THIS PROJECT IS FUNDED BY THE
EUROPEAN UNION HORIZON 2020 RESEARCH
AND INNOVATION PROGRAMME UNDER
GRANT AGREEMENT 817690

CropBooster-P

Deliverable No. 5.1

Title: The composition of the report describing the genetic deliverables of a future crop yield improvement programme

Start date of the project: **November 1st, 2018** / Duration: 36 **months**
Planned delivery date: M8 (end June 2019)
Actual submission date: 31 August 2019
Work package: WP5 / Task: 5.1

Work package leader: WU

Deliverable leader: Jeremy Harbinson

Version: Draft 1

Date of version: 31 August 2019

Dissemination level	Public

Contents

Defining the reports summarising the products of research conducted by a future crop-yield improvement programme to the end-users.

Summary Introduction

Cropbooster-P aims to devise a research road-map that will ultimately be embodied in research programme that will effectively future-proof Europe's crops. A goal of this programme will be to produce or identify new genetic potential for improving crop-yields in Europe in the future, so this requires improvements in yield that are sustainable and future-proofed (ie will be tolerant of future climates and achievable with the resources that will be available in the future). It is essential that any future crop yield improvement programme will make its discoveries clearly known to the end-users of this innovation, such as plant-breeders, crop scientists or relevant government bodies. This will require some kind of report that fully describes the discovery or innovation; metabolic or crop growth or yield models that incorporate the discovery or innovation and show its impact on biochemistry, physiology and yield; and genotypes of model and crop species that contain the discovery or innovation so that the end-user can experiment with these improved genotypes. Note that if the discovery is for valuable natural variation in a gene then the model genotypes will possibly employ gene-editing to switch the allele to (or possibly from) a more effective allele in a reference genotype and so illustrate the effectiveness of the allele. The key approach is therefore to document, to model and to demonstrate improvement. We would also stress, however, that while a well organised and comprehensive report of a discovery, and good mathematical and biological models are a powerful tool for disseminating this knowledge, the culture of the crop-yield network will itself play an important role in making available to end-users the innovations it develops. This question of how best to organise a public-private partnership of the kind that will be embodied by the future crop-yield improvement programme needs to be addressed within the management plan for the project. In this note we will address the products that will be delivered by the programme – reports, mathematical models and biological models.

The likely form of the genetic innovations produced by the crop-yield network

At this point in time European (ie the EU; in this document 'Europe' will refer to the EU and, where appropriate, Switzerland and Norway) legislation makes it difficult to grow genetically modified crops in the field so the emphasis of crop breeding in Europe is in making use of natural variation. Natural variation may include deliberately mutagenized genotypes as long as that mutagenesis does not involve transgenic, gene editing, or similar technologies at any point. The identification and use of natural variation for complex, quantitative yield related traits (such as photosynthesis) is in its infancy and currently depends on a combination of genotyped mapping populations of various kinds (eg diversity panels, biparental inbred lines, MAGIC populations), phenotyping, and then numerical analysis that correlates phenotypic diversity with genotypic diversity to identify quantitative trait loci (QTL) or associations, followed by further analysis to identify the causal gene(s) underlying the QTL or association. These QTLs or causal genes are a genetic discovery as they can be targets for conventional breeding approaches. As part of the gene identification process genetically modified or gene edited genotypes can be produced, and these may also be of value. They are also a genetic innovation. In addition a future Europe may grow to accept novel plant breeding technologies (NPBT) like gene editing or some as yet undiscovered tool. Genetic innovations may therefore also be the product of a NPBT-based genetic modification driven by a combination of genetic and physiological/biochemical knowledge. Overall, therefore, it is expected that a future crop-yield programme will deliver genetic innovation of various classes and the report structure needs to accommodate the different paths leading to innovation implied by this.

1. The report

The report will be a key document as it will describe the discovery or innovation, the background to the discovery or innovation, the research process, the evidence confirming the discovery and innovation, and the impact of the discovery or innovation on crop-yield. It should therefore include the reasoning and prior knowledge that led to the development; a description of the genetic innovation (including for example the species involved, gene identifier, gene-product identifier, metabolic pathway); the physiological and agronomic evidence for the effectiveness of the innovation; references to physiological, biochemical and crop growth models that include the phenotypic effects of the discovery; and the responsible researchers and their role in the discovery. Insofar as is possible all descriptions of genes, species, pathways etc should use both conventional names and the names or identifiers used in standard databases (eg UniProt) and these database identifiers should be linked to the appropriate website. It is envisaged that the report will be both a discrete document capable of being used stand-alone but will also be a gateway to on-line digital information that can be used to provide more detailed information to any user of the document, extending to allowing the complete re-analysis of the primary data upon the evidence is based. The report in its digital form should be easily searchable to make easier both its discovery and its use on-line and as part of the database of the any future crop-yield network. This report will attempt to describe a structure for the discovery-report.

