Variant Correlation and Wavelet Transforms: Revealing Chromosome Structure in *Populus trichocarpa*

Deborah Weighill[1,2], Stephen DiFazio[3], David Macaya-Sanz[3], Wayne Joubert[4], Gerald Tuskan[2] and Daniel Jacobson[1,2]

[1]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, 444 Greve Hall, 821 Volunteer Blvd. Knoxville, TN 37996-3394; [2]Biosciences Division, Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN 37831; [3]Department of Biology, West Virginia University, 5200 Life Sciences Building, 53 Campus Drive, Morgantown, WV 26506-6057; [4]OLCF Group: Scientific Computing, Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN 37831

Introduction

A vast collection of different data types is available for *Populus trichocarpa* [1]. A collection of ~28,000,000 Single Nucleotide Polymorphisms (SNPs) called across 882 genotypes have recently been publicly released. DNA methylation data in the form of MeDIP (Methyl-DNA immunoprecipitation) sequencing has been performed on 10 different *P. trichocarpa* tissues [2]. We have used these different data types as signals, which vary across a chromosome, for example, the gene density, SNP density, SNP correlation (LD) density, TE density, low complexity sequence density, or methylation density across a chromosomes in *P. trichocarpa* genome. Applying the Continuous Wavelet Transform signal processing technique allowed us to characterize the variation in these signals at multiple scales, and consequently, allowed us to identify the centromere (Figure 1) positions of each chromosome based on these various data signals.

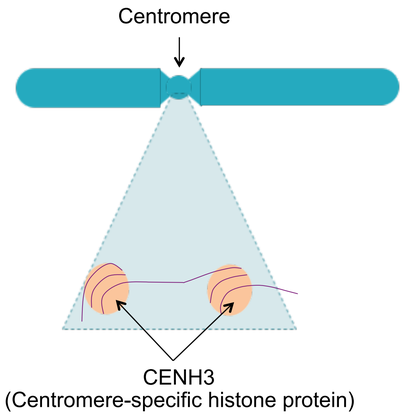

Figure 1: Diagram of chromosome depicting centromere.

Feature Construction and Wavelet Transforms

P. trichocarpa SNPs (DOI 10.13139/OLCF/1411410) available on the OLCF DOI resource were obtained, and *P. trichocarpa* genome annotations [1] as well as aligned MeDIP reads [2] were obtained from Phytozome [3]. For each chromosome, signals were constructed based on 4 features: SNP density, gap density, gene density and MeDIP read count density. Each of these signals was constructed for 10kb windows.

Figure 2: Feature Signals for Chromosome 2. The outlined region indicates where the SNP and gene signals have valleys whereas the MeDIP signal has a peak. This coincides with the putative centromere position of chromosome 2 [2].

$$W(s,\tau) = \frac{1}{\sqrt{s}} \int f(t)\psi^* \left(\frac{t-\tau}{s}\right) dt \qquad (1)$$

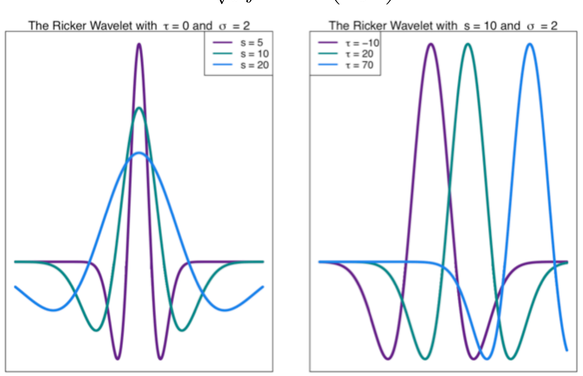

Figure 3: The Ricker wavelet shown for different values of scale *s* and translation *tau* in Equation 1.

Continuous Wavelet Transforms

Chromosome feature signals contain variation on multiple scales, including high frequency (narrow) peaks and low-frequency (broad) peaks. These different scales of peaks contain different information. Thus, techniques to analyse these signals at different scales are valuable (see [5,6]). The Wavelet Transform can be used to unpack the information in different scales of a signal. What results from a wavelet transform is a wavelet coefficient W(s), for every scale s and translation tau [8,9]. Given the peak-like shape of the wavelet (Figure 3), a wavelet coefficient will indicate "how much of a peak" is present at a particular scale and at a particular position of the signal. Thus, the wavelet transform allows us to investigate the peaks of a signal at different scales.

