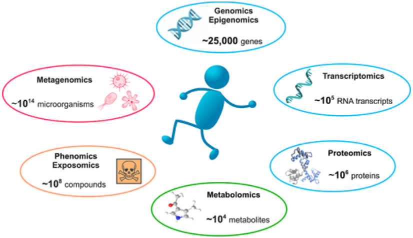
SpideR

R Package for Search, Integration and Retrieval of Big Data

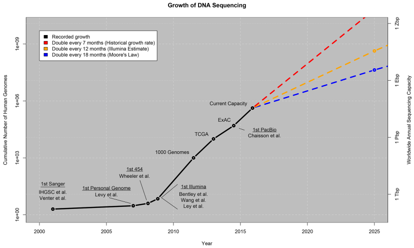
Anna M. Sozanska

Supervisors: Dr Shamith Samarajiwa, Dr Dora Bihary
MRC Cancer Unit
University of Cambridge

Biological datasets are rapidly growing in number

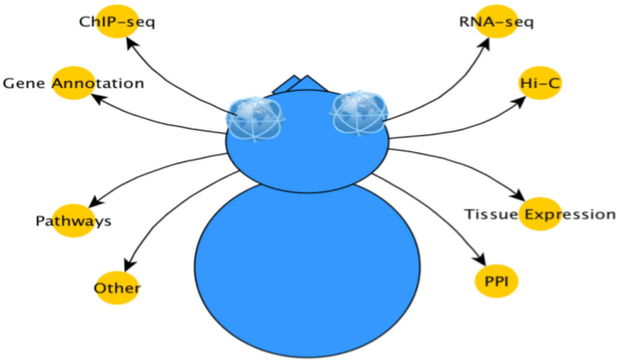


(Gligorijević et al, 2016)



(Stephens et al, 2015)

SpideR: Tool for integrated management of biological datasets

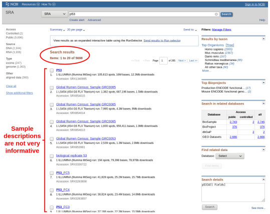


SpideR takes **minutes** to output data that it would take **months** to collect manually

Task: High-throughput bioinformatics analysis

Collect *all** ChIP-seq and RNA-seq samples for **p53** from public databases

- Identify special category samples ('inputs' for ChIP-seq)
- Integrate information from multiple databases (SRA and GEO)



NCBI Resources How To Sign in to NCBI

SRA | SRA | p53 | Search

Create alert Advanced Help

Access
Controlled (2)
Public (5,696)

Source
DNA (2,344)
RNA (3,306)

Type
exome (247)
genome (1,063)

Other
aligned data (360)

Clear all
Show additional filters

Summary 20 per page Send to:

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results
Items: 1 to 20 of 5698

<< First < Prev Page 1 of 285 Next > Last >>

1. P53
1 ILLUMINA (Illumina MiSeq) run: 105,613 spots, 16M bases, 12.3Mb downloads
Accession: SRX1340995

2. Global Rumen Census: Sample GRC0085
1 LS454 (454 GS FLX Titanium) run: 1,382 spots, 687,196 bases, 1.5Mb downloads
Accession: SRX854015

3. Global Rumen Census: Sample GRC0085
1 LS454 (454 GS FLX Titanium) run: 7,995 spots, 4.3M bases, 9Mb downloads
Accession: SRX854014

4. Global Rumen Census: Sample GRC0085
1 LS454 (454 GS FLX Titanium) run: 1,830 spots, 956,411 bases, 1.9Mb downloads
Accession: SRX854013

5. Global Rumen Census: Sample GRC0053
1 LS454 (454 GS FLX Titanium) run: 2,539 spots, 1.3M bases, 2.8Mb downloads
Accession: SRX853991

6. biological replicate 53
1 ILLUMINA (Illumina MiSeq) run: 194 spots, 79,398 bases, 78,978b downloads
Accession: SRX3330722

