The 150+ Tomato Genome (re-)Sequence Project; Lessons Learned and Potential Applications

Richard Finkers

Researcher Plant Breeding, Wageningen UR Plant Breeding, P.O. Box 16, 6700 AA, Wageningen, The Netherlands, Wageningen Campus, Building 107, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands, Tel: +31-317-484165, Fax: +31-317-418094, Skype: richardfinkers

In current breeding practice, the value of parents in potential crosses is determined almost entirely from the phenotypic performance of their progenies in earlier crosses. However, with the decreasing costs of next-generation sequencing, the extra possibilities of high-throughput phenotyping of traits across major crops and animal species, and the surge in availability of gene, protein and metabolite annotations in literature and databases, there is a demand to integrate all this information across data levels, combined with pedigree and progeny information. The precision breeding programme consists of six themes. Themes one to four focuses on statistical and bio-informatics methodology development, theme five focuses on translation of technology and theme six bundles the outreach activities. To ensure that the information obtained from the individual (PhD) projects will not be stand-alone we envisage appointing three postdocs. The postdocs have to see to it that both within a particular theme but also between different themes (1, 2 and 3 and within 4) synergy can be achieved between the different findings. The post-docs will work in close collaboration with the programme committee (which will be set up at the start of the programme). We will focus on different economically important model systems (diploid inbreeding: tomato, diploid outbreeding: cattle, pig, chicken and tetraploid outbreeding & clonally propagated: potato). Precision breeding will allow breeders to provide answers to tomorrow’s food security questions more efficiently. Statistical & bio-informatics methodology development

Theme 1: High-throughput genotyping (approximately five projects)

Developments of high-throughput (HTP) sequencing technologies, and the associated huge drop in sequencing costs, now makes whole-genome re-sequencing of large sets of individuals feasible. We expect that whole-genome sequencing-based genotyping will rapidly replace the current SNP based detection platforms in many economically important crop and animal species, such as tomato and dairy cattle. Sequencing-based genotyping also shows a great promise for species with a more complex genome, such as tetraploid potato and for the sequencing of epigenomes. Knowledge about epigenomes, which include methylated DNA sequences and small RNA populations is increasingly important for breeding as a source for genetic variation but remains largely unused due to the instable and unpredictable nature of the epigenome. Detailed investigation of genome structure will also provide insight in processes affecting breeding, such as: 1) the regulation and preferential locations of recombination, 2) stability of the genome, i.e. frequency of DNA copy number variants (CNV), and 3) will help to identify rare haplotypes in a population. These rare haplotypes can be important as they could carry alleles for important traits. However, these rare haplotypes are currently not detected, as they are not recognized as an individual haplotype. The objectives are to develop technology and strategies to improve and exploit sequencing-based (epi) genotyping technologies, aiming at a better understanding of how these technologies can be used within breeding strategies. We will focus on Genotyping-by-sequencing strategies in polyploid crops and diploid animals for reconstruction of haplotype blocks from sequencing data, analysis of the impact of structural genome variation on breeding, optimization of RNA seq approaches, the impact of epigenomic variation on breeding and haplotype reconstruction from genome-wide sequencing approaches. The deliverables from these projects are: 1) knowledge on sequence and epigenome variation, and its impact on precision breeding strategies, 2) optimal design of HTP genotyping experiments, 3) adaptation of sequencing methodology towards breeding applications, 4) statistical approaches for haplotype reconstruction and 5) efficient bio-informatics for storage & analysis of sequence-based genotyping technologies and for making these data efficiently available for further analyses (program theme four).

Theme 2: High-throughput phenotyping (approximately five projects)
An accurate and detailed description of the phenotype over multiple traits of interest is essential to discover the functional implications of haplotypes Genetic variation in haplotypes is the basis for biodiversity, directing expression and function of genes, which determines the morphology, architecture, resistance to diseases and tolerance to adverse conditions, and ultimately yield and quality. We need to monitor differences in phenotypes and link these differences to the variation in the genes. For this, we need to: 1) Monitor a large number of individuals in a relatively short time, 2) Monitor processes and traits that contribute to yield and quality in detail and 3) Account for environmental conditions that may influence phenotypic expression of these genes. New techniques for high-throughput digital imaging have been and being developed, including chlorophyll fluorescence imaging, thermal imaging, spectral imaging and 3D imaging, but also in technologies that measure concentrations of compounds inside the individual such as metabolomics, proteomics or MIR. The high frequency of data acquisition makes it possible to study development and growth, and to design and improve growth models. A major challenge in high-throughput phenotyping is processing and analysis of the data. Huge amounts of raw data are produced, that need to be converted into information that can be used for breeding. The objective is to develop methodologies to convert images and other highly multivariate data sets into more easily interpretable and manageable features that describe traits, and can be used for genetic analysis and selection, and implement these methodologies in high-throughput phenotyping systems for precision breeding. We will focus on: 1) Analysis of phenotypes, which are monitored over time, 2) Novel trait discovery and feature extraction from large datasets of digital images, 3) Methodology to account for environmental influences (e.g. fluctuations in temperature and irradiation) and correlations among traits. The deliverables from these projects are design of HTP phenotyping experi-

