The International Treaty
ON PLANT GENETIC RESOURCES FOR FOOD AND AGRICULTURE

Background Study 5

PLANT GENETIC RESOURCES AND GENOMICS: MAINSTREAMING AGRICULTURAL RESEARCH THROUGH GENOMICS

Norman Warthmann

This document has been produced by the request of the Secretariat of the International Treaty and in the context of the first expert consultation on the Global Information System on Plant Genetic Resources for Food and Agriculture to stimulate the discussion on genomics and to facilitate the consideration of appropriate technical and organizational linkages during the development and implementation of the Global Information System.

www.planttreaty.org

Author:
Dr Norman Warthmann, Lecturer in Plant Biology, Genetics and Genomics at The Australian National University

NOTE FROM THE SECRETARIAT

This study is available on line at:
http://www.planttreaty.org/content/background-study-paper-5

BACKGROUND

1. Through Article 12.1 of the Treaty, Contracting Parties agreed to facilitate access to plant genetic resources for food and agriculture under the Multilateral System and in accordance with the provisions of the Treaty.
2. Among the conditions of the transfer, Article 12.3.c. of the Treaty states that "All available passport data and, subject to applicable law, any other associated available non-confidential descriptive information, shall be made available with the plant genetic resources for food and agriculture provided".
3. Article 12.4 of the Treaty provides that facilitated access under the Multilateral System shall be provided pursuant to a Standard Material Transfer Agreement (SMTA), which was adopted the Governing Body of the Treaty, in its Resolution 1/2006 of 16 June 2006.
4. Article 3 of the SMTA states:

> "The Plant Genetic Resources for Food and Agriculture specified in Annex 1 to this Agreement (hereinafter referred to as the "Material") and the available related information referred to in Article 5b and in Annex 1 are hereby transferred from the Provider to the Recipient subject to the terms and conditions set out in this Agreement."

5. Article 17 of the International Treaty states that "*Contracting Parties shall cooperate to develop and strengthen a global information system to facilitate the exchange of information, based on existing information systems, on scientific, technical and environmental matters related to plant genetic resources for food and agriculture*".
6. At its Fifth session in Muscat in September 2013, the Governing Body of the International Treaty adopted the Resolution 10/2013, *Development of the Global Information System on plant Genetic Resources in the context of Article 17 of the International Treaty*, and requested the Secretary to call for an expert consultation.
7. In preparation of the expert consultation scheduled on January 2015 in San Diego, California, USA, the Secretariat has requested the preparation of this study as a technical input.
8. The present document is intent to bring light to the importance of plant genomics for food and agriculture and present some suggestions for the consideration of technical experts and does not intent to make recommendations on the decisions that the Governing Body will need to take, but to provide information and technical analysis that may help identify both problems and opportunities, and so support the Consultation in its task of providing advice to the Secretary for the Development of the Vision that will be later on presented to the Governing Body in October 2015.
9. The author would like to thanks the Treaty Secretariat for this opportunity and have invited comments from other experts to further elaborate this preliminary study exploring the role of genomics in its potential impact in the development of the Global Information System.

Plant Genetic Resources for Food and Agriculture and Genomics : Mainstreaming Agricultural Research through Genomics

C*rop improvement is facilitated by harnessing the gene pool of the species and related species to find genotypes and recombine genes to deliver superior plant performance in agriculture, food, energy and biomaterial production.* Henry, R. J. (2011). Next-generation sequencing for understanding and accelerating crop domestication. Briefings in Functional Genomics.

I believe plant breeders and geneticists will drive the next agricultural revolution via the web by sharing the phenotypes and genotypes of crop plants using a system that can store, manage, and allow the retrieval of data. Zamir, D. (2013). Where have all the crop phenotypes gone? PLoS Biology, 11(6), e1001595.

But the real revolutionary potential in this method lies in its power to open up the genetic bottleneck created thousands of years ago when our major crops were first domesticated. Goff, S. A., & Salmeron, J. M. (2004). Back to the future of cereals. Scientific American, 291(2), 42–49.

Table of Contents

Introduction

Motivation

The cost of genome sequencing has fallen one-million fold in the past several years. It is now inexpensive to gather genome sequence information in large numbers of individuals in timeframes much shorter than any crop's life cycle.

In principle, this wealth of genome sequence data should accelerate progress in plant breeding, and thereby help to combat hunger and malnutrition. Integrating the genomic information with crop performance, i.e., plant phenotypes, environment (weather, climate, pathogens) and management practices should transform breeding from being an art to a predictable science. Aggregating and analysing large amounts of genomic and phenotypic data across many environments and treatments would enable to connect genotypes to phenotypes, discover patterns that otherwise remain obscure, and even predict crop performance, enabling smarter choices and faster breeding.

In practice, however, we are not yet organised to seize this extraordinary opportunity. Currently, for the most part, data are collected and studied on a per experiment basis: very focused, under specific circumstances, often with unique material and hard to reproduce. The data remains isolated by crop, by environment, by year, by institution, by company, by country, etc. and is also analysed in isolation often with sample sizes too small to make robust discoveries given the amount of environmental variables. Current procedures in plant breeding do not allow for widespread comparison across studies and the sharing of information. It is hence difficult, merely impossible, to learn across datasets, experiments and breeding trials. The genomic information in its universality can serve as a nucleus and focal point for a much needed integration.

When drafting Article 17, the fathers of the ITPGRFA probably did not quite anticipate the radical technological developments that occurred and are occurring, however, they did appreciate the value of data aggregation and sharing. A **Global Information System**, as called for in Article 17, if implemented with foresight and as soon as possible, will put us on a path to take full advantage of this genomic revolution.
At present, relatively little data on PGRFA have been collected. In absence of an open and interoperable solution, closed, proprietary systems might be created. This would create a fundamental barrier to reaping the benefits of data aggregation and sharing and would hence slow progress.

It should be pointed out that there is another field that is currently revolutionised, i.e., disrupted, by the new genomics approaches: **biomedicine**. Here the goal is to reveal the genetic basis of cancer, inherited disease, infectious diseases and drug responses. Biomedicine is currently also in the need to build an information system for sharing genomic and phenotypic data. Leading researchers in the biomedical community responded to this task in 2013 with the formation of, what is now called, the "Global Alliance for Genomics &

Health (GA4GH)[1]", aka "**Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data**". Their white paper states:

"*[…] the Global Alliance aims to foster an environment of widespread data sharing that is unencumbered by competing, proprietary standards, […]. By creating a standardized framework for sharing and using genomic data, the Global Alliance will enhance the opportunities for broader study of a range of diseases while also improving information sharing globally*"[2].

This Alliance adopted a constitution[3] in Sept 2014 and currently (Oct 2014) has 191 Institutional members from 26 countries, including Google, Inc.. Google in turn recently launched a platform: "Google Genomics"[4], which has the potential to revolutionise the field. Other developments in this area include the "Public Population Project in Genomics (P3P)[5]" and Sage Bionetworks[6]. These initiatives are launched because groups of individuals are convinced of the urgent need and tremendous opportunity.

In contrast to the human medical research community, the PGRFA community is in the favourable position that it has already been agreed at the highest level to develop and implement a Global Information System. Article 17 of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA[7]) provides a framework for it, and a recent survey conducted by the Treaty Secretariat in 2014 indicates that the number of research institutions with genomics programmes on PGRFA is growing, and with it the need for coordination.

While the specific challenges in biomedicine and plant breeding will be different, the underlying organising principle of genomic information and the need to compare genetic and phenotypic variation are the same; for some crops even at the same scale: the human genome and the maize genome have the same size.

The opportunity - The genomics revolution

The biological sciences are currently undergoing a revolution. The genomics revolution is a DNA sequencing revolution. DNA sequencing is a process in which the genetic information, the genome, of an organism is deciphered, i.e., read, letter by letter.

Genetic information contained in the genome is the instruction for life and reading this code is now accessible to everyone. In the past 10 years the cost of DNA sequencing has fallen several orders of magnitude and **Figure 1** illustrates this cost decrease per raw megabase. Incremental improvements to Sanger-type DNA sequencers produced a moderate reduction in sequencing cost since its invention in the 1980s, and it was the

[1] http://genomicsandhealth.org
[2] Global Alliance for Genomics and Health, White paper (2013)
[3] http://genomicsandhealth.org/ga-constitution-about
[4] https://cloud.google.com/genomics/
[5] http://p3g.org
[6] http://sagebase.org/
[7] FAO. International Treaty on Plant Genetic Resources for Food and Agriculture. FAO. (Retrieved from ftp://ftp.fao.org/docrep/fao/011/i0510e/i0510e.pdf in July 2014)

advent of 2nd generation type DNA sequencers in the year 2007 that caused a dramatic drop in price. Main novel features of the 2nd generation sequencing machines were that they sequenced DNA in a highly parallel fashion and that they operated on complex mixtures of DNA molecules as templates.

Hence, anything that has been written about the application of genomics prior to 2007, be it in human medicine, nature conservation, or agriculture, requires revision. Not so much on the chances and opportunities genomics will provide, but certainly on time scales, project sizes, project variety, and Research & Development priorities.

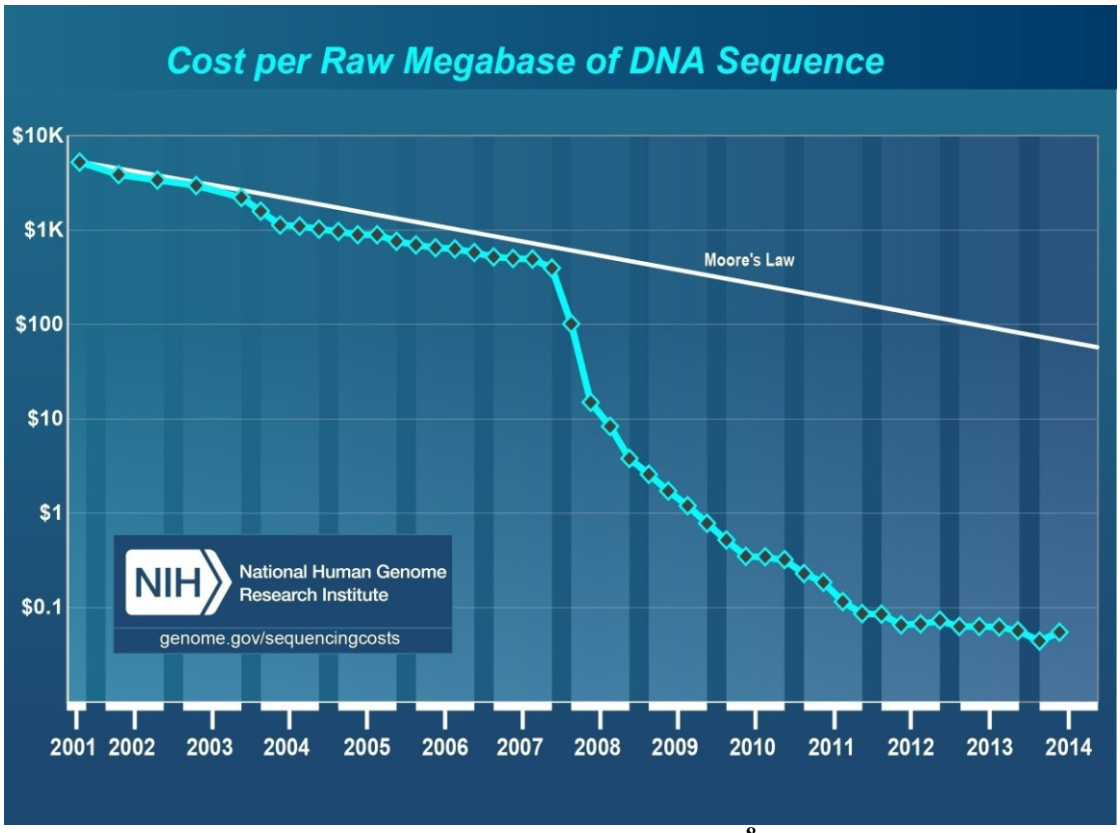

Figure 1: Cost per Raw Megabase of DNA Sequence[8]

As dramatic as the cost reduction was since 2007, close examination of the graph in Figure 1 reveals that the cost reduction has slowed down, plateaued and the cost even increased in recent years. This seems to be due to a combination of technical limitations and economic considerations of the machine manufacturers and technology providers involved. It may mean that further revolutionary improvements to 2nd generation sequencing technology are unlikely and the future price reductions will be mainly from incremental improvements. At the same time, the next generation of DNA sequencing technology –so-called 3rd generation– is emerging, but it will likely be a few more years until the technology reaches maturity and suitable analysis tools and capacity are in place to fully capitalise on the 3rd generation sequencing machines. Hence, genomics will make its big impact in the next few years through a combination of the broad, decentralised application of 2nd generation DNA sequencing technology supplemented by data from more centralised 3rd generation DNA sequencing techniques.

[8] Source: http://www.genome.gov/sequencingcosts

The genomics revolution poses chances and challenges. This document will give an introduction to why and how genomics will make difference in the conservation and use of PGRFA and will highlight the main challenges a Global Information System on PGRFA will need to address in relation to Genomics. The impact genomics will make is global and local, and most of the challenges revolve around fostering worldwide cooperation and interoperability. But because genomic information is the blueprint to life and the basis of inheritance, attaching genomic information to accessions and other PGRFA material and incorporating genomic information into the Global Information System should make the challenge more straight-forward to address rather than more difficult.

The chance

The opportunities that genomic characterisation will bring to the conservation and use of Plant Genetic Resources have been spelled out in detail frequently in the last 15 years.[9] The novelty brought about by the recent advances in genomics is that there are now fast and cheap methods to assess the genetic makeup of an organism, down to base pair resolution, if desired. Large numbers of individuals can now be assayed within timeframes shorter than the lifespan of any crop plant.

The challenge

Acquiring genomic data is cheap, especially the re-sequencing of genomes. There will be an avalanche of data from re-sequencing studies on PGRFA. The challenge is to establish the framework for data aggregation and sharing; in general, and crop specific. Despite the clear benefits of data integration, effective procedures are not yet in place to enable the widespread sharing of information and comparisons across studies on PGRFA. The consultation process established by the Governing Body of the International Treaty for the development of the Global Information System foreseen in Article 17 of the ITPGRFA may help to strengthen commitments and to trigger those procedures, which will generate clear benefits for plant breeding.

The genomics revolution is not expected to make information sharing more difficult, but rather easier. Genomic information holds the promise to **unify the type of information and approaches** and to enable the integration of information across disciplines.

[9] see for example Tanksley, S. D., & McCouch, S. R. (1997) and McCouch, et al. (2013).

Genomes and genetic variation

It is important to realise that the relevant genetic information and variation is not as vast as it may seem and certainly not intractable. As of this year, an estimated 228,000 human genomes have been completely sequenced by researchers around the globe and the number is expected to double every 12 months and reach 1.6 million genomes by 2017. The price of sequencing a single genome has dropped from the $3 billion spent by the original Human Genome Project 13 years ago to as little as $1,000. “The bottleneck now is not the cost—it’s going from a sample to an answer”[10].

The human genome has a (haploid) size of 3 billion base pairs, which is larger than many crop genomes. The 1000 Genomes Project Consortium reported in 2012 on the genetic variation detected by re-sequencing 1092 human genomes.[11] They list 38 million single nucleotide polymorphisms (SNPs), 1.4 million short insertions and deletions (InDels), and 14,000 larger deletions. Their samples where derived from 14 populations and particularly sampled as to maximise diversity.

SNPs are Single Nucleotide Polymorphisms, which means that one letter of DNA code is changed. For example at a particular position in the genome one individual might have an “A” while another has a “G”. A “deletion” is a variant where one or several bases are missing, an insertion when there are extra basses inserted. Whether or not such a variant is an insertion or deletion obviously depends on from what perspective it is viewed. An insertion in one individual can be called a deletion in the other, hence they are often denoted as InDels, which means both or either. InDels can be very large.

