Breakout Session 5:

Cloud-Based Machine Learning and Biomarker Visual Analytics for Salivary Proteomics

Dr. Marcelo Freire
Associate Professor, J. Craig Venter Institute

Human Salivary Proteome *wiki*

“Cloud-Based Machine Learning and Biomarker Visual Analytics for Salivary Proteomics”

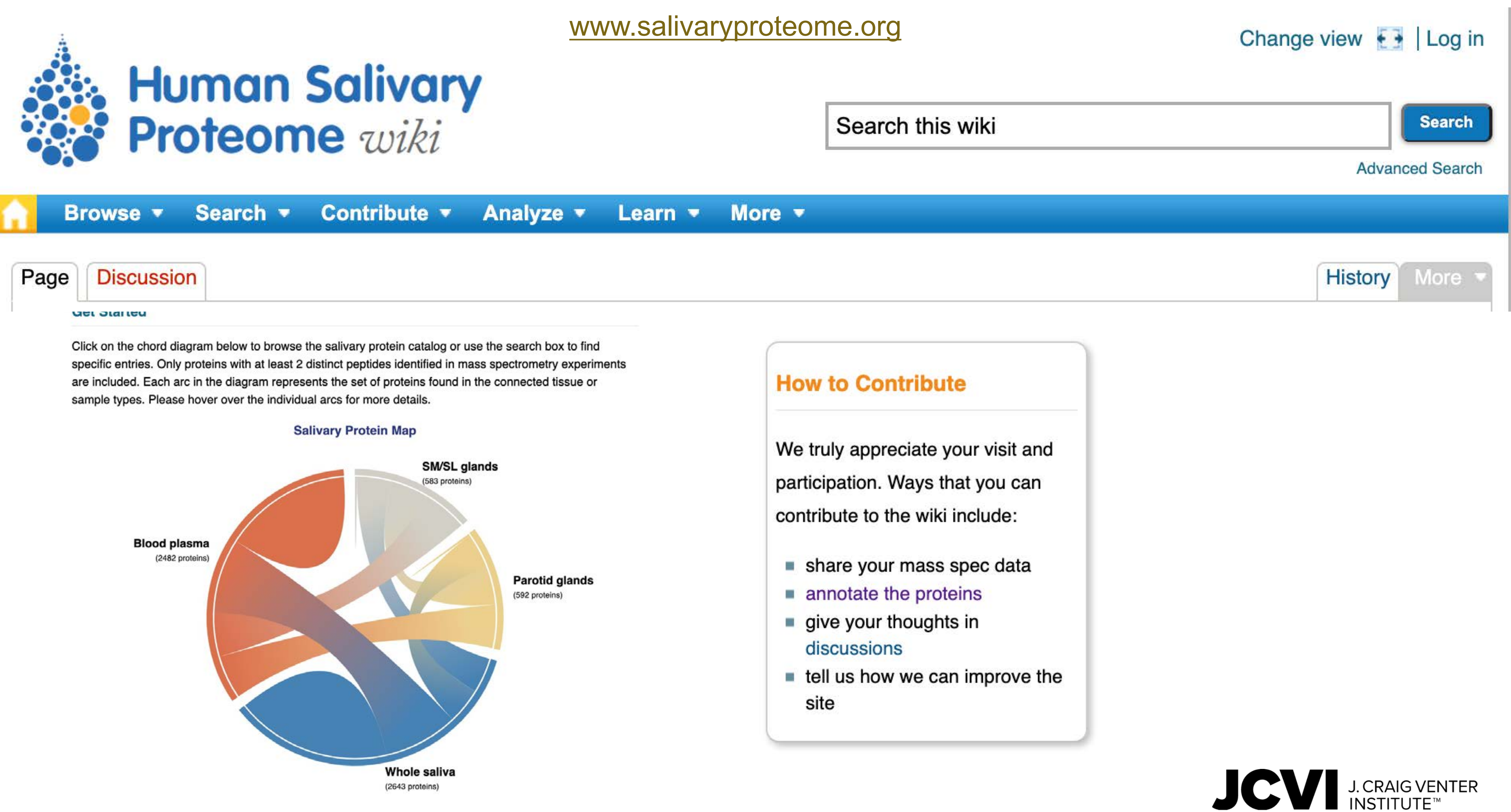
Marcelo Freire, DDS, PhD, DMedSc
Associate Professor, JCVI

Supplement grant # 3R01DE016937-18S1
(Parent grant: R01DE016937)

Date: 01-18-23

JCVI J. CRAIG VENTER INSTITUTE™

www.salivaryproteome.org



Get Started

Click on the chord diagram below to browse the salivary protein catalog or use the search box to find specific entries. Only proteins with at least 2 distinct peptides identified in mass spectrometry experiments are included. Each arc in the diagram represents the set of proteins found in the connected tissue or sample types. Please hover over the individual arcs for more details.

Salivary Protein Map

How to Contribute

We truly appreciate your visit and participation. Ways that you can contribute to the wiki include:

- share your mass spec data
- annotate the proteins
- give your thoughts in discussions
- tell us how we can improve the site

JCVI J. CRAIG VENTER INSTITUTE™

HSP WIKI Grant Aims:

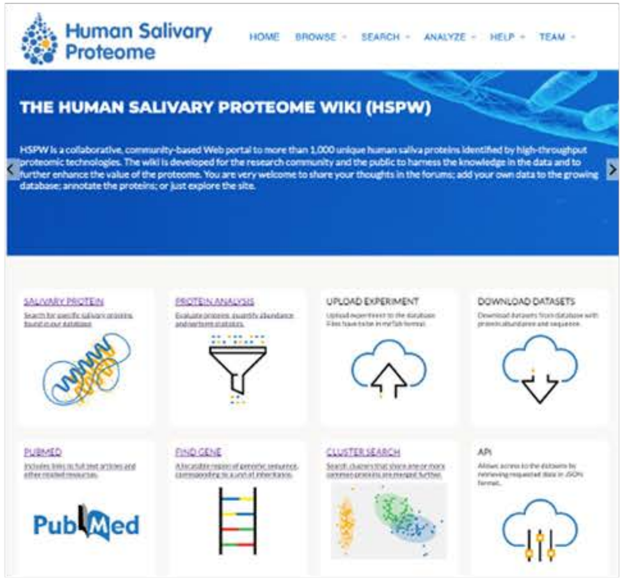
Aim 1. Transfer the HSP Wiki to a cloud-based server, maintain and develop the database computing services to provide the research community with a reliable, curated, and constantly evolving (up-to-date) data platform for evaluating salivary proteomics datasets.

Aim 2. Implement a novel computational infrastructure for salivary proteomic datasets that includes visual analytics, faster processing, and efficient measures of success.

Aim 3. Expand the HSP Wiki database to include bacterial and phage proteins.

The long-term goal of this project is to establish the resources and develop new tools to facilitate salivary research for both scientific discovery and diagnostic applications.

