WAMI: a web server for the analysis of minisatellite maps

Mohamed Abouelhoda*1,2, Mohamed El-Kalioby1 and Robert Giegerich3

Abstract

Background: Minisatellites are genomic loci composed of tandem arrays of short repetitive DNA segments. A minisatellite map is a sequence of symbols that represents the tandem repeat array such that the set of symbols is in one-to-one correspondence with the set of distinct repeats. Due to variations in repeat type and organization as well as copy number, the minisatellite maps have been widely used in forensic and population studies. In either domain, researchers need to compare the set of maps to each other, to build phylogenetic trees, to spot structural variations, and to study duplication dynamics. Efficient algorithms for these tasks are required to carry them out reliably and in reasonable time.

Results: In this paper we present WAMI, a web-server for the analysis of minisatellite maps. It performs the above mentioned computational tasks using efficient algorithms that take the model of map evolution into account. The WAMI interface is easy to use and the results of each analysis task are visualized.

Conclusions: To the best of our knowledge, WAMI is the first server providing all these computational facilities to the minisatellite community. The WAMI web-interface and the source code of the underlying programs are available at http://www.nubios.nileu.edu.eg/tools/wami.

Background

Minisatellite maps

A DNA region is categorized as a minisatellite locus if it is composed of tandemly repeated DNA stretches and spans more than 500 bp. Each of these stretches is called a unit and it holds (by most definitions) 10-100 bp. The units are not necessarily identical due to point mutations, and their number and organization may vary among individuals as a result of subsequent evolutionary events. The variant repeat mapping by PCR (MVR-PCR) is a popular technique to reveal the structure of a minisatellite locus as it enables unit typing and minisatellite map production. Unit typing is the classification of the variable units into distinct types (called variants and denoted by different symbols) according to their DNA sequences. A minisatellite map is a compact representation of the minisatellite locus, where each unit is replaced with the respective symbol. Figure 1(a) shows an example of a minisatellite locus and the respective map.

Applications of minisatellite map analysis

Minisatellite maps have manifold applications in forensics and population studies. Foster et al. [1] used minisatellite maps to resolve the dispute on the fatherhood of President Jefferson to a son of his slave. They showed that Jefferson is the biological father of her last son, but not her first son as thought before. Based on the MS205 dataset, Armour et al. [2] confirmed the African origin for modern humans, Alonso et al. [3] proved a European affiliation for the Basques, and Rogers et al. [4] dated the Eurasian population as 52000-66000 years and the oldest European as 37600-56200 years. Using the MSY1 dataset, which was first investigated by Jobling et al. [5], Brión et al. [6] showed that European lineages are more similar than North African ones. Bonhomme et al. [7] used minisatellites to study house mouse population and provided a migration map for them. Very recently, Yuan et al. [8] used the MS32 minisatellites to study the population specificity among Thai, Chinese, and Japanese. They showed that the MS32 minisatellite is an effective tool to distinguish individuals from these populations.

The functional and medical roles of minisatellites have also been addressed in many studies for the last two
there is currently just one server, called MS_ALIGN (http://www.atgc-montpellier.fr/ms_align/) for minisatellite map comparison [17]. It is, however, limited to computing all pair-wise alignments, with no post-processing and visualization of map alignments.

In this paper, we present the web server WAMI for the analysis of minisatellite maps. The server uses a recent algorithm for map alignment, improved over the one in MS_ALIGN, and provides a workflow for the execution of the four computational tasks mentioned above, including visualization. These capabilities are demonstrated here by the analysis of the MSY1 [19] and MS205 [2-4,20] datasets.

Implementation
Model of minisatellite map evolution and alignment
Minisatellite maps can be studied in an independent or a comparative fashion. In the former, a map is analyzed to identify the evolutionary history that gave rise to the observed sequence of units. In the latter, two maps are aligned together to figure out regions of common and individual evolution histories. However, both tasks are entangled, since a region of individual evolution, juxtaposed to a gap in the map alignment, must have a plausible individual history. This makes minisatellite map alignment algorithmically more challenging than the standard sequence alignment problem.

