



JORNADAS ANDALUZAS
SALUD INVESTIGA
GRANADA
29 DE OCTUBRE · 2018

‘Transformando el Sistema Público de Salud mediante el uso de los datos genómicos del paciente: perspectivas y desafíos’

Joaquín Dopazo

Director Área de Bioinformática
Fundación Progreso y Salud

Functional Genomics Node, (INB-ELIXIR-es),
Bioinformatics in Rare Diseases (BiER-CIBERER)

Sevilla



@xdopazo, @ClinicalBioinfo

Promueven



Servicio Andaluz de Salud
CONSEJERÍA DE SALUD

Organiza



Fundación Progreso y Salud
CONSEJERÍA DE SALUD

Colaboran



Escuela Andaluza de Salud Pública
CONSEJERÍA DE SALUD



Biobanco del Sistema Sanitario Público de Andalucía
CONSEJERÍA DE SALUD



Biblioteca Virtual
del Sistema Sanitario Público de Andalucía



PEDEI-UNIVERSIDAD DE GRANADA-JUNTA DE ANDALUCÍA
CENTRE FOR GENOMICS AND ONCOLOGICAL RESEARCH



RED DE FUNDACIONES GESTORAS
de la **investigación** del SSPA

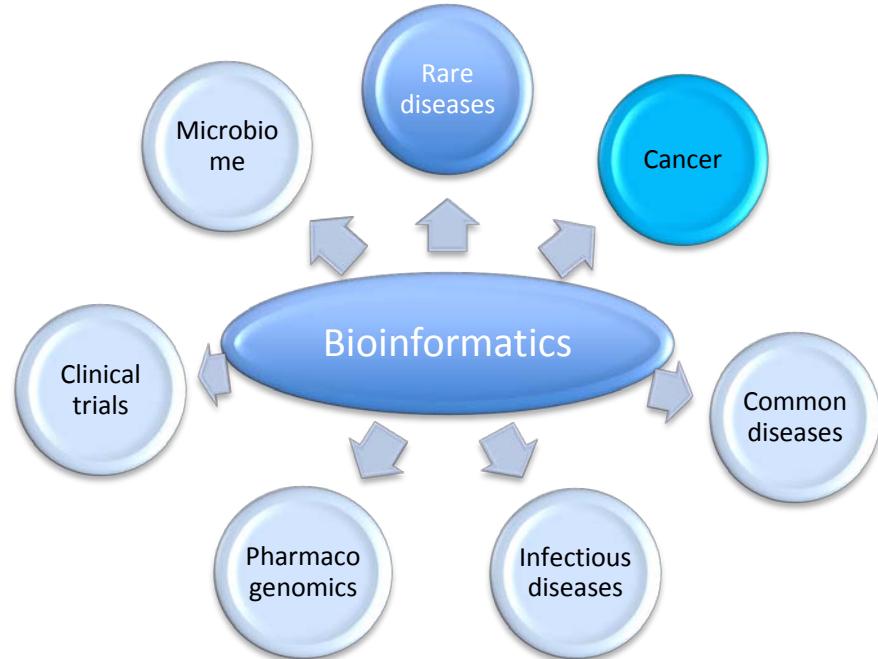
The clinical bioinformatics area

The Clinical Bioinformatics Area from the Fundación Progreso y Salud (FPS) has been conceived as a fundamental piece of the Personalized Medicine plan of the Andalusian community, with the mission of facilitating and providing the tools for the inclusion of the genomic data of the patient in the electronic health record.

This Area has the dual aim of developing innovative algorithms and methods for the analysis of genomic data of patients, combined with the production of high quality software specifically designed to be used by clinician end users, all this with a strong translational orientation. The ultimate objective of the Area is to bring to the clinician complex algorithms for the management of complex genomics data in a transparent way for them, which ultimately foster the adoption of innovative technologies in the current clinical practice.

Clinical Bioinformatics Area

Translational Bioinformatics Systems Medicine Computational Biology



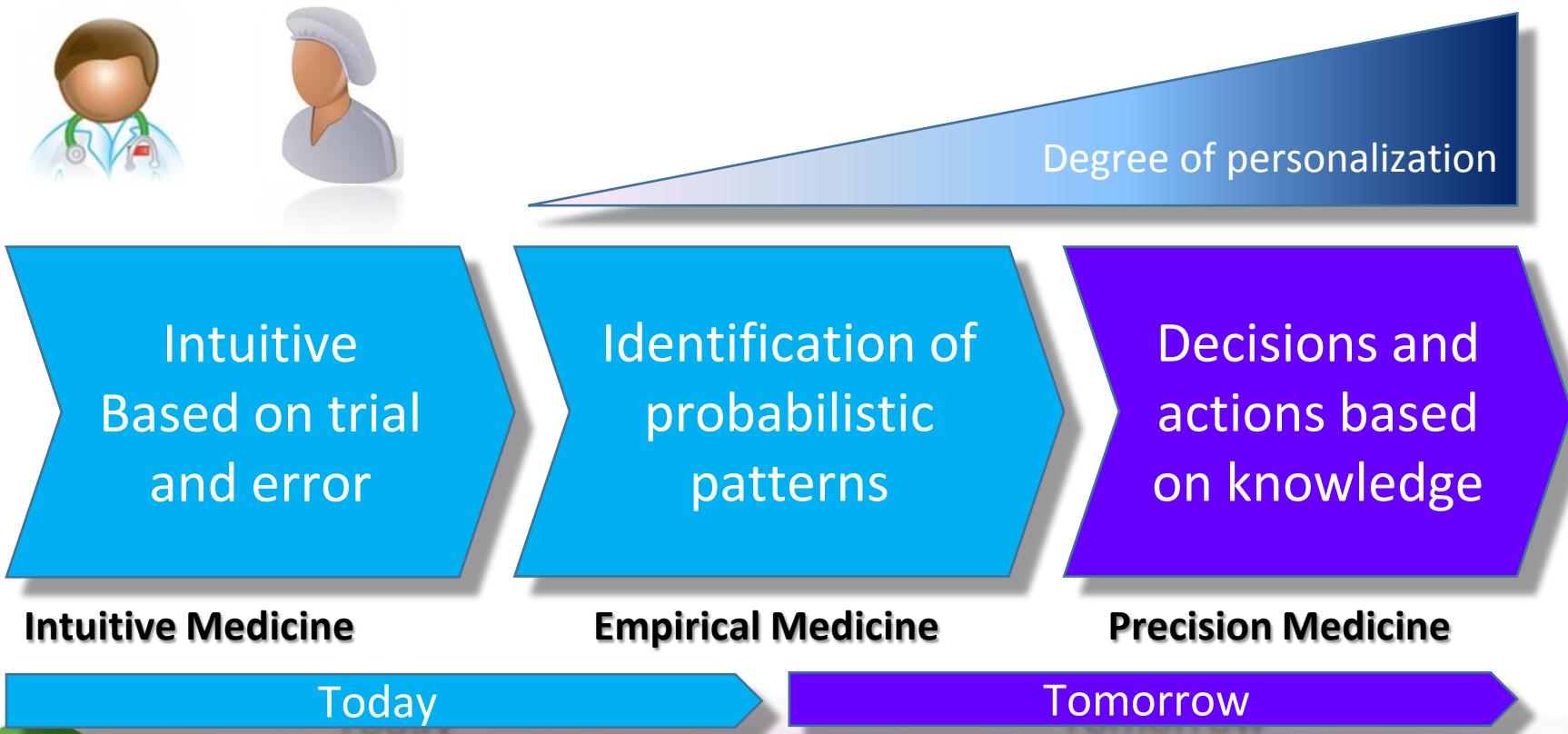
<http://www.clinbioinfosspa.es/>

The Bioinformatics Area, created in June 2016 in the Fundación Progreso y Salud, has as main goal supporting the Program of Personalized Medicine of the Andalusian Community by facilitating the use of genomic data for precision diagnostic and treatment recommendation, implementing a prospective health care functionality in the public health system .



JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

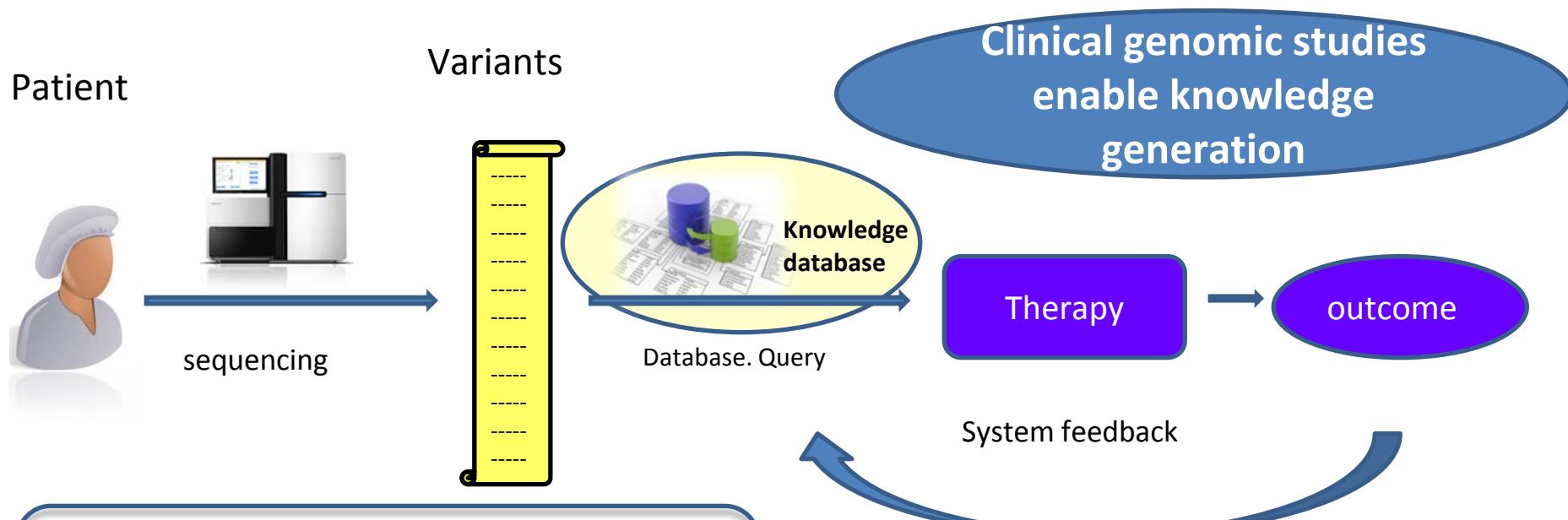
Personalized health care and the transition to precision medicine



JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

Empirical medicine

Phase I: generation of knowledge



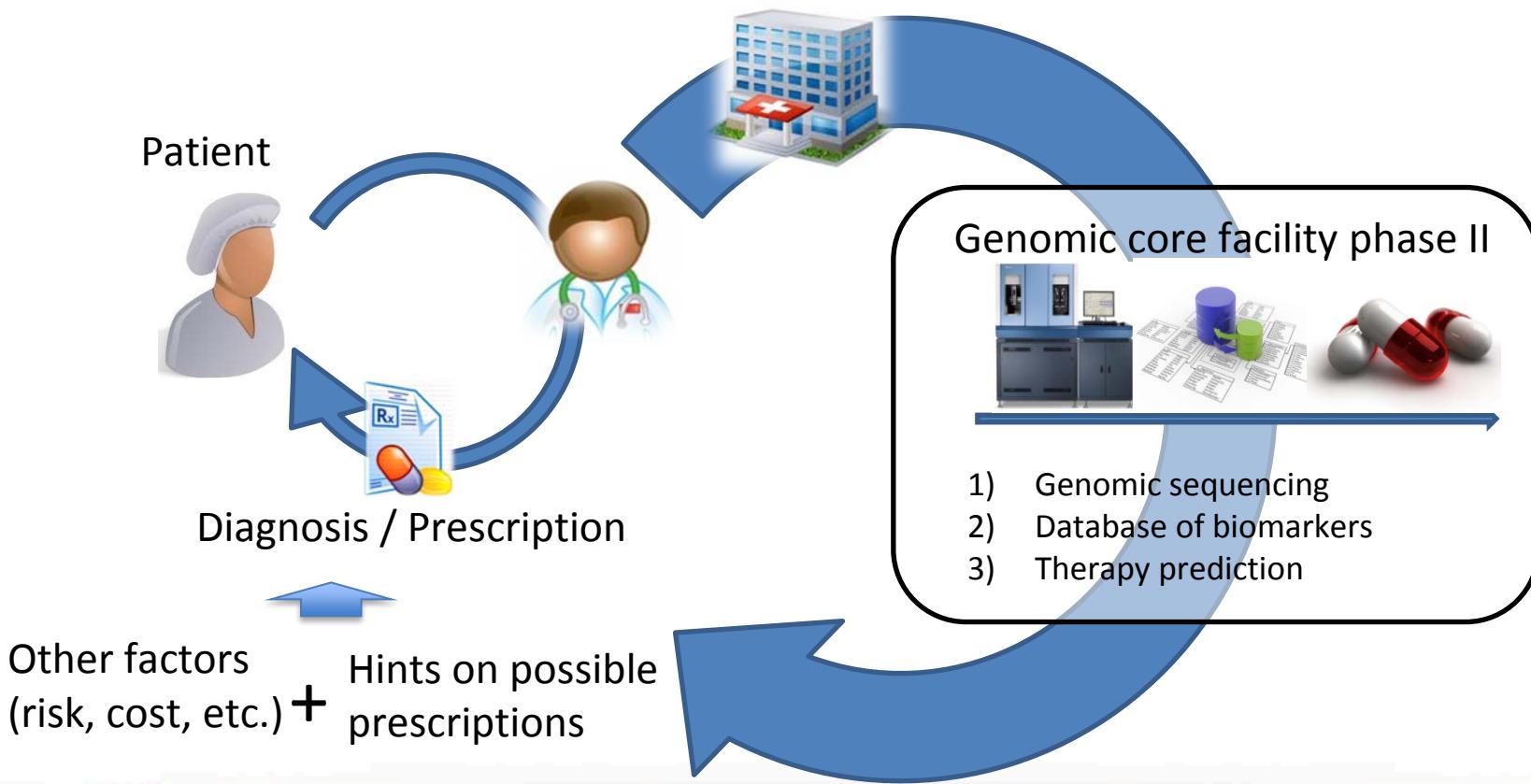
Genomic variants (biomarkers) can be quickly associated to precise diagnosis or therapy outcomes

Initially the system will need much feedback: **Knowledge generation phase.**



Precision medicine.

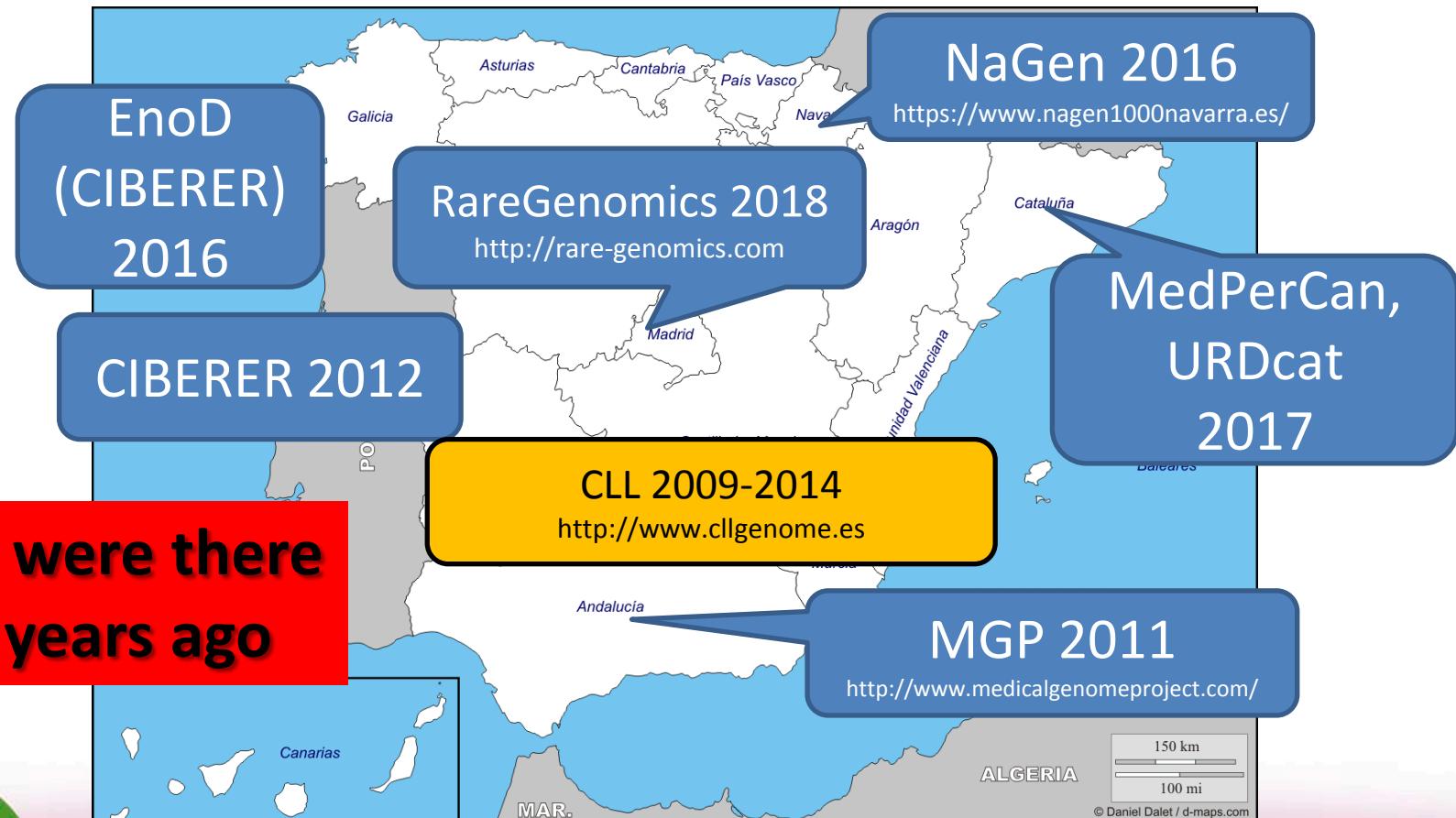
Phase II: using the knowledge database



13

JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

An historical perspective: International projects (ICGC) and the local genomic initiatives in Spain



