Vol. 27 no. 3 2011, pages 303–310
doi:10.1093/bioinformatics/btq673

Genome analysis Advance Access publication December 12, 2010

Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array

Lars Gidskehaug[1,*], Matthew Kent[1], Ben J. Hayes[2] and Sigbjørn Lien[1]

[1]CIGENE/Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Aas, Norway and [2]Biosciences Research Division, Department of Primary Industries, Victoria, Melbourne 3083, Australia

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Due to a genome duplication event in the recent history of salmonids, modern Atlantic salmon (*Salmo salar*) have a mosaic genome with roughly one-third being tetraploid. This is a complicating factor in genotyping and genetic mapping since polymorphisms within duplicated regions (multisite variants; MSVs) are challenging to call and to assign to the correct paralogue. Standard genotyping software offered by Illumina has not been written to interpret MSVs and will either fail or miscall these polymorphisms. For the purpose of mapping, linkage or association studies in non-diploid species, there is a pressing need for software that includes analysis of MSVs in addition to regular single nucleotide polymorphism (SNP) markers.

Results: A software package is presented for the analysis of partially tetraploid genomes genotyped using Illumina Infinium BeadArrays (Illumina Inc.) that includes pre-processing, clustering, plotting and validation routines. More than 3000 salmon from an aquacultural strain in Norway, distributed among 266 full-sib families, were genotyped on a 15K BeadArray including both SNP- and MSV-markers. A total of 4268 SNPs and 1471 MSVs were identified, with average call accuracies of 0.97 and 0.86, respectively. A total of 150 MSVs polymorphic in both paralogs were dissected and mapped to their respective chromosomes, yielding insights about the salmon genome reversion to diploidy and improving marker genome coverage. Several retained homologies were found and are reported.

Availability and implementation: R-package *beadarrayMSV* freely available on the web at http://cran.r-project.org/

Contact: lg@camo.no

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 7, 2010; revised on November 9, 2010, accepted on November 11, 2010

1 INTRODUCTION

Salmonid fishes have experienced several whole genome duplication (WGD) events in their evolutionary history (Danzmann *et al*., 2008). The 2R hypothesis states there have been two early WGD events (common for all jawed vertebrates), which were followed by subsequent 3R (common for all ray-finned fishes) and 4R duplications (unique for the salmonid lineage). Usually, WGD is followed by extensive modifications of both the genome and transcriptome, with a gradual return to diploidy occurring over time (Sémon and Wolfe, 2007; Wolfe, 2001). The key event during reversion to diploidy is the switch from tetrasomic to disomic inheritance, i.e. from having four chromosomes forming a quadrivalent to having two pairs each forming a bivalent during meiosis. Salmon is still in the process of returning to diploidy after the latest WGD, and the genome displays substantial evidence of retained homologies (Danzmann *et al*., 2008).

Single nucleotide polymorphism (SNP) markers have proven useful in linkage and association studies, and next-generation sequencing technologies have created exciting opportunities for genome sequencing and SNP discovery. The process of SNP discovery is not straightforward, however, and duplicated genomes in particular contain paralogous sequence variants (PSVs), which are readily mistaken for SNPs (Fredman *et al*., 2004). A PSV is created when there is a base pair difference between the sequences of two paralogs, but the substitution does not segregate within either paralogue. Another source of variation in polyploid genomes are multisite variants (MSV) which, in contrast to PSVs, segregate for a base substitution in one or both of the paralogous loci (Fredman *et al*., 2004). MSVs are polymorphic and potentially informative; however, lack of designated software for polyploid genomes makes analysis difficult. In Atlantic salmon and other salmonids, the large numbers of MSVs and PSVs have complicated the use of SNP arrays in genome-wide association (GWAS) and population genetics studies. This is also true for polyploid species of major agronomic importance such as wheat and potato.

Illumina's Infinium® technology (Illumina Inc) is one of the most widely used SNP array technologies, which allows simultaneous genotyping of thousands to millions of SNPs in samples run in parallel on a single silicon slide (Shen *et al*., 2005; Steemers and Gunderson, 2007). The assay chemistry translates the sample allele composition into red and green fluorophores [see Steemers and Gunderson (2007) for more detail].

Illumina's proprietary GenomeStudio Genotyping Analysis Module (v1.6.3 or an earlier relative) converts red and green signals for each SNP into *A* and *B* signals whose values reflect the relative abundance of arbitrarily assigned *A* and *B* alleles. Signal is converted into polar coordinates, using the Manhattan distance metric for the intensity *R*, and with $Theta \in [0,1]$ representing angle $\in [0,90]$ degrees. Each marker is clustered in Cartesian coordinates, and the genotypes $\{AA, AB, BB\}$ are assigned to samples in clusters close to $Theta = \{0, \frac{1}{2}, 1\}$. An example of a typical diploid SNP with well separated clusters is given in Figure 1a.

[*]To whom correspondence should be addressed.

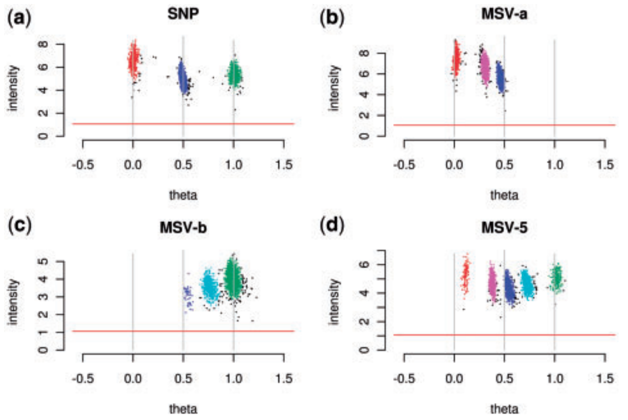

Fig. 1. Examples of how the samples group into patterns that reveal the genotype. (**a**) A diploid region SNP. The blue cluster contains heterozygotes and the red and green clusters contain homozygote *AA*s and *BB*s, respectively. (**b**) An MSV-a containing the genotypes {(*AA*,*AA*), (*AB*,*AA*), (*BB*,*AA*)} in the red, magenta and blue cluster, respectively. (**c**) An MSV-b containing the genotypes {(*AA*,*BB*), (*AB*,*BB*), (*BB*,*BB*)} in the blue, cyan and green cluster, respectively. (**d**) An MSV-5 containing all five combinations in which the two segregating paralogs can vary. The colours of the clusters reflect the B-allele ratios as above.

