



The Digital Cassava Genebank

Sarah Dyer, Bruno Santos, Mohamed Abdelhalim, Pradeep Ruperao, David Marshall & Peter Wenzl



Cassava (*Manihot esculenta*)

Widely grown in tropical and sub-tropical areas for food, feed, starch and bioethanol

3rd most important staple in tropics, feeding **>500 million people** globally

Clonally propagated, slow breeding cycle, inbreeding depression

Diploid, 770Mb genome, highly heterozygous

Reference genome: Bredeson *et al.*, Nature Biotech (2016)




www.shutterstock.com · 83105725

Accessing diversity: CIAT's cassava collection

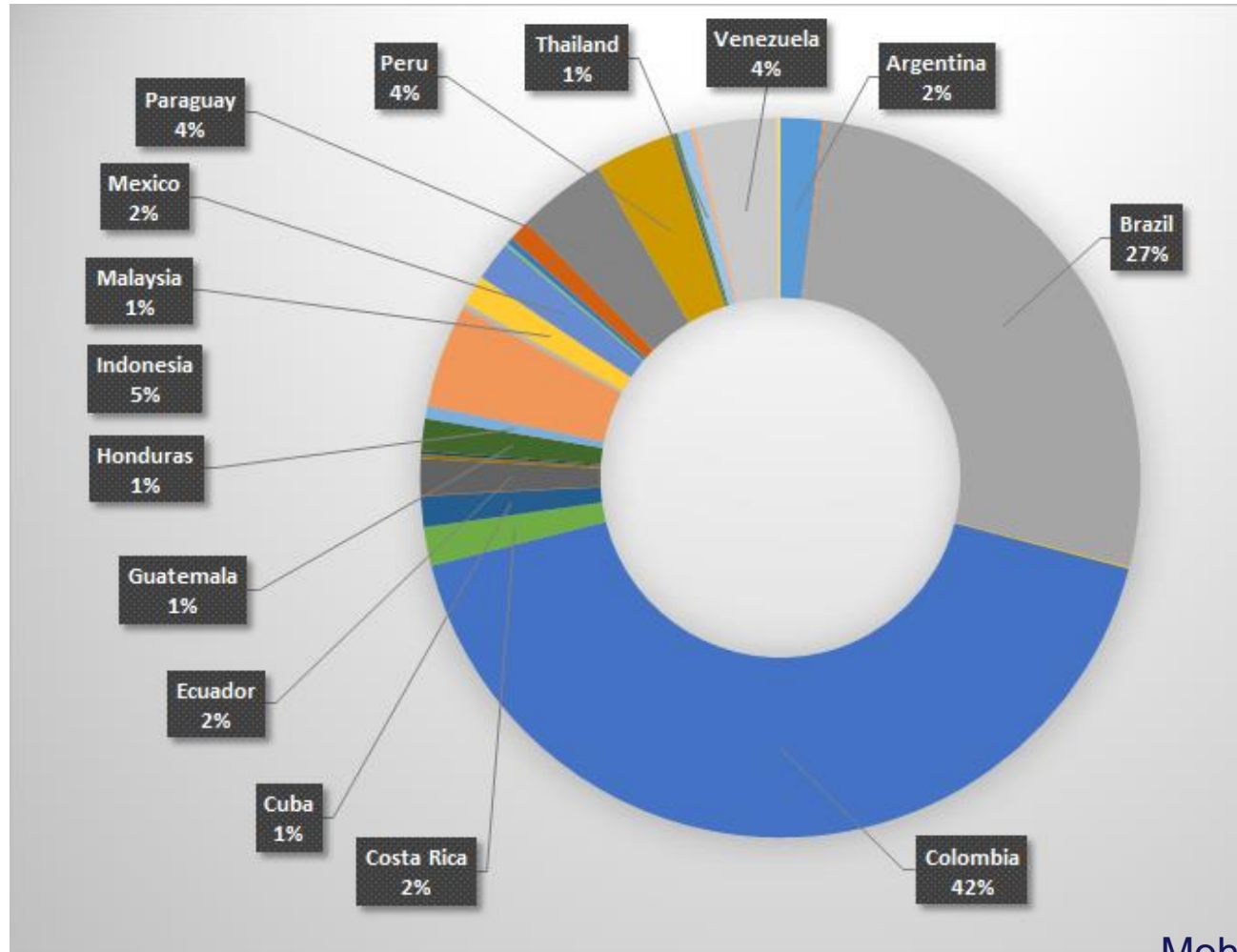
6,600 *in vitro* accessions of cassava and *Manihot* wild relatives:



