TGAC
The Genome Analysis Centre™

BBSRC Greater Norwich Development Partnership

GLOBAL CROP DIVERSITY TRUST
A FOUNDATION FOR FOOD SECURITY

Sequencing Genebank Collections: Challenges and Opportunities



Sarah Ayling
sarah.ayling@tgac.ac.uk

TGAC
The Genome Analysis Centre™

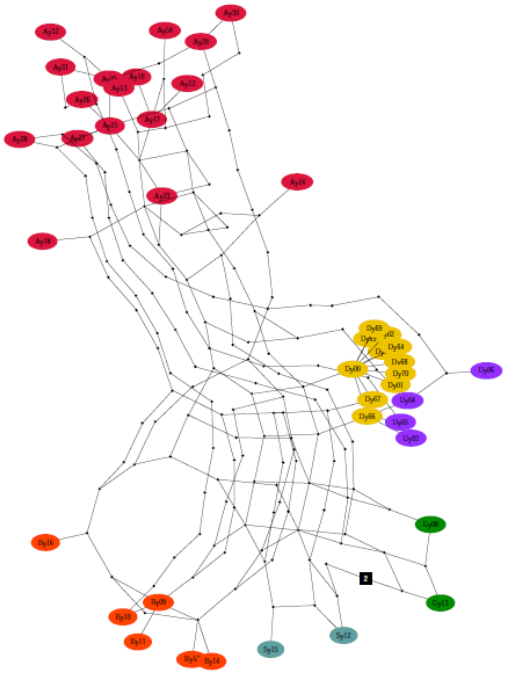
Benefits of sequencing a collection

- Genebank management
  - Regeneration of accessions
    - Check for mix-ups
    - Intra-accession diversity maintained
  - Redundancy
    - Split / merge / archive accessions
    - Add that new accession?

Photo: CGIAR/IRRI

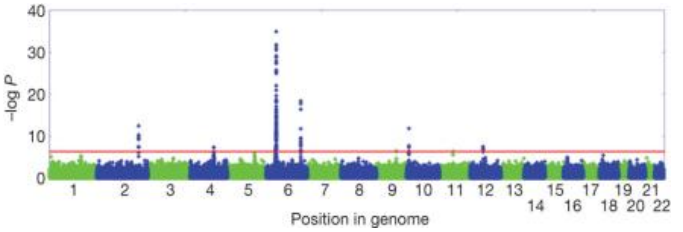
Benefits of sequencing a collection

- Development of marker sets for QTL mapping
- Diversity studies
  - Inclusion of genetic diversity into breeding programs
  - Origins of domestication and evolutionary studies
  - Ease of crossing



GWAS and GS

- Genome wide association studies (GWAS)
  - Detailed phenotyping and genotyping of unstructured population
  - Identify markers associated with phenotype



- Genomic Selection (GS)
  - Detailed phenotyping and genotyping of training set
  - Genotype a related set
  - Estimate breeding values based on genotype

TGAC
The Genome Analysis Centre™

Phenotyping challenges

- Intra-accession variability
  - SSD prior to genotyping and phenotyping
  - Increase size of collections or discard?
- Phenotyping crop wild relatives
  - Unfavourable genetic backgrounds mask allele effects
  - Multi-location phenotyping requires widely adapted materials
  - Wide crosses into widely adapted elite materials?

TGAC
The Genome Analysis Centre™

Genotyping many collections?

- Challenge: identify a single cost-effective strategy suitable for all crops listed in Annexe I of the International Treaty on Plant Genetic Resources for Food and Agriculture
- Different crop genomes have different features

