Mining for Single Nucleotide Polymorphisms in Expressed Sequence Tags of European Sea Bass

Souche E. L.[1], Hellemans B.[1], Van Houdt J. K. J.[1], Canario A.[2], Klages S.[3], Reinhardt R.[3] and Volckaert F. A. M.[1]

[1] Laboratory of Animal Diversity and Systematics, Katholieke Universiteit Leuven, Charles Deberiotstraat 32, B-3000 Leuven, Belgium
[2] Centre of Marine Sciences (CCMAR), University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal
[3] Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195 Berlin, Germany

Abstract

As a multitude of sequence data are published, discovering polymorphisms bioinformatically becomes a valid option. *In silico* Single Nucleotide Polymorphism (SNP) detection is based on the analysis of multiple alignments. Each column of an alignment is considered a slice containing one base of every sequence aligned. If a mismatch is detected, the slice is further analysed and the mismatch may be reported as a candidate SNP.
About 30,000 Expressed Sequence Tags (ESTs) of the fish European sea bass have been sequenced and processed. Since ESTs are redundant, they provide a resource for *in silico* SNP discovery. To prevent the detection of sequencing errors, a redundancy of two is chosen in order for a mismatch to be considered a candidate SNP. Among the various tools available to detect candidate SNPs, three software packages were tested: SNPServer, PolyBayes and PolyFreq. Candidate SNPs were validated in the laboratory by cloning and sequencing. From preliminary results PolyFreq outperforms both PolyBayes and SNPServer in terms of positive predictive value and SNPServer is the most sensitive tool. PolyFreq and SNPServer non-default identify respectively the fewest and highest number of candidates. Considering candidates detected by several tools seems to enhance both positive predictive value and sensitivity. Out of the 69 loci sequenced, only four were monomorphic, leading to a total of 91.3% polymorphic loci. Randomly chosen contigs will be sequenced to know whether SNP discovery tools tend to predict polymorphic fragments. Polymorphisms will be mapped, used for selection in aquaculture and the study of adaptation in natural populations.

1 Introduction

1.1 Single Nucleotide Polymorphisms

A genetic marker is a polymorphic DNA sequence that can be easily identified. Molecular markers originate from several kinds of mutations, which may occur as a result of normal cellular operations or interactions with the environment [1]. Base substitutions, insertions and deletions (indels) of nucleotide sequences within a locus, inversions of a segment of DNA within a locus and rearrange-ments of DNA segments at a locus constitute such markers.

Single Nucleotide Polymorphisms (SNPs) are point mutations occurring at the nucleotide level and producing single nucleotide differences among or within individuals of a species [2]. They are the most common in any organism, representing 90% of human variation [2] and are distributed throughout the genome [3]. They are an invaluable tool for genome mapping since they reveal hidden polymorphisms that cannot be discovered with other

markers or methods. SNPs have various impacts depending on their position within the genome; when located in a coding region, they can be synonymous or non-synonymous [4].

1.2 European Sea Bass

European sea bass (*Dicentrarchus labrax*) is an economically important marine fish for European aquaculture. However, the industry suffers from the absence of selection programs and a solid genetic background. Sea bass production is largely based on wild-caught brooders reproducing under semi-controlled conditions. After twenty years of large-scale production not a single domesticated stock has been generated [5]. Since SNPs are distributed throughout the genome and can be numerous [6], it would be interesting to incorporate this type of marker in the linkage map. Moreover SNPs located in coding regions can be used to genotype populations and to differentiate loci under selective pressure from neutral loci. The identification of such sites will provide insights in the evolutionary and functional biology of sea bass for application in aquaculture [3].

1.3 Aims

Classical molecular approaches of discovering SNPs are time and money consuming. DNA sequencing is the most accurate, the most commonly used and the most cost effective approach for SNP discovery, but requires prior knowledge of the genome [6-7]. On the other hand, more and more sequences are generated in order to discover new genes and study candidate genes [8]. It is thus feasible to use already generated sequences as a starting point for *in silico* SNP discovery [9]. Various tools using a range of strategies mine sequences for SNPs. Every new tool claims to outperform already existing tools but only a subset of candidate SNPs are molecularly validated [10-11-12]. Since about 30,000 sequences of European sea bass are available and since SNPs are in high demand, several SNPs discovery tools were compared, and their performance was evaluated.

