Mr. SymBioMath

High Performance, Cloud and Symbolic Computing in Big-Data Problems applied to Mathematical Modeling of Comparative Genomics

EU FP7 Industry-Academia Partnerships and Pathways
Project Nr. 324554

Deliverable D1.2
Data Sources

Deliverable Number	**D1.2**
Deliverable Title	**Data sources**
Type of Document	**Report**
Dissemination Level	**Public**
Workpackage	**WP1: Preliminary Analysis**
Lead Beneficiary	**UMA**
Contractual Delivery Date	**31.07.2013**
Actual Delivery Date	
Editor(s)	**O. Trelles, J. Karlsson**
Author(s)	**J.A. Cornejo, J.A. Arjona, A. Muñoz, O. Trelles**
Quality reviewer(s)	**P. Heinzlreiter**

Document Log

Version	Author(s)	Date	Modifications
V0	J.A. Cornejo, O. Trelles	13.06.2013	
V0.8.4	J.A. Cornejo, J.A. Arjona, A. Muñoz, O. Trelles, P. Heinzlreiter	18.07.2013	Updates and review
V1	J. Karlsson	24.07.2013	Review and various fixes

Table of Contents

0. Executive Summary

This deliverable aims to identify and locate the sources of data and define the benchmarking and testing procedures to validate the system using such data collections.

Regarding the content, two main sources of data will be used: Those related to biological data and those related to clinical data. From the proprietary perspective, data collections will be classified as Public data sets and Private data sets. Table 1 describes the main collections.

Type	Molecular data	Clinical data	Description / Comments (short)
Public	Genbank		Genes and genomes retrieval and annotations
	SwissProt		Protein sequences and functional annotations
	EMBL		The EMBL Nucleotide Sequence Database, is a comprehensive collection of nucleotide sequences and annotation from available public sources http://www.ebi.ac.uk/embl
		WTCCC	Welcome Trust Case Control Consortium with genotype and phenotype data
Proprietary		Unified Allergy DB	Clinical phenotype information gathered by SAS partner
	Olive DB	NGS assembled data, gene expression and metabolite data	From the Spanish National project: “Generation of genomic tools in olive and application to the analysis of fruit and oil quality and agronomical traits (OLIGEN)”
	Tomato DB	Gene expression and metabolite data	From the Spanish National project: Identification of Genes and Molecules Associated to Tomato Fruit Quality and Participation in the Sequencing of Euchromatic Regions of Chr9. A Genomic Approach (ESP-SOL)
	Strawberry DB	Gene expression and metabolite data	From the Spanish National project: “Genetical genomics for improving strawberry fruits nutritional quality” (FraGENOMICs)
	LNCC data	Assembled sequence data	Data belonging to Mycoplasma strains, fungal genomes of Sporothrix and several metagenomes from different locations.

Table 1. Main data collections in the Mr.SBM project. For detailed information see the main text.

During the implementation of work-packages (WP2 and WP3) the consortium will define the local installation of databases or their remote access.

1. Proprietary clinical datasets

The main data collection in this category corresponds to the database belonging to project partner SAS and is hosted and maintained by project partner UMA. It contains Allergic patients records linked to their molecular information about SNP (Single Nucleotide Polymorphism).

1.1 Unified allergy database

This database is maintained and exploited in the RIRAAF project coordinated by the SAS-HCH partner. The data belongs to different hospitals all over Spain and has been gathered using already validated protocols from the previous RIRAAF working period. In this integrated infrastructure the population covered represents around 5 million people distributed across different geographical areas. This implies that a process of consensus and development of protocols has occurred throughout these years, leading to the application of the same algorithms with the final registration in a data base that is available on the webpage of the RIRAAF (www.bitlab-es.com/riraaf) hosted and maintained by UMA under strict regulations that have been previously setup following the legal and ethical regulations and rules in Spain.

This has enabled to compare the data for multicentre studies. All the new clinical centers have been trained to follow the same procedures in order to reduce inter-centre variability and the procedures carried out so far will continue during the whole extension of the RIRAAF (2013-2016). This provides a continuous feedback between groups that increases the sum of the knowledge. The common cooperative structure will enable us to obtain a great number of cases per year that will provide sufficient sample sizes of patients and controls to carry out the objectives and WPs proposed in Mr.SymBioMath project. It is expected that more groups from other countries will consult the database, leading to further opportunities to establish international collaborations. All the clinical groups have a monthly meeting in order to supervise all the information, discuss the data, review how this provides input for all the ongoing projects, and propose new projects for collaborations.

The UMA group provides transverse support to all this large amount of data that will be generated in the next period, to provide relevant information on the interaction of clinical and analytical data (e.g., clinical characteristics and particular genetic, metabolic or proteomic data, as well as potential confounders) which can only be obtained with their knowledge and resources in huge multivariate data analyses.

The added value of the cooperative structure depends upon the multidisciplinary collaboration effort to elucidate both the basis of adverse reactions to drugs with an immunological basis and the basis of reactions to allergens. Because these disorders are multifactorial and because complex mechanisms are believed to be involved in their etiology, clinical course and therapeutic outcome, this data collection constitutes a privileged cooperative infrastructure that covers several potentially relevant issues. In the first period of the RIRAFF project the groups dedicated much effort to unifying the criteria, focusing on common objectives, interchanging researchers and coordinating research projects.

It is clear that genetic studies, either CGAS, GWAS or other types of studies, require a large sample size to permit reliable findings (i.e., with a high statistical power). This also permits replication of the findings in populations of diverse geographic origins or with diverse degrees of exposure. In addition, because some genetic characteristics are rare (i.e. individuals homozygous for rare variants, or double carriers of particular genetic combinations), or certain pathologies are uncommon, a very large sample size is required and this requires a coverage area of millions of individuals. This couldn't had been reached without a structure like RIRAAF ("Red de Investigación de reacciones adversas a alérgenos y fármacos (RIRAAF, Instituto Carlos III RD12/0013/0001)) where the collaborative compromise of basic groups enabled to obtain the maximum data from patient samples, including genetic, metabolic, proteomic and other analyses, and the transverse collaboration of the bioinformatics groups provided a complete system for the storage, management and exploitation of the clinical and research data.

Genome-wide case-control studies use high-throughput genotyping technologies to assay hundreds of SNPs and relate them to clinical conditions or measurable traits. To understand underlying causes of complex disease traits, it is often necessary to consider joint genetic effects (epistasis) across the genome. The concept of epistasis was introduced around 100 years ago. It is generally defined as interactions among different genes. Recently, the essential role of gene-gene interactions in the structure and evolution of genetic systems has been highlighted. Therefore, gene-gene interactions have long been recognized to be fundamentally important for understanding genetic causes of complex disease traits. At present, identifying gene-gene interactions from genome-wide case-control studies is computationally and methodologically challenging.

Another relevant issue giving added value is that many groups involved in this program already have a relevant international projection, maintaining collaboration with several national (e.g., CIBERs) and international research groups, and many researchers involved in this program belong to international advisory committees and to scientific journal editorial or review boards. All these existing interactions with research groups and structures outside the RETICS can be used to empower the research of this structure through collaborations with groups within and outside the structure, focusing on the objectives mentioned in this program. Of particular relevance, regarding the genetics program, is collaboration with international groups because often certain genetic variants and/or clinical associations are specific to an ethnic group, and collaboration with international groups that can provide samples and information from subjects with different ethnic origins or exposed to different environmental factors is compulsory to make replication studies and to validate genetic, metabolic or proteomic biomarkers.

The synergies between the groups involved in this programme have already been demonstrated in joint research projects, clinical trials and publications. Even though some groups are new to the structure, the interaction of the groups involved in this program has already produced 17 joint publications in high impact journals (see the curricula of the group coordinators for details). In the future, with this common programme, the incorporation of new groups, and with the ongoing projects corresponding to the previous period of the

RETICS, these interactions and hence the common findings and publications can be expected to increase considerably in the next few years.

1.2 Description of the Allergic application domain

Non-steroidal anti-inflammatory drugs (NSAIDs) are the drugs most frequently involved in hypersensitivity drug reactions [Demoly 2004, Johansson 2004]. Histamine is released in the allergic response to NSAIDs and is responsible for some of the clinical symptoms. The aim of this study is to analyze clinical association of functional polymorphisms in the genes coding for enzymes involved in histamine homeostasis with hypersensitivity response to NSAIDs.