Images from Wikipedia

Genome features affect sequencing strategies

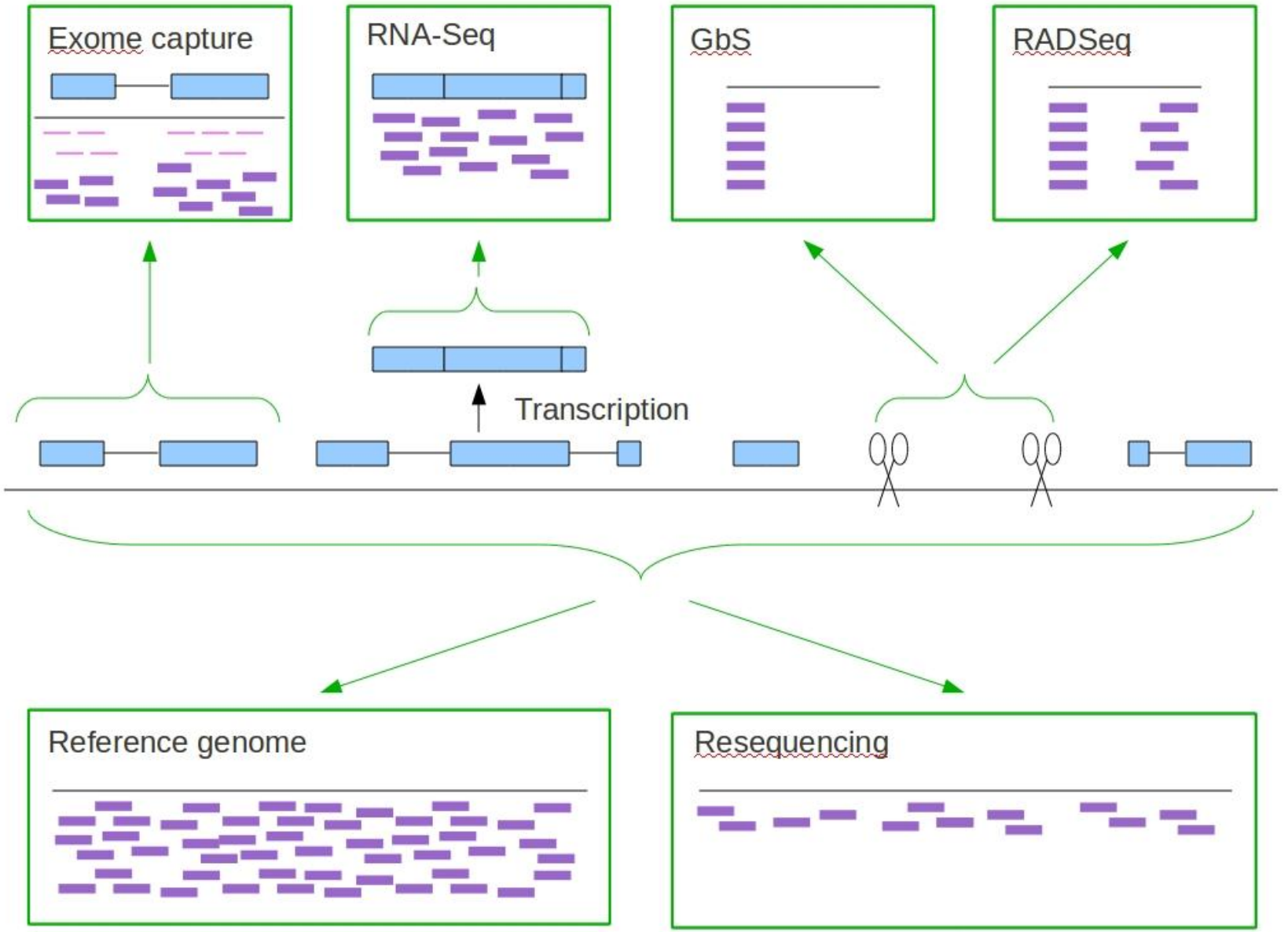
- Size and repeats
  - Large genomes require more sequencing
  - Repeats confound assemblies
- Polyploidy and heterozygosity
  - Homozygous 6x, heterozygous 13.5x (P>= 0.9975)
  - Assemblies may be chimeric
- Linkage Disequilibrium (LD)

