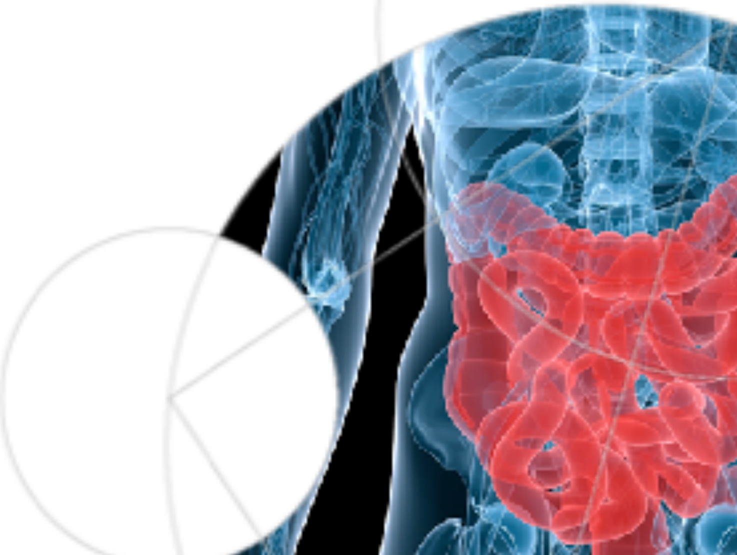
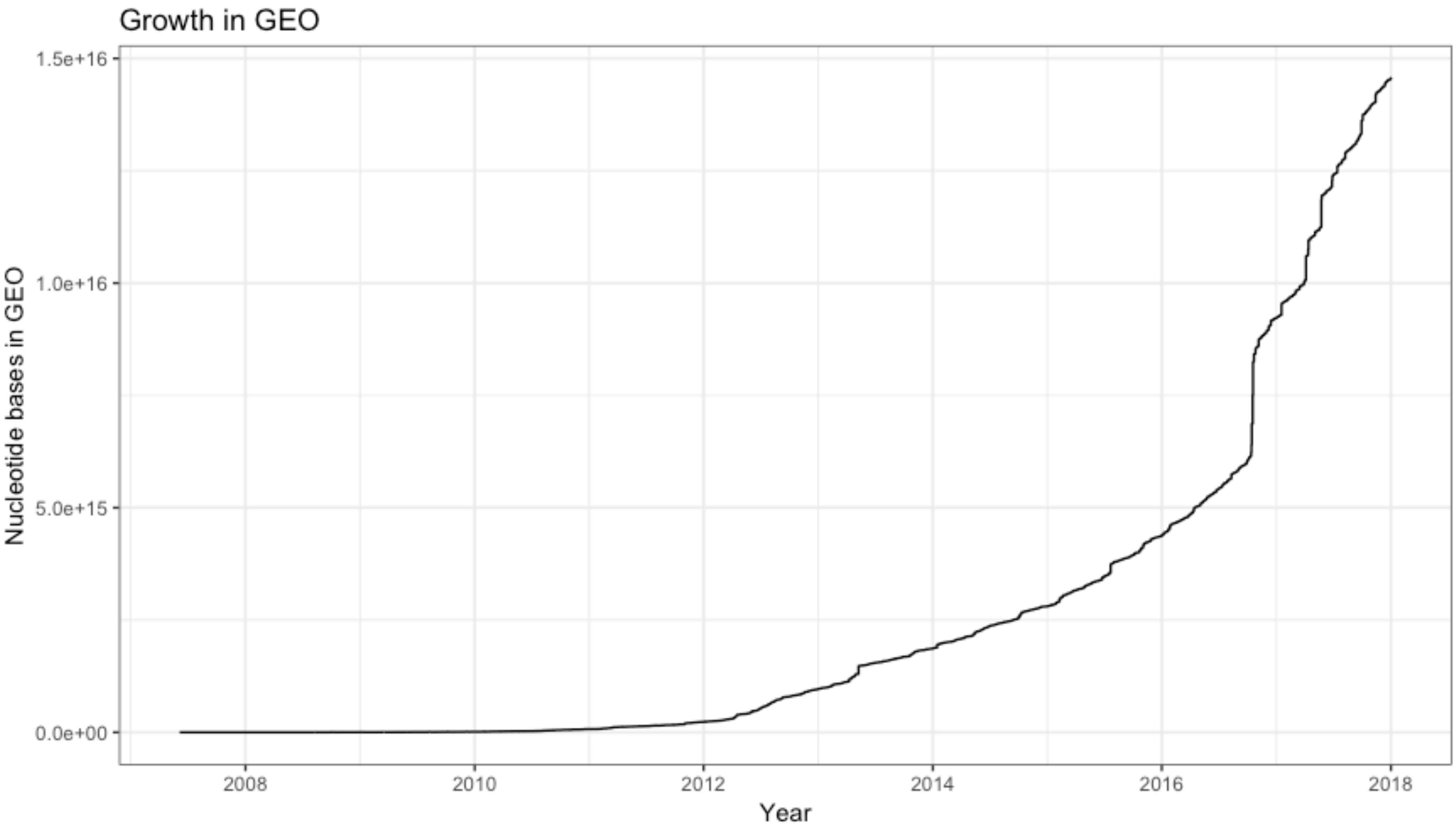
UNIVERSITY OF COPENHAGEN

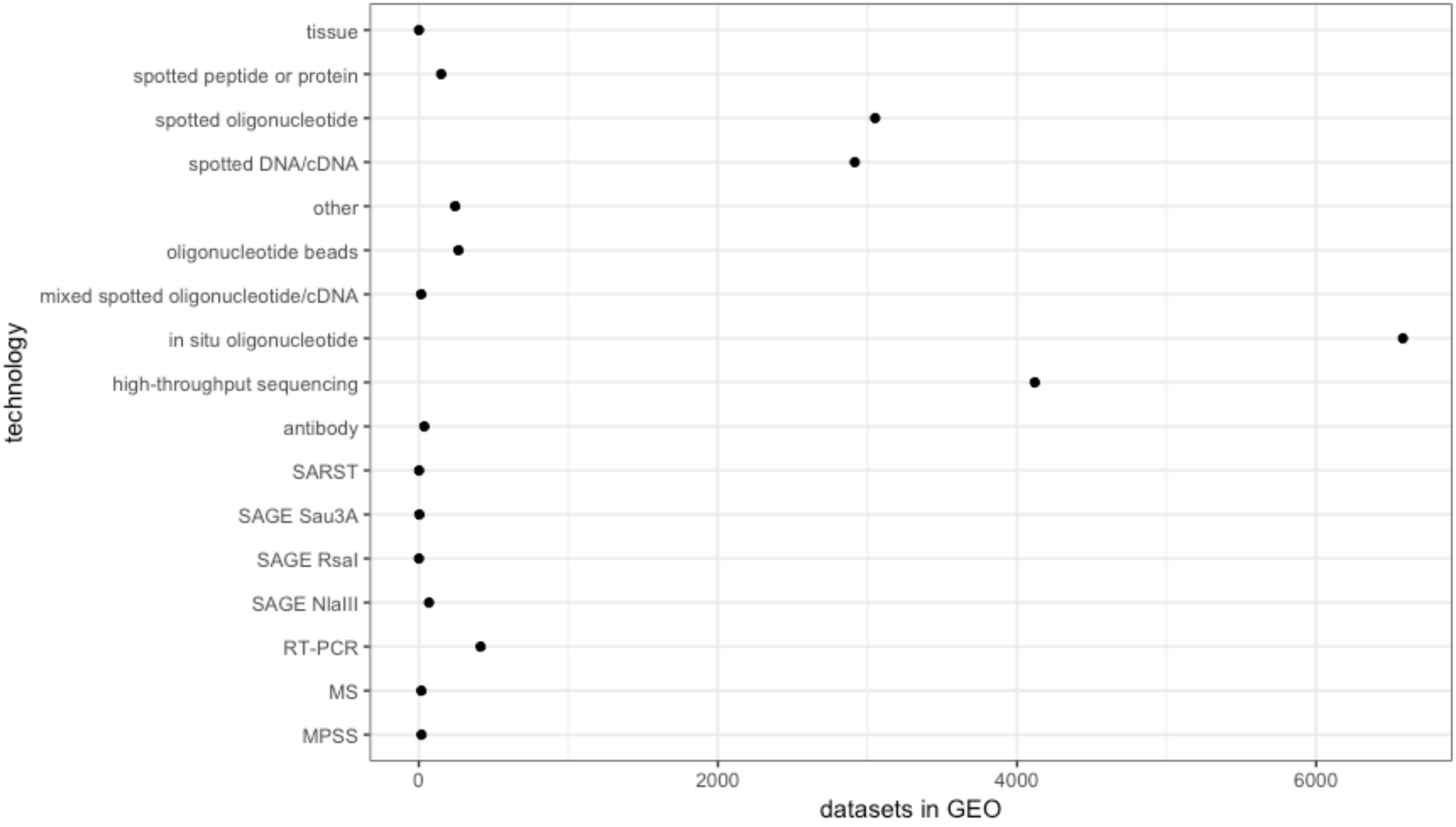
Data Integration in the Era of “Omics": Current and Future Challenges

