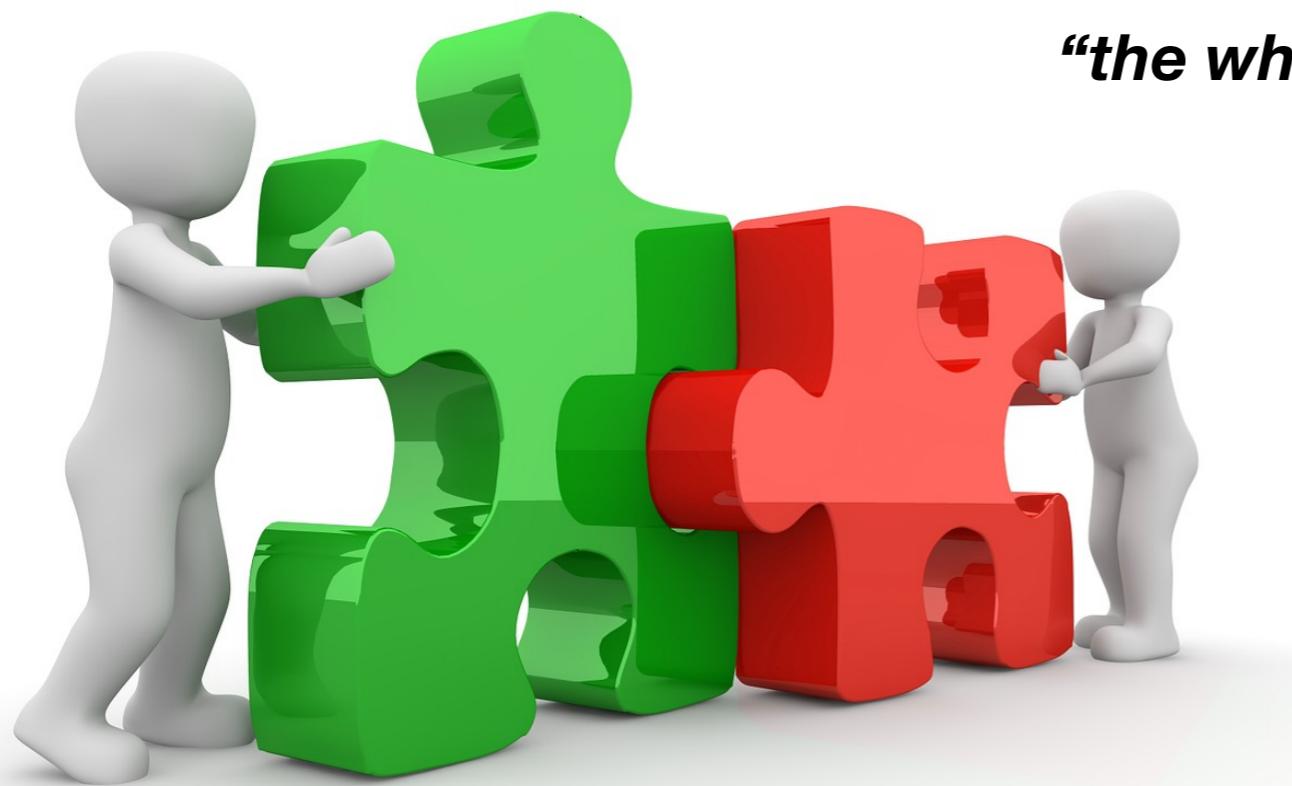




Interdisciplinary collaboration



“the whole is greater than the sum of its parts”

~ Aristotele

Olga Dethlefsen
olga.dethlefsen@nbis.se

Introduction: talking from experience

Introduction: talking from experience

Can I ask you a question? I'm trying to understand the results of this copy number variations analysis and I'd thought I would ask you as you have been working with EVERYTHING (~ NBIS colleague)

Introduction: talking from experience

Can I ask you a question? I'm trying to understand the results of this copy number variations analysis and I'd thought I would ask you as you have been working with EVERYTHING (~ NBIS colleague)



BEng/MSc in Technical Physics, PL



MSc internship, FR



PhD in Bioinformatics, UK
Medical Statistician, Oxford, UK



Post-docs experiences
KI, Sweden & RIKEN, Japan



Introduction: talking from experience

Can I ask you a question? I'm trying to understand the results of this copy number variations analysis and I'd thought I would ask you as you have been working with EVERYTHING (~ NBIS colleague)



BEng/MSc in Technical Physics, PL



MSc internship, FR



PhD in Bioinformatics, UK
Medical Statistician, Oxford, UK



Post-docs experiences
KI, Sweden & RIKEN, Japan



- ❖ at NBIS since 2015
- ❖ “bioinformatics expert”
- ❖ 40+ projects
- ❖ across multiple omics
- ❖ medical focus

Introduction: session's aim



Introduction: session's aim



*** to highlight and discuss
some of the collaboration
aspects worth thinking
about**

Introduction: session's aim



* to highlight and discuss
some of the collaboration
aspects worth thinking
about

* to introduce “reproducibility”
concepts covered more in day 2

Introduction: session's aim



* to highlight and discuss some of the collaboration aspects worth thinking about

* to introduce “reproducibility” concepts covered more in day 2

* to prepare a first draft of a “Good collaboration checklist”

Introduction: session's aim



- * to highlight and discuss some of the collaboration aspects worth thinking about

- * to introduce “reproducibility” concepts covered more in day 2

- * to prepare a first draft of a “Good collaboration checklist”

Introduction: theory

[https://docs.google.com/document/d/
1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing](https://docs.google.com/document/d/1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing)

Introduction: theory

[https://docs.google.com/document/d/
1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing](https://docs.google.com/document/d/1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing)

“the act of working together with other people or organisations to create or achieve something”

– Cambridge Dictionary

Introduction: theory

[https://docs.google.com/document/d/
1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing](https://docs.google.com/document/d/1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing)

“the act of working together with other people or organisations to create or achieve something”

– Cambridge Dictionary



Introduction: theory

[https://docs.google.com/document/d/
1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing](https://docs.google.com/document/d/1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing)

“the act of working together with other people or organisations to create or achieve something”

– Cambridge Dictionary



- trust
- attachment
- clarity and alignment
- speed
- technology, geography and culture

Introduction: theory

[https://docs.google.com/document/d/
1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing](https://docs.google.com/document/d/1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing)

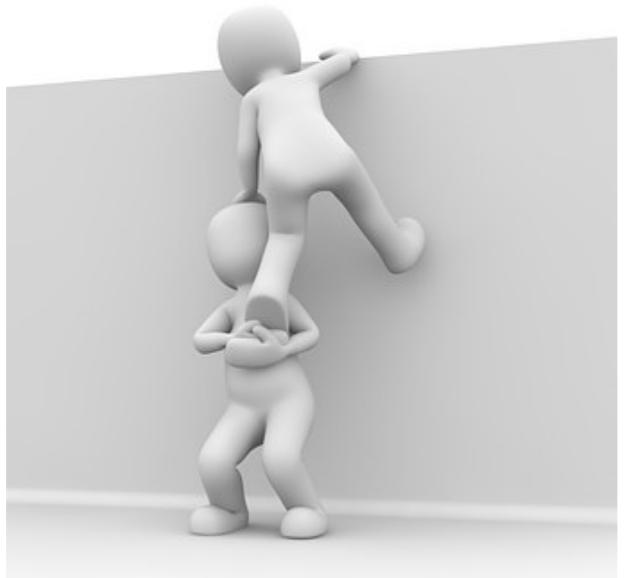
“the act of working together with other people or organisations to create or achieve something”

