White Paper

The STATegra Consortium, September 2014

1. What is STATegra

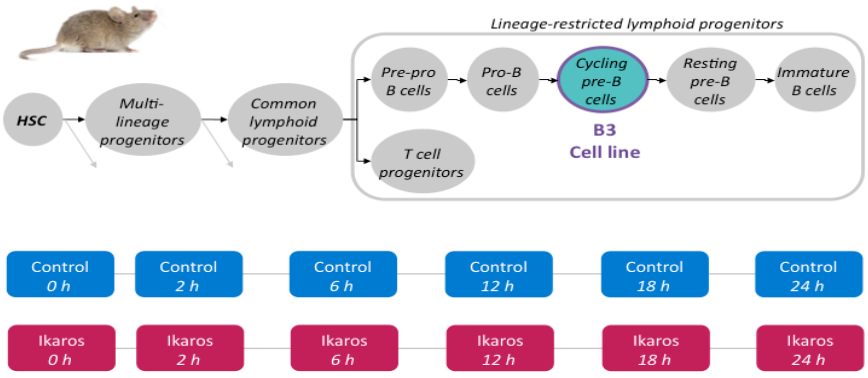
STATegra is a FP7 funded project that aims to develop statistical methods and software for the integration of NGS and omics data (http://stategra.eu). Our goal is to provide by the end of the project a set of bioinformatics resources that the genomics community can easily use to integrate and to understand their experiments involving a variety of omics measurements.
The working strategy of the project is based on a coordinated set of efforts to:

→Generate a STATegra benchmarking dataset where several omics data-types are obtained under a controlled experimental setting and use these data for method development.
→Develop integrative methods using different analysis strategies, thus leveraging the expertise of the different partners in the consortium.
→Create user-friendly tools where methods are implemented and make them available to the community.
→Disseminate the tools and results of STATegra actively to the wider genomics and life sciences community.

2. STATegra data collection

2.1 The mouse B3 cell experimental system

STATegra proposed the generation of a project dataset that will aid the development of experimental design, statistical and validation methodologies. The planned design was a time course of a cell differentiation process in mouse, specifically the differentiation of the preB-cell-like B3 cell line under the controlled induction of the transcription factor Ikaros (the "System Under Study", SuS). The consortium decided to collect a 6-point time-course with 3 replicates on two series: one with Ikaros induction by Tamoxifen and a second control series of B3 cells without inducible Ikaros. Initially we planned to obtain mRNA-seq, miRNA-seq, Methyl-seq, ChIP-seq, DNase-seq, proteomics and metabolomics data (Figure 1).



2.2 The STATegra data so far.

The time-course data collection is nearly complete with most of the planned data available (Table 1).

Technology	Planned Samples	Raw Files	Processed Files	Due complete set*
mRNA-seq	36	36 (100%)	36 (100%)	completed
miRNA-seq	36	38 (105%)	38 (105%)	completed
RRSB-seq	36	12 (33%)	12 (33%)	End September
DNase-seq	36	36 (100%)	36 (100%)	completed
ChIP-seq	NA	6 (previous data)	6 (previous data)	to be decided
proteomics	36	36 (100%)	36 (80%)	end September
metabolomics	36	48 (125%)	48 (125%)	completed

TABLE 1. STATEGRA DATA COLLECTION

2.3 Data pre-processing

While the development of routines for the pre-processing of different omics data types was not a goal of the STATegra project, we found that pre-processing was a key step in data analysis and we have spent significant time on this part of the analysis (Figure 2). This led to important insights of several issues that need to be taken into account when pre-processing large integrative datasets:

a) Batch effects: We detected batch effects in a number of technological platforms we have used, and have implemented both experimental design and analysis procedures to mitigate batch effects.

b) Normalization factors. Upon Ikaros induction cells shrinks. Depending of the technology used, this has to be taken into account differently. While metabolomics, proteomics or DNase-seq uses a fixed amount of cells as starting material, RNA-seq uses a fixed amount of RNA (involving different number of cells). Therefore a normalization on cell size needs to be applied to RNA-seq data to make these values integrable with the other omics data-types. We were able to estimate cell-size normalization factors thanks to the use of exogenous RNAs added to the RNA samples.

c) Relative vs independent measurements. While measurements for sequencing technologies provide an independent value for each sample, measurements in proteomics and metabolomics use a reference standard that is shared between each Ikaros induced sample and its matching control. This means that for these last omics technologies only Ikaros/Control ratio values and not sample-alone values are meaningful. This also affects how data must be used at integration.

