The Computational Biology and Informatics (CBI) Shared Resource Helen Diller Family Comprehensive Cancer Center https://cbi.ucsf.edu

Adam B. Olshen

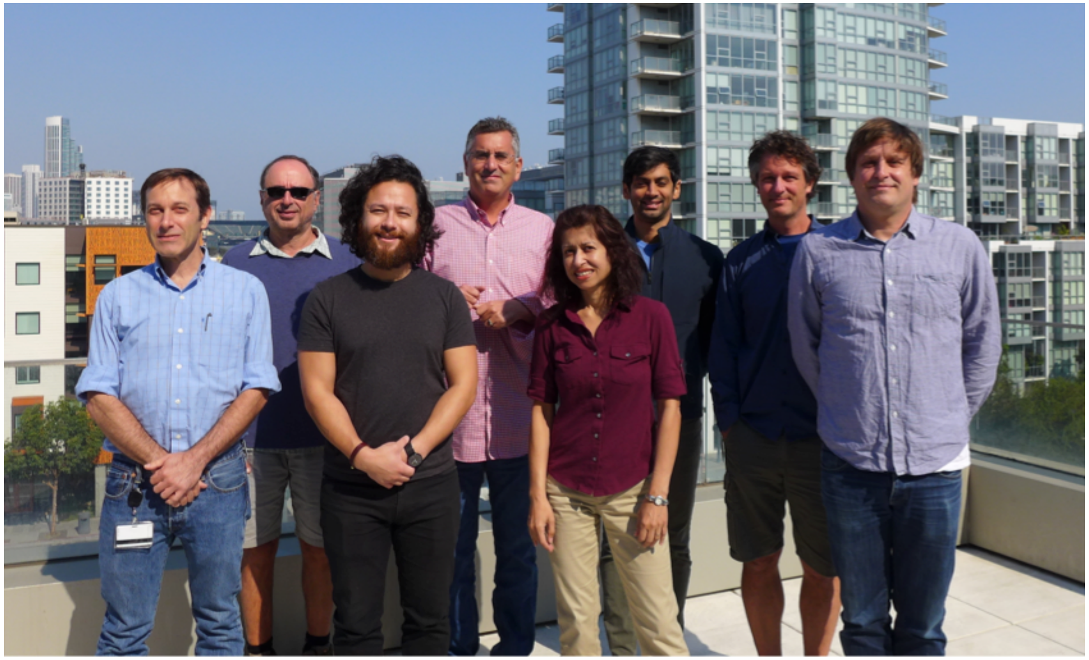
Shared Resource Leader

Professor, Department of Epidemiology and Biostatistics

11/08/21

Outline

- Come talk to us, maybe we can help your research
- Website: https://cbi.ucsf.edu
- Email: adam.olshen@ucsf.edu



UCSF
University of California
San Francisco

Our Mission

- The mission of the CBI is to collaborate on computational biology research and provide computational and data infrastructure support

- Our service is to the Helen Diller Family Comprehensive Cancer Center and to the overall cancer community at UCSF

- Our mission is both *science* and *computational infrastructure*

- I discuss infrastructure first

Old TIPCC HPC

- The TIPCC HPC system with 36 cores and 1680 nodes has been serving users since around 2011
- It has been available for all members of the HDFCCC community and has ~100 users
- The system had grown old and tired and we decided that it was a better strategy to build a new system rather than to update it

Not our cluster, but you get the idea

New C4 HPC!

Launched this January 1st

Modern software and configuration

Runs CentOS with a SLURM Scheduler

Software stacks to ease sharing of software

Allows Python3 and containerized computing

Interchangeable with campus-wide Wynton cluster

Somewhat like C4

San Francisco

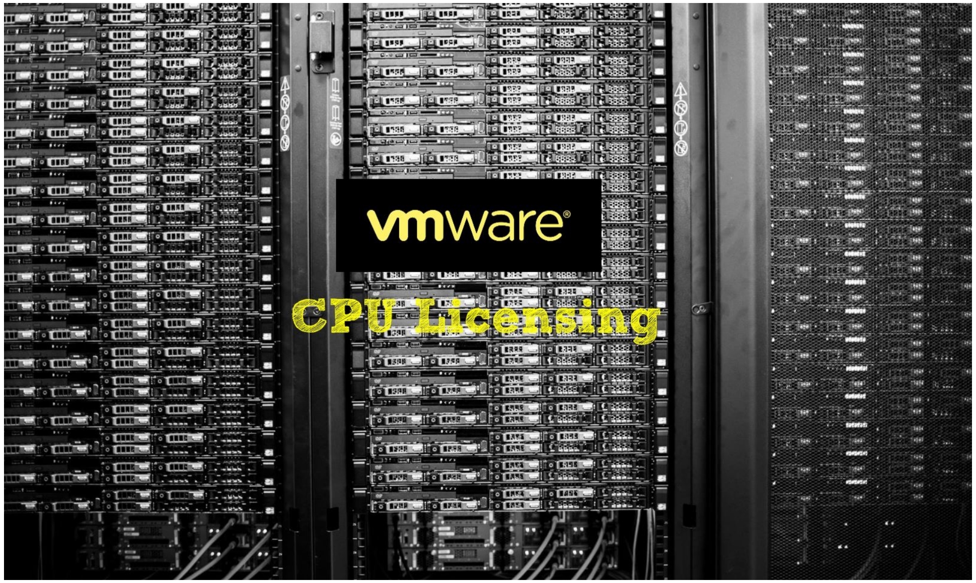
New C4 HPC!

