

Title: Gene copy number variations in adaptive evolution: the genomic distribution of gene CNVs revealed by genetic mapping and their adaptive role in an undomesticated species, white spruce (*Picea glauca*).

Authors: Julien Prunier^{*,1,2}, Sébastien Caron^{1,2}, Manuel Lamothe^{3,4}, Sylvie Blais^{1,2,4}, Jean Bousquet^{1,2,4}, Nathalie Isabel^{3,4} & John MacKay^{2,4,5}

¹ Institute for System and Integrative Biology (IBIS), 1030 av. de la médecine, Université Laval, Québec, QC, G1V 0A6, Canada

² Centre for Forest Research, Université Laval, Québec, QC, G1V 0A6, Canada

³ Laurentian Forest Centre, Canadian Forest Service, Natural Resources Canada, 1055 rue du PEPS, PO Box 10380, Stn. Sainte-Foy, Quebec, QC, G1V 4C7, Canada

⁴ Canada Research Chair in Forest Genomics, 1030 av. de la médecine, Université Laval, Quebec, QC, G1V 0A6, Canada

⁵ Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

Keywords: Gene copy number variations, adaptive evolution, genome architecture, genetic mapping, aCGH, gymnosperms

Short title: Gene CNVs in adaptive evolution of white spruce

*Corresponding author: Julien Prunier

E-mail: jprunier.1@gmail.com

28 **Abstract:**

29 Gene copy number variation (CNV) has been associated with phenotypic variability in
30 animals and plants but a genome-wide understanding of their impacts on phenotypes is
31 largely restricted to human and agricultural systems. As such, CNVs have rarely been
32 considered in investigations of the genomic architecture of adaptation in wild species.
33 Here, we report on the genetic mapping of gene CNVs in white spruce, which lacks a
34 contiguous assembly of its large genome (~20Gb), and their relationships with adaptive
35 phenotypic variation.

36 We detected 3911 gene CNVs including *de novo* structural variations by using
37 comparative genome hybridization on arrays (aCGH) in a large progeny set. We inferred
38 the heterozygosity at CNV loci within parents by comparing haploid and diploid tissues
39 and genetically mapped 82 gene CNVs. Our analysis showed that CNVs were distributed
40 over 10 linkage groups and identified four CNV hotspots that we predict to occur in other
41 species of the Pinaceae. Significant relationships were found between 29 of the gene CNVs
42 and adaptive traits based on regression analyses with timings of bud set and bud flush, and
43 height growth, suggesting a role for CNV in climate adaptation. The importance of CNVs
44 in adaptive evolution of white spruce was also indicated by functional gene annotations
45 and the clustering of 40% of the mapped adaptive gene CNVs in CNV hotspots.

46 Taken together, these results illustrate the feasibility of studying CNVs in
47 undomesticated species and represent a major step toward a better understanding of the
48 roles of CNV in adaptive evolution.

Introduction

For decades, variations in gene copy numbers involved in phenotypic variations have been reported in model organisms, mostly in human, animals or plants of agronomical interest (Cook *et al.* 2012; Perry *et al.* 2007; Sutton *et al.* 2007). However, large-scale screenings of genomes for copy number variations (CNVs), potentially including genes, only started in 2004 in human (Freeman *et al.* 2006; Redon *et al.* 2006; Sebat *et al.* 2004). These surveys were followed by the investigation of a broader set of model species including cattle (Fadista *et al.* 2010), chicken (Wang *et al.* 2010), maize (Springer *et al.* 2009; Swanson-Wagner *et al.* 2010), Arabidopsis (DeBolt 2010), alfalfa (Munoz-Amatriain *et al.* 2013), and more recently, rice (Bai *et al.* 2016), and poplar (Pinosio *et al.* 2016).

These large-scale genome studies allowed to define a wide range of genome structural variations, from large DNA segments of several kb including genes to small indels of a few base pairs. Screening entire genomes has also identified genomic regions as hotspots of CNVs (Cloup *et al.* 2012; Lupski 2007b), which have been linked to the presence of segmental duplications, also called low-copy repeats (LCRs). These LCRs are spread throughout the genome, share a high level of identity (>95%) and promote recurrent inter- and intra-chromosomal duplications and deletions, inversion and chromosome translocation, mainly by non-allelic homologous recombination (NAHR) (Gu *et al.* 2008; Hastings *et al.* 2009; Iafrate *et al.* 2004). As a result, hotspots of CNVs have been observed in the vicinity of LCRs where duplicated segments can be found side-by-side in a variety of organisms. Genome mapping of CNVs in different model organisms have further shown collocating hotspots within the mammal or the avian lineages (Cloup *et al.* 2012). In addition, duplicated DNA segments may be dispersed over several chromosomes and result from other mechanisms but to date, their characterization has been limited to the human owing to technical challenges even in a well described genome (Conrad *et al.* 2010).

Given that gene CNVs have been associated with phenotypic variation in human and plants (Cook *et al.* 2012; Lupski 2007a; Perry *et al.* 2006), among others, they may have a role in evolution in response to natural selection in wild species. Their effects are typically associated with gene dosage, sequence interruption, fusion, or position effects as reported in human diseases (Lupski & Stankiewicz 2005). Several studies have found significant

enrichment among gene CNVs for molecular functions related to defense, biotic and abiotic stress responses in both animals and plants (Clop *et al.* 2012; Conrad *et al.* 2010; Fadista *et al.* 2010; Pinosio *et al.* 2016; Prunier *et al.* 2017). Furthermore, genome-wide investigations in a variety of organisms showed that CNVs are shared between species within the same lineage in accordance with the degree of divergence, indicating a potential role for CNVs in evolution (Clop *et al.* 2012; Dumas *et al.* 2007; Fontanesi *et al.* 2011; Prunier *et al.* 2017), which could form part of the response to natural selection.

Deciphering the genomic architecture of adaptation is a major goal in evolutionary biology. Empiric and simulation data analyses showed that efficient and heritable adaptive response to divergent selection is hampered by gene flow and genomic recombination in many biological systems (Yeaman 2013). In most cases, adaptation likely requires clustering of advantageous alleles into genomic islands-of-divergence (Nosil *et al.* 2009) or even in small clusters of tightly linked favorable alleles (Yeaman 2013). We propose that studying CNVs possibly involving several genes in non-model and undomesticated species would undoubtedly shed light on their role in evolution in response to natural selection..