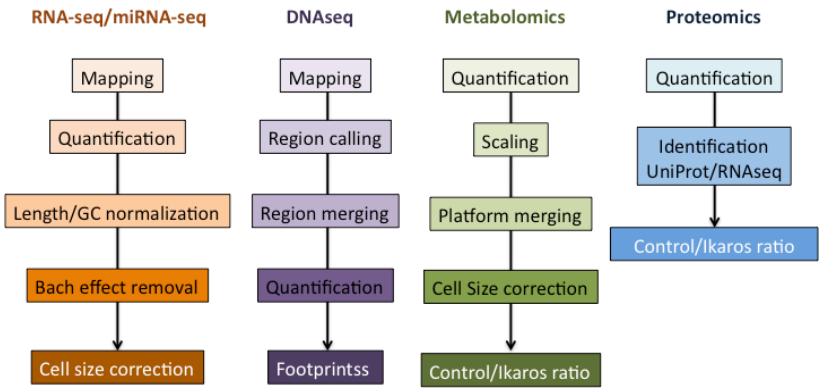


FIGURE 2 OVERVIEW STATEGRA PRE-PROCESSING PIPELINES

3. Development of the STATegra Knowledge Base

We have created a Knowledge Base (KB) to host consortium data and relevant to the SuS public. The goal of the KB is to become a repository of omics information useful for methods development that will also gather the knowledge created during the project. Next to the STATegra data, the KB integrates up to 12 different data sources, including gene expression and annotation, protein-protein interaction data, metabolomics and pathway information, gene regulatory information involving Transcription Factor and microRNAs regulation. At present the KB has defined routines for mapping genomic coordinates to gene and for linking metabolomics and protein information to gene elements. The platform implements different interfaces to retrieve data including cross database queries and different integrative visualization engines. These include Mouse gene cards displaying all joined database and project data at the gene level, and integrative local network graphs that represent graphically different layers of molecular regulation and interaction for particular genes or set of genes. For example, this visualization can represent the local molecular network around Ikaros showing all its regulating transcription factors, target genes, gene expression along different conditions, miRNA regulators, etc (Figure 3). The KB is available through the STATegra project web page, so far restricted to the members of the consortium. It will be made public at the end of the project.

