# Exploring Genome Rearrangements using Virtual Hybridization

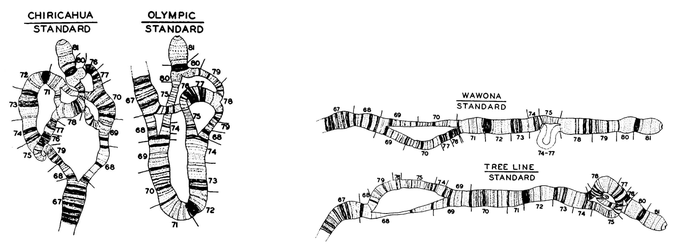
**Mahdi Belcaid**[1], **Anne Bergeron**[1], **Annie Chateau**[1], **Cedric Chauve**[1], **Yannick Gingras**[1], **Guylaine Poisson**[2]

(1) CGL, Université du Québec à Montréal, Canada (2) Information and Computer Sciences, University of Hawaii at Mānoa

## Introduction

**Visual hybridization:** Dobzhansky and Sturtevant [2]. "Genes" are sections of chromosomes identified by a combination of numbers and letters.

Relative arrangements of chromosome 3 in two strains, Standard and Santa Cruz, of *D. pseudoobscura*




$68_C$ $68_D$ 69 70 71 72 73 74 75 $76_A$ $76_B$ 77 78 $79_A$ $79_B$ 80
$68_C$ $79_B$ $76_A$ 75 74 73 72 71 70 69 $68_D$ $79_A$ 78 77 $76_B$ 80

*Top:* Loop-like configurations that occur when two homologuous rearranged chromosomes are paired. *Bottom:* fragment of the dataset constructed in 1938 by Dobzhansky and Sturtevant to compare whole genomes.

**Other techniques:**

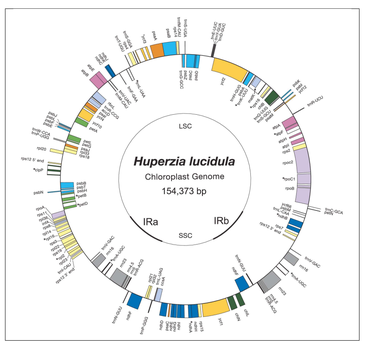
□ Chromosome painting, hybridization with probes: long and onerous;

□ Gene detection on well-annotated genomes: problems of orthologous gene identification;

□ BLAST on raw sequences cut into "conserved synteny" blocks: lot of computational ressources.

Technical and financial difficulties of reproducing independently the datasets, and of including new sequenced genomes in existing dataset, or even "revised" genomes.

**Our technique:** *Virtual hybridization* based on sets of small probes ($\approx 150$ bp), whose presence(s), absence, order and orientation can be quickly and accurately determined in a given genome. A set of probes is *useful* for a group of genomes if it captures all significant rearrangements within this group.




The chloroplast genome of *Huperzia Lucidula* ([8])

**First data set:** Chloroplast genomes are ranging from 110 kbp to 150 kbp and code typically for about 120 proteins and RNAs. The gene content is rather well conserved across species, making chloroplast genomes a very good model to study rearrangements, duplications and losses.

We found a set of 160 probes that captures most rearrangements discussed in [1, 8]. Computing the order of the probes for all the 50 choloroplast genomes available in February 2006 took ≈10 minutes on a P4 3GHz. Furthermore, this set of probes was used to confirm the reported extensive transfer of chloroplast DNA to nuclear DNA [4].

## Virtual Hybridization

Approximate string matching is defined as identifying, in a text, substrings that are *similar* to a given string $p$. In biological applications, the text is typically a genomic sequence, and similarity is defined by scoring possible aligments between $s$ and $p$. Numerous algorithms and scoring schemes are available to identify approximate occurrences of short sequences in genomic sequences, the best known being the BLAST heuristic and variations of the Smith-Waterman algorithm.

