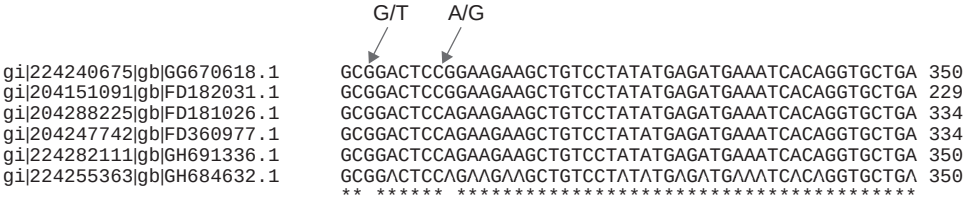

Figure 6.2 SNP visualization from ClustalW. The asterisks represent identical nucleotides at the locus. The positions indicated by arrows are the SNP site with G/T and A/G SNPs.

to add a new sequence to an old alignment, or to use secondary structure to guide the alignment process. Under Linux, ClustalW will generate three files with the command "clustalw -inputfile (FASTA sequence format)," including output file (the process of analysis), dnd file (guide tree files for construction phylogenetic tree), and aln file (multiple alignments file). The last one is the most important file used to identify SNPs.

EST Assembly

The above approaches are suiTable for single gene or small-scale SNP discovery. Large-scale SNP discovery from ESTs requires much more efficient approaches to construct multiple sequence alignments. ClustalW can provide some capability to construct multiple alignments, but it will take very long time for large-scale SNP discovery using a large number of EST sequences. The sequence assembly programs provide strong computational powers for the SNP discovery from high-throughput ESTs. Two types of assembly programs have been widely utilized on the Unix/Linux platform: (1) the first type in which the original trace file or the quality file is necessary, such as PHRAP; (2) the second type in which the trace file or quality file is not necessary, such as CAP3. In most scenarios, EST trace files are not available for scientists, such as the sequences directly retrieved from the NCBI dbEST databases. Therefore, CAP3 is quite useful.

POLYBAYES and POLYPHRED

The program POLYBAYES requires the original sequence trace files to detect SNPs, but the trace files need to be processed with the software PHRED (base calling) and PHRAP (assembly) developed by the University of Washington (PHRED/PHRAP/CONSED architecture) (Ewing and Green, 1998; Gordon et al., 1998, 2001; Ewing et al., 1998).

First, make sure that the sequence FASTA files and sequence quality files were prepared before assembly using PHRED with prompt command. All the sequences trace files are stored in the directory "chromat_dir" by specifying the location of this directory with the "-id" option. Sequence PHD format files are created in the "phd_dir" subdirectory by specifying the location of this directory with the "-pd" option. The complete command for this step is "phred -id chromat_dir (sequences trace files) -pd phd_dir (phd output files)." The program PHD2FASTA will be used to change the PHD format file to FASTA format file with option ("-os" option) of the ESTs file. Also, the program can produce a FASTA format file for the accompanying base quality values with option ("-oq" option), and one for the list of base positions that specify the location of each called nucleotide relative to the sequence trace with option ("-ob" option). The DNA sequence of the ESTs will be used in the next step, as the members of the cluster (group) of expressed sequences to be analyzed for polymorphic sites.

The application POLYBAYES packages provides two ways to construct multiple alignments for SNP identification, and all the process results will be stored in a directory named "edit_dir," which is required for CONSED. First, such as in RPBMAP (one of the applications in the POLYBAYES package), identification of SNPs is

conducted by mapping the EST sequences to the anchor sequences (reference sequences), then a multiple alignment of the EST sequences using the anchored alignment algorithm implemented within POLYBAYES is created. The CROSS_MATCH dynamic alignment program was utilized to compute the initial pairwise alignments between each of the ESTs and the genomic anchor sequence. Second, RPBACE (one of the applications in the POLYBAYES package) is used for SNP identification from *de novo* EST assembly. The assembly was required before the SNP identification if the reference sequences are not available. Vector sequence trimming was required before the assembly using CROSS_MATCH (cross_match seqs_fasta vector.seq -minmatch 12 -minscore 20 -screen > screen.out). The clean sequence file, "seqs_fasta.screen," will be generated. Then, the assembly will be performed by using PHRAP (phrap seqs_fasta.screen -new_ace > phrap.out). The program PHRAP will write the assembly results to the ace file. The ace file can be used for SNP identification using POLYBAYES (RPBACE).

