

Background and recent NSERC research relevant to this proposal

My research program is concerned with the population and quantitative genetics of plants, with an emphasis on the development and application of new statistical methods. It has used monkeyflower (*Mimulus*) as a model organism since 1982. Beginning in 2001, my work expanded broadly through genome projects involving spruce, poplar and *Arabidopsis*. I have gained experience with SNP discovery and assay, association genetics, metabolite and expression QTL (eQTL) mapping, evolutionary sequence comparisons, and structural analyses of genomes. Although to date this genomics research has been separate from my NSERC research, the time is ripe to bring these approaches directly into my NSERC program.

Four examples from my recent NSERC-funded work are particularly relevant to this proposal. (1) In the work of my student, shared with Loren Rieseberg, we identified *in-silico* 45 genes in the terpenoid biosynthetic pathway and inferred patterns of selective constraints in lineages involving *Arabidopsis*, *Populus*, *Ricinus* and *Vitis*; upstream genes had greater constraint; also a novel “pathway pleiotropy index” better explained patterns (**Ramsay** et al. submitted). (2) As an example of incorporating phylogeny into the evolution of quantitative traits, **Chen** and Ritland (submitted) was the first to incorporate phylogeny into quantitative trait locus (QTL) mapping; QTL changes along lineages leading to the evolution of two inbreeding taxa were inferred, and as expected, large effect QTL tended to lie in the ancestor lineage, and smaller effect QTL lay in derived lineages. (3) As an example of new approaches to gene expression statistics, **Albouyeh** and Ritland (2008) presented and evaluated an experimental design to estimate heritability of gene expression using parent-offspring regression with two-channel microarrays. (4) In an application of three gene identity coefficients (Part 2 of this proposal), **Thompson** et al (2008) described estimators for three-gene identity coefficients, and used these to demonstrate that clonal reproduction promotes inbreeding and spatial relatedness in yellow-cedar, *Callitropsis nootkatensis*.

Proposed research

Genomics is allowing new insights into the nature of evolution and adaptation. No longer are we confined to study quantitative traits of unknown genetic composition, nor use anonymous genetic markers for indirect inferences about adaptation and evolution. Now, with adequate genomic resources, we can survey entire gene pathways and the interactions between members of these pathways, and their changes through evolutionary time. This was not feasible even five years ago, but now, knowledge of whole genome sequences has greatly speeded up the discovery of gene and biosynthetic pathways, and inferences about their regulation. This allows us to go beyond the classic paradigm that evolution is due to structural differences (amino acid changes), to examine the role that gene regulation (changes of gene expression) plays in evolution and adaptation. In addition, the advent of high-throughput genotyping makes more elaborate population genetic inferences possible, via improved information about relatedness and patterns of diversity along the chromosome landscape.

This research will build upon my existing strengths in statistical and population genetics, by interfacing my strengths with emerging genome data. The novelty of this proposal resides in the integration of new discoveries in genomics with plant population

and evolutionary genetics. It facilitates the transfer of molecular methods to ecology and evolution, and will train HQP in the interface between evolution and bioinformatics.

Study system: The genus *Mimulus* (Scrophulariaceae) consists of about 160 species in 10-12 taxonomic sections occurring in western North and South America. Since the seminal work of Clausen, Keck and Heisey (1940), *Mimulus* has been a premiere plant model for evolution and ecology. Rapid changes in life history (Vickery 1978), edaphic traits (Macnair 1997), and pollinator attraction (Schemske and Bradshaw 1999) make *Mimulus* a tractable system to study evolution-in-action. In section *Simiolus*, the *M. guttatus* species complex has 8-12 intercrossable species centered in California (Vickery 1978). Inbreeding and a suite of characters associated with inbreeding has evolved at least twice in this group (Ritland and Ritland 1989).

