DRAFT FOR DISCUSSION

Generation Challenge Programme

Partnerships in modern crop breeding for food security

White paper on GCP research components: Genomic resources

22 AUGUST 2012

Table of Contents

Acronyms, short names and abbreviations – Genomic resources

BAC	bacterial artificial chromosome
CGIAR	No longer an acronym (*formerly* Consultative Group on International Agricultural Research)
CGIAR Consortium	CGIAR Consortium of International Agricultural Research Centers (*also* Consortium; one of two bodies of CGIAR)
CRPs	CGIAR Research Programmes
DArT	diversity arrays technology
EST	expressed sequence tag
GCP	Generation Challenge Programme (of the CGIAR)
GWAM	genomewide association mapping
GWS	genomewide selection approach
HICF	high-information-content fingerprinting
IBP	Integrated Breeding Platform (of GCP)
KASPar assay	KBioscience Competitive Allele-Specific PCR SNP genotyping system
MAS	marker-assisted selection
NGS	next-generation sequencing
PCR	polymerase chain reaction
QTLs	quantitative trait loci
SNP	single-nucleotide polymorphism
SSR	simple sequence repeat
STR	short tandem repeat
USD	United States dollar

Background and process

A series of white papers are being drafted by the Generation Challenge Programme (GCP) team in collaboration with external experts. The goals are to communicate the outputs and deliverables from each research component during 2004–2014 and to explore options for enabling and ensuring that the potential benefits of these components will be fully realised in the future. At this stage, the white papers are really a first analysis for internal use.[1] They are expected to evolve over time, shaped by progress made during GCP's remaining time and by the evolution of international agricultural research for development, particularly in terms of the 'moving landscape' of socio-economic, political and environmental issues in which operate the research portfolios of the CGIAR Consortium of International Agricultural Research Centers (CGIAR) and related CGIAR Research Programmes (CRPs). Each white paper is designed to contribute to GCP's orderly closure in 2014 by considering the following three questions:

1. What assets will be completed by the end of GCP's lifetime in December 2014?

2. What assets can best continue as integral components of the new CGIAR Research Programmes (CRPs) or elsewhere?

3. What assets may not fit within existing institutions or programmes and may require alternative implementation mechanisms?

This paper focuses on the outputs and options for GCP's genomic resources component. Outputs have been achieved through (a) collaborative work among three sets of actors: a broad network of partners in regional and country research programmes, the CGIAR and academia; and (b) capacity enhancement to assist developing-world researchers to tap into new genetic diversity and access modern breeding tools and services. GCP research activities have produced the research products described below[2].

Introduction and rationale

Accessing the genetic diversity found in genebank accessions or genetic stocks can be conducted efficiently through genetic studies that help identify new elite alleles for use in plant breeding. Molecular breeding, which combines genotypic and phenotypic information, has emerged as a powerful approach. It offers new perspectives for increasing the efficiency of breeding, reducing (among other things) the number of crop cycles required. Genetic

[1] This GCP white paper, like the others in this series, is not a conclusive, static document. Instead, it will continue to grow and evolve as the processes of evaluation and deliberation advance toward GCP's end in 2014.

[2] GCP is supported by generous funding from an array of donor organisations listed at http://www.generationcp.org/network/funders. See also descriptions of products at http://www.generationcp.org/impact/product-catalogue and of the institutions that generated them at http://www.generationcp.org/research/research-projects.

studies and molecular breeding approaches both require basic genomic resources such as molecular markers, genetic maps and sequence information. These resources were not available for several GCP target crops when GCP started in 2004, especially for the less-studied crops. In Phase II, GCP is focusing on nine of the 18 original GCP crops[3]. Those nine GCP target crops are maize, rice, sorghum, wheat, cassava, chickpeas, common beans, cowpeas and groundnuts.

Molecular markers are key tools for assaying genetic variation. Advances in molecular genetic technology have generated a range of marker types, with many reliant on the polymerase chain reaction (PCR). Some of these widely used marker systems include simple sequence repeats (SSRs), single nucleotide polymorphisms (SNPs) and diversity array technology (DArT). SSRs, also called microsatellite markers, have been, for various practical reasons, the leading marker types for breeding applications. The DArT marker system has proven extremely effective in exposing genetic diversity in germplasm collections, and is well suited for background selection in molecular breeding programmes. Finally, SNPs, which have been the prime marker platform in human and animal genetics for some time now, are rapidly gaining momentum. They are overtaking SSRs as the marker type of choice in plants as well, mainly because of their amenability for high-throughput genotyping and cost effectiveness.

