TGAC
The Genome Analysis Centre™

BBSRC EEDA Greater Norwich Development Partnership

Technical appraisal of strategic approaches to large-scale germplasm evaluation

Sarah C. Ayling

The Genome Analysis Centre,
Norwich Research Park, Norwich NR4 7UH, UK

Executive Summary

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. This report provides an overview of current sequencing technologies and strategies, applications of derived sequence data, and recommendations for maximum impact of NGS for genebanks. In addition four CGIAR projects which are already under development are described: IRRI's rice resequencing; CIMMYT's maize and wheat genotyping-by-sequencing; CIAT's cassava RAD-Sequencing; and ICRISAT's plans to resequence reference collections for chickpea and pigeonpea.

The Global Crop Diversity Trust has an interest in diverse crop species, from those with well-established genomic resources such as rice and maize, to orphan crops with little available data. The crops also have varying genome sizes, ploidy levels, and mating systems, all of which has an impact on sequencing strategy. One aim of this report is to explore to which degree it would be appropriate to develop a single strategy for the whole range of crop types of interest.

Uses of genomic data: The availability of genomic variation data (i.e. large numbers of markers) can enable association studies (**GWAS**), whereby genomic regions are associated with observed phenotypic data for simple traits. Where a reference genome is available these regions can be explored for genes relating to the observed phenotype. Markers within these regions can also be used to perform marker assisted selection within breeding programmes to detect individuals likely to contain the region of interest, reducing the need for slow/costly phenotypic evaluations. These large sets of markers may also be used to predict breeding values of individuals based on their sequence alone (genomic selection, **GS**), although this has not yet been applied to genebank materials to date.

GWAS and GS can be used to make an association between genotypic data and phenotype, however only when alleles have an observable effect on phenotype in the studied materials. Unfavourable genetic backgrounds can mask the effect of alleles on phenotype (epistatic effects), which can prevent the discovery of interesting alleles. Crosses to introduce alleles into more favourable backgrounds may expose novel associations with phenotype, but this approach is not scalable for testing all alleles.

Access to genomic information for genebank accessions can also aid **genebank management practices**. Accessions may be split, merged or archived based on genomic similarity, and potential accessions screened for allelic novelty. Mislabelled or misidentified accessions can also be identified, and accessions can be monitored during regeneration to ensure that the new seed resembles the original accession.

Publicly available **databases** which contain genotypic and phenotypic information will be key for transforming data into valuable knowledge. Phenotyping protocols described using structured vocabularies will facilitate data sharing. Variation data should be queryable from multiple entry points including marker/gene locations, traits of interest and phylogeny/pedigree data.

Genotyping approach: Currently, sequencing reduced representation libraries (**RRL**: a restriction digested reproducible fraction of the genome) or the transcriptome (**RNA-Seq**) offer the most cost-effective opportunities for large-scale genotyping of collections for species, and do not require reference genomes. Illumina is the most widely used and cost-effective technology (calculated by nucleotides per dollar).

- For RRL, the enzyme choice dictates how many loci are generated, and can be targeted to genic regions by using methylation-sensitive enzymes. Causative variants may not be included in the sequenced regions but linked SNPs can indicate genomic regions of interest and provide useful markers for breeding.

- RNA-Seq can identify large numbers of genic markers and should contain coding causative variants. RNA-Seq is less reproducible due to changes in expression levels. Target sequence regions are usually larger than in RRL, making this a costlier approach.

In the absence of reference genomes, markers must be ordered before performing GWAS, this can be done using genetic mapping and/or synteny data from related species. Whole-genome resequencing is still relatively expensive, particularly for large genomes, and assembly/mapping of repeat regions remains challenging. Neither RRL or RNA-Seq data would be re-used if whole-genome resequencing was undertaken in the future, however RNA-Seq data would remain a useful resource for genome annotation.

Accession heterogeneity: A number of seeds per accession can be pooled and genotyped to identify within-accession diversity. This information can be used to check for mistakes or genetic drift after seed regeneration.

Genetically identical seed should be used for both genotyping and phenotyping. This may require creation of a novel accession from a single seed which can be discarded once phenotyping is completed. Collections with little to no within-accession variation may be able to reuse existing phenotypic evaluation data.

Data standards: Standardised information on genotyping and phenotyping should be recorded and made publicly available. Ontology terms should be used for descriptors, being developed if necessary. A coordinated network of phenotyping sites would be advantageous for establishing data standards.

Variant calling: A Galaxy instance (or similar) containing workflows for bioinformatics analyses which can be run remotely by non-experts using a graphical interface could be established. Workflows for sequence alignment and variant calling can be shared between users, a history function can record all analyses run by each user. The instance could run on the cloud or a high-performance compute cluster to be accessed remotely via internet.

Data access and visualisation: Data should be made publicly available as soon as possible to promote use. Standard data formats should be adopted. Genome browsers can display variation data in the genomic context, on an annotated reference genome or pseudomolecules based on synteny information. Deploying a lightweight interface with a genome browser, associated variation and phenotypic data and links to accession information and ordering would allow users to mine genebanks for accessions of interest.

The Trust could play a role in supporting a dialogue between the CG centres already involved in large-scale genotyping projects, and potentially invite external centres in an advisory capacity, in order to design a single system which could be rapidly deployed to all centres, avoiding duplication of effort and development of incompatible tools. As each centre has a different sequencing approach, pre-processing of the data will vary, but the end result (variant calls) can be stored and displayed in the same way.

Pilot approach: CIAT's cassava resequencing project could be adopted as a pilot to study the impact of RRL sequencing on an entire collection. The project would be a collaboration between multi-crop genebanks (CIAT, IITA, EMBRAPA), on a relatively small collection (~6000 accessions) which is clonally propagated (avoiding within-accession heterogeneity). The largest threats to cassava from climate change are predicted to be pests and disease, which are easier to phenotype in wild relatives compared to traits such as yield. The three centres collaborating to choose traits of interest, with coordinating phenotyping activities, could produce high-impact publications to raise awareness within 3 years. Community-specific meetings could also be held to promote the use of the resource.

Conclusion: Sequencing technologies continue to improve, and there is an argument for waiting for longer reads and cheaper sequencing before attempting to sequence genebanks. However, food security is an urgent issue, and a great many marker-phenotype associations have been discovered for human disease using today's technologies. In addition, several CG centres are already embarking on whole-genebank resequencing/genotyping, and without input from the Trust it is likely that this will result in a number of independent resources which will be hard to consolidate in the future. The cassava pilot project will require a small investment, but should be a good model to test the impact of NGS on germplasm use, and identify problems before rolling out tools/protocols across all crops.

Contents

1: Introduction

With the world population reaching 7 billion in 2011 (UN, 2012), there is an urgent need to produce more food, with fewer inputs, such as water and fertiliser, under more variable/extreme climatic conditions. Whilst agronomic management practices can have a huge impact on the efficiency of crop production, improvement of genetic material may contribute at least equally (Mayes *et al.*, 2012). The current genetic base of most crops is narrow, as these species have recently (usually within the last 10,000 years) passed through the **domestication bottleneck**, and usually small numbers of individuals have contributed to modern breeding programs. A much broader range of genetic diversity can be found within landraces and crop wild relatives, which could be incorporated into breeding programs to potentially improve traits of interest and reduce susceptibility to both biotic and abiotic stresses.

The world's genebanks contain >7 million accessions of plant germplasm held within >1700 collections worldwide (FAO, 2010). The international genebanks managed by the Consultative Group on International Agricultural Research (CGIAR) contain collections for a number of crop species important for food security (ITPGRFA, 2009). These accessions are made freely available upon request to breeders and researchers throughout the international community.

With changing climatic conditions, there is an urgency to identify breeding material which can contribute traits of interest such as tolerance to cold, drought, heat, salinity, pests and pathogens, quality and yield traits. However often little is known about a genebank accession and the potential beneficial alleles it may contain. The CGIAR is attempting to standardise passport and characterization data for its materials through the GENESYS system (GENESYS, 2011), but the basic information stored per accession is limited.

Within the CGIAR, most genebanks have defined a “core collection”, typically designed to be ~10% of the total collection which aims to represent a high proportion (typically 80%) of the diversity of the full collection. Characterization and phenotyping studies are often performed on this representative set, or a subset called a mini-core (typically 20% the size of a core collection). Many papers have been written on how to construct a core collection (e.g. Hodgkin *et al.,* 1995, Grenier *et al.,* 2000, Upadhyaya *et al.,* 2001, Glaszmann *et al.,* 2010). These core collections may be defined based on niche, collection or characterization data (phenotypic or more recently molecular). DNA sequence data is favoured as it describes the inherited genetic material directly, although **epigenetic** modifications (epialleles) have also been shown to be inherited and to have an impact on phenotype but are not detected by conventional sequencing approaches (e.g. Tsaftaris *et al.,* 2005, Manning *et al.,* 2006, Becker *et al.,* 2011, Shivaprasad *et al.,* 2012). However, there are to date relatively few examples of DNA sequence evaluations across an entire collection (van Hintum, 2003, Jing *et al*., 2010, Kwon *et al*., 2012).

Until recently, genome-scale DNA sequencing projects were the privilege of well-funded international consortia, and a eukaryotic genome could cost millions of dollars to produce over a period of many years (The Arabidopsis Genome Initiative, 2000, International Human Genome Sequencing Consortium, 2001). With the advent of Next Generation Sequencing (NGS) technologies, the price of sequencing has dropped dramatically (Figure 1), whilst the speed has increased considerably. It is now possible to generate 30 Gigabases of data (~7x the human genome) in a single lane of the Illumina HiSeq machine, in 11 days, for ~$2000. However, each NGS technology has its short-comings, with varying error rates and read lengths, with no single current technology being able to produce a high quality, contiguous genome sequence in isolation.

Next generation sequencing holds the promise for a more efficient, strategic 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. This approach would dramatically reduce the amount of phenotyping required, an expensive process in both time and resources, whilst providing genomic data and predicted phenotypes for all remaining accessions. The additional information generated for all accessions would aid genebank managers when selecting materials of interest to breeders, increasing the use of germplasm collections (Kilian and Graner, 2012, McCouch *et al.,* 2012). This report explores the current feasibility of sequencing entire collections and the sequencing strategies which could be adopted to generate sufficient data.

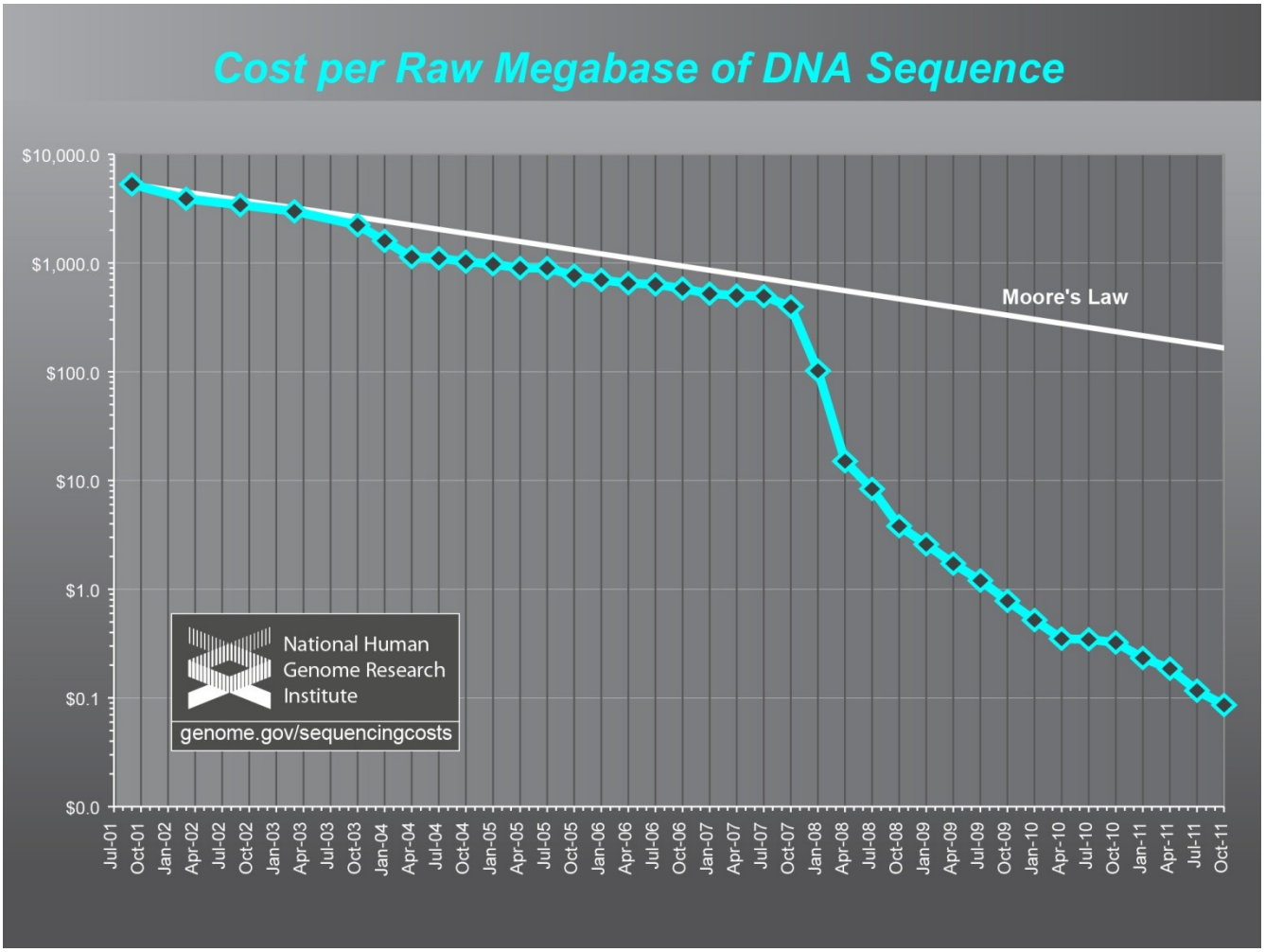

Figure 1. The cost per raw megabase of DNA sequence. Since the introduction of next generation sequencing in 2008, the average cost per megabase (Mb) of DNA produced by the National Health Genome Research Institute has dropped dramatically (NHGRI, 2012). Time is shown on the x-axis, average cost per Mb is shown on the y-axis with a logarithmic scale. Moore's law describes the doubling of compute power due to hardware improvements observed every two years, and is provided for comparison.

Section 2.1 will discuss the different sequencing technologies currently or soon to be available, their

relative advantages and disadvantages and associated costs. In order to reduce costs, it is common to sequence a subset rather than the complete genome; the information generated with respect to cost will be discussed in **Section 2.2.**

Section 3 will outline some details as to what can be achieved with this genetic data, both in terms of crop improvement and the potential benefits to genebank management. **Section 4** will detail some current NGS projects happening within the CGIAR and related projects, and finally **Section 5** will describe some recommendations for the medium term. This report has been generated in light of discussions with a number of key scientists within the field, **Appendix I** lists meetings attended as part of this work and **Appendix II** lists the individuals who have contributed information and ideas.

2: Next Generation Sequencing

2.1 NGS Technologies

Currently, there is no sequencing technology capable of producing sequence **reads** the length of an entire eukaryotic chromosome. Indeed, the early NGS technologies were characterized by having very short read lengths, initially 35 basepairs (bp) for Solexa/Illumina, which has now increased to 150bp, with Roche 454 increasing from 100 to 450bp. The genome sizes of rice, maize and wheat are 400Mb, 2.5Gb, and 17Gb respectively. These short reads need to be **assembled** into longer **contigs**, representing sections of chromosomes for genomic data or **transcripts** for transcriptomic data, and contigs may then be built up into longer structures again using read-pair information. This assembly process is hampered by regions of the genome which consist of repeated sequences, and reads from such repeat regions are often unable to be assembled, resulting in fragmented assemblies. However, genic regions are usually of most interest to researchers and breeders, and these tend to be simpler for automated assembly algorithms to build. Information from physical mapped BAC libraries or genetic maps can also be used to help order and orient contigs into longer **scaffolds**.

The original human genome sequence was generated using Sanger sequencing, which has a low error rate (<1%) and routinely gives reads of 800bp in length, but is relatively expensive. The depth of sequencing was ~7.5x (International Human Genome Sequencing Consortium 2001), meaning that on average each base of the human genome was sequenced 7.5 times, to allow the identification and correction of sequencing errors and provide overlapping sequence with which to position the reads to generate the assembly. Sequencing depth affects coverage, as the number of reads produced from a template sequence can be approximated by a Poisson distribution (Lander and Waterman, 1988). Increasing sequencing depth will increase the likelihood that all positions of the target genome are covered (i.e. represented in the set of sequencing reads). For a haploid or homozygous genome, a minimum depth of 6x is required to ensure 99.75% of bases are sequenced (Wendl and Wilson, 2008).

With Ilumina, today's most popular short read technology, depths of at least 30x are recommended for ***de novo*** genome sequencing (Schatz *et al*., 2010), as a high read depth compensates for short read lengths in the assembly process. In addition, Schatz *et al.* recommend a further 10-20x of **long mate pairs**. These are pairs of reads generated from longer DNA fragments, typically in the range of 3-20kb, and require high molecular weight DNA for **library** construction. The longer fragment sizes result in pairs of reads which can span repeat regions, enabling the organisation of contigs into longer scaffolds.

For resequencing experiments, where a reference genome is already available to **align** the reads to, the sequencing depth can be much shallower, and is determined by the ploidy of the species, heterozygosity of the sample and desired coverage and confidence in any identified polymorphisms. Each polymorphism should be identified by a minimum of two independent reads to reduce the number of false positives caused by sequencing error. Wendl and Wilson (2008) predict that for a heterozygous diploid, a depth of 13.5x is required to detect both alleles at least once for 99.75% of positions. To detect each allele at least twice, a depth of 18x would be required. When highly similar samples are sequenced together, such as offspring from a bi-parental cross, the depth for each individual can be extremely low (e.g. 1x) as missing data may be inferred from siblings or related samples (Huang *et al.,* 2010).

