Bioinformatics Advance Access published September 23, 2013

An Alignment-free Test for Recombination

Bernhard Haubold [1]*, Linda Krause [1,2], Thomas Horn [3] and Peter Pfaffelhuber [3]

[1]Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Biology, 24306 Plön, Germany
[2]Institute for Neuro- and Bioinformatics, Lübeck University, Germany
[3]Mathematical Stochastics, Mathematical Institute, Freiburg University, Germany

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: Why recombination? is one of the central questions in biology. This has led to a host of methods for quantifying recombination from sequence data. These methods are usually based on aligned DNA sequences. Here we propose an efficient alignment-free alternative.

Results: Our method is based on the distribution of match lengths, which we look up using enhanced suffix arrays. By eliminating the alignment step, the test becomes fast enough for application to whole bacterial genomes. Using simulations we show that our test has similar power as established tests when applied to long pairs of sequences. When applied to 58 genomes of *Escherichia coli*, we pick up the strongest recombination signal from a 125 kb horizontal gene transfer engineered 20 years ago.

Availability: We have implemented our method in the command-line program `rush`. Its C sources and documentation are available under the GNU General Public License from
`http://guanine.evolbio.mpg.de/rush/`

Contact: haubold@evolbio.mpg.de

1 INTRODUCTION

It is surprisingly difficult to account for the prevalence of sex and recombination in nature (Otto and Lenormand, 2002). Classical explanations for the evolution of sex are based on the realization that it can speed up adaptation (Fisher, 1930/1999, ch. 6). In addition, it removes deleterious mutations from the population (Muller, 1932, 1964). However, recombination also leads to the breakup of coadapted genes, which begs the question as to whether recombination has an evolutionary cause, or is a mere consequence of molecular mechanisms such as DNA repair (Felsenstein, 1974). Perhaps both views are correct, as the contemporary consensus is that sex and recombination would only be maintained in multicellular organisms with high mutation rates and largely negative fitness interactions between genes (Otto, 2007). Irrespective of this uncertainty about the function of recombination in unicellular organisms, horizontal gene transfer in bacteria has attracted particular attention, as it often underlies the emergence of virulent pathogens (Baquero, 2004).

*to whom correspondence should be addressed

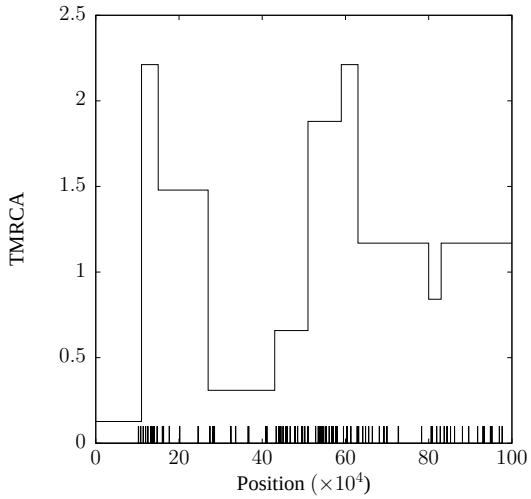

Fig. 1. Simulated time to the most recent common ancestor (TMRCA) and mutations (vertical lines) along a recombining stretch of DNA sequence. TMRCA is proportional to the population size.

The enigma of recombination, combined with its prevalence and clinical importance, has inspired the development of numerous methods for assessing genetic exchange (Posada, 2002). These fall into two broad categories: methods for detecting the presence of recombination, and methods for estimating its rate. We concentrate here on the simpler detection problem. In the past this has been solved in two ways: by looking for clustering of mutations, or by identifying multiple mutations to the same nucleotide in distinct lineages. Such recurrent mutations are called *homoplasies*. Methods based on the detection of clustered polymorphisms include the widely-used Max-χ^2 method (Maynard Smith, 1992) and the runs method implemented in the popular GENECONV program (Sawyer, 1989). Both are based on the realization that the time to the most recent common ancestor varies along a recombining sequence (Figure 1). Since the number of mutations that affect a genomic region is proportional to the time to the most recent common ancestor of the sampled sequences, the number of mutations varies along the sequence; as a result, the mutations are clustered.

Classical homoplasy-based methods rest on the assumption that any nucleotide mutates at most once thus generating three haplotypes between a pair of polymorphisms, say `10`, `01`, and `00`. The missing fourth possible haplotype, `11` in our example, can only be generated through recombination. The detection of such haplotype quartets is used to determine the minimum number of recombination events in an aligned sample of homologous sequences (Hudson, 1985). More recently, this diagnosis of either presence or absence of a recombination event between a pair of polymorphisms was generalized to inferring possibly more than one recombination in samples of more than four sequences. This has been formalized as the Φ_w statistic and implemented in `Phi`, a fast tool for detecting recombination (Bruen *et al.*, 2006). Like all homoplasy-based tests, Φ_w tends to have greater power than its rivals based on polymorphism clustering (Wiuf *et al.*, 2001).

