McGill initiative in Computational Medicine

Introduction to UNIX in RNA-seq Data Analysis

CDSI Computational and Data Systems Initiative

ISCD Initiative en systèmes computationnels et de données

D2R DNA TO RNA | HeDS

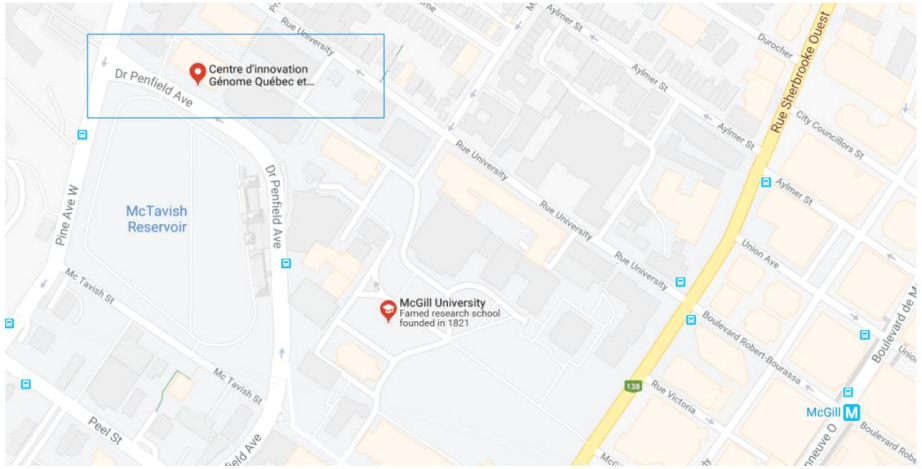
D2R DE L'ADN À L'ARN | HeDS

McGill initiative in Computational Medicine

Mission : aims to deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery

McGill.CA / MCGILL INITIATIVE IN COMPUTATIONAL MEDICINE

Contact



McGill initiative in Computational Medicine

McGill initiative in Computational Medicine
740, Dr. Penfield Avenue, Montreal, Quebec,
Canada, H3A 0G1
email: info-micm@mcgill.ca

Signup to our newsletter to receive the latest news

https://www.mcgill.ca/micm

Outline

- Setup & Troubleshooting
  - UNIX availability check
  - Fix PATH / permissions
- Module 1 — GEO Basics
  - Accession types
  - Common files
  - FTP/HTTPS & Data Download
- Module 2 — UNIX Basics
  - File management and navigation
  - File inspection
  - Pipes and redirects
  - Text search

https://github.com/QLS-MiCM/Intro-to-UNIX/tree/main

- R Troubleshooting (if time)

Setup & Troubleshooting

pwd
mkdir -p ~/workshop/data && cd ~/workshop
which bash zsh curl wget gzip tar || true

```
ztava@Zahra:~/workshop$ cd ~
ztava@Zahra:~$ pwd
/home/ztava
ztava@Zahra:~$ mkdir -p ~/workshop/data && cd ~/workshop
ztava@Zahra:~/workshop$ which bash zsh curl wget gzip tar || true
/usr/bin/bash
/usr/bin/curl
/usr/bin/wget
/usr/bin/gzip
/usr/bin/tar
```

wsl --install -d Ubuntu

wsl -l -v
wsl -d Ubuntu
exit or Ctrl+D → To exit

wsl --unregister "Ubuntu"

sudo apt update -y
sudo apt install -y curl wget gzip tar grep gawk sed

```
PS C:\Users\ztava> wsl --install -d Ubuntu
Downloading: Ubuntu
Installing: Ubuntu
Distribution successfully installed. It can be launched via 'wsl.exe -d Ubuntu'
Launching Ubuntu...
Provisioning the new WSL instance Ubuntu
This might take a while...
Create a default Unix user account: ztava
New password:
Retype new password:
No password has been supplied.
New password:
Retype new password:
passwd: password updated successfully
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ztava@Zahra:/mnt/c/Users/ztava$
```

Module 1

GEO Basics

GEO (Gene Expression Omnibus)

NCBI Resources How To Sign in to NCBI

GEO Home | Documentation | Query & Browse | Email GEO

Notice

Because of a lapse in government funding, the information on this website may not be up to date, transactions submitted via the website may not be processed, and the agency may not be able to respond to inquiries until appropriations are enacted. The NIH Clinical Center (the research hospital of NIH) is open. For more details about its operating status, please visit cc.nih.gov. Updates regarding government operating status and resumption of normal operations can be found at opm.gov.

