

RESEARCH

Open Access

# Gene families as soft cliques with backbones: *Amborella* contrasted with other flowering plants

Chunfang Zheng<sup>1</sup>, Alexey Kononenko<sup>1</sup>, Jim Leebens-Mack<sup>2</sup>, Eric Lyons<sup>3</sup>, David Sankoff<sup>1\*</sup>

From Twelfth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics

Cold Spring Harbor, NY, USA. 19-22 October 2014

## Abstract

**Background:** Chaining is a major problem in constructing gene families.

**Results:** We define a new kind of cluster on graphs with strong and weak edges: soft cliques with backbones (SCWiB). This differs from other definitions in how it controls the “chaining effect”, by ensuring clusters satisfy a tolerant edge density criterion that takes into account cluster size. We implement algorithms for decomposing a graph of similarities into SCWiBs. We compare examples of output from SCWiB and the Markov Cluster Algorithm (MCL), and also compare some curated *Arabidopsis thaliana* gene families with the results of automatic clustering. We apply our method to 44 published angiosperm genomes with annotation, and discover that *Amborella trichopoda* is distinct from all the others in having substantially and systematically smaller proportions of moderate- and large-size gene families.

**Conclusions:** We offer several possible evolutionary explanations for this result.

## Background

The automatic detection of clusters of vertices in a graph is practiced in diverse fields from image recognition to social networks, and is widely used in computational biology for the study of gene families. Conceptually, a gene family is a set of genes, in one genome or several, that includes all descendants of a single gene in some ancestral organism (i.e., homologous genes), and excludes genes descended from other ancestral genes (i.e., non-homologous genes). Operationally, lacking the historical data to identify a gene family in these terms, it is standard practice to construct gene families on the basis of DNA or protein sequence similarities. The assumption is that genes in the same family will retain much more sequence similarity than unrelated genes, though this is more of a general tendency than a strict rule. The genes belonging to a particular gene family may be identified

with the vertices of a graph, which has edges between pairs of genes exceeding a threshold similarity score.

In the present work, we will focus on gene families within a single genome. We set aside data on syntenic correspondences between orthologs as well as functional evidence relating genes, despite their usefulness in many contexts, in order to achieve the first of our two goals - the identification of the conceptual and methodological problems in the purely graph-theoretical approach, and the framing of a proposal to deal with them.

In plants, the creation, expansion and attrition of gene families through mechanisms of gene duplication, notably tandem duplication and, more spectacularly, whole genome doubling, allows rapid adaptation of populations to a broad range of niches. This motivates the second of our two goals in this paper - a comprehensive survey of gene family sizes in 44 published angiosperm genomes.

In the first part of this paper, we review some of the desiderata of clustering methods in graphs, and define a new kind of cluster: *soft cliques with backbones* (SCWiB). Though similar in some respects to methods based on Minimum Spanning Tree, SCWiB clustering controls the

\* Correspondence: sankoff@uottawa.ca

<sup>1</sup>Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada, K1N 6N5

Full list of author information is available at the end of the article

“chaining effect” characteristic of many such methods, by requiring that clusters satisfy a tolerant edge density criterion that takes into account cluster size. We present an exact algorithm based on the SCWiB concept that can handle moderate amounts of data, and that can be converted into a heuristic for realistic genomes.

We then compare SCWiB results with the output of MCL [1,2], a method which is one of the most widely used for inferring gene families. We also see how the SCWiB families compare with the curated gene families of *Arabidopsis thaliana* [3].

Finally we apply our algorithm to 44 published angiosperm genome sequences. We compare the distribution of gene family sizes, and find similar patterns are displayed in the large majority of cases. We find, however, that the earliest branching angiosperm, *Amborella trichopoda* has a distinctly different pattern, with relatively few moderate- or large-size families.

## Results

Creating gene families on the basis of similarities is essentially a kind of clustering. Well-known clustering methods like  $k$ -means [4], hierarchical methods, e.g., single-link [5], average-link [6] and complete-link, spatial methods, e.g., PCA and self-organizing maps [7], and graph-based methods, e.g., minimum spanning trees [8] and cliques, have all been used. These all have advantages and disadvantages, depending on the context. In our study of gene families, we wanted to avoid methods that produce large, inhomogeneous, families by “chaining”, such as single-link, on the one hand, and methods that are overly biased towards smaller or equal-sized families, like clique or complete-link, on the other.

Chaining is a major problem in constructing gene families, largely due to the multiple domain structure of many proteins. Some domains recur in many different families, with the result that both conceptually and operationally, there are no longer strict boundaries between families. This problem has been treated in most depth by Joseph and Durand [9,10]. Methods that construct clusters by adding objects to that cluster with an element closest to them, without accounting for the rest of the cluster, like single-link or minimum spanning tree are particularly prone to chaining and, in some applications, like recognition of objects in satellite imagery, this may be desirable [11,12]. However, in the context of constructing gene families, this amounts to the inclusion of non-homologous genes within the same family, something to be avoided in evolutionary analyses

### Gene families as soft cliques with backbones

To ensure that a gene family is determined by strong similarities connecting each of its members,

- we set a high similarity threshold  $U$  and require that a cluster be connected, in the graph theoretical sense, solely in terms of similarities exceeding  $U$ . By itself this is similar to other graph theory criteria, and like them it is susceptible to chaining, for meaningful values of  $U$  or, alternatively, to very small clusters, if  $U$  is too high. To control for chaining
- we also set a less stringent threshold  $W$ , and require that the elements in the cluster form a clique, or almost form a clique, in terms of similarities exceeding  $W$ . We cannot require that the cluster forms a full clique, since this is too stringent for high values of  $W$ , and is not restrictive enough for low values. A way of relaxing the clique criterion is
- to require the similarities in a cluster to form an  $S$ -plex [13], where  $S = sN + 1$ , the number of elements in the cluster being  $N$ , and  $0 < s < 1$  is a constant. In an  $S$ -plex, every element is of degree at least  $N - S$ .

