



3ª MESA: ¿QUÉ UTILIDAD OFRECE EL BIG DATA?



Fátima Al-Shahrour, PhD

Bioinformatics Unit

Structural Biology and Biocomputing Programme

Spanish National Cancer Research Institute. Centro Nacional de Investigaciones Oncológicas (CNIO)

¿QUÉ UTILIDAD OFRECE EL BIG DATA?

Perspectivas del uso del Big Data en Investigación Clínica

la medicina genómica

La investigación y la protección de la **salud** en la era del **Big data**:
¿Oportunidad o Mito?



Medicina Genómica



?

New Diagnosis?
New Treatments?



Human Genomic Variation

Genomic Basis of Human Diseases

Function of the Human Genome Sequence

Routine Whole-Genome Sequencing

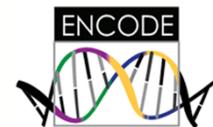
Routine Analysis of Genome Sequences



1000 Genomes
A Deep Catalog of Human Genetic Variation



GWAS



ENCODE



Molecular Biology Revolution

Genome Sequencing Revolution

Analysis Revolution

Medicine Revolution



>10 years ago

TODAY

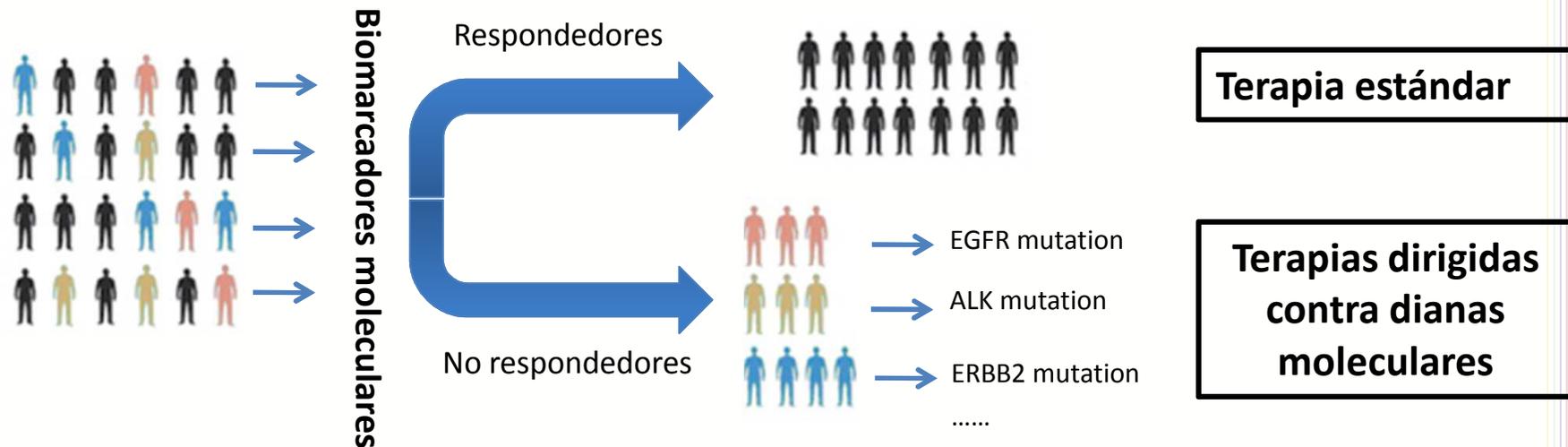
The future



La medicina de precisión en la oncología

La investigación
y la protección de la
salud en la era del
Big data
¿Oportunidad o Mito?

- La **medicina de precisión** usa el perfil molecular de un paciente y su información individual para guiar las decisiones con respecto a la prevención, diagnóstico y tratamiento del cáncer.



Proyectos del ICGC (septiembre 2016)

La investigación
y la protección de la
salud en la era del
Big data
¿Oportunidad o Mito?