2 Material and Methods

2.1 Data Description

ESTs are partial sequences of cDNA (complementary DNA) clones measuring several hundred nucleotides [13]. They are single-pass reads and have thus a high error rate (between 1 and 3 nucleotides out of 100 is expected to be wrong). Nevertheless they allow the discovery of SNPs in transcribed regions [14]. ESTs have to be processed to remove contaminating sequences (such as vector sequences) [15], to reduce their redundancy and to attempt to reconstruct the mRNA sequence they originate from [8]. ESTs representing the same mRNA transcript are pooled into a single group (or cluster) according to their similarity. For each cluster, ESTs are aligned against each other and assembled. A tentative consensus sequence or contig is built using ESTs overlaps. Ideally, each cluster represents one full mRNA sequence. However, the coverage of cDNA libraries is usually insufficient.

The European Network of Excellence Marine Genomics Europe developed 14 normalised cDNA libraries, corresponding to 14 distinct tissues, from 5 F_1 offspring from wild Atlantic parents. A total of 33,904 ESTs, of which 29,260 were of good quality, have been sequenced and processed at the Max Planck Institute for Molecular Genetics (MPI-MG). Of the 29,260 processed sequences, 55.1% (16,117 ESTs) were redundant and thus clustered; 44.9% (13,143 ESTs) remained singletons.

2.2 *In silico* SNP Detection

EST redundancy is highly advantageous for mining SNPs. If several ESTs from the same gene (or the same contig) have alignment mismatches, they may be SNPs. ESTs allow the detection of either homozygous or heterozygous SNPs because both alleles are present in the various clones of the cDNA libraries. To prevent the detection of sequencing errors, a redundancy of two was chosen for a given mismatch in order to be considered a candidate SNP [16-17]. Thus, only contigs containing four overlapping sequences or more were analysed. 975 such contigs were selected using a Perl script. However, one contig contained too many sequences (1028 ESTs): its analysis could lead to the detection of false positive candidate SNPs [18]. Therefore it was removed from the analysis.

From the clustering and assembly analysis, 974 (21.3%) contigs qualified for *in silico* SNP discovery, representing 5,548 (19%) ESTs and 477,224 overlapping base pairs.

2.2.1 SNPServer

SNPServer is an online tool developed by PGG Bioinformatics [20]: it uses the autoSNP algorithm [21], which is based on redundancy only. ESTs are clustered and assembled before processing (assembly parameters can be chosen). SNP discovery is ruled by five parameters related to five minimum redundancy scores (1 to 5). They correspond to the maximum number of sequences that must be part of an alignment for the mismatch to be considered as a candidate SNP. Default parameters are as follows: the maximum number of sequences being part of the alignment is 0, 4, 8, 12 and 20 for a minimum redundancy of respectively 1, 2, 3, 4 and 5. For example, a mismatch appearing twice will be considered as a candidate SNP in all alignments of 4 sequences but not in an alignment of 6 or 7 sequences. Therefore SNP discovery is ruled by redundancy and depth of alignment. Contigs of more than 50 reads were not processed since they may display a disproportionate number of potential SNPs [18], the mean number of candidate SNPs increasing with the number of sequences present in the contig.

2.2.2 PolyBayes

PolyBayes is the most commonly used tool for *in silico* SNP detection [22]. It contains three main functional parts: an anchored multiple sequence alignment algorithm (a reference sequence is requi-red; ideally the genomic sequence is used but the consensus sequence can also be used), duplicate sequence identification and a SNP detection algorithm. Each part may be skipped. The SNP disco-very is based on base called nucleotides and depth of alignment, as well as on quality values, base composition of ESTs and an *a priori* polymorphic rate. The latter is the probability that each nucleotide may be polymorphic. Its default value is 0.006, meaning that one permutation occurs every 166.7 base pairs. Variation probabilities are also assigned; they are the probabilities that each nucleotide may be permuted by each other one. Their default value is 0.1666. A Bayesian-statistical scheme assigns a posterior probability to discriminate real SNPs from sequencing errors. This score is the probability that a position is poly- or monomorphic. A mismatch is considered a candidate SNP if the posterior probability is higher than a threshold (which default value is 0.1).

