File Management

Qi Wang

“Managing your Research Data: Best practices in Research Data Management for Biological Sciences”

2021 Mar 19

Outline

- Data Management **Principles**
  - Research Data Life-cycle
  - Data Management Checklist

- **Techniques** to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

Research Data Lifecycle

DOI: 10.1007/s10606-019-09354-z

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

Data Management Checklist

- **What** types of data and how long?

- **Who** will be responsible for which aspect?
  - Roles & Responsibilities.

- **How** to **document**/store/back-up/share data?
  - @Study-level @Data-level @File-level
  - Restrictions?
    - Ethical Obligations & Copyright/Intellectual Property.
  - Wet lab: Electronic Lab Notebook (ELN)

Based on slide from Jing Su

Data Types Recommended by UK Data Archive

Type of data	Recommended formats
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file
Tabular data with minimal metadata column headings, variable names	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements
Geospatial data vector and raster data	ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data Geography Markup Language (.gml)
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema
Image data	TIFF 6.0 uncompressed (.tif)
Audio data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)
Video data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)
Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)

For a full table including Acceptable formats:
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx

Data Management Checklist

- **What** types of data and how long?
- **Who** will be responsible for which aspect?
  - Roles & Responsibilities.
- **How** to **document**/store/back-up/share data?
  - @Study-level @Data-level @File-level
  - Restrictions?
    - Ethical Obligations & Copyright/Intellectual Property.
  - Wet lab: Electronic Lab Notebook (ELN)

Based on slide from Jing Su

Data Management Checklist – 13 Core Questions to Consider

- **What** data will you collect or create?
- How will the data be **collect**ed or created?
- What documentation and **metadata** will accompany the data?

- How will you manage any **ethical** issues?
- How will you manage **copyright** and Intellectual Property Rights (IPR) issues?

- How will the data be **stored and backed up** during the research?
- How will you manage **access and security**?
- What is the long-term **preservation** plan for the dataset?

- Which data should be retained, **shared**, and/or preserved?
- How will you **share** the data?
- Are any **restrictions** on data sharing required?

- **Who** will be **responsible** for data management?
- **What resources will you require to deliver your plan? (people, time, hardware)**

Full Checklist with Guidance can be downloaded here:
https://www.dcc.ac.uk/news/new-checklist-data-management-plan
https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

Your needs to think about when choosing Electronic Note Book

- Cost? One-time or subscription(monthly/yearly)?
- Access control. Other Users? Collaborators?
- Types of information to record and storage space required.
- Any specialized functionality you require?
- Data protection for sensitive data.
- What happens if someone leaves the lab?
- What happens when you stop using this ELN?

Further readings about how to choose an ELN

- Kwok, Roberta. 2018. "How to pick an electronic laboratory notebook." Nature 560 (7717): 269-270. *https://doi.org/10.1038/d41586-018-05895-3*

- *The Electronic Lab Notebook in 2020: A comprehensive guide with a short list of 5 ELNs. https://www.labfolder.com/electronic-lab-notebook-eln-research-guide*

- *A Comprehensive and Up-to-date Comparison Grid from Harvard Data Management including over 30 ELNs comparing over 50 features. https://datamanagement.hms.harvard.edu/analyze/electronic-lab-notebooks*

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?


No one has
perfect data management habits,
but adopting even a few
goes a long way.