The currently available sequencing technologies differ widely in the lengths and numbers of reads they produce, the error profiles of those reads and the costs of making and sequencing the DNA libraries. The most popular technologies are described below, and these are often used in combination to achieve the best results. The costs provided are based on operating costs at TGAC from March 2012,

unless otherwise specified; however, sequencing costs continue to drop (Figure 1), due to improved chemistries resulting in increased throughput and read lengths, as well as the introduction of novel technologies. As such, the reader should bear in mind that the prices quoted here will be rapidly superseded and figures are only provided for the purpose of comparison.

Illumina/Solexa

Originally developed by Solexa, but later purchased by Illumina, this is the cheapest technology currently available in terms of price per base pair. The Illumina Genome Analyzer IIx and HiSeq2000 are widely used, and can produce 95 and 600 Gb of data per 11-day run respectively. Both **single end** and **paired end** runs can be performed (where one or both ends of the DNA fragments are sequenced) and paired end gives a significant advantage when assembling the reads, as the paired reads should be correctly oriented relative to one another and within a certain distance representing the possible range of fragment sizes determined by the DNA library preparation. The error rate is <1%, and errors are more likely to occur at the 3' end of the reads. Samples are loaded into eight **lanes** within a **flow cell**, and the HiSeq2000 can run two flow cells simultaneously. Samples can have molecular barcodes added, so that samples can be pooled for sequencing, and separated out computationally at a later stage. Illumina provides 24 barcodes, but 384 barcode systems have also been designed (e.g. NuGEN, 2012). The Beijing Genomics Institute (BGI) uses Illumina almost exclusively for its sequencing, with 100 HiSeq2000 machines between the Shenzhen and Hong Kong sites.

Library preparation costs:
To produce an Illumina Barcoded DNA library costs ~$200 (~$250 for RNA)

Sequencing costs:

One lane of 100bp paired-end reads on HiSeq:	$1900	for ~150 million pairs of reads
One lane of 100bp single-end reads on HiSeq:	$1100	for ~150 million reads
One lane of 50bp single-end reads on HiSeq:	$750	for ~150 million reads

Bioinformatics:
Adapter trimming, quality filtering and **demultiplexing** of barcodes is performed routinely, with a single lane taking up to 24 hours on an 8 processor machine with 44Gb RAM. Assembly is typically performed using a **de Bruijn graph** approach (Velvet, Zerbino and Birney, 2008; ABySS, Simpson *et al.,* 2009; SOAPdenovo Li *et al.,* 2010) which generates draft assemblies that are fragmented, particularly in repeat regions. For large genomes these programs can be very memory intensive and require access to large memory machines (e.g.>250Gb RAM).

The Amazon EC2 is a popular web-service that provides **cloud** compute facilities where users pay for the capacity which they use. Cloud computing can be an attractive option for researchers who lack access to large compute facilities. EC2 has a number of different instances available (EC2, 2012), however the largest RAM instance is currently limited to 68.4Gb of RAM which may not be sufficient for assembling large genomes.

Roche 454

The 454 sequencing technology generates longer reads than Illumina (350-450bp), with shorter run times (~10 hours), but is more expensive in terms of cost per base produced, with characteristic errors associated with **homopolymer** runs (the length of single nucleotide repeats longer than five or six contiguous bases cannot be predicted with confidence). Non-homopolymer-associated error rates are

low (~1%). Longer reads are advantageous when performing *de novo* assemblies. One strategy is to use a combination of 454 and Illumina sequence, with longer reads from 454 improving assemblies, and Illumina reads correcting homopolymer errors (e.g. Celera assembler, 2012; MIRA, 2012). Samples are run on a single plate, but the plate can be divided into halves, quarters, eighths or sixteenths, however each division results in a loss of sequence due to plate area covered by the dividers.

Library preparation costs:
To produce a 454 DNA library costs ~$350 (~$500 for mRNA)

Sequencing costs:
One plate of 454 Titanium FLX sequencing: $7600 for up to 1 million reads
(Half a plate costs half as much to sequence)

Bioinformatics:
Typically, 454 data does not get pre-processed in the same way as Illumina data, and the output files from the sequencer are used directly. Assembly is typically performed using Newbler (Newbler, 2012) or WGS (Celera assembler, 2012) but 454 reads can be used with de Bruijn graph assemblers also. Due to higher cost and the benefit of increased read length, a lower sequencing depth is typically generated, however assemblies are usually less fragmented than those obtained with Illumina reads alone. As with Illumina data, large genomes will require machines with a large memory capacity (e.g.>250 Gb RAM).

ABI SOLiD

SOLiD sequencing differs from Illumina and 454 data in that sequence is read in '**colour space**' rather than 'base space', where triplets of nucleotides are encoded as colours. This approach enables the detection of sequencing errors, resulting in very low error rates, however few downstream bioinformatics tools can work in colour space and as such SOLiD is less popular than Illumina and 454. Read lengths are currently 75bp.

Personal Genome machines

Ion Torrent and Illumina have recently released small low-throughput bench-top sequencers: the Ion Torrent PGM and MiSeq.

Ion torrent PGM

The Ion Torrent Personal Genome Machine was released in Dec 2010, and has three different chips available (314, 316 and 318) which can generate 10Mb, 100Mb and 1Gb of data respectively. The instrument run time is short (~2 hours) and read lengths range from 35-400bp. Error rate is ~1%, and the error profile is similar to 454 with problems accurately determining lengths of homopolymer runs (Glenn, 2011). Costs for Ion Torrent taken from Glenn, 2011:

Library-preparation costs:
Reagent costs per Ion Torrent '318 chip' run ~$925 in May 2011

Sequencing costs:
One run of Ion Torrent '318 chip' sequencing: ~$1200 for 4-8 million reads in May 2011

Illumina MiSeq Personal Sequencer

The MiSeq was released in 2011, and can produce 2Gb of data. Unlike the HiSeq2000, the MiSeq has a single lane on a single flow cell. The instrument run-time is 27 hours but has a simpler library preparation step when compared with Ion Torrent. Paired end reads of 150bp or single end reads of 300bp are produced, with an error rate of 0.1%. A 5 hour run is also available producing paired end reads of 25bp or single end reads of 50bp. An upgrade was announced in October 2011 that will give read lengths of 250bp, and 15 million paired end reads per run.

Library-preparation costs:
To produce an Illumina Barcoded DNA library costs ~$200 (~$250 for RNA)

Sequencing costs:
One run of 300bp MiSeq sequencing: ~$1350 for 5 million reads
One run of 50bp MiSeq sequencing: ~$900 for 5 million reads

Single molecule sequencers

PacBio RS

Pacific Bioscience's (PacBio) RS machine is a single molecule sequencer which operates in real time. The machine produces reads averaging more than 3kb in length, with 5% in the 5-10kb range and produces ~40,000 reads per 1.5 hour run. The error rate is high (~15%) but the majority of these errors (~11%) are random insertions. Initially, strobed reads were available to generate patches of sequences along a single long molecule, which could be used for scaffolding shorter reads together, however this functionality is no longer supported. An approach has been developed (PacBioToCA, 2012) which uses short Illumina reads to 'correct' the PacBio errors and then assemble the now long and accurate reads using the Celera Assembler V7.0 (Celera Assembler, 2012), which can use as input reads up to 32kb in length.

Library-preparation costs:
Reagent costs per PacBio SMRT cell ~$350

Sequencing costs:
One PacBio SMRT cell sequencing: ~$200 for 40,000 reads

Future technologies

January 2012 saw the announcement of two new sequencing machines from Life Technologies (Ion Torrent Proton II) and Illumina (HiSeq 2500), with both claiming to produce a human genome within a day.

The Ion Torrent Proton II is predicted to be able to sequence a human genome for $1000 in a day by the end of 2012. No details were given of the depth of coverage for this genome, but estimates suggest 10Gb of data, which would be 3x (SeqAnswers, 2012).

The HiSeq 2500 will have two modes, one to generate 600 million read-pairs per run in 27 hours (40x coverage of the human genome), and the other to generate 3 billion read pairs in 11 days, equivalent to the HiSeq2000 (Illumina, 2012). The machines will be available in the second half of 2012, costs per lane are estimated to be $1500 (CoreGenomics, 2012).

In February 2012 at AGBT (Advances in Genome Biology and Technology)**,** Oxford Nanopore announced two new single molecule sequencers, MinIon and GridIon (Omics!, 2012). The MinIon is a disposable USB sequencer and can sequence 512 molecules at once (one molecule per nanopore), producing 120-500 bases per minute for 6 hours (50kb reads have been described so far). The GridIon stacks 2000 nanopores (available in the 2nd half of 2012, and stacks of 8000 will be available by 2013). Each GridIon sequences ~1.4Gb per hour for up to a few days per sample. The error rate is 4%, and errors are typically deletions, which Oxford Nanopore believe can be reduced with software improvements. However, no data has been released yet to determine how accurate these projections are. Sequencing costs are estimated to be comparable with current technologies. With reads of 50kb+, the depth of sequencing required for assembling a genome would drop dramatically.

Sequencing information for the current technologies is summarised in Table 1.

Table 1: Estimated costs and volumes of sequence produced per run

Sequencing technology	**Cost per run ($)**	**Gigabases per run**	**Cost per megabase ($)**
Illumina HiSeq PE*	33600	600	0.06
Roche 454	7950	0.4	19.88
Ion Torrent PGM '318' chip**	2125	1	2.13
Illumina MiSeq*	1550	2	0.78
PacBio RS	550	0.12	4.58

* 16 lanes per run
** Information from Glenn, 2011

2.2 Sequencing strategies

Aside from the sequencing technology employed, the target you choose to sequence has an impact on the cost and volume of data generated. There are a number of ways to use NGS technology to gain genetic information about an organism, often in the absence of a reference genome. These approaches can be complementary, and are outlined below, ranging from full genome sequencing to sequencing a subset of markers, with popular non-NGS technologies included for comparison. A schematic of the different approaches is given in Figure 2.

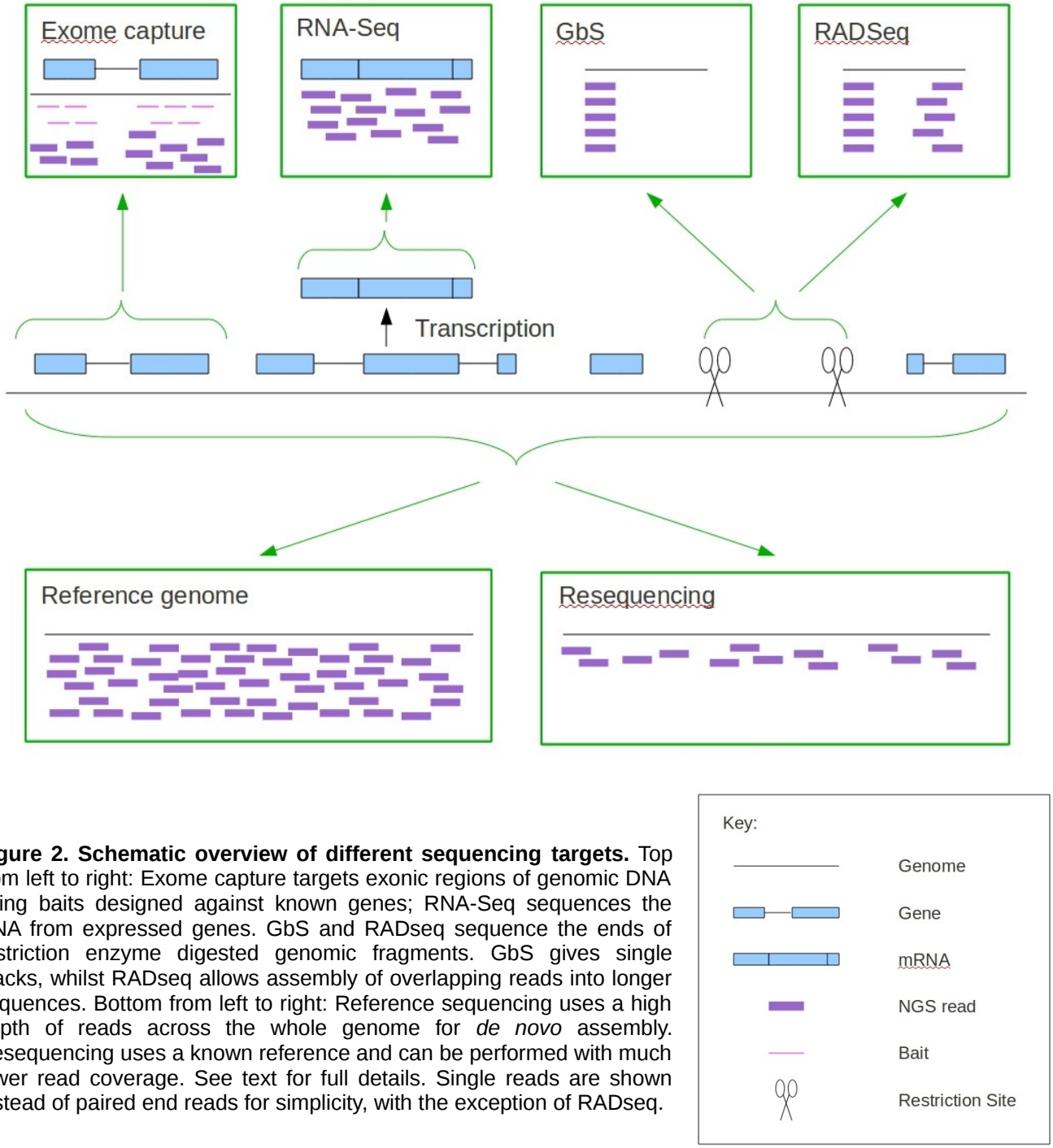

Figure 2. Schematic overview of different sequencing targets. Top from left to right: Exome capture targets exonic regions of genomic DNA using baits designed against known genes; RNA-Seq sequences the RNA from expressed genes. GbS and RADseq sequence the ends of restriction enzyme digested genomic fragments. GbS gives single stacks, whilst RADseq allows assembly of overlapping reads into longer sequences. Bottom from left to right: Reference sequencing uses a high depth of reads across the whole genome for *de novo* assembly. Resequencing uses a known reference and can be performed with much lower read coverage. See text for full details. Single reads are shown instead of paired end reads for simplicity, with the exception of RADseq.

Reference genome sequencing

Sequencing the genome of a species for the first time provides the opportunity to explore the full set of genes present within that species, the organisation of the genome, and comparisons of genomic regions shared with closely related species. However, producing a finished reference genome is expensive and time consuming. Repeat regions are difficult to assemble, and as such NGS genomes often remain highly fragmented. Technologies which produce longer reads or long mate pairs show promise to help overcome this issue, but these approaches are not yet used routinely.

To date, the best way to sequence a genome with NGS is to use a combination of technologies and library **insert sizes**. The most commonly used technologies are currently Illumina, 454 and Sanger sequencing, with a **whole genome shotgun** (WGS) approach to sequence random fragments of the genome which are assembled afterwards. For small genomes with low repeat content, this approach can be very successful, however assembling repeat regions is beyond the capabilities of current assembly tools and issues of polyploidy and heterozygosity may confound assemblers even further. For this reason, the International Wheat Genome Sequencing Consortium (IWGSC) took a 'divide and conquer' approach with the hexaploid wheat genome and are sequencing and assembling **flow-sorted chromosome arms** individually.

A finished genome sequence will reveal the exact genome for the sequenced individual, however epigenetic changes will not be revealed by standard sequencing approaches and epigenetics accounts for a number of important inherited traits (Tsaftaris *et al.,* 2005, Manning *et al.,* 2006, Shivaprasad *et al.,* 2012). In order to reveal these epigenetic modifications, specialised experiments must be performed (e.g. ChIP-seq, Robertson *et al.,* 2007; bisulphite sequencing, Darst *et al.,* 2010; Hi-C, van Berkum *et al,.* 2010). So whilst the genome sequence provides a lot of information, it does not reveal *all* genomic information which may have an impact on phenotype.

A single lane of Illumina HiSeq2000, producing 150 million 100bp paired-end reads, would give 30x coverage of a 1Gb genome for $2100 in 11 days.

Resequencing

If a reference genome is available for a species, or varietal group, low-depth resequencing can be performed relatively cheaply to discover **single nucleotide polymorphisms** (SNPs). Typically, 1x coverage of a number of homozygous diploid accessions may be generated (where only ~63% of each genome will be sequenced according to the Lander-Waterman model) and missing data may be imputed from genetically similar accessions (Huang *et al.,* 2010). If the individuals are not highly genetically similar, have undergone rearrangements or the objective is to discover rare variants, a greater sequencing depth is required to ensure that these differences are detected (cf. Section 4: IRRI rice resequencing). Required sequencing depth will be affected by the heterozygosity and ploidy of the samples. If no reference genome is available, it is advisable to generate a reference from a single individual (per varietal group) by sequencing at high depth.

In order to sequence several samples in a single lane, samples can be barcoded and multiplexed for sequencing, however this requires generation of a single library for each sample, which increases the cost per lane.

A single lane of Illumina HiSeq2000, producing 150 million 100bp paired-end reads, would give 1x coverage of 30 individuals with a 1Gb genome for $7900 ($1900 sequencing cost + 30x$200 library preparation), so ~$270 per sample.

