BMC
Genetics

## RESEARCH ARTICLE

**Open Access**

# Reconstructing past changes in locus-specific recombination rates

Murray P Cox[1]*, Barbara R Holland[2], Matthew C Wilkins[1] and Jan Schmid[1]

## Abstract

**Background:** Recombination rates vary at the level of the species, population and individual. Now recognized as a transient feature of the genome, recombination rates at a given locus can change markedly over time. Existing inferential methods, predominantly based on linkage disequilibrium patterns, return a long-term average estimate of past recombination rates. Such estimates can be misleading, but no analytical framework to infer recombination rates that have changed over time is currently available.

**Results:** We apply coalescent modeling in conjunction with a suite of summary statistics to show that the recombination history of a locus can be reconstructed from a time series of genetic samples. More usefully, we describe a new method, based on *n*-tuple dataset subsampling, to infer past changes in recombination rate from DNA sequences taken at a single time point. This subsampling strategy can correctly assign simulated loci to constant, increasing and decreasing recombination models with an accuracy of 84%.

**Conclusions:** While providing an important stepping-stone to determining past recombination rates, *n*-tuple subsampling still exhibits a moderate error rate. Theoretical limitations indicated by coalescent theory suggest that highly accurate inference of past recombination rates will remain challenging. Nevertheless, we show for the first time that reconstructing historic recombination rates is possible in principle.

**Keywords:** Recombination rate, Temporal, Reconstruction, Statistical inference

## Background

Meiotic recombination, whereby DNA variants are shuffled between homologous parental chromosomes, is a fundamental process in the evolution of genetic diversity. For many years poorly studied, the mechanisms and effects of recombination are now increasingly well understood [1]. We know that recombination rates are both heritable [2] and variable among individuals [3-7]. In other words, recombination is a Darwinian evolutionary system [8,9].

Recent studies have demonstrated convincingly that recombination rates at a given locus vary at the level of the species, population and individual. Comparisons between the chimpanzee and human genomes show poor correlation of both hotspot and background recombination rates at orthologous loci [10,11]. Similarly, recombination rates vary between human populations [11], not

only at continental scales, but also between close geographical neighbors (e.g., French and Italians) [11]. Recombination rates even vary widely between individuals drawn from the same population [3-7]. The picture now emerging is one of an extremely dynamic recombination landscape [10], with transient recombination peaks and troughs across the human genome, overlaying the better-known evolutionary variation in DNA substitution rates [12].

How changes in recombination rate are controlled is less well understood [13]. At some loci, recombination events are determined by nucleotide variation within specific DNA sequence motifs (such as the degenerate 13-bp pattern recognized by *PRDM9*) [14,15]. These *cis*-mediated recombination events often show evidence of transmission distortion [16], where biased gene conversion preferentially favors one allelic variant that can rapidly reach fixation [17]. Related mechanisms may also act as a selective force to reduce recombination around functional genomic elements [18]. However, recombination rates at most loci seem to be mediated by *trans*

* Correspondence: m.p.cox@massey.ac.nz
[1]Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North 4442, New Zealand
Full list of author information is available at the end of the article

factors [5], typically controlled by genes that coordinate DNA-protein interactions [19,20], or more generally, by regional chromatin remodeling [1,7]. These studies suggest that *trans*-mediated recombination processes dominate genome-scale recombination events and are not obviously under the influence of natural selection.

Recombination is typically detected either directly by gamete typing, or indirectly from linkage disequilibrium (LD) patterns [1]. Gamete typing surveys large numbers of recombination events within a single generation, and therefore provides an accurate (albeit costly) estimate of *contemporary* recombination rates. Conversely, statistical analysis of linkage disequilibrium patterns counts recombination events that have accrued over multiple generations, and therefore returns a long-term average estimate of *historic* recombination rates. Since recombination rates can change through time, contemporary and historic rate estimates need not agree. Regions of high recombination, as predicted from linkage disequilibrium, may be inactive when surveyed using gamete typing [21], a discrepancy that indicates the extinction of a previously high recombination region. Conversely, gamete typing may reveal regions of high recombination where none are suggested by linkage disequilibrium, thus indicating the birth of new high recombination loci [22].

The main point is that recombination rates at a genomic location can vary substantially through time. Although this fact is now widely appreciated [16], the manner in which recombination rates increase or decrease still remains completely unknown. Do changes in recombination rate occur rapidly, perhaps due to point mutations suddenly altering the action of the recombination machinery? Or are changes more gradual, occurring as regional nucleotide diversity mutates slowly over time? We currently lack any analytical framework to address these sorts of questions. Here, we determine that a suite of summary statistics can track changes in recombination rate through time. We extract temporal information about changing recombination rates, and describe some of the theoretical limitations that constrain this endeavor. More importantly, we develop a novel methodology based on *n*-tuple subsampling that has sufficient statistical power to reconstruct the recombination history of a genetic locus studied at a single time point. This approach is intended as a proof-of-concept that past changes in recombination rates can be reconstructed from contemporary data, even if reconstructing historic rates from empirical data remains challenging.

## Results
### Correlation and sensitivity of summary statistics
We first explored how different summary statistics respond to recombination events. The number of segregating sites $S$ was used as a negative control because

mean $S$ does not vary with the recombination rate. We also assembled a suite of eight summary statistics that were designed specifically to detect recombination events – $R_{min}$, $rmmg$, the number of haplotypes, haplotype diversity, Wall's $B$ and $Q$, Hudson's $C$ and $Z_{nS}$. These summaries likely recognize different aspects of recombination, although the relationships between them have not been explored. Certainly none of these summaries capture the entire recombination profile of a genetic sample (i.e., they are not statistically sufficient).

We studied the correlation matrix between summary statistics using an equal mix of datasets with linearly increasing, decreasing and constant recombination rates. *A priori*, we might expect that many of the summaries detect similar aspects of the overall recombination signal. Indeed, pairwise comparisons indicated that nearly all the summaries were correlated, albeit to different extents ($r$ values range from 0 to 0.982, mean of 0.351) (Figure 1). The smallest correlations involved $rmmg$, a conservative lower bound on the minimum number of recombination events $R_{min}$, which showed little variation among datasets under the conditions modeled here. None of the summary statistics were perfectly correlated, thus emphasizing that multiple summaries are needed to capture different aspects of the recombination profile.

To determine how these summaries respond to different recombination rates, we simulated genetic data under a wide range of constant recombination values (0–10 $\rho$/kb) (Figure 2). $S$ is shown as a negative control because its mean is invariant to the recombination rate (Figure 2, upper right). Most summary statistics varied nonlinearly across this linear range of recombination values. It follows that the usefulness of any individual summary may change with the underlying recombination rate, but in different ways. Therefore, a combination of some or all of these summaries may be more sensitive for detecting different recombination rates than any one of them alone.

