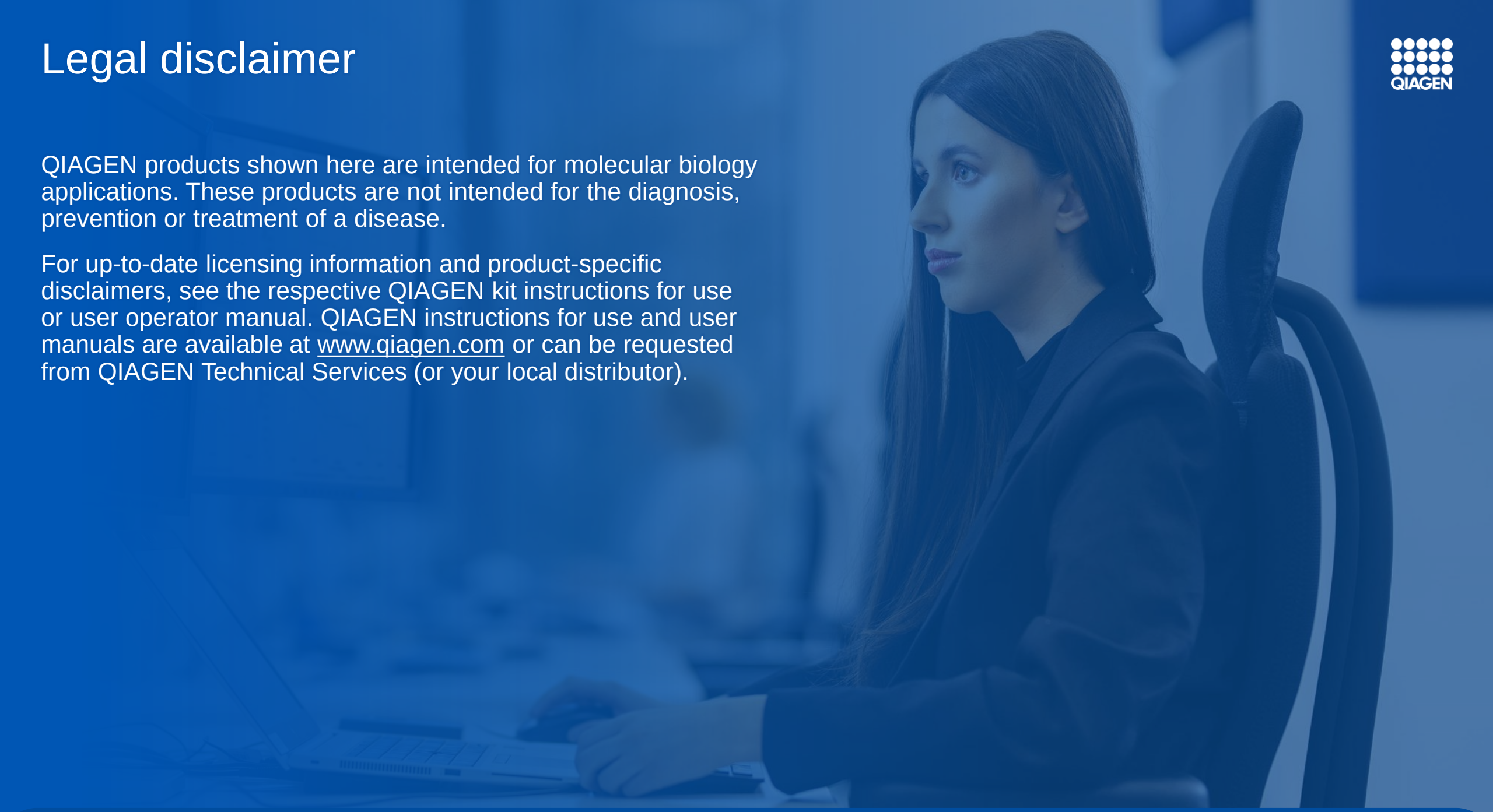
Accessing QIAGEN OmicSoft Data using the R API

Ruth.stoney@qiagen.com

Ruth Stoney
Field Applications Scientist

Legal disclaimer



QIAGEN products shown here are intended for molecular biology applications. These products are not intended for the diagnosis, prevention or treatment of a disease.

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit instructions for use or user operator manual. QIAGEN instructions for use and user manuals are available at www.qiagen.com or can be requested from QIAGEN Technical Services (or your local distributor).

Agenda

Introduction to the QIAGEN OmicSoft dataset

- QIAGEN
- OmicSoft dataset
- Why detailed curation is essential for public data

Accessing Datasets from R example (Lung Cancer)

- Finding lung cancer datasets within the database
- Finding lung cancer datasets with a specific mutation in the database
- Finding lung cancer datasets involving a specific drug

Example Queries

- Identifying differentially expressed genes
- Accessing gene expression data
- Finding genes with correlated expression

Agenda

Introduction to the QIAGEN OmicSoft dataset

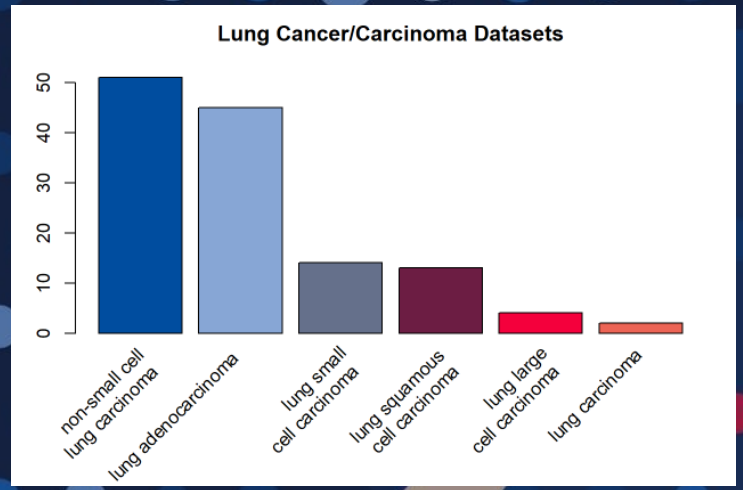
- QIAGEN
- OmicSoft dataset
- Why is detailed curation essential for public data?

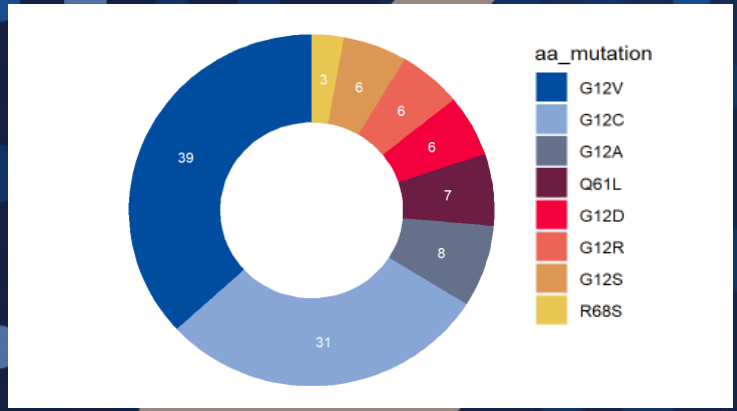
Accessing Datasets from R example (Lung Cancer)

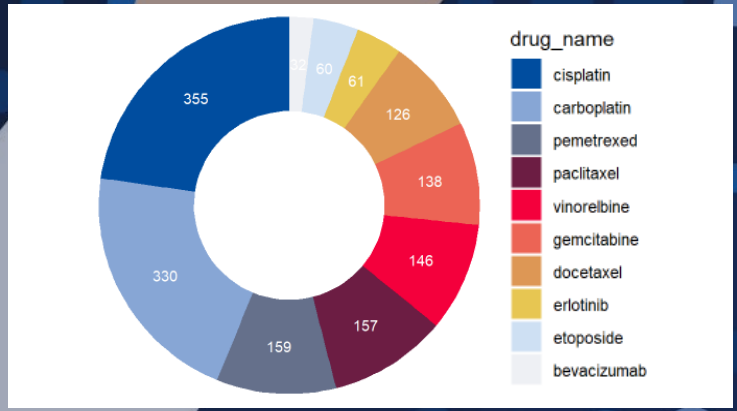
- Finding lung cancer datasets within the database
- Finding lung cancer datasets with a specific mutation
- Finding lung cancer datasets involving a specific drug

Example Queries

- Identifying differentially expressed genes
- Accessing gene expression data
- Finding genes with correlated expression







Agenda

Introduction to the QIAGEN OmicSoft dataset

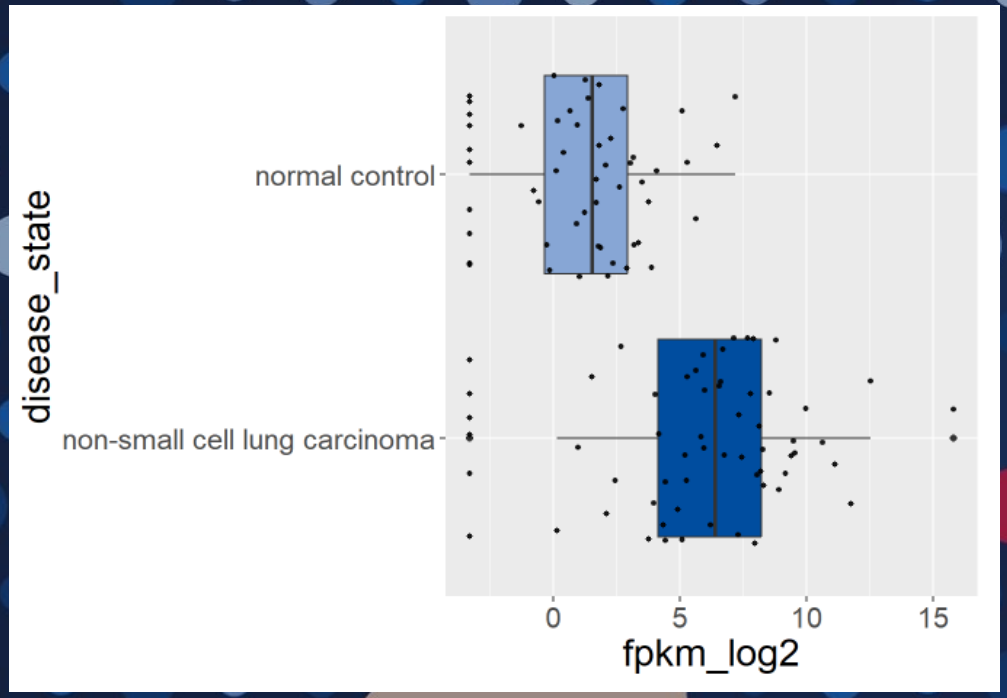
- QIAGEN
- OmicSoft dataset
- Why is detailed curation essential for public data?

