GENOME CHARACTERIZATION AND POPULATION GENETIC STRUCTURE OF WHITE PINE BLISTER RUST, *CRONARTIUM RIBICOLA*

by

Ting Pu

B.S.F., The University of British Columbia, 2011

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)

December 2014

Abstract

Rust fungi cause some of the most severe pine diseases. *Cronartium ribicola* (J. C. Fisch.), the causal agent of white pine blister rust, was introduced accidentally to North America from Europe in the late 1800s. Since then, it has devastated a large number of native, commercially valuable white pines, and is threatening alpine ecosystem stability by endangering high elevation white pines. In order to better understand the global epidemiology of this pathogen, we conducted a genome scan of a global collection of *C. ribicola* using Genotyping-by-Sequencing (GBS) to: 1) ascertain the origin and the routes of introduction of *C. ribicola*, and 2) uncover cryptic population structure of *C. ribicola* in western North America, in relation to different pine hosts, climates and landscapes. More than eight thousand single nucleotide polymorphism markers were genotyped on 192 samples of *C. ribicola* from three continents. The highest genetic and nucleotidic diversity were observed in Siberian samples, supporting the hypothesis that central Russia is the center of origin of *C. ribicola*. Diversity was reduced in all other populations and was lowest in western North America. Genetic and nucleotidic diversity were two to five times lower in western than in eastern North America. This result supports the observation of multiple introductions of the pathogen in eastern North America and contrasts with the single reported introduction in western North America. However, western populations had a larger number of rare alleles. This could represent the signature of population expansion following a bottleneck or a selective sweep. A cryptic Coast/Interior split was detected within the western cluster, most likely maintained by the scarcity of white pines in central British Columbia acting like a barrier to gene flow. Finally, western individuals with a high level of eastern admixture were discovered in two populations east of the Continental Divide. This could indicate that the eastern-western barrier to gene flow is leaky. Such information is of significance to white pine resistance breeding programs and to the monitoring of this disease.

Preface

This dissertation is an original intellectual product of the author, T. Pu. The sample collection was a collaborative work of S. Brar, A. Woods, B. Goodrich, and B. Lockman. DNA extraction, normalization and preparation for GBS experiment were done primarily by myself in the Forest Pathology Laboratory at the University of British Columbia, Point Grey campus. M.-J. Bergeron contributed to the preparation of 48 samples from eastern North America, at the Laurentian Forestry Centre, Quebec. The sequences generated by GBS were analyzed by myself, with the help of B. Dhillon, who developed the pipeline for SNP generation and filtration. All the downstream population structure analyses were my original work, with assistance from N. Feau, M. Sakalidis and A. Dale in the TAIGA Lab, UBC. I was responsible for the manuscript composition. B. Dhillon provided editorial support for this dissertation. R. C. Hamelin was the supervisory author on this project and was involved throughout the project in concept formation and manuscript edits.

Table of Contents

Abstract ii

Preface iii

Table of Contents iv

List of Tables v

List of Figures vi

Acknowledgements vii

1 Introduction 1

1.1 Literature Review 1

1.2 Genotyping-By-Sequencing (GBS) Overview 17

1.3 Objectives 20

2 Materials and Methods 21

2.1 Sampling 21

2.2 DNA Extraction 22

2.3 Experiment Design 23

2.4 Genotyping-By-Sequencing Library Construction 24

2.5 De-multiplexing and Mapping 26

2.6 SNP Calling and Filtering 26

2.7 Population Structure Analyses 28

3 Results 34

3.1 Sequencing, De-multiplexing and Mapping 34

3.2 SNP Calling and Filtering 36

3.3 Population Structure Analyses 38

4 Discussion 55

4.1 Global Introduction Pathway of Cronartium Ribicola 55

4.2 Genome-Wide Comparison Between Eastern and Western North American Populations of Cronartium Ribicola 57

4.3 Elucidating Cryptic Population Structure of Cronartium Ribicola in Western North America 60

4.4 The Suitability and Robustness of GBS for Studying Fungal Genomes 62

5 Conclusions 64

Literature Cited 66

Appendices 75

List of Tables

Table 1 Summary of sampling locations. 21

Table 2 Cronartium ribicola samples used for GBS pilot study to assess impact of DNA concentration. 25

Table 3 Statistics of SNP markers before and after filtering. 37

Table 4 Genetic diversity indices of the global Cronartium ribicola dataset, with 8,020 SNPs. 42

Table 5 Summary of all Analysis of Molecular Variance (AMOVA) results. 46

Table 6 Pairwise F_{ST} between the Korean, Russian, eastern and western North American groups of Cronartium ribicola. 47

Table 7 Genetic diversity indices of the North American dataset of Cronartium ribicola, with 4,510 SNPs. 49

Table 8 Genetic diversity indices and pairwise F_{ST} of the western North American Cronartium ribicola dataset, with 365 SNPs. 52

Table 9 Number of singletons/doubletons in the global Cronartium ribicola dataset. 75

Table 10 Number of singletons in North American Cronartium ribicola populations. 80

Table 11 Genetic diversity indices of the western North American Cronartium ribicola dataset, with 4,510 SNPs. 82

List of Figures

Figure 1 Life cycle of Cronartium ribicola, modified from Pacific Southwest Research Station. (2011). 6
Figure 2 Brief illustration of Genotyping-by-Sequencing methodology, modified from Davey et al. (2011). 19
Figure 3 North American sampling locations, blue = western North America, red = eastern North America. 24
Figure 4 Principal Component Analysis of eight samples of Cronartium ribicola genotyped with GBS at different DNA concentration, ranging from 6.25 ng to 100 ng. 35
Figure 5 Principal Component Analysis of the global Cronartium ribicola dataset, with 8,020 SNPs. 38
Figure 6 Neighbour-joining tree of the global Cronartium ribicola dataset, with 8,020 SNPs. 39
Figure 7 Discriminant Analysis of Principal Components of the global Cronartium ribicola dataset, with 8,020 SNPs. 40
Figure 8 a) Plot of k values vs. cross-validation error; b) ADMIXTURE result of the global Cronartium ribicola dataset. 41
Figure 9 Observed vs. expected heterozygosity (averaged over all loci) of each population in the global dataset of Cronartium ribicola. 43
Figure 10 Proposed spread of Cronartium ribicola across the globe from its centre of origin: Siberia, Russia. 44
Figure 11 a, b, c = Principal Component Analysis of North American samples of Cronartium ribicola analyzed with 4,510 SNPs. 48
Figure 12 Principal Component Analysis of western North American Cronartium ribicola dataset. 50
Figure 13 Histogram of Mantel test (isolation-by-distance) amongst western Cronartium ribicola populations. 51
Figure 14 Spatial Principal Component Analysis – global structure of western group of Cronartium ribicola shown on terrain map and the histogram of its associated Monte-Carlo test. 52
Figure 15 Expected heterozygosity plotted against distance to Mount Washington of western North American Cronartium ribicola populations. 53
Figure 16 Neighbour-joining tree of western North American Cronartium ribicola samples. 76
Figure 17 Neighbour-joining tree of eastern North American Cronartium ribicola samples. 77
Figure 18 Neighbour-joining tree of North American Cronartium ribicola samples. 78
Figure 19 Discriminant analysis of principal components of the North American Cronartium ribicola samples. 78
Figure 20 ADMIXUTRE result of North American Cronartium ribicola samples. 79
Figure 21 Western North American spatial Principal Component Analysis of Cronartium ribicola – global structure, laid over the distribution of white pines in western Canada. 81

Acknowledgements

I would like to express my appreciation to my supervisory committee, Dr. El-Kassaby, Dr. Ritland and Dr. Hamelin. Without your guidance, I would not have been able to accomplish this project. I am especially thankful to Dr. Hamelin, for introducing me to the field of fungal genetics and genomics, which I was never exposed to before. I am truly glad that I had the opportunity to study it. I am also grateful for the opportunity to work with such a knowledgeable group of scientists, in the TAIGA Lab. A special thank you to my mentor B. Dhillon, who has been a great help to me since the beginning of my project. You taught me all the basics about bioinformatics and generously shared all your knowledge with me. More importantly, thank you for keeping me on track. I cannot be more grateful.

This project would not have been possible without S. Brar's previous hard work, which had laid a strong foundation for me. Thank you for passing along not only your samples but also your experience and expertise in this field. I want to thank all the present and former lab mates for their technical or analytical support: A. Dale, A. Brar, S. Beauseigle, B. Lai, C. Liu, C. Tsui, D. Isidro, H. Yueh, J. Lamantia, M. Sakalidis, N. Feau, P. Herath, and S. Cervantes. I also want to thank M.-J. Bergeron, A. Woods, B. Goodrich, and B. Lockman for their help with sampling.

Thank you to my parents and partner for supporting and believing in me. This is for you.

1 Introduction

1.1 Literature Review

1.1.1 Background

Cronartium ribicola (J. C. Fisch.) is a rust fungus belonging to the order Pucciniales, Phylum Basidiomycota. It is the causal agent of white pine blister rust (*C. ribicola*) (Hunt 1983). In North America, it is an exotic pathogen that attacks five-needle pines (subsection Strobus). *C. ribicola* has been reported on all of the eight white pine species in western North America (CABI 2014). Since its introduction in North America at the turn of the 20th century, it has been associated with severe epidemics across the distribution range of its hosts (Kinloch 2003). All white pine species in North America are susceptible to *C. ribicola*, with various levels of susceptibility (Spaulding 1929). Among them, some possess high timber value such as western white pine, *Pinus monticola* Douglas ex D. Don and eastern white pine, *Pinus strobus* L. (Abrams 2001). However, because of the devastating impact of *C. ribicola*, the seed source for *P. strobus* reforestation in some parts of Ontario and the Maritime Provinces of Canada was almost eliminated (Kinloch 2003). Other white pines like whitebark pine (*P. albicaulis* Engelm.) is an ecologically important species, especially in the subalpine forest (Campbell & Antos 2000). The number of *P. albicaulis* individuals in these forests has also been dropping rapidly as a result of *C. ribicola*. Thus, it is important to understand the *C. ribicola* genetic structure as that can influence its expansion to new geographic locations and hosts. This knowledge is also critical in developing resistance and reforestation programs of the white pines.

1.1.2 History of *C. ribicola*

Spaulding (1929) believed that *C. ribicola*'s centre of origin is northern Asia. Specifically, his hypothesis was that the range stretchs from the east of the Ural Mountains of Russia, through central Siberia, to eastern Asia along the Pacific coast and to the Himalayas in the south (Leppik 1967;

Spaulding 1929). Presumably, the Asian white pines (*P. koraiensis* Siebold & Zucc., *P. pumila* (Pall.) Regel, *P. sibirica* Du Tour and *P. wallichiana* A. B. Jacks.) have co-evolved with *C. ribicola*, thus they show resistance to this disease. In Europe, co-evolution did not occur because the Weichsell Glaciation in the Eem Interglacial destroyed most of the pine forest (due to infections by *C. ribicola*), leaving behind only pines that either were non-hosts to *C. ribicola* or white pines restricted to high-elevation environments (Van Arsdel & Geils 2011). *P. strobus* is native to North America and was introduced to Europe in 1553 (Van Arsdel & Geils 2011). It was later broadly planted in the 1700s (Spaulding 1929; Moir 1924). *Ribes nigrum* L., the alternate host of *C. ribicola*, had existed in Europe long before 1750, both in nature and in gardens. The garden type - black currant is susceptible to *C. ribicola* whereas the wild type is not (Spaulding 1929). It had been hypothsized that the pathway of *C. ribicola* spread was from northern Asia to Russia (after *P. strobus* was planted in gardens), then to Europe through the movement of diseased *R. nigrum* and pine, and lastly from Europe to North America (Tubeuf 1917; Spaulding 1929; Hummer 2000).

The first official description of *C. ribicola* was made by Dietrich in 1854 in the Baltic provinces of Russia (Spaulding 1911). In 1856, the rust was reported on both *P. strobus* and *Ribes* spp. also in the Baltics (Stewart 1906). In 1861, the fungus was seen on white pine near Helsinki, Finland by W. Saellen (Moir 1924). In 1865, the disease was found on black currants in Denmark and East Prussia (Moir 1924). It was not seen in the rest of Germany until 1871 when Fischer de Waldeheim discovered it on *R. aureum* Pursh in Stralsund (Moir 1924). By 1887, it was already causing outbreaks throughout Europe (Stewart 1906), and was particularly abundant in Germany, Denmark, Belgium, Sweden, Great Britain and Holland (Spaulding 1911; Spaulding 1929). Its distribution covered the majority of Europe except the Balkan and the Hispanic peninsulas. *C. ribicola* was present on *P. monticola* all around northwestern Europe and parts of central Europe, namely Germany and Austria (Spaulding 1929). Its proposed original host is Swiss stone pine, *P. cembra* L., which was extensive in Russia and later introduced to Germany

(Stewart 1906). Globally, *C. ribicola* has been found in China, India, Iran, Japan, Korea and few other countries in Asia (CABI 2014).

1.1.2.1 Introduction to eastern North America

The introductions to eastern and western North America happened independently. Multiple introductions took place in eastern North America. In September 1906, *C. ribicola* was officially reported and documented for the first time in North America at the station currant plantation in Geneva, New York, United States, where its urediniospores were discovered on the undersurface of *Ribes* leaves. Diseased *P. strobus* was also found (Stewart 1906). Stewart pointed out that this was not the first occurrence of *C. ribicola* in North America as it had been observed in 1892 when Dr. J. C. Arthur saw the fungus on *R. aureum* collected by E. Bartholomew in Kansas (Stewart 1906). However, it was listed as *Uredo confluens* Pers at the time until Dr. Arthur corrected it to be the urediniospores of *C. ribicola* (Stewart 1906; Spaulding 1911). Another early introduction happened at Kittery Point, Maine where *R. nigrum* was imported from England (Posey & Ford 1924). Spaulding (1929) recorded occurrence of *C. ribicola* in New England in 1898, where it had been wildly distributed by 1916.

As the native North American white pines were rust-free prior to the introduction of *C. ribicola* from Europe, the pines were supposed to be more susceptible due to the absence of co-evolution. Spaulding (1929) compared the rust conditions in Europe and in North America and found that the disease was less severe in Europe because of their better sanitation practices and more scattered distribution of white pines (Spaulding 1929). The distribution of *Ribes* plants was also extremely different in these two regions. In Europe, the cultivated *Ribes* (*R. nigrum*) were more common than the wild ones, whereas in North America, it was the opposite (Spaulding 1929).

Pinus strobus seedling stocks were being imported in to the United States (US) from Europe because it was cheaper than growing them locally. Diseased 'asymptomatic' seedlings of *P. strobus* came from nurseries in Halstenbek, Germany and Ussy, Orleans, and Chatenay in France (Spaulding 1911).

Much of the German *P. strobus* stock came from a nursery centre in Belgium, where the disease was present (Spaulding 1929). This Belgium nursery centre also sold seedlings to a nursery based in Massachusetts, which was the source of the New England plantations (Spaulding 1929). Millions of white pine stocks were imported from Halstenbek alone and distributed to more than 226 locations in North America (Benedict 1981).

By 1919, *C. ribicola* had reached north to Ontario and Quebec and west to the natural limit of eastern white pine in Minnesota (Kinloch 2003; White et al. 2002). Upon the realization of the disease severity, the US government placed an embargo on the importation of white pines and *Ribes* spp. from Europe, and also on the movement of the two hosts from Eastern North America (Eastham 1923). *Ribes* cultivation is still banned or regulated in certain states (McKay 2000).

1.1.2.2 Introduction to west North America

The first documented discovery of *C. ribicola* in the western North America was made on September 10, 1921, when *R. nigrum* leaves bearing the telia were brought to a plant pathologist's office (Eastham 1923). Later, infections were found on the pines in a five-year-old *P. strobus* nursery near Stanley Park and on a branch of *P. cembra* in a garden at Point Grey (Davidson 1922). Similar to the east, a quarantine line was drawn to control the movements of the two hosts.

A disease survey was carried out in 1922. At one *P. strobus* nursery in Point Grey, all the pines imported from 1910 to 1914 were found to be infested. The owner of the nursery had imported 1,000 *P. strobus* seedlings from Ussy, France in 1910 (Davidson 1922; Spaulding 1911). Being the only documented importation of white pines from Europe and the first occurrence of *C. ribicola* on pine, it is believed to be the origin of introduction to the west (Mielke 1943). At the time of the survey, only 180 out of the 1,000 *P. strobus* were alive. The oldest canker was a stem infection found on the growth of 1910. Twelve *P. monticola* transplanted from a mile away in 1912 were growing next to the *P. strobus* and showing disease symptoms as well. There were two large *R. nigrum* growing nearby. More diseased

P. monticola were found at the transplanting site, and a young *P. albicaulis* showed pycnidia, at University farm (Davidson 1922). The fungus had spread into interior British Columbia (BC) by 1930 (Hunt 1983). By 1938, *C. ribicola* had spread from BC to northeastern Washington, northern Idaho, and western Montana (Buchanan & Kimmey 1938). By 1961, *C. ribicola* had expanded its range to the southern Sierra, California (Kinloch 2003).

1.1.3 Current distribution

Currently, in western North America, *C. ribicola* is distributed throughout its host range. In the north, it has been found in Smithers, BC, on *P. albicaulis*, in the southern Coastal BC on *P. monticola*, in Alberta and eastern South Dakota on limber pine *P. flexilis* E. James, in New Mexico on southwestern white pine *P. strobiformis* Engelm. and in southern California on sugar pine *P. lambertiana* Douglas (Zeglen et al. 2009). In 2011, the occurrence of *C. ribicola* in Arizona on *P. flexilis* was confirmed for the first time by Fairweather & Geils (2011). The general trend is that it is spreading southward from BC to the lower states of US.

C. ribicola has been gradually spreading from the west coast to central US. It was first reported in Colorado in 1998, close to the Wyoming border (Johnson & Jacobi 2002). Examination of *C. ribicola* incidence on *P. flexilis* in central and southeastern Wyoming and northern Colorado indicated that only 14.3% sampled *P. flexilis* were infected and half of the stand had *C. ribicola* incidence, with 5% mortality level (Kearns & Jacobi 2007). In October 2003, *C. ribicola* was found on Rocky Mountain bristlecone pine (*P. aristata* Engelm.) for the first time in the Great Sand Dunes National Monument in Colorado (Blodgett & Sullivan 2004). Recently in 2009, it was first identified on *P. strobiformis* in Arizona (Fairweather & Geils 2009).

1.1.4 Biology

Cronartium ribicola is a cool weather disease, endemic on white pines that grow on high latitudes and elevations in eastern Asia (Kinloch 2003). It has a heteroecious and macrocyclic life cycle, i.e. it

needs two unrelated host species to complete its life cycle, during which it produces five kinds of spores. The five spore stages are: spermatia (= pycniospores), aeciospores, urediniospores, teliospores and basidiospores. Basidiospores are usually transported only short distances due to their vulnerability to sunlight and desiccation. In contrast, thick-walled aeciospores are more durable and capable of long-distance dispersal (up to 1,000s of km) (Mielke 1943; Smith 1996; Kinloch 2003; Frank et al. 2008). White pines serve as its aecial hosts and *Ribes* and some other shrubs species act as its telial host.

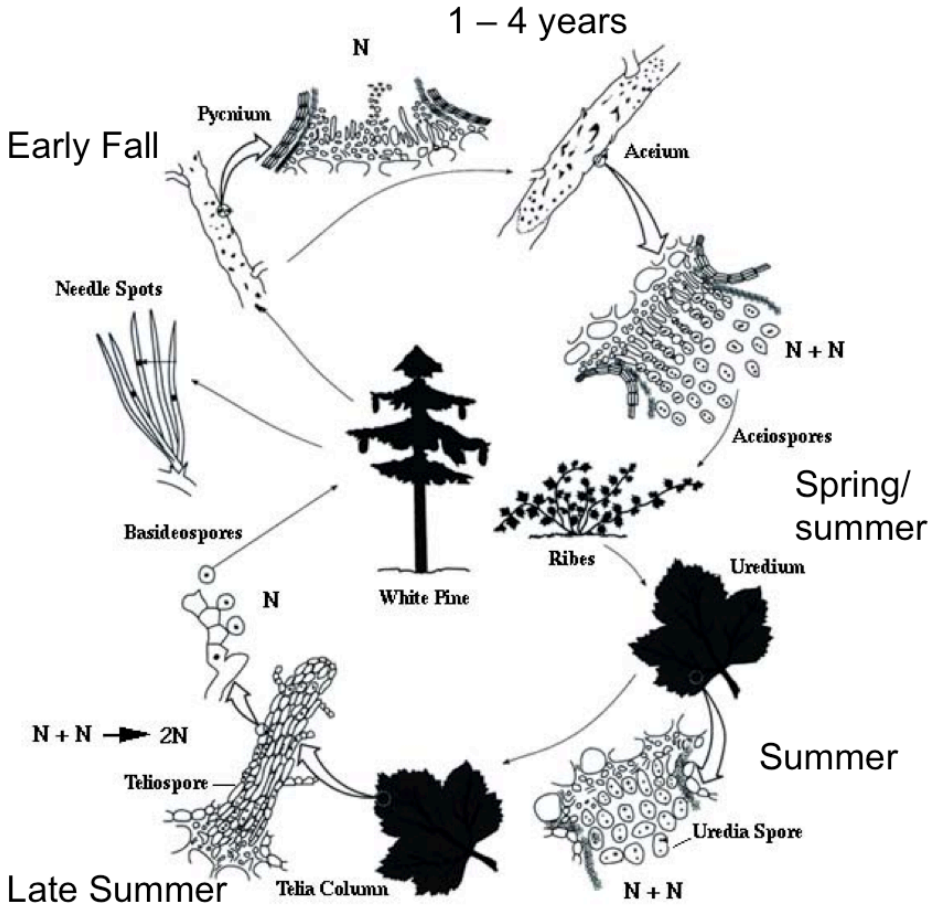

Figure 1 Life cycle of *Cronartium ribicola*, modified from Pacific Southwest Research Station. (2011).