7. P53_FC5
1 ILLUMINA (Illumina MiSeq) run: 41,828 spots, 25.2M bases, 15.7Mb downloads
Accession: SRX3263658

8. P53_FC4
1 ILLUMINA (Illumina MiSeq) run: 38,824 spots, 23.4M bases, 14.8Mb downloads
Accession: SRX3263657

9. P53_FC3
1 ILLUMINA (Illumina MiSeq) run: 37,188 spots, 22.3M bases, 15.5Mb downloads

Filters: Manage Filters

Results by taxon
Top Organisms [Tree]
Homo sapiens (2955)
Mus musculus (2367)
Danio rerio (167)
Schmidtea mediterranea (85)
Rattus norvegicus (34)
All other taxa (90)
More...

Top Bioprojects
Production ENCODE functional... (17)
Mouse ENCODE functional geno... (2)

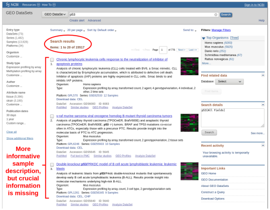
Search in related databases

Database	Access		all
	public	controlled	
BioSample	2,743	2	2,745
BioProject	376		376
dbGaP		2	2
GEO Datasets	2,666		2,666

Find related data
Database: Select
Find items

Search details
p53[All Fields]
Search See more...

Sample descriptions are not very informative



NCBI Resources How To Sign in to NCBI

GEO DataSets | GEO DataSe | p53 | Search

Create alert Advanced Help

Entry type
DataSets (73)
Series (1,482)
Samples (13,928)
Platforms (34)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (5,386)
strain (3,180)
Customize ...

Publication dates
30 days
1 year
Custom range...

Clear all
Show additional filters

Summary 20 per page Sort by Default order Send to:

Search results
Items: 1 to 20 of 15517

<< First < Prev Page 1 of 776 Next > Last >>

1. Chronic lymphocytic leukemia cells response to the neutralization of inhibitor of apoptosis proteins
Analysis of chronic lymphocytic leukemia (CLL) cells treated with BV6, a Smac mimetic. CLL is characterized by B-lymphocyte accumulation, which is attributed to defective cell death. Inhibitor of apoptosis (IAP) proteins are highly expressed in CLL cells. Smac binds to and inhibits IAP proteins.
Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 2 agent, 4 genotype/variation, 4 individual, 2 other, 2 time sets
Platform: GPL570 Series: GSE62533 12 Samples
Download data: CEL
DataSet Accession: GDS6083 ID: 6083
PubMed Similar studies GEO Profiles Analyze DataSet

2. v-raf murine sarcoma viral oncogene homolog B-mutant thyroid carcinoma tumors
Analysis of papillary thyroid carcinoma (TPOCreER; BrafV600E) and anaplastic thyroid carcinoma (TPOCreER; BrafV600E; **p53** -/-) tumors. BRAF and TP53 mutations co-occur often in ATCs, especially those with a precursor PTC. Results provide insight into the molecular basis of PTC to ATC progression.
Organism: Mus musculus
Type: Expression profiling by array, transformed count, 2 genotype/variation, 2 tissue sets
Platform: GPL6246 Series: GSE55933 10 Samples
Download data: CEL
DataSet Accession: GDS5645 ID: 5645
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

3. Double-knockout **p53**/PRKDC model of B cell acute lymphoblastic leukemia: leukemic blasts
Analysis of leukemic blasts from **p53**/Prkdc double-knockout mutants that spontaneously develop early B cell acute lymphoblastic leukemia (B-ALL). Results provide insight into molecular mechanisms underlying high-risk B-ALL.
Organism: Mus musculus
Type: Expression profiling by array, count, 3 cell type, 2 genotype/variation sets
Platform: GPL1261 Series: GSE56345 9 Samples
Download data: CEL, CHP
DataSet Accession: GDS5640 ID: 5640
PubMed Similar studies GEO Profiles Analyze DataSet

Filters: Manage Filters

Top Organisms [Tree]
Homo sapiens (9293)
Mus musculus (5825)
Danio rerio (251)
Schmidtea mediterranea (67)
Rattus norvegicus (61)
More...