MSVs differ from diploid SNPs in that the measured signal results from a mixture of four alleles rather than two. These four alleles give rise to a maximum of five different genotype combinations {*AAAA*, *AAAB*, *AABB*, *ABBB*, *BBBB*}, which would be identified by *Theta* values close to $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$, respectively. In instances where markers segregate in one of the paralogs with the other being fixed, up to three cluster positions can be recognized. A fixed *AA* paralogue will result in clusters around $Theta = \{0, \frac{1}{4}, \frac{1}{2}\}$, whereas a fixed *BB* paralogue will generate clusters around $Theta = \{\frac{1}{2}, \frac{3}{4}, 1\}$. Examples of such markers, which we term MSV-a and MSV-b, respectively, are given in Figure 1b and c, although they may be collectively referred to as MSV-3. A less frequent, but nevertheless highly informative MSV, is one which segregates at both paralogous loci simultaneously producing up to five clusters. These are called MSV-5's and display the *Theta* values given above and illustrated in Figure 1d. Though MSVs would be useful in both linkage and association studies, Illumina's GenomeStudio Genotyping Analysis Module does not recognize such variants, and MSVs would typically be discarded or genotypes incorrectly assigned. This is undesirable in Atlantic salmon and indeed other mosaic tetraploids, since it means a significant part of the genome cannot be integrated in GWAS. Genotypes from MSV-5's are especially desirable since they impart information about two paralogous loci resulting from the latest WGD event.

Our aim was to develop algorithms and software to identify and call genotypes from SNPs, PSVs and MSVs in polyploid genomes. Detailed mathematical justifications of the different steps in the analysis are outside the scope of the paper but details of the algorithms are found in the documentation of the published software package (see below). This article outlines the theory and offers empirical evidence that illustrates the functionality of the program. We emphasize that when different pre-processing and analysis schemes are used in this work, this is motivated by the need to highlight strengths and weaknesses of the available methods rather than to find a single best procedure. This enables users to decide on the best options for their own experimental setup and data, as these are expected to vary, for example depending on the extent of polyploidy. The efficacy of the software is demonstrated by the successful mapping of MSVs in Atlantic salmon, which enabled us to determine the extent of polyploidy in this genome.

A new R-package, *beadarrayMSV*, is presented here that (i) provides extensions to existing classes specifically tailored to work with Infinium data, (ii) introduces options for pre-processing of the raw data including transformation, (iii) is able to cluster and automatically call genotypes such as PSVs and MSVs and (iv) contains functionality to split the signal of MSV-5's into two individual paralogs. Quality control (QC) is performed using pedigree information combined with visual inspection of markers scoring poorly on a set of QC parameters. Interactive clustering of problematic markers is also possible. The ability to resolve and map MSVs to chromosomes enables subsequent fine-mapping of markers in areas of the genome which would be otherwise poorly covered. The technique may also be used to verify suggested homologies or propose new ones. To make processing of large datasets possible, data are sequentially read and written from and to files when necessary. The package *beadarrayMSV* is freely available from the CRAN repository (http://cran.r-project.org/) under the GNU General Public License.

2 SYSTEM AND METHODS

2.1 Genotyping

Atlantic salmon genomic DNA was extracted from fin clips provided by the Norwegian breeding company Aqua Gen. A total of 3230 fish (offspring and parents) distributed among 266 full-sib families were genotyped using an Atlantic salmon custom design iSelect SNP array containing 15 225 markers following standard protocols (Kent *et al*., 2009). In total 2648 of the markers were Infinium I, with the remainder being Infinium II. Bead arrays were scanned on an iScan reader, and the standard Infinium II scan settings protocol had been modified to record bead level intensity data in .txt format.

2.2 R-package beadarrayMSV

2.2.1 Data import and classes Two new R classes, 'BeadSetIllumina' and 'AlleleSetIllumina', are defined for the analysis of Infinium data. Both extend the Bioconductor-class 'eSet' and inherit all its methods for subsetting and working with genomic data (Gentleman *et al*., 2004). Functionality was adapted from the beadarray package (Dunning *et al*., 2007) in order to load bead summary data from a repository of scanned arrays and arrange them in an instance of the BeadSetIllumina-class. A number of normalization and transformation routines take these data structures as input.

As Infinium I markers are represented by two bead types, these must be combined to a single marker before genotype calling. The class AlleleSetIllumina is introduced to hold (marker × samples) data tables such as *intensity* and *theta*, containing the polar coordinates signal. All genotype calling procedures use instances of this class. As many genotyping experiments produce too much data for a single R-session, functions for reading and writing marker- and sample-information from and to files are provided.