### Tracking changing recombination rates using time series data
It is less obvious how summary statistics might covary with recombination rates that change over time. To explore this process, we generated coalescent simulations where recombination rates were allowed to vary over many generations. Genetic datasets were simulated using coalescent software [23] modified to allow recombination rates to change through time. We simulated data for a human-like deme: $10^4$ replicates of 10-kb autosomal sequences were drawn from a constant sized population ($N_e = 10^4$) [24-26] with a mutation rate, $\mu$, and average recombination rate, $r$, of $3.75 \times 10^{-8}$ events/ bp/generation [27]. These rates were chosen to mimic
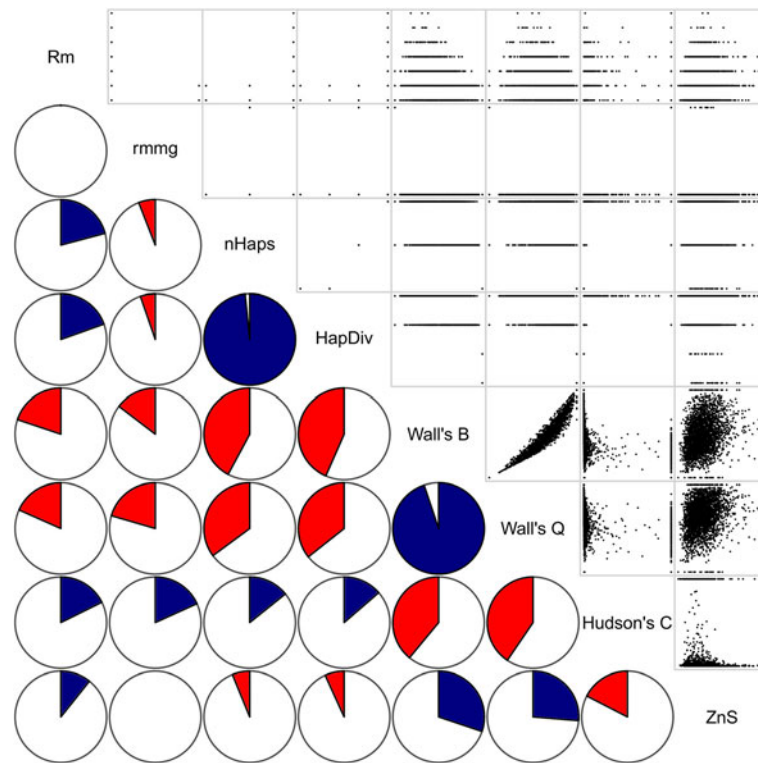
**Figure 1 Correlations between recombination summary statistics.** (Upper diagonal) Scatter plots show pairwise relationships among the summary statistics. (Lower diagonal) Pie charts show the magnitude of the correlation with blue and red indicating positive and negative values (e.g., Pearson's $r = -0.20$ for Hudson's $C$ and $Z_{nS}$). All non-zero correlations are statistically significant ($P \leq 0.05$).

regions of very strong recombination in real human populations [11]. The recombination rate was either held constant, or allowed to vary linearly, exponentially or logistically through time for $10^4$ generations (cf. [24,26]). The total amount of recombination was constrained to be identical for all models; only its distribution through time was altered.

A representative example illustrating a logistic decline in recombination rates towards the present is presented in Figure 3. Corresponding plots for constant recombination, together with recombination rates increasing and decreasing linearly, exponentially and logistically, are presented in Additional file 1: Figure S1, Additional file 2: Figure S2, Additional file 3: Figure S3, Additional file 4: Figure S4, Additional file 5: Figure S5, Additional file 6: Figure S6 and Additional file 7: Figure S7.

Most of the summary statistics tracked the changing recombination profile, albeit with notable differences in accuracy. The variance of many summaries altered with the recombination rate, thus suggesting that different summaries have greatest power to estimate recombination rates at different times. This reinforces the view that using a combination of summary statistics should maximize statistical power, although a simple linear combination may not necessarily be optimal.

Note too that summary values typically lagged changes in the recombination rate. Genetic variation observed in the present was actually laid down in the (sometimes very distant) past [28]. As recombination rates change, it takes time before this change is reflected in the genetic record. This lag effect is perhaps best illustrated in the plot showing recombination rates increasing exponentially into the past (Additional file 5: Figure S5). Although the recombination rate drops quickly, the summaries change far more slowly. Even after the recombination rate falls near zero, existing lineages still retain the signal of recombination events that occurred further back in the past. Only as these recombined lineages are lost through genetic drift is the new low recombination rate finally reflected in the summaries. This time lag places important constraints on the resolution with which recombination rates that have changed through time can be reconstructed.

