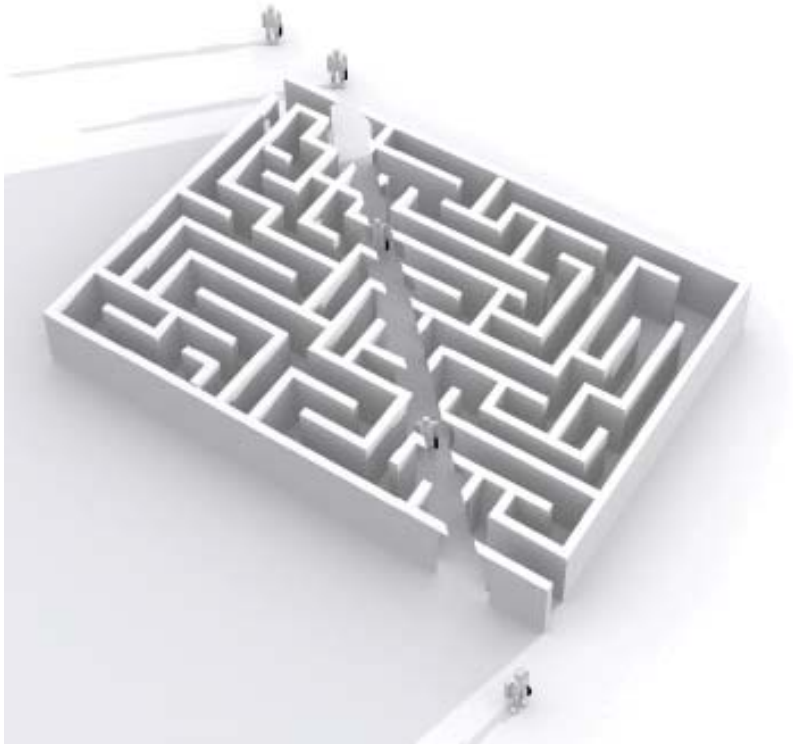
Finding a needle in a haystack: new approaches to identify disease-causing mutations in patients' genomes

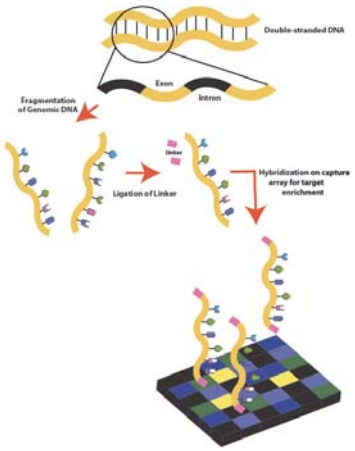
Yuval Itan

St. Giles Laboratory of Human Genetics of Infectious Diseases

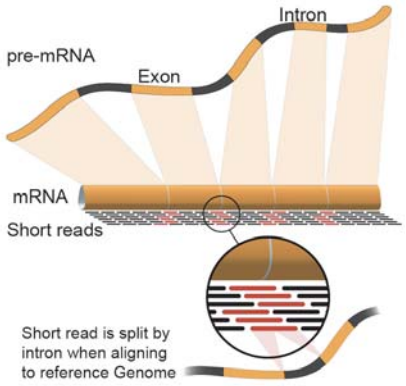


High throughput genomic/proteomic technologies

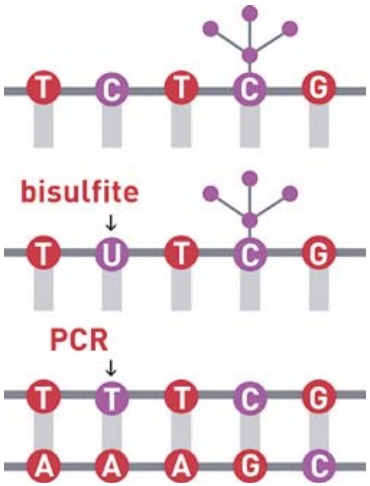
Exome/genome sequencing



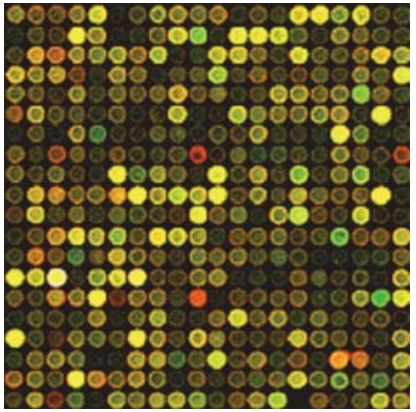
RNA sequencing



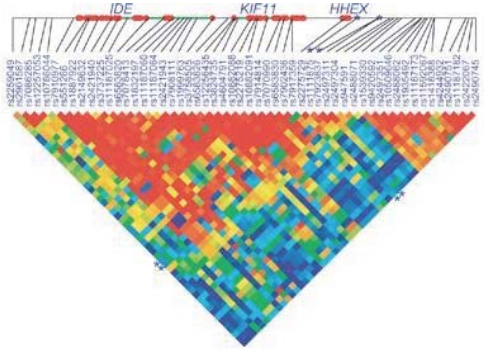
Epigenome sequencing



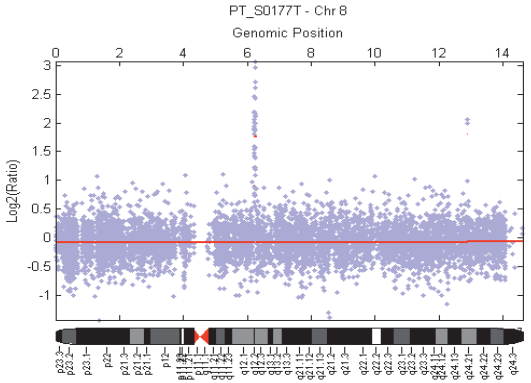
Microarray



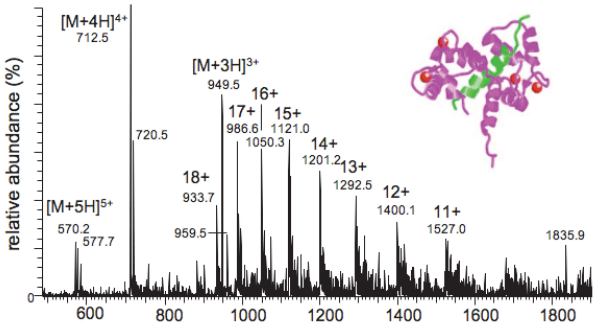
Genome-wide association



Copy number variation

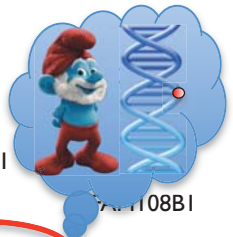


Mass spectrometry



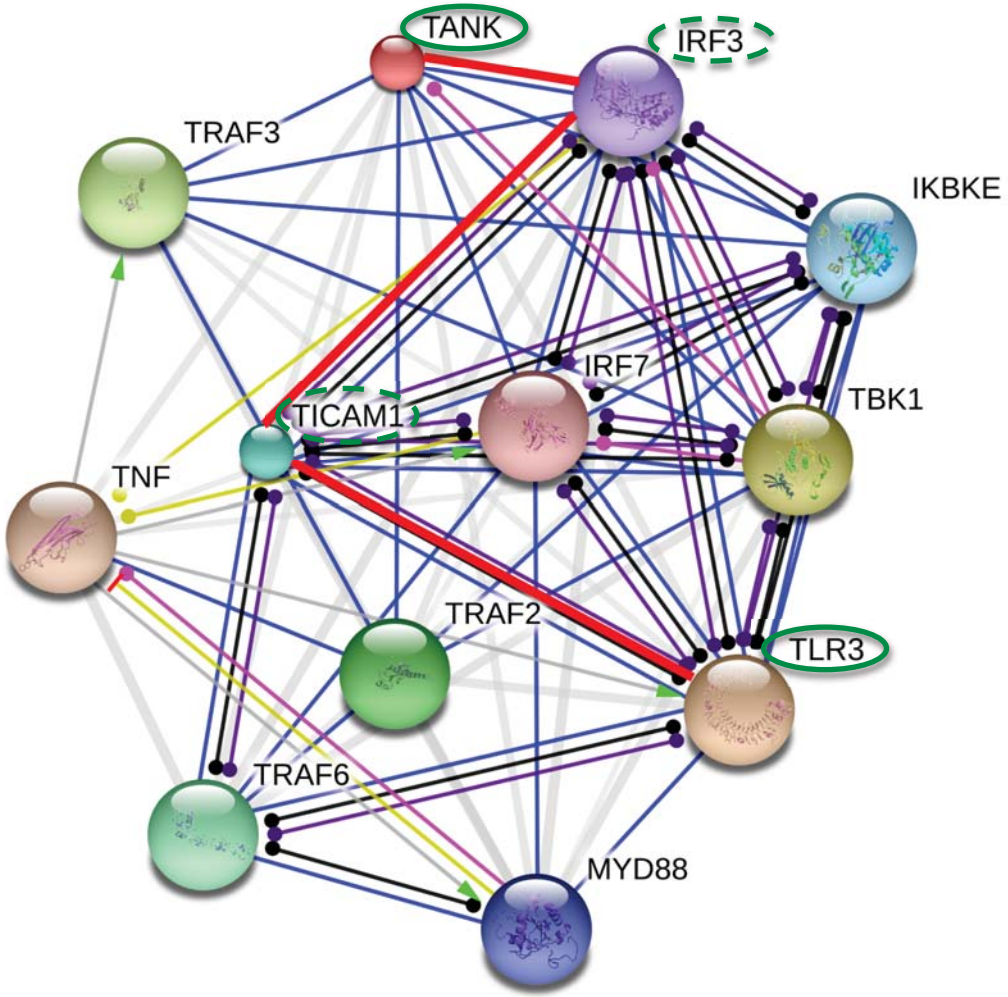
Finding a needle in a haystack: which gene is disease-causing in a patient?