13

JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

Back to 2011: the Medical Genome Project



Raw files
(FastQ)

GCGTATAG
CACGGTA
TCTGTATA
TGTTGGAAT
ATCAGCGG
GGCAGAGC
GCCAAAGT
GCGTATAG
CACGGTA
TCTGTATA
TGTTGGAAT
ATCAGCGG
GGCAGAGC
GCCAAAGT
GCGTATAG
CACGGTA
TCTGTATA
TGTTGGAAT
ATCAGCGG
GGCAGAGC
GCCAAAGT



Samples

Knowledge DB

Validated
knowledge

Analysis
Pipeline

DB

Storage

Bottleneck

Gene 1 ksdhkahckka
Gene 2 jkacaksdka
Gene 3 lkdkdcjccjdjc
Gene 4 ksfdflylvdskvjd
Gene 5 kckokksdkdsd
Gene 6 ldkdkcksdld

Gene x kcdlkckldskik
Gene Y jdksdkckdks

GCGTATAG
CACGGTA
TCTGTATA
TGTTGGAAT
ATCAGCGG

GCGTATAG
CACGGTA
TCTGTATA
TGTTGGAAT
ATCAGCGG

VCF

BAM
Processed files

Prioritization
report

Dialog with disease
experts + validations



JORNADAS ANDALUZAS SALUD INVESTIGA · GRANADA · 25 DE OCTUBRE · 2018

Current solutions for managing genomic data of patients



Bioinformatician and
biostatistician teams

- No scalability
- Expensive
- Lack of experts
- Inequity



Commercial software



JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

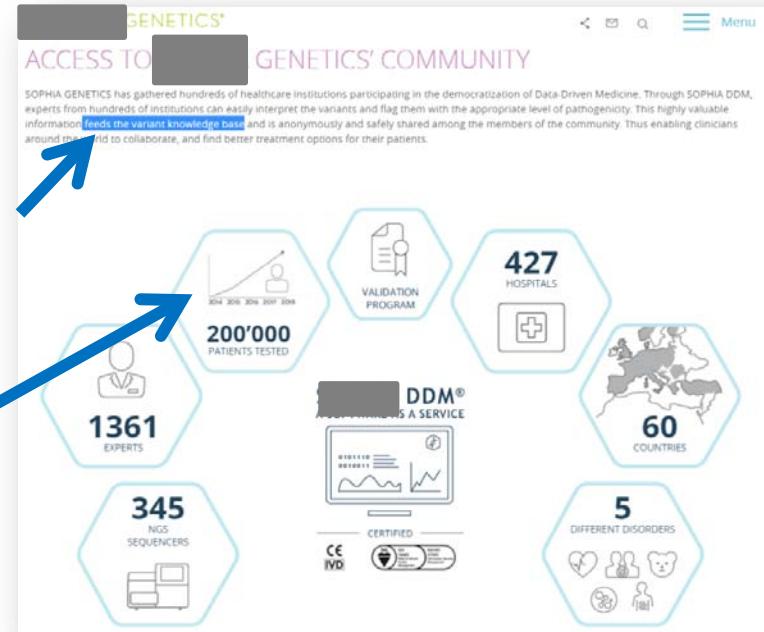
Using external software:

We generate data, pay and annotate

Companies get payed and let the database growth and be curated

"Experts from hundreds of institutions interpret variants and flag them with the appropriate level of pathogenicity ... feed the variant knowledge base"!!!!!!

"our database of 200,000 patients tested"!!!!



We are paying for the use of software of companies that collect for free genomic data and information generated by our experts (their customers), which in turn increase the value of the service offered by the companies.

Our genomic data are externalized with no value for the health system.



JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

Solutions for managing the genomic data of the patient



- Expensive (pay per use)
- GDPR non compliant
- Inequity

- No scalability
- Expensive
- Lack of experts
- Inequity

Our solution: corporative software

- **Equity**
- **Scalability**
- **Affordable**
- **End user: clinician**
- **GDPR compliant**

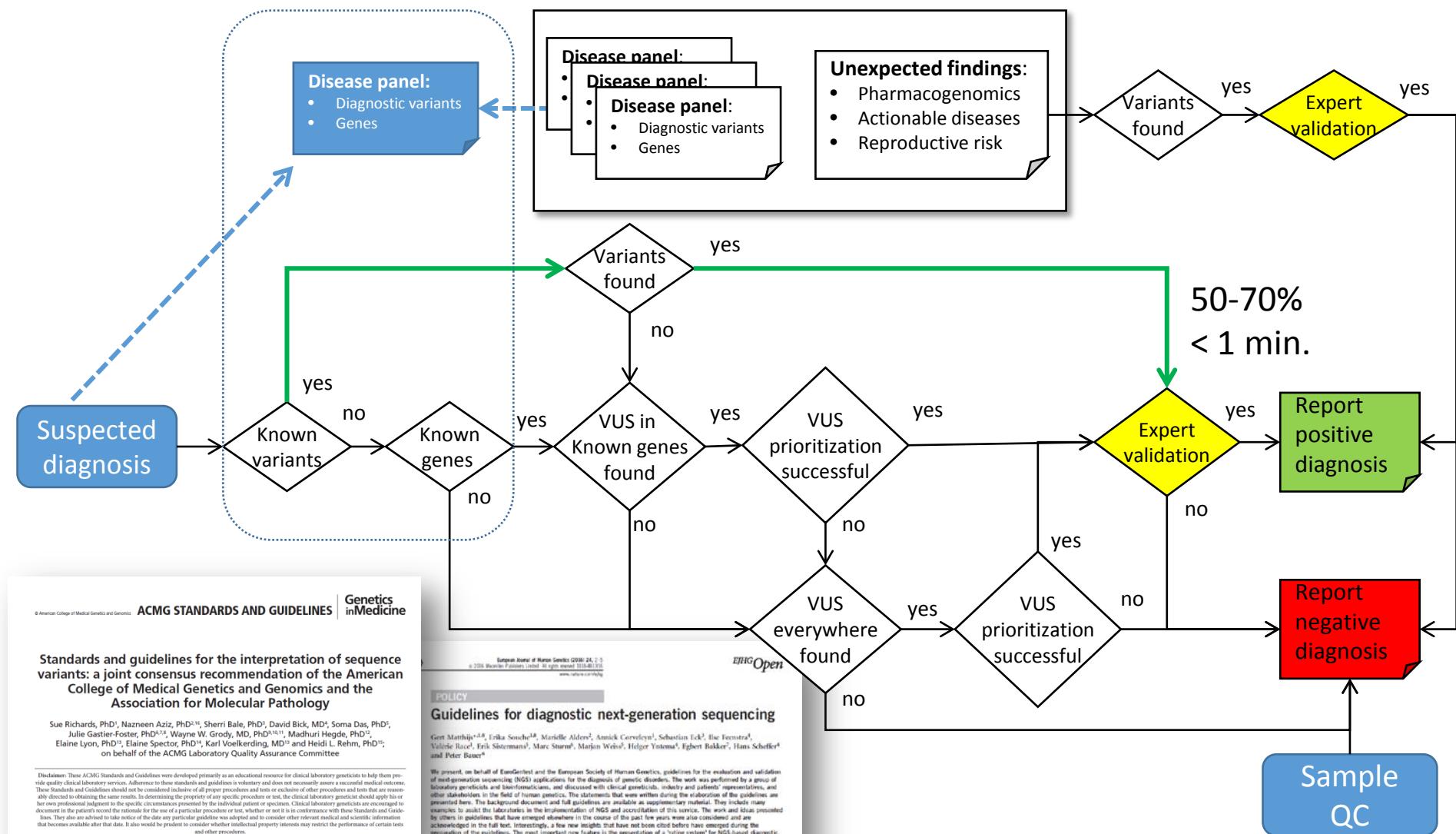
Use of genomic data in the public health system requires sustainability

User: clinician (not bioinformatician)

