Comparative sequence analysis between orthologous regions of the *Arabidopsis* and *Populus* genomes reveals substantial synteny and microcollinearity

Brigid Stirling, Zamin Koo Yang, Lee E. Gunter, Gerald A. Tuskan, and H.D. Bradshaw, Jr.

Abstract: More than 300 kb of DNA sequence from five *Populus* bacterial artificial chromosome (BAC) clones was compared with the complete sequence of the *Arabidopsis* genome to search for collinearity between the genomes of these two plant genera. Approximately 27% of the DNA sequences from the *Populus* genome were homologous to protein-coding regions in the *Arabidopsis* genome. BLAST scores and synteny were used to infer orthologous relationships between the *Populus* and *Arabidopsis* homologs. The probability that any pair of genes on a single *Populus* BAC will have orthologs on the same *Arabidopsis* chromosome is 46%–58%, substantially greater than the 20% expectation if there is no conservation of synteny between the *Populus* and *Arabidopsis* genomes. Likewise, the probability that any pair of genes on a single *Populus* BAC will have orthologs on a single *Arabidopsis* BAC is 19%–25%, much higher than the 0.1% expected if the orthologs are randomly distributed. These results provide evidence for substantial "pockets" of conserved microcollinearity between regions of the *Populus* and *Arabidopsis* genomes as well as for conservation of synteny even when local gene collinearity is not preserved during genome evolution.

Résumé : Les auteurs ont comparé plus de 300 kb de séquence d’ADN de *Populus* à la séquence complète du génome d’*Arabidopsis* afin de déterminer la colinéarité entre les génomes de ces deux genres chez les plantes. Les séquences de *Populus* ont été déterminées à partir de cinq portions du génome clonées à l’aide de chromosomes bactériens artificiels (CBA). Approximativement 27 % des séquences d’ADN de *Populus* étaient homologues aux régions codant pour des protéines du génome d’*Arabidopsis*. Les résultats de comparaison des séquences à l’aide du logiciel BLAST ainsi que la synténie ont été utilisés afin de vérifier l’orthologie entre les homologues de *Populus* et d’*Arabidopsis*. La probabilité que n’importe quelle paire de gènes d’un clone unique de CBA de *Populus* possède des orthologues sur le même chromosome d’*Arabidopsis* est de 46 à 58 %. Cette probabilité est beaucoup plus élevée que l’espérance de 20 % si la synténie n’était pas conservée entre les génomes de *Populus* et d’*Arabidopsis*. De la même façon, la probabilité que n’importe quelle paire de gènes d’un clone unique de CBA de *Populus* ait des orthologues sur un clone unique de CBA d’*Arabidopsis* est de 19 à 25 %, ce qui est plus élevé que le seuil de 0,1 % attendu si les orthologues étaient distribués aléatoirement. Ces résultats constituent des preuves que d’importantes zones conservées de microcollinéarité existent entre les régions du génome de *Populus* et celui d’*Arabidopsis*, et que la synténie demeure conservée même lorsque la colinéarité locale des gènes n’a pas été préservée durant l’évolution des génomes.

[Traduit par la Rédaction]

Introduction

Interspecific *Populus* hybrids are the fastest-growing trees in the temperate zone and have been recognized as an important source of pulp, lumber, and biofuel (Zsuffa et al. 1996; Tuskan 1998). The positional cloning of genes controlling important traits in *Populus*, and other forest trees, has been difficult (Stirling et al. 2001; Zhang et al. 2001), in large part because of their long generation time (4–10 years) and poorly known genomes. If there is substantial synteny and collinearity between the genomes of *Populus* and *Arabidopsis*, comparative genomics could provide a powerful alternative approach to gene discovery and isolation in *Populus*, given that the complete DNA sequence of the *Arabidopsis* genome is known (The Arabidopsis Genome Initiative 2000) and rapid progress is being made toward a functional understanding of all *Arabidopsis* genes (Somerville and Dangl 2000).

Our goal was to investigate the extent of synteny and microcollinearity between orthologous regions of the *Populus* and *Arabidopsis* genomes. DNA sequences from each of five *Populus balsamifera* L., ssp. trichocarpa (Torr. & A. Gray) Brayshaw bacterial artificial chromosome (BAC)
clones were compared with the *Arabidopsis thaliana* complete genome sequence, orthology relationships among the genes were inferred, and the relative positions of orthologs in each genome were determined.

