UNIVERSITÉ LAVAL

Copy number variations in the gene space of *Picea glauca*

Inheritance and spontaneous mutations

Thèse

Atef Sahli

Doctorat en sciences forestières
Philosophiae doctor (Ph.D.)

Québec, Canada

Copy number variations in the gene space of *Picea glauca*

Inheritance and spontaneous mutations

Thèse

Atef Sahli

Sous la direction de :

Jean Bousquet, directeur de recherche
John MacKay, codirecteur de recherche

RÉSUMÉ

Les variations de nombre de copies (VNCs) sont des variations génétiques de grande taille qui ont été détectées parmi les individus de tous les organismes multicellulaires examinés à ce jour. Ces variations ont un impact considérable sur la structure et la fonction des gènes et ont été impliquées dans le contrôle de différents traits phénotypiques. Chez les plantes, les caractéristiques génétiques des VNCs sont encore peu caractérisées et les connaissances concernant les VNCs sont encore plus limitées chez les espèces arborescentes. Les objectifs principaux de cette thèse consistaient i) au développement d'une approche pour la détection de VNCs dans l'espace génique de conifères arborescents appartenant à l'espèce *P. glauca*, ii) à l'estimation du taux de mutation des VNCs à l'échelle du génome et iii) à l'examen des profils de transmission des VNCs d'une génération à la suivante. Nous avons utilisé des données brutes de génotypage par puces de SNPs qui ont été générées pour 3663 individus appartenant à 55 familles biparentales, et avons examiné plus de 14 000 gènes pour identifier des VNCs. Nos résultats montrent que les VNCs affectent une petite proportion de l'espace génique. Les polymorphismes de nombre de copies observés chez les descendants étaient soit hérités soit générés par des mutations spontanées. Notre analyse montre aussi que les estimés du taux de mutation couvrent au moins trois ordres de grandeur, pouvant atteindre de hauts niveaux et variant pour différents gènes, allèles et classes de VNCs. Le taux de mutation du nombre de copies était aussi corrélé au niveau d'expression des gènes et la relation entre le taux de mutation et l'expression des gènes était mieux expliquée dans le cadre de l'hypothèse de barrière par la dérive génétique. Concernant l'hérédité des VNCs, nos résultats montrent que la plupart de ces derniers (70%) sont transmises en violation des lois mendéliennes de l'hérédité. La majorité des distorsions de transmission favorisaient la transmission d'une copie et contribuaient à la restauration rapide du génotype à deux-copies dans la génération suivante. Les niveaux de distorsion observés variaient considérablement et étaient influencés par des effets parentaux et des effets liés au contexte génétique. Nous avons aussi identifié des situations où la perte d'une copie de gène était favorisée et soumise à différentes formes de pressions sélectives. Cette étude montre que les mutations *de novo* et les distorsions de transmission de VNCs influencent la diversité génétique présente chez une espèce et jouent un rôle important dans l'adaptation et l'évolution.

ABSTRACT

Copy number variations (CNVs) are large genetic variations detected among the individuals of every multicellular organism examined so far. These variations have a considerable impact on gene structure and function and have been shown to be involved in the control of several phenotypic traits. In plants, the key genetic features of CNVs are still poorly understood and even less is known about CNVs in trees. The goals of this thesis were to i) develop an approach for the identification of CNVs in the gene space of the conifer tree *Picea glauca*, ii) estimate the rate of CNV generation genome-wide and iii) examine the transmission patterns of CNVs from one generation to the next. We used SNP-array raw intensity genotyping data for 3663 individuals belonging to 55 full-sib families to scan more than 14 000 genes for CNVs. Our findings show that CNVs affect a small proportion of the gene space and copy number variants detected in the progeny were either inherited or generated through *de novo* events. Our analyses show that copy number (CN) mutation rate estimates spanned at least three orders of magnitude, could reach high levels and varied for different genes, alleles and CNV classes. CN mutation rate was also correlated with gene expression levels and the relationship between mutation rate and gene expression was best explained within the frame of the drift-barrier hypothesis (DBH). With regard to CNV inheritance, our results show that most CNVs (70%) are transmitted from the parents in violation of Mendelian expectations. The majority of transmission distortions favored the one-copy allele and contributed to the rapid restoration of the two-copy genotype in the next generation. The observed distortion levels varied considerably and were influenced by parental, partner genotype and genetic background effects. We also identified instances where the loss of a gene copy was favored and subject to different types of selection pressures. This study shows that *de novo* mutations and transmission distortions of CNVs contribute both to the shaping of the standing genetic variation and play an important role in species adaptation and evolution.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

μ: mutation rate
μ_{AS}: allele-specific mutation rate
μ_{cg}: cross-genome mutation rate
μ_{ls}: locus-specific mutation rate
u_g: genomic mutation rate
A0: zero-copy allele
A1: one-copy allele
aCGH: array-comparative genomic hybridization
ASCN: allele-specific copy number
BIR: break-induced replication
bp: base pair
CN: copy number
CNG: copy number gain
cnTRD: copy number transmission ratio distortion
CNV: copy number variation
DBH: drift-barrier hypothesis
DNA : deoxyribonucleic acid
DNM: *de novo* mutation
DSB: double strand breakage
FDR: false discovery rate
FoSTeS: fork stalling and template switching
Gbp: giga base pair
gd: genetic distance
GO: gene ontology
HeD: heterozygous deletion
HoD: homozygous deletion
indels: insertions-deletions
LCR: low copy repeats
LOF: loss of function
MA: mutation accumulation line
Mbp: mega base pair
ME: mobile element

NAHR: nonallelic homologous recombination

NGS: next-generation sequencing

NHEJ: non-homologous end joining

P(FS|FS): proportion of full-sib dyads correctly inferred

P(HS|HS): proportion of half-sib dyads correctly inferred

P(PO|PO): proportion of parent-offspring dyads correctly inferred

P(UR|UR): proportion of unrelated individuals dyads correctly inferred

PAV: presence-absence variation

qPCR: quantitative real-time PCR

RBM: replication-based mechanism

sdTRD: sex-dependent transmission ratio distortion

siTRD: sex-independent transmission ratio distortion

SNP: single nucleotide polymorphism

SNV: single nucleotide variation

SRS: serial replication slippage

SSR: short sequence repeats

SV: structural variation

TAMH: transcription-associated mutagenesis hypothesis

TCRH: transcription-coupled repair hypothesis

TD: transmission distortion

TR(A0): transmission ration for the zero-copy allele A0

TRD: transmission ratio distortion

VNC: variation de nombre de copies

ACKNOWLEDGMENTS

"*Science is a way of life. Science is a perspective. Science is the process that takes us from confusion to understanding in a manner that is precise, predictive and reliable — a transformation, for those lucky enough to experience it, that is empowering and emotional.*" – Brian Greene.

As I look back on the journey that brought me here with all its rewards and challenges, and contemplate the path laying ahead full of excitement and expectations, I could but feel tremendously grateful and lucky to have chosen to pursue a career in science. The past few years were an intense learning period at the academic and personal level. As I progressed in my studies, it was mostly an enjoyable experience and I realize that but for the support and friendship of the people I have come to know, it may have been different. I would like to thank all my colleagues and friends who helped me bring this thesis to completion.

First and foremost, I would like to express my sincere gratitude to my supervisors Prof. John MacKay and Prof. Jean Bousquet for the opportunity to work on this PhD project and for their support and generosity with their time and expertise.

I am grateful to Isabelle Giguère and Sébastien Caron for their help with the wet-Lab experiments.

My thanks go to France Gagnon, Sylvie Blais and Patricia Lavigne for their assistance with the handling of the raw genotyping data.

I would like to extend my thanks to the members of the jury who accepted to evaluate this thesis.

Also, I would like to thank my colleagues in John Mackay's teams at Laval University and the University of Oxford for their support, constructive criticism and comments; particularly, Julien Prunier and Geneviève Parent with whom I had the most interesting and stimulating discussions.

A special thanks to Mebarek Lamara for the support and friendship he extended to me from the day I first set foot in Quebec City.

FORWARD

This thesis is based on two articles; one peer-reviewed paper accepted for publication in an international journal on April 13th 2017 (Chapter 2) and one manuscript to be submitted for review also in an international journal (Chapter 3).

Part of Chapter 1 is also planned to be published in a review article commissioned by the international journal Tree Genetics and Genomes.

The Chapter 2 is based on the peer-reviewed article accepted for publication:

Atef Sahli, Isabelle Giguère, Jean Bousquet and John MacKay (2017). High and variable copy number mutation rates in the gene space of *Picea glauca*. G3: Genes | Genomes | Genetics; manuscript accepted.

The Chapter 3 is based on a manuscript to be submitted for publication as:

Atef Sahli, Jean Bousquet and John MacKay (2017). Transmission distortions of genic copy number variants cause significant and complex frequency changes between generations. Heredity; manuscript.

In both studies, A. Sahli designed the experiments, conducted part of the manipulations in wet-Lab, performed the bioinformatics and statistical analyses, interpreted the results and drafted the manuscripts with the supervision of J. MacKay and J. Bousquet, who also provided the funding.

Chapter 1: General introduction

1.1. Preamble

Genetic diversity determines the ability of species to adapt to environmental changes, compete with other species and colonize new ecological niches. The genetic diversity that is present in a population is the result of a balance between the generation of new variants by mutation and the maintenance or purging of alleles influenced by different evolutionary forces (selection, genetic drift, migration, etc.). Genetic diversity within species manifests itself in the form of variations i) among individuals or populations and ii) across the genome of a single individual. The genetic polymorphism harbored by an individual species is determined by three main factors: mutation rate, effective population size and linked selection (Figure 1.1). These factors are in turn influenced by other elements including i) the species life history, mating system and demographic history and; ii) the genome distribution of recombination rates and gene density (Ellegren and Galtier, 2016).

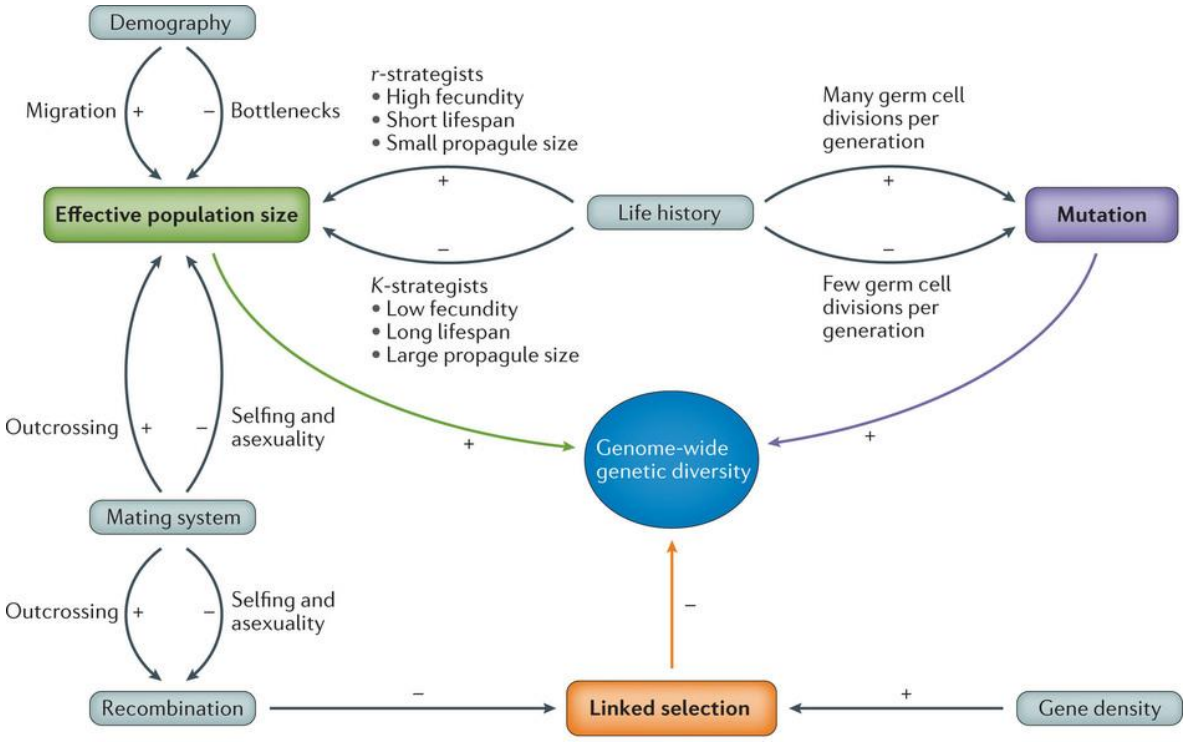


Figure 1.1: Overview of determinants of genetic diversity. Effective population size, mutation rate and linked selection are the main factors affecting diversity. These factors are in turn governed by several other parameters. The direction of correlation is indicated by the + and – symbols [Figure reproduced by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Ellegren and Galtier (2016), copyright 2016].

This thesis reports on the identification of genic copy number variations (CNVs) in the genome of white spruce (*Picea glauca* [Moench] Voss) and the characterization of their generation rates and transmission profiles. CNVs are large genetic variations believed to play an important role in adaptation and evolution but they are still under studied in forest trees. *P. glauca* is an ecologically and economically important species in Canada. It is ubiquitous throughout the boreal forest in the region (along with *Picea mariana*) and is one of the most planted species for wood and pulp production (Franceschini and Schneider, 2016). Like many conifers, its life history and genome architecture characteristics include monoecious reproduction and outcrossing mating with a delayed maturity and long generation time, long life span, large reproductive output, large population size, little inbreeding and strong inbreeding depression, high gene flow and weak natural population structure (Burns and Honkala, 1990; Jaramillo-Correa *et al.*, 2001; Bouillé and Bousquet, 2005; O'Connell *et al.*, 2006; Namroud *et al.*, 2008). The diploid genome of *P. glauca* is very large (20 Giga-bp) and rich in repeated sequences (mainly mobile elements) and; has a low G-C content (38%) and a decay of linkage disequilibrium (LD) often within gene limits (Pavy *et al.*, 2012; Birol *et al.*, 2013; Nystedt *et al.*, 2013). The gene space of white spruce encompass only 1% of the size of the genome and includes between 37491 and 56064 genes (De La Torre *et al.*, 2014; Warren *et al.*, 2015). Recent advances in forest tree genomics, including *P. glauca*, are reviewed in Parent and collaborators (2015) and Ingvarsson and collaborators (2016).

This chapter introduces CNVs as genetic variations and discuss their contribution to adaptation and evolution. Figure 1.3 summarizes the factors influencing the genetic diversity of CNVs. The three main component related to CNV generation, transmission and fate are discussed in detail. Finally, the context, objectives and hypotheses underlying the research work of this thesis are presented at the end of the chapter.

1.2. Copy number variations

Genetic variations (Figure 1.2) alter functional (coding and regulatory sequences) and non-functional (intronic, intergenic, telomeric and sub-telomeric sequences) DNA sequences and include single nucleotide variations (SNVs), small insertions and deletions (indels: 1 to 100 bp), short sequence repeats (SSRs), mobile elements (MEs) and structural variations (SVs: 50 bp to several Mbp) (Scherer *et al.*, 2007; Carvalho and Lupski, 2016). SVs are classified into two categories i) balanced SVs (with no change in copy number) such as

inversions and translocations and ii) unbalanced SVs (with a change in copy number) also called copy number variations (CNVs) which are the result of deletion, duplication or insertion events (reviewed in Alkan *et al.*, 2011).

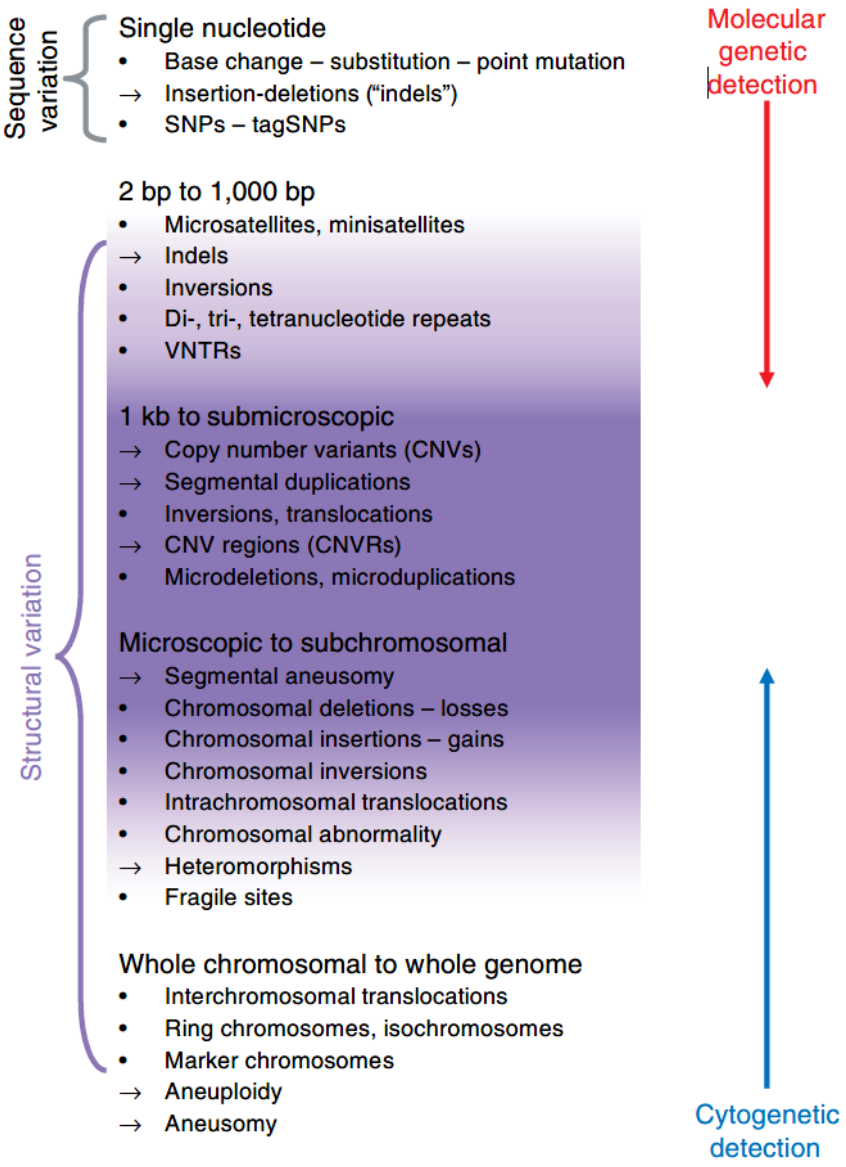


Figure 1.2: Lexicon of genomic variation. Descriptors of variation began in the realm of cytogenetics, followed by those from the field of molecular genetics. The designation of the category '1 kb to submicroscopic' is somewhat arbitrary at both ends, but is used for operational definition. In a broad sense, structural variation has been used to refer to genomic segments both smaller and larger than the narrower operational definition, as illustrated by the large bracket. The focus of recent discoveries has been the subgroup in the midrange (indicated with strong highlighting), but the gradation of shading illustrates that the biological boundaries may really encompass some forms of variation previously recognized from either cytogenetic or molecular genetic approaches [Figure reproduced by permission from Macmillan Publishers Ltd: Nature Genetics, Scherer *et al.*, 2007, copyright 2007].

1.2.1. CNV detection methods

The recent advances of genotyping and sequencing technologies allowed the detection of CNVs genome-wide for many multicellular organisms. The three main technologies used for the identification of CNVs are single nucleotide polymorphism arrays (SNP-arrays), array-comparative genomic hybridization (aCGH) and next generation sequencing (NGS) (reviewed in Alkan *et al.*, 2011). In this work, we developed a reliable approach for CNV identification from SNP-array genotyping data for thousands of white spruce trees. The aCGH technology was also recently used for the detection of genic CNVs in a moderate number of spruce trees (Prunier *et al.*, 2017). The use of NGS technologies for SV identification in trees is still limited to a few species and the studies reported so far involved the analysis of only few individuals in each case (Neves *et al.*, 2013; Warren *et al.*, 2015; Pinosio *et al.*, 2016). For conifer trees, the use of hybridization technologies (SNP-arrays and aCGH) for CNV characterization is more affordable because of their large and complex genomes, but depends on prior knowledge of the genome for probe design, requires a more extensive validation of the detected variants and provides less information than sequencing technologies. Approaches that are based on NGS data analyses, are still more expensive but offer the advantage of providing more reliable and extensive data for the study of SVs. NGS technologies can be used to i) characterize all the categories of SVs, ii) detect variants in coding and non-coding regions reliably, iii) identify the molecular mechanisms involved in SV formation and iv) simultaneously consider sequence and structural variants.

1.2.2. CNV reported in other species including plants

CNVs are among the least studied genetic variations but were widely detected in many model and non-model organisms. CNVs are commonly identified in healthy and sick (cancer and other disorders) individuals (Feng *et al.*, 2010; Mills *et al.*, 2011; Blackburn *et al.*, 2013; Gilissen *et al.*, 2014) and in response to stress conditions (Debolt, 2010). CNVs were reported for many model and crop plants including *Arabidopsis*, barley, wheat, rice, maize, sorghum, soybean and potato (reviewed in Saxena *et al.*, 2014; Zmienko *et al.*, 2014). However, little effort was invested in the characterization of their frequency, generation rate, formation mechanisms and functional impacts. In trees, CNV analyses are still scarce particularly for non domesticated species such as conifers. Aside from the present study, CNVs were described for three spruce species including white spruce

(Prunier *et al.*, 2017), loblolly pine (Neves *et al.*, 2013), poplar (Pinosio *et al.*, 2016) and to some extent *Eucalyptus* (Myburg *et al.*, 2014).

1.2.3. Functional impact of CNVs

Evidences obtained to date suggest that CNVs are usually clustered in hotspots across the genome (Perry *et al.*, 2006; Itsara *et al.*, 2009; Girirajan *et al.*, 2013). The abundance of CNVs in gene-rich regions is still controversial as CNVs were reported to be over- (Cooper *et al.*, 2007) and under- (Redon *et al.*, 2006; Korbel *et al.*, 2007) represented in genic sequences. This discrepancy may reflect i) biases due to the genotyping technology or the coverage of the genome or ii) differences between copy number loss and copy number gain events with regard to their selective effect, with gene losses expected to occur less frequently as they are supposedly more detrimental to the organism than gene gains (Conrad *et al.*, 2006). Gene redundancy can mitigate the phenotypic effect of gene loss or inactivation. Consequently, CNVs overlapping genes are expected to be more abundant in multi-gene families than in single-copy genes. Some empirical data support this assumption (She *et al.*, 2008) but since CNVs affect certain multi-gene families more than others (particularly NB-LRR and RLK gene families in plants) and are not restricted to multi-gene families (McHale *et al.*, 2012), the alternative hypothesis that CNVs can affect single-copy genes as frequently as large gene families cannot be excluded. A better characterization of the distribution of CNVs across the genome and particularly in coding regions is expected to enhance our understanding of the impact of CNVs on genome stability and the evolution of gene content.

CNVs can overlap a gene sequence entirely (full-CNVs) or partially (partial-CNVs) and can influence gene expression in different ways (Gamazon and Stranger, 2015). Full-CNVs are expected to be less deleterious than partial-CNVs and mainly alter gene expression through dosage effects (McCarroll *et al.*, 2006). Partial-CNVs on the other hand, mostly cause coding sequence disruptions through exons reshuffling, generation of splicing variants or formation of gene fusion (Korbel *et al.*, 2007; Stranger *et al.*, 2007). CNVs can also change gene expression *via* alterations of *cis* regulatory sequences (Merla *et al.*, 2006) or when it involves relocation of the gene in a different genomic context (Rodriguez-Revenga *et al.*, 2007). Expression level changes caused by CNVs acting on *trans* regulatory elements were also reported (Ricard *et al.*, 2010; Gamazon *et al.*, 2011).

Deletions, duplications and loss-of-function (LOF) mutations can cause gene dosage effects that alter gene expression in a linear or non-linear way (Gamazon and Stranger, 2015). To mitigate the negative fitness impact of gene dosage imbalance, organisms can rely on compensation mechanisms at the transcriptional and/or post-transcriptional level (Veitia *et al.*, 2013).

1.2.4. CNV role in adaptation and evolution

CNVs are less numerous than SNVs in the genome of an individual but they encompass larger DNA sequences ranging from a few hundred bases to several mega-bases. Consequently, the fraction of the genome affected by CNVs is larger than that altered by SNVs. Furthermore, CNVs are abundant, influence gene function and expression and are involved in downstream phenotypic variation which underlines their importance. Figure 1.3 summarize the contribution of CNV generation rate, transmission patterns and selective values to i) the shaping of the standing genetic variation within populations and ii) the pacing of species evolution.

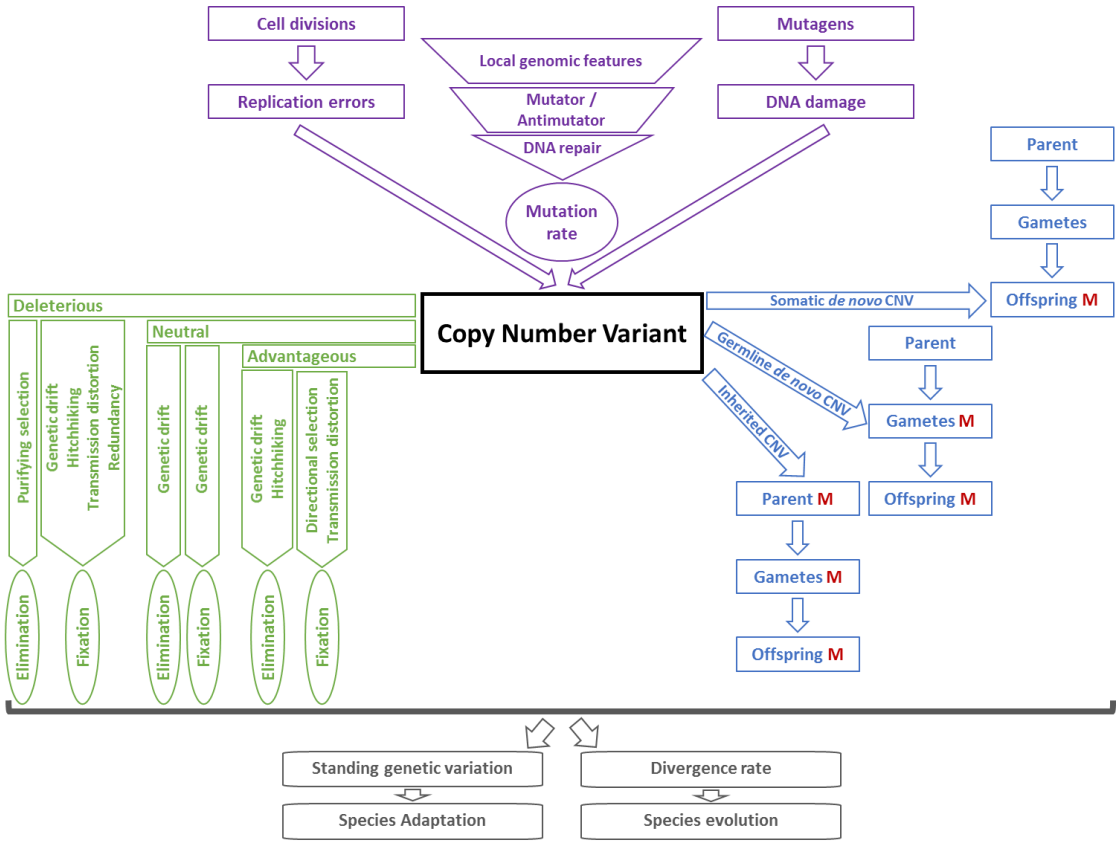


Figure 1.3: Role of CNVs in adaptation and evolution. The three factors that determine CNV diversity within species, namely CNV origin and mutation rate (in purple); transmission patterns (in blue) and fate (in green), are discussed in the following sections of this chapter.

