Analyzing genomic data

From data to insight (hopefully)

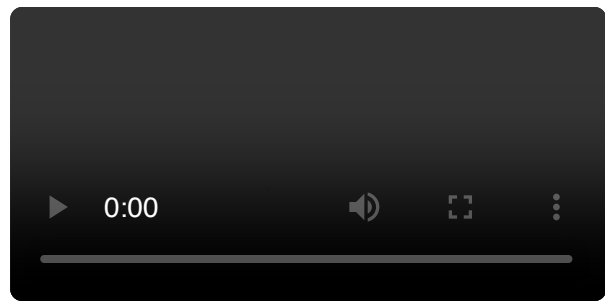
https://github.com/MatthiasZepper/Lecture-OmicsDataAnlysis

NATIONAL GENOMICS INFRASTRUCTURE

Matthias Zepper

- Life and Medical Sciences in Bonn 🇩🇪
- PhD in leukemia epigenetics in Münster 🇩🇪
- Founder (& liquidator) of start-up Nucleotidy

  nucleotidy
  embraces your bases
- Bioinformatician at the NGI, Stockholm 🇸🇪


0:00

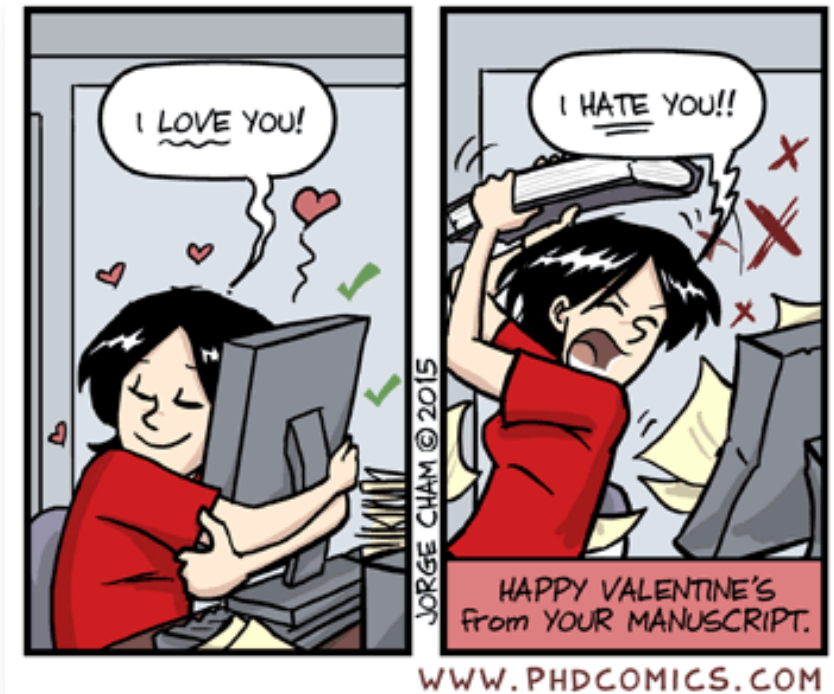
Bioinformatician

The toolbox

Genomic data formats

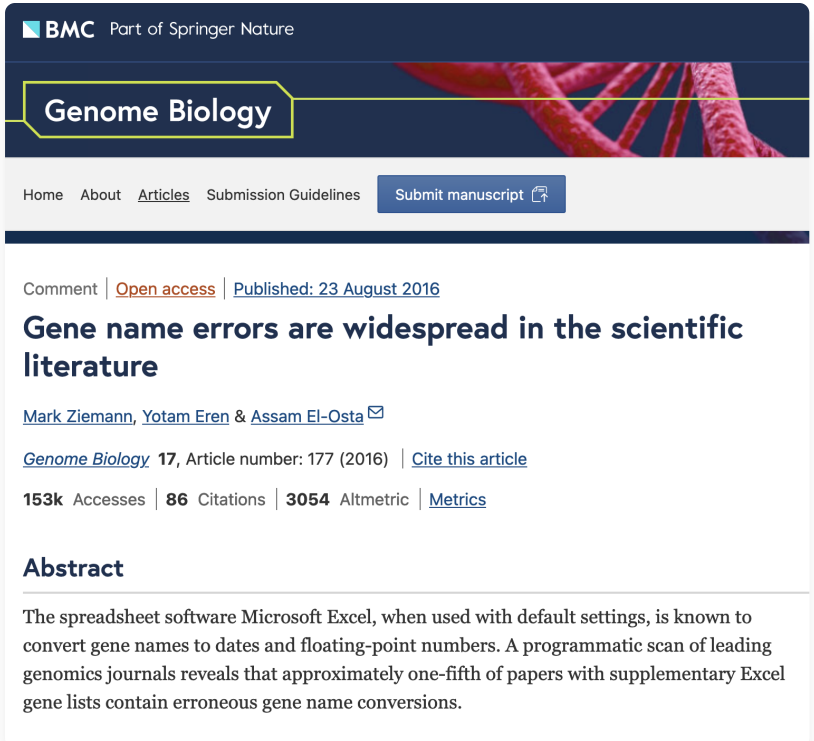
Exemplary analyses

Workflows



https://phdcomics.com/comics/archive.php?comicid=1780

Toolbox peek



BMC Part of Springer Nature

Genome Biology

Home About Articles Submission Guidelines Submit manuscript

Comment | Open access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

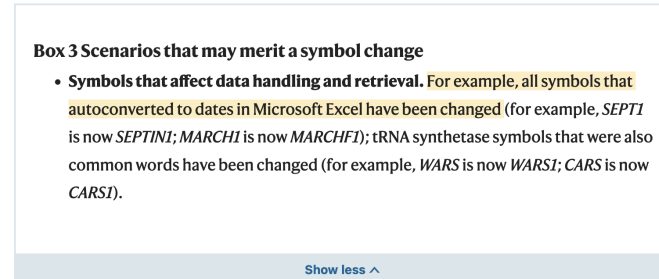
153k Accesses | **86** Citations | **3054** Altmetric | Metrics

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

https://doi.org/10.1186/s13059-016-1044-7

What I don't use



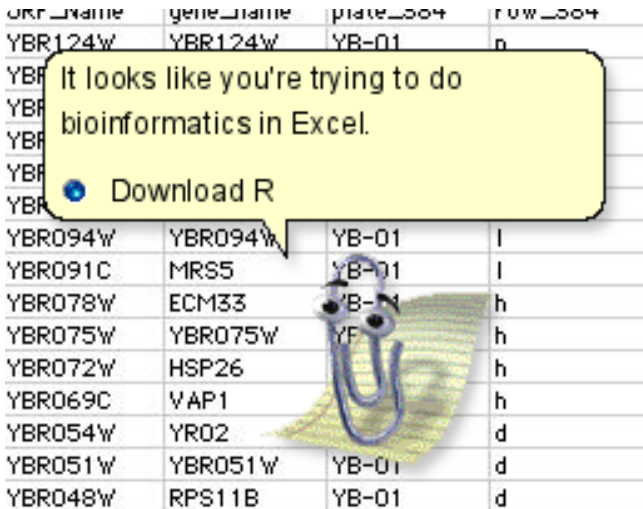
Box 3 Scenarios that may merit a symbol change

- **Symbols that affect data handling and retrieval.** For example, all symbols that autoconverted to dates in Microsoft Excel have been changed (for example, *SEPT1* is now *SEPTIN1*; *MARCH1* is now *MARCHF1*); tRNA synthetase symbols that were also common words have been changed (for example, *WARS* is now *WARS1*; *CARS* is now *CARS1*).