Directed by the consensus of the plant evolution and ecology community (Wu et al. 2008), the Joint Genomics Institute has sequenced the *Mimulus* genome to 7X coverage; the most recent assembly (June 27, 2008) now spans 322 MB. Additional genomic resources include ESTs (expressed sequence tags) for *M. guttatus* (number=300K), *M. nasutus* (33K), and *M. lewisii* (23k). Hence, this genus poses an enormous opportunity to utilize naturally occurring variation at the genome scale to understand the processes of evolution and adaptation. Here, we focus on three taxa in section *Simiolus*, taxa that differ for mating system (inbreeder *M. micranthus* vs. outbreeder *M. guttatus*), life history (annual vs. perennial *M. guttatus*) and edaphic tolerance (*M. nudatus*, serpentine specialist, vs. *M. guttatus*), applying this explosion of new tools to classic, unsolved evolutionary problems.

Part 1. Genetical genomics

The classic paradigm of evolution is that structural differences (amino acid changes) underlie adaptation and evolution. With new gene expression technology, the role of gene regulation is gaining recognition. Furthermore, systems approaches can be used to describe the complex molecular networks that plants utilize to integrate metabolic, cellular, and developmental processes for Darwinian adaptation. Ultimately we need to develop an evolutionary interpretation of this systems biology paradigm.

"Genetical genomics" uses genetic variation to study the genetic determination and evolution of gene expression, using either segregating pedigrees (Chen et al. 2007) or related species that have diverged for gene expression (Zhou and Gibson 2004). Segregating progenies are used to analyze genetic control of gene expression and infer functional interactions between genes. Causal associations between genes and phenotypes can be inferred using phenotypic QTLs as starting point. Divergence among species allows one to examine the relative roles of drift vs. selection in affecting the evolution of gene expression. This approach contrasts with the mutational studies of plant biosynthetic pathways undertaken mainly in *Arabidopsis*, where artificial mutations, as opposed to natural variation, are used in functional studies.

Genetical genomics requires a suite of interacting genes of relevance to the phenotype. The phenylpropanoid pathway has received the attention of several plant molecular evolutionary studies (e.g. Ramos-Onsins et al. 2007). This pathway produces a class of unique plant molecules from phenylalanine (Noel et al. 2005), including flavonoids (implicated in floral pigmentation and herbivory protection) and lignins (involved with cell wall properties and the perennial habit). Although traditionally

classified as “secondary compounds”, phenylpropanoid products are now recognized for their significant roles in plant growth, development, reproduction, adaptation, and defense (Tsai et al. 2006). The pathway is well-elucidated and fully-described in the Kegg database (<http://www.genome.ad.jp/kegg/pathway/ko/ko00940.html>).

Proposed activities (Part 1):

1a. *Development of marker mapping tools for Mimulus.* The first step in genetical genomics is to find a suite of polymorphic markers. My collaborator Dr. John Willis (Duke University, NC) recently developed and mapped over 800 STS markers for *M. guttatus* and relatives. These are PCR based markers wherein exon-based primers span introns; polymorphism is detected as length differences, with loci showing 30-50% heterozygosity (J. Willis, *pers. comm.*). We will develop multiplexed (5-10X) sets of evenly spaced markers for high throughput assays of pedigrees and species, using LiCor genotyping assays. We have considered more high-throughput assays such as the Illumina Beadstation (used in the conifer genome project), but these require much larger samples, and do not allow for changes of markers systems through the project.

1b. *Identification of a set of phenylpropanoid-pathway enzymes in Mimulus.* The second step of genetical genomics is to build a suite of putatively interacting genes within a pathway of functional significance. A “global” approach is used here, where as many pathway members as possible are identified, because this allows inferences about the structure of gene networks and how they change among species. (An alternative is to focus on specific genes known to alter function, such as the ABCE-function transcription factors that influence floral form in Arabidopsis, c.f., Soltis et al. 2007, but this does not allow a network approach). To construct a gene set for *Mimulus* we will follow the procedures of Ehling et al. (2005) and Hamberger et al. (2007), both from our tree genome projects at UBC, who have identified ~120 phenylpropanoid pathway genes in *Populus* and Arabidopsis.

