Extensive Structural Renovation of Retrogenes in the Evolution of the *Populus* Genome[1][W][OA]

Zhenglin Zhu, Yong Zhang, and Manyuan Long*

Center for Bioinformatics, College of Life Sciences, Peking University, Beijing 100871, China (Z.Z., M.L.); and Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 (Y.Z., M.L.)

Retroposition, as an important copy mechanism for generating new genes, was believed to play a negligible role in plants. As a representative dicot, the genomic sequences of *Populus* (poplar; *Populus trichocarpa*) provide an opportunity to investigate this issue. We identified 106 retrogenes and found the majority (89%) of them are associated with functional signatures in sequence evolution, transcription, and (or) translation. Remarkably, examination of gene structures revealed extensive structural renovation of these retrogenes: we identified 18 (17%) of them undergoing either chimerization to form new chimerical genes and (or) intronization (transformation into intron sequences of previously exonic sequences) to generate new intron-containing genes. Such a change might occur at a high speed, considering eight out of 18 such cases occurred recently after divergence between Arabidopsis (*Arabidopsis thaliana*) and *Populus*. This pattern also exists in Arabidopsis, with 15 intronized retrogenes occurring after the divergence between Arabidopsis and papaya (*Carica papaya*). Thus, the frequency of intronization in dicots revealed its importance as a mechanism in the evolution of exon-intron structure. In addition, we also examined the potential impact of the *Populus* nascent sex determination system on the chromosomal distribution of retrogenes and did not observe any significant effects of the extremely young sex chromosomes.

Retroposition is a process in which mRNAs are reverse transcribed and incorporated back into new genomic positions (Brosius, 1991). A retrocopy usually becomes a pseudogene due to the loss of its original regulatory elements like promoters. However, by chance, if the retrocopy fortunately recruits new regulatory elements, it might become a functional retrogene. Moreover, if the retrogene recruits some new coding regions that do not exist in the parental gene, it becomes a retroposed chimeric gene (Betrán et al., 2002; Wang et al., 2002; Nisole et al., 2004; Sayah et al., 2004; Zhang et al., 2004).

As a copy mechanism to generate new genes (Brosius, 1991; Kaessmann et al., 2009), retroposition is widely observed in animals, such as in mammals or Drosophila (*Drosophila melanogaster*; Betrán et al., 2002; Emerson et al., 2004; Pan and Zhang, 2009). In contrast, knowledge of retrogenes in plants is so far limited. As a pioneering study, alcohol dehydrogenase derived retrogenes were observed in *Brassicaceae* (Charlesworth et al., 1998). Recently, a genome-wide scan detected abundant retroposed chimeric genes in the rice (*Oryza sativa*) genome, revealing unexpected high protein diversity encoded by the monocot genome (Wang et al., 2006). In contrast, only one retroposed chimeric gene was observed in Arabidopsis (*Arabidopsis thaliana*; Zhang et al., 2005). These observations raised new questions: How general is the retroposition mechanism in plants? Is the observed high origination rate of chimeric genes in rice a general mechanism for generating new protein functions in other plants, e.g. in dicot organisms? Clearly, we need to investigate more plant genomes to understand these problems.

In animals, the fate of retrogenes in evolution appears to be significantly impacted by the sex determination genetic system. For example, the meiotic sex chromosome inactivation may be responsible for driving male genes out of the X chromosome in mammals, and sexual antagonisms may enhance the fixation probability of new male genes into autosomes in Drosophila (Rice, 1984; Sturgill et al., 2007). The origination of meiotic sex chromosome inactivation and the male retrogene movement were found to be associated (Potrzebowski et al., 2008). The genetic control of sex determination and related genetic processes were a consequence of a long evolutionary process in which a pair of sex chromosomes was completely developed. However, it is unclear when the patterns of sex genes, e.g. the movement of male genes from the X chromosome to autosomes, were formed during the evolution of the sex chromosomes.

After one dicot Arabidopsis and one monocot rice were sequenced, *Populus* (poplar; *Populus trichocarpa*) is the first tree whose genome had been sequenced (Tuskan et al., 2006). Unlike Arabidopsis and rice,

[1] This work was supported by the 863 grants and the MOST 973 grants, a Ministry of Education Yingzhi Program and Changjiang Adjunct Professorship Program to M.L., and a National Institutes of Health grant to M.L. to investigate new gene evolution.

* Corresponding author; e-mail mlong@uchicago.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Manyuan Long (mlong@uchicago.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.109.142984

Populus's sex determination was recently found to be genetically controlled (Tuskan et al., 2006; Yin et al., 2008). Specifically, a Z-W sex determination system occurs, i.e. the female is a heterogametic sex (ZW), while the male contains a pair of homogametic sex chromosomes (ZZ). Furthermore, the system was observed to be primitive: a dominant part of the W chromosome is today homologous to the Z chromosome, clearly showing its autosomal origination (Charlesworth et al., 2005), with only <5% of the maternal W chromosome subjected to recombination suppression (a fragment of 706 kb; Yin et al., 2008).

Thus, we scanned the *Populus* genome (Tuskan et al., 2006) to investigate all the issues above. In this report, we will show that the *Populus* genome contains >100 retrogenes, and the vast majority of them are likely functional. Furthermore, a considerable portion of these retrogenes underwent extensive structural renovation that led to origination of new chimerical genes as a consequence of recruiting preexisting exons in the target sites and new intron-containing genes resulting from intronization of previous exonic sequences of retrosequences (Irimia et al., 2008). We will finally show that the incipient sex chromosome does not have a strong effect on the expression of sex retrogenes and localization of sex genes.