Show less

In 2020, *Human Genome Gene Nomenclature Committee (HGNC)* renamed genes that were auto-converted to dates in Excel.

What I don't use



ORF_Name	gene_name	plate_384	row_384
YBR124W	YBR124W	YB-01	n
YBR094W	YBR094W	YB-01	l
YBR091C	MRS5	YB-01	l
YBR078W	ECM33	YB-01	h
YBR075W	YBR075W	YB-01	h
YBR072W	HSP26	YB-01	h
YBR069C	VAP1	YB-01	h
YBR054W	YRO2	YB-01	d
YBR051W	YBR051W	YB-01	d
YBR048W	RPS11B	YB-01	d

MICROSOFT

Microsoft Fixes Excel Feature That Forced Scientists to Rename Human Genes

Microsoft now allows users to disable automatic date conversion, which means scientists no longer have to worry about using alternative names for genes.

By **Dua Rashid** Published October 23, 2023 | Comments (16)

Image: Stephen Brashear (Getty Images)

Microsoft recently published a blog highlighting new Excel updates that allow users to disable Automatic Data Conversion. This comes as good news for the scientists, because in recent years they had to rename quite a few human gene names—since Excel was converting them to dates.

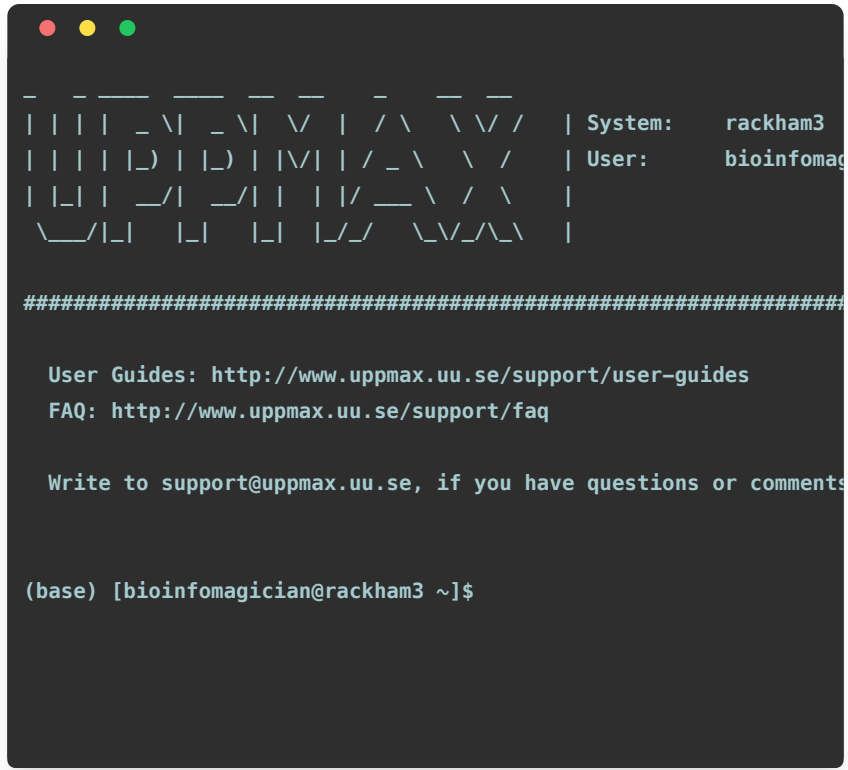
https://gizmodo.com/microsoft-fixes-excel-feature-that-forced-scientists-to-1850949443

Use suitable tools

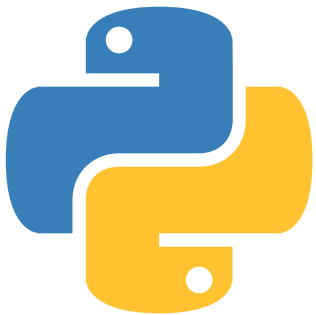
- They might not have a GUI.
- They might not run on your machine.
- For remote compute, mind data privacy!

Use a suitable OS

- GNU / Linux
- MacOS
- Windows Subsystem for Linux



Programming languages for data exploration



Python

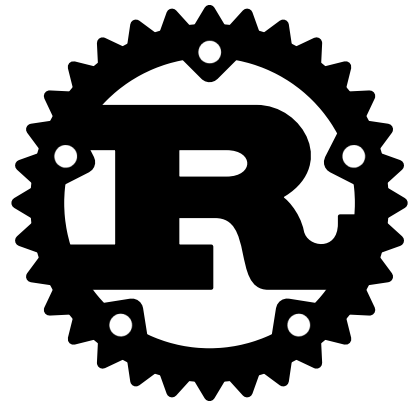
R

Julia

Programming languages for tools

C++

Go



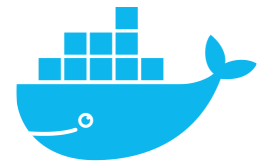
Rust

Bioinformatic ecosystem

git (version control)

Jupyter, Quarto (Notebooks)

Snakemake, Nextflow (Workflows)



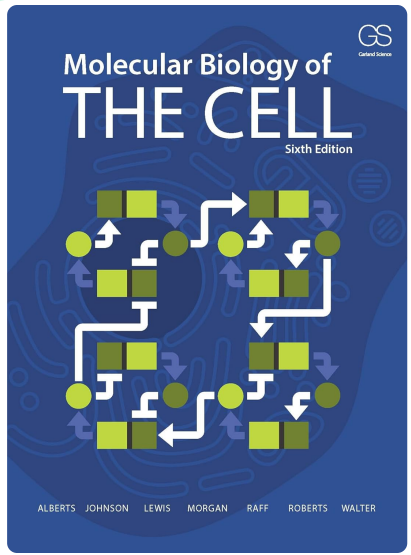
Docker, Apptainer (containers)

Bioconda (package manager)

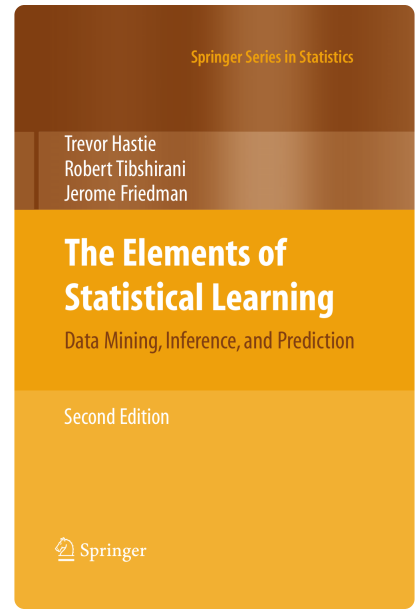
Bioconductor, Tidyverse (R packages)

BioNumPy, Pandas, Polar.rs, Apache Arrow, DuckDB (Analytics)