**Materials and methods**

**Isolation and sequencing of P. balsamifera BAC DNA**

A bacterial artificial chromosome library of 50,000 *P. balsamifera* clones with an average insert size of 120 kb (10× genome coverage; Stirling et al. 2001) was screened by polymerase chain reaction for three genes: *PHYTOCHROME B* (Howe et al. 1998) (GenBank accessions AAB81955 and AAB81954), a *Populus* homolog of maize *teosinte branched1* (Doebley et al. 1997) (GenBank accession T04347), and a *Populus* homolog of *Arabidopsis* ABCISIC ACID INSENSITIVE1 (Frewen et al. 2000). These three genes were found on poplar BACs 2c5, 16j18, and 6k8, respectively. Two additional BACs known to be linked to the poplar leaf rust resistance gene *MxC3* were isolated by genetic and physical mapping (BACs 41g18 and 47m20; Stirling et al. 2001).

DNA from *Populus* BACs 2c5, 16j18, 6k8, 41g18, and 47m20 was purified for shotgun sequencing. A single colony was picked from a freshly streaked plate and then purified with QIAquick spin columns (QIAGEN). The purified DNA was sheared to an average size of 200 µL.

The purified BAC DNA was sheared to an average size of 1 kb with a Hydro Shear (Gene Machines, San Carlos, Calif.). The sheared DNA was treated with T4 DNA polymerase (New England Biolabs, Beverly, Mass.) to make blunt ends and then purified with QIAquick spin columns (QIAGEN). The purified DNA was subcloned into the pBluescript II KS+ (Stratagene, La Jolla, Calif.), trans -formed into *Escherichia coli* DH5α, and selected on LB plates containing 100 µg ampicillin/mL, 50 µg X-Gal/mL, and 1 mmol/L isopropyl β-D-thiogalactoside. Subclones were picked and inoculated into 96-well blocks having each hole filled with 1 mL of Terrific Broth containing 100 µg ampicillin/mL and grown overnight at 37 °C. Plasmids were extracted and purified using the QIAprep 96 Turbo Miniprep Kit (QIAGEN, Valencia, Calif.) followed by two CsCl – ethidium bromide gradient centrifugation steps. After ethanol precipitation, the DNA was resuspended in a final volume of 200 µL.

The purified BAC DNA was resuspended in 10 µL of Hi-Di Formamide (Applied Biosystems) and then run on an ABI PRISM 3700 DNA analyzer (Applied Biosystems). Between 384 and 480 templates were sequenced for each BAC.

Sequence reads from each of the five *Populus* BAC clones were edited and assembled into contigs using the PHRED/PHRAP software package (Ewing et al. 1998). Contigs were deposited in GenBank (accessions 3657725–3658021).

**Determination of orthology between Populus and Arabidopsis homologs**

The nonredundant (NR) protein database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) and the *Arabidopsis* protein database (http://www.ncbi.nlm.nih.gov/BLAST/Genome/ara.html) were searched with contigs assembled from the five *Populus* BACs to identify homologs in the *Arabidopsis* genome. In the first analysis, 297 contigs assembled from the five poplar BAC DNA sequences were used as queries in BLASTX (version 2.0) (Altschul et al. 1997) searches against the NR protein database. The best alignment to an *Arabidopsis* sequence, based on a threshold of $E < 10^{-2}$, was selected and the translated *Populus* amino acid sequence used as a query in a BLASTP search against the *Arabidopsis* protein database. If multiple contigs from a *Populus* BAC matched different regions of the same protein in the initial BLASTX search, a longer *Populus* protein sequence was assembled based on these multiple alignments for subsequent analysis in the *Arabidopsis* protein database. The top four BLASTP matches (threshold $E < 10^{-19}$ and similarity score $s > 90$) in the *Arabidopsis* protein database were selected for further analysis to infer orthology. This more stringent statistical threshold and a cutoff similarity score were used in the BLASTP search of the *Arabidopsis* protein database to minimize spurious hits (Mushegan et al. 1998). For each of the top four matches, the protein-coding sequences and locations in the *Arabidopsis* genome were identified from the National Center for Biotechnology Information reference sequence (RefSeq) accession number (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html). In addition, the full protein-coding sequences for the top four alignments were compared against each other in the *Arabidopsis* protein database (via BLASTP) to identify potential gene families.