The disease cycle starts when the airborne, short-lived basidiospores (N) land on the needles of white pines at the end of the summer/beginning of fall (Figure 1). They enter through stomata on the needles and travel down to the twigs. The first noticeable symptoms are yellow needle spots, caused by damges to the chlorophyll and exposing carotene in the mesophyll tissue (Kinloch 1992). In the subsequent one to four years, the fungus grows in the phloem and bark with no visible symptoms. In the spring of the third or fourth year, a perenial canker is formed and spermatia (pcynidiospores, N) are produced in sweet nectar within. Spermatia are not involved in infections but only spermatization. *C. ribicola* is a heterothallic fungus, meaning that one canker only carries one mating type. Spermatia with the opposite mating type are carried to the canker by insects. Spermatization happens without nuclei fusion, leading to dikaryotic (N+N) aecia forming in the following spring (Richardson et al. 2008). Blisters develop into cankers within which aeciospores are produce in the following spring. The aeciospores will eventually break the blisters for dispersal. Aecia are produced annually until the bark is totally infected. The fungus keeps moving towards the main stem, at a rate about 2 inches per year, until the tree is completely girdled (Lombard & Bofinger 1999). These airborne aeciospores cannot re-infect pines but are adapted for long distance dispersal and can survive for months under good conditions, in search for the alternate *Ribes* host (Worrall 2009). In the spring, approximately ten days after infection by aeciospores, uredia are produced on the undersurface of infected *Ribes* leaves. Orange urediniospores (N+N) are produced inside uredinia. They cannot infect pines but can reinfect *Ribes*, a process of repeated infections lasting till the end of summer, thus intensifying the disease (Lombard & Bofinger 1999). During late summer/early fall, orange-brown hair-like telia will be produced, to replace uredinia. Telia are where sexual recombination and karyogamy take place – the two nuclei fuse and generate teliospores (2N). Meiosis follows and filamentous basidia (N) germinate from the teliospores, thereby completing its life cycle (Worrall 2009).

1.1.5 Hosts

1.1.5.1 White pines

White pines are a group of five-needle pines that can be classified into 25 different species. Being an obligate bio-trophic pathogen, *C. ribicola* prefers vigorous hosts (Spaulding 1929). In some stands, it can kill up to 95% white pines (Hamelin et al. 2005; Hunt 1983; Elizabeth et al. 2000; Hagle et al. 1989; Kendall & Arno 1990). The relative susceptibilities of the white pines in the subsections Cembrae and Strobi of the subgenus *strobus* have been summarized by Van Arsdel (1981). In Asia and Europe, Siberian pine (*P. sibirica*) and Dwarf Siberian pine (*P. pumila*) are susceptible to *C. ribicola*, whereas, the Japanses white pine (*P. parviflora* Siebold & Zucc.), Balkan pine (*P. peuce* Griseb), Korean pine (*P. koraiensis*), Swiss stone pine (*P. cembra*), and Himalayan pine (*P. wallichiana*) are resistant to *C. ribicola* and Armand pine (*P. armandii* Franch.) is immune to *C. ribicola* (Ekramoddoullah 2005). All native North American white pines are susceptible, ranked here from low to high susceptibility: southwestern white pine (*P. strobiformis*), bristlecone pine (*P. aristata*), eastern white pine (*P. strobus*), Mexican white pine (*P. ayacahuite* Ehrenb. ex Schltdl.), limber pine (*P. flexilis*), western white pine (*P. monticola*), sugar pine (*P. lambertiana*), and whitebark pine (*P. albicaulis*) (Hunt 1983; Kinloch 2003; Van Arsdel 1981). Some of these species were once economically important such as *P. monticola*. It was sold at a premium price for its high workability (Abrams 2001) and it can also replace Douglas fir (*Pseudotsuga menziesii* (Mirb.) Franco*)* in areas prone to laminated root disease (Zeglen et al. 2009). *P. albicaulis* is a keystone species in the Rocky Mountain ecosystem (Smith et al. 2008).

P. albicaulis is a pioneer species in subalpine forests due to its high tolerance to harsh conditions and ability to improve the microsite for other plant and animal species (Campbell & Antos 2000). Some wildlife species depend heavily on *P. albicaulis* for survival. One example is Clark's nutcracker (*Nucifraga columbiana*), which feeds on the seeds of whitebark pine and also acts as the primary dispersal agent of *P. albicaulis* (Campbell & Antos 2000). Besides Clark's nutcracker, grizzly bear (*Ursus arctos* L.) and red squirrels (*Tamiasciurus hudsonicus* E.) rely on its seeds as food source as well

(Carolin et al. 2008; Zeglen 2002). Despite the ecological importance of *P. albicaulis*, its abundance has decreased rapidly in the US and Canada in the past few decades, as a result of *C. ribicola*. A major destructive impact of *C. ribicola* on the wildlife is that it kills the upper cone-bearing branches, leading to the reduction of seeds produced (Carolin et al. 2008). In BC, *P. albicaulis* has been infected by *C. ribicola* all over its geographic range because the climatic conditions are generally suitable for the rust to reproduce and infect (Campbell & Antos 2000). Removal of white pines will affect many ecosystem functions like capacity for holding snow pack, delaying snowmelt and protecting watersheds (Kendall 1994; Keane et al. 1994).

Sampling of the BC subalpine stands revealed 21% of the trees had dead stems that could be attributed mostly to *C. ribicola*. In comparison, *C. ribicola* incidence on living trees was estimated to be 44%. *P. albicaulis* lacks natural resistance to *C. ribicola* and it was found to be much more susceptible than *P. monticola*. The occurrence of *C. ribicola* and mortality of *P. albicaulis* were considerably related to stand structure and the existence of *Ribes spp*., with little or no relationship to mountain pine beetle *Dendroctonus ponderosae* Hopk. (Campbell & Antos 2000).

Rust-killed whitebark pines have been mostly replaced by subalpine fir (*Abies lasiocarpa)* and Engelmann spruce *(Picea engelmannii)* in the subalpine forests. A large number of whitebark pines have already been infected or killed by the pathogen. Due to a low regeneration rate arising from a high susceptibility to *C. ribicola*, *P. albicaulis* trees are facing a danger of extinction. The loss of whitebark pine not only drastically changes the subalpine ecosystem, but also challenges the genetic breeding programs for rust resistance, because of reduced genetic pool (Zeglen 2002).

Limber pine (*P. flexilis*), similar to *P. albicaulis,* is an ecologically important and pioneer species, especially after disturbances like fire or avalanches (Kearns & Jacobi 2007). Although not commercially important, limber pine prevents erosion in watersheds, acts as shelter and food supply to wildlife, and adds aesthetic value to the landscape.

1.1.5.2 Alternate hosts

As mentioned earlier, being heteroecious, *C. ribicola* requires a different host besides white pines to complete its life cycle. The role of alternate host is crucial as the disease would not be able to persist or spread far without the presence of an alternate host. The availability and distribution of the telial host strongly affect the spread and intensification of *C. ribicola*.

Its most well known telial hosts belong to the genus *Ribes* L. in the Grossulariaceae family. The genus contains more than 150 described species, among which black currants (*Ribes nigrum* L. and hybrids) are a major crop and being commercially grown in northern Europe (Hummer & Dale 2010). The North American *Ribes* crops, namely black currants, red and white currants and gooseberries were domesticated from Europe and are not so economically important now due to their vulnerability to *C. ribicola*. *Ribes* cultivation is still restricted or prohibited in twelve states of America (Hummer & Dale 2010). Wild *Ribes* species native to North America are not susceptible to *C. ribicola*. Thus, the damage to *Ribes* in western North American is considered to be minimal, i.e. only late-season defoliation (Kearns et al. 2008).

Red currants were likely to have been the first species imported to North America by the British colonists before the 1500s (Hummer & Dale 2010). About 100 currant cultivars had been widely planted throughout North America prior to the European importation of *C. ribicola* from Asia. When *C. ribicola* was introduced to North America, high inoculum potential was seen on *R. nigrum,* which posed a big threat (Spaulding 1922). Species within the *Ribes* genus differ in their abilities to produce pine-infecting basidiospores. Their relative susceptibilities to *C. ribicola* were summarized by Van Arsdel & Geils (2004). Like the pathogen, the currants also prefer temperate to cool climates. Developing rust resistant cultivars is a way of re-opening the potential for currant production without threatening the pines in North America.

Besides *Ribes* (Grossulariaceae), *C. ribicola*'s alternate hosts include *Pedicularis racemosa* Dougl. ex Benth., *Casrilleja miniata* Dougl. and *Pedicularis bracteosa*, all from the Orobanchaceae family. Their status as alternate WBPR host has been confirmed by both genetic analyses and artificial innoculations (McDonald et al. 2006; Zambino et al. 2007).

1.1.6 Epidemiology

A higher incidence of *C. ribicola* seemed to be related to increase in latitude, elevation and closeness to cool waters like lakes (Van Arsdel 1972). It was determined by Van Arsdel that elevation alone does not correlate with *C. ribicola* incidence but it works together with other factors like the presence of susceptible hosts and microclimate conditions to shape the distribution of *C. ribicola*. On a micro-climatic scale, small canopy openings favour *C. ribicola* because of the accumulation of cold air (Van Arsdel 1972). A landscape analysis on the risk factors for *C. ribicola* showed that in Wisconsin, the occurrence of *C. ribicola* increases with latitude and elevation (White et al. 2002).

Summer precipitation could be one critical factor that limits the spread of *C. ribicola*, as there is less infection in California and Nevada compared to the states and provinces up North (Campbell & Antos 2000). In one study, heavy *C. ribicola* infections were seen in years with heavy rainfall (Pennington 1925). In BC, the most favourable time for *C. ribicola* infection is April to September, when there is high precipitation and cool temperatures (Spaulding 1929; Hunt 2004).

On a stand-level, on one hand, larger trees with big diameter and crown class in open stands are more likely to get infected by *C. ribicola* for the reasons that they have larger surface area for the spores to land on and fewer canopies to prevent the incoming airborne spores (Campbell & Antos, 2000; Kearns & Jacobi 2007). On the other hand, as the distance between the needles and the main stems is longer in big trees, fewer stem cankers would develop compared to smaller trees (Kearns & Jacobi 2007).

Temperature and moisture content in the surrounding environment regulate rust spores in the following aspects: production, longevity, germination, and ability to penetrate and establishment in the

host. Nocturnal air circulation has the biggest influence on spore dispersal (Van Arsdel 1972). It was shown that on a macro scale, rusts were more abundant in the northern regions, at higher elevations and close to cold water-bodies. On a micro scale, the favourable conditions are provided by very cool and wet climate in small openings, under which teliospores and basidiospores are produced on *Ribes*. Super-cooled leaves with condensed water on the undersurface produce the most favourable conditions for rust infection (Van Arsdel 1972). To summarize, in order for the spores to produce and germinate, moist environment with temperatures less than 20°C, with high relative humidity (over 97%) or free water are required (Hirt 1942; Green & Van Arsdel 1956; Krebill 1971; McDonald et al. 1981). These favourable conditions have to persist for a minimum of two days in order for the pine infections to happen (Van Arsdel et al. 1956). However, a recent whitebark pine survey in BC found that there was no relationship between the site or climatic data and the severity of *C. ribicola* infections, suggesting that the conditions throughout BC are quite favourable for the disease (Campbell & Antos 2000).

1.1.7 Resistance in white pines

Broadly speaking, plants possess two types of resistance - major gene resistance (MGR), where resistance is controlled by a single dominant gene and multigenic/quantitative resistance (quantitative resistance), where a large number of genes control a small proportion of the total resistance (Richardson et al. 2008). MGR leads to a hypersensitive reaction (HR) in the mesophyll tissue in case of an infection event by an avirulent race (Kinloch 1981). HR is also known as localized cell death, and it is one major defense mechanism carried in plants against infections by pathogens (Dangl & Jones 2001). Localized cell death not only kills the infected cells, but also prevents the nutrient flow to the pathogen (Sweeney et al. 2011) The incompatible reaction in white pines is shown as small necrotic flecks as opposed to large red or bright yellow lesions in compatible genotypes (Kinloch & Comstock 1980).

MGR is a kind of gene-for-gene interaction, which means that there exists a resistance gene in the host specific to a corresponding avirulence (Avr) gene carried by a particular pathogen race (Ekramoddoullah 2005). When incompatible, *C. ribicola* can rarely penetrate through the needle

endodermis and central vascular cylinder, which is how it grows into the bark tissues normally. When penetration does happen, it triggers an HR in the bark tissue and prevents it from infecting further (Kinloch & Comstock 1981). For instance, *Cr1* and *Cr2* are two genotypes for MGR possessed by *P. lambertiana* and *P. monticola*, respectively (Kinloch & Littlefield 1977; Kinloch et al. 1970). Upon infection, *P. monticola* carrying the *Cr2* gene would trigger host cell death and secret defense compounds in needles (Kinloch et al. 1999). Unfortunately, both *Cr1* and *Cr2* have been overcome by the virulent races of *C. ribicola* carrying *vcr1* and *vcr2* genotypes, respectively (Kinloch et al. 1999). *Vcr1* was discovered at Happy Camp and at Mountain Home Demonstration State Forest in northern California (Kinloch & Comstock 1981; Kinloch 1996). *Vcr2* was detected also at Happy Camp (Kinloch et al. 2004). *Cr3* is another MGR gene existing only in *P. strobiformis*. The corresponding virulence gene in the rust is yet to be identified (Kinloch et al. 2004; Vogler et al. 2005). As expected, artificial inoculation tests indicated that v*cr1* is avirulent against *Cr2* in western white pine and *vcr2* is avirulent against Cr1 in sugar pine (Vogler et al. 2005). Furthermore, *vcr1* and *vcr2* races were inoculated onto resistance southwestern white pine with *Cr3*, and showed no virulence either (Vogler et al. 2005). *Cr1*, *Cr2*, and *Cr3* are controlled by different alleles, and likely to be on different loci (Kinloch et al. 2003). MGR was also detected in *P. flexilis*, but the resistance gene is yet to be confirmed (Vogler et al. 2005; Jurgens et al. 2003). In BC, *Cr2* was shown to be a stable resistance gene in *P. monticola* as no virulent race of *C. ribicola* had been found in the province till 2004 (Hunt 2004).

While the MGR is generally host specific, the multigenic resistance involves multiple genes and is usually non-host specific (Ganley et al. 2008). These genes can be involved in traits like prevention of infection (Richardson et al. 2008). Different characteristics of resistance in *P. strobus* have been observed: smaller spots on the needles, spots that do not develop into stem cankers, and slower canker growth (Jurgens et al. 2003). Such kind of resistance in native *P. monticola* is found to be ontogenic, i.e. it is affected by the environment and tends to increase as plants age (Hunt 2004).

In addition to the two aforementioned types of resistance, induced resistance responses by fungal endophytes in *P. monticola* have also been described (Ganley et al. 2008). The endophytes help in reducing the damage caused by *C. ribicola* and also improve the survival rate of the seedlings (Ganley et al. 2008).

1.1.8 Management strategies

The numerous attempts to control the *C. ribicola* epidemic were characterized as the most extensive effort in forest pathology so far by Maloy (1997), even though most of them turned out to be ineffective. The most commonly employed method is the eradication of the alternate host, *Ribes*. Theoretically, removal of *Ribes* could be effective if it was permanent and covered the whole landscape (Kinloch 2003). However, it turned out that many large, well-adapted *Ribes* bushes were spreading across some steep terrains, which made it difficult to completely eliminate. The success rate of these *Ribes* eradication programs varied between areas with a higher success rate in eastern North America than in the west. Reasons of failure in the west included widespread existence of more *Ribes* species and more suitable climate for rust spread (Van Arsdel & Geils 2011). On the other hand, small-scale success in the east was debatable as it was argued that the hazard had not been high enough in eastern North America before the attempted eradication (Kinloch 2003). In brief, this method is neither practical nor economical on a regional scale.

Other options include antibiotics and silvicultural techniques. Unfortunately, neither showed any remarkable effect. As a result, pathologists have been focusing on genetic resistance in trees, which would offer the best long-term potential for controlling *C. ribicola* (Kinloch 2003). Most progress was found on the two commercially important species, sugar pine and western white pine. An accelerated deployment of resistant western white pine is being practiced in BC. A danger to the current genetic selection and breeding programs may come from reintroduction from Asia, where *C. ribicola* exists as a complex of multiple species with different alternate host affinities and can also be autoecious (Kinloch 2003).

1.1.9 Genetics studies

The population genetic structure of *C. ribicola* was not well understood until the 1990s (Kinloch 2003). Use of three different types of markers - isozymes, random amplified polymorphic DNA (RAPDs), and restriction fragment length polymorphism (RFLP) markers - found a low genetic diversity and differentiation among *C. ribicola* populations in North America (Kinloch et al. 1998). No correlation was observed between the genetic and geographic distances. This study did not find any private alleles in any western or eastern North American populations. Hence they concluded that all North American populations shared the same gene pool.

Subsequent studies found low but significant genetic differentiation among populations within a regions or stand type, although the most genetic diversity was attributed by variances within populations, suggesting no strong association between genetic distance and geographic origin or host type (Hamelin et al. 1995). The explainations for the low genetic distance between physically remote populations (1000 km apart) were proposed to be either large-scale gene flow or distribution by a common genetic pool (Hamelin et al. 1995).

Congruently, low genetic differentiation within the northeastern North American *C. ribicola* populations was detected, with an expected heterozygosity (Hw) of 0.370, which is much higher than that of among populations (Hb = 0.016) (Et-touil et al. 1999). A hierachical analysis of molecular variance (AMOVA) also showed no statistically significant differentiation among provinces or regions, but a significant differentiatin among populations within regions or provinces. This analysis confirmed that the distribution of *C. ribicola* in eastern North America is panmictic subpopulations within a metapopulation, with high gene flow across the studied area (Fst = 0.062). The isolation-by-distance hypothesis was rejected by the Mantel test, indicating again that the genetic diversity of populations is not correlated with geographic distance. It was concluded that the long-distance spore dispersal, genetic drift, and new colonization were significant mechanisms in the rust's epidemiology (Et-touil et al. 1999). Additionally,

C. ribicola was proven to be sexually outcrossing in nature (Gitzendanner et al. 1996), thus high gene flow was expected.

Later, it was determined that genetic differentiation exists between the eastern and western North American populations, with a higher genetic diversity in eastern North American populations and high migration between New Mexico and British Columbia populations (Hamelin et al. 2000). This study also suggested the presence of a barrier to gene flow between eastern and western populations, combined with founder effects (Hamelin et al. 2000).

A molecular epidemiology study of *C. ribicola* collected samples from two heavily infected white pine plantations, focusing specifically on its recombination and spatial distribution and identified multilocus haplotypes (MLHs) for the spermogonial (monokaryotic haploid) stage of *C. ribicola* (Hamelin et al. 2005). The MLHs distribution pattern at the two sites suggested that the existence of extensive sexual recombination followed by long-distance dispersal in *C. ribicola* populations. Later, it was shown that the western populations are not genetically uniform, and the selection for R-gene resistance may affect its genetic diversity and differentiation (Richardson *et al* 2008). The lowest genetic diversity at the Happy Camp site was identified by a study on the genetic structure of *C. ribicola* in relation to host resistance (Richardson et al. 2008). The highest diversity was found in a *P. monticola* plantation that utilized multigenic resistance to *C. ribicola* rust in Idaho (Richardson et al. 2008). Similarly, an earlier sampling of the *C. ribicola vcr1* race had also detected the largest distance between the *vcr1* population and the rest of the *C. ribicola* populations (Kinloch et al. 1998).

Advances in DNA sequencing technologies have allowed scientists to generate large numbers of high resolution genetic markers (Neale & Savolainen 2004). Multiple researches have used mtDNA, cpDNA, and nuclear (n)DNA to study the evolutionary relationships of *C. ribicola* (Richardson et al. 2010). One significant finding was that there are at least three distinct genetic clusters of *C. ribicola* among the Eurasian and North American populations: one cluster consisting of Korean and northeastern

Chinese isolates; Japanese isolates being an independent cluster; USA and Germany isolates forming the third cluster (Richardson et al. 2010).

1.2 Genotyping-By-Sequencing (GBS) Overview

Genome-wide markers are vital in understanding population evolutionary history, population structure and pattern, and monitoring adaptation (Davey & Blaxter 2010). Single nucleotide polymorphisms (SNPs) are small variations in DNA sequence caused by a single nucleotide change (King et al. 2013). They are abundant and highly polymorphic, which make them ideal markers for within species diversity exploration, genome-wide association studies (GWAS): genomic selection and marker-assisted breeding and selection (Baird et al. 2008; Elshire et al. 2011; Ward et al. 2013). The traditional way of SNP discovery is often costly and time-consuming, because it requires two main steps: the initial discovery of SNPs in a subset of individuals and then genotyping them in larger populations on certain platforms (Sonah et al. 2013).

Reduced representation libraries (RRLs) were first used by Altshuler et al. (2000) to build a SNP map of the human genome in search for haplotypes associated with common diseases. In recent years, the discovery of SNPs in non-model organisms has been substantially facilitated by the development of high-throughput next-generation sequencing (NGS) coupled with advances in genome complexity reduction. Restriction-site-associated DNA sequencing (RADseq), untilizes restriction enzymes (REs) to specifically digest DNA fragments flanking a cut site (similar to RFLP and AFLP) (Baird et al. 2008). Two kinds of markers are generated – presence or absence of enzyme cut site and SNPs and insertion or deletion (indels) in the tag sequences (Davey & Blaxter 2010). By sequencing (single or paired-end) these RAD tags on NGS platforms, such as Illumina, Roche and SOLid, rapid genome-wide SNP discovery and genotyping across multiple individuals can be achieved simultaneously and unbiasedly (Baird et al. 2008). *De novo* assembly and marker discovery are also possible with RADseq (Emerson et al. 2010). By looking at genome regions with exceptionally high or low differentiation among populations, selection

(either diversifying or stabilizing) can be detected too. Furthermore, genes or alleles in these regions can be extracted for functional studies (Hohenlohe et al. 2010).

