A Regression-based Approach for Estimating Recombination Rate from Population Genomic Data

Lan Zhu [a], Feng Feng [b], Carlos D. Bustamante [c]

[a]Department of Statistics, Oklahoma State University, Stillwater, OK 74078
[b]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705
[c]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

ABSTRACT

Motivation: Recently, much attention has focused on using prediction from population genetic theory to quantify variation in recombination rate along the human genome owing to the promise of association or linkage disequilibrium(LD) mapping to identify genes underlying complex traits. Current state of the art approaches to the problem estimate the local population recombination rate from patterns of LD among common single nucleotide polymorphisms(SNPs) assuming the population is randomly mating and constant in size.

Results: Here we describe an alternative method that can accommodate complex population structure and ascertainment bias. Using multiple linear regression and non-parametric bootstrap re-sampling, our method uses the variances and co-variances of un-phased SNPs at different frequencies to estimate the local recombination rate. We evaluate this new approach via Monte Carlo simulation and compare its performance with three other available methods. Our approach is less biased when the demographic assumptions of the standard neutral model are violated. We also apply our approach to the well-characterized hot spots near the human TAP2 gene and a 206-kb region on human chromosome 1q42.3 near minisatellite MS32. The results are consistent with findings in literatures.

Keywords: Recombination, Regression, Linkage Disequilibrium

Contact: lan.zhu@okstate.edu

1 INTRODUCTION

Understanding how and why recombination rates vary along a genome is a fundamental problem in genomics. From an evolutionary perspective, recombination is a rich source of novel variation and a potent force that can lead to gametic associations among positively selected mutations as well as break up associations among deleterious mutations. Recombination rates also vary dramatically among genomes with some, such as Drosophila, showing no clear fine-scale structure while others, such as humans, showing a great deal of local variation where regions of low to moderate recombination are punctuated by short 1-2 kb hotspots of meiotic exchange that can account for $50-80\%$ of all recombination events (McVean *et al.*, 2004; Myers *et al.*, 2005).

When the contributions of recombination and its interaction with selection to the process of evolution are shown essential (Cutter and Choi, 2010; Cutter and Moses, 2011), understanding recombination rate variation is also fundamentally important to the design of efficient methods for association mapping, since the degree of association among markers dictates the density and distribution of markers used for mapping (Noor *et. al.*, 2001). Classical methods for estimating recombination rates from natural populations include pedigree studies, sperm typing analysis and methods based on predictions from population genetics. In humans, the difficulty of obtaining large pedigrees limits the utility of pedigrees to estimation of large-scale (megabase) recombination rates (Kong *et al.*, 2002). Likewise, while sperm typing can provide accurate estimates of the local recombination rate in male gamete production, it is typically only applied to a few individuals and to only short regions of the genome (Greenawalt *et al.*, 2006). Moreover, it is very labor intensive and expensive. These limitations coupled with the increasing availability of genome-wide polymorphism data from humans and other species make estimation of recombination rates via population genetic theory an attractive alternative.

A number of estimators of the population recombination rate ($R = 4N_e r$, where r is the rate of crossing over for the region and N_e is the effective population size) are currently available, including moment-based (Hey and Wakeley, 1997; Hudson, 1985; Hudson, 1987; Wall, 2000), full maximum likelihood estimators (Fearnhead and Donnelly, 2001; Griffiths and Marjoram, 1996; Kuhner, *et al.*, 2000; Nielsen, 2000), and full-likelihood Markov chain Monte Carlo method (Wang and Rannala, 2008). From a statistical perspective, one would prefer to use full-likelihood methods, since these are guaranteed to capture the most amount of information in the data regarding recombination. However since it can take months of computer time to estimate recombination rate for even a modest size region using full-likelihood, there has been considerable effort to develop a litany of approximate likelihood estimators (Crawford *et al.*, 2004; Fearnhead and Donnelly, 2002; Fearnhead, *et al.*, 2004; Fearnhead and Smith, 2005; Haubold, *et. al.*, 2010; Hudson, 2001; Jiang *et. al.*, 2009; Li and Stephens, 2003; McVean, *et al.*, 2002; McVean, *et al.*, 2004). For example, the composite likelihood methods of Hudson (2001) and McVean *et al.* (2004) use pre-computation of pairwise likelihood for a given sample size to achieve speeds orders of magnitude faster than full-likelihood. Auton and McVean (2007) further constructed a pseudo-likelihood as the product of the likelihood over all pairs of SNPs in the region under consideration. To maintain the computational feasibility, SNPs separated by no more than 50 intermediate SNPs were considered to contribute to the composite likelihood.

These approaches, while quite fast, have several limitations including the need to precompute pairwise likelihood for a novel sample size or demographic model and an apparent lack of power to detect recombination hotspots that do not significantly affect linkage disequilibrium (Jeffreys, *et al.*, 2005) . The methods of Fearnhead *et al.* (2004), Li and Stephens (2003), and Fearnhead and Smith (2005) appear to have excellent power to detect hotspots, but are computationally costly (e.g., according to Fearnhead and Smith (2005) it takes their method 10-30 minutes to estimate the recombination rate for a window of six SNPs with sample

size 60 sequences). It is also important to note that the effective population size is confounded within the estimate of the population recombination rate, therefore, population genetic estimators are by definition dependent on assumptions regarding the demographic history of the sample. A limitation of many of these approaches, therefore, is that they are based on the assumption that the population under study is randomly mating and constant in size - an assumption violated by nearly all populations to which the approaches are applied. In theory, population structure and demography can be built into almost any method, but for methods such as composite likelihood that make use of a great deal of pre-computation, this will require months (or years) of computer time for each new model to generate the lookup tables used in estimation.