2.2.3 PolyFreq

One limitation of SNP discovery tools is their tendency to detect more SNPs as the number of aligned sequences increases [18]. PolyFreq is a tool designed to handle this problem, allowing an efficient mining for SNPs in alignments containing many sequences [10]. It contains five programs having their own set of parameters. The first program (GAP3) aligns the anchor

sequences in order to find paralogous sequences. The second one (DDS2) detects pairs of similar query and anchor sequences, which are aligned by the third and fourth programs (FIL and GAP22). Highly similar regions are then found and screened for candidate SNPs using the PolyFreq program. SNP discovery is based on a minimum *a priori* polymorphic rate, a minimum depth and a minimum percent of identity to keep ESTs aligned; default values are 0.001, 100 and 0.97 respectively. Aligned nucleotides must have a quality value equal or greater than a cut-off (which default value is 20). Mismatches are considered candidate SNPs if the quality of the five base pairs flanking them has a good quality value and if no more mismatches are observed in a region of 20 base pairs.

2.3 SNP Validation

SNP validation of the candidate SNPs required the development of primers. The software package Primer3 [23] was used to design primers flanking each candidate SNP. The primers had to be at least 25 bp removed from the SNP, their GC content had to be between 40 and 60% and the product length ranged between 300 and 400 bp. Other parameters were set as default. Once primer pairs were designed, they were optimised on DNA of two sea bass individuals.

Genomic DNA was extracted from the five individuals used to produce the ESTs using the NucleoSpin Extraction kit (Machery-Nagel GmBH). Each locus was amplified through Polymerase Chain Reaction (PCR) in 25 µl reaction mixture: 1 µl DNA template, 0.8 µM forward and reverse primers, 1 or 2 mM $MgCL_2$ depending on the primer pair, 0.2 mM of dNTPs, 10xPCR buffer (Silverstar), 1 U Taq polymerase (Silverstar) and mQ H_2O. After initial denaturation, 3 min at 95 °C, amplification conditions were 35 cycles of denaturation at 95 °C for 30 sec, annealing at 48-56 °C and extension at 72 °C for 1 min. A final extension step of 7 min was carried out. Amplification products were cloned into plasmid vectors using the TOPO-TA cloning kit (Invitrogen). Plasmid DNA was extracted by isolating colonies in 100 µl of water, vortexing and heating at 96 °C for 3 min. This was used as a template (10 µl) in the following PCR reaction of 50 µl: 0.8 µM standard M13 forward and reverse primers, 1.5 mM $MgCL_2$, 0.2 mM dNTPs, 10xPCR buffer (Silverstar), 1 U Taq polymerase (Silverstar) and mQ H_2O. Eight positive clones were sequenced in one direction using the BigDye Terminator version 3.1 cycle sequencing kit (Applied Biosystems) and run on an ABI 3130-Avant sequencer (Applied Biosystems). Two software packages, Gap4 and PolyPhred, were used to align the traces obtained by sequencing, to visualise them and to detect SNPs. Genotyping the individuals used for SNP detection allowed the validation of the SNP discovery tools by calculating the number of candidate SNPs that turned out to be real SNPs and the number of candidate SNPs that turned out not to be SNPs.

3 Results

3.1 SNPServer

Two different sets of parameters were used. First, 232 candidate SNPs (of which 42 were indels) were proposed by SNPServer using the default parameters. Then less stringent parameters led to the detection of 929 candidate SNPs, of which 229 were indels. Each mismatch appearing twice in a slice, with a minimum number of sequences of four and a maximum number of sequences of 50, was considered a candidate SNP. All SNPs detected by SNPServer using default parameters were also detected when less stringent parameters were used.

The use of less stringent parameters led to the detection of almost five times the number of

SNPs detected using default parameters. Both sets of candidate SNPs were used for validation in order to evaluate the effect of stringent parameters.