The detection of recombination among a sample of homologous sequences is thus a well-understood problem, as long as an alignment of the sequences is available. However, aligning genomes remains challenging in spite of great advances in this field (Bray and Pachter, 2004; Darling *et al.*, 2004). At the same time, there is interest in analyzing recombination at the scale of bacterial genomes, if not greater (Didelot *et al.*, 2010). Fortunately, the nucleotide-wise assignment of homology that defines alignments is not necessary to test for recombination. We show this by developing a fast, alignment-free test for recombination. Like Max-χ^2 (Maynard Smith, 1992) and GENECONV (Sawyer, 1989), our test is based on the detection of polymorphism clustering. But instead of scoring single nucleotide polymorphisms (SNPs), we record the lengths of exact matches between pairs of sequences. To a first approximation this corresponds to the distances between SNPs. Recombination leads to an increase in the fluctuation of match lengths when compared to no recombination. Since match lengths can be looked up efficiently using modern string algorithms (Puglisi *et al.*, 2007), our test scans a pair of *Escherichia coli* genomes totaling 10 Mb in only 8s on a contemporary laptop.

In the following we derive our test and demonstrate its sensitivity and specificity through simulation. We also use simulations to compare it to two published tests, Max-χ^2 and Φ_w. Finally, we apply our test to 58 *E. coli* genomes to search for horizontal gene transfer.

2 METHODS

2.1 Derivation

Consider the query $q =$ `ATATC` and the subject $s =$ `ACGTG`. At every position i in q we look for the shortest substring $q[i..j]$ that is absent from s. We call this SHortest Unique subSTRING "shustring" (Haubold *et al.*, 2005) at position i and denote its length by X_i. For example, `AT` is the shustring starting at the first position in q, hence $X_1 = 2$. Our approach is built on the assumption that in pairs of homologous DNA sequences the shustrings correspond to homologous matches and hence their lengths represent distances to the next SNP. This assumption is reasonable for closely related sequences such as those sampled from populations or recently diverged species.

Conceptually we identify shustrings using a suffix tree of q and s (Gusfield, 1997). We say "conceptually", because our actual implementation is based on an abstract version of a suffix tree called

	1	2	3	4	5	6
q	A	T	A	T	C	\$
s	A	C	G	T	G	\$

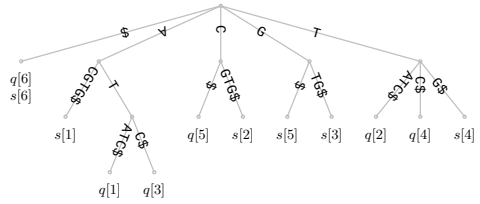

Fig. 2. Suffix tree of our example query and subject sequences s and q.

"enhanced suffix array" (Abouelhoda *et al.*, 2002). Figure 2 shows the suffix tree of q and s. Each suffix is represented by a path from the root to a leaf. For example, $q[1]$ is the leaf that corresponds to the suffix starting at position 1 in q, `ATATC`. Moreover, repeated prefixes are collapsed into single paths. For example, the suffixes $q[5..5] =$ `C` and $s[2..5] =$ `CGTG` share the prefix `C`, which appears only once in the tree. The sentinel character \$ at the end of q and s differs from every character in q and s, even from itself. Its addition ensures that a suffix such as $q[5..5] =$ `C`, which is a prefix of the suffix $s[2..5] =$ `CGTG`, is also represented by a leaf in the tree.

To identify the shustring starting at, say, $q[1]$, we visit leaf $q[1]$, and climb towards the root until we find a node, n, with a subject leaf in the subtree rooted on n. In our example we carry out two climbing steps. Then we extend the path label from the root to the node we have reached by one character towards the starting leaf, to find our first shustring, `AT`. In this way we calculate the shustrings at every position in q.

Our test statistic is based on the lengths of shustrings, X_i. Without recombination, and if sequences differ at a fraction π of their sites, every site i has probability π of differing between q and s, indepently of other sites. Using this model based on the infinite sites model from population genetics, we obtain for the average shustring length

$$\overline{X} := \frac{1}{l}\sum_{i=1}^{L} X_i, \quad E[\overline{X}] = E[X_i] = \frac{1}{\pi}.$$

Hence π can be estimated from unaligned sequences as the inverse of the average shustring length (Haubold *et al.*, 2011). Recall that recombination leads to fluctuations in the coalescence times along a sequence and therefore to clustering of polymorphisms (Figure 1). As a result, the mean shustring length increases, which might be used as an indicator of recombination (Haubold and Pfaffelhuber, 2012). Unfortunately, we found that it was impossible to infer the expected average shustring length without estimating π to a precision attainable only with alignments. However, the empirical variance of shustring lengths is a more promising statistic. Its expectation without recombination is

$$s^2 = \frac{1}{l}\sum_{i=1}^{l}(X_i - \overline{X})^2, \quad E[s^2] \approx \frac{1}{\pi^2} \equiv E[\overline{X}]^2,$$

