HERITABLE SOMATIC MUTATIONS ACCUMULATE SLOWLY IN SITKA SPRUCE
BUT INCREASE THE PER-GENERATION MUTATION RATE CONSIDERABLY

by

Vincent Charles Terrence Hanlon

B.Sc. (Hons.), Queen’s University at Kingston, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in
THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Forestry)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2018

© Vincent Charles Terrence Hanlon, 2018
Abstract

The rate and biological significance of heritable somatic mutations accumulating with vegetative growth or age in trees has long been a subject of debate. Somatic mutation rates are unknown for conifers, some of which can reach exceptional sizes and ages. I investigated the somatic mutation rate in the conifer Sitka spruce (*Picea sitchensis* (Bong.) Carr.) by sequencing 113.1 Mb of nuclear DNA from the top and bottom of 20 exceptional old growth trees averaging 76 m in height. I estimate a somatic base substitution rate of $2.7 \times 10^{-8}$ per base pair within a generation. Since this is comparable to germline mutation rates in the fastest-mutating eukaryotes, somatic mutations increase the per-generation heritable mutation rate in conifers considerably; but because these Sitka spruce are old, the increase is very small on an annual basis. I argue that although somatic mutations raise genetic loads in conifers, they generate important genetic variation, and they may also enable selection among cell lineages within individual trees.
Lay Summary

Because trees live so long they should evolve slowly; and yet, paradoxically, tree populations are well adapted to local climates and evolve effective responses to changing stresses such as herbivorous insects. A plausible solution is that trees produce more mutations than other species because they do not have a segregated germline, that is, a tree’s reproductive cells come from the same cells that build the tree’s trunk, branches, and leaves. Heritable mutations that arise in non-reproductive cells (“somatic mutations”) could alter how trees evolve by generating more variation for natural selection every generation or even by enabling selection among the branches or cells of an individual tree. I find that somatic mutations are common enough in trees that they could provide an answer to the paradox, especially over longer periods of time.
The research topic originated with Dr. Sally Aitken, and both Dr. Aitken and I developed the details of the experimental design. I conducted the fieldwork, the labwork, the bioinformatics, the analysis, the interpretation, and the writing. Dr. Aitken’s advice guided every stage of the process, and Dr. Sally Otto contributed advice on Sanger verification and the calculation of the mutation rate.
Table of Contents

Abstract ................................................................................................................................. ii
Lay Summary .......................................................................................................................... iii
Preface ....................................................................................................................................... iv
Table of Contents ...................................................................................................................... v
List of Tables ........................................................................................................................... vii
List of Figures ........................................................................................................................ viii
Acknowledgements .............................................................................................................. ix

Chapter 1: Introduction ............................................................................................................. 1
  1.1 What are somatic mutations? .......................................................................................... 1
  1.2 Causes ............................................................................................................................ 2
  1.3 Occurrence and frequency ......................................................................................... 4
  1.4 Evolutionary implications .............................................................................................. 7
  1.5 Context for the experimental design ............................................................................. 10

Chapter 2: The Somatic Mutation Rate in Sitka Spruce ......................................................... 11
  2.1 Introduction .................................................................................................................... 11
  2.2 Results ........................................................................................................................... 14
    2.2.1 Base substitutions: identification ......................................................................... 14
    2.2.2 The IGV method .................................................................................................... 16
    2.2.3 Base substitutions: characteristics ....................................................................... 16
    2.2.4 The mutation rate .................................................................................................. 19
  2.3 Discussion ...................................................................................................................... 20
  2.4 Methods ....................................................................................................................... 25
List of Tables

Table 2.1 Mutations and their characteristics ................................................................. 18
Table A.1 Data and references for Figure 2.3 ................................................................. 49
Table B.1 Geographic coordinates of trees sampled ....................................................... 50
List of Figures

Figure 2.1 Schematic diagram of the protocol for detecting somatic mutations in 20 trees........ 15
Figure 2.2 Schematic diagram of the bark samples and their consensus genotype ................... 15
Figure 2.3 Mutation rates ($\mu$) per generation (A) and per year (B) for multicellular species with sufficient data available ................................................................. 21
Acknowledgements

First, thanks to my supervisor Dr. Sally Aitken for her patience in teaching a disaffected mathematics student to do biology, and for the many ideas and people she recommended to me. I am especially grateful for the offer of such a compelling research question, particularly now that we know it leads somewhere interesting. I also appreciate the help and commentary of my committee members Dr. Rob Guy and Dr. Sally Otto, most notably Dr. Guy’s information on meristems and Dr. Otto’s work on the calculation of the size of the search space.

I am indebted to all members of the Aitken lab for answering thousands of questions, especially during the first year of research, for volunteering plenty of useful advice, for making suggestions about labwork, logistics, and bioinformatics, for friendship, and for supplying me with snacks. Thanks to Joane Elleouet and Dr. Dragana Obreht Vidakovic for help and commiseration in the lab, to Jon Degner for introducing me to UNIX and for continuous aid when all I could do was type `cd` and `ls`, to Dr. Pia Smets for instilling in me a great respect for receipts, and to Dr. Ian MacLachlan for practicality and an early trip to Carmanah.

I gratefully acknowledge Lea Zhecheva’s stoicism in the face of a broken radio, rain, and over-large boots on an abortive attempt at fieldwork on Haida Gwaii. Similarly, I thank Jon once again for help collecting samples in Carmanah, TJ Watt for his wonderful photos of trees despite a cloud of mosquitoes, and James Luce, Matthew Beatty, and Ryan Murphy for their herculean efforts while climbing 23 of the tallest trees in Canada so expertly.

Thanks to the Varsity Outdoor Club for sore muscles and a burnt nose on Mondays and for an unfocussed mind on Fridays. I also appreciate the forbearance of my friends and family whenever I mentioned somatic mutations in otherwise civilized conversation. I am sincerely grateful for the support you have all given me during the past two and a half years.
1.1 What are somatic mutations?

All mutations outside the germline are somatic mutations: in trees, these are mutations that occur in any structure except sporangia, their immediate precursors, and the cell lineages they produce by meiosis for sexual reproduction. Somatic mutations have the potential to make each tree a mosaic of genetic variation that is broadly or finely partitioned according to the tissue in question. For example, a mutation in an axillary meristem could produce an entirely mutant branch, or a mutation in the cambium could propagate through a narrow radial sector in the wood within a transverse layer of xylem and phloem cells; but a slow-dividing mutant cell in a growing leaf might give rise to just a few mutant daughter cells.

Somatic mutations can be inherited in taxa that lack a segregated germline or that reproduce asexually. However, since the popularization of Weismann’s doctrine that evolution is confined to the germline (Weissman 1892, cited in Buss 1983), as well as Magni’s observation that meiosis produces more mutations than mitosis in yeast (1963), it has been widely assumed that most heritable mutations in eukaryotes are germline mutations (Buss 1983). Yet unlike in mammals, the inheritance of somatic mutations is the rule rather than the exception in plants, fungi, and many animals (Buss 1983). Because most model organisms are small and short-lived, often with a segregated germline, the effects of heritable somatic mutations in large, long-lived organisms such as trees may be overlooked.

Researchers since the early 1980s have hypothesized that somatic mutations may be essential to the evolution of trees (Whitham and Slobodchikoff 1981; Gill et al. 1995; Petit and Hampe 2006; Schultz et al. 2011). Because apical meristems perform primary growth by
repeated cell division, they are related to the zygote that established the tree by a long series of mitoses across decades, centuries, or even millennia. They can accumulate somatic mutations as they grow, from replication errors or other DNA damage, and they may produce gametes containing the mutant alleles. If heritable somatic mutations increase the per-generation mutation rate significantly, then tree populations should experience a concomitant increase in both genetic load and standing genetic variation (Whitham and Slobodchikoff 1981). The latter is particularly interesting because it might help trees overcome the adaptive handicap of long generation times.