Gene Expression Omnibus

GEO Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession | Search

Getting Started	Tools	Browse Content	
Overview	Search for Studies at GEO DataSets	Repository Browser	
FAQ	Search for Gene Expression at GEO Profiles	DataSets:	4348
About GEO DataSets	Search GEO Documentation	Series:	264424
About GEO Profiles	Analyze a Study with GEO2R	Platforms:	27739
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples:	8058260
How to Construct a Query	Programmatic Access		
How to Download Data	FTP Site		
	ENCODE Data Listings and Tracks		

GSE251845

Series GSE251845	Query DataSets for GSE251845
Status	Public on Apr 11, 2024
Title	Identification of unique gene expression and splicing events in early-onset colorectal cancer
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	Background: The incidence of colorectal cancer (CRC) has been steadily increasing in younger individuals over the past several decades for reasons that are incompletely defined. Identifying differences in gene expression profiles, or transcriptomes, in early-onset colorectal cancer (EOCRC, < 50 years old) patients versus later-onset colorectal cancer (LOCRC, > 50 years old) patients is one approach to understanding molecular and genetic features that distinguish EOCRC. Methods: We performed RNA-sequencing (RNA-seq) to characterize the transcriptomes of patient-matched tumors and adjacent, uninvolved (normal) colonic segments from EOCRC (n=21) and LOCRC (n=22) patients. The EOCRC and LOCRC cohorts were matched for demographic and clinical characteristics. We used The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) database for validation. We used a series of computational and bioinformatic tools to identify EOCRC-specific differentially expressed genes, molecular pathways, predicted cell populations, differential gene splicing events, and predicted neoantigens. Results: We identified an eight-gene signature in EOCRC comprised of ALDOB, FBXL16, IL1RN, MSLN, RAC3, SLC38A11, WBSCR27 and WNT11, from which we developed a score predictive of overall CRC patient survival. On the entire set of genes identified in normal tissues and tumors, cell type deconvolution analysis predicted a differential abundance of immune and non-immune populations in EOCRC versus LOCRC. Gene set enrichment analysis identified increased expression of splicing machinery in EOCRC. We further found differences in alternative splicing (AS) events, including one within the long non-coding RNA, HOTAIRM1. Additional analysis of AS found seven events specific to EOCRC that encode potential neoantigens. Conclusion: Our transcriptome analyses identified genetic and molecular features specific to EOCRC which may inform future screening, development of prognostic indicators, and novel drug targets.
Overall design	Gene expression profiles of 22 surgically resected tumors and patient-matched adjacent colonic segments from colorectal cancer patients were generated with RNA-sequencing.
Web link	https://doi.org/10.3389/fonc.2024.1365762

GEO accessions: GSE (series), GSM (samples), GPL (platform)

Processed data: Series Matrix / SOFT / tabular files

Raw count data: in supplementary file

Raw reads: usually in **SRA/ENA** as SRR… → `*.fastq.gz`

Raw (FASTQ): larger; needed for alignment/QC

Contributor(s)	Marx OM, Yochum GS, Koltun WA, Mankarious MM
Citation(s)	Marx OM, Mankarious MM, Koltun WA, Yochum GS. Identification of differentially expressed genes and splicing events in early-onset colorectal cancer. *Front Oncol* 2024;14:1365762. PMID: 38680862

NIH grant(s)

Grant ID	Grant title	Affiliation	Name
R03 CA279861	Wnt/beta-catenin signaling in early-onset colorectal cancer	PENNSYLVANIA STATE UNIV HERSHEY MED CTR	Gregory S. Yochum

