



**European Cooperation
in the field of Scientific
and Technical Research
- COST -**

Brussels, 11 December 2008

Secretariat

COST 271/08

MEMORANDUM OF UNDERSTANDING

Subject : Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action TD0801: Statistical challenges on the 1000€genome sequences in plants

Delegations will find attached the Memorandum of Understanding for COST Action TD0801 as approved by the COST Committee of Senior Officials (CSO) at its 172nd meeting on 24-25 November 2008.

MEMORANDUM OF UNDERSTANDING

For the implementation of a European Concerted Research Action designated as

COST Action TD0801

STATISTICAL CHALLENGES ON THE 1000€GENOME SEQUENCES IN PLANTS

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1. The Action will be carried out in accordance with the provisions of document COST 270/07 “Rules and Procedures for Implementing COST Actions”, or in any new document amending or replacing it, the contents of which the Parties are fully aware of.
2. The main objective of the Action is to use and/or develop, through coordinated international efforts, efficient statistical and bioinformatics tools and strategies in order to produce, assemble, analyze, and integrate high-throughput genomic sequence data, aiming at a better understanding of biological systems in plants.
3. The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 64 million in 2008 prices.
4. The Memorandum of Understanding will take effect on being accepted by at least five Parties.
5. The Memorandum of Understanding will remain in force for a period of 4 years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of Chapter V of the document referred to in Point 1 above.

A. ABSTRACT AND KEYWORDS

New sequencing technologies either currently available or under development will eventually enable eukaryotic genomes to be sequenced for less than 1000 euros. This technology-push will have a major impact on plant genomics and biological research and lead to a dramatic expansion in both the availability of sequence data and the range of sequence based applications. New innovative techniques are required to unlock the information contained in the sequence data and to apply the acquired knowledge for plant science and crop improvement. The wide variety and often unique characteristics of plant genomes pose additional challenges and opportunities.

The need for and the dissemination of efficient strategies for handling and analyzing high throughput sequence data in plants requires cooperation at the international level to develop new approaches & analytical tools and share best practice. This COST Action will establish a network of researchers that coordinate, focus and strengthen national and pan-European statistical genomics and bioinformatics. It will be built on close interactions with other disciplines such as genetics, genomics and breeding. The Working Groups will arrange workshops, Short Term Scientific Missions, a website and Wiki, training courses, and publications to disseminate aims and achievements.

Keywords: Next-Generation Sequencing (NGS) technology, plants, genomes and genetics, high dimensional data, statistical analysis & data mining

B. BACKGROUND

B.1 General background

The commercial availability of next-generation sequencing (NGS) technologies that are up to 200 times faster and cheaper than conventional Sanger Sequencing is already leading to the establishment of many ambitious new genome sequencing projects. These NGS technologies will have a dramatic impact on the field of genetics and create major challenges to fast and accurate assembly, analysis and integration of genome sequences. In humans, spirited initiatives have been launched to further decrease the costs of sequencing, e.g., the 1000 dollar genome initiative. This COST Action is based on the vision that availability of the 1000 euro genome sequence for plants will also become a reality in the near future.

Currently plant molecular biologists and geneticists are enthusiastically embracing the opportunities that arise from the availability of NGS technologies. However, many groups across Europe are independently exploring the resulting analytical challenges and starting to implement solutions based on purely local knowledge and experience. In addition, the relatively, modest costs of NGS technologies are giving the opportunity for research groups and crops with only modest statistical and bioinformatics resources available to them to implement sequence based genomics solutions to their research challenges. The COST Action mechanism provides an ideal vehicle to meet the unique needs of the plant research community in providing a coordinated approach to:

- the development of a European-wide community within which both researchers in Statistical Genomics and Bioinformatics together with Plant Biologists and Geneticists can respond to the changes of Plant NGS Data
- the evaluation of existing statistical and bioinformatics tools
- the development and optimization of statistical bioinformatics methodology and algorithms to deal with plant NGS derived data

- the establishment of best practice for the analysis of plant NGS sequence data
- ensuring that early stage researchers gain support for career development in this area
- ensuring that EU Plant Research is able to efficiently exploit the opportunities that arise from the availability of both current and future NGS technologies

In the knowledge-based economy fundamental discoveries will drive the EU competitiveness and lead to innovative new products. The cutting edge of plant research is rapidly evolving from understanding the function of single genes to studying networks of genes that control complex biological processes. Full genome sequences provide the list of parts that contribute to building and sustaining plant life. However, statistical and bioinformatics tools are required to organize and validate the sequence information, and discover the fundamental gene networks that underlie plant characteristics and properties.

The resulting rapid expansion in the volume of available NGS data pose big challenges to data management and analyses. Many of the major Information and Communication Technology (ICT) challenges seem common to micro-organisms, animals and plants, and major efforts to accommodate these are ongoing (www.healthtech.com/sqe/overview.aspx). However due to both size and complexity plant genomes pose several unique challenges, for example, many crop plants are polyploid, which is a major complicating factor for identifying genes, gene duplication events and allelic diversity. Furthermore, the rapid expansion of repeat sequence families in plant genomes result in major complications in such aspects as assembly and gene finding. Major challenges in statistical genomics and bioinformatics will arise that are common to many NGS projects in a range of plant species. On the other hand, the flexibility of breeding strategies and the high genetic diversity in plants provide opportunities to design strategies and use material that can corroborate statistical methods and approaches for reliable identification of genes, genetic variation and assessment of gene function. This COST Action aims to defragment European research activities in this vibrant area and to strengthen Europe's position in plant genomics and breeding.

