EXPLORING GENOME REARRANGEMENTS USING VIRTUAL HYBRIDIZATION

M. BELCAID[1], A. BERGERON[2], A. CHATEAU[3], C. CHAUVE[2,4], Y. GINGRAS[2], G. POISSON[1] AND M. VENDETTE[2]

1) Information and Computer Sciences, University of Hawaii at Mānoa, USA
2) Comparative Genomics Laboratory, Université du Québec à Montréal, Canada
3) Laboratoire d'informatique, robotique et microélectronique de Montpellier, France
4) Department of Mathematics, Simon Fraser University, Vancouver, Canada

Genomes evolve with both mutations and large scale events, such as inversions, translocations, duplications and losses, that modify the structure of a set of chromosomes. In order to study these types of large-scale events, the first task is to select, in different genomes, sub-sequences that are considered "equivalent". Many approaches have been used to identify equivalent sequences, either based on biological experiments, gene annotations, or sequence alignments. These techniques suffer from a variety of drawbacks that often result in the impossibility, for independent researchers, to reproduce the datasets used in the studies, or to adapt them to newly sequenced genomes. In this paper, we show that carefully selected small probes can be efficiently used to construct datasets. Once a set of probes is identified – and published –, datasets for whole genome comparisons can be produced, and reproduced, with elementary algorithms; decisions about what is considered an occurrence of a probe in a genome can be criticized and reevaluated; and the structure of a newly sequenced genome can be obtained rapidly, without the need of gene annotations or intensive computations.

1. Introduction

The study of genome rearrangements started at the beginning of the last century when evidence of inversions of large segments of DNA were actually observed by Dobzhansky and Sturtevant in the chromosomes of *Drosophila pseudoobscura* [8]. Their technique, which is best described as *visual hybridization* of paired homologous rearranged chromosomes, yielded the first dataset that could be used to infer phylogenetic relationship between species using "gene" order. In that study, the word "gene" referred to sections of chromosomes and were identified by a combination of numbers and letters.

Since then, numerous techniques have been developed to compare the structure of genomes of different species. Biological experiments, such as chromosome painting [16], or hybridization with probes [11], are costly and lengthy procedures that are no longer necessary with sequenced genomes.

For well-annotated genomes, the straightforward approach of detecting whether a given species has a certain gene works only for the most elementary DNA molecules, such as animal mitochondrial genomes [4]. In bacterial genomes, for example, gene fusions lead either to the elimination of valuable information, or to the aberrant fusion of distinct gene families [13].

A way to circumvent this problem is to work directly with raw sequences, bypassing the annotation step: whole genomes are compared against each other, and the genomes of each species are cut into large blocks of "conserved synteny" [5]. This usually requires large computational resources, and if new species are added to the study, the computation must be started over again.

The main problems with these various techniques are thus the technical and financial difficulties of reproducing independently the datasets, and of including new sequenced genomes in existing dataset, or even "revised" genomes (this is the case for genome assembly projects which represent an ongoing process, and where assembly errors can easily be interpreted as large scale rearrangements [3]). It would thus be extremely valuable to have a simple and efficient method to generate datasets for the study of whole genome rearrangements.

In this paper, we propose a technique of *virtual hybridization* based on sets of small probes – up to a few hundred nucleotides –, whose presence(s), absence, order and orientation can be quickly and accurately determined in a given genome. We give two explicit sets of probes, one for the mammalian chromosome X, and one for the chloroplast genomes.

2. 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 alignments 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 [1] heuristic and variations of the Smith-Waterman algorithm [15].

In the following, *probes* refer to short sequences of nucleotides, and *virtual hybridization* refers to the detection of occurrences of these probes in a genomic sequence. We detect an occurrence of a probe p in a genomic sequence if there exists an alignment between a substring p' of p and a substring s of the sequence that with at least I % identity and such that the length of p' is at least L % of the length of p, with default values $I = 80$ and $L = 80$.

Given a chromosome C, and a set P of probes, the result of a virtual hybridization experiment is a signed sequence $p_1 p_2 \dots p_n$ which gives the order and orientation of the occurrences of probes of P in chromosome C. A probe can have more than one occurrence, or be absent from a given chromosome.