UNITED STATES

- Bladder cancer (Invasive urothelial bladder cancer)
- Blood cancer (Acute lymphoblastic leukemia/Acute myeloid leukemia/Lymphoid neoplasm diffuse large B-cell lymphoma)
- Bone cancer (Osteosarcoma)
- Brain cancer (Glioblastoma multiforme/ Lower grade glioma)
- Breast cancer (Ductal & lobular)
- Cervical cancer (Squamous)
- Colorectal cancer (Adenocarcinoma)
- Endometrial cancer (Uterine corpus endometrial carcinoma)
- Gastric cancer (Adenocarcinoma)
- Head and neck cancer (Squamous cell carcinoma/ Thyroid carcinoma)
- Liver cancer (Hepatocellular carcinoma)
- Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)
- Ovarian cancer (Serous cystadenocarcinoma)
- Pancreatic cancer (Adenocarcinoma)
- Pediatric solid tumor (Neuroblastoma)
- Prostate cancer (Adenocarcinoma)
- Renal cancer (Clear cell carcinoma/ Clear cell sarcoma/Kidney chromophobe/Papillary carcinoma/Rhabdoid tumors)
- Skin cancer (Cutaneous melanoma)
- Soft tissue cancer (Liposarcoma and multiple subtypes)

CANADA

- Pancreatic cancer (Ductal adenocarcinoma)
- Pediatric brain tumors (Medulloblastoma)
- Prostate cancer (Adenocarcinoma)

EU / UNITED KINGDOM

- Breast cancer (ER positive, HER2 negative)

UNITED KINGDOM

- Bone cancer (Osteosarcoma/ chondrosarcoma/ rare subtypes)
- Breast cancer (Triple negative/lobular/ other)
- Chronic Myeloid Disorders (Myelodysplastic syndromes, myeloproliferative neoplasms and other chronic myeloid malignancies)
- Esophageal cancer (Adenocarcinoma)
- Prostate cancer (Adenocarcinoma)

GERMANY

- Blood cancer (Malignant lymphoma, Germinal center B-cell derived lymphomas)
- Lung cancer (Multiple rare subtypes)
- Pediatric brain tumors (Medulloblastoma and Pediatric pilocytic astrocytoma)
- Prostate cancer (Early onset)

ITALY

- Rare pancreatic tumors (Enteropancreatic endocrine tumors and rare pancreatic exocrine tumors)

CHINA

- Bladder cancer (Urothelial carcinoma)
- Blood cancer (Acute myeloid leukaemia and Chronic myelogenous leukaemia)
- Brain cancer (Glioblastoma multiforme)
- Breast cancer (Triple negative)
- Colorectal cancer (Adenocarcinoma, non-Western)
- Esophageal cancer (Squamous carcinoma)
- Gastric cancer (Intestinal- and diffuse-type)
- Liver cancer (Hepatocellular carcinoma, HBV-associated)
- Lung cancer (Squamous cell carcinoma)
- Nasopharyngeal cancer (Carcinoma, Asia)
- Ovarian cancer
- Pancreatic cancer (Ductal adenocarcinoma)
- Prostate cancer
- Renal cancer (Clear cell renal cell carcinoma)
- Thyroid cancer (Papillary carcinoma)

SAUDI ARABIA

- Thyroid cancer (Papillary carcinoma)

JAPAN

- Biliary tract cancer (Multiple histological subtypes)
- Gastric cancer (Adenocarcinoma and Multiple histological subtypes)
- Liver cancer (Hepatocellular carcinoma) (Virus-associated)

FRANCE

- Blood cancer (B-cell prolymphocytic leukemia)
- Bone cancer (Ewing sarcoma)
- Breast cancer (Subtype defined by an amplification of the HER2 gene)
- Eye cancer (Retinoblastoma)
- Liver cancer (Hepatocellular adenoma/ Hepatocellular carcinoma, Secondary to alcohol and adiposity/Hepatocellular macronodules)
- Prostate cancer (Adenocarcinoma)
- Soft tissue cancer (Leiomyosarcoma)
- Uterine cancer (Carcinosarcoma)

SINGAPORE

- Biliary tract cancer (Gall bladder cancer/ Cholangiocarcinoma)
- Blood cancer (T-cell and NK-cell lymphoma)

SOUTH KOREA

- Blood cancer (Acute myeloid leukaemia)
- Breast cancer (Asian phenotype/ Very young women)
- Colorectal cancer
- Gastric cancer
- Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)

MEXICO

- Blood cancer (Diffuse large B-cell lymphoma)
- Breast cancer (Ductal carcinoma)
- Cervical cancer
- Head and neck cancer (Squamous cell carcinoma of oral cavity/oropharynx/ sinonasal cavity/ hypopharynx/larynx)
- Pediatric solid tumors

SPAIN

- Blood cancer (Chronic lymphocytic leukemia with mutated and unmutated IgVH)

BRAZIL

- Skin cancer (Melanoma)