Plants offer a wider array of possibilities for genome comparisons, functional genomics studies and experimentation than is possible in human or animal genetics: for many crops it is facile to discover and generate allelic variation at the sequence level both within and between closely or more distantly related species or within and between genomes of autopolyploid or allopolyploid species as well as between duplicated regions in diploids or ancient polyploids. For many species core collections of genotypes with known population structures are available, allowing the application of association mapping as sufficient genotype and phenotype data become available. More uniquely, it is easy to generate experimental and or pedigreed populations; in many species selfing or doubled haploidy can be used to create immortalized homozygous inbred lines on an extensive scale which can be maintained and bulked to enable replicated testing at multiple locations and over multiple time points as well as extensive multi-trait phenotyping. The same utility can be achieved in some species through facile cloning methods. Genetic modification can be studied more easily than in many animal models, both directly at the sequence level as well as downstream (the transcriptome, proteome, metabolome and more generally the phenotype). Mutant populations, based on chemical and irradiation mutagenesis or on transposable elements (e.g. in Maize) and near-isogenic lines are available, and in many agricultural crops introgressed regions from wild species are known and can be studied, in addition to these wild species themselves. The effects of homozygosity vs. heterozygosity can also be studied easily in experimental populations, e.g. in maize, by comparing inbred lines vs. F1 hybrids which allows the study of allele number and heterosis as well as additive and dominance effects.

B.2 Current state of knowledge

Each of the commercially established NGS technologies (Roche/454, Illumina/Solexa, ABI/SOLiD), together with the technologies under development, is able to produce a quantum change in the volume of sequence data compared with conventional approaches, however, this often comes at the expense of specific limitations on read sizes and quality and on the scale of the experimental unit (i.e. the need to sequence from few large pools). These NGS technologies hold the promise of enabling biologists and geneticists to have facile access to large volumes of sequence information at limited cost. This sequence information in principle would allow a much more targeted approach to link different types of biological data to the underlying molecular mechanisms.

However, this will require a paradigm shift in approaches to data handling, together with the evolutionary development of new statistical algorithms, bioinformatics tools and experimental protocols optimized for NGS technologies.

De novo assembly of a mammalian or plant genome on the basis of NGS data alone seems at the current time difficult because of the relatively short read lengths (~400bp by Roche/454; ~30bp by ABI/Solid and Illumina/Solexa), and traditional Sanger sequencing is still useful. Plants genomes often have numerous duplications and transposable elements and therefore raise specific assembly issues which may require specific designs such as a conventional/NGS hybrid approach. The use of paired-end reads does improve assembly quality of NGS data, but requires specific assembly tools. NGS already allows building fragmented assembly with draft quality. Currently, such sequences may result in complications for ab initio protein or RNA gene prediction since many existing methods assume error-free sequences and can be sensitive to such errors such as frameshifts or in-phase stop codons that may be generated by draft sequencing.

Re-sequencing (relative to existing high quality EST or genomic sequence) currently underpins the most widespread applications of NGS. Re-sequencing is used for polymorphism detection (SNP, indels or other local rearrangements) and transcriptome analysis, but also, for example, in ChIP-Seq (combining sequencing with chromatin immuno-precipitation (ChIP) assays) where ultra-short read lengths (~ 30bp) are sufficient because the reads only need to be long enough to find a unique match in the available genome sequence. Short read lengths may also be unambiguously valuable for ncRNAs such as mi- or siRNAs. Transcript tags can be useful for gene prediction, the study of alternative splicing, and for gene expression as a sensitive alternative to microarrays. Again computationally efficient statistical analysis for NGS based approaches in these areas and comparison with microarrays are needed. Currently the analysis of NGS based re-sequencing data may require significant volumes of CPU time with sequencing errors crudely handled by either massive redundancy or by rejection of incomplete or ambiguous data. These problems are exacerbated with ultra-short reads.

The first two plant genome sequences to become available (*Arabidopsis thaliana* and *Oryza sativa*) were based on conventional BAC by BAC Sanger Sequencing (SS) approaches and others such as Poplar and *Brachypodium* have made use of shotgun SS supported by paired end-read strategies. However, these strategies, despite establishing an apparent Gold Standard approach, still have some major limitations for some of the more important crop genomes due to their size and complexity. This has led to, for example, the US Maize Genome sequencing strategy which is based on targeted finishing of regions containing genes predicted from initial low pass sequence coverage of the genome.

Many efforts are currently undertaken to sequence genomes of other key food and biofuel crops and model plant species, including *Medicago*, tomato, potato, soybean, apple, barley, wheat, maize, oil palm, cassava, cocoa and many others. These include both diploid and polyploidy species. In addition, numerous EST sequence databases as well as gene-enrichment sequence libraries based on conventional SS strategies are publicly available and many laboratories are already investing in Roche 454 based EST or Solexa sequencing projects in a wide variety of species.

This rapidly accumulating reservoir of sequence data will increasingly offer unique opportunities for comparative genetics and genomics within the plant kingdom. The biodiversity between plant species and the comparison of their genomes and transcriptomes will enable the elucidation of molecular mechanisms, genes and gene networks that underpin both intra- and inter- specific diversity.

Statistical methodology supporting selection decisions as used by plant breeders is shifting away from classical phenotypic selection to more pedigree and marker based breeding value prediction and selection or a hybrid approach of the two. However, to date, most practical applications are still limited to single-marker based selection of Mendelian-inherited traits, i.e., single gene traits such as disease resistance loci. Progress in the deployment of marker-based selection on complex traits, affected by multiple genes and environmental factors, is currently limited both by lack of high density / high quality marker data and efficient statistical tools and software. The NGS data

provides a route to removal or minimization of the first hurdle but the second hurdle will still require considerable research investment by statistical geneticists and bioinformaticans to develop new tools and strategies to maximize the value from NGS and high throughput genotyping approaches.