We refer to the *beadarray* package if there is a need to analyze the scanned image files directly. Alternatively, binary summary files (idat-files) may be loaded and processed using the *crlmm* package (Ritchie *et al*., 2009). A third package *beadarraySNP* contains additional methods for normalization and analysis of genotype data (Oosting *et al*., 2007). These packages all have their unique strengths, and some methods can be used in combination with the *beadarrayMSV* package since they extend the Bioconductor classes. Note,

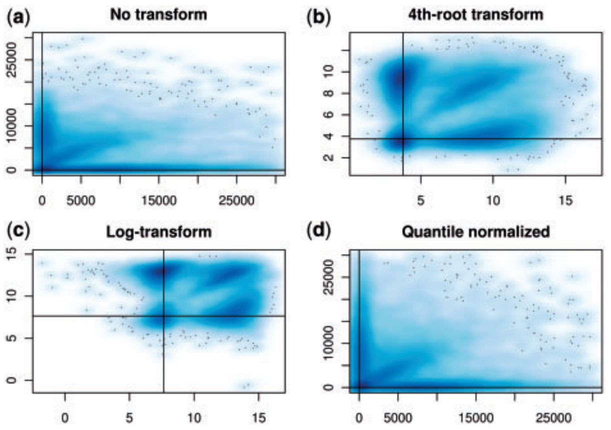

Fig. 2. The effects of different transformations and normalizations. The green versus red signal for all markers on an arbitrary array is shown, after the data have been initially sheared and rotated. Estimated new axes after transformation are indicated. The two clouds along the axes correspond to monomorphs, and the cloud protruding 45 degrees from the origin corresponds to polymorphic markers. (**a**) No transformation, (**b**) fourth root transformation with offset and (**c**) log transformation with offset. The range of the green signal (y-axis) is smaller than the range of the red signal (x-axis). There is also an observed curvature in the polymorphic cloud. (**d**) After quantile normalization of the non-transformed data, the intensities of the channels are comparable and the polymorphic cloud appears straighter.

however, that the normalization and genotyping facilities implemented in *beadarrayMSV* depend on certain required elements such as bead standard errors being available.

2.2.2 Pre-processing The initial shearing and rotation of the red and green signals performed in the GenomeStudio Genotyping Module (Peiffer *et al*., 2006) is mimicked in the package.

Fluorescence data, such as those produced by the Infinium technology, are generally heteroscedastic, i.e. the error variance increases with signal strength. Use of heteroscedastic data in analysis affects the distribution of clusters in the plot of *intensity* versus *theta* for a marker, as the clusters tend to segregate towards *theta* = *0* and *theta* = *1*. In GenomeStudio, diploid SNPs often appear with tightly clustered homozygote samples, whereas the heterozygote samples are more widely spread along the *theta* axis. This is a direct consequence of the signal in both channels being heteroscedastic and should not be mistaken for higher precision of homozygotes compared with heterozygotes. The same effect is also observed in the pooled analysis of Macgregor *et al*. (2008), where they regard and correct for *theta* as binomially distributed frequency estimates with variance *theta*(*1* − *theta*). The heteroscedasticity is of little or no practical consequence for calling regular SNPs, however MSVs, which by definition have more closely positioned clusters, may be more difficult to detect using non-transformed data. Log transformations are often used to reduce heteroscedasticity; however, these also alter the axes around zero in a manner which may be too drastic for our purposes. We have found that a fourth root transformation (or equivalent) provides a good compromise between reducing heteroscedasticity and maintaining a low uncertainty for points close to zero. Both these transformations require positive values, and an offset may be added if needed. The red and green signals after different transformations are plotted for an arbitrary array in Figure 2a–c. After transformation, a new origin in the scatter plots of green (G) versus red (R) for each array needs to be found, which is equivalent to removing the background signal. The background can be deduced using the Infinium I-beads since a single colour is used to detect [A/T] and [C/G] alleles, the colour is arbitrary since it is generated by the base following the marker and not the marker itself. This means that for each Infinium I-bead, one channel will detect nothing but background noise. The distribution of this noise can be parameterized, giving not only an origin but also the expected noise level for each array. The noise levels are subsequently used to find a detection limit below which no genotypes are called.

The average intensity of the red channel for any given array is not equal to the intensity of its green counterpart, and the Illumina software linearly scales both channels such that the centroids of candidate *AA* and *BB* homozygotes get a value of ∼1 (Peiffer *et al*., 2006). A similar scaling is implemented in *beadarrayMSV* using a quantile of all markers. Another linear option is to scale the red channel such that the median red allele frequency [medianAF; $R/(R+G)$] is 0.5 for each array. This is a modified version of the scaling suggested by Macgregor *et al.* (2008). The data from the red and green channels have systematically different distributions, which results in a pronounced curvature in the cloud of polymorphic markers when G versus R is plotted for an array (Fig. 2a–c) (Staaf *et al*., 2008). To account for such non-linear effects, quantile normalization is made available through the *limma*-package (Smyth and Speed, 2003). An example of non-transformed data after quantile normalization is given in Figure 2d.

2.2.3 Polar representation After pre-processing, the BeadSetIllumina object is converted into an instance of the class AlleleSetIllumina by merging the Infinium I-beads such that one marker is represented by a single feature rather than two bead types. After first converting the ‘bead-type’ intensities R and G to ‘marker’ intensities A and B, these are next transformed to polar coordinates *intensity* and *theta*, where *theta* = {0, 1} correspond to angles {0, 90} degrees. A distance measure is chosen depending on the transformation used. The best distance measure is one which ensures that the mean signals of the homozygote and heterozygote clusters are similar. This implies that the first quadrant unit circle should resemble the geometrical shape formed by the extent of the three clusters in each of Figure 2a–c. The Manhattan distance used in GenomeStudio is ideal for non-transformed data as the first quadrant semicircle corresponds to a straight line in Euclidean space. Drawing a straight line between the homozygote clusters after transformation would, however, lead to relatively higher intensities of the heterozygote samples (compare Fig. 2a with 2b and c). This is avoided by using a Minkowski (p-norm) distance instead, the norm p being larger than two, as the unit circle of such a distance resembles a square with rounded corners in Euclidean geometry.

