University of California
San Francisco

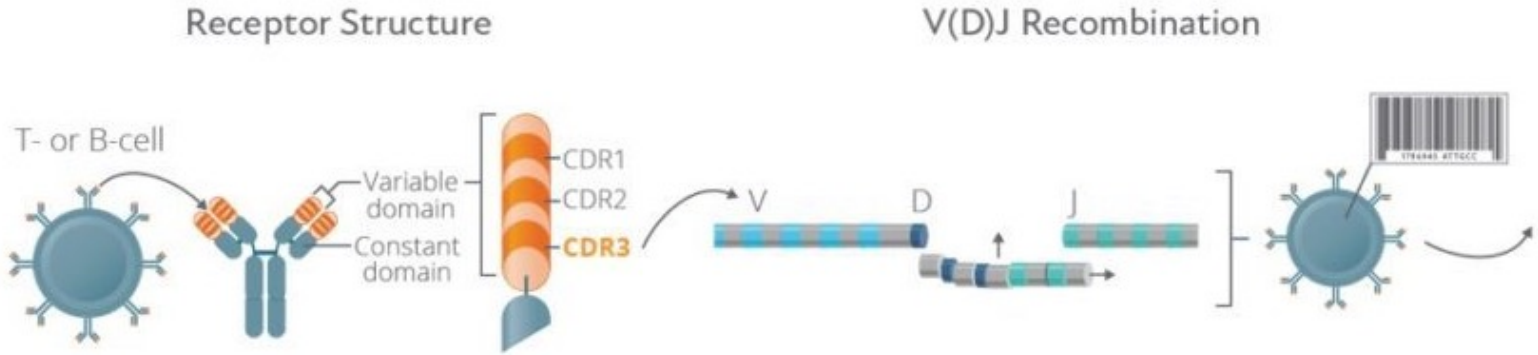
Computational Approaches to Unravel Immune Receptor Sequencing

Li Zhang
Professor of Biostatistics
Department of Medicine
Department of Epidemiology and Biostatistics
HDFCCC Biostatistics Core
University of California San Francisco
li.zhang@ucsf.edu

Outline

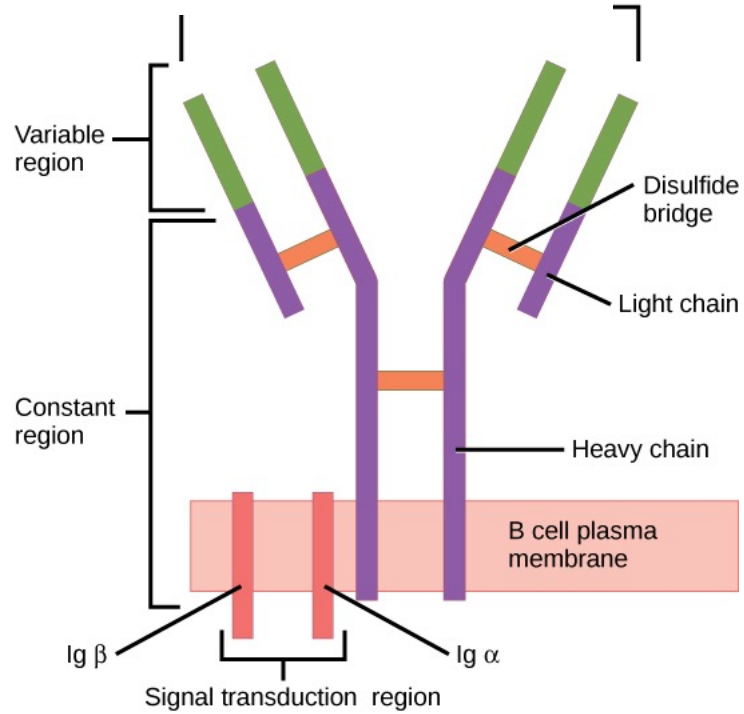
- Background and Introduction
  - T-cell/B-cell Receptor and Repertoire Sequencing
- Proposed Analysis Pipelines with Examples
- Conclusion and Future Work

T-cell Receptor (TCR) and B-cell Receptor (BCR)

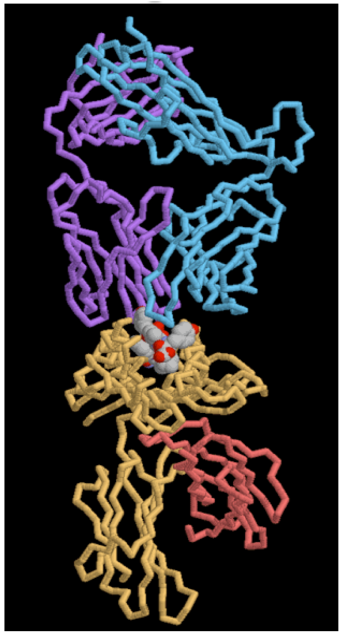


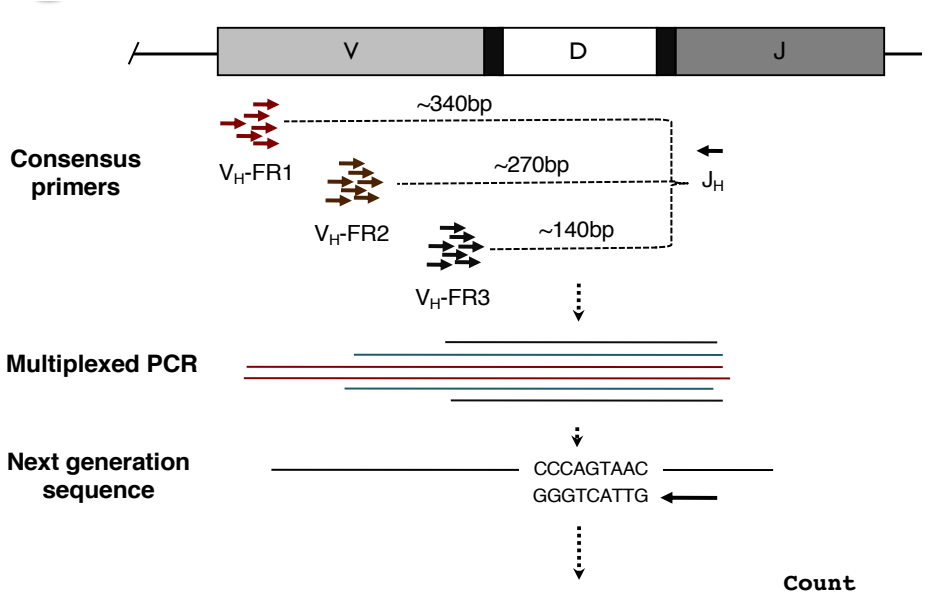
TCR is a protein complex found on the surface of T cells, or T lymphocytes, that is responsible for recognizing fragments of antigen as peptides bound to MHC molecules.