Wheat, inbreeder, high LD, require fewer markers

Maize, outcrosser, low LD, require more markers

TGAC
The Genome Analysis Centre™

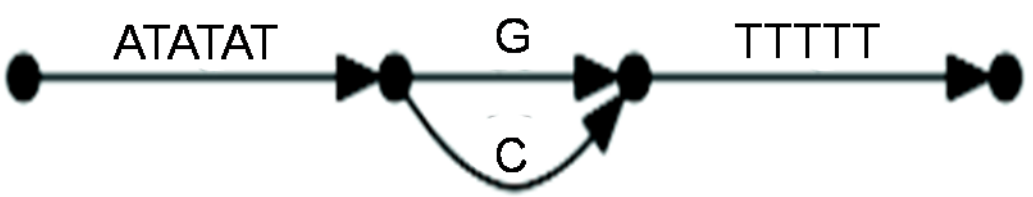
Sequencing strategies



TGAC
The Genome Analysis Centre™

Do we need a reference genome?

- Enables exploration of regions near markers
- GWAS needs ordered markers
  - Genetic map
  - Synteny-based
- Flexible non-linear reference
  - FASTG (Jaffe, MacCallum, Rokhsar & Schatz)

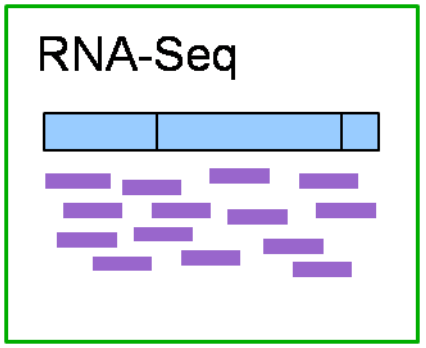


>xxx;
ATATAT G[1:alt|G,C] TTTTT

TGAC
The Genome Analysis Centre™

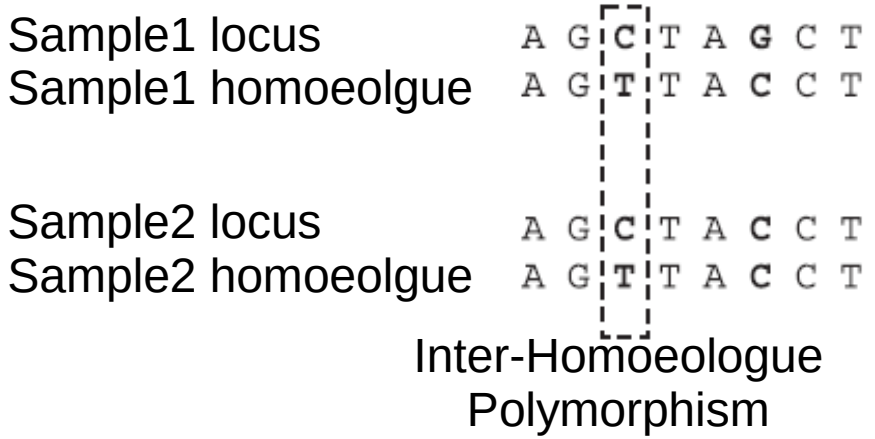
Stacks/RADtools

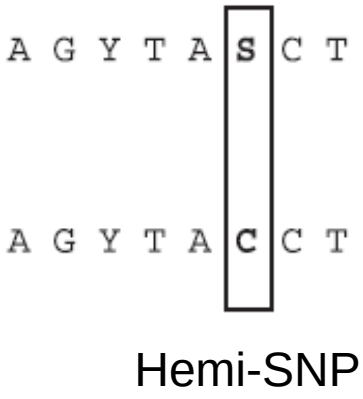
- De novo or reference guided
- Identify loci based on stacks
- Generates markers for genetic mapping
- Stacks designed for diploids
- RADtools has also been used for allopolyploids
  - e.g. *B. napus*: Bus *et al*. BMC Genomics 2012, 13:281



Associative transcriptomics

- *B. napus*: Harper et al., Nature Biotech. 2012, 30: 78–802
- Deep sequencing of juvenile leaf tissue
- De novo assembly of RNAseq data into unigenes
- Order unigenes based on linkage mapping and synteny with related species
- GWAS with SNPs and/or expression markers





