Grant Agreement Number 201550

HYPERGENES
European Network for Genetic-Epidemiological Studies

European Network for Genetic-Epidemiological Studies: building a method to dissect complex genetic traits, using essential hypertension as a disease model

Final Report

Security (distribution level)	PU
Contractual date of Delivery	Month 48
Actual date of delivery	Month 48
Document name	Final Report
Type	Report
Status and Version	Final
Number of pages	27
WP contributing to the deliverable	WP13
WP/Task responsible	IMS
Other contributors	UNIMI, specific contributions from all Partners
Author(s)	Costanza Conti
EC Project Officer	Iiro Eerola
Keywords	Final Report
Abstract (for dissemination)	Final Report for project dissemination
Document ID	HYPERGENES-WP13-IMS-DEL-D13.6-V7

Table of Contents

1 Introduction

1.1 Scope of the document

This document is the periodic progress report (PPR) related to year 2011. This report together with internal periodic progress report related to year 2010, represent the Periodic Progress Report for 3rd Reporting Period.

1.2 Applicable and reference documents

This document refers to the following documents:

Annex I - "Description of Work"

D.13.2 – Periodic Progress Report

D.13.3 – Periodic Progress Report

D.13.4 – Periodic Progress Report

Internal PPR 2010

1.3 Revision History

Version	Date	Author	Description
0	10/02/2012	C. Conti (IMS)	Report Draft
1	29/02/2012	C. Conti (IMS)	Document implementation
2	05/03/2012	M. D'Alessio (IMS)	Document revision and integration
3	05/03/2012	M. D'Alessio (IMS)	Formal revision
4	05/03/2012	E. Salvi (UNIMI)	Document revision and integration
5	05/03/2012	D. Cusi (UNIMI)	Overall revision and integration
6	06/03/2012	M. D'Alessio	Formal revision
7	06/03/2012	C. Conti (IMS) and C. Barlassina (UNIMI)	Final revision

2 Executive summary

HYPERGENES - European Network for Genetic-Epidemiological Studies: building a method to dissect complex genetic traits, using essential hypertension as a disease model is a research collaborative project financed by the EC trough the 7th Framework programme.

Its activities are structured in three steps (Discovery, Validation and Dissemination & Results Exploitation) to be realized in four years.

HYPERGENES Discovery Phase was performed during the first two years of the project and was focused on building the methodological and technical framework to support the Genome Wide Association Analysis, performed on 4,000 Caucasian subjects recruited from historical well-characterized European cohorts.

The need of integrating the observations from different studies posed significant challenges which were faced through an integrated epidemiological and bioinformatics approach. The Biomedical Information Infrastructure (BII) was developed to support the entry, persistency and retrieval of data and knowledge relevant to EH, including clinical, environmental and genotypic data.

Genotyping have been performed on high throughput Illumina technologies, thanks to the coordinated efforts of the Laboratories of UNIMI and UNIL.

Both classical and machine learning techniques were used for genetic analysis, to produce an enriched list of SNPs, that resulted associated with EH or TOD, or however other endophenotypes relevant to hypertension. The conducted case-control association study lead to hundreds of significant associations, which were only partially overlapping with the results of previous studies.

The best SNPs resulting associated to EH and TODs in the Discovery Sample together with candidate SNPs well known as being associated to the phenotypic trait of interest were used to build a custom Illumina iSelect HD chip, including 15,000 SNPs. Such tool was used to validate the results obtained in the Discovery Phase in an additional independent sample of 8,000 subjects.

Regions showing most promising results were sequenced in order to obtain a more detailed comprehension of the nucleotide sequence of each region. Sequencing was performed on 92 subjects.

A risk prediction algorithm for EH and TOD was then developed.

A lab-on-chip (LoC) meant to test the most promising SNPs found associated with microalbuminuria was developed. The LoC was therefore cross-validated on 100 samples.

The developed BII manifests the commonalities among the dozens of HYPERGENES cohorts for multi-cohort analysis, while preserving the disparities and allowing researchers to access the original cohort data. Atop of the BII two Machine Learning tools were developed: the “Bioclinical Data Mining Tool”, which is a general purpose algorithms built into a generic

framework for streaming data, and the “SNP weighting tool”, an algorithm specifically developed for utilizing the existing knowledgebase for SNP association analysis. These tools are available online for HYPERGENES partners.

These tools concurred to the development of a risk prediction disease model. Given a set of known EH and TODs risk factors for which progressive data are available, the model allows to predict the future clinical range of each parameter, exploring also the contribution of the genomic data to the prediction. To further integrate the model with genomic information, a previously developed gene network was used. Each gene in every pathway was represented by its strongest associated SNP. A good prediction performance was demonstrated for measures having sufficient samples (over thousand).

Moreover HYPERGENES performed a pathway analysis using the 29 variants identified in ICPB-GWAS, a recent genome-wide association study involving 200,000 individuals, and HYPERGENES results with the aim to verify a common pathway between the two studies. HYPERGENES added a further relevant player, NOS3, within the vasodilator pathway for blood pressure regulation.

3 Project context and main objectives

HYPERGENES project is focused on the definition of a comprehensive genetic epidemiological model of complex traits like **Essential Hypertension (EH)** and intermediate phenotypes of hypertension dependent/associated to **Target Organ Damage (TOD)**.

The discovery of the genetic component in common complex diseases is extremely challenging since most of them are multifactorial and since the genetic component is likely to be described by the interactions of several genes involved in the disease pathway, each predisposing imperceptibly to the disease. HYPERGENES adopts the **Genome Wide Association (GWA)** approach to identify common variants contributing to the inherited component of common diseases. The results of the GWA are the source to build a customized and inexpensive genetic diagnostic chip that can be validated in the project existing cohorts. HYPERGENES project is structured in three steps:

- STEP 1: Discovery
- STEP 2: Validation
- STEP 3: Dissemination & Results Exploitation

Designing a comprehensive genetic epidemiological model of complex traits foster the possibilities of translating genetic findings into improved diagnostic accuracy and new strategies for early detection, prevention and eventually personalized treatment of a complex trait.

The project's Technical and Scientific objectives are the following:

1. To identify the common genetic variants relevant for EH and TOD
2. To design and implement appropriate computational tools.
3. To develop a comprehensive Biomedical Information Infrastructure (BII).
4. To create a “Web-Based Portal” to allow access to the BII in order to allow dissemination of knowledge.
5. To develop new methods, protocols and standards for genomic association analysis, gene annotation and molecular pathways.
6. To develop a set of Decision Support Systems tools combining genetic, clinical and environmental information.
7. To develop a simple, inexpensive genetic diagnostic chip, that can be validated in our existing well-characterized cohorts.
8. To strengthen the existing clinician-basic scientist collaborative network on the genetic mechanisms of EH.
9. To generate educational tools to support professional training on all aspects of the project, favoring mobility of PhD students and post-docs.
10. To disseminate HYPERGENES achievements.
11. To exploit the results in a translational scenario.

4 Description of the main S & T results/foregrounds

The main S&T results and foreground generated from the project are listed in the following paragraphs in accordance with the project objectives listed in the previous chapter.

