BMC Bioinformatics

Software



# OrthoParaMap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies

Steven B Cannon*[1] and Nevin D Young[1,2]

Address: [1]Plant Biology Department, University of Minnesota, St. Paul, MN 55108, USA and [2]Plant Pathology Department, University of Minnesota, St. Paul, MN 55108, USA

Email: Steven B Cannon* - cann0010@umn.edu; Nevin D Young - neviny@umn.edu

* Corresponding author

Published: 02 September 2003

*BMC Bioinformatics* 2003, **4**:35

Received: 27 June 2003

Accepted: 02 September 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/35



## Abstract

**Background:** In eukaryotic genomes, most genes are members of gene families. When comparing genes from two species, therefore, most genes in one species will be homologous to multiple genes in the second. This often makes it difficult to distinguish orthologs (separated through speciation) from paralogs (separated by other types of gene duplication). Combining phylogenetic relationships and genomic position in both genomes helps to distinguish between these scenarios. This kind of comparison can also help to describe how gene families have evolved within a single genome that has undergone polyploidy or other large-scale duplications, as in the case of *Arabidopsis thaliana* – and probably most plant genomes.

**Results:** We describe a suite of programs called OrthoParaMap (OPM) that makes genomic comparisons, identifies syntenic regions, determines whether sets of genes in a gene family are related through speciation or internal chromosomal duplications, maps this information onto phylogenetic trees, and infers internal nodes within the phylogenetic tree that may represent local – as opposed to speciation or segmental – duplication. We describe the application of the software using three examples: the melanoma-associated antigen (MAGE) gene family on the X chromosomes of mouse and human; the 20S proteasome subunit gene family in *Arabidopsis*, and the major latex protein gene family in *Arabidopsis*.

**Conclusion:** OPM combines comparative genomic positional information and phylogenetic reconstructions to identify which gene duplications are likely to have arisen through internal genomic duplications (such as polyploidy), through speciation, or through local duplications (such as unequal crossing-over). The software is freely available at http://www.tc.umn.edu/~cann0010/.

## Background

To extend knowledge about genes in a model species to other related species, it is important to distinguish genes that are directly related to one another through speciation (orthologs) from genes that have duplicated independent of speciation (paralogs) [1]. Paralogs may be a result of many different types of gene duplication, including unequal crossing-over, transposon-mediated duplications, or polyploidy, and may have occurred recently or long before some speciation event of interest. One-to-one orthologous relationships at least hint at conservation of gene function, whereas functional relationships among complex many-to-many paralogous relationships are much more difficult to infer. Numerous studies highlight the fact that such many-to-many relationships are common, complicating the extrapolation from characterized

genes in model species to candidate genes in other species [1–4]. Identifying the nature of duplications is also important for investigating how different gene families evolve. One might expect that some gene families are arranged in a genome in such a way that paralogous duplications (and gene losses) are common, while other gene families have little tolerance for this sort of high turnover. For example, an important group of plant genes involved in pathogen resistance, the nucleotide binding site – leucine rich repeat (NBS-LRR) gene family, undergoes evolution through recombination and gene birth and death within large NBS-LRR clusters [5–8]. An example of a very different pattern (with little clustering and low rates of gene birth and death) is the gene family comprised of the 20S proteasome subunits (described later in this paper and [9]). Characterizing gene families in these terms requires identifying genes as having paralogous or orthologous origins.

A slightly different comparison can be made within a genome that has undergone polyploidy. Although any homologous genes in such a genome are technically paralogs, not all paralogs are equal: some originiate due to the polyploidy event, and others arise through other gene duplication mechanisms either before or after polyploidy. In a genome such as *Arabidopsis thaliana*, a history of polyploidy is greatly complicated by multiple rounds of (whole or partial) genome duplication, followed by extensive losses, rearrangements, and degradation of homoeologous (duplicated) regions [10–14]. Homologous genes within a single genome can be described as "segmental duplicates" (if they arose through polyploidy or other duplication of large genomic segments), or "tandem duplicates" (if they arose through unequal crossing-over), or "other duplicates" (for all other cases, such as ectopic duplications). Identifying nodes in a gene phylogeny as arising from segmental duplications can help to determine whether a set of internal duplications were part of a larger genomic duplication, or whether they came from independent, localized events. In turn, this type of characterization suggests mechanisms of the past and ongoing evolutionary mechanisms underlying the development of a gene family. Similarly, we can predict relative ages for the birth of various groups of genes by identifying whether gene births occurred before or after a particular segmental duplication. The same approaches can also be used in comparisons of two species to describe evolutionary patterns of gene births and deaths in a gene family relative to speciation timing.

We describe a method for integrating comparative genomic positional information and gene phylogenies to infer which nodes in a phylogeny were due to (1) speciation or internal genomic segmental duplications ("ortho-") and (2) which were due to tandem gene duplications ("para-") or other mechanisms. The method, implemented in a suite of three programs called DiagHunter, OrthoMap, and ParaMap (or OrthoParaMap (OPM) as shorthand for the suite), consists of identifying conserved, collinear regions in a two-way genome comparison ("diagonals" in a dot plot), calculating a gene family phylogeny, mapping the gene family onto the genome comparison, mapping the gene family/genome/genome comparison back onto the phylogeny, and inferring internal nodes at which segmental or tandem duplications probably occurred. By way of nomenclature, in the comparative genomic literature, large chromosomal regions that are similar in content and organization are frequently referred to as "synteny blocks" [15–19] or (in the case of a comparison of a genome to itself), "segmental duplications" [10,12,13,20]. The terms "homology" and "collinearity" are also used to describe such regions, but we will reserve "homology" to refer to gene relationships.

The problem of integrating phylogenetic and comparative genomic positional data bears some resemblance to the problem of combining information from species and gene phylogenies in order to distinguish orthologous from paralogous duplications. In a two-species gene tree, in the absence of additional gene duplications or losses, all genes should occur as ortholog pairs. In a gene tree with N species, all clades should, after some point, contain N genes (one for each species). Under the assumption that these kinds of orthologous relationships are more common than gene duplications and losses, several algorithms and program implementations have been developed to infer likely duplications and losses, given a gene phylogeny and a species phylogeny. Two programs that take this approach are GeneTree [21,22] and RIO (Resampled Inference of Orthologs) [23,24]. These "tree reconciliation" methods successfully identify likely orthologies, but do not make use of gene positional information to infer duplication mechanism, and do not distinguish segmental from tandem duplicates within a single polyploid genome. These are both objectives of OPM.

We describe an application of OPM to a gene family whose members reside primarily in the mammalian X chromosome: the melanoma-associated antigen (MAGE) genes in the mouse and human genomes [25,26]. The results show that several groups of genes within this gene family have followed very different evolutionary histories. We also apply OPM to two gene families from *Arabidopsis thaliana*, relating these to internal duplications within this genome. The approach shows that the selected gene families have followed strikingly different evolutionary trajectories, and also helps to characterize and confirm the relative age of a putative polyploidy event that gave rise to segmental duplications in the *Arabidopsis* genome.

## Implementation

The three main programatic steps in the OPM process, DiagHunter, OrthoMap, and ParaMap, require algorithm descriptions. All programs were implemented in Perl. Parts of the suite make use of the BioPerl libraries [27] and the GD graphics module [28].

### *DiagHunter*

Syntenic regions could, in principle, be identified using any of several genomic comparison programs, including PipMaker [29], MUMmer [30], VISTA [31], GRIMM s[32], BLASTZ [33], FORRepeats [34], or REPuter [35]. None of these, however, meet all of the critical criteria of 1) identifying both small and large (multi-megabase), contiguous or interrupted syntenic regions, 2) identifying synteny blocks with diverse data sets and genomes; 3) being freely available; 4) providing simple text output of gene pairs and coordinates of diagonals. Thus, we developed DiagHunter to meet these needs, and to be a part of the OPM distribution. Nevertheless, it should be emphasized that any program that can be adapted to produce gene pair coordinates appropriately can also be used as the first step in the OPM process. This might be advantageous if in specific cases a program has been well tested with a particular data set, for example. DiagHunter will be described briefly here. The algorithm and measures of sensitivity and selectivity are described in more detail a software report in Cannon et al. [36], and in the distribution at http://www.tc.umn.edu/~cann0010/Software.html.