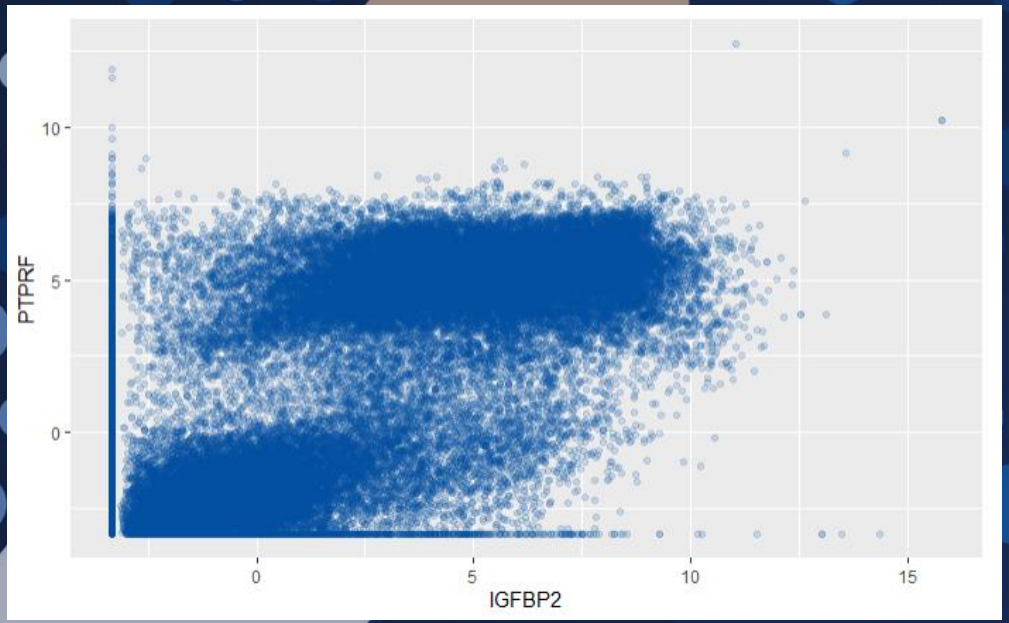
Accessing Datasets from R example (Lung Cancer)

- Finding lung cancer datasets within the database
- Finding lung cancer datasets with a specific mutation in the database
- Finding lung cancer datasets involving a specific drug

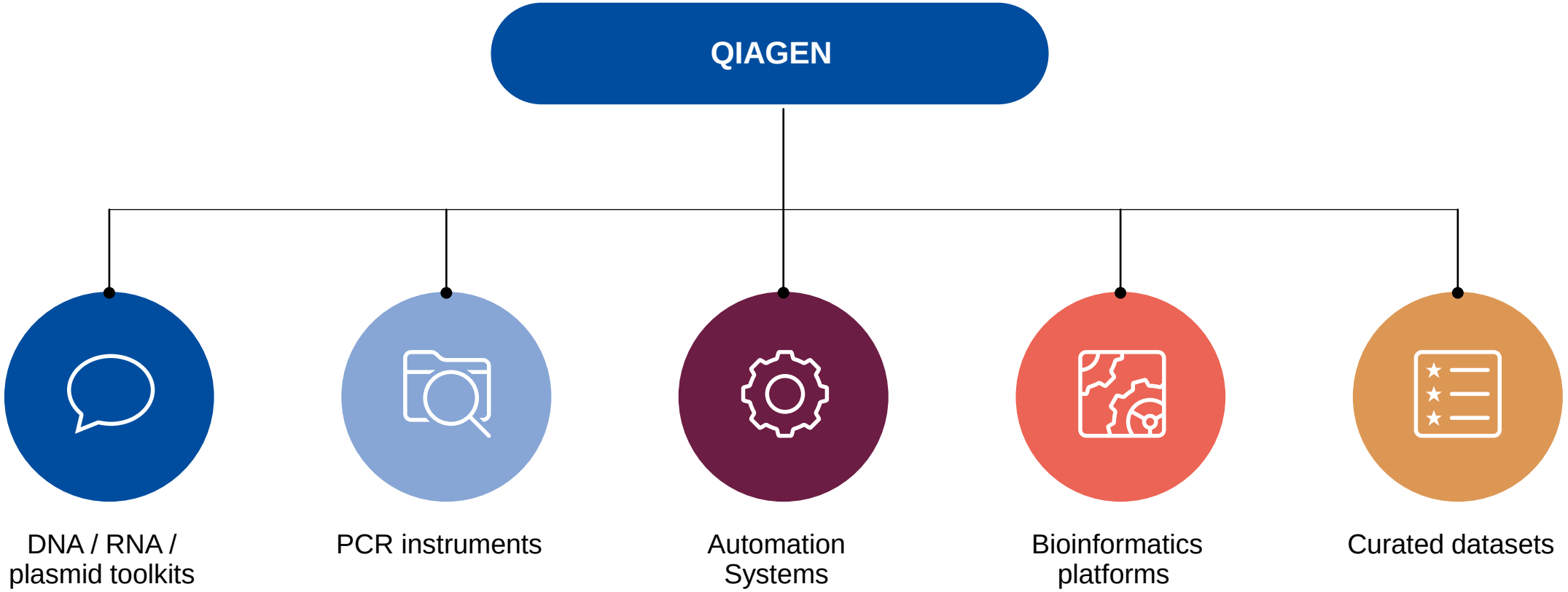
Example Queries

- Identifying differentially expressed genes
- Accessing gene expression data
- Finding genes with correlated expression