The multiple alignments are scanned for polymorphic sites. At each site, the slice of the alignment composed of nucleotides contributed by every sequence that was locally aligned is examined for mismatches. The Bayesian SNP detection algorithm calculates the probability that such mismatches are the result of true polymorphism as opposed to sequencing error. Likely, polymorphic sites are recorded as SNP candidates. The SNP detection feature is enabled with the "-screenSnps" option. The new ace file should be generated after the SNP screening, which could be opened using CONSED. Figure 6.3 shows an output file with the site of a SNP candidate in

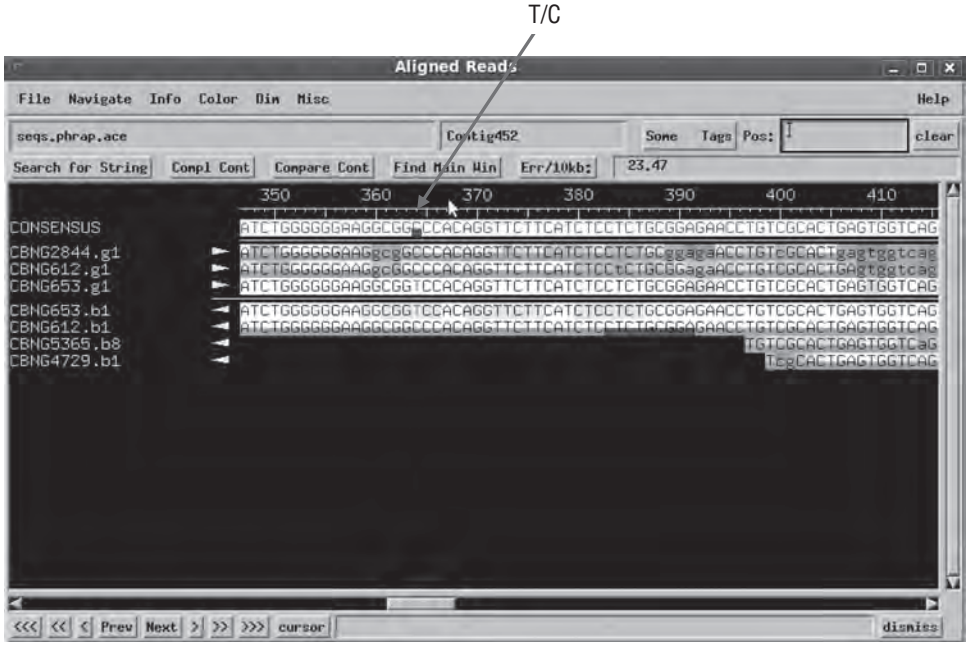

Figure 6.3 SNP visualization from POLYBAYES. The SNP identified at position 364 is a C/T SNP, which was generated from SNP screening based on multiple ESTs using POLYBAYES. See color insert.

the multiple alignments. This SNP is found within members of one alternatively spliced group of EST sequences and is automatically tagged by the SNP detection algorithm implemented within POLYBAYES. A similar procedure is applicable for a wide range of scenarios where sequence fragments (e.g., ESTs, random genomic shotgun reads, bacterial artificial chromosome (BAC)-end reads, sequenced restriction fragments) are organized with the help of genome reference sequence and compared against each other and/or with the reference sequence in search of polymorphic sites.

POLYPHRED is another SNP detection application similar to the POLYBAYES. SNP identification using POLYPHRED requires the similar steps as in using POLYBAYES. The only difference is that after base calling with PHRED, the polymorphism output file is required during the SNP analysis with POLYPHRED, but not with POLYBAYES. POLYPHRED is a program developed based on sequence fluorescence across traces obtained from different individuals to identify heterozygous sites for single-nucleotide substitutions. The functions of POLYPHRED are integrated with the use of three other programs: PHRED/PHRAP/CONSED. POLYPHRED identifies potential heterozygotes using the base calls and peak information provided by PHRED and the sequence alignments provided by PHRAP. Potential heterozygotes identified by POLYPHRED are marked for rapid inspection using the CONSED tool. POLYPHRED is very powerful to identity heterozygous individual because of fluorescence-based SNP discovery algorithm (Figure 6.4). The

