BCB 731:

Defense Against the Dark Arts

Critic: Microbiome analyses of blood and tissues suggest cancer diagnostic approach

November 29th, 2023

nature

Article | Published: 11 March 2020

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

Gregory D. Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, Jad Kanbar, Sandrine Miller-Montgomery, Robert Heaton, Rana Mckay, Sandip Pravin Patel, Austin D. Swafford & Rob Knight ✉

Nature **579**, 567–574 (2020) | Cite this article

Impact

Citations

[HTML] Microbiome analyses of blood and tissues suggest cancer diagnostic approach

GD **Poore**, E Kopylova, Q Zhu, C Carpenter… - Nature, **2020** - nature.com

Systematic characterization of the cancer microbiome provides the opportunity to develop techniques that exploit non-human, microorganism-derived molecules in the diagnosis of a

☆ Save Cite Cited by 676 Related articles All 16 versions

Science

Science Current Issue First release pa

HOME > SCIENCE > VOL. 371, NO. 6536 > THE MICROBIOME AND HUMAN CANCER

REVIEW | CANCER MICROBIOME

The microbiome and human cancer

GREGORY D. SEPICH-POORE, LAURENCE ZITVOGEL, […], AND ROB KNIGHT +3 authors

Cell

Leading Edge

CellPress

Review

Hallmarks of response, resistance, and toxicity to immune checkpoint blockade

Golnaz Morad,[1] Beth A. Helmink,[2] Padmanee Sharma,[3] and Jennifer A. Wargo[1,4,*]
[1]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[2]Department of Surgery, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
[3]Department of Genitourinary Medical Oncology and Immunology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[4]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
*Correspondence: jwargo@mdanderson.org
https://doi.org/10.1016/j.cell.2021.09.020

Tumor microenvironment: Microbial components

More recently, intratumoral microbes—yet another component of the tumor microenvironment that was heretofore underappreciated—has been shown to have significant impact on the antitumor immune responses and responses to ICB (Figure 2). Two recent studies demonstrate a high prevalence of microbes within a broad range of tumors, including those not physically associated with the aerodigestive tract and its commensal organisms (Nejman et al., 2020; Poore et al., 2020). Characterization of the

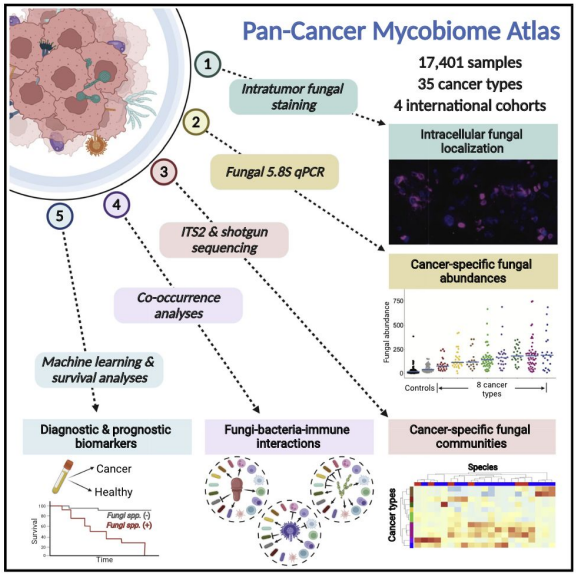
From {Bacteria, Viruses} to Fungi

Resource

Cell

Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions

Graphical abstract



Authors

Lian Narunsky-Haziza,
Gregory D. Sepich-Poore,
Ilana Livyatan, ..., Yitzhak Pilpel,
Rob Knight, Ravid Straussman

Correspondence

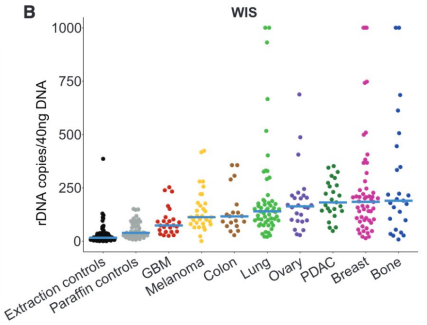
robknight@eng.ucsd.edu (R.K.),
ravidst@weizmann.ac.il (R.S.)

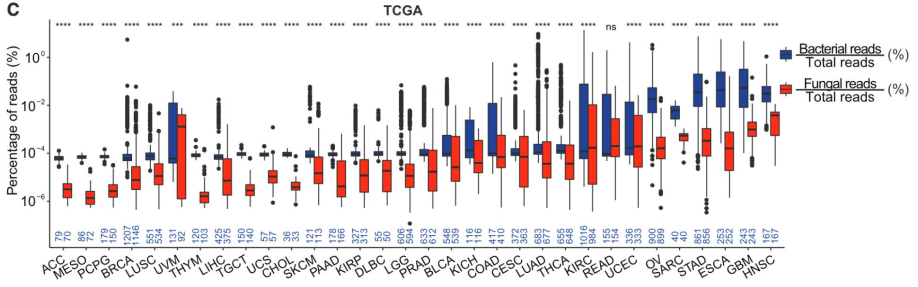
In brief

Characterization of fungi across multiple sample types and patient cohorts across 35 cancer types reveals their distribution, association with immune cell types, and potential prognostic value, including synergy with bacteria.

A

Cohort	Method	Sample types	Total # cancers	Total # samples
Weizmann (WIS)	ITS2 amplicon sequencing	Tumor, NAT, Normal, Controls	8	1183 (+295 controls)
The Cancer Genome Atlas (TCGA)	WGS, RNA-Seq	Tumor, NAT, Blood	33	15,512
Hopkins (Cristiano et al.)	WGS	Plasma, Normal	8	537
UCSD (Poore et al.)	Shotgun (WGS)	Plasma, Normal, Controls	3	169 (+52 controls)
Total	-	-	35	17,401 (+347 controls)





From Paper to Startup

Micronoma Launches with $3 Million Seed Funding

First early cancer-detection company to use microbial biomarkers in diagnostics

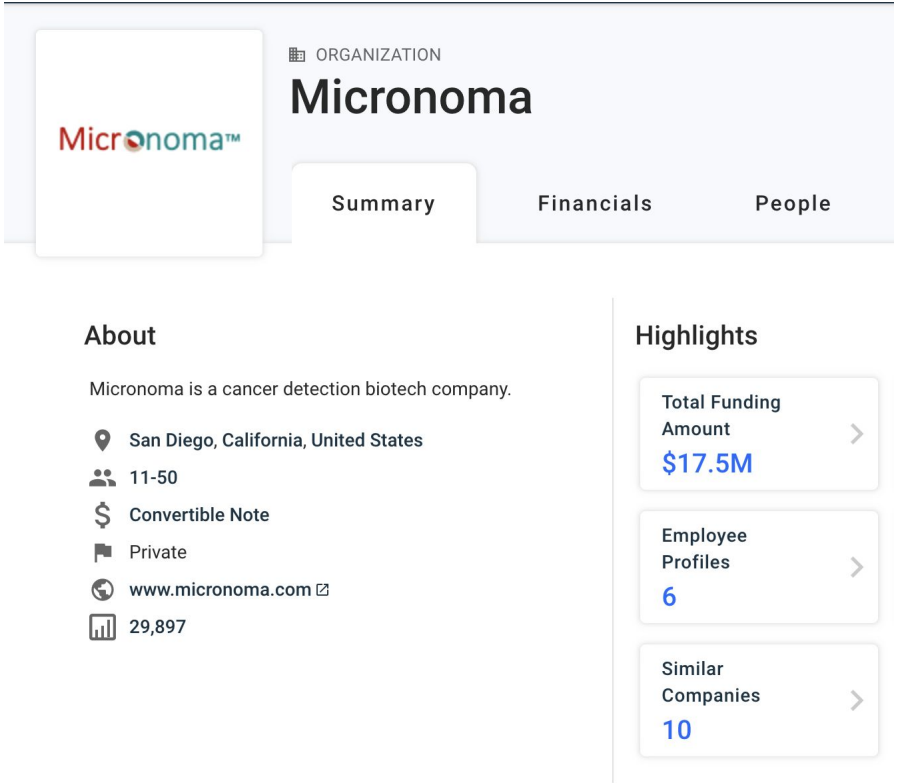
San Diego, Calif. (August 12, 2020). Micronoma, an early cancer-detection biotech that seeks to develop and commercialize a minimally invasive, microbiome-based method, announced today that it has closed a $3 million seed financing round led by microbiome-focused investor, SymBiosis, LLC.

The proceeds from funding will be used to further advance cancer detection technology with the development of pioneering microbiome research.

Micronoma was incorporated in June 2019 by its three co-founders: Sandrine Miller-Montgomery, CEO of Micronoma, previously Executive Director of the Center for Microbiome Innovation; Dr. Rob Knight, Director of the Center for Microbiome Innovation at the University of California San Diego (UC San Diego) and Greg Poore, an M.D.-Ph.D. candidate at UC San Diego School of Medicine and co-inventor. Micronoma has exclusively licensed the original IP on microbial-based cancer diagnostics and therapeutics, created by Poore and Knight, from UC San Diego.

Micronoma's technology has shown that distinct cancer types can be diagnosed sensitively and specifically solely using microbial nucleic acids in human blood and tissues. This approach was first demonstrated using a cancer sequencing database of >10,000 patients from 32 cancer types, then validated on >150 real-world patient samples from Moore's Cancer Center at UC San Diego Health, across three cancer types and comparing results to those from healthy, non-cancer controls. The results of the study were published in *Nature* in March 2020 (doi: 10.1038/s41586-020-2095-1).

Micronoma gets ~$17M in funding



Multi Cancer Early Detection Market is Growing at Rapid Pace and Expected to

Recent Developments:

- In September 2020-Orbis Biosciences, Inc., a provider of sample-to-result solutions for molecular diagnostic testing, was purchased QIAGEN N.V. With this acquisition, QIAGEN will be able to offer more sample-to-result cancer testing options.
- In September 2019 -The Guardant360 CDX liquid biopsy test for patients with advanced cancer went on sale from Guardant Health, Inc. The test is made to offer thorough genetic profiling of the tumor in a patient in order to inform therapy choices. The Guardant360 CDX test was successful in detecting actionable genetic changes in 77% of instances in a clinical trial, including more than 500 patients with advanced solid.

