

Data acquisition and data management

Sophia Derdak

Core Facilities MedUni Vienna

sophia.derdak@meduniwien.ac.at

We are going to use the interactive Q&A tool Mentimeter, please connect here
(from your phone or computer):

<https://www.menti.com>

[CODE:]

[GABIO_20240416_Data , 2 slides]



Before data analysis:



```
protein_coding Hrp115 "ENSMUSG0000033845.13" -0.0433739881207122 0.648139499180132 1 0
protein_coding Lyp1a1 "ENSMUSG0000025903.14" 0.107345588324394 0.341339394658878 1 0
protein_coding Tesc1 "ENSMUSG0000033913.15" 0.037028474292025 0.747700373418524 1 0
protein_coding Rgs20 "ENSMUSG0000002459.17" -0.928772644993741 0.384727861489713 1 0
protein_coding Aqp7v1 "ENSMUSG0000033793.12" 0.08097418454801162 0.32894109187878 1 0
protein_coding Oprk1 "ENSMUSG0000025905.14" 0.18349485111964 0.736811957745655 1 0
protein_coding Hspwr1 "ENSMUSG0000033774.4" -0.0836076788852 0.9282279641927 1 0
protein_coding Alk1c1 "ENSMUSG0000025907.14" -0.0063087426249416 0.994994942607246 1 0
protein_coding Alkal1 "ENSMUSG0000087247.3" 1.3551703524243 0.16898492782883 1 0
protein_coding Slit1 "ENSMUSG0000033740.17" -0.3723208849017 0.46334516339031 1 0
protein_coding Pcnrd1 "ENSMUSG00000051285.17" 0.0145620023281159 0.825147948234228 1 0
protein_coding Slit2 "ENSMUSG0000025909.10" 0.092382122828461 0.532823146694837 1 0
protein_coding Rrs1 "ENSMUSG0000001824.8" -0.23089274107882 0.157995268880137 1 0
protein_coding Adh1c1 "ENSMUSG0000025911.14" 0.22173732950991 0.3903366502047 1 0
protein_coding Von "ENSMUSG0000007879.3" -2.08173157681248 0.16175149154339 1 0
protein_coding Mybl1 "ENSMUSG0000025912.10" -0.209318597225185 0.47988893581395 1 0
protein_coding Vcpd1 "ENSMUSG0000004519.8" 0.013735851408464 0.70235145158066 1 0
protein_coding Spk3 "ENSMUSG0000025915.14" -0.0268293684441533 0.98748213844382 1 0
protein_coding Hmnc2 "ENSMUSG0000044501.10" 1.09339938398366 0.333765888878849 1 0
protein_coding Tcf24 "ENSMUSG0000009983.2" 0.319798298179425 0.39019737263876 1 0
protein_coding Ppp1r42 "ENSMUSG0000025916.10" 1.70644473810765 0.55021116811824 1 0
protein_coding Cspg5 "ENSMUSG0000007517.4" -0.0089728884679356 0.988192485878885 1 0
protein_coding Cspg1 "ENSMUSG00000056763.10" -0.19685847562464 0.0848846248808655 1 0
protein_coding Aif1p3 "ENSMUSG0000007851.11" 0.0024366895318727 0.92833732480865 1 0
protein_coding Cpa8 "ENSMUSG0000042501.12" -0.136718641499179 0.598676994873466 1 0
protein_coding Pfrk2 "ENSMUSG0000049369.13" 0.14409321692768 0.45103160184881 1 0
protein_coding ABR018L16R1K "ENSMUSG00000057715.13" 1.29683101094506 0.0549836680539968 1 0
protein_coding Solt1 "ENSMUSG0000016918.15" -0.08664367131271577 0.974227587209189 1 0
protein_coding Slco3a1 "ENSMUSG0000025918.10" 0.6075093360215 0.247828797093572 1 0
```

Shutterstock thumbnails

Curate lab notes and acquired data into a structured format that is computer-readable.

