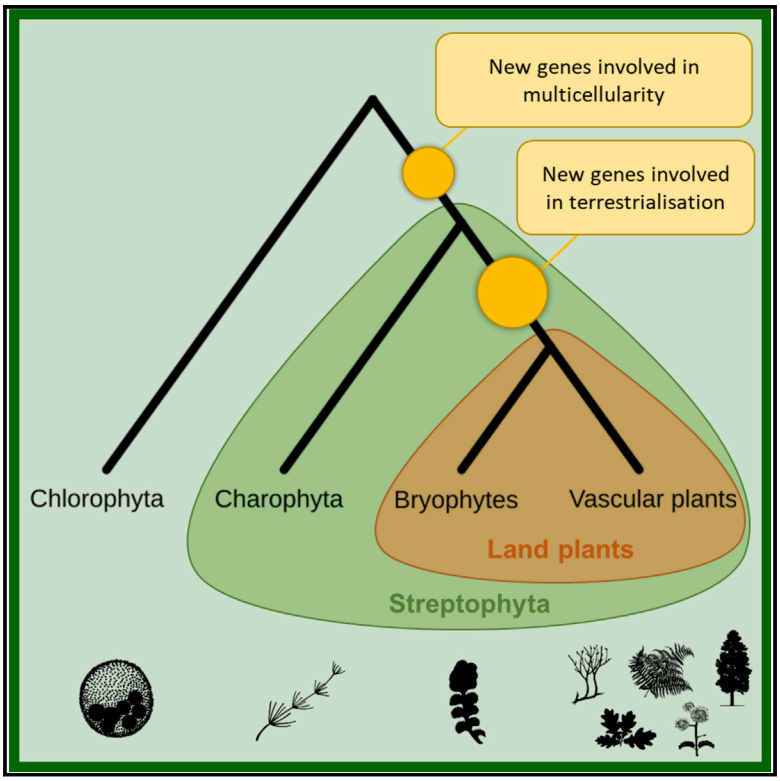
Report

Current Biology

The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty

Graphical Abstract



Highlights

- Comparing 208 genomes gives insight into the role of gene novelty in plant evolution
- Two bursts of genomic novelty played a major role in the evolution of land plants
- Functions linked to these novelties are multicellularity and terrestrialization
- The backbone of hormone signaling either predates or accompanies this transition

Authors

Alexander M.C. Bowles, Ulrike Bechtold, Jordi Paps

Correspondence

ubech@essex.ac.uk (U.B.), jordi.paps@bristol.ac.uk (J.P.)

In Brief

Bowles et al. show that two consecutive bursts of genomic novelty predate the origin of land plants. Identifying these events provides insights into the evolution of flora that has defined modern ecosystems.

Bowles et al., 2020, Current Biology *30*, 1–7
February 3, 2020
https://doi.org/10.1016/j.cub.2019.11.090

CellPress

Please cite this article in press as: Bowles et al., The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty, Current Biology (2019), https://doi.org/10.1016/j.cub.2019.11.090

Current Biology
Report

The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty

Alexander M.C. Bowles,[1] Ulrike Bechtold,[1,*] and Jordi Paps[1,2,3,4,*]
[1]School of Life Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
[2]School of Biological Sciences, University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
[3]Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK
[4]Lead Contact
*Correspondence: ubech@essex.ac.uk (U.B.), jordi.paps@bristol.ac.uk (J.P.)
https://doi.org/10.1016/j.cub.2019.11.090

SUMMARY

Over the last 470 Ma, plant evolution has seen major evolutionary transitions, such as the move from water to land and the origins of vascular tissues, seeds, and flowers [1]. These have resulted in the evolution of terrestrial flora that has shaped modern ecosystems and the diversification of the Plant Kingdom, Viridiplantae, into over 374,000 described species [2]. Each of these transitions was accompanied by the gain and loss of genes in plant genomes. For example, whole-genome duplications are known to be fundamental to the origins of both seed and flowering plants [3, 4]. With the ever-increasing quality and quantity of whole-genome data, evolutionary insight into origins of distinct plant groups using comparative genomic techniques is now feasible. Here, using an evolutionary genomics pipeline to compare 208 complete genomes, we analyze the gene content of the ancestral genomes of the last common ancestor of land plants and all other major groups of plant. This approach reveals an unprecedented level of fundamental genomic novelties in two nodes related to the origin of land plants: the first in the origin of streptophytes during the Ediacaran and another in the ancestor of land plants in the Ordovician. Our findings highlight the biological processes that evolved with the origin of land plants and emphasize the importance of conserved gene novelties in plant diversification. Comparisons to other eukaryotic studies suggest a separation of the genomic origins of multicellularity and terrestrialization in plants.

RESULTS AND DISCUSSION

Analyzing the Ancestral Plant Gene Content

Understanding the diversification of plant life on Earth is still one of the major challenges in evolutionary biology. Defining the genomic changes accompanying plant evolution is key to unraveling the molecular basis of biological innovations. Recent studies have used comprehensive taxonomic transcriptome data to understand angiosperm diversification rates and gene family expansion in the major plant groups [5, 6]. Furthermore, reduced genomic datasets have been used to investigate whole-genome duplications as well as gene family gains and losses associated with plant diversification [4, 7, 8]. However, the role of genomic novelty in the origins of distinct plant groups using an extensive sampling of complete genomes with a phylogenetically broad outgroup has not been fully evaluated.

Adapting a previously described [9, 10] comparative genomics pipeline, we compared 208 eukaryotic genomes, including a broad representation of animal (10), other unikont (11), and non-embryophyte bikont (29) genomes (STAR Methods; Data S1; Figure S1). Genome quality was assessed with BUSCO, discarding genomes with more than 15% of BUSCO missing genes, and protein sequences were compared using BLAST and MCL to identify homology groups (HGs). To reduce the error produced by the complex evolutionary dynamics of genes involved in these transitions, further dissection of HGs was not conducted [10, 11]. Therefore, a single HG is defined as a set of proteins that have distinctly diverged from others. The 208 eukaryotic genomes contain ~9 million proteins, which were clustered into ~650,000 HGs. Using scripts incorporating a phylogenetic framework to inform comparative genomics, five evolutionarily distinct classifications of HG (ancestral, ancestral core, novel, novel core, and lost) were extracted (Data S2; Figure S2). Based on these outputs, patterns of large gene gains and losses were identified across the plant phylogeny (Figure 1).