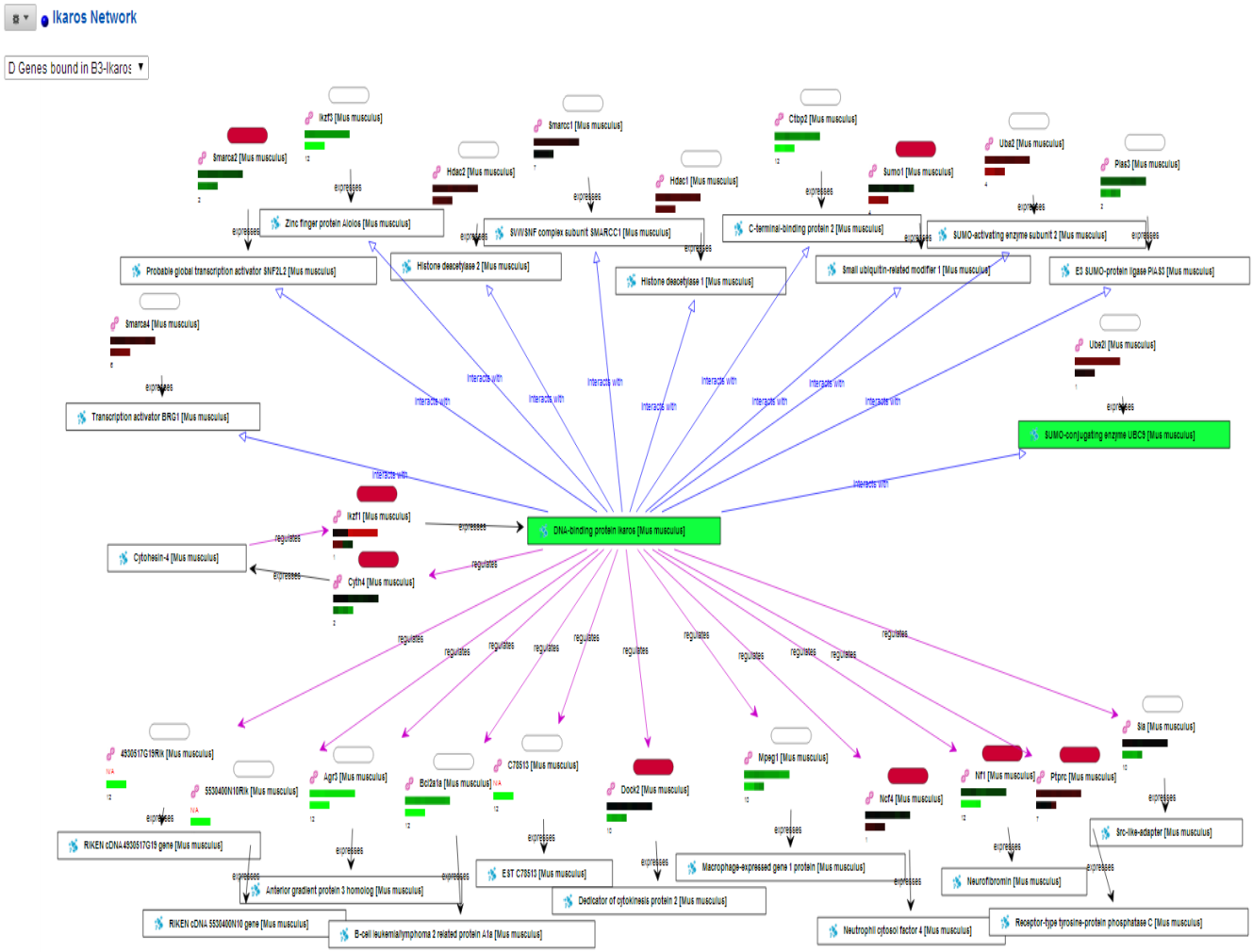


FIGURE 3 EXAMPLE IKAROS INTEGRATED NETWORK AT KB

4. Development of Analysis Methods

Methodological developments have proceeded within three different areas: methods for data integration within a defined experimental design, methods for integration of heterogeneous dataset, and analysis of noise and performance metrics in heterogeneous datasets and experimental design.

4.1. Methods for integration of datasets within the same experimental design

The development of methods for integrative omics data analysis was initiated with the utilization of public datasets and some of the pilot STATegra datasets that were obtained at the very beginning of the project. Briefly these methods and their major characteristics are:

EXPLORATIVE APPROACHES

(1) **Data Fusion Methods** is a strategy to analyze jointly the overall common and distinct variability of different omics data types measured over the same set of samples. The results of this analysis are multiple PCA models where clustering of samples for the common variability component and clustering for omics-specific variability can be visualized. The approximation is valuable to explore relationshios across multi-omics dataset.

(2) **OmicsClustering** is a clustering method based on the combined and weighted distances between genes calculated on the basis of several omics measurements. The algorithm requires a mapping startegy to assign non-gene features (such as ChiP-seq peaks) to genes. The approach is interesting to see gene associations due to different regulatory characteristics of genes.

Both DataFusion and OmicsClustering are implemented in the STATegRa R packaged that has been submitted to the Bioconductor repository (see also software section).

PATHWAY-BASED

(3) **Paintomics**. Paintomics is a data integration approach based on the joint visualization of different omics data-types on the template of KEGG pathways. The tool displays gene expression, protein, methylation or metabolic changes at protein or metabolite positions of the pathway map. The published version allowed integration of metabolomics and gene expression data, while the current version under development integrates any type of omics data that can be associated to genes. The application also incorporates a multi-omic functional enrichment test to identify significant pathways acounting for different types of omics data.

Pathway Network Analysis. It is a methodology to create a global network of pathways interactions from gene expression data. Currently we are extending the approach to include metabolomics and proteomics data. The method reveals functional relationships between pathways and key genes involved in these interconnections.

VARIABLE SELECTION

(4) Integration of omics time course datasets with maSigPro. We have applied this approach to the integration of RNA-seq and DNase-seq data. Basically we model each omics time course using the Next-maSigPro approach, developed at CIPF, and then create a classification matrix where the profile relationshiop of each gene with its associated DNase HyperSensitive (DHS) regions is categorized according to their patterns of change, for example gene expression goes up and the associated DHS regions remains flat. Genes at each expression-DHS pair category are further analyzed in terms of functional enrichment, regulatory motifs and network properties. The method is interesting to explore patterns of chromatin changes associated to gene expression changes. The approach could be generalized to model other tye of time-course omics data pair-wise relationships.

