Received: 16 March 2023 | Accepted: 24 July 2023

DOI: 10.1002/tpg2.20377

The Plant Genome

TECHNICAL ADVANCE

The conservation of gene models can support genome annotation

Cassandria G. Tay Fernandez[1] | **Philipp E. Bayer**[1] | **Jakob Petereit**[1] | **Rajeev Varshney**[2,3] | **Jacqueline Batley**[1] | **David Edwards**[1]

[1]School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, Western Australia, Australia

[2]State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, Western Australia, Australia

[3]Centre of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Telangana, India

Correspondence
David Edwards, School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia.
Email: dave.edwards@uwa.edu.au

Assigned to Associate Editor Nils Stein.

Funding information
Australian Research Council, Grant/Award Numbers: DP200100762, DP210100296

Abstract

Many genome annotations include false-positive gene models, leading to errors in phylogenetic and comparative studies. Here, we propose a method to support gene model prediction based on evolutionary conservation and use it to identify potentially erroneous annotations. Using this method, we developed a set of 15,345 representative gene models from 12 legume assemblies that can be used to support genome annotations for other legumes.

1 | INTRODUCTION

The evolution of flowering plants can be traced back to the Cretaceous period around 125–100 million years ago, with subsequent divergence and adaptation to a wide variety of habitats (Crane & Lidgard, 1989; Lupia et al., 1999). However, many structures and much of the biochemistry remain common across plant species, which reflects shared common gene sets that encode heritable features of these shared traits. Genes are classically understood to be heritable sequences that impact the characteristics of an organism when expressed (Johannsen, 1909; Schnable, 2020). However, the impact of gene expression can be difficult to assess for some genes due to redundancy and highly specialized roles. For example, the phenotype of stress-related genes may only be observed when stress is present and at specific developmental stages (Macneil & Walhout, 2011). In contrast to the classical definition of a gene, a gene model describes a region of a genome that is transcribed as RNA and translated into protein or one of the classes of noncoding RNA genes. A gene by this definition does not need to play a role in specifying the characteristics of an organism (Gerstein et al., 2007).

Many protein-coding genes share sequence identity between species and can be structurally and functionally traced back to common ancestors. Large numbers of unique genes for species would suggest that *de novo* gene birth occurs at a high rate following speciation. Additionally, these novel genes must also be lost rapidly following diversification and the evolution of subsequent species, as their persistence would lead them to no longer be species-specific. Two lines of evidence suggest that this is not the case. First, while gene birth and gene death do play vital roles in evolution, genomic studies of closely related species do not identify significant births of novel functional genes (Armisén et al., 2008;

Abbreviations: BLAST, Basic Local Alignment Search Tool; bp, base pairs; CDS, coding sequence; GC, guanine-cytosine; NCBI, National Centre for Biotechnology Information; NR, non-redundant.

Demuth & Hahn, 2009; Gu et al., 2002; Tian et al., 2009), In plants such as *A, thaliana* and *O. sativa*, many species-specific genes have been found in close relatives (Armisén et al., 2008). Second, examination of closely related species suggests that much of the gene content is conserved following speciation; this is apparent in the *Brassica* genus, where comparison of *B. oleracea, B. rapa*, and *B. napus* shows a high level of conservation for the composition and order of genes across their genomes (Golicz et al., 2016; Rana et al., 2004). The lack of evidence for the birth and death of large numbers of genes associated with speciation suggests that the majority of unique gene models predicted in individual plant genomes may be artifacts of the annotation process and are unlikely to be functional.

Methods for the prediction of gene models in a genome assembly vary depending on the resources available and the purpose of the resulting predicted gene set. A strict annotation may be applied where false-positive gene models may negatively influence subsequent analysis, while more relaxed annotation methods may be applied if the purpose is to capture all potential gene structures. Some pipelines predict gene models ab initio based on sequence features such as open reading frames JIGSAW (Allen & Salzberg, 2005) and EvidenceModeler (Haas et al., 2008), while other pipelines, such as MAKER2 (Holt & Yandell, 2011), require gene expression or other forms of external evidence that can vary depending on the availability of data (Wang et al., 2004). Genes can also be predicted through homology-based gene prediction, which involves aligning a target genome to an annotated genome (Keilwagen et al., 2016). This comparative method exploits the fact that coding genes are typically well conserved and assumes that genes that share significant sequence similarity have identical functions (Brent, 2005; Keilwagen et al., 2016; König et al., 2018; Sharma et al., 2017).