3.2 PolyBayes

The use of default parameters led to the discovery of 5,870 candidate SNPs, of which 736 were indels. Since the aim of using different SNP discovery tools was to compare their performance, only candidate SNPs appearing at least twice in an alignment of a minimum of four ESTs were studied. The selection of these candidate SNPs, called redundant SNPs, reduced the number of detected SNPs to 734 (541 mutations and 193 indels).

Other parameters were tested in order to check their influence on the large number of detected SNPs. The posterior probability threshold, the *a priori* polymorphic rate and the quality value threshold were modified, as shown in Fig 1. PolyBayes parameters are displayed as follows: the first parameter is the posterior probability threshold, the second the *a priori* polymorphic rate and the third the quality value threshold. Modification of these parameters had a clear influence on the total number of detected SNPs, the maximum number of candidate SNPs being 5,870 using default parameters and the lowest one being 2,531 using more stringent parameters (posterior probability threshold = 0.1, *a priori* polymorphic rate = 10^{-6}, and quality value threshold = 30). However there was no clear influence of the parameters on the number of redundant candidate SNPs that is to say the candidate SNPs to be validated in this study. Default parameters were thus used.

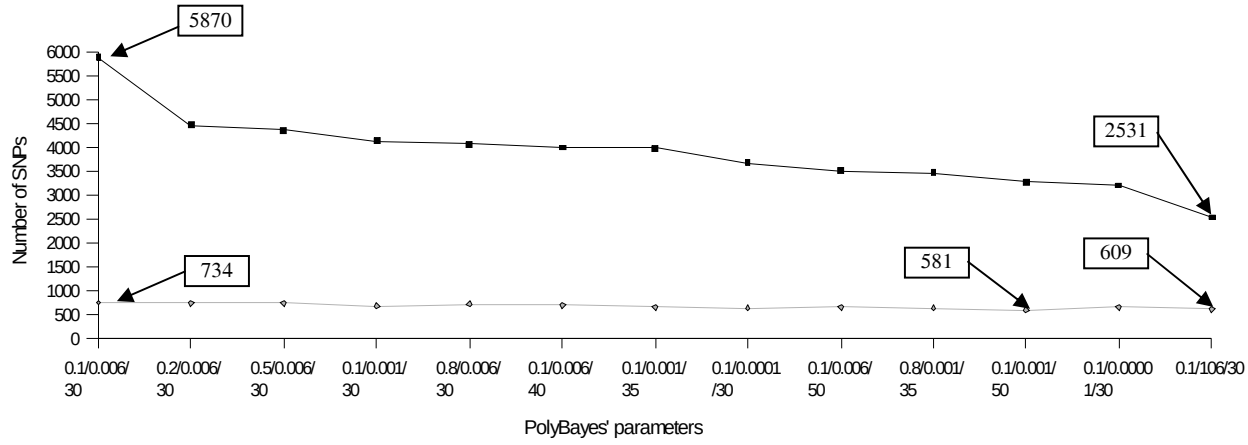

Fig 1: Number of detected SNPs given PolyBayes sets of parameters. ■ Number of SNPs detected by PolyBayes. ◇ Number of redundant SNPs detected by PolyBayes

3.3 PolyFreq

The dataset does not contain any contig of deep coverage. Using PolyFreq's default parameters, only contigs containing more than 100 sequences could be analysed. Consequently, no SNP could be detected. The minimum depth was modified and set to four: all alignments containing at least four sequences were mined for SNPs. All the other parameters were set as default. PolyFreq identified 2,002 SNPs of which two were indels. In order to compare tool performance, only candidate SNPs being present twice were further analysed. The selection of these rendundant candidate SNPs led to a total number of 219 SNPs of which two were indels.

3.4 Tool Comparison

A total of 1,059 unique candidate SNPs were detected by SNPServer (default and non-

default), PolyBayes and PolyFreq. Only 50 candidate SNPs (4.7%) were detected by all tools and sets of parameters (Table 1 in supplementary material). No SNP was unique to SNPServer default since all its candidates were also detected by SNPServer non-default. Only two candidate SNPs were unique to PolyFreq. PolyBayes and SNPServer non-default identified most of the SNPs which were detected just once. Indeed the number of candidates they predicted was three to four times higher than the number of SNPs detected by SNPServer default and PolyFreq. The repartition of candidate SNPs by the number of tools by which they were detected is given in Fig 2a.

