OMICS MASTER

Introduction to NGS Technologies for Variation Studies

PRINCIPE FELIPE
CENTRO DE INVESTIGACION
Computational · Genomics

IT4Innovations

Francisco García
fgarcia@cipf.es

Jul 2014 Ostrava, Czech Republic

Outline

1) Different Sequencing Technologies

2) Genomics Data Analysis Pipeline

3) Personalized Medicine

Outline

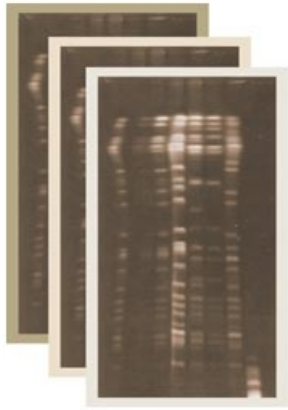
1) Different Sequencing Technologies

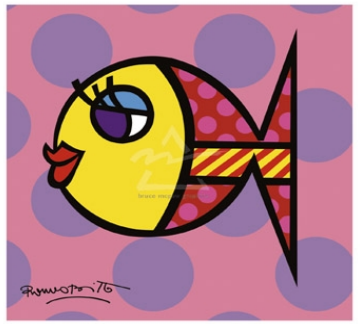
2) Genomics Data Analysis Pipeline

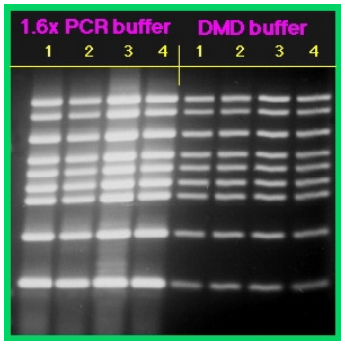
3) Personalized Medicine

Single gene- protein technologies

- 1952 Electrophoresis
- 1969 FISH
- 1972 Cloning
- 1975 Southern blot
- 1977 DNA sequencing
- 1983 RFLP realization
- 1985 PCR. DNA fingerprinting

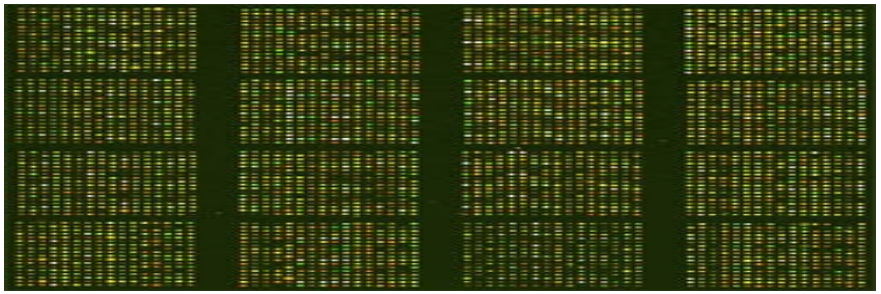


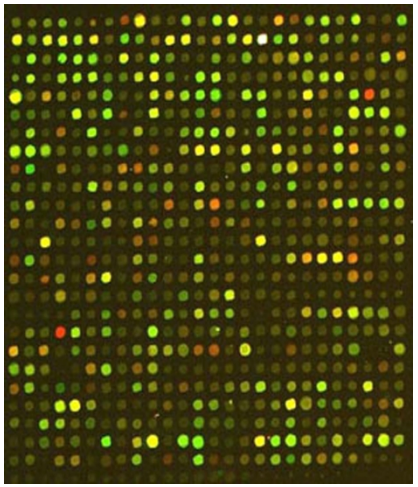




Omics technologies

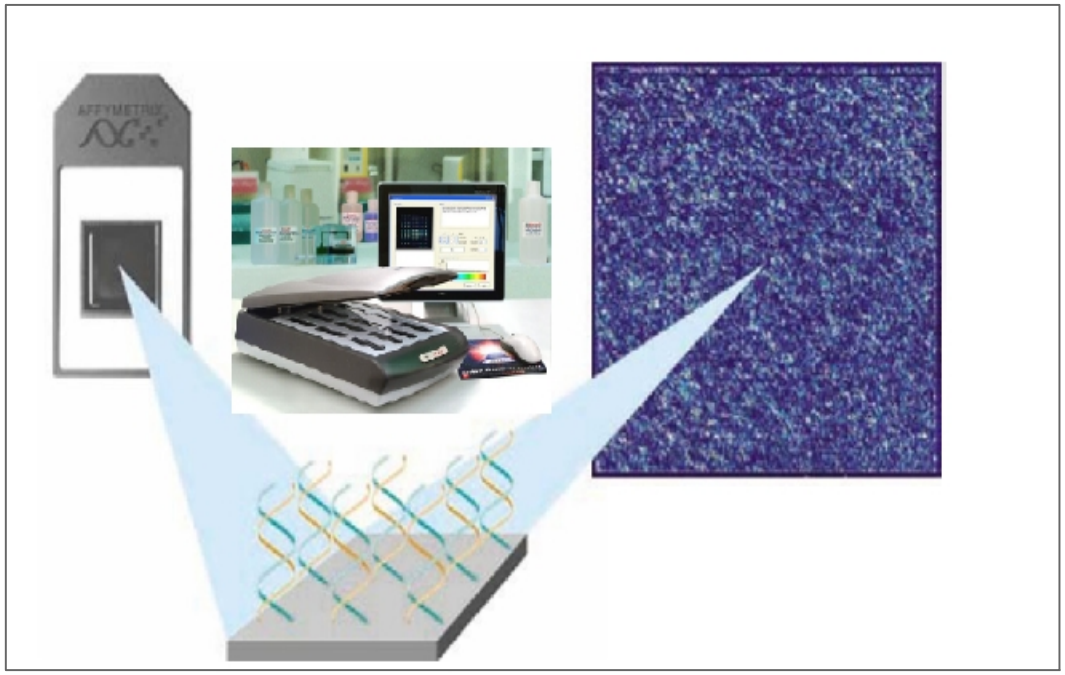
- 1988 arrayed DNAs were used
- 1991 oligonucleotides are synthesized on a glass slide through photolithography (Affymax Research Institute)
- 1995 DNA Microarrays
- 1997 Genome wide Yeast Microarray



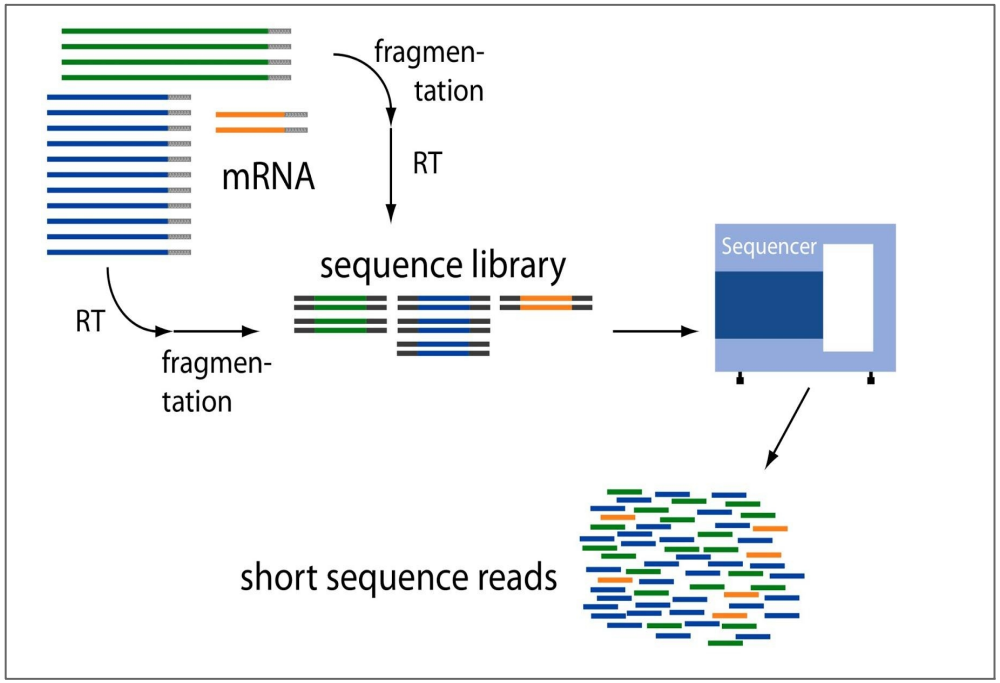


Omics technologies

- Microarrays

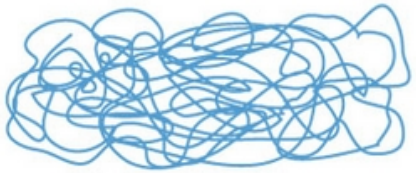


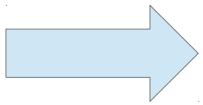
- Next Generation Sequencing (NGS)

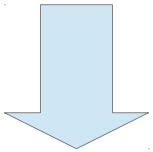


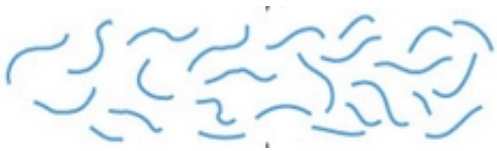
NGS technologies

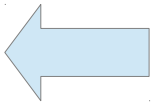
How do these technologies work ?