Here, we propose an additional method to support gene model prediction based on evolutionary conservation and comparative gene prediction. We use this approach to identify potentially erroneous annotations in 12 legume genome assemblies and establish a comparative gene set, a tool to help and improve gene model annotations.

2 | MATERIALS AND METHODS

To identify non-conserved gene models, the annotated genes for each assembly and whole genome sequences were downloaded from https://data.legumeinfo.org/ (Table 1). Each annotated coding sequence (CDS) was compared with the genome sequence for each of the other 12 species using a basic local alignment search tool (BLAST) v2.11.0+ (e-value < 0.05). Genes that aligned with genes in other species were considered conserved sequences. If a gene did not match with any of the other genomes (e-value < 0.05), the gene was compared with the National Centre for Biotechnology Information–non-redundant (NCBI-NR) database using BLASTx v2.11.0+ (e-value < 0.05; accessed June 15, 2022; Altschul et al., 1990), and genes that still failed to find a match were considered non-conserved gene models. Genes that did have a match with the NCBI-NR database with a plant other than itself were included in the conserved sequences. RepeatMasker v4.1.5, rmblastn 2.14.0, and database CONS-DFam with RepBase 3.7 (species was set to Viridiplantae) were used to see the identity of 6899 genes only present in *P. sativum*. The same parameters were used to identify TE domains in both the conserved and the non-conserved genes.

Core Ideas

- If a gene model is only found in one species, it is more likely to be an annotation artifact than a gene model found across species.
- A total of 15,345 representative gene models from 12 legume assemblies can support genome annotations for other legumes.
- Representative gene models can be established for any species to improve genome annotation.

To identify if a predicted CDS is conserved, the predicted gene sequences were compared with the whole genome sequences of the other species using BLASTx v2.11.0+ (e-value < 0.05) to avoid potential errors from differences in the annotation of the other genomes. The representative gene set was assembled by first constructing a network using networkx. Genes were represented as nodes and edges were represented as BLASTx hits (Hagberg et al., 2008). The networks were constructed using the scripts selfblast_hsp_filter.R, network_gml.py, python edgelist_generator.py, and legume_rep_gene_filter.R available at GitHub; https://github.com/AppliedBioinformatics/legume_gene_count. The gene with the most alignments in each gene cluster was designated as a representative gene, and genes sharing edges were removed. Then, the following gene with the most alignments would be designated as a representative gene, and those edges would be removed until only single nodes remained, thereby sequentially removing the redundant sequences. Gene features were visualized using ggplot2 v3.3 (Wickham, 2016).

3 | RESULTS AND DISCUSSION

We first compared all annotated genes for each assembly with the whole genome sequence of each of the other assemblies. Gene models that had no significant sequence identity

TABLE 1 The assembly versions and references of each legume used to establish a legume representative gene set.

Legume species	Assembly version	Citation
Cajanus cajan	ICPL87119.gnm1.ann1.Y27M	(Varshney et al., 2012)
Cicer arietinum	CDCFrontier.gnm1.GkHc	(Varshney et al., 2013)
Glycine max	Wm82.gnm1.ann1.DvBy	(Schmutz et al., 2010)
Lotus japonicus	MG20.gnm3.ann1.WF9B	(Sato et al., 2008)
Lupinus angustifolius	Tanjil.gnm1.ann1.nnV9	(Hane et al., 2017)
Medicago truncatula	Mt4.0v2	(Tang et al., 2014)
Phaseolus vulgaris	G19833.gnm1.ann1.pScz	(Schmutz et al., 2014)
Pisum sativum	Cameor.gnm1.ann1.7SZR	(Kreplak et al., 2019)
Trifolium pratense	MilvusB.gnm2.ann1.DFgp	(De Vega et al., 2015)
Vigna angularis	Gyeongwon.gnm3.ann1.3Nz5	(Kang et al., 2015)
Vigna radiata	VC1973A.gnm6.ann1.M1Qs	(Kang et al., 2014)
Vigna unguiculata	IT97K-499-35.gnm1.ann1.zb5D	(Lonardi et al., 2019)