In this paper, we present a novel statistical method for estimating the population recombination rate via coalescent simulations with recombination coupled with multiple linear regression (MLR) and non-parametric bootstrap. Three advantage of our method are that (1) it can readily accommodate complex demographic history, (2) provide prediction intervals for the estimated recombination rate, and (3) is computationally efficient and applicable to whole-genome data. Furthermore, since the method appears to weight heavily the variance of new mutations in estimating recombination rates, it may be able to detect recent changes in recombination rate that do not leave an explicit LD signal.

Our method is based on a readily discernible statistic of the data: the observed variability in the number of mutations at different frequencies across sub-samples of the data. It is important to note that the idea of using the variance of mutation counts in a sample to estimate recombination rates is not new. About two decades ago, Hudson (1987) introduced an estimator of the population recombination rate based on the variance of pairwise nucleotide differences among sequences in the sample. In 1997, Wakeley proposed an improved version of Hudsons (1987) estimator that has smaller bias and standard error. Our approach is loosely a generalization of Hudsons estimator in that we aim to use the most informative components of the frequency distribution to estimate the local recombination rate. A major advantage of this approach is that it does not require calculation of pair-wise linkage disequilibrium and, thus, does not require phasing of the data. Likewise, while our approach requires some pre-computation to fit the model, it is orders of magnitude less than existing approaches (roughly minutes to hours for our approach compared to days or weeks for composite likelihood). We investigate the accuracy of the approach using Monte Carlo simulations under a wide range of demographic models. We also compare the performance of our method to three commonly used approaches (Hey and Wakeley, 1997; Hudson, 1987; McVean, *et al.*, 2002).

2 METHODS

2.1 Data and Model

Consider a set of n aligned DNA sequences from a population with known demography Q (e.g., population of constant size, bottleneck, island migration, recent population growth, etc.) in which S sites are observed to be variable in the alignment. Let X_i for $i = 1 \ldots n - 1$ represent the number of SNPs at frequency i out of n in the sample. For simplicity, here the ancestral state of each SNP is assumed known (i.e., the polarized site-frequency spectrum); a model with unknown ancestral state can be easily derived in the similar way. Across independent realizations of the evolutionary process, X will vary stochastically so that for each component one has an associated variance V_i. For example, V_1 is the variance in the number of singletons that one would observe if one were to have rerun the evolutionary process and obtained an independent sampling of chromosomes at the same locus. Here we describe how recombination affects the variances and co-variances of the components of the SFS (SFS variances) in a fully predictable way and how by estimating SFS variances, one can predict the recombination rate of a genomic region for a given demographic model. For a given observed data set, however, one only has a single observed vector of frequencies, so we must first define what we mean by variance within components of the site-frequency spectrum.

Here, we consider the variance in X_i under two scenarios: (1) independent realizations of the evolutionary process (i.e., a variance that one can estimate only via simulation) and (2) bootstrap resampling of the sequences (i.e., a variance one can readily estimate via a common statistical readily applicable to the observed data). As we show in the results section, these two scenarios give different, but nearly perfectly correlated variances such that one may estimate the former given an observed value from the later.

First, let us assume that one was able to rerun the evolutionary process Q under the same recombination rate R so as to obtain Q replicate data sets, sampling an independent set of n sequence each time. From population genetic theory we expect variance and co-variance of the X_is across the Q replicates to be informative about recombination (Fu, 1995; Sawyer and Hartl, 1992; Zhu and Bustamante, 2005).

For example, for a population that evolves according to the standard neutral Wright-Fisher model, Fu (1995) derived that variance and co-variances of the X_is as a function of the population mutate rate $\theta = 4N_e\mu$ under complete linkage. Specifically, under complete linkage one can write the variance of X_i as V_i=Var(X_i)=θ/i+$\sigma_{ii}\theta^2$, where σ_{ii} is a function of i and sample size n. Under complete independence among sites, Ewens (1972) and Sawyer and Hartl (1992) showed that X_i should be Poisson distributed with mean and variance $V_i = \theta/i$. Given these two well-known results, one might posit a monotonic decrease in the variance of V_i with increasing R so that recombination acts simply to decrease the σ_{ii} term above. (These predictions are born out in Figures 1 and 2 as explained below.) The reasoning above immediately suggests a simple and potentially powerful strategy for estimating R.

2.2 Algorithms for Estimating Recombination Rate

2.2.1 Algorithm 1: estimating recombination rate across evolutionary replicates

1. Simulate data by Hudson's ms program (Hudson, 2002) under the demographic model for Q replicates keeping the matrix of site-frequency spectra (SFS) with the Q rows representing the site-frequency spectra for independent replicates (simulations can be carried out conditional on the estimated mutation rate, θ, or on the observed number of segregating sites, S):

$$\begin{vmatrix} x_{1,1} & x_{2,1} & \ldots & x_{n-1,1} \\ x_{1,2} & x_{2,2} & \ldots & x_{n-1,2} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{1,Q} & x_{2,Q} & \ldots & x_{n-1,Q} \end{vmatrix}$$

2. For each pair of columns i and k, calculate the column means $\overline{X}_i$, column variances V_i, and co-variance V_{ik} across replicates (note $V_{ii} = V_i$ in our notation above). The results of this step will constitute an $n-1$ dimensional vector of SFS means $\overline{X} = [\overline{X}_1, \overline{X}_2, ..., \overline{X}_{n-1}]$, where $\overline{X}_i = \sum_{j=1}^{Q} \frac{X_{i,j}}{Q}$ and a variance-covariance matrix with entries: $V_{ik} = \sum_{j=1}^{Q} \frac{(X_{i,j} - \overline{X}_i)(X_{k,j} - \overline{X}_k)}{Q}$.
3. Repeat above steps across a range of recombination rates (in practice we use $R \in \{1, 5, 10, 20, 50, 100, 200, 400, 1000, 2000\}$) so as to produce a set of predictor variables in the form of the $(n-1)$ variance

and $\binom{n-1}{2}$ covariance entries of the variance-covariance matrices across levels of R.