where we can estimate π from the average shustring length (Haubold and Pfaffelhuber, 2012). $E[s^2]$ is then compared to the observed

variance, s^2, to test the null hypothesis $H_0 : E[s^2] = E[\overline{X}]^2$. For this, we derived

$$\mathrm{Var}(s^2) \approx \frac{24}{l\pi^5} \quad (1)$$

and assumed that s^2 is normally distributed with expectation $\overline{X}^2$ (Supplementary Material). This means that can we use a one-sided test for the normalized difference between s^2 and $\overline{X}^2$

$$D_\mathrm{r} = \frac{s^2 - \overline{X}^2}{\sqrt{24 \cdot \overline{X}^5 / l}}, \quad (2)$$

which is approximately normally distributed with mean 0 and standard deviation 1. We also define the ratio

$$Q = \frac{s^2}{\overline{X}^2}$$

as a rough measure of recombination.

2.2 Implementation

We implemented our test in the program `rush`, which stands for "Recombination detection Using SHustrings". `rush` takes as input a query and a subject DNA sequence in FASTA format and computes Q, D_r, and the corresponding P-value. Its suffix tree construction is based on the deep-shallow algorithm and implementation by Manzini and Ferragina (2002), which is one of the most efficient string indexing methods available (Puglisi *et al.*, 2007).

2.3 Data

We downloaded complete genome sequences from GenBank for the 58 *E. coli* strains listed in Table S1. Since `rush` is restricted to the analysis of sequences consisting solely of the nucleotide designations $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, all other characters were removed prior to the analysis.

2.4 Simulations and Computer Programs

We generated samples of homologous DNA sequences using the coalescent simulation program `ms` (Hudson, 2002) in conjunction with `ms2dna` available from

`http://guanine.evolbio.mpg.de/bioBox/`

These simulations were conditioned on the population mutation rate, θ, and the population recombination rate, ρ. $\theta = 2N_\mathrm{e}\mu$, where N_e is the effective population size and μ is the mutation probability per generation. Under the infinite sites model, θ is equal to the expected number of pairwise mismatches, π. Similarly, $\rho = 2N_\mathrm{e}c$, where c is the probability of recombination per generation. For more background on coalescent theory see the excellent introduction by Wakeley (2009).

Max-χ^2 and Φ_w-values were computed using the program `Phi` (Bruen *et al.*, 2006). The distances between the *E. coli* genomes were computed using `kr` (Domazet-Lošo and Haubold, 2009) and the tree based on these distances was computed and drawn with `Phylip` (Felsenstein, 2005) using neighbor-joining and mindpoint-rooting.

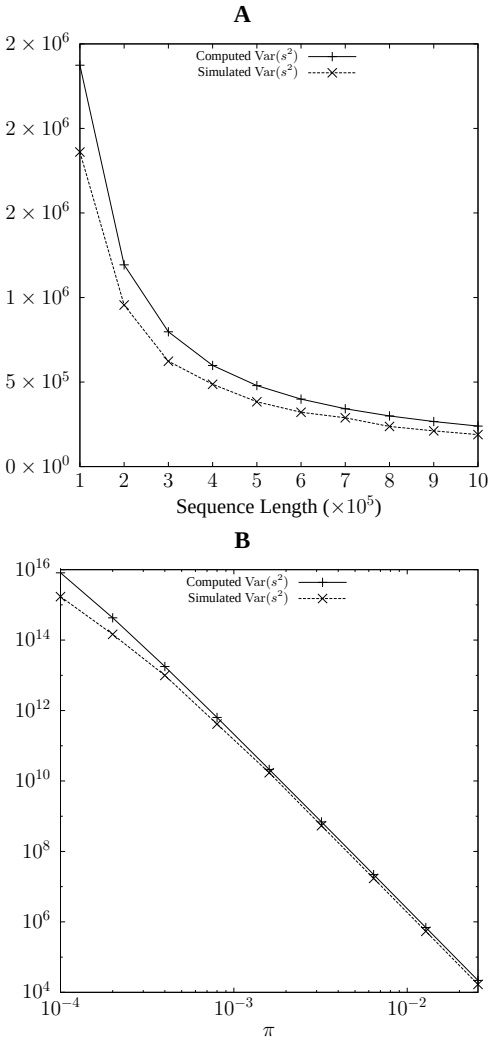

Fig. 3. Comparing the observed variance of the variance in shustring length, $\mathrm{Var}(s^2)$, simulated without recombination to $\mathrm{Var}(s^2)$ calculated according to equation (1) along the two dimensions of $\mathrm{Var}(s^2)$, sequence length (**A**) and genetic diversity, π (**B**). Number of replicates: 10^4; in **A** $\pi = 10^{-2}$; in **B** sequence length $= 10^5$.