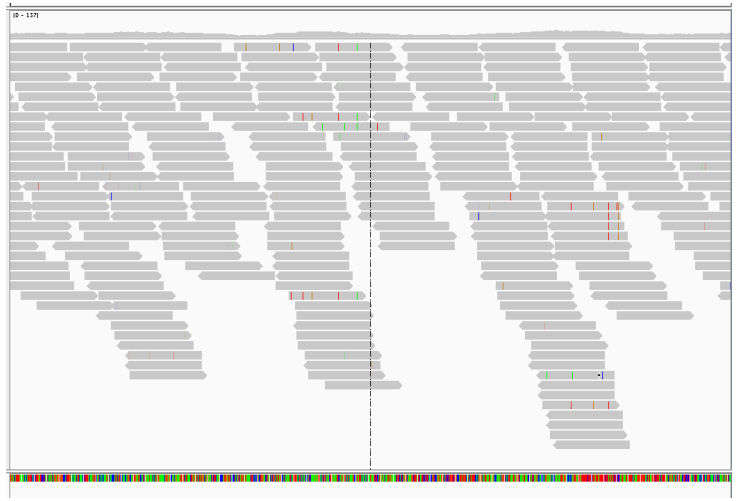








Reference genome



NGS technologies

Cost-effective
Fast
Ultra throughput
Cloning-free
Short reads

NGS technologies

Comparison

Roche 454

- Long fragments
- Errors: poly nts
- Low throughput
- Expensive

- De novo sequencing
- Amplicon sequencing
- RNASeq

Illumina

- Short fragments
- Errors: Hexamer bias
- High throughput
- Cheap

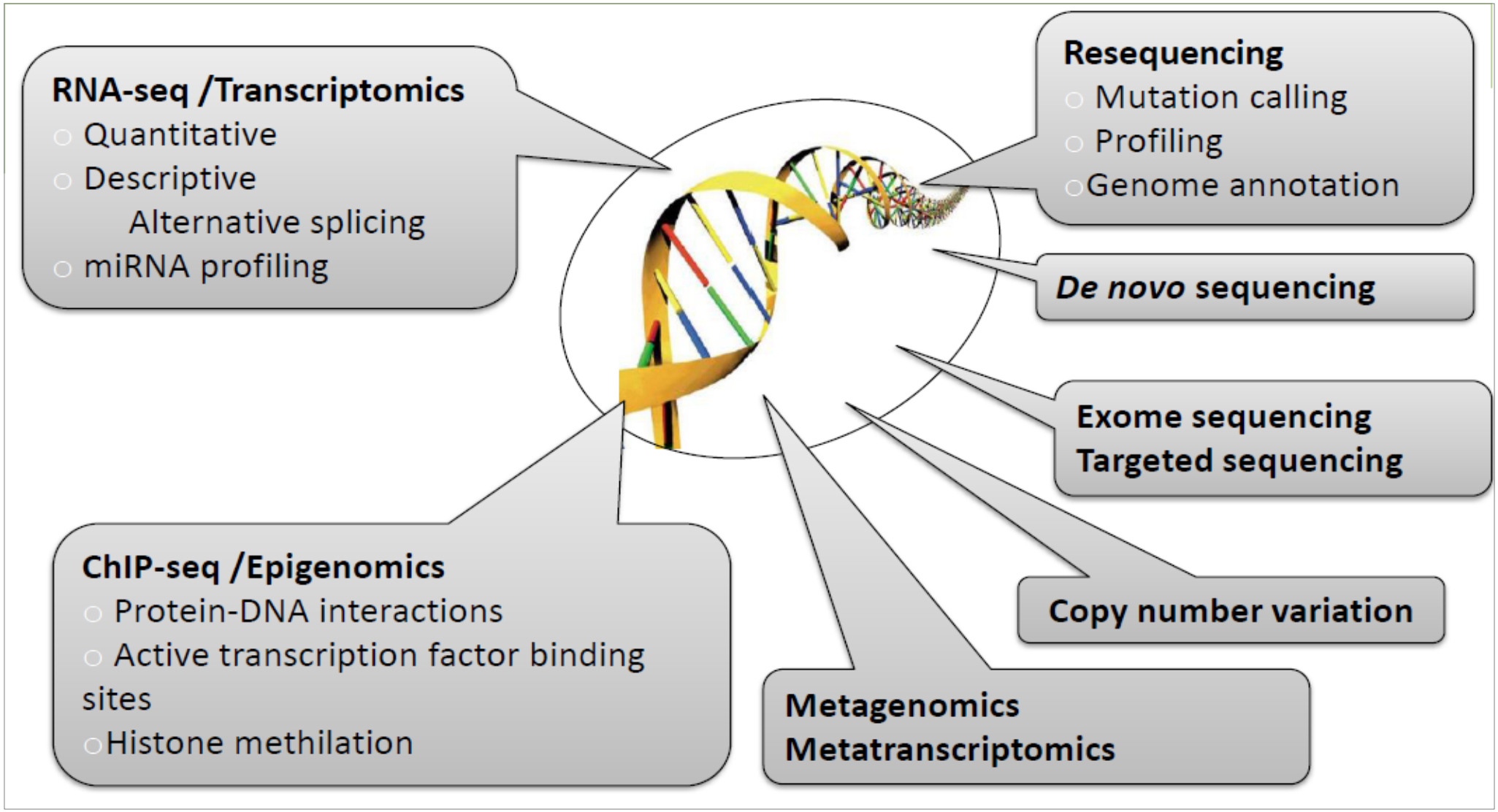
- Resequencing
- De novo sequencing
- ChipSeq
- RNASeq
- MethylSeq

SOLiD

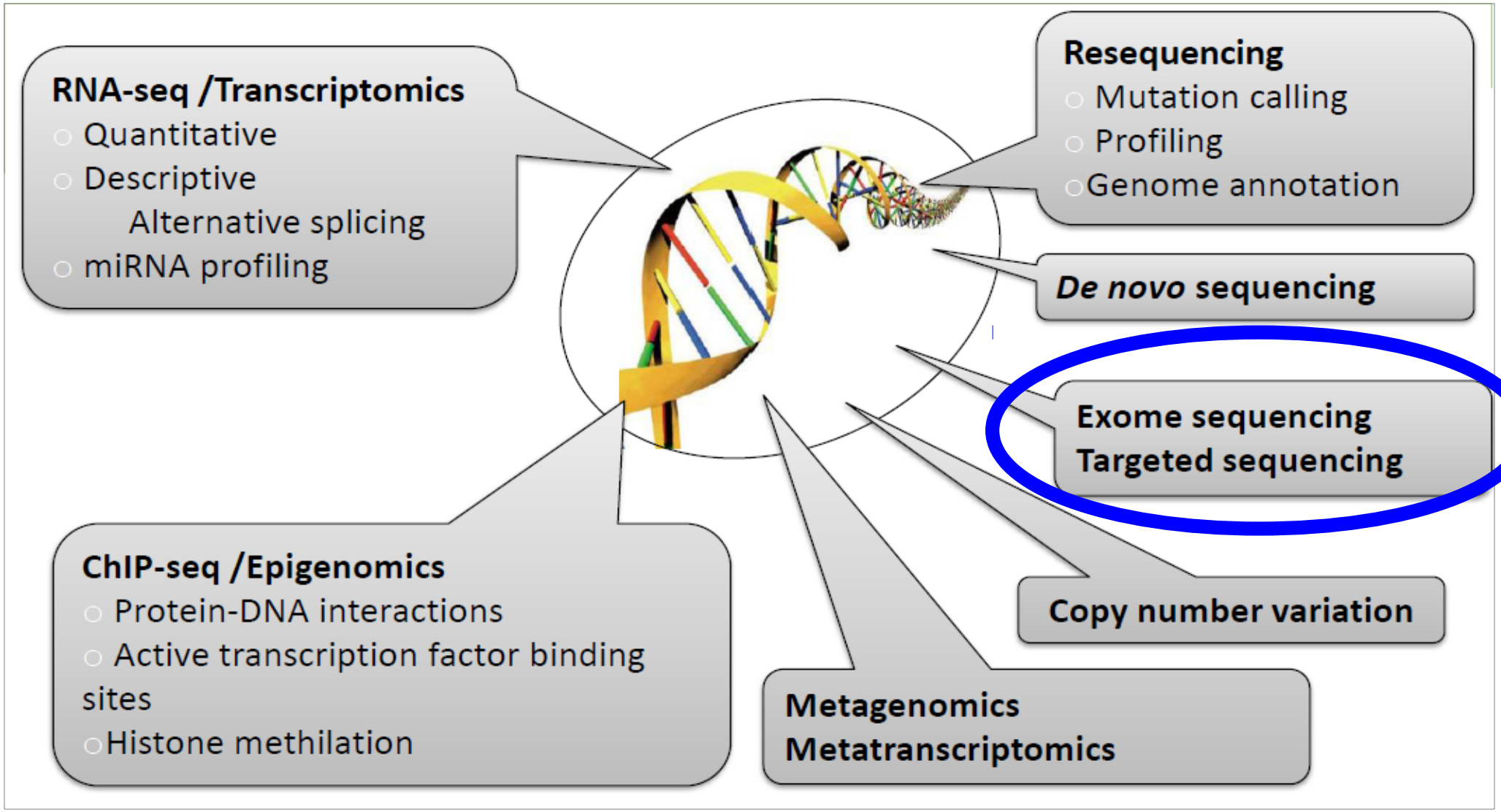
- Short fragments
- Color-space
- High throughput
- Cheap

- Resequencing
- ChipSeq
- RNASeq
- MethylSeq

Most common applications of NGS



Most common applications of NGS



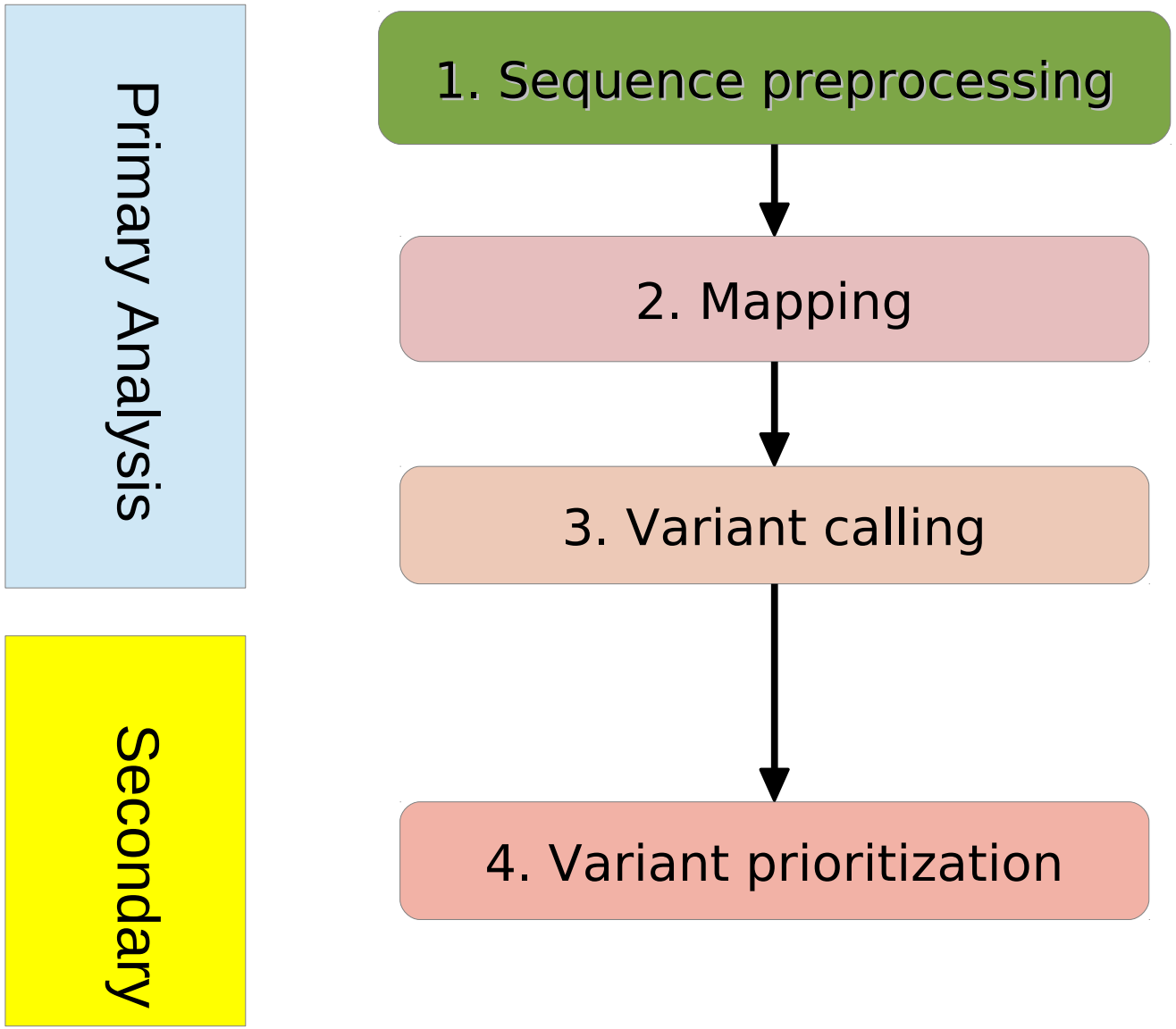
Outline

1) Different Sequencing Technologies

2) Genomics Data Analysis Pipeline

3) Personalized Medicine

Genomics Data Analysis Pipeline (1)



