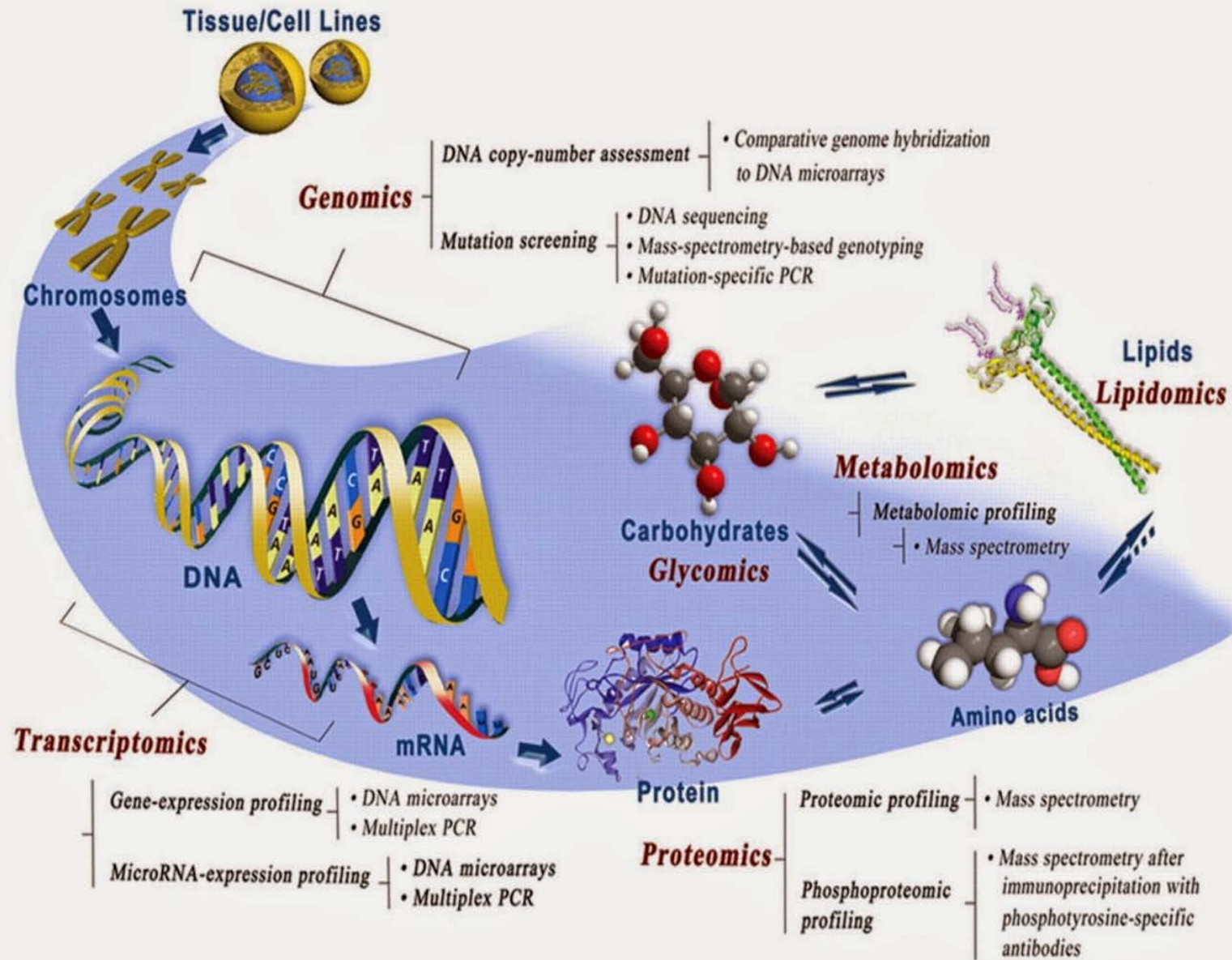


Omics technologies: genomics, transcriptomics, metabolomics, databases and big data

Vendula Pospíchalová, PhD
(pospich@sci.muni.cz)
Department of Experimental Biology
Animal Physiology and Immunology

Bi5599 Applied Biochemistry and Cell Biology Methods
2019-10-30



Schematic representation of omics technologies, their corresponding analysis targets, and assessment methods. Taken from Wu RD et al. JDR 2011; 90:561-572.

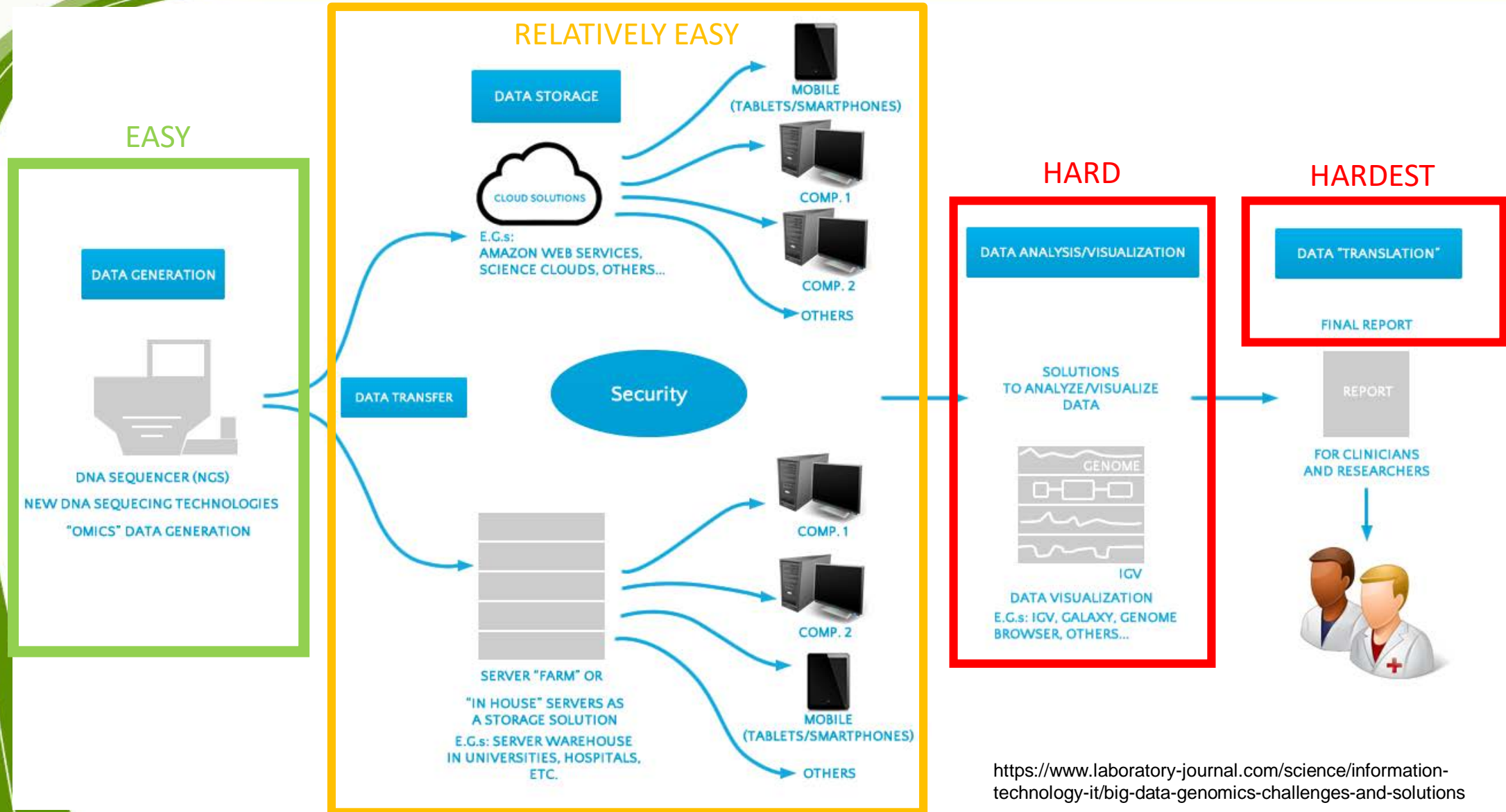
Contents

1. Introduction: what are –omics technologies + history
2. How does big data look and how to approach it
3. From –omics technologies to biomarkers and personalized medicine
4. Genomics: genomes vs exomes vs genotypes + DTC service
5. Cancer databases: COSMIC, TCGA, Oncomine and others
6. Transcriptomics: microarrays vs. RNA sequencing
7. Gene set analysis
8. Metabolomics
9. Cutting edge: single cell –omics and single cell multi-omics
10. Summary and take home messages

What are „-omics“ technologies

- **Omics** refers to a field of study in biology ending in **-omics**, such as genomics, proteomics or metabolomics
- The related suffix **-ome** is used to address the objects of study of such fields, such as the genome, proteome or metabolome
- **-ome** = many/collectivity in Latin, **-omics** = study of large sets of biomolecules
- **High-throughput experimental technologies** characterized by **automation, miniaturized assays** and **large-scale data analysis**
- Analytic part of the experiment is usually much longer than the experiment itself – bioinformatics skills needed
- Raw data is the „gem“ but usually is in user unfriendly format
- **Interpreting** functional consequences of millions of discovered events is one of **the biggest challenges**

Big -omics data challenges

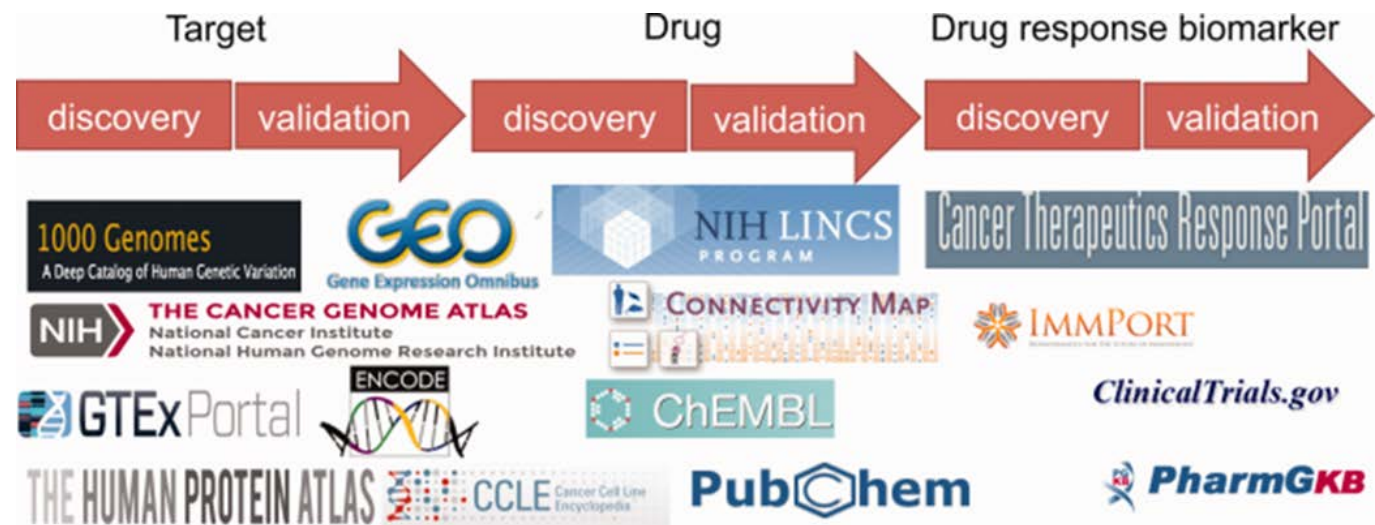


Data sharing policy

- The concepts of data sharing and open data are becoming increasingly important in science
- Funding bodies, journals and societies are now encouraging or mandating data sharing (usually the raw data)
- Sharing data publicly is an important way of improving reproducibility and showing that researchers are confident in their work
- Studies with raw data shared in a repository also receive more citations than those without publicly available data

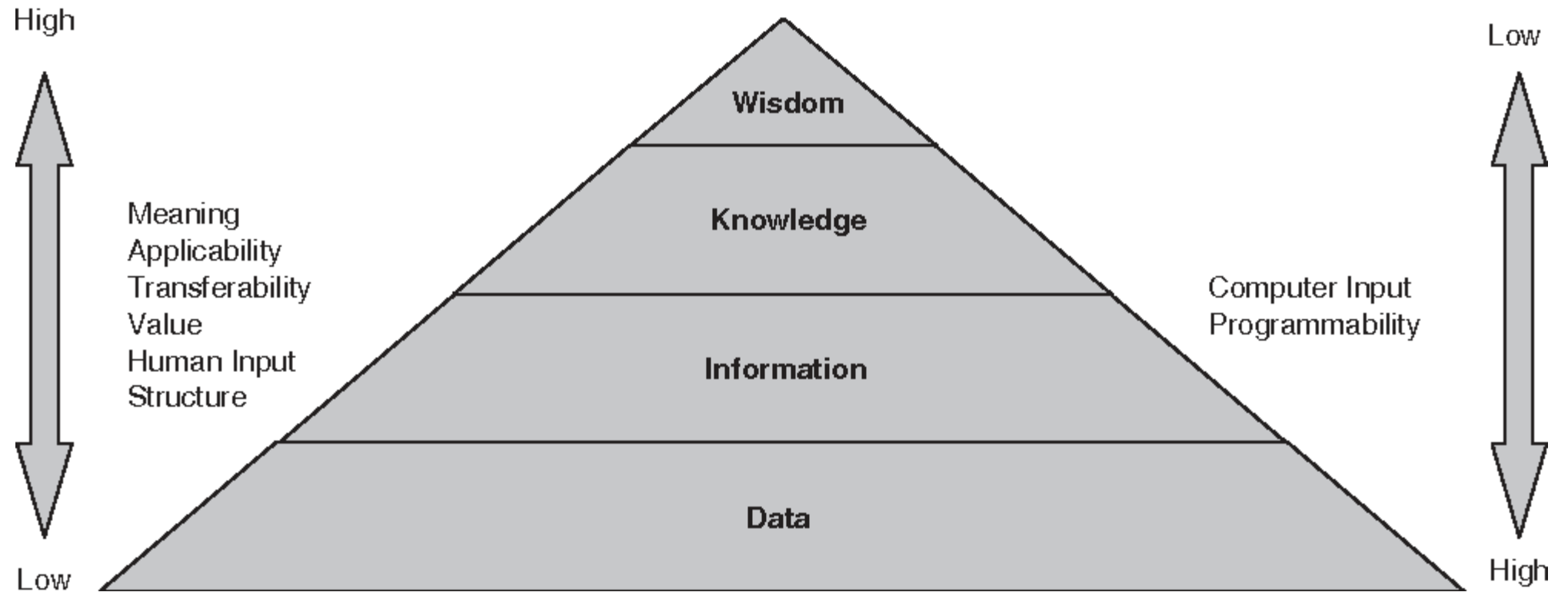
- But raw -omics data are hard to analyse, so many platforms gather the publicly available data, thoroughly analyze it, curate it and share it in a user friendly format

Leveraging Public Databases to Identify Actionable Targets



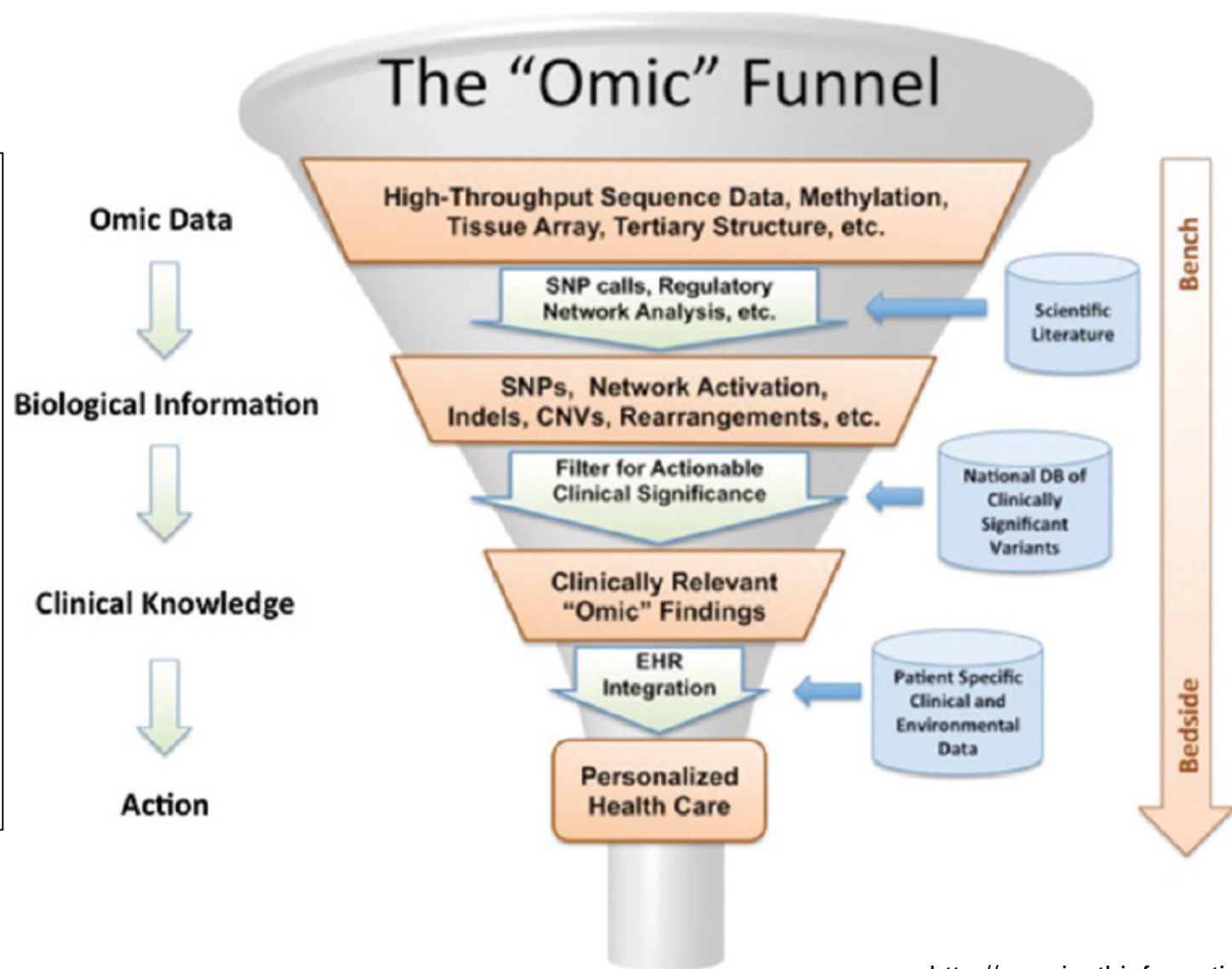
DIKW pyramide

„Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.“ – Clifford Stoll



What is the aim of OMICS technologies

In Genomics:
 Sequence of 3 billions letters in .txt file
 ↓
 information where the individual's genome varies from reference sequence
 ↓
CYP2C9 or *TPMT* genotype, which has known pharmacogenomic associations
 ↓
 individualize the dose of a new warfarin prescription



The DIKW pyramid metaphor:

"know-nothing" (Data)
 ↓
 "know-what" (Information)
 ↓
 "know-how" (Knowledge)
 ↓
 "Know-why" (Wisdom)

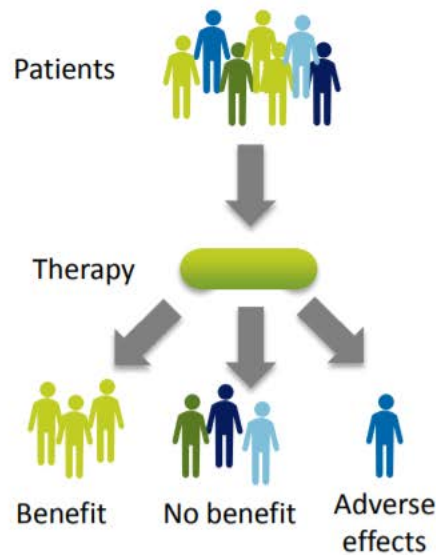
Zeleny (2005)

What is personalized health care?