Briefly, the synteny-identifying algorithm of DiagHunter walks through a pre-computed array of filtered similarity hits. The array is an M × N matrix of coordinates $(x_i, y_j)$, where $x_i$ and $y_j$ are gene "midpoint" ((5' + 3')/2) positions in genomes of sizes M and N. All coordinates are scaled by some factor to bring hits closer together (a parameter that depends on the average gene density in the genomes to be compared). At each hit, the algorithm checks in the neighborhood for hits that might either be other members of "direct" or "inverted" repeats, choosing the nearest and most favorable positions first. Once a candidate diagonal has been initiated, only hits with appropriate orientations are considered. The program follows these chains recursively, checking up to 75 possible positions in the vicinity for each step. Each position contributes a score from a scoring matrix that gives the best (lowest) scores to the nearest and most-nearly-diagonal hits. At each step, a running-average score for the candidate diagonal is computed. The program repeats the search to extend the current diagonal until a score threshold is passed or until no other candidate hits are located. If the diagonal meets the selection parameters (numbers of hits in the diagonal and average diagonal score), then it is retained, and all hits identified in this search are removed from subsequent searches. Then the program starts walking again from where it started the diagonal. Sensitivity is increased, and time to process the data is decreased, if the sparse hits are brought "closer together" by compressing the original matrix (a DiagHunter parameter). If this sort of compression is performed, then the original coordinates are recovered at the end.

### *OrthoMap*

Conceptually, this program works by identifying three-way intersections between pairs of gene family members and diagonals. Such intersections require that genes be sufficiently similar and that they have the same genomic contexts in both genomes (or in both copies of a genomic segmental duplication, in the case of a comparison of a genome with itself). For the genome self-comparison, such homologs would technically be paralogs, but as mentioned in the introduction, these share some characteristics of orthologs – and to simplify the discussion in this section, will be referred to as orthologs or ortholog candidates.

The "three-way intersection" can be visualized in terms of a dot plot comparing two chromosomes. If two genes in a gene family (one from each chromosome) "hit" within a diagonal (representing a syntenic region), then a natural assumption is that the two genes behaved like their neighbors in both chromosomes – in other words, that they are part of that diagonal, and split from one another at the same time as speciation occurred (or polyploidy or segmental duplication occurred, in the case of a comparison of a genome with itself). More specifically, the program reads a hash of "syntenic gene pair" coordinates generated by OrthoMap, then reads the positions of genes in the gene family, then BLASTs the genes in the gene family against one another, and checks whether sufficiently strong hits (coordinate pairs from the gene family) are found in or near a diagonal. "Sufficiently strong" and "near" are parameters that are specified by the user. In the analyses described here we used a BLASTP [37] E-value cutoff for the gene family of $10^{-25}$, and tested "near" cutoffs ranging from 10 kb to 250 kb in tests with mouse × human and *Arabidopsis* × self. On the basis of other tests for rates of gene duplication by distance (not shown), 50 kb was chosen as an appropriate parameter in *Arabidopsis*. The same parameter was used in the human – mouse comparison, because it was sufficient to identify most obviously clustered MAGE genes – although for general purposes, 50 kb may be conservative given the low gene density in the mouse and human genomes. More lenient values for the "near" cutoff have a beneficial effect of including more tandem duplicates as ortholog candidates, and a detrimental effect of (potentially) falsely identifying some gene pairs as members of diagonals. Identifying several tandem duplicates as ortholog

candidates does not also prevent these from being identified later as paralogs by ParaMap.

Once ortholog candidates have been identified, OrthoMap reads a phylogenetic tree file, and appends the diagonal names onto gene names in the gene phylogeny. To do this, the program takes advantage of the "extended New Hampshire" or NHX phylogeny format [38], which can include additional annotation tags at genes or internal nodes. Trees with this format can be viewed or re-annotated with the ATV application [38]. Nodes probably giving rise to orthologies are easy to spot: these are the most recent common ancestors between two genes or groups of genes that have been identified by OrthoMap as probable orthologs.

#### *ParaMap*

Local gene duplication appears to be common in many gene families. Inferring at which point in a gene phylogeny a local duplication has occurred is tedious, and is made difficult by the fact that relationships may be non-transitive (A is close to B and B to C, but A is not directly close to C), several genes in a clade may be genomically close but not appear to be near one another in a gene tree, and the definition of "close" may need to be changed for different genome comparisons. The ParaMap program reads a rooted, bifurcating tree file in which gene positions have been appended to gene names. It reads the tree from root to tip (from most ancient common ancestor out to individual genes), and recursively asks whether any two subtrees contain genes that are genomically near. If so, and if the depth in the tree is no greater than the depth specified as a parameter (for example, half of the average tree height from root to tips), then that node is annotated as a candidate origin of a local duplication. A final script combines NHX tags from "ortholog" and "paralog" trees to produce a tree with orthologous and paralogous duplications identified.

## Results and Discussion

#### *Overview of Results and Discussion*

In this section, we describe 1) objectives of the OPM suite; 2) application of the method in the MAGE gene family in mouse and human 3) application of the method to the 20S proteasome gene family in *Arabidopsis*; 4) application of the method to another gene family, the "major latex proteins" in *Arabidopsis*. 4) discussion of performance.

#### *Objectives of the OrthoParaMap suite*

Briefly, the OPM procedure consists first of pre-processing of genomic data to obtain BLAST [37] (or other similarity) scores between all genes in two genomes, followed by a search for syntenic regions, then gene family phylogeny construction, and finally "mapping" of the gene family onto the genome comparison and back onto the phylogeny to identify potential orthologs. All scripts and programs are freely available at http://www.tc.umn.edu/~cann0010/Software.html. These steps are as follows (details are in the Methods section):

• Construct a fasta file that contains ID lines with unique gene names, species and chromosome identifiers, and gene positions.

• BLAST all gene sequences from two genomes against one another (or in the case of the *Arabidopsis* example, against itself), and parse to give a similarity or "hit" matrix.

• Find syntenic regions (or "diagonals" in a genome × genome dot plot). This is accomplished with DiagHunter [36] (also briefly described below).

• For a gene family of interest, identify all genes from the genomes to be compared.

• Construct final, trimmed alignments, and a phylogeny.

• Plot gene family members in both genomes, calculate similarities between gene family members from both genomes, and map the gene family similarities onto the genome comparison. This was done with OrthoMap. The general approach consists of identifying three-way intersections between pairs of gene family members and synteny blocks from two genomes: these represent probable orthologs.

• Infer nodes responsible for orthologous gene duplications or paralogous duplications. This is done for the paralogs that appear to have arisen through local duplications, using the ParaMap program. This recursively walks through the tree, identifying internal nodes that give rise to genes or other nodes that are physically near one another on the chromosome.

#### *MAGE family example in mouse and human*

The melanoma associated antigen (MAGE) gene family [25,26] serves as a convenient and informative case study for several reasons. Most of the 20+ gene family members come from the X chromosome in both mouse and human. Because X is a sex chromosome, the MAGE and other genes on this chromosome have essentially been trapped on the chromosome (in the human genome, we identified 24 MAGE genes on the X chromosome and four on chromosomes 3 or 15; and in the mouse genome, we identified 19 MAGE genes on the X chromosome, and four on chromosomes 7 or 19). Rearrangements within this chromosome have also not been extensive [16,39,40], so synteny and collinearity have been retained to an unusual degree between the mouse and human X

chromosomes. The MAGE gene family also nicely illustrates at least two distinct evolutionary patterns.

We used the OPM programs first to identify syntenic regions between the mouse and human X chromosomes. We then used these syntenic regions to identify orthologous and paralogous duplications that have given rise to the mouse / human MAGE phylogeny. The results are shown in Figure 1, which shows predicted syntenic regions for a portion of the comparison between mouse and human X chromosomes.

If two MAGE genes are found within a syntenic region in the two genomes (a diagonal feature in the dot plot), these should simply be one of the hits or dots in the diagonal that represents this region. Because local insertions, transpositions, deletions and tandem duplications are common in genomes, it is important also to consider hits that may fall slightly off the predicted diagonal. This is done by considering hits that fall within a specified "off-diagonal threshold" for a vertical and horizontal distance in which a gene may be considered to be part of a diagonal. With a threshold of zero, ten pairs of MAGE genes are identified as having a speciation origin. This means that these genes were also identified by DiagHunter as being on a diagonal path, and are therefore parts of the synteny block predicted by that algorithm. Because small inversions and duplications are common, a higher off-diagonal threshold is desirable (though this parameter depends on features of the genomes being compared, including gene density and time to speciation, affecting amount of rearrangement and duplication). The analysis in Figure 1 used an off-diagonal threshold of 50 kb, which resulted in 17 pairs of MAGE genes being identified as having a speciation origin.