– Cambridge Dictionary



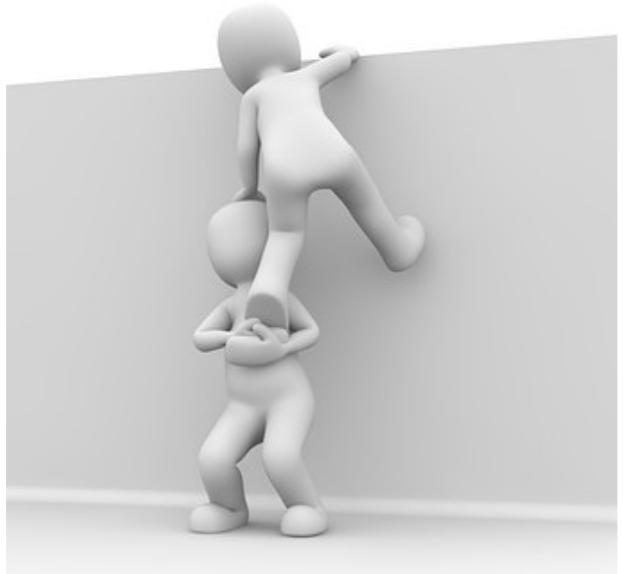
- trust
- attachment
- clarity and alignment
- speed
- technology, geography and culture

- diversity
- speed
- engagement
- productivity



Alignment on a collective goal

- missing details
- conflicting incentives
- conflicting prioritisation
- conflict avoidance

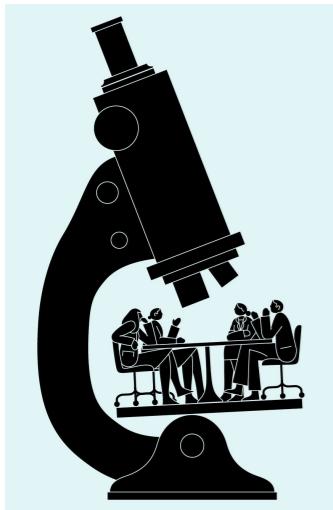


Alignment on a collective goal

- missing details
- conflicting incentives
- conflicting prioritisation
- conflict avoidance

Team players

- core team
- others joining as needed
- roles and responsibilities
 - e.g. facilitators, evaluators
 - e.g. coordinators, communicators
- engage leadership support
- setting the team for success



Google

THE WORK ISSUE

What Google Learned From Its Quest to Build the Perfect Team

New research reveals surprising truths about why some work groups thrive and others falter.

- Equality of distribution of conversational turn-taking
- Highly average social sensitivity



re:Work

<https://rework.withgoogle.com/print/guides/5721312655835136/>

<https://www.nytimes.com/2016/02/28/magazine/what-google-learned-from-its-quest-to-build-the-perfect-team.html?smid=pl-share>

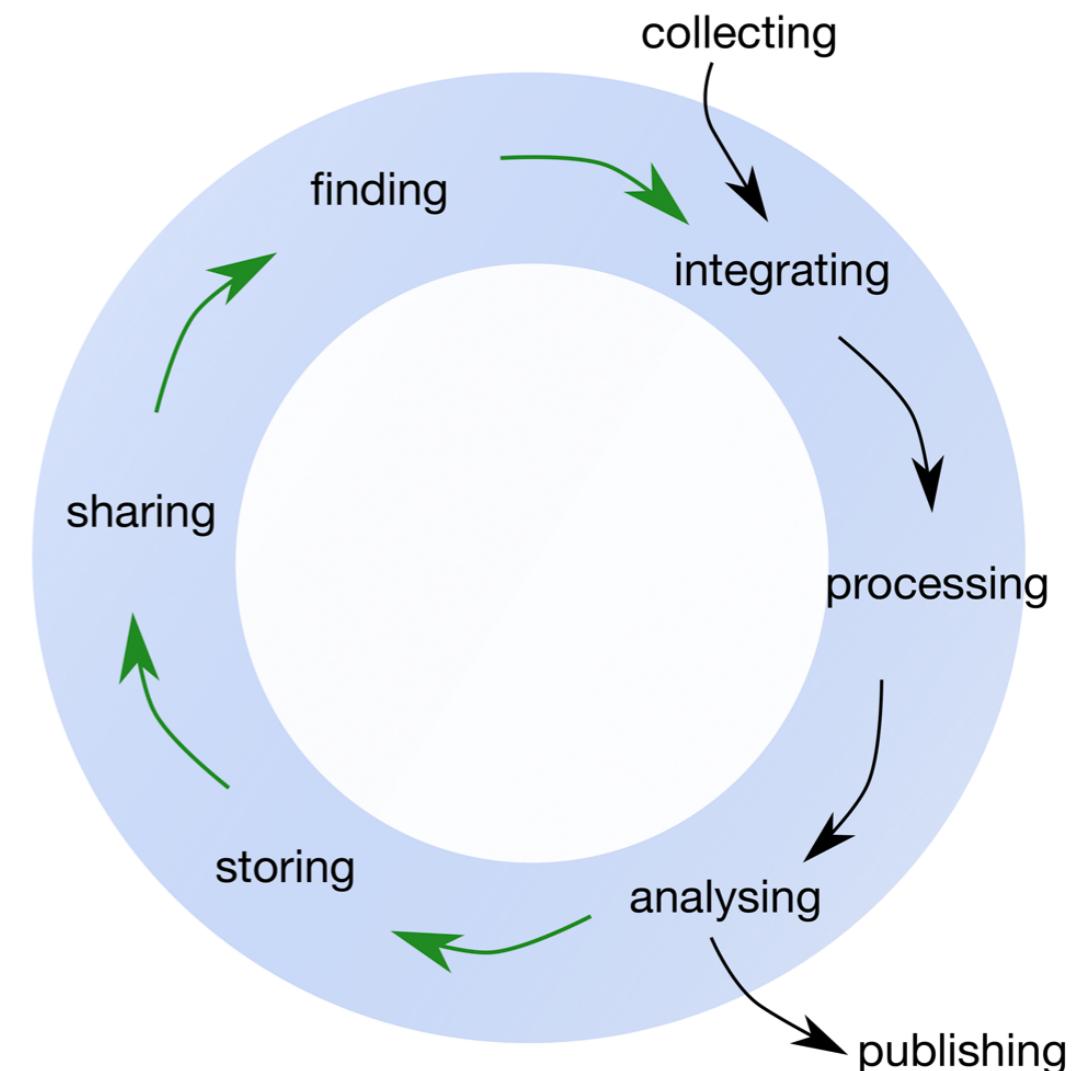
Project management main steps

- defining project**
- listing tasks**
- estimate times and costs**
- assess risk and prepare action plans**
- monitor progress & costs**
- review**

Project management main steps

- defining project
- listing tasks
- estimate times and costs
- assess risk and prepare action plans
- monitor progress & costs
- review

Simplified Data Life Cycle framework for bioscience, biomedical and bioinformatics data

