Identifying Genetic Load By Whole Genome and Transcriptome Sequencing of Cassava Breeding Lines

Ramu Punna, Fei Lu, Janu Verma and Edward S. Buckler

Institute of Biotechnology
Cornell University
Ithaca, NY - 14853

Identifying Genetic Load By Whole Genome and Transcriptome Sequencing of Cassava Breeding Lines

CASSAVA
(Manihot esculenta)

- 2n=36
- Highly heterozygous species
- Genome size – 770 Mb
- Reference genome – AM560-2 (CIAT – Colombia)

- Clonally propagated crop
- Less/no recombination events - purge deleterious mutations

Accumulate deleterious mutations in every generation

Deleterious mutations are enriched in low recombination regions

Deleterious mutations are at the heart of inbreeding depression

Maize well known to show inbreeding depression
(Jones 1924; Neale 1935)

Effect of deleterious mutations in Cassava – is very severe

- Clonal propagation
- Accumulation of deleterious mutations/generation – high
- No recombination events

50 – 65 % reduction in cassava root yield in S1 generation

Constant Mutational Pressure

- 90 single nucleotide mutations in maize/generation
- 9 of these mutations are likely deleterious every generation
- Deleterious alleles are ubiquitous yet rare
**Objectives**

- To identify deleterious mutations in functional domains of cassava genome
- Build models to predict the performance of these deleterious polymorphisms

**Whole-genome sequencing**

- 20-30x coverage
- 3-10 kb contigs
- Align to reference genome
- To call large proportion of deleterious variants

**Variant calling**

- Maize HapmapV2 pipelines
- DISCOVAR variant calling
- GATK
- Freebayes

**Diverse cassava clones selected for WGS to develop ‘cassava hapmap’**

<table>
<thead>
<tr>
<th>Population</th>
<th>Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>IITA</td>
<td>32</td>
</tr>
<tr>
<td>NaCRRI</td>
<td>32</td>
</tr>
<tr>
<td>NRCRI</td>
<td>44</td>
</tr>
<tr>
<td>CIAT</td>
<td>44</td>
</tr>
<tr>
<td>Wild</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
</tr>
</tbody>
</table>

- 16 clones (50 Mb of contigs)
- > 20 clones (> 20 Mb)

**‘Annotation’ of each and every variant in cassava hapmap with key ‘attributes’**

- Allele frequency
- Conservation with other species
- Transition and transversion
- Gene importance: Expression level in 5 tissues
- More ...

- Will be available in CassavaBase and Phytozome
- GWAS/Candidate gene analysis

**Genomic Evolutionary Rate Profiling (GERP) scores**

- Identifying evolutionary constraints - invariants
- Mutations at evolutionarily conserved sites - higher likelihood of being deleterious.

**Evolutionary constraints**

- Human ~ 5% (150 Mb)
- Drosophila ~ 42% (50 Mb)
- Cassava > humans and flies

More % of cassava genome is evolutionarily constrained
GERP scores are enriched in coding regions

Chromosome 12

Functional annotations from RNA sequencing

Current status: Sequencing

Modeling of Genomic constraints

Machine learning algorithms with minimum of 25 attributes

Make prediction for each minor allele about how likely it is to be deleterious

Sum for all deleterious mutants in a clone to estimate ‘genetic burden index’

Genetic burden will be tested empirically with inbreeding populations

Test and evaluate machine learning algorithms

Existing genomic selection models developed based on high allele frequency

Enhanced Genomic Prediction?

Complementary to ongoing GS models, breeding: 15x faster, currently 2-3x
Easy to purge deleterious mutations

- Purge genetic load/gamete – over next 5-10 generations
  - Model: Need not be perfectly accurate
  - Increase accuracy over 100-1000 variants
    (reducing genetic load in this manner we might achieve what occurred over first 50 years of 20th century maize breeding)

- Alternative – Genome editing
  - E.g.: CRISPR technology
    (100s at a time)

Conclusions

- Cassava – rich in deleterious mutations – affects fitness of the trait
- Machine learning algorithms with inclusion of index from annotations – increases the accuracy in prediction

- Even if predictions of deleterious mutations fail
  - variants and annotations available to cassava community/NextGen Cassava through CassavaBase
    - population genomics, functional genomics
    - evolutionary studies
    - GWAS and GS analysis

Acknowledgements

Cornell University
Ed Buckler
Fei Lu
Janu Verma
Eli Rodgers-Melnick
Lab members

NACRI
Robert Kawuki
Esuma Williams

IITA
Ismail Rabbi

NRCRI
Chiedosie Egesi

Thank you...