Screening for CNVs at the genome level has been mainly limited to a few model organisms to date because most of the methods that have been developed for CNV detection require a high-quality contiguous reference genome assembly (Carter 2007; Carvalho *et al.* 2004). As a first step toward understanding the impacts of CNV in evolution of undomesticated conifers, we developed an approach based upon comparative genomic hybridization on arrays (aCGH) that is centered on gene sequences (Prunier *et al.* 2017). We developed a microarray chip comprised of ~60bp probes specifically targeting thousands of gene sequences and carried out comparative hybridizations of a ‘test’ genomic DNA to a ‘reference’ genomic DNA. We thus identified CNVs in 3612 genes in spruce progenies (*Picea spp.*) with a low false discovery rate (Prunier *et al.* 2017), despite spruces having very large genomes (~20 Gb) (Birol *et al.* 2013; Murray 1998; Nystedt *et al.* 2013) and lacking contiguous reference genome sequences (De La Torre *et al.* 2014). An enrichment was observed among gene CNVs for functional annotations related to defense and abiotic stress responses, suggesting an overall role in adaptation. Adaptation to climate is complex in temperate and boreal perennial plants and involves various

111 physiological features including the timing of vegetative bud burst and bud set, which are
112 crucial for avoiding the highly damaging effects of late-spring and early-fall frosts on
113 apical meristems (Bigras & Colombo 2001). Bud burst and bud set are strongly heritable
114 and have been frequently investigated in forest trees (Howe *et al.* 2003; Jaramillo-Correa
115 *et al.* 2015). Although spruce gene CNVs were previously identified in white spruce (*Picea*
116 *glauca* [Moench] Voss; Prunier *et al.* 2017), the sample size was too small for genetic
117 mapping or investigating the relationships between CNVs and adaptive variation.
118 Therefore, the objectives of this study were to 1) follow the segregation of CNVs within a
119 large full-sib progeny and genetically map them; 2) investigate whether gene CNVs have
120 an impact upon adaptation in spruce based on adaptive trait variation data, namely the
121 timings of vegetative bud set and vegetative bud burst, and annual growth (Fig.1). Genetic
122 mapping is restricted to side-by-side copy number variations because co-segregation is
123 expected only in such situations, which are nevertheless more prone to occur in CNVs
124 hotspots. Thus, we expected to uncover CNV hotspots that may exist even though the
125 genomic location of many dispersed gene CNVs would remain unknown. We also expected
126 to identify gene CNVs involved in adaptation in white spruce, which may apply to other
127 perennial boreal and temperate plant species sharing the same adaptive traits.

Material and methods

Plant material and adaptive trait phenotypes

White spruce is a diploid conifer tree of the Pinaceae. Here, 120 individuals from a single full-sib family produced by a single bi-parental mating ($\text{♀}77111 \times \text{♂}2388$) were sampled in a common garden test site (Québec province, Canada) as described (Pelgas *et al.* 2011). The individuals analysed here were previously assessed for bud phenology traits (bud burst and bud set timings) and annual height growth at the ages of 3, 4 and 5 years (Pelgas *et al.* 2011). Vegetative bud burst timings were evaluated annually in outdoor conditions by visual inspection three times a week during spring and using a rating on a scale of 0 to 6, from completely closed vegetative buds (=0) to completely flushed buds with elongating needles (=6). Similarly, vegetative bud set timings were evaluated annually by visual inspection once a week during summer and using a scale ranging from 0 to 5, from budset initiation (=0) to a complete budset (=5). As all trees were scored several many times, bud flush and bud set scores were each annually averaged over the period of measurement, lower scores indicating latter timings. Height growth was measured each year after bud set completion in falls. Whole shoot tips were collected during summer 2012 and kept at -80°C until used for DNA extractions.

Array and experimental designs

The array comprised 5629 gene sequences of 500 bp or more, which were each targeted by a minimum of 6 probes. The probes were designed by Genotypics (<http://www.genotypic.co.in>) from a genome capture assembly that includes mostly exons and introns < 1Kbp of ~23000 genes (Stival Sena *et al.* 2014). Probes were selected based on the following criteria: single hit against the target, alignment length of 60bp, less than 3 mismatches and 2 gaps were allowed (Prunier *et al.* 2017). The genes represented fell into three categories. A total of 3612 gene CNVs were from a previous study of 19 individuals of the pedigree studied here by using arrays of 4×180k probes as described (Prunier *et al.* 2017). In the present study, all of the probes for the entire set of previously discovered gene CNVs were incorporated in a new array design of 8×60k probes allowing to test a large number of individuals at an affordable cost. In addition, we included

the probes of genes that were previously tested but scored negatively when testing for CNV in our previous experiment to obtain a diploid baseline of gene copy numbers. This further set included 1612 randomly selected gene sequences and a set of 405 genes that have been well-studied in spruces such as cytochrome P450, MYB, and glycosyl-hydrolase. SurePrint G3 Custom CGH Microarrays 8×60k were printed by the Agilent sureDesign platform (Agilent Technologies).

Three types of genome DNA comparisons were performed using aCGH (Fig. 2), aside from the self-self comparisons in which the same gDNA is used as both test and reference samples so detection parameters can be adjusted to minimize the False Discovery Rate (FDR). First, a total of 120 of the full-sib progenies were individually compared to the reference parent to determine a relative gene copy number ‘genotype’ of each individual (Fig.2A). Second, each parent was compared to the megagametophyte from one of its seeds to infer the within parent heterozygosity at each gene CNV locus (Fig.2B). As the megagametophyte has a haploid genome derived from the product of meiosis in the parent, only one of the two possible alleles is expected within one individual. We used the same total amount of labelled gDNA from the megagametophyte and the parent for the hybridization, thus the megagametophyte DNA is equivalent to a pseudo-diploid and necessarily homozygous sample. When both DNA (diploid and pseudo-diploid) intensities were similar (log ratio close to 0), we inferred that the parent was homozygous regarding the copy number of the targeted DNA segment, i.e. it had the same copy number on homologous chromosomes. In contrast, detection of a CNV between the megagametophyte and the parent was indicative of a difference between the two chromosomes in the parent. Third, the two parent gDNAs were compared to each other to identify additional gene CNVs possibly resulting in phenotypic variation within the progeny (Fig.2C).