Map evolution
Our evolutionary model of minisatellite maps includes the following operations acting on the unit level:

- **Unit mutation**: This is the change of one unit type into another. For example, the unit b in the map abdc mutates into e leading to the map acde. The unit mutation is a consequence of point mutations (substitution and indels) acting on the DNA sequence of the units. In the example of Figure 1, the differences between the three unit types are attributed to nucleotide substitutions.
- **Duplication**: Duplication (also known as expansion or amplification) is the generation of new copies of the units by tandem duplications. Replication slippage, reciprocal exchange (unequal crossover or unequal sister chromatid exchange), and gene conversion (including synthesis-dependent strand annealing, abbreviated by SDSA) are potential mechanisms for unit duplications. The first is suggested for short segments while the others are for long ones; see [21-25] for more details. Figure 1(b) illustrates the unequal cross over mechanism, where the paired homologous chromosomes exchange unequal segments during the cell division. This results in the duplication of the unit b in one chromosome and the deletion (contraction) of it in the other. The single-copy duplication model assumes that one unit can

| Model of minisatellite map evolution and alignment |
|--------------------------------------------------|
| Minisatellite maps can be studied in an independent or a comparable fashion. In the former, a map is analyzed to identify the evolutionary history that gave rise to the observed sequence of units. In the latter, two maps are aligned together to figure out regions of common and individual evolution histories. However, both tasks are entangled, since a region of individual evolution, juxtaposed to a gap in the map alignment, must have a plausible individual history. This makes minisatellite map alignment algorithmically more challenging than the standard sequence alignment problem. |

**Map evolution**
Our evolutionary model of minisatellite maps includes the following operations acting on the unit level:

- **Unit mutation**: This is the change of one unit type into another. For example, the unit b in the map abdc mutates into e leading to the map acde. The unit mutation is a consequence of point mutations (substitution and indels) acting on the DNA sequence of the units. In the example of Figure 1, the differences between the three unit types are attributed to nucleotide substitutions.
- **Duplication**: Duplication (also known as expansion or amplification) is the generation of new copies of the units by tandem duplications. Replication slippage, reciprocal exchange (unequal crossover or unequal sister chromatid exchange), and gene conversion (including synthesis-dependent strand annealing, abbreviated by SDSA) are potential mechanisms for unit duplications. The first is suggested for short segments while the others are for long ones; see [21-25] for more details. Figure 1(b) illustrates the unequal cross over mechanism, where the paired homologous chromosomes exchange unequal segments during the cell division. This results in the duplication of the unit b in one chromosome and the deletion (contraction) of it in the other. The single-copy duplication model assumes that one unit can
duplicate at a time while the multiple-copy duplication model assumes that multiple adjacent units can duplicate at a time. For example, the adjacent units bc in the map abbc can duplicate in one event, leading to the map abbcabc.

- **Insertion/Deletion**: Insertion is the appearance of unit types, possibly due to errors or translocation events. For example, insertion of unit z in the map abc leads to the map abcabc. A dual operation to insertion is deletion where one unit disappears, leading also to map contraction. Potential mechanisms for these events include the ones mentioned above except for replication slippage.

Each of these operations is assigned a cost to reflect the relative rate at which it occurs in nature. The cost of a unit mutation is proportional to the Hamming/edit distance between the nucleotide sequences of the units. We write $d_M(x,y)$ to denote this cost between two units $x$ and $y$. (Of course, $d_M(x,y) = 0$ if $x = y$.) In Figure 1, $d_M(a,b) = 1$ because of one mismatch at the last nucleotide, and $d_M(a,c) = 2$ because of mismatches at the fourth and the last nucleotide. The costs of duplication, insertion, and deletion are arbitrary and usually chosen such that the duplication is less than the mutation, deletion and insertion cost.

**Reconstruction of evolutionary history**

The evolutionary history of a map is the series of evolutionary operations leading to the observed sequence of units. This history is also called duplication history, because the duplication is the main event contributing to map evolution. The cost of a duplication history is the total cost of the occurring operations. An optimal (most parsimonious) history is one with the minimal cost. For example, one history of the map bceaceabbc originated from the leftmost unit b is as follows: The leftmost unit b duplicated three times to the right leading to the sub-map bbbbb. Then the second b mutates into c leading to the sub-map bcbbb. The unit c duplicates two times to the right producing the sub-map bcecbbc. The second c mutates into a and the last c duplicates once again to the right leading to the final observed map. Assuming that $d_M(a,b) = d_M(b,c)$ and $d_M(a,b) < d_M(a,c)$, we leave it as an exercise for the reader to verify that this scenario is indeed an optimal one.