To identify probable orthologs among the *Arabidopsis* homologs, two criteria were applied sequentially. First, if the *E* value for the best BLASTP match was at least 1020 fold lower than the next highest scoring alignment, or all four top matches were on the same *Arabidopsis* BAC, then the best match was selected as a “single-hit” ortholog representing a one-to-one relationship between the genomes of *Populus* and *Arabidopsis*. Second, if the BLASTP *E* values were within a factor of 1020 among members of an *Arabidopsis* protein family, but one of the family members showed conserved synteny or collinearity with *Arabidopsis* and *Populus* genes whose orthology relationship was based on a single hit, then this family member was inferred to be the ortholog.

Although analysis of synteny and collinearity is often done by simple graphical alignment of orthologous regions of the genomes (e.g., Cavell et al. 1998; Acarkan et al. 2000; Ku et al. 2000; Rossberg et al. 2001), we wished to have a quantitative measure. Accordingly, the probability that any
pair of genes from a single Populus BAC was syntenic or collinear with its orthologs in Arabidopsis was calculated for each of the five Populus BACs. Considering all pairwise combinations of genes from each Populus BAC, an estimate of the probability of synteny with the Arabidopsis genome can be calculated as 2S(n(n – 1)) where S is the number of syntenic gene pairs, i.e., where a pair of genes from a Populus BAC has orthologs on the same Arabidopsis chromosome, and n is the total number of genes identified on the Populus BAC. Analogously, the probability of microcollinearity is calculated as 2C(n(n – 1)) where C is the number of pairs of orthologous genes found on both a single Populus BAC and a single Arabidopsis BAC and n is the total number of genes identified on the Populus BAC.

Results

Identification of orthologs between Populus and Arabidopsis

DNA sequence (approximately 2x depth) from BACs 2c5, 16j18, 6k8, 41g18, and 47m20 was assembled into 297 contigs with an average size of 1007 bp. The length of the contigs is 301 541 bp, with an average contig coverage of 60 328 bp/BAC. Based on the BAC library’s average insert size of 120 kb (Stirling et al. 2001), this low-pass sequencing is expected to find approximately half of the Populus sequence in the five BAC clones.

Of the 297 Populus contigs assembled from the five BACs, 80 (27%) had significant BLASTX matches (E < 10^-5) to individual proteins or protein families in the NR database. Of the remaining 217 contigs, 200 (67.3%) had no significant similarities at the established threshold, and 17 (5.7%) were homologous to transposon-like elements. On BAC 47m20, there were no genes found other than putative transposons.

A total of 46 different proteins or protein families were homologous to the 80 Populus contigs that had significant hits to coding regions in the initial BLASTX search. When this set of 46 proteins/protein families was reexamined in the Arabidopsis protein database, 33 had significant hits in the Arabidopsis protein database at the stricter statistical threshold (E < 10^-15) and similarity score cutoff (s > 90) and so were analyzed further. Of these 33 Populus proteins, 19 produced single hits in the Arabidopsis database and thus represent one-to-one relationships presumed to be due to orthology. The remaining 14 Populus proteins had two or more hits in the Arabidopsis database and thus represent protein families with more than one possible ortholog. On the basis of conserved sequence and synteny or microcollinearity (Doyle and Gaut 2000), orthologs were inferred for these 14 gene family members. The proteins encoding these 33 orthologs and their locations in the Arabidopsis genome are listed in Table 1.