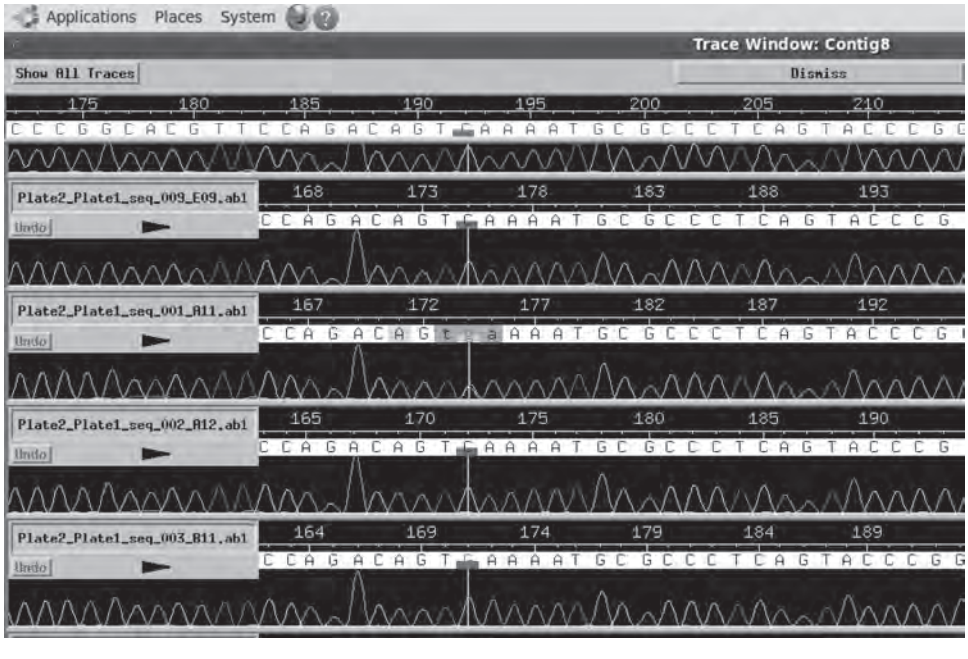

Figure 6.4 SNP visualization from POLYPHRED. The SNP identified at position 192 is a C/G SNP, which was generated from SNP screening based on individual fish. The sample E09 and A12 has homozygous allele C, the sample B11 has homozygous allele G, and the sample A11 is heterozygous with both allele C and allele G. See color insert.

command for PHRED is "phred -id chromat_dir (sequences trace files) -pd phd_dir (phd output files) -dd poly_dir (polymorphism output files)." All the other following steps before SNP discovery are similar to the process for the POLYBAYES. The command for SNP discovery with POLYPHRED is "polyphred -a Input.ace -o polyphred.out" for the default settings. The output report can be viewed with text editor, or CONSED can be utilized to view the tags added by POLYPHRED.

AUTOSNP

The CAP3 program includes a number of improvements and new features (Huang and Madan, 1999). The program has a capability to clip 5′ and 3′ low-quality regions of EST sequence reads. It uses base quality values in computation of overlaps between reads, construction of multiple sequence alignments of reads, and generation of consensus sequences. The program also uses forward-reverse constraints to correct assembly errors and link contigs. PHRAP often produces longer contigs than CAP3, whereas CAP3 often produces fewer errors in consensus sequences than PHRAP. It is easier to construct scaffolds with CAP3 than with PHRAP on low-pass data with forward-reverse constraints. CAP3 also has the capability to assemble the ESTs without any quality files or sequence trace files.

The CAP3 program can be used directly with all FASTA format files. CAP3 is easy to use by inputting the command "cap3 input_file" without giving any settings. The CAP3 program will use all the default parameters to assemble all input sequences and generate contig and singleton sequences. If one prefers to set up more stringent or loose parameters for the assembly, one can give the parameters by himself or herself. The most common parameters for the assembly is "-o," which specify the cutoff length of sequence overlap (default 40 bp) and "-p," which specify the cutoff percent of identity within the overlap sequence (default 80). If the sequence quality file is available along with the sequence file, it can be used to provide more accurate assembly with the quality score. However, the disadvantage of using quality score is the requirement for more system memory to perform the assembly. CAP3 assembly generates several files; the most important file is the ace file, which includes all the assembly information. The ace file will be utilized for the next step in SNP identification.