Before data analysis:



Gene	ENSG	chr	start	end	score	strand	type	count	norm
proteincoding	Hrpl15	"ENSMUSG0000033845"	13	-0.0433739881207122	0.648139499100132	1	0		
proteincoding	Lyp1a1	"ENSMUSG0000025903"	14	-0.107345588324394	0.341339394508078	1	0		
proteincoding	Ten1	"ENSMUSG0000033913"	15	-0.037028476290225	0.747700373762524	1	0		
proteincoding	Rgs20	"ENSMUSG0000002459"	17	-0.928772644993741	0.384727661507713	1	0		
proteincoding	Qpmyh	"ENSMUSG0000033793"	12	-0.0809741645001162	0.93944109187878	1	0		
proteincoding	Oprk1	"ENSMUSG0000025905"	14	-0.18349485511964	0.736011957745555	1	0		
proteincoding	Habr1	"ENSMUSG0000033774"	4	-0.0836076788652	0.9262279641927	1	0		
proteincoding	Alct1	"ENSMUSG0000025907"	14	-0.0006309742549491	0.99499492497046	1	0		
proteincoding	Atkal1	"ENSMUSG0000008724"	3	1.3551703524243	0.16898492762083	1	0		
proteincoding	S12	"ENSMUSG0000033740"	17	-0.37232088490917	0.46334516330031	1	0		
proteincoding	Pcncd1	"ENSMUSG0000005128"	17	-0.0145620023201159	0.825147948234228	1	0		
proteincoding	S12	"ENSMUSG0000025909"	10	-0.09238222320461	0.530232146694637	1	0		
proteincoding	Rrs1	"ENSMUSG0000001024"	8	-0.23089274107082	0.15799526880137	1	0		
proteincoding	Adh1c1	"ENSMUSG0000025911"	14	-0.22173732295991	0.3903366502047	1	0		
proteincoding	Von	"ENSMUSG0000007079"	3	-2.981371557081248	0.1617549154339	1	0		
proteincoding	Mybl1	"ENSMUSG0000025912"	10	-0.209318597225185	0.47988993581395	1	0		
proteincoding	Vcpd1	"ENSMUSG0000004510"	8	-0.0137345813490454	0.70231541530606	1	0		
proteincoding	Sgk3	"ENSMUSG0000025915"	14	-0.026829360441533	0.907482138443482	1	0		
proteincoding	Hmc2	"ENSMUSG0000004401"	10	1.0933938398956	0.32376508078849	1	0		
proteincoding	Tcf24	"ENSMUSG0000009983"	2	-0.319798298179425	0.39019737263076	1	0		
proteincoding	Ppp1r42	"ENSMUSG0000025916"	10	1.7064474810705	0.55021156815024	1	0		
proteincoding	Cops5	"ENSMUSG0000025917"	4	-0.0089728864507956	0.9801928507084	1	0		
proteincoding	Cspol1	"ENSMUSG0000005270"	10	-0.196858475682464	0.0848862628036655	1	0		
proteincoding	Rifp2	"ENSMUSG0000007051"	11	-0.0024366095310727	0.92833733600065	1	0		
proteincoding	Cpa8	"ENSMUSG0000004201"	12	-0.136718641499179	0.59867098487346	1	0		
proteincoding	Rfnd2	"ENSMUSG0000004009"	13	1.40091521092708	0.4510316104005	1	0		
proteincoding	ABR018L16R1K	"ENSMUSG0000005771"	13	1.29683101094506	0.0549836600539960	1	0		
proteincoding	Solt1	"ENSMUSG0000016918"	15	-0.08664367131217577	0.974227507209109	1	0		
proteincoding	Ucp5a1	"ENSMUSG0000025918"	10	-0.60750993606215	0.247082794932572	1	0		

Shutterstock thumbnails

Curate lab notes and acquired data into a structured format that is computer-readable.

