On the Detection of Recombination in Minisatellite Data

Eric Rivals, Ezekiel Adebiyi

On the Detection of Recombination in Minisatellite Data

Ezekiel Adebiyi
Department of Computer and Information Sciences, Covenant University, PMB 1023, Ota, Nigeria.
and
Eric Rivals
L.I.R.M.M., UMR 5506 CNRS, Université de Montpellier II, 161 rue Ada, F34392 Montpellier Cedex 5, France.

Tandem repeats are repeated sequences whose copies are adjacent along the chromosomes. They account for large portion of eukaryotic genomes and are found in all types of living organisms. Among tandem repeats, those with repeat unit of middle size are called minisatellites. These loci depart from classical loci because of the propensity to vary in size due to the adjunction or the removal of one or more repeat units. Because of this polymorphism, they prove useful in genetic mapping, in population genetic and forensic medicine. Moreover, some specific loci are involved in diseases, like the insulin minisatellite which is implicated in type I diabetes. Those loci also undergo complex recombination events. Presently, program to compare minisatellite alleles exist and yield good results when recombination is absent, but none treats correctly recombinant minisatellite alleles. Our goal is to develop an adequate tool for the detection of recombinant among the minisatellite sequences. By combining a multiple alignment tool and a method based on phylogenetic profiling, we design a first solution, called *MS_PhylPro*, for this task. The method has been implemented, tested on real data sets from the insulin minisatellite, and proves to detect recombinant alleles.

General Terms: multiple alignment, insulin minisatellite, phylogenetic profile, recombinant, cross-over, turnover.

Additional Key Words and Phrases: computational analysis of minisatellite sequence data, also called *map*

1. INTRODUCTION

The genome length in base pairs (bps) displays huge variations among species: from about 10^5 bps for an archebacteria to 3.10^9 for humans, or even to more than 10^{11} bps for the protozoa *Amoeba dubia*. These differences are explained by the presence of regions, called repeats, that occur many times in the genome. Some molecular mechanisms allow the cell to duplicate a genome region. Among different classes of repeats, those whose copies are located one next to the other on the chromosome are termed *tandem repeats*. Particularly in tandem repeats, duplication and its dual events, contraction, may occur at very high frequencies, letting these loci be the most variable (polymorphic) regions for instance in the human genome. *Minisatellites* are tandem repeats whose repeat unit ranges between 7 and 100 bps. The length variability of minisatellites made them markers of choice to study genome variation inside or across population (population genetics), in genetic mapping, and in forensic medicine for individual identification or paternity testing [Jeffreys et al. 1985]. Indeed, the repeat sequences (*alleles*) observed at a given locus in two individuals may be different. Along time, the repeat copies also undergo point mutations (substitutions, insertions, and deletions), all copies are not identical. The sequence of the variants of the repeated unit of a minisatellite can be charted by a technique named *Minisatellite Variant Repeat PCR* [Jeffreys et al. 1991a], the result of which is a sequence over the alphabet of variants (not the DNA alphabet). We will give more details later. From the medical view-point, a lot of interest has been devoted to minisatellite since some loci are involved in disease development [Bois and Jeffreys 1999]. For instance, the insulin minisatellite has proven to be an important genetical factor in polycystic ovary syndrome, obesity, and type I diabetes [Stead and Jeffreys 2000]. Like in hypervariable minisatellites, and in many human minisatellites, the evolution of the INS locus involves recombination, which leads to exchanges of groups of variants between alleles [Buard and Vergnaud 1994].

Recently, progress has been made towards computational analysis of minisatellite sequence data, also called *map*. The main need is to compare the maps that represents the alleles of two individuals. The comparison should measure the differences between the two maps by accounting the number of mutations needed to transform one into the other. Solutions to the minisatellite map alignment problem have been proposed for the case where mutations include point mutations, duplications, and contractions [Bérard and Rivals 2003; Sammeth et al. 2005; Bérard et al. 2006]. Basically these works extend dynamic programming approaches for classical sequences (which do not undergo duplications contractions) to account for long range dependencies in the maps. The pairwise alignment program of [Bérard et al. 2006] has been used to construct a multiple alignment program called *MS_Alimul*[Rivals]. *MS_Alimul* starts with pairwise alignments, and grows them by adding new maps that are close to the maps already in the alignment. The choice of the maps is performed