Cassava Digital Genebank – Phase I

- **Genotype** 4,000 accessions
- **Diversity analysis** – understand collection, select accessions for WGS, identify duplicates, compare to core collection
- **WGS** – identify more variants, track genome segments back to progenitors
- **Database** : 

4,000 accessions by country of origin



DArTseq data

- Read files for **4,074** accessions (plus 903 replicates)
 - 178 replicated wells & 725 DArT technical replicates
- **75,548** raw SNP calls & **74,524** PAVs

Filtering:

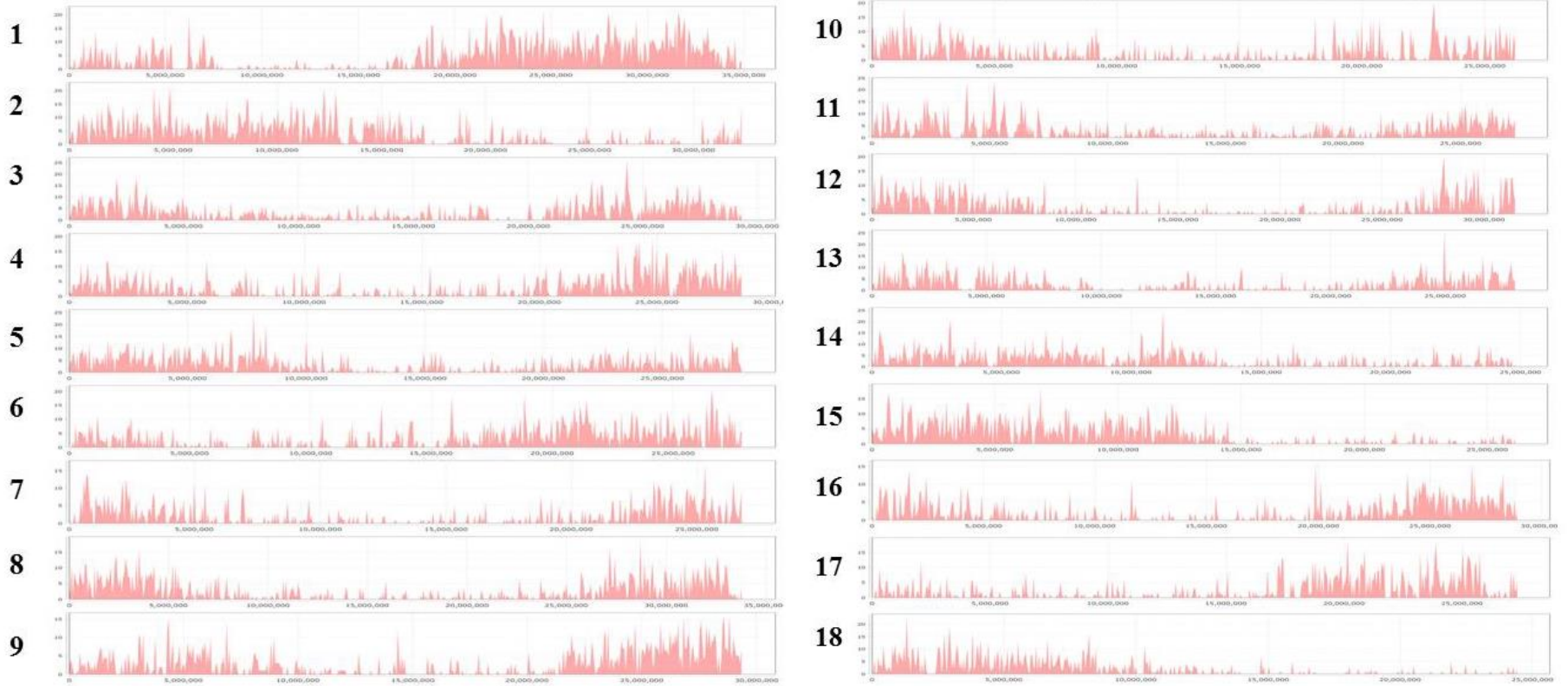
Reason	Samples removed
Failed replicates	22
Wild samples	51
>15% missing data	47