The HG categorization juxtaposes between the traditional gene classification (e.g., gene families and classes) and their evolutionary dynamics. Therefore, a HG can either contain genes traditionally designated as subfamilies (e.g., GA3ox), gene families (e.g., allene oxide cyclase), or gene superfamilies. This recovery of traditional gene classifications demonstrates the reliability of this clustering approach (Data S3). There are limitations shared with other BLAST-based analyses, such as the impact of gene fusion, fission, and lateral gene transfer. However, genes in broad HGs are less likely to be misassigned than orthologs and paralogs (e.g., OrthoMCL) [12]. The pipeline approach also tackles biases seen in tree reconciliation methods, which are prone to inaccurate assignments of gene gains and losses [13].

The Role of Highly Conserved Gene Groups in Plant Evolution

The evolutions of Embryophyta (land plants) and Streptophyta (land plants and their closest algal relatives, Charophyta) are arguably the most dramatic transitions in the history of plants.

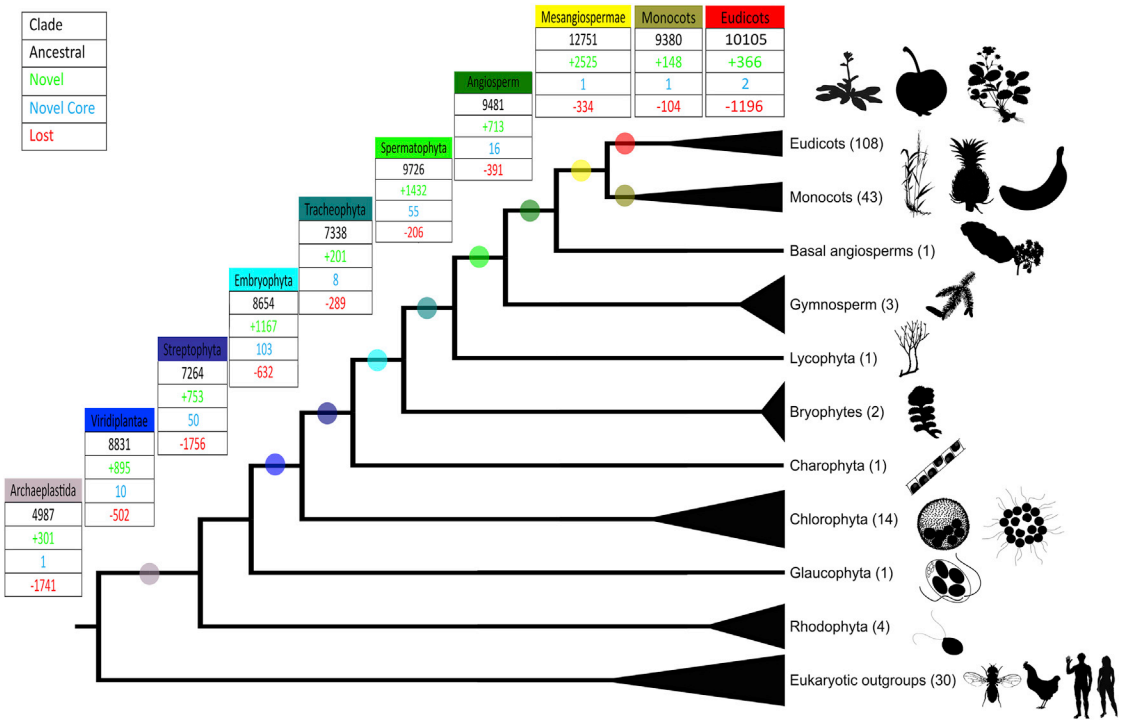


Figure 1. Analysis of the Gene Content of Ancestral Plant Genomes

The number of HGs of different categories indicated at each node for all major plant groups. Evolutionary relationships of these groups can be found in Data S1. Organism silhouettes were sourced from http://phylopic.org. See also Figures S1 and S2 and Table S1.

These events have previously been linked with the expansion of many processes and developmental traits, including embryogenesis [14], plant hormones [15], and symbiotic interactions with arbuscular mycorrhizae and rhizobacteria [16]. Our analyses revealed that there was a substantial increase in the number of highly retained gene novelties in the last common ancestor (LCA) of Streptophyta and the LCA of Embryophyta with 50 and 103 novel core HGs identified, respectively (Figure 1). Gene Ontology (GO) analyses using *Arabidopsis thaliana*, which has comprehensive GO annotations, were used to explore the modern functions of descendants of genes from novel core HGs (Data S4; Figure 2). The protein class category was used, as this classification is less prone to false assignments and biases [10]. All other GO categories, including molecular function, biological process, and pathway were produced (Data S4). HGs present in the LCA of embryophytes are abundant in classes involved in protein modification (e.g., transferase, oxidoreductase, and ligase) and protein transport (e.g., transporter proteins and membrane traffic proteins), whereas HGs present in the LCA of streptophytes are abundant in gene regulation (e.g., transcription factor) and cell structure, movement, and division (e.g., cytoskeletal proteins). The origins of Streptophyta were accompanied by the evolution of many plant-specific transcription factors (e.g., HD-ZIP) and an increasingly complex cell wall corresponding to the high number of the protein class hits seen in the Streptophyta novel core (NC) HGs [8, 14, 17].

