MAKING ANCESTRAL TREES USING BAYESIAN INFERENCE TO IDENTIFY DISEASE-CAUSING GENETIC VARIANTS

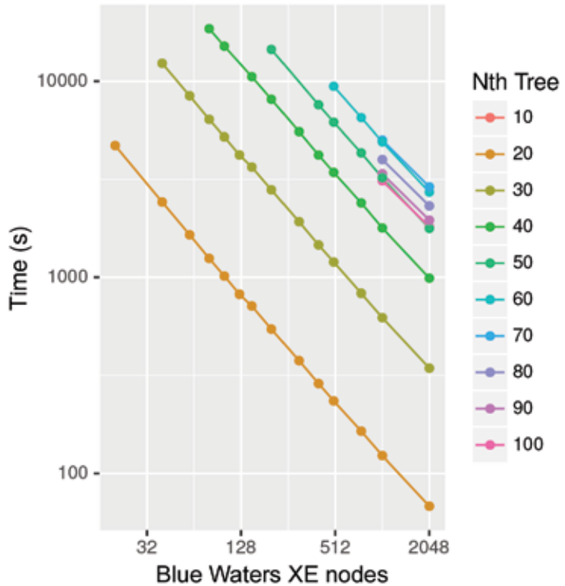
Allocation: Illinois/50.0 Knh
PI: Don Armstrong[1]
Collaborators: Derek Wildman[1] and Monica Uddin[1]

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

A complex network of different interacting genetic variants contributes to many common human diseases. Discovering effective treatments for these diseases requires distinguishing the genetic variants that cause diseases from those variants that are just associated with the disease. Ancestral trees will enable us to identify variants which are associated merely due to ancestry, and decrease the rates of false positives in case-control studies due to admixture. The size of the human genomes (3.2 Gbp x 2000 ≈ 6 TiB) coupled with the gigantic number of possible trees (2000!! or $\approx 4\times10^{2868}$) requires supercomputing on the scale of Blue Waters. In a preliminary allocation, we were able to demonstrate the feasibility of utilizing Blue Waters to generate ancestral trees on large datasets (6 TB) using maximum likelihood and were able to identify bottlenecks in the Bayesian approach.

FIGURE 1: Linear scaling of time in seconds to calculate 10 trees after the 10th, 20th, etc. tree with number of Blue Waters XE nodes starting with 1 million variants and 1000 individuals with the same run parameters.



INTRODUCTION

Many human diseases such as diabetes, cancer, cardiovascular disease, and mental illness are caused in part by a complex network of genetic variants which interact with each other and environmental factors. Ancestral trees, which depict the descent from ancestors of a set of genomic regions, enable us to: 1) distinguish between common and rare variants; 2) identify variants which are associated with disorders; 3) identify sets of cases and controls which are ethnically matched for a particular genomic region. However, the generation of trees is nondeterministic polynomial time (NP) complex; the number of rooted binary trees for 2000 individuals is 2000!!, or $\approx 4\times10^{2868}$. When coupled with the size of the human genome (3.2 Gbp) by 2000 individuals (6 TB of information, uncompressed), and the number of variables (mutation rate, selective forces) which can vary at each position and over ancestral time, the computational problem becomes enormous. Making accurate estimates of ancestral trees requires computational resources on the order of Blue Waters.

METHODS & RESULTS

We used two existing programs which use Bayesian inference (MrBayes) and maximum likelihood (ExaML[1]) to estimate ancestral trees on smaller regions of genomes in 1000 individual genomes from the 1000 Genomes Project [2]. Initial experiments identified scaling issues with MrBayes beyond four nodes which we are currently working on resolving using an XSEDE allocation. ExaML was able to scale one million variants linearly to 2048 nodes (Fig.1), and additional scaling to larger numbers of nodes and variants should be possible once we resolve issues with the checkpoint code and initial tree generation. The trees generated from this analysis are congruent with the ethnicity of the individuals (Fig. 2) and we expect to find similar patterns when we can calculate the trees across the entire genome using a full allocation. During this preliminary allocation, we utilized 44,000 node hours.

WHY BLUE WATERS

Blue Waters is one of the only systems that has the computational and I/O resources at the scale necessary to calculate enough trees to examine the tree space in sufficient detail necessary to obtain trees with global maximum likelihood as opposed to a tree with local maximum likelihood. Even our preliminary analyses are beyond the scale of other existing computational resources.

NEXT GENERATION WORK

We hope to extend the initial tree that we generated with additional whole genome sequences as projects like H3 Africa expand our knowledge of the complete variation and population migrations represented through human history.

FIGURE 2: Preliminary ancestral tree from 1 million variants which was calculated from the longest run (8 hours) on 2048 Blue Waters XE nodes. Colors represent reported ethnicities from a subset of the 1000 Genomes Project and are relatively congruent with similar ethnicities and known human migration patterns.