List of Prominent Players in the Multi-Cancer Early Detection Market:

- Micronoma Inc
- Anpac Bio
- EarlyDiagnostics, Inc
- Early is Good (EIG)
- Cansense
- Freenome Holdings, Inc.
- Oncocyte Corporation
- SeekIn
- Naveris
- VESEN, Inc.
- Grail, LLC (Illumina, Inc.),

Forbes 30 Under 30

Micronoma
November 29, 2022 ·

Our co-founder, Greg Sepich-Poore made the Forbes 30 Under 30 Healthcare list!

We are so proud of his unwavering drive toward revolutionizing the cancer diagnostics field; developing the first microbiome-driven liquid biopsy designed to detect cancer earlier than conventional screening methods, starting with lung cancer. And this is only the beginning! Congratulations, Greg, a well deserved recognition. Learn more from the scientists and entrepreneurs who are taking on some of healthcare's biggest challenges, here.
https://bit.ly/3FcC3QN

...and grants for academic collaborators

Micronoma, NYU Partner to Develop Microbial Biomarkers of Lung Cancer

May 19, 2023 | staff reporter

NEW YORK — Micronoma said on Thursday that it has partnered with New York University on a National Cancer Institute-funded project to identify microbial biomarkers that can be used to predict non-small cell lung cancer (NSCLC) and its chance of recurrence.

The grant was awarded to NYU Grossman School of Medicine researcher Leopoldo Segal to identify microbial and host biomarkers, obtained from blood and lower airway samples of patients with lung nodules, that can predict early-stage NSCLC, according to its abstract. Potential biomarkers will be used to develop diagnostics — including ones based on next-generation sequencing, metabolite measurement, and custom-made NanoString panels — that can identify patients at high risk of NSCLC and disease recurrence following complete surgical resection.

The five-year grant is worth $703,424 in its first year.

Micronoma said that it is currently working with NYU Langone Health to identify the blood-based biomarkers. The San Diego-based company said it could potentially commercialize a test under CLIA regulations later this year based on its OncobiotaLUNG microbiome-based liquid biopsy platform.

FDA Breakthrough designation

FDA Grants Breakthrough Device Designation to OncobiotaLUNG Assay for Lung Cancer Detection

January 12, 2023
Conor Killmurray

Article

The FDA has granted the OncobiotaLU

The FDA has provided the Onc
Micronoma.[1]

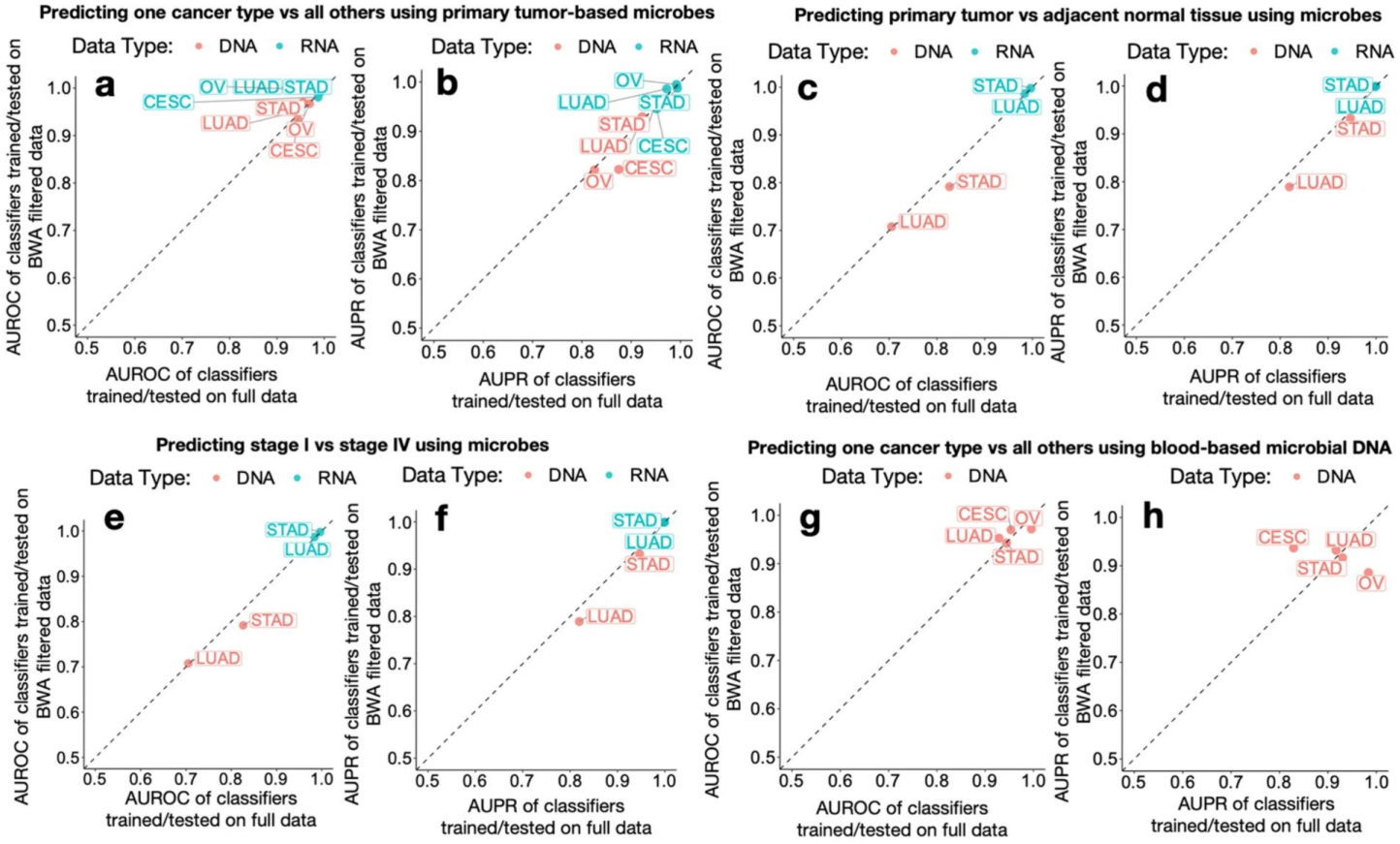
Whole-genome and whole-transcriptome sequencing studies were re-examined in *The Cancer Genome Atlas* (TCGA) from 33 types of cancer in a total of 18,116 samples of treatment naïve patients. They found unique microbial signatures in both tissue and blood samples among the most major types of cancer. The blood signatures then remained predictive of disease for patients with stage I to stage IIa. According to the research, findings suggest there are widespread associations between specific microbiota across multiple cancer types and the potential for a diagnosis.

By looking at these, the OncobiotaLUNG assay can determine the risk factor of a nodule while providing a minimally invasive way of collecting the information compared with the standard of care biopsy in lung cancer. However, more widespread research is warranted beyond the FDA's newest designation.

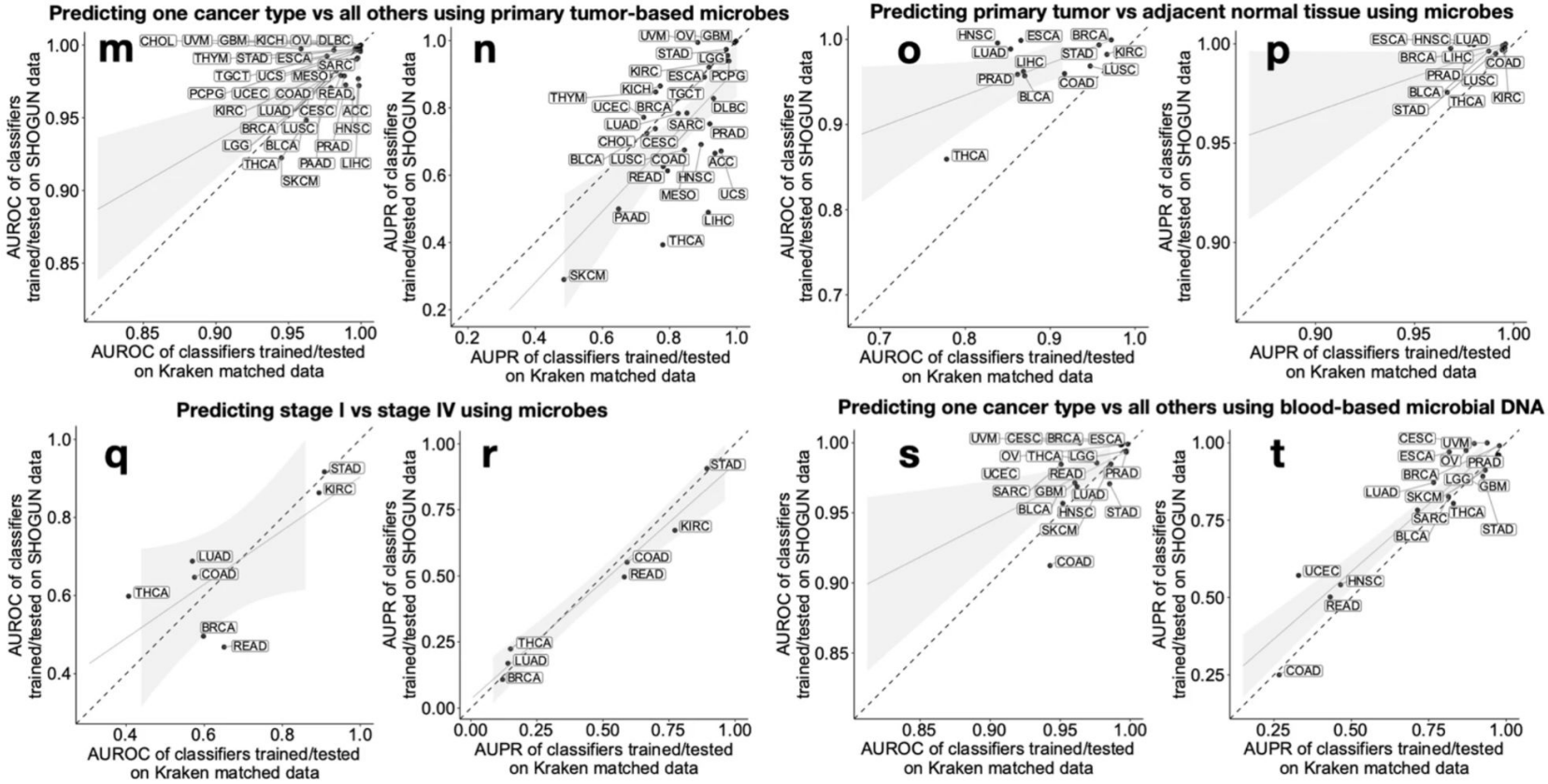
The OncobiotaLUNG assay is microbiome-driven liquid biopsy platform used for the early detection of lung cancer. The device categorizes lung nodule samples into those that are considered high or low-risk for malignancy based off of a blood draw from the patient. According to the new designation, the company is expecting prioritized reviews from the agency regarding upcoming trials further analyzing the effectiveness of the assay.

