BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 8 2009, pages 1056–1062
doi:10.1093/bioinformatics/btp104

Genetics and population analysis

Array-based genotyping in *S.cerevisiae* using semi-supervised clustering

Richard Bourgon[1,*], Eugenio Mancera[2], Alessandro Brozzi[1], Lars M. Steinmetz[2] and Wolfgang Huber[1]

[1]EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Received on December 28, 2008; revised on February 1, 2009; accepted on February 17, 2009

Advance Access publication February 23, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Microarrays provide an accurate and cost-effective method for genotyping large numbers of individuals at high resolution. The resulting data permit the identification of loci at which genetic variation is associated with quantitative traits, or fine mapping of meiotic recombination, which is a key determinant of genetic diversity among individuals. Several issues inherent to short oligonucleotide arrays—cross-hybridization, or variability in probe response to target—have the potential to produce genotyping errors. There is a need for improved statistical methods for array-based genotyping.

Results: We developed ssGenotyping (ssG), a multivariate, semi-supervised approach for using microarrays to genotype haploid individuals at thousands of polymorphic sites. Using a meiotic recombination dataset, we show that ssG is more accurate than existing supervised classification methods, and that it produces denser marker coverage. The ssG algorithm is able to fit probe-specific affinity differences and to detect and filter spurious signal, permitting high-confidence genotyping at nucleotide resolution. We also demonstrate that oligonucleotide probe response depends significantly on genomic background, even when the probe's specific target sequence is unchanged. As a result, supervised classifiers trained on reference strains may not generalize well to diverged strains; ssG's semi-supervised approach, on the other hand, adapts automatically.

Availability: The ssGenotyping software is implemented in R. It is currently available for download (www.ebi.ac.uk/~bourgon/yeast_genotyping/ssG) and is being submitted to Bioconductor.

Contact: bourgon@ebi.ac.uk

Supplementary information: Supplementary data and a version including color figures are available at *Bioinformatics* online.

1 INTRODUCTION

During meiosis, homologous copies of the chromosomes align, and the repair of programmed double-stranded breaks in the DNA leads to recombination: the reciprocal exchange of DNA between homologs (crossovers), or the non-reciprocal modification of one homolog, using the other as a template (non-crossover gene conversion). As a consequence, the genome of each meiotic product, or 'segregant', is a mosaic of the two parental genotypes (Fig. 1). A recent study in *Saccharomyces cerevisiae* used the array-based genotyping methodology presented here to create a genome-wide map of crossover and non-crossover gene conversion with the highest resolution to date (Mancera *et al.*, 2008).

Oligonucleotide microarrays provide an accurate and cost-effective means of identifying and genotyping polymorphic loci. Oligonucleotide microarray probes hybridize more efficiently to targets whose sequence is exactly complementary than to targets which only partially or imperfectly match the probes. Winzeler *et al.* (1998) used this fact to identify several thousand polymorphic positions in the same two yeast strains we consider here. Since then, numerous authors have made use of these so-called 'single feature polymorphisms' (SFPs)—in yeast (Brem *et al.*, 2002; Deutschbauer and Davis, 2005; Gresham *et al.*, 2006; Steinmetz *et al.*, 2002; Winzeler *et al.*, 2003), and also in other organisms (Albert *et al.*, 2005; Borevitz *et al.*, 2003; Rostoks *et al.*, 2005; Turner *et al.*, 2005). With the exception of Brem *et al.* (2002), these authors have taken a supervised approach to the problem, training a genotyping classifier on samples of known genotype and then applying the classifier to new samples. Winzeler *et al.* (1998) hybridized parental genomic DNA from each of the two strains to standard yeast expression arrays. Then, after preprocessing, analysis of variance (ANOVA) was used to identify probes whose observed log-scale fluorescence intensities appeared to be better fit by a model with two means than by a model with one. Such probes were deemed to be SFPs. To genotype segregants from a cross, a posterior probability was computed using the estimated Gaussian densities from the parental-array ANOVA, plus a uniform prior on the two genotypes:

$$\hat{P}(\text{Genotype } g \,|\, \text{intensity} = x) = \frac{\hat{p}_g(x)}{\hat{p}_1(x) + \hat{p}_2(x)}. \qquad (1)$$

Variants on this procedure soon emerged. The 1- versus 2-mean ANOVA is equivalent to a two-sample t-test for difference in means, and Borevitz *et al.* (2003) proposed an alternative t-test for identification of SFPs, using the *ad hoc* moderated t-statistic of SAM (Tusher *et al.*, 2001). Brem *et al.* (2002)—whose data included hybridizations from numerous segregants of unknown genotype, as well from parental samples of known genotype—further augmented this approach: using parental data, candidate SFPs were identified on the basis of a high moderated t-statistic. Then, known parental

*To whom correspondence should be addressed.

