OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing

Paul M. Magwene[1]*, John H. Willis[2], John K. Kelly[3]

1 Department of Biology and IGSP Center for Systems Biology, Duke University, Durham, North Carolina, United States of America, **2** Department of Biology, Duke University, Durham, North Carolina, United States of America, **3** Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas, United States of America

Abstract

We describe a statistical framework for QTL mapping using bulk segregant analysis (BSA) based on high throughput, short-read sequencing. Our proposed approach is based on a smoothed version of the standard G statistic, and takes into account variation in allele frequency estimates due to sampling of segregants to form bulks as well as variation introduced during the sequencing of bulks. Using simulation, we explore the impact of key experimental variables such as bulk size and sequencing coverage on the ability to detect QTLs. Counterintuitively, we find that relatively large bulks maximize the power to detect QTLs even though this implies weaker selection and less extreme allele frequency differences. Our simulation studies suggest that with large bulks and sufficient sequencing depth, the methods we propose can be used to detect even weak effect QTLs and we demonstrate the utility of this framework by application to a BSA experiment in the budding yeast *Saccharomyces cerevisiae*.

Citation: Magwene PM, Willis JH, Kelly JK (2011) The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. PLoS Comput Biol 7(11): e1002255. doi:10.1371/journal.pcbi.1002255

Editor: Adam Siepel, Cornell University, United States of America

Received April 15, 2011; **Accepted** September 13, 2011; **Published** November 3, 2011

Funding: This research was supported by NIH grant P50GM081883-04 (to PMM), NIH grant R01-GM073990 (to JKK and JHW), NSF grant DEB-10-19753 (to PMM) and NSF grant IOS-10-24966 (to JHW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: paul.magwene@duke.edu

Introduction

Bulk segregant analysis (BSA; [1]) is a QTL mapping technique for identifying genomic regions containing genetic loci affecting a trait of interest. Starting with a segregating population from a genetic cross, individuals are assayed for the focal trait and two pools (bulks) of segregants are created by selecting individuals from the tails of the phenotypic distribution (other sampling designs can also be used as discussed below). Genotype frequencies are estimated for the two bulks, either via genotyping of individuals or via the creation of pooled DNA samples from which allele frequencies are estimated. Allele frequencies should be approximately equal between the two bulks in genomic regions without loci affecting the trait. Regions of the genome containing causal loci should exhibit allele frequency differences between bulks. BSA is most effective with high marker density and accurate allele frequency estimation within bulks [2]. The former was effectively addressed with the application of microarray based genotyping to BSA [3–8]. More recently, investigators have begun to use massively parallel sequencing methods to estimate allele frequencies for BSA studies [9–11], which has a number of advantages. For organisms with moderately sized genomes, next generation sequencing can provide essentially single base-pair resolution. In such cases rather than simply observing markers in linkage with causal loci the BSA-sequencing approach should allow one to observe allelic biases at the causal loci themselves. For larger genomes where high coverage of the entire genome is less practical, BSA-sequencing still has many potential advantages. For example, it does not require the design of new genotyping arrays for new crosses and may provide greater resolution than array based genotyping. Furthermore, sequencing data yields counts of alleles at polymorphic loci and thus provides a simple and intuitive way of estimating allele frequencies.

In bulk segregant studies based on high-throughput sequencing there are two sources of variation that affect allele frequency estimates. The first is variation due to the sampling of segregants that constitute the bulks themselves. This source of variation can be minimized by increasing both the size of the segregant population and the size of the bulk samples. The second source of variation is a consequence of the measurement technique used to estimate allele frequencies in the bulks. In the case of sequencing of pooled DNA samples, the sources of variation of this second type include, but are not limited to, library preparation, sequencing chemistry, sequencing coverage, post-sequencing alignment of reads, and base/allele calling algorithms. Here again, some of these sources of variation can be minimized by standardization of experimental protocols and analysis pipelines. However some of these sources of variation, particularly stochasticity in sequencing coverage, are an inherent property of short-read sequencing methods.

In this paper, we develop explicit statistical models to describe the sources of variation that should be considered in the analysis of BSA-sequencing data. We first develop test statistics based on the classic G -statistic accounting for the two phase sampling inherent to BSA. We then propose an analysis pipeline for whole-genome studies and present a proof-of-concept example with data from yeast. A combination of simulation and empirical application demonstrate the utility of this analytical framework.

Author Summary

Quantitative or complex phenotypes are traits that are under the control of multiple genes and environmental factors. Identifying the parts of the genome that contribute to variation in complex traits (Quantitative Trait Loci or QTLs), and ultimately the genes and alleles that are mechanistically responsible for trait variation, is a primary challenge in animal and plant breeding, population studies of human health and disease, and evolutionary genetics. In this study we describe an analytical framework that allows investigators to marry a QTL mapping approach called "bulk segregant analysis" (BSA) with high-throughput genome sequencing methodologies in order to map traits quickly, efficiently, and in a relatively inexpensive manner. This framework provides a statistical basis for analyzing BSA experiments that use next-generation sequencing and will help to accelerate the identification of QTLs in both model and non-model organisms.

Results

Theory and Analytical Framework

Expected distribution of G for BSA-sequencing data. Consider the experimental design with an F_2 population consisting of N individuals, each of which is measured for a phenotype of interest. A set of n_s individuals from each of the tails of the distribution (low and high) are collected. DNA bulks are prepared by combining equal amounts of tissue/cells from individuals within each bulk followed by DNA extraction, or by extracting DNA from each individual and combining equal amounts. Following preparation of DNA bulks, genomic libraries are prepared and sequenced at average coverage C per SNP. Thus for each SNP the data is four allele counts that can be summarized in a 2×2 table, where A_1 is the allele from the high parent (Table 1). The n_i-values in the table are counts of alleles not individuals. The observed allele frequency of A_1 in the low bulk is $p_1=n_3/(n_1+n_3)$; that in the high bulk is $p_2=n_4/(n_2+n_4)$. If the SNP is close to a QTL with effects in the expected direction (i.e. the 'high allele' increases trait values), then we expect $p_2 \gg p_1$.

The counts in Table 1 are determined by two levels of hierarchical of sampling. The first sample is the $2n_s$ chromosomes that constitute each bulk (assuming diploid inheritance). Second, there is random variation in the number of reads per allele within each bulk due to the stochastic nature of next-generation sequencing. Let ρ_1 and ρ_2 be the expected ('true') frequency of the high allele in each bulk. The realized frequencies (p_1^*, p_2^*) differ from ρ_1 and ρ_2 in each bulk due to binomial sampling:

$$2n_s p_1^* \sim \text{Binomial}(2n_s, \rho_1) \tag{1}$$

Table 1. The summary of data from a single variable site.