Personalized medicine, sometimes referred to as *precision* or *individualized* medicine, is an emerging field of medicine that uses diagnostic tools to identify specific biological markers, often genetic, to help assess which medical treatments and procedures will be best for each patient.

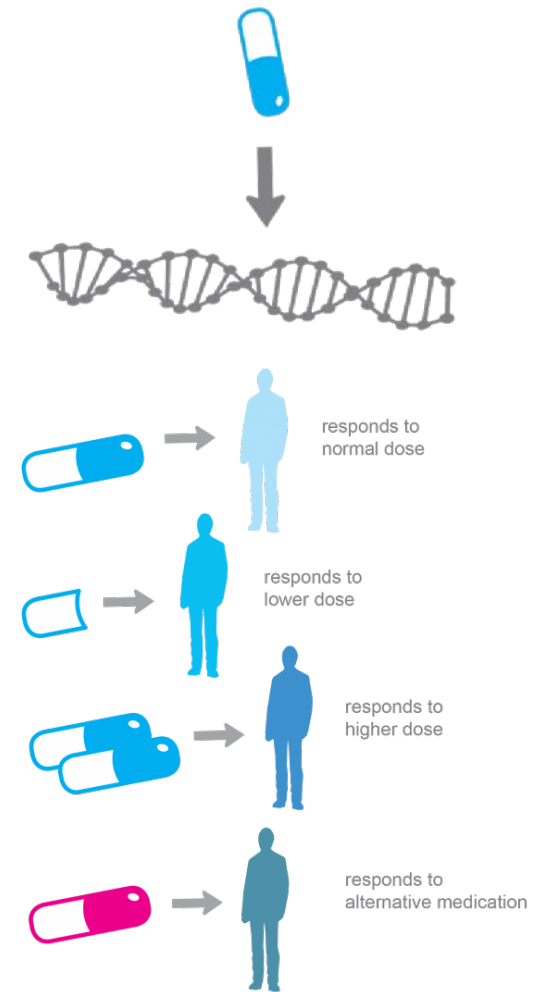
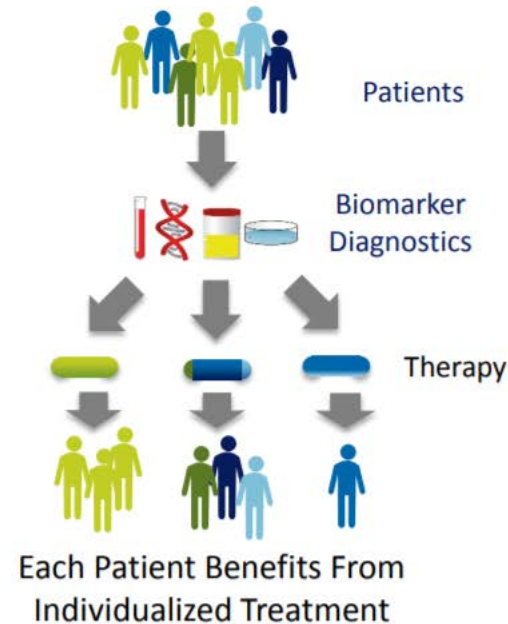
Without Personalized Medicine:

Some Benefit, Some Do Not



With Personalized Medicine:

Each Patient Receives the Right Medicine For Them



Value of personalized medicine



\$5 billion

Estimated annual cost of wasted prescription drugs in the US.

\$3 billion

Estimated cost of wasted hospital cancer drugs.



27%

of all NMEs approved by the FDA in 2016 are personalized medicines.



50%

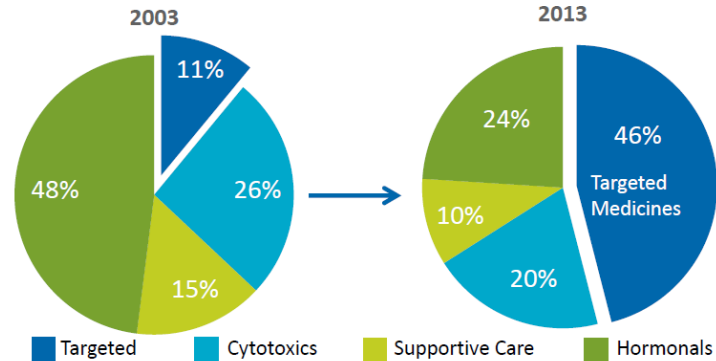
of the personalized medicines approved by the FDA in 2016 are oncology drugs.

<https://invivo.pharmaintelligence.informa.com/>

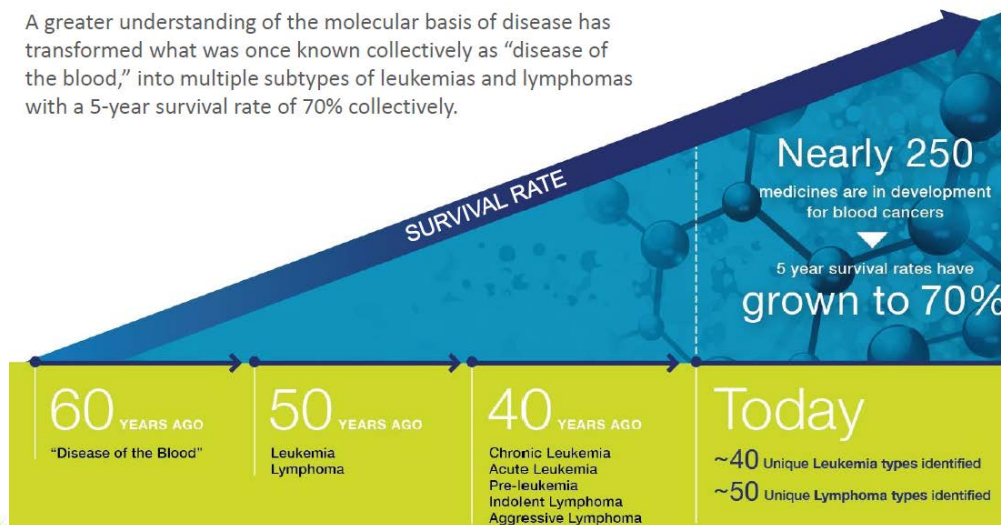
Oncology is on the Leading Edge of Personalized Medicine

In ten years, cancer patients have seen a four-fold increase in their personalized medicine treatment options.

*Breakdown of Oncology Treatment Modalities, Global Market share 2003-2013**



A greater understanding of the molecular basis of disease has transformed what was once known collectively as “disease of the blood,” into multiple subtypes of leukemias and lymphomas with a 5-year survival rate of 70% collectively.



Personalized Medicine Can Create Efficiencies in the Health Care System

Breast Cancer



Reduction in chemotherapy use would occur

If women with breast cancer receive a genetic test of their tumor prior to treatment.

Metastatic Colorectal Cancer



In annual health care cost savings would be realized

If patients with metastatic colorectal cancer receive a genetic test for the KRAS gene prior to treatment.

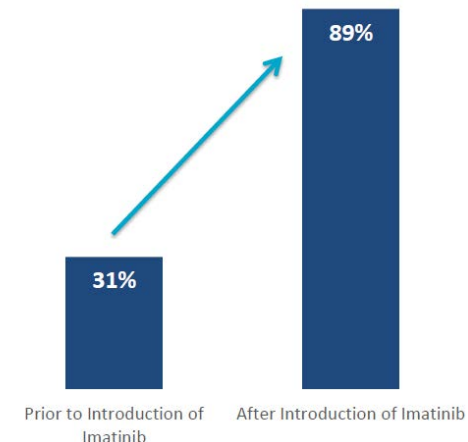
Stroke



Strokes could be prevented each year

If a genetic test is used to properly dose blood thinners

5-Year Survival Rates for CML Patients Nearly Triple After Introduction of Imatinib

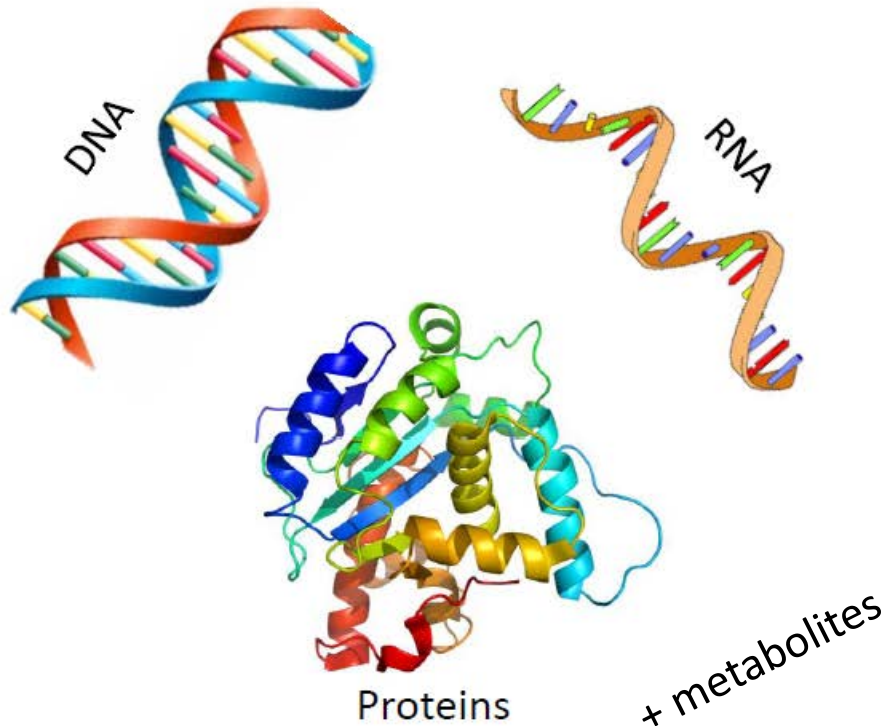


And many more examples, see <http://www.personalizedmedicinecoalition.org> for more detailed information on PM

What is a biomarker?

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, disease processes, or biological responses to a therapeutic intervention. Biomarkers can be used to reduce uncertainty and guide clinical care.

Molecular Biomarkers Can Include:



Biomarkers Help Inform Medical Decisions:

- Prevention measures?
- Which diagnosis?
- Treat or don't treat?
- What dose?

How Do You Detect a Biomarker?

- Diagnostics
 - Blood draw
 - Microscopic analysis
 - Gene sequencing
 - Biopsy
 - Protein analysis

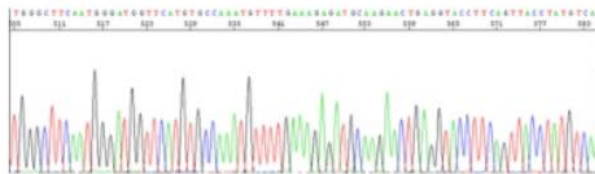
-OMICS technologies and their integration is crucial for biomarker discovery and validation

History of „-omics“ technologies

- Genome – central part of all –omics technologies
- NGS = next generation sequencing

Sanger VS NGS

	Bases	Genes
Human Genome	3.3×10^9	~20,000



Sequencing of the human genome using Sanger technology took more than a decade and cost an estimated \$70 million dollars



In 3 days (one run), Illumina HiSeq 4000 is able to produce $1,680 \times 10^9$ bases for ~\$32,000

Brief History of DNA Sequencing

1953: Discovery of DNA structure by Watson and Crick

1973: First sequence of 24 bases published

1977: Sanger sequencing method published

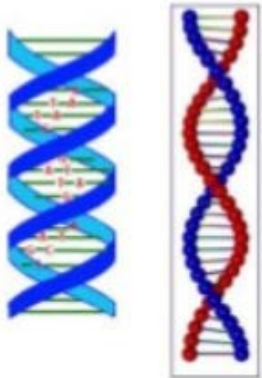
1982: GenBank started

1987: 1st automated sequencer: Applied Biosystems Prism 373 (up to 600 bases)

1996: First Capillary sequencer: ABI310

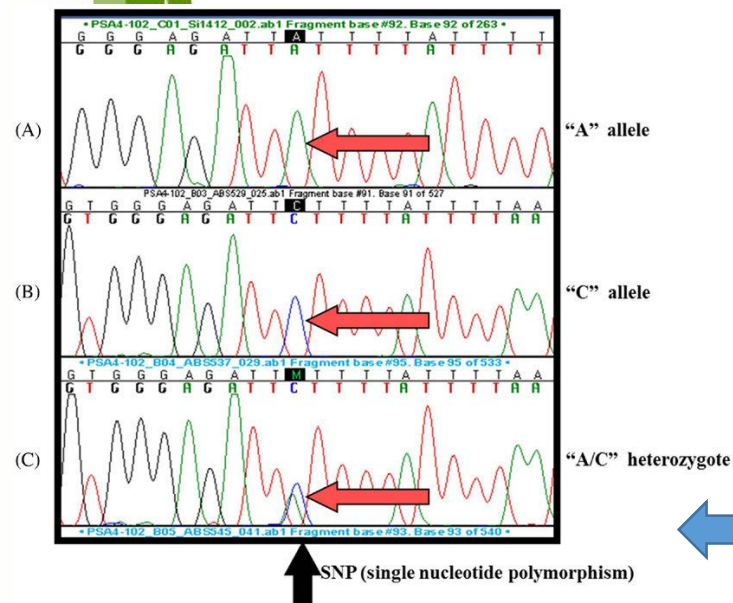
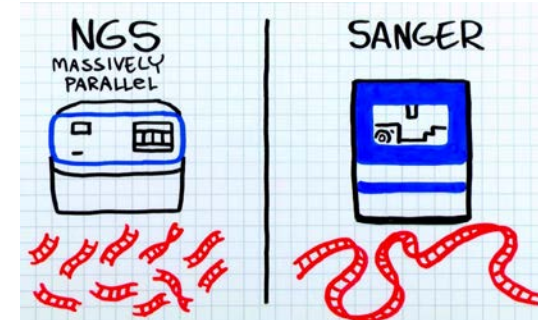
2000-2003: Human Genome Sequenced

2005- : First NGS sequencers 454 Life Sciences, Solexa/Illumina, Helicos, Ion Torrent



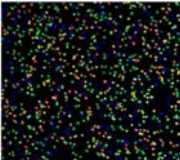
Sanger vs next generation sequencing

- Sanger sequencing
 - <https://www.youtube.com/watch?v=e2G5zx-OJlw>
- Next generation sequencing (Illumina is shown as an example)
 - <https://www.youtube.com/watch?v=9YxExTSwgPM>

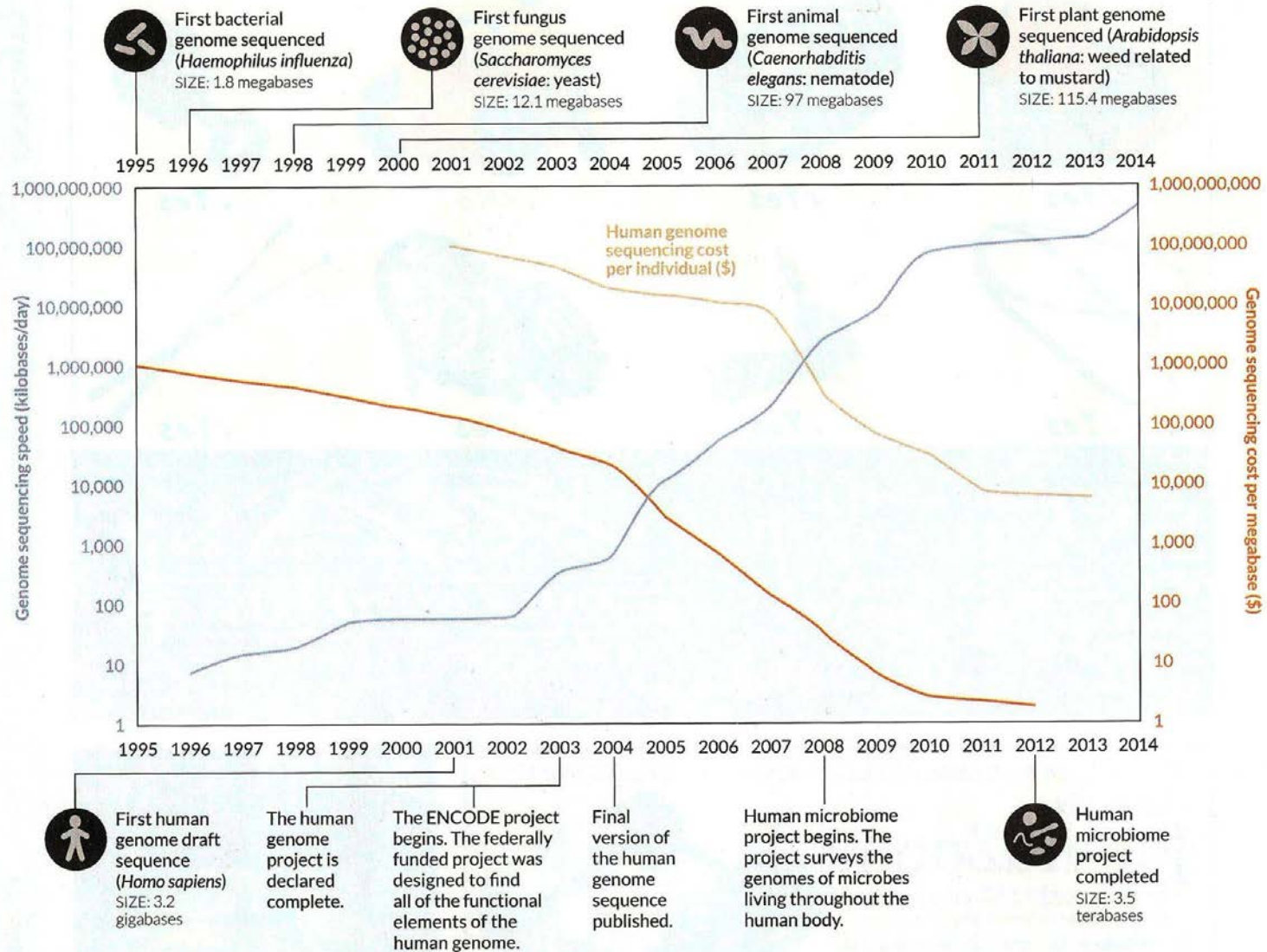


Sanger Sequencing

Illumina Sequencing

Sanger Sequencing		Illumina Sequencing	
Advantages	Disadvantages	Advantages	Disadvantages
Lowest error rate (1.5%)	High cost per base	Low error rate	Must run at very large scale
Long read length (~750 bp)	Long time to generate data	Lowest cost per base	Short read length (50-75 bp)
Can target a primer	Need for cloning	Tons of data	Runs take multiple days
Used to confirm NGS results	Amount of data per run	 An image of hundreds of extended molecules	High startup costs
Seeing is believing			De Novo assembly difficult

Cost of sequencing over time



DTC (direct-to-customer) genetic testing

ancestry

FREE TRIAL SIGN IN >



Give the gift that has connected
20 million members to a deeper family story.