RNA-seq

To detect expressed genes and to identify genic markers, total RNA can be extracted and sequenced. This has the advantage that only RNA from transcribed genes is sequenced which represents a fraction of the genomic sequence. Samples can be multiplexed and run in a single lane to reduce sequencing costs. This approach will identify fewer SNPs per Mb of sequence as coding regions are more highly conserved, however these markers are considered to be very valuable as they will be tightly linked to the gene in which they are located. Transcriptome sequencing for SNP discovery has been performed successfully in polyploid species oilseed rape (*Brassica napus)* and wheat (*Triticum aestivum),* generating tens of thousands of SNPs, sufficient for use in genome-wide association studies (GWAS) (I. Bancroft pers. comm.).

It is common to apply different treatments to individuals or to sample tissues from different growth/developmental stages to try to increase the number of genes expressed in order to sequence transcripts from more loci. However, the majority of expressed genes are housekeeping genes, and as such are expressed at all times, albeit at varying levels. For this reason, a tissue such as juvenile leaf will contain mRNA from the majority of expressed genes (I. Bancroft pers. comm.). The cost of experimental treatments and time required for sampling different growth stages should be taken into account when considering this option, as greater sequencing depth of a single tissue may reveal a greater diversity of transcripts.

The main difference when working with RNA instead of DNA is that transcript expression levels are highly variable, and high copy RNA will be sequenced much more frequently. The proportion of rRNA in a sample can be decreased by isolating mRNA using the polyA tail, or by ribo-depletion of rRNA. RNA samples are also less stable than DNA, and must be stored at -80 °C and transported on dry ice, as opposed to DNA which can be stored at -20 °C and transported at room temperature. Whilst mRNA normalization techniques can be used to reduce the frequencies of highly-expressed mRNA transcripts (Ekblom *et al.,* 2012), their application can sometimes have a negative impact on sequencing quality (I. Bancroft, pers. comm.). Unnormalized samples also have the advantage that differential expression values can be calculated, allowing the exploration of changes in expression levels between individuals or time points. For expression studies, the depth of sequencing is typically high (e.g. ~18 million reads per sample, Brown *et al.,* 2012) although the ENCODE project recommends 100-200 million reads (ENCODE, 2011). Biological replicates are essential to assign confidence to differential expression values.

A complicating factor when dealing with RNAseq data is **alternative splicing** (AS). AS has been predicted in ~95% of human multi-exon genes (Pan *et al.,* 2008), and 61% of Arabidopsis multi-exon genes (Marquez *et al.,* 2012). Expression of splice variants can confuse standard assemblers, but several pieces of assembly software have been released recently to deal with this problem; some use reference-guided alignments (e.g. Cufflinks, Trapnell *et al,.* 2008) whilst others assemble *de novo* (e.g. Trinity, Grabherr *et al.,* 2011; Trans-ABySS, Robertson *et al.,* 2012; Oases, Schulz *et al.,* 2012). For ease of assembly, 454's longer reads may be advantageous and have been used to exploit the pigeonpea transcriptome to identify Simple Sequence Repeat (SSR) markers (Dutta *et al.,* 2011), however Illumina data is cheaper to produce, and the additional depth of sequencing allows the discovery of a greater number of transcripts, although the assemblies produced are likely to be more fragmented (D. Swarbreck pers. comm.). Use of longer reads from sequencers such as the PacBio RS should improve assembly of AS transcripts even further.

Transcriptome samples can be barcoded and multiplexed for sequencing. For expression studies in a plant using the Illumina HiSeq2000, eight samples could be multiplexed per lane to give 19 million 100bp paired-end reads per sample at a cost of $490 per sample. The cost of generating RNA samples will vary depending on the species and experimental conditions and mRNA enrichment techniques used.

Target enrichment/Exome capture

Target enrichment is a process by which **baits** are designed to pull out sequence fragments of interest, which can then be sequenced. For well annotated genomes, it is possible to design baits which are tiled across the exons of transcripts, known as exome capture. These baits are used in pull-down assays to capture genomic DNA fragments of typically 200-600 base pairs. The ends of these fragments are then sequenced and aligned to a reference to allow discovery of variants in the context of genes. Several companies offer ready-made human exome sequencing platforms (e.g. Agilent, Illumina and Nimblegen) which were compared by Clark *et al.,* (2011). Per reaction, human-exome prices ranged from $400-$1000.

For species without a well annotated reference genome, cDNA evidence can be used when designing the baits, but without knowledge of the exon/intron structure baits are likely to fail where they span exon junctions; overlapping baits can address this problem. The resulting sequence reads can be aligned against the initial cDNA set, however reads which were generated from intronic sequence will not align, and as such coverage surrounding the exon junctions is likely to drop. Assembling the captured sequences can provide partial intronic sequences, allowing the generation of a new reference (composed of exons and flanking intronic sequence). Reads can then be remapped to the new reference to enable alignment beyond the exon junctions (R. Enriquez Gasca pers. comm.).

Exome capture is used with genomic DNA, and as such does not require varying experimental conditions. However, the bait design requires prior knowledge of transcript structures and will enrich for known targets. For species with fully-annotated exomes, baits can potentially be designed against all exonic regions, however when using exome capture on individuals that are genetically divergent from the individual for which the capture was designed, novel exons will not be enriched. Species with little genomic information will be reliant on available EST or cDNA data, which will most likely bias the capture towards a subset of possible genes. Combining exome capture with a lower-throughput sequencer such as MiSeq could be a cost-effective way to perform allele mining on highly-multiplexed samples for a select number of loci.

Reduced representation approaches

Several approaches have been developed to sequence fragments of the genome produced by enzyme digestion. The choice of enzyme in combination with the genome itself will dictate how many fragments are generated by the digestion, but unlike approaches utilising random **shearing**, the results will be reproducible. In addition, methylation sensitive enzymes may be employed which will cut in unmethylated regions typically associated with genes, to give an enrichment of genic fragments. A size selection step and/or PCR stage may be performed, generating fragments several hundred basepairs in length which are then sequenced (Davey *et al.,* 2011*)*. As the genome is digested prior to sequencing, reads from different fragments will not overlap and therefore cannot be used to produce a genome assembly. These approaches are most commonly used for SNP detection/discovery.

There are two main protocols in use for producing reduced representation libraries: RADseq (Davey and Blaxter, 2010) and Genotyping by Sequencing (GbS) (Elshire *et al.,* 2011, Poland *et al.,* 2012). Both approaches require digestion of genomic DNA followed by a PCR step and sequencing of the resulting fragments. In GbS, the fragments are sequenced at a single end, and can then be aligned to the reference genome and analysed using TASSEL (Bradbury *et al.*, 2007). For RADseq, fragments of size 300-700bp are selected, then sheared such that for a given digested fragment after sequencing, all sheared fragments produce one read at the same position, in a stack, but the other read varies in position. These variable-position reads can be assembled using an assembly algorithm such as Velvet (Zerbino *et al.,* 2008) or ABySS (Simpson *et al.,* 2009), to generate longer fragments of ~500bp.

Software to analyse RADseq data has been developed (Stacks, Catchen *et al.,* 2011; RADtools, Baxter *et al.,* 2011). Both GbS and RADseq can be performed when no reference sequence is available, however the assembled-end of the RADseq approach provides a larger region to design probes against for genotyping systems such as Illumina's GoldenGate (GoldenGate, 2012) and KBioscience's KASPar (KASP, 2012), which require 50bp either side of a SNP. A third method, DArTSeq, is available from DArT PL (A. Kilian, pers. comm.) which has adapted the DArT approach to use NGS sequencing in place of microarrays to detect presence/absence variations and SNPs (Sansaloni *et al*., 2011).

As a relatively small amount of the genome is sequenced with these approaches, samples are always barcoded and multiplexed. Multiplexing 384 GbS samples on a single HiSeq2000 lane costs $9 per sample (Bucklerlab, 2012), and would generate ~500,000 reads per sample. An enzyme which cuts frequently will result in low coverage of the sequenced tags, and large amounts of missing data per sample. For instance, using ApeKI in maize results in 40% of the 680,000 SNPs identified with GbS being observed per sample. However, as a large proportion of the maize genome is not shared between accessions (~23%) this is equivalent to ~52% of the observable SNPs being sequenced per accession for a 384-plex run (E. Buckler pers. comm.). For closely related samples, imputation can be employed to reduce missing data, however for highly diverse samples this approach may not be suitable as missing data will be replaced by alleles from the nearest neighbours (Huang *et al.*, 2010). In these cases rare alleles can be missed, unless they are present in a similar accession. If rare alleles are of interest, an enzyme which cuts less frequently may be employed to increase the number of reads per site. Alternatively, multiplexing can be reduced or additional lanes run to increase coverage, although this will also increase cost per sample. Using PstI, a less frequent cutter, only 60,000 SNPs are obtained for maize, with 10x coverage from a single 384-plex lane. The coverage per SNP is high, but too few markers are generated. Using ApeKI, 2-4 million reads are required to approach full coverage of the observable 680,000 SNPs, so for landraces the 384-plex set is typically run on four lanes (E. Buckler pers. comm.).

RADseq library preparation costs range from $6 to $13 per sample based on 384 to 20 multiplexed samples. Sequencing costs are $1900 per lane, giving a total cost of $11 per 384-plexed or $108 per 20-plexed sample.

SNP genotyping

For some species which have many known Single Nucleotide Polymorphisms (SNPs), high density oligonucleotide arrays, or chips, have been developed for high throughput **genotyping** (e.g. RiceSNPs, 2012; MaizeSNP50, 2010). Other popular SNP genotyping assays include Kbioscience's KASPar (KASP, 2012) and Illumina's GoldenGate assay (GoldenGate, 2012). SNP genotyping has the advantage that the same set of SNPs is interrogated over all samples and the results generated by the various platforms are relatively easy to interpret with little further bioinformatics analysis required. However, the initial investment to generate the set of SNPs is large if SNP resources are not yet available for a given species, and may necessitate the sequencing of a number of diverse lines in order to identify SNPs (e.g. Ammiraju *et al.,* 2006, McNally *et al.,* 2009). In addition, if the accessions used to construct the SNP set do not represent the diversity of the entire species, the results may give a skewed impression of the total diversity when applied to other samples. A further limitation is that SNP assays will only reveal those variants included in the design, and as such novel variants will remain undetected by these approaches.

The cost of designing 200 SNP assays with KASPar technology for use on the Fluidigm genotyping platform is $134 per SNP for 2,500 samples, giving a cost of $11 per sample for 200 SNPs (KASP, 2012)

The cost of designing the oligonucleotide set (Oligo Pool All, OPA) for the BeadXpress is currently $27,000 to genotype 1500 individuals for 384 SNPs, a cost of $18 per sample.

A rice 1M SNP chip is currently being developed by Affymetrix in collaboration with Susan McCouch at Cornell. The price to genotype 1 million SNPs for a single sample is expected to be ~$450 (M. Lorieux pers. comm.)

SSRs

SNPs are the most abundant genomic markers, with more than 60 million simple genetic polymorphisms detected between human genomes to date and 5.4 million identified for rice (dbSNP, 2012). SNPs can be readily used for genotyping via sequencing or SNP assays as outlined above. However, the amount of information per SNP marker is low (i.e. there is a maximum of four possible bases for a given SNP, and most reported SNPs are biallelic) when compared with SSR (Simple Sequence Repeat) markers where the number of repeats often vary by tens of copies (e.g. Singh *et al.,* 2010). SSR analysis will be confounded by pooling of individuals heterozygous for a given locus, as such the analysis must be performed on individual seeds in cases where an accession may contain diversity (P. Isaac pers. comm.).

The automated detection of SSRs can be performed with capillary sequencers, and usually twenty to thirty loci are sufficient to characterise the diversity within a set of germplasm. Genic SSRs may be favoured over inter-genic SSRs (Dutta *et al.*, *2011),* although these are often less variable than inter-genic SSRs. Also the number of genes containing SSRs will be far fewer than those containing SNP markers, as such SNP markers have a greater application for breeding and use in fine mapping. For species without available SSR markers, transcriptome sequencing of a small number of individuals may be an attractive approach to identify candidate genic SSRs.

To genotype SSRs is cheap in terms of reagents, but can be labour intensive. SSR markers can be multiplexed to target ten loci at once for ~$1 per sample (excluding labour costs). Initial investment to identify SSR loci, and map them to ensure an even distribution is an additional cost.

Summary

Availability of a reference genome sequence can enable study of the basic biology for the species but alone does not aid breeding programs. The availability of genetic information from a number of individuals is essential to bring the impact of genomics to breeders. With several hundred markers and a mapping population, a genetic map can be developed. The availability of genetic markers linked to known phenotypes can enable marker assisted breeding. With the addition of a reference genome, markers can be anchored in the genomic context, providing opportunities to perform fine mapping and identify candidate genes underlying the phenotypic differences.

When considering which sequencing approach to use to obtain information for large numbers of individuals, there are several factors to consider. The size of the target genome will have an impact on cost, and ploidy may affect the strategy you need to use (for example, sequencing wheat chromosome arms individually), the relatedness of the individuals may affect sequencing depth required and the availability of existing resources such as SNP chips may also influence the choice. However, one should always keep in mind the purpose of the data when deciding how to generate it.

For a diversity study of an entire germplasm collection, a set of SSR markers may be sufficient.

However, to identify markers for association studies or marker assisted breeding, a greater depth of markers will be required which can readily be discovered by reduced representation approaches such as GbS or RADseq. These will identify SNP markers throughout the genome, but only a small percentage of the genome is targeted. RNAseq provides a method to sequence expressed transcripts which can provide sets of genic SNPs. Exome capture and SNP chips can genotype large sets of markers, but will be limited to variants within 'known' regions. Resequencing will provide markers spread evenly throughout the genome, but accessions may need to be sequenced to some depth if rare variants are of interest, as you might expect when exploring wild germplasm and landraces for diverse traits. The ideal case would be to perform full genome sequencing on all accessions, to identify all variants and rearrangements between individuals, however this is still expensive and the analysis is non-trivial. In addition, standard sequencing will not identify epigenetic variations which may be important for certain traits.

3: Using genomic data

Genotypic data generated from NGS technologies can be combined with phenotype data to predict loci associated with phenotypic traits, or generate estimates of **breeding values**, via genome-wide association studies (GWAS) and genomic selection (GS) respectively. Section 3.1 briefly describes the application of GWAS and GS to genebank materials, and describes some of the issues surrounding phenotyping of wild relatives. Section 3.2 explores how genomic data can impact genebank management strategies.

3.1 Genotype to Phenotype

The routine detection of large numbers of variants which can be used as molecular markers has provided new tools to breeders for the characterization of genetic content of individuals and the tracking of regions passed on from parents to offspring. Where markers have been associated with a phenotype of interest, marker assisted selection (MAS) can be employed to identify individuals likely to exhibit that phenotype. Many quantitative trait loci (QTLs) have been discovered for a range of traits, and some have been successfully introgressed into breeding lines using marker assisted breeding (e.g. Neeraja *et al.,* 2007, Suh *et al.,* 2011).

Genome wide association studies (GWAS) use large numbers of markers in hundreds of individuals from a population to detect loci statistically associated with the phenotype of interest (Klein *et al.,* 2005). The structure of the population can confound the analysis, so precautions must be taken to account for this (Price *et al.,* 2010). The rate at which **linkage disequilibrium** (LD) decays in the population will affect the granularity of the loci identified, with slow LD decay giving larger regions containing more candidate genes and as such, the analysis requires fewer markers than in species which exhibit a high rate of LD decay. A GWAS performed in 373 indica rice landraces returned 80 association signals for 14 agronomic traits (Huang *et al.,* 2010).

GWAS is widely used to detect individual variants which have a large (main) effect on phenotype, however epistatic (interaction) effects can also have a significant impact. These effects are more difficult to detect, requiring larger sample sizes and increased computational resources to test possible combinations of variants without loss of statistical power (Cantor *et al.,* 2010). In order to reduce the number of tests performed, SNPs associated with main effects are often prioritized to be tested for epistatic effects, however SNPs with epistatic effects do not always exhibit significant main effects (Hu *et al.,* 2011). Cantor *et al.* (2010) recommended performing simple score tests to identify variant combinations with significant effects, and then performing more sophisticated and computationally intensive tests on those candidates to estimate the effect size.

For more complex traits controlled by large numbers of loci, for example yield or human height, many QTLs of small effect have been proposed. In the case of human height, a highly heritable trait, >50 known QTLs can account for only 5% of the heritability (Hill, 2010). For traits such as these, genomic selection (GS) may be a preferable approach as it uses all available markers to predict the breeding value of individuals. A number of statistical approaches have been proposed with different statistics suitable for different populations (Heffner *et al.,* 2009). To date GS, has only been carried out within a breeding program, so the applicability of the statistical approaches to more distantly related individuals, as would be the case with landraces and crop wild relatives in a genebank collection, remains to be seen (Meuwissen, 2009). Recently, an exploration of the prediction accuracy of 390 SNP markers for five traits of interest in 358 cassava hybrids, cultivars and landraces was conducted by de Oliveira *et al.*, (2012). They found that the prediction accuracy increased if a subset of informative SNPs identified by GWAS was used as input for GS rather than all SNPs. For these traits,

phenotype still gave more accurate predictions than GS, however the reduction in number of improvement cycles when GS was used was predicted to outperform phenotype-based methods per unit time.

A three year joint CIRAD-CIAT project has recently started to explore the accuracy of genomic estimated breeding values (GEBV) for a rice breeding population based on two training populations: one population used in the development of the breeding population; and one which is a diversity panel of 200 tropical japonica lines unrelated to the breeding population. The training populations will be genotyped with high density SNP genotyping assays and phenotyped in multiple locations for yield components and grain quality under favourable upland conditions and yield and canopy temperature under drought conditions. The study will explore the impact on GEBV predictions of relatedness between the training and breeding populations, through comparison with the true breeding values of individuals for the traits under consideration (Grenier *et al.,* 2012).

