A roadmap to low-coverage whole genome sequencing (LC-WGS) for population genomics

R. Nicolas Lou[1], Arne Jacobs[1], Aryn Wilder[2], Nina Therkildsen[1]
1. Department of Natural Resources, Cornell University, Ithaca, NY
2. Institute for Conservation Research, San Diego Zoo, San Diego, CA
Email: rl683@cornell.edu

A guide on the design of LC-WGS

- Low-coverage whole genome sequencing (LC-WGS) is a cost-effective sequencing strategy for population genomic studies
- It costs as low as $20 per sample for a species with a 1Gbp genome when sequenced at 1x coverage
- It is not yet clear how we can best balance between its inference accuracy and cost
- We use forward genetic simulation to assess the power of different types of population genomic inference under different LC-WGS designs
- We compare LC-WGS with other widely-used sequencing strategies, including RAD-seq and pool-seq

A review on the methods for data analysis

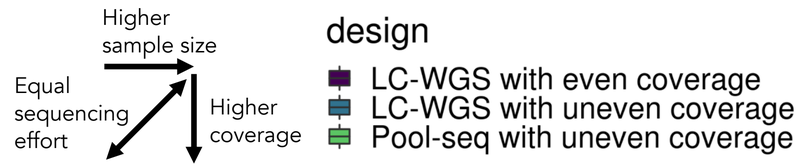
- A second part of this project aims to introduce, compare, and contrast existing methods that have been developed for LC-WGS data; please stay tuned
- We have deposited customizable scripts to implement and streamline many of these methods in our two GitHub repositories: one for data processing and the other for data analysis

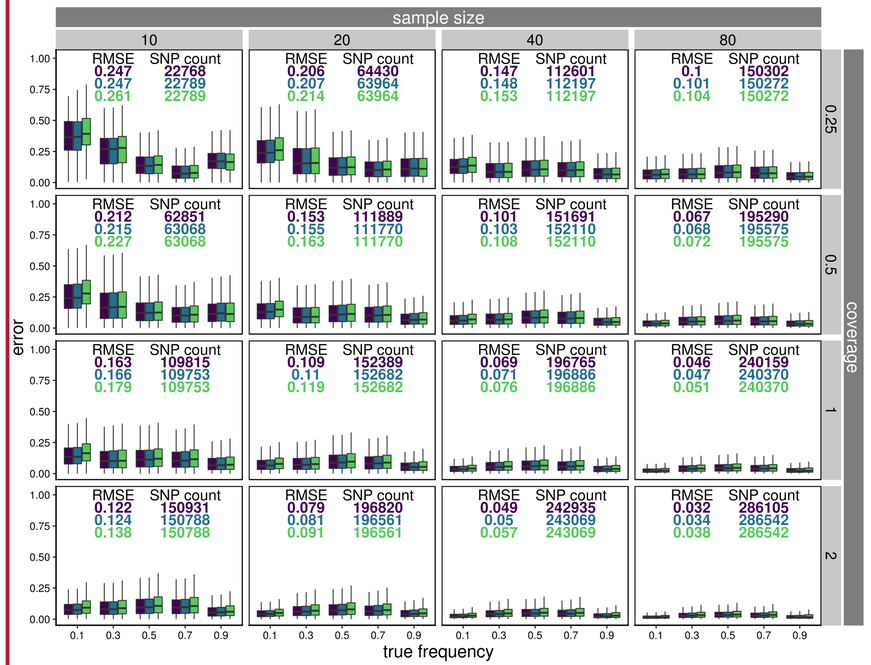
Simulation and analysis pipeline

Forward genetic simulation (SLiM[1])	High-throughput sequencing simulation (ART[2])	LC-WGS data analysis (ANGSD[3])
30Mbp chromosome, **θ**=0.0024, ~600,000 SNPs	150bp paired-end reads, 0.125-8x coverage, 5-160 samples per population	genotype likelihood, SNP calling, allele frequency, F_{ST}, PCA

Allele frequency estimation

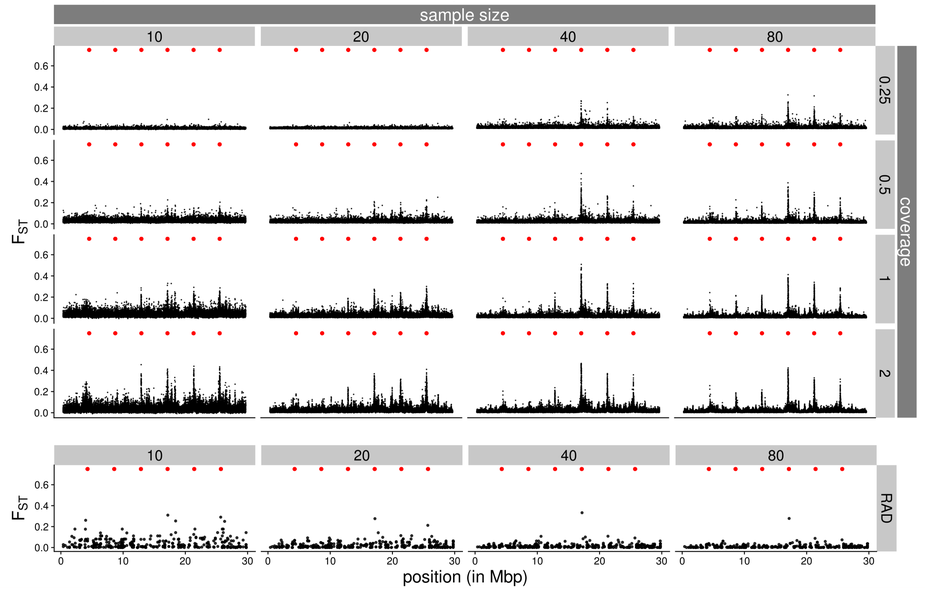
- It is preferable to increase sample size rather than sequencing coverage given the same sequencing effort in total
- LC-WGS can better account for uneven coverage among samples as compared to pool-seq





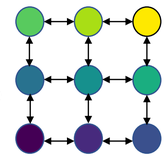
Genome-wide scan for divergent selection

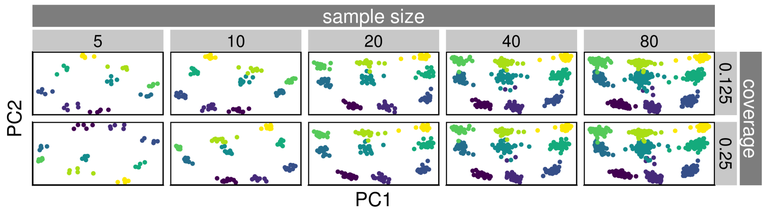
- Divergent selection in high-gene flow systems can create narrow genomic islands of differentiation[4] (F_{ST} peaks) surrounding the sites under selection (positions marked by **red dots**)
- With reasonable coverage and sample size, LC-WGS (**top panel**) can detect many of these signals, but RAD-seq (**bottom panel**) tends to miss them
- Distributing the same amount of sequencing effort across more samples tends to reduce false positive signals



Characterization of spatial population structure

- The true population structure, shown on the right, can be accurately inferred from PCA[3,5] despite extremely low coverage and sample size, since LC-WGS can take advantage of the aggregated genome-wide signal





Literature cited

1. Haller and Messer, 2019. Mol Biol Evol
2. Huang et al., 2012. Bioinformatics
3. Korneliussen et al., 2014. BMC Bioinformatics
4. Turner et al., 2005.. PLoS Biology
5. Patterson et al., 2006. PLOS Genetics