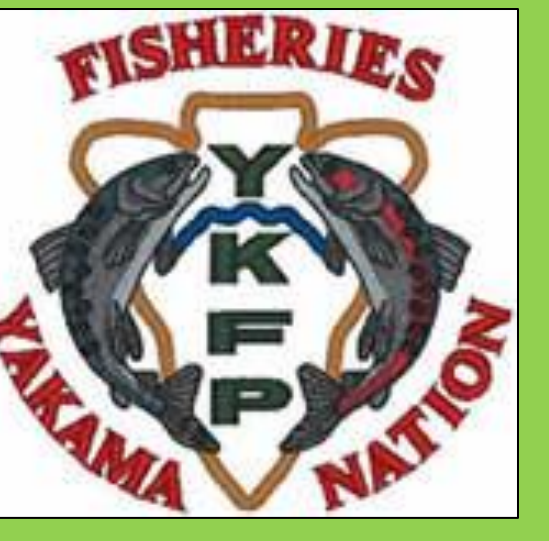


Genetic Basis of Adult Migration-Timing within a Population of Steelhead (*Oncorhynchus mykiss*)



Jon E. Hess¹, Joseph S. Zendt², Amanda R. Matala¹, and Shawn R. Narum¹

¹Columbia River Inter-Tribal Fish Commission, 3059-F National Fish Hatchery Rd, Hagerman, ID 83332

²Yakama Nation Fisheries Program, Yakima/Klickitat Fisheries Project, P.O. Box 151, Toppenish, WA 98948, USA.

Introduction

- Migration traits are presumed to be complex and to involve interaction among multiple genes of minor effect, therefore univariate analyses may lack the power to detect associations.
- Steelhead (*Oncorhynchus mykiss*) is an anadromous salmonid species which exhibits two divergent freshwater migration-timing life histories, winter- and summer- run.
- Our study focused on a tributary of the Columbia River, the Klickitat River, where population divergence between summer and winter run types appears minimal and this rare condition is beneficial to a study of the genetic basis of this trait.

Objective

- Use univariate analyses as well as a random-forest (RF) machine learning algorithm to conduct association mapping of ~15k single nucleotide polymorphisms (SNPs) for adult migration-timing phenotype in steelhead.

Methods Part I

- Adult steelhead were trapped, and then released in the Lyle Falls fish ladder (Rkm 2.4, Fig. 1c) in the Klickitat River.
- Biological data were collected from individual fish including fork length, sex, and a clip of tissue from the caudal fin. A dataset of 180 SNPs was analyzed with a much larger set of steelhead (hatchery and natural origin) that were sampled at Lyle Falls trap between 2007-2013.
- Structure analyses were run with K1-20, and we observed an early plateau at K=6 and then later at K=10. To be conservative, we chose K=10. We removed stray hatchery fish and used a sample of 237 natural-origin fish native to the Klickitat River.

- We conducted Restriction-site Associated DNA sequencing (RAD-seq) and after quality filtering there were 15,239 SNPs that had minor allele frequency >3% and had data for >80% of the 237 individuals.