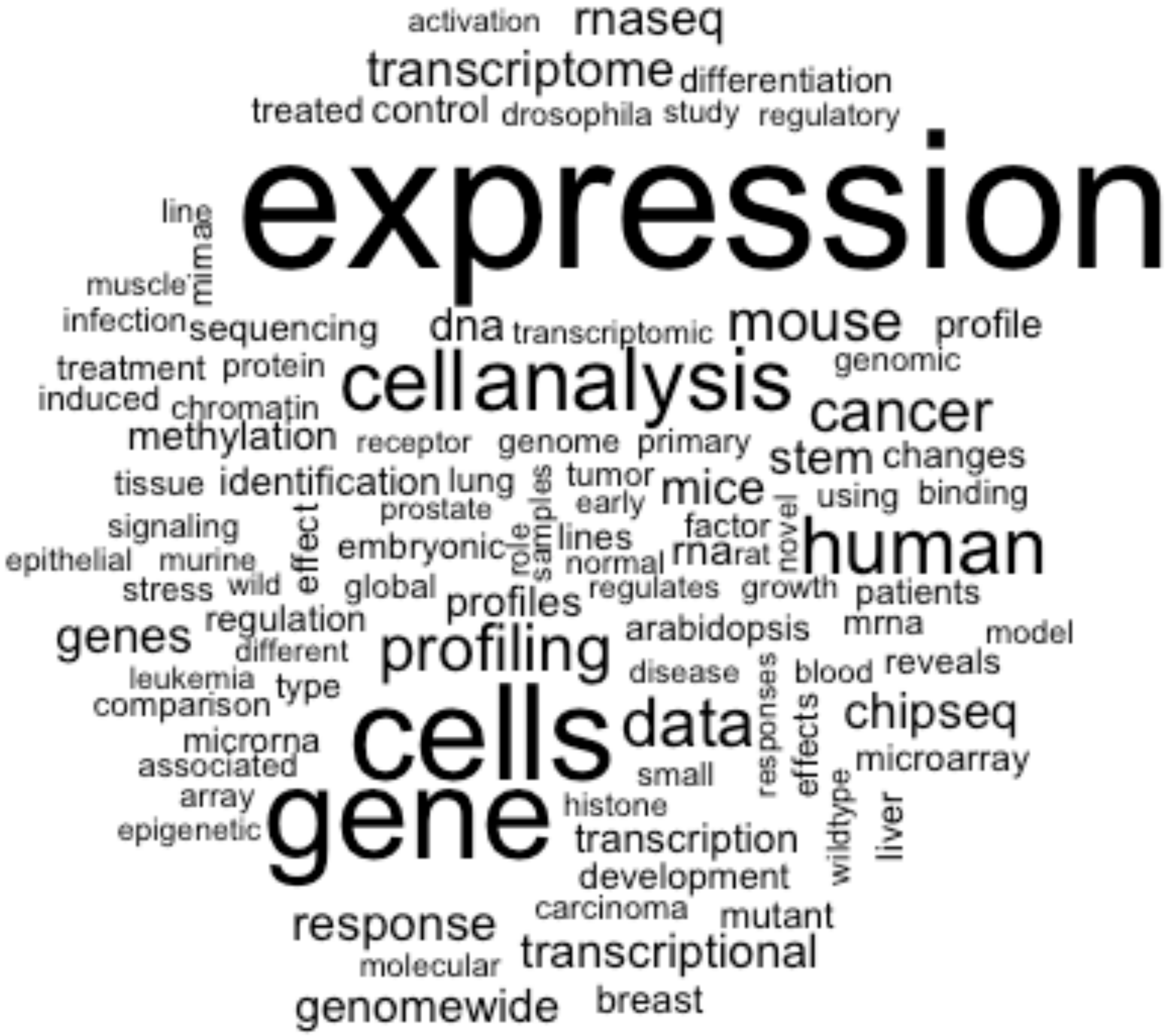
Sandelin Lab
The Bioinformatics Centre, Dept. of Biology
Biotech Research and Innovation Center (BRIC),
University of Copenhagen



The classical “why you should get a degree in bioinformatics” slide




technology
tissue
spotted peptide or protein
spotted oligonucleotide
spotted DNA/cDNA
other
oligonucleotide beads
mixed spotted oligonucleotide/cDNA
in situ oligonucleotide
high-throughput sequencing
antibody
SARST
SAGE Sau3A
SAGE RsaI
SAGE NlaIII
RT-PCR
MS
MPSS
0
2000
4000
6000
datasets in GEO



There is a reason that we give away our data and put it in a database - it is so that other people can use it. Combining data might produce insights that would not be possible with one data set alone.

There are many good reasons to integrate data…

There are large sets of data available, for free.

Re-using data is cheap and efficient: most data is only surface analysed - maybe you can find new things?

Conversely: cheap to produce very rich, new data

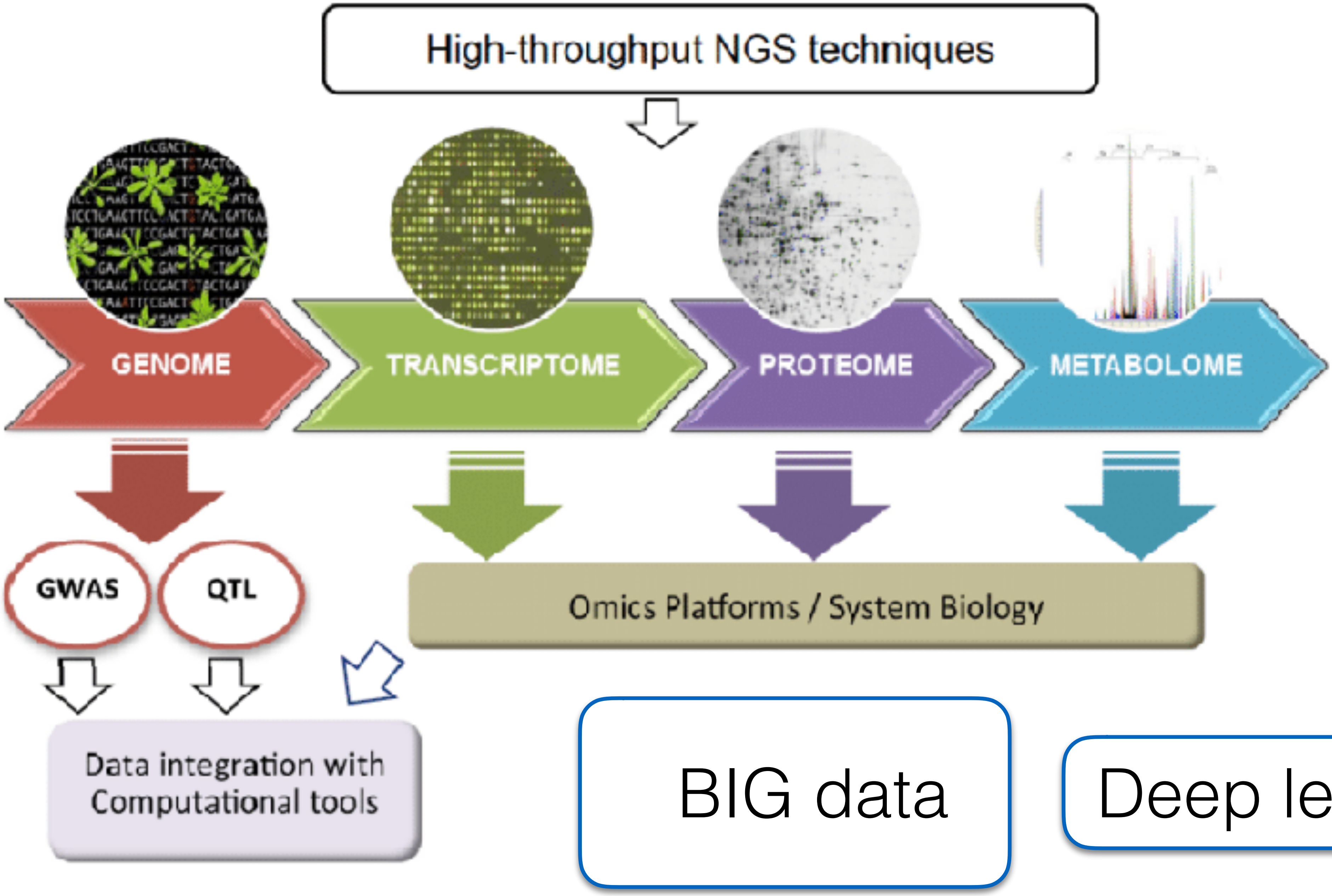
The most interesting data analysis is when you can combine two or more data sets

Same experiment from many individuals
Same experiment from many tissues/cells
Different experiments on same individual
Different experiments from many tissues/cells

3 scenarios:

- Design your own large study
- Take data from other big study (or possibly add to it)
- Take data from multiple studies (single or large studies)

But how to do this? The typical image around this is to make impressive arrows to a box called "Data integration" or "Systems Biology"


High-throughput NGS techniques
GENOME
TRANSCRIPTOME
PROTEOME
METABOLOME
GWAS
QTL
Omics Platforms / System Biology
Data integration with Computational tools