DNA extraction, labelling and hybridizations

Tissues were ground to a fine powder using a MixerMill MM 300 (Retsch, <http://www.retsch.com/>). Genomic DNA (gDNA) from each sample was extracted using the DNeasy Plant Mini Kit (QIAGEN, <http://www.qiagen.com/>) and following manufacturer’s instructions. DNA concentration was determined using a NanoDrop 1000

(Thermo Scientific, <http://www.thermoscientific.com/>) and quality was assessed by agarose gel electrophoresis (1%, TAE buffer) of 100 ng gDNA per sample. Array CGH hybridizations were performed with the SureTag DNA Labeling Kit (Agilent, protocol version 7.2), following manufacturer's instructions for the 8×60K array. Arrays were scanned using the Tecan powerscanner and images were analyzed to extract raw Cy3, Cy5 and background fluorescence intensities.

Data treatment

CGH arrays and intensities analyses

The fluorescence intensities were analyzed following the approach and the pipeline of R and Python scripts developed and described in Prunier *et al.* (2017) (available at <https://bitbucket.org/jprunier/git-CNVgene-discovery>). This approach takes into account variability among slides and dyes using a LOESS correction performed using the 'limma' R-package (Smyth 2005). In addition, the effect of GC content and probe nucleotide composition were tested and taken into account to adjust intensities ratio for each probe. An adjusted intensity ratio above or below a significant threshold for a given probe was considered indicative of a difference in copy number of the targeted DNA sequence between the individual and the parent. The array design of 8×60k probes represented an average number of 10.9 probes per gene which allowed to repeatedly test the same genomic regions with a minimum of 6 probes. Thus, the proportion of significant probes per gene was an additional parameter that was explored to robustly detect gene CNVs. We tested a series of detection parameters to analyze self-self hybridizations of reference gDNA and assess the FDR. An absolute value of 0.58 as a significant log2 ratio threshold for each probe (corresponding to a variation of three over two copies of the targeted DNA segment) and a minimum proportion of 85% significant probes per gene resulted in a FDR << 1% for gene CNVs detection. These parameters were used for CNV detection described in subsequent sections. Lower criteria for intensity ratios and significant probe rates greatly inflated the FDR (results not showed).

Genetic mapping

Our rationale for the genetic mapping analysis was that side-by-side gene copies are co-inherited in the next generation. Thus, a locus presenting a copy number variation can be used as a genetic marker and the copy number in a progeny can be estimated from the median fluorescence intensity ratio (FIR) for a given gene.

The F1 progeny were analyzed following the “two-way pseudo-testcross” mapping approach (Grattapaglia & Sederoff 1994), where markers heterozygous in only one parent and segregating with the expected 1:1 ratio are used to build parental genetic maps that are combined afterwards to provide a consensus genetic map that is deemed to be representative of the species. Here, we developed here an original approach to genetically map gene CNVs from aCGH data (Fig.1). First, we compared each (diploid) parent to the megagametophyte in order to determine for each gene whether the parent is homozygous or heterozygous in copy numbers (see Fig.2). Next, the 114 individuals that produced aCGH hybridization data suitable for CNV detection were compared to the reference parent and clustered into groups according to their gene-median FIR for each gene CNV heterozygous in only one parent. Only those that gave two clusters of progeny representing a 1:1 segregation were used for genetic mapping (Fig.1).

The most recent spruce genetic map includes 8793 gene SNP markers localized over 1895 cM by genotyping 1976 individuals from the same full-sib family that was used in the present study (Pavy *et al.* 2012). Genetic recombination rates between ~8.8K markers cannot be satisfactorily estimated using a sample size of 114 individuals. Thus, we began by integrating the CNVs into a previously published spruce genetic map of 1801 gene SNPs spread over 2083 cM that was built from the same progeny (Pavy *et al.* 2012); we did this by using the command “assign ungrouped loci to SCL-Groups” in Joinmap® 4 (Van Ooijen 2006) for the CNV loci, followed by three rounds of linkage. Parental maps were then combined to the ~8.8k markers map using the “LPmerge” R-package (Endelman & Plomion 2014) to produce a consensus genetic map including gene CNVs distributed among markers including SNPs from both previous maps and gene CNVs. The “LPmerge” function was used by estimating maps from a range of maximum interval sizes varying between 1 and 8, and weighting the importance of each input maps according to number of progenies used to produce them. For each chromosome, we retained the map with the

lowest mean and variance for the root mean-squared error (RMSE) . Density distributions over the genetic map between all markers and only CNV markers were compared using the Kolmogorov-Smirnov test using the “ks.test” R function.

Testing gene CNVs relationships with variation in adaptive traits

The set of mapped gene CNV was limited to those heterozygous in only one parent and presenting side-by-side copies; therefore, we used an approach allowing to test a broader set of loci to investigate the relationships between gene CNVs and adaptive variation. The analyses were carried out by means of regression between relative gene copy numbers and adaptive trait values.

The median intensity ratio between an individual progeny and the reference parent tree was estimated for each gene CNV that was heterozygous in parents (N=366) and gene CNVs found between parents using the same CNV detection approach, which yielded an additional set of 44 genes in CNVs (Fig.2B). The link between adaptive traits and gene CNVs (N=410) was subsequently tested using linear regression analyses among intensity ratios as predictors and adaptive phenotypes as dependant variables. All of the p-values were adjusted for multiple testing (q-values) (Storey & Tibshirani 2003) using R (release 3.2.3). Testing within a full-sib family would result in correlation among neighbouring loci because of genetic linkage; therefore, multiple regressions including all significant gene CNVs as predictors were conducted and a stepwise regression testing for the effect of each predictor on the model AIC (Akaike Information Criterion) was used to identify a subset of the most informative predictors (i.e. gene CNVs) as implemented in R using the “step” function.

The functional annotations and Gene Ontology assignments were based on the closest homologous sequence in TAIR as described in the white spruce gene catalog (Rigault *et al.* 2011). Enrichment tests were conducted using FatiGO in Babelomics 4.3 (<http://v4.babelomics.org/functional.html>) (Medina *et al.* 2010).

Results

Detecting gene CNVs within a large full-sib family

We tested 5629 genes with an average of 10.9 probes evenly distributed along the gene sequence and detected 3911 gene CNVs with a low FDR (<0.0001) by successfully comparing 114 individuals to the maternal parent. The average number of gene CNVs found within an individual was 111.5 with a few individuals ($n=10$) presenting an outlier number of 353 to 858 (Fig. 3A). As a result, 43% of the gene CNVs were detected in only one parent-offspring comparison (Fig. 3B) and the outlier individuals represented 31.5% of the detected gene CNVs.