Genomics Data Analysis Pipeline (2)

Preprocessing + QC

- Sequence cleansing
- Base quality
- Remove adapters
- Remove duplicates

FASTQ file

Mapping + QC

- Mapping (HPG)
- Remove multiple mapping reads
- Remove low quality mapping reads
- Realigning
- Base quality recalibrating

BAM file

Variant calling + QC

- Calling and labeling of missing values
- Calling SNVs and indels (GATK) using 6 statistics based on QC, strand bias, consistence (poor QC callings are converted to missing values as well)
- Create multiple VCF with missing, SNVs and indels

VCF file

Variant and gene prioritization + QC

- Counts of sites with variants
- Variant annotation (function, putative effect, conservation, etc.)
- Inheritance analysis (including compound heterozygotes in recessive inheritance)
- Filtering by frequency with external controls (dbSNP, 1000g, ESP) and annotation
- **BiERapp / TeAM**

Primary analysis

Gene prioritization

Genomics Data Analysis Pipeline (3)

Fastq file, up to hundreds of GB per run

QC and preprocessing

QC stats, filtering and preprocessing options

HPG suite

High-Performance Genomics

HPG Aligner, short read aligner

Double mapping strategy:
Burrows-Wheeler Transform (*GPU Nvidia CUDA*) + Smith-Waterman (*CPU OpenMP+SSE/AVX*)

More info at:

http://bioinfo.cipf.es/docs/compbio/projects/hpg/doku.php

SAM/BAM file

QC and preprocessing

QC stats, filtering and preprocessing options

Other analysis (HPC4Genomics consortium)

Variant calling analysis

GATK and SAM mPileup HPC Implementation.
Statistics genomic tests

VCF file

- RNA-seq
- DNA assembly
- Methyl-seq
- Copy Number
- Structural variation
- Transcript isoform
- Etc.

Variant VCF viewer

HTML5+SVG Web based viewer

QC and preprocessing

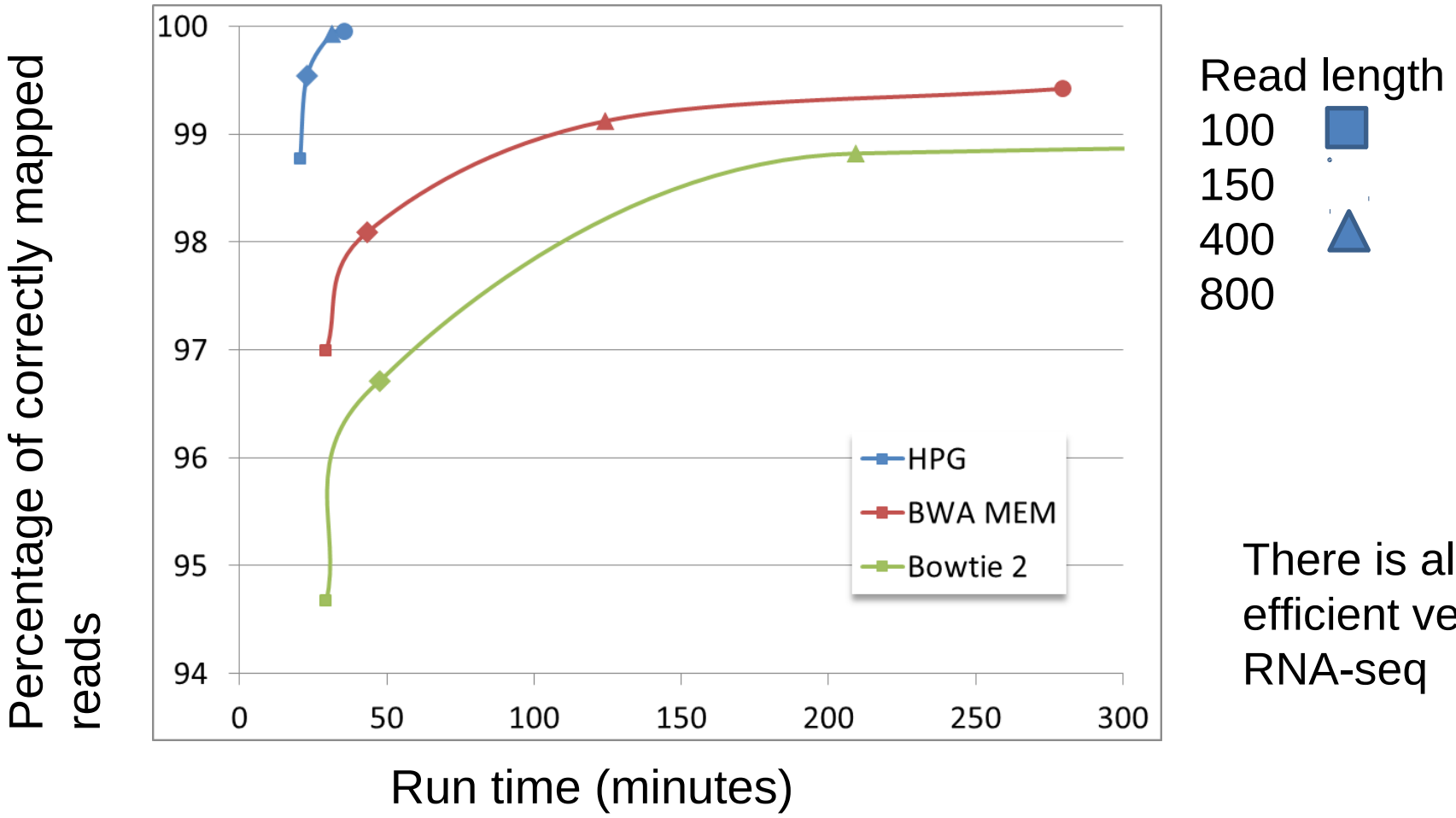
QC stats, filtering and preprocessing options

HPG Variant, Variant analysis

Consequence type, GWAS, regulatory variants and system biology information

Developing tools to improve the pipeline

HPG Aligner is free and open source. Documentation and software are available at http://www.opencb.org/projects/hpg/doku.php?id=aligner:overview



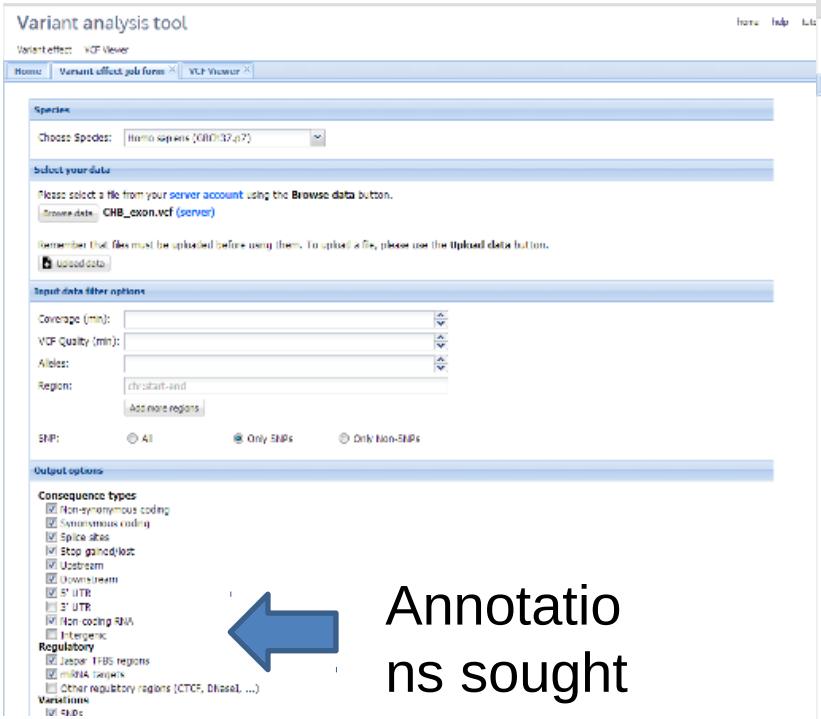
There is also a highly efficient version for RNA-seq

Developing tools to improve the pipeline

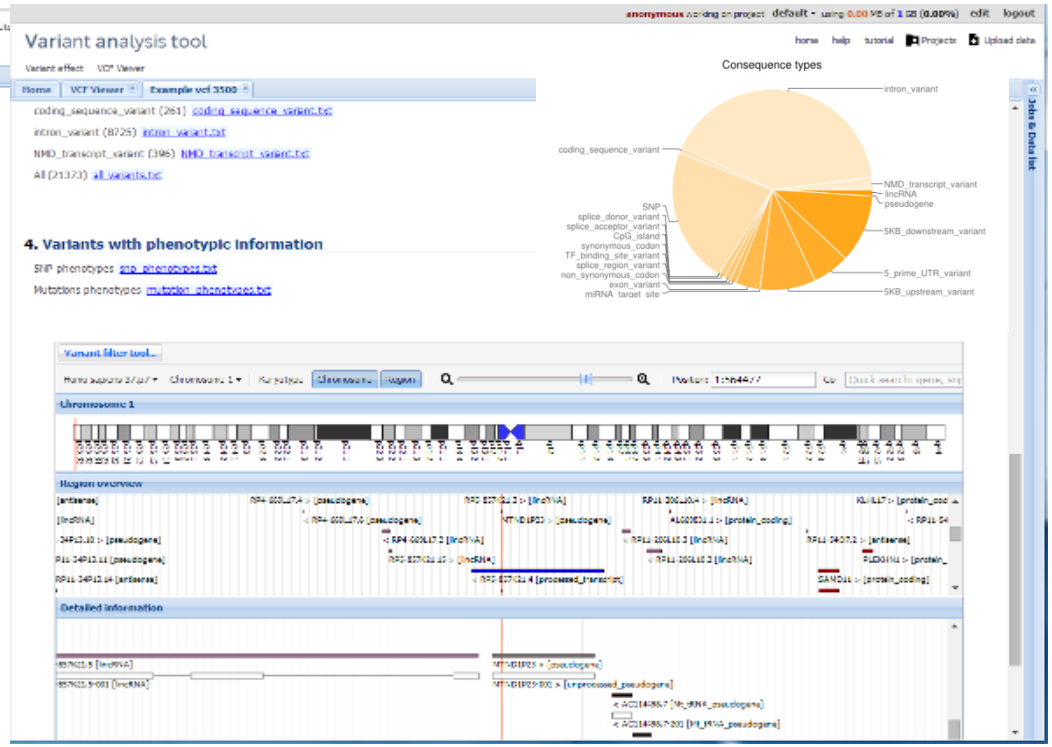
HPG Variant, a suite of tools for HPC-based genomic **variant annotation** **VARIANT** = ***VARI***ant ***AN***notation ***T***ool. Tools implemented using *OpenMP*, *Nvidia CUDA* and *MPI* for large clusters.