Both GWAS and GS involve the statistical association of genotypic and phenotypic data within a training population. This information is then applied to a wider set of individuals, based on genotypic information alone. For GWAS, this means predicting loci associated with a particular phenotype, which allows the screening of other individuals at these loci to predict their phenotype for the trait of interest. For GS, an estimated breeding value is assigned to individuals taking into account all marker information. GS has been adopted by the dairy cattle breeding industry to predict breeding values of bulls with accuracies equivalent to traditional progeny testing for some traits (Hayes *et al.,* 2009). Avoiding progeny testing in cattle could double the rate of genetic gain by enabling breeding of bulls at the age of two years instead of five.

Impact of genomic characteristics

The genomic characteristics of a species can have a large impact upon which sequencing strategy should be adopted; genome size, ploidy, heterozygosity levels and **linkage disequilibrium** (LD) all have an impact on the number of markers required. In addition the study type plays a role, with diversity and phylogeny studies typically requiring far fewer markers than GWAS or GS.

When performing diversity or phylogeny studies, typically tens of multi-allelic markers such as SSRs are sufficient to give an overview of the composition of the population and genetic diversity between individuals. Alternately hundreds of biallelic SNP markers may be employed. In the case where so few markers are being used, provided these are spread out within the genome, differences in genome size and LD will have little effect, as the distances between markers will always be large. Ploidy however can have an impact, where polyploids may have different variants in each genome. This can cause problems for SSR genotyping, where each genome may contribute different numbers of repeats for orthologous loci. Marker probes may be designed to target a single genome or all genomes within a polyploid, however in the latter case variations in the flanking regions can cause variable detection efficiencies in each genome. Heterozygosity poses similar challenges, where heterozygous loci have a similar impact to polyploidy. For this reason probes are typically designed where the bases flanking a marker site are highly conserved. The mating system affects the level of heterozygosity observed, and for outcrossers more sites may need to be genotyped as the level of inter-accession diversity can be low when compared to intra-accession diversity.

For analyses such as GWAS and GS, usually upwards of tens of thousands of SNP markers are used, and samples are genotyped either by SNP chips or sequencing (reduced representation or resequencing). LD has a large effect on the number of markers required, where species with low LD requiring many more markers as the linkage blocks are smaller. The number of markers required has an impact on sequencing depth when using reduced representation libraries, in order to increase marker number, a more frequently cutting enzyme is selected; this in turn will increase the number of

fragments being sequenced, and additional lanes or reduced multiplexing may be required to maintain the desired coverage. Similarly, larger genome sizes will require additional sequencing depth to maintain coverage for both resequencing and also reduced representations (as more restriction sites are likely to be identified in larger genomes). Polyploidy and heterozygosity also require increased sequence depth to detect multiple homoeologues/alleles. As before, outcrossers may need additional markers as the level of inter-accession diversity can be low when compared to intra-accession diversity, and they may exhibit more heterozygous loci.

Use of wild relatives

Accurate phenotyping is important for both GWAS and GS, and generating high quality phenotypic data is often a limiting factor for these approaches. Accurately determining phenotypes for wild relatives is particularly challenging as wild plants may not be adapted to grow in available testing environments due to differences in day length, temperature or susceptibility to pests. In addition, the phenotype may not reflect alleles which are masked by epistatic effects from other loci. This is particularly evident in yield-related traits which are usually impossible to measure in wild relatives that have never undergone selection for agricultural yield.

One approach to identify wild relatives with resistance to a particular stress is to select accessions collected from environments characterised by that stress. This often provides individuals which exhibit tolerance/resistance to that stress as selection pressure will have removed those individuals unfit for survival under those conditions. However, individuals harbouring favourable alleles may also be found in environments where that stress does not occur, and these individuals would be missed by such an approach. Restricting the search to areas outside of the centre of origin often implies searching within materials which have passed through a genetic bottleneck when the new population was established, reducing the available diversity.

Crossing wild germplasm to an elite line (top-cross) provides one way to enable the evaluation of wild alleles in a more favourable genetic background. When choosing the elite parent it is best to choose an accession which is widely adapted to allow evaluation of phenotypes in multiple location trials (S. Beebe pers. comm.). Using multiple elite accessions will allow the exploration of wild alleles in different backgrounds, providing an indication of the stability of the effect. This approach can overcome some of the challenges of phenotyping wild accessions,but the success rate of making wide-crosses varies depending on the species and most phenotyping operations cannot manage more than a few hundred individuals at a time. With such a small sample of all possible combinations of wild and elite loci, it is inevitable that only a subset of possible phenotypes can be evaluated per cross. Combined with the large number of accessions available within genebanks for many species, even the large-scale application of this approach is unlikely to lead to the identification of all favourable alleles.

The 1001 Genomes Project aims to sequence 1001 wild Arabidopsis accessions using next generation sequencing by the end of 2012 (1001genomes, 2012). Several publications have been published to date, and there are already 471 released genomes available. A major finding of this project has been the variation present in wild accessions. Large numbers of SNPs have been detected, in almost all functional genes, with one third being disrupted by deletions or premature stop codons, however the majority of these are thought to be compensated for by the presence of an alternate gene model (Gan *et al.,* 2011). When comparing the wild sequences to the Arabidopsis reference genome (col-0; TAIR, 2012), large numbers of insertions and deletions have also been discovered. A similar phenomenon is observed in maize, where 50-77% of the maize genome is shared between any two varieties (50% is estimated from BAC-by-BAC sequencing, 77% is estimated from GbS which is biased to the less presence/absence variable portion of the genome, BucklerLab, 2012). The presence of novel regions may be associated with novel phenotypes. To observe this level of variability points towards the need for full genome resequencing to truly gain a clear picture of the

complexity within each species.

Whilst the process of domestication results in a genetic bottleneck, there may still be variability in domesticated species which has not yet been fully exploited by breeders. The Illinois maize selection for kernel oil content experiment provides one such example, and has been ongoing since 1896. High and low-content lines have been selected for, and the high-content lines have steadily increased in oil-content for more than 100 generations (Hill, 2010) with no signs of slowing. The increase was shown to be mainly due to variation in the founder lines, but subsequent mutations may also have contributed (Hill, 2010). If the maize results can be replicated in other species (Hill cites cattle and chicken as similar examples), breeding strategy and selection pressure on existing breeders materials alone may achieve some of the necessary increases in yield.

The breeding company Ceres (Ceres, 2012) uses genetic transformation to increase yield in crop plants. Starting with Arabidopsis genes, they over-express single copy genes and screen plants for large (>20%) changes in phenotype which have no detrimental effects and are inherited in a Mendelian fashion. A number of such genes which have a large impact on Arabidopsis phenotype have been successfully transferred to rice. Traits of interest include biomass, plant architecture, tolerance to biotic and abiotic stresses, and nitrogen use efficiency, illustrating that changing the expression of single genes may generate large changes in important phenotypes (Flavell, 2010). Therefore, altering expression levels of single genes may also achieve large yield increases using only the alleles found within existing breeders' materials.

Summary

High density genotype information can be combined with phenotypic data to create predictions of alleles associated with traits via GWAS or estimated breeding values via GS. The efficacy of GS when trained on diverse materials unrelated to the testing material is unknown, but the CIRAD-CIAT study described above should provide some interesting insights. GWAS works well on unrelated individuals to identify loci associated with phenotypic traits, making it suitable for use with germplasm collections. GWAS works best for simple traits, associated with few loci, and these make good targets for breeders to introgress, typically reducing the amount of non-elite genome being introduced into elite backgrounds.

An essential part of both GWAS and GS is having reliable phenotypic data for the training materials. Phenotyping wild accessions can be difficult, and some domesticated traits such as yield are not measurable in wild species. Top-crosses can introduce some alleles into elite backgrounds, where they may display a measureable phenotype. Alleles which are masked by epistatic interactions will not produce discernable phenotypic differences, and therefore cannot be detected by GS or GWAS. Both the Illinois maize experiment and the over-expression of genes at Ceres illustrate the progress that may be achieved using alleles already found within breeding materials. However, for characteristics which cannot be found within domesticates, crop wild relatives represent a valuable potential source of useful alleles.

3.2 Genebank management

Having genomic information available would not only impact how accessions are used, but also how germplasm banks operate, influencing which accessions are maintained (and how) in order to ensure cost-effective safeguarding of the diversity of alleles stored within the collection.

Maintenance of accessions

In order to maintain viable seed within genebanks, germination testing is routinely performed, and stocks are regenerated when germination levels drop below a certain threshold (75% germination for seeds in Kew's Millennium Seed Bank, Kew 2012) or when seed stocks become low as a result of supplying demand to the user community. For species distributed *in vitro* such as cassava, new plantlets must be generated regularly (every 6-18 months for cassava) to maintain a collection of viable plants for distribution (IITA, 2012). Different species remain viable for different lengths of time; some such as common bean (*Phaseolus vulgaris*) can be viable for 30 years without undergoing regeneration (D. Debouck pers. comm.).

When regenerating collections, one consideration is the size of the population required to accurately represent the diversity of alleles within the original accession. For accessions with very low genetic diversity, a few seeds may suffice to capture all of the variation within the accession, however for more diverse accessions larger numbers are required. If too few plants are selected from a diverse accession, this will result in genetic drift and the resulting seeds will no longer give an accurate picture of the original genetic diversity which was collected. Mating systems will also have an impact on regeneration sizes with more individuals required when regenerating outcrossers to avoid inbreeding depression. Genotypic information on intra-accession variation can help genebank managers to efficiently regenerate seed, without loss of genetic diversity.

In addition, whilst regenerating seed novel alleles may be introduced into an existing accession. Modern genebanks will have safeguards to reduce the likelihood of these errors, using electronic barcoding systems at each step of the process, having controlled plots to prevent introgression from nearby accessions (a problem for outcrossers) and phenotypic checks to discard seed which does not resemble the accession. Having genotype information would provide a more reliable system to verify that the seed produced at the end of the process is representative of the accession from which it purportedly came. For each period of regeneration, if the genotypes of the original accessions are known, a small set of distinguishing markers (fingerprints) can be chosen to ensure in a cost-effective manner that there is no mix-up of seed, or introgression of alleles.

Conservation of accessions

Of the 7 million accessions held in *ex situ* genebanks worldwide, ~2 million are thought to be distinct, with the rest being duplicates (FAO, 2010). Some duplication is intentional, as parts of collections are mirrored at multiple genebanks as a physical backup; in case of disaster at one site, these can be quickly recovered. However, other accessions may be duplicated, where accessions have the same, or highly similar, genotypes but are unintentionally maintained separately. For many species, particularly those which are expensive to maintain or regenerate, genebanks will usually have attempted to reduce the presence of duplicates based on passport data, agronomic information, and marker information where available.

The availability of genotype information will allow the identification of duplicates with much higher confidence, although epigenetic variations would not be sampled. In addition to identifying duplicate

accessions, genotypic information may also indicate where mixed accessions should be split and maintained separately in future. For accessions with high intra-accession diversity, classifying two samples as duplicates or mixed is not clear-cut. It has been proposed (McCouch *et al.,* 2012) that for these accessions a threshold of acceptable intra-accession, as opposed to inter-accession, diversity must be defined, but this will vary depending on species, and type of variety (i.e. traditional, elite or hybrid).

Once accessions have been classified as duplicates, a decision can be taken as to how to handle the duplication. McCouch *et al.* propose two approaches, the first being to combine duplicate accessions, providing this does not raise the intra-accession variability beyond the acceptable level. The second is to “archive” one of the accessions, removing it from active management. Maintaining an archived accession is much more economical than maintaining an active accession, but prevents the loss of alleles (although the accession will not remain viable in the archive indefinitely). If the archived accession has full genotypic data recorded which suggest some interesting properties at a future date, then the accession can be restored to active management.

In addition to streamlining existing collections, genotypic information can help genebank managers to assess which newly acquired samples will add diversity to their collections and should therefore be included as additional accessions. Furthermore, mislabelled or misidentified accessions can be detected by performing phylogenetic analyses and identifying anomalies (van Hintum, 2003).

For sequencing of accessions, it is preferable to use a single plant as the source of DNA (Tung *et al.,* 2010); homozygous plants also require a lower sequencing depth. A single seed may be chosen from an accession and purified by single seed descent (SSD) prior to sequencing. The advantage of this approach is that there is a supply of seed with known genotype which can be phenotyped to allow the more accurate association of genotypes with phenotypes. However, if this purified seed is to be maintained by the genebank, this could result in a doubling of the genebank size (assuming one seed taken from each accession). This is the strategy so far adopted by IRRI's rice resequencing project, USDA's lettuce project and JIC's pea genotyping program (Mike Ambrose pers. comm.). In order to prevent the maintenance of all of the purified seeds as new accessions, McCouch *et al.* propose to maintain accessions only if the material will be used for phenotyping as part of genotype-phenotype association studies. Purified seed which is sequenced without planned evaluations may be discarded, and only the information retained.

Databases

For genotypic data to have an impact on germplasm use, the storage and presentation of the information is critical, both for internal use by curators and external use for members of the research and breeding communities. Not all genebanks have online catalogues, and those that do currently display limited information. Passport data are often shown, along with perhaps some basic characterization data (GENESYS, 2011; Cassava Registry, 2012), but much of the data on evaluations which have been performed remains locked away in private institutional databases, or lab books. While evaluation information is potentially very valuable, it has often been collected over many years, by different individuals, who have recorded differing amounts of meta-data, such as information on how the trial was planned, the protocols used for measuring the traits of interest and relevant environmental information. Automated phenotyping approaches including image-based analyses are currently being developed which reduce opportunities for human error and subjectivity when recording results. Whilst ontologies for molecular biology have been in use for many years (e.g. The Gene Ontology Consortium, 2000), the Plant Ontology is comparatively new (Avraham *et al.,* 2008), the Crop Ontology has recently been published (Shrestha *et* al., 2012) and the Trait Ontology (TO, 2012) is currently under development. As such, awareness within the breeding community may be limited (Crop Ontologies for Agronomic Traits, 2011). Increased use of ontological terms coupled with

publication of, and compliance with, standardised phenotyping protocols will increase the utility of evaluation data and facilitate data sharing among institutes.

Existing genebank databases have mainly been developed for an individual genebank, with a couple of exceptions (GENESYS, 2011; GRIN-Global, 2012), and once the genebank's users are familiar with a particular system, there may be some reluctance to change and adopt an international standard. Currently there are no genebanks which display comprehensive genotypic data, and few, if any, currently have the resources to maintain, analyse and display genotypic data in an informative way.

There are two main groups of users who will be interested in accessing genotypic data within genebanks: molecular biologists and breeders, and each will have different objectives and strategies for using the information available.

Most molecular biologists will be familiar with sequence and variation data, and visualisation tools such as genome browsers (Ensembl, Flicek *et al.,* 2012; UCSC, Kent *et al.,* 2002; Gbrowse, Stein *et al.,* 2002). These users may want to download SNPs for use in other software for performing phylogenetic analyses or GWAS for instance, or they may have a locus of interest and want to explore variation within a certain region or gene, known as allele mining (Kilian and Graner, 2012). The visualisation of genomic variation within large numbers of accessions is a pressing issue for bioinformaticians, and the human 1000 genomes (1000 genomes, 2012) and Arabidopsis 1001 genomes (1001 genomes, 2012) projects are currently leading development in this area, however scalable visualisations of non-SNP variations (e.g. insertions/deletions, rearrangements and copy number variations) are still needed.

Developing access for breeders requires a different approach as their interests will be based on phenotypic traits, associated markers and pedigree/phylogeny information. Phylogeny can indicate ease of crossing for inter-specific or wide crosses. Associated markers can be employed in MAS. Genotypic data can provide information on phylogeny, however without phenotypic data there will be few known markers and few accessions associated with traits, limiting the data's usefulness to breeders. Some breeder-oriented tools to access genotypic and phenotypic data have been developed for barley (THT, 2012) and more recently cassava (Cassavabase, 2012). The Generation Challenge Programme‘s (GCP) Integrated Breeding Platform (IBP, 2012) provides information, tools and services for integrated plant breeding. Many areas of the site and tools are still under development, however it is possible to access crop-specific ontologies, and sets of 1000-2000 KASPar markers developed for ten important crop species.

Impact of genomic characteristics

The genomic characteristics of a species can also have a large impact upon genebank management strategies. When assessing intra-accession diversity, SSRs can be used, however non-identical seeds should not be pooled, and polyploids may cause problems. GbS may be a more effective strategy, as it can be easily automated and allows the multiplexing of many samples (currently up to 384) which can come from one or more accessions. Where individuals within an accession are homozygous diploids, multiple individuals may be pooled with a single barcode to assess the diversity, although the identity of which seeds exhibit which alleles would be lost. Sequencing depth should also be increased to ensure all alleles at each locus are detected. Species with low LD may benefit from larger numbers of markers to allow a more thorough assessment. Outcrossers are expected to exhibit higher intra-accession than inter-accession diversity, so more markers may be required to explore intra-accession diversity than for selfers.

Once intra-accession diversity has been assessed, this information can be used to identify sets of

SNP combinations which may separate one accession from another. These SNPs can then be used to check for problems with cross-pollination during regeneration and mix-ups. Again, species exhibiting lower LD may need more markers than those with high LD. For accessions with high intra-accession diversity, seed should also be checked post-regeneration to ensure that the original alleles are still present. The number of markers required will depend on the level of diversity within each accession.

Summary

NGS data has the potential to aid genebank management, through monitoring of intra-accession variability throughout the regeneration process, and helping inform decisions on the splitting and merging of accessions. The requirement for genetically identical plants for use in genotyping and phenotyping activities potentially puts an added strain on genebanks, through generation of novel accessions via SSD. However, only maintaining those accessions which will be phenotyped for such analyses reduces the number of additional accessions being stored.