EU / FRANCE

- Renal cancer (Renal cell carcinoma) (Focus on but not limited to clear cell subtype)

INDIA

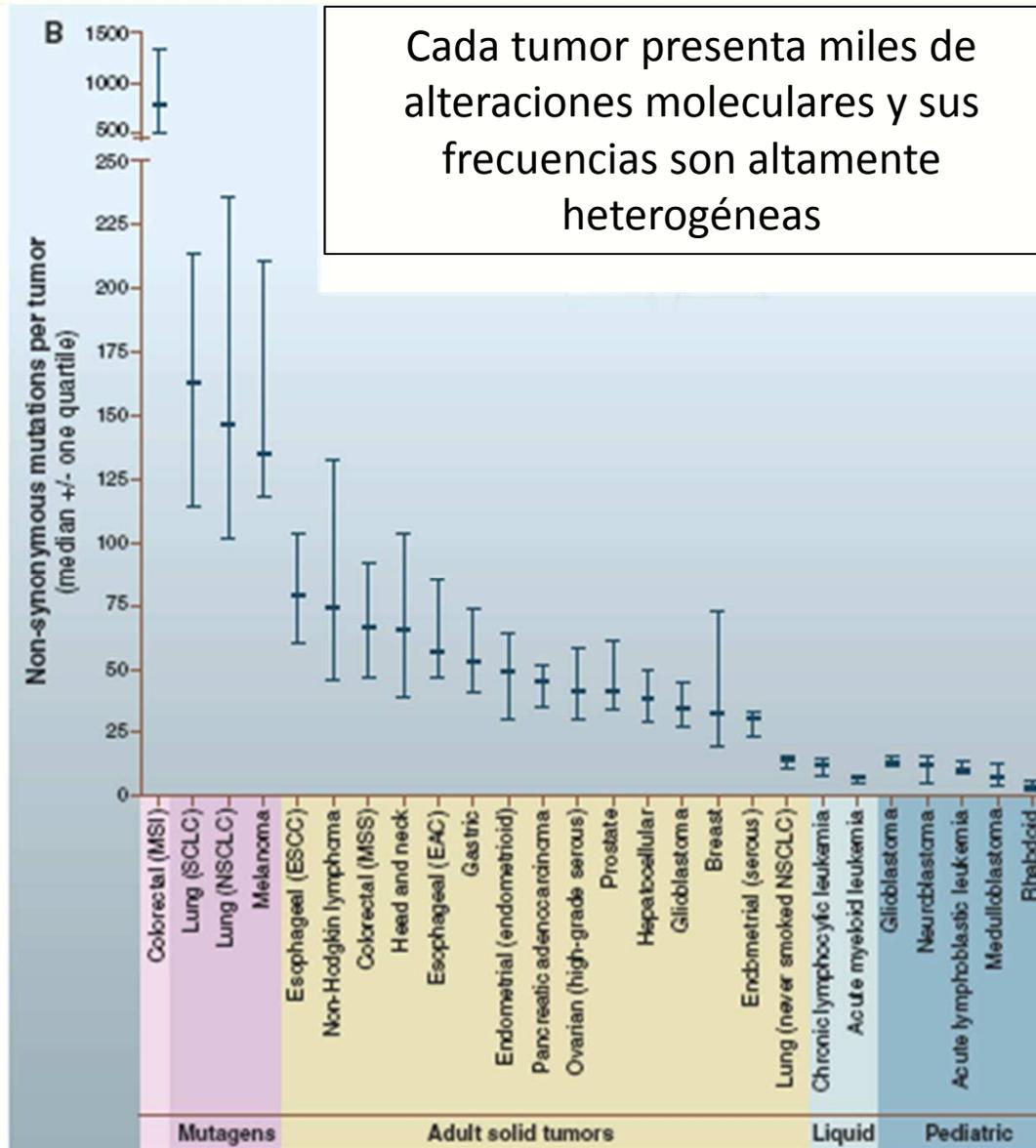
- Oral cancer (Gingivobuccal)

AUSTRALIA

- Ovarian cancer (Serous cystadenocarcinoma)
- Pancreatic cancer (Ductal adenocarcinoma/ Endocrine neoplasms)
- Prostate cancer
- Skin cancer (Melanoma)

Grey = Collaboration

El genoma del cáncer



Biomarcadores y nuevas terapias dirigidas

La investigación
y la protección de la
salud en la era del
Big data
¿Oportunidad o Mito?