RESULTS

Identification of Retrogenes

By integrating and improving previous strategies (Emerson et al., 2004; Wang et al., 2006; Zhang et al., 2006b), we developed a new pipeline to identify retrocopies in the *Populus* genome. We searched 45,555 predicted gene models (Joint Genome Institute [JGI] *Populus trichocarpa* v1.1; Tuskan et al., 2006) against the genome and found 106 retrocopies (Fig. 1). To estimate the performance of our pipeline, we identified 83 retrocopies in Arabidopsis with the same pipeline and compared this result with the previous report (Zhang et al., 2005). Out of the 69 cases previously identified, only 32 are covered by our data set. As for the remaining 37 cases, they either failed to pass our stringent criteria or were lost due to the database update (for details, see Supplemental Table S1).

Significant Functionality of Retrogenes

We combined three complementary methods to test the functionality of 106 *Populus* retrogenes. First, we implemented the likelihood ratio test in the PAML package (Nei and Gojobori, 1986; Yang, 1997) to infer whether the ratio between the nonsynonymous sub-

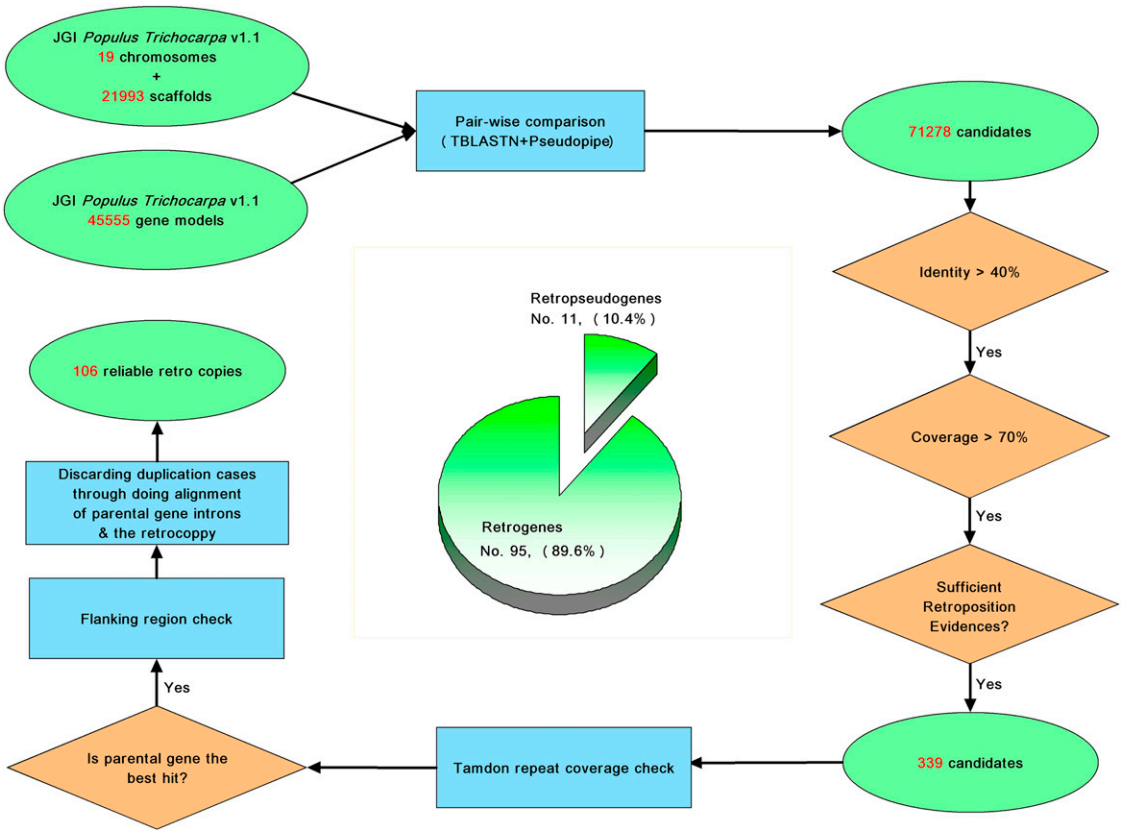

Figure 1. A schema of our pipeline (for details, see "Materials and Methods").

stitution rate and the synonymous substitution rate (*Ka/Ks*) would be significantly smaller than 0.5 if we compare the parental gene and retrogene. In this context, the cutoff of 0.5 is established as a stringent signature of evolutionary constraint between paralogous genes (Betrán et al., 2002; Emerson et al., 2004). A total of 75 cases (70.8%) show significant sequence constraint (likelihood ratio test, $P \leq 0.05$).

Second, we searched for evidence of transcription. We scanned PopulusDB, a comprehensive EST database (Sterky et al., 2004) that covers 19 samples from different tissues or development stages. Meanwhile, we also searched the microarray tissue profiling data GSE13990 (Wilkins et al., 2009), which covers nine tissues. As a result, we identified 36 retrogenes (33.9%) that overlap with at least one EST or show presence in all three replicates for at least one tissue by microarray.

Finally, we mapped retrogenes to the protein-coding gene annotation (JGI *Populus trichocarpa* v1.1; Tuskan et al., 2006) according to the chromosomal coordinates. A retrogene will be recognized as functional if it overlaps with one protein-coding gene on the same strand and the overlapping open reading frame region contributed to at least 50 amino acids of the protein. As a result, we found 90 retrogenes (84.9%) that were annotated as protein-coding genes.

We identified a total of 95 retrogenes (89.6% out of 106 retrocopies) showing signature of transcription, gene annotation, or undergoing significant purifying selection. In other words, the majority of retrocopies in this dicotyledonous species was associated with significant functionality. This is in sharp contrast to the human genome, in which only 16% of retrocopies are potentially functional, while the remaining majority seems to be functionless (Vinckenbosch et al., 2006).

Extensive Structural Renovation of Retrogenes

As a main source of functional diversity, retrogenes often generate a chimeric structure by recruiting nearby preexisting exons (for example, see Fig. 2), introns, or intergenic regions (Long and Langley, 1993; Long et al., 2003). Out of 90 retrogenes overlapping with annotated gene models, we identified 12 (13%) chimeric genes (Table I), nine of which are involved in fusion of the coding region. In comparison, the chimeric retrogenes account for 23% (19 out of 83) in Arabidopsis, possibly due to its more comprehensive annotation of gene structure. Again, the majority (13 out of 19) has a chimeric coding region. Such data suggest a frequent functional divergence on the protein level.