**Map alignment**

The alignment of the minisatellite maps includes the operations of replacement (match/mismatch, where mismatch corresponding to mutation), free insertion/deletion (indel), and duplication. Given a cost for each operation, an optimal alignment is the one of minimum cost. An efficient algorithm for finding optimal map alignments is ARLEM [18].

An optimal map alignment delivers a three-stage scenario: The aligned units (match/mismatch) refer to common ancestors, the duplications refer to differences in the individual duplication histories, and the indels may refer to errors or translocations. Figure 2 (right), shows an alignment of two maps where the replaced (matched/mismatched) characters are put above each other and the units evolved by duplications are attached to arcs. In this representation, an arc connecting two identical units corresponds to a duplication event, and an arc connecting two different units corresponds to a duplication followed by a mutation. In this alignment, the sub-map bccacbcb has emerged as a result of duplication/mutation events from the leftmost unit b. This sub-map is the example given above in the duplication history, and no indels exist.

Biologically, the map alignment, compared to individual map analysis, provides clues about the timing and direction of map evolution as well as the type of operation. In the alignment, we can conclude that the replaced units had appeared before the units evolved by duplications because of inheritance. We can also conclude that the evolved regions emerged from the inherited units that occur either on its left or right side. Furthermore, if we know that one sequence is the ancestor of the other, then we can distinguish better between the loss and gain of units, i.e. contraction versus duplication and insertion versus deletion.

**Extension beyond the single-copy model**

For WAMI, we extended ARLEM with a simple heuristic algorithm to account for double-copy duplications, where at most two different units can duplicate at a time, e.g., bcTbcbc. The idea of our algorithm is to pre-process the map to locate each sub-array of units of the form "xyxyx...", where $x \neq y$. We then create a new type $X = xy$ and replace the units in this array with the new type to yield the array "XX...". The distance between the new type $X$ and each original type $z$ is the cost of optimal duplication history of $xy$ emerged from or contracted to $z$. The distance between two new types $X = xy$ and $X' = x'y'$ is the cost of aligning $xy$ and $x'y'$. Finally, the alignment algorithm of ARLEM runs on the transformed map using the new distances between the map units. In WAMI, the use of double copy model is optional, because the single copy duplication model is already sufficient for many data sets.

Computationally, it is infeasible to infer a history under the still more general, multiple copy duplication, model involving arbitrary number of copies [26].

**Four tasks supported by WAMI**

**Fast computation of pairwise map alignment**

The basic step in WAMI is the computation of all pairwise alignments of the input maps, uploaded or edited online in multi-fasta format. The user can use default
parameters (costs of each operation) or specify other ones through the use of a cost file uploaded to the server.

The map alignment model implemented by ARLEM allows that aligned units duplicate either to the left or to the right. For example, the sub-map $dd$ in the aligned lower sequence of Figure 2 (right) was originated from the inherited unit $c$ on the right, where $c$ duplicated to the left to produce $cc$, then the leftmost $c$ mutated into $d$ which eventually duplicated to the left to produce another $d$. Previous programs allowed duplication only in left-to-right direction, where such a scenario cannot be modeled. This leads to an alignment of higher cost. This symmetric feature is crucial for studying the direction of map elongation, discussed below.

**Phylogenetic tree construction**

WAMI uses the program BIONJ [27] to construct the phylogenetic tree from the pairwise distances computed by ARLEM. BIONJ is based on a neighbor joining algorithm. The program njplot [28] is then used to visualize the tree.

**Analysis of structural variation**

In studying structural variation, researchers try to identify highly variable regions of the map. Most previous studies showed that map extremities are more variable than other map regions, a phenomenon known as (bi)polar variability [2,5,20]. WAMI can automatically provide evidence of (bi)polarized variation for a given dataset based on a scramble (randomization) test.
The program ARLEM was augmented with an extra option that determines the location associated with half of the optimal score in the aligned maps. We call this location the pivot-point. The rationale of the pivot point is that if the variations were accumulated at one end, then the pivot-point would be shifted towards this end. The pivot-points are calculated for all pairwise alignments and normalized with respect to the respective sequence lengths. A histogram for the pivot-points is then generated. To qualitify the results, WAMI computes another histogram for a randomized dataset obtained by shuffling the units in each map of the input dataset. It is expected that the histogram for uniformly distributed unit types along the maps is close to the Gaussian distribution, centered around the value 0.5. WAMI produces a single plot containing the two histograms overlaid on each other. The results section contains examples of applying this procedure to MS205 and MSY1 datasets.