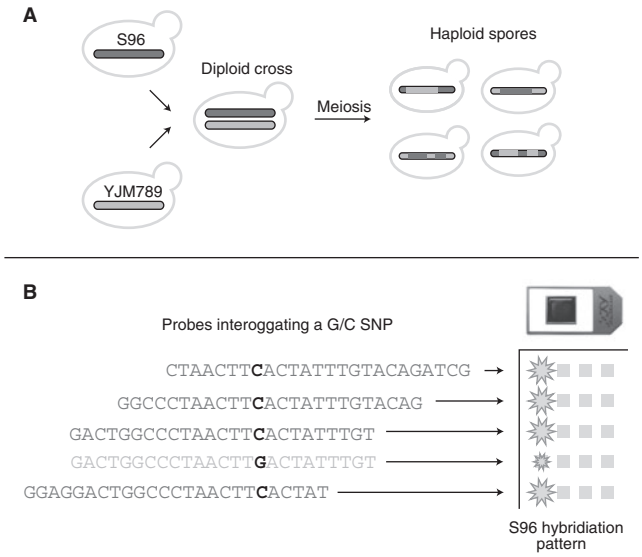

Fig. 1. Meiotic recombination genotyping assay. (**A**) Meiosis was induced in a diploid cross derived from the highly polymorphic S96 and YJM789 strains. Haploid parents and meiotic products ('segregants') were used for genotyping. (**B**) Five probes—four S96-specific and one YJM789-specific—interrogate a C/G SNP. When presented with S96 genomic DNA, the S96-specific probes are bright while the YJM789-specific probe exhibits reduced fluorescence intensity. For YJM789 genomic DNA, the pattern would be reversed.

genotype labels were temporarily set aside, and the combined parental and segregant data were subjected to k-means clustering ($k=2$). Candidate SFPs were only retained if the parental samples were correctly separated by the resulting clusters. Further, Brem *et al.* estimated the Gaussian densities required in (1) from *all* data in the clusters, rather than only from parental observations of known genotype.

The more recent, multivariate approach of Gresham *et al.* (2006)—designed for high-density tiling microarrays—is quite different: the authors considered the set of probes which interrogate a given position, and they modeled the decrease in fluorescence intensity caused by a SNP as a function of (i) the SNP's position within each probe, (ii) known response of the probes to reference sequence and (iii) various aspects of the probes' base composition. Their algorithm, SNPscanner, was trained on a set of 'high-quality' known SNPs to produce two predictions for probe set behavior: one which corresponds to reference sequence, and the other, to sequence with a variant base at the given position. Observed behavior on new arrays was compared with the two predictions, and genotype was assigned on the basis of which model fits best.

In the remainder of this article, we introduce ssGenotyping (ssG) as an alternative to SNPscanner, and show that it provides both more specific and more sensitive genotyping in the context of a meiotic recombination dataset. In addition, we use the comparison between the methods to illustrate two points which are important for successful array-based genotyping in any context: (i) the extent to which probe behavior—cross-hybridization behavior, in particular—is sensitive to genomic background, and (ii) the ability of predictive models to describe probe behavior in a complex setting.

2 APPROACH AND METHODS

2.1 Motivation

We developed the ssG algorithm to genotype over 50 000 polymorphic markers in 220 segregants—51 wild-type tetrads and 5 *msh4* deletion mutant tetrads—resulting from the sporulation of a diploid cross of two substantially diverged strains of *S.cerevisiae* (see Supplementary Methods). One strain, S96, is isogenic with the common laboratory strain S288c, for which the whole-genome sequence is known; the other, YJM789, is a clinical isolate that has recently been sequenced (Gu *et al.*, 2005; Wei *et al.*, 2007). The segregation patterns of the markers provided detailed information about local recombination rates, patterns of crossover interference, and the size and spatial distribution of gene conversion events (Mancera *et al.*, 2008).

Genomic DNA from the segregants as well as from 25 parental samples was hybridized to Affymetrix tiling microarrays which provide dense coverage of the reference S288c genome, typically at 4 bp resolution. The arrays also include probes which interrogate YJM789 sequence, at positions where this sequence differs from the S288c reference (see Supplementary Methods). Comparison of the aligned sequences from the two strains revealed ≈61 000 putative polymorphisms—single nucleotide polymorphisms (SNPs), insertions or deletions—of which ≈52 000 were interrogated by distinct sets of one or more uniquely mapping probes. Given the tiling design, the vast majority of these polymorphisms were interrogated by sets of overlapping probes (Supplementary Figure S1). We therefore selected a multivariate approach which is able to accommodate correlation arising from the overlap among the probes in a probe set (Fig. 2).