Within a species, large parts of the genomes of individuals will be identical. As seen in the human genome example of the 1000 genomes project above: in a diversity maximised sample of highly heterozygous genomes, 38 Million SNPs in 3 billion bp is about 1 variant in 100 bp. For genetics, it is merely these differences between the genomes that are of interest. In addition, most of the genetic variation within a species is shared. This means, not every individual or variety or cultivar (whatever the unit is that is compared) harbours unique genetic variation. For the most part, an individual is the combination of common variation, which presents itself as haplotypes, which are blocks of linked genetic variants. In the case of self-fertilising crop plants, in contrast to humans, the level of heterozygosity is expected to be low. This is certainly true for the mega-varieties of our major crops and it is a feature of a uniform crop, that within a cultivar all individuals are identical. This is probably less so for landraces. Landraces may harbour residual heterozygosity and haplotypes will have frequencies in the population different from 100%. Those frequencies can change from year to year in the field, but nonetheless, the genomes of individuals within a landrace will be a combination of common haplotypes.

It is this **combinatorial nature of genetic diversity** that allows geneticists to detect patterns, and enable Genome Wide Association Studies (GWAS). They do demand, however, to compare large numbers of individuals and cultivars with each other and, because the genetic makeup interacts with the environment, in as many environmental conditions as possible.

[10] Regalado, A. (2014, September 24)
[11] McVean, et al. (2012)

Genomic information will allow to uncouple the haplotypes from the particular individual or variety that was analysed. This makes studies comparable at the haplotype level, even when different sets of individuals and cultivars were analysed. Different experiment will certainly also differ in the environmental conditions the plants experienced. This has always been the challenge of crop phenotyping. Environmental conditions are difficult and often impossible to control. The hope is, that this can be accounted for by large sample numbers and monitoring the actual environmental conditions. Recording high-resolution genomic data on PGRFA in a **Global Information System** will allow to integrate data across experiments, which will facilitate reaching the sample sizes needed to make robust discoveries given the amount of environmental variables.

Genomics

DNA Sequencing

A plant's nuclear genome is organised in chromosomes, which are very long DNA molecules; each is millions and dozens of millions of base pairs (bp) long. So far there is no sequencing technology capable of producing sequence reads the length of an entire eukaryotic chromosome.
At present there exist high-throughput machines that produce **short reads**, which are generally shorter than 400 bp, and "**long read**" sequencing technologies, which can read up to several thousand bp of continuous sequence. The different sequencing technologies have their own unique strengths and are used complementary to one another.

Short read sequencing is cheap, widely available and hence easily accessed. Many universities and research institutes around the world possess their own machines. In addition, there exists an industry of commercial sequencing service providers. **Figure 2** shows a map with 2558 machines situated in 920 sequencing centres.

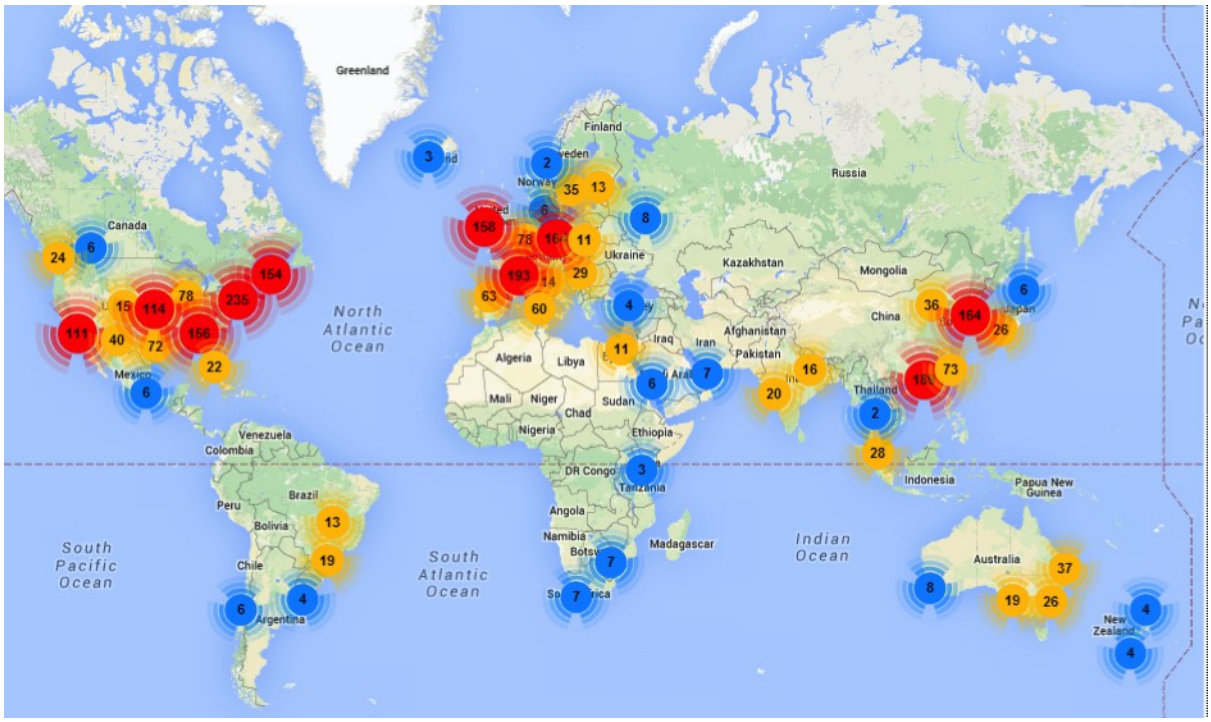

Figure 2: World Map of High-throughput Sequencers[12]

[12] by James Hadfiled and Nick Loman. For an interactive map of global sequencing capacity see http://omicsmaps.com/- (last accessed November 2014).

About 90% of the world's sequencing data today is produced using Illumina's short read technology[13]. These high-throughput machines produce large amounts of reliable data that readily capture SNP and small InDel variation, which accounts for the bulk of the variation within a species. Even though the reads are short, molecular biology and bioinformatic procedures have been developed that are capable of assembling larger tracts of sequence; see chapter "Genome Assembly".

Long read sequencing is currently more expensive and preparing the samples for sequencing is more challenging. Capacity in this area is mainly found in Europe, the USA, Japan and China at specialised research institutes and commercial sequencing service providers. The technology is hence also widely accessible. The main technology in use in mid-2014 to produce long reads was Pacific Bioscience.

Despite the high error rate of the current long-read sequencing process, long reads are valuable, because they enable reading through complex genomic regions and provide higher confidence and precision for calling structural variants, which are large insertions, deletions and rearrangements. With long read information, variants are more easily phased into haplotypes and, for building reference genomes, they improve assemblies into even longer scaffolds, providing order and closing gaps.

Long and short reads together provide a very comprehensive view suitable for most genomes. For exceptionally large, repetitive genomes, **reduced representation sequencing** strategies might be the most cost effective.

Technologies and machines

Currently, mid 2014, there are really only three sequencing technologies available and widely used: Illumina, Ion Torrent, and Pacific Biosciences. The main method of choice is high-throughput short-read sequencing using Illlumina sequencers, complemented by long-read sequencing with PacBio sequencing machines. DOE/JGI operates in this mode since 2012[14], while the Beijing Genomics Institute **(**BGI**)** uses Illumina almost exclusively for its sequencing, with >128 HiSeq2000 machines between the Shenzhen and Hong Kong sites.

[13] Regalado, A. (2014, September 24)
[14] 2013 DOE JGI Progress Report

The most recent account of the history of 2nd generation DNA sequencing instrument development of the last 10 years can be found in McPherson, J. D. (2014).[15] In addition, Sarah Ayling reviewed the different sequencing technologies from a practical perspective in 2013 for the DivSeek Initiative[16].

Short Read sequencing technology

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. Illumina recently released the MiSeq, a bench top type sequencing instrument.
Illumina machines can perform single-end and paired-end runs, where one or both ends of the same DNA fragment are sequenced. Paired-end offers a significant advantage. In genome assembly, the paired reads should be correctly oriented relative to one another and within a certain distance.
The error rate is <1%, and errors are more likely to occur at the 3' end of the reads. Sequencing is performed on eight lanes within a flow cell. Samples can have molecular barcodes added (so-called indexing), so that many samples can be pooled for sequencing. They are separated computationally at a later stage. Illumina currently provides 96 indices, but with user supplied indices the number of samples can be increased at will.
For Illumina sequencing it is rather unlikely that the sequencing cost drops much further, certainly not another order of magnitude. Read lengths might incrementally increase, but certainly also not by an order of magnitude.

Long read technologies, Single molecule sequencers

Current technologies which attempt to sequence very long reads are sometimes called "single-molecule sequencing" or 3rd generation sequencing. They have in common that, as opposed to reading the sequence by monitoring the synthesis of the 2nd DNA strand, a single DNA strand is pulled through a pore, which is just big enough to allow the DNA thread to pass through. This DNA strand is then "read" en passant. Two technologies are currently being used successfully: Pacific BioScience (PacBio) and Oxford Nanopore.

[15] McPherson, J. D. (2014)
[16] Sarah C Ayling (2013), *Technical appraisal of strategic approaches to large-scale germplasm evaluation*.

PacBio® RS II DNA Sequencing System: Single Molecule, Real-Time (SMRT®) Sequencing technology

Pacific Bioscience's (PacBio) RS II system is a single molecule sequencer, which operates in real time. Pacific Biosciences' SMRT Sequencing technology currently achieves the industry's longest read lengths. The P6-C4 Chemistry is advertised with producing 500 Mb to 1 Gb per run in read length of up to 40 kb, with the top 5% of reads longer than 24 kb and 50 % of the reads longer than 14 kb. Sequencing happens by analysing the kinetics of the polymerisation reaction. As part of the sequencing process the technology is able to detect many types of DNA base modifications (e.g., methylation) on-the-fly. The error rate, however, is rather high, 15%. There exist protocols to apply PacBio for RNAseq, which allows the full-length sequencing of (intact) transcripts.

Complementary uses with short read sequencing include to use short Illumina reads to 'correct' the errors in PacBio reads and then assemble these, now long and accurate, reads using the capable, long read assemblers. Alternatively, the assembly can be based on Illumina reads alone and then the long, erroneous PacBio reads are used to determine the correct order of the high quality, but short 'contigs' and fill the gaps in between.

Oxford Nanopore offer their technology in 3 different instruments: MinIon, PromethIon, and GridIon, which all use the same principle, sequencing single molecules by shuttling them trough pores, one molecule per nanopore at a time. The machines differ in size and customability. Read length, accuracy and read number metrics are emerging and the technology looks promising, but is in early stages and not in widespread use (mid 2014).

Sequencing strategies

There are a number of ways to employ DNA sequencing to gain genetic information about an organism. The choice of the approach is currently a balance between desired resolution and cost. The approaches differ mostly in **the actual sequencing cost** and volume of data generated. Instead of shotgun sequencing the entire genome (the default application), methods have been developed to only sequence a fraction. These methods are known as "complexity reduction" or "reduced representation" strategies. However, it is important to realise that sequencing, or applying any sort of genomic analysis for that matter, involves several discrete steps, each incurring costs. For the average plant, the actual sequencing accounts only for a fraction of the cost.

The steps are:

1) the sample (collection/storage/access)
2) DNA isolation (including quality control and storage)
3) preparing sequencing libraries (one or several different libraries)
4) the actual sequencing (i.e., running the sequencing machine)
5) data processing (including data storage and transfer)
6) data analysis (mainly assembly or alignment)

When comparing genomics and DNA sequencing approaches, the total cost need to be considered and as sequencing gets cheaper, the actual sequencing cost might soon be irrelevant. There are several different methods to prepare sequencing libraries. This is a dynamic area and there is competition in the marketplace for ever more effective methods. This is an area were further cost reduction is likely.

Whole genome shotgun re-sequencing to a high genome coverage/depth

This is the default application and the most informative approach. In a whole genome shotgun, DNA gets randomly sheared and a random subset of the resulting fragments is then sequenced. DNA shotgun is a Poisson process[17]. Hence, two parameters improve with increased genome coverage: (1) the average coverage per sequenced nucleotide increases, and with it the confidence in a particular base call. This is particular valuable when sequencing heterozygous genomes or regions. (2) the regions of the genome that did not get sequenced (coverage 0) decreases. At six fold genome coverage (6x) 99.75% of bases of a haploid genome are sequenced. For a heterozygous diploid, a coverage 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 is required.[18]

For genome re-sequencing, one sequencing library needs to be prepared and coverage is increased by simply sequencing more of this library.

Whole genome shotgun sequencing to a low genome coverage

Because genetic variation within a species is common, confidence in variant calls can be attained across individuals. This means that, when sequencing entire populations of similar genomes, each individual can be sequenced to very low coverage without severely affecting the confidence of the variant calls. Due to indexing options with molecular barcodes, arbitrary numbers of individuals can be sequenced simultaneously in one lane on an Illumina

[17] Lander and Waterman (1988)
[18] Wendl and Wilson (2008)

sequencer. This reduces the actual sequencing cost. Other costs remain the same.

Reduced representation methods

The actual sequencing cost can be reduced by sequencing only a fraction of the genome, which, however, reduces the resolution. Since genetic variation segregates in larger blocks, as haplotypes, getting a snapshot every few thousand base pairs along a chromosome can be sufficient for some applications. While the actual sequencing procedure is largely the same, sequencing libraries are prepared differently or standard sequencing libraries undergo additional preparation steps.

Very popular methods of complexity reduction are **GbS** (Genotyping-by-Sequencing) and **RADseq** (Restriction site associated DNA (RAD) sequencing). Both methods require to digest the genome with restriction enzymes, which cut DNA in a sequence specific manner. Subsequently, only the ends of those fragments get sequenced. While RADseq, in principle, yields the end-sequences adjacent to all restriction enzyme sites, GbS further reduces that number to a subset of sites which are in close proximity to one another. This can pose challenges when comparing datasets from different sources as the subsets can differ. In both methods, the resolution is adjusted by choice of the restriction enzyme. When paired-end sequencing is applied to RADseq fragments, the reverse reads of a particular RAD can be assembled into longer sequences, which can be advantageous in some cases. The resulting data can be analysed in different ways. Popular software tools for GbS data are: TASSEL[19] and for RADseq: Stacks[20] and RADtools[21]. The sequence reads may also be aligned to a reference genome.

With **RNAseq**, actively transcribed genes can be sequenced. Instead of DNA, RNA has to be extracted. RNA molecules can then be reverse-transcribed into DNA (so-called cDNA) and sequenced. Since only a very small fraction of the genome is transcribed, RNAseq reduces the sequencing space on the genome. However, it poses a number of challenges, which discourage routine used for genotyping: Handling RNA samples is more involved, because RNA is less stable than DNA and the reverse transcription adds cost. Transcript expression levels are highly variable, and while some molecules are highly abundant, others are missing entirely. Normalisation methods exist, but they again add cost and can negatively impact sequencing quality.

[19] Bradbury et al. (2007)
[20] Catchen et al. *(*2011)
[21] Baxter et al. (2011)

If prior sequence information is available, e.g., reference genome sequence(s), then **target capture** provides an alternative method to reduce the sequence space. Input are standard whole genome shotgun sequencing libraries and regions of interest (targets) are then enriched by capturing them out of the mixture. This method exploits the property that DNA will associate to complimentary DNA strands. Capture kits are commercially available and consist of DNA oligonucleotides attached to magnetic beads. The DNA oligos serve as baits that bind complementary DNA from the genome, and are then recovered using a magnet. The kits can be purchased customised with millions of different oligonucleotides and can, for example, be targeted against the exome, then it is called "Exome Capture". However, arbitrary regions can be chosen as targets for target capture and it can be performed on multiplexed samples. It reduces actual sequencing cost, but does add to the cost and complexity of the sequencing library preparation step. Several companies offer target capture kits (e.g., Agilent, Illumina and Nimblegen).