1.2 Causes

It is generally thought that somatic mutations in plants originate from DNA replication during mitosis, and the hypothesis that the mitotic rate predicts the per-generation mutation rate across plants of different sizes or ages has been repeatedly proposed (Antolin and Strobeck 1985; Scofield and Schultz 2006; Lanfear et al. 2013). However, this hypothesis is not yet well supported (Gaut et al. 2011). The literature on human cancer identifies UV light, cigarette smoke, and the deamination of methylated cytosine as particularly important sources of somatic mutations (see Alexandrov et al. 2013). Analogously, in trees, mutagens and spontaneous damage could cause somatic mutations that accumulate steadily with time rather than successive cell divisions. If this is so, the implications for trees that may live for thousands of years are intriguing.

An insightful analytic model by Gao et al. (2016) shows how difficult it is to discern the causes of mutation. In particular, under plausible conditions, DNA damage accruing with absolute time can couple the mutation rate with the rate of cell division if damaged sites are repaired quickly—that is, if damage only escapes repair if it occurs immediately before
replication. Conversely, under more restrictive conditions, replication errors can accumulate as mutations at a constant rate per time or nearly so. Since the evidence that replication drives mutation comes from indirect correlations between neutral substitution rates and inferred rates of mitosis (Baer et al. 2007; Smith and Donoghue 2008; Arnheim and Calabrese 2009; Lanfear et al. 2013; Bromham et al. 2015), and since such relationships are easily confounded, the causes of mutations in plants may not be as well established as the literature on the subject suggests.

Knowledge of the causes of mutations could help predict variation in the somatic mutation rate among species and populations. If somatic mutations are not caused by replication, then factors like the proportion of methylated sites, UV exposure, or metabolic rate could predict mutation rates. Conversely, if replication is the primary culprit, we might expect ancient but stunted species (e.g. high-elevation subalpine larch (Larix lyallii) or bristlecone pine (Pinus longaeva)) to have lower per-generation mutation rates than large but comparatively short-lived species (e.g. valley-bottom Sitka spruce). Morphological characteristics such as the architecture of apical meristems and the average length of the cells they produce—or, for clonal species, the tissue of origin of the ramet—might also be predictive of somatic mutation rates, all else being equal.

Of course, all else is not equal. Differences in average generation time, effective population size (Lynch 2010b), genome size (Bromham et al. 2015), and other unknown mechanisms, such as variation in the susceptibility to infection (Ranade et al. 2014), can inform or complicate such inferences. Differences in genomic architecture and composition among taxa could also alter mutation rates. For example, members of the Pinaceae have more interspersed repeats (40–86% of the genome; Wegrzyn et al. 2014) than angiosperms such as Arabidopsis (14%; Arabidopsis thaliana; The Arabidopsis Genome Initiative 2000). Conversely,
chromosomal rearrangements are less common in the Pinaceae than in angiosperms, along with duplications (Bowers et al. 2003), with 3 times more collinearity among homologous genes shared by *Picea glauca* and *Pinus taeda* than those shared by Arabidopsis and *Populus trichocarpa* despite similar divergence dates around 100 million years ago (Wang et al. 2009; Pavy et al. 2012; Wang et al. 2012). Such differences in genomic architecture could alter mutational processes, especially since repeats such as transposable elements can be mutagenic (Kidwell and Lisch 1997).

### 1.3 Occurrence and frequency

Conspicuous phenotypic variation within individuals provided the first evidence of somatic mutations in plants. Agriculturally-important examples include the origins of pink grapefruits and nectarines from mutant sectors of wild-type grapefruit trees (*Citrus x paradisi*; Hartmann and Kester 1975) and peach trees (*Prunus persica*; Philp and Davis 1936, cited in McGregor 1976), as well as the high frequency of albinism in embryos produced by old mangroves (*Rhizophora mangle*; Klekowski and Godfrey 1989). Induced or spontaneous mutations affecting the colour or shape of leaves or flowers are also well documented throughout the last century (e.g. Punnett 1919, cited in Gates 1920; Lindstrom 1933; Cuany et al. 1958, cited in Mericle and Mericle 1965; Heiken 1960; Whitham and Slobodchikoff 1981). More recently, direct genetic evidence has come from estimated rates of somatic mutation at microsatellite loci in a variety of trees (Pla et al. 2000; Cloutier et al. 2003; Lian et al. 2004; O’Connell and Ritland 2004; Burg et al. 2006; Helmersson et al. 2008; Ally et al. 2008; Kuchma et al. 2011; Ranade et al. 2014), although such rates can be difficult to interpret because microsatellite loci are generally non-functional and because of rate variation among loci.
Nonetheless, new morphological work suggests that heritable somatic mutations may be comparatively uncommon. Assuming they are caused by replication errors, a low rate of mitosis would be evidence for a low somatic mutation rate. A recent study found exactly this in the meristems of tomato (*Solanum lycopersicum*) and Arabidopsis during branching (Burian et al. 2016): that is, few cell divisions are necessary to produce an axillary bud from an apical meristem. If we extrapolate to conifers, however, it may not follow that there are few cell divisions—and hence perhaps few mutations—in apical meristems, since the production of long internodes or of apical buds could conceal many more cell divisions than branching events.

A study of Arabidopsis pursuing a similar line of evidence found that small plants undergo only 13% fewer meristematic cell divisions than large plants (Watson et al. 2016). Sequencing the parents and progeny of large and small mismatch-repair-deficient mutants, the authors also showed that there was no significant difference among plant types in the number of new mutations. Again, as detailed above, we might infer that large increases in plant size require only small increases in the number of cell divisions, and if somatic mutations result from replication errors, this could limit the somatic mutation rate even in very large trees.

Little purchase on the problem was technically feasible before next-generation sequencing methods became widely available, and in 2016, Xie et al. reported the first per-generation mutation rate in a tree from sequence data. They estimated $7.77 \times 10^{-9}$ substitutions per base pair per generation from the leaf of a young peach tree to the leaves of its very young progeny, including germline mutations. Their mutation identification protocol likely excluded most somatic mutations by requiring that each mutation be unique among progeny from different branches, but an additional application of their method showed at most a modestly elevated
mutation rate even when the parent was 200 years old—that is, it did not indicate a strong effect of somatic mutation.

The most up-to-date work on the somatic mutation rate in trees is a new estimate for an angiosperm, a 234-year-old common oak (*Quercus robur*) in Switzerland (Schmid-Siegert et al. 2017). Whole-genome sequencing of two branches revealed 17 heterozygous mutations and a base substitution rate of $4.2\text{--}5.2 \times 10^{-8}$ per base pair within a generation. This is the highest mutation rate reported for a multicellular species per generation. But although an oak undergoes much more vegetative growth during its lifetime than Arabidopsis, for example, the difference between their mutation rates is not proportionately large. This comparison suggests either that oak meristems perform fewer mitoses per meter of growth than Arabidopsis or perhaps that replication errors do not drive somatic mutation. Regardless, if oaks are similar to conifers—despite the fact that angiosperms evolve much faster than conifers in general (De La Torre et al. 2017; see below)—then we have a broad idea of how the somatic mutation rate behaves: conifers may generate many more new mutations each generation than most taxa, but we expect a low rate of mutation per year because they are so long-lived.