Drug	Company	Type	Target	Status
▶ ARQ 736	ArQule	small molecule	BRAF	Phase I
▶ Selumetinib	Array BioPharma	"	MEK	Phase II
▶ PD-0325901	Pfizer	"	MEK	Phase I
▶ PX-866	Oncothyreon	"	PI3K	Phase II
▶ BEZ235	Novartis	"	PI3K	Phase I
▶ BKM120	Novartis	"	PI3K	Phase II
▶ Tivantinib	ArQule	"	MET	Phase II



Biomarcadores farmacogenómicos en Oncología:

En 2013 -> 39

En 2015 -> 55

En 2017 -> 81

▶ Ramucirumab	ImClone Systems	"	VEGFR-2	Phase II
▶ AMG 386	Amgen	biologics (peptibody)	Tie-2R	Phase II
▶ GL-ONC1	Genelux	biologics (oncolytic virus)	Tumoral cells	Phase I/II
▶ ColoAd1	PsiOxus	"	"	Phase I/II
▶ NV1020	Medigene	"	"	Phase II
▶ Reolysin	Oncolytics	"	"	Phase I/II
▶ JX594	Jennerex	"	"	Phase I/II

Simon R et al Nat Rev Drug Discov 2013
Wheeler HE et al. Nat Rev Genet 2013

Cost per Genome

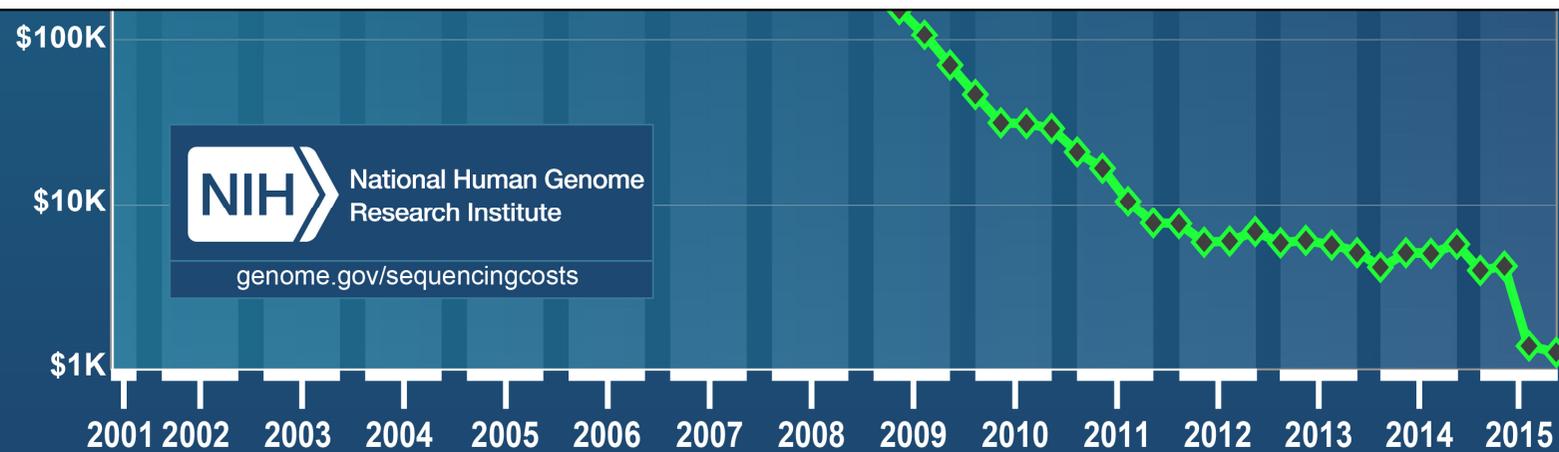
Mardis Genome Medicine 2010, 2:84
<http://genomemedicine.com/content/2/1/84>



MUSINGS

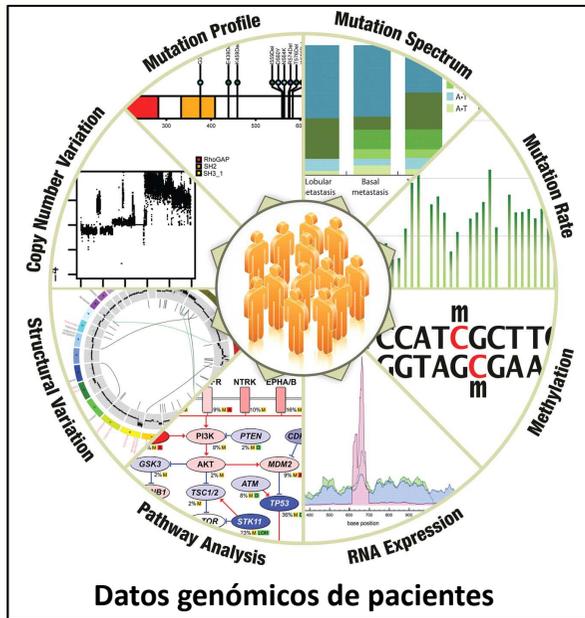
The \$1,000 genome, the \$100,000 analysis?