Once pairs of genes in the gene family have been mapped to predicted synteny blocks, those pairs can also be identified in a phylogeny for that family. This can be done conveniently using ParaMap, taking advantage of the "extended New Hampshire" format (NHX), described in [38]. The extended format allows additional tags to be associated with nodes or sequence names. OrthoMap determines the intersections of genes and diagonals (synteny blocks or segmental duplications), and then inserts NHX tags that contain diagonal names and chromosome identifiers, into a phylogenetic tree for the gene family. For example, at the top of Figure 1, there are two clades that each contain one Hx and one Mx sequences (standing for human X and mouse X chromosomes). Each of these would be natural ortholog candidates, simply on the basis of this phylogenetic pattern (simple one-to-one relationships, with one human sequence corresponding to one mouse sequence). The 14b and 14a tags on these sequences indicate that these came from diagonal 14 (outside the portion of the dot plot shown in Figure 1, but a similar case is shown for diagonal 9, with members near the center of the tree). The human sequence comes from the "b" (vertical) axis, and the mouse sequence comes from the "a" (horizontal) axis. There are also three other cases of likely speciation origin for MAGE genes outside of the X chromosome. These are apparent as three H15/M7 doublets, from syntenic regions on human chromosome 15 and mouse chromosome 7 (data not shown, but synteny relationships depicted in [16]). In both the comparison of MAGE genes on the X chromosomes and on 15 and 7, The combination of positional data and phylogenetic context makes it clear that many of these gene pairs are very likely to be orthologs. Once the diagonal-name tags have been added, it is a simple matter to use ATV to manually add "S" tags (for speciation or segmental duplication, depending on the genomic comparison, and distinct from tandem duplications) at internal nodes.

The last major step is to infer which internal nodes may represent local or "tandem" gene duplications. This does not require two-way genome comparisons, because tandem duplication occurs in only one genome, but the inference cannot be made from a straightforward visual inspection of the phylogeny, even if positional information is included in the phylogeny. The reason is that all pairs of sequences within subtrees (defined in terms of a portion of the total tree depth or in absolute distance terms) need to be evaluated for "closeness," and the most recent common ancestor of "close" sequences in a clade then needs to be flagged as a candidate origin of a tandem duplication. These candidate nodes are indicated in the phylogeny by "t" at tandem duplication nodes. Two such nodes are highlighted in the top half of the figure, and many are present in the bottom half of the figure. The two tandem duplications in the top half are interesting because each significantly predates speciation duplications in duplication blocks 14 and 9.

Duplication patterns in the bottom portion of the tree are dramatically different than in the top or middle of the tree. The bottom-most clade of 12 human genes has been described as MAGE family A [25,26], which maps to three adjacent clusters within 3.5 Mb centered near chromosomal band Xq28 [25]. The next clade of 4 human genes has been described as MAGE family B, which maps to a cluster at chromosomal band Xp21 [41]. ParaMap identifies likely tandem duplication origins for most of these genes. All fall into three clusters within a 3.5 Mb region. Likewise, all of the genes in the MAGE B group are identified by ParaMap as originating through tandem duplications. OrthoMap identifies the members of family A as "orthologous"to five paralogs in mouse (using a relaxed definition of orthology that allows for many-to-many relationships for genes in two species). It should be noted

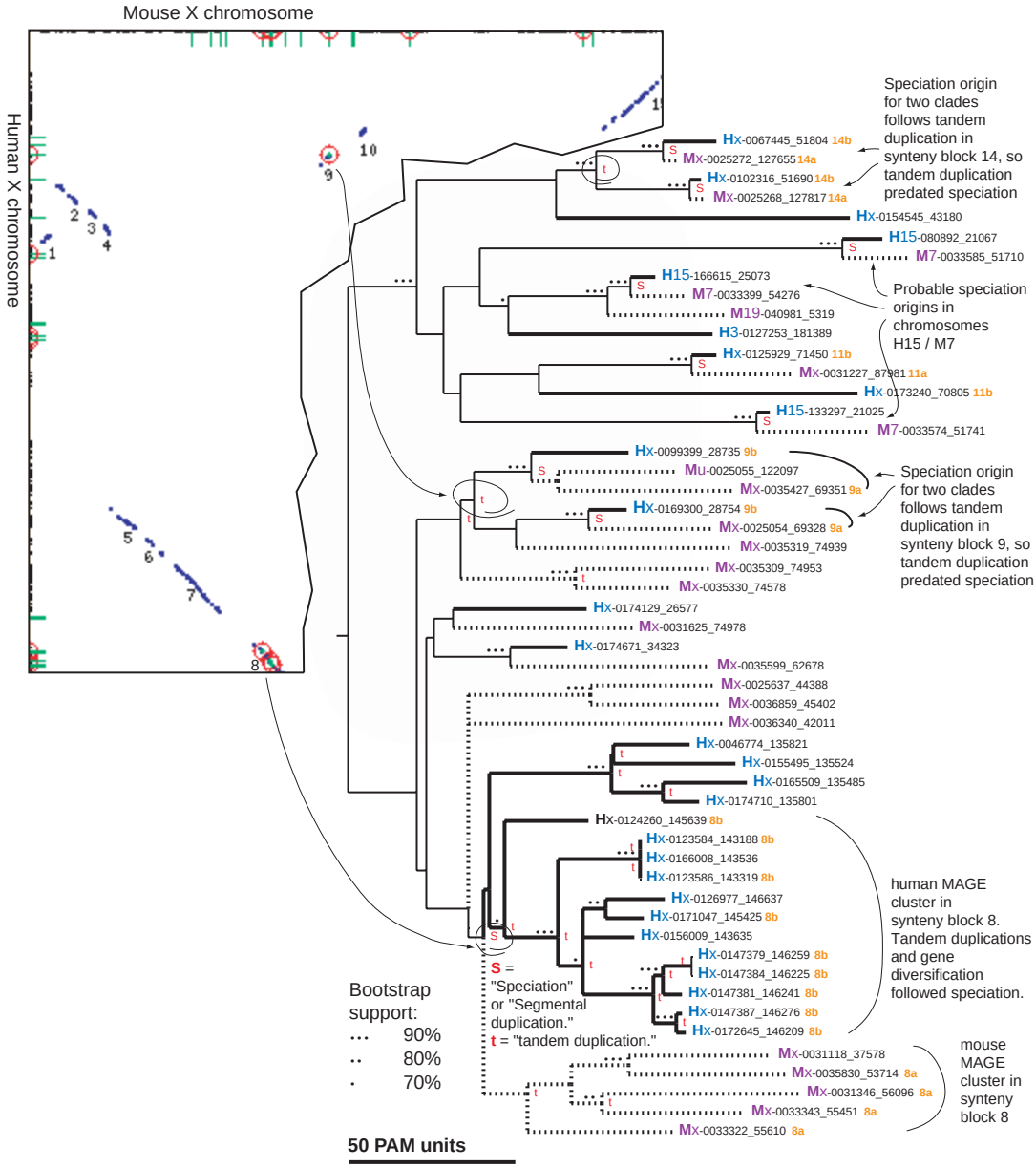


**Figure 1**

**A comparison of the mouse and human X chromosomes and the MAGE gene phylogeny**. Dot plot on the left shows the mouse X chromosome on the horizontal axis and the human X chromosome on the vertical axis, with the 5'-end of both chromosomes at the upper left corner of the dot plot. Syntenic regions predicted by DiagHunter [36,72] are highlighted in blue and numbered in the order that the program identified them. The locations of MAGE [25,26] genes are shown with short green lines on the axes. Where two of these genes from mouse and human intersect with a diagonal, they are highlighted with bulls-eyes, both on the diagonal and both axes. These points represent candidate orthologs. OrthoMap uses these diagonal names to annotate the phylogeny, shown on the right. Names in the phylogeny have the form "Mx-0035427_69351 9a". First character indicates Mouse or Human; second character or digit(s) indicates chromosome number (with u being undetermined); middle digits (after the dash) are the last seven digits of the Ensembl gene ID; digits after the underscore are the gene midpoint position in Kb; and last characters (e.g. 9a) correspond to diagonal numbers from the dot plot on, with a or b signifying horizontal or vertical axis/chromosome origin, respectively. In the tree, S indicates inferred speciation, and t indicates inferred tandem duplication (as inferred by ParaMap). Lines drawn between the middle dot plot and nodes in the phylogeny show where segmental duplications have been "mapped" between the genomic dot plot and the phylogenetic analysis. Two cases of ancient (pre-speciation) tandem gene duplications are indicated on the tree, as are cases of tandem duplications that have occurred in mouse and human after speciation.

that probable recombination and gene conversion in these regions may decrease certainty (and bootstrap values) in this portion of the tree. Clearly, this group of genes is evolving much more rapidly than those at the top of the phylogeny – both in terms of gene births through tandem duplications, and in terms of rate of change in coding sequence.

Examples of known functions for some family members suggest ways that biological roles may have shaped the evolutionary history and vice versa. The two topmost human genes in the phylogeny, Ensembl gene IDs 67445 and 102316, are the MAGE-D1 and MAGE-D2 genes. These appear to play key developmental roles in the brain [42,43] – and therefore, might be expected to be highly conserved. The large cluster of human MAGE A genes, centered near Xq28 [25,26], have been found to be highly expressed in tumors of various histological types [26], particularly in melanoma and breast carcinoma cell lines [44]. Proteins coded by these genes are also the targets of autoantibodies from patients with systemic lupus erythematosus, and so may play a role in autoimmune diseases [44]. A natural speculation is that proteins important in both autoimmune disease and cancer-recognition might require (or acquire) a nimble evolutionary strategy, analogous to the clustered and rapidly-evolving Major Histocompatibility Complex (MHC) gene family in mammals [45,46] – and in contrast to the MAGE-D1 genes that are involved in early brain development.