Variation among sites or regions within a genome—for instance, associated with recombination rates (Lercher and Hurst 2002)—is an important caveat to any discussion of mutation rates. Not only are mutations more common in some nucleotide contexts, but mutation rates in different contexts are also differently correlated with replication time, recombination, and other processes, at least in mammals (Hodgkinson and Eyre-Walker 2011). Moreover, in humans, different tissues can have vastly different somatic mutation rates (Lynch 2010a). Any study of the somatic mutation rate must therefore contend with possible heterogeneity within species.
1.4 Evolutionary implications

Research on somatic mutations is motivated by the potential consequences of an unusually high dose of heritable mutation for large or long-lived species. Although many hypotheses exist, few have been thoroughly tested. It has been suggested that somatic mutations may permit trees to fine-tune local adaptation to current or changing climates (Whitham and Slobodchikoff 1981), maintain a coevolutionary arms race with insect herbivores (Whitham and Slobodchikoff 1981), create a branch-by-branch mosaic of plant defenses (Whitham and Slobodchikoff 1981), allow natural selection among branches and cell lineages within individual trees (Gaul 1964), contribute to genetic load (Klekowski 1988), or force the evolution of outcrossing mating systems (Scofield and Schultz 2006). There is some theory on the latter and even more on selection within the individual (Klekowski and Kazarinova-Fukshansky 1984; Otto and Orive 1995; Otto and Hastings 1998; Pineda-Krch and Fagerstrom 1999; Orive 2001; Folse and Roughgarden 2012), as well as on the implications of somatic mutations for senescence (Klekowski 2003; Ally et al. 2008; Brutovská et al. 2013; Groot and Laux 2016), but molecular studies are restricted to plant defense mosaics (Padovan et al. 2013).

In order to understand many of these hypotheses, more detailed knowledge is required of the physiological processes that determine how somatic mutations spread through a tree and the effects they may have. For instance, selection among cell lineages within apical meristems is predicted to be most efficient if both the number of apical initials and the number of mitoses is high (Klekowski and Kazarinova-Fukshansky 1984), if cells are haploid, and if the meristem is unstructured (Otto and Orive 1995). Although the basic organization of conifer meristems shows that they fulfill few of these requirements (Evert 2006), the implications for selection within the
individual will be unknown until meristems are described in finer detail. And even if meristem organization is not conducive to such selection, the physiology of branching events (Burian et al. 2016) might make selection among cell lineages efficient nonetheless.

Similarly, the efficiency of selection among branches (Antolin and Strobeck 1985) depends both on the genetic homogeneity of individual branches (produced by patterns of cell division in meristems) and also on the capacity of vigorous branches to generate or requisition more resources and produce more cones or flowers than weaker branches. For example, although strong apical dominance limits the independence of individual branches in conifers, mutant branches could obtain more or less resources for cone production by photosynthesizing more efficiently or across greater leaf area, by producing cytokinin phytohormones to draw in more resources (Logan et al. 2013; Rob Guy, personal communication March 2018), by surviving stresses such as defoliation (Edwards et al. 1990), or simply by growing faster or slower than their neighbours. All of these processes are determined by physiology, not only by the amount of genetic variation within a tree.

All the same, data on somatic mutation rates is also needed to evaluate all the evolutionary hypotheses listed above. So far, the best available data for most plants are indirect inferences from neutral substitution rates. Simple but compelling arguments show that the neutral substitution rate is equal to the average heritable mutation rate per year, including mutations from meiosis, regardless of effective population size (Kimura 1983) or linked selection (Birky and Walsh 1988; but see Phung et al. 2016). If most heritable mutations in trees are somatic, then the neutral substitution rate is an informative upper bound for the somatic mutation rate. Estimates of the neutral substitution rate are confounded, however, by the difficulty of identifying truly neutral loci (Shields et al. 1988). The neutral substitution rate also
incorporates historical alterations to the mutation rate due, perhaps, to variation in generation time, the appearance of mutator alleles, or changes in the efficiency of selection on the mutation rate itself, and this could decouple past substitution rates from the current mutation rate.

A number of studies have estimated that substitution rates in gymnosperms are $0.5–1 \times 10^{-9}$ per base pair per year, with variation among genera, and higher in angiosperms at $4–6.5 \times 10^{-9}$ per base pair per year (Willyard et al. 2006; Buschiazzo et al. 2012; Chen 2012; De La Torre et al. 2017). The inferences that can be drawn from these data are limited. Perhaps angiosperm trees have a higher somatic mutation rate than conifers of a similar age and size because of differences in metabolism, meristematic architecture, genome composition, or any number of other factors. *Picea*, the genus considered in this study, displays a faster substitution rate than most other gymnosperms, so we may suspect that it accumulates more somatic mutations as well. However, the main implication is that since conifers have low neutral substitution rates compared with other plants, somatic mutations do not compensate for the slow pace that long generation times impose on evolution.

If we want to evaluate hypotheses about the evolutionary effects of somatic mutations, then we need better estimates than neutral substitution rates provide. Phenotypic observations and neutral substitution rates are a poor substitute for direct evidence. The available estimates of mutation rates in trees are encouraging, but information from multiple species of angiosperms and gymnosperms, as well as knowledge both of the causes of somatic mutations and of the variation in the somatic mutation rate with life history traits, is necessary before a complete picture of their evolutionary significance emerges.
1.5 Context for the experimental design

To provide an estimate of the somatic mutation rate in a gymnosperm, I sequenced tissue from the top and bottom of exceptionally tall Sitka spruce. I chose this approach over sequencing two branch tips from each tree because the number of mutations between the top and bottom is a better approximation of the total contribution of heritable somatic mutations to the per-generation mutation rate, because samples from two branch tips cannot be separated by as much vegetative growth (i.e. as many mitoses), and because obtaining samples from the ends of branches on tall trees would be challenging. I also chose to focus on base substitutions rather than indels, rearrangements, or other mutations, since base substitutions should be both more common and easier to identify against the highly-fragmented reference genome available (Birol et al. 2013)—making detection more straightforward.

Because most literature on the subject suggests that somatic mutations are caused by replication errors (see above), and since greater height could imply a greater number of mitoses in primary apical meristems, I chose exceptionally tall rather than exceptionally old trees for this research. Sitka spruce is among the tallest tree species in the world, occurs locally, and has better genomic resources than the other local tall tree, Douglas-fir (*Pseudotsuga menziesii*). I sampled trees from the well-known population of seral, old growth Sitka spruce on the floodplain of the Carmanah River valley (Carmanah Walbran Provincial Park), which is maintained by regular erosion when the river changes course and which contains the tallest tree in Canada (Little et al. 2013). Although these striking trees may not be representative of the average Sitka spruce, their height and age increase the odds of observing somatic mutations at all; and this advantage is necessary given the technical challenges of detecting such very rare events.
Chapter 2: The Somatic Mutation Rate in Sitka Spruce

2.1 Introduction

Evolution depends on heritable mutations. Because new mutations increase genetic variation, can speed up adaptation, and also contribute to genetic load, any process that might alter the heritable mutation rate is of fundamental interest in evolutionary biology. The inheritance of somatic mutations is one such process in many animals and in all plants (Buss 1983). Without a segregated germline, mutations that arise in a plant’s stem-cell-like apical meristems during vegetative growth can be propagated through new or growing branches, incorporated into the gametes those branches produce, and inherited by offspring.