1c. *Expression QTL mapping of a biosynthetic pathway.* The phenylpropanoid-pathway microarray will be used for expression QTL (eQTL) mapping and evolutionary comparisons using F2 pedigrees. All will involve the same *M. guttatus* population, an annual ecotype from Tulloch Reservoir, in the interior of California, representing the ancestral phenotype (this will allow lineage specific inferences to be made, as described below). This population is at the center of the *M. guttatus* distribution. Two members from this population will be crossed to (1) selfing *M. micranthus* (2) perennial *M. guttatus* (Pt. Reyes, Calif.), and to (3) the serpentine endemic *M. nudatus* (Lake Co., Calif.), then F2 progenies generated by crossing between the two unrelated F1s.

Each cross will involve genotyping ~100 progeny for ~200 markers and phenotyping for gene expression using real-time PCR on RNA extracts. In the first cross, RNA will be sampled from flowers; in the second, from stems, and in the third, from leaves. With rtPCR, 96 individuals can be assayed for 8 genes (4 runs, two dyes) per day. We have also considered constructing a custom microarray, but this limits us in the flexibility of adding new genes mid-stream, and also costs more.

For these wet-lab activities, we have a fully equipped molecular population genetics lab, the CFI-funded “Genetic Data Center” (GDC; www.forestry.ubc.ca/gdc). We have full access to microarray scanners and real-time PCR from equipment purchased from the Genome Canada/BC genome grant (to be transferred to GDC). Adequate growth chamber facilities exist in the Forest Sciences Center at UBC.

Undirected gene pathways will be inferred by using the co-location of QTL as a measure of interaction in social network reconstruction programs such as Pajek (de Hooy et al 2005). Directed networks will also be inferred using the principle that for a co-located pair, the downstream gene shows no partial correlation with the marker (spurious QTL) whereas the upstream gene still shows correlation (Chen et al. 2007). I have experience in these analyses in the current spruce genome project, where we are inferring eQTLs and constructing networks in two spruce experiments, using a 22,800 member cDNA microarray.

Also, with an assembled genome sequence, *cis* vs. *trans* acting eQTLs will be identified. Finally, when more than one pedigree is finished, we will compare networks across crosses to detect evolutionary changes of regulation. A modification of the method developed by **Chen** and Ritland (submitted) will be used to identify lineage specific changes, using the *M. guttatus* population as a common reference.

1d. *Evolutionary changes of pathway expression.* Evolution of gene expression in this pathway will be examined among eight species of the yellow monkeyflower species complex. For each species, 10 individuals will be sampled for RNA from flowers, stems and leaves (30 total). The experiment will be designed with the help of Rick White, Dept. of Statistics, UBC. Evolutionary changes of gene expression will be evaluated using an ANOVA design. We ask: (1) Is coordinated expression of genes (expression clusters; Ehltng et al., 2005) conserved between species? (2) Are the networks involving various functional groups conserved; if not, what effect is this having on the phenotype? (3) Are there greater changes of gene expression of the end products of pathways, suggesting increased positive selection?

Each species will also be genotyped using the above multiplexed STS markers, and inferred genetic divergence will be used to correct for decreased hybridization affinity of species less related to *M. guttatus* (this and other concerns about cross-species hybridizations are reviewed in Bar-Or et al. 2007). Tests for the relative importance of selection vs. drift will then be conducted (c.f. Fay and Wittkopp 2008)

1e. *Distant comparison of expression patterns.* Using reciprocal protein BLASTs, we will identify homologues of *Mimulus* genes with other published plant whole genome sequences. We will then compare gene expression variation between *Mimulus* and these species (to the extent that expression databanks are available) to deduce changes that have occurred since divergence of conifers and angiosperms.