AUTOSNP is a program to detect SNPs and insertion/deletion polymorphisms (indels) with EST data (Barker et al., 2003). AUTOSNP is a perl-based program that can use *d2cluster* or CAP3 to cluster and align EST sequences. The program can also directly read the ace file generated by CAP3 and rebuild the multiple alignments within each contig. The program uses redundancy to differentiate between candidate SNPs and sequence errors. Candidate polymorphisms are identified as occurring in multiple reads within an alignment. For each candidate SNP, two measures of confidence are calculated—the redundancy of the polymorphism at a SNP locus and the codetection of the candidate SNP with other SNPs in the alignment (sort of cosegregation due to close linkage, or the assumption of being within the same haplotype). The mismatch will be identified from the multiple alignments based on the parameters. The default parameters for SNP identification are as follows: (1) a sequence variation is declared as a SNP whenever a mismatch is identified within contigs with four or fewer sequences; (2) a sequence variation is declared as a SNP

when the minor allele sequence existed at least twice within contigs with five to six sequences; (3) a sequence variation is declared as a SNP when the minor allele sequence existed at least three times within contigs with seven to eight sequences; (4) similarly, a sequence variation is declared as a SNP when the minor allele sequence existed at least four times within contigs with 9–12 sequences, and (5) when the minor allele sequence existed at least five times within contigs with 13–16 sequences, and so on.

Before using AUTOSNP, the CAP3 program is necessary and needs to be set promptly, especially if FASTA file is used; the command should be "cap3SNP.pl -f input_file (fasta file)." If the assembly ace file have already been generated by CAP3, the command should be "cap3SNP.pl -a input_file (ace file)." The memory requirement for the program depends on the number of sequence used in the assembly, 4–8 GB of memory were recommended for SNP identification with 100,000–200,000 sequences.

After execution of SNP detection, AUTOSNP will generate a folder, usually named "results," that will hold all the HTML files including the sequence information, SNP information, and sequence alignments for each contig (Figure 6.5). A summary HTML file will also be generated including the assembly and SNP identification summary information. If text files are desired (by using "-t" option), three text files will be generated including contig.txt, snps.txt, and sequences.txt. The file contig.txt includes the number of sequences and SNP identified in each contig

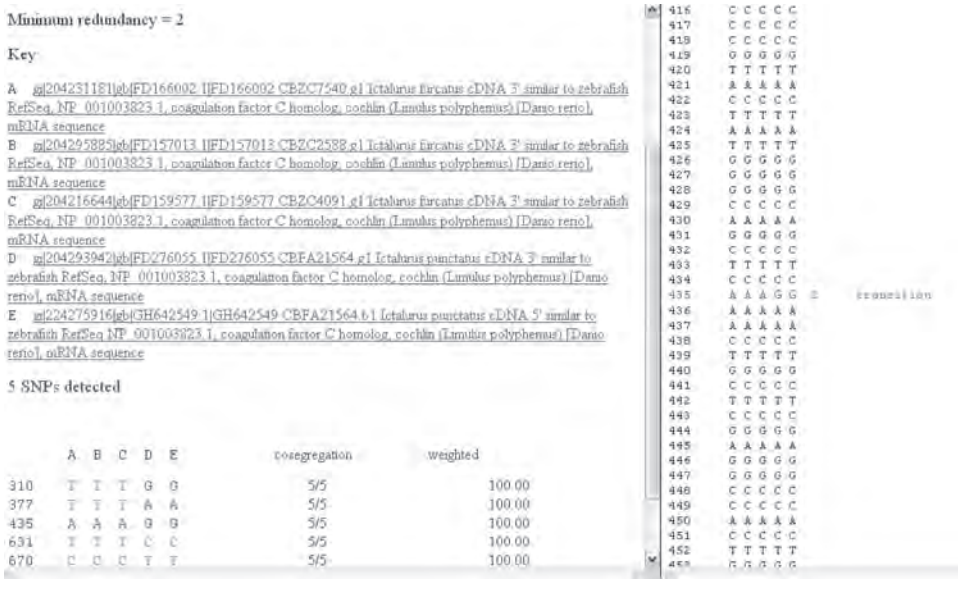

Figure 6.5 SNP visualization from AUTOSNP. The left upper panel displays sequence information, for example, GenBank accession numbers and putative gene identities. The left lower panel displays SNP summary information, for example, at position 310, the SNP is a T/G SNP, at position 377, the SNP is a T/A SNP with a sequence ratio of 3:2 each. The right panel displays the sequences alignment information, highlighting the SNP at position 435, an A/G SNP with a sequence ratio of 3:2. See color insert.

