High-throughput data generation

for
biomedical applications

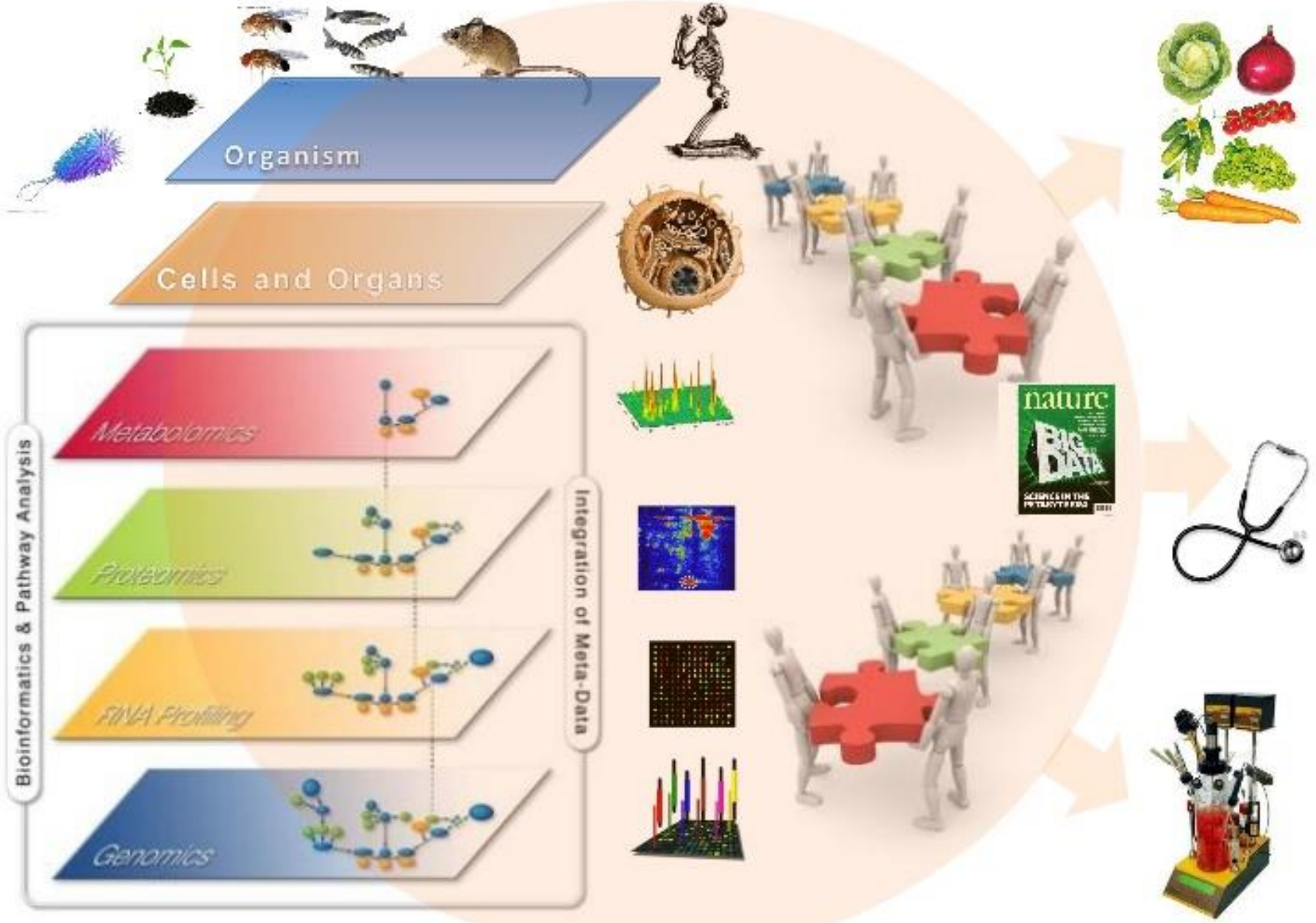
Dr Vytaute Starkuviene-Erfle
„HiCell – High content analysis of cell“
„Integrin trafficking networks“
Heidelberg University

Topics of today

1. Large scale data for biomedical applications
2. An overview: high throughput data generation
3. A closer look: phenotypic screening
4. Large data sets for in depth information

High throughput data analysis – next lecture, Dr Apic

Life Science data: Multi-omics, multi-technology, multi organism, multi dimensional



Drug discovery - one of the most complicated projects of the humankind

When did it started?

Kahun Gynaecological Papyrus (1800 BC, Egypt)

Ebers papyrus (1500 BC, Egypt) contained more than 800 remedies from herbs

Start of modern drug discovery

Developments in pharmaceutical manufacturing

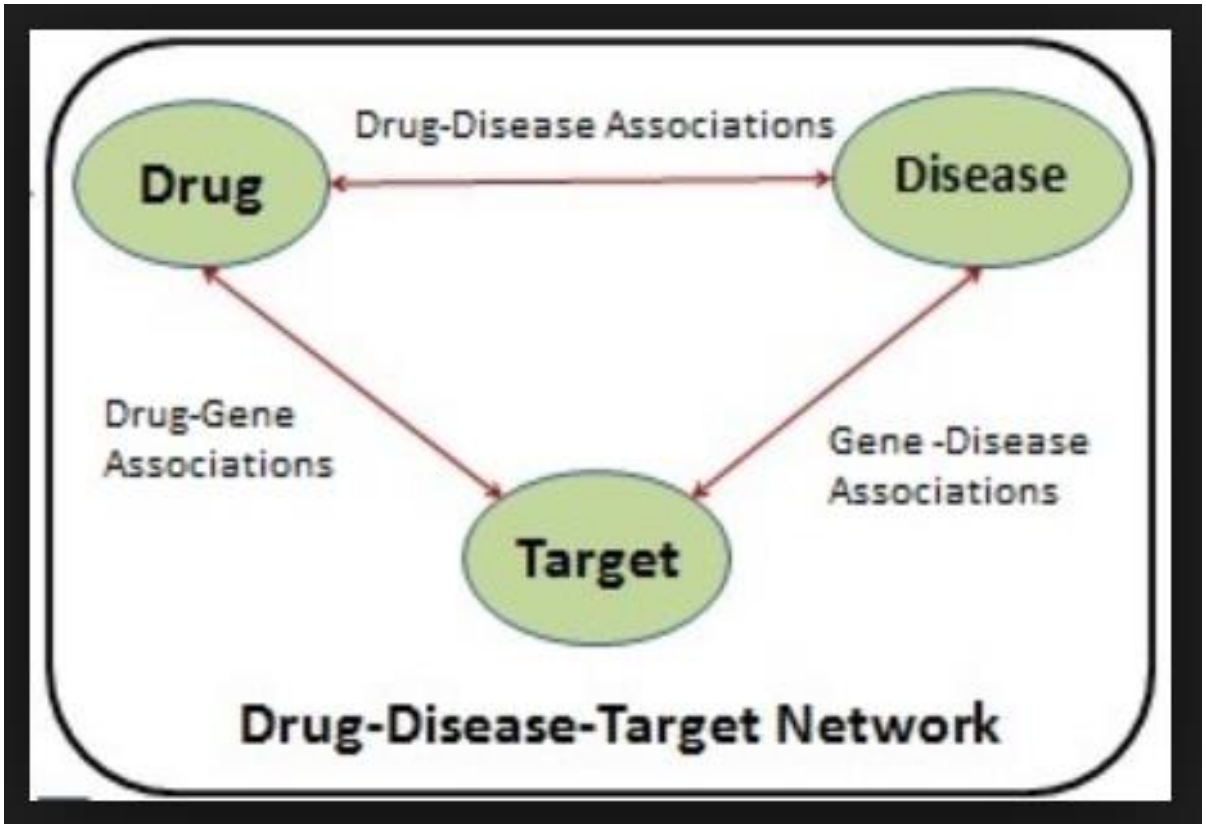
Merck is established (1668) – as the „Angel Pharmacy“.

Morphine was isolated and purified from opium extract (1805) to treat pain

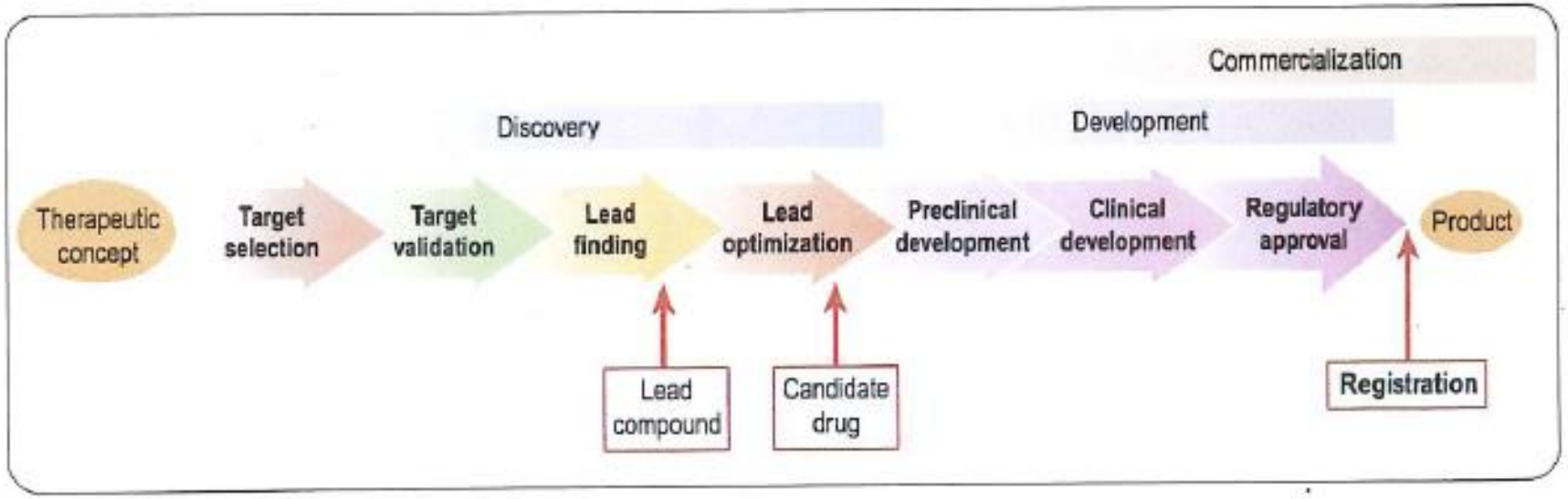
Merc commercially produced morphine since 1827 and other substances, calling the activities "Cabinet of Pharmaceutical and Chemical Innovations.

Bayer developed heroin (1898) as a non-addictive alternative to morphine. It is the first derivative of the natural product (diacetylmorphine)

Bayer developed aspirin (1899)


Drug
Drug-Disease Associations
Disease
Drug-Gene
Associations
Gene -Disease
Associations
Target
Drug-Disease-Target Network

Phases of modern drug discovery

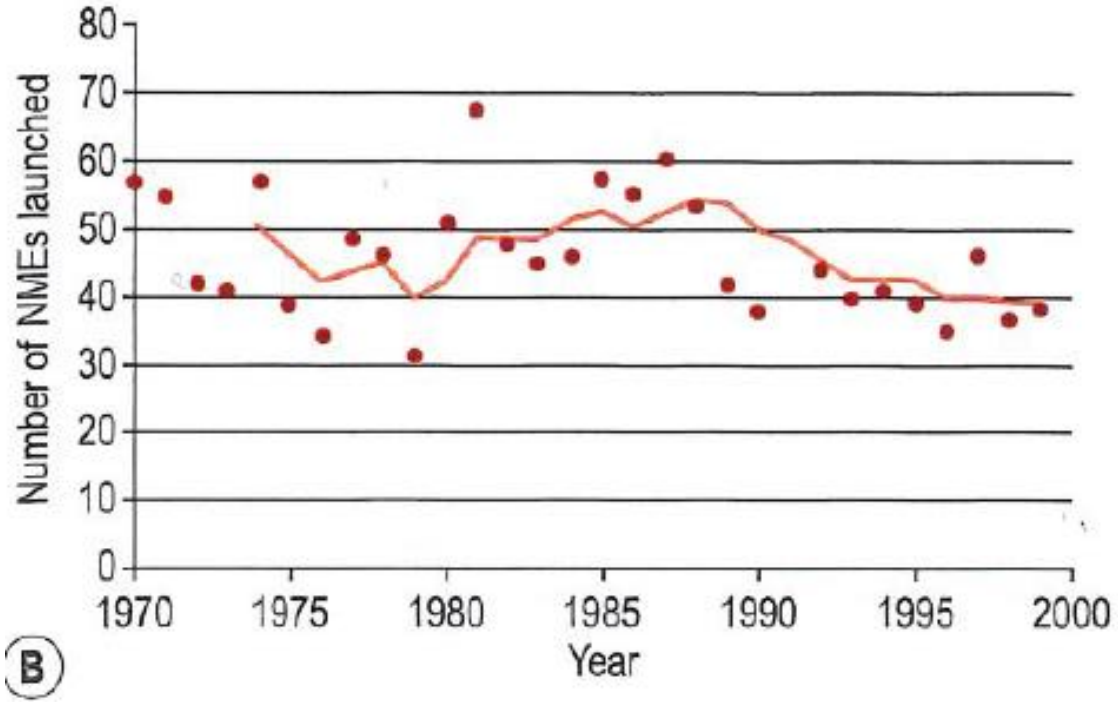


Phase I – drug discovery – from the concept to molecule
Phase II – drug development – from molecule to registered product
Phase III – commercialization – from product to therapeutic application to sales

