A TASTE OF YEAST MOBILOMICS

Giulia Menconi[1], Giovanni Battaglia[2], Roberto Grossi[2], Nadia Pisanti[2] and Roberto Marangoni[2,3]
[1]*Istituto Nazionale di Alta Matematica, Roma, Italia*
[2]*Dipartimento di Informatica, Universitá di Pisa, Pisa, Italia*
[3]*Istituto di Biofisica, CNR, Pisa, Italia*

Keywords: Mobile elements, Multiple genome comparison.

Abstract: *Mobilomics* calls for detecting all the mobile elements in a genome so as to understand their dynamic behavior. We devise and apply a method that extends a pairwise strain comparison tool for mobile genetic elements (MGE) inference, and perform experiments on a whole dataset of 39 complete genomes of as many yeast (*S.cerevisiae*) strains. We locate *a priori* all the MGEs regions that are annotated in the reference sequence at hand, and map all the putative MGEs in all the other (non-annotated) strains. Interestingly, evolutionary relation among the strains based on the presence/absence of candidate MGEs, turns out to be quite close to that inferred by classic phylogenetic methods based on SNPs analysis.

1 INTRODUCTION

The *mobilome* (Siefert, 2009) is the whole collection of the mobile genetic elements (MGEs) hosted in a genome. MGEs can vary in length, sequence content, and copy number. They behave as parasites as they replicate by exploiting the resources of the host (Kidwell and Lisch, 2001). They can even express their own genes, and by doing this they can destabilize the host organism, as the mutations induced by their jumps or replications can result in gene inactivation or modification. The relation between MGEs and the host genome is complex and still debated: there are also other (than the usual parasitic relation) kinds of interactions, such as direct competition or, at the opposite, cooperation towards synergizing MGEs and their host (see (Leonardo and Nuzhdin, 2002) for a detailed discussion on this subject).

Mobilomics, that is the task of investigating the mobilome, is a comprehensive approach for providing rapid and exhaustive methods and tools for the identification of all the mobilome elements in an organism, for tracking their movements (including replications and deletions) during evolution, and - as a long term goal - for developing dynamic models able to forecast the fate of the relations between the mobilome and the host genome.

In the literature, population genomists study the mobilome paying particular attention to the dynamics of the MGEs, while evolutionary biologists attempt to define the contribution of the mobilome in the evolution of the host organisms. Some evidence supporting the conjecture that the mobilome has a great impact on the fate of the host, has been found in all the living kingdom, from prokaryotes (Rankin et al., 2010) to higher eukaryotes (Koszul et al., 2004; Bennetzen, 2000; Bourque, 2009) including human (Brittten, 2010). The authors of (Menconi et al., 2011) designed a tool for finding the mobilome of two genomes by performing a suitable pairwise comparison and extracting and mapping the complement of the shared (thus assumed immotile) DNA. In this paper we devise a pipeline (Section 2.1) that actually extends the mobilome inference method for two-genomes to the case of a collection of genomes.

The method aligns homologous chromosomes and marks the non-homologous "island" surrounded by homologous sequences as putative mobile elements (PMEs) to indicate the possibility of an occurrence of an MGE *m*. This approach is prone to errors: on one hand, an MGE *m* that did not move within a small set of organisms, would not be marked this way; on the other hand, a chromosomal mutation uncorrelated with the mobilome could be marked as PME. By progressively extending the above alignment to other strains (or even organisms), as we do in this paper, the set of PMEs becomes more and more populated by all the MGEs that actually moved or replicated.

Furthermore, in this paper we show (Section 2.2) the results of the application of our approach to the

Menconi G., Battaglia G., Grossi R., Pisanti N. and Marangoni R..
A TASTE OF YEAST MOBILOMICS.
DOI: 10.5220/0003702302710274
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2012), pages 271-274
ISBN: 978-989-8425-90-4

whole data set of 39 strains of the yeast *Saccharomyces cerevisiae*, the genome sequences recently released by (Liti et al., 2009). They have a low coverage (one-to-fourfold), and so they are unannotated and rich of *unresolved* regions (i.e. sequences of unspecified bases). Our choice of studying the yeast is motivated by the large availability of its sequenced strains and by the observation that it is probably the most known organism from a molecular point of view. To have a referral point, we adopt the `S288C` strain, called `RefSeq` hereafter, as it is fully sequenced and annotated in the SGD database (SGD, 2010), along with its MGEs. We obtain a mapping of all the PMEs in all the strains (Section 2) that turns out to include all the annotated MGEs. We remark that this method to detect the mobilome makes use of sequence information only, can deal with whole chromosome at a time, and does not require any preprocessing of the data (like for example a partition of coding and non coding DNA fragments) and does not use any database information, and therefore it is independent from the type of organism it deals with. In particular, for example, the applicability of this tool are more general than those of mGenomeSubtractor (Shao et al., 2010) that is specifically for bacteria and requires a preprocessing of the data for the different purpose of detecting genetic variants.