The most important tools



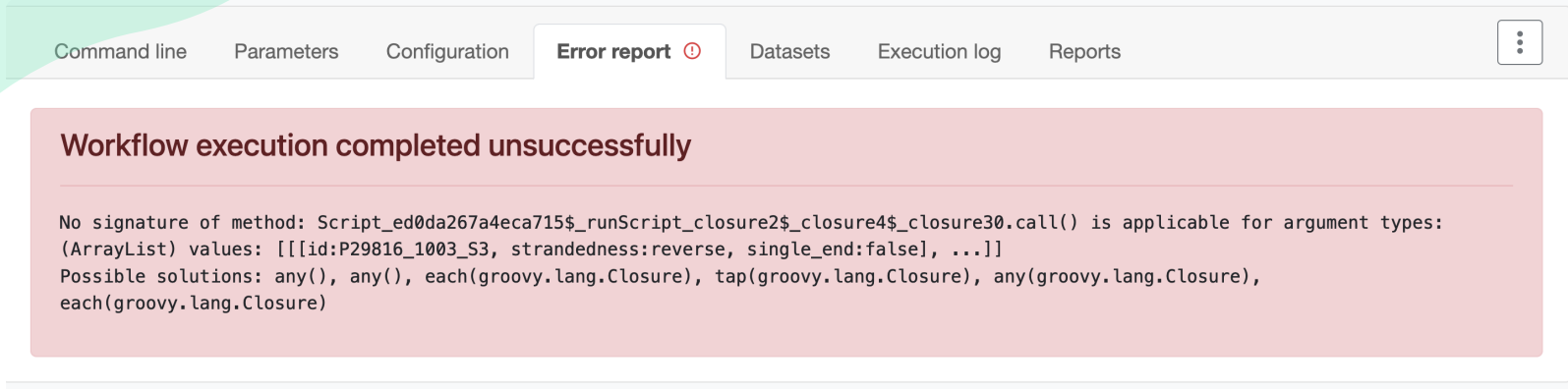
Biological understanding



Statistical knowledge

(free fulltext)

Syntax errors are easy to debug



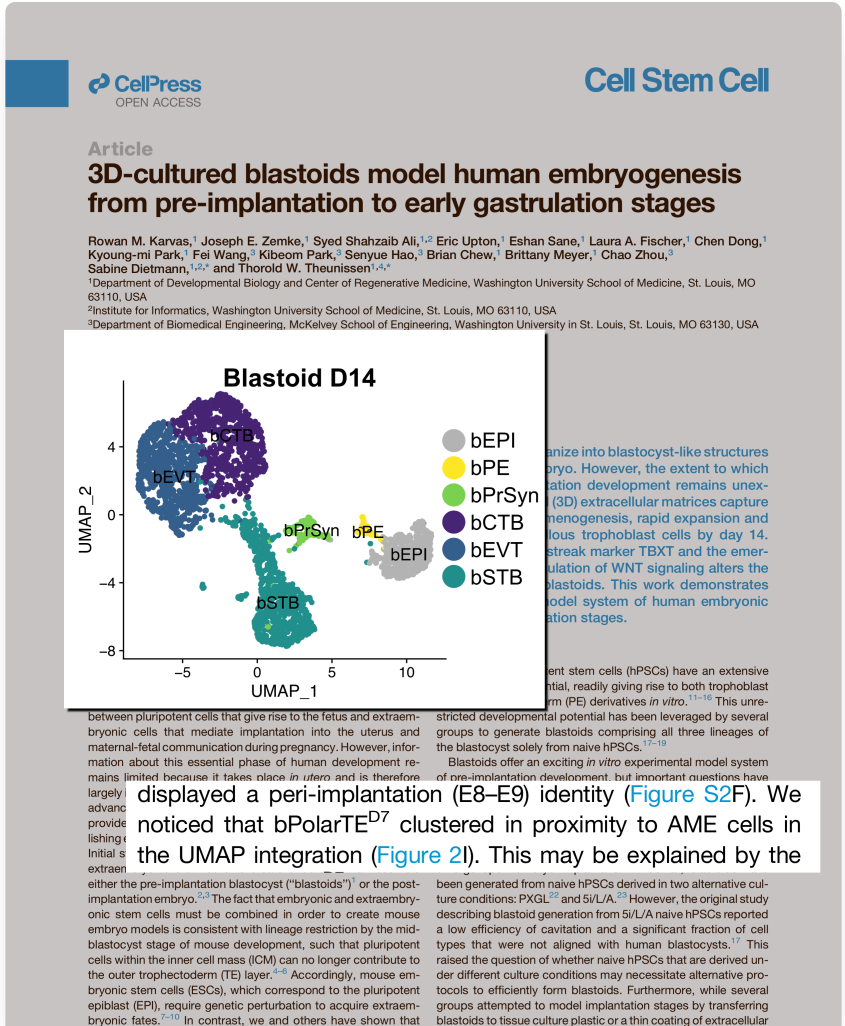
but it frequently happens that

tools output *something* arbitrarily.

Understand the methods you apply

Example from j.stem.2023.08.005:

- Findings backed by wet lab results.
- Distances in 2D projections of UMAP / t-SNE are not directly interpretable.
- Their loss functions are invariant with respect to rotations.
- More details at Understanding UMAP



CellPress
OPEN ACCESS

Cell Stem Cell

Article

3D-cultured blastoids model human embryogenesis from pre-implantation to early gastrulation stages

Rowan M. Karvas,[1] Joseph E. Zemke,[1] Syed Shahzaib Ali,[1,2] Eric Upton,[1] Eshan Sane,[1] Laura A. Fischer,[1] Chen Dong,[1] Kyoung-mi Park,[1] Fei Wang,[3] Kibeom Park,[3] Senyue Hao,[3] Brian Chew,[1] Brittany Meyer,[1] Chao Zhou,[3] Sabine Dietmann,[1,2,*] and Thorold W. Theunissen[1,4,*]

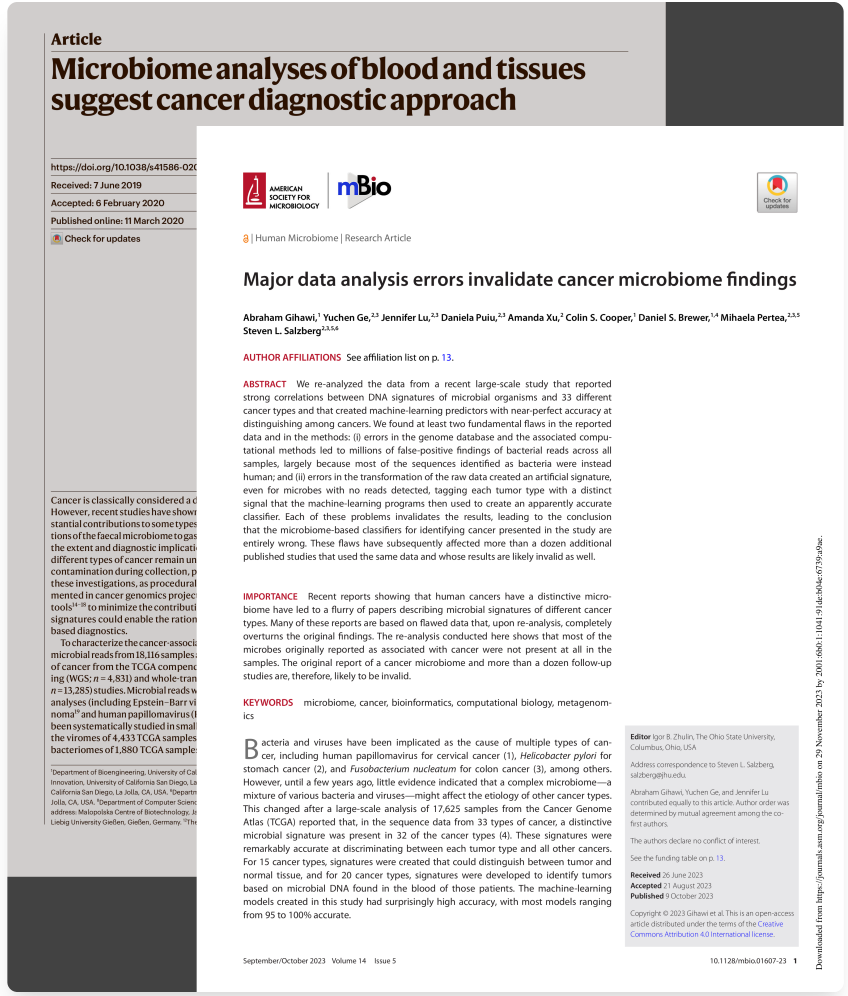
[1]Department of Developmental Biology and Center of Regenerative Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
[2]Institute for Informatics, Washington University School of Medicine, St. Louis, MO 63110, USA
[3]Department of Biomedical Engineering, McKelvey School of Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA

displayed a peri-implantation (E8–E9) identity (Figure S2F). We noticed that bPolarTE^{D7} clustered in proximity to AME cells in the UMAP integration (Figure 2I). This may be explained by the