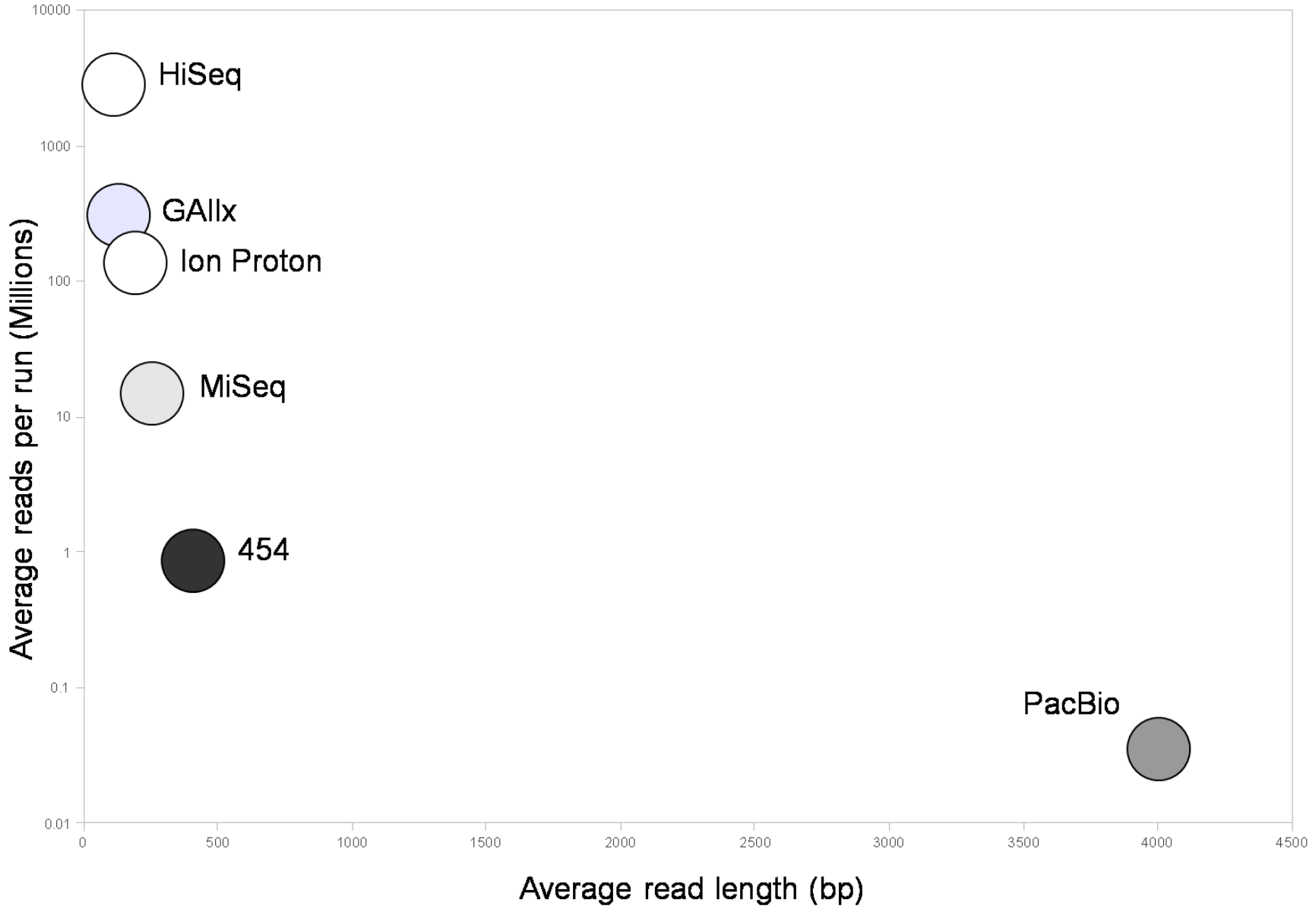
Genotyping-by-Sequencing

TGAC
The Genome Analysis Centre™

- Barley and wheat: Poland *et al*. PloS One 2012, 7:e32253
- Two enzyme GbS protocol
- Bi-parental populations of double haploid lines
- Bi-allelic SNPs identified and filtered based on
  - Frequency in population >20%
  - Expect allelic SNPs to be mutually exclusive
- Added markers to reference genetic maps
  - 34K to barley and 20K to wheat

TGAC
The Genome Analysis Centre™

Sequencing technologies



TGAC
The Genome Analysis Centre™

Novel sequencing technologies

- Single molecule sequencers coming…
- General trends:
  - Longer read lengths
  - Faster run times
  - Cheaper per base
- Should we wait longer?

