Improved conversion rates for SNP genotyping of nonmodel organisms

Darrell Conklin[1,2], Iratxe Montes[3], Aitor Albaina[3], and Andone Estonba[3]

[1] Department of Computer Science and Artificial Intelligence
University of the Basque Country UPV/EHU, San Sebastian, Spain
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[3] Department of Genetics, Physical Anthropology and Animal Physiology
University of the Basque Country UPV/EHU, Leioa, Spain

Abstract. Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variation and are highly adaptable to large-scale automated genotyping and population genetics studies. For nonmodel organisms, many SNP discovery projects are based on sequencing and assembly of a transcriptome and the calling of sequence variation in contigs. This paper develops a new method for avoiding intron/exon boundaries in genotyping primer design, based on the mapping of raw genome reads to a transcriptome assembly. Applied to a project in European anchovy SNP discovery, the attained conversion rate of 91.6% is the highest yet reported for a nonmodel teleost SNP discovery project.

Keywords: de novo sequencing, assembly, SNP discovery, SNP genotyping, conversion rate

1 Introduction

Next generation sequencing (NGS) technologies have led to a revolution in population genetics with the potential of high-throughput discovery of DNA sequence variation. Our project is concerned with the discovery and use of single nucleotide polymorphisms (SNPs) in the European anchovy (*Engraulis encrasicolus*) for traceability of geographic origin and understanding of population evolution. The research in *E. encrasicolus* was initiated to understand the causes of the recruitment failures in the Bay of Biscay and to assess genetic diversity. The scarcity of available genetic information for this species has limited the potential for genetic marker development.

Many approaches to SNP discovery rely on deep sequencing of a transcriptome, alignment of the transcript reads (either by mapping to a reference genome sequence or to a *de novo* transcriptome assembly), and attempted separation of true sequence variation from sequencing errors. Prior to their use in genetics studies, discovered SNPs must be validated as polymorphic markers.

SNP validation may be performed experimentally by SNP genotyping arrays which can assay thousands of individuals on panels of up to 256 SNPs [1]. A standard measure for quantifying success on genotyping array design is the SNP *conversion rate*. This is the proportion of all genotyped SNPs which show clear genotyping clusters, and therefore enough power for genotype discrimination. More precisely, SNPs with genotyping errors can be grouped into two categories: *failed*, SNPs for which no amplification has been detected; and *disperse*, where amplification is detected but no clear genotyping clusters are detected. The conversion rate is defined as the proportion of successful SNPs in an experiment. It is important to distinguish the converted SNPs, which may include monomorphic SNPs and apparently heterozygous SNPs within paralogous sequence variants (PSV) [2], from the *validated* SNPs which are those verified to be polymorphic in a population. The primary concern of this paper is with improving conversion rates.

Due to the technologies of SNP genotyping arrays, to minimize genotyping failures it is imperative to avoid intron/exon boundaries (IEB) both within the genotyping primers and probe and indeed anywhere within the expected amplified PCR product: this is due either to prohibited primer annealing or an expected product too large for amplification [3]. In an experiment with genotyping 374 EST-derived SNPs, Wang et al. [4] suggested that 64% of genotyping failures may have been due to the proximity of SNPs to IEB, and other studies have reached similar conclusions [3, 5, 6].

For nonmodel organisms such as the European anchovy and many fishes, where a complete reference genome does not exist, the location of IEB must be computationally inferred. The most common strategy is the IEB by homology approach: BLASTing transcripts to a closely related annotated genome and inferring analogous IEB locations based on coordinates of significant matches to a presumed orthologous gene.