Finally, we perform some mobilomics experiments by doing comparative analysis of the strains based on their PMEs. Interestingly, clustering the binary vectors obtained by marking the presence/absence of candidate MGEs in each of the strains provides an evolutionary relation among the strains that is quite close to that inferred by classic phylogenetic methods based on SNPs analysis (Section 3).

2 MOBILOME INFERENCE ON 39 STRAINS

In this section we explain (Subsection 2.1) how the method of REGENDER (Menconi et al., 2011) can be extended to be applied to data sets larger than two genomes, and we describe (Subsection 2.2) an application of the new method to the whole available data set of 39 yeast strains.

2.1 Finding the Mobilome in more than Two Genomes

The pipeline we devise in this paper is aimed at extracting the putative mobilome from a vast collection of genomes, and mapping the PMEs on each strain. This method does not require any template sequence for the sought MGEs and it can be applied to infer MGEs also for low coverage genomes with unspecified bases, where traditional approaches are largely ineffective.

REGENDER (Menconi et al., 2011) was designed for detecting mobile elements that could be inferred from the comparison of two genomes. It performs a two-phase processing of all the possible chromosomes' pairs: first, it finds the common n-grams and, second, it aggregates consecutive n-grams in a greedy fashion using some user-defined parameters that control when the next conserved region begins.

The extension to more than two strains is critical due to the high computational cost of the multiple alignment that should actually be performed; furthermore, the presence in all the input sequences (except `RefSeq`) of unresolved bases makes the whole task even more arduous. Taking advantage of the efficiency of REGENDER, we actually perform a progressive star-like multiple alignment centered at `RefSeq`. The main steps are the following, that are performed separately for each one of the 16 chromosomes:

1. REGENDER is applied to `RefSeq` against each one of the other 38 strains, in a progressive way.
2. Complement the output of Step 1 and find the genomic coordinates of non-conserved segments in all the strains of the collection.
3. Remove from the list the non-conserved segments (in any strains) which in `RefSeq` are in telomeric regions.
4. Remove very short indels/mutations: non-conserved segments shorter than 200 nucleotides in all strains.

Step 1 performs the simultaneous extraction of conserved regions from the whole collection, using the `RefSeq` chromosome as an outgroup to align all the others. Once the segment-based pairwise alignments between `RefSeq` and each other input chromosome have been computed by REGENDER, we only report the segments that are conserved in all the input chromosomes, by intersecting the conserved segments. This choice, in particular, implies that the result is independent from the order in which the strains are taken into account.

Step 2 filters out everything that is not conserved as this is presumed to be resident genome (i.e. not mobilome).

Step 3 was motivated by the fact that telomeres of any chromosome of any strain different from `RefSeq` mainly contain unresolved bases, because the presence of long repeats is a source of noise for the assembly phase.

Finally, Step 4 is due to the fact that very short indels or mutations are known not to be related to MGEs nor to chromosomal rearrangements.

2.2 Application to the Yeast Dataset

We applied the pipeline described in Section 2.1 to the whole dataset of 39 *S.cerevisiae* strains (Liti et al., 2009). The richness of the data sets gives a new - somehow more realistic - insight on the characterization of PMEs as actual MGEs, with respect to the case of complementing only the fragments of genomes that result immotile after a single pairwise comparison.

After numbering the 16 chromosomes by $N = 1,\ldots,16$, we take into account the 39 homologous chromosomes N, denoted $\texttt{Chr}N_1$, ..., $\texttt{Chr}N_k$, ..., $\texttt{Chr}N_{39}$ for as many as strains $k = 1,2,\ldots,39$ (where strain 1 is RefSeq). We mark a large set of PMEs, which vary in their length. To collect information about their real linkage with the MGEs, and also to deal with unspecified bases, we again refer to the accurate annotation available for RefSeq. We therefore map on RefSeq all the sequences that are detected as PMEs, and examine their possible annotations.