Hill and Rang, Drug Discovery and Development

Productivity of drug discovery

NME – number of molecular entities



„not much so far“

tight safety and efficiency regulations
long development time (up to 20 years)
high costs (up to 2Mrd USA $)

Hope: combination of new disciplines !

genomics, proteomics, bioinformatics, structural genomics, high-throughput, computational chemistry

Hill and Rang, Drug Discovery and Development

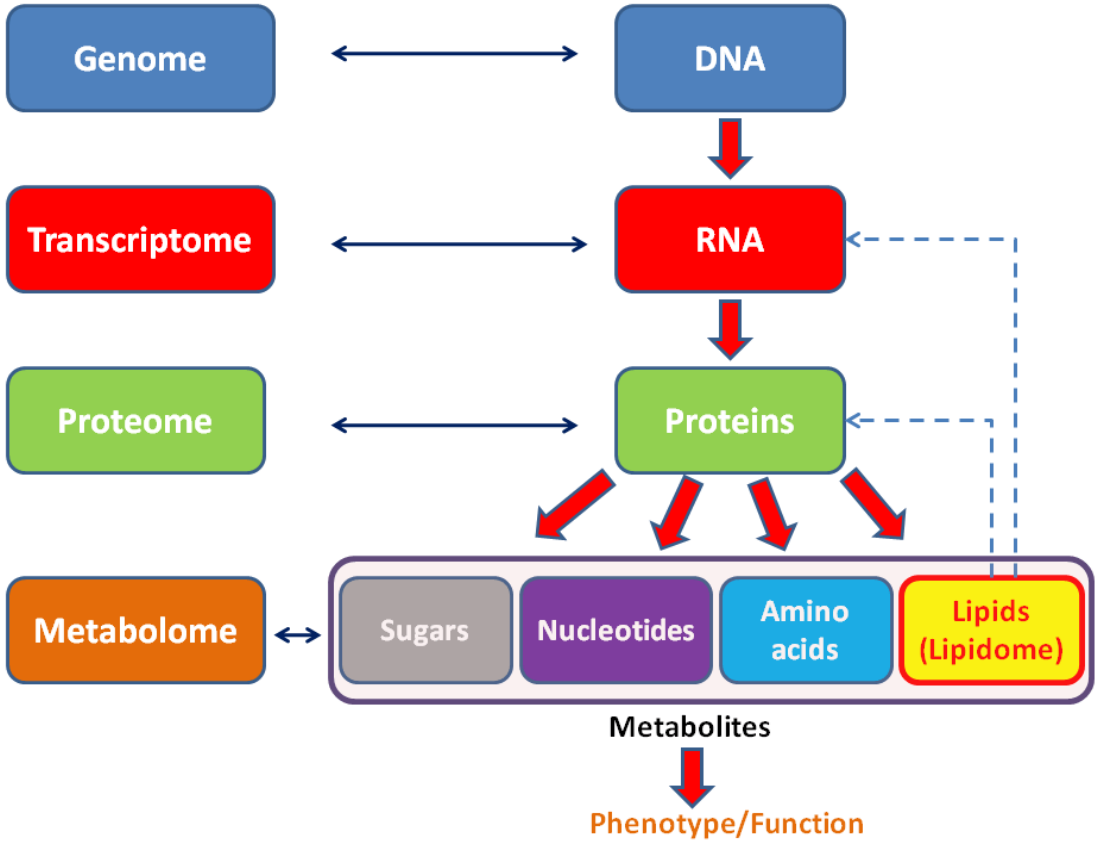
High-throughput biology or „omics“ research

What can happen
~20 000 genes

What appears to be happening
~150 000 transcripts

What makes it happen
~ 1000 000 proteins

What actually happens
~3000 metabolites

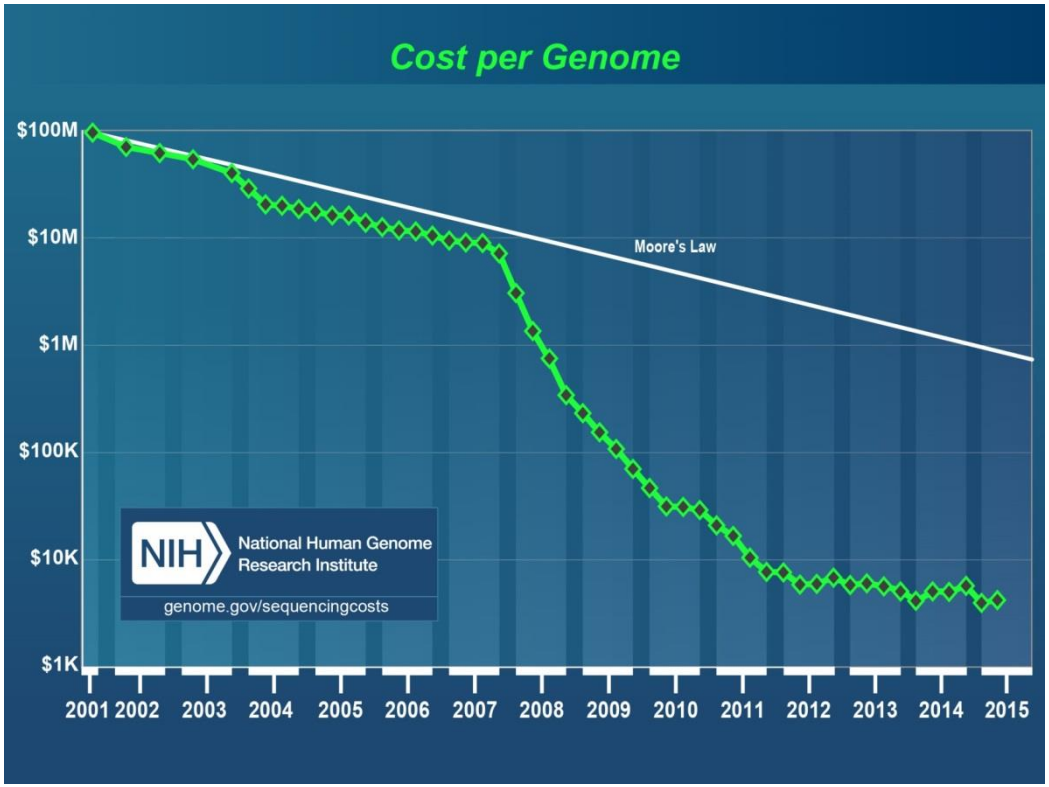


Genomics - static information on all information encoded in the genome
Genomes the term coined 1986
The major method: DNA sequencing

a 96-capillary sequencer: 10 Kbp
1 human genome, 30x coverage, ~80 years

5 billion read pairs, 1,5 Tbp
8 human genomes, 30x, ~4 days

Sequencing costs



"… the first human genome sequence, announced in April 2002, utilized the expertise, infrastructure, and people from 20 institutions and took 13 years of work and about $3 billion to determine the order of approximately three billion nucleotides. Now we can sequence a human genome for $1,000, and we can generate more than 320 genomes per week."

A century's worth of Roche R&D data were more than doubled in 2011–2012 in a single large-scale experiment to sequence hundreds of cancer cell lines.

May, 2014

TARA

http://oceans.taraexpeditions.org/en/m/about-tara/

The mission of the expedition is to bring back quantitative and qualitative data on the composition of these ecosystems as a function of geographic position and environmental conditions. The goal of this collection of samples and data is **3 fold**:

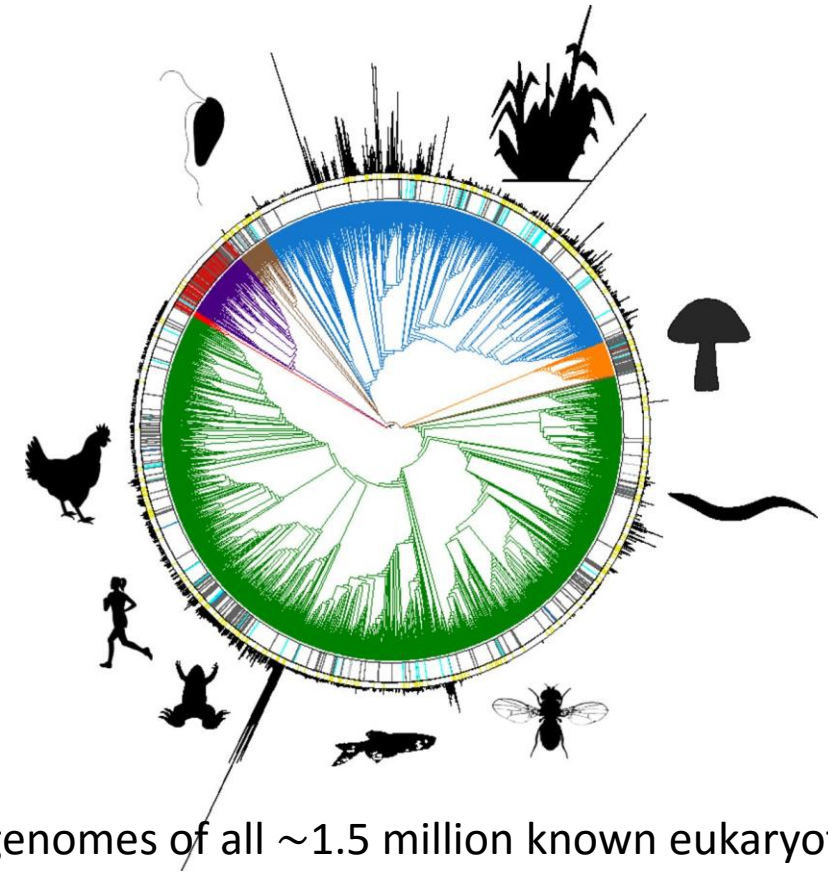
to feed morpho-genomic analyses of marine ecosystems in order to better understand the nature of the organisms and genes expressed in a given oceanic environment,

to better understand the evolution of marine organisms and

to feed models of the co-evolution of these ecosystems with the hydro-climate.

revealing tens of thousands of new eukaryotic species, along with 40 million mostly novel genes from the viruses, bacteria and single-celled creatures that were collected

Earth BioGenome Project



The Earth BioGenome Project (EBP), a *moonshot* for biology, aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years.

genomes of all ~1.5 million known eukaryotes, up to 100,000 new eukaryotic species can be sequenced to a high level of completeness and accuracy for approximately US $4.7 billion

Lewin et al., 2018, https://www.earthbiogenome.org/

Genomics Big Data

Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Stephens et al., 2015

Dynamic information encoded in the genome

Transcriptome: gene expression profiling

Only a part of the genome is expressed in any given moment in both physiological and pathological conditions.

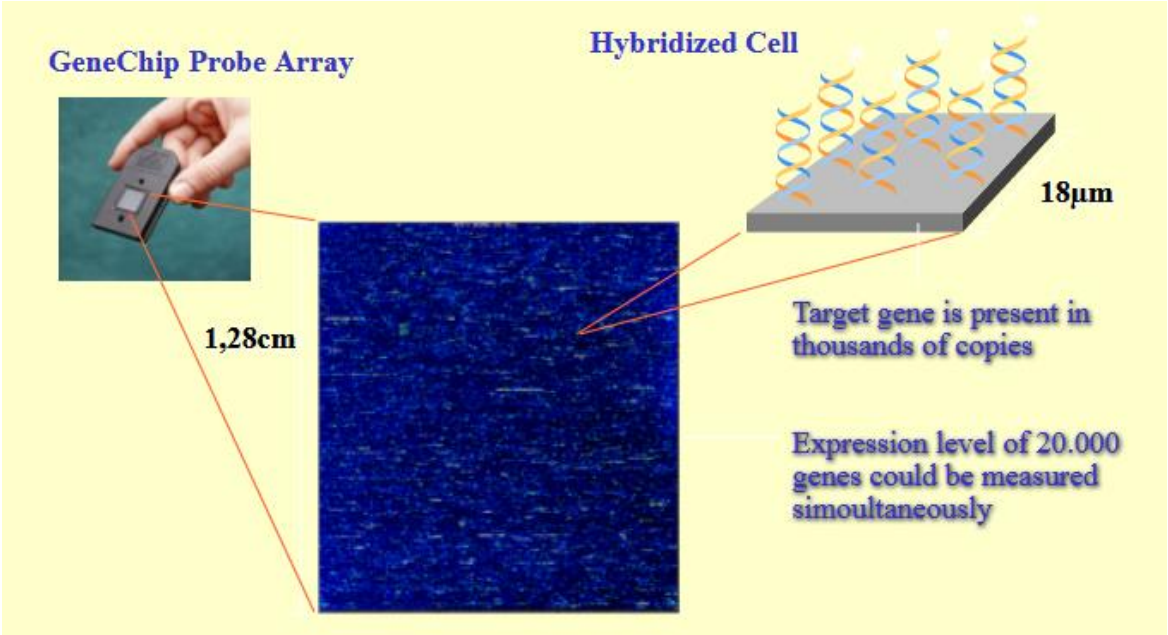
Identification of difference in expression profile between physiological and pathological states could lead to the identification of new targets.

