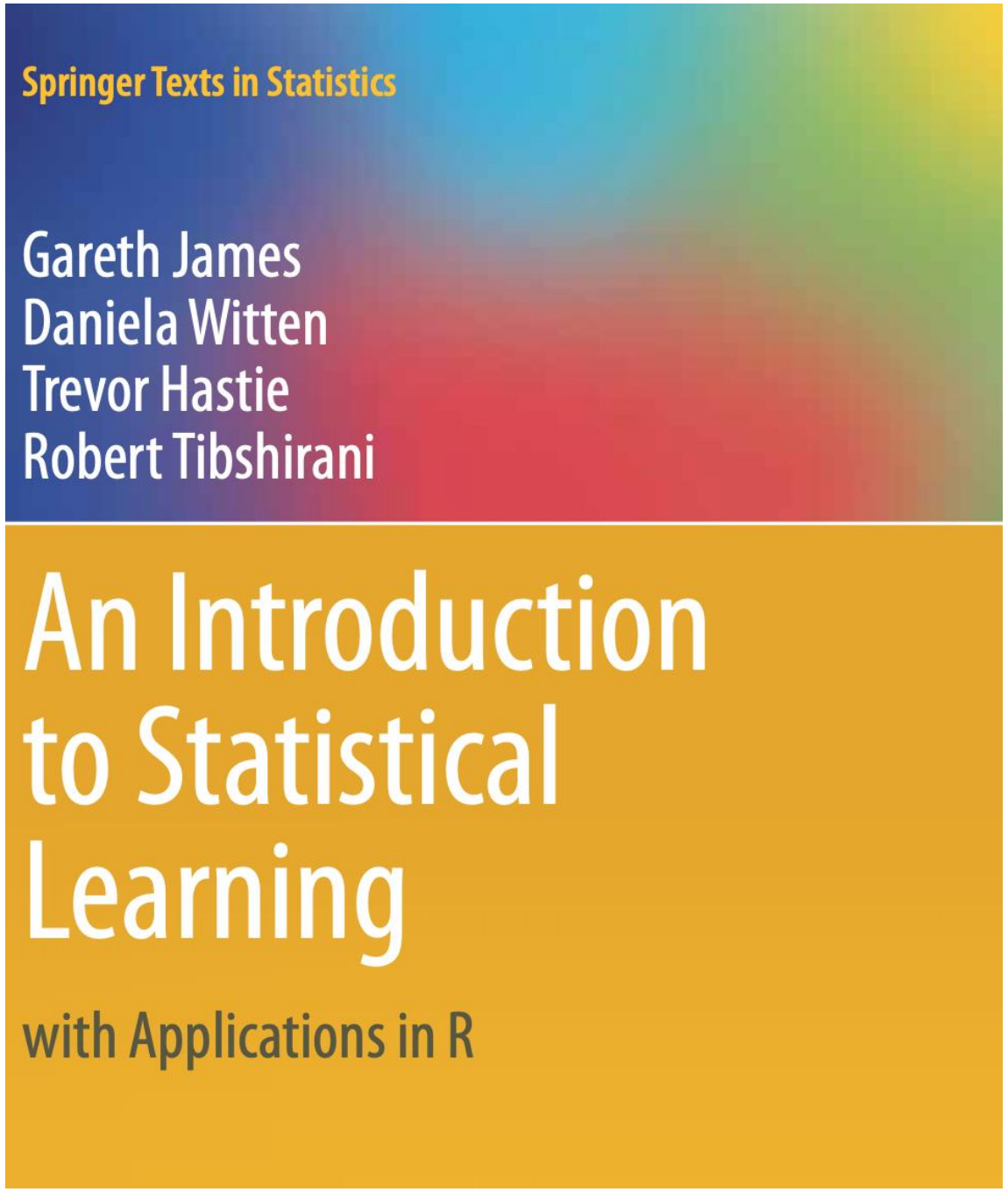
Data Handling and Reproducible Research



ISL is based on the following four premises.

1. *Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.* We believe that many contemporary statistical learning procedures should, and will, become as widely available and used as is currently the case for classical methods such as linear regression. As a result, rather than attempting to consider every possible approach (an impossible task), we have concentrated on presenting the methods that we believe are most widely applicable.

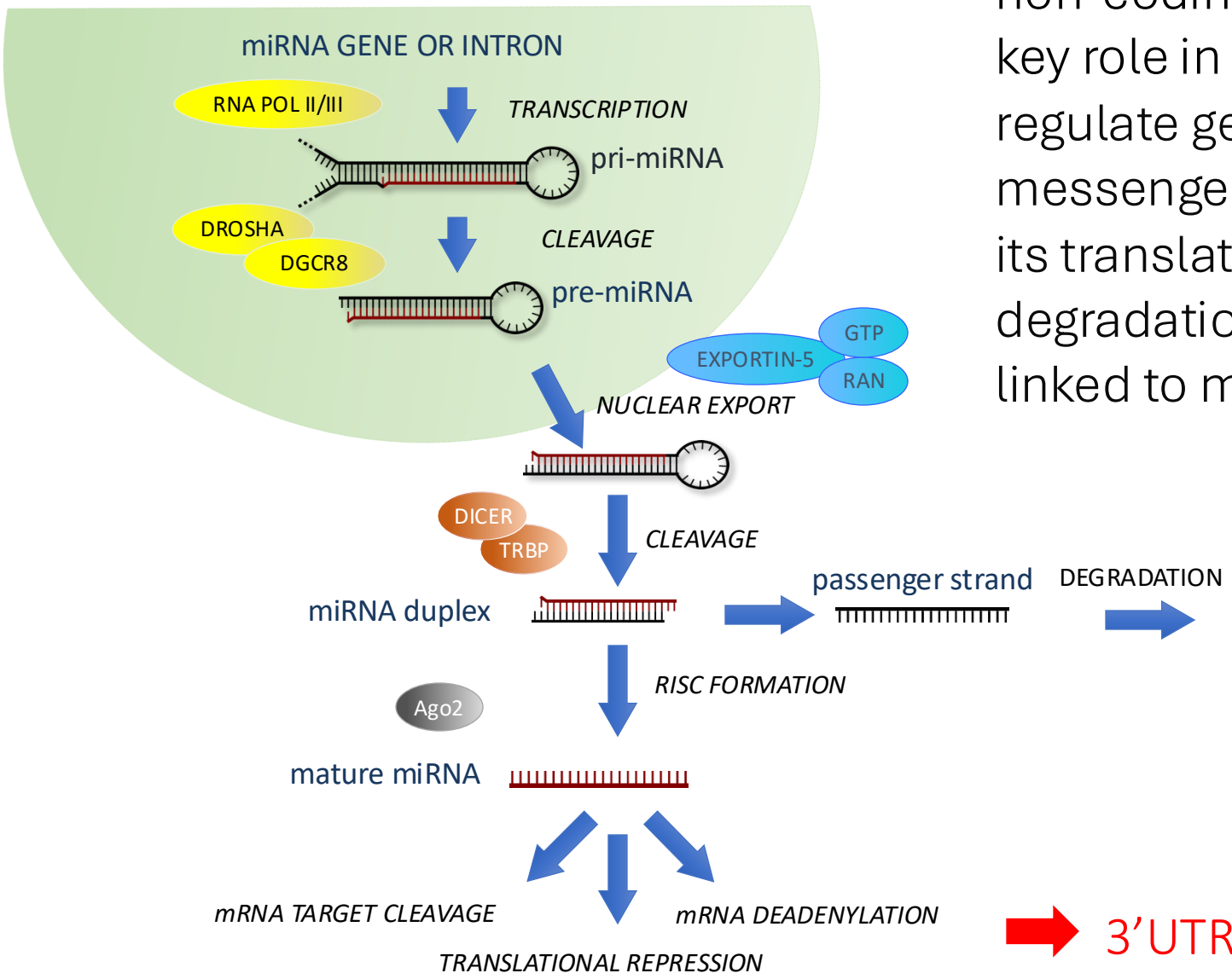
2. *Statistical learning should not be viewed as a series of black boxes.* No single approach will perform well in all possible applications. Without understanding all of the cogs inside the box, or the interaction between those cogs, it is impossible to select the best box. Hence, we have attempted to carefully describe the model, intuition, assumptions, and trade-offs behind each of the methods that we consider.

3. *While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine inside the box!* Thus, we have minimized discussion of technical details related to fitting procedures and theoretical properties. We assume that the reader is comfortable with basic mathematical concepts, but we do not assume a graduate degree in the mathematical sciences. For instance, we have almost completely avoided the use of matrix algebra, and it is possible to understand the entire book without a detailed knowledge of matrices and vectors.

Data

Results

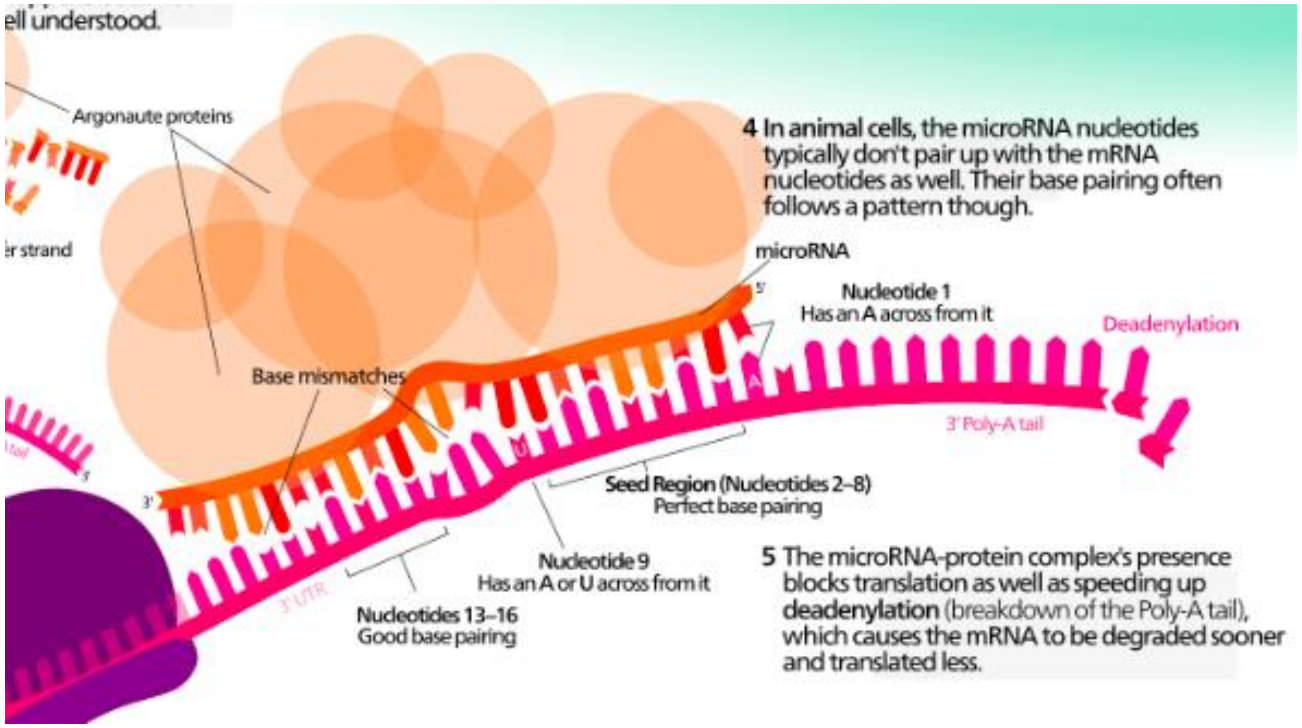
microRNAs (miRNAs)

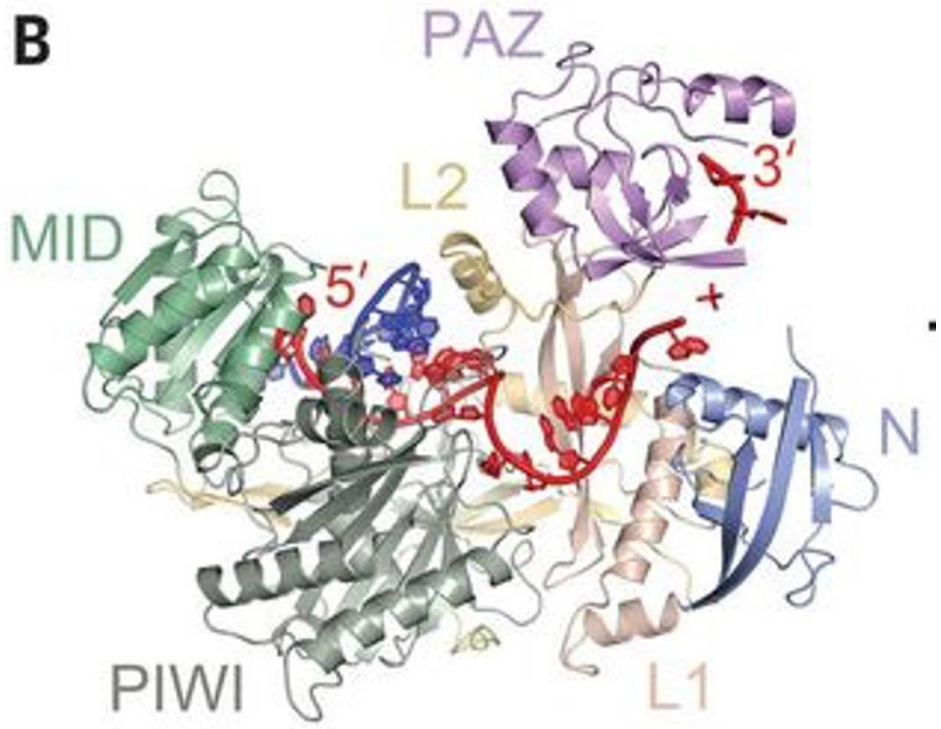


miRNAs are short, single-stranded, non-coding RNA molecules that play a key role in gene regulation. They regulate gene expression by binding to messenger RNA (mRNA) to either block its translation into a protein or cause its degradation. Their dysregulation is linked to many diseases

➡ 3'UTR TARGETING

A microRNA (abbreviated miRNA) is a small non-coding RNA molecule (containing about 22 nucleotides) found in plants, animals and some viruses, that functions in RNA silencing and post-transcriptional regulation of gene expression by binding to mRNA targets.