In plants, most progress in modeling sequence to phenotype has been reported in the model plant *Arabidopsis thaliana* because of its unique biological properties and extensive genetic resources that are available, including the high quality finished genome sequence. Comparative approaches are likely to be increasingly productive with the availability of new sequence information. Polymorphisms in coding sequences may affect protein function and polymorphisms in regulatory regions may affect expression levels of genes both of which may introduce phenotypic variation. Integrative analyses of multiple levels of organization of biological processes may well be required to unravel the genetic architecture of phenotypic traits in plants.

B.3 Reasons for the Action

The plant industry forms a key role in the economy and food production in all European states. Full genome sequences and sequence based genomics resources funded at national and international levels will significantly advance plant research, resulting in new plant varieties with higher yields, increased nutritional value, new properties for industrial use, and more stable and sustainable growth. This will enable the plant breeding and processing industries to respond more rapidly to new demands for food and raw material resources in the face of pressures such as climate change. To make full use of the potential of sequence information within the EU, innovative statistical and bioinformatics tools need to be developed and utilized efficiently. This COST Action will establish a network to facilitate communication and transfer of knowledge between different research groups developing and utilizing statistical and bioinformatics methodologies and software tools for plant sequence data that are working largely independently all over Europe.

This COST Action is very timely as it identifies the need for a collaborative effort to tune, develop and optimize statistical and bioinformatic tools for NGS prior to the potential of NGS to generate significant volumes of sequence data in plants being fully realized. This will serve to reduce the duplication of research within Europe and will create synergy between scientists from different disciplines (working on different topics) by bringing them together to identify and advance unified approaches. This Action will also provide a framework for the rapid dissemination of research results and the development of best practice for utilizing NGS data to end users, such as plant geneticists, molecular biologists and breeders.

The main advantages of this COST Action may be summarized as:

- Avoiding duplication of research in different European countries.
- Better communication between the research groups involved.
- Faster decision making with regard to the prioritization of research areas and directions
- Facilitation of sharing of novel statistical and bioinformatics techniques and software between different groups and efficient deployment of research tools between different labs in the COST member states via Short Term Scientific Missions (STSM).
- Establishment of training schools to share know-how and training of early-stage independent researchers in advanced statistical and bioinformatics concepts relevant to next-generation sequence (NGS) data in a plant context.
- Efficient dissemination of acquired knowledge and increased public awareness of this key area through a dedicated website and wiki.

B.4 Complementarity with other research programmes

No other current or planned European research programme exists that has the same objectives and benefits as this COST Action. However, this COST Action is complementary to several ongoing EU Framework research projects, such as:

- EU-SOL - Integrated Project (FOOD-CT-2006-016214, www.eu-sol.net) that aims to develop high quality tomato and potato varieties with improved traits important for consumers, processors and producers.
- EU-EVOLTREE - Network of Excellence (www.evoltree.eu/) & EU-NOVELTREE - Collaborative Project that aims to analyze the impacts of climate change on forest ecosystems from an evolutionary perspective.
- EU-SPICY - Collaborative Project (SP7-KBBE-2007-1) to develop a suite of tools for molecular breeding of crop plants for sustainable and competitive agriculture. The tools help the breeder in predicting phenotypic response of genotypes for complex traits under a range of environmental conditions.
- EU-TRANSISTOR - RTN project (www.transistor-arabidopsis.org/) that aims to elucidate the regulatory gene networks controlling flowering processes in plants and to train young scientists in this area.
- EU-TriAnnot Life grid project (<http://urgi.versailles.inra.fr/projects/TriAnnot/>) that aims to develop a semi-automated pipeline for wheat genome annotation.

In addition, the COST Action is complementary to several worldwide research programs, such as the:

- Potato Genome Sequencing Consortium (www.potatogenome.net), with the primary objective to elucidate the complete DNA sequence of the potato genome (850 Mbp) by the end of 2010.

- Tomato Genome Sequencing Initiative: In a worldwide initiative all tomato chromosomes are currently sequenced.
- International Wheat Genome Sequencing Consortium (www.wheatgenome.org) that aims to enhance our knowledge of the structure and function of the wheat genome to enable plant scientists and breeders to accelerate wheat improvement to meet the challenges of the 21st century.
- International Barley Genome Sequencing Consortium (www.barleygenome.org). The objective of the IBSC is to physically map and sequence the barley gene space to accelerate crop improvement.

Furthermore, this COST Action will complement ongoing many trilateral / bilateral / national research projects, e.g., Centre for BioSystems Genomics, SYNBREED , GABI-BeetSeq, Celiac Disease Consortium, Whole genome sequencing in Apple and Grape; Decoding the grape genome; GrapeReSeq; TRANSNET, etc.

C. OBJECTIVES AND BENEFITS

C.1 Main/primary objectives

The main objective of the Action is to use and/or develop, through coordinated international efforts, efficient statistical and bioinformatics tools and strategies in order to produce, assemble, analyze, and integrate high-throughput genomic sequence data, aiming at a better understanding of biological systems in plants.

C.2 Secondary objectives

[WG1] Designs and analysis of NGS data

- I) Develop and maintain information on cost-effective designs for NGSbased genome sequencing and re-sequencing, SNP identification, transcriptomics and ChIP-Seq analysis for a wide diversity of plant species and their characteristics, including comparison with alternative micro array technologies.
- II) Promote the uptake and development of tools and algorithms to efficiently analyze high volume NGS data, including assembly, gene prediction, analysis of splice variants, SNP and indel identification, EST and TAG -based transcriptomics.
- III) Evaluate the robustness of existing and novel designs and statistical tools in the context of sequencing errors and investigate error handling approaches for dealing with NGS data as indicated in II.