1.3. CNV generation

Cell division requires the replication of the genetic material. Errors that occur during this process can introduce new variants that are transmitted to the daughter cells. Genetic exchanges between chromatids or chromosomes (crossing-overs) during meiosis can also generate new mutations that are transferred to the gametes then to the individuals of the next generation. Exogenous and endogenous mutagens frequently cause DNA damage that can give rise to new genetic variations if not repaired. The rate at which *de novo* mutation events occur and the efficiency of the DNA repair machinery determine both the rate at which new variants are introduced to the genetic pool of a population (see purple section in Figure 1.3).

1.3.1. CNV formation mechanisms

Through the examination of genomic rearrangements at high-resolution in human disorders and model organisms, Carvalho and Lupski (2016) classified SVs into two categories: recurrent and nonrecurrent rearrangements. Recurrent SVs are variants with simple structures that are found to be similar in size and content when unrelated individuals are compared. These rearrangements are the product of recombination events (mainly ectopic crossovers) that occur during meiosis in the germline. A molecular mechanism involved in their formation known as nonallelic homologous recombination (NAHR) usually requires the presence of low copy repeats (LCR) sequences in the vicinity and was shown to be biased toward the generation of more copy number losses than gains (Chen *et al.*, 2010). On the other hand, nonrecurrent SVs are mainly somatic variants with simple or complex structure and with unique size and content at a given locus among unrelated individuals. Their formation takes place during mitosis and is the result of template slippage or double strand breakage (DSB) repair errors associated with the DNA replication process. Different replication based mechanisms (RBM) can generate nonrecurrent SVs in a LCR dependent or independent fashion including non-homologous end joining (NHEJ), break-induced replication (BIR), serial replication slippage (SRS) and fork stalling and template switching (FoSTeS) (reviewed in Hastings *et al.*, 2009). The local genomic architecture (base content, methylation, recombination rate and repeats' organization: size, orientation, density and distribution) influences greatly the rate of SV formation (Saxena *et al.*, 2014; Makova and Hardison, 2015; Carvalho and Lupski, 2016).

1.3.2. Evolutionary consequences of mutation rate variation

De novo mutations (DNM) generate new variants that fuel the evolutionary process (Hodgkinson and Eyre-Walker, 2011; Jiang *et al.*, 2014; Ness *et al.*, 2015). The standing genetic variation in a population results from an equilibrium between the rate of introduction of new variants through mutation (mutation rate μ), and the fixation or elimination of the available variants by genetic drift or selection (Baer *et al.*, 2007; Katju and Bergthorsson, 2013). Furthermore, differences between the mutation rates of alleles can shape the evolutionary outcome, and the fate of alleles can be determined not only by their effect on fitness but also by the order and rate of their generation (Yampolsky and Stoltzfus, 2001). The observed mutation rate is an evolvable trait (Lynch, 2010a) and was shown to vary overtime (Latta *et al.*, 2013; Bromham *et al.*, 2015) and, between individuals (Haag-Liautard *et al.*, 2007; Conrad *et al.*, 2011) and loci (Hodgkinson and Eyre-Walker, 2011). Different factors contributes to the mutation rate variation at the intraspecific level including i) the fluctuation of selection pressure overtime, ii) the effective population size Ne, iii) the individual genetic background and particularly mutation rate modifiers (mutators and antimutators), iv) the molecular mechanisms involved in the formation of the mutation and v) the local features of genome architecture (Sniegowski *et al.*, 2000; Chen *et al.*, 2010; Latta *et al.*, 2013; Raynes and Sniegowski, 2014; Ness *et al.*, 2015; Sung *et al.*, 2016). The equilibrium mutation rate reached in a population is a tradeoff between the need to generate new variants for adaptation and the imperative of reducing the disruptive or even harmful effect of most mutations (Sniegowski *et al.*, 2000). Accurate estimates of the mutation rate genome-wide and for different types of genetic variations and a proper empirical characterization of the level and source of μ variation are essentials for a better understanding of species evolution and adaptation.

1.3.3. Methods for CN mutation rate estimation

Five methods have been used for the estimation of copy number (CN) mutation rates. (1) Direct estimates for single locus where mutations induce a quantifiable change in genotype or phenotype provided accurate per-locus rate estimations (Lam and Jeffreys, 2007; Watanabe *et al.*, 2009) but since these rates are locus specific, they may be biased and do not inform on the mutation rate spectrum across the genome. Estimates were also derived (2) from CNV frequency in populations using a mutation-selection balance population genetic theory (Lupski, 2007) and (3) from whole genome sequencing data

using the age distribution of duplicated genes (Lynch and Conery, 2000, 2003). But since these two indirect methods rely on hypotheses that are not necessarily true in natural populations, the estimates they provide are likely to be underestimated in some situations and overestimated in others. (4-5) The two direct approaches that are believed to have provided the most accurate genome-wide estimates used mutation accumulation line (MA) (Lipinski *et al.*, 2011; Schrider *et al.*, 2013) or pedigree (trios) (Itsara *et al.*, 2010) data. MA experiments allow the estimation of the mutation rate in conditions where purifying selection has a minimal effect but this method has the significant disadvantages of i) identifying only neutral or slightly deleterious variants, ii) drawing conclusions only for a limited number of genetic backgrounds and iii) being applicable only for species with short generation times that are not highly sensitive to inbreeding. The use of pedigrees is straightforward but is contingent on the availability of high-quality genotyping data for large data sets of trios or full-sib families. For species with large and complex genomes and a long generation time such as conifer trees, the use of pedigree data is the only option for the accurate estimation of CN mutation rate genome-wide.

1.3.4. Mutation rate estimates for different genetic variations

The different forms of genetic variations present in the genome are generated by diverse molecular processes. In addition, they are subject to repair mechanisms with different efficiency levels and have distinct distribution profiles across the genome. Consequently, their *de novo* mutation rates are expected to be very different. Table 1.1 includes estimates for the mutation rate of different types of genetic variations in multicellular organisms and shows that these rates can vary several orders of magnitude.

Table 1.1: Mutation rate estimates for different genetic variations.

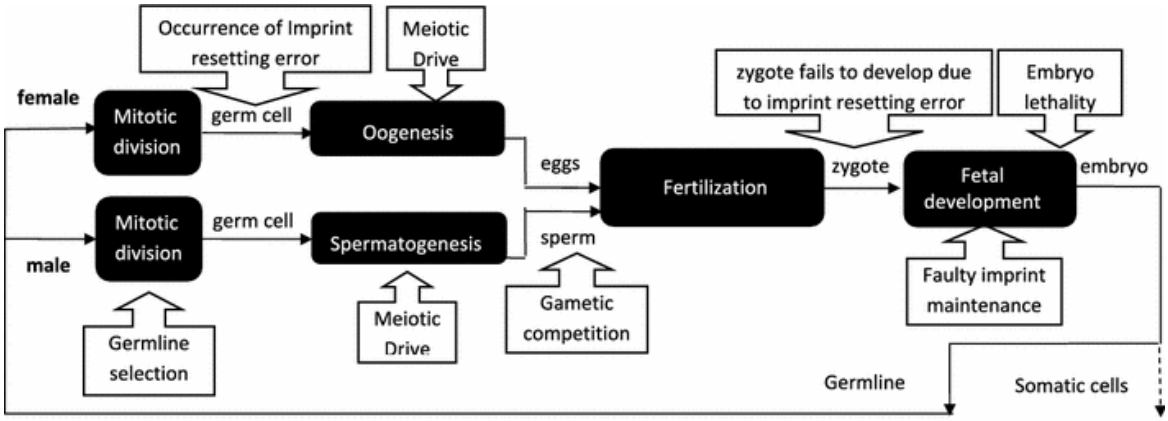
Genetic variation	Mutation rate estimates (mutation per generation)	References
SNVs	$1 \times 10^{-8} – 1.7 \times 10^{-8}$	Bouillé and Bousquet, 2005 Conrad *et al.*, 2011 Kong *et al.*, 2012
SSRs	$2.7 \times 10^{-4} – 1 \times 10^{-3}$	Sun *et al.*, 2012
Indels	$2 \times 10^{-10} – 5.8 \times 10^{-10}$	Lynch, 2010b
CNVs	$2.5 \times 10^{-6} – 3.6 \times 10^{-2}$	Turner *et al.*, 2008 Hehir-Kwa *et al.*, 2011
MEs	$3 \times 10^{-3} – 4.6 \times 10^{-2}$	Ray and Batzer, 2011 Stewart *et al.*, 2011

1.4. CNV transmission through generations

Genetic variations observed in the genome of an individual are either inherited from its parents or acquired through *de novo* mutation events. Part of the variants carried by the parents are transmitted to the gametes then to the offspring after fecundation. *De novo* mutations generated in the germline cells or during the genesis of gametes *via* meiosis are also transmitted to the next generation. Somatic *de novo* variants introduced in the genome of the parents (excluding the germline cells) are not transmitted to the descendants in animal species. In plants however, somatic mutations can be transmitted to the next generation because there is no clear separation between germline and somatic cells and the mutations that accumulate in the vegetative tissues during the growth of the individual can be transmitted to the gametes and then to the offspring when the vegetative meristem (within the bud) differentiates into reproductive organs (see blue section in Figure 1.3).

1.4.1. Transmission distortions: definition, causes and consequences

Genetic variations can be transmitted to the next generation according to Mendel's laws of inheritance or not (called Mendelian or non-Mendelian segregation, respectively). Transmission distortions (TDs) occur when an allele is preferentially transmitted to the next generation at the expense of alternative alleles. This departure from Mendelian expectations is observable in heterozygous individuals and is the consequence of disruptive mechanisms operating during the gametic or zygotic stages of development (Figure 1.4). TDs can be the result of germline selection, meiotic drive, gametic completion, male-female incompatibilities (pollen and pistil), embryo lethality (due to deleterious genotypes), mother-embryo incompatibility or faulty imprint resetting and maintenance (Huang *et al.*, 2013). TDs are under genetic control (Lyttle, 1991) and are the result of complex mechanisms usually involving a responder locus (target of the distortion) and one or more distorter locus(ci) (linked or unlinked modifiers of the level of distortion). TDs can cause a wide range of frequency changes: from mild distortions to a complete skew of transmission in favor of one allele (Koide *et al.*, 2012; Huang *et al.*, 2013). Some balancing forces (recombination, mutation, genetic drift) can interfere and counter TD effects on frequency changes leading to the preservation of both alleles (Polański *et al.*, 1998).



(1) **Germline selection** - Germ cell life cycle begins when a mature embryo is formed. The germ cells first start division through mitosis. During mitosis, mechanisms such as mutation, recombination and gene conversion, collectively called germline selection mechanisms cause cells with certain genotypes to be produced at a higher proportion than others. Hence, germ cells entering the next stage, meiosis, have an imbalanced genotype ratio.

(2) **Meiotic drive** - Female meiosis is called oogenesis, and male spermatogenesis. Since oogenesis is asymmetric by nature, only one of the four chromatids becomes a functional gamete, and the others become polar bodies and are eliminated. The chromatid of the haplotype with structural advantage in facilitating the orientation and replication during meiosis tends to be transmitted more. This mechanism is called meiotic drive. Although rare, meiotic drive can occur in male eukaryotes as well. There is another type of meiotic drive called sex chromosome drive that occurs during spermatogenesis, which leads to unequal production of X- or Y-bearing gametes.

(3) **Gametic competition** - In some male organisms, sperms survived through meiotic drive tend to compete with each other to achieve fertilization. This is called gametic selection. Well-studied models of gametic selection include t-haplotype system in mouse and segregation distorter in drosophila.

(4) **Imprinting errors** - Imprint resetting occurs during the postimplantation stage, where parental imprints are erased and re-established. When an error occurs during imprint resetting, the resulting embryo may be incompatible for survival. Faulty imprint maintenance during embryonic development can also lead to the death of embryos.

(5) **Embryo lethality** - After the embryo is formed, there are other mechanisms of selection termed embryo lethality. One example of embryo lethality is the Rh+ system where mother and fetal blood types are incompatible. During delivery when the placenta ruptures, upon the blending of maternal blood with fetal blood stream, the fetus dies.

Figure 1.4: Underlying biological mechanisms behind transmission distortions [Figure reproduced by permission from Springer: Human Genetics, Huang *et al.*, 2013, copyright 2012].

1.4.2. Transmission distortions reported in other species

Transmission distortions are still largely understudied. The research reported to date is mostly specific to a single locus and has often focused on the particular case of meiotic drive (Didion *et al.*, 2015). More genome-wide analyses of this phenomenon would help to better understand the plethora of mechanisms involved in non-Mendelian transmissions. In human and model organisms, TDs were identified for SNVs, inversions and genic or intronic sequences (Huang *et al.*, 2013). In plants, TDs were linked to interpopulation genetic divergence and reproductive isolation (Leppälä *et al.*, 2008, 2013; Matsubara *et*

al., 2011). The effect of parental sex on TDs was investigated in rice and a few loci displaying sex-independent TDs were reported in this species (Koide *et al.*, 2008, 2012). TDs caused by meiotic drive have been described in maize (Buckler *et al.*, 1999) and monkeyflower (Fishman and Willis, 2005). In these examples, knob structures (acting as artificial centromeres) are responsible for the preferential transmission of selfish DNA elements through a physical mechanism. The most extensive analysis of TDs that we are aware of, was conducted in *Arabidopsis thaliana* where 130 distorted loci were identified and linked to different phenotypes associated with fecundation and early embryo development (Pagnussat *et al.*, 2005).

1.4.3. Examples of transmission distortions involving CNVs

Three cases where CNVs contributed to TDs as responder or distorter loci were reported in mice and worm. Table 1.2 summarizes the features of these distortion systems.

Table 1.2: Systems where CNVs contribute to transmission distortions.

Species	*Mus musculus*	*Mus musculus*	*Caemorhabditis elegans*
System	*Om* (*ovum mutant*)	R2d2^{WSB}	peel-1/zeel-1
CNV type	PAV[a]	CNG[b]	PAV[a]
CNV location	Distorter locus	Responder locus	Distorter locus
TD level	0.6	0.6 – 1	0 – 1
Responder locus[c]	*Om*	*R2d2*	*peel-1*
Distorter locus[d]	*Li-ch1*	Multiple	zeel-1
Linked distorter	No	No	Yes
Sperm dependent	Yes	Unknown	Yes
Reference	Pardo-Manuel de Villena *et al.*, 2000	Didion *et al.*, 2015	Seidel *et al.*, 2011

[a]Presence-absence variation; [b]Copy number gain; [c]locus subject to transmission distortion; [d]locus that control the level of transmission distortion of the responder locus.

1.5. The fate of CNVs

CNVs can be detrimental, neutral or advantageous to the organism harboring them. The fate of CNVs (fixation or elimination) is determined in part by their effect on fitness and selective value, the effective size of the population and the degree of their linkage with other genetic variations in the genome. Neutral CNVs can be fixed or purged under the action of genetic drift. Deleterious CNVs are often eliminated by purifying selection but occasionally they can be maintained *via* genetic drift, hitchhiking (if linked to advantageous loci), transmission distortion or genetic redundancy (if gene function can be performed fully or partially by other genes). Advantageous CNVs are mainly retained in the genome through directional selection and transmission distortion but, can also be eliminated as a consequence of genetic drift or hitchhiking (if linked to deleterious or lethal loci) (see green section in Figure 1.3).

1.5.1. Examples of deleterious, neutral and advantageous CNVs

The general trend observed for CNVs shows that they are i) underrepresented in coding sequences and gene-rich regions (Redon *et al.*, 2006; Korbel *et al.*, 2007), ii) more frequently detected in multi-gene families and genes involved in environmental responses rather than in genes involved in basic cellular functions (Korbel *et al.*, 2009), iii) more likely to overlap genes with low connectivity than highly connected genes (hub genes) within metabolic networks (Kim *et al.*, 2007). Selective forces also act differently according to the class of CNVs; for instance, copy number losses (or deletion) are expected to be more deleterious and under stronger purifying selection than copy number gains (or duplications). Empirical data show that deletions i) occur less frequently in coding sequences relative to duplications (Emerson *et al.*, 2008), ii) are detected in introns more than in exons (Emerson *et al.*, 2008), iii) display lower frequencies in natural populations compared to duplications (Locke *et al.*, 2006). These observations support the afore mentioned hypothesis.

CNVs with deleterious effects on phenotype were shown to be involved in cancer and human disease (reviewed in Shlien and Malkin, 2009; Girirajan *et al.*, 2011). Advantageous CNVs were also identified in humans. Perry and collaborators (2007) associated copy number gains of the human salivary amylase gene *AMY1* with the dietary content in starch, which suggest that this CNV is under positive selection. The gene *UGT2B17* that metabolizes steroids and foreign compounds displays CNVs that are under

balancing and positive selection in European and Asian populations, respectively (Xue *et al.*, 2008). Copy number variation in the olfactory receptor genes on the other hand were shown to be neutral (Nozawa *et al.*, 2007).

1.5.2. Biological processes and gene functions associated with CNVs

Studies of the association between CNVs and quantitative traits are still lacking in plants. However, a few examples were reported in crop plants (Zmienko *et al.*, 2014). Examples of CNVs involved in adaptive and phenotypic traits control include i) flowering time (Díaz *et al.*, 2012) and height (Li *et al.*, 2012) in wheat, ii) tolerance to submergence in rice (Xu *et al.*, 2006), iii) tolerance to aluminum toxicity in maize (Maron *et al.*, 2013), iv) biotic resistance, seed composition, flowering and maturity time, organ size and final biomass in soybean (Cook *et al.*, 2012; Li *et al.*, 2014), seedlessness in grapevine (Di Genova *et al.*, 2014) and reproductive morphology in cucumber (Zhang *et al.*, 2015).

In a survey of presence-absence variations (PAV) in 80 accessions of *Arabidopsis thaliana*, Tan and collaborators (2012) found that functional classes involved in basic biological processes (such as heat shock proteins and ABC transporter) and transcription regulation (such as Myb and HLH) are less affected by PAVs than gene categories involved in stress response and disease resistance.

Gene ontology (GO) annotations and functional enrichment analyses for genes displaying CNVs showed that these genetic variations are associated with diverse biological processes including: stress response and protein modification in sorghum (Shen *et al.*, 2015); responses to stresses, cell death, protein phosphorylation and defense response in rice (Yu *et al.*, 2013; Bai *et al.*, 2016) and; disease resistance and protein kinases in barley (Muñoz-Amatriaín *et al.*, 2013). In trees, recent studies show that CNVs are enriched in genes involved in response to stress, defense response, cell death and protein modification processes in North American spruces (Prunier *et al.*, 2017) and; resistance to abiotic and biotic stresses in poplar (Pinosio *et al.*, 2016).

Debolt (2010) submitted *A. thaliana* plants to temperature and biotic stresses for five consecutive generations and then quantified the appearance of CNVs relatively to the genome of the original accession. Genes displaying CNVs as a result of plant growth in stress conditions included resistance genes (from NBS-LRR class), Leucine-rich kinases

(involved in hormone mediated signalling), F-box proteins (involved in heat acclimation) and auxin response genes.

1.6. Project context, objectives and hypotheses

The work presented in this thesis is part of an effort dedicated to the i) development of a better understanding of the evolution of conifer trees in comparison to other plant species; ii) identification of the mechanisms involved in the genetic adaptation of forest trees to environmental changes; iii) development of resources to support genetic and phenotypic analyses of spruce and pine trees and iv) formulation of strategies and design of tools for marker assisted tree improvement and conservation (under the projects SMarTForests and GenAC co-directed by Prof. John MacKay and Prof. Jean Bousquet, Université Laval, Canada).

The motivation for this research project came in part from the emergence of new genomic resources generated for conifer trees including large white spruce SNP-genotyping data sets from custom-made chips (Pavy *et al.*, 2013; Beaulieu *et al.*, 2014) that were obtained for a large number of individuals. At the same time, reference genome assemblies were also generated for three conifer species (Birol *et al.*, 2013; Nystedt *et al.*, 2013; Neale *et al.*, 2014); the gene catalog available for white spruce was extended and improved upon (Rigault *et al.*, 2011); a genetic map with higher resolution was published (Pavy *et al.*, 2017) and; large sets of expression data were analyzed for different tissues (Raherison *et al.*, 2012, 2015). The genomic analyses conducted so far however, relied mainly on SNP data and; mitochondrial and chloroplastic DNA to some extent. At the onset of the present work, little was known about CNVs in trees in general and almost no data were available for conifers.

In this study, we took advantage of the availability of large data sets of SNP-array raw intensity data to scan the gene space of *P. glauca* for CNVs. The data were initially generated for genetic mapping, genetic association studies, and genomic selection modelling based on gene SNPs (Pavy *et al.*, 2013; Beaulieu *et al.*, 2014), and we reanalyzed the raw intensity data using CNV detection methods. Each of the SNPs was positioned within a protein coding gene; these genes in turn were distributed throughout the genome (Pavy *et al.*, 2013, 2017). Our goal was to identify CNVs by examining genotyping data for more than 14 000 genes in 55 full-sib families (for 3663 individuals),

provide the first estimates of CN mutation rates in trees and investigate the inheritance of copy number variants.

1.6.1. Objectives

The specific objectives of this Ph.D. thesis project were to:

- Develop an approach for CNV detection in white spruce.

- Characterize genic CNVs abundance and identify their classes.

- Estimate the copy number mutation rate genome-wide.

- Discover CNVs inheritance patterns.

- Examine the potential role of CNVs in the shaping of the standing genetic variation.

1.6.2. Hypotheses

Knowledge about CNVs in trees is still limited. Hence, it is challenging to formulate hypotheses regarding the biology of these genetic variations, particularly in light of the distinct features of conifers life history and genome architecture. Nevertheless, based on the data available for multicellular organisms and particularly human, we propose the following hypotheses:

Hypothesis 1: Genic CNVs affect a small proportion of the gene space.
Since CNVs can have considerable impact on gene function and downstream phenotype, genic CNVs are expected to be rare.

Hypothesis 2: Gene copy losses are expected to be more abundant than gene copy gains.
For many species, reported CNVs are mostly copy number losses. We see no reason to expect a different trend in *P. glauca*.

Hypothesis 3: Copy number mutation rate is low and variable across the genome.
Hypothesis 4: Copy number mutation rate is associated with gene expression.
Previous work in other species suggest that CNVs are mainly deleterious and the rate of their formation is dependent on the local genomic architecture and associated with gene expression. Consequently, we expect the copy number mutation rate to be low, variable

across the genome and linked to gene expression according to one of the following hypotheses: i) transcription-coupled repair hypothesis (TCRH); ii) transcription-associated mutagenesis hypothesis (TAMH) or iii) drift-barrier hypothesis (DBH).

Hypothesis 5: Transmission distortions (TDs) are expected to be frequent as a mechanism of restoration of the normal two-copy genotype in a diploid organism and cause significant frequency changes between generations.

Hypothesis 6: TDs are genetically controlled.

CNVs are important genetic variations with considerable impacts on gene function and downstream phenotypes. The work presented in this thesis is a contribution for a better understanding of the biology of these variations outside the context of human disease. For economically and ecologically important species such as conifers, we believe that the application of this knowledge within tree improvement and conservation programs will be extremely valuable in the long term.

1.7. References

Alkan C, Coe BP, Eichler EE (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.

Baer CF, Miyamoto MM, Denver DR (2007). Mutation rate variation in multicellular eukaryotes: Causes and consequences. *Nat Rev Genet* **8**: 619–631.

Bai Z, Chen J, Liao Y, Wang M, Liu R, Ge S, *et al.* (2016). The impact and origin of copy number variations in the Oryza species. *BMC Genomics* **17**: 261.

Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J (2014). Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genomics* **15**: 1048.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA *et al.* (2013). Assembling the 20 Gb white spruce (*Picea glauca)* genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**: 1492–1497.

Blackburn A, Göring HH, Dean A, Carless MA, Dyer T, Kumar S *et al.* (2013). Utilizing extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. *Eur J Hum Genet* **21**: 404–409.

Bouillé M, Bousquet J (2005). Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): Implications for the long-term maintenance of genetic diversity in trees. *Am J Bot* **92**: 63–73.

Bromham L, Hua X, Lanfear R, Cowman PF (2015). Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *Am Nat* **185**: 507–524.

Buckler ESI, Phelps-Durr TL, Buckler CSK, Dawe RK, Doebley JF, Holtsford TP (1999). Meiotic drive of chromosomal knobs reshaped the maize genome. *Genetics* **153**: 415–426.

Burns RM, Honkala BH (1990). *Silvics of North America, Vol. 1 Conifers,* US Department of Agriculture, Forest Service: Washington DC.

Carvalho CMB, Lupski JR (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**: 224–238.

Chen JM, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP (2010). Genomic rearrangements in inherited disease and cancer. *Semin Cancer Biol* **20**: 222–233.

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**: 75–81.

Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F *et al.* (2011). Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**: 712–714.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM *et al.* (2012). Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* **338**: 1206–1209.

Cooper GM, Nickerson D a, Eichler EE (2007). Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* **39**: S22-S29.

Debolt S (2010). Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* **2**: 441–453.

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012). Copy number variation affecting the *Photoperiod-B1* and *Vernalization-A1* genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS One* **7**: e33234.

Didion JP, Morgan AP, Clayshulte AMF, Mcmullan RC, Yadgary L, Petkov PM *et al.* (2015). A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLoS Genet* **11**: e1004850.

Ellegren H, Galtier N (2016). Determinants of genetic diversity. *Nat Rev Genet* **17**: 422–433.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* **320**: 1629–1631.

Feng J, Fu W, Sun F (2010). *Frontiers in Computational and Systems Biology*, Springer: London, Dordrecht, Heidelberg, New York.

Fishman L, Willis JH (2005). A novel meiotic drive locus almost completely distorts segregation in *Mimulus* (monkeyflower) hybrids. *Genetics* **169**: 347–353.

Franceschini T, Schneider R (2016). Factors affecting plantation grown white spruce (*Picea glauca)* acoustic velocity. *J For* **114**: 629–637.