4. Natural log-transform both the predictor (V_{ik} for different levels of R) and predicted variables (R).
5. Use stepwise selection or best subset methods to choose the model that is sufficient to explain the relationship among $log(R)$ and the log of the components of the variance-covariance matrix. Formally, the full model would have $\binom{n-1}{2} + (n-1) + 1$ terms of the form:
$log(R_j) = \alpha + \sum_{i=1}^{n-1} \sum_{k=1}^{i} \beta_{ik} log(V_{ik,j}) + e_j$
where $e_j \sim N(0, \sigma^2)$. In the model above, α is the intercept of the regression, β_{ik} are regression coefficients under the saturated model, and e_j are independent and identically distributed error terms for the residual variance for j = 1 ... J where J is the number of levels of recombination used to fit the model. In practice, we use stepwise selection and best subset methods to search over the space of models so as to identify the subset of β_{ik} terms that are sufficient to explain the data.
6. Check all assumptions for fitting a linear regression model, including normality, equal variance of residuals, and independence among residuals.

2.2.2 Algorithm 2: estimating recombination rate by bootstrap-based regression(BSTReg) across k-subset replicates A potential problem of applying the above method to real data is that for a given data set, one only has a single observed site-frequency spectrum, X. In order to generate estimates of the variance/covariance matrix across replicates of the evolutionary process we need to use a resampling scheme such as non-parametric bootstrap resampling of the data. Since the estimated variances under the bootstrapping procedure use correlated data, we expect estimates of V_{ik} to be affected. Therefore, we need to modify our MLR fitting procedure as follows:

1. Sample a single data set with n sequences under a demographic model of interest Θ, and label the data set q.
2. Divide the n sequences into k subsets of equal size, calculate the SFS for each subset, then modify the above step so that the mean and variance-covariance matrix are now calculated across the k site-frequency spectra.
3. Repeat this k-subset division sampling for the same n sequences for B bootstrap replicates to obtain B variance-covariance matrices.
4. Let $V_i^{(q)} = \frac{1}{B} \sum_{k=1}^{B} V_{i,k}$ be the average variance of component i and $Cov_{ij}^{(q)} = \frac{1}{B} \sum_{k=1}^{B} Cov_{ijk}$ the average covariance across the B replicates of the subsetting approach. (In practice, we use $n = 60$ and $k = 10$. If the data is unphased, resample individuals; if the data is phased, resample phased haplotypes.

Repeat steps 1-4 for Q replicate data sets to obtain the bootstrap estimated variance-covariance matrix V_{bs} for a given model Θ, where $V_{i\_bs} = \frac{1}{Q} \sum_{q=1}^{Q} V_i^{(q)}$ and $Cov_{ij\_bs} = \frac{1}{Q} \sum_{q=1}^{Q} Cov_{ij}^{(q)}$.

3 RESULTS

3.1 Estimating recombination rate when θ is known or S is fixed under evolutionary replication

We first consider the problem of predicting the population recombination rate from polymorphism data arising under a known demographic model. Using standard coalescent algorithms, we simulated $10,000$ replicate samples for each of 10 levels of recombination rate $R \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000\}$ under a fixed mutation rate $\theta = 4N_e\mu = 30$ where μ is the regional mutation rate per chromosome. (These parameter values correspond roughly to a $30kb$ region in humans with recombination rate varying from 2.5×10^{-4} cM to $0.25cM$.) or a fixed number of segregating sites $S = \{10, 20, 30, 50, 100\}$. For a given level of recombination, we calculate the vector of SFS variances $V = \{V_1, V_2, ..., V_{n-1}\}$ across the Q replicate data sets as explained in the method description above.