EFFECT: A CLI and web application, it's a *cloud*-based genomic variant ***effect*** predictor tool has been implemented (http://variant.bioinfo.cipf.es, *Medina 2012 NAR*)

VCF*: C library and tool*: allows to analyze large VCFs files with a low memory footprint: stats, filter, split, merge, etc. Example: *hpg-variant vcf –stats –vcf-file ceu.vcf*



Annotatio
ns sought

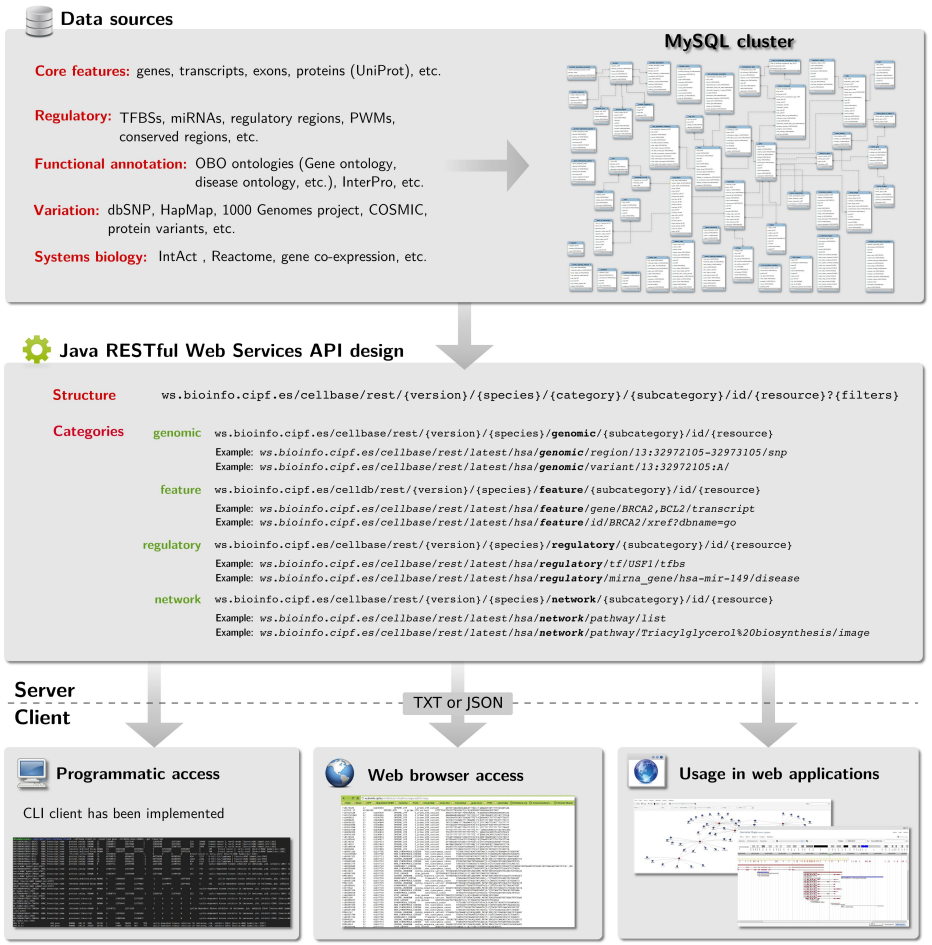


Developing tools to improve the pipeline

- ***CellBase*** (*Bleda, 2012, NAR*), a comprehensive integrative database and *RESTful Web Services API*, more than 250GB of data and 90 tables exported in TXT and JSON:
  - Core features: genes, transcripts, exons, cytobands, proteins (UniProt),...
  - Variation: dbSNP and Ensembl SNPs, HapMap, 1000Genomes, Cosmic, ...
  - Functional: 40 OBO ontologies (Gene Ontology), Interpro, etc.
  - Regulatory: TFBS, miRNA targets, conserved regions, etc.
  - System biology: Interactome (IntAct), Reactome database, co-expressed genes.
- **NoSQL** and scales to TB

Project: http://bioinfo.cipf.es/compbio/cellbase

Wiki: http://docs.bioinfo.cipf.es/projects/cellbase/wiki



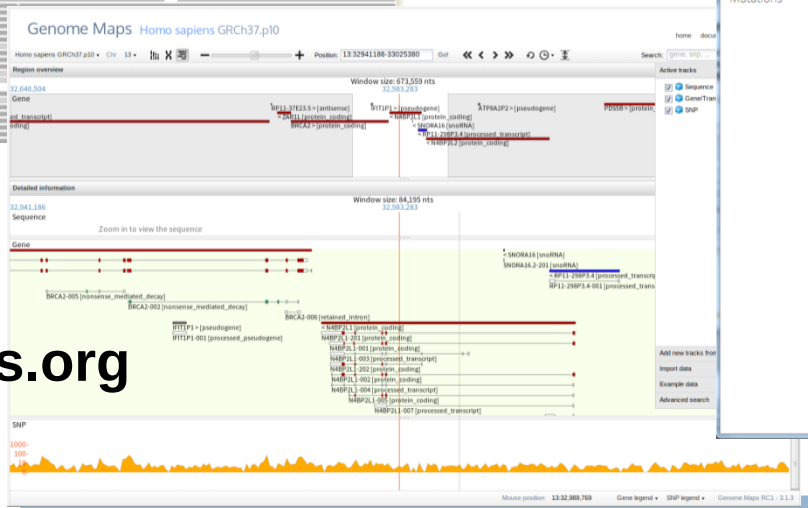
Developing tools to improve the pipeline

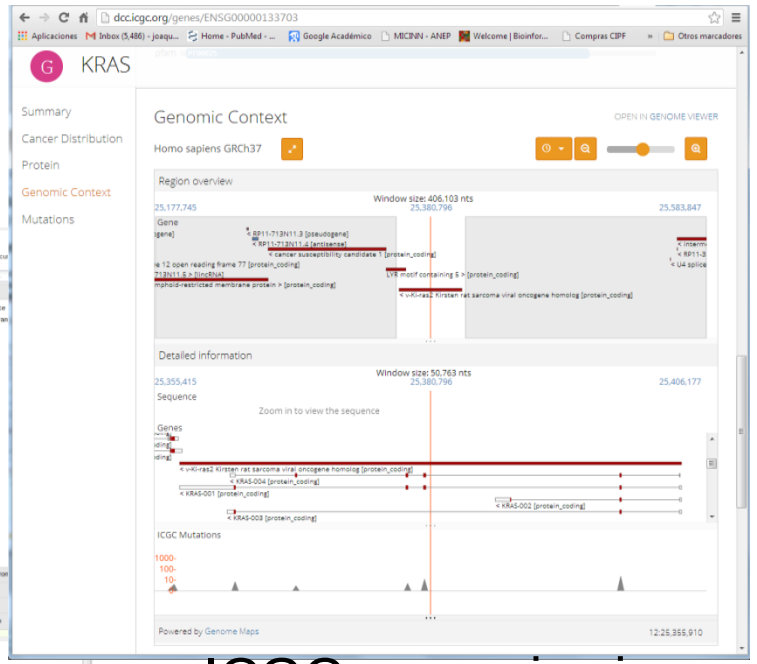
- Genome scale data **visualization** plays an important role in the data analysis process. It is a big data management problem.
- Features of ***Genome Maps*** (Medina, 2013, NAR; ICGC data analysis portal)
  - First 100% HTML5 web based: ***HTML5+SVG*** (inspired in Google Maps)
  - Always updated, **no browser plugins or installation**
  - Data taken from ***CellBase***, ***remote NGS data***, ***local files*** and ***DAS servers***: genes, transcripts, exons, SNPs, TFBS, miRNA targets, etc.
  - Other features: Multi species, API oriented, easy integration, plugin framework, etc.

BAM viewer

VCF viewer

www.genomemaps.org





ICGC genomic viewer

Outline

1) Different Sequencing Technologies

2) Genomics Data Analysis Pipeline

3) Personalized Medicine

Personalized Medicine

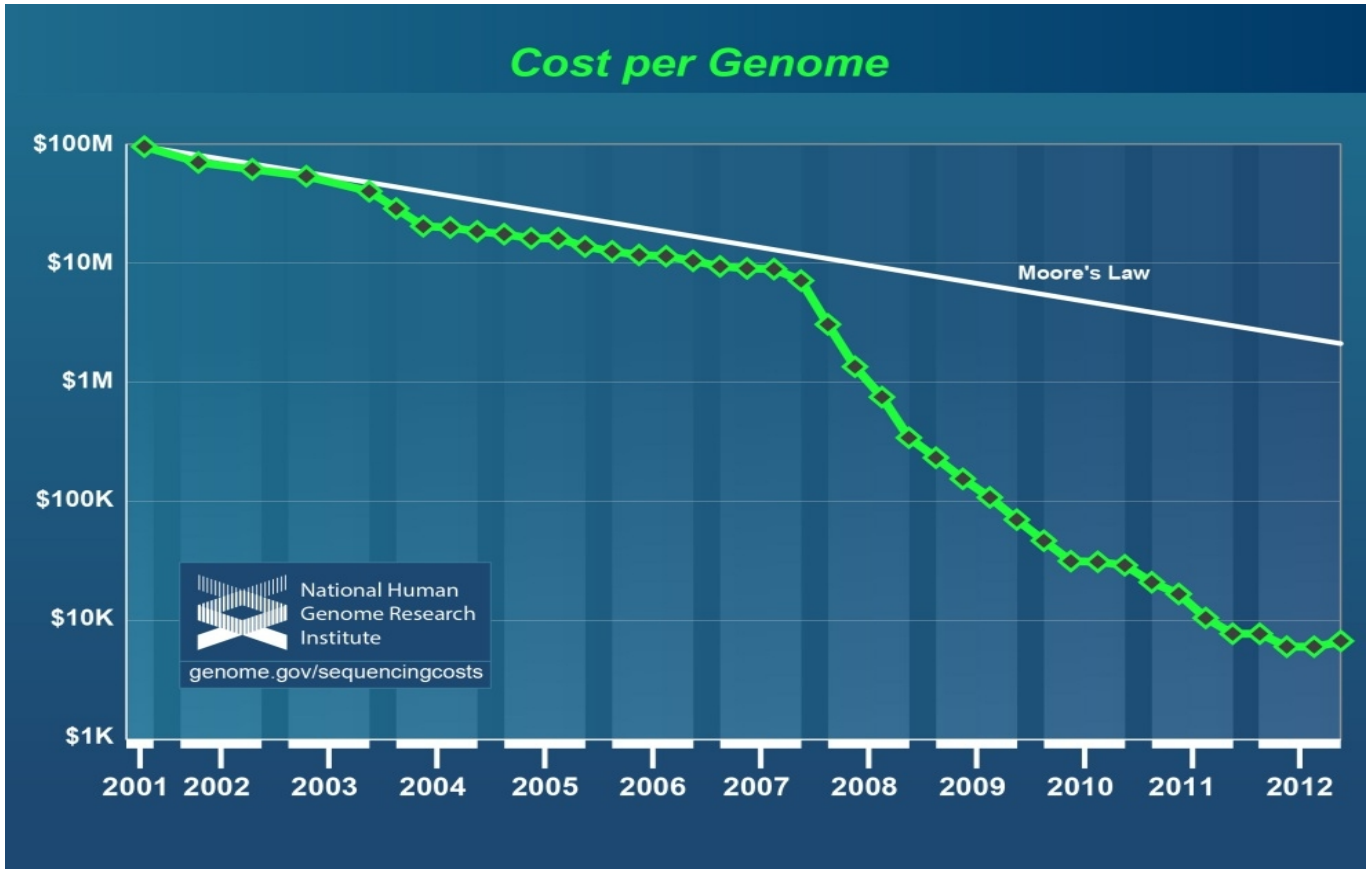
Personalized medicine: just about a better understanding of the **relationship phenotype-genotype**