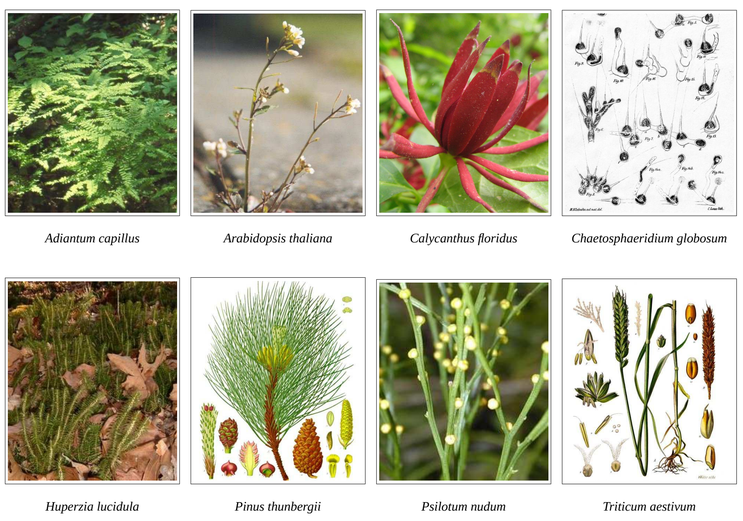
In the following, *probes* refer to sequences between 60 and 250 bp long, and *hybridization* refers to the detection of occurrences of these probes in a genomic sequence. In the current study, we detect an occurrence of a probe $p$ in a genomic sequence if there exist an aligment between $p$ and a substring $s$ of the sequence that has 80 % identity over 80 % of the length of $p$.

## Probe selection

Given a set $P$ of probes, virtual hybridization can be used to compare the presence, order, and orientation of the probes of $P$ in several genomes. A *good* set of probes should capture significant rearrangements between genomes, such as inversions, transpositions, translocations, duplications and losses.

Our first goal was to obtain a good set of probes for chloroplast genomes. Given the relatively small size of these genomes, most of the initial work was done by hand.

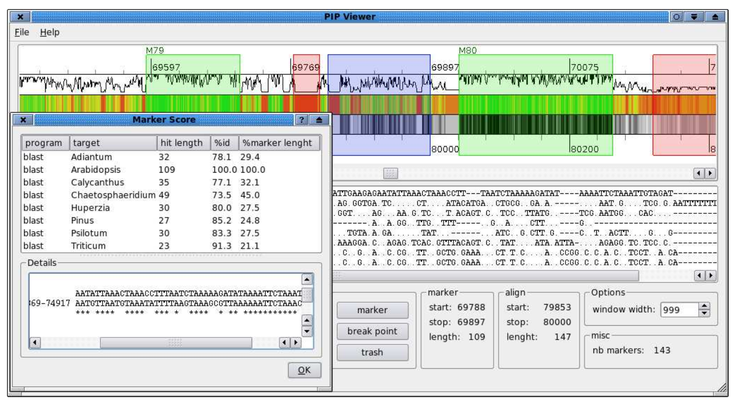
**First step:** we identified a set of *candidate probes* using global alignments of the non-duplicated regions of the following 8 references genomes:



*Adiantum capillus* | *Arabidopsis thaliana* | *Calycanthus floridus* | *Chaetosphaeridium globosum*

*Huperzia lucidula* | *Pinus thunbergii* | *Psilotum nudum* | *Triticum aestivum*

The global alignments were obtained with MultiPipMaker [7] whose output was analyzed with a graphic tool that allows the selection of a subsequence $s$ of the first genome in the list, and the computation of the virtual hybrydization score of $s$ on the remaining genomes.




Candidate probes were selected with the following criteria:

1. Good scores on at least two of the reference genomes.
2. A well defined separation between high and low scores.

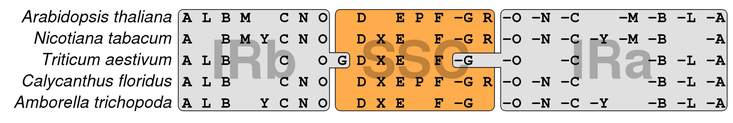
This initial set of candidate probes was then checked for adequate coverage of annotated genes

**Second step**: we eliminated candidates. We used the containment clustering algorithm icaAss [5] to detect total or partial containment. Members of each cluster were hybridized on the eight reference genomes, and the most specific probe was selected. The resulting set of probes has currently 160 members, ranging from 65 bp to 288 bp, with average length 144 bp. The following table gives the number of occurrences of probes in each of the 8 reference genomes. Note that a probe can have more than one occurrence, thus the total number of occurrences can be greater than 160.

| Genome | Single hits | Double hits | Triple hits | Total |
|---|---|---|---|---|
| *Arabidopsis thaliana* | 110 | 34 | 0 | 178 |
| *Calycanthus floridus* | 115 | 31 | 0 | 176 |
| *Pinus thunbergii* | 102 | 6 | 0 | 114 |
| *Triticum aestivum* | 96 | 29 | 2 | 160 |
| *Adiantum capillus* | 40 | 14 | 0 | 68 |
| *Psilotum nudum* | 74 | 19 | 0 | 112 |
| *Huperzia lucidula* | 87 | 16 | 0 | 119 |
| *Chaetosphaeridium globosum* | 52 | 13 | 0 | 78 |