Elaine R Mardis*



Genómica integrativa

La investigación y la protección de la **salud** en la era del **Big data**:
¿Oportunidad o Mito?



Anotaciones clínicas

Hospital Universitario Ramón y Cajal
Hospital Universitario 12 de Octubre
Hospital Universitario Clínico San Carlos
Hospital Universitario de Fuenlabrada
Hospital Universitario La Paz

Datos experimentales in vitro e in vivo

Diagram of a genetic construct: shRNA, pLKO1, and pBS ori.

Bases de datos

PubMed, GEO, OMIM, CCLC Cancer Cell Line Encyclopedia, COSMIC Catalogue of somatic mutations in cancer, THE CANCER GENOME ATLAS, International Cancer Genome Consortium, International HapMap Project, 1000 Genomes, ENCODE, KEGG.

Análisis Bioinformático

cnio Centro Nacional de Investigaciones Oncológicas, Bioinformatics CNIQ Bioinformatics Unit, Welcome to PANDRUGS A novel method for prioritizing therapies using individual genomic data.

Identificación de variantes

Gen-proteína predicción de función

Farmacogenómica: Dianas terapéuticas y biomarcadores

Aplicación → Interpretación clínica → Medicina de precisión

Limitaciones y retos

La investigación
y la protección de la
salud en la era del
Big data
¿Oportunidad o Mito?

Generación del “informe genómico” del paciente

Eficiencia en tiempo y coste (*real time*)

Área tecnológica

Generación datos
(plataformas “ómicas”)

Almacenamiento de
datos seguro y anónimo

Análisis computacional

Acceso y visualización

Área clínica

Recogida muestras

Historias clínicas
electrónicas

Data manager

Data sharing

Área recursos humanos

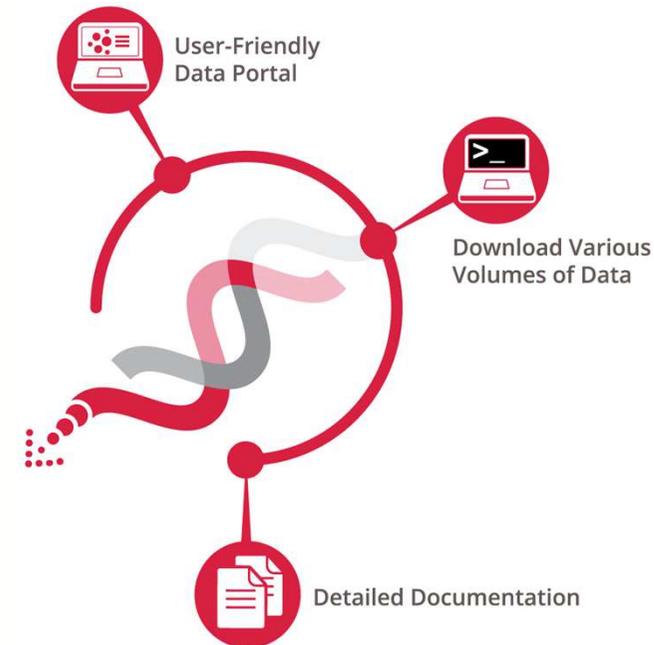
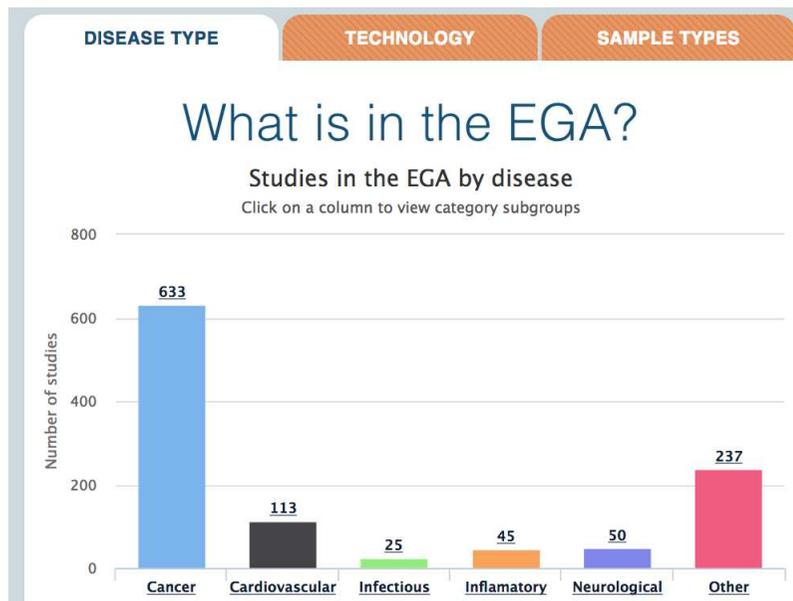
Equipos
multidisciplinares