- **The future of personalized medicine is strongly based on genomics**
- Personalized medicine is based on the availability of **diagnostic biomarkers**
- Genome sequencing offers **ALL** this information (if properly analyzed)

NGS prices will be soon affordable

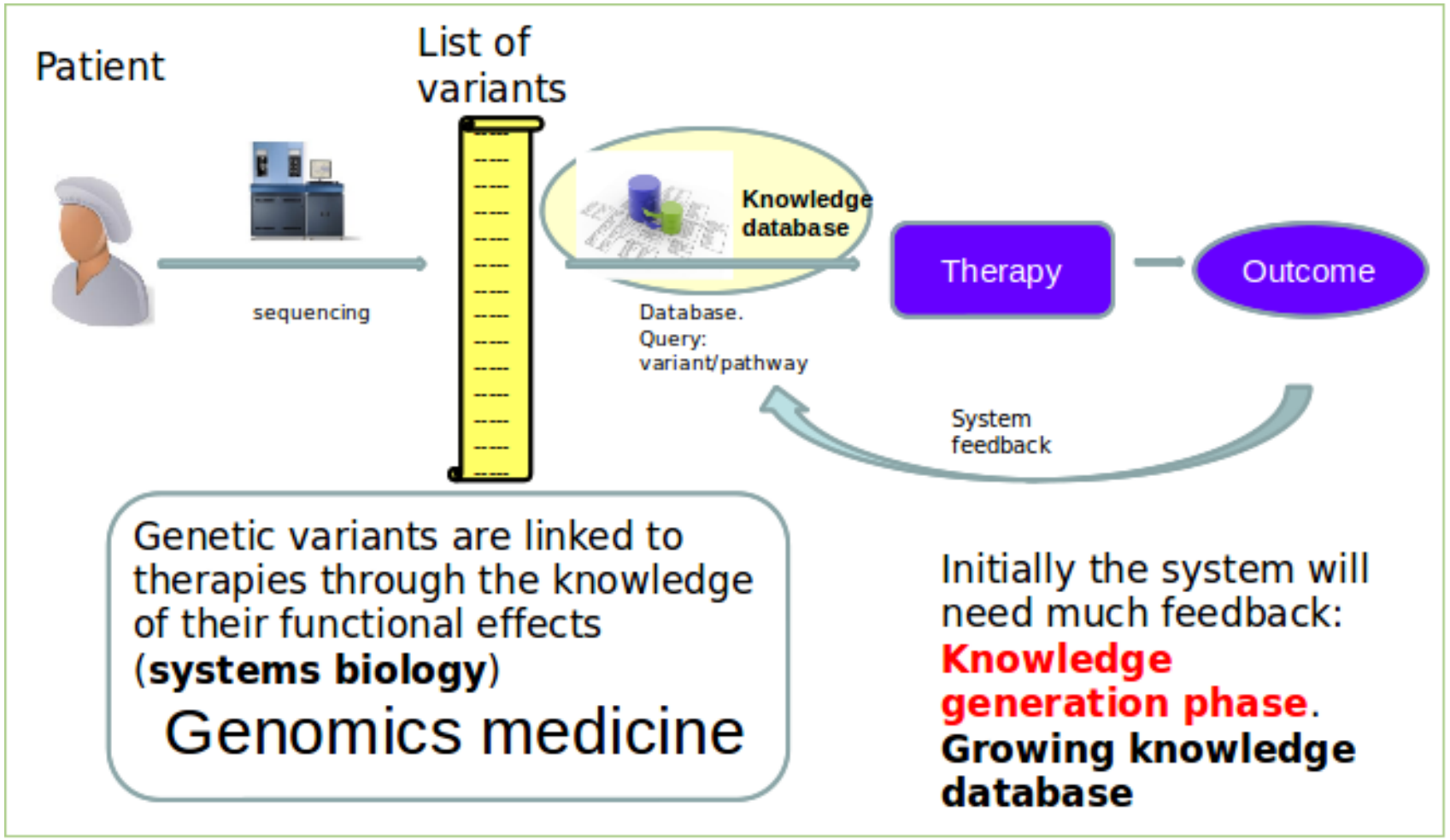
- While the cost falls down, the amount of data to manage and its complexity raise exponentially. Soon, costs will be competitive enough to be used in clinic