Ezekiel Adebiyi, Department of Computer and Information Sciences, Covenant University, PMB 1023, Ota, Nigeria. Email address: adebiyi@dkfz-heidelberg.de
Eric Rivals, L.I.R.M.M., UMR 5506 CNRS, Université de Montpellier II, 161 rue Ada, F34392 Montpellier Cedex 5, France. E-mail address: rivals@lirmm.fr

following a *guide tree* obtained by hierarchical clustering of the maps according to their pairwise alignment distances (all pairwise alignments are first precomputed). Alignments are progressively "merged" using alignment aligning procedure from [Kececioglu and Zhang 1998], this is performed iteratively until one obtains a complete multiple alignment. This strategy is known as *progressive alignment* in the literature.

Both pairwise and multiple alignment make sense from the biological point of view if the maps under scrutiny are homologous, that is derive from a common ancestor. The assumption is that mutations represent a small portion of the sequences and it is possible to find a putative series of mutations to transform one sequence in the other. Recombination is able to create a new allele, the *recombinant*, from pieces of two different alleles; these parent alleles may not be homologous or their common ancestor may be so ancient that their sequences are now dissimilar. Recombinant alleles are therefore align-able by pieces, but not completely, with each of its parent or with any map or cluster of maps that similar to one of its parent. When aligning pairwise a recombinant with such a sequence yields an alignment of high similarity in regions of homology and low similarity in others regions. In the case of multiple alignment it pertubates the construction of the progressive alignment by inserting many gaps in the non-homologous region.

A long biological literature is devoted to the detection of recombinant sequences in non repetitive genetic sequences, and these have been reviewed and compared in [Posada 2002] (see references therein). Proposed approaches either (1) rely on the knowledge of a phylogenetic tree, or (2) use the pattern of nucleotidic substitution site-by-site, or (3) computed a distance based criterion on a window of a multiple alignment. In latter class, one finds the *Phylogenetic profile* method of [Weiller 1998] (also developed in RAT [Etherington et al. 2005]).

Up to now, for repeat sequence in general and for minisatellite data in particular, none of the available comparison approaches deals with recombinant alleles. As the availability of a "phylogenetic" tree of the minisatellites alleles is unlikely, and as in minisatellite, variant duplication is much more frequent than nucleotidic substitutions (and makes it more difficult to obtain surely homologous sites in a column of a multiple alignment), approaches of type (1) and (2) are unadapted to the detection of recombination in minisatellite. Thus, we turn ourselves towards a distance based method. Here, we present a first algorithm to detect if an allele is a recombinant and to propose which alleles could be its parents alleles. For this we adapt the phylogenetic profile method[Weiller 1998] to minisatellite data and combine it with the multiple alignment procedure to detect putative recombination positions in a map. We have implemented our method in a program dubbed *MS_PhylPro* and tested it on different recombinant maps of the INS minisatellite provided in [Stead and Jeffreys 2000].

The sequel of the paper is organized as follows: in Section 2 we summarize the phylogenetic profile approach and an improvement on the distance criterion used. In Section 3, we explain *MS_PhylPro*, and while the tests on real data are described in Section 4. A conclusion and some perspectives are discussed in Section 5.

2. DETECTION OF RECOMBINATION USING PHYLOGENETIC PROFILES

2.1 A brief discussion of PhylPro

We describe the phylogenetic profile method in general terms, for any type of sequence.

One is given a test sequence t and a set of possible parent sequences P. Once t and set P are aligned using any multiple alignment algorithm (*e.g. clustalw*), the goal is to find positions in the test sequence that are recombination points, that is, the sequence on the left (left sequence) comes from a different parent other than the sequence on the right (right sequence) comes from. For a given position, the idea is to test the sequence similarity profile between t and each s in P on windows at the left- and at the right of the current position. The sequence similarity profile consists in a vector of distance (*e.g.*, alignment distance) between t and any $s \in P$. If the vector of distances on the right window correlate well with the one of the left windows, then the corresponding region of t probably comes from the same parents and the current position is not a recombination position. If on the contrary the correlation is low, it might be position where the cross-over took place. So *PhylPro* evaluates this test for each position of the test sequence and searches for position where the vectors correlation is minimum. Of course, it remains difficult to choose appropriate window size, distance measure, and correlation coefficient (see [Weiller 1998; Posada 2002; Etherington et al. 2005] for discussions on these questions). The window size is automatically set to the total number of variable sites divided by 1.5.