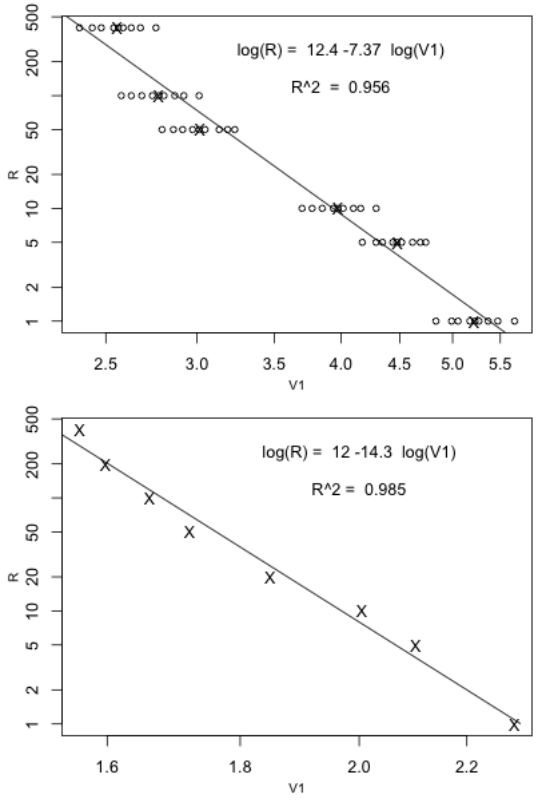

Fig. 1. Linear regression of log transformed recombination rate ($logR$) and log transformed variance in the number of singletons in the sample. Top: 200 replicates of data sets each with sample size $n = 6$, $S = 10$ were simulated independently under the standard neutral Wright-Fisher model. Each points represents the V_1 quantiles $\{0.025, 0.10, 0.20, 0.40, 0.5, 0.60, 0.80, 0.90, 0.975\}$ corresponding to R in the range of $\{1, 5, 10, 50, 100, 400\}$. Cross signs are the means of $logV_1$ over 200 replicates; Bottom: Linear model is fitted by $logR$ on the average of $logV_1\_bs$ by k-subset bootstrap resampling over 1000 replicates.

When we perform the multiple linear regression of R on all V_{ik}s including all pairwise covariances among SFS components and use both stepwise selection and best subset methods, all terms are dropped except for the variance of singletons (V_1) in the model. Scatter plot of R versus the V_1 across simulated data sets shows a curvilinear relationship suggesting that linear regression of log-transformed data could be used to estimate R from a linear combination of the components in V. Using a step-wise addition rule, we find that $log(V_1)$ alone is a sufficient predicator variable for the population recombination rate with the best fit linear regression explaining over 95% of the variance in either fixed θ or S scenarios, as shown in Figure 1 (top) when $S = 10$. Diagnostic tests (linearity, constant variance, normality, independence) for validation of the model were performed and none of the tests suggests a violation of the assumptions. (Note: for all regressions performed, diagnostic tests were checked and no violation is found,

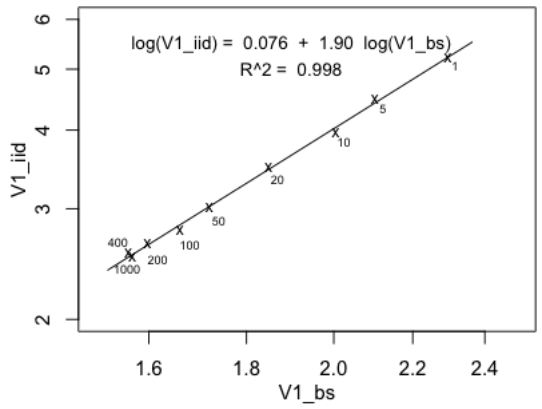

Fig. 2. Relationship between the average of bootstrap estimated variance in the number of singletons (V_1_bs) and that from independent sampling (V_1_iid). Sample size $n = 6$, $S = 10$ under the standard neutral Wright-Fisher model.

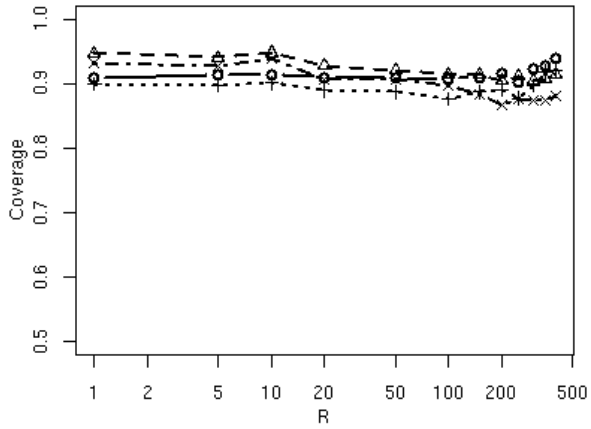

Fig. 3. Coverage of predicted local recombination rate using our bootstrap-based linear regression method with sample size $n = 60$, $S = \{10, 20\}$, $k = 10$. X-axis is plotted in log scale. Linear regression model is fitted in the range of $R = \{1, 5, 10, 20, 50, 100, 200, 1000, 2000\}$ with equation $log(R) = 13.644 - 16.612 * log(V_1)$ for $S = 10$ ($R^2 = 0.933$) and $log(R) = 17.544 - 9.517 * log(V_1)$ for $S = 20$ ($R^2 = 0.959$). Coverage is defined as the percentage of replicates that have 90% or 95% predicted intervals cover true recombination rate.

results not shown). This simple example shows that for a fixed level of the mutation rate or a fixed number of segregating sites, the transformed recombination rate and the first component of SFS variances are highly correlated. By choosing a fixed number of segregating sites in a genomic region, one can reliably predict the recombination rate for the region using the observed SFS variances across samples.

3.2 Estimating recombination using bootstrap re-sampling and k-subseting (BSTReg)

For a real data set, however, one only has a single observed SFS vector. To estimate the SFS variances, one therefore needs to couple a re-sampling step such as non-parametric bootstrapping

Table 1. Multiple linear regression output for estimating $log(R)$ on $log(V_1\_bs)$ for $Q = 1,000$ replicate simulated data set, each with $n = 60$, $S = 10$, $R \in \{1, 5, 10, 20, 50, 100, 200, 400\}$. Each V_1_bs was estimated by K-subset non-parametric bootstrap sampling as described in the method session. Here $K = 10$.