After data acquisition and analysis:

European Nucleotide Archive

ARTICnetwork

National Library of Medicine
National Center for Biotechnology Information

ProteomeXchange

EUROPEAN GENOME-PHENOME ARCHIVE

NATIONAL CANCER INSTITUTE
GDC Data Submission Portal

Gene Expression Omnibus

FLOW Repository

Publication and FAIR sharing of well-annotated data with the scientific community.

Outline

0. Before the experiment: Collect samples and sample metadata.
1. What data do I have?
2. What information do I want to extract?
3. How to search for software tools that suit my needs?
4. How do I know whether a software tool is a good choice - before even downloading it?
5. How to make my data “AI-compatible”.
6. After data analysis: Publication and FAIR sharing of data.

o. Before the experiment: Collect samples and sample metadata.



What is your area of research?

<https://www.menti.com>

[CODE:]

[GABIO_Data_20240416]

o. Before the experiment: Collect samples and sample metadata.



Which technologies do you use to generate data?

<https://www.menti.com>

[CODE:]

[GABIO_Data_20240416]

o. Before the experiment: Collect samples and sample metadata.



Use one sample identifier throughout all steps of the experiment.

Use the same metadata grid for all samples in the experiment.

Label your samples carefully and in a meaningful and systematic way.

o. Before the experiment: Collect samples and sample metadata.



- Which samples do I have? Do I have several sample groups (experimental conditions...)?
- What are the parameters I will measure for my samples?
- Which instrument(s) will be used for the measurements?

o. Before the experiment: Collect samples and sample metadata.

	A	B	C	D	E	F	G	H	I	J
1	sampleDescription	sample_ID	samplegroup	clinicalParameterA	clinicalParameterB	clinicalParameterC	measurementA	measurementB	measurementC	file from instrumentA
2	normal sample	SD01001	normal	0.60	0.19	0.90	493	1	NA	
3	external sample #12@1497AB	SD01002	normal	0.87	0.83	0.34	914	0	NA	
4	new sample	SD01003	treatment	0.87	0.87	0.43	862	1	9.2	
5	SAMPLE5	SD01004	treatment	0.91	0.29	0.47	120	1	1.1	
6	control sample	SD01005	treatment	0.22	0.94	0.07	852	1	9.7	
7		SD01006	normal	0.62	0.52	0.86	577	0	6.0	
8										
9										

existing data

missing data

free text
sample
description
(incl. different
colours, fonts)

structured
sample
identifier

structured
samplegroup
label

1. What data do I have?



Know which files are generated, how big these files are and how to transfer and store them.

Keep raw data well-labelled for long-term storage or later publication.

1. What data do I have?



- Which files are generated on the acquisition instrument computer for each sample?
- Where are these files generated on the instrument computer?
- Are the data in a standard format or in tabular format?
- How big are these data in terms of computer storage?
- For how many samples do I usually generate data in an experiment?

1. What data do I have?

Example: Illumina sequencing data acquisition for genomic variant analysis

- Illumina format raw sequencing output per flowcell
- “sampleSheet.csv” in Illumina format



Illumina NextSeq2000 sequencing system

<https://cna.uga.edu>

1. What data do I have?

- Illumina format raw sequencing output per flowcell

Genomics_data > Illumina_NextSeq > NextSeq-Backup > 211004_NB501433_0169_AHTVNMBGXC