Analista de datos

Formación en genética
& bioinformática

Infraestructuras y tecnología

La investigación
y la protección de la
salud en la era del
Big data
¿Oportunidad o Mito?



EGA es un servicio para archivar y compartir permanentemente todos los tipos de datos genéticos y fenotípicos resultantes de proyectos de investigación biomédica.

GDC permite almacenar, compartir y analizar los datos incluyendo la información clínica y la caracterización genómica

Infraestructuras y tecnología

La investigación
y la protección de la
salud en la era del
Big data
¿Oportunidad o Mito?

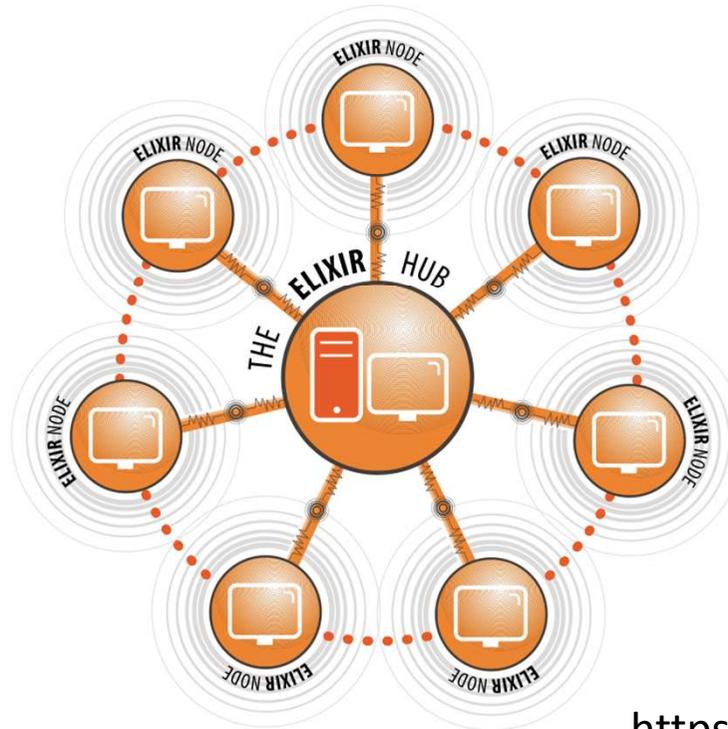


A distributed infrastructure to scale with the challenge