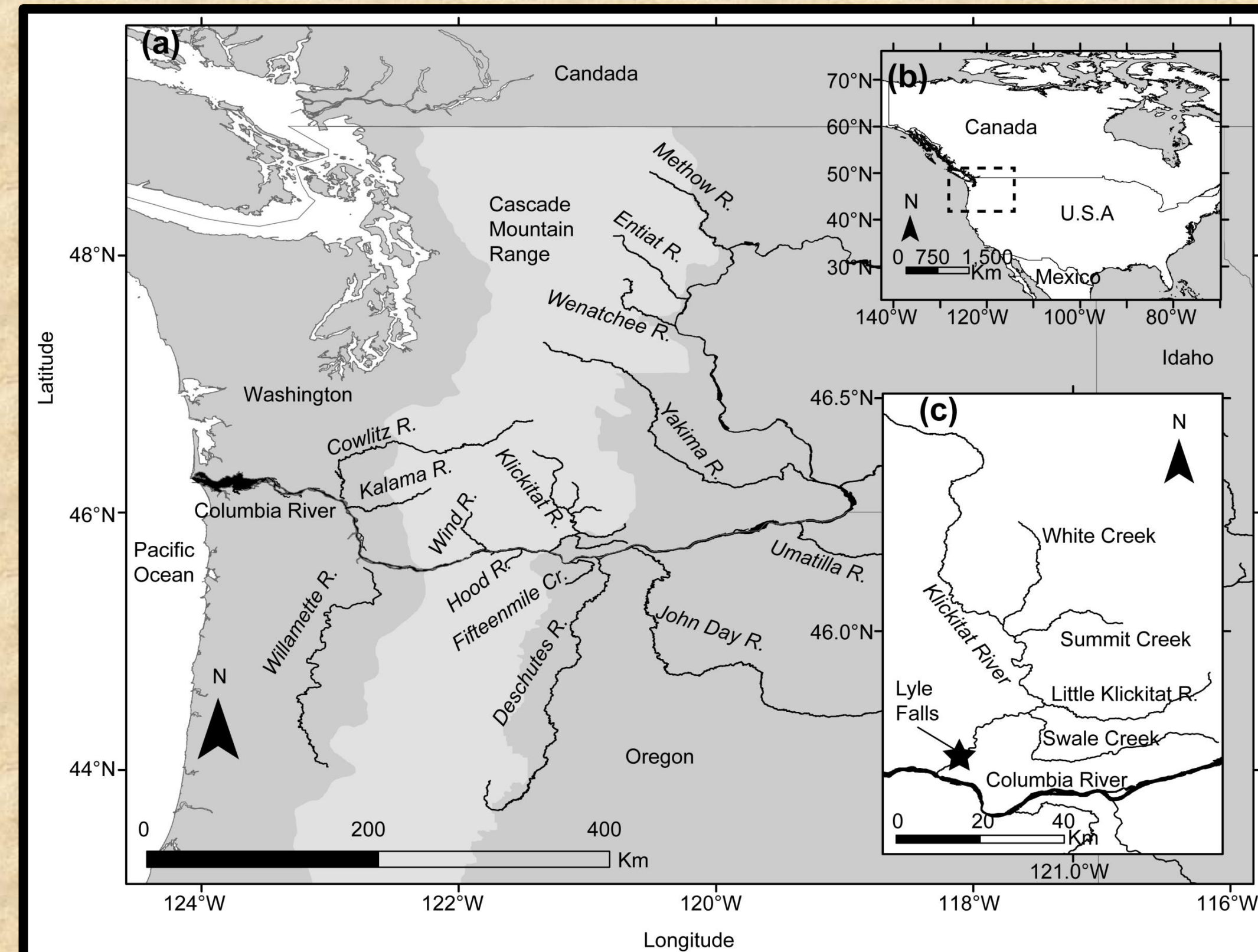


Figure 1. Map of the study site. (a) The Klickitat River tributary and neighboring tributaries of the Columbia River Basin. (b) North America with a dashed box to indicate the study region in the Columbia River Basin. (c) The Klickitat River with labels on secondary tributaries within the subbasin and the site where steelhead were collected at Lyle Falls.