2.2 ssG: raw genotype calls

Figure 2 shows that (i) segregant data, as one might expect, typically produce two distinct clusters, and (ii) parental data are informative, but not always representative of the segregants with which they share a genotype. With both points in mind, the ssG algorithm fits a single model to the combined parental and segregant data, using known genotype labels for the parents and soft class assignments for the segregants. More concretely, the preprocessed log-scale intensity data from the set of d probes which interrogate a given polymorphism—i.e. which overlap the polymorphism in at least one of their 25 bases—may be represented with vectors $X_1,\ldots,X_n$, each of dimension d. To assign genotypes to the segregants, we fit a Gaussian mixture model to all arrays simultaneously, computing maximum likelihood estimates via the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). The algorithm proceeds as follows:

Let $Y_i \in \{1,2\}$ denote the genotype of sample i at the polymorphism in question. If A, B and S denote the indices corresponding to the two parental types and the segregants, respectively, then Y_i is known whenever $i \in A \cup B$, but unknown for $i \in S$. We postulated that $X|_Y \sim \mathcal{N}_d(\mu_Y, \Sigma_Y)$, and that (X_i, Y_i) was independent of $(X_{i'}, Y_{i'})$ for $i \neq i'$. Importantly, we did not require μ_{1j} to differ from μ_{2j} for every j—reflecting the fact that the marginal behavior of some probes in a probe set may not distinguish between the two genotypes. Further, Σ_1 and Σ_2 were not assumed to be equal nor diagonal. Figure 2 provides a typical example of the marginal and joint behavior of probes in a eight-probe set, and shows that both of these issues are relevant.

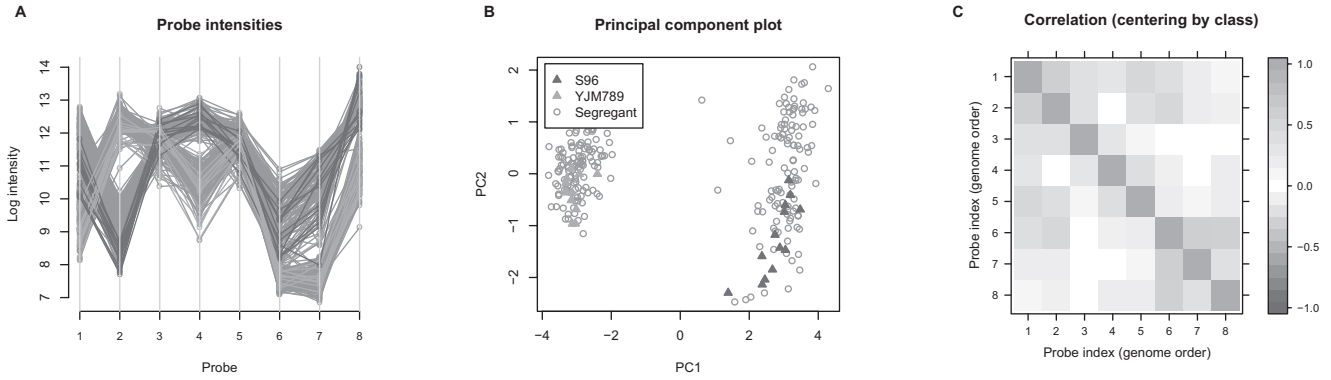

Fig. 2. A probe set associated with a SNP (ID 180) on chromosome I. (**A**) Behavior of the eight probes in the set, ordered by genomic position. S288c-specific probes (e.g. 1) produce higher intensity on S96 parental arrays (red), while YJM789-specific probes (e.g. 2) are brighter on YJM789 parental arrays (blue). Note that some probes (3 and 5) fail to distinguish between the parental data, and that within-class variance is not constant. (**B**) Projection of eight-dimensional data into the first two principal components (PCs). Each point in B corresponds to one connected set of segments in A. Data form clear clusters, although parental behavior is not fully representative of segregant behavior. PC1 aligns with a vector whose entries are ± 1 for informative probes—with sign determined by the allele to which the probe is specific—and 0 for the two uninformative probes ($r = 0.97$). Projection in this direction corresponds to comparing average log intensities from informative probes and ignoring the two non-responders. (**C**) Residual pair wise correlation, after centering within class. There are blocks along the diagonal—corresponding to consecutive, overlapping probes—with significant positive and negative correlation.

In the Gaussian mixture case, the M step of the EM algorithm—which maximizes an estimate of the conditional expectation of the log likelihood—only requires estimates of $P(Y_i = g|X_i)$ for all $i \in S$. To initialize these conditional probabilities (hereafter denoted p_{ig}), we applied a simple clustering algorithm—k-means, with the two clusters seeded with parental observations—to the combined parental and segregant data, and then set each $p_{ig}^{(0)}$ to either 0 or 1, depending on the outcome of this clustering. (Alternately, one could begin with the E step and initialize the $\hat{\mu}_g$ and $\hat{\Sigma}_g$ using the parental data; this strategy produced identical results.) Defining $p_{\cdot g} \equiv \sum_i p_{ig}$, it is straightforward to show that the M step's objective function is maximized in μ_g by

$$\hat{\mu}_g^{(1)} = \frac{1}{p_{\cdot g}^{(0)}} \sum_{i=1}^{n} p_{ig}^{(0)} X_i, \tag{2}$$

i.e. by a weighted average of the observations, with weights determined by (estimated) conditional probability of membership in class g. The objective function is maximized in Σ_g by