ion ▾ Tools ▾ Settings

Name ▾	Size	File Type	Modified Date
Config		Folder	2021-10-04 15:4...
Data		Folder	2021-10-04 19:0...
Images		Folder	2021-10-04 18:1...
InstrumentAnalyticsLogs		Folder	2021-10-05 22:1...
InterOp		Folder	2021-10-05 22:1...
Logs		Folder	2021-10-05 22:1...
Recipe		Folder	2021-10-04 15:4...
RTALogs		Folder	2021-10-05 22:1...
CopyComplete.txt	0 bytes	TXT File	2021-10-05 22:1...
RTAComplete.txt	49 bytes	TXT File	2021-10-05 22:1...
RTAConfiguration.xml	6.1 KB	XML File	2021-10-04 15:5...
RTARead1Complete.txt	36 bytes	TXT File	2021-10-05 05:5...
RTARead2Complete.txt	36 bytes	TXT File	2021-10-05 08:2...
RTARead3Complete.txt	36 bytes	TXT File	2021-10-05 09:1...
RTARead4Complete.txt	37 bytes	TXT File	2021-10-05 22:1...
RunCompletionStatus.xml	927 bytes	XML File	2021-10-05 22:1...
RunInfo.xml	28 KB	XML File	2021-10-04 15:4...
RunParameters.xml	26.3 KB	XML File	2021-10-04 15:4...
SampleSheet.csv	841 bytes	CSV File	2021-10-04 15:4...



Illumina NextSeq2000 sequencing system

<https://cna.uga.edu>

~ 90 GB
Do not extract or delete
subdirectories!

1. What data do I have?

- “sampleSheet.csv” in Illumina format

[Header]

Date,20211006

ExperimentName,Jeitler_20211006

Workflow,GenerateFastQ

[Reads]

150

150

[Settings]

InstrumentType,NextSeq

Adapter,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

[Data]

Sample_ID,Sample_Name,I5_Index_ID,index2,I7_Index_ID,index

MJ04001,MJ04001,506,TAAGATTA,701,ATTACTCG

MJ04002,MJ04002,506,TAAGATTA,702,TCCGGAGA

MJ04003,MJ04003,506,TAAGATTA,703,CGCTCATT

MJ04004,MJ04004,506,TAAGATTA,704,GAGATTCC

MJ04005,MJ04005,506,TAAGATTA,705,ATTCAGAA

MJ04006,MJ04006,506,TAAGATTA,706,GAATTCGT



Illumina NextSeq2000 sequencing system

<https://dna.uga.edu>

2. What information do I want to extract?



Store raw data as backup and for publication.

Convert proprietary formats into portable standard or tabular formats for downstream analysis.

2. What information do I want to extract?

Example: a Core Facility

- Which files should be stored as long-term backup?
- Which files are transferred to the users?
- Which files can be shared with Tech Support?



2. What information do I want to extract?



Example: a Core Facility

- Which files should be stored as long-term backup?
- Which files are transferred to the users?
- Which files can be shared with Tech Support?

Example: Core Facility users ~ researchers

- Which files can I use directly for analysis?
- Which files can I use to extract data formatted for downstream analysis?
- Which files do I need to deposit in public repositories upon publication?

2. What information do I want to extract?

Genomic variants:

.fastq files can be further analysed with numerous community-developed and commercial analysis tools.

In the analysis workflow to detect genomic variants, the sequences in .fastq format are first aligned to the reference genome of the organism of origin:



Linux command-line tools for alignment to the genome: e.g. bwa (<https://bio-bwa.sourceforge.net>)

2. What information do I want to extract?

From the alignments, genomic variants are extracted and stored in **.vcf** = Variant Call Format:

1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

1.1 An example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:..
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

Linux command-line tools for variant calling: e.g. GATK (<https://gatk.broadinstitute.org/hc/en-us>)

3. How to search for software tools that suit my needs?



Be aware of the type of data and scientific question at hand.

Find out what are established software tools and workflows in the community.

When choosing an analysis workflow from a paper, make sure their and my data and research conditions are comparable.