The MGEs in RefSeq are almost all LTR-retrotransposons, that we denote with Ty. Instead, we simply denote with LTR (*Long Terminal Repeats*) what is often called solo-LTR: the sequences of about 300b delimiting both ends of a LTR-retrotransposon. We distinguish PMEs basing on their length: PME-LTR candidates, having length 300b (compatible with LTR elements), and PME-Ty candidates, longer than 4000b (compatible with a complete Ty element).

Mapping the PME-LTR candidates to the annotated LTRs in RefSeq leads to an uncertain situation, since only about 44% of the known LTRs are actually marked as putative LTRs. This might have two motivations. First, the large amount of undetected LTR elements derives from the low probability that a LTR moves. Second, it is not rare to have a chromosomal mutation that spans from 300b to 4000b in a dataset of 39 strains, and this populates the class of putative LTRs that do not match LTR annotations. Therefore, the comparative genomics approach is ineffective for discovering LTRs, while motif-search based approaches might perform better.

The scenario for PME-Ty candidates is much different instead: we are able to detect 77 non-conserved regions longer than 4000b that are also in RefSeq. Our careful inspection pays a particular attention to the annotations involved in genomic mutations or rearrangements, apart from the MGE annotations already taken into account. In particular, we have considered: meiotic recombination hotspots (Gerton et al., 2000), evolutive and experimental breakpoints (Di Rienzi et al., 2010), autonomously replicating sequences (Di Rienzi et al., 2010), tRNA genes (SGD, 2010), γ-H2A rich loci (Szilard et al., 2010), and replication termination loci (Fachinetti et al., 2010). Only 2 regions do not host any feature. Out of the remaining 75 regions, 44 of them host at least one full-Ty annotation, 12 of them at least one LTR annotation, and 19 of them host some of the above markers of genome rearrangements, different from Ty and LTR. Many regions (31) do not involve active MGEs but correspond to loci prone to chromosomic recombination, rearrangement or fragility. We remark that all the known Tys are correctly marked as PMEs: the only Ty not recognized is the unique copy of Ty5 that appears in the telomere of Chromosome III and, because of its localization, it is ruled out from this investigation.

We then inspected the frequency of movements, by building a binary matrix B as follows: for each PME-Ty candidate $i = 1,2,\ldots,77$ and for each yeast strain $k = 1,2,\ldots,39$, we report '1' in $B[i,k]$ if this candidate occurs in that strain, and '0' if it does not. Note that we report a '1' whenever we find a undefined (or highly mutated) sequence of length compatible with a Ty. If we sum up the '1' values in B by candidates (rows), and sort the candidates according to their sums, we observe that 33 out of 44 candidates correspond to annotated Tys with score *strictly less* than 39 (i.e. there is at least one strain where the candidate is missing), whereas 31 out of 33 of the non-Ty annotated PMEs have a *full* score of 39. In other words, the non-Ty annotated PMEs do not move across all the examined strains, even though their sequence is not conserved in the genomes. Among the annotated Tys, there are some of them that appear to maintain their position across the genomes, possibly with a change in their sequence, but the large gap between the frequencies of jumps allows us to conclude that false positive candidates tend to be resident.

3 YEAST MOBILOMICS

Our last and most interesting goal is to employ the data gathered in Section 2 to compare and possibly cluster the yeast strains according to the topology of their MGEs. The fact that different sequences marked as PME-Ty candidates have a different degree of presence in the different strains, suggests us to try to understand the dynamics of these movements. We represent the topology of the MGEs of the 39 strains by creating as many binary vectors as follows.

Let p_N denote the number of non-conserved

regions within chromosome $\texttt{Chr}N$. Let $S_k(N) = (\texttt{Chr}N_k[i_1, j_1], \ldots, \texttt{Chr}N_k[i_p, j_{p_N}])$ denote the sequence of non-conserved regions in left-to-right order in chromosome $\texttt{Chr}N_k$ of the kth strain. For each $k = 1, \ldots, 39$, and for each $N = 1, \ldots, 16$, we construct a binary vector $\hat{S}_k(N)$ of the same size as $S_k(N)$, where the nth component is $'0'$ if the segment $\texttt{Chr}N_k[i_n, j_n]$ is smaller than the user-supplied size threshold d, and $'1'$ otherwise. We use default thresholds $d = 4000$ (called Ty-c threshold) and $d = 300$ (called LTR-c threshold) according to whether we want to detect just Tys or also LTRs, respectively.