$$\hat{\Sigma}_g^{(1)} = \frac{1}{p_{\cdot g}^{(0)}} \sum_{i=1}^{n} p_{ig}^{(0)} (X_i - \hat{\mu}_g^{(1)})(X_i - \hat{\mu}_g^{(1)})', \tag{3}$$

i.e. by a weighted version of the standard empirical covariance estimate. In the meiotic recombination context, it is natural to assume that, for a given polymorphism, a segregant is equally likely to inherit either of the two parental alleles, so we fixed $\pi_1 = \pi_2 = 0.5$. In other contexts, π_1 and π_2 can easily be estimated. To begin the next iteration, we updated p_{ig} for all $i \in S$, by

$$p_{ig}^{(t)} = \frac{\phi(X_i; \hat{\mu}_g^{(t)}, \hat{\Sigma}_g^{(t)})}{\phi(X_i; \hat{\mu}_1^{(t)}, \hat{\Sigma}_1^{(t)}) + \phi(X_i; \hat{\mu}_2^{(t)}, \hat{\Sigma}_2^{(t)})}, \tag{4}$$

and continued until a convergence criterion was met. Here, $\phi(\cdot; \mu, \Sigma)$ denotes the density of a multivariate normal distribution with mean μ and covariance matrix Σ. We also define $\hat{\phi}_g$ to be $\phi(\cdot; \hat{\mu}_g, \hat{\Sigma}_g)$. Final assignment of genotype for the segregants was then obtained by comparing p_{i1} and p_{i2}. This is analogous to (1), although the two distributions are now multivariate, and the parameter estimates are derived from a combination of the parental and offspring data rather than from parental data alone. The contrast between the two fit types (semi-supervised versus supervised parental-only) can be substantial, as illustrated in Figure 3.

2.3 ssG: filtering

After fitting distributions to all probe sets, we applied quality filtering at the (i) array, (ii) polymorphism and (iii) individual call levels. Genotype calls for four arrays implied a huge increase (more than an order of magnitude above what was typically observed) in genotype switching along the associated segregants' chromosomes, so these four arrays were set aside. The distributional estimates $\hat{\phi}_1$ and $\hat{\phi}_2$ returned by the EM algorithm admit natural polymorphism- and call-level filtering as well. Figure 3B shows that the distributions were not always well separated; further, due to errors in genomic sequence or alignment, some putatively polymorphic loci may not actually be polymorphic. We used $\hat{\phi}_1$ and $\hat{\phi}_2$ to compute expected misclassification rates, and set aside probe sets for which this rate was too large (>1%). Finally, individual calls which were intermediate with respect to $\hat{\phi}_1$ and $\hat{\phi}_2$—producing p_{ig} which were too far from both 0 and 1—were removed. Individual calls with unambiguous p_{ig} but which were nonetheless outliers with respect to their assigned class were also removed. (See Supplementary Methods for additional details.)

Supplementary Figure S2 depicts a further problem found in a small fraction (0.7%) of probe sets: behavior which is inconsistent with the biological and statistical models for meiotic recombination. Such behavior may be due to cross-hybridization from sequence at an unlinked locus, or to unanticipated translocations in our S96 parental strain, relative to the S288c reference sequence. To address this issue, we computed auxiliary fits for each probe set—by using only parental or only offspring data, or by fitting more than two

