Dear Editor,

Willows (Salix) and poplars (Populus) are known worldwide as woody species with diverse uses [1, 2]. Although these two genera diverged from each other around the early Eocene [3], they share numerous traits, including the same chromosome number of 2n = 38 and the common ‘Salicoid’ genome duplication with a high macrosynteny [4, 5]. However, most willow species flower early in their lives with short, small and sometimes indistinct stems, and thus differ from poplars in their life histories and habits [2]. In addition, multiple inter- and intrachromosomal rearrangements have been detected involving chromosomal regions present in both lineages [6], suggestive of the likely genomic divergence after the common genome duplication.

In order to test this hypothesis, we sequenced the genome of a shrub willow S. suchowensis, which flowers within two years [1, 2]. Genome sequencing was conducted with a combined approach using Roche/454 and Illumina/HiSeq-2000 sequencing technologies (Supplementary information, Table S1A and S1B), and the statistics of the genome assembly were listed in Supplementary Table S1C. The size of the S. suchowensis genome was estimated to be ~425 Mb and ~429 Mb based on 17-mer analysis and flow cytometry, respectively (Supplementary information, Figure S1A and Table S1D), about 60 Mb smaller than that of P. trichocarpa (approximately 485 ± 10 Mb) [4], which is largely consistent with the previous genome measurements of other willow species [7]. The final length of the assembled sequence was amounted to about 71% (303.8/425 Mb) of the estimated genome size. Sequencing depth distribution showed that ~94% of the assemblies had greater than 20-fold coverage, ensuring a high level of single-base accuracy (Supplementary information, Figure S1B). A detailed analysis found that the ~120 Mb of unassembled genomic sequence consisted mainly of repetitive sequences, and the protein-coding regions were assembled to near completeness. The S. suchowensis genome assembly is of comparable quality to other sequenced plant genomes based on next-generation sequencing (Supplementary information, Table S1E). We further evaluated the assembled genome using two independent EST datasets. The first comprises the EST assemblies from various tissues of the sequenced individual, including tender root, young leaves, bark, non-lignified shoot and vegetative buds. The second consists of EST assemblies from flower buds collected from natural willow stands. These two analyses identified 97.6% and 94.2% of the two EST assemblies in our S. suchowensis genome assembly, separately, with > 95% identity and > 50% coverage of query length, confirming the nearly complete coverage of genic regions (Supplementary information, Table S1F).

Core eukaryotic genes identified by CEGMA were further mapped to the predicted gene set, and 97.8% of the NCBI euKaryotic clusters of Orthologous Groups (KOgs) are covered by the predicted willow gene set, confirming the nearly completeness of the genic regions by the present genome assembly (Supplementary information, Table S1G). Assessment of the quality of gene prediction revealed that the overall gene length distributions were similar between S. suchowensis and other plant species with genomes sequenced (Supplementary information, Figure S1C).

In total, we identified 26 599 putative protein-coding genes (Supplementary information, Table S1H and Data S1), 20 261 of which constitute homologous gene pairs with the P. trichocarpa reference gene set [4]. Synteny and collinearity analyses indicate that the chromosomal structures are highly similar between the willow and poplar genomes (Figure 1A and 1B), supporting the concept that these two genera share an additional common (Salicoid) whole genome duplication (WGD) event in their evolutionary history [4]. Estimation with four-fold synonymous third-codon transversion (4DTv) values of the orthologous pairs suggests that the divergence between Salix and Populus took place around 52 million years ago, approximately 6 million years after the additional WGD. The shared WGD was indicated by sharp peaks in 4DTv values of 0.2842 and 0.2052 for S. suchowensis and P. trichocarpa, respectively (Figure
Figure 1 Similarity and divergence between the genomes of *Salix* suchowensis and *Populus* trichocarpa. (A) Gene collinearity between *S. suchowensis* and *P. trichocarpa*. The x-axis corresponds to the *Populus* chromosomes; and the y-axis corresponds to the *Salix* scaffolds. (B) Nucleotide alignments between *S. suchowensis* scaffolds and 19 *P. trichocarpa* chromosomes. The 19 poplar chromosomes were scaled in their physical length and displayed in a numeric order. For each chromosome, the black bar on the top represents the *Populus* chromosome, and the black bar under it represents the *Salix* scaffolds. (C) 4DTv values were separately calculated with paralogs of *S. suchowensis* and *P. trichocarpa*, and with orthologs between *S. suchowensis* and the other four species, including *P. trichocarpa*, *A. thaliana*, *V. vinifera* and *O. sativa*. (D) The mean substitution rates were estimated using single-copy genes for *S. suchowensis*, *P. trichocarpa*, *A. thaliana*, *V. vinifera* and *O. sativa*.

1C). These different 4DTv values might be caused by different evolutionary rates of nucleotide substitution in these two lineages. Using 1 232 single-copy orthologous genes (Supplementary information, Table S1I), we estimated mean substitution rates of $1.09 \times 10^{-9}$ and $0.67 \times 10^{-9}$/site/year for *S. suchowensis* and *P. trichocarpa*, respectively.
respectively. This estimated substitution rate is largely consistent with a previous study of other *Salix* species based on fewer genes [6]. Our comparisons suggest that *S. suchowensis* has a significantly higher substitution rate than *P. trichocarpa* (P < 2.2e-16). However, substitution rates of both species were substantially lower than those of *Arabidopsis thaliana* and *Oryza sativa* (Figure 1D).

The great differences in evolutionary rates between these two species are highly correlated with their flowering habits: the early-flowering species has faster substitution rates than the long-generation one. This finding is consistent with previous expectations [8, 9].