ONLY **\$59*** ~~\$99~~

Give AncestryDNA®

*Offer ends 11/21. Excludes taxes and shipping.

 Build a family tree to see your story emerge.

Learn more


Genotyping vs Sequencing

- **Genotyping** - determining which genetic variants an individual possesses through a variety of different methods, especially genotyping chips (based mostly on SNPs – single nucleotide polymorphisms)
 - cheap, but require prior identification of the variants of interest

THANKSGIVING FAMILY OFFER

RECOMMENDED

Ancestry Service




Experience your ancestry in a new way! Get a breakdown of your global ancestry by percentages, connect with DNA relatives and more. [learn more](#)

~~\$99~~ **\$49**
when you buy 2+ kits*

add to cart

NOW WITH **150+** REGIONS

Health + Ancestry Service



Get an even more comprehensive understanding of your genetics. Receive 90+ online reports on your ancestry, traits and health - and more. **New BRCA1/BRCA2 (Selected Variants)* report just added!** [learn more](#)

\$199

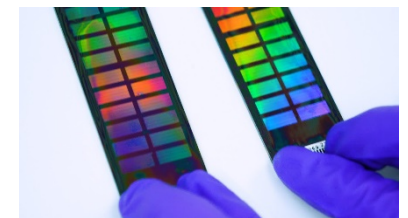
add to cart

Methods

We use genotyping technology to look at specific genetic variants in the genome that can be most informative about an individual's health and ancestry.

Unlike sequencing which analyses all nucleotides in a gene to identify changes, genotyping detects specific known variants within the genome. 23andMe uses a custom Illumina HumanOmniExpress-24 format chip that analyses approximately half a million variants. This custom chip has been designed to include variants:

- In medically relevant genes
- Involved in drug metabolism, efficacy and side effects
- With known disease associations
- Associated with traits
- Used to assign genetic ancestry and ethnicity



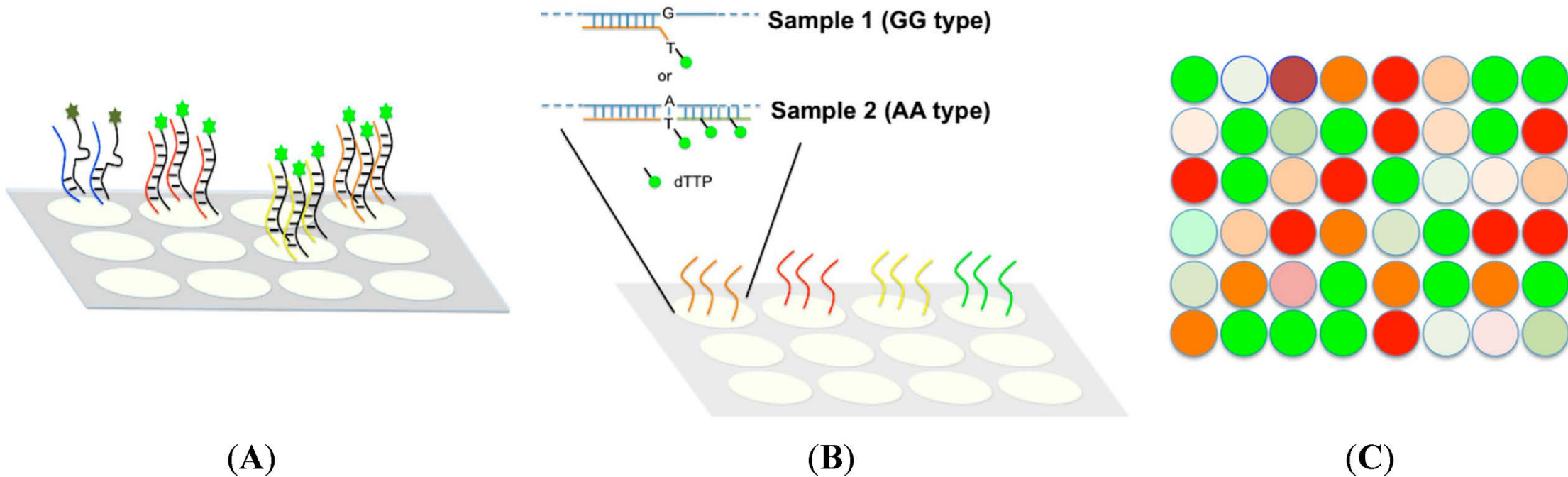
<https://www.23andme.com/>

How SNP genotyping works

- https://www.youtube.com/watch?v=Naona1y_I2U
- For more information see YouTube Channel Useful Genetics:
<https://www.youtube.com/channel/UcTxCrx28msMBQ-vFUIOIReA>

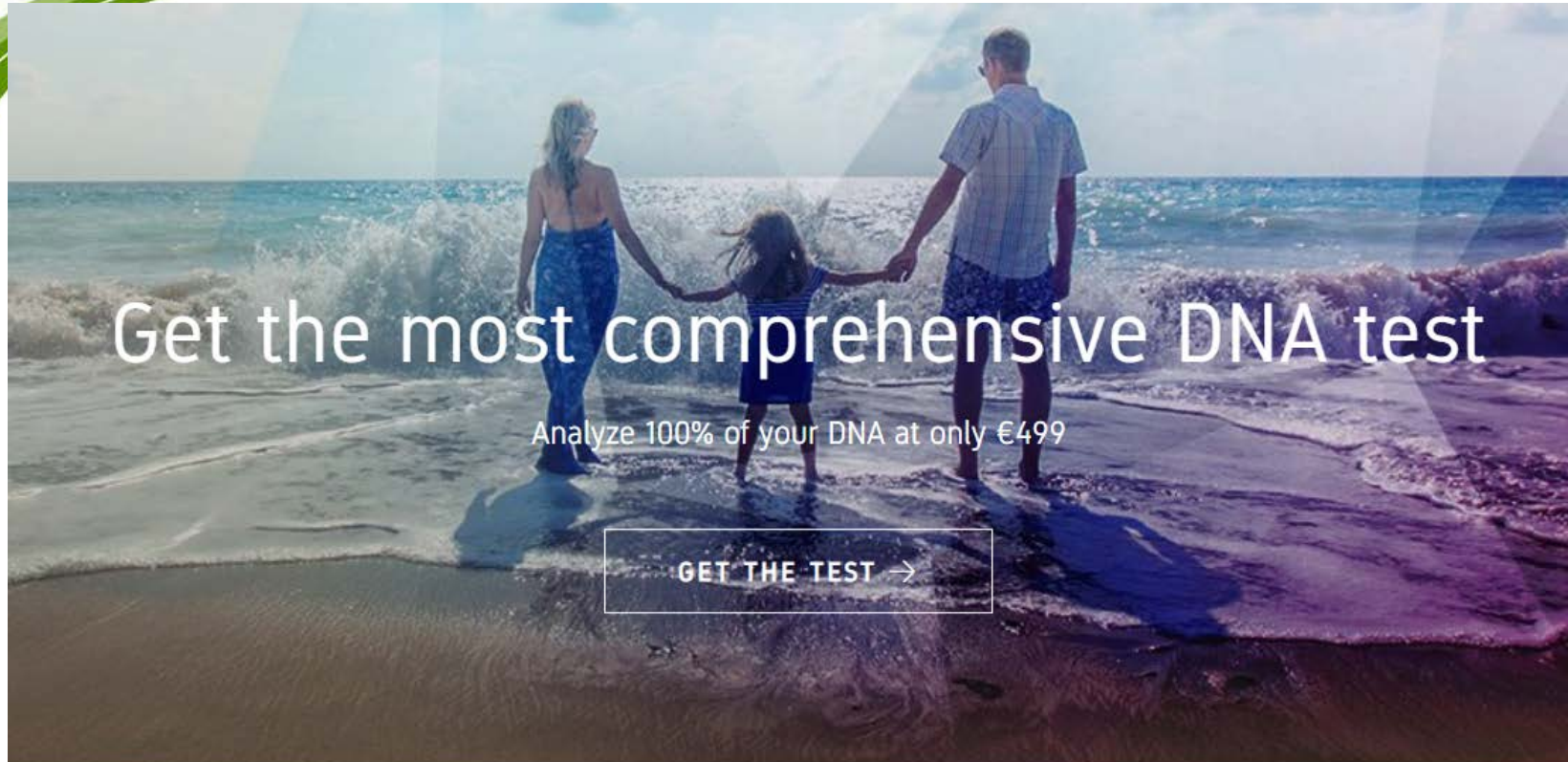


How SNP genotyping works



There are two types of microarray commonly used in multiplexing SNP analysis: allele-specific oligonucleotide (ASO) hybridization and allele-specific primer (ASP) extension. **(A)** ASO hybridization: The allele-specific oligonucleotide for every SNP is synthesized and separately immobilized onto the glass plate. Fluorescence labeled targets containing SNP sites are produced from a PCR reaction and plotted separately into each well to conduct the hybridization reaction. The mismatched base pair between target and oligonucleotide can decrease the binding strength with the fluorescence-labeled target removed after a stringent washing. A fluorescence signal is detected on a perfectly matched base pair; **(B)** Allele-specific primer (ASP) extension: The specific primer for SNP location is designed and separately immobilized onto a microarray. A different fluorescence labeled dNTP is individually used in an extension reaction. The extended fragment showing fluorescence signal can only be found when the 3' end of primer pair is perfectly matched (AA type in this case) in contrast to the mismatched primer pair (GG type in this case); **(C)** The SNP genotype can be determined according to fluorescent intensity from the products/target DNA. <https://doi.org/10.3390/microarrays4040570>

DTC genome sequencing as popular demand



Get the most comprehensive DNA test

Analyze 100% of your DNA at only €499

[GET THE TEST →](#)

Coverage (or depth) in sequencing

```
Read 1: CGGATTACGTGGACCATG (read length of 18)
Read 2: ATTACGTGGACCATGAATTGCTGACA
Read 3: ACCATGAATTGCTGACATTCGTCA
Read 4: TGAATTGCTGACATTCGTCA
Depth: 1112222222233334433333333332222221
```

WHAT YOU GET

Dante Labs analyzes 100% of your DNA, so that we can give you reports on predispositions on any genetic disease. You will receive easy reports for you and your doctor, as well as raw data to explore.



My Full DNA: Whole Genome Sequencing with mtDNA

€449.00 EUR ~~€850.00-EUR~~
YOU SAVE €401.00 EUR

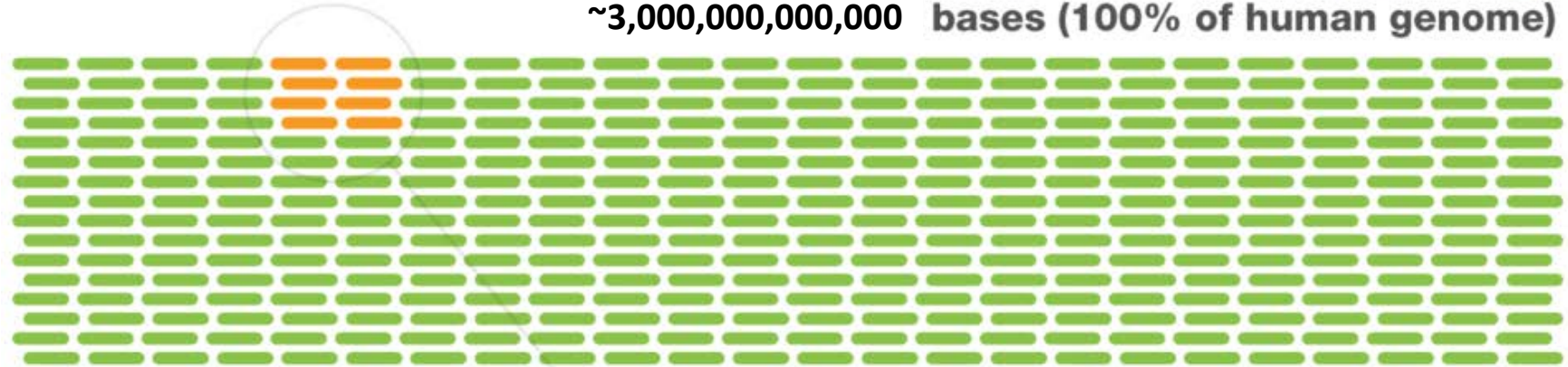
<https://www.dantelabs.com>

Sequencing – WGS and WES

- Determining the exact DNA sequence

Whole Genome Sequencing

~3,000,000,000,000 bases (100% of human genome)

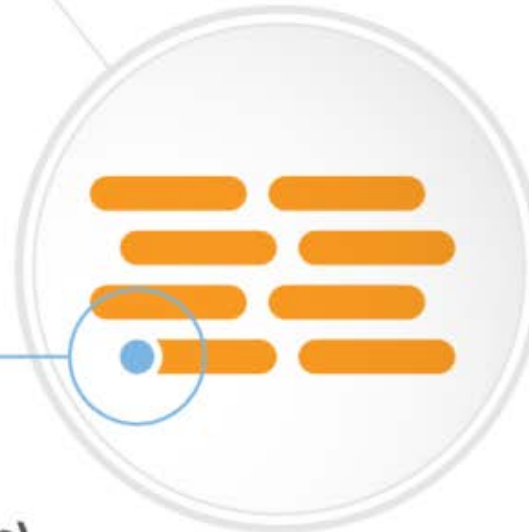


Whole Exome Sequencing

~60,000,000 bases
(~2% of human genome)

Large Scale Genotyping

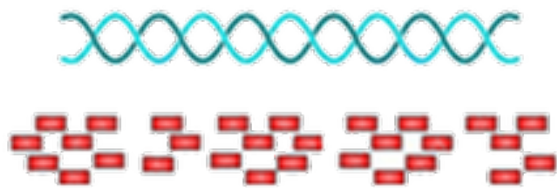
~1,000,000 bases
(~0.03% of human genome)



“Non-coding DNA” was long thought of as junk DNA, but as we understand more about our genetics we now know these regions play a hugely important role in regulating the coding portions of our DNA. Our understanding of these regions and their interactions is relatively poor compared to our knowledge of the DNA coding regions.

Genomes vs exomes vs genotypes

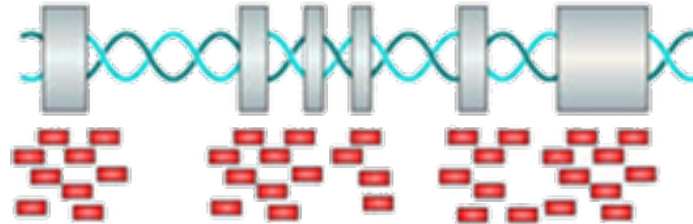
Whole genome sequencing



- Sequencing region : whole genome
- Sequencing Depth: >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

- Results are sometimes challenging to interpret

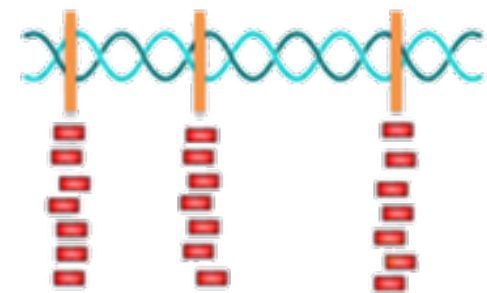
Whole exome sequencing



- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

- Good alternative to WGS in terms of clinical use

Targeted sequencing



- Sequencing region: specific regions (could be customized)
- Sequencing Depth : >500X
- Identify all kinds of variants including SNPs, INDELs in specific regions
- Most Cost effective

- Most sensitive – able to detect rare tumor cells in biopsy

What to expect

- Genetic testing provided by most of the companies is moreless for fun (ancestry, health and wellness, nutrigenetics, skincare, sports,...)



Health and Wellness

The 24Genetics Health DNA Test can make a real difference to your health. Your genetic information is vital when making some of the most important decisions in your life ...



Nutrigenetics

What is Nutrigenetics? You have surely noticed that you have certain reactions to some foods. You have probably noticed, for example, that while some lose weight eating a set of foods, you ...



Skin Care

In this Genetic Test for Skin Care we analyse how your genetics influence skin characteristics, such as hydration, elasticity and antioxidant capacity, which ...



Sports

To reach the highest level in the sports world, training hard is not enough. You have to train smart. Knowing oneself is the best starting point, and these genetic tests are the key ...

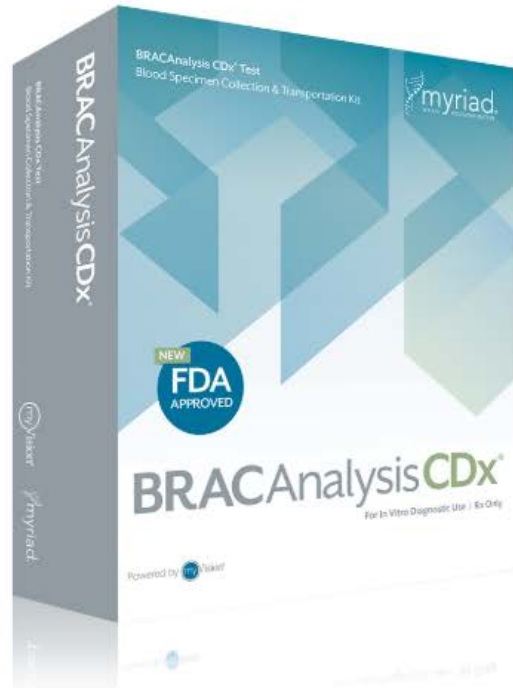
- More expensive, and complete, sequencing like the one provided by Illumina can be used for medical investigation
- Do not expect your genome sequencing to tell you how long is your life expectation, whether you are likely to get cancer and so on
- So far our knowledge on the “implication” of the genome are quite limited
- What we can already do in health care is to look at the genome once you have been diagnosed a specific ailment and look for specific genes that would make one cure more effective than another (this has become normal practice in some form of cancer cure)

Example of genetic testing in clinical practise

- *BRCA* genes testing for PARP inhibitor treatment

BRACAnalysis CDx[®] Ovarian Cancer

Overview



Mutations in *BRCA1* or *BRCA2* cause Hereditary Breast and Ovarian Syndrome (HBOC). Now mutations in the *BRCA1* and *BRCA2* genes provide an indication for treatment with Lynparza™ (olaparib) for patients with ovarian cancer. Specifically, BRACAnalysis CDx[®] is the only FDA-approved laboratory developed test approved to be used to inform treatment decisions for the PARP inhibitor, Lynparza. A positive BRACAnalysis CDx result in patients with ovarian cancer is also associated with enhanced progression-free survival (PFS) from Zejula™ (niraparib) maintenance therapy.^{1,2,3}

Learn More

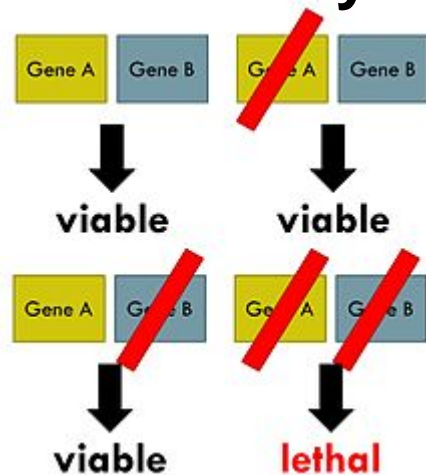
Order BRACAnalysis CDx

- <https://www.youtube.com/watch?v=ilwMGRH276M>

PARP inhibitors

- In December 2014, the drug olaparib (Lynparza) became the first of a new class of treatments known as PARP (poly(ADP-ribose)polymerase) inhibitors to be licensed for clinical use, heralding in a new era for personalised, targeted treatment—and turning the promise of ‘synthetic lethality’ into reality.