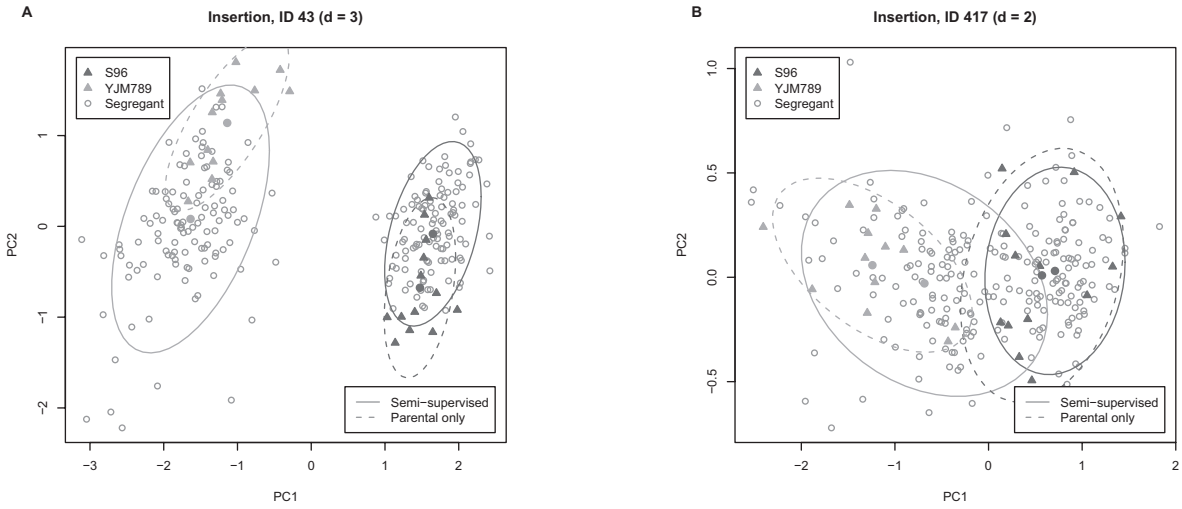

Fig. 3. Parental-only (dashed) versus semi-supervised (solid) fitted distributions for two probe sets, with projection into the first two PCs for visualization. Contours represent ± 2 SDs in each PC. Clusters are well separated in (**A**) but not (**B**). For both probe sets, assuming $\Sigma_1 = \Sigma_2$ is not justified for either fit type. The parental-only fits do not describe segregant behavior well, particularly for (**A**). For (**B**), the supervised parental-only fits exaggerate the degree of separation between the classes—potentially leading to retention of an error-prone probe set.

clusters—and compared these fitted distributions with the main semi-supervised results. Strong disagreement between the fit types, or a significant improvement in fit quality when three or four clusters were used, permitted identification and removal of these aberrant probe sets.

2.4 Comparison to SNPscanner

We compared our ssG results to those of the supervised classifier, SNPscanner (Gresham *et al.*, 2006). The purpose of this comparison was two fold. First, we were interested in exploring the extent to which SNPscanner's statistical model, trained on parental data, could predict the behavior of probes in a different genomic context. Second, we were interested in knowing which algorithm provided better genotyping data.

The SNPscanner algorithm was designed to work with arrays which only interrogate the reference genome (S288c). In addition, the SNPscanner algorithm uses loess-type (Cleveland, 1979) normalization instead of the experiment-wide VSN we used in Mancera *et al.* (2008), and was originally trained on a different array design. To facilitate comparison, we carried out a secondary ssG analysis using only S288c-specific probes, and following the SNPscanner normalization strategy. We retrained SNPscanner on our parental hybridization data, using only SNPs meeting its 'high-quality' criterion. We then used the so-trained model to genotype our segregant arrays, and passed results through the same SNPscanner quality filters used in Gresham *et al.* (2006) (see Supplementary Methods). Filtered ssG and SNPscanner genotyping results were compared on the basis of call rate, concordance and accuracy.

3 RESULTS

Application of ssG to the data described in Section 2.1 permitted assessment of (i) the relationship between supervised and semi-supervised approaches to the genotyping task, (ii) the importance of quality filtering for genotyping accuracy and (iii) the differences between ssG and SNPscanner's model-based, supervised approach.

3.1 Parental versus offspring hybridizations

Probe set behavior in parental hybridizations—the only source of training data available to a supervised classifier—was often not representative of behavior in offspring hybridizations. Figures 2 and 3 provide typical examples. In Figure 3A, parental distributions, while inaccurate, nonetheless lead to correct classification; in Figure 3B, however, a substantial fraction of offspring would be classified as S96 using the parental data but as YJM789 using the semi-supervised approach. Further, the parental data are better separated than the offspring, leading to an overly optimistic assessment of confidence in the probe set's genotype calls.

3.2 Filtering

One objective of Mancera *et al.* (2008) was the characterization of short non-crossover gene conversion events. The number of putative small events seen in unfiltered ssG (Fig. 4A) or SNPscanner (data not shown) calls, however is far too large given our understanding of the biology. The ssG filters discussed in Section 2.3 removed most small events (Fig. 4B). Importantly, these filters are based only on properties of the inferred distributions $\hat{\phi}_1$ and $\hat{\phi}_2$, not on event size; therefore, they are not biased against small events.

As validation, sequencing-based genotype calls were obtained for 283 markers involved in or immediately adjacent to putative small events observed in the unfiltered ssG data (see Supplementary Methods). Figure 5 shows that $\approx$10% of these unfiltered ssG calls were genotyping errors, but that the ssG filters removed all mistakes but one. As a second validation of our algorithm, array data for one wild-type tetrad were produced twice, in separate laboratories. Among the $\approx$163 000 genotype calls which passed ssG filters in both cases, one disagreement was found. The set of filtered calls, however, differed substantially: 23.8% of calls were made in one