- Tools for end users, which involves **hiding the complexity** of the analysis to the clinician
- A solution for the management of genomic data must be **integrated** the same way other analyses of the health system are.
- **Genomic** data must be **stored** in the system, **linked** to **clinical** data the same way that other data are for further potential prospective clinical studies



Complexity of the general diagnosis protocol (rare diseases)



Our approach: hiding the complexity

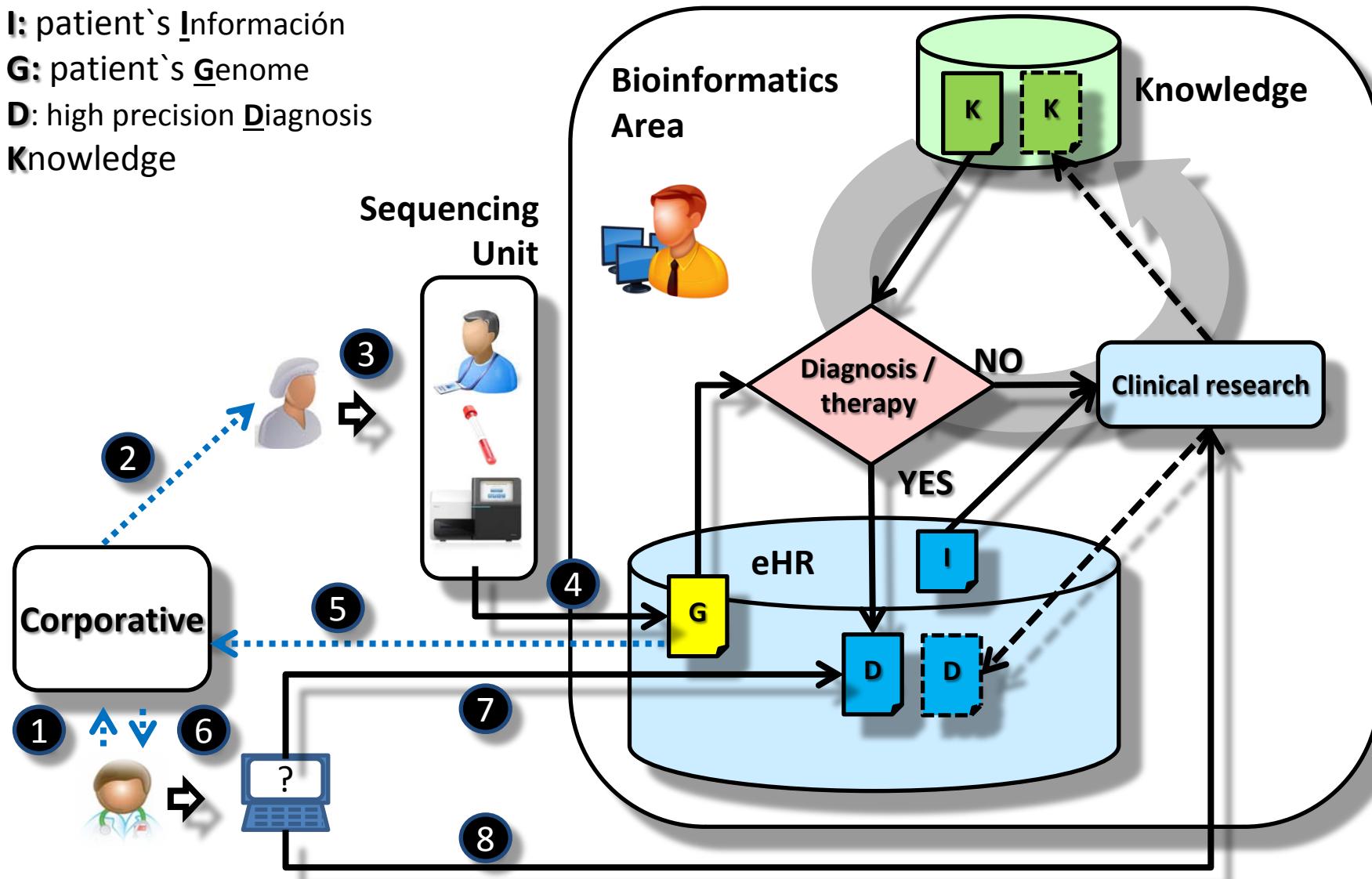
Decision support systems democratize the use of complex
(genomic) data

I: patient's Información

G: patient's Genome

D: high precision Diagnosis

Knowledge



Front end: Personalized Medicine Module (MMP)

This screenshot shows the 'Sample Results' section of the IVA-ACCI v0.9.0 Clinical Analysis interface. It displays a table of samples with columns for Sample name, Individual ID, Date, Status, Sex, Diagnosis, Father, Mother, and Cell Line. A large blue arrow points from this screen to the 'Variant prioritization' screen.

Sample selection

This screenshot shows the 'Prioritization' section of the IVA-ACCI v0.9.0 Clinical Analysis interface. It displays a table of variants with columns for Sample, Variant, Disease, and various prioritization scores. A large blue arrow points from the 'Variant prioritization' screen to the 'Selection of variants for the report' screen.

Variant prioritization

This screenshot shows the 'Report Generation' section of the IVA-ACCI v0.9.0 Clinical Analysis interface. It displays a table of variants with columns for Sample, Variant, Disease, and various scores. A large blue arrow points from the 'Report generation' screen to the 'Selection of variants for the report' screen.

Report generation
(sent to the eHR)

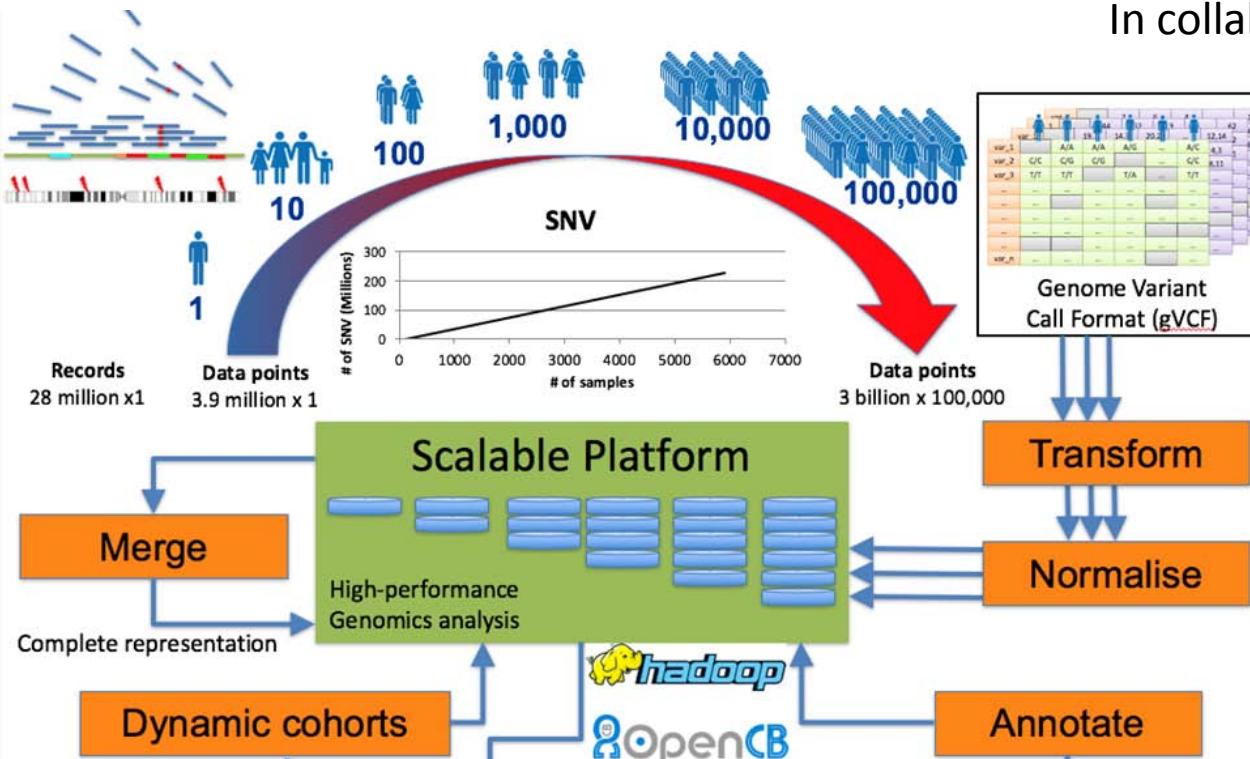
This screenshot shows the 'Interpretation Analysis' section of the IVA-ACCI v0.9.0 Clinical Analysis interface. It displays a summary of the analysis results, including variants and their scores, and a table of reported variants. A large blue arrow points from the 'Selection of variants for the report' screen to this final report screen.