Each cluster is thus validated on two levels, as a set of strongly connected elements, at level  $U$ , that is not built by chaining, due to the  $S$ -plex condition at level  $W$ .

### Exact algorithm

The algorithm generates all possible SCWiBs in a graph. The output can then be post-processed to find a compatible subset of these to satisfy any one of several criteria.

#### Algorithm SCWiB

**Parameters:** thresholds  $0 < W < U < 1$  and tolerance coefficient  $0 < s < 1$

**Input:** graph  $G(V, E)$  with edge-weights  $w(\cdot)$

**Output:** the list  $\mathcal{L}$  of the possible SCWiBs in  $G$ .

#### Steps:

define  $E_U = \{e \in E \mid U \leq w(e)\}$ ,  $E_W = \{e \in E \mid W \leq w(e) < U\}$

order vertices by increasing degree

while there are more vertices

    select the first vertex  $v$  from the ordered list of vertices

    call **ListGeneFamily**( $v, G, s, L_1$ )

    store  $L_1$  results in  $\mathcal{L}$ .

    remove  $v$  from  $G$

#### Algorithm ListGeneFamily( $v, G, s, L_1$ )

**Input:** vertex  $v$ , graph  $G(V, E)$ ,  $E = E_U \cup E_W$

**Output:** the list  $L_1$  of the all the SCWiBs in  $G$  that contain  $v$

#### Steps:

let  $d$  be the degree of vertex  $v$

maximum size  $m$  of SCWiB that can contain  $v$  is  $m = \frac{d}{s} + 1$

create a queue  $Q$ , insert the subgraph with only one vertex  $v$  into  $Q$

while there are more subgraphs in  $Q$

    pop out the first subgraph  $sg$  and store it into  $L_1$

if  $|sg| < m$ ,  
for each vertex  $u \in V$  (and  $\notin sg$ ) joined with  
an edge in  $E_U$  to the last vertex of  $sg$ ,  
check if the subgraph  $G'(V', E')$  of  $G$  is a  
SCWiB cluster, ( $V' =$  vertices of  $sg \cup \{u\}$ .  
 $E'$  is the edge set induced by  $V'$ ) if yes,  
insert  $G'$  into  $Q$ .

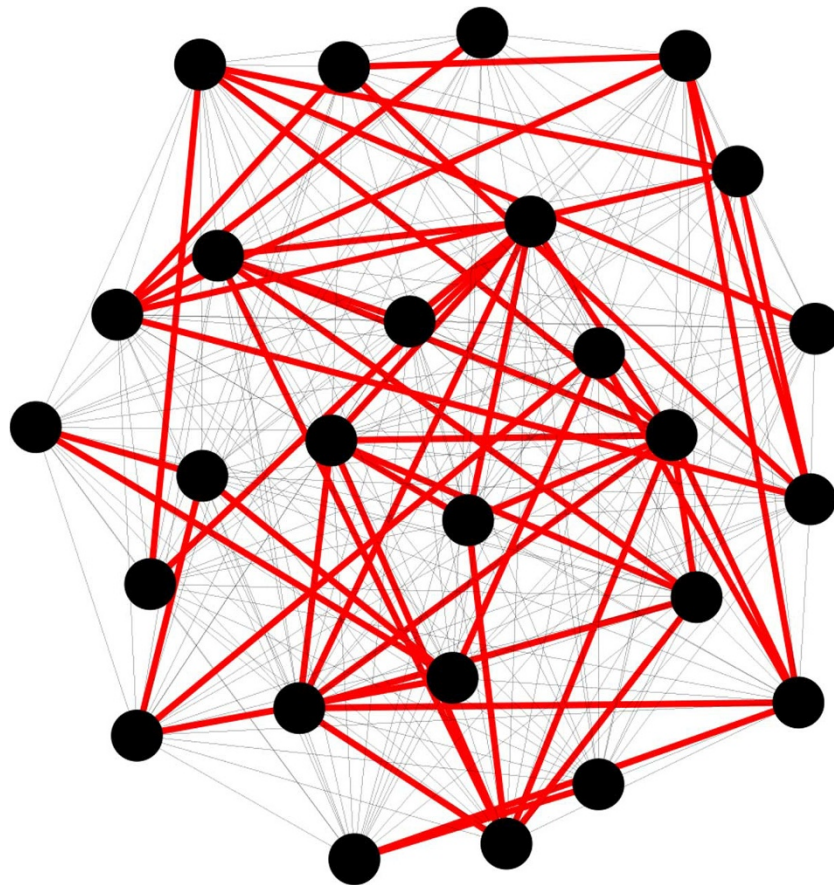
The SCWiB algorithm presented here shows how the clusters can be calculated naturally, despite two independent levels of control on cluster quality. This is an exact algorithm as it constructs all possible clusters and then picks the largest, the next disjoint largest and so on. It requires exponential time, since the number of possible clusters can be exponential. As displayed it is simple, but unnecessarily inefficient; the algorithm can be sped up enormously by reordering the vertices after a specified number of calls to the inner algorithm. It can also be made more efficient by temporarily postponing the construction of clusters that threaten to require excessive time, and by a number of other devices. For use on large genomes, it can be converted into a heuristic by replacing

the exhaustive exploration of all search paths by a large enough sample of these paths.