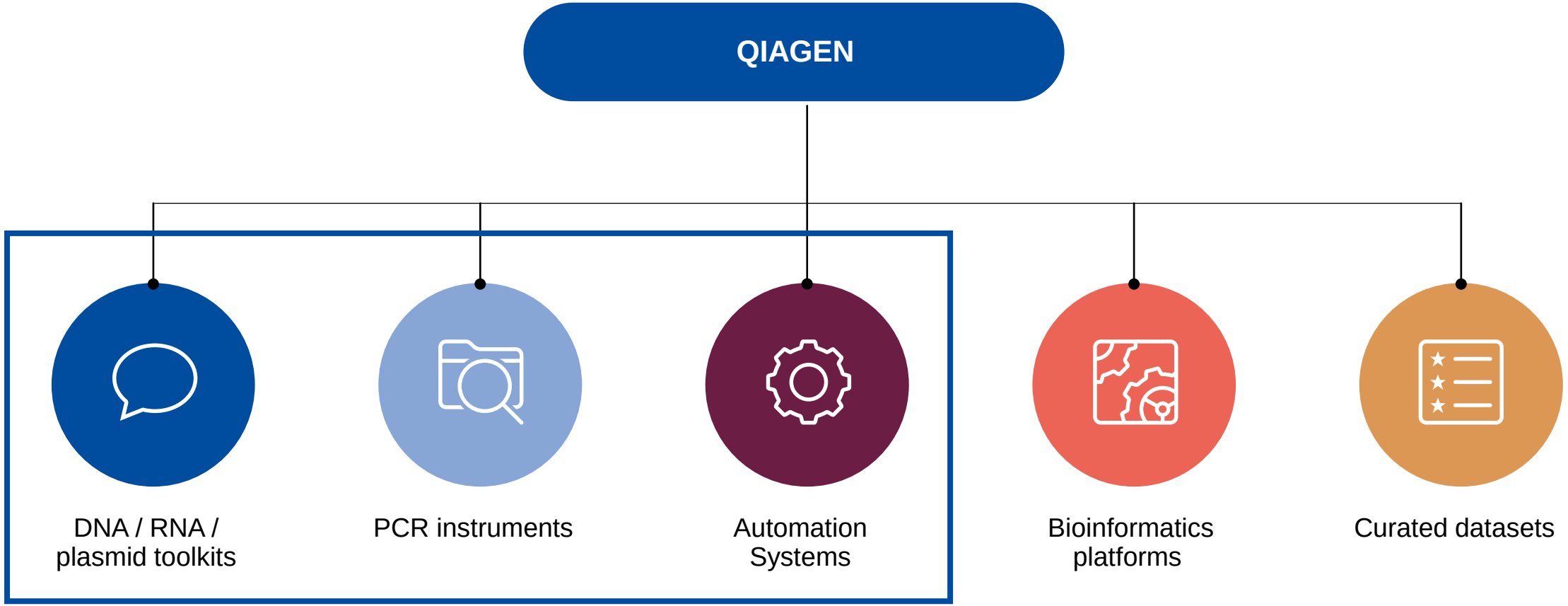




From sample to insight

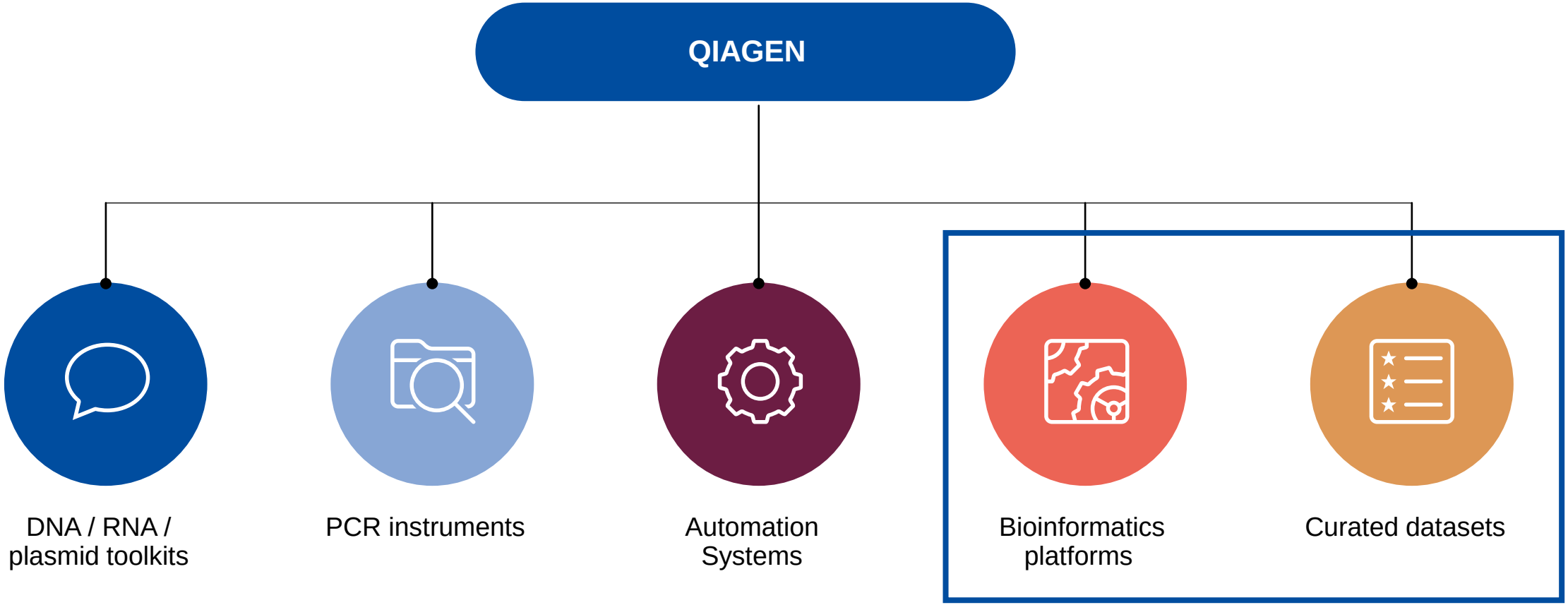


From sample to insight

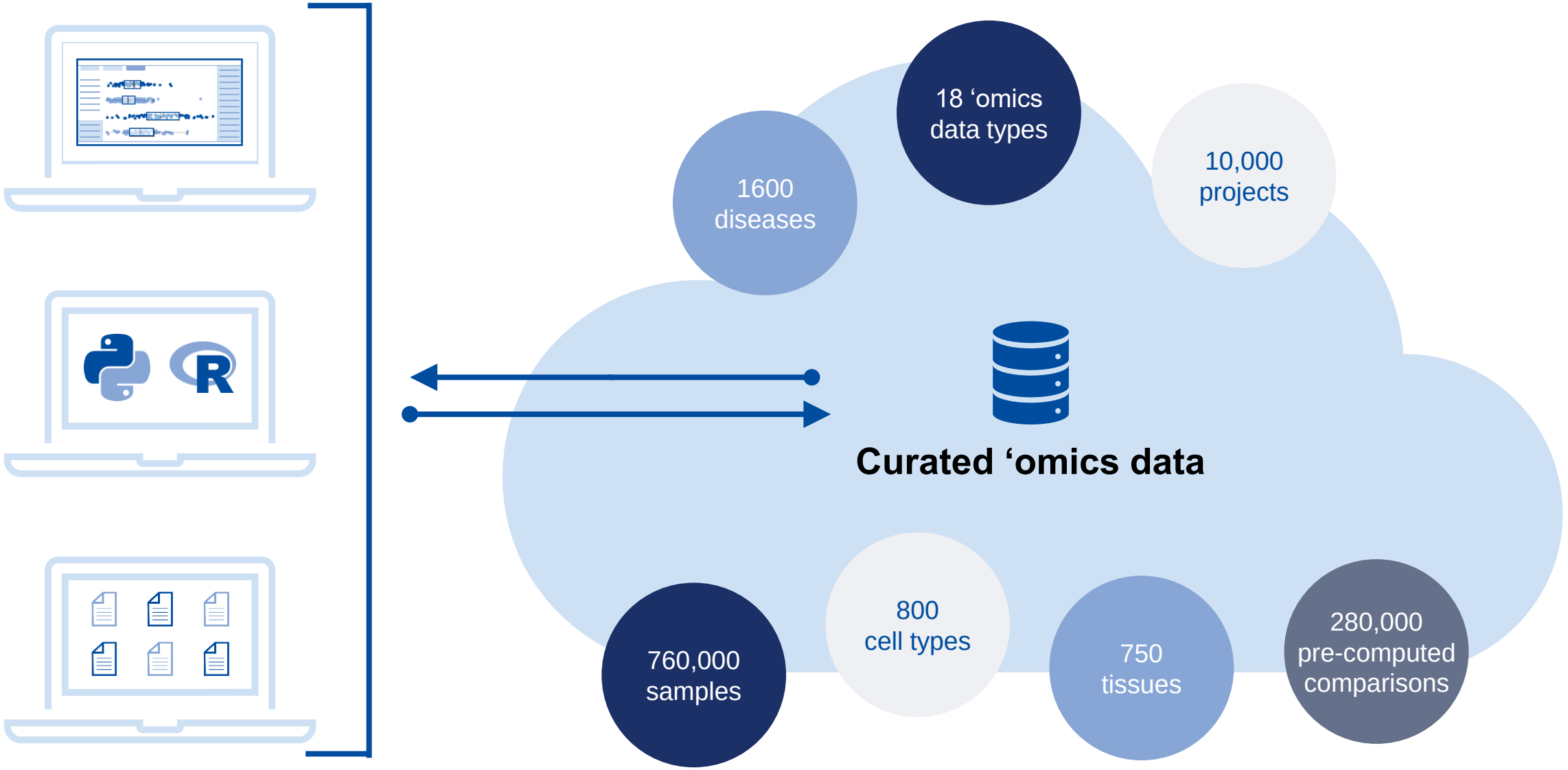


Process samples for Next Generation Sequencing

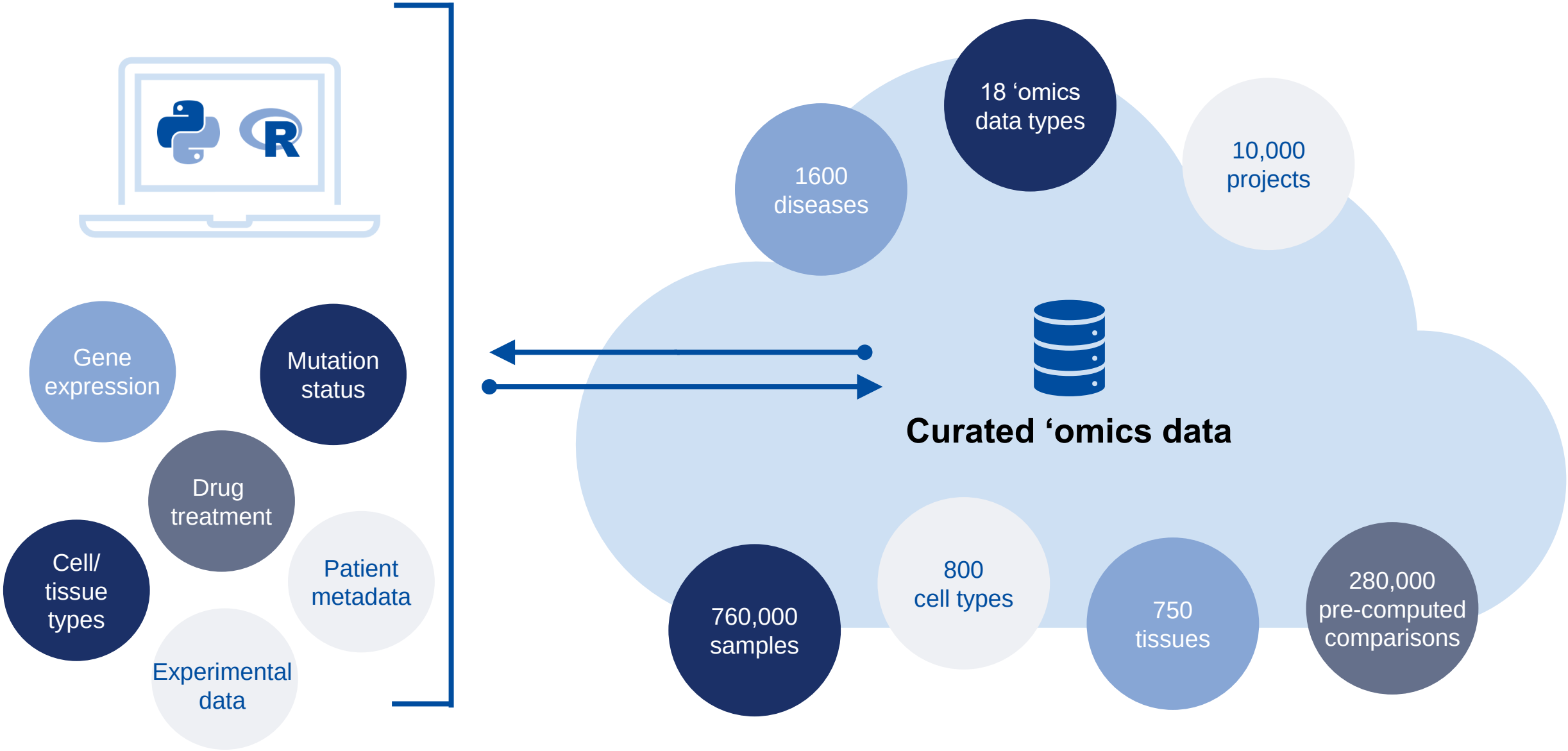
From sample to insight



OmicSoft Collection

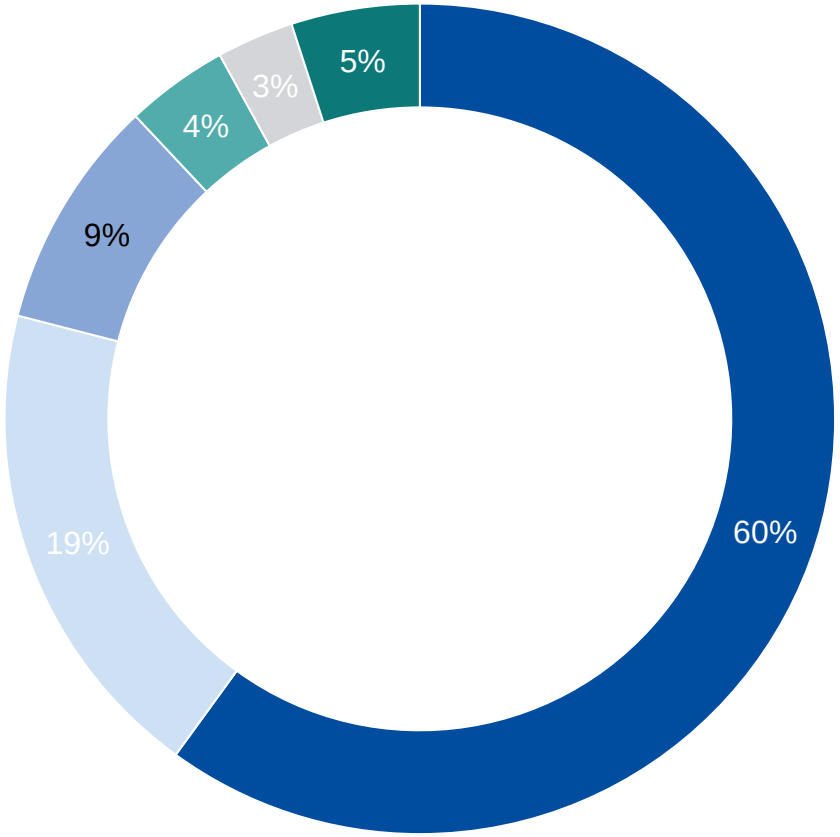


OmicSoft Collection



Why not use Open-Source Data?

Data scientists and bioinformaticians spend ~80% of their time collecting, cleaning and processing the data.