Selection of
variants for the
report



JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

Backend: OpenCGA, a scalable storage and genomic data management platform



13

JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

We share the backend with the 100.000 genomes project

The screenshot shows the Genomics England website. At the top, there's a navigation bar with links for 'About Us', '100,000 Genomes Project', 'Taking Part', 'For Healthcare Professionals', 'Research', 'Industry Partnerships', and 'News & Events'. Below the navigation is a search bar with 'Google Custom Search' and a magnifying glass icon. The main content area has a teal header bar with the text 'Home > News > Posts > Genomics England uses MongoDB to power the data science behind the 100,000 Genomes Project'. The main article title is 'Genomics England uses MongoDB to power the data science behind the 100,000 Genomes Project'. Below the title is a timestamp 'Posted on April 10, 2018 at 10:35 am'. The text of the article discusses how Genomics England is using MongoDB to power the data science behind the 100,000 Genomes Project, mentioning a partnership with MongoDB that allows for faster processing times.

Genomics England uses MongoDB to power the data science behind the 100,000 Genomes Project

Posted on April 10, 2018 at 10:35 am

Genomics England is using data platform [MongoDB](#) to power the data science that makes the [100,000 Genomes Project](#) possible. Our partnership with MongoDB allows the processing time for complex queries to be reduced from hours to milliseconds, which means scientists can discover new insights more quickly.



Published online 12 June 2012

Nucleic Acids Research, 2012, Vol. 40, Web Server issue W609-W614
doi:10.1093/nar/gks575

CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources

Marta Bleda^{1,2}, Joaquín Tarraga^{1,3}, Alejandro de María¹, Francisco Salavert^{1,2}, Luz García-Alonso¹, Matilde Celma⁴, Ainhoa Martín⁴, Joaquín Dopazo^{1,2,3,*} and Ignacio Medina^{1,3,*}

¹Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), 46012 Valencia, Spain, ²CIBER de Enfermedades Raras (CIBERER), 46010 Valencia, Spain, ³Functional Genomics Node (INB) at CIPF, 46012 Valencia, Spain and ⁴Research Center on Software Production Methods (ProS), DSIC Universitat Politècnica de València (UPV), 46007 Valencia, Spain

Received March 23, 2012; Revised May 18, 2012; Accepted May 21, 2012

CellBase, the Knowledge base and **OpenCGA**, the genomic data management engine are projects initiated in our group by 2010. Now are the backbone of the GEL

Ignacio Medina, Head of Computational Biology Lab HPC Service, University of Cambridge, and Head of Bioinformatics Databases at Genomics England has been building many of the applications that sit on top of MongoDB. He said:

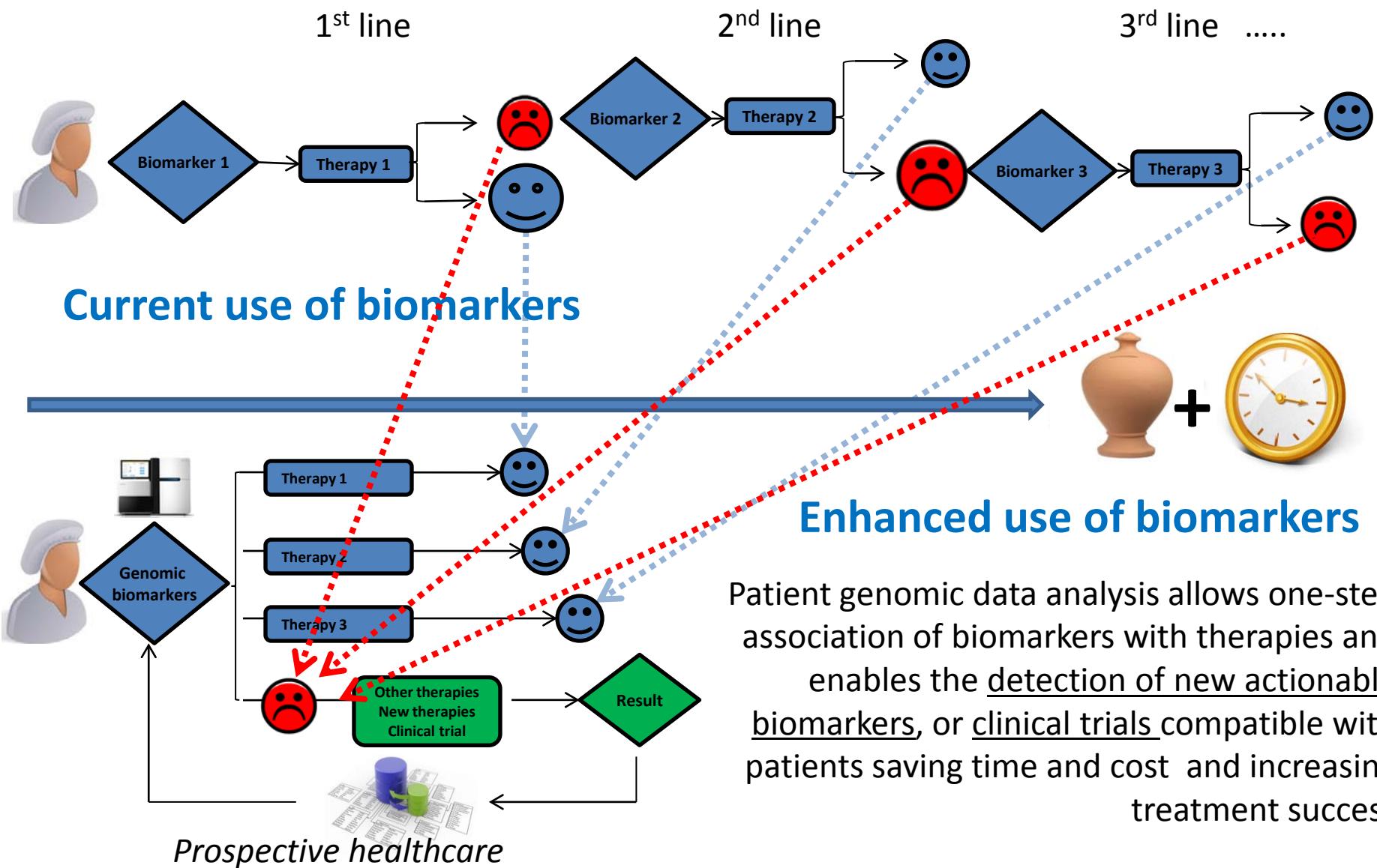
“*MongoDB is performing beautifully for us. From the beginning of the project it's been fantastic for our developers to iterate quickly. Now that the 100,000 Genomes Project is running at scale, MongoDB is also helping us extend our experience on to the scientists and clinicians who access the data, making it easier and faster for them to find critical insights in the data.”*



Ignacio Medina

Two of the important projects also utilising MongoDB are [Cellbase](#) and [OpenCGA](#) (Computational Genomics Analysis). Cellbase is a data warehouse and open API that stores reference genomic data from public resources such as Ensembl, Clinvar, and Uniprot. By relying on MongoDB, Cellbase can typically run sophisticated queries in an average of 40 milliseconds or less, and complex aggregations in less than one second – down from six hours using previous filesystem-based querying and storage. Importantly, it can annotate about 20,000 variants per second, making it compatible with whole genome sequencing data throughput requirements, while also returning a rich set of annotations that helps scientists better understand the data.

Personalized Medicine in cancer



Niveles de evidencia para la recomendación de tratamiento en cáncer

Nivel 1	Biomarcador de tratamiento estándar predictivo de respuesta a un fármaco indicado para este cáncer	Implicaciones terapéuticas estándar
Nivel 2A	Biomarcador predictivo de respuesta a tratamiento con un fármaco indicado para una mutación de este cáncer*	*Incluye biomarcadores de ensayos basket
Nivel 2B	Biomarcador de tratamiento estándar predictivo de respuesta a un fármaco indicado para otro cáncer	
Nivel 3A	Biomarcador con evidencia clínica convincente de respuesta de este cáncer a un fármaco que no es indicación terapéutica	Terreno de investigación terapéutica.
Nivel 3B	Biomarcador con evidencia clínica convincente de respuesta de otro cáncer a un fármaco que no es indicación terapéutica	Posibilidad de reclutamiento en ensayos clínicos
Nivel 4	Biomarcador con evidencia biológica convincente de respuesta de algún cáncer a un fármaco que no es indicación terapéutica	Implicaciones terapéuticas hipotéticas
Nivel R1	Biomarcador de tratamiento estándar predictivo de resistencia a tratamiento con un fármaco indicado para este cáncer	Implicaciones terapéuticas estándar
Nivel R2	Biomarcador no estándar con evidencia clínica convincente de ser predictivo de resistencia a tratamiento con un fármaco	Implicaciones terapéuticas hipotéticas basadas en datos preliminares no clínicos
Nivel R3	Biomarcador no estándar con evidencia biológica convincente de ser predictivo de resistencia a tratamiento con un fármaco	