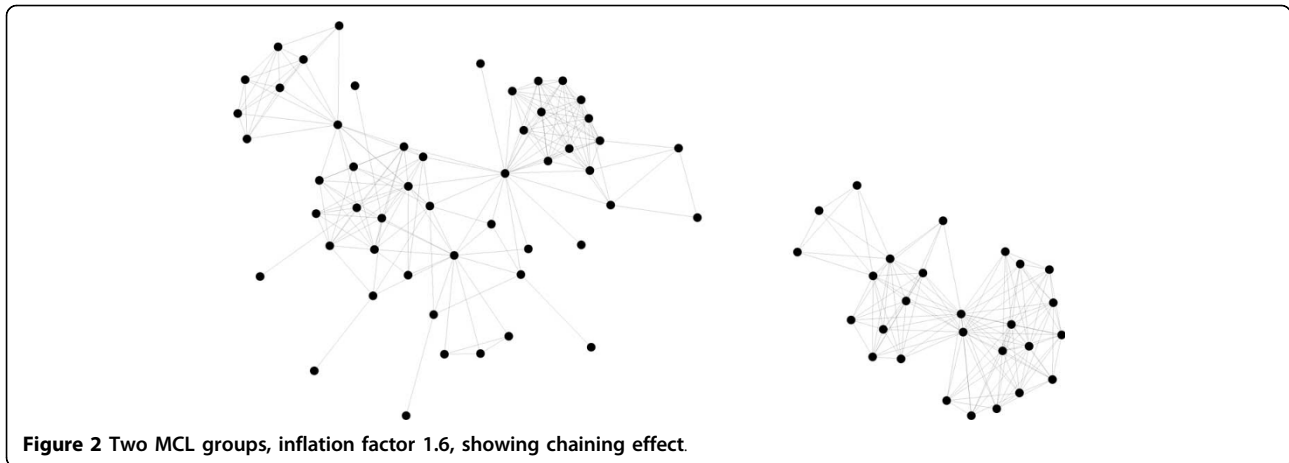
#### Comparison with MCL

Figure 1 shows an example of a SCWiB cluster emerging from an analysis of the *Arabidopsis thaliana* genome. It can be seen that although some of the genes are connected to the cluster by only one or two edges of similarity greater than  $U$ , any tendency to chaining is controlled by the  $S$ -plex condition at level  $W$ , with every vertex having a high degree within the cluster.

MCL [1,2] is one of the most widely used methods for inferring gene families. Its basic principle is the iteration of a procedure that strengthens certain heavily weighted edges and weakens those with lesser weight. With appropriate parameter settings, MCL and SCWiB can produce very similar distributions of cluster sizes. The lack of any cluster quality criterion influencing the MCL process, however, results in many of its clusters, including some of the largest ones, having very few internal edges, as in Figure 2, while the SCWiB construction explicitly prohibits this.



**Figure 1** SCWiB cluster containing part of the NAC transcription factor family [60]. Dots represent genes. Red edges constitute the "backbone" with similarity greater than  $U$ , black edges indicate similarity greater than  $W$ , less than  $U$ .



### Comparison with TAIR

The most comprehensive, though very incomplete, inventory of curated gene families for a plant pertains to the *Arabidopsis thaliana* genome [3]. This is a collection of gene families, found on the *Arabidopsis* Information Resource (TAIR) gene family page, contributed by individual scholars and groups, based largely on function within the cell as well as sequence similarities. It is not an attempt to partition the entire set of *Arabidopsis* genes into clusters, and there is no requirement that the families are disjoint. Furthermore, the functional groups are not intended to correspond perfectly with gene families as defined by common ancestry. Nonetheless, we compare these families with those produced by SCWiB. We find that many of the gene similarities in large functionally-determined families do not meet the SCWiB criteria, which therefore splits them into a number of subfamilies. The same holds for the comparison of the functional families with MCL clusters. This implies the limitations of purely similarity-based methods for gene family detection. Nevertheless, many functional families are in almost one-to-one correspondence with gene families determined by SCWiB.

In Figure 1, only part of the NAC transcription factor family is in the cluster; other parts are in other SCWiB clusters. This family has been diverging in the land plants long before the emergence of the angiosperms, so that different ancient NAC transcription factor subfamilies are not connected at the  $U = 70\%$  level that we used. Of interest is that in an MCL analysis of this same data, with inflation factor fixed at 1.6 to achieve the same total number of gene families as SCWiB, this cluster is fragmented among five MCL families, none of them containing more than nine of the 26 genes.