Drug hypersensitivity reactions (DHRs) are a frequent reason for consultation in allergy departments. They include immunologically mediated reactions, where the mechanisms involved may be either immunoglobulin (Ig)–E mediated or T-cell dependent [Roberts 2001, Szczeklik 2009], and non-immunologically mediated reactions, the most frequent of which involve cross intolerance of non-steroidal anti-inflammatory drugs (NSAIDs) [Gomes 2005, Szczeklik 2009, Sanches-Borges 2010]. It has been difficult to determine the true prevalence of DHRs because of difficulties concerning a precise definition and identification of reactions, as well as a lack of population studies [Gruchalla 2003]. Figures reported vary and it has been estimated that DHRs account for 3% to 6% of all hospital admissions and that they occur in 10% to 15% of hospitalized patients [Thong 2011]. However, several biases exist, such as differences in study populations and diagnostic criteria and methods [Demoly 2002, Adkinson 2002, Mockenhaupt 2007]. DHRs are associated with a high use of health care services, particularly in adults. Indeed, in Spain drug allergy is the third most common reason for consultation in allergy departments, after rhinitis and bronchial asthma [Gamboa 2009]. The diagnosis of DHR is usually based on clinical history, skin testing, and to a lesser extent in vitro testing [Romano 2012]. Clinical history, however, is often not reliable [Messaad 2004], and reagents used in skin testing and/or in vitro diagnosis are seldom standardized, and even when appropriate, if the reaction occurred a long time previously, sensitivity can be lost or the test can show negative results [Blanca 1999].

Thus drug provocation testing (DPT) often remains the sole alternative [Aberer 2003]. However, DPT is cumbersome, often dangerous, and sometimes non-definitive [Blanca 2009]. New diagnostic tools, such as the basophil activation test (BAT) for IgE-dependent reactions [Abuaf 1999, Sanz 2002, Torres 2004, Ebo 2006, Aranda 2011] and lymphocyte stimulation studies [Luque 2001, Nyfeler 1997] have been proposed, though they are only available at a few centers. Epidemiological studies of DHRs report varying results because of differences in diagnostic methods [Bousquet 2009, Torres 2003]. Drug allergy is not a static process; it varies over the years and is related to changes in patterns of drug consumption, the introduction of new drugs and the withdrawal of others, and the establishment of new indications [England 2003, Dietrich 2009, Gomes 2005, Blanca 1995, Doña 2011].

1.3 The Unified Allergy data model

The implemented database unifies patient data, biological samples, clinical data, diagnostic tests and agents extracted from the associated centers (hospitals).

The patient data is the main information stored in the database, containing information that could be useful for statistical or GWAS studies (see Figure 1). The information stored contains an unique value (anonymized) that identifies the patient inside the database (patient id), the year when the doctor evaluates the patient, the list of doctors that are treating the patient, the patient's birth year, the actual location of the patient, the gender (Male/Female), the job (Farmer, Helper, Unknown, etc.) and a plain text field with information that could be useful to determine the source of patient's allergy.

There are also another data related with the information of the patient (personal and familiar history, samples extracted, allergic episodes and tests). The personal history contains information such as for example if the patient is diabetic, hypertensive, smoker, if has suffered any surgery, if is allergic to insect stings, if has chronic urticaria or food intolerance. The previous knowledge apart from giving extra information about the patient, could be useful to determine relations between for example, smokers patients that are allergic to some allergens (i.e. identify the influence of smoking in the allergy reaction. The family history contains one entry per component of the family (father, mother, brother, sister, etc.). The relationship field contains the relation of the patient with the family component. There are other fields inside the familiar history to indicate if the family has experienced food, drug, inhalants or hymenopterous allergy. The table of the samples extracted contains information about two types of samples (DNA and serum), the extraction date and a code to identify the SNPs file associated. In the case of the allergic episodes that the patient has experienced we store the description given by the patient, the starting year, the causing agents, the diagnosis and the frequency of these episodes. Finally in the tests table we have information about the test type (skin prick test, total IgE, etc.), the date and the obtained results.

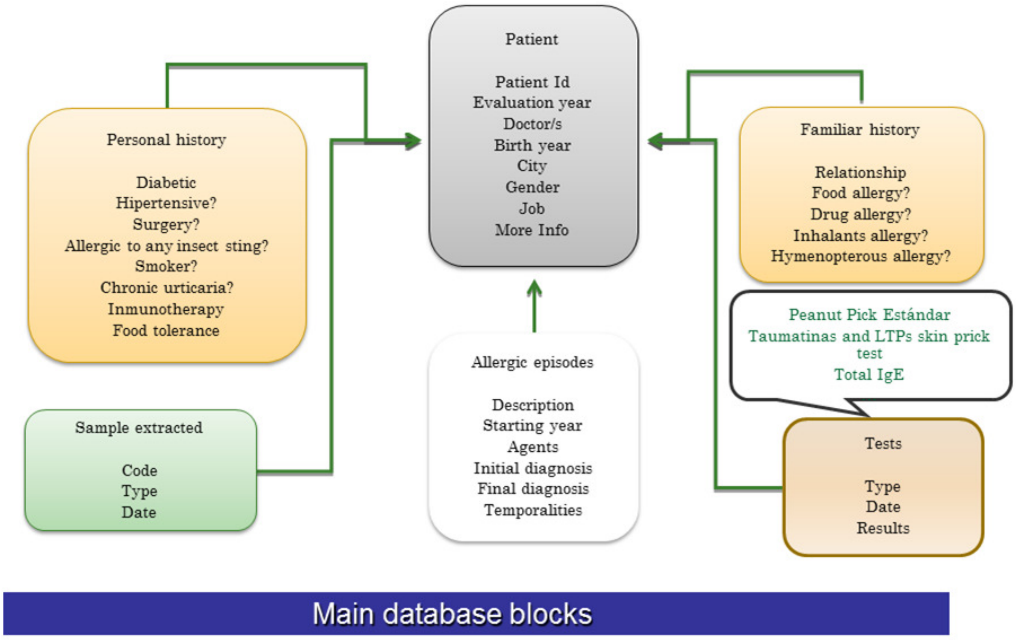


Figure 1. Main functional blocks in the 'Unified Allergic Database', including patient, familiar history, allergic episodes, trials, samples and tests.

1.4 Estimated number of records in the Unified Allergy database

An initial indicator of the available patient data is provided in Table 2. These data sets and their volumes will be promptly meet.

Study	Samples	Number patients	Determinations per patient	TOTAL
ISAC	Food and inhalant allergens	6000	Specific IgE to allergens 100	60000
SNPs	NSAIDs hypersensitivity	1000	686.000 (Affimetrix)	$6x10^8$
SNPs	Food allergy	1800	197000 HumanImmuno BeadChip (Illumina)	$3.5x10^8$
Protein Arrays		1800	40	72000
Skin test		1800	60	108000
SNPs	Drug allergy	500	197000 HumanImmuno BeadChip (Illumina)	$9.85x10^7$
LAR GWAS	LAR	800	570.000 Affymetrix (Imputation of 6 Million SNPs)	$4.8x10^9$

Table 2. **Proprietary databases from SAS: Estimated number of patients.**

1.5 Molecular data for allergic patients

At present SAS-HCH has available an initial collection of molecular data corresponding to the genotyping of 124 patients (cases and controls). It is expected the number of patients genotyping data will increase along the lifetime of the project

Non-steroidal anti-inflammatory drugs (NSAIDs) are the most consumed medicines worldwide because of their efficacy and utility for the treatment of pain and clinical symptoms of inflammatory diseases. However, they are associated with a broad range of adverse events including hypersensitivity reactions (HRs).

HRs to NSAIDs are complex because they can be triggered by both immunological specific and pharmacological mechanisms [Kowalski 2011]. The first group is mediated by specific IgE antibodies or T cells, and are called selective reactions (SR) because is only one culprit drug or chemical group is involved [Canto 2009]. Those reactions mediated by pharmacological mechanisms are known as cross-intolerance (CI) [Stevenson 2006], because chemically different NSAIDs activate metabolic pathways that lead to the release of preformed (tryptase, histamine, chimase and proteoglicans) [Kowalski 2011] and de novo synthetized (prostaglandins and leukotrienes) inflammatory mediators, as well as cytokines, chemokines and growth factors, that are responsible for vasodilation, increase in vascular permeability, smooth muscle contraction and bronchospasm, and the development of urticaria/angioedema [Minai-Fleminger 2009]. CI is the most frequent type of HRs to NSAIDs [Doña 2011].