While manually checking the structure of retrogenes, we found a second structural renovation category, intronization. Instead of recruiting nearby regions, retrocopy could splice out one segment of previous protein-coding retrosequences and generate a new intron. We found a total of 11 (10%) cases in *Populus* and 18 (22%) cases in Arabidopsis (Table II). As shown in Figure 3, a hydrolase, AT1G66860, generated a retroposed copy, AT1G15040. This new gene has two

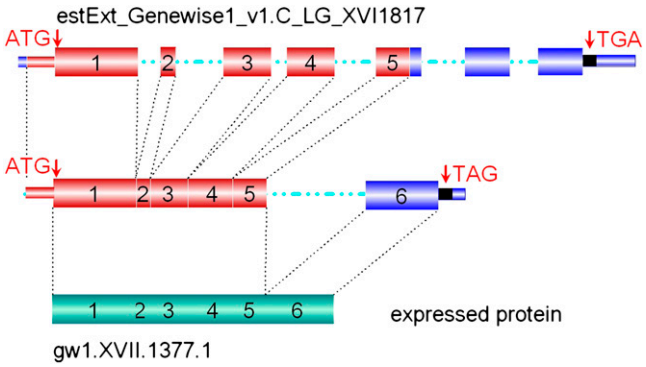

Figure 2. An example of chimeric gene in *Populus*. Thick bars indicate coding exonic regions, while thin bars represent noncoding exonic regions. The black thin bar is the stop codon area. Retroposed regions are marked in red, and nonretroposed regions are marked in blue. The region spanning five exons (1–5) of the parental gene estExt_Genewise1_v1.C_LG_XVI1817 are retroposed and fused with a nearby coding exon 6, creating a new chimeric structure retrogene, gw1.XVII.1377.1, which might be involved in transcription initiation (KO annotation; PopulusDB; Sterky et al., 2004).

distinct isoforms from alternative splicing due to an intron retention event (Fig. 4). The spliced isoform is the major form since 10 cDNA or EST sequences unambiguously support this form, and only two cDNA sequences are compatible with the retention form.

Both the protein-level and nucleotide-level sequence alignments between the parental gene and retrogene show this new intron was derived from the exonic region of the parental gene (Fig. 5). However, as expected, with intronization the divergence between the parental gene and this new intron is much higher compared to the immediate flanking region. The nucleotide sequence alignment shows the splicing donor GT is generated de novo, while the splicing acceptor AG is shared by both the parental gene and the retrogene (Supplemental Fig. S1). Sequence alignment shows this intron is conserved in Arabidopsis and *Arabidopsis lyrata* (Supplemental Fig. S2). The spliced isoform serves as the major isoform, which also suggests the functionality of this new intron.

Structure Renovation Occurs at a High Speed

Next, we estimated the ages of these chimeric genes or intronized retrogenes. Since the synonymous mutation rate (*Ks*) is sometimes distorted by gene conversion or inconstancy of the molecular clock, we dated their ages by investigating the related synteny based on genome-level alignments in related species whose genome sequences are available ("Materials and Methods").

For the 18 *Populus* retrogenes that undergo either chimerization and or intronization, we found that eight retrogenes are *Populus* specific and do not exist in the other dicot species, Arabidopsis, *Vitis vinifera*, and the monocot species rice. Considering the diver-

Table I. *List of chimeric retrogenes at Arabidopsis and Populus*

A. T., Arabidopsis; P. T., *P. trichocarpa*.

Species	Retrogene		Parental Gene		Hallmarks	
	Accession No.	Synteny[a]	Accession No.	Synteny[a]	*Ka/Ks*	Type[b]
A. T.	AT1G20000.1	N;N;N;N;Y	AT4G20280	Y;Y;N;Y;Y	0.20	C
	AT1G34130	Y;Y;Y;Y;Y	AT5G19690	Y;Y;Y;Y;Y	0.01	C
	AT1G73050	Y;Y;Y;Y;Y	AT1G72970	Y;Y;Y;Y;Y	0.04	C
	AT1G77130	Y;Y;Y;Y;Y	AT3G18660	Y;Y;Y;Y;Y	0.03	C
	AT2G01180.2	Y;Y;Y;Y;Y	AT3G02600	Y;Y;Y;Y;Y	0.03	C
	AT2G24810	N;N;N;N;Y	AT1G19320	N;Y;Y;Y;Y	0.35	C
	AT4G11485	N;N;N;N;Y	AT4G11760	N;N;N;N;Y	0.52	C
	AT4G11760	N;N;N;N;Y	AT4G11485	N;N;N;N;Y	0.74	C
	AT4G14250	N;N;N;N;Y	AT1G14570	Y;Y;Y;Y;Y	0.19	C
	AT4G35680	N;N;N;N;Y	AT4G01590	N;N;N;N;Y	0.33	C
	AT5G39840	Y;Y;Y;Y;Y	AT4G14790	Y;Y;Y;Y;Y	0.07	C
	AT5G56720	Y;Y;N;Y;Y	AT1G04410	Y;Y;Y;Y;Y	0.02	C
	AT5G59240	Y;N;Y;Y;Y	AT5G20290	Y;Y;N;Y;Y	0.05	C
	AT3G47520	Y;Y;Y;Y;Y	AT5G09660	Y;Y;Y;Y;Y	<0.01	N
	AT3G62350	N;N;N;N;Y	AT1G71320	N;N;N;N;Y	0.52	N
	AT5G02920	N;N;N;N;Y	AT5G02930	N;N;N;N;Y	0.58	N
	AT5G16510	Y;Y;Y;Y;Y	AT3G02230	Y;Y;Y;Y;Y	0.02	N
	AT5G54940	Y;N;Y;Y;Y	AT4G27130	Y;Y;Y;Y;Y	0.08	N
	AT5G63370	Y;Y;Y;Y;Y	AT1G67580	Y;Y;Y;Y;Y	0.13	N
P. T.	gw1.II.4117.1	Y;Y;N	estExt_Genewise1_v1.C_640287	Y;Y;Y	0.23	C
	fgenesh4_pm.C_LG_III000107	N;N;N	fgenesh4_pg.C_LG_I003004	N;N;N	0.37	C
	grail3.0019034401	Y;Y;N	gw1.IX.2754.1	Y;N;N	0.15	C
	gw1.VIII.2743.1	Y;Y;Y	eugene3.00010969	Y;Y;Y	0.01	C
	gw1.X.1338.1	Y;N;N	gw1.III.1437.1	Y;Y;Y	0.17	C
	estExt_Genewise1_v1.C_LG_XVII1045	Y;Y;Y	eugene3.00100771	Y;Y;Y	0.01	C
	gw1.XVII.1377.1	N;Y;N	estExt_Genewise1_v1.C_LG_XVI1817	Y;Y;N	<0.01	C
	gw1.148.143.1	Y;Y;Y	estExt_fgenesh4_pm.C_LG_XVIII0265	Y;Y;N	0.21	C
	fgenesh4_pg.C_scaffold_188000004	N;N;N	fgenesh4_pg.C_LG_X000230	N;N;N	0.75	C
	estExt_Genewise1_v1.C_LG_XIII1276	Y;Y;N	eugene3.01450035	Y;Y;N	0.24	N
	grail3.0020003201	Y;Y;Y	estExt_Genewise1_v1.C_1820077	Y;Y;N	<0.01	N
	grail3.0124002901	Y;Y;Y	estExt_fgenesh4_pg.C_LG_I0109	Y;Y;Y	0.02	N