- Cleaning and organizing data: 60%
- Collecting datasets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Building training sets: 3%
- Others: 5%

Data source: Sarih, H., Tchangani, A. P., Medjaher, K. and Pere, E. (2019) Data preparation and preprocessing for broadcast systems monitoring in PHM framework. 6th International Conference on Control, Decision and Information Technologies (CoDIT). 1444–1449.

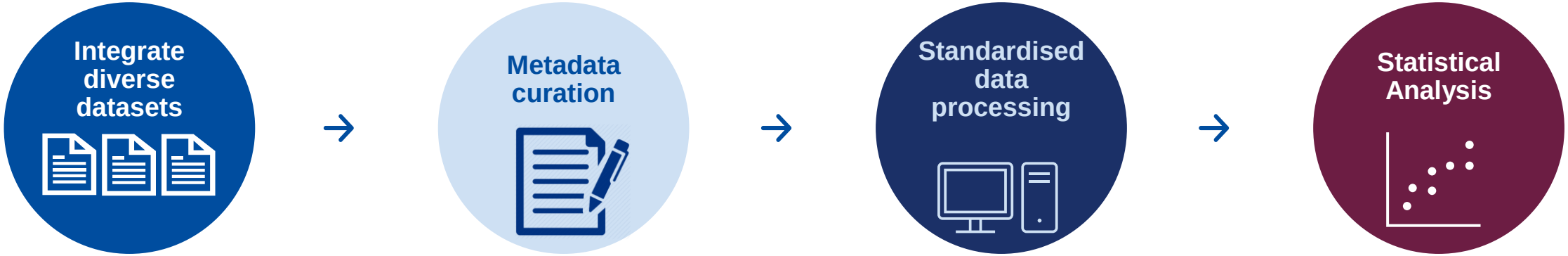
OmicSoft pipeline

Disease relevant **omics samples** are gathers from **hundreds of thousands of published projects** and popular **consortia**.

Manual metadata curation using controlled vocabularies

Extracted data is processed using **consistent bioinformatics pipelines**

All results are included in our models to generate **statistical analysis**



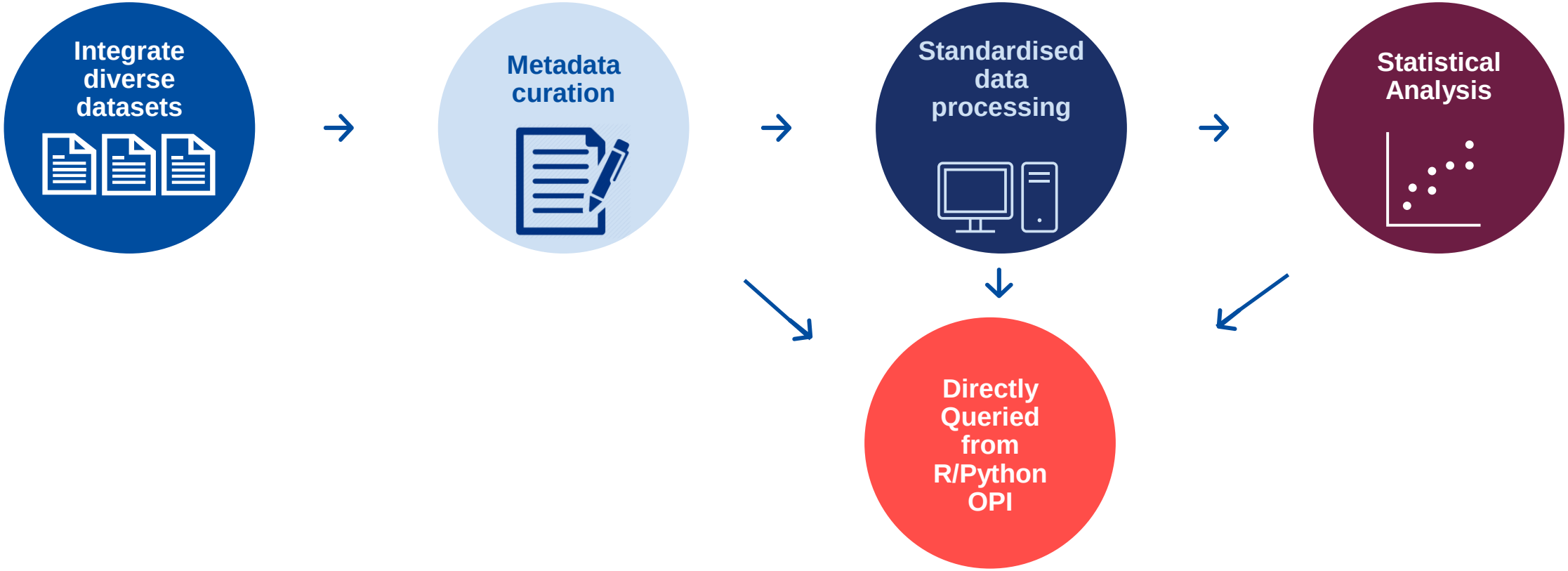
OmicSoft Land pipeline

Disease relevant **omics samples** are gathers from **hundreds of thousands of published projects** and popular **consortia**.

Manual metadata curation using controlled vocabularies

Extracted data is processed using **consistent bioinformatics pipelines**

All results are included in our models to generate **statistical analysis**



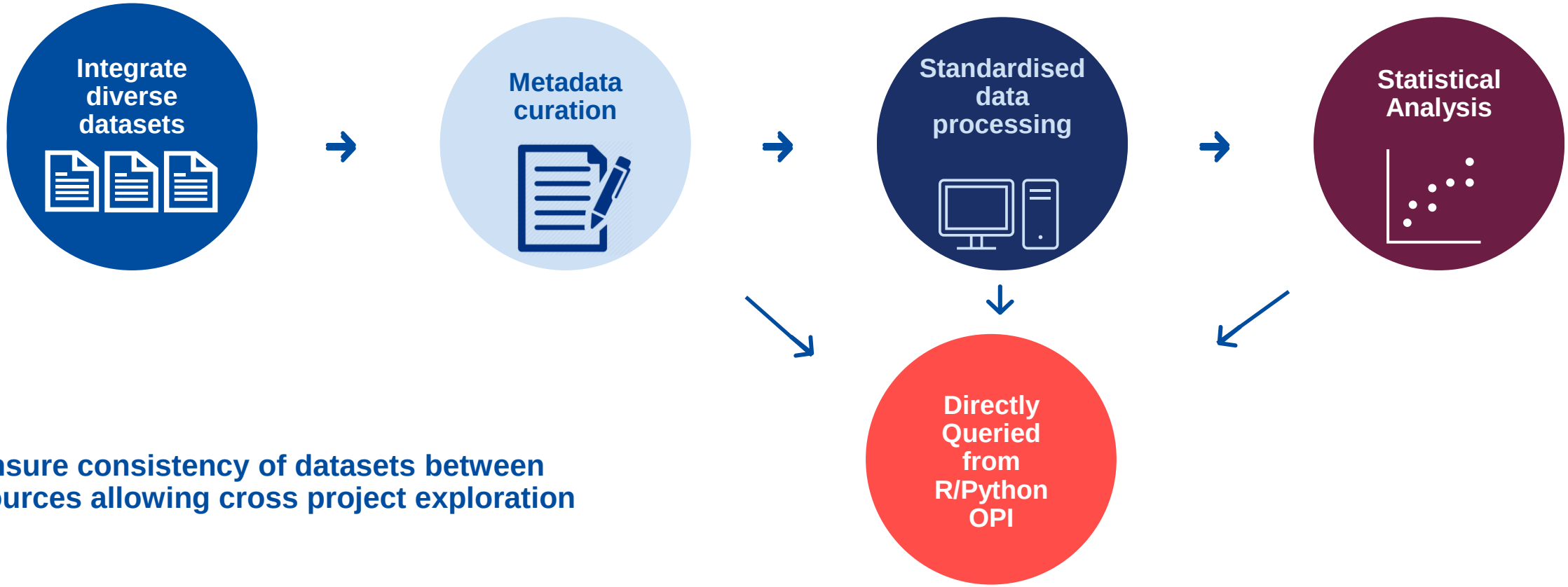
OmicSoft Land pipeline

Disease relevant **omics samples** are gathers from **hundreds of thousands of published projects** and popular **consortia**.

Manual metadata curation using controlled vocabularies

Extracted data is processed using **consistent bioinformatics pipelines**

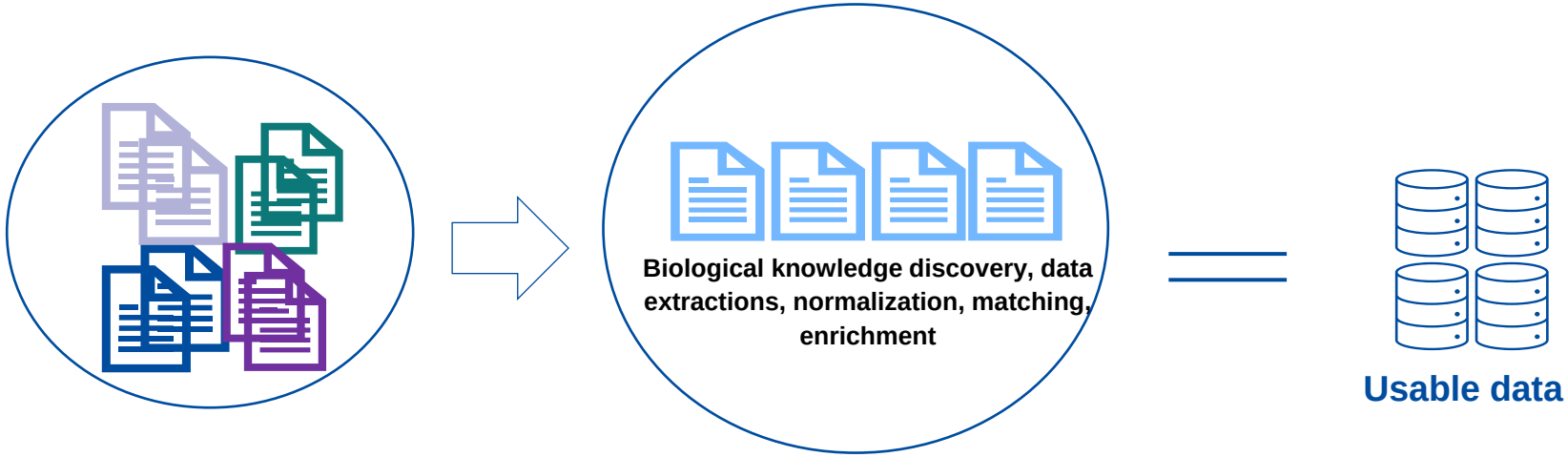
All results are included in our models to generate **statistical analysis**