- You can test out and use C4 common resources for free
- If you want additional resources your lab can buy its own computational nodes or storage
- We have extensive documentation at https://www.c4.ucsf.edu/
- Contact Harry Putnam (harry.putnam@ucsf.edu) for an account and help getting started

UCSF
University of California
San Francisco

Our VM Infrastructure

- We have recently installed a VM farm
- If you have a website, database or other specialized need, please talk to us
- We would recharge based on the needs of the application



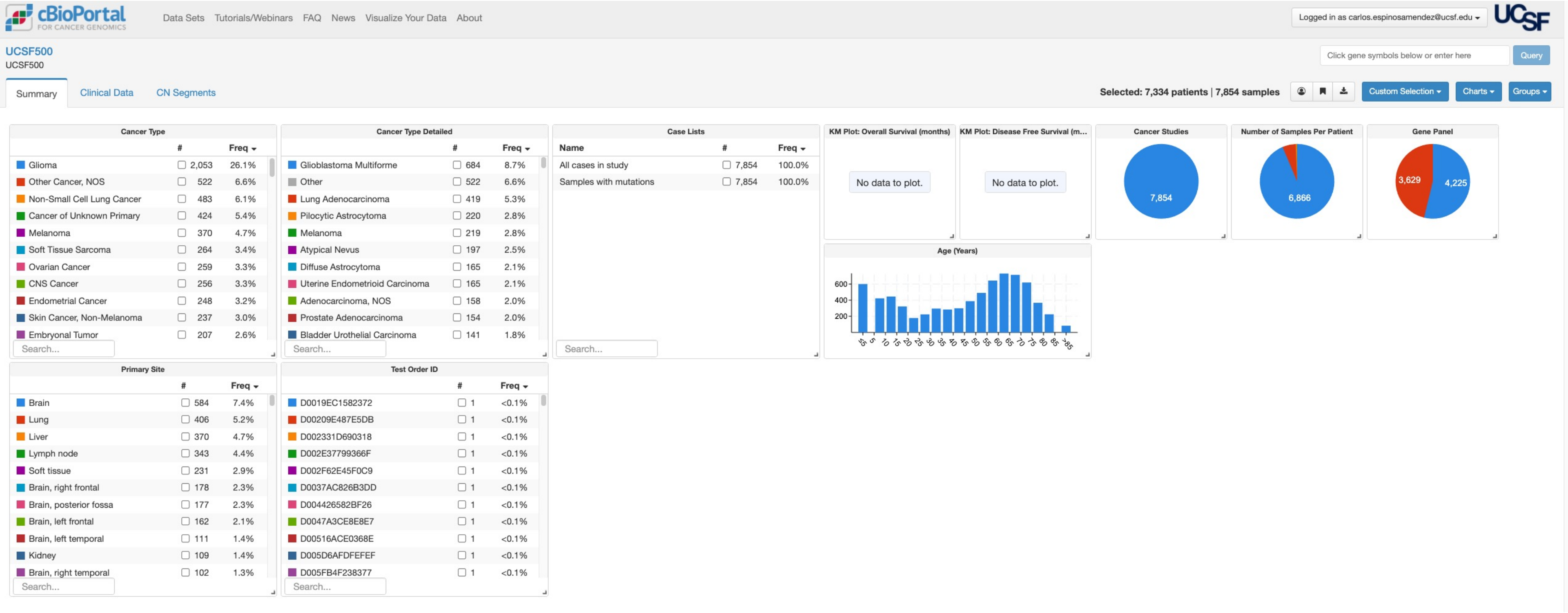
Bigger than our farm

University of California
San Francisco

Organizing UCSF500 Data on the cBioPortal

- The UCSF500 cancer gene panel is used to measure mutations of 500+ cancer genes
- It is used on all types of cancer and now has been run on >7k cancer patients at UCSF
- While it has been used to guide patient care, the data had not been organized for research
- Partnering with Alejandro Sweet-Cordero through the Molecular Oncology Initiative, we have organized the data on cBioPortal for UCSF

UCSF500 Data on the cBioPortal



UCSF500
UCSF500

Click gene symbols below or enter here

Selected: 7,334 patients | 7,854 samples

Cancer Type

	#	Freq
Glioma	☐ 2,053	26.1%
Other Cancer, NOS	☐ 522	6.6%
Non-Small Cell Lung Cancer	☐ 483	6.1%
Cancer of Unknown Primary	☐ 424	5.4%
Melanoma	☐ 370	4.7%
Soft Tissue Sarcoma	☐ 264	3.4%
Ovarian Cancer	☐ 259	3.3%
CNS Cancer	☐ 256	3.3%
Endometrial Cancer	☐ 248	3.2%
Skin Cancer, Non-Melanoma	☐ 237	3.0%
Embryonal Tumor	☐ 207	2.6%

Search...

Cancer Type Detailed

	#	Freq
Glioblastoma Multiforme	☐ 684	8.7%
Other	☐ 522	6.6%
Lung Adenocarcinoma	☐ 419	5.3%
Pilocytic Astrocytoma	☐ 220	2.8%
Melanoma	☐ 219	2.8%
Atypical Nevus	☐ 197	2.5%
Diffuse Astrocytoma	☐ 165	2.1%
Uterine Endometrioid Carcinoma	☐ 165	2.1%
Adenocarcinoma, NOS	☐ 158	2.0%
Prostate Adenocarcinoma	☐ 154	2.0%
Bladder Urothelial Carcinoma	☐ 141	1.8%

Search...

Case Lists

Name	#	Freq
All cases in study	☐ 7,854	100.0%
Samples with mutations	☐ 7,854	100.0%

Search...

KM Plot: Overall Survival (months)

No data to plot.

KM Plot: Disease Free Survival (m...

No data to plot.

Cancer Studies

Number of Samples Per Patient

Gene Panel

Age (Years)

Primary Site

	#	Freq
Brain	☐ 584	7.4%
Lung	☐ 406	5.2%
Liver	☐ 370	4.7%
Lymph node	☐ 343	4.4%
Soft tissue	☐ 231	2.9%
Brain, right frontal	☐ 178	2.3%
Brain, posterior fossa	☐ 177	2.3%
Brain, left frontal	☐ 162	2.1%
Brain, left temporal	☐ 111	1.4%
Kidney	☐ 109	1.4%
Brain, right temporal	☐ 102	1.3%

Search...