Understand the methods you apply

Example from 10.1038/s41586-020-2095-1:

- In this case, both the analysis strategy and the understanding of the used methods was inadequate.
- Overconfident broad generalization of the findings.
- More details at 10.1128/mbio.01607-23



Article

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

https://doi.org/10.1038/s41586-020

Received: 7 June 2019

Accepted: 6 February 2020

Published online: 11 March 2020

Check for updates

Cancer is classically considered a d
However, recent studies have shown
stantial contributions to some types
tions of the faecal microbiome to gas
the extent and diagnostic implicati
different types of cancer remain un
contamination during collection, p
these investigations, as procedural
mented in cancer genomics projec
tools[14–16] to minimize the contributi
signatures could enable the ration
based diagnostics.

To characterize the cancer-associa
microbial reads from 18,116 samples
of cancer from the TCGA compend
ing (WGS; $n = 4,831$) and whole-tran
$n = 13,285$) studies. Microbial reads w
analyses (including Epstein–Barr vi
noma[16] and human papillomavirus (
been systematically studied in smal
the viromes of 4,433 TCGA samples
bacteriomes of 1,880 TCGA sample

[1]Department of Bioengineering, University of Ca
Innovation, University of California San Diego, La
California San Diego, La Jolla, CA, USA. [4]Departm
Jolla, CA, USA. [5]Department of Computer Scienc
address: Malopolska Centre of Biotechnology, J
Liebig University Gießen, Gießen, Germany. [11]Th

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

Check for updates

Human Microbiome | Research Article

Major data analysis errors invalidate cancer microbiome findings

Abraham Gihawi,[1] Yuchen Ge,[2,3] Jennifer Lu,[2,3] Daniela Puiu,[2,3] Amanda Xu,[2] Colin S. Cooper,[1] Daniel S. Brewer,[1,4] Mihaela Pertea,[2,3,5] Steven L. Salzberg[2,3,5,6]

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT We re-analyzed the data from a recent large-scale study that reported strong correlations between DNA signatures of microbial organisms and 33 different cancer types and that created machine-learning predictors with near-perfect accuracy at distinguishing among cancers. We found at least two fundamental flaws in the reported data and in the methods: (i) errors in the genome database and the associated computational methods led to millions of false-positive findings of bacterial reads across all samples, largely because most of the sequences identified as bacteria were instead human; and (ii) errors in the transformation of the raw data created an artificial signature, even for microbes with no reads detected, tagging each tumor type with a distinct signal that the machine-learning programs then used to create an apparently accurate classifier. Each of these problems invalidates the results, leading to the conclusion that the microbiome-based classifiers for identifying cancer presented in the study are entirely wrong. These flaws have subsequently affected more than a dozen additional published studies that used the same data and whose results are likely invalid as well.

IMPORTANCE Recent reports showing that human cancers have a distinctive microbiome have led to a flurry of papers describing microbial signatures of different cancer types. Many of these reports are based on flawed data that, upon re-analysis, completely overturns the original findings. The re-analysis conducted here shows that most of the microbes originally reported as associated with cancer were not present at all in the samples. The original report of a cancer microbiome and more than a dozen follow-up studies are, therefore, likely to be invalid.

KEYWORDS microbiome, cancer, bioinformatics, computational biology, metagenomics

Bacteria and viruses have been implicated as the cause of multiple types of cancer, including human papillomavirus for cervical cancer (1), *Helicobacter pylori* for stomach cancer (2), and *Fusobacterium nucleatum* for colon cancer (3), among others. However, until a few years ago, little evidence indicated that a complex microbiome—a mixture of various bacteria and viruses—might affect the etiology of other cancer types. This changed after a large-scale analysis of 17,625 samples from the Cancer Genome Atlas (TCGA) reported that, in the sequence data from 33 types of cancer, a distinctive microbial signature was present in 32 of the cancer types (4). These signatures were remarkably accurate at discriminating between each tumor type and all other cancers. For 15 cancer types, signatures were created that could distinguish between tumor and normal tissue, and for 20 cancer types, signatures were developed to identify tumors based on microbial DNA found in the blood of those patients. The machine-learning models created in this study had surprisingly high accuracy, with most models ranging from 95 to 100% accurate.

Editor Igor B. Zhulin, The Ohio State University, Columbus, Ohio, USA

Address correspondence to Steven L. Salzberg, salzberg@jhu.edu.

Abraham Gihawi, Yuchen Ge, and Jennifer Lu contributed equally to this article. Author order was determined by mutual agreement among the co-first authors.

The authors declare no conflict of interest.

See the funding table on p. 13.