4.1 Identify common genetic variants relevant for EH and TOD

The HYPERGENES Project pursued a two-stage study to investigate novel genetic determinants of essential hypertension or Target Organ damage related to hypertension. Cases and controls were recruited from extensively characterized cohorts over many years in different European regions using standardized clinical ascertainment.

The discovery phase consisted of 1,865 cases and 1,750 controls genotyped with the 1 Million SNPs Illumina array (Genome Wide Association Study, GWAS). Best hits were followed up in a validation panel of 1,385 additional cases and 1,595 controls that were genotyped with a custom array of 14,055 SNPs. The SNP selection for custom chip was based on: a) The list of BEST SNPs from Discovery case-control analysis at a genome wide level (p-value $<1 \cdot 10^{-4}$); b) A list of candidate genes and SNPs historically studied in hypertension or genes selected according to their functional role and involvement in biological pathways relevant in hypertension.

The following paragraphs provide further details on the study phases.

Discovery phase

Concerning sample characteristics, HYPERGENES Consortium decided to maintain a very neat separation between cases and controls, selecting cases among well defined hypertensive patients and controls among normotensives with little chances to develop hypertension later in life. Particular care was devoted to the of control subjects. A large proportion of the sample has been followed for 5–10 years after DNA collection, allowing for the exclusion of controls that developed high blood pressure at a later age, thereby defining the hyper-normal controls.

The genotyping has been performed using high throughput Illumina technologies, thanks to the coordinated efforts of the Laboratories of Universities of Milan and Lausanne. The genetic analysis performed during the project followed different methodologies, including classical and Machine Learning techniques, to produce an enriched list of SNPs, that resulted associated with EH or TOD. More in details, after reaching an agreement among domain experts, HYPERGENES focused on the following TODs or endophenotypes: left Ventricular Mass Index, Interventricular Septum in diastole, Estimated Glomerular Filtration Rate, Body Mass Index, Low Density Lipoproteins and Microalbuminuria. Other

endophenotypes relevant to hypertension, such as pharmagenomic profiles for two widely used antihypertensive drugs (losartan and hydrochlorothiazide), were considered as well.

From Genetic analysis, conducted through Classical and Machine Learning approach, emerged that ethnicity is the main component of heterogeneity within the sample used.

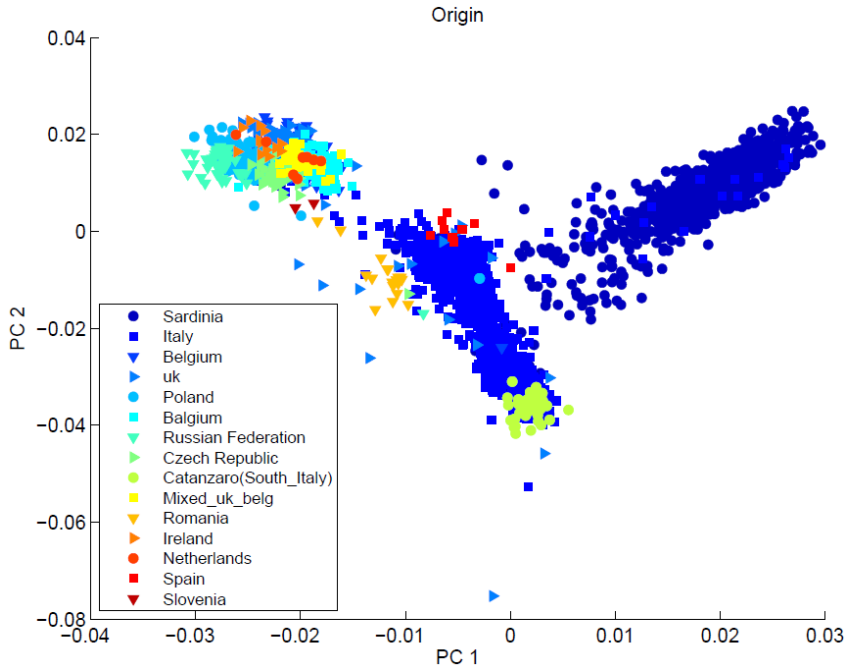

Figure 1: Clusterisation of the HYPERGENES Discovery Sample according to Principal Components Analysis (PCA). The graph shows the clusterisation of the different ethnic groups.

As shown in the graph above, HYPERGENES sample clusterises in three major groups. On the upper left the cluster refers to the North Europeans, on the center to the South Europeans and on the upper right to the Sardinians.

Interestingly, this clusterisation reflects the geographical distribution of subjects.

In the discovery phase, 90 SNPs (57% intragenic) with p-value < $1 \cdot 10^{-4}$ were identified after genomic control correction, only partially overlapping with the results of previous studies. Moreover, the signals of SNPs previously presented in literature are in our study in the same direction as the original studies showing evidence of a marginally significant association in HYPERGENES

Quantitative phenotypes were considered for TOD and other intermediate-phenotypes.

Preliminary results include hundreds of TOP SNPs, with p-value ≤ 10^{-5}, for Left Ventricular Mass, Inter Ventricular Septum, Intima Media Thickness, Microalbuminuria, eGFR, HOMA index, LDL Cholesterol, Body Mass Index.

Some of the genes identified were already described in previous studies.

Based on the Discovery Phase findings, an Illumina custom chip (including 14055 SNPs) was designed to re-test specific genetic variations that have been associated with Hypertension or Target Organ Damage under the HYPERGENES project or in previous studies.

Validation phase

Essential Hypertension: In the validation stage, ***rs3918226*****,** which maps to the promoter region of the *eNOS* gene (-665 C>T), was confirmed to be associated with hypertension in Caucasians, reaching a combined (discovery and validation) P-value of $2.58x10^{-13}$. Minor Allele Frequency (MAF) of T allele was 13.75% in cases and 8.95% in controls. The Consortium investigated in a prospective cohort study the impact of the identified rs3918226 polymorphism on incident hypertension and on longitudinal blood pressure changes. In whole cohort (n=2445) followed up for 7.58 years, the increases in systolic and diastolic blood pressure from baseline to follow-up were 11.12 mm Hg increase in systolic blood pressure in both additive (p=0.040) and recessive (p=0.023) models. For diastolic blood pressure the increase associated with TT homozygosity was 7.76 mm Hg in both additive (p=0.015) and recessive (p=0.005) models.

Target Organ Damage

After evaluation of the population stratification in each TOD sample, we performed a quantitative trait association analysis using a linear model by correcting for stratification. We obtained the adjusted chi-squared statistics and relative p value for the SNPs that showed a significant association to each TOD. A threshold of p≤1E-4 was used to select the most significant associations and to create a preliminary list of ranked SNPs.

Our SNPs strongly associated to TOD, did not show association with EH (P>0.05), thus they may act through different pathways and not via increasing blood pressure. The most promising replication outcomes were observed for LDL Cholesterol and Interventricular Septum (IVS) phenotypes. The LDL hit rs10901555 showed promising discovery signal (beta=-6.92, P=1.10E-05) and replicated successfully in the validation cohort, yielding a combined P-value of 2.91E-06 (beta=-5.84, SE=1.25). This SNP has not yet been reported in previous GWAS studies, hence is worth further replication in independent cohorts to reach genome-wide significance. For the IVS phenotype our top hit rs2133471 reached a P-value of 5E-07 in discovery, however was not successfully genotyped in the validation cohort. This SNP needs further evidence before we can claim significance.