(5) Machine learning approaches. In this approach we first apply classification trees to identify different clusters of genes with similar "regulatory programmes", understood as the set of regulatory variables (methylation, microRNAs, DHS regions, etc) that predict the expression of their associated gene. Then we use structural equations to model the expression of each gene as a function of its candidate regulators**.**

(6) **Mens for Machina (MxM)**. This package developed by FORTH implements the Statistically Equivalent Signatures (SES) algorithm, i.e., a feature selection method that aims at discovering the minimally-sized set(s) of variables that are needed for optimally predict a given output.

NETWORK MODELLING

(7) Dynamic Network Modelling. The method models analyzes ChIP-seq and DNase-seq data to infere the binding of TFs to specific gene related hypersensitivity regions and create a time-resolved network of TF-target binding interactions. The approach incorporates gene expression data to monitor which genes are actually expressed and set them as seed of new chromatin interactions in sucessive time points.

Method	On data	Output	Class
Time course models	RNA-seq/other	Selection of genes with specific correlation patterns	Gene selection
Multi-random forest	All gene regulation omics	The regulators of the expression of a gene	Local networks
Dynamic networks	DNAse-seq/ RNA-seq	Regulatory program	Global networks
Data Fusion	All omics	Combined PCA analysis	Data set overview
Omics Clustering	All omics	Cluster of genes according to several omics data types	Gene Clusters
Paintomics	RNA-seq, proteomics , metabolomics	Integrated visualization over pathways	Enriched pathways
PANA	RNA-seq, proteomics , metabolomics	Network of pathways	Global networks

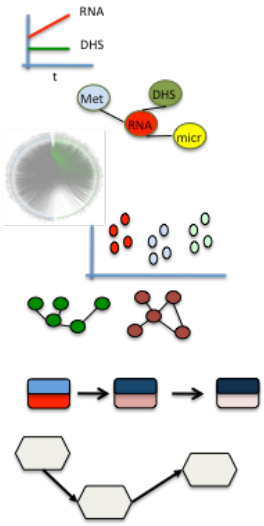


FIGURE 4 STATEGRA METHODS FOR INTEGRATIVE ANALYSIS

4.2 Methods for causal discovery using data from heterogeneous sources

This WP addresses the problem of integrating datasets of diverse origins, typically found in the public domain, that target the same experimental system but have not a shared experimental design. Three methodologies have been proposed:

(1) **COmbINE**. This algorithm infers causal structures from the integrative analysis of collections of data sets that measure overlapping sets of variables under different experimental conditions. The method assumes that there are common underlying causal relationships in the data measured in different experiments and can infer relationships between variables even if they have not been measured together

(2) **CNMA**. The **Causal Network Meta-Analysis** integrate datasets that share experimental conditions and variables to the aim of creating a larger dataset that can be efficiently used in causal network inferential analysis. The goal in such case is to account for possible bacth effect that reveal the source of each dataset and need to be removed to allow proper causal network analysis.

(3) Methods for **integrating prior knowledge and data** that learn causal networks in the context of prior causal knowledge, e.g., that X causally affects, directly or indirectly, Y. These algorithms can be used in combination with the previous ones, particularly with the Causal Network Meta-Analysis algorithms.

5. Methods for integrative experimental designs

In this WP we aim to provide insights of the different aspects that need to be taking into account when designing experiments involving multiple omics data types. Our first approach has been de study and definition of the **Figure of Merits** (FoM) that characterize each technology as a way to uniform languages and be able to assess performance metrics for each data type. FoM evaluated include Sensitivity, Reproducibility, Selectivity, Detection Limit, Dynamic Range, Coverage and Identification. Some of these metrics have been used for comparative performance analysis across omics data types of the STATegra and public datasets to perform a systematic analysis of **noise in omics technologies**. We show that noise and performance levels can significantly vary across data types and that this has an effect on the power size calculations.

6. Software development

STATegra aims at generating user-friendly software for the scientific community.
Apart from the KB, the following tools have been developed or are under development:

6.1 STATegra Experiment Management System

The STATegraEMS is a java application for the annotation, storage and tracking of omics experiments. The idea behind this development is to provide a tool for experimentalists who set up multiple sequencing and omics projects or obtain datasets from different omics providers, to organize and document their experiments. The system is built around three modules: the Experiment module that annotates data on experiments such as the experimental design and omics platform involved. The second part is the Sample module, where all information about biological conditions, replicates and analytical extractions are documented, and third part is the Analysis module that annotates information on raw data production, intermediate steps of data processing and final "clean" data (Figure 5). The application can be downloaded from the STATegra web site and installed locally at the user's site.