ments, efficient bio-informatics for storage of phenotyping data and statistical approaches for analysis of high-throughput phenotyping data and tools to utilize these approaches efficiently (program theme four).

Theme 3: Integration of knowledge on gene function within Precision Breeding approaches (approximately three projects)

Current marker-assisted (MAS) breeding approaches like, linkage disequilibrium (LD) analysis or genomic selection (GS) are used to improve plants and animals not regarding the processes responsible for expressing the desired phenotype(s). Nowadays, functionally annotated genome sequences are available for major species, including cattle, chicken, tomato and potato. The objective of this subprogram is to investigate strategies to incorporate functional knowledge about biological systems into breeding approaches. We will focus on strategies to identify genes, which are candidates affecting our phenotype of interest and to incorporate these functional annotations into breeding approaches. Such a strategy is important as a genomic region, associated with a trait of interest can easily contain over 1000 genes. Semantic database integration approaches and ontologies will be used to aggregate "meaningful" annotations for each functionally annotated domain or gene (e.g. automatically obtain the metabolic pathway in which a gene functions, and from that determine whether it is likely to have impact on the trait of interest). This will be combined with bioinformatics methods that provide sequence- and network-based annotation of gene functions. Knowledge about gene regulation by transcription factors should be captured in such a scheme as well. In addition, approaches that prioritize sequence variation based on e.g. known functional sites and predicted impact on functional properties will be applied. In order to successfully integrate this knowledge with data for a phenotype, terminology for phenotyping should also be standardized. The deliverable is a tool, which will provide "meaningful" annotations and assigns an "interest" or "priority" score for all genes in a given region, thus quantifying the possible involvement of genes in the traits. These priority scores can then be used within programme theme four for further analysis.

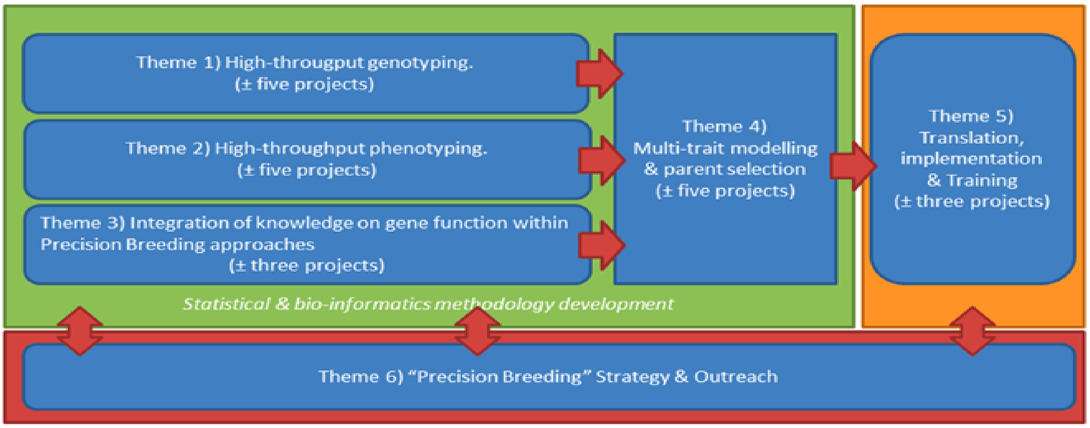
Theme 4: Multi-trait modelling & parent selection (approximately three projects). Genomic selection (GS) has grown into an important tool within animal breeding while in plant breeding marker-assisted selection (MAS) based on marker-trait relation resulting from methods like QTL mapping or association analysis (LD) mapping is more often used. Genomic selection aims to identify those candidate individuals that have the highest breeding value for a trait based on the fullgenome genotype. In contrast, MAS aims to obtain a desired combination of traits by selecting for specific alleles that have been identified as associated to the traits. Although these approaches are fundamentally different, each strategy can be used successfully in both plant and animal systems. A shortcoming of GS, QTL and LD mapping is that they aim at analysis of one trait at a time, whereas practical breeders need to improve multiple traits simultaneously, while environmental factors also have an influence on expression of a phenotype. Pedigree data, including phenotypic observations for these genotypes, can be analysed to identify “negative” or “positive” epistatic interactions (the combination of alleles, which do not lead to the in phenotypic value as predicted from additivity). For new and expensive traits to measure (e.g. disease data in an animal population), the challenge is optimal use of information from relatively few but very important phenotypes. This requires design of optimal recording strategies, and methodology that combines information from “genomically” correlated traits that are widely recorded. The objectives are to develop and evaluate methodology that will be able to predict multi-trait phenotypic values, for a specified environment, using: 1) haplotype block data (Theme 1), 2) parameters obtained using high-throughput phenotyping experiments (Theme 2), 3) information from ~omics technologies (metabolites, proteins, gene expression, Theme 2). Meaning to these predictions will be provided via database integration approaches (Theme 3). We will focus on integration of phenotyping datasets, multi-trait association modelling, prediction of optimal genotypes for a combination of traits for a specific environment, and methodologies to select the optimal parental genotypes to breed for such a genotype. The deliverables are: 1) efficient statistical methodology for multi-trait modelling and integration of high-throughput phenotyping experiments, 2) methodology to predict optimal parents to obtain individuals or varieties with a specific set of traits (where those traits would be defined by the market goals, including the environment where the individual or variety would be grown) using the strategies developed within the program themes 1 to 4, and 3) the breeding strategy achieve this goal and deliver a proof-of-concept.