User access to the data is a critical factor for its impact, and the requirements of different users must be taken into consideration during database and interface design processes. Adoption of international standards and ontologies, especially for phenotyping, will facilitate sharing of data, increasing the value of individual datasets beyond the life of each experiment. Guidelines emerging from the transPLANT project on infrastructure for plant genomics projects may aid this process (transPLANT, 2012).

4: Case studies

The following section details six projects employing large scale genotyping for genebank collections. Three projects are currently underway within the CGIAR: IRRI's 10,000k rice resequencing; CIMMYT's Seeds of Discovery project; and CIAT's genotyping of cassava collections from CIAT, IITA and EMBRAPA. ICRISAT and GCP's proposal to resequence reference sets is also outlined. In addition, two large-scale genotyping projects are described which have been completed in different lettuce collections. Finally, the WISP wheat pre-breeding project is described, which is making use of landraces, synthetics and wild relatives of wheat from collections for pre-breeding purposes, in the absence of NGS information.

IRRI – 10k rice resequencing

November 2011 saw the announcement of a collaboration between IRRI, BGI-Shenzhen (formerly at the Beijing Genomics Institute) and CAAS (Chinese Academy of Agricultural Science) to sequence the first 3,000 of 10,000 rice accessions selected from IRRI's genebank of 119,000 accessions (BGI, 2011a). The 10,000 have been selected to cover the diversity of the rice genebank collection, and contain accessions from the five varietal groups of cultivated rice: indica; temperate japonica; tropical japonica; aromatics; and aus, plus accessions of wild *Oryza rufipogon* and *Oryza nivara.* The cultivated accessions are comprised of both landraces and improved materials from breeding programs. While work on the first phase is well underway, work on the second phase will depend on securing funding for 2nd or 3rd generation sequencing.

The 10,000 accessions have been purified by single seed descent (SSD) and are stored in the IRRI genebank as separate accessions, increasing the size of the genebank collection by 10%. Sequencing of the 3,000 accessions was completed at BGI in January 2012. Full genome resequencing was performed to a depth of ~7x. Resequencing was selected in order to enable the identification of rare alleles that may be of interest. Alternative strategies such as the rice 1M SNP chip from Cornell would not be able to identify novel alleles, and genotyping by sequencing (GbS) may miss low frequency alleles, especially when there is a high proportion of missing data. In parallel to this resequencing effort, 2,000 lines are being genotyped with the 1M SNP chip, and these data will be used in conjunction with phenotyping information to identify SNPs associated with traits of interest via GWAS. There will be a subset of 200 lines both resequenced and genotyped with the 1M SNP chip.

Detailed phenotyping of the 2,000 1M SNP chip genotyped lines will be performed with a focus on abiotic stress tolerance (i.e. drought, extreme temperatures, submergence and salinity), resistance to pests and diseases, yield and grain quality. This phenotypic data will be used to predict gene-phenotype relationships, through collaboration with researchers from Cornell University. The aim being that these predictions can be extrapolated to those accessions with genotype data but no phenotypic information, helping to guide genebank curators and breeders in the selection of suitable materials for inclusion into breeding programs.

In addition, a further 50 lines will be sequenced to a greater depth (~30x) using Illumina paired-end sequencing with a variety of library-insert sizes (170, 500, and 800 bp) plus a 5kb mate pair library. Lines will be selected from each varietal group (except temperate japonica) in order to generate representative reference genomes for each group. Initially "pan-genomes" to represent the entire genome space of all individuals per varietal group will be constructed from these sequences. These sequences will be assembled *de novo* by BGI using SOAPdenovo (Li *et al.,* 2010) and IRRI will explore use of reference-guided assemblies using the temperate japonica Nipponbare reference (MSU, 2012). These pan-genomes will then be improved by the incorporation of regions assembled

from the lower-coverage 3,000 genomes.

The sequence reads from the 3,000 genomes will be aligned to the appropriate pan-genome, to identify variations (SNPs, small indels, copy number variations). As the pan-genomes are improved by the addition of novel regions from the 3,000 genomes, SNP coordinates relative to the reference will need to be adjusted; this will require the development of a novel dynamic coordinate system and data storage approach. Where missing data remains, it may be imputed to improve statistical power (Huang *et al.,* 2010). Haplotypes will be identified and accessions will be explored for novel alleles at known loci of interest. Accessions displaying novel alleles will be selected for phenotyping.

The data produced by this project will all be made publicly available through a portal developed by IRRI and hosted in the cloud. Raw reads will be deposited in EMBL's sequence read archive (ENA 2012), and the remaining data (assembled genomes, variants, and phenotypes) will be available for download from a cloud instance. In addition, there will be an annotation effort to generate predicted gene models on the pan-genomes involving the National Institute of Agrobiological Sciences, Japan (NIAS). These annotations can then be transferred to the lower-coverage accessions, allowing prediction of variant effects (VEP, 2012).

The annotated genomes will be displayed in a genome browser (e.g. Gbrowse, Stein *et al.,* 2002; UCSC, Kent *et al.,* 2002), along with tools to visualise genotypes (e.g. Flapjack, Milne *et al.,* 2010). Pre-computed GWAS analyses will be available, along with tools to allow users to perform GWAS analyses on their own phenotypic data (e.g. TASSEL, Bradbury *et al.,* 2007). Breeder friendly interfaces will be developed, taking advantage of progress made in these areas by USDA's Coordinated Agricultural Projects (e.g. Sol Genomics, 2012; THT, 2012).

Beyond the 10,000: The project proposes to sequence the entire germplasm collection, and as more data becomes available it may be possible to more accurately determine the minimum sequencing depth required. In general, the purified accessions used for the sequencing will only be maintained for phenotyping purposes. If a sequence from a new accession displays a novel allele it may be phenotyped, if not then the purified seed will be discarded and the purification and genotyping repeated if and when phenotyping will be performed, as this process is less costly than creating and maintaining a new accession. In the majority of cases, these accessions will only be sequenced and the phenotype will be predicted from the models derived from the initial 10,000.

CIMMYT – Diversity survey and association mapping in wheat and maize

CIMMYT's Seeds of Discovery (SeeD) project was launched at the end of 2010 with the aim of exploring the genetic diversity within international triticeae and maize collections, including CIMMYT's genebank comprising 125,000 wheat and 27,000 maize accessions (CIMMYT, 2012), maize genebanks at Mexican partner organizations such as INIFAP, and ICARDA's collection of wheat progenitors. SeeD is one of the four components of the Mexican-government funded MasAgro initiative, which aims to promote innovation across the value chain, starting at genetic resources and ending in extension work focusing on the promotion of conservation agriculture. It brings together national research partners in Mexico and international collaborators such as the James Hutton Institute (JHI), Cornell University and Diversity Arrays Technology Pty Ltd (DArT PL) as an industrial partner to establish a genetic-analysis service in Mexico to cover the project's genotyping needs and service Mexico's agricultural R&D community.

Due to differences in the nature of the wheat and maize genomes, and differences in the amount of currently available genomic information for these crops, distinct approaches are being undertaken for the two crops.

Wheat

At 17Gb, with A, B and D genomes and high numbers of repeated retroelements, hexaploid wheat is a complicated genome to sequence and assemble. The International Wheat Genome Sequencing Consortium (IWGSC) has adopted a 'divide and conquer' approach, sequencing and assembling flow-sorted chromosome arms individually, to avoid misassemblies due to **homoeologous** regions. With current sequencing technologies, it is not yet possible to sequence and assemble complete hexaploid wheat genomes.

As full genome sequencing is not suitable, CIMMYT plans to perform genotyping-by-sequencing (GbS, Elshire *et al.,* 2011) on CIMMYT's entire wheat collection to sample the diversity. GbS is a mechanism to sequence a reproducible fraction the genome (genome representation) through the use of restriction enzyme digestion, adapter ligation and PCR amplification of small restriction enzyme digested fragments (See Section 2.2: Reduced representation approaches). The combination of enzymes used for the digestion determines how many fragments are generated. For plants with low linkage disequilibrium (LD), like maize (Heffner *et al.,* 2009), a large number of fragments are required to allow association analyses (>0.5 million loci). For wheat the LD is higher, and as such associations can be determined from lower marker densities (i.e. fewer sequenced fragments, 40k-100k). This is fortuitous, as more sequencing capacity can thus be allocated towards increasing sequencing depth, resulting in datasets with less missing data and enabling the scoring of presence/absence variation or DArT markers and the classification of heterozygotes.

CIMMYT and DArT PL have selected PstI as their restriction enzyme for the wheat project to provide backward compatibility to the widely used DArT marker platform for wheat (genetically mapped DArT markers continue to be scored by sequencing through the GbS platform). PstI is methylation sensitive, providing enrichment of genic regions in plants. This is critical for wheat as non-genic regions are composed of repeated retroelements, which would make the identification of homologous SNPs in these regions impossible. DArT PL has been successfully using PstI to generate fragments for use in microarrays in a large number of species (50-60 DArT PL website). The SeeD project is using the same approach, however instead of hybridising the resulting fragments to an array, they are using NGS to detect markers in a high throughput manner (DArTseq). Use of GbS has an advantage over SNP chips or exome capture systems as no prior knowledge is required, making GbS a less biased approach, which is important when exploring unknown genetic diversity.

As more wheat samples are being analyzed by GbS, a 'consensus genome representation' comprising all GbS loci is being assembled and regularly updated. This consensus representation is stored in a database to be used as a reference against which new samples are being analyzed. Due to the hexaploid nature of the wheat genome it is necessary to differentiate between inter-homoeologue polymorphisms (IHPs) and true single nucleotide polymorphisms (SNPs), which is achieved by genotyping biparental populations, both from CIMMYT and other DArT customers, and identifying true SNPs which segregate in a Mendelian fashion. Loci containing true SNPs are continuously being discovered by this process and annotated in the database, so that more SNPs are being reported over time for new samples or samples for which the sequencing data are re-analyzed. Availability of a reference genome sequence for the A, B and D genomes will reduce the need for genetic mapping to determine whether polymorphisms are SNPs or IHPs; however if accessions are highly divergent from the reference this approach may still be required. This approach is, of course, also applicable to other polyploid crops.

In addition, subsets of varying size (200 – 20,000) of CIMMYT's wheat genebank are being evaluated in field trials for key agricultural traits prioritized by breeders (heat and drought tolerance, phosphorous-use efficiency, spot blotch, tan spot and blast resistance, and several quality traits), to generate datasets that will be subjected to association-mapping analyses.

Maize

Within the 27,000 maize accessions in the CIMMYT genetic resources unit, 25,500 accessions can be regarded as populations due to within-accession variation. It has been speculated that in some cases intra-accession variation may be higher than inter-accession variation. To explore within accession variation, 48 individuals from each accession will be selected, and DNA will be extracted from a leaf disc for each. This DNA will be pooled, and GbS performed using a restriction-enzyme combination that produces fewer amplifiable fragments than ApekI to achieve a higher sequencing depth to estimate allele frequencies per accession. Regions of low diversity will be explored as these may represent loci which are under selection.

A second analysis involves the selection of an individual seed from each of 5,000 accessions (landraces) from CIMMYT's breeders' core collection for maize. These individuals have been crossed with one of a set of six testers, stratified according to growth regions (two tropical, two subtropical and two highland). Some landraces of intermediate adaptation were crossed to two testers from two adaptation zones. The F1 progeny will be phenotyped in field trials taking place at a number of sites within Mexico. The 5,000 landrace parents and the testers will be genotyped at ultra-high density to perform genome-wide association studies (GWAS) to identify favorable haplotypes. In addition, the genome profiles will be used to identify individuals with high estimated genomic breeding values (GEBVs) by using the phenotyping data to train a genomic-selection (GS) model.

Computing infrastructure

The James Hutton Institute (JHI) is developing a non-relational database for ultra-high density GbS profiles, a data warehouse to store processed data and metadata for publication via a web portal, and summarization, visualization and query tools for the web portal. DarT PL have already built a data-access (middleware) layer and backend databases that hold phenotyping data, genotyping data with up to 100,000 SNP loci per sample, and environmental data. These JHI and DArT components will be integrated into an IT platform for storage, management and dissemination of genotyping, phenotyping and environmental data generated from the SeeD project. A pilot version of the web portal for the project will be available early in 2013.

CIAT – Sequencing the cassava collection

In December 2011, CIAT and BGI announced a collaborative agreement towards sequencing 5,000 cassava accessions (BGI, 2011b). These accessions will include *Manihot esculenta* landraces, improved varieties, and wild relatives from the CIAT, EMBRAPA and IITA germplasm banks.

CIAT's Genetic Resources Program currently has 6,592 cassava accessions maintained *in vitro*. The collection is comprised of 5,709 clones of *M. esculenta*, 5,301 of which are landraces (GRU, 2012). In addition, there are 883 genotypes from 33 of the 98 wild *Manihot* species. IITA and EMBRAPA have additional holdings of cassava germplasm, making approximately 9,000 accessions in these three genebanks combined. The objective is to obtain sequence for 5,000 accessions, which will represent the entire collection due to redundancy present between these genebanks (Joe Tohme pers.comm.). This will generate comparable data from the three main cassava germplasm banks worldwide.

Initially the 5,000 accessions will be sequenced using RADseq, to provide sequence tags from enzyme-digested genomic DNA. Currently a pilot project is underway with 100 accessions. The data will be made available through a custom database which will be developed by CIAT's bioinformatics group.

Once the 5,000 have been genotyped, a number of accessions will be selected for full-genome resequencing; the draft cassava genome having been completed in November 2009 by JGI and Roche 454 (Phytozome, 2012). These accessions will be selected to address a number of outstanding research questions in four areas:

1) Insights into domestication – where the genus and domesticated species originated, exploring phylogeography of the *Manihot* genus, and relationships between the wild and domesticated species.
2) Genetic diversity across regions and adaptation zones – exploring founder effects in Africa and Asia, and detecting variation related to eco-geographic adaptation and biotic stress resistance.
3) Trait-based analysis – looking for variations associated with yield components, morphological traits, resistance/susceptibility to pests and diseases, and root quality traits.
4) New breeding strategies – e.g. identification of heterotic groups and gene pyramiding strategies

Selecting lines to be sequenced based on biological questions of importance to the community aims to ensure that the data generated will be of immediate use for cassava research and breeding programs.

ICRISAT – re-sequencing reference sets

The concept of a core collection, to represent the diversity present within the sampled germplasm of a species based on morphological, phenotypic and geographic data is well established and has been described above. The GCP coordinated the genotyping of 32,000 accessions from the global composite/core collections of 21 species with 14-50 SSR markers in order to develop reference sets (Glaszmann *et al.*, 2010, Varshney *et al.*, 2010). These reference sets are designed to capture ~80% of the molecular diversity in a reduced set of lines. These materials have been made into purified genetic stocks (from one individual per accession), and have been stored as new accessions within the corresponding genebank (R. Varshney pers. comm.). Some phenotyping has also been performed on these reference sets (Glaszmann *et al.,* 2010).

With the reduction in sequencing costs and publication of relevant reference sequences, ICRISAT together with GCP is planning to re-sequence accessions from the reference sets of chickpea and pigeonpea. The strategy proposed is to re-sequence 300 accessions from each species to a depth of 5x with Illumina sequencing supplied by either BGI or Macrogen. These reads will be aligned to the reference to identify variants. The variation data will be combined with existing phenotypic data to perform Genome-wide association studies (GWAS). In addition, elite lines from breeders in developing countries will be genotyped using RAD-seq.

In addition to chickpea and pigeonpea, GCP is hoping to perform the same analysis for the cassava, common bean, cowpea and sorghum reference sets in collaboration with the corresponding CG centres.

Once the variants have been identified, these will be placed into a database with a genome browser interface, and links to the phenotypic data. The data format has not yet been decided, however ICRISAT plans to coordinate with ongoing projects and adopt standard data formats. The database will be accessible either via the GCP's Integrated Breeding Platform (IBP), or through ICRISAT's webpage.

Lettuce – genotyping two collections

Between 1997 and 2000, the Centre for Genetic Resources, Netherlands, characterised their entire lettuce collection using a mixture of AFLPs and microsatellites (van Hintum, 2003). The collection consisted of 2,323 accessions, 64% *Lactuca sativa*, and the remainder from 18 wild species and 4 species from related genera. Between two and thirty plants were sampled from each accession.

From the AFLP data, differences could be seen among individuals from the same accession. In addition, 20% of accessions could not be uniquely identified. While some of these accessions may be true duplicates, it is likely that the AFLP methodology was insufficient to differentiate between closely related accessions. However, the data were sufficient to allow the exploration of structure within the wild species, and the identification of some errors in passport information. In addition, these data allowed the identification of diverse material based on the AFLP fingerprints.

In 2012, the entire USDA-ARS National Plant Germplasm System lettuce collection was genotyped using a custom oligonucleotide set (OPA) of 384 SNPs with the Illumina GoldenGate assay on the BeadXpress platform. The 384 SNPs were a subset of the SNPs identified from 80,000 EST sequences from a variety of genotypes by the Compositae Genome Project (CGP, 2012). The SNPs were chosen such that they displayed a high level of polymorphism against a diversity panel of 36 cultivars, and that were widely spaced within the nine genetic linkage groups (Kwon *et al*. 2012). The USDA-ARS lettuce collection consists of 2,078 accessions, 72% *Lactuca sativa*, with the remainder from 28 other species.

Initially a pilot study was performed using 380 lettuce accessions (cultivars and landraces) from the five common horticultural types (Kwon *et al*., 2012). The authors found a high level of heterogeneity within accessions in spite of lettuce being predominantly self-pollinated. The presence of mixed homozygote genotypes could be due to accessions being collected as mixed genotypes or accidental mixing of seed during the regeneration process. Heterozygote genotypes however may have come from cross-pollination in the field (wind or insect) or be residual in accessions derived from bi- or multi-parental crosses. The authors found that the OPA used for genotyping gave highly reproducible results and that this OPA is suitable for rapid assessment of genetic diversity and population structure within the lettuce collection.