TLL1	SOX30	CCDC99	LCP2	TSPAN17	NOP16	RRM2B	ARHGEF5	MAPK9	COL23A1
OGFR	LTK	HIPK2	CD84	TP53BP1	POLB	FRMPD1	NDST1	AP3M2	CAMK2A
C18orf55	VDAC3	CPB2	CNOT4	UBA5	DUSP12	MRPL22	GEMIN5	NFE2L1	TRPM3
UIMC1	REM1	TPX2	PDRG1	CRLS1	DEFB127	F11	SNX5	RBBP9	DZANK1
AQR	GABRA1	SLAMF7	B4GALT7	ANK1	EFCAB1	RPL26L1	FLT4	MFAP3	MAT2B
RAD51	THOC3	MSMO1	FAM114A2	CYFIP2	C4orf27	TBXAS1	PARP12	PET112	CTDP1
TCOF1	ARHGAP10	HMMR	FCGR2B	STK10	FBXW11	TBX21	CDHR2	SNCB	KCNQ2
KAT6A	GRHL2	TDRD3	NUFIP1	FAT1	IGSF9	AKR1B1	FAT2	RBM22	ADAMTS2
KIF16B	TRMT6	CHGB	MRPS33	NDUFB2	TF	CDV3	TYRO3	WDR76	CAPN3
SNAP23	COL9A3	GABRP	NTSR1	C20orf20	TCFL5	DIDO1	C20orf11	SLC17A9	BIRC7
NKAIN4	ARFGAP1	COL20A1	CHRNA4	EEF1A2	PTK6	GMEB2	TRIB3	CSNK2A1	C20orf54
RPS10L	ANGPT4	HM13	MYLK2	SEC23B	FERMT1	XKR7	HAO1	C20orf160	**CC2D1A**
HCK	TM9SF4	POFUT1	PAK7	KIF3B	ZNF516	ADNP2	DHRS12	OLFM4	RP11-723C11.2
RBFA	RP3-324O17.4	KIAA0226L	HTR2A	FNDC3A	MLNR	CDADC1	CAB39L		INTS6
RPAP1	EHD4	TMEM87A	ZFP106	TGM5	BMF	DNAJC17	EIF3J		RHOV
VPS18	OIP5	SLC30A4	PLDN	IKBKB	PLAT	DKK4	EIF3E		NCALD
UBR5	MCM4	NAMPT	WDR91	MTPN	PTN	ATP6V0A4	ZC3HAV1		SSBP1
EPHB6	CASP2	NOBOX	CHCHD3	CNTNAP3	PIP5K1B	APBA1	SHB	[illegible]108B1	EXOSC3
RGS9	KPNB1	PNPO	CDK5RAP3	CBX1	HOXB6	DDX5	**SMURF2**	KLHL2	CPE
ANXA10	CLCN3	AADAT	GALNT7	TRIM2	FBXW7	UFSP2	FAM149A	SPARC	AKR1D1



Starting point: defining core genes (such as *TLR3* for herpes simplex encephalitis), then looking for genes "closest" to core genes

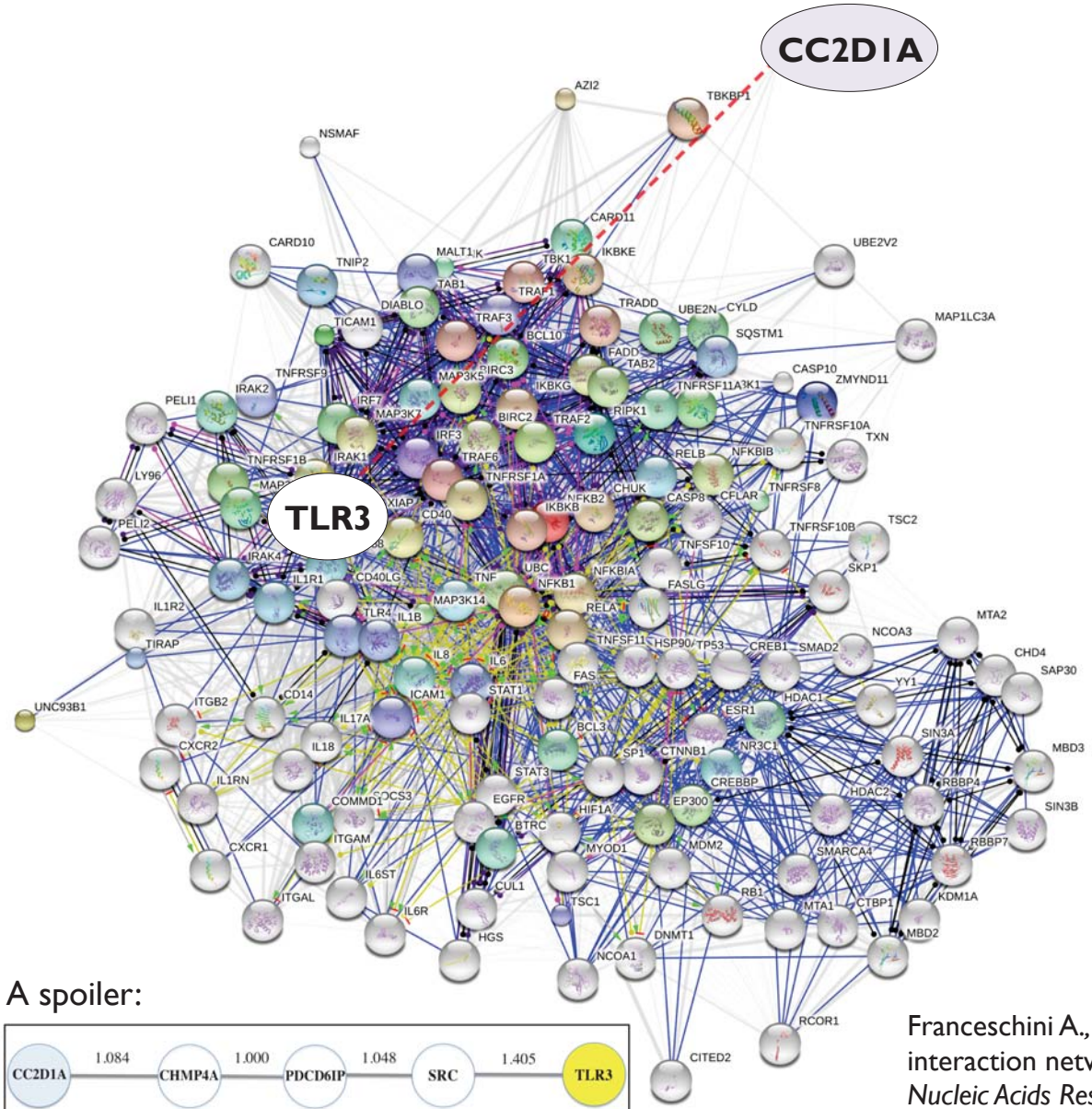
Estimating the relatedness of *TANK* (candidate gene) and *TLR3* (core gene)



- *TANK*: a "low-hanging fruit"
- Finding biologically plausible route from *TANK* to *TLR3*:
  - Hundreds of possible routes
  - Days of String and Pubmed research

Szklarczyk, D., *et al.* (2011) The **STRING** database in 2011: functional interaction networks of proteins, globally integrated and scored, *Nucleic Acids Res*, **39**, D561-568.

Estimating the relatedness of *CC2D1A* (candidate gene) and *TLR3* (core gene)

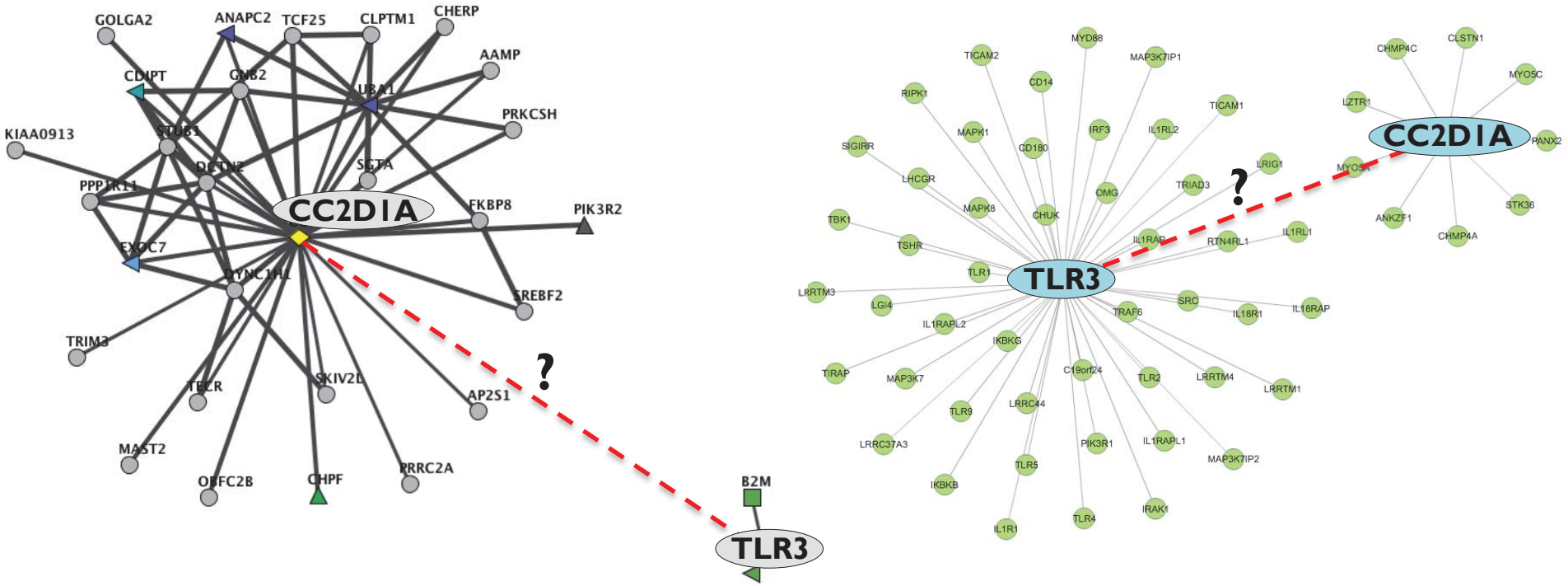


- *CC2D1A*: a *TLR3* pathway gene
- Finding biologically plausible route from *CC2D1A* to *TLR3*:
  - Millions of possible routes
  - Years of String and Pubmed research

A spoiler:

Franceschini A., *et al.* (2011) **STRING** v9.1: protein-protein interaction networks, with increased coverage and integration *Nucleic Acids Res*, D808-15.