**Analysis of duplication dynamics**

Determining the direction in which the units duplicate is an interesting issue that can help in inferring the evolutionary processes and the source/origin unit of the map. For the MSY1 dataset, for example, Jobling et al. [5] conjectured that Type 4 (4a) is the source of the map and assumed that the units preferably duplicate in the 3' → 5' direction. WAMI has a procedure that can test this kind of hypothesis based on another scramble test.

In ARLEM, units are allowed to duplicate towards the left or towards the right to achieve the best alignment score, while accommodating the most parsimonious series of duplication events. We added an option to ARLEM to restrict the duplications to originate either from the leftmost or from the rightmost unit of a map interval with duplication events. For example, if only the option imposing left-to-right duplication origin was set, then the sub-map "dd" in the aligned lower sequence of Figure 2 (right) could not have been originated from the unit "c" on its right, leading probably to increased alignment cost under this restriction.

To detect directional bias, WAMI invokes ARLEM three times on the dataset: 1) with both duplication directions allowed, 2) with only left-to-right duplications allowed, and 3) with only right-to-left duplications allowed. The latter two cases tend to yield higher cost than the first, because the duplications may be forced to follow a non-parsimonious scenario. Then the number of alignments in the second invocation with cost higher than the optimal one (as determined by the first invocation) is counted. Let $E_i$ denote this number. The analogous number $E_j$ for the third invocation is also computed. A normalized value combining both figures $E_n = (E_i - E_j)/(E_i + E_j)$ is then computed. The idea is that if $E_j$ was different from $E_i$ and $E_j$ was small, then $E_n$ would converge to 1, and one could argue that the duplications in the right-to-left direction are almost sufficient to yield alignments close to the optimal ones. Hence, right-to-left duplications appear preferred in the evolution of the minisatellites at hand. To further validate the results, WAMI runs a scramble test and computes the normalized $E_n$ values for many random datasets, obtained by shuffling the map units. Finally, the $E_n$ values are summarized in a histogram and plotted along with a peak representing $E_n$ of the original dataset. For random datasets where duplications to the left and to the right occur in an equal rate, it is expected that the distribution of $E_n$ is close to the Gaussian distribution centered around the value zero.

The scramble test is compute intensive, because the map alignment phase is repeated many times over scrambled datasets of the same size as the original. To speed up the computation, we use an approximation technique. We reduce each map to its modular structure, which is the sequence of distinct units in the map. For example, the modular structure of the map aaabbc is abc. This is reasonable because transitions between unit types strongly indicate the direction of duplication. Because the modular structure is typically much shorter than the map, a significant speed up is achieved.

**User Interface**

WAMI has an easy to use and intuitive interface. The main web-page contains four examples to help the user format map data and cost file. (One example is about the real dataset for the President Jefferson's fatherhood case, mentioned above. Other two examples about some published maps of the MSY1 [4] and MS205 [2-4,20]. datasets.) Tool tips and a help menu are also provided. For sustainability of service, we attached a blog to the website, to collect user feedback and learn about new features requested by the community. A part of the main interface is shown in Figure 2 (left).

Upon job termination, the user is directed to the results page, where pairwise alignments are displayed and one can toggle between them, see Figure 2 (right). The duplication events within optimal alignments are represented by arcs. The images depicting the alignments are produced based on LaTeX. (The respective Tex files are included in the batch download). If the phylogeny option was chosen, the tree in Newick/JPEG/PDF format can be retrieved. The results of structural variations and duplication dynamics options are summarized and presented to the user in the form of histograms. For datasets larger than 50 sequences, the user is prompted to enter an email address to receive a notification when the job terminates. All these results can be downloaded as a compressed file.
Computational efficiency
The program ARLEM uses a highly optimized algorithm for map alignment. It is based on a compression technique to save redundant computations and its speed is not affected by any increase in the number of types. In [18], we reported that ARLEM is 18 to 24 times faster than the previously available algorithm MS_ALIGN, using real and artificial datasets. For further speed-up, the options for computing phylogeny, analyzing structure variations, and duplication dynamics run in parallel over a computer cluster of four nodes, where each node contains two Quadcore CPUs (2.5 GHz each) with 64 GB RAM.

Results and Discussion
The examples given in the sequel are based on the minisatellite datasets MSY1 [5] and MS205 [2-4,20]. The former dataset is composed of 345 maps and the number of distinct unit types is eight; the types are assigned the codes {0, 1, 1a, 2, 3, 3a, 4, 4a}. The latter dataset is composed of 653 maps of which 429 valid maps belong to haplotype C [4]. The number of distinct unit types is two and the types are assigned the codes {A,T}.