The application of RADseq has been demonstrated in non-model species like threespine stickleback, *Gasterosteus aculeatus,* (with reference genome), which aimed to study the genetic diversity and differentiation among its wild populations. Genomic regions and candidate genes significant to population adaptation were successfully identified (Hohenlohe et al. 2010). Similarly, RADseq has been used in other non-model organisms e.g. fungal species *Neurospora crassa* (without reference genome) (Baird et al. 2008); in phylogeny study of *Wyeomyia smithii* (the pitcher plant mosquito, without reference genome) (Emerson et al. 2010); and SNP generation in rainbow trout (*Oncorhynchus mykiss*) and cutthroat trout (*O. clarkii lewisi*) (Hohenlohe et al. 2011).

Genotyping-by-sequencing (GBS) was originally developed by Elshire et al. (2011) for maize, an organism with large genome and high genetic diversity. The detailed methodology is thoroughly described by Davey (2011) and Elshire et al. (2011). Similar to RADseq, it produces a reduced representation of the genome by only targeting sequences flanking REs cut sites, but its library construction process is much simplified. In RADseq, after REs digestion, P1 adaptors are added to the sticky end overhangs. Then the samples go through a pooling, random shearing and size selection process (300-700 bp). Afterwards, Y-shaped P2 adaptors will be added and only fragments with both P1 and P2 adaptors will be amplified and sequenced (Davey et al. 2011). As for GBS, following digestions, both barcoded adaptors (4-9 bp) and common adaptors are ligated to the overhangs at the same time (Figure 2). Samples are pooled together and passed straight to the PCR process on the Illumina Genome Analyzer flow-cell. Thus, less DNA is required (100 ng). Only fragments with both barcoded and common adaptors (150-350 bp) will be sequenced (Davey et al. 2011; Elshire et al. 2011). Therefore, GBS is both labour and cost effective. It is also quite versatile as it allows users to choose enzyme combinations, according to their desired marker density. Sonah et al. (2013) used the one-enzyme (*Ape*KI) protocol on soybean and

discovered 10,120 high quality SNPs. They then randomly validated 24 SNPs using Sanger sequencing, with a success rate of 98%.

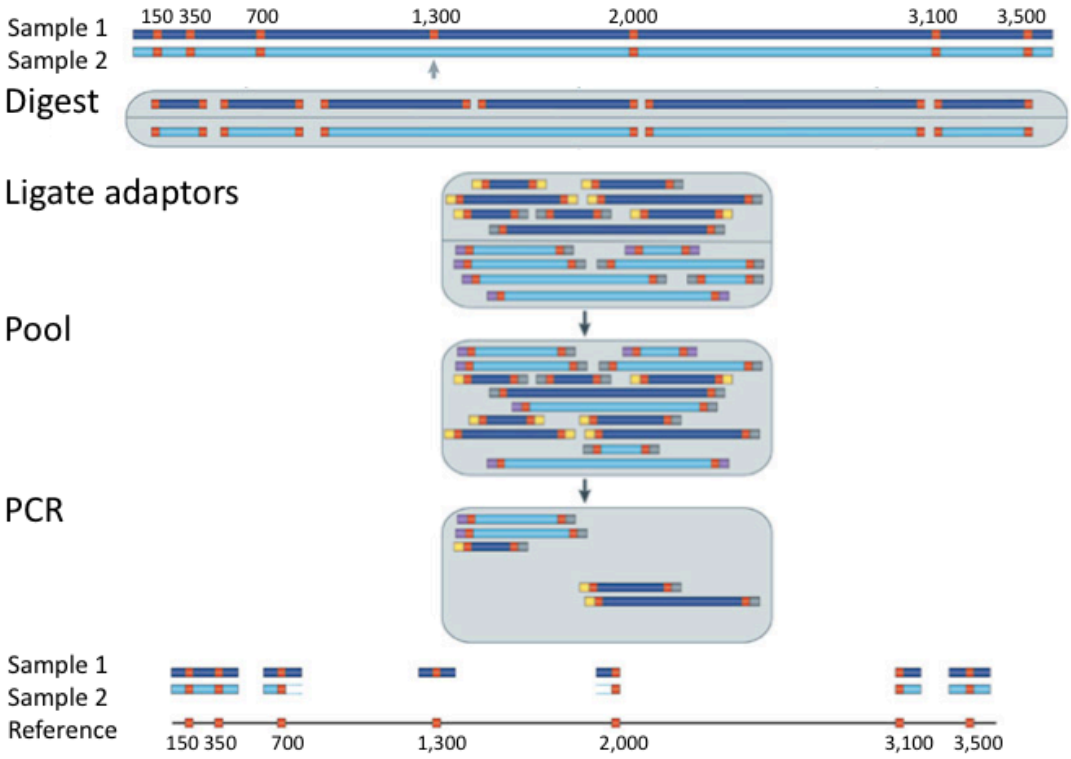

Figure 2 Brief illustration of Genotyping-by-Sequencing methodology, modified from Davey et al. (2011).

Poland et al. (2012) modified the original GBS protocol to a two-enzyme approach and successfully applied it in barley and wheat. The major difference is that it uses a "rare-cutter" restriction enzyme along with a "common-cutter". In their study, the rare-cutting enzyme is *Pst*I (CTGCAG), to which the barcoded forward adapters are designed. The Y-shaped reverse adapters are designed to match the common-cutting enzyme *Msp*I (CCGG) overhangs. Because the Y-adapters contain the same base pairs (instead of complements) as the reverse PCR primer, binding sites for the reverse primer are not created until the extension from the first PCR amplification cycle (happens from the forward side) reaches the reverse side. This design assures that the GBS libraries are uniform and only *Pst*I-*Msp*I fragments are being amplified (Poland et al. 2012).

High levels of missing data tend to be associated with GBS (Elshire et al. 2011). The likely sources of missing data in GBS are: 1) the site maybe absent in some samples, in which case, missing data can be treated as a presence/absence marker; 2) unsuccessful digestion and PCR due to poor DNA quality or other technical issues; 3) low read depth at certain sites. To maintain low genotyping cost, GBS libraries are generally sequenced at low coverage. Therefore, increasing coverage is one way of reducing the amount of missing data. Another alternative is to impute the missing data, i.e. using statistical methods to replace missing values with estimated values (Davey et al. 2011). Many imputation methods are being developed for GBS for species with or without reference genome, and the imputation results are quite accurate and promising (Huang et al. 2014; Rutkoski et al. 2013; Ward et al. 2013).

1.3 Objectives

The objectives of this study were to use genome-wide SNP markers generated using GBS to 1) reconstruct disease spread pathways using global collections, by studying the centre of genetic diversity; 2) contrast populations of *C. ribicola* in eastern and western North America; 3) elucidate fine-level populations structure in western NA with regard to landscape features and host species. The hypotheses are that the highest diversity to be found in northern Asian samples, and decreases gradually through Russia, Europe and North America; higher diversity in eastern North America than western North America because of multiple and earlier introductions; more population structure in the western epidemiological unit to be detected due to the presence of more susceptible hosts and variations in landscape features.

2 Materials and Methods

2.1 Sampling

Samples of this study came from a variety of sources, aiming to capture the broad distribution range of the fungus: 1) continent-wide sampling (Hamelin et al. 2000); 2) international samples from collaborators; 3) intensive sampling in western Canada (Brar 2012), with a focus on covering different landscapes and host species; 4) additional sampling in 2012 and 2014, primarily from Smithers, BC, the northern most limit of *C. ribicola* and white pine distribution, where a private SNP was identified in Brar's study in 2012. The final dataset used for population analyses is summarized in Table 1.

Table 1 Summary of sampling locations.

AREA	REGION	NUMBER OF
Asia	Russia (RU)	3
	Korea (KO)	9
	China (CH)	1
Subtotal		**13**
Eastern North America	Maine (MA)	1
	Vermont (VT)	1
	Wisconsin (WI)	1
	Québec (QC)	28
	Ontario (ON)	2
	New York (NY)	1
	Nova Scotia (NS)	2
	New Hampshire (NH)	1
	Newfoundland (NF)	3
	Minnesota (MN)	2
Subtotal		**42**
Europe	France (FR)	1
	Finland (FI)	1
Subtotal		**2**
Western North America	British Columbia (BC)	110
	New Mexico (USNM)	3
	Montana (USMR)	10
	Alberta (AB)	12
Subtotal		**135**
Total		**192**

Generally, dikaryotic aeciospores were collected on white pines, with a few exceptions where urediniospores were produced on *Ribes* from either an aecial or a telial source. To make sure aeciospores are pure and genetically uniform, samples were collected in late spring/early summer before the aecial blisters open. Each blister was ruptured open by toothpicks and dry spores inside were collected into a 1.5 ml eppendorf tube. On average, three blisters per canker were collected and labeled. To minimize contamination, the blisters were ruptured from bottom up on each canker. On each site, ten or more trees were sampled when possible. During transportation, the tubes were stored in 96-well boxes, which were placed in plastic containers together with bags of calcium sulphate (dryrite) as desiccant. When the samples reached the lab, before transferring them to cold storage in -20˚C freezer, they were processed in desiccation chambers, at the bottom of which was a layer of water saturated with calcium chloride.

2.2 DNA Extraction

Two DNA extraction protocols were tested. The first extraction method, herein referred to as the CTAB method, was adapted from a previously described standardized rust DNA extraction protocol (Zolan & Pukkila 1986). Briefly, approximately 10mg of lyophilized spores, 2 tungsten beads (QIAGEN) and 10 mg of diatomaceous earth (Sigma Chemical Co., St. Louis) were added to a 2 ml SafeLock (Eppendorf) tube. The tube was then frozen in liquid nitrogen for a few seconds and disrupted using a mixer-mill for 1 min at 26 Hz. This step was repeated for getting better cell wall breakage. After the mixer-mill step, 600 µl of Cetyl Trimethyl Ammonium Bromide (CTAB) buffer (2% CTAB, 0.1 M Tris (pH 8.0), 1.4 M NaCl, 0.02 M EDTA, 0.2% β-mercaptoethanol) was added to the tube. The samples were then incubated at 65˚C for one hour and vortexed every 15 minutes during the process.

Subsequently, the samples were extracted with 600 µl of phenol:chloroform:isoamyl alcohol (25:24:1), vortexed and centrifuged at 13,000 rpm for 5 min. The supernatant was pipetted into a fresh 1.5 ml SafeLock tube and washed with 600 µl of chloroform:isoamyl alcohol (24:1). The mix was vortexed and centrifuged once more at 13,000 rpm for 5 min. The upper layer was transferred to a new 1.5 ml SafeLock tube and digested with 1 µl of RNAse A (100 mg/ml). This mix was incubated at 37˚C for one

hour. The samples were washed with 600 µl of phenol:chloroform:isoamyl alcohol (25:24:1) followed by 600 µl of chloroform:isoamyl alcohol (24:1) again to remove the impurities. The upper phase was pipetted into a new 1.5 ml SafeLock and the DNA was precipitated with 150 µl of 7.5 M NH_4OAc and 800 µl ice-cold isopropanol and incubating for at least 30 min at -20˚C. The DNA pellet was formed by centrifuging at 13,000 rpm for 5 min (at 4˚C) and washing with 800 µl ice-cold 70% ethanol. The pellet was then air dried and re-suspended in 20 µl H_2O (Sigma-Aldrich, Inc). The DNA was stored at -20˚C after extraction.

The second DNA extraction protocol used the Qiagen DNeasy Plant Mini extraction kit® (Qiagen Inc., Toronto, Ontario, Canada) as previously described (Brar 2012). DNA yield from both protocols was quantified on a Qubit® 2.0 Fluorometer (Invitrogen™, Life Technologies) using the Qubit® dsDNA BR assay kit.

One of the main requirements for GBS is decent DNA amount and quality. Rusts being obligate biotrophs, these cannot be cultured outside their hosts. As a result the starting material to be used for DNA extraction is quite limited (less than 20 mg). This raises a challenge to obtain high DNA yield from the available spores. For this reason, two DNA extraction protocols – Qiagen kit and CTAB method - were tested and the amount of DNA extracted was compared. The CTAB method proved to be more effective as we got a higher DNA yield with the same amount of starting material. Therefore, all the samples were extracted with the CTAB protocol and quantified for GBS library preparation.

2.3 Experiment Design

Based on their geographic origin, 192 inter-continental samples were selected for GBS (Table 1). The North American sample collection sites are illustrated in Figure 3. At least one technical replicate was incorporated in each region's data setup. For subsequent library preparation and sequencing, the 192 samples were equally divided into four plates. Plates 1-3, which mainly consisted of western NA samples, were prepared at the University of British Columbia. Plate 4, assembled at Laurentian Forestry Centre,

Québec, contained the eastern North American and international samples. An additional whole genome amplification (WGA) step was performed for 15 samples that did not have enough DNA. To visualize the effects of WGA, replicates of the same sample before and after WGA were included.

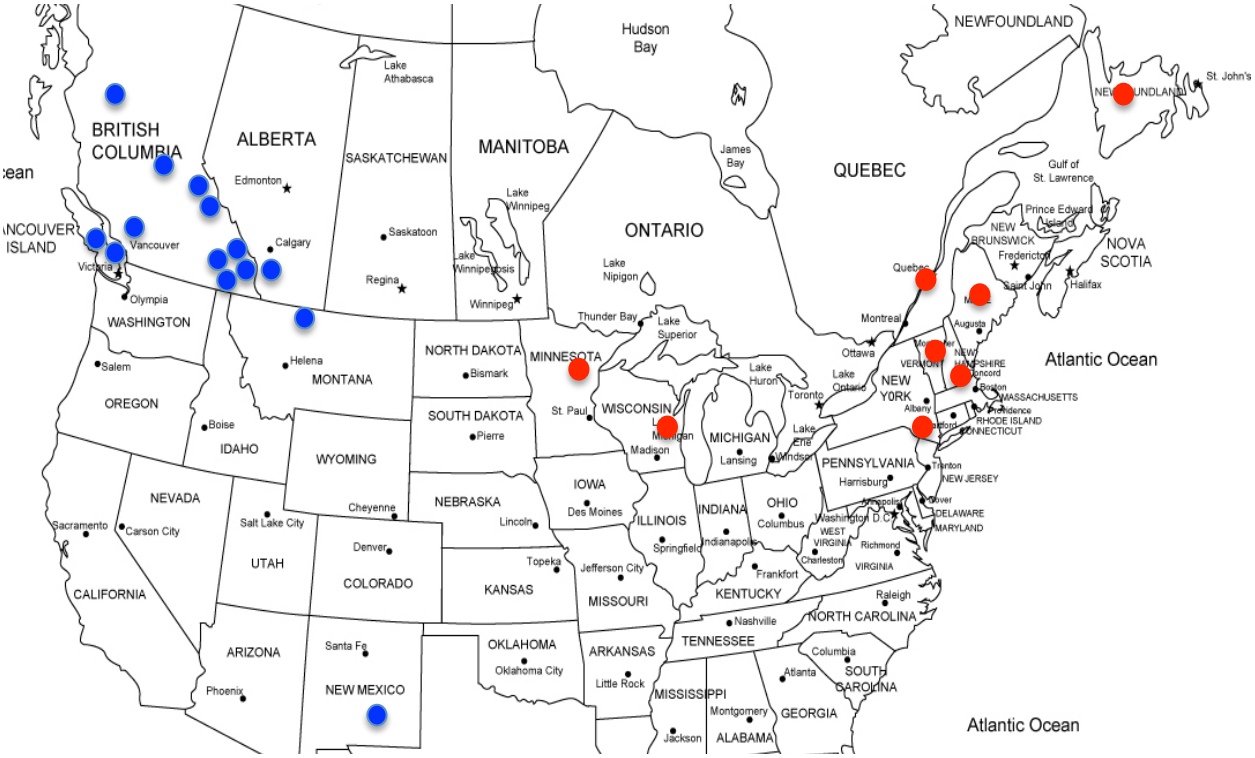

Figure 3 North American sampling locations, blue = western North America, red = eastern North America.

2.4 Genotyping-By-Sequencing Library Construction

2.4.1 GBS pilot run

A pilot GBS study was conducted in 2013 to test the suitability of this technique for SNP discovery in *C. ribicola*. The choice of REs is critical in GBS because the digested fragments do not go through the size selection process, which means that the REs should maximize the number of cut fragments that are 100-400 bp long for sequencing (Sonah et al. 2013). The main objectives were to

determine 1) whether GBS data would yield true SNPs in *C. ribicola*; 2) if the default two enzyme combination would produce sites for sequencing; 3) and to determine the optimum working DNA concentration necessary for GBS library preparation and subsequent sequencing.

For this study, single *C. ribicola* aecia collections collected in 2009 and 2010 were used. Seven out of the eight samples came from western Canada while one came from the east. DNA was extracted using CTAB protocol standardized for rusts. To meet the minimal DNA amount (200 ng) required for GBS, up to four samples from nearby locations were pooled together (Table 2).

Table 2 *Cronartium ribicola* samples used for GBS pilot study to assess impact of DNA concentration.

	SAMPLE MIX	LOCATION (S)	REGION	DNA PURITY
1	M 4-1-2 MW 30-1-3	McBride Mount Washington	Interior BC	Mix
2	WGA-SME-2B	Québec	East	Pure / WGA
3	B-26-3	Texada Island	West Coast	Pure
4	P 1-1-2	Alberta	Alberta	Pure
5	K7-201 B, Q-6-3	Alberta	Alberta	Mix
6	Bulk 7	BC	BC	Mix
7	V-5-3 MK-054-2, MK-054-5	Valemont Smithers	Interior BC, High altitude	Mix
8	PR-9-5, PR-7-4 B-30-2, B-3-9	Powel River Texada Island	West Coast	Mix

The GBS library preparation step was done at Institut de Biologie Intégrative et de Systèmes (IBIS), Université Laval. To test different DNA amounts for library construction, a dilution series was made for each sample: 6.25 ng (column 1), 12.5 ng (column 2), 25 ng (column 3), 50 ng (column 4) and 100 ng (column 5) (except for sample 2, which did not have enough DNA amount for the 100 ng concentration). Individual restriction and ligation steps were performed on each sample/dilution combination (4 columns with 8 and 1 column with 7 samples = 39 reactions), following a two-enzyme

(*Pst*I and *Msp*I) protocol (Poland et al. 2012). Samples with the same concentration (column pools, 8 samples) were pooled together and submitted for Illumina sequencing. Library concentration was quantified using Nanodrop 1000 (Thermo scientific, Wilmington, DE 19810 USA). Quality of the library construction step was assessed in terms of the fragment size distribution using a 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA).

2.4.2 Final GBS library preparation and sequencing

Four 48-plex libraries (192 individuals) were prepared and sent to McGill University and Génome Québec Innovation Centre for single-end sequencing on Illumina HiSeq 2000 platform. Each sample was tagged with a unique barcode for identification purpose.

2.5 De-multiplexing and Mapping

Reads from each sequencing run were de-multiplexed using FASTX-Toolkit (HannonLab 2014). After splitting the Illumina reads in to separate files, the barcodes (first nine bases from the 5' end) were trimmed using PRINSEQ (Schmieder & Edwards 2011). Statistics about the read quality were generated using FastQC (Andrews 2014). Reads from individual samples were mapped onto the *C. ribicola* reference genome (GenBank ID: GCA_000500245.1) assembled by the Tree Aggressors Identification using Genomic Approaches (TAIGA) Lab using Burrows-Wheeler Aligner (BWA) (Li & Durbin 2009). The resulting BAM files were then sorted and indexed using SAMtools (Li et al. 2009). Mapping quality was checked by both SAMtools ('flagstat' function) and Qualimap (García-Alcalde et al. 2012).

2.6 SNP Calling and Filtering

SAMtools (Li et al. 2009) were used for generating SNPs. Before SNP calling, the reference sequences were indexed by using the 'faidx' function. SNPs were called and stored in BCF file using the 'mpileup' function. BCF format was converted to VCF format by 'bcftools view' function. To avoid ascertainment bias, four SNP sets were generated independently for studying global, continental and regional structures, respectively. The global set includes all the 192 individuals that were sequenced. The

continental set contains only the North American (NA) samples while the regional set was further divided into two sets: the eastern and western North America.

Stringent filtering parameters were applied on all of the four SNP sets using VCFtools along with associated perl scripts (Danecek et al. 2011). High quality SNP sets were constructed by meeting the following ten criteria: 1) read depth for each genotype: minimum (min) - 4, maximum (max) - 130; 2) mean depth for a site across all individuals min - 4, max - 200; 3) genotype quality cut-off - 30 (99.9% accuracy); 4) root mean square (RMS) mapping quality cut-off - 30; 5) base quality cut-off - 20 (99% base call accuracy); 6) exclude SNPs within 10 bp around a gap; 7) keep bi-allelic loci only; 8) remove INDELs; 9) exclude heterozygotes only sites; 10) keep SNPs present in at least 80% of the individuals. Two additional filters were only applied on the regional datasets: 1) minor allele account – 2; 2) exclude sites that are not in Hardy-Weinberg equilibrium (HWE). This step assured that the SNPs for in-depth population studies were in strictly high quality and there would be no singletons. Genotypes that did not meet the above criteria were filtered out and replaced with a dot representing missing data.

After initial filtering, some individuals had extremely high percentage of missing data, which could introduce potential artifacts in the downstream analyses. As a result, 26 individuals with over 30% missing data were discarded. Unfortunately, the one sample from France fell into this category so the only European individual left was the Finnish sample. Therefore, the European population was not covered in this study. Similarly, only one New Mexico sample passed the filtering process, thus it was excluded from population study as well. Eventually, 164 samples were retained from the total dataset with eight being international samples and 156 coming from North America. The North American dataset itself comprised of 118 western and 38 eastern samples.

Some basic statistics of the filtered SNP sets were extracted using the built-in functions of VCFtools, including observed heterozygosity, individual missingness, singletons and private doubletons

(i.e. SNPs where the minor allele only occurs in a single individual and that individual is homozygotic for that allele).