Three kinds of SNPs were compared: transitions, transversions and indels (Fig 2b and Table 2 in supplementary material). PolyBayes and SNPServer (default and non-default) discovered a similar proportion of mutations, which is comparable with earlier *in silico* SNP discovery studies [17-18-24]. Indeed, transitions counted for about 50% of detected SNPs, transversions for 25 to 30%, and indels for 20 to 25%. PolyFreq showed different proportions since only two indels were detected. More transitions were detected (68%); transversions counted for 31% and indels for 1%. The proportions of mutations actually found in the sea bass genome are 57% of transitions, 36% of transversions and 7% of indels. PolyBayes and SNPServer tended to overestimate the number of indels whereas PolyFreq overestimated the number of transitions and underestimated the number of indels.

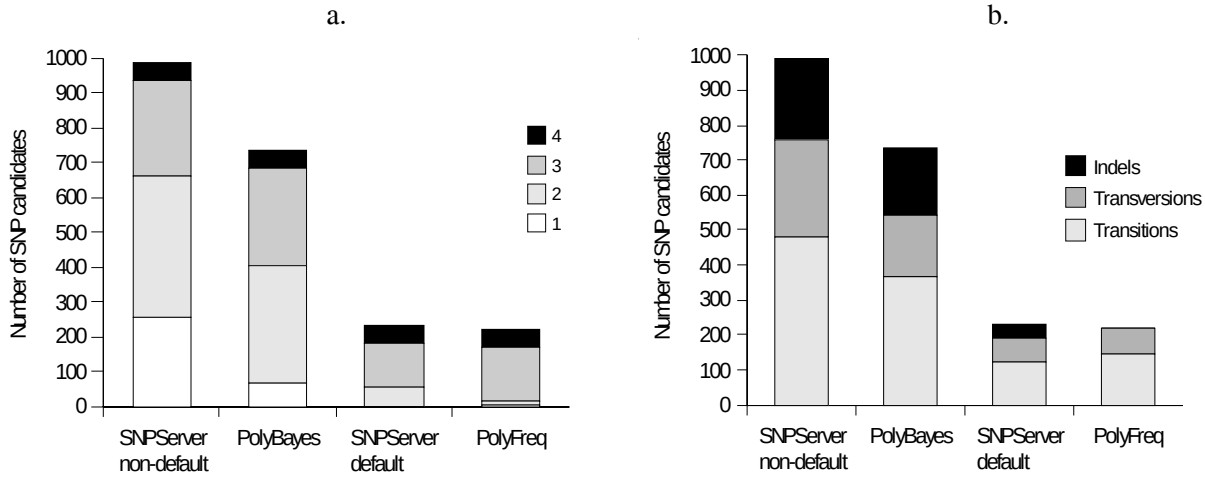

Fig 2: Repartition of the candidate SNPs by tool. a. Repartition by the number of tools by which SNP candidates are detected. b. Repartition by the number of transitions, transversions indels that are detected.

3.5 Molecular Validation

So far 69 loci were cloned and a total of 181 SNPs detected. Sixty three (91.3%) of the selected loci were polymorphic. Out of the 112 tested SNPs, only 59 (52.7%) turned out to be real SNPs (or true positives, TP) (Table 3 supplementary material). The remaining 53 candidate SNPs turned out not to be SNPs (or false positives, FP). PolyFreq outperformed PolyBayes and SNPServer by predicting 21 true positive SNPs out of 25 candidates. This represented a positive predictive value (PPV) of 84%. However the number of SNPs detected by other tools but missed by PolyFreq (false negative, FN) was the highest. ESTs are cloned transcripts randomly sequenced. It may happen that one individual is sequenced several times at a given position when other individuals are not sequenced. If only one individual out of five is polymorphic at one locus and not sequenced at that locus, the SNP will be missed. Moreover a mismatch was considered a candidate SNP if it was present at least twice in the dataset. This limited the number of real SNPs that could be detected. Using real data, the