It is possible that the bursts of conserved genomic novelty could be explained by the presence of one or multiple whole-genome duplications (WGDs). Inferring WGDs in these ancestral nodes is difficult with no events currently identified in the LCA of these groups [18, 19]. Analysis of over 1,000 transcriptomes has identified 244 WGDs across the green plant phylogeny [6]. These mostly occur after the origin of vascular plants and do not appear to coincide with the bursts of novelty seen in this study. This supports the theory that there was a change in strategy from gene family birth and expansion to WGD along the backbone of the plant phylogeny. Another contributing factor that might explain the origins of some novel core HGs is the presence of horizontal gene transfer (HGT). BLAST searches against the Swissprot database confirmed the absence of all novel core HGs in outgroup taxa, validating the outputs of the pipeline approach (BLAST outputs on Github: https://github.com/AlexanderBowles/Plant-Evomics/tree/master/Extended%20Data). Queries using the pipeline approach revealed that 323 HGs were present in fungal and land plant genomes but absent in all other taxa in this study's dataset (Data S1), suggesting widespread HGT in plants [20, 21]. The last eukaryotic common ancestor (LECA) is the ancestor that connects all eukaryotes, including plants and fungi. Either these HGs were in LECA and lost from all eukaryotic representatives aside from fungi and land plants or they are the product of HGT [22]. GO analysis of 25 of the HGs that contained at least 100 embryophyte taxa revealed that they were associated with gene regulation and protein modification (Data S5). Other possible HGT events that could explain the marked distribution of these novel core HGs include parasitism by other plants, symbiosis with other plants (e.g., transfer of a photoreceptor gene from bryophytes to ferns), and symbiosis with rhizobacteria [21, 23].

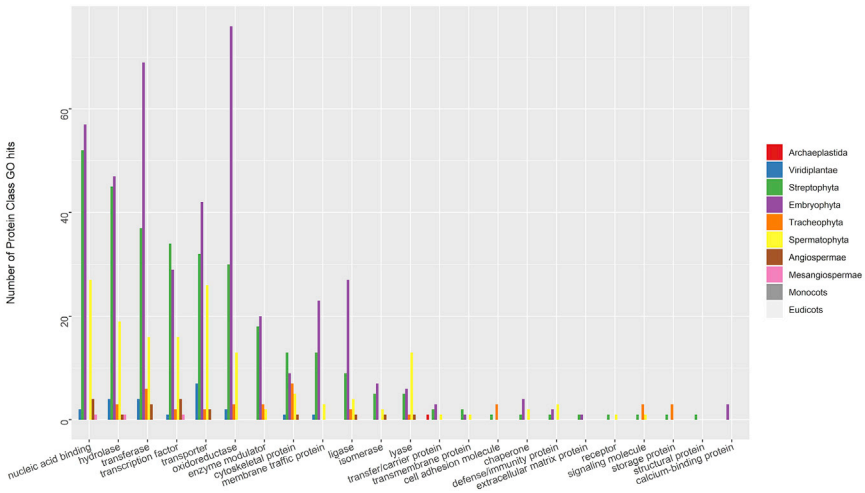


Figure 2. Gene Ontology Annotations of Novel Core HGs

Using *Arabidopsis thaliana* genes as an extant representative, protein classes were assigned for all novel core HGs. All other GO annotations (e.g., molecular function, biological process, cellular component, and pathways) were produced. See also Data S4.

The Functions of Highly Conserved Gene Groups

In streptophytes, novel core HGs were implicated in root, multicellular, and lateral organ development (Data S6; Figure 3). These terms were assigned based on the functions in extant *Arabidopsis thaliana* genes. In some cases, the evolutionary emergence of HGs predates the origin of the function with which they are often associated. For example, there is no evidence of roots outside Tracheophyta, yet genes associated with root development are found in older nodes [24, 25]. Therefore, these HGs are potential examples of co-option of old genes for new processes (Figure 3).

Other key functions include the increased complexity of the cell wall, which is crucial for multidimensional cell growth [26]. Further indicators of multicellularity in the predecessor of land plants are HGs involved in the regulation of transcription, cell adhesion, and division. The findings here also support an expansion of cellular signal transduction pathways associated with growth, development, and stress responses in streptophytes.

Many of the novel core HGs identified in our study have not previously been associated with the origin of land plants. These include proteins involved in plant organ development, cell wall construction, and host microbe interactions [27]. Other HGs are related to terrestrialization, with functions related to the synthesis of lignin, UV light protection, and cell signaling. The latter comprise plant hormones (phytohormones) linked with growth, such as auxin (body plan definition) [28], brassinosteroids (photomorphogenesis) [29], and gibberellins, as well as those associated with environmental responses, such as abscisic acid (ABA), salicylic acid, and jasmonic acid (primordial root growth) [30]. Several novel core HGs, including basic-helix-loop-helix (bHLH) transcription factors, receptor like kinases (LRR-RLKs), and three families of heavy-metal-associated isoprenylated plant proteins (HIPPs), have been previously linked to the origin of embryophytes, further validating our results (Data S6) [31].

The Evolution of Phytohormone Signaling

Some of these innovations have evolved in an incremental fashion. For example, phytohormone signaling genes identified as novel core to Streptophyta include ethylene-overproduction protein 1 (ETO1) and ethylene insensitive 3 (EIN3) (Figure S3). However, genes involved in ethylene signaling have been shown to originate before (1-aminocyclopropane-1-carboxylate synthase [ACS]) and after (1-aminocyclopropane-1-carboxylate oxidase [ACO]) this point in the evolutionary history of plants [14]. Therefore, these assigned functions do not demonstrate an establishment of these features but the additive developments contributing to their origin and evolution.

Using the same comparative genomics approach, we infer the evolutionary origins and conservation of phytohormone pathways in plants (Figure S3). The fundamental backbone of the biosynthesis and signaling pathways of all phytohormones either predates or accompanies the land plant transition [14, 32–34]. Genes involved in gibberellic acid production and signaling originate with plant terrestrialization (Figure 4). However, the role of hormones may have changed during land plant evolution, as recently highlighted for ABA signaling [39]. Important innovations in land plants include tightly controlled responses to drought and salt stresses, which require the production and perception of ABA. Our results show that ABA biosynthesis and perception evolved earlier than previously thought and are highly conserved across the plant phylogeny (Figure 4). The ABA receptor, PYL, has recently been identified in *Zygnema circumcarinatum* but is absent in other streptophyte algae [32]. In combination with the analysis presented here, this confirms that PYLs are conserved across Zygnematophyceae and Embryophyta. PP2Cs and SnRK2s, known to be present across Viridiplantae, are here supported as an Archaeplastida novelty [33]. Identifying these HGs is a significant step in understanding the evolution of phytohormones and their implications for plant diversification.