SNP genotyping

On a whole genome scale, DNA samples can be **genotyped** by hybridisation based approaches. They have been developed for the purpose of high-throughput, cheap **genotyping**. After a custom preparation step, which brings about a complexity reduction and/or labelling of some sort, DNA is hybridised to oligonucleotides housed on high-density arrays, often on glass slides or silicium wavers, called 'chips'. Differential, or selective hybridisation of DNA to these oligonucleotides is the signal that is interpreted into genotypes. The technology does require specialised equipment, however, operation is straight-forward, and can be outsourced. Examples for this technology include Illumina's BeadArray technology (e.g., Infinium HD assay, GoldenGate Genotyping Assay), Affymetrix Axiom Genotyping Arrays and Diversity Arrays Technology (DArT).
The SNP arrays can be customised and for some species predesigned genotyping chips are commercially available off-the-shelf (e.g., Illumina's BeadChips & Bead Sets: e.g., MaizeSNP50. Affymetrix's "Axiom Genotyping Solutions for Agrigenomics": e.g., Lettuce, rice, strawberry, etc.).
Development of these high-density oligonucleotide arrays requires prior knowledge of SNPs so that the marker assays can be developed. This information is usually gathered by DNA sequencing and comparing a number of diverse individuals[22]. Design and production of the assays (chips, or other) requires up-front investment. The approach is hence attractive in cases where

[22] e.g., Ammiraju et al. (2006), McNally et al. (2009)

very large numbers of samples are to be assayed so that the initial investment amortises, or in cases where the genome in question is large. In any case, by design, the approach can only genotype at markers that are known and set a priori; additional variants will remain undetected. And because the chosen SNP-marker set may not represent the diversity of the entire species or population, it carries ascertainment biases.
The available options for crop genotyping with arrays has recently been reviewed in Ganal et al. (2012)[23].
Other popular SNP genotyping assays include Kbioscience's KASPar, based on competitive allele specific PCR (KASP, 2012), TaqMan® Assays (ABI/life technologies), and DArT assays (Diversity Arrays Technology Pty Ltd). It is worth noting that Diversity Arrays Technology Pty Ltd (DArT PL) has adapted the DArT approach to now use DNA Sequencing in place of microarrays to detect presence/absence variations and SNPs[24].

File formats

The data produced by genomics approaches are mainly stored in files, many of them large. The relevant files are text based files and a handful of file types are emerging as de-facto standard for their particular purpose. File types and purpose are briefly described below. Some, more recent file formats such as BAM and VCF are under continued development and their specifications might change in the future. But they are actively maintained and managed, which should ensure backwards compatibility.
The recent survey undertaken between July and August 2014 in the context of the Global Information System on Plant Genetic Resources for Food and Agriculture [25]shows that the listed formats are familiar to the Treaty's community working on genomics research.

DNA Sequence
FASTA, FASTQ

Pure DNA sequence is stored as a continuous word (chain of letters) in a text file in fasta format. A sequence entry has a header line and there can be many sequences in any one file. The fasta file format is also used for RNA and Protein sequences. The universally accepted nucleic acid notation was

[23] Ganal, M. W., et al. (2012)
[24] Sansaloni et al. (2011)
[25] Survey: FAO document ANALYSIS OF THE LANDSCAPE AND GENOMICS SURVEYS IN THE CONTEXT OF THE GLOBAL INFORMATION SYSTEMS

formalized by the International Union of Pure and Applied Chemistry (IUPAC) and is called the "IUPAC code"; A, C, G and T represent the four nucleotides, with additional letters used to express ambiguities; e.g., N means 'any base'.
Often, especially if it is output from a sequencing machine, each baseball, i.e., each letter of sequence that has been read, has a quality score associated with it. The quality score indicates how confident the sequencing process is that the call is correct. This is important information for downstream software such as aligners, assemblers and SNP-callers as they are enabled to resolve discrepancies by weighing conflicting information differently. In a FASTQ file, each individual sequence is represented by 4 lines of text[26]: a header line, the sequence line, the header again, and a line containing the corresponding quality scores.

Annotations
BED, GFF, GFF3, gbk

Annotation is concerned with annotating features encoded by DNA sequences on the DNA sequences. A particular challenge are the representation of nested features. Annotation files are simple text files and conventions have been developed.
In the past, a format called GFF (Genome Feature Format) was widely used, but it has since been superseded by GFF3. It seems that the BED-file format is becoming the preferred format to describe genomic features, which may have to do with its simplicity. The BED file consists of one line per feature, each containing 3-12 columns of data. Three are mandatory, which is 1) the identifier of the DNA sequence the feature is on, and 2) start- and 3) end-position. Additional columns can be added of which column 4 is then interpreted as the name of the feature. There exist tools for conveniently manipulating BED files to curate annotations[27]. GFF3 is a hybrid file which contains a tabular section containing the features and a sequence section containing corresponding DNA sequence[28]. Another type of a hybrid file format is the GenBank file format (gbk)[29]. Genome features can be visualised by genome viewers: the UCSC genome browser[30] uses BED files, while GBROWSE[31] requires GFF3 files, while the Integrative Genomics Viewer (IGV)[32] reads both.

[26] Further information on the format and its history: http://en.wikipedia.org/wiki/FASTA_format
[27] bedtools: http://bedtools.readthedocs.org
[28] For more details on GFF3: http://www.sequenceontology.org/gff3.shtml
[29] http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
[30] http://genome.ucsc.edu
[31] http://www.gbrowse.org/
[32] http://www.broadinstitute.org/igv/

Alignment to a reference genome
SAM, BAM, CRAM, where BAM and CRAM files are compressed versions of Sam.

SAM files are text files produced by software that aligns sequences to a reference genome. These files store the information where a sequencing read can be aligned to the reference genome and how good the individual match is. Sequence and quality information of the original sequence is retained. These files can be very large, even in their compressed forms. SAM files can be compressed to BAM or CRAM files[33].

Sequence Variants
VCF, BCF, where BCF is the binary encoding of VCF, most recent specification is BCF2.

VCF files are text files that hold genetic variation data in tabular format. Variants are SNPs or small insertion or deletions (InDels) compared to a reference. The file can contain variant information for more than one sample; it tabulates variants and samples. The variants and associated features such as allele counts and the confidence and statistical support of the variant call in lines, one line of text per variant, and samples in columns[34].

SAM and VCF file formats where introduced by the 1000 Genomes Project and have since become widely used. It is interesting to note that the specifications were until recently maintained by the 1000 Genomes Project, but the group now leading the management and expansion of the format is the Global Alliance for Genomics and Health Data Working group file format team[35].

Assemblies
FASTG

Genomes are traditionally provided as FASTA files containing the consensus sequence derived from assemblies. This is unsatisfying for some applications,

[33] The most recent specifications can be found at https://github.com/samtools/hts-specs and http://www.htslib.org
[34] The most recent specification can be found at https://github.com/samtools/hts-specs.
[35] http://ga4gh.org/#/fileformats-team

because the information on quality and possible alternative assemblies is not contained. Attempts are made to establish file formats that represent assemblies more directly, as graphs, but they have not gained much traction yet. This may change as new ways or representing and storing population level genome variation data are being explored. One example of a file format that has been suggested to hold sequence assembly graphs is the FASTG format[36].

Example of a standard analysis workflow: re-sequencing one or several individuals

Raw sequencing data comes off the sequencing machine as FASTQ files. Modest 250 million sequencing reads, as are routinely produced by one lane of sequencing on the Illumina HiSeq sequencing machine, means that the resulting text file has 1 billion lines of text. Mapping these reads to a reference genome, which is stored as a FASTA file, will produce a SAM file, with at least one line of text per read indicating where it matched the reference and several lines, if it matched at several locations; it is also recorded if it did not match. The read quality information is also retained in these files, hence SAM files can be really large; dozens of GigaByte for one lane of sequencing. They are large, even when compressed to BAM or CRAM, and file transfer other than within a local network is a challenge. Variants are then called on one or several SAM files to produce a variant file in VCF format. These files are then again small, listing one line per variant.

Data Analysis - Genomic information

Assembly vs. re-sequencing

When using or hearing the term "(**whole) genome sequencing**" it is important to distinguish between two very different things with completely different requirements in terms of preparation of the material, sequencing breath and depths, and analysis:
1) **De-novo sequencing a genome** with the goal of fully **assembling the genome** of a species for the first time, for example as a **reference genome sequence,** and

[36] for specification see http://fastg.sourceforge.net/

2) **Whole-genome re-sequencing** of an individual of a species, where a reference genome sequence already exists, with the goal of identifying the genetic variation, i.e., where it differs from said reference and from other individuals.

Knowing the genome sequence of a species is very valuable. It provides opportunity to explore genes and features present within that species, the organisation of its genome, and allows for comparisons of genomic regions shared with closely related species (comparative genomics). In addition, it provides a reference, to which DNA sequences from other members of the species can be compared against to unravel the genetic variation between individuals and ultimately within a species.
The vast majority of future sequencing data will be genome re-sequencing data. Producing a reference genome sequence will only be required once per species (and sub-genome). Producing a high-quality, finished reference genome, however, is challenging; how challenging depends on the genome in question. As established earlier, a plant's nuclear genome is comprised of large DNA molecules, chromosomes, each millions and dozens of millions of base pairs (bp) long. So far there is no sequencing technology capable of producing sequence reads of the length of an entire eukaryotic chromosome. Therefore, to produce a reference genome sequence, this sequence will need to be **assembled** from much smaller sequence reads in a process called "Genome assembly".

Genome assembly

Since no sequencing technology to date is able to produce chromosome size sequencing reads, the pseudo-molecule sequence has to be produced by assembling shorter reads into longer and ever longer contigs. The most commonly used approach is a whole genome shotgun (WGS) approach where random fragments of the genome get sequenced and then assembled based on them overlapping each other. The sequence of steps is depicted in **Figure 3**.

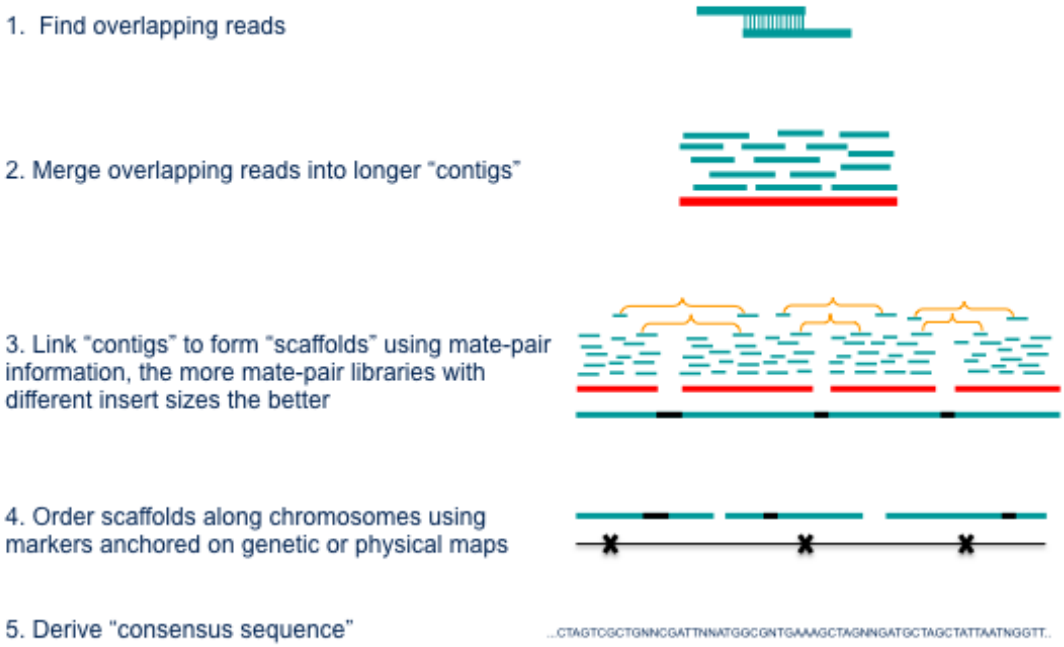

Figure 3: Overview of computational steps involved in genome assembly

In the past, commonly used technologies were Sanger sequencing, which later got supplemented with 454 and Illumina reads. The original human genome sequence (announced completed in 2000) was generated using exclusively Sanger sequencing, which has a low error rate (<1%), at read length of 800bp, but is expensive. The cost of the Human Genome project is frequently said to have amounted to 3 Billion Dollars, or about 1 Dollar per base pair.[37] The depth of sequencing was ~7.5x[38], meaning that, on average, each base of the human genome was sequenced 7.5 times.

To date, the most effective approach is to use the Illumina sequencing platform. The reads are still shorter, but high genome coverage (read depth) compensates for short read lengths in the assembly process. For assembly, a combination of **paired-end** and **mate-pair** libraries with different insert sizes are used: While paired-end refers to sequencing both ends of a small fragment, typically 300-600 bp, mate-pairs are pairs of reads generated from the opposite ends of long DNA fragments, typically in the range of 3-40kb. Knowing the sequence of both ends and the length of the particular DNA fragment greatly increases its utility of short read sequencing for assembly.

To produce mate-pair sequencing libraries requires high-molecular weight DNA and additional equipment, but their advantage is that they span regions

[37] Compare to Figure 1: Cost per Raw Megabase of DNA Sequence
[38] International Human Genome Sequencing Consortium (2001)

that might not have been sequenced or were difficult to assemble with short reads alone due to their repetitive nature. This enables to order otherwise separate contigs into longer scaffolds. For de novo genome sequencing with Illumina, sequencing depths of at least 30x are recommended which should be supplemented by 10-20x coverage from long mate-pairs[39].
The assembly itself is performed by software, most of which use an approach based on de Bruijn graphs (Velvet[40], ABySS[41], SOAPdenovo[42]). Already for medium size genomes these programs can be very memory intensive and require access to large memory machines (e.g., >250Gb RAM).

Genomes assembled solely from 2nd generation sequencing reads (**short reads**) often are highly fragmented. **Long reads** can greatly improve assemblies and hence, in the most recent sequencing projects, the Illumina reads are supplemented by long reads, mostly generated with the Pacific Bioscience platform (PacBio). The currently relatively high error rate of PacBio (15%) is not of much concern in this application as the high coverage achieved through the Illumina reads compensates. Two approaches are currently employed: (1) mapping the short Illumina reads to long PacBio reads to correct errors and then assembling the long, and now corrected, PacBio reads or (2), and this seems the more prevalent, perform a shotgun assembly with the Illumina reads and use the PacBio reads for scaffolding. The resulting draft assemblies are still fragmented, particularly in repeat regions. Hence another valuable addition to a genome assembly are a few thousand BAC-end sequences, which are essentially mate-pair sequences from fragments larger than 100,000 bp. However, producing BAC libraries is non-trivial and is frequently left to specialists and commercial service providers. **Finishing a genome**, that means to obtain continuous sequence of the length of an entire chromosome arm requires additional resources and data; usually a higher order map, such as a physical or genetic map produced by other means.

Several features of plant genomes make genome assembly technically challenging: Among them are **genome size, ploidy, the large size of gene families, repetitive features**, etc. The challenges have been the topic of recent scientific reviews[43].

[39] Schatz, et al. (2010)
[40] Zerbino and Birney (2008)
[41] Simpson et al. (2009)
[42] Li et al. (2010)
[43] E.g., Schatz, M. C., et al (2012) and Morrell, P. L., et al (2011).