Using the IEB by homology approach, Studer et al. [7] achieved a notably high conversion rate of 77% on SNP discovery in an assembly of perennial ryegrass (*Lolium perenne L.*) ESTs. The location of IEB was inferred by analogy with the rice (*Oryza sativa*) genome. To quantify the relatedness between the two genomes, we have independently performed an assembly of $\approx 33,000$ ryegrass ESTs with the CAP assembler [8] of all ryegrass ESTs. Subsequent BLASTn against the rice genome showed that $\approx 60\%$ of the assemblies have a significant match and therefore presumably a detectable orthologous gene. Similarly, for the Atlantic herring *Clupea harengus* a sample of 59 mRNA BLASTned against genomes of bony fishes (*Teleostei*) yielded 35 with a significant match. Thus the homology approach used by Helyar et al. [5] is expected to have some utility in inferring IEB in the Atlantic herring.

For the anchovy genome, the picture is entirely different: in a sample of 1000 transcriptome assemblies created from 454 reads (see Methods), $\approx 90\%$ had no significant BLASTn match to any fish genomic sequence. Therefore for such nonmodel organisms a new approach is required. This paper develops a new approach based on genome read mapping which can be applied when the transcriptome reads are supplemented by short sequence reads from the genome of the same organism.

Table 1. A sample of approaches to SNP discovery in teleost species showing their genotyping conversion rate.

Study	Organism	Sequences	IEB method	Converted	Total	Rate
Wang et al. [4]	catfish	EST	none	266	384	69.3
Roberts et al. [3]	herring	454	none	46	96	47.9
Helyar et al. [5]	herring	454	homology	779	1536	50.7
Moen et al. [9]	cod	EST	none	410	594	69.0
Milano et al. [6]	hake	454/GAII	homology	705	1628	43.3
Hubert et al. [10]	cod	GAII	none	2291	3072	74.6
this study	anchovy	454/Hiseq	read mapping	427	466	91.6

Table 1 shows the conversion rates on a sample of genotyping studies for nonmodel fishes, including the results reported in this paper, showing their sequencing platform and IEB detection method.

2 Methods

2.1 Sequencing, assembly, and putative SNP discovery

The IEB detection method has been applied in a SNP discovery pipeline, using an assembly created from $\approx 800,000$ European anchovy reads (average trimmed read length 275bp) from 10 pooled individuals using the Roche 454 GS-FLX, supplemented with $\approx 1.5 \times 10^9$ 100bp genome reads from the Illumina HiSeq 2000 for the same 10 individuals. Both 454 and HiSeq reads were mapped back to the assembly using bowtie2 (local, k=2) [11] and the resulting BAM files were processed efficiently using the Perl Bio-Samtools library (v1.36).

Trimmed 454 reads were assembled using the Assembler v2.6 software ("Newbler", 454 Life Sciences), and subsequently all contigs containing a statistically extreme number of reads per contig were removed from further consideration as they were assumed to represent incorrectly assembled reads or PSVs.

All locations of sequence variation in the 454 read assembly, representing a large number of potential SNPs, were then enumerated by samtools "mpileup" and the bcftools package [12]. A sequence of stringent logical filters were designed to reveal only those SNPs in positions of specified coverage and specified minor allele frequency (MAF). A SNP was called only if it appeared as a variant in both the 454 and the HiSeq read bowtie mappings. SNPs within any contig containing a read that maps also to another contig were removed as the two contigs could be PSVs.

2.2 IEB detection

The key observation of our method is that the alignment of short genome reads to a transcriptome assembly produces a statistical signal at areas of IEB. These signals are seen at positions on the reference sequence where a large number of genome reads initiate or terminate an alignment with the assembled transcript sequence. The genome reads will retain intron sequence and a location of discontinuity may be seen in the genome read mappings. To illustrate this, Figure 1 depicts a portion of an assembly and associated HiSeq reads, in the Tablet assembly visualizer [13], with a clear IEB visible in the centre of the figure, and additional IEB indicated by arrows.