Other state-of-the-art tools: direct connections only



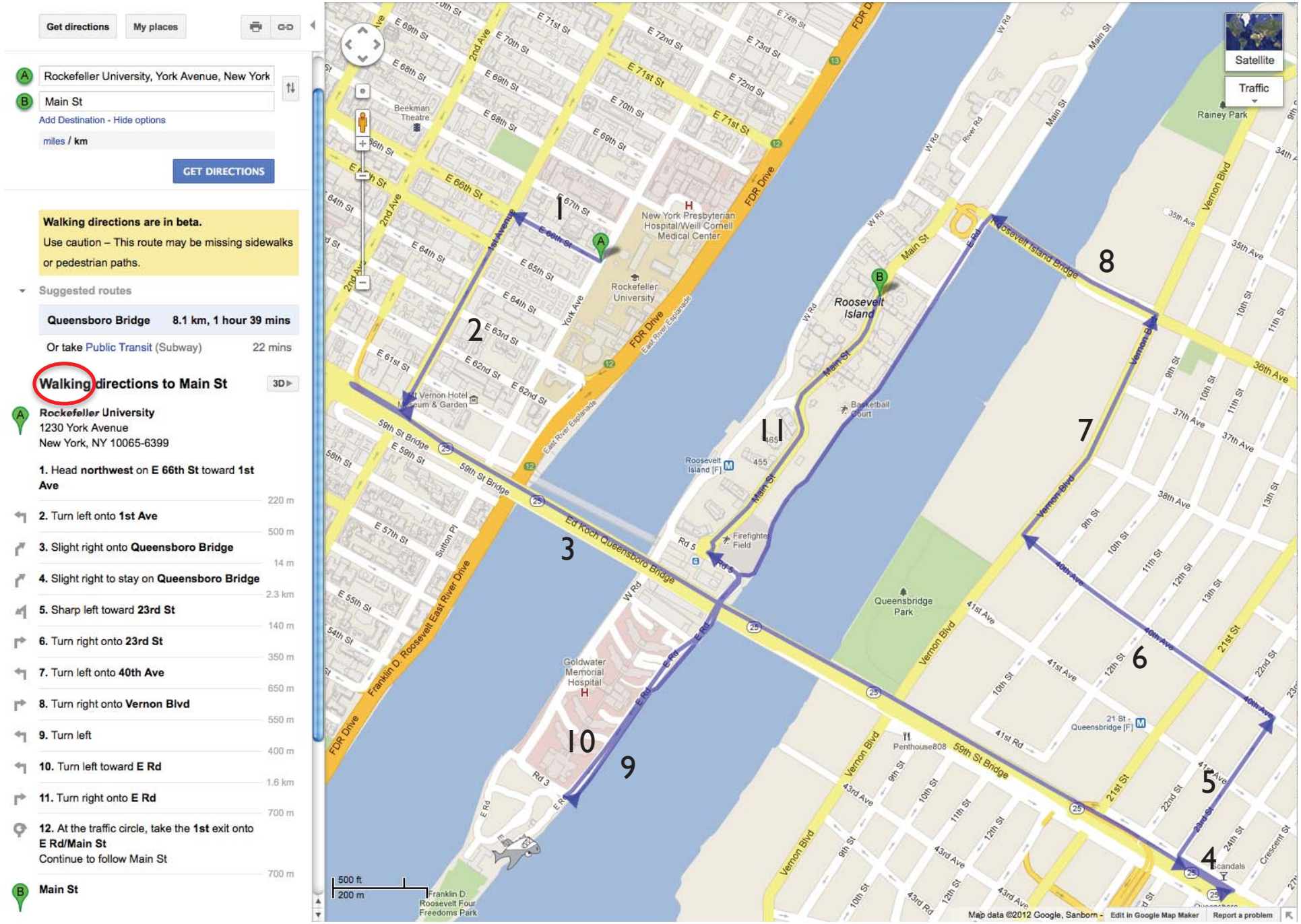
Funcoup

Alexeyenko, A., *et al.* (2012) Comparative interactomics with Funcoup 2.0, Nucleic Acids Res, **40**, D821-D828.

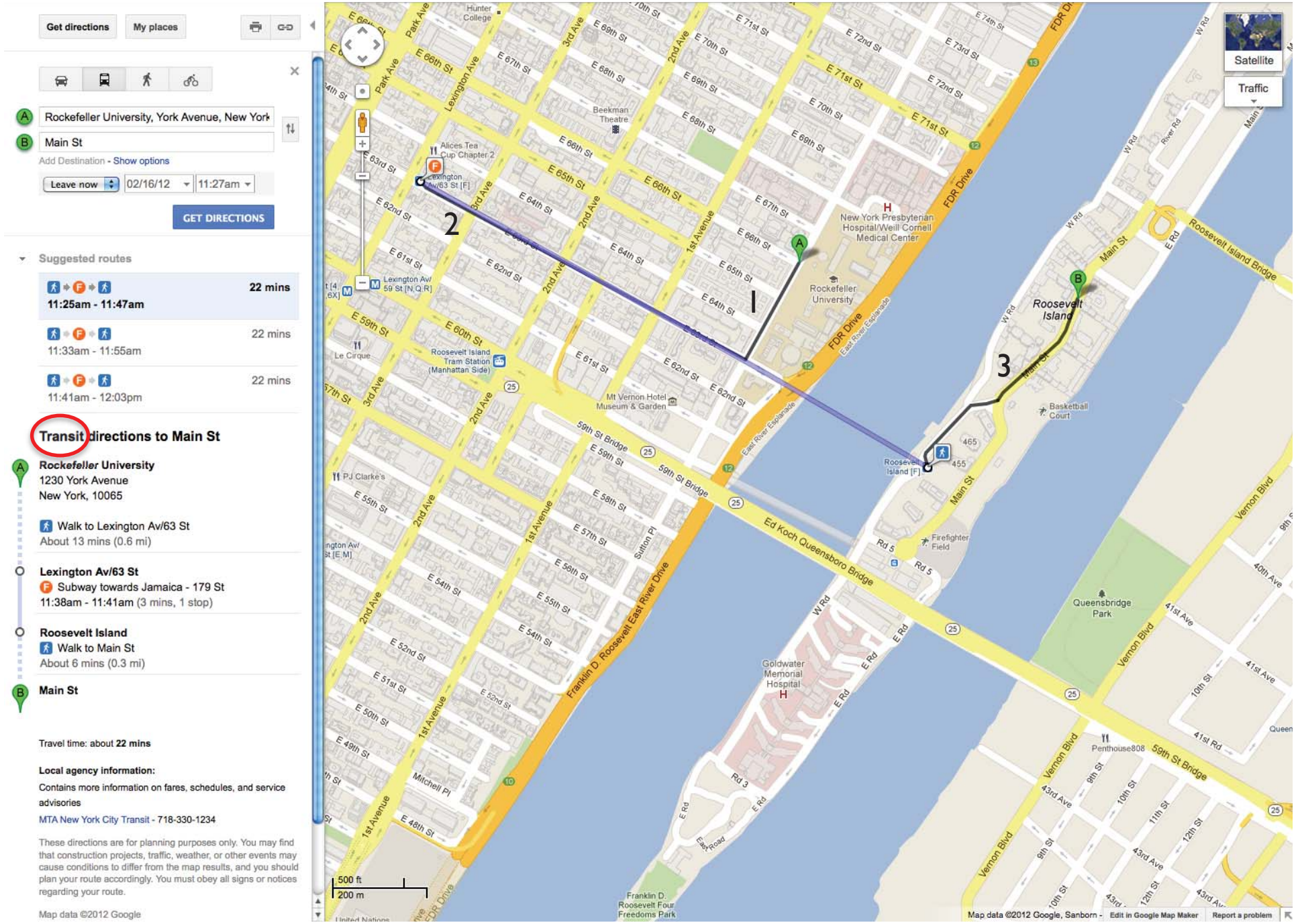
HumanNet

Lee, I., *et* al. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome Res. **7**,1109-1021

Planning a holiday at Roosevelt Island



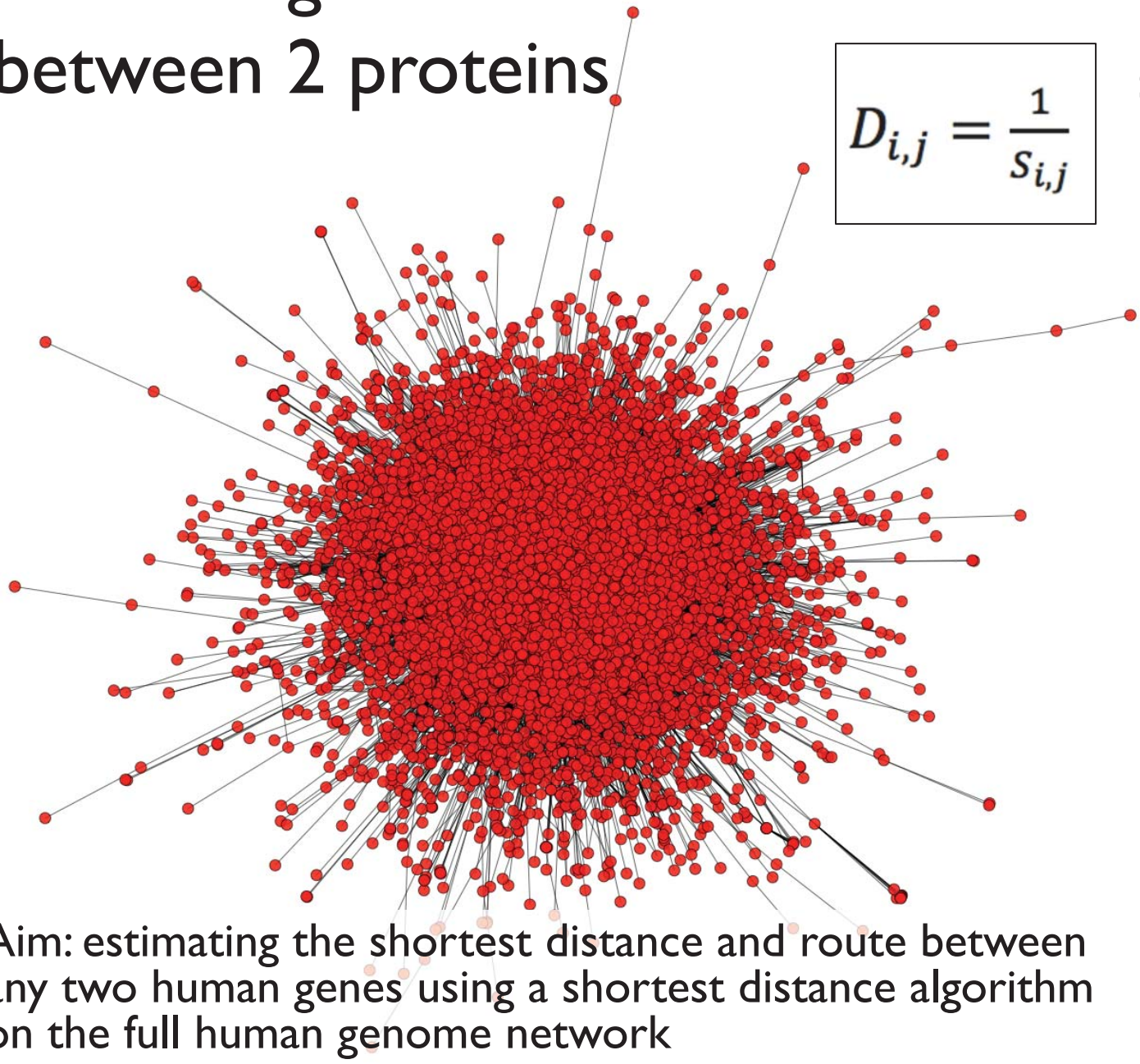
Degrees of separation between A and B = 11