Other Evolutionarily Distinct Gene Groups of Ancestral Plant Genomes

Genomic novelty is considered to have an important role in the establishment of new features during the origins of land plants and other taxa. Genomic novelty in the LCA of distinct plant groups was substantial (Figure 1). In the LCAs of Streptophyta and Embryophyta, 753 and 1,167 novel HGs were identified, respectively, similar to values found in other studies (Data S4) [7, 14]. In contrast to other plant nodes, these values are relatively low compared to the 2,525 HGs identified in the origin of Mesangiospermae. As mentioned, WGD in plants is common and multiple events have been identified across the angiosperm phylogeny [3]. Two WGD events have been established in the ancestors of seed plants (Spermatophyta) and flowering plants

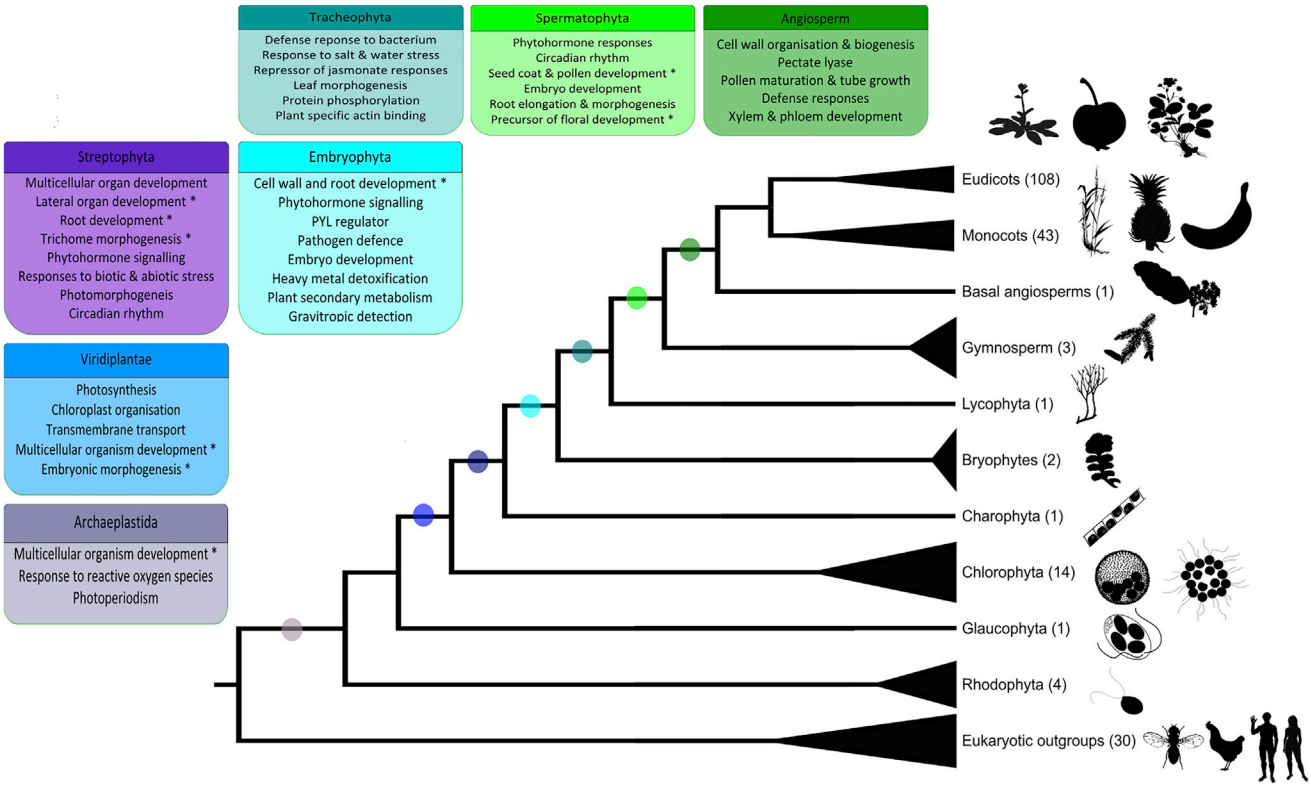


Figure 3. Biological Functions of Novel Core HGs based on *A. thaliana* Genes

Each box, color coded by phylogenetic group, is a summary of the modern day biological processes that are associated with each set of novel core HGs. An asterisk denotes an assigned biological term that is known to predate the origin of the function. Organism silhouettes were sourced from http://phylopic.org.

(Angiospermae), which could explain the 1,432 and 713 novel HGs identified in these nodes [40, 41].

Our analyses also identify that the LCA of extant land plants (Embryophyta) contained at least 8,654 ancestral HGs (Data S4). This number is likely lower than the total number of gene families present in the ancestral Embryophyta gene content because a HG can contain multiple genes, and HGs and genes can be lost from all extant representatives. *Arabidopsis thaliana* and *Brachypodium distachyon* genomes contain 27,655 and 34,310 genes clustered into 13,345 and 14,235 HGs, respectively, with 60%–70% of their genes present in the LCA of land plants. 2,254 of these ancestral HGs were retained (ancestral core) by at least 157 of the embryophyte genomes, demonstrating extensive gene loss has occurred across land plant evolution (Data S4). GO analysis revealed genes derived from HGs present in the LCA of embryophytes are abundant in gene regulation (e.g., nucleic acid binding and transcription factors) and protein modification (e.g., hydrolase and transferase; Data S4).

Furthermore, our analyses recognize HG losses (Data S4). *Drosophila melanogaster* was used as a representative of a well-annotated non-plant genome in the GO analyses of HGs lost in plant evolution. A total of 1,756 HGs were absent in the LCA of Streptophyta comprising protein classes involved in gene regulation (e.g., nucleic acid binding and transcription factor), cell signaling (e.g., enzyme modulator and signaling molecules), and catalytic activity (e.g., hydrolase and oxidoreductase). Lost HGs were also identified in Embryophyta, suggesting that gene turnover was prolific during the evolution of the ancestors of streptophytes and land plants (Figure 1). Large losses were also identified in branches leading to the LCA of eudicots and Archaeplastida with 1,196 and 1,741 HGs, respectively.