Analyze with GEO2R **Download RNA-seq counts**

Submission date	Dec 21, 2023
Last update date	Apr 04, 2025
Contact name	Gregory Yochum
E-mail(s)	gyochum@pennstatehealth.psu.edu
Organization name	Penn State College of Medicine
Department	Biochemistry, Colorectal Surgery
Street address	700 HMC Crescent Road
City	Hershey
State/province	PA
ZIP/Postal code	17033
Country	USA

Platforms (1) GPL24676 Illumina NovaSeq 6000 (Homo sapiens)

Samples (44) More...
GSM7988989 24C
GSM7988990 24N
GSM7988991 27C

Relations
BioProject PRJNA1055547

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE251845_RAW.tar	9.0 Mb	(http)(custom)	TAR (of TXT)
GSE251845_htseq_raw_counts.csv.gz	2.1 Mb	(ftp)(http)	CSV

SRA Run Selector

SRA Run Selector

Filters List

1 Bases
2 Bytes
3 create_date
4 source_name
5 tissue
6 tumor_stage

Accession PRJNA1055547 Search

Common Fields

BioProject	PRJNA1055547
Consent	PUBLIC
Assay Type	RNA-Seq
AvgSpotLen	117
Center Name	BIOCHEMISTRY, COLORECTAL SURGERY, PENN STATE COLLEGE OF MEDICINE
Collection_Date	missing
DATASTORE filetype	FASTQ, RUN.ZQ, SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.us-east1, ncbi.public, s3.us-east-1

Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	44	52.75 Gb	152.51 G	Metadata or Accession List		

	Run	BioSample	age_at_surgery	Bases	Bytes	Experiment	Library Name	create_date	Sample Name	source_name	tissue	tumor_stage
1	SRR27320655	SAMN39058850	88.75	2.30 G	830.65 Mb	SRX22997924	GSM7989032	2023-12-21 21:33:00Z	GSM7989032	Rectum	Rectum	NA
2	SRR27320656	SAMN39058851	88.75	3.40 G	1.16 Gb	SRX22997923	GSM7989031	2023-12-21 21:36:00Z	GSM7989031	Rectum	Rectum	1
3	SRR27320657	SAMN39058852	68	4.74 G	1.63 Gb	SRX22997922	GSM7989030	2023-12-21 21:38:00Z	GSM7989030	Hepatic flexure	Hepatic flexure	NA
4	SRR27320658	SAMN39058853	68	3.06 G	1.06 Gb	SRX22997921	GSM7989029	2023-12-21 21:34:00Z	GSM7989029	Hepatic flexure	Hepatic flexure	2

NCBI > GEO > **Accession Display** Not logged in | Login

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSM7988989 GO

Sample GSM7988989

Query DataSets for GSM7988989

Status	Public on Apr 11, 2024
Title	24C
Sample type	SRA
Source name	Sigmoid
Organism	Homo sapiens
Characteristics	tissue: Sigmoid age at_surgery: 52.91666667 tumor stage: 2
Extracted molecule	polyA RNA
Extraction protocol	Tumors and adjacent colonic segments were surgically resected and aliquots were stored in RNAlater (Thermo Fisher) preservative. Samples were homogenized with a pestle and chloroform was added. The samples were centrifuged at 12,000 x g for 15 minutes. The aqueous layer was purified with RNAEasy Mini Kit (Qiagen). Samples were selected for poly-adenylation, and cDNA libraries were made and sequenced on MiSeq2500 (Illumina). RNA libraries were prepared for sequencing using standard Illumina protocols
Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina NovaSeq 6000
Description	24C
Data processing	Base calling with Illumina MiSeq2500 Alignment with STAR version 2.7.3 HTSeq counting of aligned reads Assembly: hg38 Supplementary files format and content: Count files contain raw counts for sequences aligning to each transcript.