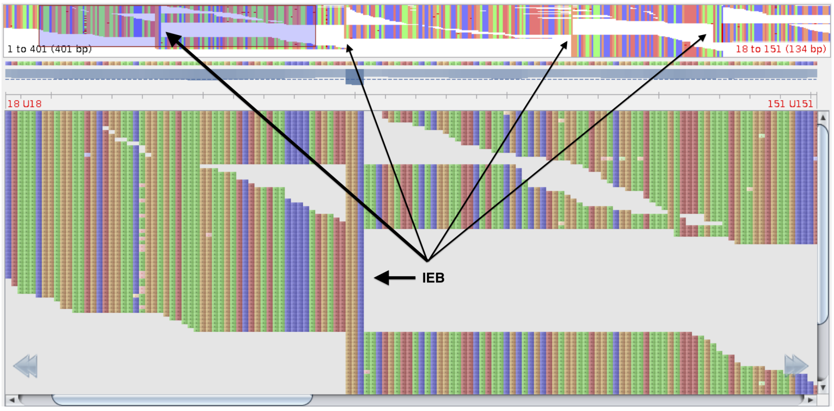

Fig. 1. An illustration of an IEB signal produced by aligning HiSeq reads to a 454 read contig.

For a given reference position the *change point count* is defined as the number of reads where a segment alignment starts or ends at that position. Consider a position with change point count k at a reference position with total coverage C. The p-value (probability of a change point count of k or more at the position) is computed using the cumulative Poisson distribution:

$$p(X >= k) = 1 - p(X < k) = 1 - e^{-\lambda} \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} \qquad (1)$$

with $\lambda = 2/100 \times C$ the expected number of change points at the position (each read length 100bp, and 2 possible change points in each read). Low p-values are suggestive of an IEB at the reference position.

The change point counts can be efficiently tabulated by iterating over the reads mapped to a contig, rather than over every position of the contig. This is because a change point can only occur at a position where a read begins or ends its alignment to the contig. Therefore a key point in the efficiency of the algorithm is the avoidance of iteration through every position within all assembly contigs.

2.3 Anchovy samples

For SNP validation, six sampling points, comprising six anchovy populations, were selected. Following [14], each sampling point corresponded to one anchovy population, comprising the whole species genetic variability. From each locality 30 individuals were sampled and stored in 99% ethanol at 20°C until DNA extraction was performed. Genomic DNA was extracted from 50 to 75 mg of anchovy muscle tissue using NucleoSpin® 96 Tissue kit (Macherey-Nagel). The amount and quality of DNA from each sample was subsequently quantified in a NanoDrop ND-8000 spectrophotometer (Thermo Fisher Scientific).

2.4 Genotyping

The 180 achovies from the six sampling points were screened for SNPs with TaqMan® OpenArray® Genotyping System (Life Technologies). DNA concentrations and reactions for amplification and detection of the SNPs were established according to the TaqMan® OpenArray® Genotyping System User Guide. Finally, genotypes were scored using TaqMan® Genotyper Software v1.2 (Life Technologies).

After default clustering was performed, data was viewed in the scatter plot and genotype calls were reviewed and manually adjusted for producing the final cluster assignments. Based on these assignments, SNPs were classified as failed (no amplification), dispersed (less than 80% of individuals assigned to a cluster), monomorphic (MAF < 0.01), polymorphic (MAF ≥ 0.01), and PSV ($\geq 99\%$ of all individuals heterozygous).

3 Results

3.1 Sequencing and assembly

The 454 reads were assembled into $\approx 25,000$ contigs, of which $\approx 5,000$ contigs were filtered due to excessive read count as described in Methods. The filtered assembly contains $\approx 10^9$ reference bases. Bowtie mapping of the all 454 reads for all individuals to this assembly produces a coverage of ≈ 15 (≈ 77 for the Hiseq read mappings) reads per reference base, as computed by the samtools "depth" command.

3.2 Putative SNP discovery

In the $\approx 20,000$ filtered contigs, a total of $\approx 210,000$ locations of sequence variation were enumerated as described in Methods. For SNP discovery a selective p-value threshold of 10e-9 (Eqn. 1) was chosen by visual identification of IEB within 100 randomly selected transcript assemblies. Approximately $2,000$ contigs contain one or more predicted IEB.