Human Salivary Proteome Wiki 2.0



THE HUMAN SALIVARY PROTEOME WIKI (HSPW)

HSPW is a collaborative, community-based Web portal to more than 1,000 unique human saliva proteins identified by high-throughput proteomic technologies. The wiki is developed for the research community and the public to harness the knowledge in the data and to further enhance the value of the proteome. You are very welcome to share your thoughts in the forums; add your own data to the growing database; annotate the proteins; or just explore the site.

SALIVARY PROTEIN
Search for specific salivary proteins found in our database.

PROTEIN ANALYSIS
Evaluate proteins, quantify abundance and perform statistics.

UPLOAD EXPERIMENT
Upload experiment to the database. Files have to be in mzTab format.

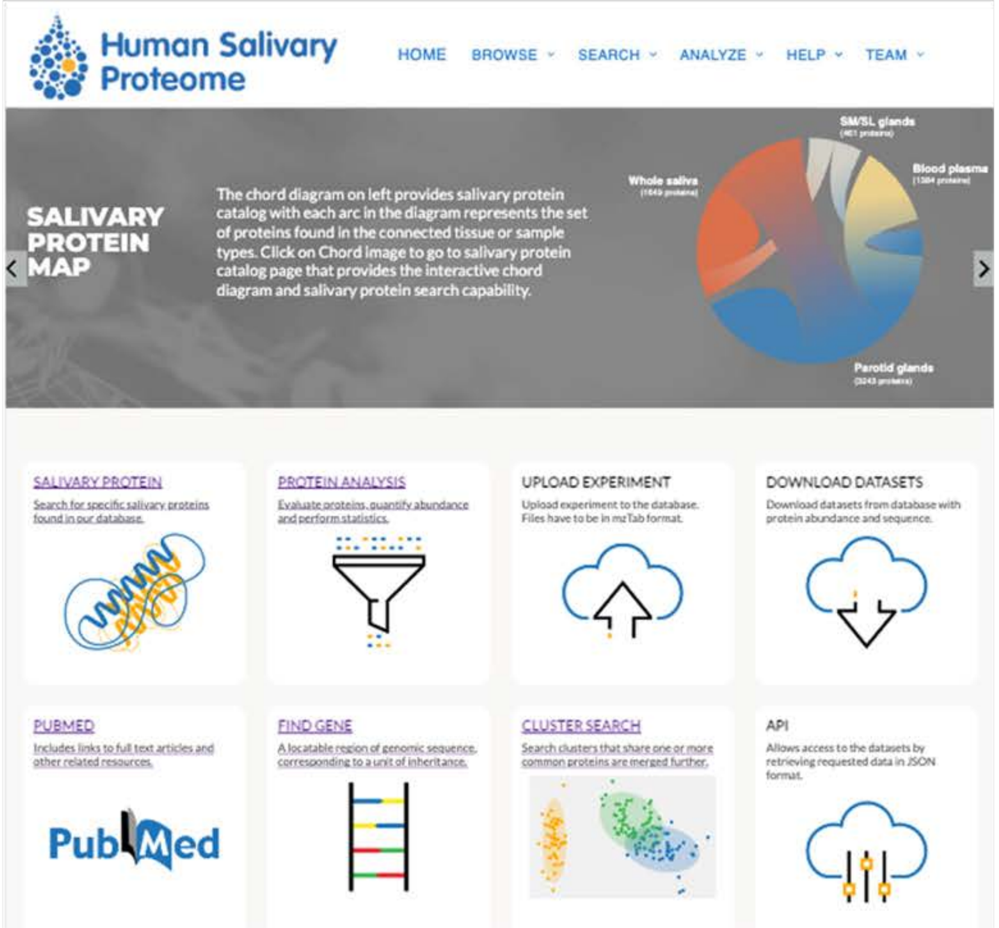
DOWNLOAD DATASETS
Download datasets from database with protein abundance and sequence.

PUBMED
Includes links to full text articles and other related resources.

FIND GENE
A locatable region of genomic sequence, corresponding to a unit of inheritance.

CLUSTER SEARCH
Search clusters that share one or more common proteins are merged further.

API
Allows access to the datasets by retrieving requested data in JSON format.



SALIVARY PROTEIN MAP

The chord diagram on left provides salivary protein catalog with each arc in the diagram represents the set of proteins found in the connected tissue or sample types. Click on Chord image to go to salivary protein catalog page that provides the interactive chord diagram and salivary protein search capability.

SALIVARY PROTEIN
Search for specific salivary proteins found in our database.

PROTEIN ANALYSIS
Evaluate proteins, quantify abundance and perform statistics.

UPLOAD EXPERIMENT
Upload experiment to the database. Files have to be in mzTab format.

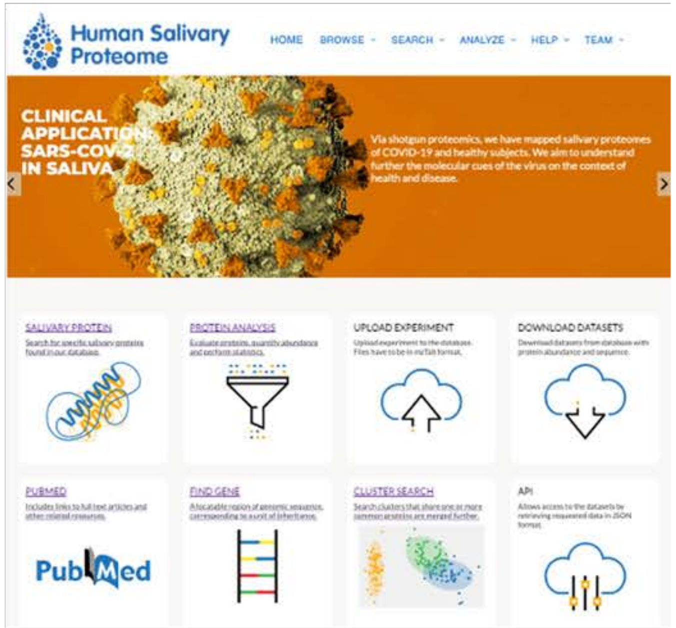
DOWNLOAD DATASETS
Download datasets from database with protein abundance and sequence.

PUBMED
Includes links to full text articles and other related resources.

FIND GENE
A locatable region of genomic sequence, corresponding to a unit of inheritance.

CLUSTER SEARCH
Search clusters that share one or more common proteins are merged further.

API
Allows access to the datasets by retrieving requested data in JSON format.



CLINICAL APPLICATION: SARS-COV-2 IN SALIVA

Via shotgun proteomics, we have mapped salivary proteomes of COVID-19 and healthy subjects. We aim to understand further the molecular cues of the virus on the context of health and disease.

SALIVARY PROTEIN
Search for specific salivary proteins found in our database.

PROTEIN ANALYSIS
Evaluate proteins, quantify abundance and perform statistics.

UPLOAD EXPERIMENT
Upload experiment to the database. Files have to be in mzTab format.

DOWNLOAD DATASETS
Download datasets from database with protein abundance and sequence.