Additionally, we assessed the quality of our data by checking how well already published GWAS hits replicate in our cohort. As can be seen from the massive inflation of the QQ-plots, our data is robustly replicating a substantial fraction of the published hits.

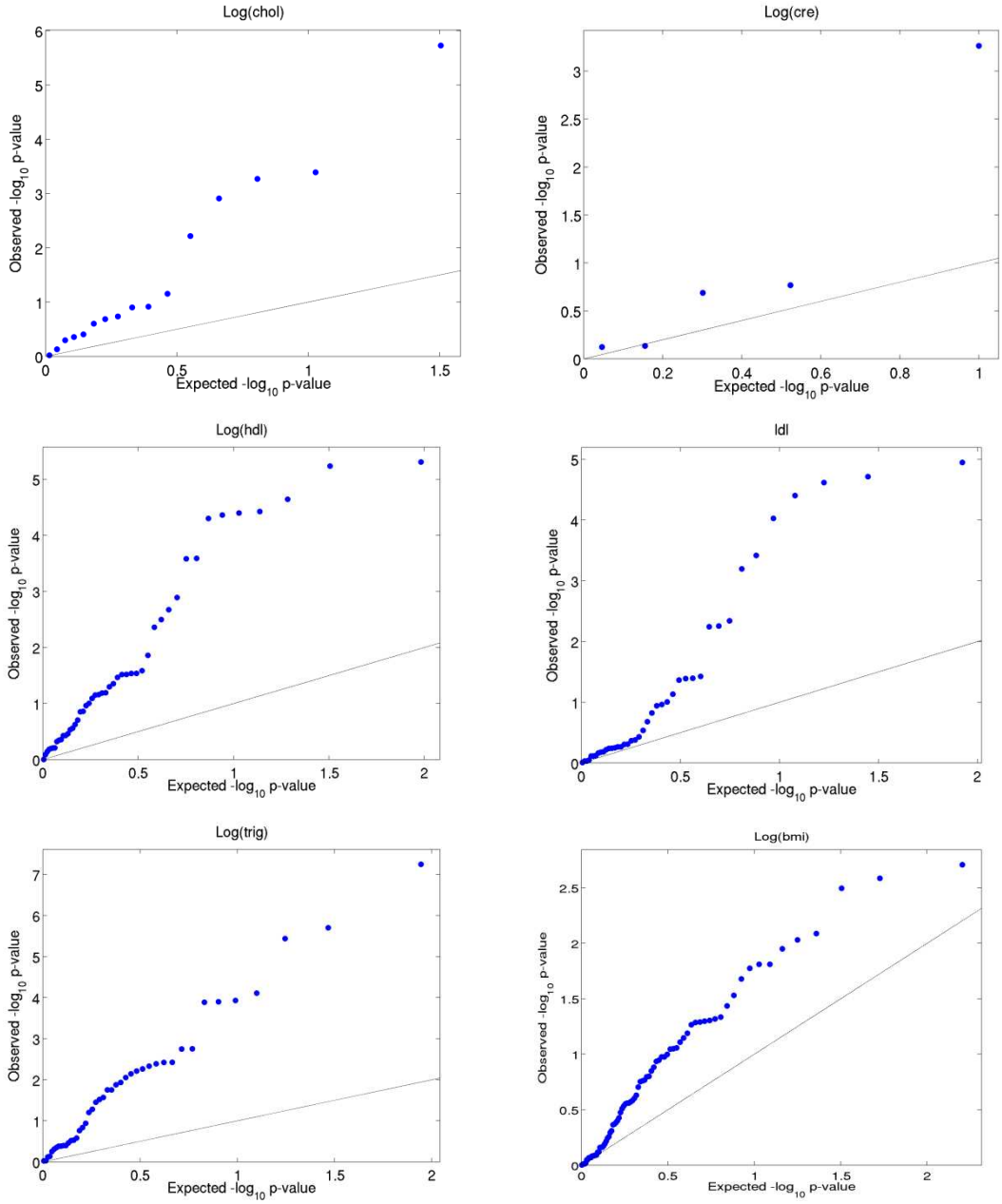

Figure 2: QQ-plots of the different TODs including published hits only. Quantile-quantile plots show the ordered observed and expected P-values with the ones one would obtain if no association were present (null hypothesis). Under the null hypothesis dots are expected to fall on the diagonal. (Positive) deviations from the diagonal indicate evidence for association.

TODs and longitudinal Analysis

The randomly selected EPOGH-FLEMENGO prospective population cohorts were followed-up for many years allowing the recruitment of some TODs data in different time period (Body Mass Index (BMI), Glomerular Filtration Rate (GFR) and Low Density Lipoproteins (LDL)). We investigated also the effect of the 15K SNPs (custom chip) on the longitudinally changes of BMI, LDL and eGFR starting from the 15K custom chip list by applying single-SNP linear regression models. With this analysis, we identified different important candidates genes (AGT, PRKG1 SLC12A1, SLC8A1, WNK1, NEDD4L, ENPP1 and CRP) that need to be further investigated in other studies or meta-analyses for their possible relation with hypertension and endo-phenotypes.

Analyzing interactions in the associations of Essential Hypertension (EH) and the following TODs: BMI, GFR, Serum Creatinine, Serum Insulin and HOMAi.

These analyses describe the significant association for 2 SNPs with serum creatinine, which is worthy of replication, since it exceeded the genome-wide significance threshold of 5x10-8.

Regions showing most promising results were sequenced (**target re-sequencing**) in order to obtain a more detailed comprehension of the nucleotide sequence of each region. Sequencing was performed on 92 subjects, hypertensives and normotensives selected for the presence/absence of the “risk” alleles for hypertension at the SNPs significantly found associated to hypertension in the discovery phase. Using the Agilent SureSelect enrichment Kit we captured for each DNA a region of approximately 1.3 Mb and prepared libraries for sequencing. The sequenced regions included:

- eNOS: endothelial Nitric Oxide Synthase gene
- PLG: plasminogen gene, the inactive precursor of plasmin
- SLC9A1: gene encoding for a Na-H antiporter involved in pH regulation and in the clearance of acids generated by active metabolism.
- SLC24A4: gene encoding for a potassium dependent sodium-calcium exchanger
- KCNMA1: gene encoding for a potassium channel activated by calcium
- ARHGEF15: gene encoding for a RhO GTPase with a fundamental role in processes triggered by G-protein-coupled receptors.

For the SNPs shared between 1M chip and the sequencing platform, the concordance genotyping rate was 98%. For these SNPs the UNIMI group also calculated the allele frequency between cases and controls and compared it to the one of 1M chip. The sequencing data suggested rs3918226 as the only candidate for hypertension in this region and the data was also supported by imputation using 1000Genome database (release June 2011).

Generation of risk predictive algorithm

Promising results were observed for some phenotypes, even if the results need to be interpreted with caution.