BCR is composed of immunoglobulin molecules that form a type 1 transmembrane receptor protein usually located on the outer surface of B cells.



Overview of TCR Repertoire Sequencing





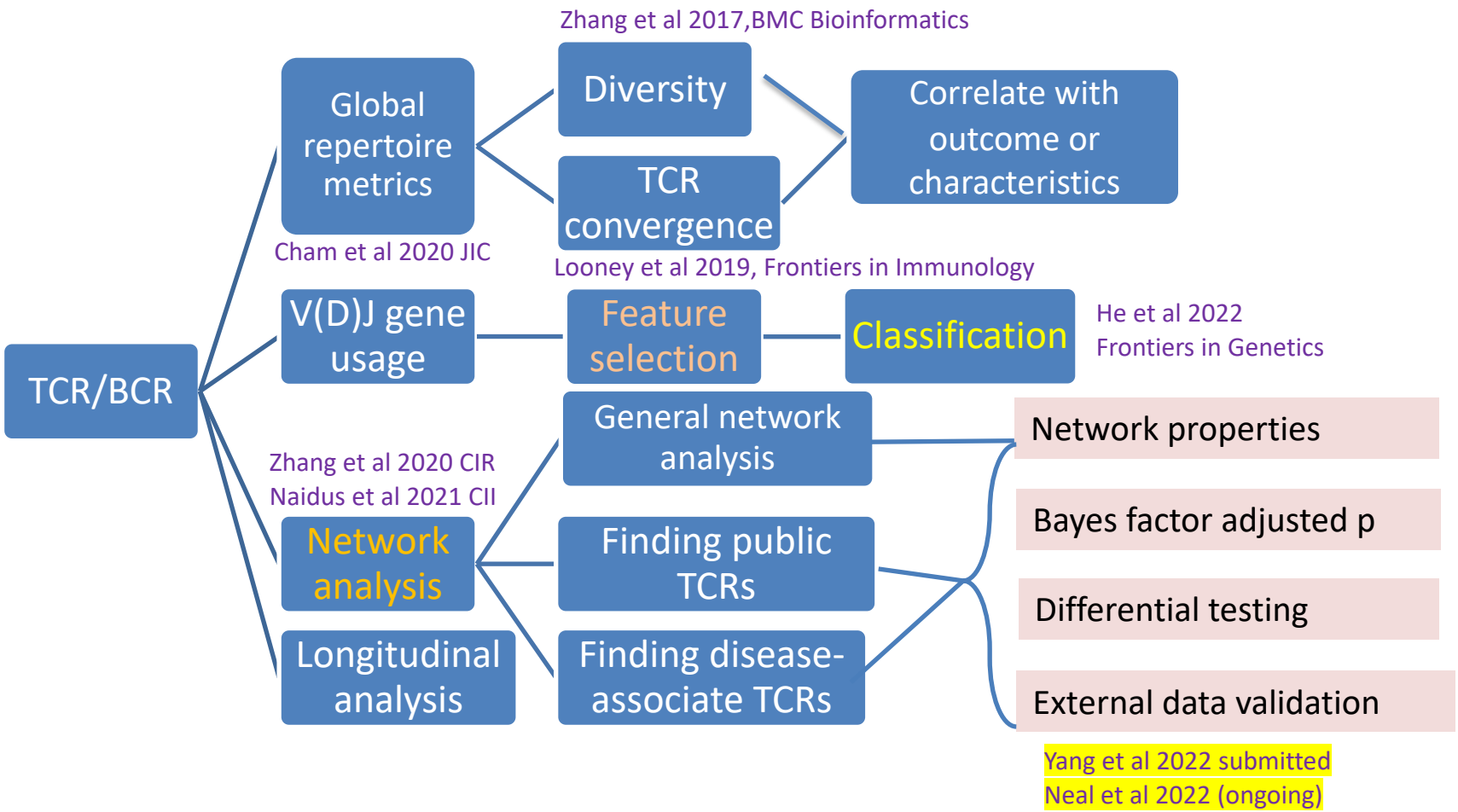
		Count
1	AAAGCGACATTGGGATCTGTCAGTTGTCATTCGCG	1321
2	GCGGTTTTGTAGAAGGTTAGGGGAATAGGTTAGAT	1122
3	TGAGTGGCTTAAGAATGTAAAATCTGGGATTATAG	901
4	TGTAGTAATCTCTGATTAACGGTGACGGTTTTAAG	534
5	GAAGAATAATTAAGAAAAAAGCACCCCTCGTCGCC	421
6	TAGAATTACCTACCGCGGTCCACCATACCTTCGAT	132
7	TATCGCGCCCACTCTCCCATTAGTCGGCAGAGGTG	113

10^{6-} 10^{8} Independent Sequence Reads Per Run

(Adapted from Aaron Logan)