Differential display

Subtractive cDNA library

S.A.G.E Serial analysis of gene expression

Microarray

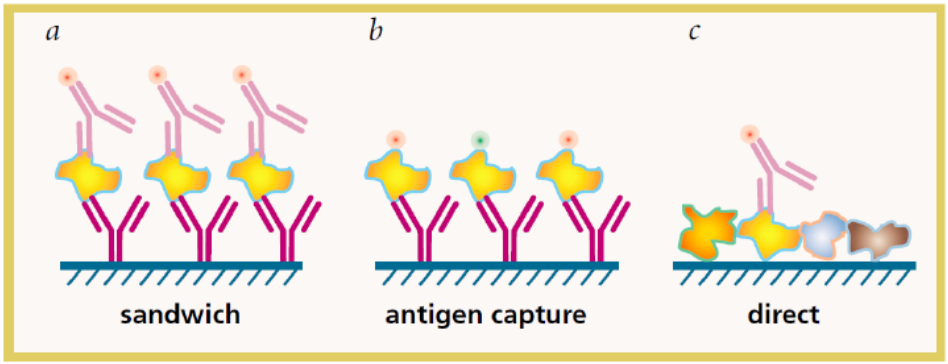


Proteomics

- Proteomics: a collection of various technical disciplines:
  - imaging: microscopy techniques
  - **protein microarrays/ chip experiments**
  - mass spectrometry-based proteomics

Protein Microarrays

Different array strategies:

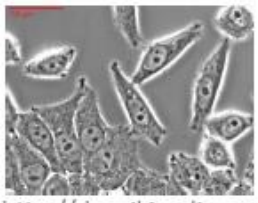


Advantages:
- relatively fast
- less costly than
  e.g. MS-based methods

Disadvantages:
- Antibodies not always available
- lower resolution than MS-based methods
  → cross-reactions
- epitopes often unknown (e.g. for PTM)
- lower sample size than MS-based methods

courtesy of Nadine Veith

Mass spectrometry-based proteomics

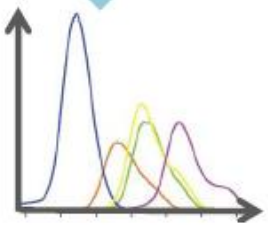


http://de.wikipedia.org/

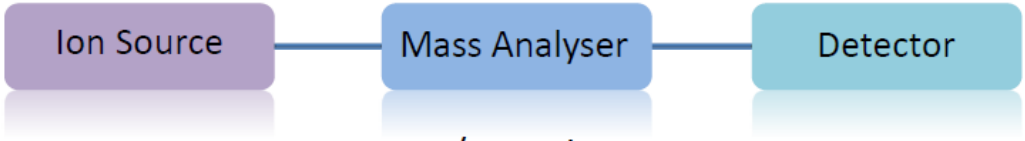
protein extraction

protein digestion

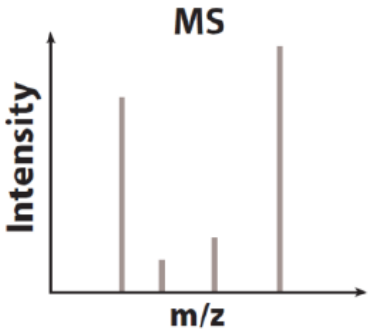
fractionation



courtesy of Nadine Veith



m/z - ratio

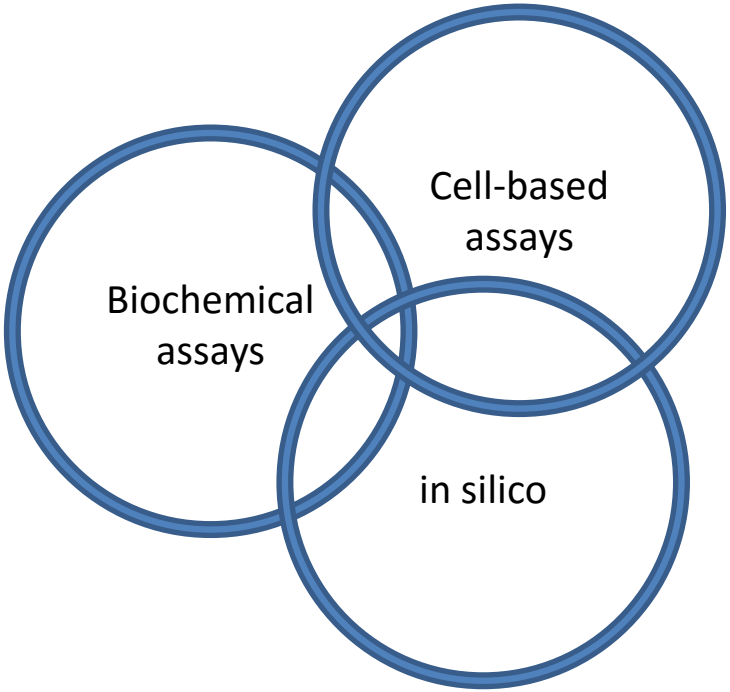


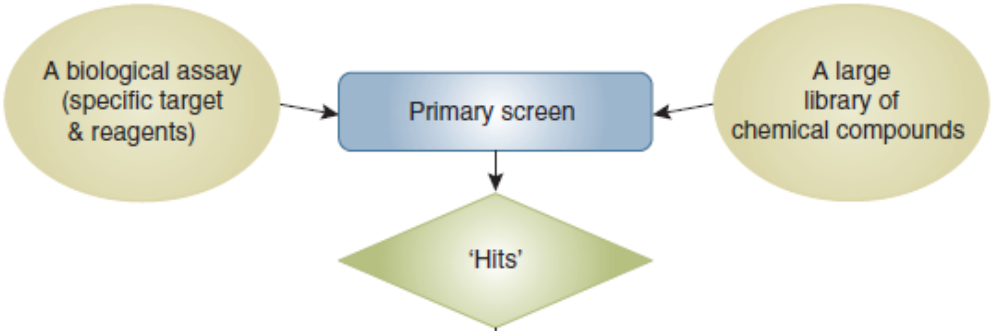
MS ion spectra

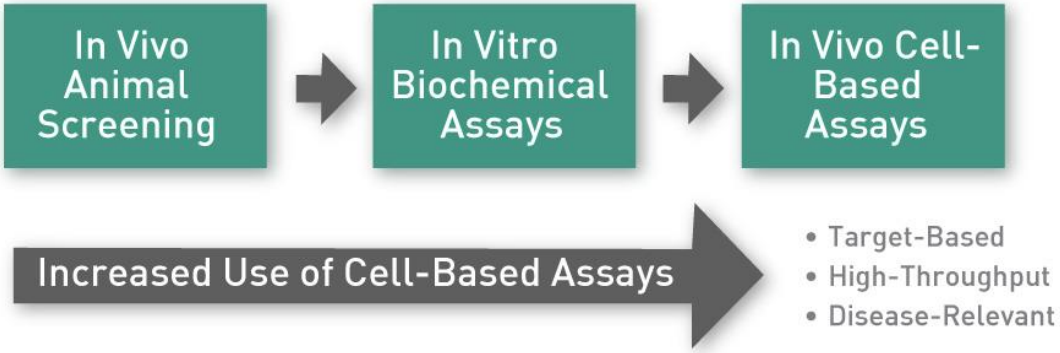
Clinical sequencing and transcription analysis, but no clinical proteomics yet

HTS – High Throughput Screening

Genetic or chemical screening - analyzing a large number of biological or chemical compounds against specific targets or phenotypes, or to identify specific targets.