Long-lived or large trees could have drastically elevated mutation rates if age or successive mitoses drive the accumulation of somatic mutations (Whitham & Slobodchikoff 1981). This possibility has inspired a variety of tantalizing hypotheses: Within a single tree, somatic mutations could permit selection among cell lineages to purge deleterious mutations (Gaul 1964), expose different combinations of new mutations to selection in offspring derived from different branches, and slow adaptation in insect herbivores by confronting them with a genetic mosaic of plant defenses (Whitham & Slobodchikoff 1981; but see Folse and Roughgarden 2012). Within a species, the extra genetic variation that somatic mutations generate could permit closer adaptation to local or changing conditions (Whitham & Slobodchikoff 1981), increase genetic load and promote outcrossing mating systems (Klekowski 1988; Scofield &

1. Chapter 2 is a manuscript for publication to which Sally Aitken and Sally Otto also contributed, so “we” is used rather than “I”.
and allow for coevolution with short-lived insect herbivores (Whitham & Slobodchikoff 1981). There is some evidence that the somatic mutation rate is low per year and—at most—high per generation in conifers (Petit and Hampe 2006). Namely, the neutral nucleotide substitution rate—which equals the mutation rate per year, irrespective of effective population size (Kimura 1983) or linked selection (Birky & Walsh 1988)—is lower in conifers than the per-generation mutation rate in annual rice and Arabidopsis (De La Torre et al. 2017; Ossowski et al. 2010; Yang et al. 2015). Thus, somatic mutations should contribute more to per-generation processes in conifers (e.g. selection within the individual) than to per-year processes (e.g. arms races with herbivores). A recent first estimate of the somatic mutation rate between two branch tips of a 234-year-old common oak (Quercus robur) corroborates this interpretation for angiosperm trees (Schmid-Siegert et al. 2017), but it remains to be seen whether it extends to conifers as well.

Early researchers argued that the ubiquity of mutant phenotypes in horticulture and the expectation of a high mitotic rate imply a high somatic mutation rate in trees (see Gill et al. 1995). Evidence of somatic mutations in nature soon followed. Strikingly, Klekowski and Godfrey (1989) observed that the incidence of mutant chlorophyll-deficient embryos in long-lived mangroves (Rhizophora mangle) is roughly 25 times higher than in short-lived barley (Hordeum vulgare) and buckwheat (Fagopyrum esculentum); and Edwards et al. reported a presumed-mutant Eucalyptus melliodora tree with a branch uniquely resistant to herbivory (Edwards et al. 1990). But recent work on cell divisions and apical meristem organization in Arabidopsis (Arabidopsis thaliana) and tomato (Solanum lycopersicum) suggests that increases in body size may not entail concomitant increases in the number of mitoses, implying that
somatic mutations may make a minor contribution to the total heritable mutation rate (Watson et al. 2016; Burian et al. 2016).

However, elegant theory cautions that the causes of mutations—whether mitosis or time-dependent DNA damage—are nearly indistinguishable (Gao et al. 2016). If mitosis is not the main source of mutation, then the number of mitoses may be a poor predictor of the somatic mutation rate. Regardless, conifers and angiosperms have dramatically different physical architectures, since conifers typically have a single dominant stem and little bifurcative branching, and the two taxa are separated by 300–350 million years of evolution each (Wang & Ran 2014): that is, the number of mitoses in annual angiosperms may not predict the somatic mutation rate in conifers. Although estimates of the somatic mutation rate in conifers exist for fast-mutating microsatellite loci (e.g. O’Connell & Ritland 2004), the rate of heritable somatic mutations in protein-coding regions is unknown.

Here we estimate the somatic mutation rate in 20 exceptionally tall, old-growth Sitka spruce (*Picea sitchensis*) in the range of 220–500 years old (Little et al. 2013). We targeted and sequenced 10.2 Mb of DNA, primarily in or adjacent to exons, in samples separated by an average of 74 m of vegetative growth and 3–4 branching events from the top and bottom of each tree, compared their sequences to identify somatic mutations, and verified the results using Sanger sequencing.
2.2 Results

2.2.1 Base substitutions: identification

We took 2 samples of cambium and phloem tissues (“bark samples”) from the base of each tree as well as a sample of young foliage from as high in the crown as possible, divided into 2 technical replicates (Figure 2.1). To include only heritable mutations, the bark samples were taken from opposite sides of the trunk. Because their most recent common ancestor is a cell in the centre of the tree, near to the tree’s meristem when it was a young seedling, this guarantees that any genotype shared by both bark samples represents the ancestral state, that is, the genotype young seedling itself (Figure 2.2). We extracted DNA separately for all four samples from each of the 20 trees, sequenced reduced-representation libraries that targeted exon-rich regions totaling 10.2 Mb, aligned the data to the draft genome of a closely-related natural hybrid (Picea glauca x engelmannii; Birol et al. 2013; Warren et al. 2015), and called and filtered SNPs to obtain a set of high-confidence variants. Overall, we obtained 276 Gb of Illumina sequence data, with a read depth of 34 over the target region and 71% of target bases covered to a depth of at least 5 in all four samples from each tree.

Comparing consensus bark genotypes with consensus foliage genotypes within each tree yielded a conservative pool of 24 candidate base substitutions, of which Sanger sequencing verified 3, discarded 16 as false positives, and was inconclusive for 5. We used the Integrative Genomics Viewer (IGV; Robinson et al. 2011) to manually examine raw sequence alignments (see The IGV method, below), identifying an additional 3 likely mutations from a broader pool of 141 candidate base substitutions procured by relaxing filtering criteria. Sanger sequencing verified 2 of these 3 and discarded the remaining false positive, bringing the total number of verified base substitutions to 5.
Figure 2.1 Schematic diagram of the protocol for detecting somatic mutations in 20 trees. DNA from samples and replicates was extracted separately. Illustration by Matt Strieby.

Figure 2.2 Schematic diagram of the bark samples and their consensus genotype. To infer the genotype of the tree when it was a young seedling, we discard loci for which the genotypes of the two bark samples differ, thus eliminating non-heritable mutations in the vascular cambium.
2.2.2 The IGV method

To assess candidate mutations from the conservative pool that failed either PCR amplification or Sanger sequencing despite multiple attempts, as well as to pre-screen the broader pool of 141 candidate base substitutions, we developed a set of criteria based on Keightley et al. (2014) to distinguish between true and false positives based. In brief, we rejected a candidate mutation if unfiltered, aligned reads showed 3 haplotypes or if the mutant allele was present in putatively wild type samples (see Methods). We tested the criteria by making predictions for the original 24 candidate base substitutions and then Sanger sequencing them: the predictions were supported in every case.

Based on the IGV method, we discarded all 5 of the candidate mutations we could not Sanger sequence. Raw alignments for 3 showed heterozygotes with roughly equal numbers of reads supporting each allele in all four samples. The other 2 candidate mutations were at adjacent nucleotides in the same tree and displayed 3 well-supported haplotypes, even though Sitka spruce is diploid. All four samples for this tree contained at least 1 read supporting the 2 mutant alleles, and in every case the supporting read also contained a unique allele at a nearby locus—indicating that the candidate mutation resulted from the capture and sequencing of a paralogous locus outside the draft reference.

For the broader pool of 141 candidate base substitutions, we used the IGV method to identify all except the 3 likely base substitutions—and the 3 base substitutions from the conservative pool, which is a subset of the broader pool—as false positives.

2.2.3 Base substitutions: characteristics

For all of the 5 verified somatic mutations, a homozygote mutated to a heterozygote containing an allele that was unique among all trees for the locus. No tree or scaffold contained
more than 1 mutation. Three of the 5 base substitutions were G/C to A/T transitions and 2 of these 3 occurred at C-containing dipyrimidine sites (Table 2.1), which are known to be susceptible to UV-induced mutations (Friedberg et al. 2006). Moreover, 2 of the 5 mutations were non-synonymous and 1 was synonymous, although the remaining 2 did not align well to the hybrid spruce transcriptome and may have occurred in introns or in non-genic sequences outside the target region. The genes in which the synonymous and non-synonymous mutations occurred were predicted from a transcriptome assembly (Yeaman et al. 2014), and as such the functional consequences of these mutations are unknown.