Gamazon ER, Nicolae DL, Cox NJ (2011). A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet* **7**: e1001292.

Gamazon ER, Stranger BE (2015). The impact of human copy number variation on gene expression. *Brief Funct Genomics* **14**: 352–357.

Di Genova A, Almeida AM, Muñoz-Espinoza C, Vizoso P, Travisany D, Moraga C *et al.* (2014). Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol* **14**: 7.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH *et al.* (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**: 344–347.

Girirajan S, Campbell CD, Eichler EE (2011). Human copy number variation and complex genetic disease. *Annu Rev Genet* **45**: 203–226.

Girirajan S, Dennis MY, Baker C, Malig M, Coe BP, Campbell CD *et al.* (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am J Hum Genet* **92**: 221–237.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D *et al.* (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445**: 82–85.

Hastings P, Lupski J, Rosenberg S, Ira G (2009). Mechanisms of change in gene copy number. *Nat Rev Genet* **10**: 551–564.

Hehir-Kwa JY, Rodriguez-Santiago B, Vissers LE, de Leeuw N, Pfundt R, Buitelaar JK *et al.* (2011). *De novo* copy number variants associated with intellectual disability have a paternal origin and age bias. *J Med Genet* **48**: 776–778.

Hodgkinson A, Eyre-Walker A (2011). Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**: 756–766.

Huang LO, Labbe A, Infante-Rivard C (2013). Transmission ratio distortion: Review of concept and implications for genetic association studies. *Hum Genet* **132**: 245–263.

Ingvarsson PK, Hvidsten TR, Street NR (2016). Towards integration of population and comparative genomics in forest trees. *New Phytol* **212**: 338-344.

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D *et al.* (2009). Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* **84**: 148–161.

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ *et al.* (2010). *De novo* rates and selection of large copy number variation. *Genome Res* **20**: 1469–1481.

Jaramillo-Correa JP, Beaulieu J, Bousquet J (2001). Contrasting evolutionary forces driving population structure at expressed sequence tag polymorphisms, allozymes and quantitative traits in white spruce. *Mol Ecol* **10**: 2729–2740.

Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP (2014). Environmentally responsive genome-wide accumulation of *de novo Arabidopsis thaliana* mutations and epimutations. *Genome Res* **24**: 1821–1829.

Katju V, Bergthorsson U (2013). Copy-number changes in evolution: Rates, fitness effects and adaptive significance. *Front Genet* **4**: 273.

Kim PM, Korbel JO, Gerstein MB (2007). Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci USA* **104**: 20274–20279.

Koide Y, Ikenaga M, Sawamura N, Nishimoto D, Matsubara K, Onishi K *et al.* (2008). The evolution of sex-independent transmission ratio distortion involving multiple allelic

interactions at a single locus in rice. *Genetics* **180**: 409–420.

Koide Y, Shinya Y, Ikenaga M, Sawamura N, Matsubara K, Onishi K *et al.* (2012). Complex genetic nature of sex-independent transmission ratio distortion in Asian rice species: the involvement of unlinked modifiers and sex-specific mechanisms. *Heredity* **108**: 242–247.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G *et al.* (2012). Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**: 471–475.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Korbel JO, Kim PM, Chen X, Urban AE, Snyder M, Gerstein MB (2009). The current excitement about copy-number variation: How it relates to gene duplication and protein families. *Curr Opin Struct Biol* **18**: 366–374.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM *et al.* (2014). Insights into conifer giga-genomes. *Plant Physiol* **166**: 1724–1732.

Lam K-WG, Jeffreys AJ (2007). Processes of *de novo* duplication of human *α*-globin genes. *Proc Natl Acad Sci USA* **104**: 10950–10955.

Latta LC, Morgan KK, Weaver CS, Allen D, Schaack S, Lynch M (2013). Genomic background and generation time influence deleterious mutation rates in *Daphnia*. *Genetics* **193**: 539–5444.

Leppälä J, Bechsgaard JS, Schierup MH, Savolainen O (2008). Transmission ratio distortion in *Arabidopsis lyrata*: effects of population divergence and the S-locus. *Heredity* **100**: 71–78.

Leppälä J, Bokma F, Savolainen O (2013). Investigating incipient speciation in *Arabidopsis lyrata* from patterns of transmission ratio distortion. *Genetics* **194**: 697–708.

Li Y, Xiao J, Wu J, Duan J, Liu Y, Ye X *et al.* (2012). A tandem segmental duplication (TSD) in green revolution gene *Rht-D1b* region underlies plant height variation. *New Phytol* **196**: 282–291.

Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z *et al.* (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**: 1045–1052.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U (2011). High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr Biol* **21**: 306–310.

Locke DP, Sharp AJ, McCarroll S a, McGrath SD, Newman TL, Cheng Z, *et al.* (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* **79**: 275–290.

Lupski JR (2007). Genomic rearrangements and sporadic disease. *Nat Genet* **39**: S43-S47.

Lynch M (2010a). Evolution of the mutation rate. *Trends Genet* **26**: 345–352.

Lynch M (2010b). Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* **107**: 961–968.

Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lynch M, Conery JS (2003). The evolutionary demography of duplicated genes. *J Struct Funct Genomics* **3**: 35–44.

Lyttle TW (1991). Segregation distorters. *Annu Rev Genet* **25**: 511–557.

Makova KD, Hardison RC (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**: 213–223.

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ *et al.* (2013). Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proc Natl Acad Sci USA* **110**: 5241–5246.

Matsubara K, Ebana K, Mizubayashi T, Itoh S, Ando T, Nonoue Y *et al.* (2011). Relationship between transmission ratio distortion and genetic divergence in intraspecific rice crosses. *Mol Genet Genomics* **286**: 307–319.

McCarroll S a, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC *et al.* (2006). Common deletion polymorphisms in the human genome. *Nat Genet* **38**: 86–92.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL *et al.* (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol* **159**: 1295–1308.

Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT *et al.* (2006). Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am J Hum Genet* **79**: 332–341.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C *et al.* (2011). Mapping copy number variation by population scale genome sequencing. *Nature* **470**: 59–65.

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B *et al.* (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol* **14**: R58.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J *et al.* (2014). The genome of *Eucalyptus grandis*. *Nature* **510**: 356–362.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008). Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Mol Ecol* **17**: 3599–3613.

Neale DB, Wegrzyn JL, Stevens K a, Zimin A V, Puiu D, Crepeau MW *et al.* (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol* **15**: R59.

Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD (2015). Extensive *de novo* mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Res* **25**: 1739–1749.

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013). A high-density gene map of loblolly pine (*Pinus taeda L.*) based on exome sequence capture genotyping. G3 *Genes|Genomes|Genetics* **4**: 29–37.

Nozawa M, Kawahara Y, Nei M (2007). Genomic drift and copy number variation of sensory receptor genes in humans. *Proc Natl Acad Sci USA* **104**: 20421–20426.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG *et al.* (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.

O'Connell LM, Mosseler a, Rajora OP (2006). Impacts of forest fragmentation on the mating system and genetic diversity of white spruce (*Picea glauca*) at the landscape level. *Heredity* **97**: 418–426.

Pagnussat GC, Yu HJ, Ngo QA, Rajani S, Mayalagu S, Johnson CS *et al.* (2005). Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development* **132**: 603–614.

Pardo-Manuel de Villena F, De La Casa-Esperon E, Briscoe TL, Sapienza C (2000). A genetic test to determine the origin of maternal transmission ratio distortion: Meiotic drive at the mouse *Om* locus. *Genetics* **154**: 333–342.

Parent GJ, Raherison E, Sena J, MacKay JJ (2015). Forest tree genomics: Review of progress. In: Adam-Blondon A-F, Plomion C (eds) *Land Plants - Trees*, Elsevier: San Diego. Vol 74, pp 39–92.

Pavy N, Namroud M-C, Gagnon F, Isabel N, Bousquet J (2012). The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity* **108**: 273–284.

Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B *et al.* (2013). Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Mol Ecol Resour* **13**: 324–336.

Pavy N, Lamothe M, Pelgas B, Gagnon F, Birol I, Bohlmann J *et al.* (2017). A high resolution reference genetic map positioning 8.8K genes for the conifer white spruce: Structural genomics implications and correspondence with physical distance. *Plant J* **90**: 189–203.

Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Caceres AM *et al.* (2006). Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci USA* **103**: 8006–8011.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R *et al.* (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**: 1256–1260.

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC *et al.* (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Mol Biol Evol* **33**: 2706–2719.

Polański A, Chakraborty R, Kimmel M, Deka R (1998). Dynamic balance of segregation distortion and selection maintains normal allele sizes at the myotonic dystrophy locus. *Math Biosci* **147**: 93–112.

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.

Raherison E, Rigault P, Caron S, Poulin PL, Boyle B, Verta JP *et al.* (2012). Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. *BMC Genomics* **13**: 434.

Raherison ESM, Giguère I, Caron S, Lamara M, Mackay JJ (2015). Modular organization

of the white spruce (*Picea glauca*) transcriptome reveals functional organization and evolutionary signatures. *New Phytol* **207**: 172–187.

Ray DA, Batzer MA (2011). Reading TE leaves: New approaches to the identification of transposable element insertions. *Genome Res* **21**: 813–820.

Raynes Y, Sniegowski PD (2014). Experimental evolution and the dynamics of genomic mutation rate modifiers. *Heredity* **113**: 375–380.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD *et al.* (2006). Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Ricard G, Molina J, Chrast J, Gu W, Gheldof N, Pradervand S *et al.* (2010). Phenotypic Consequences of copy number variation: Insights from Smith-Magenis and Potocki-Lupski syndrome mouse models. *PLoS Biol* **8**: e1000543.

Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, Mackay JJ (2011). A white spruce gene catalog for conifer genome analyses. *Plant Physiol* **157**: 14–28.

Rodriguez-Revenga L, Mila M, Rosenberg C, Lamb A, Lee C (2007). Structural variation in the human genome: The impact of copy number variants on clinical diagnosis. *Genet Med* **9**: 600–606.

Saxena RK, Edwards D, Varshney RK (2014). Structural variations in plant genomes. *Brief Funct Genomics* **13**: 296–307.

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP *et al.* (2007). Challenges and standards in integrating surveys of structural variation. *Nat Genet* **39**: S7–S15.

Schrider DR, Houle D, Lynch M, Hahn MW (2013). Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics* **194**: 937–954.

Seidel HS, Ailion M, Li J, van Oudenaarden A, Rockman M V., Kruglyak L (2011). A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in *C. elegans*. *PLoS Biol* **9**: e1001115.

She X, Cheng Z, Zöllner S, Church DM, Eichler EE (2008). Mouse segmental duplication and copy number variation. *Nat Genet* **40**: 909–914.

Shen X, Liu ZQ, Mocoeur A, Xia Y, Jing HC (2015). PAV markers in *Sorghum bicolour*: genome pattern, affected genes and pathways, and genetic linkage map construction. *Theor Appl Genet* **128**: 623–637.

Shlien A, Malkin D (2009). Copy number variations and cancer. *Genome Med* **1**: 62.

Sniegowski PD, Gerrish PJ, Johnson T, Shaver A (2000). The evolution of mutation rates: Separating causes from consequences. *BioEssays* **22**: 1057–1066.

Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM *et al.* (2011). A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**: e1002236.

Stranger B, Forrest M, Dunning M, Ingle C (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–853.

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S *et al.* (2012). A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**: 1161–1165.

Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS *et al.* (2016). Evolution of the insertion-deletion mutation rate across the tree of life. *G3 Genes|Genomes|Genetics* **6**: 2583–2591.

Tan S, Zhong Y, Hou H, Yang S, Tian D (2012). Variation of presence/absence genes among *Arabidopsis* populations. *BMC Evol Biol* **12**: 86.

Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, *et al.* (2008). Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40**: 90–95.

Veitia RA, Bottani S, Birchler JA (2013). Gene dosage effects: Nonlinearities, genetic interactions, and dosage compensation. *Trends Genet* **29**: 385–393.

Warren RL, Keeling CI, Yuen MM Saint, Raymond A, Taylor GA, Vandervalk BP *et al.* (2015). Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J* **83**: 189–212.

Watanabe Y, Takahashi A, Itoh M, Takano-Shimizu T (2009). Molecular spectrum of spontaneous *de novo* mutations in male and female germline cells of *Drosophila melanogaster*. *Genetics* **181**: 1035–1043.

Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R, Heuer S *et al.* (2006). *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442**: 705–708.

Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M *et al.* (2008). Adaptive evolution of *UGT2B17* copy-number variation. *Am J Hum Genet* **83**: 337–346.

Yampolsky LY, Stoltzfus A (2001). Bias in the introduction of variation as an orienting factor in evolution. *Evol Dev* **3**: 73–83.

Yu P, Wang CH, Xu Q, Feng Y, Yuan XP, Yu HY *et al.* (2013). Genome-wide copy number variations in *Oryza sativa L*. *BMC Genomics* **14**: 649.

Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J *et al.* (2015). Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell* **27**: 1595–1604.

Zmienko A, Samelak A, Kozlowski P, Figlerowicz M (2014). Copy number polymorphism in plant genomes. *Theor Appl Genet* **127**: 1–18.

Chapter 2: High and variable copy number mutation rates in the gene space of *Picea glauca*

[Atef Sahli, Isabelle Giguère, Jean Bousquet and John MacKay (2017). High and variable copy number mutation rates in the gene space of *Picea glauca*. G3: Genes | Genomes | Genetics; manuscript accepted]

2.1. Abstract

Copy number variations (CNVs) are large genetic variations detected among the individuals of every multicellular organism examined so far. Knowledge of the copy number (CN) mutation rate (μ) spectrum is fundamental for understanding evolution and adaptation, but it is lacking for many species. In plants, the key characteristics of CNVs are poorly understood and even less is known about the rates at which they are generated. In this work, we developed an approach to identify genic CNVs in the conifer tree *Picea glauca*, a species with a large and complex genome and a long generation time. We used SNP-array raw intensity data for 3663 individuals belonging to 55 full-sib families to scan the gene space for CNVs and estimate the genome-wide CN mutation rate. Our findings show that CNVs affect a small proportion of the gene space and are predominantly copy number losses. CNVs were either inherited or generated through *de novo* events. CN mutation rate estimates span at least three orders of magnitude, can reach high levels and vary for different genes, alleles and CNV classes. Analysis of this broad range of CN mutation rates identified correlations with gene expression levels and the relationship between μ and gene expression is best explained within the frame of the drift-barrier hypothesis. This study shows that *de novo* mutations not only generate new copy number variants frequently in trees, which are generally well equipped to handle the resulting mutational load, but can also contribute as an orienting force that determines the fate of alleles.

2.2. Résumé

Les variations de nombre de copies (VNCs) sont des variations génétiques de grande taille qui ont été détectées dans tous les organismes multicellulaires examinés à ce jour. Les connaissances sur le spectre des taux de mutation (μ) du nombre de copies (NC) sont importantes pour mieux comprendre l'évolution et l'adaptation mais sont jusque-là très

limitées pour plusieurs espèces. Pour les plantes, les caractéristiques clés des VNCs sont peu connues en particulier le taux auquel elles sont générées. Dans ce travail, on a développé une approche pour l'identification des VNCs géniques chez le conifère arborescent *Picea glauca*, une espèce qui possède un génome large et complexe et un temps de génération long. On a utilisé des données brutes de puces de génotypage obtenues pour 3663 individus appartenant à 55 familles bi-parentales pour détecter des VNCs dans l'espace génique et estimer leur taux de mutation à travers le génome. Nos résultats montrent que les VNCs affectent une petite proportion de l'espace génique et sont majoritairement des pertes de nombre de copies. Les VNCs identifiées chez les descendants ont été soit héritées des parents ou générées *via* des évènements *de novo*. Les estimés du taux de mutation du NC couvrent au moins trois ordres de grandeur, peuvent atteindre des niveaux élevés et varient pour différents gènes, allèles et classes de VNCs. L'analyse du spectre des taux de mutation a permis d'identifier des corrélations entre le taux de mutation et le niveau d'expression des gènes et la relation entre μ et l'expression des gènes est mieux expliquée dans le cadre de l'hypothèse de barrière par la dérive génétique. Cette étude montre que les mutations *de novo* non seulement génèrent fréquemment de nouveaux polymorphismes de nombre de copies chez les arbres, mais peuvent aussi contribuer comme force évolutive dirigeante déterminant la destinée des allèles.

2.3. Introduction

Copy number variations (CNVs) are specific sequences (100 bp to few Mbp) that are present in variable numbers among individuals and are believed to play an important role in the evolution and adaptation of species (Katju and Bergthorsson, 2013). Although CNVs have been frequently detected in healthy and unhealthy populations of multicellular organisms (Zhang *et al.*, 2009; Swanson-Wagner *et al.*, 2010; Blackburn *et al.*, 2013; Zichner *et al.*, 2013; Chain *et al.*, 2014), little is known about the rate of their generation. The genome-wide copy number (CN) mutation rate spectrum has been reported for only a handful of model organisms [Human (Itsara *et al.*, 2010), mice (Egan *et al.*, 2007) and *Drosophila* (Schrider *et al.*, 2013)] mainly because their accurate estimation (which entails the reliable detection of rare mutation events) requires i) the analysis of a large number of individuals and ii) stringent criteria for variant calls from genotyping data (Fu *et al.*, 2010). Genome-scale identifications of CNVs have been reported in many plant species (Saxena *et al.*, 2014; Zmienko *et al.*, 2014); however, analyses of CN mutation rates are still scarce

particularly for non-model and perennial plants [CN mutation rates reported only for *Arabidopsis* (Ossowski *et al.*, 2010; Yang *et al.*, 2015) and maize (Jiao *et al.*, 2012)]. This is due in part to the additional challenges of estimating the rates of *de novo* mutations in plants where i) the use of mutation accumulation lines (MA) is often unpractical or even impossible (due to the long generation time of certain species), ii) the lack of genome-wide genotyping data (array or sequencing data) for large families- or trios- data sets, iii) the lack of reference genomes for species with large and complex genomes and iv) the relative difficulty of CNV identification in polyploid species.

To our knowledge, no attempt was made to estimate CN mutation rates in trees. White spruce (*Picea glauca* (Moench) Voss) is a perennial outcrossing monoecious gymnosperm with a long generation time (around 50 years) and a diploid giga-genome (20 Gbp) enriched in repeated sequences (Birol *et al.*, 2013; De La Torre *et al.*, 2014; Warren *et al.*, 2015). The long generation time and the size and complexity of the genome (together with the lack of a contiguous reference genome) prevents the use of whole-genome sequencing of individuals derived from MA lines for the estimation of genome-wide CN mutation rates. A family based approach coupled with the use of genome-scale array genotyping data on the other hand, will allow for the direct estimation of CN mutation rates for different lineages (and genetic background), genes and CNV classes. Estimation of mutation rates from the analysis of trios is expected to give underestimates because it would reflect newly generated somatic and germline mutation minus the proportion of *de novo* variants eliminated by the purifying selection or undetected due to technical limitations (Egan *et al.*, 2007; Katju and Bergthorsson, 2013). In this work, we took advantage of the availability of raw intensity data obtained from a SNP-array for 14 000 genes and 55 two-generation families (3663 individuals in total) to scan the gene space of the conifer tree *Picea glauca* for CNVs. Our objectives were: 1) to identify a high quality CNV set based on stringent criteria for variants calling; 2) classify CNVs as inherited or *de novo* variants; 3) estimate CN mutation rates and; 4) characterize their variation for different genes, alleles and CNV classes. This work allowed the testing of three hypotheses to explain the relationship between the mutation rate and gene expression levels.

2.4. Material and methods

2.4.1. Data sets

For the purpose of identifying genic CNVs using a cross-sample approach (Marioni *et al.*, 2007), we selected two subsets of raw intensity data previously generated for SNP genotyping analyses of 55 white spruce families (Pavy *et al.*, 2013; Beaulieu *et al.*, 2014). In the two data sets designated 54F and 1LF, *Picea glauca* trees from two generations pedigrees were genotyped using PGLM3 SNP-array (14 140 probes targeting 14 058 genes). The design of PGLM3 Infinium SNP-array (Illumina, San Diego, California) and the genotyping protocol are described in (Pavy *et al.*, 2013).

The data set 54F consist of the genotyping data of 54 full-sib families (family size range: 28 to 32) with their respective parents (total of 1650 offsprings + 37 parents). Each of the 37 parents was involved in one to five crosses and was used as male and female indistinctively in different crosses (Supplemental Table S2.9.1). This data set includes also two technical replicates for 24 trees, genotyped on different arrays for quality control. The data set 1LF correspond to the genotyping data of a single large family. The 1974 offsprings of the ♀77111 × ♂2388 cross were genotyped along with five technical replicates of six individuals and the parents 77111 and 2388 were genotyped 12 times each on separate arrays.

2.4.2. Copy number inference

X/Y intensities (corresponding to A/B alleles' probes, respectively) were normalized using Illumina proprietary software Genome Studio V2011.1 (Illumina, San Diego, California). The normalized signal intensity data were then exported for copy number inference using the algorithms PlatinumCNV (Kumasaka *et al.*, 2011) and GStream (Alonso *et al.*, 2013). Both algorithms were used with default parameters except for the call rate (in PlatinumCNV) were a conservative threshold of 0.999 was used instead of 0.99 (default value).

A two-steps approach was applied for quality control. First, CNV calls displaying reproducibility between technical replicates below 95% were excluded. Second, the copy numbers inferred for each individual at each locus by both algorithms were compared. CNV calls showing a consistency between the two algorithms below 95% were excluded.

For the two data sets analyzed, on average 68% of the initially called CNVs were retained for further analysis.

2.4.3. CNVs validation with real-time qPCR

To validate the discovered CNVs, quantitative real time PCR was performed for 15 genes (10% of the discovered CNV set) displaying copy number variations (three homozygous deletions, eight heterozygous deletions and four copy number gains). A detailed description of the quantification procedure with qPCR is provided in the Supplemental File S2.

Briefly, qPCR data were imported in the software REST 2009 (Qiagen, Hilden, Germany) (Pfaffl *et al.*, 2002) for analysis after quality control. Copy number ratios were calculated using the *BT102965* gene as reference, the parent of each sample as calibrator and an efficiency correction for each reaction. A randomization test (with 10,000 iterations and a significance level $\alpha = 0.05$) was used to identify significant differences in copy numbers between samples (22 to 44 samples for each gene).

2.4.4. Pedigree reconstruction

Forty-three pedigrees were selected representing the 54 full-sib families genotyped in the 54F data set. Thirty-three of these pedigrees encompassed half-sib families sharing a common parent (two to five families per pedigree). The remaining 10 pedigrees involved two to four unrelated families each. Pedigree reconstructions, based on Allele Specific Copy Numbers (ASCN) from 23 (inherited CNVs only) and 79 (inherited and *de novo* CNVs) loci, were performed separately using the maximum-likelihood method implemented in the software Colony 2.0 (Wang and Santure, 2009; Jones and Wang, 2010; Wang, 2013). The reconstruction of each pedigree was performed in three independent runs (with different seeds to start each run) in order to check the convergence of runs toward the same optimal solution. Medium length runs with allele frequency update and no sibship prior were used. The optimal pedigree configuration was identified using the full-likelihood method. The accuracy of pedigree reconstruction using ASCN genotypes was estimated through the parameters P(FS|FS), P(HS|HS), P(UR|UR), P(PO|PO) defined in Wang and Santure (2009).

2.4.5. Statistical analyses

The distribution of mutation rates (μ_{ls}) for different genes was characterized using two approaches i) computation of the Gaussian Kernel Density (GKD) using the function *density* in R (R Core Team, 2016) and ii) fitting of a Gaussian Mixture Model (GMM) using the package *mclust* in R (R Core Team, 2016). Since both approaches provided similar results, we chose to present only the GKD distributions in this paper.

The function *cor.test* in R (R Core Team, 2016) was used to calculate the one-tailed Pearson correlation coefficient (Cor) between the mutation rate (μ_{ls}) and the average gene expression level, and associated p-values.

2.5. Results

We detected copy number variations (CNVs) in the *P. glauca* gene space by using SNP-array raw intensity data for 14 058 genes in 3663 individuals (for details on the CNV calls, see methods). We estimated and characterized the *de novo* mutation rates of CNVs by analyzing 55 two-generation families (54 small families in the 54F data set and one large family in the 1LF data set). Hereafter, we consider a copy number (CN) of two as the normal state for a gene (*P. glauca* being a diploid organism) and variants (also called non-two-copy genotypes) as homozygous deletions HoD (CN = 0), heterozygous deletions HeD (CN = 1) or copy number gains CNG (CN = 3 or 4).

2.5.1. Detection and validation of genic CNVs in pedigree populations

We identified CNVs affecting 143 different genes among individuals (Table 2.1). The genic CNVs detected in each data set represent a small proportion of the 14 058 genes inspected (0.5% on average). Most of the variants (90%) are CN losses (homozygous and/or heterozygous deletions) and only 10% are CN gains (Table 2.1). No two-way (tri-allelic) CNVs were detected.

Table 2.1: Detected CNVs.

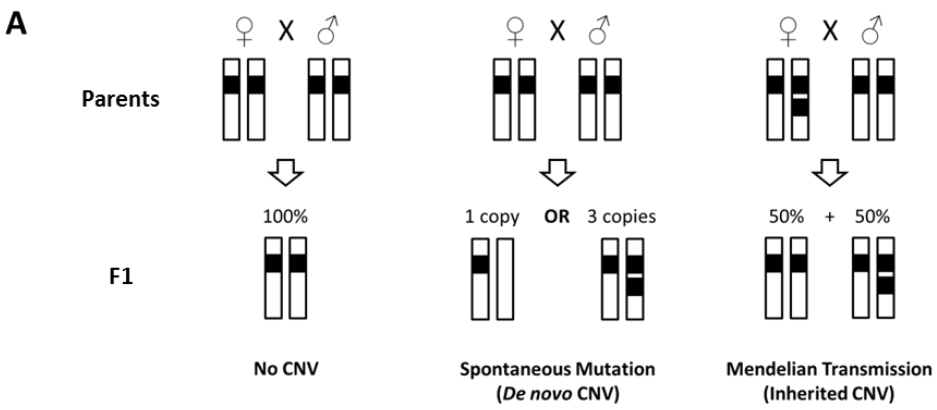
Data set	# individuals	CN Loss	CN Gain	Total	% targeted genes
54F	1687	79	3	82	0.6 %
1LF	1976	50	11	61	0.4 %
Total	3663	129	14	143	1.0 %

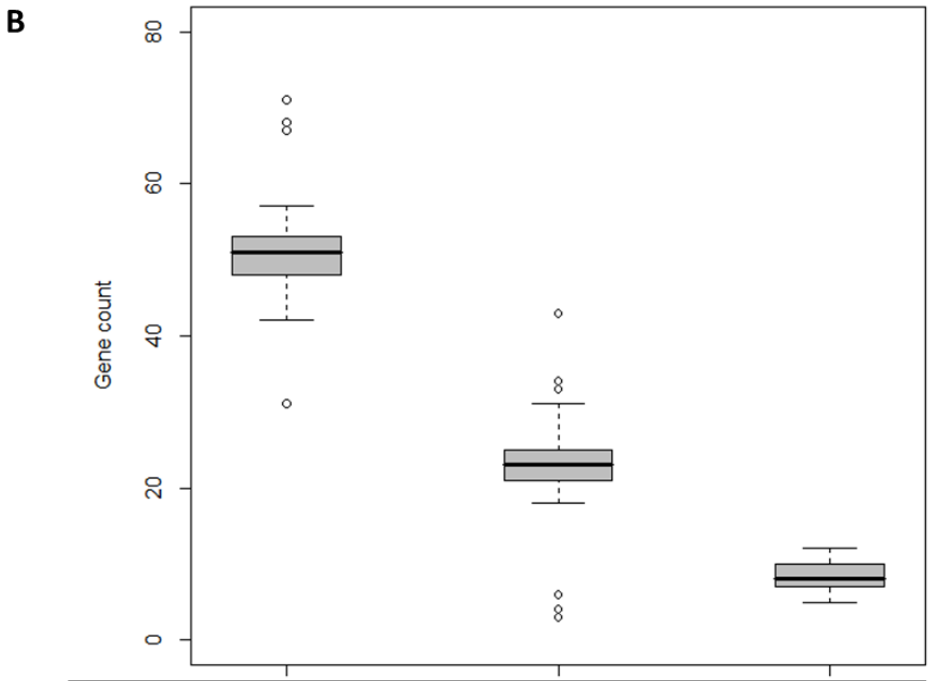
CN: copy number.