Find related data
Database: Select
Find items

Search details
p53[All Fields]
Search See more...

Recent activity
Your browsing activity is temporarily unavailable.

Important Links
GEO Home
GEO Documentation
About GEO DataSets
Construct a Query
Download Options

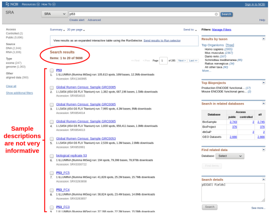
More informative sample description, but crucial information is missing

SpideR takes **minutes** to output data that it would take **months** to collect manually

Task: High-throughput bioinformatics analysis

Collect *all** ChIP-seq and RNA-seq samples
for **A FEW HUNDRED FACTORS** from public databases

- Identify 'inputs' for ChIP-seq
- Integrate information from multiple databases (SRA and GEO)



Access
Controlled (2)
Public (5,696)

Source
DNA (2,344)
RNA (3,306)

Type
exome (247)
genome (1,063)

Other
aligned data (360)

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results
Items: 1 to 20 of 5698

1. P53
   1 ILLUMINA (Illumina MiSeq) run: 105,613 spots, 16M bases, 12.3Mb downloads
   Accession: SRX1340995
2. Global Rumen Census: Sample GRC0085
   1 LS454 (454 GS FLX Titanium) run: 1,382 spots, 687,196 bases, 1.5Mb downloads
   Accession: SRX854015
3. Global Rumen Census: Sample GRC0085
   1 LS454 (454 GS FLX Titanium) run: 7,995 spots, 4.3M bases, 9Mb downloads
   Accession: SRX854014
4. Global Rumen Census: Sample GRC0085
   1 LS454 (454 GS FLX Titanium) run: 1,830 spots, 956,411 bases, 1.9Mb downloads
   Accession: SRX854013
5. Global Rumen Census: Sample GRC0053
   1 LS454 (454 GS FLX Titanium) run: 2,539 spots, 1.3M bases, 2.8Mb downloads
   Accession: SRX853991
6. biological replicate 53
   1 ILLUMINA (Illumina MiSeq) run: 194 spots, 79,398 bases, 78,978b downloads
   Accession: SRX3330722
7. P53_FC5
   1 ILLUMINA (Illumina MiSeq) run: 41,828 spots, 25.2M bases, 15.7Mb downloads
   Accession: SRX3263658
8. P53_FC4
   1 ILLUMINA (Illumina MiSeq) run: 38,824 spots, 23.4M bases, 14.8Mb downloads
   Accession: SRX3263657
9. P53_FC3
   1 ILLUMINA (Illumina MiSeq) run: 37,188 spots, 22.3M bases, 15.5Mb downloads

Filters: Manage Filters

Results by taxon
Top Organisms [Tree]
Homo sapiens (2955)
Mus musculus (2367)
Danio rerio (267)
Schmidtea mediterranea (85)
Rattus norvegicus (34)
All other taxa (90)
More...

Top Bioprojects
Production ENCODE functional... (17)
Mouse ENCODE functional geno... (2)

Search in related databases

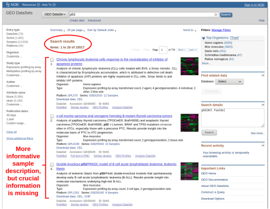
Database	Access		all
	public	controlled	
BioSample	2,743	2	2,745
BioProject	376		376
dbGaP		2	2
GEO Datasets	2,666		2,666

Find related data
Database: Select

Search details

```
p53[All Fields]
```

Sample descriptions are not very informative



Entry type
DataSets (73)
Series (1,482)
Samples (13,928)
Platforms (34)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (5,386)
strain (3,180)
Customize ...