$log(R) = 11.9959 - 14.2855 * Log(v_1)$	
RSquare	0.9846
RSquare Adj	0.9820

Parameter Estimates

Term	Estimate	Std. Error	t value	Prob > |t|
Intercept	11.9959	0.45210	26.53	1.89e-07
$log(v_1)$	-14.2855	0.7294	-19.59	1.15e-06

to the MLR procedure. K-subset bootstrap sampling as described in the method session results in a predictive relationship between $log(R)$ and the average of $log(V_{1\_bs})$ over 1000 replicates as shown in Figure 1 (bottom). The output of the regression is shown in Table 1. In simulations we have also found that non-parametric bootstrap estimates of variances are systematically smaller than the evolutionary variance since the bootstrap procedure only considers variability across samples with the same population history instead of the evolutionary variance across random populations; however, there is a clear linear relationship between these two variance on a log-log scale. Figure 2 shows the near-perfect linear correlation between the average $log(V_{1\_iid})$ and average $log(V_{1\_bs})$ as indicated by the cross-signs. This provides us the flexibility of using i.i.d. samples to estimate the relationship between $log(R)$ and the SFS variances for a given demographic model. We can then estimate $log(V_{1\_bs})$ from $log(V_{1\_iid})$ greatly speeding up the computation.

3.3 Comparing BSTReg to existing methods

Figure 3 shows the coverage (the percentage of replicates that have 90% or 95% prediction intervals cover true recombination rate) of predicted local recombination rate using our bootstrap based linear regression model with sample size $n = 60$, segregating sites $S = \{10, 20\}$ under the standard neutral Wright-Fisher model. The method performs well with coverage close to or greater than 90% at all level of R in the range of 1 to 400. Moreover, the coverage increases with the number of segregating sites where more information is included in the data. Mean square errors (MSEs) in figure 4 are low and uniform by our new method compared with Hudson's (1987) approach. LDhat results in the lowest MSE. Due to the limit of the maximum R that can be estimated by LDhat software, $R > 100$ are not explored here. We did not include Hey-Wakeley's (1997) performance in MSE comparison because for data with $S = 10$, Hey-Wakeley's (1997) failed to output the estimates and for samples with $S = 20$, the estimators were quite under estimated. This can be seen in figure 5 where we report the ratio of the median estimates over the true parameters for four methods. We can see that Hey-Wakeley's (1997) estimator is uniformly downwardly biased for all levels of the recombination rate in the range of $R \in \{1, 5, 10, 20, 50, 100, 200, 400\}$ while Hudson's (1987) is upwardly biased for $R \leq 50$ and performs better for larger R. Our new BSTReg approach performs almost equally

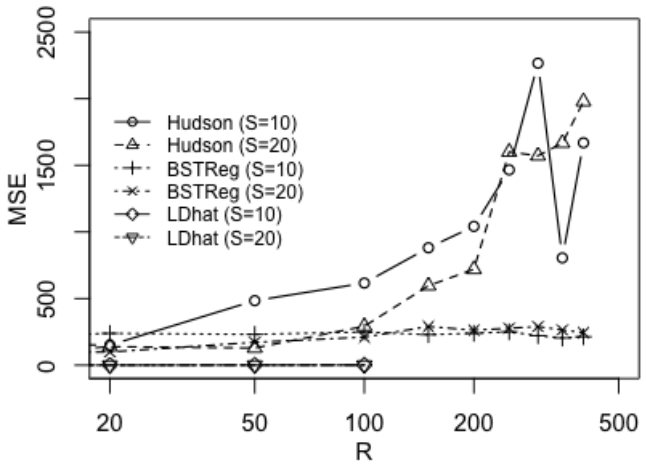

Fig. 4. Comparison of mean square errors (MSE) of predicted local recombination rate over 1000 replicates using Hudson's (1987) method, LDhat (McVean 2004) and our bootstrap based linear regression method. Sample size $n = 60$, segregating sites $S = \{10, 20\}$, $k = 10$. X-axis is plotted in log scale. Linear models are the same as used in the coverage evaluation. Due to the limit of the maximum R that can be estimated by LDhat software, $R > 100$ are not explored.

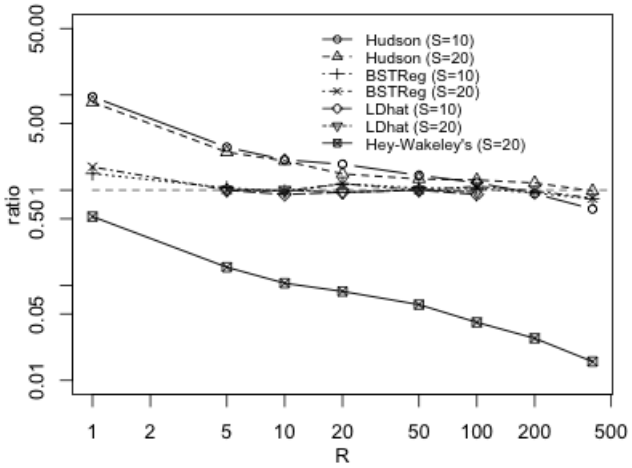

Fig. 5. Performance comparison of Hudson's (1987), Hey-Wakeley's(1997), LDhat (McVean 2004) and our bootstrap-based linear regression method in terms of the ratio of the median of predicted local recombination rates over 1000 replicates to the true recombination rate. Sample size $n = 60$, segregating sites $S = \{10, 20\}$, $k = 10$. Both axes are plotted in log scale. (For $S = 10$, Hey-Wakeley's (1997) method fails to work due to not enough informative segregating sites, results are not included in the figure; results of $R > 100$ from LDhat are not explored as well).

well as LDhat when the population size is constant and without structure: the ratios are around 1 for all levels of R.