	Low bulk	High bulk	Total
A_0	n_1	n_2	n_1+n_2
A_1	n_3	n_4	n_3+n_4
Total	n_1+n_3	n_2+n_4	

The n_i represent counts of alleles A_0 and A_1 generated from sequencing of the segregant bulks.

doi:10.1371/journal.pcbi.1002255.t001

$$2n_s p_2^* \sim \text{Binomial}(2n_s, \rho_2). \tag{2}$$

If we assume that sequencing coverage is approximately Poisson, then the conditional distributions of the observed allele counts are:

$$n_1 | p_1^* \sim \text{Poisson}(C[1-p_1^*]) \tag{3}$$

$$n_2 | p_2^* \sim \text{Poisson}(C[1-p_2^*]) \tag{4}$$

$$n_3 | p_1^* \sim \text{Poisson}(Cp_1^*) \tag{5}$$

$$n_4 | p_2^* \sim \text{Poisson}(Cp_2^*) \tag{6}$$

A natural statistic to characterize the data at each SNP is the standard G -statistic:

$$G = 2\sum_{i=1}^{4} n_i \ln\left(\frac{n_i}{\hat{\boldsymbol{n}}_i}\right) \tag{7}$$

where $\hat{\boldsymbol{n}}_i$ is the 'expected value' for count n_i. The null hypothesis is that there is no QTL close to the focal SNP. This implies the standard expected counts for a 2×2 contingency table, e.g. $\hat{\boldsymbol{n}}_1=(n_1+n_2)(n_1+n_3)/(n_1+n_2+n_3+n_4)$. If the null hypothesis is correct, $E[n_1]=E[n_2]$ and $E[n_3]=E[n_4]$. If we further assume no segregation distortion and equal (average) sequencing coverage of each bulk, then $E[n_1]=E[n_2]=E[n_3]=E[n_4]=C/2$. See the supplementary materials (Text S1) for a generalization that includes segregation distortion.

However, due to the hierarchical sampling scheme, the usual expectation that G follows a χ_1^2 distribution (chi-square with 1 d.f.; [12]) does not hold in the present situation. The mean and variance of G are inflated relative the χ_1^2 even when the null hypothesis is true (i.e. there is no QTL). Based on the arguments in Text S1 we approximate the mean and variance of G as:

$$E[G] \approx 1 + \frac{C}{2n_s} \tag{8}$$

$$\text{Var}[G] \approx 2 + \frac{1}{2C} + \frac{1+2C}{n_s} + \frac{C^2(4n_s-1)}{8{n_s}^3} \tag{9}$$

These equations predict convergence on χ_1^2 under certain parameter sets. In particular, if $n_s \gg C \gg 1$, then $E[G]\to 1$ and $\text{Var}[G]\to 2$, as expected from χ_1^2.

A simulation model was used to test the accuracy of approximate equations (8) and (9). We simulated genetic data for a chromosomal region of 10 cM in recombinational length. Informative markers were uniformly distributed along this chromosome with d SNPs per cM. The causal locus (QTL) was located at the center of the chromosome and was thus flanked by $5d$ SNPs on each side. Alternative homozygotes at the QTL differ by $2a$ phenotypic units on average (additive gene action) and simulations of the null hypothesis (no QTL) were done with $a=0$. In each simulation run, we first established the genotypes and phenotypes of the N distinct F_2 segregants. Each individual was

assigned a QTL genotype according to Mendelian probabilities (0.25, 0.5, 0.25) and the phenotype was assigned as the genotypic value plus a normal deviate. Individuals were then ranked by phenotype and n_s were selected from each tail. The full haplotype of these individuals was then established by working out from each allele at the QTL and allowing recombination to occur probabilistically according to the linkage map. Given the haplotypes in each bulk, we simulated an independent Poisson number for each count of Table 1 for each SNP. These data were used to calculate G at each SNP, and also G' as described below, within windows around each SNP. For the latter we needed to specify a window size in centimorgans. For each parameter set, this entire procedure was repeated 10,000 times. Table 1 in Text S1 reports simulation results for the null hypothesis ($a=0$) for a range of reasonable combinations of C and n_s. There is a close correspondence of observed means and variances of G with the values predicted by equations (8) and (9). As expected, in these simulations the distribution of G is right skewed with a mean and variance exceeding the χ^2_1 expectations.

The full distribution of G values is depicted for one parameter set ($n_s=150$, $C=50$) in Figure 1a. The gray histogram shows the distribution of G under the null hypothesis ($a=0$) while the overlapping red histogram shows the corresponding distribution in the case of a weak QTL ($a=0.02$). Focusing first on the null distribution, because the distribution is right skewed (mean = 1.19, variance = 2.93), if we compare this distribution to critical values of χ^2_1 the observed false positive rate is somewhat elevated (6.98% at $p=0.05$; 1.98% at $p=0.01$). However when C approaches n_s the mean and variance of G far exceed the χ^2_1 expectation and type I error rates increase dramatically. Perhaps even more problematic is the inability of G to detect a QTL based on the naïve χ^2_1 expectation. For the weak QTL case, where the QTL explains 2% of the phenotypic variance, the causal SNP is significant at a $p=0.05$ in only 34.9% of the simulations, and in only 16.8% of simulations at $p=0.01$. The application of the naïve χ^2_1 thus suffers from a lack of power.

G', A Smoothed Version of G. A substantial source of variation in G is the random margin in Table 1, $n_1+n_3, n_2+n_4 \sim \text{Poisson}(C)$. To deal with this variation we propose the use of a weighted average of G across neighboring SNPs. Averaging G values across SNPs is sensible because the real signal of divergence in allele frequency between bulks is conserved between closely linked sites but random noise due to variable sequencing read coverage is not. We suggest the following average test statistic for each SNP:

$$G' = \sum_{j \text{ in } W} k_j G_j \tag{10}$$

where the sum includes all SNPs within the window W bracketing the SNP. This type of weighted moving average, where the weights are given by a kernel function, k, is also known as Nadaraya-Watson kernel regression [13,14]. Nadaraya-Watson kernel regression acts as a smoothing function, with the amount of smoothing increasing with larger window size W [15]. The simplest scheme for k_j would be to give equal weight to all SNPs within W (a rectangular kernel). We opt instead to apply the tri-cube kernel fuction:

$$k_j = \frac{(1-D_j^3)^3}{S_W} \tag{11}$$

where D_j is standardized distance, with value 0 at the focal position and value 1 at the edge of the window. S_W is the sum of $(1-D_j^3)^3$ for all SNPs in W. The tri-cube kernel is commonly used in local polynomial regression methods like LOESS [16] and gives greater weight to observations that are close to the focal SNP. Any other weighting kernel that decreases smoothly to 0 as D_j goes to 1 could be used as well. We discuss the choice of the kernel window size, W, below.