The sequence quality at confirmed mutant loci was high. Read depth after filtering ranged from 21 to 141 in each of the four samples from a mutant tree, and no reads in the homozygous samples supported the mutant allele. Nonetheless, errors in alignment and SNP calling seem to be responsible for the large number of false positives among candidate mutations, presumably because the reference genome is fragmented and incomplete.
Table 2.1 Mutations and their characteristics. Positions given for the PG29 v4.1 reference genome.

<table>
<thead>
<tr>
<th>Contig</th>
<th>Position</th>
<th>Mutation</th>
<th>CpG site</th>
<th>C-containing dipyrimidine site</th>
<th>Effect</th>
<th>Transcriptome reading frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALWZ04S2162146.1</td>
<td>18479</td>
<td>G/C to A/T</td>
<td>yes</td>
<td>yes</td>
<td>non-synonymous</td>
<td>serine to leucine</td>
</tr>
<tr>
<td>ALWZ04S2011522.1</td>
<td>5666</td>
<td>G/C to A/T</td>
<td>no</td>
<td>no</td>
<td>-</td>
<td>no alignment</td>
</tr>
<tr>
<td>ALWZ04S1889965.1</td>
<td>2816</td>
<td>G/C to A/T</td>
<td>no</td>
<td>yes</td>
<td>synonymous</td>
<td>proline</td>
</tr>
<tr>
<td>ALWZ04S2907496.1</td>
<td>48235</td>
<td>G/C to T/A</td>
<td>no</td>
<td>no</td>
<td>non-synonymous</td>
<td>asparagine to lysine</td>
</tr>
<tr>
<td>ALWZ04S1913287.1</td>
<td>11681</td>
<td>A/T to T/A</td>
<td>no</td>
<td>no</td>
<td>-</td>
<td>incomplete alignment</td>
</tr>
</tbody>
</table>
2.2.4 The mutation rate

We estimate a somatic base substitution rate of $2.7 \times 10^{-8}$ [95% CI: $7.2 \times 10^{-9} – 7.6 \times 10^{-8}$] per base pair within a generation, calculated as the number of verified mutations from the conservative pool (3) divided by the size of the search space, with an exact Poisson confidence interval. The 2 additional verified mutations from the broader pool, which we screened using the IGV method, were excluded from this estimate because we could not assess the IGV method’s false negative rate. However, including both the conservative pool and the broader pool (all 5 verified mutations), as well as expanding the search space to account for the relaxed filters that produced the broader pool, yields a similar somatic base substitution rate estimate of $2.5 \times 10^{-8}$ [95% CI: $1.0 \times 10^{-8} – 5.8 \times 10^{-8}$] per base pair within a generation.

The cumulative search space across all trees was estimated to be 113.1 Mb for the conservative pool and 197.6 Mb for the broader pool, calculated as a sum of the number of genotypes contained in invariant sites and the number of genotypes contained in variable sites that were checked for mutations. Some filters could only be applied to variable sites, and we reduced the number of invariant sites in the search space proportionally: since 47% of all genotypes contained in variable sites passed filters and were checked for mutations, 47% of all genotypes contained in invariant sites were included in the search space. After accounting for diplody, a 7% correction was required to compensate for bias against heterozygotes in the genotype-level filters. Because verified mutations in our study consist of two homozygous genotypes (bark) and two heterozygous genotypes (needles), such bias means that the filters retain mutant genotypes at a slightly different rate than wild type genotypes (see Methods).

Somatic mutations discarded in the process of filtering are not false negatives for the mutation rate, since the corresponding search space was reduced in proportion to the number of
genotypes that did not pass filters. Nevertheless, we compared the frequency of verified mutations in the broader pool (relaxed filters) and the conservative pool (more stringent filters) to confirm that candidate mutations discarded by the conservative pool’s stringent filters were largely false positives. In general, the relaxed filters targeted a subset of the quality statistics targeted by the stringent filters, and the relaxed filters excluded less of the genome. The frequency of true mutations in the conservative pool of 24 candidate base substitutions was 7 times higher than among the new genotypes in the broader pool of 141 base substitutions.

### 2.3 Discussion

We report a somatic mutation rate in old growth Sitka spruce that is among the highest rates of heritable base substitution per generation in any eukaryote, even without mutations from error-prone meiosis (Magni 1963). Differences in age and height within a tree species (Whitham & Slobodchikoff 1981), among other effects (e.g. Sharp and Agrawal 2012), are predicted to produce broad variation in somatic mutation rates among individuals and populations. Our population of Sitka spruce has outlived the average generation time, so we expect an elevated per-generation mutation rate; and this increase could be non-linear if older trees are more or less susceptible to mutagenic infections or other stress (Ranade et al. 2014).

In contrast, the annual somatic mutation rate in Sitka spruce is remarkably low: \(7.4 \times 10^{-11}\) base substitutions per base pair per year, assuming an average age of 360 years for the trees sampled (Figure 2.3). Perhaps this is unsurprising, since height and genome size are negatively correlated with neutral substitution rates in both gymnosperms (De La Torre et al. 2017) and angiosperms (Bromham et al. 2015), and Sitka spruce is both among the tallest tree species and has a ~21 Gb genome (Birol et al. 2013). Meiosis aside, the per-year mutation rate in
Figure 2.3 Mutation rates (μ) per generation (A) and per year (B) for multicellular species with sufficient data available. Per-year rates were calculated using a reported age, an estimate of generation time, or an average of two estimates of generation time (Table A.1, Appendix A). For the trees, μ was estimated in exceptionally long-lived specimens; and for Quercus and Picea, μ excludes the contribution of meiosis, subsequent mitoses, and early growth.

Sitka spruce is one-hundredth of that reported for rice (Oryza sativa; Yang et al. 2015), for instance, as well one-third of the per-year somatic mutation rate in oak (Schmid-Siegert et al. 2017), consistent with the much lower neutral substitution rates reported for gymnosperms (De La Torre et al. 2017). Finally, we note the discrepancy of nearly an order of magnitude between the annual somatic mutation rate for Sitka spruce and the higher average neutral substitution rate reported for Picea (average of P. glauca, P. sitchensis, and P. abies; De La Torre et al. 2017). This might be explained by the addition of meiosis and subsequent mitoses to the substitution
rate, interspecific differences, or possibly by episodes of faster mutation in the history of the genus.

That the somatic mutation rate is high per generation but low per year suggests that it should be interpreted according to the timescale of the evolutionary process in question. For example, somatic mutations may explain the high genetic load of most trees (Klekowski 1988). Although there are examples of nearly loadless species, a median of 7 lethal equivalents per zygote in loblolly pine (Pinus taeda) and 10 in Douglas-fir (Pseudotsuga menziesii) are considered representative for conifers—compared with just 2 lethal equivalents per zygote in humans—and considerable genetic load has also been documented in angiosperms such as Eucalyptus regnans (Klekowski 1988). Since an increase to the deleterious mutation rate directly increases mutation load (Haldane 1937), mutation load due to the accumulation of somatic mutations likely accounts for much of the observed genetic load in trees.

Selection among or within a tree’s branches (i.e. among or within apical meristems) is likewise a per-generation (or per-cell-generation) process. Although the extent of variation within an apical meristem is unknown, our results indicate that the somatic mutation rate between two distant branches on an average Sitka spruce could well be near the human germline mutation rate (1.20 x 10⁻⁸; Kong et al. 2012). This is a difference of 4–5 base substitutions in coding regions and UTRs (untranslated regions), assuming that Sitka spruce’s transcriptome is the same size as that of Picea glauca x engelmannii natural hybrids (182.2 Mb; Yeaman et al. 2014). Indels, inversions, and structural changes, as well as further base substitutions in regulatory regions and introns, would generate additional dissimilarities. If selection on such variation is effective, it could skew the distribution of fitness effects for heritable mutations by removing variants that are deleterious at the cell or branch level (Gaul 1964). In fact, one caveat
to our estimate is that selection within or among branches may have already depressed the somatic mutation rate.