Comparisons with Animal Evolution

A previous study using the same comparative approach used in our study revealed an increase of genomic novelty during the origin of the animal kingdom, with an increase of conserved genomic novelty (novel core HGs) in a single node: the LCA of metazoans, which comprises 25 novel core HGs associated with multicellular processes; this represents a 5-fold increase from previous ancestors [10]. The origin of land plants shows two nodes with an increase of conserved genomic novelty: one in the LCA of streptophytes (in the Ediacaran; 629 mya) [1] and another in the LCA of land plants (Ordovician; 473 mya) [1]. Moreover, plants show higher numbers of conserved gene novelties than animals, representing a 10-fold increase compared to older ancestors (e.g., novel core HGs originating in the respective ancestors of Viridiplantae and Archaeplastida). In green plants, multicellularity has multiple independent evolutionary origins, with chlorophycean and charophycean algae showing a patchy

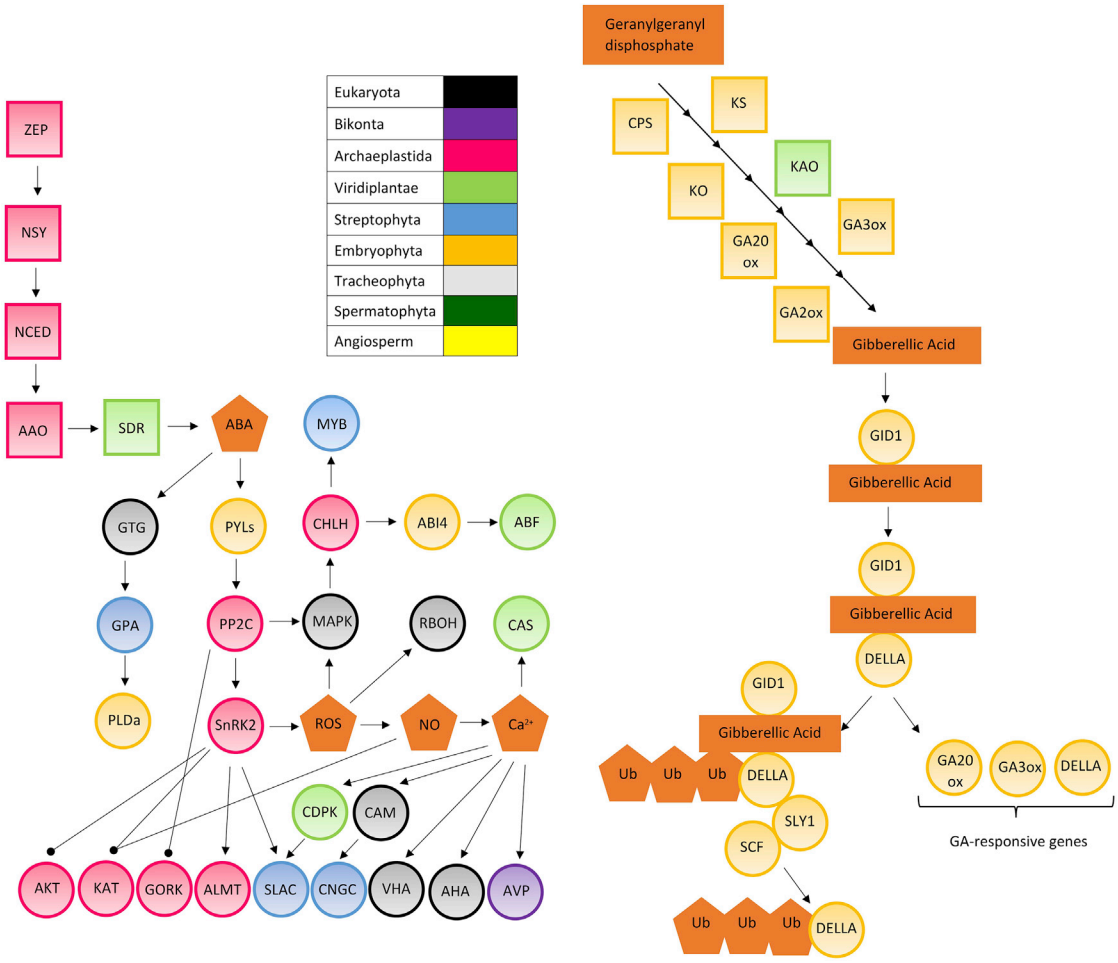


Figure 4. Evolution of Abscisic Acid (ABA) and Gibberellic Acid (GA) Biosynthesis and Signaling

Squares indicate genes that are involved in biosynthesis and circles indicate genes involved in signaling. Dark orange shapes indicate non-genetic elements. Color coding demonstrates that a gene was present in at least the last common ancestor of a clade. Arrows indicate positive regulation, and circle ended lines indicate negative regulation. Acronyms for genes: ABA biosynthesis: *AAO*, *ABA-ALDEHYDE OXIDASE*; *NCED*, *9-CIS-EPOXYCAROTENOID DIOXYGENASE*; *NSY*, *NEOXANTHIN SYNTHASE*; *SDR*, *SHORT-CHAIN ALCOHOL DEHYDROGENASE/REDUCTASE*; *ZEP*, *ZEAXANTHIN EPOXIDASE*. ABA signaling: *ABF*, *ABA RESPONSIVE ELEMENT-BINDING FACTOR*; *ABI4*, *ABA INSENSTIVE4*; *AHA*, *ARABIDOPSIS PLASMA MEMBRANE H^+-ATPASE*; *AKT*, *SER/THR KINASE1*; *ALMT*, *ALUMINUM-ACTIVATED MALATE TRANSPORTER*; *AVP*, *ARABIDOPSIS VACUOLAR H^+-PYROPHOSPHATASE*; *CAS*, *CALCIUM SENSING RECEPTOR*; *CHLH*, *PROTOPORPHYRIN IX MAGNESIUM CHELATASE*, *SUBUNIT H*; *CNGC*, *CYCLIC NUCLEOTIDE GATED CHANNEL*; *GORK*, *GATED OUTWARDLY RECTIFYING K^+ CHANNEL*; *KAT*, *GUARD CELL INWARDLY RECTIFYING K^+ CHANNEL*; *MAPK*, *MITOGEN ACTIVATED KINASE-LIKE PROTEIN*; *MYB*, *MYB DOMAIN PROTEIN*; *PLDa1*, *PHOSPHLIPASE Dα1*; *PP2C*, *PROTEIN PHOSPHATASE 2C*; *RBOH*, *RESPIRATORY BURST OXIDASE HOMOLOG PROTEIN*; *SLAC*, *SLOW ANION CHANNEL*; *VHA*, *VACUOLAR H^+-ATPASE*. GA biosynthesis: *CPS*, *ENT-COPALYL DIPHOSPHATE SYNTHASE*; *KS*, *KAURENE SYNTHASE*; *KO*, *ENT-KAURENE OXIDASE*; *KAO*, *ENT-KAURENOIC ACID OXIDASE*; *GA20ox*, *GIBBERELLIN 20 OXIDASE 1*; *GA3ox*, *GIBBERELLIN 3-BETA-DIOXYGENASE*; *GA2ox*, *GIBBERELLIN 2-BETA-DIOXYGENASE*. GA signaling: *GID1*, *GIBBERELLIN-INSENSITIVE DWARD PROTEIN 1*, *DELLA*; *SLY1*, *SLEEPY1*; *SCF*, *SKP1-CULLIN-F-BOX*. This figure has been adapted from previous publications for ABA [35, 36] and GA [37, 38]. See also Figure S3.