We validated the CNV calls using quantitative real time PCR as independent technique. Fourteen out of the 15 tested genes (Supplemental Table S2.9.2) displayed CNVs with both techniques and the estimated False Discovery Rate (FDR) was 6.6%. Careful examination of the gene *BT102213* displaying discrepancies between the two technologies showed that while the qPCR primers target a conserved region of the gene, the array probe is located in an LRR1 domain that can be present in one or two copies in different variants of the gene. We were unable to design a probe and primers that target the same region of the gene because of the different technical requirements of the genotyping array and the qPCR. The genotyping accuracy assessed through qPCR was 80% on average but depended on the CNV class. It was high for deletions, i.e. 85% and 87% for heterozygous and homozygous deletions, respectively, and it was low for CN gains (63%) mainly due to a lack of sensitivity of the SNP-array technology for gains (the median sensitivity for CN gain is 20%).

2.5.2. Classification of CNVs as inherited or *de novo*

CNVs were classified into two categories i) inherited CNVs and ii) *de novo* CNVs that reflect the source of the CN variants observed in the offspring generation (Figure 2.1-A). Inherited CNVs are observed when a non-two-copy genotype is detected in at least one of the parents and among the offspring. *De novo* CNVs on the other hand, are observed when both parents have a two-copy genotype and a non-two-copy genotype is detected among the offspring, which is presumed to result from a germline or somatic mutation event (loss or gain of a copy). For the 54F data set, 23 (28%) of the identified CNVs were transmitted from parents to their offspring and 59 (72%) of the CNVs were detected as *de novo* events. Each of the families displayed eight inherited CNVs and 23 *de novo* CNVs on average. The remaining genes (51) were maintained at two copies for all the family members (Figure 2.1-B). The narrow whisker-boxes in figure 2.1-B show consistent proportions of inherited versus *de novo* CNVs within the different families. A similar profile was observed in the 1LF data set where seven inherited CNVs and 54 *de novo* CNVs were detected (Figure 2.1-C). This observation indicated that the estimates obtained from small families were not considerably biased. The larger number of *de novo* CNVs identified in the large family is to be expected due to the very large sample size.





C

Data set	No CNV	*De novo* CNV	Inherited CNV
54F (Median)	51	23	8
1LF	-	54	7

Figure 2.1: CNV classification. Definition of CNV categories according to their status in the parental and F1 generations (A). CNV genes distribution within the three CNV categories for 54 full-sib families (B). Number of *de novo* and inherited CNVs per family (C).

We proceeded to the reconstruction of two-generations pedigrees from the 54F data set using a maximum likelihood approach based on allele specific copy numbers (ASCN). The pedigrees reconstruction using the 23 inherited CNV genes only was achieved with an average accuracy ranging from 91.7 to 95.5% depending on the nature of the relation between individuals (Figure 2.2). The proportion of dyads correctly inferred was 91.7, 95.5, 92 and 94% for full-sib, half-sib, unrelated individuals and parent-offspring dyads, respectively. In an independent simulation, we used both inherited and *de novo* CNV

genotypes for the reconstruction of the same pedigrees. As might be expected, this decreased the accuracy of full-sib, half-sib and parent-offspring dyads inference and increased the inference accuracy for dyads of unrelated individuals (Figure 2.2). This result highlights the quality of CNV genotyping using raw data from SNP-arrays and the proper classification of the observed CNVs into inherited and *de novo* CNVs that ensued.

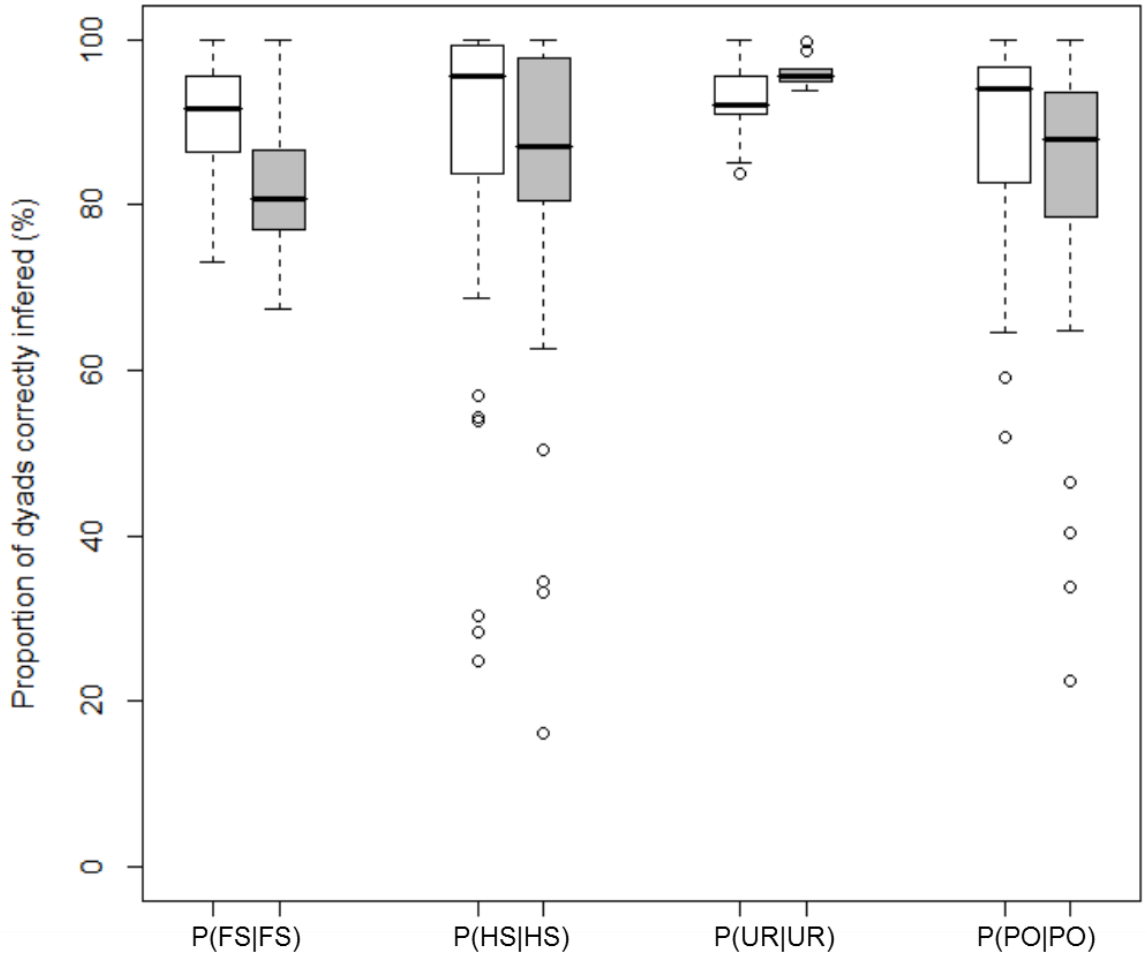


Figure 2.2: Pedigrees reconstructions from CNV data. The accuracy of pedigrees reconstruction using allele specific copy number (ASCN) genotypes for inherited CNVs only (white boxes) and for inherited and *de novo* CNVs (grey boxes) is estimated with four parameters: the proportion of full-sib [P(FS|FS)], half-sib [P(HS|HS)], unrelated [P(UR|UR)] and parent-offspring [P(PO|PO)] dyads correctly inferred.

2.5.3. High rates of *de novo* copy number mutations

We analyzed the 3624 trios (corresponding to 7248 meiotic generations) in 54F and 1LF data sets to estimate the rates of *de novo* copy number changes from parent to offspring. The locus specific mutation rate μ_{ls} at the 113 genes displaying *de novo* CNVs ranged

from 2.6 x 10^{-4} to 9.3 x 10^{-2} mutation per generation which is nearly two orders of magnitude broader than the range observed for mammals using the same experimental approach (Table 2.2). Assuming that *de novo* CNVs have an equal chance of occurring at any location across the gene space, we estimate the mutation rate μ_{cg} is 3 x 10^{-5} mutation per gene per generation based on the data obtained for the 14 058 genes targeted in our study. This estimate of the cross-genome mutation rate μ_{cg} is one to two orders of magnitude higher than that observed for unicellular and multicellular eukaryotes (Table 2.2).

Considering the 14 058 genes examined here, we found that individuals in 54F and 1LF data sets have inherited non-two-copy number in five genes on average and had 1/6 chance of harboring one to two additional gene(s) with non-two-copy number resulting from *de novo* mutations. Using a Poisson distribution to predict the number of *de novo* events occurring in the whole gene space (37491 to 56064 genes (De La Torre *et al.*, 2014; Warren *et al.*, 2015)) we estimate a genomic mutation rate u_g of 1.4 ± 0.36 per haploid genome per generation which is higher than what was observed in unicellular and multicellular eukaryotes (Table 2.2) with *P. glauca* u_g 3 and 23 times higher than *H. sapiens* and *A. thaliana* u_g respectively.

2.5.4. Variable CN mutation rates between genes and for different CNV classes

A closer inspection of the mutation rates for different genes revealed a bimodal distribution (mode 1 with $\mu_{ls} < 10^{-2}$ and mode 2 with $\mu_{ls} > 10^{-2}$ mutation per generation) (Figure 2.3-A). The μ_{ls} estimates from the large family were more widely spread (Figure 2.3-B) while estimates from the 54 smaller families covered a narrower range (Figure 2.3-C). This trend was expected given that the 1LF data set contained around 2000 trees, which should facilitate the detection of rare and recurrent events while the 54F data set may reveal mutation events that are common in the population. Taken together, the estimates from both data sets should provide a more complete picture of the spontaneous mutation dynamics in the *P. glauca* gene space.

The *de novo* mutation rates varied for different CNV classes. For copy number gains and heterozygous deletions, the mutation rates were mostly in the range of mode 1 with only 17 and 35% of the genes, respectively, in the range of mode 2 (Figure 2.3-E,F). On the other hand, mutation rates for homozygous deletions were restricted to mode 1 (Figure 2.3-D).

Table 2.2: Copy number mutation rate estimates in *Picea glauca* and for other eukaryotes.

Species	Mutation rate	Reference
	μ_{ls} (mutation per generation)	
P. glauca	$2.6 \times 10^{-4} – 9.3 \times 10^{-2}$	This Study
H. sapiens	$6.5 \times 10^{-3} – 1.2 \times 10^{-2}$	Itsara *et al.,* 2010
M. musculus	$3.6 \times 10^{-3} – 1.1 \times 10^{-2}$	Egan *et al.*, 2007
	μ_{cg} (mutation per gene per generation)	
P. glauca	3.0×10^{-5}	This Study
S. cerevisiae	2.1×10^{-6}	Lynch *et al.*, 2008
M. musculus	1.2×10^{-6}	Egan *et al.*, 2007
D. melanogaster	9.4×10^{-7}	Schrider *et al.*, 2013
C. elegans	2.2×10^{-7}	Lipinski *et al.*, 2011
	u_g (mutation per haploid genome per generation)	
P. glauca	1.4	This Study
H. sapiens	0.4	Sung *et al.*, 2016
M. musculus	0.08	Sung *et al.*, 2016
A. thaliana	0.06	Sung *et al.*, 2016
C. elegans	0.04	Sung *et al.*, 2016
D. melanogaster	0.04	Sung *et al.*, 2016
S. cerevisiae	0.001	Sung *et al.*, 2016

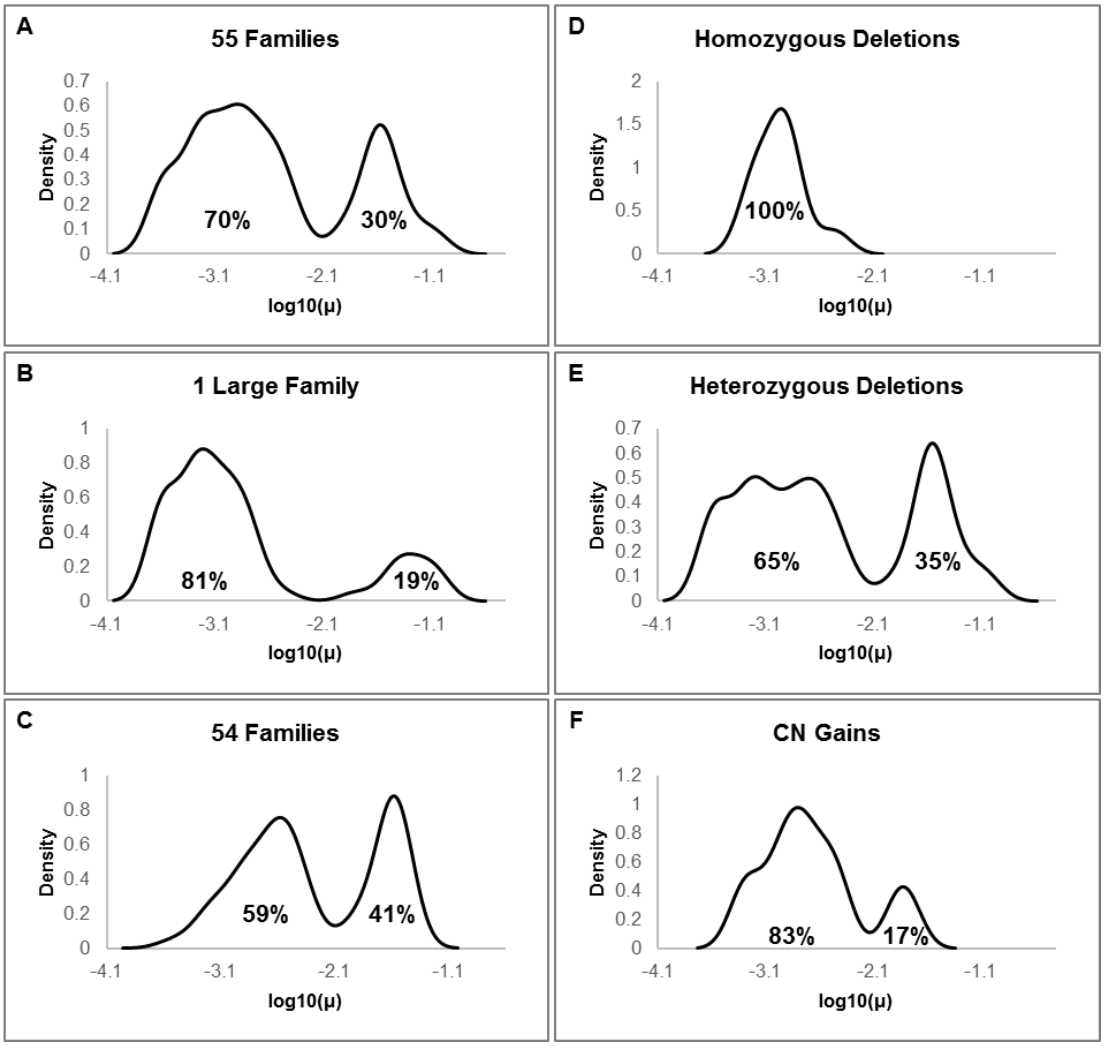


Figure 2.3: Spontaneous mutation rate distribution. The bimodal distribution of μ_{ls} for all of the families (A), the large family data set (B) and the 54 small families' data set (C). The distribution of μ_{ls} for different CNV classes: homozygous deletions (D), heterozygous deletions (E) and copy number gains (F). The proportion of CNV genes is shown for each mode (%).

2.5.5. Allele specific CN mutation rates

We estimated that allele specific mutation rates μ_{AS} for CNVs were an order of magnitude lower than μ_{AS} for Single Nucleotide Variations SNVs (Figure 2.4-A). This observation was based on the analysis of seven genes for which crosses between two homozygote parents allowed us to estimate the μ_{AS} for CNVs and SNVs in the same individuals. Figure 2.4-B also shows that differences between the mutation rates of two alleles of the same gene $\Delta\mu_{AS}$ can be as high as 10^{-2} mutation per generation for CNVs and SNVs.

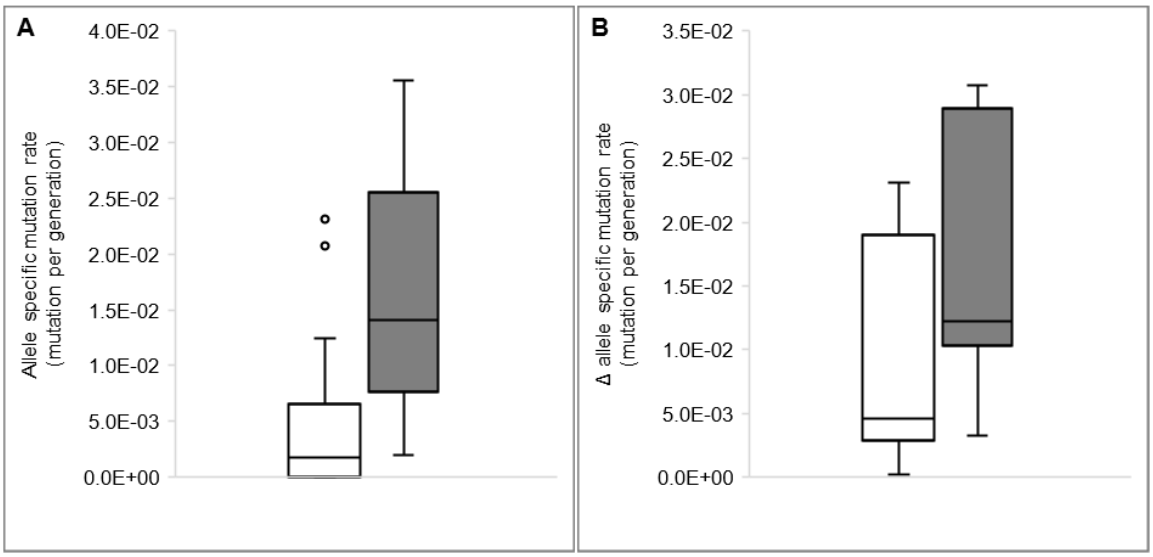


Figure 2.4: Allele specific mutation rates. Range of the allele specific mutation rates μ_{AS} for the CNVs (white box) and SNVs (grey box) of the seven loci for which homozygote individuals were cross-bred (A). Differences between the mutation rates of the two alleles on a locus by locus basis ($\Delta\mu_{AS}$) for CNVs (white box) and SNVs (grey box) (B).

2.5.6. Relationship between CN mutation rates and gene expression

Expression levels for the 113 genes displaying *de novo* CNVs are available for eight *P. glauca* tissues (Raherison *et al.*, 2012) and were analyzed here. We checked if there was an association between the locus specific mutation rate μ_{ls} and transcript accumulation levels. In both data sets (54F and 1LF), we show that the μ_{ls} of mode 2 genes was negatively correlated with gene expression level (Figure 2.5-A,B) indicating that highly and broadly expressed genes have a lower μ_{ls}. This observation does not hold true for mode 1 genes as there was no correlation between μ_{ls} and gene expression levels (Figure 2.5-C, D).

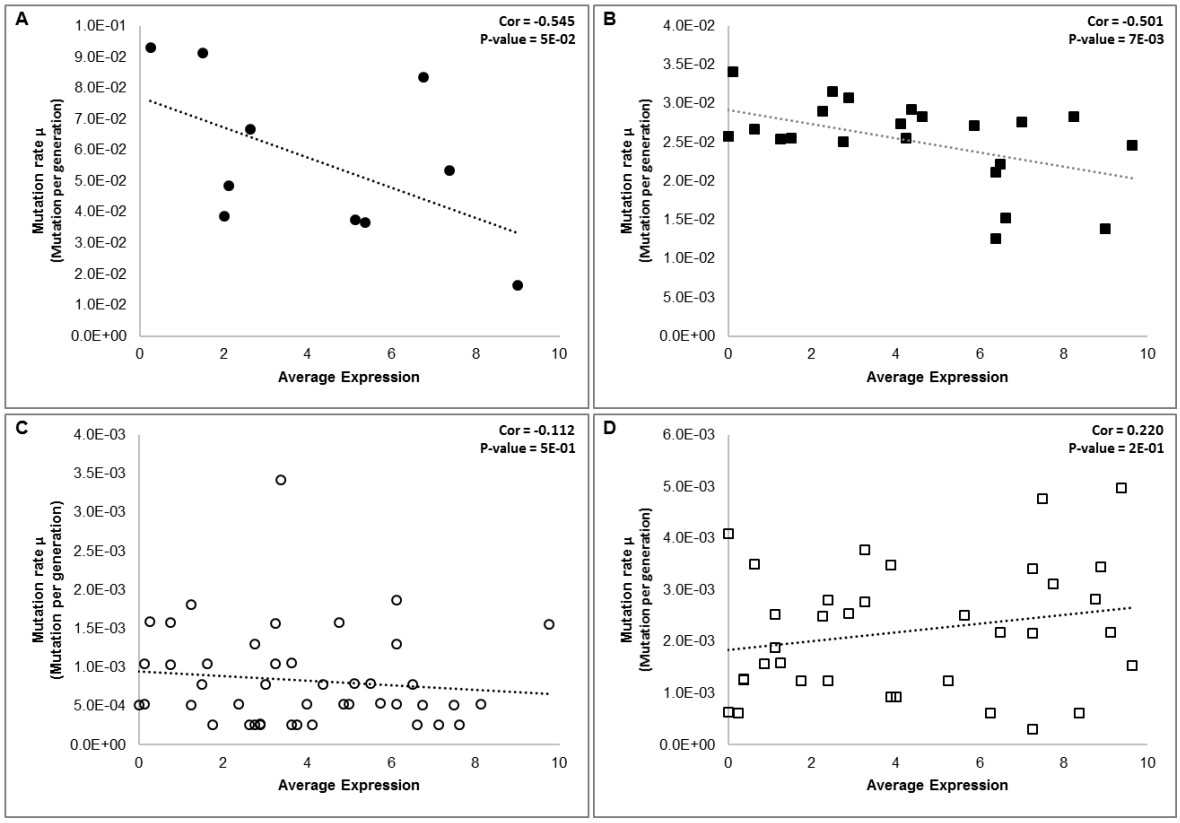


Figure 2.5: Correlation between mutation rates and gene expression. Negative correlations were identified between the mutation rates and expression levels for mode 2 genes ($\mu_{ls} > 10^{-2}$) (A, B) and not for mode 1 genes ($\mu_{ls} < 10^{-2}$) (C, D) both for the data sets 1LF (A, C) and 54F (B, D). Average expression was calculated based on the relative transcript accumulation class (1 being lowest and 10 being highest) reported in Raherison and colleagues (2012).

2.6. Discussion

In this work, we applied a cross-sample strategy (to call CNVs) coupled with a family-based approach (to differentiate inherited CN variants from *de novo* CNVs) that allowed us to detect genic CNVs and directly estimate their mutation rate in the conifer tree *Picea glauca*. The use of large data sets (14 058 genes genotyped in 3624 trios) and stringent criteria for data quality control and CNV inference contributed to the identification of a high-quality CNV set (as shown by the independent validation with qPCR and the accurate reconstruction of pedigrees) and the reliable characterization of the *de novo* mutation rate spectrum across the gene space. The approach is transferable to other species, including non-model organisms and species with large and complex genomes or long generation times.

2.6.1. CNVs in the *P. glauca* gene space

CNVs can affect gene structure and expression, and impact downstream phenotypes, fitness and reproductive success (Tang and Amon, 2013). Consequently, many CNVs are expected to be deleterious and under strong purifying selection (Schrider *et al.*, 2013). Our data showed that only 0.5 to 1% of the 14 058 targeted genes displayed CNVs even though thousands of individuals were examined. These data allow us to predict that each white spruce individual will have 17 to 20 genes (0.04% of the gene space) with non-two-copy genotypes on average. These observations agree with the afore-mentioned hypothesis of purifying selection. The majority of CNVs identified in this study were copy number losses (90%), and bi-allelic variations were more abundant by far than multi-allelic variations, as previously observed in other species including human (Kato *et al.*, 2010; Mills *et al.*, 2011), stickleback fish (Chain *et al.*, 2014), bovine (Cicconardi *et al.*, 2013) and plants (Swanson-Wagner *et al.*, 2010; Yu *et al.*, 2011; McHale *et al.*, 2012), to cite a few. The copy number losses were nine times more abundant than copy number gains in our data set. This estimate is similar to ratios of seven and height reported for maize by Swanson-Wagner and collegues (2010) and Liu and collegues (2012), respectively. Rice (Yu *et al.*, 2011) and soybean (McHale *et al.*, 2012) also showed a bias toward copy number losses. By comparison, the ratio of deletions to duplications in human is considerably lower (two to three times) (Chen *et al.*, 2010) than reported in plants. The detection of more abundant copy number losses can be explained from technical and/or biological perspectives. For large copy numbers, signal intensities are more noisy and the relationship between copy number and signal intensity is not linear (Cantsilieris *et al.*, 2013), which decreases the likelihood of detecting copy number gains. More copies also increases the potential for mismatches affecting probe and primer hybridization due to sequence polymorphisms. From a biological perspective, several molecular mechanisms involved in CNV formation favor sequence losses over duplications (Chen *et al.*, 2010), particularly non-allelic homologous recombination (NAHR), which was shown to be the dominant mechanism for CNV formation in *A. thaliana* (Lu *et al.*, 2012). Another, non-exclusive, hypothesis is that losses are the result of the segregation of non-allelic homologs. In maize, single-copy homologous sequences located in non-allelic positions in the genomes of two crossed hemizygous parents, segregate in the offspring as zero, one or two copies genotypes (Liu *et al.*, 2012), which is consistent with our observations for most of the genes in the present set of *P. glauca* families.