TGAC
The Genome Analysis Centre™

CGIAR projects underway

- Rice resequencing
  - IRRI
- Maize and wheat GbS
  - CIMMYT
- Cassava RADSeq
  - CIAT, IITA and EMBRAPA
- Chickpea and pigeonpea reference set resequencing
  - ICRISAT






TGAC
The Genome Analysis Centre™

Data standards and documentation

- Sequencing metadata
- Variant call format – VCF, SHORE
- Detailed documentation of analyses
- Phenotyping protocols and metadata
- Crop and trait ontologies
- Links to accessible seed

TGAC
The Genome Analysis Centre™

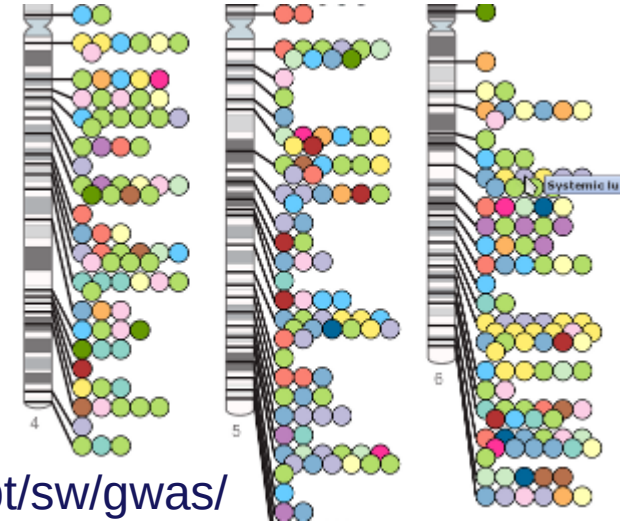
Tools and Resources

- Bioinformatics and compute capacity limited
- Data sets are Tb+
- Central location for analysis
  - e.g. Cloud, iPlant, etc
- Best practices for analysis are still evolving…
- Galaxy
  - Share workflows
  - Histories

Data visualisation and access

- Query data by accession, trait, gene, SNP etc.
- Integrate large data sets and visualisation tools
  - Polyploid genome browser
  - Visual summaries of data for thousands of accessions
  - Links to germplasm and pedigree/collection
  - QTL and associations,

    gene annotation etc
  - Genome view to locus view



http://www.ebi.ac.uk/fgpt/sw/gwas/

TGAC
The Genome Analysis Centre™

Developing a standard interface

- Reusable for different crops / genebanks
- Easier for users of >1 genebank
- Extendable interface, community development
- Standardised data formats will enable mirroring of data between centres
  - e.g. Ensembl, UCSC and NCBI