The pilot was followed by the genotyping of the complete collection. Several plants were grown per accession and grouped by phenotype. Within each accession, one plant from each phenotypic group was chosen for genotyping. The data is currently being analysed (J. Hu pers. comm.). Selfed seeds from a subset of plants with homozygous genotypes will be harvested and seed increased as a “pure-line” collection for storage, distribution, multi-location phenotyping and GWAS (Hu, 2012). This will ensure maintenance of maximal diversity within the collection and allow phenotyping and genotyping to be performed on identical materials for increased accuracy in association studies.

The original lettuce collection accessions will also be maintained to allow the maintenance of diversity “as is”, and large numbers of plants will be used in each regeneration cycle to minimise the risk of alleles being lost due to genetic drift or selection during the regeneration process. Whilst researchers are likely to be interested in pure-line accessions, breeders are less concerned with purity provided that the accession contains the alleles related to the trait of interest (Hu, 2012).

WISP - Enhancing diversity in UK wheat through a public sector pre-breeding programme

In 2011, the BBSRC (Biotechnology and Biosciences Research Council) funded the first 3-year phase of an initiative to re-establish a wheat pre-breeding programme in the UK's public sector. The objective is the development of pre-breeding germplasm, characterised for key traits, and the identification of genic markers for selecting these traits, for use both in commercial breeding programmes and for academic research. The project is a collaboration between UK universities and research institutes including the John Innes Centre (JIC), the National Institute of Agricultural Botany (NIAB), Nottingham and Bristol Universities, Rothamsted Research and the Institute of Biological, Environmental and Rural Sciences (IBERS) at Aberystwyth University. In addition, the wheat breeding industry is well represented on the steering committee, and aims to produce elite wheat cultivars from the germplasm developed throughout the project (WISP, 2011).

The project is divided into three sub-projects, utilising diversity from landraces, synthetic and wild relatives.

Landraces

In order to broaden the genetic base of wheat for the benefit of UK farming, the project aims to identify useful genetic variation from diverse sources of wheat germplasm to accelerate the genetic improvement of modern UK wheat. These sources include the Watkins collection (JIC, 2012) which contains 831 landrace accessions collected by British consulate staff from 33 countries between 1929-1932. The Watkins collection provides a snapshot of the diversity of global wheat landraces before domination by a few elite varieties. Accessions of the Watkins collection were found to have heterogenous phenotypes when tested for height, heading date and vernalisation requirement as part of the Defra funded Wheat Genetic Improvement Network (WGIN, 2009). Four seeds were grown from each accession and the resulting seed tested in field trials. All of these sub-accessions have been genetically fixed and multiplied and are maintained at the JIC seedbank. In addition, materials from the Gediflux collection comprised of >500 Western European winter wheat varieties that individually have occupied over 5% of national acreage from 1940 onwards are included, plus a collection of lines with extreme phenotypes from JIC, and non-UK parents of existing mapping populations, including key varieties from CIMMYT.

Phenotyping is being carried out on these materials to explore the genetic diversity and guide selection of parents for crossing and QTL mapping. Photoperiod insensitive and winter genotypes are removed at the F2 stage by marker assisted selection (MAS) to reduce the heading date window, which will increase the phenotyping precision. Traits of interest include increased biomass, enhanced nitrogen and phosphorous use efficiency, and resistance to aphids, bulb fly and Take-All. High-throughput marker platforms will be used with bulked segregants to identify QTLs efficiently. These QTLs will be confirmed with population screens. Data from 2,000 landraces and exotic varieties and 75 segregating populations will be produced. Potentially useful alleles will be introgressed into Paragon, an elite spring wheat, and a subset introgressed into a wheat diversity panel to measure the effects compared to current elite varieties. Diagnostic markers will be identified within the regions of interest to enable MAS.

Synthetics

The D genome of bread wheat contains little diversity, due to the difficulty of crossing hexaploid AABBDD with diploid DD genomes, whereas crosses between the hexaploid and tetraploid AABB are easier. CIMMYT has developed a number of synthetic wheats where *T. turgidum* (AABB) and diploid *Ae. tauschii* (DD) genomes have been combined to create synthetic hexaploid genomes. These will be

crossed with elite varieties to introduce D genome variation. At the same time, crosses between the elite varieties and a range of tetraploid donors will increase variation in the A and B genomes. In addition, mapping populations for wild emmer wheat (*T. dicoccoides;* AABB) and single chromosome founding lines for *Ae. tauschii* and emmer will be developed.

Wild relatives

The Ph1 locus controls pairing of homoeologous chromosomes during meiosis. A Ph1 mutant of the elite hexaploid bread wheat Paragon will be crossed with four wild diploid species, to generate inter-specific F1s. The wild species have been selected for a range of target traits: *T. urartu* (wheat A genome donor, implicated in photosynthetic capacity and disease resistance etc), *Thinopyrum bessarabicum* (highly salt tolerant, potential donor of genes for heat tolerance, drought and disease resistance), rye (a specific genotype which has resistance to all known rust diseases, heat tolerant, drought tolerant, resistance to acid soils etc) and *Ae. speltoides* (wheat B-genome donor, disease resistance, potentially insect resistance). These F1s will then be backcrossed to the wild-type Paragon giving BC1 plants with introgressions from the wild species, and these will then be selfed to produce plants homozygous for the introgression. The introgressed regions will be identified using DarT markers designed against the wild parents, and accessions with overlapping introgressed regions from the same parent can be crossed to produce smaller introgressed segments, reducing linkage drag from undesirable flanking genes. Illumina-based genotyping will be employed to determine more precisely the extent of the introgressed regions. First, cDNA libraries will be generated for the four wild and elite parents, with the aim of identifying 10,000 genic SNPs between the wheat and wild species. A SureSelect capture assay will be designed using these SNPs, and the captured genomic DNA from 140 selected introgression lines will be sequenced with Illumina to identify which parent the line matches at each SNP position.

NGS sequencing

This project aims to make use of genetic variation from landraces, and related species to improve domesticated wheat. The explicit use of NGS is mainly as a tool to identify the introgressions from wide crosses, where parents have been selected based on phenotypic data alone. However, the initial detection of genic SNPs is performed by sequencing normalised cDNA libraries (transcriptome sequencing) to identify variants. Twenty-four additional lines will be sequenced during this project taken from the Watkins collection, synthetics and wild species above. True varietal SNPs are identified bioinformatically by in-house software developed at Bristol University and used for genotyping. The predominant approach is to develop sets of SureSelect probes. A variety of different probe sets will be developed to provide a range of SNP sets of different sizes for use with a range of accession sizes, and different probe sets will be developed for the Watkins collection, the synthetics and characterised UK breeding materials.

The genotypic and phenotypic data will be made available via a custom-built relational database under development at JIC and Bristol University. The project also has a sizeable training component designed to attract young scientists into wheat genetics from breeding to basic research involving the underlying mechanisms of phenotypes.

Summary

The four CGIAR projects listed above are each using different approaches to discover genetic variants within genebank accessions. IRRI are performing resequencing to a depth of ~7x, which is an expensive strategy, but reflects the advanced position that the rice community is in with respect to genomics. The finished *Oryza sativa* genome was published in 2005 (International Rice Genome

Sequencing Project, 2005), large-scale SNP detection projects have been performed, and SNP chips have been developed (RiceSNPs, 2012). Rice is also a diploid with a small genome (~400Mb), making resequencing feasible. The IRRI 10K project chose resequencing to detect rare variants and structural variations which would be missed by other approaches. This project should provide an excellent resource for the rice community, however the approach is costly and not yet feasible for crops with larger genomes and higher ploidy, such as wheat. ICRISAT is also planning to perform resequencing to a depth of ~5x for chickpea and pigeonpea reference sets.

CIMMYT's SeeD project and CIAT's cassava project are using reduced representation libraries. These will provide large numbers of genomic markers but involve sequencing a fraction of the genome, therefore this is a more cost-effective approach. CIMMYT is employing bi-parental crosses to overcome the problem of differentiating between IHPs and SNPs, and performing studies on intra-accession diversity within maize. CIAT is using RAD-seq for its approach, and as such should be able to assemble larger regions (~500bp) against which KASPar or GoldenGate assays may be designed. CIAT intends to select materials for sequencing in order to answer a set of biological questions, to ensure that the data has immediate impact on its research program. The CIAT project is at a very early stage, and although a reference genome was published in 2009 (Phytozome, 2012) the community is not advanced in terms of genomic resources. As such the progress made by CIAT should provide insight into the benefits and limitations of genotyping germplasm resources for a relatively orphan crop. Lessons learned may be transferable to other species in similar positions.

Two non-CGIAR projects were also presented here, for lettuce and wheat. Two lettuce collections have been completely genotyped in 2000 and 2012. The first study used AFLP data, while the second used 384 SNPs developed by the Compositae Genome project. Neither set have high marker density, however the AFLPs were sufficient for diversity studies and to detect errors within the genebank records. The second set is currently under analysis, but it will be interesting to see what impact this set of SNPs has on genebank operations. The final project presented was the WISP pre-breeding project. This project was included as an example of a large project utilizing genetic resources in the absence of NGS data. Some sequencing is being performed, however the accessions will be chosen based on phenotyping alone, and NGS will be used to support other aspects of the project.

5: Recommendations

The world's genebanks contain a variety of different crops important for food security and their wild relatives. These species can have hugely varying genome sizes (e.g. 400Mb rice and 17Gb wheat) a range of ploidy, and different levels of intra-accession and intra-specific diversity relating to the domestication and reproductive processes. One thing that these species have in common, is that relatively little is known about the majority of accessions which comprise the genebank collections, beyond basic passport information and characterization data. Genebank managers are therefore required to help identify accessions which can contribute traits of interest to breeding programs for specific environments based on incomplete knowledge.

Beyond analysis of the passport information, two activities are available which can help increase knowledge of the collections: phenotyping and genotyping. Phenotyping the entire collection for a specific trait of interest remains the best way to determine a set of accessions with useful alleles, however those alleles which are not expressed due to the accession's genetic background will be missed by this approach. Test crosses to elite materials may reveal some of these missing alleles. Phenotyping is an essential tool, but the traits of interest to breeders are numerous and changing, and there are numerous and changing methods to measure them. Phenotyping an entire collection for all traits is a monumental task, however such large-scale phenotyping has been performed for a subset of traits. The National Bureau of Plant Genetic Resources, India, is currently performing field trials on the entire Indian national wheat germplasm collection, with all 22,000 accessions planted in three locations to perform evaluations for resistance against rusts and foliar diseases, terminal heat tolerance and characterization under optimum conditions (K.C. Bansal, pers. comm.).

Genotyping collections is much more tractable, although the extent to which heterogenous accessions should be genotyped is an area for discussion. Depending on the strategy, genotyping will provide sequence information for varying portions of the genome. This information can be used in isolation as a measure of diversity within the collection, allowing selection of materials for use based on genetic distance. Combining genotypic information with phenotypic information for a subset of accessions enables the identification of genomic regions associated with phenotypes of interest via GWAS (for low-complexity traits) or allows the estimation of breeding values via GS, although the success of GS on more distantly related samples remains to be seen. These approaches open the possibility of using genotype to predict the phenotype of material that has not been evaluated.

Genotyping approach

Of the genotyping strategies presented in Section 2.2, GbS and RNA-seq are the two most suitable to use for diverse collections without annotated reference genomes. These approaches do not require prior knowledge (unlike exome capture, SNP genotyping chips/assays or SSRs), so may identify novel variants. However, unlike resequencing, only a fraction of the genome is targeted, and as such these approaches are significantly cheaper, particularly for large genomes, with both GbS and RNA-seq generating sufficient SNP markers for use with GWAS or GS approaches. Whilst resequencing may be an option with today’s sequencing technology for species with small genomes and a good quality reference, the presence of large repetitive regions will continue to make the analysis of resequencing data problematic until long and accurate reads are available.

GbS will provide markers distributed throughout the genome, and is more than 10 times cheaper* than RNA-seq, however the tags returned are short and may not be able to be assigned to the correct homoeologs in polyploid species, unless a tag covers a known IHP. For this reason, the SeeD project includes a number of biparental crosses to try and identify SNPs based on segregation. In addition, if the decision is taken to resequence GbS genotyped lines in the future, the GbS tags will not contribute

significantly to the assembly, and as such cannot really be considered to be an investment towards more complete genome coverage in the future.

RNA-seq will provide large numbers of genic SNPs, but the same set of transcripts will not necessarily be identified in each accession, which may be due to variation in expression at time of sampling rather than genomic differences. RNA-seq may be >10 times as expensive as GbS*, and will generate fewer SNPs. In addition, bi-parental crosses may be required to generate ordered pseudomolecules from related species to allow GWAS to be performed if no reference is available (Harper *et al.,* 2012). RNA-seq is suitable for use in polyploids, and transcriptome assembly tools should continue to improve as more effort is invested in detection of alternative-splicing. In addition, the generation of transcriptome sequence is something that will remain valuable for annotation purposes, even if the decision is taken to perform genome sequencing of the same lines in the future.

*Based on 4 lanes of Illumina HiSeq2000 384-plex GbS at $9 per sample, compared to 1 lane of 8-plex RNA-Seq at $490 per sample

Accession heterogeneity

The presence of non-identical seed within an accession introduces noise when associating genotype with phenotype, and as such the most popular approach has been to select one or more seeds for genotyping, and perform SSD for each of the genotyped seeds to generate genetically identical seed which can be used for phenotyping. In some cases, this has resulted in the creation of novel accessions representing the purified seed, however McCouch *et al.* (2012) suggest an approach whereby purified seed is only maintained if it will be used for phenotyping. For GWAS or GS, only a proportion of the accessions will be phenotyped, and the predictions transferred to other accessions via genotype information, so this would only generate novel accessions for a subset of the collection. Should phenotyping be required on additional accessions, an individual would be selected, as above, for genotyping and phenotyping, as it will be cheaper to genotype an additional individual as opposed to creating and storing a novel accession long term.

The exploration of intra-accession variation may lead to splitting and merging of accessions, depending on the level of intra-accession heterogeneity acceptable to genebank managers. This requires the genotyping of multiple individuals per accession to determine the variability. The knowledge of variability could also be used to ensure that alleles are not lost through the regeneration procedure. A number of individuals could be pooled per accession and genotyped, which would give an estimate of allele frequencies within each accession. This information would be useful if a known marker is desired, as accessions with the highest frequency for the marker of interest may be favoured for inclusion within a breeding program. The sampling of multiple individuals per accession could be performed during the routine viability testing of accessions, when individual seeds are grown to test germination and then discarded.

For all accessions, knowledge of the intra-accession variability is important, for genebank management and to estimate allele frequencies within heterogenous accessions for breeders. For researchers performing GWAS and GS, individual SSD plants from a subset of accessions are required for genotyping and phenotyping. Individuals selected for SSD should be prioritised for genotyping and phenotyping, as the results from these studies will inform the predictions of phenotype for the other genotyped accessions. For collections where there is little intra-accession diversity, selecting individuals from existing core collections may be advantageous, as these are likely to have been chosen as diverse representatives and have existing phenotypic data which may be used for GWAS studies. For collections with high intra-accession variability, phenotypic data from the SSD individuals will be more reliable, and caution should be exercised when using existing phenotypic data from unknown genotypes. If there is no advantage in using existing phenotypic data, novel training

sets may be developed, and these should represent the widest possible diversity within the collection. For collections with no previous genotypic information, these selections will, by necessity, follow traditional approaches for developing core collections based on collection and characterisation data. As additional genetic diversity is discovered through genotyping the collection, individuals can be selected for SSD and added to this set for phenotyping.

Data standards

Genotyping and phenotyping a collection will generate novel data sets pertaining to those accessions. Where these activities are carried out by different centres, adopting standardised approaches for data generation and documentation will ensure that the data obtained can be widely used, and will have maximum impact.

A number of standards are already available, or emerging, for different data types. For sequencing data to be submitted to the public databanks (e,g. ENA, NCBI, DDBJ) there is a minimum set of information which must be recorded (e.g. SRA, 2012) describing how the samples were sequenced. Once the set of genomic variants has been identified, Variant Call Format (VCF) developed by the 1000 genomes project is fast becoming the standard way to record genomic variation data (VCF, 2012). However, the 1001 genomes project database is using both VCF and SHORE format (SHORE, 2012) (Fitz, J pers. comm.) which are then converted to an internal data representation for storage in a relational database (Polymorph, 2012).

Phenotypic data can be recorded in many different ways, and use of ontology terms to allow comparisons of datasets is not yet commonplace, although work is ongoing to establish ontology terms for different crop species (Crop Ontology, 2012; Shrestha *et al.,* 2012) and traits (TO, 2012). There is currently no ‘minimum information’ standard for recording phenotypic data sets registered in the MIBBI project (Minimum Information for Biological and Biomedical Investigations, Taylor *et al.*, 2008, MIBBI, 2012). One objective of the transPLANT project (transPLANT, 2012) is to establish data standards for recording phenotypic information.

In order to measure environmental effects on phenotype, phenotyping should be carried out at a number of locations, and over multiple years. Having a co-ordinated global network of phenotypic evaluation sites where would be an advantage, particularly where trials are difficult to conduct due to risks of pest/pathogen escape in regions which have not been previously exposed, or due to political pressures (e.g. for genetically modified organisms). Standardised documentation of methodology, recording of results, monitoring of soil and climatic conditions could be implemented, increasing confidence in the comparison of trials conducted at multiple sites.

Several projects are underway to develop databases to store and mine information related to phenotypic trials. CropStoreDB is a component of the InterStoreDB (Love *et al.*, 2012) which manages information on genetic, QTL and trait measurement data, and is being used by the brassica research community. The Agtrials database (Agtrials, 2012) provides an interface to store and access agricultural trial results and associated environmental meta-data e.g. weather and soil information. The Ephesis project (Ephesis, 2011) is exploring the integration of genotypic, phenotypic and environmental data to study genotype by environment interactions.