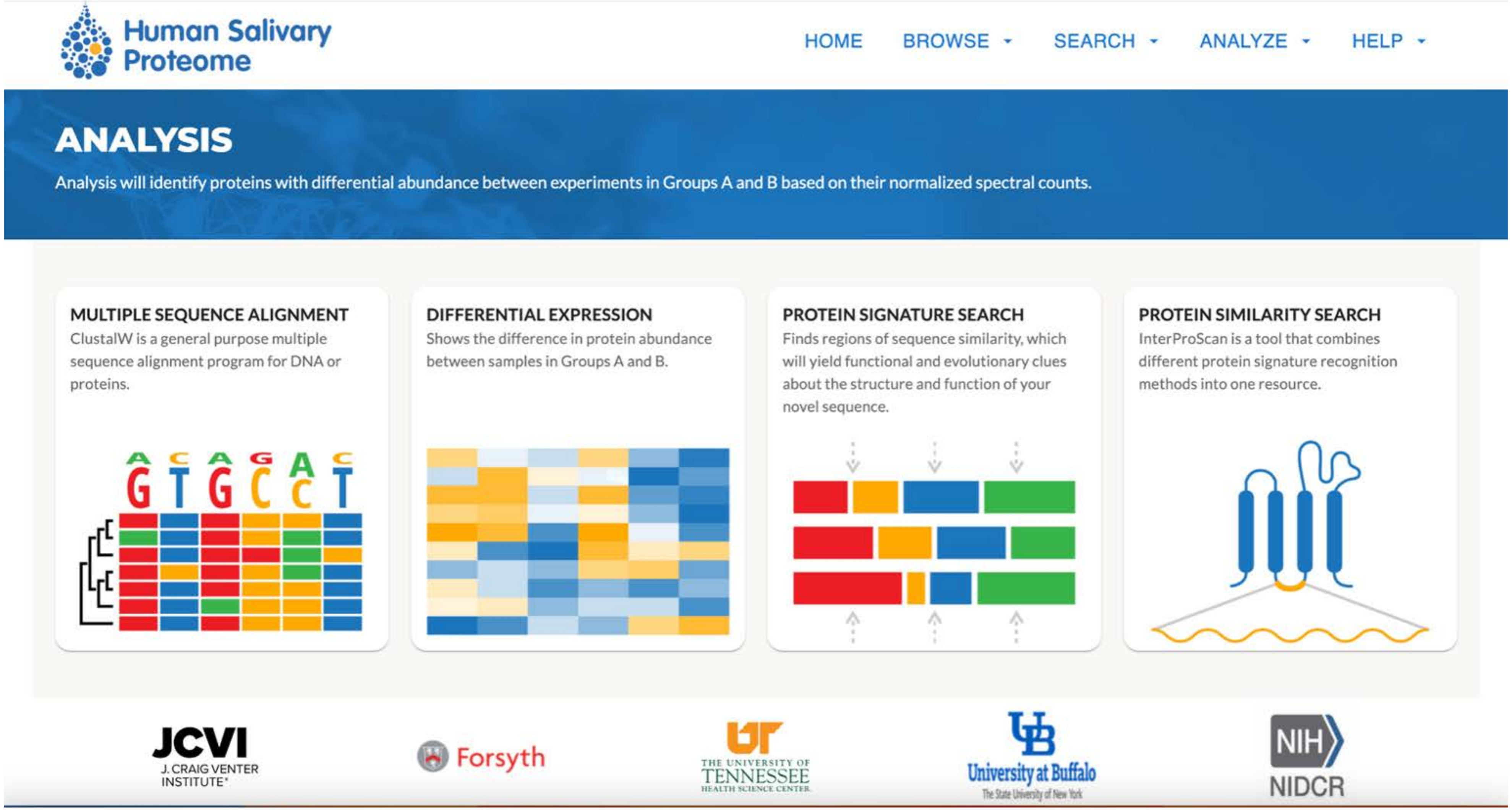
PUBMED
Includes links to full text articles and other related resources.

FIND GENE
A locatable region of genomic sequence, corresponding to a unit of inheritance.

CLUSTER SEARCH
Search clusters that share one or more common proteins are merged further.

API
Allows access to the datasets by retrieving requested data in JSON format.

ANALYSIS PAGE


Human Salivary Proteome
HOME
BROWSE
SEARCH
ANALYZE
HELP
A C A G A C
G T G C C T
JCVI J. CRAIG VENTER INSTITUTE
Forsyth
THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER
University at Buffalo The State University of New York
NIH NIDCR

ANALYSIS

Analysis will identify proteins with differential abundance between experiments in Groups A and B based on their normalized spectral counts.

MULTIPLE SEQUENCE ALIGNMENT

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins.

DIFFERENTIAL EXPRESSION

Shows the difference in protein abundance between samples in Groups A and B.

PROTEIN SIGNATURE SEARCH

Finds regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of your novel sequence.

PROTEIN SIMILARITY SEARCH

InterProScan is a tool that combines different protein signature recognition methods into one resource.

Cloud Grant Aims:

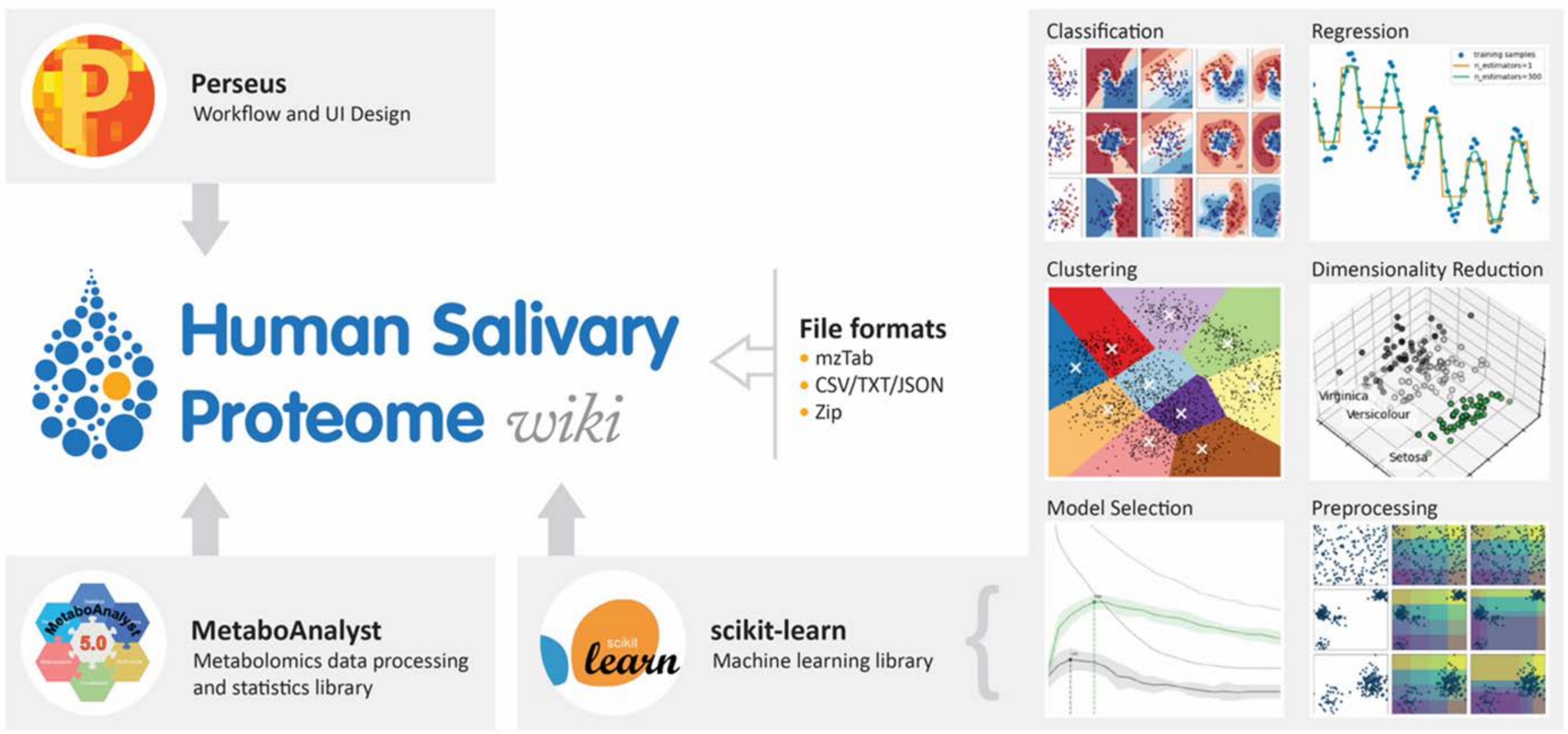
Aim 1: Develop a novel proteomics data analytics platform to support explainable and optimal machine learning identification of salivary proteomic biomarkers through visual analytics and parallel computing.