Reason	Loci removed
<80% call rate	11,824
<98% reproducibility	3,166
Monomorphic loci	27,850

4,835 samples (3,979 acc) and 32,708 SNP loci



Genomic distribution of alleles



94% of allele tags aligned to genome v6.1

Challenges

- Choice of technology / provider – DArT-Seq, GbS, skim-seq, WGS etc.
- Lack of integration between genebank and lab?
 - Who will perform the extractions? And any re-extractions?
- Handling large numbers of samples
 - Procedures for sample tracking/barcoding – automation?
- Data QC
- Data release

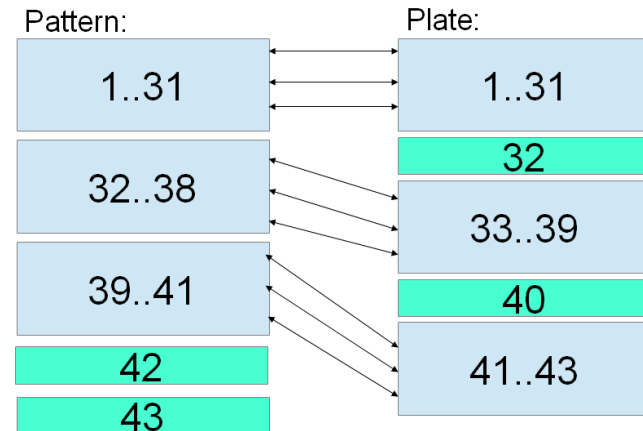
Data QC checklist (1)

- Is the data from your samples? Check species first (& contaminants)!
- Cross-check sample names with original samples
 - typos & handwritten labels e.g. COL8098 = COL809B
- Sample replicates
 - Same sample in different wells, do they match?
 - Same extraction or re-extraction?
- Does data agree with prior knowledge?

If replicates don't match...

Same sample in different wells, should look similar

- 178 replicates, distributed across 42/43 plates
 - Plates 1-31 looked fine
 - Plates 32-43 something wrong...
- Compared replicate patterns across all plates
 - identified plate re-ordering problem
 - Excel issue traced and resolved



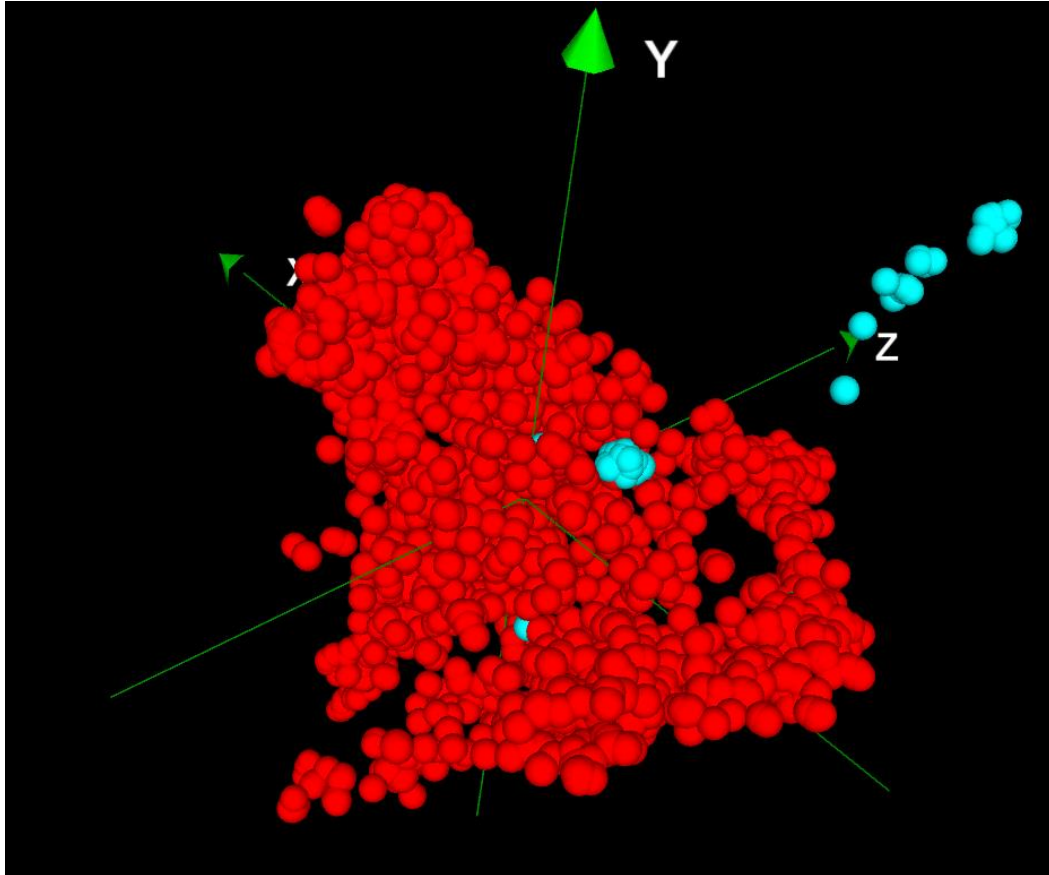
Plan plate layouts wisely...