### Reconstructing past recombination rates from data taken at a single time point

Tracking variable recombination rates using time series data may be feasible for some fast evolving systems (e.g., exploring the loss of sexual competency in yeast), but it is not practical for long-lived organisms like humans. To explore whether past recombination rates can be
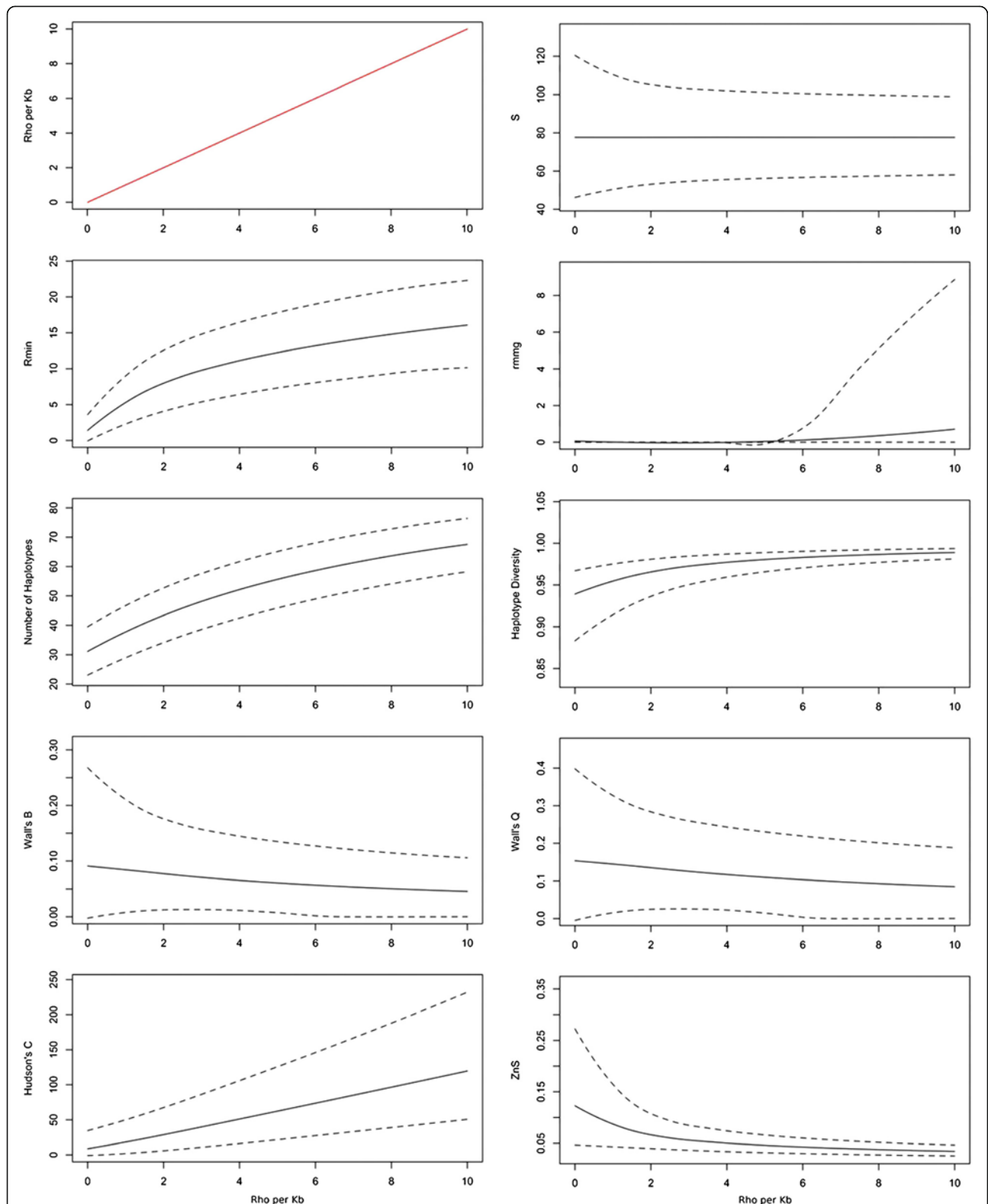
**Figure 2 Sensitivity of summary statistics to different constant recombination rates.** Black lines show the mean (solid) and 95% confidence intervals (dotted) of summary statistic values. The red line indicates different constant recombination rates (ρ per kb). Note that summary statistics mostly vary nonlinearly with linear change in recombination rates.
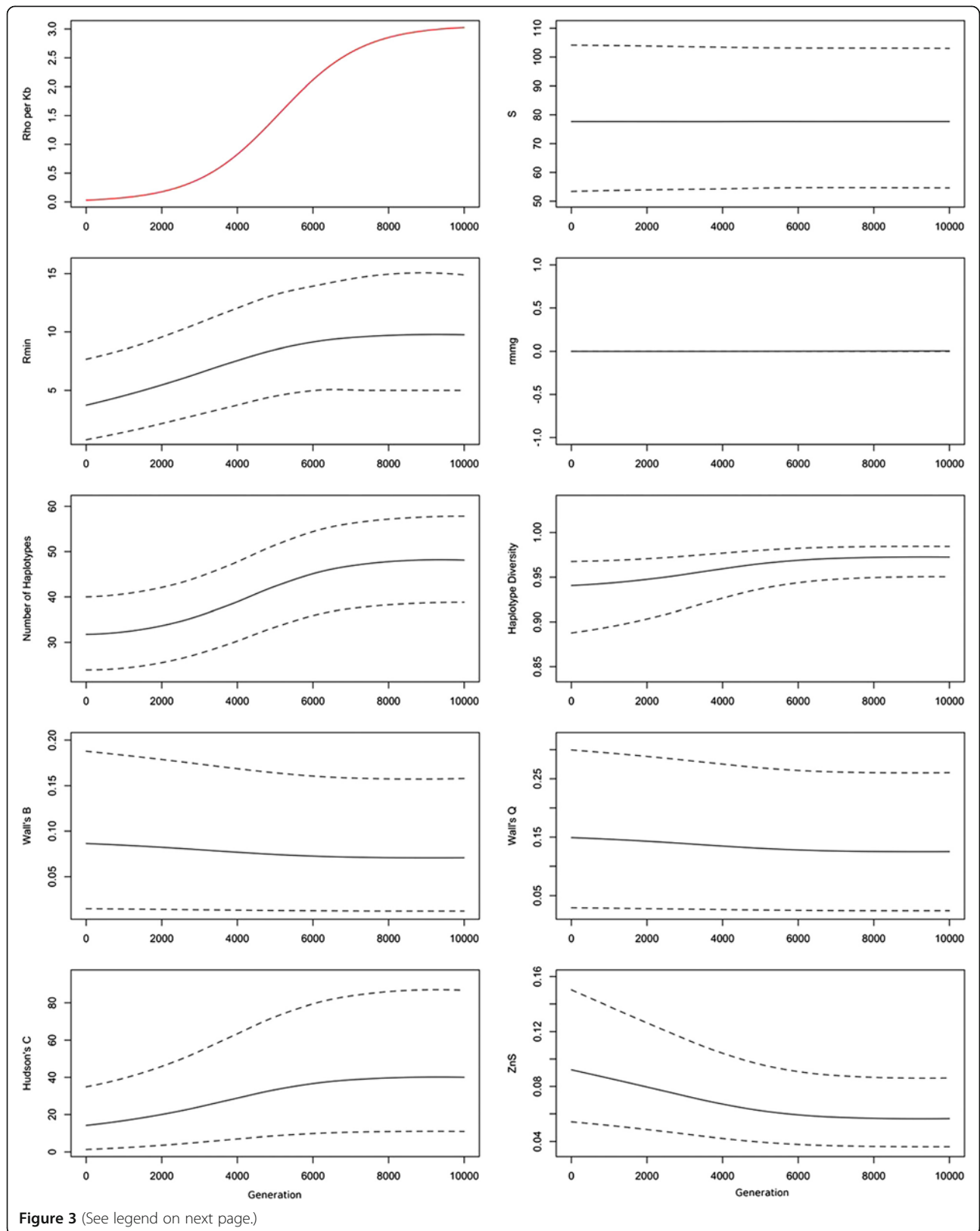
**Figure 3** (See legend on next page.)