GCP aims to access new diversity and promote the use of molecular breeding approaches to enable breeders in developing countries to increase the efficiency of their crop improvement programmes. For this reason, it was imperative for GCP to develop genomic resources that allow the genetic dissection of target traits and the pyramiding of elite alleles at key regulatory loci. The main objectives of the genomic resources component are therefore to:

- Develop sufficient numbers of molecular markers to conduct meaningful genetic studies for each GCP target crop.

- Develop high-throughput marker technologies for the more effective application of molecular breeding techniques to improve cultivar performance in target environments.

- Generate sequencing information for studies on genomewide association mapping (GWAM) and genomewide selection (GWS) in some GCP target crops.

To achieve these objectives, GCP commissioned about a dozen projects during Phase I (2004–2009) to develop SSR and DArT markers for orphan crops. For Phase II (2010–2014), the focus is more on developing SNP markers to enable high-throughput genotyping, a requirement for the type of molecular breeding being conducted by the six GCP crop Research Initiatives. The genotyping services of a commercial provider (KBioscience[4]) are also being used.

[3] The original 18 crops were barley, cassava, chickpeas, coconut, cowpeas, groundnuts, lentils, maize, millets, *Musa* spp, *Phaseolus* spp, pigeonpeas, potato, rice, sorghum, sweet potato, wheat and yam. Diversity studies, since completed, also included Andean roots and tubers, faba beans and soya beans.
[4] https://www.integratedbreeding.net/kbiosciencegcp-kaspar-snp-platform

To complement the large sequencing efforts conducted in the public sector to sequence crop genomes, GCP has contributed to the sequencing of BAC libraries (Phase I) or full genomes (Phase II) for cassava, chickpeas, common beans, *Musa* spp, pigeonpeas and sorghum. GCP also generated new markers and physical maps, thus contributing to the development of reference genetic maps and opening doors to genomewide approaches.

Overall, GCP invested about USD 15 million to the development of genomic resources. This sum represents about 10% of the Programme's total budget of USD 150 million spread over 11 years.

Project activities and outputs

Depending on the availability of their genomic resources, the world's main food crops can be grouped into three tiers (Annex 1). These crops are relevant to developing countries and are also mandate crops of the CGIAR Consortium of International Agricultural Research Centers (CGIAR). Considering the importance of the less-studied crops (tiers 2 and 3) for balanced diets or income in developing countries, several international initiatives have contributed significantly to the development of genomic resources for those crops. The following section focuses on GCP's contribution to this general effort.

Molecular markers: SSRs and DArTs

Species listed in tier 1 have benefited from adequate supplies of molecular markers, particularly SSRs (Annex 1). A significant effort was therefore made towards developing SSRs for crops listed in tiers 2 and 3. Two major approaches were deployed: the first was to construct SSR-enriched clone libraries, and the second centred on an *in silico* analysis of end sequences of clones from bacterial artificial chromosome (BAC) libraries or expressed sequence tag (EST) sequences. As a result, hundreds of SSR markers have become available for several crops such as chickpeas, common beans, groundnuts and pigeonpeas (Annex 1). However, the cost of developing SSR markers, especially via the enriched library route, remains high.

The need for a whole genome assay, to both characterise diversity and develop genetic maps, has led to the application of DArT technology to several crops listed in tiers 2 and 3 (http://www.diversityarrays.com/genotypingserv.html). While the first set of DArT arrays were developed for cassava, coconut, *Musa* spp, pearl millet, potato, sweet potato and yam, the expansion of existing DArT arrays was also accomplished for chickpeas, groundnuts and pigeonpeas (Annex 2). DArT markers have since proved useful in the areas of germplasm diversity assessment, genome mapping and gene tagging. They are also expected to be useful in marker-assisted selection (MAS), especially for monitoring the recovery of the recurrent parent genome.