sequence. The file snps.txt provides the detailed information for every SNP including the location of SNP in the contig, number of minor allele frequency, SNP cosegregation information, and all the alleles identified in each SNP. The file sequences.txt provides the sequences alignment information for each contig. The .txt file is relatively easy to access, which could be combined with Access or MySQL database for data management, especially for large-scale sequence assemblies and SNP identification. The HTML file is more intuitive and combines the contig, sequences, and SNP information together in one file.

SNP Discovery Using Transcript Sequences Generated by Next Generation Sequencing

Recently, transcript sequences have been generated using the next generation sequencing platforms such as Illumina Genome Analyzer and the Roche 454 sequencer. These sequencing platforms generate relatively short sequences with a huge number of sequence reads. As a result, NCBI has deposited these short sequence reads into a special database called Short Reads Archive (SRA). Here in this section, we will present methods for SNP detection using SRA sequences.

GIGABAYES

In order to adapt the next generation sequencing data from 454 and Illumina sequencing platforms, SNP discovery applications are required to optimize the characteristics of new sequencing platforms. New SNP discovery application packages (PYROBAYES/MOSAIK/GIGBAYES) for SRA sequences have been developed based on POLYBAYES, which was originally developed for Sanger sequencing, as discussed above. The application package includes three applications: PYROBAYES, for 454 pyrosequencing base calling (Quinlan et al., 2008); MOSAIK, for Reference Guided Read Aligner/Assembler; and GIGABAYES, for short-read polymorphism detection (Smith et al., 2008). The alignment visualization applications EAGLEVIEW and GAMBIT have been developed to adapt the next generation sequencing assembly and alignment.

The package does not include any *de novo* assembly application, which means that if reference sequences are not available, *de novo* assembly is required before SNP discovery. In addition to the assembly applications developed along with the next generation sequencing platforms, several independent *de novo* assembly applications have been developed, such as Celera Assembler (Rausch et al., 2009), Velvet for both 454 and Illumina (Zerbino and Birney, 2008), and SOAPdenovo for Illumina (Li et al., 2008, 2009).

The next generation sequencing SNP discovery using GIGABAYES requires several steps:

1. Prepare the reference sequence; if the reference sequence is not available, *de novo* assembly is required to produce reference sequences.

2. Once the reference sequences have been generated, which could be used for alignment, MOSAIK package is required to build the alignment before the SNP discovery by using MosaikBuild, MosaikAligner, MosaikSort, and MosaikAssembler. The MOSAIK program package is briefly introduced below, but the detailed information for the use of this package is not provided here as it can be found in the manual of the software package.
   a. In order to speed up the assembly process, MosaikBuild can convert external read formats, including FASTA, FASTQ, and sequence read format (SRF), to a compressed archive format that the aligner can readily use. In addition to processing reads, the program also converts reference sequences from a FASTA format to an efficient binary format.
   b. MosaikAligner performs pairwise alignment between every read in the read archive and the reference sequences. Then, MosaikSort takes the pairwise alignment output and sort the alignment. For single-ended reads, MosaikSort simply resorts the reads in the order they occur on each reference sequence. For mate-pair/paired-end reads, MOSAIK resolves the reads according to user-specified criteria (the fragment length confidence interval) before re-sorting the reads in the order they occur on each reference sequence.
   c. MosaikAssembler convert the sorted alignment file to a multiple sequence alignment that is saved in an assembly file format. At the moment, MosaikAssembler saves the assembly in the phrap ace format and the GigaBayes gig format.
3. After the alignment, GIGABAYES can read the assembly output and generate SNP discovery results with the GFF format. GIGADUMP can convert the GFF format output to ace format output, which can be used by CONSED (-nophd option is required) for visualization. EAGLEVIEW (Huang and Marth, 2008) is another alternative visualization program for reading the results.