As suggested in Figures 1 and 2, the clusters of homozygotes are centred around *theta* = {0, 1} rather than being confined within these limits. This is to allow clusters to be centred on the expected cluster positions, which is a better criterion after transformation. If B versus A were plotted in Cartesian coordinates, the arc length of the first quadrant of a p-norm circle with radius given by *intensity* would increase with the radius depending on the value of p. When *intensity* versus *theta* on the other hand is plotted in Cartesian coordinates, the corresponding arc length is always one, as it is defined by the distance between *theta* = 0 and *theta* = 1. For the standard errors to be representative also in this plot, they need to be scaled with the intensity-dependent arc length. The standard errors are pooled between the channels and divided with the p-norm arc length found by numerical integration. One can imagine a low-intensity point whose uncertainty spans the origin in a B versus A scatter plot, this point would be given a large uncertainty in the *intensity* versus *theta* plot. Similarly, a point with high intensity would be given a low uncertainty.

2.2.4 Clustering and calling of genotypes Genotype calling for each marker is based on k-means clustering in the two dimensions defined by *intensity* and *theta*. The clustering is simplified by recognizing that only a few more cluster combinations are allowed when we extend the analysis to account also for tetraploid loci. Still, the utility of k-means depends on the existence of clearly defined clusters, and sparsely populated and widespread overlapping clusters may yield inaccurate results. We identify seven genotype classes in which to place all the markers; these are MONO-a, MONO-b, SNP, PSV, MSV-a, MSV-b and MSV-5. The first two denote monomorphic *AA* and *BB*, respectively, the remainder were defined earlier. Only those data points exceeding the estimated noise level are analyzed. Further, samples with the largest standard error may be eliminated from the clustering to provide more accurate clusters. This is useful when many, redundant data points are available to represent each cluster. Removed samples are reintroduced once cluster patterns have been defined.

Based on a histogram of *theta* values and knowledge about the seven allowed cluster combinations, a ranked list of the most likely genotype classes is suggested. This approach improves the probability of detecting very small clusters, and it ensures reproducible clustering as the starting conditions will be the same in every analysis. The suggested cluster combinations are in turn subjected to a series of tests until one is found which fits the following criteria: (i) the maximum deviation of a cluster centre from its theoretical *theta* position cannot exceed *devCentLim*, (ii) the maximum within cluster spread cannot exceed *wSpreadLim*, (iii) the probability of Hardy–Weinberg (HW) equilibrium must exceed *hwAlpha*, (iv) the clusters cannot overlap as defined by their previously calculated centres and spread in the *theta* direction. Markers passing the quality control are subjected to a Hotelling's T^2-test (Gidskehaug *et al*., 2007), which effectively superimposes an ellipse on each cluster and discards all data points falling outside its boundaries or within overlapping ellipses. The extents of the clusters are controlled by the significance level *clAlpha* of the T^2-tests in such a way that small *clAlpha* levels yield large ellipses. The call rate is then required to be larger than a predefined threshold *detectLim*. Finally, if the within cluster spread in the *theta* direction compared to the *intensity* direction is too large, the algorithm moves on to the next most likely cluster combination. If a marker passes all the tests for one of the candidate cluster combinations, the genotype is called, otherwise the marker is failed.

A key difficulty with assessing MSV-5's is that the allele frequencies in the individual paralogs are not directly measured. Rather, it is the mean B allele frequency (BAF) across both paralogs that is inferred from the clustering. The most likely individual BAFs are, however, estimated and reported in the HW test. The quality of these estimates depends on the probability of HW equilibrium.

When parental genotypes are available for the sampled animals, tracking the inherited alleles from parents to offspring is the best method for determining the quality of the clustering and genotype calls. Pedigree checking is included for both SNPs and MSVs, and an overview of the informative meioses and pedigree errors for MSV-5's is provided in Supplementary Material S1. Discrepancies may sometimes be detected close to the cluster borders, and if a large number of pedigree errors are found it may be due to erroneous clustering. Functions for interactive reassignments of clusters, or in extreme cases a manual reclustering, are implemented. These make use of the package *rggobi*, an R-interface to the dynamic graphics package GGobi (www.ggobi.org/rggobi/). The ability to interactively visualize and modify clustering results is an important quality control step also in Illumina's software, and automatic clustering results should always be validated by the researcher. This subjective, final assessment is the best guarantee that the estimated genotypes are as accurate as possible.

Several genotype calling and testing metrics are returned that collectively indicate the clustering quality for each marker. These include the maximum deviation from expected cluster positions, the maximum within cluster spread, the probability of HW equilibrium and the call rate. If pedigree validation has been performed, the number of pedigree errors for each marker may also be found. No overall probability is returned, as this would have to depend on many factors and the estimate would likely be uncertain. Rather, the markers may be ranked using the different quality estimates in order to plot or interactively recluster any questionable markers.

With a little added functionality, the genotype calling algorithm could in theory be utilized for analysing higher ploidy species as well. It would, however, require very accurate data to reliably detect clusters that are closer together than the tetraploid MSVs in the current dataset.

2.2.5 MSV-5 mapping Initially, all calls are identified by the B-allele ratios of the markers. For instance, $call = \frac{1}{2}$ means genotype *AB* or *AABB* depending on whether the marker is a SNP or an MSV/PSV, respectively. For all genotype classes except MSV-5, these calls can be directly translated into genotypes once the relevant marker- and strand-information is known.