2.1. *Probe Selection*

The construction of a set of probes can be done in several different ways, the easiest being the use of already identified sets of markers common to different species. This approach is used in Section 3 in order to construct sets of probes for the mammalian chromosomes X. A alternate approach is described in Section 4 in which we present a software tool that can assist the selection procedure.

The two approaches to probe selection are first based on a multiple alignment of a small set of genomes, called *reference genomes*, in which probes are selected. The selected probes can then be hybridized with genomes different from the reference genomes. For example, in the chromosome X study, the reference genomes are the human, mouse and rat assemblies used in [5]. The set of probes was then used to analyze rearrangements in the dog and Rhesus monkey chromosomes X.

Any method of probe selection implies a series of choices that can be discussed and revised. However, once a set of probes is fixed, the information obtained in the comparison of genomes is easily and completely reproducible. The sets of probes discussed in this paper, and software to generate datasets, are available at cgl.bioinfo.uqam.ca/vhybridization.

2.2. *Probe Usefulness*

Genome rearrangement studies are all ultimately based on datasets that are signed sequences of markers. These markers can be genes, introns, exons, domains, probes or larger segments of DNA. An *occurrence* of a marker in a genome is specified by its start and end points, and its orientation (+ or -). We assume that the markers are non-overlapping in each genome in the study. The *dataset* D of the study is thus a set of signed sequences corresponding to the order and orientation of occurrences of the markers in various chromosomes.

Definition 2.1. A set P of probes is *useful* with respect to a given study if the dataset D of the study can be reconstructed using virtual hybridization.

We say that a probe – or its reverse complement – *detects* a marker m in a chromosome if: 1) It has exactly one occurrence within each occurrence of m, and the orientation of both occurrences are equal. 2) It has no occurrences outside of occurrences of m.

In order to prove that a set of probes can reconstruct the dataset of a study, it suffices to show that each marker of the study is detected by at least one probe. Given a set of n different probes that detect a set of n different markers, if $C = (m_1, m_2, \ldots ..., m_k)$ is a sequence in dataset D that describes a chromosome, then the virtual hybridization of the set of n probes on this chromosome will yield the sequence C. In Figure 1, for example, the set of probes {a, e, h} can be used to reconstruct the sequence $(m_1, m_2, m_3, -m_1, m_3)$, while capturing more rearrangements.

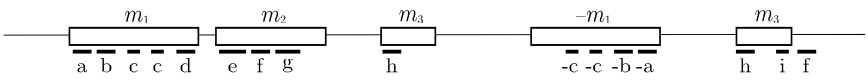

Figure 1. An example of the relations between small probes, in black, and larger markers, in white. The subset of probes {a, e, h} can be used to reconstruct the order of the markers.

For the chromosome X study, we constructed three different sets of probes, all of which can be used to reconstruct the order of synteny blocks of [5]. However, since genome assembly are often revised, we could apply the virtual hybridization procedure to the most

recent assemblies of the three reference genomes, even if the probes were constructed using the older assemblies.

For chloroplast genomes, we made sure that each annotated gene of the chloroplast of *Arabidopsis thaliana* was detected by at least one probe. As a result, we can reconstruct the datasets of studies such as [7], that use the set of annotated genes common to chloroplasts.

Adapting datasets to revised genomes, or reconstructing existing datasets is a first step. The real challenge is to be able to identify rearrangements in newly sequenced genomes, or genomes that are different from the reference genomes. We will discuss some aspects of this problem in the next section.

3. From Chromosome X Anchors to Sets of Probes

In order to develop sets of probes to investigate rearrangements in the mammalian chromosomes X, we used the set of 12866 three-way anchors identified in the comparison of the human, mouse and rat chromosomes X [5]. We selected the corresponding sequences in the human chromosome X (Apr. 2003 assembly), and retained only anchors that were longer than 75 nucleotides. This initial set of probes was hybridized against the most recent assemblies of the human (Mar. 2006), mouse (Feb. 2006) and rat (Nov. 2004) chromosomes X. With a threshold of 80% identity over 80% of the length of the probes, the initial set of anchors was reduced to 1593, after duplicate and missing hits were removed. This set of probes is called P-1593 in the following experiments.