Received 26 June 2023
Accepted 21 August 2023
Published 9 October 2023

Common genomic data formats

FastQ: Format for sequencing reads

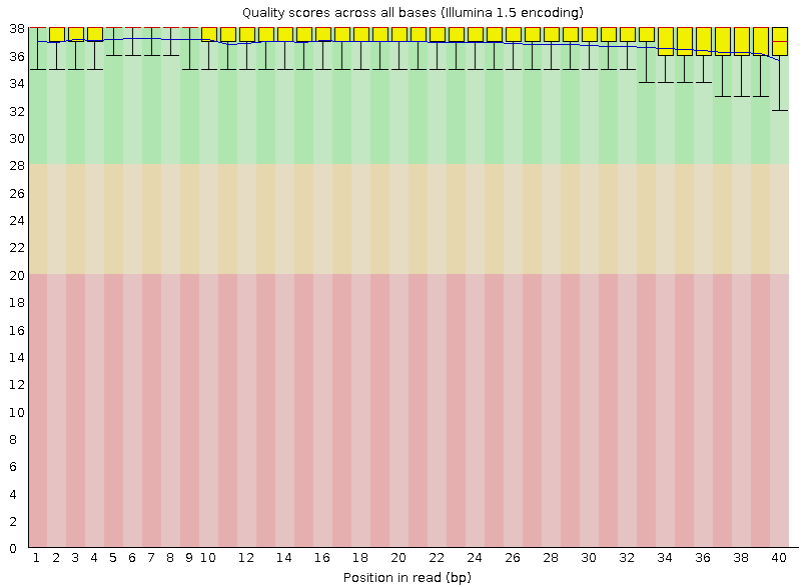
FastQ

FastQ: Format for sequencing reads

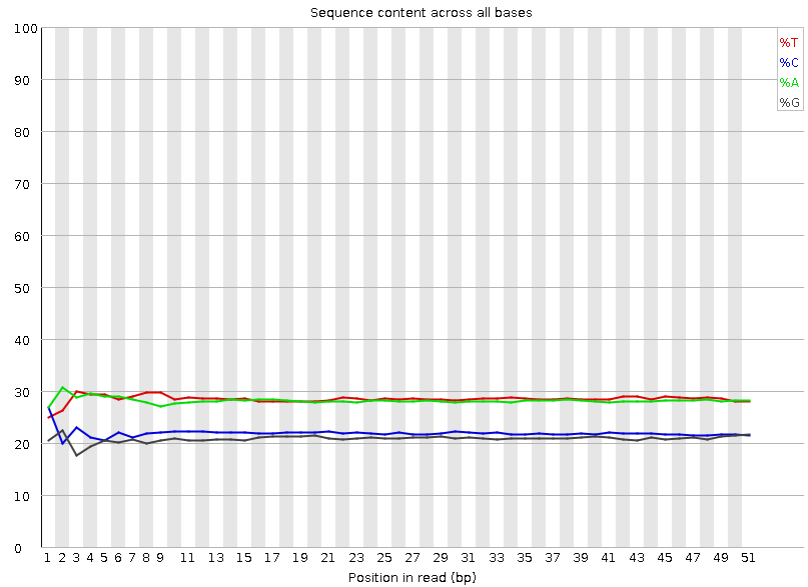
- Plain text format
- Each read is represented by four consecutive lines:
  1. Sequence identifier and an optional description
  2. The sequence
  3. + (optional)
  4. The base call quality

```
@SCILIFELAB:500:NGISTLM:1:1101:32832:1016 1:N:0:GCTTCAGGGT+AAGGTAGCGT
TCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGGGACGAAAAATGGTTTTTAGAGAACGAGAAGACGGTTACGCAG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF
```

Quality control: Good sequencing quality

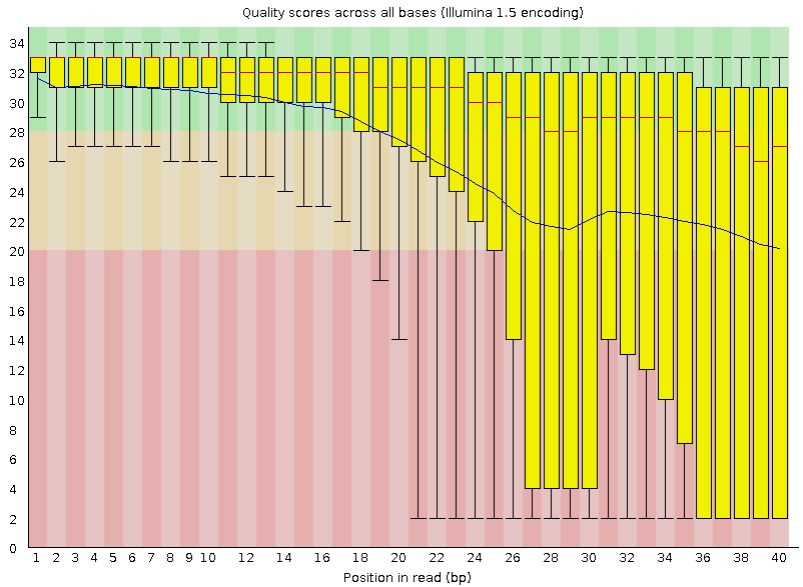


Base call quality is high along the full read

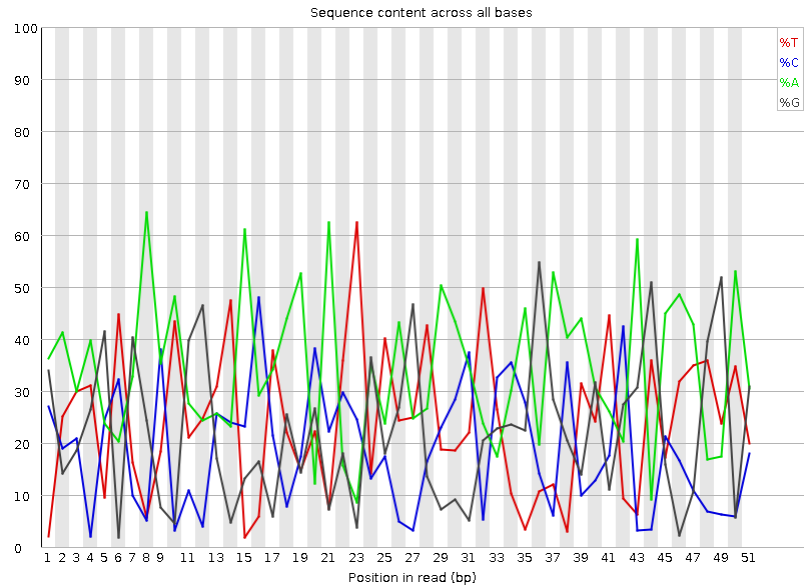


The base composition is balanced

Quality control: Poor sequencing quality



Base call quality drops off dramatically



The base composition is heavily skewed

Common tasks

(pairwise) Alignment

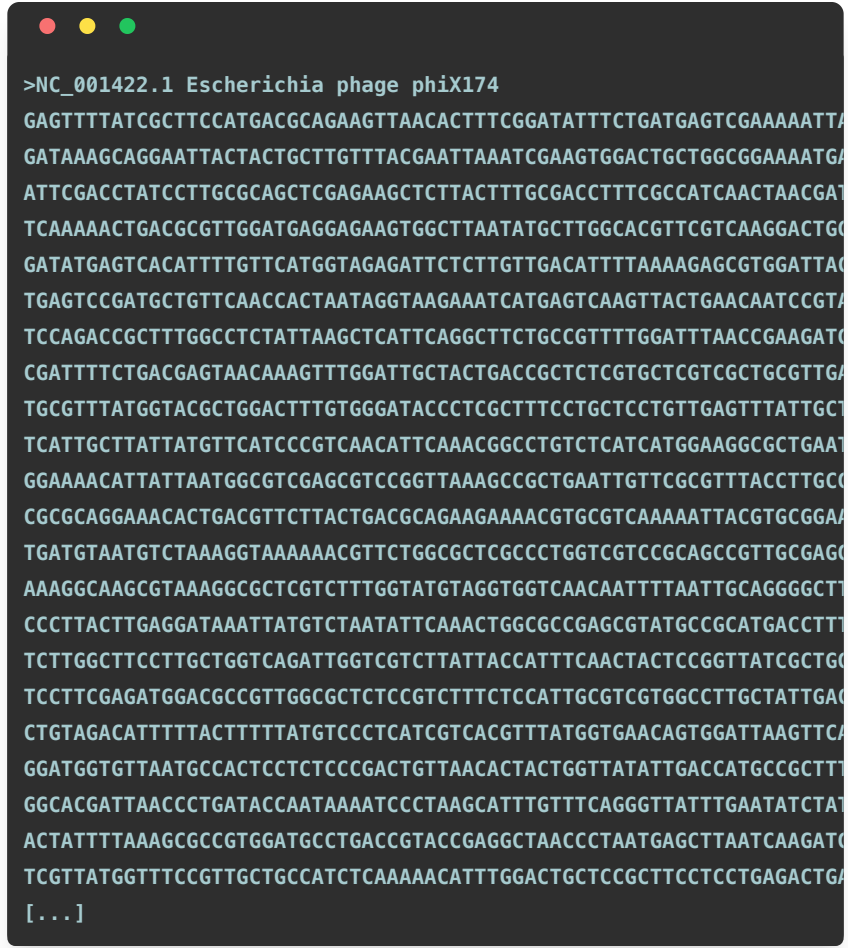
Find the exact origin of a short fragment in a long reference

Quasi-mapping

Which reference is the most-likely origin?