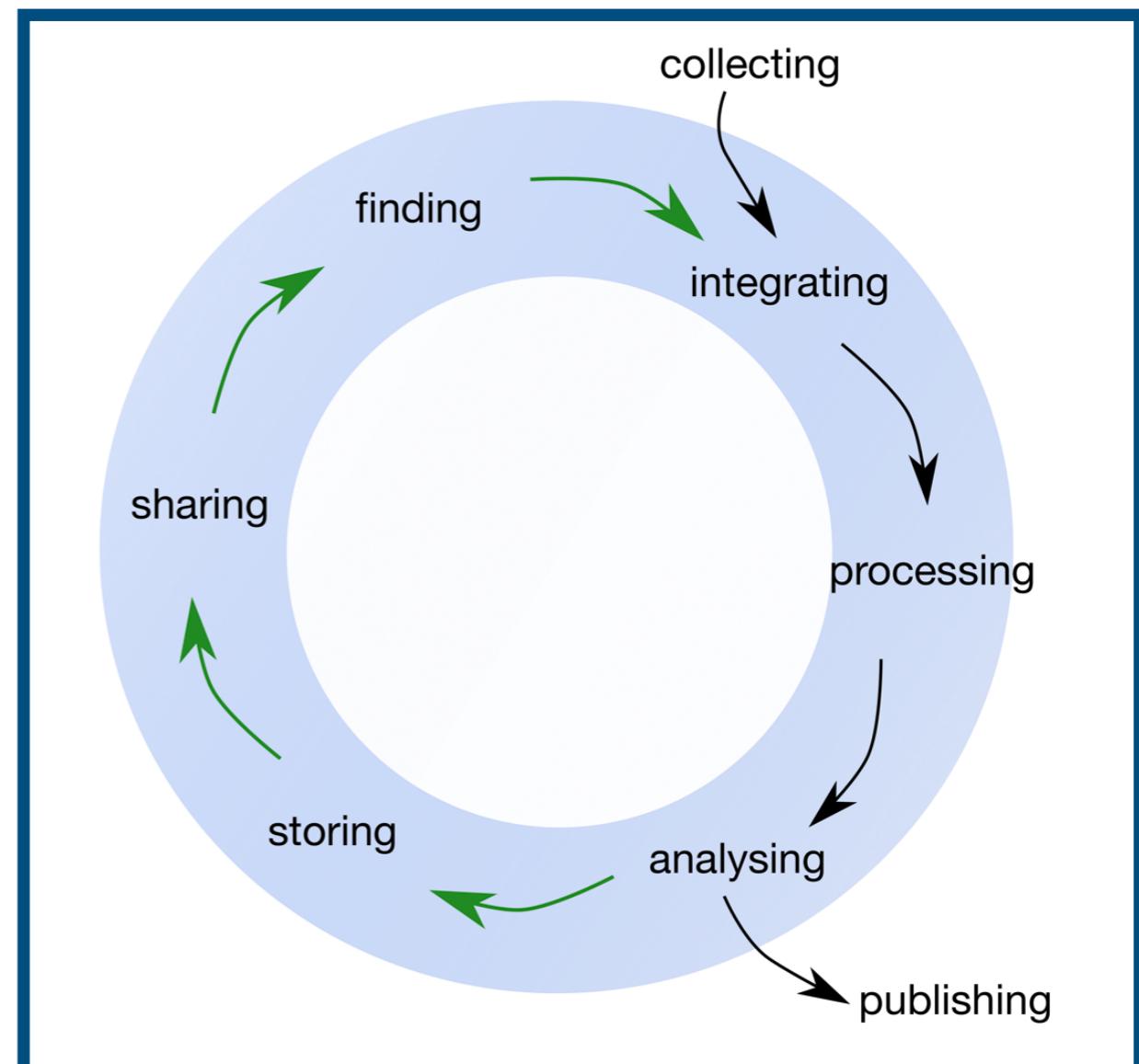


Project management main steps

- defining project
- listing tasks
- estimate times and costs
- assess risk and prepare action plans
- monitor progress & costs
- review

Simplified Data Life Cycle framework for bioscience, biomedical and bioinformatics data

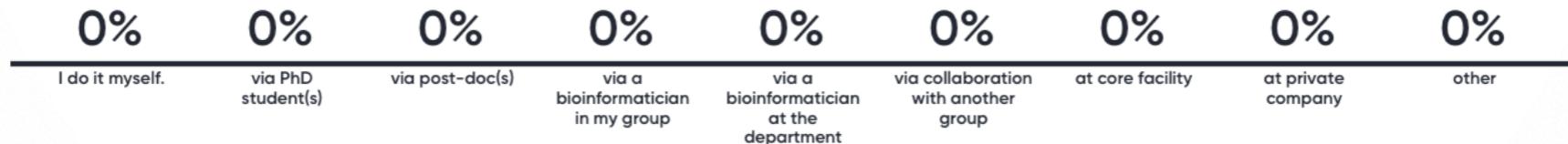
potential bioinformatics needs



OPTIONS

Go to www.menti.com and use the code 86 28 89

How do you get bioinformatics done?



OPTIONS

[https://docs.google.com/document/d/](https://docs.google.com/document/d/1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing)

1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing

OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely



OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely

via a PhD student(s)

is there a second supervisor? asking for troubles, if not



OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely

via a PhD student(s)

is there a second supervisor? asking for troubles, if not

via a post-docs(s)

it may work short-term if a post-doc(s) has some experience



OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely

via a PhD student(s)

is there a second supervisor? asking for troubles, if not

via a post-docs(s)

it may work short-term if a post-doc(s) has some experience

via a group / depart.
bioinformatician

it may be a good set-up for the group, and hopefully for the
bioinformatician; is anyone discussing with him / her career development?



OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely

via a PhD student(s)

is there a second supervisor? asking for troubles, if not

via a post-docs(s)

it may work short-term if a post-doc(s) has some experience

via a group / depart.
bioinformatician

it may be a good set-up for the group, and hopefully for the
bioinformatician; is anyone discussing with him / her career development?

via a group
collaboration

may be a win-win, long-lasting collaboration



OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely

via a PhD student(s)

is there a second supervisor? asking for troubles, if not

via a post-docs(s)

it may work short-term if a post-doc(s) has some experience

via a group / depart.
bioinformatician

it may be a good set-up for the group, and hopefully for the
bioinformatician; is anyone discussing with him / her career development?

via a group
collaboration

may be a win-win, long-lasting collaboration

at a sequencing
centre

wait, have you double-checked that the default pipelines and workflows
are applicable to your project, worth time and money?



OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely

via a PhD student(s)

is there a second supervisor? asking for troubles, if not

via a post-doc(s)

it may work short-term if a post-doc(s) has some experience

via a group / depart.
bioinformatician

it may be a good set-up for the group, and hopefully for the
bioinformatician; is anyone discussing with him / her career development?

via a group
collaboration

may be a win-win, long-lasting collaboration

at a sequencing
centre

wait, have you double-checked that the default pipelines and workflows
are applicable to your project, worth time and money?

at a core facility

may be great, but what really motivates these bioinformaticians? And how
does the core facility work?



OPTIONS

I do it by myself

most likely unrealistic, admirable, lonely

via a PhD student(s)

is there a second supervisor? asking for troubles, if not

via a post-doc(s)

it may work short-term if a post-doc(s) has some experience

via a group / depart.
bioinformatician

it may be a good set-up for the group, and hopefully for the
bioinformatician; is anyone discussing with him / her career development?

via a group
collaboration

may be a win-win, long-lasting collaboration

at a sequencing
centre

wait, have you double-checked that the default pipelines and workflows
are applicable to your project, worth time and money?

at a core facility

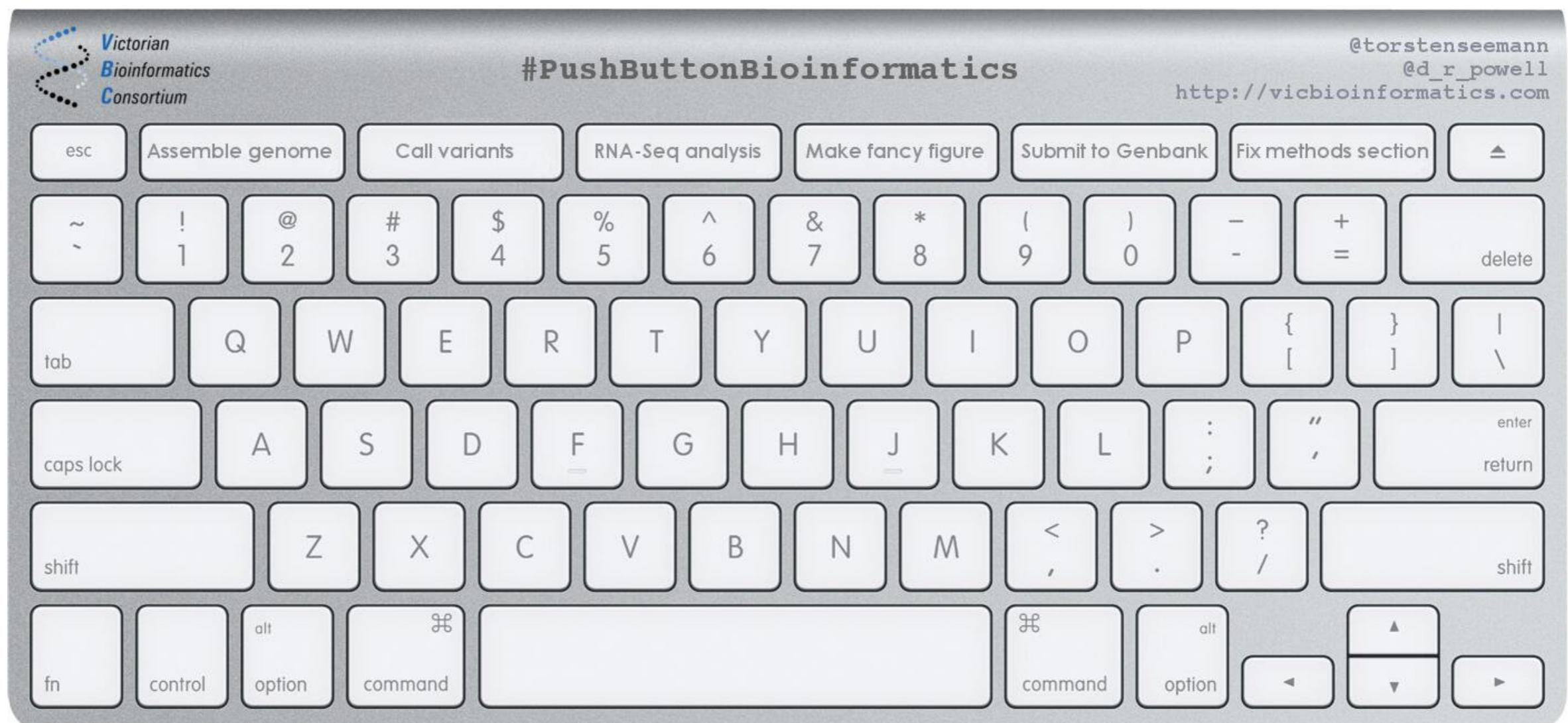
may be great, but what really motivates these bioinformaticians? And how
does the core facility work?