For most of the traits, when considering the published plus newly identified hits in total approach, ~4% explained phenotypic variance, which is far from any practical application for prevention or prediction, except for a very small subgroup of individuals carrying large number of risk alleles.

4.2 Design and implement computational tools

Atop of the Biomedical Information Infrastructure (BII), implemented for the project purposes, two Machine Learning tools were developed, namely *SNP weighting tool*,, an algorithm specifically developed for utilizing the existing knowledgebase for SNP association analysis, and *Bioclinical Data Mining Tool*, which is a general purpose machine learning algorithms built into a generic framework for streaming data.

We have constructed a web **SNP weighting tool**, for computing weights for SNPs, that is freely available online. The user can:

- choose from a collection of commercial SNP chips
- select annotations from a pre-specified list with the default being the annotations presented in this paper
- can add additional SNPs possibly marked as trait-associated SNPs (TAS) or non-TAS, beyond the SNPs specified on the chip
- control the list of diseases that are used for computing the requested SNP weights.

The freedom to control the list of diseases is useful when the user has knowledge about a subset of diseases in the webpage list that better resembles the new disease under study.

The following figure depicts the processes implemented for SNP weights:

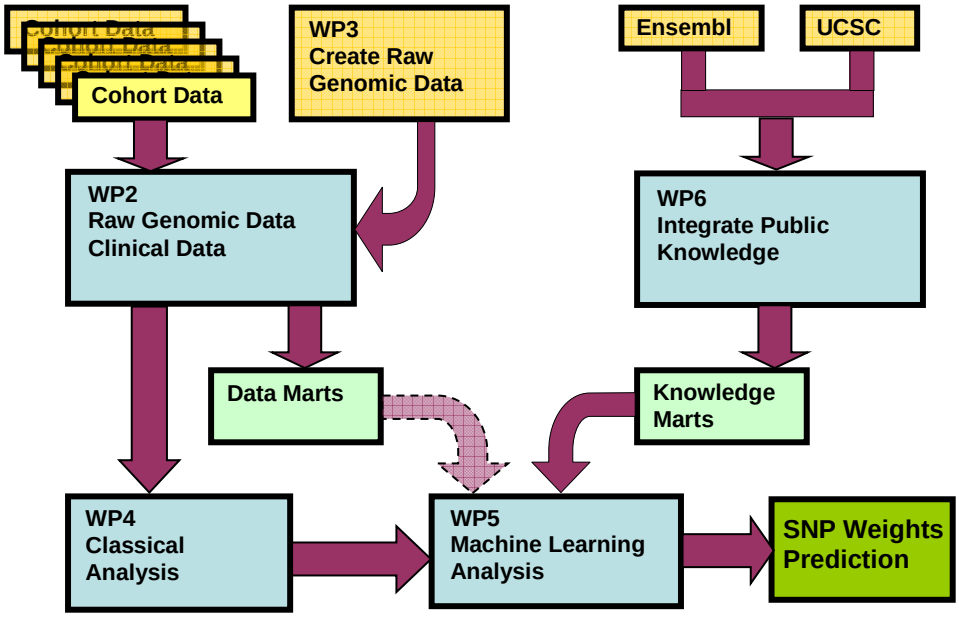

Figure 3:The figure is a commented screen shot taken from the BII Portal and illustrates the use of the "Weigh your SNPs" feature. This screen shot shows the initial selection of SNPs and annotations.

The **Bioclinical Data Mining Tool (BDM)** allows users to execute machine learning and data mining algorithms on large data sets. Through a simple XML file, the user may combine various algorithmic building blocks in a workflow that performs the desired task; in addition, new blocks can be written using a simple Java interface.

The BDM can operate with many built-in data types, starting from generic types such as integers, doubles and booleans, through more complex types as maps and lists, to the SNP data type, specifically relevant for association studies.

In the context of this task, a few algorithms have been implemented as BDM blocks. First is the Chi-square test, useful for feature selection in general, and for detection of SNPs associated with a disease in case-control data. Second is the logistic regression algorithm, which is a very common approach for binary classification problems. Third, is the k-means clustering algorithm. This is a popular clustering approach that can later be integrated with various metrics defined for the data. In the specific GWAS context, the Hardy-Weinberg test was also implemented.

Moreover blocks for reading data from various sources e.g. plink files (in tped format), text files, and a block for loading data from the BII and writing back to it are also implemented. As the algorithm development process usually starts with experimenting with simulated data, random data generators, for generating sequences of numbers, random vectors, and random Gaussian mixtures are also included.

The IBM BDM is freely available at http://www.alphaworks.ibm.com/tech/bdm.

Another tool implemented during HYPERGENESs SNP Ranker: a tool for SNP mining which can take into account experimental data and knowledge information about genomic regions. It is aimed at mining SNPs knowledge and genes' annotation to score SNPs and rank markers with respect to genes and diseases. With this tool it is possible to set up specific knowledge features for each SNP (for example SNP localization into the gene, MAF, etc) and balance this feature importance within the genomic region of interest. The evaluation score for each SNP is derived from a mining approach that leverages on data-integration. The tool also uses expansion to achieve final score of SNP features. The software is available at www.itb.cnr.it/snpranker.

Furthermore a novel dynamic and user-friendly management system that fully supports data integration and visualization, data access and analysis was developed to enhance the Genetic Labs data storing capacities, the **ADBioDB**:
Ad2BioDB is an open source project based on Adempiere ERP. Adempiere ERP was chosen as starting base for the platform since it natively supports data warehousing and business logics for enterprises. Above Adempiere ERP, a plugin function for Bioinformatics and genomic data management has been integrated. This feature allows having a unique software suite both for administrative needs and scientific purposes.
Ad2BioDB is designed in three main components (3-tierd): (1) an application server, that is devoted to process all user-system interactions, (2) a database server for all data queries that triggers data user actions, and (3) a processing server, that manages data manipulation, import/export, and jobs' management.
The prototype version of Ad2BioDB is currently plugged into a high performance server that processes all user requests through a cluster queuing system: this approach grants reliability, scalability and flexibility and balances computational loads to all available resources.

4.3 Develop a Biomedical Information Infrastructure

HYPERGENES approach foresaw the creation of a Biomedical Information Infrastructure (BII), providing the project itself infrastructure. Such data warehouse had to store existing and newly created harmonized and standardized information (data and knowledge), at the same time providing efficient access to it.
The BII enabled the collection, integration, harmonization and correlation of data described in diverse formats and vocabularies, scattered in disparate geographies. Furthermore, the BII design targets both research and clinical environments by having a single standard-based warehouse as a source of multiple marts that serve specific needs in research and healthcare. These marts contain data and knowledge that stem from the warehouse and

have back-references to the original data in the warehouse. Figure 4 shows the landscape of the BII, with the warehouse at the heart of it and the multiple marts on the top of warehouse, utilized by research and clinical environments. Input is fed to the warehouse through standard interfaces, mainly HL7 for the data and RDF for knowledge.