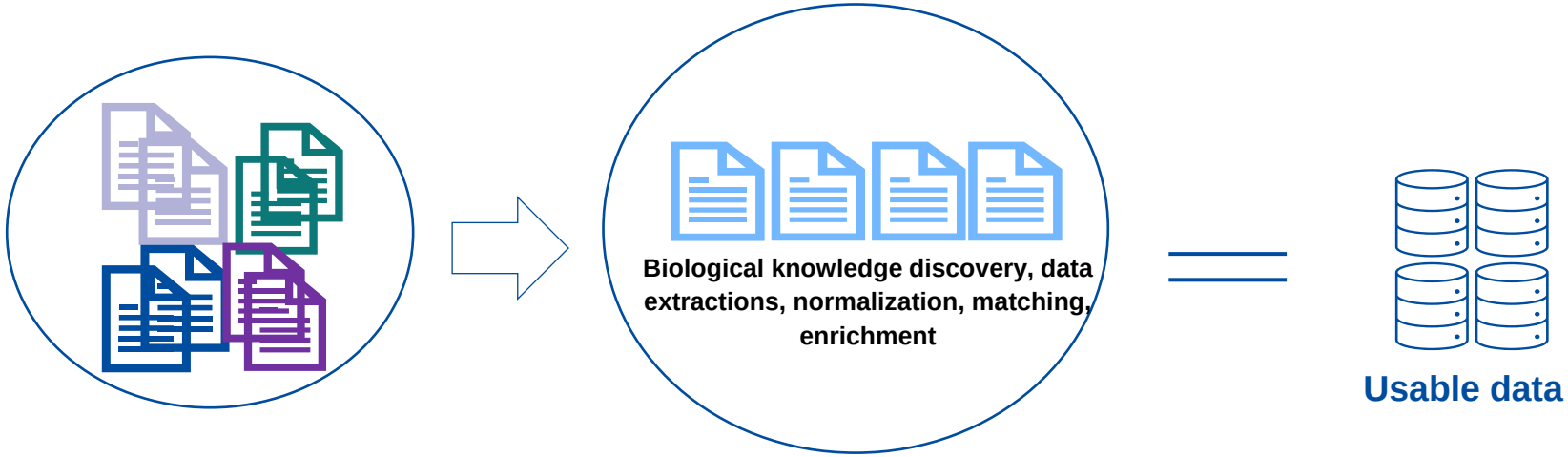
Ensure consistency of datasets between sources allowing cross project exploration

The OmicSoft collection

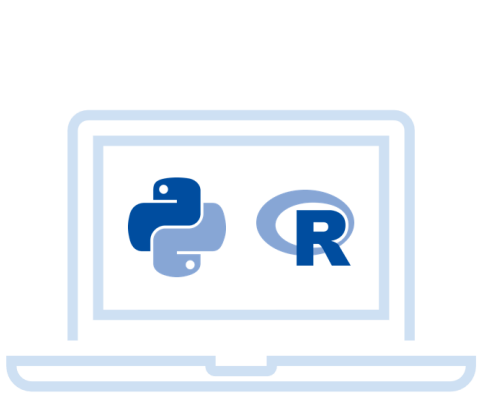
OmicSoft


Biological knowledge discovery, data extractions, normalization, matching, enrichment
Usable data
Heterogeneous data
Curation and processing

OmicSoft


Biological knowledge discovery, data extractions, normalization, matching, enrichment
Usable data
Heterogeneous data
Curation and processing

R example: Lung cancer

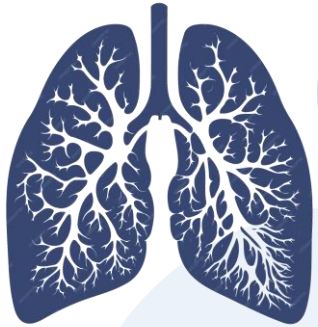


Gene expression

Mutation status

Cell/ tissue types

Drug treatment



18 ‘omics data types

1600 diseases

10,000 projects

Curated ‘omics data

760,000 samples

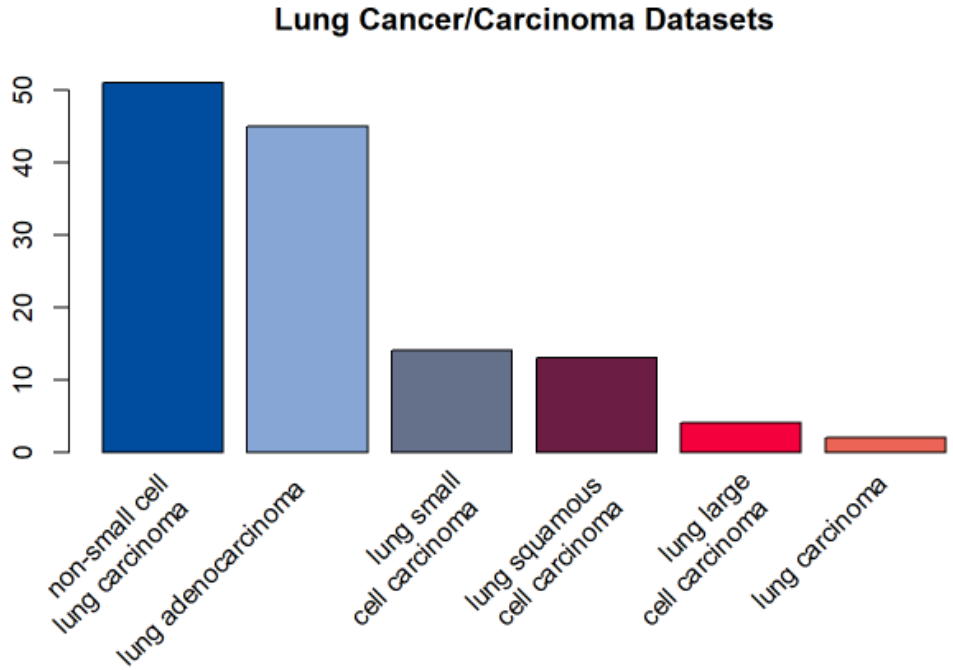
800 cell types

750 tissues

280,000 pre-computed comparisons

Finding Datasets related to Lung Cancer

```
datastore <- opi()
sql <- "
SELECT disease_state, _db_ as database, project_id, tissue, organism,
  count(*) AS no_samples
FROM samples
WHERE (LOWER(disease_state) LIKE '%lung%' AND
  LOWER(disease_state) LIKE  '%carcinoma%')
GROUP BY disease_state, _db_, project_id, tissue, organism
ORDER BY no_samples DESC
"

result<- datastore$query(sql)
counts <- sort(table(result[result$no_samples>30,'disease_state']),
               decreasing = TRUE)
```

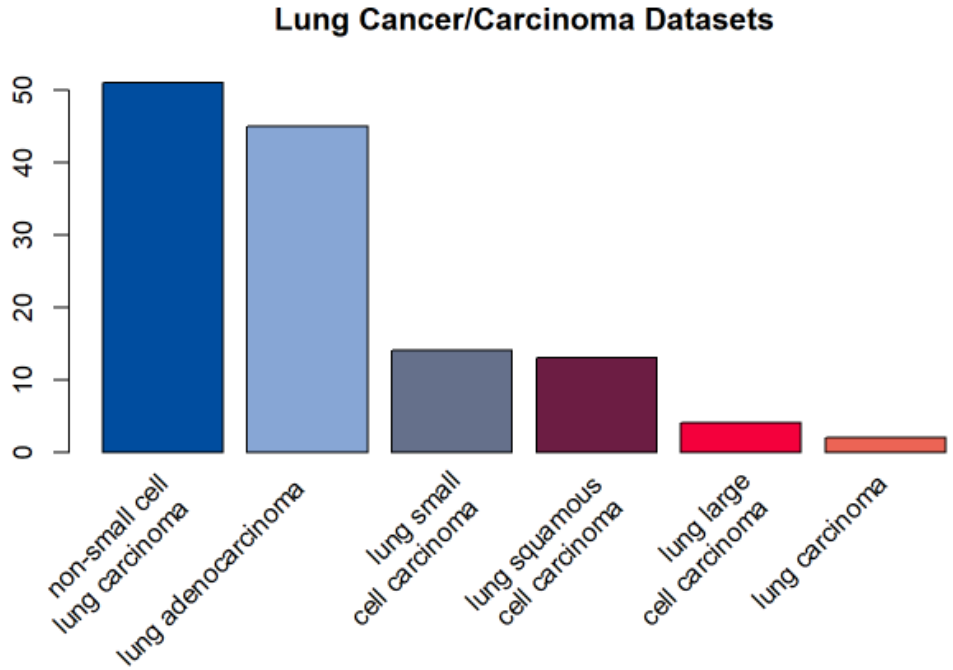


Lung Cancer Datasets

disease_state	database	project_id	tissue	organism	no_samples
lung adenocarcinoma (LUAD)	lincs_b38_gc33	LINCS_B38_GC33	lung	human	19558
lung adenocarcinoma (LUAD)	tcga_b38_gc33	TCGA_LUAD	lung	human	767
lung squamous cell carcinoma (LUSC)	tcga_b38_gc33	TCGA_LUSC	lung	human	747
non-small cell lung carcinoma	trace_rx_b38_gc33	TRACERx_2020R3	lung	human	447
lung adenocarcinoma (LUAD)	onco_human_b38_gc33	GSE72094	lung	human	442

Finding Datasets related to Lung Cancer

```
datastore <- opi()
sql <- "
SELECT disease_state, _db_ as database, project_id, tissue, organism,
  count(*) AS no_samples
FROM samples
WHERE (LOWER(disease_state) LIKE '%lung%' AND
  LOWER(disease_state) LIKE  '%carcinoma%')
GROUP BY disease_state, _db_, project_id, tissue, organism
ORDER BY no_samples DESC
"

result<- datastore$query(sql)
counts <- sort(table(result[result$no_samples>30,'disease_state']),
               decreasing = TRUE)
```