3 RESULTS

3.1 Simulations

We explored the accuracy of the newly derived equation (1) through simulations. In Figure 3A we varied sequence length. The simulated values of $\mathrm{Var}(s^2)$ were always smaller than equation 1, but the two curves are quite close. Similarly, when varying π in Figure 3B, we found that equation (1) is greater than the simulated value, though not much. Notice also that as expected from equation (1), a 10-fold change in sequence length corresponds to a 10-fold change in $\mathrm{Var}(s^2)$ (Figure 3A), while a 10-fold change in π corresponds to an enormous 10^5-fold change in $\mathrm{Var}(s^2)$ (Figure 3B).

This comparison between simulated and computed $\mathrm{Var}(s^2)$ values suggests that our test is conservative. Accordingly, we found that for the low rate of recombination of $\rho/\mathrm{bp} = 10^{-7}$ the null

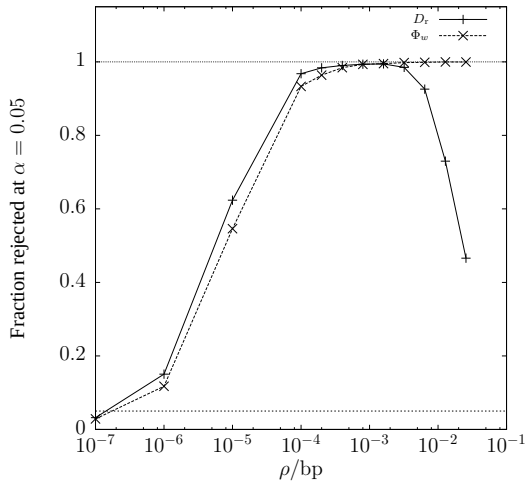

Fig. 4. D_{r} has similar sensitivity as the alignment-based Φ_w (Bruen *et al.*, 2006). The graph shows the fraction of hypothesis tests rejected with significance $P \leq 0.05$ as a function of the rate of recombination, ρ. Horizontal line: 0.05; Sequence length $= 10^6$, $\pi = 10^{-3}$, replicates $= 10^4$; sample size for Φ_w: 4.

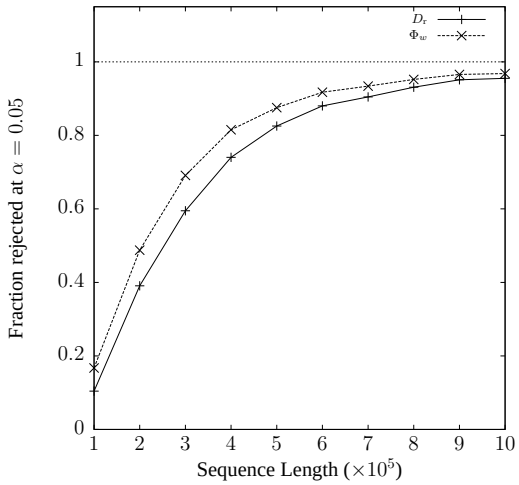

Fig. 5. Rejection rate as a function of sequence length for D_{r} compared to Φ_w (Bruen *et al.*, 2006); $\pi = 10^{-3}$, $\rho = 100$ for entire region, replicates $= 10^4$.

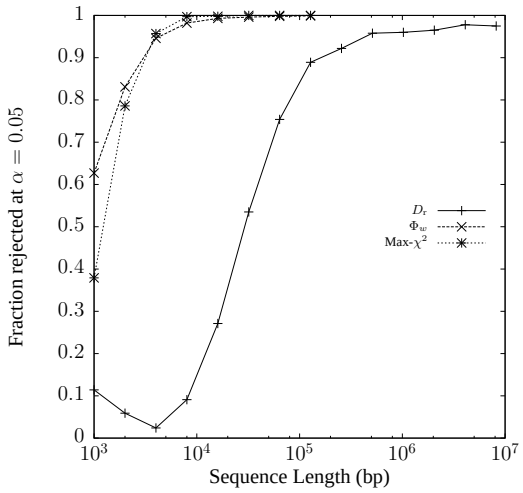

Fig. 6. Comparing rejection frequencies between Φ_w, Max-χ^2, and D_{r} as a function of the length of the input sequences. The simulation parameters are described in the text.

hypothesis of no recombination is rejected at $\alpha = 0.05$ with a frequency of only 0.032 rather than the expected 0.05 (Figure 4). However, with increasing recombination the rejection rate grows to greater than 0.99 until $\rho \approx \pi$. When ρ exceeds π, the rejection rate declines again, as our test statistic D_{r} is maximal if $\pi \approx \rho$. We also compared the rejection rate with D_{r} to that with the homoplasy-based test statistic Φ_w, which is computed from four or more aligned sequences (Bruen *et al.*, 2006). Figure 4 shows that Φ_w has a similar rejection curve as D_{r}, with slightly less sensitivity at the simulated parameter combination.

However, the sensitivity of D_{r} and Φ_w depends on the number of mismatches sampled. Figure 5 shows the rejection rate as a function of sequence length at $\pi = 10^{-3}$ and a length-invariant expected number of recombination events of 100. When sequence length is 10^5, D_{r} and Φ_w have little power, while ten times longer sequences yield rejection in $\geq 97\%$ of cases, with Φ_w marginally outperforming D_{r}.