### The Angiosperm genomes

The emergence of new genes and new functions for existing genes is a major aspect of evolutionary divergence of species. In animals, especially the mammals, a

key mechanism for such innovation is alternative splicing, which affects at least 50% of genes [14]. In plants, however, this phenomenon is thought to be much less important, impacting just 5-10% [15], while the creation, expansion and attrition of gene families through mechanisms of gene duplication, notably tandem duplication and, more spectacularly, whole genome doubling, may spur rapid adaptation of populations to a broad range of niches. We extracted all the data available on angiosperm genomes in the CoGe database [16,17]. We required genomes to be published, publicly available, and have associated structural gene annotations. The genomes included *Amborella*, soybean, *Brachypodium distachyon*, *Setaria*, peach, cassava, *Capsella rubella*, sorghum, eucalyptus, common bean, grapevine, cacao, banana, turnip, papaya, *Arabidopsis thaliana*, tomato, potato, *Arabidopsis lyrata*, *Leavenworthia alabamica*, *Sisymbrium irio*, *Aethionema arabicum*, strawberry, *Thellungiella parvula*, watermelon, sacred lotus, *Utricularia*, *Spirodela polyrhiza*, date palm, pigeonpea, sweet orange, poplar, rice, *Ricinus communis*, clementine, lotus, flax, maize, cucumber, kiwifruit, *Mimulus*, *Medicago*, pepper and *Eutrema parvulum* [18-59]. We could not exercise any control on the quality of the sequencing, the assembly, or the annotation, and we will discuss the possible consequences of this on our results in the Conclusions.

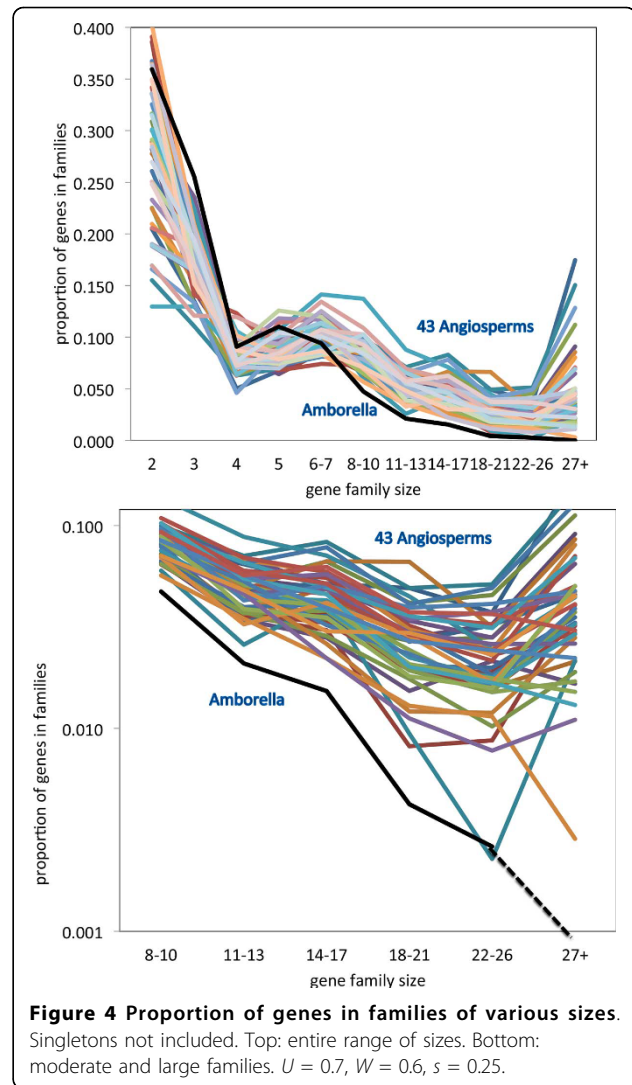
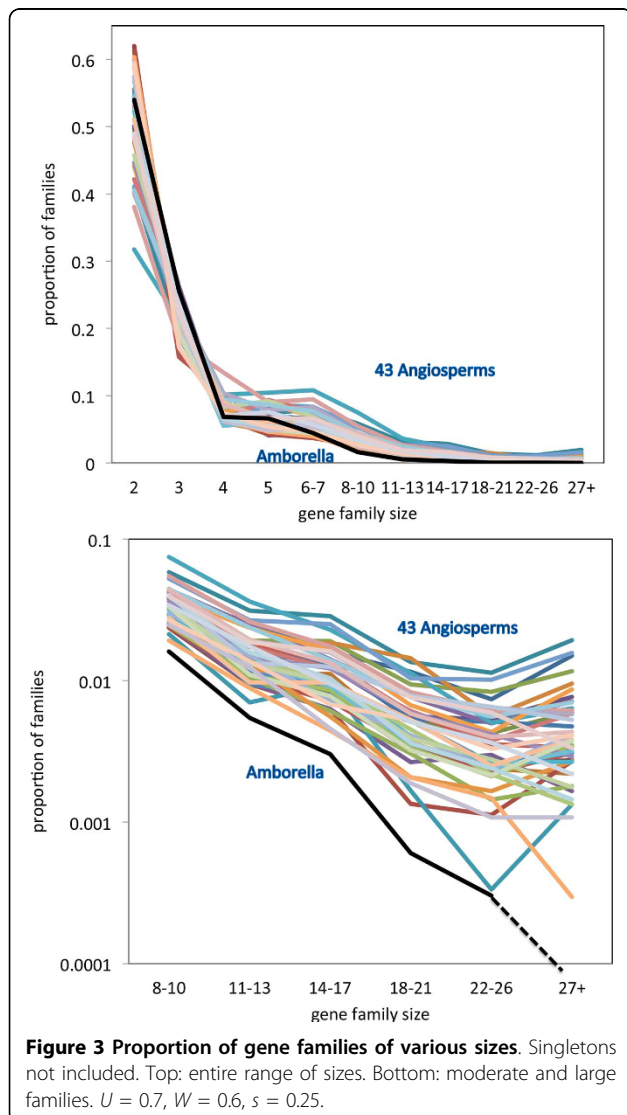
We used the SynMap tool in CoGe to run a comparison of each genome with itself in order to construct a complete set of gene duplicates. We disregarded syntenic context (pertinent only to WGD duplicates), by setting the minimum block length to 1. From the unfiltered results, we eliminated duplicates with similarities less than  $W = 0.6$ .