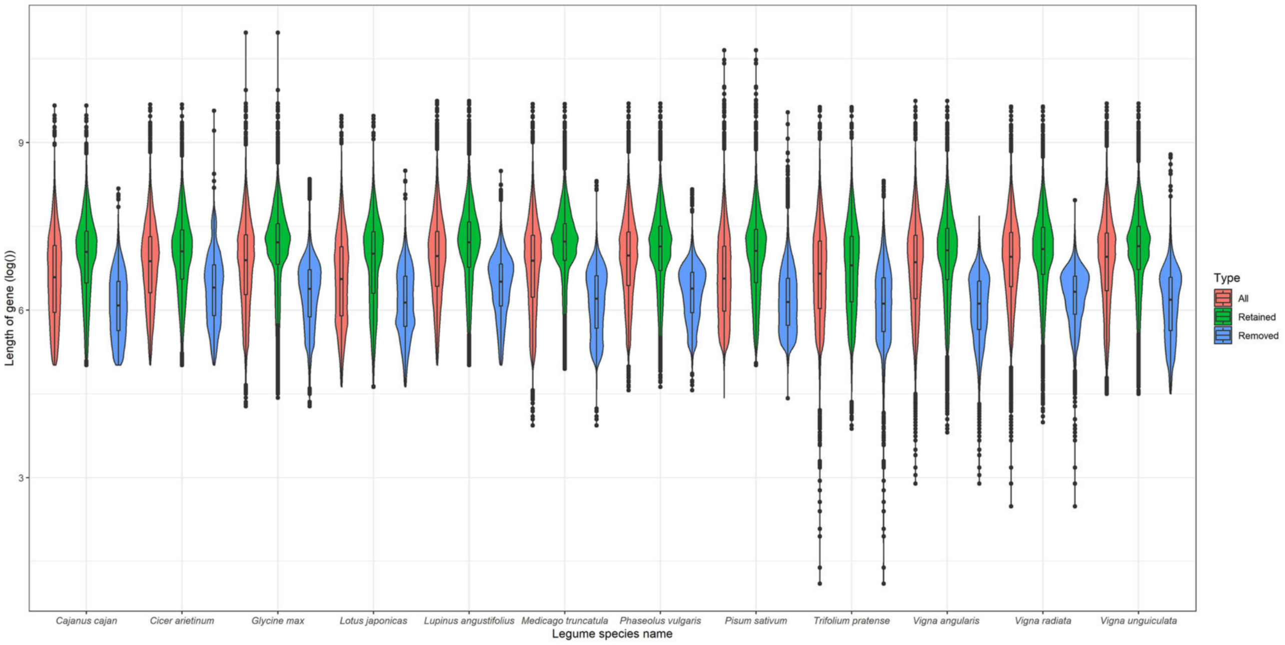

FIGURE 1 Comparison of gene length between 12 different legume species based on the reference sequence. The spread of gene lengths were taken for all genes (red) and the genes after filtering. If a gene shared sequence identity with the BLAST NCBI-NR database, or another legume, the gene was retained (green). Genes that did not were removed (blue).

with any of the other genomes were then compared with the NCBI-NR database (NCBI Resources Coordinators, 2018), and genes that still failed to find a match were considered species-specific or potential misannotations (Figure 1; Table 2). This process led to the identification of between 18 genes in *Cajanus cajan* and 6899 genes for *Pisum sativum* that are not present in the other legume genomes or NCBI-NR. Genes with no hits in legumes but hits in NCBI-NR were mostly contaminated. For example, in the *C. cajan* annotation, 80 gene models with no hits in the other legume genomes matched *Acinetobacter* sp. gene models deposited in NCBI-NR.

Many *P. sativum* genes were not found in other legumes (Table 2). This was attributed to the pea reference genome used, which focused on capturing as many genes as possible and used AUGUSTUS and Fgnesh to identify gene models. However, these programs were trained on the *M. truncatula* gene matrix (Kreplak et al., 2019; Tang et al., 2014) hence many gene models were called for the pea genome, even if there was less confidence in the genes (Kreplak et al., 2019).

TABLE 2 A summary of legume gene models present within 12 legume species.

Species	Number of reference gene models	Number of gene models with no identity with the other legume genomes	Number of gene models with no identity with other legume genomes and no match in NCBI-NR
Cajanus cajan	40,071	100	18
Cicer arietinum	28,269	1149	1143
Glycine max	54,175	726	504
Lotus japonicus	48,097	2721	2670
Lupinus angustifolius	33,070	608	582
Medicago truncatula	38,256	2062	1165
Phaseolus vulgaris	27,197	68	67
Pisum sativum	44,756	6955	6899
Trifolium pratense	39,916	1012	851
Vigna angularis	26,845	131	130
Vigna radiata	22,364	29	29
Vigna unguiculata	29,771	382	377

Abbreviations: NCBI, National Centre for Biotechnology Information; NR, non-redundant.