number of false negative is underestimated since every SNP present in the data set but missed by all tools is not considered a false negative. The sensitivity (Sens) of PolyFreq was thus the lowest: only 36% of the SNPs validated were discovered by PolyFreq. The probabilities were calculated as follows:

$$PPV = \frac{TP}{TP + FP} * 100 \qquad Sens = \frac{TP}{TP + FN} * 100$$

When non default parameters were used, SNPServer detected more true positives than when default parameters were used. However, more false positives were detected. The positive predicted value and sensitivity of SNPServer non-default were 56% and 97% respectively compared to 63% and 78% for SNPServer default. Finally, PolyBayes seemed to be a compromise between positive predictive value and sensitivity with intermediate values of respectively 66% and 88%. These results are summarised in Fig 3a.

One additional test was made by considering only candidate SNPs detected by several tools. The results are shown in Fig 3b (Table 4 in supplementary material). Using SNPs detected by three and four tools seemed to increase performance. A total of 60 candidate SNPs were detected and 46 of them were true positives. The positive predictive value is then 77% and the sensitivity 78%.

The 69 tested loci were annotated and their amino acid sequence retrieved. Out of the 112 tested SNPs, 41 (36.6%) were situated in amino acid sequences. The number of synonymous (dS) and non-synonymius (dN) SNPs, as well as the ratio dN/dS was calculated for each tool (Table 5 in supplementary material). While dS was well predicted, the dN was overestimated. The observed ratio dN/dS is about twice as low as the predicted one.

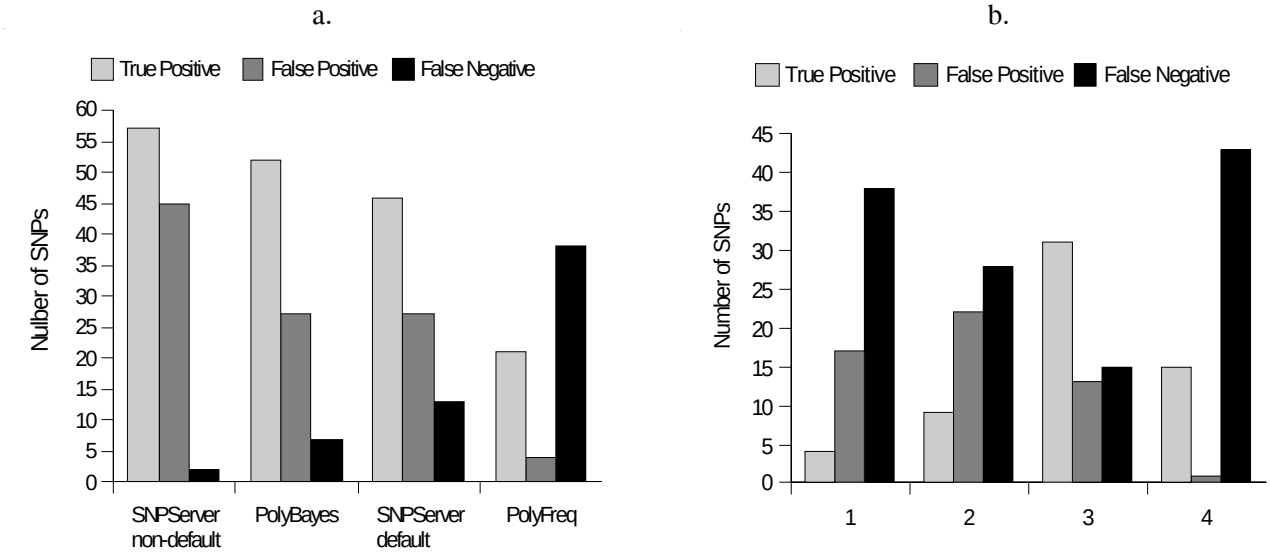

Fig 3: Performance of tools mining for SNPs. a. Number of true positives, false positives and false negatives given by tool. b. Number of true positives, false positives and false negatives given the number of tools used for SNPs detection.