GC Calc

```
$ tree -d -L 2 lecture2__ReproducibleResearch
lecture2__ReproducibleResearch
├── code
├── FilePaths.Rmd
│   ├── gccalc
│   │   └── sorting
└── data
```

You probably signed up for this course looking to find out how to analyze your data

Why should you care about data handling and reproducible research?

Researchers at TAE Technologies in California and at Google are using machine learning to optimize equipment that produces a high-energy plasma.

Three pitfalls to avoid in machine learning

As scientists from myriad fields rush to perform algorithmic analyses, Google's **Patrick Riley** calls for clear standards in research and reporting.

https://doi.org/10.1038/d41586-019-02307-y


THE
ROBERT DE NIRO
DEER HUNTER
JOHN CAZALE
JOHN SAVAGE
MERYL STREEP
CHRISTOPHER WALKEN

M134 Minigun is a 7.62×51mm NATO six-barrel rotary machine gun with a high rate of fire (2,000 to 6,000 rounds per minute).

M134 Minigun is a 7.62×51mm NATO six-barrel rotary machine gun with a high rate of fire (2,000 to 6,000 rounds per minute).

	Deer	Bunnies
Hits	100 %	100 %
Misses	0 %	0 %

```
$ tree -d -L 2 lecture2__ReproducibleResearch
lecture2__ReproducibleResearch
├── code
│   ├── gccalc
│   └── sorting
└── data
```

Reproducible Research

You probably signed up for this course looking to find out how to analyze your data

Why should you care about data handling and reproducible research?

From an experimental perspective, you probably already consider reproducible research

Researchers at TAE Technologies in California and at Google are using machine learning to optimize equipment that produces a high-energy plasma.

Three pitfalls to avoid in machine learning

As scientists from myriad fields rush to perform algorithmic analyses, Google's **Patrick Riley** calls for clear standards in research and reporting.

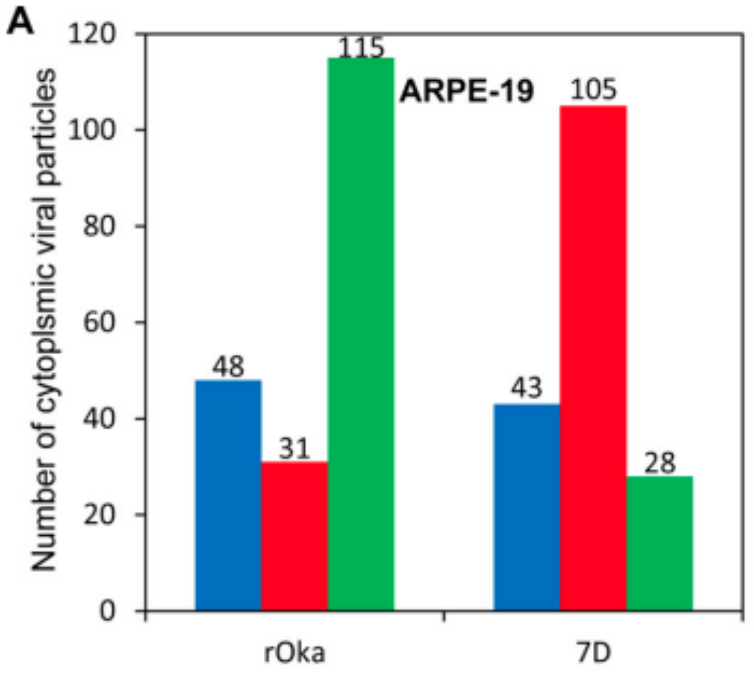
https://doi.org/10.1038/d41586-019-02307-y

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

ORF7 of Varicella-Zoster Virus Is Required for Viral Cytoplasmic Envelopment in Differentiated Neuronal Cells

Hai-Fei Jiang,[a,b] Wei Wang,[c] Xuan Jiang,[a] Wen-Bo Zeng,[a] Zhang-Zhou Shen,[a] Yi-Ge Song,[a,b] Hong Yang,[a,b] Xi-Juan Liu,[a,b] Xiao Dong,[a] Jing Zhou,[a,b] Jin-Yan Sun,[a] Fei-Long Yu,[d] Lin Guo,[d] Tong Cheng,[c] Simon Rayner,[e] Fei Zhao,[a] Hua Zhu,[f] Min-Hua Luo[a,b]

ABSTRACT Although a varicella-zoster virus (VZV) vaccine has been used for many years, the neuropathy caused by VZV infection is still a major health concern. Open reading frame 7 (ORF7) of VZV has been recognized as a neurotropic gene *in vivo*, but its neurovirulent role remains unclear. In the present study, we investiga
effect of ORF7 deletion on VZV replication cycle at virus entry, genome rep
gene expression, capsid assembly and cytoplasmic envelopment, and trans
transmission in differentiated neural progenitor cells (dNPCs) and neuroblasto
SY5Y (dSY5Y) cells. Our results demonstrate that the ORF7 protein is a compo
the tegument layer of VZV virions. Deleting ORF7 did not affect viral entry, v
nome replication, or the expression of typical viral genes but clearly impacte
plasmic envelopment of VZV capsids, resulting in a dramatic increase of en
defective particles and a decrease in intact virions. The defect was more se
differentiated neuronal cells of dNPCs and dSY5Y. ORF7 deletion also impaired trans-
mission of ORF7-deficient virus among the neuronal cells. These results indicate that ORF7 is required for cytoplasmic envelopment of VZV capsids, virus transmission among neuronal cells, and probably the neuropathy induced by VZV infection.



Complicated lab work, but simple data and analysis
(Number of viral particles, simple statistical test)

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

ORF7 of Varicella-Zoster Virus Is Required for Viral Cytoplasmic Envelopment in Differentiated Neuronal Cells

Hai-Fei Jiang,[a,b] Wei Wang,[c] Xuan Jiang,[a] Wen-Bo Zeng,[a] Zhang-Zhou Shen,[a] Yi-Ge Song,[a,b] Hong Yang,[a,b] Xi-Juan Liu,[a,b] Xiao Dong,[a] Jing Zhou,[a,b] Jin-Yan Hua Zh
State Key

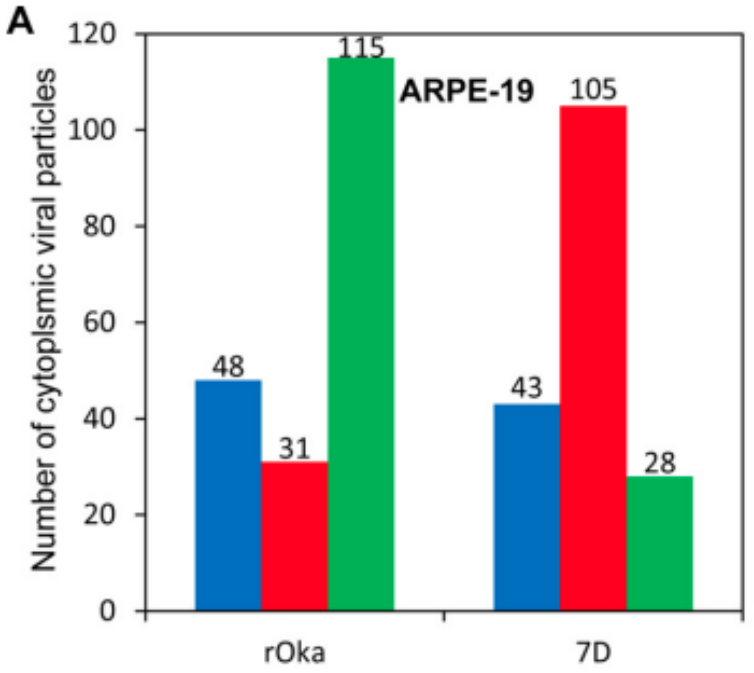
MATERIALS AND METHODS

Ethics statement. The Wuhan Institute of Virology Institutional Review Board approved (WIVH10201202) the isolation of primary human NPCs from postmortem fetal embryonic tissue and waived the need for consent. The isolated NPCs are maintained in our laboratory and complied with the rule that NPCs must be less than nine passages (49).

Cells and cell culture. ARPE-19 cells (ATCC, CRL-2302) were grown in Dulbecco modified Eagle medium (DMEM) with 10% fetal bovine serum (FBS) and penicillin-streptomycin (100 U/ml and 100 μg/ml), all from Gibco/Life Technology. NPCs were isolated from the postmortem neonate brai and cultured as described previously (21). To differentiate NPCs toward neurons, monolayer NP cultured in the presence of 25 ng/ml human basic fibroblast growth factor 2 (FGF-2), 20 ng/n growth factor (NGF), and 10 ng/ml brain-derived neurotrophic factor (BDNF), all from Prospec; dibutryl cyclic AMP (Selleck); and 1 μM retinoic acid (Sigma) for 10 days as described previously (differentiated NPCs were designated dNPCs. SH-SY5Y (designated SY5Y, ATCC, CRL-2266) ce maintained in DMEM-F12 containing 10% FBS. To differentiate SY5Y cells toward neurons, ce treated with 50 μM retinoic acid for 5 days, followed by treatment with neurotropic growth fact nM NGF and 50 nM BDNF) for 7 days (23). The differentiated SY5Y cells were designated dSY5

Viruses and infection. rOka, 7D, and 7R were all derived from the parental wild-type Oka s rOka contains a GFP coding gene and a luciferase gene in the genome and has been shown growth kinetics similar to those of the parental Oka strain (51). 7D is constructed by deleting the ORF7 gene from the rOka genome, and 7R is a revertant virus of 7D (7).

Statistical analysis. For growth curve, virus entry, qPCR, and qRT-PCR analyses, each experiment was performed in triplicate, and the results are presented as means ± the standard deviations (SD) from three independent experiments. A Student *t* test was performed to analyze the statistical significance between different virus infections. Differences were considered to be significant when $P < 0.05$.



The materials and methods provide sufficient details to allow the reader to reproduce the experiments and the statistical analysis

Here, there is a simple process to analyze the data

nature COMMUNICATIONS

ARTICLE

Received 18 Aug 2014 | Accepted 18 Nov 2014 | Published 9 Jan 2015

DOI: 10.1038/ncomms6901 OPEN

Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung[1,*], Andrea Pellagatti[2,*], Luca Malcovati[3,4], Aristoteles Giagounidis[5], Matteo G. Della Porta[3,6], Martin Jädersten[7], Hamid Dolatshad[2], Amit Verma[8], Nicholas C.P. Cross[9], Paresh Vyas[10], Sally Killick[11], Eva Hellström-Lindberg[7], Mario Cazzola[3,4], Elli Papaemmanuil[1], Peter J. Campbell[1] & Jacqueline Boultwood[2]

Methods

Samples. The study was approved by the ethics committees (Oxford C00.196, Bournemouth 9991/03/E, Duisburg 2283/03, Stockholm 410/03, Pavia 26264/2002) and informed patient consent was obtained. A total of 159 MDS samples and 17 healthy controls were studied; no samples were excluded in the statistical analysis. Gene expression data of 43 bone marrow samples from MDS patients without published expression data were obtained using the protocol described in ref. 12. In brief, CD34 + cells were enriched from mononuclear cells using CD34 MicroBeads (Miltenyi Biotec, Bergisch Gladbach, Germany). RNA was extracted using TRIZOL (Invitrogen, Paisley, UK). Fifty nanograms of RNA was subsequently amplified and biotin labelled. A total of 10 μg of labelled cRNA was hybridized to Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays (Affymetrix, Santa Clara, CA, USA), which were scanned on an Affymetrix GeneChip Scanner 3000 (ref. 12).

This provides us with some basic information about how the microarray experiments were carried out. Importantly, it also points us to more detailed protocols in reference 12

11. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–3627 (2013).
12. Pellagatti, A. *et al.* Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells. *Leukemia* **24**, 756–764 (2010).
13. Mills, K. I. *et al.* Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood* **114**, 1063–1072 (2009).

Leukemia (2010) 24, 756–764

www.nature.com/leu

ORIGINAL ARTICLE

Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells

A Pellagatti[1], M Cazzola[2], A Giagounidis[3], J Perry[1], L Malcovati[2], MG Della Porta[2], M Jädersten[4], S Killick[5], A Verma[6], CJ Norbury[7], E Hellström-Lindberg[4], JS Wainscoat[1] and J Boultwood[1]

Materials and methods

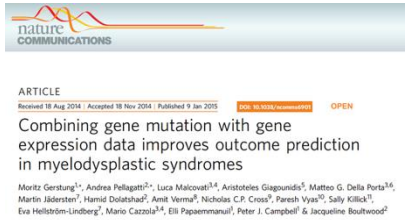
Sample collection and cell separation

A total of 183 patients with MDS and 17 healthy controls were included in the study. Classification of MDS patients was according to the French–American–British criteria,[5] with RAEB patients further subdivided into RAEB1 and RAEB2. At the time of sample, 55 patients had RA, 48 RARS, 37 RAEB1 and 43 RAEB2. The MDS patient samples were collected from several centers: Oxford and Bournemouth (UK), Duisburg (Germany), Stockholm (Sweden) and Pavia (Italy). The study was approved by the ethics committees (Oxford C00.196, Bournemouth 9991/03/E, Duisburg 2283/03, Stockholm 410/03, Pavia 26264/2002) and informed consent was obtained. Bone marrow samples were obtained and CD34+ cells isolated from MDS patients and healthy controls. Mononuclear cells were separated using Histopaque (Sigma-Aldrich, Gillingham, UK) density gradient centrifugation, labeled with CD34 MicroBeads, and then CD34+ cells were isolated using MACS magnetic cell separation columns (Miltenyi Biotec, Bergisch Gladbach, Germany) according to the manufacturer's recommendations. The purity of CD34+ cell preparations was evaluated with FACS and was ⩾90%.