at an external
company

will you ever see the code and be able to explain M&M?

NO OPTIONS

#PushButtonBioinformatics



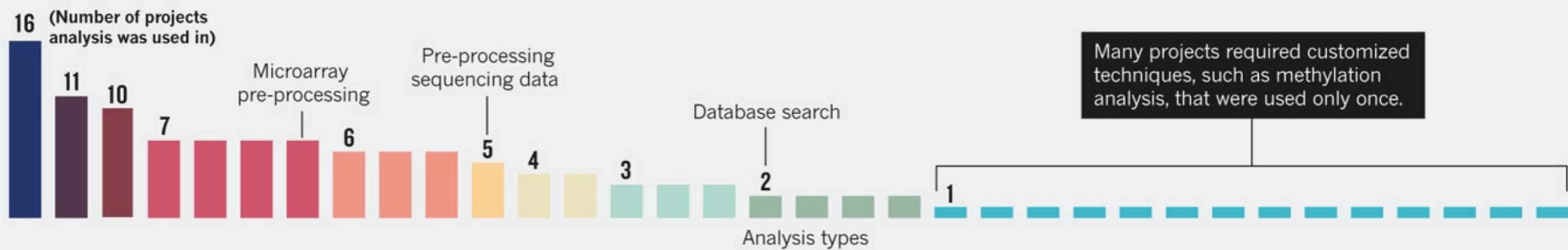
OPTIONS



“Biological data will continue to pile up unless those who analyse it are recognised as creative collaborators in need of career paths”

– Jeffrey Chang, 2015

Over 18 months, 46 data-analysis projects undertaken at the bioinformatics core of the University of Texas Health Science Center at Houston required 34 different types of analysis — most were used infrequently. Each project demanded unique combinations of analyses, demonstrating how bioinformaticians must be versatile, creative and collaborative.



Core services: Reward bioinformaticians; *Nature* **520**, 151–152 (09 April 2015) doi:10.1038/520151a

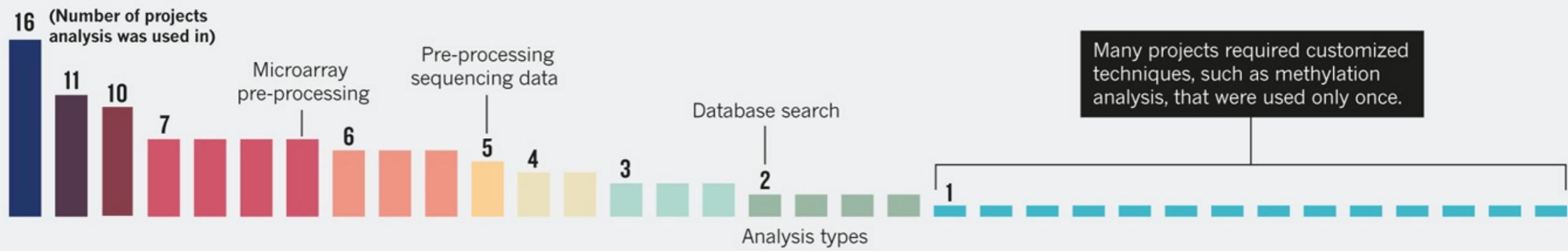
OPTIONS

“Biological data will continue to pile up unless those who analyse it are recognised as creative collaborators in need of career paths”

– Jeffrey Chang, 2015



Over 18 months, 46 data-analysis projects undertaken at the bioinformatics core of the University of Texas Health Science Center at Houston required 34 different types of analysis — most were used infrequently. Each project demanded unique combinations of analyses, demonstrating how bioinformaticians must be versatile, creative and collaborative.



Core services: Reward bioinformaticians; *Nature* 520, 151–152 (09 April 2015) doi:10.1038/520151a

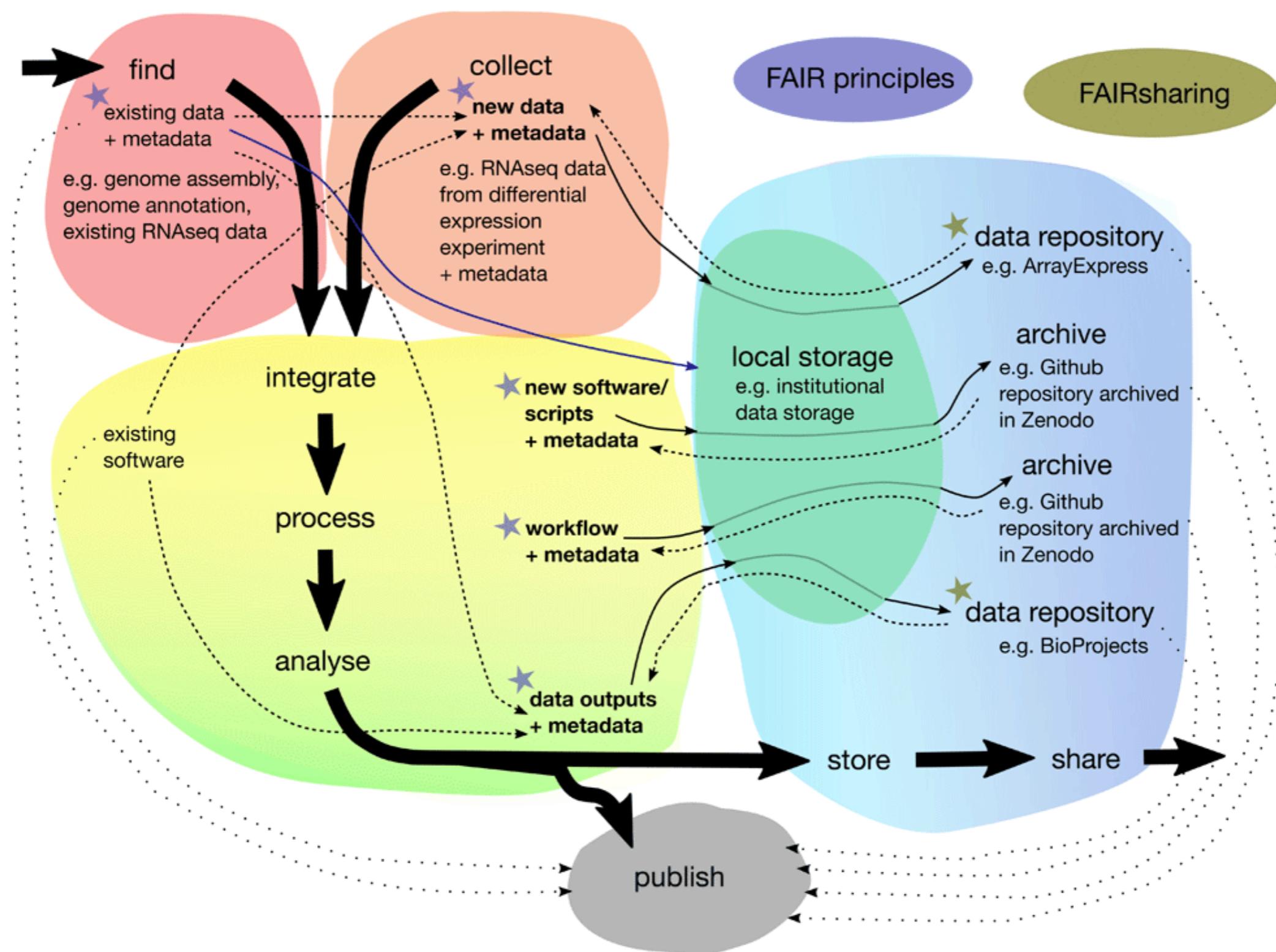
“87% of analysis time was spent on projects that had the characteristics of research...These findings suggest that to foster team-based multidisciplinary research, institutions must adopt policies that recognise contributions to research by applied bioinformatics scientists.”