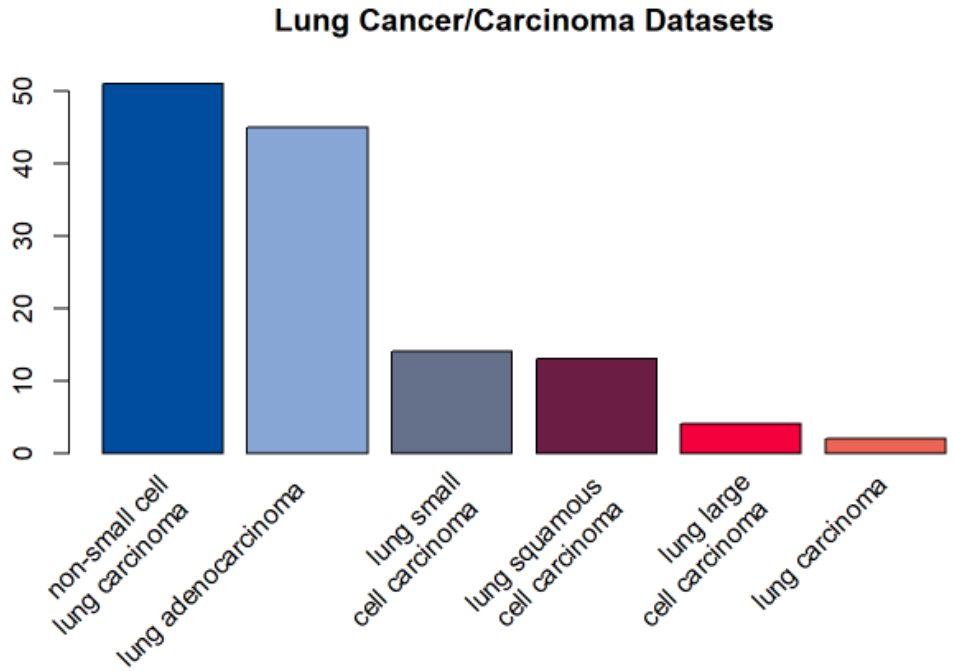


Lung Cancer Datasets

disease_state	database	project_id	tissue	organism	no_samples
lung adenocarcinoma (LUAD)	lincs_b38_gc33	LINCS_B38_GC33	lung	human	19558
lung adenocarcinoma (LUAD)	tcga_b38_gc33	TCGA_LUAD	lung	human	767
lung squamous cell carcinoma (LUSC)	tcga_b38_gc33	TCGA_LUSC	lung	human	747
non-small cell lung carcinoma	trace_rx_b38_gc33	TRACERx_2020R3	lung	human	447
lung adenocarcinoma (LUAD)	onco_human_b38_gc33	GSE72094	lung	human	442

Finding Datasets related to Lung Cancer

```
sql <- "
SELECT disease_state, _db_ as database, project_id, tissue, organism,
  count(*) AS no_samples
FROM samples
WHERE (LOWER(disease_state) LIKE '%lung%' AND
  LOWER(disease_state) LIKE  '%carcinoma%')
GROUP BY disease_state, _db_, project_id, tissue, organism
ORDER BY no_samples DESC
"

result<- datastore$query(sql)
counts <- sort(table(result[result$no_samples>30,'disease_state']),
               decreasing = TRUE)
```

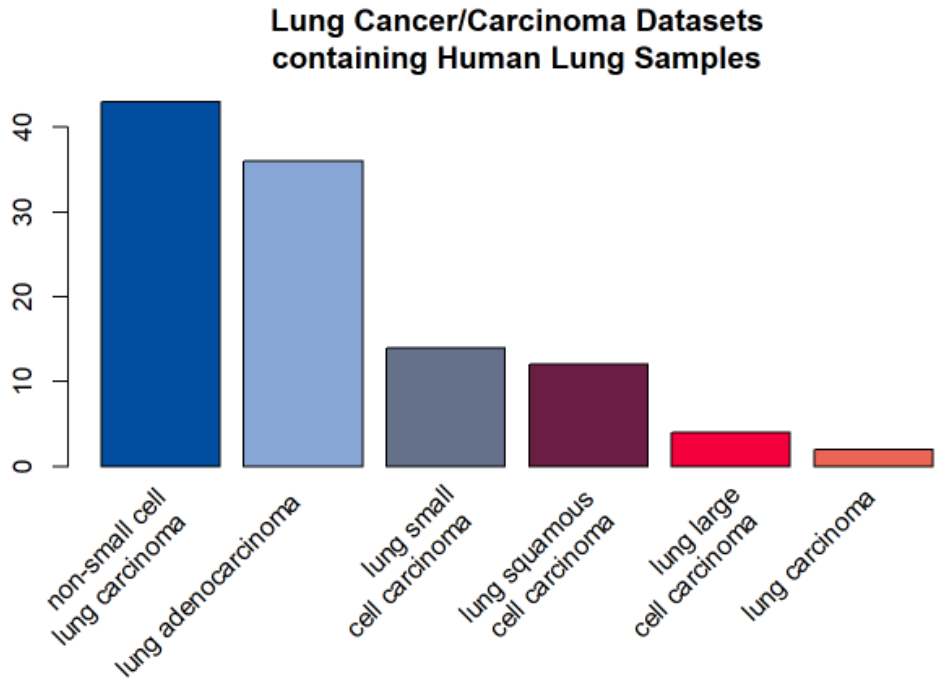


Lung Cancer Datasets

disease_state	database	project_id	tissue	organism	no_samples
lung adenocarcinoma (LUAD)	lincs_b38_gc33	LINCS_B38_GC33	lung	human	19558
lung adenocarcinoma (LUAD)	tcga_b38_gc33	TCGA_LUAD	lung	human	767
lung squamous cell carcinoma (LUSC)	tcga_b38_gc33	TCGA_LUSC	lung	human	747
non-small cell lung carcinoma	trace_rx_b38_gc33	TRACERx_2020R3	lung	human	447
lung adenocarcinoma (LUAD)	onco_human_b38_gc33	GSE72094	lung	human	442

Finding Datasets related to Lung Cancer

```
sql <- "
SELECT disease_state, _db_ as database, project_id, tissue, organism,
  count(*) AS no_samples
FROM samples
WHERE (LOWER(disease_state) LIKE '%lung%' AND
  LOWER(disease_state) LIKE  '%carcinoma%') AND
  organism = 'human'AND
  tissue = 'lung'
GROUP BY disease_state, _db_, project_id, tissue, organism
ORDER BY no_samples DESC
"

result<- datastore$query(sql)
counts <- sort(table(result[result$no_samples>30,'disease_state']),
               decreasing = TRUE)
```



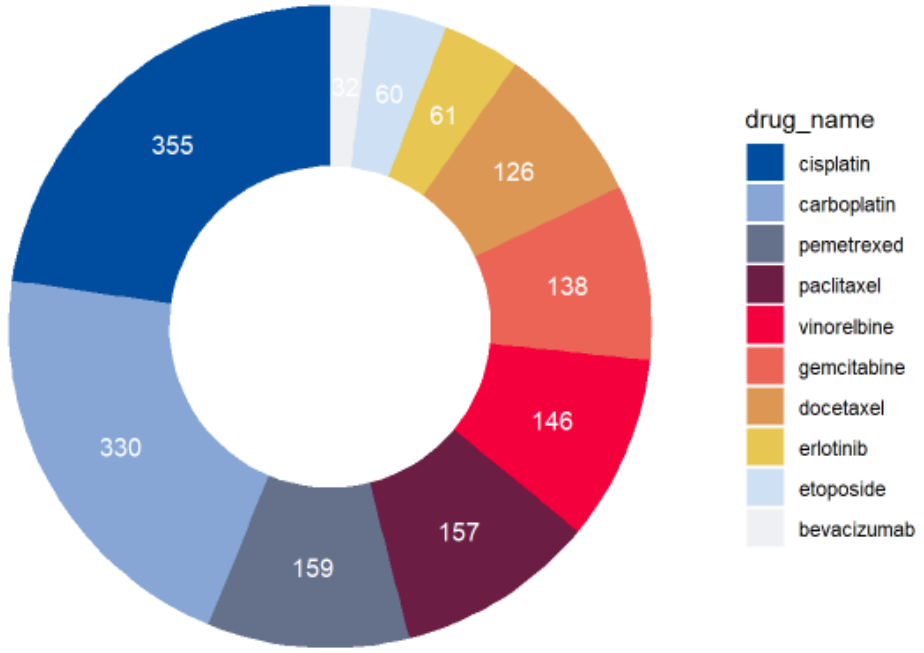
Lung Cancer Datasets

disease_state	database	project_id	tissue	organism	no_samples
lung adenocarcinoma (LUAD)	lincs_b38_gc33	LINCS_B38_GC33	lung	human	19558
lung adenocarcinoma (LUAD)	tcga_b38_gc33	TCGA_LUAD	lung	human	767
lung squamous cell carcinoma (LUSC)	tcga_b38_gc33	TCGA_LUSC	lung	human	747
non-small cell lung carcinoma	trace_rx_b38_gc33	TRACERx_2020R3	lung	human	447
lung adenocarcinoma (LUAD)	onco_human_b38_gc33	GSE72094	lung	human	442

Finding Drug Datasets

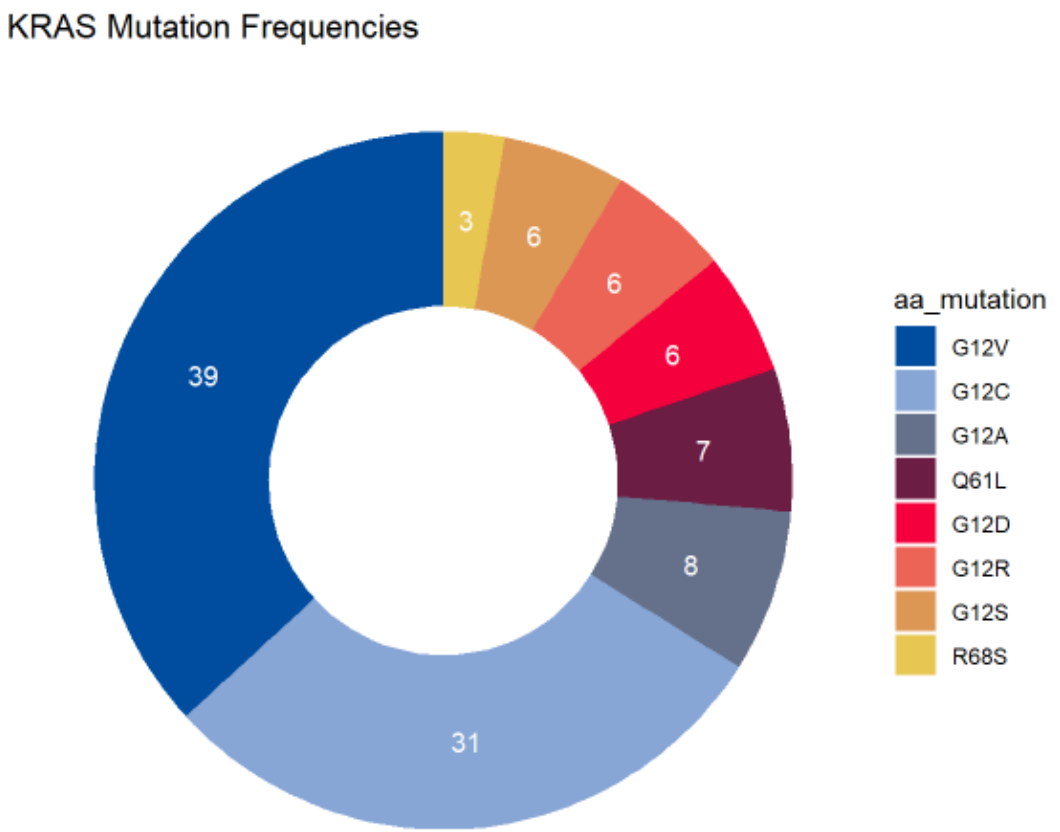
```
sql <- "
SELECT
     project_id, disease_state, tissue, drug_name,
     count(*) AS count
FROM
    samples
LEFT JOIN (
    SELECT sample_id, CAST(value as VARCHAR) AS drug_name
    FROM clinical_triplets
    WHERE attribute = 'drug_name'
    ) USING (sample id)
WHERE
    LOWER(disease_state) LIKE '%lung%carcinoma%' AND
    LOWER(drug_name) not LIKE 'NA'
GROUP BY
    project_id, disease_state, tissue, drug_name
ORDER BY count DESC
"

drugs<- datastore$query(sql)
drugs2 <- drugs %>%
  separate_rows(drug_name, sep=";")

result <- drugs2 %>%
  select(drug_name, count) %>%
  group_by(drug_name) %>%
  summarise(total_samples= sum(count)) %>%
  arrange(desc(total_samples))
```

```
result<- result[1:10,]
result$drug_name <- reorder(result$drug_name, - result$total_samples)

# Create the doughnut plot
ggplot(result, aes(x = 2, y =total_samples, fill = drug_name)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  xlim(0.5, 2.5) + theme_void() +
  theme(legend.position = "right") +
  geom_text(aes(label = paste0(total_samples)),
            position = position_stack(vjust = 0.5), color = "white") +
  scale_fill_manual(values = my_cols) + ggtitle("Drug Treatment Samples")
```