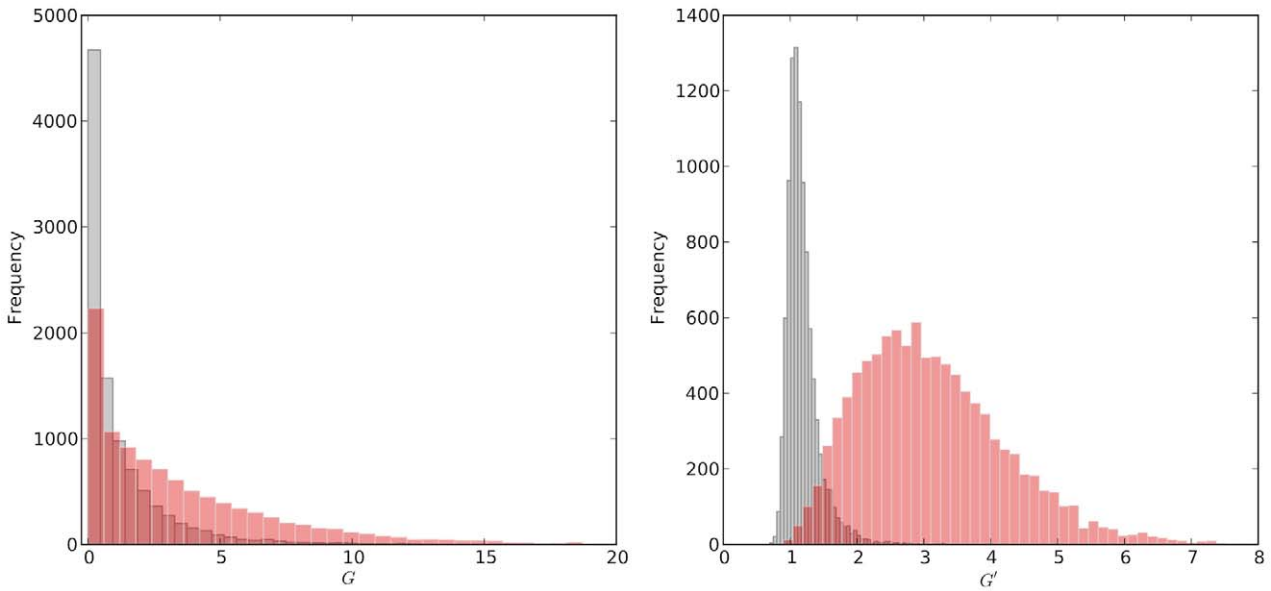

Figure 1. The distribution of G (A) and G' values (B) from 10,000 simulations. The gray histograms depict the observed distributions of G and G' for the null case (no QTL), while the red distributions depict the distributions in the case of a weak QTL that explains 2% of the phenotypic variance.
doi:10.1371/journal.pcbi.1002255.g001

A methodological issue arises when kernel smoothing is used – at the beginning or end of a data series it can produce a biased estimate because the data included in the kernel bandwidth is asymmetric. The simplest way to deal with this is to append a reflected version of the values that fall within the right half-bandwith (at the beginning of the series) and left half-bandwidth (at the end of the series), run the kernel smoother as normal, and then trim the appended values from the output.

Expected distribution of G' for BSA-sequencing data. The null expectation of G' is given by equation (8). The variance of G' depends on the variance of individual G values (equation 9) and the covariance between SNPs within a window. In Text S1 we show that $\mathrm{Var}[G']$ can be approximated as:

$$\mathrm{Var}[G'] = \left(2 + \frac{1}{2C} + \frac{1+2C}{n_s} + \frac{C^2(4n_s-1)}{8n_s^3}\right)\sum_{j \text{ in } W} k_j^2 + \sum_{j \text{ in } W}\sum_{i \neq j} \frac{C^2(4n_s-1)}{8n_s^3}(1-2r_{ij})^2 k_i k_j \quad (12)$$

where i indexes all SNPs other than j contained within the window.

Figure 1b illustrates the distribution of G' for the same parameters as Figure 1a (plus window size $W = 20$ cM and SNP density $d = 10$ per cM). The difference between the null distributions in Figure 1a and 1b is due to the normalizing effect of averaging. The predicted mean and variance of G' (1.17 and 0.066) are reasonably close to the observed moments (1.18 and 0.056). The distribution of G' is still right skewed but the right tail can reasonably predicted from log-normal densities with parameters derived from $E[G']$ and $\mathrm{Var}[G']$ (Figure S1 and Text S2). The observed false-positive rates (using a log-normal density estimation) are: 5.14% at $p = 0.05$ and 1.86% at $p = 0.01$). Unlike the use of the naive G -test based on χ^2_1, the type I error does not increase dramatically as C approaches n_s. Furthermore, G' has good power to detect QTLs. For the example illustrated in Figure 1b the causal SNP is significant in 94.3% of the simulations at $p = 0.05$, and in 88.0% and 77.2% of simulations at $p = 0.01$ and $p = 0.001$ respectively.

Non-parametric estimation of the null distribution of G'. In addition to the theoretical expectations discussed above, an empirical estimate of the null distribution of G' can be derived from the observed data itself. We assume that the observed data, $X_{G'}$, is a mixture of the null distribution (non-QTL regions) and several contaminating distributions (QTLs). As discussed above, the null distribution of G' ($\theta_{G'}$) is right-skewed with a tail density reasonably predicted from a log-normal distribution, $\theta_{G'} \sim \ln \mathcal{N}(\mu, \sigma^2)$. We also assume the contaminating distributions have higher means than the null distribution. Our goal is to estimate μ and σ^2 in a manner that is not unduly influenced by the contaminating distributions.

Recall that for a log-normal distribution: $\mathrm{Median} = e^{\mu}$ and $\mathrm{Mode} = e^{\mu - \sigma^2}$ [17]. Thus if we can estimate the median and mode of $\theta_{G'}$ can use those to estimate μ and σ^2. To do so we propose the folowing steps:

1. Let $W_{G'} = \ln[X_{G'}]$
2. Let $s_W = \mathrm{MAD}_l(W_{G'})$, the left median absolute deviation (MAD) of W_G where MAD_l is defined as

$$\mathrm{MAD}_l(Y) = \mathrm{Median}(|y_i - \mathrm{Median}(Y)|) \quad \text{for all } y_i \leq \mathrm{Median}(Y)$$

3. Use Hampel's rule [18] to identify outliers, $\mathbf{O}_W$, as all w_i in $W_{G'}$ that satisfy:

$$w_i - \mathrm{Median}(W_{G'}) > g(N,\alpha_N)\mathrm{MAD}_l(W_{G'})$$

where $g(N,\alpha_N)$ defines the limits of the outlier regions [18] and is usually taken to be 5.2 for normally distributed data.