Alignment and phylogeny
Table 1 shows the running times for real and artificial datasets of varying sizes and for different scramble test parameters. The number of iterations is the number of random datasets analyzed for studying the directional bias based on the modular map structure. The number of iterations based on the non-modular structure is a multiplication of the alignment time. The time for constructing the phylogenetic trees is not shown in the table, because it is in the range of seconds, i.e., negligible compared to other steps. The alignment time of the MSY1 dataset is higher than that of the MS205 because the average length of the MSY1 maps is higher than that of the MS205. But in analysis of directional bias, MS205 takes more time because the average length of its modular structure is three times the one of MSY1 with much higher variability, and our approximation technique described above is less effective for MS205. (The average modular structure lengths is approximately 13 and 4 for the MS205 and MSY1, respectively.) The random datasets were generated such that each map has an average length of 80 units (minimum and maximum are 60 and 100 units, respectively) with average modular structure length of 12 units to simulate difficult scenarios.

Figure 3 shows two phylogenetic trees produced by WAMI for a subset of the MS205 and MSY1 datasets. In these trees, we see that individuals from the same population are clustered together, which is in accordance with published results [2,3,6].

| Task               | Dataset | Num. | Iteration | Time |
|--------------------|---------|------|-----------|------|
| Alignment          | MS205   | 91806| 1         | 2.5  |
|                    | MSY1    | 59340| 1         | 3.25 |
|                    | Random100 | 5000 | 1         | 0.3  |
|                    | Random200 | 20000| 1         | 1.3  |
|                    | Random400 | 80000| 1         | 5    |
| Structural variation | MS205 | 91806| 2         | 4    |
|                    | MSY1    | 59340| 2         | 6    |
|                    | Random100 | 5000 | 2         | 0.6  |
|                    | Random200 | 20000| 2         | 2.3  |
|                    | Random400 | 80000| 2         | 9    |
| Directional bias   | MSY1    | 59340| 10        | 0.3  |
|                    | MSY1    | 59340| 25        | 1    |
|                    | MSY1    | 59340| 50        | 1.8  |
|                    | MS205   | 91806| 10        | 2.3  |
|                    | MS205   | 91806| 25        | 5.5  |
|                    | MS205   | 91806| 50        | 11   |
|                    | Random100 | 5000 | 10        | 0.2  |
|                    | Random100 | 5000 | 25        | 0.5  |
|                    | Random100 | 5000 | 50        | 1    |
|                    | Random200 | 20000| 10        | 1    |
|                    | Random200 | 20000| 25        | 2.1  |
|                    | Random200 | 20000| 50        | 4.1  |
|                    | Random400 | 80000| 10        | 3.5  |
|                    | Random400 | 80000| 25        | 8    |
|                    | Random400 | 80000| 50        | 16   |

Running times in minutes on WAMI for the MSY1 and MS205 (haplotype C) datasets. The column titled “Dataset” contains the dataset used. “Random100”, “Random200”, and “Random400” are datasets with 100, 200, and 400 artificial maps, respectively. The column titled “Num.” contains the number of pairwise map alignments which need to be computed. The column titled “Iteration” includes the number of randomization steps (and hence increased data size) in the analysis of duplication dynamics. The number of iterations in the task of analyzing structural variation is 2 because it runs one time on the original dataset and one time on randomized dataset of the same size.

Structural variation
We applied WAMI to both datasets to investigate structural variation. When studying structure variation with MS205, Armour et al. [2,20] noticed polarized variability at the 3’ end, where most of the differences between the alleles (individual maps) accumulate at the 3’ end. Figure 4(a) and 4(b) shows the histograms of the pivot points obtained for the original MS205 dataset and a subset of it.
including haplotype C. It is clear that the histograms of the original datasets are biased to the right in comparison to that of a randomized datasets. This bias indicates polar variability towards the 3' end. These plots confirm the results obtained by Armour et al. [2]. (The presented results of MS205 are obtained using the double-copy option, but the results under the single-copy model are very similar.)