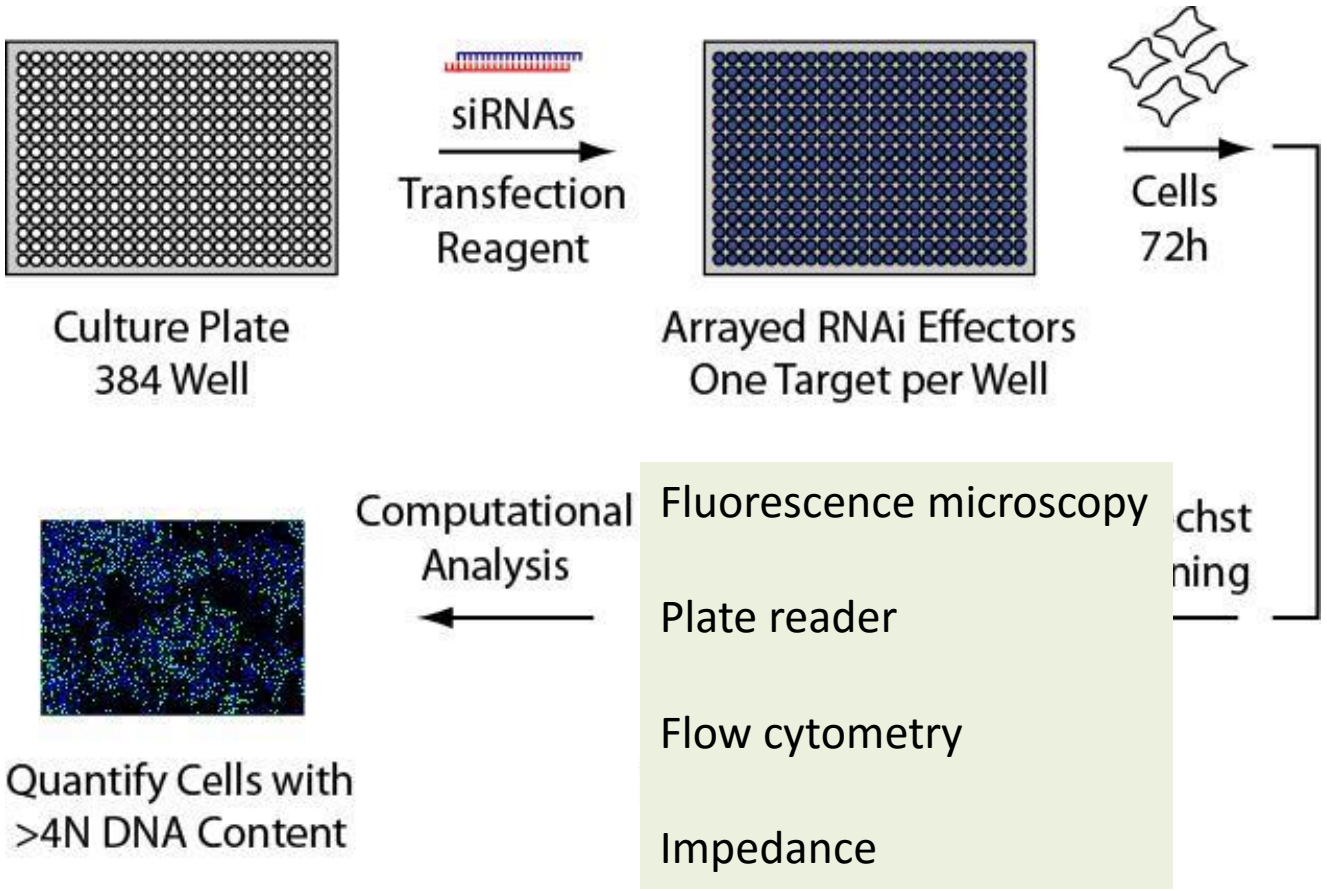






Janzen, Chem Biol, 2014

High-throughput phenotypic screening

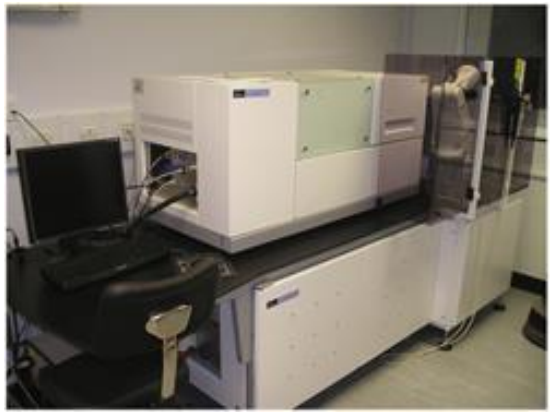


Boutros et al., 2015

Microscopy advantages in cell-based screening

in vitro – cell cultures

in vivo – model organisms

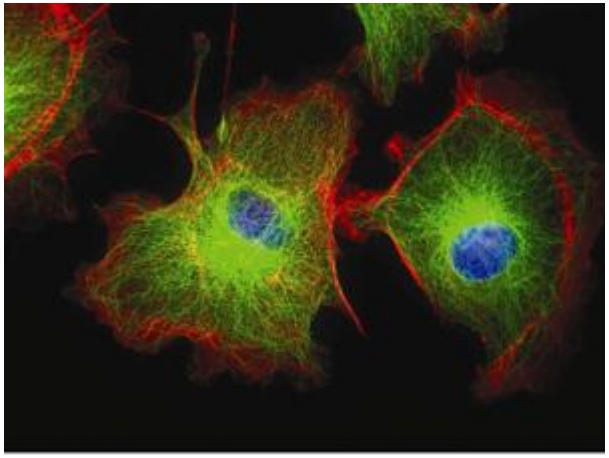


high- throughput

ultra-high-throughput

high-content

high-resolution



Individual cells

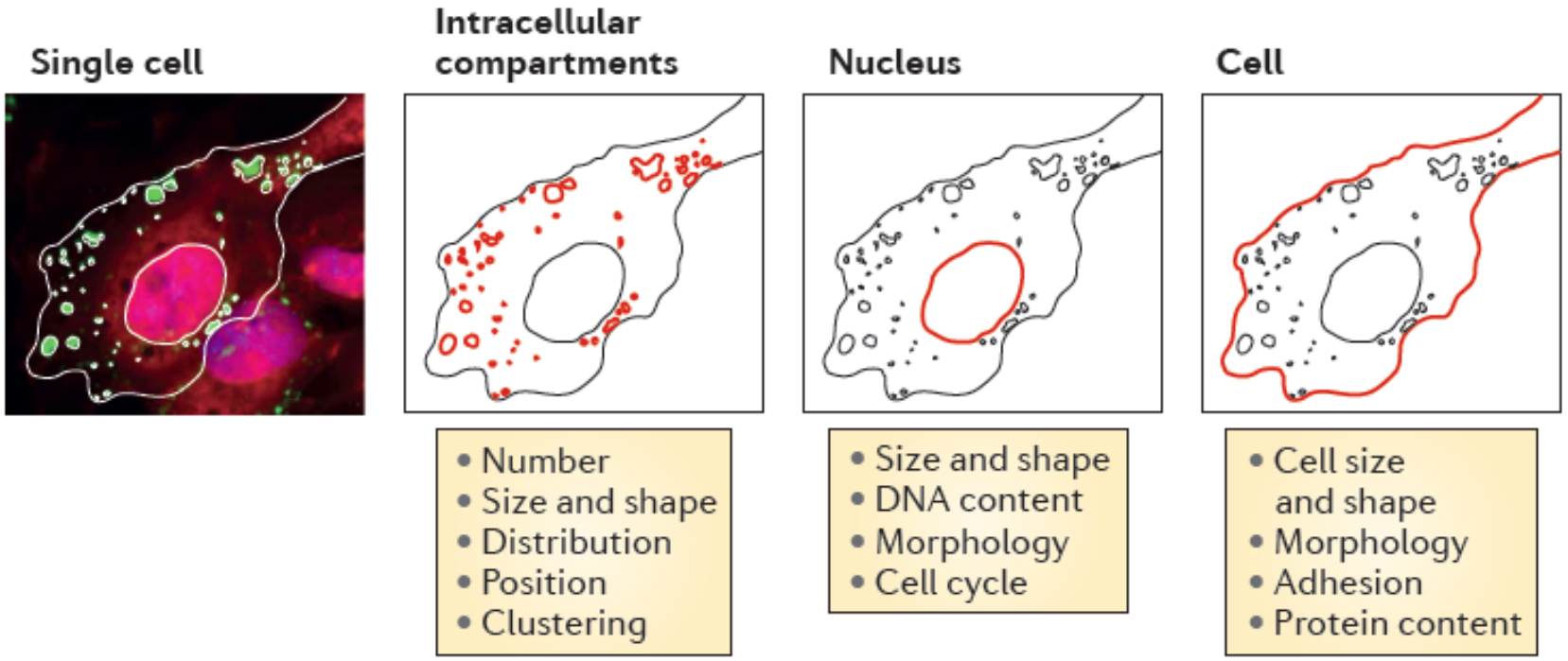
Living cells

Subcellular organelles

Multiplexing

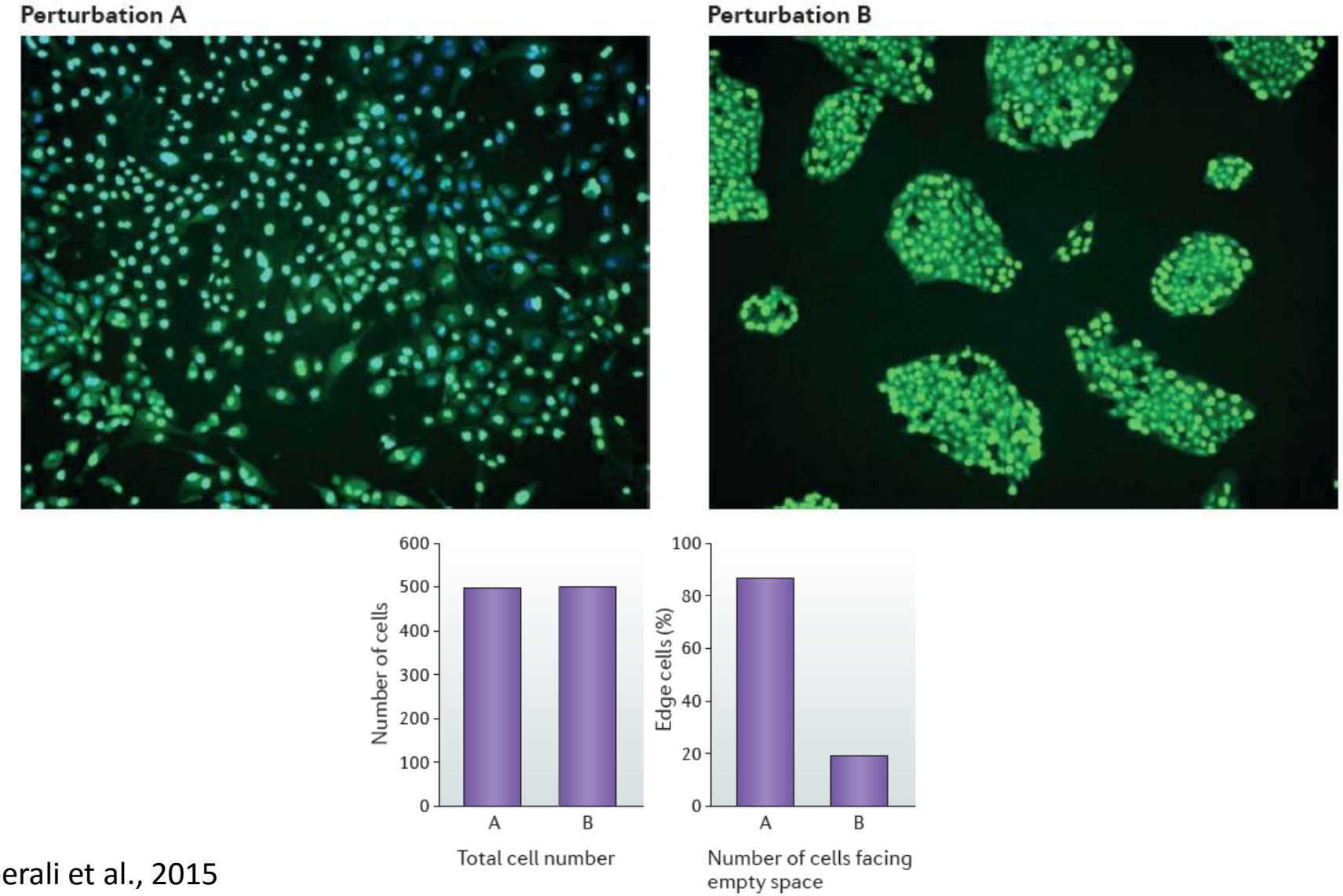
Multiparametric feature extraction

difficult to achieve by other methods



Liberali et al, 2014

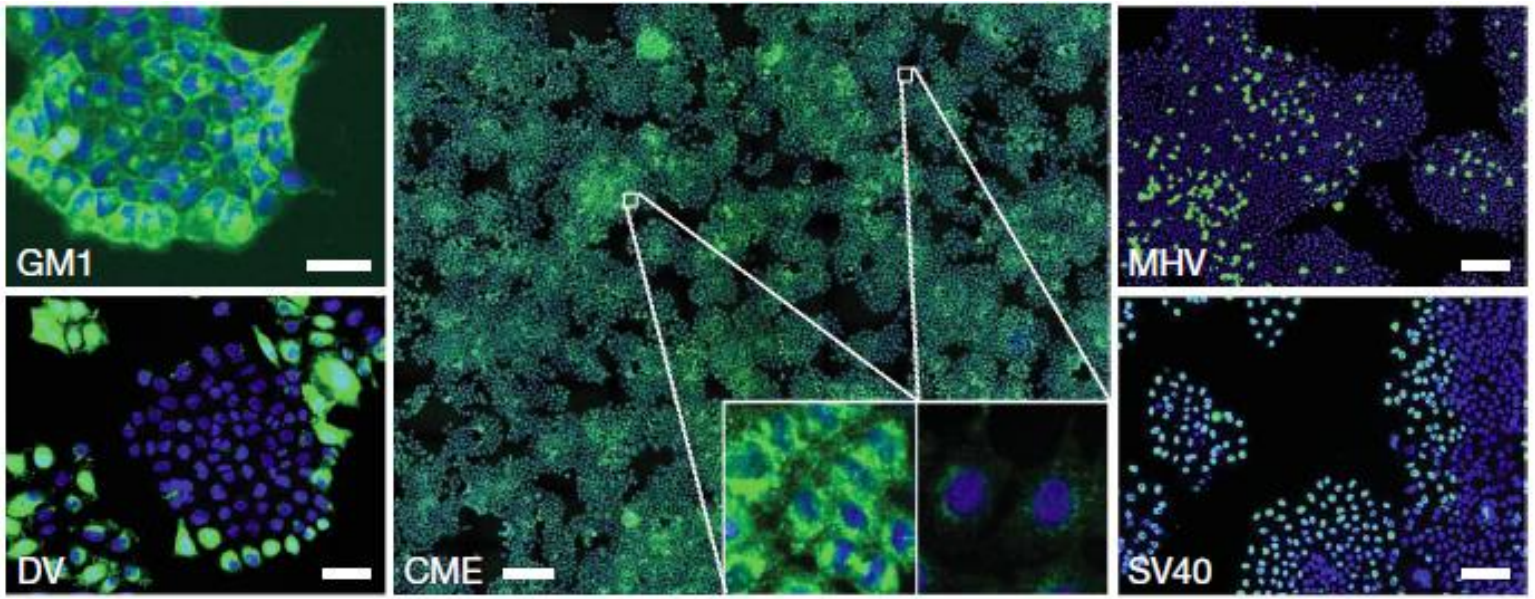
Variations of cell populations



Liberali et al., 2015

Impact of variations of cell populations

Genetically identical cells display variable activity and behaviour



enrichment on edges

crowded versus spare regions

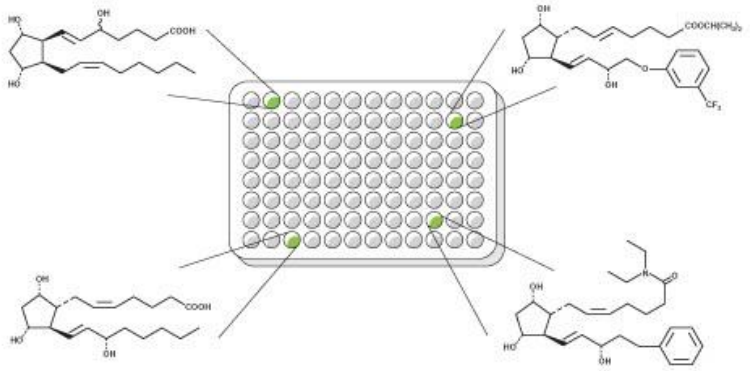
Heterogenity of viral infection can be traced to individual cell states

Snijder et al., Nature, 2009

DV – dengue virus
CME – clathrin-mediated endocytosis
MHV – mouse hepatitis virus
GM1 - sphingolipid

Interaction disease - drug

- **Phenotypic screening** is applied for nearly 75% of cases in the current drug discovery activities



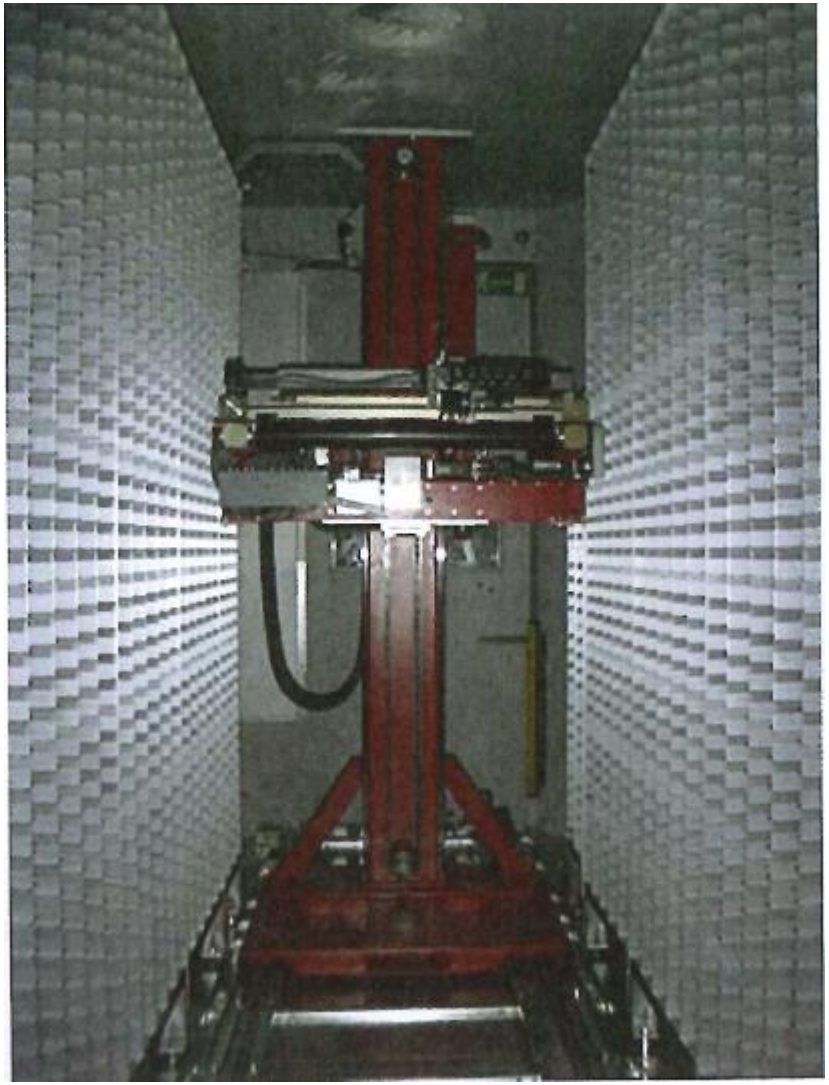
Janzen, Chem Biol, 2014

The first chemical screening

Ehrlich and Bertheim, 1906 – synthesized Salvarsan (arsphenamine) to treat syphilis
The compound 606 was selected from more than 600 compounds based on their effect to heal the infection and was marketed by Hoechts company