Genotyping arrays were ordered from Life Technologies for 466 SNPs. In this set, 145 contained one or more predicted IEB. For SNPs in these IEB-containing contigs which were not within 35bp of a predicted IEB, the sequence presented for genotyping primer design was clipped from the 5′ IEB (or, start of sequence) to the 3′ IEB (or, end of sequence). To reduce the possibility of non-specific genotyping primers, sequences were also masked in all other positions of sequence variation in either the 454 read alignments or the HiSeq read alignments.

3.3 Genotyping results

The 180 anchovy individuals were screened for 466 SNPs on the TaqMan® OpenArray® Genotyping System (Life Technologies). Figure 2 shows an example of four different outcomes from SNP genotyping. Each individual is represented as a point in the plot and colours represent cluster assignments. Red individuals are homozygous for allele 1, green individuals are heterozygous for both allele 1 and 2, dark blue means homozygous for allele 2, yellow individuals have not amplified, black points are inconsistent with the call and are assigned to an unknown, and light blue are NTC (no-template control) or negative controls (water).

As illustrated by Table 2, 427 SNPs (conversion rate 91.6%) yielded a signal and clear genotyping clusters. Only 39 (8.4%) SNPs were failed or dispersed. In the set of 427 converted SNPs, 136 are SNPs from contigs that contain one or more predicted IEB. Of the converted SNPs, further detail is given regarding their validation as polymorphic SNPs. Thus a total validation rate of $385/466 = 82.6\%$ has been achieved by our methods.

Table 2. Detail on genotyping results of 466 anchovy SNPs.

Category		Number of SNPs
converted		427 (91.6%)
	PSV/CNV	2
	monomorphic (MAF < 0.01)	40
	polymorphic	385
failed		39 (8.4%)
	no signal	11
	disperse	28

4 Discussion and conclusions

This study has produced the highest conversion rates yet reported for teleost SNP discovery, and importantly has provided a source of validated SNP markers to be used in future studies. The IEB detection is central to the success: within the converted set of 427 SNPs, assuming that all SNPs from IEB-containing contigs would fail, the conversion rate would drop to $291/427 = 62.4\%$, which is

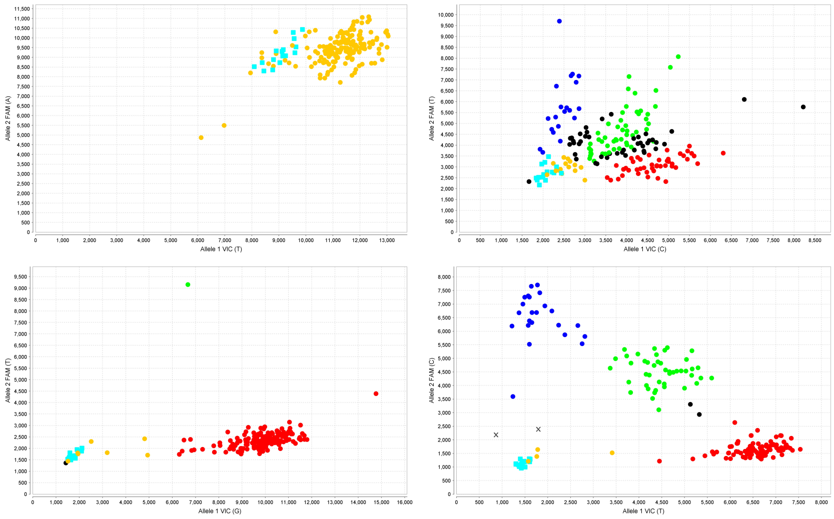

Fig. 2. Four genotyping results. Top left: no amplification; top right: disperse with no clear genotyping clusters; bottom left: a converted but monomorphic SNP; bottom right: a converted and validated (polymorphic) SNP.

a figure in line with other SNP discovery approaches (see Table 1) that do not employ an IEB detection method.