3. How to search for software tools that suit my needs?

- Will I use the software myself or will I have help with data analysis?
- Will the software work on my operating system?
- Is the software license-dependent or open access?
- Is the software compatible with my type of data?
- Was the software developed with a similar research question in mind?
- Take note from publications and talks on what other researchers in the field use, look out for review/benchmarking articles
- Look up repositories of open source software packages

3. How to search for software tools that suit my needs?

- Tools mentioned at conferences and in publications:



- Repositories of open source software packages:



- Institutes that develop many frequently used computational tools:

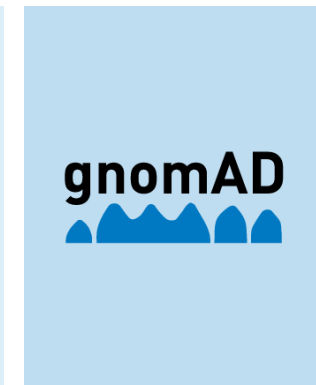
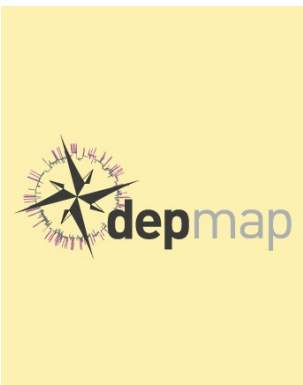


Career
About us ▾ Research ▾ Centers ▾ Education a

HOME

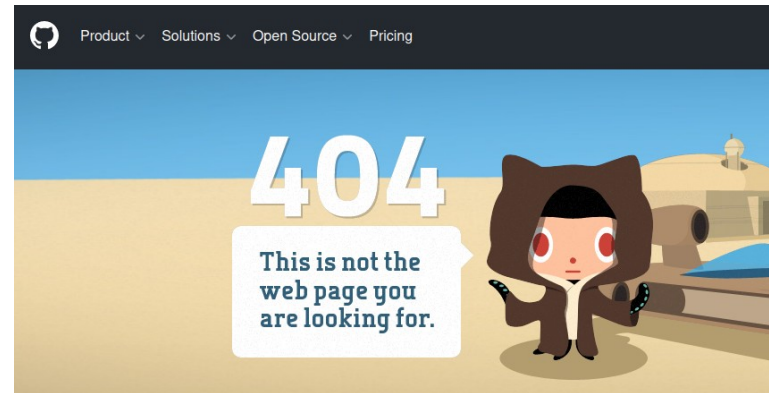
Resources, services, and tools

Key scientific datasets and computational tools developed by our scientists and their collaborators.



...

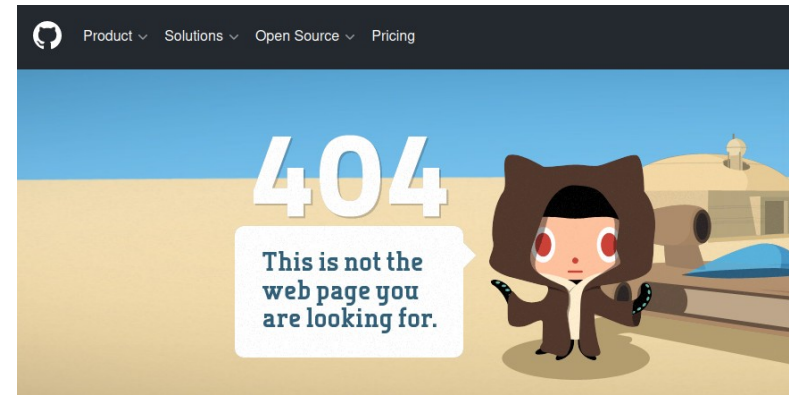
4. How do I know whether a software tool is a good choice - before even downloading it?



Do not waste time with ill-documented or discontinued computational tools.

4. How do I know whether a software tool is a good choice - before even downloading it?

- Is the download page available (for open source tools) when linking from a publication?
- On the download page, when was the latest update?
- Is there documentation and user guide available? Does the user guide include example use cases? Do these use cases fit my needs?
- Find out from the documentation whether the tool will work with the data format I have as an input.
- Is there a way to contact the developers/tech support of the tool by email?