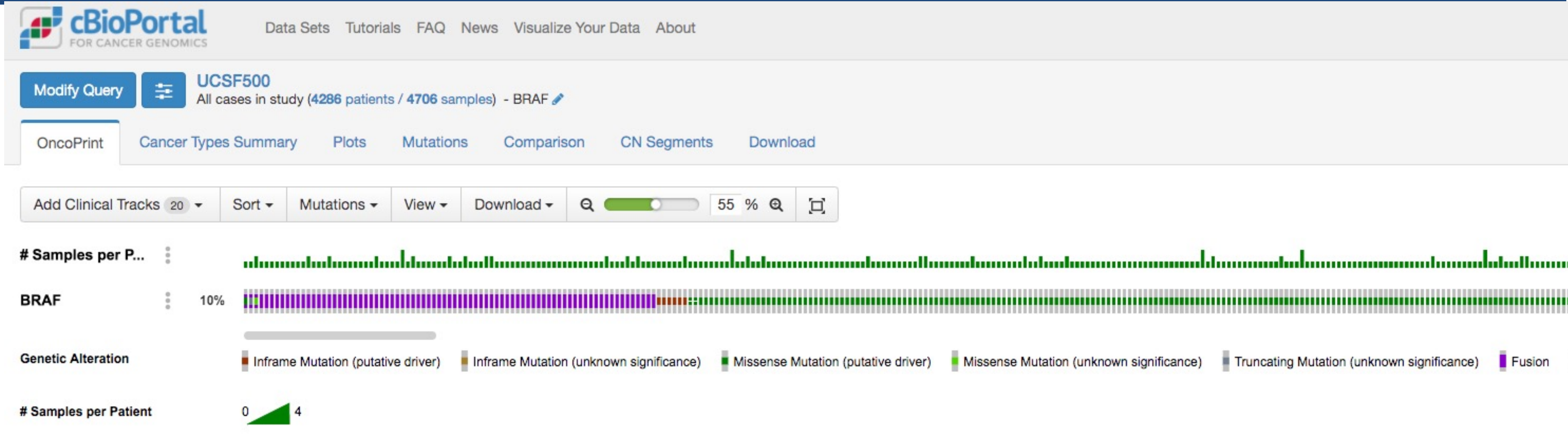
Test Order ID

	#	Freq
D0019EC1582372	☐ 1	<0.1%
D00209E487E5DB	☐ 1	<0.1%
D002331D690318	☐ 1	<0.1%
D002E37799366F	☐ 1	<0.1%
D002F62E45F0C9	☐ 1	<0.1%
D0037AC826B3DD	☐ 1	<0.1%
D004426582BF26	☐ 1	<0.1%
D0047A3CE8E8E7	☐ 1	<0.1%
D00516ACE0368E	☐ 1	<0.1%
D005D6AFDFEFEF	☐ 1	<0.1%
D005FB4F238377	☐ 1	<0.1%

Search...

University of California
San Francisco

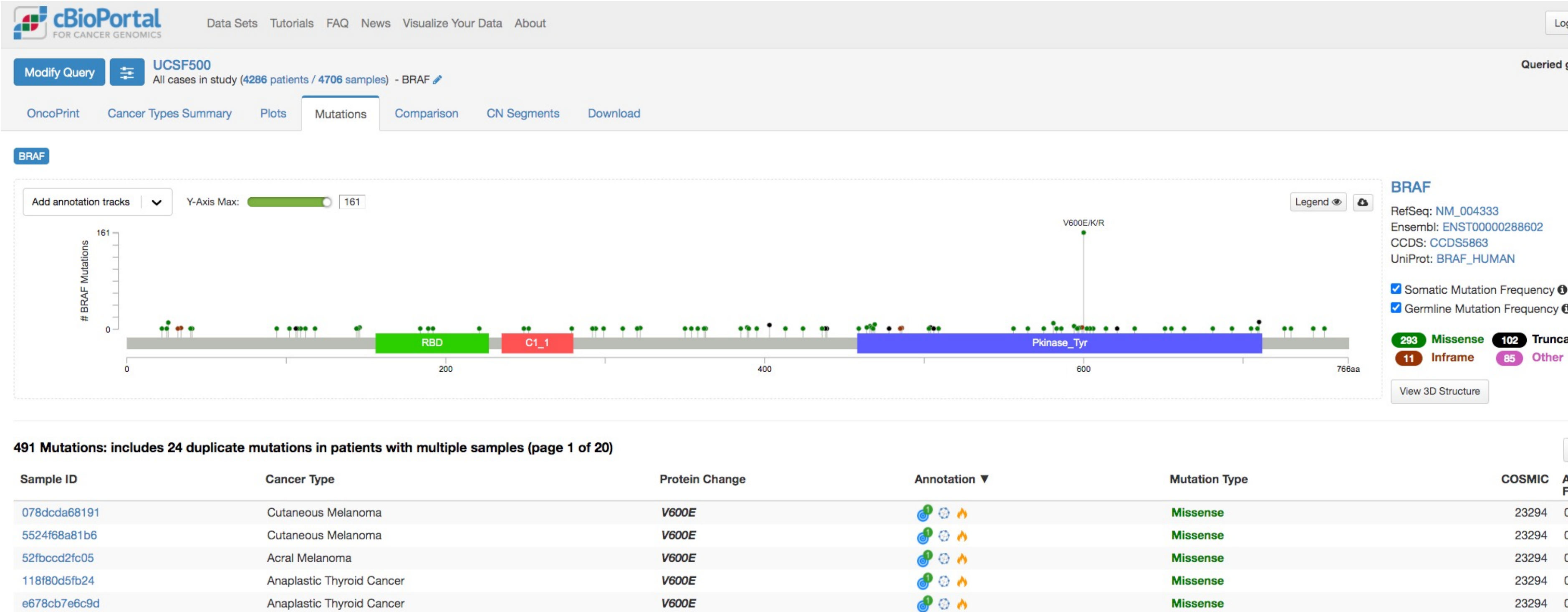
Views of BRAF



UCSF500
All cases in study (4286 patients / 4706 samples) - BRAF

Genetic Alteration: Inframe Mutation (putative driver), Inframe Mutation (unknown significance), Missense Mutation (putative driver), Missense Mutation (unknown significance), Truncating Mutation (unknown significance), Fusion