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

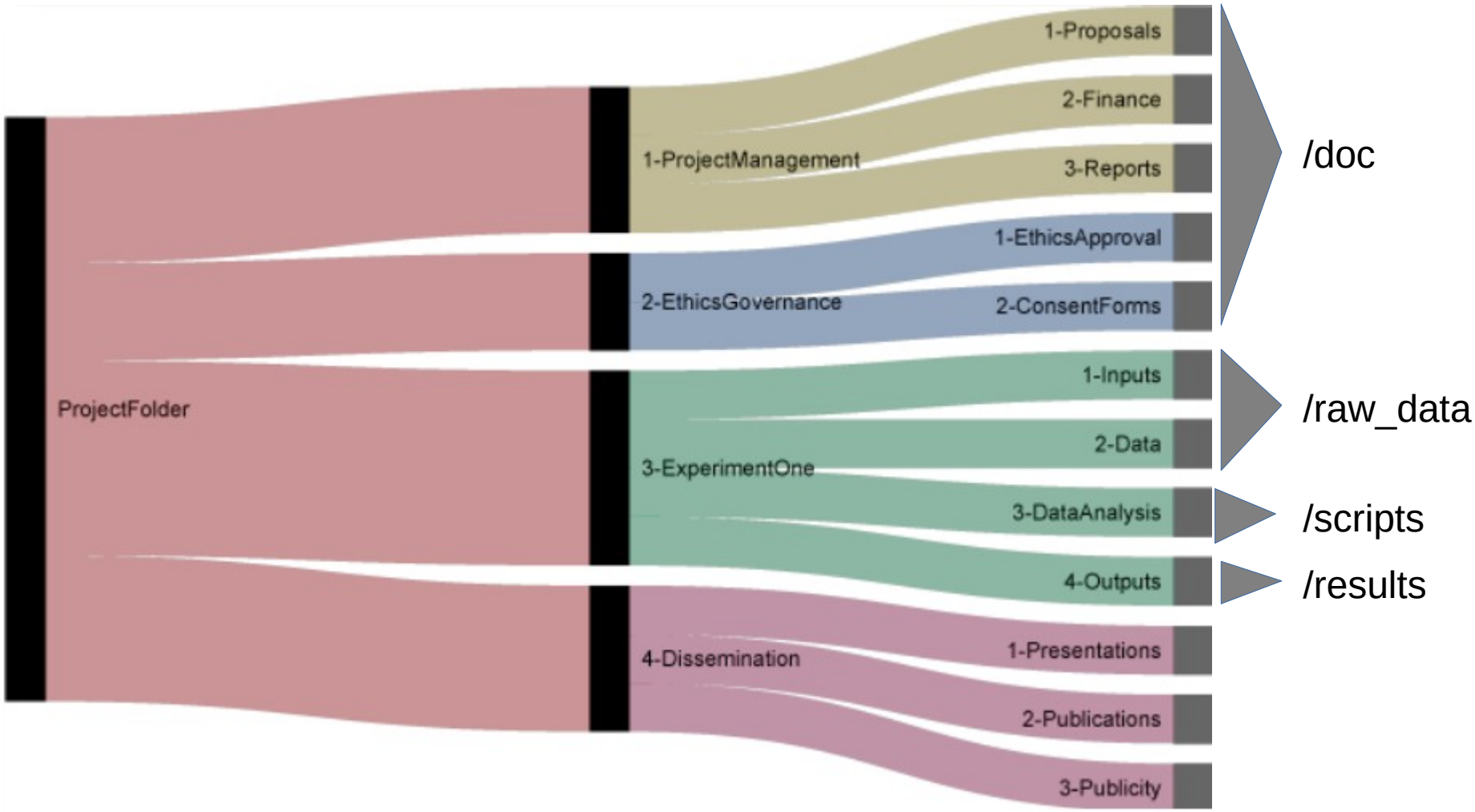
File Structure

- Ways to organize electronic files

- **Hierarchical**
  - Files organized in folders and sub-folders

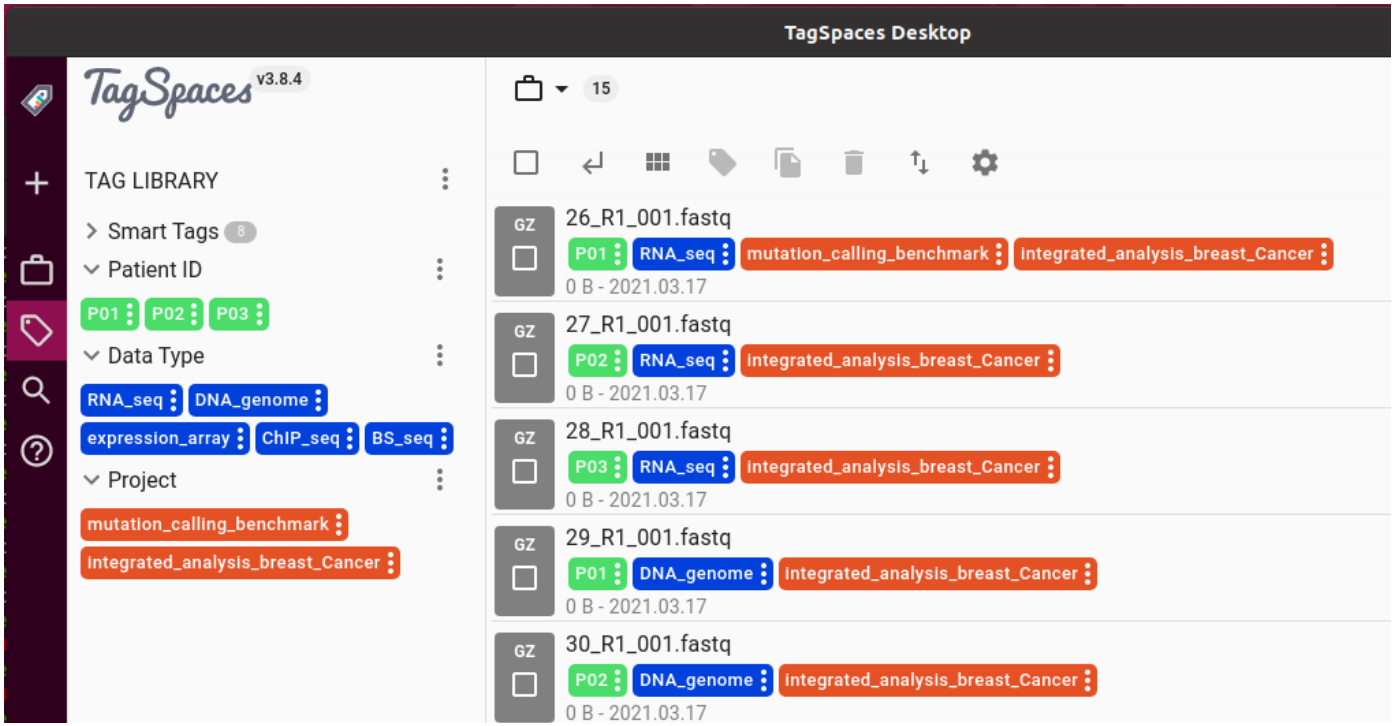
- **Tag-based**
  - Each file assigned one or more tags

Folder Structure Examples – Hierarchical



http://nikola.me/folder_structure.html