Publication dates
30 days
1 year
Custom range...

Search results
Items: 1 to 20 of 15517

1. Chronic lymphocytic leukemia cells response to the neutralization of inhibitor of apoptosis proteins
   Analysis of chronic lymphocytic leukemia (CLL) cells treated with BV6, a Smac mimetic. CLL is characterized by B-lymphocyte accumulation, which is attributed to defective cell death. Inhibitor of apoptosis (IAP) proteins are highly expressed in CLL cells. Smac binds to and inhibits IAP proteins.
   Organism: Homo sapiens
   Type: Expression profiling by array, transformed count, 2 agent, 4 genotype/variation, 4 individual, 2 other, 2 time sets
   Platform: GPL570 Series: GSE62533 12 Samples
   Download data: CEL
   DataSet Accession: GDS6083 ID: 6083
   PubMed Similar studies GEO Profiles Analyze DataSet
2. v-raf murine sarcoma viral oncogene homolog B-mutant thyroid carcinoma tumors
   Analysis of papillary thyroid carcinoma (TPOCreER; BrafV600E) and anaplastic thyroid carcinoma (TPOCreER; BrafV600E; **p53** -/-) tumors. BRAF and TP53 mutations co-occur often in ATCs, especially those with a precursor PTC. Results provide insight into the molecular basis of PTC to ATC progression.
   Organism: Mus musculus
   Type: Expression profiling by array, transformed count, 2 genotype/variation, 2 tissue sets
   Platform: GPL6246 Series: GSE55933 10 Samples
   Download data: CEL
   DataSet Accession: GDS5645 ID: 5645
   PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet
3. Double-knockout **p53**/PRKDC model of B cell acute lymphoblastic leukemia: leukemic blasts
   Analysis of leukemic blasts from **p53**/Prkdc double-knockout mutants that spontaneously develop early B cell acute lymphoblastic leukemia (B-ALL). Results provide insight into molecular mechanisms underlying high-risk B-ALL.
   Organism: Mus musculus
   Type: Expression profiling by array, count, 3 cell type, 2 genotype/variation sets
   Platform: GPL1261 Series: GSE56345 9 Samples
   Download data: CEL, CHP
   DataSet Accession: GDS5640 ID: 5640
   PubMed Similar studies GEO Profiles Analyze DataSet

Filters: Manage Filters

Top Organisms [Tree]
Homo sapiens (8293)
Mus musculus (5825)
Danio rerio (251)
Schmidtea mediterranea (67)
Rattus norvegicus (61)
More...

Find related data
Database: Select

Search details

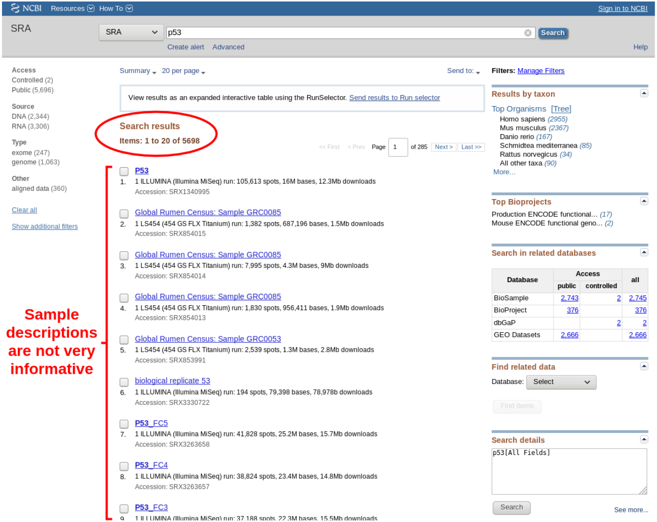
```
p53[All Fields]
```

Recent activity
Your browsing activity is temporarily unavailable.