Evidence for conserved synteny and microcollinearity between the genomes of Populus and Arabidopsis

Under the null hypothesis that there is no conservation of synteny between the genomes of Populus and Arabidopsis, the expectation is that the probability of any pair of Populus genes from the same BAC having orthologs on a single chromosome in Arabidopsis (where the haploid number of chromosomes is 5) is one-fifth or 20%. Observed synteny values higher than 20% for all possible pairs of genes indicate that there is conservation of synteny between the genomes of Populus and Arabidopsis. The most conservative estimate of synteny between the Populus and Arabidopsis genomes (i.e., using only the 19 orthologous relationships inferred by single hits in the BLASTP search and taking the mean across all four gene-containing BACs) is 46% (range 0%–100%) (Table 2). If the 14 gene family members whose orthology relationships were inferred by synteny and collinearity are added to the single-hit orthologs, the estimated probability of synteny between any pair of Populus and Arabidopsis genes rises to 58% (range 33%–83%) (Table 2). In either case, the average observed degree of synteny between Populus and Arabidopsis is much higher than the 20% expected if orthologs are distributed randomly between the two genomes.

At a finer genomic scale, the null expectation is that the probability that any pair of Populus genes drawn from a single BAC has orthologs on a single Arabidopsis BAC is just 0.1% (approximately 120 kb/BAC in an approximately 120-Mb Arabidopsis genome), if the genomes of Populus and Arabidopsis have diverged so much that their orthologs are randomly distributed. The average observed degree of microcollinearity is 19% (range 0%–33%) for the 19 single-hit orthologs and 25% (range 16%–33%) if all 33 inferred orthologous relationships are included in the calculation (Table 2). Both of these estimates far exceed the expectation under the null hypothesis of no collinearity.

Discussion

General structure of the Populus genome

The number of putative protein-coding sequences identified for each of the five BAC clones varied substantially, showing that gene density is not uniform within the Populus genome. For example, 19 putative protein-coding regions were identified for BAC 2c5 in the initial BLASTX search of the NR protein database. In contrast, the only coding sequences identified for BAC 47m20 were homologous to transposon-like elements. Approximately 9% of the Populus sequences had homology to transposons, similar to Arabidopsis where approximately 10% of the genome corresponds to transposable elements (The Arabidopsis Genome Initiative 2000). Transposons are common in plant genomes and are generally indicative of gene-poor regions with low recombination (Martienssen 1998). Interestingly, 47m20 is located in a region of the Populus genome with >25-fold suppressed recombination (Stirling et al. 2001). BAC 41g18, also derived from the same region of the Populus genome as BAC 47m20, had relatively few protein-coding sequences compared with the BAC clones derived from other regions of the Populus genome (Table 1).

The Populus BACs were not chosen at random but were known or suspected to have genes in them. It may well be that a random collection of Populus BACs would contain a larger proportion of retrotransposons, since these elements seem to be responsible for genome expansion in many plants (Bennetzen 2000).

Populus BAC clones 2c5, 16j18, and 6k8 are known to carry Populus homologs of PHYTOCHROME B, teosinte branched1, and ABSCISIC ACID INSENSITIVE1, respec-
As shown in Table 1, orthologs for PHYTOCHROME B and ABSCISIC ACID INSENSITIVE1 were both identified with sequenced contigs. Sequenced contigs from BAC 16j18 did not include a teosinte branched1 ortholog in the Arabidopsis protein database at the established threshold, even though Arabidopsis is known to contain two homologs of maize teosinte branched1. This is presumably due to the incomplete (approximately 50%) sequence coverage for each BAC.