BIG data

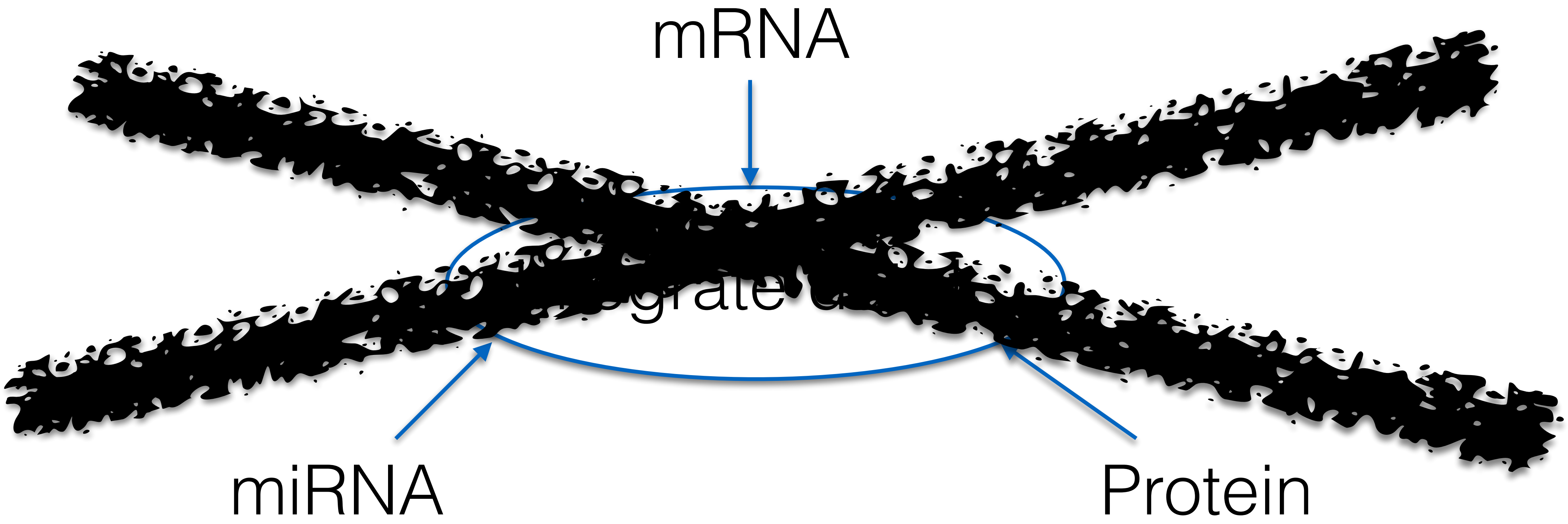
Deep learning

Lesson 1: Start with the question(s), not the data

What is the goal of your data integration? What type of questions do you want to answer?

Then: what data do you need, and how do you imagine these things should be analysed?

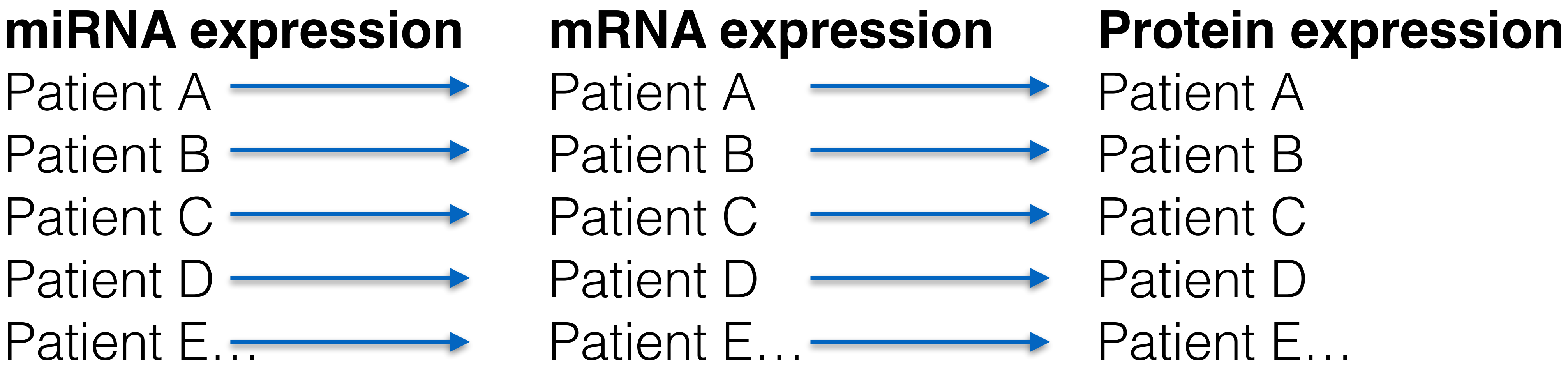
Example: I want to find out miRNA regulation of gene expression and protein expression, across individuals within same cell type



Analysis (example):

- Find genes with sites for each miRNA
- For each such gene, correlate mRNA expression, protein expression with of miRNA expression across samples.
- Hope to find convincing negative correlations

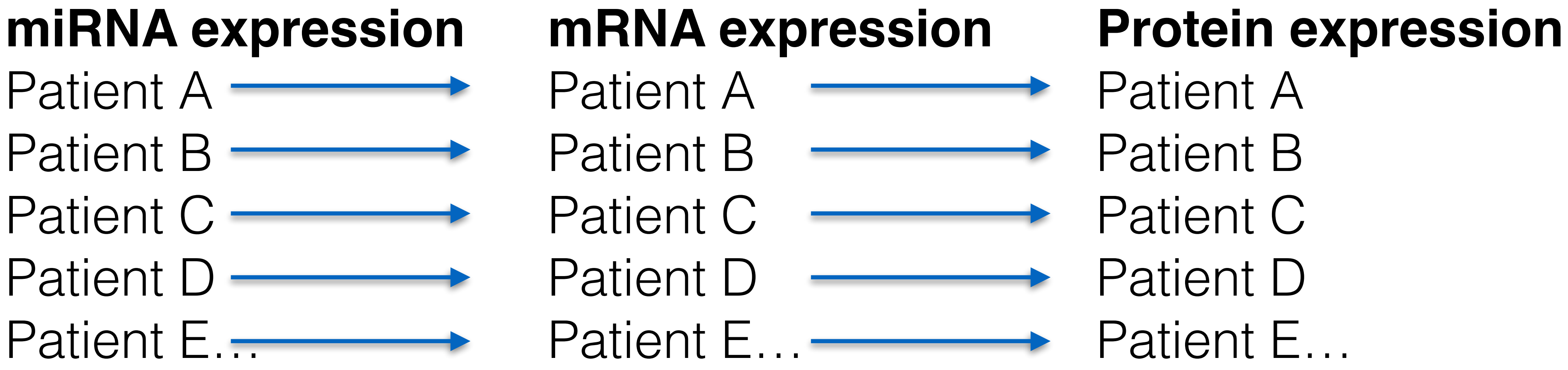
Data needed



miRNA expression	mRNA expression	Protein expression
Patient A	Patient A	Patient A
Patient B	Patient B	Patient B
Patient C	Patient C	Patient C
Patient D	Patient D	Patient D
Patient E…	Patient E…	Patient E…

…from same samples, and same gene models

Lesson 2.1 If you don’t have comparable data, don’t integrate!



miRNA expression		mRNA expression		Protein expression
Patient A	→	Patient A	→	Patient A
Patient B	→	Patient B	→	Patient B
Patient C	→	Patient C	→	Patient C
Patient D	→	Patient D	→	Patient D
Patient E…	→	Patient E…	→	Patient E…

miRNA expression	mRNA expression	Protein expression
Patient A	Patient F	Patient H
Patient B	Patient G	Patient I
Patient C	Patient H	Patient J
Patient D	Patient I	Patient K
Patient E…	Patient J…	Patient L…

If analysis require linked samples, anything less than linked samples wont make it

ENCODE pilot example

- The ENCODE pilot aimed to test out many different technologies on just 1% on the genome, defined to capture both known and unknown loci (selected and random)

Problems:

- Each laboratory used their own technique, often their own (different) cell lines. And often published their results in advance of the integrative study

- Led to large problems for the analysis persons, and for selling the whole project as a joint and merged data set

- Experiment type A shows B in a certain region in cell C
- Experiment type D shows E in same region in cell G

Indecipherable

- It meant the integration just became cumulative: so and so many bases have evidence for experiments of this and this type, and there was very limited narrative. Made it very hard to publish - was envisioned as 5 Nature papers but ended as 1.

Lesson 2.2 . Beware of batch effects and confounders - leads to non-comparability

All data contains signal and noise.

