Opportunities and challenges in forest tree genomics are seemingly as diverse and as large as the trees themselves; however, here, we have chosen to focus on the potential significant impact on all of tree biology research if only an open-access culture and comparative-genomics infrastructure were developed. In earlier articles [1,2], we argued that the great diversity of forest trees found in both the undomesticated and domesticated state provides an excellent opportunity to understand the molecular basis of adaptation in plants and furthermore that comparative-genomic approaches will greatly facilitate discovery and understanding. We identified several priority research areas towards realizing these goals (Box 1), such as establishing reference genome sequences for important tree species, determining how to apply sequencing technologies to understand adaptation, and developing resources for storing and accessing forestry data. Significant progress has been made in many of these priorities, with the exception of investments in database resources and understanding ecological functions. Here, we briefly summarize the rapid progress in developing genomic resources in a small number of species and then offer our view on what we believe it will take to realize the final two priorities.

**The great diversity found in forest trees**

There are an estimated 60,000 tree species on earth, and approximately 30 of the 49 plant orders contain tree species. Clearly, the tree phenotype has evolved many times in plants. The diversity of plant structures, development, life history, environments occupied and so on in trees is nearly as broad as higher plants in general, but trees share the common characteristic that all are perennial and many are very long lived. Because of the sessile nature of plants, each tree must survive and reproduce in a specific environment over the seasonal cycles of its lifetime. This tight association between individual genotypes and their environment provides a powerful research setting, just as it has driven the evolution of a plethora of uniquely arboreal adaptations. Understanding these evolutionary strategies is a long-standing area of study of tree biologists, with many broader biological implications.

Completed and current genome-sequencing projects in forest trees are limited to about 25 species from just 4 of more than 100 families: Pinaceae (pines, spruces and firs), Salicaceae (poplars and willows), Myrtaceae (eucalyptus) and Fagaceae (oaks, chestnuts and beeches). Large-scale sequencing projects such as the 1000 Human Genomes [3], 1000 Plant Genomes (1KP) [4] or the 5000 Insect Genome (i5k) [5] projects have not yet been proposed for forest trees.

**Rapidly developing genomic resources in forest trees**

Genome resources are developing rapidly in forest trees in spite of the challenges associated with working with large, long-lived organisms and sometimes very large genomes [2]. Complete genome sequencing, however, has been slow to advance in forest trees owing to funding limitations and the large size of conifer genomes. Black cottonwood (Populus trichocarpa Torr. & Gray) was the first forest tree genome to be sequenced by the US Department of Energy Joint Genome Institute (DOE/JGI) [6] (Table 1). Black cottonwood has a relatively small genome (450 Mb) and is a target feedstock species for cellulosic ethanol production, and thus fits into the DOE/JGI priority of sequencing bioenergy feedstock species.
Box 1. Research priorities in forest tree genomics identified in earlier Opinion papers

From Neale and Ingvarsson [1]:
- Deep expressed-sequence tag (EST) sequencing in many species
- Comparative resequencing in many species
- Reference genome sequence for pine

From Neale and Kremer [2]:
- Reference genome sequences for several important species
- Greater investment in diverse species towards understanding ecological function
- Application of next-generation sequencing technologies to understand adaptation using landscape genomic approaches
- Greater investment in database resources and cyber-infrastructure development
- Development of new and high-throughput phenotyping technologies

The genus *Populus* has 30+ species (aspens and cottonwoods) with genome sizes of approximately 500 Mb. Several species are being sequenced by DOE/JGI, and other groups around the world, and it seems likely that all members of the genus will soon have a genome sequence (Table 1). The next forest tree to be sequenced was the flooded gum (*Eucalyptus grandis* BRASUZ1, which is a member of the Myrtaceae family), again by DOE/JGI. Eucalyptus species and their hybrids are important commercial species grown in their native Australia and many regions throughout the southern hemisphere. Several more eucalyptus species are being sequenced (Table 1), each with relatively small genomes (500 Mb), but it will probably take many years before all 700+ members of this genus are completed. Several members of the Fagaceae family are now being sequenced (Table 1). Members of this group include the oaks, beeches and chestnuts, with genome sizes less than 1 Gb.