De-novo assembly

Create a long reference from short fragments



🎄🎶❄️ mode on...

Pairwise alignment

- Unique: ng spirits brigghing all the w
- Multi-mapper: Jingle bel
- Base error: what pun it is
- Indels: Jingggge bls

Quasi-mapping

- Within scaffold: Bells on bob tail open sleigh. Hey!
- Within reference: Sankta Lucia

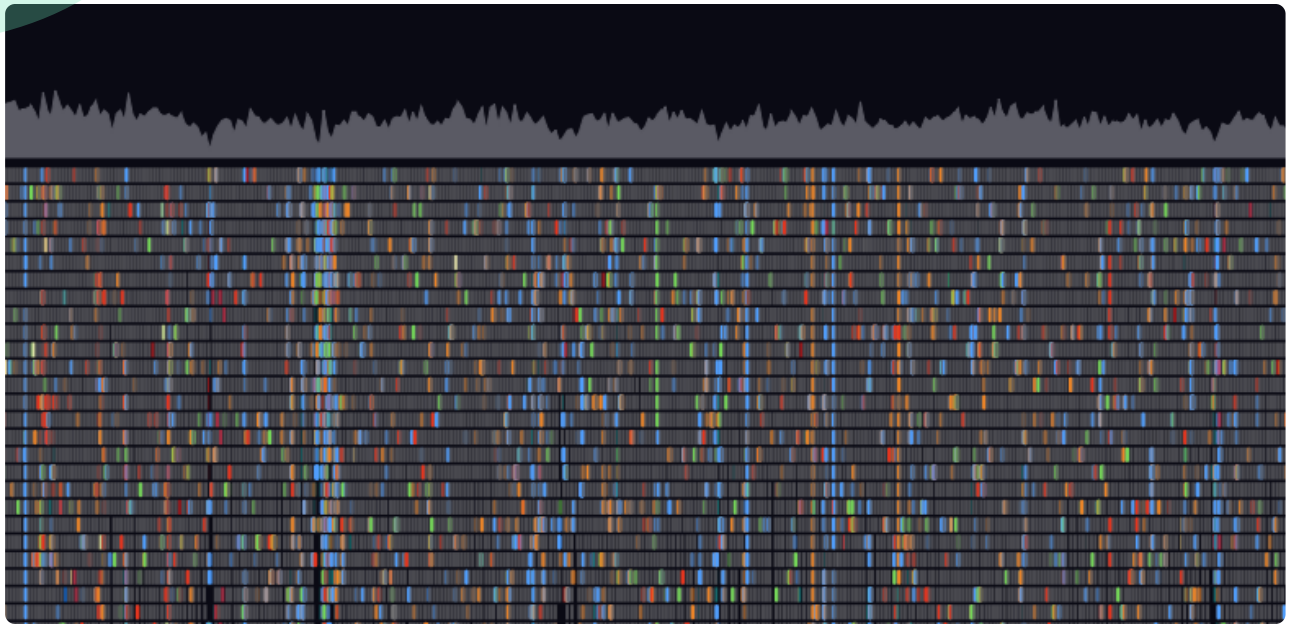
```
Dashing through the snow
In a one-horse open sleigh
O'er the fields we go
Laughing all the way
Bells on bob tail ring
Making spirits bright
What fun it is to ride and sing
A sleighing song tonight! Oh!

Jingle bells, jingle bells,
Jingle all the way.
Oh! what fun it is to ride
In a one-horse open sleigh. Hey!

Jingle bells, jingle bells,
Jingle all the way;
Oh! what fun it is to ride
In a one-horse open sleigh.
```

SAM/BAM/CRAM: Format for pairwise alignments

- Plain text format (SAM)
- Binary & compressed format (BAM/CRAM)
- Contains a header with metadata about reference and aligner
- Prints one alignment per line
- May contain secondary alignments

```
@HD VN:1.0  SO:coordinate
@SQ SN:chr1 LN:197195432
[...]
@PG ID:Bowtie   VN:1.1.2    CL:"bowtie --wrapper basic-0 --threads 4 -v 2 -m 10 -a /ifs/mirror/genome
[...]
SRR2057595.665063_CGCCG 16  chr19   3486359 255 63M *   0   0   *   *   XA:i:0  MD:Z:63 NM:i:0  UG:i:
SRR2057595.1043355_CGCCG    16  chr19   3486359 255 63M *   0   0   *   *   XA:i:0  MD:Z:63 NM:i:0  U
SRR2057595.2024535_CGCCG    16  chr19   3486359 255 63M *   0   0   *   *   XA:i:0  MD:Z:63 NM:i:0  U
SRR2057595.3828487_CGCCG    16  chr19   3486359 255 63M *   0   0   *   *   XA:i:0  MD:Z:63 NM:i:0  U
```