2.7 Population Structure Analyses

2.7.1 Principal component analysis (PCA)

In Principal Component Analysis (PCA), correlated variables are converted to values of linearly uncorrelated variables through orthogonal transformation, resulting in data points represented on certain principal components that best explain the variance in the dataset. It is a fast and straightforward way of visualizing evident data structure. In this study, PCA was performed as the first step to validate the trueness of the SNP markers by comparing it to previously established results.

PCA was performed in R (R Core Team 2014) using the 'adegenet' package (Jombart et al. 2014). First, to make the large genome-wide SNP sets readable by R, the VCF files were converted to PLINK format using PLINK v1.07 (Purcell 2009; Purcell et al. 2007). Then the 'adegenet' "read.PLINK" function was used to import data as a 'genlight' object. "glPCA" is a function specially designed for massive SNP datasets for faster computations. By default, it produces a centered unscaled PCA with missing data replaced by means of available observations (i.e. mean allele frequency). Number of principal components retained was chosen according to the eigenvalue screeplot. Usually, the first two axes were kept before there was a rapid decrease in eigenvalues.

2.7.2 Neighbour-joining tree

Neighbour-joining (NJ) tree shows the phylogenetic relationships among the individuals. The R 'ape' package (Paradis et al. 2004) was used to build NJ-trees (Saitou & Nei 1987). Euclidean distance matrix was computed by using the "dist" function. Missing data was excluded from all computations involving the rows within which they occur. When columns were excluded because of missing data, the sum was scaled up in proportion to the number of columns used.

2.7.3 Discriminant analysis of principal components (DAPC)

Discriminant Analysis of Principal Components (DAPC) is designed to investigate the clustering of biological populations, as a part of the 'adegenet' package (Jombart et al. 2014). It is a multivariate statistical approach that maximizes the between-group variance and minimizes within-group variance. Data transformation through PCA is the first step followed by discriminant analysis (DA), which identifies the clusters. Compared to Bayesian methods like STRUCTURE (Pritchard et al. 2000), a big advantage of this approach is that it does not presume panmixia in the dataset.

To discover any previously unrevealed structure, DAPC was implemented without *a priori*. The clusters were identified by using the 'adegenet' "find.cluster" function. Serial numbers of clusters (*k*) were tested with associating goodness of fit values (Bayesian information criterion, BIC) produced. The optimal *k* would be the one yielding the highest BIC.

2.7.4 Spatial principal component analysis (sPCA)

Spatial principal component analysis (sPCA) is designed to incorporate spatial patterns into genetic variability. It has been shown to reveal cryptic structure that could not be detected by PCA (Jombart et al. 2008). It uses a centered PCA; the scores produced not only summarize genetic variability, but also are also spatially autocorrelated. This is accomplished by using the product of data variance and Moran's *I* (Moran 1950; Moran 1948). Two types of spatial structure can be detected by sPCA. Positive components indicate global structure (such as patches and clines), i.e. a strong variance and highly positive spatial autocorrelation. In contrast, local structure (genetically different neighbours) is shown by negative components that correspond to a strong variance and a highly negative spatial autocorrelation. Thus, the sPCA method proved particularly useful for one of my research objectives - to study the population structure of *C. ribicola* in western North America in relation to the landscape features.

The "spca" function in 'adegenet' was used for this analysis. First, the regional SNP dataset was imported in to R as a "genind" object (genotypes) with missing values replaced with mean observed

genotype. It was then converted into a 'genpop' object (allelic frequency), with geographic coordinates added for each population. First, an overall isolation-by-distance (IBD) test was done with a Mantel Monte-Carlo test. In sPCA, Delaunay triangulation (type 1) connection network was chosen and the first positive and first negative components were selected. The sPCA results were overlaid on terrain maps by using two R packages, 'RgoogleMaps' (Loecher 2014) and 'png' (Urbanek 2013).

2.7.5 ADMIXTURE

Individual ancestry was assessed using ADMIXTURE 1.23 (Alexander et al. 2009). It is a maximum likelihood model-based estimation tool optimized for large SNP datasets. Both ADMIXTURE and STRUCTURE (Pritchard et al. 2000) use ancestry proportions and population allele frequencies to estimate the probability of observed genotypes. The difference is that in ADMIXTURE faster convergence is achieved by a block relaxation algorithm (Leeuw 1994) followed by a novel quasi-Newton method (Zhou et al. 2011).

ADMIXTURE also takes PLINK format as input. All four datasets were run with *K* values (the believed number of ancestral populations) ranging from 1-5, with 2,000 bootstrap replicates. The default termination criterion was chosen. The resulting Q values (ancestry coefficient matrix) for the *K* value with the lowest cross-validation (CV) error were plotted.

2.7.6 Arlequin

Several analyses, including standard diversity indices, population pairwise genetic differences, locus-by locus analysis of molecular variance (AMOVA), and *F-statistics* were done with the software Arlequin ver. 3.5.1.3 (Excoffier & Lischer 2010). Arlequin version 'arlumstat_linux' was run on a linux system while the program settings and parameters were chosen using arlequin windows version 'winarl35'. Most of the settings were kept as default, except that 100% missing data was allowed. Input files were created from filtered VCF files by a custom python script, 'VCF2Arlequin' (Taiga Lab 2014). A list of population names was fed to the program. Group structure was edited manually once an arlequin

format input file was generated. Once the input and setting files were both complete, the analyses were run using 'VCF2Arlequin'.

2.7.6.1 Standard diversity indices

In this section, expected heterozygosity per locus for a given population was calculated as $\hat{H} = \frac{n}{n-1}(1-\sum_{i=1}^{k} p_i^2)$. The expected heterozygosity for a population was simply the arithmetic mean of across all the loci. This was compared to the observed heterozygosity, which was calculated by dividing the observed heterozygotic sites/total sites per individual, and then averaging across all the individuals for one population.

Number of polymorphic sites *(S)* summarized the loci with more than one observed allele. Nucleotide diversity over *L* loci was measured by the probability two randomly chosen nucleotides are different, which is equivalent to gene diversity.

$$\hat{\pi}_n = \frac{\sum_{i=1}^{k}\sum_{j<i} p_i p_j \hat{d}_{ij}}{L}$$

$$\mathrm{V}(\hat{\pi}_n) = \frac{n+1}{3(n-1)L}\hat{\pi}_n + \frac{2(n^2+n+3)}{9n(n-1)}\hat{\pi}_n^2$$

The formulas used were (Tajima 1983; Saitou & Nei 1987). Theta ($\theta = 2M\mu$, $M = 2$ N (N = diploid population size), μ = per generation mutation rate) was estimated based on number of segregation sites (S) as a population genetic diversity measure using the Tajima (1983) method implemented in Arlequin (Excoffier & Lischer 2010). Lastly, genetic distances between DNA sequences were computed as pairwise difference – number of loci for which two haplotypes are different.

2.7.6.2 Locus-by-locus analysis of molecular variance (AMOVA)

Analysis of Molecular Variance (AMOVA) was used to describe the partitioning of genetic variation within/among different hierarchical levels. In this study, the model employed 'within individual' (WI), 'among individuals within populations' (WP), 'among populations within groups' (AP/WG) and 'among groups' (AG). Four groups were classified - Korea, Russia, and eastern and western NA. Populations represented the provenances of the samples. Four types of fixation indices were produced: F_{IT} corresponds to WI; F_{IS} denotes WP; F_{SC} shows AP/WG; lastly, F_{CT} denotes AG. F_{IT} and F_{IS} are also known as inbreeding coefficients.

σ^2_T = total molecular variance

σ^2_a = covariance component due to differences among the G populations

σ^2_b = covariance component due to differences among individuals in different populations within a group

σ^2_c = covariance component due to differences among individuals within a populations

σ^2_d = covariance component due to differences within individuals

Significance of *F-statistics* was tested by 1,000 non-parametric permutations. Population specific F_{IS} was computed to check inbreeding. The pairwise Euclidean squared distance matrix was chosen for AMOVA.

Due to missing genotypes in the SNP dataset, a locus-by-locus AMOVA was preferred over a standard AMOVA (Excoffier & Lischer 2010). In locus-by-locus AMOVA, variance and *F-statistics* are produced for each locus individually. When there is missing data for one of the two alleles in a diploid individual, that individual is removed from this locus's analysis. Synthetic *F-statistic* estimators for each level were summarized as ratios between the sum of variance components at the given level and the total (Weir 1996; Weir & Cockerham 1984). The results from this analysis with missing data will differ from the standard AMOVA because the degrees of freedom would vary by loci.

For the global and continental dataset, AMOVA was run on pre-defined groups, without within-individual level variance. In this case, $\mathbf{F_{CT}} = \sigma^2_a / \sigma^2_T$, $\mathbf{F_{ST}} = (\sigma^2_a + \sigma^2_b) / \sigma^2_T$, $\mathbf{F_{SC}} = \sigma^2_b / (\sigma^2_b + \sigma^2_c)$. For the regional dataset, AMOVA analysis was done on a finer scale for the western North American populations to compare Coastal and Interior groups; for the eastern cluster, AMOVA was used to test the difference between the two groups separated by the great lakes. Within-individual level variance was considered. This makes $\mathbf{F_{ST}} = \sigma^2_a / \sigma^2_T$, $\mathbf{F_{IT}} = (\sigma^2_a + \sigma^2_b) / \sigma^2_T$, $\mathbf{F_{IS}} = \sigma^2_b / (\sigma^2_b + \sigma^2_c)$.

2.7.6.3 Population pairwise genetic differences

Wright's fixation index (1965; 1951; 1931) is one of the most widely used measures for genetic differentiation in population studies. Arlequin adopted the version developed by Weir & Cockerham (1984), which is based on Wright's model but takes sample size effects into account. Pairwise F_{ST} (Weir & Cockerham 1984) was calculated and interpreted as short-term genetic distances between populations. The significance of obtained F_{ST} values was assessed by permuting haplotypes 110 times between populations and generating an associated *P*-value.

3 Results

3.1 Sequencing, De-multiplexing and Mapping

3.1.1 GBS pilot run

We conducted a pilot run of GBS to test whether it was applicable to *C. ribicola*. Since aeciospore collections always yield lower amounts of DNA, we wanted to assess the effect of DNA concentration on the outcome. Initially, 200 ng of DNA was required by the sequencing platform for GBS. This pilot run determined that as low as 6.25 ng of DNA per sample was adequate for a GBS reaction as different DNA concentrations did not cause much variation in the library profiles (results not shown). In order to include more samples, yet still retain a good quality of the DNA, we aimed for a final amount of 20 ng per sample. Depending on the original concentrations of the extractions, samples > 2 ng/µl DNA were normalized to 2 ng/µl (x 10 µl).

Principal components analysis (PCA) (Figure 4) shows that the different concentrations of the same sample overlapped on the PCA, indicating that the identification of the underlying SNPs is not influenced by DNA concentration. In addition, a separation between the eastern and western samples was observed. This is consistent with previous findings which demonstrated that the eastern and western *C. ribicola* populations are genetically differentiated (Hamelin et al. 2000).

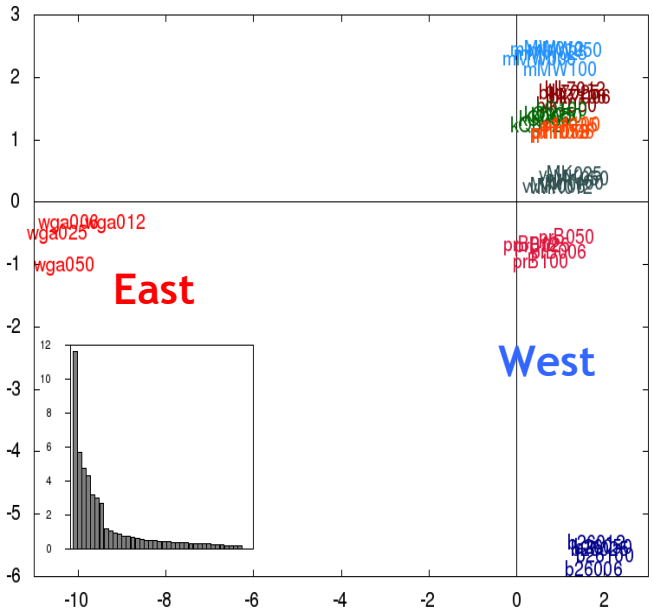

Figure 4 Principal Component Analysis of eight samples of *Cronartium ribicola* genotyped with GBS at different DNA concentration, ranging from 6.25 ng to 100 ng.

The conclusions drawn from the pilot run are: i) GBS library of *C. ribicola* can be successfully constructed with different DNA concentrations without affecting the outcome and that even low DNA amounts, as little as 6.25 ng yields reliable results; ii) Standard GBS *Pst*I/*Msp*I protocol is suitable for *C. ribicola*; iii) GBS is a feasible and inexpensive approach for SNP discovery in *C. ribicola*.

3.1.2 Final GBS dataset

The 192 *C. ribicola* samples were split into four groups of 48 that were indexed and sequenced on the Illumina platform. The four independent Hiseq sequencing runs generated a total of 158.6 Gb of data comprising of 623.5 million raw reads. The indexed reads were de-multiplexed into separate files, each representing one individual. A total of 97.3% (606.6 million) of the reads were retained after de-multiplexing and the remaining 2.7% unmatched reads were discarded. Out of these 606.6 million reads 478.5 million (78.9%) were successfully mapped onto the reference *C. ribicola* genome. The percentage

of reads mapped per individual ranged from 0.51% to 98.2%, with an average of 82.3%. Among the 192 individuals, 159 (82.8%) had more than 70% reads mapped.

The *C. ribicola* reference genome contains 71.2 million bp spread across 19,901 contigs. As this study utilized GBS approach we expected a reduced coverage of the reference genome. On average, the total mapped reads represented ~2% of the *C. ribicola* reference genome, with at least 2-fold coverage. Generally, less than 1% of the genome had over 50-fold coverage. Genome coverage per individual varied, with the lowest being 0.02%, highest 13%, and a mean of 3%.

3.2 SNP Calling and Filtering

Initially, SNPs were called upon all the samples and 180,548 raw SNPs were identified. After applying the filters as described in the Methods section, 25 individuals displayed high levels of missing data (>30%) and were eliminated. An additional individual from Smithers, which showed excessive amount of heterozygotes, were removed from the dataset.

For further analyses, we generated three datasets (Table 3): 1) the global dataset comprised 135,905 raw SNPs from 166 individuals from Korea, Russia, and North America that yielded 8,020 (5.90%) SNPs after filtering; 2) the North American dataset comprised 106,606 raw SNPs from 157 individuals from eastern and western North America, yielding 4,510 (4.23%) SNPs after filtering; and 3) the regional dataset was filtered more stringently with two extra filters compared to the above datasets. The two filters eliminated SNPs that only occurred once in the whole sample set (minor allele count of one) and loci that were in Hardy-Weinberg disequilibrium. As a result, 897 (1.79%) out of 50,196 SNPs passed the filters in the eastern North American set (38 individuals); the western North American set had 118 individuals and only 365 (0.41%) out of 88,027 SNPs were retained after filtration. For all datasets, filtered SNP sets were used for subsequent analyses. Two neighbour-joining trees (Figure 16 & 17 Appendix) showed that the lab replicates (with or without WGA) always fell next to the original samples

and formed exclusive clades, which indirectly proved the reliability and consistency of the SNPs generated by GBS.

Table 3 Statistics of SNP markers before and after filtering.

DATASET	N	STARTING # OF VARIANTS	SNPS AFTER FILTERING	% REMAINING
RAW	191	180,548	N/A	N/A
GLOBAL*	166	135,905	8,020	5.90%
NORTH AMERICA*	156	106,606	4,510	4.23%
EASTERN NORTH AMERICAN*	38	50,196	897	1.79%
WESTERN NORTH AMERICAN*	118	88,027	365	0.41%

*Removed individuals with >30% missing data and SM18 (a hybrid individual) from the raw SNP set.

3.3 Population Structure Analyses

3.3.1 Global structure of *Cronartium ribicola*

First, to reveal the general structure of all the individuals from Asia, North America and Europe, a PCA was performed on the global dataset. The first principal component (PC1) explained 33.3% of the total variation and separated Asian samples, with positive scores from North American ones (Figure 5). The second principal component (representing 5.4% of the variation) separated the eastern and western North American individuals into two independent clusters. Both the Russian and Finnish samples had positive PC2 values and were most closely related to eastern North American samples, while Korean samples had slightly negative values, similar to the western North American cluster.

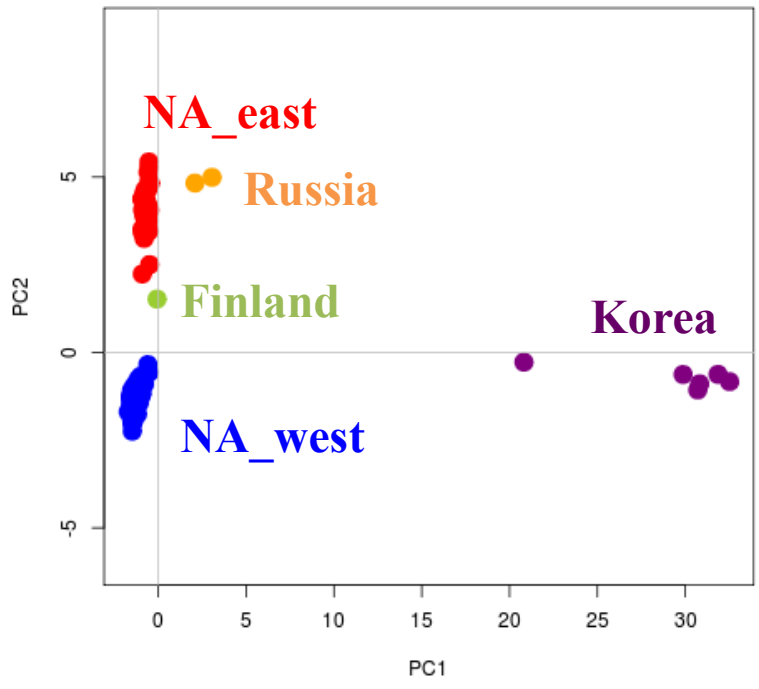

Figure 5 Principal Component Analysis of the global *Cronartium ribicola* dataset, with 8,020 SNPs.

The Neighbour-joining (NJ) tree (Figure 6) displays a similar profile - the eastern and western samples form distinct clusters; the Finnish and Russian samples are nested within the eastern cluster while the Korean samples are the most distant group, with a long branch.

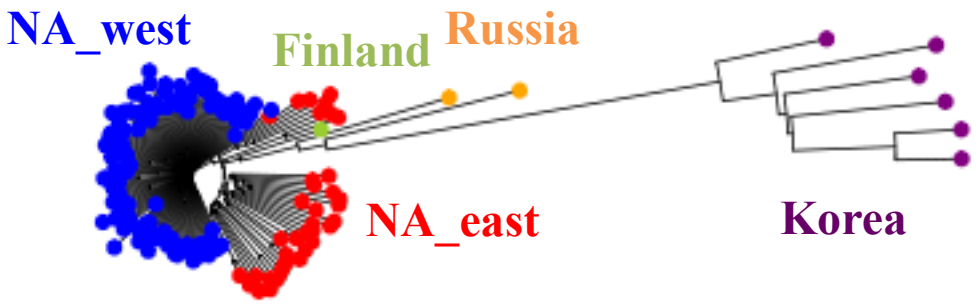

Figure 6 Neighbour-joining tree of the global *Cronartium ribicola* dataset, with 8,020 SNPs.

In Discriminant Analysis of Principal Components (DAPC), the lowest Bayesian Information Criterion (BIC) value was found to be associated with $K = 4$. Therefore, four clusters were chosen in DAPC (Figure 7). This is congruent to results by PCA and NJ-tree. Again, the only Finnish sample fell into the eastern North American cluster in DAPC.

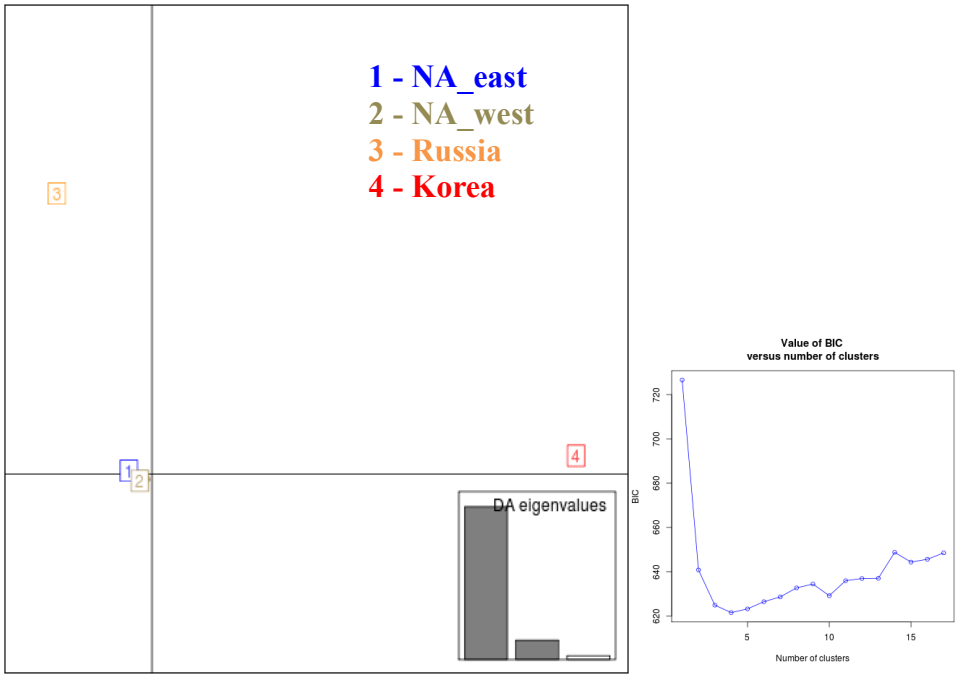

Figure 7 Discriminant Analysis of Principal Components of the global *Cronartium ribicola* dataset, with 8,020 SNPs.