2.6.2. Copy number mutational features

The mutation rate determines the rate of generation of new variants necessary for species evolution and adaptation. To date, estimates of the mutation rate for CNVs have been limited to a few model organisms and suffer from biases related to the targeted region in the genome, the individuals sampled (families or populations), and the estimation approach (Itsara *et al.*, 2010; Katju and Bergthorsson, 2013). Therefore, we expect that a better characterization of the spectrum of CN mutation rates in different species will provide new insights into the evolution process.

Here we show that CN mutation rates in *P. glauca* cover a wide range (at least three orders of magnitude) and can reach as high as 10^{-2} mutation per generation for some genes. High *de novo* mutation rates (locus specific and genomic estimates) are expected in plants and perhaps even more in trees. In plants, many cell divisions occur during a single generation, which increase the probability of mutation events during DNA replication and repair (Petit and Hampe, 2006; Scofield and Schultz, 2006). More importantly, there is no clear separation between germline and soma which both contribute to the estimated mutation rates in plants. Somatic mutations are particularly high in plants and are frequently generated under stress; for example, a two-fold increase in μ was found for stress induced mutations (Debolt, 2010; Jiang *et al.*, 2014). Mutations generated during plant growth, accumulate in the meristem and are transmitted to the gametes. This phenomenon may be more pronounced in perennials (like *P. glauca*) compared to annuals due to their longevity. In conifers, the egg cell for fertilization differentiates from the megagametophyte through about 11 rounds of mitotic division from the megaspore and the sperm nucleus is formed *via* five mitotic divisions from the microspore (Williams, 2009). These rounds of division between meiosis generating the megaspore or the microspore, and the fecundation increase the chances of spontaneous mutations.

The high mutation rates reported here can also be explained by two specific features of the genome, common to many plants including conifers such as *P. glauca*: a high A-T content (62%) and an abundance of repeated sequences (70%), particularly transposons (Birol *et al.*, 2013; Nystedt *et al.*, 2013). In other species, transposons were found to actively promote high mutation rates (Woodruff *et al.*, 1984; Bégin and Schoen, 2006; Lu *et al.*, 2012; Pinosio *et al.*, 2016), and the examination of deletions breakpoints identified the presence of A-T rich sequences in the vicinity of these variations (Chen *et al.*, 2010).

Copy number mutation rates in the *P. glauca* gene space followed a bimodal distribution with the majority of genes (70%) subject to low mutation rates and the rest (30%) associated with high mutation rates (above 10^{-2} mutation per generation). This spectrum of mutation rates could reflect local differences in the genome or differences in the selection pressure. Local features of genome architecture such as base composition, short repeats density, mobile elements, recombination rates and methylation can influence the frequency at witch mutations are generated (reviewed in Baer *et al.*, 2007). Also, genes involved in basic metabolic functions are expected to be under strong selection pressure. On the other hand, redundant genes, genes associated with compensation mechanisms and genes involved in adaptation are likely to be under relaxed selection and tolerate more frequent mutations (Tang and Amon, 2013). In this work, we have shown that homozygous deletions (complete gene losses) are rare (confined in mode 1 only). Since the complete loss of a gene is expected to be more deleterious than a partial loss (heterozygous deletion) or a duplication, we can presume that homozygous deletions are under strong purifying selection.

We found that CNVs have lower mutation rates (an order of magnitude on average) than SNVs for the same genes. In *A. thaliana* and human, the mutation rate for SNVs is one and three order(s) of magnitude higher than for CNVs respectively (Itsara *et al.*, 2010; Ossowski *et al.*, 2010). The effects of CNVs on gene structure and/or expression are likely to be more detrimental on average than those of single nucleotide mutations (for example synonymous SNVs are mostly slightly deleterious or neutral). Hence, a strong selection pressure is expected to drive CN mutation rates to lower levels (Schrider *et al.*, 2013).

We found a negative correlation between mutation rate and average gene expression level only for the genes at the high end of the mutation rate spectrum (mode 2). This was taken as an indication that selection pressure maintains the mutation rate at a lower level for highly and broadly expressed genes, which are presumed to be more essential for cellular function. In eukaryotes, the relationship between gene transcription levels and the mutation rates is still not clear. The transcription-coupled repair hypothesis TCRH (Hendriks *et al.*, 2010; Fidantsef and Britt, 2011) suggests that highly expressed genes should be associated with lower mutation rates based on the observation that DNA repair is more efficient for actively expressed genes and for the transcribed strand rather than the non-transcribed strand. On the other hand, the transcription-associated mutagenesis hypothesis TAMH (Park *et al.*, 2012; Sollier *et al.*, 2014; Heinäniemi *et al.*, 2016) proposes

that transcription promotes spontaneous mutations based on the observation that highly expressed genes are more frequently associated with mutations (SNVs, intra-genic deletions or double strand breakages). Since we observed a negative correlation only in mode 2, neither the TCRH nor the TAMH hypotheses explain our data entirely. Alternatively, Lynch (2011) proposed the drift-barrier hypothesis DBH, which stipulates that selection will drive μ down from high mutation rates to lower levels while at the lower bound of observed mutation rates, genetic drift will circumvent selection and the mutation rate will evolve randomly toward higher or lower values. The DBH fits our data and explains the observed relationship between CN mutation rate and gene expression in both mode 1 and mode 2.

Mutations are the source of variations that fuel the evolutionary process but because mutation events are rare, their contribution to the changes of allele frequencies and to the determination of evolutionary outcomes has been underestimated. Yampolsky and Stoltzfus (2001) proposed an origin-fixation model in which mutation rates can be an orienting factor in evolution. In this model the fate of two alleles is not determined by their relative effects on fitness alone but also depends on the order of the appearance of alleles, their respective mutation rates and the effective population size (Ne). Here we report empirical estimates of the differences between the mutation rates associated with two alleles of the same locus. A pair of alleles may have mutation rates that differ by as much as an order of magnitude in favor of one allele relatively to the other, which will have a large impact on the chances of fixation of the two alleles. The mutation rate also determines how long an allele will remain in the genome. Alleles with lower deletion rates will be retained for longer increasing their chance of accumulating more mutations or being converted to alternative allelic forms.

2.6.3. Evolutionary consequences of high and variable CN mutation rates in *P. glauca*

In the present study, we primarily detected copy number losses and estimated that CNVs affect a small proportion of the gene space, which supports the hypothesis that CNVs are mainly deleterious and should be under strong purifying selection. *De novo* CNV formation occurs at a lower rate than SNV (expected to be less deleterious than structural variations on average), particularly when compared to complete gene losses (homozygous deletions). Still, copy number mutation rates can reach high levels for *P. glauca* and the

genomic mutation rate we report here is higher than in other organisms examined to date. This high mutation rate imposes a mutational load that seems to be tolerated because frequent mutations would fuel the standing genetic variation of the population, contributing to adaptation to environmental changes, which is a well-known feature for perennial trees (Hamrick, 2004; Petit *et al.*, 2004). New mutations are less harmful and have lower impact on phenotype and fitness in diploid organisms such as *P. glauca* (sheltered mutational load) which may explain in part why it can tolerate such high mutation rates. Also, the pattern of stem cell divisions in the meristem described by Burian and colleagues (2016) seems to i) slowdown the mutational meltdown resulting from the accumulation of somatic mutations, ii) reduce the chances of transmission of new mutations to the next generation and iii) maximize the genetic heterogeneity within a tree (in the form of nested sectors) to allow for survival and rapid adaptation to a changing environment.

Our data further show that alleles of the same gene can have considerably different mutation rates and consequently the fate of alleles can be determined based on the order and rate of their generation and maintenance in the genome in addition to their respective effects on fitness. From this perspective, spontaneous mutations are not only a source of new variants but play also a role as an orienting factor that can determine the fate of alleles.

2.7. Acknowledgments

We thank Sylvie Blais (Canada Research Chair in Forest Genomics at Université Laval, Quebec, Canada) for help with data treatment. We also thank Julien Prunier and Mebarek Lamara (Université Laval) for help with the collection of samples. We thank the Ministry of Forests, Wildlife and Parcs (Quebec, Canada) for the maintenance of the experimental plantations where the samples were collected. We are thankful to Cyril Van Ghelder from the University of Oxford (Oxford, United Kingdom) for his help with the *in silico* prediction of the domains of the gene *BT102213*.

Funding was received from Genome Canada and Genome Quebec for the SMarTForests project. AS received financial support from the SMarTforests project and from Quebec Ministry of Economic Development, Innovation and Exports through the GenAC project.

2.8. References

Alonso A, Marsal S, Tortosa R, Canela-Xandri O, Julià A (2013). GStream: Improving SNP and CNV coverage on genome-wide association studies. *PLoS One* **8**: e68822.

Baer CF, Miyamoto MM, Denver DR (2007). Mutation rate variation in multicellular eukaryotes: Causes and consequences. *Nat Rev Genet* **8**: 619–631.

Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J (2014). Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genomics* **15**: 1048.

Bégin M, Schoen DJ (2006). Low impact of germline transposition on the rate of mildly deleterious mutation in *Caenorhabditis elegans*. *Genetics* **174**: 2129–2136.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA *et al.* (2013). Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**: 1492–1497.

Blackburn A, Göring HH, Dean A, Carless MA, Dyer T, Kumar S *et al.* (2013). Utilizing extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. *Eur J Hum Genet* **21**: 404–409.

Burian A, Barbier de Reuille P, Kuhlemeier C (2016). Patterns of stem cell divisions contribute to plant longevity. *Curr Biol* **26**: 1385–1394.

Cantsilieris S, Baird PN, White SJ (2013). Molecular methods for genotyping complex copy number polymorphisms. *Genomics* **101**: 86–93.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M *et al.* (2014). Extensive copy-number variation of young genes across stickleback populations. *PLoS Genet* **10**: e1004830.

Chen JM, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP (2010). Genomic rearrangements in inherited disease and cancer. *Semin Cancer Biol* **20**: 222–233.

Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P *et al.* (2013). Massive screening of copy number population-scale variation in *Bos taurus* genome. *BMC Genomics* **14**: 124.

Debolt S (2010). Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* **2**: 441–453.

Egan CM, Sridhar S, Wigler M, Hall IM (2007). Recurrent DNA copy number variation in the laboratory mouse. *Nat Genet* **39**: 1384–1389.

Fidantsef AL, Britt AB (2011). Preferential repair of the transcribed DNA strand in plants. *Front Plant Sci* **2**: 105.

Fu W, Zhang F, Wang Y, Gu X, Jin L (2010). Identification of copy number variation hotspots in human populations. *Am J Hum Genet* **87**: 494–504.

Hamrick JL (2004). Response of forest trees to global environmental changes. *For Ecol Manage* **197**: 323–335.

Heinäniemi M, Vuorenmaa T, Teppo S, Kaikkonen MU, Bouvy-Liivrand M, Mehtonen J *et al.* (2016). Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots. *eLife* **5**: e13087.

Hendriks G, Calléja F, Besaratinia A, Vrieling H, Pfeifer GP, Mullenders LHF *et al.* (2010). Transcription-dependent cytosine deamination is a novel mechanism in ultraviolet light-induced mutagenesis. *Curr Biol* **20**: 170–175.

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ *et al.* (2010). *De novo* rates and selection of large copy number variation. *Genome Res* **20**: 1469–1481.

Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP (2014). Environmentally responsive genome-wide accumulation of *de novo Arabidopsis thaliana* mutations and epimutations. *Genome Res* **24**: 1821–1829.

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, *et al.* (2012). Genome-wide genetic changes during modern breeding of maize. *Nat Genet* **44**: 812–815.

Jones OR, Wang J (2010). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Mol Ecol Resour* **10**: 551–555.

Katju V, Bergthorsson U (2013). Copy-number changes in evolution: Rates, fitness effects and adaptive significance. *Front Genet* **4**: 273.

Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R, Shapero MH *et al.* (2010). Population-genetic nature of copy number variations in the human genome. *Hum Mol Genet* **19**: 761–773.

Kumasaka N, Fujisawa H, Hosono N, Okada Y, Takahashi A, Nakamura Y *et al.* (2011). PlatinumCNV: a Bayesian Gaussian mixture model for genotyping copy number polymorphisms using SNP array signal intensity data. *Genet Epidemiol* **35**: 831–844.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM *et al.* (2014). Insights into conifer giga-genomes. *Plant Physiol* **166**: 1724–1732.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U (2011). High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr Biol* **21**: 306–310.

Liu S, Ying K, Yeh CT, Yang J, Swanson-Wagner R, Wu W *et al.* (2012). Changes in genome content generated via segregation of non-allelic homologs. *Plant J* **72**: 390–399.

Lu P, Han X, Qi J, Yang J, Wijeratne AJ, Li T *et al.* (2012). Analysis of *Arabidopsis* genome-wide variations before and after meiosis and meiotic recombination by resequencing *Landsberg erecta* and all four products of a single meiosis. *Genome Res* **22**: 508–518.

Lynch M (2011). The lower bound to the evolution of mutation rates. *Genome Biol Evol* **3**: 1107–1118.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, *et al.* (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* **105**: 9272–9277.

Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H *et al.* (2007). Breaking the waves: Improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* **8**: R228.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL *et al.* (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol* **159**: 1295–1308.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C *et al.* (2011). Mapping copy number variation by population scale genome sequencing. *Nature* **470**: 59–65.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG *et al.* (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG *et al.* (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**: 92–94.

Park C, Qian W, Zhang J (2012). Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep* **13**: 1123–1129.

Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B *et al.* (2013). Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Mol Ecol Resour* **13**: 324–336.

Petit RJ, Bialozyt R, Garnier-Géré P, Hampe A (2004). Ecology and genetics of tree invasions: From recent introductions to Quaternary migrations. *For Ecol Manage* **197**: 117–137.

Petit RJ, Hampe A (2006). Some evolutionary consequences of being a tree. *Annu Rev Ecol Evol Syst* **37**: 187–214.

Pfaffl MW, Horgan GW, Dempfle L (2002). Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res* **30**: e36.

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC *et al.* (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Mol Biol Evol* **33**: 2706–2719.

R Core Team (2016). R: A Language and Environment for Statistical Computing.

Raherison E, Rigault P, Caron S, Poulin PL, Boyle B, Verta JP *et al.* (2012). Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. *BMC Genomics* **13**: 434.

Saxena RK, Edwards D, Varshney RK (2014). Structural variations in plant genomes. *Brief Funct Genomics* **13**: 296–307.

Schrider DR, Houle D, Lynch M, Hahn MW (2013). Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics* **194**: 937–954.

Scofield DG, Schultz ST (2006). Mitosis, stature and evolution of plant mating systems: low-phi and high-phi plants. *Proc R Soc L B Biol Sci* **273**: 275–282.

Sollier J, Stork CT, García-Rubio ML, Paulsen RD, Aguilera A, Cimprich KA (2014). Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Mol Cell* **56**: 777–785.

Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS *et al.* (2016). Evolution of the insertion-deletion mutation rate across the tree of life. *G3 Genes|Genomes|Genetics* **6**: 2583–2591.

Swanson-Wagner R a, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D *et al.* (2010). Pervasive gene content variation and copy number variation in maize and its

undomesticated progenitor. *Genome Res* **20**: 1689–1699.

Tang YC, Amon A (2013). Gene copy-number alterations: A cost-benefit analysis. *Cell* **152**: 394–405.

Wang J (2013). An improvement on the maximum likelihood reconstruction of pedigrees from marker data. *Heredity* **111**: 165–174.

Wang J, Santure AW (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* **181**: 1579–1594.

Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP *et al.* (2015). Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J* **83**: 189–212.

Williams CG (2009). *Conifer Reproductive Biology*, Springer: Heidelberg, London, New York.

Woodruff RC, Thompson JN, Seeger MA, Spivey WE (1984). Variation in spontaneous mutation and repair in natural population lines of *Drosophila melanogaster*. *Heredity* **53**: 223–234.

Yampolsky LY, Stoltzfus A (2001). Bias in the introduction of variation as an orienting factor in evolution. *Evol Dev* **3**: 73–83.

Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen J-Q *et al.* (2015). Parent–progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**: 463–467.

Yu P, Wang C, Xu Q, Feng Y, Yuan X, Yu H *et al.* (2011). Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics* **12**: 372.

Zhang F, Gu W, Hurles ME, Lupski JR (2009). Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**: 451–81.

Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavo E, Braun M, *et al.* (2013). Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res* **23**: 568–579.

Zmienko A, Samelak A, Kozlowski P, Figlerowicz M (2014). Copy number polymorphism in plant genomes. *Theor Appl Genet* **127**: 1–18.

2.9. Supplementary information

2.9.1. Supplemental File S1

The 54 families analyzed in the 54F data set involve 37 parents that contributed as female and/or male in different crosses. The partial diallel crossing scheme used is described in the following table:

Supplemental Table S2.9.1: Partial diallel crossing scheme used to generate the 54 families analyzed in the 54F data set.

Parent contributing as male	Parent contributing as female																																						
	79101	79102	79104	80101	80102	80105	80106	80109	80110	80112	80113	80114	80115	80116	80118	80119	80120	80123	80124	80131	80132	81101	81102	81103	81104	81105	81106	81107	81108	81113	81114	81115	821016	821019	821025	821061	821084
79101												X																										X
79102													X																									
79104															X																							
80101					X	X																															
80102											X																											
80105							X																															
80106	X																														X						
80109									X																													
80110	X									X																												
80112								X								X																						
80113			X			X																																
80114										X																												
80115																					X	X																
80116															X			X							X		X											
80118																														X								
80119															X																							
80120				X					X																													
80123																																						
80124																		X		X																		
80131													X																									
80132																			X	X																		
81101																								X									X					
81102																								X														
81103																					X																	
81104														X																									
81105																														X					X		X	
81106		X																							X										X			
81107																																						
81108														X											X											X			
81113																																	X					X
81114											X																				X							
81115												X																								X			
821016																																						
821019																																						
821025																																						
821061																			X																			
821084																																	X					

The X symbol indicate that the corresponding two parents were crossed and 28 to 32 of their offsprings were used to identify inherited genic CNVs and characterize their transmission patterns.

2.9.2. Supplemental File S2

2.9.2.1. CNVs validation with real-time qPCR (detailed protocol)

Primers for 15 CNV genes and 6 candidate reference genes (Supplemental Table S2.9.2) were designed using Primer3 algorithm (Koressaar and Remm, 2007; Untergasser *et al.*, 2012). Primers properties (self-hybridization, hairpin loop formation and dimers formation) were assessed *in silico* using OligoCalc (Kibbe, 2007). Hybridization temperature (Tm) and DNA concentration were optimized for each qPCR assay. The high-resolution melting curves were inspected for each assay and the amplicons were sequenced to check that each reaction amplify the intended target sequence in the genome.

Six candidate reference genes were selected based on their stable expression in *Picea glauca* (Beaulieu *et al.*, 2013). The software geNorm (Vandesompele *et al.*, 2002) was used to test the copy number stability of the candidate reference genes in our DNA samples, and the gene *BT102965* (coding for a GTP binding elongation factor) was selected as reference gene for the validation of CNV calls using qPCR.

To validate the discovered CNVs, quantitative real-time PCR was performed for 15 genes with 22 to 44 samples for each gene. qPCR reactions were prepared in 384 micro-well plates using epMotion 5075 automated liquid handler (Eppendorf, Hamburg, Germany). Each 15 µl reaction contained 20 ng genomic DNA, 7.5 µl Qiagen Fast Protocol Master Mix (1x) (Qiagen, Hilden, Germany) and forward/reverse primers (300 nM final concentration). qPCR reactions were run in four replicates on a LightCycler480 instrument (Roche Life Science, Penzberg, Germany) and thermal cycling conditions were 95°C for 5 min followed by 50 cycles of 94°C for 10 s and 62°C for 1 min. At the end of the 50th cycle, a high resolution melting curve was generated as follow: 95°C for 1 min, 40°C for 1 min, and a final step of continuous temperature increase from 55°C to 95°C with a 0.02°C/s ramp rate.

The efficiency of each qPCR reaction was estimated using the linear regression method described in (Boyle *et al.*, 2009). The Cp (Crossing point) values and efficiency estimates were imported in the software REST 2009 (Qiagen, Hilden, Germany) (Pfaffl *et al.*, 2002) for further analysis. Copy number ratios were calculated using the *BT102965* gene as reference, the parent of each sample as calibrator and an efficiency correction for each

reaction. A randomization test (with 10,000 iterations and a significance level α = 0.05) was used to identify significant differences in copy numbers between samples.

Supplemental Table S2.9.2: Target and reference genes used for CNV validation.

GenBank Acc	Function	CNV Class	Genotyping Accuracy	Stability Rank
DR591843	Heat shock protein	HoD	93 %	-
BT101196	F-box family protein (MEE66)	HoD	87 %	-
BT109539	Cysteine/Histidine-rich C1 domain family protein	HoD	81 %	-
BT106719	NmrA-like negative transcriptional regulator family protein	HeD	96 %	-
BT101142	Aminophospholipid ATPase 3	HeD	90 %	-
BT110689	Unknown	HeD	88 %	-
BT109608	Chaperone DnaJ-domain superfamily protein	HeD	87 %	-
BT108501	Cellulose synthase 5	HeD	85 %	-
BT119696	ARM repeat superfamily protein	HeD	83 %	-
BT110740	Unknown	HeD	75 %	-
BT105558	RNI-like superfamily protein	HeD	74 %	-
BT115697	Regulatory particle triple-A ATPase 3	CNG	64 %	-
BT107108	Protein kinase	CNG	59 %	-
BT110964	G-protein-coupled receptor	CNG	59 %	-
BT102213	ADR1-like 1	CNG	No CNV	-
BT102965	GTP binding elongation factor	Reference	-	1
BT115988	Ubiquitin-specific protease	Reference	-	1
BT112014	Novel cap-binding protein	Reference	-	3
BT119125	Lipin family protein	Reference	-	4
BT109864	Ubiquitin-conjugating enzyme	Reference	-	5
BT108451	Villin	Reference	-	6

HoD: homozygous deletion; HeD: heterozygous deletion; CNG: copy number gain.

2.9.2.2. References

Beaulieu J, Giguère I, Deslauriers M, Boyle B, MacKay J (2013). Differential gene expression patterns in white spruce newly formed tissue on board the International Space Station. *Adv Sp Res* **52**: 760–772.

Boyle B, Dallaire N, MacKay J (2009). Evaluation of the impact of single nucleotide polymorphisms and primer mismatches on quantitative PCR. *BMC Biotechnol* **9**: 75.

Kibbe WA. (2007). OligoCalc: An online oligonucleotide properties calculator. *Nucleic Acids Res* **35**: 43–46.

Koressaar T, Remm M (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**: 1289–1291.

Pfaffl MW, Horgan GW, Dempfle L (2002). Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res* **30**: e36.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M *et al.* (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res* **40**: e115.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A *et al.* (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**: RESEARCH0034.

Chapter 3: Transmission distortions of genic copy number variants cause significant and complex frequency changes between generations

[Atef Sahli, Jean Bousquet and John MacKay (2017). Transmission distortions of genic copy number variants cause significant and complex frequency changes between generations. Heredity; manuscript to be submitted]

3.1. Abstract

Copy number variations (CNVs) are among the least studied genetic variations despite their abundance and impact on gene structure and function. The transmission of genetic variants from a generation to the next (Mendelian or non-Mendelian inheritance) shapes the genetic variation within species. Transmission distortion systems are still poorly understood particularly for CNVs. Here, we examined the transmission profile of CNVs identified in the gene space of 1650 *Picea glauca* trios and investigated the factors that influence distortion levels. Our findings show that most of the inherited CNVs (70%) are transmitted from the parents in violation of Mendelian expectations. The observed distortion levels vary considerably and are influenced by parental, partner genotype and genetic background effects. We also identified instances where the loss of a gene copy is favored and subject to different types of selection pressures. This study shows that transmission distortions can contribute to considerable and complex frequency changes between generations and have significant evolutionary consequences on the standing genetic variation.

3.2. Résumé

Malgré leur abondance et leur impact sur la structure et le fonctionnement des gènes, les variations de nombre de copies (VNCs) restent parmi les variations génétiques les moins étudiées à ce jour. La transmission des polymorphismes génétiques d'une génération à la suivante (selon une hérédité mendélienne ou pas) contribue à la détermination de la diversité génétique présente chez une espèce. Les systèmes de distorsion des transmissions génétiques sont encore peu caractérisés en particulier pour les VNCs. Dans cette étude, on a examiné le profil de transmission de VNCs identifiées dans l'espace génique de 1650 trios d'arbres appartenant à l'espèce *Picea glauca* et on a étudié les

facteurs qui influencent le niveau de distorsion. Nos résultats montrent que la majorité des VNCs (70%) sont transmis des parents en violation des lois de l'hérédité mendélienne. Les niveaux de distorsion observés varient considérablement et sont influencés par des effets parentaux, de génotype du partenaire et de contexte génétique. On a aussi identifié des situations où la perte de copies d'un gène est favorisée et sujette à différents types de pression de sélection. Cette étude démontre que les distorsions de transmission génétique peuvent contribuer à des changements de fréquences alléliques considérables et complexes entre les générations successives, et qu'elles ont des conséquences évolutives importantes sur la diversité génétique maintenue chez une espèce.

3.3. Introduction

Genetic variants generated through *de novo* mutations or introduced through sexual reproduction are expected to be transmitted randomly to the next generation. Transmission distortion (TD) is the preferential transmission of an allele to the next generation at the expense of alternative alleles. This departure from the Mendelian expectations is observable in the offspring of heterozygous individuals and is often the consequence of disruptive mechanisms operating during the gametic or zygotic stages of development (Huang *et al.*, 2013). TDs are under genetic control (Lyttle, 1991) and are the result of complex mechanisms usually involving a responder locus (target of the distortion) and one or more distorter locus(ci) (linked or unlinked modifiers of the level of distortion) (examples of distortion systems are presented in Didion and collaborators 2015).

TDs can cause a wide range of frequency changes: from mild distortions to complete skew of transmission in favor of one allele (Chevin and Hospital, 2006; Koide *et al.*, 2012). The favored allele will eventually reach fixation while the other allele will be purged from the population unless some balancing force (recombination, mutation, genetic drift) counters the TDs effects, which may lead to the maintenance of the alleles (Polański *et al.*, 1998). TD is an important evolutionary force that shapes the standing genetic variation within populations but is poorly understood due to the lack of appropriate data sets (large and accurate genotyping data sets for trios or pedigrees). With the recent progress of genotyping and sequencing technologies, the dissection of additional distortion systems is now feasible at reasonable cost and will provide new insights into evolution and adaptation.