BAC libraries

BAC libraries are an important component of physical mapping, map-based cloning and analysing gene and genome structure and function. BAC libraries, providing extensive genome coverage, have been established in most well-studied crops such as barley, maize,

rice, sorghum, soya beans and wheat. GCP focuses its efforts on constructing and sequencing BAC libraries in several less high-profile crop species such as cassava, cowpeas and groundnuts.

Two independent cowpea BAC libraries, comprising about 74,000 clones and a 17X genome coverage, have been developed. High-information-content fingerprinting (HICF) analysis of 60,000 of these clones has led to their assembly into 790 contigs (http://phymap.ucdavis.edu/cowpea/). A similar project in cassava produced a BAC library of 72,000 clones that represented about 10X coverage of the genome. The HICF method also assembled about 58,000 clones into 2,104 contigs (http://cassava.igs.umaryland.edu/). Although a 6.5X groundnut library had already been established, two new BAC libraries were recently constructed for fingerprinting: one was derived from the crop's *A* genome progenitor *Arachis duranensis*, and the other from its *B* genome progenitor, *A ipaensis*.

Finally, the physical maps of several tropical crop species are now being developed. As these maps become integrated with genetic maps, they will provide the framework for full genome sequencing.

Sequence-based, next-generation genomic resources

Recent years have witnessed the advent of several high-throughput sequencing and genotyping technologies. These approaches have great potential to facilitate the development of genomic resources, even in those crops that have few genomic resources. Such potential and applications of next-generation sequencing (NGS) technologies include Roche's 454 and Illumina (Solexa) sequencing. In fact, because of NGS technologies, whole genome sequences are becoming available in plant species that did not have even basic genomic resources such as a repertoire of markers or genetic maps. These approaches have therefore been used for the fast-pace development of sequence-based genomic resources for some of the crops listed in tiers 2 and 3 that previously had been considered as orphan crops (Annex 1). GCP has contributed significantly to the genome sequencing of chickpeas, *Musa* spp, pigeonpeas and sorghum.

SNP resources and high-throughput genotyping platforms

International collaborative efforts have generated large-scale SNP resources and genotyping platforms in barley, maize (www.maizegenetics.net), rice (www.ricesnp.org), soya beans and wheat (http://wheat.pw.usda.gov/SNP/new/index.shtml). Initiatives like GCP have extended SNP technology to several tropical crop species. Two main approaches have been applied for SNP discovery: the first is based on available EST and short tandem repeat (STR) sequences, and the second on allele re-sequencing of candidate genes or BAC ends. The cowpea SNP programme has delivered about 10,000 SNPs from an *in silico* comparison of about 183,000 ESTs generated from fewer than ten cultivars (http://harvest.ucr.edu/). Re-sequencing an average of 600 genes per species among the mapping parents used in chickpeas, common beans, cowpeas, pigeonpeas and a diploid progenitor of groundnuts has identified about 16,000 SNPs in these tropical legume species. Building on this large effort to identify SNPs, GCP participated in a collaborative effort with KBioscience (a genotyping service provider) to develop a set of SNP markers for GCP target crops for Phase II. Annex 3 summarises the number of SNP markers from KBioscience that are available today for 11 crops and the number of markers expected to be available by December 2014 for 15 crops.

Measuring success

When GCP started in 2004, the genetic studies of several crops listed in tier 2 and most of those listed in tier 3 had little application because the crops had few, if any, genomic resources. Today, this bottleneck has been overcome, together with that of accessing genotyping facilities and high-throughput marker technologies. All GCP target crops now have sufficient genomic resources to conduct meaningful genetic studies and molecular breeding (Annex 1).

Post-GCP sustainability and projected impact

Genomic resources are in a very special position, compared with other kinds of GCP products, as most of them are publically available and easily accessible through various websites such as those referred to above. Based on sequence information, markers can be synthesised in-house or through service providers. Access and sustainability for those products are therefore not really issues. Even for high-throughput markers, after supporting GCP funding the development of a first batch for each marker, the next batch can now be negotiated to be included in the genotyping price to be paid by users. Up-front payment to develop KASPar (KBioscience Competitive Allele-Specific PCR SNP genotyping system) markers is therefore no longer needed.