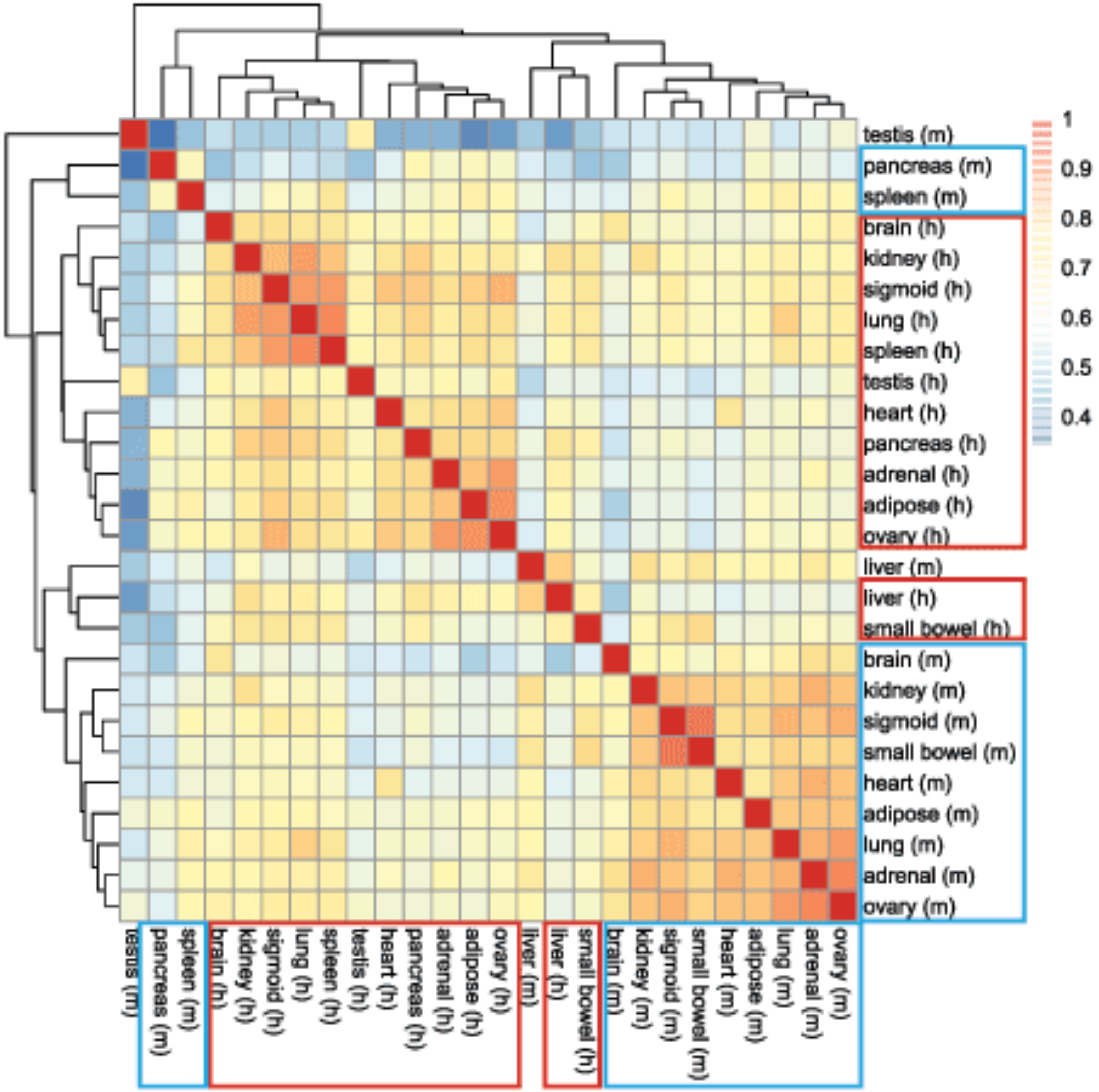
Noise can be random or systematic. Systematic noise is hugely effected by experimental design: especially batch effects and confounding effects

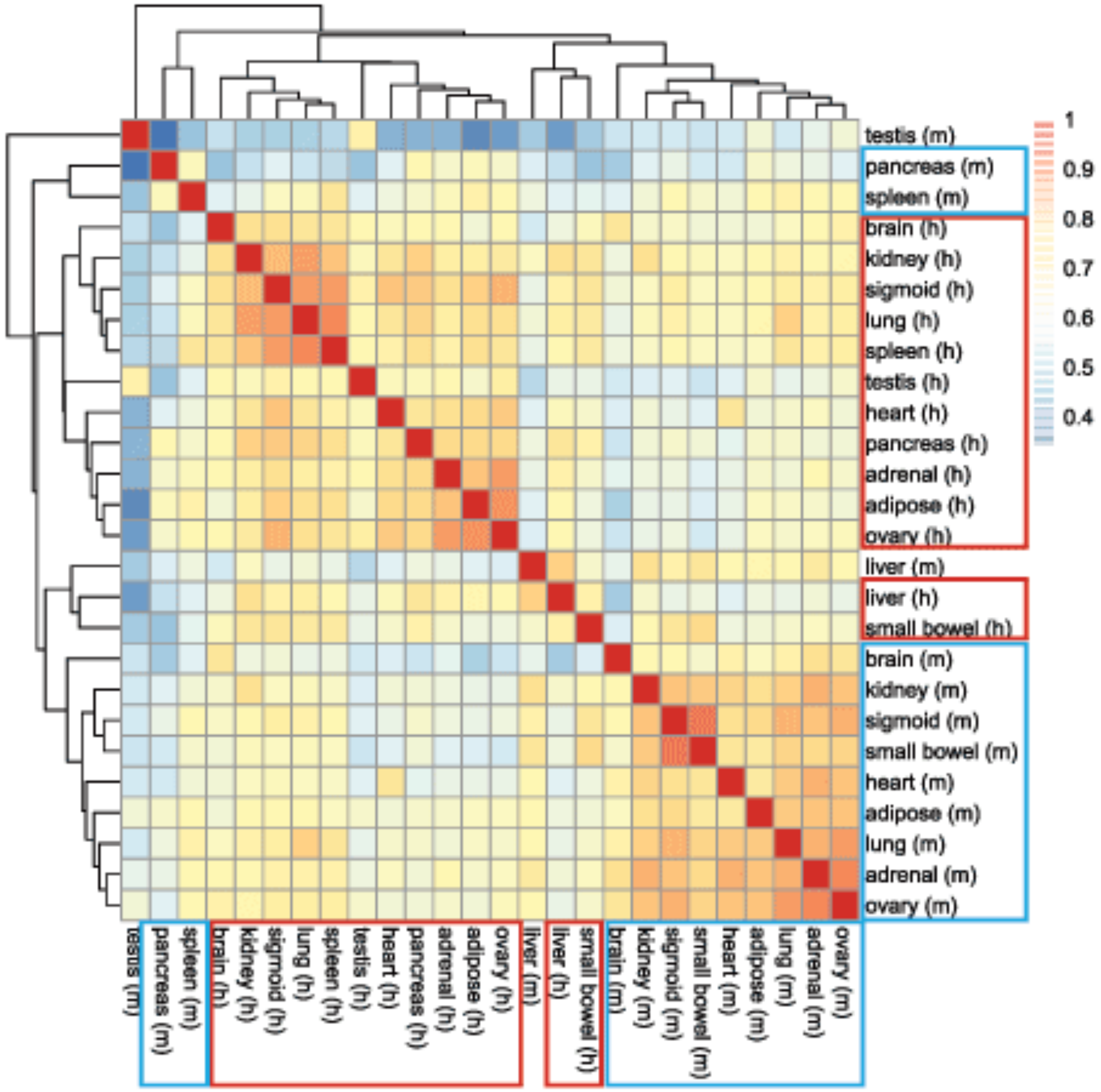
When dealing with large experiments, across multiple platforms, good experimental design becomes even more important - and sanity check afterwards.

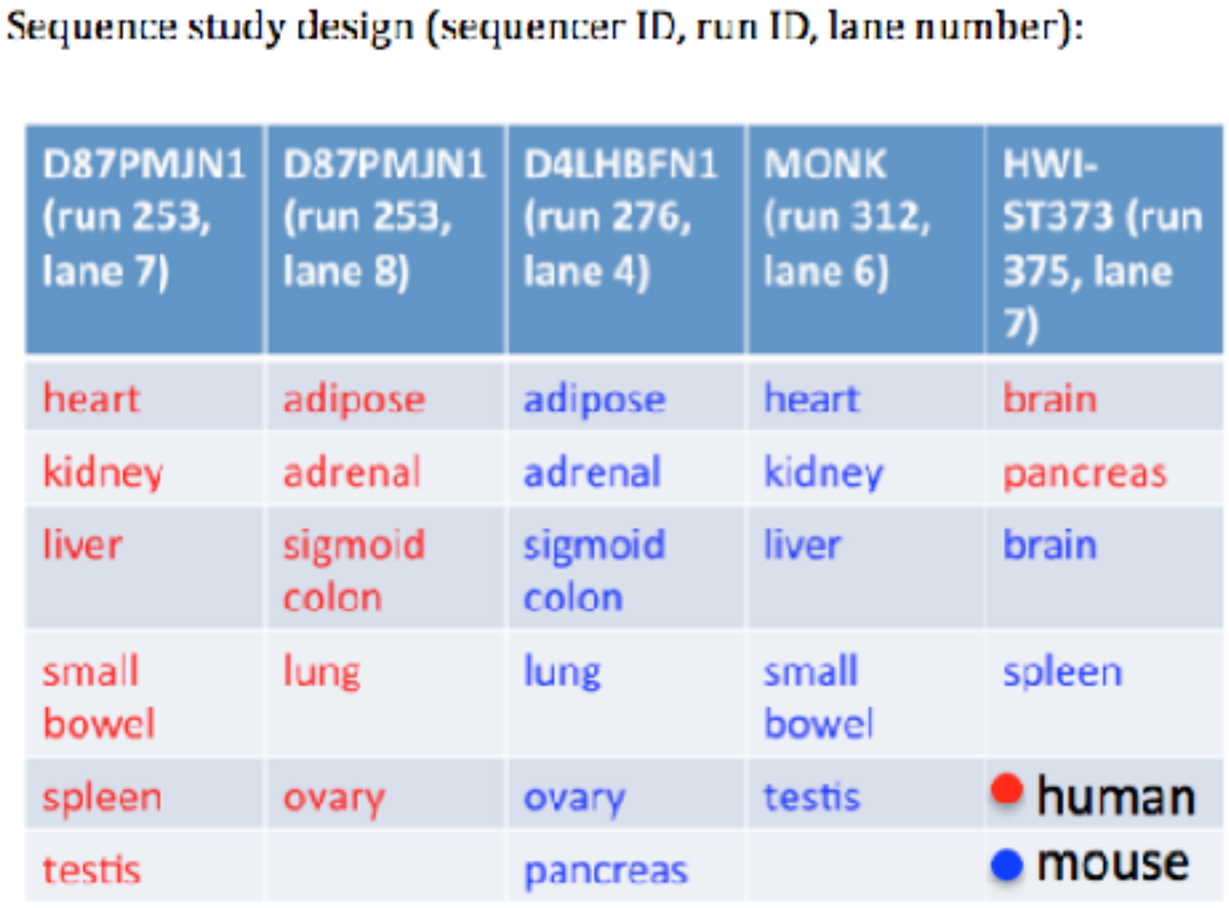
This is important as you may start reporting things that you are actually not interested in…