Affymetrix experiments

Total RNA was extracted using TRIZOL (Invitrogen, Paisley, UK) following the manufacturer's protocol. The quality of the RNA samples was evaluated using Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). For each sample, 50 ng of total RNA were amplified and labeled with the Two-Cycle cDNA Synthesis and the Two-Cycle Target Labelling and Control Reagent packages (Affymetrix, Santa Clara, CA, USA). 10 µg of biotin-labeled fragmented cRNA was hybridized to GeneChip Human Genome U133 Plus 2.0 arrays (Affymetrix), covering over 47 000 transcripts representing 39 000 human genes. Hybridization was performed at 45 °C for 16 h in Hybridization Oven 640 (Affymetrix). Chips were washed and stained in a Fluidics Station 450 (Affymetrix) and scanned using a GeneChip Scanner 3000 (Affymetrix).

This gives a reasonably good overview and should allow us to repeat the experiment



But the Gerstung paper is more than the experiments, it’s about analysis and interpretation of the data.

In this sense, it’s a different perspective from a small-scale biology experiment

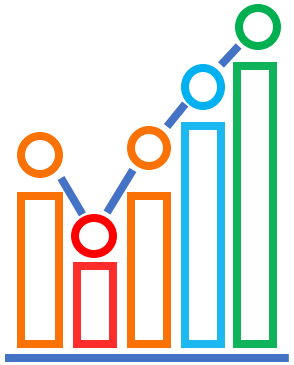
So, the analysis Gerstung paper is more complicated

DATA

THE DATA CAME FROM MANY DIFFERENT PLACES

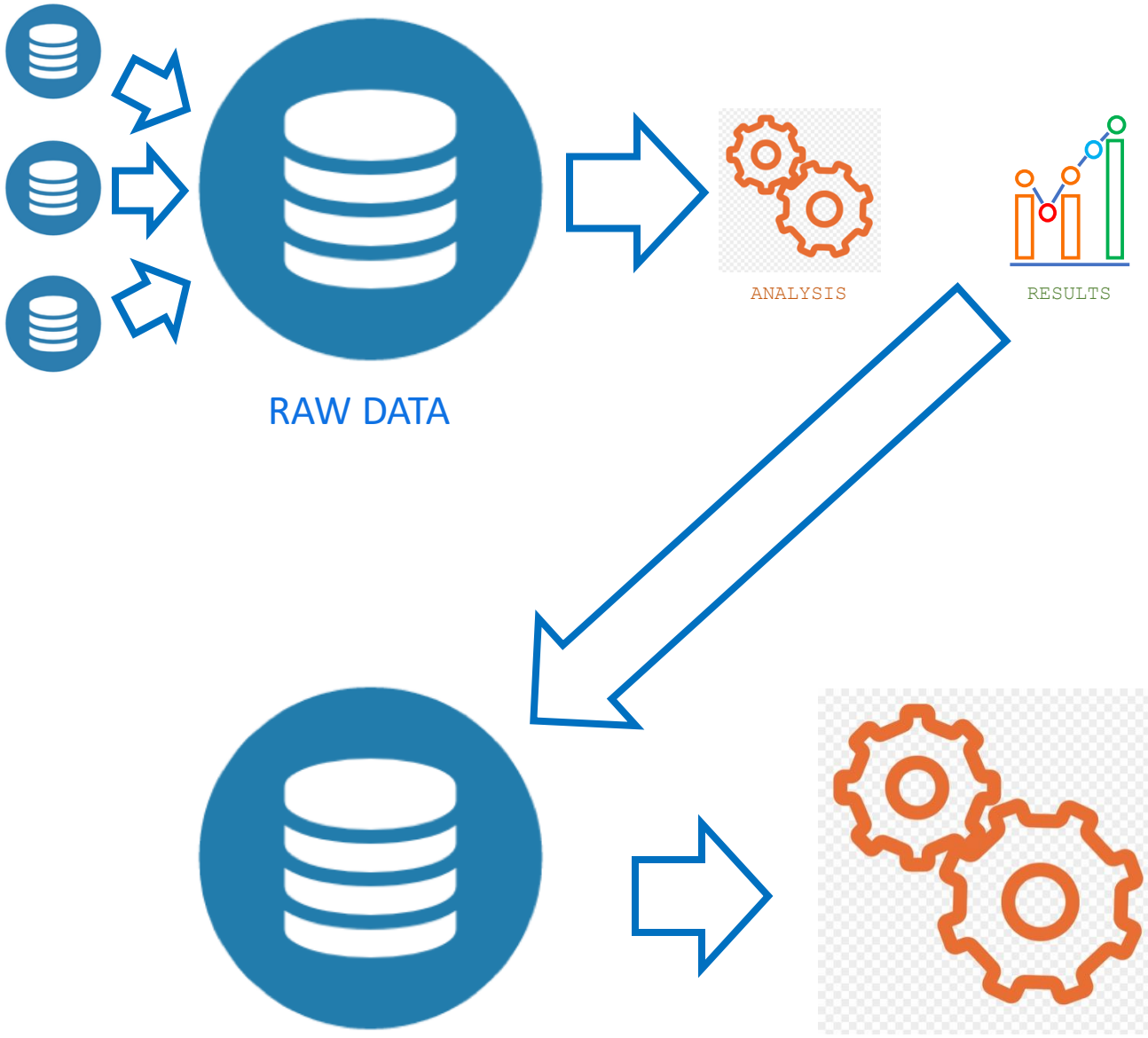
ANALYSIS

THERE ARE MANY DIFFERENT ANALYSES

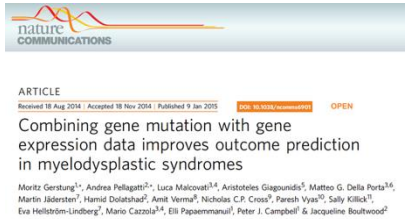


RESULTS

THERE ARE MULTIPLE RESULTS AND THERE ARE LOTS OF THEM

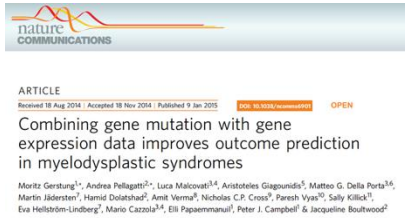


In reality it's even more complicated as the data we are working with isn't the raw data



The Gerstung paper describes the analysis steps by providing the R code in a Markdown document

```
library(limma)
library(org.Hs.eg.db)

library(RColorBrewer)
library(AnnotationDbi)
library(affy)
library(gcrma)
library(hgu133plus2.db )

library(VennDiagram)

library(org.Hs.eg.db)
library(GenomicRanges)

library(GenomicFeatures)
library(rtracklayer)
library(biomaRt)
library(glmnet)

library(survival)

library(Hmisc)

library(randomForestSRC)

set1 = c(brewer.pal(9,"Set1"), brewer.pal(8, "Dark2")) source("suppData/mg14.R")
celFiles <- dir("GSE58831", pattern = ".CEL", full.names = T) celFiles
affyBatch <- read.affybatch(filenames = celFiles)
gset = gcrma(affyBatch)

samples = sub("_.+","", sampleNames(gset))
sampleNames(gset) = samples

#Now merge probes to genes by the means of all probes mapping to a particular entrez id
tab <- select(hgu133plus2.db, keys = keys(hgu133plus2.db), columns = c("ENTREZID"))
```



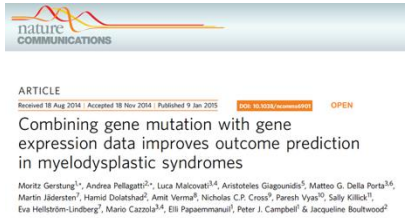
The last piece of the puzzle is the integration of the data

For the Gerstung data, this is a combination of clinical data and microarray data.

The microarray data is stored in the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/)

Unfortunately, there isn't a straightforward way to search using R, so we have to do it manually

We can get the GEO accession number from the paper


nature COMMUNICATIONS

ARTICLE

Received 18 Aug 2014 | Accepted 18 Nov 2014 | Published 9 Jan 2015 DOI: 10.1038/ncomms6901 OPEN

Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung[1,*], Andrea Pellagatti[2,*], Luca Malcovati[3,4], Aristoteles Giagounidis[5], Matteo G. Della Porta[3,6], Martin Jädersten[7], Hamid Dolatshad[2], Amit Verma[8], Nicholas C.P. Cross[9], Paresh Vyas[10], Sally Killick[11], Eva Hellström-Lindberg[7], Mario Cazzola[3,4], Elli Papaemmanuil[1], Peter J. Campbell[1] & Jacqueline Boultwood[2]

P2 top right

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms6901

patients and 17 normal individuals in total. Combined expression and mutation data were available for 124/159 MDS patients (Table 1; Supplementary Data 1). Outcome data were also available and are released with the gene expression data (GEO accession GSE58831). In addition, we release all code associated with implementing the detailed statistical analyses that follow (Supplementary Data 2), in order that this study can be replicated on this data set and extended to other tumour types.

We can look this up in the Gene Expression Omnibus


NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE58831 GO

Series GSE58831 Query DataSets for GSE58831

Status	Public on Jan 06, 2015
Title	Gene expression data from bone marrow CD34+ cells of patients with myelodysplastic syndromes (MDS) and healthy controls
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	We aimed to determine the impact of the common mutations on the transcriptome in myelodysplastic syndromes (MDS). We linked genomic data with gene expression microarray data and we deconvoluted the expression of genes into contributions stemming from each genetic and cytogenetic alteration, providing insights into how driver mutations interfere with the transcriptomic state. We modelled the influence of mutations and expression changes on diagnostic clinical variables as well as survival.
Overall design	159 patients with MDS patients and 17 healthy controls were included in the study. CD34+ cells were isolated from bone marrow samples obtained from MDS patients and healthy controls. Samples were hybridized to Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays
Contributor(s)	Pellagatti A
Citation(s)	Gerstung M, Pellagatti A, Malcovati L, Giagounidis A et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun* 2015 Jan 9;6:5901. PMID: 25574665
Submission date	Jun 25, 2014
Last update date	Mar 25, 2019
Contact name	Andrea Pellagatti
E-mail(s)	andreapellagatti@yahoo.co.uk
Phone	00441865222911
Organization name	University of Oxford
Department	NDCLS, RDM
Lab	LLR Molecular Haematology Unit
Street address	John Radcliffe Hospital
City	Oxford
ZIP/Postal code	OX3 9DU
Country	United Kingdom
Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (176)	GSM1420393 MDS009
More...	GSM1420394 MDS011

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58831

Go to the bottom of the page

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58831

Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (176)	GSM1420393 MDS009
+ More...	GSM1420394 MDS011
	GSM1420395 MDS012

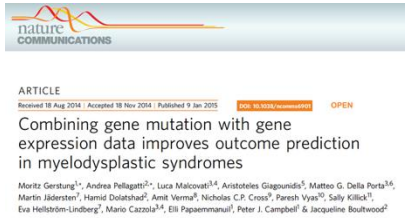
Relations

BioProject PRJNA253626

Analyze with GEO2R

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

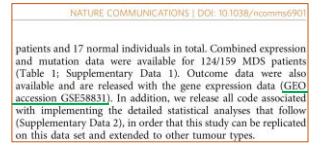
Supplementary file	Size	Download	File type/resource
GSE58831_RAW.tar	818.8 Mb	(http)(custom)	TAR (of CEL)



So, we can **find** and **access** (download) the microarray data in the GEO database, **integrate** it with clinical data (and we will **reuse** it in the class)

We can get the GEO accession number from the paper

P2 top right



We can look this up in the Gene Expression Omnibus

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58831

Go to the bottom of the page

FAIR Data

FAIR Guiding Principles seek to improve the findability, accessibility, interoperability, and reuse of digital assets.

The FAIR principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

FAIR Principles

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

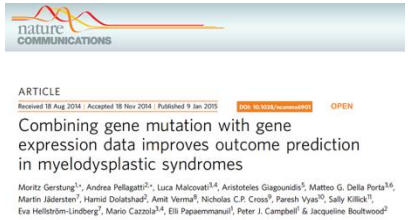
The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

https://www.go-fair.org/fair-principles/

However, in addition to the microarray data, there is also clinical data.

How does that look?