Deployment of those genomic resources will also be significantly enhanced through virtual platforms, such as the Integrated Breeding Platform (IBP) at https://www.integratedbreeding.net/, initiated in mid-2009 by GCP in collaboration with the Bill & Melinda Gates Foundation. Further aided by the information and communication technology revolution, breeders in developing countries now have better access to genomic resources, advanced laboratory services, and robust analytical and data management tools.

Until a few years ago, genomic resources for chickpeas, cowpeas, pearl millet, pigeonpeas and sweet potato were either non-existent or displayed an inadequate level of marker density. The far-reaching improvement in molecular marker technology witnessed over the past few years has helped put in place serviceable genetic maps for all these species (Annex 1). In accordance with the availability of genomic resources such as markers, genetic maps and transcriptomic resources, QTL identification or gene-cloning projects have been used for several less-studied crops for both biotic and abiotic stresses (Annex 1).

Suitable genomic resources, available in a user-friendly way through IBP for example, would allow molecular breeding to contribute significantly and positively to crop productivity in developing countries. However, for this to happen in tier 2 countries, several hurdles must be overcome: limited human resources, inadequate field infrastructure and limited capacity in information management. Although large-scale molecular breeding activities is unlikely to be conducted in the near-term in most developing countries, prospects are bright for breeders in these countries to take advantage of large international initiatives that will ensure access to germplasm, data, tools and methodology, thus enabling them to conduct efficient molecular breeding.

On a relative scale of 1 to 5, where 5 represents the largest impact across all kinds of GCP products, regardless of activity or crop, and 0 no impact, all GCP's efforts to develop and deploy genomic resources are estimated to have an impact factor of 4. Such a score indicates that the genomic resources development component has a highly significant impact on genetic study and plant breeding efficiency in developing countries.

Analysing the post-GCP placement of the genomic resources component

Annexes I, II and III indicate that GCP's objectives in developing genomic resources are most likely to be successfully attained by December 2014. That is, the development of the genomic resources effort as envisioned in the original workplan for GCP will be finished by December 2014. Indeed, for Phase II, GCP had considerably reduced its investment in that area because major achievements had already been made in Phase I. The focus in Phase II is on developing high-throughput genotyping markers and facilitating their use through IBP, thus providing access to next-generation sequencing (NGS) technologies.

Conclusion

We can therefore claim "job done" for GCP crops. Access to genomic resources for implementing molecular breeding in developing countries is no longer an issue. Nor is there need to think of extending activities to CRPs, Centres or other institutions.

However, going beyond GCP's own scope and internal assessment on meeting Programme objectives, an assessment that includes the high degree of crops diversity cultivated in developing countries would most certainly indicate that the job is not completed for all subsidiary crops.

GCP remains committed to its mission and community to the last day of the Programme and will work with partners along the delivery chain to maximize successful implementation of the delivery plans developed for each Research Initiative. GCP will also closely engage with its partners until its very sunset to ensure – as far as will be possible – the integration, extension, and expansion of activities as may be required. The Programme will even help initiate related new activities that build on GCP's achievements, should there be clear added value and demand for such activities. In this way, the Programme is working to secure a broad and sustainable use of its products well beyond 2014 while mitigating against the loss of gains made this far.

Annex 1: The availability of genomic resources in the public domain for each of the CGIAR/GCP mandate crops