Samples per Patient



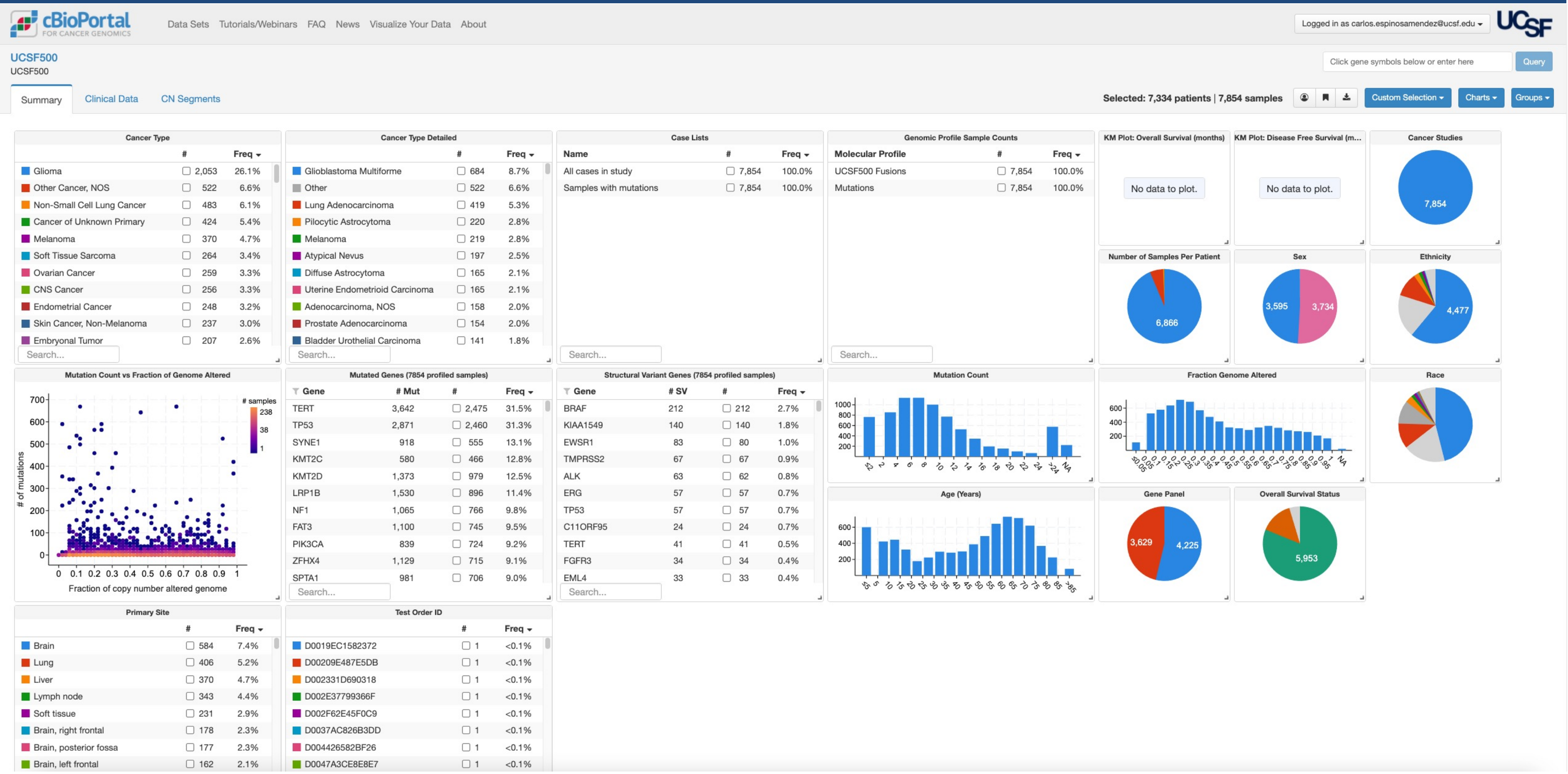
UCSF500
All cases in study (4286 patients / 4706 samples) - BRAF

BRAF
RefSeq: NM_004333
Ensembl: ENST00000288602
CCDS: CCDS5863
UniProt: BRAF_HUMAN

491 Mutations: includes 24 duplicate mutations in patients with multiple samples (page 1 of 20)

Sample ID	Cancer Type	Protein Change	Annotation ▼	Mutation Type	COSMIC
078dcda68191	Cutaneous Melanoma	*V600E*		Missense	23294
5524f68a81b6	Cutaneous Melanoma	*V600E*		Missense	23294
52fbccd2fc05	Acral Melanoma	*V600E*		Missense	23294
118f80d5fb24	Anaplastic Thyroid Cancer	*V600E*		Missense	23294
e678cb7e6c9d	Anaplastic Thyroid Cancer	*V600E*		Missense	23294

Now Including Clinical Information



UCSF500
UCSF500

Selected: 7,334 patients | 7,854 samples

Cancer Type

	#	Freq
■ Glioma	□ 2,053	26.1%
■ Other Cancer, NOS	□ 522	6.6%
■ Non-Small Cell Lung Cancer	□ 483	6.1%
■ Cancer of Unknown Primary	□ 424	5.4%
■ Melanoma	□ 370	4.7%
■ Soft Tissue Sarcoma	□ 264	3.4%
■ Ovarian Cancer	□ 259	3.3%
■ CNS Cancer	□ 256	3.3%
■ Endometrial Cancer	□ 248	3.2%
■ Skin Cancer, Non-Melanoma	□ 237	3.0%
■ Embryonal Tumor	□ 207	2.6%

Search...

Cancer Type Detailed

	#	Freq
■ Glioblastoma Multiforme	□ 684	8.7%
■ Other	□ 522	6.6%
■ Lung Adenocarcinoma	□ 419	5.3%
■ Pilocytic Astrocytoma	□ 220	2.8%
■ Melanoma	□ 219	2.8%
■ Atypical Nevus	□ 197	2.5%
■ Diffuse Astrocytoma	□ 165	2.1%
■ Uterine Endometrioid Carcinoma	□ 165	2.1%
■ Adenocarcinoma, NOS	□ 158	2.0%
■ Prostate Adenocarcinoma	□ 154	2.0%
■ Bladder Urothelial Carcinoma	□ 141	1.8%

Search...

Case Lists

Name	#	Freq
All cases in study	□ 7,854	100.0%
Samples with mutations	□ 7,854	100.0%

Search...