Aim 2 : Develop a scalable and on-demand cloud pipeline to identify protein modification, protein structure prediction using AlphaFold2, and 3D models comparison tool using MolStar.

Specific Aim 1: Develop a novel proteomics data analytics platform to support explainable and optimal machine learning identification of salivary proteomic biomarkers through visual analytics and parallel computing.

The goals are:
A) Review and integrate the advantages of the existing different general analytical methods and systems

B) Design and produce a specific but efficient and user-friendly platform on the Cloud for machine learning salivary proteomics data analysis



Perseus
Workflow and UI Design

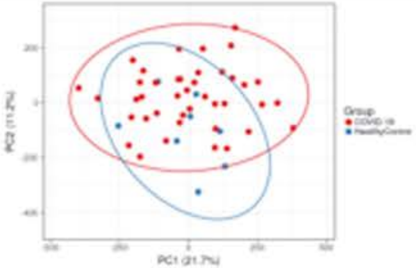
Human Salivary Proteome wiki

File formats
- mzTab
- CSV/TXT/JSON
- Zip

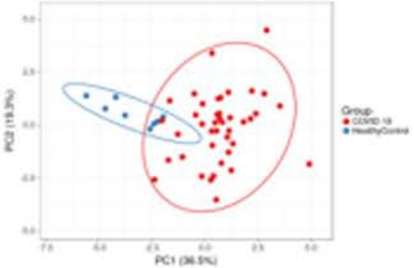
MetaboAnalyst
Metabolomics data processing and statistics library

scikit-learn
Machine learning library

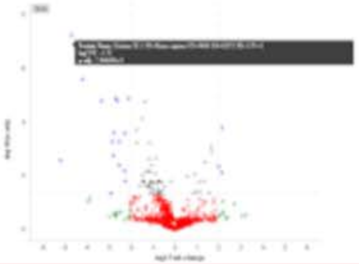
Goals

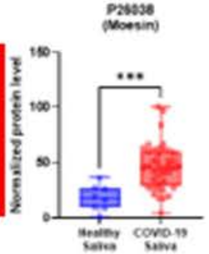


Without feature selection



With feature selection (top 12)





Normalization → Feature Selection

Classification → Predictive Modeling

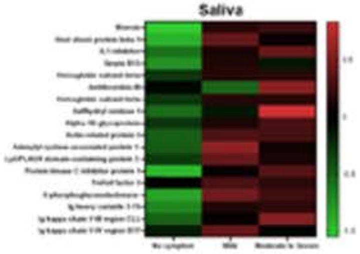
Cluster Analysis → Pathway Enrichment

Network Analysis → Correlation/Causal Relationship

Biomarker Identification

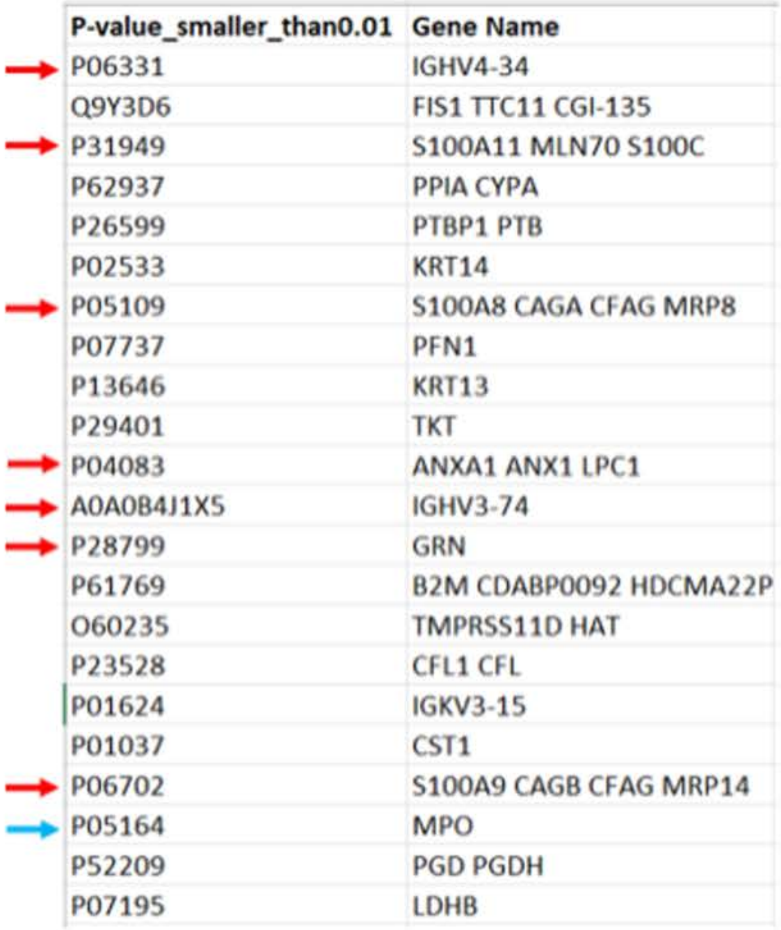
Wilcoxon Rank Sum test
Wald-Wolfowitz test
Kolmogorov-Smirnov test
Shapiro-Wilk test
Anderson-Darling test

Principal component analysis
t-SNE and UMAP nonlinear embedding visualization
Linear discriminant analysis