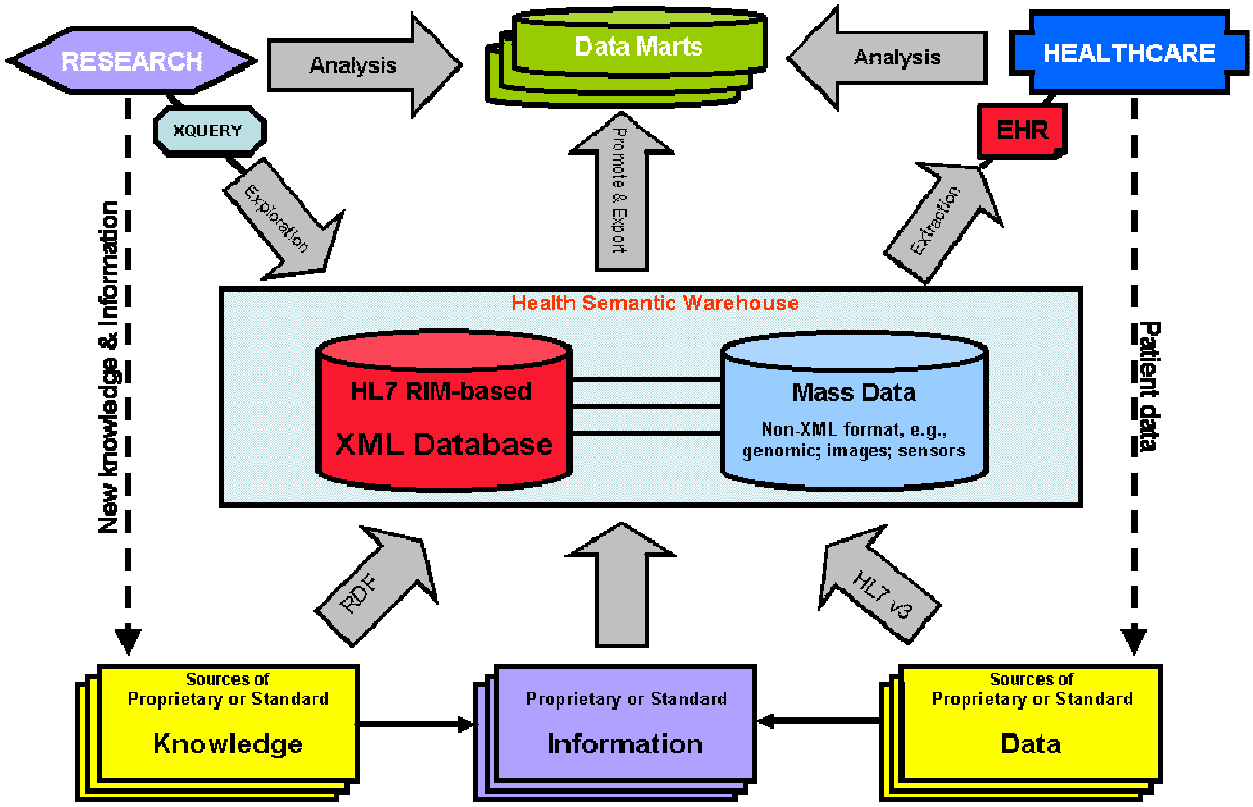

Figure 4: The Biomedical Information Infrastructure Landscape – a unified infrastructure serving both research and clinical environments.

BII allows integration, harmonization and standardization of clinical, environmental and genomic data. The warehouse service stores data in its richest format using a set of constrained internationally-recognized standards such as HL7 Clinical Document Architecture (CDA), the Pedigree and Genetic Variation standards. It manifests the commonalities among the dozens of HYPERGENES cohorts for multi-cohort analysis, while preserving the disparities and allowing researchers to access the original cohort data, yet in a standardized way using the CDA flexibility for representing the heterogeneous phenotypic data. This flexibility has been enabled as a CDA implementation guide and in the last year this guide has been ported into an open source tool (MDHT) which uses UML + OCL to represent the base model and its constraints.

The BII has been largely disseminated and is available online for HYPERGENES partners through its portal and offline through a number of data marts.. The data marts have been

coupled with knowledge generated in HYPERGENES as well as fetched from publicly-available knowledge sources. The main challenges faced includedrepresenting knowledge from various sources as well as coupling it with data marts for further analysis. In particular, various knowledge representations have been experimented with, in an attempt to serve the goals of developing alternative disease models.

The essence of the BII lays in its methodology, which can be replicated in future projects similar to HYPERGENES, and it is open to other potential uses for example for biobanks warehousing, integration of electronic health records and pharmaco-surveillance.

4.4 Create a web-based portal to allow to allow access to the BII to disseminate knowledge

As already indicated, BII data wharehouse is available online for HYPERGENES partners. During HYPERGENS, the final users have been testing the RDF-based marts, in order to extract data for monitoring and analytics purposes. The process involved the development of relational datamarts and tools for querying the data wharehouse.
Atop of the BII are available web accessible Machine Learning tools described in paragraph 4.2.

Furthermore the Consortium developed a special portal within the BII to allow experimental use of the knowledge gained in the course of the HYPERGENES project.
The portal allows viewing of selected clinical and demographic parameters for each subject from any cohort.
Specifically, for demographic parameters, we demonstrated subject's credentials (id, nlab_id, cohort name), study phase, affection status, gender, birth date, smoking status and first date of antihypertensive drug reception (see figure 7 below).
Selected clinical parameters (BMI, Blood Pressure and Cholesterol) are presented in the graphical form. Moreover, if available then Blood Pressure and Cholesterol predictions, generated from disease model constructed by IBM (Machine Learning group) are also displayed.
Additionally, a link is provided to a page presenting historical and clinical guideline treatment recommendations. The user can overview Clinical Guidelines treatment recommendations for the similar patients as well as their actual treatments and the treatment outcome.

4.5 Develop new methods, protocols and standards for genomic association analysis, gene annotation and molecular pathways

After consultation with domain experts, HYPERGENES revisited the Guyton model in light of the available data and analysis results deposited into the BII. The working group concluded that HYPERGENES needed to develop a more focused and simpler disease model where to introduce their GWAS analysis results. In particular, it was decided that only those portions of the Guyton model relevant to essential hypertension known to be associated with genetic variants could be considered for representation in the BII and for inclusion in the final disease model.

Consequently, several efforts have been concentrated on patophysiological mechanisms in which the genetic variants showing most promising results in our association analysis seem to be involved.

The clinical genomic model developed for hypertension aims to predict the future clinical range of known risk factors for hypertension. The objective was to integrate both clinical and genomic information for this purpose.

One representation of the genomic information in the context of this disease model utilizes the network-based disease model for essential hypertension previously developed. Under this representation, each gene in every pathway is represented by its strongest associated SNP. Following, every individual is represented by a vector corresponding to the set of pathways in the model. The feature value of each pathway is the number of risk alleles along the pathway carried by the individual. Thus, overall the genomic information in the network is represented by the 23 pathways over 149 genes that were found to be significantly associated with essential hypertension.

In addition, a more clinical-oriented effort took place to model the key concepts and respective data and knowledge that are relevant to essential hypertension.

Furthermore, using GeneMANIA as prediction software (http://genemania.org) and a gene list that combined top results of the two studies as input file, the project developed a GO pathway annotation. The software extends the input list with functionally similar genes that are identified using available genomics and proteomics data. The Consortium identified several pathways related to blood pressure homeostasis that include ICPB-GWAS (a recent genome-wide association study involving 200,000 individuals) and HYPERGENES genes.