Over short timescales, however, the annual addition of heritable somatic mutations is just 0.03 mutations per year in coding regions and UTRs (in 182.2 Mb; see above). In comparison, given its per-generation mutation rate (Ossowski et al. 2010), an annual plant such as Arabidopsis could generate more than 80 times as many mutations per year in an equivalent amount of sequence. All else equal, it is improbable that such a minute boost to the mutation rate in Sitka spruce could sidestep its long generation times and let it adapt to fast-changing climates (Aitken et al. 2008) or keep pace with coevolving insect herbivores (Whitham and Slobodchikoff 1981).

Nevertheless, it is possible that the somatic mutations observed in Sitka spruce comprise an elevated proportion of beneficial alleles. If selection among branches or cell lineages within a tree has removed many deleterious mutations, then the beneficial mutation rate may be higher than it appears, and somatic mutations may do more to facilitate adaptation over short timescales than the comparison with Arabidopsis suggests. The structure of conifer apical meristems determines the importance of this process (Otto and Orive 1995). If deleterious mutations depress the rate of cell division in mutant apical initials by just 20% in a structured meristem, selection could remove as much as two-thirds of deleterious mutations (Otto and Orive 1995). Conifer meristems have more structure than the structured meristem considered by Otto and Orive (Evert 2006), however, and this may make them resistant to such selection. Alternatively, it is also possible that enough (beneficial) mutations arise early in the development of conifers—meaning they appear in most reproductive branches—that they make an outsized contribution to adaptation by appearing in gametes at a higher rate per year than expected.
Even if the adaptive contribution of somatic mutations over short timescales is negligible, somatic mutations could generate enough variation to strengthen local adaptation over centuries or generations. When adaptation requires a very rare beneficial mutation or sequence of mutations, for example, somatic mutations improve the odds that it will appear in a population, although beneficial mutations are easily lost to drift when they first arise. Selection within the individual could even promote a beneficial somatic mutation if it improves the reproductive success of the branches that bear it. Similarly, since conifers often have large effective population sizes (Dodd & Silvertown 2000), a higher mutation rate can produce considerably stronger adaptation when many loci underlie adaptive traits (Yeaman 2015).

A striking similarity between Sitka spruce and other plants is the preponderance of G/C to A/T transitions (Yang et al. 2015; Xie et al. 2016; Schmid-Siegert et al. 2017; Exposito-Alonso et al. 2018), which often result either from UV damage or the deamination of methylated cytosine at CpG sites (Friedberg et al. 2006). Although mitosis can transform such damage into mutation, mitosis does not cause either methylation damage or UV damage, and so the common doctrine that replication errors are the dominant cause of mutation in plants deserves more scrutiny than it receives. For example, if we accept predictions of the mitotic rate based on the number of mitoses during branching events in Arabidopsis (assumed to be bifurcative; Burian et al. 2016), just 18–24 mitoses from an embryonic meristem to a third- or fourth-order branch in Sitka spruce yielded 4 times more mutations than both meiosis and 30–50 mitoses do over the life of an Arabidopsis plant (Ossowski et al. 2010). Alternatively, if the number of mitoses is proportional to the length of vegetative growth, then Arabidopsis, which might be less than 0.3 m tall, produced nearly a quarter of the mutations that occurred in 74 m of Sitka spruce. In this example, then, estimates of the mitotic rate do not explain the mutation rate.
We offer the first direct estimate of the somatic mutation rate in a conifer. The somatic mutation rate is very high in large, old Sitka spruce within a generation but very low both per year and per meter of growth, and we argue that this distinction determines the relevance of somatic mutation for evolutionary processes. This interpretation supports the hypotheses that somatic mutations both cause high genetic load in trees and also provide significant genetic variation for both selection within the individual and also adaption over long timescales. It does not, however, support the hypothesis that somatic mutations allow trees to maintain adaptation under fast-changing conditions. Future research on intraspecific variation of the somatic mutation rate with tree age, height, and environment could identify useful correlates for predicting interspecific differences in the mutation rate.

2.4 Methods

2.4.1 Samples and processing

Twenty Sitka spruce trees were chosen for exceptional height along a one-kilometer stretch of the Carmanah River on Vancouver Island, Canada (48° 66’ N, 124° 69’ W; Table B.1, Appendix B). A previous study of the population (Little et al. 2013) found that the largest Sitka spruce in the Carmanah River valley are between 220 and 500 years old, and we used 360 years (the midpoint of the range) as an estimate for the age of the trees we sampled (e.g. in Figure 2.3). Two bark samples were taken from opposite sides of the base of each tree with a leather punch. The outer bark and as much of the phloem as possible were shaved off, leaving a mixture of cambium and phloem to be dried in silica beads in the field, and DNA from each sample was extracted in the laboratory using a CTAB protocol modified from Zeng et al. (2002). In particular, we omitted the additives used during grinding and added 5 μL β-mercaptoethanol
with the 3% CTAB. Each sample of needles was taken from a single twig as high in the tree as possible, similarly dried, and then divided into 2 technical replicates before DNA extraction (NucleoSpin). Sequence capture probes (Roche Nimblegen) were designed for a subset of the exon-rich hybrid spruce contigs targeted in a previous study (Suren et al. 2016). We targeted the contigs that were captured and sequenced most successfully across the most individuals using Suren et al.’s sequence capture probes on Sitka spruce (Joane Elleouet, unpublished data). Prepared libraries were barcoded and sequenced on 4 lanes of an Illumina HiSeq 4000 with 100 bp paired-end reads at Genome Quebec (Montreal, QC).

2.4.2 SNP calling and filtering

We aligned the reads from each sample to the PG29 v4.1 hybrid spruce draft reference genome (Birol et al. 2013) using NextGenMap v0.5.0 (Sedlazeck et al. 2013) and marked PCR duplicates with Picard MarkDuplicates v2.8.1. The contigs of the highly-fragmented reference genome were first assembled into pseudo-scaffolds separated by 400 “N”s to facilitate the use of bioinformatics programs downstream. We then applied HaplotypeCaller and GenotypeGVCFs from GATK v3.7 with the flag -all_sites to create a VCF file of both variable and invariant sites (Van der Auwera et al. 2013).

Because we found that genotype calls at sites with a high minor allele frequency (MAF) across all trees were less reliable, we divided sites according to a MAF threshold of 0.05 and processed low- and high-MAF sites separately, with more stringent filters on high-MAF sites. We then applied a series of unbiased filters. We discarded indels, multiallelic sites, and sites with more than 85% missing data before filtering loci by site depth (high-MAF: <6000 reads across all samples; low-MAF: <7000 reads) and genotype depth (high-MAF: >11 and <71 reads per sample; low-MAF: >11 and <76 reads) to obtain a basic set of good-quality sites (the initial set,
below). We then separated invariant and variable sites, filtered the latter by GATK’s ReadPosRankSumTest (high-MAF: \(-0.6\) and \(<0.7\); low-MAF: \(-1.3\) and \(<1.6\)) and BaseQualityRankSumTest (high-MAF: \(-0.6\) and \(<1.0\); low-MAF: \(-0.7\) and \(<1.3\)), and further filtered high-MAF sites using the strand bias tests StrandOddsRatio (\(<1.4\)) and FisherStrand (\(<12\)). We avoided stringent filtering in order to procure a generous search space for candidate mutations.

