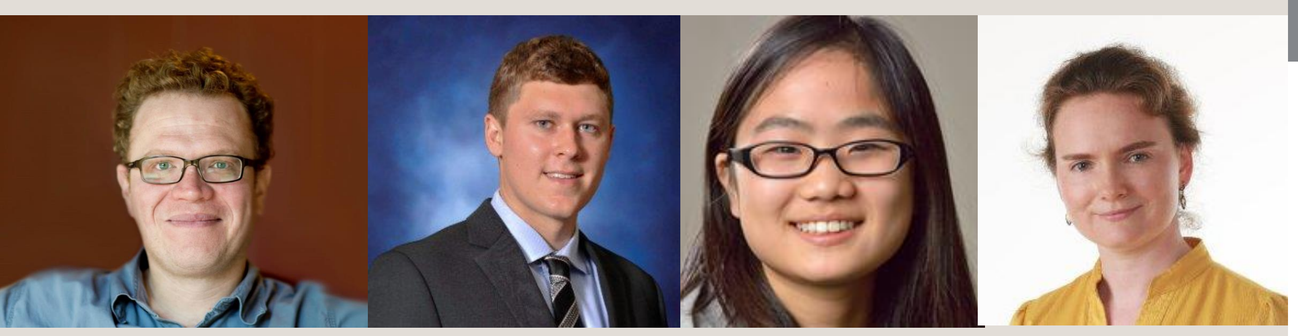
Variant calling by assembly:

complex variants and repeats in populations of complex genomes

Matthew Hudson, Matthew Kendzior, Junyu Li, Liudmila S. Mainzer

AN NSF INDUSTRY/UNIVERSITY COOPERATIVE RESEARCH CENTER



NEED AND INDUSTRIAL RELEVANCE

Genome complexity hinders the construction of quality reference genomes, leading to difficulty in detecting large structural variants and important repeats. For example, many crop genomes are highly repetitive and polyploid due to whole-genome duplication events. Polyploidy can range from 3-ploid (triploid) in bananas to 12-ploid in sugarcane. Tumor cells in human cancers exhibit chromosomal instability, aneuploidy, and high mutation rates. Thus, reference based variant calling pipelines can have high rates of false negative variant calls, as structural variants can be easily missed or miscalled. For example, variant sequences can align to other regions of a polyploid genome in the case of deletions, or not align to the reference at all in the case of insertions.

Variant calling by assembly can lead to the detection of sample variants that may have otherwise gone undiscovered by alignment to reference. The algorithms used here perform de Bruijn graph comparison between samples, in order to characterize complex structural variants that may not appear in the reference sequence. However, the computational requirements to employ variant calling by assembly can be very demanding, especially in organisms with highly repetitive and polyploid genomes.

Objectives

1. Employ and scale the *Cortex_var* software on for use on iForge, the NCSA's supercomputer for Industry.
2. Develop a workflow specifically for use in complex genomes, such the crops.
3. Identify and validate difficult-to-call genomic loci.

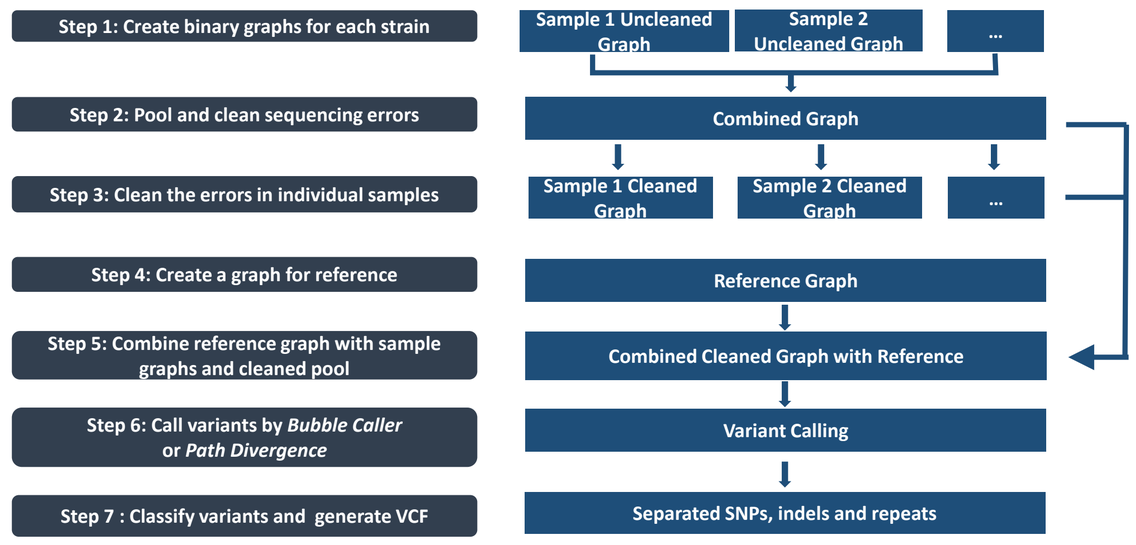
Computational Requirements

iForge queues:

"normal"	24 Intel "Haswell"cores	64 GB of RAM per node
"big_mem"	20 Intel "Ivy Bridge" cores	256 GB of RAM per node
"super_mem"	60 Intel "Ivy Bridge" cores	1.5 TB of RAM per node

Memory Usage	4 samples of 8,791,954,193 bp or less	1 sample of 14,314,137,130 bp
Step 1	112 GB	900GB
Step 2	225 GB	901 GB
Step 3	150 GB	1201 GB
Step 4	112 GB	112 GB
Step 5	225 GB	1201 GB
Step 6	225 GB	1201 GB

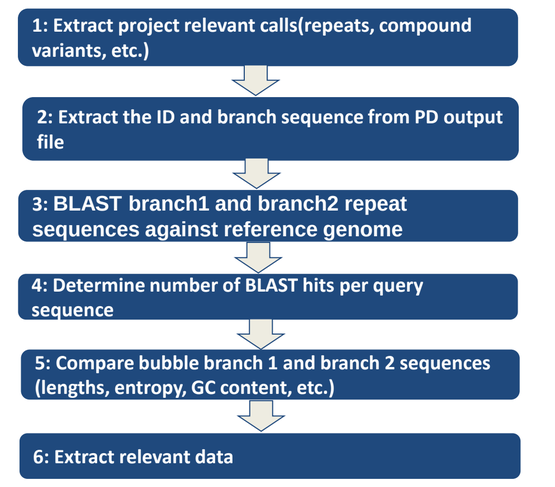
Cortex_var iForge Workflow



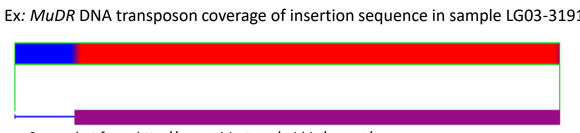
- *Cortex_var* takes short sequencing reads from several samples and assembles them simultaneously into de Bruijn graphs, which are then compared to look for divergences along the traversal path. These divergences are classified as potential SNPs, complex variants, or genomic repeats.
- Variants can be identified in a completely reference free manner, but if there is a reference available, it can be used to roughly place where the samples diverge from the reference.

Making Sense of *Cortex_var* Output

- Using raw sequencing data from the soybean nested association mapping (SoyNAM) population, we have located many large insertions in samples relative to the reference. So far, most are the result of transposon movement.
- We optimized this "downstream" workflow to detect novel repetitive insertions in samples or groups of samples relative to the reference genome.
- BLAST was performed on insertion sequences determine an approximate genomic location and how frequently the sequence appears in the genome.

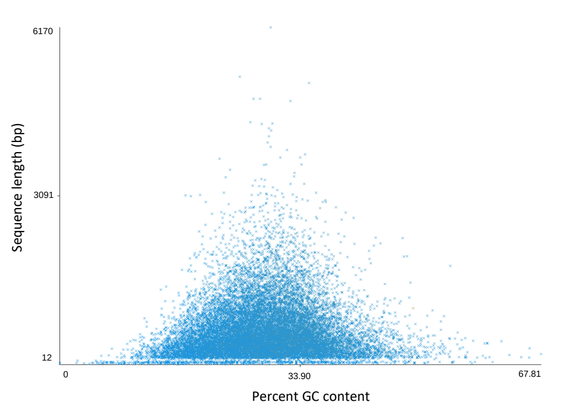


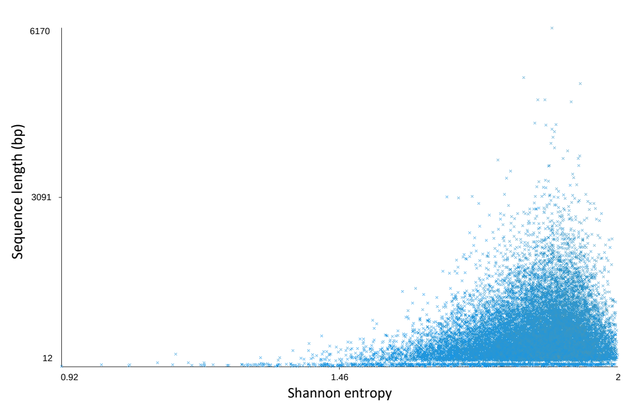
Ex*: MuDR* DNA transposon coverage of insertion sequence in sample LG03-3191.



Screenshot from: http://www.girinst.org/cgi-bin/censor/

Sequence properties of genomic repeat insertions in SoyNAM samples relative to the soybean reference genome





Impact

A combination of population genomics, de Bruijn graph comparison, and workflow automation in appropriate HPC environments, will enable faster and cheaper detection of complex differences between large genomes.

1. **De novo assembly and genotyping of variants using colored de Bruijn graphs.** Z Iqbal(*), M Caccamo(*), I Turner, P Flicek, G McVean, **Nature Genetics** (2012) doi:10.1038/ng.1028
2. http://www.soybase.org/SoyNAM/
3. Bao, W., Kojima, K.K., Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA, 2015;6:11
4. http://www.ncsa.illinois.edu/industry/iforge