This original method provides correlation coefficient as a score of the propensity of a position to be a recombination junction, but lacks a significance measure. An empirical evaluation of the significance of the correlation coefficient was proposed and implemented in [Posada 2002] as follows. A user-defined number n (in this work, we use $n = 1000$, unless otherwise noted) of randomized alignments are computed by shuffling the columns of the multiple alignment. Then, one computes a P-value for the null hypothesis of no recombination, taken as the number of times the minimum distance vector correlation was smaller than the ones computed from the randomized alignments. The location of such minimum distance vector correlation likely is a recombination junction.

2.2 Influence of the distance

In the case of non-repetitive sequences, the use of different sensitive distance measures is mentioned, but not studied in depth in [Weiller 1998]. Here, we show the effect of using two distance measures on the resulting profiles on 24 nucleotide data sets of [Posada 2002]. In half of them, recombination is assumed to be absent, while for the other 12, literature has suggested the presence of recombination. The first classical distance simply distinguish between identity and mismatch, while the second differentiate mismatches in transitions and transversions.

TableI in Appendix summarizes the results. The use of the transition/transversion scoring matrix instead of the identity matrix used in [Weiller 1998] implementation of Posada[Posada 2002] improves the sensitivity of phylpro to identify some likely recombinant data sets DmelCytB, WolfCR and Armillaria-mtDNA. Interestingly, no other tools identify recombinant likelihood of DmelCytB data set.

The results between the two distances disagree and as *PhylPro* is conservative in recombination detection, we suggest that both measures could be used to complement each other.

To validate the results obtained above, we apply *PhylPro* that uses identity matrix and the one that uses the transition/transversion scoring matrix on simulated datasets obtained using a program, named *ms*(see [Hudson 2002] and its supplementary material (user manual)). This will allow us to evaluate the power and false positives of the two distance measures applied above using *PhylPro*.

The program *ms* can generate many independent replicate samples under a variety of neutral models about migration, recombination rate and population size. In this program, the standard coalescent approach is used to generate our samples. The random genealogy of the sample is first generated and then mutations are randomly place on the genealogy.

In this work, the sequences representing our simulated data, using appropriate parameters (we will give them later), were evolved under the Hasegawa-Kinshino-Yano model[Hasegawa et al. 1985] of evolution using the gene trees output of *ms* that represent the history of the sampled chromosomes. We program used for this purpose is *seq-gen* of Rambaut and Grassly[Rambaut and Grassly 1997]. The parameters used in the simulation include Θ: the neutral mutation rate per site (otherwise known as the nucleotide diversity), ρ: the levels of recombination and α: the strength of rate variation among sites. Note that when $\alpha = \infty$, there is no rate variation among sites. And the smaller the α, the stronger the rate variation. Furthermore, for $\rho = 1$, note that 10% of the datasets simulated contain no recombination events.

TableII of the appendix contains the power and rate of false positives of *PhylPro* under the identity and the transition/transversion (TT) matrices respectively. We observed that the results obtained under the identity matrix i similar to that of the transition/transversion matrix except that TT is conservation. We can therefore conclude that its greater power experienced in the Posada biological datasets above is not due to greater false positive rate.

3. A NEW ALGORITHM FOR DETECTING RECOMBINATIONS IN MINISATELLITES

3.1 Adaptation of *PhylPro* to minisatellites

In *PhylPro*, the input is a multiple alignment of the sequences and the recombination test is performed individually on each sequence of the alignment, which is then compared to the remaining set.

In *MS_PhylPro*, the input consists of a test allele t, and a multiple alignment M of a cluster of maps P that are possible progenitors of t. Such a cluster contains sequences that can be the progenitor of only one part of t. The recombination test using the phylogenetic profile is applied to the test allele t, and t is compared with the sequences in the multiple alignment. To compute the distances between t and each s in P, we first align t with M using *MS_Alimul*. Indeed, the progressive strategy implies that *MS_Alimul* uses a procedure to align any two multiple alignment, even in the case where one alignment is a single sequence. We use this procedure of *MS_Alimul*. The distance in a window between two maps in the alignment is simply the number of columns in which the character differ for the two maps (of course, here nucleotidic transversion/ transition cannot be accounted for). The multiple alignment M of the cluster P is obtained using *MS_Alimul*. The window size W is set automatically as in *PhylPro*. This algorithm is summarized in Figure 2.