The 6899 genes were examined through RepeatMasker and most genes were not repeats. While, on average, 2.8% of genes were covered with repeats, only 68% of those genes were covered by more than 90% by TEs. The most common repeat type was simple repeats, with 84 genes being identified as such. InterProScan was used to look for TE-related gene candidates but found no genes with TE-like domains amongst the low-confidence pea genes. In contrast, InterProScan did identify genes with TE-like domains in the representative gene set.

Separately, 6899 genes were compared to the Yang et al. (2022) pea reference using blastn (e-value 1e-10). The Yang et al. (2022) assembly use an independent, *de novo* annotation, and we assumed that genes that did not have a match with the other legumes would be misannotations and would not have a match with this assembly. A total of 1007 genes had a hit in the new annotation with an e-value < 1e-10. Of those 1007 hits, 144 genes had 100% identity but only 14 had a 100% query coverage and 100% identity. Note that, 12 of 14 genes were not covered by any repeats and two genes were covered by 21% and 16% of repeats. *Pisum sativum* has more genes because the annotation parameters used gene were relaxed, allowing for more potential gene candidates to be captured. Under scrutiny, only 14 genes were able to be identified as “real genes”. These genes are flagged for further study.

We compared the potential false-positive gene models with conserved gene models. On average, the potential false-positive gene models (removed genes) were shorter than the retained genes (median 510 bp compared with 1209 bp, $p < 0.05$ *t*-test; Figure 1). This is most prominent in *T. pratense, V. angularis*, and *V. radiata* where the lower threshold of the gene length of all genes corresponds with the lower threshold of removed genes. Across all species, longer genes tend to be retained (Figure 1). The retained genes had a similar guanine-cytosine (GC) content (median 42.3% compared with 42.6%) to the conserved gene models (Figure S1). The similar GC content suggests that false positive gene models are not always due to contamination with DNA from other species during the genome assembly process.

To establish a non-redundant representative set of conserved gene models, we first removed non-conserved gene models and then used a graph-based approach to identify representative gene models across the 12 genomes. This resulted in a total of 15,345 representative gene models that would produce at least one significant match when searched with each conserved gene model from each of the 12 genomes. For example, repeats and gene families that share similar sequences would thus be represented by a single representative in the final set.

Studying the representative set (made from conserved genes) showed that 9% of genes were masked. Of the 15,345 genes, 292 genes were covered by repeats (at least 90%). These repeats are mostly Gypsy/Copia and the presence of these repeats could be attributed to R-genes which are mostly made up of repeats. We looked at the non-conserved genes and could not identify any TEs genes, likely because TEs are usually removed from annotation and assembly. Of the 6954 non-conserved pea genes, 4048 genes were identified as single-copy and 2858 were identified as paralogs (48 non-conserved genes likely being contaminations, e-value > 0.01).