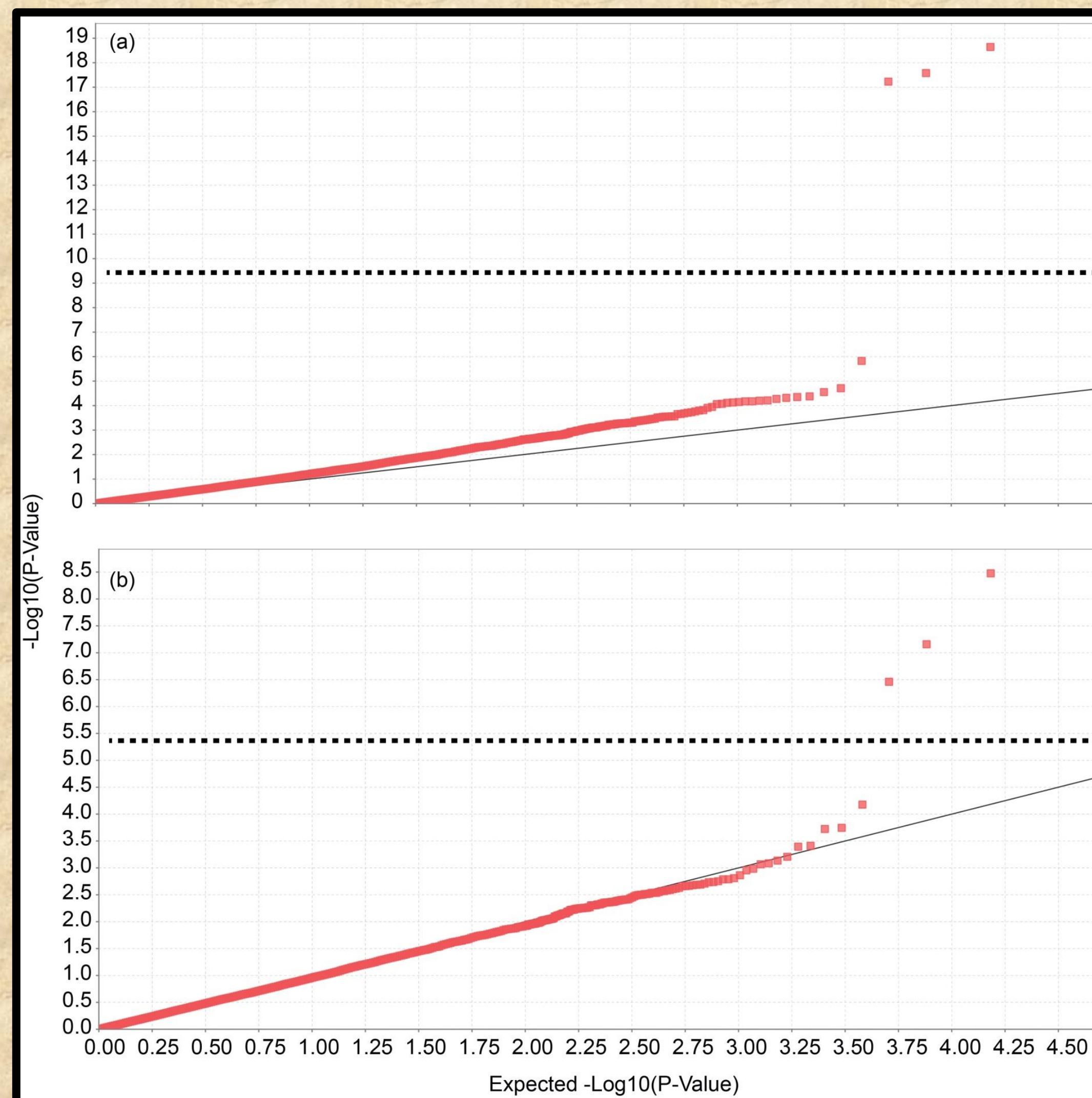


Figure 2. QQ plots showing the Expected $-\log_{10}(P\text{-Value})$ vs. $-\log_{10}(P\text{-Value})$ for (a) general linearized model and (b) mixed linearized model (MLM) of the run timing trait. MLM, which takes into account kinship effects, appears to be a better fit. The heavy dashed lines indicate the Bonferroni corrected alpha level of 0.05.