[a]Synteny follows the format "rice; *Populus*; papaya; *V. vinifera*; *A. lyrata*" in Arabidopsis, and "rice; *V. vinifera*; Arabidopsis" in *Populus*, while Y and N indicate presence and absence, respectively. [b]Chimeric gene type. C indicates coding exon fusion chimeric gene, which means coding sequence stems both from the retrocopy and the host gene, while N represents noncoding fusion chimeric gene, where the coding sequence only originates from the retrocopy (Vinckenbosch et al., 2006).

gence time of these species (Tuskan et al., 2006), their ages should be younger than 100 million years. This gives a rate of structural innovation of around 0.1 events per million years per genome, comparable with the rate of retrogene chimerization in hominoid lineages (0.14; Marques et al., 2005; Vinckenbosch et al., 2006).

For Arabidopsis, we have closer outgroups to estimate the ages of the 31% (26 out of 83) of retrogenes with structural renovation. We found that 15 retrogenes occurred after the divergence between Arabidopsis and papaya (*Carica papaya*) 72 million years ago (Ming et al., 2008). Two of them even emerged after the split of Arabidopsis and *A. lyrata* five million years ago (Tang et al., 2007).

Such data not only proved that our pipeline really screened out those evolutionary young genes, but suggest chimerization occurs with a speed as high as that in the human genome (Marques et al., 2005; Vinckenbosch et al., 2006) and intronization happens at a higher speed compared to sporadically reported mammalian cases (Roy et al., 2003).

There Is No Retrogene Traffic Out of the Sex Chromosome in *Populus*

We did not find an excess of retrogene traffic out of the sex chromosome. There are no retrocopies duplicated from the parental gene encoded by XIX. Moreover, only two retrocopies jump into XIX, which is not very different with regard to the expectation based on the number of all genes. Contrary to this pattern, X chromosomes of humans and *Drosophila* generate disproportional retrogenes (Betrán et al., 2002; Emerson et al., 2004). Moreover, out of 32 retrogenes with unique probes, we found 6 (18.8%) retrogenes transcribed in female catkin, and two (6.3%) retrogenes transcribed in male catkin. The abundance is lower than that of root and mature leaf (28.1% for both). In contrast, >50% of retrogenes express in testis for both human (Emerson et al., 2004) and fly (Dai et al., 2006). These data indicate that the incipient Z/W chromosomes do not play a significant role with respect to retrogene origination.