The gymnosperm forest trees (such as the conifers) were the last to enter the world of genome sequencing. This was entirely due to their very large genomes (10 Gb and greater) as they are extremely important economically and ecologically, and phylogenetically they represent the ancient sister lineage to that of angiosperm species. Genome resources needed to support a sequencing project were reasonably well developed, but it was not until the introduction of next-generation sequencing (NGS) technologies that sequencing conifer genomes became tractable. Currently, there are at least ten conifer (Pinaceae) genome-sequencing projects under way (Table 1).

Aside from reference genome sequencing in forest trees, there is significant activity in transcriptome sequencing and resequencing for polymorphism discovery (Tables 2 and 3). We have only listed the transcriptome and resequencing projects in Table 1 that are associated with a species that has an active genome-sequencing project.

**The opportunity for comparative-genomic approaches in forest trees**

The power of comparative-genomic approaches for understanding function in an evolutionary framework is well established [7-13]. Comparative genomics can be applied to sequence data (nucleotide and protein) at the level of individual genes or genome-wide. Genome-wide approaches provide insight into both chromosome evolution and the diversification of biological functions and interactions.

Understanding of gene function in forest tree species is challenged by the lack of standard reverse-genetic tools routinely used in other systems - for example, standard marker stocks, facile transformation and regeneration - and by the long generation times. Thus, comparative genomics becomes the more powerful approach to understanding gene function in trees.

Comparative genomics requires not only data availability but also cyber-infrastructure to support exchange and analysis. The TreeGenes database is the most comprehensive resource for comparative-genomic analyses in forest trees [14]. Several smaller databases have been created to facilitate collaborations, including: Fagaceae genomics web, hardwoodgenomics.org, Quercus portal, PineDB, ConiferGDB, EuroPineDB, PopulusDB, PoplarDB, EucalyptusDB and Eucanext (Tables 1, 2, and 3). These resources vary greatly in their scope, relevance and integration. Some are static and archival, whereas others focus on current sequence content for a specific species or a small number of related species. This results in overlapping and conflicting data among repositories. In addition, each database uses its own custom interfaces and back-end database technology to serve sequence to the user. The US National Science Foundation funding for large-scale infrastructure projects, such as iPlant, is leading efforts aimed towards centralizing resources for research communities [15]. Without centralized resources, researchers are forced to employ inefficient data-mining methods through queries of independently maintained databases or inconsistently formatted supplemental files on journal websites. Specific areas of interest for the forest tree genomic community include the ability to connect sequence, genotype and phenotype to individual, geo-referenced trees. This type of integration can only be achieved through web services that allow disparate resources to communicate in ways that are transparent to the user [16]. With the recent increase of genome sequences available for many of these species, there is a
| Family      | Genus     | Species                  | Sequence access | Genome                                | Related publications |
|------------|-----------|--------------------------|-----------------|---------------------------------------|----------------------|
| Pinaceae   | Pinus     | *taeda* (loblolly pine)  | [14,19]         | Resequenced amplicons                 | [21,22]              |
|            |           |                          | [20]            | BACs                                  |                      |
|            |           |                          | [14,19]         | Fosmids                               |                      |
|            |           |                          | [23]            | Draft genome complete                 |                      |
|            | Pinus     | *lambertiana* (sugar pine) | [14,19]       | Resequenced amplicons                 | [24]                 |
|            |           |                          | [23]            | Draft genome (in progress)            |                      |
|            | Pseudotsuga | *menziesii* (Douglas-fir) | [14,19]       | Resequenced amplicons                 | [25]                 |
|            | Pinus     | *sylvestris* (Scots pine) | [26]           | Draft genome (in progress)            |                      |
|            | Pinus     | *pinaster* (maritime pine) | [26]           | Draft genome (in progress)            |                      |
|            | Pinus     | *sibirica* (Siberian pine) | [14,19]       | Resequenced amplicons                 | [27]                 |
|            | Pinus     | *radiata* (Monterey pine) | [14,19]       | Draft genome (in progress)            |                      |
|            | Picea     | *abies* (Norway spruce)   | [14,19]        | Resequenced amplicons                 | [29]                 |
|            | Picea     | *glauca* (white spruce)   | [32]           | Draft genome complete                 |                      |
|            | Larix     | *sibirica* (Siberian larch) | [28]         | Draft genome (in progress)            |                      |
| Salicaceae | Populus   | *trichocarpa* (black cottonwood) | [34]     | Genome complete                      | [6]                  |
|            |           |                          | [19]           | BACs                                  | [35]                 |
|            |           |                          | [19]           | Genome resequencing (restricted)      | [36]                 |
|            | Populus   | *tremula* (European aspen) | [37]           | Draft genome complete                 |                      |
|            | Populus   | *tremula x tremuloides*   | [37]           | Draft genome complete                 |                      |
|            | Populus   | *tremuloides* (quaking aspen) | [37]       | Draft genome complete                 |                      |
|            |           |                          | [19]           | BACs                                  | [38]                 |
|            | Populus   | *grandidentata* (bigtooth aspen) | [37]   | Draft genome complete                 |                      |
|            | Populus   | *nigra* (black poplar)    | [39]           | Draft genome complete (restricted)    |                      |
|            | Salix     | *purpurea* (purpleosier willow) | [40]       | Draft genome complete (restricted)    |                      |
| Myrtaceae  | Eucalyptus | *grandis* (rose gum)    | [19]           | BACs                                  | [41]                 |
|            |           |                          | [34]           | Draft genome complete                 |                      |
|            | Eucalyptus | *globulus* (blue gum)    | [42]           | Draft genome (in progress)            |                      |
|            | Eucalyptus | *carnaldulensis* (river red gum) | [43]     | Draft genome complete                 | [44]                 |
|            | Corymbia  | *citriodora* (lemon-scented gum) | [45]       | Draft genome complete (restricted)    |                      |
| Fagaceae   | Quercus   | *robur* (English oak)    | [19]           | BACs                                  | [46,47]              |
|            |           |                          | [48]           | Draft genome (in progress)            |                      |
|            | Castanea  | *mollissima* (Chinese chestnut) | [49]       | Draft genome (in progress)            |                      |
|            |           |                          | [50]           | BACs                                  | [51]                 |
| Betulaceae | Betula    | *nana* (dwarf birch)     | [52]           | Draft genome complete                 | [53]                 |
| Oleaceae   | Fraxinus  | *excelsior* (European ash) | [54]         | Draft genome complete                 |                      |