The alternative IEB by homology approach employed until now by SNP discovery projects in nonmodel organisms has the weakness that between divergent species it is difficult to detect BLAST matches at the nucleotide level and therefore to separate true orthologs from inter-species paralogs. Therefore obtaining a high recall rate of IEB can be difficult with the homology approach.

This paper has shown that the availability of raw genome reads, which have not necessarily been assembled into a high quality reference sequence, permits the detection of IEB within transcriptome assemblies. Future work will study, by simulation on an annotated genome, how the precision and recall of IEB detection varies with the coverage provided by the genome reads. The p-value threshold for IEB detection can thereby be calibrated to particular projects and the method applied to other nonmodel organism transcriptomes.

Acknowledgments This research was supported by the Ikerbasque Foundation, by the project ECOGENBAY (MICINN CTM2009-13570-C02-02) funded by the Ministry of Science and Research of the Government of Spain, by the project BERDEPOP (SAIOTEK S-PE12UN098) funded by the Basque Government, and by a Research Grant (3571/2008) from the University of the Basque Country UPV/EHU. We are indebted to the Sequencing and Genotyping Service (University of the Basque Country UPV/EHU) for their technical assistance.

Bibliography

[1] J Seeb, G Carvalho, L Hauser, K Naish, S Roberts, and L Seeb. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, 11:1–8, 2011.

[2] D Fredman, S White, S Potter, E Eichler, J Den Dunnen, et al. Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics*, 36(8):861–866, 2004.

[3] S Roberts, L Hauser, L Seeb, and J Seeb. Development of genomic resources for pacific herring through targeted transcriptome pyrosequencing. *PLoS ONE*, 7(2):e30908, 02 2012.

[4] S Wang, Z Sha, T. S Sonstegard, H Liu, P Xu, B Somridhivej, E Peatman, H Kucuktas, and Z Liu. Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics*, 9(1):450, 2008.

[5] S Helyar, M Limborg, D Bekkevold, M Babbucci, van Houdt J, et al. SNP discovery using Next Generation transcriptomic sequencing in Atlantic Herring (*Clupea harengus*). *PLoS ONE*, 7(8):e42089, 2012.

[6] I Milano, M Babbucci, F Panitz, R Ogden, R Nielsen, et al. Novel Tools for Conservation Genomics: Comparing Two High-Throughput Approaches for SNP Discovery in the Transcriptome of the European Hake. *PLoS ONE*, 6 (11):e28008, 11 2011.

[7] B Studer, S Byrne, R. O Nielsen, F Panitz, C Bendixen, M. S Islam, M Pfeifer, T Lübberstedt, and T Asp. A transcriptome map of perennial ryegrass (*Lolium perenne L.*). *BMC Genomics*, 13:140, 2012.

[8] X Huang and A Madan. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9:868–877, 1999.

[9] T Moen, B Hayes, F Nilsen, M Delghandi, K Fjalestad, S.-E Fevolden, P Berg, and S Lien. Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *BMC Genetics*, 9(18), 2008.

[10] S Hubert, B Higgins, T Borza, and S Bowman. Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics*, 11:191, 2010.

[11] B Langmead and S Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359, 2012.

[12] H Li, B Handsaker, A Wysoker, et al. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25:2078–2079, 2009.

[13] I Milne, M Bayer, L Cardle, P Shaw, G Stephen, F Wright, and D Marshall. Tablet - next generation sequence assembly visualization. *Bioinformatics*, 26(3):401–402, 2010.

[14] I Zarraonaindia, M Iriondo, A Albaina, M Pardo, C Manzano, W Grant, X Irigoien, and A Estonba. Multiple SNP Markers Reveal Fine-Scale Population and Deep Phylogeographic Structure in European Anchovy (*Engraulis encrasicolus L.*). *PLoS ONE*, 7(7):e42201, 2012.