Genomic Profile Sample Counts

Molecular Profile	#	Freq
UCSF500 Fusions	□ 7,854	100.0%
Mutations	□ 7,854	100.0%

Search...

Mutated Genes (7854 profiled samples)

Gene	# Mut	#	Freq
TERT	3,642	□ 2,475	31.5%
TP53	2,871	□ 2,460	31.3%
SYNE1	918	□ 555	13.1%
KMT2C	580	□ 466	12.8%
KMT2D	1,373	□ 979	12.5%
LRP1B	1,530	□ 896	11.4%
NF1	1,065	□ 766	9.8%
FAT3	1,100	□ 745	9.5%
PIK3CA	839	□ 724	9.2%
ZFHX4	1,129	□ 715	9.1%
SPTA1	981	□ 706	9.0%

Search...

Structural Variant Genes (7854 profiled samples)

Gene	# SV	#	Freq
BRAF	212	□ 212	2.7%
KIAA1549	140	□ 140	1.8%
EWSR1	83	□ 80	1.0%
TMPRSS2	67	□ 67	0.9%
ALK	63	□ 62	0.8%
ERG	57	□ 57	0.7%
TP53	57	□ 57	0.7%
C11ORF95	24	□ 24	0.7%
TERT	41	□ 41	0.5%
FGFR3	34	□ 34	0.4%
EML4	33	□ 33	0.4%

Search...

Primary Site

	#	Freq
■ Brain	□ 584	7.4%
■ Lung	□ 406	5.2%
■ Liver	□ 370	4.7%
■ Lymph node	□ 343	4.4%
■ Soft tissue	□ 231	2.9%
■ Brain, right frontal	□ 178	2.3%
■ Brain, posterior fossa	□ 177	2.3%
■ Brain, left frontal	□ 162	2.1%

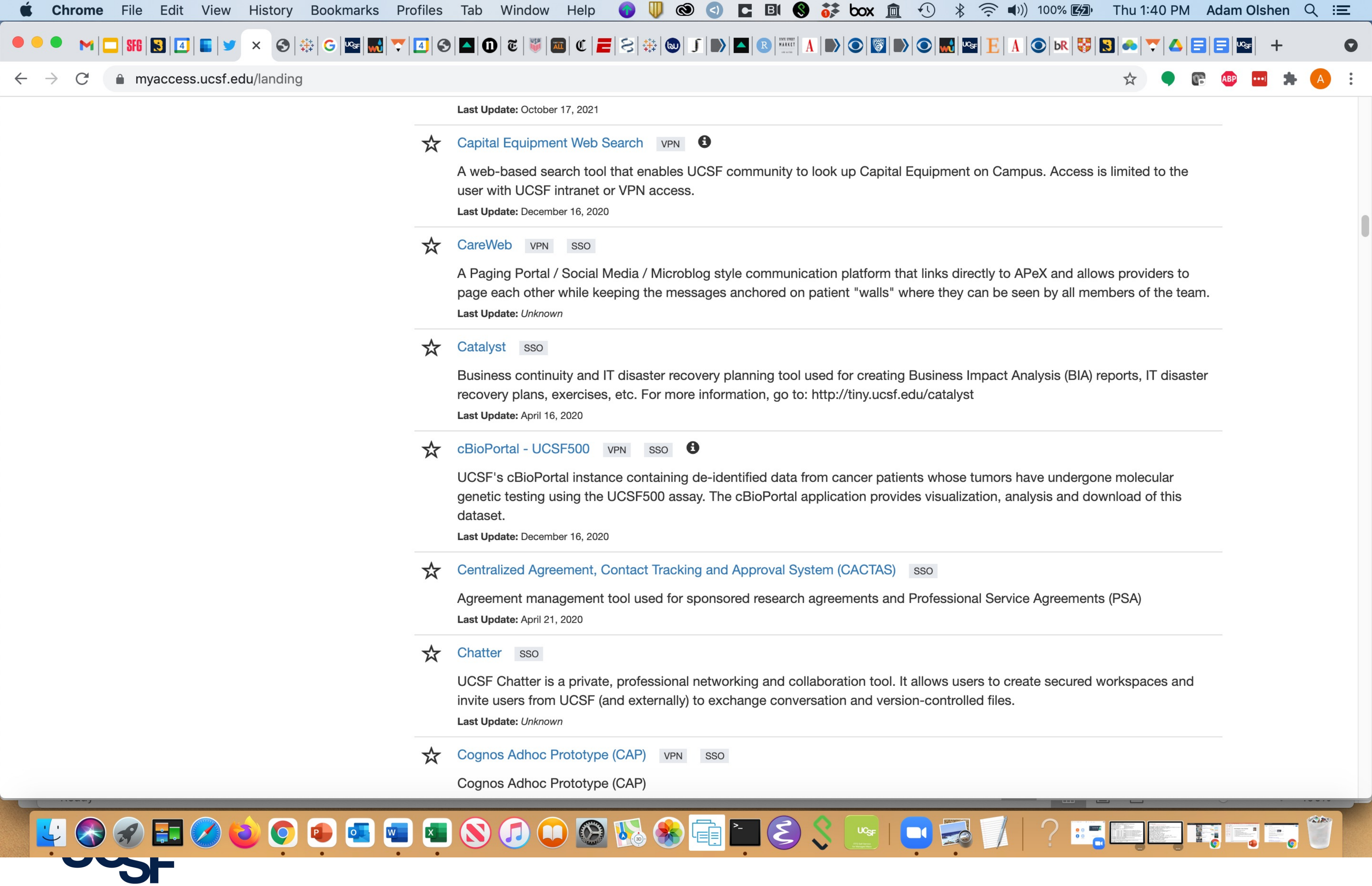
Test Order ID

	#	Freq
■ D0019EC1582372	□ 1	<0.1%
■ D00209E487E5DB	□ 1	<0.1%
■ D002331D690318	□ 1	<0.1%
■ D002E37799366F	□ 1	<0.1%
■ D002F62E45F0C9	□ 1	<0.1%
■ D0037AC826B3DD	□ 1	<0.1%
■ D004426582BF26	□ 1	<0.1%
■ D0047A3CE8E8E7	□ 1	<0.1%

Newest version includes Foundation Medicine data on 2k

UCSF
University of California
San Francisco

Accessing UCSF500 Data on the cBioPortal


Chrome File Edit View History Bookmarks Profiles Tab Window Help
100%
Thu 1:40 PM
Adam Olshen
myaccess.ucsf.edu/landing

Last Update: October 17, 2021

☆ Capital Equipment Web Search VPN

A web-based search tool that enables UCSF community to look up Capital Equipment on Campus. Access is limited to the user with UCSF intranet or VPN access.