Folder Structure Examples – Tag-based



*Note: **Tags** themselves can have **hierarchical** structure.*

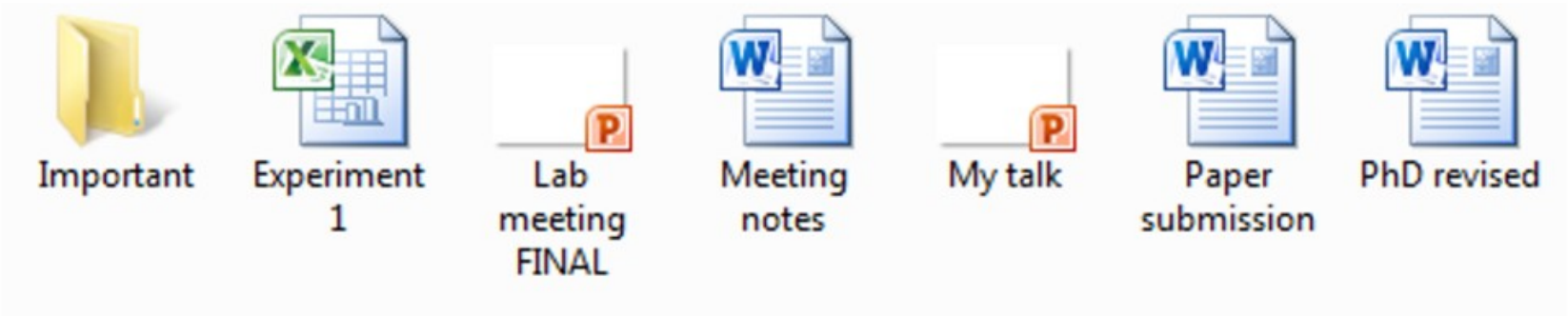
More about Tagging: https://libguides.mit.edu/metadataTools/

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Structures/Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

File Naming – does it matter?