Degrees of separation between A and B = 3

The String database: direct connections between 2 proteins

$$D_{i,j} = \frac{1}{S_{i,j}}$$



Strong connection
→ short direct biological distance

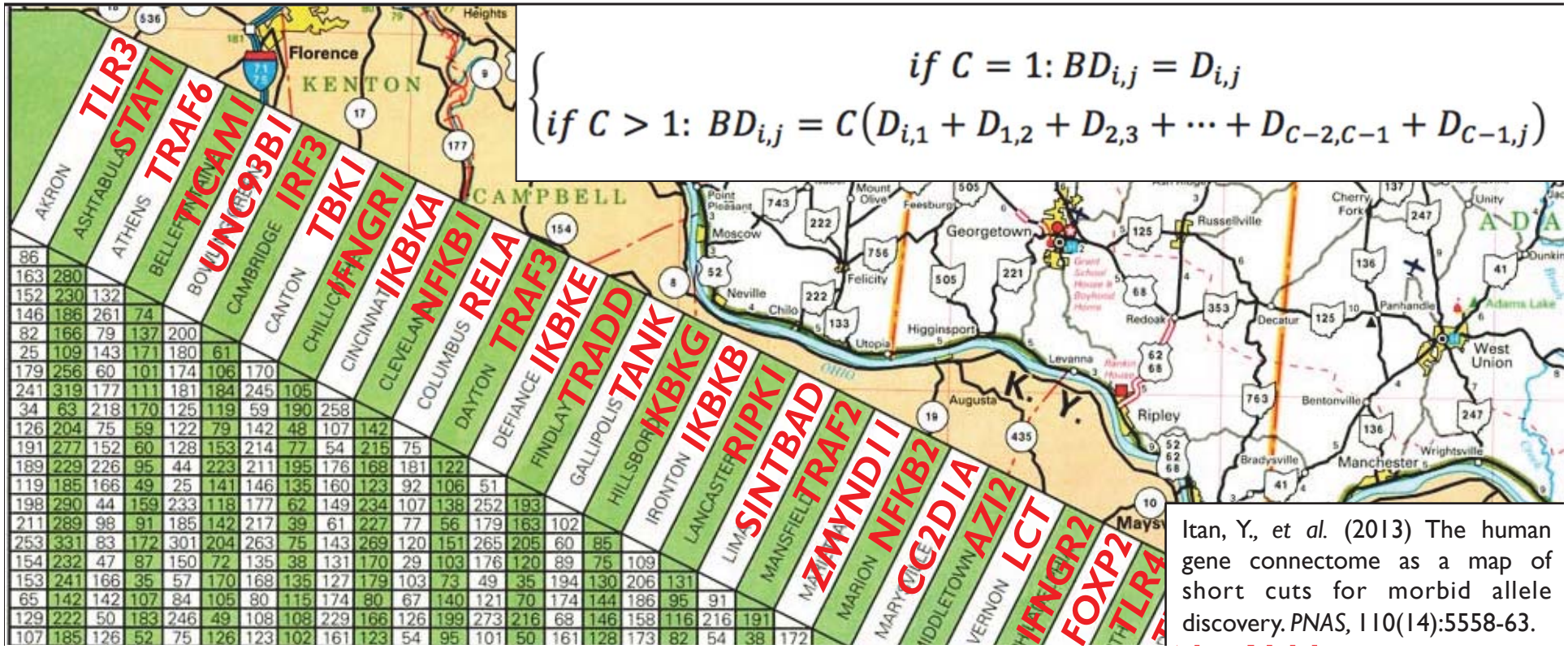
Protein A	Protein B	Confidence Score
ENSP00000000233	ENSP00000254584	18
ENSP00000000233	ENSP00000296557	33
ENSP00000000233	ENSP00000350199	42
ENSP00000000233	ENSP00000355153	51
ENSP00000000233	ENSP00000262305	79
ENSP00000000233	ENSP00000255194	344
ENSP00000000233	ENSP00000204517	776
ENSP00000000412	ENSP00000221957	9
ENSP00000000412	ENSP00000338297	29
ENSP00000000412	ENSP00000236671	37
ENSP00000000412	ENSP00000302665	38
ENSP00000000412	ENSP00000259711	42
ENSP00000000412	ENSP00000341344	68
ENSP00000000412	ENSP00000311962	92
ENSP00000000412	ENSP00000306920	99
ENSP00000000412	ENSP00000245541	178
ENSP00000000412	ENSP00000344401	249
ENSP00000000412	ENSP00000259470	262
ENSP00000000412	ENSP00000361125	304
ENSP00000000412	ENSP00000278412	399
ENSP00000000412	ENSP00000281527	400
ENSP00000000412	ENSP00000295297	400
ENSP00000000412	ENSP00000343348	400
ENSP00000000412	ENSP00000357981	400
ENSP00000000412	ENSP00000217244	421
ENSP00000000412	ENSP00000216297	517
ENSP00000000412	ENSP00000262506	517
ENSP00000000412	ENSP00000351671	517
ENSP00000000412	ENSP00000395546	517
ENSP00000000412	ENSP00000415615	517
ENSP00000000412	ENSP00000164133	754
ENSP00000000412	ENSP00000221138	754
ENSP00000000412	ENSP00000261461	754
ENSP00000000412	ENSP00000261475	754
ENSP00000000412	ENSP00000264977	754
ENSP00000000412	ENSP00000300413	754
ENSP00000000412	ENSP00000311344	754
ENSP00000000412	ENSP00000333905	754
ENSP00000000412	ENSP00000335083	754
ENSP00000000412	ENSP00000336591	754
ENSP00000000412	ENSP00000337641	754
ENSP00000000412	ENSP00000370113	754
ENSP00000000412	ENSP00000375080	754
ENSP00000000412	ENSP00000381100	754
ENSP00000000412	ENSP00000382021	754
ENSP00000000412	ENSP00000417963	754
ENSP00000000412	ENSP00000418447	754
ENSP00000000442	ENSP00000264867	9
ENSP00000000442	ENSP00000254227	21

Aim: estimating the shortest distance and route between any two human genes using a shortest distance algorithm on the full human genome network

The human genes connectome (HGC)

- Calculated on the human genes network using the Dijkstra algorithm. A matrix sized 12,009 X 12,009 X 2
- Shortest biologically plausible **distances** between all human gene pairs
- Shortest biologically plausible **routes** between all human gene pairs

$$\begin{cases} if\ C = 1: BD_{i,j} = D_{i,j} \\ if\ C > 1:\ BD_{i,j} = C\left(D_{i,1} + D_{1,2} + D_{2,3} + \cdots + D_{C-2,C-1} + D_{C-1,j}\right) \end{cases}$$



Itan, Y., *et al.* (2013) The human gene connectome as a map of short cuts for morbid allele discovery. *PNAS*, 110(14):5558-63.

A gene-specific connectome (*TLR3*)

Target gene	Distance from *TLR3*	Rank	*P*-value (percentile)	Median ratio	Average ratio	Sphere	Predicted route to *TLR3*	Degrees of separation with *TLR3*
TLR3	0.000	0	0.000	0.000	0.000	0	TLR3	0
TRIF (TICAM1)	1.001	1	<0.001	0.059	0.050	0	TLR3[1.001]TRIF	1
MYD88	1.002	2	<0.001	0.059	0.050	0	TLR3[1.002]MYD88	1
IRAK2	1.023	3	<0.001	0.060	0.051	0	TLR3[1.023]IRAK2	1
UNC93B1	1.081	4	<0.001	0.064	0.054	0	TLR3[1.081]UNC93B1	1
HMGB1	1.128	5	<0.001	0.066	0.056	0	TLR3[1.128]HMGB1	1
TICAM2	1.134	6	<0.001	0.067	0.056	0	TLR3[1.134]TICAM2	1
LY86	1.187	7	0.001	0.070	0.059	0	TLR3[1.187]LY86	1
RNF216	1.363	8	0.001	0.080	0.068	0	TLR3[1.363]RNF216	1
SRC	**1.405**	**9**	**0.001**	**0.083**	**0.070**	**0**	**TLR3[1.405]SRC**	**1**
PIK3R1	1.615	10	0.001	0.095	0.080	0	TLR3[1.615]PIK3R1	1
IL1B	4.004	11	0.001	0.236	0.199	0	TLR3[1.002]MYD88[1.0]IL1B	2
IL1R1	4.004	12	0.001	0.236	0.199	0	TLR3[1.002]MYD88[1.0]IL1R1	2
IRAK1	4.004	13	0.001	0.236	0.199	1	TLR3[1.002]MYD88[1.0]IRAK1	2
IRAK4	4.004	14	0.001	0.236	0.199	1	TLR3[1.002]MYD88[1.0]IRAK4	2
IRF3	4.004	15	0.001	0.236	0.199	1	TLR3[1.001]TRIF[1.001]IRF3	2
IRF7	**4.004**	**16**	**0.001**	**0.236**	**0.199**	**1**	**TLR3[1.002]MYD88[1.0]IRF7**	**2**
TLR4	4.004	17	0.001	0.236	0.199	1	TLR3[1.002]MYD88[1.0]TLR4	2
IRAK3	4.006	18	0.001	0.236	0.199	1	TLR3[1.002]MYD88[1.001]IRAK3	2
FADD	4.008	19	0.002	0.236	0.199	1	TLR3[1.002]MYD88[1.002]FADD	2
TLR9	4.008	20	0.002	0.236	0.199	1	TLR3[1.002]MYD88[1.002]TLR9	2
IRF5	4.010	21	0.002	0.236	0.199	1	TLR3[1.002]MYD88[1.003]IRF5	2
TBK1	**4.010**	**22**	**0.002**	**0.236**	**0.199**	**1**	**TLR3[1.001]TRIF[1.004]TBK1**	**2**
TLR2	4.012	23	0.002	0.236	0.199	1	TLR3[1.002]MYD88[1.004]TLR2	2

Sort any list of genes by biological proximity to the phenotype's known core gene(s)

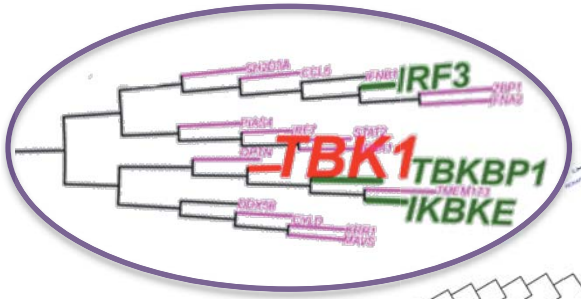
Functional Genomic Alignment (FGA): clustering by biological distance