(See figure on previous page.)
**Figure 3 Response of summary statistics to recombination rates changing logistically over time.** Black lines show mean (solid) and 95% confidence intervals (dotted) of summary statistic values. The red line indicates how the recombination rate changes over time ($\rho$ per kb). Note the lag between changes in the recombination rate and changes in summary statistic values. Compare responses for constant recombination, as well as linearly, exponentially and logistically increasing and decreasing recombination rates, in Additional file 1: Figure S1, Additional file 2: Figure S2, Additional file 3: Figure S3, Additional file 4: Figure S4, Additional file 5: Figure S5, Additional file 6: Figure S6 and Additional file 7: Figure S7.

reconstructed from genetic data taken at a single time point, we developed a novel bootstrapping methodology that we call *n*-tuple subsampling.

Mutations occur randomly through time. In any given dataset, some polymorphisms will be old and most modern lineages will carry them. Others will be young, and will therefore be found in only one or two individuals. By determining whether recombination events affect young or old polymorphisms, we can theoretically obtain snapshots of recombination rates through time.

This concept is best shown graphically (Figure 4). Imagine we repeatedly subsample a group of four lineages (i.e., an *n*-tuple of four, or a 'quartet') from a given dataset. These individuals may be closely related (Figure 4A), in which case they contain information about recombination rates in the recent past. Alternately, the individuals may be only distantly related (Figure 4C), in which case they may carry both young and old recombination events. By repeatedly resampling the dataset, the recombination rate at different times can be inferred. We emphasize that *n*-tuple subsampling has a natural confound. Young *n*-tuples carry information about recent recombination rates, but old quartets contain a mix of information about old and recent recombination events. The statistical power of this approach is therefore unclear and we explore this issue in detail below.

The use of resampling methods, such as the bootstrap and jackknife [29], to estimate sampling distributions is widespread in statistics in general, but especially so in evolutionary biology [30]. More recently, interest has grown in so-called ensemble methods (also referred to as 'bagging and boosting') that seek to improve overall

accuracy by combining the predictions of many weak classifiers, each of which is shown a slightly perturbed version of the data [31]. The following implementation of *n*-tuple subsampling differs from the ensemble method as we use only a single classifier. However, it is a related concept in that the classifier is shown many subsets of the data, which in our case is used to develop combinations of summary statistics that capture information about recombination rates over different time periods.

To ascertain whether *n*-tuple subsampling has sufficient power to estimate past recombination rates, we created a test system where datasets derived from only one of three recombination models: recombination that is constant, linearly increasing or linearly decreasing through time ($10^4$ datasets each) (Additional file 1: Figure S1, Additional file 2: Figure S2 and Additional file 3: Figure S3). $10^3$ quartets were generated for each dataset, and the suite of summary statistics was calculated for each subsample. The mean, variance and maximum of these summary statistic distributions were recorded.

Although powerful Bayesian and maximum likelihood methods have been developed to perform inference on such datasets [32,33], these approaches are extremely computationally expensive (i.e., tens to thousands of CPU hours to analyze a single dataset) [25,26,34,35]. As we needed to run power analyses for thousands of test cases, a more pragmatic analytical framework was necessarily adopted (cf. [36]). We instead employed discriminant analysis [37], a routine statistical technique for data classification, with jackknife cross-validation to evaluate the accuracy of the classifier. Discriminant analysis infers the combination of weighted summaries (i.e., the optimal transform) that best distinguishes the recombination models. In jackknife cross-validation, model labels were removed, and each blinded dataset was instead assigned to a recombination model using the discriminant function. Assignment accuracy was calculated by determining the proportion of datasets that were assigned to the correct recombination model (i.e., datasets derived under a constant recombination model should be assigned back to the constant model). Both linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) were tested. These related methods respond differently to unequal covariance among models, as well as different sizes of sample and training sets [38]. LDA proved to return more accurate classifications in this instance.
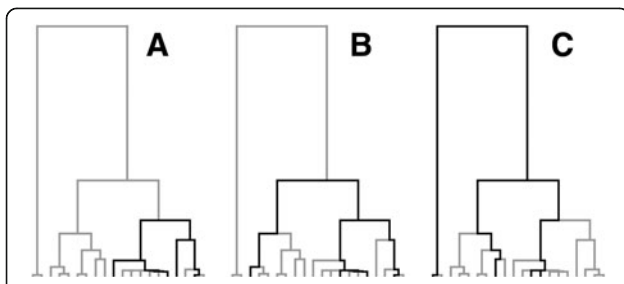


**Figure 4 Variable ages of quartets.** Randomly selected quartets (black lines) capture information about (**A**) young, (**B**) medium and (**C**) old time depths. For visual clarity, quartets are shown on a non-recombining genealogy, but the principle holds equally for ancestral recombination graphs.

LDA was performed on all datasets from all three recombination models. Each dataset was sequentially excluded, the optimal transform inferred by LDA was applied, and each dataset was reassigned back to a recombination model. As we have three models, assignment rates of one-third are expected just by chance. Assignment rates approaching one indicate increasingly accurate assignments.

Table 1 shows observed assignment rates. Because the mean of $S$ is invariant to recombination (Figure 2, upper right), assignments using $S$ alone are presented as a negative control. As expected, the mean, variance or maximum number of segregating sites performed no better than chance. The best individual summary, the mean number of haplotypes, was much more accurate (60%). The best result was obtained by combining the mean, variance and maximum values of all summaries (68%), although 32% of datasets were still placed incorrectly.

### Scaling subsamples by *n*-tuple Age

These assignments were obtained using information about the amount of recombination in each *n*-tuple, but not its age. When recombination rates change over time, the amount of recombination and the age of each *n*-tuple should be correlated. We would therefore prefer to use summaries that capture information about both factors simultaneously. To develop such summaries, we scaled the recombination summaries by $S$, which is a robust proxy for *n*-tuple age (Additional file 8). Pairwise correlations indicated that most scaled summaries are positively correlated ($r$ values from 0.055 to 1, mean of 0.436) (Additional file 9: Figure S9). The cross-validation test was repeated, and surprisingly, the scaled summaries often performed more poorly than their unscaled versions. Nevertheless, using the mean, variance and maximum values of both the scaled and unscaled summaries returned the best overall result (71% correct assignment). This suggests two key conclusions. First, genetic datasets do record retrievable information about past changes in recombination rate. Second, scaled and unscaled recombination summaries do capture very slightly different information from the recombination profile.