Last Update: December 16, 2020

☆ CareWeb VPN SSO

A Paging Portal / Social Media / Microblog style communication platform that links directly to APeX and allows providers to page each other while keeping the messages anchored on patient "walls" where they can be seen by all members of the team.

Last Update: *Unknown*

☆ Catalyst SSO

Business continuity and IT disaster recovery planning tool used for creating Business Impact Analysis (BIA) reports, IT disaster recovery plans, exercises, etc. For more information, go to: http://tiny.ucsf.edu/catalyst

Last Update: April 16, 2020

☆ cBioPortal - UCSF500 VPN SSO

UCSF's cBioPortal instance containing de-identified data from cancer patients whose tumors have undergone molecular genetic testing using the UCSF500 assay. The cBioPortal application provides visualization, analysis and download of this dataset.

Last Update: December 16, 2020

☆ Centralized Agreement, Contact Tracking and Approval System (CACTAS) SSO

Agreement management tool used for sponsored research agreements and Professional Service Agreements (PSA)

Last Update: April 21, 2020

☆ Chatter SSO

UCSF Chatter is a private, professional networking and collaboration tool. It allows users to create secured workspaces and invite users from UCSF (and externally) to exchange conversation and version-controlled files.

Last Update: *Unknown*

☆ Cognos Adhoc Prototype (CAP) VPN SSO

Cognos Adhoc Prototype (CAP)

University of California
San Francisco

cBioPortal Service

- We have plans to set up local instances of the cBioPortal as a recharge service
- Contact us if interested

Henrik Bengtsson: Areas of Expertise

- Statistics and methods development
- Reproducible science
- Small and large-scale analysis
- High-performance computational methods
- Compute cluster design, support & usage (TIPCC, C4, and UCSF Wynton)
- Scientific software development, maintenance, and support
- Most things in R
- Mottos: Open access, sharing & helping, correctness, reproducibility

Henrik Bengtsson: Deeply involved with R

- Member of **The R Foundation**

  - Goals: Support development of R, exploration of new methodology, teaching and training of statistical computing, and the organization of meetings and conferences with a statistical computing orientation.

  - The R language and software. ~2 million users (2016), Worldwide UseR! conference (~1,500 ppl)

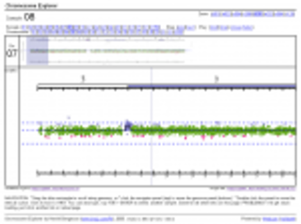
- Chair of **The R Consortium Infrastructure Steering Committee (ISC)**

  - Goals: Advance worldwide promotion of and support for R. Create and organize infrastructure projects, technical and infrastructure collaboration initiatives, support specific initiatives

  - Industry sponsored, 150,000 USD/year budget to support community R grants, R-Hub, R/Medicine, R/Pharma, R FDA Submissions, R Certification, …, 94 R User Groups in 38 countries (~70,000 ppl),
  212 R Ladies chapters in 60 countries (~80,000 ppl)

- **Contributor** to the core R code & community engagement

Henrik Bengtsson: Scientific Software



- 25 years of experience from world-wide collaborations
- Developed & maintains > **30 scientific R packages (CRAN & Bioconductor)**
- Some examples:
  - **Aroma Project - Small to Large-scale Affymetrix Analysis** in R, e.g. expression, genotyping, and copy number (peak 2005-2015)
  - **PSCBS - Parent-Specific Copy Numbers,** and **QDNAseq - CNs with FFPE DNAseq**
  - **matrixStats** - **Efficient Matrix Calculations**
    (top 0.8% most downloaded; 500,000+ downloads/month)
  - **future** - **A Unifying Parallelization Framework in R for Everyone**
    (top 1.0% most downloaded; 250,000+ downloads/month)

Henrik Bengtsson: Futureverse.org

- **A Unifying Parallelization Framework in R for Everyone**
- **Worry free** - lets researcher go from **exploratory** method's development to a **scalable pipeline**, e.g. prototyping new HiC method on local laptop, do **very minor code updates** to make use "futures", and then, with a single-change of settings, run on local computer, on a computer cluster, or in the cloud.
- **Write once - parallelize anywhere!**, e.g.

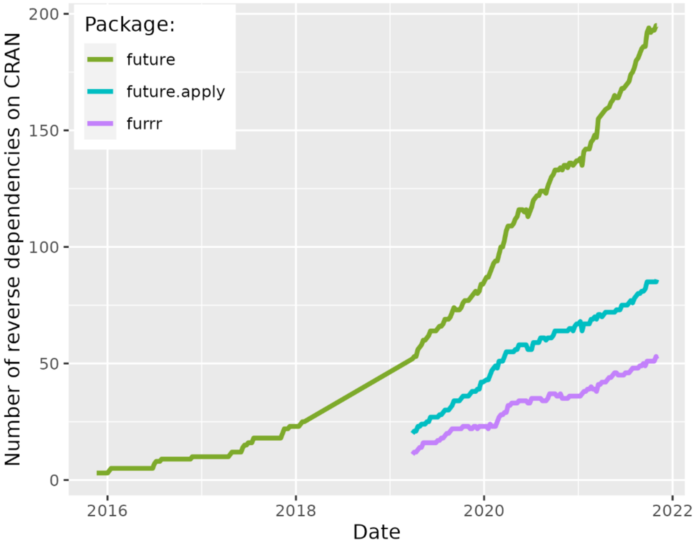
```
y <- lapply(X, slow_fcn)          ## original
y <- future_lapply(X, slow_fcn)   ## future version
```

- **Rapid update**

  - **Top-1% most downloaded R package**. 250,000+ downloads/months

  - 200+ packages depend on it directly, e.g. Shiny, Seurat, EpiNow2, and ml3r

  - A **Chan Zuckerberg Initiative Essential Open-Source Software** (CZI EOSS) - two-year grant