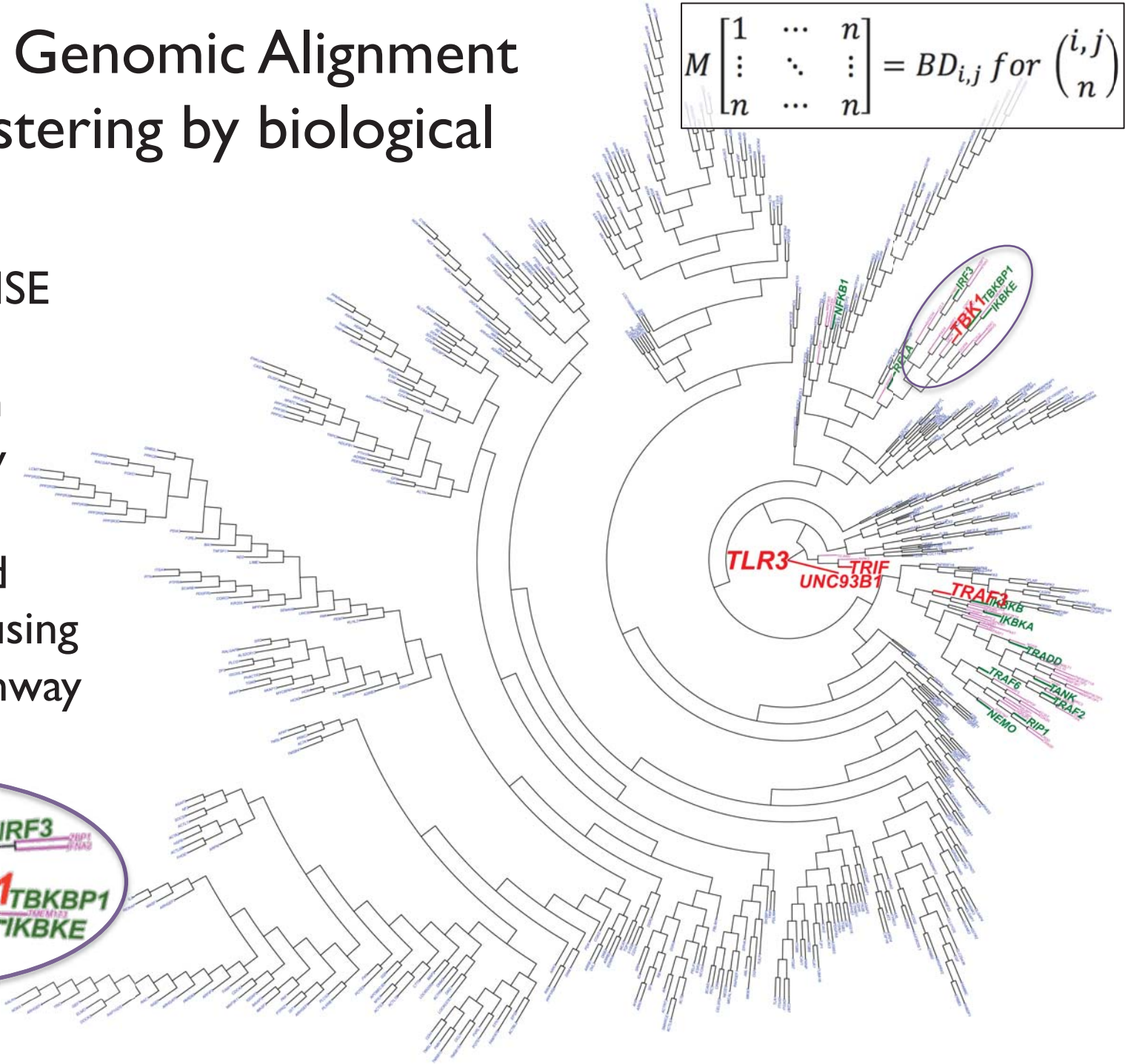
$$M\begin{bmatrix} 1 & \cdots & n \\ \vdots & \ddots & \vdots \\ n & \cdots & n \end{bmatrix} = BD_{i,j}\ for\ \binom{i,j}{n}$$

Red: known HSE causing genes

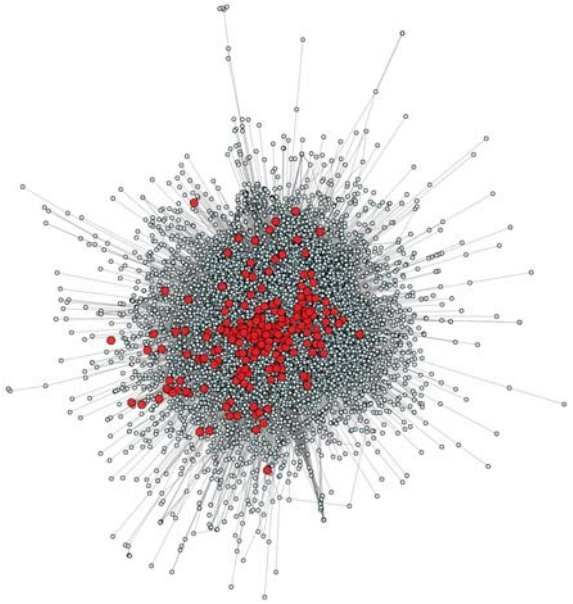
Green: known TLR3 pathway genes

Pink: expected novel HSE causing and TLR3 pathway genes





New experimentally validated disease-causing genes predicted by HGC



Itan, Y. and Casanova, JL. (2015). Novel primary immunodeficiency candidate genes predicted by the human gene connectome. *Front. Immunol*, 6:142.

Predicted PID gene candidate	Known PID gene candidate	Biological Distance between Candidate and Known	Rank of candidate in known	P-value (percentile) of candidate in known	Route between Candidate and Known
BCL10	CARD11	1.001	1	0.00007	CARD11←→BCL10
IRF7	MYD88	1.001	1	0.00007	MYD88←→IRF7
IL21	IL21R	1.001	1	0.00007	IL21R←→IL21
CTLA4	ICOS	1.616	3	0.00021	ICOS←→CTLA4
STING (TMEM173)	TBK1	1.001	3	0.00021	TBK1←→TMEM173
NFKB1	NFKBIA	1.001	4	0.00028	NFKBIA←→NFKB1
NLRC4	NOD2	1.183	5	0.00035	NOD2←→NLRC4
NIK (MAP3K14)	CD40	1.064	8	0.00057	CD40←→MAP3K1
TPP1	TINF2	1.191	18	0.00127	TINF2←→TPP1
JAGN1 (JAG1)	CHD7	2.38663	27	0.00191	CHD7←→JAG1
TGFBR2	C8B	6.44124	27	0.00191	C8B←→CLU←→TGFBR2
TGFBR1	C8B	6.44124	28	0.00198	C8B←→CLU←→TGFBR1
INO80	ACTB	1.582	35	0.00248	ACTB←→INO80
DOCK2	RAC2	1.61	35	0.00248	RAC2←→DOCK2
STAT4	STAT3	1.10375	35	0.00248	STAT3←→STAT4
ADA2 (TADA2A)	TBX1	7.228	83	0.00587	TBX1←→C11orf30←→TADA2A
IFIH1	FADD	4.141	87	0.00616	FADD←→MAVS←→IFIH1

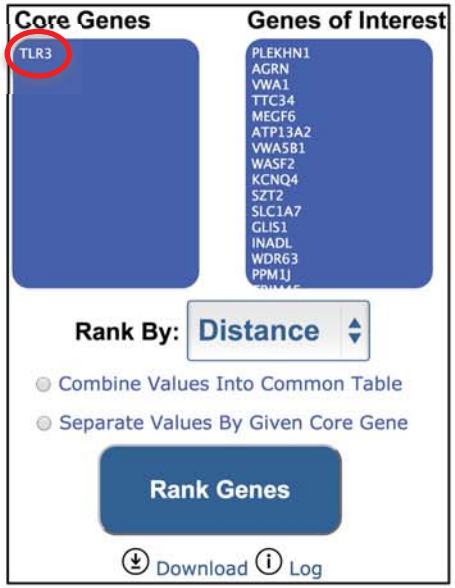
How many experimental biologists do we lose when having only a command-line option?

- A small survey: losing 4 out of 5 potential users due to command line terminal phobia

The human gene connectome server (HGCS): an easy to use online interface for prioritizing genes by biological distance. http://hgc.rockefeller.edu/

- **Instructions:**
  1. Put core genes in left box
  2. Put genes of interest in right box
  3. Press "rank genes" button

A



B

Core Gene	Target	Distance	Rank	P-Value (percentile)	BRP	Median Ratio	Average Ratio	Sphere	Route	Degrees Separation	Full Gene Name
TLR3	TRAF3	1.75747	23	0.00163	0.00163	0.10984	0.10553	1	TLR3[1.75747]TRAF3	1	TNF receptor-associated factor 3
TLR3	RBCK1	5.25832	280	0.01981	0.01981	0.32865	0.31576	2	TLR3[1.01626]TRAF6[1.6129]RBCK1	2	RanBP-type and C3HC4-type zinc finger containing 1
TLR3	IRF4	5.72062	491	0.03475	0.01776	0.35754	0.34352	2	TLR3[1.61031]MYD88[1.25]IRF4	2	interferon regulatory factor 4
TLR3	CD3E	6.44124	628	0.04444	0.04444	0.40258	0.38679	2	TLR3[1.61031]PIK3R1[1.61031]CD3E	2	CD3e molecule, epsilon (CD3-TCR complex)
TLR3	CD5	6.44124	637	0.04508	0.03177	0.40258	0.38679	2	TLR3[1.61031]PIK3R1[1.61031]CD5	2	CD5 molecule
TLR3	ACTN1	6.44642	640	0.04529	0.04529	0.4029	0.3871	2	TLR3[1.61031]SRC[1.6129]ACTN1	2	actinin, alpha 1
TLR3	CDC42BPG	6.71494	695	0.04918	0.01387	0.41968	0.40322	2	TLR3[1.6]TBK1[1.75747]CDC42BPG	2	CDC42 binding protein kinase gamma (DMPK-like)
TLR3	RAD18	9.08214	829	0.05867	0.05867	0.56763	0.54537	3	TLR3[1.01626]TRAF6[1.0]UBC[1.01112]RAD18	3	RAD18 homolog (S. cerevisiae)
TLR3	POLH	9.09447	844	0.05973	0.05973	0.5684	0.54611	3	TLR3[1.01626]TRAF6[1.0]UBC[1.01523]POLH	3	polymerase (DNA directed), eta
TLR3	FANCI	9.10689	854	0.06043	0.06043	0.56918	0.54686	3	TLR3[1.01626]TRAF6[1.0]UBC[1.01937]FANCI	3	Fanconi anemia, complementation group I

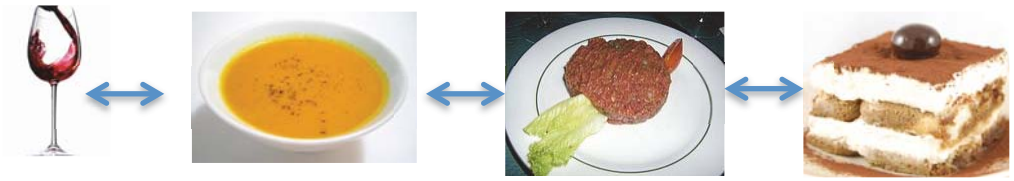
Itan, Y., *et al.* (2014) HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics*, 15:256.

But… What if there isn't a known core gene, and we have a cohort of **30 patients** having the same Mendelian disease, **each has 1000 genes with variants**, assuming some genetic heterogeneity?

- **Hypothesis**: given properly filtered lists of genes, the biologically smallest cluster of genes in a cohort of patients includes the disease causing gene for each of these patients

What is the best possible meal (including wine, 1st course, main and dessert) in a restaurant?