Adaptado de: Chakravarty et al. JCO precision oncology 2017

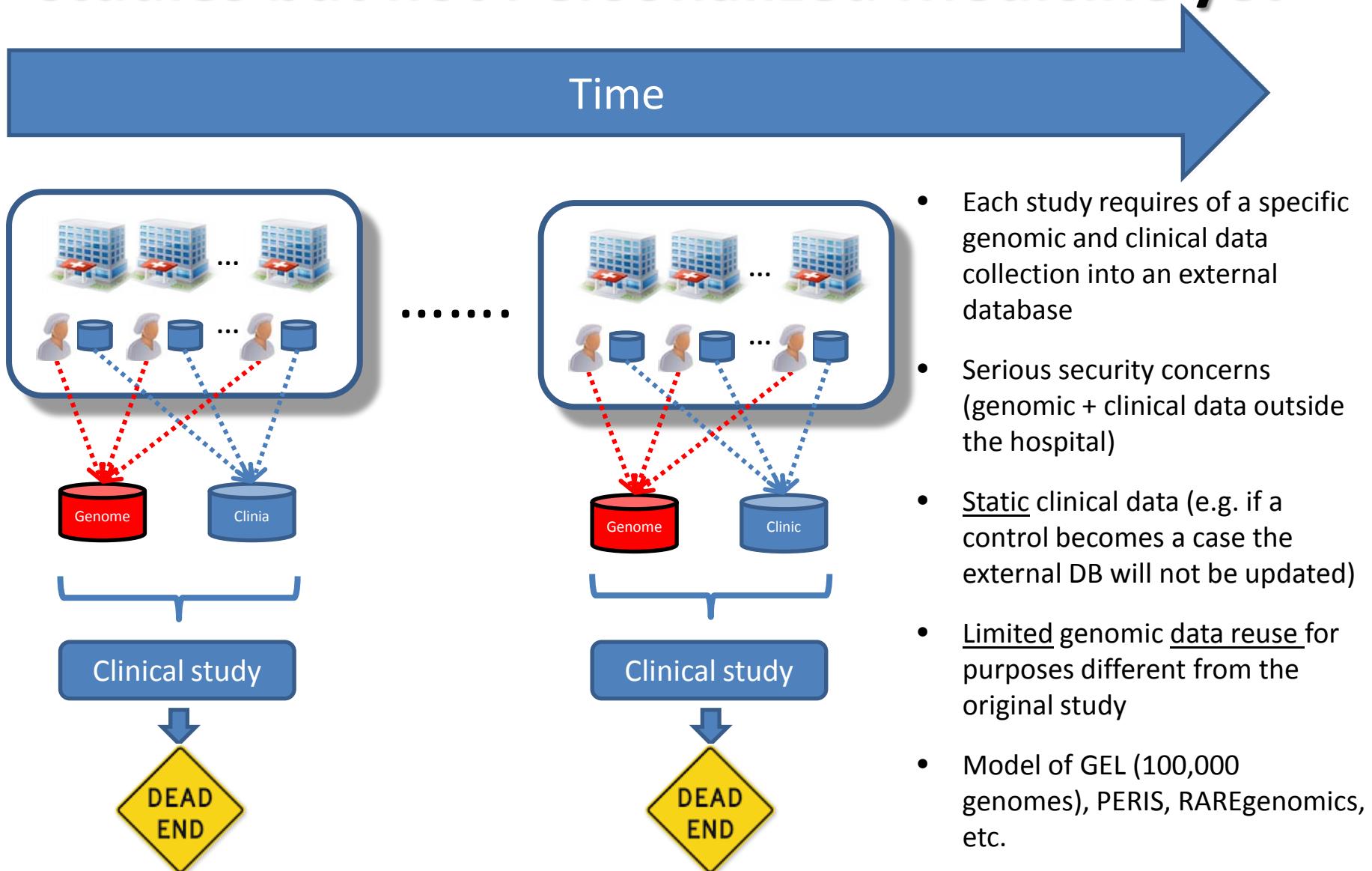
However, precision diagnostic is only the beginning to implement personalized medicine in the health system

- Precision diagnostic using genomic data can be carried out everywhere (only a sequencer and the appropriate software is required)
- Genomic data generation in a disconnected health system generates **silos** of data at different hospitals or even departments, which limits the sample size for clinical studies

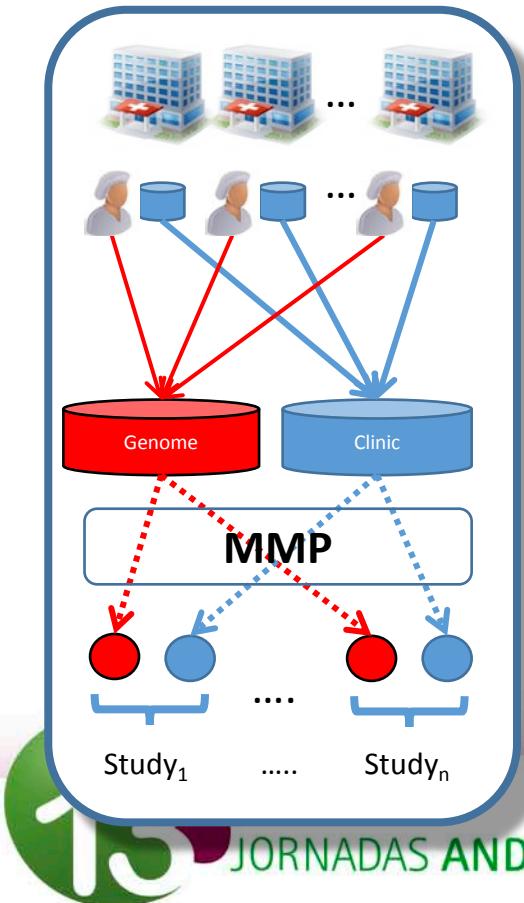


JORNADAS ANDALUZAS SALUD INVESTIGA | GRANADA · 29 DE OCTUBRE · 2018

Actually, genomic initiatives are just clinical studies but not Personalized Medicine yet

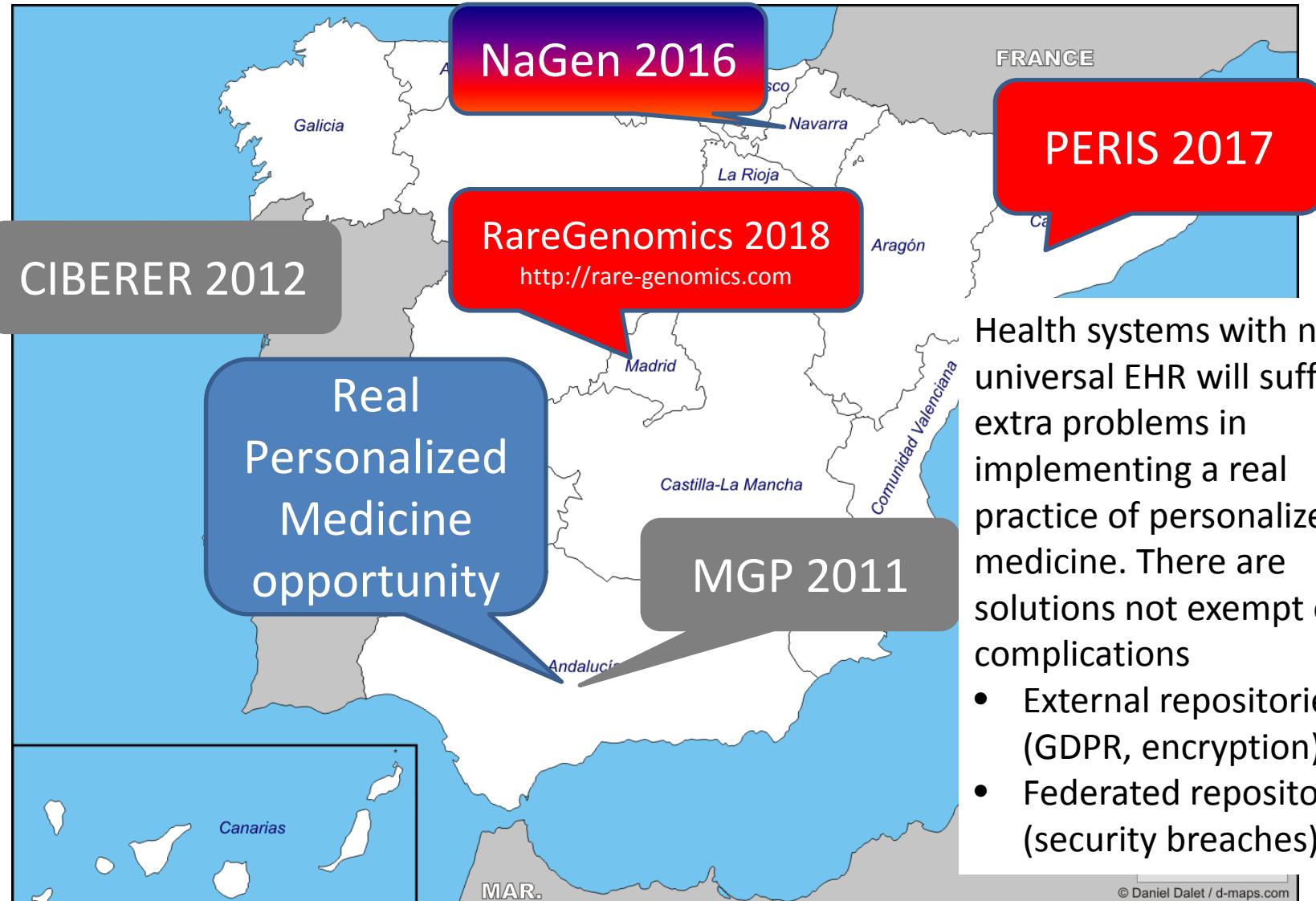


The real implementation of Personalized Medicine is facilitated by a model that integrates genomic data and universal EHR