In 3 years' time would you know what these are?

File Naming – 3C principles

Can your collaborator (or yourself 5 years from now) identify the content without opening the file?

- **Clear**
  - Objective: ~~my~~, ~~current~~, ~~latest~~, ~~final~~
  - Meaningful: He?
- **C**oncise
  - ~~the~~, ~~and~~
- **C**onsistent
  - Systematic naming/Standard

https://www.jisc.ac.uk/guides/managing-information/good-file-name

File Naming – Other Tips

- Use **underscores** "_ " to separate elements
  - avoid special characters, *e.g.*, "@" and spaces " "
  - Periods ". " only before the file extension
- Use **leading zero** for consistent sorting

Without Leading Zero

Name

- datafile_number_1.txt
- datafile_number_2.txt
- datafile_number_3.txt
- datafile_number_4.txt
- datafile_number_5.txt
- datafile_number_6.txt
- datafile_number_7.txt
- datafile_number_8.txt
- datafile_number_9.txt
- datafile_number_10.txt
- datafile_number_11.txt
- datafile_number_12.txt
- datafile_number_13.txt
- datafile_number_14.txt
- datafile_number_15.txt
- datafile_number_16.txt
- datafile_number_17.txt
- datafile_number_18.txt
- datafile_number_19.txt
- datafile_number_20.txt

```
qw254@qw254-desktop:~/t
datafile_number_10.txt
datafile_number_11.txt
datafile_number_12.txt
datafile_number_13.txt
datafile_number_14.txt
datafile_number_15.txt
datafile_number_16.txt
datafile_number_17.txt
datafile_number_18.txt
datafile_number_19.txt
datafile_number_1.txt
datafile_number_20.txt
datafile_number_2.txt
datafile_number_3.txt
datafile_number_4.txt
datafile_number_5.txt
datafile_number_6.txt
datafile_number_7.txt
datafile_number_8.txt
datafile_number_9.txt
```

Not consistent sorting.

With Leading Zero

Name

- datafile_number_01.txt
- datafile_number_02.txt
- datafile_number_03.txt
- datafile_number_04.txt
- datafile_number_05.txt
- datafile_number_06.txt
- datafile_number_07.txt
- datafile_number_08.txt
- datafile_number_09.txt
- datafile_number_10.txt
- datafile_number_11.txt
- datafile_number_12.txt
- datafile_number_13.txt
- datafile_number_14.txt
- datafile_number_15.txt
- datafile_number_16.txt
- datafile_number_17.txt
- datafile_number_18.txt
- datafile_number_19.txt
- datafile_number_20.txt

```
qw254@qw254-desktop:~/t
datafile_number_01.txt
datafile_number_02.txt
datafile_number_03.txt
datafile_number_04.txt
datafile_number_05.txt
datafile_number_06.txt
datafile_number_07.txt
datafile_number_08.txt
datafile_number_09.txt
datafile_number_10.txt
datafile_number_11.txt
datafile_number_12.txt
datafile_number_13.txt
datafile_number_14.txt
datafile_number_15.txt
datafile_number_16.txt
datafile_number_17.txt
datafile_number_18.txt
datafile_number_19.txt
datafile_number_20.txt
```

Consistent!

File Naming Examples

How about the following file name?

my Data @DryValley November 15 2010.v2.dat

How would you revise it?