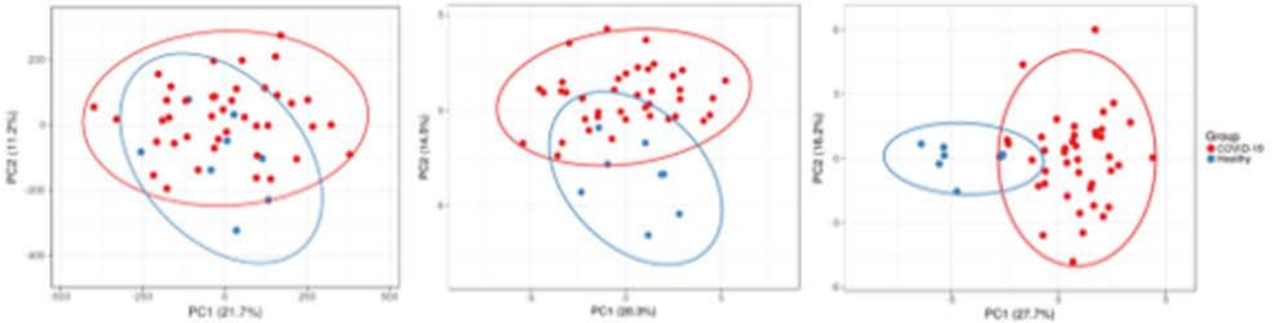
A metadata-driven **visual analytics** interface and platform for Web-based exploration and analysis of salivary proteomics data

Preliminary Data

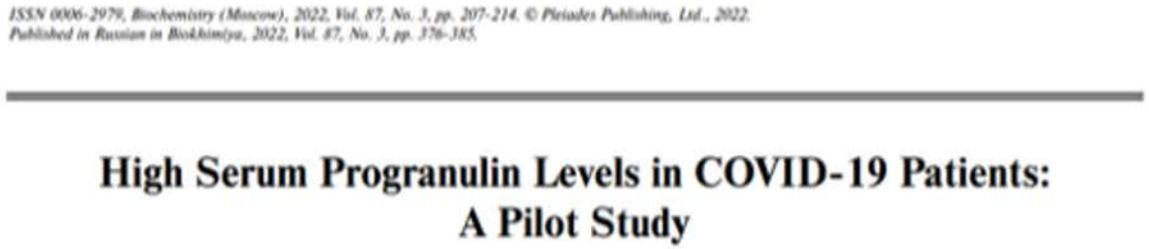


	P-value_smaller_than0.01	Gene Name
→	P06331	IGHV4-34
	Q9Y3D6	FIS1 TTC11 CGI-135
→	P31949	S100A11 MLN70 S100C
	P62937	PPIA CYPA
	P26599	PTBP1 PTB
	P02533	KRT14
→	P05109	S100A8 CAGA CFAG MRP8
	P07737	PFN1
	P13646	KRT13
	P29401	TKT
→	P04083	ANXA1 ANX1 LPC1
→	A0A0B4J1X5	IGHV3-74
→	P28799	GRN
	P61769	B2M CDABP0092 HDCMA22P
	O60235	TMPRSS11D HAT
	P23528	CFL1 CFL
	P01624	IGKV3-15
	P01037	CST1
→	P06702	S100A9 CAGB CFAG MRP14
→	P05164	MPO
	P52209	PGD PGDH
	P07195	LDHB

Rank sum test p-value < 0.01

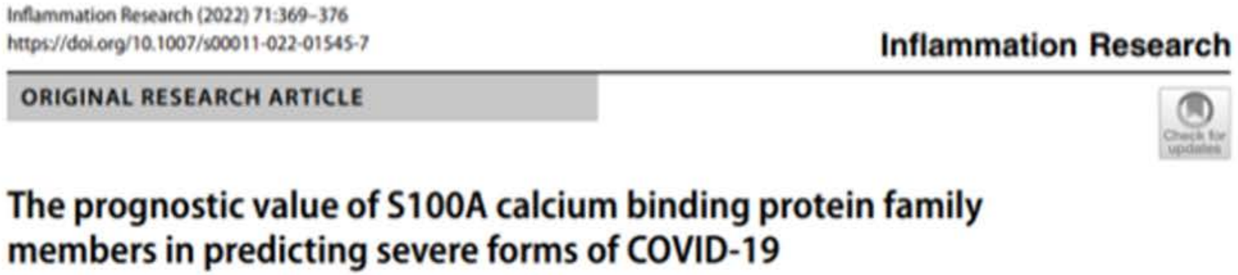


Without feature selection Wald-Wolfowitz test Wilcoxon rank sum test



ISSN 0006-2979, Biochemistry (Moscow), 2022, Vol. 87, No. 3, pp. 207-214. © Pleiades Publishing, Ltd., 2022.
Published in Russian in Biokhimiya, 2022, Vol. 87, No. 3, pp. 376-385.

High Serum Progranulin Levels in COVID-19 Patients: A Pilot Study



Inflammation Research (2022) 71:369–376
https://doi.org/10.1007/s00011-022-01545-7