ELIXIR data infrastructure for Europe's life science research sector

ELIXIR Nodes build local bioinformatics capacity throughout Europe

ELIXIR Nodes build on national strengths and priorities



<https://www.elixir-europe.org/>

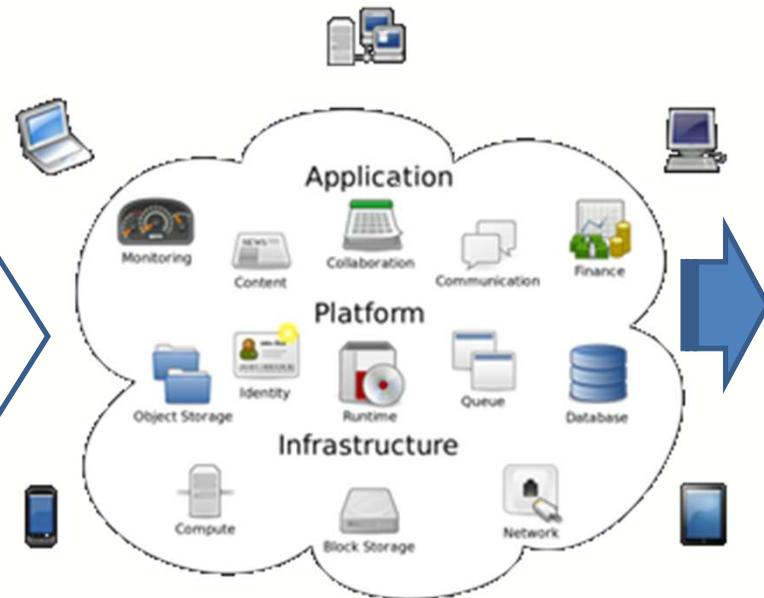
Área clínica

La investigación
y la protección de la
salud en la era del
Big data:
¿Oportunidad o Mito?



Cohortes de pacientes prospectivas y retrospectivas
200.000 pacientes en 2024

- Datos demográficos
- Factores ambientales
- Datos genómicos
- Datos clínicos
- Tratamientos
- Evolución



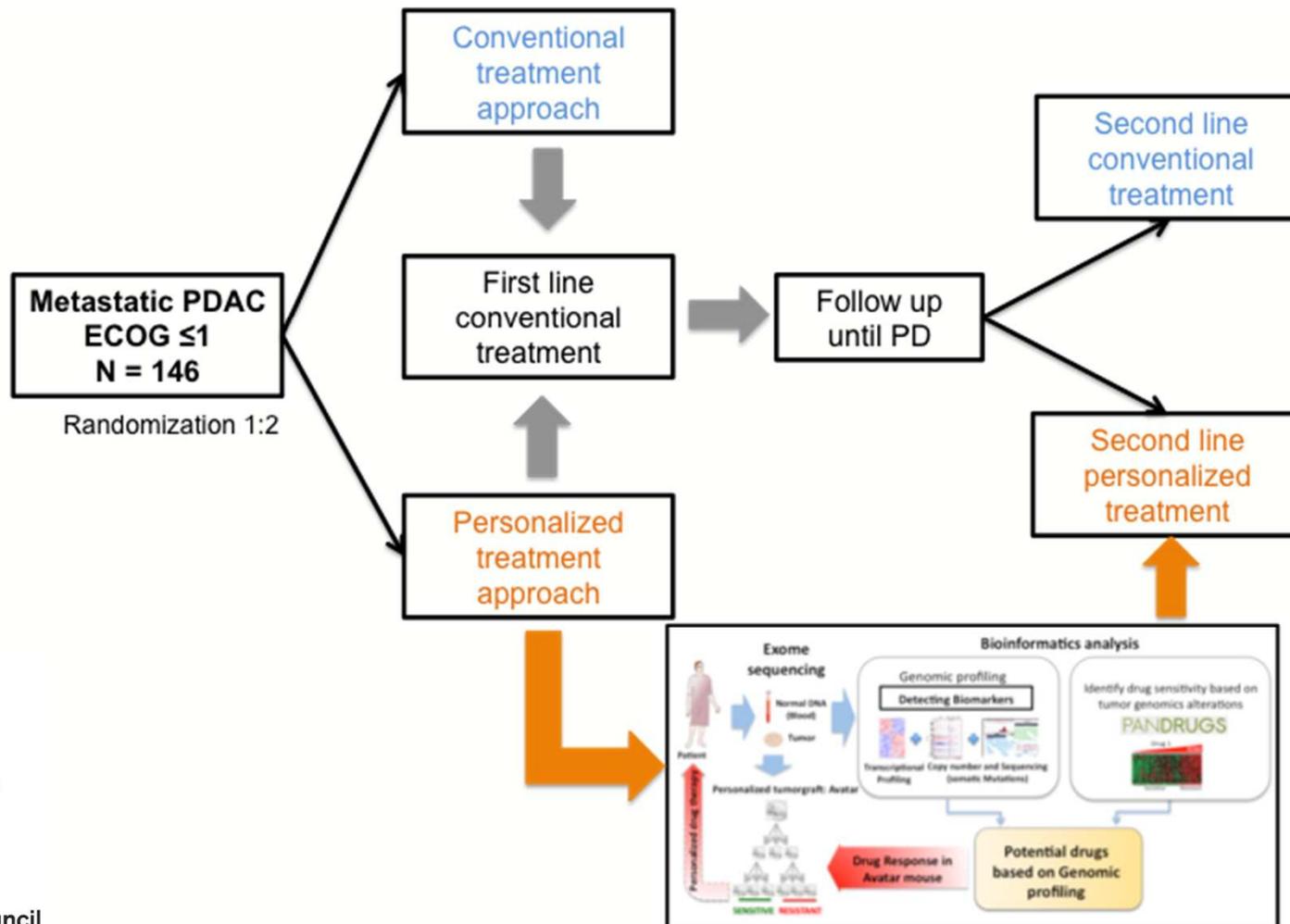
Correlación entre genómica (y biología) y la evolución de la enfermedad