4 Discussion

As more sequence data are generated, computational extraction of relevant information from these sequences becomes an efficient strategy. In this study we checked the efficiency of *in silico* SNP discovery in European sea bass ESTs using three software packages. A total of 112

candidate SNPs in 69 loci were validated by direct sequencing. Only half of them appeared to be true SNPs. PolyFreq detected less false positives than PolyBayes and SNPServer, but the number of candidates detected by PolyFreq was very low. This is due to the fact that PolyFreq has been designed to reduce the number of candidate SNPs detected in sequences of deep coverage. From the literature records PolyFreq detected less false positives than PolyBayes [10]. However, the number of true positives and false negatives were reported to be similar whereas they were respectively lower and higher with PolyFreq than with PolyBayes. The fact that the number of candidates detected by PolyFreq and tested in the laboratory was about three times less than the number of candidates tested for PolyBayes may explain this difference. The number of confirmed true positives while using PolyBayes was low compared to what has been described for example in maritime pine (*Pinus pinasterspecies*) ESTs where PolyBayes detected 83.1% of true SNPs [24]. The ESTs originated from five cDNA libraries using more than 350 individuals. The number of individuals used while mining for SNPs might be a critical factor. A filtering strategy based on substitution redundancy and good quality values detected 63% of true SNPs in 10% of the candidate SNPs of human ESTs from 19 cDNA libraries [17]. However, other studies detected lower percentages of true positives, the lowest being 7.8% with PolyBayes in soybean sequences [25] and 8% with autoSNP in human ESTs [11]. Both of them tried to increase this number either by using another tool or an additional selection.

The total number of SNPs found in ESTs was about three times higher than the number of SNPs detected computationally. This is partly due to the criterion of redundancy. Moreover some true SNPs have been missed since automated SNP discovery tends to fail detection of less common alleles [24].

Not all candidate SNPs have been molecularly validated, but preliminary results suggest that PolyFreq gives the best positive predictive value, outperforming PolyBayes and SNPServer whatever the settings. PolyBayes outperformed SNPServer. Nevertheless the majority of SNPs detected by PolyFreq (89%) and PolyBayes (87%) were not redundant and thus not taken into account for further evaluation. Some of the confirmed SNPs detected by SNPServer were not detected by the two other tools; a number of SNPs must have been missed. This can be assessed by calculating the sensitivity of each tool; SNPServer non-default had the highest sensitivity and PolyFreq the lowest. A good SNP discovery tool has to have both a high positive predictive value and a high sensitivity; PolyBayes outperforms the other tools. However, SNP detection can be further enhanced by considering candidate SNPs detected by three and four tools.

Through the detection of candidate SNPs, software packages such as SNPServer, PolyBayes and PolyFreq predict which contigs are polymorphic. Molecular validation showed that 91.3% of the contigs supposedly containing SNPs were actually polymorphic, even if the expected SNPs did not turn out to be real ones. In case this percentage is lower when sequencing contigs randomly, SNP discovery tools would select polymorphic contigs more accurately than candidate SNPs.

Acknowledgements

This research was funded by the European FP6 network of excellence Marine Genomics Europe (contract no. GOCE- CT-2004-505403) and EU FP6 STREP project Aquafirst (contract no. 513692). The EST dataset was provided by Marine Genomics Europe. We also thank G. T. Marth for access to PolyBayes; D. Edwards for SNPServer; P. Green, B. Ewing, D. Gordon and M. Stephens for PHRED, PHRAP, CONSED and PolyPhred; the Staden Package for the pre GAP4 and GAP4.

References

[1] Z. J. Liu and J. F. Cordes. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238:1-37, 2004.

[2] R. T. Brumfield, P. Beerli, D. A. Nickerson and S. V. Edwards. The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology and Evolution, 18:249-256, 2003.

[3] P. A. Morin, G. Luikart, R. K. Wayne and The SNP Workshop Group. SNPs in ecology, evolution and conservation. Trends in Ecology and Evolution, 19:208-216, 2004.

[4] A.-C. Syvänen. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature Reviews Genetics, 2:930-942, 2001.

[5] D. A. Chistiakov, B. Hellemans, C. S. Haley, A. S. Law, C. S. Tsigenopoulos, G. Kotoulas, D. Bertotto, A. Libertini and F. A. M. Volckaert. A Microsatellite Linkage Map of the European Seabass *Dicentrarchus labrax* L. Genetics, 170:1821-1826, 2005.