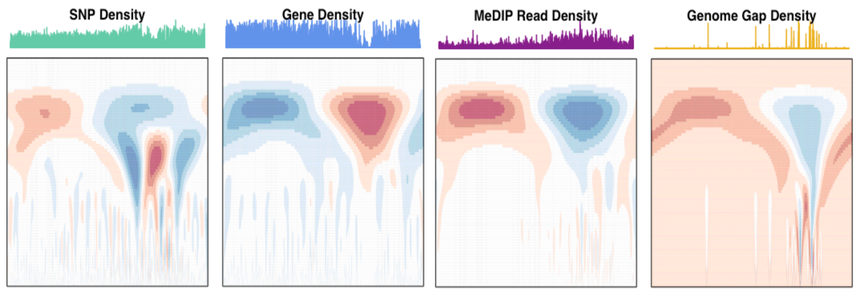

Figure 4: CWT Coefficient landscapes of chromosome 2 for (A) SNP density, (B) gene density, (C) methylation (MeDIP-Seq read density, internode explant tissue) and (D) genome gap density. X-axes represent the bp dimension of the signals, Y-axes represent scales (s in Equation 1). Blue regions indicate positive coefficients and red regions indicate negative coefficients.

Centromere Position Identification

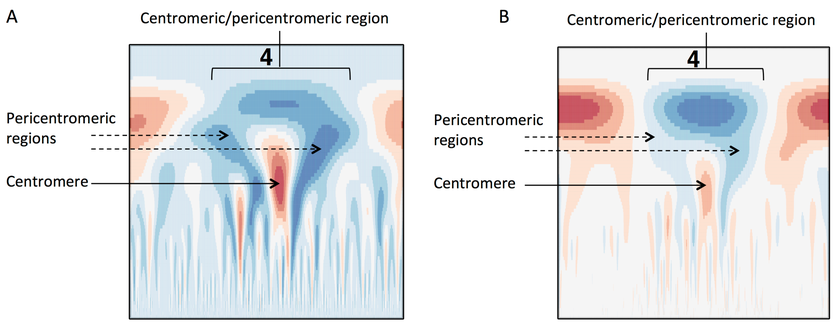

Figure 5: "Tooth x-ray" centromeric signature for (A) SNP density and (B) methylation density, consisting of a broad-scale peak encompassing the centromeric/pericentromeric regions, and the lower scale valley within the large peak indicating the centromeric region.

1. Identify the position of the maximum wavelet coefficient in the upper third of the methylation landscape (**pericentromere scale**).
2. Find the putative pericentromere borders as the zeros on their side of this maximum.
3. Identify the the minimum coefficient in the lower two thirds of the SNP wavelet landscape, between the approximate pericentromere borders (**centromere scale**).
4. Extract the SNP wavelet coefficient vector at centromere scale and the methylation wavelet coefficient vector at pericentromere scale.
5. Mean center (mean = 0) and scale these vectors to have standard deviation 1, and find the approximate centromere location as the position of maximum difference between these two vectors.

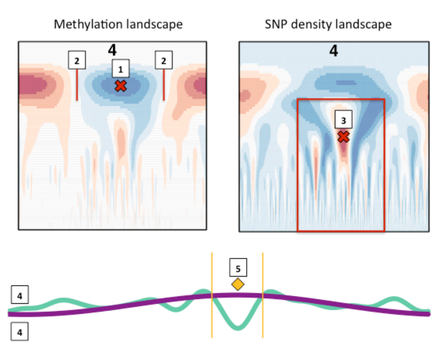

Figure 6: Steps in the centromere Identification strategy.

Centromere Repeat Sequences

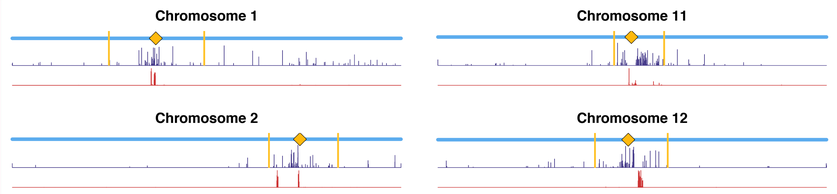

Figure 7: Putative centromere positions (yellow diamonds) using methylation and SNP wavelet coefficients, as well as the density of BLAST matches of plant centromere repeat sequences (navy bars) and putative *P. trichocarpa* centromere repeat sequences (red bars) [11].