El conocimiento vendrá de la agregación de la diversidad (enfermedades, etnias, geografía, exposiciones/ambiente, tratamientos)
¡Y compartir!

Medicina de precisión en PDAC

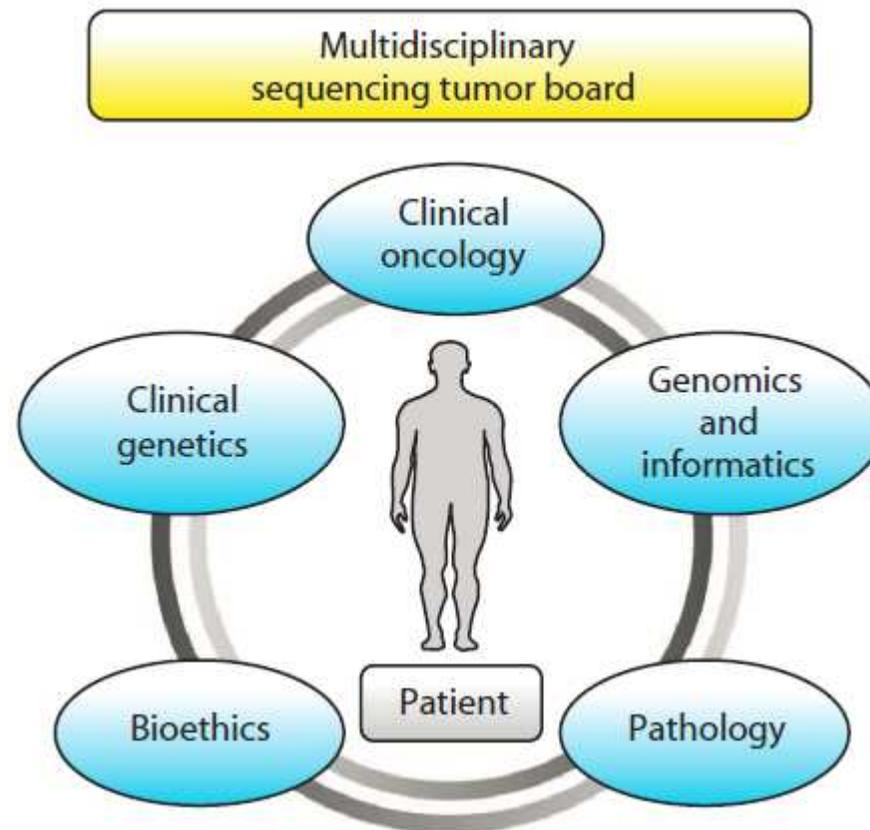
La investigación y la protección de la salud en la era del **Big Data**:
¿Oportunidad o Mito?

- Ensayo clínico multicéntrico aleatorizado en pacientes con cáncer de páncreas metastásico.
- Tratamiento personalizado basado en la integración de datos genómicos: ¿mejora al tratamiento convencional?



Recursos humanos

- Enfoque multidisciplinario para la práctica médica.
- Mayor interacción entre oncólogos, patólogos, genetistas, biólogos moleculares y bioinformáticos.



Riesgo
Prevención
Diagnóstico
Pronóstico
Tratamiento

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants



DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



Conclusiones



La investigación
y la protección de la
salud en la era del
Big Data:
¿Oportunidad o Mito?

- Las nuevas tecnologías han generado una enorme cantidad de datos: era del Big Data.
- Estos datos han de ser producidos de manera rápida, eficiente y accesible.
- Son necesarios grandes proyectos colaborativos internacionales para fijar los estándares.
- Implementación de plataformas ómicas en el entorno hospitalario.
- Equipos multidisciplinares.
- Formación del personal clínico.
- Compartir datos (moleculares y clínicos) es la clave

¡Gracias por su atención!