- **The problem is… are we ready to deal with this?**



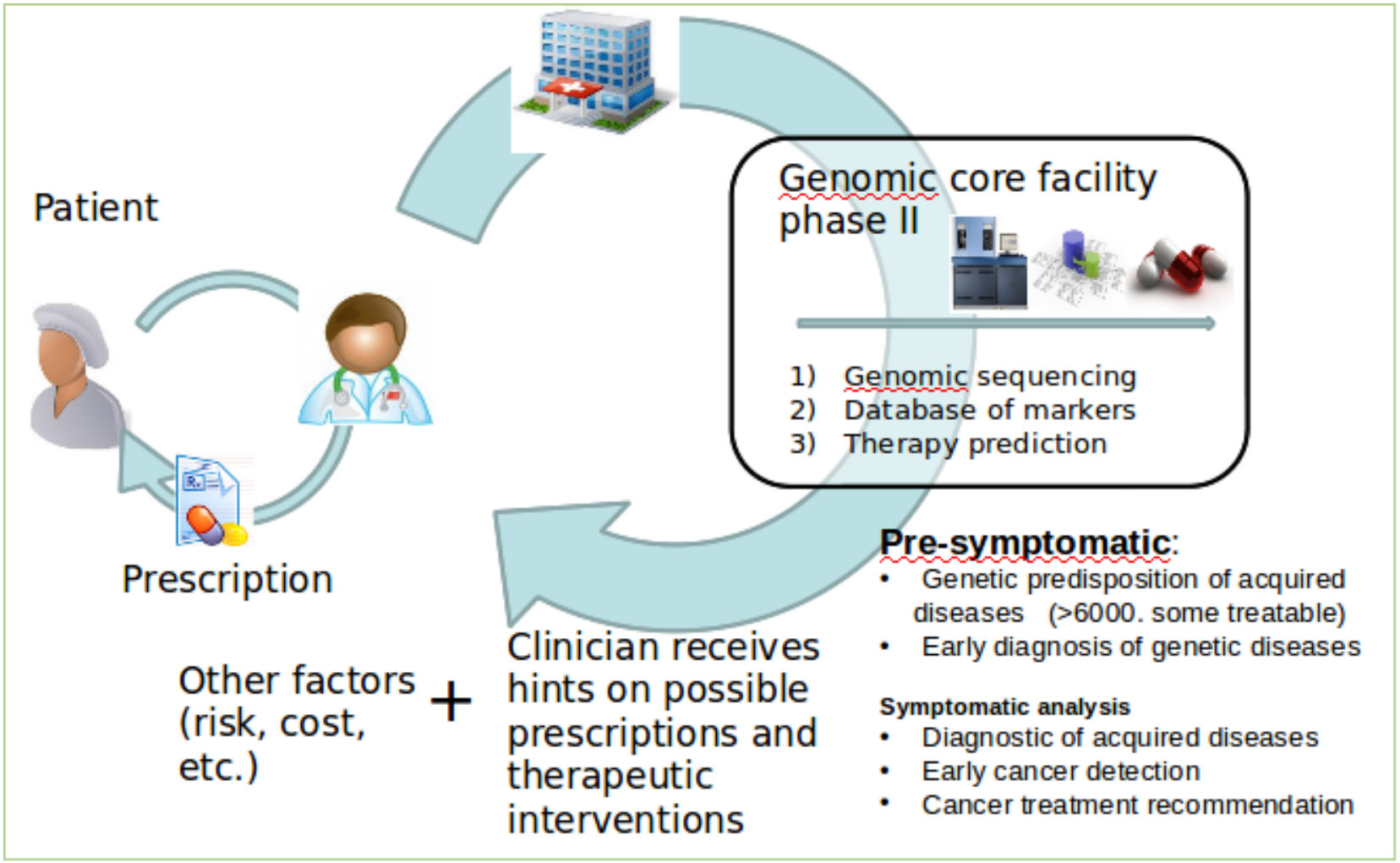
Personalized Genomic Medicine

Phase I: generating the knowledge database

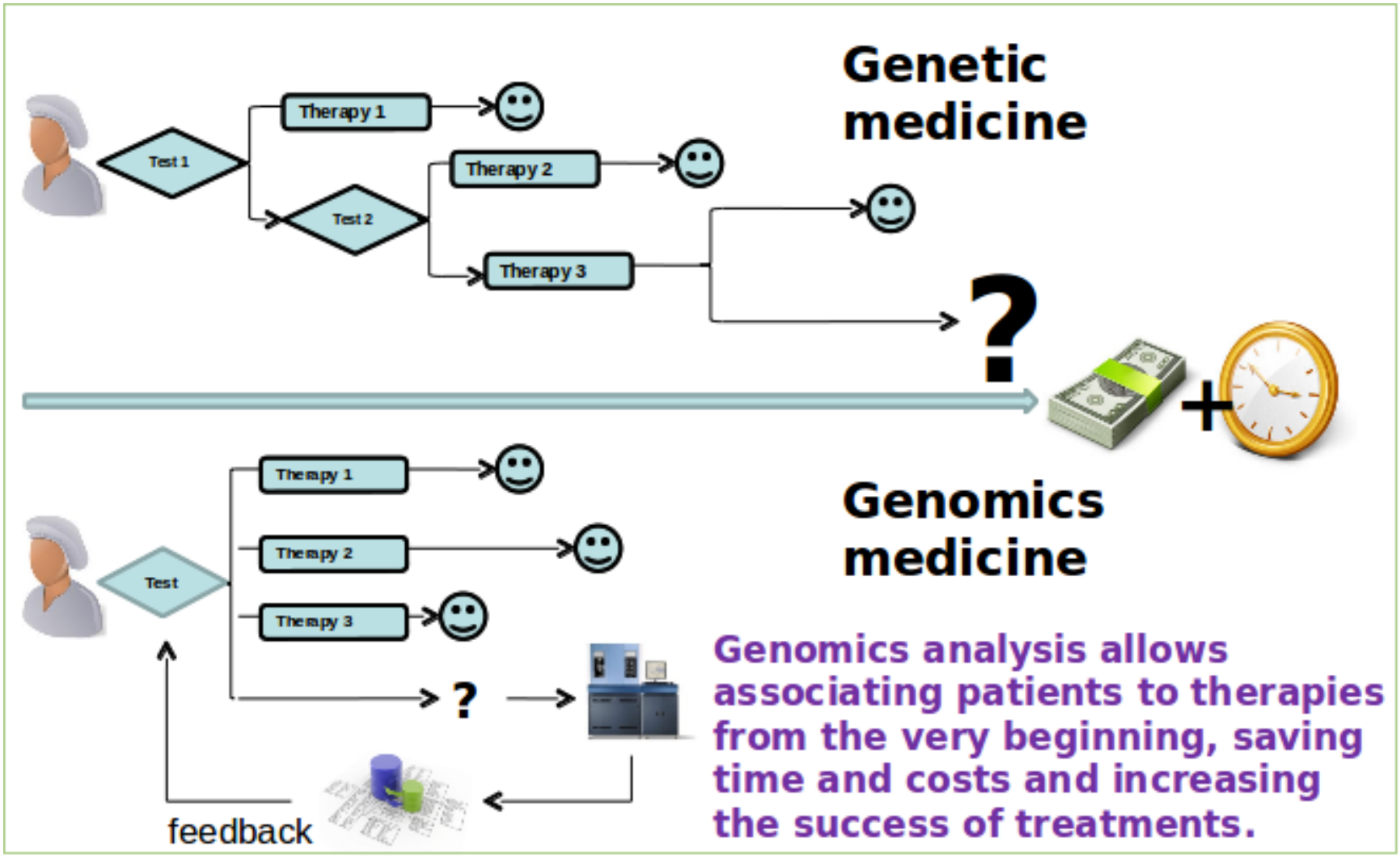


Personalized Genomic Medicine

Phase II: applying the knowledge database



From genetic to genomic medicine

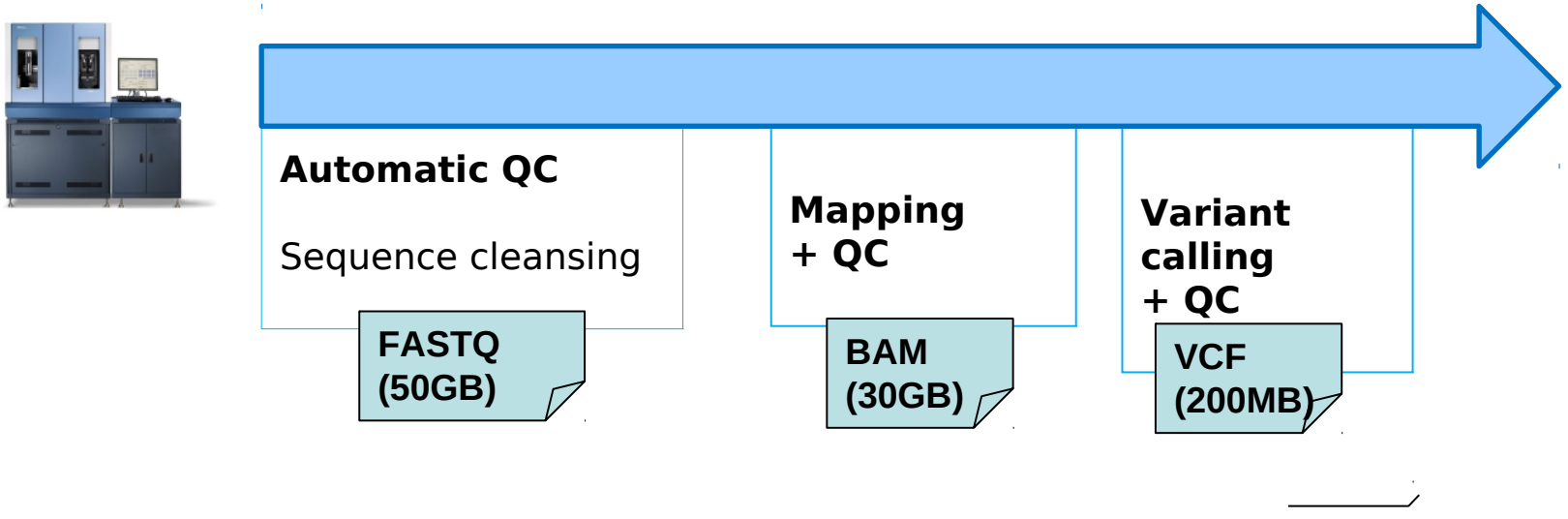


Some examples

- Low initial investment
- Already existent infrastructure
- Quick implementation
- Easily implementation as a cloud service that guarantees sustainability

	Conventional sequencing	**NGS (with capture)**
Marfan syndrome	1300€ 2 genes, 75 exons	900€ 3 genes, 237 exons
Hereditary deafness	12500€ 36 genes 1500 exons	1100€ 38 genes > 1500 exons

What are the real storage requirements?



Raw data 80GB/genome

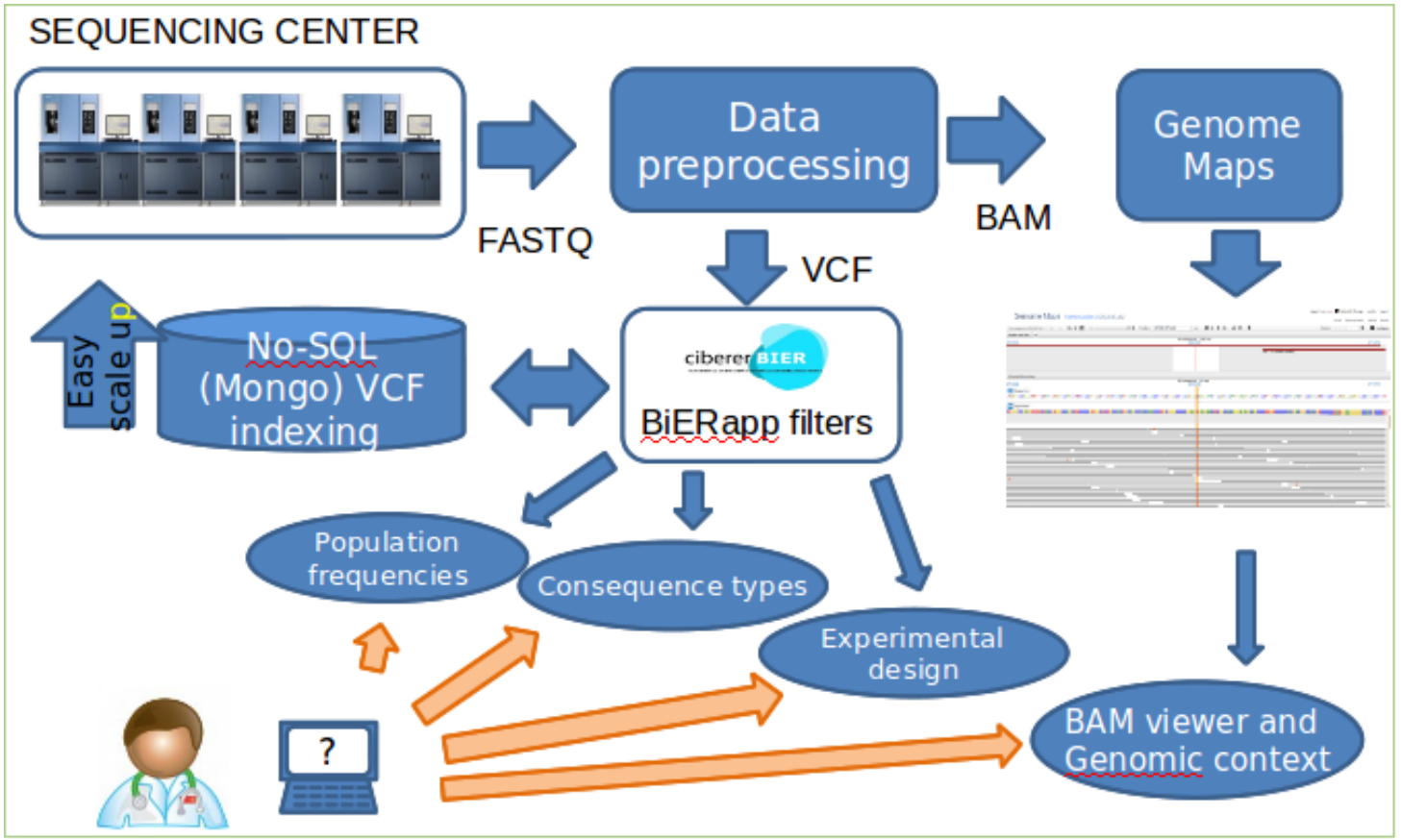
Processed data 200MB/genome

Hereditary diseases: **1 patient = 1 genome**
Cancer: **1 patient = 1 genome + x disease genomes**