Coalescent theory tells us that the power to detect recombination events should decline exponentially into the past (see details later). Therefore, the linearly increasing and decreasing models are mostly dominated by low and high recombination rates, respectively, while the constant model is intermediate. We were concerned that our cross-validation test might simply be detecting low, medium and high recombination rates rather than distinguishing constant recombination from recombination rates that change through time. We therefore repeated the cross-validation test with four recombination models: constant high, constant low, linearly increasing and linearly decreasing recombination rates. Assignment accuracy was only slightly lower than for the three-model test (64% vs 71%). We conclude that *n*-tuple subsampling can distinguish changing recombination from constant recombination, as well as rates that increase or decrease through time.

### Effect of *n*-tuple size on classification accuracy

Thus far, *n*-tuple subsampling has been performed using just four sequences (a quartet). Quartets have found many uses in phylogenetics because this is the minimum number required for unrooted trees to possess distinguishable topologies. However, quartets may not be optimal for reconstructing past recombination rates within a population. We therefore varied the subsample size from 4 to the total sample size, $4 \le n \le 100$ (Figure 5). The three-model system was used, and the optimal LDA transform was recalculated for each value of $n$.

As before, assignment rates started at 71% for quartets, initially improved with increasing $n$, before declining again as the subsample size approached the total

**Table 1 Assignment accuracy using linear discriminant analysis on quartets**

|  | Mean | Variance | Maximum | Combined |
|---|---|---|---|---|
| **$S$** | 0.36 | 0.32 | 0.36 | |
| **$R_{min}$** | 0.43 | 0.45 | 0.38 | |
| **rmmg** | 0.43 | 0.43 | 0.45 | |
| **nHaps** | 0.60 | 0.59 | 0.33 | |
| **HapDiv** | 0.59 | 0.57 | 0.33 | |
| **Wall's $B$** | 0.41 | 0.47 | 0.34 | |
| **Wall's $Q$** | 0.40 | 0.37 | 0.33 | |
| **Hudson's $C$** | 0.43 | 0.42 | 0.33 | |
| **$Z_{nS}$** | 0.48 | 0.34 | 0.48 | |
| **All unscaled** | 0.66 | 0.65 | 0.52 | 0.68 |
| **$S$ x $R_{min}$** | 0.36 | 0.36 | 0.34 | |
| **$S$ x rmmg** | 0.43 | 0.41 | 0.45 | |
| **$S$ x nHaps** | 0.32 | 0.36 | 0.36 | |
| **$S$ x HapDiv** | 0.32 | 0.36 | 0.36 | |
| **$S$ x Wall's $B$** | 0.37 | 0.33 | 0.31 | |
| **$S$ x Wall's $Q$** | 0.37 | 0.34 | 0.32 | |
| **$S$ x Hudson's $C$** | 0.42 | 0.37 | 0.37 | |
| **$S$ x $Z_{nS}$** | 0.42 | 0.34 | 0.43 | |
| **All scaled** | 0.64 | 0.59 | 0.53 | 0.67 |
| **All combined** | | | | 0.71 |

Proportions of datasets assigned correctly to constant, linearly increasing and linearly decreasing recombination models using jackknife cross-validation. Values of one-third indicate assignments no better than chance; values approaching one indicate improving assignment rates. $10^4$ datasets consisting of 10-kb of sequence for 100 individuals were generated under each model. Assignments were made using the mean, variance and maximum value of summary statistics for $10^3$ quartets for each dataset.

sample size. The best assignment accuracy (84%) was obtained with an *n*-tuple of size 77. While subsampling a large proportion of the dataset seemed to be most informative under this model system, optimal subsample size appears to vary from dataset to dataset in practice. When recombination rates are low, recombination events are recorded by few lineages and subsampling a greater proportion of the dataset improves detection of these rare occurrences. However, we can only detect recombination rates from their effects on DNA polymorphisms, so recombination events will often pass undetected if nucleotide diversity is low. Both low recombination rates and low genetic diversity therefore favor larger subsamples, while smaller subsamples are preferable for highly diverse or recombination-rich regions. Apart from these general guidelines, it seems that optimal subsample sizes must be determined empirically for each dataset.

Assignment accuracy was maximized at 84% across all analyses performed here. Although considerably better than chance, the error rate is still moderate. Because power levels are relatively modest, reconstructing historic



**Figure 5 Effect of subsample (*n*-tuple) size on assignment accuracy.** Subsample sizes range from 4 to the sample size (*n* = 100). The curve shows a local regression through the data points.

recombination rates for real genomic loci is expected to remain difficult even when *n*-tuple subsampling is employed. The highly constrained testing environment used here (e.g., a simple and perfectly known demography) emphasizes this point. In practice, complex demographic processes can produce patterns of genetic variation that might otherwise be attributed to processes of recombination [39]. Still, *n*-tuple subsampling is directly amenable to statistical methods that infer model likelihoods by simulating data across a parameter space (e.g., approximate Bayesian computation [32,33]). Like *n*-tuple subsampling, these methods typically employ a suite of summary statistics, and because they are based on Monte Carlo simulation, they can readily be modified to accommodation the novel bootstrapping process that we propose.

## Discussion

We show that information about past changes in recombination rate can be extracted from genomic data using a suite of summary statistics coupled with lineage subsampling to provide proxy information about recombination events at different time depths. Simulated datasets can be correctly assigned to different models of historic recombination with high accuracy (84%).

Why is the power of *n*-tuple subsampling not greater? Coalescent theory suggests several possible reasons. The coalescent describes how pairs of lineages sequentially share a common parent and merge ("coalesce") until only one ancestral lineage remains. This process is analogous to genetic drift, where lineages are lost by chance over time. The key point is that individuals existing today are represented by fewer and fewer ancestral lineages moving backwards into the past (Figure 6).