TGAC
The Genome Analysis Centre™

Community involvement

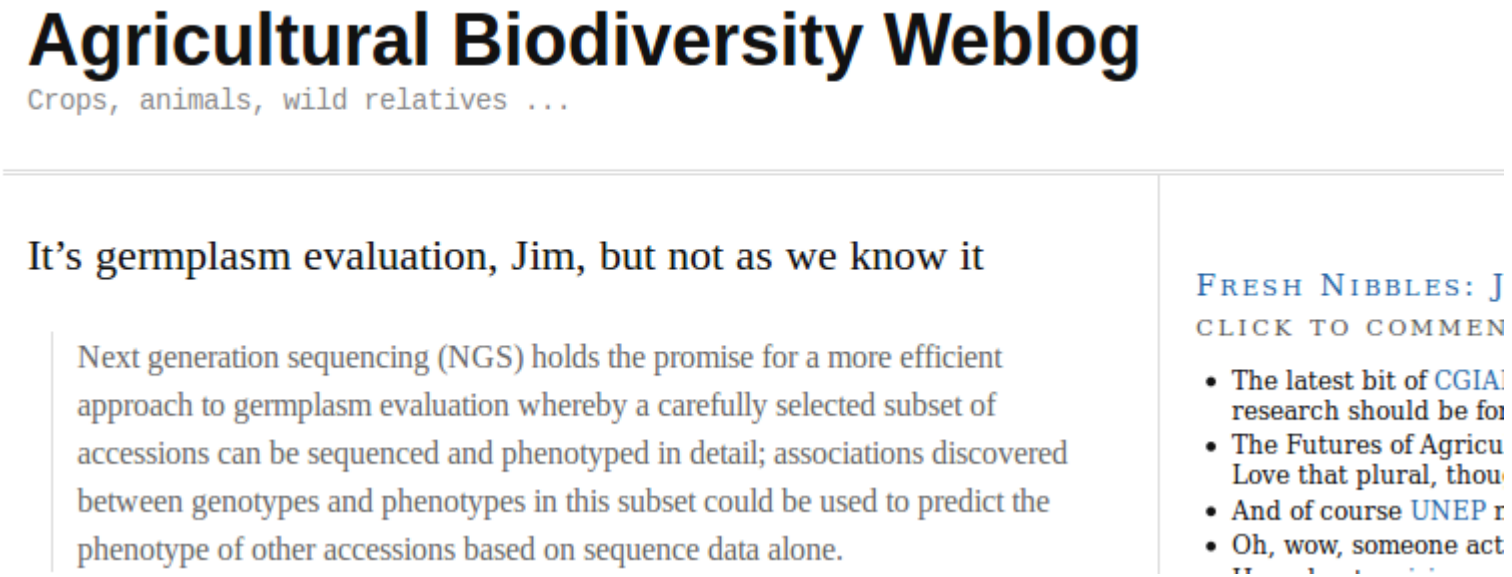
- User requirements
- Training workshops
- Community annotation fed back in
- Open access to data and tools
- Accelerate genebank research

Add your thoughts

- Initial report can be found here:

http://agro.biodiver.se/its-germplasm-evaluation-jim-but-not-as-we-know-it/

- Strategy for pilot project – ideas and feedback welcome



Agricultural Biodiversity Weblog

Crops, animals, wild relatives ...

It's germplasm evaluation, Jim, but not as we know it

Next generation sequencing (NGS) holds the promise for a more efficient approach to germplasm evaluation whereby a carefully selected subset of accessions can be sequenced and phenotyped in detail; associations discovered between genotypes and phenotypes in this subset could be used to predict the phenotype of other accessions based on sequence data alone.

FRESH NIBBLES: J

CLICK TO COMMEN

- The latest bit of CGIA research should be fo
- The Futures of Agricu Love that plural, thou
- And of course UNEP r
- Oh, wow, someone act

TGAC
The Genome Analysis Centre™

Acknowledgements

CIAT: Steve Beebe, Daniel Debouck, Cecile Grenier, Mathias Lorieux, Joe Tohme

CIMMYT: Peter Wenzl

Cornell: Ed Buckler, Jean-Luc Jannink, Susan McCouch

Generation Challenge Program: Jean-Marcel Ribaut

Global Crop Diversity Trust: Hannes Dempewolf, Luigi Guarino, Geoff Hawtin

ICRISAT: Rajeev Varshney

IDna Genetics: Peter Isaac

IRRI: Hei Leung, Ramil Mauleon, Ken McNally, Ruaraidh Sackville Hamilton

JIC: Mike Ambrose, Ian Bancroft, Cristobal Uauy

TGAC: Mario Caccamo, Donatien Chedom-Fotso, Matt Clark, Nizar Drou, Rocio Enriquez Gasca, Darren Heavens, Kirsten McLay, David Swarbreck, Chris Watkins

USDA: Jinguo Hu, Chris Richards

GLOBAL CROP DIVERSITY TRUST
A FOUNDATION FOR FOOD SECURITY

BBSRC
bioscience for the future