### Arabidopsis *20S proteasome comparison and internal genomic duplications*

Making sense of the relationship between gene function and evolution in the 20S proteasome gene family requires an understanding of the structure of the proteasome. The proteasome is responsible in eukaryotes for recycling proteins through degradation of ubiquitin-tagged proteins [9,47]. The proteasome is a large complex, consisting of a 28-subunit catalytic cylindrical structure, called the 20S proteasome, and an ATP-dependent 19S "regulatory particle," consisting of an additional set of approximately 18 subunits [48]. The combination of regulatory particle and 20S proteasome constitutes the 26S proteasome. The 20S proteasome (the catalytic core the 26S proteasome) is made up of four stacked rings. The two inner rings are each composed of seven 20S beta polypeptides, and these rings are sandwiched between two alpha rings each composed of a ring of seven alpha 26S alpha polypeptides, giving an alpha 7 beta 7 beta 7 alpha 7 structure [9]. Given this type of arrangement, it might be expected that the minimum number of kinds of proteins making up the 20S proteasome would be two: alpha and beta. Alternatively, each of the seven alpha and beta subunits might be somewhat different from one another, requiring 14 kinds of protein to make up the 20S proteasome. The simpler arrangement is seen in the archeon *Thermoplasma acidophilum*, which has a single alpha-type subunit and a single beta-type subunit [49,50]. The more complex arrangement is seen in yeast (*Saccharomyces cerevisiae*), which has seven distinct alpha-type and seven beta-type subunits [51].

In the *Arabidopsis* 20S proteasome there are 23 genes encoding subunits of the 20S proteasome [47,52–54] – but why 23 rather than 14? The phylogeny in Figure 2 suggests the answer. There are 14 groups of 20S proteasome subunits, in two groups of seven. Most (18) of the 23 proteins occur in pairs, and five are singletons. As might be expected, the two groups of seven correspond with the alpha-type subunits (top of Figure 2) and beta-type subunits (bottom of Figure 2) [9,53]. It appears that there are two nearly complete sets of alpha and beta subunits in *Arabidopsis*. OPM identifies seven of the nine pairs of subunit types as having segmental duplication origins. Furthermore, all seven of the duplication blocks involved likely occurred during the same polyploidy event. By the analysis of Blanc et al. [10,55], these duplicated segments all come from a probable round of polyploidy that they estimate occurred before the *Arabidopsis*/*Brassica rapa* split and probably during the early emergence of the crucifer family (24–40 Mya) [10]. The timing of this polyploidy episode is in agreement with other recent analyses [11,14,17]. Ermolaeva et al. [14] place this event at roughly 30 – 35 Myr, after the Brassicaceae / Malvaceae split [17], and before the *Arabidopsis*/*Brassica* split.

This relative timing of duplications in the *Arabidopsis* 20S proteasome gene family is supported when sequences from other plant species are considered. The bottom of Figure 2 shows two *Arabidopsis* proteasome subunits and probable orthologs from soybean, *Medicago*, tomato, potato, and maize. These sequences come from TIGR EST "Tentative Consensus" sequences (TCs) [56], so are inherently error-prone (i.e., may contain mis-reads or mis-assemblies). Nevertheless, in this highly expressed family of relatively small protein subunits, TCs do appear to be of high enough quality to use for approximate phylogenetic work: most TCs from these species in this gene family cover the full gene length, and most of the 14 20S subunits do have at least one representative from each included species. In the sample shown at the bottom of Figure 2, the two *Arabidopsis* sequences group together, as do the *Medicago* and *Glycine* sequences and the *Lycopersicon* and *Solanum* sequences. Furthermore, the *Zea* sequences are placed basally in the phylogeny, as would be expected of any monocot. Likewise, the solanaceous and legume sequences placements recapitulate the species phylogenies for these taxa. Though not all clades of orthologs in the phylogeny are this tidy, this example is generally typical of the remaining multi-species clades for

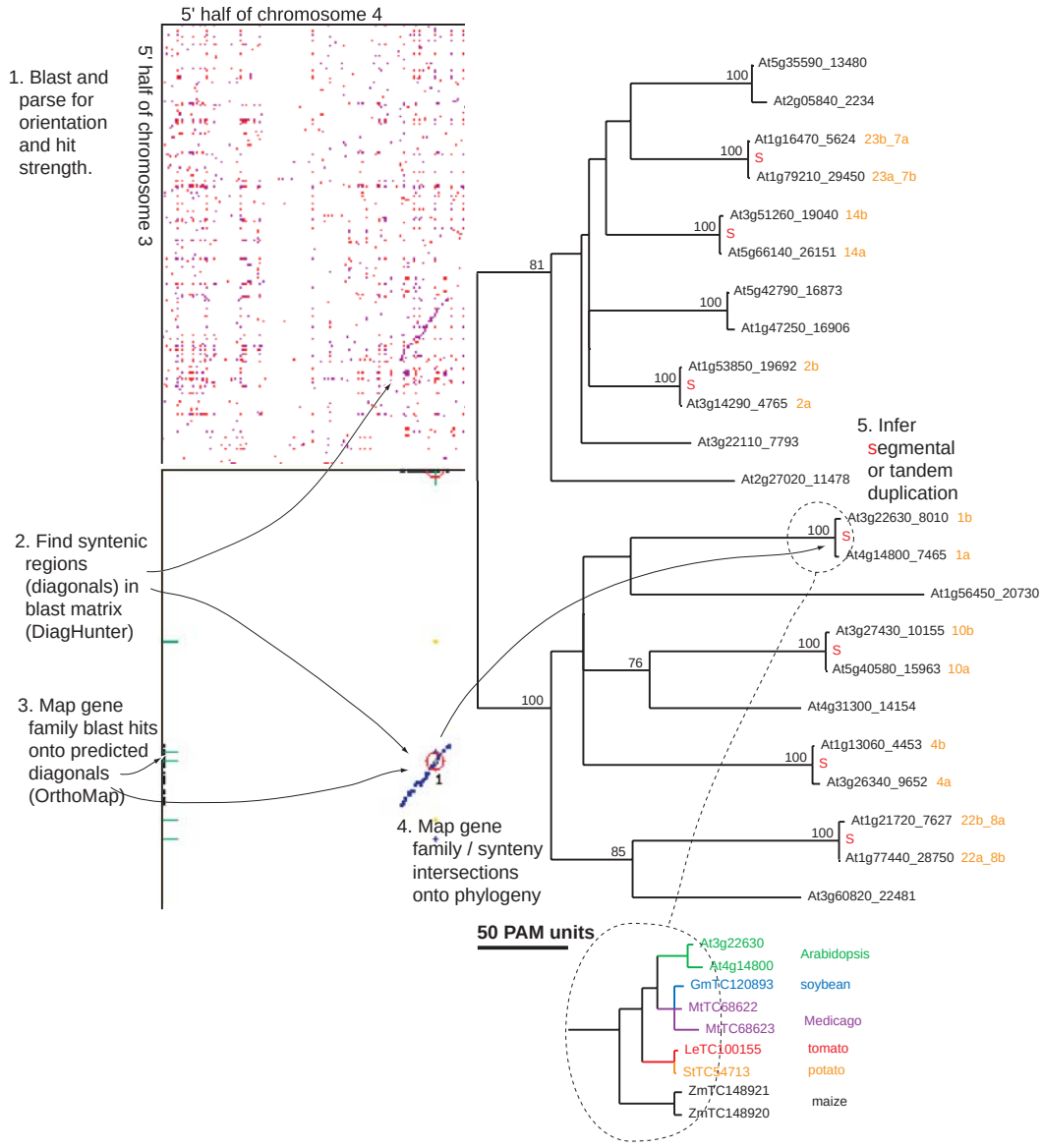


**Figure 2**

**20S proteasome gene family from *Arabidopsis*.** The OPM procedure, which involves predicting synteny blocks (dot plot and diagonal figure on the left) and mapping gene tree data onto these diagonals (phylogeny on the upper right). The dot plot in the upper left shows approximately half of *Arabidopsis* chromosomes 4 and 3 (on the horizontal and vertical axes, respectively). Each dot represents the coordinates of two proteins with bit score cutoffs of 560 (expect value of $10^{-68}$ for these data). Red dots indicate homologous proteins with the same orientations in both chromosomes, and blue indicate proteins with opposite orientations. This information is used by DiagHunter to predict synteny blocks, which are reported as text coordinates and as images (lower left). Where two 20S proteasome members intersect with a diagonal, they are highlighted with bulls-eyes, both on the diagonal and on both axes. In this version, other hits are also highlighted, with percent identity indicated using a yellow-to-black color scheme (black = 100% identity). Hits between gene family members and a diagonal represent candidate orthologs. OrthoMap uses these diagonal names to annotate the phylogeny, shown on the right. Gene names have the form "At3g22630_8010 1b", where first nine characters are the *Arabidopsis* Genome Initiative name; the number after the underscore is the position in kb on the chromosome (indicated after 'At'), and the orange numbers/letters after the space indicate the diagonal name from a chromosome comparison with that gene. Nodes giving rise to tandem gene duplications would be inferred by ParaMap and shown by a red 't' (none were found in this phylogeny). Nodes giving rise to segmental duplicates were manually inferred using the OrthoMap tags and annotated using the ATV tree-viewing program [38]. The small phylogeny at the bottom shows the positions of EST consensus sequence homologs from soybean, *Medicago truncatula*, tomato, potato, and maize (see text). These help to pinpoint when the segmental duplication occurred in this clade in *Arabidopsis*.

the gene family. For example, eight of the nine *Arabidopsis* 20S gene pairs are reciprocally their phylogenetically nearest neighbors (and in the one exception, a *Solanum* sequences nests with the two *Arabidopsis* sequences). This argues strongly for common origin of the 20S proteasome duplicates in Brassicaceae, after the split from Solanaceae and other dicot families – and consistent with current estimates for this *Arabidopsis* polyploidy episode [10,17].