Among the 5629 genes tested, 70% have been ascribed to 1557 different gene families according to the white spruce gene catalog (Rigault *et al.* 2011) and the remaining 30% were classified as unknowns (Supplementary table1). CNVs were highly represented in the following families: hydrolases (48 genes), disease resistance protein of the leucine-rich repeats class (30) and leucine-rich repeat kinases (22), kinases (27), TPR-like proteins (23), glycosyltransferases (19) and cytochrome P450 (18). These functional annotations are in accordance with recent literature indicating that multigene families such as NB-LRR and kinases appear to be prone to CNV in plants (Meyers *et al.* 2003; Żmieńko *et al.* 2014). No significant enrichment of gene families was observed among the gene CNVs, which was expected because these same families were the most highly represented in the probe array by design.

Genomic locations of gene CNVs

We identified 267 and 180 gene CNVs that were heterozygous in the maternal and paternal parent of the bi-parental mating, respectively; 366 gene CNVs were heterozygous in only one of the two parent. The heterozygous gene CNVs were identified by comparing each of the two diploid parents to the haploid megagametophyte of one of their progeny in an aCGH analysis. This subset of uniquely heterozygous gene CNVs could potentially be located on a genetic map using the two-way pseudo-test cross approach and was investigated for progeny clustering into two groups of fluorescence intensity ratios (FIR) representing the expected 1:1 segregation. The progeny clustering into two groups was

successful for 89 of the gene CNVs and 82 of them were successfully positioned on an updated *P. glauca* genetic map of 9,468 markers (Pavy *et al.* 2017) (Fig.4, Supplementary Table 2). Distances were recalculated after integrating the new markers and the overall map size did not change substantially (1,990 cM) from map sizes described in Pavy *et al.* (2012 and 2017). Gene CNVs were distributed over 10 of the 12 spruce linkage groups (LGs) (no gene CNVs for LG5 and LG6; Fig. 4A). There was an average of 6.92 gene CNVs per LG and a maximum of 20 gene CNVs on LG8 (Fig. 4B). The distribution of gene CNVs over the genome was significantly different than the overall marker distribution (K-S test, p-value < 0.05) and showed gene CNV hotspots on LG2, LG3 and LG8 (Fig.4A, green histogram). In addition, 7 clusters of 2 to 3 gene CNVs separated by less than 1cM were observed upon LG3, LG7, LG8 and LG9 (Fig. 4A, inner track).

We found that the mapped gene CNVs were associated with a wide variety of gene family annotations and 38% of the gene CNVs had unknown gene functions or were not assigned to known gene families (Supplementary Table 2). The most represented gene families were the large leucine-rich repeats (LRR) gene family, the cytochrome P450 and the H(+)-ATPase family, with 4, 3, and 3 gene CNVs, respectively.

Gene CNVs related to variation in adaptive traits

We tested for relationships between CNVs and three adaptive traits, i.e. the timings of vegetative bud flush and bud set, and annual height growth, which were surveyed during three consecutive years as reported by Pelgas *et al.* (2011). The phenotypic data were obtained for the 114 progeny and regressions were carried between the 9 quantitative traits (3 traits \times 3 years) and 410 gene CNVs that were heterozygous in one or the other parent, or that varied between parents. By using linear regressions, we identified 94 significant relationships involving 77 different gene CNVs after correction for multiple testing (q-value < 0.10). Based on our genetic mapping results, we expected that correlations between loci due to genetic linkage in this full-sib family was likely to inflate the number positive associations. Therefore, we use of a stepwise regression procedure and this reduced to 29 the number of distinct gene CNVs in significant relationships between adaptive traits and gene CNVs (in 31 relationships) (Table 1). A majority of the significant regressions were between gene CNVs and annual height growth at the age of 5 years (n=18). We also found

339 that one gene CNV was involved in multiple traits, i.e. bud set timing at the age of 4 years
340 and growth both at the age of 4 and 5 years. The proportion of variance explained per
341 individual CNVs (PVE) ranged from 7.4 to 13.9% (Table 1). A total 13 gene CNVs related
342 to these adaptive traits were also genetically mapped on LG2, LG3, LG7, LG8, LG9 and
343 LG10 in the present study (Fig. 4, red dots). The linear models considering all of the
344 significant gene CNVs for one trait explained 7.6 to 22.2% of the variation. Functional
345 annotations were available for 17 of genes and represented NBS-LRR, glycosyltransferase,
346 tubby-like, hydrolase or heavy metal ATPase gene families (Table 1).

Discussion

Outlier individuals and *de novo* gene CNVs within a large full-sib family

The present study revealed the heterogeneity in gene copy content that may exist between homologous chromosomes by comparing each of the parent's diploid DNA obtained from needles with the haploid DNA obtained from the megagametophyte tissue of one of its progeny. This approach allowed to identify 180 and 267 gene CNVs in the two parents that were analysed, which represents respectively 3.5 and 5% of the tested genes.

Next, we obtained and analysed data for a set of 114 diploid full-sib individuals and found an average of 111.5 gene CNVs per individual. A few outlier individuals (8.7%) were found that had many more gene CNVs (Fig. 3A), as previously observed when testing 19 individuals in black spruce and white spruce (Prunier *et al.* 2017). Another similarity between the present study and the report of Prunier *et al.* (2017) is that the gene CNVs found in outlier individuals accounted for 31.5% of 3911 gene CNVs found in the present work, and a significant portion of gene CNVs were found in only one parent-offspring comparison (Fig. 3B). The overall consistency between the two reports supports the existence of such outlier individuals and gene CNVs involving one out of 114 progeny. Since there was no replication of any progeny/parent comparison, these gene CNVs found in only one individual might result from uncontrolled technical bias. However, using an average number probes (>10) targeting the same gene allowed to repeatedly test the same gene sequence which was well above the minimal number of 3 probes recommended in the literature (Carter 2007). As a result, this offered the opportunity to delineate robust gene CNVs detection parameters resulting in a very low FDR. Another possibility is the existence of these gene CNVs in other but not sampled individuals, although the sampling size ($N > 100$) minimized the likelihood of such occurrence and they are not shared by a sufficient number of progenies to support a transmission from parent genomes. Thus, even though the possibility of false-positive cases could not be completely discarded, individual-specific gene CNVs likely encompass *de novo* variations that were formed between the two generations. Consequently, these *de novo* CNVs could not be genetically mapped since they did not conform to Mendelian expectations. They were also unsuitable to test for

possible relationships with adaptive traits since they were unlikely to have been influenced by natural selection yet, given their likely recent origin.