Copy number variations (CNVs) are among the least studied genetic variations despite their abundance in natural populations (Jakobsson *et al.*, 2008; Kato *et al.*, 2010; Mills *et al.*, 2011; Tan *et al.*, 2012; Blackburn *et al.*, 2013; Chain *et al.*, 2014), the considerable proportion of the genome they affect (12-15% of the human genome; Sebat *et al.*, 2004; Redon *et al.*, 2006) and their impacts on gene function (Korbel *et al.*, 2009; Debolt, 2010; Conrad *et al.*, 2010; Mills *et al.*, 2011), gene expression (Stranger *et al.*, 2007; Schlattl *et al.*, 2011) and downstream phenotypes (McCarroll and Altshuler, 2007; Beckmann *et al.*, 2007). TDs involving CNVs at responder and/or distorter loci have been identified in a few model organisms including the loci *om* and *WSB* in mice (Pardo-Manuel De Villena *et al.*, 2000; Didion *et al.*, 2015) and the *peel-1* locus in *C. elegans* (Seidel *et al.*, 2011). A more comprehensive analysis of the transmission of CNVs and the occurrence of copy number transmission distortions at the genome scale will help to better understand how genetic variations are transmitted and maintained in a population.

In this work, we examined the transmission profile of genic copy number variants in a large set of trios from the conifer tree white spruce (*Picea glauca* [Moench] Voss). Our goals were to i) investigate the frequency of TD occurrence for these genetic variations, ii) estimate the levels of distortions, iii) identify the factors (parental, partner genotype and genetic background effects) that influence the TDs and iv) evaluate the contribution of TDs to frequency changes between generations.

3.4. Material and methods

3.4.1. Data set

In a recent work (Sahli *et al.,* 2017; a copy of the manuscript accepted for publication was inserted in Chapter 2), we examined 14 058 genes to identify genic CNVs in *P. glauca*. The pedigree population analyzed included 54 full-sib families with 28 to 32 progeny per family for a total of 1650 offspring (Beaulieu *et al.*, 2014). The crossing scheme (Supplemental Table S2.9.1) was a partial diallel implicating 37 mature individuals used as paternal and/or maternal parents in crosses with different partners (each parent was involved in 1 to 5 different crosses). The analysis allowed us to identify 82 genes displaying CNVs among the offspring that were either inherited from the parents (23 CNV genes) or generated through *de novo* events (59 CNV genes). In this paper, we examined the transmission patterns (Mendelian or non-Mendelian segregations) of the 23 inherited CNV loci identified previously. These 23 genes are present in the form of zero-copy

(A0/A0), one-copy (A0/A1) and two-copy genotypes (A1/A1) in the offspring. A0 and A1 are copy number alleles and designate zero-copy allele and one-copy allele respectively.

3.4.2. Statistical analyses

Transmission ratio of the copy number allele A0 was estimated as follow:

$$\mathrm{TR(A0)} = \frac{b}{b+c} \qquad (1)$$

Where *b* and *c* are the number of A0 (zero-copy allele) and A1 (one-copy allele) transmissions from a heterozygous parent to his offspring, respectively. Significant transmission distortions (defined as TR(A0) departures from Mendelian expectations of 0.5) were identified using a two-tailed exact binomial test (α = 0.05). The TR(A0) 95% confidence interval and the test p-value were calculated using the function *binom.test* in R (R Core Team, 2016).

The distribution of TR(A0) for different families was characterized by computing the Gaussian Kernel Density (GKD) using the function *density* in R (R Core Team, 2016).

To examine the potential relationship between the transmission ratio TR(A0) and the genetic distance between parents, we selected 8452 high quality SNPs (GenTrain Score = 0.5 and Call Rate = 1) from the Illumina SNP-array PGLM3 (Pavy *et al.*, 2013). Genotypes for these SNPs were available for the 37 parents analyzed in this study (Beaulieu *et al.*, 2014). We calculated pairwise genetic distances *gd* between parents as the proportion of loci at which the two genotypes being compared were different (Nei and Kumar, 2000). The two-dimensional clustering analysis between TR(A0) and *gd* was conducted using the package mclust v5.2.3 in R (Fraley and Raftery, 2002; R Core Team, 2016).

The dependence of Δp (evolution of the copy number allele A0 frequency between the parental generation n and the next generation n+1) on TR(A0) and *pq* (product of the frequencies of the copy number alleles A0 et A1 in the parental generation, respectively) values was demonstrated through i) fitting of the data to the model proposed by Chevin and Hospital (2006) and ii) ANOVA using the function *aov* in R (R Core Team, 2016).

The function *cor.test* in R (R Core Team, 2016) was used to calculate the one-tailed Pearson correlation coefficient (Cor) between Δp and TR(A0), and the associated p-value.

3.5. Results

In this work, we investigated the inheritance of genic copy number variants by examining a set of 23 High-confidence CNV genes in 1687 individuals from 54 full-sib families. We identified the high-confidence genic CNVs by screening 14 058 genes and selecting those that we detected in at least one parent and several of the full-sib progeny (see methods). We estimated the levels of transmission distortion (TD), identified potential factors involved in the control of CNV inheritance and evaluated the impact of TD on frequency change between generations and on the maintenance or elimination of variants.

3.5.1. Most CNVs are associated with transmission distortions

We examined the transmission profile of the 23 inherited CNVs in bi-parental crosses where at least one parent was a heterozygote for one-copy (A1) and zero-copy (A0) alleles. A situation of transmission distortion (TD) is a departure from the expected transmission ratio 0.5 under Mendel's laws of inheritance. We found that 16 (70%) of the 23 CNV genes displayed transmission ratio distortions (TRDs), while the remaining seven (30%) were transmitted according to the Mendelian expectations or the number of trios examined was too small to detect TRDs (Table 3.1). Preferential transmission of one-copy (allele A1) was found in 13 (81%) of the TRD genes and preferential transmission of zero-copy (allele A0) was found in three (19%) of the TRD genes.

The data showed that the transmission distortions can depend on i) whether the heterozygous parent was the paternal or maternal contributor; a parental effect was observed in 44% of the genes with TRDs, with 31% as maternal effects and 13% as paternal effects and/or ii) the copy number genotype of the partner in the cross (by this we mean whether the individual crossed with the heterozygote parent, otherwise called partner throughout the manuscript, has a zero-copy, one-copy or two-copy genotype for the gene under examination for TD), which was observed in 54% of the genes with TRDs (Table 3.1).

Table 3.1: Parental and partner genotype effects on copy number transmission ratio distortion (cnTRD).

GenBank Acc	♀He x ♂Ho	♀Ho x ♂He	♀He x ♂He	He x N	Favored CN	PE	PGE
BT103424	+	-	+	+	0	D (ME)	I
BT109608	+	-	-	+	0	D (ME)	D
BT101196	-	+	-	-	0	I (PME)[1]	D
BT105201	+	+	NA	+	1	I (PME)	ND
BT107150	+	+	+	+	1	I (PME)	I
CO253730	+	+	+	+	1	I (PME)	I
BT110689	+	+	+	+	1	I (PME)	I
BT119696	-	+	NA	+	1	D (PE)	ND
BT113426	-	+	-	+	1	D (PE)	D
BT111788	+	-	+	+	1	D (ME)	I
BT106719	+	-	-	+	1	D (ME)	D
BT106118	+	-	-	+	1	D (ME)	D
DV985927	-	-	NA	+	1	I	ND
BT109181	-	-	-	+	1	I	I
BT116133	-	-	+	+	1	I	D
DR563872	-	-	+	-	1	I	D
DR587158	-	-	NA	-	ND	ND	ND
BT103050	-	-	-	-	ND	ND	ND
BT119776	-	-	NA	-	ND	ND	ND
BT101202	-	-	NA	-	ND	ND	ND
BT114401	-	-	-	-	ND	ND	ND
BT110740	-	-	NA	-	ND	ND	ND
BT109138	-	-	-	-	ND	ND	ND

[1]More details in Table 3.
He: heterozygote, Ho: homozygote, N: heterozygote or homozygote, CN: copy number allele, PE: parental effect, PGE: partner genotype effect, TRD: transmission ratio distortion, +: significant TRD, -: non-significant TRD, NA: not available, ND: not determined, D: dependent, I: independent, (PE): paternal effect, (ME): maternal effect, (PME): paternal and maternal effects. A more detailed version of this simplified table is presented in Table S1.

We also observed that the level of transmission distortion is dependent on the genetic background of the parents. For a parent participating in different crosses, the level of transmission distortion will vary for different partners as shown in Figure 3.1-A for the gene *BT101196* and Figure S1 for all the genes. This was confirmed by the variable distributions of the transmission ratios TR(A0) for different families, ranging from broad to narrow and mono-modal to bi-modal (Figure S2).

An additional factor that may influence the level of transmission distortion is the genetic distance (gd) between the two parents in a cross. Our analysis shows that TR(A0) values

are distributed into two clusters (Figure 3.1-B) with average pairwise genetic distances of 0.24 and 0.28. The two clusters have the same average distortion level (t-test p-value = 6.2E-01) but the TR(A0) variance in cluster 2 (crosses with more distant parents) was twice that of cluster 1, suggesting that crosses involving more distant individuals can give rise to more extreme TRDs. No significant correlations were found between TR(A0) and genetic distance (Cor = -0.20; p-value = 2.4E-01 in group 1 and Cor = 0.06; p-value = 4.3E-01 in group 2).

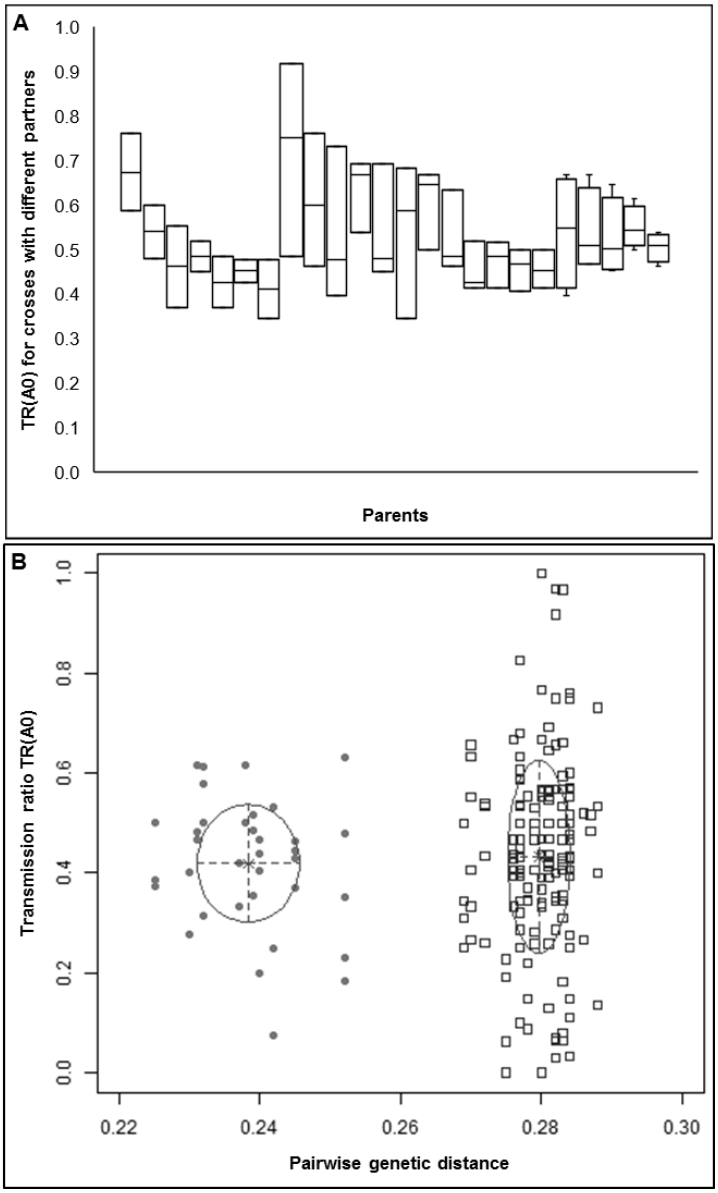


Figure 3.1: Effects of genetic background and genetic distance between parents on copy number transmission ratio distortion (cnTRD). Transmission ratio TR(A0) range when a parent is crossed with different partners for the gene *BT101196* (A). Plot of the transmission ratio TR(A0) versus pairwise genetic distances between the parents involved in each cross; cluster 1 (dots) and cluster 2 (squares) (B).

3.5.2. Transmission distortions contribute to CN allele frequency changes between generations

TDs have the potential to change allele frequencies considerably on a scale of very short evolutionary periods. Here we show that TRDs can contribute to changes in CN allele frequencies ranging between 0.001 and 0.08 in a single generation. We also observed that the CN allele frequency change between the parental generation (n) and the next generation (n+1) delta p(A0) was well correlated with transmission ratios TR(A0) (Cor = 0.73; p-value = 4.2E-05) (Figure 3.2-A). That being said, the level of distortion TR(A0) was not the only factor influencing the CN allele frequency changes between generations (Figure 3.2-B). Chevin and Hospital (2006) proposed a linear model that links the change in allele frequency *Δp* with the transmission ratio *TR* and the product of alleles frequencies in the parental generation *pq*:

$$\Delta p = (2TR - 1)pq \text{ (equation 1 in Chevin and Hospital 2006)} \qquad (2)$$

We found that the model fits our data well (adjusted R^2 = 0.91; p-value = 2.7E-06) for the 13 genes where transmission ratios (TR) are independent of the parental and partner genotype effects. When we considered all of the 23 inherited genic CNVs, the fit (p-value = 9.0 E-07) was less strong (adjusted R^2 = 0.73) due to the interference of the parental and partner genotype effects on TR(A0) levels for some genes. These results were also confirmed by an ANOVA with p-value = 4.0E-06 for the effect of TR(A0) on delta p(A0) and p-value = 5.8E-0.4 for the interaction between TR(A0) and pq.

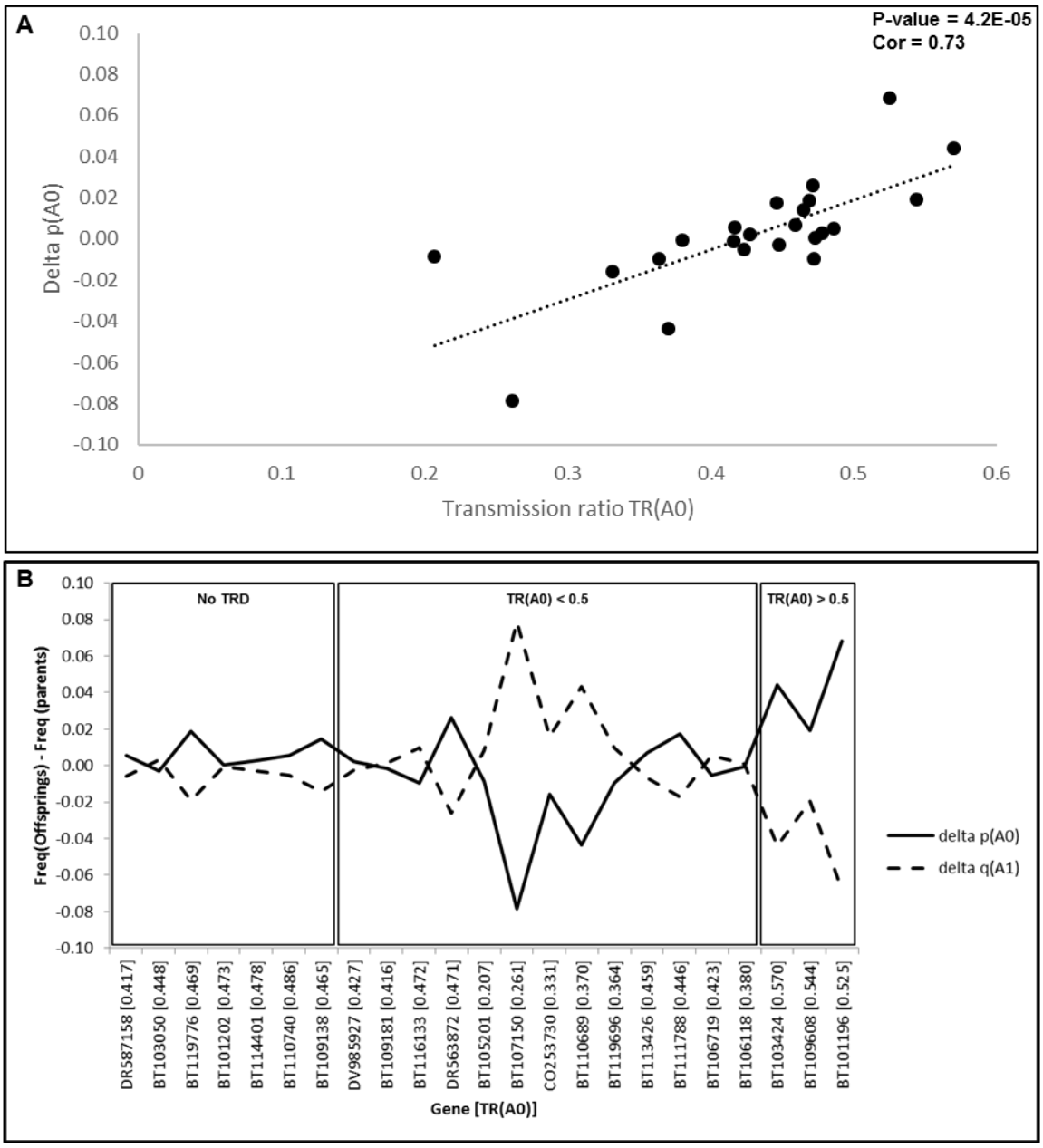


Figure 3.2: Evolution of copy number allele frequencies from the parental generation to the offspring generation is function of the transmission ratio TR(A0). Correlation between delta p(A0) (A0 frequency in generation (n+1) – A0 frequency in generation n) and transmission ratio TR(A0) (A). Evolution of A0 (zero-copy) and A1 (one-copy) alleles frequencies p(A0) and q(A1) respectively, between generations for the 23 inherited CNVs (B).

3.5.3. Genes with preferential transmission of zero copy

For three of the genes (*BT101196*, *BT103424* and *BT109608*) where TDs favored the allele A0 (zero-copy), we identified three different patterns of selection based on the genotypes frequencies (Table 3.2). These patterns were inferred from the examination of genotypes and alleles frequencies in the parents and offspring generations in three steps.

First, we considered all the crosses (Table S2) and made three observations: i) there was a significant departure from Mendelian expectations for the gene *BT101196*, even if the proportion of crosses with TRD was only 31%, because it had a high level of transmission distortion; ii) the level of distortion observed was lower for the gene *BT103424* but was significant in a pedigree population that included 42% of crosses with TRD and 58% crosses with no TRD (those with two homozygote parents and those where the heterozygote parent was male); and iii) the effect of transmission distortions was diluted in the population for the gene *BT109608* because of the interference of both parental and partner genotype effects on TD levels and the presence of crosses between two homozygote parents (76% of crosses with no TRD).

Next, we considered only the crosses with at least one heterozygous parent (Table S3) and observed that there was a significant departure from the genotype frequencies for Mendelian expectations for the three genes analyzed, although the detected level of distortion was moderate due to the interference of the parental and partner genotype effects that still remained in the pedigree population.

Finally, we examined only the crosses with significant TDs (Table S4) and were able to quantify the effect of transmission distortions on changes in genotypes frequencies between generations without the interference of double homozygote crosses, parental or partner genotype effects. Again, there were significant deviations from the expected genotypes frequencies for the three genes and the levels of distortion were higher than those observed in the second step of the analysis.

The findings of these analyses (Table 3.2) show that the frequency of the zero-copy allele (A0) for the gene *BT101196* (highly similar to F-box proteins) increased in the offspring. This resulted from more of the one-copy genotype (A0/A1) and fewer of the two-copy genotype (A1/A1). This observation suggests that this gene is under balancing selection with a heterozygote advantage. On the other hand, the increased frequency of the allele

A0 was due to more of the zero-copy genotype (A0/A0) for the two other genes *BT103424* (unknown function) and *BT109608* (Chaperone DnaJ-domain superfamily protein). This suggests that these genes are under directional selection favoring the allele (A0) and the homozygote genotype (A0/A0) at the expense of the genotype (A1/A1) or both the genotypes (A0/A1) and (A1/A1) for the genes *BT103424* and *BT109608*, respectively.

Table 3.2: Three patterns of selection on the copy number genotypes of genes favoring the transmission of zero copy.

Gene	***BT101196***	***BT103424***	***BT109608***
Predicted function	F-box protein	Unknown	Chaperone DnaJ-domain protein
Selection pressure	Balancing selection	Directional selection	Directional selection
Selection pattern	1	2	3
A0/A0 genotype (Zero copy)	Neutral	Advantageous	Advantageous
A0/A1 genotype (One copy)	Advantageous	Neutral	Deleterious
A1/A1 genotype (Two copies)	Deleterious	Deleterious	Deleterious

3.5.4. The case of the F-box gene *BT101196*

The gene *BT101196* is homologous to an *A. thaliana* gene (*MEE66*, *AT2G02240*) which also displays transmission distortions but follows a distinct pattern of that observed in *P. glauca*. The *A. thaliana* gene is involved in embryo development arrest (Pagnussat *et al.*, 2005). In *P. glauca*, the transmission of the zero-copy allele is favored (TR(A0) > 0.5) and is influenced by the genotype of the partner. We observed that transmission distortions occur only in crosses involving a heterozygous parent and an individual harboring two copies of the gene, i.e. there is no distortion when the partner has a zero- or one- copy genotype (Table 3.3). The data also show simultaneous paternal and maternal effects on TR(A0) with transmission distortions being observed whether the heterozygous parent contributed as male or female (Table 3.3). Like the other genes displaying TRDs, the transmission of the gene *BT101196* was dependent on the parents' genetic background (Figure 3.1-A) but not their pairwise genetic distance (Figures S3).

Table 3.3: Parental and partner genotype effects on copy number transmission ratio distortion (cnTRD) for the F-box gene *BT101196*.

Partner genotype effect				
Crosses	**N**	**TR(A0)**	**CI (95%)**	**P-value**
(A0/A1) x (A1/A1)	458	0.598	0.552 – 0.643	3.0E-05
(A0/A1) x (A0/A1)	244	0.488	0.423 – 0.552	7.5E-01
(A0/A1) x (A0/A0)	402	0.494	0.459 – 0.529	7.5E-01
Maternal effect				
Crosses	**N**	**TR(A0)**	**CI (95%)**	**P-value**
♀ (A0/A1) x ♂ (A1/A1)	239	0.577	0.512 – 0.641	2.0E-02
♀ (A0/A1) x ♂ (A0/A0)	168	0.476	0.399 – 0.555	5.9E-01
Paternal effect				
Crosses	**N**	**TR(A0)**	**CI (95%)**	**P-value**
♀ (A1/A1) x ♂ (A0/A1)	219	0.621	0.553 – 0.686	4.2E-04
♀ (A0/A0) x ♂ (A0/A1)	76	0.513	0.396 – 0.630	9.1E-01

A main difference in the transmission distortion of the two genes is that A0 was favored in the *P. glauca* gene (*BT101196*) and A1 was favored in its *A. thaliana* homolog (*MEE66*, *AT2G02240*) (Pagnussat *et al.*, 2005). This and other TRD features, along with gene expression and embryo viability phenotypes, are summarized in Table 3.4.

Table 3.4: F-box gene copy number transmission in *P. glauca* and *A. thaliana*.

	***P. glauca* (*BT101196*)**	***A. thaliana* (*AT2G02240*)**
Favorably transmitted allele	A0	A1
Parental effect	Maternal effect = Yes Paternal effect = Yes	Maternal effect = Yes Paternal effect = ? (not tested)
Partner genotype effect	Yes A0/A1 x A1/A1 (TRD) A0/A1 x A0/A1 (no TRD)	No A0/A1 x A1/A1 (TRD) A0/A1 x A0/A1 (TRD)
Phenotype	Embryo viability not tested No expression (0/10) in embryogenic cells Low expression (1/10) in megagametophyte Low expression (2/10) in buds	Embryo development arrest Low viability of both gametophytes
Selection pressure on genotypes	A0/A0 (zero copy) viable A0/A1 (one copy) heterozygote advantage A1/A1 (two copies) viable	A0/A0 (zero copy) lethal A0/A1 (one copy) deleterious A1/A1 (two copies) advantageous

3.6. Discussion

Alleles are maintained in a population according to various factors including their effect on the fitness and reproductive success of the individuals carrying them. Transmission distortions (TDs) may circumvent selection pressure and promote the transmission of an allele, even if it is deleterious to the organism. TD is supposed to be a transient state that will lead to a rapid fixation of the favored allele unless antagonistic forces intervene to maintain allelic polymorphism (Taylor and Ingvarsson, 2003). TDs are very common in plants. In *A. lyrata*, 50% of the inspected loci displayed transmission ratio distortions (TRDs) (Kuittinen *et al.*, 2004). In conifers, depending on the species, 2 to 79% (with an average of 20%) of the loci examined by Krutovskii and colleagues (1998) were associated with transmission distortions.

Our analysis identified transmission distortions for 70% of inherited CNVs in *P. glauca*, which is at the upper end of the range reported by Krutovskii and colleagues (1998) for conifers. In contrast, for *P. glauca* gene SNPs, the rate of significant transmission distortions was about 3% (Pavy *et al.*, 2012). Out of the 16 CNV genes with TRDs, the majority (81%) favored the transmission of one-copy allele (A1) instead of zero-copy allele (A0) which helps to maintain two-copy genotypes and counteracts the accumulation of copy number losses by drift and mutations.

TRD levels reported in other species are between 0.3 and 0.6 except in a few cases (reviewed in Huang *et al.*, 2013). In our study, TRD levels for preferential transmission of one-copy (A1) and zero-copy (A0) alleles were 0.06 – 1.00 and 0.34 – 1.00, respectively, for different families. For the aggregated data, these ranges are 0.53 – 0.79 and 0.53 – 0.57, respectively. These observations show that even within a single species, TRD levels vary widely and can be as extreme as the systematic transmission of one allele at the expense of the other (for TR approaching 0 or 1). Allele frequency changes contributed by TRDs can reach 0.08 within one generation for some *P. glauca* CNV genes (present study), which predicts that the favored allele could reach fixation in less than ten generations. This indicates that transmission distortion may be a strong evolutionary force capable of shaping the genetic diversity in a short evolutionary period.

TRDs can result from mechanisms operating during the gametic (pre- or post- meiosis) or the zygotic stage (pre- or post- fecundation). TRDs can also be parent-sex dependent sdTRDs (paternal effect only or maternal effect only) or independent siTRDs (both,

paternal and maternal effects). sdTRDs have a gametic origin and are frequent at the intra-population level. On the other hand, siTRDs have a zygotic origin, play a large role at the inter-population level, are rare within a species and display higher distortion levels than sdTRDs (Koide *et al.*, 2008, 2012; Leppälä *et al.*, 2008; Huang *et al.*, 2013). The majority of TRDs reported in *A. lyrata* (Leppälä *et al.*, 2013) were sdTRDs; however, 46% of the identified TRDs were sex-independent in *A. thaliana* (Pagnussat *et al.*, 2005). In *P. glauca*, siTRDs accounted for 42% and were associated with higher levels of distortion compared to sdTRDs for the TRDs favoring the transmission of one-copy (Table S1). For TRDs favoring the transmission of zero-copy, sdTRDs levels were slightly higher than siTRDs, although the number of observations was too small to draw clear conclusions (Table S1).