HRs to NSAIDs include a heterogeneous group of entities, so the classification of patients is complex [Blanca 2012]. SR comprise immediate reactions (IgE-mediated), with clinical entities being classified as single drug-induced urticaria/angioedema/anaphilaxis [Stevenson 2001, Szczeklik 2003], and delayed reactions, which are mediated by different cell types (T lymphocytes, cytotoxic T cells, NK cells, among others) [Kowalski 2011]. In CI three different groups of clinical entities can be described:

1. NSAIDs-induced rhinitis/asthma (also known as aspirin-exacerbated respiratory disease or aspirin-induced asthma, AIA).
2. NSAIDs-exacerbated urticaria/angioedema in patients with chronic urticaria (CU), and
3. Multiple NSAIDs-induced urticaria/angioedema (MNSAID-AUA) in patients without history of underlying chronic skin and/or respiratory diseases [Kowalski 2011].

Some patients with CI present a mixed pattern that involves both skin and airways, also called blended reactions according to the terminology proposed by Stevenson and Szczeklik, which combines categories 1 and 3 [Doña 2011, Blanca 2012].

The HCH group coordinates the Spanish Network “Red de Investigación de reacciones adversas a alérgenos y fármacos (RIRAAF, Instituto Carlos III RD12/0013/0001), enabling us to access an important number of samples from patients with HRs to NSAIDs. We have provided for this project molecular data (genotypes) of patients with MNSAID-AUA and healthy controls from a genome-wide association study (GWAS).

To be included all patients had to have experienced episodes with more than 2 different NSAIDs without NSAID-exacerbated CU. However, in those instances where this criterion was not met, diagnosis was confirmed by oral provocation test in a single-blind procedure as described [Kowalski 2011]. On the first day, placebo capsules were given at different time intervals. At least 1 week later, increasing doses of acetylsalicylic acid were administered orally at intervals of 90 minutes up to a total of 2-4 administrations. The procedure was stopped if any cutaneous and/or respiratory symptoms or alterations in vital signs (rhythm modifications, decrease in peak expiratory flow rate or hypotension) appeared, and patients were evaluated and treated. If no symptoms occurred, the therapeutic dose was achieved by taking a course of two additional days giving acetylsalicylic acid 500 mg every eight hours. Those patients who responded to one single drug and showed good tolerance to a strong COX-1 inhibitor were considered selective responders and therefore excluded from the study. Considering potential interaction between food allergy and NSAIDs, patients with a clinical history of food allergy or positive IgE antibodies for food allergens despite having no history of food allergy were not included. The controls comprised individuals without any previous history of HRs to any drug that usually take NSAIDs.

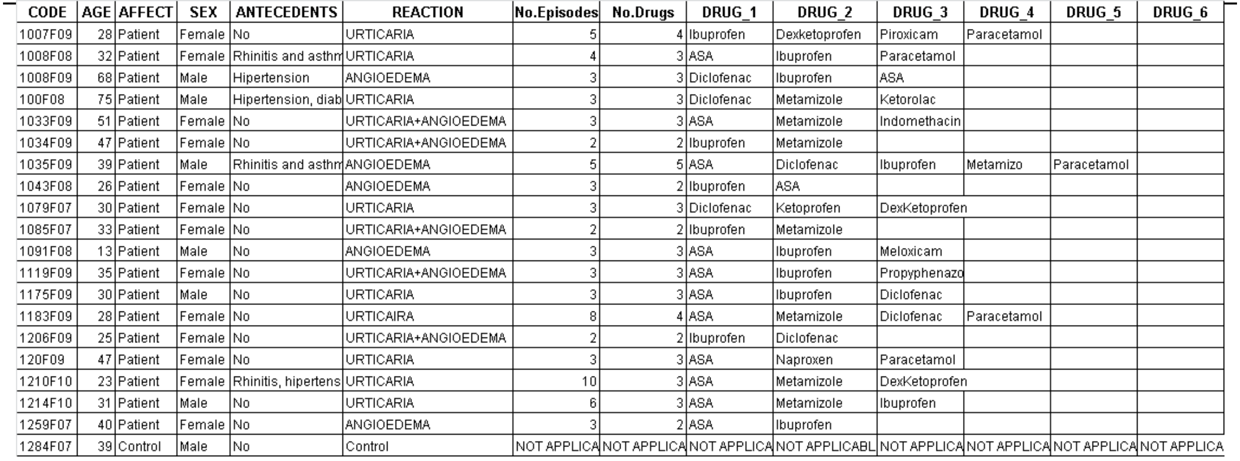
As copy number variations (CNVs) account for a large amount of genetic variation in the human genome and are still relatively under-ascertained. We plan to perform a copy number variation analysis with cn.farms [Clevert 2011]. cn.farms is powerful tool –developed by JKU partner- to detect this kind of variation with a low false discovery rate. Therefore to perform this analysis the Affymetrix SNP 6 raw data (CEL files) are needed. More information about technical details is given in [Affymetrix].

```
#CHP File=D:\SNP6\0091\0091(263sp)CHP\1316F09.birdseed-v2.chp
#Exec GUID=00006243-4b68-44ee-58ed-005ced000be5
#GenomeWideSNP_6.na32.annot.db
#%genome-version-ucsc=hg19
#%genome-version-ncbi=GRCh37.1
Probe Set ID Call Codes   Forward Strand Base Calls dbSNP RS ID
SNP_A-2131660BB     TT    rs2887286
SNP_A-1967418AB     AG    rs1496555
SNP_A-1969580BB     GG    rs41477744
SNP_A-4263484BB     TT    rs3890745
SNP_A-1978185AB     GA    rs10492936
SNP_A-4264431AB     GA    rs10489588
SNP_A-1980898AB     GC    rs2376495
SNP_A-1983139AA     AA    rs4648462
SNP_A-4265735AA     GG    rs10492939
SNP_A-1995832AB     CG    rs9424283
SNP_A-1995893AB     AG    rs2154068
SNP_A-1997689BB     GG    rs12060299
SNP_A-1997709AA     TT    rs10909802
SNP_A-1997896AA     AA    rs16824230
SNP_A-1997922AB     TG    rs17404435
SNP_A-2000230NoCall ---   rs12059199
SNP_A-2000332AB     CG    rs6702000
SNP_A-2000337AA     TT    rs16838547
SNP_A-2000342AA     GG    rs16838549
SNP_A-4268173AB     CT    rs3912752
SNP_A-2002663AB     GA    rs505933
SNP_A-2004169AB     CT    rs4654438
SNP_A-2004249BB     CC    rs676853
SNP_A-4268681AA     CC    rs583027
SNP_A-2004332AA     GG    rs350165
SNP_A-4268770BB     TT    rs10489135
SNP_A-4268887BB     CC    rs12120353
SNP_A-2005859AB     TG    rs241212
SNP_A-2006248AA     CC    rs17349020
SNP_A-2007744NoCall ---   rs3766969
SNP_A-2008162BB     GG    rs41355644
```

Table 3 Genome wide data for close to 1M SNPs. First column correspond to the SNP identifier, second column the values (nucleotide) in each allele and the RS value.

There is a file for each patient (30MB approximately each one). These files are associated with phenotype information (exported from the Unified allergy database) with the following minimal format (see Table 4):

FIELD	Description
CODE	link to patient and SNP data file
AGE	in number of years
AFFECTED	Control / Patient
SEX	Male / Female
ANTECEDENTS	Description (with Controlled vocabulary)
REACTION	Description (with Cotrolled vocabulary)
No. EPISODES	Numeric value
No. DRUGS	Number of different drugs prescribed
DRUG_1...N	Drug IDs