Important Links
GEO Home
GEO Documentation
About GEO DataSets
Construct a Query
Download Options

More informative sample description, but crucial information is missing

SRA Database: Search results



Access
Controlled (2)
Public (5,696)

Source
DNA (2,344)
RNA (3,306)

Type
exome (247)
genome (1,063)

Other
aligned data (360)

Clear all

Show additional filters

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results

Items: 1 to 20 of 5698

1. □ P53
1 ILLUMINA (Illumina MiSeq) run: 105,613 spots, 16M bases, 12.3Mb downloads
Accession: SRX1340995

2. □ Global Rumen Census: Sample GRC0085
1 LS454 (454 GS FLX Titanium) run: 1,382 spots, 687,196 bases, 1.5Mb downloads
Accession: SRX854015

3. □ Global Rumen Census: Sample GRC0085
1 LS454 (454 GS FLX Titanium) run: 7,995 spots, 4.3M bases, 9Mb downloads
Accession: SRX854014

4. □ Global Rumen Census: Sample GRC0085
1 LS454 (454 GS FLX Titanium) run: 1,830 spots, 956,411 bases, 1.9Mb downloads
Accession: SRX854013

5. □ Global Rumen Census: Sample GRC0053
1 LS454 (454 GS FLX Titanium) run: 2,539 spots, 1.3M bases, 2.8Mb downloads
Accession: SRX853991

6. □ biological replicate 53
1 ILLUMINA (Illumina MiSeq) run: 194 spots, 79,398 bases, 78,978b downloads
Accession: SRX3330722

7. □ P53_FC5
1 ILLUMINA (Illumina MiSeq) run: 41,828 spots, 25.2M bases, 15.7Mb downloads
Accession: SRX3263658

8. □ P53_FC4
1 ILLUMINA (Illumina MiSeq) run: 38,824 spots, 23.4M bases, 14.8Mb downloads
Accession: SRX3263657

9. □ P53_FC3
1 ILLUMINA (Illumina MiSeq) run: 37,188 spots, 22.3M bases, 15.5Mb downloads

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens (2955)
Mus musculus (2367)
Danio rerio (167)
Schmidtea mediterranea (85)
Rattus norvegicus (34)
All other taxa (90)
More...

Top Bioprojects

Production ENCODE functional... (17)
Mouse ENCODE functional geno... (2)

Search in related databases

Database	Access		all
	public	controlled	
BioSample	2,743	2	2,745
BioProject	376		376
dbGaP		2	2
GEO Datasets	2,666		2,666

Find related data

Database: Select

Find items

Search details

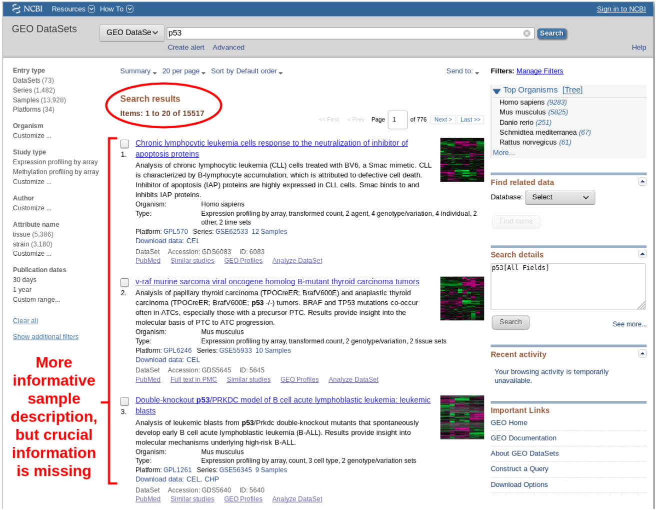
```
p53[All Fields]
```

Search

See more...

Sample descriptions are not very informative

GEO Database: Search results



NCBI Resources How To Sign in to NCBI

GEO DataSets GEO DataSe p53 Search

Create alert Advanced Help

Entry type
DataSets (73)
Series (1,482)
Samples (13,928)
Platforms (34)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (5,386)
strain (3,180)
Customize ...