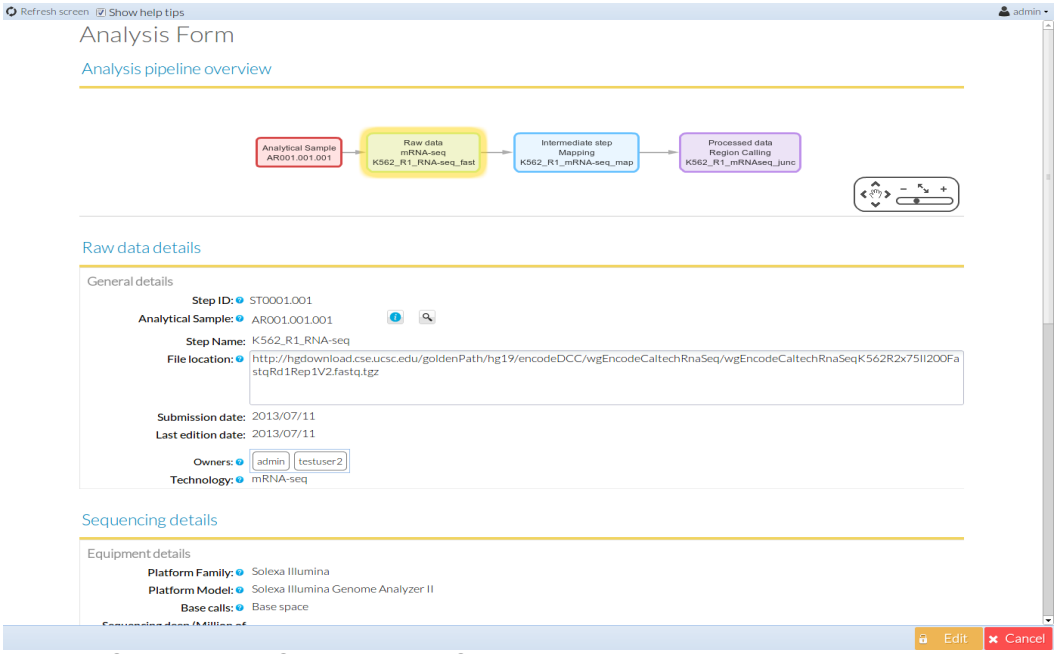


FIGURE 5 SCREEN-SHOT STATEGRA EMS DATA ANALYSIS FORM

6.2 The STATegRa R package

The STATegra project is committed to provide all new statistical methodologies as R packages to be disseminated through the Bioconductor repository among the scientific community. We decided to create one unified R package –STATegRa- to include all developments to enhance the impact and visibility of the project. A first release of package has been created and submitted to Bioconductor. The package defines methods and classes to call, process and plot multiple omics types and contains two basic functions for integrated visualization of omics dataset: OmicsClustering and OmicsPCA (DataFusion). Future developments of the consortium will be integrated in this software platform.

6.3. The Systems Biology CLC bio platform

The Systems Biology CLC bio platform is conceived as an update of the current CLCbio workbench, which accepts, integrates and processes multiple omics data-types. An alpha version has been released for evaluation by partners that include the central additions to the existing CLCbio Genomic Workbench to make it amenable to integration with STATegra developments. These additions are:

* R plug-in, for integration of the STATegRa package
* Biomax plug-in, for integration with the KB and retrieval of a priori information from the KB into the CLCbio platform
* Development of a generic Peak-caller to be able to deal with peak-based STATegra data, such as ChIP-seq and DNase-seq data.

7. Dissemination activities

STATegra is holds an active dissemination of its scientific activities. The project can be followed at social media and thought the workshops organized by the Consortium:
facebook, twitter

- Website: http://stategra.edu
- Facebook: STATegra
- Twitter: @stategra
- Worshops:
    - "High-Throughput Omics and Data Integration" in collaboration with SeqAhead, Barcelona March 2013
    - "The next NGS challenge Conference: data processing and integration" in collaboration with ISCB, SeqAhead and EMBnet, Valencia, May 2013
    - "Massive Data Analysis Course", Valencia, March 2014
    - "Workshop in Experimental pipelines and post-analysis in NGS and omics data", Amsterdam,March 2014.
    - "Statistical Methods for Omics Data Integration and Analysis", Crete November 2014) with special invitation to EU Omics Consortia.
- Collaborations with EU projects involving omics data analysis: CostAction SeaAhead, ALLBIO, Profilic and Cost Action EpiConcept
- We have established an agreement with BMC to publish STATegra series in their Open Access journals. A first special issue has been published in BMC Systems Biology (Volumen 8, Supplement 2 : "Selected articles from the High-Throughput Omics and Data Integration Workshop").