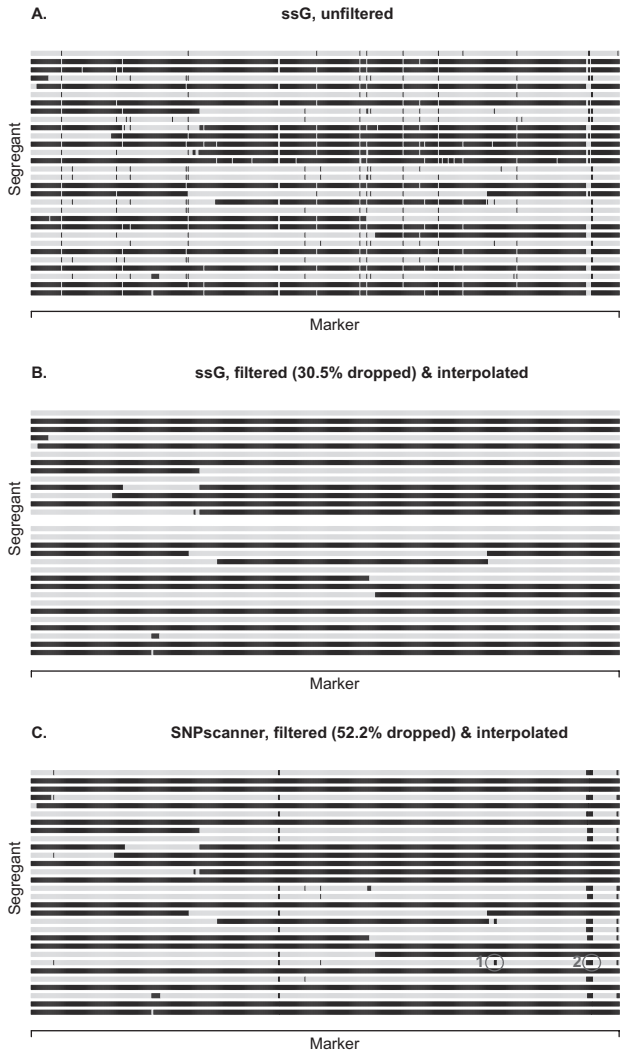

Fig. 4. Typical genotype call behavior (30 segregants and 600 markers on chromosome IV). (**A**) Unfiltered ssG genotype calls include numerous putative single-marker genotype switches, as well as multi-marker regions with a large excess of one genotype. (**B**) Array-, polymorphism- and call-level filtering reduced marker density (by 30.5%), but also substantially reduced the error rate, as shown in Figure 5. Most short events vanish, even though none of ssG's filters explicitly removes short events. (**C**) SNPscanner's heuristic filters discard substantially more calls (52.5%). Results are largely in agreement with those of ssG, but more putative short events remain. For two, identified in red, we examine probe behavior more closely in Figure 6.

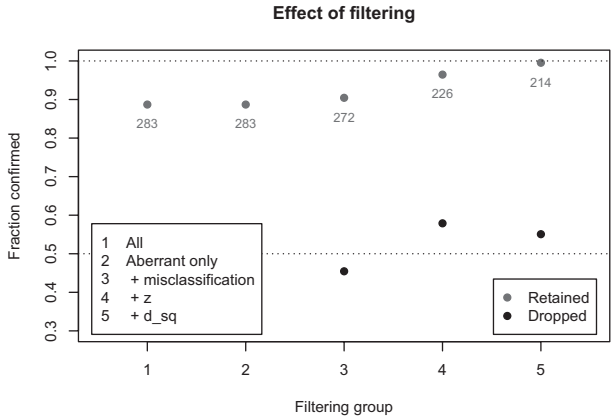

Fig. 5. The effect of post-processing filters on concordance with 283 sequencing-confirmed genotype calls. Filters are applied sequentially; red numbers show calls passing applied filters. Group 1 is unfiltered data. Groups 2 and 3 use polymorphism-level filters based on, respectively, aberrant (three-class) behavior and excess distributional overlap. Groups 4 and 5 use call-level filters based on, respectively, proximity to decision boundary between classes and outlier status with respect to assigned class. Black plot corresponds to calls *removed* by the filters: accuracy here is near 50%, consistent with random guessing, suggesting that we are not over-filtering.

replicate but filtered in the other. Thus, ssG's filters were able to effectively adapt to varying array behavior.

3.3 Comparison to SNPscanner

When both ssG and SNPscanner employed their native filters, ssG made 45% more calls than SNPscanner, producing significantly denser effective marker coverage. A visual comparison of Figure 4B and C shows, however, that ssG and SNPscanner typically made the same genotype call. Indeed, the methods only disagreed in 0.1% of the cases in which both made a genotype call. While this fraction is small, it constitutes thousands of potentially spurious short gene conversion events in our experiment—a number which would overwhelm the amount of true positive signal expected—or thousands false positives in the SNP discovery context.

Are the short events identified by SNPscanner in Figure 4C correct? Figure 6 shows SNPscanner's estimated distributions for two such events, circled in red in Figure 4C. In the first case, segregants formed two clusters, but YJM789 probe behavior did not match SNPscanner's prediction, and sequencing confirmed that SNPscanner's genotype call was in error. In fact, such discrepancies between SNPscanner's predictions and observed behavior were apparent for most probe sets. Often, segregants carrying the YJM789 allele produced data lying *closer* to the misplaced YJM789 estimate than to the S96 estimate, so SNPscanner made a correct genotype call. In other cases like Figure 6A, however, errors arose; and in all cases, one expects that polymorphism- and call-level filtering was negatively impacted by mis-estimation of the distributions. In the second circled example in Figure 4C, SNPscanner assigned nearly all tetrads the S96 genotype. Figure 6B, however, strongly suggest that this locus is not actually polymorphic. If so, then SNPscanner's calls are correct in a sense; accepting such calls at face value, however creates erroneous short events for segregants which have inherited YJM789 alleles in this region. ssG avoids this problem by applying a polymorphism-level filter which sets aside markers which fail to generate two well-separated distributions.