distribution, but is a trait that is conserved in all embryophytes [42, 43]. Here, we speculate that analysis of the gene content of the ancestral genomes of the plant kingdom (Viridiplantae) supports a decoupling between the emergence of multicellularity (streptophytes) and terrestrialization (embryophytes), which is in contrast to a single burst of novelty in the animal kingdom (Metazoa), whose origins did not involve a change of environment. In the future, the inclusion of new genomes may change the reconstruction of HGs at each node. Specifically, recent sequencing of the first two fern genomes and a second charophyte genome would help to fill phylogenetic gaps [7, 14]. Results from BLAST searches of novel core HGs against these phylogenetically important genomes supported the pipeline outputs, further validating our analyses (BLAST outputs available on Github: https://github.com/AlexanderBowles/Plant-Evomics/tree/master/Extended%20Data). In addition, this study solely focuses on protein-coding genes; however, non-coding genes, regulatory regions, and epigenetic modifications most likely contributed to the diversification of plant life. The analysis presented here, which incorporates genomic data for 208 taxa from across the tree of life, provides new insight into the composition of ancestral

plant genomes and emphasizes the role of genome evolution in the emergence of terrestrial flora.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Compiling genomic dataset
  - Homology assignment
  - Phylogenetically Aware Parsing Script
  - Novel Core HG validation
  - Functional annotation
  - Inferring Horizontal Gene Transfer
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2019.11.090.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Phillip Mullineaux, Prof. Leonard Schalkwyk, Prof. Peter W.H. Holland, Prof. Phillip Donoghue, Dr. Nacho Maeso, Dr. Ferdinand Marlétaz, and Dr. Sarah F. Worsley for their comments on the manuscript. We also would like to thank Stuart Newman for his support of the Genomics HPC server. A.M.C.B., U.B., and J.P. received funding from the School of Life Sciences (University of Essex). U.B. was in part funded by BBSRC grant BB/N016831/1.

AUTHOR CONTRIBUTIONS

Conceptualization, A.M.C.B., U.B., and J.P.; Formal Analysis, A.M.C.B.; Visualization, A.M.C.B.; Writing – Original Draft, A.M.C.B., U.B., and J.P.; Writing – Review & Editing, A.M.C.B., U.B., and J.P.; Supervision, U.B. and J.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 13, 2019
Revised: October 15, 2019
Accepted: November 29, 2019
Published: January 16, 2020

REFERENCES

1. Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., and Donoghue, P.C.J. (2018). The timescale of early land plant evolution. Proc. Natl. Acad. Sci. USA *115*, E2274–E2283.
2. Christenhusz, M.J.M., and Byng, J.W. (2016). The number of known plant species in the world and its annual increase. Phytotaxa *261*, 201–217.
3. Clark, J.W., and Donoghue, P.C.J. (2018). Whole-genome duplication and plant macroevolution. Trends Plant Sci. *23*, 933–945.
4. Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. *24*, 1334–1347.
5. Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C., and Soltis, P.S. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. *105*, 348–363.
6. Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., et al.; One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature *574*, 679–685.
7. Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.-M., Eily, A., Koppers, N., Kuo, L.-Y., Li, Z., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat. Plants *4*, 460–472.
8. Wilhelmsson, P.K.I., Mühlich, C., Ullrich, K.K., and Rensing, S.A. (2017). Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. Genome Biol. Evol. *9*, 3384–3397.
9. Dunwell, T.L., Paps, J., and Holland, P.W.H. (2017). Novel and divergent genes in the evolution of placental mammals. Proc. Biol. Sci. *284*, 20171357.
10. Paps, J., and Holland, P.W.H. (2018). Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat. Commun. *9*, 1730.
11. Holland, P.W.H., Marlétaz, F., Maeso, I., Dunwell, T.L., and Paps, J. (2017). New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20150480.
12. Pett, W., Adamski, M., Adamska, M., Francis, W.R., Eitel, M., Pisani, D., and Wörheide, G. (2019). The role of homology and orthology in the phylogenomic analysis of metazoan gene content. Mol. Biol. Evol. *36*, 643–649.
13. Hahn, M.W. (2007). Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol. *8*, R141.
14. Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K.K., Haas, F.B., Vanderstraeten, L., Becker, D., Lang, D., et al. (2018). The Chara genome: secondary complexity and implications for plant terrestrialization. Cell *174*, 448–464.e24.
15. Wang, C., Liu, Y., Li, S.-S., and Han, G.-Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. Plant Physiol. *167*, 872–886.
16. Field, K.J., Pressel, S., Duckett, J.G., Rimington, W.R., and Bidartondo, M.I. (2015). Symbiotic options for the conquest of land. Trends Ecol. Evol. *30*, 477–486.
17. Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., et al. (2014). Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. Nat. Commun. *5*, 3978.
18. Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. Nat. Rev. Genet. *18*, 411–424.
19. Zwaenepoel, A., and Van de Peer, Y. (2019). Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. Mol. Biol. Evol. *36*, 1384–1404.
20. Lutzoni, F., Nowak, M.D., Alfaro, M.E., Reeb, V., Miadlikowska, J., Krug, M., Arnold, A.E., Lewis, L.A., Swofford, D.L., Hibbett, D., et al. (2018). Contemporaneous radiations of fungi and plants linked to symbiosis. Nat. Commun. *9*, 5451.
21. Yue, J., Hu, X., Sun, H., Yang, Y., and Huang, J. (2012). Widespread impact of horizontal gene transfer on plant colonization of land. Nat. Commun. *3*, 1152.
22. Margulis, L., Chapman, M., Guerrero, R., and Hall, J. (2006). The last eukaryotic common ancestor (LECA): acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. Proc. Natl. Acad. Sci. USA *103*, 13080–13085.
23. Wickell, D.A., and Li, F. (2020). On the evolutionary significance of horizontal gene transfers in plants. New Phytol. *225*, 113–117.