(type your proposal in zoom chat window)

Reminder to check:

- **C**lear
  - Objective
  - Meaningful
- **C**oncise
- **C**onsistent
  - standard

File Naming Examples - revised

Original:
my Data @DryValley November 15 2010.v2.dat

Revised (one of the possibilities):
DV_ICPOES_20101115_JDS_v02.dat

- DV: site code (Dry Valley)
- ICPOES: instrument name
- 20101115: date of data generation
- JDS: initial of the scientist
- V02, second version (leading zero)

https://data.research.cornell.edu/content/file-management

Batching Renaming Tools

Windows:

- Ant Renamer: http://www.antp.be/software/renamer
- Bulk Rename Utility: http://www.bulkrenameutility.co.uk/
- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html

Mac:

- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
- Renamer4Mac : http://renamer4mac.com/
- Name Mangler: http://manytricks.com/namemangler/

Linux/Unix:

- GNOME Commander: http://www.nongnu.org/gcmd/
- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
- Use *grep, sed and awk* to search for and change

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

Version Control

Why?

- Track changes
- Enable reversing to earlier version

How?

- **File naming (maually)**
  - Date
    - Template_soil _testing_**20120319**.xlsx
  - Author's name
    - Template_soil _testing_by_**AS**.xlsx
  - Version number
    - *v01, v02 for major edit; v01_0, v01_1, v01_2 for minor edit*
    - Template_soil _testing_**v03_02**.xlsx
- **Version control tools** (automatically)
  - *Wet lab*
    - Electronic Lab Notebooks(ELN)
    - Laboratory Information Management System(LIMS)
  - *Dry lab*
    - Git (GitHub/GitLab)

Based on slide from Jing Su

Version Control Example

VERSION CONTROL TABLE FOR A DATA FILE	
Title:	Vision screening tests in Essex nurseries
File Name:	VisionScreenResults_00_05
Description:	Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007
Created By:	Chris Wilkinson
Maintained By:	Sally Watsley
Created:	04/07/ 2007
Last Modified:	25/11/ 2007
Based on:	VisionScreenDatabaseDesign_02_00

VERSION	RESPONSIBLE	NOTES	LAST AMENDED
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

https://ukdataservice.ac.uk/media/622417/managingsharing.pdf

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

Why meta data?

- What is metadata?
  - Description that help someone else understand the contents and organization of your files *in your absence*
  - Usually stored in top-level folders

- What does metadata include?
  - What?
  - Who?
  - How?
  - Where & When?

What are in meta data? (I)

- **What?**
- **Who?**
- How?
- Where & **When**?

DCC DATA RELEASES

DCC / Filter by file name...

Name

- README.txt
- current
- PCAWG
- release_28
- release_20
- release_19
- release_18
- release_17
- release_16
- release_15
- release_14

README.txt

ICGC - DCC DATA RELEASES

These are the DCC Data Releases of the International Cancer Genome Consortium (ICGC). Release 28 also contains PCAWG mutation data. Please see below for more information on the **PCAWG publication policy and embargo status.**

Current DCC Data Releases

Directory	Contents	Release Date
Release_28	DCC Data Release 28	03/27/2019
Release_27	DCC Data Release 27	04/30/2018
Release_26	DCC Data Release 26	12/08/2017
Release_25	DCC Data Release 25	06/08/2017
Release_24	DCC Data Release 24	05/17/2017

ICGC Publication and Embargo Policy

Contact

https://dcc.icgc.org/releases

What are in meta data? (II)

DCC / release_28 / Filter by file name...

Name	File Size	Date
README.txt	3.23 KB	Nov 26, 2019
Summary	--	Nov 26, 2019
Projects	--	Nov 26, 2019

README.txt

CORRECTIONS

LICA-FR sequencing-based expression data

November 26, 2019

The current sequencing expression data (exp_seq.LICA-FR_corrected.tsv.gz) at https://dcc.icgc.org/releases/current/Projects/LICA-FR contains **incorrect** raw read count values. Please download the corrected "exp_seq.LICA-FR.tsv.gz" file at https://dcc.icgc.org/releases/Supplementary/LICA-FR/corrected_data.

ICGC DATA PORTAL RELEASE 28

This is the Data Portal data Release 28 of the International Cancer Genome Consortium (ICGC). NOTE: This Release also contains PCAWG mutation data. Please see below for more information on the **PCAWG publication policy and embargo status**.