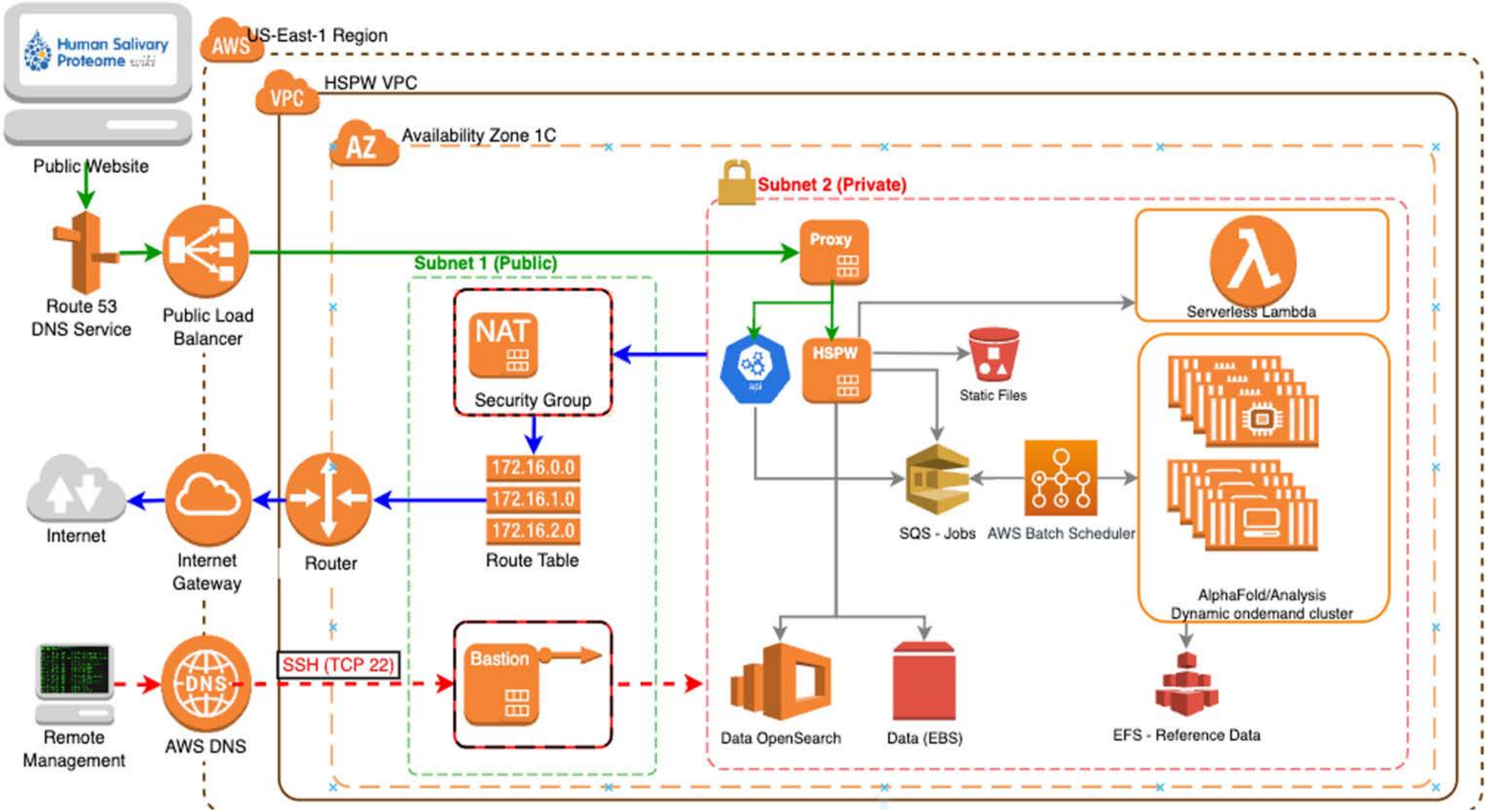
Inflammation Research

ORIGINAL RESEARCH ARTICLE

Check for updates

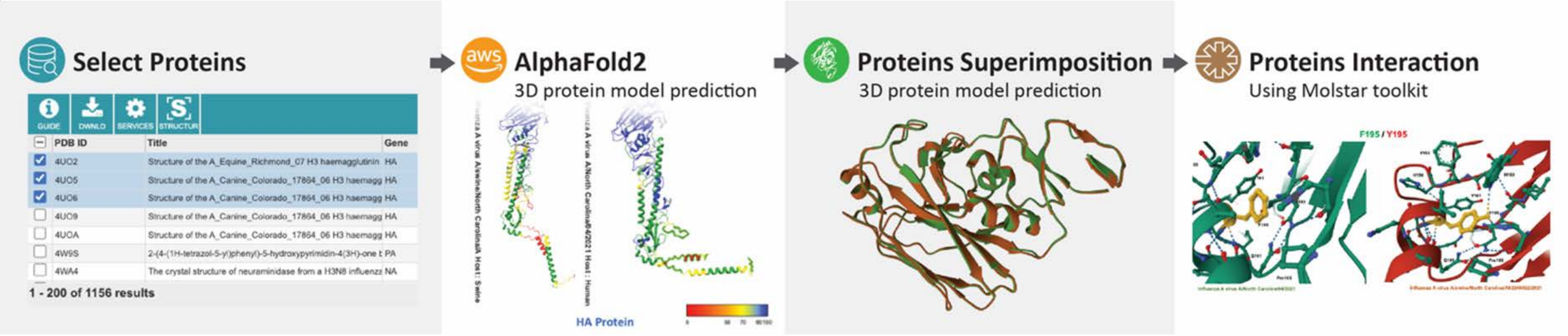
The prognostic value of S100A calcium binding protein family members in predicting severe forms of COVID-19

Infrastructure



JCVI J. CRAIG VENTER INSTITUTE™

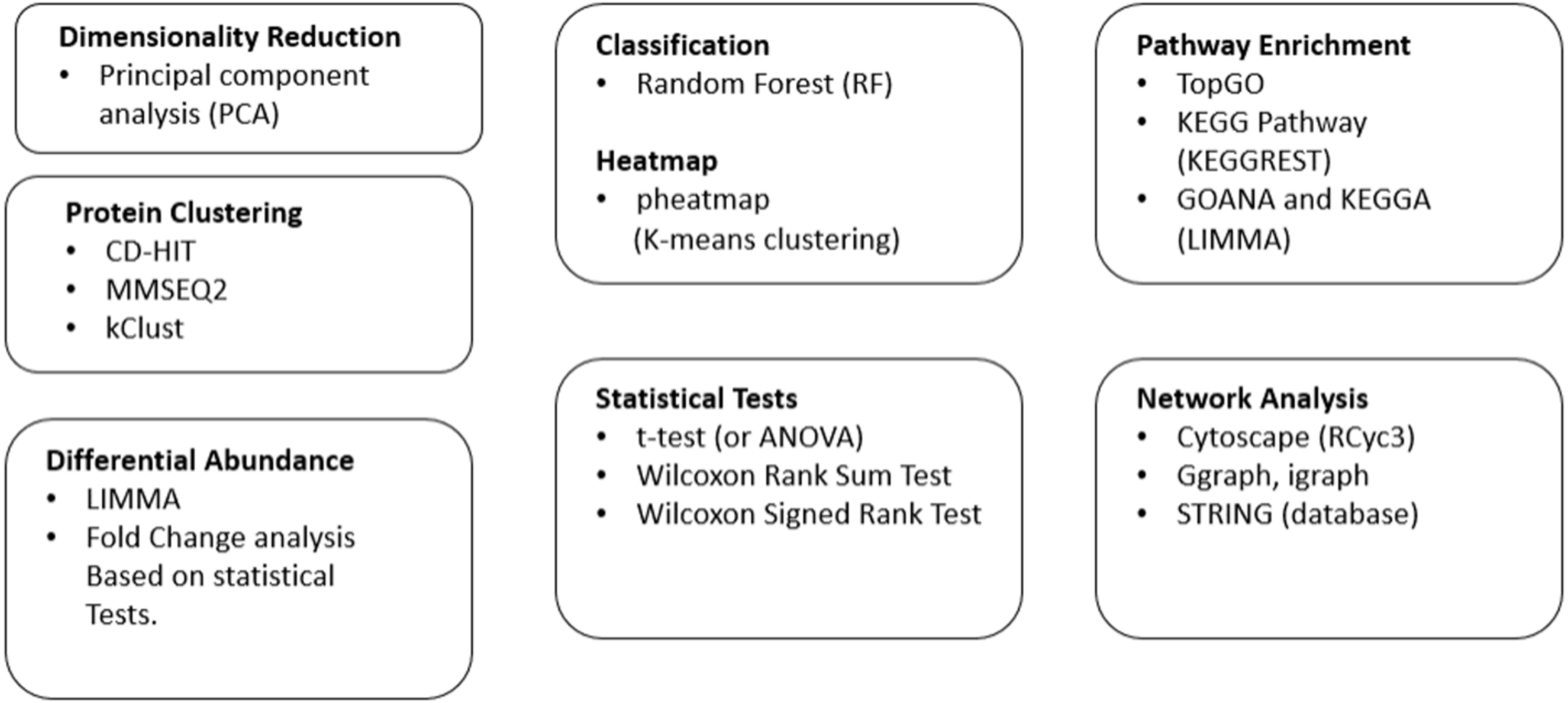
Specific Aim 2: Develop a scalable and on-demand cloud pipeline to identify protein modification, protein structure prediction using AlphaFold2, and 3D models comparison tool using MolStar.