- Replicates on n-1 plates (>1 if possible)
- Different positions per replica pair / plate
- Sample ordering in wells
 - easier to identify splash and mix-ups if neighbours are dissimilar
 - consider your pipetting methods for most likely sources of error



Does it match with prior knowledge?

- Principal Co-ordinate Analysis

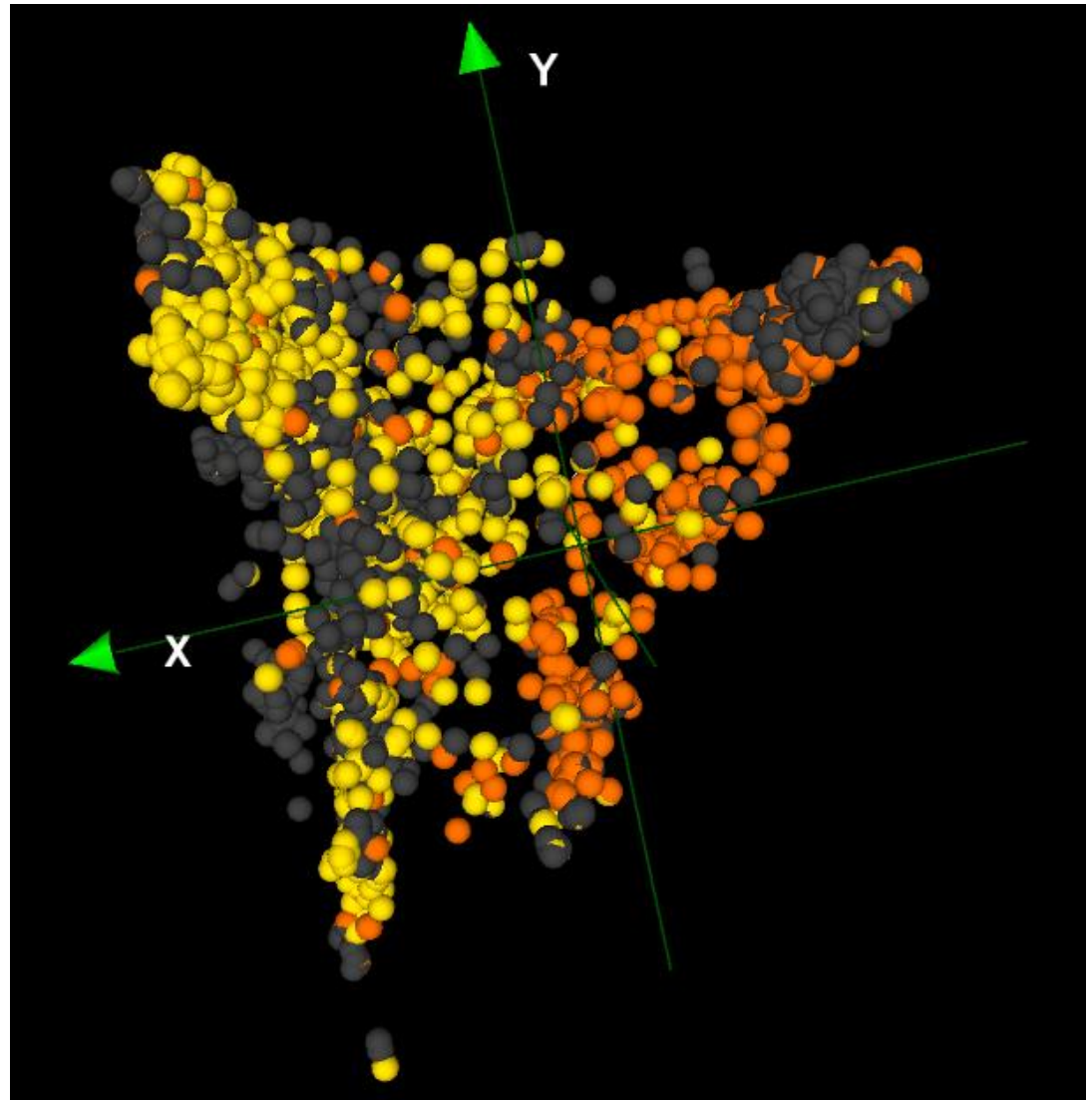


- Domesticated cassava