Lots of Sanity Checks in the Paper

Not an artifact of kmers (or DNA/RNA)



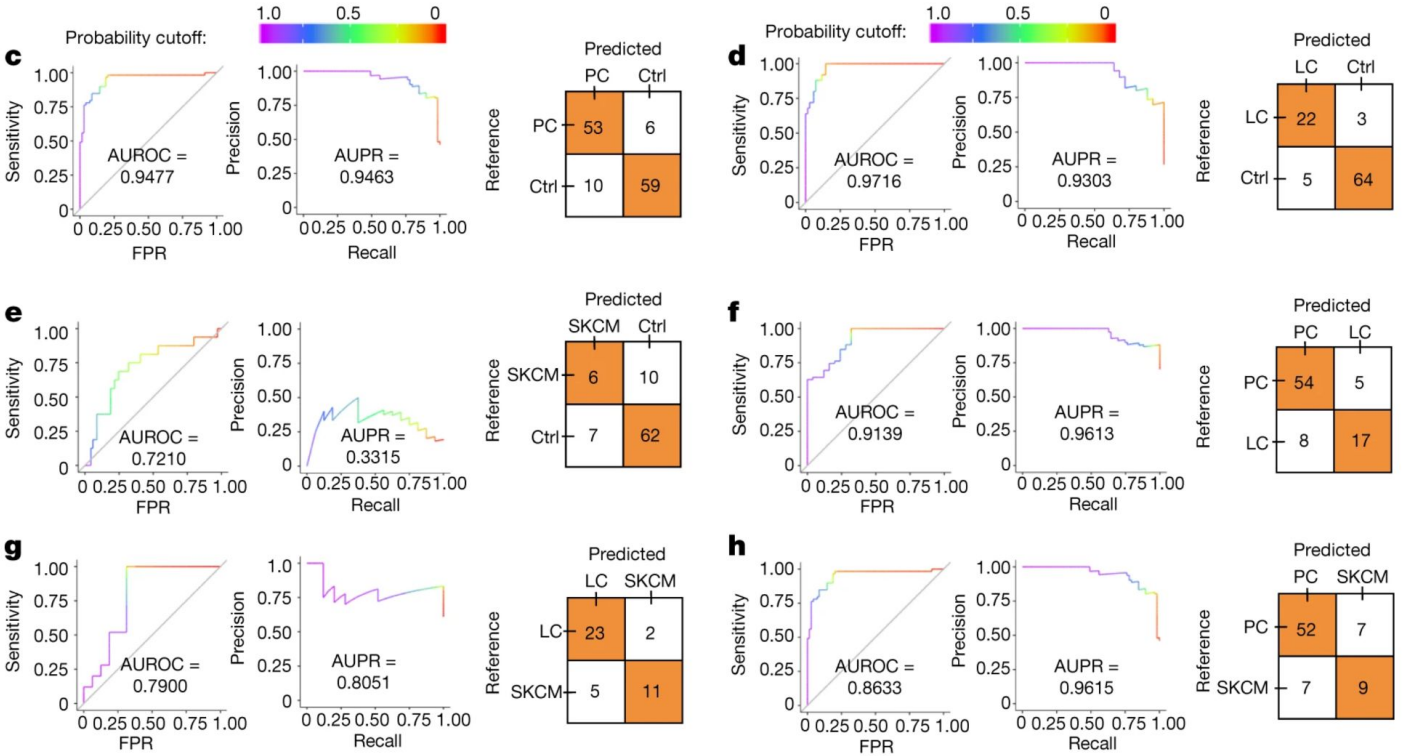
Not an artifact of particular kmer tool



Validation study: cell-free DNA

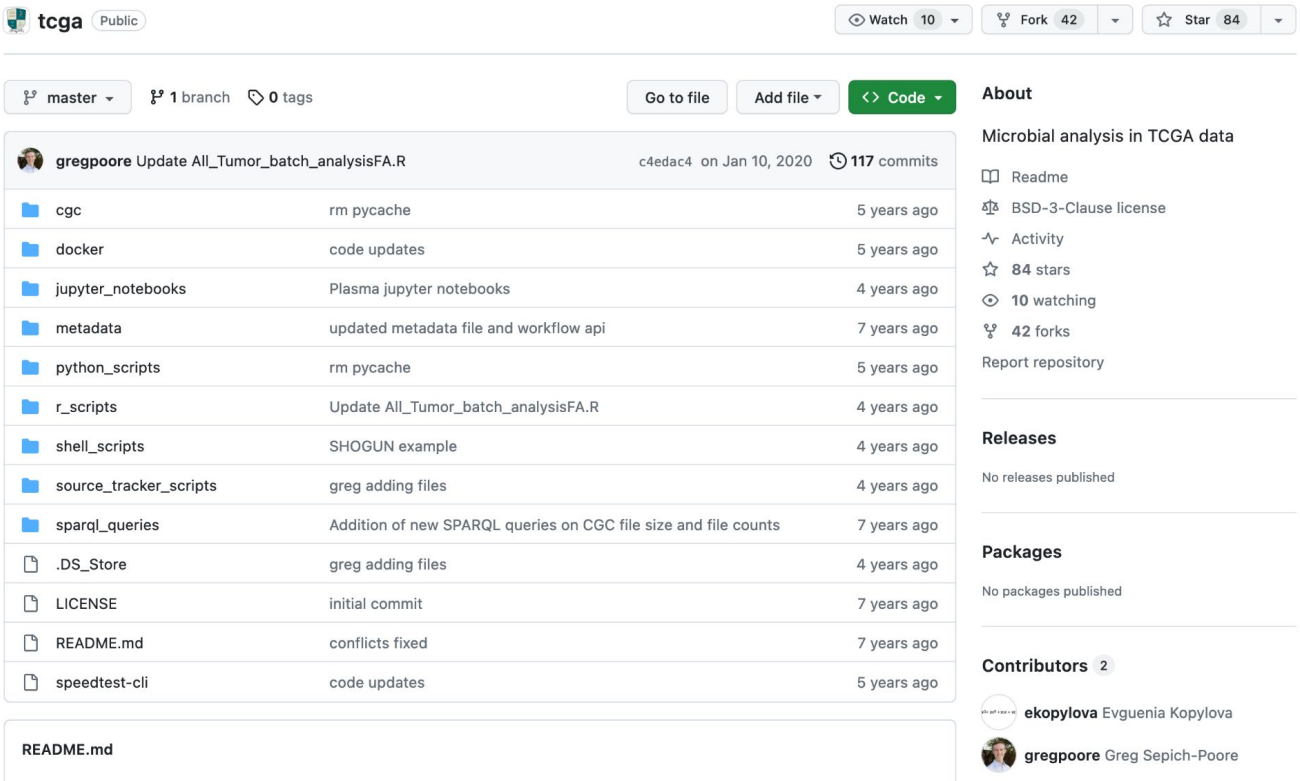
Variable	Control	PC	LC	SKCM
Sample size (*n*)	69	59	25	16
Age (years) (mean ± s.d.)	45.00 ± 12.80	69.70 ± 9.63	69.50 ± 9.88	58.50 ± 13.00
Sex (% female)	24.6%	0%	68%	12.5%
Known condition(s)/ subtype(s) (*n*)	HIV-free (69)	HSPC (32) CRPC (27)	LUAD (15) LUSC (5) Sarcomatoid (1) Large cell (1) NOS (3)	SKCM (16)

Validation performance



Open Data / Open Code

A real GitHub repository!



tcga Public

Watch 10 | Fork 42 | Star 84

master | 1 branch | 0 tags | Go to file | Add file | Code

gregpoore Update All_Tumor_batch_analysisFA.R — c4edac4 on Jan 10, 2020 — 117 commits

cgc	rm pycache	5 years ago
docker	code updates	5 years ago
jupyter_notebooks	Plasma jupyter notebooks	4 years ago
metadata	updated metadata file and workflow api	7 years ago
python_scripts	rm pycache	5 years ago
r_scripts	Update All_Tumor_batch_analysisFA.R	4 years ago
shell_scripts	SHOGUN example	4 years ago
source_tracker_scripts	greg adding files	4 years ago
sparql_queries	Addition of new SPARQL queries on CGC file size and file counts	7 years ago
.DS_Store	greg adding files	4 years ago
LICENSE	initial commit	7 years ago
README.md	conflicts fixed	7 years ago
speedtest-cli	code updates	5 years ago

README.md

About

Microbial analysis in TCGA data

Readme
BSD-3-Clause license
Activity
84 stars
10 watching
42 forks
Report repository