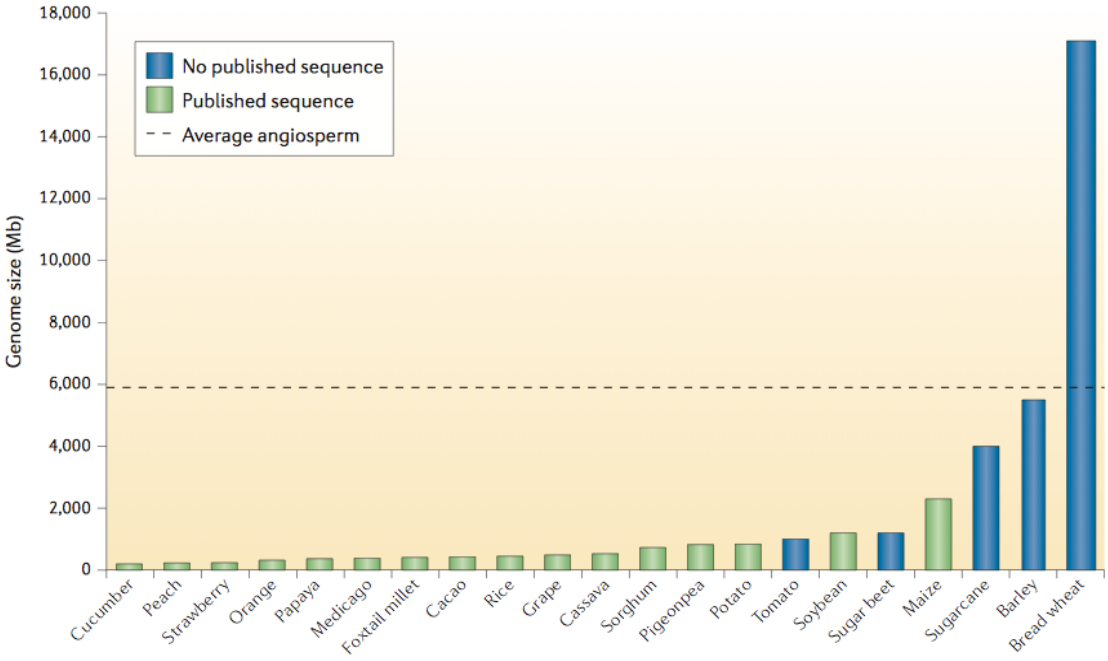

Figure 1 | **Crop genome size.** Genome size of all published crop genomes (shown in green) and the five most important production crops with unpublished genome sequences (shown in blue). The average angiosperm genome size of ~6 Gb is shown by the dotted line for comparison.

Figure 4: Genome sizes of crop species[44]

Plant **genome sizes** span several orders of magnitude[45]. The genome sizes of rice, maize and bread wheat are 400Mb, 2.5Gb, and 17Gb respectively. Larger genomes obviously require more sequencing, increasing the actual sequencing cost. Very large genomes with high ploidy levels such as bread wheat (6n) require very sophisticated approaches: To enable genome sequencing and assembly, the International Wheat Genome Sequencing Consortium (IWGSC) sorted the chromosomes of the hexaploid bread wheat genome by flow-cytometry and sequences the chromosome arms individually.
Genomes contain highly repetitive DNA sequences, such as ribosomal repeats, centromeres, telomeres, and entire families of transposable elements (TE) of varying copy numbers. Some plant genomes are packed with TEs to an extent where they constitute the majority of the genome. Genomes sequenced to date range from 3 to 85% repetitive sequence[46]! Repeats that are longer than the maximum sequence read length cannot be precisely assembled, but rather collapse and break a contig. Repetitive regions longer than the inserts of the longest mate-pair sequencing library are difficult to unambiguously scaffold, especially if there are several repeats of similar sequence. Some repetitive

[44] Figure is taken from Morell et al., 2011.
[45] See Figure 4 and Kew Royal Botanic Garden's C-value database online at http://data.kew.org/cvalues/
[46] Michael, T. P., & Jackson, S. (2013)

elements are missed entirely during the assembly and it should be pointed out that capturing and annotating repetitive genomic components may turn out to be important as they have been shown to function in gene regulation and as structural components of the genome[47].

Genome assembly quality

In recent years a trend is observed from mainly descriptive genome studies with formulaic descriptions of the assembly, gene numbers, repeats, and other genomic features to publish a newly sequenced genome to transport novel biological insights. This shift is mainly driven by demands of publication in high impact journals. The worrying aspect of this trend is that assemblies become less rigorous and result in "sequenced" genomes of limited use to the a broader scientific and plant breeding communities[48]. Incomplete and "error-filled" assemblies then result in erroneous annotations. This effect has recently been quantified by comparing several draft genomes against completed versions of the same sequences, and it was found to be huge[49]. This is worrisome because, increasingly, genomes that have been declared as "sequenced" or "finished" are rarely revisited and improved. The notable exception are the model organisms and among them japonica rice, where an update to the Nipponbare reference genome sequence was released recently, almost 10 years after its original release and publication[50].

One **measure of genome assembly quality** is the contiguity or the length of contigs and scaffolds at which 50% of the assembly can be found; this is commonly referred to as **N50**. An analysis of the first 50 published plant genomes found that genome assemblies conducted by the JGI and based on Sanger sequencing have a median contig N50 for all assemblies of 25.6 kb[51]. Illumina based assemblies, primarily from BGI, have a similar median N50 length (25.9 kb), which points to the fact that assembly strategies using different sized libraries together with the massive sequencing depth of Illumina sequencing begins to be of similar quality. Another measure of a genome assembly is the **amount of the genome captured** in the assembly in per cent (%) of total genome size. While this fraction for individual genomes varies widely, it was found that of all published genomes until 2013, an average plant genome assembly captures 85% of the genome space

[47] Shapiro, J. A., & Sternberg, von, R. (2005)
[48] Michael, T. P., & Jackson, S. (2013)
[49] Denton et al. (2014)
[50] Kawahara (2013)
[51] Michael, T. P., & Jackson, S. (2013)

represented as thousands of contigs with an N50 of 20 kb and tens of scaffolds with an N50 of 1 Mb[52].

We can summarise that today performing a quick shotgun assembly of a genome is cheap and straight-forward, but significant additional work is needed to produce a high-quality assembly and even more to "finish a genome". However, a finished, high-quality reference genome is of enormous value for any crop research, genetics and breeding community. The support of projects aiming at producing high-quality reference genomes should become a funding priority in the short term.

Genome Re-sequencing

For re-sequencing 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 the identified polymorphisms. Ideally, a polymorphism should be identified by several independent sequencing reads covering the variant. This increases the confidence of the call as a single occurrence of a variant might very well be a sequencing error. It has been predicted 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[53].

However, is worth noting here that sequence variation is common. For the purpose of describing species wide diversity and genome wide association studies, i.e., when sequencing large numbers of individuals of a single species, then the coverage per individual can be very low, because the confidence in the call results from the joint coverage from all samples. In other words, increasing sample numbers and diversity can compensate for sequencing depth in any one individual. At low coverage, the genomic information for any given individual will likely have gaps. However, again, since genomic variation tends to occur in the form of haplotypes, the missing information is readily inferred as has been done recently in rice[54]. The same is true, when closely related samples are sequenced together, such as offspring from a bi-parental cross, which is the standard method for genetic mapping either a monogenic or polygenic (QTL) traits: the sequencing depth for each

[52] Michael, T. P., & Jackson, S. (2013)
[53] Wendl MC, Wilson RK (2008)
[54] Huang, et al. (2010)

individual can be extremely low (i.e., less than 1x), because at any given locus there are only 2 haplotypes segregating and segregation was produced by only a small number of recombination events: On average, one recombination event per individual, per generation, per chromosome.

The organised re-sequencing of all accessions currently held in public genebanks is certainly within reach. The human genome has a (haploid) size of 3 billion base pairs. Since humans are highly heterozygous, informative sequencing must assay about twice as much and yet: "As of this year, an estimated 228,000 human genomes have been completely sequenced by researchers around the globe". And the projection is that this number will double every 12 months, reaching 1.6 million human genomes by 2017[55]. The human genome has about the size of maize genome, which is a medium sized crop genome[56]. The re-sequencing of accessions of a given species can be viewed as sequencing of populations where variation is common. The return in terms of new genetic variation detected by additional sequencing will diminish as more samples get sequenced. Hence, the number of samples to undergo whole-genome sequencing to in order to capture the diversity within a species will be much lower than the number of accession actually held in collections.

As the actual sequencing cost is only a fraction of the total cost, genome size will be of lesser concern as prices further drop. However, exceptions at present will be crops with very large genomes, such as bread wheat. For those genomes, complexity reduction methods may be the method of choice for a few more years. Crops with ploidy levels higher than two (diploid) where there are more than 2 similar instances for each locus in the genome also currently pose a challenge. High-quality reference genomes with which the short reads can reliably be anchored to the respective homeolog and the application of long read sequencing technology will be needed to address those.

One of the current frontiers in the field of population re-sequencing studies is **to develop methods for efficient representation of re-sequencing data of population-size datasets**. Current practice in the characterisation of diversity within a species is aligning re-sequencing data to a reference genome (or reference contigs), calling variants in comparison to this reference and then deriving tables that list variants for each individual anchored to the base pair position in said reference. This makes the comparison of individuals and varieties possible and meaningful, but

[55] Regalado, A. (2014, September 24)
[56] See Figure 4 and Morrell, et al. (2011)

can obviously only interrogate sequences present in the reference and cannot capture the full spectrum of genetic diversity. Furthermore these tables will not scale to thousands of individuals.

Ideally, however, every bit of re-sequencing data of a given variety or individual is used to improve our understanding of the genome space of the particular crop species. This is not happening at present, simply because the required smart data structures have not been developed and are hence not in place to facilitate such integration. In addition, future -yet to be developed- data structures should be able to receive and integrate re-sequencing data from many different sources in many different quantities and qualities, because the generation of re-sequencing data will very likely be fragmented and decentralised with many researchers all over the world contributing sequences of varieties and cultivars.

This decentralisation has a parallel in the human genetics field with cancer research leading the development; great expectations are placed on genomics reflected by terms such as "personalised medicine", and "genomics to bedside". As sequencing becomes readily available for hospitals all around the world, they will sequence genomes for the patient's benefit. A tremendous amount of data will be produced, with currently no way to put it or jointly analyse it. The human genetics community realised this shortcoming and leading institutions in the field have forged an alliance with the goal of seizing upon this opportunity. Main task besides addressing the privacy concerns is creating standards for data sharing and integration. The most recent development are graph-like representations of species wide variation[57], which, in principle, could be used for data collection and on-the-fly analysis. It is interesting to note that the plant science community had experimented with such representations in Arabidopsis[58], however, I am unaware of currently active developments in that area in plants. Note that the researchers that currently develop these graph structures for population level sequencing (Gil McVean et al.) are largely the leaders of the Global Alliance for Genomics and Health[59]. New data structures are crucial for what the Alliance is trying to achieve and it will be for the crop science communities. While the needs of the human genetics research communities in regards to representing and sharing genomic data are very similar to the needs in crop genetics, they will be different with respect to the technological details. Levels of heterozygosity, ploidy, effective population size, life history traits and strategies, confidentiality of data, etc. are all

[57] Gil McVean February 18th, 2014
[58] Schneeberger et al. (2009)
[59] http://genomicsandhealth.org

different and the integration of whole genome prediction tools to predict breeding outcomes may not be meaningful in the human context. Supporting the development of data structures that enable pan-genome representation for crop species should be an area of **strategic investment.** The human field is blasting the trail, however, simple adoption of the human approaches will not be possible. The plant crop science community should develop a system very similar and possibly in collaboration with the Global Alliance for Genomics and Health (http://genomicsandhealth.org). The **Global Alliance for Genomics and Health** issued their white paper in 2013[60]. It is remarkable that the plant research community, and in particular crop research is not far behind. January 2014 saw the formation of **DivSeek,** which in the meantime also released a white paper[61]**.** As opposed to the Human genetics community, the PGRFA community already agreed to realise a **Global Information System**.

The Transcriptome

The complete set of RNA transcripts produced by an organism at any one time is the "transcriptome". Transcription denotes the process of synthesizing RNA using DNA as a template, and it is the first step of the "expression" of genomes towards phenotypes. Transcriptomics is the study of "gene expression" (or better: RNA abundance) and its spatial and temporal patterns.

RNA is a macromolecule very similar to DNA. RNA is synthesised by DNA-dependent RNA-polymerases. These enzymes physically move along the DNA and the DNA serves as a 1:1 template for producing the RNA. Similar to a stamp is used to produce imprints, or in photography a negative can produce many positives, the cell uses DNA to cast many RNA molecules. While DNA is mainly present in a cell as a double stranded molecule, RNA is single stranded and highly susceptible to decay by RNA digesting enzymes, i.e., RNAses. Working with RNA requires a more sophisticated laboratory setup, procedures and hygiene to keep RNAses out to prevent RNA decay.

Transcription is normally regulated. An interplay of chromatin structure, epigenetic mechanisms and transcription factors determine whether or not a particular section of the genome is transcribed. With the exception of the basic

[60] Global Alliance for Genomics and Health, White paper (2013)
[61] DivSeek Initiative, White paper (2014)

cellular machinery, most transcription is specific to tissue, time, and developmental stage. There exists genotypic variation in the transcriptome.

RNA molecules can have several roles:
a) A subset of them code for **protein-coding** genes. Hallmark of these RNA molecules is a so-called "open reading frame", which in turn serves as template for protein synthesis. Almost all enzymes, transporters, receptors, transcription factors, etc. in a living organism are proteins.
b) as structural components. Examples are tRNAs and rRNA subunits. The latter are by far the most abundant RNA molecules in a living cell.
c) as regulatory molecules. Some RNAs serve as, or are processed into RNAs with regulatory function, e.g. the entire group of "small RNAs": miRNAs, siRNAs, piRNAs, etc.. They are currently mainly implicated in post-transcriptional control of gene expression and epigenetic mechanisms. Other small RNAs, i.e., snoRNAs, guide the chemical modifications of other RNA molecules.
d) long non-coding RNAs (lncRNAs). In recent years, and mainly through 2nd generations sequencing, a class of RNAs was discovered which is apparently not processed into small RNAs but also does not contain an open reading frame. Their abundance is probably in the tens of thousands, however, so far only a small portion is characterised. They have been found to have a variety of roles[62] and certainly deserve much more attention by research in the future.

The Transcriptome can be assayed with DNA sequencers: Until recently, RNA molecules where detected and quantified on a whole genome scale with array-based technologies. These methods still exist, but the field is now dominated by sequencing. The specific method is called RNAseq. It uses the same instruments that are used to sequence DNA, but requires tailored upstream protocols. Since transcription is the direct readout for what portion of the genome a particular cell is actually using to fulfil its function, the importance of RNAseq for functional genomics cannot be overestimated. With RNAseq, the transcriptome can be assayed without prior knowledge of genes, and in fact the output can be and is frequently used to "annotate" a genome, that is: to identify regions in the genome that are transcribed and can hence be suspected to exert a function. RNAseq is very sensitive. It can be performed on small amounts of input material. There are methods under development to reduce the input requirements down single cell equivalents. Quantification is possible across a large dynamic range with little ambiguity.

Gene "expression" itself, the abundance of the primary gene product in time and space has increasingly been viewed as a phenotype. Genetic studies

[62] http://www.lncrnadb.org/

elucidating the genetic architecture of these expression traits are called eQTL studies.

Transcriptomics - Gene Expression

The value of sequencing transcripts for finding genes and facilitate genome annotation has been discussed above. The following concerns the **quantification of RNA transcripts** produced by an organism at any one time. The complete set is often called the transcriptome. RNA, and especially mRNA, is the link between the genome and the (conventional) phenotype. In the case of protein-coding genes, via a protein complement. Gene transcription in an organism is highly regulated in time and space and the term transcriptomics is often used to denote the study of **gene expression patterns.**

Prior to the advent of high-througput technologies, transcription was assayed on a gene-by-gene basis using PCR based approaches. Generating a comprehensive whole genome view of the transcriptome was until recently the domain of **DNA microarray** technology. While microarrays are still in use, transcription is increasingly measured using RNAseq, which is a term that encompasses various protocols to assay RNA on 2nd generation sequencing machines.

Gene expression patterns can be, and are, **viewed as a phenotypes** themselves and differences in expression patterns between individuals can help to discern complex phenotypes. A method of establishing causal connections between differences in gene expression and the underlying, causal genes is eQTL-mapping.