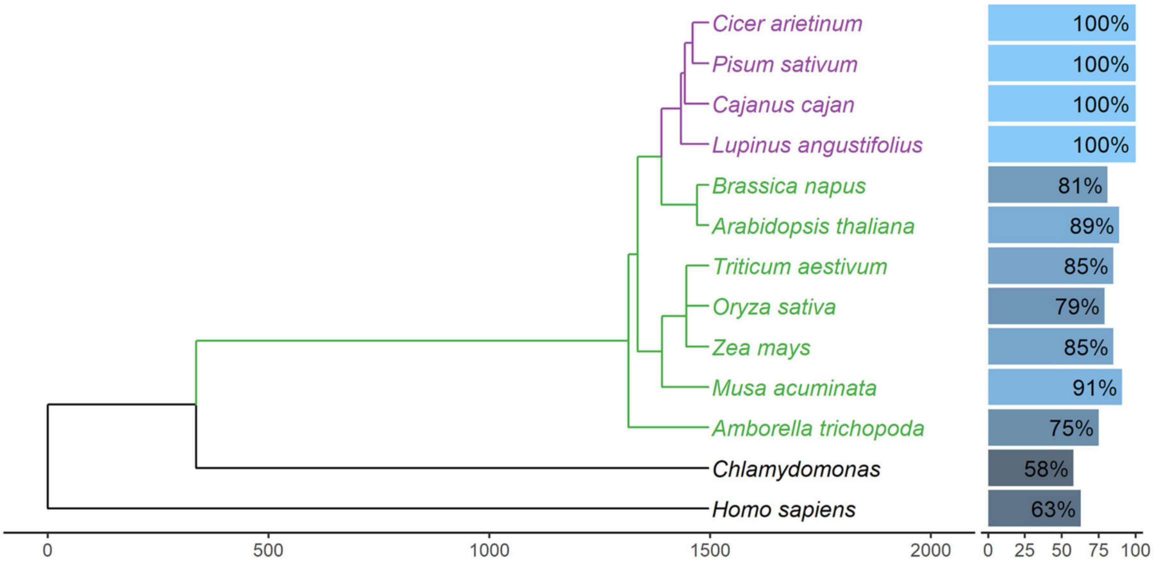

FIGURE 2 Percentage of plant, algae, and eukaryote annotation gene sets aligning with the representative legume gene set. Tip labels of legumes are purple, tip labels of angiosperms are green, and the remainder are black. The phylogeny is based on Kumar et al. (2017) with the *x* axis showing divergence time in Mya.

Similar proportions were found through the non-conserved genes of the other species.

We compared the annotated gene models from other plant species, *Chlamydomonas reinhardtii*, and *Homo sapiens* with the representative gene models to assess how representative the gene set is across kingdoms (Figure 2). As expected, within the 12 legumes the gene sets align 100% with the conserved gene models, with fewer genes aligning with increasing evolutionary distance.

Brassica napus is more closely related to legumes than *Musa acuminata* but has less sequence identity than *M. acuminata* (Figure 2). This was attributed to the annotation of *B. napus* used which reported 101,040 annotated genes (Hurgobin et al., 2018). We compared the representative legumes to a more recent *B. napus* annotation (Lee et al., 2020) using tblastx (e-value 1e-2) and found that only 72.26% of genes had sequence similarity (e-value 1e-02). We assume this is also caused by the annotation being more relaxed than for the other species aiming to predict all possible gene models and accepting a relatively high rate of erroneous gene calls.

The establishment of representative gene sets supports standardized annotation procedures for related species and robust comparative genomic analysis. While a predicted gene having sequence identity with a representative gene does not necessarily support a functional role, the prediction of a novel gene that lacks sequence identity to the representative set suggests that it may be an artifact and should be examined more closely for evidence to support a role (RNA-seq, proteomic data, etc.). Other approaches are required to assess if a gene is functional, as predicted genes may have a sequence identity to pseudogenes or gene fragments. This representative gene set is valuable to support comparative genomic analysis among the 12 legumes studies here, but additional representative genes may be included with the expansion of high-quality genome assemblies for application to a broader range of species.

AUTHOR CONTRIBUTIONS

Cassandria G. Tay Fernandez: Investigation; writing—original draft. **Philipp E. Bayer**: Investigation; writing—review and editing. **Jakob Petereit**: Investigation; writing—review and editing. **Rajeev K. Varshney**: Writing—review and editing. **Jacqueline Batley**: Project administration; writing—review and editing. **David Edwards**: Conceptualization; funding acquisition; project administration; writing—review and editing.

ACKNOWLEDGMENTS

This work is funded by the Australia Research Council (Projects DP210100296, DP200100762, and DE210100398). This work was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

A fasta file of the representative legume proteins is available at https://dx.doi.org/10.26182/n6b5-zx38.

ORCID

Philipp E. Bayer https://orcid.org/0000-0001-8530-3067
Rajeev Varshney https://orcid.org/0000-0002-4562-9131
Jacqueline Batley https://orcid.org/0000-0002-5391-5824
David Edwards https://orcid.org/0000-0001-7599-6760

REFERENCES

Allen, J. E., & Salzberg, S. L. (2005). JIGSAW: Integration of multiple sources of evidence for gene prediction. *Bioinformatics*, *21*(18), 3596–3603. https://doi.org/10.1093/bioinformatics/bti609

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Armisén, D., Lecharny, A., & Aubourg, S. (2008). Unique genes in plants: specificities and conserved features throughout evolution. *BMC Evolutionary Biology*, *8*, 280–280. https://doi.org/10.1186/1471-2148-8-280