Releases

No releases published

Packages

No packages published

Contributors 2

ekopylova Evguenia Kopylova
gregpoore Greg Sepich-Poore

All the preprocessing actually online

tcga / r_scripts /

All_Tumor_ML_predict_LvsH_tumor_stageFA.R	greg adding files
All_Tumor_ML_predict_cancer_1VsAllFA_Patent.R	greg adding files
All_Tumor_ML_predict_tumor_vs_normal_FA_Patent.R	greg adding files
All_Tumor_batch_analysisFA.R	Update All_Tumor_batch_analysisFA.R
All_Tumor_clinical_analysesFA.R	greg adding files
Bubble-Plots-of-Spiked-Pseudo-Contaminant-Feature-Imp...	Bubble plots script
Decontamination-by-Plate-Center-Combination.R	Plate-Center decontamination script
Machine-Learning-Performance-Vs-Minority-Class.R	ML performance vs minority class size
Plasma-Bootstrap500-Machine-Learning.R	R script examples for revised paper submission
Plasma-Subsampling-Machine-Learning-100Iterations.R	Plasma subsampling script
Plasma-Voom-SNM-Normalize-Age-and-Sex.R	R script examples for revised paper submission
Plasma-age-regression-loocv.R	R script examples for revised paper submission
Plasma-permutation-testing-age-and-sex.R	R script examples for revised paper submission
Run-All-TCGA-Machine-Learning-Comparisons-Across-All-...	R script examples for revised paper submission
TCGA-Simulation-for-Estimating-Plasma-Experiment-Samp...	R script examples for revised paper submission
Tumor_Microbe_SubtypingFA.R	greg adding files
Validation_Data_Splits_and_Model_Testing.R	greg adding files
cgc_sparql_metadata_r_api.R	File organization improvements and addition of Python script for pull...
predCancer1VsAllFA_Patent.R	greg adding files
predTumorVsNormalFA_Patent.R	greg adding files

tcga / python_scripts /

..	
tests	code updates
add_metadata_columns.py	code updates
cgc_compare_files.py	code updates
cgc_compute_stats.py	update scripts
cgc_create_tcga_workflow_task.py	code updates
cgc_create_viral_output_task.py	code updates
cgc_get_download_links.py	get download links for stats files
cgc_metadata_to_qiime_mapping_file.py	Updating docker and python files
cgc_python_api_by_disease_type.py	Completed unittest for cgc_python_api_by_disease_type.py file. As an ...
cgc_samtools_bam2fasta_worklfow_task.py	code updates
cgc_sparql_metadata_python_api.py	File organization improvements and addition of Python script for pull...
commands.txt	code updates
kraken_to_fasta.py	kraken to fasta complete
make_merged_CDE_table.py	Conversion of Python scripts from 2.7 to 3.5
parse_kraken_to_biom.py	update docker
process_fasta_stats.py	code updates
rename_sample_ids.py	code updates
repophlan_genomeID_taxonomy.py	code updates

External tools run with reproducible pipelines / Docker containers

tcga / docker /

ekopylova code updates

Name	Last commit message
..	
braken_docker	Addition of other docker container directories
centrifuge_docker	Addition of other docker container directories
docker-parse-kraken-to-biom	code updates
docker-parse-kraken-to-fasta	code updates
docker-qiime	code updates
kraken_docker	Addition of other docker container directories
meta2qiime_docker	Addition of other docker container directories
.DS_Store	code updates

Detailed notebooks

In [200…

```
# Save dict/DataFrame as a pickle file
pickle.dump(caseGroupedDict, open("caseGroupedDict.p", "wb"), protocol=4)
```

In [201…

```
# Save dict/DataFrame as a pickle file
pickle.dump(fileGroupedDict, open("fileGroupedDict.p", "wb"), protocol=4)
```

In [202…

```
# Load pickle files to demonstrate that the Pandas formatting is not lost in the bina
caseTesting = pickle.load(open("caseGroupedDict.p", "rb"))
fileNameTesting = pickle.load(open("fileGroupedDict.p", "rb"))
```

In [203…

```
# Test for equality between saved and loaded file as a sanity check.
# Note that values saved as 'NaN' will show as 'False'.
caseTestingKey = '40C217EE-429F-41F4-ADC6-DB9186622B17'
caseTesting[caseTestingKey] == caseGroupedDict[caseTestingKey]
```

Out[203…

	ageAtDiagnosis	aliquot_name	case_name	clinical_m_label	clinical_n_label	clinical_t_la
16759	True	True	True	True	True	
16760	True	True	True	True	True	

2 rows × 59 columns

In [204…

```
# Test for equality between saved and loaded file as a sanity check.
# Note that values saved as 'NaN' will show as 'False'.
fileTestingKey = '01249B8C-2E9E-4CEA-B39A-485863AB231A'
fileNameTesting[fileTestingKey] == fileGroupedDict[fileTestingKey]
```

All the Data!

cfDNA	Jan 24, 2020 at 12:00 AM	--
Kraken	Jan 24, 2020 at 12:00 AM	--
Kraken-Plasma-Validation-Raw-Data.csv	Jan 24, 2020 at 12:00 AM	233 KB
Kraken-Plasma-Voom-SNM-Age-And-Sex-Data.csv	Jan 24, 2020 at 12:00 AM	1.4 MB
Metadata-Plasma-Filtered-For-Analysis.csv	Jan 24, 2020 at 12:00 AM	21 KB
Metadata-Plasma-For-Decontam-With-Negative-And-Positive-Controls.csv	Jan 24, 2020 at 12:00 AM	24 KB
SHOGUN	Jan 24, 2020 at 12:00 AM	--
SHOGUN-Plasma-Validation-Raw-Data.csv	Jan 24, 2020 at 12:00 AM	402 KB
SHOGUN-Plasma-Validation-Voom-SNM-Age-And-Sex-Data.csv	Jan 24, 2020 at 12:00 AM	2.6 MB
TCGA	Jan 24, 2020 at 12:00 AM	--
BWA	Jan 24, 2020 at 12:00 AM	--
BWA-Filtered-TCGA-DNA-Raw-Data.csv	Jan 24, 2020 at 12:00 AM	4.7 MB
BWA-Filtered-TCGA-DNA-Voom-SNM-Data.csv	Jan 24, 2020 at 12:00 AM	16.4 MB
BWA-Filtered-TCGA-RNA-Raw-Data.csv	Jan 24, 2020 at 12:00 AM	11.4 MB
BWA-Filtered-TCGA-RNA-Voom-SNM-Data.csv	Jan 24, 2020 at 12:00 AM	77.3 MB
Metadata-TCGA-BWA-DNA-4Cancers.csv	Jan 24, 2020 at 12:00 AM	641 KB
Metadata-TCGA-BWA-RNA-4Cancers.csv	Jan 24, 2020 at 12:00 AM	1.6 MB
Kraken	Apr 4, 2020 at 12:00 AM	--
BWA-Filtered-TCGA-DNA-Raw-Data.csv	Apr 4, 2020 at 12:00 AM	4.7 MB
BWA-Filtered-TCGA-DNA-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	16.4 MB
BWA-Filtered-TCGA-RNA-Raw-Data.csv	Apr 4, 2020 at 12:00 AM	11.4 MB
BWA-Filtered-TCGA-RNA-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	77.3 MB
Kraken-Matched2BWA-TCGA-DNA-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	17.3 MB
Kraken-Matched2BWA-TCGA-RNA-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	79.1 MB
Kraken-Plasma-Validation-Raw-Data.csv	Apr 4, 2020 at 12:00 AM	233 KB
Kraken-Plasma-Voom-SNM-Age-And-Sex-Data.csv	Apr 4, 2020 at 12:00 AM	1.4 MB
Kraken-TCGA-50-50-Validation-Split-1-Raw-Data.csv	Apr 4, 2020 at 12:00 AM	40.4 MB
Kraken-TCGA-50-50-Validation-Split-1-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	270.6 MB
Kraken-TCGA-50-50-Validation-Split-2-Raw-Data.csv	Apr 4, 2020 at 12:00 AM	40.3 MB
Kraken-TCGA-50-50-Validation-Split-2-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	272.1 MB
Kraken-TCGA-Matched2SHOGUN-Raw-Data.csv	Apr 4, 2020 at 12:00 AM	51.6 MB
Kraken-TCGA-Matched2SHOGUN-Voom-SNM-Quantile-Data.csv	Apr 4, 2020 at 12:00 AM	396.2 MB
Kraken-TCGA-Raw-Data-17625-Samples.csv	Apr 4, 2020 at 12:00 AM	80.5 MB
Kraken-TCGA-Raw-Data-All-18116-Samples.csv	Apr 4, 2020 at 12:00 AM	82.9 MB
Kraken-TCGA-Simulations-Broad-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	3.7 MB
Kraken-TCGA-Simulations-HMS-Voom-SNM-Data.csv	Apr 4, 2020 at 12:00 AM	10.6 MB
Kraken-TCGA-Voom-SNM-All-Putative-Contaminants-Removed-Data.csv	Apr 4, 2020 at 12:00 AM	427.3 MB
Kraken-TCGA-Voom-SNM-Full-Data.csv	Apr 4, 2020 at 12:00 AM	561.3 MB
Kraken-TCGA-Voom-SNM-Likely-Contaminants-Removed-Data.csv	Apr 4, 2020 at 12:00 AM	491.4 MB

A Nice Data Portal

Cancer Microbiome

CENTER FOR microbiome INNOVATION

KNIGHTLAB

UC San Diego Moores Cancer Center

- Kraken microbe abundances in TCGA cancer types
- Shogun microbe abundances in TCGA cancer types
- Kraken TCGA model performance and feature list
- Shogun TCGA model performance and feature list
- Plasma cell-free microbial abundances (validation)

TCGA Abbreviation	Cancer Type
TCGA-ACC	Adrenocortical Carcinoma
TCGA-BLCA	Bladder Urothelial Carcinoma
TCGA-LGG	Brain Lower Grade Glioma
TCGA-BRCA	Breast Invasive Carcinoma
TCGA-CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma

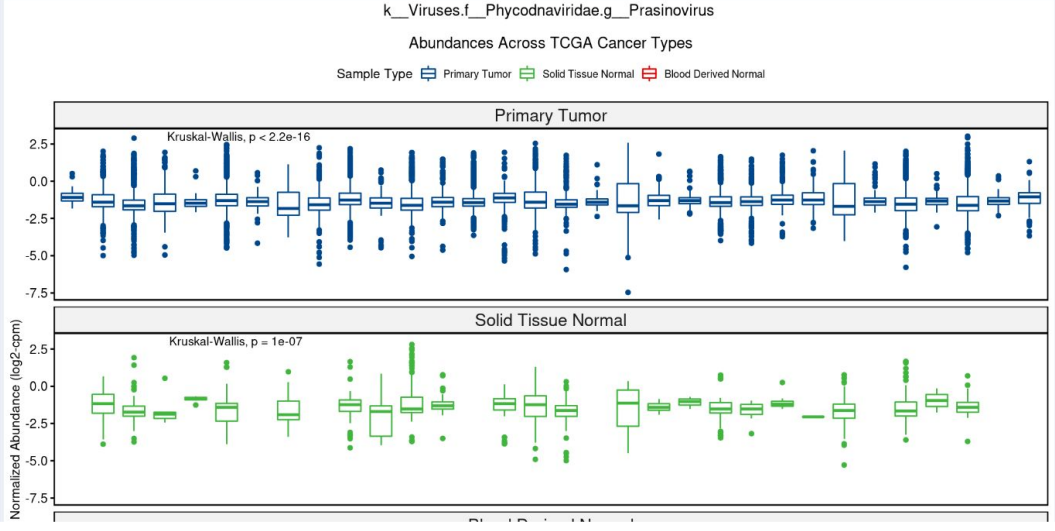
Kraken-Derived Normalized Microbial Abundances in TCGA

Select microbe of interest to display its normalized abundance distribution across TCGA cancer types.