CODE	AGE	AFFECT	SEX	ANTECEDENTS	REACTION	No.Episodes	No.Drugs	DRUG_1	DRUG_2	DRUG_3	DRUG_4	DRUG_5	DRUG_6
1007F09	28	Patient	Female	No	URTICARIA	5	4	Ibuprofen	Dexketoprofen	Piroxicam	Paracetamol		
1008F08	32	Patient	Female	Rhinitis and asthm	URTICARIA	4	3	ASA	Ibuprofen	Paracetamol			
1008F09	68	Patient	Male	Hipertension	ANGIOEDEMA	3	3	Diclofenac	Ibuprofen	ASA			
100F08	75	Patient	Male	Hipertension, diab	URTICARIA	3	3	Diclofenac	Metamizole	Ketorolac			
1033F09	51	Patient	Female	No	URTICARIA+ANGIOEDEMA	3	3	ASA	Metamizole	Indomethacin			
1034F09	47	Patient	Female	No	URTICARIA+ANGIOEDEMA	2	2	Ibuprofen	Metamizole				
1035F09	39	Patient	Male	Rhinitis and asthm	ANGIOEDEMA	5	5	ASA	Diclofenac	Ibuprofen	Metamizo	Paracetamol	
1043F08	26	Patient	Female	No	ANGIOEDEMA	3	2	Ibuprofen	ASA				
1079F07	30	Patient	Female	No	URTICARIA	3	3	Diclofenac	Ketoprofen	DexKetoprofen			
1085F07	33	Patient	Female	No	URTICARIA+ANGIOEDEMA	2	2	Ibuprofen	Metamizole				
1091F08	13	Patient	Male	No	ANGIOEDEMA	3	3	ASA	Ibuprofen	Meloxicam			
1119F09	35	Patient	Female	No	URTICARIA+ANGIOEDEMA	3	3	ASA	Ibuprofen	Propyphenazo			
1175F09	30	Patient	Male	No	URTICARIA	3	3	ASA	Ibuprofen	Diclofenac			
1183F09	28	Patient	Female	No	URTICAIRA	8	4	ASA	Metamizole	Diclofenac	Paracetamol		
1206F09	25	Patient	Female	No	URTICARIA+ANGIOEDEMA	2	2	Ibuprofen	Diclofenac				
120F09	47	Patient	Female	No	URTICARIA	3	3	ASA	Naproxen	Paracetamol			
1210F10	23	Patient	Female	Rhinitis, hipertens	URTICARIA	10	3	ASA	Metamizole	DexKetoprofen			
1214F10	31	Patient	Male	No	URTICARIA	6	3	ASA	Metamizole	Ibuprofen			
1259F07	40	Patient	Female	No	ANGIOEDEMA	3	2	ASA	Ibuprofen				
1284F07	39	Control	Male	No	Control	NOT APPLICA	NOT APPLICA	NOT APPLICA	NOT APPLICABL	NOT APPLICA	NOT APPLICA	NOT APPLICA	NOT APPLICA

Table 4. Phenotype data: Each row represents a patient (i.e. "sick") or control (not "sick"). Some demographic data and previous history of concomitant pathologies are described. In the case of the patients, the data includes the number of previous episodes (number of times the patient had a reaction), number of drugs involved in the treatment and which those drugs are. Obviously this is not included for the controls

2. Public clinical datasets

Correspond to the **WTCCC data collection**. This ***dataset*** was generated by the Wellcome Trust Sanger Institute in collaboration with the 1958 BC, but is being distributed as part of the ***WTCCC***.

> “The primary purpose of the WTCCC is to accelerate efforts to identify genome sequence variants influencing major causes of human morbidity and mortality, through implementation and analysis of large-scale genome wide association studies. Additional objectives include the development and validation of informatics and analytical solutions appropriate to the scale and nature of the project, as well as use of the data generated to answer important methodological and biological questions relevant to association studies in general, and in the UK in particular (for example issues of population substructure).”
>
> More information in : http://www.wtccc.org.uk/info/access_to_data_samples.html

The Wellcome Trust Case Control Consortium (WTCCC) was established with an aim to harness the power of newly-available genotyping technologies to improve our understanding of the aetiological basis of several major causes of global disease. The consortium has gathered genotype data for up to 500,000 sites of genome sequence variation (single nucleotide polymorphisms or SNPs) in samples ascertained for the disease phenotypes. Analysis of the genome-wide association data generated has led to the identification of many SNPs and genes showing evidence of association with disease susceptibility, some of which will be followed up in future studies.

2.1 Population sub-structure

It has been known for some time that geographical population structure (i.e. differences in allele frequencies in different geographical regions) and geographical variation in disease prevalence can lead to false positive, and false negative results in population-based disease association studies. For studies of this size, it has been shown recently that population structure within the British Caucasian population can result in poorly calibrated tests of association. In the statistical analysis, geographical sub-region information was used to assess the extent and nature of any population structure present in Great Britain, and to advise on design strategies and analysis methods that efficiently and accurately allow for this. Information on the results of this analysis are described in full in the WTCCC paper.

2.2 Access to WTCCC genotype data

The primary purpose of the WTCCC is to accelerate efforts to identify genome sequence variants influencing major causes of human morbidity and mortality, through implementation and analysis of large-scale genome wide association studies. Additional objectives include the development and validation of informatics and analytical solutions appropriate to the scale and nature of the project, as well as use of the data generated to answer important methodological and biological questions relevant to association studies in general, and in the UK in particular (for example issues of population substructure).

More information in: http://www.wtccc.org.uk/info/access_to_data_samples.html

> ***Note**: Although UMA and JKU have permissions to exploit this data set we need to ask for specific permission to use the collection in this project. Access to the data will require the completion of a Data Access Agreement. A specific Data Access Agreement for the United States is available: Applications can include collaborators, but each Institution must submit a signed Data Access Agreement.*

This data collection will be used to test and benchmark large GWAS studies (it includes more than 50 thousand patients).

3. Proprietary molecular datasets

3.1 Olive data collection.

Contains NGS assembled data, gene expression and metabolite data, gathered during the lifetime of the Spanish National project: “Generation of genomic tools in olive and application to the analysis of fruit and oil quality and agronomical traits (OLEAGEN)”

These sequencing data sets belongs to the OLEAGEN project aimed to generate genomic tools in olive and apply this technologies to the analysis of fruit and oil quality and agronomical traits. For that reason transcriptomic sequencing is a priority to discriminate between genes with real agronomical interest. A genomic assembly is of course necessary but for the gene selection with a transcriptomic assembly is enough and is a time saving for the project.

The sequencing dataset used for the test is composed by 12 different libraries obtained from Sanger and 454 sequencing technologies. In addition 454 GS/FLX sequencing technology was added during the process. As a result, libraries from buds (5,6,7 and 8) were sequenced with the old version (which produces shorter read lengths), and the 4 last 454 libraries were obtained with the new version (that produces average lengths around 450 pb). Libraries are shown in Table 5.

#	Library	Tissue	Variety	Technology	N. reads	Av. length(bp)	N. reads after trimming	Av. length after trimming (bp)	% clean reads
1	OLmeso	Mesocarp	Lechin’	Sanger	14,688	989,91	13,477	522,74	91,76%
2	OLmer	Buds	‘Picual’ x ‘Arbequina’	Sanger	10,174	783	8,138	323	79.99%
3	OLroot	Roots	Mix	Sanger	11,136	753	8,324	368	74.74%
4	OLrest	Rest of tissues	Mix	Sanger	11,52	680	8,244	344	74.03%
5	MIA	Inactive buds	‘Arbequina’	454	171,762	142	147,826	138	86.06%
6	MAA	Active buds	‘Arbequina’	454	178,755	237	176,062	235	98.49%
7	MIP	Inactive buds	‘Picual’	454	79,473	276	78,001	270	98.15%
8	MAP	Active buds	‘Picual’	454	230,737	241	225,652	239	97.80%
9	MID1	Mesocarp green fruit	‘Picual’	454	273,519	305	253,023	435	92.51%
10	MID2	Seeds green fruit	‘Arbequina’ x ‘Picual’	454	411,421	290	375,851	406	91.35%
11	MID3	Mesocarp turning fruit	Arbequina’	454	292,02	336	256,35	498	87.79%
12	MID4	Mesocarp turning fruit	‘Picual’	454	247,142	325	230,085	482	93.10%

Table 5. Characterization of libraries

3.2 Tomato data collections.

Contains Gene expression and metabolite data gathered during the lifetime of the Spanish National project: Identification of Genes and Molecules Associated to Tomato Fruit Quality and Participation in the Sequencing of Euchromatic Regions of Chr9. A Genomic Approach (ESP-SOL).

During the ESP-SOL project 231 hybridizations were made using the technology of 70mers with the commercial TOM2 microarray (http://ted.bti.cornell.edu/cgi-bin/TFGD/order/clone_info.cgi)
.
Microarrays were distributed according to the hybridizations in:

- 36 arrays hybridized with samples of plants with extreme values in acidity.
- 30 arrays hybridized with samples with extreme values for organiz sugars (6 for each type: glucose, fructose, sucrose, malic acid, citric acid).
- 84 arrays hybridized with samples with extreme values of brix degrees (%of sucrose in volume).
- 39 arrays hybridized with samples with high/low levels in carotenoids.
- 39 arrays hybridized with extreme lines of vitamin C content.
- 3 arrays hybridized with samples of one ancestor (pimpinellifolium).