Finding NSCLC with KRAS Mutations

```
sql<- "
SELECT
    project_id, tissue, disease_state, mc.gene_name, mc.aa_mutation
FROM
    samples
JOIN
    dna_seq_somatic_mutation dssm ON samples.sample_index = dssm.sample_index
JOIN
    mutations_canonical mc on mc.mutation_index = dssm.mutation_index
WHERE
    LOWER(disease_state) = 'non-small cell lung carcinoma' AND
    mc.gene_name = 'KRAS'
"

mutations<- datastore$query(sql)
```



Identifying Differentially Expressed Genes
Pre-computed Comparison: Disease vs. Normal

```
sql <- "
SELECT  gene_name, log2_fold_change, adjusted_p_value, project_id
FROM human_disease_b38_gc33.comparison_data
JOIN human_disease_b38_gc33.comparisons USING (comparison_index)
JOIN human_disease_b38_gc33.gene_annotation USING (gene_index)
WHERE
     LOWER(case_disease_state) = 'non-small cell lung carcinoma' AND
     comparison_category = 'Disease vs. Normal' AND
     log2_fold_change > 3  AND
     p_value < 0.001 AND
     case_sample_size > 50 AND
     control_sample_size > 50 AND
     sample_data_mode = 'RnaSeq_Transcript'
ORDER BY  log2_fold_change DESC
"

result<- datastore$query(sql)
```

Lung Cancer Differential Expression - Gene vs Normal

gene_name	log2_fold_change	adjusted_p_value	project_id
AC000093.1	6.9448	0.00e+00	GSE68086
PRR7	5.6663	0.00e+00	GSE68086
GP1BB	5.3220	0.00e+00	GSE68086
IGFBP2	5.0468	0.00e+00	GSE68086
SH2B2	4.8870	0.00e+00	GSE68086
LYL1	4.4718	0.00e+00	GSE68086
EVA1B	4.2421	0.00e+00	GSE68086
NUDT4B	4.1178	0.00e+00	GSE68086
ANKRD9	3.9237	0.00e+00	GSE68086
TPGS1	3.8363	8.61e-05	GSE68086

Identifying Differentially Expressed Genes
Pre-computed Comparisons

```
sql <- "
SELECT  gene_name, log2_fold_change, adjusted_p_value, project_id
FROM human_disease_b38_gc33.comparison_data
JOIN human_disease_b38_gc33.comparisons USING (comparison_index)
JOIN human_disease_b38_gc33.gene_annotation USING (gene_index)
WHERE
     LOWER(case_disease_state) = 'non-small cell lung carcinoma' AND
     comparison_category = 'Disease vs. Normal' AND
     log2_fold_change > 3  AND
     p_value < 0.001 AND
     case_sample_size > 50 AND
     control_sample_size > 50 AND
     sample_data_mode = 'RnaSeq_Transcript'
ORDER BY  log2_fold_change DESC
"

result<- datastore$query(sql)
```

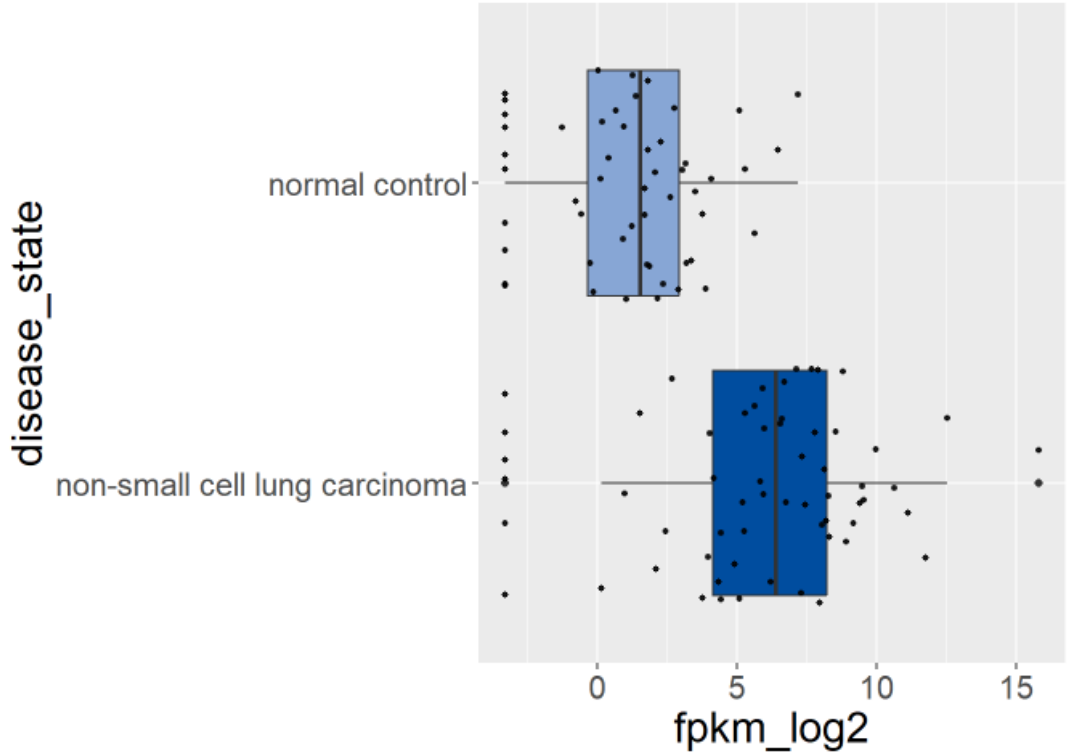
Lung Cancer Differential Expression - Gene vs Normal

gene_name	log2_fold_change	adjusted_p_value	project_id
AC000093.1	6.9448	0.00e+00	GSE68086
PRR7	5.6663	0.00e+00	GSE68086
GP1BB	5.3220	0.00e+00	GSE68086
IGFBP2	**5.0468**	**0.00e+00**	**GSE68086**
SH2B2	4.8870	0.00e+00	GSE68086
LYL1	4.4718	0.00e+00	GSE68086
EVA1B	4.2421	0.00e+00	GSE68086
NUDT4B	4.1178	0.00e+00	GSE68086
ANKRD9	3.9237	0.00e+00	GSE68086
TPGS1	3.8363	8.61e-05	GSE68086

Accessing Gene Expression Data

```
sql <- "
SELECT _db_, gene_name, fpkm, disease_state, project_id
FROM gene_fpkm
JOIN samples USING (sample_index)
JOIN gene_annotation USING (gene_index)
WHERE
      gene_name = 'IGFBP2' AND
      project_id = 'GSE68086' AND
      (LOWER(samples.disease_state) = 'non-small cell lung carcinoma' OR
      LOWER(samples.disease_state) LIKE '%control%')
"

result <- as.data.frame(datastore$query(sql))
result$fpkm_log2 <- log2(result$fpkm + 0.1)
```

```
# plot boxplot
plot <- ggplot(result,
               aes(x=fpkm_log2, y=disease_state, fill=disease_state)) +
  geom_boxplot(show.legend = F) +
  geom_jitter(color="black", size=1, alpha=0.9, show.legend = F) +
  scale_fill_manual(values = my_cols[1:2]) +
  theme(axis.text.x= element_text(size = 15),
        axis.text.y= element_text(size = 15),
        axis.title.x= element_text(size = 20),
        axis.title.y= element_text(size = 20),
        )
```



Gene Co-expression

```
sql <- "
SELECT gene, gene_annotation.gene_name,
CORR(log_fpkm_1,log_fpkm_2) AS correlation
FROM
(
    SELECT
        gene1_fpmk.gene_index AS gene,
        log2(gene1_fpmk.fpkm+0.1) AS log_fpkm_1,
        log2(gene2_fpmk.fpkm+0.1) AS log_fpkm_2
    FROM
        human_disease_b38_gc33.gene_fpkm gene1_fpmk,
        human_disease_b38_gc33.gene_fpkm gene2_fpmk
    WHERE
        gene2_fpmk.gene_index = 9112 AND
        gene1_fpmk.sample_index = gene2_fpmk.sample_index
)
JOIN human_disease_b38_gc33.gene_annotation ON gene = human_disease_b38_gc33.gene_annotation.gene_index
GROUP BY gene, gene_annotation.gene_name, gene_annotation.gene_index
ORDER BY correlation DESC
LIMIT 11
"
result <- datastore$query(sql)
```

Lung Cancer Differential Expression - Gene vs Normal

gene_name	correlation
PTPRF	0.7213568
MDK	0.7185268
ADCY6	0.7088722
MMP15	0.7009866
TMEM98	0.6950555
BCAR1	0.6925772
SHROOM3	0.6871723
TSPAN6	0.6805749
EFNB2	0.6798894
VWA1	0.6788953

Gene co-expression

```
sql <- "
SELECT gene, gene_annotation.gene_name,
CORR(log_fpkm_1,log_fpkm_2) AS correlation
FROM
(
    SELECT
        gene1_fpmk.gene_index AS gene,
        log2(gene1_fpmk.fpkm+0.1) AS log_fpkm_1,
        log2(gene2_fpmk.fpkm+0.1) AS log_fpkm_2
    FROM
        human_disease_b38_gc33.gene_fpkm gene1_fpmk,
        human_disease_b38_gc33.gene_fpkm gene2_fpmk
    WHERE
        gene2_fpmk.gene_index = 9112 AND
        gene1_fpmk.sample_index = gene2_fpmk.sample_index
)
JOIN human_disease_b38_gc33.gene_annotation ON gene = human_disease_b38_gc33.gene_annotation.gene_index
GROUP BY gene, gene_annotation.gene_name, gene_annotation.gene_index
ORDER BY correlation DESC
LIMIT 11
"
result <- datastore$query(sql)
```