24. Raven, J.A., and Edwards, D. (2001). Roots: evolutionary origins and biogeochemical significance. J. Exp. Bot. *52*, 381–401.
25. Hetherington, A.J., and Dolan, L. (2018). Stepwise and independent origins of roots among land plants. Nature *561*, 235–238.
26. Becker, B., and Marin, B. (2009). Streptophyte algae and the origin of embryophytes. Ann. Bot. *103*, 999–1004.
27. Raffaele, S., Mongrand, S., Gamas, P., Niebel, A., and Ott, T. (2007). Genome-wide annotation of remorins, a plant-specific protein family: evolutionary and functional perspectives. Plant Physiol. *145*, 593–600.
28. Finet, C., Berne-Dedieu, A., Scutt, C.P., and Marlétaz, F. (2013). Evolution of the ARF gene family in land plants: old domains, new tricks. Mol. Biol. Evol. *30*, 45–56.
29. Zhu, J.-Y., Sae-Seaw, J., and Wang, Z.-Y. (2013). Brassinosteroid signalling. Development *140*, 1615–1620.
30. Briggs, G.C., Mouchel, C.F., and Hardtke, C.S. (2006). Characterization of the plant-specific BREVIS RADIX gene family reveals limited genetic redundancy despite high sequence conservation. Plant Physiol. *140*, 1306–1316.
31. Liu, P.L., Du, L., Huang, Y., Gao, S.M., and Yu, M. (2017). Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. BMC Evol. Biol. *17*, 47.
32. de Vries, J., Curtis, B.A., Gould, S.B., and Archibald, J.M. (2018). Embryophyte stress signaling evolved in the algal progenitors of land plants. Proc. Natl. Acad. Sci. USA *115*, E3471–E3480.
33. Bowman, J.L., Kohchi, T., Yamato, K.T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., et al. (2017). Insights into land plant evolution garnered from the Marchantia polymorpha genome. Cell *171*, 287–304.e15.
34. Ju, C., Van de Poel, B., Cooper, E.D., Thierer, J.H., Gibbons, T.R., Delwiche, C.F., and Chang, C. (2015). Conservation of ethylene as a plant hormone over 450 million years of evolution. Nat. Plants *1*, 14004.
35. Cai, S., Chen, G., Wang, Y., Huang, Y., Marchant, D.B., Wang, Y., Yang, Q., Dai, F., Hills, A., Franks, P.J., et al. (2017). Evolutionary conservation of ABA signaling for stomatal closure. Plant Physiol. *174*, 732–747.
36. Yamauchi, Y., Ogawa, M., Kuwahara, A., Hanada, A., Kamiya, Y., and Yamaguchi, S. (2004). Activation of gibberellin biosynthesis and response pathways by low temperature during imbibition of Arabidopsis thaliana seeds. Plant Cell *16*, 367–378.
37. Freschi, L. (2013). Nitric oxide and phytohormone interactions: current status and perspectives. Front. Plant Sci. *4*, 398.
38. Middleton, A.M., Úbeda-Tomás, S., Griffiths, J., Holman, T., Hedden, P., Thomas, S.G., Phillips, A.L., Holdsworth, M.J., Bennett, M.J., King, J.R., and Owen, M.R. (2012). Mathematical modeling elucidates the role of transcriptional feedback in gibberellin signaling. Proc. Natl. Acad. Sci. USA *109*, 7571–7576.
39. McAdam, S.A.M., Brodribb, T.J., Banks, J.A., Hedrich, R., Atallah, N.M., Cai, C., Geringer, M.A., Lind, C., Nichols, D.S., Stachowski, K., et al. (2016). Abscisic acid controlled sex before transpiration in vascular plants. Proc. Natl. Acad. Sci. USA *113*, 12862–12867.
40. Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.
41. Ruprecht, C., Lohaus, R., Vanneste, K., Mutwil, M., Nikoloski, Z., Van de Peer, Y., and Persson, S. (2017). Revisiting ancestral polyploidy in plants. Sci. Adv. *3*, e1603195.
42. De Clerck, O., Kao, S.-M., Bogaert, K.A., Blomme, J., Foflonker, F., Kwantes, M., Vancaester, E., Vanderstraeten, L., Aydogdu, E., Boesger, J., et al. (2018). Insights into the evolution of multicellularity from the sea lettuce genome. Curr. Biol. *28*, 2921–2933.e5.
43. Umen, J.G. (2014). Green algae and the origins of multicellularity in the plant kingdom. Cold Spring Harb. Perspect. Biol. *6*, a016170.
44. Catarino, B., Hetherington, A.J., Emms, D.M., Kelly, S., and Dolan, L. (2016). The stepwise increase in the number of transcription factor families in the Precambrian predated the diversification of plants on land. Mol. Biol. Evol. *33*, 2815–2819.
45. Banks, J.A., Nishiyama, T., Hasebe, M., Bowman, J.L., Gribskov, M., dePamphilis, C., Albert, V.A., Aono, N., Aoyama, T., Ambrose, B.A., et al. (2011). The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science *332*, 960–963.
46. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
47. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.
48. Rambaut, A.D.A. (2012). FigTree version 1.4. http://tree.bio.ed.ac.uk/software/figtree.
49. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.
50. The Inkscape Project (2019). Inkscape v0.92.4. https://inkscape.org/.
51. Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. *28*, 45–48.
52. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. *45* (D1), D183–D189.
53. R Development Core Team. (2014). R: A language and environment for statistical computing (R Foundation for Statistical Computing).
54. Wickham, H., and Henry, L. (2018). tidyr: easily tidy data with "spread()" and "gather()" functions. https://cran.r-project.org/package=tidyr.
55. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
SWISSPROT	[46]	https://www.uniprot.org/
Genome Data	Data S1	N/A
Software and Algorithms		
BUSCO v3	[35]	https://busco.ezlab.org/
BLAST 2.7	[36]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
mcl-14-137	[37]	https://micans.org/mcl/
Phylogenetic Aware Parsing Script	[10]	https://github.com/PapsLab/Phylogenetic_Aware_Parsing_Script
Panther GO v11	[47]	http://www.pantherdb.org/
R 3.4.2; R - tidyr; R - GGplot2	[44, 45, 48]	https://www.r-project.org/
PAPS Plant-Evomics	https://github.com/AlexanderBowles/Plant-Evomics	N/A