Variant calling

The purpose of resequencing many individuals within a species is to identify genetic variations. As previously discussed there are several ways this can be achieved, and which analyses should be performed will depend on what was sequenced (genomic or transcriptomic), with which sequencing

technology and strategy, and whether a reference genome is available which is appropriate to use for those samples. Once the sequencing strategy has been performed, there will still be a number of pieces of software to choose from when performing the analysis. Analysing next generation sequencing data is still a relatively young field, and very dynamic. There are several popular assemblers, aligners and variant detection methods. For groups with little experience, knowing which software to try for your data set can be challenging. In addition, the installation of these pieces of software (especially when using open source options) is not always straightforward, and having access to hardware appropriate for running these programs on large data sets can be another stumbling block.

One approach which may be appropriate is to put together a set of analyses for the most commonly performed tasks and share these amongst genebanks using an interface such as Galaxy (Galaxy 2012). Galaxy provides an intuitive interface for users to run workflows on their data sets; it maintains a history of the steps which were performed, and allows sharing of workflows and results with other users. Galaxy can be run in the cloud, or on an institute's cluster. If the Galaxy instance is coupled with the sequencing centre it would also prevent the need to download large data files from the sequencing centre to the user, who would then need to upload them in order to perform the analysis. This is the model currently being investigated by TGAC, as a way to make their hardware available to external users. The iPlant platform provides access to high performance computing for plant science research and is currently developing cyberinfrastructure to enable high throughput analysis of genotype to phenotype data (IPG2P, 2012). There are also a number of Grid computing collaborations which have been developed to facilitate computationally demanding projects by running them on distributed hardware. This has been common for analysis of data within the physics community, but the same model can be applied to bioinformatics analyses where we now face similar challenges with large data.

Once an individual has been genotyped, the sequence data associated with that accession is unlikely to change, unless a new round of sequencing is performed with a different strategy. The variant calling however, is less static, as software to perform variant calling is still under development. Therefore, whilst the current best approaches may be used to detect SNPs and small structural changes, it is likely that as additional approaches are developed, the raw sequence data may be reanalysed to give improved results. Using a Galaxy instance (or similar) for these types of analysis will reduce the burden on centres when re-analysing data, which would provide standardised high quality automated analysis.

Data access and visualization

Data generated from resequencing genebank accessions will be most valuable to the research community if it is made public. Whilst there may be a temptation to keep the information private to the genebank which generated it to enable its researchers to have a head-start on publications, if the aim of the sequencing is to increase use of genebank materials, making the data freely available is the best way to achieve this goal. In cases where the sequencing has been funded by public money, publishing the data may well be a prerequisite. As such, data access and visualisation are important factors to consider.

If the raw sequence data has been deposited within a public repository, there will be no need for individual genebanks to provide the data to users directly, a link to the repository would suffice. Variant Call format (VCF) files can also be deposited with public repositories, but where these simple text files are small they could also be provided for download from the genebank's website. If files are made available directly from the genebank, these should be in standardised formats to maximise their utility.

SNPs and small structural variants can be visualised within genome browsers, where a reference exists. Where no reference is available, for instance where RAD-seq or GbS has been performed on a species without a reference genome, some of the variants may be arranged into pseudomolecules based on synteny (e.g. Mayer *et al.*, 2011) or using genetic mapping approaches. Visualisation of sequence tags which are not ordered by these approaches will most likely be of limited interest to users.

A number of genome browers exist and different users often prefer different browsers. The EnsEMBL and UCSC browsers (Flicek *et al.,* 2012; Kent *et al.,* 2002) have existed for a number of years, with different user communities accessing shared data sets through their preferred interface. Where data can be shared in standardised formats, there is no reason that different genebanks cannot provide access to the same data through different interfaces, although unnecessary duplication of effort is not desirable. The development of a simple lightweight open source interface, which links accession information and ordering, to associated phenotypic and variant data, with visualisation in a genome browser would provide access to the basic information users would require. Once available, centres could use this interface, and as time and resources permit, they could contribute new modules to improve the usability. Generating this interface as part of the GMOD community would be an easy way to make use of existing experience and expertise (GMOD, 2012).

As with all public data repositories, data must be secure, have stable unique identifiers (especially when data is mirrored between sites), be reasonably fast to access and be presented through a user-friendly interface. User feedback should be solicited throughout the development process, and training should be considered to increase use. Use of online video tutorials for how to perform routine queries should be made available. Allowing users to contribute comments and annotations (crowd-sourcing) can also be valuable, enriching the data sets and increasing the sense of ownership the community feels in its resource.

There are currently three large-scale genotyping approached being undertaken within the CGIAR; two of which will produce reduced-representation data (CIMMYT GbS, CIAT RADseq), and the third which will produce resequenced genomes (IRRI). In addition, ICRISAT recently announced a collaborative partnership with BGI on applied genomics research and molecular breeding (ICRISAT, 2012) and is planning to resequence a number of germplasm reference sets in collaboration with GCP. CIMMYT is the most advanced in terms of its strategy for database storage and data access, having partnered with DarT PL and JHI. The other centres (CIAT, ICRISAT and IRRI) are not yet advanced in this area, and as such this provides an opportunity for these centres to work together to develop a single open-source solution that could be used not only by CGIAR genebanks but also beyond.

Due to the different approaches being adopted by these centres, preprocessing of the data will be project-specific. However, the end result will be similar, with variants associated with each accession, and phenotypic data associated with a subset of accessions. International standards should be adopted (or developed where necessary) for phenotyping protocols and recording of meta-data, using ontology terms. Data should be available for download, or visualisation using web-based tools, and groups such as Cornell, EBI, MIPS, TGAC or URGI could help in this area.

CIMMYT's approach to data management should be evaluated for suitability to the other projects, and due to the early stage at which the SeeD project is, there may be an opportunity to make adjustments to provide a solution for all four CG centres if necessary. The Trust could support this dialogue and perhaps invite institutes external to the CGIAR to advise, such as Cornell or USDA. A single unified solution would allow cross-talk for species genotyped at multiple genebanks, and multiple species genotyped at the same genebank, and avoid duplication of effort whereby each centre develops its own approach independently.

Pilot approach

For the Global Crop Diversity Trust, developing a strategy which is applicable across crop types is important, in terms of generating and also using the data. While the final objective may be to increase the knowledge and usefulness of accessions within genebanks for all of the Trust's crops of interest, a pilot study may be useful to explore the feasibility and scalability of the proposed approaches for a single genebank or crop. An ideal use-case would be for a species with little access to genomic resources, and a small collection, with a research community who would benefit from the availability of increased genomic information.

Perhaps the most immediate impact to the breeding and research community from genotyping a genebank's entire collection can come from GWAS studies identifying alleles of interest for simple traits which can be introgressed into breeding programs. Traits which can be evaluated in wild germplasm will eliminate the need for top crosses, reducing the time and scale required to perform the phenotyping; for example resistance to pests and disease.

Development of a project targeting two or three simply inherited but agriculturally important traits, in collaboration with two or three research groups (preferably with access to field trials to allow simultaneous phenotyping of multiple traits at multiple locations), would have potential to discover useful alleles relating to traits of interest. Initial successful use of the data would encourage further projects, perhaps targeting more complex traits, and therefore increase use and visibility of the genotyped collection. The progress of the pilot can be used to improve the strategy for further collections, should the pilot prove successful; success being measured by increased utility of germplasm for research and breeding.

To have impact, the pilot project would need to engage with genebank users. Having groups actively using the data will be one approach, through high-impact open-access publications of success stories. In addition, community meetings could be organised to raise awareness of the data and tools. The accessibility and usability of user interfaces to the data will be critical to ensure maximum gain from this resource, as previously discussed.

Cassava as a pilot

Of the four existing CGIAR projects, CIAT's cassava genotyping project is most similar to the pilot project proposed here. Whilst there is a draft cassava genome sequence available, there are very few molecular markers for breeding, and little resequencing data, so the impact of genotyping the cassava collection could be a good indicator of success for the majority of the Trust's crops of interest, in comparison with rice, maize or wheat which have many pre-existing resources. The global cassava germplasm collection is small (~9,000 accessions) and the traits which are predicted to be most important with respect to climate change are related to pests and disease (A. Jarvis, pers. comm.), which may be more amenable to phenotyping in exotic materials than other traits. In addition, as cassava is maintained *in vitro* and clonally propagated, there is no issue of within-accession diversity to contend with.

Cassava is a mandate crop for CIAT, EMBRAPA and IITA. If each institute were to propose one trait of interest, and each has the capacity for field trials, there could be three projects making immediate use of the data as proof of principle. Resulting open-access high impact publications would increase visibility within the community, and for other genebanks.

Development of a prototype interface, utilising data standards and ontology terms, could be coordinated between the three centres, as CIAT has experience with developing and maintaining a genebank database, and also has a bioinformatics group experienced in software development. IITA has strong links with Cornell through the Cassavabase project, where there is also expertise in

genomic selection (GS). EMBRAPA has also been exploring GWAS and GS in cassava (de Oliveira *et al.*, 2012). The analyses could be performed in conjunction with centres or projects with large compute facilities, such as iPlant, TGAC or the Grid Colombia project (Grid Colombia, 2012).

An additional advantage of involving both CIAT and IITA genebanks in the pilot study, is that both of these genebanks support multiple crops. These centres would be able to bring their experience to discussions over adoption of data standards and interfaces for their other mandate species also.
CIAT's cassava genotyping project is already underway, and as such the Trust can see what impact the project makes without having any input. However, without the Trust's involvement it is unclear whether CIAT will involve the other groups in development of this resource, which could be a missed opportunity. Also the timeframe for publication of this data is unknown, and the Trust may have to wait several years to see the impact on the community, during which time other projects are likely to have begun within the CGIAR. Ensuring that these independent projects use common standards and approaches may be difficult without the Trust's direct involvement.

Timeline

2012: Coordinate database discussions between CIAT, CIMMYT, ICRISAT and IRRI.

Seek collaborators and funds for pilot study

2013: Begin pilot project on a single crop (cassava)

SSD of core collection (not necessary for cassava), genotyping and phenotyping at multiple locations

2014: Continued phenotyping at multiple locations, begin genotyping non-core accessions (pooled if have intra-accession diversity), data visualisation

2015: Continued phenotyping at multiple locations, GWAS, publications and community meeting. Evaluate success.

2016: Start fundraising for other crops if pilot was successful

Start working with breeders to introgress loci from GWAS

References

1000 genomes (2012) http://www.1000genomes.org cited on 27th May 2012

1001genomes (2012) http://1001genomes.org cited on 27th May 2012

Agtrials (2012) www.agtrials.org cited 27th August 2012

Ammiraju JS, Luo M, Goicoechea JL, Wang W, Kudrna D, Mueller C, Talag J, Kim H, Sisneros NB, Blackmon B, Fang E, Tomkins JB, Brar D, MacKill D, McCouch S, Kurata N, Lambert G, Galbraith DW, Arumuganathan K, Rao K, Walling JG, Gill N, Yu Y, SanMiguel P, Soderlund C, Jackson S, Wing RA (2006) The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. Genome Research 16:140–147

Avraham S, Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, PujarA,Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, VincentL, Zapata F, Ware D (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Research 36:D449-D454

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, Blaxter ML (2011) Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism. PLoS ONE 6(4):e19315

Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D (2011) Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature 480:245-249

BGI (2011a) http://www.bgisequence.com/eu/newsandevents/news/bgi-unveils-significant-new-global-research-collaborations-at-th cited on 28th May 2012

BGI (2011b) http://bgiamericas.com/collaboration-for-large-scale-genome-sequencing-of-cassava-%E2%80%93-fourth-major-food-crop-in-developing-world/ cited on 13th May 2012

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633-2635

Brown AP, Kroon JTM, Swarbreck D, Febrer M, Larson TR, Graham IA, Caccamo M, Slabas AR (2012) Tissue-Specific Whole Transcriptome Sequencing in Castor, Directed at Understanding Triacylglycerol Lipid Biosynthetic Pathways. PloS One 7(2):e30100

Bucklerlab (2012) http://www.maizegenetics.net/gbs-overview cited on 26th May 2012

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 86:6–22

Cassavabase (2012) http://cassavabase.org/ cited on 27th May 2012

Cassava Registry (2012) http://istest.ciat.cgiar.org/cassavaregistry/JSPX/home.jspx cited on 27th May 2012

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and

Genotyping Loci De Novo From Short-Read Sequences. G3 1(3):171-182

Celera Assembler (2012) http://wgs-assembler.sf.net/ cited on 20th May 2012

Ceres (2012) http://www.ceres.net/ cited on 31st May 2012

CGP (2012) The Compositae Genome Project http://compgenomics.ucdavis.edu/ cited on 19th May 2012

CIMMYT, 2012 http://www.cimmyt.org/en/services/seed/wellhausen-anderson-plant-genetic-resources-center/the-facility cited on 31st May 2012

Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M (2011) Performance comparison of exome DNA sequencing technologies. Nature Biotechnology 29(10): 908-914

CoreGenomics (2012) http://core-genomics.blogspot.co.uk/2012/05/hiseq-2500-how-much-will-genome-in-day.html cited on 26th May 2012

Crop Ontologies for Agronomic Traits (2011) http://transplantdb.eu/?q=node/68 cited on 27th May 2012

Crop Ontology (2012) http://www.cropontology.org/ cited 27th August 2012

Darst RP, Pardo CE, Ai L, Brown KD, Kladde MP (2010) Bisulfite sequencing of DNA. Current Protocols in Molecular Biology 91:7.9.1–7.9.17

Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. Briefings in Functional Genomics 9(5-6):416-423

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Baxter ML (2010) Genome-wide genetic marker discovery and genotyping using next-generation sequencing, Nature Reviews Genetics 12:499-510

dbSNP (2012) http://www.ncbi.nlm.nih.gov/snp cited on May 26th 2012

de Oliveira EJ, de Resende MDV, da Silva Santos V, Ferreira CF, Oliveira GAF, da Silva MS, de Oliveira LA, Aguilar-Vildoso CI (2012) Genome-wide selection in cassava. Euphytica DOI: 10.1007/s10681-012-0722-0

Dutta S, Kumawat G, Singh BP, Gupta DK, Singh S, Dogra V, Gaikwad K, Sharma TR, Raje RS, Bandhopadhya TK, Datta S, Singh MN, Bashasab F, Kulwal P, Wanjari KB, Varshney RK, Cook DR, Singh NK (2011) Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [Cajanuscajan (L.) Millspaugh]. BMC Plant Biology 11:17

EC2 (2012) http://aws.amazon.com/ec2/instance-types/ cited on 20th May 2012

Ekblom R, Slate J, Horsburgh GJ, Birkhead T, Burke T (2012) Comparison between normalized and unnormalised 454-sequencing libraries for small-scale RNA-seq studies. Comparative and Functional Genomics 2012: 281693

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6(5):

e19379

ENA (2012) European Nucleotide Archive http://www.ebi.ac.uk/ena/home cited on 28th May 2012

ENCODE (2011) http://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf cited on 26th May 2012

Ephesis (2011) http://urgi.versailles.inra.fr/Projects/URGI-softwares/Ephesis cited 27th August 2012

FAO (2010) The Second Report on THE STATE OF THE WORLD's PLANT GENETIC RESOURCES FOR FOOD AND AGRICULTURE http://www.fao.org/docrep/013/i1500e/i1500e00.htm cited on 27th May 2012

Flavell, R (2010) Improving feedstock crops for biofuels. Oral presentation at the AgriGenomics World Congress meeting,Brussels.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Singh Riat H, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suárez XM, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SMJ (2012) Ensembl 2012. Nucleic Acids Research 40: D84-D90

Galaxy (2012) https://main.g2.bx.psu.edu/ cited 27th August 2012

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osbourne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Ratsch G, Mott R (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477:419-423

GENESYS (2011) http://www.genesys-pgr.org/ cited on 27th May 2012

Glaszmann JC, Kilian B, Upadhyaya HD, Varshney RK (2010) Accessing genetic diversity for crop improvement. Current Opinion in Plant Biology 13:167–173

Glenn TC (2011) Field guide to next-generation DNAsequencers. Molecular Ecology Resources 11: 759–769

GMOD (2012) http://gmod.org cited 27th August 2012

GoldenGate (2012) http://www.illumina.com/technology/goldengate_genotyping_assay.ilmn cited on 29th May 2012

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644–652

Grenier C, Bramel-Cox PJ, Noirot M, PrasadaRao KE, Hamon P (2000) Assessment of genetic diversity in three subsets constituted from the ICRISAT sorghum collection using random vs non-

random sampling procedures A. Using morpho-agronomical and passport data. Theoretical and Applied Genetics 101(1-2): 190-196

Grenier C, Châtel M, Ospina Y, Cao T, Guimaraes EP, Martinez CP, Tohme J, Courtois B, Ahmadi N (2012) Population Improvement through Recurrent Selection in Rice.Prospect for Maker Assisted Recurrent Selection and Genomic Selection.Oral presentation at the PAG-XX meeting San Diego. Grid Colombia (2012) http://www.gridcolombia.org/ cited on 27th August 2012

GRIN-Global (2012) http://www.grin-global.org cited on 27th May 2012

GRU (2012) http://isa.ciat.cgiar.org/urg cited on 13th May 2012

Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, Hattori C, Werner P, Bancroft I (2012) Associative transcriptomics of traits in the polyploid crop species Brassica napus. Nature Biotechnology 30:798–802

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92(2):433–443

Heffner EL, Sorrells ME, Janninck J-L (2009) Genomic Selection for Crop Improvement. Crop Science 49:1-12

Hill WG (2010) Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society B 365(1537):73-85

Hodgkin T, Brown AHD, van Hintum TJL, Morales EAV (eds) (1995) Core collections of Plant Genetic Resources. John Wiley & Sons, UK.