ImmunoSEQ Assay

NAIR: Proposed Analysis Pipeline

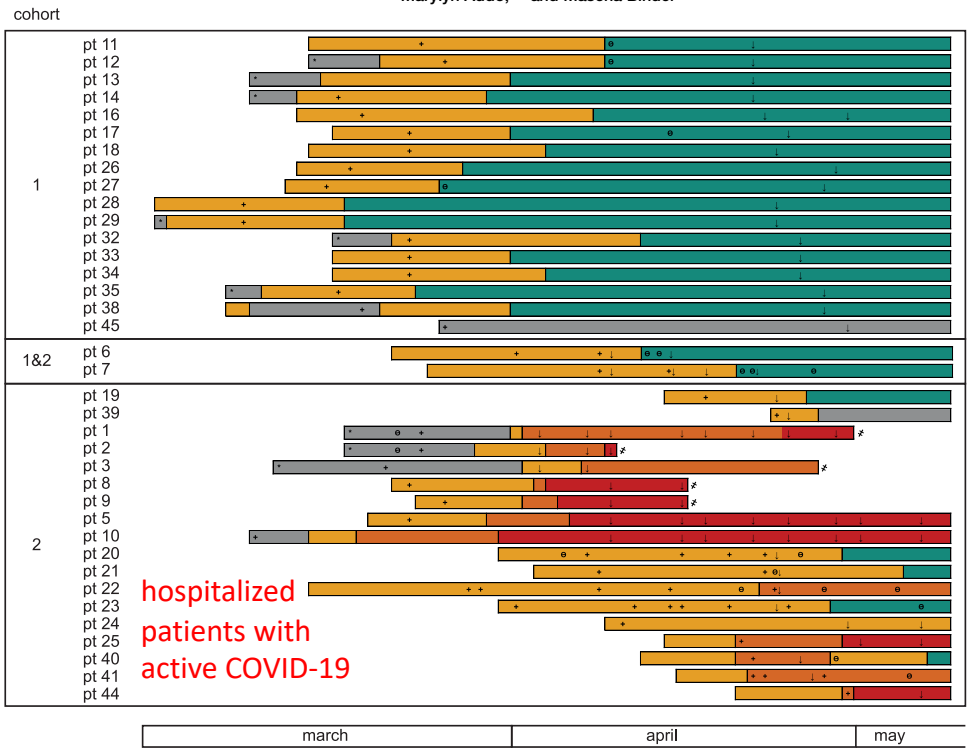


TCR Repertoire Sequences European COVID-19 Patients

Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease

Christoph Schultheiß,[1,11] Lisa Paschold,[1,11] Donjete Simnica,[1,11] Malte Mohme,[2] Edith Willscher,[1] Lisa von Wenserski,[1] Rebekka Scholz,[1] Imke Wieters,[3] Christine Dahlke,[4,5] Eva Tolosa,[6] Daniel G. Sedding,[7] Sandra Ciesek,[8,9,10] Marylyn Addo,[4,5] and Mascha Binder[1,12,*]

recovered without medical intervention



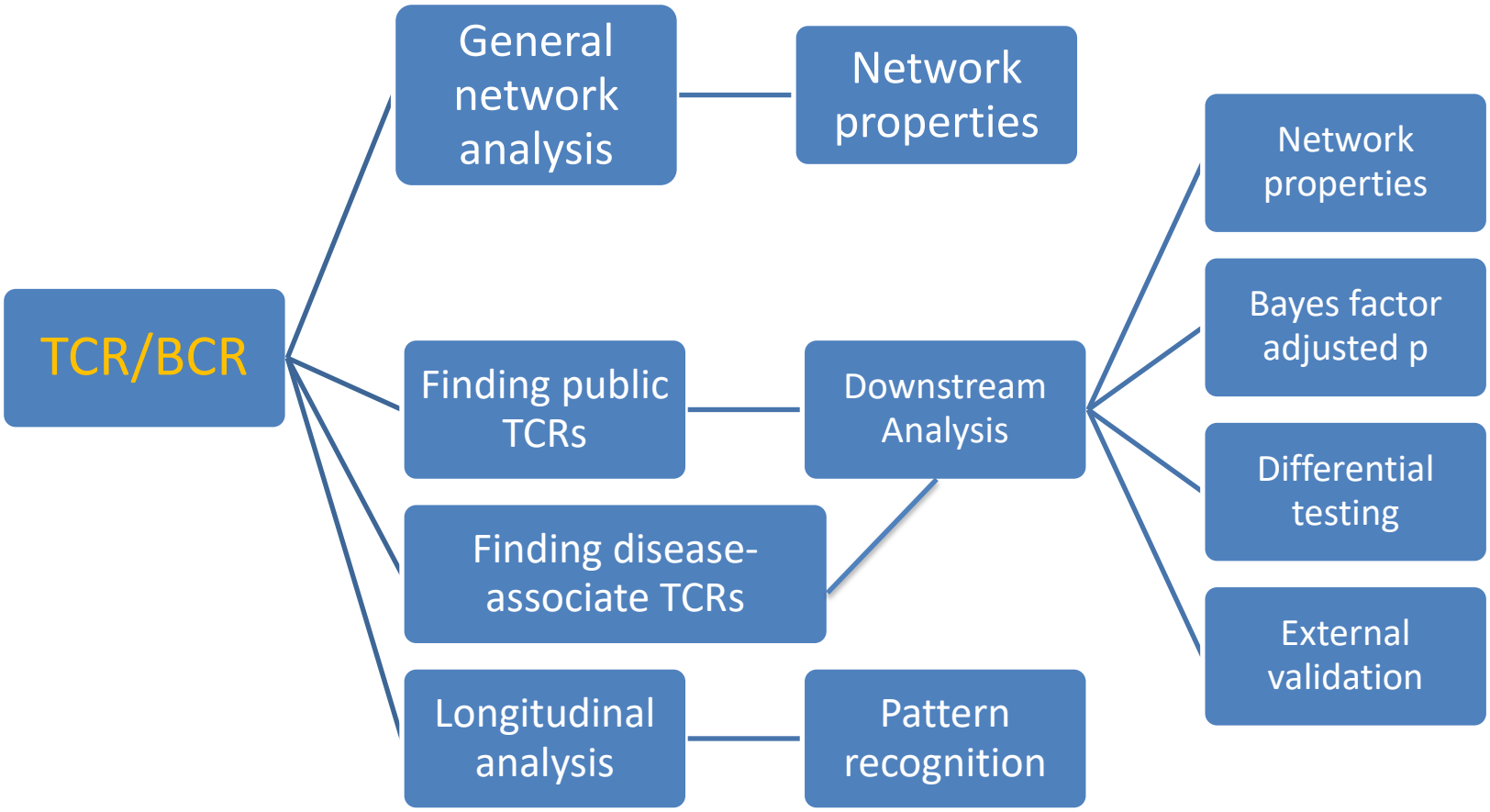
cohort	patient ID	age range [y]	sex	diagnosis	severity	respiratory status	duration of sympt.[d]	relevant risk factors[$]
1	11	30-39	m	PCR	mild	spont. breath.	25	none
	12	20-29	f	PCR	mild	spont. breath.	19	none
	13	40-49	f	serological	mild	spont. breath.	16	none
	14	50-59	m	PCR	mild	spont. breath.	16	none
	16	20-29	m	PCR	mild	spont. breath.	15	HTN
	17	30-39	f	PCR	mild	spont. breath.	25	none
	18	30-39	m	PCR	mild	spont. breath.	20	none
	26	40-49	m	PCR	mild	spont. breath.	14	none
	27	20-29	m	PCR	mild	spont. breath.	13	none
	28	30-39	f	PCR	mild	spont. breath.	16	none
	29	30-39	m	PCR	mild	spont. breath.	15	none
	32	20-29	f	PCR	mild	spont. breath.	21	none
	33	30-39	m	PCR	mild	spont. breath.	15	none
	34	40-49	m	PCR	mild	spont. breath.	18	none
	35	60-69	f	PCR	mild	spont. breath.	13	none
	38	20-29	f	PCR	mild	spont. breath.	13	none
	45	30-39	m	PCR	asymptomatic	spont. breath.	NA	none
1&2	6	60-69	f	PCR	moderate*	spont. breath.	21	HTN, age
	7	70-79	f	PCR	moderate*	spont. breath.	26	HTN, DM, age
2	19	20-29	m	PCR	moderate*	spont. breath.	12	none
	39	30-39	m	PCR	moderate*	spont. breath.	4	none
	1	60-69	m	PCR	fatal	ECMO	28	cancer, age
	2	60-69	m	PCR	fatal	ECMO	12	cancer
	3	70-79	m	PCR	fatal	mech. vent.	25	cancer, age
	8	40-49	m	PCR	fatal	ECMO	25d	HTN
	9	60-69	m	PCR	fatal	ECMO	23	HTN, CVD, age
	5	60-69	m	PCR	severe[#]	ECMO	42+	HTN, DM, age
	10	60-69	m	PCR	severe[#]	ECMO	47+	CRD, age
	20	50-59	m	PCR	moderate*	spont. breath.	29	HTN, CVD
	21	50-59	f	PCR	moderate*	spont. breath.	31	DM., HTN, CVD
	22	70-79	m	PCR	severe[#]	mech. vent.	54+	HTN, CVD, DM, age
	23	70-79	m	PCR	moderate*	spont. breath.	28	CVD, age
	24	80-89	f	PCR	moderate*	spont. breath.	29+	HTN, DM,age
	25	60-69	m	PCR	severe[#]	ECMO	19+	age
	40	70-79	f	PCR	severe[#]	mech. vent.	24	HTN
	41	70-79	m	PCR	severe[#]	mech. vent.	22+	CVD
	44	70-79	m	PCR	severe[#]	ECMO	18+	HTN, CVD, CRD, age