Raw data is in Gerstung_ncomms6901-s2.xlsx



43 genes / 10 alterations

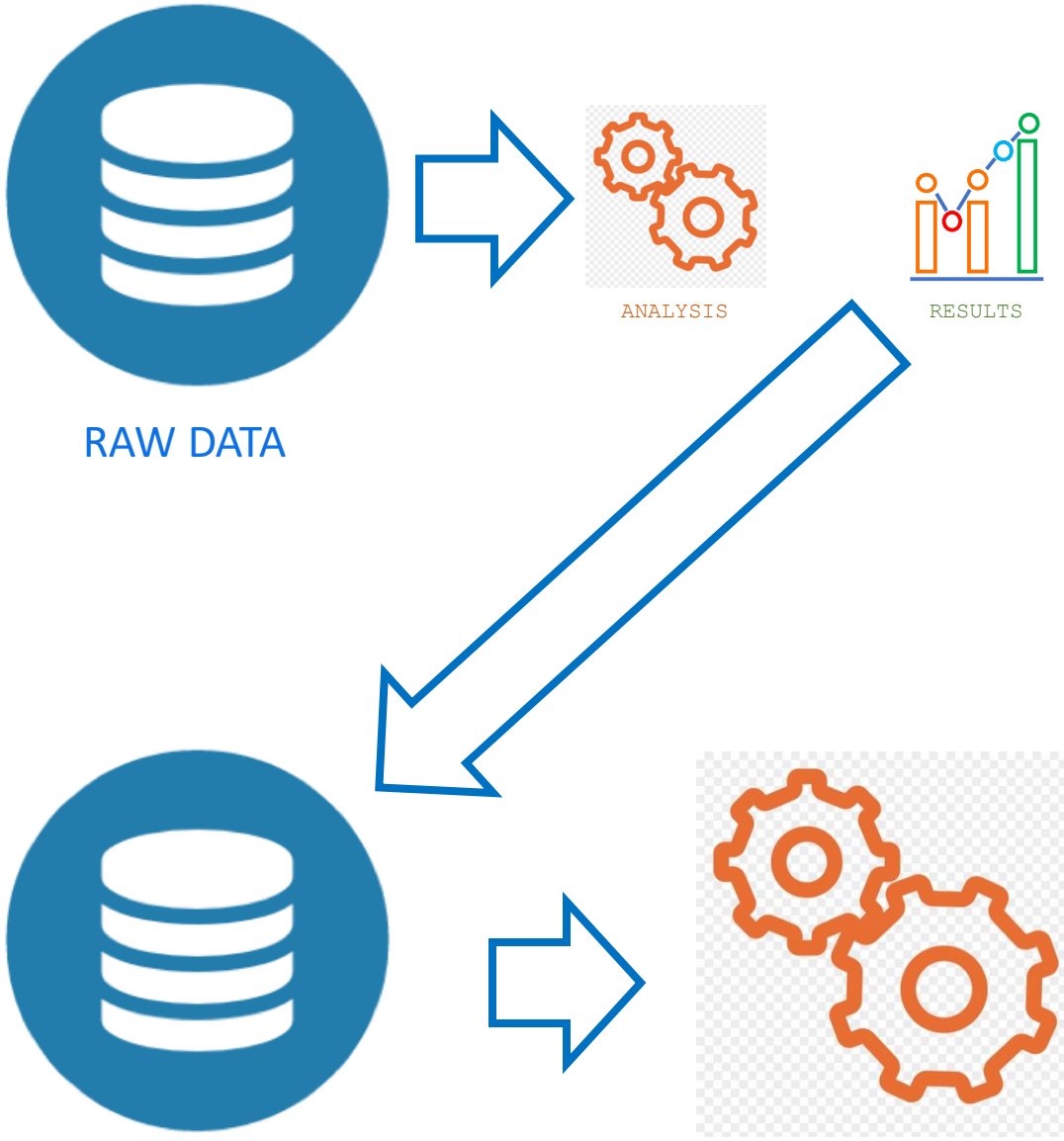
176 patients

This data is not standardized
(data is randomly assigned to columns)

In part 2 of the Gerstung paper they examine data from TCGA (The Cancer Genome Atlas)

90 parameters

Obviously, there is some additional data pre-processing somewhere



So, in the Gerstung paper, we do have all the bits in the puzzle

Reproducible Research

WORLD VIEW *A personal take on events*

No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says **Philip Stark**.

SCIENCE SHOULD BE **'SHOW ME'**, NOT **'TRUST ME'.**

In computational science, 'reproducible' often means that enough information is provided to allow a dedicated reader to repeat the calculations in the paper for herself. In biomedical disciplines, 'reproducible' often means that a different lab, starting the experiment from scratch, would get roughly the same experimental result.

"Science may be described as the art of systematic oversimplification
- the art of discerning what we may with advantage omit."

•DOI: 10.1038/d41586-018-05256-0

WORLD VIEW *A personal take on events*

No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says **Philip Stark**.

SCIENCE SHOULD BE **'SHOW ME'**, NOT **'TRUST ME'.**

- Results that generalize to all universes (or perhaps do not even require a universe) are part of mathematics.
- Results that generalize to our Universe belong to physics.
- Results that generalize to all life on Earth underpin molecular biology.
- Results that generalize to all mice are murine biology.
- Results that hold only for a particular mouse in a particular lab in a particular experiment are arguably not science

•DOI: 10.1038/d41586-018-05256-0

WORLD VIEW *A personal take on events*

No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says **Philip Stark**.

SCIENCE SHOULD BE 'SHOW ME', NOT 'TRUST ME'.

Most papers fail to report many aspects of the experiment and analysis that we may not with advantage omit — things that are crucial to understanding the result and its limitations, and to repeating the work.

We have no common language to describe this shortcoming. I've been in conferences where scientists argued about whether work was reproducible, replicable, repeatable, generalizable and other '-bles', and clearly meant quite different things by identical terms. Contradictory meanings across disciplines are deeply entrenched.

•DOI: 10.1038/d41586-018-05256-0

WORLD VIEW *A personal take on events*

No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says **Philip Stark**.

SCIENCE SHOULD BE 'SHOW ME', NOT 'TRUST ME'.

The lack of standard terminology means that we do not clearly distinguish between situations in which there is not enough information to attempt repetition, and those in which attempts do not yield substantially the same outcome.

To reduce confusion, I propose an intuitive, unambiguous neologism: 'preproducibility'. An experiment or analysis is *preproducible* if it has been described in adequate detail for others to undertake it.

Preproducibility is a prerequisite for reproducibility, and the idea makes sense across disciplines.

We should aim for pre-producibility

DOI: 10.1038/d41586-018-05256-0

WORLD VIEW *A personal take on events*

Give every paper a read for reproducibility

I was hired to ferret out errors and establish routines that promote rigorous research, says **Catherine Winchester**.

Catherine Winchester *is the grants- and research-integrity adviser at the Cancer Research UK Beatson Institute in Glasgow.*
e-mail: c.winchester@beatson.gla.ac.uk

Feedback about my reviews has been positive, especially because, as a fresh set of eyes, I can sometimes spot mistakes that someone closer to the work might not see. I've pointed out duplicated image panels, missing data and mislabelled images, among other problems.

THE BEST WAY
TO BOOST RESEARCH
QUALITY
IS TO
DISCUSS IT
OFTEN
AND
FREELY.

DOI: 10.1038/d41586-018-05140-x

PERSPECTIVE

SCIENTIFIC INTEGRITY

What does research reproducibility mean?

Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of "research reproducibility" are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for "truth."

Results reproducibility – How much variation in the results ?
Robustness and generalizability – How reproducible are the results to other situations?
Inferential reproducibility – How reproducible are the conclusions?

DOI: 10.1126/scitranslmed.aaf5027

https://www.cos.io/rpcb

Project Overview

The *Reproducibility Project: Cancer Biology* was an 8-year effort to replicate experiments from high-impact cancer biology papers published between 2010 and 2012. The project was a collaboration between the Center of Open Science and Science Exchange with all papers published as part of this project available in a collection at *eLife* and all replication data, code, and digital materials for the project available in a collection on OSF.

When preparing replications of **193 experiments** from **53 papers** there were a number of challenges.

2%
experiments with open data

70%
of experiments required asking for key reagents

69%
of experiments needing a key reagent original authors were willing to share

0%
of protocols completely described

32%
of experiments the original authors were not helpful (or unresponsive)

41%
of experiments the original authors were very helpful

Science, 374 (6573), DOI:10.1126/science.acx9788

BIOMEDICINE

Key cancer results failed to be reproduced

Project to replicate high-impact preclinical cancer studies delivers sobering verdict

By **Jocelyn Kaiser**

An ambitious project that set out 8 years ago to replicate findings from top cancer labs has drawn to a discouraging close. The Reproducibility Project: Cancer Biology (RP:CB) reported this week that when it attempted to repeat experiments drawn from 23 high-impact papers published about 10 years ago, fewer than half yielded similar results.

The findings pose "challenges for the credibility of preclinical cancer biology," says psychologist Brian Nosek, executive director of the Center for Open Science (COS), a co-organizer of the effort. The project also points to a need for authors to share more details of their experiments so others can try to reproduce them, he and others involved argue. Indeed, vague protocols and uncooperative authors, among other problems, ultimately prevented RP:CB from completing replications for 30 of the 53 papers it had initially flagged, the team reports

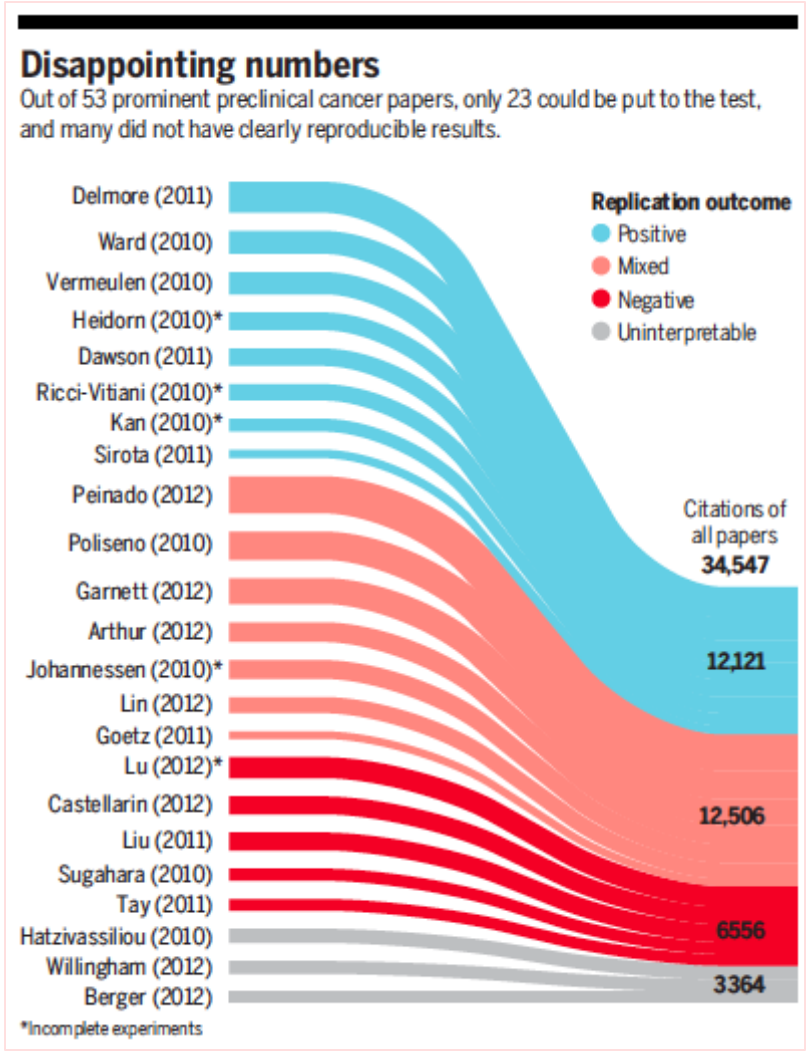
reagent sources. When authors were contacted for this information, many spent months tracking down details. But only 41% of authors were very helpful; about one-third declined or did not respond. Other problems surfaced when labs began experiments, such as cells that did not behave as expected in a baseline study.

The project ended up paring the initial list of 53 papers, comprising 193 key experiments, to just 23 papers with 50 experiments. They completed all replications for 18 of those papers and some experiments for the rest; starting in 2017, the results from each one have been published, mostly as individual papers in *eLife*. All told, the experimental work cost $1.5 million.

Results from only five papers could be fully reproduced. Other replications yielded mixed results, and some were negative or inconclusive. Overall, only 46% of 112 reported experimental effects met at least three of five criteria for replication, such as a change in the same direction—increased cancer cell growth or tumor shrinkage, for example. But even when the effects reappeared, their magnitude was usually much more modest, on average just 15% of the original effect. "That has huge implications for the success of these things moving up the pipeline into the clinic. [Drug companies] want them to be big, strong, robust effects," says Tim Errington, project leader at the COS.

The findings are "incredibly important," says Michael Lauer, deputy director for extramural research at the National Institutes of Health (NIH). At the same time, Lauer notes the lower effect sizes are not surprising because they are "consistent with ... publication bias"—that is, the fact that the most dramatic and positive effects are the most likely to be published. And the findings don't mean "all science is untrustworthy," Lauer says.

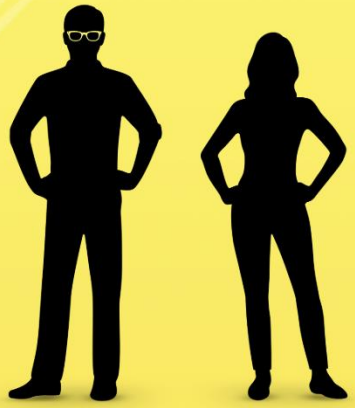
Indeed, labs have reported findings that support most of the papers, including some that failed in RP:CB. And two animal studies that weren't replicated by RP:CB



Ten Years Reproducibility Challenge

RESCIENCE SPECIAL ISSUE

FREE TO READ - FREE TO PUBLISH



Would you dare to run the code from your past self ?

(the one that does not answer mail)

SUBMISSION DEADLINE 01/04/2020

http://rescience.github.io/ten-years

In association with Inria, CNRS, Software Heritage, ReScience, Comité pour la Science Ouverte, URFIST Bordeaux & Mission de la pédagogie et du numérique pour l'enseignement supérieur.

TECHNOLOGY FEATURE · 24 AUGUST 2020

Challenge to scientists: does your ten-year-old code still run?

Missing documentation and obsolete environments force participants in the Ten Years Reproducibility Challenge to get creative.

https://www.nature.com/articles/d41586-020-02462-7

So, let's begin looking at the Gerstung paper from the perspective of reproducibility (without getting into too many details at this point)

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms6901

patients and 17 normal individuals in total. Combined expression and mutation data were available for 124/159 MDS patients (Table 1; Supplementary Data 1). Outcome data were also available and are released with the gene expression data (GEO accession GSE58831). In addition, we release all code associated with implementing the detailed statistical analyses that follow (Supplementary Data 2), in order that this study can be replicated on this data set and extended to other tumour types.

Supplementary information

Supplementary Figures

Supplementary Figures 1-2 (PDF 292 kb)

Supplementary Data 1

Clinical and sequencing data (XLSX 78 kb)

Supplementary Data 2

MDS analysis report (ZIP 2025 kb)

Supplementary Data 3

Gene-level expression fold changes and test results (TXT 8050 kb)

Supplementary Data 4

TCGA AML analysis report (ZIP 1007 kb)

Supplementary Data 5

Curated TCGA AML data (XLSX 170 kb)

https://www.nature.com/articles/ncomms6901#Sec17

If we download supplementary data 2 and unzip, then it's an html file, not the original markdown, so we have to copy and paste into a text file.