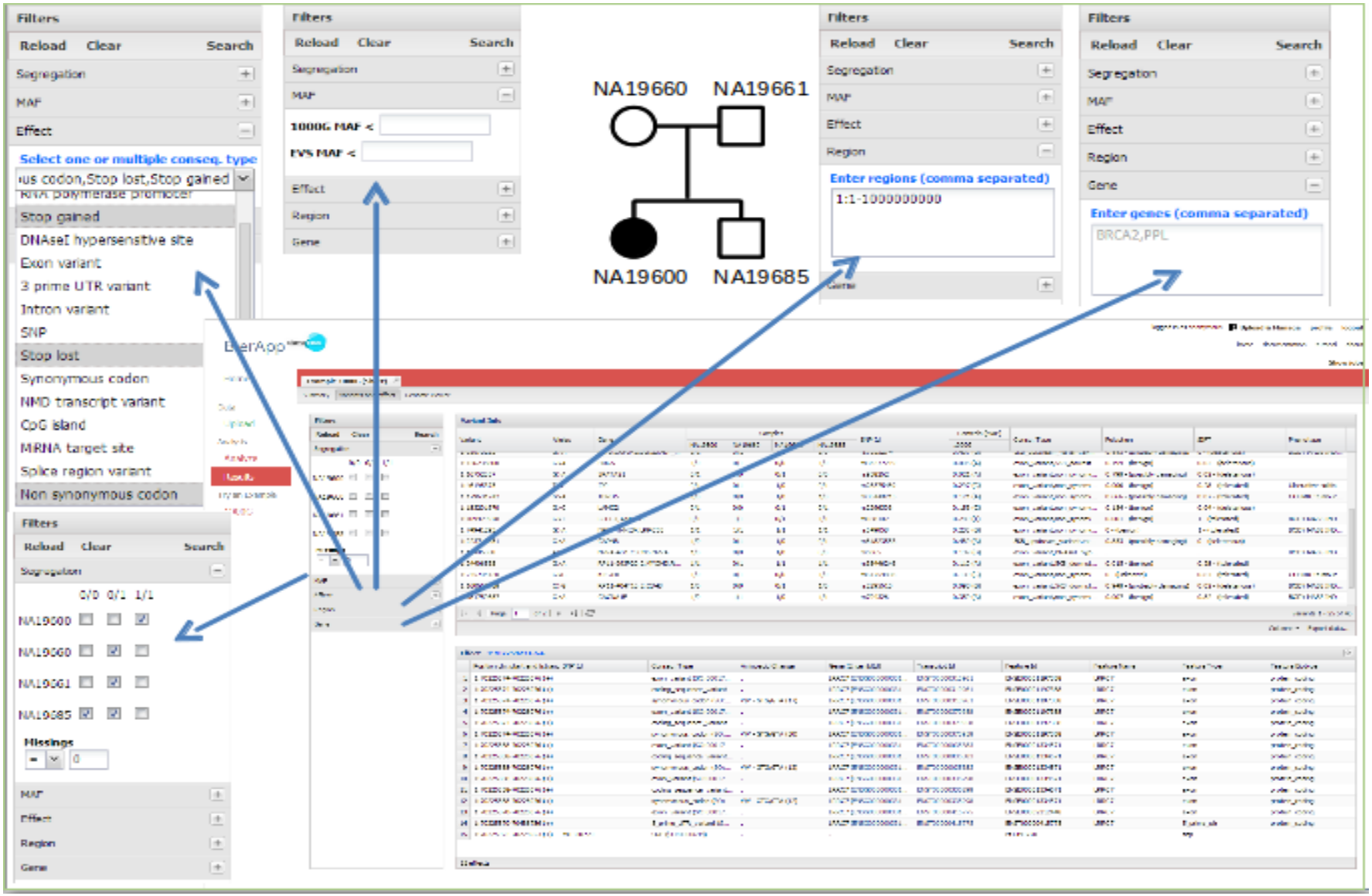
Now we store everything (>80GB/genome).
Once QC and software reach an acceptable standard of quality we will store only VCF files (or similar)

BiERapp: Discovering Variants

Interactive web-based tool for easy candidate prioritization by successive filtering



BiERapp: the interactive filtering tool for easy candidat priorization



New variants and disease genes found with WES and sucessive filtering

Human Mutation
Variation, Informatics, and Disease

Research Article

Two Novel Mutations in the *BCKDK* (Branched-Chain Keto-Acid Dehydrogenase Kinase) Gene Are Responsible for a Neurobehavioral Deficit in Two Pediatric Unrelated Patients

Angels García-Cazorla[1,†], Alfonso Oyarzabal[2,†], Joana Fort[3,†], Concepción Robles[4], Esperanza Castejón[5], Pedro Ruiz-Sala[2], Susanna Bodoy[3], Begoña Merinero[2], Anna Lopez-Sala[1], Joaquín Dopazo[6,7,8], Virginia Nunes[9], Magdalena Ugarte[2], Rafael Artuch[10,‡], Manuel Palacín[3,‡], Pilar Rodríguez-Pombo[2,‡,*], Working Group: (Patricia Alcalde[2], Rosa Navarrete[2], Paloma Sanz[2], Mariona Font-Llitjós[7], Mª Antonia Vilaseca[9], Aida Ormaizabal[9], Anna Pristoupilova[11], Sergi Beltran Agulló[11])

Issue

Human Mutation

Early View (Online Version of Record published before inclusion in an issue)

SEARCH

In this issue

Advanced > Saved Searches >

ARTICLE TOOLS

Get PDF (928K)

Save to My Profile

E-mail Link to this Article

Export Citation for this Article

Get Citation Alerts

Request Permissions

Share

Article first published online: 5 MAR 2014

DOI: 10.1002/humu.22513

© 2014 WILEY PERIODICALS, INC.

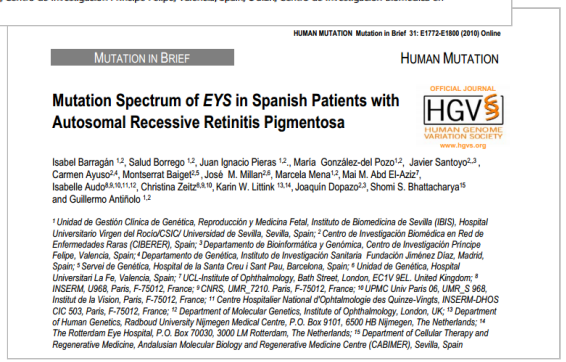
OPEN ACCESS Freely available online

PLOS ONE

Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies

Marta de Castro-Miró[1,2,3], Esther Pomares[1,2¤], Laura Lorés-Motta[1], Raul Tonda[4], Joaquín Dopazo[5,6], Gemma Marfany[1,2,3], Roser Gonzàlez-Duarte[1,2,3*]

1 Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain, 2 Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III, Barcelona, Spain, 3 Institut de Biomedicina (IBUB), Universitat de Barcelona, Barcelona, Spain, 4 Centre Nacional d'Anàlisi Genòmica, PCB, Barcelona, Spain, 5 Department of Computational Genomics, Centro de Investigación Príncipe Felipe, Valencia, Spain, 6 BIER, Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Valencia, Spain



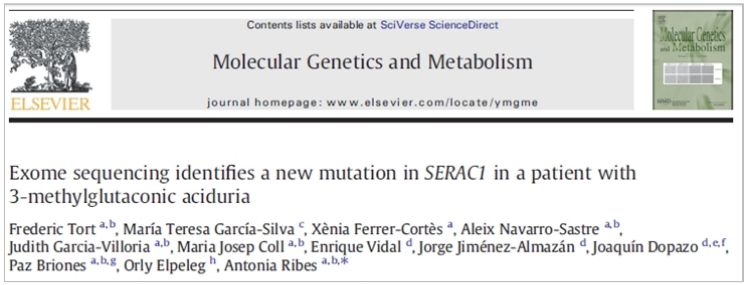
HUMAN MUTATION Mutation in Brief 31: E1772-E1800 (2010) Online

MUTATION IN BRIEF

HUMAN MUTATION

Mutation Spectrum of *EYS* in Spanish Patients with Autosomal Recessive Retinitis Pigmentosa

Isabel Barragán [1,2], Salud Borrego [1,2], Juan Ignacio Pieras [1,2], María González-del Pozo[1,2], Javier Santoyo[2,3], Carmen Ayuso[2,4], Montserrat Baiget[2,5], José M. Millan[2,6], Marcela Mena[1,2], Mai M. Abd El-Aziz[7], Isabelle Audo[8,9,10,11,12], Christina Zeitz[8,9,10], Karin W. Littink [13,14], Joaquín Dopazo[2,3], Shomi S. Bhattacharya[15] and Guillermo Antiñolo [1,2]



Contents lists available at SciVerse ScienceDirect

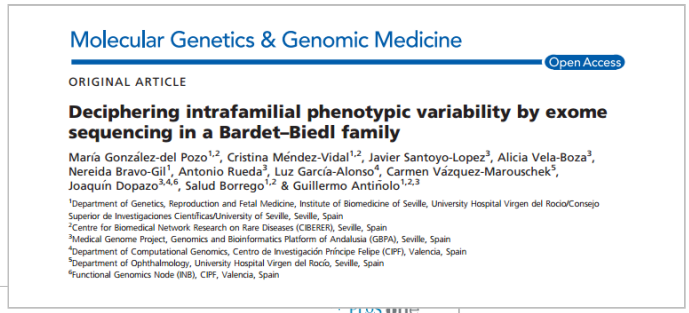
Molecular Genetics and Metabolism

journal homepage: www.elsevier.com/locate/ymgme

ELSEVIER

Exome sequencing identifies a new mutation in *SERAC1* in a patient with 3-methylglutaconic aciduria

Frederic Tort [a,b], María Teresa García-Silva [c], Xènia Ferrer-Cortès [a], Aleix Navarro-Sastre [a,b], Judith Garcia-Villoria [a,b], Maria Josep Coll [a,b], Enrique Vidal [d], Jorge Jiménez-Almazán [d], Joaquín Dopazo [d,e,f], Paz Briones [a,b,g], Orly Elpeleg [h], Antonia Ribes [a,b,*]



Molecular Genetics & Genomic Medicine

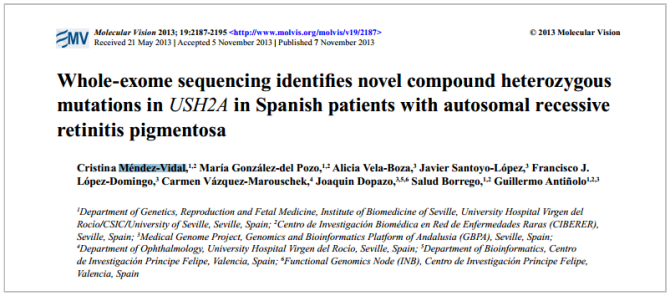
Open Access

ORIGINAL ARTICLE

Deciphering intrafamilial phenotypic variability by exome sequencing in a Bardet–Biedl family

María González-del Pozo[1,2], Cristina Méndez-Vidal[1,2], Javier Santoyo-Lopez[3], Alicia Vela-Boza[3], Nereida Bravo-Gil[1], Antonio Rueda[3], Luz García-Alonso[4], Carmen Vázquez-Marouschek[5], Joaquín Dopazo[3,4,6], Salud Borrego[1,2] & Guillermo Antiñolo[1,2,3]

[1]Department of Genetics, Reproduction and Fetal Medicine, Institute of Biomedicine of Seville, University Hospital Virgen del Rocío/Consejo Superior de Investigaciones Cientificas/University of Seville, Seville, Spain
[2]Centre for Biomedical Network Research on Rare Diseases (CIBERER), Seville, Spain
[3]Medical Genome Project, Genomics and Bioinformatics Platform of Andalusia (GBPA), Seville, Spain
[4]Department of Computational Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain
[5]Department of Ophthalmology, University Hospital Virgen del Rocío, Seville, Spain
[6]Functional Genomics Node (INB), CIPF, Valencia, Spain



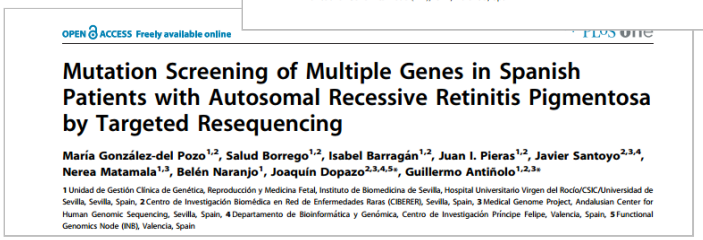
Molecular Vision 2013; 19:2187-2195 <http://www.molvis.org/molvis/v19/2187>
Received 21 May 2013 | Accepted 5 November 2013 | Published 7 November 2013