Next, we defined homozygous-reference, heterozygous, and homozygous-alternate genotypes by setting requirements for the proportion of reads supporting the reference and alternate alleles (from the hybrid spruce reference). In particular, heterozygotes were called if 40–60% of filtered reads supported each allele (high-MAF: 23–77% for low-MAF); and homozygotes or homozygote-alternates were called if \(>98\%\) of filtered reads supported one allele (high-MAF: \(>97\%\) for low-MAF). Such requirements are biased because they skew the proportion of heterozygous genotypes and hence also the proportion of true mutations (bias corrected below). We also filtered by GATK’s genotype likelihood values (high-MAF: \(>53\); low-MAF \(>15\)). The result was two final sets of high-confidence genotypes for low-MAF sites (containing \(n_{low}\) genotypes, below) and high-MAF sites (ditto \(n_{high}\)). We considered any locus for which both bark samples have the same genotype, both foliage samples have the same genotype, and the bark and foliage genotypes differ to be a candidate somatic mutation.

Candidate mutations were verified using Sanger sequencing. We designed primers for each mutation, amplified it, and sent it to the Sequencing and Bioinformatics Consortium (the University of British Columbia) for sequencing. If the mutant allele was not found in a heterozygous sample, we considered the candidate mutation a false positive. If the mutant allele
was both confirmed in a heterozygous sample and also absent in a homozygous sample, we considered the candidate mutation a true positive.

2.4.3 The IGV criteria

We examined raw BAM files for the four relevant samples to evaluate select candidate mutations. If we define the candidate mutant allele to be whichever allele is present in the heterozygous samples but not the homozygous samples, then we consider a candidate mutation to be a false positive if one of the following is true:

i. All reads that show the candidate mutant allele also show a unique allele at a nearby locus in the heterozygous samples, and at least 1 of the homozygous samples contains reads supporting both alleles.

ii. At the candidate locus and a nearby polymorphic locus there are 3 haplotypes in the heterozygous samples, each supported by multiple reads; and/or,

iii. All four samples contain at least 1 read supporting the candidate mutant allele.

2.4.4 Estimating the size of the search space

The calculation of the search space accounts for the fact that not all filters could be applied to invariant sites and corrects for the bias in the filtering process against heterozygotes. The bias against heterozygotes was observed as a decrease in heterozygosity after the genotype definitions (above) were applied to variable sites. Since a mutation consists of two homozygous bark samples and two heterozygous needles samples, this implies a bias against mutations as well. Thus, we reduced the size of the search space by 7% to bring it in line with the slightly lower pass rate of mutant loci (Appendix C).

More precisely, the search space was calculated as

\[ 2S^{-1}(n_{high}h_{high} + n_{low}h_{low}) \]
where $n_{\text{high}}$ and $n_{\text{low}}$ are the high- and low-MAF high-confidence variable genotypes, $S$ is the fraction of all genotypes that are contained in variable sites—counted in the initial set, above—and the factor of 2 adjusts for diploidy. Finally, $h_{\text{low}}$ and $h_{\text{high}}$ constitute the heterozygosity correction described above.

2.4.5 Synonymous and non-synonymous substitutions

Mutant codons were identified using BLAST (Altschul et al. 1990) to align 200 bp around each mutation to the *Picea glauca x engelmannii* transcriptome (Yeaman et al. 2014), retaining only high-confidence matches ($E$ value < $10^{-40}$). We then mapped mutations onto the longest predicted open reading frames produced by Yeaman et al. (2014) and determined the impact on the predicted protein (Table 2.1).
Chapter 3: Conclusion

I estimated a somatic base substitution rate of $2.7 \times 10^{-8}$ per base pair within a generation in 20 exceptionally tall old-growth Sitka spruce trees. This is among the highest known mutation rates for any species on a per-generation basis, but it corresponds to a remarkably low rate per year because of the great age of these trees (Figure 2.3). Most of the mutations I identified were G/C to A/T transitions (Table 2.1), as in other plants, and they may have been caused by UV damage or the deamination of methylated cytosine.

One caveat is that the mutation rate reported here is probably higher than average for Sitka spruce. Their mean generation time and mitotic rate are unknown, but if mutations accumulate linearly with time or cell divisions, we might expect most individuals to mutate at half or less of the rate estimated here because they are much smaller on average—although heritable somatic mutations would still be exceptionally common per generation. Likewise, since I sampled 20 trees, I could only sequence a small fraction of Sitka spruce’s ~21 Gb genome. Whole-genome sequencing (as in Schmid-Siegert et al. 2017) might have provided enough data to reveal a clearer excess of G/C to A/T transitions, as well as a more precise estimate of the somatic mutation rate, but would have required resources far in excess of those available.

In order to disentangle the causal effects of time and vegetative growth on the somatic mutation rate, one could study conifers of similar heights but different ages and vice versa. For example, one could choose a species such as *Abies amabilis* that grows across a wide variety of elevations at the same latitude, since high-elevation trees both grow more slowly and are exposed to more UV light. If taller, younger trees mutate more frequently than older, smaller trees, then replication is the culprit; but if the converse is true, then time-dependent DNA damage is more likely (although patterns of mitosis in apical meristems under different
conditions, as well as the joint increase in age and UV intensity at higher elevations, could confound these inferences). Alternatively, since the pattern of mitosis in the cambium is well studied (e.g. Newman 1956; Bannan 1957; Murmanis 1970; Larson 1994), it would be interesting to sequence tissue culture derived from single cambial cells on opposite sides of the trunk and evaluate whether the per-cell-division mutation rate (of non-heritable mutations) is constant for trees of different ages.

Similarly, an initial estimate of the somatic mutation rate is far from sufficient to evaluate their significance for conifer evolution. My result can only suggest that somatic mutations are likely to be of greater interest for per-generation processes than per-year processes. For instance, recent theory on alleles of small effect explores the adaptive consequences of a boost to the mutation rate such as somatic mutations provide (Yeaman 2015). In particular, Yeaman’s simulations show that for multilocus traits, transient alleles can replace each other to maintain adaptation even as some are lost to drift, and that even a modest increase to the per-generation mutation rate can enhance local adaptation of this sort. The implication is that somatic mutations might contribute to the remarkable local adaptation that trees display—but over many generations and not just many years. Unless selection within the individual drastically increases the proportion of beneficial mutations that a tree produces, it is unlikely that a tiny annual contribution of less than $10^{-10}$ mutations per base pair could do much to accelerate adaptive responses to the rapidly changing conditions created by species invasions or increasing drought frequency with climate change, for example.

The proposal that selection lowers the mutation rate until drift obscures the benefit of any further decrease is one hypothesis that is clearly locked to the per-generation mutation rate: this is the drift-barrier hypothesis of mutation rate evolution (Lynch 2010), and the implication is that
in species with large effective population sizes such as many conifers, the heritable mutation rate per generation should be low. But because the accumulation of somatic mutations results in a high per-generation mutation rate in some conifers despite their large effective population size, they may offer a chance to test the generality of the hypothesis. That is, are mutation rates similar among conifers and non-conifer species with similar effective population sizes?

For many of these investigations, better genomic resources and more knowledge of basic population genetic parameters are required. In trees these are scarce—although they could be obtained for fast-growing angiosperm trees with small genomes—but since somatic mutations are heritable in many species, perhaps some of these ideas could be evaluated in well-studied model organisms. Nonetheless, many apply only to large or long-lived species, and so trees provide one of the better systems for research. In particular, a comparison of the distribution of fitness effects for somatic mutations between large, old trees and seedlings could assess the efficiency of selection within the individual, and the methods described above could elucidate both the causes of somatic mutations and their adherence to the drift-barrier hypothesis. I hope that early estimates of the somatic mutation rate in conifers will be of use for further investigation of these problems.
References


Bannan MW. 1957. The relative frequency of the different types of anticlinal divisions in conifer cambium. Can J Bot. 35:875–884.