- Wine list (patient #1): red, white, whiskey
- 1st course list (patient #2): salad, soup
- Main course list (patient #3): steak, chicken
- Dessert (patient #4): cake, tiramisu, cognac



36 possible combinations

```
1   ['white', 'salad', 'steak', 'cake']
2   ['white', 'salad', 'steak', 'tiramisu']
3   ['white', 'salad', 'steak', 'congnac']
4   ['white', 'salad', 'chicken', 'cake']
5   ['white', 'salad', 'chicken', 'tiramisu']
6   ['white', 'salad', 'chicken', 'congnac']
7   ['white', 'soup', 'steak', 'cake']
8   ['white', 'soup', 'steak', 'tiramisu']
9   ['white', 'soup', 'steak', 'congnac']
10  ['white', 'soup', 'chicken', 'cake']
11  ['white', 'soup', 'chicken', 'tiramisu']
12  ['white', 'soup', 'chicken', 'congnac']
13  ['red', 'salad', 'steak', 'cake']
14  ['red', 'salad', 'steak', 'tiramisu']
15  ['red', 'salad', 'steak', 'congnac']
16  ['red', 'salad', 'chicken', 'cake']
17  ['red', 'salad', 'chicken', 'tiramisu']
18  ['red', 'salad', 'chicken', 'congnac']
19  ['red', 'soup', 'steak', 'cake']
20  ['red', 'soup', 'steak', 'tiramisu']
21  ['red', 'soup', 'steak', 'congnac']
22  ['red', 'soup', 'chicken', 'cake']
23  ['red', 'soup', 'chicken', 'tiramisu']
24  ['red', 'soup', 'chicken', 'congnac']
25  ['whiskey', 'salad', 'steak', 'cake']
26  ['whiskey', 'salad', 'steak', 'tiramisu']
27  ['whiskey', 'salad', 'steak', 'congnac']
28  ['whiskey', 'salad', 'chicken', 'cake']
29  ['whiskey', 'salad', 'chicken', 'tiramisu']
30  ['whiskey', 'salad', 'chicken', 'congnac']
31  ['whiskey', 'soup', 'steak', 'cake']
32  ['whiskey', 'soup', 'steak', 'tiramisu']
33  ['whiskey', 'soup', 'steak', 'congnac']
34  ['whiskey', 'soup', 'chicken', 'cake']
35  ['whiskey', 'soup', 'chicken', 'tiramisu']
36  ['whiskey', 'soup', 'chicken', 'congnac']
```

For some customers (diseases) the best meal (smallest cluster) would be a different order of courses (genes)

```
467    ('steak', 'cake', 'red', 'soup')
468    ('steak', 'cake', 'soup', 'red')
469    ('cake', 'red', 'soup', 'steak')
470    ('cake', 'red', 'steak', 'soup')
471    ('cake', 'soup', 'red', 'steak')
472    ('cake', 'soup', 'steak', 'red')
473    ('cake', 'steak', 'red', 'soup')
474    ('cake', 'steak', 'soup', 'red')
476    ['red', 'soup', 'steak', 'tiramisu']
476    ('red', 'soup', 'steak', 'tiramisu')
477    ('red', 'soup', 'tiramisu', 'steak')
478    ('red', 'steak', 'soup', 'tiramisu')
479    ('red', 'steak', 'tiramisu', 'soup')
480    ('red', 'tiramisu', 'soup', 'steak')
481    ('red', 'tiramisu', 'steak', 'soup')
482    ('soup', 'red', 'steak', 'tiramisu')
483    ('soup', 'red', 'tiramisu', 'steak')
484    ('soup', 'steak', 'red', 'tiramisu')
485    ('soup', 'steak', 'tiramisu', 'red')
486    ('soup', 'tiramisu', 'red', 'steak')
487    ('soup', 'tiramisu', 'steak', 'red')
488    ('steak', 'red', 'soup', 'tiramisu')
489    ('steak', 'red', 'tiramisu', 'soup')
490    ('steak', 'soup', 'red', 'tiramisu')
491    ('steak', 'soup', 'tiramisu', 'red')
492    ('steak', 'tiramisu', 'red', 'soup')
493    ('steak', 'tiramisu', 'soup', 'red')
494    ('tiramisu', 'red', 'soup', 'steak')
495    ('tiramisu', 'red', 'steak', 'soup')
496    ('tiramisu', 'soup', 'red', 'steak')
497    ('tiramisu', 'soup', 'steak', 'red')
498    ('tiramisu', 'steak', 'red', 'soup')
499    ('tiramisu', 'steak', 'soup', 'red')
501    ['red', 'soup', 'steak', 'congnac']
501    ('red', 'soup', 'steak', 'congnac')
502    ('red', 'soup', 'congnac', 'steak')
503    ('red', 'steak', 'soup', 'congnac')
```

899 possible combinations

The number of possible clusters of one gene per patient

- Number of patients P=30
- Number of polymorphic genes per patient N=1000

~= 1000x1000x…x1000x(1x2x…x30) = 1000^{30}x30! =
~= **$2.6x10^{122}$ → NP-complete**

Number of atoms in the universe: 10^{80}
Amount of information contained in the universe: 10^{90}

Amount of time to calculate all possible clusters if all the computers in the world are used: $8.24x10^{97}$ years

Conclusion: a medium cohort of patients contains more information than the universe

Expanding Core Mendelian Clustering (ECMC) of HSE patients' gene variants – a hypothetical example

Patient 1	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**
TBK1	*UBA5*	*TRIF*	*OLFM4*	*TTN*
PAK7	*LTK*	*NOP16*	*MAPK9*	*MUC3*
HMI3	*SMURF2*	*X11*	*FAT2*	*TRPM3*
PNPO	*TYRO3*	*XKR7*	*CONT4*	*TLR3*
LCP2	*TRAF3*	*SHB*	*UNC93B1*	*RHOV*

- The biological distance between all human genes is known and pre-calculated
- Starting with a temporary core gene (*TBK1*) – gene #1 in patient 1

Expanding Core Mendelian Clustering (ECMC) of HSE filtered genes – a hypothetical example

Patient 1	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**
TBK1	*UBA5*	*TRIF*	*OLFM4*	*TTN*
PAK7	*LTK*	*NOP16*	*MAPK9*	*MUC3*
HMI3	*SMURF2*	*X11*	*FAT2*	*TRPM3*
PNPO	*TYRO3*	*XKR7*	*CONT4*	*TLR3*
LCP2	*TRAF3*	*SHB*	*UNC93B1*	*RHOV*

- **Q:** Which gene in all patients other than patient 1 is biologically closest to *TBK1*?
  - ***A:*** *TRAF3*

Expanding Core Mendelian Clustering (ECMC) of HSE filtered genes – a hypothetical example

Patient 1	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**
TBK1	*UBA5*	*TRIF*	*OLFM4*	*TTN*
PAK7	*LTK*	*NOP16*	*MAPK9*	*MUC3*
HMI3	*SMURF2*	*X11*	*FAT2*	*TRPM3*
PNPO	*TYRO3*	*XKR7*	*CONT4* →	*TLR3*
LCP2	*TRAF3*	*SHB*	*UNC93B1*	*RHOV*

- **Q:** Which gene in all patients other than patients 1 and 2 is biologically closest to *TBK1* and *TRAF3*?
  - ***A:*** *TLR3*

Expanding Core Mendelian Clustering (ECMC) of HSE filtered genes – a hypothetical example

Patient 1	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**
TBK1	*UBA5*	*TRIF*	*OLFM4*	*TTN*
PAK7	*LTK*	*NOP16*	*MAPK9*	*MUC3*
HMI3	*SMURF2*	*X11*	*FAT2*	*TRPM3*
PNPO	*TYRO3*	*XKR7*	*CONT4*	*TLR3*
LCP2	*TRAF3*	*SHB*	*UNC93B1*	*RHOV*

- **Q:** Which gene in all patients other than patients 1, 2 and 5 is biologically closest to *TBK1*, *TRAF3* and *TLR3*?
  - ***A:*** *UNC93B1*

Expanding Core Mendelian Clustering (ECMC) of HSE filtered genes – a hypothetical example

Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
TBK1	*UBA5*	*TRIF*	*OLFM4*	*TTN*
PAK7	*LTK*	*NOP16*	*MAPK9*	*MUC3*
HMI3	*SMURF2*	*X11*	*FAT2*	*TRPM3*
PNPO	*TYRO3*	*XKR7*	*CONT4*	*TLR3*
LCP2	*TRAF3*	*SHB*	*UNC93B1*	*RHOV*

- **Q:** Which gene in all patients other than patients 1, 2, 4 and 5 is biologically closest to *TBK1*, *TRAF3*, *TLR3* and *UNC93B1*?
  - ***A:*** *TRIF*

Moving on to gene #2 in patient 1 as the temporary core gene

Patient 1	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**
TBK1	*UBA5*	*TRIF*	*OLFM4*	*TTN*
PAK7	*LTK*	*NOP16*	*MAPK9*	*MUC3*
HMI3	*SMURF2*	*X11*	*FAT2*	*TRPM3*
PNPO	*TYRO3*	*XKR7*	*CONT4*	*TLR3*
LCP2	*TRAF3*	*SHB*	*UNC93B1*	*RHOV*

- **Q:** Which gene in all patients other than patient 1 is biologically closest to *PAK7*?
  - …

After rigorous conventional filtering: testing with 11 HSE patients with known HSE-causing genes and 39 HSE patients with unknown etiology. Top clusters polluted by highly mutated genes