Details current genome sequencing projects in forest trees with sequence access information and relevant publications.
Table 2. Transcriptome resources in forest trees

| Family   | Genus   | Species                             | Sequence access | Transcriptome                      | Related publications |
|----------|---------|-------------------------------------|-----------------|-----------------------------------|----------------------|
| Pinaceae | Pinus   | *taeda* (loblolly pine)             | [14,19,55]      | EST sequencing (Sanger)           | [56-59]              |
|          |         |                                     |                 | EST sequencing (454)              | [60]                 |
|          |         |                                     |                 | Exome resequencing                | [61]                 |
|          | Pinus   | *ponderosa* (sugar pine)            | [14,19,60]      | EST sequencing (454)              | [60]                 |
|          | Pseudotsuga | *menziesii* (Douglas-fir)       | [14,19]         | EST sequencing (Sanger)           |                      |
|          |         |                                     |                 | EST sequencing (454)              | [60,63,64]           |
|          | Pinus   | *sylvestris* (Scots pine)           | [14,19,55]      | EST sequencing (Sanger)           |                      |
|          | Pinus   | *pinaster* (maritime pine)          | [14,19,65]      | EST sequencing (Sanger/454)       | [65,66]              |
|          | Pinus   | *radiata* (Monterey pine)           | [14,19,55]      | EST sequencing (Sanger)           | [67-72]              |
| Picea    | *abies* (Norway spruce)             | [14,19,60]      | EST sequencing (454)              | [60]                 |
|          |         |                                     |                 | EST sequencing (Next-Gen)         | [73]                 |
| Picea    | *glauca* (white spruce)             | [19]            | UniGenes (Sanger/454)             | [74]                 |
| Salicaceae | *Populus* | *trichocarpa* (black cottonwood)   | [14,19,75]      | EST sequencing (Sanger)           | [75-78]              |
|          |         |                                     |                 | Exon capture                      | [79]                 |
|          |         |                                     |                 | UniGenes (Sanger)                 | [80]                 |
|          | *Populus* | *tremula* (European aspen)         | [14,19,75]      | EST sequencing (Sanger)           | [75,81]              |
|          | *Populus* | *tremula x tremuloides*           | [14,19,75]      | EST sequencing (Sanger)           | [75,76]              |
|          | *Populus* | *tremuloides* (quaking aspen)       | [14,19]         | EST sequencing (Next-Gen)         | [82]                 |
|          | *Populus* | *nigra* (black poplar)              | [14,19]         | UniGenes (Sanger)                 | [83]                 |
| Myrtaceae | *Eucalyptus* | *grandis* (rose gum)         | [14,19]         | EST sequencing (Sanger)           | [84-86]              |
|          |         |                                     |                 | EST sequencing (NextGen)          | [87]                 |
|          | *Eucalyptus* | *globulus* (blue gum)        | [14,19]         | EST sequencing (NextGen)          | [41,88]              |
|          | *Eucalyptus* | *carnulagensis* (river red gum)  | [14,19]         | EST sequencing (RNA-Seq)          | [89]                 |
| Fagaceae | *Quercus* | *robur* (English oak)             | [19,90]         | EST sequencing (454)              | [91]                 |
|          | *Castanea* | *mollissima* (Chinese chestnut)    | [50]            | EST sequencing (454)              | [92,93]              |
| Oleaceae | *Fraxinus* | *excelsior* (European ash)        | [19]            | EST sequencing (454)              | [94]                 |