**Table 1.** List of 33 poplar contigs and their Arabidopsis orthologs.

<table>
<thead>
<tr>
<th>Populus clone</th>
<th>Protein^a</th>
<th>Arabidopsis chromosome</th>
<th>Arabidopsis BAC clone</th>
<th>Approximate map position on the Arabidopsis chromosome (bp)</th>
<th>NCBI^b RefSeq accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>2c5</td>
<td>RGA1-like protein</td>
<td>III</td>
<td>T21P5</td>
<td>820 300</td>
<td>NM_111216</td>
</tr>
<tr>
<td>2c5</td>
<td>Putative protein kinase</td>
<td>III</td>
<td>F7018</td>
<td>274 640</td>
<td>NM_111341</td>
</tr>
<tr>
<td>2c5</td>
<td>Putative betaine aldehyde dehydrogenase^a</td>
<td>III</td>
<td>T8E24</td>
<td>2 097 242</td>
<td>NM_111545</td>
</tr>
<tr>
<td>2c5</td>
<td>Unknown protein^a</td>
<td>III</td>
<td>T8E24</td>
<td>2 089 253</td>
<td>NM_111543</td>
</tr>
<tr>
<td>2c5</td>
<td>Putative ATP-citrate lyase</td>
<td>III</td>
<td>T8E24</td>
<td>2 084 704</td>
<td>NM_111541</td>
</tr>
<tr>
<td>2c5</td>
<td>Putative protein kinase</td>
<td>III</td>
<td>T8E24</td>
<td>2 076 476</td>
<td>NM_111540</td>
</tr>
<tr>
<td>2c5</td>
<td>Putative E2 ubiquitin conjugating enzyme^a</td>
<td>III</td>
<td>K14A17</td>
<td>5 798 415</td>
<td>NM_112576</td>
</tr>
<tr>
<td>2c5</td>
<td>Expressed protein^a</td>
<td>III</td>
<td>K14A17</td>
<td>5 803 326</td>
<td>NM_112578</td>
</tr>
<tr>
<td>2c5</td>
<td>Hypothetical protein^a</td>
<td>III</td>
<td>K14A17</td>
<td>5 806 537</td>
<td>NM_112579</td>
</tr>
<tr>
<td>2c5</td>
<td>Unknown protein^a</td>
<td>III</td>
<td>K14A17</td>
<td>5 810 960</td>
<td>NM_112580</td>
</tr>
<tr>
<td>2c5</td>
<td>Putative protein^a</td>
<td>III</td>
<td>F3A4</td>
<td>18 595 140</td>
<td>NM_114872</td>
</tr>
<tr>
<td>2c5</td>
<td>PHYB</td>
<td>III</td>
<td>MSF3</td>
<td>8 090 711</td>
<td>NM_127435</td>
</tr>
<tr>
<td>16j18</td>
<td>Unknown protein^a</td>
<td>I</td>
<td>F14G11</td>
<td>9 003 400</td>
<td>NM_102370</td>
</tr>
<tr>
<td>16j18</td>
<td>Unknown protein^a</td>
<td>I</td>
<td>F14G11</td>
<td>9 021 923</td>
<td>NM_102374</td>
</tr>
<tr>
<td>16j18</td>
<td>Putative cytochrome b561</td>
<td>I</td>
<td>F14G11</td>
<td>9 023 356</td>
<td>NM_102375</td>
</tr>
<tr>
<td>16j18</td>
<td>Expressed protein^a</td>
<td>I</td>
<td>F14G11</td>
<td>9 025 967</td>
<td>NM_102376</td>
</tr>
<tr>
<td>16j18</td>
<td>Putative nuclear matrix constituent protein</td>
<td>I</td>
<td>F14K14</td>
<td>25 489 000</td>
<td>NM_105552</td>
</tr>
<tr>
<td>16j18</td>
<td>Putative DNA-binding protein^a</td>
<td>I</td>
<td>F14K14</td>
<td>25 513 980</td>
<td>NM_105555</td>
</tr>
<tr>
<td>16j18</td>
<td>Expressed protein^a</td>
<td>I</td>
<td>F14K14</td>
<td>25 519 240</td>
<td>NM_105556</td>
</tr>
<tr>
<td>16j18</td>
<td>Hypothetical protein^a</td>
<td>I</td>
<td>T5M7</td>
<td>9 360 858</td>
<td>NM_113469</td>
</tr>
<tr>
<td>6k8</td>
<td>Rab geranylgeranyl transferase-like protein</td>
<td>IV</td>
<td>F22K18</td>
<td>11 621 180</td>
<td>NM_118582</td>
</tr>
<tr>
<td>6k8</td>
<td>Protein phosphatase ABI1</td>
<td>IV</td>
<td>F20B18</td>
<td>12 185 439</td>
<td>NM_118741</td>
</tr>
<tr>
<td>6k8</td>
<td>Putative mitochondrial carrier protein</td>
<td>IV</td>
<td>F20B18</td>
<td>12 225 553</td>
<td>NM_118751</td>
</tr>
<tr>
<td>6k8</td>
<td>Putative protein kinase</td>
<td>IV</td>
<td>F4I10</td>
<td>14 926 435</td>
<td>NM_119462</td>
</tr>
<tr>
<td>6k8</td>
<td>Aminopeptidase-like protein^a</td>
<td>IV</td>
<td>F4I10</td>
<td>14 932 685</td>
<td>NM_119463</td>
</tr>
<tr>
<td>6k8</td>
<td>Putative protein^a</td>
<td>IV</td>
<td>F4I10</td>
<td>14 940 748</td>
<td>NM_119466</td>
</tr>
<tr>
<td>6k8</td>
<td>Lysine decarboxylase-like protein V</td>
<td>IV</td>
<td>F14F18</td>
<td>3 855 954</td>
<td>NM_121233</td>
</tr>
<tr>
<td>6k8</td>
<td>Putative protein^a</td>
<td>V</td>
<td>F14F18</td>
<td>3 860 332</td>
<td>NM_121234</td>
</tr>
<tr>
<td>6k8</td>
<td>Putative receptor-like kinase V</td>
<td>V</td>
<td>F14F18</td>
<td>3 875 470</td>
<td>NM_121238</td>
</tr>
<tr>
<td>6k8</td>
<td>Unknown protein^a</td>
<td>II</td>
<td>TID16</td>
<td>11 103 115</td>
<td>NM_128178</td>
</tr>
<tr>
<td>41g18</td>
<td>Pectinesterase-like protein^a</td>
<td>IV</td>
<td>FIN20</td>
<td>10 629 479</td>
<td>NM_118322</td>
</tr>
<tr>
<td>41g18</td>
<td>Putative protein</td>
<td>IV</td>
<td>FIN20</td>
<td>10 638 520</td>
<td>NM_118324</td>
</tr>
<tr>
<td>41g18</td>
<td>Expressed protein^a</td>
<td>I</td>
<td>F25C20</td>
<td>3 986 248</td>
<td>NM_101052</td>
</tr>
</tbody>
</table>