Publication dates
30 days
1 year
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

Search results

Items: 1 to 20 of 15517

<< First < Prev Page 1 of 776 Next > Last >>

□ 1. Chronic lymphocytic leukemia cells response to the neutralization of inhibitor of apoptosis proteins

Analysis of chronic lymphocytic leukemia (CLL) cells treated with BV6, a Smac mimetic. CLL is characterized by B-lymphocyte accumulation, which is attributed to defective cell death. Inhibitor of apoptosis (IAP) proteins are highly expressed in CLL cells. Smac binds to and inhibits IAP proteins.

Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 2 agent, 4 genotype/variation, 4 individual, 2 other, 2 time sets
Platform: GPL570 Series: GSE62533 12 Samples
Download data: CEL
DataSet Accession: GDS6083 ID: 6083
PubMed Similar studies GEO Profiles Analyze DataSet

□ 2. v-raf murine sarcoma viral oncogene homolog B-mutant thyroid carcinoma tumors

Analysis of papillary thyroid carcinoma (TPOCreER; BrafV600E) and anaplastic thyroid carcinoma (TPOCreER; BrafV600E; **p53** -/-) tumors. BRAF and TP53 mutations co-occur often in ATCs, especially those with a precursor PTC. Results provide insight into the molecular basis of PTC to ATC progression.

Organism: Mus musculus
Type: Expression profiling by array, transformed count, 2 genotype/variation, 2 tissue sets
Platform: GPL6246 Series: GSE55933 10 Samples
Download data: CEL
DataSet Accession: GDS5645 ID: 5645
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

□ 3. Double-knockout **p53**/PRKDC model of B cell acute lymphoblastic leukemia: leukemic blasts

Analysis of leukemic blasts from **p53**/Prkdc double-knockout mutants that spontaneously develop early B cell acute lymphoblastic leukemia (B-ALL). Results provide insight into molecular mechanisms underlying high-risk B-ALL.

Organism: Mus musculus
Type: Expression profiling by array, count, 3 cell type, 2 genotype/variation sets
Platform: GPL1261 Series: GSE56345 9 Samples
Download data: CEL, CHP
DataSet Accession: GDS5640 ID: 5640
PubMed Similar studies GEO Profiles Analyze DataSet

Filters: Manage Filters

Top Organisms [Tree]
Homo sapiens (9283)
Mus musculus (5825)
Danio rerio (251)
Schmidtea mediterranea (67)
Rattus norvegicus (61)
More...

Find related data
Database: Select
Find items

Search details
p53[All Fields]
Search See more...

Recent activity
Your browsing activity is temporarily unavailable.

Important Links
GEO Home
GEO Documentation
About GEO DataSets
Construct a Query
Download Options

More informative sample description, but crucial information is missing

GEO and SRA use different accession hierarchies

SRA Name	SRA Acc.	SRA ↔ GEO	GEO Acc.	GEO Name
🟩 Project	🟥 SRP	$\Longleftrightarrow$	🟥 GSE	🟩 Series
🟩 **Sample!**	🟥 SRS		🟥	🟩
🟩 Exp.	🟥 SRX	$\Longleftrightarrow$	🟥 GSM	🟩 **Sample!**
🟩 Run	🟥 SRR		🟥	🟩

Database Accession Numbers

SRA #	SRA	SRA → GEO #	GEO	GEO #
98 028	SRP	17 300	GSE	80 782
2 974 554	SRS	396 513		
2 377 047	SRX	400 975	GSM	2 231 166
3 366 463	SRR	510 890		

Main SpideR Functions

searchForTerm

- IN:
  - library_strategy [R]
  - gene [*]
  - antibody [*]
  - cell_type [*]
  - treatment [*]
  - species [*]
  - platform
  - secondary_lib._strat.
- OUT:
  files for db & pipeline

For example:

Find human HiC data

Find ChIP-seq data
with STAT1 antibody

searchForAccession

- IN: accession list
- OUT: df or files for db & pipeline

For example:

Find runs within SRP052871

Find runs within GSE34715

Other SpideR Functions

superseriesFinder

- IN: GSM list
  List of samples of interest
- OUT: GSM list
  List of all other samples within the same superseries as the samples of interest

convertAccession

- IN: accession list
  In **one** accession format
- OUT: accession list
  In **all** accession formats

inputDetector

Labels inputs in a sample sheet

controlDetector

Labels controls in a sample sheet

dbExtractGenerator

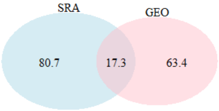
Generates db extract from df

sampleSheetGenerator

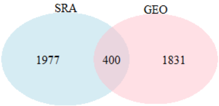
Generates sample sheet from df

SpideR uses a **custom database** to solve the non-trivial problem of SRA/GEO mapping

Projects (SRP/GSE)



Samples (SRX/GSM)



All numbers in 1000s of entries.

- No 1:1 mapping of samples
- Different accession hierarchy
- Different information in each database
- Mapping information scattered across different columns of the database

Exploring hidden sample hierarchy can help find related experiments

Task

Find related ChIP-seq and RNA-seq experiments

Case 1
Relationship easy to establish

SRP#1

- A (ChIP)
- B (RNA)
- C (RNA)

GSE#1

- A (ChIP)
- B (RNA)
- C (RNA)

Case 2
Relationship difficult to establish

SRP#1

- A (ChIP)

SRP#2

- B (RNA)
- C (RNA)

GSE#1

- A (ChIP)

GSE#2

- B (RNA)
- C (RNA)

Superseries?

Overcoming superseries challenge can help find hidden hierarchies

Format in the database

GSE	GSM
GSE1, GSE3	A
GSE1, GSE3	B
GSE1, GSE3	C
GSE2, GSE3	**D**
GSE2, GSE3	**E**
GSE2, GSE3	**F**
GSE2, GSE3	**G**

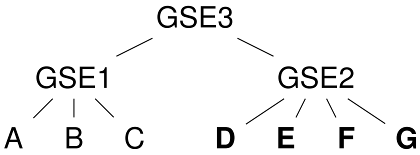
Approach

Given a list of samples (GSMs), get all the samples that belong to the same GSEs (excluding the original samples).

Overcoming superseries challenge can help find hidden hierarchies

Format in the database

GSE	GSM
GSE1, GSE3	A
GSE1, GSE3	B
GSE1, GSE3	C
GSE2, GSE3	**D**
GSE2, GSE3	**E**
GSE2, GSE3	**F**
GSE2, GSE3	**G**



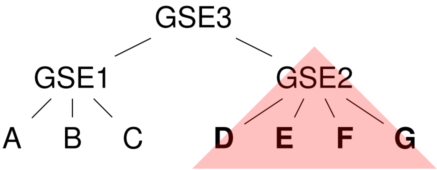
Approach

Given a list of samples (GSMs), get all the samples that belong to the same GSEs (excluding the original samples).

Overcoming superseries challenge can help find hidden hierarchies

Format in the database

GSE	GSM
GSE1, GSE3	A
GSE1, GSE3	B
GSE1, GSE3	C
GSE2, GSE3	**D**
GSE2, GSE3	**E**
GSE2, GSE3	**F**
GSE2, GSE3	**G**



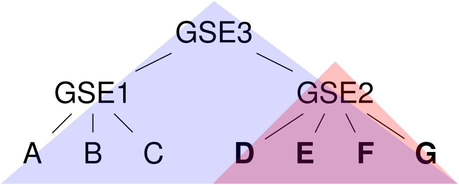
Approach

Given a list of samples (GSMs), get all the samples that belong to the same GSEs (excluding the original samples).