It is reassuring that D_{r} is highly sensitive in the limit of long sequences. This is equivalent to saying that D_{r} is sensitive when applied to polymorphic sequences as long as π is small. This proviso follows from the assumption that shustrings are homologous. The greater π, the more shustrings occur that are generated by random, non-homologous matches between query and subject.

Traditional methods for detecting recombination such as Max-χ^2 were designed for sequences no longer than a few kb. To compare D_{r} to Max-χ^2, we used the simulation scheme by Bruen *et al.* (2006): For Φ_w and Max-χ^2 we simulated samples of 10 sequences with $\theta/\mathrm{bp} = 0.01$ and $\rho = 16$. For 1 kb sequences Φ_w was more sensitive than Max-χ^2, in agreement with Bruen *et al.* (2006). For D_{r} we simulated pairs of sequences with the same θ, and a ρ value that resulted in the same expected number of recombination events as in the sample of ten; that is, we used $\rho = 16 \times \sum_{i=1}^{9} 1/i = 45.3$ (Hudson and Kaplan, 1985). For kb-length sequences, D_{r} has little power. However, the power of all methods to detect recombination increases with sequence length (Wiuf *et al.*, 2001). For sequences of 100 kb or more D_{r} is quite sensitive. Incidentally, this is the sequence length where we stopped the Max-χ^2 simulations as they became too slow.

The Φ_w statistic was developed as a fast alternative to methods such as Max-χ^2. We therefore simulated sequence quartets of lengths 10^4–10^7 bp and timed `Phi` *given* an alignment. The slope of the resulting run time curve was 2.2, that is, doubling the sequence length increases the run time $2^{2.2} = 4.6$-fold. We compared this to `rush` when applied to sequence pairs of the same lengths. The slope of its run time curve is 1.2. This is still a bit steeper than the optimal slope of 1, but means that with double the input data `rush` takes 2.3

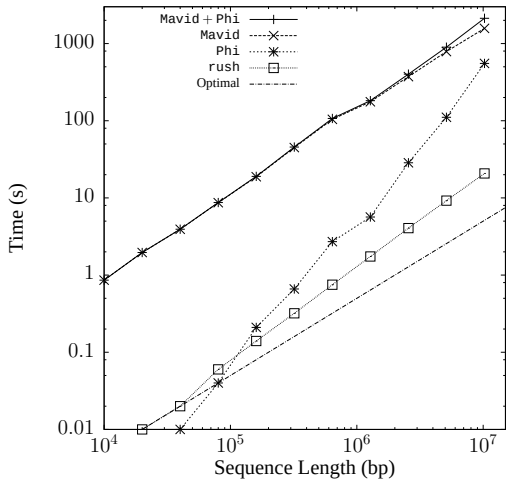

Fig. 7. Comparing the run times of `Phi` and `rush` as a function of sequence length. $\theta/\mathrm{bp} = 0.01$, $\rho = 1000$.

times longer. As a consequence, `Phi` is eventually outperformed by `rush` for sequences longer than 100 kb (Figure 7).

Since `Phi` requires aligned sequences as input, we also timed aligning sequence quartets with `Mavid` (Bray and Pachter, 2004). `Mavid` is very fast and the slope of its run time curve is only 1.1, making it virtually optimal. Still, `rush` is roughly 100 times faster than alignment-based `Phi`.

In our final set of simulations we investigated the potentially confounding effects of codon data and changes in GC-content, and found that the rejection rate was unaffected by either (Figures S1 and S2).

3.2 Horizontal Gene Transfer in *E. coli*

Having established the accuracy and robustness of `rush` through simulations, we applied it to the 58 fully sequenced *E. coli* genomes available from GenBank at the time of writing (Table S1). Figure 8 shows a cluster diagram of the strains computed from their complete genomes.

The $58 \times 57 = 3306$ pairwise recombination tests between the 58 *E. coli* strains took `rush` 15 hours, 9 minutes, and 15 seconds on an Intel Xeon 2.40GHz CPU, an average of 16.5s per test. 97% of these tests were rejected with significance $P \leq 0.05$, indicating that the model assumption of uniformly distributed mutations is usually violated across whole bacterial genomes. Figure 9 shows the distribution of the 3306 values of our recombination measure Q; its median is 2.951 with a huge range of 0.587 to 40.939. We focused on the two strains with the largest Q, KO11 *vs.* K12_MG1655, which are marked by arrows in the cluster diagram (Figure 8).