Translation of technology; Theme 5: Translation, implementation & Training (approximately three projects)

The methodology development programs deliver fundamental knowledge and analyses methodologies to analyse high-throughput genotyping and phenotyping experiments. The objective is to translate these technologies into solutions that can be implemented and used by the community (scientists and breeders alike). Methodology developed for high-throughput analysis (subprogram 1 & 2) will be translated into solutions that will be used in real-life breeding practice. Bio-informaticians or scientific programmers from institutions and or companies in our user community will, with the support of the appointed post-doc develop dedicated environments. This training & development on-the -job mode will ensure that: 1) these environments are embedded in the infrastructure of the involved institution or company and 2) can be maintained and further developed after the program has ended. The focus will be on: 1) Development of Virtual Machines (VM) for secure deployment on computational clouds. The developed VM can contain one or more "Problem Solving Environments", each capable of executing a different type of analysis and 2) Development of a system of in house genomic databases and off -line storage of raw data. Methodology developed for integrating knowledge of gene function, multi-trait modelling and parent selection will be implemented within breeding database systems, such as BreeDB (http://www.plantbreeding.wur.nl/UK/software_breedb.html). BreeDB is a relational database system that aims to support breeding for quantitative and qualitative traits. This database can be explored through a web-based interface, which provides a number of data exploration tools. The focus of this project will be on implementing methodology for optimal parent selection within breeding database systems such as BreeDB and training of end-users. The deliverables of these projects are translation of the developed methodology into user-friendly solutions **& training of end-users.**

Outreach; Theme 6: Outreach

This project's objective is to ensure that both internally and externally sufficient exposure will be delivered of the obtained results. Each project will contribute to the overall goals of precision breeding by delivering knowledge, protocols, tools and/or strategies on how to efficiently use HTP genotyping, HTP phenotyping, functional annotations, and advanced statistical modelling to achieve predefined precision breeding goal(s). We plan to organize a yearly symposium where program members, and international keynote speakers, will present their findings and to discuss the impact on breeding strategies with the user community. Keynote researchers and PhD students appointed to the program will be asked to write a review on how their work contributes to the precision breeding strategy. These contributions will be bundled and published as an (e)book. To further educate PhD students in this field, the obtained information will be reformulate into modules which will be offered as training courses within existing networks of graduate schools like PERC, EPS, and WIAS. Likewise, within Wageningen University we will strive to link and incorporate essential knowledge into the concurrent MSc programmes where possible to ensure the education of N.


Theme 1) High-througput genotyping.
(± five projects)
Theme 2) High-throughput phenotyping.
(± five projects)
Theme 3) Integration of knowledge on gene function within Precision Breeding approaches
(± three projects)
Theme 4)
Multi-trait modelling
& parent selection
(± five projects)
Theme 5)
Translation,
implementation
& Training
(± three projects)
Statistical & bio-informatics methodology development
Theme 6) "Precision Breeding" Strategy & Outreach