Put more formally, the coalescent times $T_i$ of *m* sampled lineages are exponentially distributed with mean and variance [40-42].

$$E[T_i] = \frac{2}{i(i-1)} \quad for \quad i = 2, \ldots, m \tag{1}$$

$$Var[T_i] = \left(\frac{2}{i(i-1)}\right)^2 \tag{2}$$

In the sampling limit $m \to \infty$, this exponential process implies that coalescence of the final two lineages ($T_2$) takes, on average, half the time to the most recent common ancestor of the sampled dataset (Figure 6A). While these equations only hold for constant sized populations, the general principle is true for most demographic scenarios of relevance to human history. Many lineages record the recent history of a locus, but older time depths are represented by fewer and fewer lineages that still exist today. Indeed, only two lineages will be present for approximately half the total age of the tree. These
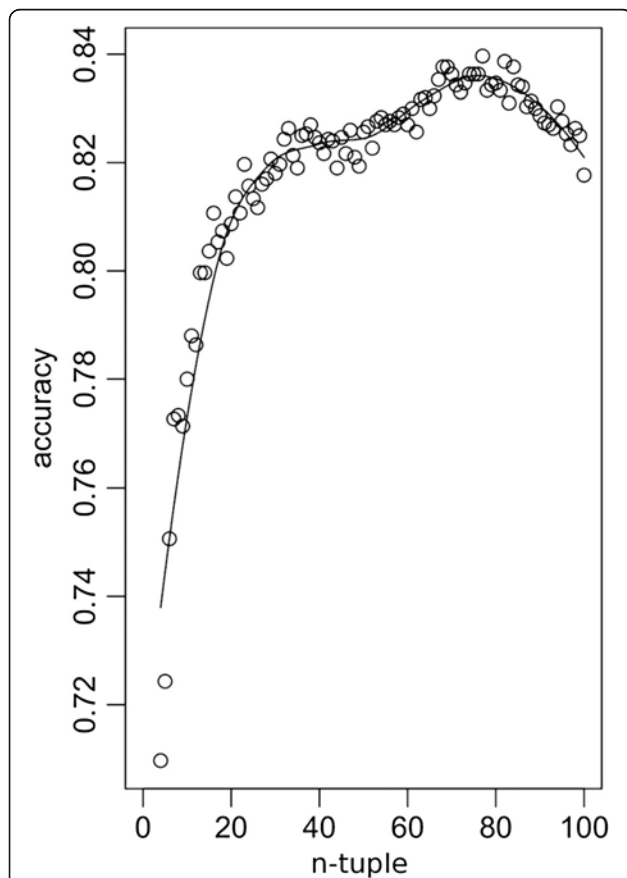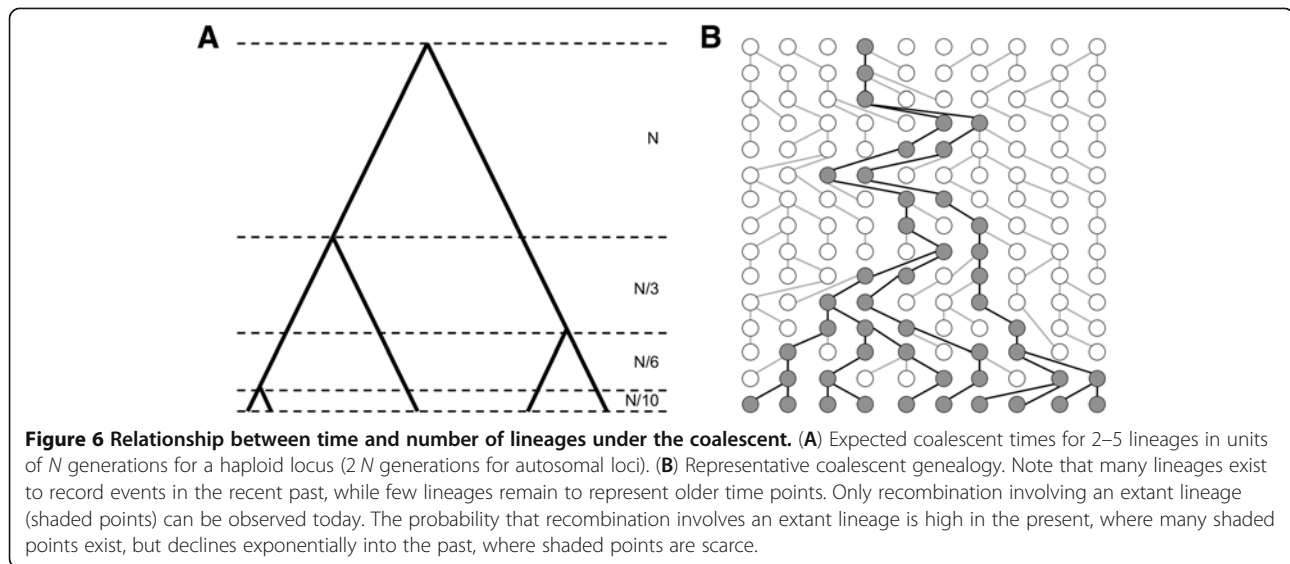
**Figure 6 Relationship between time and number of lineages under the coalescent. (A)** Expected coalescent times for 2–5 lineages in units of *N* generations for a haploid locus (2 *N* generations for autosomal loci). **(B)** Representative coalescent genealogy. Note that many lineages exist to record events in the recent past, while few lineages remain to represent older time points. Only recombination involving an extant lineage (shaded points) can be observed today. The probability that recombination involves an extant lineage is high in the present, where many shaded points exist, but declines exponentially into the past, where shaded points are scarce.

rootward branches are expected to provide most of the signal for changes in historic recombination rates. However, recombination events from distant times can only be detected in the modern dataset if they directly affect the few old lineages that survive to the present (Figure 6B). Information about historic recombination events therefore declines exponentially into the past.

Further, there is a high probability of observing these two deepest branches, even with very small subsample sizes [43,44].

$$P(root|n) = \frac{n-1}{n+1} \qquad (3)$$

Given four randomly chosen subsamples (i.e., a quartet), *P*(root | *n* = 4) = 3/5 = 0.60. Even with just four quartets, the probability that the oldest two branches are not observed is less than 5%. This equation gives the average expectation across all trees, but the general principle holds even for highly unbalanced trees. At the *RRM2P4* locus, which putatively introgressed from archaic hominins into modern humans, 232 individuals are observed on one side of the tree's basal split, while only 21 are found on the other [35]. The probability of observing the root with a single quartet is still high, *P*(root | *n* = 4) = 0.29. Importantly, all *n*-tuples that sample these two oldest branches return exactly the same information about these oldest recombination events (Figure 4C). It follows that simply observing a larger number of *n*-tuples is not sufficient to obtain more information about recombination in the oldest half of the ancestral recombination graph. Conversely, young quartets offer many possible sampling permutations (Figure 4A), and each of these can potentially provide independent information about recent recombination events. All datasets therefore

record more information about recent recombination rates, while power to detect old recombination events declines exponentially backwards into the past.