Transparency and traceability of evidence.

A value of report describing an innovation will depend on the evidence upon which it is based. Normally the experimental evidence supporting a conclusion is presented in a refined way in the form of graphs and tables which normally contain derived data that supports the reasoning behind the claims. Rarely is the primary data, and often not most of the derived data, made publicly available. This leads to questions surrounding the transparency of data and the traceability of the evidence. The future crop-yield programme will have a comprehensive data management allowing all primary and derived data to be stored. All of this stored data will be uniquely identified by means of a DOI (digital object identifier), allowing it to be electronically retrieved. Within the report model we propose that all evidence will be connected via a DOI to either the source data (both primary and derived) and the methods used to analyse that data, or if the source data or methods are too extensive to allow them to be practically linked to from the report, to a supplementary document that will catalogue in a meaningful way the source data and methods. So, for example, if a graph is included in the report all data within the graph will traceable via DOIs to archived tables (etc) from which the graph was plotted, and the data in these tables will be similarly traceable to more primitive data all the way to primary data. In this way it will be possible for anyone to verify any piece of evidence presented in the report. This will, by implication, mean that the report will be integrated with, and dependent on, the data management structure of the programme and it will serve, along with other publications, as a top-level object in the data pyramid.

Searchability.

In order to maximise the usefulness of the report it will need to be in a form that maximises the searchability of the document while still retaining its readability. Searchability will allow the report to be identified or reclassified based on its content. Currently the PDF format seems to give high quality formatting, embedding of URLs, portability and archivability. While the PDF format is not so simple to search as, for example, LaTeX source files, it can be searched, so currently we believe that PDF format should be the format for the downloadable version of the report. It is also a format that is viewable on-line so to maintain parity between the downloadable and on-line viewable forms of the document we advise using PDF for all final versions of the report.

Suggested structure of the report.

The report is expected to be a live document whose content will progress as research is conducted upon the innovation. The reason for this is to accelerate communication with end-users and make easier feed-back from stakeholders. To be sure, the first version of the report should describe an innovation which is clearly at the point where it will be useful to end-users; there should be solid evidence that a specific genetic change, however achieved, can result in an increase in yield. It is accepted, however, that continued research upon this innovation will increase understanding of the utility and applicability of the change. This later research will be enhanced by discussion with end-users so early publication of the discovery is important. The results and consequences of this subsequent research should be included in later, updated versions of the report, with earlier versions continuing to be available. This will make the progress of research supporting the development of the innovation transparent.

The reports should include the following sections containing at least the specific information listed:

1. **A title page (or pages)**, with the title of the report, the version number, the date of completion, the current lead author (ie the person who will ultimately be accountable for the report) and their address followed by past lead authors of earlier versions, and the names and addresses of all contributors to the report (all versions), including those responsible for the research embodied by the report. It is recommended that in addition to their name all authors use an employee code, or similar, so they can be more unambiguously identified.
2. **A keyword page or pages**; this will contain a table summarising and classifying the innovation described in the report. The contents of this page will need to be finalised at the beginning of the programme and will subsequently need to be updated in accord with developments in science, technology, agronomy and commerce. The keywords should be

sufficient to allow the innovation described in the report to be broadly classified so should include (if known), *inter alia*: the target species; the nature of the innovation in the target species (GM, gene-editing, natural variation); the names of all species and cultivars/accessions used in the research, or population names; the target trait (in text), the gene code or QTL, the gene-product name, the pathway affected, the germplasm code for all stable genotypes produced in the research and the nature of the genetic change in these genotypes (GM, gene-edited, natural crossing, etc); the names of field trial sites used.

3. **A summary of the discovery.** A brief but complete description of the discovery including at least the species and genes involved, its impact on crop physiology or development, and its impact on yield. It should be made clear if the main discovery relates to natural variation of a gene or depends on genetic modification using a NPBT; if it is primarily a discovery of natural variation the availability of genotypes created by NPBT *en route* to testing the identity of the gene should be reported. Where possible sequence level data should be provided or linked to. All metabolic and crop growth models incorporating the effect of the genetic change should be named and linked to. The names of all genotypes of model and crop species that demonstrate the discovery or innovation should be provided.
4. **Patents and formal publications associated with the discovery.** The significance of any patents will be clear – intellectual property must be protected and the extent of this protection should be made clear. Formal publications should include any refereed publication or non-refereed publication (such as book chapters) of a similar quality. These publications will provide another perspective on the discovery and can also be seen as a measure of the quality of the work.
5. **Background and Introduction.** This should explain the background science and knowledge that led to following the research leading to the discovery. It should be written in the style of an Introduction to a paper or grant application, and should be supported by references. This Introduction should make clear the evidence and reasoning that inspired the research that produced the discovery.
6. **Background Knowledge.** The background knowledge and other intellectual property of each participant (or group of participants) that contributed to the discovery and of which account must be taken in any valorisation of the discovery.
7. **A report of the discovery.** This section should explain in detail the experimental procedures and material used in the research, the evidence produced by this research, and