The 1593 probes define three signed permutations that exhibit 100 conserved segments, meaning that, in all three reference chromosomes, the order and orientation of these segments are conserved. In each of these conserved segments, we chose the probe that had maximal percentage of identity with the mouse genome. The resulting set of probes is called P-100. The average length of the probes in P-100 is 277, ranging from 76 to 1548 nucleotides.

Finally, we repeated the above selection process with a threshold of 70% identity over 70% of the length of the probes, yielding 6858 probes common to the three reference genomes, that regrouped into 334 conserved segments. Again, we chose the probes that had maximal percentage of identity with the mouse genome, yielding the set of probes P-334.

We first investigated how these sets of probes captured the rearrangements of the three reference chromosomes compared to the 16 synteny blocks defined in [5]. Table 1 shows that even the set P-100 captures much more rearrangements than the 16 blocks. Note that the distances are equal for the sets P-100 and P-1593, which is a consequence of how the set P-100 was constructed.

The distances obtained in Table 1 for the three sets of probes are similar to distances that take into account both macro and micro rearrangements [5]. Lowering the threshold to 70% identity over 70% of the length predictably increases the inversion distance, since the permutations obtained by hybridization with the P-334 set have more than three times the number of breakpoints of the permutations obtained by hybridization with the P-100 set.

Using the three sets of probes, we next hybridized the dog (Jul. 2005) and Rhesus

Table 1. Inversion distances between reference chromosomes according to different sets of probes.

Pair of species	16 synteny blocks	P-1593	P-100	P-334
Human and mouse	10	33	33	115
Human and rat	10	59	59	166
Mouse and rat	10	45	45	134

monkey (Jan. 2006) chromosomes X, with the same thresholds that were used in the construction of the probes. In each experiment, about a third of the probes were not found either in the dog or the Rhesus chromosomes X. Since these are still draft assemblies, we did not investigate further the missing probes. Table 2 gives the inversion distance between pairs of genomes with respect to each of the three sets of probes.

Table 2. Inversion distances between pairs of species according to different sets of probes.

Pair of species	P-1593	P-100	P-334
Human and Rhesus	3	2	4
Human and dog	14	5	18
Rhesus and dog	13	3	14

Interestingly, for each experiment, detected rearrangements were all non-overlapping inversions. It was also possible to assign each inversion to a specific lineage. Table 2 raises some questions on the size and construction of the set of probes. Clearly, the method of selection of the P-100 set has a considerable impact in assessing the rearrangements of the dog compared to the primates. For such comparisons, the set P-1593 seems more appropriate, since some of the conserved segments between the human and rodents appear to have been broken in the dog lineage.

4. Ab-initio Probes for Chloroplast Genomes

A second project was to obtain a set of probes for chloroplast chromosomes. Given the relatively small size of these sequences, we used a semi-automated approach that relies on visual inspection. We first identified a set of *candidate probes* using global alignments of the non-duplicated regions of the references chloroplast chromosomes of Table 3.

Table 3. Reference chloroplast chromosomes

Species	Accession	Sequence (gi)
Arabidopsis thaliana	NC_000932	7525012
Calycanthus floridus	NC_004993	32480822
Pinus thunbergii	NC_001631	7524593
Triticum aestivum	NC_002762	14017551
Adiantum capillus	NC_004766	30352011
Psilotum nudum	NC_003386	18860289
Huperzia lucidula	NC_006861	60117151
Chaetosphaeridium globosum	NC_004115	22711893

A global alignment was obtained with MultiPipMaker [14]. The sequence of *Arabidopsis thaliana* was chosen as base sequence for the multiple alignment, which explains that most of the probes belong to the *Arabidopsis thaliana* chloroplast genome.