Scientific Collaborations

1. Scientific consultation for problems relating to computational biology
2. Data analysis for such projects
3. Grant development, including design of experiments, initial data analyses and writing
4. Education, in both formal and informal settings
5. Software development (when necessary) and training

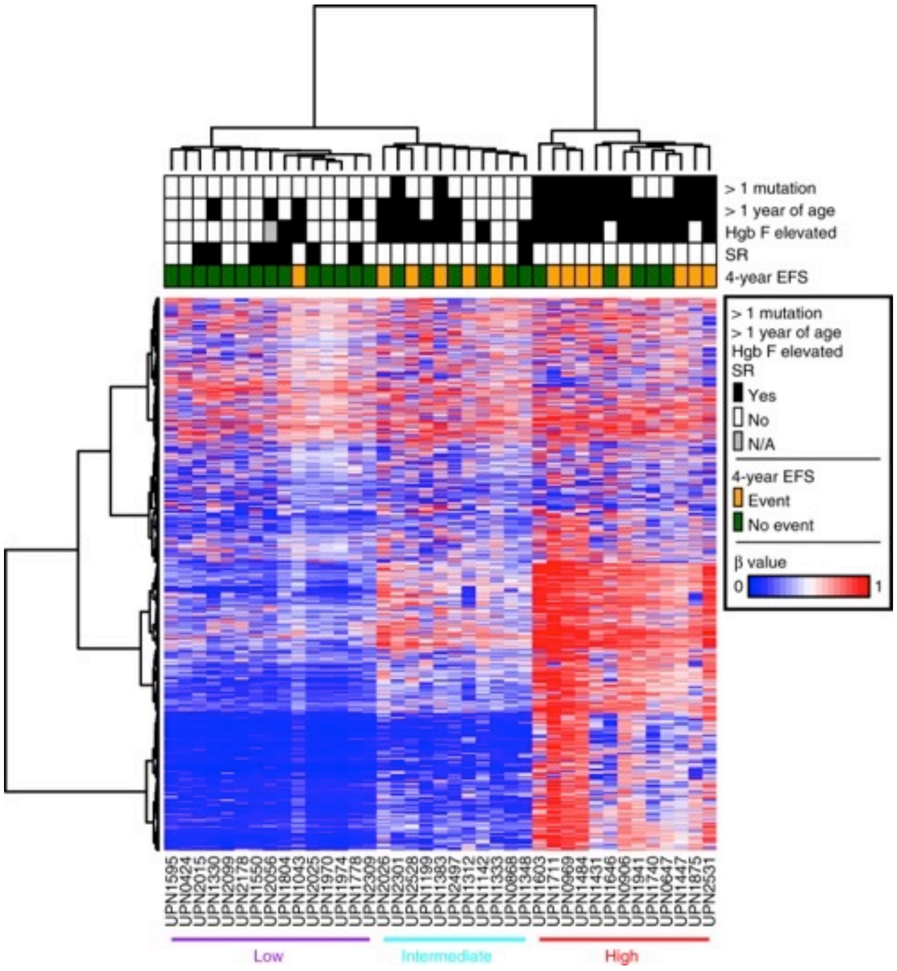
JMML Methylation Project

- Juvenile myelomonocytic leukemia (JMML) is a myeloproliferative disorder of childhood caused by mutations in the Ras pathway that occur in hematopoietic stem cells
- Outcomes vary: from spontaneous resolution with little or no treatment to relapse after stem cell transplantation
- We undertook a study of 39 training patients and 40 validation patients utilizing the 450k methylation array

(Stieglitz et al., *Nature Medicine*, 2017)

JMML Dendrogram shows clear patterns

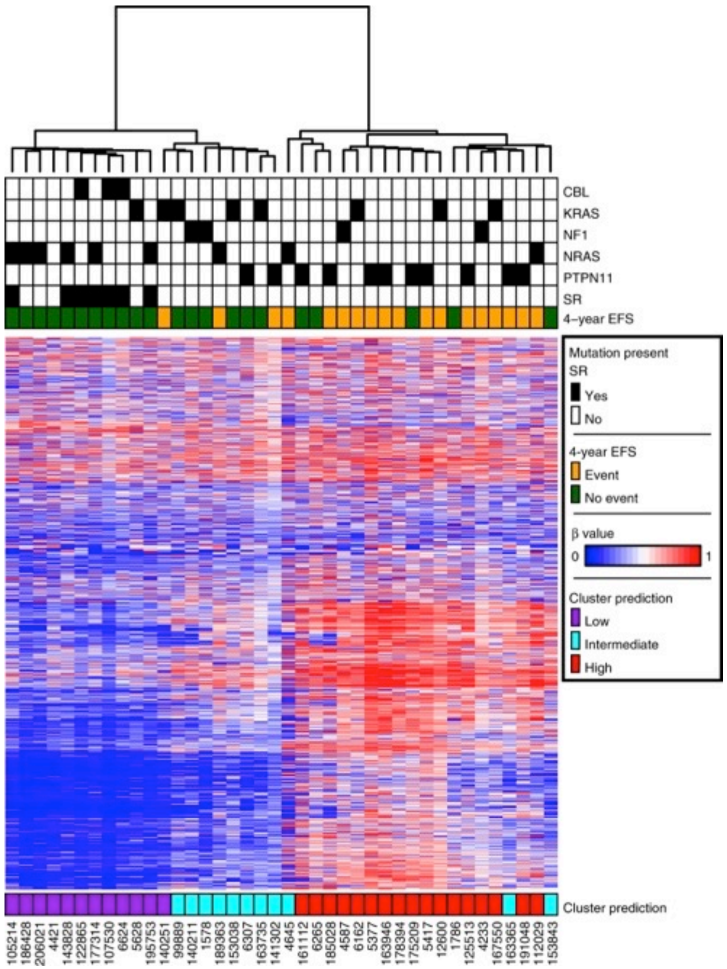
- 39 samples, 1500 markers
- Lower methylation more blue, higher methylation more red
- Samples naturally split into low, intermediate and high methylation groups
- Sample groups appear related to survival and other factors
- Markers split into four or so groups