Supplementary file	Size	Download	File type/resource
GSE251845_RAW.tar	9.0 Mb	(http)(custom)	TAR (of TXT)
GSE251845_htseq_raw_counts.csv.gz	2.1 Mb	(ftp)(http)	CSV

SRA Run Selector

Raw data are available in SRA

Custom GSE251845_RAW.tar archive:

Supplementary file	File size
☑ GSM7988989_24C_htseq.out.txt.gz	210.5 Kb
☐ GSM7988990_24N_htseq.out.txt.gz	211.3 Kb
☐ GSM7988991_27C_htseq.out.txt.gz	209.0 Kb
☐ GSM7988992_27N_htseq.out.txt.gz	206.9 Kb
☐ GSM7988993_29C_htseq.out.txt.gz	214.4 Kb
☐ GSM7988994_29N_htseq.out.txt.gz	205.9 Kb
☐ GSM7988995_30C_htseq.out.txt.gz	204.9 Kb
☐ GSM7988996_30N_htseq.out.txt.gz	207.2 Kb
☐ GSM7988997_31C_htseq.out.txt.gz	214.6 Kb
☐ GSM7988998_31N_htseq.out.txt.gz	210.4 Kb
☐ GSM7988999_32C_htseq.out.txt.gz	208.7 Kb
☐ GSM7989000_32N_htseq.out.txt.gz	208.3 Kb
☐ GSM7989001_33C_htseq.out.txt.gz	208.7 Kb
☐ GSM7989002_33N_htseq.out.txt.gz	206.4 Kb
☐ GSM7989003_34C_htseq.out.txt.gz	210.6 Kb
☐ GSM7989004_34N_htseq.out.txt.gz	208.2 Kb
☐ GSM7989005_35c_htseq.out.txt.gz	206.3 Kb
☐ GSM7989006_35N_htseq.out.txt.gz	210.2 Kb
☐ GSM7989007_36C_htseq.out.txt.gz	206.5 Kb
☐ GSM7989008_36N_htseq.out.txt.gz	205.8 Kb

☐ **Select All** Cancel Download **1 file(s), 210.5 Kb**

Exercise 1.1

Pick any GSE you like (or use the example on the slide). Using your browser only, download one of:

- ❑ Processed matrix — the Series Matrix (or SOFT) from the GSE page.
- ❑ Counts table (if provided) — a Supplementary file named like `*_counts.txt`, `htseq_counts`, or `featureCounts` (tab/CSV).
- ❑ Raw reads (one run) — a single small SRR FASTQ (from SRA Run Selector or the ENA "Run" page).

FTP / HTTPS Download

Pick one tool: `curl -L -O` <URL> or `wget` <URL>

```
ztava@Zahra:~/workshop$ wget https://ftp.ncbi.nlm.nih.gov/geo/series/GSE251nnn/GSE251845/suppl/GSE251845%5Fhtseq%5Fraw%5Fcounts.csv.gz
--2025-10-06 17:26:09--  https://ftp.ncbi.nlm.nih.gov/geo/series/GSE251nnn/GSE251845/suppl/GSE251845%5Fhtseq%5Fraw%5Fcounts.csv.gz
Resolving ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)... 130.14.250.12, 130.14.250.7, 130.14.250.31, ...
Connecting to ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2185236 (2.1M) [application/x-gzip]
Saving to: ‘GSE251845_htseq_raw_counts.csv.gz’

GSE251845_htseq_raw_counts.cs 100%[===============================================>]   2.08M  4.47MB/s    in 0.5s

2025-10-06 17:26:10 (4.47 MB/s) - ‘GSE251845_htseq_raw_counts.csv.gz’ saved [2185236/2185236]
```

```
ztava@Zahra:~/workshop$ curl -L -O https://ftp.ncbi.nlm.nih.gov/geo/series/GSE251nnn/GSE251845/suppl/GSE251845%5Fhtseq%5Fraw%5Fcounts.csv.gz
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 2134k  100 2134k    0     0   793k      0  0:00:02  0:00:02 --:--:--  793k
```

Exercise 1.2

Recreate your manual download using a command:

In your data folder:

```
cd ~/workshop/data
```

1. Run one:

```
curl -C - -L -O "<URL-from-Exercise-1.1>"
wget -c "<URL-from-Exercise-1.1>"
```

1. Verify:

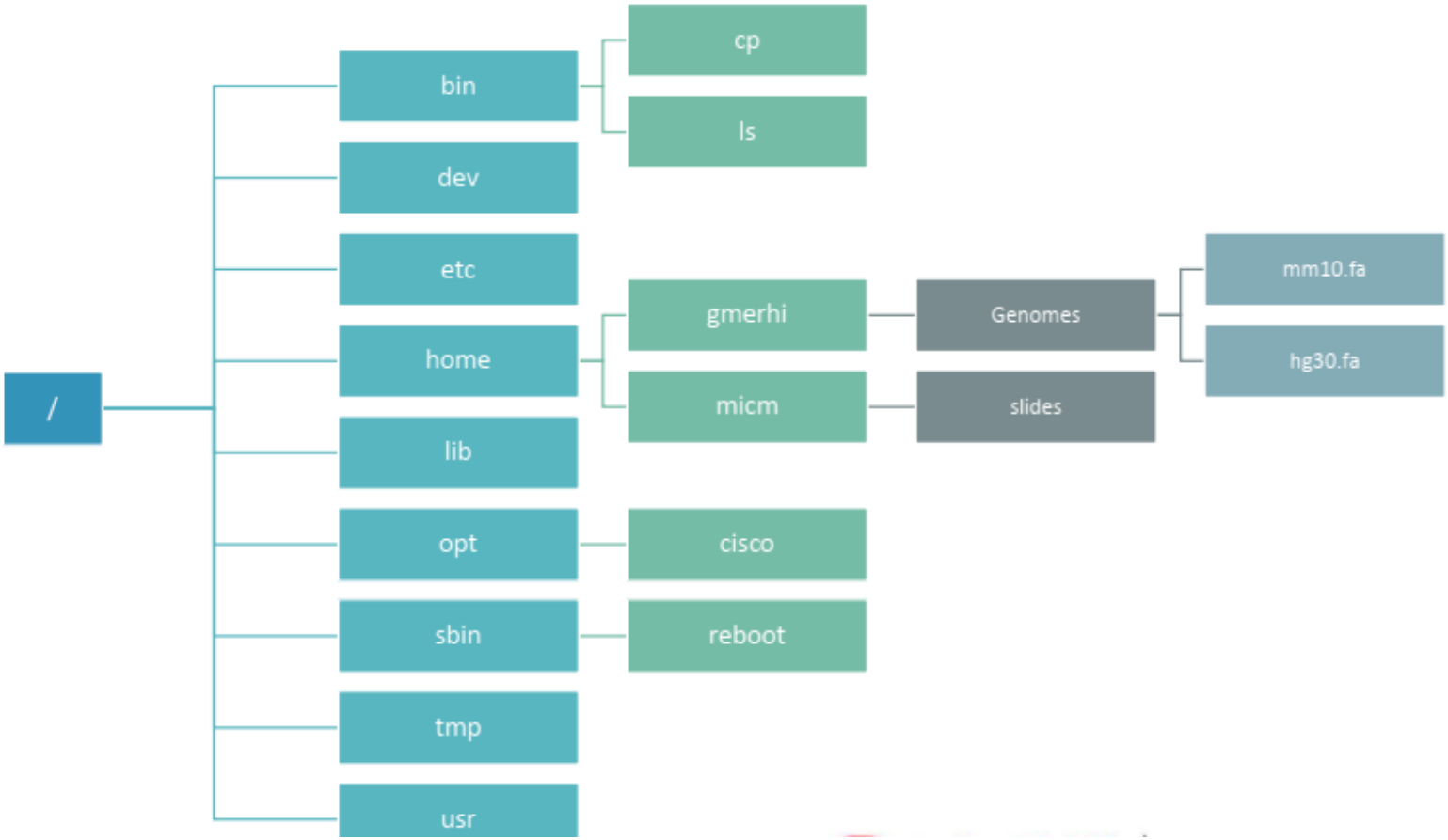
```
ls -lh
```

Module 2

UNIX Basics

https://github.com/QLS-MiCM/Intro-to-UNIX/tree/main

File system structure



Keyboard shortcuts

Ctrl+A	Ctrl+E	Ctrl+C	q
Go to the start of the line	Go to the end of the line	Stop the current process	Exist a child process (less, more, etc)