^aProtein identification based on results from BLASTP (E < 10^-19, s > 90) searches against the Arabidopsis protein database. Single-hit orthologs (see text) are followed by an asterisk.

^bNational Center for Biotechnology Information.

tively. As shown in Table 1, orthologs for PHYTOCHROME B and ABSCISIC ACID INSENSITIVE1 were both identified with sequenced contigs. Sequenced contigs from BAC 16j18 did not include a teosinte branched1 ortholog in the Arabidopsis protein database at the established threshold, even though Arabidopsis is known to contain two homologs of maize teosinte branched1. This is presumably due to the incomplete (approximately 50%) sequence coverage for each BAC.

**Genome synteny and collinearity between Populus and Arabidopsis**

Although Populus and Arabidopsis differ greatly in many evolutionarily interesting characters such as growth habit (woody perennial versus herbaceous annual), adult size (2 x 10^6 versus 2 x 10^-1 g), mating system (dioecious outcrosser versus selfer), and ecology (dominant, perennial, shade-intolerant pioneer versus subdominant, annual, shade-intolerant ruderal), they are both members of the Eurosid clade within the Core Eudicots (Solit et al. 1999; Stevens 2001). *Populus* is in the order Malpighiales within the Eurosid I clade, while *Arabidopsis* is in the order Brassicales within the Eurosid II clade. The relatively close phylogenetic relationship between *Populus* and *Arabidopsis*, combined with the drastic contrasts in their adaptive phenotypes, makes comparative genomics a particularly attractive ap-
proach to understanding how trees evolve from herbaceous ancestors.

As a beginning for comparative genomics between *Populus* and *Arabidopsis*, sequenced segments derived from five different *Populus* BAC clones were compared with their counterparts in the *Arabidopsis* genome to assess the degree of synteny and microcollinearity between the genomes of these two species. Orthology between *Populus* and *Arabidopsis* genes was inferred from BLAST search results. This is probably an overly simplistic view of orthology (Theissen 2002), but until more plant genomes from across the phylogenetic spectrum are sequenced in their entirety, it will be difficult to reconstruct orthology relationships in a more sophisticated manner.

Gene duplication and the consequent proliferation of large gene families in plant genomes can make it difficult to determine the orthology of genes derived from different species (Doyle and Gaut 2000). For large gene families, collinearity in the flanking regions combined with sequence similarity could help to determine orthology. This approach was followed for the *Le-A* and *Le-D* genes located in the *Lateral suppressor* region of tomato chromosome 7 (Schumacher et al. 1999; Rossberg et al. 2001). We also used a similar approach to determine the most probable *Arabidopsis* orthologs for protein families identified with sequenced segments from the *Populus* genome. Candidate orthologs were selected based on high sequence similarity and conserved microcollinearity with proteins that had a clear one-to-one orthologous relationship.