4. Construct a trimmed data set $X_T = \{x_i\}$ for all i such that $w_i \notin \mathbf{O}_W$
5. Let $\hat{\mu}_\theta = \ln[\mathrm{Median}(X_T)]$ and $\hat{\sigma}^2_\theta = \hat{\mu}_\theta - \ln[\mathrm{Mode}_r(X_T)]$ where $\mathrm{Mode}_r(X_T)$ is a robust estimator of the mode for continuous variables (see [19] for several such estimators)

The logic of this procedure is as follows. The median and MAD are robust estimators of location and spread respectively [20]. In the absence of contaminating distributions $W_{G'}$ should be approximately normally distributed, and hence the median and MAD of $W_{G'}$ can be used as robust estimates of the mean and spread of $\theta_{G'}$ ($\mathrm{MAD}_l \approx MAD$ for a symmetric distribution). Hampel's rule is a commonly used procedure to identify likely outliers in a set of data based on the median and MAD; if the underlying distribution is normally distributed and $g(N,\alpha_N) = 5.2$ this is approximately equivalent to identifying outliers as those observations with p-values < 0.001 (we use a one-sided test in the procedure above). When contaminating distributions (QTLs) are present, $\mathrm{Median}(W_{G'})$ lies to the right of the true mean of the null distribution. Thus, $\mathrm{Median}(W_{G'})$ and $\mathrm{MAD}_l(W_{G'})$ are conservative estimators of $\mathrm{Median}(\theta_{G'})$ and $\mathrm{MAD}(\theta_{G'})$. We then use Hampel's procedure to identify observations likely to be drawn from the contaminating distributions and create a trimmed data set, X_T, with those outlying observations removed. From the trimmed data set we estimate $\mathrm{Median}(\theta_{G'})$ and $\mathrm{Mode}(\theta_{G'})$.

For the null simulations in Figure 1b the observed false-positive rate estimated using this non-parametric approach are 3.18% at $p = 0.05$ and 0.76% at $p = 0.01$. In general, the non-parameteric procedure tends to be slightly more conservative than our proposed parametric estimators but not greatly so. Because this non-parametric approach makes few distributional assumptions (other than approximate log-normality of the null distribution) it might be preferred in cases where one suspects the sampling (either of segregants or alleles) grossly violates the hierarchical model described above.

Choosing W. A weighted moving average is a type of low-pass filter; the larger the window size the lower the frequncy of signals that are rejected by the filter. The choice of smoothing width, W, is therefore a tradeoff between filtering out high-frequency deviations in G due to variable sequence coverage and SNP density and attenuating the signal of real QTLs. We want to pick a W that minimizes noise while maximizing the underlying signal. The matched filter theorem [21] suggests that the filter that maximizes the signal-to-noise ratio of a symmetric signal is one which matches the shape of the signal. A simple measure of the shape of a symmetric signal is the full-width at half maximum (FWHM). The ratio of the width of the kernel to the peak FHWM ('smoothing ratio') is a useful metric for quantifying the effects of smoothing [22]. As a rule of thumb, using a smoothing kernel with a smoothing ratio of approximately two provides a good signal-to-noise ratio [22]. However, the matched filter may fail to distinguish multiple peaks when there are two or more signals in the input [23] as we would expect in cases of multiple QTLs with overlapping regions of elevated G. Specifically, peaks separated by less than twice the FWHM of the filter will be merged [24].

Therefore, to distinguish overlapping signals requires filters with smoothing ratios significantly smaller, perhaps as small as 0.7.

In Text S2 we derive the expected shape of G around a single causal SNP. For the case in which the causal allele is fixed in one bulk and has a frequency of 0.5 in the other bulk, the half-bandwidth ($h = W/2$) at half-maximum corresponds to $\sim$12.42 cM ($r \sim 0.11$). More extreme allelic biases between the bulks favor slightly smaller bandwidths, while less extreme differences favor larger bandwidths. SNP density also affects the optimal kernel bandwidth, with higher SNP density favoring narrower bandwidths. In simulations and applied to real data we have found that kernels with smoothing ratios in the range 1–1.5 produce smoothed estimators with good signal-to-noise ratios and which are neither strongly over- or undersmoothed. In terms of mapping distances this corresponds to kernels with W in the range $\sim$24.8–37.25 cM.

Since recombination rates vary across genomes, a given genetic distance will correspond to a range of physical distances. In terms of the choice of smoothing width, higher recombination rates favor smaller window sizes (in physical distance). If regional recombination rates are known this can be incorporated into the analysis; however the use of average chromosomal or genomic recombination rates to choose a single physical size for the smoothing window should not be problematic unless recombination rates vary widely. In such cases, one can calculate G' using a range of smoothing widths to explore whether peak estimates are strongly affected by over- or undersmoothing.

Proposed Analytical Pipeline

Based on the arguments developed above, we propose the following analytical pipeline for the analysis of BSA-sequencing data sets. We assume that sequencing reads have been aligned to a reference genome where physical distances between polymorphic sites and (approximate) rates of recombination are known. We assume that all sites are biallelic. Following alignment of reads to a reference genome, per site counts of each allele are generated from the reads. Our recommended analysis pipeline for estimating QTLs is as follows:

1. For each variable site, calculate G based on the observed number of reads for each allele in each of the two pools
2. At each site calculate G' using a smoothing kernel with bandwidth W bases where W is chosen based on known or estimated rates of recombination. Bandwidths should typically correspond to genetic map distances in the range 25–40 cM.
3. Estimate parameters of the log-normal null-distribution (i.e. no QTL) of G', $\theta_{G'} \sim \ln\mathcal{N}(\mu_{G'}, \sigma^2_{G'})$, based on either theoretical expectations (equations (8) and (12 and Text S2) or using the robust empirical estimator of the null distribution inferred from the observed G'.
4. Using $\theta_{G'}$ estimate p-values directly using the log-normal CDF. Alternately log-transform G' and calculate Z scores $G'_{Z,i} = (\ln(G'_i) - \mu_{G'})/\sigma_{G'}$ and corresponding p-values at each site.
5. Use a false discovery rate approach (FDR; [25,26]) to account for multiple comparisons and estimate an appropriate p-value threshold (or the corresponding G' threshold) to determine sites that deviate significantly from the background null distribution
6. Define candidate QTL regions as continuous runs of significant sites