The resulting alignment was parsed using a visualization software called PipViewer. This software tool provides a representation of the multiple alignment with a color gradient, from red to green, standing respectively for low to good score. We developed PipViewer to quickly display large portions of a multiple alignment, and to select and mark blocks of contiguous nucleotides in the base sequence. When a block s is selected, PipViewer computes the virtual hybridization scores of s on the remaining sequences.

A good score is a non-ambiguous answer to the question "Does the probe hybridize at this place in the considered species ?". The first two columns of Table 4 show an example of a candidate probe of length 186 that hybridizes well with all species except *Pinus thunbergii*. The last two columns show an example of a rejected candidate probe of length 286: percentages of identity between 55 % and 70 % are considered ambiguous and yield to the rejection of the candidate probe.

Table 4. Examples of accepted and rejected candidate probes.

	Accepted candidate ($l = 186$)		Rejected candidate ($l = 286$)	
Genome	% Identity	% Probe length	% Identity	% Probe length
Arabidopsis thaliana	100.0	100.0	100.0	100.0
Calycanthus floridus	93.0	99.5	80.9	100.0
Pinus thunbergii	92.1	54.3	68.4	43.5
Triticum aestivum	93.5	100.0	79.4	100.0
Adiantum capillus	81.7	100.0	65.8	99.2
Psilotum nudum	82.7	99.5	67.6	100.0
Huperzia lucidula	91.9	100.0	71.8	100.0
Chaetosphaeridium globosum	88.2	100.0	75.8	100.0

Additional probes were added to this initial set to cover annotated genes of the reference chromosomes that were not detected by the initial set of candidates. The resulting set of candidate probes had 212 elements.

The second phase of the selection procedure was to eliminate overlapping candidates. We used the containment clustering algorithm implemented in ICAass [12] to detect total or partial containment between probes. Members of each cluster were hybridized on the eight reference chromosomes, and the most specific probe was selected. The resulting set of probes has currently 160 elements, ranging from 65 bp to 288 bp, with average length 144 bp. Table 5 gives the number of occurrences of probes in each of the 8 reference chromosomes. Note that a probe can have more than one occurrence, thus the total number of occurrences can be greater than 160.

4.1. *Investigating rearrangements in chloroplast inverted repeats*

In most chloroplast chromosomes, the presence of a large inverted repeat, with variable gene content, is a challenge to current models of genome rearrangements. One of our main

Table 5. Hits of the 160 probes on the reference chromosomes.

Genome	Single hits	Double hits ($\times 2$)	Triple hits ($\times 3$)	Total
Arabidopsis thaliana	110	34	0	178
Calycanthus floridus	115	31	0	177
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

goal in developing the virtual hybridization technique was to create a common dataset to study these types of rearrangements.

Chloroplast chromosomes are usually depicted as circular molecules divided in 4 regions (Fig. 2): a long single copy (LSC), a short single copy (SSC), and two repeated regions (IRa and IRb). For example, in the *Arabidopsis thaliana* chloroplast chromosome, the two repeated regions have 100 % identity over 26264 bp long. However, there is ample evidence [2] that chloroplast chromosome molecules exist in many other confi gurations, such as the right part of Figure 2.

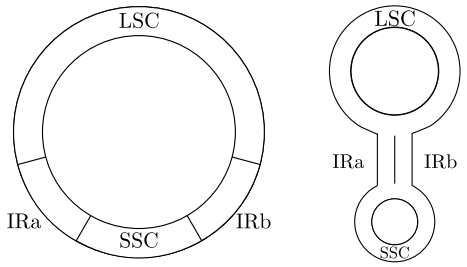

Figure 2. Chloroplast chromosomes are often depicted as round molecules divided in 4 regions: LSC, SSC, IRa, and IRb. Regions IRa and IRb are the exact inverted Watson-Crick complement of each other, thus the gene content of a chloroplast chromosome can be analyzed using the configuration on the right hand side.

Among the 160 probes, 23 of them cover the SSC region and parts of the neighboring IR region. This subset is particularly suitable to study rearrangements that occur in the IRa-SSC and SSC-IRb junctions. Table 6 gives the addresses of these 23 probes, together with a one letter code that will allow us to represent the order of these probes.