To search for evidence of horizontal gene transfer, we looked for regions in the KO11 genome that were more closely related to K12_MG1655 than to its closest relative. The cluster diagram (Figure 8) indicates that the closest relative of KO11 is KO11FL. However, this is an artifact of the clutering algorithm, which guarantees finding the correct tree only for distances that actually fit a tree. Empirical distances are often not strictly tree-like, especially in the presence of recombination. When we look up the raw pairwise

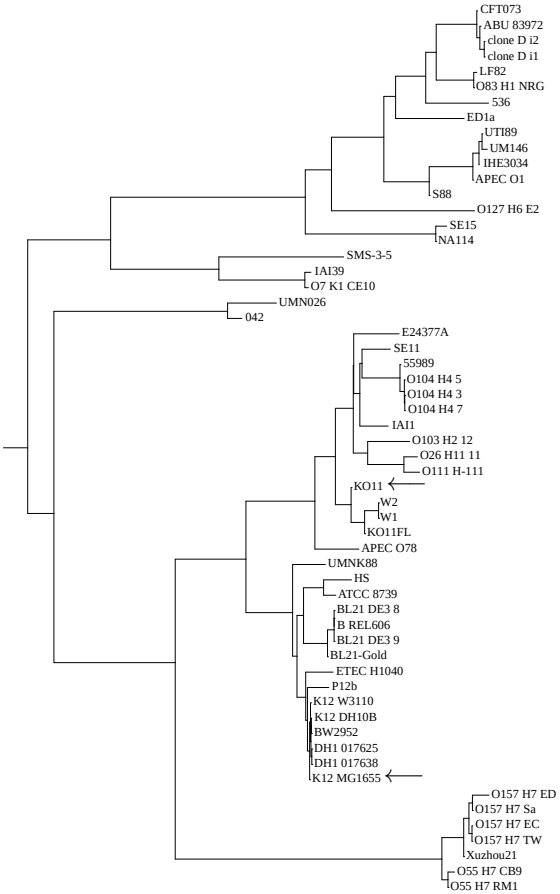

Fig. 8. Midpoint-rooted neighbor joining tree of 58 *Escherichia coli* strains. The two genomes with the highest recombination measure Q, K12_MG1655 and KO11, are indicated by arrows.

distances, we find that the distance between KO11 and KO11FL is 2.1×10^{-5}, while that between KO11 and W1 or W2 is below the sensitivity of `kr`, the program we used for estimating the substitution rates between genomes (Domazet-Lošo and Haubold, 2009). Hence we took W1 as the closest relative of KO11. Figure 10 shows that at position 4.4 Mb KO11 contains large fragments more similar to K12_MG1655 than to W1. The longest of these spans 102 kb and is located at 4,385,866–4,487,685. When blasting this region against W1, the best hit is 46.9 kb long and contains 367 mismatches and 13 gaps. In contrast, the best hit in K12_MG1655 is 78.6 kb long with just 8 mismatches and 1 gap. The next best hit in K12_MG1655 is 23.5 kb long with 4 mismatches and 0 gaps. This is strong evidence for horizontal gene transfer. It looks as if it affected a common ancestor of KO11 and KO11FL, because in KO11FL the 102 kb query generates two top hits, one 68.2 kb long, the other 33.5 kb with just 2 mismatches and 0 gaps.

4 DISCUSSION

To make the most of genomic sequences, we would ideally use software that allows us to query them interactively. Our fast,

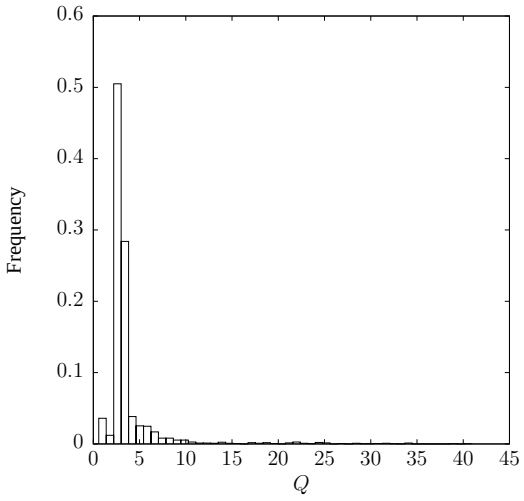

Fig. 9. Histogram of all 3306 measures of recombination, Q, computed in pairwise comparisons between the 58 *E. coli* strains shown in Figure 8.

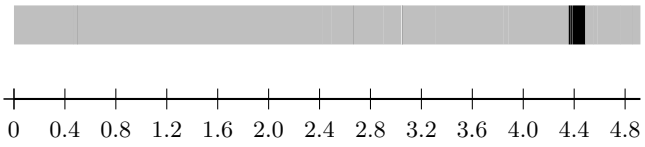

Fig. 10. Comparison between *E. coli* strain KO11 as query and the two strains K-12_MG165 and W1 as subject using the program `alfy` (Domazet-Lošo and Haubold, 2011). Regions where KO11 is most closely related to W1 are shown in light gray, regions closer to K-12_MG165 in black; query regions with no close homologue in the subject sequences are shown in white. X-axis gives positions in Mb.

alignment-free test of recombination is intended as a step toward this goal. `rush` achieves its speed by using similar string-indexing techniques as applied in the genome-aligner `MUMmer` (Kurtz *et al.*, 2004). The central feature of this approach to sequence analysis is that it is "optimal" in the sense that in theory it runs in time linear in the length of the sequences analyzed. In practice, programs based on modern string indexing techniques are very fast and memory efficient (Puglisi *et al.*, 2007). We have been interested for some time now in bringing the power of these algorithms to biology by combining them with modeling the distribution of shortest unique substrings (shustrings), which we take as proxy for the distribution of the distances to the next polymorphism (Haubold *et al.*, 2005, 2009; Domazet-Lošo and Haubold, 2009, 2011). In particular, in our previous derivation of an alignment-free estimator of genetic diversity we noted that the mean shustring length is sensitive to recombination (Haubold and Pfaffelhuber, 2012).