UCSF
University of California
San Francisco

JMML Dendrogram shows clear patterns

- 40 samples, same 1500 markers
- Samples again cluster into low, intermediate and high clusters
- Prediction based on nearest training set centroid
- Relationship to survival repeated

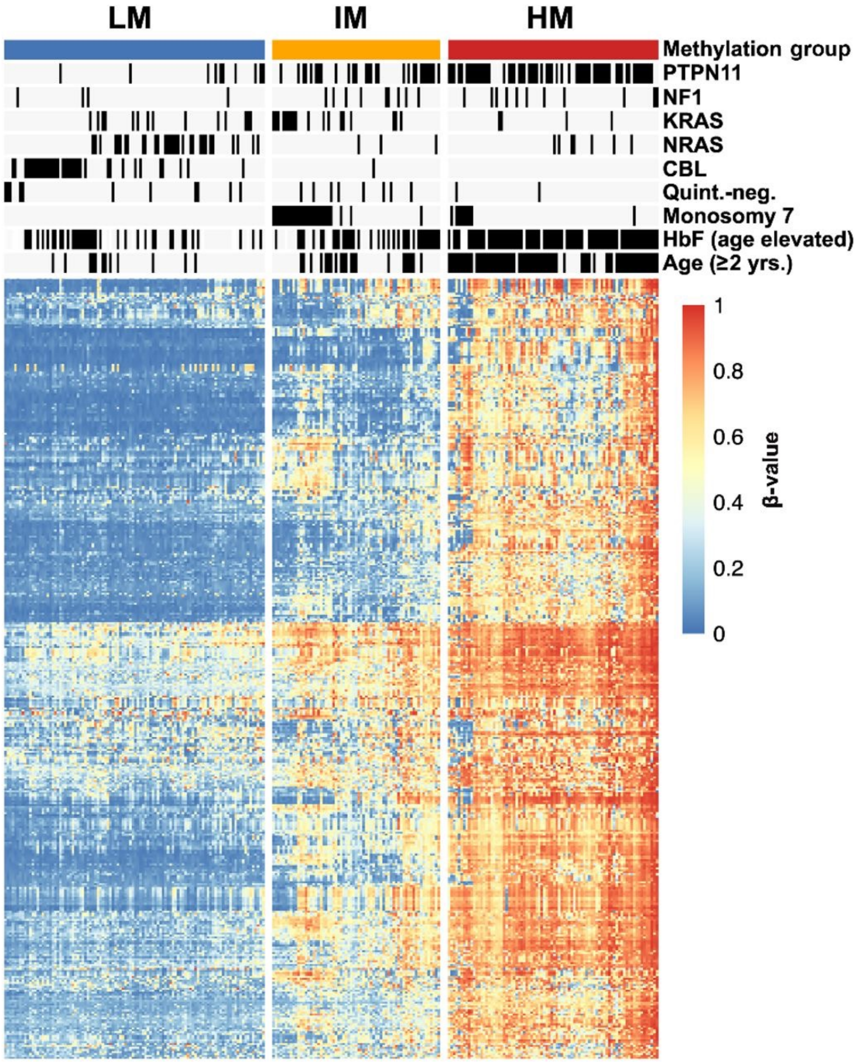


UCSF
University of California
San Francisco

Similar Methylation Groups Found Across Sites

- International Consortium including groups from Japan and Germany found the same three methylation groups (Schönung et al., *Clinical Cancer Research*, 2021)

- Being made into a clinical test to risk-stratify patients



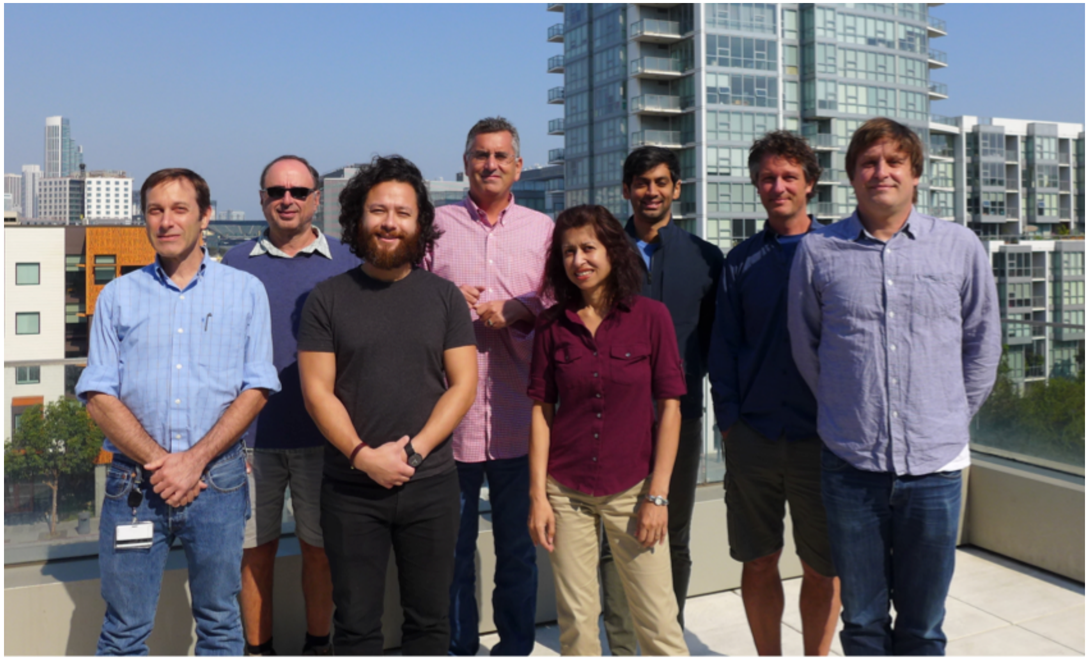
University of California San Francisco

How we work

- Pre-award - Supported through CCSG
- Long-term collaborations- Usually supported by post-award
- Short-term projects-Supported through hourly support
- Talk is always free

Conclusion

- Come talk to us, maybe we can help your research
- Website: https://cbi.ucsf.edu
- Email: adam.olshen@ucsf.edu



University of California
San Francisco