Figure 3 gives the linear order of the 23 probes in seven chloroplast chromosomes, illustrating the complex dynamics of rearrangements around the SSC region. These rearrangements cannot be explained by the classic models of inversions, tandem duplications and losses in linear sequences.

However, using a representation similar to the right-hand side of Figure 2, the relative order of the 23 probes can be compared with stem-loop diagrams (Fig. 4) that show the "Ebb and flow of the chloroplast inverted repeat" [9] in a very clear way.

Table 6. Addresses of the 23 probes that span the SSC region

Code	Sequence (gi)	Address	Code	Sequence (gi)	Address
A	14017551	98822-98897	B	7525012	123386-123510
C	7525012	123051-123169	D	22711893	99559-99740
E	30352011	125951-126108	F	7525012	111571-111716
G	7525012	112035-112194	H	7524593	106257-106331
I	7525012	114271-114351	J	32480822	115274-115356
K	14017551	106037-106135	L	7525012	116136-116278
M	7525012	117387-117565	N	7525012	117964-118080
P	7525012	119384-119489	Q	7525012	120115-120388
R	7525012	120856-120989	S	7525012	121481-121619
T	7525012	122013-122151	U	7525012	122648-122796
V	30352011	126930-127041	W	7525012	127115-127231
X	18860289	107775-107877			

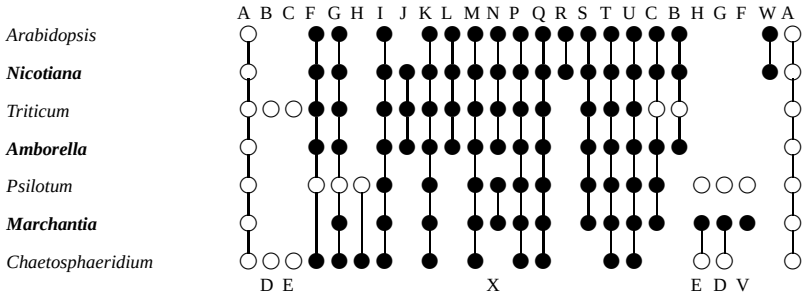

Figure 3. Linear order of 23 probes on 7 chloroplast chromosomes, showing numerous gains and losses. The SSC region of each chromosome is represented by black dots, white dots represent probes that belong to the inverted repeat region. Names in bold indicate species that are not in the reference chromosomes.

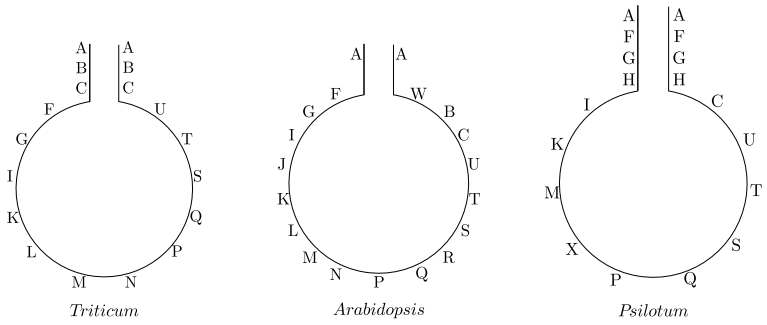

Figure 4. Respective order of probes of the SSC region and part of the inverted repeat of *Triticum*, *Arabidopsis*, and *Psilotum*. Probes B and C slip from the inverted repeat of *Triticum* to the right of SSC region of *Arabidopsis*, while probes F and G slip from the inverted repeat of *Psilotum* to the left of SSC region of *Arabidopsis*.

5. Conclusion

This paper presented a new approach to the construction of datasets used in genome rearrangement studies. We began to develop this approach when it became clear that it was

extremely difficult to share or to reproduce the data used in published papers. Many decisions must be made when producing permutations or sequences that compare gene orders in different species. Our goal was to be able to give, in a compact way, all the tools necessary to reproduce our experiments.

Chloroplast chromosomes are small, and we could do probe selection and validation using elementary tools. The corresponding set of probes seems to be able to capture most of the rearrangements occurring in chloroplasts.