These genes are involved in vasodilator and natriuretic mechanisms, electrolyte homeostasis and hypertrophic processes. The combined top findings of HYPERGENES and ICPB-GWAS highlighted genes and functional pathways concerning vasodilator and natriuretic mechanisms, electrolyte homeostasis and hypertrophic processes. In particular ICPB-GWAS identified GUCY1A3 and GUCY1B3 nitric oxide receptors involved in a pathway known to

influence blood pressure, but didn't identify NOS3. The use of the 1M Illumina array in HYPERGENES allowed the identification of rs3918226 in NOS3 as major finding of the study. By combining the two studies HYPERGENES could add a further relevant player, NOS3, within the vasodilator pathway for blood pressure regulation.

4.6 Develop a set of Decision Support Systems tools combining genetic, clinical and environmental information

During the last part of the project the Consortium simulated a healthcare scenario where the BII serves clinical decision support (CDS) applications used at the point of care. For example, a patient would like to know his/her risk of developing essential hypertension. The results are sent back to the referring clinician, encapsulating the values of each SNP alleles in an HL7 Genetic Variation message payload.

In a later phase of this healthcare scenario, a CDS application (running as part of the computing environment of that clinician) is parsing the raw data encapsulated in the aforementioned message payload and attempts to associate the genetic data with interpretations of the data. The CDS is accessing the BII to get the various analyses that identified the SNPs and ranked them based on p-value of their association with essential hypertension.

CDS applications can also look at the GWAS analysis results in the context of the patient's EHR which provides a broader view to the interpretation process, for example by looking at observed clinical data related to hypertension and use that for validation. Once interpretations are generated, a CDS application can use the same HL7 structure that carried the raw data, and add to it 'interpretive phenotypes' objects associated with the patient's genotype.

It is important to note that multiple interpretation algorithms might be invoked by the same CDS application or perhaps by the EHR system that uses CDS web services hosted elsewhere, in order to validate and generate new interpretive phenotypes based on the same genotypic data.

Selected clinical parameters (BMI, Blood Pressure and Cholesterol), which are among the most represented in HYPERGENES sample, can be presented in the graphical form (see figure 5), together with predicted values generated from disease model constructed by Machine Learning group.

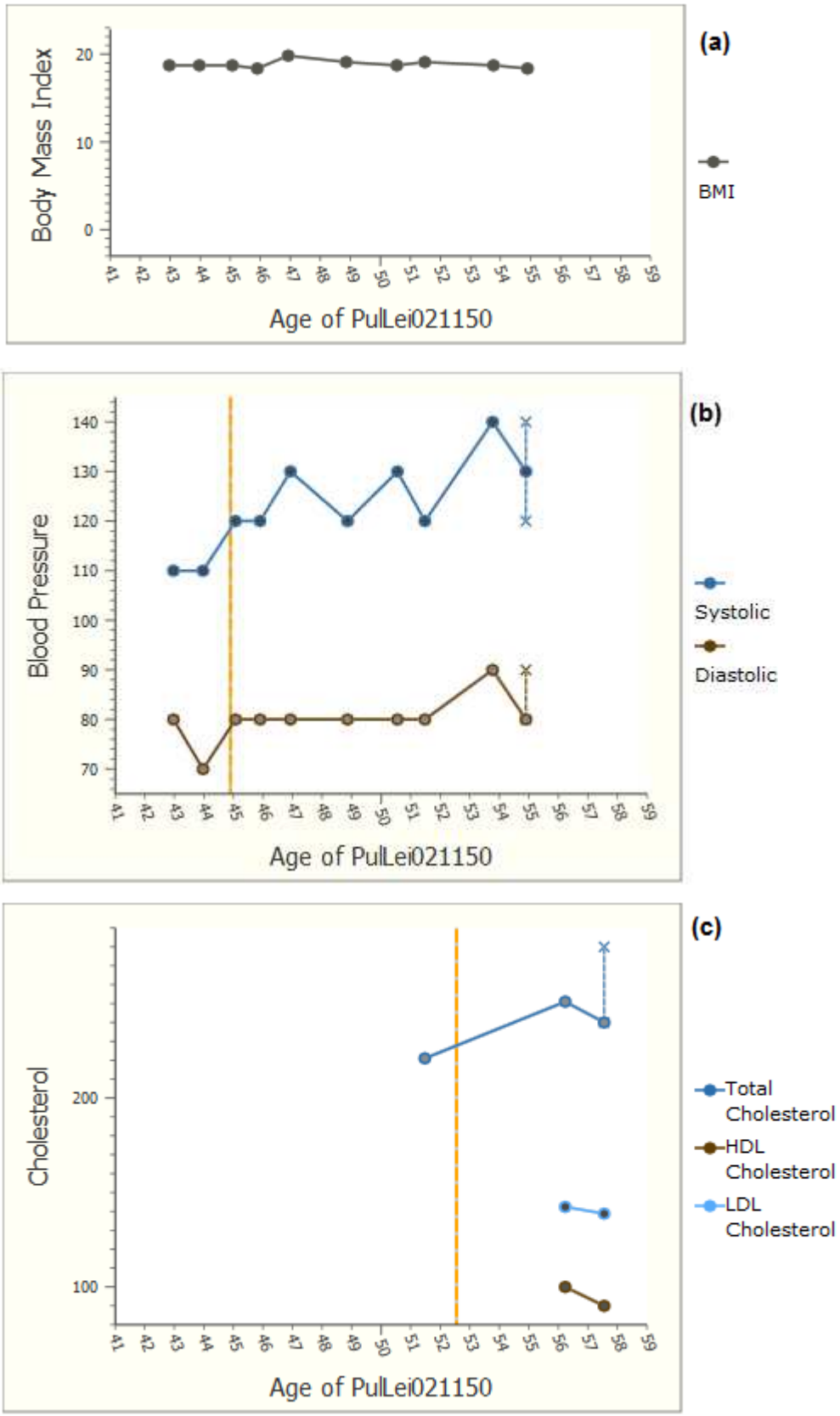

Figure 5: Patient Clinical Parameters: (a) BMI; (b) Blood Pressure: Systolic and Diastolic; (c) Cholesterol: Total Cholesterol, HDL, LDL. Yellow vertical line indicates the prediction reference date, which is the date, when prediction was produced. Dashed vertical lines demonstrate the prediction range.

As example, figure 5 shows that the BMI of the sample patient (PulLei021150) remains stable during ages 43-55. For what concerns Blood Pressure measurement, based on prediction algorithms developed during HYPERGENES, at age 45 we generated predictions for the blood pressure values at age of 55. The system predicted slight increase in both systolic and diastolic blood pressure values.

The utility of predicted values relies on the fact that when the system predicts possible development of hypertension, physician can recommend to the patient to increase the frequency follow up visits.