Such large numbers of *de novo* CNVs in outlier individuals may either come from a few structural events (duplications or deletions) of large DNA segments including several genes, or from a higher level of genome instability leading to increased duplications or deletions. The latter hypothesis might be more likely in conifer species, such as white spruce, because of their very large intergenic regions (Neale *et al.* 2014; Nystedt *et al.* 2013).

The most widely used markers for genetic mapping are single nucleotide polymorphisms (SNPs) and they are usually selected to conform to Mendelian expectations within a progeny set. As a result, SNPs embedded in CNVs are usually discarded since they are likely to result in uniform heterozygosity or low quality genotype calls (Prunier *et al.* 2017). Our data are highly consistent with this expectation since there was no overlap between the gene CNVs genetically mapped here from aCGH data and those mapped from SNP genotypes, even when considering the recently expanded genetic map (8.8K gene SNPs; Pavy *et al.* 2017). However, SNPs affected by *de novo* CNVs are less likely to be discarded as the new copy is identical to the original and likely remains undetected during the SNP-genotyping phase. As predicted, we were able to retrieve the genomic location for 34.7% of the *de novo* gene CNVs identified in outlier individuals based on SNP genotypes (Table 2). Only 19% of them grouped within 3 to 23 small mapping clusters each comprised of 2 to 3 gene CNVs depending on the outlier individual (Table 2), while the remainder appeared independently distributed across the genome. This observation is in accordance with previous reports indicating that small CNVs are more abundant than large ones (Yu *et al.* 2011). Even though these mapping clusters may include other gene CNVs considering that ~70% of the white spruce genes are still unmapped (Pavy *et al.* 2017) and likely represent large CNVs, the majority (81%) of the mapped *de novo* gene CNVs support the hypothesis of a higher genome instability in these outlier individuals.

Most mechanisms leading to CNV formation result in equal proportions of duplications and deletions but it was proposed that deletions should be slightly more frequent than duplications because one of the mechanisms responsible for the formation of CNV, namely intra-chromatid rearrangement, can only lead to copy losses (Gu *et al.* 2008). This trend

was supported by empirical data from many CNV studies that identified more losses than gains (Springer *et al.* 2009; Swanson-Wagner *et al.* 2010). However, it has also been argued that copy losses may have been overestimated with regards to copy gains owing to reference genome bias (Beló *et al.* 2010) and the stronger signal differential produced by an allele loss compared to an allele gain in a diploid species (Carter 2007). From an evolutionary standpoint, duplications are expected to be more frequent than deletions because the latter represent a high risk of loss of function and elimination by purifying selection (Brewer *et al.* 1999). Supporting this hypothesis, a population survey of large CNVs in human detected only ~1% of large deletions including genes (Conrad *et al.* 2010). Interestingly, CNVs were often identified in plants by comparing varieties, inbred lines or accessions, each represented by only one individual, but rarely by testing populations (Żmieńko *et al.* 2014). The *de novo* gene CNVs found in outlier individuals affords the opportunity to estimate the relative number of gains and losses by assessing fluorescence ratios between the progenies and the reference parent. We found that outlier individuals in white spruce had a significantly more positive than negative ratios for *de novo* CNVs (Table 2), which is indicative of a larger number of gene copy gains (*U*-test, *p*-value = 0.0089, Sup. Fig.2). Highly deleterious gene copy deletions have likely disappeared from the full-sib population analysed in the present study since only healthy individuals that germinated and survived were included. Therefore, the prevalence of duplications over deletions would support the hypothesis of purifying selection against gene copy deletions, even in favorable environmental conditions.

Identification of CNVs hotspots by genetic mapping of aCGH data

Here we identified a total of four (or five) putative CNV hotspots which contained between *x* and *y* gene CNVs. The hotspots were identified by locating 82 gene CNVs on the white spruce genetic map, which revealed clusters on linkage groups 2, 3 and 8, with the caveat that genetic mapping does not translate into physical distances (bp), (Fig 4 and Supplementary Table 2).

Mechanisms leading to the occurrence of CNVs were described in a variety of model systems and shown to be common to many if not all living organisms, from bacteria to cattle and human (Clop *et al.* 2012; Fadista *et al.* 2010; Hastings *et al.* 2009). A major

mechanism responsible for recurrent duplications and deletions is non-allelic homologous recombination (NAHR) linked to the presence of low-copy repeats (LCRs) (Cloup *et al.* 2012; Gu *et al.* 2008; Żmieńko *et al.* 2014). The LCRs do not mediate but rather stimulate duplications or deletions such that they are often found in the vicinity of genomic regions described as CNV hotspots (Gu *et al.* 2008). Thus, CNV hotspots are important sources of large genetic novelty that can be favored or discarded by natural selection (Gokcumen *et al.* 2011). Studies in great apes, domesticated mammals and poultry have shown that CNV hotspots tend to persist during the course of evolution owing to ancestral LCRs (Cloup *et al.* 2012; Perry *et al.* 2006). Detecting CNV hotspots has been a main focus in few plant studies (Żmieńko *et al.* 2014); nonetheless, they have been shown by comparing 13 inbred lines in maize (Beló *et al.* 2010) and 2 subspecies in rice (Yu *et al.* 2011). In *Arabidopsis*, comparisons between progenies suggested that NAHR during meiosis is the main mechanism responsible for CNVs, which should result in hotspot formation (Lu *et al.* 2012). In contrast, the barley genome harbors the signature of double-strand breaks repaired by single-stranded annealing (Munoz-Amatriain *et al.* 2013).

Genome mapping of CNVs to delineate hotspots is a difficult endeavour in non-model species because a high quality contiguous genome assembly is rarely available. Notwithstanding improvements in sequencing technology, genome assemblies developed for the vast majority of non-model species including white spruce (Warren *et al.* 2015) are of draft quality, which does not enable CNV mapping by direct alignment. Consequently, we used genetic mapping based on aCGH data as an alternative approach and thus expand structural investigation of CNVs to wild species.