March 27, 2019

Summary

This release includes:

86 Cancer Projects
22 Cancer primary sites
22,330 Donors with molecular data in DCC
24,289 Total Donors
81,782,588 Simple Somatic Mutations
57,905 Mutated genes

https://dcc.icgc.org/releases

What are in meta data? (III)

- **What?**
- **Who?**
- How?
- **Where** & When?

DCC / release_28 / Projects / Filter by file name...

Name

- README.txt
- [ALL-US] Acute Lymphoblastic Leukemia - TARGET, US
- [AML-US] Acute Myeloid Leukemia - TARGET, US
- [BLCA-CN] Bladder Cancer - CN
- [BLCA-US] Bladder Urothelial Cancer - TCGA, US
- [BOCA-FR] Soft Tissue cancer - Ewing sarcoma - FR
- [BOCA-UK] Bone Cancer - UK
- [BPLL-FR] B-Cell Prolymphocytic Leukemia
- [BRCA-EU] Breast ER+ and HER2- Cancer - EU/UK

ALL-US Acute Lymphoblastic Leukemia - TARGET, US
AML-US Acute Myeloid Leukemia - TARGET, US
BLCA-CN Bladder Cancer - CN
BLCA-US Bladder Urothelial Cancer - TCGA, US
BOCA-FR Soft Tissue cancer - Ewing sarcoma - FR
BOCA-UK Bone Cancer - UK
BPLL-FR B-Cell Prolymphocytic Leukemia - FR
BRCA-EU Breast ER+ and HER2- Cancer - EU/UK

File Descriptions

Open-access analyzed data:

clinical.[ICGC project code].tsv.gz: contains aggregated clinical donor, specimen and sample information
exp_array.[ICGC project code].tsv.gz: gene expression measured at the transcriptional level (mRNA) using array-based platforms
exp_seq.[ICGC project code].tsv.gz: gene expression measured at the transcriptional level (mRNA) using sequencing-based platforms

File Formats

Controlled-Access Analyzed Data

Access to Raw Data (ie. BAM, FASTQ files)

https://dcc.icgc.org/releases

What are in meta data? (IV)

Details of the columns in Table S3:

1. sample_id
   tumor sample id (aliquot id)
2. ttype
   Tumor type name
3. chr
   Chromosome number
4. position
   Chromosome position
5. ref
   Reference allele
6. alt
   Alternate allele
7. gene
   Gene name
8. driver
   information related to 'mutational' driver type, in particular whether the driver mutation is in [promoters_core, 5utr, 3utr, enhancers, cds, ncRNA, mirna_pre, lncrna_promoters_core, splice_sites]
9. driver_statement
   information related to 'mutational' drivers, whether the driver mutation is known_driver, driver_by_rank, driver_by_rule or germline pathogenic variant
10. category
    subclasiffication of driver alteration types: CNA:[coding_amplification, coding_deletion], SV:[cis-activating_SV, coding_tsg_breakpoint, gene_fusion], mutational:['coding', 'noncoding']
11. top_category
    driver alteration types: somatic copy number alterations (CNA), somatic structural variant (SV), somaic point mutations, or germline pathogenic variants.

https://dcc.icgc.org/releases

What if your file transfer

got **interrupted**

without any warning message?

What are in meta data? (V)

- Avoid pitfalls in data transfer using md5sum check

file name	md5sum
PCAWG16.consensus.virus.genus.normal.2out3.v3.icgc.controlled.tsv.gz	854b6a4dce3b46891c8cc4afc65a40d3
PCAWG16.consensus.virus.genus.normal.3out3.v3.icgc.controlled.tsv.gz	82f20aa61129522672fb8e1d7036cdfc
PCAWG16.consensus.virus.genus.tumour.2out3.v3.icgc.controlled.tsv.gz	1787e28e61651b19701cfbb9c108b908
PCAWG16.consensus.virus.genus.tumour.3out3.v3.icgc.controlled.tsv.gz	054200b756d059fc435c6f39ae9646b3
PCAWG16.consensus.virus.genus.normal.2out3.v3.tcga.controlled.tsv.gz	bba31c95dad98dc3b796c6937969a4e7
PCAWG16.consensus.virus.genus.normal.3out3.v3.tcga.controlled.tsv.gz	af0d91d2be2263f68c40e10a7780aced
PCAWG16.consensus.virus.genus.tumour.2out3.v3.tcga.controlled.tsv.gz	f5c5c6b6b09a2f2eb1372cdfd85077b9
PCAWG16.consensus.virus.genus.tumour.3out3.v3.tcga.controlled.tsv.gz	8e1352617fff430d5bedfcaa8fd3362f