Here we have used the test statistic D_{r} defined in equation (2), which is based on the variance of the shustring length. Specifically, we compared its observed value, s^2, to its expectation, $\overline{X}^2$. We tested the null hypothesis that $D_{\mathrm{r}} = 0$ by deriving $\mathrm{Var}(s^2)$ and constructing a parametric hypothesis test. This contrasts with the approach taken by the authors of `Phi`, our main point of reference: They test the null hypothesis of no recombination using a permutation test (Bruen *et al.*, 2006). This was not an option for us, because generating the null distribution of shustring lengths by shuffling polymorphisms requires an alignment. Moreover, a parametric approach is faster, albeit sometimes less accurate, than its Monte Carlo equivalent. In our case, $\mathrm{Var}(s^2)$ is always slightly larger than its simulated counterpart (Figure 3). Our test is thus conservative when applied to simulated data. This is illustrated by Figure 4, where no recombination leads to 3% rejections with $\alpha = 5\%$. However, a 3% rejection rate is reasonably close to the expected 5%, giving us additional confidence, that the test statistic D_{r} behaves as desired.

We compared Max-χ^2 to D_{r} using one of the parameter combinations of Bruen *et al.* (2006). Our 1 kb results in Figure 6 were compatible with theirs; our results also document that D_{r} is not a replacement of established methods, but rather complements them for Mb-length sequences. Application to sequences of this length is possible because `rush` runs approximately 100 times faster than `Phi` plus alignment (Figure 7).

Given that our test is conservative (Figure 4), it might come as a surprise that 97% of the pairs of *E. coli* genomes tested had a significant D_{r} (Figure 9). This points to a weakness our method shares with all methods for detecting recombination that are based on the identification of clustered polymorphisms. These methods are sensitive to variations in the rate of mutation. This is the reason for the superiority of homoplasy-based methods including Φ_w over cluster-based methods (Bruen *et al.*, 2006).

In the present study we compared in particular the alignment-based test statistic Φ_w to our alignment-free test statistic D_{r}. An important difference between them is that Φ_w is applied to at least four aligned sequences, while D_{r} compares unaligned pairs of query and subject sequences. In our example application the subject was always a single genome, but it could also consist of several concatenated sequences.

Statistical tests exist independently of their implementations. However, to analyze simulated and experimental data, we rely on the tests' implementations, which may or may not be the best possible. With this proviso in mind, we computed Φ_w using the published program `Phi` (Bruen *et al.*, 2006), and D_{r} using our new program `rush`. `rush` is always faster than the alignment step necessary for applying `Phi`. However, for sequences longer than 10^5 `rush` is even faster than `Phi` *given* an alignment (Figure 7). Its ability to work without alignment makes `rush` not only fast, it also facilitates its application to genomes that are available only as sets of contigs. This is useful, because genomes are increasingly published in this form rather than as fully assembled chromosomes. In our study, not aligning enabled us to efficiently search for the genome pair with the largest value of $Q = s^2/\overline{X}^2$, which was KO11 as query and K12_MG1655 as subject. The closest relative of KO11 is strain W. Together with strains K12, B, and C this belongs to the select group of four *E. coli* strains classified as Risk Group 1 organisms; these are safe to use in the lab. It turns out that KO11 was engineered in 1991 from W to produce ethanol (Ohta *et al.*, 1991; Turner *et al.*, 2012). In the process, two chunks of DNA were transferred into what became KO11: three genes for ethanol production from *Zymonas mobilis*, and 125 kb of the laboratory work horse *E. coli* K12_MG1655. These 125 kb comprise the *uvrA–mutL* region of the K12_MG1655 chromosome. The 102 kb fragment we investigated starts inside the *uvrA* locus. In other words, we have picked up a region that was

Fig. S1.

Fig. S2.

Table S1.

acquired from K12_MG1655 20 years ago. During these 20 years KO11 was serially transferred and evolved into KO11FL. We found that KO11FL had diverged more from its ancestor KO11 than from their common ancestor W. This was also reported by the team that sequenced KO11FL (Turner *et al.*, 2012).

CONCLUSION

Our fast method for detecting recombination from unaligned genomes is accurate when applied to simulated data. Applied to bacterial genomes, it diagnoses recombination too frequently, due to variation in mutation rate across long sequences. This is a weakness D_{r} shares with other methods to detect recombination based on identifying polymorphism clustering (Bruen *et al.*, 2006). Nevertheless, using the recombination measure Q, we discovered the strongest signal for the pair of *E. coli* genomes that had undergone an engineered 125 kb horizontal gene transfer 20 years ago.