In total, more than 15 millions of expression data were analyzed in the ESP-SOL project and are available for Mr.SBM. Figure 2 shows the different comparisons that were performed in the project.

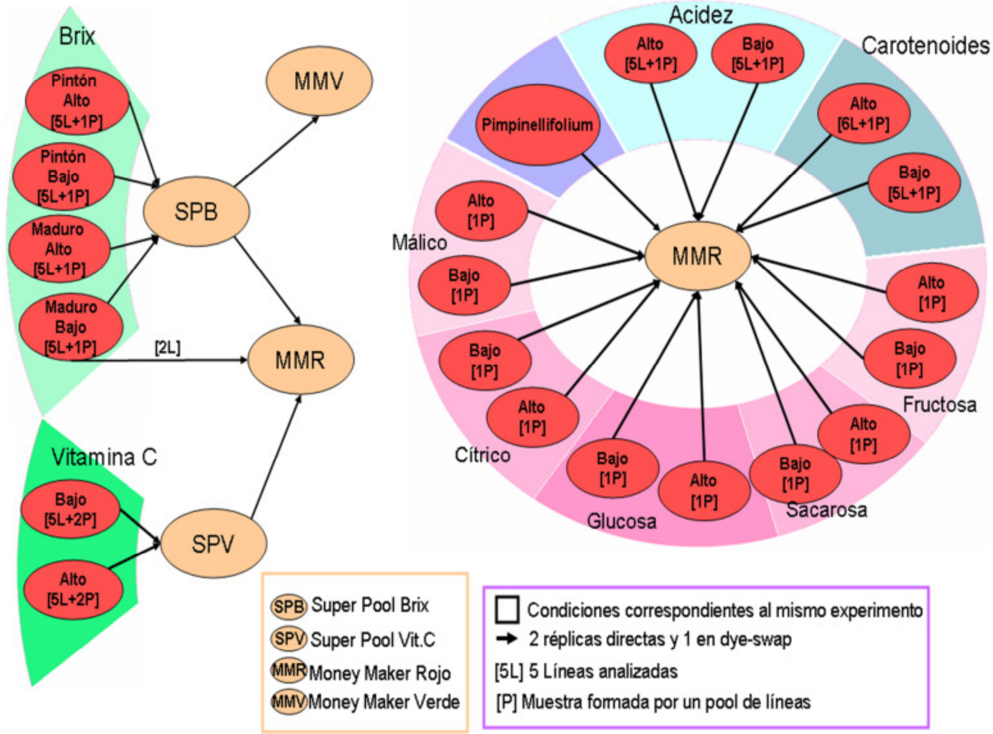


Figure 2. Hybridization schema of the ESP-SOL project. Balls in beige were the samples taken as reference in each experiment. In red balls the number of hybridizations is shown. The number of lines (L) and pools (P) used appear in brackets.

During the ESP-SOL project several measures from agronomical traits of interest were also taken, and used to select the lines to hybridize in microarrays. Appart from this, a set of volatile substances were analysed by means of an electronic nose that was able to detect levels 1000 times less than the human nose could recognize. This set of substances was composed by:

- 2-hexenal
- 3-methylbutanol
- 6-methyl-5-hepten-2-one
- methylsalicilate

This data set is an alternative to tests the phenotype-genotype correlations

3.3 Strawberry data collections.

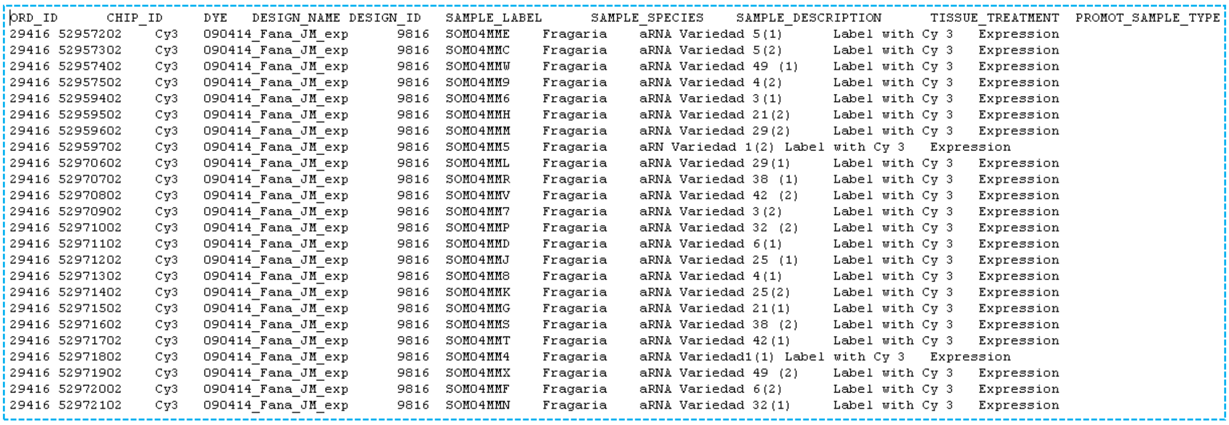
Contains Gene expression and metabolite data gathered during the lifetime of the Spanish National project: “Genetical genomics for improving strawberry fruits nutritional quality” (FraGENOMICs); in which the University of Malaga was the bioinformatics group in charge of data processing.

Relationship between expression data and metabolic data in strawberry varieties

The experiment was designed over 12 different wild strawberry varieties, for which different measures have been performed in both, expression and metabolic levels. Microarrays were generated by Nimblegen, and the raw data are available at the Call_Files.rar file.

This file contains all the single outputs for the microarrays values and a general file 090414_Fana_JM_exp.calls with all the values together by merging the individual ones.

The file SampleKey.txt attached to the raw data has the association between the array ID and the variety that has been hybridized in, see Figure 3. Each variety has two replicates that have been hybridized in a different array, so the experiment is composed of two values for each variety.



```
ORD_ID      CHIP_ID     DYE   DESIGN_NAME DESIGN_ID   SAMPLE_LABEL      SAMPLE_SPECIES     SAMPLE_DESCRIPTION      TISSUE_TREATMENT  PROMOT_SAMPLE_TYPE
29416 52957202    Cy3    090414_Fana_JM_exp      9816  SOM04MME     Fragaria     aRNA Variedad 5(1)      Label with Cy 3   Expression
29416 52957302    Cy3    090414_Fana_JM_exp      9816  SOM04MMC     Fragaria     aRNA Variedad 5(2)      Label with Cy 3   Expression
29416 52957402    Cy3    090414_Fana_JM_exp      9816  SOM04MMW     Fragaria     aRNA Variedad 49 (1)    Label with Cy 3   Expression
29416 52957502    Cy3    090414_Fana_JM_exp      9816  SOM04MM9     Fragaria     aRNA Variedad 4(2)      Label with Cy 3   Expression
29416 52959402    Cy3    090414_Fana_JM_exp      9816  SOM04MM6     Fragaria     aRNA Variedad 3(1)      Label with Cy 3   Expression
29416 52959502    Cy3    090414_Fana_JM_exp      9816  SOM04MMH     Fragaria     aRNA Variedad 21(2)     Label with Cy 3   Expression
29416 52959602    Cy3    090414_Fana_JM_exp      9816  SOM04MMM     Fragaria     aRNA Variedad 29(2)     Label with Cy 3   Expression
29416 52959702    Cy3    090414_Fana_JM_exp      9816  SOM04MM5     Fragaria     aRN Variedad 1(2) Label with Cy 3   Expression
29416 52970602    Cy3    090414_Fana_JM_exp      9816  SOM04MML     Fragaria     aRNA Variedad 29(1)     Label with Cy 3   Expression
29416 52970702    Cy3    090414_Fana_JM_exp      9816  SOM04MMR     Fragaria     aRNA Variedad 38 (1)    Label with Cy 3   Expression
29416 52970802    Cy3    090414_Fana_JM_exp      9816  SOM04MMV     Fragaria     aRNA Variedad 42 (2)    Label with Cy 3   Expression
29416 52970902    Cy3    090414_Fana_JM_exp      9816  SOM04MM7     Fragaria     aRNA Variedad 3(2)      Label with Cy 3   Expression
29416 52971002    Cy3    090414_Fana_JM_exp      9816  SOM04MMP     Fragaria     aRNA Variedad 32 (2)    Label with Cy 3   Expression
29416 52971102    Cy3    090414_Fana_JM_exp      9816  SOM04MMD     Fragaria     aRNA Variedad 6(1)      Label with Cy 3   Expression
29416 52971202    Cy3    090414_Fana_JM_exp      9816  SOM04MMJ     Fragaria     aRNA Variedad 25 (1)    Label with Cy 3   Expression
29416 52971302    Cy3    090414_Fana_JM_exp      9816  SOM04MM8     Fragaria     aRNA Variedad 4(1)      Label with Cy 3   Expression
29416 52971402    Cy3    090414_Fana_JM_exp      9816  SOM04MMK     Fragaria     aRNA Variedad 25(2)     Label with Cy 3   Expression
29416 52971502    Cy3    090414_Fana_JM_exp      9816  SOM04MMG     Fragaria     aRNA Variedad 21(1)     Label with Cy 3   Expression
29416 52971602    Cy3    090414_Fana_JM_exp      9816  SOM04MMS     Fragaria     aRNA Variedad 38 (2)    Label with Cy 3   Expression
29416 52971702    Cy3    090414_Fana_JM_exp      9816  SOM04MMT     Fragaria     aRNA Variedad 42(1)     Label with Cy 3   Expression
29416 52971802    Cy3    090414_Fana_JM_exp      9816  SOM04MM4     Fragaria     aRNA Variedad1(1) Label with Cy 3   Expression
29416 52971902    Cy3    090414_Fana_JM_exp      9816  SOM04MMX     Fragaria     aRNA Variedad 49 (2)    Label with Cy 3   Expression
29416 52972002    Cy3    090414_Fana_JM_exp      9816  SOM04MMF     Fragaria     aRNA Variedad 6(2)      Label with Cy 3   Expression
29416 52972102    Cy3    090414_Fana_JM_exp      9816  SOM04MMN     Fragaria     aRNA Variedad 32(1)     Label with Cy 3   Expression
```