When applied to the replicated tetrad discussed in the previous section, ssG produced just one discrepant call across the four spores. SNPscanner, on the other hand, produced discrepancies for 13% of the markers at which it made a call in both replicates. SNPscanner's filters were also very sensitive to laboratory effect. When applied to the replicate hybridized in the same laboratory as its training data, SNPscanner filtered S96 and YJM789 calls in roughly equal

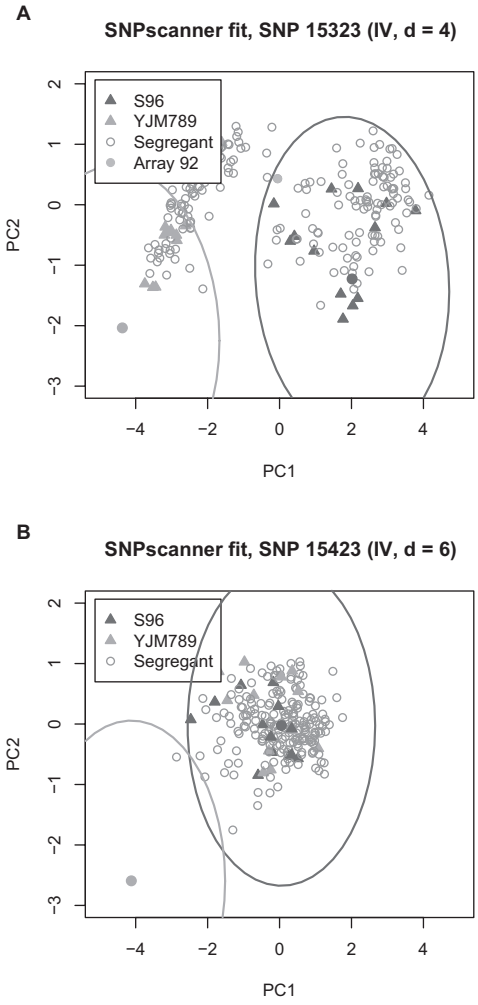

Fig. 6. SNPscanner distributional estimates, projected into the first two PCs. (**A**) The YJM789 prediction does not correctly capture probe behavior for segregants with that genotype, although most apparent YJM789 segregants are *closer* to this distribution, and thus, are genotyped correctly. Sequencing confirmed that SNPscanner's call 1 in Figure 4C, shown here in orange, was in error. Because this call was intermediate to the two empirical distributions, it was dropped by ssG as error prone. (**B**) Call 2 in Figure 4C corresponds to a locus which does not actually appear to be polymorphic.

proportion; when applied to the other replicate, however, it filtered out almost all S96 calls—most likely due to a shift in distributional locations caused by the different conditions. Gresham *et al.* (2006) suggest that, when using SNPscanner to genotype new samples, the training data provided by the authors are sufficient, i.e., that it is not necessary to retrain the model on locally produced training data. The observed sensitivity to laboratory effect for our replicated tetrad, however, suggests that this is not always the case.

3.4 Application: gene conversion

After filtering, it is straightforward to infer the crossover and gene conversion history for each tetrad, on each chromosome. Figure 7 provides one example. In total, our analysis infers approximately 2126 observable, high-confidence gene conversion events and 4163 crossovers. A detailed analysis of these results and their implications for meiotic recombination is reported in Mancera *et al.* (2008).

4 DISCUSSION AND CONCLUSION

Classification and clustering algorithms have traditionally been called supervised and unsupervised approaches, respectively. Supervised classification learns model parameters from labeled training data in one step, then attempts to assign labels to new data in a separate step. Unsupervised clustering, on the other hand, is not given labeled training data; instead, it attempts to divide unlabeled data into sensible groups in a single step.

In this article, we present the multivariate ssG algorithm. While many previously proposed array-based genotyping methods are supervised classifiers, ssG takes a *semi-supervised* approach: it clusters data by genotype in a single step, but in a way that takes advantage of the limited amount of labeled parental data. It is clear from Figures 2 and 3 that parental hybridizations are not always representative, and that offspring intensity data contain a great deal of information about cluster membership and distributional shape. The semi-supervised approach exploits this information—information which is not available to a supervised classifier—and uses it to make more accurate distributional estimates and genotype calls.

We also contrast ssG with SNPscanner, a recently proposed supervised classifier which is also based on multivariate Gaussian mixtures. SNPscanner employs a parametric model to predict the impact of polymorphisms on probe behavior, while ssG uses no such model, relying instead on empirical distributions derived from the clusters it identifies. By using a probe behavior model, SNPscanner attempts to shift statistical testing from the asymmetric case ($H_0: \theta = \theta_0$ vs. $H_A: \theta \neq \theta_0$) to the simpler symmetric case ($H_0: \theta = \theta_0$ vs. $H_A: \theta = \theta_1$). Such a shift is only possible if one can correctly specify θ_1. Figure 6, however, shows that model-based specification is still not a realistic option. Indeed, models which accurately capture important quantities related to microarray probe behavior have remained elusive.