Analysis of gene expression across tissues , human vs mice: Do tissue or species pair up?

Original study (Lin *et al.*: *PNAS.* 2014) found that all almost human tissue clustered together, so species is more important than tissue. This was very surprising as it went against many previous studies





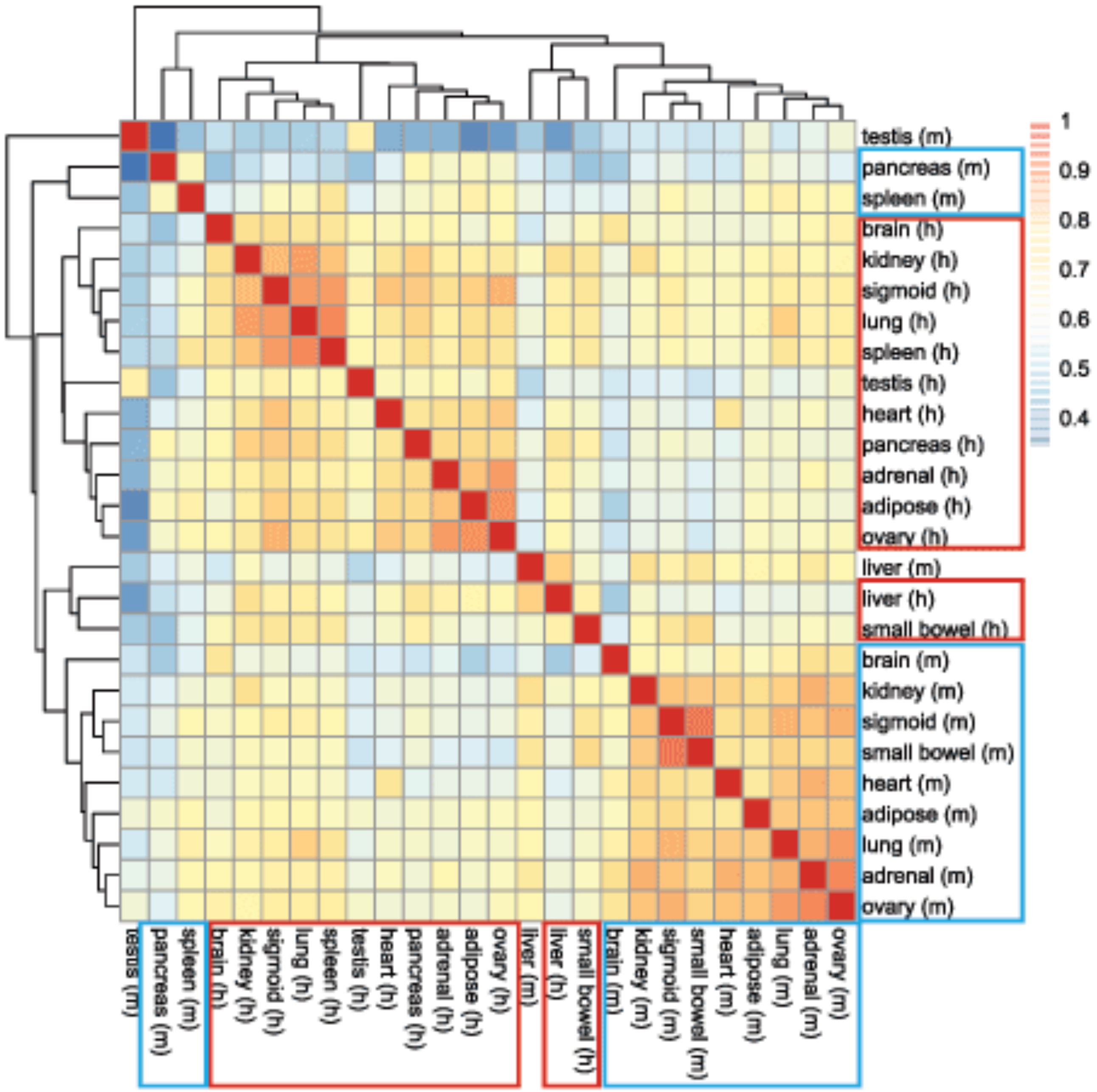


Sequence study design (sequencer ID, run ID, lane number):

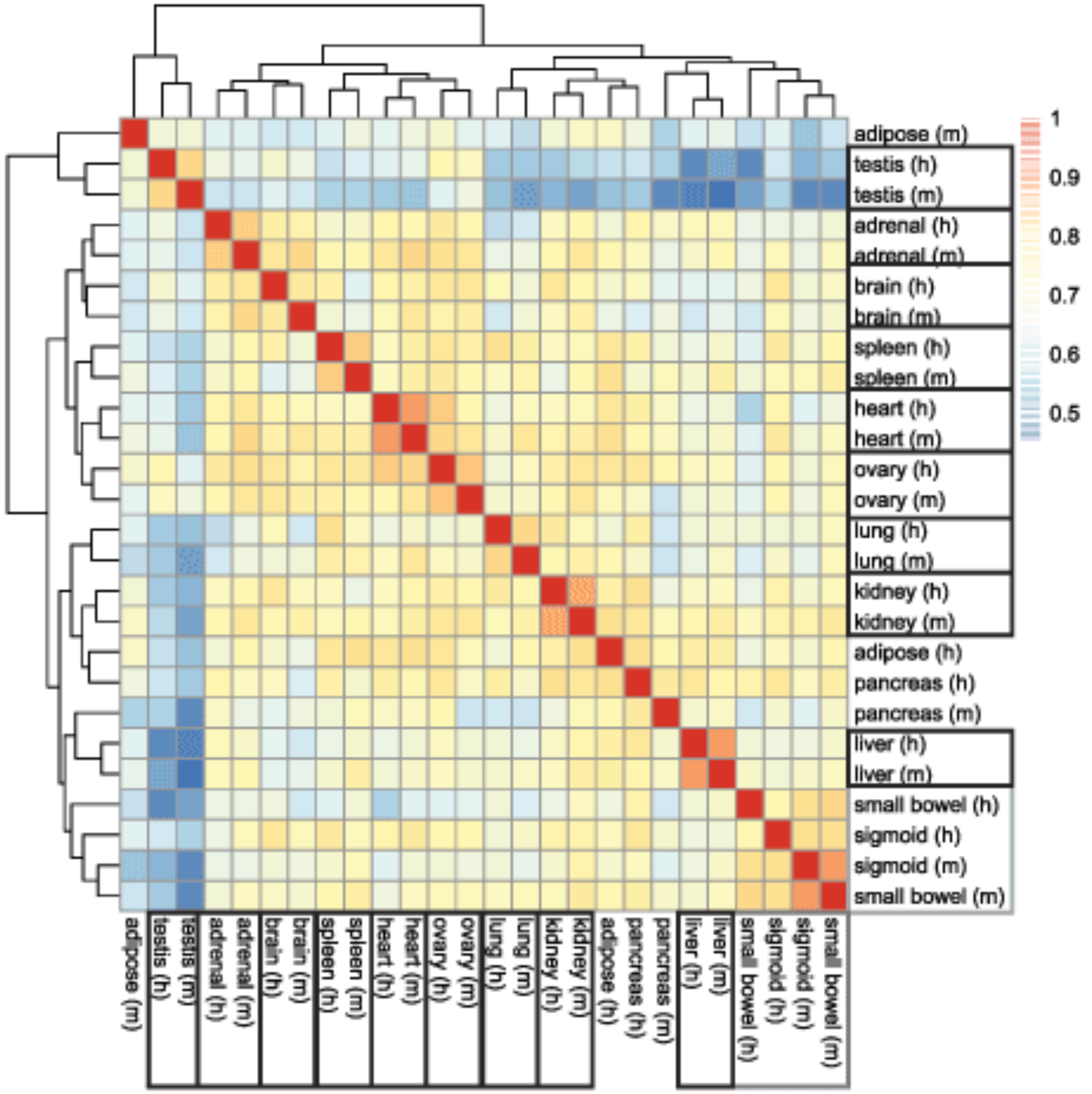
D87PMJN1 (run 253, lane 7)	D87PMJN1 (run 253, lane 8)	D4LHBFN1 (run 276, lane 4)	MONK (run 312, lane 6)	HWI-ST373 (run 375, lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● human
testis		pancreas		● mouse