The optimal number of clusters provided by ADMIXTURE was $K = 3$ (Figure 8a). The three groups defined by ADMIXTURE could be broadly classified as Korea, eastern North America and western North America, even though there was some admixture between some eastern and western North American samples (Figure 8b). The Russian samples were assigned into eastern North American cluster with some admixture from the Korean group. When repeating the analysis using $K = 4$, the western group further divided into two sub-groups, but it still failed to separate the Russian and Finnish samples from the eastern North American cluster, possible due to the small sample size (result not shown).

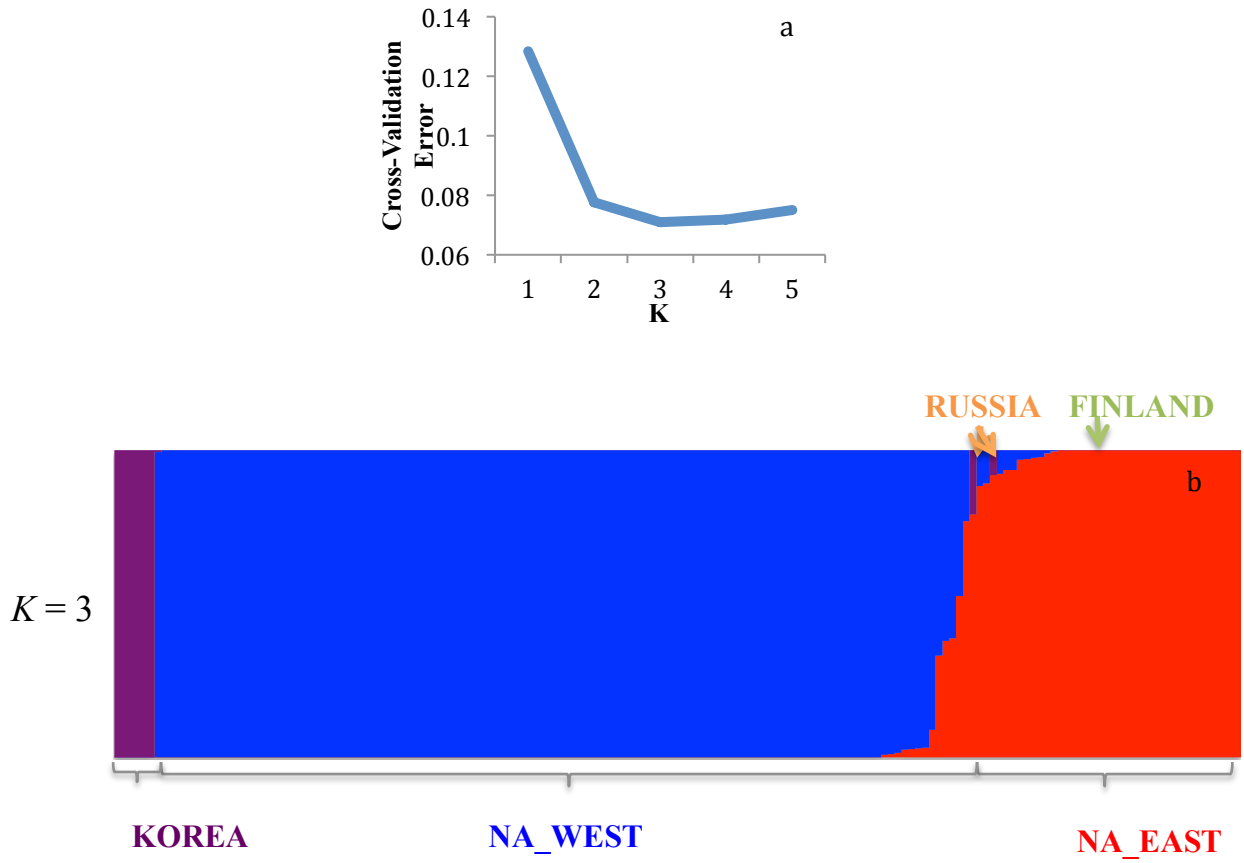

Figure 8 a) Plot of k values vs. cross-validation error; b) ADMIXTURE result of the global *Cronartium ribicola* dataset.

Admixture was observed in some North American samples as well. The eastern individuals generally had a low amount of admixture from the western group, with the highest level of admixture at 20%. In contrast, some western individuals were heavily admixed with the eastern group. Specifically, USMR06 and ABB02 showed the highest proportion of admixture (~50%) from the eastern group; PR06, ABPH01, and ABCR02 were moderately admixed (~35%); the rest of them (ABPM03, PE03, TE06, TE09, SM02, MW07, PE05) were lightly admixed (less than 10% of eastern profile). These individuals came from two broad locations –Alberta/Montana and Vancouver Island.

The highest observed heterozygosity and nucleotide diversity were observed in Russian samples, followed by Korea, and then eastern North America (Table 4). Western North American samples

possessed the lowest diversity. θ_s values did not follow the exact pattern, especially in the western North American group, where it is higher than the eastern North American group and even the Korean group. Among the eastern samples, Minnesota and Wisconsin had the lowest diversity, similar to the western ones. For all populations, the expected heterozygosities were very close to the observed heterozygosity (Figure 9), with the exception of the Asian samples where there is a slight excess of expected heterozygosity.

Table 4 Genetic diversity indices of the global *Cronartium ribicola* dataset, with 8,020 SNPs.

	N	% POLYMORPHIC SITES	Pi	θ_s	H_o	H_e	H_{op}	H_{ep}
RUSSIA	2	13.9%	0.070	0.076	0.075	0.083	0.600	0.598
KOREA	6	16.1%	0.043	0.053	0.052	0.061	0.390	0.376
EASTERN NORTH AMERICA	38	10.1%	0.024	0.021	0.030	0.027	0.325	0.276
WESTERN NORTH AMERICA	118	37.2%	0.013	0.062	0.018	0.018	0.057	0.050

N = sample size, % poly. sites = proportion of sites that are polymorphic, Pi = nucleotide diversity averaged over all loci, θ_s = number of segregation sites averaged over all loci, H_o, H_e = observed/expected heterozygosity averaged over all loci, H_{op}, H_{ep} = observed/expected heterozygosity averaged for polymorphic loci only.

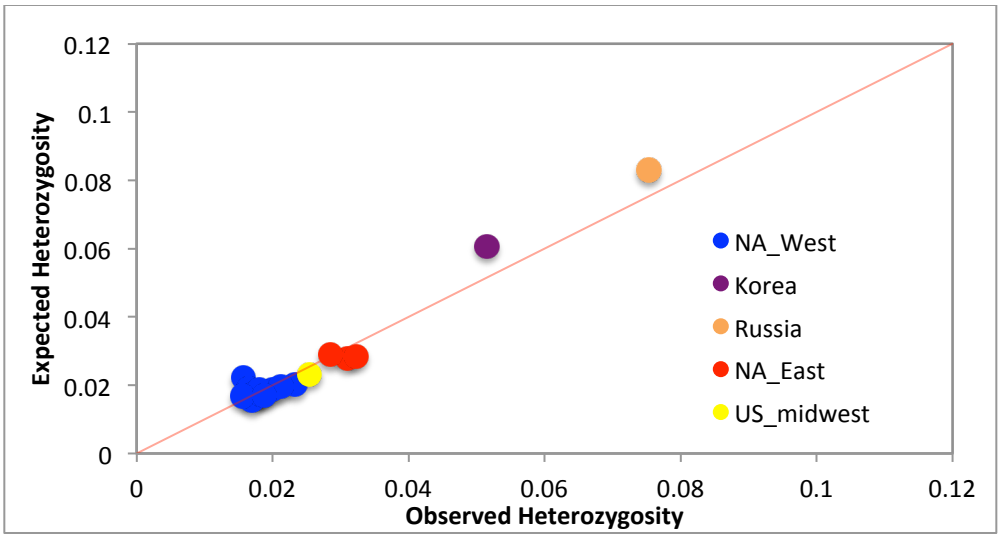

Figure 9 Observed vs. expected heterozygosity (averaged over all loci) of each population in the global dataset of *Cronartium ribicola*.

Singletons refer to the SNPs that only appear once throughout the whole dataset (private alleles) and doubletons denote homozygotic private alleles. Individually, the highest numbers of singletons/doubletons were found in the two Russian samples (294 and 174 respectively), followed by three Korean samples (range between 104 and 121, Table 9 Appendix). Population-wise, the Korean group contained the most total number of singletons, followed by Russia and western North America.

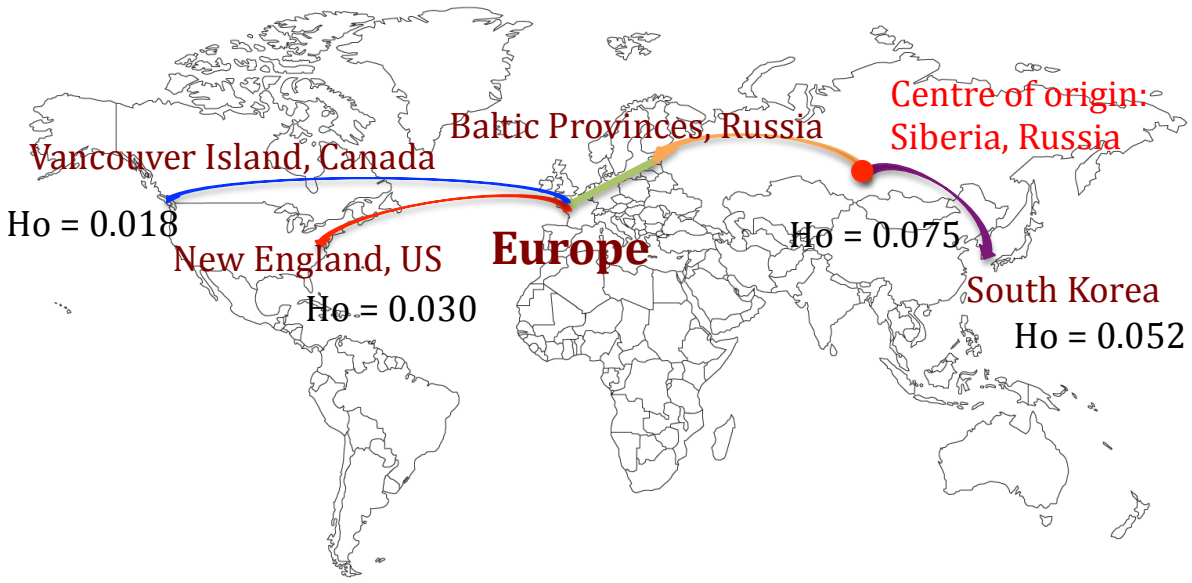

Figure 10 Proposed spread of *Cronartium ribicola* across the globe from its centre of origin: Siberia, Russia.

An analysis of molecular variance (AMOVA) confirms the patterns observed in previous analyses. Most of the variation resides among groups (Russia, Korea, eastern and western North America (57.34%, Table 5). Genetic differentiation among groups (F_{CT}) was 0.573 ($P < 0.001$), which indicates that the four groups are distinct clusters (Hartl & Clark 1997). The differentiation among populations within groups was small (0.003), yet significantly different from 0. Lastly, a high amount of variation lay within populations (42.5%) and all populations were strongly differentiated ($F_{ST} = 0.575$, $P < 0.001$).

All pairwise F_{ST} values were statistically significant except the one between Russia and Korea (Table 6). Korea is the most distant group from the rest, with the largest distance lying between Korea and western North America (0.890). Russia was also found to be distinct from the North American groups. The genetic distance between western and eastern North America was 0.164, which suggests genetic differentiation but not as large as between North American and Asian samples.

Table 5 Summary of all Analysis of Molecular Variance (AMOVA) results.

HIERARCHICAL STRUCTURE	SOURCE OF VARIATION	SUM OF SQUARES	VARIANCE COMPONENTS	% VAR.	*F-STATISTICS*	P-VALUES
NA_WEST VS. NA_EAST VS. KOREA VS. RUSSIA	Among groups (Va)	14460.84	111.33	57.34	F_{CT} = 0.573	**< 0.001**
	Among populations within groups (Vb)	1284.37	0.28	0.14	F_{SC} = 0.003	**< 0.001**
	Within populations (Vc)	21906.40	82.55	42.51	F_{ST} = 0.575	**< 0.001**
	Total	37651.62	194.16	100.00		
NA_WEST VS. NA_EAST	Among groups (Va)	1960.47	17.95	17.31	F_{CT} = 0.173	**< 0.001**
	Among populations within groups (Vb)	1489.13	1.10	1.06	F_{SC} = 0.013	**< 0.001**
	Within populations (Vc)	21395.38	84.65	81.63	F_{ST} = 0.184	**< 0.001**
	Total	37651.62	194.16	100.00		
AMONG NA_WEST POPULATIONS	Among populations (Va)	461.26	0.91	3.58	F_{ST} = 0.036	> 0.001
	Among individuals within populations (Vb)	2160.60	0.01	0.05	F_{IS} = 0.001	> 0.001
	Within individuals (Vc)	2480.00	24.54	96.37	F_{IT} = 0.036	**< 0.001**
	Total	5101.85	25.46	100.00		
NA_WEST: COASTAL VS. INTERIOR	Among groups (Va)	61.47	0.23	0.90	F_{CT} = 0.009	**< 0.001**
	Among populations within groups (Vb)	399.78	0.79	3.10	F_{SC} = 0.031	**< 0.001**
	Within populations (Vc)	4640.60	24.55	96.00	F_{ST} = 0.040	**< 0.001**
	Total	5101.85	25.57	100.00		
NA_WEST: AB+USMR VS. BC	Among groups (Va)	41.51	0.17	0.66	F_{CT} = 0.007	> 0.001
	Among populations within groups (Vb)	419.75	0.88	3.42	F_{SC} = 0.034	**< 0.001**
	Within populations (Vc)	4640.60	24.55	95.92	F_{ST} = 0.041	**< 0.001**
	Total	5101.85	25.60	100.00		

HIERARCHICAL STRUCTURE	SOURCE OF VARIATION	SUM OF SQUARES	VARIANCE COMPONENTS	% VAR.	*F-STATISTICS*	P-VALUES
NA_WEST: COMPARISON AMONG PINE HOSTS	Among groups (Va)	49.97	0.12	0.48	F_{CT} = 0.005	> 0.001
	Among populations within groups (Vb)	411.28	0.86	3.35	F_{SC} = 0.034	**< 0.001**
	Within populations (Vc)	4640.60	24.55	96.17	F_{ST} = 0.038	**< 0.001**
	Total	5101.85	25.53	100.00		
NA_WEST: NATURAL STAND VS. PLANTATION	Among groups (Va)	121.37	-0.08	-0.32	F_{CT} = -0.003	> 0.001
	Among populations within groups (Vb)	339.89	0.97	3.80	F_{SC} = 0.038	**< 0.001**
	Within populations (Vc)	4640.60	24.55	96.52	F_{ST} = 0.035	**< 0.001**
	Total	5101.85	25.43	100.00		
NA_EAST: AMONG POPULATIONS	Among populations (Va)	602.92	6.86	4.66	F_{ST} = 0.047	> 0.001
	Among individuals within populations (Vb)	4118.35	-7.45	-5.06	F_{IS} = -0.053	> 0.001
	Within individuals (Vc)	5153.00	147.75	100.40	F_{IT} = -0.004	> 0.001
	Total	9874.27	147.16	100.00		
NA_EAST: MN+WI VS. THE REST	Among groups (Va)	328.52	17.72	11.20	F_{CT} = 0.11	**< 0.001**
	Among populations within groups (Vb)	274.41	-0.28	-0.18	F_{SC} = -0.002	> 0.001
	Within populations (Vc)	9271.35	140.77	88.98	F_{ST} = 0.11	**< 0.001**
	Total	9874.27	158.21	100.00		

Table 6 Pairwise F_{ST} between the Korean, Russian, eastern and western North American groups of *Cronartium ribicola*.

	KOREA	RUSSIA	NA_WEST	NA_EAST
KOREA	0			
RUSSIA	0.683	0		
NA_WEST	0.890	0.582	0	
NA_EAST	0.825	0.409	0.164	0

Red indicating statistically significant

3.3.2 North American population structure of *Cronartium ribicola*

To focus on the population structure of *C. ribicola* in North America, we generated a new dataset that excludes the Asian samples. This allowed us to discern patterns that were not apparent in the analysis of the entire dataset.

Principal component analysis shows that the western and eastern North American populations of *C. ribicola* are clearly separated by PC1, which explains the highest variance in the dataset (5.86%, Figure 11a). The western North American samples form a distinct genetic cluster, distinct from the eastern North American samples, which are more scattered. Besides this obvious pattern, PC2 (representing 1.04% of the total variation), separates several eastern individuals from the rest of the cluster (i.e. QC24, QC15, MN02, WI01, QC01, MN01, Figure 11 a, b & c).

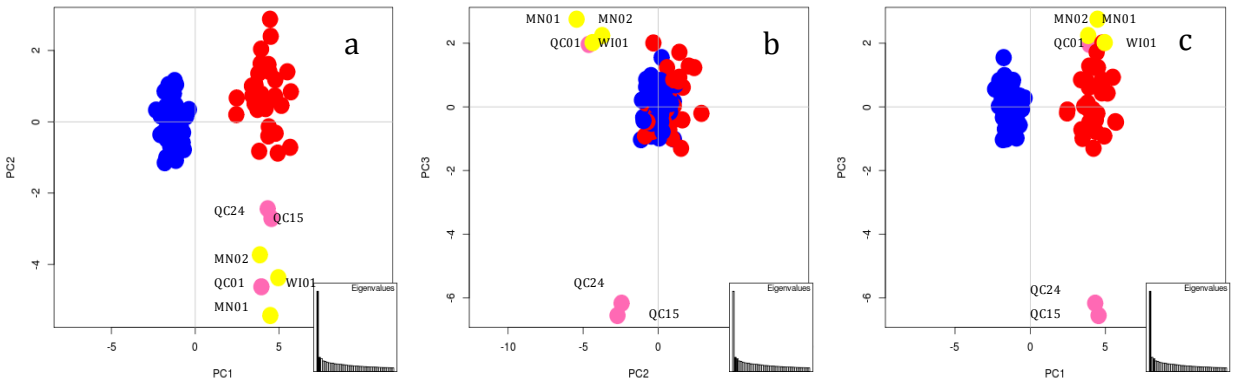

Figure 11 a, b, c = Principal Component Analysis of North American samples of *Cronartium ribicola* analyzed with 4,510 SNPs.

The NJ-tree (Figure 18 in the Appendices) also clusters eastern North American and western North American separately. DAPC and ADMIXTURE (Figure 19 & 20 in the Appendices), one using multivariate statistical approach while the other using maximum-likelihood model-based methods, both supported $K = 2$ as the optimal clustering. DAPC successfully assigned each individual back to its geographic origin (eastern or western North America).

All measures of heterozygosity (observed and expected) and nucleotide diversity were higher (two to five times) in eastern than in western populations (Table 7). However, θ_s showed the opposite trend, being approximately three times larger in western than in eastern populations. θ_s is almost five times higher than the nucleotide diversity within the western group, a trend absent from eastern populations. This result can be explained by the observation that more singletons were found in the western than in the eastern samples. Within the western group, Smithers possessed the highest number of singletons (Table 10 in the Appendices). The genetic diversities of each North American population were summarized in Table 11 in the Appendices.

However, there were three times more individuals sampled in western North America. Hence, to account for any possible sampling bias, 20 random subsets of 38 individuals (equal the population size of the eastern group) from the western population were chosen randomly and the average proportion of polymorphic sites was calculated. More polymorphic sites were found in all these subsamples than in eastern samples (average over the 20 subsets ~28%).

Table 7 Genetic diversity indices of the North American dataset of *Cronartium ribicola*, with 4,510 SNPs.

REGIONS	N	% POLYMORPHIC SITES	Pi	θ_s	H_o	H_e	H_{op}	H_{ep}
NA_WEST	118	66.7%	0.024	0.110	0.032	0.034	0.058	0.051
NA_EAST	38	19.0%	0.045	0.039	0.056	0.052	0.318	0.273

N = sample size, % poly. sites = proportion of sites that are polymorphic, Pi = nucleotide diversity averaged over all loci, θ_s = number of segregation sites averaged over all loci, H_o, H_e = observed/expected heterozygosity averaged over all loci, H_{op}, H_{ep} = observed/expected heterozygosity averaged for polymorphic loci only.

The difference between the eastern and western groups was tested without the Asian samples by AMOVA. Results show that 17.31% of the total variation was significantly attributed to the differentiation between the two groups (Table 5). The variation among populations within groups was much smaller (1.06%). Again, most of the variation was found within populations (81.63%). The *F-Statistics* on all three levels were significant.

3.3.3 Regional population structure in western North American populations

In accordance with our hypothesis that there is more structure in western North America due to the various landscape and hosts, we looked further into the western dataset separately. There was no obvious pattern in the PCA of the western samples (Figure 12), except for a few outlier individuals. Among these, there is a contrast between the Coastal outliers and the Interior ones. The Coastal individuals (green circles) all had negative PC1 values and positive PC2 values; whereas the orange/red circles represent interior BC samples which had positive PC1 values but negative PC2 values. No clear structure was seen in the NJ-tree (Figure 16 in the Appendices).