Brent, M. R. (2005). Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Research*, *15*(12), 1777–1786. https://doi.org/10.1101/gr.3866105

Crane, P. R., & Lidgard, S. (1989). Angiosperm diversification and paleolatitudinal gradients in cretaceous floristic diversity. *Science*, *246*(4930), 675–678. https://doi.org/10.1126/science.246.4930.675

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å., Rognli, O. A., Jones, C., Swain, M., Geurts, R., Lang, C., Mayer, K. F. X., Rössner, S., Yates, S., Webb, K. J., Donnison, I. S., Oldroyd, G. E. D., Wing, R. A., Caccamo, M., … Skøt, L. (2015). Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Scientific Reports*, *5*(1), Article 17394. https://doi.org/10.1038/srep17394

Demuth, J. P., & Hahn, M. W. (2009). The life and death of gene families. *BioEssays*, *31*(1), 29–39. https://doi.org/10.1002/bies.080085

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., & Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, *17*(6), 669–681. https://doi.org/10.1101/gr.6339607

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*(1), Article 13390. https://doi.org/10.1038/ncomms13390

Gu, Z., Cavalcanti, A., Chen, F. C., Bouman, P., & Li, W. H. (2002). Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Molecular Biology and Evolution*, *19*(3), 256–262. https://doi.org/10.1093/oxfordjournals.molbev.a004079

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hagberg, A., Swart, P., & Chult, S. D. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of 7th Python in Science Conference (SciPy2008)* (pp. 11–15). SciPy.

Hane, J. K., Ming, Y., Kamphuis, L. G., Nelson, M. N., Garg, G., Atkins, C. A., Bayer, P. E., Bravo, A., Bringans, S., Cannon, S., Edwards, D., Foley, R., Gao, L.-L., Harrison, M. J., Huang, W., Hurgobin, B., Li, S., Liu, C.-W., McGrath, A., … Singh, K. B. (2017). A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: Insights into plant–microbe interactions and legume evolution. *Plant Biotechnology Journal*, *15*(3), 318–330. https://doi.org/10.1111/pbi.12615

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491. https://doi.org/10.1186/1471-2105-12-491

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-K. K., Tirnaz, S., Dolatabadian, A., Schiessl, S. V., Samans, B., Montenegro, J. D., Parkin, I. A. P., Pires, J. C., Chalhoub, B., King, G. J., Snowdon, R., Batley, J., & Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal*, *16*(7), 1265–1274. https://doi.org/10.1111/pbi.12867

Johannsen, W. (1909). *Elemente der exakten Erblichkeitslehre*. Fischer.

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K., Jun, T. H., Hwang, W. J., Lee, T., Lee, J., Shim, S., Yoon, M. Y., Jang, Y. E., Han, K. S., Taeprayoon, P., Yoon, N., Somta, P., Tanya, P., Kim, K. S., … Lee, S.-H. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature Communications*, *5*(1), Article 5443. https://doi.org/10.1038/ncomms6443

Kang, Y. J., Satyawan, D., Shim, S., Lee, T., Lee, J., Hwang, W. J., Kim, S. K., Lestari, P., Laosatit, K., Kim, K. H., Ha, T. J., Chitikineni, A., Kim, M. Y., Ko, J.-M., Gwag, J.-G., Moon, J.-K., Lee, Y.-H., Park, B.-S., Varshney, R. K., & Lee, S.-H. (2015). Draft genome sequence of adzuki bean, *Vigna angularis*. *Scientific Reports*, *5*(1), Article 8069. https://doi.org/10.1038/srep08069

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, *44*(9), e89. https://doi.org/10.1093/nar/gkw092

König, S., Romoth, L., & Stanke, M. (2018). Comparative genome annotation. *Methods in Molecular Biology*, *1704*, 189–212. https://doi.org/10.1007/978-1-4939-7463-4_6

Kreplak, J., Madoui, M.-A., Cápal, P., Novák, P., Labadie, K., Aubert, G., Bayer, P. E., Gali, K. K., Syme, R. A., Main, D., Klein, A., Bérard, A., Vrbová, I., Fournier, C., d'Agata, L., Belser, C., Berrabah, W., Toegelová, H., Milec, Z., … Burstin, J. (2019). A reference genome for pea provides insight into legume genome evolution. *Nature Genetics*, *51*(9), 1411–1422. https://doi.org/10.1038/s41588-019-0480-1