the logic that this evidence points to the conclusion that there has, in fact, been a useful discovery. It should specify clearly what this discovery is in terms of genetics – such as the gene code and allele number that produces the improved phenotype. It is important that the report is transparent which means that the reader should be able to find within the report, or have access via the report to linked files, the details of experimental procedures used the equipment used should be described, the primary experimental data and metadata, and how the data was analysed. This means making clear all the numerical methods or software used and ensuring that any user-written scripts, spreadsheets etc are archived and accessible via hyperlinks embedded in the document. The results arising from the analysis should be clearly connected to primary data and intermediate data, if possible within the report itself but if this is not possible because of the space required then the chain of data should be documented in a linked document. These linked documents or data should all have DOI numbers and be located within programme data-archive (ie the report will be integrated with the data archive). It is important that anyone reading the document be able to, if they would wish to, redraw any graph or recreate any table starting from any point in the data hierarchy that begins with primary data, and they should be able to understand how that primary was obtained. It should be clear who authored each section of the report of the discovery.

2. The mathematical models

The future crop-yield improvement programme will make use of mathematical models both as heuristic tools and as a means to combine, collate and distribute knowledge. These models will extend across scales and will connect the basic processes of biology, such as metabolic processes or, in the case of photosynthesis, the even more basic photophysical processes, up to the level of crop-canopy level processes. This acknowledges the reality that while we manipulate and analyse plants at the molecular level, the canopy is the level of crop production, so we need to extrapolate the molecular scale to that of the crop canopy. This analysis across scales needs to be reflected in the reporting of the effect of any genetic discovery or innovation. While the report will be a comprehensive description of the discovery, describing it in detail, including the background, methods, evidence and outcomes, mathematical models will report the impact of the genetic change at the level of the models used to describe the operation of the plant. These models will allow any informed user to evaluate to a limit set by the state of the art in modelling the impact of the genetic discovery at the level of metabolism and crop growth and yield. We envisage that the effect of the change will be reflected in specific modules of metabolic, and other, models. These modified modules alone will define the impact of the genetic change at the mechanistic operational level. They should also be portable *in silico* allowing, for example, the impact of a change to be explored within a model built for another species. It is acknowledged that modelling at this level will be challenging but we see this approach to crop improvement as being a natural development and an extension of approaches that have been proven to be useful in identifying options for improvement of photosynthesis, for example.

3. Biological models and assets.

It is obvious that a product based on a genetic innovation should be demonstrated using genotypes that embody the innovation. In the case of an innovation based on a novel plant breeding technology (and note that even a discovery based on natural variation will probably make use of genetic variation at some stage) the genotypes will be genetically modified individuals with the transgene, edited gene (etc). These genotypes could be of model species (eg *Arabidopsis*), but should include, in a timely manner, crop models (eg *Nicotiana*) or crop species. In the case of a discovery based on natural variation there should be genotypes that demonstrate the effect of the elite allele. This requirement may well be fulfilled by genetically modified etc genotypes produced en route to proving the identity of the causal gene. If not, then a genotype that shows the effect of the elite allele must be produced by either a NPBT (preferably by gene editing) or by conventional breeding.

It is interesting that while we are making progress on ensuring that data is well curated and open, the guarantees that the biological materials associated with this data will remain available are weaker. As a general principle we believe that biological materials should remain available but biological materials are diverse and are generally perishable while data can be translated to binary assets and stored safely and cheaply in the long term. The complexities of systematically storing and consequently making available biological material following a reasonable request are challenging. Clearly seed of any plant model connected to the genetic discovery (this may embody the discovery but may have been created for other reasons) should be stored and treated as a germplasm asset. Storing seed is relatively easy but not all genetic modifications are stable. Materials can be stored in cold storage but there are practical to the amount that can be stored in this way. Biological materials could also be regenerated using descriptions of their composition (eg nucleotide sequences). Given the complexity and possible cost of archiving biology in the comprehensive we envisage archiving and making available data it is probably best to convene an expert group to consider what might be practical. For practical reasons it will probably be necessary to have some assessment of the biological assets created during a research programme with a view to deciding what should be stored and what can be discarded.