[WG2] Sequence To Phenotype Integration

- IV) To list and assess the existing statistical tools and algorithms for forward genetics (phenotype to sequence), including QTL mapping and genomic selection tools in a context of NGS data.
- V) To increase our knowledge of effective and efficient statistical modeling of sequence (candidate genes) to phenotype (bottom-up), frequently including integration of data from intermediate levels of organization, e.g., transcriptomics, proteomics and metabolomics, leading to a systems biology approach.
- VI) To evaluate the effectiveness of NGS data in efficient plant population structures, e.g., Mutants, Near Isogenic Lines, segregating populations, which will allow study and validation of various hypotheses on the modeling of sequence to phenotype relationships.

Although the Action covers a large number of problems at different levels (as structured in the WGs), they all carry as their central theme the problems and benefits that come from the extensive sequence volumes that will increasingly be possible through the use of NGS technology. The network will allow people to share data and experiences, as well as modeling and solving issues in this competitive and challenging area. Integrating this data at a systems level by statistical modeling will be a major scientific challenge that is beyond the skills and resources of any one research team.

Thus, this COST Action aims to establish and extend a network among European scientists to collaboratively evaluate and develop novel strategies to overcome inadequacies in statistical genomics and bioinformatics to enable extensive exploration and exploitation of the large volumes of NGS sequence data that are becoming available and harvest their full potential effectively and efficiently.

C.3 How will the objectives be achieved?

This COST Action will maximize its outcomes by:

- The development of a collaborative research community of both methodology developers and current and potential users bringing the required disciplines (statistics, bioinformatics, ICT, genetics, breeding, etc.) closer together.
- An evaluation of existing statistical and bioinformatics methodologies and the identification of research priorities to meet the developing needs of the plant research community in response to the rapid development and deployment of NGS technologies through meetings, focused workshops and the development and maintenance of whitepapers on the web-portal wiki.

- The use of Short-Term Scientific Missions (STSM) to meet specific challenges and to provide a vehicle to support the training of early stage researchers.
- The development of a web-portals to act as a vehicle for discussion and development of ideas and as a primary focus for best practice guidelines and recommendations and to provide access to the scientific literature, algorithms and software tools.

C.4 Benefits of the Action

This COST Action will lead to:

- A better understanding of the bioinformatics and statistical tools that are currently available or yet to be developed for the analysis of large volumes of NGS data, and their integration with other genomics data types (from the transcriptome, proteome, metabolome and more general phenotypic traits), at the level of single species or for comparative studies across species.
- Guidelines for the development of methodology as well as software to carry out such analyses.
- An effective European network that can deal with the new statistical and bioinformatics challenges posed by high-throughput NGS technologies.
- Dissemination to the scientific community, policy makers and the industry (e.g. breeding companies) of the current and future possibilities of NGS.

The availability of NGS and effective statistical tools will allow much faster determination of effects of alleles on plant phenotype and of assessing allelic variation across accessions; this will be beneficial to researchers and breeders. For example, results from a QTL analysis can much faster be transferred across large numbers of accessions that are or can be used in breeding programs.

In addition, interactions of specific alleles can be assessed and predicted more easily and with the availability of the sequence information experimental populations can be generated to study these.

C.5 Target groups/end users

The target audience may be categorized into three main groups:

- **Developers:** statisticians, bioinformaticians, computer scientists that develop new algorithms and make these accessible by user-friendly software tools. Documents and guidelines produced as a result of expert discussion and debate during this COST Action will be of immense importance to the researchers in this field. It will allow moving towards a set of standardized conditions for experimentation and comparison between results obtained from different labs in different COST countries. It will minimize waste of European tax money by preventing duplication of the development of new analysis tools in different laboratories.
- **Experimenters:** plant biologists and geneticists utilizing NGS data to study biological systems in plants. Recommendations for cost-efficient designs and suitable analysis tools will be disseminated among plant biologists accelerating plant biology research.
- **Implementers:** plant breeders policy makers. The results obtained will improve selection for complex traits and catalyze innovative breeding strategies. It will enable the breeders to accurately and efficiently breed for improved crops and vegetables. Breeding programs that target complex traits such as stress tolerance, yield and durable resistance are expected to benefit the most. Documents and guidelines prepared as a result of this COST Action will be valuable to policy makers both at the level of national governments or the European level. These guidelines can be used for decision making regarding research funding activities in member states at national and European level. In addition, the results can be further used in planning and directing plant breeding activities in European community.

In addition, this COST Action will be beneficial to:

- **The environment.** Better and improved breeding programs will be devised, which will lead to efficient production strategies yielding genetically improved plant varieties and reducing the need for pesticides and other chemicals.
- **Natural resources.** Efficient use of NGS data will enable efficient development of crop varieties that require less input (water, chemicals, nutrients), which reduces strain on the environment and on the earth's natural resources
- **Public health.** The results obtained will improve quality traits in plants to develop and produce sufficient, diversified and affordable high-quality plant raw materials for food products.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

The COST Action will help facilitate the efficient unlocking of the extensive genomic datasets that will arise from new cost-effective next-generation sequencing (NGS) technologies. The scientific program is focused on two research areas that exhibit a logical progression but with strong links and interrelationships (see also Figure 1).

(I) Designs and analysis of NGS data

WG1 will coordinate the development of new powerful designs together with efficient statistical and bioinformatics tools to address the challenges raised by NGS technologies in different areas of application. The data volumes, read lengths and error profiles are specific to each existing NGS technology. Specific designs and analysis are needed, taking into account unique features of plant genomes as well as the heterogeneity of NGS data as well as utilizing hybrid data generation strategies.