Table II. *List of intronized retrogenes at Arabidopsis and Populus*

A. T., Arabidopsis; P. T., *P. trichocarpa*.

Species	Retrogene		Parental Gene		Hallmarks		
	Accession No.	Synteny[a]	Accession No.	Synteny[a]	*Ka/Ks*	Intron (−)[b]	Intron (+)[b]
A. T.	AT1G03300	N;N;N;N;Y	AT2G47230	N;N;N;N;Y	0.45	7	1
	AT1G08120	Y;N;N;N;N	AT5G26110	Y;Y;Y;Y;Y	0.16	5	1
	AT1G08135	N;N;Y;N;Y	AT1G08140	N;N;Y;N;Y	0.15	3	2
	AT1G15040[c]	N;Y;Y;Y;Y	AT1G66860	N;Y;Y;Y;Y	0.09	4	1
	AT1G30455[c]	N;N;N;N;N	AT2G45100	N;N;N;N;Y	0.89	10	1
	AT1G34130[c]	Y;Y;Y;Y;Y	AT5G19690	Y;Y;Y;Y;Y	0.01	17	1
	AT1G45100	N;N;N;N;Y	AT5G41690	N;N;N;N;Y	0.91	20	1
	AT1G72850[c]	N;N;N;N;Y	AT4G09420	N;N;N;N;Y	0.44	1	1
	AT1G77130[c]	Y;Y;Y;Y;Y	AT3G18660	Y;Y;Y;Y;Y	0.03	3	1
	AT2G24810	N;N;N;N;Y	AT1G19320	N;Y;Y;Y;Y	0.35	1	2
	AT3G05860[c]	N;Y;Y;Y;Y	AT2G28700	N;N;N;N;Y	0.36	2	1, 2[d]
	AT3G53550[c]	N;N;N;N;Y	AT1G05080	N;N;N;N;Y	0.48	2	4
	AT4G14250	N;N;N;N;Y	AT1G14570	Y;Y;Y;Y;Y	0.19	7	1
	AT4G16680	Y;Y;Y;Y;N	AT1G32490	Y;Y;Y;Y;Y	0.14	20	1
	AT4G19240	N;N;N;N;Y	AT3G43290	N;N;N;N;N	1.24	2	1
	AT4G35680	N;N;N;N;Y	AT4G01590	N;N;N;N;Y	0.33	2	1
	AT5G03980	N;N;N;N;Y	AT1G28580	Y;Y;N;Y;Y	0.10	4	1
	AT5G10880	Y;Y;Y;Y;Y	AT3G62120	Y;Y;Y;Y;Y	0.15	10	2
P. T.	fgenesh4_pg.C_LG_I001866	N;N;N	fgenesh4_pg.C_LG_I001869	N;N;N	0.14	5	1
	fgenesh4_pm.C_LG_III000107	N;N;N	fgenesh4_pg.C_LG_I003004	N;N;N	1.01	2	2
	eugene3.00050136	Y;Y;Y	gw1.V.2779.1	Y;Y;Y	0.43	1	1
	eugene3.00061026	Y;Y;N	gw1.145.154.1	Y;Y;N	0.49	4	1
	fgenesh4_pg.C_LG_VIII001079	N;N;N	eugene3.00151048	N;N;N	0.34	2	1
	gw1.XIII.3155.1	N;N;N	gw1.I.8037.1	Y;Y;Y	0.06	8	1
	estExt_Genewise1_v1.C_LG_XVII1045 [c]	Y;Y;Y	eugene3.00100771	Y;Y;Y	79.99	11	1
	eugene3.01250039	Y;Y;N	gw1.XIII.3170.1	N;N;N	0.96	4	1
	fgenesh4_pg.C_scaffold_15221000001	N;N;N	fgenesh4_pg.C_scaffold_218000001	N;N;N	0.43	4	1
	fgenesh4_pg.C_scaffold_5145000001	N;N;N	fgenesh4_pg.C_LG_XI001324	N;N;N	0.41	2	1
	eugene3.53350001	N;N;N	eugene3.00181086	Y;Y;N	0.38	2	1

[a]Synteny follows the format in Table I. [b]Intron gain/loss of retrogenes. Columns "Intron (−)" and "Intron (+)" list the number of introns absent with respect to the parental gene and the number of newly gained introns in the retrogene, respectively. [c]Newly gained introns of these retrogenes have EST evidence supporting the splicing junction. [d]The number of newly gained introns is 1 (AT3G05860.2, AT3G05860.3) or 2 (AT3G05860.1) because of alternative splicing.

DISCUSSION

In this report, we develop an efficient pipeline to scan retrocopies in the whole genome while excluding possible DNA level duplications and transposon-derived duplicates. Detailed comparison with previous reported datasets in Arabidopsis shows the specificity and sensitivity of our pipeline. With such an improved pipeline, we have identified 106 retrocopies in the *Populus* genome and found that most of them were associated with functional signatures. Previously, genome duplication was widely observed in plant genomes to be a dominant evolutionary process that significantly impacts important components of molecular evolution, e.g. gene contents and substitution rates. Our analysis of retrogenes in two dicots added a new element of gene evolution to this general picture of plant genome evolution.

Our analysis revealed that 17% of these retrogenes undergo structural renovation through recruitment of preexisting exons and intronization of previous coding regions in the retrosequences. These evolutionary changes in gene structure consequently generate new chimeric proteins and/or new expression patterns, leading to the rise of novel gene functions. Intriguingly, it seems common to recruit internal exonic sequences as new introns for these retrogenes in *Populus*, although there were exceptional cases reported in humans before (Lahn and Page, 1999; Baertsch et al., 2008). This finding adds a new concept to the conventional notion that retrogenes should not be always expected to be single-exon genes derived from retrosequences. Thus, if we search retrogenes based on comparisons of single-exon genes and multiple-exon genes, we might overlook these respliced retrogenes. More than that, it is hypothesized that a transient phase might be mandatory, during which gene regions are neither fully exonic nor fully intronic, like alternative splicing, since an instantaneous intronization is unlikely (Catania and Lynch, 2008). Although the aforementioned case in Arabidopsis does support this view, most of the cases we found in both plants

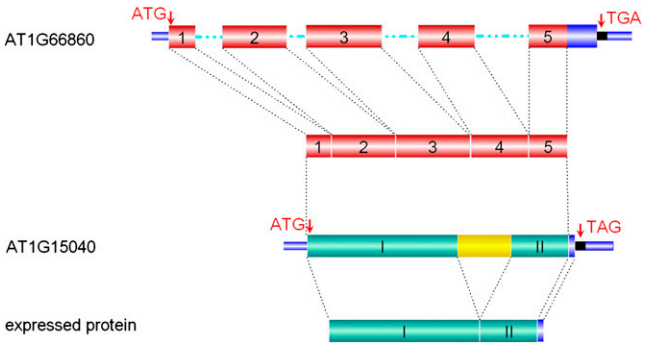

Figure 3. Intronization at retrogene AT1G15040. This figure follows the convention of Figure 2 except with a yellow bar marking the intronized region. Five exons of the parental gene AT1G66860 marked from 1 to 5 are duplicated as the retrogene AT1G15040. Then, an intronization event occurs: the previous retrocopy is spliced again and two new exons are created, which were marked with I and II, respectively. Furthermore, a chimeric event occurred in that the 3′ terminal coding region including the stop codon was recruited from the nearby region.

have only one spliced form. Nevertheless, given the gene redundancy, the new retroposed form is very likely to get a new intron in a short time due to the relaxation of functional constraint. Finally, regarding the largely unknown issue of intron origination (Roy and Irimia, 2008), the retrocopy provides a well-controlled system since all the preexisting introns get lost.