For the MSY1 dataset, lying on the Y chromosome, Jobling et al. [5] noticed high variability at the 5' end in contrast to the autosomal MS205 dataset, and they noticed also that Types 4 and 4a, existing almost solely at the 3’ side, causes another source of variation at this end. This suggests bi-polar variability of this dataset. For us it was interesting to see how WAMI can thus help in spotting not only polar but also bi-polar variability. Figure 4(c) shows our observations for the MSY1 dataset. The resulting histogram has peaks at both ends. This indicates that the variations are bi-polar. To further verify our procedure on the MSY1 dataset, we removed Type 4 and 4a from the 3’ end and repeated the experiment. Figure 4(d) shows biased histogram to the 5' end. That is, both extremities of the MSY1 maps are highly variable, and the unit types 4 and 4a already introduces another source of variation, verifying the observation of Jobling et al. [5].

**Duplication dynamics**

We used WAMI to study duplication dynamics with the MSY1 and MS205 datasets. Figure 5(left) shows the resulting histogram for MSY1. The peak value on the right shows \( E_n \) of the real dataset, where \( E_l = 876 \), and \( E_r = 0 \). It is clear that this value is far from the \( E_n \) values of the randomized datasets with expected equal rates of left-to-right and right-to-left duplications. That is, the plot indicates that left and right duplications do not contribute equally to the duplication history, and the units duplicate preferably in the direction 3’ T 5’, as conjectured by Jobling et al. [5]. In Figure 5 (right), we show the histogram for the MS205 dataset (haplotype C), which also shows directional bias, but this time towards the right ( \( E_l = 1940 \), and \( E_r = 13318 \)). These results for both datasets may indicate the existence of unknown (chromosome-specific) dynamic constraints governing the duplication...
of the minisatellite units. Hence, they call for further investigation.

**Conclusions**

In this paper, we presented WAMI, a web server for comprehensive analysis of minisatellites. The server provides many of the functionalities needed by researchers in this area. Future versions of the server are planned to provide data-mining functionalities for associating the map comparison results to other features, like age, ethnicity, or genetic markers on the chromosomes.

The algorithms of WAMI for minisatellite map analysis can also be used for comparing arrays for tandemly repeated units within proteins or gene sequences; the work of Rivals et al. [29] shows an example of this application. The alignment part of WAMI can also be used to compare parent/son microsatellite datasets, provided that the microsatellite units are mapped to symbols, in analogy to the unit typing step of minisatellites. In addition to its applications in parental tests, this comparison helps in studying the mutation rates in association with other map characteristics and helps in estimating parental ages. The work of Dupuy et al. [30] is an example of such studies.

In this paper, we rely on a map evolution model based on single- and double-copy duplications. In spite of the computational difficulty, it is still interesting to incorporate the multiple copy duplication model in map alignment, eventually through heuristic algorithms. Furthermore, it is also desirable to incorporate recently suggested evolutionary operations, such as boundary switch and modular structure change [31] appearing in some minisatellite datasets. These operations could be modeled by block exchange within the map, in an analogous way to the block exchange operation in genome rearrangement studies. But a practical solution to this problem is algorithmically challenging and remains a subject of future research.

**Availability and requirements**

**Project name:** WAMI: A Web Server for the Analysis of Minisatellite Maps.

**Project home page:** http://www.nubios.nileu.edu.eg/tools/wami.

**Operating system(s):** Platform independent (web server).

**Programming language:** Perl, C, Java script, JSF

**Other requirements:** Better viewed on the browsers FireFox, Internet Explorer 8 (IE8), Safari, and Opera. For local installation, Tomcat 6.0 or more, JDK 1.5 or more, Apache Ant 1.7 or more are needed.

**License:** Free for academics. Authorization license needed for commercial usage (Please contact the corresponding author for more details).

**Any restrictions to use by non-academics:** No restrictions.

**Authors’ contributions**

MA and RG contributed to theoretical developments which form the basis of WAMI. MA and MEK developed and tested the software. All authors wrote and approved the manuscript.

**Acknowledgements**

We would like to thank the anonymous reviewers for valuable comments that helped us improve this paper. This work was supported by a grant from ITIDA and Nile University. We also thank Mark Jobling and John Armour for providing us with the MSY1 and MS205 datasets, respectively.

**Author Details**

1. Center for Informatics Sciences, Nile University, Giza, Egypt. 2.Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt. 3.Practical Computer Science Department, Faculty of Technology, Bielefeld University, Bielefeld, Germany.