**Comparison of the *Populus* and *Arabidopsis* genomes**

An initial BLASTX search revealed that 27% of the *Populus* sequences had significant homology to *Arabidopsis* protein-coding sequences in the NR protein database, suggesting that *Arabidopsis* and *Populus* share many genes. The sequences with no significant matches may represent genes unique to *Populus* that have no orthologs in *Arabidopsis*, as may be the case for up to 45% of rice genes (Goff et al. 2002; Yu et al. 2002). Alternatively, the sequence with no significant matches could represent noncoding regions, which are likely to be more expanded in *Populus* because of its larger genome (approximately 550 Mb; Bradshaw and Stettler 1993) compared with *Arabidopsis*. In support of this idea, Ku et al. (2000) found that both introns and intergenic spacer regions were longer in tomato (approximately 900-Mb haploid genome; Galbraith et al. 1983; Tanksley 1987). It is also possible that these “unique” regions of the genome represent faster-evolving genes; hence, homologs are no longer recognizable. Comparative analyses of eukaryotic genomes have demonstrated that approximately 30% of the predicted proteins in every organism bear no similarity to either other members of its own protein-coding sequences or the protein-coding sequences of other organisms (Rubin et al. 2000). Given the probable polyploid origin of the *Populus* genome, it is also possible that these “unique” regions represent fast-evolving pseudogenes following genome duplication. All plant genomes may not just be slight variants of the *Arabidopsis* gene set but may include a wide variety of genes that have no orthologs in *Arabidopsis*. Improved gene prediction algorithms and more refined genome annotation based on experimental verification of genes believed to lack orthologs between pairs of species will be needed to resolve this issue.

**Table 2. Extent of synteny and microcollinearity based on the number and location of orthologs identified in the *Arabidopsis* genome with sequences from five *Populus* BAC clones.**

<table>
<thead>
<tr>
<th><em>Populus</em> BAC Clone</th>
<th>Single-hit orthologs</th>
<th>Orthologs inferred by synteny</th>
<th><em>Arabidopsis</em> observed chromosomal syntenic regions (%)</th>
<th><em>Arabidopsis</em> observed microcollinearity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2c5</td>
<td>7</td>
<td>4</td>
<td>III, 100&lt;sup&gt;b&lt;/sup&gt;, 83&lt;sup&gt;c&lt;/sup&gt;</td>
<td>33&lt;sup&gt;c&lt;/sup&gt;, 18&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>16j18</td>
<td>0</td>
<td>1</td>
<td>I, 67&lt;sup&gt;b&lt;/sup&gt;, 75&lt;sup&gt;c&lt;/sup&gt;</td>
<td>27&lt;sup&gt;c&lt;/sup&gt;, 32&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>6k8</td>
<td>1</td>
<td>0</td>
<td>II, 17&lt;sup&gt;b&lt;/sup&gt;, 40&lt;sup&gt;c&lt;/sup&gt;</td>
<td>17&lt;sup&gt;b&lt;/sup&gt;, 16&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>41g18</td>
<td>1</td>
<td>0</td>
<td>I, 0&lt;sup&gt;b&lt;/sup&gt;, 33&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0&lt;sup&gt;b&lt;/sup&gt;, 33&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>47m20f</td>
<td>1</td>
<td>1</td>
<td>IV, 0&lt;sup&gt;b&lt;/sup&gt;, 33&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0&lt;sup&gt;b&lt;/sup&gt;, 33&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>14</td>
<td>Average, 46&lt;sup&gt;b&lt;/sup&gt;, 58&lt;sup&gt;c&lt;/sup&gt;</td>
<td>19&lt;sup&gt;b&lt;/sup&gt;, 25&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>*Arabidopsis* chromosomes lacking orthologs of genes on a *Populus* BAC are not shown.

<sup>b</sup>Estimate based on the 19 orthologous relationships inferred by single hits in the BLASTP search of the *Arabidopsis* protein database.