Only a small portion (<5%) of the sex chromosomes in the popular genome was differentiated from autosomes (Yin et al., 2008), suggesting that sex determination of *Populus* evolved very recently. This may explain our observation that there are no significant retrogene movements out of the sex chromosome. Moreover, in contrast to the dominant retrogene expression in the testes of mammals (Vinckenbosch et al., 2006), we did not find an enrichment of female catkin expression. These data reveal that this Z-W system is too young to impact the chromosomal distribution and expression of most retrogenes.

MATERIALS AND METHODS

Data Source

Populus trichocarpa genome data were downloaded from JGI (http://genome.jgi-psf.org; Tuskan et al., 2006), while Arabidopsis (*Arabidopsis thaliana*) genome data were from The Arabidopsis Information Resource (TAIR; www.arabidopsis.org; Poole, 2007). The EST data were from PopulusDB (www.populus.db.umu.se; Sterky et al., 2004), while the microarray data were from the Gene Expression Omnibus (Barrett et al., 2006) with accession number GSE13990 (Wilkins et al., 2009).

Identification of Retrocopies

By modifying previous strategies (Emerson et al., 2004; Wang et al., 2006; Zhang et al., 2006b), we developed an automatic pipeline to identify retrocopies in the genome. First, we mapped *Populus* protein sequences to genomes using TBLASTN (Altschul et al., 1997). Then, we implemented the Pseudopipe package (Zhang et al., 2006b) to process the raw alignments in merging BLAST high-score blocks, inferring the conceptual open reading frame based on FASTY (Pearson and Lipman, 1988) and identifying poly(A) tracts. We kept Pseudopipe's parameters, such as TBLASTN e-value cutoff (1e-10), coverage cutoff (70%), and identity cutoff (40%). Moreover, we corrected several bugs in the Pseudopipe package possibly caused by the update of the BLAST or FASTA package.

After that, we used several BioPerl-based (Stajich et al., 2002) scripts to scan the absence of parental introns, which map within the alignments between parents and retrocopies. Compared with Marques' strategy (Marques et al., 2005), which discarded small introns shorter than 80 bp, we retained small introns with the support of the canonical splicing site (GT-AG) and surveyed the distribution of noncanonical splicing sites because genuine tiny introns exist (Deutsch and Long, 1999; Chamary and Hurst, 2005). We kept small introns with the noncanonical splicing site GC-AG, for it contributes to 94 annotated introns with a total abundance only less than that of GT-AG. To account for possible intron loss for old nonprocessed duplicates, we kept all the cases with more than three introns absent in the retrocopy or only two introns absent in the retrogene, but having either a *Ks* value smaller than 2 or an identifiable poly(A) track. We also retained cases where only one parental intron is absent if they have a *Ks* value smaller than 2.0 and an identified poly(A) tract.

At this stage, we identified a pair of genes with one having fewer introns. However, a manual check showed a lot of them did not seem like retrogenes. Possibly because the *Populus* genome is full of DNA level duplications, like multiple rounds of genome-level duplication (Tuskan et al., 2006), the gene structure or intron loss predicted based on FASTY was not that reliable. Thus, we added five more filters. First, we discarded all the retrocopies with at least 50% of regions overlapped with repeats of RepPop (Zhou and Xu, 2009). Second, we discarded all the retrocopies with flanking genes similar to the parental gene's flanking region. Third, for retrocopies with one or two parental introns absent, we extended the introns 20 bp in both directions and aligned them with the retrocopies. If the whole region aligned well, we also considered it as a DNA-level duplication. Finally, because of possible retroposition followed by DNA-level duplication or recombination of a retrocopy with an intron-containing allele, multiple retrocopies might share one parental gene. In this case, we kept the retrocopies if they are more similar to the parental gene compared to all other retrocopies.

Identification of Chimerical Retrogene

To generate the complete structure of retrocopies, we compared their full-length parental transcripts and the retrocopies with their 10,000-bp flanking regions, considering we only used protein sequence as the query initially. Based on such sequence comparison, we inferred the borders for retropositions. Given the coordinates of the retroposition borders and annotated gene borders, we checked whether retrocopy already becomes chimeric by recruiting some new regions.

EST Profiling and Microarray Analysis

It is possible that an EST sequence can align well both with a retrogene and its parental gene. To ensure that an EST is derived from a retrogene, we followed a relatively complicated pipeline (Zhang et al., 2006a) to retain high-quality mappings. First, we mapped 141,158 EST sequences of PopulusDB (Sterky et al., 2004) to genome sequences using BLAT (Kent, 2002). With the default BLAT score of 30, 94% (133,284) ESTs were retained. Then, we discarded low-quality mappings that failed to meet the following criteria: mapping length ≥150 bp, identity ≥96%, coverage within mapping ≥97%, and coverage within whole transcript ≥75%. Only 86,695 ESTs were retained. If a transcript was mapped to multiple genomic loci, only the best mapping was retained; if more than one nearly identical best mapping existed (difference in BLAT scores <2%), the transcript was discarded to avoid ambiguity. Finally, we only kept 82,278 ESTs (58%). Subsequently, given chromosomal coordinates of 45,555 gene models and 309 retrocopies, we assigned ESTs to genes by checking whether ESTs had exonic overlap on the same strand after accounting for EST reading orientation. ESTs associated with multiple genes were discarded.