Chromosomes X, on the other hand, are huge molecules. Consequently, rearrangements occur at very different scales. A small set of probes, such as P-100, can be used to detect large scale rearrangements in species that are close to the reference genomes. However, we saw that such a set of probes becomes insufficient to analyze rearrangements in farther species such as the dog, and that the set P-1593 was more adequate. The influence of the phylogenetic spectrum spanned by both the reference genomes used to select probes, and the analyzed genomes, seems then to be an issue that should be addressed, in particular the question of when a new set of probes needs to be constructed for the current set of genomes.

In this work, we did not consider bacterial genomes. However, the nature of evolutionary events that affect them – duplications, lateral transfer and gene losses in particular – induces many non trivial gene families. The analysis of bacterial gene orders is thus challenging (see [6]), and involves sophisticated algorithms. It would be interesting to use the principle of virtual hybridization, instead of all-against-all comparisons of protein sequences, with such genomes.

References

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402, 1997.
2. Bendich, A.: Circular Chloroplast Chromosomes: The Grand Illusion. *The Plant Cell*, Vol 16, 1661–1666, 2004.
3. Bérard, S., Bergeron, A., Chauve, C.: Conserved structures in evolution scenarios. *Comparative Genomics RECOMB 2004 Workshop, LNCS/LNBI*, 3388:1–15, 2005.
4. Boore, J.L.: Animal mitochondrial genomes. *Nucleic Acids Research*, 27(8):1767–1780, 1999.
5. Bourque, G., Pevzner, P.A., Tesler, G.: Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. *Genome Research*, 14(4):507–516, 2004.
6. Blin, G, Chateau, A., Chauve, C., Gingras, Y. Inferring positional homologs with common intervals of sequences. *Comparative Genomics RECOMB 2006 Workshop, LNCS/LNBI*, 4205:24–38, 2006.
7. Cui, L., Yue, F., dePamphilis, C., Moret, B.M.E., and Tang, J.: Inferring ancestral chloroplast genomes with inverted repeat. *Proceedings of the 2006 International Conference on Bioinformatics and Computational Biology (Biocomp'06)*, Las Vegas, 75–81, 2006.
8. Dobzhansky, T., Sturtevant, A.T.: Inversions in the Chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23:28–64, 1938.
9. Goulding, S.E., Olmstead, R.G., Morden, C.W., Wolfe, K.H.: Ebb and flow of the chloroplast inverted repeat. *Molecular and General Genetics*,252:195–206, 1996.
10. Matsuoa, M., Itob, Y., Yamauchib, R., Obokataa, J.: The Rice Nuclear Genome Continuously

Integrates, Shuffles, and Eliminates the Chloroplast Genome to Cause ChloroplastNuclear DNA Flux. *The Plant Cell*, 17:665–675, 2005.

11. Olmstead, R.G., Palmer, J.D.: A chloroplast DNA phylogeny of the Solanaceae: subfamilial relationships and character evolution. *Annals of the Missouri Botanical Garden*, 79:346–360, 1992.
12. Parsons, J.D.: Improved Tools for DNA Comparison and Clustering. *Computer Applications in the Biosciences*, 11:603–613, 1995.
13. Pasek, S., Bergeron, A., Risler, J.-L., Louis, A., Ollivier, E., Raffinot, M.: Identification of genomic features using microsyntenies of domains: domain teams. *Genome Research*, 15(6):867–74, 2005.
14. Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., Miller, W.: MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Research*, 31(13):3518–3524, 2003.
15. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
16. Speicher, M.R., Ballard, S.G., Ward, D.C.: Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, 12:368–376, 1996.
17. Wolf, P.G., Karol, K.G., Mandoli, D.F., Kuehl, J., Arumuganathan, K., Ellis, M.W., Mishler, B.D., Kelch, D.G., Olmstead, R.G., Boore, J.L.: The first complete chloroplast genome sequence of a lycophyte, Huperzia lucidula (Lycopodiaceae). *Gene*, 350(2):117–28, 2005.