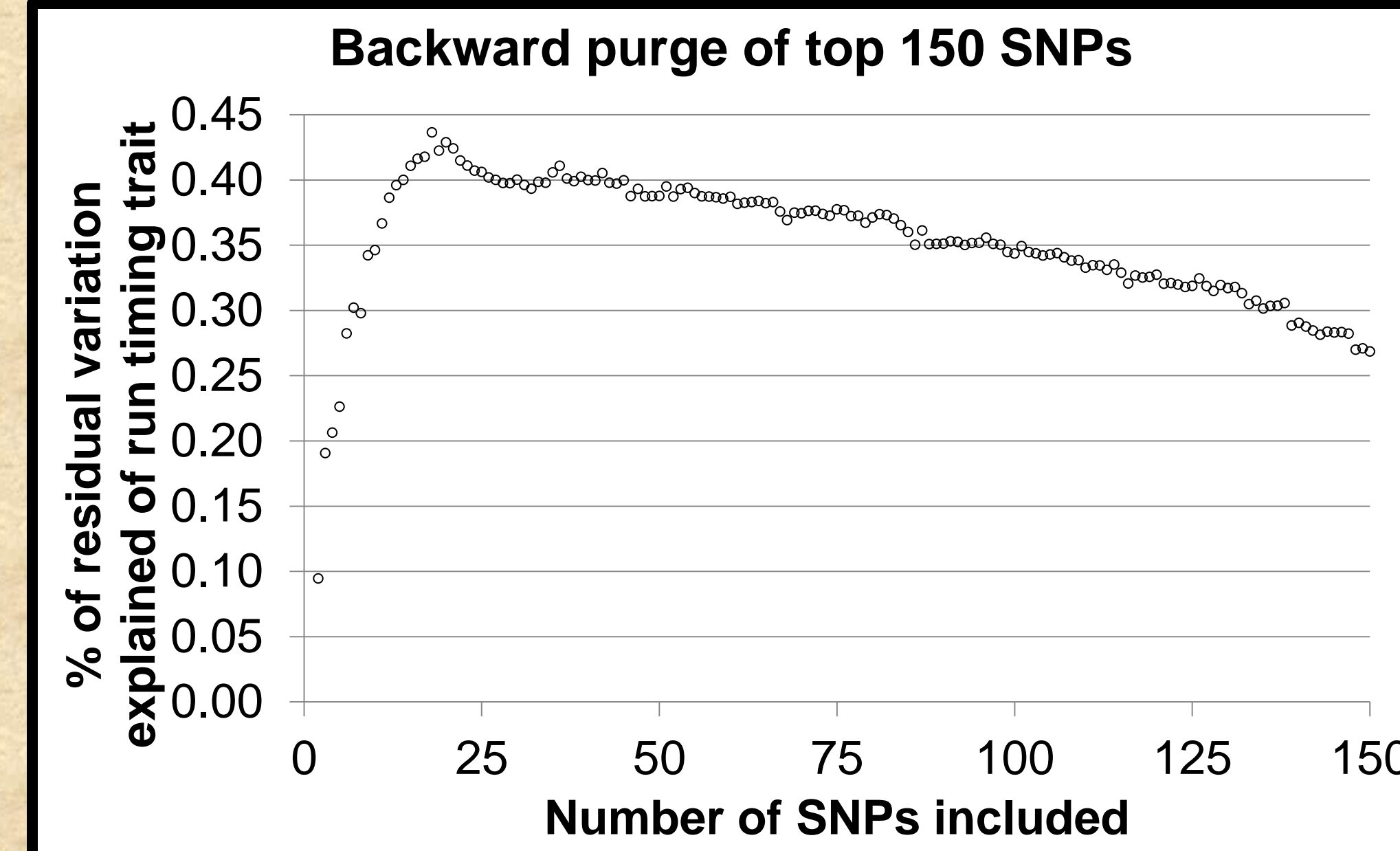


Figure 3. Percentage of residual variation (i.e. after kinship, population structure, and year effects were accounted for) explained in run timing with the number of SNPs included in the RF model. The top 150 SNPs were selected from those with the lowest p values from the MLM univariate test.