As for microarray data analysis, it is essential to generate high-quality mapping between gene and probe. We took a strategy similar to assigning ESTs to genes. Briefly, we downloaded the chromosomal location annotation of probes from the Affymetrix Web site (www.affymetrix.com). We assigned probes with nonambiguous chromosomal mappings to genes based on chromosomal coordinates and strand information. We excluded probes overlapping with multiple genes. For genes overlapping with multiple probes, we

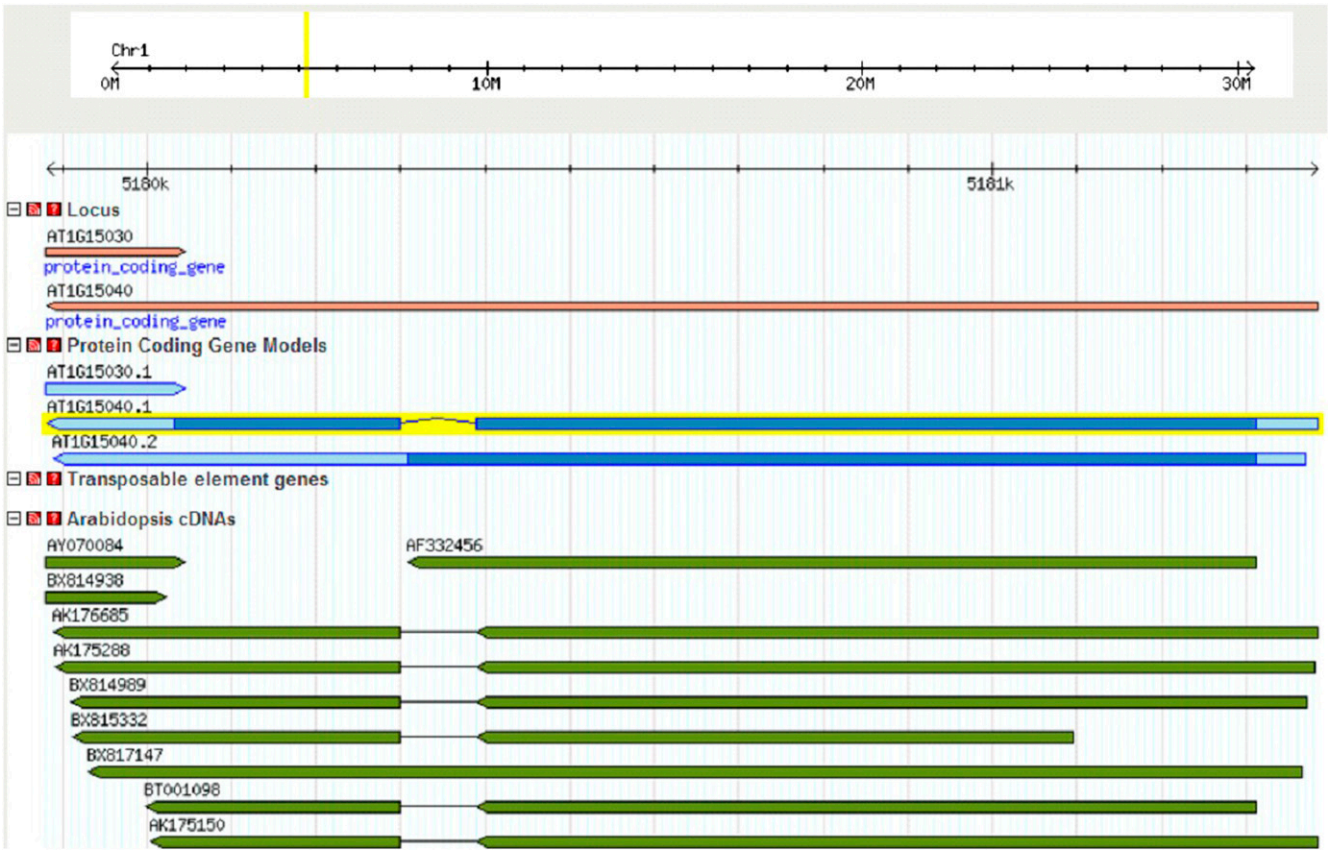

Figure 4. Expression evidence for intronization at retrogene AT1G15040 from TAIR (Poole, 2007). This snapshot of TAIR shows the gene model of AT1G15040 and its supporting cDNA tracks. The arrowed bars mark exons with the transcription orientation, while the thin connecting lines indicate the introns.

retained only the one with the highest alignment identity. If some probes mapped to the same gene with the same identity, probes with the "_at" would have the highest priority. The priority of "_s_at" and "_x_at" follows since their specificity decreased.

We downloaded raw microarray data from the Gene Expression Omnibus (Barrett et al., 2006) and processed them using the R-Bioconductor platform (Gentleman et al., 2004). Detection call was generated based on the MAS5 package (Pepper et al., 2007).

Function Analysis of Retrocopies

We investigated the functionality of *Populus* retrocopies using three strategies. First, we used the codeml program in the PAML package (Nei and Gojobori, 1986; Yang, 1997) to infer whether the *Ka/Ks* is significantly smaller than 0.5. A retrocopy is recognized as a functional candidate if it meets the following criteria ($Ka/Ks < 0.5$ and likelihood ratio test $P < 0.01$). Second, as described above, we generated the transcription profile of retrocopies using both ESTs and microarray data. Finally, if one retrocopy overlaps with the coding region of one cDNA sequence (JGI *Populus trichocarpa* v1.1; Tuskan et al., 2006) on the same strand and the overlapping open reading frame region contributed to at least 50 amino acids, we considered it to be functional.

Synteny Check of Retrocopies

We checked the synteny information of all retrocopies in Phytozome (www.phytozome.net). Specifically, we checked whether there is some VISTA (Frazer et al., 2004) genome alignment plot in the outgroup for the 40-kb window centered with the gene of our interest, like one *Populus* retrogene or one Arabidopsis retrogene X. If we find that gene X and at least one of its neighboring genes could form a syntenic chain (a series of VISTA alignment blocks), this gene should be shared by *Populus* or Arabidopsis with this outgroup species.