Gene CNVs with side-by-side copies are likely to be inherited together in the next generation and can be analyzed as markers to locate the CNVs. White spruce is a highly heterozygous outcrossing species where genetic mapping is most efficient by means of the two-way pseudo test cross approach using markers segregating in a 1:1 ratio. This approach imposes limitations that reduced the proportion of gene CNVs that could be mapped. Here, clustering gene-median FIR in only two groups was successfully attempted for 82 gene CNVs with 114 progeny; however, attempting to cluster in more groups (i.e. 1:2:1 or 1:1:1:1 segregations) would definitely require the use of a larger progeny set ($n > 150$). Nevertheless, the approach was successful and it could be applied to other undomesticated

species where control crosses are feasible and this includes autogamous plants where a higher rate of genetically mapped gene CNVs could be expected. Starting from an exome-capture assembly, we develop here a robust approach that first use a comparison between haploid and diploid tissues from the parents to ensure that one parent is heterozygous while the other is homozygous for one variable site. However, probes targeting genes could likely be designed from an RNA-seq assembly and systematically selecting gene CNVs with a two-mode distribution for gene-median FIR within an F1, F2 or back-cross progeny should be feasible for any diploid species. Such experiments would reveal genomic distributions of the gene CNVs harbouring side-by-side copies and contribute to expanding our knowledge of CNVs to a wider set of species.

Gene CNVs included in hotspots are most likely to originate from NAHR. It is also likely that those CNVs that locate outside hotspots are formed as a result of other genomic mechanisms (Hastings *et al.* 2009). The CNVs hotspots found in white spruce encompassed 39.7% of the mapped gene CNVs, therefore supporting the hypothesis of NAHR is a major mechanism responsible for recurrent CNVs in plants. Furthermore, 48% (N=16) of the gene CNVs located in hotspots were also detected in both black spruce and white spruce in a previous study of gene CNVs in the spruce lineage (Prunier *et al.* 2017) while only 23% (N=11) of the gene CNVs outside hotspots were also detected in black spruce. These observations together with findings from studies of other plant lineages support and underscore the importance of NAHR in CNV formation. Furthermore, we can also expect to find these CNV hotspots in other conifers, particularly in the Pinaceae family including members of the genus *Pinus* and other species of the genus *Picea* given the persistence of CNV hotspots within lineages.

CNVs in adaptive evolution

In the present study, a conservative set of 29 out of a total of 410 gene CNVs was identified as likely involved in adaptive trait variation. This proportion is consistent with findings from other studies on adaptive genetic variations – mostly based on SNPs analyses – from a set of candidate polymorphisms delineated from gene annotations (Hall *et al.* 2016).

Copy number variants that are transmitted through the germline to the next generation may impact phenotypes and be favored or eliminated by natural selection. As a result, some CNVs involving genes become fixed in a lineage-specific manner (Dumas *et al.* 2007) while others remain variable in populations in accordance with divergent selection (Perry *et al.* 2007). CNVs involving several genes at a time are more likely to substantially impact the phenotype. For example, copy number variations in three functional genes found in tandem in the *Glycine max* genome have been associated with nematode resistance (Cook *et al.* 2012). Similarly, coat color in horses, pigs and sheeps is linked to CNVs involving two to three genes (Cloup *et al.* 2012). In these examples, additional functional copies were associated with up-regulation of gene expression, which translated into substantial phenotypic variation. Similarly, studies of gene expression regulation in model organisms showed correlated expression profiles in response to stress between neighbouring genes (Williams & Bowles 2004), which can be easily explained by identical copies of many functional genes.

In the present study, we identified genomic clusters of gene CNVs less than 1cM apart, by using direct genetic mapping of heritable gene CNVs and indirect genetic mapping of *de novo* CNVs. These clusters included two to three genes that were either duplicated or deleted altogether compared to the reference parent. These clusters included 18% of the heritable gene CNVs and 19% of the *de novo* gene CNVs, which is in the same range as rates observed by comparing maize inbreds and teosinte genotypes (15-31% of gene CNVs found in clusters of mostly 2 or 3 genes, Swanson-Wagner *et al.* 2010) or 13 maize inbred lines (16% of gene CNVs within similar clusters, Beló *et al.* 2010).

We directly mapped inherited gene CNVs and found significant relationships between 77 of them and adaptive trait variation by means of linear regressions within the full-sib family. A stepwise regression procedure reduced the number CNVs to 29, which may represent more convincing cases of adaptive gene CNVs, but this methodology would typically select only one of the genes representative of an entire cluster. This point is illustrated by the observation of a cluster of two gene CNVs located within a narrow window of 1cM that have the same significant effect on growth at age 5 (similar PVE of 8.9%) and show a highly positive correlation in terms of relative copy numbers (Fig.5D). Other similar cases of clustered gene CNVs may potentially be found among the unmapped

532 adaptive gene CNVs. On the other hand, markers involved in quantitative trait variation
533 are embedded in a region including several other markers also showing significant
534 relationships because of the genetic linkage between neighbouring loci in a full-sib
535 progeny. Although the genetic disequilibrium between neighbouring markers in natural
536 populations extends less far than in related individuals, the same difficulty may arise since
537 adaptive alleles tend to hitchhike surrounding neutral alleles (Nosil *et al.* 2009), hence
538 possibly creating clusters of alleles (CNVs in the present case) involved in adaptation.
539 Therefore, testing for the effect of gene CNVs clustering on adaptive trait variation would
540 require transgenic experiments in which the copy number is artificially increasing for one
541 gene at a time and measuring phenotypic effects.

542 The set of 410 gene CNVs tested here were selected for their copy number variability
543 in the progeny under investigation and not on prior knowledge of their involvement in
544 adaptation. Climate adaptation in trees is critical for forest sustainability in the face of
545 climate change and involves a wide range of physiological processes (Bigras & Colombo
546 2001), underpinned by several metabolic pathways and the perception of environmental
547 signals. As a result, most studies have discovered large sets of genetic markers linked to
548 climate adaptation in forest trees and found that each marker had a small effect on adaptive
549 variation (Howe *et al.* 2003; Jaramillo-Correa *et al.* 2015). Accordingly, the portion of the
550 explained variation (PVE) is often low, usually in the range of 1 – 5% in association studies
551 and 6 – 15% in QTL mapping (Hall *et al.* 2016). We found PVE estimates for gene CNVs
552 in the progeny (7 – 14%) in the range of allelic effects measured in QTL mapping studies
553 based on SNP markers. The effects of gene CNVs on the adaptive variation may be
554 complex and not straightforward to model; however, we found several linear relationships
555 between gene copy numbers and adaptive traits, which is strongly suggestive of dosage
556 effects (Fig. 5 A, B, and C). The detection of dosage effects are expected when the
557 physiological pathway leading from genetic to phenotype is up- or down-regulated because
558 of variations in gene expression related to functional gene copies.