Note: Kraken-derived data were generated using Kraken 1 against 59,974 quality filtered microbial genomes, originally downloaded from RepoPhlan on 14 June 2016 (original size was 71,782 before quality filtering). This included bacteria, archaea, and viruses. This database is *different* than the 'Web of Life' database (accepted for publication; https://biocore.github.io/wol/) used for generating Shogun-derived data in TCGA. The Voom-SNM normalized data are plotted below.

Either (i) select from the drop-down list or (ii) click on the name and hit backspace, then start typing the name of your microbe of interest (it will autocomplete). After selecting your microbe of interest, please wait a second for the plot to appear below. A basic Kruskal-Wallis test is performed for each sample type to show if the microbe varies across cancer types. The genus for the HPV virus (Alphapapillomavirus) is pre-selected for you.

k__Viruses.f__Phycodnaviridae.g__Prasinovirus



Some Red Flags

What's that first taxon?

Cancer Microbiome

CENTER FOR microbiome INNOVATION

KNIGHTLAB

UC San Diego Moores Cancer Center

- Kraken microbe abundances in TCGA cancer types
- Shogun microbe abundances in TCGA cancer types
- Kraken TCGA model performance and feature list
- Shogun TCGA model performance and feature list
- Plasma cell-free microbial abundances (validation)

TCGA Abbreviation	Cancer Type
TCGA-ACC	Adrenocortical Carcinoma
TCGA-BLCA	Bladder Urothelial Carcinoma
TCGA-LGG	Brain Lower Grade Glioma
TCGA-BRCA	Breast Invasive Carcinoma
TCGA-CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma

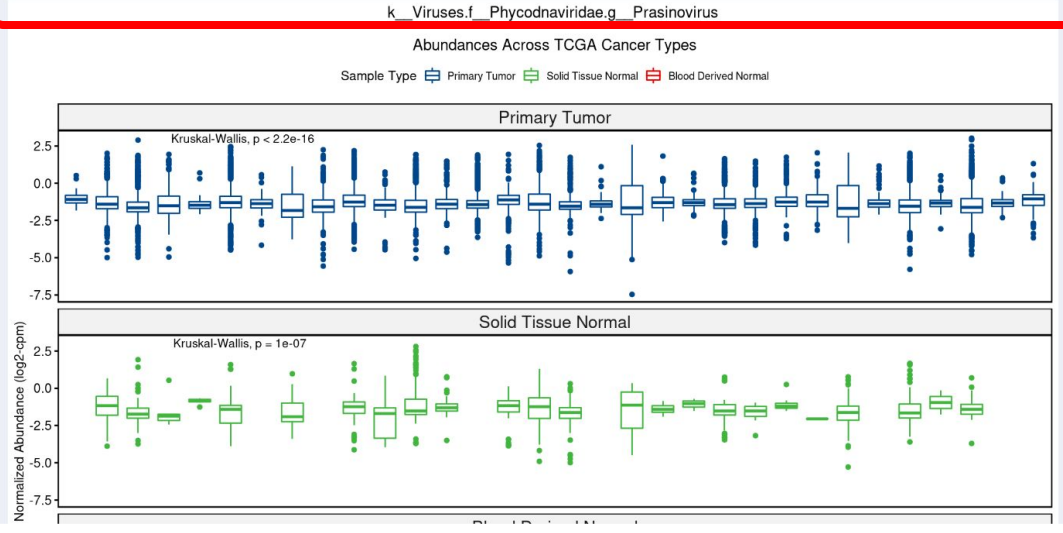
Kraken-Derived Normalized Microbial Abundances in TCGA

Select microbe of interest to display its normalized abundance distribution across TCGA cancer types.

Note: Kraken-derived data were generated using Kraken 1 against 59,974 quality filtered microbial genomes, originally downloaded from RepoPhlan on 14 June 2016 (original size was 71,782 before quality filtering). This included bacteria, archaea, and viruses. This database is *different* than the 'Web of Life' database (accepted for publication; https://biocore.github.io/wol/) used for generating Shogun-derived data in TCGA. The Voom-SNM normalized data are plotted below.

Either (i) select from the drop-down list or (ii) click on the name and hit backspace, then start typing the name of your microbe of interest (it will autocomplete). After selecting your microbe of interest, please wait a second for the plot to appear below. A basic Kruskal-Wallis test is performed for each sample type to show if the microbe varies across cancer types. The genus for the HPV virus (Alphapapillomavirus) is pre-selected for you.

k__Viruses.f__Phycodnaviridae.g__Prasinovirus



Huh, that's weird...

Prasinovirus

Article Talk

From Wikipedia, the free encyclopedia

Prasinovirus is a genus of large double-stranded DNA viruses, in the family *Phycodnaviridae* that infect phytoplankton in the *Prasinophyceae*. There are three groups in this genus,[1][2] including *Micromonas pusilla virus SP1*, which infects the cosmopolitan photosynthetic flagellate *Micromonas pusilla*.[3]

There is a large group of genetically diverse but related viruses that show considerable evidence of lateral gene transfer.[4][5]

Huh, that's weird...

Prasinovirus

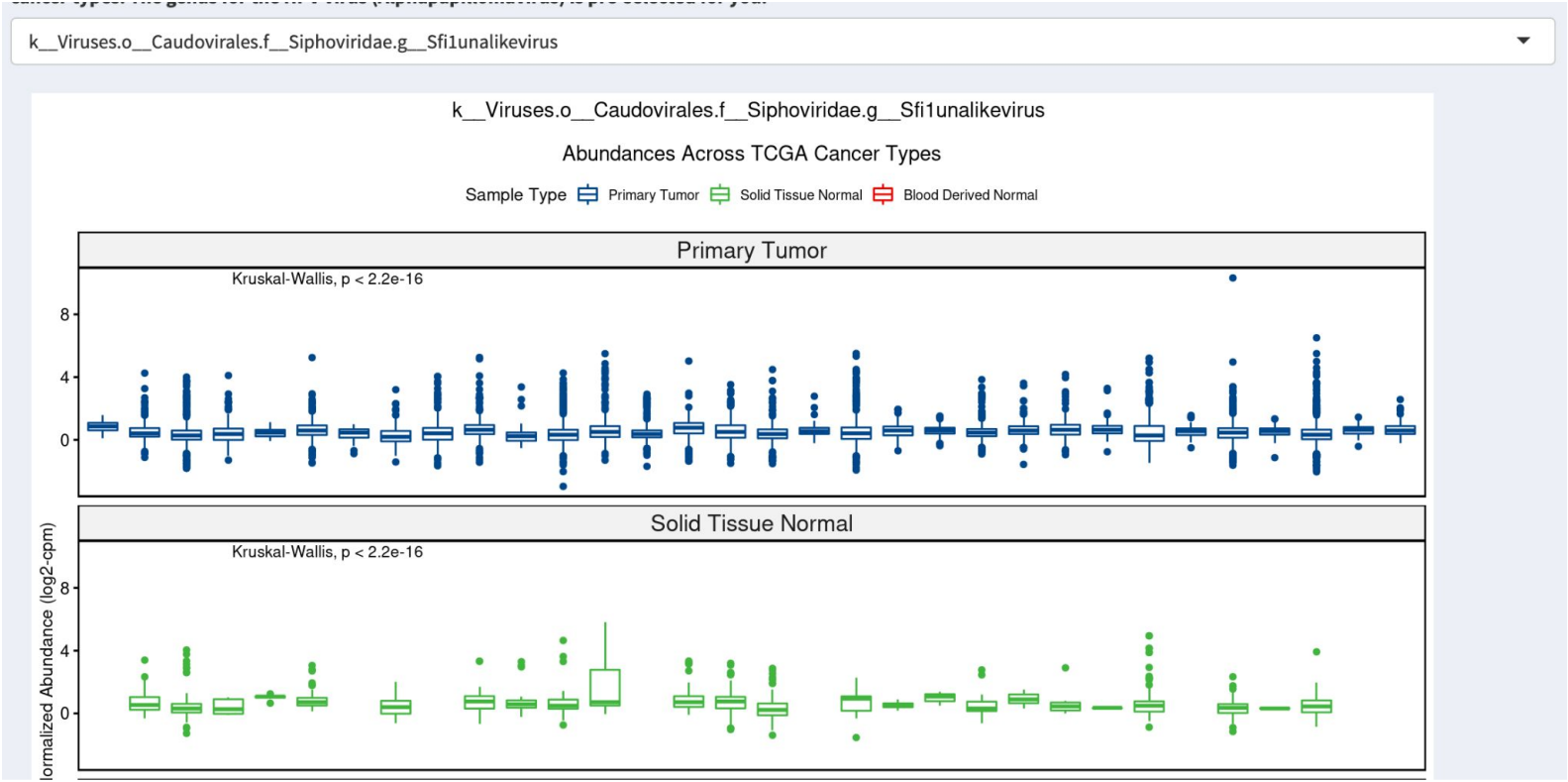
Article Talk

From Wikipedia, the free encyclopedia

Prasinovirus is a genus of large double-stranded DNA viruses, in the family *Phycodnaviridae* that infect phytoplankton in the *Prasinophyceae*. There are three groups in this genus,[1][2] including *Micromonas pusilla virus SP1*, which infects the cosmopolitan photosynthetic flagellate *Micromonas pusilla*.[3]

There is a large group of genetically diverse but related viruses that show considerable evidence of lateral gene transfer.[4][5]

What's the next one down?