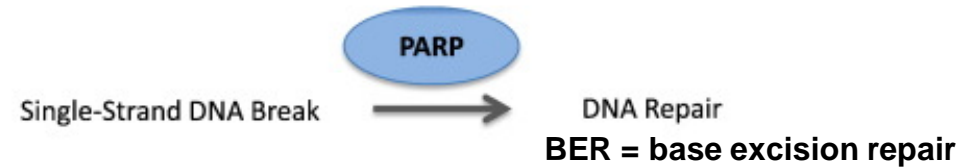
Synthetic lethality concept



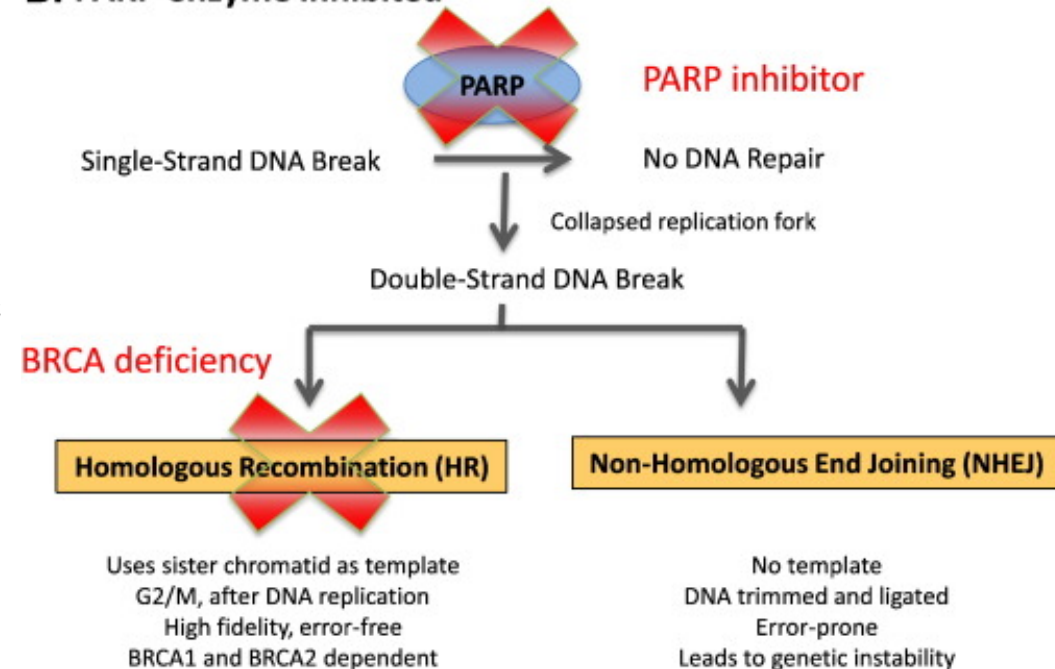
More info on PARPi:

<https://www.youtube.com/watch?v=mgW30YyaJz4>

A. Functioning PARP enzyme



B. PARP enzyme inhibited



C. Deficiency in HR and BER together lead to synthetic lethality

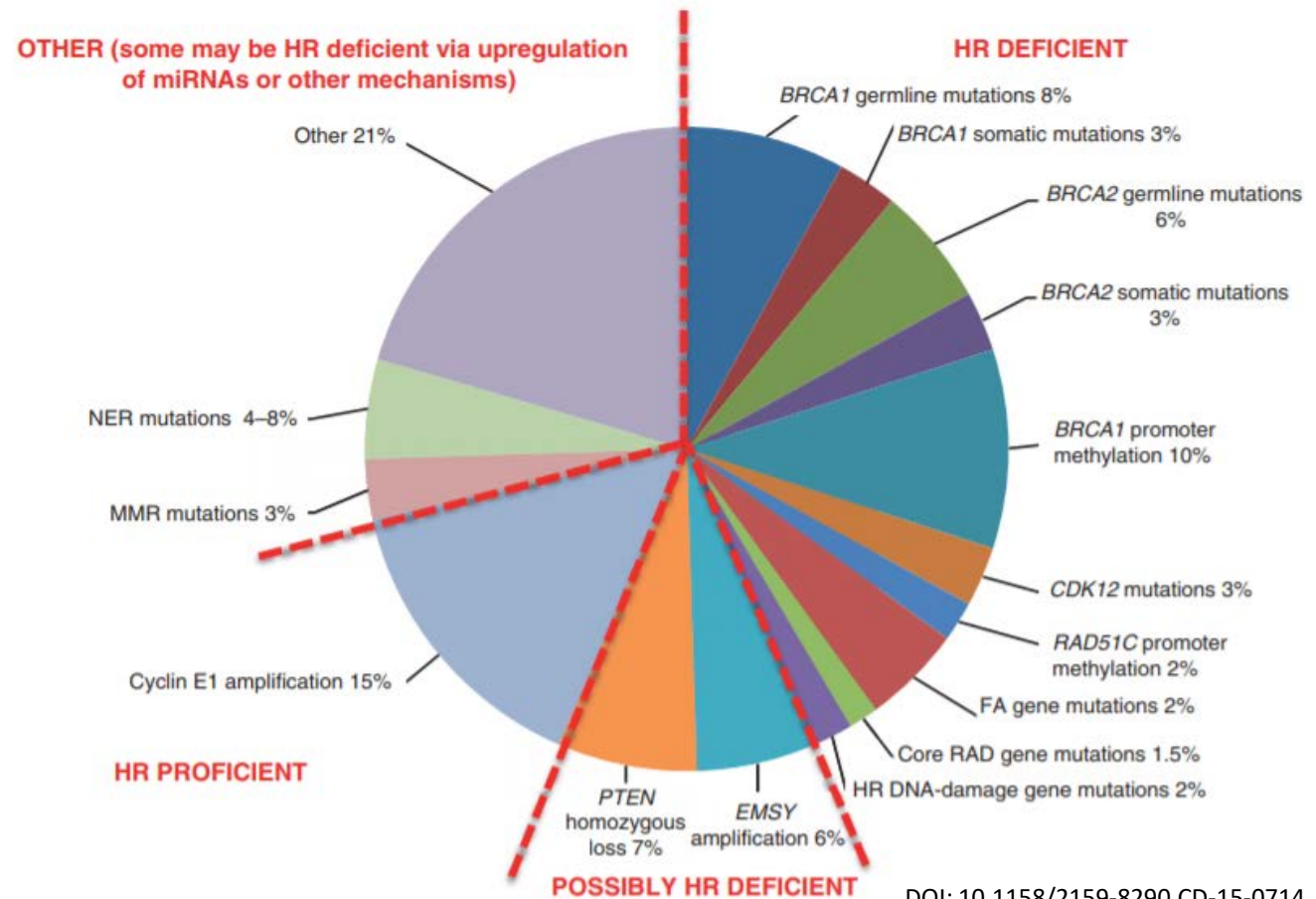
Condition	HR	BER	Outcome
Normal cells	+	+	Viable
BRCA deficient	-	+	Viable
Normal cells, PARP inhibitor	+	-	Viable
BRCA deficient, PARP inhibitor	-	-	Cell Death

<https://doi.org/10.1016/j.ygyno.2015.02.017>

BRCAness ovarian tumors and PARPi

- Only women with mutation in *BRCA* genes are now eligible for PARPi treatment
- *BRCA* mutations underlie only a small portion of tumors defective in HR
- But 50% of ovarian tumors are HR deficient = 'BRCAness' phenotype
- PARPi are effective also in BRCAness tumors
- *Can we identify the BRCAness tumors/patients and provide them also with this novel and highly promising treatment option?*

Homologous Recombination Deficiency:
Exploiting the Fundamental Vulnerability of Ovarian Cancer



The Present and Future of Genome Sequencing

- Genomics England - **100,000 patients** with rare diseases, their families, and cancer patients
- Precision Medicine Initiative (PMI) **1-million-volunteer** health study, data including genetics and lifestyle factors
- GenomeAsia **100K** - genomic data for Asian populations
- ... a many more initiatives
- How to handle such huge amount of data and the ethical implications?
- In the US, the Genetic Information Nondiscrimination Act (2008) but mostly no act in other countries and somewhat grey legal position in Europe



<https://labiotech.eu/features/genome-sequencing-review-projects/>

COSMIC: Catalogue of Somatic Mutations in Cancer



<https://cancer.sanger.ac.uk/cosmic/about>



Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Genome Version ▾ Search COSMIC... **SEARCH** Login ▾

What is COSMIC?

COSMIC – the Catalogue of Somatic Mutations in Cancer – is the world's largest source of expert manually curated somatic mutation information relating to human cancers. Here we outline that data in terms of structure, content and scope making it easier for you to evaluate what you will find in COSMIC, and how best to access it to fulfill your research needs.

Overview

COSMIC comprises the COSMIC database and the Cell Lines Project, two separate but related resources. This page discusses [COSMIC](#); please see the [About Cell Lines](#) page for more information on the Cell Lines Project.

The COSMIC database combines two main types of data:

High Precision Data, Manually Curated by Experts:

- Targeted gene-screening panels
- Over 25,000 peer reviewed papers
- Metadata (environmental factors and patient history)
- Focused on known and suspected cancer genes and mutations
- Objective frequency data as a result of mutation negative samples
- Full details of the curation process and data captured

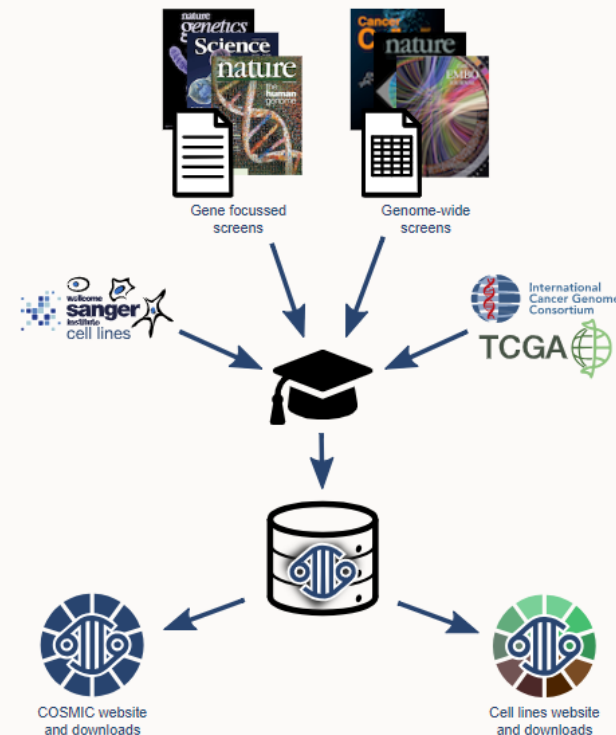
Genome-wide Screen Data:

- Over 32,000 genomes, consisting of:
 - peer reviewed large scale genome screening data
 - other databases such as [TCGA](#) and [ICGC](#)
- Provides unbiased, genome-level profiling of diseases
- Objective frequency data, by interpreting non-mutant genes across each genome
- Can be used to discover novel driver genes

Together, this compilation of data provides extensive coverage of the cancer genomic landscape from a somatic perspective. New and potentially significant data are continually captured and made available through four significant updates to COSMIC each year.

For more information on COSMIC, read more about our [curation processes](#) and the [analyses](#) that we run on mutation data, or see our answers to [frequently asked questions](#) about curation, histology, and mutation syntax.

Find out more about licensing COSMIC data



- How to use COSMIC database:

<https://www.youtube.com/watch?v=2FD5RabgK6o>, <https://www.youtube.com/watch?v=k477uAiKx74>

TCGA: The Cancer Genome Atlas



1-800-4-CANCER Live Chat Publications Dictionary

ABOUT CANCER CANCER TYPES RESEARCH GRANTS & TRAINING NEWS & EVENTS ABOUT NCI search

Home > About NCI > NCI Organization > CCG > Research > Structural Genomics



TCGA

Program History +

TCGA Cancers Selected for Study

Publications by TCGA

Using TCGA +

Contact

The Cancer Genome Atlas Program

The Cancer Genome Atlas (TCGA), a landmark [cancer genomics](#) program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between the National Cancer Institute and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already led to improvements in our ability to diagnose, treat, and prevent cancer, will remain [publicly available](#) for anyone in the research community to use.



TCGA Outcomes & Impact

TCGA has changed our understanding of cancer, how research is conducted, how the disease is treated in the clinic, and more.



TCGA's PanCancer Atlas

A collection of cross-cancer analyses delving into overarching themes on cancer, including cell-of-origin patterns, oncogenic processes and signaling pathways. Published in 2018 at the

https://www.youtube.com/watch?time_continue=249&v=epsZjJ_A1y4

<https://cancergenome.nih.gov/>

TCGA: Overview

- Initiated in 2005
- A joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI).
- 27 participating Institutes in US and Canada.
- The overarching goal of TCGA is to improve our ability to diagnose, treat and prevent cancer, through the application of genome analysis technologies, including large-scale genome sequencing.
- The Cancer Genome Atlas Network have published more than 20 papers since the project began

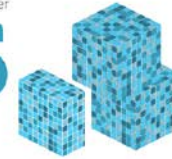
(<https://tcga-data.nci.nih.gov/docs/publications/>)

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over

2.5
PETABYTES
of data



TCGA data describes

33
DIFFERENT
TUMOR TYPES

including

10
RARE
CANCERS

...based on paired tumor and normal tissue sets collected from

11,000
PATIENTS

...using

7 DIFFERENT
DATA TYPES

To put this into perspective, 1 petabyte of data is equal to

212,000
DVDs



TCGA RESULTS & FINDINGS



MOLECULAR
BASIS OF
CANCER

Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies



TUMOR
SUBTYPES

Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*



THERAPEUTIC
TARGETS

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM



20
COLLABORATING
INSTITUTIONS
across the United States
and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

TCGA Data Portal

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects

Exploration

Analysis

Repository

Q e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary

Data Release 13.0 - September 27, 2018

PROJECTS

43

PRIMARY SITES

69

CASES

33 096

FILES

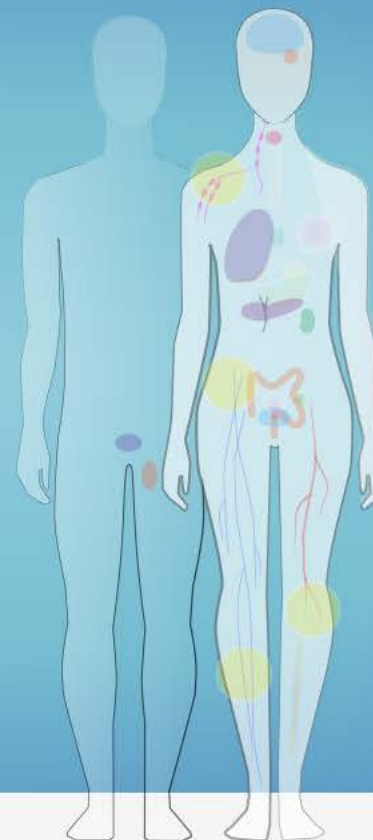
358 092

GENES

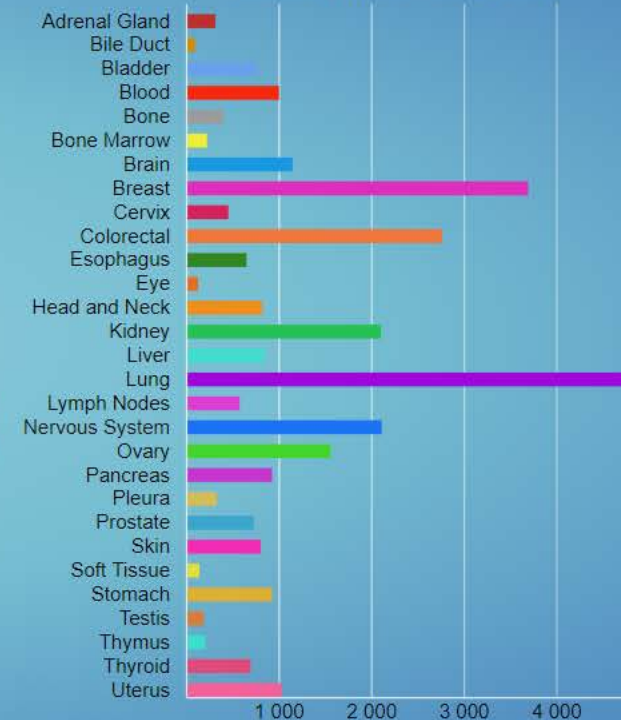
22 872

MUTATIONS

3 142 246



Cases by Major Primary Site



GDC Applications

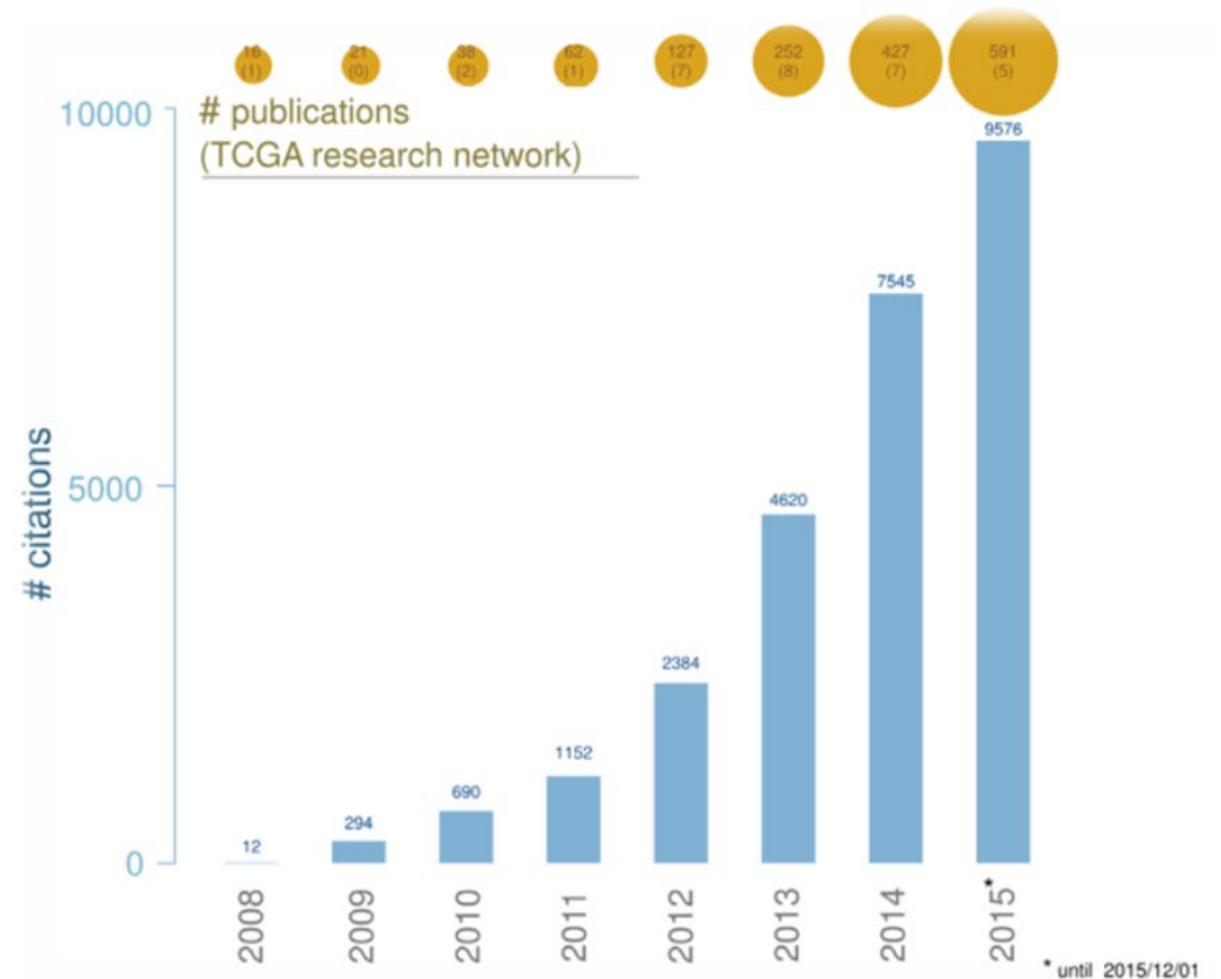
The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