Hill and Rang, Drug Discovery and Development

High Throughput Screening for Drugs

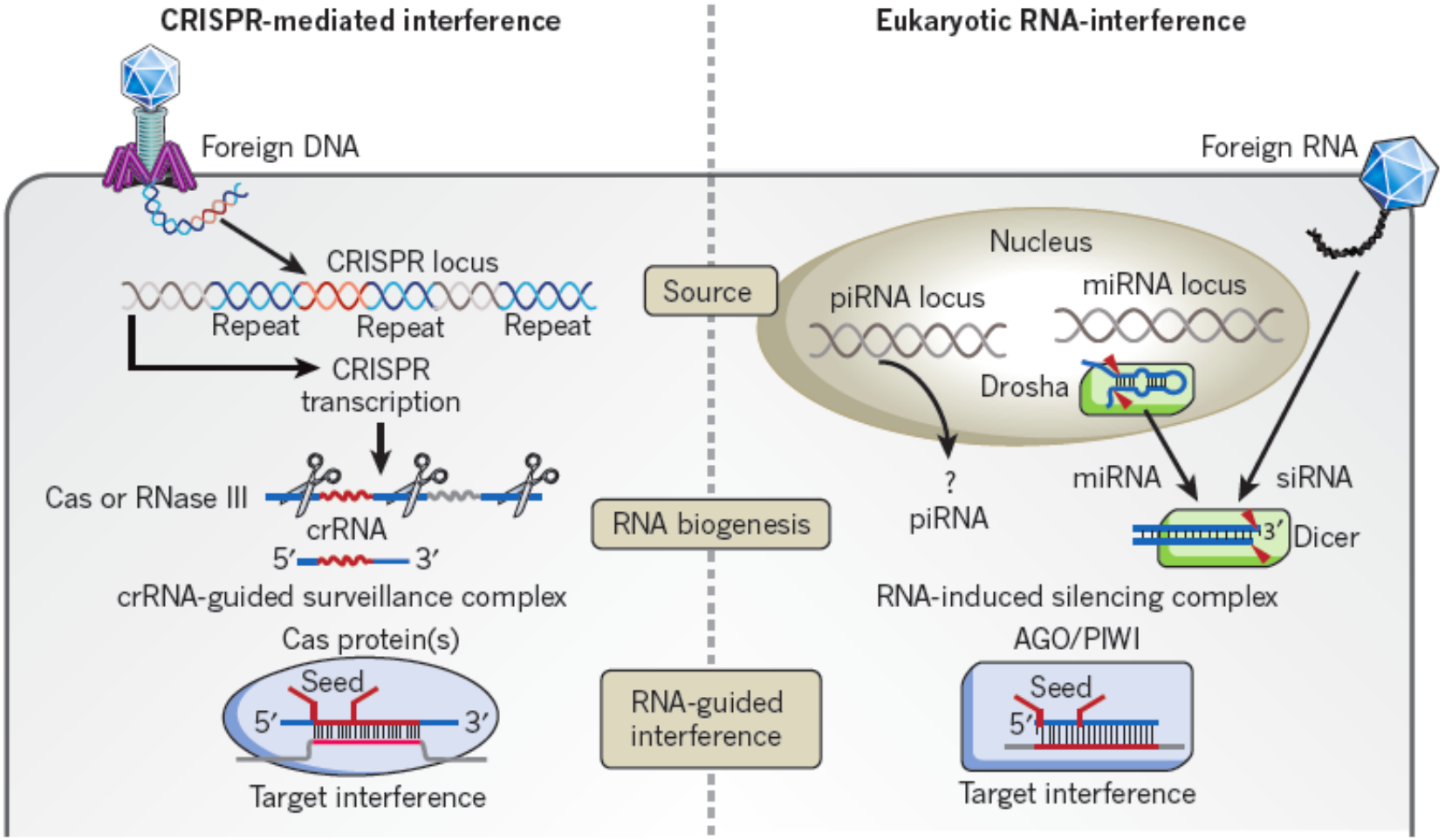


High degree of automation and robotics in handling, storage and data analysis

Hill and Rang, Drug Discovery and Development

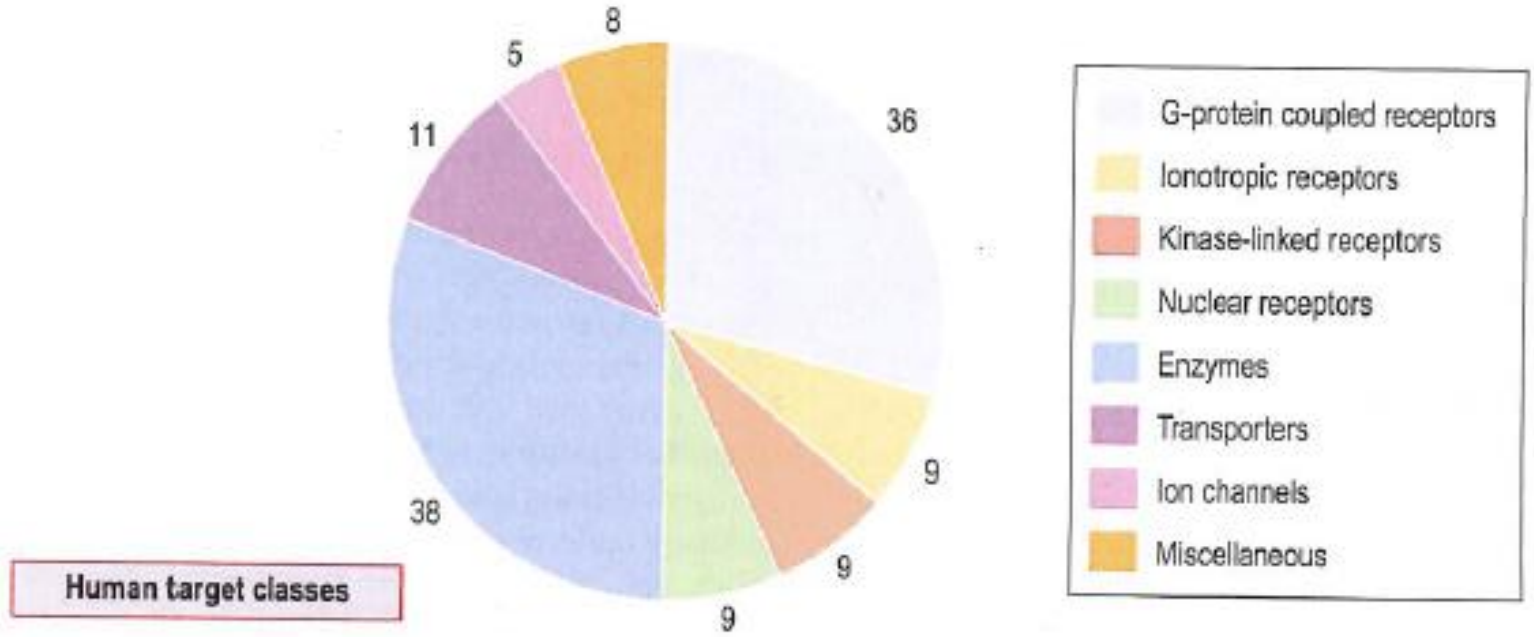
Interaction disease - target

RNA interference and CRISPR-Cas9



Wiedenheft et al, Nature, 2012

Do we need more drug targets?

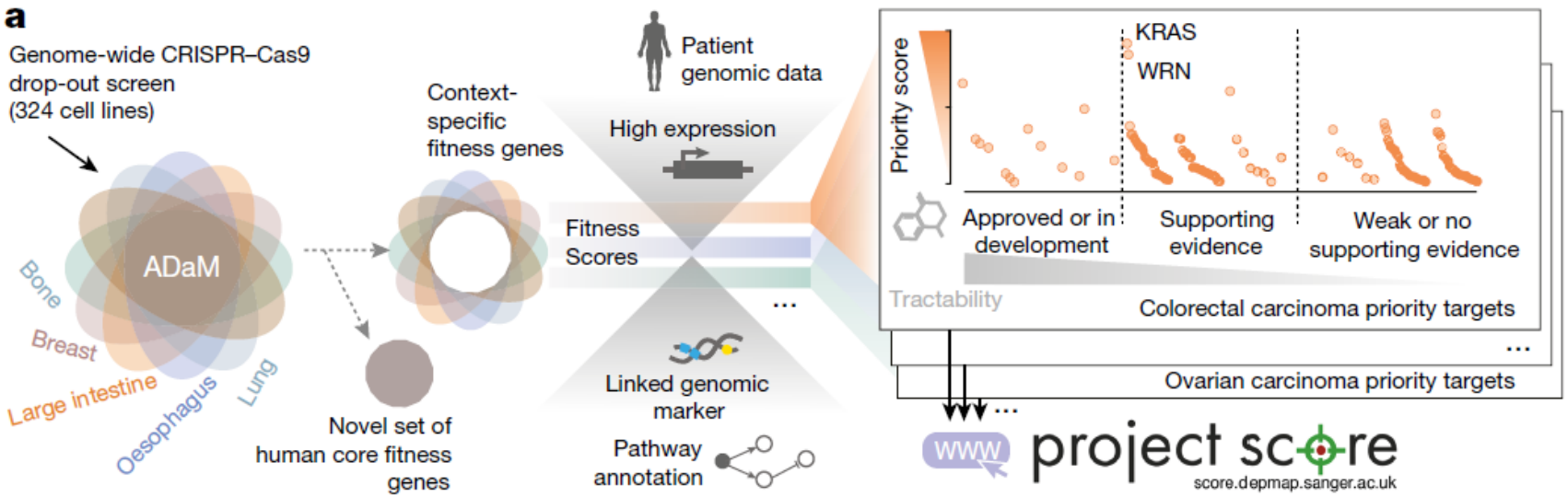


Drews and Ryser, 1997 – 500 targets addressed by the available drugs
Hopkins and Groom, 2001 – 120 targets
Zambrowicz and Sands, 2003 – 100 targets

100 best selling drugs target 43 proteins

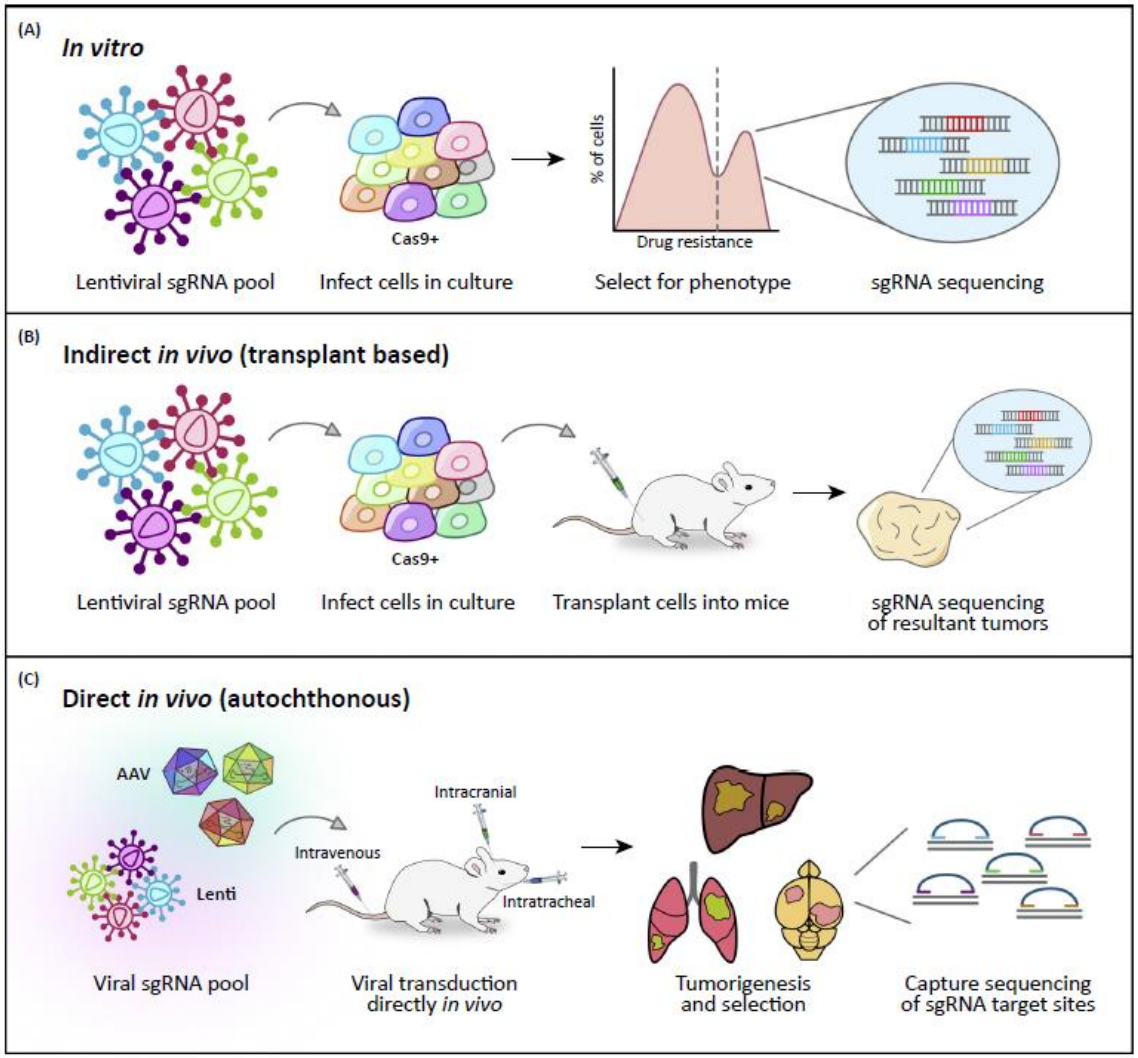
Hill and Rang, Drug Discovery and Development

Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens



Behan et al, 2019

Examples of large data sets

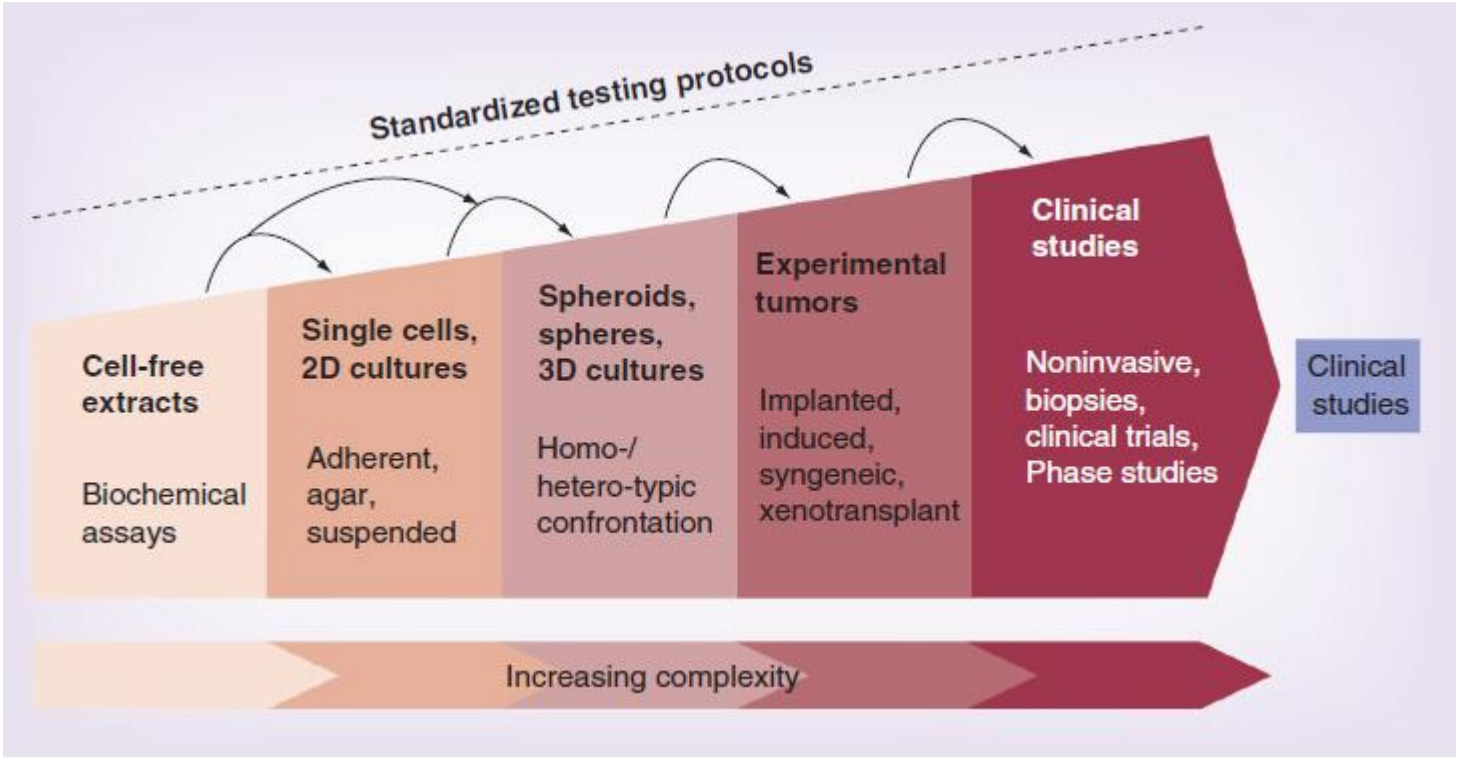


Many samples:
High throughput
High accuracy
Limited direct clinical applications

Few samples:
Disease-relevant
Poor coverage
Poor statistics

Chow and Chen, 2018

Large sets of data for in depth information



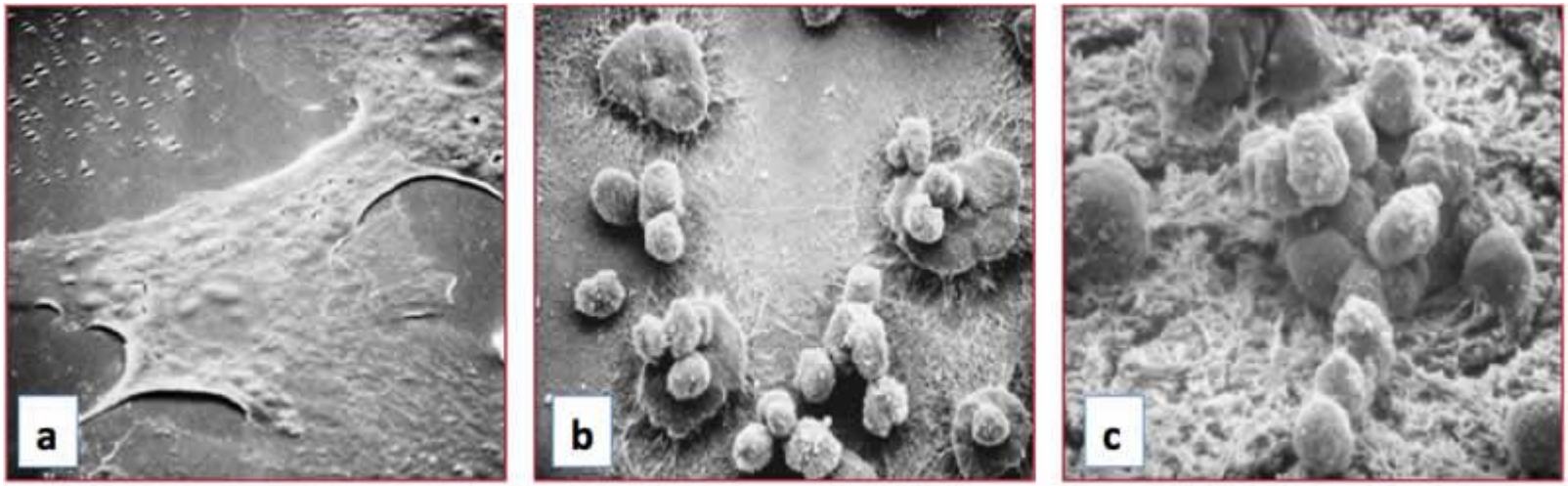
Benien and Swarni, 2014

Comparisson between 2D and 3D cultures

2D cultures loose the properties of their original tissues. Cell adhere only by the side, which is in contact to the surface (~50% of the cell surface)

Flat = 3 μm thick

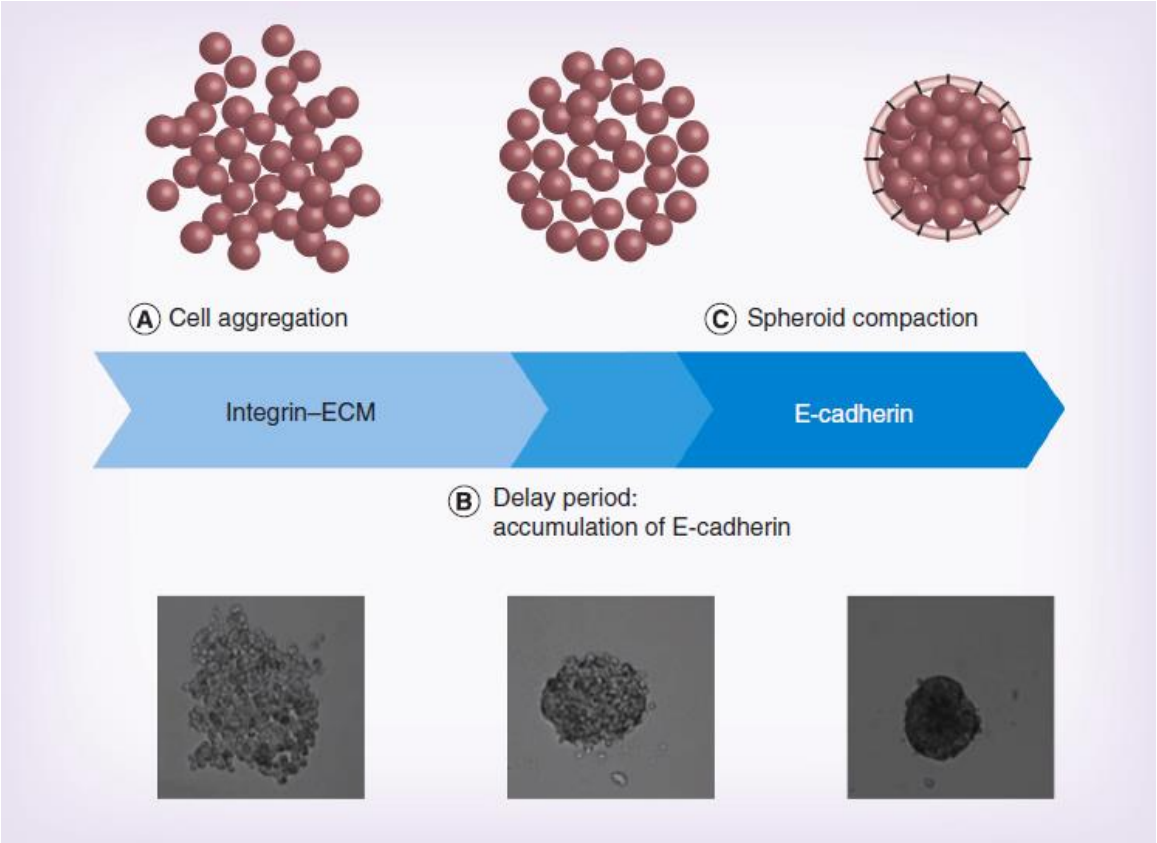
Elipsoid = 10-30 μm thick



Collagen I (2D thin coat) **Collagen I (3D gel)** **Matrigel® matrix (3D gel)**

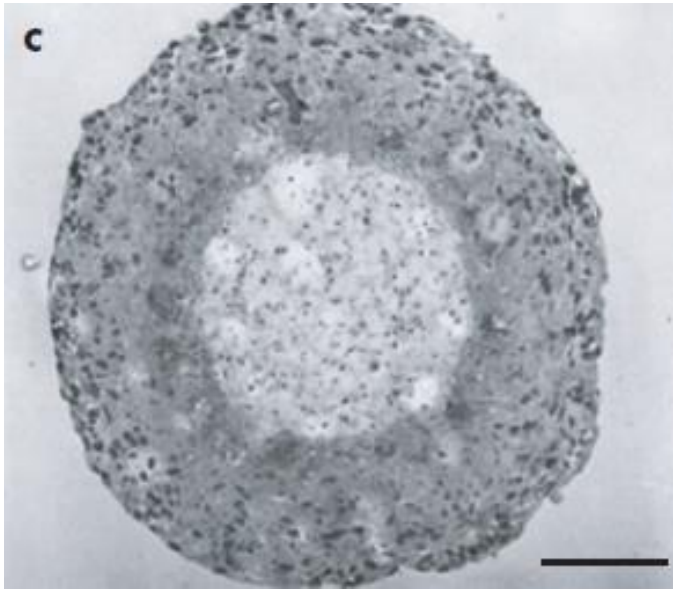
Sanyal, 2014

Cell spheroid - one of the simpliest 3D cultures



Benien and Swami, 2014

Spheroids mimic solid tumours

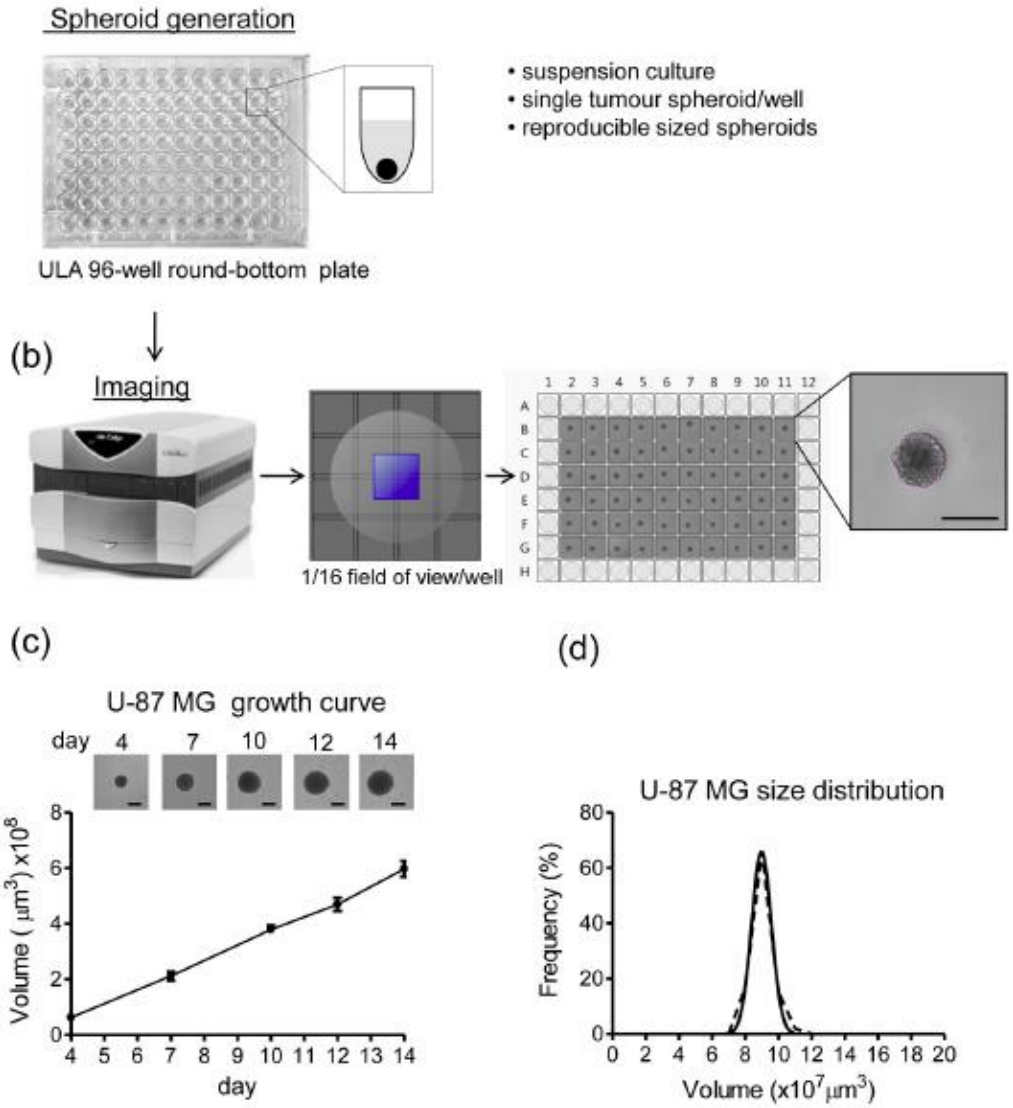


Electron microscopy cross-section of lung cell spheroid.