If the patient was genotyped, then we also present content of the risk assessed SNP alleles (see figure 6)

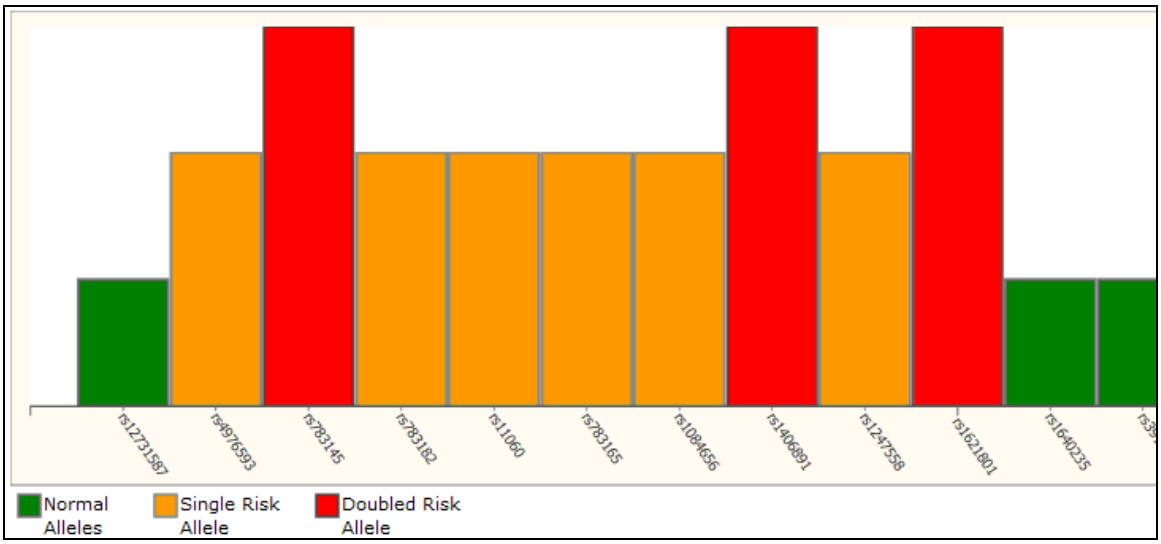

Figure 6: Patient Risk Allele measurements Displayed values for 15 risk-assessed SNPs, ordered according to their p-values. Two normal alleles are indicated by low green bar; single risk allele is indicated by intermediate orange bar; two risk alleles are indicated by high red bar.

Additionally, a link is provided to a page presenting historical and clinical guideline treatment recommendations. We can overview the Clinical Guidelines treatment recommendations for a group of similar patients as well as their actual treatments and the treatment outcome.

4.7 Develop a simple, inexpensive genetic diagnostic chip, that can be validated in our existing well-characterized cohorts

The *lab-on-chip (LoC)* is a device that integrates benefits and strengths from different established technologies. It is a microfluidic device, consisting in a PCR micro reactor and a planar array. By coupling the benefits of these technologies with a flexible and easy to use

platform, these solution represent an ideal approach to target areas where was difficult to implement molecular diagnostics till today.

HYPERGENES LoC was developed based on STMicroelectronics *In-Check*® platform. The *In-Check*® platform was designed to operate in a random access mode, with low-medium test throughput and in a simplified operation setting in terms of infrastructures and skills needed. This makes the platform and an ideal tool for the decentralized clinical laboratories or sub-district level laboratories.

Within HYPERGENES, the current microarray technology detection was for the first time extensively applied to the investigation of SNPs in human DNA, posing significant challenges, since the technology is partially adapted to this kind of application. In facts, it cannot guarantee a full discrimination between the wild type and the mutated gene, but can assure a degree of signal separation between the expected and not expected probe.

The LoC design was based on six human gene target sequences that include specific SNPs resulting positively associated with the risk for developing microalbuminuria (a parameter strongly associated with the risk for future cardiovascular events), from analysis conducted within the HYPERGENES project.

The final LoC prototype is the result of three subsequent experimental phases, as described below:

- first study cycle, in which the probes and primers were designed and the PCR protocols were optimized,
- second study cycle, during which “tails” were added to the probes to try to increase the discriminating ability of the tool. After the second study cycle the LoC still posed problems in the ability of discrimination. In facts, it should be noted that the current LoC methodology foresees a correct identification of both Alleles (wild and mutated type).
- third study cycle, meant to improve the LoC predictive ability, in which the probes were redesigned (varying the SNPs positions in the microarray layout) and the researchers worked on varying the stringency of the conditions for the protocols (varying the hybridizations temperatures, salt concentrations in the washing buffer, PCR reagents concentrations, washing buffer stringency).

In these conditions, specific thresholds must be set to discriminate the signals and, therefore, there is the need to set the platform diagnostics rules on a set of data from real samples. Based on this evidence, that the HYPERGENES project has the possibility to explore, STMicroelectronics is working on changing the approach to detection for SNPs detection applications, focusing on microarray technologies capable to univocally identify the nucleotide variation.

This last phase brought very promising results and the protocol was finalized so that the LoC discriminating ability in testing context was satisfied.
The validation performed on the final version of the LoC prototype, by the manufacturer, in testing conditions, brought to the following results: excellent performance for two targets, medium and improvable performance for two targets and below 80% of performance for the remaining two.

The LoC was then **tested in double blind** by two independent experimental laboratories, to estimate its efficiency in a "real working" conditions. Genotype data obtained from LOC were compared with known genotype data for detection efficiency.
The performance of the LoC working conditions, measured independently by the two laboratories, was overall lower than expected:

- The aggregated fraction of detected genotypes was only 48% on average,
- Among positive probes only ~70% on average called the genotype correctly, taking the Infinium-based genotype as a reference.

The use of the LoC for diagnostic purposes, considering its discriminating ability, is still not optimal for all the targets at this stage of feasibility. We must however consider that the test has been limited to a cluster of genes associated to microalbuminuria only.
The results obtained within the frame of the activities related to the porting of a set of genes into the LoC constitute an advanced phase of feasibility. Not all of the selected targets comprised in the LoC configuration responded properly (none of them with good performance) and different efficiency among the targets have been observed.

4.8 Strengthen the existing clinician-basic scientist collaborative network on the genetic mechanisms of EH

HYPERGENES Consortium brought together worldwide experts in the field of CVD genetic epidemiology and clinics. It has constituted a unique opportunity to generate a huge amount of genetic information on well-characterized European cohorts and to build the basis for edge-research on the field.
The HYPERGENES project, throughout its lifespan, focused on the generation and integration of genetic, phenotypic and environmental data from already existing well-characterized population cohorts. This allowed the generation of huge amount of information that proved to be highly valuable in the context on the project, for other studies on the cardiovascular diseases and also for the scientific community in a broader sense. In this context, the HYPERGENES Consortium established a relevant number of collaborations with

researchers from the international community, being at times contacted from other Consortia and at times proposing itself as a collaborating partner.

It is worth to say that the genetic knowledge generated, boosted by the innovative tools generated within HYPERGENES allowed the creation of a large amount of knowledge that has only been partially exploited within the project's lifespan. Far from being comprehensive, the results obtained steer future research on the topic of cardiovascular disease and pose the basis to future collaborations that will broaden the network of excellence established within HYPERGENES Consortium.