*requiring hospitalization; [#] requiring ventilation support; [$]according to Wu, C., et al. 2020a; HTN – hypertension; CVD – cardiovascular disease; DM – diabetes; CRD – chronic respiratory disease

asymptomatic	+ positive SARS-CoV-2 PCR
symptomatic	* COVID-19 contact
ventilation	ϴ negative SARS-CoV-2 PCR
ECMO	↓ sample collection
recovered	≠ death

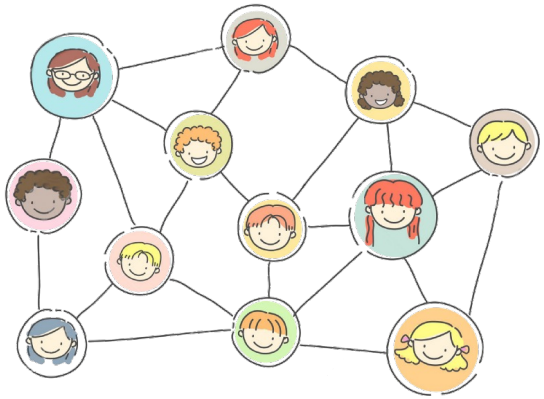
contained sequences from a total of 37 patients, including 69 time points, and overall >6.2 million BCR and >8.3 million TCR sequences

Recall Major Pipelines



Network Analysis

Node	Account	Nucleotide clone
Distance	Minimum number of accounts between two accounts	Number of nucleotide differences between two clones
Edge	Relationship	Only one nucleotide change between two nodes
Distance matrix	Friendship info among a group	Pairwise distances among clones
Attributes	Photos or posts	Meta data in nucleotide clone
Cluster	Groups in FB	A group of clones having direct or indirect connection



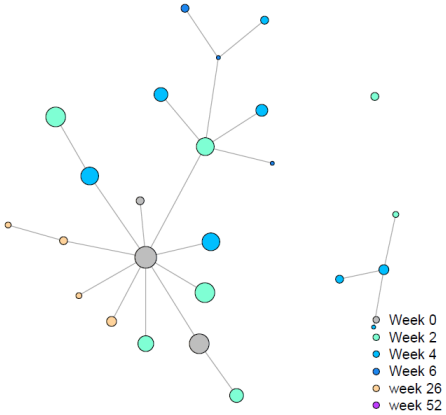


Image source: https://www.freecodecamp.org/news/deep-dive-into-graph-traversals-227a90c6a261/