Another adaptation was made to *PhylPro*. In the former, sites of the alignment that are too polymorphic caused the recombinant detection to be less reliable. The reason is that such site may have undergone several substitutions in one sequence and are often badly aligned. Thus, *PhylPro* also includes a procedure to detect such sites and removed them from the alignment prior to the recombinant detection. In our adaptation of *PhylPro*, we included this as an option. We show in Section 4 that removing polymorphic sites in minisatellite maps distort the detection of recombination in datasets that truly contain recombinants.

3.2 Progenitor selection

We have designed another procedure for the progenitor selection to accelerate the prediction of parent alleles. We use the tree attribute 'depth' to determine how representatives of a class are automatically selected as possible progenitors. Graphically, if we want to use depth equal one to select a representative for a class, then any of sequences $S1$ and $S2$ and sequences $S3$ and $S4$ of the following guide tree (fig. 1), can be picked to represent each pair. For example, we can select $S1$

and $S3$. So that the dataset containing sequences $S1$, $S2$, $S3$, $S4$, and $S5$ can optimally be represented (for example) by the sequences $S1$, $S3$, and $S5$. And progressively, if we prefer depth equal to two, the class will be represented (for example) by the sequences $S1$ and $S5$.

This automated selection can be done during a depth-first traversal by evaluating the depth of the resulting left subtree. If the depth of the resulting subtree is equal to value of the depth of interest, one of the sequences in this subtree will be randomly selected to represent the subtree and the selection procedure back-track in a depth-first way to find representives for the other subtrees of the guide tree. Let us call this selection procedure, *cut and find current depth* (henceforth, (CF_CD(T, depth))), where T is the guide tree. Procedure CF_CD will return set P, which is the set of potential progenitors. Possible range of values of depth that can be use to determine P is $0 < depth < O(\log l)$, where l is the total number of alleles (leaves of the guide tree T).

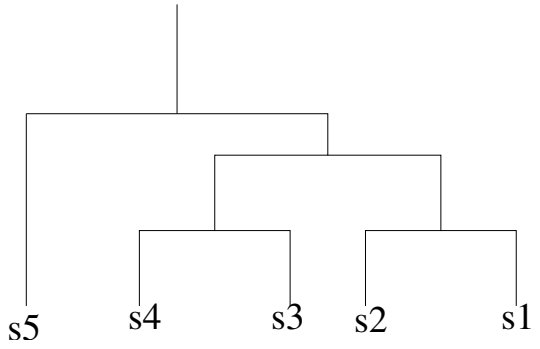

Figure 1. A guide tree for sequences S1, S2, S3, S4, and S5.

Using the foregone details, we design an algorithm *MS_PhylPro* to detect recombinant in minisatellites using MS_ALIMUL and PhylPro. This algorithm is encapsulated in Figure 2. Given is a class C (the class naturally is made of aligned minisatellite sequences), a guide tree T , and an test allele t that needs explanation on its recombinant candidacy status. The variable A contains either *yes* or *no* to indicate if t is a recombinant or not.

1. *MS_PhylPro*(C, t, $depth$)
2. $T \leftarrow MS\_ALIMUL(C, t)$
3. $P \leftarrow CF\_CD(T, depth)$
4. P-value←PhylPro(P, t)
5. A ←Is-Significant(P-value)

Figure 2. The algorithm *MS_PhylPro*.

4. IMPLEMENTATION AND EXPERIMENTATION

The above algorithm *MS_PhylPro* (with depth of interest equal 1) has been implemented in C and tested on eleven (11) datasets using a PC with Pentium IV CPU and 512MB RAM. Our datasets are collected from the set of manually determined recombinants of Stead and Jeffreys [Stead and Jeffreys 2000] (in fig. 5 of their paper). Briefly, these datasets, obtained from the insulin minisatellite, consist of dispersion patterns of A-, B-, C-, E-, F- and H-type repeats plus o-type repeats (unamplifiable repeats due to additional unknown variants). Note that minisatellite variant repeat mapping by PCR (MVR-PCR)[Jeffreys et al. 1991b] system equipped with the capacity to detect these repeat variants was used to derived this insulin minisatellite.