TCGA: A Valuable Resource for Research Community

TCGA Data Types

- Clinical data
- DNA sequencing
- miRNA sequencing
- Protein expression
- mRNA sequencing
- Total RNA sequencing
- Array-based expression
- DNA methylation
- Copy number variations

+ Computational tools



OncoMine database

The screenshot shows the OncoMine website interface. At the top, the OncoMine logo is on the left, the URL www.OncoMine.org is in the center, and 'upgrade' and 'contact' buttons are on the right. The main content area features a large banner with the text 'Design better experiments. Gain more insights. Prepare to publish faster.' and a background image of a hand holding a globe with data overlays. Below the banner, there are three columns of text describing the platform's capabilities. On the left side, there is a 'login' form with fields for 'USER ID' (containing 'pospich@sci.muni.cz') and 'PASSWORD', along with links for 'Forgot password?' and 'Not a user? Register now!'. Below the login form are sections for 'news' and 'events'.

www.OncoMine.org

upgrade @ contact

OncoMine™ Research Edition: 715 datasets and 86,733 samples



With OncoMine™ Research Premium Edition, you can:

Design better experiments... Answer more questions with fewer experiments, select the most promising gene or cell line, and test your hypothesis.

Gain more biological insights... Discover novel targets for therapeutic development, interrogate gene expression profiles, and identify drug and biological interactions.

Prepare to publish faster... Validate your results faster, visualize data easier and make connections to clinical significance.

The OncoMine™ Platform—from web applications to translational bioinformatics services—provides solutions for individual researchers and multinational companies, with robust, peer-reviewed analysis methods and a powerful set of analysis functions that compute gene expression signatures, clusters and gene-set modules, automatically extracting biological insights from the data. It has become an industry-standard tool cited in more than 1,100 peer-reviewed journal articles. The OncoMine Platform has been used as a foundation for groundbreaking discoveries with unique features that include:

ion torrent
by Thermo Fisher Scientific

The Origin of the OncoMine Platform
The OncoMine Platform was conceived by physicians, scientists, and software engineers at the University of Michigan. It was commercialized by Arul Chinnaiyan and Dan Rhodes in February 2006 with the goal of building a version that would have a greater ability to impact drug development and clinical practice.

How to use OncoMine:
<https://www.youtube.com/watch?v=b8ckDiVNrFE>

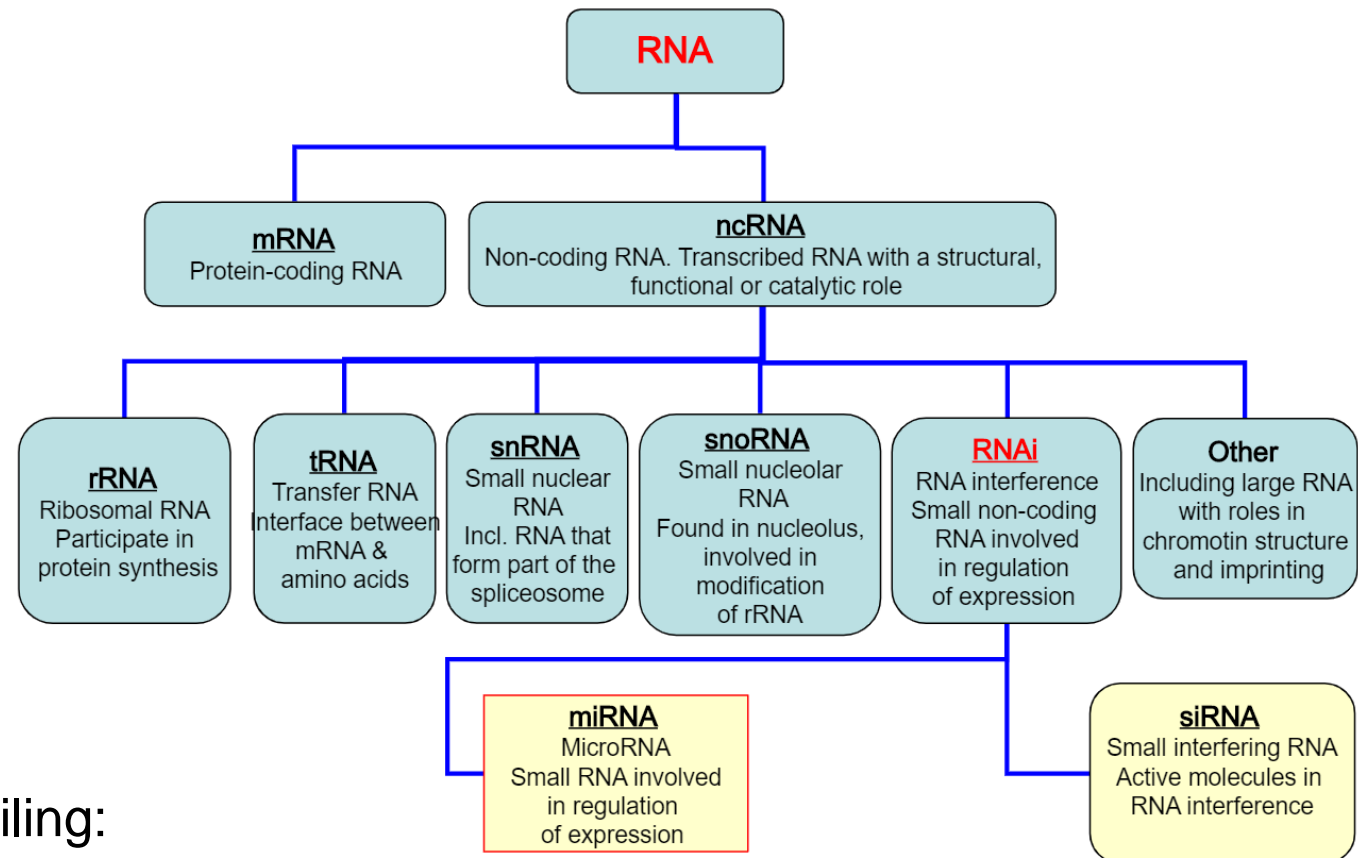
Transcriptomics

- Study of transcriptome, the sum of all RNA transcripts
- Two most widely studied types of RNA

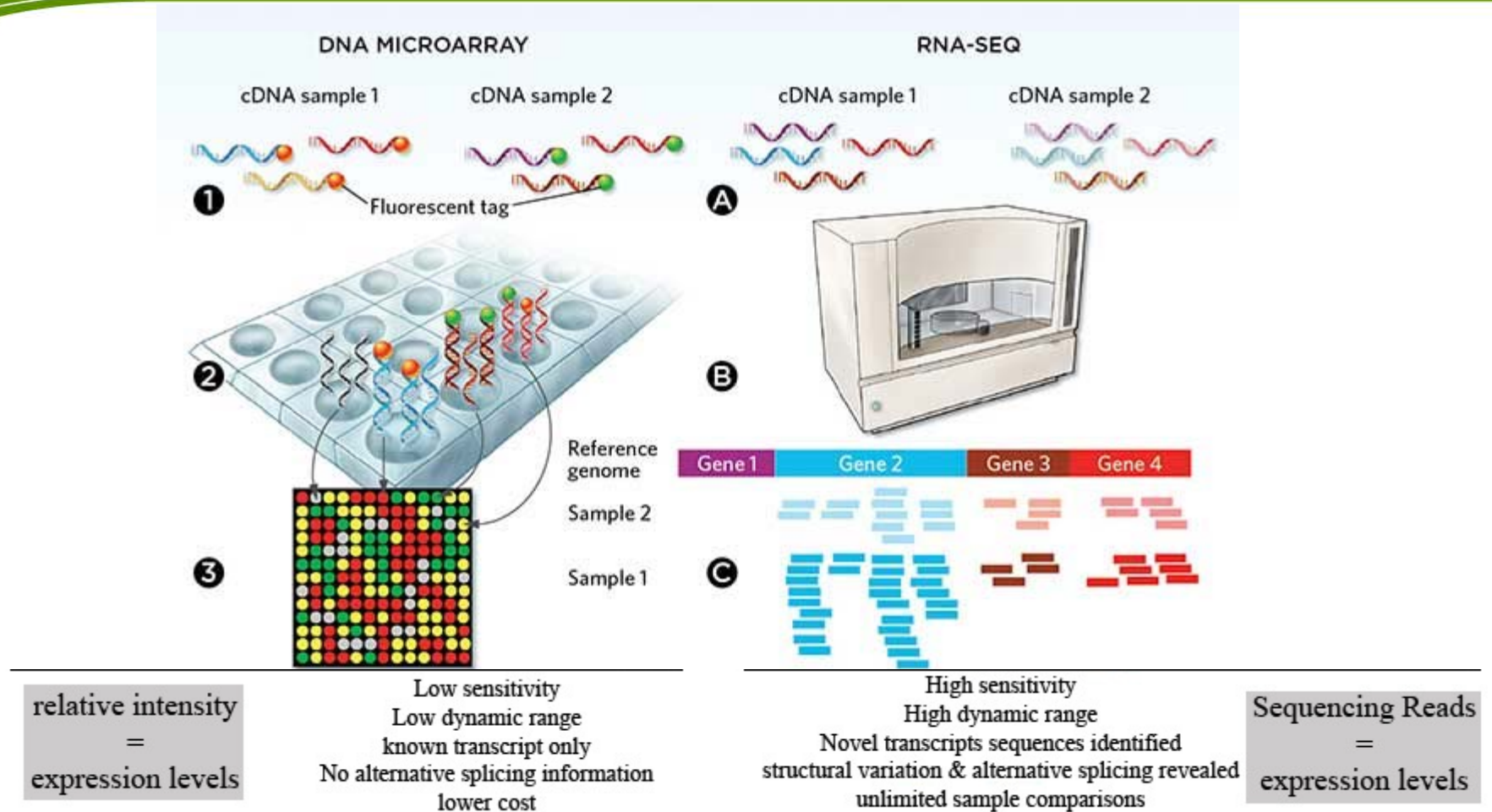
- mRNA - transcriptome or the expressed genes. Usually contains genes with poly A tail.
- miRNA - Small non-coding RNA (containing about 21-25 nucleotides), important in gene regulation.

- Array-based Expression Profiling:
- <https://www.youtube.com/watch?v=6ZzFihESjp0>

Type of RNA molecules

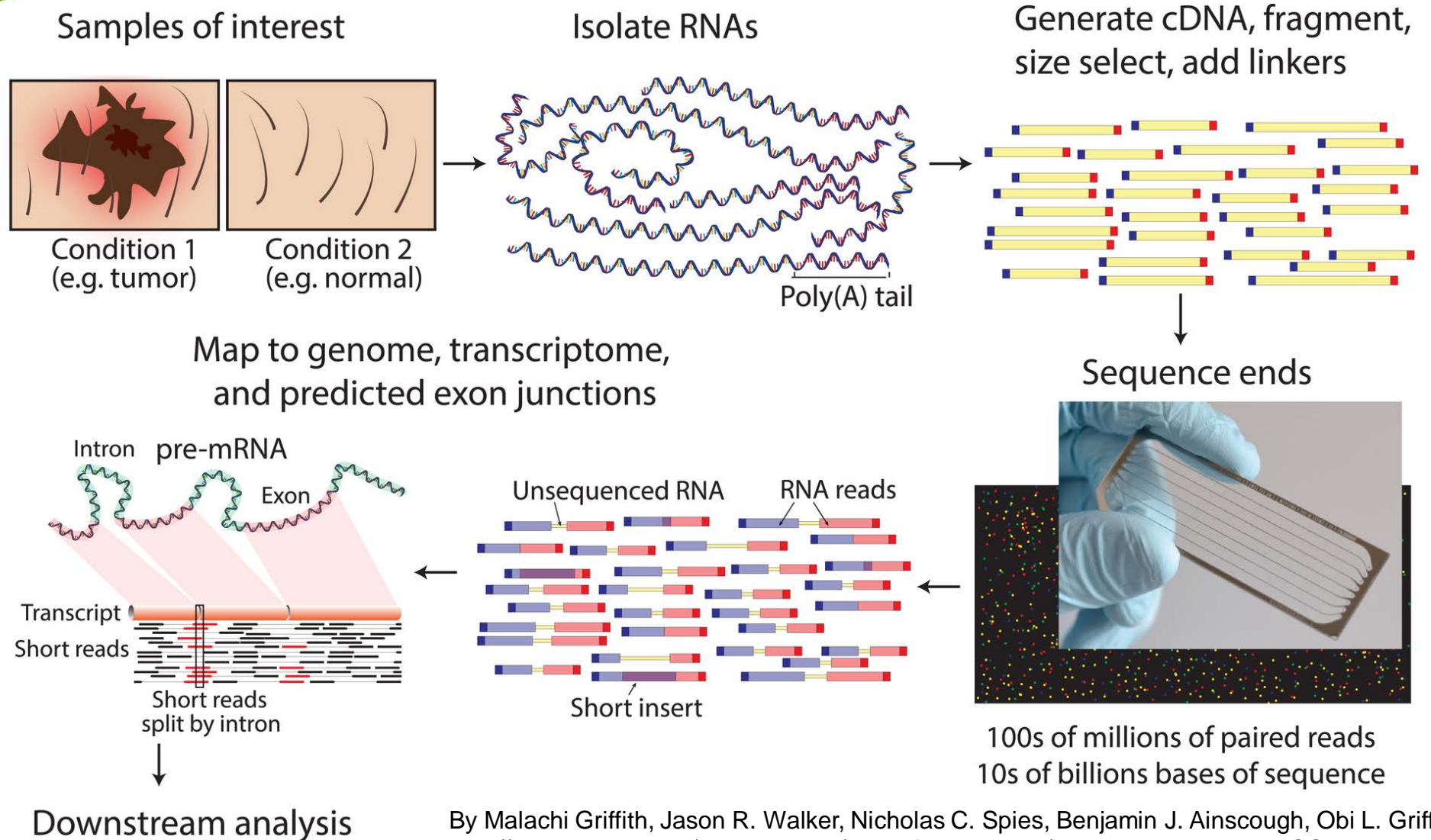


Microarrays vs RNA-seq



- While methods for analyzing microarray data are fully mature and straightforward, there is no consensus on which pipelines—or series of computational steps—to use to analyze RNA-seq data.