Power Analysis

We used simulations to conduct a simple power analysis of our proposed methodology. In this analysis we used the mean G'_Z at a causal site as measure of power for given values of N, n_s, C, window size (W), SNP density, and for different magnitudes of QTL effect on phenotype. Figure 2 summarizes results for two different values of N, corresponding to large ($N = 1{,}000$) and very large ($N = 10{,}000$) F_2 populations. We find that increasing coverage, C, is advantageous until $C > n_s$, but has minimal effect beyond that. A somewhat counterintuitive result is that larger bulk size, n_s, is generally beneficial as long as sequencing coverage is modest to high. This is despite the fact that larger bulks imply weaker selection for a given N (and hence a smaller allele frequency divergence among bulks). Based on these findings we recommend bulks consisting of at least 10% and as perhaps as high

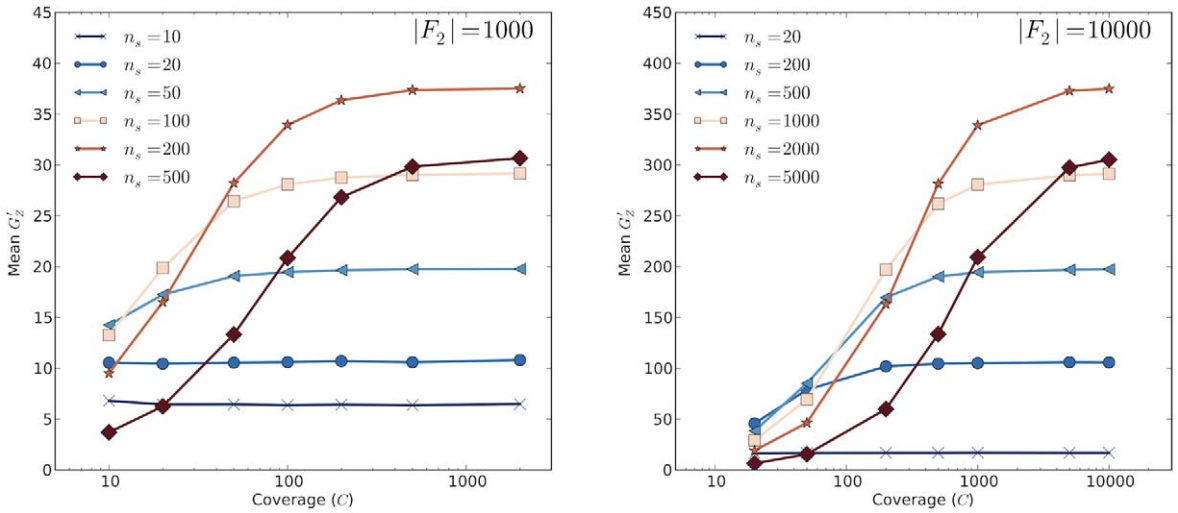

Figure 2. Power analysis. Average G'_Z at a causal site as a function of sequencing coverage, C, and bulk size, n_s, for two different F_2 population sizes (left, $N = 1{,}000$; right, $N = 10{,}000$). Note the difference in scales between the two figures.
doi:10.1371/journal.pcbi.1002255.g002

as 20% of the F_2 segregant population in order to maximize power to detect QTLs.

An Application to Yeast

To demonstrate the correspondence between theory and data we here draw on a BSA-sequencing data set generated to identify loci that contribute to variation in colony morphology in the budding yeast *Saccharomyces cerevisiae* [27]. A full description and analysis of these data will appear elsewhere (Granek et al., in prep). Here, these data serve to illustrate the utility of both our theoretical framework and the associated robust estimators for data analysis.

The yeast data consist of a low and high bulk, each composed of 288 homozygous diploid segregants drawn from an F_2 population of size $N=960$ generated by sporulating a naturally heterozygous diploid strain [28]. The low bulk consists of segregants with simple colony morphology, while the high bulk consists of segregants with complex colony morphology (see [27] for a description of morphology scoring). Creation of DNA pools, sequencing, and mapping of reads is described in the Methods section. Because each segregant is homozygous, the effective number of alleles sampled for each bulk is n_s instead of 2 n_s. In total 44,066 polymorphic sites were analyzed with a mean interval between sites of approximately 280 bp. Below we refer to the two sequencing runs for the low bulks as l_1 and l_2, and those for the high bulks as h_1 and h_2. The coverage per SNP (C) for each sequencing run was as follows: $l_1=48.5$, $l_2=53.8$, $h_1=55.5$, and $h_2=54.2$. For each of the analyses below, we used a smoothing window width of $W=80\,\mathrm{Kb}$ ($\sim$30 cM), and took the average coverage of each bulk being compared as the estimate of coverage, C.

Because there are two sequencing runs per DNA pool, variation in allele frequency estimates between sequencing runs from the same segregant bulk should be exclusively due to stochastic aspects of the sequencing reaction and primary bioinformatics analyses (base calling, read alignment). The structure of this data set is thus useful for dissecting the impact of sequencing variation on estimates of G and G', and the subsequent impact of this variability on the inference of QTL regions and peaks. We use these data to explore both the null model (no QTL; by analyzing the low-vs-low and high-vs-high comparisons) as well as the case where QTLs are expected (comparing low-vs-high bulks). In the null case, the differences in allele frequencies are subject to only one source of variation because the bulks are fixed but sequencing is variable. The non-null analyses are individually affected by both sources of variation (bulking and sequencing), but when comparing the results from comparable analyses (e.g. comparing QTL peak locations between the l_1-vs-h_1 and l_2-vs-h_2 analyses), the differences are again simply a function of sequencing variation.

Null comparisons: Variation in G and G' due to sequencing. The two low samples (l_1 and l_2) and the two high samples (h_1 and h_2) represent independent sequencing runs of the same low and high segregant bulks respectively. Using G and G' from a comparison of l_1 vs. l_2 and h_1 vs. h_2 we can estimate the impact of sequencing on the variation of these statistics. When the two bulks differ only due to read number variation, there is only one source of variation, and the statistics of G should should be approximately χ^2_1 with $E[G]\rightarrow 1$ and $\mathrm{Var}[G]\rightarrow 2$. By invoking a weighted version of the central limit theorem [29], we find the distribution of G' should be approximately normal with $E[G']\rightarrow 1$ and $\mathrm{Var}[G']\rightarrow 2/a_n$ where $a_n=k_1^2+\cdots+k_v^2$, the sum of the v squared kernel weights in the smoothing window (a_n converges to v in the case of a square kernel). As illustrated in Table 2 the observed data for the null-comparisons conform well to the asymptotic expectations.