One question that arises is: how sensitive are other approaches to the demographic assumptions of the standard neutral model? In figure 6, we report the ratio of median estimates to the true paramter by our BSTReg method and LDhat across a range of recombination rates for $1,000$ simulated data sets under two island migration ($4N_em = 12$) and population growth ($rate = 5.0$). We note that our approach has less bias, presumably since it can incorporate the demographic details explicitly in the estimating equations.

3.4 Application to the TAP2 and MS32 recombination hotspots

We have also used our approach to estimate fine-scale recombination rate variation around two recombination hotspots in the human genome characterized through sperm typing (haplotype sequences were kindly provided by Professor Sir Alec J. Jeffreys). For the TAP2 gene region, a total of 60 sequences with 48 SNPs were included in the analysis. According to Jeffreys *et al.* (2000), 81% of the sperm crossover breakpoints in the data were localize to the $1.4kb$ region between markers T15 and T30 (depicted as grey box from position $4,017$ to $5,417$ in Figure 7). We estimated the recombination rate between adjacent pairs of SNPs (as well as associated prediction intervals) using a sliding window approach with 10 SNPs in each window as described in the Methods section. Figure 7 shows the mean and lower bound of the 95% prediction interval of the recombination rate along the TAP2 genomic region before the SNP ascertainment bias correction. As we see from Figure 7, the hot spots regions identified by our approach are completely consistent with the results from both sperm typing and haplotype analysis (Jeffreys, *et al.*, 2000). That is we detect a strong signal of dramatically active recombinational exchange in the regions between markers $T16(4180)$ and $T18(4553)$, $T23(4917)$ and $T24(4934)$, and $T27(5188)$ and $T30(5417)$. After the ascertainment bias correction, the same hot spots regions are identified (result not shown); but without correcting the ascertainment bias will results in more conservative estimation. In this case, the ascertainment bias increased the variance of singletons about 1.55 fold.

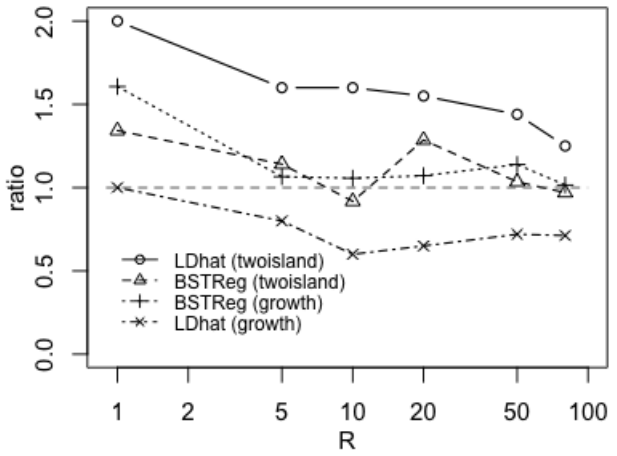

Fig. 6. Median estimates over the true recombination rate ratio over 1000 replicates by LDhat (McVean 2004) and our bootstrap-based linear regression methods under twoisland migration model ($4N_em = 12$) and population exponentially growing model (growth rate $G = 5.0$). Sample size $n = 60$, segregating sites $S = 10$. X-axis is in log-scale.

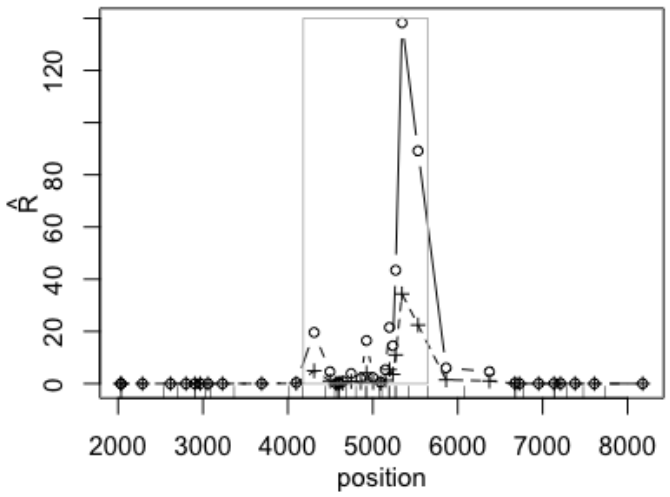

Fig. 7. The mean and lower bound of 95% prediction interval of recombination rate along the TAP2 region. The regression model in Table 1 is used for the prediction. SNPs marker positions are consistent with those in Jeffreys *et al.* (2000). Region in the grey box is the location where sperm crossover breakpoints were highly clustered (Jeffreys *et al.*, 2000).