We first align the whole set of assumed non recombinant alleles and build a Neighbor Joining tree from the resulting pairwise distances for this sake. In this tree, the four classes of alleles defined by Stead and Jeffreys in [Stead and Jeffreys 2000] appear as the four main clusters (each cluster being monophyletic). We could then use these clusters as sets of potential progenitor. On these 11 datasets, *MS_PhylPro* was always able to detect recombination in the recombinant maps. In table III of the appendix, we show for each dataset, the size of the test allele t (recombinant) and set M (possible progenitors of t) in term of the number of the dispersion patterns (repeat variants) contained.

First, we tested which influence has the option that removes polymorphic sites from the input multiple alignment. Table IV in Appendix summarizes the comparison with (2nd column) and without (3rd column) this option. Recombination was never detected with removal of polymorphic sites and always without. It clearly shows that this option does not apply for minisatellites and its finds application only in DNA sequences.

In a second experiments, we have tested *MS_PhylPro* with procedure for selection of progenitor. In Table V in Appendix, we list the results; these illustrate the effect of the procedure $P \leftarrow CF\_CD(T, depth)$ that is used to automatically select the possible progenitors for any given test sequence. The first two columns show the results and running times of *MS_PhylPro* without this procedure and the last two columns with it. One observes that it improves the running time without sacrificing the sensitivity of *MS_PhylPro* for detecting recombination.

5. CONCLUSION AND FUTURE WORK

We provide the first program *MS_PhylPro* to detect recombinant sequences among a set of minisatellite data. Our approach take advantage of the multiple alignment program *MS_Alimul* and of an adapted phylogenetic profile method for recombinant detection. This tool was tested on several real datasets and shown to give positive results.

We wish to test thoroughly the sensitivity *MS_PhylPro* on different minisatellite data like the Mouse minisatellite [Bois and Jeffreys 1999] and some hypervariable human minisatellites [Buard and Vergnaud 1994]. Further works include the implementation of other automatic determination of progenitor sets, and dealing with multiple recombination in a single sequence. We also hope to extend *MS_PhylPro* to estimate the amount of recombination in a given data set.

ACKNOWLEDGMENTS

We thank Weiller, Posada and Etherington for useful discussions. We are grateful to Posada for sending us his C implementation of *PhylPro*. Part of this work was done while the first author was at LIRMM, Montpellier, France on a CNRS-NEPAD Bioinformatics Initiative special grant from the CNRS.

REFERENCES

BÉRARD, S., NICOLAS, F., BUARD, J., GASCUEL, O., AND RIVALS, E. 2006. Fast and specific alignment method for minisatellite maps. *Evolutionary Bioinformatics Online (in press)*, x–x.

BÉRARD, S. AND RIVALS, E. 2003. Comparison of minisatellites. *Journal of Comp. Biol. 10,* 3-4, 357–372.

BOIS, P. AND JEFFREYS, A. 1999. Minisatellite instability and germline mutation. *Cellular and Molecular Life Sciences 55,* 12, 1636–1648.

BUARD, J. AND VERGNAUD, G. 1994. Complex recombination events at the hypermutable minisatellite ceb1 (d2s90). *Embo J.* 13, 3203–10.

ETHERINGTON, G., DICKS, J., AND ROBERTS, I. 2005. Recombination analysis tool (rat): a program for the high-throughput detection of recombination. *Bioinformatics 21,* 3, 278–281.

HASEGAWA, M., KISHINO, K., AND YANO, T. 1985. *J. Mol. Evol.* 22, 160–174.

HUDSON, R. R. 2002. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338.

JEFFREYS, A. ET AL. 1985. Individual-specific 'fingerprints' of human dna. *Nature* 316, 76–79.

JEFFREYS, A. ET AL. 1991a. Minisatellite repeat coding as a digital approach to dna typing. *Nature* 354, 204–209.

JEFFREYS, A. J. ET AL. 1991b. Minisatellite repeat coding as a digital approach to dna typing. *Nature* 354, 204–209.

KECECIOGLU, J. D. AND ZHANG, W. 1998. Aligning alignments. In Proceedings of the 9th Symposium on Combinatorial Pattern Matching. *Lecture Notes in Computer Science* 1448, 189–208.