- The whole health system becomes a enormous potential prospective clinical study
- Clinical data dynamically associated to genomic data
- Possibility of many clinical studies by reanalyzing genomic data under diverse perspectives (with no extra investment)
- Growing genomic DB with increasing study possibilities

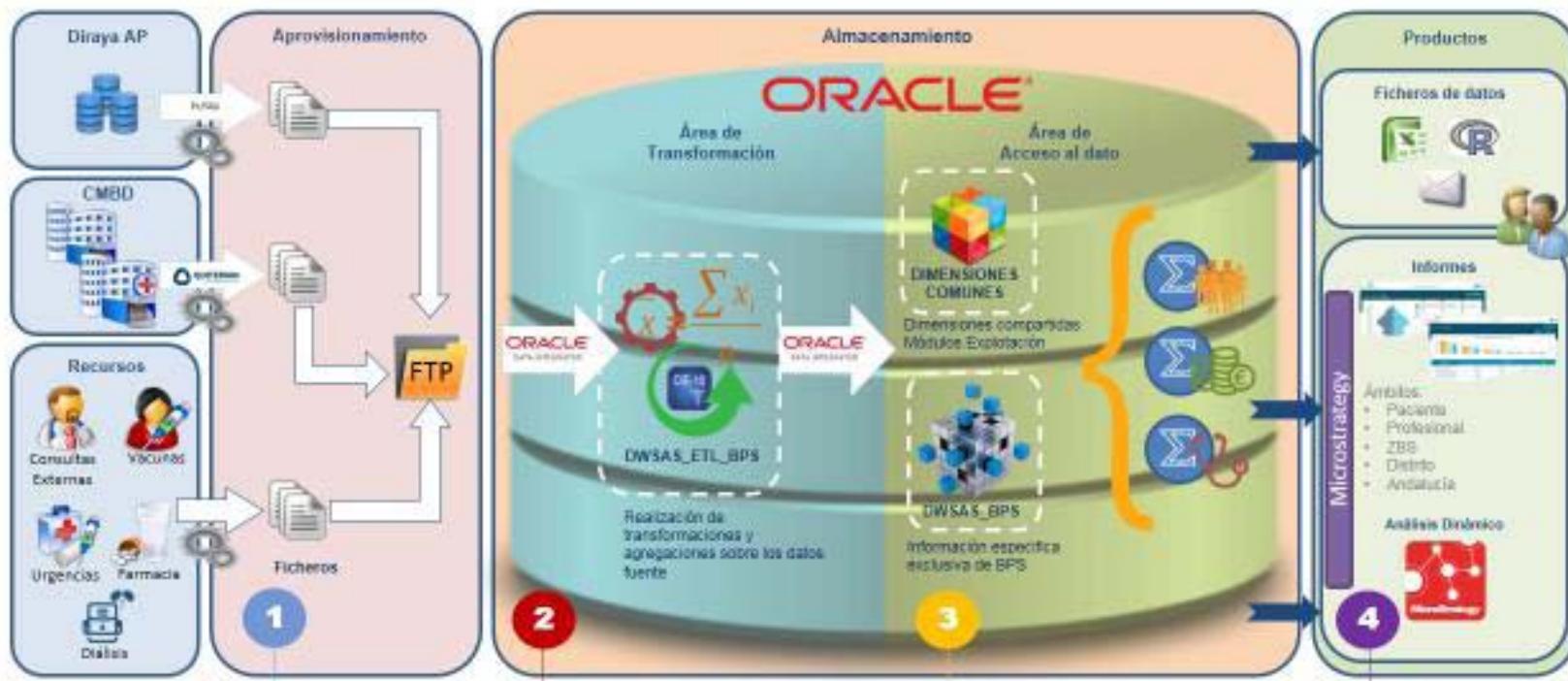
Real opportunities for personalized medicine



The population health database



Possibly the largest database ever created with detailed clinical data, storing information on 12.083.681 patients since 2001



En el área de extracción se obtienen ficheros desde los sistemas operacionales que almacenan la información susceptible de ser incorporada. Las fuentes de información susceptibles de ser incorporadas son:

- DIRAYA_AP
- CMBD
- CMED_URG
- INFHOS
- VACUNAS
- FARMACIA
- DIÁLISIS

En el Área de transformación se enriquecen los datos procedentes de los ficheros para favorecer su explotación. Las acciones que se llevan a cabo son:

- Codificación de diagnósticos.
- Agrupadores (ACG y GMA).
- Cálculo de indicadores.
- Aplicación de reglas de negocio.

En esta área se almacenan las **tablas de dimensiones** y **tablas de hechos** de los modelos en estrella definidos. Además, contiene tablas agregadas para dar soporte a la explotación y favorecer el rendimiento de los informes. Las tablas se almacenarán en distintos esquemas para cada uno de los datos que componen el sistema.

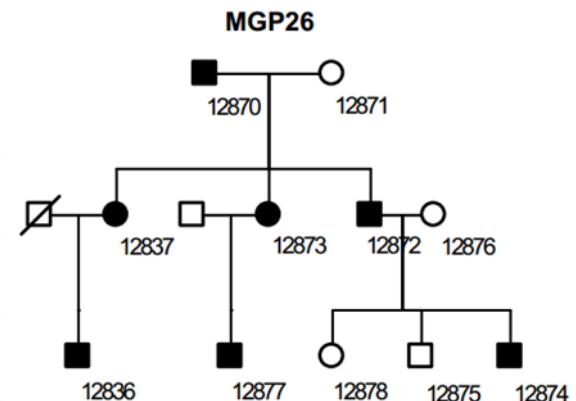
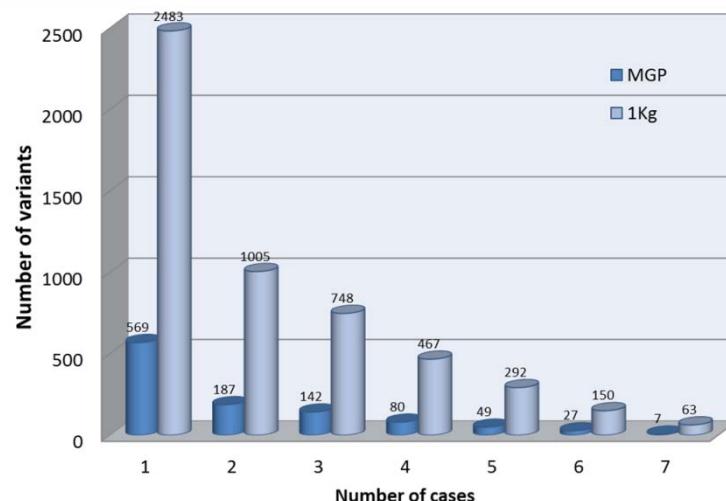
El sistema proporciona:

- Ficheros de datos.
- Informes a distintos niveles, desde paciente a Andalucía.
- Herramientas para análisis.

Lessons learned from MGP: the importance of local variability

The screenshot shows the Oxford Journals website for Molecular Biology and Evolution. The article title is "267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation". The authors listed include Joaquín Dopazo, Alicia Amadoz, Marta Bleda, Luz García-Alonso, Alejandro Alemán, Francisco García-García, Juan A. Rodríguez, Josephine T. Daub, Gerard Muntané, Antonia Rueda, Alicia Vela-Boza, Francisco J. López-Domínguez, Javier P. Florido, Pablo Arce, Macarena Ruiz-Ferrer, Cristina Méndez-Vidal, Todd E. Arnold, Olivia Spleiss, Miguel Alvarez-Tejado, Arcadi Navarro, Shomi S. Bhattacharya, Salud Borrego, Javier Santoyo-López, and Guillermo Antúñolo. The article is Open Access, with links to the abstract, full text (HTML and PDF), and supplementary data. The journal's November 2016 issue is also displayed.