The mapping of MSV-5's in *beadarrayMSV* is a three-step process which involves: (i) splitting of markers into two paralogs for all informative meioses within half-sib families, (ii) naming the individual paralogs with unique names reflecting their chromosome numbers and (iii) merging the linkage information across families for both parents and supplementing with additional, remaining meioses. The mapping of MSV-5 paralogs starts by creating individual markers with names referring to paralogue 1 and 2. These names are arbitrarily chosen and therefore unique only within the half-sib families. Those calls that can be split according to the tables in Supplementary Material S1 are then filled in for each half-sib family. This results in two sparsely populated data tables, one for the fathers and one for the mothers. Step two of the analysis involves associating the individual paralogs with several markers of known positions in the genetic map (S. Lien *et al*., manuscript in preparation, a dense SNP-based linkage map for Atlantic salmon (Salmo salar): recombination patterns and chromosome homeologies enlightening the extent of diploidization in the salmon genome). If a matching offspring is registered each time an informative allele in a paralogue corresponds with an informative allele in a mapped marker, the degree of association between the two is determined by counting the number of matches. Associations supported by too few informative meioses are filtered away. The total number of matches across markers within each chromosome is then divided by the number of tested markers, such that the chromosomes with the highest average number of matches can be found. For each MSV-5, up to two chromosomes are identified if the number of matches per marker exceeds some threshold. The individual paralogs are given names reflecting the chromosome they map to, and the merged linkage information from both parents can be used in a single analysis across families. Tentative chromosome positions are also provided by *beadarrayMSV*; however, its main utility in this setting is to provide the split MSV-5's for linkage mapping using other software.

2.3 Data analysis

Two analyses, 'run 1' and 'run 2', were performed with different pre-processing of the data. For both runs, the raw data for each channel were initially sheared and rotated. In run 1, an offset of 200 was then added to the data before fourth root transformation. The channels were subsequently medianAF-normalized, and the genotype of each marker was called using the default settings. In run 2, no additional transformation of the data was performed, but the channels were quantile normalized. Default genotype-settings were used except for the parameters *devCentLim* and *wSpreadLim*, controlling the maximum allowed cluster centre deviation and within-cluster spread. These were in run 2 set to 0.4 and 0.15, respectively, to account for the uneven separation and spread of clusters resulting from heteroscedastic data. All settings were considered optimal for each transformation.

3 RESULTS AND DISCUSSION

3.1 Genotype calling

A summary of the results is given in Table 1. From the total number of around 15K markers on the array, more than 50% were classified as MONO, PSV or FAIL. This reflects challenges related to SNP discovery in polyploid species (Moen *et al*., 2008; Sánchez *et al*., 2009), but it is important to note that many of the monomorphic markers in this population are polymorphic in other populations (Kent *et al*., 2009).

Some quality scores are also presented in Table 1 to give an overall impression of the accuracy of the calls. These scores alone should not be used to infer which pre-processing is universally best as the results depend heavily on the data quality, input parameters and

Table 1. Summary and quality of genotype calling

	MONO-a	MONO-b	PSV	SNP	MSV-a	MSV-b	MSV-5	FAIL
Run 1: Fourth root transformation (offset 200), medianAF-normalization								
Sum	2923	2877	1900	4268	544	795	132	1786
Acc	1.00	1.00	0.97	0.97	0.89	0.84	0.91[a]	–
Ped	–	–	–	0.98	0.93	0.88	0.94	–
nP	–	–	–	1.97	4.61	6.25	7.35	–
Run 2: No transformation, quantile-normalization								
Sum	3313	3395	1359	4522	426	580	139	1491
Acc	0.97	0.95	0.94	0.95	0.95	0.92	0.88[a]	–
Ped	–	–	–	0.96	0.91	0.88	0.92	–
nP	–	–	–	3.3	4.71	6.35	5.12	–

The variable ‘Sum’ gives the total number of called markers, ‘Acc’ is the estimated fraction of correct calls based on visual inspection of up to 220 markers in each class, ‘Ped’ is an alternative accuracy estimated as the fraction of calls with 15 offspring pedigree errors or less and ‘nP’ is the average number of pedigree errors per marker (from the full set of 2915 offspring). The genotype categories ‘MONO-a’, ‘MONO-b’, ‘PSV’, ‘SNP’, ‘MSV-a’, ‘MSV-b’ and ‘MSV-5’ are defined above, and non-assigned markers are denoted ‘FAIL’.

[a] Based on the full set of MSV-5 markers.

the subjective assessment of which calls are correct. The accuracy (‘Acc’) of the calls, defined as the fraction of correct calls among all markers assigned to a category, was assessed through visual inspection. The first 220 markers within each category were selected for inspection, except for MSV-5 where all were included. The selection of the first 220 markers is expected to be unbiased as there is no theoretical or discernible trend between marker performance and their alphabetical order. The accuracy was high in both runs, though more PSVs and MSVs (and less monomorphics and SNPs) were found in run 1 than in run 2. With a notable exception for PSVs, there was a strong tendency that more calls within a category implied both more correct assignments and more false positives. A more objective quality criterion than visual inspection (though not necessarily more accurate) is to declare all markers with more than 15 offspring pedigree errors as false calls. This gave rise to the alternative accuracy ‘Ped’ in Table 1, which is based on the full set of markers. The two estimates of accuracy are similar in both runs, except ‘Acc’ is higher for the MSV-a and -b in run 2. The average number of pedigree errors per marker (‘nP’) was less for SNPs than for MSVs, and the median number of pedigree errors was zero in all cases. Error rates were well within what is expected given the genotyping error rate.

The above accuracies are not exact, as the true genotypes were not known. However, when a large sample is used, it is usually clear from visual inspection whether the clustering successfully distributes the data into genotype groups. In addition, information about sample pedigree is very effective at highlighting instances of poor clustering. In some instances, it was difficult to distinguish between an MSV-3 and a SNP; however, such errors do not influence the subsequent mapping. Based on markers that were unambiguously called we estimated that the percentage of incorrect calls was at most 4–5% for SNPs and 8–10% for MSVs. Significant improvements to these accuracies are unlikely since some assays fail to work correctly, possibly due to secondary mutations in SNP flanking regions.

Some selected markers are plotted in Figure 3, with run 1 in the first row and run 2 in the second row. The SNP to the left was correctly called in both instances, but the blue cluster of heterozygotes was far to the right from its ideal position. Combined with elongated clusters in the *intensity* direction, this led to wrong assignments of some heterozygotes to the green cluster in run 2. The yellow cross indicates one such offspring that failed the pedigree check, and its parents are identified by yellow circles.