[6] A Vignal, D. Milan, M. Sancristobal and A. Eggen. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics, selection, evolution, 34:275-305, 2002.

[7] P.-Y. Kwok and X. Chen. Detection of Single Nucleotide Polymorphisms. Curr. Issues Molecular Biology 5:43-60, 2003.

[8] H. Xu, L. He, Y. Zhu, W. Huang, L. Fang, L. Tao, Y. Zhu, L. Cai, H. Xu, L. Zhang, H. Yu, and Y. Zhou. EST Pipeline System: Detailed and Automated EST Data Processing and Mining. Genomics, Proteomics & Bioinformatics 1:236-242, 2003.

[9] F. J. Useche, G. Gao, M. Hanafey and A. Rafalski. High-Throughput Identification, Database Storage and Analysis of SNPs in EST Sequences. Genome Informatics, 12:194–203, 2001.

[10] J. Wang and X. Huang. A method for finding single-nucleotide polymorphisms with allele frequencies in sequences of deep coverage. BMC Bioinformatics, 6, 2005.

[11] J. Tang, B. Vosman, R. E. Voorrips, C. G. Van Der Linder and J. AM Leunissen. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics, 7, 2006.

[12] S. Weckx, J. Del-Favero, R. Rademakers, L. Claes, M. Cruts, P. De Jonghe, C. Van Broeckhoven and P. De Rijk. novoSNP, a novel computational tool for sequence variation discovery. Genome Research 15:436-422, 2005.

[13] A. D. Baxevanis and B. F. F. Ouelette. Bioinformatcs A practical guide to the analysis of genes and proteins, second edition. John Wiley & Sons, Inc, New York, 2001.

[14] G. T. Marth. Computational SNP discovery in DNA sequence data. P. Y. Kwok (ed.), Single Nucleotide Polymorphisms, Methods and Protocols, New Jersey, 2003.

[15] H.-H. Chou and M. H. Holmes. DNA sequence quality trimming and vector removal. Bioinformatics 17:1093-1104, 2001.

[16] D. G. Cox, C. Boillot and F. Canzian. Data Mining: Efficiency of Using Sequence Databases for Polymorphism Discovery. Human mutation, 17:141-150, 2001.

[17] L. Picoult-Newberg, T. E. Ideker, M. G. Pohl, S. L. Taylor, M. A. Donaldson, D. A. Nickerson and M. Boyce-Jacino. Mining SNPs From EST Databases. Genome Research, 9:167-174, 1999.

[18] J. Batley, G. Barker, H. O' Sullivan, K. J. Edwards and D. Edwards. Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. Plant Physiology 132:84-91, 2003.

[19] B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. G. Muller, T. Wetter and S. SUHAI. Using miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. Genome Research 14:1147-1159, 2004.

[20] D. Savage, J. Batley, T. Erwin, E. Logan, C. G. Love, G. A. C. Lim, E. Mongin, G. Barker, G. C. Spangenberg and D. Edwards. SNPServer: a real-time SNP discovery tool. Nucleic Acids Research, 33:493-495, 2005.

[21] G. Barker, J. Batlet, H. O' Sullivan, K. J. Edwards and D. Edwards. Redundancy based detection of

sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics Oxford Journals, 2002.

[22] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, LaD. Hillier, P. Y. Kwok and W. R. Gish. A general approach to single-nucleotide polymorphism discovery. Nature Genetics 23:452-456, 1999.

[23] S. Rozen, http://frodo.wi.mit.edu/primer3/primer3_code.html, 2006.

[24] L. Le Dantec, D. Chagné, D. Pot, O. Cantin, P. Garnier- Géré, F. Bedon, J.-M. frigerio, P. Chaumeil, P. Léger, V. Garcia, F. Laigret, A. De Daruvar and C. Plomion. Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. Plant Molecular Biology 54:461-470, 2004.

[25] L. K. Matukumalli, J. J. Grefenstette, D. L. Hyten, I. Y. Choi, P. B. Cregan and C. P. Van Tassell. Application of machine learning in SNP discovery. BMC Bioinformatics, 7, 2006.