<sup>c</sup>Estimate when the 14 gene family members whose orthologous relationships were inferred by synteny and collinearity are included with the 19 single-hit orthologs.

<sup>d</sup>No genes were found on BAC 47m20.

**Synteny and collinearity between the genomes of *Arabidopsis* and distantly related taxa**

Most comparative genomic studies have focused on comparing genomes between very closely related taxa, such as species within the same genus (Tanksley et al. 1992; Lagercrantz and Lydiate 1996) or family (Kowalski et al. 1994; Chen et al. 1997; Lagercrantz 1998; Acarkan et al. 2000; O’Neill and Bancroft 2000). Attempts to analyze genome organization between more distantly related species have been difficult, primarily because of reliance on genetic linkage maps for comparison. Nonetheless, a few studies have demonstrated genome collinearity for distantly related
species (Paterson et al. 1996; Ku et al. 2000; Grant et al. 2000; Mayer et al. 2001; Rossberg et al. 2001). For distantly related species, the length of conserved gene content and order is expected to be small (Paterson et al. 1996).

In a study of genome collinearity between Arabidopsis and tomato (Solanum, order Solanales within the Asterid clade of Core Eudicots), Ku et al. (2000) compared all of the genes encoded in a 105-kb BAC clone located on tomato chromosome 2 with its homoeologous counterparts in Arabidopsis. Rather than aligning to a single region of the Arabidopsis genome, the tomato BAC clone showed conservation of gene content and order with four different segments of Arabidopsis chromosomes II and V.

At an even greater phylogenetic distance, the completion of a draft of the rice genome has permitted analysis of synteny and collinearity between a monocot (rice) and a eudicot (Arabidopsis) (Goff et al. 2002; Yu et al. 2002). The approximately 200 × 10^6 years of evolution separating these taxa has scrambled but not completely randomized the chromosomal positions of their orthologous genes. While just 2% of rice gene pairs have immediately adjacent orthologs in Arabidopsis, more than half of rice gene pairs have orthologs separated by fewer than 150 intervening genes in Arabidopsis (Goff et al. 2002; see Web link 12, http://www.science.org/cgi/content/full/296/5565/92/DC1).

It is clear that there is substantial collinearity between the Populus and Arabidopsis genomes, at least on the scale of BAC-sized chromosomal fragments (approximately 120 kb). Although the exact order of genes within the Populus BACs is unknown, because the low-pass sequence data could not be assembled into a single contig for each BAC, the presence of Arabidopsis orthologs in a BAC-sized (approximately 120 kb) piece of the genome was considered good evidence for microcollinearity. Roughly 25% of the time, any adjacent pair of genes on a Populus BAC will have orthologs on a single BAC in Arabidopsis, suggesting that DNA sequence and gene location data in Arabidopsis could be used to inform positional cloning efforts in Populus in a large proportion of cases. For example, QTL maps in Populus could be anchored with markers based on Populus–Arabidopsis orthologous genes; then the regions containing the Populus QTLs could be scanned in the orthologous region of the Arabidopsis genome to develop a list of candidate genes for the Populus QTL.

Our estimates of the extent of microcollinearity between the Populus and Arabidopsis genomes are conservative, being downwardly biased by the potential for nonoverlap between two BACs cloned from each genome and by our inability to recognize rapidly diverging genes as orthologs. The Populus genome will be among the next to be completely sequenced (Mann and Plummer 2002; http://bahama.jgi-pacif.org/projdb/bin/populus/home.populus.cgi), at which time, a more comprehensive comparison with the Arabidopsis genome will be possible.

Acknowledgements

We thank Carol Loopstra and Julia Vrebalov for the construction of the Populus BAC library, Maynard Olson and Kerry Bubb for their assistance with the PHRAP software program, Barbara Frewen, Xuesong Yu, and Brian Watson for their assistance in the laboratory and greenhouse, and two anonymous reviewers and the Associate Editor for helpful comments on the manuscript. This work was supported by members of the Poplar Molecular Genetics Cooperative and by the U.S. Department of Energy Bioenergy Feedstock Development Program (contract ST806-19).

References


© 2003 NRC Canada
Stirling et al.


Lagercrantz, U. 1998. Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics, 150: 1217–1228.


© 2003 NRC Canada