A subset of heterozygous samples deviated from the main cluster for the second marker in Figure 3. Possible reasons for this deviation could be differences in DNA quality, sample preparation or lab processing. Both runs correctly called this marker a SNP; however, incorrect assignment of the deviating cluster in run 2 gave rise to pedigree errors. The MSV-b (third example from the left) was not identified in the second run at all, as three clusters could not be distinguished by the algorithm. The last example (to the far right) is an MSV-5 which was correctly called in both runs, but where run 2 performed much better in terms of both interpretation and assignments of samples close to the cluster borders. Due to overlapping clusters in run 1, many of the samples were called incorrectly, while the quantile normalization in run 2 contributed to the clusters being upright and well defined. Overall, run 1 performed better for these data in terms of more identified MSVs and a higher occurrence of compact clusters. For the rest of the article, we will therefore use and refer to the results of run 1.

3.2 Mapping of MSV-5’s

The set of 132 detected MSV-5’s (Table 1) was supplemented with 61 initially failed markers visually found to resemble MSV-5’s. The full set was subjected to quality control using the interactive plotting tools in *beadarrayMSV*, and manual clustering or re-assignments of clusters were performed where needed. A small number of the ambiguous markers were included in the set as we wanted to identify as many homeologies as possible. Pedigree validation and visual inspection resulted in a list of 150 MSV-5’s which were split into paralogs, in the majority of cases these were successfully mapped to chromosomes. An example is given in Figure 4, which reveals that the paralogs of marker ‘ESTNV_29659_1478’ are likely found on chromosomes 7 and 17. Phillips *et al.* (2009) confirm that the q-arm of chromosome 7 is homeologous to a segment of chromosome 17 as result of the recent genome duplication event in salmonids. The full set of suggested homeologies or single marker duplications is given in Table 2. Note that this method is not able to distinguish true homeologies from smaller scale duplications; however, it is an indication of previous genome duplication if a large number of MSV-5’s are found that map to the same pairs of chromosomes. In that sense, these data confirmed the ‘2/5’-, ‘7/17’-, ‘4/8’- and ‘3/6’-homeologies. The ‘11/26’-homeology was supported by only two markers in this study; however, this is still more than the previously reported numbers.

Five suggested homologies were supported by a single MSV-5 marker only. Of these, two corresponded to verified homeologies; however, three remained uncertain. The ‘1/6’-pair had a small peak for chromosome 1 that might be a false positive (data not shown). The ‘13/15’-pair had a peak for chromosome 13 supported by the fathers only. A single duplication on these chromosomes is, however, reported by Danzmann *et al.* (2005,2008). The ‘19/29’-pair had two significant peaks based on the linkage information from the mothers only. In addition to the pairs presented in Table 2,

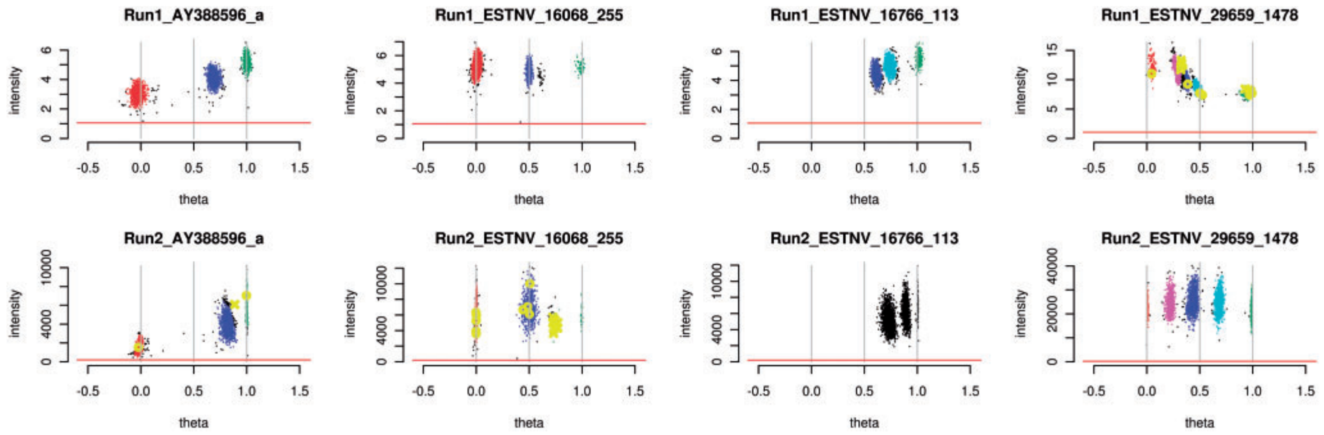

Fig. 3. Comparison of genotype calling results based on four example markers after two different pre-processing settings. For run 1 (upper row), the default fourth root transformation and medianAF channel normalization were performed. For run 2 (bottom row), no transformation was performed, but the channels were quantile normalized. In general, the heteroscedasticity of the data in run 2 leads to smaller spread in *theta* for the homozygote compared with the heterozygote clusters. Still, the classifications of the markers are identical between the runs, except for 'ESTNV_16766_113', which is failed in run 2. Also, there are apparently non-systematic differences in the number of pedigree errors between the runs. The lower variance of points close to $theta = \{0, 1\}$ in run 2 seems to complicate the clustering of 'ESTNV_16766_113', which is an MSV-3.