– Jeffrey Chang, 2019

Practicalities: key aspects



A short intro to how we work & reproducibility



* DATA



Sharing data with a statistician / bioinformatician

- ❖ The raw data
- ❖ The tidy / clean dataset
- ❖ A code book describing each variable and its values in the tidy data set
- ❖ An explicit and exact recipe to go from 1 to 2 and 3

* DATA



Sharing data with a statistician / bioinformatician

- ❖ The raw data
- ❖ The tidy / clean dataset
- ❖ A code book describing each variable and its values in the tidy data set
- ❖ An explicit and exact recipe to go from 1 to 2 and 3

Expecting back

- ❖ An analysis script that performs each of the analyses
- ❖ The exact computer code
- ❖ All output files and figures generated

Read more: <https://github.com/jtleek/datasharing>

* Bioinformatics File Formats

* Bioinformatics File Formats

FASTA

*.fa, *fasta, *.sa

A simple way to represent nucleotide or amino acid sequences of nucleic acids and proteins; 2 lines per entry

```
>XR_002086427.1 Candida albicans SC5314 uncharacterized ncRNA (SCR1), ncRNA
```

```
TGGCTGTGATGGCTTTAGCGGAAGCGCGCTGTCGCGTACCTGCTGTTGAAATTAAAGAGCAAAGTGTCCGGCTCGATCCCTGCGAATTGAATTCTGAACGCTAGAGT  
AATCAGTGTCTTCAAGTTCTGTAATGTTAGCATAACCACTGGAGGGAAGCAATTCAAGCACAGTAATGCTAATCGTGGTGGAGGCGAACCGGATGGCACCTGTTGA  
TAAATAGTGCAGGTATCTAGTGTGCAACTCTATTTT
```

* Bioinformatics File Formats

FASTA

*.fa, *fasta, *.sa

A simple way to represent nucleotide or amino acid sequences of nucleic acids and proteins; 2 lines per entry

```
>XR_002086427.1 Candida albicans SC5314 uncharacterized ncRNA (SCR1), ncRNA
```

```
TGGCTGTGATGGCTTTAGCGGAAGCGCGCTGTCGCGTACCTGCTGTTGAAATTAAAGAGCAAAGTGTCCGGCTCGATCCCTGCGAATTGAATTCTGAACGCTAGAGT  
AATCAGTGTCTTCAAGTTCTGTAATGTTAGCATAACCACTGGAGGGAAGCAATTCAAGCACAGTAATGCTAATCGTGGTGGAGGCGAACCGGATGGCACCTGTTGA  
TAAATAGTGCAGGTATCTAGTGTTGCAACTCTATTTT
```

FASTQ

*.fastq, *fq, *sanfastq

Puts together sequence and its quality score Q; 4 lines per entry

```
@K00188:208:HFLNGBBXX:3:1101:1428:1508 2:N:0:CTTGTA  
ATAATAGGATCCCTTCTGGAGCTGCCTTAGGTAATGTAGTATCTNATNGACTGNCCANANGGCTAAAGT  
+  
AAAFFJJJJJJJJJJJJJJJJFJJFJJJJJJJJJJJJJJJJ#FJ#JJJJF#F#FJJ#F#JJFJJJJ
```

SAM / BAM *.sam, *.bam

Sequence Alignment Map, generated following mapping of the reads to reference sequence; BAM a binary equivalent; header lines (@) followed by 1 line per entry

Example :

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ
1:497:R:-272+13M17D24M	113	1	497	37	37M	15	100338662	0	CGGGTCTGACCTGAGGAG
AACTGTGCTCCGCCTTCAG	0;====9;>>>>>>>>>=>>>>>>					XT:A:U	NM:i:0	SM:i:37	AM:i:0 X0:i:1 X1:i:0 XM
:i:0 X0:i:0 XG:i:0 MD:Z:37									
19:20389:F:275+18M2D19M	99	1	17644	0	37M	=	17919	314	TATGACTGCTAATAATACCTACACAT
GTTAGAACCAT	>>>>>>>>>>>><>>><>>>4:>>:<9				RG:Z:UM0098:1	XT:A:R	NM:i:0	SM:i:0 AM:i:0 X0:i:4 X1	
:i:0 XM:i:0 X0:i:0 XG:i:0 MD:Z:37									
19:20389:F:275+18M2D19M	147	1	17919	0	18M2D19M	=	17644	-314	GTAGTACCAACTGTAAGT

SAM / BAM

*.sam, *.bam

Sequence Alignment Map, generated following mapping of the reads to reference sequence; BAM a binary equivalent; header lines (@) followed by 1 line per entry

Example :

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ
1:497:R:-272+13M17D24M	113	1	497	37	37M	15	100338662	0	CGGGTCTGACCTGAGGAG
AACTGTGCTCCGCCTTCAG	0;====9;>>>>>>>>>=>>>>>>					XT:A:U	NM:i:0	SM:i:37	AM:i:0 X0:i:1 X1:i:0 XM
:i:0 X0:i:0 XG:i:0 MD:Z:37									
19:20389:F:275+18M2D19M	99	1	17644	0	37M	=	17919	314	TATGACTGCTAATAATACCTACACAT
GTTAGAACCAT	>>>>>>>>>>>><>><>>4:>>:<9				RG:Z:UM0098:1	XT:A:R	NM:i:0	SM:i:0 AM:i:0 X0:i:4 X1	
:i:0 XM:i:0 X0:i:0 XG:i:0 MD:Z:37									
19:20389:F:275+18M2D19M	147	1	17919	0	18M2D19M	=	17644	-314	GTAGTACCAACTGTAAGT

VCF

*.vcf

Variant Calling Format/File, used to store gene sequence variations, header lines (##) followed by 1 liner per entry

```
##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 2/2:3
5:4
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
```

GFF/GTF *.gff, *.gff2, *.gff3, *.gtf

General Feature Format / Gene Transfer Format, used for describing genes and other features of DNA, RNA and protein sequences

GTF

```
1 transcribed_unprocessed_pseudogene gene      11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "1 processed_transcript transcript 11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "DDX11L1"; gene
```

GFF

```
X Ensembl Repeat 2419108 2419128 42 . . hid=trf; hstart=1; hend=21
X Ensembl Repeat 2419108 2419410 2502 - . hid=AluSx; hstart=1; hend=303
X Ensembl Repeat 2419108 2419128 0 . . hid=dust; hstart=2419108; hend=2419128
X Ensembl Pred.trans. 2416676 2418760 450.19 - 2 genscan=GENSCAN00000019335
X Ensembl Variation 2413425 2413425 . + .
X Ensembl Variation 2413805 2413805 . + .
```

Fields

Fields must be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '!'.