Overcoming superseries challenge can help find hidden hierarchies

Format in the database

GSE	GSM
GSE1, GSE3	A
GSE1, GSE3	B
GSE1, GSE3	C
GSE2, GSE3	**D**
GSE2, GSE3	**E**
GSE2, GSE3	**F**
GSE2, GSE3	**G**



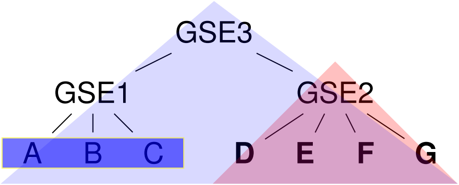
Approach

Given a list of samples (GSMs), get all the samples that belong to the same GSEs (excluding the original samples).

Overcoming superseries challenge can help find hidden hierarchies

Format in the database

GSE	GSM
GSE1, GSE3	A
GSE1, GSE3	B
GSE1, GSE3	C
GSE2, GSE3	**D**
GSE2, GSE3	**E**
GSE2, GSE3	**F**
GSE2, GSE3	**G**



Approach

Given a list of samples (GSMs), get all the samples that belong to the same GSEs (excluding the original samples).

Other SpideR Features

SpideR tackles inconsistencies in categories

For example, in sample_attribute field which contains comma-separated-categories.
E.g. tissue, source name, cell, cell_type, cell type - same/similar???

SpideR provides easily manipulable output and is easily **reproducible**

.*Rda* outputs with **data frames**
.*tab* outputs with **data frames**
.*Rda* output with the **function call** .*tab* output with function **call parameters**

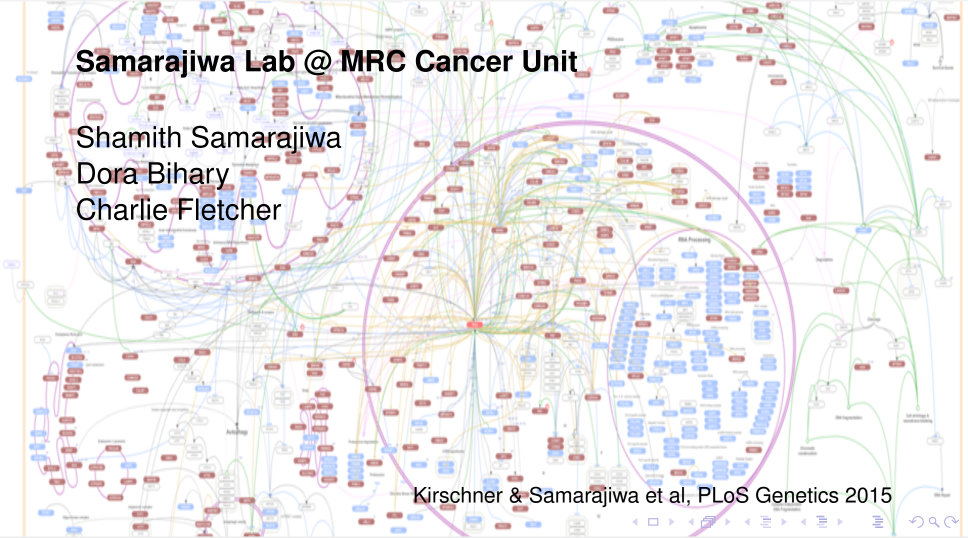
SpideR addresses most challenges with database design and integration

☑ Misleading fields in SRAdb and GEOmetadb
☑ Inconsistent attributes or categories
☑ Superseries in GEO
☑ SRA-GEO conversion
□ Erroneous or inconsistent entries

Plans for future development

- A new function for searching everywhere in the database (to get a list of all *potentially matching* samples, including those rejected by the more specific **searchForTerm** function)
- Integration of *elasticsearchr*
- Linking the search functions to gene synonyms database
- SQL(ite) database for storing samples of interest

Acknowledgements



Samarajiwa Lab @ MRC Cancer Unit

Shamith Samarajiwa
Dora Bihary
Charlie Fletcher

Kirschner & Samarajiwa et al, PLoS Genetics 2015