Unix commands

- Programs built in the shell that perform specific actions

	Command	-[Options]	[Argument(s)]
$	ls	-lt	~/Onedrive

Basic commands

top	man	history	clear
See active processes and the resources they're using	Shows the manual page of a command	List your previous commands	Clear your terminal window
% htop	% man ls % man cd % man htop	% history	% clear

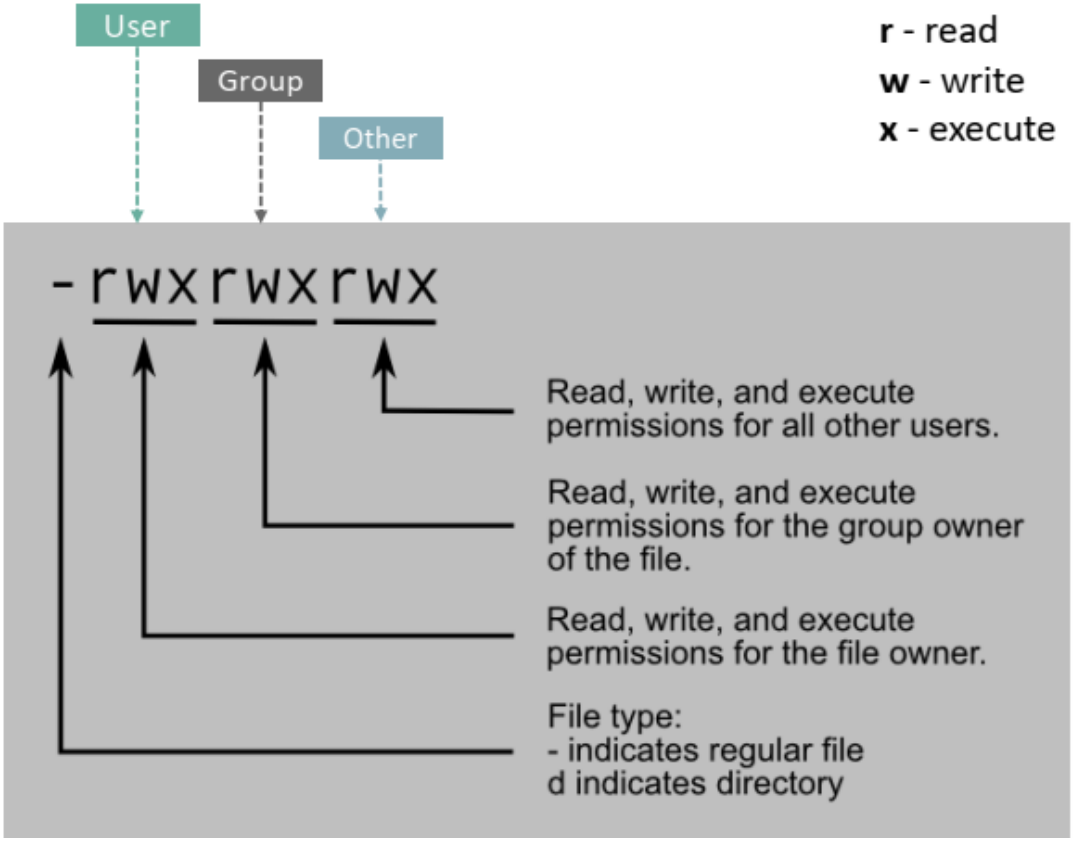
Basic commands

whoami	pwd	ls	cd
Displays current user id	Shows the current directory	Prints the current directory	Access a directory
% whoami	% pwd	% ls folder1	% cd folder1

Special characters

.	..	*	~
Current directory	Parent directory	Match any and all characters	Home directory
% ls .	% ls ..	% ls *	% ls ~