LEAD CONTACT AND MATERIALS AVAILABILITY

Genome sources and software (e.g., BLAST) are listed (Data S1) and referenced (Figure S1) and all scripts used are available on Github listed below. Further information about the study and intermediary files (BLAST and MCL outputs) should be directed to the Lead Contact, Jordi Paps (jordi.paps@bristol.ac.uk). This study did not generate any new, or unique reagents.

METHOD DETAILS

Compiling genomic dataset

A detailed description of the pipeline utilized here can be found elsewhere [10]. Briefly, the pipeline uses the protein coding genes of whole genome sequences to identify homologous groups of proteins within and between species (Figure S1). Broad taxonomic sampling of genomic data was implemented to be able to accurately infer the phylogenetic origin of different HGs (Data S1). 208 eukaryotic genomes were downloaded equating to 9,204,593 predicted proteins including 178 Archaeplastida genomes (including 158 land plant genomes) and 30 from a diverse representation of eukaryotic outgroups (Data S1). BUSCO analysis was used to assess the quality of the genome annotation, using a < 15% of missing genes in the BUSCO Eukaryota dataset as a benchmark to accept a genome for further analysis (Data S1) [49].

Homology assignment

Sequence similarity for all predicted proteins was identified with an all-versus-all BLAST [46] (version 2.7.1) using an e-value of 10^{e-5}, resulting in 84,724,532,295,649 comparisons with 3,680,714,880 significant BLAST hits. The BLAST search was launched on 7th February 2018 and therefore any genomes published after this date were not included in the analysis. Within the MCL protocols, it is recommended to assess the effects of changing of the granularity score which is the fineness of the clusters produced [47]. Outputs for granularity scores 1.2, 2, 4 and 6 were used to compare the phylogenetic appearance and clustering of plant gene families against published datasets of Banks et al. [45] and the transcription factor families from Catarino et al. [44] (Data S3). After testing the impact of altering this inflation value, BLAST outputs were clustered using MCL with the default granularity score (I = 2.0, Data S3) [47]. This approach identified 661,545 groups of homologous genes across all proteins.

Phylogenetically Aware Parsing Script

The MCL output was processed by modifying the Perl scripts described in Paps and Holland [10] with Perl version 5. In the form of three Perl scripts, the pipeline can be used to identify the origin or loss of homologous groups of proteins (HGs) based upon their taxonomic occupancy (Data S2). Different sets of HGs can be analyzed (initially defined in Paps and Holland [10]);

- Ancestral (HGs present in the Last Common Ancestor of a clade),
- Ancestral Core (HGs present in every representative species within a clade or absent only in one genome),
- Novel (HGs present in the Last Common Ancestor of a clade and absent in all outgroup taxa),

- Novel Core (HGs present in every representative species within a clade or absent only once and absent in all outgroup taxa),
- Lost (HGs lost in the Last Common Ancestor of a clade).

A more detailed explanation of these query terms with examples is available (Figure S2). The main tree figures were made in FigTree [48] and edited in Inkscape [50].

Novel Core HG validation

To confirm accurate identification of conserved gene novelties, *Arabidopsis thaliana* (and *Brachypodium distachyon* for Liliopsida novelties) genes for each HG were tested, by performing BLASTP searches against the Swissprot database [51] (25th July 2018) excluding in-group sequences with the option negative_gilist [46]. This offers the maximum breadth of taxonomic sampling possible. Based on sequence similarity, e-value, and taxonomic occupancy, BLAST searches further validated the identification of novel core Homology Groups.

Three evolutionarily significant genomes have recently been published, the first two fern genomes [7] and the second charophyte genome [14]. Novel Core HGs from all groups were BLASTP searched against the protein coding genes of these genomes (Data S4). Based on sequence similarity, e-value, and taxonomic occupancy, these BLAST searches refined the number of Novel Core HGs identified (Table S1).

Functional annotation

To obtain a functional description for all types of HG for every Archaeplastida node, their *Arabidopsis thaliana* genes were assessed using Panther GO [52] (Version 11). The number of Gene Ontology hits for all GO classifications were collated: Protein Class, Molecular Function, Biological Process, Cellular Component, Pathways (Data S4). A literature search further revealed the functions of the Novel Core Homology Groups (Data S6). Graphics were produced in R [53] using packages tidyr [54] and GGplot2 [55].

Inferring Horizontal Gene Transfer

Inferences about potential HGT were made. Based on the taxon sampling in the dataset, the pipeline was used to produce the query: Atleast1-fungi present, Atleast1-Embryophyta present and Outgroups absent. 323 HGs were identified which were subsequently whittled down to 25 HGs by stipulating that at least 100 land plant taxa must be present. Similar to the above, GO analysis was used to reveal the functions of these HGs (Data S5).

DATA AND CODE AVAILABILITY

All genomic data used in the study is publically available with sources listed in Data S1. The code used to process the outputs of MCL and extract the 5 evolutionarily distinct Homology Groups is available on Github at https://github.com/AlexanderBowles/Plant-Evomics. Also available on Github are the BLASTs of all Novel Core HGs against the SwissProt database and the results of the BLASTs against the protein coding genes of *Chara braunii, Azolla filiculoides* and *Salvinia cucullata* (https://github.com/AlexanderBowles/Plant-Evomics/tree/master/Extended%20Data).