5. How to make my data “AI-compatible”

Sample data and metadata in a spreadsheet:

	A	B	C
1	sample_ID	samplegroup	clinicalParameterA
2	SD01001	normal	0.97
3	SD01002	normal	0.84
4	SD01003	treatment	0.88
5	SD01004	treatment	0.04
6	SD01005	treatment	0.15
7	SD01006	normal	0.15

[. . .]

H	I	J
measurement	.fastq file	.vcf file
NA	/my/filesystem/project1/fastq/SD01001.fastq.gz	/my/filesystem/project1/vcf/SD01001.vcf.gz
NA	/my/filesystem/project1/fastq/SD01002.fastq.gz	/my/filesystem/project1/vcf/SD01002.vcf.gz
7.7	/my/filesystem/project1/fastq/SD01003.fastq.gz	/my/filesystem/project1/vcf/SD01003.vcf.gz
5.1	/my/filesystem/project1/fastq/SD01004.fastq.gz	/my/filesystem/project1/vcf/SD01004.vcf.gz
9.5	/my/filesystem/project1/fastq/SD01005.fastq.gz	/my/filesystem/project1/vcf/SD01005.vcf.gz
2.8	/my/filesystem/project1/fastq/SD01006.fastq.gz	/my/filesystem/project1/vcf/SD01006.vcf.gz

Or in .csv plain text:

```
sample_ID,samplegroup,clinicalParameterA,,,,measurementC,.fastq file,.vcf file
SD01001,normal,0.97,,,,NA,/my/filesystem/project1/fastq/SD01001.fastq.gz,/my/filesystem/project1/vcf/SD01001.vcf.gz
SD01002,normal,0.84,,,,NA,/my/filesystem/project1/fastq/SD01002.fastq.gz,/my/filesystem/project1/vcf/SD01002.vcf.gz
SD01003,treatment,0.88,,,,7.7,/my/filesystem/project1/fastq/SD01003.fastq.gz,/my/filesystem/project1/vcf/SD01003.vcf.gz
SD01004,treatment,0.04,,,,5.1,/my/filesystem/project1/fastq/SD01004.fastq.gz,/my/filesystem/project1/vcf/SD01004.vcf.gz
SD01005,treatment,0.15,,,,9.5,/my/filesystem/project1/fastq/SD01005.fastq.gz,/my/filesystem/project1/vcf/SD01005.vcf.gz
SD01006,normal,0.15,,,,2.8,/my/filesystem/project1/fastq/SD01006.fastq.gz,/my/filesystem/project1/vcf/SD01006.vcf.gz
```



Your analysis here!



6. After data analysis: Publication and FAIR sharing of data.

Everyone can benefit from well-annotated public datasets (not only computational biologists and data analysts!).

6. After data analysis: Publication and FAIR sharing of data.

F indability:

unique dataset identifier, in a searchable resource, with rich metadata

A ccessibility:

retrievable using a standardized communications protocol, when necessary with authentication

I nteroperability

in “standardized” formats, metadata use controlled vocabulary

R eusability

detailed metadata and explicit about authors and data usage license

www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

6. After data analysis: Publication and FAIR sharing of data.

Which public data repositories have you used or heard about?

<https://www.menti.com>

[CODE:]

[GABIO_20240416_Data , last slide]

6. After data analysis: Publication and FAIR sharing of data.



Filter

- Subjects ⊕
- Content Types ⊕
- Countries ⊕
- AID systems ⊕
- API ⊕
- Data access ⊕
- Data access restrictions ⊕
- Database access ⊕
- Database access restrictions ⊕
- Database licenses ⊕
- Data licenses ⊕
- Data upload ⊕
- Data upload restrictions ⊕
- Enhanced publication ⊕
- Institution responsibility type ⊕
- Institution type ⊕
- Keywords ⊕
- Metadata standards ⊕
- PID systems ⊕
- Provider types ⊕
- Quality management ⊕
- Repository languages ⊕
- Software ⊕

gene expression

Search

Toogle short help

← Previous 1 2 3 4 Next →

Sort by ▾

Found 88 result(s)