Might be helpful to group the Bioconductor and Regular R packages to help the user with the installation process

Even after installing all these packages, we still have a problem with this line – we don't have the code

```
source("suppData/mg14.R")
```

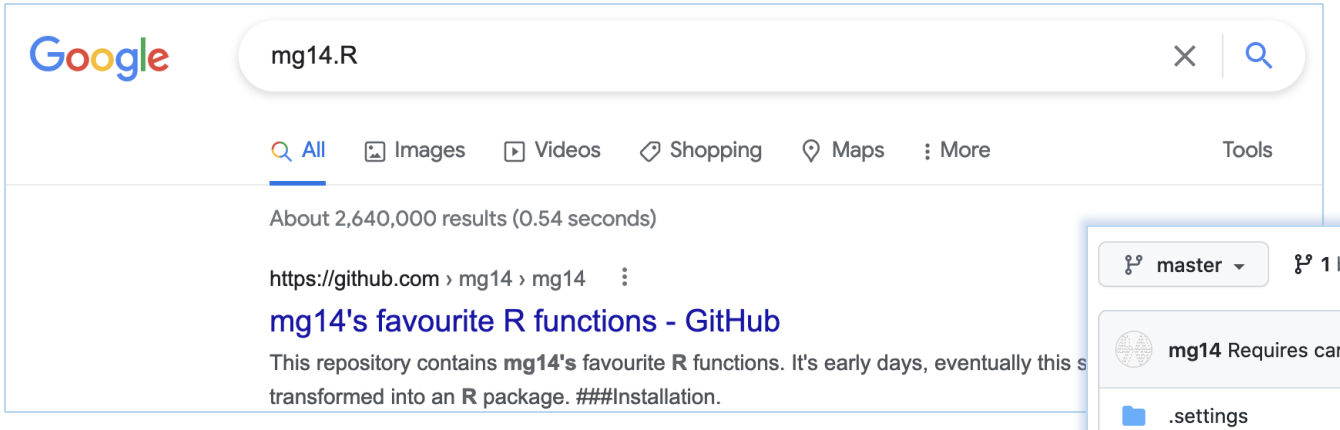
We also have another missing file

2. Load mutation and clinical data

Load clinical data for 159 MDS patients and 17 normals from Supplementary Table S1.

```
mdsData <- read.table("suppData/SuppTableS1GEO.txt", sep="\t", header=TRUE, check.names=FALSE) ## A tab-delimited version of Supplementary Table S1
head(mdsData)
```

We only have the XLS file, so will have to export as a tsv at some point



master 1 branch 0 tags Go to file Add file Code

mg14 Requires car ee1aecb on Aug 17, 2019 **33** commits

.settings	-n	8 years ago
R	Some updates to colTrans	4 years ago
man	Added numericalize function	6 years ago
.gitignore	update	8 years ago
.project	-n	8 years ago
DESCRIPTION	Requires car	2 years ago
NAMESPACE	Added numericalize function	6 years ago
README.md	Included .Rd	7 years ago

README.md

mg14's favourite R functions

This repository contains mg14's favourite R functions. It's early days, eventually this shall be transformed into an R package.

###Installation

```
> library(devtools); install_github("mg14/mg14")
```

Moritz Gerstung

mg14

Follow

67 followers · **1** following · ☆ **5**

European Bioinformatics Institute EM...

Cambridge, UK

http://mg14.github.io

Achievements

Overview | Repositories 14 | Projects | Packages

Popular repositories

CoxHD — Public archive

This repo won't be updated. Please see https://github.com/gerstung-lab/CoxHD for the most recent version

HTML ☆ 5 11

mg14

mg14's favourite R functions

R ☆ 4 5

deepSNV-old — Public archive

A bioconductor package for subclonal variant calling

R ☆ 1 2

AML-multistage

Forked from gerstung-lab/A

HTML ☆ 1 2

MDS-expression — Public archive

Forked from gerstung-lab/MDS-expression

R code for Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

R ☆ 1

python-photo-mosaic

Forked from dvdtho/python-

Python script for creating pl

Python ☆ 1 1

17 contributions in the last year

master 1 branch 1 tag Go to file Add file Code

This branch is even with gerstung-lab:master. Contribute

mg14 Fixed compatibility with recent CoxHD `0620aa6` on Mar 23, 2016 **4** commits

README.md	Added README.md	6 years ago
Supplementary-Data-2-MDS.R	Enabled reading .xlsx Supp Tables from nature.com	6 years ago
Supplementary-Data-4-TCGA-AML.R	Fixed compatibility with recent CoxHD	6 years ago

README.md

Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

master 1 branch 1 tag

Go to file | Add file | Code

This branch is even with gerstung-lab:master.

mg14 Fixed compatibility with recent CoxHD

README.md	Added README.md
Supplementary-Data-2-MDS.R	Enabled reading .xlsx Supp Tables
Supplementary-Data-4-TCGA-AML.R	Fixed compatibility with recent Co

Clone

HTTPS SSH GitHub CLI

https://github.com/mg14/MDS-expression

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

README.md

Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

However, this is the R source code, not the R Markdown file

Also, we said we needed to export the Excel file containing the clinical to a TSV

2. Load mutation and clinical data

Load clinical data for 159 MDS patients and 17 normals from Supplementary Table S1.

```
mdsData <- read.table("suppData/SuppTableS1GEO.txt", sep="\t", header=TRUE, check.names=FALSE) ## A tab-delimited version of Supplementary Table S1
head(mdsData)
```

However, if we look at the R code, it has been changed

```
63  #' ### 2. Load mutation and clinical data
64  #' Load clinical data for 159 MDS patients and 17 normals from Supplementary Table S1.
65  library(xlsx)
66  tmp <- tempfile()
67  download.file("http://www.nature.com/ncomms/2015/150109//ncomms6901/extref/ncomms6901-s2.xlsx", tmp) # Supp Table 2 from the paper
68  numericalize <- function(x){if(class(x)!='factor') return(x) else if(all(is.na(as.numeric(levels(x))))) return(x) else return(as.numeric(as.character(x)))}
69  mdsData <- as.data.frame(sapply(read.xlsx(tmp, sheetIndex=1, startRow=2, check.names=FALSE), numericalize, simplify=FALSE))
70  ix <- setdiff(na.omit(match(samples, mdsData$GEOID)), which(is.na(mdsData$PDID))) ## All MDS samples with expression and seq data
71  normalSamples <- as.character(mdsData$GEOID[mdsData$Type=="Normal"])
```

This is not good technique. And to make things worse, they have erased the change history from the GitHub repository

CBGOUS

2 followers · Oslo · http://pinga.no · @pingano1

Follow

Overview | Repositories 39 | Projects | Packages | Teams | People 14 | Settings

Popular repositories

CAJA_CLIP — Public
codes for CLIP data analysis
Python

frekner_bagcore — Public
core classes used in the Frerkner analysis
Java

frekner_SmallRNAPipeline — Public
snapshot of small RNA analysis pipeline used to analyse the Frerkner data
Java

WillowNazgul — Public
Willow program to calculate GDR etc.
Python · 1

smallRNAseqPipeline_Liu — Public
small RNAseq Pipeline for lung cells

mengjie3429 — Public
Forked from mengjie3429/mengjie3429
Config files for my GitHub profile.

Repositories

Find a repository... | Type | Language | Sort | New

View as: **Public**
You are viewing the README and repositories as a public user.

You can create a README file or pin repositories visible to anyone.

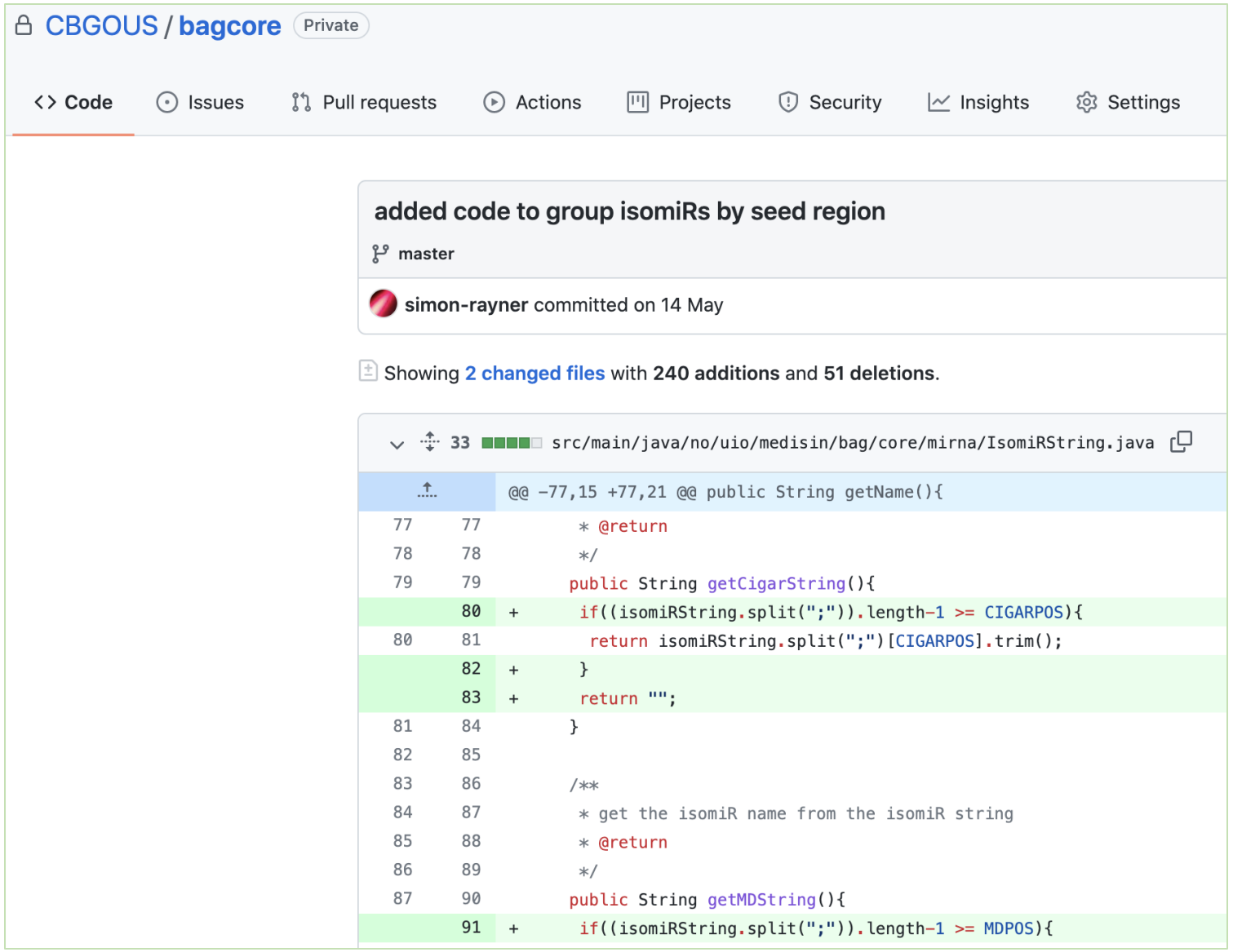
Get started with tasks that most successful organizations complete.

People

Invite someone

Top languages

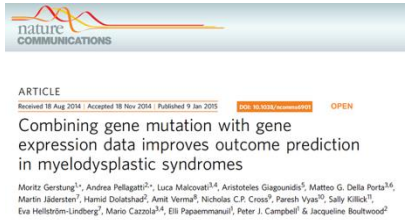
Python · Java · HTML



This is what a change history in GitHub is supposed to look like

There are other issues too

If we go back to the clinical data in the Gerstung paper



nature COMMUNICATIONS

ARTICLE

Received 18 Aug 2014 | Accepted 18 Nov 2014 | Published 9 Jan 2015 DOI: 10.1038/ncomms6901 OPEN

Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung[1,*], Andrea Pellagatti[2,*], Luca Malcovati[3,4], Aristoteles Giagounidis[5], Matteo G. Della Porta[3,6], Martin Jädersten[7], Hamid Dolatshad[2], Amit Verma[8], Nicholas C.P. Cross[9], Paresh Vyas[10], Sally Killick[11], Eva Hellström-Lindberg[7], Mario Cazzola[3,4], Elli Papaemmanuil[1], Peter J. Campbell[1] & Jacqueline Boultwood[2]

Cancer is a genetic disease, but two patients rarely have identical genotypes. Similarly, patients differ in their clinicopathological parameters, but how genotypic and phenotypic heterogeneity are interconnected is not well understood. Here we build statistical models to disentangle the effect of 12 recurrently mutated genes and 4 cytogenetic alterations on gene expression, diagnostic clinical variables and outcome in 124 patients with myelodysplastic syndromes. Overall, one or more genetic lesions correlate with expression levels of ~20% of all genes, explaining 20–65% of observed expression variability. Differential expression