The ability to assay the whole transcriptome relatively cost-effectively has led to the development of plant transcriptomic resources, often called Gene Expression Atlases. They have an important role in hypothesis generation in basic plant research and can contain important hints for establishing genotype-phenotype associations. Rensink & Buell[63] provide a comprehensive listing of the databases publicly available from genome-wide expression platforms that existed in 2005. They include: Arabidopsis, barley, Brassica, Citrus, grape, maize, Medicago, Populus, potato, rice, soybean, sugarcane, tomato and wheat. Most of these research projects that led to these atlases had been undertaken by consortia, and this is changing. With the advent of

[63] Rensink, W. A., & Buell, C. R. (2005)

RNAseq the resource requirements for undertaking large-scale gene expression studies are no longer prohibitive and a wealth of gene expression data covering a plethora of genotypes, tissue types, from different developmental stages, under different environments and management practices will come online.

To what extend a **Global Information System** should integrate transcriptome information will need to be discussed. Integrating gene expression datasets from the expected large variety of source poses its own set of challenges, because of the large influence of the exact experimental conditions they were obtained in.

Epigenetics

The paradigm is that phenotypic variation is attributable to genetic and environmental variation. Often times this leaves parts of the phenotypic variation unexplained. Epigenetic phenomena might contribute to some of this missing heritability.

Epigenetic effects are caused by chemical modifications to DNA and/or the proteins around which the DNA is packed: the histones. These chemical modifications include *methylation* of cytosine residues in DNA, and *methylation* and *acetylation* of specific amino acid residues in histones. These modifications have bee shown to modulate the probability of DNA transcription and to be causative to epigenetic phenomena such as some forms of gene silencing, the silencing of transposable elements, genome imprinting, and even the transgenerational inheritance of adapted (i.e., acquired) phenotypes[64].

The fields of epigenetics and epigenomics is making fast progress, largely also due to the modern sequencing technologies. DNA methylation can directly be read by a method called *bisulfite sequencing*[65] and histone modifications are readily assayed by *ChIPseq*[66], which is the high-throughput sequencing of DNA fragments bound by modified histones, which can be isolated using antibodies. As with RNAseq, ChiPseq is yet another example how current DNA sequencing technologies are unifying previously separated approaches and assays in molecular biology.

[64] For examples see: Manning et al., (2006), Shivaprasad et al. (2012).
[65] Darst et al. (2010)
[66] Robertson et al., (2007)

It is important to note that the existence of epigenetic phenomena is only another, additional layer of information. It does not change the genotype-phenotype paradigm and can be incorporated into the **Global Information System** where relevant.

Another phenomenon that is accessible to researchers by the new high-throughput sequencing approaches is the actual 3D structure of the genome in the nucleus. Several, related, methods have been established, the most recent is Hi-C[67], which has recently also been applied in plants[68].

[67] van Berkum et al. (2010)
[68] (Wang, C., Liu, C, et al. (2014)

Data sharing

Data sharing - Technical issues

Genomics data (in fact, any kind of -omics data) are rapidly increasing in volume. Current and future approaches that use genomic information for genetic research profit from large sample numbers. It is desirable, and often mandatory for publicly funded research, that genomic date on plant genetic resources is shared, i.e., publicly available. It will foster collaboration, effective knowledge transfer and reproducibility. Efficient data sharing requires robust standards which need to be enforced.

While genotypic differences between individuals can be quite complicated in nature, **genomics data is not particularly diverse**. They are all text files in a handful of file formats that have exact specifications, which are actively maintained and backwards compatible. Those files are produced during analysis by a handful of software programs well known in the field. Running the same analysis again with the same input data and run-time parameters will return the same results. However, the results are susceptible to slight changes in input and/or run-time parameters. This is particularly true for current methods to detect genetic variants. Results can substantially differ when different groups of individuals are analysed together. This is because the software programs must evaluate a variety of parameters to distinguish signal from noise. In different groups of individuals the software will reach different conclusions as to what is signal and what is noise. This can have the consequence that an individual gets assigned different variants, depending on which other individuals it is analysed with. This is of particular concern when sequencing to low coverage in a limited number of samples, which is common practice. Variant callers perform better, and the resulting variants are more reliable, when more samples are analysed together. While the broader research community will be mainly interested in the reference genome, annotation, and the genetic variants, it will not be enough to only store and provide the variant files. Furthermore, the field of detecting genetic variants is dynamic with methods constantly improving. It is hence desirable to be able to reanalyse samples. To that end, the original DNA sequence reads files must be available. To recapitulate a completed analysis will in addition require all relevant information about software programs, version numbers and run-time parameters that had been used.

Plant Science communities have approached the problem of data sharing by establishing online data repositories where information is stored, organised,

maintained and distributed. Value addition through data curation by skilled staff is desirable but highly variable as is the funding. The landscape is fragmented and tends to be organised around a particular crop. The repositories draw on some common methods and software modules but largely had to and did develop their own. The scope, content and usability of each crop repository varies greatly and largely depends on financial means. This is an area in high demand for intervention by policy as targeted funding is needed to develop or adopt procedures, methods, data standards, and example repositories, which can serve as templates and/or skeletons for customised crop repositories in the context of the **Global Information System**. The requirements for different crops will be different and, ideally, customisation is achieved by combining modules.

As mentioned earlier, in terms of storing and making accessible genomic information of population-sized datasets, **new approaches are needed**. Current methods do not scale. When framing the Global Information System, the scientific and technical community working under the umbrella of the International Treaty will be able to learn from experience form other online repositories, examples are listed below, and from existing initiatives. The most significant are DivSeek, The Global Alliance for Genomic and Health, the African Orphan Crop Consortium, iPlant, transPLANT and Google, Inc..

Current Access to Plant Genomes

General Plant Genome Resources (all accessed 24 July 2014)

https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes
http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html
https://genomevolution.org/CoGe/
http://gramene.org/
http://jgi.doe.gov/data-and-tools/
http://www.phytozome.net/
http://bioinformatics.psb.ugent.be/plaza/
http://mips.helmholtz-muenchen.de/plant/genomes.jsp
http://www.plantgdb.org/
http://pgn.cornell.edu/
http://www.gramene.org/

Plant/Crop Specific Genome Resources (all accessed 24 July 2014)

Arabidopsis
http://www.arabidopsis.org/
http://1001genomes.org/

Maize
http://www.maizegdb.org/
http://www.maizesequence.org
Rice
http://rice.genomics.org.cn
http://rice.plantbiology.msu.edu/
Cassava
http://www.cassavabase.org/
Tomato
http://solgenomics.net/

Data sharing - other issues

Intellectual property rights

An important consequence of the sharp drop in price of DNA sequencing has been and is the **democratisation of sequencing,** moving large-scale sequencing out of the few large genome centres, which had been the major contributors of genomic information in the past. This trend will continue and likely accelerate with the use of bench-top sized sequencing instruments that are now affordable even for small labs and with the increasing DNA sequencing capacity available through commercial DNA sequencing providers. DNA sequence data on plant genetic resources can now be attained/produced by anyone, everywhere. Channelling this, truly globally, decentrally produced genomic information into a Global Information System has technical challenges, as discussed above, but it also requires addressing the issues of **Intellectual Property Right**s (IPR), freedom to operate and access and benefit sharing.

It has been argued that "genome sequences (and their functional characterisation) may be deemed to be international public goods, particularly for those crops covered under the multilateral system of the ITPGRFA.[69]".

From a genomic science perspective, broad participation in data sharing is key to reach large sample numbers. Uncertainty with respect to IPR will certainly hinder participation in data sharing, hence addressing the **Intellectual Property Rights associated with genomic information** gathered from PGR is central to encourage participation. Anyone holding DNA sequence data

[69] Robin Fears (2007)

should be able to make it publicly available without risking legal consequences. If this is not the case, then participation might suffer. From a technical standpoint, it is not desirable to require **curation of data** prior to upload, because a requirement of curation will **increases the cost** of data sharing. Data upload from sequencing machines into public repositories will soon likely be an automated process. In addition, since much of the downstream analysis on the data is of statistical nature, the information contained in the **Global Information System** should be as correct, complete and unbiased as possible. Selective **omissions and removal of data** should be avoided as it may distort analysis and results.

As the example from the free, and **open-source software movements** shows, the sum of global, decentralised and small contributions frequently leads to outcomes comparable and often superior to commercial products. Examples are the operating system Linux with its plethora of applications and the web browser Mozilla Firefox, currently the most popular web browser. A key enabling invention was the GNU general public license[70], which requires published software to be free, the software's source code to be open and attaches copyleft to this code. Adopting an enabling set of rules allowing and demanding the free sharing and use of genomic information related to PGRFA should be discussed, as well as the consequences under the existing regulatory frameworks and options when no policy decision is made.

Access and Benefit Sharing

Most genetic variation is common within a species and populations differ merely in their allele frequencies. With high-resolution genomic information it is straight-forward to infer population structure and pedigrees between individuals of a species and hence between varieties and cultivars of a plant genetic resource. These inferences are based on the amount of shared genetic variation and the coalescence of haplotypes. It is hence possible to determine the origin of particular variants and haplotypes, at least within the space of samples for which high-resolution genomic information is available. It is then, in principle, possible to attempt to quantify the contribution, in terms of genetic variation, a particular individual, variety, landrace, breeding line, etc., has made to the global gene pool and/or to breeding programs. In this way, genomics may be used to assign value to PGRFA and thus may influence the way policy makers regulate access and benefit-sharing and impact ongoing and future negotiations. It will need to be discussed whether quantifying genetic variation in this context is desirable.

[70] http://www.gnu.org/licenses/

Cyberinfrastructures for Analysis of Genomic Data of Plants

Governments around the world have realised the pressing computational and data sharing need of the genomics community. There are publicly funded cyberinfrastructures in place especially geared towards genomic data and plant research. Most notably are **iPlant** (USA) and **transPlant** (EU). Particularly geared towards breeders is the **Integrated Breeding Platform** (IBP), which is in turn hosted by iPlant. Brief descriptions of these cyberinfrastructures are provided below. They will be ideal partners for implementation of a Global Information System for Plant Genetic Resources.

iPlant[71]

Country: USA, funded by NSF

iPlant:
IPlant is a "virtual organization lead by The University of Arizona, Texas Advanced Computing Center, Cold Spring Harbor Laboratory, and University of North Carolina at Wilmington. It was established by the U.S. National Science Foundation (NSF) in 2008 to develop cyberinfrastructure for life sciences research and democratize access to U.S. supercomputing capabilities. iPlant develops the national (USA) cyberinfrastructure for data-intensive biology driven by high-throughput sequencing, phenotypic and environmental datasets. It wants to provide powerful extensible platforms for data storage (including cloud services), data exploration (bioinformatics analysis, image analyses), data exchange and APIs. iPlant makes broadly applicable cyberinfrastructure resources available across the life science disciplines (e.g., plants, animals, and microbes). iPlant deliberately invites "feedback"; it understands itself as community driven: *iPlant is of, by, and for the community; community-driven needs and requirements shape and focus iPlant's mission*.

The "iPlant Collaborative" was set-up in 2008 as a US-$ 50 million, five-year project to create the cyber- (or computer-) infrastructure needed to tackle "grand challenge" questions in plant biology. It got since renewed. *iPlant works with the community to support the storage, access, and analyses of data for collaborative and individual research* [...] *and fosters innovation across the biological, education, and computer science communities through its education, outreach and training activities*. Overarching motivation is the challenge to feed the growing human population, *while the amount of available farmland decreases, food cultivation competes with fuel production and climate change and energy*

[71] http://www.iplantcollaborative.org

sustainability impact agriculture, ecology, and biodiversity. iPlant reasons that *in order to successfully address these issues, we need to understand the mechanisms through which organisms' appearance, physiology and behaviour are shaped by the interactions between their genetic makeup and the environment*.

Acknowledging that this is mainly a big-data problem, iPlant's goal is to "*enable biologists to do data-driven science by providing them with powerful computational infrastructure for handling huge datasets and complex analyses*. [...] *iPlant does not set, nor pursue, its own scientific agenda, but rather builds an infrastructure that allows community members to pursue their own ends, in collaboration with the project and, more importantly, with each other*.
Originally put on place as cyberinfrastructure tailored for the plant science research community, following a recommendation of the NSF, iPlant has extended its scope beyond plants. iPlant provides cyberinfrastructure services for a large number of projects[72], including, the Arabidopsis Information Portal (AIP), BigPlant, the Integrated Breeding Platform (IBP), Bioextract Server, Galaxy.org, CIPRES, CoGe, Gramene, NEVP, 1,000 Plants (OneKP), BIEN, MaizeGDB, SoyKB and IRRI; and has plans to do so with ICRISAT, the African Orphan Crops Consortium (AOCC) and AgMIP.

iPlant, is actually a member of the African Orphan Crops Consortium (AOCC), where it provides storage for sequencing data and a partner in the 1000 Plant Transcriptomes (1KP Project) project, which aims to sequence and assemble the mRNA complement of over 1,000 plant species. There are partnerships in place with KBase, NCGAS, ELIXIR, and transPLANT.

transPLANT[73]

Country: EU, funded by EU

transPLANT is a consortium of 11 European partners gathered to develop a trans-national infrastructure for **plant genomic science**. It is in place to provide computational resources and address the challenges arising from "the quantity, diversity and dispersed nature of data in need of integration". transPLANT is "committed to establishing the broadest possible international collaborations for data and standards".

Partners[74] include research institutions: EMBL-EBI (Europe), the Helmholtz Gemeinschaft (Germany), the Gregor Mendel Institute for Molecular Plant

[72] source http://www.iPlantcollaborative.org/about-iPlant/the-organization/strategic-initiatives
[73] http://www.transplantdb.eu/
[74] Source: http://www.transplantdb.eu/partners

Biology (Austria), IPK Gatersleben (Germany), INRA (France), iGRpan (Poland), Plant Research International Wageningen (Netherlands), TGAC of BBSRC (UK), Barcelona Supercomputing Centre (Spain), and private sector companies: biogemma (France), keygene (Netherlands).

The motivation behind **transPLANT** are the "significant opportunities for crop improvement through plant breeding and increased understanding of plant biology" that are "opening up" due to the "falling cost of nucleotide sequencing. Many plant genomes are large and have complex evolutionary histories, making their analysis theoretically challenging and highly demanding of computational resources. Issues include genome size, polyploidy, and the quantity, diversity and dispersed nature of data in need of integration […] transPLANT will develop integrated standards and services, undertaking research and development to capitalise on the sequencing revolution across the spectrum of agricultural and model plant species." transPLANT is "committed to establishing the broadest possible international collaborations for data and standards".

Project objectives[75] range from Research and Development, database and storage to community outreach. Specific goals are to "develop a computational infrastructure for plant genomic science, a portal to provide integrated, interactive access to a broad range of databases, services and tools, develop new methods for the large-scale analysis of genotype-phenotype associations and for genomic analysis, develop and maintain a common set of reference plant genomic data, **explore the mechanisms required for the analysis and storage of genomic complexity in plant species**, develop a new infrastructure for the archiving of genomic variation, provide a new search engine, integrating reference bioinformatics databases and physical genetic materials". In order "to endorse and develop standards for extant and emerging data types" transPLANT's aims to engage with the widest possible communities." transPLANT reaches out via a series of networking activities and meetings with experts from related fields, to exploit experiences and explore synergies, host training events to familiarise the plant science community with the use of cutting edge tools."

The Integrated Breeding Platform (IBP)[76]

[75] http://www.transplantdb.eu/project
[76] https://www.integratedbreeding.net/

The Integrated Breeding Platform (IBP) is a web-based one-stop shop for information, analytical tools and related services to design and carry out integrated breeding projects. It is hosted by the "iPlant Collaborative" and jointly funded by the Bill&Melinda Gates foundation, UKaid, the European commission, the CGIAR and the Swiss Agency for Development and Cooperation (SDC). IBP is coordinated by the CGIAR Generation Challenge Programme, the development of the Platform is a project bringing together numerous partners and several key funders.