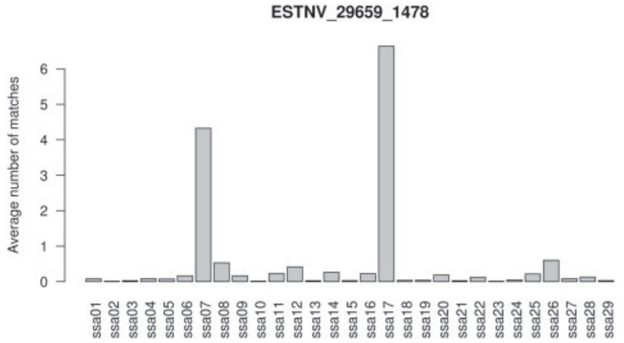

Fig. 4. The total number of matches between MSV-5 paralogs and the linkage map markers for each chromosome, divided by the number of markers representing that chromosome. The example shows an MSV-5 marker whose paralogs map to the homeologous regions of chromosomes 7 and 17.

42 markers were mapped to 19 single chromosomes. These may represent repeated elements on the same chromosomes, but many are likely due to an insufficient number of informative meioses for the second chromosome. This may happen when the minor allele frequency for one of the paralogs is low. For eight of the 150 markers, no significant peaks could be found. This is likely due to limited meiotic information or false positive MSV-5 calls.

The 242 resolved paralogs were included in the linkage map using CRI-MAP (Green *et al.*, 1990) and are reported in Supplementary Material S2. Representations of chromosomes 2 and 5, including the mapped MSV-5 paralogs, are plotted in Figure 5. The approximate shapes and relative sizes of the chromosomes were adapted from Phillips *et al.* (2009). The relevant parts of the (female) genetic map were superimposed on the illustrations in Figure 5. The illustration shows that all 39 '2/5' pairs from Table 2 are located on the p-arm of chromosome 2 and the q-arm of chromosome 5. In addition, the q-arm of chromosome 2 holds one of the paralogs of the '2/12' pair.

Table 2. Suggested duplications based on 150 MSV-5's

Suggested chrom. pairs	Verified homeologies[a]	Number of duplicated markers			
		MSV-5	Phillips[b]	Danzmann-08[c]	Danzmann-05[d]
2/5	2p/5q	39	9	9	7
7/17	7q/17qa	33	5	5	2
4/8	4p/8q	14	4	4	–
3/6	3q/6p	7	9	9	1
11/26	11qb/26	2	–	1	–
1/6	–	1	–	–	–
2/12	2q/12qb	1	5	5	2
13/15	–	1	–	1	1
16/17	16qa/17qb	1	5	5	3
19/29	–	1	–	–	–

[a] Highly supported homeologous chromosome arms (Phillips *et al.*, 2009). The number of supporting markers is compared to those of [b] Phillips *et al.* (2009), [c] Danzmann *et al.* (2008) and [d] Danzmann *et al.* (2005).

These results verify two of the homeologies reported by Phillips *et al.* (2009). Paths have been drawn between corresponding paralogs for a few selected markers. Though a degree of imprecision due to small genetic distances in the relevant regions is expected, there are indications that the homeologous segment of chromosome 5 is flipped compared with chromosome 2.

4 CONCLUSIONS

SNP array genotyping software (such as that supplied by Illumina) typically call genotypes at diploid SNP markers very accurately; however, it is unreliable when markers are in duplicated genome regions. While MSV-3's can be called correctly, this requires manual inspection and is a daunting task when thousands of markers are