Raw data is in Gerstung_ncomms6901-s2.xlsx

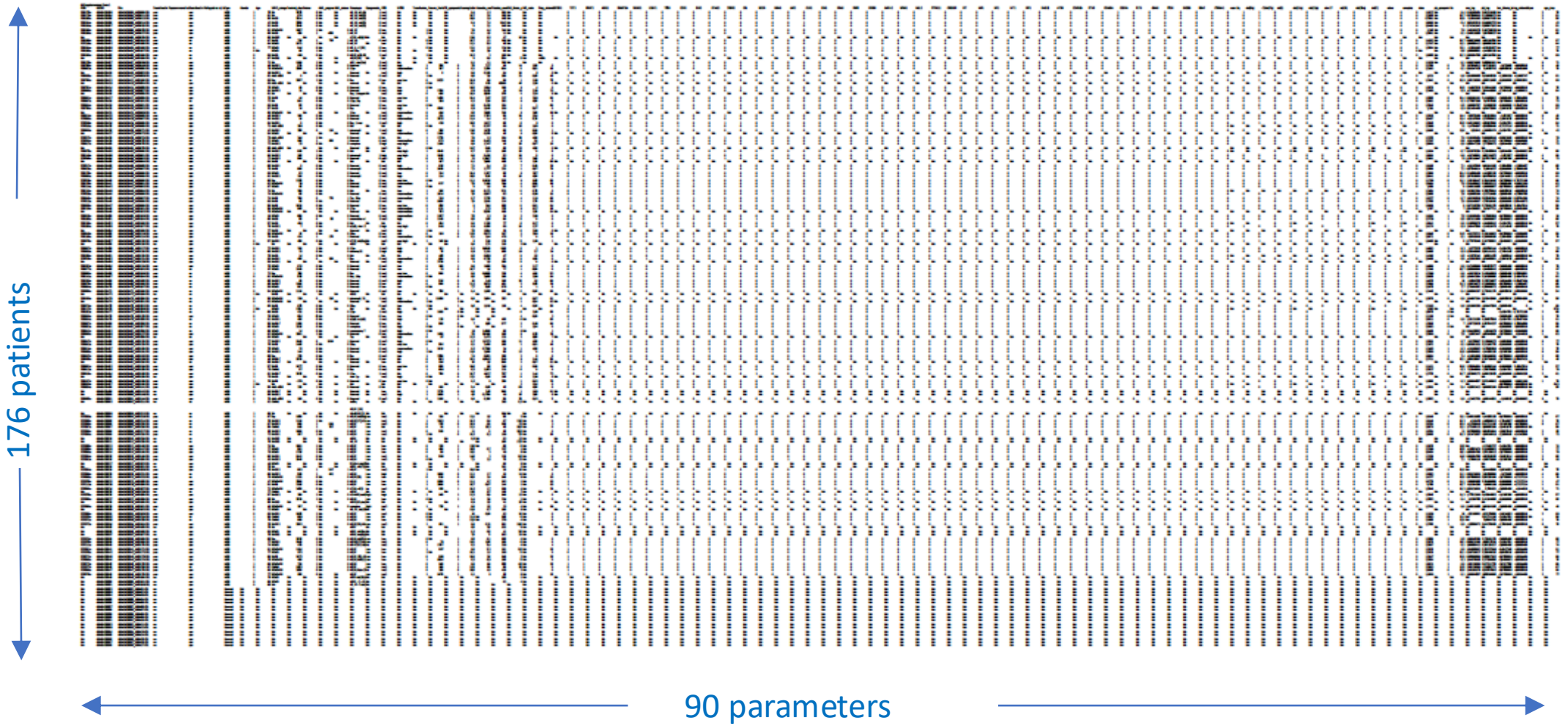
nature COMMUNICATIONS

ARTICLE

Received 18 Aug 2014 | Accepted 18 Nov 2014 | Published 9 Jan 2015 DOI: 10.1038/ncomms6901 OPEN

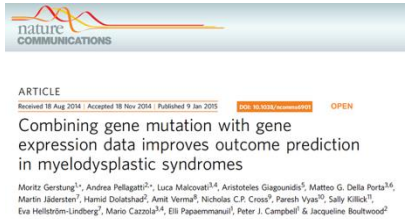
Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung[1,*], Andrea Pellagatti[2,*], Luca Malcovati[3,4], Aristoteles Giagounidis[5], Matteo G. Della Porta[3,6], Martin Jädersten[7], Hamid Dolatshad[2], Amit Verma[8], Nicholas C.P. Cross[9], Paresh Vyas[10], Sally Killick[11], Eva Hellström-Lindberg[7], Mario Cazzola[3,4], Elli Papaemmanuil[1], Peter J. Campbell[1] & Jacqueline Boultwood[2]



How did they get from 43 genes/10 alterations down to 12 genes/4 alterations?

It’s hidden in the analysis code (Gerstung_ncomms6901-s3.zip)



2. Load mutation and clinical data

Load clinical data for 159 MDS patients and 17 normals from Supplementary Table S1.

```
mdsData <- read.table("suppData/SuppTableS1GEO.txt", sep="\t", header=TRUE, check.names=FALSE) ## A tab-delimite
head(mdsData)
```

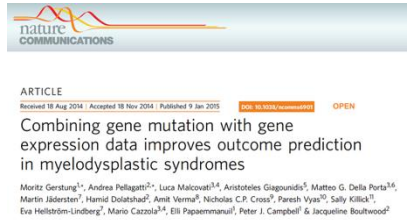
3. Match expression and clinical data

Incrementally construct the design matrix

```
design = cbind(offset=1,mdsData[ix, grep("SF3B1|TET2|SRSF2|ASXL1|DNMT3A|RUNX1|U2AF1|TP53|EZH2|IDH2|STAG2|ZRSR2|(
minF=5 ## Minimal number of alterations
design = design[,colSums(design)>=minF]
rownames(design) <- mdsData$GEOID[ix]
```

“Strikingly, overlaying the status of 12 recurrent (> 5 patients) genetic and 4 cytogenetic alterations…”
(p2, col 2, para 3)

Raw data is in Gerstung_ncomms6901-s2.xlsx



43 genes / 10 alterations

176 patients

Everywhere where things are hardcoded needs to be updated

3. Match expression and clinical data

Incrementally construct the design matrix

```
design = cbind(offset=1,mdsData[ix, grep("SF3B1|TET2|SRSF2|ASXL1|DNMT3A|RUNX1|U2AF1|TP53|EZH2|IDH2|STAG2|ZRSR2|(
minF=5 ## Minimal number of alterations
design = design[,colSums(design)>=minF]
rownames(design) <- mdsData$GEOID[ix]
```

90 parameters

Obviously, there is some additional data pre-processing somewhere

What is a PDID?

	A	B	C	D	E	F	G	
1	## Supplementary Data 1							
2	PDID	GEOID	File	Described in Papaemmanuil et	Described in Pellagatti et al,	Type	Gender	Ag
3	PD6175a	GSM1420393	GSM1420393_MDS009.CEL	yes	yes	MDS	1	
4	PD6173a	GSM1420394	GSM1420394_MDS011.CEL	yes	yes	MDS	1	
5	PD6185a	GSM1420395	GSM1420395_MDS012.CEL	yes	yes	MDS	1	
6	PD6184a	GSM1420396	GSM1420396_MDS014.CEL	yes	yes	MDS	1	
7	PD6183a	GSM1420397	GSM1420397_MDS015.CEL	yes	yes	MDS	1	
8	PD6198a	GSM1420398	GSM1420398_MDS025.CEL	yes	yes	MDS	0	
9	PD6188a	GSM1420399	GSM1420399_MDS027.CEL	yes	yes	MDS	0	
10	NA	GSM1420400	GSM1420400_MDS029.CEL	no	yes	MDS	0	
11	PD6189a	GSM1420401	GSM1420401_MDS184.CEL	yes	no	MDS	1	
12	PD7116a	GSM1420402	GSM1420402_MDS185.CEL	yes	no	MDS	0	
13	PD6194a	GSM1420403	GSM1420403_MDS186.CEL	yes	no	MDS	0	
14	PD6195a	GSM1420404	GSM1420404_MDS187.CEL	yes	no	MDS	0	NA
15	PD6190a	GSM1420405	GSM1420405_MDS188.CEL	yes	no	MDS	0	

Is 0 male or female?

2. Load mutation and clinical data

Load clinical data for 159 MDS patients and 17 normals from Supplementary Table S1.

```
mdsData <- read.table("suppData/SuppTableS1GEO.txt", sep="\t", header=TRUE, check.names=FALSE) ## A tab-delimite
head(mdsData)
```

3. Match expression and clinical data

Incrementally construct the design matrix

```
design = cbind(offset=1,mdsData[ix, grep("SF3B1|TET2|SRSF2|ASXL1|DNMT3A|RUNX1|U2AF1|TP53|EZH2|IDH2|STAG2|ZRSR2|
minF=5 ## Minimal number of alterations
design = design[,colSums(design)>=minF]
rownames(design) <- mdsData$GEOID[ix]
```

nature COMMUNICATIONS

ARTICLE

Received 18 Aug 2014 | Accepted 18 Nov 2014 | Published 9 Jan 2015 DOI: 10.1038/ncomms6901 OPEN

Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung[1,*], Andrea Pellagatti[2,*], Luca Malcovati[3,4], Aristoteles Giagounidis[5], Matteo G. Della Porta[3,6], Martin Jädersten[7], Hamid Dolatshad[2], Amit Verma[8], Nicholas C.P. Cross[9], Paresh Vyas[10], Sally Killick[11], Eva Hellström-Lindberg[7], Mario Cazzola[3,4], Elli Papaemmanuil[1], Peter J. Campbell[1] & Jacqueline Boultwood[2]

This should be in the methodology

It isn't an ideal coding solution.

- if you want to change the gene list, you must change the code
- If you want to change the dataset, you must change the code

Nevertheless, Gerstung provides code that could be adapted to input any combination of phenotype and genetic data for further investigation by statistical analysis techniques.

These are challenges everyone faces when analyzing their data

Data Packages

Example 2: screening data from tuberculosis study – 3 groups

Sample ID	Age	Sex	Weight	Height	Place of sample collection	Type of specimen (sputum/FNA/Biopsy	Form of TB (SPTB/EPTB)	X-ray findings	Duration of cough	Hemoptysis (Y/N)	self reported HIV status	Tested HIV status	BCG vaccination	GenXpert gra	AFB grading	week at which TB culture growth first observed	Drug resistanc	Drug resistance-LPA (S/R/INH/RIF only MDR)	SIT number	
5		M	weight loss		GU	sputum	SPTB	No	2 weeks and above	No	None	negative		MTB Detected			RSSS			
6		F	weight loss		GU	sputum	SPTB	No	2 weeks and above	No	None	negative		MTB Detected			RSSS			
7		M	weight loss		GU	sputum	SPTB	No	2 weeks and above	No	None	negative		MTB Detected			SRSS			
38		M	weight loss		GY	sputum	SPTB	No	2 weeks and above	No	None	negative		MTB Detected			SSSS			
72		M	weight loss		GU	sputum	SPTB	No	2 weeks and above	Yes	None	Positive		MTB Detected			SSSS			

Sample ID	Drug resistance-LPA (S/R/INH/RIF only MDR)	Age	Sex	Weight	Height	Place of sample colelction	Type of specimen (sputum/FNA /Biopsy	Form of TB (SPTB/EPTB)	X-ray findings	Duration of cough	Hemoptysis (Y/N)	self reported HIV status	Tested HIV status	BCG vaccination	GenXpert gra	AFB grading	week at which TB culture growth fir observed	Drug resistanc	-Phynotypic	SIT number
CH001	ND		Male	59	177	HC	sputum	SPTB	cavitatio				Negative	n	ND	POS	NA	ND		53
CH092	ND		Male	50	168	HC	sputum	SPTB		week	Yes		Negative		ND	POS	NA	ND		26
CH984	ND		Female	51	152	HC	sputum	SPTB	HML	week	No			n	ND	NEG	NA	ND		53
CH3294	ND		Female			HC	sputum	SPTB		month	Yes	negtive		n	ND	NEG	NA	ND		ND
CH43	ND		Male	54	170	HC	sputum	SPTB		month	No		Posetive	n	ND	NEG	NA	ND		41

Sample ID	Age	Sex	Weight	Height	Place of sample colelction	Type of specimen (sputum/FNA/Biopsy	Form of TB (SPTB/EPTB)	X-ray findings	Duration of cough	Hemoptysis (Y/N)	self reported HIV status	Tested HIV status	BCG vaccination	GenXpert gra	AFB grading	week at which TB culture growth first observed	Drug resistanc	Drug resistance-LPA (S/R/INH/RIF only MDR)	SIT number	Lineage
WH039		male	N/A	N/A	WDY	Sputum	PTB	N/A	>3 weeks			N/A	N/A		scanty	4	N/A	MDR	SIT 4	L4
WH78		female	N/A	N/A	WDY	Sputum	PTB	N/A	>3 weeks			N/A	N/A	pos		3	N/A	INH res	SIT 149	L4
WH135		male	N/A	N/A	WDY	Sputum	PTB	N/A	>3 weeks			N/A	N/A	pos		4	N/A	Rif res	SIT 21	L3
FH13		male	N/A	N/A	FLK	Sputum	PTB	N/A	>3 weeks			N/A	N/A		"+2"	4	N/A	Rif res	SIT 53	L4

GENDER DATA IS DESCRIBED USING THREE STANDARDS

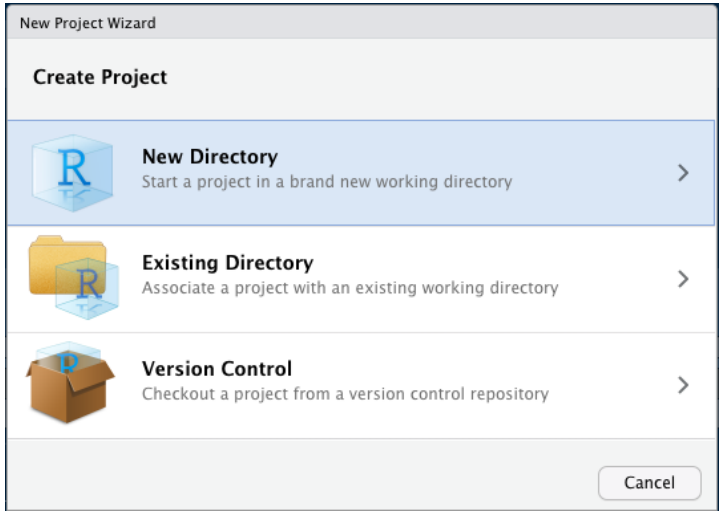
DRUG RESISTANCE DATA IS ALL OVER THE PLACE

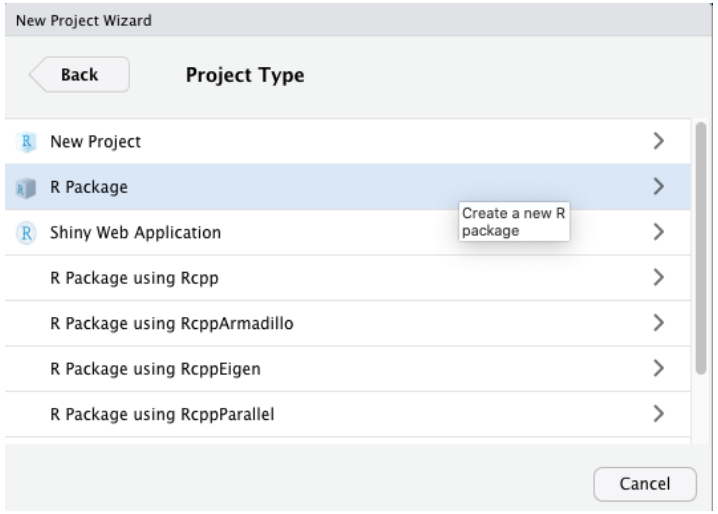
What we need is a standardized way of providing information

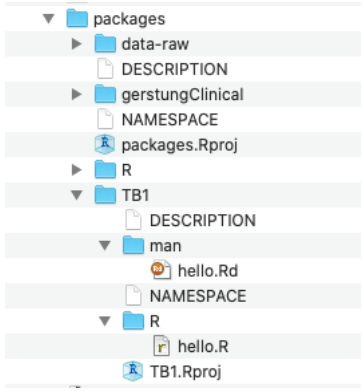
For example, the Gerstung paper is based around data collected from microarray experiments.