McGregor SE. 1976. Insect pollination of cultivated crop plants. USDA.


### Appendices

#### Appendix A

Table A.1 Data and references for Figure 2.3

<table>
<thead>
<tr>
<th>Species</th>
<th>( \mu ) per gen.</th>
<th>( \mu ) per year</th>
<th>Estimated age (years)</th>
<th>Notes</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Heliconius melpomene</em></td>
<td>2.9 x 10^{-9}</td>
<td>1.16 x 10^{-8}</td>
<td>0.25</td>
<td>Assumes 6-month lifespan and continual reproduction</td>
<td>(Boggs 1979, cited in Naisbit et al. 2001; Keightley et al. 2015)</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>5.40 x 10^{-9}</td>
<td>2.16 x 10^{-8}</td>
<td>0.25, 1.0</td>
<td>Average generation time</td>
<td>(e.g. Castiglia and Capanna 1999; e.g. Geraldes et al. 2008; Uchimura et al. 2015)</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>5.49 x 10^{-9}</td>
<td>5.49 x 10^{-8}</td>
<td>0.1, 0.25</td>
<td>Average generation time</td>
<td>(e.g. Nei et al. 1975; e.g. Keightley 1994; Schrider et al. 2013)</td>
</tr>
<tr>
<td><em>Apis mellifera</em></td>
<td>6.80 x 10^{-9}</td>
<td>4.53 x 10^{-9}</td>
<td>1.5</td>
<td>Assumes 3-year lifespan and yearly swarming</td>
<td>(Seeley 1979; Yang et al. 2015)</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>7.00 x 10^{-9}</td>
<td>1.40 x 10^{-8}</td>
<td>0.5, 1.3</td>
<td>Average generation time</td>
<td>(Thompson 1994; Falahati-Anbaran et al. 2014; Ossowski et al. 2010)</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>7.10 x 10^{-9}</td>
<td>1.42 x 10^{-8}</td>
<td>0.9</td>
<td>Harvest frequency</td>
<td>(Davis et al. 2016; Yang et al. 2015)</td>
</tr>
<tr>
<td><em>Prunus mira</em></td>
<td>9.48 x 10^{-9}</td>
<td>4.74 x 10^{-11}</td>
<td>200</td>
<td></td>
<td>(Xie et al. 2016)</td>
</tr>
<tr>
<td><em>Pan troglodytes</em></td>
<td>1.20 x 10^{-8}</td>
<td>5.00 x 10^{-10}</td>
<td>24</td>
<td></td>
<td>(Venn et al. 2014)</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>1.20 x 10^{-8}</td>
<td>4.04 x 10^{-10}</td>
<td>29.7</td>
<td></td>
<td>(Kong et al. 2012)</td>
</tr>
<tr>
<td><em>Picea sitchensis</em></td>
<td>2.65 x 10^{-8}</td>
<td>7.36 x 10^{-11}</td>
<td>360</td>
<td>Age estimated from population dendrochronology data</td>
<td>(Little et al. 2013)</td>
</tr>
<tr>
<td><em>Quercus robur</em></td>
<td>4.25 x 10^{-8}</td>
<td>1.82 x 10^{-10}</td>
<td>234</td>
<td></td>
<td>(Schmid-Siegert et al. 2017)</td>
</tr>
</tbody>
</table>
Appendix B

Table B.1 Geographic coordinates of trees sampled. The accuracy of these GPS coordinates is questionable, since they were taken under a large forest canopy and because some of the trees were close together. In some cases identical GPS coordinates are given for multiple trees. The trees are best identified by photographs, which can be provided upon request.

<table>
<thead>
<tr>
<th>Latitude (° N)</th>
<th>Longitude (° W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>48.6627</td>
<td>124.6955</td>
</tr>
<tr>
<td>48.6619</td>
<td>124.6954</td>
</tr>
<tr>
<td>48.6625</td>
<td>124.6918</td>
</tr>
<tr>
<td>48.6627</td>
<td>124.6916</td>
</tr>
<tr>
<td>48.6629</td>
<td>124.6911</td>
</tr>
<tr>
<td>48.6629</td>
<td>124.6911</td>
</tr>
<tr>
<td>48.6629</td>
<td>124.6911</td>
</tr>
<tr>
<td>48.6630</td>
<td>124.6910</td>
</tr>
<tr>
<td>48.6634</td>
<td>124.6907</td>
</tr>
<tr>
<td>48.6634</td>
<td>124.6907</td>
</tr>
<tr>
<td>48.6634</td>
<td>124.6907</td>
</tr>
<tr>
<td>48.6632</td>
<td>124.6909</td>
</tr>
<tr>
<td>48.6634</td>
<td>124.6907</td>
</tr>
<tr>
<td>48.6634</td>
<td>124.6907</td>
</tr>
<tr>
<td>48.6632</td>
<td>124.6909</td>
</tr>
<tr>
<td>48.6557</td>
<td>124.6951</td>
</tr>
<tr>
<td>48.6559</td>
<td>124.6948</td>
</tr>
<tr>
<td>48.6559</td>
<td>124.6948</td>
</tr>
<tr>
<td>48.6560</td>
<td>124.6951</td>
</tr>
<tr>
<td>48.6557</td>
<td>124.6951</td>
</tr>
</tbody>
</table>
Appendix C

The heterozygosity correction to the size of the search space is derived from the observed decrease in heterozygosity caused by the genotype-level filters. If these filters discard a proportion \( p \) of homozygous genotypes and a proportion \( p + q \) of heterozygous genotypes (where a genotype is specific to a particular locus and sample from a tree), and if all four genotypes for any locus in a tree must pass all filters if the locus itself is to pass filters, then \((1 - p)^4\) of 4-sample homozygous loci and \((1 - p - q)^4\) of 4-sample heterozygous loci would pass filters. Both quantities can be directly estimated from the numbers of heterozygous and homozygous consensus genotypes (i.e. all four samples for a tree agree) before and after filtering, so \( p \) and \( q \) can be calculated. Since mutations consist of two heterozygous and two homozygous genotypes, \((1 - p)^2(1 - p - q)^2\) of them pass filters.

Let \( n_{het} \) and \( n_{homo} \) be the known number of heterozygotes and homozygotes, respectively, among the 4-sample genotypes that are checked for mutations (i.e. that have passed all filters), and let \( n'_{het} \) and \( n'_{homo} \) be the same among the genotypes that have passed site-level filters but not yet genotype-level filters. Then,

\[
\sigma = n_{het} \frac{(1 - p)^2(1 - p - q)^2}{(1 - p - q)^4} + n_{homo} \frac{(1 - p)^2(1 - p - q)^2}{(1 - p)^4}
\]

is the size of the search space that would have resulted if all loci passed filters at the rate mutations do? Substituting expressions for \( p \) and \( q \) in terms of the four known quantities, we obtain

\[
\sigma = (n_{het} + n_{homo}) \cdot h
\]

where
\[ h = \frac{x \cdot n_{het} + x^{-1} \cdot n_{homo}}{n_{het} + n_{homo}} \]

and

\[ x = \frac{n_{homo} \cdot n'_{het}}{\sqrt{n_{het} \cdot n'_{homo}}} \]

And if we restrict \( n_{het} \) and \( n_{homo} \) (and \( n'_{het} \) and \( n'_{homo} \)) to genotypes contained in either low-MAF or high-MAF sites, then we obtain \( h = h_{low} \) and \( h = h_{high} \), respectively.