IBP's centrepiece is the Breeding Management System (BMS, previously known as the Integrated Breeding Workflow System, IBWS). Development partners are the iPlant Collaborative, Plant Research International of the Wageningen University and Research Centre, Software development service provider "Leafnode" (private company) and data analysis software supplier VSN International Limited (VSNi, private company).

IBP is conceived as a vehicle for dissemination of knowledge and technology, enabling broad access to and proactive distribution of crop genetic stocks and breeding material; molecular, genomics and informatics technology and information; cost-effective high-throughput laboratory services; and capacity building programmes. Its primary clients are developing-country breeders.

The following text is taken from the IBP website:
"The Integrated Breeding Platform (IBP) is conceived to help plant breeders accelerate the creation and delivery of new crop varieties in the context of an increasing global demand for food. It does so by giving access to vanguard technology and quality services – for both traditional and modern breeding activities – at low to no cost. As cornerstone elements of its deployment approach, the IBP also provides training opportunities, responsive technical support and community space for meaningful exchanges with peers and other experts. The IBP is not a simple software or service provider. It is firmly committed in democratising and facilitating the adoption of today's tools for tomorrow's crops by plant breeders across world regions and economies, anywhere from emerging national programmes to well-established companies. To that end, you will find essential resources on our pages to optimise modernise your plant breeding programme: downloadable, comprehensive software tools: the Breeding Management System (BMS) and more tools from our partners a network of accessible and reliable breeding service providers; a resource library with products and information for over 10 crops, including diagnostic markers and trait dictionaries; training material and activities for an optimal use of our technology as well as for integrating good breeding practises; support through peer communities and dedicated technical assistance."

Other platforms

KBase, US Department of Energy Systems Biology Knowledgebase
The US Department of Energy Systems Biology Knowledgebase (KBase) is a large-scale software and data platform that aims to enable researchers to predict and ultimately design biological function. KBase enables secure sharing of data, tools, and conclusions in a unified, extensible system that allows researchers to collaboratively generate, test, and share hypotheses about molecular and cellular functions. KBase is not currently a data repository. It relies on and interacts with existing public databases. At KBase, users are enabled to *discover genetic variations within plant populations and map these to complex organismal traits*. KBase is an open platform where external developers can integrate their analysis tools.

Bioplatforms Australia[77]
Bioplatforms Australia provides services and scientific infrastructure in the specialist fields of genomics, proteomics, metabolomics and bioinformatics. It supports Australian life science research with crucial investments in state-of-the-art technologies and cutting edge expertise[78].

[77] http://www.bioplatforms.com.au
[78] more information: http://www.bioplatforms.com.au/about-us/what-we-do

Relevant Initiatives

DivSeek

http://www.divseek.org

DivSeek is an emerging consortium with the goal "to unlock the potential of crop diversity stored in genebanks around the globe and make it available to all so that it can be utilised to enhance the productivity, sustainability and resilience of crops and agricultural systems."[79]. "DivSeek will bring together genebanks, breeders, plant and crop scientists, database and computational experts to enhance the use of genebank materials, promote effective utilization of genetic variation in plant improvement, and to better understand how components of genetic variation contribute to plant performance (i.e., growth, development, yield and nutritional composition) in diverse climatic environments[80]."

The Global Crop Diversity Trust[81] hosts and implements the facilitation unit of DivSeek jointly with the Secretariat of the International Treaty on Plant Genetic Resources for Food and Agriculture[82], and operates it on a day to day basis with additional inputs provided by the CGIAR Consortium, the Global Plant Council[83] and other experts. DivSeek wants to facilitate the curation, integration and utilisation of relevant data and germplasm, and promote international exchange.

In their white paper[84], DivSeek outlined a road map detailing 2 phases over 5 years and in the first phase hopes to subsequently: first, connect the physical germplasm resource, and associated passport information of Plant Genetic Resource stored in genebanks with genomic/genotypic information, then second, intends "to enrich genebank data with large-scale phenotypic data"; these linkages between "germplasm, genotypic and phenotypic information" can then be used to, third, facilitate "allele mining". By running pilot projects, DivSeek will establish common protocols for said data integration during this first phase.

It appears that the DivSeek Initiative will address many of the same genomics-related issues that the **Global Information System** will need to

[79] http://www.divseek.org/mission-and-goals/
[80] http://www.divseek.org/who-are-we/
[81] http://www.croptrust.org
[82] FAO, http://www.planttreaty.org
[83] http://globalplantcouncil.org
[84] DivSeek Initiative, White paper (2014)

address. The necessary technical coordination mechanisms with DivSeek will need to be put in place or strengthened as the initiative evolves.

Global Alliance for Genomics and Health (GA4GH)

http://genomicsandhealth.org

From the GA4GH website[85]:

"What is the Global Alliance? The Global Alliance for Genomics and Health (Global Alliance) is an international coalition, dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing. The promise of genomic data to revolutionize biology and medicine depends critically on our ability to make comparisons across millions of human genome sequences, but this requires coordination across organizations, methods, diseases, and even countries. The members of the Global Alliance for Genomics and Health are working together to create interoperable approaches and catalyze initiatives that will help unlock the great potential of genomic data."
"What is the Global Alliance doing? Since its formation in 2013, the Global Alliance for Genomics and Health is leading the way to enable genomic and clinical data sharing. The Alliance's Working Groups are producing high-impact deliverables to ensure such responsible sharing is possible, such as developing a Framework for Data Sharing[86] to guide governance and research and a Genomics API[87] to allow for the interoperable exchange of data. The Working Groups are also catalyzing key collaborative projects that aim to share real-world data […]".

The GA4GH published a white paper[88] in 2013 and adopted a constitution[89] in Sept 2014 and has currently (Oct, 2014) 191 Institutional members from 26 countries, including Google Inc.. The GA4GH has established working groups to address the challenges of sharing genomic information and launched a series of targeted initiatives, among them a "Metadata Task Team", a "File Formats Task Team", an initiative towards "Phenotype Ontologies, etc.. The full list of current initiatives can be found are listed on their website[90]. The GA4GH is formed and headed by the world's leading researchers in the field

[85] http://genomicsandhealth.org
[86] http://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data
[87] http://ga4gh.org/#/api
[88] Global Alliance for Genomics and Health, White paper (2013)
[89] http://genomicsandhealth.org/ga-constitution-about
[90] http://genomicsandhealth.org/our-work/initiatives

of human genetics. Whatever technical solutions they come up with is likely to be the standard for years to come.

African Orphan Crop Consortium

http://www.mars.com/global/african-orphan-crops.aspx

The African Orphan Crops Consortium (AOCC) is an international effort. The goal of the consortium is "to sequence, assemble and annotate the genomes of 100 traditional African food crops, which will enable higher nutritional content for society over the decades to come. The resulting information will be put into the public domain, with the endorsement of the African Union". The consortium is lead by Mars, Incorporated and has numerous high profile members including: The African Union - New Partnership for Africa's Development (NEPAD), LGC (for Plant sampling kits, DNA extraction and KASP genotyping chemistry), World Agroforestry Centre, BGI (who will do the initial sequencing), Life Technologies Corporation (donor of sequencing equipment), World Wildlife Fund (WWF), University of California, Davis, iPlant Collaborative (which will manage the data produced), Biosciences eastern and central Africa, International Livestock Research Institute (ILRI).

It should be mentioned that Mars, Incorporated had organised a whole genome project before and delivered in 2010 the sequenced, assembled and annotated cacao (cocoa) genome and made these data publicly available on the internet open to the public. Given their experience, and high profile partners, there is little doubt that the projects will succeed.

Impact of Genomics on Plant Genetic Resources for Food and Agriculture (PGRFA)

The chapter "Genomics" established that by using current genomics approaches the genetic make-up of an individual organism and the genetic variation within and between species can be known to unprecedented detail, down to nucleotide resolution if required. This knowledge has potential impact on the management of Plant Genetic Resources for Food and Agriculture (**PGRFA**) in genebanks and is changing the process of plant breeding. The genetic information is also an effective integrator with which the plant breeding, research and genebank communities can interact. This impact and the interactions are illustrated in **Figure 4 The central role of genomics for Research and Development on Plant Genetic Resources for Food and Agriculture**. Two major research cycles on Plant Genetic Resources are highlighted: the genebank management research cycle (blue) and the research cycle on applied plant breeding (red). Some of the many interactions are indicated by connections and arrows. The figure is supposed to illustrate the central role **genomics** is expected to play in facilitating advanced breeding, effective **pre-breeding** as well as **targeted breeding**, and **informed genebank management**. The message is, that, with comprehensive genotype information, the genetic data can serve as an integrator.

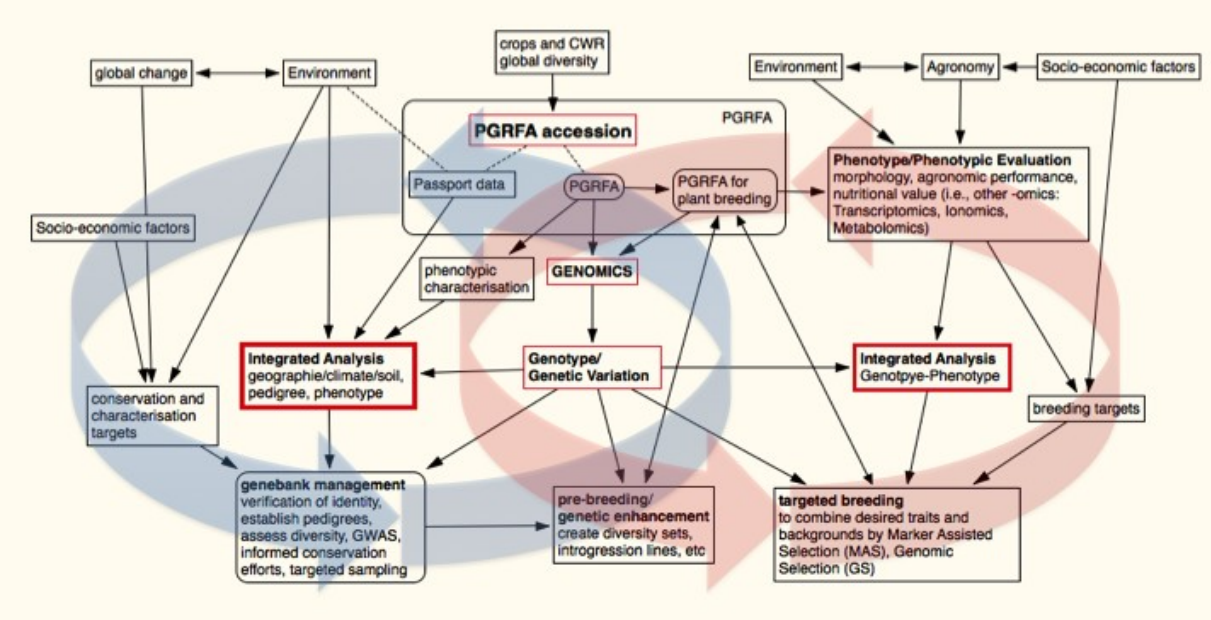

Figure 4: The central role of genomics for Research and Development on Plant Genetic Resources for Food and Agriculture

Explanatory text for **Figure 4: The central role of genomics for Research and**

Development on Plant Genetic Resources for Food and Agriculture; Key terms from the figure are printed in bold font in the text.

PGRFA accession: An accession of a Plant Genetic Resource for Food and Agriculture held in a genebank (**PGRFA accession**) is a bag of material (**PGRFA**) that can be used to reproduce said accession, i.e., seeds or tissue for vegetative propagation. It is drawn from the **global diversity** of crop plants and crop wild relatives (**CWR**) and usually enters the system of Plant Genetic Resources for Food and Agriculture (**PGRFA**) by being collected from farmer's fields or in the wild. Every accession has **passport data** associated with it, which at minimum contains information about its origin. Time and location of the collection links the **PGR accession** to the actual **environment** it is/was living in.
From an accession bag stored in a genebank (**PGR**), material can be drawn for characterisation or plant breeding and, in the illustration, it becomes a **PGR for plant breeding**. In case it is sent out to researchers or plant breeders it leaves the genebank. Using **GENOMICS**, the genome of a **PGR for plant breeding** can be compared at any time to the genome of its source **PGR**, to establish identity and differences. If the information where a particular **PGR for plant breeding** originated from is lost or unknown, **Genomics** can be used to identify the most likely source PGR. This requires a comprehensive catalog of **PGR** and their associated **Genotype** information. Given such a catalog, detailed pedigrees of all **PGRFA** can be established.

Genebank management research cycle (blue):
G**enotype** data of **PGRFA**s on their own already allow for establishing pedigrees and analyse relatedness of **PGRFA**s in the collections. Clones, duplicate entries and mislabels are readily identified, which allows to consolidate the collection, which should reduce operation cost. **Passport data,** especially the geographic information, allows for a plethora of inference about the environmental envelope a particular crop and its varieties/cultivars exist in and thrive or strive. Information of past and current distribution of a **PGRFA** is used to detect gaps in the collection based on geography. Adding **genotype** data to allows to map the **genetic variation** of a crop onto the geographic landscape. This makes obvious regions of high and low genetic diversity and helps to prioritise future collection efforts. Last but not least, based on genotypic information the maximum informative allele combinations can be selected for phenotypic evaluations by the gene bank.

Information on the species-wide **genetic variation** then allows for efficient **pre-breeding** and **genetic enhancement**. This is an area where genebank management and plant breeding overlap. Ideally, a catalog exists of **PGRFA** and their associated **genotypes**. Based on this information about **genetic variation** within a crop, the maximal informative sets of **PGRFA** for phenotypic evaluation can be selected, or created and these diversity sets and introgression lines can then enter the breeding research cycle, denoted in the illustration as "**PGRFA for plant breeding**".

Applied, targeted Plant Breeding research cycle (red):
Targeted breeding is concerned with phenotype(s). Starting point of analysis are phenotypic evaluations of a variety of genotypes in a variety of environments and/or under different agronomic management practices. With **Genomics**, genotype and phenotype data can be analysed together by methods such as genetic mapping, QTL-mapping and Genome-Wide Association Studies (GWAS). This yields **Genotype-**

Phenotype associations, which establish causality between a particular trait and one or several loci in the genome. With this information the breeding outcomes can be predicted, breeding material for particular breeding targets can be selected in-silico, and it provides markers to reliably follow the inheritance of many traits simultaneously through the generations and will hence greatly reduce phenotyping requirements in the process. All of the above will vastly accelerate breeding by shortening the breeding cycle, reducing cost and making accessible PGRFA otherwise overlooked due the desired phenotypes being masked.

The impact of genomics on genebank management

The impact of genomics on genebank management, including the technical aspects, has been thoroughly reviewed last year (2013) for the **DivSeek Initiative**[91]. A major goal of genebank management is to assess how much of the global diversity is present in the collection and to help identify accessions, which can contribute traits of interest to breeding programs for the respective crop. Genomics will be another instrument available to genebank managers besides analysing passport information and phenotypic characterisation. It will be an enabling technology and will reduce operating cost.

Phenotypic characterisation of genebank material is essential. Entire collections can be and have been screened for specific traits. But phenotyping entire collections for many traits, repeatedly in many environments quickly becomes a prohibitively large task. In addition, phenotypic evaluation frequently misses valuable variation, because alleles can be masked in particular genetic backgrounds.

Genotyping collections on the other hand is much more tractable and is now more cost-effective than ever through next generation sequencing technologies. While the resolution will depend on the strategy, the genotype data alone can already be used to **measure diversity** within collections. Genomics will reveal the genotype independent of origin of the material or how it was labelled. After entire collections have undergone genomic characterisation it is straightforward to confirm the identity of accessions, build pedigrees, and reveal misidentifications and mislabellings. Genomic information can be screened for signatures of selection pointing out regions in the genome of possible importance for the breeding process, e.g. domestication genes.