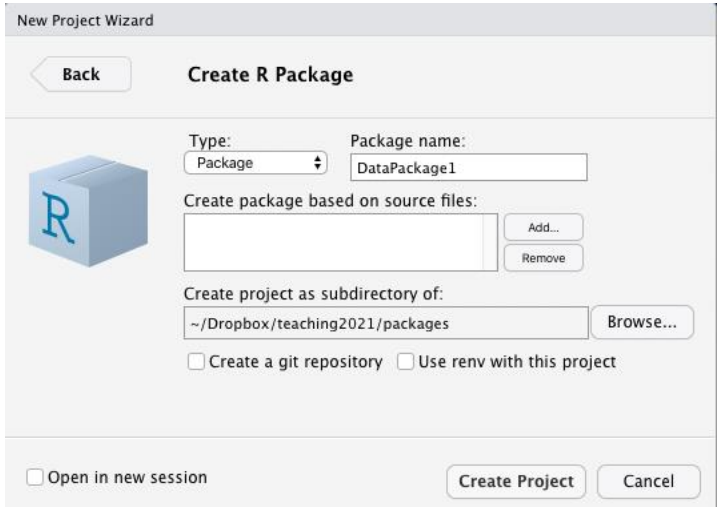
In order to interpret the data, we need to know what feature is on which spot on the microarray

In R, We can access this through an **AnnotationDB** object


New Project Wizard
Create Project
New Directory
Start a project in a brand new working directory
Existing Directory
Associate a project with an existing working directory
Version Control
Checkout a project from a version control repository
Cancel

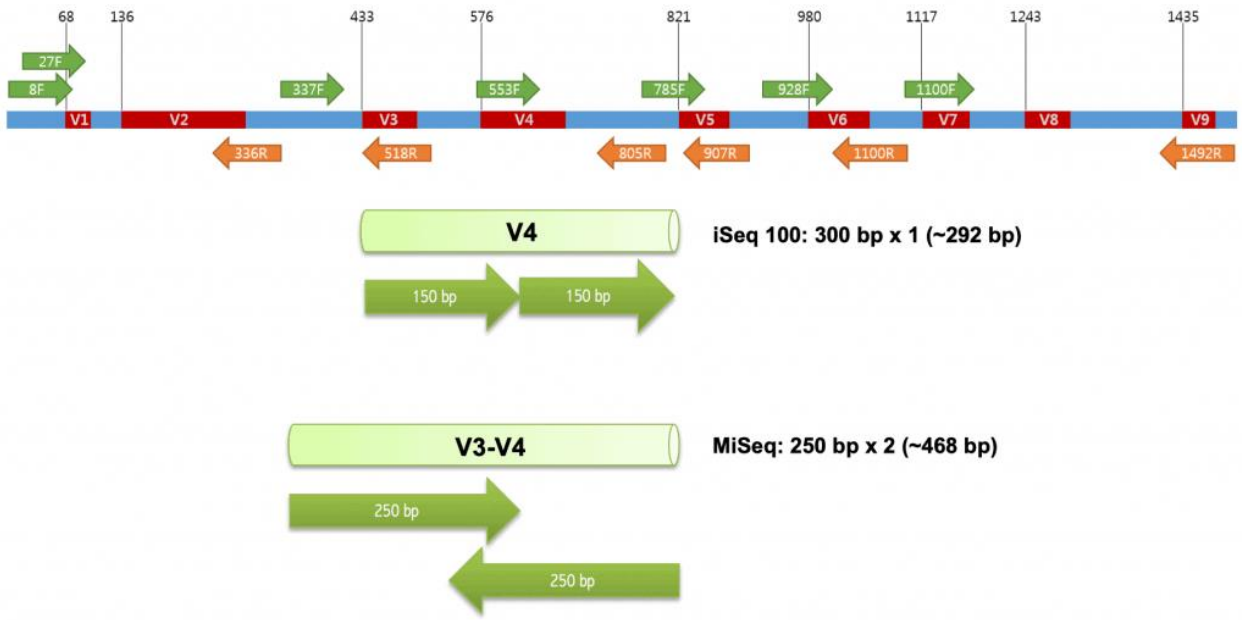

New Project Wizard
Back
Project Type
New Project
R Package
Shiny Web Application
R Package using Rcpp
R Package using RcppArmadillo
R Package using RcppEigen
R Package using RcppParallel
Create a new R package
Cancel


packages
data-raw
DESCRIPTION
gerstungClinical
NAMESPACE
packages.Rproj
R
TB1
DESCRIPTION
man
hello.Rd
NAMESPACE
R
hello.R
TB1.Rproj


New Project Wizard
Back
Create R Package
Type:
Package
Package name:
DataPackage1
Create package based on source files:
Add...
Remove
Create project as subdirectory of:
~/Dropbox/teaching2021/packages
Browse...
Create a git repository
Use renv with this project
Open in new session
Create Project
Cancel

Data

Metagenomics



illumina

Search Illumina.com

Products Learn Company Support Recommended Links

16S Metagenomic Sequencing Library Preparation

NORWEGIAN SEQUENCING CENTRE

Home About Illumina services Illumina Submission PacBio services PacBio Submission Publications Forms FAQ

Illumina Submission

- 1. General information
- 2. How to fill the submission form

Ribosomal 16S DNA library preps

Prep. type	Min. amount	Recommend amount	Max. volume	Accepted buffers
DNA for 16S prep	200 ng per sample	200 - 2000 ng	100 µl	10 mM Tris

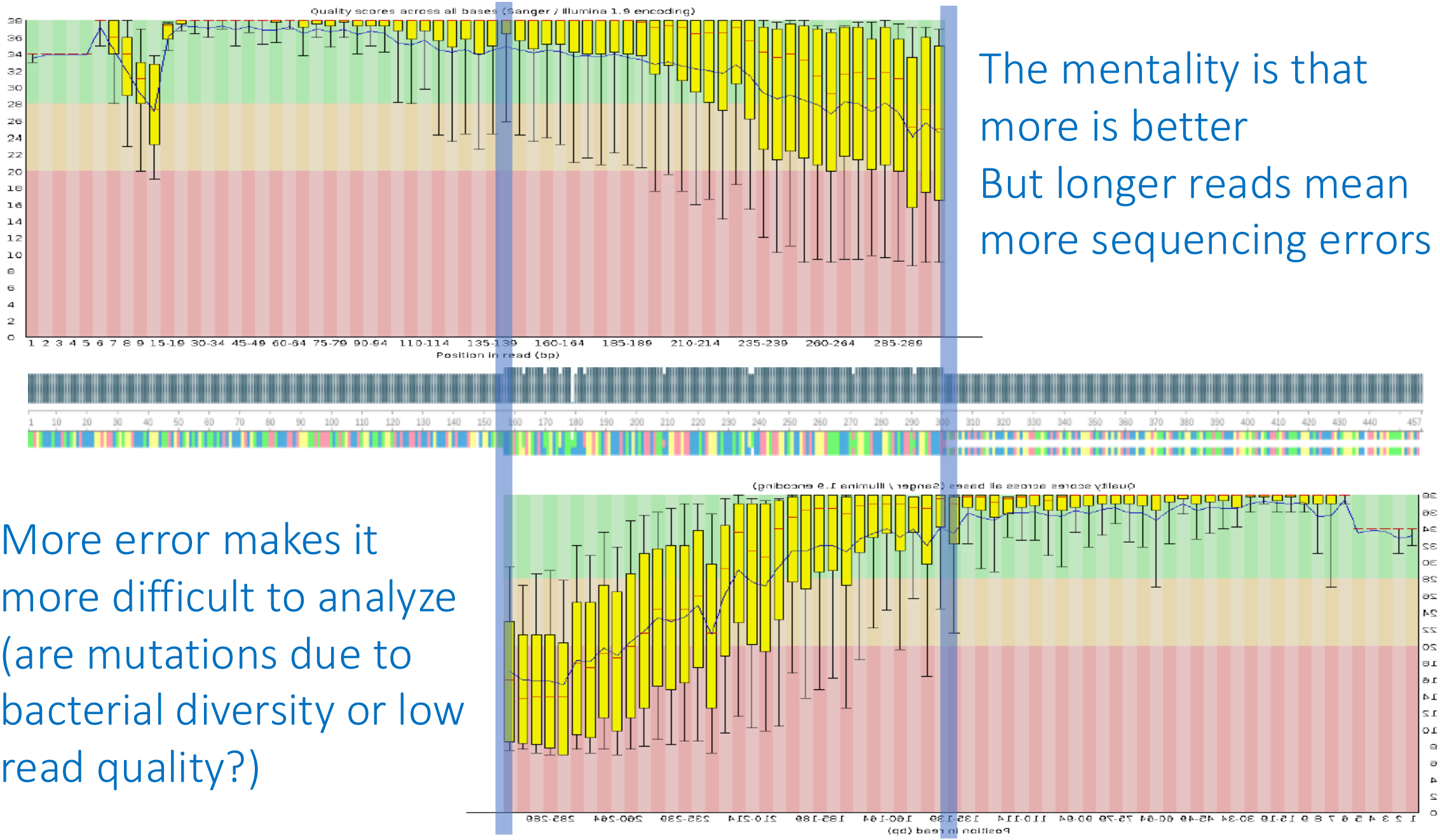
Products > NGS Library Preparation > 16S rRNA Library Prep Kits

16S V3-V4 Library Preparation Kit for Illumina

Cat. 70400, 70410, 70420, 70430, 70440

The reagents and components required for library preparation of the 16S V3-V4 amplicon libraries to be used for next-generation sequencing on Illumina platforms

- Protocol optimized for DNA isolated from a diversity of samples including stool, soil, water, saliva, plant, urine, skin, and more
- Simple and quick workflow: library could be prepared in less than 5 hours
- Component of Norgen's metagenomics workflow
- A single NGS run can be prepared with up to 384 unique dual-index libraries

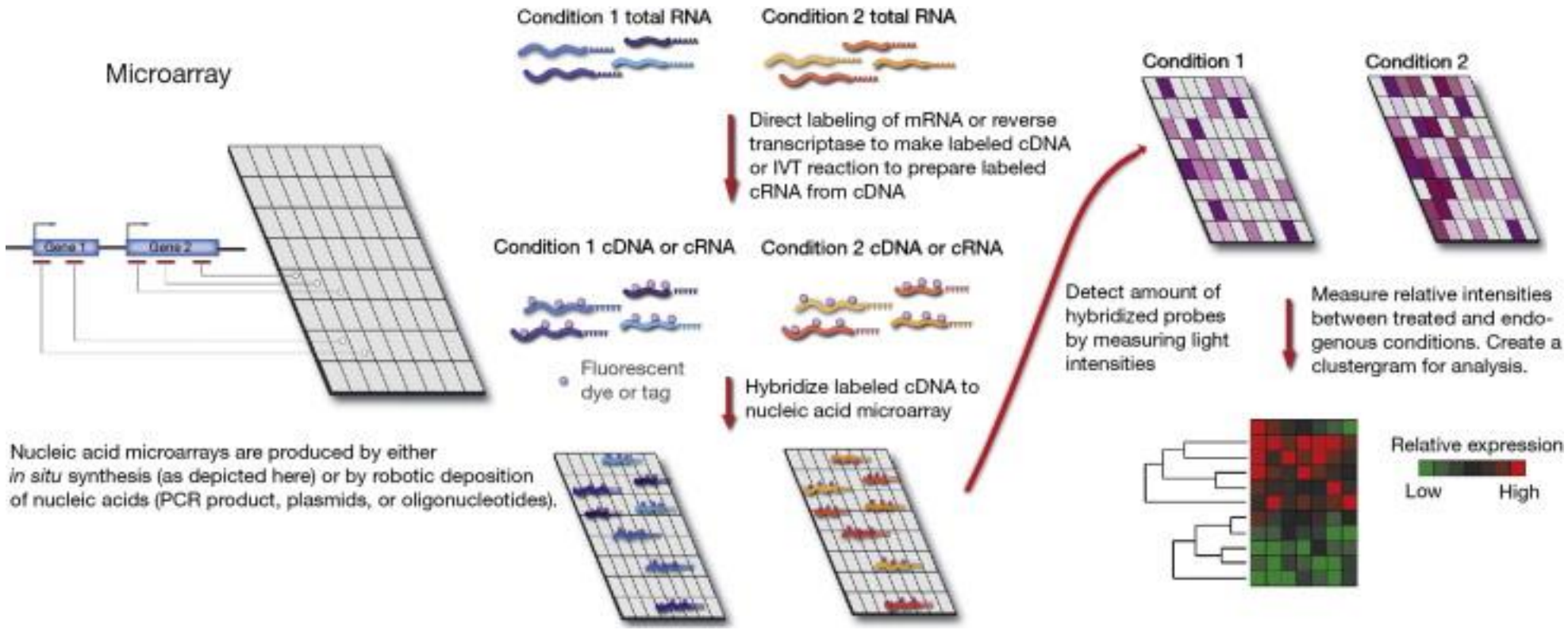


The mentality is that more is better
But longer reads mean more sequencing errors