We decomposed the set of resulting set of duplicates into SCWiBs with parameters  $U = 0.7$ ,  $W = 0.6$ ,  $s = 0.25$ . We used a local optimization criterion, finding the largest possible SCWiB first, then re-applying the

method on successively small graphs that result from removing the vertices in the previously generated clusters. It should be noted that SynMap, as we used it produces a large peak of duplicate genes with similarities from 60-64. This had little if any consequence for our results, since almost all of these duplicates would be eliminated by the  $U$  criterion, although they could provide support for the  $S$ -plex criterion.

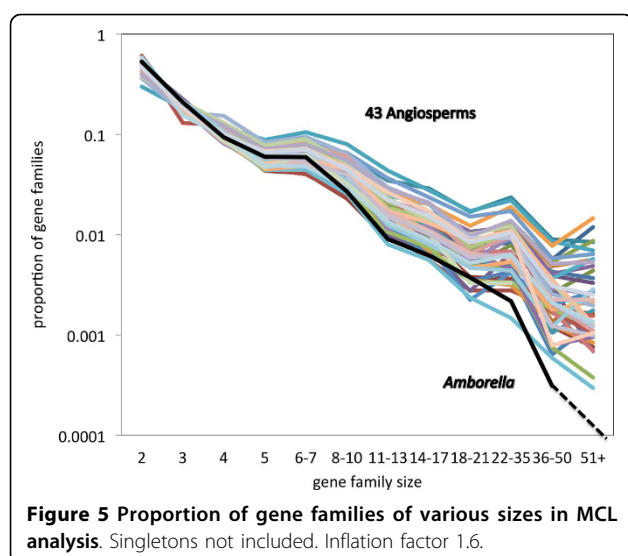
Based on  $U = 0.7$ , our gene families would largely have origins within the angiosperms, or be subfamilies of ancient plant gene families diversifying within the angiosperms.

Figure 3, displaying relative numbers of families of each size, and Figure 4, with the total number of genes in these families, show broadly similar gene family size distribution across the angiosperms, but also show a remarkable trend involving the *Amborella trchopoda*



genome. Whether we measure it according to number of gene families of a given size, or according to the proportion of genes in gene families of a given size, *Amborella* has fewer gene families of moderate (starting at 8-10 members) or of large size (22-26, 27 or more), than any of the other genomes.

As a control, we carry out an experiment on the same set of gene pairs for each genome, but using the MCL method. Exactly the same genes are involved. To ensure that the number of gene families were comparable, we used an inflation factor of 1.6 for the MCL. Figure 5 shows that distribution of MCL family sizes is more spread out than in the SCWiB case in Figure 3. However, the anomalous lack of large gene families of *Amborella* still stands out. This pattern emerges clearly, although the distinction is not as clear as with SCWiB. Another genome, cucumber, also has small numbers of moderate-sized families.



## Discussion and conclusion

The parameters of SCWiB directly control the connectedness and density characteristics of the clusters; we can predict the results of changes in each one. This contrasts with program parameters like the inflation factor in MCL, whose effects are largely indirect and unpredictable at the level of individual clusters. Although SCWiB involves three numerical factors, they enter into the algorithm in a simultaneous way to assure both connectedness and density. SCWiB clusters can also be generated by heuristics derived from generic search strategies such as branch-and-bound, and we have implemented this as a check on results from the **algorithm SCWiB**-derived heuristics.

The remarkably distinct pattern apparent in the *Amborella* distribution of gene family sizes will have to be validated in future studies. Most immediate is the role of the specific parameter values for  $U$ ,  $W$  and  $s$ . Does the pattern hold up when one or more of these are shifted? Preliminary results, not shown here, are positive: increasing  $s$  from 0.25 to 0.35 increases the number of larger gene families (size  $>26$ ) for all the genomes, but the distinction between *Amborella* (which only sees one family achieving a size of 30) and the other 43, is amplified. A systematic answer to this question will require considerable computing time to experiment with different values of  $U$  and  $W$ , but without any change in methodology. Another question is whether the pattern we observe is somehow dependent on the SCWiB definition. We have shown that the MCL method, which differs from SCWiB in almost every way possible, reproduces the distinct pattern of *Amborella* with respect to the other genomes, with almost no large gene families and a small number of moderate-sized one.

Another question arises because of the great heterogeneity of methods used over more than a decade of genome sequencing, particularly with regard to gene annotation. Most pertinent is the attention paid to identify gene families that are in fact families of transposons. And indeed, the annotation of the recently sequenced *Amborella* genome zealously pursued the identification and exclusion of such families from the set of bonafide gene families. Nevertheless, while this may have ensured a deficit of large families in the data from this genome, it could not account for the observed deficit in families with 8 to 27 genes.