## Conclusions

A natural limit places important constraints on our ability to reconstruct past changes in recombination rates. If the change occurred recently, sufficient extant lineages may still record the event, and *n*-tuple subsampling is likely to be an informative technique. Moving further in time from the change, the power to reconstruct the recombination profile decreases exponentially. If the change occurs beyond the coalescent (i.e., the most recent common ancestor of the dataset), it obviously cannot be reconstructed at all. Extremely detailed changes in past recombination profiles, particularly for more distant events or complex genomic loci, will remain challenging. However, we show that *n*-tuple subsampling does have sufficient power to reconstruct some aspects of past changes in recombination rates, especially for relatively recent events.

## Methods
### Simulations

The coalescent simulation software *ms* [23] was modified to allow recombination rates to change through time. The C source code of the resulting program, *ms_recomb*, is available from the authors on request. Simulations focus on the most common *trans*-mediated recombination events, and we therefore model changes in recombination without selection (i.e., no transmission distortion).

Genetic datasets were simulated using Kingman's *n*-coalescent [40,41]. To ground the simulations in a realistic framework, model parameters were chosen to reflect

biologically meaningful values for humans. We purposely simulated a generic human-like deme rather than any specific population. Simulations were generated for a single Wright-Fisher deme with a constant effective population size ($N_e = 10^4$) (i.e., the estimated global effective population size of modern humans) [24-26]. A sample of 10-kb autosomal sequences was simulated for 100 individuals with a mutation rate, $\mu$, and average recombination rate, $r$, of $3.75 \times 10^{-8}$ events/bp/generation [27]. Unless otherwise noted, this process was iterated $10^4$ times for each model.

The recombination rate was either held constant, or allowed to vary linearly, exponentially or logistically through time for $10^4$ generations (cf.[24,26]). Linear rates were incremented by the reciprocal of the generation units per generation, exponential rates were fitted to a curve with $\lambda = 5 \times 10^{-4}$, and logistic rates were fitted to a curve with $K = 100$, $N = 1$ and $r = 9 \times 10^{-4}$. (Note that these curves are for exploratory purposes only. They are not intended to represent real rates of change in human populations). The total amount of recombination was constrained so as to be identical for all models, but was apportioned through time according to the constant, linear, exponential and logistic distributions described above. Overall population recombination rates (i.e., $\rho = 4N_er = 15$) were chosen to mimic regions of very strong recombination in real human groups [11]. Low and high constant rates were defined as 15% and 85% of the maximum rate under the corresponding linear models.

To infer past recombination rates, samples were taken at a single time point and surveyed using $n$-tuple subsampling (see main text for details). To determine how summary statistics respond to changing recombination rates, variation in summary statistics was tracked over a time span of $10^4$ generations by taking $10^4$ independent coalescent simulations at each of 500 20-generation intervals.

## Summary statistics
Summary statistics were calculated using functions from the libsequence library [45]. The C++ source code of the resulting program, *msstats_recomb*, is available from the authors on request. The number of segregating sites, $S$, controls for the population mutation rate $\theta$ ($= 4N_e\mu$) and summarizes the total length of the genealogy [46]. A suite of eight additional summary statistics was employed to capture different aspects of the recombination profile: $R_{min}$, the minimum number of recombination events calculated from observed four-gamete violations [47]; *rmmg*, a conservative lower bound on $R_{min}$ proposed by Myers and Griffiths (equation four in [48]); *nHaps*, the number of observed unique sequence haplotypes; *HapDiv*, the haplotype diversity, expected heterozygosity, or probability that two sequences chosen

randomly from the sample are different [49]; Wall's $B$ and $Q$, variant estimators of the number of congruent polymorphic sites (i.e., segregating sites in complete linkage disequilibrium) [50]; Hudson's $C$, an estimator of the population recombination rate $\rho$ ($= 4N_er$) estimated from the variance of pairwise sequence differences [51]; and $Z_{nS}$, the mean pairwise $r^2$ estimate of linkage disequilibrium across all polymorphic sites [52].

## Statistics
Correlations between scaled and unscaled summary statistics, and discriminant analyses, were calculated using the statistical software $R$ [53]. Local regressions were performed using a polynomial of degree 2, a smoothed-particle hydrodynamics (SPH) kernel, and a 50% nearest neighbor bandwidth with a 10% constant component.

## Additional files

**Additional 1: Figure S1.** Response of summary statistics to constant recombination rates.

**Additional 2: Figure S2.** Response of summary statistics to recombination rates decreasing linearly.

**Additional 3: Figure S3.** Response of summary statistics to recombination rates increasing linearly.

**Additional 4: Figure S4.** Response of summary statistics to recombination rates decreasing exponentially.

**Additional 5: Figure S5.** Response of summary statistics to recombination rates increasing exponentially.

**Additional 6: Figure S6.** Response of summary statistics to recombination rates decreasing logistically.

**Additional 7: Figure S7.** Response of summary statistics to recombination rates increasing logistically.

**Additional 8: Determining the optimal scaling factor to capture $n$-tuple age.**

**Additional 9: Figure S8.** Correlations between recombination summary statistics scaled by $S$.

### Author details
[1]Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North 4442, New Zealand. [2]School of Mathematics and Physics, University of Tasmania, Hobart, Australia.