Figure 3. Association between Array ID and the variety

In this file, the interesting columns are denoted as CHIP_ID, where the microarray identifier appears, and the SAMPLE_DESCRIPTION, where is shown the number of the variety and in brackets the number of the replicate.

The excel file expression-varieties.xlsx contains the information related to the normalized data coming from the expression values. This normalization was performed by Nimblegen using RMA method.

In excel file the columns from D to AA has the original normalized data for individual replicates. Then an average for each gene has been calculated in column AD and the ratios for each replicate respect to the average are shown in columns AF to BC.

Columns from BE to BP contains values for each variety after combining the two replicates for each one.

In these columns, BE to BP, it also appears the metabolic values for different metabolites along the different varieties. These values are shown from row 25132 to 25151.

3.4 LNCC data collections.

These data collections correspond to assembled sequence data gathered in the LNCC project that will be delivered to the consortium for testing procedures. The data belongs to different sources of information from simple bacterial genomes, higher species genomes as fungus or woody plants, till systems of high complexity as metagenomes in which the sequencing is performed over a mix of genomes that belong to the same environment as could be a river, a soil sample or the microbiome present in gut.

The samples used are:

- 29 mycoplasma genomes belonging to different species or strains that produce different symptoms in human diseases (the average size is 1 Mb).
- Two fungus genomes corresponding to the already sequenced specie Sporothrix schenckii and the new one to be compared Sporothrix brasiliensis. (Average size 30Mb). 2 Fecal microbiome of Human from distal gut of healthy adults (20 Mb).

4. Public molecular datasets

A catalogue of public databases will be needed in the project. A decision making process is needed to choose between installing local version (data warehouse), or to provide remote access to them. Among the different source of data we mention the following (not limited to):

Note: at the end of the document it has been included a detailed list of publicly available datasets

UNIPROT is likely the best quality and complete protein database [Consortium 2007]. A human-based curation process ensures high quality level, in particular in the Swiss-Prot section, which is manually annotated, in contrast to the automatic annotation of the second important section, TrEMBL.
Swissprot: (quoted) UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB). It is a high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. This database can be downloaded in plain text from the EBI ftp site (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/) in his traditional format .dat compressed as .dat.gz. The info contained in this plain text includes references to several cross databases. Additional information about the database composition and the way to download the data can be found in this readme file (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/README). This database can be downloaded also in FASTA format to perform sequence-sequence comparisons by mean of BLAST (for example) where it is necessary to give a specific format to the database using the script present in Blast installation (*formatdb* in older versions and *makeblastdb* in newer versions).

GenBank / **EMBL.** The former is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (***Nucleic Acids Research***, 2013 Jan;41(D1):D36-42); and it comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI.
The different versions of the GenBank can be downloaded from: ftp://ftp.ncbi.nih.gov/blast/db/.

Gene Ontology (GO): The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. The terms are classified in a hierarchical distribution with dependencies between the different levels. There are three categories to classify the different terms (Biological process, Cellular component and molecular function).

The automatic annotator SMA3s developed in UMA uses as a main source of data, the UniProt database in plain-text format (*file*.dat) and specifically the taxonomic division to which the organism under study belongs. The UniProt fields used by SMA3s are:

- *Gene Ontology.* Gene Ontology [Ashburner 2000] provides a controlled vocabulary to describe gene and gene product attributes which are organized in three ontologies of biological terms: molecular function, biological process and cellular component. Standardized annotations of GO terms are described in the cross-reference (DR) field of UniProt.

- *InterPro.* InterPro [Mulder 2007] is an integrated documentation resource for protein families, domains and sites. It combines a number of different databases with complementary information about sequence patterns. UniProt provides InterPro identifiers also from the DR field.

- *Swiss-Prot Keywords.* The Swiss-Prot keywords constitute a well-defined and controlled vocabulary of terms used to annotate a UniProt protein entry. These keywords describe functions, biological processes, structure, cellular localization and other protein characteristics, and are included in KW field in UniProt.

- *Pathway annotation.* This annotation provides a description of the metabolic pathway(s) in which a protein is involved. It is obtained from the comment (CC) field and is formed by a list of descriptors that illustrate —from generic to specific— the metabolic pathway (e.g.: PATHWAY: Nucleotide metabolism; purine metabolism). Sma3s gathers the annotations regarding the most generic level (e.g. “Nucleotide metabolism” in the previous example). This annotation type is very useful because co-expressed genes working in the same metabolic pathway might be expected.

Additional to the databases used for functional annotation we will use (remotely) the following databases to extract information for the Use Case on Comparative Genomics

5. Genomes and metagenomes

Additionally to the public datasets being freely available, during the project the computational infrastructure will be populated with the most frequently used datasets; which includes –but is not limited to:

- Higher mammalian genomes: *Homo sapiens*; Pan troglodytes, Macaca mulatta, Canis familiaris, Mus musculus, Rattus norvegicus, and Bos Taurus (Human, chimpanzee, macaca, dog, rat, mouse, cow, respectively).
- Mycoplasma genomes collections. This data collection is composed by more than 25 genomes corresponding to mycoplasma organisms
- Soil microbial communities from switchgrass rhizosphere

- metagenomes corresponding to water samples from Rio and Bahia (Brasil)

6. References

[Demoly 2004] Demoly P, Hillarie-Buys D. Classification and epidemiology of hypersensitivity drug reactions. Immunol Allergy Clin N Am. 2004 Aug;24:345-6.

[Johansson 2004] Johansson SG, et al. Revised nomenclature for allergy for global use: report of the nomenclature review committee of the World Allergy Organization, October 2003. J Allergy Clin Immunol. 2004 May;113(5):832-6.

[Roberts 2001] Roberts LJ, Morrow JD. “Analgesic-antipyretic and antiinflammatory agents and drugs employed in the treatment of gout”. In: Hardman JG, Limbird LL, Goodman Gilman A, eds. The pharmacological basis of therapeutics, 10th edn. New York: McGraw-Hill, 2001; 687–731.

[Szczeklik 2009] Szczeklik A, Nizankowska E, Sanak M., "Hypersensitivity to aspirin and non-steroidal antiinflammtory drugs". In: Adkinson NF, Bochner BS, Busse WW, Holgate S, Lemanske RF, Simons FE, eds. Middelton’s allergy, principles and practice, 7th edn. Philadelphia: Mosby, 2009; 1227–43.

