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

Richard Bourgon1,∗, Eugenio Manca2, Alessandro Brozzi1, Lars M. Steinmetz2 and Wolfgang Huber1

1EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and 2European 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 (Manca 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 loci. 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 targets.

Several 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 targets.

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.

Let’s denote by \( x \) the value of some probe, \( \hat{x} \) the expected value of \( x \) under a Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \), \( \hat{x} = \frac{1}{\sigma} \). Then, the probability of observing \( x \) is given by

\[
P(G = g | \text{intensity } = x) = \frac{P(x)}{P(x) + \hat{P}(x)}
\]

Finally, the estimated Gaussian densities from the parental-array ANOVA, plus a uniform prior on the two genotypes:

\[
P(G = g | \text{intensity } = x) = \frac{\hat{P}(x)}{\hat{P}(x) + \hat{P}(x)}
\]

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 as 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...
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 mls0 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 \(A, B, S\) 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 \(X_i \sim N(\mu_1, \Sigma_1)\) for \(i \in A\), \(X_i \sim N(\mu_2, \Sigma_2)\), and that \((X_i, Y_i)\) was independent of \((X_j, Y_j)\) for \(i \neq j\). Importantly, we did not require \(\mu_1\) and \(\mu_2\) to differ from \(\mu_j\) 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, \(\Sigma_1\) and \(\Sigma_2\) 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.
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 $P(Y_i = \text{g}(X_i))$ for all
$i \in S$. To initialize these conditional probabilities (hereafter denoted
$\hat{p}_{\text{g}}$), we applied a simple clustering algorithm—k-means, with
the two clusters seeded with parental observations—so the combined
parental and segregant data, and then set each $\hat{p}_{\text{g}}^{(0)}$ 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 $\Sigma_g$ using the
parental data; this strategy produced identical results.) Defining $p_{\text{g}} = \sum \hat{p}_{\text{g}}$, it is straightforward to show that the M step’s objective function is
maximized in $\hat{\mu}_g$ by
\begin{equation}
\hat{\mu}_g^{(1)} = \frac{1}{P^g} \sum_{i=1}^{n} p_{\text{g}}^{(0)} X_i,
\end{equation}
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 $\Sigma_g$ by
\begin{equation}
\hat{\Sigma}_g^{(1)} = \frac{1}{P^g} \sum_{i=1}^{n} p_{\text{g}}^{(0)} (X_i - \hat{\mu}_g^{(1)}) (X_i - \hat{\mu}_g^{(1)})^T,
\end{equation}
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
inhabit either of the two parental alleles, so we fixed $\pi_1 = \pi_2 = 0.5$.
In other contexts, $\pi_1$ and $\pi_2$ can easily be estimated. To begin the
next iteration, we updated $\hat{p}_{\text{g}}$ for all $i \in S$, by
\begin{equation}
\hat{p}_{\text{g}}^{(1)} = \frac{\phi(X_i; \hat{\mu}_g^{(1)}, \hat{\Sigma}_g^{(1)})}{\phi(X_i; \hat{\mu}_g^{(0)}, \Sigma_1^{(0)}) + \phi(X_i; \hat{\mu}_g^{(0)}, \Sigma_2^{(0)})},
\end{equation}
and continued until a convergence criterion was met. Here, $\phi(\cdot; \mu, \Sigma)$
denotes the density of a multivariate normal distribution with mean
$\mu$ and covariance matrix $\Sigma$. We also define $\phi_2$ to be $\phi(\cdot; \hat{\mu}_g, \hat{\Sigma}_g)$.

Final assignment of genotype for the segregants was then obtained
by comparing $p_{1,2}$ and $p_{3,4}$. 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 calls 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 $p_{1,2}$ which were
too far from both 0 and 1—were removed. Individual calls with unambiguous $p_{1,2}$ 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
supervised approaches to the genotyping task, (ii) the importance
quality filters used in Gresham et al. (2008), and was originally trained on a different
Mancera et al. (2008), and was originally trained on a different
semi-supervised results. Strong disagreement between the fit types,
clusters—and compared these fitted distributions with the main
2.4 Comparison to SNPscanner
trained SNPscanner's model-based, supervised approach.
3.1 Parental versus offspring hybridizations
one wild-type tetrad were produced twice, in separate laboratories.
the parental-only fits exaggerate the degree of
of quality filtering for genotyping accuracy and (iii) the differences
among the ≈163,000 genotype calls which passed ssG filters in
3 RESULTS
Mutation, ID 43 (d = 3)
Semi−supervised
Parental only
Mutations, ID 417 (d = 2)
Semi−supervised
Parental only
3.2 Filtering
one wild-type tetrad were produced twice, in separate laboratories.
Among the ≈163,000 genotype calls which passed ssG filters in
the ssG analysis using only S288c-specific probes, and following the
Supplementary Methods). Filtered ssG and SNPscanner genotyping results were
compared on the basis of call rate, concordance and accuracy.
3.3 Parental versus offspring hybridizations
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.
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
Prob 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 \( \phi_1 \) and \( \phi_2 \), 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 ≈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.

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
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 2,126 observable, high-confidence gene conversion events and 4,163 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 $\theta_1$. 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.

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
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