- Md5sum output are "**fingerprints**" to files. They are hash values derived using the whole file as input.
- Changes to a file will cause md5sum output to change. Conversely, if md5sum outputs are the same the files are identical.

Note : If you are worried that the data is maliciously altered instead of accidental corruption, there are more advanced options: SHA-256 (sha256sum), SHA-512 (sha512sum) or BLAKE2(b2sum).

https://dcc.icgc.org/releases

Further reading about meta data:

Cornell University has excellent README file guidelines:
https://data.research.cornell.edu/content/readme

Putting It Together

```
Project_20190102_SuJ_BC_RNASeq_5490/
    │
    ├── README
    │
    ├── meta/
    │       ├── ExperimentalDesign_20190105.doc
    │       ├── Platelayout.doc
    │       ├── SampleSheet.csv
    │       ├── FileList.csv
    │       └── MeetingNotes_20190114.doc
    │
    ├── data/
    │   ├── bam/
    │   │   ├── Sample1.aligned.bam
    │   │   └── Sample2.aligned.bam
    │   └── fastq/
    │       ├── Sample1_R1.fq
    │       ├── Sample1_R2.fq
    │       ├── Sample2_R2.fq
    │       └── Sample2_R2.fq
    │
    ├── reference/
    │   └── human/
    │       ├── Grch38_genome.fa
    │       └── Gencode26_genes.gtf
    │
    ├── scripts/
    │     ├── 1.subread_align_grch38.sh
    │     ├── 2.featureCounts_mRNA.sh
    │     ├── 3.edgeR_DE.R
    │     ├── 4.GSEA_mSigDB13.sh
    │     └── 5.backup_scratca_to_nas.sh
    │
    └── results/
          ├── counts/
          │   └── count_files.txt
          └── DE_genelists
              ├── PrimaryTumour_v_normal_DE_genelist.csv
              └── Metastatic_v_Primary_DE_genelist.csv
```

- File Structures/Organization
- File Naming
- Version Control
- Meta Data

Jing Su

Outline

- Data Management Principles
  - Research Data Life-cycle
  - Data Management Checklist

- Techniques to help organize your research data
  - File Organization
  - File Naming
  - Version Control
  - Meta Data (ReadMe)
  - Running Low in Storage Space?

Running out of Space – Do we need all the files?

