Towards Mobilome Inference in Yeast Genomes

Menconi G[2], Battaglia G[1], Pisanti N[1], Grossi R[1], Marangoni R[1,3]

Motivation

Mobile genomic elements (collectively mobilome) are found in large number in eukaryotic genomes: they represent between 15% and 20% of the total human DNA. Their relationships with the resident genome can be viewed as a competition of different species in an ecosystem. Most of the mobilome is constituted by transposons, which are sequences able to replicate and jump over the genome. Transposons can cause phenotypic variations between individuals and between cells in the same individual. Indeed, some complex pathologies whose molecular mechanisms and global inheritance are hard to explain by common inheritance laws, turned out to be correlated to transposons' translocations. A contemporary challenge in comparative genomics aims at understanding the dynamics of the mobilome, so as to design a model able to describe (and even forecast) the logic followed by mobile elements to decide when and where to transpose. To this aim, it is necessary to study different lineages of organisms, to identify and locate all the mobile elements on the whole genome, and to compare the obtained results. This task is computationally challenging since the classical alignment has two main drawbacks: a) alignment algorithms do not perform efficiently on large repositories of whole chromosomes in practice; b) most of sequenced strains have unresolved regions exactly in correspondence with transposable elements. In this work, we face these drawbacks using fast algorithmic techniques that we have experimented for the analysis of different yeast strains' genomes made publicly available [Liti et al., Nature 2009].

Methods

The yeast genome contains 16 chromosomes. Our preliminary hypothesis is that the major chromosomal differences are caused by transposons' movements, since the chromosomal mutations between different strains of the same specie occur mostly for these reasons. The high similarity in the available data allows us to compare the same chromosome in two different strains by searching for L-grams shared by the two chromosomes, say, S and T, thus creating a map of these correspondences. Then, we join the L-grams located within a given distance threshold into larger runs. This final map allows us to detect insertions or deletions existing between S and T. We employ hashing based on cyclic polynomials in order to search for all the L-grams of T; this turns out to be very effective on our

[1] Dipartimento di Informatica, University of Pisa. [2] Istituto Nazionale di Alta Matematica, Roma. [3] Istituto di Biofisica, CNR, Pisa

datasets in practice. It processes a whole chromosome in just 6.5s with the longest sequence (IV, 1.5Mb) on a standard PC. We compared all the chromosomes of the yeast RefSeq@SGD to the corresponding chromosomes of two yeast strains, Y55 and YPS128.

Results

Our approach is able to identify chromosomal mutations: Figure 1 shows transposons deletions for chromosomic regions in RefSeq@SGD and the strains Y55 and YPS128. Here, the correspondences between L-grams are represented by green straight lines; deletions with respect to RefSeq@SGD appear as downward white triangles, and insertions as upward white triangles. Black rectangles on the bottom sequences indicate runs of unresolved bases ("N"). A detailed count of the number of chromosomal mutations for all the 16 chromosomes in the two strains is provided. Our results fully justify the initial hypothesis: almost all the detected mutations are indeed related to mobile elements annotated in RefSeq@SGD. The few indels apparently not related to the mobilome can be attributed to genomic rearrangements or to un-annotated mobile elements. Moreover, our results highlight that the genomic segments left unassigned after the genome sequencing are almost always represented by mobile elements. This is caused by two reasons: (a) like all repeats, mobile elements are hard to be exactly located during sequencing; (b) since RefSeq is used as a reference when assembling, a failure may occur for transposons which are not annotated in RefSeq.

Contact e-mail

menconi@mail.dm.unipi.it

Image

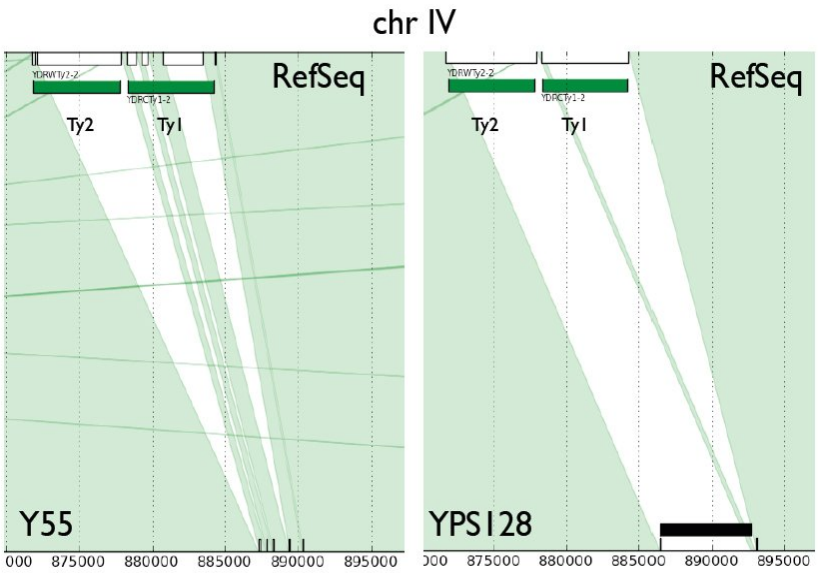

Fig.1. The plot shows an example of different transposons deletion/conservation in chr IV RefSeq@SGD and the Y55 and YPS128 strains