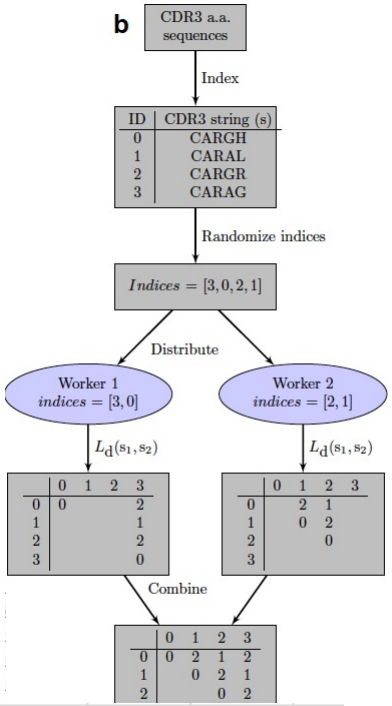
Distance Matrix

Levenshtein distance

- Cat → fat (transformation) distance = 1
- Health → healthy (insertion) distance = 1
- Sunny → sun (deletion) distance = 2

Similar as

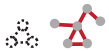
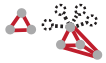
- ATCG → ATGG (transformation) distance = 1
- ATCG → ATTCG (insertion) distance = 1
- ATCG → ACG (deletion) distance = 1



ASSQDTGNTEAF	A	S	S	Q	D		T	G	N	T	E	A	F
ASSQDRGNTEAF	A	S	S	Q	D		R	G	N	T	E	A	F
ASSQDRMNTEAF	A	S	S	Q	D		R	M	N	T	E	A	F
ASSQDWTGNTEAF	A	S	S	Q	D	W	T	G	N	T	E	A	F
ASSQDWTGYTEAF	A	S	S	Q	D	W	T	G	Y	T	E	A	F
ASSQITGNTEAF	A	S	S	Q		I	T	G	N	T	E	A	F
ASSQDLRMNTEAF	A	S	S	Q	D	L	R	M	N	T	E	A	F

	V26	V27	V28	V29	V30	V31	V32
V26	0	0	0	0	0	0	0
V27	0	0	0	0	0	0	0
V28	0	0	0	1	1	0	0
V29	0	0	1	0	0	0	0
V30	0	0	1	0	0	0	0
V31	0	0	0	0	0	0	1

Network Properties

Network property	Definition (unit*)	Illustration
Node (vertex)	The fundamental unit of which graphs are formed: v	
Edge (link)	An unordered pair of distinct vertices: $\{v, w\}$	
Degree	The number of edges incident to a vertex e: $deg(v)$	
Largest component	Largest subgraph in which any two vertices are connected	
k-core	A maximal subgraph of a graph in which all vertices have degree of at least k	
Clique	A complete subgraph in a graph	
Diameter	The length of the "longest shortest path" between any two vertices: $\max_{(v,w)} d(v, w)$	
Assortativity coefficient	Pearson correlation coefficient of degree between pairs of linked nodes r ={-1,1}	
Cluster size, number	Connected component of a graph in which any two nodes are connected	
Clustering coefficient (transitivity)	The probability that the adjacent vertices of a vertex are connected	
Density	The ratio of the number of edges and the number of possible edges	
Centralization	Centrality score based on node-level centrality c: sum(max$(c(w), w) - c(v), v)$	
Average Degree	The average number of degrees per node: $2e/v$	
Neighborhood	Set of all the nodes that are adjacent to a node v: $N(v)$	

Supplementary Table 1. Network global properties.

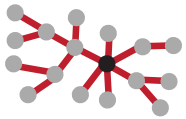
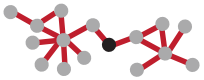
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09278-8 OPEN

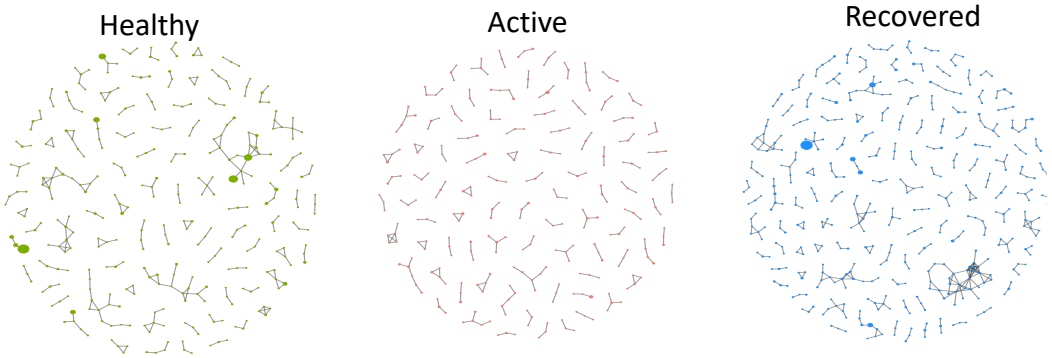
Large-scale network analysis reveals the sequence space architecture of antibody repertoires

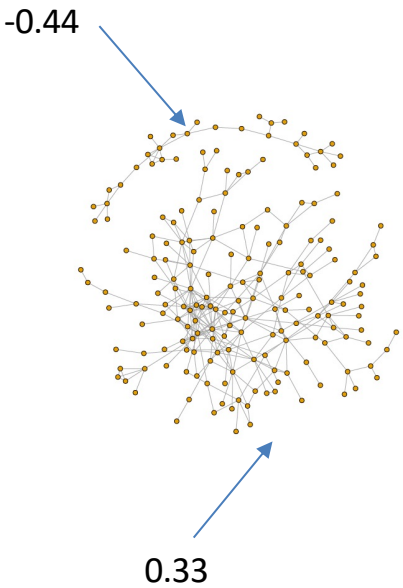
Enkelejda Miho[1,2,3], Rok Roškar[4], Victor Greiff[5] & Sai T. Reddy[1]