- Wild Manihot

Principal Coordinate Analysis - Geography

● Colombia

● Brazil



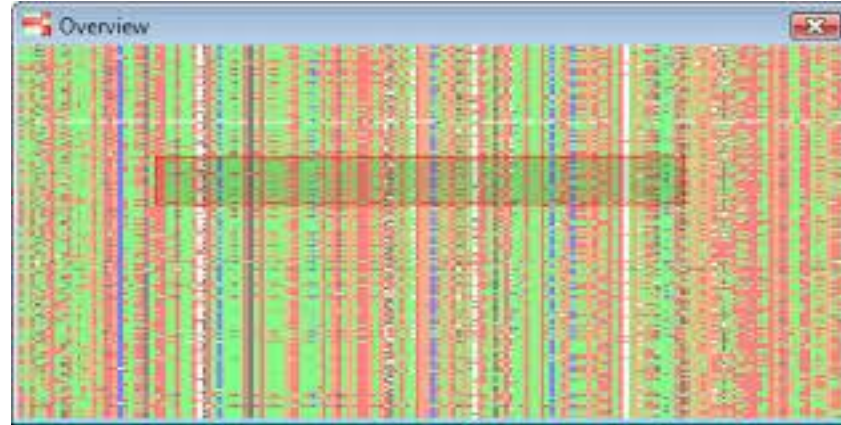
Data QC checklist (2)

- Batch effects – do you see similar trends across batches/plates/rows/columns?

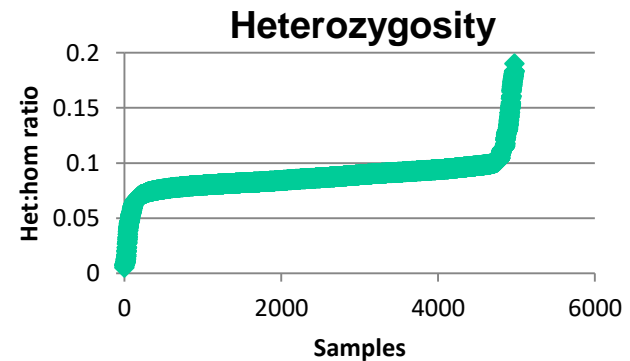
Flapjack:



- Check heterozygosity



- high could mean mixed sample - check neighbours
- low could mean data quality issue