[Gomes 2005] Gomes ER, Demoly P. Epidemiology of hypersensitivity drug reactions.Curr Opin Allergy ClinImmunol. 2005 Aug;5(4):309-16.

[Sanchez-Borgez 2010] Sanchez-Borges M. NSAID Hypersensitivity (Respiratory, Cutaneous, and Generalized Anaphylactic Symptoms). Med Clin N Am. 2010 Jul;94(4): 853–64.

[Gruchalla 2003] Gruchalla RS. Drug allergy. J Allergy ClinImmunol. 2003 Feb;111(2 Suppl) 2:S548-59.

[Thong 2011] Thong BY & Tan TC. Epidemiology and risk factors for drug allergy. Br J Clin Pharmacol. 2011 May;71(5):684-700.

[Demoly 2002] Demoly P, Bousquet J. Drug allergy diagnosis work up. Allergy. 2002;57 Suppl 72:S37-40.

[Adkinson 2002] Adkinson NF Jr, et al.; Health and Environmental Sciences Institute Task Force. Task force report: future research needs for the prevention and management of immune-mediated drug hypersensitivity reactions. J Allergy ClinImmunol. 2002 Mar;109 Suppl (3):S461-78.

[Mockenhaupt 2007] Mockenhaupt M. Epidemiology and causes of severe cutaneous adverse reactions to drugs. In: Pichler WJ, editor. Drug Hypersensitivity. Basel: Karger; 2007. p. 18-31.

[Gamboa 2009] Gamboa PM. The epidemiology of drug allergyrelated consultations in Spanish Allergology Services: Alergológica-2005. J Investig Allergol Clin Immunol. 2009; 19 Suppl 2:45-50.

[Romano 2012] Romano A, Demoly P. Recent advances in the diagnosis of drug allergy. Curr Opin Allergy Clin Immunol. 2007 Aug;7(4):299-303. 369 J Investig Allergol Clin Immunol 2012; Vol. 22(5): 363-371

[Messaad 2004] Messaad D, et al.; Drug provocation tests in patients with a history suggesting an immediate drug hypersensitivity reaction. Ann Intern Med. 2004 Jun 15;140(12):1001-6.

[Blanca 1999] Blanca M, et al. Natural evolution of skin test sensitivity in patients allergic to beta-lactam antibiotics. J Allergy Clin Immunol. 1999 May;103(5 Pt 1):918-24.

[Aberer 2003] Aberer W, et al.; European Network for Drug Allergy (ENDA); EAACI interest group on drug hypersensitivity. Drug provocation testing in the diagnosis of drug hypersensitivity reactions: general considerations. Allergy. 2003 Sep;58(9):854-63.

[Blanca 2009] Blanca M, et al.; Update on the evaluation of hypersensitivity reactions to betalactams. Allergy. 2009 Feb;64(2):183-93.

[Abuaf 1999] Abuaf N, et al. Validation of flow cytometry assay detecting in vitro basophil activation for the diagnosis of muscle relaxant allergy. J Allergy Clin Immunol. 1999 Aug;104 (2 Pt 1):411-8.

[Sanz] Sanz ML, et al.; Flow cytometric basophil activation test by detection of CD63 expression in patients with immediate-type reactions to betalactam antibiotics. Clin Exp Allergy. 2002 Feb;32(2):277-86.

[Torres 2004] Torres MJ, et al. “The diagnostic interpretation of basophil activation test in immediate allergic reactions to betalactams”. Clin Exp Allergy. 2004 Nov;34(11):1768-75.

[Ebo 2006] Ebo DG, et al. “Flow-assisted diagnostic management of anaphylaxis from rocuronium bromide”. Allergy. 2006 Aug;61(8):935-9.

[Aranda 2011] Aranda A, Mayorga C, Ariza A, Doña I, Rosado A, Blanca-Lopez N, Andreu I, Torres MJ. In vitro evaluation of IgE-mediated hypersensitivity reactions to quinolones. Allergy. 2011 Feb;66(2):247-54.

[Luque 2001] Luque I, Leyva L, Torres MJ, Rosal M, Mayorga C, Segura JM, Blanca M, Juárez C. In vitro T-cell responses to b-lactam drugs in immediate and nonimmediate allergic reactions. Allergy. 2001 Jul;56(7):611-8.

[Nyfeler 1997] Nyfeler B, Pichler WJ. The lymphocyte transformation test for the diagnosis of drug allergy: sensitivity and specificity. Clin Exp Allergy. 1997 Feb;27(2):175-81.

[Bousquet 2009] Bousquet PJ, et al.; Global Allergy, Asthma European Network (GALEN) and Drug Allergy and Hypersensitivity Database (DAHD) and the European Network for Drug Allergy (ENDA). Pharmacovigilance of drug allergy and hypersensitivity using the ENDA-DAHD database and the GALEN platform. The Galenda project. Allergy. 2009 Feb;64(2):194-203.

[Torres] Torres MJ, et al.; ENDA; EAACI Interest Group on Drug Hypersensitivity. Diagnosis of immediate allergic reactions to beta-lactam antibiotics. Allergy. 2003 Oct;58(10):961-72.

[England 2003] England RW, et al. Inpatient consultation of allergy/immunology in a tertiary care setting. Ann Allergy Asthma Immunol. 2003 Apr; 90(4):393-7.

[Dietrich 2009] Dietrich JJ, et al. Reasons for outpatient consultation in allergy/immunology. Allergy Asthma Proc. 2009 Jan-Feb; 30(1):69-74.

[Gomes 2005] Gomes ER, Demoly P. Epidemiology of hypersensitivity drug reactions. Curr Opin Allergy Clin Immunol. 2005 Aug;5(4):309-16.

[Blanca 1995] Blanca M. Allergic reactions to penicillins. A changing world? Allergy. 1995 Oct;50(10):777-82.

[Doña 2011] Doña I, Blanca-López N, Cornejo-García JA, Torres MJ, Laguna JJ, Fernández J, Rosado A, Rondón C, Campo P, Agúndez JA, Blanca M, Canto G. Characteristics of subjects experiencing hypersensitivity to non-steroidal anti-inflammatory drugs: patterns of response. Clin Exp Allergy 2011;41(1):86-95.

[Kowalski 2011] Kowalski ML, Makowska JS, Blanca M, Bavbek S, Bochenek G, et al. (2011) Hypersensitivity to nonsteroidal anti-inflammatory drugs (NSAIDs) - classification, diagnosis and management: review of the EAACI/ENDA(#) and GA2LEN/HANNA*. Allergy 66: 818-829.

[Gómez 2009] Gómez E, Blanca-Lopez N, Torres MJ, Requena G, Rondon C, Canto G, Blanca M, Mayorga C. Immunoglobulin E-mediated immediate allergic reactions to dipyrone: value of basophil activation test in the identification of patients. Clin Exp Allergy 2009;39(8):1217-24.

[Blanca 1989] Blanca M, Perez E, Garcia JJ, Miranda A, Terrados S, Vega JM, Suau R. Angioedema and IgE antibodies to aspirin: a case report. Ann Allergy 1989;62(4):295-8.

[Levi 2009] Levi N, Bastuji-Garin S, Mockenhaupt M, Roujeau JC, Flahault A, Kelly JP, Martin E, Kaufman DW, Maison P. Medications as risk factors of

Stevens-Johnson syndrome and toxic epidermal necrolysis in children: a pooled analysis. Pediatrics 2009;123(2):e297-304.

[Mockenhaupt 2003] Mockenhaupt M, Kelly JP, Kaufman D, Stern RS; SCAR Study Group. The risk of Stevens-Johnson syndrome and toxic epidermal necrolysis associated with nonsteroidal antiinflammatory drugs: a multinational perspective. J Rheumatol 2003;30(10):2234-40.

[Roujeau 1995] Roujeau JC, Kelly JP, Naldi L, Rzany B, Stern RS, Anderson T, Auquier A, Bastuji-Garin S, Correia O, Locati F, et al. Medication use and the risk of Stevens-Johnson syndrome or toxic epidermal necrolysis. N Engl J Med 1995;333(24):1600-7.

[Canto 2009] Canto MG, Andreu I, Fernandez J, Blanca M. Selective immediate hypersensitivity reactions to NSAIDs. Curr Opin Allergy Clin Immunol 2009;9(4):293-7.

[Stevenson 2006] Stevenson DD, Szczeklik A. Clinical and pathologic perspectives on aspirin sensitivity and asthma. J Allergy Clin Immunol 2006;118(4):773-86.