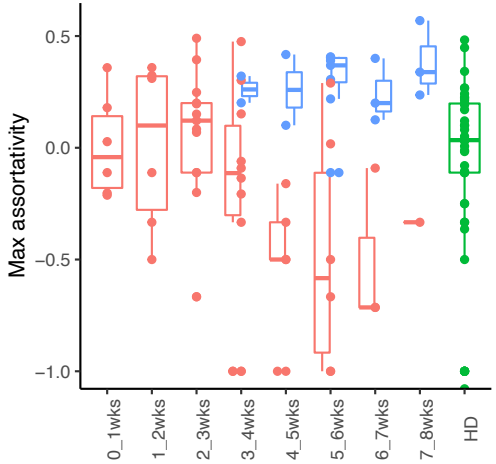
Network property	Definition*	Illustration
Eigenvector	Principal eigenvector of $t(A)*A$, where A is the adjacency matrix of the graph: $x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t$	
Authority	Principal eigenvector of $t(A)*A$, where A is the adjacency matrix of the graph	
PageRank	Principal eigenvector of the normalized matrix of the graph	
Closeness	Node centrality in a graph: $C(v) = \frac{1}{\sum_w d(v,w)}$	
Betweenness	Number of shortest paths through v: $B(v) = \sum_{s \neq v \neq t} \frac{\delta_{st}(v)}{\delta_{st}}$	

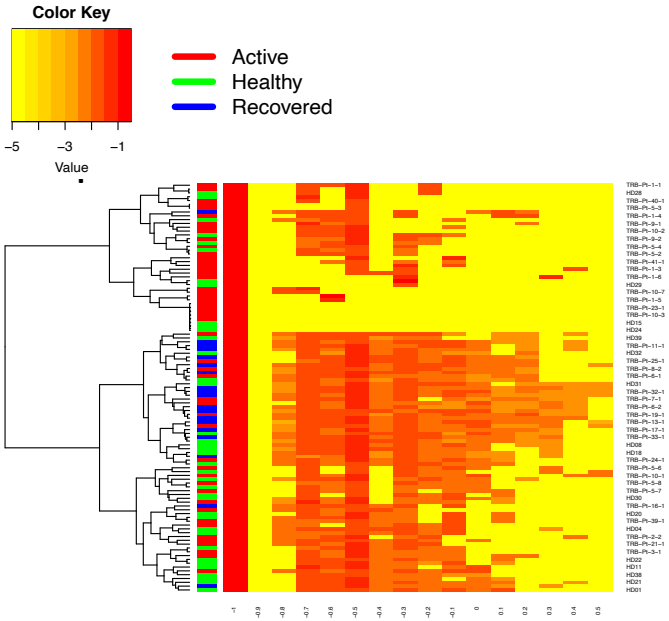
Supplementary Table 2. Network local properties. *These properties are dimensionless.

Network Properties and Immunological Features



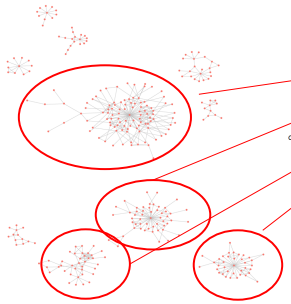




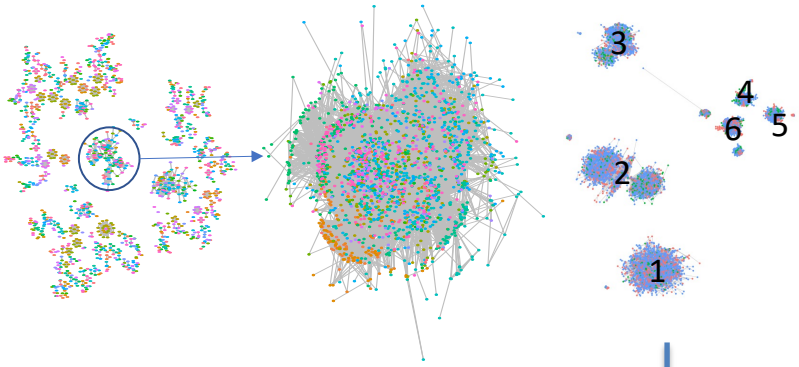


Finding Public Clusters Workflow

Build the network for each sample → Pick the top K largest clusters or single node with large abundance within each sample → Within each cluster, identify a representative clone → Build a new network based on those selected clones → Generate public clusters → Assign global membership to the public clusters