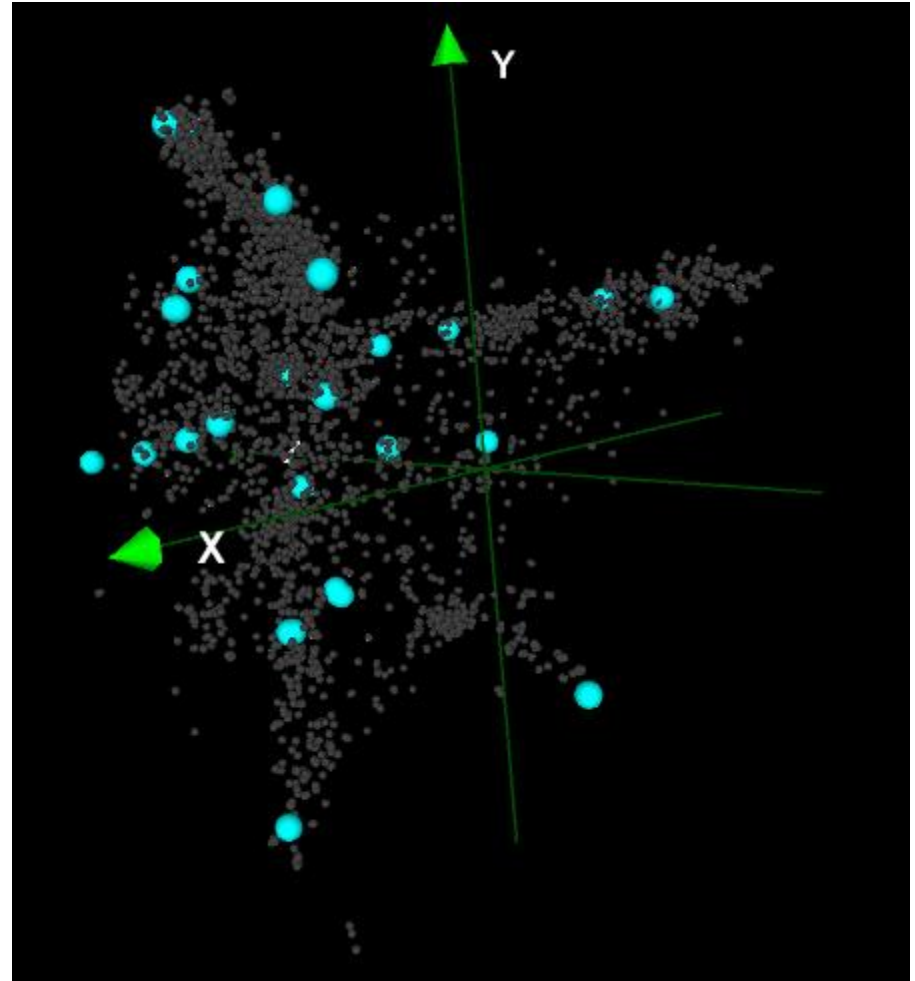
Further checks planned

- We are in the process of selecting samples to re-extract and re-genotype using the following criteria:
 - Failed replicates – testing consistency in extraction pipeline
 - High proportion of rare alleles
 - High and low heterozygosity
 - Random from all plates to check for consistency
 - Other peculiarities?


Selection of samples for WG sequencing

- DArT-seq passed QC
- Samples with known resistance to pests e.g. whitefly, thrips & green mite
- Parents of elite lines
- Frequently requested samples
- Genetically diverse

- | |
|--|
|  Selected |
|  Not selected |



Next steps

- Sequencing 25 samples for WGS
- Re-genotyping for sample QC
- Check DArT-seq genotype calls vs other SNP calling pipelines & public data
- Explore PAVs
- Check for duplicates within germplasm collection - cryopreservation
-  Germinate 3
- Complete the whole collection

GCRF: Developing a natural variation platform for pest-resistant cassava breeding

- **Phenotype** 100 cassava wild relatives (24 sp.) in response to: whitefly, cassava frogskin disease and bacterial blight
- **Resequence** the 100, calling variants, identifying novel regions and annotating
- **Software development** for genomic browsing and filtering tools for genebank users – Germinate 3, Cassava Genome Hub and CassavaBase



Acknowledgements



Bruno Santos
Mohamed
Abdelhalim
Pradeep Ruperao
Faraz Khan
Mario Caccamo



Peter Wenzl
Joe Tohme
Luis Augusto Becerra
Lopez-Lavalle
Anestis Gkanogiannis



David Marshall