Overview of RNA-seq

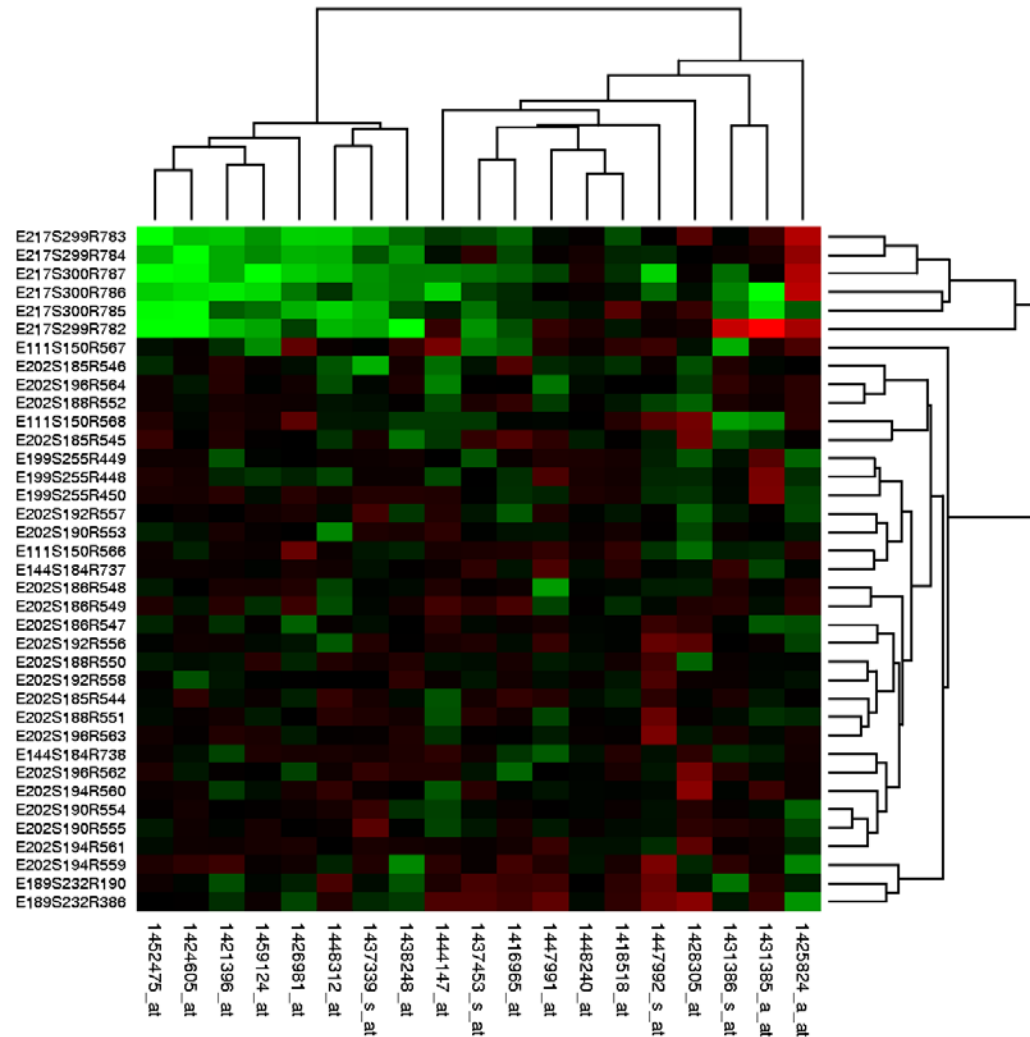


By Malachi Griffith, Jason R. Walker, Nicholas C. Spies, Benjamin J. Ainscough, Obi L. Griffith - <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>, CC BY 2.5, <https://commons.wikimedia.org/w/index.php?curid=53055894>

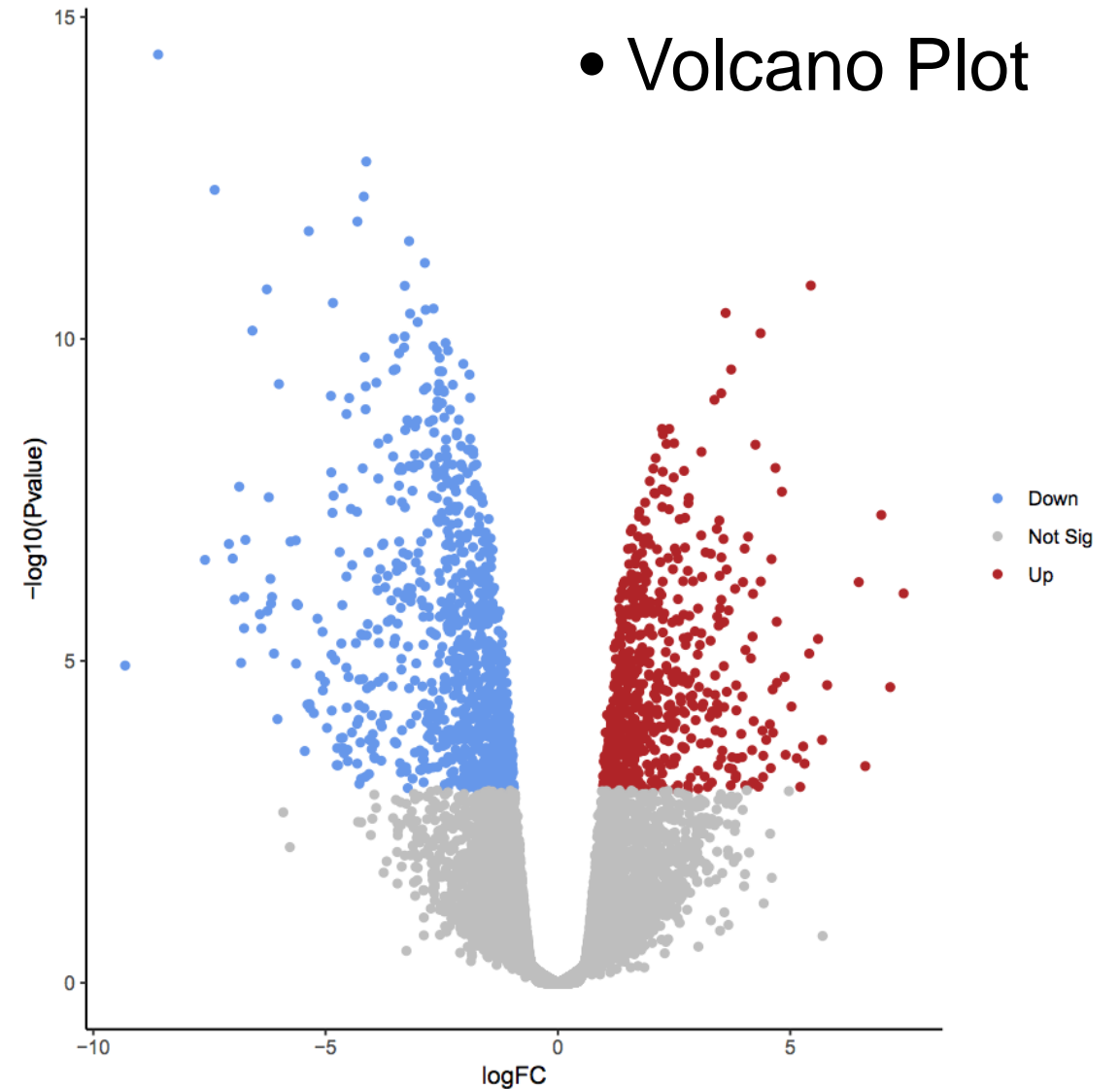
RNA sequencing downstream analysis

- <https://www.youtube.com/watch?v=tlf6wYJrwKY> (from 13:10)
- More info about microarray vs. RNA-seq at:
<https://www.youtube.com/watch?v=2c3t3tDEmsU>
- More info RNA seq at:
- https://www.youtube.com/watch?v=MFRkwXq6v_I
- Useful detailed info about anything connected to RNA-seq
- <https://www.rna-seqblog.com>

Examples of transcriptomics data outputs



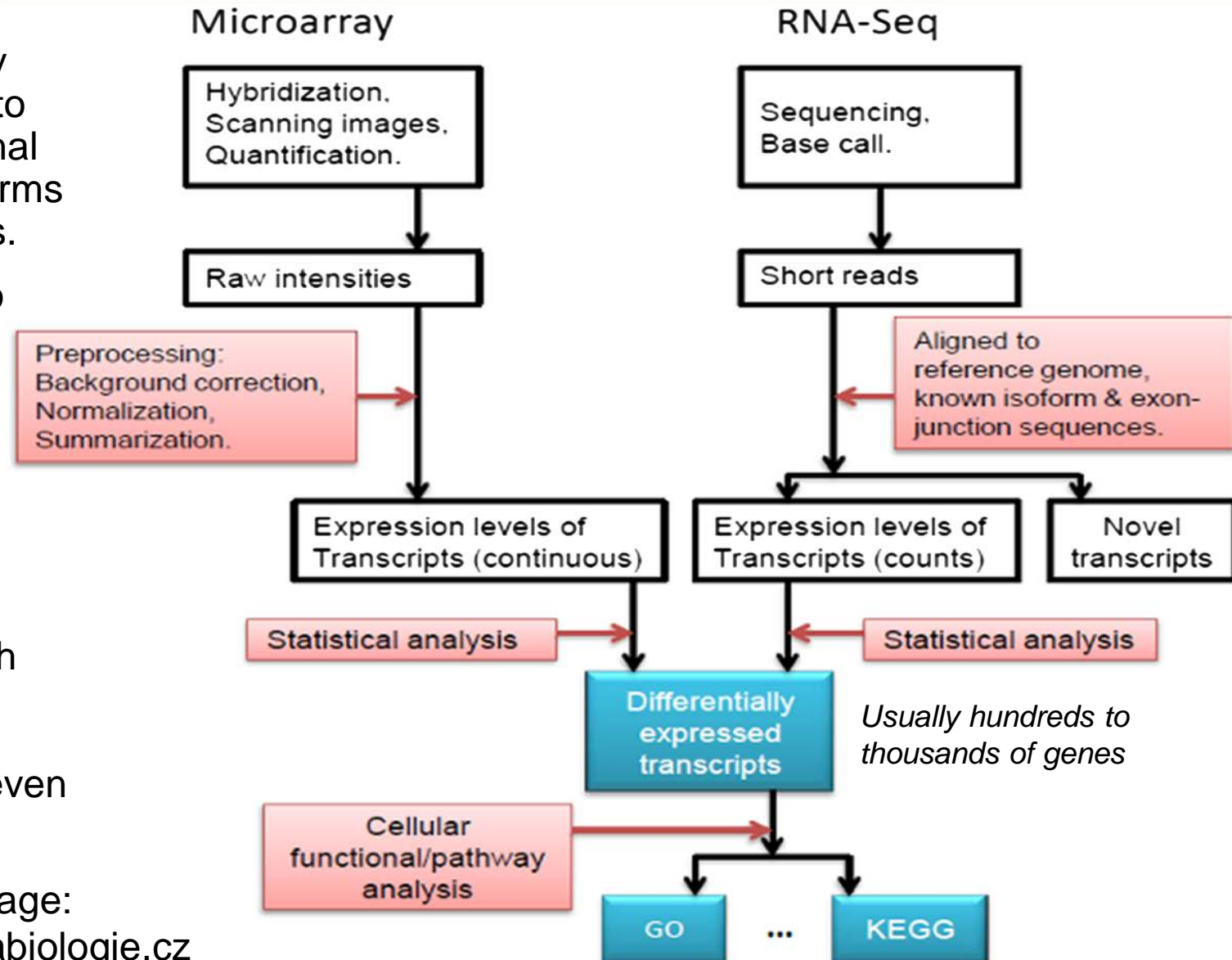
• Heat map



• Volcano Plot

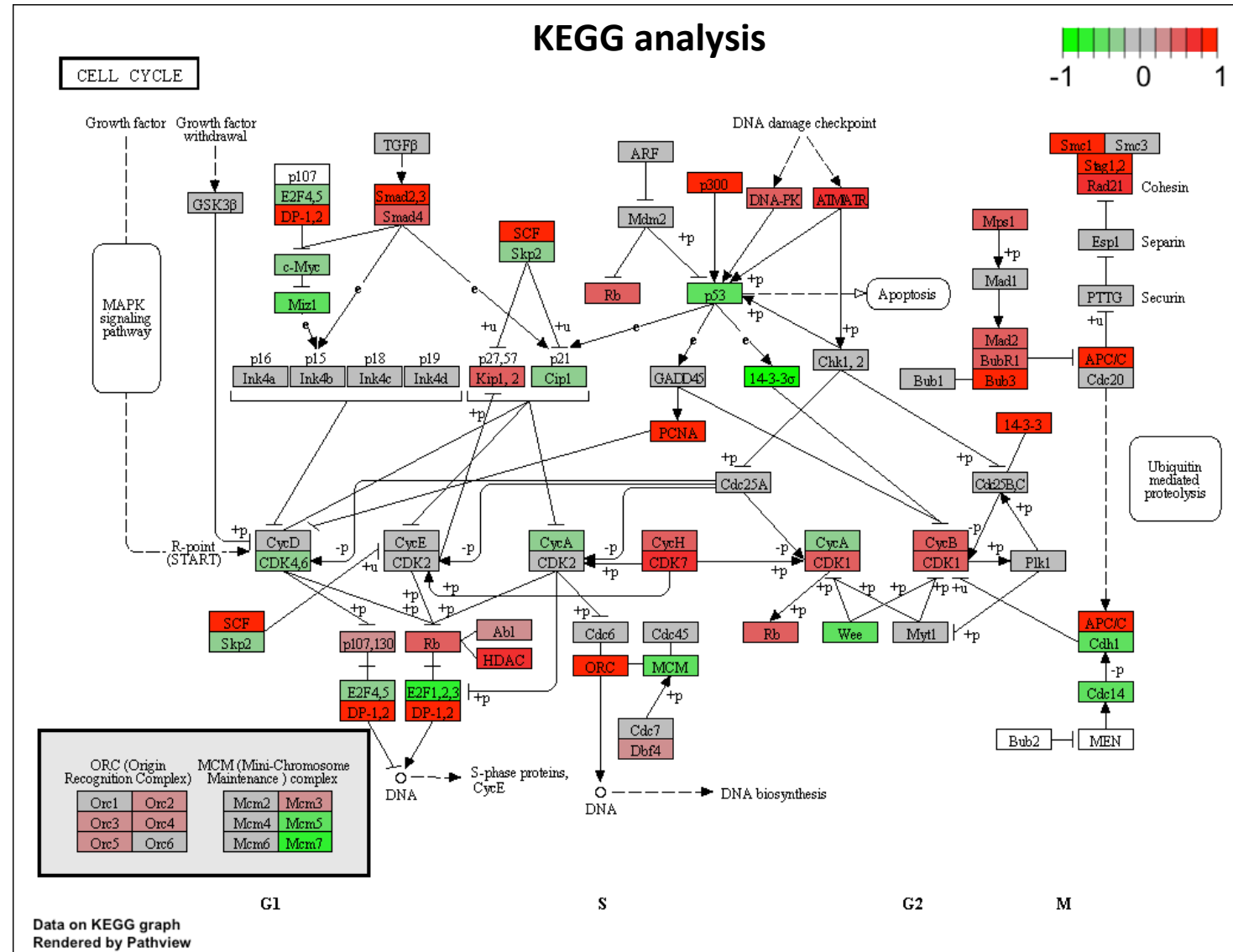
Cellular/functional/pathway analysis

- Cellular/functional/pathway analysis is a valuable tool to summarize high-dimensional gene expression data in terms of biologically relevant sets.
- Genes are aggregated into gene sets on the basis of shared biological or functional properties as defined by a reference knowledge base.
- Knowledge bases are database collections of molecular knowledge which may include molecular interactions, regulation, molecular product(s) and even phenotype associations.
- Useful info in Czech language: <https://portal.matematickabiologie.cz>



Database resources for understanding high-level functions and utilities of the biological system

- Database tools:
 - **KEGG** (Kyoto Encyclopedia of Genes and Genomes)
 - (<https://www.genome.jp/kegg/>)
 - Disadvantage – does not provide statistical significance of particular pathways
 - And many others available online



Gene-set analysis (GSA)/Pathway analysis

- Gene Ontology (GO) analysis (<http://geneontology.org/>)

Current release 2019-10-07: 44 733 GO terms | 7 330 378 annotations
1 405 197 gene products | 4 493 species (see statistics)

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

Search GO term or Gene Product in AmiGO ...

Any ● Ontology ● Gene Product

GO Enrichment Analysis ?

Powered by PANTHER

Your gene IDs here...

biological process

Homo sapiens

Examples Launch

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs



The network of biological classes describing the current best representation of the "universe" of biology. The molecular functions, cellular locations, and processes gene products may carry out.



Statements, based on specific, traceable scientific evidence, asserting that a specific gene product is a real exemplar of a particular GO class.



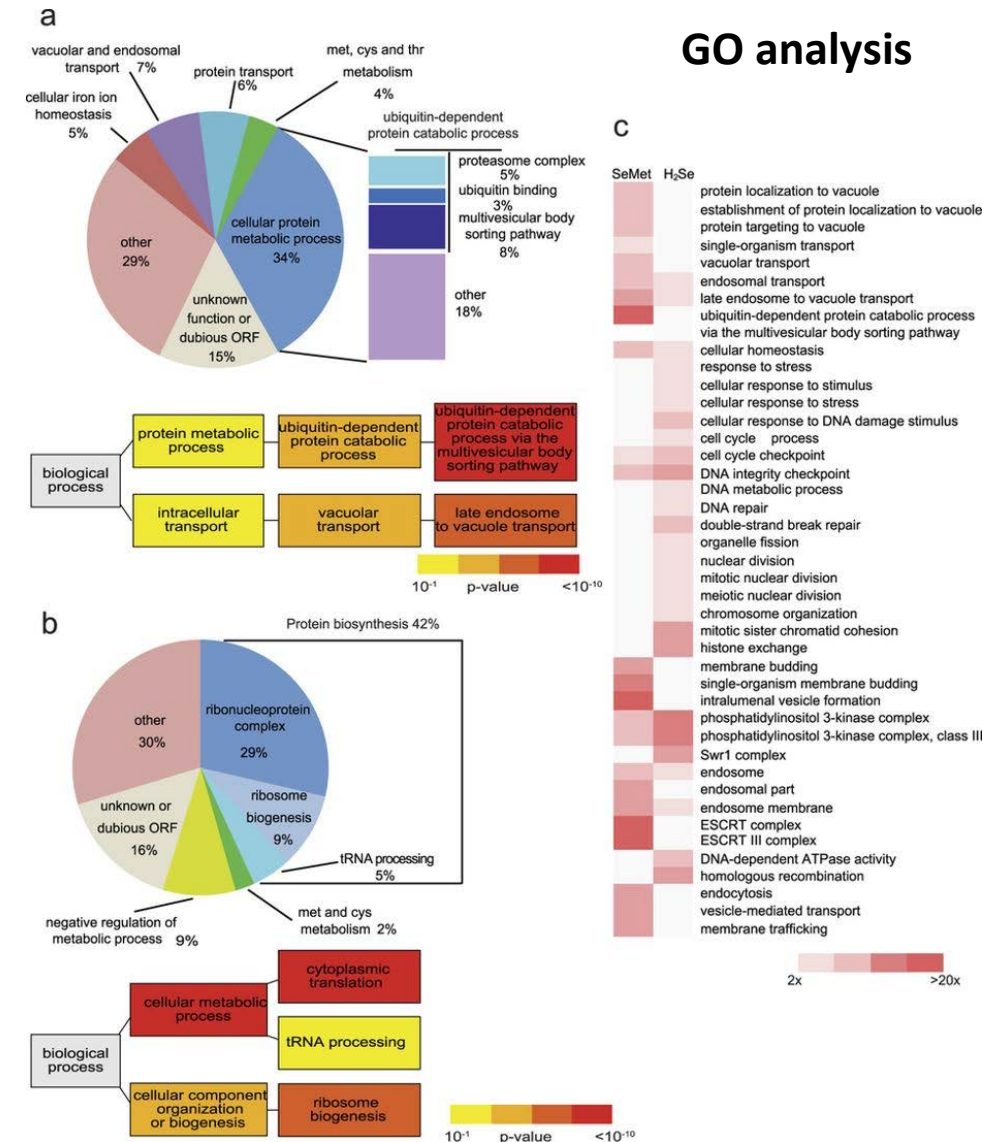
GO Causal Activity Model (GO-CAM) provides a structured framework to link standard GO annotations into a more complete model of a biological system.



Tools to curate, browse, search, visualize and download both the ontology and annotations. Bioinformatic Guides (Notebooks) and simple API access to integrate GO into your research.

Example data of GO enrichment analysis

- **GO enrichment analysis**
- One of the main uses of the GO is to perform enrichment analysis on gene sets. For example, given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set.
- 3 main GO aspects (molecular function, biological process, cellular component)
- <http://geneontology.org/docs/go-enrichment-analysis/>



Reactome Knowledgebase

Why Reactome

Reactome is a free, open-source, curated and peer-reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education.