CENH3 Co-evolution

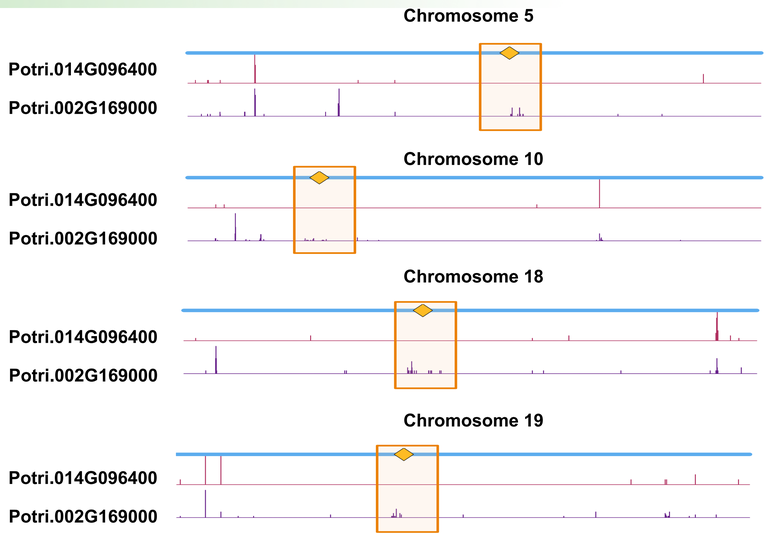

Figure 8: SNP density profiles across a selection of chromosomes involving SNPs which which correlate with SNPs in putative CENH3 genes, Potri.002G169000 and Potri.014G096400 across a population of *P. trichocarpa* genotypes. One can clearly see the clusters of SNPs in the centromeric regions which are correlating with SNPs within these CENH3 genes.

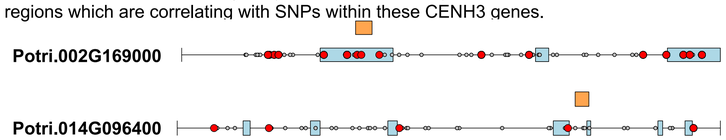

Figure 9: SNPs in putative *P. trichocarpa* CENH3 genes Potri.002G169000 and Potri.014G096400. Exons (blue) for Potri.014G096400 were determined from the v3.0 genome annotation on Phytozome and from mapped ESTs on Phytozome Jbrowse for Potri.002G169000. Grey circles represent SNPs, red circles represent SNPs that correlate with SNPs in centromeric regions. Orange rectangles indicate the location of the histone domain as determined using NCBI CDScan.

Acknowledgements

Funding provided by The BioEnergy Science Center (BESC) and The Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science. This research was also supported by the Plant-Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov) in the Genomic Science Program, the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science, and by the Department of Energy, Laboratory Directed Research and Development funding (7758), at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725. An award of computer time was provided by the INCITE program. This research also used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. This research also used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE- AC05-00OR22725. Support for the Poplar GWAS dataset was provided by The BioEnergy Science Center (BESC) and The Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science The Poplar GWAS Project used resources of the Oak Ridge Leadership Computing Facility and the Compute and Data Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. Support for DOI 10.13139/OLCF/1411410 dataset is provided by the U.S. Department of Energy, project BIF102 under Contract DE-AC05-00OR22725. Project BIF102 used resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725

References

[1] Tuskan, Gerald A., et al. "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)." *science* 313.5793 (2006): 1596-1604.
[2] Vining, Kelly J., et al. "Dynamic DNA cytosine methylation in the ###Populus trichocarpa genome: tissue-level variation and relationship to gene expression." BMC Genomics 13.1 (2012): 27.
[3] Goodstein, David M., et al. "Phytozome: a comparative platform for green plant genomics." *Nucleic acids research* 40.D1 (2012): D1178-D1186.
[5] Climer, Sharlee, et al. "A Custom Correlation Coefficient (CCC) Approach for Fast Identification of Multi-SNP Association Patterns in Genome-Wide SNPs Data." *Genetic epidemiology* 38.7 (2014): 610-621.
[6] Spencer, Chris CA, et al. "The influence of recombination on human genetic diversity." *PLoS Genet* 2.9 (2006): e148.
[7] McCormick, Ryan F., et al. "The Sorghum bicolor reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization." *bioRxiv* (2017): 110593.
[8] Percival, Donald B., and Andrew T. Walden. *Wavelet methods for time series analysis*. Vol. 4. Cambridge university press, 2006.
[9] Leavey, C. M., et al. "An introduction to wavelet transforms: a tutorial approach." *Insight-Non-Destructive Testing and Condition Monitoring* 45.5 (2003): 344-353.
[10] Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13.11 (2003): 2498-2504.
[11] Cossu, Rosa Maria, et al. "A computational study of the dynamics of LTR retrotransposons in the Populus trichocarpa genome." *Tree genetics & genomes* 8.1 (2012): 61-75.
[12] Joubert, Wayne, et al. "Parallel Accelerated Custom Correlation Coefficient Calculations for Genomics Applications." *arXiv preprint arXiv:1705.08213* (2017).
[13] **Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction. Wayne Joubert, Deborah Weighill, David Kainer, Sharlee Climer, Amy Justice, Kjiersten Fagnan, Daniel Jacobson. Gordon Bell Prize Submission.**