559 Several of gene CNVs that we identified had functional annotations in accordance with
560 a putative role in adaptation. For instance, several NBS-LRR proteins have a demonstrated
561 role in disease resistance in plants (Belkhadir *et al.* 2004). Similarly, several members of
562 the glycosyltransferase gene families are involved in stress and defence responses in

Arabidopsis (Langlois-Meurinne *et al.* 2005). One of the CNVs involved a gene (*GQ04107_K11*; Table1) of the Tubby-like protein family and was correlated with both bud set timing and height growth in our study. In *Arabidopsis*, members of the Tubby-like protein family act in abscisic acid signaling and their over-expression alters seed dormancy and germination (Lai *et al.* 2004). A similar role in spruce may influence the phenology of vegetative dormancy and annual growth.

Our findings indicate that gene CNVs have an effect in genetic adaptation that may be as important as that reported for SNPs. Furthermore, many adaptive gene CNVs were not located on the genetic map but four (13.7%) adaptive gene CNVs were located in CNV hotspots, further highlighting the potential importance of hotspots of CNVs in adaptive evolution.

Conclusion

Here, we genetically mapped gene CNVs and assessed putative relationships between CNVs and adaptation in an undomesticated species. This investigation allowed to reveal the wide distribution of gene CNVs over the genome and to identify a few of CNV hotspots. The prevalence of duplications over deletions among *de novo* CNVs support the occurrence of purifying selection against gene copy deletions. Our findings indicate that gene CNVs contribute to genetic adaptation in white spruce and their function may be explained by gene dosage effects. Furthermore, the potential importance of hotspots in the genomic architecture of adaptation in spruce is supported by the fact that 40% of the mapped adaptive genes are within CNV hotspots. .

Acknowledgement

This work was financially supported by Genome Canada and Genome Quebec for the SmarTForests project (JM, JB) and by the Quebec Ministry the Economy, Science and Innovation (JM, JB). The authors thank Marie-Claude Gros-Louis (Canadian Forest Service, Natural Resources Canada) for sampling, trait measurements and field work, Jérôme Laroche (Univ. Laval) for informatics support, Atef Sahli, Isabelle Giguère, Geneviève Parent, Sébastien Gérardi and Ilga Porth (Univ. Laval) for helpful discussion

and insights. We also thank two anonymous reviewers and the subject editor for helpful comments on the first version of the manuscript.

Data accessibility

The raw datasets were compressed in one file available in the GEO repository, at the accession no. GSE92329. All R and Python scripts and codes as well as a demonstration input file can be found at <https://bitbucket.org/jprunier/git-CNVgene-discovery/>.

Figure captions:

Fig.1: Approach developed for genetic mapping of gene CNVs in a highly heterozygous wild species using the two-way pseudo-test cross approach within a large full-sib progeny, and testing relationships between CNVs and variation in adaptive traits by linear regression analyses.

Fig.2: The three types of comparisons performed during the experiment. Red and green colors symbolize the labeling in either Cy5 or Cy3. A) Parent ...

Fig.3: Detection of gene CNVs within a large full-sib family. A) Number of gene CNVs found comparing each progeny gDNA to its maternal gDNA. B) Density distribution of gene CNVs according to the number of progeny in which they have been detected.

Fig.4: Genome distribution of gene CNVs over the spruce genetic map. A) Genetic map of gene CNVs and gene SNPs. From the outward to the inward tracks: the genetic map (SNPs in blue, CNVs in orange, LG: linkage groups, graduations in cM), histogram density for SNPs (in grey), histogram density for CNVs (in green), adaptive gene CNVs (red dots), clusters of CNVs (less than 1cM apart). B) Distribution of gene CNVs density per linkage group.

623 Fig.5: Relationships between adaptive traits and CNV. A, B and C) Linear model
624 relationships between variations in adaptive traits and relative copy numbers for gene
625 CNVs (expressed by the median of fluorescence intensities ratios). D) Strong copy number
626 correlation between two gene CNVs similarly associated to growth and clustered together
627 upon linkage group 3.

References

- Bai Z, Chen J, Liao Y, *et al.* (2016) The impact and origin of copy number variations in the *Oryza* species. *BMC Genomics* **17**, 261.
- Belkhadir Y, Subramaniam R, Dangl JL (2004) Plant disease resistance protein signaling: NBS–LRR proteins and their partners. *Current opinion in plant biology* **7**, 391-399.
- Beló A, Beatty MK, Hondred D, *et al.* (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics* **120**, 355.
- Bigras F, Colombo SJ (2001) *Conifer Cold Hardiness* Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Birol I, Raymond A, Jackman SD, *et al.* (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, btt178.
- Brewer C, Holloway S, Zawalnyski P, Schinzel A, FitzPatrick D (1999) A chromosomal duplication map of malformations: regions of suspected haplo-and triplolethality—and tolerance of segmental aneuploidy—in humans. *The American Journal of Human Genetics* **64**, 1702-1708.
- Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39**, S16-21.
- Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol* **57**, 644-646.
- Clop A, Vidal O, Amills M (2012) Copy number variation in the genomes of domestic animals. *Anim Genet* **43**, 503-517.
- Conrad DF, Pinto D, Redon R, *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712.
- Cook DE, Lee TG, Guo X, *et al.* (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* **338**, 1206-1209.
- De La Torre A, Birol I, Bousquet J, *et al.* (2014) Insights into conifer giga-genomes. *Plant Physiol*, pp. 114.248708.
- DeBolt S (2010) Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* **2**, 441-453.
- Dumas L, Kim YH, Karimpour-Fard A, *et al.* (2007) Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* **17**, 1266-1277.
- Endelman JB, Plomion C (2014) LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* **30**, 1623-1624.
- Fadista J, Thomsen B, Holm L-E, Bendixen C (2010) Copy number variation in the bovine genome. *BMC Genomics* **11**, 284.
- Fontanesi L, Beretti F, Martelli P, *et al.* (2011) A first comparative map of copy number variations in the sheep genome. *Genomics* **97**, 158-165.
- Freeman JL, Perry GH, Feuk L, *et al.* (2006) Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949-961.
- Gokcumen O, Babb PL, Iskow RC, *et al.* (2011) Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol* **12**, R52.
- Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* **137**, 1121-1137.
- Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4.