In Phytozome, we used three outgroups for *Populus* (Arabidopsis, *Vitis vinifera*, and rice [*Oryza sativa*]) and five outgroups for Arabidopsis (*Arabidopsis lyrata*, papaya [*Carica papaya*], *Populus*, *V. vinifera*, and rice). Based on the phylogenetic tree of all these species in Phytozome, we dated when genes emerged by following a parsimony rule.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Nucleotide-level sequence alignment (program Water of Emboss; Rice et al., 2000) indicates this new intron is derived from the exonic region of the parental gene.

Supplemental Figure S2. A snapshot of the alternative intron region at retrogene AT1G15040 between Arabidopsis and *A. lyrata* in www.phytozome.net.

Supplemental Table S1. Detailed comparison of our retrocopy screen results in Arabidopsis (83) with the previous report (Zhang et al., 2005).

ACKNOWLEDGMENTS

We thank Liping Wei, Jingchu Luo, Ge Gao, Lei Kong, Li Zhang, and the members of the Center for Bioinformatics for valuable discussion. We also thank Jan Karlsson and Stefan Jansson for help with PopulusDB.

Received June 16, 2009; accepted September 26, 2009; published September 29, 2009.

Figure 5. Protein-level alignment (tBLASTN; Altschul et al., 1997) of retrogene AT1G15040 and its parental gene AT1G66860. "Query" and "Sbjct" indicate parental gene and retrogene, respectively. Red downward arrows indicate the exon-exon border of parental genes, while the red upward arrows and light-blue connecting line mark the new intron in the retrogene.

LITERATURE CITED

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J (2008) Retrocopy contributions to the evolution of the human genome. BMC Genomics **9:** 466

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2006) NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res **35:** D760–D765

Betrán E, Thornton K, Long M (2002) Retroposed new genes out of the X in Drosophila. Genome Res **12:** 1854–1859

Brosius J (1991) Retroposons—seeds of evolution. Science **251:** 753

Catania F, Lynch M (2008) Where do introns come from? PLoS Biol **6:** e283

Chamary JV, Hurst LD (2005) Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet **21:** 256–259

Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity **95:** 118–128

Charlesworth D, Liu FL, Zhang L (1998) The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). Mol Biol Evol **15:** 552–559

Dai H, Yoshimatsu TF, Long M (2006) Retrogene movement within- and between-chromosomes in the evolution of Drosophila genomes. Gene **385:** 96–102

Deutsch M, Long M (1999) Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res **27:** 3219–3228

Emerson JJ, Kaessmann H, Betrán E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. Science **303:** 537–540

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res **32:** W273–W279

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW (2008) Origin of introns by 'intronization' of exonic sequences. Trends Genet **24:** 378–381

Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet **10:** 19–31

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res **12:** 656–664

Lahn BT, Page DC (1999) Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. Nat Genet **21:** 429–433

Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet **4:** 865–875

Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimerical processed functional gene in Drosophila. Science **260:** 91–95

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol **3:** e357

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature **452:** 991–997

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3:** 418–426

Nisole S, Lynch C, Stoye JP, Yap MW (2004) A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. Proc Natl Acad Sci USA **101:** 13324–13328

Pan D, Zhang L (2009) Burst of young retrogenes and independent retrogene formation in mammals. PLoS One **4:** e5040

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA **85:** 2444–2448

Pepper SD, Saunders EK, Edwards LE, Wilson CL, Miller CJ (2007) The utility of MAS5 expression summary and detection call algorithms. BMC Bioinformatics **8:** 273

Poole RL (2007) The TAIR database. Methods Mol Biol **406:** 179–212

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, Kaessmann H (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol **6:** e80

Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet **16:** 276–277

Rice WR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution **38:** 735–742

Roy SW, Fedorov A, Gilbert W (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci USA **100:** 7158–7162

Roy SW, Irimia M (2008) Mystery of intron gain: new data and new models. Trends Genet **25:** 67–73

Sayah DM, Sokolskaja E, Berthoux L, Luban J (2004) Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature **430:** 569–573

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, et al (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res **12:** 1611–1618

Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, et al (2004) A Populus EST resource for plant functional genomics. Proc Natl Acad Sci USA **101:** 13951–13956

Sturgill D, Zhang Y, Parisi M, Oliver B (2007) Demasculinization of X chromosomes in the Drosophila genus. Nature **450:** 238–242

Tang C, Toomajian C, Sherman-Broyles S, Plagnol V, Guo Y, Hu TT, Clark RM, Nasrallah JB, Weigel D, Nordborg M (2007) The evolution of selfing in *Arabidopsis thaliana*. Science **317:** 1070–1072

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604

Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA **103:** 3220–3225

Wang W, Brunet FG, Nevo E, Long M (2002) Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. Proc Natl Acad Sci USA **99:** 4448–4453

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell **18:** 1791–1802

Wilkins O, Nahal H, Foong J, Provart NJ, Campbell MM (2009) Expansion and diversification of the Populus R2R3-MYB family of transcription factors. Plant Physiol **149:** 981–993

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13:** 555–556

Yin T, Difazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, et al (2008) Genome structure and emerging evidence of an incipient sex chromosome in Populus. Genome Res **18:** 422–430

Zhang J, Dean AM, Brunet F, Long M (2004) Evolving protein functional diversity in new genes of Drosophila. Proc Natl Acad Sci USA **101:** 16246–16250

Zhang Y, Li J, Kong L, Gao G, Liu QR, Wei L (2006a) NATsDB: Natural Antisense Transcripts DataBase. Nucleic Acids Res **35:** D156–D161

Zhang Y, Wu Y, Liu Y, Han B (2005) Computational identification of 69 retroposons in Arabidopsis. Plant Physiol **138:** 935–948

Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M (2006b) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics **22:** 1437–1439

Zhou F, Xu Y (2009) RepPop: a database for repetitive elements in *Populus trichocarpa*. BMC Genomics **10:** 14