Could this be due to bath effects? Gilad, 1000Research 2015, 4:121 tried to remove this effect using the COMBAT algorithm

Non-batch-corrected



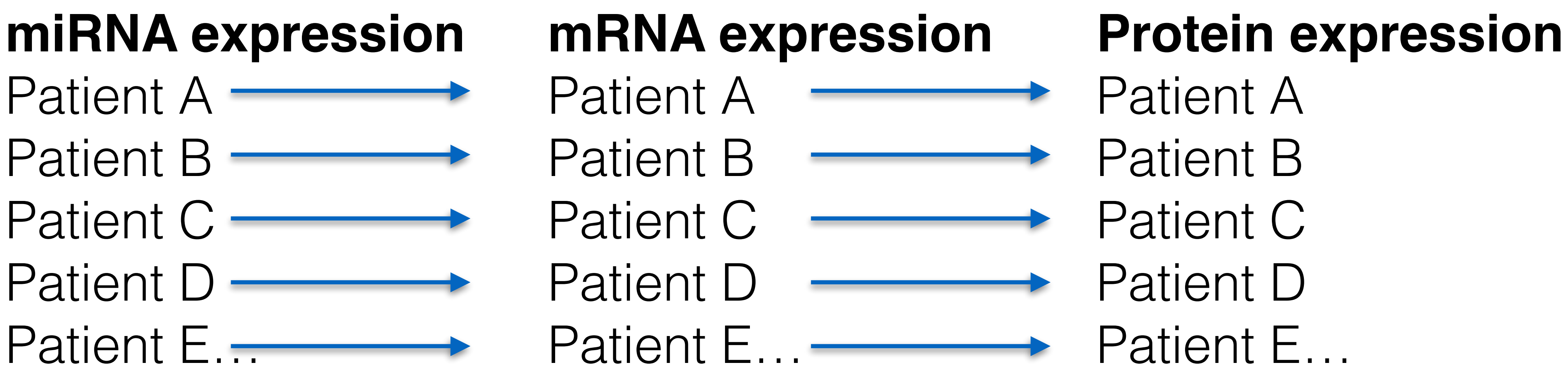
Batch-corrected



Which one is correct?

Actually, we can never know.

Lesson 3: Common notation, reference points and agreed data simplifications is critical



miRNA expression	mRNA expression	Protein expression
Patient A	Patient A	Patient A
Patient B	Patient B	Patient B
Patient C	Patient C	Patient C
Patient D	Patient D	Patient D
Patient E…	Patient E…	Patient E…

- How can we really compare miRNAs of Patient A to mRNA expression in Patient A? It requires an agreed standard on genes and how they are measured and defined

- Is splicing important? Should I consider different gene isoforms? Or do I simply want one value per "gene"?

- These definitions are in effect agreed simplifications or conceptualisations of the data, and will shape all analysis downstream

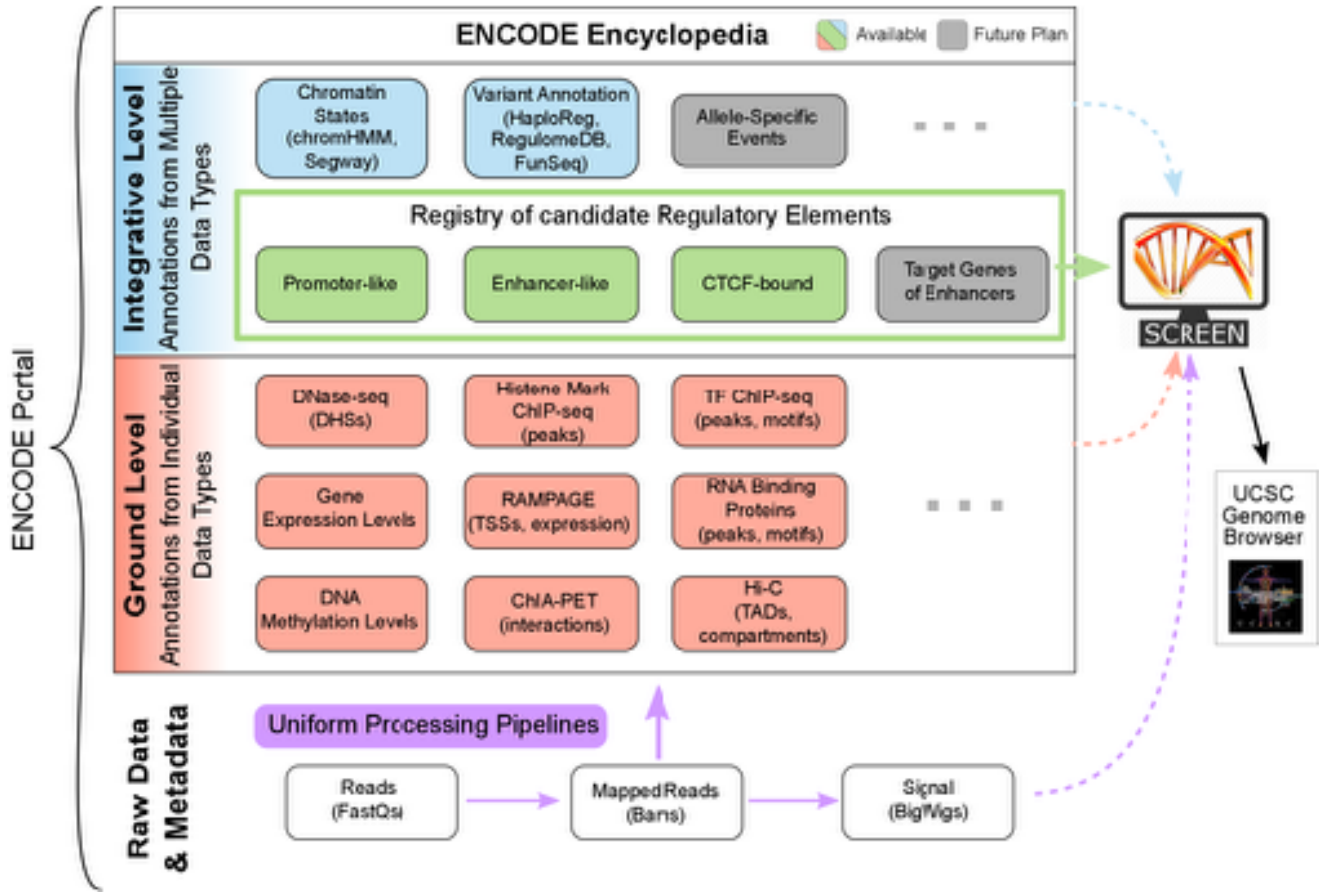
Three examples of large scale data consortial and their data integration efforts - not their analysis but rather their overall approach

ENCODE
FANTOM
Cancer Genome Atlas

Example 1: ENCODE phase 2

"Encyclopaedia" of genome elements (mostly cell lines) using a wide variety of techniques - mostly focused on gene regulation and gene expression, sometimes combined with knockdowns

Has a very clear strategy for data integration on multiple levels - starting from raw data, annotation from different data types and then integrated levels where relevant data sets are combined (combine many chromatin experiments into specific states, for instance)


ENCODE Encyclopedia
Available
Future Plan
ENCODE Portal
Integrative Level
Annotations from Multiple Data Types
Chromatin States (chromHMM, Segway)
Variant Annotation (HaploReg, RegulomeDB, FunSeq)
Allele-Specific Events
...
Registry of candidate Regulatory Elements
Promoter-like
Enhancer-like
CTCF-bound
Target Genes of Enhancers
SCREEN
UCSC Genome Browser
Ground Level
Annotations from Individual Data Types
DNase-seq (DHSs)
Histone Mark ChIP-seq (peaks)
TF ChIP-seq (peaks, motifs)
Gene Expression Levels
RAMPAGE (TSSs, expression)
RNA Binding Proteins (peaks, motifs)
...
DNA Methylation Levels
ChIA-PET (interactions)
Hi-C (TADs, compartments)
Raw Data & Metadata
Uniform Processing Pipelines
Reads (FastQs)
Mapped Reads (Bams)
Signal (BigWigs)

Common reference points

1) For most experiments, the genome is the reference - peaks or blocks that then can be combined between experiments