SSAHA

SSAHA (Ning et al., 2001) is developed for mapping large Sanger sequences data set to reference sequences with fast alignment algorithm by the Sanger Institute, which has been used in the HapMap project. The new version SSAHA2 has been developed to adapt the program to next generation sequencing reads including 454 and Illumina sequencing, and a variety of output formats are supported: SSAHA2, SAM, CIGAR, GFF, PSL, and so on. SNP discovery using SSAHA2 package also takes several steps:

1. ssaha2Build: builds a hash Table for reference sequences stored in the file as subject file;
2. ssaha2: maps next generation sequencing reads in the FASTQ format file as query file against the reference sequence;
3. ssaha2SNP: detects the SNPs and indels by aligning next generation sequencing reads to the reference sequences. The quality of base will also be considered with variation to reduce the false discovery rate, as well as the quality values in the neighboring bases.

Windows-Based Platforms for EST Assembly and SNP Identification

The programs introduced above all require the use of the Linux system, with which many users are not familiar. Several Windows-based programs are available for analysis of relatively small data sets based on a 32-bit Windows system. The programs developed for a 64-bit Windows system can take the advantage of large memory and will provide much power for large data sets. As most users are familiar with the Windows system, these programs may find their way for applications.

CLC Genomics Workbench

CLC Genomics Workbench is a very comprehensive package that integrates analysis functions for nucleotide and protein sequences such as sequence assembly, multiple alignment, BLAST searches, and gene expression analysis. It is especially well suited for next generation sequencing data analysis, including Roche/454, Illumina Genome Analyzer, and ABI SOLiD. The CLC Genomics Workbench has both Windows and Linux versions commercially available (32-bit and 64-bit).

SNP detection is one of the applications that could be used to identify putative SNPs from EST sequences or other short sequences generated from next generation sequencing platforms. The principle of SNP discovery is based on the sequence coverage and base quality score, which can be adjusted by users. Before the SNP discovery, the sequence files need to be imported through the "NGS import" function built within the software, which can recognize the raw sequence trace file (Sanger, Roche 454, Illumina Genome Analyzer, or ABI SOLiD), FASTA, FASTQ, or the ace file with the preassembly information. If the sequence file is utilized for SNP discovery, the *de novo* or reference assembly is required first. Once the assembly is finished, the SNP detection function can be directly performed under the Toolbox "High-throughput sequencing" built within the software.

The parameters for the SNP detection from sequences without any quality score is based on the sequences coverage of SNP position: (1) two minimum minor alleles for four or five sequences; (2) three minimum minor alleles for six to eight sequences; (3) four minimum minor allele for 9–11 sequences; (4) five minimum minor alleles for 12 sequences or more. The SNP detection application will generate the information related to the SNPs including, SNP position, allele variants, allele frequency, allele counts, sequences coverage, and alignment information (Figure 6.6).

As a demonstration project, we have used the CLC Genomics Workbench for the analysis of SNPs with carp transcript sequences generated with 454 sequencing. A total of 3157 SNP were identified from 242,261 Carp 454 sequencing reads (FASTQ format), which downloaded from the NCBI SRA database (SRX007427). The *de novo* assembly was conducted first and then SNP detection was directly applied on the assembly results. The whole SNP detection processes took approximately 30 min. Both sequence coverage and quality score were utilized for SNP detection. The quality score was set for Q20 for center base and Q15 for surrounding base.