We discovered some 12,000 “Spanish” polymorphisms not present in other databases. The filtering efficiency enormously increases using local population data



The CSVS is a crowdsourcing project

Welcome to the Collaborative Spanish Variant Server. CSVs was created to provide information about the variability of the Spanish population to the scientific/medical community. It is useful for detecting polymorphisms and local variations in the process of prioritizing candidate disease genes. CSVs currently stores information on 1582 unrelated Spanish individuals. We accept submissions from WES or WGS. See the protocol for sending samples.

Supported by:

Note:
CSVs web application makes an intensive use of the HTML5 standard and other cutting-edge web technologies such as Web Components, so only modern web browsers are fully supported; these include Chrome 34+, Firefox 32+, IE 10+, Safari 7+ and Opera 24+.

Chromosome	Location	Allele	Ref	rsID	n	2	1	0	0.197	0.003	0.003	0.360	0.430	0.280	0.440	0.400	0.450	0.438	0.430	0.305	0.442		
10	43572512	C>G	RET	rs13990297	575	2	1	0	0.197	0.003	0.003	0.360	0.430	0.280	0.440	0.400	0.450	0.438	0.430	0.305	0.442		
10	43572812	G>A	RET	rs12287480	508	62	18	0	0.124	0.078	0.078	0.350	0.400	0.270	0.330	0.400	0.368	0.418	0.410	0.305	0.442		
10	43515808	CT>G	RET		575	3	0	0	0.197	0.003	0.003												
10	43515818	C>T	RET		575	3	0	0	0.197	0.003	0.003												
10	43515837	T>C	RET		474	71	33	0	0.881	0.119	0.119												
10	43515837	T>	RET		474	71	33	0	0.881	0.119	0.119												
10	43515968	A>G	RET	rs1800858	348	74	156	0	0.464	0.334	0.334	0.280	0.220	0.470	0.090	0.300	0.246	0.253	0.308	0.473	0.280	0.300	0.473
10	43516002	G>A	RET		577	1	0	0	0.999	0.001	0.001												
10	43516179	G>A	RET	rs2435251	380	355	36	7	0.801	0.199	0.199	0.380	0.260	0.100	0.070	0.260	0.335	0.280	0.288	0.118	0.442		

<http://csvs.babelomics.org/>

Allelic population frequencies obtained from 1,600 exomes are currently available in CSVS

Scenario: Sequencing projects of healthy population are expensive and funding bodies are reluctant to fund them

CSVs Aim: To offer increasingly accurate information on variant frequencies characteristic of Spanish population.

CSVs Main use: Frequency-based filtering of candidate variants

Main data source: Sequencing projects of individual researchers (CIBERER and others)

Problem: Most of the contributions correspond to patient exomes

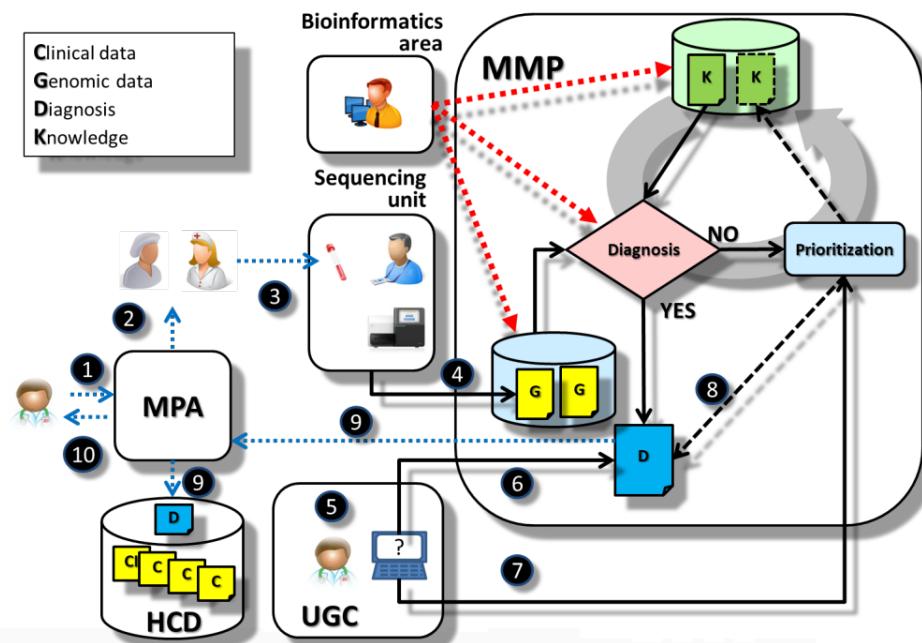
Idea: Patients of disease A can be considered healthy **pseudo-controls** for disease B (providing no common genetic background exist between A and B)

Beacon: CSVS has a Beacon server

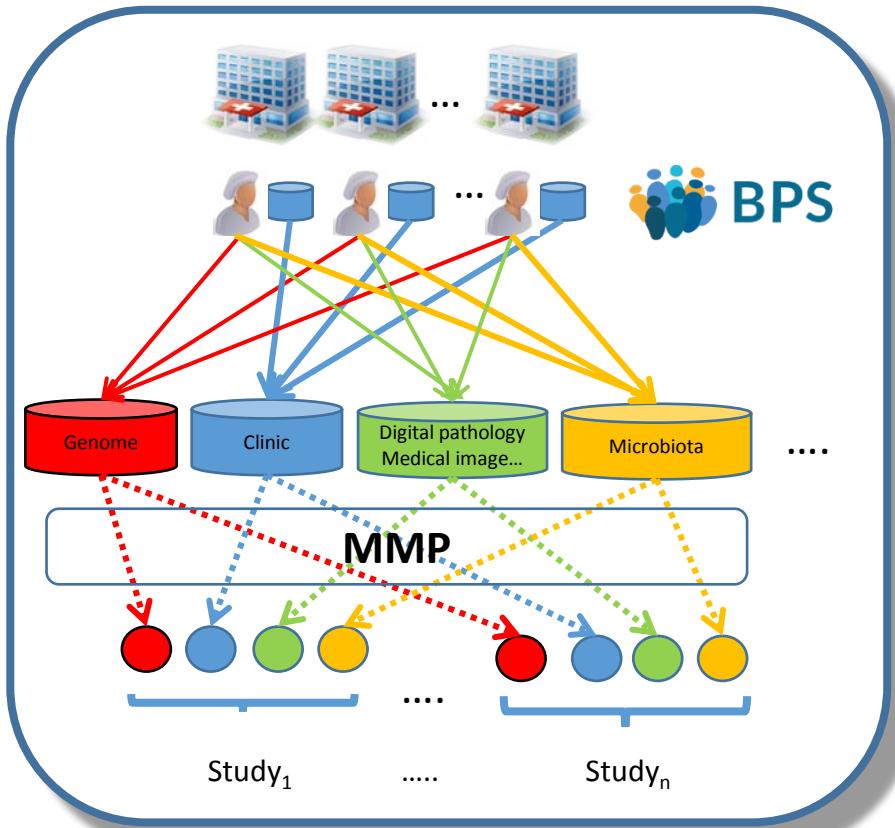
GDPR compliance

The system has been designed in a way that is compliant with EU and Spanish General Data Protection Regulation

- Clinicians requesting for a genomic diagnostic have access to eHR and only get the result of the test.
- Geneticists have access to eHR and can query the genomic data (but never extract them)
- IT have access to anonymized genomic data but not to eHR.



Future vision involves *big data* integration: Genomic data are especially relevant for discovering the genetic determinants of diseases, but not the only useful *big data*



- Other *big data* are being collected (medical image, digital pathology, wearable devices, etc.)
- Microbiota in the future (CR cancer screening)
- Clinical data In the BPS will be dynamically associated to different *big data*
- The whole health system becomes a enormous potential prospective clinical study
- Immense possibility for data reusability
- Growing genomic DB with increasing study possibilities

Clinical Bioinformatics Area

Fundación Progreso y Salud, Sevilla, Spain, and...

...the INB-ELIXIR-ES, National Institute of Bioinformatics
and the BiER (CIBERER Network of Centers for Research in Rare Diseases)



<https://www.slideshare.net/xdopazo/>



Follow us on
twitter
[@xdopazo](https://twitter.com/xdopazo)
[@ClinicalBioinfo](https://twitter.com/ClinicalBioinfo)





JORNADAS ANDALUZAS
SALUD INVESTIGA

GRACIAS POR SU ATENCIÓN

Promueven _____



Servicio Andaluz de Salud
CONSEJERÍA DE SALUD

Organiza _____



Fundación Progreso y Salud
CONSEJERÍA DE SALUD

Colaboran _____



Escuela Andaluza de Salud Pública
CONSEJERÍA DE SALUD



Biobanco del Sistema Sanitario Público de Andalucía
CONSEJERÍA DE SALUD



Biblioteca Virtual
del Sistema Sanitario Público de Andalucía



FUERTE UNIVERSIDAD DE GRANADA-IBS DE ANDALUCÍA
CENTRE FOR GENOMICS AND ONCOLOGICAL RESEARCH