In this study, we showed that a considerable proportion of TRDs (54%) was influenced by the genotype of the partner at the TRD locus. Also, consistent with observations in other plants (Buckler *et al.*, 1999; Koide *et al.*, 2012; Leppälä *et al.*, 2013), we found that TRDs levels vary considerably for different genetic backgrounds in the parents. Transmission distortion is a genetically controlled mechanism that can be influenced by the partner genotype at the TRD locus, as well as by the genetic background and the genetic distance between the partners. Incompatibilities between alleles of the TRD locus (e.g. the *S* locus in *A. lyrata* (Leppälä *et al.*, 2008) and the *D* locus in monkeyflower (Fishman and Willis, 2005)) or of linked and/or unlinked loci located elsewhere in the genome (e.g. in wheat (Friebe *et al.*, 2003) and in rice (Koide *et al.*, 2012)) can influence TRD levels. This dependence of TRD levels on the genetic background can be the result of different non-exclusive mechanisms: i) the action of a linked or unlinked driver (distorter or suppressor), ii) the epistatic interactions between different loci in the genome or iii) the effects of cytoplasmic factors. We were unable to identify which mechanism is implicated in the genetic control of TRDs in *P. glauca* because we lacked information on the co-segregation of TRD loci with other genomic markers in reciprocal crosses.

Our analysis in *P. glauca* showed no significant association between TRD levels and pairwise genetic distance between parents. It was suggested that transmission distortions contribute to the establishment of reproductive barriers and TRD levels should increase linearly (Leppälä *et al.*, 2013) or exponentially (snow-ball effect) (Moyle and Nakazato, 2010) with the genetic distance between parents and the divergence between populations. So far though, the existence of this relationship is controversial because it was detected in some species (Koide *et al.*, 2008; Matsubara *et al.*, 2011; Leppälä *et al.*, 2013) but not in

others (Leppälä *et al.*, 2013). We observed that TRD levels were more variable with higher genetic distances; therefore, the influence of genetic distance on TRD level could manifest itself more clearly in more diverged populations.

Three cases where heterozygote parents preferentially transmitted a zero-copy allele were identified in *P. glauca.* The genes involved likely responded to different selection pressures that promoted their partial or complete suppression from the population. The transmission of the *BT103424* (Unknown protein) and *BT109608* (Chaperone DnaJ-domain protein) genes was subject to a maternal effect (and most likely have a gametic origin). The one-copy allele (A1) of both genes was under negative selection while the zero-copy allele (A0) was under positive selection (accumulating in the form of the homozygote genotype A0/A0 at the expense of the heterozygote genotype A0/A1 and/or homozygote genotype A1/A1). On the other hand, the TRD for the gene *BT101196* (F-box protein) was sex-independent (with both paternal and maternal effects), dependent on the partner genotype (TRD observable only when a heterozygote is crossed with an individual with two copies of the gene) and most likely had a zygotic origin. Its zero-copy (A0) and one-copy (A1) alleles appear to be under balancing selection with a heterozygote advantage. The *BT101196* transcripts were not detected in the embryo but were expressed at low levels in the megagametophyte and vegetative buds (Raherison *et al.*, 2012). This pattern suggests a case of sheltered load were the alleles A0 et A1, which have opposing effects on the fitness at different life stages, are both maintained in the population in the form of heterozygote genotypes. The genotype frequencies observed for the gene *BT101196* suggest that the allele A1 is likely deleterious (but not lethal) for the embryo and its expression indicates that it is necessary in vegetative buds later in development. In *A. thaliana*, transmission of a homologous gene sequence (*AT2G02240*, designated by *MEE66*) was also shown to be distorted under a maternal effect, but in contrast to *P. glauca*, the allele A1 was preferentially transmitted to the next generation (Pagnussat *et al.*, 2005) and was under positive selection (Table 3.4). The different transmission behaviors of these F-box genes, in the two species, could potentially reflect different evolutionary processes operating in annual angiosperms and perennial gymnosperms such as the conifer *P. glauca*.

Our results show that the transmission of genic CNVs from a generation to the next is distorted most of the time. The majority of the observed transmission distortions favor the transmission of the one-copy allele and the restoration of the two-copy genotype in the

next generation. This pattern is expected in order to maintain the integrity and stability of the diploid spruce genome. On rare occasions, transmission distortions would promote the inheritance of the zero-copy allele either because this variant would behave selfishly or the copy number loss would be advantageous (variant under balancing or positive selection). The present study shows that transmission distortions can cause large allele frequency changes on short evolutionary periods, and that the level of distortion is genetically controlled, which contributes to the maintenance of a substantial standing genetic variation in the population. Future studies of TDs for variants in non-coding sequences (supposedly neutral or at least less deleterious) in species with different life habits and at different developmental stages are expected to provide valuable new insights into the mechanisms underlying the inheritance of genetic variants and the role of transmission distortions in species adaptation and evolution.

3.7. Acknowledgment

Funding was received from Genome Canada and Genome Quebec for the SMarTForests project (JM, JB) and from the Quebec Ministry of Economic Development, Innovation and Exports through the GenAC project (JM, JB).

3.8. References

Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J (2014). Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genomics* **15**: 1048.

Beckmann JS, Estivill X, Antonarakis SE (2007). Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* **8**: 639–646.

Blackburn A, Göring HH, Dean A, Carless MA, Dyer T, Kumar S *et al.* (2013). Utilizing extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. *Eur J Hum Genet* **21**: 404–409.

Buckler ESI, Phelps-Durr TL, Buckler CSK, Dawe RK, Doebley JF, Holtsford TP (1999). Meiotic drive of chromosomal knobs reshaped the maize genome. *Genetics* **153**: 415–426.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M *et al.* (2014). Extensive copy-number variation of young genes across stickleback populations. *PLoS Genet* **10**: e1004830.

Chevin LM, Hospital F (2006). The hitchhiking effect of an autosomal meiotic drive gene. *Genetics* **173**: 1829–1832.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y *et al.* (2010). Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–

712.

Debolt S (2010). Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* **2**: 441–453.

Didion JP, Morgan AP, Clayshulte AMF, Mcmullan RC, Yadgary L, Petkov PM *et al.* (2015). A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLoS Genet* **11**: e1004850.

Fishman L, Willis JH (2005). A novel meiotic drive locus almost completely distorts segregation in *Mimulus* (monkeyflower) hybrids. *Genetics* **169**: 347–353.

Fraley C, Raftery a E (2002). Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* **97**: 611–631.

Friebe B, Zhang P, Nasuda S, Gill BS (2003). Characterization of a knock-out mutation at the *Gc2* locus in wheat. *Chromosoma* **111**: 509–517.

Huang LO, Labbe A, Infante-Rivard C (2013). Transmission ratio distortion: Review of concept and implications for genetic association studies. *Hum Genet* **132**: 245–263.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H-C *et al.* (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998–1003.

Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R, Shapero MH *et al.* (2010). Population-genetic nature of copy number variations in the human genome. *Hum Mol Genet* **19**: 761–773.

Koide Y, Ikenaga M, Sawamura N, Nishimoto D, Matsubara K, Onishi K *et al.* (2008). The evolution of sex-independent transmission ratio distortion involving multiple allelic interactions at a single locus in rice. *Genetics* **180**: 409–420.

Koide Y, Shinya Y, Ikenaga M, Sawamura N, Matsubara K, Onishi K *et al.* (2012). Complex genetic nature of sex-independent transmission ratio distortion in Asian rice species: the involvement of unlinked modifiers and sex-specific mechanisms. *Heredity* **108**: 242–247.

Korbel JO, Kim PM, Chen X, Urban AE, Snyder M, Gerstein MB (2009). The current excitement about copy-number variation: How it relates to gene duplication and protein families. *Curr Opin Struct Biol* **18**: 366–374.

Krutovskii K V., Vollmer SS, Sorensen FC, Adams WT, Knapp SJ, Strauss SH (1998). RAPD genome maps of Douglas-fir. *J Hered* **89**: 197–205.

Kuittinen H, De Haan AA, Vogl C, Oikarinen S, Leppälä J, Koch M *et al.* (2004). Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **168**: 1575–1584.

Leppälä J, Bechsgaard JS, Schierup MH, Savolainen O (2008). Transmission ratio distortion in *Arabidopsis lyrata*: Effects of population divergence and the S-locus. *Heredity* **100**: 71–78.

Leppälä J, Bokma F, Savolainen O (2013). Investigating incipient speciation in *Arabidopsis lyrata* from patterns of transmission ratio distortion. *Genetics* **194**: 697–708.

Lyttle TW (1991). Segregation distorters. *Annu Rev Genet* **25**: 511–557.

Matsubara K, Ebana K, Mizubayashi T, Itoh S, Ando T, Nonoue Y *et al.* (2011). Relationship between transmission ratio distortion and genetic divergence in

intraspecific rice crosses. *Mol Genet Genomics* **286**: 307–319.

McCarroll SA, Altshuler DM (2007). Copy-number variation and association studies of human disease. *Nat Genet* **39**: S37–S42.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C *et al.* (2011). Mapping copy number variation by population scale genome sequencing. *Nature* **470**: 59–65.

Moyle LC, Nakazato T (2010). Hybrid incompatibility 'snowballs' between Solanum species. *Science* **329**: 1521–1523.

Nei M, Kumar S (2000). *Molecular Evolution and Phylogenetics*, Oxford University Press: Oxford, New York.

Pagnussat GC, Yu HJ, Ngo QA, Rajani S, Mayalagu S, Johnson CS *et al.* (2005). Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development* **132**: 603–614.

Pardo-Manuel de Villena F, De La Casa-Esperon E, Briscoe TL, Sapienza C (2000). A genetic test to determine the origin of maternal transmission ratio distortion: Meiotic drive at the mouse *Om* locus. *Genetics* **154**: 333–342.

Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J (2012). A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biol* **10**: 84.

Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B *et al.* (2013). Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Mol Ecol Resour* **13**: 324–336.

Polański A, Chakraborty R, Kimmel M, Deka R (1998). Dynamic balance of segregation distortion and selection maintains normal allele sizes at the myotonic dystrophy locus. *Math Biosci* **147**: 93–112.

R Core Team (2016). R: A Language and Environment for Statistical Computing.

Raherison E, Rigault P, Caron S, Poulin PL, Boyle B, Verta JP *et al.* (2012). Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. *BMC Genomics* **13**: 434.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD *et al.* (2006). Global variation in copy number in the human genome. *Nature* **444**: 444–54.

Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO (2011). Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res* **21**: 2004–2013.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P *et al.* (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Seidel HS, Ailion M, Li J, van Oudenaarden A, Rockman M V., Kruglyak L (2011). A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in *C. elegans*. *PLoS Biol* **9**: e1001115.

Stranger B, Forrest M, Dunning M, Ingle C (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–853.

Tan S, Zhong Y, Hou H, Yang S, Tian D (2012). Variation of presence/absence genes among *Arabidopsis* populations. *BMC Evol Biol* **12**: 86.

Taylor DR, Ingvarsson PK (2003). Common features of segregation distortion in plants and animals. *Genetica* **117**: 27–35.

3.9. Supplementary information

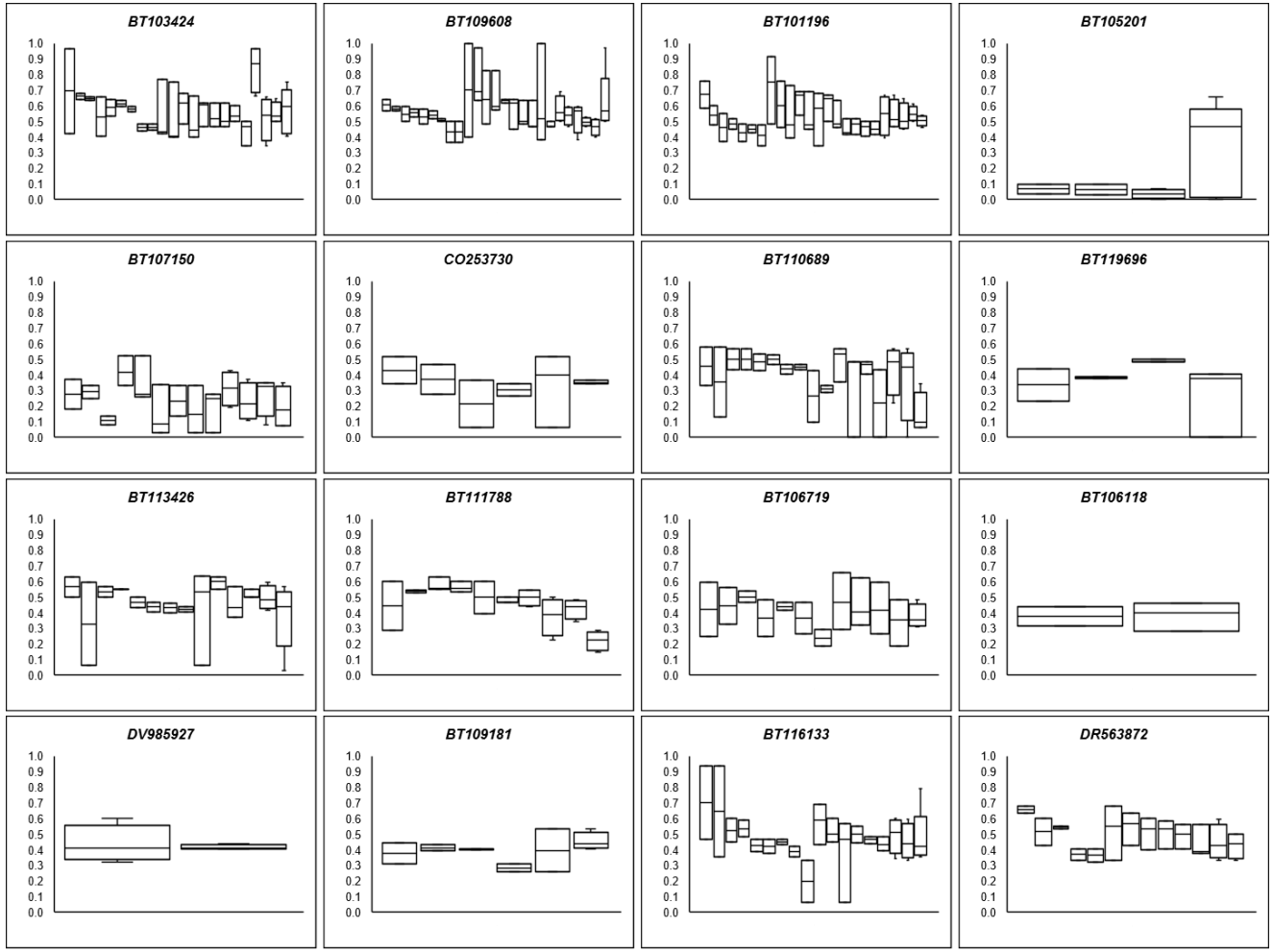


Figure S1: Genetic background effect on copy number transmission ratio distortion (cnTRD). Transmission ratio TR(A0) range when a parent is crossed with different partners for the 16 genes displaying significant transmission distortions. x-axis: parents involved in two or more crosses, y-axis: TR(A0).

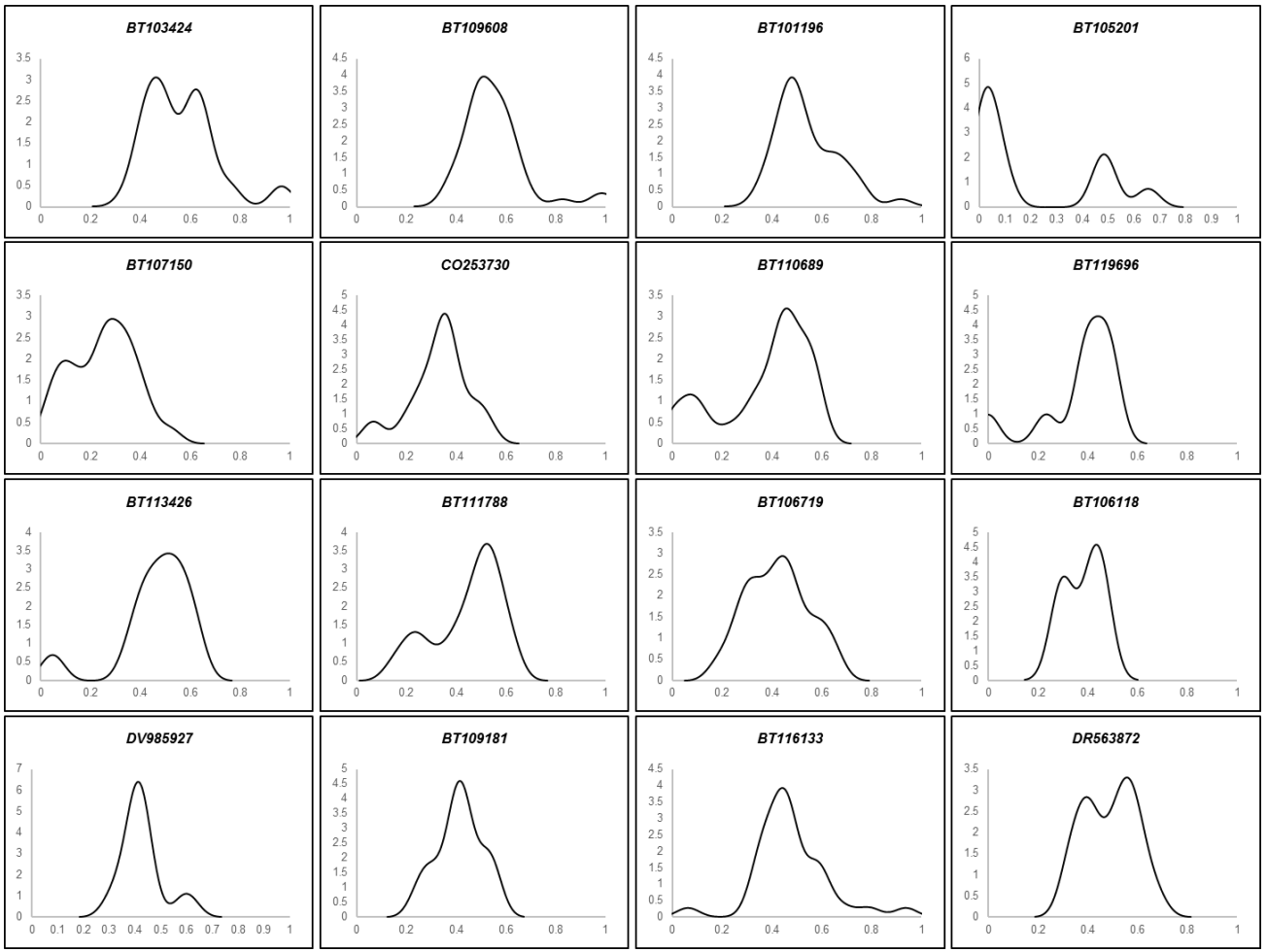


Figure S2: Genetic background effect on copy number transmission ratio distortion (cnTRD). The distribution of transmission ratio TR(A0) for different families for each of the 16 genes displaying significant transmission distortions. x-axis: TR(A0), y-axis: density.

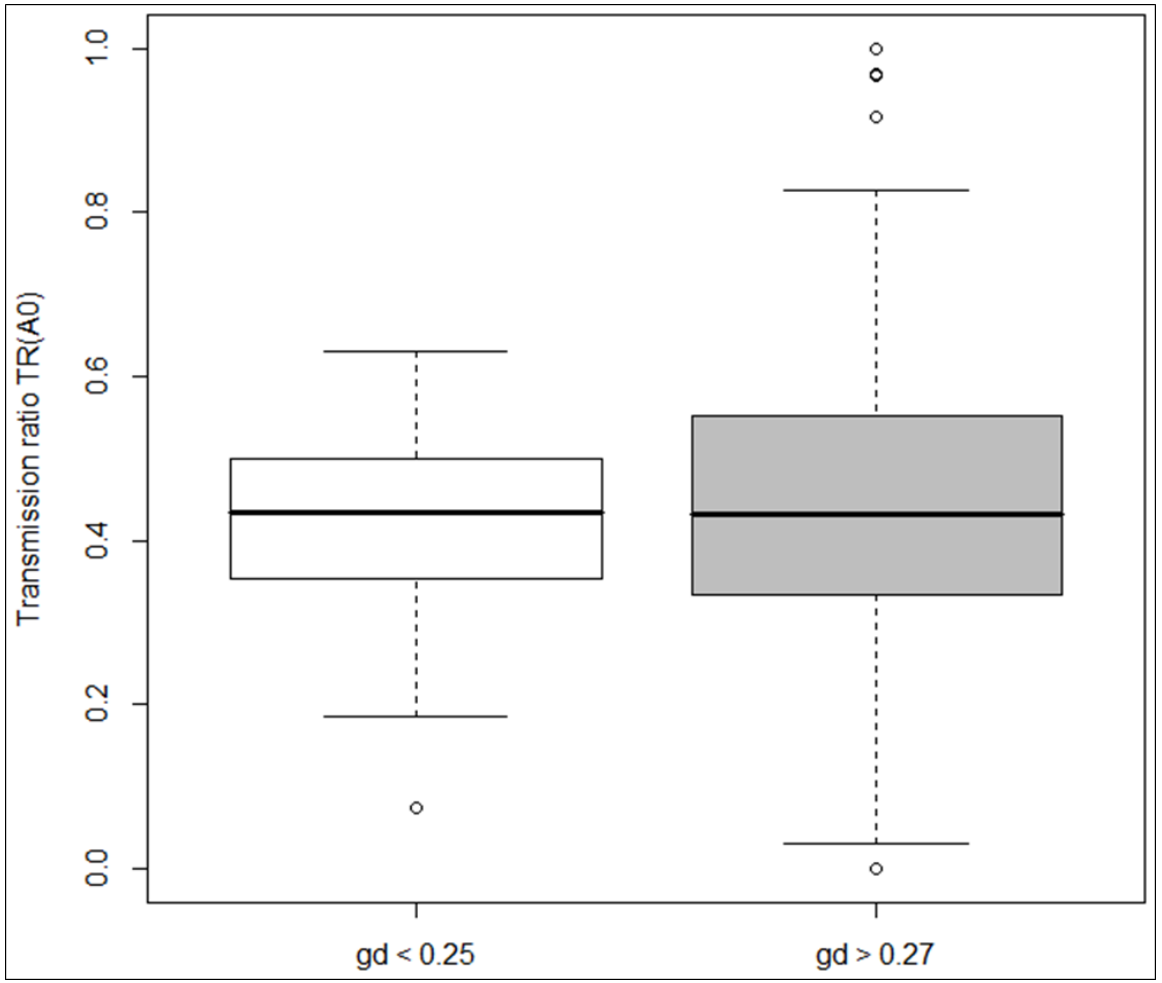


Figure S3: Genetic distance between parents' effect on copy number transmission ratio distortion (cnTRD) for the F-box gene *BT101196*. Distribution of transmission ratio TR(A0) for crosses where the two parents are genetically close (white box) or more distant (grey box). gd (pairwise genetic distance between parents).

Table S1: Parental and partner genotype effects on copy number transmission ratio distortion (cnTRD).

GenBank Acc	♀He x ♂Ho crosses				♀Ho x ♂He crosses				♀He x ♂He crosses				He x N crosses				Favored CN	PE	PGE
	n	TR(A0)	CI (95%)	P-value	n	TR(A0)	CI (95%)	P-value	n	TR(A0)	CI (95%)	P-value	n	TR(A0)	CI (95%)	P-value			
BT103424	392	0.554	0.502 - 0.603	3.80E-02	421	0.542	0.492 - 0.589	9.70E-02	308	0.601	0.560 - 0.639	6.60E-07	1121	0.57	0.544 - 0.596	1.20E-07	0	D (ME)	I
BT109608	384	0.604	0.553 - 0.653	5.20E-05	387	0.543	0.491 - 0.593	1.00E-01	464	0.52	0.487 - 0.553	2.20E-01	1235	0.544	0.520 - 0.568	2.70E-04	0	D (ME)	D
BT101196	407	0.536	0.485 - 0.584	1.70E-01	295	0.593	0.534 - 0.649	1.60E-03	402	0.494	0.458 - 0.528	7.50E-01	1104	0.525	0.498 - 0.550	6.00E-02	0	I (PME)	D
BT105201	150	0.147	0.094 - 0.213	< 2.2e-16	213	0.249	0.192 - 0.312	1.10E-13	NA	NA	NA	NA	363	0.207	0.166 - 0.251	< 2.2e-16	1	I (PME)	ND
BT107150	276	0.236	0.186 - 0.290	< 2.2e-16	250	0.208	0.159 - 0.263	< 2.2e-16	153	0.327	0.274 - 0.382	1.30E-09	679	0.261	0.231 - 0.292	< 2.2e-16	1	I (PME)	I
CO253730	150	0.353	0.277 - 0.435	4.10E-04	185	0.308	0.242 - 0.380	1.90E-07	32	0.344	0.229 - 0.473	1.70E-02	367	0.331	0.284 - 0.379	1.30E-11	1	I (PME)	I
BT110689	319	0.386	0.331 - 0.441	5.20E-05	316	0.358	0.304 - 0.413	4.70E-07	190	0.368	0.319 - 0.419	3.30E-07	825	0.37	0.340 - 0.400	< 2.2e-16	1	I (PME)	I
BT119696	183	0.432	0.358 - 0.506	7.60E-02	89	0.225	0.142 - 0.325	1.80E-07	NA	NA	NA	NA	272	0.364	0.306 - 0.424	8.50E-06	1	D (PE)	ND
BT113426	365	0.479	0.427 - 0.532	4.60E-01	326	0.436	0.381 - 0.491	2.30E-02	31	0.468	0.339 - 0.598	7.00E-01	722	0.459	0.423 - 0.495	2.90E-02	1	D (PE)	D
BT111788	258	0.407	0.346 - 0.469	3.40E-03	178	0.506	0.429 - 0.581	9.40E-01	249	0.446	0.401 - 0.490	1.70E-02	685	0.446	0.414 - 0.479	1.20E-03	1	D (ME)	I
BT106719	298	0.383	0.327 - 0.440	6.00E-05	251	0.47	0.407 - 0.533	3.80E-01	86	0.424	0.349 - 0.501	5.60E-02	635	0.423	0.386 - 0.460	4.10E-05	1	D (ME)	D
BT106118	96	0.344	0.249 - 0.447	2.90E-03	28	0.464	0.275 - 0.661	8.50E-01	30	0.4	0.275 - 0.534	1.60E-01	154	0.38	0.310 - 0.454	1.50E-03	1	D (ME)	D
DV985927	119	0.42	0.330- 0.514	9.90E-02	122	0.434	0.344 - 0.527	1.70E-01	NA	NA	NA	NA	241	0.427	0.364 - 0.492	2.80E-02	1	I	ND
BT109181	87	0.414	0.309 - 0.524	1.30E-01	115	0.409	0.317 - 0.504	6.20E-02	54	0.426	0.331 - 0.524	1.50E-01	256	0.416	0.360 - 0.473	3.70E-03	1	I	I
BT116133	355	0.482	0.428 - 0.535	5.20E-01	244	0.484	0.419 - 0.548	6.50E-01	365	0.463	0.426 - 0.499	5.00E-02	964	0.472	0.444 - 0.499	4.20E-02	1	I	D
DR563872	298	0.47	0.412 - 0.528	3.20E-01	270	0.544	0.482 - 0.604	1.60E-01	94	0.367	0.298 - 0.440	3.30E-04	662	0.471	0.434 - 0.507	1.20E-01	1	I	D
DR587158	29	0.414	0.235 - 0.610	4.60E-01	55	0.418	0.286 - 0.558	2.80E-01	NA	NA	NA	NA	84	0.417	0.309 - 0.529	1.60E-01	ND	ND	ND
BT103050	146	0.493	0.409 - 0.577	9.30E-01	146	0.432	0.349 - 0.515	1.20E-01	26	0.365	0.236 - 0.510	7.00E-02	318	0.448	0.394 - 0.501	5.90E-02	ND	ND	ND
BT119776	58	0.397	0.270 - 0.533	1.50E-01	119	0.504	0.411 - 0.597	1.00E+00	NA	NA	NA	NA	177	0.469	0.393 - 0.545	4.50E-01	ND	ND	ND
BT101202	218	0.505	0.436 - 0.572	9.50E-01	171	0.433	0.357 - 0.510	9.20E-02	NA	NA	NA	NA	389	0.473	0.422 - 0.523	3.10E-01	ND	ND	ND
BT114401	272	0.467	0.406 - 0.528	3.00E-01	297	0.502	0.443 - 0.559	1.00E+00	88	0.455	0.379 - 0.531	2.60E-01	657	0.478	0.441 - 0.514	2.40E-01	ND	ND	ND
BT110740	84	0.476	0.366 - 0.588	7.40E-01	58	0.5	0.365 - 0.634	1.00E+00	NA	NA	NA	NA	142	0.486	0.401 - 0.571	8.00E-01	ND	ND	ND
BT109138	92	0.543	0.436 - 0.647	4.70E-01	61	0.443	0.315 - 0.575	4.40E-01	31	0.371	0.251 - 0.503	5.60E-02	184	0.465	0.397 - 0.534	3.40E-01	ND	ND	ND

He: heterozygote, Ho: homozygote, N: heterozygote or homozygote, CN: copy number allele, PE: parental effect, PGE: partner genotype effect, n: number of offsprings examined, TR(A0): transmission ratio for the copy number allele A0 (zero copy), CI (95%): 95% confidence interval, P-value: p-value for two-tailed exact binomial test, NA: not available, ND: not determined, D: dependent, I: independent, (PE): paternal effect, (ME): maternal effect, (PME): paternal and maternal effects.