The scenario suggested by the combination of biological, phylogenetic, and genome contextual information from OPM is that following polyploidy, roughly 20–40 Mya [10,11,14,17], nearly all of the members of the "extra" proteasome subunits have been maintained. Therefore, in the current *Arabidopsis*, there are two (nearly) complete sets of 20S subunits. Though there appear to have been five losses in the gene family since polyploidy, no "double losses" have been tolerated – which would have brought the number of alpha or beta subunits below seven. It is also worth noting that we see no tandem (local) duplications in the gene family. The low rates of loss or duplication in the gene family, apart from one probable round of whole-family duplication, suggest that maintenance of the stoichiometry of the 20S components is important, and that there is a significant cost to the loss or duplication of a single component. This pattern of low rates of gene duplication or loss in highly conserved, multi-subunit proteins contrasts with the following example, which shows rapid turnover of gene family members, including loss of major gene lineages in some plant families.

### Arabidopsis *Major Latex Protein gene family and internal genomic duplications*

The Major Latex Protein (MLP) family encodes proteins that were originally isolated from the latex of opium poppy, with high levels of RNA expression in poppy capsules and lactifers [57,58]. MLPs are have also been found in a wide range of plants and tissues ([59] and references therein). Functions of MLP are not known, but the MLPs do show significant similarity to a group of intracellular pathogenesis-related (IPR or PR10) proteins [60]. The IPRs typically show increased expression following wounding, pathogen attack, or stress, and several have been shown to have antibacterial, antifungal, or ribonuclease activity [60–63]. The two gene families (MLP and IPR-PR10) show only about 25% identity, but have similar structures, sizes, and pI's, and sequence and structural analyses indicate that they are similar enough to be considered to be part of a single superfamily and to be included in a single phylogenetic analysis [60].

The top phylogeny in Figure 3 shows all MLP homologs in *Arabidopsis*, with the same OPM analysis as described in the MAGE and 20S proteasome gene families. The bottom phylogeny also includes MLP homologs from *Glycine*, *Medicago*, and tomato, with a rooting at approximately the node leading to the IPR/PR10 subfamily (not shown). Interestingly, there are no *Arabidopsis* homologs that group with the IPR/PR10 subfamily [60].

Differences between the MLP and proteasome families are immediately apparent. The MLP phylogeny is "bushy" and uneven, in contrast to the regular, deeply divided, paired structure in the proteasome family. Also, while the proteasome family has no tandem duplications and at least seven (and perhaps nine) segmental duplications, the MLP family has 11 tandem and three segmental duplications. Distances to predicted segmental duplications are greater in the MLP than in the proteasome family. In the MLP family, protein distances to segmental duplications range from about 15 to 60 PAM units [64], but in the proteasome 20S family, range from 0 to about 4 PAM. Nevertheless, the MLP duplications do appear to come from the same polyploidy event as was observed in the proteasome phylogeny. By the analysis of Blanc et al. [10,55], the three duplication blocks identified in Figure 3 are all part of the "recent" polyploidy event that occurred early in the evolution of the Brassicaceae. In fact, duplication blocks 10a/21b in clades B and C in Figure 3 are part of the same duplication complex on chromosome 1 where two of the proteasome 20S duplications reside (blocks 7a/23b from the alpha subunits, and 8a/22b from the beta subunits).

Clearly, the MLP members have been evolving much more rapidly following polyploidy than have the proteasome 20S subunits. This evolution has consisted of whole-gene duplications as well as nonsynonymous changes, as is apparent from the 11 tandem duplications among the 24 MLP members.

As with the proteasome subunits, adding sequences from related taxa helps to provide evolutionary context. For example, it is possible to determine whether predicted segmental duplications may have occurred before the split between Fabaceae and Brassicaceae (not expected, if the segmental duplications are from the more recent polyploidy event rather than a more ancient event). In fact, all legume (and, for that matter, tomato) sequences have a basal placement relative to segmental duplications in the *Arabidopsis* clades, supporting the hypothesis of a more recent polyploidy in the *Arabidopsis* genome. The use of sequences from several species also provides some indication of rates of gene birth in the MLP family. The *Glycine*, *Medicago*, and tomato sequences are all EST-derived, so are subject to under-sampling and/or sequence errors, but do provide at least crude indications that genes have duplicated in particular lineages (and have probably been lost from others) following separation of these plant families.

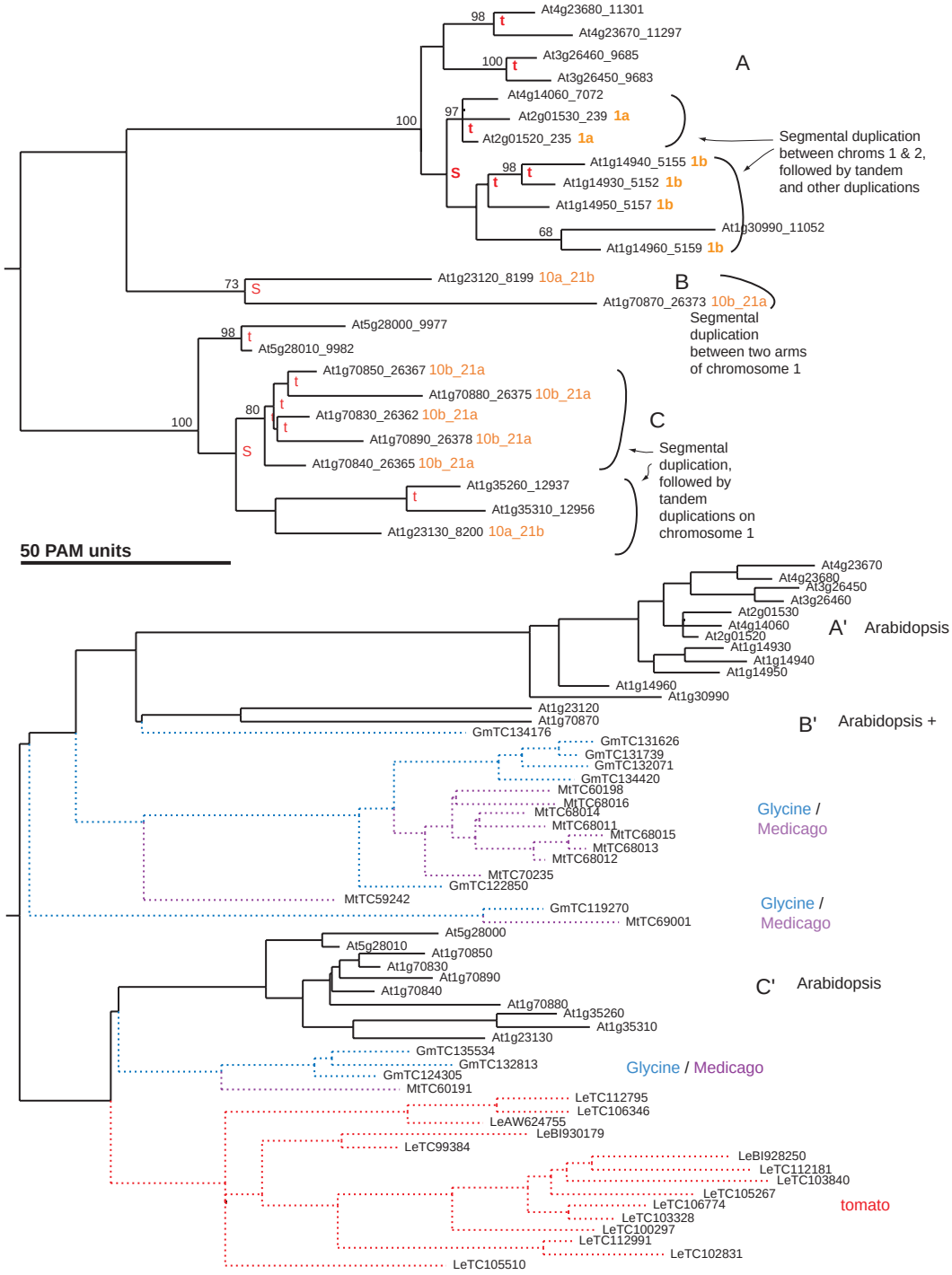


**Figure 3**

**The "major latex protein" (MLP) gene family from Arabidopsis and three comparison species**. The "major latex protein" (MLP) gene family from *Arabidopsis* is shown in the top phylogeny. In the bottom phylogeny, the same gene family is shown, but including sequences from tomato, soybean, and *Medicago truncatula* are also included. The top phylogeny was constructed using one of nine equally parsimonious topologies from a maximum parsimony search, followed by branch length estimation using Tree-Puzzle. The bottom phylogeny uses the consensus of 22 equally parsimonious topologies from a maximum parsimony search, followed by branch length estimation using Tree-Puzzle. Gene nomenclature is the same as described in Figure 2. Both trees were rooted at the approximate location of the connecting branch to the intracellular pathogenesis-related (IPR or PR10) proteins (not shown).