© 2013 Molecular Vision

Whole-exome sequencing identifies novel compound heterozygous mutations in *USH2A* in Spanish patients with autosomal recessive retinitis pigmentosa

Cristina Méndez-Vidal,[1,2] María González-del Pozo,[1,2] Alicia Vela-Boza,[3] Javier Santoyo-López,[3] Francisco J. López-Domingo,[3] Carmen Vázquez-Marouschek,[4] Joaquín Dopazo,[3,5,6] Salud Borrego,[1,2] Guillermo Antiñolo[1,2,3]

[1]*Department of Genetics, Reproduction and Fetal Medicine, Institute of Biomedicine of Seville, University Hospital Virgen del Rocío/CSIC/University of Seville, Seville, Spain;* [2]*Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Seville, Spain;* [3]*Medical Genome Project, Genomics and Bioinformatics Platform of Andalusia (GBPA), Seville, Spain;* [4]*Department of Ophthalmology, University Hospital Virgen del Rocío, Seville, Spain;* [5]*Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Valencia, Spain;* [6]*Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe, Valencia, Spain*



OPEN ACCESS Freely available online

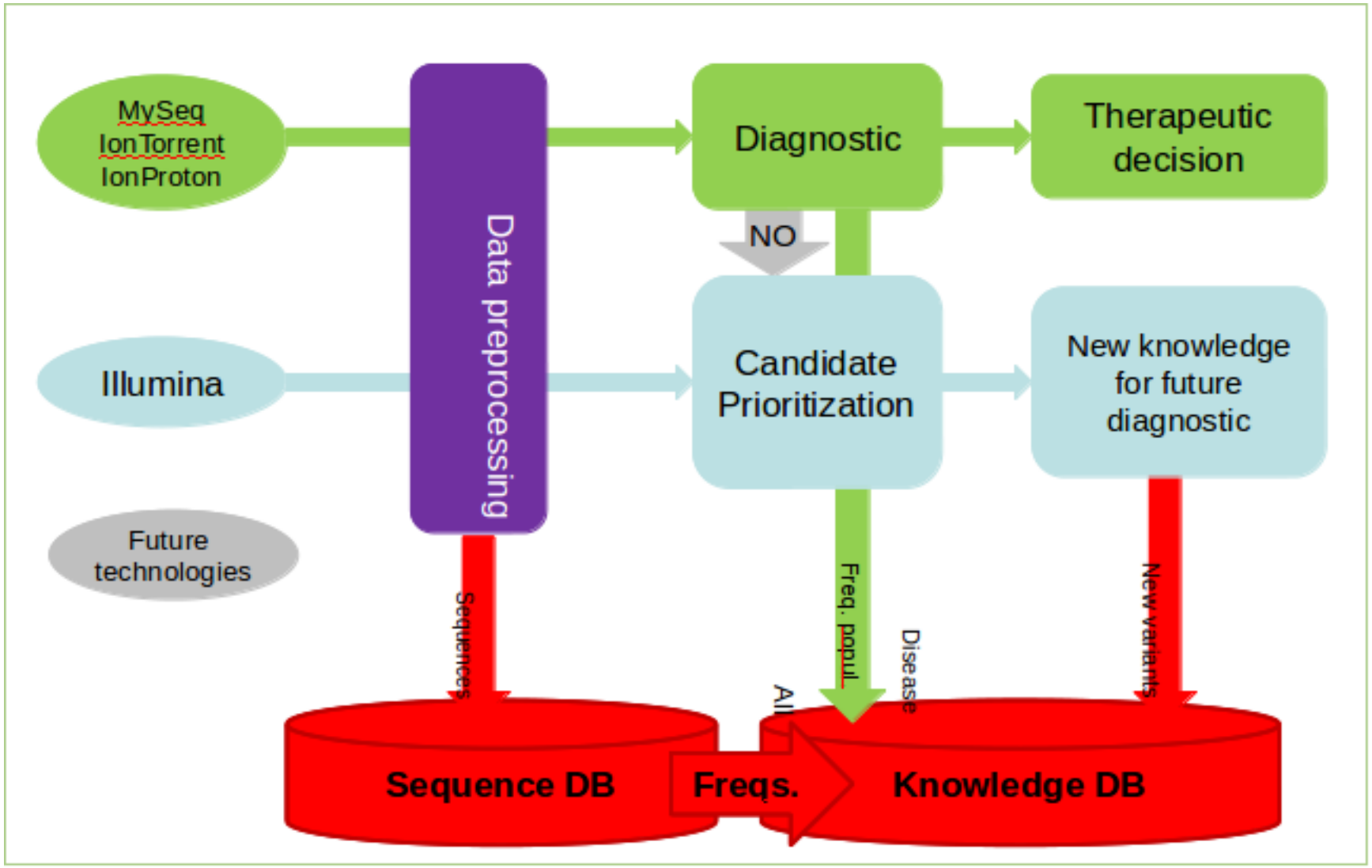
PLOS ONE

Mutation Screening of Multiple Genes in Spanish Patients with Autosomal Recessive Retinitis Pigmentosa by Targeted Resequencing

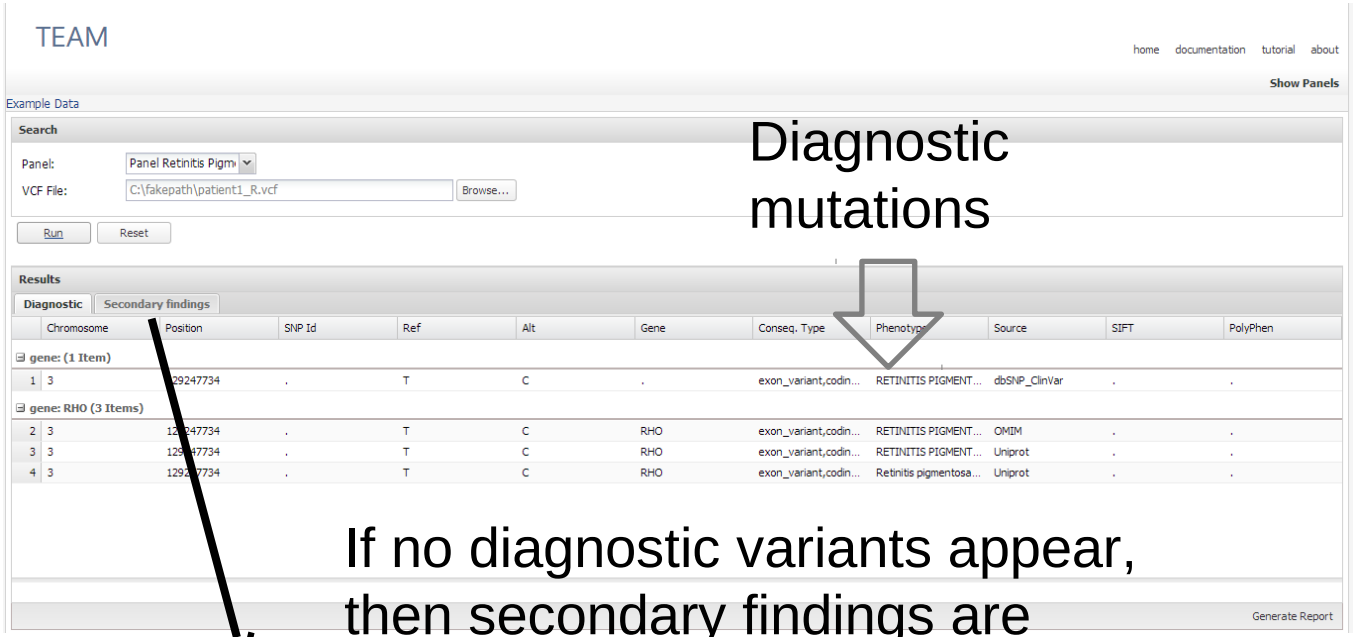
María González-del Pozo[1,2], Salud Borrego[1,2], Isabel Barragán[1,2], Juan I. Pieras[1,2], Javier Santoyo[2,3,4], Nerea Matamala[1,3], Belén Naranjo[1], Joaquín Dopazo[2,3,4,5*], Guillermo Antiñolo[1,2,3*]

1 Unidad de Gestión Clínica de Genética, Reproducción y Medicina Fetal, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Sevilla, Spain, 2 Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Sevilla, Spain, 3 Medical Genome Project, Andalusian Center for Human Genomic Sequencing, Sevilla, Spain, 4 Departamento de Bioinformática y Genómica, Centro de Investigación Príncipe Felipe, Valencia, Spain, 5 Functional Genomics Node (INB), Valencia, Spain

The final schema: diagnostic and discovery



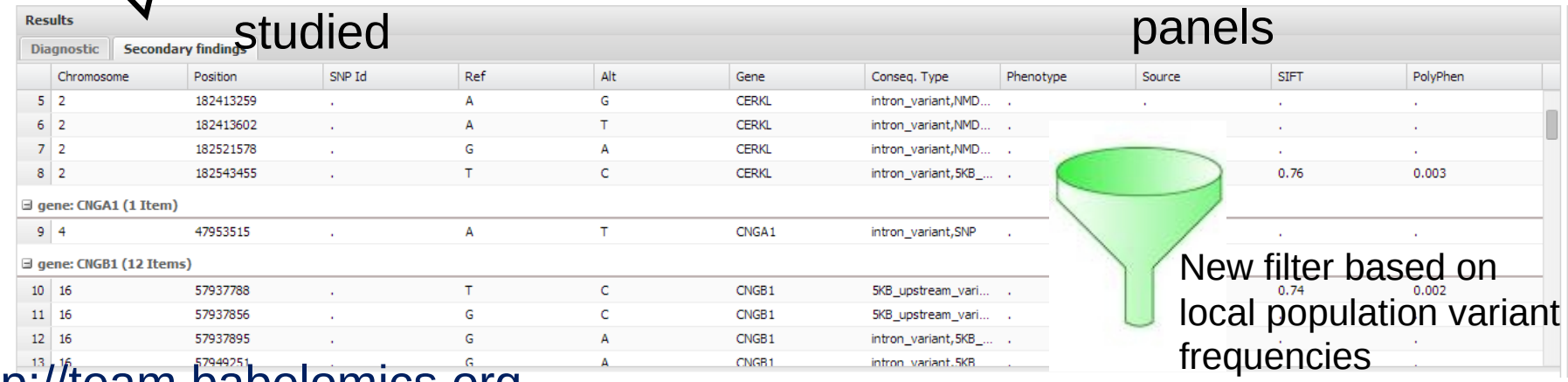
Diagnostic by targeted sequencing (panel of genes)



Diagnostic mutations

If no diagnostic variants appear, then secondary findings are studied

Tool for defining panels



New filter based on local population variant frequencies

http://team.babelomics.org

Future is now

Gattaca (1997)

Conclusions

- **NGS technologies** allow us to deal different problems in **omics scenarios**

- An optimal **pipeline for Genomics Data Analysis** + modules for **discovering and diagnostic** are **efficient resources** in several areas: Biology, Medicine...

- Genomics data are a **challenge in the big data context**