Huh! (part deux)

Siphoviridae

Article Talk

From Wikipedia, the free encyclopedia

Siphoviridae is a family of double-stranded DNA viruses in the order *Caudovirales*. Bacteria and archaea serve as natural hosts. There are 1,166 species in this family, assigned to 366 genera and 22 subfamilies.[2][3] The characteristic structural features of this family are a nonenveloped head and noncontractile tail.

Some people noticed

Caution Regarding the Specificities of Pan-Cancer Microbial Structure

Abraham Gihawi, Colin, S. Cooper, Daniel S. Brewer

Affiliations

Abraham Gihawi - A.Gihawi@uea.ac.uk - 1

Colin, S. Cooper - C.Cooper@uea.ac.uk - 1

Daniel, S. Brewer - D.Brewer@uea.ac.uk - 1,2

1 - Bob Champion Research & Education Building, Norwich Medical School, University of East Anglia, Norwich, UK NR4 7UQ

2 - Earlham Institute, Norwich Research Park, Colney Lane, Norwich, UK, NR4 7UG

The top features are weird

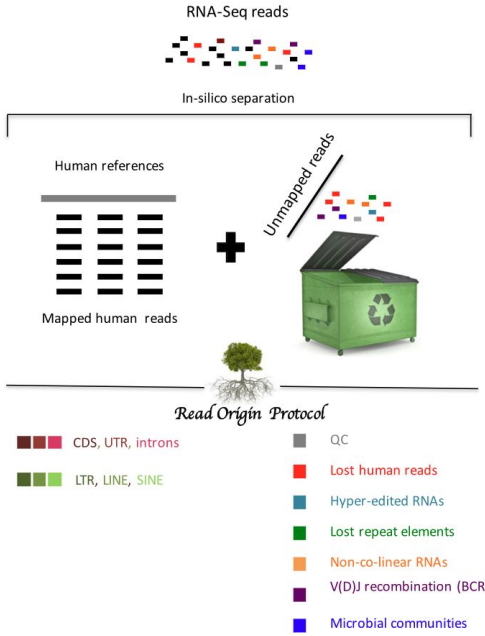
Genus	Top Feature in cancer type model	Details
Leucothrix	Bladder Cancer	Bacteria from marine macroalgae[5]
Thalassomonas	Uveal Melanoma	Bacteria causes disease in coral[4]

Velarivirus	Cervical Cancer	Grapevine is natural host[8]
Tritimovirus	Colon Cancer	Known to infect cereals[9]
Dinovernavirus	Renal Clear Cell Carcinoma	Contains insect viruses[10]
Bacillarnavirus	Lung Squamous Cell Carcinoma	Infects algae[11]

Velarivirus	Cervical Cancer	Grapevine is natural host[8]
Tritimovirus	Colon Cancer	Known to infect cereals[9]
Dinovernavirus	Renal Clear Cell Carcinoma	Contains insect viruses[10]
Bacillarnavirus	Lung Squamous Cell Carcinoma	Infects algae[11]

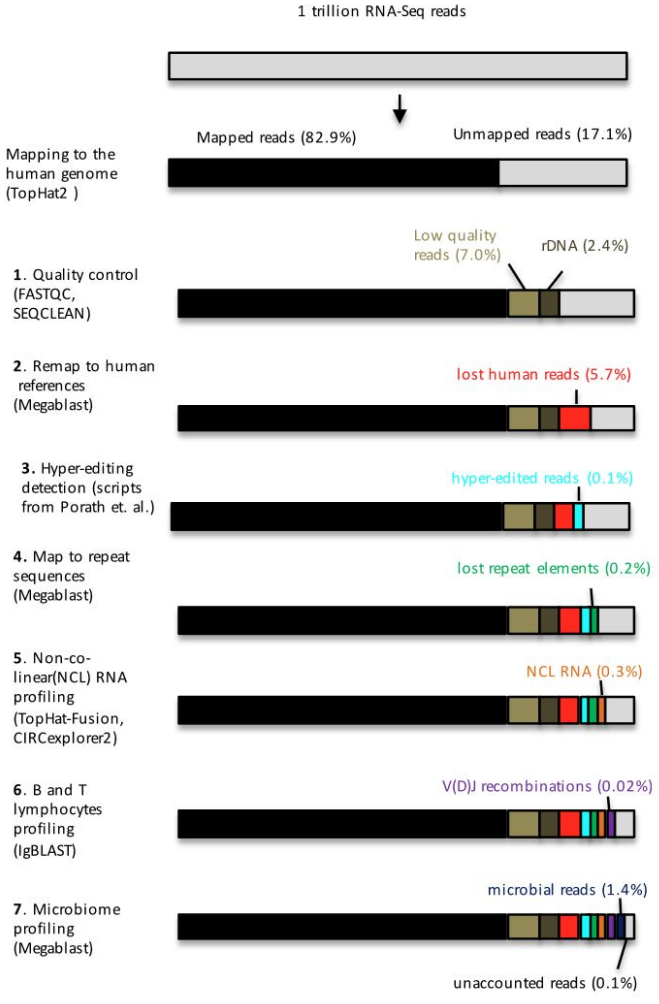
Rymovirus	Ovarian serous	Infects species of grass[12]
Ignicoccus	Prostate	Identified in marine hydrothermal vents[13]
Salinimicrobium	Testicular Cancer	Halophilic genus identified from marine environments[14]

Where do they come from?
Clue #1: remember ROP



ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues

Serghei Mangul[1,2*], Harry Taegyun Yang[1], Nicolas Strauli[3], Franziska Gruhl[4,5], Hagit T. Porath[6], Kevin Hsi Linus Chen[7], Timothy Daley[8], Stephanie Christenson[9], Agata Wesolowska-Andersen[10], Roberto Spreafic Cydney Rios[10], Celeste Eng[11], Andrew D. Smith[8], Ryan D. Hernandez[12,13,14], Roel A. Ophoff[15,16,17], Jose Rodriguez Santana[18], Erez Y. Levanon[6], Prescott G. Woodruff[9], Esteban Burchard[22], Max A. Seibold Sagiv Shifman[21†], Eleazar Eskin[1,16†] and Noah Zaitlen[9*†]



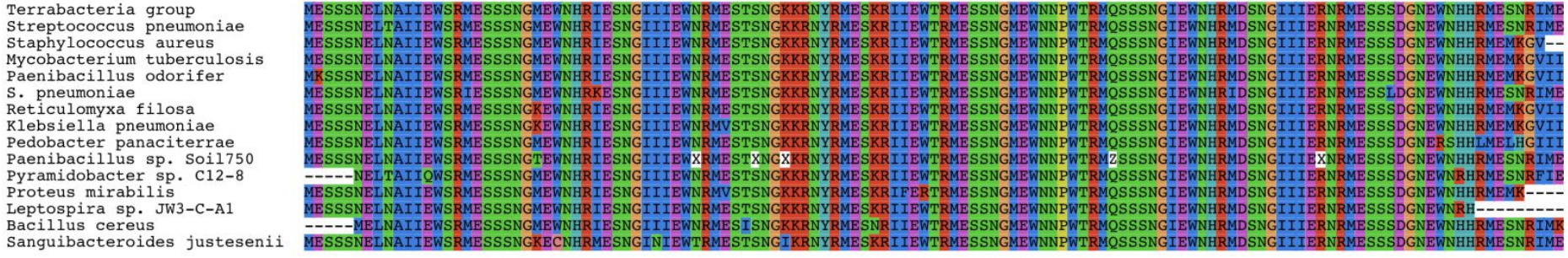
Clue #2: metagenomic reference missing human sequences

Sequencing reads that did not align to known human reference genomes (based on mapping information in the raw BAM files) were mapped against all known bacterial, archaeal, and viral microbial genomes using the ultrafast Kraken algorithm[23]. A total of 71,782 microbial genomes were downloaded using RepoPhlan (https://bitbucket.org/nsegata/repophlan) on 14 June 2016, of which 5,503 were viral and 66,279 were bacterial or archaeal. On the basis of prior literature, bacterial and archaeal genomes were filtered for quality scores of 0.8 or better[58], which left 54,471 of them for subsequent analysis, or a total of 59,974 microbial genomes.

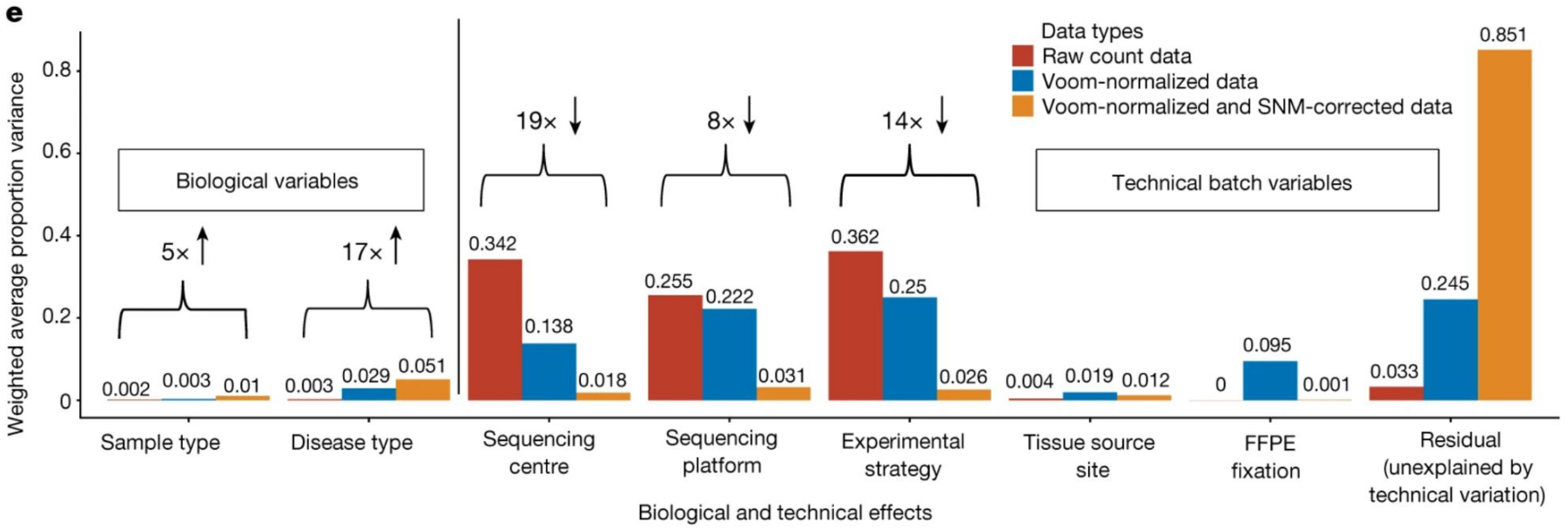
Clue #3: microbial genomes full of human sequences