## References

1. Paigen K, Petkov P: **Mammalian recombination hot spots: properties, control and evolution.** *Nat Rev Genet* 2010, **11**:221–233.
2. Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G, Sigurdardottir S, Richardsson B, Jonsdottir J, Thorgeirsson T, Frigge ML, *et al*: **Recombination rate and reproductive success in humans.** *Nat Genet* 2004, **36**(11):1203–1206.
3. Calabrese P: **A population genetics model with recombination hotspots that are heterogeneous across the population.** *Proc Natl Acad Sci USA* 2007, **104**(11):4748–4752.
4. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL: **Comprehensive human genetic maps: individual and sex-specific variation in recombination.** *Am J Hum Genet* 1998, **63**(3):861–869.
5. Neumann R, Jeffreys AJ: **Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation.** *Hum Mol Genet* 2006, **15**(9):1401–1411.
6. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, *et al*: **A high-resolution recombination map of the human genome.** *Nat Genet* 2002, **31**(3):241–247.
7. Khil PP, Camerini-Otero RD: **Variation in patterns of human meiotic recombination.** *Genome Dyn* 2009, **5**:117–127.
8. Smukowski CS, Noor MAF: **Recombination rate variation in closely related species.** *Heredity* 2011, **107**(6):496–508.
9. Webster MT, Hurst LD: **Direct and indirect consequences of meiotic recombination: implications for genome evolution.** *Trends Genet* 2012, **28**(3):101–109.
10. Jeffreys AJ, Neumann R: **The rise and fall of a human recombination hot spot.** *Nat Genet* 2009, **41**(5):625–629.
11. Laayouni H, Montanucci L, Sikora M, Melé M, Dall'Olio GM, Lorente-Galdos B, McGee KM, Graffelman J, Awadalla P, Bosch E, *et al*: **Similarity in recombination rate estimates highly correlates with genetic differentiation in humans.** *PLoS One* 2011, **6**(3):e17913.
12. Spencer CCA, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G: **The influence of recombination on human genetic diversity.** *PLoS Genet* 2006, **2**(9):e148.
13. Úbeda F, Wilkins JF: **The Red Queen theory of recombination hotspots.** *J Evol Biol* 2011, **24**(3):541–553.
14. Myers S, Freeman C, Auton A, Donnelly P, McVean G: **A common sequence motif associated with recombination hot spots and genome instability in humans.** *Nat Genet* 2008, **40**(9):1124–1129.
15. Zheng J, Khil P, Camerini-Otero RD, Przytycka T: **Detecting sequence polymorphisms associated with meiotic recombination hotspots in the human genome.** *Genome Biol* 2010, **11**(10):R103.
16. Coop G, Myers SR: **Live hot, die young: transmission distortion in recombination hotspots.** *PLoS Genet* 2007, **3**(3):e35.
17. Hellenthal G, Pritchard JK, Stephens M: **The effects of genotype-dependent recombination, and transmission asymmetry, on linkage disequilibrium.** *Genetics* 2006, **172**(3):2001–2005.
18. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV: **Genetic recombination is directed away from functional genomic elements in mice.** *Nature* 2012, **485**:642–645.
19. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P: **Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination.** *Science* 2010, **327**(5967):876–879.
20. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B: **PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice.** *Science* 2010, **327**(5967):836–840.
21. Kauppi L, Stumpf MPH, Jeffreys AJ: **Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region.** *Genomics* 2005, **86**(1):13–24.
22. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: **Human recombination hot spots hidden in regions of strong marker association.** *Nat Genet* 2005, **37**(6):601–606.
23. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation.** *Bioinformatics* 2002, **18**(2):337–338.
24. Cox MP, Hammer MF: **A question of scale: human migrations writ large and small.** *BMC Biol* 2010, **8**:98.
25. Cox MP, Morales DA, Woerner AE, Sozanski J, Wall JD, Hammer MF: **Autosomal resequence data reveal late stone Age signals of population expansion in sub-saharan african foraging and farming populations.** *PLoS One* 2009, **4**(7):e6366.
26. Cox MP, Woerner AE, Wall JD, Hammer MF: **Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow.** *BMC Genet* 2008, **9**:e76.
27. Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF: **A novel DNA sequence database for analyzing human demographic history.** *Genome Res* 2008, **18**:1354–1361.
28. Schmid J, Cannon RD, Holland BR: **A futile act? Thoughts on the reproductive biology of *Candida albicans*.** *Mycologist* 2004, **18**(4):158–163.
29. Efron B, Tibshirani R: *An Introduction to the Bootstrap.* Boca Raton, FL: Chapman & Hall; 1993.
30. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39**(4):783–791.
31. Hastie T, Tibshirani R, Friedman J: *The elements of statistical learning: data mining, inference and prediction.* Berlin: Springer Verlag; 2009.
32. Beaumont MA, Nielsen R, Robert C, Hey J, Gaggiotti O, Knowles L, Estoup A, Panchal M, Corander J, Hickerson M, *et al*: **In defence of model-based inference in phylogeography.** *Mol Ecol* 2010, **19**(3):436–446.
33. Csilléry K, Blum MGB, Gaggiotti OE, François O: **Approximate Bayesian Computation (ABC) in practice.** *Trends Ecol Evol* 2010, **25**(7):410–418.
34. Cox MP: **Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models.** *Hum Biol* 2008, **80**:335–357.
35. Cox MP, Mendez FL, Karafet TM, Metni Pilkington M, Kingan SB, Destro-Bisol G, Strassmann BI, Hammer MF: **Testing for archaic hominin admixture on the X chromosome: model likelihoods for the modern human *RRM2P4* region from summaries of genealogical topology under the structured coalescent.** *Genetics* 2008, **178**:427–437.
36. Hey J, Nielsen R: **Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*.** *Genetics* 2004, **167**(2):747–760.
37. Fisher RA: **The use of multiple measurements in taxonomic problems.** *Ann Eug* 1936, **7**(2):179–188.
38. Wu W, Mallet Y, Walczak B, Penninckx W, Massart DL, Heuerding S, Erni F: **Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data.** *Anal Chim Acta* 1996, **329**:257–265.
39. Johnston HR, Cutler DJ: **Population demographic history can cause the appearance of recombination hotspots.** *Am J Hum Genet* 2012, **90**(5):774–783.
40. Kingman JFC: **On the genealogy of large populations.** *J Appl Prob* 1982, **19**:27–43.
41. Kingman JFC: **The coalescent.** *Stochastic Process Appl* 1982, **13**:235–248.
42. Wakeley J: *Coalescent Theory: An Introduction.* Greenwood Village, Colorado: Roberts & Company Publishers; 2008.
43. Kliman RM, Hey J: **DNA sequence variation at the *period* locus within and among species at the *Drosophila melanogaster* complex.** *Genetics* 1993, **133**:375–387.
44. Saunders IW, Tavaré S, Watterson GA: **On the genealogy of nested subsamples from a haploid population.** *Adv Appl Prob* 1984, **16**:471–491.
45. Thornton K: **libsequence: a C++ class library for evolutionary genetic analysis.** *Bioinformatics* 2003, **19**(17):2325–2327.
46. Watterson GA: **On the number of segregating sites in genetical models without recombination.** *Theor Popul Biol* 1975, **7**(2):256–276.
47. Hudson RR, Kaplan NL: **Statistical properties of the number of recombination events in the history of a sample of DNA sequences.** *Genetics* 1985, **111**(1):147–164.
48. Myers SR, Griffiths RC: **Bounds on the minimum number of recombination events in a sample history.** *Genetics* 2003, **163**(1):375–394.
49. Nei M: *Molecular evolutionary genetics.* New York: Columbia University Press; 1987.
50. Wall JD: **Recombination and the power of statistical tests of neutrality.** *Genet Res, Camb* 1999, **74**:65–79.
51. Hudson RR: **Estimating the recombination parameter of a finite population model without selection.** *Genet Res* 1987, **50**:245–250.
52. Kelly JK: **A test of neutrality based on interlocus associations.** *Genetics* 1997, **146**:1197–1206.
53. R Development Core Team: *A Language and Environment for Statistical Computing.* 2012. http://www.r-project.org.