POSADA, D. 2002. Evaluation of methods for detecting recombination from dna sequences: Empirical data. *Mol. Biol. Evol. 19,* 5, 708–717.

RAMBAUT, A. AND GRASSLY, N. C. 1997. Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.,* 13, 235–238.

RIVALS, E. Multiple alignment of minisatellite maps: Ms alimul (cluster). *In preparation*.

SAMMETH, M., WENIGER, T., HARMSEN, D., AND STOYE, J. 2005. Alignment of tandem repeats with excision, duplication, substitution and indels (edsi). *Workshop on Algorithm in Bioinformatics (WABI)*.

STEAD, J. D. H. AND JEFFREYS, A. J. 2000. Allele diversity and germline mutation at the insulin minisatelliete. *Human Molecular Genetics 9,* 5, 713–723.

WEILLER, F. W. 1998. Phylogenetic profiles: A graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol. 15,* 3, 326–335.

A. APPENDIX A

Data	Identity matrix	Transition/Transversion matrix
	P-value	*P*-value
DmelCytB	0.8440(N)	0.0490(Y)
FusariumTri101	0.9020(N)	0.7840(N)
HumanHRVI	0.8450(N)	0.1070(N)
GymnND4	0.8930(N)	0.6320(N)
DaphniaC01	0.5950(N)	0.8300(N)
MammPGK	0.2700(N)	1.0000(N)
WolfCR	0.3120(N)	0.0050(Y)
MammPDH	1.000(N)	0.9900(N)
Armillaria-mtDNA	0.4470(N)	0.0830(Y)
InsectaC011	0.7760(N)	0.2970(N)
Perom12S	0.9900(N)	0.5200(N)
VertebC01	1.0000(N)	0.2770(N)
Candidula 16S	0.0330(Y)	0.1840(N)
Petunias-RNase	0.4250(N)	0.4360(N)
HIV(B)EnvNR	0.7010	0.8990(N)
HIVEnvNR	0.0030(Y)	0.1900(N)
Maiz ACT	0.2760(N)	0.1420(N)
NeisseriaArgF	0.0000(Y)	0.0700(Y)
BoletalesATP6	0.0000(Y)	0.2940(N)
HGVgenome	0.0520(Y)	0.8910(N)
HumanDRB1	0.0580(Y)	0.9860(N)
Fusaruim3	0.0000(Y)	0.3860(N)
Candida-mtDNA	0.0000(Y)	0.0000(Y)
HIVEnv	0.0000(Y)	0.3540(N)

Table I. Comparison of recombination detection with two distance measures on classical nucleotidic data sets from [Posada 2002]. The second column shows the result obtained in [Posada 2002], while the third indicated the *P*-values obtained for using transition/transversion matrix. Data sets in plain font indicate recombination was assumed to be absent, while those in bold font have being shown to contained recombinant sequence(s). The *P*-value indicate recombination was detected if less than 0.05. But marginally significant, if $0.05 > P$-value > 0.01 and marginally nonsignificant if $0.10 > P$-value> 0.05. We indicate the present of recombination significantly, marginally significantly or insignificantly using a Yes (*Y*) and non-present of it using a No (*N*).

	POWER	
	Identity matrix	Transition/Transversion matrix
Parameters	*P*-value	*P*-value
$\Theta = 10$, $\alpha = \infty$, $\rho = 1$	0(Y)	0.66(N)
$\Theta = 10$, $\alpha = \infty$, $\rho = 4$	0(Y)	0.30(N)
$\Theta = 10$, $\alpha = \infty$, $\rho = 10$	0(Y)	0.02(Y)
$\Theta = 200$, $\alpha = \infty$, $\rho = 1$	0(Y)	0.08(N)
$\Theta = 200$, $\alpha = \infty$, $\rho = 4$	0(Y)	0.88(N)
$\Theta = 200$, $\alpha = \infty$, $\rho = 10$	0(Y)	0.86(N)
	FALSE POSITIVES	
	Identity matrix	Transition/Transversion matrix
Parameters	*P*-value	*P*-value
$\Theta = 10$, $\rho = 0$, $\alpha = \infty$	1(N)	0.96(N)
$\Theta = 10$, $\rho = 0$, $\alpha = 2$	1(N)	0.44(N)
$\Theta = 10$, $\rho = 0$, $\alpha = 0.5$	1(N)	0.72(N)
$\Theta = 10$, $\rho = 0$, $\alpha = 0.05$	0.02(Y)	0 (Y)
$\Theta = 50$, $\rho = 0$, $\alpha = \infty$	0.44(N)	0.6(N)
$\Theta = 50$, $\rho = 0$, $\alpha = 2$	0.34(N)	0.62(N)
$\Theta = 50$, $\rho = 0$, $\alpha = 0.05$	0.7(N)	0.5(N)