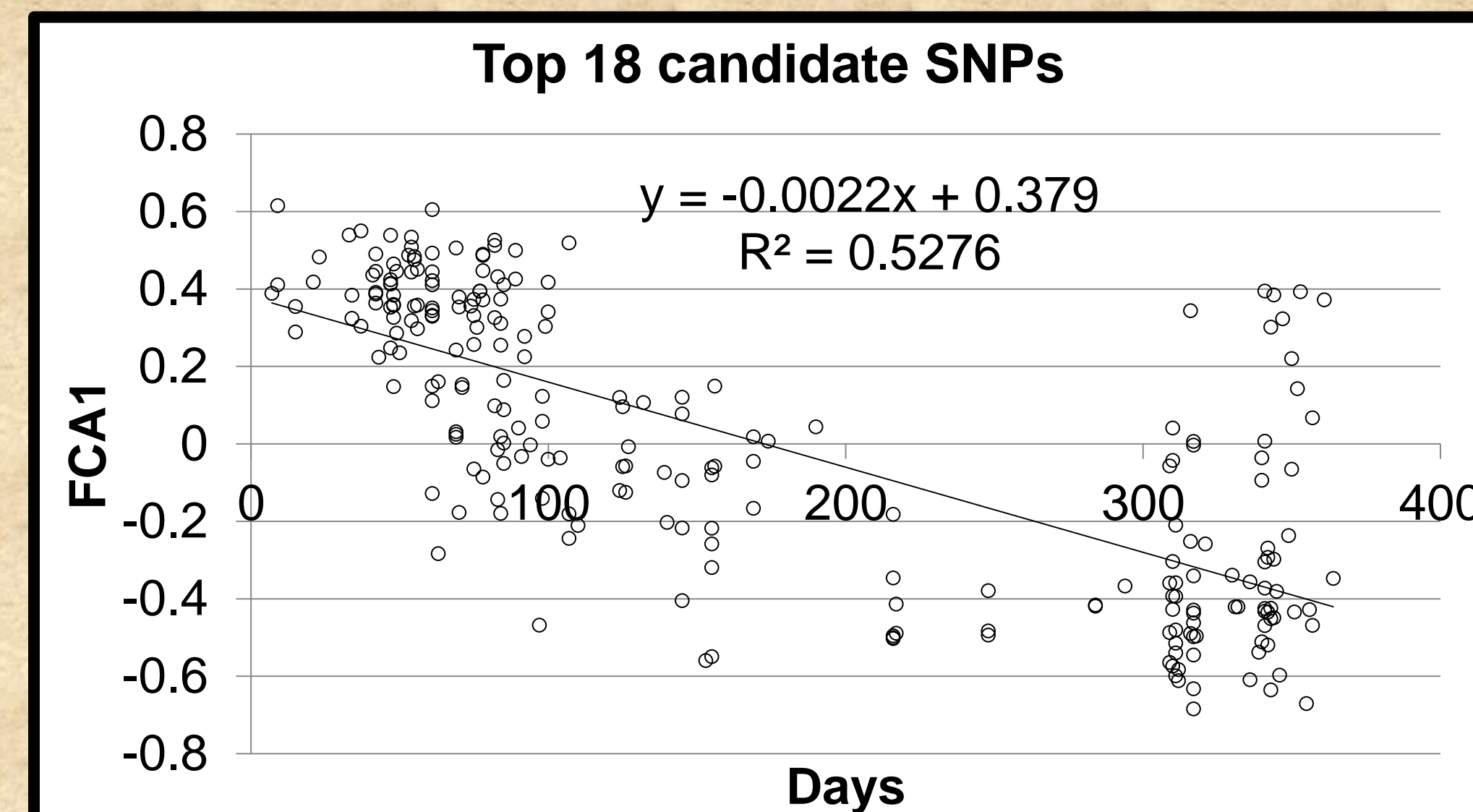


Figure 4. Correlation of factorial correspondence analysis (FCA) axis 1 scores using the top 18 candidate SNPs versus day-of-passage (as ordered by seasonal migration summer to winter).