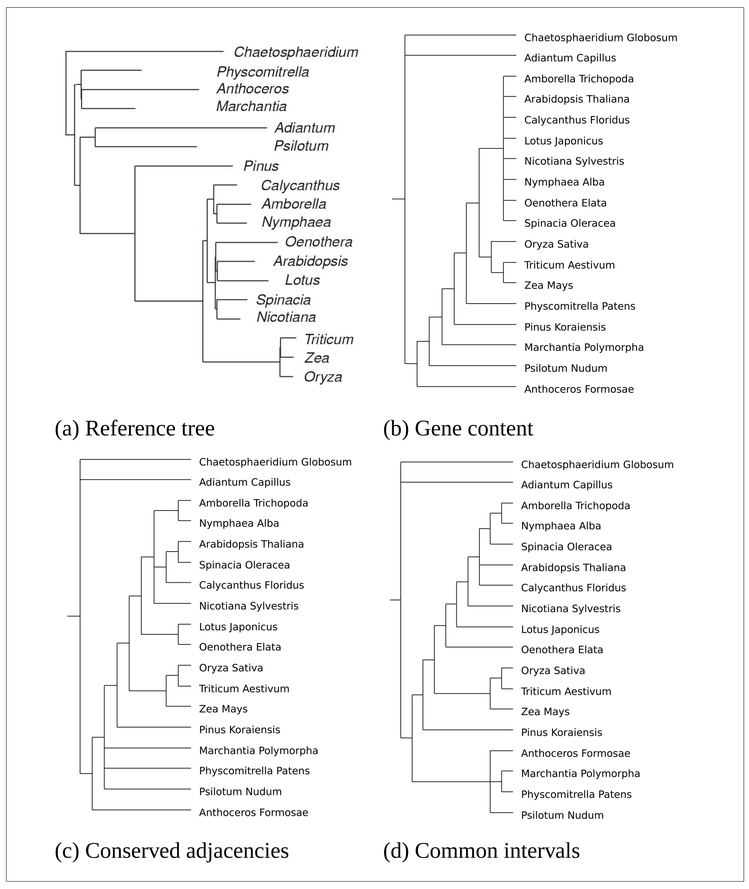
## Conserved max-gap clusters

Chloroplast genome evolution is punctuated with gene losses and duplications. Understanding these events requires computational tools that can reveal both the conserved and the variable parts of the different genomes. We used the DomainTeam algorithm [6] to identify max-gap clusters of probes. With 11 genomes, we found 588 clusters present in at least two genomes. A striking example is depicted below. It shows the SSC region flanked by parts of the inverted repeat regions, with rearrangements involving the border between regions.




## Phylogeny of 18 chloroplasts

We studied the phylogeny of 18 chloroplasts using our probes. First, chloroplasts genomes were represented as sequences of integers, based on the result of the virtual hybridization process. Next, these sequences were used to define binary characters matrices based on three models of gene order analysis (gene content, conserved adjacencies, common intervals), where we used successfull hybridizations instead of genes. Finally, these matrices were analysed with PAUP, using the Neighbor-Joining algorithm, the Hamming distance and 100 replicates of bootstrap. The reference tree is described in [3].




(a) Reference tree (b) Gene content

(c) Conserved adjacencies (d) Common intervals

## References

[1] L. Cui, J. Tang, B.M.E. Moret, and C. dePamphilis. Inferring ancestral chloroplast genomes with duplications. In preparation.

[2] Th. Dobzhansky and A, T. Sturtevant. Inversions in the Chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23:18, 1938.

[3] W. Martin, O. Deusch, N. Stawski, N. Grunheit, V. Goremykin. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 9(10):477–83, 2004.

[4] M. Matsuoa, Y. Itob, R. Yamauchib and J. Obokataa. The Rice Nuclear Genome Continuously Integrates, Shuffles, and Eliminates the Chloroplast Genome to Cause ChloroplastNuclear DNA Flux. *The Plant Cell*, 17:665–675, 2005.

[5] J.D. Parsons Improved Tools for DNA Comparison and Clustering. Comput. Applic. Biosci., 11:603–613, 1995.

[6] S. Pasek, A. Bergeron, J.-L. Risler, A. Louis, E. Ollivier and M. Raffinot. Identification of genomic features using microsyntenies of domains: domain teams. *Genome Research*, 15(6):867-74, 2005.

[7] S. Schwartz, et al.. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences, Nucl. Acids Res., 31(13):3518–3524, 2003.

[8] P.G. Wolf, et al.. The first complete chloroplast genome sequence of a lycophyte, Huperzia lucidula (Lycopodiaceae). *Gene.*, 350(2):117-28, 2005.

**Address for correspondence: Départment d'Informatique, UQÀM, CP 8888, succ. centre-ville, Montréal (QC), H3C 3P8**