In order to confirm that the smaller number of genes annotated in *S. suchowensis* (26 599) than in *P. trichocarpa* (40 303) was the result of divergent evolution between the two genera after the common genome duplication, we conducted the following analyses. First, the collapsed sequences may lead to the reduced gene number through the fused assembly of homologous genes with multiple copies into a ‘single’ one. Under this scenario, the collapsed sequences will manifest themselves to those regions of the assembled genome which have elevated read coverage. In order to determine how this artifact might affect our gene annotation, we analyzed the read coverage across the genome. We found a small number of regions with elevated read coverage, covering approximately 38.8 Mb, indicating the presence of collapsed sequences in the assembled genome. Of these, 19.8 Mb were located in the repeat regions, whereas the genic region contained only 308.2 Kb of sequences, which were correlated with just 315 annotated genes. Thus, the collapsed assembly of homologous genes contributed little to the reduced number of genes in the assembled willow genome in the present study.

A second possibility is that *Populus* retained more gene copies than *Salix* following genome duplication. To test this, we compared the shrinkage and expansion of gene families within each genome. We extracted the gene family clusters generated by the OrthoMCL pipeline in *S. suchowensis* and *P. trichocarpa*. Of the 14 772 gene families shared by the two genomes, 3 434 families contain more genes in poplar than in willow. In contrast, only 627 families contain more genes in willow than in poplar. The remaining 10 711 gene families contain the same number of genes in both species. We found that 1 034 families were unique in *P. trichocarpa*, while 179 families were specific in *S. suchowensis*. Comparisons of gene expansions between willow and poplar also suggest that fractions of genes associated with WGD or segmental duplications were higher in poplar than in willow, suggesting that poplar retained more *Salicoid* duplicates than willow after their divergence (Supplementary information, Table S1J). Even in the families with more genes in willow, these genes were rarely associated with WGD or segmental duplications, but were mainly derived from expansions through lineage-specific tandem duplications and transposons (Supplementary information, Table S1J). These comparisons suggest that willow might have lost more genes than poplar after the common genome duplication.

Both gene variation and genome duplication have contributed greatly to our current species and life diversity [10]. The genomic divergences between willow and poplar illustrated here highlight their divergent evolution since the common genome duplication, which laid the genetic bases for development of their divergence in life history, habit and other traits [2]. In addition, willows are promising sources for bioenergy crops due to their high biomass yields [1]. The availability of this willow genome will help improve productivity and quality of such woody crops for biofuel and other uses.

**Acknowledgments**

This work was supported by the Key Forestry Public Welfare Project (201304102), the National Basic Research Program of China (2012CB114505), the National Natural Science Foundation of China (31125008) and the Program for Innovative Research Team of the Educational Department of China and in Universities of Jiangsu Province. It was also funded by the Doctorate Fellowship Foundation and PAPD (Priority Academic Program Development) program at Nanjing Forestry University and the BioEnergy Science Center, which is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the Department of Energy.

Xiaogang Dai1,*, Quanjun Hu2,*, Qingle Cai3,*, Kai Feng1,*, Ning Ye1,*, Gerald A Tuskan4,*, Richard Milne5, Yingnan Chen1, Zhbing Wan1, Zefu Wang2, Wenchun Luo2, Kun Wang2, Dongshi Wan2, Mingxiu Wang1, Jun Wang3, Jianquan Liu2, Tongming Yin1

1*The Southern Modern Forestry Collaborative Innovation Center, Nanjing Forestry University, Nanjing 210037, China; 2State Key Laboratory of Grassland Agro-Ecosystem, College of Life Science, Lanzhou University, Lanzhou 730000, China; 3BGI-Shenzhen, Shenzhen 518083, China; 4Bio-Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; 5Institute of Molecular Plant Science, The University of Edinburgh, Daniel Rutherford Building, King’s Buildings, Mayfield Road, Edinburgh EH9, 3JH, UK

*These six authors contributed equally to this work.

Correspondence: Tongming Yin1, Jianquan Liu2, Jun Wang3

*E-mail: tmyin@njfu.edu.cn

*E-mail: liujq@lzu.edu.cn

*E-mail: wangji@genomics.org.cn
References

1. Djomo SN, Kasmioui OE, Ceulemans R. Global Change Biol 2011; 3:181-197.
2. Eckenwalder JE. Systematics and evolution of *Populus*. In: Stettler RF, Brandshaw HD Jr, Heilman PE, Hinckley TM, eds. Biology of *Populus* and its implications for management and conservation. NRC Research Press: Ottawa 1996:7-32.
3. Manchester SR, Judd WS, Handley B. Int J Plant Sci 2006; 167:897-908.
4. Tuskan GA, DiFazio S, Jansson S, *et al.* Science 2006; 313:1596-1604.
5. Hanley S, Mallott M, Karp A. Tree Genet Genomes 2006; 3:35-48.
6. Berlin S, Lagercrantz U, von Arnold S, *et al.* BMC Genomics 2010; 11:129.
7. Thibault J. Can J Bot 1998; 76:157-165.
8. Gaut B, Yang L, Takuno S, *et al.* Annu Rev Ecol Evol Syst 2011; 42:245-266.
9. Smith SA, Donoghue MJ. Science 2008; 322:86-89.
10. DeBodt S, Maere S, Van de Peer Y. Trends Ecol Evol 2005; 20:591-597.

(Supplementary information is linked to the online version of the paper on the Cell Research website.)

This work is licensed under the Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0