1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note:** the seqname must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Let's discuss: share your experiences in any of the following

- ❖ locating data
- ❖ describing data (metadata)
- ❖ coding data (continuous, ordinal, categorical, missing, censored)
- ❖ sharing data



* Unix-like operating systems

Unix, 1960s, Bell Labs

“Unix philosophy” of creating small, modular utilities that do one thing and do them well.

Commonly used for: working with files and directories; checking file sizes, previewing data, summary statistics

```
# List directory content
[olga@rackham3 Fastq]$ ls -lh
total 23G [REDACTED]
-rw-rw---- 1 5.8G Sep 10 09:18 P12516_101_R1.fastq
-rw-rw---- 1 5.8G Sep 10 09:18 P12516_101_R2.fastq
-rw-rw---- 1 5.4G Sep 10 09:18 P12516_102_R1.fastq
-rw-rw---- 1 5.4G Sep 10 09:18 P12516_102_R2.fastq
```

```
# Output the first part of files
head P12516_101_R1.fastq [REDACTED]
@A00187:119:H72F7DRXX:2:1101:1072:1000 1:N:0:ATTACTCG
CAATGTTCTGCATGGTTATCGATCCGGAGGCTGCTTAGCTTCCAGCCAGAC
+
FFFFFFFFFFFFFFFF::FFFFFFFFFFFFFFFFFFFFFF
```

* Unix-like operating systems

Unix, 1960s, Bell Labs

“Unix philosophy” of creating small, modular utilities that do one thing and do them well.

Commonly used for: working with files and directories; checking file sizes, previewing data, summary statistics

```
# List directory content
[olga@rackham3 Fastq]$ ls -lh
total 23G [REDACTED]
-rw-rw---- 1 5.8G Sep 10 09:18 P12516_101_R1.fastq
-rw-rw---- 1 5.8G Sep 10 09:18 P12516_101_R2.fastq
-rw-rw---- 1 5.4G Sep 10 09:18 P12516_102_R1.fastq
-rw-rw---- 1 5.4G Sep 10 09:18 P12516_102_R2.fastq
```

```
# Output the first part of files
head P12516_101_R1.fastq [REDACTED]
@A00187:119:H72F7DRXX:2:1101:1072:1000 1:N:0:ATTACTCG
CAATGTTCTGCATGGTTATCGATCCGGAGGCTGCTTCCAGCCAGAC
+ [REDACTED]
FFFFFFFFFFFFFFFFFF::FFFFFFFFFFFFFFFFFFFFF
```

* High-performance computers & large-scale storage



Tetralith, NSC, Linköping University

e.g. <https://www.uppmx.uu.se>

UPPMAX (*Uppsala Multidisciplinary Center for Advanced Computational Science*) is Uppsala University's resource of high-performance computers, large-scale storage and know-how of high-performance computing (HPC)

Commonly used as: computational infrastructure and project/temporary data storage

* Bioinformatics tools

e.g. FastQC, SAMtools, BEDTools

```
# Use samtools to generate stats  
samtools stats example.sam
```

```
# Sort bed file  
bedutils sort example.bed
```

```
# Mapping reads to a reference  
star --genomeDir /proj/uppstore2019092/NBIS/Index/Star --readFilesIn /proj/snici2019-8-218/private/NBIS/Cutadapt/v03/Out/FC5_9.trim8.fastq.gz --readFilesCommand zcat --outFileNamePrefix /proj/uppstore2019092/NBIS/Star/Out/v03/FC5_9/FC5_9_ --runThreadN 4 --outFilterMultimapNmax 999 --outFilterMismatchNmax 3 --outSAMtype SAM --outSAMunmapped Within --outFilterMultimapScoreRange 1 --seedSearchStartLmax 15
```

* Bioinformatics tools

e.g. FastQC, SAMtools, BEDTools

```
# Use samtools to generate stats
samtools stats example.sam
```

```
# Sort bed file
bedutils sort example.bed
```

```
# Mapping reads to a reference
star --genomeDir /proj/uppstore2019092/NBIS/Index/Star --readFilesIn /proj/snac2019-8-218/private/NBIS/Cutadapt/v03/0ut/FC5_9.trim8.fastq.gz --readFilesCommand zcat --outFileNamePrefix /proj/uppstore2019092/NBIS/Star/0ut/v03/FC5_9/FC5_9_ --runThreadN 4 --outFilterMultimapNmax 999 --outFilterMismatchNmax 3 --outSAMtype SAM --outSAMunmapped Within --outFilterMultimapScoreRange 1 --seedSearchStartLmax 15
```



bio.tools

[Database \(Oxford\)](#). 2014; 2014: bau069.
Published online 2014 Jul 14. doi: [10.1093/database/bau069](https://doi.org/10.1093/database/bau069)

PMCID: PMC4095679
PMID: [25024350](https://pubmed.ncbi.nlm.nih.gov/25024350/)

OMICtools: an informative directory for multi-omic data analysis

Vincent J. Henry,¹ Anita E. Bandrowski,² Anne-Sophie Pepin,³ Bruno J. Gonzalez,¹ and Arnaud Desfeux^{3,*}

► Author information ► Article notes ► Copyright and License information [Disclaimer](#)

This article has been [cited by](#) other articles in PMC.

Abstract

Go to:

Recent advances in ‘omic’ technologies have created unprecedented opportunities for biological research, but current software and database resources are extremely fragmented. OMICtools is a manually curated metadatabase that provides an overview of more than 4400 web-accessible tools related to genomics, transcriptomics, proteomics and metabolomics. All tools have been classified by omic technologies (next-generation sequencing, microarray, mass spectrometry and nuclear magnetic resonance) associated with published evaluations of tool performance. Information about each tool is derived either from a diverse set of developers, the scientific literature or from spontaneous submissions. OMICtools is expected to serve as a useful didactic resource not only for bioinformaticians but also for experimental researchers and clinicians.

Database URL: <http://omictools.com/>

<https://omictools.com>

[Sci Data](#). 2018; 5: 180023.
Published online 2018 Feb 27. doi: [10.1038/sdata.2018.23](https://doi.org/10.1038/sdata.2018.23)
Article

PMCID: PMC5827688
PMID: [29485625](https://pubmed.ncbi.nlm.nih.gov/29485625/)

Datasets2Tools, repository and search engine for bioinformatics datasets, tools and canned analyses

Denis Torre,¹ Patrycja Krawczuk,¹ Kathleen M. Jagodnik,¹ Alexander Lachmann,¹ Zichen Wang,¹ Lily Wang,¹ Maxim V. Kuleshov,¹ and Avi Ma'ayan^{a,1}

► Author information ► Article notes ► Copyright and License information [Disclaimer](#)

This article has been [cited by](#) other articles in PMC.

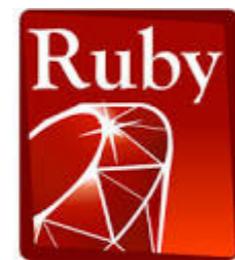
Abstract

Go to:

Biomedical data repositories such as the Gene Expression Omnibus (GEO) enable the search and discovery of relevant biomedical digital data objects. Similarly, resources such as OMICtools, index bioinformatics tools that can extract knowledge from these digital data objects. However, systematic access to pre-generated ‘canned’ analyses applied by bioinformatics tools to biomedical digital data objects is currently not available. Datasets2Tools is a repository indexing 31,473 canned bioinformatics analyses applied to 6,431 datasets. The Datasets2Tools repository also contains the indexing of 4,901 published bioinformatics software tools, and all the analyzed datasets. Datasets2Tools enables users to rapidly find datasets, tools, and canned analyses through an intuitive web interface, a Google Chrome extension, and an API. Furthermore, Datasets2Tools provides a platform for contributing canned analyses, datasets, and tools, as well as evaluating these digital objects according to their compliance with the findable, accessible, interoperable, and reusable (FAIR) principles. By incorporating community engagement, Datasets2Tools promotes sharing of digital resources to stimulate the extraction of knowledge from biomedical research data. Datasets2Tools is freely available from: <http://amp.pharm.mssm.edu/datasets2tools>.