If you use Reactome in Asia, we suggest using our Chinese mirror site at reactome.ncpsb.org.

EMBL-EBI NYU Langone Health OICR

The development of Reactome is supported by grants from the US National Institutes of Health (P41 HG003751 and 1U54GM114833-01), CFREF Medicine by Design, and the European Molecular Biology Laboratory.

Latest News

- Version 70 Released
- ReacFoam: Genome-wide pathway overview based on Voronoi tessellation
- ORCID: Claim your works
- Version 69 Released
- Season of Docs

Tweets


reactome @reactome
Interested in pathway enrichment analysis or GSEA! Read our new "Perform Pathway Enrichment Analysis Using ReactomeFIViz" #Methods in #Molecular #Biology chapter: ncbi.nlm.nih.gov/pubmed/31583638 & try out the #cytoscape app #oaweeek #thanksOA #OAWeeek2019

OCTOBER 21-27 OPEN ACCESS WEEK 2019

Perform Pathway Enrichment Analysis Using ReactomeFIViz

Embed View on Twitter


- More info at:
- <https://www.youtube.com/user/Reactome/videos>

 Version 70 released on September 9, 2019



2,287
Human Pathways


12,608
Reactions


10,860
Proteins

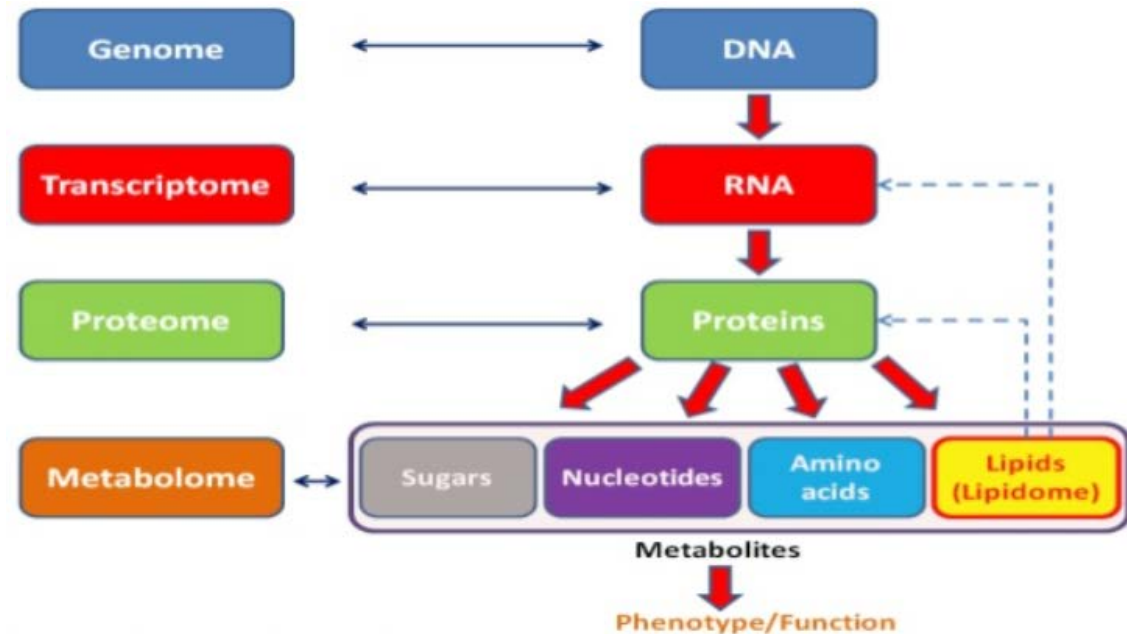
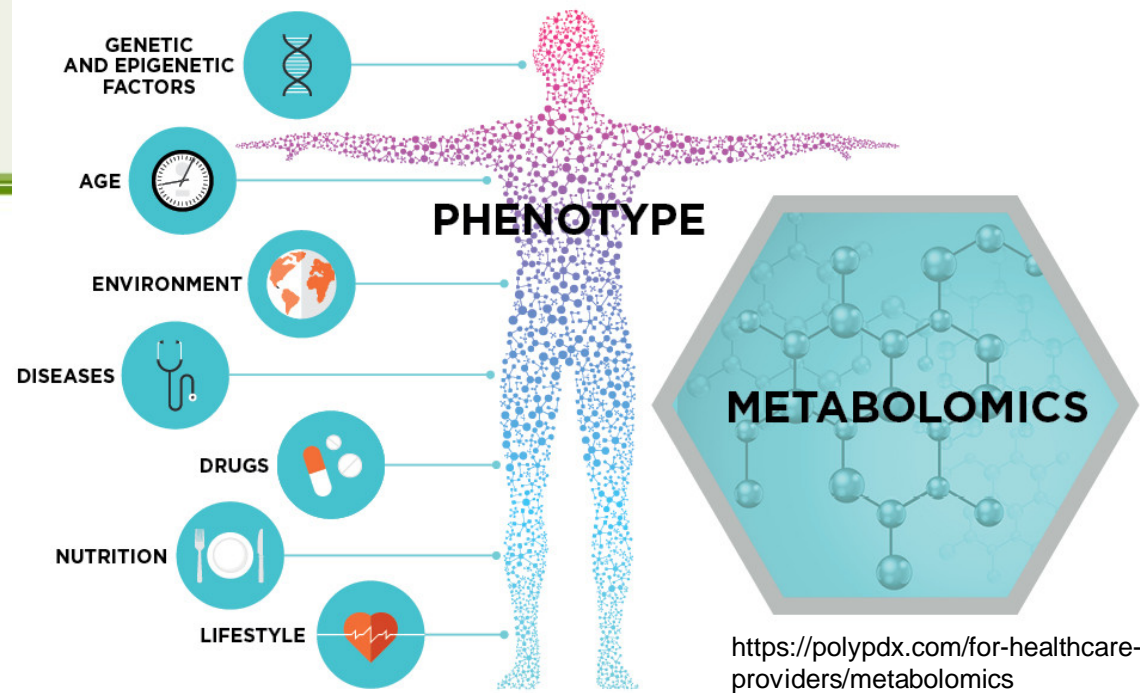

1,856
Small Molecules


222
Drugs


30,398
Literature References

Metabolomics

- Metabolomics – large-scale systematic study of the metabolome
- Metabolome - total complement of metabolites present in a biological sample under given genetic, nutritional or environmental conditions
 - the unique biochemical fingerprint of all cellular processes
- Metabolite - low molecular (usually 50 – 1,500 Da) weight organic compound, typically involved in a biological process as a substrate or product.
- Metabolomics yield many insights into basic biological research in areas such as systems biology, metabolic modelling, pharmaceutical research, nutrition and toxicology

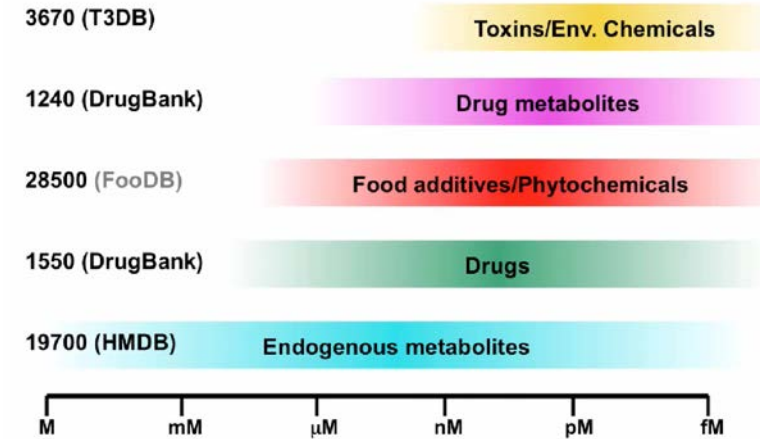


Metabolites are important

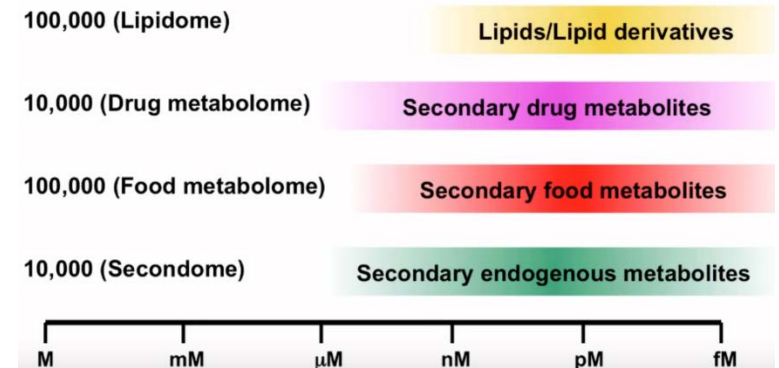
- **>95% of all diagnostic clinical assays test for small molecules**
- **89% of all known drugs are small molecules**
- **50% of all drugs are derived from pre-existing metabolites**
- **30% of identified genetic disorders involve diseases of small molecule metabolism**
- **Small molecules serve as cofactors and signaling molecules to 1000's of proteins**

Metabolomics can therefore be seen as bridging the gap between genotype and phenotype

Human Metabolomes (2015)



Theoretical Human Metabolomes

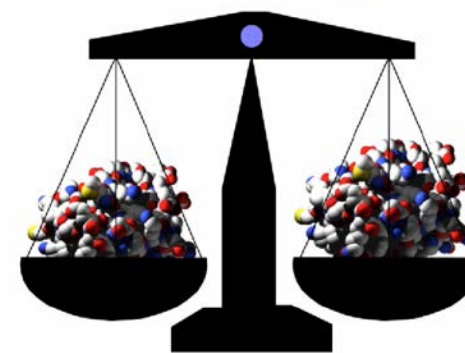


Metabolomics technologies

- UPLC, HPLC
- CE/microfluidics
- LC-MS
- FT-MS
- QqQ-MS
- NMR spectroscopy
- X-ray crystallography
- GC-MS
- FTIR

Mass Spectrometry

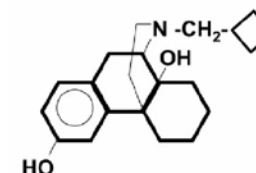
Analytical method to measure the molecular or atomic weight of samples



MS Principles

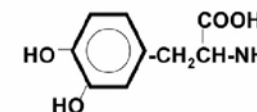
- Different compounds can be uniquely identified by their mass

Butorphanol



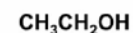
MW = 327.1

L-dopa



MW = 197.2

Ethanol



MW = 46.1

Metabolomics – ,a snapshot‘ in time

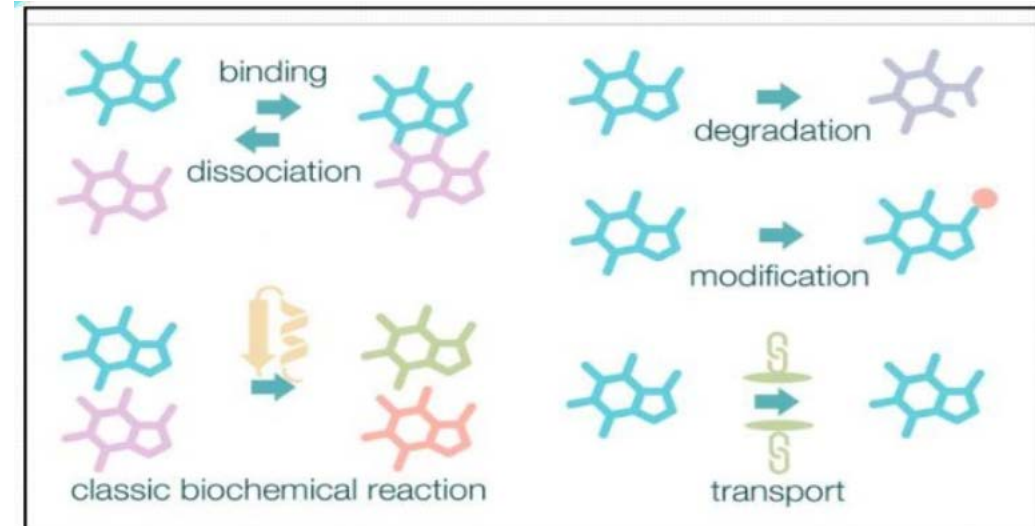
Conceptual approaches in metabolomics:

- Target analysis: has been applied for many decades and includes the determination and quantification of a small set of known metabolites (targets) using one particular analytical technique of best performance for the compounds of interest.

- Metabolite profiling: aims at the analysis of a larger set of compounds, both identified and unknown with respect to their chemical nature. This approach has been applied for many different biological systems using GC-MS, including plants, microbes, urine, and plasma samples.

- Metabolomics: employs complementary analytical methodologies, for example, LC-MS/MS, GC-MS, and/or NMR, in order to determine and quantify as many metabolites as possible, either identified or unknown compounds.

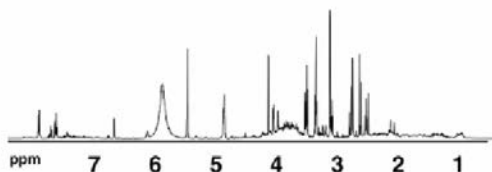
- Metabolic fingerprinting: a metabolic “signature” or mass profile of the sample of interest is generated and then compared in a large sample population to screen for differences between the samples. When signals that can significantly discriminate between samples are detected, the metabolites are identified and the biological relevance of that compound can be elucidated, greatly reducing the analysis time.



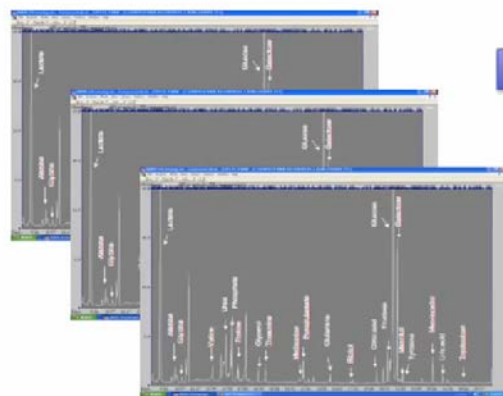
A diagram showing the main different types of metabolic reactions that take place in a cell. These are shown as they are represented in the database *Reactome*.

Metabolomics data analysis

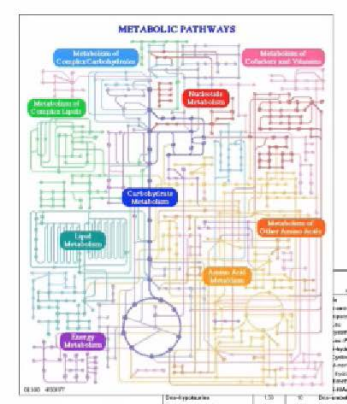
From Spectra to Lists



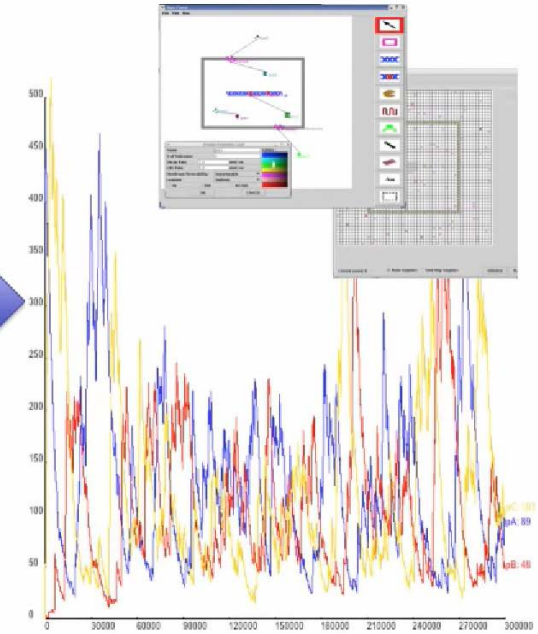
From Lists to Pathways



Compound	Retention Time (min)	Conc. in Urine (µM)	Compound	Retention Time (min)	Conc. in Urine (µM)
Dns-o-phospho-L-leusine	0.92	<0.L*	Dns-Ile	6.35	25
Dns-o-phospho-L-tyrosine	0.95	<0.L	Dns-β-aminovaleric acid	6.44	0.5
Dns-adenosine monophosphate	9.09	<0.L	Dns-pipecolic acid	6.50	0.5
Dns-α-phospho-β-alanine	1.06	90	Dns-L-iso	6.54	54
Dns-glucosamine	1.06	22	Dns-cystathionine	6.54	0.3
Dns-α-phospho-L-threonine	1.09	<0.L	Dns-L-iso-Pho	6.60	0.4
Dns-β-dimethyl-histamine putres	1.30	<0.L	Dns-5-hydroxylysine	6.65	1.6
Dns-3-methyl-histidine	1.22	80	Dns-Cytidine	6.73	100
Dns-leucine	1.25	534	Dns-N-acetylserine	6.81	0.1
Dns-carnitine	1.34	28	Dns-5-hydroxyproline	7.17	<0.L
Dns-dag	1.53	96	Dns-dimethylamine	7.33	205
Dns-Asn	1.55	133	Dns-5-HIAA	7.46	18
Dns-hypoxanthine	1.58	10	Dns-samboliferone	7.47	1.9
Dns-homocysteine	1.61	3.9	Dns-2,3-diaminopropanoic acid	7.63	<0.L
Dns-guanidine	1.62	<0.L	Dns-L-ornithine	7.70	15
Dns-Gln	1.72	633	Dns-4-acetylaminophenol	7.73	51
Dns-allantoin	1.83	3.8	Dns-proline	7.73	8.9
Dns-L-cystathionine	1.87	2.9	Dns-β-homocysteine	7.76	3.3
Dns-1-(or 3-β-methylhistamine	1.94	1.9	Dns-acetaminophen	7.97	82
Dns-adenosine	2.06	2.6	Dns-Phe-Phe	8.03	0.4
Dns-methylglycine	2.20	<0.L	Dns-5-methyl-xylylic acid	8.04	3.1
Dns-Ser	2.24	511	Dns-L-lys	8.16	194
Dns-aspartic acid amide	2.44	26	Dns-argin	8.17	<0.L
Dns-β-hydroxy-proline	2.56	2.3	Dns-β-ala-Phe	8.22	0.3
Dns-Orn	2.67	21	Dns-His	8.35	1860
Dns-Asp	2.90	90	Dns-4-thiobutane	8.37	<0.L
Dns-Thr	3.03	107	Dns-benzylamine	8.38	<0.L
Dns-epinephrine	3.05	<0.L	Dns-1-naphthol	8.50	0.6
Dns-ethanolamine	3.11	471	Dns-tryptamine	8.53	0.4
Dns-aminoadipic acid	3.17	90	Dns-pyridoxamine	8.94	<0.L
Dns-Gly	3.43	2510	Dns-β-methyl-cysteamine	9.24	<0.L
Dns-Asa	3.88	638	Dns-β-hydroxyprophane	9.25	0.32
Dns-aminolevulinic acid	3.97	30	Dns-1,3-diaminopropane	9.44	0.23
Dns-α-amino-β-tyrosic acid	3.98	4.6	Dns-putrescine	9.60	0.5
Dns-β-amino-hippuric acid	3.98	2.9	Dns-1,2-diaminopropane	9.66	0.1
Dns-5-hydroxymethylfural	4.08	1.9	Dns-tyrosinamide	9.75	29
Dns-β-hydroxybutamide	4.70	5.5	Dns-dopamine	10.06	140
Dns-β-alanine	4.75	<0.L	Dns-capsaicine	10.08	0.66
Dns-5-aminopentanoic acid	4.78	1.6	Dns-histamine	10.19	0.4
Dns-ascorbic acid	4.81	7.2	Dns-3-methoxy-tyramine	10.19	9.2
Dns-3-amino-isobutyrate	4.81	85	Dns-Tyr	10.28	321
Dns-2-aminobutyric acid	4.91	17	Dns-cysteamine	10.41	<0.L