Details current transcriptome sequencing projects in forest trees with sequence access information and relevant publications.

need to facilitate community-level annotation and research support.

**The need for a better-developed open-access culture in forest tree genomics research**

The Human Genome Project established a culture of open access and data sharing in genomics research for both humans and animal models that has been extended to many other species, including *Arabidopsis*, rat, cow, dog, rice, maize and more than 500 other eukaryotes. Beginning in the late 1990s, these large-scale projects released data very rapidly to the scientific community, often years before publication. This rapid release of data with few restrictions has allowed thousands of scientists to begin work on specific genes and gene families, and on functional studies, long before the genome papers have appeared. One of the driving motivations for this culture, and the reason that many scientists support it, is that large-scale sequencing can be done most efficiently when centers that have expertise in sequencing technology take the lead. With all the sequencing concentrated, the body of data needs to be shared freely in order to get it in the
hands of the widely distributed experts. This open-access culture has dramatically accelerated scientific progress in biological research.

The path to success avoids delays
Careful inspection of Table 1 reveals that forest tree genome projects are very slow to release sequence data into the public domain. Once a project is finished and submitted for publication, a draft genome becomes available - for example, the poplar genome was released and published in 2006. However, pre-publication releases are infrequent, exceptions being the PineRefSeq project that has made three releases and the SMarTForest project that has made one (Table 1). This is unfortunate because good-quality sequence contigs and scaffolds could be made available years before publication, delivering an extremely important resource to the community. This delay can be understood from privately financed projects seeking commercial advantages, but nearly all the projects listed in Table 1 are financed by public funds whose stated mission is advancing science and development of community resources. Publication rights are easily protected by data-use policy statements such as the Ft Lauderdale [17] and Toronto agreements [18], but unfortunately these conventions are not often used and data access is restricted by password-protected websites (Tables 1, 2, and 3). We hope the opinion offered here will lead to a discussion in the forest tree community, to a more open-access culture and thus to a more vibrant and rapidly advancing research area.

Competing interests
The authors declare they have no competing interests.

Acknowledgements
Writing of this paper was funded by US Department of Agriculture, National Institute of Food and Agriculture grant #2011-67009-30030.

Abbreviations
EST, expressed-sequence tag; Mb, mega-base; NGS, next-generation sequencing.

Author details
1Department of Plant Sciences, University of California, Davis, CA 95616, USA.
2Department of Evolution and Ecology, University of California, Davis, CA 95616, USA.
3Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD 21205, USA.