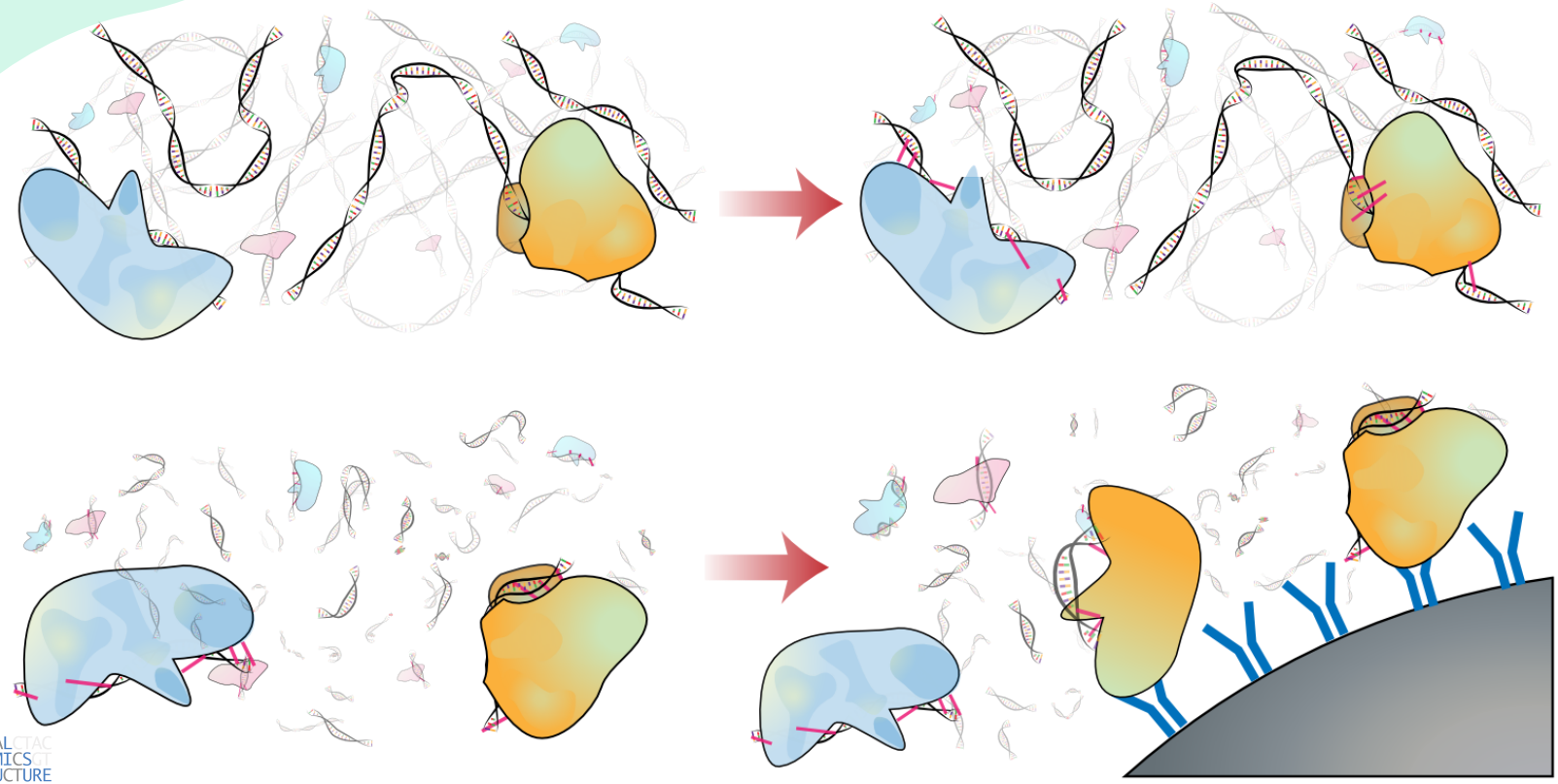
Genome browsers for viewing



Agglomerated errors may represent individual variations (mind the ploidy)

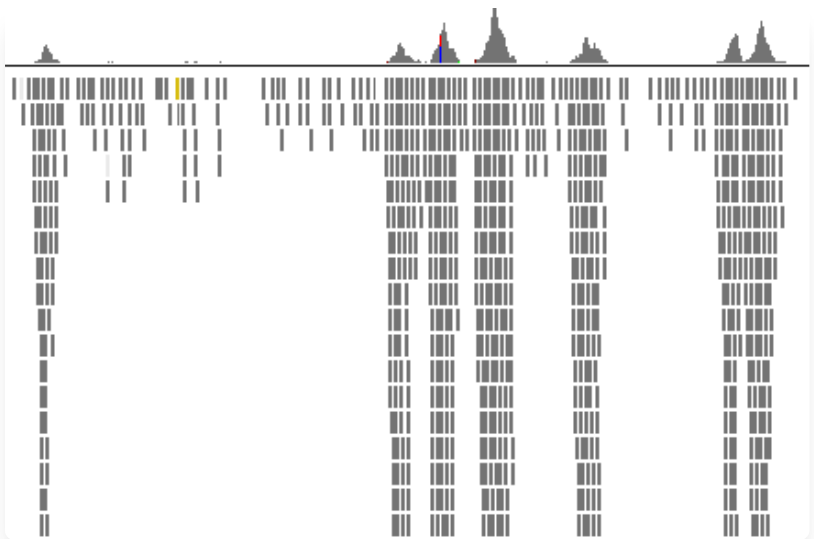
Exemplary analysis

ChIP-seq: Location of DNA-binding proteins



NATIONAL GENOMICS INFRASTRUCTURE

?



ChIP-seq analysis

1. **FastQ generation**
    - Basecalling
    - De-multiplexing of samples
2. **Quality control**
3. **Pairwise alignment**
4. **Peak-calling**

    Discriminate true signal from false positives

ChIP-seq analysis

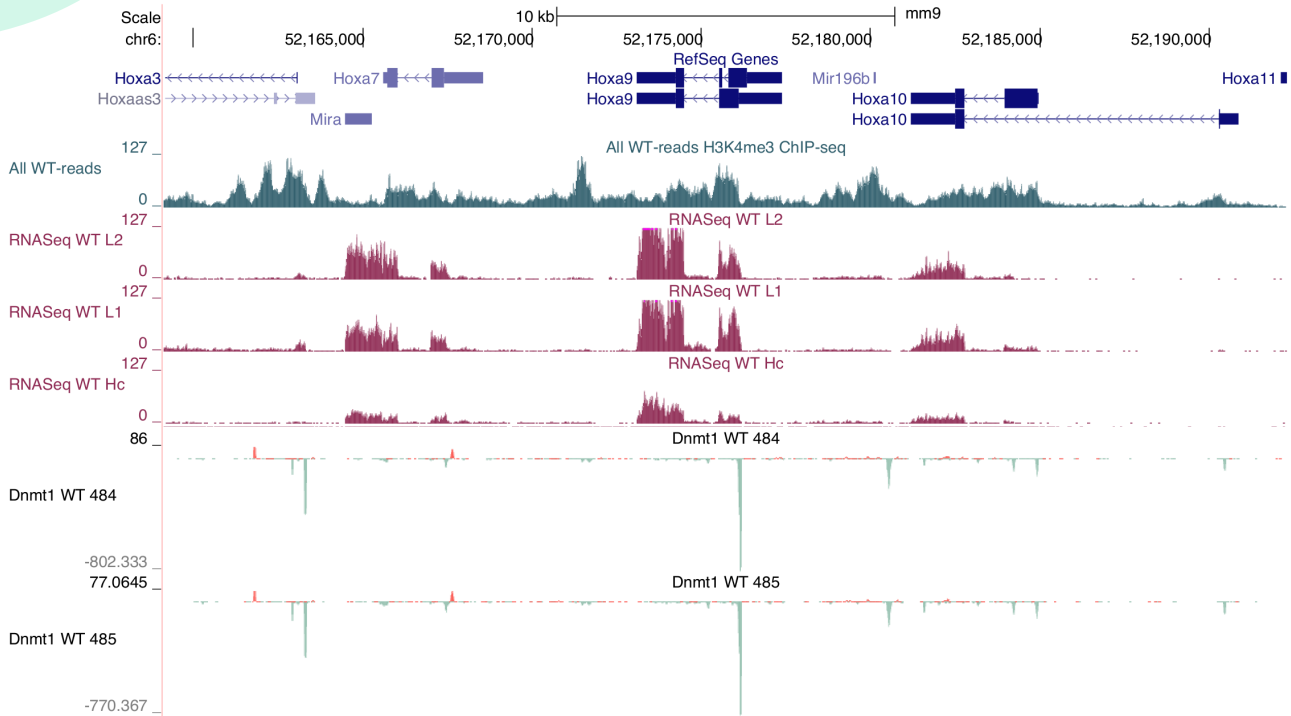
5. **Motif analysis**

(Somewhat outdated by now)