<http://amp.pharm.mssm.edu/datasets2tools/>

* Scripts and programs putting commands together



C++



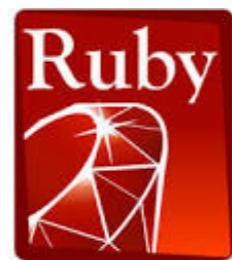
* Scripts and programs putting commands together



BASH
THE BOURNE-AGAIN SHELL



Main three (interpreted)



C++



THE
C
PROGRAMMING
LANGUAGE



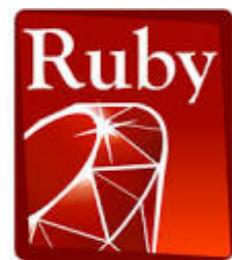
* Scripts and programs putting commands together



BASH
THE BOURNE-AGAIN SHELL



Main three (interpreted)



C++



THE
C
PROGRAMMING
LANGUAGE



Compiled languages

* Scripts and programs putting commands together



BASH
THE BOURNE-AGAIN SHELL



Main three (interpreted)



C++



THE
C
PROGRAMMING
LANGUAGE



Compiled languages

Website development

* Scripts and programs putting commands together



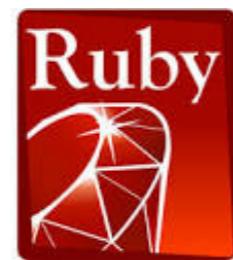
BASH
THE BOURNE-AGAIN SHELL



Main three (interpreted)



perl
Programming Language



Ruby



Java
Programming

C++



php



JavaScript

Website development

Compiled languages

open-source initiatives



biopython



BioPerl

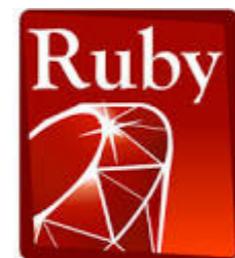
* Scripts and programs putting commands together



BASH
THE BOURNE-AGAIN SHELL



Main three (interpreted)



C++



THE
C
PROGRAMMING
LANGUAGE



Website development

Compiled languages

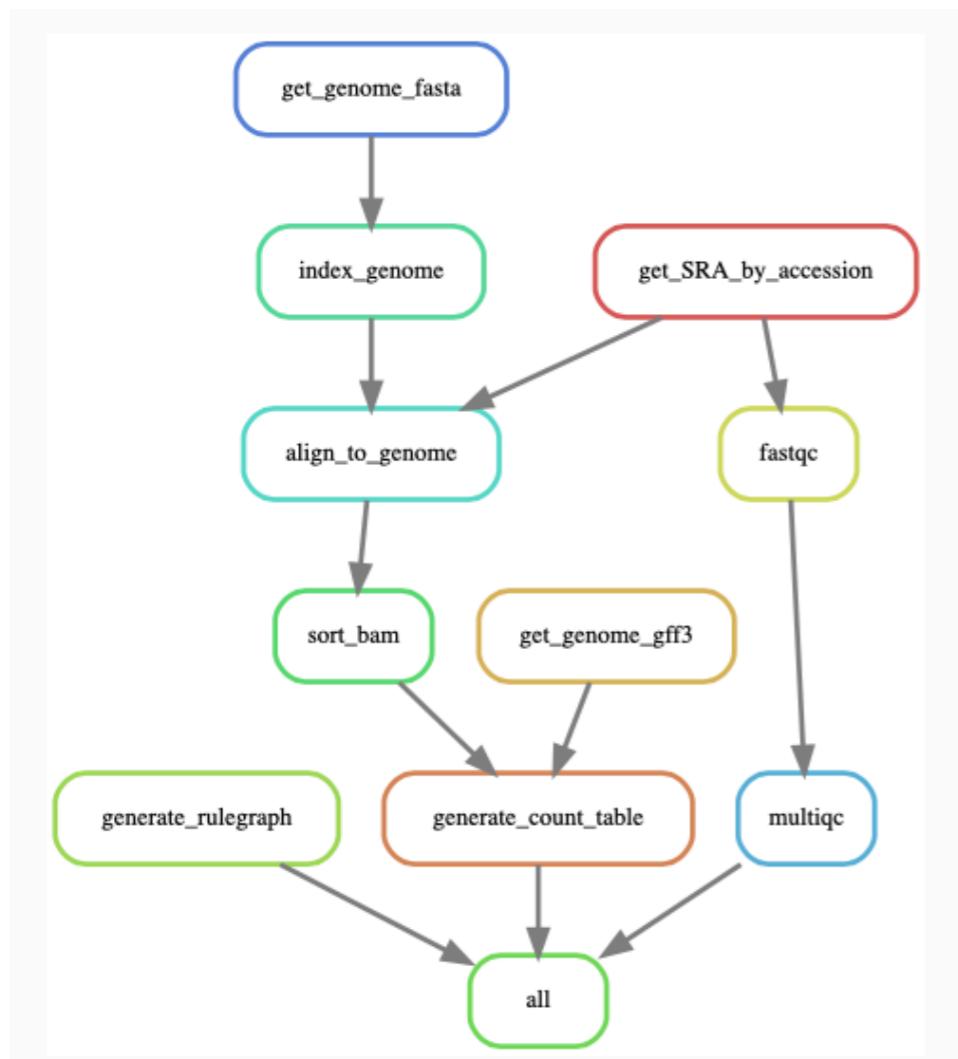
open-source initiatives

- ❖ Interpreted vs. compiled languages
- ❖ Script vs. program vs. tool vs. software

Read more:
<http://omgenomics.com/programming-languages/>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2267699/>