De novo sequencing and assembly of complete plant genomes: Due to their high level of duplications and extensive complex repeat expansions, many key crop plant genomes are at the boundaries of what is possible with current NGS technologies and software. Adapted cost-effective designs/protocols and algorithms for complex plant genome sequencing and assembly are needed, though a number of the new de Bruin graph-based approaches have considerable potential. The performance of existing gene prediction tools on plant NGS has not been fully evaluated although it is clear that different types of errors such as indels in coding regions will dramatically influence their performance. However gene, mRNA and protein prediction will increasingly be based on the production of draft and possibly fragmented snapshots of NGS derived genomic data for many plant species. Since this resource represents the fundamental pieces of information required for system biology approaches (WG2), there is a need to evaluate the ability of current gene prediction tools (ab initio, or more integrative) to handle these types of data. Due to the great taxonomic diversity of crop plants, the absence of closely related genomes is likely to be a more serious issue than that encountered in humans and domestic animals. This will confound a number of the problems that will arise from NGS technologies and increase the pressure to develop new statistical models and tools capable of dealing with sequencing induced errors and other unique features of NGS data.

Re-sequencing: Classical alignment tools have not been optimized to deal with the unique features of NGS data or the high volumes of data. Alignment itself may be more difficult because of the higher level of SNP polymorphism observed in plants (even within species) together with the high frequency of duplications and other types of re-arrangement, including those induced by transposons. There is therefore a need for a comprehensive review of the existing approaches for polymorphism detection, depending on the type of polymorphism analyzed (SNP, indels etc.) and on the re-sequencing technology used. Furthermore, error rates and technology specific sequencing chimeras may obfuscate polymorphism detection. There is a need to develop an understanding of the error characteristics of NGS data through statistical modeling for each technology in order to avoid erroneous experimentally induced polymorphisms. In relation with WG2, association screening is facilitated by designs based on DNA pools and existing dedicated designs for NGS data should be reviewed and developed.

Transcriptomic applications of NGS: NGS has considerable potential for applications that involve RNA sequencing including those in which ESTs or short sequence tags are derived from cDNA or alternatively those based on de novo sequencing of short regulatory miRNAs, siRNAs or other species of non-coding RNAs. The increased depth of sequencing coverage that is possible for NGS - either of single pass EST equivalent or short (SAGE or MPSS equivalent) tags - allows for the detection of relatively rare transcripts from unique tissues or developmental stages and novel possibilities for quantification of transcript abundance as a sensitive alternative to microarray analyses. Here again there are major issues of experimental design and analysis methodology to be resolved. The resulting data will increasingly be utilized in the formulation of regulatory networks based on differential expression of both protein coding and regulatory RNA genes.

Phylogenomics: The availability of whole genome data opens up new opportunities for inferring species relationships and for improving functional predictions of uncharacterized genes by evolutionary analysis. Whole-genome phylogenetic inference approaches (using concatenation, consensus and supertree estimation strategies) can be improved by the application of recent model-based methods including a hierarchical substitution/tree incongruence model and a Bayesian lineage sorting model that accounts for the stochastic variation expected for gene trees from multiple unlinked loci sampled from a single species history after a coalescent process. Within the concatenation approach, there are opportunities for improved substitution model choice by optimizing across all loci rather than each locus individually. The incidence of hybridization and introgression in plants requires improved visualization of gene trees from all loci to check for lateral transfer events which may lead to errors in phylogenetic inference if undetected. Detailed study of multi-gene families can lead to improved understanding of function and lead to improved annotation. One methodological opportunity is though the detection of positive selection by improved modeling of rate variation in synonymous positions.

(II) Sequence To Phenotype Integration

A highly exciting aspect of the new sequencing era is the integration of genomic sequence data with extensive phenotypic data. Currently, considerable effort of statistical genetics is aimed at finding quantitative trait loci in genomes, i.e. regions in the genomic DNA that are somehow associated with changes in phenotype.

Molecular marker Phenotypic trait association studies. A variety of approaches have recently been presented to analyze multiple related populations jointly or, to study gene by gene interactions, to study gene by environment interactions or gene response curves and to combine linkage analysis and linkage disequilibrium mapping. The number of genetic markers in this type of analyses is rapidly increasing both in association mapping studies and in QTL linkage studies (e.g. www.gramene.org). For example, the seminal Wellcome Trust Case Control Consortium used the Affymetrix GeneChip 500K Mapping Array Set for a genome-wide association study in common human diseases, reported single-point P-values for association signals. A multipoint imputation method for genome-wide association may allow meta-analysis but will dramatically increase the dimensionality of the search space. The huge numbers will require more efficient statistical analysis tools that integrate the many studies using different populations, traits and locations. Genomic selection has recently put forward to use all markers simultaneously in a Bayesian model selection framework. While some suggestions, such as LASSO and non-parametric Bayesian model selection, have recently been made, further improvements will be fundamental to the efficient harvest of the available sequence polymorphism data. Bulked Segregant Analysis (BSA) involves comparing only two pooled samples and thus requires only a few sequencing runs. Each bulk consists of members from a segregating population that are alike for a particular trait or genomic region and heterogeneous at unlinked loci. The major challenge in BSA is to find the right balance between the number of bulked individuals and genome coverage by NGS to fine-map the trait of interest. In all types of association study approaches the determination of significance thresholds, such as permutation tests or False Discovery Rates, becomes an even bigger challenge because of the high dimensionality of the NGS data.