From Pathways & Lists to Models & Biomarkers



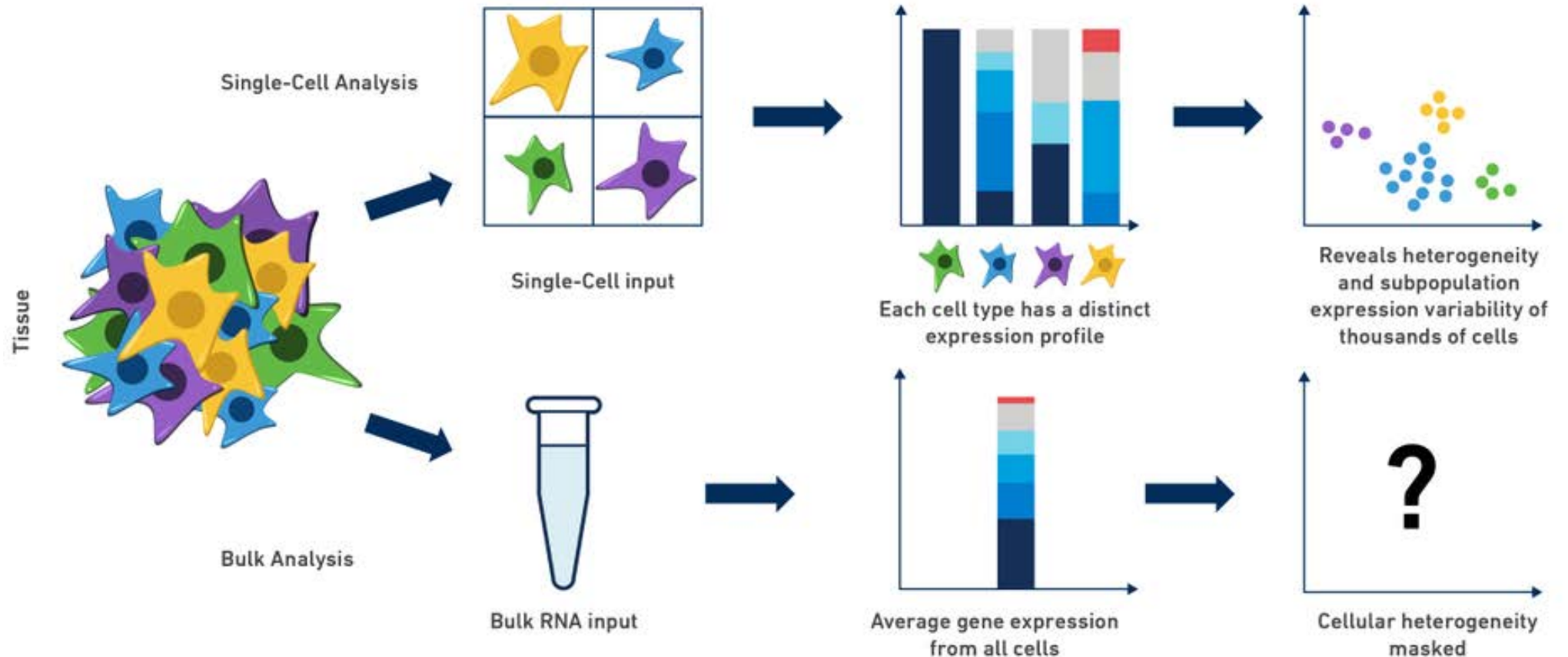
Where to look for metabolomics data

Metabolic pathway databases

- ▶ Pathway viewers KEGG (<http://www.genome.ad.jp/kegg/>),
- ▶ Atomic Reconstruction of Metabolism database (<http://www.metabolome.jp/>),
- ▶ BioCyc (<http://biocyc.org>) (Paley and Karp 2006),
- ▶ MetaCyc (<http://metacyc.org/>) (Caspi et al. 2006),
- ▶ AraCyc (<http://www.Arabidopsis.org/tools/aracyc/>) (Zhang et al. 2005), MapMan (<http://gabi.rzpd.de/projects/MapMan/>)
- ▶ (Thimm et al. 2004), KaPPA-View (<http://kpv.kazusa.or.jp/kappa-view/>) (Tokimatsu et al. 2005) and
- ▶ BioPathAT (<http://www.ibr.wsu.edu/research/lange/public%5Ffolder/>) (Lange and Ghassemian 2005),
- ▶ the data model for plant metabolomics experiments ArMet (<http://www.armet.org/>)

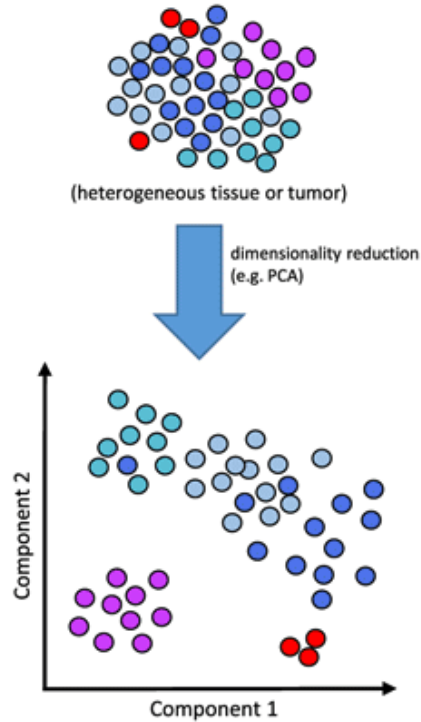
Cutting edge: Single cell -omics

- Application of whole genome, whole transcriptome sequencing and other -omics methods to single cells, sc RNA-seq is now the top method

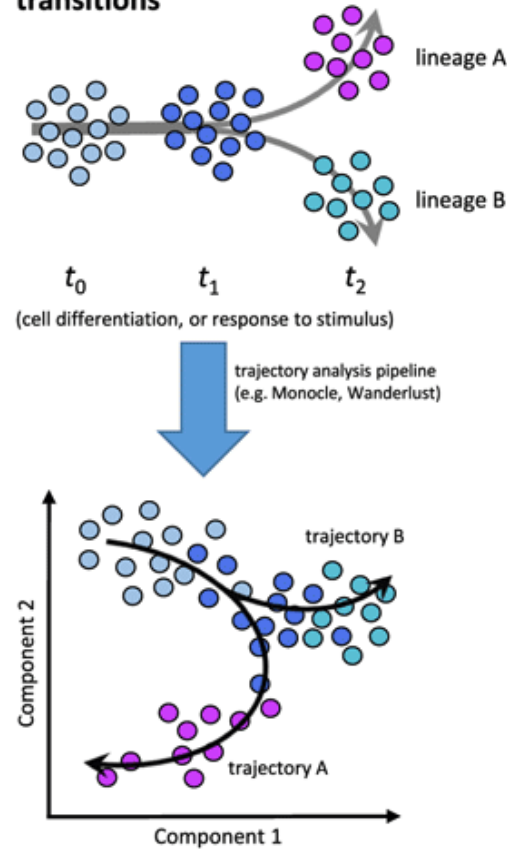


Common applications of sc RNA-seq

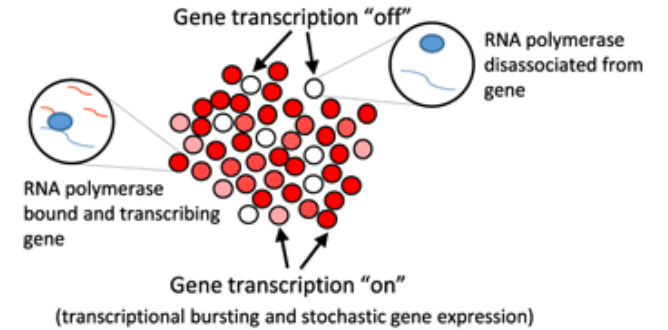
a) Deconvolving heterogeneous cell populations



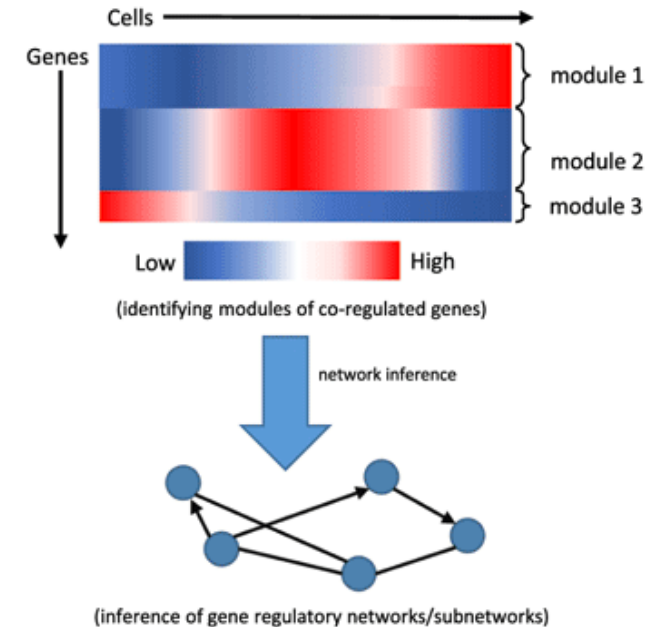
b) Trajectory analysis of cell state transitions



c) Dissecting transcription mechanics



d) Network inference



<https://f1000research.com/articles/5-182/v1>

For more info go at: <https://omicstools.com>

Single-cell multi-omics

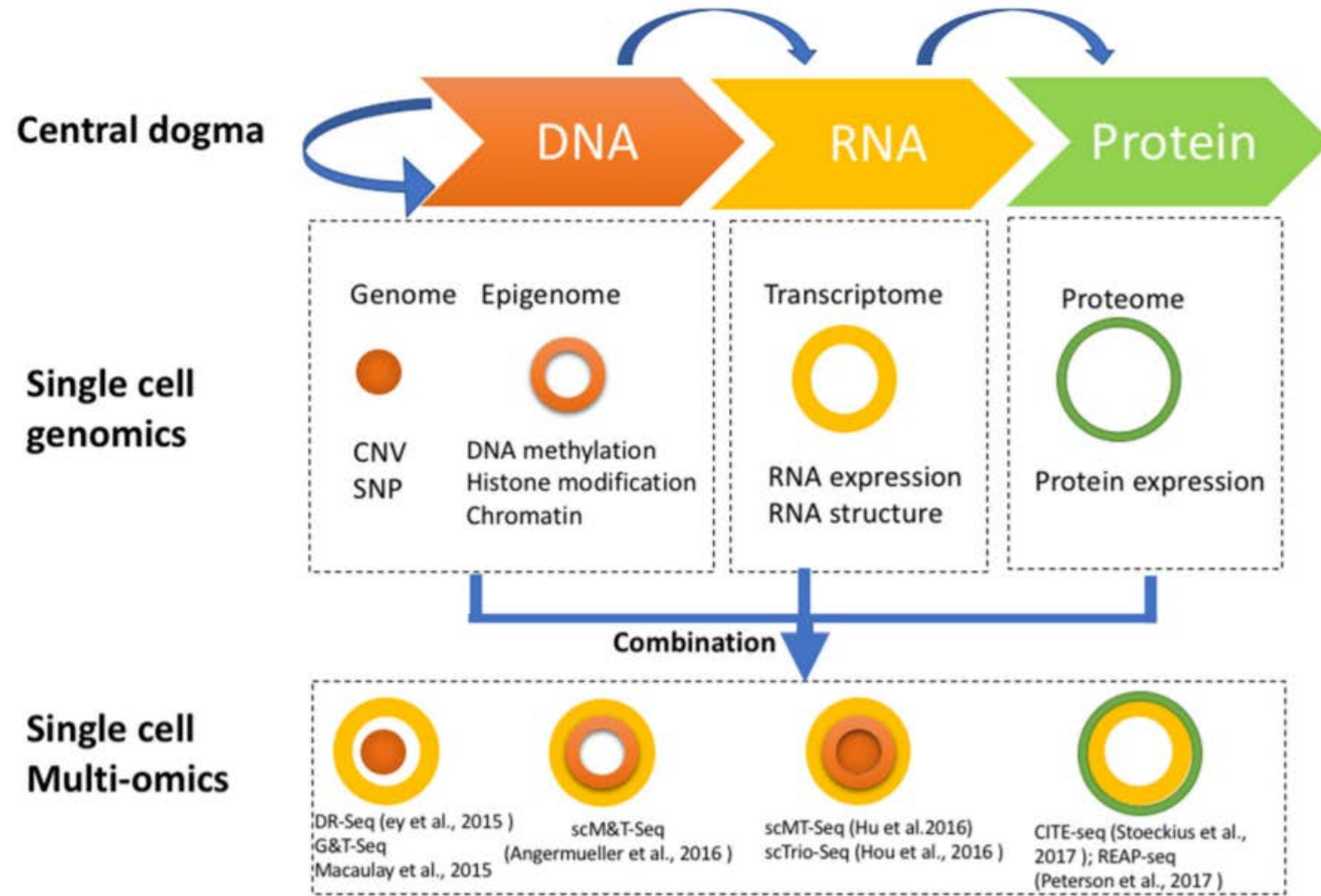


FIGURE 2 | Strategies for multi-omics profiling of single cells. Three major types of molecules relating to biological central dogma (Top). Single cell genomics methods profiling the genome, epigenome, transcriptome, and proteome are shown by different shapes with variable colors (Middle). Single cell multi-omics methods are built by combining different single cell sequencing methods to simultaneously profile multiple types of molecules of a single cell genome wide (Bottom). For example, G&T-seq was built by combining genome (orange) and transcriptome (yellow) to simultaneously detect DNA and RNA of the same cell genome wide.

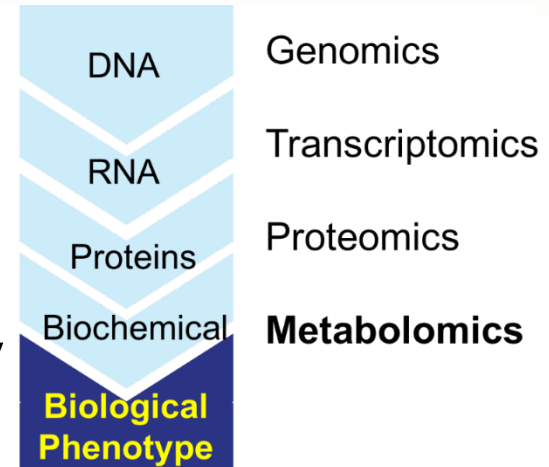
Challenges:

- There are no commercial kits available yet for any single-cell multi-omics techniques, and many are technically challenging.
- Researchers must modify existing single-cell protocols so that they're compatible with multiple types of molecules and take great care to minimize the loss or contamination of samples

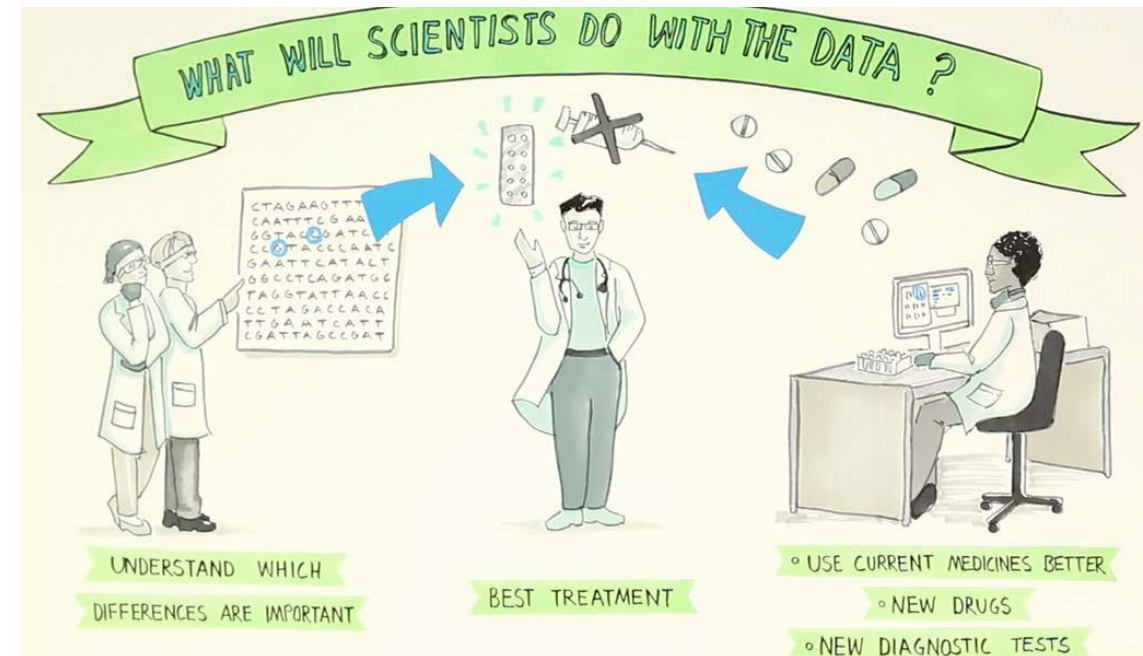
<https://www.the-scientist.com/lab-tools/integrating-multiple--omics-in-individual-cells-64829>

Summary

- Omics technologies - „the data deluge“
- Genomics and Transcriptomics rely on two main approaches: microarrays (hybridization) and NGS (sequencing by synthesis)
- Proteomics and Metabolomics rely heavily on mass spectrometry



- Omics technologies are revolutionizing science and medicine
- From data to actionable knowledge - Integrated Omics data
- Precision medicine is the ultimate goal of many –omics efforts
- Despite the progress made we have still a long way to go ...



Take home messages

- We have been generating Big data, but we hardly understand it 😞
- Big data is publicly available, go through the databases before you even start even planing your experiment – it can save you enourmous time and money
- Databases contain huge datasets of patients you would never be able to gather by yourself, test your hypothesis in silico before the „wet-lab“ work
- If you cannot find the „yes/no“ or „a few genes“ answer, use the Cellular/functional/pathway analyses to help you out 😊
- Learning bioinformatics skills (e.g. programing in R) is a good investment plan for your future career

Thank you for your attention

Any Questions?

- Jay Flatley, Executive Chairman of Illumina:
- *„Everyone is going to get sequenced, it is gonna be part of their health record and it will be used to manage their health care throughout their lifetime“.*