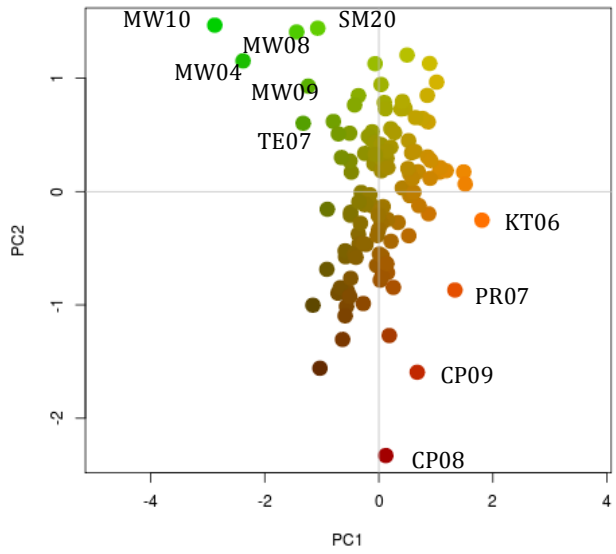

Figure 12 Principal Component Analysis of western North American *Cronartium ribicola* dataset.

The "find.clusters" function in the 'adegenet' package was unable to detect any noticeable clustering in either the western or eastern groups. Therefore, Discriminant Analysis of Principal Components analysis was not performed. The histogram of the isolation-by-distance test demonstrates that the observed value falls right in the middle of the simulated ones, with a *P*-value of 0.555 (Figure

13). Thus, it failed to reject the null hypothesis of absence of spatial structure in the whole data, meaning that there is no significant correlation between genetic distance and geographic distance.

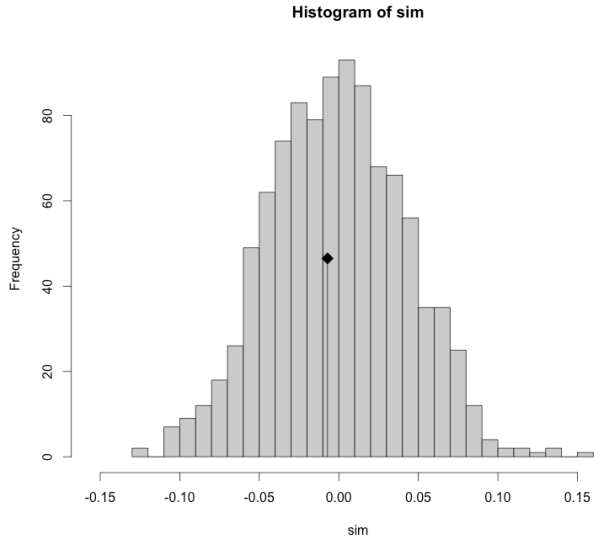

Figure 13 Histogram of Mantel test (isolation-by-distance) amongst western *Cronartium ribicola* populations.

Despite the overall lack of spatial structure, the global structures of the western dataset were significant at the 5% level (I = 0.676, P = 0.014), but not the local ones (I = -0.281, P = 0.948). Global structures describe positive correlation between geographic distance and genetic distance. Local structures, on the other hand, appear when neighbouring sites appear to be dissimilar, corresponding to negative spatial autocorrelation.

The global structures in western North America are shown as a major separation between the Coastal and Interior groups. All Coastal populations, together with Pemberton and Prince George had positive sPCA scores while the rest of the populations were mainly negative. What is surprising is that the northernmost population Smithers has more similarity to the Rocky mountain populations than to the Coastal populations or its closest neighbour Prince George (Figure 14). This pattern is cryptic and was not detected formerly by PCA or DAPC.

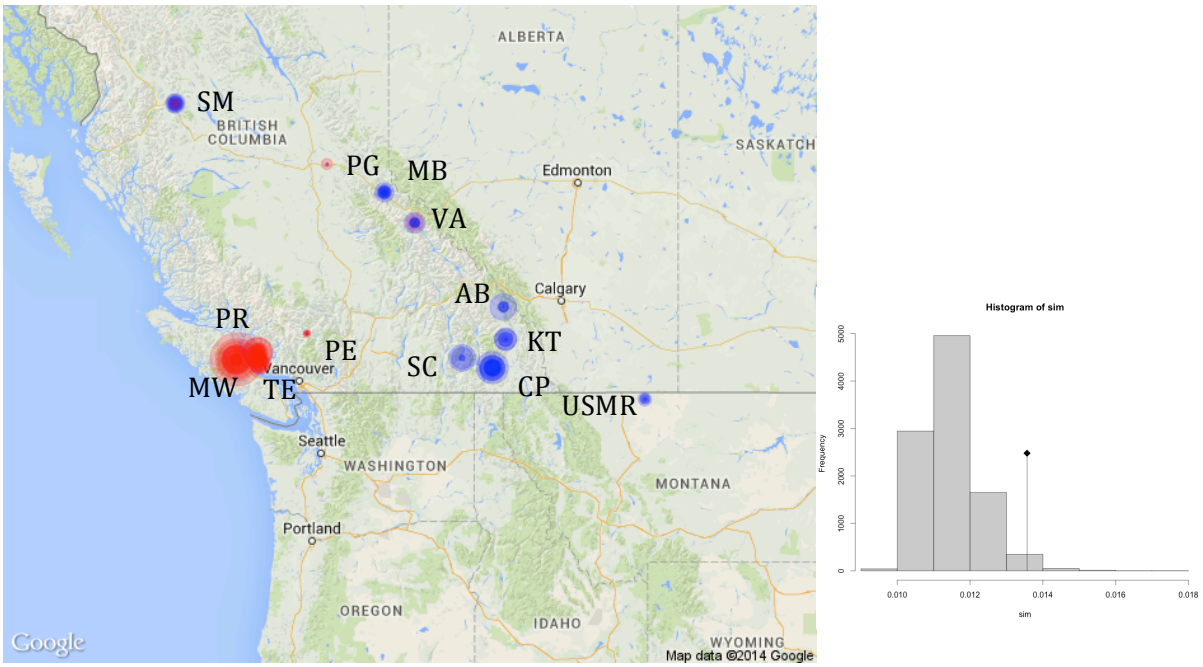

Figure 14 Spatial Principal Component Analysis – global structure of western group of *Cronartium ribicola* shown on terrain map and the histogram of its associated Monte-Carlo test.

Red = positive scores, blue = negative scores, size of the circles proportional to the scores.

Based on the clustering information provided by sPCA (Figure 14), the western cluster was further divided into Coastal and Interior groups. Nucleotide diversity, θ_s and heterozygosity are all higher in the Coastal than in the interior *C. ribicola* samples. The pairwise F_{ST} between the two groups was 0.013 and significant (Table 8). On a population-level, the highest expected heterozygosity was found in Mount Washington (MW), and it gradually decreases from there (Figure 15). The only exception is Montana, US, that possesses diversity comparable to the Coastal populations.

Table 8 Genetic diversity indices and pairwise F_{ST} of the western North American *Cronartium ribicola* dataset, with 365 SNPs.

REGIONS	N	Pi	θ_s	H_e	PAIRWISE F_{ST}	COAST	INTERIOR
COAST*	46	0.113	0.172	0.148	**COAST**	0	
INTERIOR	72	0.094	0.163	0.132	**INTERIOR**	**0.013**	0

N = sample size, Pi = nucleotide diversity averaged over all loci, θs = number of segregation sites averaged over all loci, He = expected heterozygosity averaged over all loci. *Coast: PE, PR, TE, MW, and PG

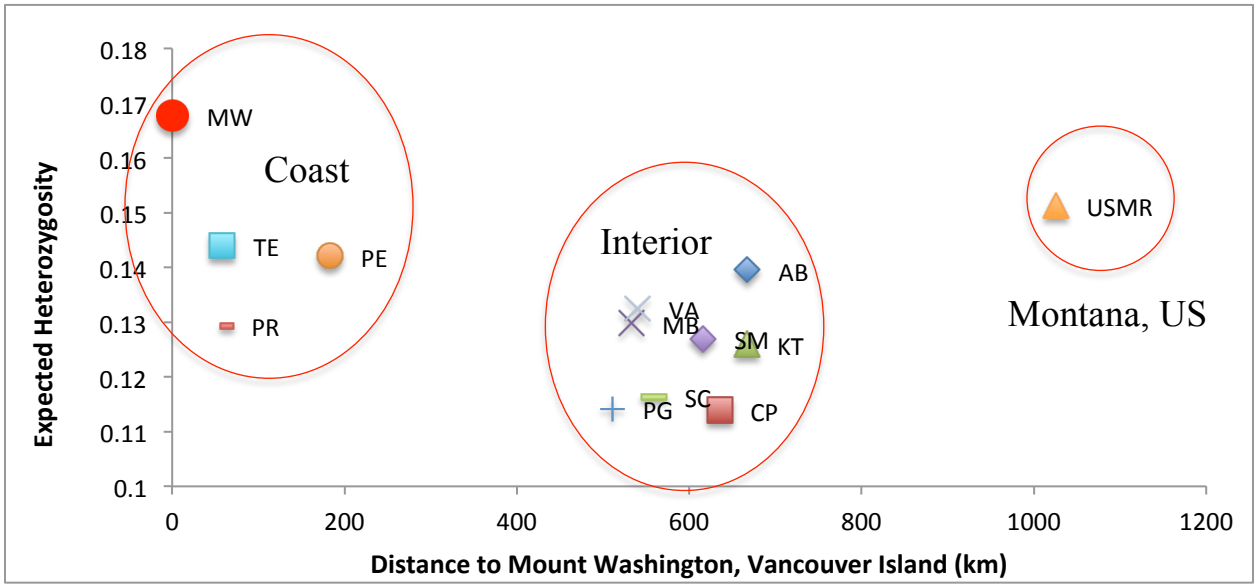

Figure 15 Expected heterozygosity plotted against distance to Mount Washington of western North American *Cronartium ribicola* populations.

AMOVA was performed for the western group on a population level (Table 5). Small amount of variation (3.58%) was found among populations, and the associated F_{ST} turned out to be insignificant. F_{IS} was not significant either, indicating that there is no inbreeding within populations. The within-individual variation was very large (96.37%), and F_{IT} was 0.036 ($P < 0.001$).

To test the statistical difference between the Coastal and Interior groups, another AMOVA was designed and implemented. Although the amount of variation between the two groups was relatively small (0.74%), it was statistically significant from zero ($F_{CT} = 0.0074$, $P < 0.001$). The low F_{CT} value suggests that little genetic difference exits between these two groups. Both F_{SC} and F_{ST} were significant, indicating that the populations within the two groups were inbreeding. Lastly, the pairwise F_{ST} value between the two groups was 0.0089 (Table 5) and it was statistically significant.

We investigated further the populations from Alberta (AB) and Montana (USMR) that comprised the most heavily admixed western individuals. AMOVA was utilized to check whether the AB/USMR

group was significantly different from the rest of the western populations. The results show that there was no significant difference between this group and the rest. Similarly, no difference was found among various pine species or between natural and plantation sites (Table 5).

4 Discussion

4.1 Global Introduction Pathway of *Cronartium Ribicola*

This is the most extensive population genetic study of *Cronartium ribicola*, with the largest number of markers covering 19 populations from three continents. Using 8,020 SNPs generated from GBS, we found that clustering was strongly influenced by geographic origin, with Korean, Russian, western and eastern North American populations of *C. ribicola* forming four distinct genetic populations.

Our hypothesis was that the center of diversity for *C. ribicola* would be Asia, where the pathogen was historically thought to originate. The first historical record of *C. ribicola* was made by Dietrich in 1856 (Spaulding 1911), in an annotated list of fungi of the Baltic provinces of Russia (the westernmost part of Russia). There were documented movements of diseased *R. nigrum* and pine from Russia to Europe. An Asian origin of *C. ribicola* has been speculated before. Leppik (1967) and Spaulding (1929) believed that the origin of *C. ribicola* was in northeastern Asia, ranging from east of the Ural Mountains of Russia, through central Siberia, to eastern Asia along the Pacific coast and to the Himalayas in the south. Stewart (1906) proposed that the original host of *C. ribicola* is *P. cembra*, a common pine species in Russia. Tubeuf (1917) also believed that the rust had originated in northern Asia. Spaulding (1929) hypothesized the origin to be Siberia, which is within the range of *P. cembra* and some *Ribes* species. He suggested that it would have been possible for the early Russian travelers and explorers to carry *C. ribicola* from Siberia to some botanical gardens in Russia, where *P. strobus* was present, and later to western Europe through the movement of diseased *P. strobus* seedlings. Although not discussed explicitly, Spaulding (1929) implied that the spread from Russia to Asia was highly likely due to the favourable climatic conditions and availability of five-needled pines.

The highest genetic diversity is expected at the center of diversity where a species originates (Vavilov 1926). Our study for the first time generated estimates of genetic diversity from samples from

all three continents, including the proposed center of diversity, allowing this comparison to be made. The pattern of diversity we observed (heterozygosity 3-5 times higher in Asian samples than in the North American or European ones, Table 4 & Figure 9) is consistent with our hypothesis that the pathogen originated from Asia and subsequently migrated to Europe and from there to North America. Although our sampling size is small, observed heterozygosity is independent of sample size. According to above results, the proposed global introduction pathway of *C. ribicola* from Siberia is illustrated in Figure 10. The two Russian samples included in the current study were collected in Tomsk and Siberia, respectively. Tomsk is located in the western part of Russia, closer to the Ural Mountains, whereas Siberia is in central Russia. Although they were assigned to the same clusters and the PCA and NJ analyses grouped them closely, the Siberian sample had a higher heterozygosity than the Tomsk sample. This would be in agreement with the proposed western migration from a Siberian origin. From there, it spread to the Baltic provinces and later to Europe. From Europe, it was introduced to North America, on diseased *P. strobus* seedlings (Figure 10).

The Korean samples were strikingly different from all other samples. This agrees with previous studies (Richardson 2008). It is possible that these samples represent a distinct species or subspecies. Distinct species of *C. ribicola* have been reported, but not formerly described (Kim et al. 2010). Prior to this, the phylogeographic structure among Asian, European, and North American populations of *C. ribicola* was studied using DNA sequences from four nuclear loci (Richardson et al. 2008). Maximum likelihood, Bayesian and maximum parsimony all produced similar results, where three clades were identified: Korea and China made up the first and most distant clade, Japan formed one clade on its own, and the last clade comprised samples from the US and Germany. The source of *C. ribicola,* however, was not covered in their sampled locations in Korea, China or Japan. Thus, the source of the Eurasian and North American spread of *C. ribicola* remained unclear. It is possible that they were derived from the Russia isolates, as the Korean samples possess a lower heterozygosity than the Russian samples.

4.2 Genome-Wide Comparison Between Eastern and Western North American Populations of *Cronartium Ribicola*

The large genetic differentiation that we observed was reported before using RAPD markers and a small set of SNPs (Hamelin et al. 2000; Brar 2012). Our results using over 4000 SNPs confirm this pattern. However, previous studies did not have the power to discover novel patterns of variation or to quantify heterozygosity. As expected and previously observed, the expected and observed heterozygosity was approximately twice higher in eastern than in western North America. This is consistent with record of importation of several millions of white pine seedlings in eastern North America from several locations in Europe (Spaulding 1929), but contrasts with the single recorded introduction to western North America from France (Mielke 1943). This would have created a founder effect, and possibly followed by population bottleneck, resulting in reduced genetic diversity and the patterns of diversity that we observed in western North America.

We observed a larger percentage of polymorphic loci and more private alleles in western than in eastern samples, reduced in θ_s being larger than Pi. This result presents the hallmark of a population bottleneck following by population expansion. An additional possible explanation is that the Western populations of *C. ribicola* are exposed to more hosts and more climate variation than their eastern counterparts. This could create selection pressure that would favour rare alleles that provide selective advantages and eventually generate novel adaptations to these conditions. Another possible explanation is that hybridization between *C. ribicola* and a native rust, *C. comandrae*, could introduce novel variants, especially if such events lead to introgression. The fact that high diversity and admixture were found in areas where the hybrid rust has been observed Alberta and Montana would support this explanation.

Samples from the US Midwest (Minnesota and Wisconsin) had heterozygosities that were comparable to those of western populations. Similar results were observed before (Brar 2012). In the present study, the population specific F_{IS} was found to be -0.166, with observed heterozygosity slightly

higher than expected, indicating that inbreeding was not happening within this group. This could be explained by a recent genetic drift followed by a population bottleneck. The Great Lakes were suggested as a barrier to gene flow between Midwest and other eastern populations (Brar 2012). Alternatively, it is possible that *C. ribicola* in the Midwest had slightly different origins than the main eastern North American genetic population.

Within the western genetic cluster, Alberta and Montana populations had high heterozygosity, comparable to those east populations (Table 11). Also, a few individuals from these two populations were heavily admixed with the eastern ones. Coincidentally, the two most heavily admixed individuals were from Montana and Alberta, respectively. Such admixture suggests that there could be some long-distance gene flow between the western and eastern North American populations. *Cronartium ribicola* can travel over distances of up to thousands of kilometers. The rust in the west originated from Vancouver and migrated to southern distant states including California and even further to New Mexico and Arizona (Hawksworth 1990; Fairweather & Geils 2011). The distance between Alberta and Minnesota is ~2000 km, the gene flow between the western and eastern populations is unlikely to have occurred naturally because of the absence of host (Hamelin et al. 2000). Intensive agriculture in the Great Plains has limited the number of natural five-needled pines and *Ribes*, *C. ribicola*'s aecial and telial hosts, respectively. This lack of hosts (both aecial and telial) in the Great Plains is the most likely factor contributing to the east/west split of *C. ribicola* in North America (Hamelin et al. 2000).

However, the increased cultivation of *Ribes* in this area may serve as a bridge for exchange of genetic material between the eastern and western rust populations (Hamelin et al. 2000). According to the ADMIXTURE results mentioned above, this may have already happened. *Ribes* fruits have always been commercially important for producing juice, jams and other products. Many *Ribes* species are native and wildly distributed throughout North America. For example, American black currant (*R. americanum* Mill.), a native species, is grown from Alberta to New Brunswick in Canada (United States Department of Agriculture 2014). Therefore, the whole range of native species in North America is suitable for *Ribes*

cultivation (Barney 1996). Since the introduction of *C. ribicola* to North America, *Ribes* cultivation was stopped in the early 1900s to prevent spread of the pathogen. In the 1990s, *Ribes* cultivation (especially black currants *R. nigrum*) regained its popularity as the legislations to control *C. ribicola* were revoked and commercial five-needled pine plantations declined (Dale 2000). US and Canada both prohibit importations of *Ribes* from other countries, but the North American varieties are allowed to move freely between these two countries (Dale 2000).

In the current cultivation practices, the potential threat posed by *C. ribicola* is being countered by utilizing resistant cultivars, such as 'Consort', 'Crusader', and 'Titania' (Department of Horticulture 2014). However, recently a new race of *C. ribicola* that is virulent on previously immune *Ribes* was reported in eastern North America (Tanguay et al. 2013). First discovered as early as 2008 in Connecticut, this race is present across the entire range of *C. ribicola* in eastern North America (Quebec, Connecticut, New Hampshire, New York, Nova Scotia, and Prince Edward Island). This new race resulted from either a novel mutation or a DNA recombination event instead of a new introduction (Tanguay et al. 2013). It demonstrates the capacity of the rust to overcome host's resistance through time. As the movement of *Ribes* plants across the continent is not restricted, rust can be carried along with the presumably immune plants. This is an important implication to the policy makers in terms of regulating *Ribes* cultivation for *C. ribicola* control.

4.3 Elucidating Cryptic Population Structure of *Cronartium Ribicola* in Western North America

Cryptic population structure could be important in explaining patterns of differentiation that are not apparent using other analyses (Jombart et al. 2008). The development of spatial PCA is a useful method to highlight such cryptic structure. Since western North American populations are subjected to more landscape and climate variation than their eastern counterpart, we hypothesized that cryptic structure will be observed. No obvious population structure was identified by PCA, DAPC or ADMIXTURE. Nevertheless, a cryptic yet significant Coastal/Interior separation was found by sPCA. This global structure describes a positive spatial autocorrelation, meaning that neighbouring populations are genetically more similar than distant ones. This grouping pattern agrees with the geography, grouping all Coastal populations separately from the interior ones. Nucleotide diversity was higher in the Coastal than in the interior groups. Specifically, Mount Washington (MW), a Vancouver Island population, possessed the highest overall nucleotide diversity. This is concordant with the records of Eastham (1923) and Mielke (1943) that the single introduction to western North America from France took place on the BC Coast. One can also see the gradual decrease of diversity as *C. ribicola* spread eastward to the Rockies and southward to interior BC (Figure 15).

However, there were two exceptions to the Coast/Interior pattern. Firstly, the rust population from Prince George, located in interior BC, was more similar to the Coastal group than its close neighbours - Smithers or McBride. As the *C. ribicola* in Prince George was sampled from white pines in plantations, (Figure 21 Appendix), the most likely cause of such pattern is its seedling source. The plantations in Prince George most likely share the same seedling source as the Coastal plantations. Therefore, if the seedlings had been previously infested with *C. ribicola* at an early stage (without showing symptoms) before getting transplanted to Prince George, the spread of the rust would mimic its host. This can explain the seemingly unusual similarity between the Prince George population and the Coastal group. The second exception was that the Smithers population was found to be genetically more similar to the

Interior than to the Coastal group. This result is slightly surprising given the fact that the distribution of *P. albicaulis* is continuous on high elevations along the coastline of southern BC but there is a patch between the BC Coast and the Interior where no white pines (including *P. monticola* and *P. flexilis*) occur naturally (Figure 21 Appendix).

White pine blister rust is well known for its ability to disperse over long distances. For example, it traveled hundreds of kilometers aerially from California to the Sacramento Mountains of New Mexico about 1970 (Frank et al. 2008). This was possible because the areas in between are almost always covered by at least one of the four local white pine species - *P. lambertiana*, *P. flexilis*, *P. strobiformis,* and *P. aristata*. In contrast, the scarcity of white pines in central BC may act like a barrier to gene flow between BC Coast and Interior, which would explain the cryptic division between the Coastal and Interior rust populations that we observed.

Combined with the previously discussed east/west split of the *C. ribicola* populations in North America due to the absence of white pines in the Great Plains, it can be concluded that the continental and regional population structure of *C. ribicola* is strongly influenced by the distribution and availability of its aecial hosts.