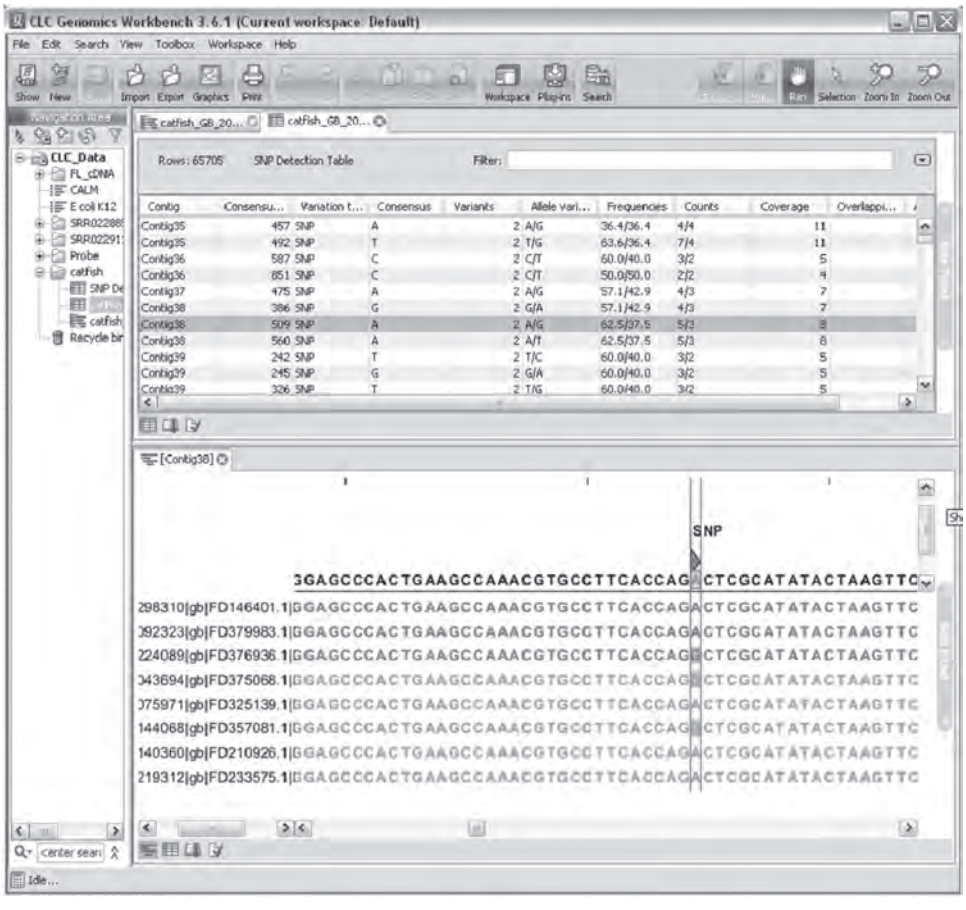

Figure 6.6 SNP visualization from the CLC Genomics Workbench. The left panel is a navigation area including all the files and results information. The right upper panel displays SNP summary information, for example, contig number, consensus sequence length, consensus base at the SNP site (majority rule), SNP allele bases, sequence count (count) and ratio (frequency), and total number of sequences (coverage). The right lower panel displays the sequences alignment information and the SNP sites of the selected contig. In this example, sequence alignments of contig 38 are given, with forward sequence being shown in red and reverse sequence being shown in green. See color insert.

NextGENe

NextGENe is a commercial available software package developed by SoftGenetics for the data analysis of next generation sequencing, including Roche 454, Illumina Genome Analyzer, and the ABI SOLiD system. The program has both 32-bit and 64-bit versions, and a 64-bit Windows system with at least a Quad Core processor and at least 8GB of RAM is recommend.

The NextGENe next generation sequences data analysis includes four steps: (1) select the instrument type (454, Illumina, and SOLiD) based on next generation

sequencing data and the application type (*de novo* assembly, SNP/Indel discovery, transcriptome, ChIP-seq, SAGE, and other); (2) load the sequences data (FASTA format only) to the software and define the results output folder, and the program also provides a tool to convert different sequences format to FASTA format; (3) select the condensation tool based on the sequence instrument type to correct the sequencing error and increase the assembly and alignment efficiency; (4) select algorithm and set up the parameters for the assembly and alignment based on the sequences data and instrument type.

For SNP/Indel discovery, the reference sequences are required for this application. If no reference sequences are available, the *de novo* assembly is required first. The SNP discovery process takes four steps: (1) select the instrument type and application "SNP/Indel"; (2) load next generation sequences and reference sequences generated during the *de novo* assembly; (3) select the default condensation tool; and (4) select the parameter for the SNP discovery by defining sequence coverage, minimum minor allele counts, and frequency. After the SNP discovery, the results will be displayed by the SequenceAlignment software (Figure 6.7), then go to drop-down menu "Report," select "Mutation Report," and the SNP information can be exported to a text file.

Figure 6.7 SNP visualization from NextGENe. The upper panel displays a global view of the project, and the lower panel displays the sequence alignment and sequence variation with SNPs highlighted in blue. See color insert.

Table 6.2 Comparison of different SNP identification applications.

Application	Input format			Output with SNP	Preassembly	Platform		Free
	Trace	FASTA	Ace			PC	Linux	
BLAST	—	Y	—	N/A	—	Y	Y	Y
CLUSTALW	—	Y	—	N/A	—	Y	Y	Y
POLYBAYES	R	R	R	Y	Y	—	Y	Y
POLYPHRED	R	R	R	Y	Y		Y	Y
AUTOSNP	—	Y	Y	Y	Y	—	Y	Y
CLC Genomics Workbench	Y	Y	Y	Y	NR	Y	Y	N
NextGENe	—	Y	—	Y	NR	Y	—	N

Y, yes; N, no; N/A, not available; R, required; NR, not required.