Our results have focused on genotyping in the context of meiotic recombination, but the ssG algorithm is immediately applicable to other contexts. It can be applied to individual probes as well as to probe sets, and can be used with other array designs—i.e. non-tiling arrays, or arrays which interrogate a single genome. Because sequence data were available for both strains considered in this study, it was natural to define probe sets on the basis of known polymorphisms. In general, sequence for the second strain is not required: probe sets may be defined simply on the basis of shared regions of interrogation. In such a case, most probe sets will interrogate non-polymorphic sequence and thus be uninformative, but standard model selection procedures (e.g. Bayesian information criterion (BIC)) appear to be sufficient for identification of sets which exhibit two-class behavior. As shown above, however, two-class behavior is necessary but not sufficient for effective genotyping: some probe sets corresponding to known polymorphisms do not clearly distinguish between the alleles; in other cases, varying genomic background and cross-hybridization may create two-class behavior even when there is no polymorphism at the interrogated locus.

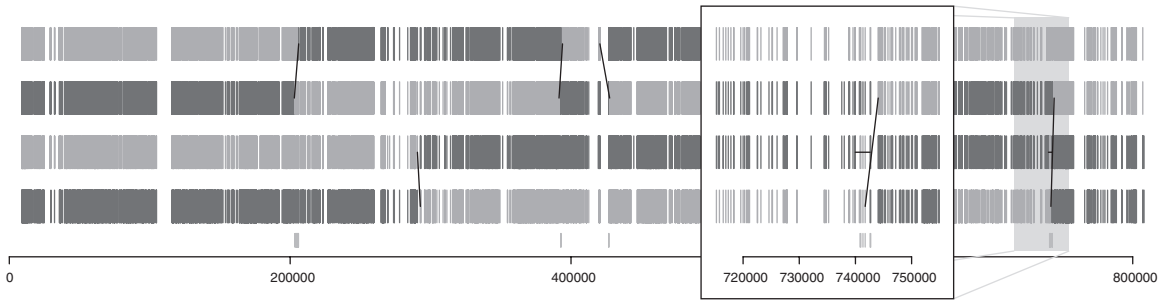

Fig. 7. Inferred recombination events for one tetrad, across chromosome II. Orange bars denote markers with a non-Mendelian ratio, arising from crossover-associated or non-crossover gene conversion. Zoomed regions shows that marker density is sufficient to capture genotype between the two Holliday junctions producing the crossover (see, e.g. de Massy, 2003). Note the unexpected region of apparent gene conversion on a third strand not involved in the crossover.

Our results have important implications for the detection of polymorphisms in novel, unsequenced strains. Detection is typically accomplished by testing the null hypothesis that the novel strain's data have arisen from the same distribution seen in the reference strain. The discrepancies between parental and segregant behavior seen in Figure 3, however, suggest that significant deviation may be observed even in the absence of polymorphism at the interrogated locus. If the novel and reference strains differ little genetically, use of the reference strain distribution is likely to be appropriate, provided that other sources of experimental variation have been appropriately controlled or normalized out. If, on the other hand, the novel strain is substantially diverged, naive comparisons to the reference strain distribution may run into serious false positive problems.

ACKNOWLEDGEMENTS

We thank contributors to the Bioconductor (www.bioconductor.org, Gentleman *et al.*, 2004) and R (www.R-project.org) projects for their software. We also thank Zhenyu Xu and Paul McGettigan for providing insight in the early stages of the project.

Funding: Deutsche Forschungsgemeinschaft; National Institutes of Health (to L.M.S.).

Conflict of Interest: none declared.

REFERENCES

Albert,T.J. *et al.* (2005) Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nat. Methods*, **2**, 951–953.

Borevitz,J.O. *et al.* (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.*, **13**, 513–523.

Brem,R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **39**, 1–38.

Deutschbauer,A.M. and Davis,R.W. (2005) Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat. Genet.*, **37**, 1333–1340.

de Massy,B. (2003) Distribution of meiotic recombination sites. *Trends Genet.*, **19**, 514–522.

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

Gresham,D. *et al.* (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science*, **311**, 1932–1936.

Gu,Z. *et al.* (2005) Elevated evolutionary rates in the laboratory strain of saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. USA*, **102**, 1092–1097.

Mancera,E. *et al.* (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, **454**, 479–485.

Rostoks,N. *et al.* (2005) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol.*, **6**, R54.

Steinmetz,L.M. *et al.* (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, **416**, 326–330.

Turner,T.L. *et al.* (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.*, **3**, e285.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wei,W. *et al.* (2007) Genome sequencing and comparative analysis of saccharomyces cerevisiae strain yjm789. *Proc. Natl Acad. Sci. USA*, **104**, 12825–12830.

Winzeler,E.A. *et al.* (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.

Winzeler,E.A. *et al.* (2003) Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics*, **163**, 79–89.