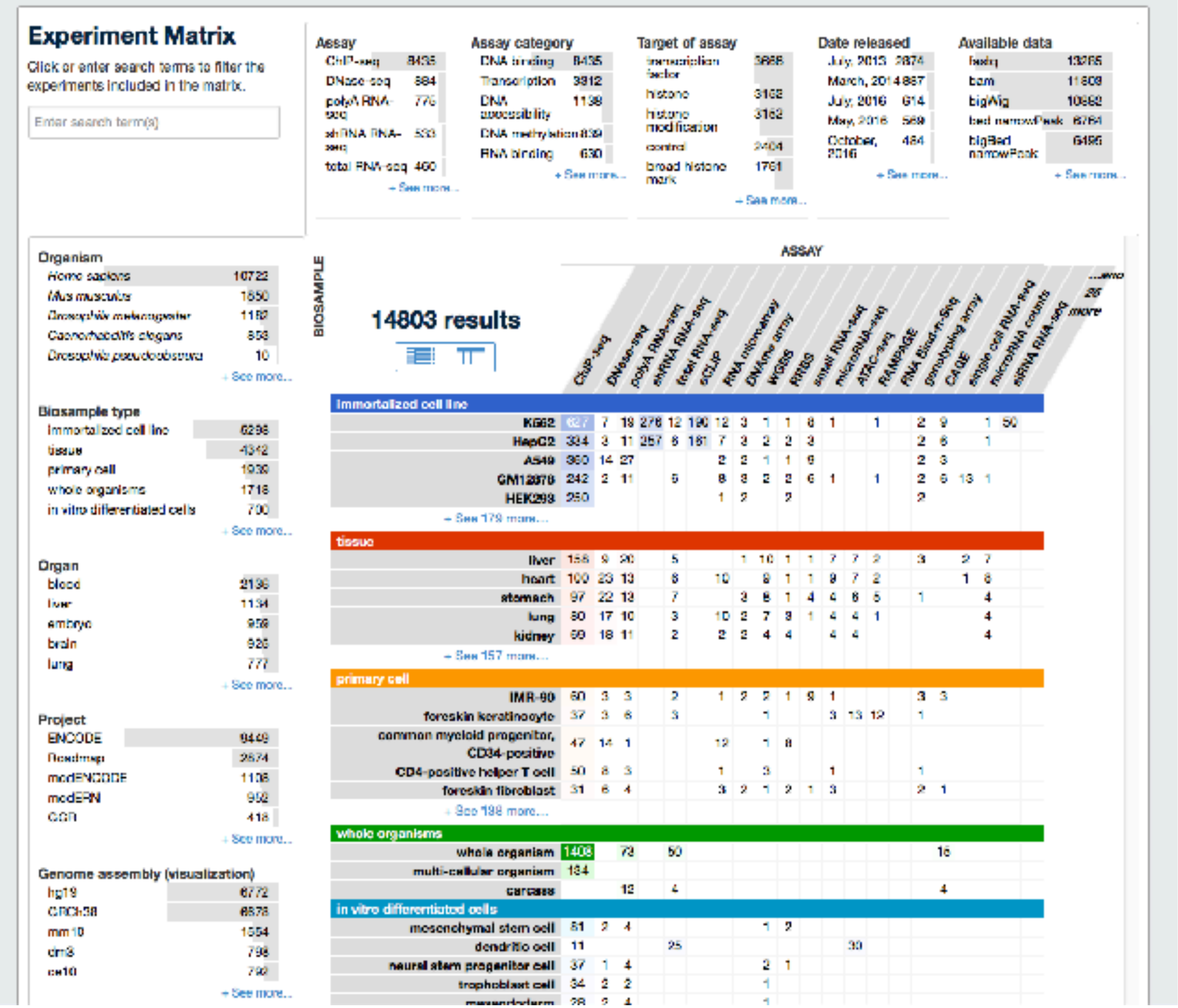
2) GENCODE transcript models

Also has fantastic download section, and all data is “free before publishing”

From an data integration setting, ENCODE is worth studying since it is one of the largest project which started out very chaotic but then got to something that is perhaps the most mature

The ENCODE project also led to a lot of standards for eg ChIP data was established - extremely valuable

Possible downsides: very descriptive with somewhat unclear goals (narrative?), many labs contributing data (good or bad?)



Experiment Matrix

Click or enter search terms to filter the experiments included in the matrix.

Enter search term(s)

Assay	
ChIP-seq	8435
DNase-seq	884
polyA RNA-seq	776
shRNA RNA-seq	533
total RNA-seq	460

+ See more...

Assay category	
DNA binding	8435
Transcription	3812
DNA accessibility	1138
DNA methylation	839
RNA binding	630

+ See more...

Target of assay	
transcription factor	3868
histone	3162
histone modification	3162
control	2404
broad histone mark	1761

+ See more...

Date released	
July, 2013	2874
March, 2014	887
July, 2016	614
May, 2016	569
October, 2016	484

+ See more...

Available data	
fastq	13265
bam	11808
bigWig	10882
bed narrowPeak	8764
bigBed narrowPeak	6495

+ See more...

Organism	
Homo sapiens	10722
Mus musculus	1650
Drosophila melanogaster	1182
Caenorhabditis elegans	858
Drosophila pseudoobscura	10

+ See more...

Biosample type	
immortalized cell line	6296
tissue	4342
primary cell	1939
whole organisms	1718
in vitro differentiated cells	700

+ See more...

Organ	
blood	2136
liver	1134
embryo	958
brain	926
lung	777

+ See more...

Project	
ENCODE	9449
Roadmap	2674
modENCODE	1106
modERN	952
GGR	418

+ See more...

Genome assembly (visualization)	
hg19	6772
GRCh38	6578
mm10	1554
dm3	798
ce10	792

+ See more...

14803 results

ASSAY

BIOSAMPLE

	ChIP-seq	DNase-seq	polyA RNA-seq	shRNA RNA-seq	total RNA-seq	eCLIP	RNA microarray	DNAme array	WGBS	RRBS	small RNA-seq	microRNA-seq	ATAC-seq	RAMPAGE	RNA Bind-n-Seq	genotyping array	CAGE	single cell RNA-seq	microRNA counts	siRNA RNA-seq	...and 28 more
Immortalized cell line																					
K562	627	7	18	276	12	190	12	3	1	1	8	1		1		2	9		1	50	
HepG2	324	3	11	257	8	161	7	3	2	2	3					2	6		1		
A549	360	14	27				2	2	1	1	9					2	3				
GM12878	242	2	11		5		8	3	2	2	6	1		1		2	6	13	1		
HEK293	250						1	2		2						2					
+ See 179 more...																					
tissue																					
liver	156	9	20		5			1	10	1	1	7	7	2		3		2	7		
heart	100	23	13		8		10		8	1	1	9	7	2				1	6		
stomach	97	22	13		7			3	8	1	4	4	8	5		1			4		
lung	80	17	10		3		10	2	7	3	1	4	4	1					4		
kidney	69	18	11		2		2	2	4	4		4	4						4		
+ See 157 more...																					
primary cell																					
IMR-90	60	3	3		2		1	2	2	1	9	1				3	3				
foreskin keratinocyte	37	3	6		3				1			3	13	12		1					
common myeloid progenitor, CD34-positive	47	14	1				12		1	8											
CD4-positive helper T cell	50	8	3				1		3			1				1					
foreskin fibroblast	31	6	4				3	2	1	2	1	3				2	1				
+ See 198 more...																					
whole organisms																					
whole organism	1408		73		50												16				
multi-cellular organism	184																				
carcass			12		4												4				
in vitro differentiated cells																					
mesenchymal stem cell	81	2	4						1	2											
dendritic cell	11				25								30								
neural stem progenitor cell	37	1	4						2	1											
trophoblast cell	34	2	2						1												
mesendoderm	28	2	4						1												

Example 2: FANTOM5 - atlas of transcription start sites

Question/Goal: Use CAGE-seq to profile transcription start site locations and usage across nearly all human cell types

Difference vs ENCODE: Only one technique, but a lot more samples. All data made at the same place, with many sample contributors, and many analysis spread over the world

Advantage: Much higher data control. On the other hand all data is locked down until a large paper is published