Summary

Identification of SNP from ESTs requires local alignment techniques that are unperturbed by exon–intron punctuation and alternatively spliced sequence variants. Once a multiple alignment is constructed, nucleotide differences among individual sequences can be analyzed. The programs BLAST and CLUSTALW were not originally designed for SNP identification, but these approaches can be utilized under certain scenarios. POLYBAYES, POLYPHRED, AUTOSNP, GIGABAYES, CLC Genomics Workbench, and NextGENe SNP detection were very powerful and efficient SNP identification platforms. POLYBAYES and AUTOSNP require the preassembly step for SNP identification. The trace files were necessary for the POLYBAYES and POLYPHRED, but not for AUTOSNP. A comprehensive comparison of all applications introduced in this chapter is listed in Table 6.2. Putative SNP discovery are not limited to the programs mentioned above; many customized SNP detection pipelines have been developed based on these applications. The intention of this chapter is to provide some basic methods and protocols for researchers who wish to discover SNPs.

Owing to the presence of sequencing errors, not every nucleotide position with mismatches automatically implies a SNP. The success of SNP projects depends heavily on the ability to discriminate true SNPs from sequencing errors, especially without trace file or sequence quality score. This is usually accomplished by statistical considerations that take advantage of measures of sequence accuracy accompanying the analyzed sequences (Marth et al., 1999). The result, ideally, should be a set of candidate SNPs, each with an associated SNP score that indicates the confidence of the prediction. The confidence values can be useful for researchers in selecting which SNPs to use for follow-up studies. The issues related to SNP quality assessment are discussed in Chapter 7.

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*, 215:403–410.

Barker G, Batley J, O'Sullivan H, Edwards KJ, and Edwards D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, 19:421–422.

Ewing B and Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8:186–194.

Ewing B, Hillier L, Wendl MC, and Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8:175–185.

Gordon D, Abajian C, and Green P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res*, 8:195–202.

Gordon D, Desmarais C, and Green P. 2001. Automated finishing with autofinish. *Genome Res*, 11:614–625.

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.

He C, Chen L, Simmons M, Li P, Kim S, and Liu ZJ. 2003. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. *Anim Genet*, 34:445–448.

Higgins DG and Sharp PM. 1988. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene*, 73:237–244.

Higgins DG, Thompson JD, and Gibson TJ. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*, 266:383–402.

Huang W and Marth G. 2008. EagleView: A genome assembly viewer for next-generation sequencing technologies. *Genome Res*, 18:1538–1543.

Huang X and Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Res*, 9:868–877.

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, and Gibson TJ. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem Sci*, 23:403–405.

Li R, Li Y, Kristiansen K, and Wang J. 2008. SOAP: Short oligonucleotide alignment program. *Bioinformatics*, 24:713–714.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, and Wang J. 2009. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*, 25:1966–1967.

Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, and Gish WR. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat Genet*, 23:452–456.

Moen T, Hayes B, Baranski M, Berg PR, Kjoglum S, Koop BF, Davidson WS, Omholt SW, and Lien S. 2008. A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. *BMC Genomics*, 9:223.

Nickerson DA, Tobe VO, and Taylor SL. 1997. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*, 25:2745–2751.

Ning Z, Cox AJ, and Mullikin JC. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res*, 11:1725–1729.

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, and Boyce-Jacino M. 1999. Mining SNPs from EST databases. *Genome Res*, 9:167–174.

Quinlan AR, Stewart DA, Stromberg MP, and Marth GT. 2008. Pyrobayes: An improved base caller for SNP discovery in pyrosequences. *Nat Methods*, 5:179–181.

Rausch T, Koren S, Denisov G, Weese D, Emde AK, Doring A, and Reinert K. 2009. A consistency-based consensus algorithm for *de novo* and reference-guided sequence assembly of short reads. *Bioinformatics*, 25:1118–1124.

Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, et al. 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res*, 18:1638–1642.

Thompson JD, Higgins DG, and Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–4680.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, and Higgins DG. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, 25:4876–4882.

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.

Zerbino DR and Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*, 18:821–829.