Received: 8 November 2009 Accepted: 6 June 2010
Published: 6 June 2010

**References**

1. Foster E, Jobling M, Taylor P, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C: Jefferson fathered slave’s last child. Nature 1998, 396:27-28.
2. Armour J, Anttinen T, May C, Vega E, Sajantila A, Kidd J, Bertranpetit J, Paabo S, Jeffreys A: Minisatellite diversity supports a recent African origin for modern humans. Nature Genetics 1996, 13:154-160.
3. Alonso S, Armour J: MS205 minisatellite diversity in Basques: evidence for a pre-neolithic component. Genome Research 1998, 8:1289-1298.
4. Rogers E, Shone A, Alonso S, May C, Armour J: Integrated analysis of sequence evolution and population history using hypervariable compound haplotypes. Human Molecular Genetics 2000, 9(18):2675-2681.
5. Jobling M, Bouzakri N, Taylor P: Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Human Molecular Genetics 1998, 7(4):643-653.
6. Brión M, Cao R, Salas A, Lareu M, Carucaico A: New method to measure minisatellite variant repeat variation in population genetic studies. American Journal of Human Biology 2002, 14:421-428.
7. Bonhomme F, Rivals E, Orth A, Grant G, Jeffreys A, Bois P: Species-wide distribution of highly polymorphic minisatellite markers supports past and present genetic exchanges among house mouse subspecies. Genome Biology 2007, 8:B90.
8. Yuan A, Tanaka A, Kasznyski R, Iino M, Okuno T, Tsuromaya T, Yamamoto T, Jeffreys A, Tamaki K: Minisatellite MS32 alleles show population specificity among Thai, Chinese, and Japanese. Molecular Evolution 2009, 68:126-133.
9. Thierry A, Bouchier C, Dujon B, Richard G: Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Research 2008, 36(18):5970-5982.
10. Vafiadis P, Bennett S, Todd J, Nadeau J, Grabs R, Goodyer C, Ruhlandting E, Paabo S, Jeffreys A: Parental age determination in fathers: the use of Dupuy et al. [30] is an example of such studies.
11. Foster E, Jobling M, Taylor P, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C: Jefferson fathered slave’s last child. Nature 1998, 396:27-28.
12. Armour J, Anttinen T, May C, Vega E, Sajantila A, Kidd J, Bertranpetit J, Paabo S, Jeffreys A: Minisatellite diversity supports a recent African origin for modern humans. Nature Genetics 1996, 13:154-160.
13. Alonso S, Armour J: MS205 minisatellite diversity in Basques: evidence for a pre-neolithic component. Genome Research 1998, 8:1289-1298.
14. Rogers E, Shone A, Alonso S, May C, Armour J: Integrated analysis of sequence evolution and population history using hypervariable compound haplotypes. Human Molecular Genetics 2000, 9(18):2675-2681.
15. Jobling M, Bouzakri N, Taylor P: Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Human Molecular Genetics 1998, 7(4):643-653.
16. Brión M, Cao R, Salas A, Lareu M, Carucaico A: New method to measure minisatellite variant repeat variation in population genetic studies. American Journal of Human Biology 2002, 14:421-428.
17. Bonhomme F, Rivals E, Orth A, Grant G, Jeffreys A, Bois P: Species-wide distribution of highly polymorphic minisatellite markers supports past and present genetic exchanges among house mouse subspecies. Genome Biology 2007, 8:B90.
18. Yuan A, Tanaka A, Kasznyski R, Iino M, Okuno T, Tsuromaya T, Yamamoto T, Jeffreys A, Tamaki K: Minisatellite MS32 alleles show population specificity among Thai, Chinese, and Japanese. Molecular Evolution 2009, 68:126-133.
19. Thierry A, Bouchier C, Dujon B, Richard G: Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Research 2008, 36(18):5970-5982.
20. Vafiadis P, Bennett S, Todd J, Nadeau J, Grabs R, Goodyer C, Ruhlandting E, Paabo S, Jeffreys A: Parental age determination in fathers: the use of Dupuy et al. [30] is an example of such studies.
21. Foster E, Jobling M, Taylor P, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C: Jefferson fathered slave’s last child. Nature 1998, 396:27-28.
22. Armour J, Anttinen T, May C, Vega E, Sajantila A, Kidd J, Bertranpetit J, Paabo S, Jeffreys A: Minisatellite diversity supports a recent African origin for modern humans. Nature Genetics 1996, 13:154-160.