The inner core of spheroids exhibits a hollow lumen resembling the necrotic areas of *in vivo* cancers that are larger than 500 μm in size.
It forms due to low pH, accumulation of metabolites and hypoxic condtions.

Pampaloni et al., 2007

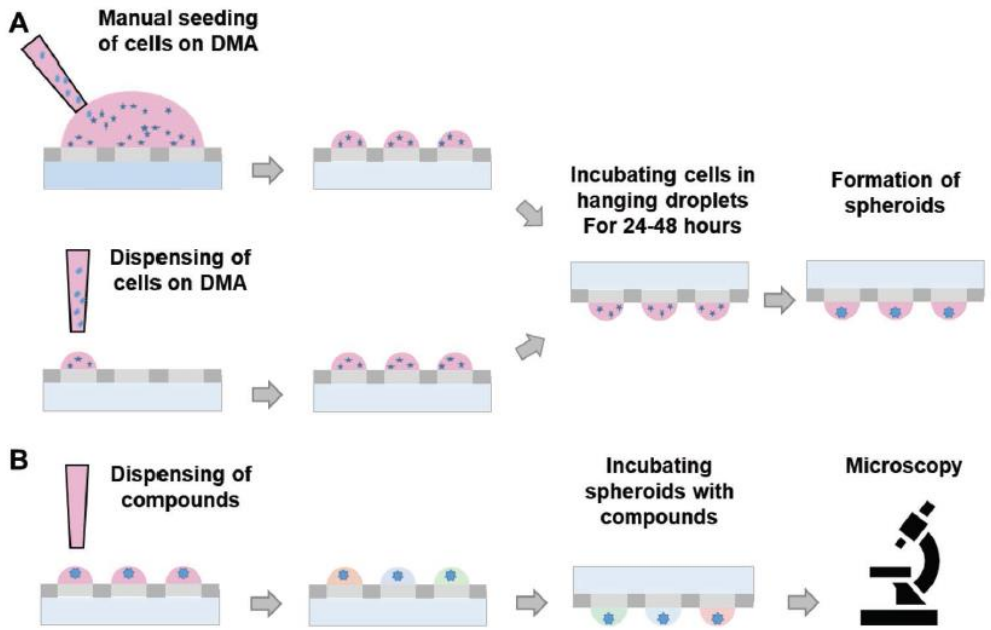
Spheroids in high-throughput

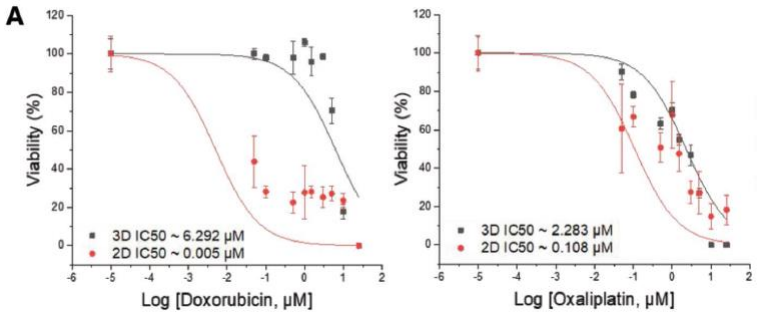


Vinci et al., 2012

Droplet array – spheroid formation

588 spheroids , ~80 nl droplets, 60-80 starting cells/spheroid



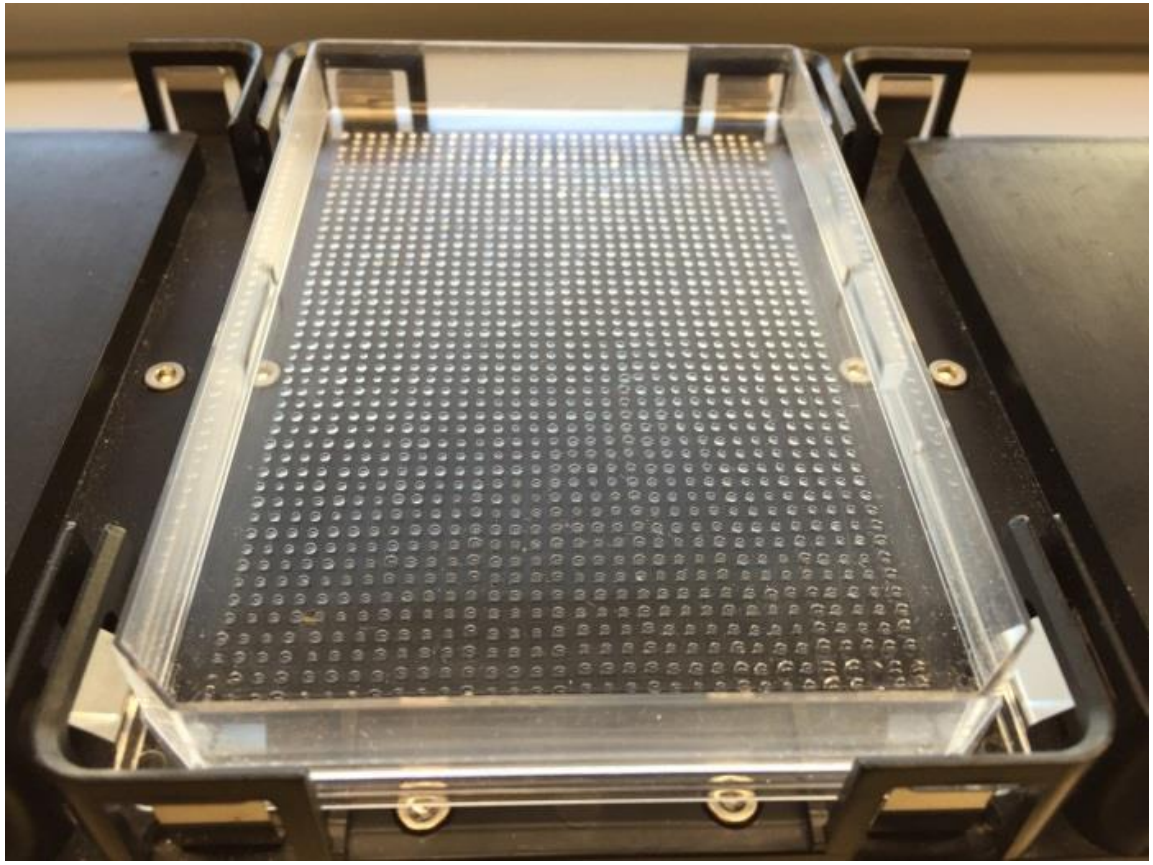


Facile One Step Formation and Screening of Tumor Spheroids Using Droplet-Microarray Platform

Anna A. Popova, Tina Tronser, Konstantin Demir, P. Haitz, Karolina Kuodyte, Vytaute Starkuviene, Piotr Wajda, and Pavel A. Levkin**

Popova et al., 2019

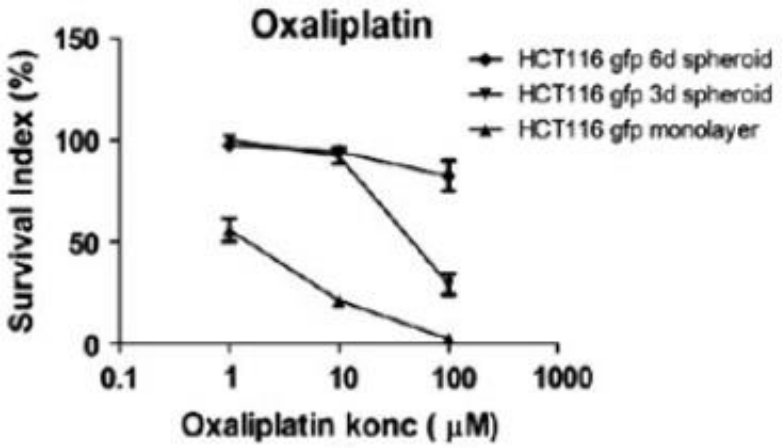
Matrix-spheroids arrays

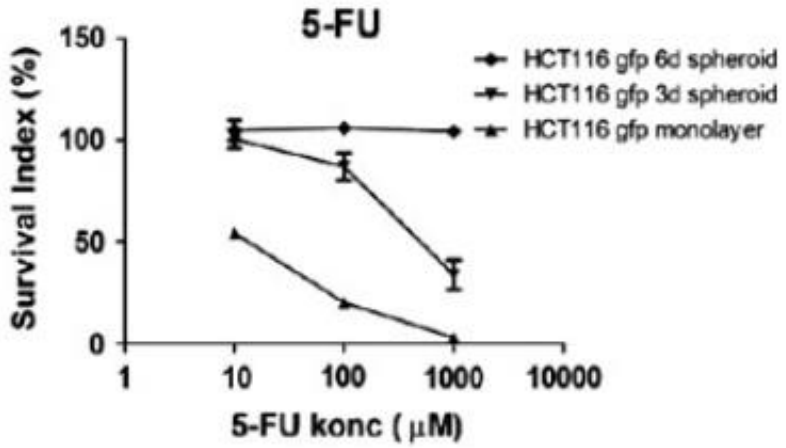


Beneke, personal communication

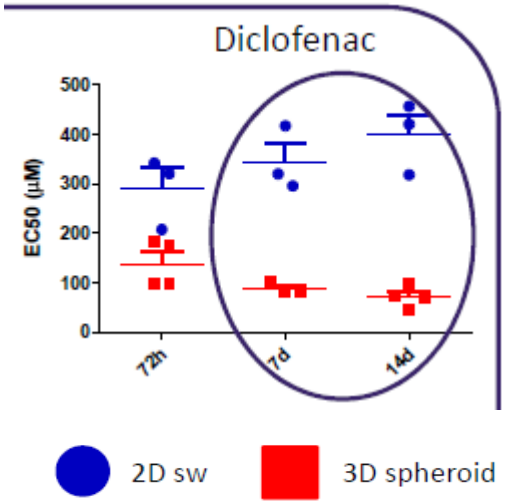
Spheroids as a model for drug activity

Cell in 2D grows in homogenous populations, which usually responds stronger to cytotoxic drugs than heterogenous 3D cultures