With genotype information, the most informative sets of accessions for a phenotypic characterisation projects can be selected. In cases where the desired allele combinations to be tested by phenotyping are not realised within any accession, the best suited allele donors for test crosses can be cherry-picked. More generally, diversity sets for phenotyping can be selected, or created, to maximise allelic diversity and for replication purposes in phenotypic trials. Creating diversity-sets and introgression lines by targeted crosses and making otherwise considered "exotic"

[91] Sarah C Ayling, 2013, *Technical appraisal of strategic approaches to large-scale germplasm evaluation*

germplasm available to plant breeding is often referred to as **pre-breeding**[92].

Combining genotype information with passport data, especially with information about where and when an accession was collected, allows to map the genetic variation onto the landscape. "With detailed original source information, genetic assessments of germplasm collections can go beyond the basic measurements of collection diversity and breeding for simple traits to assessments of natural variation in environmental contexts."[93]. Volk and Richards give detailed suggestions on what passport data to record and how to place "functional variation into a spatial context", which can then "lead to a more complete understanding of genes that result in adaptation" and to unravel the "natural variation of potential use for agriculture"[94]. Joining spatial data with genomics information quickly reveals where the centres of diversity are, where collection gaps exist, and in which areas sampling was exhaustive. This will help to prioritise future collection efforts. Overlaying this map of genetic diversity on the landscape with past biotic and abiotic stresses, such as climate scenarios, disease pressure or the genetic variation of a pathogen of interest, allows for targeted identification of populations in which valuable genes and alleles to combat these stresses most likely segregate and what those genes and alleles are.

Combining genotypic information with phenotypic data enables the identification of causal relationships between genomic regions and associated phenotypes using routine forward genetic approaches such as genetic mapping, QTL mapping and Genome Wide Association Studies (GWAS). This knowledge is very powerful, as it helps to overcome several constrains of phenotypic evaluation. Accessions can then be selected for breeding programs based on genotype. In principle then, based on these associations, attempts can be made to predict the phenotype of unobserved genotypes.

A **Global Information System** should reliably link an accession stored in a genebank and its associated passport data to its genomic information. Ideally, entire genebank collections undergo genomic characterisation. Very detailed suggestions have been made on the technical aspects of how to interconnect passport and genomics data[95]. To enable associating traits to genes, it is desirable that the PGRFA and the genomic information is linked to the results from all phenotypic evaluation trials this material ever underwent; ideally, this reaches out into the plant breeding domain. It is hence desirable that

[92] http://www.croptrust.org/content/pre-breeding
[93] Volk, G. M., & Richards, C. M. (2011)
[94] Volk, G. M., & Richards, C. M. (2011)
[95] Finkers et al. (2014), Volk, G. M., & Richards, C. M. (2011)

descendant material of a PGRFA in a genebank can be traced through the Global Information System, including material that leaves the genebank; this will be especially valuable, if a PGRFA for which high-density genotype data exists, enters the plant breeding domain and undergoes additional phenotypic evaluations. While relatedness of individuals can be easily established using genomic tools, it requires access to tissue of the individuals in question, which his not always practical. In order to compare phenotypic datasets, it would hence be helpful to record identity and descent by an identifier (ID). Ideally, the ID system aids tracing PGRFA between the genebank and breeding domains and between breeding programs. It should be considered to routinely genotypical characterise PGRFA that leave genebanks.

The impact of genomics on plant breeding

The opportunities that genomic characterisation will bring to the conservation and use of Plant Genetic Resources including plant breeding have been spelled out in detail frequently in the last 15 years[96]. The hope is, that with the modern genomic tools, supported by an effective **Global Information System**, plant breeding can draw from a greater variety of PGRFA, and will become more predictive and hence faster and cheaper.

It is the within-species genetic diversity that is most readily exploited by plant breeding. Species are usually reproductively isolated, which makes the inclusion of closely related, but different species, so-called Crop Wild Relatives (CWR) into a breeding program difficult and usually requires **pre-breeding** efforts.

In plant breeding, plant breeders cross (interbreed) individuals of different varieties, cultivars or strains with one another and phenotypically evaluate the progeny. Superior offspring is produced through repeated rounds of **phenotypic selection**. It is mainly a process of trial and error. **Genomics** can assist this process by making the outcome predictable and has the potential of changing the trade from an art to a scientific endeavour of prediction and validation. A prerequisite for predictions are established causal relationships between a particular phenotype and a locus in the genome, often a particular allele of a particular gene. A phenotype is said to be *heritable* if it is transmitted to the progeny. A heritable phenotype is called a trait, and its heritability is the basis for genetics. **A gene** is a section of DNA that is transcribed and fulfils a function. While phenotypes are ultimately produced by their respective products, geneticists often treat genes as black boxes, which the majority of them still are. Assigning a phenotype genetically to a gene or a locus can be and often is ignorant to gene function. While individuals by and large share the same set of genes, they often have different versions of them. The different versions of a gene are called *alleles*.

Although a **trait** is a **heritable phenotype**, it can be very different in different genetic backgrounds and environments. Genes or alleles interact with each other and respond to environmental signals to produce a phenotype. The underlying genetic structure of traits can be very different in complexity. Some traits are conferred by a **single gene** (or a particular allele), and

[96] E.g., in Tanksley, S. D., & McCouch, S. R. (1997), McCouch, et al. (2013) and Appendix 3, *The state-of-the-art: methodologies and technologies for the identification, conservation and use of plant genetic resources for food and agriculture*, in: FAO 2010. The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Rome

presence or absence of a gene (or having a different alleles of this gene) expresses itself directly. Examples are many resistances to pathogens or the various traits in Mendel's experiments with peas. These genes are said *to have large effects*. At the other end of the spectrum are so-called **quantitative phenotypes**, where the interactions of many genes of small effect together produce a phenotype. Rather than exhibiting a discrete, binary phenotype, an individual will show a phenotype within a continuous distribution. The more genes involved the closer this distribution resembles a normal distribution in a population. Examples include body height in humans of the same gender. The situation is complicated by the fact that alleles can be **dominant or recessive**. A diploid organism has two sets of chromosomes: It has two copies for most genes, and they can be different alleles. An allele is said to be *dominant*, if one copy of this allele is enough to confer the phenotype. It is said to be *recessive* when not. In diploid organisms, *recessive* has the consequence, that both copies of this gene have to be of the respective allele in order to produce the phenotype. Genes and alleles can also interact with each other in a non-additive fashion. This phenomenon is called **epistasis**. Genes can also have more than one function in different spatial, temporal or developmental contexts or tissues, and alleles can hence influence multiple traits. This is known as **pleiotropy**.

A useful crop combines may useful, desired characteristics and crop performance is shaped by the abiotic and biotic interactions, a large component of which –in agriculture- are management practices. Throughout much of agricultural history, crops were improved through **phenotypic selection**. But since the advent of genomics, choosing PGRFA for breeding and selecting the progeny in a breeding program solely based on phenotype is a suboptimal approach.

Ideally, before breeding even starts, the underlying causal, functional relationships between genes and traits are unravelled and the knowledge of these relationships is then used for **rational improvement**. To achieve sustainable intensification of agriculture, the plant breeding approach will have to be more tailored and problem oriented, rather than to continue selecting the best performing plants in a given year at a given location. Problem orientation means to respond to the specific needs of crop and location, which determines the biotic and abiotic conditions and the available management practices.

The key activities that a **Global Information System** will need to support are:

a) To link phenotypes to genetic variants. This is achieved by genetics experiments, which make use of high-density genotype data and data from phenotyping trials.
b) To support the selection of ideal parental lines as allele donors for the desired traits. The breeding targets are still combinations of phenotypes. But since traits have been associated to genomic loci, they can now be translated genome segments that need to be combined.
c) Monitor inheritance of the genomic segments in the progeny with the use of molecular markers. The molecular markers are developed based on the genomic information

a) Genotype – Phenotype Associations

Genetics is used to find causal genes and establish genotype-phenotype associations. Going from the phenotype to the gene is referred to as *forward genetics*; as opposed to *reverse genetics* where a gene is tested for its function. Frequently used direct methods to establish correlations and causality between a trait and it underlying genetic architecture are: genetic mapping, QTL-mapping and Genome Wide Association Studies (GWAS). High-throughput DNA sequencing vastly increases the speed and resolution of genetic studies.

Limiting factors in forward genetics are the availability of genetic markers along the chromosomes and the size of the segregating populations. With whole genome sequencing, markers can be quickly detected; to saturation, if required and desired. Because the tools can be cost-effectively applied, large numbers of individuals and hence chromosomes can be assayed. Large numbers of individuals are necessary in genetic mapping, because it is the cross-over events during meiosis when producing the segregating population that is used to determine where along the genome the causal locus is, or the causal loci are. The more chromosomes in the experiment, the more cross-over events, the greater the resolution.

It is important to stress that, for breeding, progress in genomic technologies alone will not be sufficient to bring about the reduction to practice. Key will be well-phenotyped PGRFA, which will allow assigning traits to genomic loci and genes. The costs for applying genomics is now low and **quantity and quality of phenotyping data are becoming the rate-determining steps** in capitalising on the genetic knowledge in plant breeding. But phenotyping trials themselves also profit from genomics. Based on genotype information the lines for phenotype trials can be selected based on allelic diversity, and replication can happen on the level of alleles. The integrative power and the replication of alleles could, in

principle, enable to jointly analyse apparently independent phenotyping trials as one experiment.

All of the mentioned forward genetic approaches (i.e., genetic mapping, QTL mapping, and GWAS) are enabled by genomic tools, but frequently suffer from small sample sizes. The hope is that a **Global Information System** will enable joining datasets across experiments to reach sample sizes and the statistical power needed to establish the robust genotype-phenotype associations that are required.

b) selecting parental lines

Given solid phenotype-genotype associations, a breeding target, and a large database of genotypes, the selection of two or more parental lines for the breeding crosses that will contribute the desired alleles should be straightforward. Ideally, parents are selected such that the progeny is made up mostly of desired haplotypes, i.e., the linkage drag is minimal. It should be noted however, that the prediction of the actual phenotypic outcome of crosses is afflicted with uncertainties and currently one of the great frontiers in genetics. Reasons are the effects of dominance, recessiveness, epistasis and pleiotropy of genes mentioned above. However, as data accumulates and phenotyping trials can be viewed as replication of alleles in different genetic background, our knowledge about the interaction of genes will refine and predictions will improve.

The **Global Information System** may also help to determine breeding goals. With a catalog of allelic and phenotypic variation of a crop species, it can be known ahead of time whether or not a desired trait actually segregates in the species and hence if the breeding goal is realistic. It also enables translating knowledge between species and crops and their wild relatives, which may then be used as donors for the desired trait.

c) marker assisted selection

It is currently not possible to predict or influence where along the chromosomes the crossover events will occur in meiosis. This will have to be established after the fact using molecular markers. Molecular markers in the traditional sense are molecular biology based assays able to interrogate known variants. Applying molecular markers -also called: genotyping- is fairly cheap hence large numbers of individuals can be assayed. The process of using molecular markers to determine inheritance of one or several

genomic loci in plant breeding is called “Marker Assisted Selection” (MAS). With MAS the genetic makeup of the progeny can be tested, which abolishes the need for phenotyping. The **Global Information System** could provide means to translate the genomic information into markers for MAS.

The impact of genomics on pre-breeding

Opening genetic bottlenecks.
There is ample evidence, scientific and anecdotal, that landraces harbour alleles and pathways highly valuable for crop improvement breeding programs, yet, this variation is currently un- and under-exploited[97]. One of the great promises of genomic tools applied to genebank material and plant breeding is that it opens breeding programs to landraces and so enables to overcome the **genetic bottleneck** of our crop plants. Please see the textbox on Genetic Bottlenecks.

In plant breeding, the desired enrichment of alleles that confer superior agronomic performance comes at the cost of reduced genetic diversity. The loss may concern other desirable characteristics, such as nutritional value, but most importantly, with the loss of genetic variation the ability to quickly adapt to environmental challenges is impaired. Examples for severe bottleneck events during the development of our major food crops are the a) original domestication event and b) The Green Revolution.

Genetic Bottleneck

A bottleneck event is a significant reduction in population size. In a bottleneck event variation is purged, because the survivors only contain a subset of the variation that was present in the original population. When the population recovers and population size increases, it can only draw from the genetic variation that is left. In addition, during recovery, allele frequencies are likely changed. In the illustration alleles are represented by balls. Note that, after the bottleneck event, the red balls are lost completely and the ratio between yellow and the blue balls in the new population is vastly different from what it was in the source population.

[97] Tanksley, S. D., & McCouch, S. R. (1997), McCouch, et al. (2013)

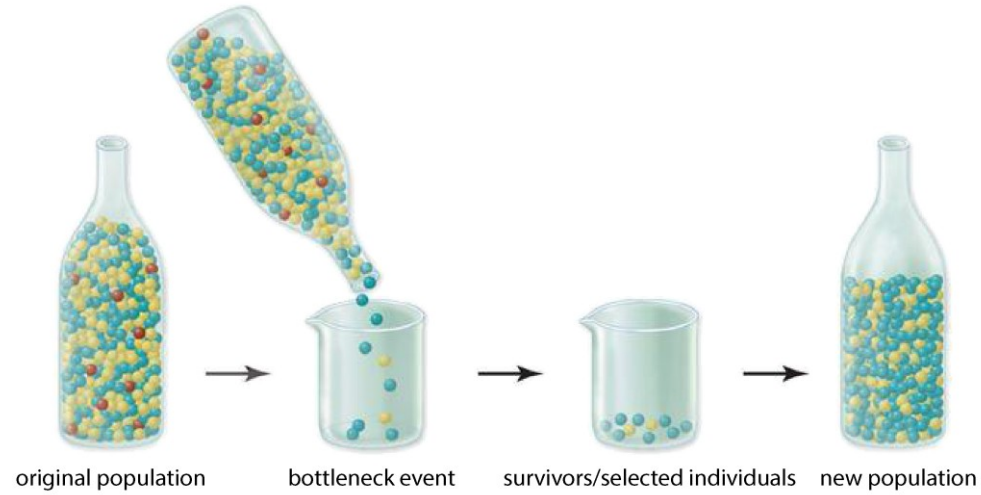

Figure 6: Genetic drift associated with a bottleneck event.[98]

Bottleneck events during crop development

a) The original domestication event
With domestication begins the narrowing of a crop plant's genome. The first farmers drew a sample from the wild gene pool and this sample contained only a subset of the genetic variation found in the progenitor. Centuries of selecting and propagating formed landraces. Every landrace has a subset of genetic variation of the source population.They often contain significant amounts of heterozygosity. Hence, some traits still segregate and allele frequencies can fluctuate from year to year due to different fitness of different genotypes. Landraces typically have been cultivated for centuries in a particular geographic region. Part of the harvest was kept for the next cultivation cycle. It is hence often argued that the allelic diversity in landraces is suited for the conditions of said geographic region.

b) The Green Revolution and the creation of elite and mega varieties.
Modern scientific plant breeding created another severe bottleneck. The desire for uniformity of the agricultural product with focus on traits of commercial value, such as high yield and suitability to mechanisation, almost dictate low genetic diversity. The current, highly successful elite or even mega varieties represent a successful combination of many desired genes, allele and traits, but almost zero genetic diversity. New traits will have to be deliberately bred into them.