Alberta and Montana populations were compared to the rest of the western populations for three reasons: 1) the admixture pattern in some individuals; 2) geographic location on the eastside of the Continental Divide; 3) the high infection and mortality rate in a *C. ribicola* field survey targeting *P. albicaulis* populations along the Rocky Mountains (Carolin et al. 2008). The higher levels of infection on *P. albicaulis* was attributed to both the greater abundance of *Ribes* spp. and also the nearby existence of two other white pine species – *P. monticola* and *P. flexilis*. Therefore, in the current study, an AMOVA was designed specifically to contrast these populations. However, our results do not support the presence of distinct populations across the Continental Divide or on different pine hosts. This confirms that the *C.*

ribicola populations on either side of the Canadian Rocky Mountains, on different pine hosts, belong to the same genetic cluster.

What is unique about the western cluster is the remarkable difference between its nucleotide diversity (Pi) and number of segregating sites (θs, Table 7 & 8). This is an indication that there are rare alleles present at low frequencies, which could have been caused by either population expansion following a recent bottleneck or selective sweep (Larsson et al. 2013). The presence of rare alleles can be verified by the higher number of singletons observed in the western cluster, together with a lower overall heterozygosity (Table 10 & 12 Appendix). The single recorded introduction from France in 1910 would explain the population bottleneck following the founder effect. Since the highest number of singletons was observed in the Smithers population (Table 10 Appendix), which is an edge population for both the rust and white pines, it could imply that the rare alleles are maintained in the population and are under selection pressure for adapting to extreme conditions.

4.4 The Suitability and Robustness of GBS for Studying Fungal Genomes

GBS is a powerful method to generate large SNP datasets on non-model species. After stringent filtering, GBS successfully generated up to 8,020 high quality SNP markers for the global dataset, ~ 4000 for the North American dataset, and hundreds of markers for the regional datasets, covering ~2% of the whole genome. However, it has some shortcomings that should be addressed: 1) dependence on large amount of good quality DNA; 2) choice of restriction enzyme(s) to maximize number of short fragments; 3) sequencing error; 4) limited range of alleles; 5) high linkage disequilibrium; and 6) missing data due to either sampling or biological reasons (Buckler 2011). Some of the above potential problems can be corrected with proper SNP filtering procedure, or improved coverage with paired-end sequencing.

The biggest challenge faced when generating GBS data for a rust fungus is that the quantity of starting materials is limited, due to the biotrophic nature of rust fungi. Our pilot study proved that we could lower the amount of DNA without affecting the outcome. One major weakness of GBS is its

tendency to generate missing data, often due to preferential amplification of smaller fragments (Davey et al. 2011). With this in mind, a two-enzyme protocol (*PstI* (CTGCAG) and *MspI* (CCGG)) was chosen in order to control the number and size of the fragments (Poland et al. 2012). Because *C. ribicola* has a relatively small genome size (~72 Mbp) compared to crop species, two-enzyme digestion increased the number of short fragments thus increasing the read depth. *Cronartium ribicola* was shown to be highly outcrossed (Hamelin et al. 1998), which means heterozygous genotypes were expected. Higher read depth allowed us to score such genotypes with more confidence. Nevertheless, raw SNPs were filtered for minimum and maximum read depth to eliminate sequencing errors. We tolerated up to 20% missing data in the SNP datasets in order to maximize the number of markers for analysis. In majority of the subsequent analyses (PCA, DAPC, sPCA and Neighbouring-joining tree), missing data were replaced with the mean observed genotype.

For functional studies and genome-wide association studies (GWAS) several algorithms including, FastPhase, NPUTE, and BEAGLE, might be used to impute missing data (Wang et al. 2012). Alternatively, the quality of the data can be improved by generating paired-end reads.

Our results show that the SNP datasets we generated with GBS are reliable. All SNPs in our dataset were in Hardy-Weinberg equilibrium and the technical repeats included in this study always clustered next to each other, which demonstrates the repeatability and consistency of these markers. GBS proved to be a reliable and inexpensive way to generate thousands or markers to answer biological and epidemiological questions about white pine blister rust.

5 Conclusions

This study has demonstrated the power of genomic tools in understanding the evolutionary history of forest pathogens and highlighted the advantages of reduced-representation library sequencing methods, in this case, genotyping-by-sequencing. By studying the genome of this invasive, highly destructive pathogen, important implications about the management strategies have been revealed. First and foremost, it was confirmed for the first time that, *C. ribicola*'s likely centre of origin is in the central east part of Russia, near the Central Siberian Plateau. From there, it spread westward to Europe and southward to China, Korea and Japan. Secondly, there was evidence for cryptic gene flow between the eastern and western populations in North America, possibly assisted by the resumed plantation of *Ribes* spp. in the Great Plains area. This poses a threat to the remaining white pine forests in both eastern and western North America. So far, *C. ribicola* in the two regions forms two genetically distinct clusters. If more long-distance migration persists, exchange of genetic material between the eastern and western populations would introduce new alleles, thereby changing the current state of the disease. Furthermore, the discovery of new virulent races of *C. ribicola* in eastern North America is another reason to be concerned. Therefore, if *Ribes* plantation is to be continually permitted, vigilant disease monitoring needs to be in place to maintain the existing barrier to gene flow.

The existence of cryptic Coast/Interior pattern in western North America is a novel finding uncovered by GBS. The initial hypothesis was that because of the more diverse landscape in western Canada, population structure of *C. ribicola* would be somewhat correlated with geography, host species or climate. In the current study, the only noticeable pattern was the Coast/Interior split, most likely due to disrupted host connectivity in south central BC. Accordingly, the null hypothesis cannot be rejected.

Cryptic population structure in western North America, long-distance gene flow between the east/west North America and existence of interspecific hybrids at high altitude/latitude indicate that *C.*

ribicola is a dynamic pathogen that could be slowly adapting to extreme environments. These results provide useful and important information for both the forest pest management and the white pine resistance breeding programs.

Literature Cited

Abrams, M.D., 2001. Eastern white pine versatility in the presettlement forest. *Bioscience*, 51, pp.967–979.

Alexander, D.H., Novembre, J. & Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9), pp.1655–1664.

Altshuler, D. et al., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803), pp.513–516.

Andrews, S., 2014. FastQC A Quality Control tool for High Throughput Sequence Data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Van Arsdel, E.P., 1972. Environment in relation to white pine blister rust infection. *USDA For. Serv. Tech. Rep. Misc. Publ.*, 1221, pp.479–494.

Van Arsdel, E.P. & Geils, B.W., 2011. Blister Rust in North America : What We Have Not Learned in the Past 100 Years. In *Proceedings of the 58th Annual Western International Forest Disease Work Conference*. Valemount, BC, pp. 61–69.

Van Arsdel, E.P. & Geils, B.W., 2004. *The Ribes of Colorado and New Mexico and Their Rust Fungi*, Rep. FHTET 04-13. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Forest Health Technology Enterprise Team. p. 32. [Online].

Van Arsdel, E.P., Riker, A.J. & Patton, R.F., 1956. The effects of temperature and moisture on the spread of white pine blister rust. *Phytopathology*, 46, pp.307–318.

Baird, N.A. et al., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, 3(10), p.e3376.

Barney, D.L., 1996. *Ribes* production in North America: Past, present, and future. *HortScience*, 31(5), p.774.

Benedict, W.V., 1981. *History of white pine blister rust control - a personal account*, Washington, D. C.: U.S. Dept. of Agriculture, Forest Service.

Blodgett, J.T. & Sullivan, K.F., 2004. First Report of White Pine Blister Rust on Rocky Mountain Bristlecone Pine. *Plant Disease*, 88(3), p.311.

Brar, S., 2012. *Landscape genetics of Cronartium ribicola*. M. Sc. thesis, The University of British Columbia.

Buchanan, T.S. & Kimmey, J.W., 1938. Initial tests of the distance of spread to and intensity of infection on Pinus Monticola by *Cronartium ribicola* from *Ribes lacustre* and *R. viscosissimum*. *Journal of Agricultural Research*, 56, pp.9–30.

Buckler, E.S., 2011. Why can GBS be complicated? Tools for filtering, error correction and imputation. , pp.19–37.

CABI, 2014. Invasive Species Compendium. *CAB International*. Available at: http://www.cabi.org/isc/datasheet/16154 [Accessed May 1, 2014].

Campbell, E.M. & Antos, J.A., 2000. Distribution and severity of white pine blister rust and mountain pine beetle on whitebark pine in British Columbia. *Canadian Journal of Forest Research*, 30(7), pp.1051–1059.

Carolin, T. et al., 2008. Whitebark pine and white pine blister rust in the Rocky Mountains of Canada and northern Montana. *Canadian Journal of Forest Research*, 38(5), p.982.

Dale, A., 2000. Potential for Ribes Cultivation in North America. *HortTechnology*, 10(3), pp.548–554.

Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15), pp.2156–2158.

Dangl, J.L. & Jones, J.D.G., 2001. Plant pathogens and integrated defence response to infection. *Nature*, 411(14 June), pp.826–833.

Davey, J.W. et al., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, 12(7), pp.499–510.

Davey, J.W. & Blaxter, M.L., 2010. RADSeq: next-generation population genetics. *Briefings in functional genomics*, 9(5-6), pp.416–423.

Davidson, A.T., 1922. Western white pine blister in British Columbia (Report on Canadian conditions and work). In *Proceedings and Recommendations of the 3rd western white pine blister rust conference*. Portland, Oregon.

Department of Horticulture, 2014. Minor Fruits Gooseberries and Currants Ribes. spp. *Cornell University*. Available at: http://www.fruit.cornell.edu/mfruit/gooseberries.html.

Eastham, J.W., 1923. White-pine blister-rust in B. C. *The Agricuture Journal (BC, Canada)*, 7(29), p.41.

Ekramoddoullah, A.K.M., 2005. Molecular tools in the study of the white pine blister rust [*Cronartium ribicola*] pathosystem. *Canadian Journal of Plant Pathology*, 27, pp.510–520.

Elshire, R.J. et al., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5), p.e19379.

Emerson, K.J. et al., 2010. Resolving postglacial phylogeography using high-throughput sequencing. *PNAS*, 107(37), pp.16196–16200.

Et-touil, K. et al., 1999. Genetic Structure of *Cronartium ribicola* Populations in Eastern Canada. *Phytopathology*, 89(10), pp.915–919.

Excoffier, L. & Lischer, H.E.L., 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10, pp.564–567.

Fairweather, M.L. & Geils, B.W., 2011. First Report of the White Pine Blister Rust Pathogen, *Cronartium ribicola*, in Arizona. *Plant Disease*, 95(4), p.494.

Frank, K.L. et al., 2008. Synoptic climatology of the long-distance dispersal of white pine blister rust II . Combination of surface and upper-level conditions. *International journal of biometeorology*, 52, pp.653–666.

Ganley, R.J., Sniezko, R.A. & Newcombe, G., 2008. Endophyte-mediated resistance against white pine blister rust in *Pinus monticola*. *Forest Ecology and Management*, 255(7), pp.2751–2760.

García-Alcalde, F. et al., 2012. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics (Oxford, England)*, 28(20), pp.2678–2679.

Gitzendanner, M.A. et al., 1996. Genetics of *Cronartium ribicoIa*. III. Mating system. *Canadian Journal of Botany*, 74(1852-1859).

Green, V.E.J. & Van Arsdel, E.P., 1956. *State Project 680*, Florida. In Florida Agr. Exp. Sta. Annu. Rep. 1956, pp. 224-225.

Hamelin, R.C. et al., 2000. Barrier to Gene Flow Between Eastern and Western Populations of *Cronartium ribicola* in North America. *Phytopathology*, 90(10), pp.1073–1078.

Hamelin, R.C. et al., 2005. Molecular epidemiology of white pine blister rust: recombination and spatial distribution. *Phytopathology*, 95(7), pp.793–799.

Hamelin, R.C., Beaulieu, J. & Plourde, A., 1995. Genetic diversity in populations of *Cronartium ribicola* in plantations and natural stands of Pinus strobus. *Theoretical and Applied Genetics*, 91, pp.1214–1221.

Hamelin, R.C., Dusabenyagasani, M. & Et-Touil, K., 1998. Fine-level genetic structure of white pine blister rust populations. *Phytopathology*, 88(11), pp.1187–1191.

HannonLab, 2014. FASTX Toolkit. *http://hannonlab.cshl.edu/fastx_toolkit/index.html*. Available at: http://hannonlab.cshl.edu/fastx_toolkit/.

Hartl, D.L. & Clark, A.G., 1997. *Principles of population genetics* Third Edi., Sunderland, Massachusetts: Sinauer Associates, Inc. Publishers. p. 542

Hawksworth, F.G., 1990. White pine blister rust in southern New Mexico. *Plant Disease*, 74(11).

Hirt, R.R., 1942. *The relation of certain meteorological factors to the infection of eastern white pine by the blister - rust fungus*, State Univ. ColI. For., Syracuse Univ. Tech. Publ. pp. 59, 65

Hohenlohe, P.A. et al., 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular ecology resources*, 11 Suppl 1, pp.117–22.

Hohenlohe, P.A. et al., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics*, 6(2), p.e1000862.

Huang, B.E. et al., 2014. Efficient imputation of missing markers in low-coverage genotyping-by-sequencing data from multiparental crosses. *Genetics*, 197(1), pp.401–404.

Hummer, K.E., 2000. History of the Origin and Dispersal of White Pine Blister Rust. *Mycologia*, 10(September), pp.515–517.

Hummer, K.E. & Dale, A., 2010. Horticulture of *Ribes*. *Forest Pathology*, 40(3-4), pp.251–263. Available at: http://doi.wiley.com/10.1111/j.1439-0329.2010.00657.x [Accessed August 24, 2014].

Hunt, R.S., 2004. *Blister-Rust-Resistant Western White Pines for British Columbia*, Info. Rep. BC-X-397. Victoria, BC: Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre. p. 18.

Hunt, R.S., 1983. White Pine Blister Rust in British Columbia. II. Can stands be hazard rated? For. Chron. 59, pp. 30–33.

Jombart, T. et al., 2014. adegenet: an R package for the exploratory analysis of genetic and genomic data. Available at: http://cran.r-project.org/web/packages/adegenet/adegen

Jombart, T. et al., 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, 101(1), pp.92–103.

Jurgens, J.A. et al., 2003. Histology of White Pine Blister Rust in Needles of Resistant and Susceptible Eastern White Pine. *Plant Disease*, 87(9), pp.1026–1030.

Keane, R., Morgan, P. & Menakis, J., 1994. Landscape assessment of the decline of whitebark pine (Pinus albicaulis) in the Bob Marshall Wilderness Complex. *Northwest Sci.*, 68, pp.213–229.

Kearns, H.S.J. et al., 2008. Distribution of Ribes , an alternate host of white pine blister rust , in Colorado and Wyoming. *Journal of Torrey Botanical Society*, 135(3), pp.423–437.

Kearns, H.S.J. & Jacobi, W.R., 2007. The distribution and incidence of white pine blister rust in central and southeastern Wyoming and northern Colorado. *Canadian Journal of Forest Research*, 37, pp.462–472.

Kendall, K.C., 1994. Whitebark pine conservation in North American National Parks. In W. C. S. and F. K. Holtmeier, ed. *International Workshop on Subalpine Stone Pines and Their Environment: the Status of Our Knowledge*. St. Moritz, Switzerland: USDA For. Serv. Gen. Tech. Rep. INT-GTR-309, pp. 302–307.

Kim, M.-S. et al., 2010. White pine blister rust in Korea, Japan and other Asian regions: comparisons and implications for North America. *Forest Pathology*, 40(3-4), pp.382–401.

King, R.C., Mulligan, P.K. & Stansfield, W.D., 2013. Single nucleotide polymorphism (SNP). In *A Dictionary of Genetics*. "Oxford University Press."

Kinloch, B.B.J., 2003. White Pine Blister Rust in North America: Past and Prognosis. *Phytopathology*, 93, pp.1044–1047.

Kinloch, B.B.J. & Comstock, M., 1980. Cotyledon test for major gene resistance to white pine blister rust in sugar pine. *Canadian Journal of Botany*, 58(17), pp.1912–1914.

Kinloch, B.B.J. & Littlefield, J.L., 1977. White pine blister rust: hypersensitive resistance in sugar pine. *Canadian Journal of Botany*, 55, pp.1148–1155.

Kinloch, B.B.J., Parks, G.K. & Fowler, C.W., 1970. White pine blister rust: simply inherited resistance in sugar pine. *Science*, 167, pp.193–195.

Kinloch, B.B.J., Sniezko, R.A. & Dupper, G.E., 2003. Origin and distribution of cr2, a gene for resistance to white pine blister rust in natural populations of Western white pine. *Phytopathology*, 93(6), pp.691–694.

Kinloch, B.B.J., 1992. Distribution and frequency of a gene for resistance to white pine blister rust in natural populations of sugar pine. *Canadian Journal of Botany*, 70, pp.1319–1323.

Kinloch, B.B.J. & Comstock, M., 1981. Race of Cronartium ribicola virulent to major gene resistance in sugar pine. Pl. Dis. Rep. 65, pp. 604-605.

Kinloch, Jr., B.B., Sniezko, R.A. & Dupper, G.E., 2004. Virulence gene distribution and dynamics of the white pine blister rust pathogen in Western north america. *Phytopathology*, 94(7), pp.751–758.

Krebill, R.G., 1971. Effect of low temperature on germination of teliospores of *Cronartium-ribicola*. *Phytopathology*, 61(8), p.889.

Larsson, H. et al., 2013. Distribution of long-range linkage disequilibrium and Tajima's D values in Scandinavian populations of Norway Spruce (Picea abies). *G3 (Bethesda, Md.)*, 3(5), pp.795–806.

Leeuw, J. De, 1994. Block-relaxation Algorithms in Statistics. In H.-H. et al. Bock, ed. *Information Systems and Data Analysis*. Heidelberg: Springer-Verlag Berlin, pp. 308–324.

Leppik, E.E., 1967. Phylogeny of rust fungi. *Mycologia*, 59, pp.568–579.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Loecher, M., 2014. RgoogleMaps: Overlays on Google map tiles in R. Available at: http://cran.r-project.org/web/packages/RgoogleMaps/RgoogleMaps.pdf.

Lombard, K. & Bofinger, J., 1999. White Pine Blister Rust (*Cronartium Ribicola*): Infection Incidence for Selected Areas of New Hampshire. UNH Cooperative Extension, 1999.

McDonald, G.I. et al., 2006. *Pedicularis* and *Castilleja* are natural hosts of *Cronartium ribicola* in North America: a first report. *Forest Pathology*, 36(2), pp.73–82.

McDonald, G.I., Hoff, R.J. & Wykoff, W.R., 1981. Computer simulation of white pine blister rust epidemics. I. Model formulation. *U.S. Department of Agriculture, Forest Service, Intermountain Research Station Research Paper*, INT-258, p.77.

McKay, S., 2000. State regulation of *Ribes* to control white pine blister rust. *HortTechnology*, 10(3), pp.562–564.

Mielke, J.L., 1943. White pine blister rust in western North America. *Yale Shcool of Forestry Bulletin*, 52, pp.118–155.

Moir, W.S., 1924. White-pine blister rust in western Europe. *USDA Bulletin*, 1186.

Moran, P.A.P., 1950. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2), pp.17–23.

Moran, P.A.P., 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2), pp.243–251.

Neale, D.B. & Savolainen, O., 2004. Association genetics of complex traits in conifers. *Trends in plant science*, 9(7), pp.325–330.

Pacific Southwest Research Station, 2011. Life Cycle of White Pine Blister Rust. United States Forest Service. Available at: http://www.fs.fed.us/psw/topics/forest_genetics/wpbr/life.shtml [Accessed September 1, 2013].

Paradis, E., Claude, J. & Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, pp.20:289–290.

Pennington, L.H., 1925. Relation of weather conditions to the spread of white pine blister rust in the Pacific Northwest. *Journal of Agricultural Research*, 30, pp.593–607.

Poland, J.A. et al., 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS one*, 7(2), p.e32253.

Posey, G.B. & Ford, E.R., 1924. Survey of blister rust infection on pines at Kittery Point, Maine and the effects of *Ribes* eradication in controlling the disease. *Journal of Agricultural Research*, 28(12), pp.1253–1358.

Pritchard, J.K., Stephens, M. & Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155, pp.945–959.

Purcell, S., 2009. PLINK. Available at: http://pngu.mgh.harvard.edu/purcell/plink/.

Purcell, S. et al., 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. Available at: http://www.r-project.org.

Richardson, B.A. et al., 2010. Current and future molecular approaches to investigate the white pine blister rust pathosystem. *Forest Pathology*, 40(3-4), pp.314–331.

Richardson, B.A. et al., 2008. Influence of host resistance on the genetic structure of the white pine blister rust fungus in the western United States. *Phytopathology*, 98(4), pp.413–420.

Richardson, B.A. et al., 2008. Tracking the Footsteps of an Invasive Plant-Pathogen: Intercontinental Phylogeographic Structure of the White-pine-blister-rust Fungus, *Cronartium ribicola*. In *Breeding and Genetic Resources of Five-Needle Pines: Ecophysiology, Disease Resistance and Developmental Biology*. pp. 56–60.

Rutkoski, J.E. et al., 2013. Imputation of unordered markers and the impact on genomic selection accuracy. *G3 (Bethesda, Md.)*, 3(3), pp.427–439.

Saitou, N. & Nei, M., 1987. The Neighbour-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology Evolution*, 4(4), pp.406–425.

Schmieder, R. & Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, 27(6), pp.863–864.

Smith, C.M. et al., 2008. Whitebark pine and white pine blister rust in the Rocky Mountains of Canada and northern Montana. *Canadian Journal of Forest Research*, 38, pp.982–995.

Sonah, H. et al., 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PloS one*, 8(1), p.e54603.

Spaulding, P., 1922. Investigations of the white-pine blister rust. *Bulletin of the U.S. Department of Agriculture*, 957, p.100.

Spaulding, P., 1911. The blister rust of white pine. *Bureau of plant industry - bulletin*, 206.