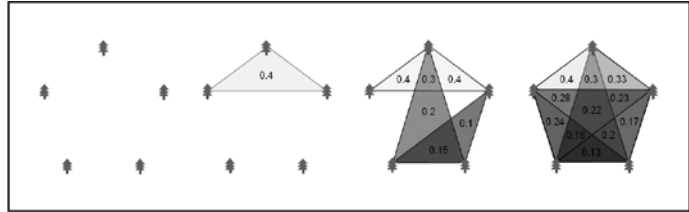
We will also address the hypothesis that the overall structure the pathway has been highly conserved at the earlier stages of the pathway (stabilizing selection), but has undergone adaptive evolution at the later stages (e.g. diversifying selection). This is similar to what we did with nucleotide sequence evolution in **Ramsey** et al. (submitted).

Part 2. Statistical population genomics

The advent of high-throughput genotyping makes more informative inferences about population history and the genetic determination of complex traits. Two-gene coefficients of identity have been used to estimate numerous population genetic parameters, but we extend this to three-gene coefficients. An example of this extension is **Thompson** et al. (2008). Relationships involving three entities considered simultaneously can reveal interactions not detected with pairwise comparisons; interactions in gene networks can also be better resolved.

Proposed activities (Part 2):

2a. *Spatio-genetic triangulation, or landscape genetics with gene trios.* Measures of genetic identity between individuals are adequate for describing patterns of geographic structure in a one-dimensional or linear fashion (STRUCTURE, GENELAND, GENECLUST, TESS; c.f. Francois et al. 2006). To represent genetic relationships over a two-dimensional surface, identities among three genes can be used. As illustrated, when composites of three-gene identities are overlaid and averaged on a cartographic plane, they can lend insight into regional patterns of genetic structure.



In collaboration with Antoine Kremer

(INRA, Bordeaux; who has agreed with this collaboration), this method will be developed and applied to a pan-European chloroplast haplotype dataset from over 2600 populations of European oaks.

2b. *Development of genetic distance measures that take into account progenitor-derivative taxa.* The "effective selfing rate" (E) of plant X mated to plant Y is the probability that a randomly chosen allele from plant Y is identical by descent to either homologous allele in plant X (Ritland 1984). It is an asymmetrical measure in that E for X may not equal E for Y. An analogous measure can be derived for genetic distance, where the distance from population X to Y may not equal that for Y to X. This occurs when taxa differ in diversity, with one tending to have a subset of diversity of the other. This will be applied to the yellow monkeyflower species complex, using the STS data previously generated (eight species each assayed for 400 STS loci in 20 individuals).

Since the map positions of these STS markers are known, we can also look for chromosome specific patterns of derivation. We will look for regions of reduced diversity as evidence for hitchhiking of variants involved in adaptive evolutionary change, and characterize the "landscape" of genetic diversity among the eight species. This includes the separation of linkage disequilibrium from variation of heterozygosity as underlying the correlation of gene diversity between linked loci (Ritland, unpublished).

2c. *Detection of epistatic interactions.* A three gene approach will be used to detect pairwise epistasis between SNPs for a quantitative trait. In this model, two of the genes are SNPs and the third gene is the quantitative trait. This model will be developed and applied to a large association study conducted by Dr. David Neale, UC Davis (a co-funder of the spruce genome project), on 2,000 loblolly pine (*Pinus taeda*) trees interrogated with a 7,600 member Infinium SNP chip. The quantitative traits include wood properties, disease resistance, drought tolerance, gene expression and metabolomic profiles. Dr. David Neale has agreed with this collaboration.

Training aspects

From my current NSERC, I fund and supervise five graduate students. I culture independence and originality of my students in their research. At UBC, we have interactions with students/post-docs in Zoology, Botany, and Land/Food systems, via seminars, discussion groups, and shared lab space. This is especially fostered by our CFI-funded "Genetic Data Center", a lab available to students throughout UBC (and beyond) to conduct laboratory based molecular population genetics research. Dr. Carol Ritland runs this facility; who trains many people in the required research techniques.