[98] Image source: studyblue.com
https://classconnection.s3.amazonaws.com/749/flashcards/1672749/jpg/bottleneck_effect1342125075025.jpg , original figure text removed

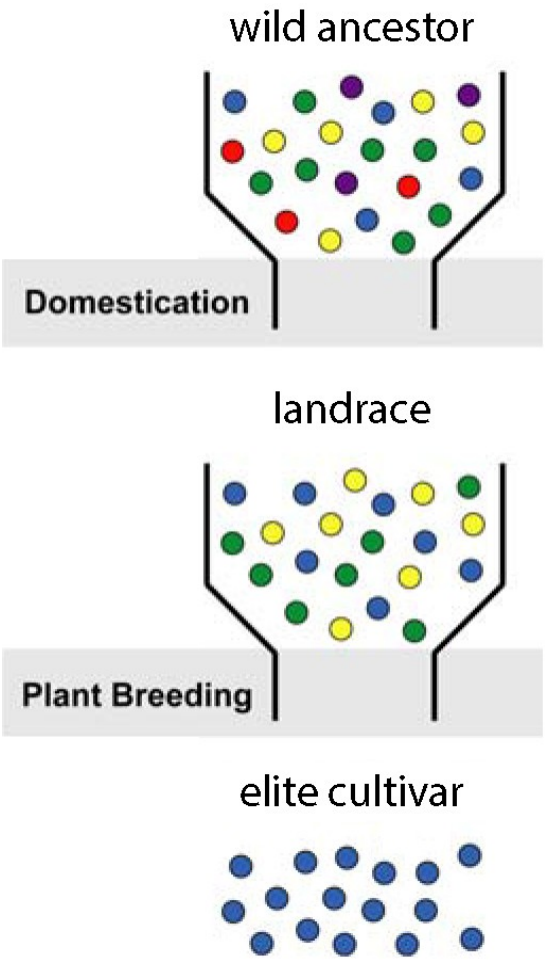

Figure 5: The consequence of plant breeding on the genetic variation in crops[99]

Breeders are reluctant to include highly diverse (aka exotic) material into their breeding programs, because of the so-called **linkage drag**. Linkage drag is unwanted genetic material (parts of the genome) that is co-inherited from the wild progenitor along with the trait(s) of interest. This unwanted genetic material has to be subsequently removed from the breeding population by further crossing, which can take several iterations, and, without marker assisted selection, each step requiring additional phenotypic evaluation of the progeny.

Cost-effective genomics tools and a **Global information System** will assist in opening the genetic bottleneck in several ways. Given a database of PGRFA genotyped to high resolution, allele donors can be selected in a targeted fashion. A breeder can select ideal allele donors that bring the least amount of unwanted genetic material. In a more systematic way, this database can be

[99] Cartoon adapted from Yamasaki, et al. (2005)

used to select and create diversity sets and allele-mining sets. A process often called pre-breeding[100]. Often times not all desired crosses can be realised due to incompatibility barriers between strains, even within a species. If encountered, then the database will allow for selecting alternatives. With the possibility to quickly and cost-effectively genotype progeny with dense sets of markers, chances are that linkage drag will be less of a concern in the future.

Finish the domestication of landraces

Many minor crops exist only as landraces, many of which very robust to environmental stresses, but low yielding. Plant research has unravelled many of the large effect genes and alleles exploited during domestication processes. The genomic information of landraces in those minor crops will reveal whether or not similar domestication genes and alleles exist and pre-breeding can attempt to introduce or genetically fix them. This will make a contribution to maintain and increase agrobiodiversity.

(Re)creating allelic combinations

If a PGR or a breeding line gets lost, but its genotype is known and a database of genetic data from other PGR is available, then it is rather straight-forwards to identify the closest relative, which can then be used as replacement. More generally, the boundary between genomic information and actual PGR might blur in the future: by knowing the genomic composition of a PGR's genome it can potentially be recreated, by crossing the likely ancestors again and selecting the progeny with the desired composition or, in the not too distant future, possibly by DNA synthesis. Chromosome-size DNA molecules have already been and are being assembled in yeast[101].

[100] http://www.croptrust.org/content/pre-breeding
[101] Karas, B. J. (2012), and http://syntheticyeast.org

Recommendations

The future landscape
Genomics has been and is a game changer. The widespread availability and easy access to modern DNA sequencing technologies puts high-density genotyping in the reach of everyone, everywhere. The near future will bring whole genome reference sequences for dozens and hundreds of crop species and varieties. A great proportion of accession held in major genebank collections (on the order of hundreds of thousands) will soon be genetically characterised in detail, many by DNA sequencing. Large projects on rice, maize, wheat, are underway, and they will soon be followed by more specialised collections, and additional crops.

The resolution of the genomic characterisation will vary and will mainly be dictated by technical obstacles posed by the genome in question and available funding. Nonetheless a plethora of genomic data will be created in the near future and to maximise utility, this data should be made publicly available for the advancement of Research and Development and to add value to the PGRFA. The current technology of DNA sequencing and the analysis pipelines are mature. There will be incremental improvements to the technology, but the price for sequencing is not likely to drop another order of magnitude any time soon. While the type of genomic data is also not likely to change soon, methods of storage, representation and analysis likely will.

General Recommendations
A **Global Information System** for PGRFA as foreseen in the International Treaty on Plant Genetic Resources for Food and Agriculture should contain genotype data, preferably sequence data, because such data greatly enhances the value of PGRFA. High resolution genomic information increases efficiency and reduces cost of genebank management and enables breeding methodologies that shorten breeding cycles, and enables much better and more targeted breeding by targeted access to genetic variation for plant breeding. In addition, genomic information can be mined to develop molecular markers that can then be used to follow the inheritance of linked traits during the breeding cycle, i.e., by Marker Assisted Selection (MAS), reducing phenotyping requirements.

The Governing Body of the International Treaty in interaction with the various stakeholders should conceptualise a **Global Information System** that can receive, store and make publicly available genotype and phenotype and any other relevant data for work with PGRFA. This will

require the development of data standards, incentives to contribute, and a data sharing policy.

Genomic data is rather uniform in nature, which makes the genomic data a great integrator. Challenges arise from the shear data volume and the varying quantity, quality and the choice of parameters during analysis. However, in case of DNA sequencing, when the raw sequence data is available, it can be easily (re-)analysed at any time; also in different context and using more advanced analysis tools and software programs. The Global Information system needs to be capable of receiving, collating and storing data from diverse, decentralised sources. In case of DNA sequencing, the most valuable is the raw date, i.e., the original sequencing reads. For derived data meta-data requirement maybe put in place that at least comprise the original data and all relevant information about software programs, version numbers, and run-time parameters used. Since the power of analysis and confidence in the results increases with the number of individuals analysed and because novel analysis tools will become frequently available, the Global Information System should provide the means to enable data to be (re-)analysed in the future.

Data sharing

The Global information System should make data accessible in meaningful and comprehensible ways. In human readable formats as well as in formats suitable for analysis by a variety, including third party, software tools. Ideally, the Global Information System has an efficient and open Application Programming Interface (API). In the field of genomics, the data volumes are large and researchers interact with the data exclusively through computer programs. An exception are genome viewers with which genomic data can be explored and inspected by the human eye, however, in this case the viewer will have to have automated access to the data.

Despite the requirements of the Treaty's Standard Material Transfer Agreement (SMTA) to share all non-confidential data associated with the PGRFA transferred within the Multilateral System[102], there is the risk is that stakeholders may be reluctant to comply if no additional incentives or mechanisms are put in place. As the sharing of data will be a cornerstone for the practicability and the success of the Global Information System, The Treaty must encourage participation and data sharing. This can be achieved by showcasing the advantages of data sharing through success

[102] ITPGRFA, FAO. Standard Material Transfer Agreement, Article 6, 6.9

stories, for the current rate of progress in plant breeding is currently too slow and will remain too slow without the sharing of data.

To enable data sharing, the issue of Intellectual Property Rights of genomic data, in particular whole genome DNA sequence data of PGR must be addressed. PGRFA holders producing genomics data and, in particular, DNA sequence data from a PGRFA within the Multilateral System should be put in a position to make this information publicly available without risking legal consequences. Uncertainty in this are will negatively impact participation.

Support of high-quality reference genomes including genome annotation
A finished, high-quality reference genome is of enormous value for any crop research, genetics and breeding community. Projects aiming at producing high-quality reference genomes should be supported by the Programme of Work of the Global Information System. Where appropriate, The Treaty should consider initiating the sequencing, assembly and annotation of **reference genomes**.

Support the development of data structures for re-sequencing data
The current practice of storing genetic variation in large tables will soon become a bottleneck as it does not sufficiently scale to large sample numbers. Hence, besides high-quality reference genomes, **supporting the development of data structures** that enable efficient representation of crop pan-genomes should be an area of **strategic investment.**

Re-sequencing of genomes already is and will become more so, highly de-centralised. For any given crop, researchers all over the world contribute data at different quantities an qualities. An efficient data structure needs to be developed that can receive this data, allows for cost-efficient storage, and easy, fast and flexible retrieval. It further needs to provide straightforward ways to update the population wide variation information on a regular basis. Relevant developments in this area are coming from the Global Alliance for Genomics and Health[103], where data storage based one genome reference graphs are suggested and developed[104]. Whatever the Global Alliance will implement, the crop science community should develop a system very similar and possibly in collaboration, but extending and adopting it to accommodate the peculiarities of plant genomes, such as genome size, ploidy levels, genome structure and repeats, organellar genomes, etc..

[103] http://genomicsandhealth.org
[104] Gil McVean February 18th, 2014

Support phenotyping projects and standards
Phenotyping projects need to be comparable such that the data can be aggregated to reach the numbers required to establish statistically significant genotype-phenotype associations. They, and data from environmental monitoring, will become the starting point for applied, targeted breeding. The Treaty, in framing the Global Information System, could initiate and promote the development of standards and Crop Ontologies, possibly in partnership with FAO and existing international projects and initiatives. Expertise in the field can be found within the CGIAR system especially at Bioversity International[105].

Partners
In developing and implementing the **Global Information System**, The Treaty should liaise with key partners who can contribute resources and knowledge. Relevant partners to liaise with on the technical aspects relating to Genomics of the Global Information System are the **DivSeek initiative**[106], the **African Orphan Crop Consortium**[107], the **Global Alliance for Genomics and Health**[108], the cyberinfrastructures **iplant**[109] and **transPLANT**[110]**,** and possibly **Google**[111]**.** The organisational structure of the Global Alliance for Genomics and Health may serve as example.

Capacity building
As established, DNA sequencing is becoming mainstream. It is very accessible through commercial DNA sequencing providers to anyone who has samples and funding. While the generation of sequences is readily outsourced, capacity needs to be build for the work upstream and downstream, which concerns the preparation and handling of DNA and the interpretation of the data. Advances in genomics will enhance the scope and efficiency of plant breeding, but it will still be *breeding* in the traditional sense. Attention needs to be paid to the global scarcity of knowledgeable plant breeders, people able to work with whole plants in the field. The skills required will be mainly in phenotypic characterisation, environmental monitoring and principles of genetics to enable informed use of the genotype information.

[105] Descriptors http://www.bioversityinternational.org/research-portfolio/information-systems-for-plant-diversity/descriptors/
[106] http://www.divseek.org
[107] http://www.mars.com/global/african-orphan-crops.aspx
[108] http://genomicsandhealth.org
[109] http://www.iplantcollaborative.org
[110] http://www.transplantdb.eu
[111] https://cloud.google.com/genomics/

Bibliography

2013 DOE JGI Progress Report http://jgi.doe.gov/wp-content/uploads/2013/11/JGI_Progress_Report_2013.pdf

Sarah C Ayling (2013) Technical appraisal of strategic approaches to large-scale germplasm evaluation – prepared for the DivSeek Initiative, online available at http://agro.biodiver.se/wp-content/uploads/2012/12/Technical-appraisal-NGS-for-genebanks-please-comment.pdf, last accessed Nov 2014.

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

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

Bennett MD, Leitch IJ. 2012. Angiosperm DNA C-values database (http://data.kew.org/cvalues)

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

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

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. PLoS Computational Biology, 10(12), e1003998. doi:10.1371/journal.pcbi.1003998

DivSeek Initiative, White paper (2014): Harnessing the power of crop diversity to feed the future. It can be retrieved from http://www.divseek.org/white-paper/

FAO. International Treaty on Plant Genetic Resources for Food and Agriculture. FAO. (Retrieved from ftp://ftp.fao.org/docrep/fao/011/i0510e/i0510e.pdf in July 2014)

FAO 2010. The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Rome

Finkers, R., Chibon, P.-Y., Van Treuren, R., Visser, R., & Van Hintum, T. (2014). Genebanks and genomics: how to interconnect data from both communities? Plant Genetic Resources, 1–4. doi:10.1017/S1479262114000689

Ganal, M. W., Polley, A., Graner, E.-M., Plieske, J., Wieseke, R., Luerssen, H., & Durstewitz, G. (2012). Large SNP arrays for genotyping in crop plants. Journal of Biosciences, 37(5), 821–828.

Gil McVean February 18th, 2014: A Population Reference Graph for Human Genetic Variation http://simons.berkeley.edu/talks/gil-mcvean-2014-02-18

Global Alliance for Genomics and Health, White paper (2013): Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data (2013), edited by Peter Goodhand, Marian Orfeo, and David Altschuler

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

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

ITPGRFA, FAO. Standard Material Transfer Agreement, ftp://ftp.fao.org/ag/agp/planttreaty/agreements/smta/SMTAe.pdf (accessed July, 2014)

Karas, B. J., Molparia, B., Jablanovic, J., Hermann, W. J., Lin, Y.-C., Dupont, C. L., et al. (2012). Assembly of eukaryotic algal chromosomes in yeast. Journal of Biological Engineering, 7(1), 30–30. doi:10.1186/1754-1611-7-30)

Kawahara, Y., la Bastide, de, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice, 6(1), 4. doi:10.1186/1471-2164-8-278

Lander ES, Waterman MS: Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 1988, 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

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

McCouch, S. R., McNally, K. L., Wang, W., & Sackville Hamilton, R. (2012). Genomics of gene banks: A case study in rice. American Journal of Botany, 99(2), 407–423. doi:10.3732/ajb.1100385

McCouch, S., Baute, G. J., Bradeen, J., Bramel, P., Bretting, P. K., Buckler, E., et al. (2013). Agriculture: Feeding the future. Nature, 499(7456), 23–24. doi:10.1038/499023a

McPherson, J. D. (2014), A defining decade in DNA sequencing. Nature Methods, 11(10), 1003–1005.

McVean, G. A., Altshuler Co-Chair, D. M., Durbin Co-Chair, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56–65. doi:10.1038/nature11632

Michael, T. P., & Jackson, S. (2013). The first 50 plant genomes. The Plant Genome, 6(2), 1–7.

Morrell, P. L., Buckler, E. S., & Ross-Ibarra, J. (2011). Crop genomics: advances and applications. Nature Reviews Genetics, 13(2), 85–96. doi:10.1038/nrg3097

Reece, J. D., & Haribabu, E. (2007). Genes to feed the world: The weakest link? Food Policy, 32(4), 459–479.

Regalado, A. (2014, September 24). EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced This Year. MIT Technology Review. Retrieved October 16, 2014, from http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/

Rensink, W. A., & Buell, C. R. (2005). Microarray expression profiling resources for plant genomics. Trends in Plant Science, 10(12), 603–609. doi:10.1016/j.tplants.2005.10.003

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

Robin Fears (2007) Genomics and Genetic Resources for Food and Agriculture - prepared for the Commission on Genetic Resources for Food and Agriculture, FAO.

Rosamond Naylor and Richard Manning (2005): Unleashing the Genius of the Genome to Feed the Developing World, Proceedings of the American Philosophical Society, Vol. 149, 515-528

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, 5(Suppl 7), P54.

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

Schatz, M. C., Witkowski, J., & McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biology, 13(4), 243. doi:10.1186/gb4015

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., & Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. Genome Biology, 10(9), R98. doi:10.1186/gb-2009-10-9-r98

Shapiro, J. A., & Sternberg, von, R. (2005). Why repetitive DNA is essential to genome function. Biological Reviews of the Cambridge Philosophical Society, 80(2), 227–250.

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

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

Tanksley, S. D., & McCouch, S. R. (1997). Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. Science, 277(5329), 1063–1066.

van Berkum 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 Oct 2014

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., et al. (2014). Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. Genome Research. doi:10.1101/gr.170332.113

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

Yamasaki, M., Tenaillon, M. I., Bi, I. V., Schroeder, S. G., Sanchez-Villeda, H., Doebley, J. F., et al. (2005). A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. The Plant Cell, 17(11), 2859–2872. doi:10.1105/tpc.105.037242

Zamir, D. (2013). Where have all the crop phenotypes gone? PLoS Biology, 11(6), e1001595. doi:10.1371/journal.pbio.1001595

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