Select Proteins
GUIDE
DWNLD
SERVICES
STRUCTUR
AlphaFold2
3D protein model prediction
...nza A virus A/swine/North Carolina/A Host : Swine
...nza A virus A/North Carolina/4/2021 Host : Human
HA Protein
Proteins Superimposition
3D protein model prediction
Proteins Interaction
Using Molstar toolkit
F195 / Y195

⊟	PDB ID	Title	Gene
☑	4UO2	Structure of the A_Equine_Richmond_07 H3 haemagglutinin	HA
☑	4UO5	Structure of the A_Canine_Colorado_17864_06 H3 haemagg	HA
☑	4UO6	Structure of the A_Canine_Colorado_17864_06 H3 haemagg	HA
☐	4UO9	Structure of the A_Canine_Colorado_17864_06 H3 haemagg	HA
☐	4UOA	Structure of the A_Canine_Colorado_17864_06 H3 haemagg	HA
☐	4W9S	2-(4-(1H-tetrazol-5-yl)phenyl)-5-hydroxypyrimidin-4(3H)-one t	PA
☐	4WA4	The crystal structure of neuraminidase from a H3N8 influenzz	NA

1 - 200 of 1156 results

Tools Needed



Dimensionality Reduction

- Principal component analysis (PCA)

Protein Clustering

- CD-HIT
- MMSEQ2
- kClust

Differential Abundance

- LIMMA
- Fold Change analysis Based on statistical Tests.

Classification

- Random Forest (RF)

Heatmap

- pheatmap (K-means clustering)

Statistical Tests

- t-test (or ANOVA)
- Wilcoxon Rank Sum Test
- Wilcoxon Signed Rank Test

Pathway Enrichment

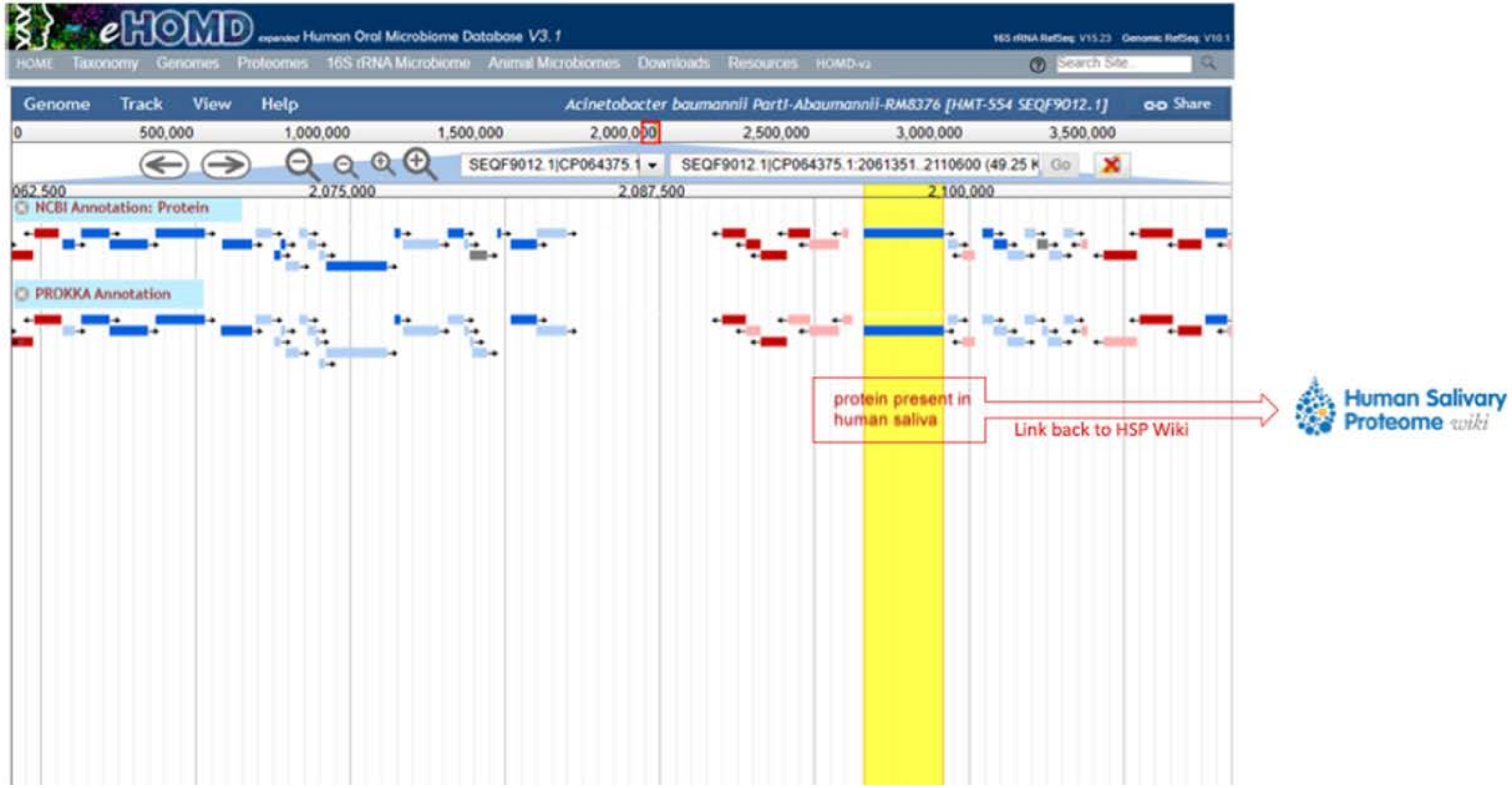
- TopGO
- KEGG Pathway (KEGGREST)
- GOANA and KEGGA (LIMMA)

Network Analysis

- Cytoscape (RCyc3)
- Ggraph, igraph
- STRING (database)

Future Insights: Host-Microbial Proteins

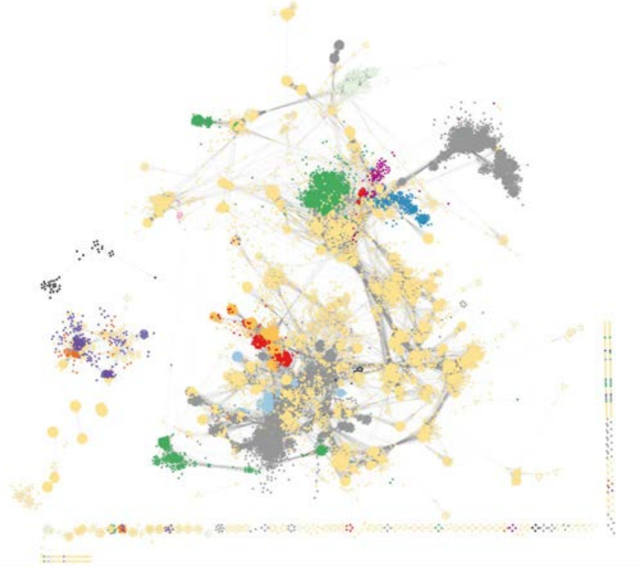
Genes encoding the protein can be displayed in HOMD Genome Browser, with link back to WiKi