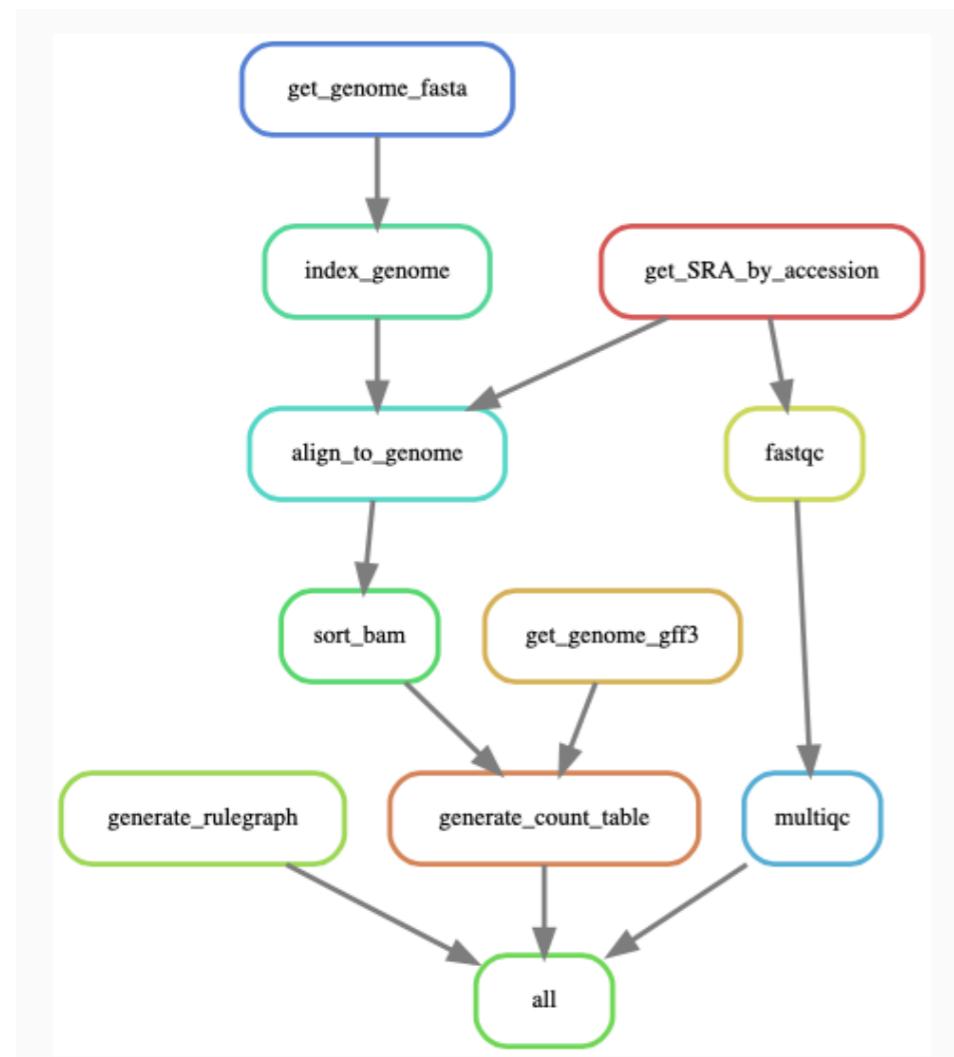
* Workflows

automating tasks
tracking provenance



* Workflows

automating tasks
tracking provenance



TOOLBOX · 02 SEPTEMBER 2019

Workflow systems turn raw data into scientific knowledge

How workflow tools can make your computational methods portable, maintainable, reproducible and shareable.



Snakemake

nextflow

Galaxy

* Literate computing & authoring

writing self-contained documents that include narrative and code used to generate both text and graphical results

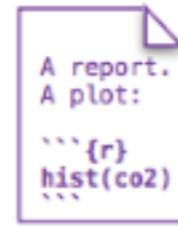
i. **Open** - Open a file that uses the .Rmd extension.



ii. **Write** - Write content with the easy to use R Markdown syntax



iii. **Embed** - Embed R code that creates output to include in the report



iv. **Render** - Replace R code with its output and transform the report into a slideshow, pdf, html or ms Word file.



* Literate computing & authoring

writing self-contained documents that include narrative and code used to generate both text and graphical results

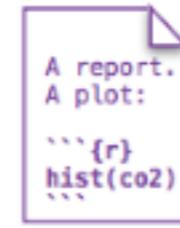
i. Open - Open a file that uses the .Rmd extension.



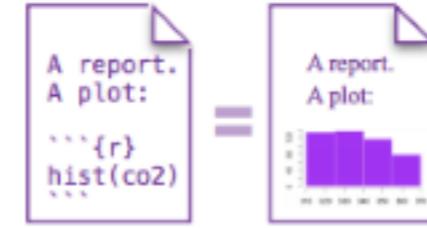
ii. Write - Write content with the easy to use R Markdown syntax



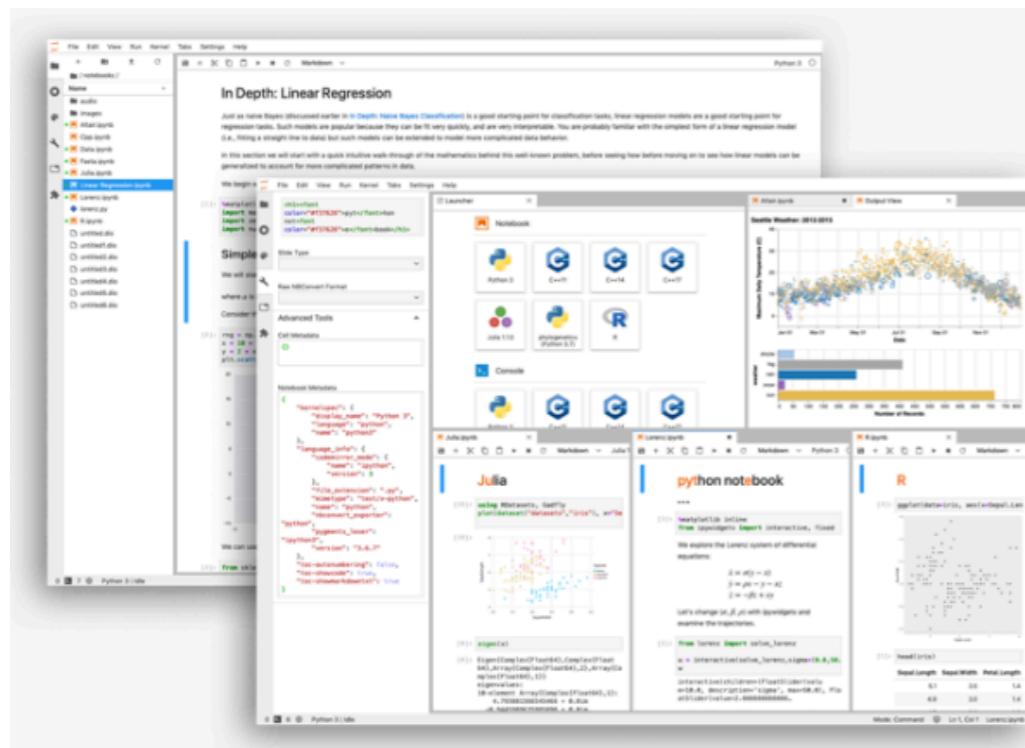
iii. Embed - Embed R code that creates output to include in the report



iv. Render - Replace R code with its output and transform the report into a slideshow, pdf, html or ms Word file.



The LATEX Project



JupyterLab 1.0: Jupyter's Next-Generation Notebook Interface

JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. JupyterLab is extensible and modular: write plugins that add new components and integrate with existing ones.

Try it in your browser

Install JupyterLab



Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

* Version control

keeping a record of file changes over time
collaborating on a code development

The screenshot shows the official Git website. At the top left is the Git logo (a red diamond with a white 'g'). To its right is the word "git" in a large, bold, black sans-serif font. Below the logo is the tagline "--distributed-even-if-your-workflow-isnt". To the right of the tagline is a search bar with a magnifying glass icon and the placeholder text "Search entire site...". The main content area contains two paragraphs. The first paragraph describes Git as a "free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency." The second paragraph highlights Git's ease of learning, fast performance, and various features like local branching, staging areas, and multiple workflows. To the right of the text is a diagram illustrating Git's distributed nature. It shows six white server icons, each containing a stack of four files with colored bars (green, blue, red, yellow). These servers are interconnected by a network of colored lines: a teal line connects the top-left and middle-left servers; a red line connects the middle-left and middle-right servers; a blue line connects the middle-right and bottom-right servers; a yellow line connects the bottom-right and bottom-left servers; and a red line connects the bottom-left and top-left servers. This visualizes how multiple repositories can exist independently while being connected through a shared network.

<https://git-scm.com>

Git repository hosting services



* Capturing the computational environment



Package and environment manager

- ❖ As a package it enables installing a wide range of tools using one command `conda install`
- ❖ As an environment manager it allows creating and managing multiple different environments, e.g. different versions of the same package

* Capturing the computational environment



Package and environment manager

- ❖ As a package it enables installing a wide range of tools using one command `conda install`
- ❖ As an environment manager it allows creating and managing multiple different environments, e.g. different versions of the same package



Containers

- ❖ Full control of environment. Can be used to package entire scientific workflows, software, libraries and data, by isolating everything in a “container”

Summary



Team work and project planning

- ❖ align on a common goal
- ❖ value team members
- ❖ communicate

Summary



Team work and project planning

- ❖ align on a common goal
- ❖ value team members
- ❖ communicate



Data sharing & bioinformatics

- ❖ planning ahead when collecting and describing data
- ❖ a min. knowledge of unix skills and bioinformatics terms

Summary



Team work and project planning

- ❖ align on a common goal
- ❖ value team members
- ❖ communicate



Data sharing & bioinformatics

- ❖ planning ahead when collecting and describing data
- ❖ a min. knowledge of unix skills and bioinformatics terms



Planning ahead:

- ❖ reproducible research
- ❖ reproducible publications
- ❖ sustainability and long-term growth

“Good collaboration checklist”

[https://docs.google.com/document/d/
1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing](https://docs.google.com/document/d/1MpefgEukIXooeo86Q5bX_wjD4RQxMHT9OIY1jEQA2q4/edit?usp=sharing)

Thank you!