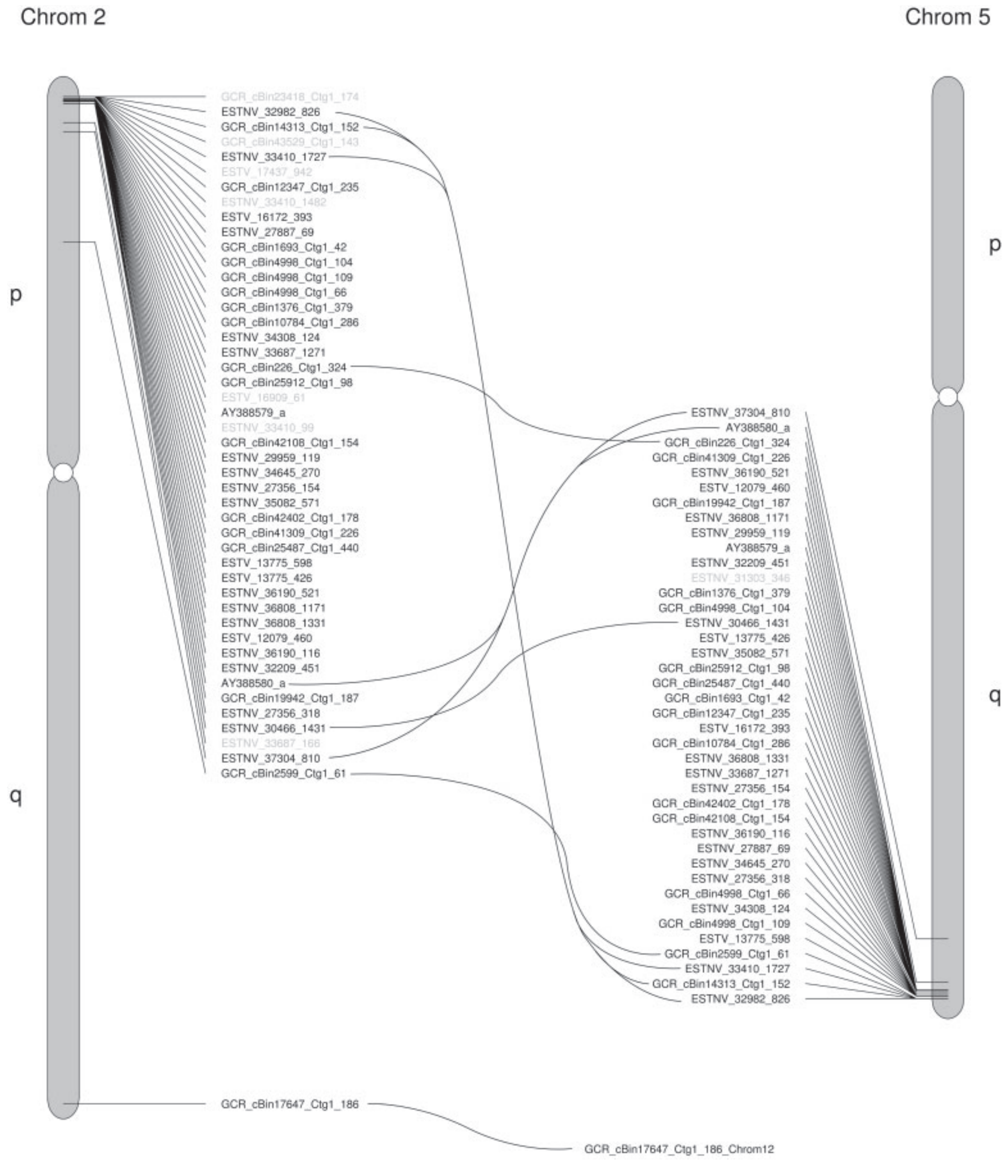

Fig. 5. Illustration of the homeologous chromosome arms 2p and 5q, including the mapped MSV-5 paralogs. The female genetic map has been used as a proxy for the physical map. A single MSV-5 from the 2q/12qb homeology was also found. All paralogs in black text have been positioned on both chromosomes, whereas grey text indicates that the position of the alternate paralogue is unknown. Lines connect a few of the known paralogue pairs. The shapes and relative sizes of the chromosomes are approximately taken from Phillips *et al*. (2009).

genotyped, in contrast to MSV-5's that cannot be called at all. Consequently, an automated routine for analyzing such data is highly desirable.

The main objective for developing *beadarrayMSV* was to enable analysis of Illumina BeadArrays in the partly tetraploid Atlantic salmon. This has resulted in a flexible R-package with demonstrated merit for duplicated genomes. The methods should also be useful for genomes typically characterized as diploid to identify duplicated regions. The data structures provided are useful for any data genotyped on the Infinium platform, and the tools developed for normalization, transformation and plotting have general applicability.

The work towards a complete reference genome for Atlantic salmon is still in progress; however, the mosaic tetraploid nature of the salmon genome complicates assembly. Tools which enable calling and mapping of markers of duplicated regions will represent a valuable contribution towards this activity. From a more practical point of view, prior to designing a new SNP-chip it is very hard to tell whether a putative marker is a SNP or an MSV, especially when no reference genome is available. Having tools that enable us to utilize MSV markers on chips means that available resources for SNP discovery and genotyping can be expanded to a greater proportion of the genome. By identifying MSV-markers on the salmon Illumina SNP array, we were able to increase the number of useable polymorphic markers by 35% compared to using the SNP markers only. Subsequent GWAS and genomic selection (Meuwissen *et al.*, 2001) studies in Atlantic salmon are likely to benefit from these improvements. Lastly, the MSVs and MSV-5's in particular hold important clues to how the genome of salmon has evolved since the last duplication event. Further studies using this information may shed additional light on the processes involved in diploidization and speciation in general.

ACKNOWLEDGEMENTS

Aqua Gen AS (Trondheim, Norway) is gratefully acknowledged for providing the high-quality sample material including extensive pedigree information used in this work. Prof. Mike Goddard (Department of Primary Industries, Melbourne, Australia) is thanked for his helpful contributions regarding mathematical challenges encountered during method development.

Funding: The Norwegian Research Council (NFR) (grant numbers 177036/S10, 183607/S10).

Conflict of Interest: none declared.

REFERENCES

Danzmann,R.G. *et al.* (2005) A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). *Genome*, **48**, 1037–1051.

Danzmann,R.G. *et al.* (2008) Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon). *BMC Genomics*, **9**, 557.

Dunning,M.J. *et al.* (2007) beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**, 2183–2184.

Fredman,D. *et al.* (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.*, **36**, 861–866.

Gentleman,R. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Gidskehaug,L. *et al.* (2007) A framework for significance analysis of gene expression data using dimension reduction methods. *BMC Bioinformatics*, **8**, 346.

Green,P. *et al.* (1990) *Documentation for CRI-MAP, Version 2.4*. Washington University School of Medicine, St Louis, MO, USA.

Kent,M.P. *et al.* (2009) Development of 16.5 K SNP chip for Atlantic Salmon, *Proceedings of the 17th Plant and Animal Genome Conference*, January 10–14, 2009 San Diego, CA. http://www.intl-pag.org/

Macgregor,S. *et al.* (2008) Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. *Nucleic Acids Res.*, **36**, e35.

Meuwissen,T.H.E. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Moen,T. *et al.* (2008) A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. *BMC Genomics*, **9**, 223.

Oosting,J. *et al.* (2007) High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays. *Genome Res.*, **17**, 368–376.

Peiffer,D. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.

Phillips,R.B. *et al.* (2009) Assignment of Atlantic salmon (Salmo salar) linkage groups to specific chromosomes: conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (Oncorhynchus mykiss). *BMC Genetics*, **10**, 46.

Ritchie,M.E. *et al.* (2009) R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. *Bioinformatics*, **25**, 2621–2623.

Sánchez,C.C. *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Sémon,M. and Wolfe,K.H. (2007) Consequences of genome duplication. *Curr. Opin. Genet. Dev.*, **17**, 505–512.

Shen,R. *et al.* (2005) High-throughput SNP genotyping on universal bead arrays. *Mutat. Res.*, **573**, 70–82.

Smyth,G. and Speed,T. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.

Steemers,F.J. and Gunderson,K.L. (2007) Whole genome genotyping technologies on the BeadArray^{TM} platform. *Biotechnol. J.*, **2**, 41–49.

Staaf,J. *et al.* (2008) Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics*, **9**, 409.

Wolfe,K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.*, **2**, 333–341.