### *Discussion of performance*

There are several areas in which the OPM approach and implementation might be improved. The identification of syntenic regions using DiagHunter currently is strongly dependent on parameters that will differ from genome to genome. For example, the much lower gene density in the mouse and human genomes than in *Arabidopsis* meant that we needed to compress the mammalian "hit matrix" more than for the *Arabidopsis* hit matrix. Evaluations of DiagHunter specificity and selectivity are described at [36]. There is also no built-in statistical evaluation of parameters. This might be improved using a maximum likelihood or Bayesian framework.

The assignment of gene pairs to diagonals is quite straightforward, though is subject to some sources of error. A pair of genes might by chance have a hit that is near a diagonal, or might be inappropriately be judged to be outside of a diagonal, perhaps because of local rearrangements in one genome. The probability that a gene falls within any given region can be though of as a Poisson process, so it should also be a Poisson process for the coordinates of a pair of genes to randomly fall within a specified region (the space near a diagonal) in a two-dimensional comparison of genomes. The larger the gene family in comparison to the proteome, the more likely it is that a false positive will be identified. These statistical analyses could potentially be incorporated into the analysis, however, determining the sources of most (probable) errors is essentially an empirical task, and will differ between genome comparisons. For example, ongoing transposing duplications can occur in *Arabidopsis*, but cannot occur between mouse and human, so the amount of background noise should be inherently higher in the single-genome comparison – making it difficult to produce general estimates of error. Some, but not all, false positives or false negatives will be apparent in the context of a phylogeny. For example, in the 20S proteasome, it appears that two duplications in the alpha subunit subfamily probably originated as part of the same duplication that gave rise to the rest of the paired genes. Presumably, these come from duplicated regions that were too small to be picked up by DiagHunter.

An additional potential source of error is in the inference of tandem duplication origins (by ParaMap), or of segmental duplication origins. Both inferences depend on the correctness of the phylogenetic reconstruction. The MLP trees in Figure 3 illustrate the problem. These clustered, rapidly-evolving sequences, have probably undergone recombination and conversion – processes that will tend to make a phylogeny more uncertain [65]. This is evident in the number of branches with poor bootstrap support the MLP *Arabidopsis* tree, and in the alternate (equally parsimonious) solutions in clade A in the top phylogeny and the bottom multi-species phylogeny. In the lower tree, At1g14960 is placed basally in the clade, but in a derived position in the upper tree. This poses a dilemma as to which node was the actual segmental duplication origin. What is more certain is that there was some segmental duplication event near the base of this clade. There are similar difficulties in the assignment of tandem duplications. One way to approach this problem would be to calculate duplication origins for resampled bootstrap data, and then to assign segmental or tandem duplications and the frequencies of observed duplication origins to a consensus tree. The current program (ParaMap) does not currently perform these kinds of bootstrap calculations. Given this limitation, therefore, it is important to keep in mind that inferences about internal character states (duplication origins) are, in fact, reconstructions, so trees should be interpreted as possible and not absolute explanations of the data.

How do the OPM results compare with other phylogeny-based orthology identification programs such as GeneTree [21,22] or RIO [23,24]? RIO asks how often, in phylogenies calculated from a resampled gene family alignment, genes from different species appear to be paralogs or orthologs. GeneTree attempts to identify a minimum number of gene duplications needed to reconcile a gene tree with a species tree. Examining GeneTree in more detail with the MAGE data for Figure 1, the species tree has only two terminal nodes: mouse and human. Figure 4 shows a portion of the MAGE phylogeny from duplication region 9, with the OPM results on top and the GeneTree results on the bottom. In the GeneTree figure, duplications are nodes with squares, speciations are nodes without squares, and additional inferred genes are in gray. GeneTree and RIO generally appear to do a better job of identifying probable gene duplications and speciations than OPM. In this example, GeneTree identifies five probable duplications and four probable speciations vs. whereas OPM identifies three tandem duplications and two speciations. However, OPM also provides different, complementary information: it suggests mechanisms of gene duplications by taking into account gene location and synteny information, it provides a means of mapping phylogenetic data into a comparative genomic view and vice versa, and it provides data that can be used to test predictions of orthology. OPM can also be used to describe patterns of gene family evolution within a single genome, as was shown in the MLP and proteasome examples.

## Conclusions

We have described a suite of programs called OrthoParaMap (OPM) that combines comparative genomic positional information and phylogenetic reconstructions of gene families to identify which gene duplications are likely to have arisen through internal genomic duplications (such as polyploidy), or through speciation, or

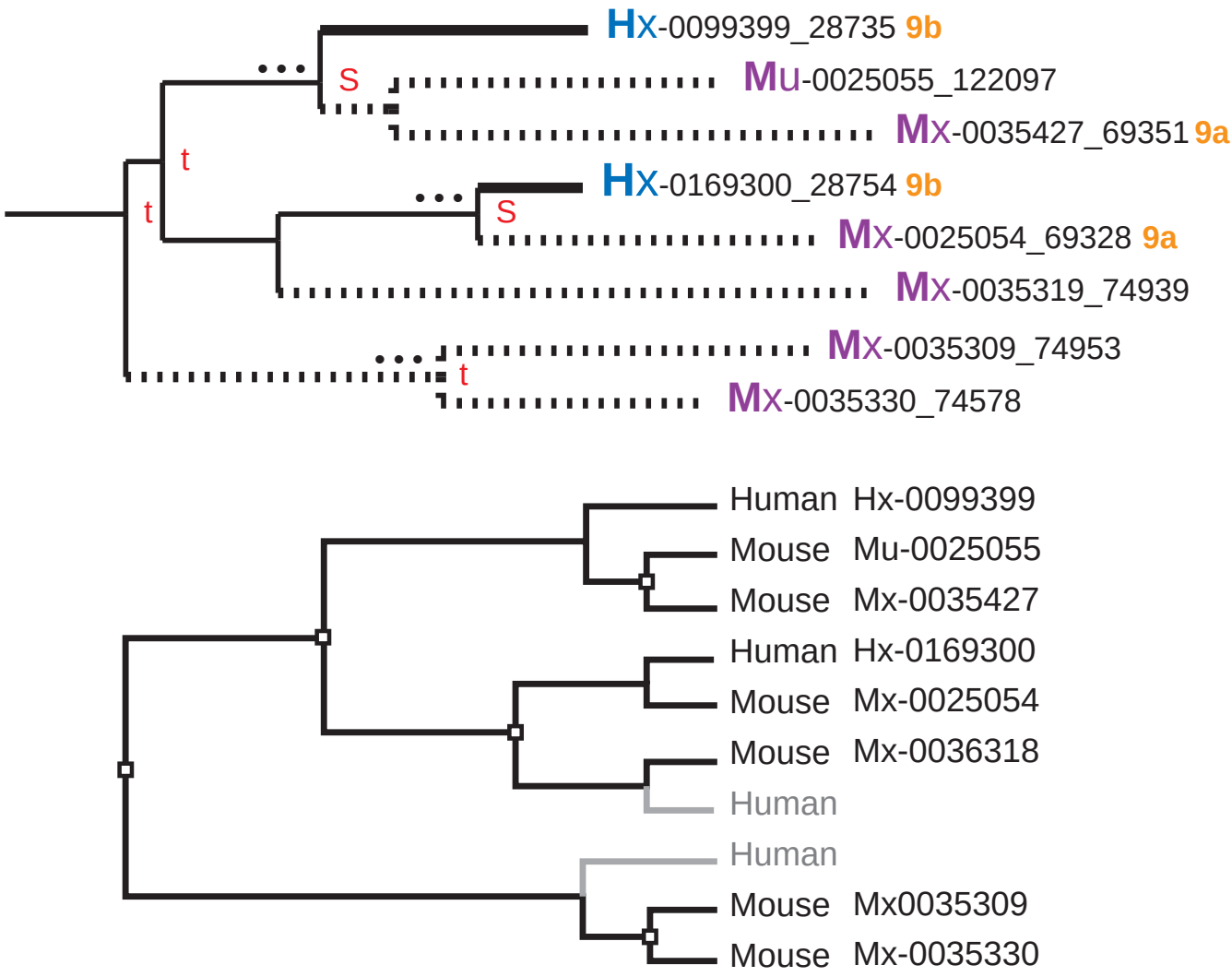


**Figure 4**

**Comparisons of approaches of GeneTree and OrthoParaMap**. The top tree is a clade from the middle of the MAGE phylogeny in Figure 1. The bottom tree shows the "reconciled" species and gene family tree predicted by GeneTree [22]. Gene duplications in the GeneTree prediction are indicated with small squares, and all other nodes are predicted speciations. Inferred sequence losses are shown in gray. The two procedures provide complementary results: OPM identifies which genes are part of synteny blocks or clusters, and GeneTree predicts some additional gene duplications or losses based on speciation and gene family trees.