Human contamination in bacterial genomes has created thousands of spurious proteins

Profile	Description	Profile length (bp)	Complete	Draft	Total
LINEs	Long interspersed nuclear elements, >15% of human genome	~6000	2	1066	1068
Alu family	Most abundant SINEs, about 10% of human genome	~300	3	746	749
Satellites	Satellite repeats ALR, BSR, HSATII	~170	0	910	910
LTRs	Long terminal repeats from endogenous retroviruses	~200–5000	0	228	228
DNA transposon	Tigger1 DNA transposon	2418	1	20	21



Clue #4: all that normalization



What's their classifier?

Algorithm 2: Stochastic Gradient_TreeBoost

1 $F_0(\mathbf{x}) = \arg\min_\gamma \sum_{i-1}^{N} \Psi(y_i, \gamma)$.

2 For $m = 1$ to M do:

3 $\quad \{\pi(i)\}_1^N = \text{rand\_perm } \{i\}_1^N$

4 $\quad \tilde{y}_{\pi(i)m} = -\left[\frac{\partial \Psi(y_{\pi(i)}, F(\mathbf{x}_{\pi(i)}))}{\partial F(\mathbf{x}_{\pi(i)})}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, \; i = 1, \tilde{N}$

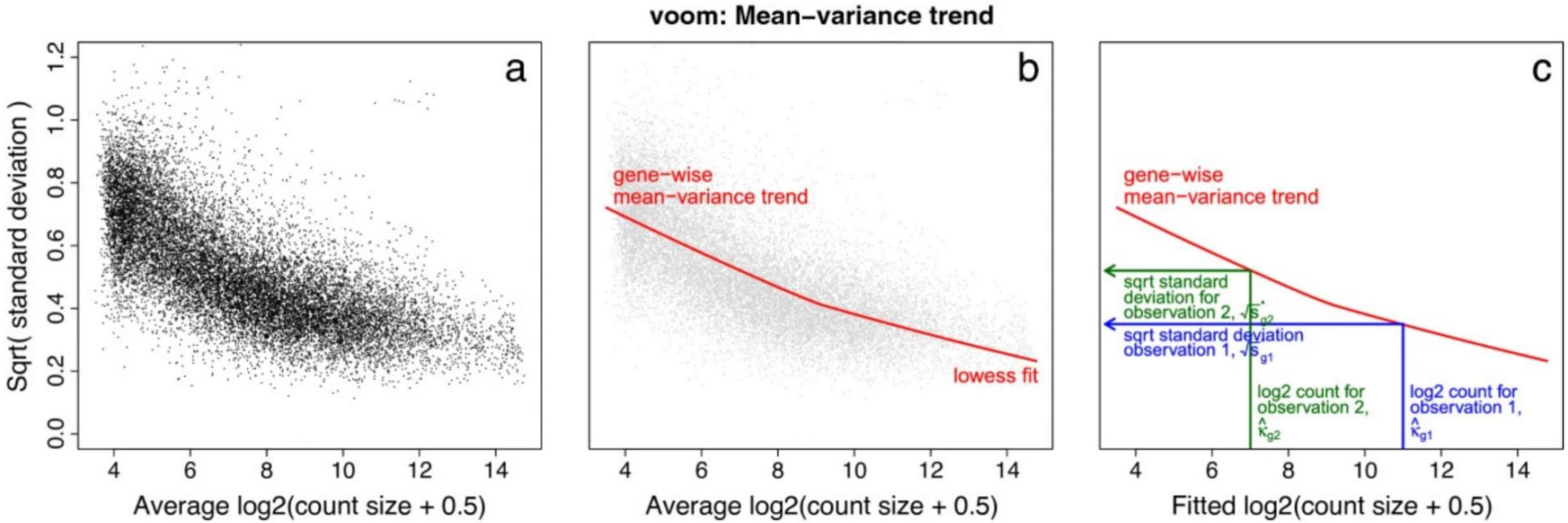
5 $\quad \{R_{lm}\}_1^L = L - \text{terminal node } tree(\{\tilde{y}_{\pi(i)m}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}})$

6 $\quad \gamma_{lm} = \arg\min_\gamma \sum_{\mathbf{x}_{\pi(i)} \in R_{lm}} \Psi(y_{\pi(i)}, F_{m-1}(\mathbf{x}_{\pi(i)}) + \gamma)$

7 $\quad F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} 1(\mathbf{x} \in R_{lm})$

8 endFor.

Pre-Normalization: Voom



Normalization: Limma

Limma fits a linear model to each gene.

Linear models include analysis of variance (ANOVA) models, linear regression, and any model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The covariates X can be:

- a continuous variable (pH, RIN score, age, weight, temperature, etc.)
- Dummy variables coding a categorical covariate (like cultivar, time, and group)

More Normalization: SNM

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 10 2010, pages 1308–1315
doi:10.1093/bioinformatics/btq118

Gene expression

Advance Access publication March 31, 2010

Supervised normalization of microarrays

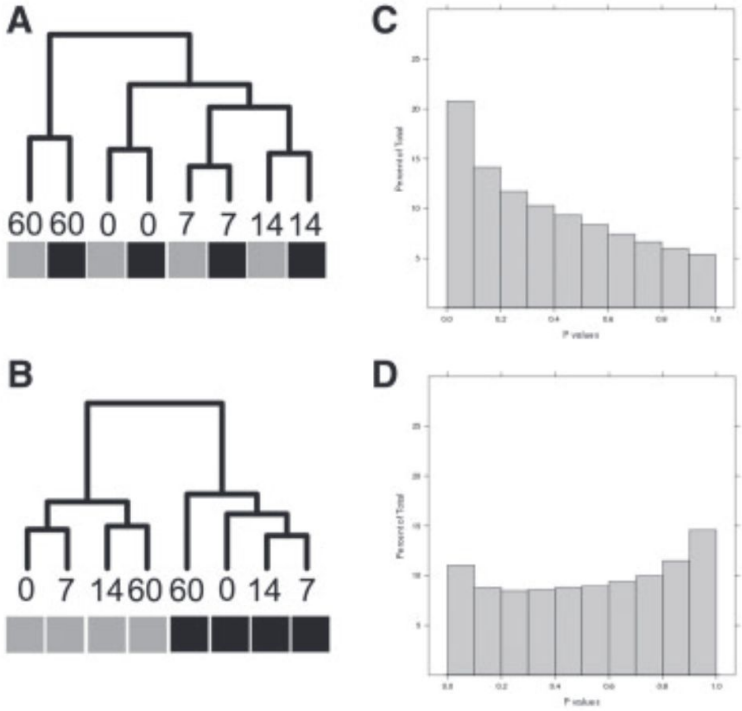
Brigham H. Mecham[1], Peter S. Nelson[1,2] and John D. Storey[3,*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, [2]Divisions of Human Biology and Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, WA 98109 and [3]Lewis-Sigler Institute for Integrative Genomics and Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

We can write model (1) for probe i data across all n arrays, $\mathbf{y}_i$, as

$$\mathbf{y}_i = \mathbf{b}_i\mathbf{X} + \mathbf{a}_i\mathbf{Z} + \sum_{t=1}^{r_f} f_t(\mathbf{b}_i\mathbf{X} + \mathbf{a}_i\mathbf{Z}) + \mathbf{e}_i, \quad (2)$$

Effect of SNM on p-values



Controlling for *how many* conditions?

```
 [1] "sample_typeSolid Tissue Normal"
 [2] "sample_typeAdditionI
 [3] "sample_typeAdditionM
 [4] "sample_typeBlood De
 [5] "sample_typeMetastat
 [6] "sample_typePrimary
 [7] "sample_typeRecurren
 [8] "data_submitting_cen
 [9] "data_submitting_cen
[10] "data_submitting_cen
[11] "data_submitting_cen
[12] "data_submitting_cen
[13] "data_submitting_cen
[14] "data_submitting_cen
[15] "platformHiSeq X Ten
[16] "platformIllumina GA
[17] "platformIllumina Hi
[18] "platformIllumina Mi
[19] "platformLS 454"
[20] "experimental_strate
[21] "tissue_source_site_
[22] "tissue_source_site_
```

```
[187] "tissue_source_site_labelUniversity of Sydney"
[188] "tissue_source_site_labelUniversity of Texas MD Anderson Cancer Center"
[189] "tissue_source_site_labelUniversity of Ulm"
[190] "tissue_source_site_labelUniversity of Utah"
[191] "tissue_source_site_labelUniversity of Washington"
[192] "tissue_source_site_labelValley Hospital"
[193] "tissue_source_site_labelVanderbilt"
[194] "tissue_source_site_labelVanderbilt University"
[195] "tissue_source_site_labelWake Forest University"
[196] "tissue_source_site_labelWalter Reed"
[197] "tissue_source_site_labelWashington University"
[198] "tissue_source_site_labelWashington University - Alabama"
[199] "tissue_source_site_labelWashington University - CALGB"
[200] "tissue_source_site_labelWashington University - CHUV"
[201] "tissue_source_site_labelWashington University - Cleveland Clinic"
[202] "tissue_source_site_labelWashington University - Emory"
[203] "tissue_source_site_labelWashington University - Mayo Clinic"
[204] "tissue_source_site_labelWashington University - NYU"
[205] "tissue_source_site_labelWashington University - Rush University"
[206] "tissue_source_site_labelWashington University - St. Louis"
[207] "tissue_source_site_labelWashington University St. Louis"
[208] "tissue_source_site_labelWills Eye Institute"
[209] "tissue_source_site_labelYale"
[210] "tissue_source_site_labelYale University"
[211] "portion_is_ffpeYES"
```