Genetical genomics, genetics of regulatory networks, epigenetics, and systems biology: The direct use of sequence information triggers a completely new brand of statistical genetics as applied to plant data where phenotypic information will be linked directly with the allelic information of the genes or regulatory elements themselves. Genome-wide expression sequence analysis of fully sequenced genomes will enable the identification of correlations between phenotype and transcriptional regulation. Such type of analysis should also be capable of handling interactions with environmental and genetic background effects. Similarly, the recent development of genetical genomics is aimed at the simultaneous exploitation of sequence polymorphisms observed in different genetically related genomes, phenotypic variations and transcriptomic (or proteomic and metabolomic) data in order to unravel the gene networks based on common genetic regulation. The dimension of this complex puzzle may be further extended by including epigenetic control of transcription. Substantial progress in regulatory network construction was already made by expression and metabolic profiling of deletion strains or using double mutants. The application of such a detailed genome-wide analysis to plant species becomes even more challenging because of higher complexity of genetic architecture and cellular composition. NGS will become a natural competitor of the popular microarray data to infer gene regulatory networks. The problem may be statistically similar for microarray and NGS data, although the methods developed for microarray will need to be adapted (1) to be efficient for very large data sets; (2) to deal with counts of sequences rather than with a continuous signal; and (3) to handle the (still unknown) bias of NGS data.

Experimental designs for sequence to phenotype integration: Plant population genomic data provide an excellent means to study sequence to phenotype relations, as the size of progeny families can easily be manipulated to be large, the generation cycle for many plants is relatively short, there exists a wide variety of reproduction mechanisms and the phenotypic responses can efficiently be recorded and studied under a large number of different environments. The usefulness of combining quantitative genetics in mapping populations and large-scale -omics analyses has been illustrated by the successful construction of genetic regulatory networks. Novel statistical approaches for integrative multi-level modelling have recently been suggested, however, further improvements are required to successfully exploit NGS data.

D.2 Scientific work plan – methods and means

The scientific goals of this project will be realized by five key methods: organizing specialized workshops and meetings, training of young-researchers, short term scientific missions, dissemination of information via dedicated website(s), and publication of specialized documents as white papers from the Working Groups.

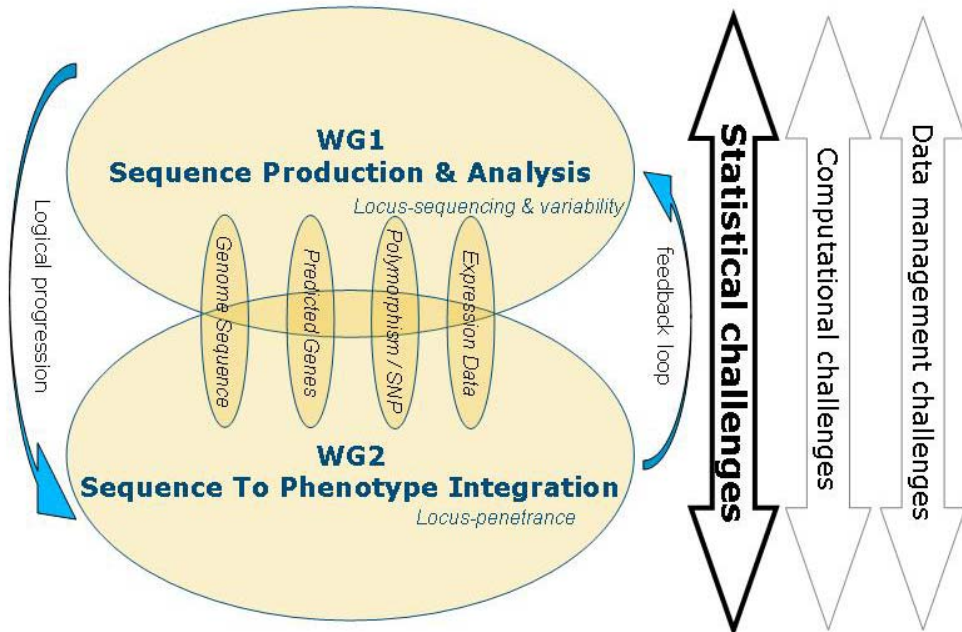


Figure 1. Sequence and interactions of Working Groups

Thus WG1 will work on designing experiments to produce, assemble and analyze plant data while WG2 targets efficient integration of data from sequence level up to phenotypic trait level. The WGs will be the driving forces behind the Action. Their objectives and deliverables are:

WG1. Designs and analysis of NGS data

The objectives of this WG are objectives (I) to (III) mentioned in section C.2 Secondary Objectives. These objectives will be achieved via the following deliverables,

- [A] To organize scientific workshops on designs and analysis of NGS data at year 1 (objective a1) and at year 3 (objective a2). The arrival of NGS data defines new design and analysis problems in all areas; each workshop will be dedicated to the three objectives mentioned in section C.2. The first workshop will concentrate on the performance and limitations of currently available statistical tools. The second workshop will focus on the development of new ideas and approaches to overcome encountered and expected limitations. Workshops will be co-located in space and time with workshops of WG2, preferably within a large plant genetics/genomics conference (e.g. PlantGEMS) to provide opportunities for exchange.
- [B] A whitepaper will be produced and developed and maintained on the PlantNGSWiki after each workshop (objectives b1 and b2) summarizing the gaps between existing techniques and specific needs for NGS designs and sequence analysis as well as plant specific issues.
- [C] On demand, but at least yearly a competitive call for Short Term Scientific Missions (STSMs) (objective c1-4) supported by the Action will be held, with preference given to early-stage scientists, to facilitate knowledge transfer on the design and analysis of NGS data.
- [D] A frequently updated list of references to state of the art designs and analysis tools and their applications to NGS data on plants (objective d) will be collaboratively maintained on PlantNGSWiki attached to the COST Action web-portal, providing links to key publications and as well as existing designs and software tools.