Is the *Amborella* pattern phylogenetically significant? Most of the 43 other genomes are core eudicots, but there is a good number of Poaceae and other monocots, as well as the basal eudicot *Nelumbo*, and these all share the same pattern as the core eudicots. Sequenced genomes of other land plants, like *Selaginella* and *Pinus taeda* are not included in our analysis, and preliminary analyses show other, inconsistent, differences in family size distribution from the angiosperms, but no dearth of large gene families. There is thus no evidence that *Amborella* conserves some ancestral, pre-angiosperm pattern of gene family sizes, but this will question will require further genomic data to settle. A similar question, whether *Amborella* represents a pre-core eudicot pattern among angiosperms, will also require further data from other early branching plants, but already we know that *Nelumbo* as well as the monocots, all have the typical pattern. Another factor may lie in the fact that *Amborella* is the only genome to have escaped whole genome duplication since the origins of the angiosperms; this may also be associated with a lesser tendency to amplify and diversify gene families. Finally, the paucity of large families in *Amborella* may be an acquired feature, and not a conserved one. The current restricted ecological range of this plant may reflect a long history of isolation, of small populations, and little advantage to genetic innovation.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

All authors participated in the research, wrote the paper, read and approved the manuscript.

## Acknowledgements

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

## Declarations

The publication charges for this article were funded by the Canada Research Chair in Mathematical Genomics, and by the University of Ottawa. This article has been published as part of *BMC Genomics* Volume 15 Supplement 6, 2014: Proceedings of the Twelfth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on

Comparative Genomics. The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcgenomics/supplements/15/S6>.

#### Authors' details

<sup>1</sup>Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada, K1N 6N5. <sup>2</sup>Department of Plant Biology, University of Georgia, Athens, GA 30602-7271, USA. <sup>3</sup>The School of Plant Sciences, University of Arizona, Tucson, AZ 85721 USA.

Published: 17 October 2014

#### References

- Van Dongen S: Graph clustering by flow simulation, Ph. D. Thesis, University of Utrecht. 2000.
- Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 2002, **30**(7):1575-1584.
- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al: The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 2012, **40**(D1):1202-1210.
- Kelarev A, Kang B, Steane D: Clustering algorithms for ITS sequence data with alignment metrics. *AI 2006: Advances in Artificial Intelligence* Springer, ???; 2006, 1027-1031.
- NCBI: Blastclust: 2002.
- Sperisen P, Pagni M: JACOP: a simple and robust method for the automated classification of protein sequences with modular architecture. *BMC Bioinformatics* 2005, **6**(1):216.
- Kohonen T: Self-organizing Maps. In *Springer Series in Information Sciences. Volume 30*. Springer; 1995.
- Zahn CT: Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers* 1971, **100**(1):68-86.
- Joseph JM, Durand D: Family classification without domain chaining. *Bioinformatics* 2009, **25**(12):45-53.
- Joseph JM: On the identification and investigation of homologous gene families, with particular emphasis on the accuracy of multidomain families. PhD thesis, Carnegie Mellon University; 2012, Technical Report CMU-CB-12-103.pdf.
- Foggia P, Percannella G, Sansone C, Vento M: A graph-based algorithm for cluster detection. *International Journal of Pattern Recognition and Artificial Intelligence* 2008, **22**(05):843-860.
- Foggia P, Percannella G, Sansone C, Vento M: A graph-based clustering method and its applications. *Advances in Brain, Vision, and Artificial Intelligence* Springer; 2007, 277-287.
- Guo J, Komusiewicz C, Niedermeier R, Uhlmann J: A more relaxed model for graph-based data clustering: s-plex editing. *Algorithmic Aspects in Information and Management* Springer, ???; 2009, 226-239.
- Brett D, Pospisil H, Valcárcel J, Reich J, Bork P: Alternative splicing and genome complexity. *Nature Genetics* 2002, **30**(1):29-30.
- Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al: Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* 2003, **31**(19):5654-5666.
- Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, et al: Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiology* 2008, **148**(4):1772-1781.
- Lyons E, Freeling M: How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* 2008, **53**(4):661-673.
- Amborella Genome Project: The *Amborella* genome and the evolution of flowering plants. *Science* 2013, **342**(6165):1241089.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto MDG, Shu SDGKB, Futrell-Griggs MBA, Du J, Tian Z, Zhu L, et al: Genome sequence of the palaeopolyploid soybean. *Nature* 2010, **463**(7278):178-183.
- Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail L, Tice H, Grimwood J, Mckenzie N, Bevan MW, Huo N, Gu YQ, Lazo GR, Anderson OD, You FM, Luo M, Dvorak J, Wright J, Febrer M, Idziak D, Hasterok R, Linndquist E, Wang M, Fox SE, Priest HD, Filichkin SA, et al: Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 2010, **463**(7282):763-768.
- Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli M, Estep AC, Feng L, Vaughn JN, Grimwood J, Jenkins J, Barry K, Lindquist E, Hellsten U, Deshpande S, Wang X, Wu X, Mitros T, Triplett J, Yang X, Ye C, Mauro-Herrera M, Wang L, Li P, Sharma M, Sharma R, Ronald PC, Panaud O, Kellogg EA, Brutnell TP, Doust AN, et al: Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology* 2012, **30**(6):555-561.
- The International Peach Genome Initiative, Verde I, Abbott AG, Scalabrini S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, Zuccolo A, Rossini L, Jenkins J, Vendramin E, Meisel LA, Decroocq V, Sosinski B, Prochnik S, Mitros T, Policriti A, Cipriani G, Dondini L, Ficklin DM, an Goodstein A, Xuan P, Del Fabbro C, Aramini V, Copetti D, Gonzalez S, et al: The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics* 2013, **45**(5):487-494.
- Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S: The cassava genome: current progress, future directions. *Tropical Plant Biology* 2012, **5**(1):88-94.
- Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo Y, Steige K, Platts AE, Escobar JS, Newman LK, Wang W, Mandáková T, Vello E, Smith LM, Henz SR, Steffen J, Takuno S, Brandvain Y, Coop G, Andolfatto P, Hu M, Blanchette TT, Clark RM, Quesneville H, Nordborg M, Gaut BS, Lysak MA, Jenkins J, Grimwood J, Chapman J, et al: The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics* 2013, **45**(7):831-835.
- Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haber G, Hellsten U, Miltos T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Godwik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otiilar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, et al: The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 2009, **457**(7229):551-556.
- Myburg A, Grattapaglia D, Tuskan G, Jenkins J, Schmutz J, Mizrahi E, Hefer C, Pappas G, Sterck L, De Peer YV, Hayes R, Rokhsar D: The *Eucalyptus grandis* genome project: Genome and transcriptome resources for comparative analysis of woody plant biology. In *BMC Proceedings. Volume 5*. BioMed Central Ltd; 2011:20.
- Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MMS, Miklas PN, et al: A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics* 2014.
- Jailion O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Huguency P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, et al: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007, **449**(7161):463-467.
- Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gousy J, Alegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto F, an Sabot JF, Kudrna D, Ammiraju JSS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gellay L, Shi Z, Bérard A, et al: The genome of *Theobroma cacao*. *Nature Genetics* 2011, **43**(2):101-108.
- D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, B N, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardí C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam SRMM, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, et al: The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 2012, **488**(7410):213-217.
- The Brassica rapa Genome Sequencing Project Consortium, Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai YJHM, Bancroft IFC, Huang S, Li X,

- Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson B, Chalhouh AH, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Tong C, Song C, *et al*: **The genome of the mesopolyploid crop species *Brassica rapa***. *Nature Genetics* 2011, **43**(10):1035-1039.
32. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang BV, ad Ly W, Lewis KLT, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, *et al*: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)**. *Nature* 2008, **452**(7190):991-996.
33. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**(6814):796.
34. Tomato Genome Consortium, Sato S, Tabata S, Hirakawa H, Asamizu E, Shirsawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, Egholm M, Knight J, Bogden R, Li C, Han B, Feng Q, Fanm D, Shuang Y, Xu X, Pan S, Cheng S, Liu X, Ren Y, Wang J, Albiero A, Dal F, Todesco S, Van Eck J, Buels RM, *et al*: **The tomato genome sequence provides insights into fleshy fruit evolution**. *Nature* 2012, **485**(7400):635-641.
35. Potato Genome Sequencing Consortium, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zheng G, Yang S, Li R, Wand J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, Del la Cruz G, Chakrabarti SK, Patil VU, Skryabin BB, amd Kuznetsov KG, Ravin NV, Kolganova AVB, Mardanov AV, Di Genova A, Bolser DM, Martin DMA, Li G, Yang Y, *et al*: **Genome sequence and analysis of the tuber crop potato**. *Nature* 2011, **475**(7355):189-195.
36. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer JD, Hollister Gand, Ossowski S, Ottillar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KFX, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL: **The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change**. *Nature Genetics* 2011, **43**(5):476-481.
37. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, Dewar K, Stinchcombe JR, Schoen DJ, Wang X, Schmutz J, Town CD, Edger PP, Pires JC, Schumaker KS, Jarvis DE, Mandáková T, Lysak MA, van den Bergh E, Schranz ME, Harrison PM, Moses AM, Bureau TE, Wright AI, Blanchette M: **An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions**. *Nature Genetics* 2013, **45**(8):891-898.
38. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton JM, Rees DJG, Williams KP, Holt SH, Rojals JJR, Chatterjee M, *et al*: **The genome of woodland strawberry (*Fragaria vesca*)**. *Nature Genetics* 2011, **43**(2):109-116.
39. Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, Yun DJ, Bressan RA, Zhu JK, Bohnert HJ, Cheeseman JM: **The genome of the extremophile crucifer *Thellungiella parvula***. *Nature Genetics* 2011, **43**(9):913-918.
40. Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Yi Z, Mao L, Ren Y, Wang Z, Min J, Guo X, Murat F, Ham BK, Zhang Z, Gao S, Huang M, Xu Y, Zhong S, Bombarely A, Mueller LA, Zhao H, He H, Zhang Y, Zhang Z, Huang S, Tan T, Pan E, Lin K, Hu Q, *et al*: **The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions**. *Nature Genetics* 2013, **45**(1):51-58.
41. Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li LT, Zhang Q, Kim MJ, Schatz MC, Campbell M, Li J, Bowers JE, Tang H, Lyons E, Ferguson AA, Narzisi GRND, Blaby-Haas CE, Gschwend AR, Jiao Y, Der JP, Zeng F, Han J, Min XJ, Hudson KA, Singh R, Grennan AK, Karpowicz SJ, Watling JR, Ito K, *et al*: **Genome of the long-living sacred lotus (*Nelumbo nucifera Gaertn.*)**. *Genome Biology* 2013, **14**(5):41.
42. Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Anahí Pérez-Torres C, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juárez MJA, Simpson J, Fernández-Cortés A, Arteaga-Vázquez M, Góngora-Castillo A, Acevedo-Hernández G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Pérez SA, Ortega-Estrada MdJ, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L: **Architecture and evolution of a minute plant genome**. *Nature* 2013, **498**(7452):94-98.
43. Wang W, Haberer G, Gundlach H, Giasser C, Nussbaumer TCLM, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J, Byrant DW, Mockler TC, Appenroth KJ, Grimwood J, Jenkins J, Chow J, Choi C, Adam C, Cao XH, Fuchs J, Schubert I, Rokhsar D, Schmutz J, Michael TP, Mayer KFX, Messing J: **The *Spirodela polyrhiza* genome reveals insights into its neotenuous reduction fast growth and aquatic lifestyle**. *Nature Communications* 5(2014).
44. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chakuvadi S, Pontaroli AC, DeBarry J, Arondel V, Ohlrogge J, Saie Ul, Suliman-Elmeer KM, Bennetzen JL, Kruegger RM, Malek JA: **De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*)**. *Nature Biotechnology* 2011, **29**(6):521-527.
45. Varshney RK, *et al*: **Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers**. *Nature Biotechnology* 2012, **30**(1):83-89.
46. Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP, Chen J, Gao S, Xing F, Lan H, Chang JW, Ge X, Lei Y, Hu Q, Miao Y, Wang L, Xiao S, Biswas MK, Zeng W, Guo F, Cao H, Yang X, Xu XW, Cheng YJ, Xu J, Liu JH, *et al*: **The draft genome of sweet orange (*Citrus sinensis*)**. *Nature Genetics* 2013, **45**(1):59-66.
47. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Putnam N, Ralph S, Rombauts S, Salamov A, Schein JLS, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson JMC, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, *et al*: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)**. *Science* 2006, **313**(5793):1596-1604.
48. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia YPB, Zhong J, Miguel T, Paszkowski U, Zhang S, ColBERT MWL, Cooper B, Park S, Wood TC, Mao L, Quail P, *et al*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)**. *Science* 2002, **296**(5565):79-92.
49. Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD: **Draft genome sequence of the oilseed species *Ricinus communis***. *Nature Biotechnology* 2010, **28**(9):951-956.
50. Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrini S, Terol J, Takita MA, Labadie K, Poulain J, Couloux A, Jabbari K, Cattonaro F, Del Fabbro C, Pinosio S, Zuccolo A, Chapman J, Grimwood J, Tadeo FR, Estornell LH, Munõz-Sanz JV, Ibanez V, Herrero-Ortega A, Aleza P, Pérez-Pérez J, Ramón D, Brunel D, *et al*: **Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication**. *Nature Biotechnology* 2014.
51. Sato S, Nakamura Y, Kaneko T, Asamizu A, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, KAtoh M, Kohara M, kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, HAYashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S: **Genome structure of the legume, *Lotus japonicus***. *DNA Research* 2008, **15**(4):227-239.
52. Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Dalta R, Lambert G, Galbraith DW, Grassa CJ, Geraldies A, Cronk QC, Cullis C, Dash PK, Kumar PA, Cloutier S, Sharpe AG, Wong GKS, Want J, Deyholos MK: **The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads**. *The Plant Journal* 2012, **72**(3):461-473.
53. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zheng J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott M, Cotton RM, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
54. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, J H, Jia Z, Ren Y, G T, Lu Y, Ruan J, Qian W, Wang M, *et al*: **The genome of the cucumber, *Cucumis sativus* L.** *Nature Genetics* 2009, **41**(12):1275-1281.
55. Huang S, Ding J, Deng D, Tang W, Sun H, Liu D, Zhang L, Niu X, Zhang X, Meng M, Yu J, Liu J, Shi W, Zhang D, Cao S, Wei Z, Cui Y, Xia Y, Zeng H, Bao K, Lin K, Min Y, Zhang M, Miao , Tang X, Zhu Y, Sui Y, Li G, Sun H, *et al*: **Draft genome of the kiwifruit *Actinidia chinensis***. *Nature Communications* 4(2013).
56. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan T, Wessler SR, Schmutz J, Willis JH, Rokhsar DS: **Fine-scale variation in meiotic recombination in**