eHOMD expanded Human Oral Microbiome Database V3.1
16S rRNA RefSeq V15.23 Genomic RefSeq V10.1
HOME Taxonomy Genomes Proteomes 16S rRNA Microbiome Animal Microbiomes Downloads Resources HOMD-v2
Search Site...
Genome Track View Help
Acinetobacter baumannii PartI-Abaumannii-RM8376 [HMT-554 SEQF9012.1]
Share
0 500,000 1,000,000 1,500,000 2,000,000 2,500,000 3,000,000 3,500,000
SEQF9012.1|CP064375.1
SEQF9012.1|CP064375.1:2061351..2110600 (49.25 K
Go
062,500 2,075,000 2,087,500 2,100,000
NCBI Annotation: Protein
PROKKA Annotation
Human Salivary Proteome wiki

protein present in human saliva

Link back to HSP Wiki

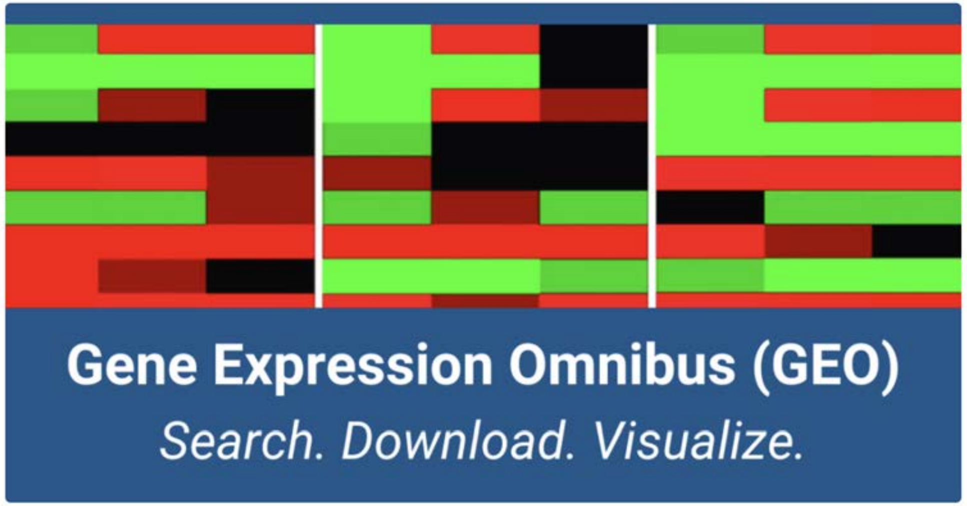
From: Drs. Jessica Mark Welch and George Chen (Forsyth Institute)



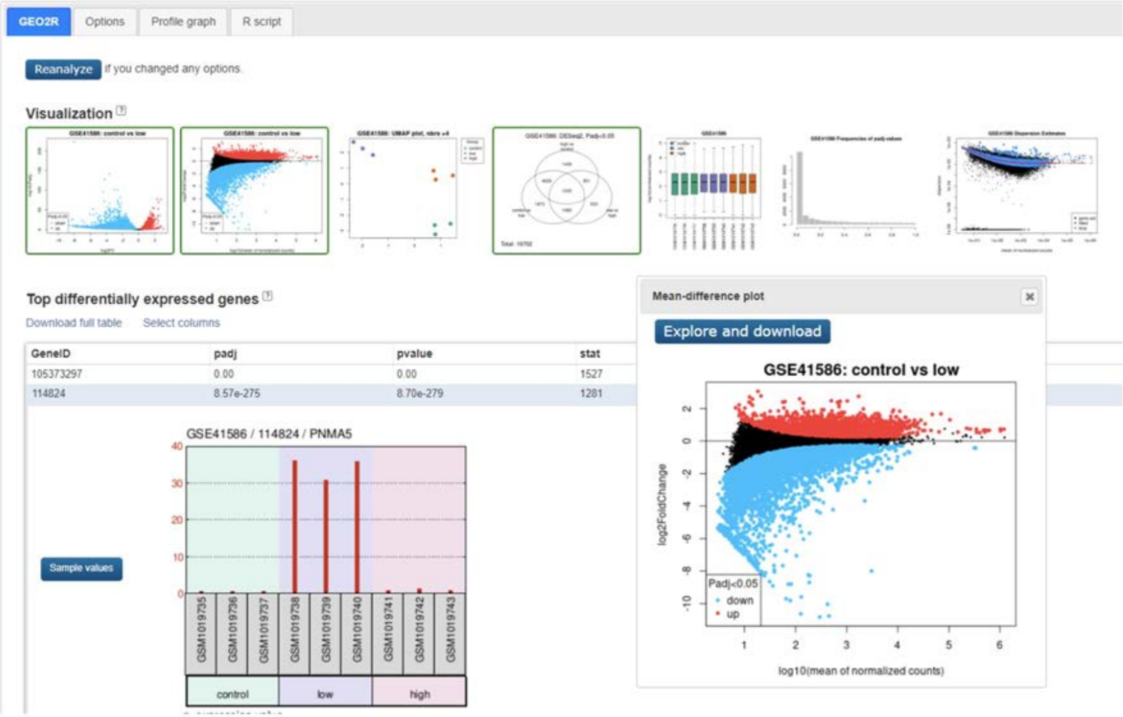
Overview of 16,745 predicted prophage-like entities in HOMD bacterial genomes (HOMD v0.1 draft). Nodes represent predicted prophage genomes; selected phages are colored by the host in which they reside (e.g. *Streptococcus*: green; *Tannerella*: orange; *Prevotella*: purple; *Aggregatibacter*: red; *Fusobacterium*: blue).

From: Dr. Kauffmm (U at Buffalo)

Open Source Visual Analytics (NCBI)



Search, Download, and Visualize Human RNA-Seq Gene Expression Data in NCBI's Gene Expression Omnibus (GEO)



Reanalyze if you changed any options.

Top differentially expressed genes

Download full table Select columns

GeneID	padj	pvalue	stat
105373297	0.00	0.00	1527
114824	8.57e-275	8.70e-279	1281

Figure 1: Screenshot of GEO2R differential gene expression analysis results, including Volcano, Mean difference, Mean variance, UMAP, Venn, Boxplot, and Histogram plots.

Source: NCBI

Acknowledgements
- Team: HOMD+HSP Wiki

JCVI

- Marcelo Freire
- Indresh Singh
- Harinder Singh
- Wan Hin (Marco) Choi
- Max Qian

The Forsyth Institute

- Jessica Mark Welch
- Floyd Dewhirst
- Tsute (George) Chen
- Markus Hardt

University at Buffalo

- Stefan Ruhl
- Kathryn Kauffman

University of Tennessee

- Yanhui Zhang

National Institute of De
and Craniofacial Resea

- NIH / NIDCR Support:

R01DE016937-16S1

R01DE016937

JCVI J. CRAIG VENTER INSTITUTE™

Questions

mfreire@jcvi.org

@drmfreire

"Enhancing human immunity to impact global disparities"

JCVI J. CRAIG VENTER INSTITUTE™