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Lee, H., Chawla, H. S., Obermeier, C., Dreyer, F., Abbadi, A., & Snowdon, R. (2020). Chromosome-scale assembly of winter oilseed rape *Brassica napus*. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00496

Lonardi, S., Muñoz-Amatriaín, M., Liang, Q., Shu, S., Wanamaker, S. I., Lo, S., Tanskanen, J., Schulman, A. H., Zhu, T., Luo, M.-C., Alhakami, H., Ounit, R., Hasan, A. M., Verdier, J., Roberts, P. A., Santos, J. R. P., Ndeve, A., Doležel, J., Vrána, J., … Close, T. J. (2019). The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *Plant Journal*, *98*(5), 767–782. https://doi.org/10.1111/tpj.14349

Lupia, R., Lidgard, S., & Crane, P. R. (1999). Comparing palynological abundance and diversity: Implications for biotic replacement during the cretaceous angiosperm radiation. *Paleobiology*, *25*(3), 305–340. http://www.jstor.org/stable/2666001

Macneil, L. T., & Walhout, A. J. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, *21*(5), 645–657. https://doi.org/10.1101/gr.097378.109

NCBI Resources Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095

Rana, D., van den Boogaart, T., O'Neill, C. M., Hynes, L., Bent, E., Macpherson, L., Park, J. Y., Lim, Y. P., & Bancroft, I. (2004). Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. *Plant Journal*, *40*(5), 725–733. https://doi.org/10.1111/j.1365-313X.2004.02244.x

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., 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., … Tabata, S. (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Research*, *15*(4), 227–239. https://doi.org/10.1093/dnares/dsn008

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178–183. https://doi.org/10.1038/nature08670

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., Torres-Torres, M., Geffroy, V., Moghaddam, S. M., Gao, D., Abernathy, B., Barry, K., Blair, M., Brick, M. A., Chovatia, M., … Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, *46*(7), 707–713. https://doi.org/10.1038/ng.3008

Schnable, J. C. (2020). Genes and gene models, an important distinction. *New Phytologist*, *228*(1), 50–55. https://doi.org/10.1111/nph.16011

Sharma, V., Schwede, P., & Hiller, M. (2017). CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation. *Bioinformatics*, *33*(24), 3985–3987. https://doi.org/10.1093/bioinformatics/btx527

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., Gentzbittel, L., Childs, K. L., Yandell, M., Gundlach, H., Mayer, K. F., Schwartz, D. C., & Town, C. D. (2014). An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics*, *15*, Article 312. https://doi.org/10.1186/1471-2164-15-312

Tian, X., Pascal, G., Fouchécourt, S., Pontarotti, P., & Monget, P. (2009). Gene birth, death, and divergence: The different scenarios of reproduction-related gene evolution. *Biology of Reproduction*, *80*(4), 616–621. https://doi.org/10.1095/biolreprod.108.073684

Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., Donoghue, M. T. A., Azam, S., Fan, G., Whaley, A. M., Farmer, A. D., Sheridan, J., Iwata, A., Tuteja, R., Penmetsa, R. V., Wu, W., Upadhyaya, H. D., Yang, S.-P., Shah, T., … Jackson, S. A. (2012). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnology*, *30*(1), 83–89. https://doi.org/10.1038/nbt.2022

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., Cannon, S., Baek, J., Rosen, B. D., Tar'an, B., Millan, T., Zhang, X., Ramsay, L. D., Iwata, A., Wang, Y., Nelson, W., Farmer, A. D., Gaur, P. M., Soderlund, C., … Cook, D. R. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology*, *31*(3), 240–246. https://doi.org/10.1038/nbt.2491

Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics*, *2*(4), 216–221. https://doi.org/10.1016/s1672-0229(04)02028-5

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Yang, T., Liu, R., Luo, Y., Hu, S., Wang, D., Wang, C., Pandey, M. K., Ge, S., Xu, Q., Li, N., Li, G., Huang, Y., Saxena, R. K., Ji, Y., Li, M., Yan, X., He, Y., Liu, Y., Wang, X., … Zong, X. (2022). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nature Genetics*, *54*(10), 1553–1563. https://doi.org/10.1038/s41588-022-01172-2

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tay Fernandez, C. G., Bayer, P. E., Petereit, J., Varshney, R., Batley, J., & Edwards, D. (2023). The conservation of gene models can support genome annotation. *The Plant Genome*, *16*, e20377. https://doi.org/10.1002/tpg2.20377