[Minai-Fleminger 2009] Minai-Fleminger Y, Levi-Schaffer F. Mast cells and eosinophils: the two key effector cells in allergic inflammation. Inflamm Res 2009;58(10):631-8.

[Blanca 2012] Blanca M, Thong B. Progress in understanding hypersensitivity drug reactions: an overview. Curr Opin Allergy Clin Immunol 2012;12(4):337-40.

[Stevenson 2001] Stevenson DD, Sanchez-Borges M, Szczeklik A. Classification of allergic and pseudoallergic reactions to drugs that inhibit cyclooxygenase enzymes. Ann Allergy Asthma Immunol 2001;87(3):177-80.

[Szczeklik 2003] Szczeklik A, Stevenson DD. Aspirin-induced asthma: advances in pathogenesis, diagnosis, and management. J Allergy Clin Immunol 2003;111(5):913-21.

[Clevert 2011] Clevert DA, et al. cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate, Nucl. Acids Res. (2011) 39 (12).

[Affymetrix] Genome-Wide Human SNP Array 6.0, Affymetrix, http://www.affymetrix.com/browse/products.jsp?productId=131533&navMode=34000&navAction=jump&aId=productsNav#1_3)

[Consortium 2007] Consortium TU: The Universal Protein Resource (UniProt). Nucleic acids research 2007, 35(Database issue):D193-197

[Ashburner 2000]	Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 2000, 25(1):25-29
[Mulder 2007]	Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R et al: New developments in the InterPro database. Nucleic acids research 2007, 35(Database issue):D224-228.

7. Appendix I

Databases of importance for Mr.SymBioMath

Name	Description	More information	License / Price	Priority (H/M/L)
	Medical Databases			
BIC	Breast cancer Information Core (public database)	http://research.nhgri.nih.gov/bic/	Free for academic	High
OMIM	Online Mendelian Inheritance in Man, a database of human genes and genetic disorders	http://www.ncbi.nlm.nih.gov/omim/	Free for research	High
	Gene Expression Databases			
BASE	A LIMS system (ref: Lao H. Saal, Carl Troein, Johan Vallon-Christersson, Sofia Gruvberger, Åke Borg and Carsten Peterson BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data)	http://base.thep.lu.se/	Free	High
ArrayExpress	MIAME compliant repository of published microarray datasets (EBI)	http://www.ebi.ac.uk/arrayexpress/ http://www.ebi.ac.uk/miamexpress	OS	High
GEO -Gene Expression Omnibus.	A gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval. (Microarray database with good search engine and export of data)	http://www.ncbi.nlm.nih.gov/geo/	OS	High
Oncomine	Microarray Database Oncomine (includes some tools for profiling)	www.oncomine.org		

OS	Medium			
CleanEx	Comparative database of published microarray datasets.	http://www.cleanex.isb-sib.ch/	Free	Low
SMD	Stanford Microarray Database (Database of Stanford arrays with export of data and some tools for filtering and first analysis)	http://genome-www5.stanford.edu/	OS	Medium
	Nucleotide Sequence Databases			
EMBL Bank	European Nucleotide Sequence Database	www.ebi.ac.uk/embl	Free	High
EnsEMBL	Integrated nucleotide sequence knowledge base	http://www.ensembl.org/	Free	High
GenBank	American Nucleotide Sequence Database	www.ncbi.nih.gov/ www.ncbi.nlm.nih.gov/Genbank/	Free	High
UniGene	Database of clusters of GenBank sequences.	http://www.ncbi.nih.gov/	Free	High
DDBJ	DNA Data Bank of Japan	http://sakura.ddbj.nig.ac.jp/		Low
	Protein Databases			
Swissprot	Protein knowledge base	http://www.expasy.org/sprot/	Free	High
UniProt	Universal protein resource: Consolided DB from : Swissprot + TrEMBL + REMTrEBML + PIR).	http://www.uniprot.org	free to copy, distribute, ...	High
PIR	Protein information resource	http://pir.georgetown.edu/	Free	Medium
	3D Structure databases			
PDB	Protein data bank	http://www.rcsb.org/pdb/	Free	High
CATH	Protein structure classification. CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H).	http://cathwww.biochem.ucl.ac.uk/		Medium
PDBsum	Putative protein-protein binding sites, ligand binding sites, and protein-DNA binding sites by homology with those observed in crystallized protein structures.	http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/	freely available	Medium

FSSP	Fold classification based on structure-structure assignments	http://www.ebi.ac.uk/dali		Medium
DSSP	secondary structure assignments for all PDB-protein entries (it is also a programm)			Medium
HSSP	DB of homology-derived secondary structure of proteins	http://swift.cmbi.kun.nl/gv/hssp/		Medium
IntAct	Protein interaction data	http://www.ebi.ac.uk/intact/	freely available	Medium
	Distance Matrices (Distance matrices are needed in all programs that perform sequence comparison)			
PAM	Point Accepted Mutation matrices (from PAM10 to PAM450)		free	High
BLOSUM	Block alignment derived substitution matrices (from Blosum 30 to Blosum90)		free	High
	Ontology Databases			
GO: Gene Ontology	Function, Biological process, and Cellular component	http://www.geneontology.org/	free	High
GOA	Gene Ontology Annotation @ EBI, provides association between GO terms and genes	http://www.ebi.ac.uk/GOA/	Open access	High
AMIGO	Gene Ontology database	http://www.godatabase.org/cgi-bin/amigo/go.cgi	OS	Medium
	Pathway Databases			
KEGG	Kyoto Encyclopedia of Genes and Genomes: pathways map	http://www.genome.ad.jp/kegg/	Licenses for non-academic users	High
EBI Databases repositorie	Public databases and tools	http://www.ebi.ac.uk/	Free	High

	Motif Databases			
Pfam.	Protein Families and domains database (includes multiple sequence alignments and HMM models	http://www.sanger.ac.uk/Software/Pfam/		Medium
Prosite, BLOCKs, PRODOM, PRINTS..	Sequence motifs databases; protein domains, etc	www.expasy.org/prosite/ http://protein.toulouse.inra.fr/prodom/current/html/home.php		Medium
	Scientific literature			
PubMed	Bibliographic references (including MeSH terms) . PubMed is a service of the U.S. NLM that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s.	http://www.ncbi.nlm.nih.gov/PubMed/	Public domain access	High
NCBI databases	Public databases and tools	http://www.ncbi.nlm.nih.gov/	OS	Medium
BEA	Biovista	http://www.biovista.com	Commercial	High
	DAS servers (data integration)			
Ensembl	system which produces and maintains automatic annotation on selected eukaryotic genomes	http://www.ensembl.org		
free access	High			
UniProt	Protein features annotations	http://www.ebi.ac.uk/uniprot-das/ (DAS server)	free access	High
KEGG	Pathway maps	http://www.genome.jp/kegg/soap/ (Java)	Licenses for non-academic users	High
GO: Gene Ontology	Function, Biological process, and Cellular component	http://www.geneontology.org/GO.tools.shtml (different annotation tools)	freely available to all the public	High
Pubmed	Bibliographic references (including MeSH terms)	http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html	Public domain	High

			information (NLM)	
CATH	Annotates PDB structures with CATH structural domains	http://www.biochem.ucl.ac.uk/bsm/cath/ (DAS)		Medium
PDBsum	Putative protein-protein binding sites, ligand binding sites, and protein-DNA binding sites by homology with those observed in crystallized protein structures.	http://www.ebi.ac.uk/das-srv/proteindas/das/sasprot/ (DAS)	freely available	Medium
Phenotypes	Phenotypes associated directly or via orthologues or protein families. Use the Ensembl,Gene_ID databases.	http://www.ebi.ac.uk/das-srv/genedas/das/phenotypes/ (DAS)		Medium
Catalytic Site Atlas (CAS)	Manually curated collection of catalytic sites (and predicted by homology) described in the literature	http://www.ebi.ac.uk/das-srv/proteindas/das/csalit/ and http://www.ebi.ac.uk/das-srv/proteindas/das/csaextended/ (DAS)		Medium
SMART	Domain annotations for Uniprot/Ensembl	http://smart.embl.de/smart/das/smart/ (DAS)	http://smart.embl-heidelberg.de/	Medium
BIND, DIP, MINT, PIM, etc.	Protein interactions	http://www.blueprint.org/bind/bind_relateddatabases.html		Medium