Between replicate comparisons of G and G' in the presence of a QTL. In addition to tests of the null model, the design of the yeast experiment facilitates a between replicate comparison of G and G' in the presence of QTLs. There are four possible low-vs-high comparisons; here we focus on two of those, l_1-vs-h_1 and l_2-vs-h_2. Figure 3 illustrates the relationships for G and G' at each SNP for l_1-vs-h_1 and l_2-vs-h_2. The between replicate correlation for G is $\sim$0.677, while that between G' is $\sim$0.996. This illustrates the ability of the smoothing kernel to act as a low-pass filter on the G -statistic, filtering out the high-frequency noise associated with variation in read counts, while preserving the underlying signal of QTLs and increasing the repeatability of the analysis.

Using the false discovery rate approach outline above, we estimated cutoff values for G' using a FDR of 0.01 based on both our theoretical results (equations 8 and 12) and the corresponding non-parametric estimators. For the parametric estimate we used the following parameter values: $n_s=144$, $C=52$, $v=200$. The estimated G' cutoff values are as follows: l_1-vs-h_1 : 2.59 [parametric], 3.51 [non-parameteric]; l_2-vs-h_2 : 2.58 [parametric], 3.91 [non-parametric].

Using the theoretical G' cutoff of 2.59 we find 7,845 SNPs have significant G' values for the l_1-vs-h_1 comparison, and 8,011 significant SNPs for the l_2-vs-h_2 comparison, representing approximately 17% of the polymorphic sites. Nearly 38% of the significant sites are on chromosome XIII which appears to have multiple overlapping peaks leading to elevated G' values across much of the chromosome. The number of significant sites shared between the replicates is 7,330. We identified 12 significant regions (QTLs) in the two replicates (Figure 4). The QTLs are nearly identical between the replicates except for a marginal QTL on chromosome 7, where one of the replicates is significant but the other is just short of significance. To assess the variability in QTL location we compared the distance between peaks (using the single largest peak in cases of multiple peaks per chromosome). The mean and median absolute distances between nine comparable QTL peaks from the two comparisons are 5.08 Kb and 4.97 Kb respectively. The root mean square deviation (RMSD) between comparable QTL peaks is 6.7 Kb. Using the RMSD as a measure of spread and applying the 3σ rule of thumb, a conservative confidence interval for QTL peak is $\pm$20 Kb ($\pm$7.4 cM) around

Table 2. Null comparisons for the yeast data set.

Comparison	Theoretical $E[G]$, $\mathrm{Var}[G]$	Observed $E[G]$, $\mathrm{Var}[G]$	Theoretical $E[G']$, $\mathrm{Var}[G']$	Observed $E[G']$, $\mathrm{Var}[G']$
l_1-vs-l_2	1.000, 2.000	1.018, 2.050	1.000, 0.0124	1.020, 0.0115
h_1-vs-h_2	1.000, 2.000	1.015, 2.077	1.000, 0.0124	1.014, 0.0117

Theoretical and observed means and variances of G and G' for the null comparisons in the yeast data set.
doi:10.1371/journal.pcbi.1002255.t002

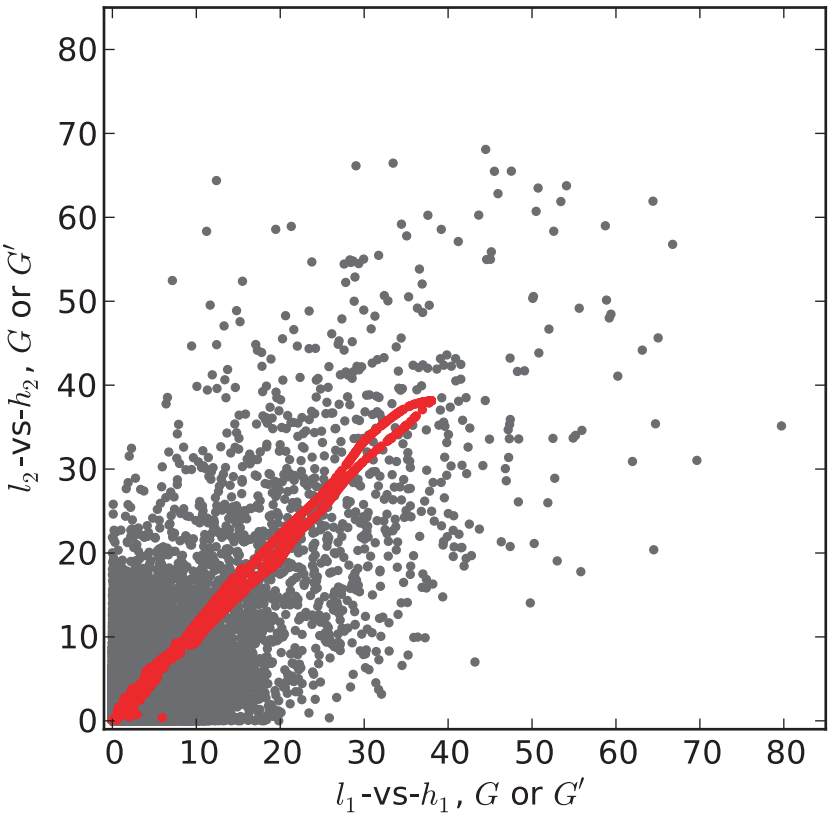

Figure 3. Comparison of G and G' between technical replicates. The correspondence of raw G (black) and smoothed G' values (red) for different sequencing runs of the same low-vs-high bulks from the yeast data set.
doi:10.1371/journal.pcbi.1002255.g003

the observed peak. The size of this confidence interval is a function of read depth and SNP density, and *is a measure of variability in peak estimation due to sequencing only*. This confidence interval doesn't include variation that would arise from the bulking of segregants.

As will be described elsewhere, candidate genes corresponding to several of the major peaks in this analysis have been functionally validated to affect yeast colony morphology (J. Granek and P. Magwene, unpublished data).

Discussion

The use of a test based on the G -statistic provides a straightforward framework for analyzing BSA-sequencing data. The G -statistic has several advantages over the use of allele frequency differences as the basis for QTL estimation (e.g. [11]). For example, as shown in the supporting information (Text S2), G is expected to decrease much more rapidly around the causal site than bias in allele frequencies, implying narrower intervals of support around QTLs. Also in contrast to statistics based on the divergence of allele frequencies, G takes into account the strength of evidence related to sample size. This feature of the G -statistic can also potentially complicate analyses, as variance in read depth contributes to variance in G over relatively small spatial scales. However, as we show above, weighted averaging of G effectively smooths out 'high frequency' noise associated with sequencing variation.