Table II. Power evaluation is shown in the upper part of the table, while the lower part contains the rate of false positives. In the results above, we remove polymorphic and due to the time limitation (before submitting the final version of this paper), number of permutations considered was 50. Recall that Θ is nucleotide diversity, α is rate variation among sites and ρ is the levels of recombinations in the simulated datasets. And when $\alpha = \infty$, there is no rate variation among sites and the smaller the α, the stronger the rate variation. Furthermore, for $\rho = 1$, note that 10% of the datasets simulated contain no recombination events. The *P*-value indicate recombination was detected if less that 0.05. But marginally significant, if $0.05 > P$-value> 0.01 and marginally nonsignificant if $0.10 > P$-value > 0.05. We indicate the present of recombination significantly, marginally significantly with a Yes (Y) and recombination insignificantly using a No (N).

Data	allele t	Set M
	Total size	# allele/Total size
Dataset 1	46	36/1524
Dataset 2	46	113/12336
Dataset 3	66	50/3433
Dataset 4	58	50/3433
Dataset 5	56	50/3433
Dataset 6	56	50/3433
Dataset 7	54	50/3433
Dataset 8	51	50/3433
Dataset 9	48	50/3433
Dataset 10	47	50/3433
Dataset 11	46	50/3433

Table III. The second column shows the size of each dataset in term of the number of the dispersion patterns/repeat variants, each dataset contained, while the third column shows two values. The first value is the number of alleles undergoing consideration as possible progenitors for allele t and the second value shows total size of these allele in term of repeat variants, they contained.

Data	Remove Polymorphic sites P-value	Don't remove Polymorphic sites P-value
Dataset 1	0.9440(N) ($W = 6$)	0.0020(Y) ($W = 34$)
Dataset 2	0.1330(N) ($W = 10$)	0.0000(Y) ($W = 114$)
Dataset 3	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 4	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 5	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 6	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 7	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 8	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 9	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 10	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)
Dataset 11	1.0000(N) ($W = 2$)	0.0000(Y) ($W = 100$)

Table IV. The P-value indicate recombination was detected if less than 0.05. But marginally significant, if $0.05 > P$-value> 0.01 and marginally non-significant if $0.10 > P$-value > 0.05. We indicate the present of recombination significantly, marginally significantly with a Yes (Y) and recombination insignificantly using a No (N). The width of the window used is denoted with W in columns 2 and 3 above.

	Without	CF_CD(T,depth)	Using	CF_CD(T,depth)
Data	P-value	Run time(sec)	P-value	Run time(sec)
Dataset 1	0.0230(Y)	10.92	0.0170(Y)	4.23
Dataset 2	0.0000(Y)	448.72	0.0000(Y)	190.11
Dataset 3	0.0000(Y)	76.44	0.0000(Y)	34.25
Dataset 4	0.0000(Y)	74.90	0.0000(Y)	33.72
Dataset 5	0.0000(Y)	79.33	0.0000(Y)	33.79
Dataset 6	0.0000(Y)	81.89	0.0000(Y)	34.20
Dataset 7	0.0000(Y)	88.15	0.0000(Y)	34.32
Dataset 8	0.0000(Y)	76.85	0.0000(Y)	34.22
Dataset 9	0.0000(Y)	75.24	0.0000(Y)	33.05
Dataset 10	0.0000(Y)	75.82	0.0000(Y)	33.04
Dataset 11	0.0000(Y)	84.80	0.0000(Y)	34.28

Table V. In the results above, we do not remove polymorphic. The P-value indicate recombination was detected if less than 0.05. But marginally significant, if $0.05 > P$-value> 0.01 and marginally nonsignificant if $0.10 > P$-value > 0.05. We indicate the present of recombination significantly, marginally significantly with a Yes (Y) and recombination insignificantly using a No (N).