23. Alonso S, Armour J: MS205 minisatellite diversity in Basques: evidence for a pre-neolithic component. Genome Research 1998, 8:1289-1298.
24. Rogers E, Shone A, Alonso S, May C, Armour J: Integrated analysis of sequence evolution and population history using hypervariable compound haplotypes. Human Molecular Genetics 2000, 9(18):2675-2681.
25. Jobling M, Bouzakri N, Taylor P: Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Human Molecular Genetics 1998, 7(4):643-653.
26. Brión M, Cao R, Salas A, Lareu M, Carucaico A: New method to measure minisatellite variant repeat variation in population genetic studies. American Journal of Human Biology 2002, 14:421-428.
27. Bonhomme F, Rivals E, Orth A, Grant G, Jeffreys A, Bois P: Species-wide distribution of highly polymorphic minisatellite markers supports past and present genetic exchanges among house mouse subspecies. Genome Biology 2007, 8:B90.
28. Yuan A, Tanaka A, Kasznyski R, Iino M, Okuno T, Tsuromaya T, Yamamoto T, Jeffreys A, Tamaki K: Minisatellite MS32 alleles show population specificity among Thai, Chinese, and Japanese. Molecular Evolution 2009, 68:126-133.
29. Thierry A, Bouchier C, Dujon B, Richard G: Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Research 2008, 36(18):5970-5982.
30. Vafiadis P, Bennett S, Todd J, Nadeau J, Grabs R, Goodyer C, Ruhlandting E, Paabo S, Jeffreys A: Parental age determination in fathers: the use of Dupuy et al. [30] is an example of such studies.
31. Foster E, Jobling M, Taylor P, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C: Jefferson fathered slave’s last child. Nature 1998, 396:27-28.
SMYD3 is a risk factor for human cancers. Nature Genetics 2005, 37(10):1104-1107.
15. Vergnaud G, Deneud F: Minisatellites: mutability and genome architecture. Genome Research 2000, 10(7):899-907.
16. Wikigenes: MeSH: Minisatellite repeats. [http://www.wikigenes.org/e/mesh/e/14291.html].
17. Bérand S, Nicolas F, Buard J, Gascuel O, Rivals E: A fast and specific alignment method for minisatellite maps. Evolutionary Bioinformatics 2006, 2:327-344.
18. Abouelhoda M, Giegerich R, Behzadi B, Steyaert J: Alignment of minisatellite maps based on run length encoding scheme. Journal of Bioinformatics and Computational Biology 2009, 7(2):287-308.
19. Andreassen R, Lundsted J, Olaisen B: Mutation at minisatellite locus DYF1S55: Allele length mutation rate is affected by age of progenitor. Electrophoresis 2002, 23(15):2377-2383.
20. Armour J, Harris P, Jeffreys A: Allelic diversity at minisatellite MS205 (D16S309): evidence for polarized variability. Human Molecular Genetics 1993, 2:1137-1145.
21. Richard G, Pâques F: Mini- and microsatellite expansions: the recombination connection. EMBO reports 2000, 1(2):122-126.
22. Strachan T, Read A: Human molecular genetics. Garland Science Third edition. 2003.
23. Jeffreys A, Tamaki K, MacLeod A, Monckton D, Neil D, Armour J: Complex gene conversion events in germline mutation at human minisatellites. Nature Genetics 1994, 6(2):136-145.
24. Bois P, Jeffreys A: Minisatellite instability and germline mutation. Cell Molecular Life Science 1999, 55(12):1636-1648.
25. May C, Jeffreys A, Armour J: Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). Human Molecular Genetics 1996, 5(11):1823-1833.
26. Sammeth M, Stoye J: Comparing tandem repeats with duplications and excisions of variable degree. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2006, 3(4):395-407.
27. Gascuel O: BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution 1997, 14:585-695.
28. Perrière G, Gouy M: WWW-query: An on-line retrieval system for biological sequence banks. Biochimie 1996, 78(5):367-369.
29. Rivals E, Bruyère C, Toffano-Hicche C, Lechertey A: Formation of the arabidopsis pentatricopeptide repeat family. Plant Physiology 2006, 141:825-839.
30. Dupuy B, Stenersen M, Engeland T, Olaisen B: Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Human Mutation 2004, 23(2):117-124.
31. Shanks M, May C, Dubrova V, Balaresque P, Rosser Z, Adams S, Jobling M: Complex germline and somatic mutation processes at a haploid human minisatellite shown by single-molecule analysis. Mutation Research 2008, 65(1-2):H6-53.