More error makes it more difficult to analyze (are mutations due to bacterial diversity or low read quality?)

https://mothur.org/blog/2014/Why-such-a-large-distance-matrix/

Microarray


Microarray
Gene 1
Gene 2
Condition 1 total RNA
Condition 2 total RNA
Direct labeling of mRNA or reverse transcriptase to make labeled cDNA or IVT reaction to prepare labeled cRNA from cDNA
Condition 1 cDNA or cRNA
Condition 2 cDNA or cRNA
Fluorescent dye or tag
Hybridize labeled cDNA to nucleic acid microarray
Nucleic acid microarrays are produced by either *in situ* synthesis (as depicted here) or by robotic deposition of nucleic acids (PCR product, plasmids, or oligonucleotides).
Condition 1
Condition 2
Detect amount of hybridized probes by measuring light intensities
Measure relative intensities between treated and endogenous conditions. Create a clustergram for analysis.
Relative expression
Low
High

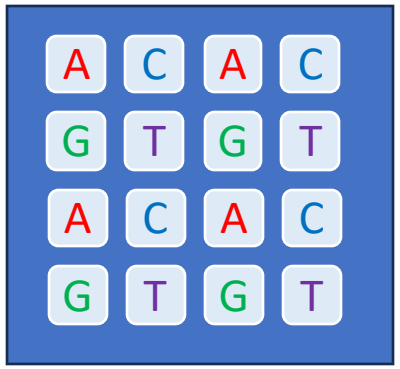
THE GENOME *Bacteria minimus* HAS A SIZE OF 200 NUCLEOTIDES. WHOLE GENOME SEQUENCING FOLLOWED BY SEQUENCE ANALYSIS DISCOVERS 4 GENES

Gene 1: CGCTGAAAAAAAAAA

Gene 2: CGCTGACCCCCCCCC

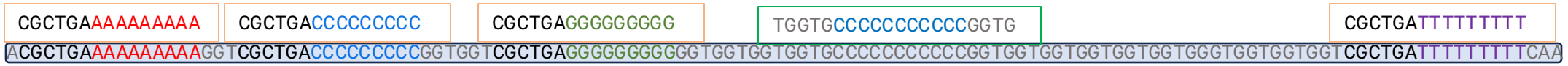
Gene 3: CGCTGAGGGGGGGGG

Gene 4: CGCTGATTTTTTTTT



ARSEYMETRICS V1.0

WE PERFORM RNA-SEQ ON *Bacteria minimus* AND DISCOVER A LONG CODING RNA

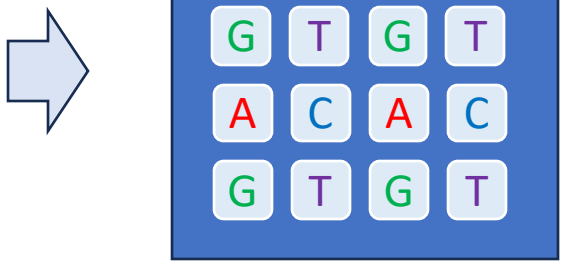


Gene 1: CGCTGAAAAAAAAAA

Gene 2: CGCTGACCCCCCCCC

Gene 3: CGCTGAGGGGGGGGG

Gene 4: CGCTGATTTTTTTTT



ARSEYMETRICS V1.0

lncRNA 1: TGGTGCCCCCCCCCCCGGTGG

v1 OF THE MICROARRAY CANNOT DISTINGUISH BETWEEN Gene 1 and lncRNA 1

CREATE A NEW MICROARRAY BASED ON THIS UPDATED INFORMATION

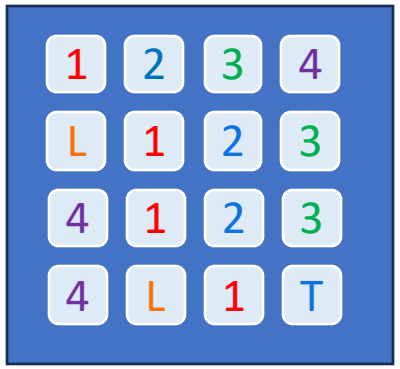
Gene 1: CGCTGAAAAAAAAAA

Gene 2: CGCTGACCCCCCCCC

Gene 3: CGCTGAGGGGGGGGG

Gene 4: CGCTGATTTTTTTTT

lncRNA 1: TGGTGCCCCCCCCCCCGGTGG



ARSEYMETRICS V2.0

DIFFERENT LAYOUT ON THE ARRAY, BUT IT NOW SCREENS FOR FIVE FEATURES

STUDY 1 ON *Bacteria minimus* IDENTIFIES CHANGES IN EXPRESSION OF GENE 1 BETWEEN CONTROL AND ANTIBIOTIC TREATMENT

A SECOND STUDY ON *Bacteria minimus* FINDS NO SIGNIFICANT CHANGES IN EXPRESSION OF GENE 1 BETWEEN CONTROL AND ANTIBIOTIC TREATMENT

STUDY 1 ON *Bacteria minimus* USING THE ARSEYMETRICS MICROARRAY PLATFORM IDENTIFIES CHANGES IN EXPRESSION OF GENE 1 BETWEEN CONTROL AND ANTIBIOTIC TREATMENT IN 100 SAMPLES

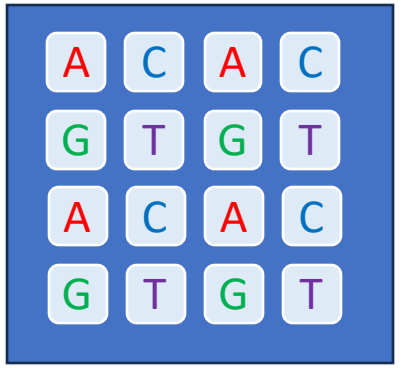
A SECOND STUDY ON *Bacteria minimus* USING THE ARSEYMETRICS MICROARRAY PLATFORM FINDS NO SIGNIFICANT CHANGES IN EXPRESSION OF GENE 1 BETWEEN CONTROL AND ANTIBIOTIC TREATMENT BUT WITH ONLY 20 SAMPLES

WHICH STUDY DO WE BELIEVE?

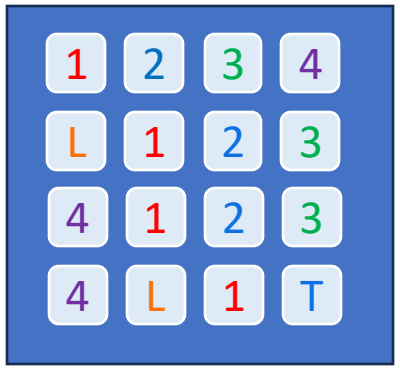
STUDY 1 ON *Bacteria minimus* USING THE ARSEYMETRICS MICROARRAY PLATFORM IDENTIFIES CHANGES IN EXPRESSION OF GENE 1 BETWEEN CONTROL AND ANTIBIOTIC TREATMENT IN 100 SAMPLES

A SECOND STUDY ON *Bacteria minimus* USING THE ARSEYMETRICS MICROARRAY PLATFORM FINDS NO SIGNIFICANT CHANGES IN EXPRESSION OF GENE 1 BETWEEN CONTROL AND ANTIBIOTIC TREATMENT BUT WITH ONLY 20 SAMPLES

WHICH STUDY DO WE BELIEVE?



ARSEYMETRICS V1.0



ARSEYMETRICS V2.0

WHICH VERSION OF THE ARSEYMETRICS ARRAY WAS USED IN EACH STUDY?

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58831

Platforms (1) GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (176) GSM1420393 MDS009
More... GSM1420394 MDS011
GSM1420395 MDS012

Relations

BioProject PRJNA253626

Analyze with GEO2R

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
GSE58831_RAW.tar	818.8 Mb	(http)(custom)	TAR (of CEL)

NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP

NCBI > GEO > **Accession Display**

GEO help: Mouse over screen elements for information.

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GPL570

Platform GPL570

Query DataSets for GPL570

Status	Public on Nov 07, 2003
Title	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Technology type	in situ oligonucleotide
Distribution	commercial
Organism	Homo sapiens
Manufacturer	Affymetrix
Manufacture protocol	see manufacturer's web site

Complete coverage of the Human Genome U133 Set plus 6,500 additional genes for analysis of over 47,000 transcripts

All probe sets represented on the GeneChip Human Genome U133 Set are identically replicated on the GeneChip Human Genome U133 Plus 2.0 Array. The sequences from which these probe sets were derived were selected from GenBank®, dbEST, and RefSeq. The sequence clusters were created from the UniGene database (Build 133, April 20, 2001) and then refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database (April 2001 release).

In addition, there are 9,921 new probe sets representing approximately 6,500 new genes. These gene sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from the UniGene database (Build 159, January 25, 2003) and refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the NCBI human genome assembly (Build 31).

Applied Biosystems™

GeneChip™ Human Genome U133A 2.0 Array

The GeneChip™ Human Genome U133A 2.0 Array is a single array representing 14,500 well-characterized human genes that can be used Read more

Have Questions? **Contact Us**

Change view

Catalog Number	Number of Arrays
900471	2 arrays

Catalog number 900471

Price (NOK) / 2 arrays

8 336,00

Online exclusive

~~10 420,00~~

Save 2 084,00 (20%)

1

Estimated availability date
Pending

Add to cart

Number of Arrays: 2 arrays

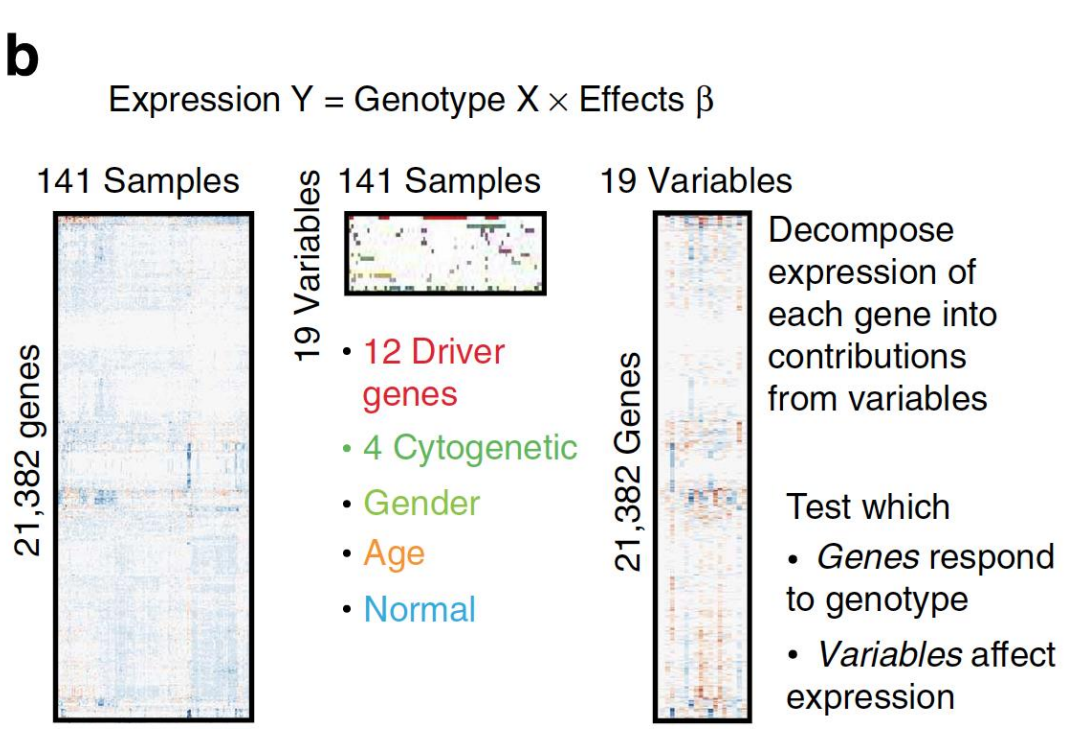
The GeneChip™ Human Genome U133A 2.0 Array is a single array representing 14,500 well-characterized human genes that can be used to explore human biology and disease processes.

THE CURRENT ENSEMBL ANNOTATION RELEASE IS v113
BASED ON GRCh38.p14 CONTAINS > 21 000 GENES

http://www.ensembl.org/index.html

Why should we care about this?

MODELING THE DATA



observed expression for gene k in patient k

$$Y_{ik} = \sum_{j=1} X_{ij}\beta_{ij} + \beta_{0i}$$

where:

X_{ij} is the mutation matrix for patient k and mutation j

$X_{ij} = 1$: patient i has an oncogenic mutation j

$X_{ij} = 0$ no mutation

for gender $X_{ij} = 1 =$ female; $X_{ij} = 0 =$ male

for age X_{ij} takes integer values.

β_{jk} is the the expression change in gene k induced by the presence of mutation j.

β_{0j} denotes the baseline expression level of gene j

glmPrediction <- geneExprlm$coefficients %*% t(design)

Usage

```
lmFit(object, design=NULL, ndups=1, spacing=1, block=NULL, correlation, weights=NULL, method="ls", ...)
```

Arguments

object	A matrix-like data object containing log-ratios or log-expression values for a series of arrays, with rows corresponding to genes and columns to samples. Any type of data object that can be processed by `getEAWP` is acceptable.
design	the design matrix of the microarray experiment, with rows corresponding to arrays and columns to coefficients to be estimated. Defaults to the unit vector meaning that the arrays are treated as replicates.

The advantage with using **lmFit** is you can simply provide gene expression data and a design.
Trying to use the in-built function would require more effort.
Interactions can be handled but they must be specified in separate commands in **Limma**