File permissions



r - read
w - write
x - execute

File management commands

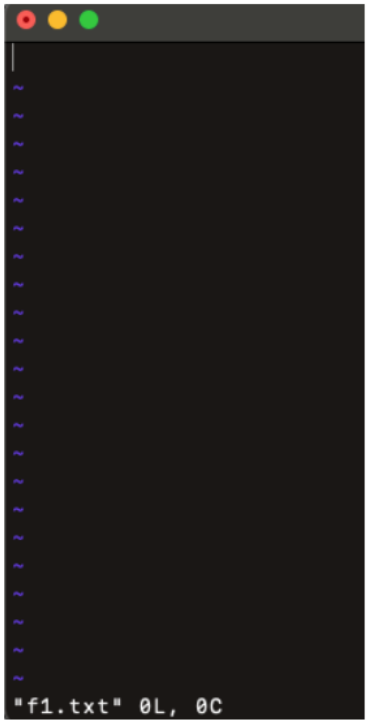
touch	mkdir	chmod	echo
Creates a new file	Creates a new directory	Change file permissions chmod **[ugo][+-][rwx]** u: user +: grant w: write g: group -: revoke r: read o: other x: execute	Prints something to the terminal
% touch f1.txt	% mkdir ~/intro_unix	% chmod o-w f1.txt	% echo “hello world”

File management commands

cp	mv	cat	zcat
Copy a file	Move a file or rename it	Print the contents of a file(s) to the terminal	Print the contents of a zipped file(s) to the terminal
% cp f1.txt f1_copy.txt	% mv f1_copy.txt f1.txt	% cat f1.txt	% zcat f1.txt.gz

How to edit files in the terminal

	nano	vi
	Recommended for plain text files	Recommended for code files/scripts
To open the file	% nano f1.txt	% vi f1.txt
To close the file	Press Ctrl+X Type Y Press Enter	Press Esc then hold Shift+zz Type :xq then press Enter



File management commands

gzip	gunzip	tar	rm
Compress a file	Decompress a file	Bundle files with compression (optional)	Removes file(s)
% gzip cars.tsv	% gunzip *	% tar –cvzf cars.tgz f*	% rm cars.tsv

File management commands

head	tail	more	wc
Print the first N lines of a file	Print the last N lines of a file	View contents of a file	Count words, characters lines or bytes
% head -3 cars.tsv	% tail -3 cars.tsv	% more cars.tsv	% wc cars.tsv

Text processing commands

cut	sort	uniq	paste
Extract columns of file	Order elements	Get set of uniq elements	Paste two files column-wise
% cut –f1 cars.csv > col1.tsv	% sort cars.tsv > cars.sort.tsv	% uniq cars.sort.tsv	% paste col1.txt cars.tsv

Exercise 2

1. Print the current directory and list all the contents of the directory
2. Create a directory named intro_unix in the home directory
3. Create a directory called data within intro_unix and sub-directories ho1 ho2 and ho3
4. Create a directory called folder1 under data/ho1
5. Go into folder1 and create two files: f1.txt and .f2.txt
6. Write the numbers from 1 to 10 in f1.txt (one number per line)
7. Change the name of .f2.txt to f2.txt
8. Write only the first 10 lines of f1.txt and all the lines in f2.txt to a new file called f3.txt

Pipes

- A way to connect the end of something with the start of something else
- They are specified with the control operator “ | ”

```
% cat cars.tsv | head -10
```

Redirect output

>	>>	<
Will send the output of the command to a NEW file	Will APPENND the output of the command to a file	Redirect output of one command as input to another command
% ls folder1 > files.txt	% ls folder1 >> files.txt	% head -1 <(cat files.txt)

Pattern matching

grep

Search a pattern in a file

```
% grep Male
happiness.csv
```

```
Country Gender Mean N=
AT Male 7.3 471
AT Female 7.3 570
AT Both 7.3 1041
BE Male 7.8 468
BE Female 7.8 542
BE Both 7.8 1010
BG Male 5.8 416
BG Female 5.8 555
BG Both 5.8 971
```

Regular expressions

Groups and ranges

.	Any character except new line (\n)
(a|b)	a or b
(...)	Group
(?:...)	Passive (non-capturing) group
[abc]	Range (a or b or c)
[^abc]	Not (a or b or c)
[a-q]	Lower case letter from a to q
[A-Q]	Upper case letter from A to Q
[0-7]	Digit from 0 to 7
\x	Group/subpattern number "x"