WG2. Sequence To Phenotype Integration

The objectives of this WG are objectives (IV) to (VII) mentioned in section C.2 Secondary Objectives. These objectives will be achieved via the following deliverables,

- [A] To organize focused workshops: one at year 1 (objective a1) and one at year 3 (objective a2) where the 1st workshop will concentrate on the performance and limitations of currently available statistical tools to link phenotypic variation to sequence variation (i.e., large amounts of SNPs) or transcript abundance. The 2nd workshop will put forward new ideas and approaches to overcome encountered and expected limitations.
- [B] To develop and maintain a whitepaper on the PlantNGS Wiki covering existing statistical tools and algorithms for modeling phenotype to sequence, including QTL mapping and genomic selection (objective b).
- [C] To yearly hold a competitive call for STSMs (objective c1-4) , with preference given to early-stage scientists, to facilitate knowledge transfer with regard to statistical modeling of sequence to phenotype integration within the Action.
- [D] To organize in year 4 an inter-disciplinary Training School on Systems Biology focusing on sequence to phenotype modeling with multiple organization levels (objective d).
- [E] To develop and maintain on the Plant NGS Wiki a whitepaper specifying guidelines for designing efficient plant population structures that are useful when modeling sequence to phenotype relationships (objective e)

E. ORGANISATION

E.1 Coordination and organization

This COST Action is a means of coordinating several research projects in different laboratories in different European states that are already financed by their respective governments. Via this COST Action, these efforts will be coordinated to increase knowledge of and accessibility to sound statistical tools to efficiently exploit biological information from high volume NGS sequence plants

data to the mutual benefit of the broader European Plant Research Community. The coordination of research and exchange of data between laboratories in different European countries will be achieved by realizing the common milestones/deliverables of the different WGs as described in section D.2 Scientific work plan methods and means.

The organization of this Action will follow rules and regulations set by COST guidelines (COST doc. 270/07) to be applied in Action specific ways as appropriate.

The Management Committee (MC) will be convened by representatives of the participating countries as described in the COST guidelines. The MC will elect a Chair and a Vice Chair by majority vote. The MC will set up two Working Groups (WG). MC will appoint a leader and deputy leader for each WG. Meetings of the MC will take place at least once each year, preferably linked to a WG meeting / workshop. This will ensure efficient use of budget, coordination of the activities and discussion about the objectives and critical issues of the Action.

The Working Groups (WGs) will also meet during MC meetings and discuss the progress and coordination of their specific WG objectives. WG leaders will report the progress of their respective WG based on the predefined timetable to MC. The WG leaders are responsible for achieving goals and milestones of each WG and for reporting their achievements to MC.

A dedicated website/portal will be established to provide information about the COST Action as well as to document its progress and the achievement of its objectives. It will be the responsibility of the MC Chair to organize the website and keep its contents up-to-date with the progress of the Action. The website will contain a restricted access page for registered participants of the COST Action. In addition, dedicated mailing lists will be established to keep different MC members informed regarding the progress of the COST Action and as a mean of communication between different members of MC and WGs. A key feature of the web-site will be the development of a dedicated PlantNGSWiki that participants can utilize to develop reports and whitepapers based on the outcome of workshops and discussions between participants as well as share experience and best practice through FAQs and blogs.

An ad-hoc Short Term Scientific Mission (STSM) evaluation committee will be appointed by the MC with one coordinator and one representative of each WG. This committee will favor the mobility and training of young researchers. Dependent on the budgetary possibilities a number of STSMs will be financed each year for exchange of technologies and training between labs in different member states as defined in the rules of COST. The applications for the STSMs will be sent to the Chair of the Evaluation Committee. The approval process is subject to the COST rules and guidelines.

The planning of workshops in the different WGs will be such to stimulate exchange of information and ideas among WG. That is, workshops should be organized at the same time and sufficient time between consecutive workshops within each WG is required to build upon acquired knowledge and put forward new ideas at the next workshop.

An inter-disciplinary systems biology Training School will be organized by the WG 2. The MC Chair and members will help WG 2 in organization of this Training School and will follow COST rules and procedures for its organization and selection of the participants. Again priority will be given to early stage researchers in taking part in the Training School.

E.2 Working Groups

The structure of the COST Action will be based on two main WGs;

WG 1. Designs and analysis of NGS data

WG 2. Sequence To Phenotype Integration

MC members based on their specialty and interest will take part in different WGs. Each WG will elect a leader and a deputy leader to guide the scientific progress in their respective Working Group. WG leaders will report progress on WG objectives and deliverables to MC.

The main tasks of the WG leaders are:

- Participate in plenary and restricted MC meetings and report WG progress
- Plan the appropriate scientific meetings
- Coordinate the activities within their WG in order to meet scientific objectives
- Promote the set-up of joint research (through STSM) and publications

Meetings of the WG will be organized on a yearly basis at different partner locations. These frequent gatherings are scheduled in order to have an optimal exchange of ideas. This COST Action aims to organize two to three days meetings for each WG. The first one to two days would be devoted to specific WG activities. This allows the exchange of information and ideas, encourage the collaboration between scientists and institutes, stimulate the planning of joint experimental work and will address WG specific topics. The last day of the meeting would be combined with the other WG. This would greatly enhance integration of activities from the different fields, and promote interface between WG. Indeed, the results obtained in the first Working Group on the production and analysis of NGS data will be extremely helpful during the sequence to phenotype integration in WG2. Feedback from the partners of WG2 will also be essential to define new challenges and technical requirement for the WG1.

E.3 Liaison and interaction with other research programmes

Inter-COST activities such as meetings and training will be stimulated. Particular synergy is anticipated with related COST Actions on plant genomics, e.g., FA0603 Plant proteomics in Europe (www.costfa0603.eu) and FA0604 Triticeae genomics for the advancement of essential European crops (tritigen.ari.gov.cy).