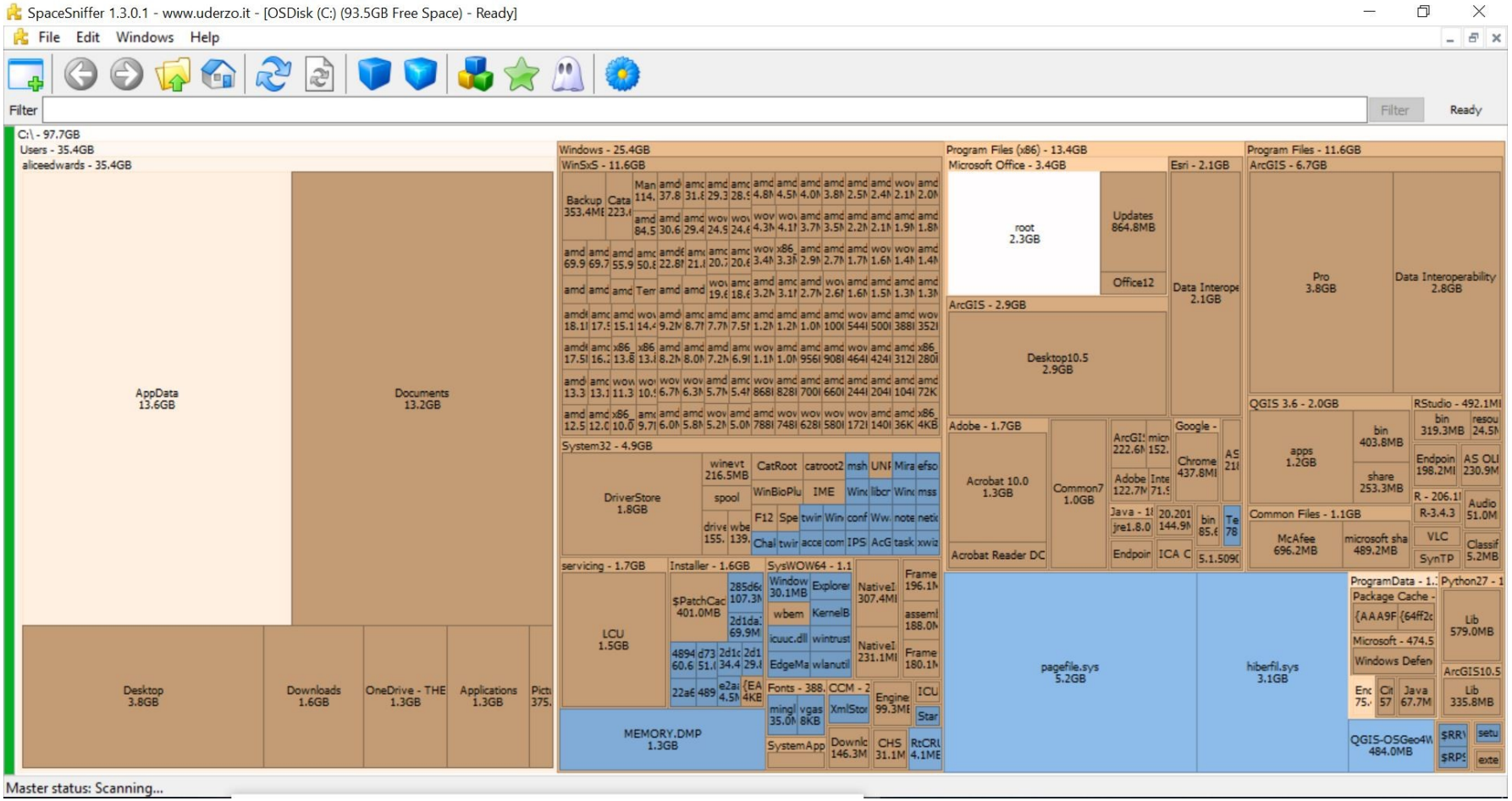
Five steps to decide what data to keep:

1. Identify **purposes** that the data could fulfill
2. Identify data that **must** be kept
3. Identify data that **should** be kept
4. Weight up the **costs**
5. Complete the data appraisal

For more details:
https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep

Running out of Space – for Windows

SpaceSniffer (http://www.uderzo.it)



Abigail Edwards

Running out of Space – for Mac & other Linux?

For Mac:

Disk Inventory X (http://www.derlien.com/)

Linux command line (bash):

du -sh # shows you how much disk space the current folder takes

du -h -d 1 | sort -h # sort all folders in the current directory by size

Summary

- The Complete Research Data Life-cycle

  Can someone else (as well as yourself years from now) understand the contents and organization of your files *in your absence*.

- Data Management Checklist
  - **W**hat?
  - **W**ho?
  - **H**ow?

- Data Management Techniques
  - File Structure & ReadMe & File Name (3C)
    - add **W**here and **W**hen
  - Keep track of changes with Version Control
  - Avoid pitfalls in data transfer using md5sum check

Further readings:

Good Practice in Bioinformatics Analysis

Wilson, G. (2017). Good enough practices in scientific computing. PLoS Computational BiologyS Computational Biology. (https://doi.org/10.1007/BF02378113)