Hu Z, Li Y, Song X, Han Y, Cai X, Xu S, Li W (2011) Genomic value prediction for quantitative traits under the epistatic model. BMC Genetics 12:15

Hu (2012) Germplasm Management in the Post-genomics Era - a case study with lettuce. Oral presentation at the PAG-XX meeting San Diego.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics 42(11): 961-967

IBP (2012) www.integratedbreeding.net cited on 7th October 2012

ICRISAT (2012) http://www.icrisat.org/newsroom/news-releases/icrisat-pr-2012-media7.htm cited on 31st May 2012

IITA (2012) Cassava in vitro processing and gene banking http://www.iita.org/c/document_library/get_file?uuid=b5ec2a5b-ef6a-4ccc-8ba2-5c91cacf6963&groupId=25357 cited on 27th May 2012

Illumina (2012) http://www.illumina.com/systems/hiseq_systems.ilmn cited on 26th May 2012

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436(7052): 793-800

IPG2P (2012) http://www.iplantcollaborative.org/challenge/iplant-genotype-phenotype cited on 27th August 2012

IT PGRFA (2009) The International Treaty on Plant Genetic Resources for Food and Agriculture, Annexe I ftp://ftp.fao.org/docrep/fao/011/i0510e/i0510e.pdf cited on 31st May 2012

JIC (2012) http://www.jic.ac.uk/corporate/media-and-public/historic-grain.htm cited on 14th May 2012

Jing R, Vershinin A, Grzebyta J, Shaw P, Smykal P, Marshall D, Ambrose M, Ellis THN, Flavell AJ (2010) The genetic diversity and evolution of field pea (Pisum) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. BMC Evolutionary Biology10:44

KASP (2012) http://www.kbioscience.co.uk/reagents/KASP/KASP.html cited 29th May 2012

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Research 12(6):996-1006

Kew (2012) http://www.kew.org/science-conservation/save-seed-prosper/millennium-seed-bank/saving-seeds-worldwide/saving-seeds-at-the-seed-bank/checking-germination/index.htm cited on 27th May 2012

Kilian B, Graner A (2012) NGS technologies for analyzing germplasm diversity in genebanks. Briefings in Functional Genomics 11(1):38-50

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 308(5720):385–389

Kwon SJ, Truco MJ, Hu J (2012) LSGermOPA, a custom OPA of 384 EST-derived SNPs for high-throughput lettuce (Lactuca sativa L.) germplasm fingerprinting. Molecular Breeding 29(4): 887-901

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics 2: 231–239

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Research 20: 265-272

Love CG , Andongabo AE, Wang J, Carion PWC, Rawlings CJ, King GJ (2012) InterStoreDB: A Generic Integration Resource for Genetic and Genomic Data. Journal of Integrative Plant Biology 54(5): 345–355

MaizeSNP50 (2010) http://www.illumina.com/documents/products/datasheets/datasheet_maize_snp50.pdf cited on 26th May 2012

Manning K, Tor M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nature Genetics 38: 948-952

Marquez Y, Brown JWS, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Research doi: 10.1101/gr.134106.111

Mayer KFX, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, Kubalakova M, Suchankova P, Murat F, Felder M, Nussbaumer T, Graner A, Salse J, Endo T, Sakai H, Tanaka T, Itoh T, Sato K, Platzer M, Matsumoto T, Scholz U, Dolezel J, Waugh R, Stein N (2011) Unlocking the Barley Genome by Chromosomal and Comparative Genomics. The Plant Cell 23: 1249–1263

Mayes S, Massawe FJ, Alderson PG, Roberts JA, Azam-Ali SN, Hermann M (2012) The potential for underutilized crops to improve security of food production. Journal of Experimental Botany 63(3): 1075–1079

McCouch SR, McNally KL, Wang W, Sackville Hamilton R (2012) Genomics of gene banks: A case study in rice. American Journal of Botany 99(2):407-423

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Ratsch G, Buell R, Leung H, Leach JE (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. PNAS 106(30):12273-12278

Meuwissen THE (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution 41:35

MIBBI (2012) http://mibbi.sourceforge.net/portal.shtml cited 27th August 2012

Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WTB, Flavell AJ, Marshall D (2010) Flapjack – graphical genotype visualization. Bioinformatics 26(24):3133-3134

MIRA (2012) http://sourceforge.net/projects/mira-assembler/ cited on 20th May 2012

MSU (2012) Unified rice pseudomolecules 7 http://rice.plantbiology.msu.edu/ cited on 28th May 2012

Neeraja C, Maghirang-Rodriguez R, Pamplona A, Heuer S, Collard B, Septiningsih E, Vergara G, Sanchez D, Xu K, Ismail A, Mackill D (2007) A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. TAG Theoretical and Applied Genetics 115(6):767-776

Newbler (2012) http://my454.com/products/analysis-software/index.asp cited on 20th May 2012

NHGRI (2012) http://www.genome.gov/sequencingcosts/ cited on 20th May 2012

NuGEN (2012) http://www.nugeninc.com/nugen/index.cfm/products/pl/library-preparation/encore-384-multiplex-system/ cited on 20th May 2012

Omics! (2012) http://omicsomics.blogspot.co.uk/2012/02/oxford-nanopore-doesnt-disappoint.html cited on 11th March 2012

PacBioToCA (2012) http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=PacBioToCA cited on 20th May 2012

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 40(12):1413-1415

Phytozome (2012) www.phytozome.org cited on 13h May 2012

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE 7(2): e32253

Polymorph (2012) http://polymorph.weigelworld.org cited 27th August 2012

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11:459-463

RiceSNPs (2012) http://www.ricesnp.org/snpchips.aspx cited on 26th May 2012

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods 4:651-657

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A-L, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I (2010) De novo assembly and analysis of RNA-seq data. Nature Methods 7:909-912

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proceedings 5(Suppl 7):P54

Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Research 20: 1165-1173

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics doi: 10.1093/bioinformatics/bts094

SeqAnswers (2012a) http://seqanswers.com/forums/showthread.php?t=16709 cited on 11th March 2012

Shivaprasad PV, Dunn RM, Santos BACM, Bassett A, Baulcombe DC (2012) Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs. The EMBO Journal 31: 257–266

SHORE (2012) http://sourceforge.net/apps/mediawiki/shore/index.php?title=Shore_consensus#Prediction_file_formats cited on 27th August 2012

Shrestha R, Matteis L, Skofic M, Portugal A , McLaren G, Hyman G, Arnaud E (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Frontiers in Plant Physiology 3:326

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: A parallel assembler for short read sequence data. Genome Research 19:1117-1123

Singh H, Deshkumh RK, Singh A, Singh AK, Gaikwad K, Sharma TR, Mohapatra T, Singh NK (2010) Highly variable SSR markers suitable for rice genotyping using agarose gels. Molecular Breeding 25(2):359-364

Sol Genomics (2012) http://solgenomics.net/breeders/index.pl cited on 28th May 2012

SRA (2012) http://www.ebi.ac.uk/ena/about/sra_submissions cited on 27th August 2012

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. Genome Research 12(10):1599-1610

Suh J-P, Yang S-J, Jeung J-U, Pamplona A, Kim J-J, Lee J-H, Hong H-C, Yang C-I, Kim Y-G, Jena KK (2011) Development of elite breeding lines conferring Bph18 gene-derived resistance to brown planthopper (BPH) by marker-assisted selection and genome-wide background analysis in japonica rice (Oryza sativa L.). Field Crops Research 120:215-222

TAIR (2012) http://www.arabidopsis.org/ cited on 31st May 2012

Taylor CF, Field D, Sansone S-A, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz P-A, Bogue M, Booth T, Brazma A, Brinkman RR, Clark AM, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian Jr RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, Leebens-Mack J, Lewis SE, Lord P, Mallon A-M, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert Jr CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnology 26(8): 889-896

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25(1):25-9

THT (2012) The Hordeum Toolbox http://hordeumtoolbox.org/ cited on 27th May 2012

TO (2012) Gramene's Trait Ontology http://www.gramene.org/plant_ontology/ontology_browse.html#to cited on 27th May 2012

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28:511–515

transPLANT (2012) http://transplantdb.eu/ cited on 31st May 2012

Tsaftaris AS, Polidoros AN, Koumproglou R, Tani E, Kovacevic N, Abatzidou E (2005) Epigenetic mechanisms in plants and their implications in plant breeding. In: Tuberosa R., Phillips R, et al., (eds). In the wake of the double helix: from the green revolution to the gene revolution. pp. 157-171. Avenue Media. Bologna. Italy

Tung C, Zhao K, Wright MH, Ali ML, Jung J, Kimball J, Tyagi W, Thompson MJ, McNally K, Leung H,

Kim H, Ahn S-N, Reynolds A, Scheffler B, Eizenga G, McClung A, Bustamante C, McCouch SR (2010) Development of a Research Platform for Dissecting Phenotype–Genotype Associations in Rice (Oryza spp.). Rice 3(4):205–217

UN (2012) www.un.org/apps/news/story.asp?NewsID=40257 cited on 31st May 2012

Upadhyaya HD, Ortiz, R (2001) A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theoretical and Applied Genetics 102(8): 1292-1298

vanBerkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES (2010) Hi-C: a method to study the three-dimensional architecture of genomes. Journal of Visualized Experiments 39 http://www.jove.com/details.php?id=1869 cited on 31st May 2012

vanHintum (2003) Molecular characterisation of a lettuce germplasm collection. Eucarpia Leafy Vegetables.eds. Th.J.L. van Hintum, A. Lebeda, D. Pink, J.W. Schut: 99-104

Varshney RK, Glaszmann J-C, Leung H, Ribaut J-M (2010) More genomic resources for less-studied crops. Trends in Biotechnology 28: 452-460

VCF (2012) Variant Call Format http://www.1000genomes.org/node/101 cited on 9th August 2012

VEP (2012) Variant Effect Predictor http://www.ensembl.org/info/docs/variation/vep/index.html cited on 28th May 2012

Wendl MC, Wilson RK (2008) Aspects of coverage in medical DNA sequencing. BMC Bioinformatics 9:239

WGIN (2009) http://www.wgin.org.uk/resources/researchresults.php cited on 14th May 2012

WISP (2011) http://www.wheatisp.org/Documents/DOC_WISP.php cited on 14th May 2012

Yu J, Hu S, Wang J, Wong GK-S et al. (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica) Science 296(5565):79-92

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821-829

Appendix

I. Trips

Plant and Animal Genome XX, San Diego: January 14th - 18th 2012
- Attended 'Genomics of Genebanks' session
- Met with Ken McNally, Ruaraidh Sackville Hamilton and Ramil Mauleon (IRRI)
- Met with Chris Richards (USDA)

Cornell, Ithaca: January 18th – January 20th 2012
- Met with Jean-Luc Jannink, Ed Buckler and Susan McCouch

The Global Crop Diversity Trust, Rome: January 30th 2012
- Met with Hannes Dempewolf and Luigi Guarino

CIAT, Colombia: March 27th – April 2nd 2012
- Attended the 'Expert consultation workshop on the use of crop wild relatives for pre-breeding common bean, lima bean and tepary bean'
- Met with Daniel Debouck, Joe Tohme and Geoff Hawtin

II. 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
- HannesDempewolf
- 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

III. Abbreviations

AGBT – Advances in Genome Biology and Technology
AFLP – Amplified Fragment Length Polymorphism
ARS – Agricultural Research Service, USA
AS – Alternative Splicing

BBSRC – Biotechnology and Biosciences Research Council, UK
BGI – Beijing Genomics Insititute, China
bp – base pairs

CAAS – Chinese Academy of Agricultural Science, China
cDNA – complementary DNA
CGIAR – Consultative Group on International Agricultural Research
CIAT – Centro Internacional de Agricultura Tropical, Colombia
CIMMYT – Centro Internacional de Mejoramiento de Maiz Y Trigo, Mexico
CIRAD – Centre decoopération Internationale en Recherche Agronomique pour le Développement, France

DarT PL – Diversity Arrays Technology Pty Ltd, Australia
DNA – Deoxyribonucleic acid

EBI – European Bioinformatics Institute, UK
EMBL – European Molecular Biology Laboratory, Germany

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária, Brazil
EST – Expressed Sequence Tag

Gb – Gigabases
GbS – Genotyping by Sequencing
GCP – Generation Challenge Programme
GEBV – Genomic Estimated Breeding Values
GS – Genomic Selection
GWAS – Genome-wide Association Study

IBERS – Institute of Biological, Environmental and Rural Sciences, UK
IBP – Integrated Breeding Platform
ICARDA – International Center for Agricultural Research in the Dry Areas, Syrian Arab Republic
ICRISAT – International Crops Research Institute for the Semi-Arid Tropics, India
IHP – Inter-Homoeologue Polymorphisms
IITA – International Institute of Tropical Agriculture, Nigeria
INIFAP – Instituto Nacional de Investigaciones Forestales, Agricolas y Pecuarias, Mexico
IRRI – International Rice Research Institute, Philippines
IWGSC – International Wheat Genome Sequencing Consortium

JHI – James Hutton Institute, UK
JIC – John Innes Centre, UK

kb – kilobases

LD – Linkage Disequilibrium

MAS – Marker Assisted Selection
Mb – Megabases
MIPS – Munich Information Center for Protein Sequences, Germany
mRNA – messenger RNA

NGS – Next Generation Sequencing
NIAB – National Institute of Agricultural Botany, UK
NIAS – National Institute of Agrobiological Sciences, Japan

OPA – Oligo Pool All

PCR – Polymerase Chain Reaction

QTL – Quantitative Trait Loci

RADseq – Restriction site Associated DNA sequencing
RAM – Random-Access Memory
RNA – Ribonucleic acid
rRNA – ribosomal RNA

SNP – Single Nucleotide Polymorphism
SSD – Single Seed Descent
SSR – Simple Sequence Repeat

TGAC – The Genome Analysis Centre, UK

URGI – Unité de Recherche Génomique Info, France
USDA – U.S. Department of Agriculture, USA

VCF – Variant Call Format

WGS – Whole Genome Shotgun
WISP – Wheat Improvement Strategic Program

IV. Glossary

Adapter – short nucleotide sequences which are attached to fragmented DNA molecules in preparation for the sequencing process.

Align – the positioning of two or more sequences based on regions of sequence similarity. For instance, sequence reads are often aligned onto reference genomes.

Alternative splicing - the process by which exons transcribed from a single gene may be spliced together in different ways to produce multiple differently spliced transcripts.

Assemble – Genome assembly is the process by which shorter overlapping DNA sequences are combined to generate longer stretches of DNA sequence.

Baits – These are nucleotide probes which are complementary to target DNA and are labelled with beads, allowing the selection of DNA fragments of interest which can be sequenced after bead removal.

Breeding value – a value placed on an individual to describe the performance of its progeny.

Cloud – computing capacity and storage delivered remotely as a service.

Colour space – ABI's SOLiD sequencer reads pairs of nucleotides at a time, and encodes the combination as a colour, this encoding is referred to as 'colour space'.

Contigs – a DNA sequence made up from shorter overlapping sequences.

Demultiplexing – the process of separating samples which have been sequenced in a single lane/plate based on the sequence barcodes they have been tagged with.

de Bruijn graph – a data representation often used in sequence assembly algorithms. Sequence reads are represented as k-mers, subsequences of length k, generated by a sliding window, such that for a read of length l, l-(k-1) k-mers can be generated. K-mers are represented as nodes in the graph, with edges connecting k-mers where bases 2..k in the first k-mer match bases 1..k-1 in the second k-mer. Neighbouring k-mers generated from a single read will satisfy this property.

de novo – using no prior information, for de novo assembly this means assembling in the absence of guidance from a reference sequence.

Domestication bottleneck – the reduction in genetic diversity caused by the selection of a reduced number of individuals who exhibit domestication traits.

Epigenetic – heritable changes in gene expression or phenotype which are not caused by

modifications to the DNA sequence, e.g. methylation.

Flow cell – a slide into which samples are loaded for use with Illumina sequencing technologies. Illumina flow-cells currently have up to 8 separate lanes or channels.

Flow-sorted chromosome arms – flow cytometry is used to isolate individual chromosome arms from aneuploid plants.

Genotyping – identification of alleles at a set of loci for an individual.

Homoeologous – the relationship between chromosomes in a polyploid individual which were homologous in the ancestral species.

Homopolymer – a sequence of identical bases e.g. AAAAAAAA.

Insert sizes – this is the size of the DNA fragment to be sequenced minus the length of the adapter sequences.

Lane – one channel on a flow cell.

Library – a sequencing library is the processed sample which is ready to be sequenced. Library preparation steps include DNA fragmentation, addition of adapter sequences and amplification.

Linkage disequilibrium – the non-random association of alleles within a population, at multiple loci which are not physically linked.

Long mate pairs – short reads sequenced from ends of a long fragment of DNA, typically 3-20Kb.

Paired end – short reads sequenced from ends of a short fragment of DNA, typically <800bp.

Reads – short segments of DNA sequence read by the sequencing machines, typically 35-450bp.

Scaffolds – a DNA assembly composed of contigs and gaps (represented by 'N's), contigs are arranged based on information from long range sequences e.g. long mate pairs.

Shearing – the mechanical breakage of DNA e.g. by sonication.

Single end – only one end is sequenced from a short fragment of DNA.

Single nucleotide polymorphism (SNP) – single nucleotide sequence variation found within members of a population.

Transcripts – RNA sequences which are complementary to regions of the genic DNA from which they are transcribed.

Whole genome shotgun sequencing – genomes are fragmented and random fragments are then sequenced. These sequence fragments require assembly to reconstruct the original genome.