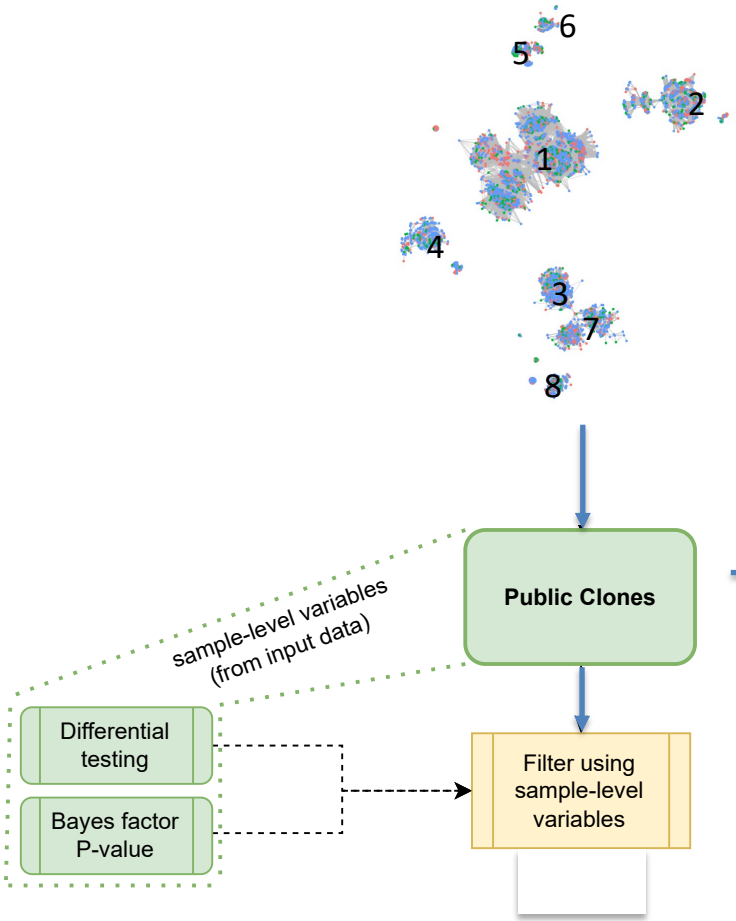


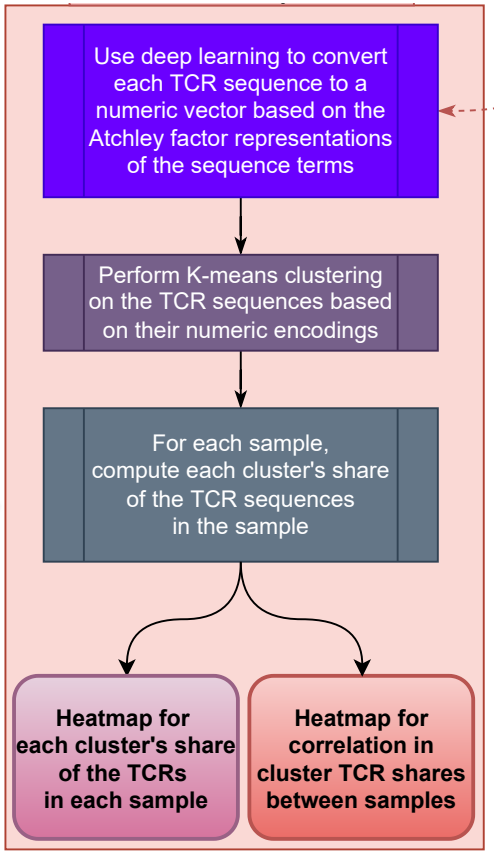
membership	node_count	motif_top_deg_beta	n deg	betaCDR3_le	cou
2	79	CASSLGSYEQYF	10.96	12.52	
1	74	CASSLGTDTQYF	7.7	12.96	
3	45	CASSLGETQYF	12.76	11.89	
4	26	CASSLGTGELFF	5.38	12.81	
27	26	CASSLGGNTEAFF	11.15	13.12	
6	25	CASSIEGQLSTDTQYF	16.4	16	
5	17	CASSLGNEQFF	4.71	11.71	
28	14	CASSLGGNQPQHF	4.43	13	
22	13	CASSLGGNYGYTF	4.62	13.08	



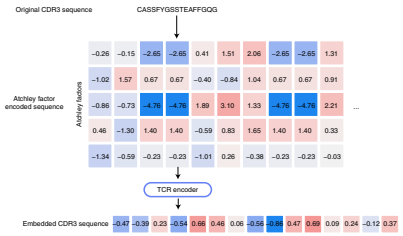
Downstream Analysis

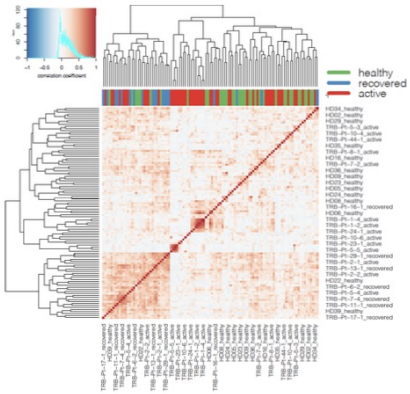
Downstream Analysis





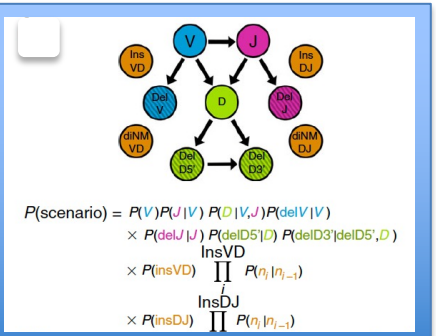
TESSA Software





Bayes Factor Adjusted Pvalue

It probabilistically annotates sequences to evaluate which specific sequences are likely to be generated and found in repertoires.



Pgen (OLGA)

Normalized Frequency

Bayes factor

$$BF_c(j) = \frac{P(M_c|D)/P(M_c)}{P(M_j|D)/P(M_j)}$$

$c \neq j$ and $c, j = 1, \ldots, K$,

$BF_c(1)$
$\vdots$
$BF_c(j)$
$\vdots$
$BF_c(K)$

$X = \log_{10}(BF_c(j)) \geq x_0$

Z = the number of $X \geq x_0$

$X \sim N(0, \sigma^2)$

$Z \sim Bin(K-1,p)$,

$p = 1 - \Phi(x_0)$

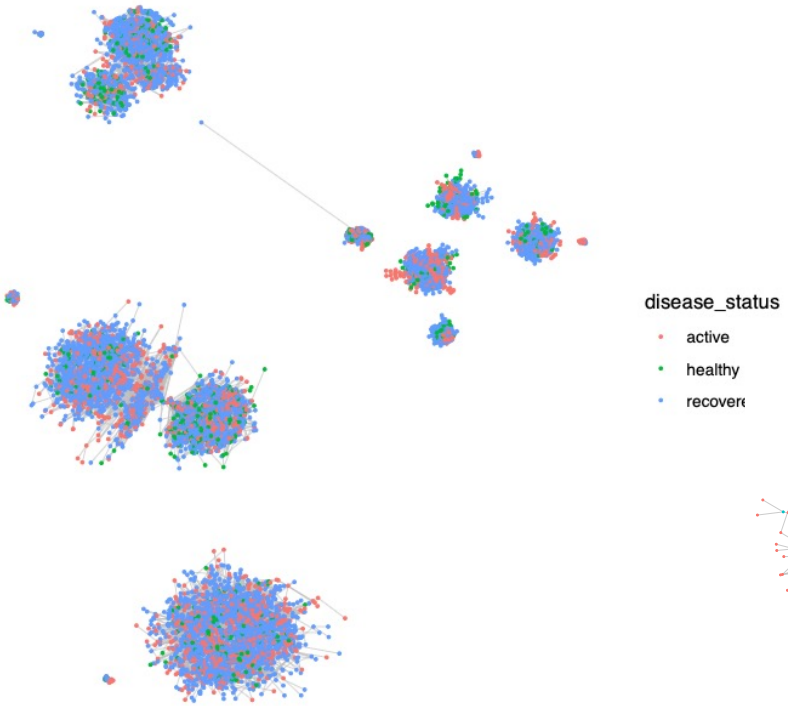
Pseudo-pvalue

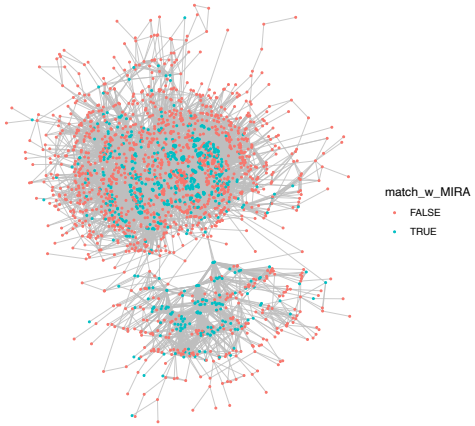
P_{BF}^c (FDR_{BF}^c)