ACKNOWLEDGEMENT

We are grateful to Paul Rainey for helpful comments.

Funding: PP is supported by the Deutsche Forschungsgemeinschaft through grant Pf672/3-1.

REFERENCES

Abouelhoda, M., Kurtz, S., and Ohlebusch, E., 2002. The enhanced suffix array and its applications to genome analysis. In *Proceedings of the Second Workshop on Algorithms in Bioinformatics*, 449–463. Lecture Notes in Computer Science 2452, Springer-Verlag.

Baquero, F., 2004. From pieces to patterns: evolutionary engineering in bacterial pathogens. *Nature Reviews Microbiology* 2, 510–518.

Bray, N. and Pachter, L., 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Research* 14, 693–699.

Bruen, T. C., Philippe, H., and Bryant, D., 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172, 2665–2681.

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangement. *Genome Research* 14, 1394–1403.

Didelot, X., Lawson, D., Darling, A., and Falush, D., 2010. Inference of homologous recombination in bacteria using whole genome sequences. *Genetics* 186, 1435–1449.

Domazet-Lošo, M. and Haubold, B., 2009. Efficient estimation of pairwise distances between genomes. *Bioinformatics* 25, 3221–3227.

Domazet-Lošo, M. and Haubold, B., 2011. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* 27, 1466–1472.

Felsenstein, J., 1974. The evolutionary advantage of recombination. *Genetics* 78, 737–756.

Felsenstein, J., 2005. PHYLIP (phylogeny interference package) version 3.6.

Fisher, R. A., 1930/1999. *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, Variorum edition.

Gusfield, D., 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge.

Haubold, B. and Pfaffelhuber, P., 2012. Alignment-free population genomics: an efficient estimator of sequence diversity. *Genes, Genomes, Genetics* 2, 883–889.

Haubold, B., Pfaffelhuber, P., Domazet-Lošo, M., and Wiehe, T., 2009. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology* 16, 1487–1500.

Haubold, B., Pierstorff, N., Möller, F., and Wiehe, T., 2005. Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics* 6, 123.

Haubold, B., Reed, F. A., and Pfaffelhuber, P., 2011. Alignment-free estimation of nucleotide diversity. *Bioinformatics* 27, 449–455.

Hudson, R. R., 1985. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* 109, 611–631.

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

Hudson, R. R. and Kaplan, N. L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111, 147–164.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S., 2004. Versatile and open software for comparing large genomes. *Genome Biol* 5, R12.

Manzini, G. and Ferragina, P., 2002. Engineering a lightweight suffix array construction algorithm. In *ESA '02: Proceedings of the 10th Annual European Symposium on Algorithms*, 698–710. Springer-Verlag, London, UK.

Maynard Smith, J., 1992. Analysing the mosaic structure of genes. *Journal of Molecular Evolution* 34, 126–129.

Muller, H. J., 1932. Some genetic aspects of sex. *American Naturalist* 66, 118–138.

Muller, H. J., 1964. The relation of recombination to mutational advance. *Mutation Research* 1, 2–9.

Ohta, K., Beall, D. S., Mejia, J. P., Shanmugarn, K. T., and Ingram, L. O., 1991. Genetic improvement of Escherichia coli for ethanol production: chromosomal integration of zymomonas mobilis genes encoding pyruvate decarboxylase and alcohol dehydrogenase ii. *Appl. Environ. Microbiol.* 57, 893–900.

Otto, S. P., 2007. Unravelling the evolutionary advantage of sex: a commentary on 'Mutation-selection balance and the evolutionary advantage of sex and recombination' by Brian Charlesworth. *Genet. Res. Camb.* 89, 447–449.

Otto, S. P. and Lenormand, T., 2002. Resolving the paradox of sex and recombination. *Nature Reviews Genetics* 3, 252–261.

Posada, D., 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular Biology and Evolution* 19, 708–717.

Puglisi, S. J., Smyth, W. F., and Turpin, A. H., 2007. A taxonomy of suffix array construction algorithms. *ACM Comput. Surv.* 39, 4.

Sawyer, S. A., 1989. Statistical tests for detecting gene conversion. *Molecular Biology and Evolution* 6, 526–538.

Turner, P. C., Yomano, L. P., Jarboe, L. R., York, S. W., Bagett, C. L., Moritz, B. E., Zent, E. B., Shanmugam, K. T., and Ingram, L. O., 2012. Optical mapping and sequencing of the *Escherichia coli* KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the *Zymomonas mobilis pdc* and *adhB* genes. *J. Ind. Microbiol. Biotechnol.* 39, 629–639.

Wakeley, J., 2009. *Coalescent Theory: An Introduction*. Roberts & Company, Colorado.

Wiuf, C., Christensen, T., and Hein, J., 2001. A simulation study of the reliability of recombinaiton detection methods. *Molecular Biology and Evolution* 18, 1929–1939.