STATegra publications:
All STATegra publications except for one are Open Access papers:

11. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics. 2014 Sep 15;30(18):2598-602.
10. Sansoni V, Casas-Delucchi CS, Rajan M, Schmidt A, Bönisch C, Thomae AW, Staege MS, Hake SB, Cardoso MC, Imhof A (2014) The histone variant H2A.Bbd is enriched at sites of DNA synthesis. Nucleic Acids Res.
9. Gomez-cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., … Tegnér, J. (2014). Data integration in the era of omics : current and future challenges. BMC Systems Biology, 8(Suppl 2), I1. doi:10.1186/1752-0509-8-S2-I1 (Highly Accessed Paper)
8. Conesa, A., & Mortazavi, A. (2014). The common ground of genomics and systems biology. BMC Systems Biology, 8(Suppl 2), S1. doi:10.1186/1752-0509-8-S2-S1
7. Diego, R. H. De, Boix-chova, N., Gómez-cabrero, D., Tegner, J., Abugessaisa, I., and Conesa, A. (2014). STATegra EMS : an Experiment Management System for complex next-generation omics experiments. BMC Systems Biology, 8(Suppl 2), S9. doi:10.1186/1752-0509-8-S2-S9
6. Ponzoni, I., Nueda, M., Tarazona, S., Götz, S., Montaner, D., Dussaut, J., … Conesa, A. (2014). Pathway network inference from gene expression data. BMC Systems Biology, 8(Suppl 2), S7. doi:10.1186/1752-0509-8-S2-S7
5. Reshetova, P., Smilde, A. K., van Kampen, A. H., & Westerhuis, J. a. (2014). Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. BMC Systems Biology, 8(Suppl 2), S2. doi:10.1186/1752-0509-8-S2-S2
4 Schmidt, A., Forne, I., & Imhof, A. (2014). Bioinformatic analysis of proteomics data. BMC Systems Biology, 8(Suppl 2), S3. doi:10.1186/1752-0509-8-S2-S3
3. Mortazavi A, Pepke S, Jansen C, Marinov GK, Ernst J, Kellis M, Hardison RC, Myers RM, Wold BJ (2013) Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. Genome Research.
2. de la Rica L, Urquiza JM, Gómez-Cabrero D, Islam AB, López-Bigas N, Tegnér J,Toes RE, Ballestar E. (2013) Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. Journal of Autoimmunity 2013 Gen 7. (Not OpenAccess)
1. Giorgos Borboudakis, Ioannis Tsamardinos (2013) Scoring and Searching over Bayesian Networks with Informative, Casual and Associative Priors", Uncertainty in Artificial Intelligence (UAI).

8. Future plans

Our immediate plans in the project are:

→Publication STATegra data. We plan to release a first publication of the STATegra data with basic description of the dataset and an analysis of first insights in the B3 Ikaros system from the integrative analysis. The data will be available from the STATegra website, submitted to public repositories and we will create a R data package.

→Next to the STATegra data collection, we are working already working on different types of manuscripts:

* Software: Publication of the STATegRa R package and the multi-omics version of the Paintomics tool. We also expect to publish the Knowledge Base as such.

* Experimental design: there will be two papers on the analysis of Figures of Merit and considerations for experimental designs of multi-omics experiments

* Statistical methods: We are expecting between 2 and 4 papers per statistical group describing novel methodologies.

* Collaboration papers.

→ Events. Planned dissemination actions are:

Workshop on Statistical Methods for Omics Data Integration (SMODIA, Crete, November 2014) and special issue BMC Bioinformatics

STATegra summer school (2015)

STATegra news letter

Closing sysmposium, September 2015.

→ Additionally, the second part of the project will address additional goals, such the building of an integrated network for the B3 experimental system, designing validation experiments, the development of methods for feeding back results in the public domain, and the release of the Systems Biology Platform.