Mouse Atlas of Gene Expression



Subject(s)

Microbiology, Virology and Immunology Animal Genetics, Cell and Developmental Biology

Biomedical Technology and Medical Physics Medicine Biology Life Sciences Zoology Medicine

Content type(s)

Plain text Structured text Scientific and statistical data formats Structured graphics Databases Software applications

other

Country

Canada

<<<!!!<<< This repository is no longer available >>>!!!>>>

C. Elegans Gene Expression



Subject(s)

Animal Genetics, Cell and Developmental Biology Zoology Biology Life Sciences

Content type(s)

Networkbased data Scientific and statistical data formats Databases other

Country

Canada

!!!<<< Genome data generated by BC Genome Sciences Centre is no longer available through this site as it is regularly deposited into controlled data

6. After data analysis: Publication and FAIR sharing of data.

You deposit data:

The screenshot shows the EGA website interface. At the top, there is a search bar and navigation tabs: ABOUT, SUBMISSION, BROWSE, ACCESS, DOWNLOAD, METADATA. A dropdown menu is open under 'SUBMISSION', listing options like 'Submission Terms', 'Public Keys', 'Submission Guide', 'Quick Guide', 'Sequence data', 'Array', 'Phenotype', 'Metagenomics', 'DAC', 'Tools', 'Data Use Conditions', and 'FAQ'. Below the menu is a bar chart titled 'Number of studies' showing the distribution of studies across disease categories: Cancer (2238), Cardiovascular (201), Infectious (60), Inflammatory (235), Neurological (100), and Other. A 'Latest' section highlights a study titled 'Mapping the t architecture o' from 2021-08-21. At the bottom, there are buttons for 'I want to access data' and 'I have data to submit'.

Disease Category	Number of Studies
Cancer	2238
Cardiovascular	201
Infectious	60
Inflammatory	235
Neurological	100
Other	-

Get your submission account

Fill the **submission form** and provide details of the data type(s), used platform(s) and estimated size of your submission.
If you are affiliated to an existing consortium, such as the International Cancer Genome Project (ICGC), please add this information in the comments.

Register Study/DAC

Use **Submitter Portal** to register your study, samples, Data Access Committee (DAC) and Policy.

Upload data

Encrypt your data files using the **EgaCryptor** or **encrypt your files locally** and upload it using default FTP clients or Aspera.

Register experiments and runs

Associate each data file to a registered sample and study by **Linking files to samples**. Details of the experimental procedure you followed must be provided.

Finish your submission

Group your runs/analysis into datasets and link them to your new or existing DAC and policy using **Submitter Portal** or an **XML based programmatic** submission. Data request are done at a dataset level, thus files within a datasets must share release conditions.

Release and admin

Instruct our Helpdesk to release your study. All registered studies are automatically placed on hold until the named submission or DAC contact instructs our Helpdesk for the study to be released.

6. After data analysis: Publication and FAIR sharing of data.

You source data:

Attention GDC Users: The [GDC Legacy Archive](#) is retiring soon. SOME FILES WILL NO