Published: 24 June 2013

References
1. Neale DB, Ingvarsson PK. Population, quantitative and comparative genomics of adaptation in forest trees. Curr Open Plant Biol 2008, 11:149-155.
neale et al. Genome Biology 2013, 14:120
http://genomicbiology.com/2013/14/6/120
Page 6 of 8

2. Neale DB, Kremer A. Forest tree genomics: growing tools and applications. Nat Rev Genet 2011, 12:111-122.
3. Genomes Project C, Abecasis GR, Auer A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marsh GT, McEvan GA. An integrated map of genetic variation from 1,092 human genomes. Nature 2012, 491:56-65.
4. The 1KP Project [http://onekpop.com/]
5. iSK Insect and other Arthropod Genome Sequencing Initiative [http://arthropodgenomes.org/wiki/ISK]
6. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev IV, Hellsten U, Putnam N, Ralph S, Romberg S, Salamov A, Schmutz J, Salse J, Stenzel U, vonholdt BM, Weeden NF, Wang X, Wetherall J, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 2006, 313:1566-1604.
7. Arabidopsis Genome Initiative Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 2000, 408:796-815.
8. Batzoglou S, Flicek P, Mesirov JP, Berger B, Lande ES. Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res 2000, 10:950-958.
9. Zdobnov EM, von Mering C, Letunic I, Torrens D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kozolko E, Kraft CJ, Lai Z, Lewis S,ouis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, tops T, Winters K, Przeworski M, et al. Comparative genomics of Anopheles gambiae and Drosophila melanogaster. Science 2002, 298:149-159.
10. El-Sayed NM, Myles PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science 2005, 309:404-409.
11. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, yIv VN, Pollard DA, Saccone TT, Larramendi AM, Singh ND, Abad JP, Axt DN, Adrian B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, et al. Comparative genomics of the trypanosomatid parasitic protozoa. Science 2005, 309:404-409.
12. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler DD, Matasci N, Vasquez-Gross HA, Yu JJ, Figueroa B, Gessler DD, Neale DB, Wegrzyn JL: The 1KP Project [http://onekpop.com/]
13. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW, GenBank. Nucleic Acids Res 2013, 41:D36-D42.
14.Accelerating Pine Genomics [http://www.pine.msstate.edu/]
15. Kelleher CT, Chiu R, Shin H, Bosdet IA, Krzywinski MI, Fjell2 CD, Wilkin J, Yin T, Difazio SP, Ali LS, Jakubowicz K, Chao Y, Pongpipat N, Weng Y, Zhen Y, Kato T, Umehara M, et al. Comparative genomics of the Populus genome: functional annotations, comparative genomics and integration with the plant genome database. Nature 2009, 458:407-413.
98. Pavy N, Pelgas B, Beauseigle S, Bias S, Gagnon F, Gosselin I, Lamothe M, Isabel N, Bossuqet J: Enhancing genetic mapping of complex genomes through the design of highly multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics 2008, 9:21.

99. Dryad [http://datadryad.org/]

100. Geraldes A, Pang J, Thissien N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann M, Birol I, Jones SJ, Cronk QC, Douglas CJ: SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing. Mol Ecol Resour 2011, 11 Suppl 1:81-92.

101. Wegrzyn J, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai CJ, Lamothe M, Neale DB: Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem. New Phytol 2010, 188:515-532.

102. Isabel N, Lamothe M, Thompson SL: A second-generation diagnostic single nucleotide polymorphism (SNP)-based assay, optimized to distinguish among eight poplar (Populus L.) species and their early hybrids. Tree Genet Genomes 2013, 9:621-626.

103. Guerra FP, Wegrzyn JL, Sykes R, Davis MF, Stanton BJ, Neale DB: Association genetics of chemical wood properties in black poplar (Populus nigra). New Phytol 2013, 197:162-176.

104. Diversity Arrays Technology Pty Ltd (DArT P/L) [http://www.diversityarrays.com/]

105. Sansaloni CP, Petrolt CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A: A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. Plant Methods 2010, 6:11.

106. Grattapaglia D, Silva-Junior OB, Kirst M, de Lima BM, Faria DA, Pappas GJ Jr: High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC Plant Biol 2011, 11:65.

107. Hendre PS, Kamalakannan R, Varghese M: High-throughput and parallel SNP discovery in selected candidate genes in Eucalyptus camaldulensis using Illumina NGS platform. Plant Biotechnol J 2012, 10:646-656.

doi:10.1186/gb-2013-14-6-120

Cite this article as: Neale DB, et al. Open access to tree genomes: the path to a better forest. Genome Biology 2013, 14:120.