This COST Action will team up with the European Plant Science Organisation (EPSO, www.epsoweb.org), one of the initiators of the first European Technology Platforms Plants for the Future. EPSO also contributed to the establishment of the ERA NET Plant Genomics initiative (www.erapg.org). The ERA-PG already made an inventory of national plant genomics programs, research landscape, and Europe's competitiveness.

Collaboration is planned with the European Association for Research on Plant Breeding (www.eucarpia.org) to disseminate the scientific results to the plant breeding community.

E.4 Gender balance and involvement of early-stage researchers

This COST Action will respect an appropriate gender balance (for example with regard to the appointment of MC) in all its activities and the Management Committee will place the promotion of gender balance as a standard item on all its MC agendas. The Action will also be committed to the active support and promotion of Early-Stage Researchers. This item will also be placed as a standard item on all MC agendas. Furthermore, in the selection process of STSM proposal and admittance to workshops and training courses priority will be given to favor gender balance and Early-Stage Researchers.

F. TIMETABLE

The total time of this COST Action is 4 years. The time table (in months) of milestones is as follows:

Timeline	Milestones	Deliverables
M 01 - 06	- Initial MC meeting, election Chairs, etc. - Launching website - 1st round STSM applications	
M 07 - 12	- WG meetings & workshops	WG1.a1; WG2.a1; WG.c1
M 13 - 18	- 2nd round STSM applications	WG1.b1; WG2.b
M 19 - 24	- WG meetings & workshops	WG.c2
M 25 30	- 3rd round STSM applications	
M 31 36	- WG meetings & workshops	WG1.a2; WG2.a2; WG.c3
M 37 42	- 4th round STSM applications	WG2.e
M 43 48	- Final meeting / conference	WG.c4

The planning of workshops in the different WGs will be such to stimulate exchange of information and ideas among WGs. That is, workshops should be planned at the same time and sufficient time between consecutive workshops within WGs is required to build upon acquired knowledge and put forward new ideas at the next workshop.

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: Austria (AT); Belgium (BE); Finland (FI); France (FR); Germany (DE); Ireland (IE); Israel (IL); Italy (IT); Latvia (LV); Netherlands (NL); Norway (NO); Poland (PL); Spain (ES); Sweden (SE); Switzerland (CH); United Kingdom (UK). On the basis of national

estimates, the economic dimension of the activities to be carried out under the Action has been estimated at 64 Million € for the total duration of the Action. This estimate is valid under the assumption that all the countries mentioned above but no other countries will participate in the Action. Any departure from this will change the total cost accordingly.

Over 70 scientists from 35 different institutions (both academic, government and industry) in 16 European countries have actively participated in the preparation of the Action or otherwise indicated their interest. Scientists from China, New Zealand, and United States have also shown interest to cooperate and collaborate with this Action.

H. DISSEMINATION PLAN

H.1 Who?

The target audiences of this COST Action can be categorized as Developers; Experimenters; and Implementers (as specified under section C.5). Thus, next to statisticians, bioinformaticians, computer scientists, also plant geneticists and biologists will be targeted by this Action. The third group of target audience is the plant breeding industry and European and government level policy makers. Finally, the general public will be targeted as the public awareness and acceptance is key to successful implementation and exploitation of knowledge generated via this COST Action.

H.2 What?

Website: A public website/portal will be made available to disseminate a wide variety of information regarding the activities and achievements of this COST Action. This website will be open to the general public and all target audience groups of this Action. The latest results of the Action will become available through articles that are understandable for the general public. A dedicated password-protected section of the website will be used to post specialized working documents among participants of this Action. A key feature of the COST web site will be the development of a PlantNGSWiki which will be used to:

- Develop and maintain whitepapers on experimental design and analysis methodology
- Summarize properties of extant methods and analysis tools and provide routes to accessing them

Maintain lists of relevant publications

- Share experience of best practice and known issues and problems with designs and methodologies.
- Identify expertise within the European Scientific community.
- Identify priority areas for future research and development.

Publications: As described in section F production of several technical and scientific documents are important deliverables for WG in this COST Action. The Action will aim to publish these manuscripts in peer-reviewed scientific journals or proceedings of international conferences. Shorter versions of these documents will be presented in a less technical format to the general audience and plant breeders in more popular magazines. Several scientists interested in this Action are members of editorial board of different scientific journals.

International conferences: Knowledge gained from this COST Action activities will be integrated and presented at international conferences, for example those organised by European Plant Science Organisation (EPSO, www.epsoweb.org) and European Association for Research on Plant Breeding (www.eucarpia.org), especially the section Biometrics in Plant Breeding. This will promote the European know-how and increase the international collaboration.

Electronic communication network: A dedicated e-mail network allowing multiple levels will be established. One will be available for the general public, policy makers and scientists not directly involved in the Action but interested to be informed on recent achievements of the Action. A second network will facilitate specific groups such as different WG and MC. This will be used for communication and discussion between the members of these groups.

Events: The WG will organize focused workshops to integrate and promote scientific knowledge by bringing scientific experts together at a European level. Symposia and training courses will be organized to disseminate specialized knowledge to the scientific community, especially to early-stage scientists. More general and open meetings will be organized to present and disseminate new insights to the end-users, for example plant breeders.

H.3 How?

The main aim of this COST Action is to progress the efficient discovery of biological knowledge in plants derived from the substantial volumes of next-generation sequencing data than will be generated over the next 5 to 10 years. Financial support for the generation of the plant NGS data is already ongoing does not appear currently to be rate limiting. In sharp contrast, the availability of optimized statistical tools to analyze and integrate NGS data and their supported adoption by the plant biology user community is already a limiting factor. The establishment of a network of scientific experts through this COST Action will create synergy to standardize and advance expertise statistical genomics and bioinformatics to unlock the biological knowledge held within this NGS data.