NIH NATIONAL CANCER INSTITUTE
GDC Data Portal

Harmonized Cancer Datasets
Genomic Data Commons Data Portal

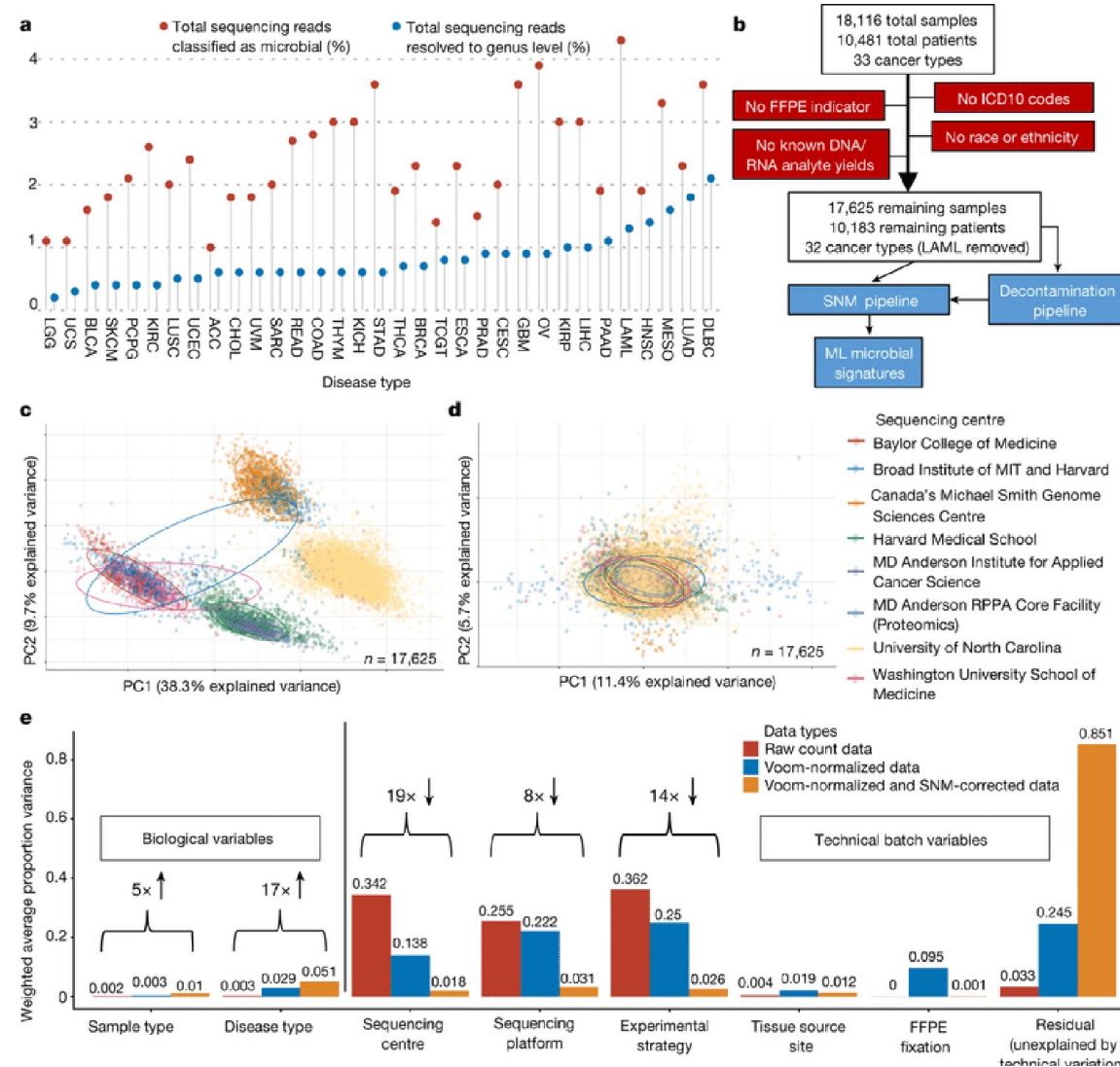
Get Started by Exploring:

Projects Exploration Analysis Repository

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

Data Portal Summary [Data Release 37.0 - March 29, 2023](#)

PROJECTS 78
PRIMARY SITES 68
FILES 931,947
GENES 22,501



Poore GD et al.: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 2020.

Store data forever?



Thank you!

Sophia Derdak

Core Facilities MedUni Vienna

sophia.derdak@meduniwien.ac.at