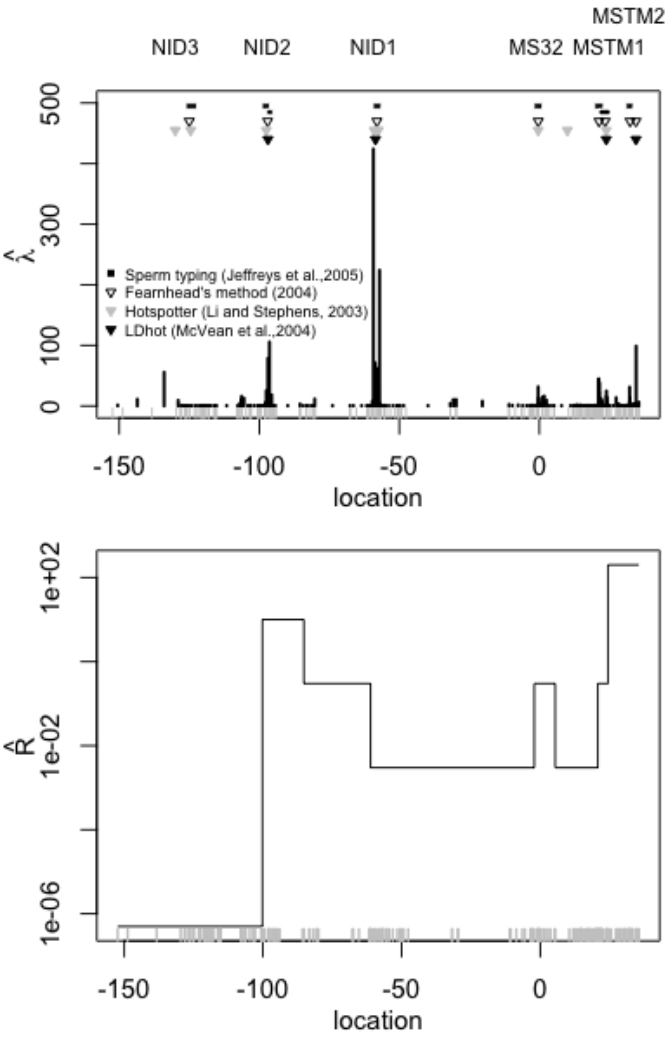

Fig. 8. Top: Ratio of recombination rate estimates to the background values in the 206 kb interval surrounding minisatellite $MS32$ on chromosome $1q42.3$. Putative hotspots identified by sperm typing (Jeffreys *et al.*, 2005), Fearnheads method (2004), Hotspotter (Li and Stephens, 2003), and LDhot (McVean *et al.*, 2004), respectively, are also shown as reported in Jeffreys *et al.* (2005), Figure 1b. Bottom: Estimated background recombination rate along the region. Data from: http://www.le.ac.uk/ge/ajj/MS32/MS32%20genotypes%20file.html.

We have also applied this approach to a 206 kb region on human chromosome $1q42.3$ which contains several well-characterized autosomal crossover hotspots around the highly variable minisatellite $MS32$ (Jeffreys, *et al.*, 1998). Due to the complexity of the SNPs identification in this data set, we only estimated the recombination rate without correcting the ascertainment bias. For this analysis, 80 individuals with 214 SNPs were included (we again use a $w = 10$ SNP window). Figure 8(top) shows the mean ratio of predicted recombination rate to the estimated background rate (the estimated background rates along the region which are the average rates of the local predicted rates exclude the putative hotspot regions are shown in figure 8 bottom) as well as the location of predicted hotspots by several approaches as reported in figure 1b of Jeffreys *et al.* (Jeffreys, *et al.*, 2005). The black rectangles in our figure 8(top) show the location of recombination hotspots as estimated by sperm typing (figure 1b, Jeffreys et al., 2005). As demonstrated in Jeffreys *et al.* (2005), the approximate likelihood method of Fearnhead *et al.* (2004) (white triangles) and the PAC likelihood method of Li and Stephens (2003) (grey triangles) do an excellent job of identifying the location of the hotspots in the region as evidence by the strong concordance with hotspots estimated from sperm typing. Both of these approaches are very computationally intensive and require hours to run on the data set, and are thus not currently viable options for genome-wide estimation of recombination rate variation. We note that our approach (which takes about 70 seconds by a Power Mac G5 with $2.5GHz$ CPU speed and 4GB memory to run on the same region) shows clear signatures of recombination rate variation near the six putative hotspots ($NID1$, $NID2$, and $NID3$ in and near the NID gene, as well as $MS32$, $MSTM1$ and $MSTM2$).

4 DISCUSSION

We proposed a new bootstrap-based linear regression approach to estimate the population recombination rate. While the algorithm we have presented is fast, flexible, and scalable to the whole genome level, a few caveats must be raised. In order to make inference, we must still presuppose some demographic model for the data. Our preliminary results confirm the predictions of population genetic theory that recombination rate estimates will be sensitive to the demographic model used in the MLR fitting step. This sensitivity is not likely unique to our approach and probably holds for the majority of algorithms currently in use. At the same time, it also appears that our approach is robust to demography for the problem

of detecting recombination rate variation. Secondly, our method can currently only deal with uniform ascertainment schemes. When ascertainment differs dramatically among SNPs in the same region, however, this may likely cause problems for any method aiming to discover variation in recombination rate.

It is important to note that the choice of the window size on the regression region may affect rate estimation. Windows significantly overlap when we move one SNP site step by step. If the window size is too large, rate estimates are upwardly or downwardly affected by adjacent SNPs, especially when the window ranges from no or low recombination rate region to a hot spots region. From our experience with this model, we suggest that a window size between 10 to 20 SNPs appears to be an optimal trade-off between signal of recombination rate variation and noise due to stochastic variation of individual SNPs.

Lastly, we have assumed (as all other methods) that the SNPs in our sample are evolving neutrally. Since natural selection is known to affect both the patterns of linkage disequilibrium as well as the site-frequency spectrum in a region, our method is likely sensitive to this assumption. For example, a region that has experienced a recent selective sweep is expected to have low levels of nucleotide variation as well as a skew towards rare alleles. If the variance of singletons in the region is also reduced, then one may overestimate the recombination rate. One possible way to distinguish these two factors is to test explicitly for evidence of a selective sweep in the region (which is expected to leave a characteristic spatial pattern of reduced variation around the target of selection). For regions that show strong evidence of a sweep other approaches such as direct sperm typing may be necessary for accurate estimation of recombination rate variation.