Finally, let $\hat{S}_k = \hat{S}_k(1)\hat{S}_k(2)\cdots\hat{S}_k(16)$ be the binary sequence corresponding to the concatenation of the 16 chromosomes of the kth strain. Note that, since the number of conserved regions is the same in each input chromosome, then also the number of non-conserved regions is the same over all the chromosomes. It follows that the 39 binary vectors $\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_{39}$ have the same size. Using them as representatives of the mobilome topology of the corresponding 39 yeast strains, we apply the clustering package of the `scipy` scientific library (Jones et al., 2001) to perform a hierarchical clustering. The chosen metric is the Hamming distance, while the selected linkage method is `UPGMA`. In this way, we generate a tree which we call the *mobilome tree*.

The resulting mobilome tree reveals the clusters among strains obtained by minimizing the movements of PMEs. It is really interesting to compare the mobilome tree with the tree obtained by standard phylogenetic approaches based on SNPs comparison on a set of suitably identified genes (Liti et al., 2009). Almost all of the clades determined by the two trees coincide: this would support the recently established paradigm that Tys are able to drive the evolution of organisms, as reported in (Kazian, 2004). It is remarkable that the amount of information needed for our approach is really minimal, and can be obtained *a priori*. An interesting side observation is that the mobilome tree does not change when employing the binary vectors $\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_{39}$ using the LTR-c threshold $d = 300$ (rather than the Ty-c threshold): also in this case, the large majority of clades are identical to those of the classical phylogenetic tree.

REFERENCES

Bennetzen, J. (2000). Transposable elements contribution to plant gene and genome evolution. *Plant Mol Biol*, 42:251–269.

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics and Development*, 19:607–612.

Brittten, R. (2010). Transposable element insertions have strongly affected human evolution. *PNAS*, 107:19945–19948.

Di Rienzi, S., Collingwood, D., Raghuraman, M., and Brewer, B. (2010). Fragile genomic sites are associated with origins of replication. *Genome Biology and Evolution*, 1(0):350.

Fachinetti, D., Bermejo, R., Cocito, A., Minardi, S., Katou, Y., Kanoh, Y., Shirahige, K., Azvolinsky, A., Zakian, V., and Foiani, M. (2010). Replication Termination at Eukaryotic Chromosomes Is Mediated by Top2 and Occurs at Genomic Loci Containing Pausing Elements. *Molecular Cell*, 39(4):595–605.

Gerton, J., DeRisi, J., Shroff, R., Lichten, M., Brown, P., and Petes, T. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11383.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.

Kazian, H. H. (2004). Mobile elements: Drivers of genome evolution. *Science*, 303:1626–1632.

Kidwell, M. G. and Lisch, D. R. (2001). Perspective: transposable elements, parasitic dna, and genome evolution. *Evolution*, 55:1–24.

Koszul, R., Caburet, S., Dujon, B., and Fischer, G. (2004). Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.*, 23:234–243.

Leonardo, T. and Nuzhdin, S. (2002). Mobile elements and disease. *Genet Res*, 80:155–161.

Liti, G., Carter, D. M., Moses, A. M., and *et al.* (2009). Population genomics of domestic and wild yeast. *Nature*, 458:337–341.

Menconi, G., Battaglia, G., Grossi, R., Pisanti, N., and Marangoni, R. (2011). Inferring mobile elements in S.cerevisiae strains. In *International Conference on Bioinformatics Models, Methods and Algorithms*. SciTePress. ISBN 978-989-8425-36-2.

Rankin, D., Bichsel, M., and Wagner, A. (2010). Mobile dna can drive lineage extinction in prokaryotic populations. *Journal of Evolutionary Biology*, 23:2422–2431.

SGD (2010). Saccharomyces Genome Database. http://www.yeastgenome.org/.

Shao, Y., He, X., Harrison, E. M., Tai, C., Ou, H.-Y., Rajakumar, K., and Den, Z. (2010). mgenomesubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes. *Nucleic Acids Research*, 38:W194–W200.

Siefert, J. L. (2009). Defining the mobilome. *Methods in Molecular Biology*, 532:13–27.

Szilard, R., Jacques, P., Laramée, L., Cheng, B., Galicia, S., Bataille, A., Yeung, M., Mendez, M., Bergeron, M., Robert, F., et al. (2010). Systematic identification of fragile sites via genome-wide location analysis of γ-H2AX. *Nature Structural & Molecular Biology*.