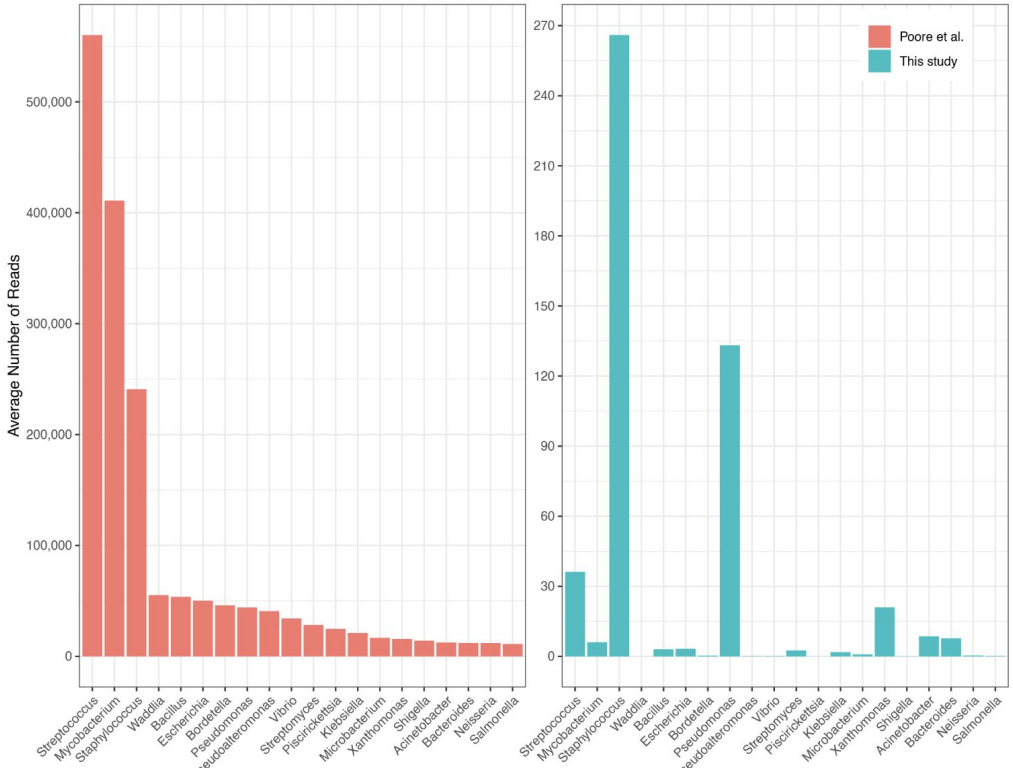
Team Critic Strikes Again

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.

How many bacterial reads, really?



Very specific values for abundances...

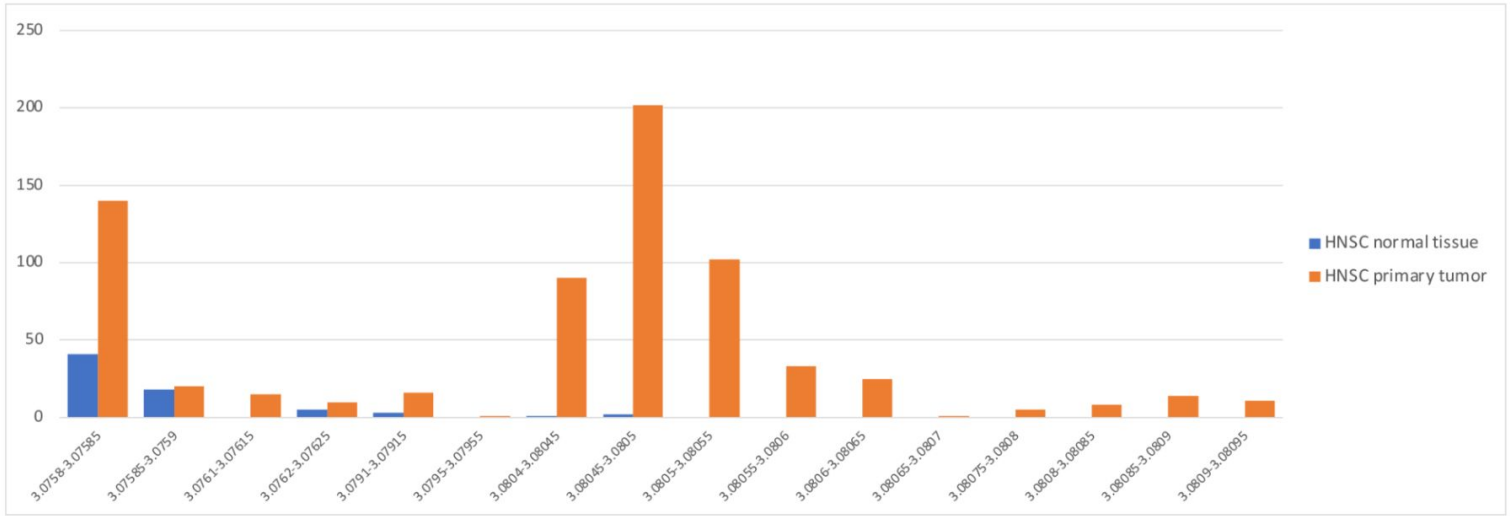
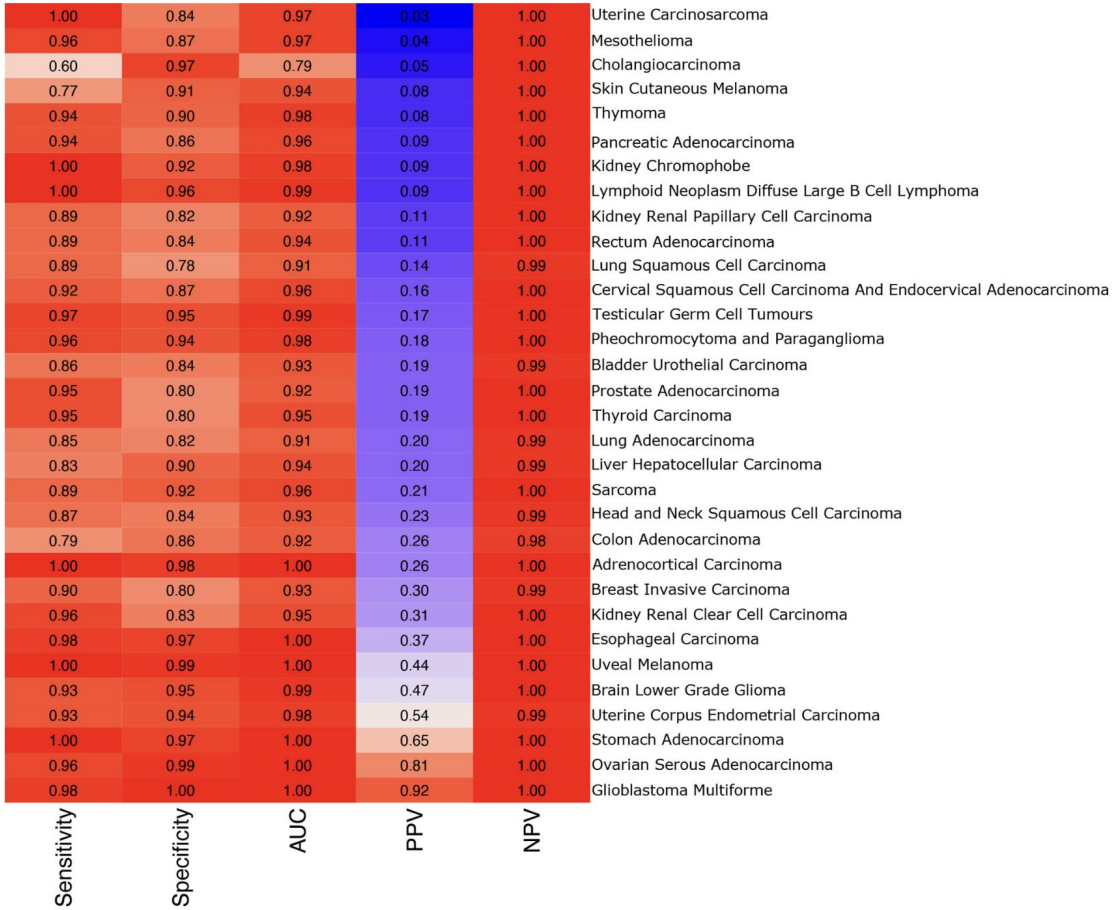


FIG 5 Distribution of normalized counts for *Mulikevirus* reads in head and neck squamous cell cancer (orange) and normal (blue) samples. All raw values were zero.

“Machine learning” works even when all original counts zero



What data leakage did they use for the plasma healthy vs. cancer validation?

(the labels!)

```
# List biological and normalization variables in model matrices
bio.var <- model.matrix(~disease_type_consol,
                        data=qcMetadata)

adj.var <- model.matrix(~host_age +
                            sex,
                          data=qcMetadata)

colnames(bio.var) <- gsub('([[:punct:]])|\\s+','',colnames(bio.var))
colnames(adj.var) <- gsub('([[:punct:]])|\\s+','',colnames(adj.var))
print(dim(adj.var))
print(dim(bio.var))
print(dim(t(vdge$E)))
print(dim(covDesignNorm))

snmDataObjOnly <- snm(raw.dat = vdge$E,
                      bio.var = bio.var,
                      adj.var = adj.var,
                      rm.adj=TRUE,
                      verbose = TRUE,
                      diagnose = TRUE)
snmData <<- t(snmDataObjOnly$norm.dat)
```


Fin