ACKNOWLEDGEMENT

We thank Charles Aquadro, Andrew Clark, Martin Wells, and Scott Willimson for advice and help on earlier drafts of this paper. This work was funded by the National Science Foundation grant 0516310 to CDB.

REFERENCES

[1]Auton, A. and McVean, G. (2007) Recombination rate estimation in the presence of hotspots. *Genome Res.*, **17**, 1219-1227.

[2]Crawford, D.C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A. and Stephens, M. (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.*, **36**, 700-706.

[3]Cutter, AD., Choi, J.Y. (2010) Natural selection shapes nucleotide polymorphism across the genome of the nematode Caenorhabditis briggsae. *Genome Res*, **20**, 11031111.

[4]Cutter, A.D. and Moses, A.M. (2011) Polymorphism, divergence, and the role of recombination in saccharomyces cerevisiae genome evolution. *Mol. Biol. Evol.*, **28**, 1745-1754.

[5]Ewens, W. (1972) The sampling Theory of selectively natural alleles, *Theor. Pop. Biol.*, **3**, 87-112.

[6]Fearnhead, P. and Donnelly, P. (2001) Estimating recombination rates from population genetic data, *Genetics*, **159**, 1299-1318.

[7]Fearnhead, P. and Donnelly, P. (2002) Approximate likelihood methods for estimating local recombination rates, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **64**, 657-680.

[8]Fearnhead, P., Harding, R.M., Schneider, J.A., Myers, S. and Donnelly, P. (2004) Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots, *Genetics*, **167**, 2067-2081.

[9]Fearnhead, P. and Smith, N.G.C. (2005) A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes, *Am. J. Hum. Genet.*, **77**, 781-794.

[10]Fu, Y.X. (1995) Statistical properties of segregating sites, *Theor. Pop. Biol.*, **48**, 172-197.

[11]Greenawalt, D.M., Cui, X., Wu, Y., Lin, Y., Wang, H.Y., Luo, M., Tereshchenko, I.V., Hu, G., Li, J.Y., Chu, Y., et al. (2006) Strong correlation between meiotic crossovers and haplotype structure in a 2.5-Mb region on the long arm of chromosome 21. *Genome Res.*, **16**, 208-214.

[12]Griffiths, R.C. and Marjoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination, *J. Comput. Biol.*, **3**, 479-502.

[13]Hey, J. and Wakeley, J. (1997) A coalescent estimator of the population recombination rate, *Genetics*, **145**, 833-846.

[14]Haubold, B., Pfaffelhuber, P., and Lynch, M. (2010) mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology.*, **19**, 277-284.

[15]Hudson, R.R. (1985) The sampling distribution of linkage disequilibrium under an infinite allele model without selection, *Genetics*, **109**, 611-631.

[16]Hudson, R.R. (1987) Estimating the recombination parameter of a finite population model without selection, *Genet. Res.*, **50**, 245-250.

[17]Hudson, R.R. (2001) Two-locus sampling distributions and their application, *Genetics*, **159**, 1805-1817.

[18]Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation, *Bioinformatics*, **18**, 337-338.

[19]Jeffreys, A.J., Murray, J. and Neumann, R. (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot, *Mol. Cel.*, **2**, 267-273.

[20]Jeffreys, A.J., Neumann, R., Panayi, M., Myers, S. and Donnelly, P. (2005) Human recombination hot spots hidden in regions of strong marker association, *Nat. Genet.*, **37**, 601-606.

[21]Jeffreys, A.J., Ritchie, A. and Neumann, R. (2000) High-resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot, *Hum. Mol. Genet.*, **9**, 725-733.

[22]Jiang, R., Tavare, S., and Majoram, P. (2009) Population genetic inference from resequencing data. *Genetics*, **181**, 187-197.

[23]Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R. and Stefansson, K. (2002) A high-resolution recombination map of the human genome, *Nat. Genet.*, **31**, 241-247.

[24]Kuhner, M.K., Beerli, P., Yamato, J. and Felsenstein, J. (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters, *Genetics*, **156**, 439-447.

[25]Li, N. and Stephens, M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, *Genetics*, **165**, 2213-2233.

[26]McVean, G., Awadalla, P. and Fearnhead, P. (2002) A coalescent-based method for detecting and estimating recombination from gene sequences, *Genetics*, **160**, 1231-1241.

[27]McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004) The fine-scale structure of recombination rate variation in the human genome, *Science*, **304**, 581-584.

[28]Noor, M.A.F., Cunningham, A.L., and Larkin, J.C. (2001) Consequences of recombination rate variation on quantitative trait locus mapping studies: simulations based on the drosophila melanogaster genome. *Genetics*, **159**, 581-588.

[29]Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005) A fine-scale map of recombination rates and hotspots across the human genome, *Science*, **310**, 321-324.

[30]Nielsen, R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms, *Genetics*, **154**, 931-942.

[31]Sawyer, S.A. and Hartl, D.L. (1992) Population genetics of polymorphism and divergence, *Genetics*, **132**, 1161-1176.

[32]Wall, J.D. (2000) A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.*, **17**, 156-163.

[33]Wang, Y. and Rannala, B. (2008) Bayesian inference of fine-scale recombination rates using population genomic data. *Phil. Trans. R. Soc. B*, **363**, 3921-3930.

[34]Zhu, L. and Bustamante, C.D. (2005) A composite-likelihood approach for directing selection from DNA sequence data, *Genetics*, **170**, 1411-1421.