through local gene duplications. We described the application of the software using three examples: the melanoma-associated antigen (MAGE) gene family on the X chromosomes of mouse and human; the 20S proteasome subunit gene family in *Arabidopsis*, and the major latex protein gene family in *Arabidopsis*. In the MAGE family, the software effectively identifies orthologs and paralogs, and highlights parts of the gene family that have undergone rapid evolution and expansion since speciation, as well as parts of the family that are highly conserved. Tandem duplications are evident both before and after speciation. In the two examples from *Arabidopsis*, OPM identifies duplications that occurred as a result of polyploidy and those that occurred due to local gene duplications, illustrating strikingly different evolutionary patterns in the two gene families.

## Methods

### *Data Sources*

Predicted mouse and human proteins from Chromosome X of both species, and Predicted MAGE genes, were retrieved from Ensembl [18] in late February, 2003. Predicted *Arabidopsis thaliana* proteins are all from the May 11, 2002 release of the MIPS *Arabidopsis thaliana* database [66]. Predicted amino acid sequences for *Glycine max*, (Gm) *Medicago truncatula* (Mt), *Solanum tuberosum* (St), *Zea mays* (Zm), and *Lycopersicon esculentum* (Le) in Figures 2 and 3 are the result of TBLASTN [37] searches of the respective TIGR EST unigene sets. TBLASTN searches were carried out using each of the *Arabidopsis* sequences in the proteasome and MLP gene families, and for each target, the longest stop-free translation was used for inclusion in the alignment and phylogeny. The TIGR Gene Indexes were May, 2003 releases: GmGI 8.0; MtGI 6.0; StGI 8.0, ZmGI 12, LeGI 9.0.

### *Alignment and Phylogeny Parameters*

Similar alignment and phylogenetic methods were used for the three gene families examined in this study. Initial alignments were constructed using T-Coffee [67]. Poorly aligning sequences were removed. HMMs were generated using HMMER [68], using the hmmalign program. For *Arabidopsis*, the HMMs were used to re-search the full set of predicted *Arabidopsis* proteins (using the hmmsearch command in HMMER), to search for other gene family members. Resulting sequences were re-aligned to the HMM (using the hmmalign command in HMMER). During HMM construction, stringent parameters were used for alignment "match states" (archpri = .7 and gapmax = .3). This assigns all remaining residues in an alignment to "insertion states" or "deletion states," and these indel sites were removed prior to tree-making steps. All alignments (ClustalX and hmmalign alignments with and without with indel sites removed) are available at a site containing similar analyses of 50 large gene families, with similar OPM treatments, at http://www.tc.umn.edu/~cann0010/genefamilyevolution.

Phylogenetic trees were constructed using both maximum parsimony and bootstrapped neighbor joining (NJ) techniques. These gave generally similar tree topologies, and all are available at http://www.tc.umn.edu/~cann0010/genefamilyevolution. Topologies generated by each of these methods were then used as the basis for computing maximum likelihood branch lengths. Parsimony trees were calculated using the Phylip 'protpars' program [69]. This produced a single most-parsimonious tree, which was fed to the Tree-Puzzle program [70] for calculating maximum likelihood branch lengths. The model of substitution was of Adachi and Hasegawa [71], amino acid frequencies were calculated from the input trees, and rate heterogeneity was allowed with 8 Gamma rate categories.

## List of abbreviations

OrthoParaMap (OPM); melanoma-associated antigen (MAGE); nucleotide binding site – leucine rich repeat (NBS-LRR); Major Latex Protein (MLP); intracellular pathogenesis-related protein (IPR); million years ago (MYA); point-accepted mutation (PAM); hidden Markov model (HMM); neighbor joining (NJ)

## Authors' contributions

SBC wrote the OPM software, carried out the analyses, and drafted the manuscript. NDY provided guidance, support, and advice throughout the project, and participated in writing of the manuscript. All authors read and approved the final manuscript.

## Availability and requirements

The software is freely available at http://www.tc.umn.edu/~cann0010/Software.html. All programs were implemented in Perl, under the Perl Artistic License. Parts of the suite make use of the BioPerl libraries [27] and the GD graphics module [28]. DiagHunter has been tested on Linux, Mac OS X, and Windows. OrthoMap and ParaMap have been tested on OS X, and should run without modification on other Unix platforms.

## Acknowledgements

This work was supported in part by a USDA National Needs fellowship and a University of Minnesota Plant Molecular Genetics Institute fellowship to SC. Special thanks to Andy Baumgarten, Georgiana May, and Alexander Kozik for conversations and advice along the way, to Kevin Roberg-Perez for suggesting use of the MAGE gene family in the mouse – human comparison, and to Martina Stromvik for suggestions on the MLP family.

## References

1. Doyle JJ and Gaut BS: **Evolution of genes and taxa: a primer.** *Plant Mol Biol* 2000, **42:**1-23.
2. Martienssen R and Irish V: **Copying out our ABCs: the role of gene redundancy in interpreting genetic hierarchies.** *Trends Genet* 1999, **15:**435-437.
3. Sankoff D: **Gene and genome duplication.** *Curr Opin Genet Dev* 2001, **11:**681-684.
4. Lynch M and Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
5. Baumgarten AM, Cannon SB,, Spangler R and May G: **Genome-level evolution of NBS-LRR resistance genes in Arabidopsis thaliana.** *Genetics* 2003, **[in press]:**.
6. Michelmore R and Meyers BC: **Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process.** *Genome Res* 1998, **8:**1113-1130.
7. Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW and Young ND: **Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily.** *Plant J* 1999, **20:**317-332.
8. Cannon SB, Zhu H, Baumgarten AM, Spangler R, May G, Cook DR and Young ND: **Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies.** *J Mol Evol* 2002, **54:**548-562.