675 Hall D, Hallingbäck HR, Wu HX (2016) Estimation of number and size of QTL effects in forest tree
 676 traits. *Tree Genetics & Genomes* **12**, 110.
 677 Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number.
 678 *Nat Rev Genet* **10**, 551-564.
 679 Howe GT, Aitken SN, Neale DB, *et al.* (2003) From genotype to phenotype: unraveling the
 680 complexities of cold adaptation in forest trees. *Canadian Journal of Botany* **81**, 1247-1266.
 681 lafrate AJ, Feuk L, Rivera MN, *et al.* (2004) Detection of large-scale variation in the human genome.
 682 *Nat Genet* **36**, 949-951.
 683 Jaramillo-Correa JP, Prunier J, Vázquez-Lobo A, Keller SR, Moreno-Letelier A (2015) Chapter Eight-
 684 Molecular Signatures of Adaptation and Selection in Forest Trees. *Advances in Botanical*
 685 *Research* **74**, 265-306.
 686 Langlois-Meurinne M, Gachon CM, Saindrenan P (2005) Pathogen-responsive expression of
 687 glycosyltransferase genes UGT73B3 and UGT73B5 is necessary for resistance to
 688 *Pseudomonas syringae* pv tomato in Arabidopsis. *Plant Physiol* **139**, 1890-1901.
 689 Lu P, Han X, Qi J, *et al.* (2012) Analysis of Arabidopsis genome-wide variations before and after
 690 meiosis and meiotic recombination by resequencing Landsberg erecta and all four
 691 products of a single meiosis. *Genome Res* **22**, 508-518.
 692 Lupski JR (2007a) Genomic rearrangements and sporadic disease. *Nat Genet* **39**, S43-47.
 693 Lupski JR (2007b) Structural variation in the human genome. *N Engl J Med* **356**, 1169-1171.
 694 Lupski JR, Stankiewicz P (2005) Genomic disorders: molecular mechanisms for rearrangements
 695 and conveyed phenotypes. *PLoS Genet* **1**, e49.
 696 Medina I, Carbonell J, Pulido L, *et al.* (2010) Babelomics: an integrative platform for the analysis
 697 of transcriptomics, proteomics and genomic data with advanced functional profiling.
 698 *Nucleic Acids Res* **38**, W210-W213.
 699 Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-
 700 LRR-encoding genes in Arabidopsis. *The Plant Cell* **15**, 809-834.
 701 Munoz-Amatriain M, Eichten SR, Wicker T, *et al.* (2013) Distribution, functional impact, and origin
 702 mechanisms of copy number variation in the barley genome. *Genome Biol* **14**, R58.
 703 Murray BG (1998) Nuclear DNA amounts in gymnosperms. *Ann Bot* **82**, 3-15.
 704 Neale DB, Wegrzyn JL, Stevens KA, *et al.* (2014) Decoding the massive genome of loblolly pine
 705 using haploid DNA and novel assembly strategies. *Genome Biol* **15**, 1.
 706 Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic
 707 divergence. *Molecular ecology* **18**, 375-402.
 708 Nystedt B, Street NR, Wetterbom A, *et al.* (2013) The Norway spruce genome sequence and
 709 conifer genome evolution. *Nature* **497**, 579-584.
 710 Pavy N, Lamothe M, Pelgas B, *et al.* (2017) A high-resolution reference genetic map positioning
 711 8.8 K genes for the conifer white spruce: structural genomics implications and
 712 correspondence with physical distance. *The Plant Journal* **90**, 189-203.
 713 Pavy N, Pelgas B, Laroche J, *et al.* (2012) A spruce gene map infers ancient plant genome
 714 reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant
 715 conifers. *Bmc Biology* **10**.
 716 Pelgas B, Bousquet J, Meirmans PG, Ritland K, Isabel N (2011) QTL mapping in white spruce: gene
 717 maps and genomic regions underlying adaptive traits across pedigrees, years and
 718 environments. *BMC Genomics* **12**, 145.
 719 Perry GH, Dominy NJ, Claw KG, *et al.* (2007) Diet and the evolution of human amylase gene copy
 720 number variation. *Nat Genet* **39**, 1256-1260.
 721 Perry GH, Tchinda J, McGrath SD, *et al.* (2006) Hotspots for copy number variation in chimpanzees
 722 and humans. *Proceedings of the National Academy of Sciences* **103**, 8006-8011.

- Pinosio S, Giacomello S, Faivre-Rampant P, *et al.* (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology and Evolution*, msw161.
- Prunier J, Caron S, MacKay J (2017) CNVs into the wild: screening the genomes of conifer trees (*Picea* spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. *BMC Genomics* **18**, 97.
- Redon R, Ishikawa S, Fitch KR, *et al.* (2006) Global variation in copy number in the human genome. *Nature* **444**, 444-454.
- Rigault P, Boyle B, Lepage P, *et al.* (2011) A white spruce gene catalog for conifer genome analyses. *Plant Physiol* **157**, 14-28.
- Sebat J, Lakshmi B, Troge J, *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-528.
- Smyth GK (2005) Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor. New York: Springer; 2005. P397 – 420.
- Springer NM, Ying K, Fu Y, *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* **5**, e1000734.
- Stival Sena J, Giguere I, Boyle B, Rigault P, Birol I, Zuccolo A, Ritland K, ritland C, Bohlmann J, Jones S, *et al.* Evolution of gene structure in the conifer *Picea glauca*: a omparative analysis of the impact of intron size. *BMC Plant Biol.* 2014; 14:95
- Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445.
- Sutton T, Baumann U, Hayes J, *et al.* (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* **318**, 1446-1449.
- Swanson-Wagner RA, Eichten SR, Kumari S, *et al.* (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res* **20**, 1689-1699.
- Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N (2010) An initial map of chromosomal segmental copy number variations in the chicken. *BMC Genomics* **11**, 351.
- Warren RL, Keeling CI, Yuen MMS, *et al.* (2015) Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *The Plant Journal* **83**, 189-212.
- Williams EJ, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res* **14**, 1060-1067.
- Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci U S A* **110**, E1743-1751.
- Yu P, Wang C, Xu Q, *et al.* (2011) Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics* **12**, 372.
- Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant genomes. *Theoretical and Applied Genetics* **127**, 1-18.