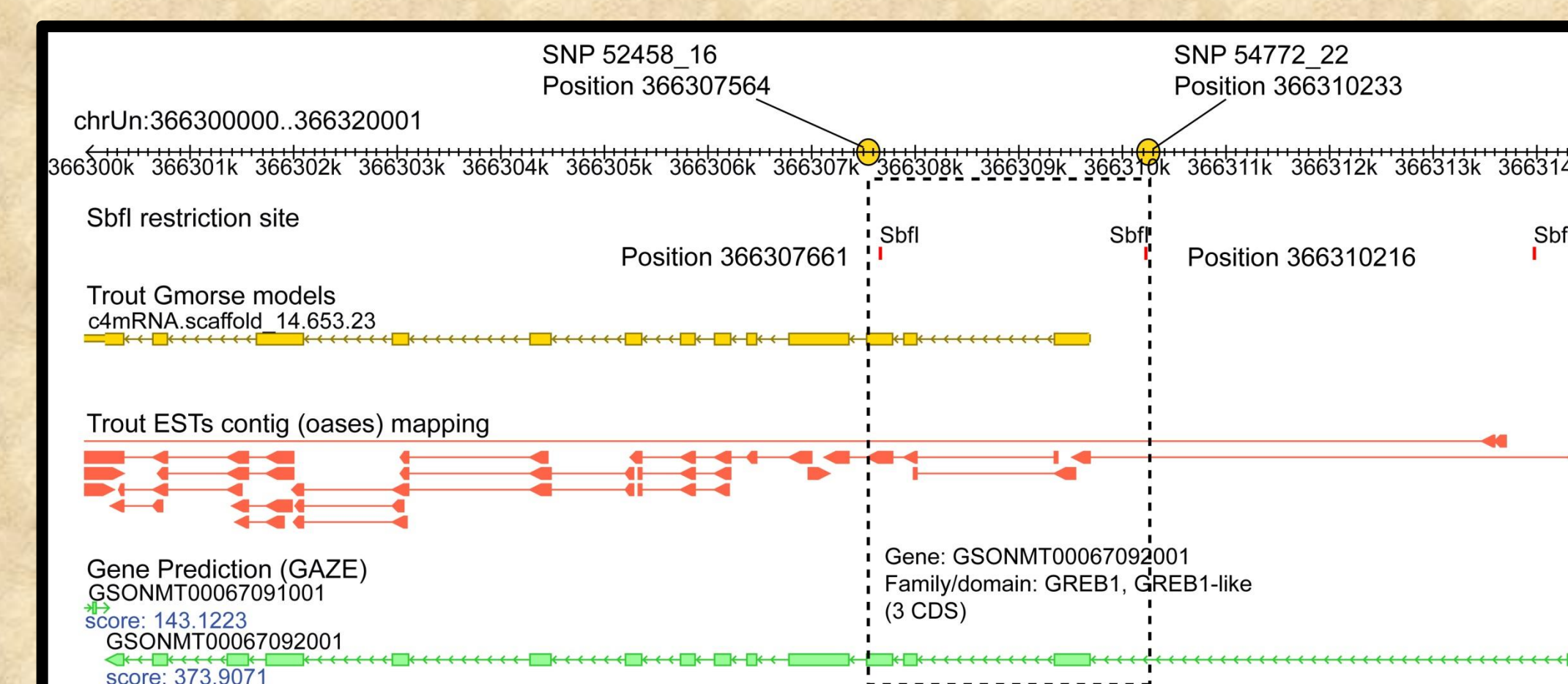


Figure 5. A graphical representation of locations of the 2 candidate SNPs 52458_16 and 54772_22 in the *O. mykiss* unknown chromosome scaffold (chrUn). The region between the SNPs contains 3 CDS of a predicted gene in the GREB1/GREB1-like family. Included in this annotation are the 2 restriction sites for SbfI, which was the enzyme used in the RAD-seq methods to identify these SNPs

Methods Part II

- TASSEL v. 5.0.8 (Bradbury et al. 2007) was used to perform univariate analyses; i.e., general linearized model (GLM) and mixed linearized model (MLM).
- A multivariate type analysis using the RANDOMFOREST package (Liaw & Wiener 2002) was performed in R.
- Genetic variation of 18 top candidate SNPs was analyzed with a factorial correspondence analysis in GENETIX v. 4.03 (Belkhir et al. 2004).

Results

- A univariate MLM found 3 SNPs to be significantly associated with migration-timing (Fig. 2).
- The same 3 SNPs were ranked high in importance values based on RF and explained 46% of trait variation (7% residual).
- However, RF identified 18 minor SNPs were required to reach a maximum ~70% explained trait variation (44% residual, Fig. 3).
- The 18 top minor SNPs exhibit strong correlation with migration-timing (Fig. 4).
- One SNP localized to GDF11, a growth differentiation factor, and 2 of these SNPs were spaced 3000 bp apart within GREB1-like (an estrogen target gene, Fig. 5).

Discussion

- These 18 top candidate SNPs may provide the ability to predict the adult migration-timing of steelhead and facilitate conservation management of this species.
- This study underscores the benefits of an RF multivariate approach for identifying SNPs in minor effect genes of complex traits that may otherwise escape detection in univariate association analyses.

Work cited

Bradbury et al. 2007. *Bioinformatics* 23:2633–2635.
Belkhir, et al. 2004. University of Montpellier II, Montpellier.
Liaw & Wiener. 2002. *R News* 2:18–20.