4.9 Support professional training on all aspects of the project, favoring mobility of PhD students and post-doc

The HYPERGENES partners have also promoted and participated to several activities enhancing the mobility of European PhD students on clinical phenotyping, genomic, bioinformatics and genetic epidemiology. About 60 visits for research activities concerning HYPERGENES have been performed by researchers and PhD students involved in the project.

4.10 Dissemination

HYPERGENES gave great relevance to dissemination activities, through publication of many papers in peer reviewed journals, participation to international conferences and workshops and constant update of the project website.

The HYPERGENES project website (www.hypergenes.eu), *on-line* since March 2008, has been periodically updated and enriched with information on project activities and results.

All sections have been constantly updated. With regards to the Dissemination page (http://www.hypergenes.eu/dissemination.html), it has been updated with the latest information on the activities held by the Consortium to create awareness on project advancements, with a specific focus on the related scientific publications issued by the Consortium.

4.11 Translation of the project results

Most of the project results, as described in the previous paragraphs, offer significant translational opportunities. However it must be pointed out that all HYPERGENES outcomes need to go through a validation or re-engineering process, before they can be used in a day-by-day healthcare context.

The BII prototype developed within HYPERGENES, has been highly disseminated through a number of publications and presentations of papers in top quality scientific venues of the field

of biomedical informatics and IBM Research Lab in Haifa makes an effort to suggest it to IBM Software and Services Groups as the next generation of clinical genomics solutions.

During HYPERGENES, STMicroelectronics had the chance to explore the potentiality of the *In-Check*® platform current detection approach towards the SNPs detection, and considering non optimal results, it is triggering a series of activities to both improve the current technology and look for an alternative methodology to avoid/reduce false diagnosis.

The main discovered genetic polymorphism associated to EH, according to the project results confers a significantly higher risk of developing hypertension in individuals which are homozygous for the risk allele. Such findings could be translated both in new therapeutic intervention as well as integrated in a diagnostic tool detecting the risk for EH.

Furthermore a huge amount of data has been generated within the project: in particular genetic data of more than 12,000 subjects from well phenotyped cohorts including the entire FLEMENGHO-EPOGH cohort, a random population sample recruited in 6 European countries and followed up for many years with the collection of many data on target organ damage at different periods. This will allow further genotype-phenotype association analyses and collaborations for data sharing with other Consortia involved in genetic studies.

5 Project Impact

HYPERGENES results apply to a variety of fields, as expected from the inherent project multidisciplinarity. They can be summarized per areas as follows:

IT tools to support multidisciplinary research

The **BII developed during HYPERGENES** that allows integration, harmonization and standardization of clinical, environmental and genomic data. The BII was primarily developed as a data warehouses, established in an attempt to accomplish such integration and support patient-centric care as well as secondary use of the data such as analysis of aggregated data in the context of clinical research. In the context of HYPERGENES project clinical and environmental data as well as genotypes of more than 12,000 patients, collected according different protocol and stored in different databases have been integrated and represented in the data wharehouse. Besides preserving data, and allowing advanced queries, the system has the advantage of offering services that reach out to Ensembl, UCSC and HapMap and extracting SNP's annotations. Furthermore the BII has developed some functionality that allow to persist analysis results.

Besides, the BII propose some features (CSD) that are meant to ease the Clinicians daily practice.It allows the selection of clinical parameters that can be presented in the graphical form, and the access of existing Clinical Guidelines. Moreover, if available then Blood Pressure and Cholesterol predictions, generated from disease model constructed and integrated in the BII, are also displayed.

The main advances of the BII are the potential to serve both research and clinical requirements; the alignment of multiple data marts a with standards-based warehouse; and the disruptive technology of XML databases with native indexing to avoid shredding original XML data to relational and limited structures in the warehouse.

In addition, some tools were developed during the project to serve specific genetic researchers needs. Such tools might be highly valuable for small research groups. Among the most promising tools.

- Development of a SNP-Ranker software, that accounts for experimental data and knowledge information about genomic regions
- Creation of a tool (IBM Bioclinical Data Mining Tool) that allows users to execute machine learning and data mining algorithms on large data sets.

Predictive biomarkers for the design and the production of diagnostic chips

HYPERGENES identified a ranked list of SNPs associated with EH and TODs in a sample of 4,000 well characterized subjects, through a high density array and validation of the results in an independent sample of 8,000 individuals.

One of the most promising HYPERGENES discoveries is a SNP in the promoter region of the eNOS gene (endothelial nitric oxide synthase), significantly associated with hypertension (OR, 1.54; 95% CI, 1.37 to 1.73; p-value = $2.58 \cdot 10^{-13}$). The result was confirmed by meta-analyzing in-silico data from ASCOT/AIBIII/NBS, BRIGHT, EPIC-Turin, HYPEST and NORDIL/MDC samples. eNOS, which catalyses the synthesis of nitric oxide (NO) by vascular endothelium, is responsible for the vasodilator tone that is fundamental for the regulation of blood pressure. Furthermore, eNOS is a critical mediator of cardiovascular homeostasis through regulation of blood vessels diameter and of the maintenance of an anti-proliferative and anti-apoptotic environment.

Such finding may open new opportunities for drug discovery and for the detection of risk factors associated to the development of EH later in life. In this context genetic risk factors allow for an evidence-based prevention strategy, and could be used to define personalized follow up frequencies.

Other genetic polymorphisms which showed promising results need further validation in independent samples.

The project applied the LoC methodology for the analysis of Human DNA, and developed a LoC for the detection of genetic polymorphisms associated to microalbuminuria. The STMicroelectronics *In-Check*® platform, which was the technology applied during the process, was designed to operate in a random access mode, with low-medium test throughput and in a simplified operation setting in terms of infrastructures and skills needed. This makes the platform and an ideal tool for the decentralized clinical laboratories or sub-district level laboratories

Model allowing prediction about EH and associated TOD:

HYPERGENES devoted lots of efforts in the creation of models for EH and associated TOD, in order to elucidate disease mechanisms and put the basis for the development of new therapeutic strategies.

The main outcomes include :

- a model that predicts the future clinical range of each measure from the clinical data available integrating the contribution of the genomic data to the prediction. Such model, based on Machine Learning approach, demonstrated a good prediction performanceIdentification of several functional pathways that concern vasodilator and

natriuretic mechanisms, electrolyte homeostasis and hypertrophic processesDevelopment of two strategies for clinical decision support systems that rely on mining patient data

HYPERGENES results will have a strong impact on the scientific community involved in research on EH and other complex diseases. A unique European cohort, gathering a great amount of genetic and well-characterized phenotypic data, has been built and it represents a valuable resource for future projects. Thanks to HYPERGENES, strong collaborations and synergies between different centers worldwide have been strengthened and the project effort will contribute to build the European leadership in the domain of advanced genomics.

The dissemination of HYPERGENES results will support the definition of new strategies for prevention and treatment in EH. Furthermore they will play a central role in enhancing the appropriateness of the therapeutic protocols, toward a more personalized treatment and the reduction of side effects of drugs and pharmaceutical health expenses.

Concerning the commercial impact, the Consortium collected and analyzed the entire foreground generated, identifying the best exploitation strategy for each of them. One patenting procedure is ongoing.