Spaulding, P., 1929. White-pine blister rust: A comparison of European with North American conditions. *USDA Tech. Bull.*, 87.

Stewart, F.C., 1906. An outbreak of the European currant rust (*Cronartium ribicola* Dietr.). *Bulletin of the U.S. Department of Agriculture*, 2, pp.61–74.

Sweeney, K. et al., 2011. Are Needle Reactions in Resistance to *Cronartium ribicola* a Hypersensitivity Response? In *Proceedings of the 4th International Workshop on Genetics of Host-Parasite Interactions in Forestry*. pp. 368–371.

Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 104, pp.437–460.

Tanguay, P. et al., 2013. The origin of a new race of *Cronartium ribicola*, virulent on previously immune blackcurrant cultivars, and rapidly spreading in eastern North America. In *APS MSA Joint Meeting*. Austin, Texas: The American Phytopathological Society, p. 441.

Tubeuf, C.F. von, 1917. Uber das Verhaltnis der Kiefern- Peridermien zu Cronartium. II. Studein uber die lnfekion der Weymouthskiefer. *Naturwissenschaftliche Zeitschrift fOr Forst- und Landwirtschaft*, 15(7), pp.274–307.

United States Department of Agriculture, 2014. *Ribes americanum* Mill. American black currant. *Natural Resources Conservation Service*. Available at: http://plants.usda.gov/core/profile?symbol=RIAM2 [Accessed September 15, 2014]

Urbanek, S., 2013. png: Read and write PNG images in R. Available at: http://www.rforge.net/png/

Vavilov, N., 1926. *Studies on the origin of cultivated plants*, Bull. Appl. Bot. Plant Breed., Leningrad, USSR 16, pp. 139–248.

Vogler, D.R., Delfino-mix, A. & Schoettle, A.W., 2005. White Pine Blister Rust in High-Elevation White Pines:Screening for Simply-Inherited , Hypersensitive Resistance. In J. C. (compiler) Guyon, ed. *Proceedings of the 53rd Western International Forest Disease Work Conference*. Jackson, WY: USDA Forest Service, Intermountain Region, Odgen UT, pp. 73–82.

Wang, Y. et al., 2012. Fast accurate missing SNP genotype local imputation. *BMC research notes*, 5, p.404.

Ward, J.A. et al., 2013. Saturated linkage map construction in Rubus idaeus using genotyping by sequencing and genome-independent imputation. *BMC genomics*, 14, p.2.

Weir, B.S., 1996. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. In Sunderland, MA, USA: Sinauer Assoc., Inc.

Weir, B.S. & Cockerham, C.C., 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6), pp.1358–1370.

White, M.A., Brown, T.N. & Host, G.E., 2002. Landscape analysis of risk factors for white pine blister rust in the Mixed Forest Province of Minnesota , U . S . A . *Canadian Journal of Forest Research*, 1650(319), pp.1639–1650.

Worrall, J., 2009. White Pine Blister Rust. *forestpathology.org*. Available at: http://www.forestpathology.org/dis_wpbr.html [Accessed May 29, 2014].

Wright, S., 1931. Evolution in Mendelian populations. *Genetics*, 16, pp.97–159.

Wright, S., 1951. The genetical structure of populations. *Ann. Eugen.*, 15, pp.323–354.

Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, 19, pp.395–420.

Zambino, P.J., Richardson, B.A. & McDonald, G.I., 2007. First Report of the White Pine Blister Rust Fungus, Cronartium ribicola, on Pedicularis bracteosa. *Plant Disease*, 91(April), p.467.

Zeglen, S., 2002. Whitebark pine and white pine blister rust in British Columbia, Canada. *Canadian Journal of Forest Research*, 32(7), pp.1265–1274.

Zeglen, S., Hunt, R. & Cleary, M., 2009. British Columbia's forests: White Pine Blister Rust Forest Health Stand Establishment Decision Aid. *BC Journal of Ecosystem and Management*, 10(1), pp.97–100.

Zhou, H., Alexander, D. & Lange, K., 2011. A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistics and computing*, 21(2), pp.261–273.

Zolan, M.E. & Pukkila, P.J., 1986. Inheritance of DNA methylation in Coprinus cinereus. *Molecular and cellular biology*, 6(1), pp.195–200.

Appendices

Table 9 Number of singletons/doubletons in the global *Cronartium ribicola* dataset.

REGIONS	POP	# OF SINGLETONS/DOUBLETONS	# OF INDS	AVERAGE
ASIA	RU	468	2	234
	KO	516	6	86
NA_WEST	AB	324	8	41
	USNM	33	1	33
	SC	130	4	33
	PG	110	4	28
	USMR	131	5	26
	KT	129	5	26
	PR	267	11	24
	VA	229	10	23
	SM	379	21	18
	MW	209	12	17
	TE	150	9	17
	MB	93	10	9
	PE	79	9	9
	CP	63	9	7
EUROPE	FI	16	1	16
NA_EAST	NY	6	1	6
	NF	5	1	5
	MA	3	1	3
	MN	5	2	3
	QC	68	28	2
	NH	2	1	2
	WI	2	1	2
	VT	1	1	1

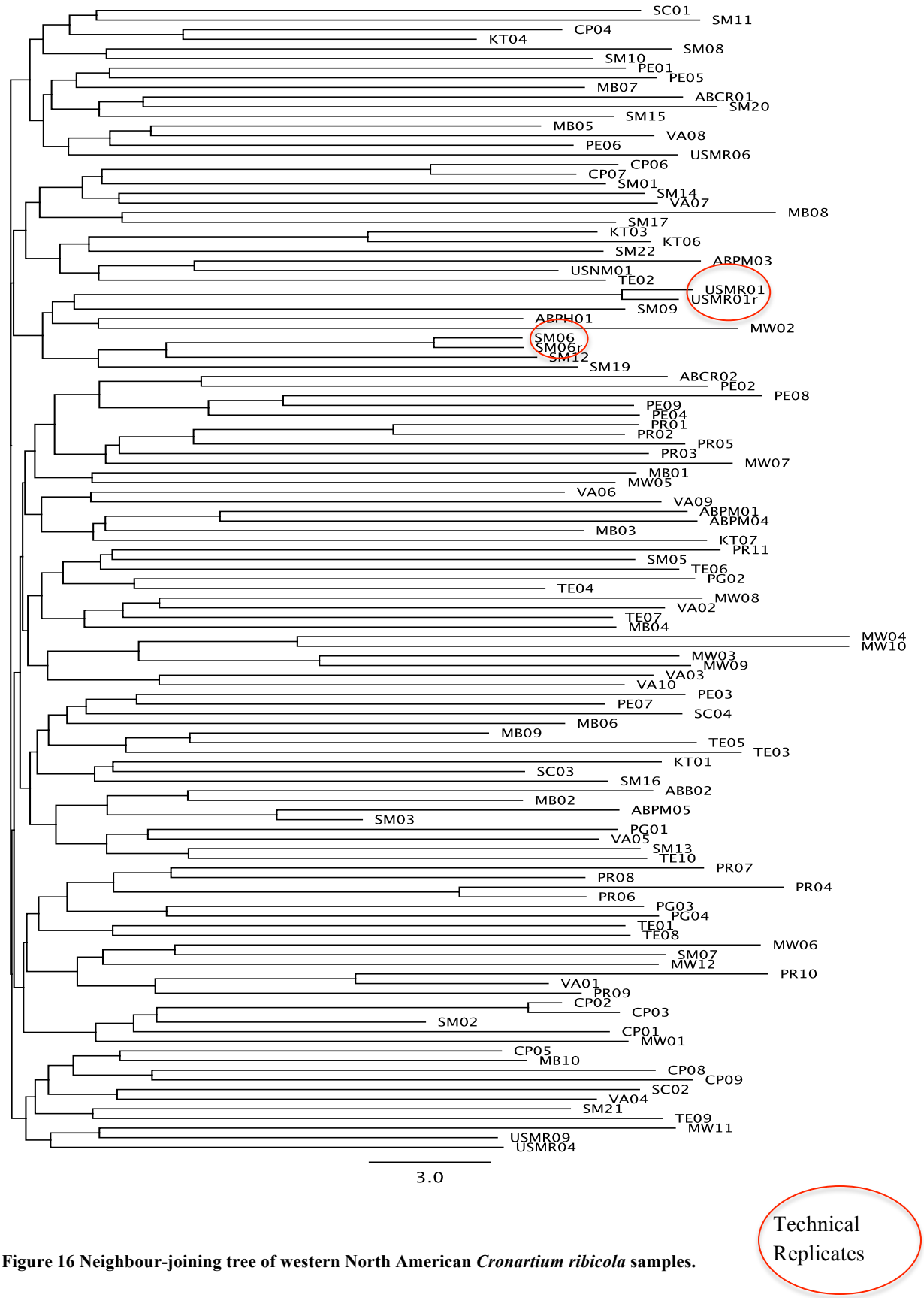

Figure 16 Neighbour-joining tree of western North American *Cronartium ribicola* samples.

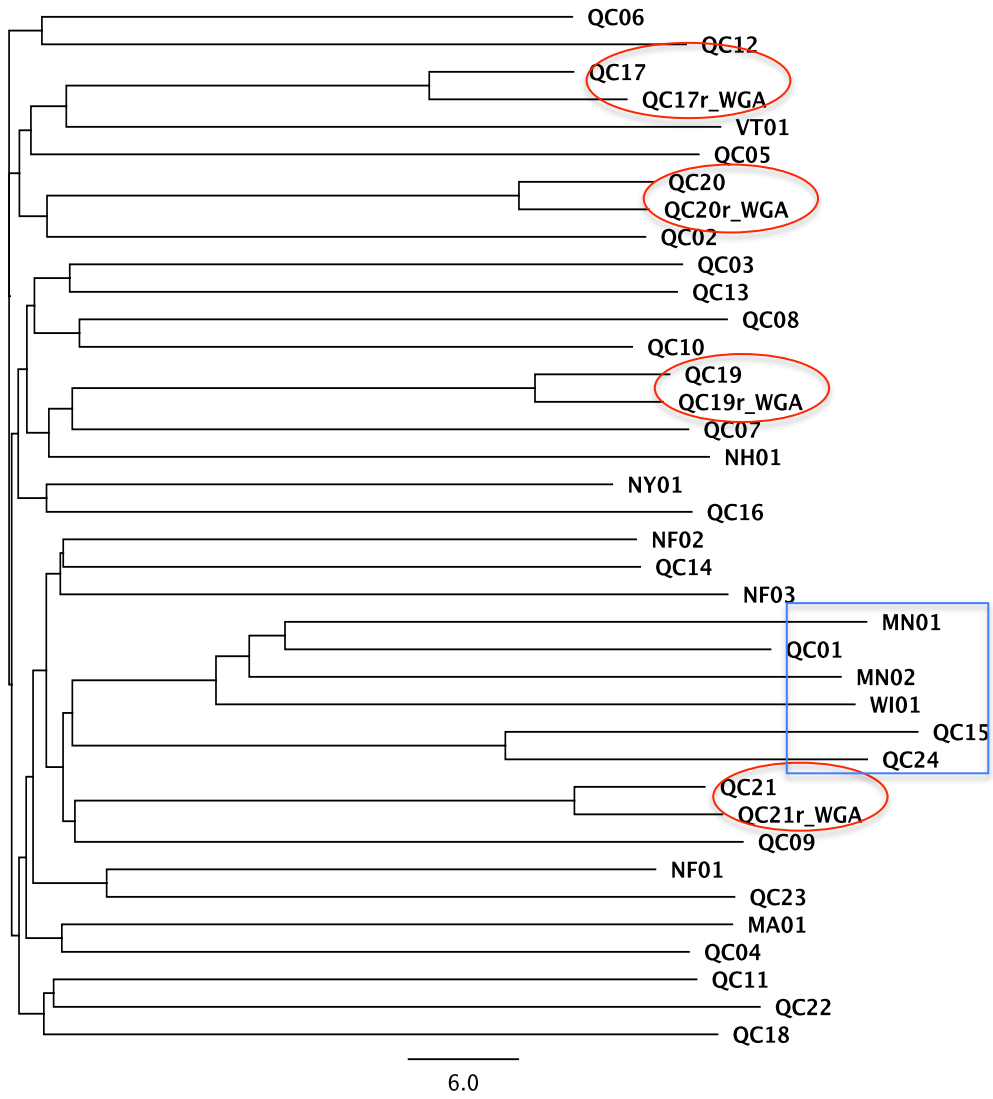

Figure 17 Neighbour-joining tree of eastern North American *Cronartium ribicola* samples.

Outliers within NA_East

Technical Replicates

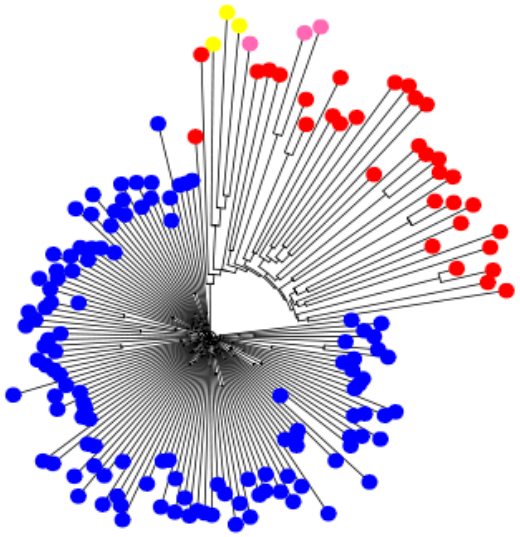

Figure 18 Neighbour-joining tree of North American *Cronartium ribicola* samples.

Blue = NA_West, red = NA_East, yellow = US_Midwest, pink = outliers from QC in the PCA

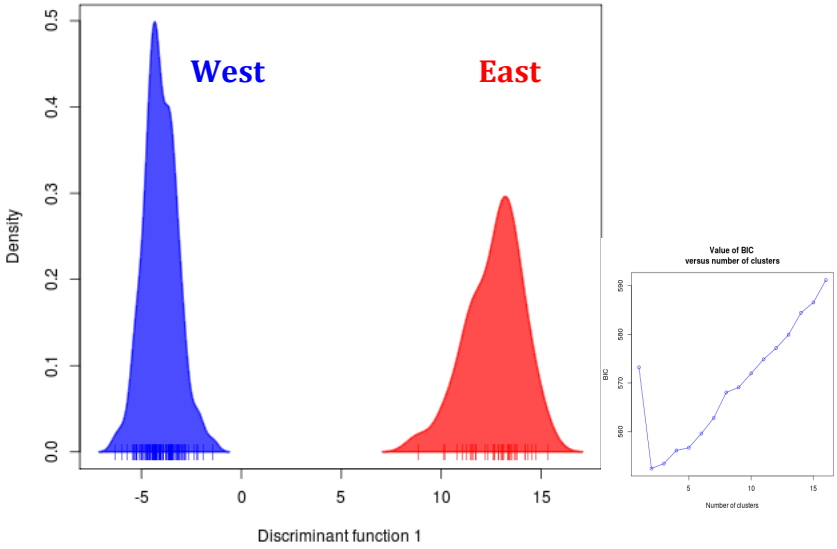

Figure 19 Discriminant analysis of principal components of the North American *Cronartium ribicola* samples.

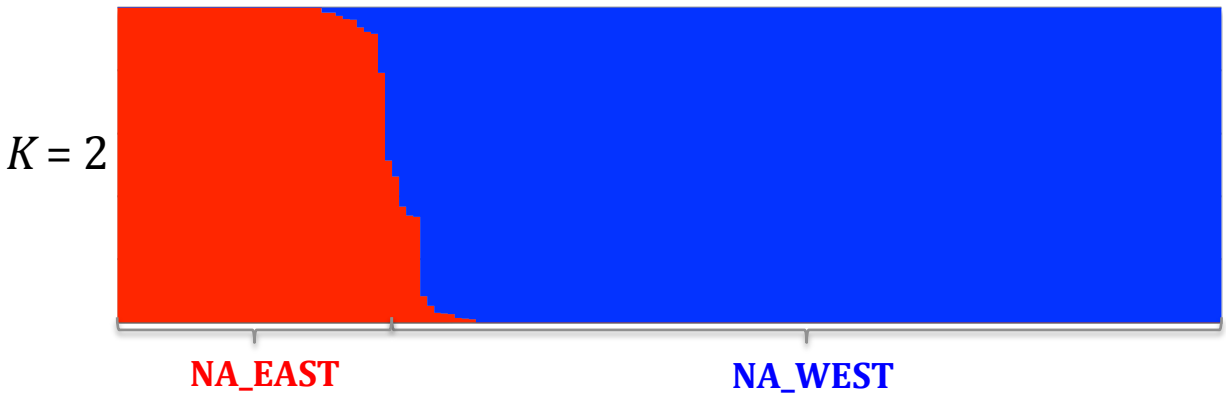

Figure 20 ADMIXUTRE result of North American *Cronartium ribicola* samples.

Table 10 Number of singletons in North American *Cronartium ribicola* populations.

REGIONS	POPULATION	N	SINGLETONS/DOUBLETONS
EAST	QC	28	83
	East_Coast	4	14
	US_Midwest	3	9
	NF	3	4
	Total		**110**
WEST	SM	21	390
	AB	8	327
	PR	11	274
	MW	12	232
	VA	10	227
	TE	10	156
	USMR	5	139
	SC	4	129
	KT	5	127
	PG	4	113
	PE	9	106
	MB	20	95
	CP	9	66
	Total		**850***

*Based on the average of 20 random selections of 38 western samples

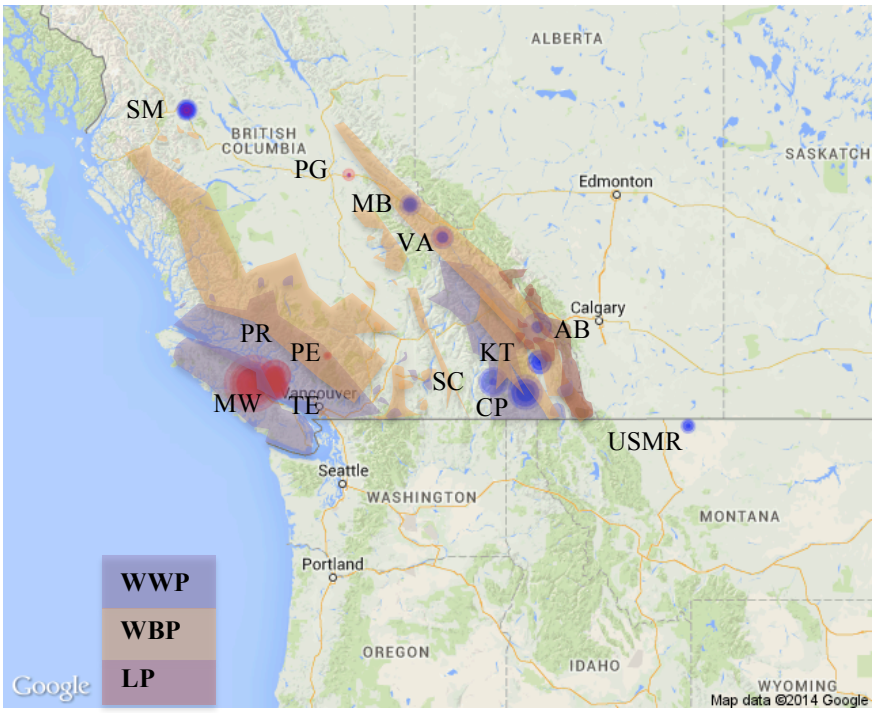

Figure 21 Western North American spatial Principal Component Analysis of *Cronartium ribicola* – global structure, laid over the distribution of white pines in western Canada.

Table 11 Genetic diversity indices of the western North American *Cronartium ribicola* dataset, with 4,510 SNPs.

REGION	POPS	HOST	STAND TYPE	N	% POLY. SITES	H_e	Pi	θs	F_{IS}
NA_WEST COAST	MW	WWP	Natural	12	14.24%	0.036	0.029	0.038	-0.09
	PE	WWP	Plantation	9	10.18%	0.032	0.025	0.030	-0.15
	PG	WWP	Plantation	4	8.20%	0.033	0.027	0.032	-0.10
	PR	WWP	Plantation	11	14.10%	0.035	0.022	0.039	-0.05
	TE	WWP	Plantation	10	11.42%	0.033	0.025	0.032	-0.10
	TOTAL			**46**					
NA_WEST INTERIOR	AB	WBP/LP	Natural	8	14.72%	0.040	0.021	0.044	0.02
	CP	WBP	Natural	9	8.03%	0.029	0.022	0.023	-0.17
	KT	WWP	Natural	5	8.89%	0.034	0.023	0.031	-0.06
	MB	WBP	Natural	10	9.56%	0.031	0.020	0.027	-0.08
	SC	WWP	Natural	4	8.76%	0.034	0.030	0.034	-0.14
	USMR	WBP/LP	Natural	5	10.00%	0.037	0.030	0.035	-0.24
	SM	WBP/WWP/ EWP	Natural	21	17.41%	0.032	0.023	0.040	-0.05
	VA	WWP	Natural	10	12.57%	0.034	0.023	0.035	-0.08
	TOTAL			**72**					
NA_EAST	East_Coast	EWP	Natural	4	11.91%	0.053	0.049	0.046	-0.20
	US_Midwest	EWP	Plantation	3	8.91%	0.043	0.041	0.039	-0.17
	Newfoundland	EWP	Natural	3	10.00%	0.053	0.036	0.044	-0.30
	QC	EWP	Natural/ Plantation	28	18.16%	0.052	0.045	0.040	-0.11
	TOTAL			**38**					

Pops = population names, WWP = western white pine, WBP = whitebark pine, EWP = eastern white pine, LP = limber pine, N = sample size, % poly. sites = proportion of sites that are polymorphic, H_E = expected heterozygosity over all loci, Pi = nucleotide diversity, θs = number of segregation sites averaged over all loci, F_{IS} = population specific inbreeding coefficient.