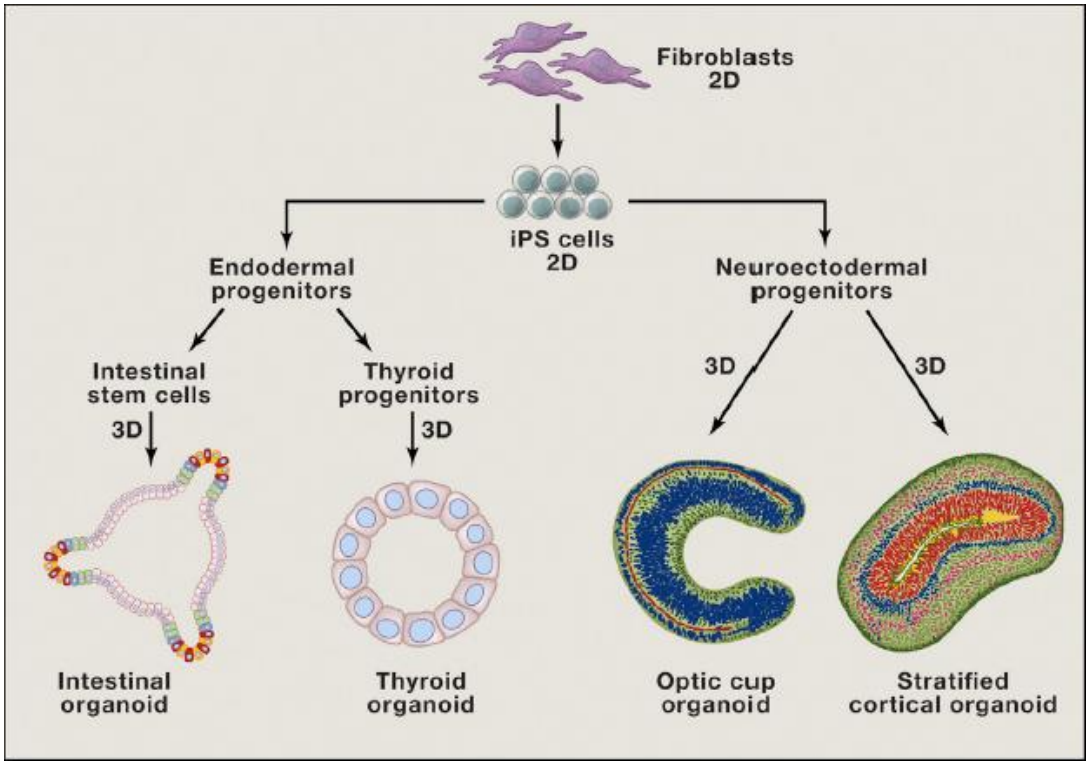


Drug candidates turns to be innefective *in vivo* or clinical tests



Edmonsdon et al., 2014, cortesy of Hewitts, Merc

Organotypic cultures



Origins: organ-specific or embrionic stem cells

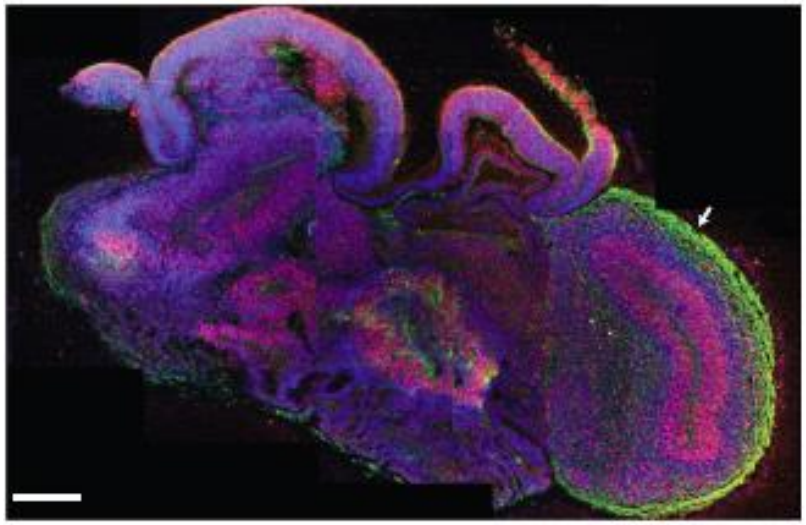
3D aggregates are transfered into appropriate induction media (signaling factors, ECM) that resembles the conditions of development

grown in spinning bioreactors up to 4 mm in size, but are limited by the lack of circulatory systems to sustain further growth

Organoids of lungs, stomach, thyroid, intestinal, eyes, ears, kidneys and brain are cultivated

Bershteyn et al., 2013, Nature Methods, January, 2018

Cerebrial organoids



Similarities to neuroepithilium:

apical-basolateral polarity
formation of ventricum
formation of cortical layers
symetrical or unsimetrical
cell division
survive for up to 9 months

Development and diseases of human brain can not be reflected in mouse models.

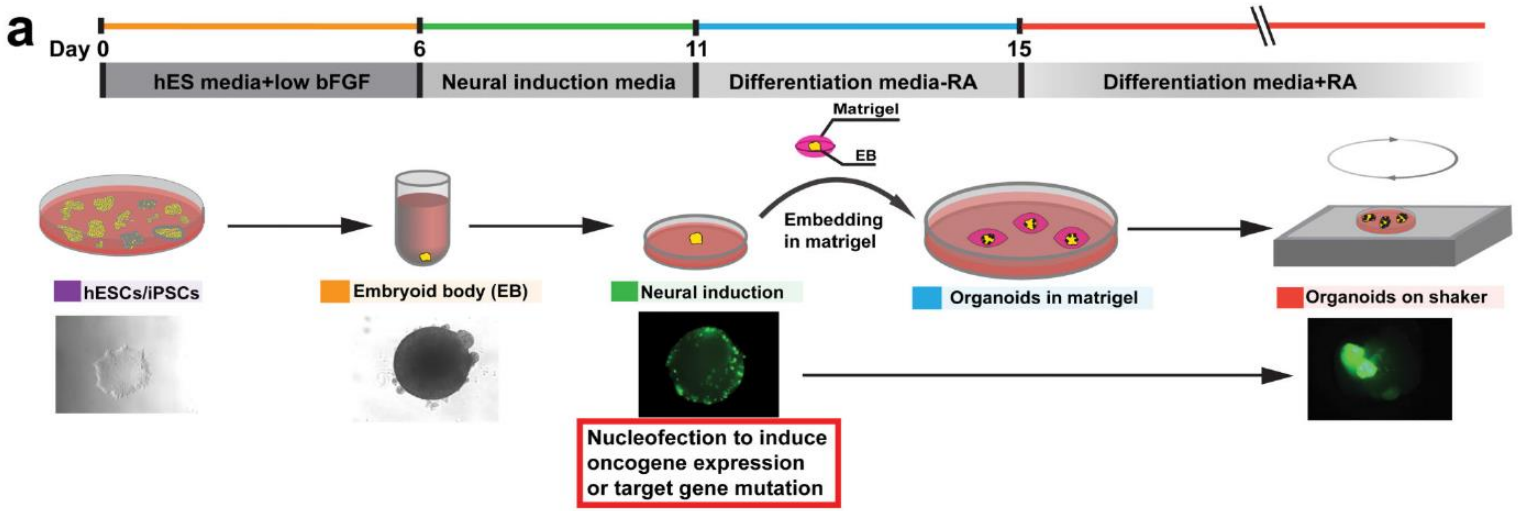
Example case: microcephalie. The mutated gene CDK5RAP2 in mouse does not induce the disease. Organoid cultures from the patient cells are smaller than that of the healthy individuals.

Knoblich et al., 2014

Functional analysis of cerebrial organoids

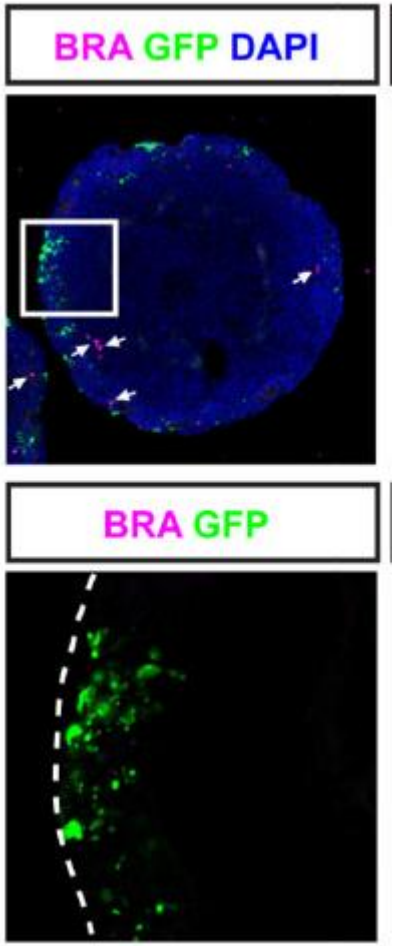
Quadrato et al., Nature, 2017: The researchers analysed the gene-expression profiles (the transcriptomes) of more than 80,000 cells from 3- or 6-month-old organoids — the most comprehensive single-cell analysis of organoid composition performed so far

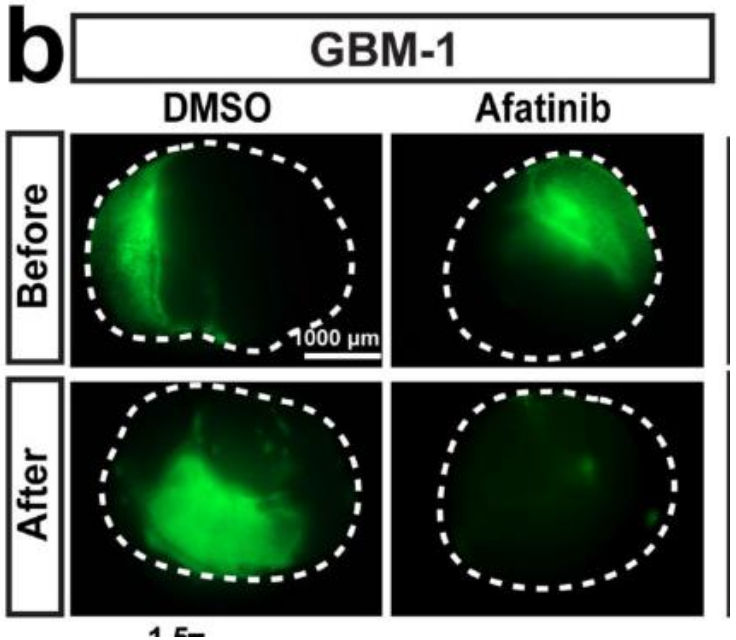
Bian et al., Nature Methods, 2018: Activation of oncogene expression via CRISPR-Cas9 mediated gene editing and re-capitulating human brain tumours



Quadrato et al., 2017, Bian et al., 2018

neoCOR – neoplastic cerebrial organoid

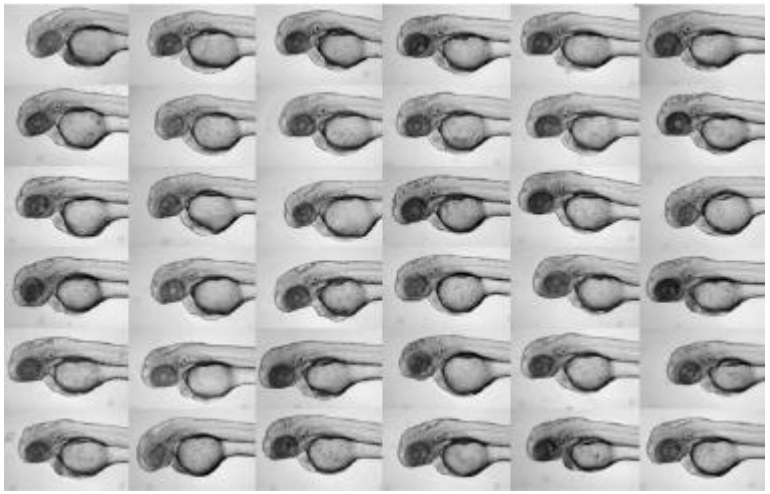


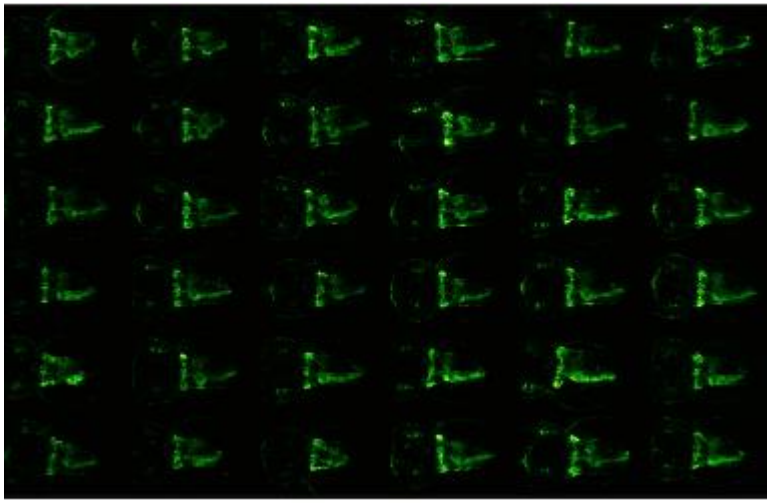


Yoon et al., 2019, Nature Methods:
reliable drug testing platform

Bian et al., 2018

Whole animal screening



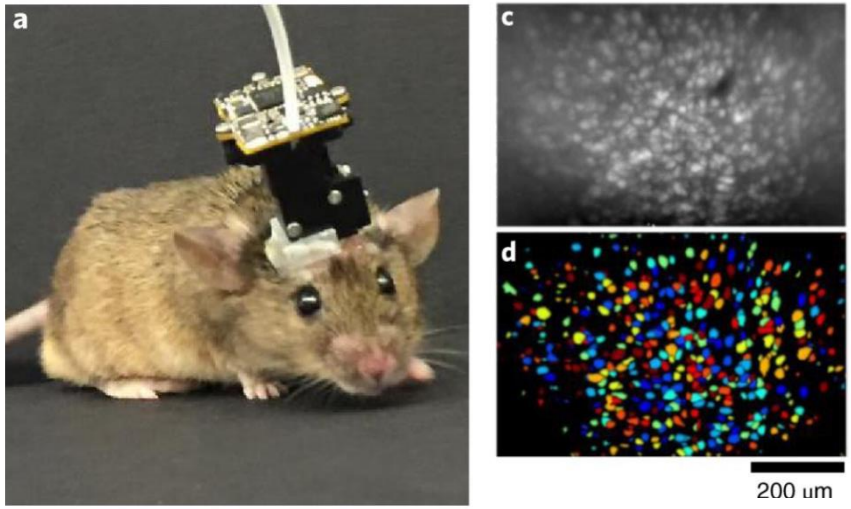


Main application: studies for toxins and environmental polutants

Wittbrod et al., 2014

Nature Method of the year 2018:

Imaging in freely behaving animals



Aharoni et al., 2018

Needs for large-scale data collection

1. Handling complexity of the information and linking it to the particular topic

2. Storage and access

3. Rules for data sharing and protection

4. Standartization of experimental models

5. Ethical rules for gene editing and animal models