Bulk Size and Sequencing Considerations

Our simulations suggest that for the experimental design considered here using bulk sizes as large as 15–20% of the phenotyped segregant population increases power to detect causal QTLs despite the fact that this means relatively smaller allele frequency differences between bulks. This is due to tradeoffs between bulk-size, selection intensity, and the variance of allele frequencies under the hierarchical sampling. Consider, for example, a single locus with alleles A_0 and A_1, where the effect of A_1 is additive and the two homozygotes differ by $2a$ units on average. Assuming no segregation distortion, and an F_2 population generated from inbred lines, the change in the allele frequency of A_1 in the high bulk after truncation selection is approximately $\Delta q = \frac{1}{8} i \frac{2a}{\sigma_p}$ [30,31] where i is the intensity of selection, and $\frac{2a}{\sigma_p}$ is the 'standardized effect of the locus' (these quantities can be related to the selection coefficient, s, by $s \approx i\frac{2a}{\sigma_p}$). Given truncation selection on a normal distribution, the intensity of selection is given by $i = z/p$ where p is the proportion of selected individuals and z is the probability density function at the truncation point [31]. Since the intensity of selection increases at a rate much less than $1/p$ (e.g. see [31], Fig. 11.3), an n-fold decrease in p results in a much less than n-fold change in the intensity of selection. For example, let $\frac{2a}{\sigma_p} = 0.2$ and consider truncation on the upper 20%, 10%, and 1%, of the phenotypic distribution. The increase in the frequency of A_1 in the high bulk given these truncation points is approximately 3.5%, 4.4%, and 6.7% respectively (translating to allele frequency differences of 7%, 8.8%, and 13.4% in the two-bulk case). On the other hand, the variance of the realized frequencies of the alleles in each bulk is inversely proportional to bulk size ($\mathrm{Var}[p_1^*] = \frac{\rho_1(1-\rho_1)}{2n_s}$). Thus, a twenty-fold decrease in bulk size translates to less than a two-fold increase in allele

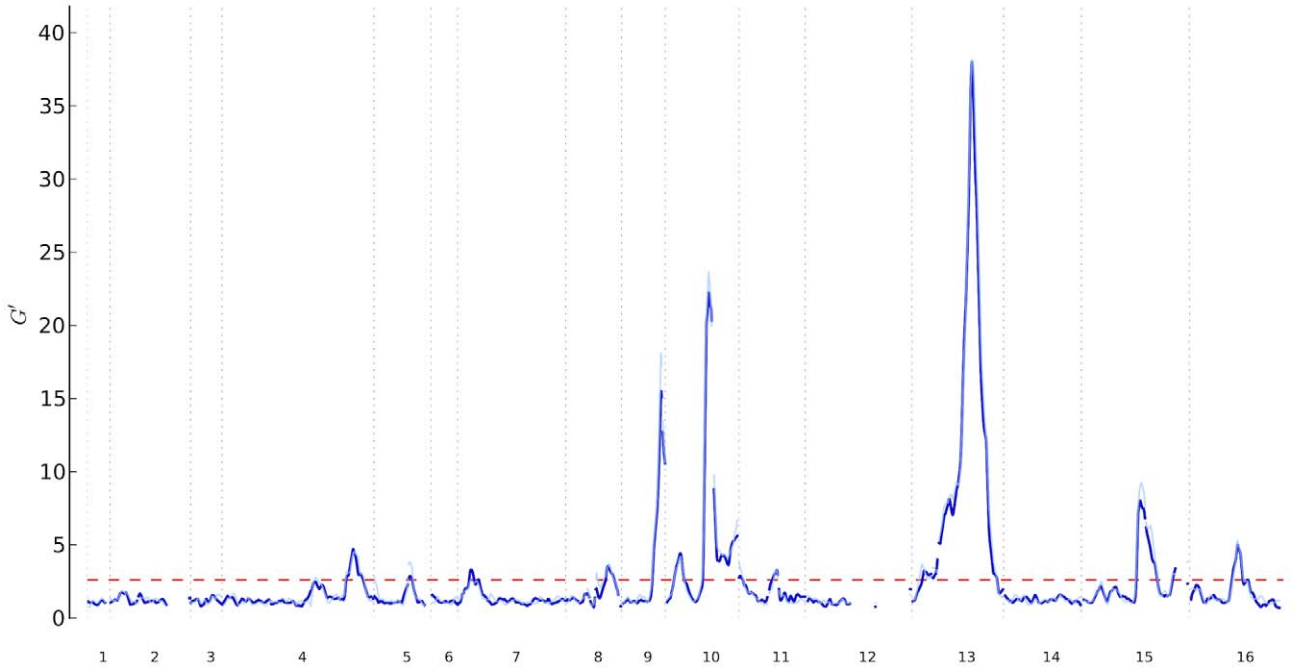

Figure 4. Yeast QTL Peaks. Chromosomal distributions of G' for the l_1-vs-h_1 (dark blue) and l_2-vs-h_2 (light blue) data sets. The dashed red line indicates the estimated G' threshold corresponding to a FDR of 0.01. Regions above the red line are QTL regions; the highest point in each QTL region was called as the QTL peak.
doi:10.1371/journal.pcbi.1002255.g004

frequency divergence, but a twenty-fold increase in the variance of allele frequencies. As long as average coverage, C, is moderate to large, the benefit of increasing n_s offsets the relatively smaller penalty resulting from a decrease in selection intensity. However, there is little benefit to increasing sequencing coverage beyond the size of the bulks.

Sequencing can introduce complications such as biases toward particular nucleotide calls; however in general this should effect both segregant bulks in the same direction. Due to the averaging affect of G', unless such biased sites are common over very large map distances they are unlikely to have substantial affects on results derived under our proposed framework. Similarly, a low percentage of mismapped reads or miscalled SNP calling are unlikely to be problematic for our framework, again because of the averaging affect of G'. However caution should be exercised in genomic regions that are particularly problematic in this regard, such as repeat rich regions.

Other Experimental Designs

In this paper we have focused on QTL mapping with an F_2 experimental design, but clearly our framework can be extended to other designs. Common alternatives include mapping populations produced by imposing one or more generations of inbreeding on an F_2, such as Recombinant Inbred Lines (RILs). The increased homozygosity of such populations should also be taken into consideration, as it increases the expected change in allele frequency due to selection but it also decreases the number of independent chromosomes that are sampled for a given number of selected individuals. Chromosomes in such RILs experience as much as twice the number of crossovers as do F_2 populations so the physical size of the smoothing window W should be reduced to take this reduced linkage disequilibrium into account. Even greater reductions of linkage disequilibrium can be accomplished by an alternative design that imposes additional generations of random mating, rather than inbreeding, on an F_2, resulting in more precise localization of QTLs. Additional generations of outcrossing (beyond the F_2) will likely magnify deviations of the null allele frequency from 0.5 owing to segregation distortion and/or inadvertent selection. This can be accommodated by application of formulas in Text S1 with q estimated from all sites within a genomic window.