Lung Cancer Differential Expression - Gene vs Normal

gene_name	correlation
PTPRF	0.7213568
MDK	0.7185268
ADCY6	0.7088722
MMP15	0.7009866
TMEM98	0.6950555
BCAR1	0.6925772
SHROOM3	0.6871723
TSPAN6	0.6805749
EFNB2	0.6798894
VWA1	0.6788953

Gene co-expression

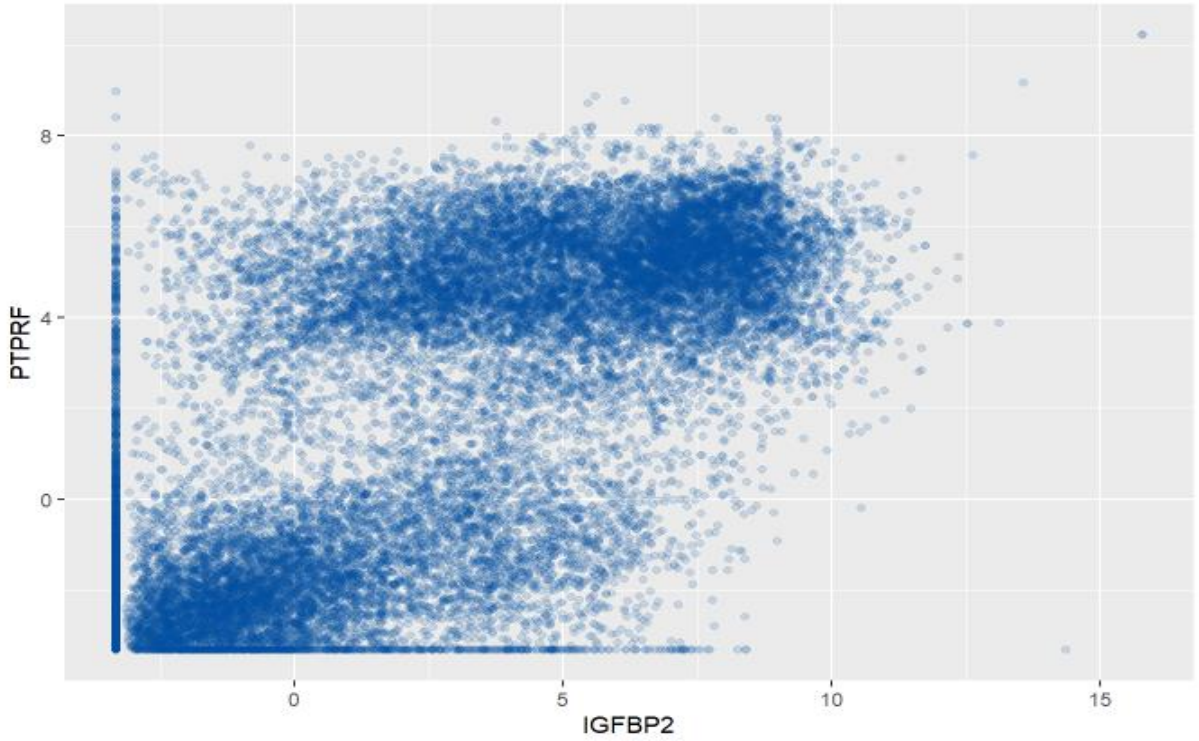
```
sql <- "
SELECT gene, gene_annotation.gene_name,
CORR(log_fpkm_1,log_fpkm_2) AS correlation
FROM
(
    SELECT
        gene1_fpmk.gene_index AS gene,
        log2(gene1_fpmk.fpkm+0.1) AS log_fpkm_1,
        log2(gene2_fpmk.fpkm+0.1) AS log_fpkm_2
    FROM
        human_disease_b38_gc33.gene_fpkm gene1_fpmk,
        human_disease_b38_gc33.gene_fpkm gene2_fpmk
    WHERE
        gene2_fpmk.gene_index = 9112 AND
        gene1_fpmk.sample_index = gene2_fpmk.sample_index
)
JOIN human_disease_b38_gc33.gene_annotation ON gene = human_disease_b38_gc33.gene_annotation.gene_index
GROUP BY gene, gene_annotation.gene_name, gene_annotation.gene_index
ORDER BY correlation DESC
LIMIT 11
"
result <- datastore$query(sql)
```

Lung Cancer Differential Expression - Gene vs Normal

gene_name	correlation
PTPRF	0.7213568
MDK	0.7185268
ADCY6	0.7088722
MMP15	0.7009866
TMEM98	0.6950555
BCAR1	0.6925772
SHROOM3	0.6871723
TSPAN6	0.6805749
EFNB2	0.6798894
VWA1	0.6788953

Gene co-expression

```
sql <- "
SELECT gene_name, log2(fpkm+0.1) AS log_fpkm, sample_index, project_id
FROM human_disease_b38_gc33.gene_fpkm
JOIN human_disease_b38_gc33.samples USING (sample_index)
JOIN human_disease_b38_gc33.gene_annotation USING (gene_index)
WHERE
      (gene_name = 'IGFBP2' OR gene_name = 'PTPRF') AND
      (LOWER(samples.disease_state) = 'non-small cell lung carcinoma' OR
      LOWER(samples.disease_state) LIKE '%control%')
"

result <- datastore$query(sql)
```

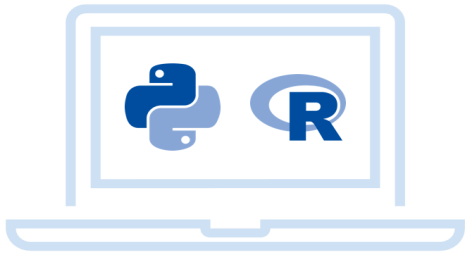


Lung Cancer Differential Expression - Gene vs Normal

gene_name	correlation
PTPRF	0.7213568
MDK	0.7185268
ADCY6	0.7088722
MMP15	0.7009866
TMEM98	0.6950555
BCAR1	0.6925772
SHROOM3	0.6871723
TSPAN6	0.6805749
EFNB2	0.6798894
VWA1	0.6788953

OmicSoft Summary

Extensive Database of Disease Relevant Omics Data



Gene expression

Mutation status

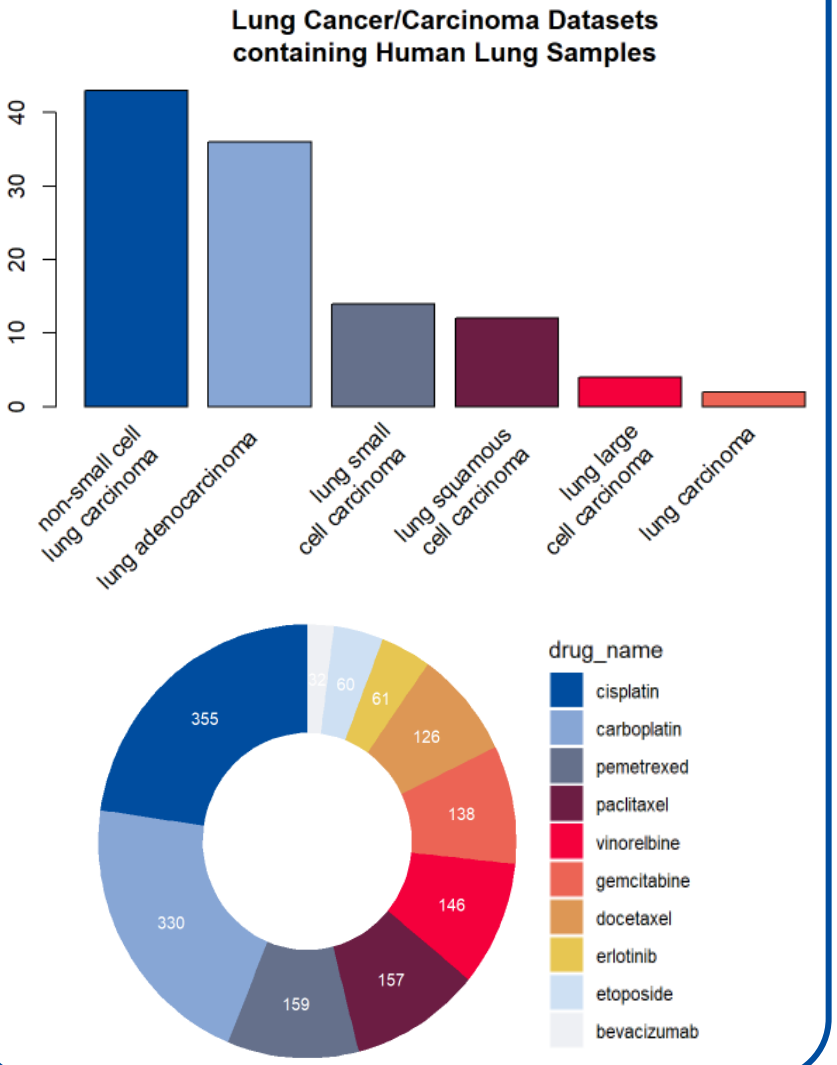
Drug treatment

Cell/ tissue types

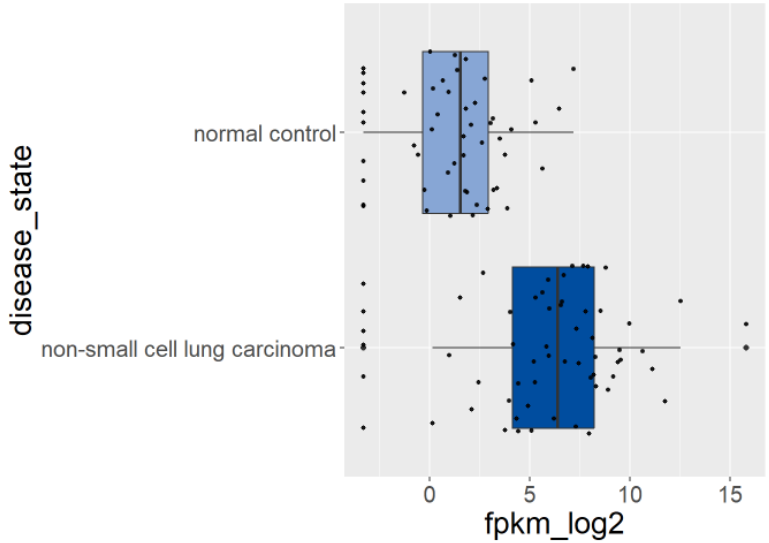
Patient metadata

Experimental data

Easy API Access to Datasets



Pre-computed Multi-dataset Analysis



Please get in touch

We are happy to provide training and participate in collaborative projects