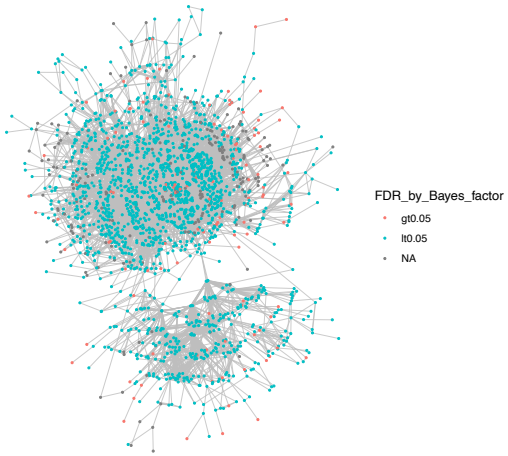
TCR Sequence	Count	Normalized frequency
TGTGCGAGAGGGCGATACTTTGACTACTGG	91	0.00606667
TGTGCGAGACTGGGGGGGCAGTGGCTGGTCAGGGACGGTATGGACGTCTGG	69	0.0046
TGTGCGAGAGAGAAGTCCTACGGTATGGACGTCTGG	66	0.0044
TGTGCGAGAGGAACGGGGAGCAGCTCGAACTACTACATGGACGTCTGG	66	0.0044
TGTGCGACCTTGGGTTGGTCCGAGGGCTGGTTCGACCCCTGG	66	0.0044
TGTGCCAGAGCTTACGGTGACTACGTGGAATACTGG	62	0.00413333
TGTGCGAGAGCGTATAGCAGCTCGTCCCATGTTGACTACTGG	62	0.00413333
TGTGCGAGAGGCCTAGCCGGTTCGCATGGCTACTGG	62	0.00413333
TGTGCGCATGGGTACGGTGACTTCGGGTACTGG	61	0.00406667
TGTGCGAGAGATCAGCCGGGCTGGGGCACGTCGGACTGG	61	0.00406667
TGTGCACGATGCTTACGTCCCCGCCTGGACTACTACTACATGGACGTCTGG	60	0.004
TGTGCGAGAGCGAGCAGCAGCTGGTACGGGAACTGGTTCGACCCCTGG	59	0.00393333

Summary of Public Clusters

Public Cluster ID[1]	No. of TCRs	Motif[2]	No. of HD Samples[3]	No. of Active COVID Samples[4]	No. of Recovered COVID Samples[5]	Estimate (95%CI) Pvalue[6]			Coreness[7] Median [Min,Max]	The % of significant TCRs based on Bayes factor[8]	Correlation of Atchley factor[9] Median [IQR]	The % of TCRs matched with MIRA[10]
						Active COVID vs. HD	Recovered COVID vs. HD	Recovered COVID vs. Active COVID				
1	2092		12	39	19	0.33 (0.02, 0.64) p= 0.039	0.7 (0.38, 1.02) p <0.001	0.37 (0.11, 0.63) p= 0.005	1[1,6]	84.6%	0.37 [0.2,0.53]	28.7%







Conclusion & Discussion

- Used network analysis, other advanced machine learning techniques and statistical approaches, to interrogate and measure immune repertoire architecture in a clinical context.
- Developed customized search algorithms to identify disease associated clones and public shared clones.
- Implemented the proposed methods on different types of datasets that have a wealth of diverse and rich data to demonstrate the flexibility and power of the proposed tools.
- Developed a comprehensive user-friendly bioinformatics tool with visualization to tackle the complexity of the immunosequencing data in a translational fashion.

Future Work

- Incorporate the abundance into network analysis
- Adapt more features for scRNA-seq data
- A lot more…..

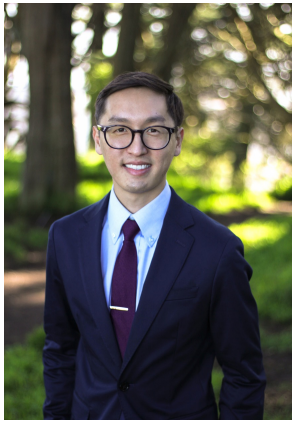
Acknowledgements

- HDFCCC Biostatistics Core
- Division of Hematology/Oncology
- Fong Lab
- NIH/NCI R21R21CA264381 (2021-2023)
- NIH/NLM R01LM013763-01A1 (2022-2026)
- UCSF Prostate Cancer Pilot Award (2021-2022)

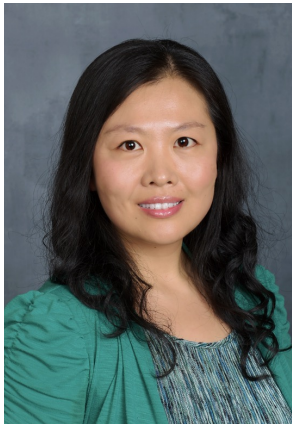
Lawrence Fong, MD
Professor of Medicine, UCSF

Hai Yang, MS
Senior Statistician
Zhang Lab, UCSF

Brian Neal, MS
Student
Zhang Lab, UCSF



Jason Cham, MD
Resident Physician
Scripps Clinic



Tao He, PhD
Associate Professor
SFSU

University of California
San Francisco