Crop	Botanical name	Chromosome number (n) and ploidy level (×) (genome size in Mbp)	Marker resources (mainly SSRs and SNPs)	Genetic map density	Physical map	Transcript-omic resources and gene discovery	Genetic resources and diversity character-isation	Trait mapping and molecular breeding
Tier 1								
Barley	*Hordeum vulgare*	2n=2x=14 (5,500)	++++	+++	+++	+++	+++	++
Maize	*Zea mays*	2n=2x=20 (2,500)	++++	++++	++++	+++	+++	+++
Rice	*Oryza sativa*	2n=2x=24 (430)	++++	++++	++++	+++	+++	+++
Sorghum	*Sorghum bicolor*	2n=2x=20 (750)	+++	+++	+++	+++	+++	+++
Wheat	*Triticum aestivum*	2n=6x=42 (16,000)	++++	+++	+++	+++	+++	+++
Tier 2								
Chickpeas	*Cicer arietinum*	2n=2x=16 (740)	+++	+++	-	+++	+++	+++
Common beans	*Phaseolus vulgaris*	2n=2x=22 (~637)	+++	+++	++	++	+++	+++
Cowpeas	*Vigna unguiculata*	2n=2x=22 (620)	+++	+++	++	++	+++	++
Ground-nuts	*Arachis hypogaea*	2n=4x=40 (2,890) 2n=2x=20 (A and B genomes: 1,260)	++	++	++	+	+++	++
Potato	*Solanum tuberosum*	2n=4x=48 (850- 1000)	+++	++	++	++	+++	++
Tier 3								
Banana	*Musa acuminata*	2n=2x=22 (552-607)	++	++	+	++	++	+
Cassava	*Manihot esculenta*	2n=2x=36 (760)	++	++	++	++	++	+
Coconut	*Cocos nucifera*	2n=2x=32 (3,600)	++	++	-	-	++	-
Lentils	*Lens culinaris*	2n=2x=14 (4,063)	+	+	-	+	+	+
Pearl millet	*Pennisetum glaucum*	2n=2x=14 (2,450)	++	++	-	+++	++	+
Pigeon-peas	*Cajanus cajan*	2n=2x=22 (858)	++	+	-	++	++	-
Sweet potato	*Ipomoea batatas*	2n=6x=90 (2,200- 3,000)	++	++	-	++	++	-
Yam	*Dioscorea alata*	2n=2x=40 (550)	+	+	-	-	+	-

Availability of resources is shown by '+' signs, as follows: one '+' = basic, two '+' = moderate, three '+' = good, four '+'= excellent. Red '+' represent GCP's contributions; black '+' indicate resources that were already developed and publicly available.

Annex 2: CGIAR/GCP crop species for which DArT markers are available

Crop	Genotypes used for developing library	Features on array	Diversity panel surveyed	Polymorphic features[a]	Marker loci on genetic map
Cereal crops					
Barley	>200	> 50,000	>700	3,072	~3,000
Pearl millet[b]	96	7,680	96	1,500	In progress
Rice	>150	>30,000	>300	1,900	~1,000
Sorghum[b]	>200	>20,000	>1,000	2,500	2,000
Wheat	>300	>80,000	>1,000	7,680	~5,000
Legume crops					
Chickpeas[b]	250	21,500	300	5,400	>700
Common beans	150	>22,000	150	2,500	–
Groundnuts[b]	164	15,360	300	>5,000	~200
Pigeonpeas	300	29,000	400	>5,000	>400
Root, tuber and other crops					
Cassava[b]	270	18,000	450	2,500	450
Coconut[b]	130	7,680	130	400	–
Musa spp[b]	100	15,360	300	>5,000	>500
Potato[b]	>150	>20,000	>300	9,000	>4,000
Sweet potato[b]	96	7,680	96	2,000	800
Yam[b]	94	6,890	94	>2,500	–

[a]The number of informative markers in chickpeas, groundnuts and pigeonpeas includes related wild species. The number for yam is a projection of current work.
[b]Species for which the research was supported by GCP. A complete list of crop species for which DArT arrays are available is posted at http://www.diversityarrays.com/.
SOURCE: A Kilian, DArT Pty Ltd, Australia

Annex 3: KASPar SNP markers available for different crops today and projected numbers for December 2014

Crop		Current number of KASPar SNPs as of July 2012	Potential total number of KASPar SNPs as of December 2014
1	Barley	0	1,900
2	Cassava	1,740	2,140
3	Chickpeas	2,068	2,468
4	Common beans	1,497	2,197
5	Cowpeas	1,122	1,522
6	Finger millet	0	1,900
7	Groundnuts	91	1,991
8	Maize	1,250	1,650
9	Pigeonpeas	1,616	2,016
10	Potato	0	1,900
11	Rice	2,015	2,415
12	Sorghum	1,503	1,903
13	Soya beans	1,082	1,482
14	Sweet potato	0	1,900
15	Wheat	2,714	3,114
	Total	**16,698**	**30,498**