Common reference points

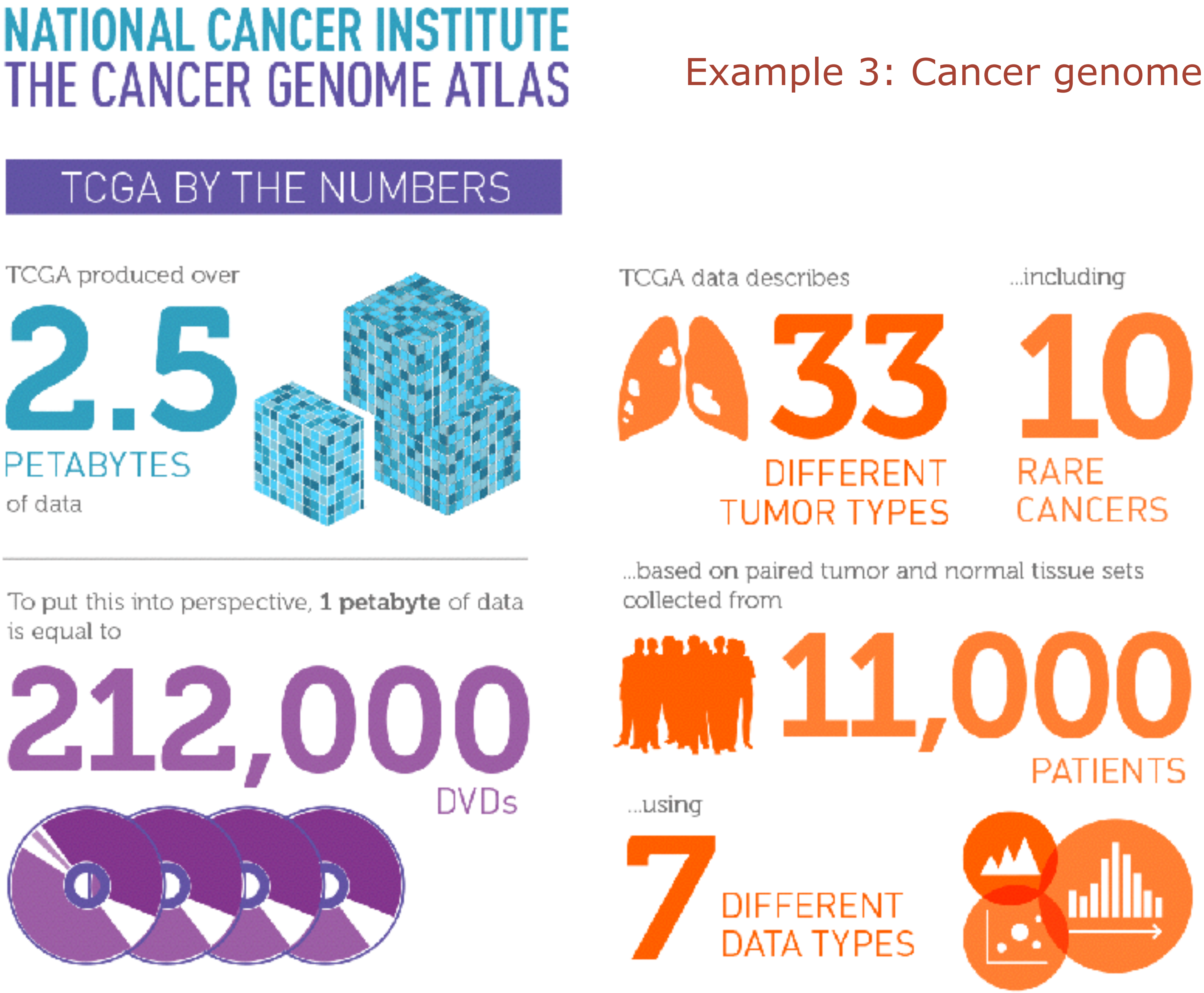
Promoters, or really clusters of CAGE tags across all experiments - a very large effort went into making these

A genome browser with all linked data

A wiki for collaborators with all raw and processed data

A data warehouse inc published

Example 3: Cancer genome atlas “TCGA”



NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs

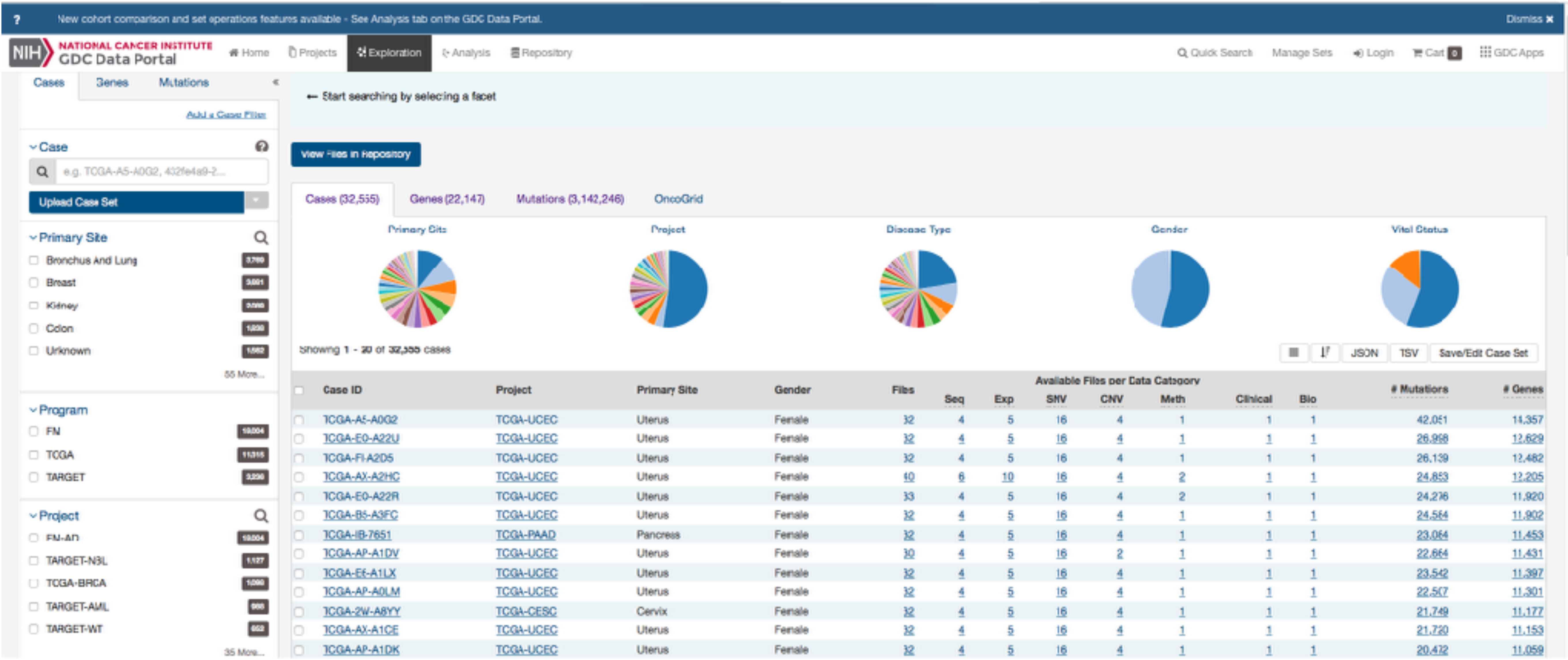
TCGA data describes **33** DIFFERENT TUMOR TYPES

...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

Much more complex reference points: on many levels: patients, cancer types tissue, genes, mutations



? New cohort comparison and set operations features available - See Analysis tab on the GDC Data Portal. Dismiss ✖

NIH NATIONAL CANCER INSTITUTE GDC Data Portal | Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

Cases | Genes | Mutations

Add a Case Filter

← Start searching by selecting a facet

Case

e.g. TCGA-A5-A0G2, 432fe4a9-2...

Upload Case Set

View Files in Repository

Primary Site

- □ Bronchus And Lung 3,789
- □ Breast 3,081
- □ Kidney 2,098
- □ Colon 1,820
- □ Unknown 1,582

55 More...

Program

- □ FM 18,004
- □ TCGA 11,315
- □ TARGET 3,236

Project

- □ FM-AD 18,004
- □ TARGET-NBL 1,127
- □ TCGA-BRCA 1,098
- □ TARGET-AML 988
- □ TARGET-WT 652

35 More...

Cases (32,555) | Genes (22,147) | Mutations (3,142,246) | OncoGrid

Showing 1 - 20 of 32,555 cases

JSON | TSV | Save/Edit Case Set

□	Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category								# Mutations	# Genes
						Seq	Exp	SNV	CNV	Meth	Clinical	Bio			
□	TCGA-A5-A0G2	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		42,051	14,357
□	TCGA-E0-A22U	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		26,958	12,629
□	TCGA-FI-A2D5	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		26,139	12,482
□	TCGA-AX-A2HC	TCGA-UCEC	Uterus	Female	40	6	10	16	4	2	1	1		24,853	12,205
□	TCGA-E0-A22R	TCGA-UCEC	Uterus	Female	33	4	5	16	4	2	1	1		24,276	11,920
□	TCGA-B5-A3FC	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		24,564	11,902
□	TCGA-IB-7651	TCGA-PAAD	Pancreas	Female	32	4	5	16	4	1	1	1		23,064	11,453
□	TCGA-AP-A1DV	TCGA-UCEC	Uterus	Female	30	4	5	16	2	1	1	1		22,664	11,431
□	TCGA-E6-A1LX	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		23,542	11,397
□	TCGA-AP-A0LM	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		22,507	11,301
□	TCGA-2W-A8YY	TCGA-CESC	Cervix	Female	32	4	5	16	4	1	1	1		21,749	11,177
□	TCGA-AX-A1CE	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		21,720	11,153
□	TCGA-AP-A1DK	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1		20,472	11,059

Fantastic overview on their strategy for different levels of data here:

https://cancergenome.nih.gov/abouttcga/aboutdata/datalevelstypes

Summary

Data integration important, but often challenging - especially for ad-hoc data

No magic: Careful planning and setting goals, vision and narrative important. Once narrative is in place, integration “only” become a technical issue

Experimental design, data quality, outlier and bath identification is an important part of the integration - and it strangely enough becomes even more important if you collect other peoples data

Budget much more than you think to integration and integrative analysis - it takes time and good people

Consider what data that should be integrated: why should it be, what is the end goal. Avoid integration because of the “hype”

There are data and projects where data integration makes zero sense.

Credits

LUNDBECKFONDEN

novo nordisk fonden

DET FRIE FORSKNINGSRÅD
DANISH COUNCIL FOR INDEPENDENT RESEARCH

Innovationsfonden

MARIE CURIE ACTIONS

elixir