Character classes

\s	White space
\S	Not white space
\d	Digit
\D	Not digit
\w	Word
\W	Not word

Regular expressions

Anchors

^	Start of string, or start of line in multi-line pattern
\A	Start of string
$	End of string, or end of line in multi-line pattern
\Z	End of string
\b	Word boundary
\B	Not word boundary
\<	Start of word
\>	End of word

Quantifiers

*	0 or more	{3}	Exactly 3
+	1 or more	{3,}	3 or more
?	0 or 1	{3,5}	3, 4 or 5

Special characters

^	[	.	$
{	*	(	\
+	)	|	?
<	>		

The escape character is usually \

Example

- All lines that start with a letter

```
% grep -E "^[A-Z]+" happiness.csv
```

- Lines that do not start with a letter

```
% grep -v -E "^[A-Z]+" happiness.csv
% grep  -E "^\W" happiness.csv
```

- Lines with a country from the list

```
% grep -f countries.txt happiness.complete.tsv
```

awk

Scripting language that provides much more flexibility for text processing

Built-in functions

- length(string)
- tolower(string)
- toupper(string)
- match(string,pattern)

Built-in variables

- Entire line - $0
- Fields (specified by delimiter) - $1,$2,…

Example

- Print the 3rd column and then the 1st column

```
% awk '{print $3 "\t" $1}' happiness.complete.txt
```

- Print columns 1,3 and 2 if they contain the word “Female”

```
% awk '/Female/ {print $1 "\t" $3 "\t" $2}' happiness.complete.txt
```

- Count the number of characters in “Female” entries only

```
% awk -F "," '/Female/ { print length($0) "\t" $1 "\t" $2}'
happiness.complete.csv
```

sed

Command or streamline editor with multiple text processing functionalities

Built-in functions

- s/search/replace/ for pattern substitutions
  - /g – replace all occurrences
  - /1,/2,… - specifying which occurrence to replace
  - /I – Ignore case
  - /w – write to a file with /w filename
- -e to run multiple commands
  - sed -e 's/a/A/' -e 's/b/B/'

Examples

- Delete the first line

```
% ls -l | sed 1d
```

- Replace capital A for lowercase a

```
% sed 's/A/a/g' happiness.complete.txt | head
```

- Print the first line every 3 lines

```
% awk 'NR % 3 == 0' happiness.csv
```

Exercise 3.1

1. Preview & row count (*zcat/ gunzip*)

- Show the first 10 and last 5 lines.
- Report the number of data rows (exclude header).

2. Print the header with column indices so you can choose sample columns.

zcat GSE251845_htseq_raw_counts.csv.gz | head -n 1 | awk -F',' '{for(i=1;i<=NF;i++) printf("%d\t%s\n", i, $i)}'

```
ztava@Zahra:~/workshop$ zcat GSE251845_htseq_raw_counts.csv.gz | head -n 1 | awk -F',' '{for(i=1;i<=NF;i++) printf("%d\t%s\n", i, $i)}'
1       ""
2       "24C_htseq.out"
3       "24N_htseq.out"
4       "27C_htseq.out"
```

3. Pick & extract

- Choose two sample columns (your choice) plus the gene ID column.
- Save to `subset_cols.csv`.

4. Sort (*sort -t',' -k3,3nr)*

- Sort `subset_cols.csv` descending by one of your sample columns.
- Keep the header; save as `sorted_subset.csv`.

Exercise 3.1

5. Low-information checks

- ANY-zero genes: genes with at least one zero across samples.

awk -F',' 'NR>1 && ($2==0 || $3==0){c++} END{print c}' # ANY-zero

- ALL-zero genes: genes with all zeros across samples.
- Also report highly expressed genes for one chosen sample column (e.g., counts ≥ 1000).

Exercise 3.2

Move/copy your downloaded `*.fastq` files into your working directory.

Count reads (headers) — no decompression needed (hint: @)

Find how many reads contain ambiguous base ‘N’

Top 10 most common 6-mers at the 5' end

Motif search (choose a short motif, e.g., ACGTAC)

Take the first 100 reads

With thanks to our collaborators

CDSI Computational and Data Systems Initiative

ISCD Initiative en systèmes computationnels et de données

D2R DNA TO RNA | HeDS

D2R DE L'ADN À L'ARN | HeDS