*Mimulus* inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences* 2013, **110**(48):19478-19482.

57. Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, Van der Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Nudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KAT, Tang H, Rombauts S, Zhao PX, Zhou P, et al: **The *Medicago* genome provides insight into the evolution of rhizobial symbioses.** *Nature* 2011, **480**(7378):520-524.
58. Qin C, Yu C, Shen Y, Fang X, Chen L, Min J, Cheng J, Zhao M, ad Xu S, Luo Y, Yang Y, Wu Z, Mao L, Wu H, Ling-Hu C, Zhou H, Lin H, González-Morales DL, ad Trejo-Saavedra S, Tian H, Tang X, Zhao M, Huang Z, Zhou A, Yao X, Cui J, Li W, Chen Z, Feng Y, Niu Y, et al: **Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization.** *Proceedings of the National Academy of Sciences* 2014, **111**(14):5135-5140.
59. Yang R, Jarvis E, David , Chen H, Beilstein MA, Grimwood J, Jenkins J, Shu S, Prochnik S, Xin M, Ma C, Schmutz J, Wing RA, Mitchell-Olds T, Schumaker KS, Wang X: **The reference genome of the halophytic plant *Eutrema salsugineum*.** *Frontiers in Plant Science* 2013, **4**.
60. Olsen AN, Ernst HA, Leggio LL, Skriver K: **NAC transcription factors: structurally distinct, functionally diverse.** *Trends in Plant Science* 2005, **10**(2):79-87.

doi:10.1186/1471-2164-15-S6-S8

**Cite this article as:** Zheng et al.: Gene families as soft cliques with backbones: *Amborella* contrasted with other flowering plants. *BMC Genomics* 2014 **15**(Suppl 6):S8.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