Other experimental designs, such as backcrosses, will not have allele frequencies of 0.5. For these situations the null expected distributions of G and G' can be approximated using the equations presented in Text S1, although in this case it will be necessary to know the parental origin of the SNP alleles. Similarly, since G can be generalized to an arbitrary number of classes [12], one-tailed scenarios (e.g. [9]) involving comparison to either a theoeretical population or a random sampling of segregants can be addressed in this framework.

Methods

Sequencing of Yeast Bulks

To create the bulked DNA pools each segregant was grown overnight in liquid medium to saturation ($\sim 10^8$ cells/ml) and equal volumes of each culture were mixed to form cell bulks. Genomic DNA was isolated from the cell bulks and single Illumina DNA sequencing libraries were prepared from each bulk, using standard protocols as described in [28]. Each bulk DNA pool was sequenced twice using 50 bp reads on an Illumina GAII sequencing instrument. Approximately 15 M reads were generated in each sequencing run. Reads were aligned to the yeast reference genome (obtained from the Saccharomyces Genome Database, January 2010) using the program BWA [32] and polymorphic sites were called using SAMtools [33]. For each

sequencing run, SAMtools was used to create a pileup file giving the alleles at each polymorphic site, from which allele counts were derived using scripts written in Python.

Supporting Information

Figure S1 Simulations results for the null distribution of G' based on 10,000 simulations with ($n_s = 150$, $C = 50$, $a = 0$). The gray histogram represents the observed distribution of G', corresponding to Figure 1b. The dashed lines represent log-normal distributions estimated from theoretical expectation (red line) or via the non-parametric approach described in the text (black line). Both the parametric and non-parametric approaches provide good control of type I error (right tail of the distribution).
(PDF)

Text S1 Generalization of theoretical results to include segregation distortion.
(PDF)

Text S2 Miscellaneous information. This file includes information on: 1) estimation of the parameters of a log-normal distribution from the expected mean and variance of a variable of interest; 2) the expected shape of the G around at a QTL; and 3) A summary table of expected and observed means and variances of G' based on simulations of the null hypothesis (no QTL).
(PDF)

Acknowledgments

We thank Joshua Granek and Debra Murray who helped to generate the yeast BSA data set. We thank Stuart McDonald for conversations and feedback. We thank the Duke University Institute for Genome Sciences & Policy Sequencing Facility for the sequencing of genomic libraries.

Author Contributions

Conceived and designed the experiments: PMM JHW JKK. Performed the experiments: PMM JKK. Analyzed the data: PMM JKK. Contributed reagents/materials/analysis tools: PMM JKK. Wrote the paper: PMM JHW JKK.

References

1. Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A 88: 9828–9832.
2. Ehrenreich IM, Gerke JP, Kruglyak L (2009) Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the byxrm cross. Cold Spring Harb Symp Quant Biol 74: 145–153.
3. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, et al. (1998) Direct allelic variation scanning of the yeast genome. Science 281: 1194–1197.
4. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, et al. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13: 513–523.
5. Brauer MJ, Christianson CM, Pai DA, Dunham MJ (2006) Mapping novel traits by array-assisted bulk segregant analysis in saccharomyces cerevisiae. Genetics 173: 1813–1816.
6. Segr AV, Murray AW, Leu JY (2006) High-resolution mutation mapping reveals parallel experimental evolution in yeast. PLoS Biol 4: e256.
7. Boer VM, Amini S, Botstein D (2008) Inuence of genotype and nutrition on survival and metabolism of starving yeast. Proc Natl Acad Sci U S A 105: 6930–6935.
8. Demogines A, Smith E, Kruglyak L, Alani E (2008) Identification and dissection of a complex dna repair sensitivity phenotype in baker's yeast. PLoS Genet 4: e1000123.
9. Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, et al. (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature 464: 1039–1042.
10. Wenger JW, Schwartz K, Sherlock G (2010) Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from saccharomyces cerevisiae. PLoS Genet 6: e1000942.
11. Parts L, Cubillos FA, Warringer J, Jain K, Salinas F, et al. (2011) Revealing the genetic structure of a trait by sequencing a population under selection. Genome Res 21: 1131–1138.
12. Sokal RR, Rohlf FJ (1994) Biometry W. H. Freeman.
13. Nadaraya EA (1964) On estimating regression. Theor Probab Appl 9: 141–142.
14. Watson GS (1964) Smooth regression analysis. Sankhaya 26: 175–184.
15. Schucany WR (2004) Kernel smoothers: An overview of curve estimators for the first graduate course in nonparametric statistics. Statist Sci 19: 663–675.
16. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Amer Stat Assoc 74: 829–826.
17. Mohn E (1979) Confidence estimation of measures of location in the log normal distribution. Biometrika 66: 567–575.
18. Davies L, Gather U (1993) The identification of multiple outliers. J Amer Stat Assoc 88: 782–792.
19. Bickel DR, Frühwirth R (2006) On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. Comput Stat Data An 50: 3500–3530.
20. Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. J Amer Stat Assoc 88: 1273–1283.
21. Turin GL (1960) An introduction to matched filters. IEEE Trans Inform Theory 6: 311–329.
22. Enke CG, Nieman TA (1976) Signal-to-noise ratio enhancement by least-squares polynomial smoothing. Anal Chem 48: 705–712A.
23. Gu H, Gao R (1997) Resolution of overlapping echoes and constrained matched filter. IEEE Trans Signal Proc 45: 1854–1857.
24. Mikl M, Marecek R, Hlustk P, Pavlicov M, Drastich A, et al. (2008) Effects of spatial smoothing on fmri group inferences. Magn Reson Imaging 26: 490–503.
25. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Sci, B 57: 289–300.
26. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188.
27. Granek JA, Magwene PM (2010) Environmental and genetic determinants of colony morphology in yeast. PLoS Genet 6: e1000823.
28. Magwene PM, Ömür Kayıkçı, Granek JA, Reininga JM, Scholl Z, et al. (2011) Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in saccharomyces cerevisiae. Proc Natl Acad Sci USA 108: 1987–1992.
29. Weber M (2006) A weighted central limit theorem. Stat Probabil Lett 76: 1482–1487.
30. Kimura M, Crow JF (1978) Effect of overall phenotypic selection on genetic change at individual loci. Proc Natl Acad Sci U S A 75: 6168–6171.
31. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edition Longman.
32. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754–1760.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The sequence alignment/map format and samtools. Bioinformatics 25: 2078–2079.