6. **Create context from annotations**

   Find nearby genes or regulatory elements

Different methods result in different signals



Annotations (e.g. gene positions) aid the interpretation

Covers from the 2001 draft sequence release

Reference genomes

Linear reference genomes

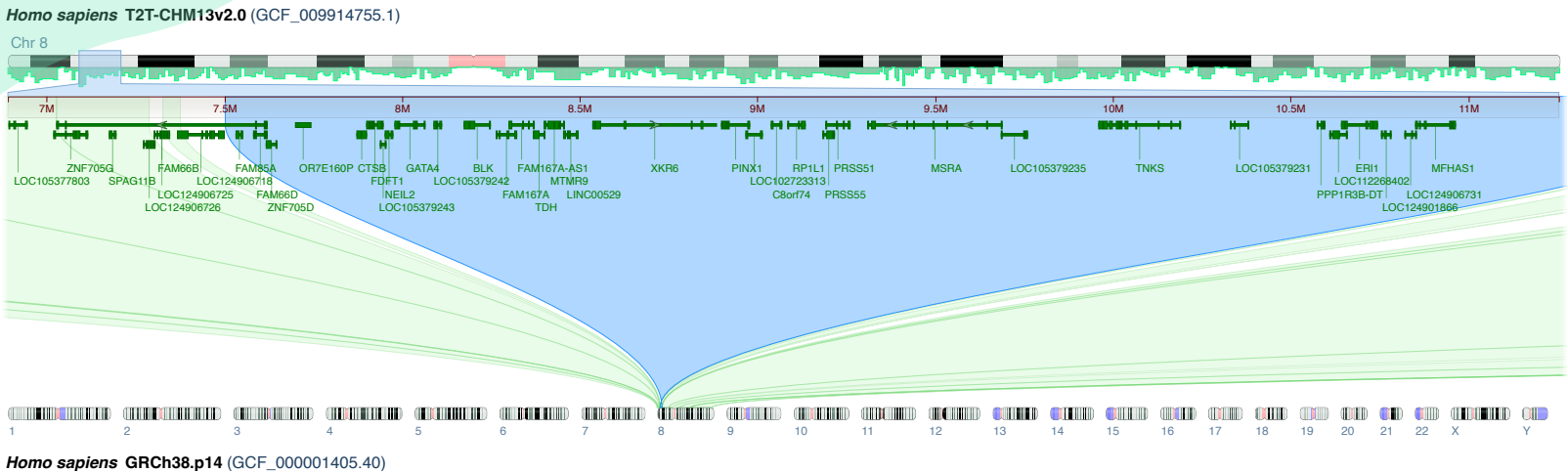
- Are versioned in major (GRCh38, hg38) and minor releases (GRCh38.p14)
- Come in different flavors
- Used for most applications.

T2T assemblies (Human: 2022)

Pangenomes (Human: 2023)

- Combine multiple linear references in a graph representation

T2T vs. "regular" reference genome

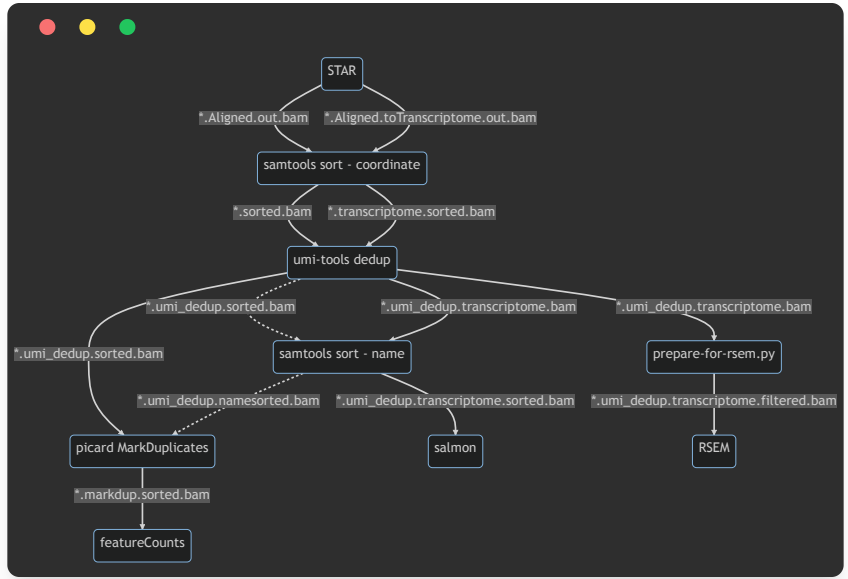


Long-reads filled gaps and revealed inversions

Workflow management systems

Workflow management systems

- Scale analyses to a large number of samples
- Allow for parallel processing
- Agnostic of the compute infrastructure



A workflow (pipeline)

- A sequence of interdependent processes
- Outputs are consumed by other steps

Example

- One process
- Input is a list of three greetings
- The process is run for each input

```
#!/usr/bin/env nextflow
nextflow.enable.dsl=2

process sayHello {
  input:
    val x
  output:
    stdout
  script:
    """
    echo '$x world!'
    """
}

workflow {
  Channel.of('Bonjour',
             'Hej',
             'Hello') | sayHello | view
}
```

Workflow systems

Several hundred workflow systems exist, but in bioinformatics it boils down to those:

Domain-specific language (Pythonic)

Domain-specific language (Groovy, Java)

Honorable mention: Reflow, Workflow Description Language

Workflow "philosophies"

Batch-processing

- Optimised for finite batches of data (one sequencing run)
- Snakemake, Nextflow ...

Stream-processing

- Optimised for a constant stream of data (sensors, stock prices)
- Kafka, Red Panda, Rising Wave ...

Workflow "philosophies"

Dataflow model (inspired)

- Isolated processes linked by dependencies (Directed acyclic graphs)
- Conceptually no dimension for time
- Snakemake, Nextflow ...

Imperative

- Specify a sequence of steps explicitly
- Airflow...

Data-asset lineages

- Each output is the materialization of a task sequence.
- Data assets are "aware" of their pedigree
- Dagster...

nf-core

- Bioinformatic workflow community
- Nextflow pipelines
- 93 public pipelines
- More than 1ooo modules
- A friendly Slack space for questions
- Watch Beginner's guide to nf-core

NATIONAL GENOMICS INFRASTRUCTURE

nf-core

Pipelines

Browse the 93 pipelines that are currently available as part of nf-core.

Filter & Sort

rnafusion ✓ 118

New release!

RNA-seq analysis pipeline for detection gene-fusions

fusion fusion-genes gene-fusion rna rna-seq

3.0.1 released about 17 hours ago

Anvi'o

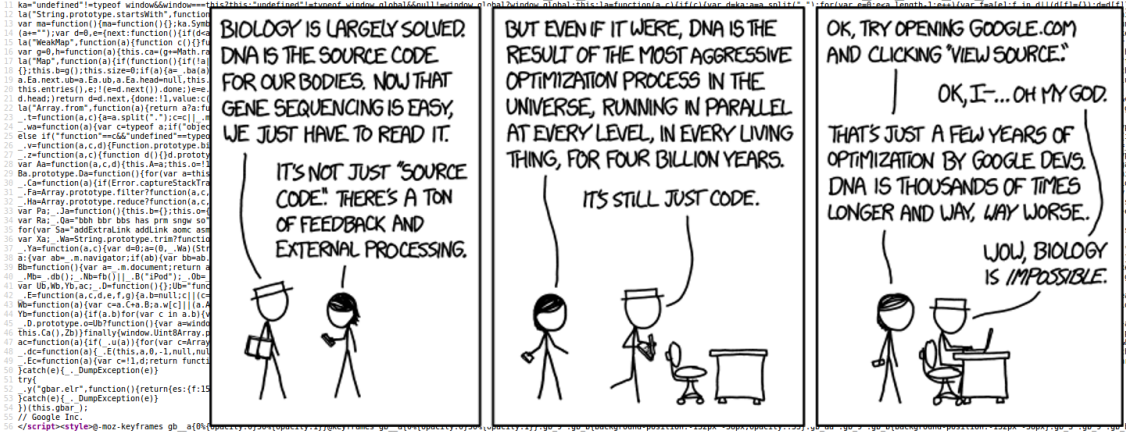
- Microbial omic's community
- Build on top of Snakemake
- Reproducible exploratory analyses with artifacts and workflows
- Publish figures with provenance
- A friendly Discord space for questions
- Watch a video tutorial.

NATIONAL GENOMICS INFRASTRUCTURE

Integrated multi-omics at scale

An open-source, community-driven **an**alysis and **vi**sualization platform for microbial **'o**mics.

UAG (stop)



https://github.com/MatthiasZepper/Lecture-OmicsDataAnlysis