Table S2: Genotypes and alleles frequencies in the parental and offspring generations for three genes favoring the transmission of zero copy for all crosses.

	Gene *BT101196*					Gene *BT103424*					Gene *BT109608*				
	Genotypes			Alleles		Genotypes			Alleles		Genotypes			Alleles	
	A0/A0	A0/A1	A1/A1	A0	A1	A0/A0	A0/A1	A1/A1	A0	A1	A0/A0	A0/A1	A1/A1	A0	A1
Parental generation (P)	0.13	0.50	0.37	0.38	0.62	0.36	0.44	0.20	0.58	0.42	0.44	0.52	0.04	0.70	0.30
Offspring generation (E)	0.15	0.46	0.39	0.38	0.62	0.36	0.43	0.21	0.58	0.42	0.51	0.40	0.09	0.71	0.29
Offspring generation (O)	0.17	0.56	0.27	0.45	0.55	0.40	0.44	0.16	0.62	0.38	0.53	0.39	0.08	0.72	0.28
O - E	0.02	0.10	-0.12	0.07	-0.07	0.04	0.01	-0.05	0.04	-0.04	0.02	-0.01	-0.01	0.01	-0.01
P(χ^2)	1.0E-16			1.0E-16		1.0E-16			1.0E-16		1.7E-01			2.0E-01	

E: Expected under Mendelian inheritance (no transmission distortion); O: observed; P(χ^2): Chi-square test p-value.

Table S3: Genotypes and alleles frequencies in the parental and offspring generations for three genes favoring the transmission of zero copy for crosses with at least one heterozygote parent.

	Gene *BT101196*					Gene *BT103424*					Gene *BT109608*				
	Genotypes			Alleles		Genotypes			Alleles		Genotypes			Alleles	
	A0/A0	A0/A1	A1/A1	A0	A1	A0/A0	A0/A1	A1/A1	A0	A1	A0/A0	A0/A1	A1/A1	A0	A1
Parental generation (P)	0.11	0.68	0.21	0.45	0.55	0.22	0.63	0.15	0.53	0.47	0.28	0.68	0.04	0.62	0.38
Offspring generation (E)	0.20	0.50	0.30	0.45	0.55	0.28	0.50	0.22	0.53	0.47	0.37	0.50	0.13	0.62	0.38
Offspring generation (O)	0.20	0.54	0.26	0.47	0.53	0.33	0.50	0.17	0.58	0.42	0.41	0.49	0.10	0.66	0.34
O - E	0.00	0.04	-0.04	0.02	-0.02	0.05	0.00	-0.05	0.05	-0.05	0.04	-0.01	-0.03	0.04	-0.04
P(χ^2)	2.0E-02			6.0E-02		1.0E-04			1.0E-16		5.0E-03			1.0E-16	

E: Expected under Mendelian inheritance (no transmission distortion); O: observed; P(χ^2): Chi-square test p-value.

Table S4: Genotypes and alleles frequencies in the parental and offspring generations for three genes favoring the transmission of zero copy for crosses displaying transmission distortions.

	Gene *BT101196*					Gene *BT103424*					Gene *BT109608*				
	Genotypes			Alleles		Genotypes			Alleles		Genotypes			Alleles	
	A0/A0	A0/A1	A1/A1	A0	A1	A0/A0	A0/A1	A1/A1	A0	A1	A0/A0	A0/A1	A1/A1	A0	A1
Parental generation (P)	0.00	0.50	0.50	0.25	0.75	0.17	0.72	0.11	0.53	0.47	0.42	0.50	0.08	0.67	0.33
Offspring generation (E)	0.00	0.50	0.50	0.25	0.75	0.28	0.50	0.22	0.53	0.47	0.43	0.50	0.07	0.68	0.32
Offspring generation (O)	0.00	0.60	0.40	0.30	0.70	0.34	0.51	0.15	0.59	0.41	0.51	0.45	0.04	0.74	0.26
O - E	0.00	0.10	-0.10	0.05	-0.05	0.06	0.01	-0.07	0.06	-0.06	0.08	-0.05	-0.03	0.06	-0.06
$P(\chi^2)$	1.0E-16			1.0E-16		1.0E-16			1.0E-16		7.0E-04			4.0E-04	

E: Expected under Mendelian inheritance (no transmission distortion); O: observed; $P(\chi^2)$: Chi-square test p-value.

Chapter 4: Conclusions

The main goals of the present thesis were to estimate the rate at which genic CNVs are generated and investigate their transmission from a generation to the next in the conifer *P. glauca*. To date, knowledge about CNVs genetic properties is still limited particularly for non-model organisms including trees. This study represents the first attempt to estimate the mutation rate of CNVs (Chapter 2) and describe the transmission patterns of these genetic variations (Chapter 3) in trees. In addition, the approach that we applied in the present work is based on SNP genotyping array hybridization data and to our knowledge, had not been attempted in trees previously, although it is commonly used in human. To reach our goals, we used genome-wide genotyping data collected from a large multiparental pedigree population. Our findings provided new insights into the generation and inheritance of CNVs and their contribution to the evolutionary process and lay the ground for future investigations of these important genetic variations in natural populations and their potential use in tree breeding programs. In this chapter, we summarize and critically assess the main findings of our study (section 4.1) and present some perspectives and future research directions that will allow for a better understanding of the contribution of CNVs to trees' adaptation and evolution (section 4.2).

4.1. Major findings and critical assessment of the thesis work

4.1.1. CNV detection and classification

The first main objective of this PhD thesis was to develop an approach for the detection of genic CNVs in *P. glauca*, estimate the prevalence of these variations in the gene space and proceed to their classification. The postulated hypotheses were:

Hypothesis 1: Genic CNVs affect a small proportion of the gene space.

Hypothesis 2: Gene copy losses are expected to be more abundant than gene copy gains.

For this thesis, we used SNP-array raw intensity data available for 3663 individuals to scan 14 058 genes for CNVs. CNV identification from SNP-array data can be performed using cross-genome or cross-sample analyses, as the two main approaches (Marioni *et al.*, 2007). The former requires a prior knowledge of the position of probes on the chromosomes and relies on the examination of signal intensity of adjacent probes to infer

CNVs accurately. Typically, a copy number variant may be called if at least six adjacent probes provide the same signal intensity. The latter is based on the examination of the signal intensity of each probe separately and the accuracy of CNV calls relies on the analyses of many individuals. When a large number of individuals is examined, the signal intensity clusters are well defined and the theoretical model fits the empirical data better.

The genome of *P. glauca* is very large (20 Gbp) and enriched in repeated sequences (50 to 70%) which prevented the assembly of complete chromosome reference sequences using the available sequencing technologies (the current reference assembly of *P. glauca* genome although representing the whole genome, is still fragmented in about four million scaffolds) (Birol *et al.*, 2013; Warren *et al.*, 2015). Consequently, for *P. glauca* the position of the array probes on chromosomes is not available. Hence the cross-genome approach cannot be applied. However, the cross-sample approach, can be used reliably in this situation because it does not require a prior knowledge of probe positions and the number of individuals analyzed is large (3663 trees).

Two algorithms (PlatiniumCNV (Kumasaka *et al.*, 2011) and GStream (Alonso *et al.*, 2013)) were used to identify CNVs based on the cross-sample approach, which offered two advantages i) the opportunity to compare the results obtained from the two algorithms and ii) the possibility of inferring allele-specific CNVs. We applied conservative criteria for CNV identification including a 95% threshold for i) reproducibility between technical replicates and ii) consistency between the two algorithms used. The validation of CNV calls using an independent technique (qPCR) allowed the estimation of a False Discovery Rate (FDR) of 6.6% and an average genotyping accuracy of 86%. These validation metrics are similar to what was observed for other species when SNP-array, aCGH and NGS technologies were used for CNV detection (Kato *et al.*, 2010; Cicconardi *et al.*, 2013; Chain *et al.*, 2014).

We also performed a pedigree reconstruction analysis based on the copy number genotypes obtained for each individual and were able to infer the correct pedigree structure with an accuracy ranging between 92 and 96%. The validation and pedigree reconstruction results suggest that the inferred copy number genotypes are quite accurate and the high-confidence CNV set obtained can be used to estimate the mutation rate of CNVs and analyze their transmission patterns. That being said, we also recognize that the genotyping accuracy of CN losses and gains is not the same and the proposed method for

CNV detection is less reliable when it comes to identifying CN gains (average genotyping accuracy of 63% for CN gains). However since the majority (90%) of the detected CNVs represent CN losses, we believe that the uncertainty associated with CN gain genotyping should not affect dramatically the overall findings of this study.

Using the above-mentioned approach, we identified a set of 143 genes displaying CNVs among individuals. This set represent less than 1% of the genes we examined, which indicates that CNVs affect a small proportion of the gene space in accordance with Hypothesis 1. Most of the detected CNVs (90%) are CN losses, which is consistent with what was observed in other species using different technologies (Kato *et al.*, 2010; Swanson-wagner *et al.*, 2010; Mills *et al.*, 2011; Yu *et al.*, 2011; McHale *et al.*, 2012; Cicconardi *et al.*, 2013; Chain *et al.*, 2014). However, since the cause of this bias can be technical or biological (as discussed in Chapter 2), it is unclear if the data support Hypothesis 2 or not and additional analyses are required in order to establish that CN losses are more abundant than CN gains.

The examination of the CN genotypes for the offspring and their respective parents allowed us to distinguish between CNVs that are inherited and those that are generated through *de novo* mutation events. At the individual level, we estimated that each individual will harbor on average less than 20 non-two-copy genotypes with 17 inherited from the parents and zero to three resulting from *de novo* events. These proportions are consistent among the 54 analyzed families and a similar number of inherited CNVs was detected in the large family (around 2000 offspring) and the small families (around 30 offspring per family) analyzed, which suggests that the observed pattern regarding the origin of CNVs is not biased because of the size of the analyzed families.

4.1.2. CN mutation rate estimation

The second main objective of this PhD thesis was to estimate CN mutation rates across the genome and investigate the relationship between CN mutation rate and gene expression. The postulated hypotheses were:

Hypothesis 3: Copy number mutation rate is low and variable across the genome.

Hypothesis 4: Copy number mutation rate is associated with gene expression.

In this study, we report the first estimates of the mutation rate for genic CNVs in trees. We used a family-based approach to directly estimate the mutation rate for CNVs located in 113 different genes. Our results show that the average mutation rate across the gene space is high (10^{-5} mutation per gene per generation). The mutation rate spectrum also showed that μ distribution is bi-modal with a threshold mutation rate of 10^{-2} mutation per generation separating the two modes regardless of the data set or the CNV class (HoD, HeD or CN gain). The mutation rate covers a range which spans three orders of magnitude (2.6×10^{-4} to 9.3×10^{-2} mutation per generation) but since our experimental design (mainly the number of genotyped individuals) does not allow for the detection of mutation rates lower than 2.5×10^{-4} mutation per generation, the lower bound of μ distribution is still unknown. Our results also show that the mutation rate varies for different genes, alleles and CNV classes. These variations may reflect i) different local genomic features near CNV loci or ii) different selection pressures acting on CN variants. These findings partially support Hypothesis 3, which proposed that μ should be low (rejected supposition) and variable across the genome (accepted supposition).

We also examined the relationship between the level of gene expression and mutation rate. This relationship remains controversial and different hypotheses were proposed to explain it. The transcription-coupled repair hypothesis TCRH (Hendriks *et al.*, 2010; Fidantsef and Britt, 2011) proposes that highly expressed genes are associated with lower mutation rates. On the other hand, the transcription-associated mutagenesis hypothesis TAMH (Park *et al.*, 2012; Sollier *et al.*, 2014; Heinäniemi *et al.*, 2016) proposes that highly expressed genes are associated with higher mutation rates. Our results show that μ distribution is bi-modal: in mode 1, μ is below 10^{-2} mutation per generation, and in mode 2, μ is above 10^{-2} mutation per generation. When we inspected the relationship between gene expression and μ in each mode separately, we found no significant correlation in mode 1 (low mutation rates) while in mode 2 (high mutation rates), there was a significant negative correlation between gene expression and μ. These observations suggest that the TAMH is not valid for both modes and the TCRH is valid only for mode 2. Lynch (2010, 2011) proposed the drift-barrier hypothesis (DBH) to describe μ evolution: selection drives the mutation rate down until μ reaches a lower bound where selection is overcome by the power of genetic drift (μ can go up or down randomly from there).

> The lower bound on the mutation rate is not set by physiological or biochemical limitations, but by the intrinsic inability of selection to push the rate any lower. The power of random genetic drift (*1/2Ne* for diploid organisms, where *Ne* is the genetic effective population size) ultimately constrains what natural selection can accomplish with any trait, and once the mutation rate is pushed to such a low level that any further incremental improvement conveys a fitness advantage smaller than the power of drift, selection will be incapable of reducing the rate any further. - Lynch (2010).

When the relationship between µ and gene expression is considered within the DBH framework, we expect that µ will be negatively correlated with gene expression for high mutation rates, while the relationship between µ and gene expression will be random (no correlation) for low mutation rates. Our results conform to these expectations, which suggests that the relationship between µ and gene expression is best described in the context of the DBH.

4.1.3. CNV Inheritance

The third main objective of this PhD thesis was to characterize the transmission patterns of CNVs from the parents to their progeny. The postulated hypotheses were:

Hypothesis 5: Transmission distortions (TDs) are frequent and cause significant frequency changes between generations.

Hypothesis 6: TDs are genetically controlled.

For 23 genes, we examined the transmission of CN variants from the parental generation to the offspring generation using 1650 trios. The results show that 70% of these genic CNVs are transmitted in violation of Mendelian expectations. The majority of the detected TDs (81%) promote the transmission of the one-copy allele (A1) and the restoration of the two-copy genotype in the next generation. The estimated TRD levels for CNVs ranged between 0 and 1 and caused significant frequency change between generations according to the model proposed by Chevin and Hospital (2006). TDs were also subject to parental effects and controlled by genetic factors (genetic background and partner genotype but not the genetic distance between parents). These findings are consistent with the Hypotheses 5 and 6.

4.1.4. Unanswered questions

In this work, we developed a reliable approach for CNV identification in the gene space of *P. glauca* using SNP-array raw signal intensity data. This approach is transferable to other non-model organisms, particularly those with large and complex genomes and a long generation time, provided that reference genic sequences are available. We used the inferred CN genotypes to provide new insights into the generation and transmission of CNVs. However, several questions remain unanswered regarding the biology of CNVs. For instance, hybridization technologies such as SNP arrays do not provide information about the size of CNVs and the molecular mechanisms responsible for their formation. Also, the probes used in our experiments only target genic sequences and we still do not know the prevalence of CNVs in non-coding regions. Without sequencing data, it is not possible to inspect the genomic sequences in the vicinity of newly generated CN variants. Hence, the impact of local genomic features on the mutation rate could not be investigated. Our analysis identified several cases of TDs but did not explore the underlying causes of these distortions.

4.2. Research perspectives and potential applications of this study

To further improve our understanding of the role of CNVs in conifer trees adaptation and evolution and complement the findings generated in this study, we propose some future research directions that could be considered.

NGS technologies allow the identification of CNVs even if a reference genome is not available and provide more complete information about CNVs than hybridization based methods (Alkan *et al.*, 2011). However, the size and structure of conifer trees genome make it difficult to analyze a large number of individuals using NGS but this technology can be useful to explore CNV features and variation within a single genome. In an ongoing project, we proceeded to sequence haploid genomes isolated from the seeds of *Pinus taeda* (loblloly pine). Our primary goal is to examine SVs in general (including balanced and unbalanced variations) in other conifer species. The wealth of information generated by the NGS technology used will allow us to i) detect SVs in coding and non-coding regions; ii) estimate the size of SVs present in the genome and iii) identify the molecular mechanisms involved in SV formation through the examination of their breaking-points. The simultaneous identification of sequence and structural variation within the same genome can also be useful to detect selection signatures for genes displaying SVs. A

comparative analysis of SVs between the genome of gymnosperms (e.g. conifers) and angiosperms (e.g. poplar or eucalyptus) trees could provide new insights into the role of SVs in the long-term evolution of species with long life-span, which could be achieved thanks to the NGS data available in public data-bases.

The study presented in this thesis was based on the analysis of multi-parental pedigree populations. One potential fruitful next step would be to investigate CNVs in natural populations. In a preliminary work, we have already applied the same approach for CNV detection described here to 2386 trees collected from natural populations of *P. glauca* in eastern Canada (Beaulieu *et al.*, 2014). The data indicate that CNVs are five times more abundant in natural populations than in pedigree populations. However, the identified CNVs are present at a low frequency in the population; we speculate that this is because of their deleterious effect, as we suspect that genic CNVs in natural populations are under strong purifying selection. The CNVs show an enrichment in genes whose functional annotations are linked to response to stress and chemical stimuli. We also carried out a preliminary association analysis between CNVs and seven environmental variables and the results suggest the involvement of CNVs in i) the response to different abiotic stresses including dehydration, salt, osmotic and light intensity stresses and ii) the regulation of organ development such as flowers and leaves.

There is also scope for further analysis of CN mutation rates. The mutation rates reported in this work reflect the pace of accumulation of germline and somatic mutations. These two types of mutations are generated by different mechanisms and have different impacts on the standing genetic variation. Estimation of separate mutation rates for germline and somatic mutations will provide a better understanding of the role of mutations in evolution and adaptation. Germline mutations affect directly the gametes and are transmitted to the next generation. Hence, they directly impact the reproductive success of parents and the viability and fitness of the offspring. Quantifying the rate of accumulation of somatic mutations is also important for several reasons including i) stress induced mutations are common in plants (Debolt, 2010); ii) the accumulation of somatic mutations during the aging of perennial organisms influences their life-long reproductive output and the senescence process (Petit and Hampe, 2006; Williams, 2009) and iii) the respective rates of accumulation of somatic mutations for pathogens and trees (the hosts) determine the ability of trees to survive epidemics and the outcome of the coevolution process (Fenning, 2014). Estimation of the mutation rate for somatic *de novo* CNVs can be achieved by a

thorough analysis of a small number of individuals, for example by comparative genotyping of several samples from the same tree that are collected at different levels of the canopy and at different nodes on a single branch.

In the present study, we reported that the majority of inherited CNVs are transmitted in violation of Mendelian expectations but we did not identify the underlying causes responsible for these TDs. Transmission distortions are caused by mechanisms taking place during the gametic and/or zygotic stages of development (Huang *et al.*, 2013). Identifying the origin of TDs particularly for CNVs with large impact on phenotypes has a considerable practical importance. TDs with a gametic origin impact directly the reproductive capability and the competitiveness of gametes. A better characterization of these mechanisms can potentially help identify the best progenitors and seed lots to use for artificial propagation in commercial nurseries and plantations. On the other hand, TDs with a zygotic origin determine the viability of embryos and their study may help improve the protocols of trees propagation *via* somatic embryogenesis. Admittedly, the identification and characterization of TDs is challenging in trees mainly because of their long generation time but creative approaches relying partially on model organisms can be conceived. The F-box gene identified in this study as being transmitted in violation of Mendelian expectations in both *P. glauca* and *A. thaliana*, is suspected of being involved in embryo development. A more detailed functional analysis of this TD system can start by studying it in *A. thaliana*.

Practical applications have yet to be developed from the study of CNVs in forest trees. We have already eluded to using the knowledge developed in this project for assessing genetic resources for TDs. The literature also indicates that CNVs can be involved in the control of diverse quantitative traits of economic importance in crop plants (Zmienko *et al.*, 2014) and play an important role in plant resistance to biotic and abiotic stresses (Cook *et al.*, 2012; Maron *et al.*, 2013; Li *et al.*, 2014). Additional studies of the functional impact of CNVs whether through QTL, landscape genomics, association with environmental variables or knockout analyses can shed light on the contribution of CNVs to phenotype regulation. This knowledge could contribute to producing applied impacts in conjunction with the development of molecular markers for i) tree breeding programs, ii) the conservation efforts of the boreal forest and iii) the management of genetic resources in response to environmental changes.

4.3. References

Alkan C, Coe BP, Eichler EE (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.

Alonso A, Marsal S, Tortosa R, Canela-Xandri O, Julià A (2013). GStream: Improving SNP and CNV coverage on genome-wide association studies. *PLoS One* **8**: e68822.

Beaulieu J, Doerksen T, Clément S, Mackay J, Bousquet J (2014). Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity* **113**: 343–352.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA *et al.* (2013). Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**: 1492–1497.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M *et al.* (2014). Extensive copy-number variation of young genes across stickleback populations. *PLoS Genet* **10**: e1004830.

Chevin LM, Hospital F (2006). The hitchhiking effect of an autosomal meiotic drive gene. *Genetics* **173**: 1829–1832.

Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P *et al.* (2013). Massive screening of copy number population-scale variation in *Bos taurus* genome. *BMC Genomics* **14**: 124.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM *et al.* (2012). Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* **338**: 1206–1209.

Debolt S (2010). Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* **2**: 441–453.

Fenning T (2014). *Challenges and Opportunities for the World's Forests in the 21st Century*, Forestry Sciences 81, Springer: Dordrecht, Heidelberg, New York, London.

Fidantsef AL, Britt AB (2011). Preferential repair of the transcribed DNA strand in plants. *Front Plant Sci* **2**: 105.

Heinäniemi M, Vuorenmaa T, Teppo S, Kaikkonen MU, Bouvy-Liivrand M, Mehtonen J *et al.* (2016). Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots. *eLife* **5**: e13087.

Hendriks G, Calléja F, Besaratinia A, Vrieling H, Pfeifer GP, Mullenders LHF *et al.* (2010). Transcription-dependent cytosine deamination is a novel mechanism in ultraviolet light-induced mutagenesis. *Curr Biol* **20**: 170–175.

Huang LO, Labbe A, Infante-Rivard C (2013). Transmission ratio distortion: Review of concept and implications for genetic association studies. *Hum Genet* **132**: 245–263.

Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R, Shapero MH *et al.* (2010). Population-genetic nature of copy number variations in the human genome. *Hum Mol Genet* **19**: 761–773.

Kumasaka N, Fujisawa H, Hosono N, Okada Y, Takahashi A, Nakamura Y *et al.* (2011). PlatinumCNV: a Bayesian Gaussian mixture model for genotyping copy number polymorphisms using SNP array signal intensity data. *Genet Epidemiol* **35**: 831–844.

Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z *et al.* (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**: 1045–1052.

Lynch M (2010). Evolution of the mutation rate. *Trends Genet* **26**: 345–352.

Lynch M (2011). The lower bound to the evolution of mutation rates. *Genome Biol Evol* **3**: 1107–1118.

Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H *et al.* (2007). Breaking the waves: Improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* **8**: R228.

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ *et al.* (2013). Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proc Natl Acad Sci USA* **110**: 5241–5246.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL *et al.* (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol* **159**: 1295–1308.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C *et al.* (2011). Mapping copy number variation by population scale genome sequencing. *Nature* **470**: 59–65.

Park C, Qian W, Zhang J (2012). Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep* **13**: 1123–1129.

Petit RJ, Hampe A (2006). Some evolutionary consequences of being a tree. *Annu Rev Ecol Evol Syst* **37**: 187–214.

Sollier J, Stork CT, García-Rubio ML, Paulsen RD, Aguilera A, Cimprich KA (2014). Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Mol Cell* **56**: 777–785.

Swanson-wagner RA, Eichten SR, Kumari S, 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.

Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP *et al.* (2015). Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J* **83**: 189–212.

Williams CG (2009). *Conifer Reproductive Biology*, Springer: Heidelberg, London, New York.

Yu P, Wang C, Xu Q, Feng Y, Yuan X, Yu H *et al.* (2011). Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics* **12**: 372.

Zmienko A, Samelak A, Kozlowski P, Figlerowicz M (2014). Copy number polymorphism in plant genomes. *Theor Appl Genet* **127**: 1–18.