9. Fu H, Doelling JH, Arendt CS, Hochstrasser M and Vierstra RD: **Molecular organization of the 20S proteasome gene family from Arabidopsis thaliana.** *Genetics* 1998, **149:**677-692.
10. Blanc G, Hokamp K and Wolfe KH: **A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome.** *Genome Res* 2003, **13:**137-144.
11. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M and Van de Peer Y: **The hidden duplication past of Arabidopsis thaliana.** *Proc Natl Acad Sci U S A* 2002, **99:**13627-13632.
12. Vandepoele K, Simillion C and Van de Peer Y: **Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice.** *Trends Genet* 2002, **18:**606-608.
13. Vision TJ, Brown DG and Tanksley SD: **The origins of genomic duplications in Arabidopsis.** *Science* 2000, **290:**2114-2117.
14. Ermolaeva MD, Wu MM, Eisen JA and Salzberg SL: **The age of the Arabidopsis thaliana genome duplication.** *Plant Mol Biol* 2003, **51:**859-866.
15. Delcher AL: **MUMmer.** 2002 [http://www.tigr.org/software/mummer/].
16. Pevzner P and Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes.** *Genome Res* 2003, **13:**37-45.
17. Bowers JE, Chapman BA, Rong J and Paterson AH: **Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events.** *Nature* 2003, **422:**433-438.
18. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, Kasprzyk A, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I and Birney E: **Ensembl 2002: accommodating comparative genomics.** *Nucleic Acids Res* 2003, **31:**38-42.
19. Ku HM, Vision T, Liu J and Tanksley SD: **Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny.** *Proc Natl Acad Sci U S A* 2000, **97:**9121-9126.
20. Ziolkowski PA, Blanc G and Sadowski J: **Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome.** *Nucleic Acids Res* 2003, **31:**1339-1350.
21. Page RD and Charleston MA: **From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem.** *Mol Phylogenet Evol* 1997, **7:**231-240.
22. Page RD: **GeneTree: comparing gene and species phylogenies using reconciled trees.** *Bioinformatics* 1998, **14:**819-820.
23. Zmasek CM and Eddy SR: **A simple algorithm to infer gene duplication and speciation events on a gene tree.** *Bioinformatics* 2001, **17:**821-828.
24. Zmasek CM and Eddy SR: **RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs.** *BMC Bioinformatics* 2002, **3:**14.
25. Rogner UC, Wilke K, Steck E, Korn B and Poustka A: **The melanoma antigen gene (MAGE) family is clustered in the chromosomal band Xq28.** *Genomics* 1995, **29:**725-731.
26. De Plaen E, Arden K, Traversari C, Gaforio JJ, Szikora JP, De Smet C, Brasseur F, van der Bruggen P, Lethe B, Lurquin C and et al.: **Structure, chromosomal localization, and expression of 12 genes of the MAGE family.** *Immunogenetics* 1994, **40:**360-369.
27. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD and Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
28. Stein LD: **GD.pm perl module.** 2003 [http://stein.cshl.org/WWW/software/GD/].
29. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R and Miller W: **PipMaker--a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10:**577-586.
30. Delcher AL, Phillippy A, Carlton J and Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic Acids Res* 2002, **30:**2478-2483.
31. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS and Dubchak I: **VISTA : visualizing global DNA sequence alignments of arbitrary length.** *Bioinformatics* 2000, **16:**1046-1047.
32. Tesler G: **GRIMM: genome rearrangements web server.** *Bioinformatics* 2002, **18:**492-493.
33. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D and Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13:**103-107.
34. Lefebvre A, Lecroq T, Dauchel H and Alexandre J: **FORRepeats: detects repeats on entire chromosomes and between genomes.** *Bioinformatics* 2003, **19:**319-326.
35. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J and Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale.** *Nucleic Acids Res* 2001, **29:**4633-4642.
36. Cannon SB, Kozik A, Chan B, Michelmore R and Young ND: **DiagHunter: a program for genomic comparisons and large-scale synteny-discovery.** *Genome Biology* .
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
38. Zmasek CM and Eddy SR: **ATV: display and manipulation of annotated phylogenetic trees.** *Bioinformatics* 2001, **17:**383-384.
39. DeBry RW and Seldin MF: **Human/mouse homology relationships.** *Genomics* 1996, **33:**337-351.
40. Carver EA and Stubbs L: **Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale.** *Genome Res* 1997, **7:**1123-1137.
41. Lurquin C, De Smet C, Brasseur F, Muscatelli F, Martelange V, De Plaen E, Brasseur F, Monaco AP and Boon T: **Two members of the human MAGEB gene family located in Xp21.3 are expressed in tumors of various histological origins.** *Genomics* 1997, **46:**394-408.
42. Salehi AH, Roux PP, Kubu CJ, Zeindler C, Bhakar A, Tannis LL, Verdi JM and Barker PA: **NRAGE, a novel MAGE protein, interacts with the p75 neurotrophin receptor and facilitates nerve growth factor-dependent apoptosis.** *Neuron* 2000, **27:**279-288.
43. Pold M, Zhou J, Chen GL, Hall JM, Vescio RA and Berenson JR: **Identification of a new, unorthodox member of the MAGE gene family.** *Genomics* 1999, **59:**161-167.
44. McCurdy DK, Tai LQ, Nguyen J, Wang Z, Yang HM, Udar N, Naiem F, Concannon P and Gatti RA: **MAGE Xp-2: a member of the MAGE gene family isolated from an expression library using systemic lupus erythematosus sera.** *Mol Genet Metab* 1998, **63:**3-13.
45. Anzai T, Shiina T, Kimura N, Yanagiya K, Kohara S, Shigenari A, Yamagata T, Kulski JK, Naruse TK, Fujimori Y, Fukuzumi Y, Yamazaki M, Tashiro H, Iwamoto C, Umehara Y, Imanishi T, Meyer A, Ikeo K, Gojobori T, Bahram S and Inoko H: **Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence.** *Proc Natl Acad Sci U S A* 2003, **100:**7708-7713.
46. Garcia-Lora A, Algarra I and Garrido F: **MHC class I antigens, immune surveillance, and tumor immune escape.** *J Cell Physiol* 2003, **195:**346-355.
47. Vierstra RD: **The ubiquitin/26S proteasome pathway, the complex last chapter in the life of many plant proteins.** *Trends Plant Sci* 2003, **8:**135-142.
48. Fu H, Doelling JH, Rubin DM and Vierstra RD: **Structural and functional analysis of the six regulatory particle triple-A ATPase subunits from the Arabidopsis 26S proteasome.** *Plant J* 1999, **18:**529-539.
49. Lowe J, Stock D, Jap B, Zwickl P, Baumeister W and Huber R: **Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution.** *Science* 1995, **268:**533-539.
50. Zwickl P, Grziwa A, Puhler G, Dahlmann B, Lottspeich F and Baumeister W: **Primary structure of the Thermoplasma proteasome and its implications for the structure, function, and evolution of the multicatalytic proteinase.** *Biochemistry* 1992, **31:**964-972.
51. Hochstrasser M, Johnson PR, Arendt CS, Amerik AYu, Swaminathan S, Swanson R, Li SJ, Laney J, Pals-Rylaarsdam R, Nowak J and Connerly PL: **The Saccharomyces cerevisiae ubiquitin-proteasome system.** *Philos Trans R Soc Lond B Biol Sci* 1999, **354:**1513-1522.
52. von Arnim AG: **A hitchhiker's guide to the proteasome.** *Sci STKE* 2001, **2001:**PE2.
53. Parmentier Y, Bouchez D, Fleck J and Genschik P: **The 20S proteasome gene family in Arabidopsis thaliana.** *FEBS Lett* 1997, **416:**281-285.

54. Gray WM and Estelle I: **Function of the ubiquitin-proteasome pathway in auxin response.** *Trends Biochem Sci* 2000, **25:**133-138.
55. Blanc G and Wolfe K: **Paralagons in Arabidopsis thaliana.** 2002 [http://wolfe.gen.tcd.ie/athal/].
56. Quackenbush J, Liang F, Holt I, Pertea G and Upton J: **The TIGR gene indices: reconstruction and representation of expressed gene sequences.** *Nucleic Acids Res* 2000, **28:**141-145.
57. Nessler CL and Burnett RJ: **Organization of the major latex protein gene family in opium poppy.** *Plant Mol Biol* 1992, **20:**749-752.
58. Nessler CL: **Sequence analysis of two new members of the major latex protein gene family supports the triploid-hybrid origin of the opium poppy.** *Gene* 1994, **139:**207-209.
59. Stromvik MV, Sundararaman VP and Vodkin LO: **A novel promoter from soybean that is active in a complex developmental pattern with and without its proximal 650 base pairs.** *Plant Mol Biol* 1999, **41:**217-231.
60. Osmark P, Boyle B and Brisson N: **Sequential and structural homology between intracellular pathogenesis-related proteins and a group of latex proteins.** *Plant Mol Biol* 1998, **38:**1243-1246.
61. Bufe A, Spangfort MD, Kahlert H, Schlaak M and Becker WM: **The major birch pollen allergen, Bet v 1, shows ribonuclease activity.** *Planta* 1996, **199:**413-415.
62. Flores T, Alape-Giron A, Flores-Diaz M and Flores HE: **Ocatin. A novel tuber storage protein from the andean tuber crop oca with antibacterial and antifungal activities.** *Plant Physiol* 2002, **128:**1291-1302.
63. Moiseyev GP, Fedoreyeva LI, Zhuravlev YN, Yasnetskaya E, Jekel PA and Beintema JJ: **Primary structures of two ribonucleases from ginseng calluses. New members of the PR-10 family of intracellular pathogenesis-related plant proteins.** *FEBS Lett* 1997, **407:**207-210.
64. Dayhoff MO: **Atlas of Protein Sequences and Structure. Volume 5, Supplement 3, pp. 353-358.** *Washington, DC, USA, National Biomedical Research Foundation*; 1979.
65. White J and Crother BI: **Gene conversions may obscure actin gene family relationships.** *J Mol Evol* 2000, **50:**170-174.
66. Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW and Mayer KF: **MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome.** *Nucleic Acids Res* 2002, **30:**91-93.
67. Notredame C, Holm L and Higgins DG: **T-COFFEE: an objective function for multiple sequence alignments.** *Bioinformatics* 1998, **14:**407-422.
68. Eddy SR: **HMMER: Profile hidden Markov models for biological sequence analysis: The HMMER User's Guide (http://hmmer.wustl.edu/).** 2001.
69. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genetics, University of Washington, Seattle.** 2000.
70. Schmidt HA, Strimmer K, Vingron M and von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
71. Adachi J and Hasegawa M: **Model of amino acid substitution in proteins encoded by mitochondrial DNA.** *J Mol Evol* 1996, **42:**459-468.
72. Cannon SB: **DiagHunter web site.** 2003, **2003:**.