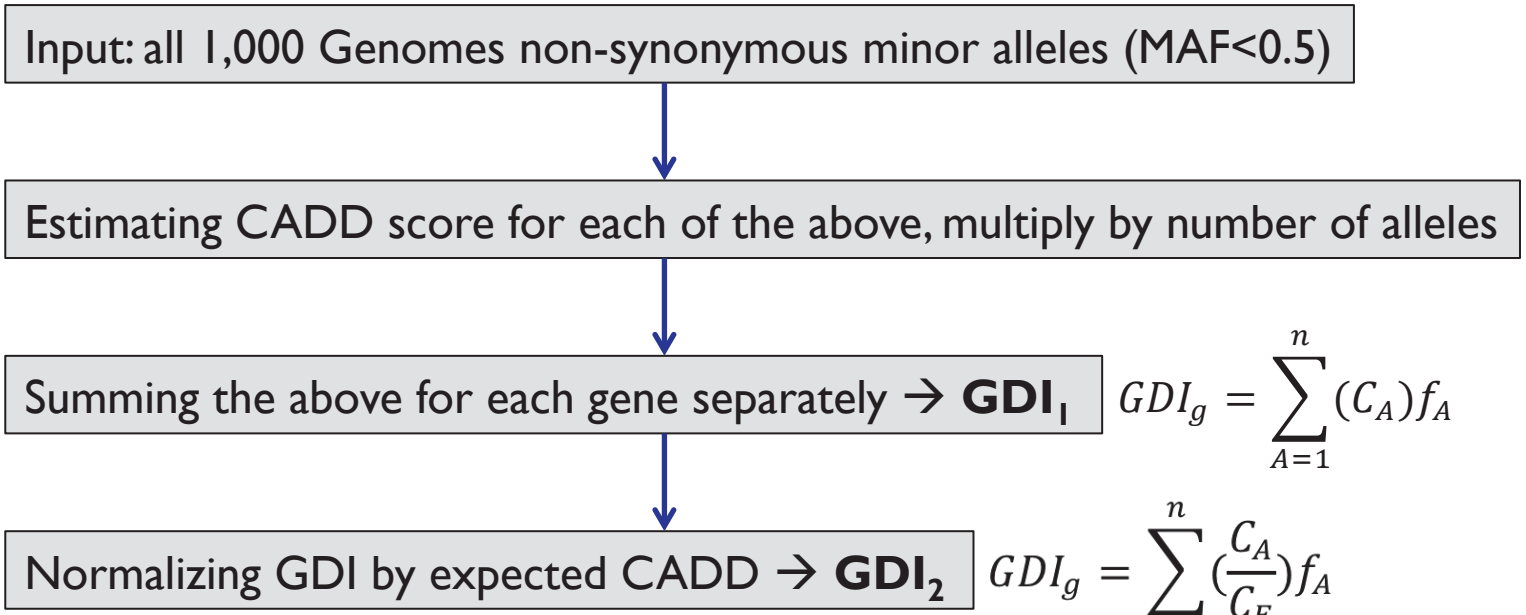
Core_gene	Patient1	Patient2	Patient3	• • •														
TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	OBSL1	TRIM55	TTN	TTN	TTN	TTN	NEB	PDLIM
MYH3	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
MYBPC2	TTN	MYBPC2	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
MYOM2	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	MYOM2	MICAL
OBSL1	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	OBSL1	TRIM55	TTN	TTN	TTN	TTN	NEB	PDLIM
TPM1	TTN	TPM1	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
NEB	TTN	TTN	TTN	TTN	TTN	NEB	TTN	TTN	NEB	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
OR4S2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR4S2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR4S2	OR4S2	OR51B4	OR2AG2	OR52M1	OR6K
ACTN1	ACTN1	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	ACTN1	OBSL1	MAGI2	TTN	TTN	TTN	TTN	NEB	IQGA
OR4C15	OR4C15	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR3A1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR3A1	OR51B4	OR3A1	OR52M1	OR6K
OBSCN	TTN	TTN	OBSCN	TTN	OBSCN	OBSCN	TTN	TTN	TTN	OBSCN	ANK1	SPTA1	OBSCN	OBSCN	TTN	TTN	MYOM2	SPTBN
OR4C3	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR7A5	OR7A5	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR56B1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4B1	OR4C3	OR4B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR2AJ1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR52N2	OR4C3	OR56B1	OR52N2	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR1N2	OR4C3	OR56B1	OR1N2	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR14A2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR51D1	OR4C3	OR56B1	OR2AJ1	OR51D1	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5D13	OR4C3	OR56B1	OR2AJ1	OR14A2	OR5D13	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR6C74	OR4C3	OR56B1	OR2AJ1	OR14A2	OR6C74	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10H2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR10H2	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5P2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR2Z1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR2Z1	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10G8	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10P1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10P1	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR1Q1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR1Q1	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5M8	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR51A7	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR52W1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR52W1	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR9K2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR9K2	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4K17	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR4K17	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10G3	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR10G3	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5H2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR5H2	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR52N4	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4N2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR4N2	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4K1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR4K1	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR6K3	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K

Patients' WES variants data "polluted" by false positives: genes that are highly polymorphic in the general population and unlikely to be disease-causing

Word size represents the number of patients' WES filtered rare variants in the gene

"Data don't make any sense, we will have to resort to statistics."

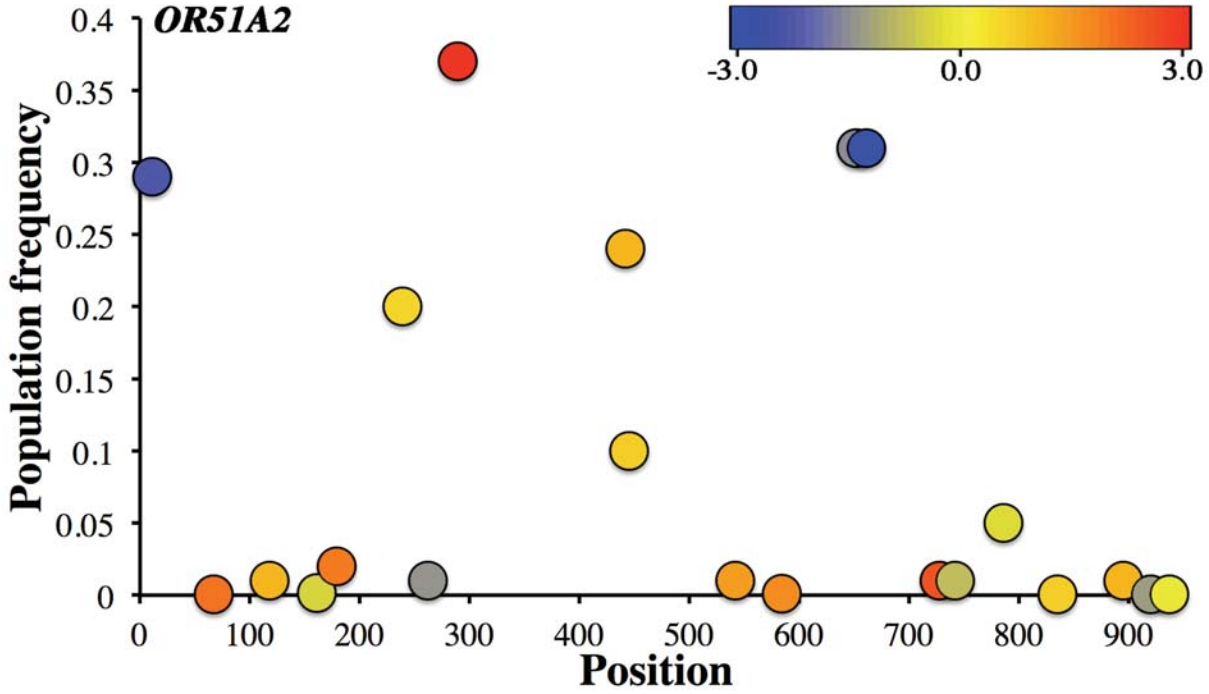
The gene damage index (GDI): accumulated mutational damage of each human gene in the general population



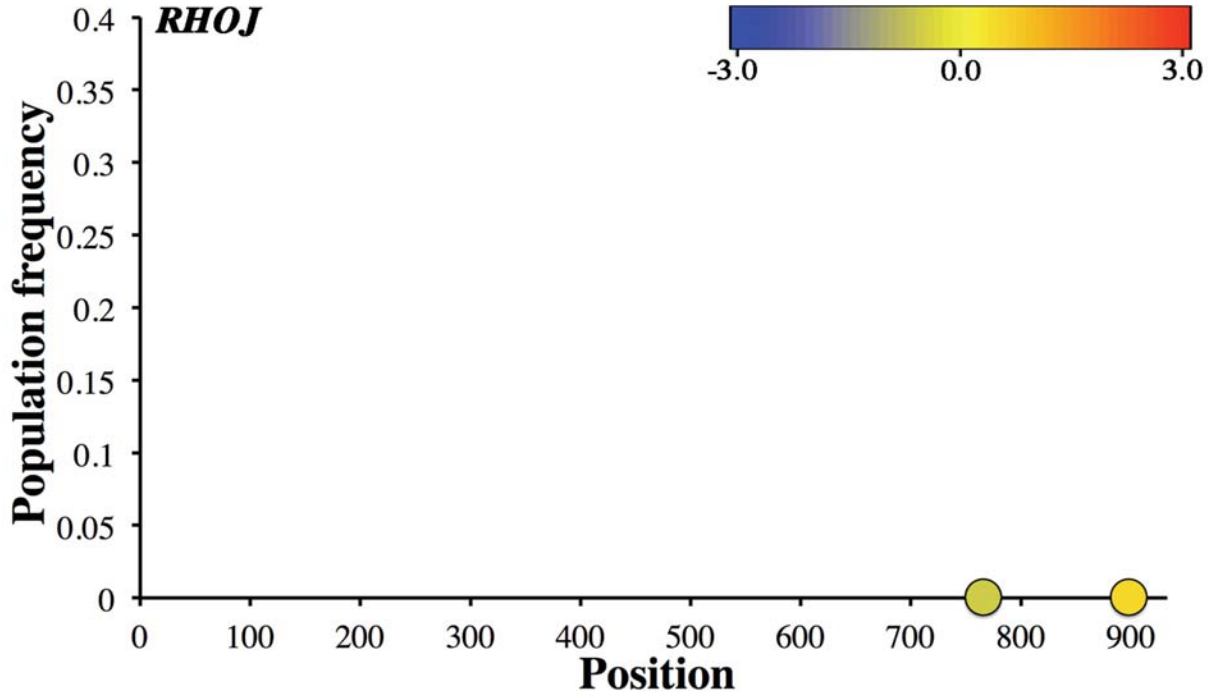
$$GDI_g = \sum_{A=1}^{n} (C_A) f_A$$

$$GDI_g = \sum_{A=1}^{n} (\frac{C_A}{C_E}) f_A$$

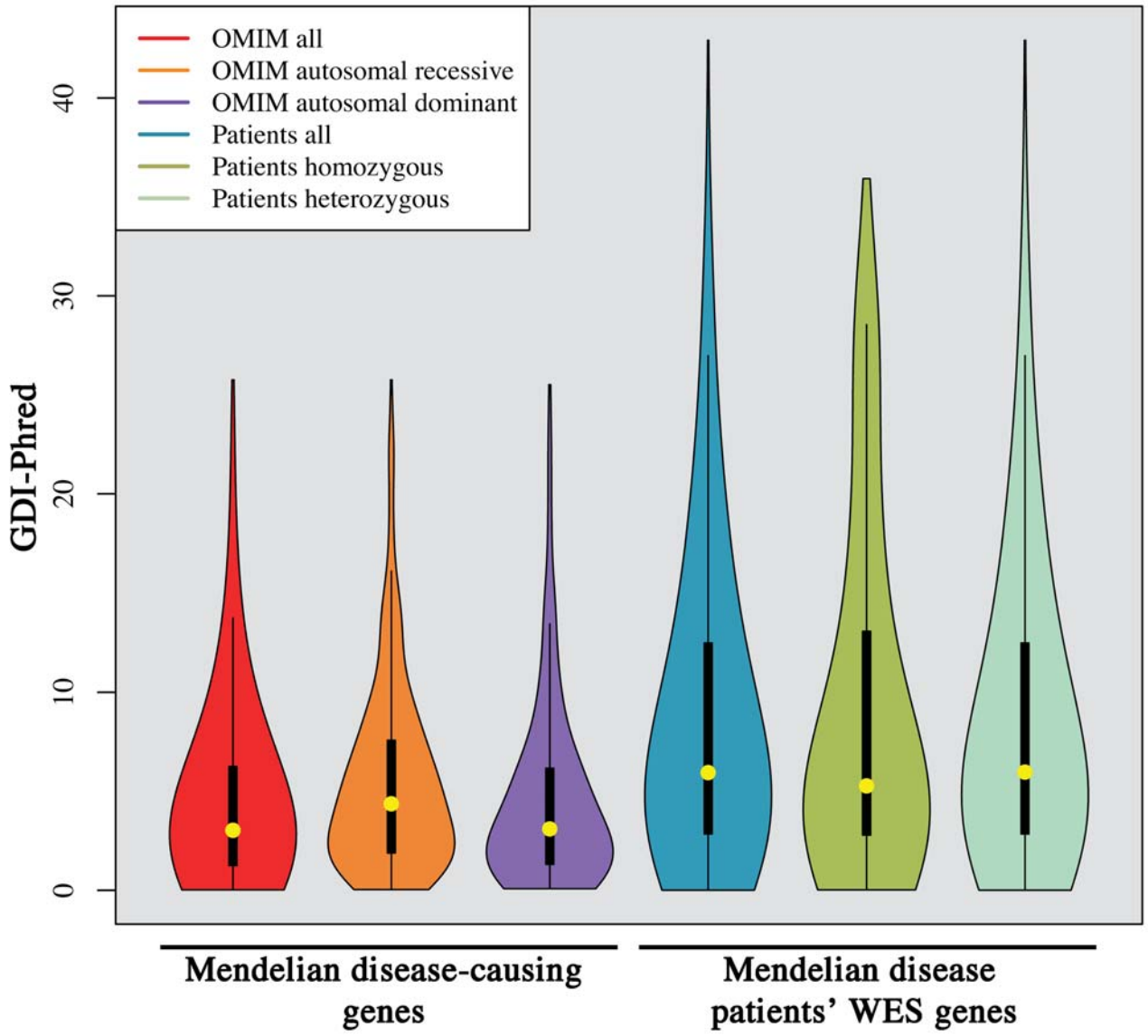
A highly damaged human gene:



A lowly damaged human gene:

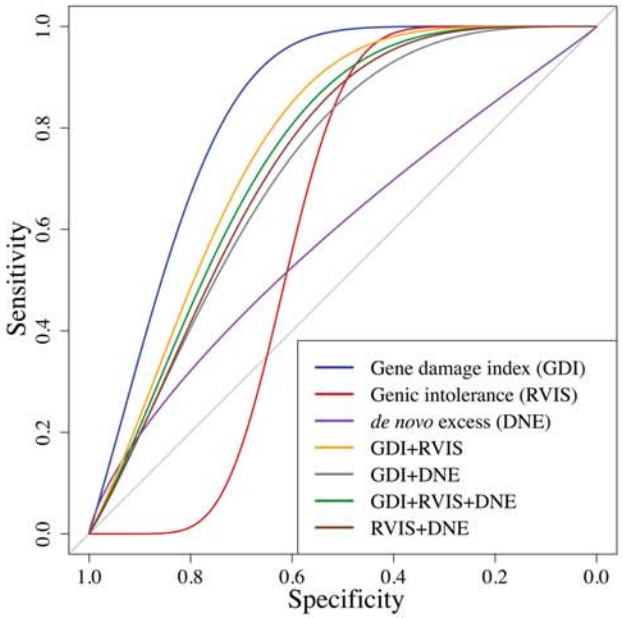


GDI difference between disease-causing genes and corresponding patients' WES data of filtered variants

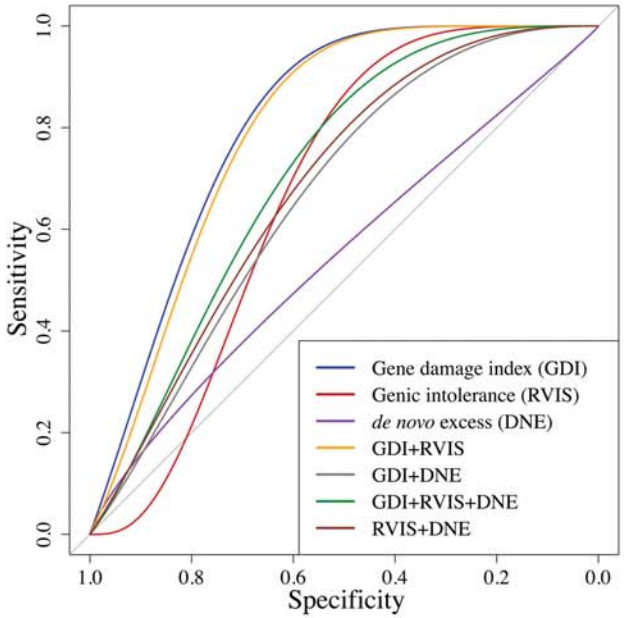


Formal ROC curves comparisons between GDI and other gene-level methods

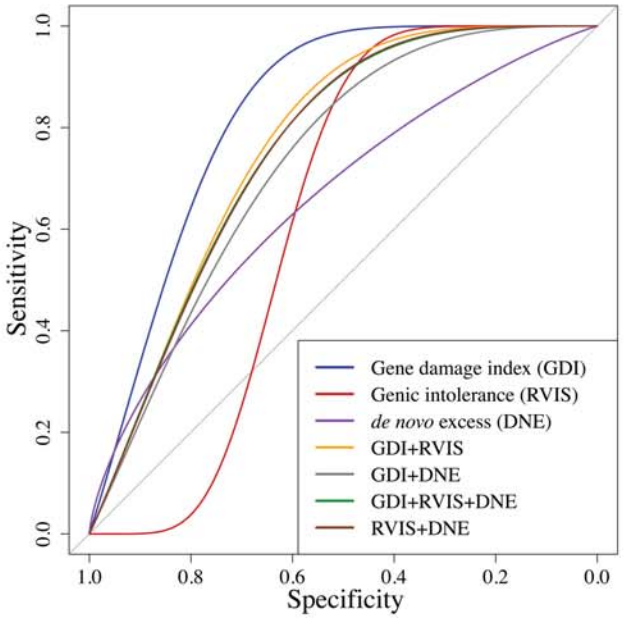
ALL



AR



AD



- GDI shows better performance to detect FP variants, whereas RVIS and *de novo* excess are better to detect TP mutations

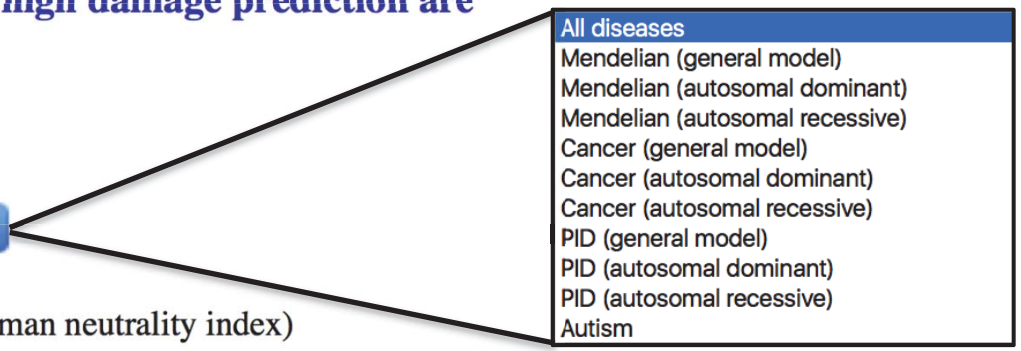
The GDI server: http://lab.rockefeller.edu/casanova/GDI

The Gene Damage Index (GDI) Server

Estimate the GDI for a list of genes. Genes with high damage prediction are unlikely to be disease-causing.

Link to the main GDI project webpage

Disease Type: All diseases



☐ In addition, display selective pressure (McDonald-Kreitman neutrality index)

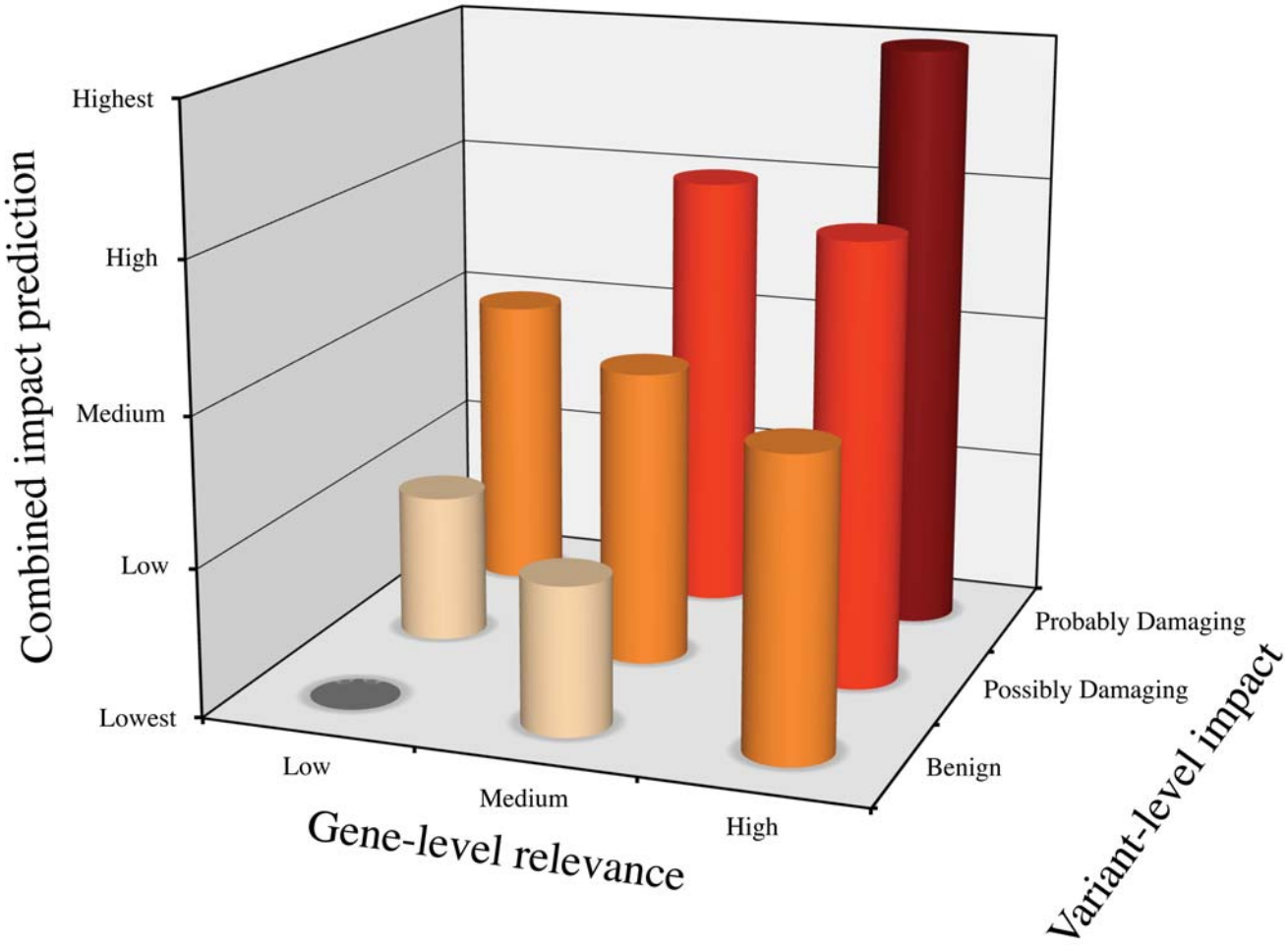
Type in genes separated with space, tab or comma in one or multiple rows

Submit

- GDI and selective pressure estimates for lists of human genes
- Filtering 20%-60% of variants in genes irrelevant to disease

Itan, Y., *et al.* (2015) The human gene damage index as a gene-level approach to prioritizing exome variants, *PNAS*, 112(44): 13615-20.

Proposed variant-level and gene-level phenotypic impact estimate

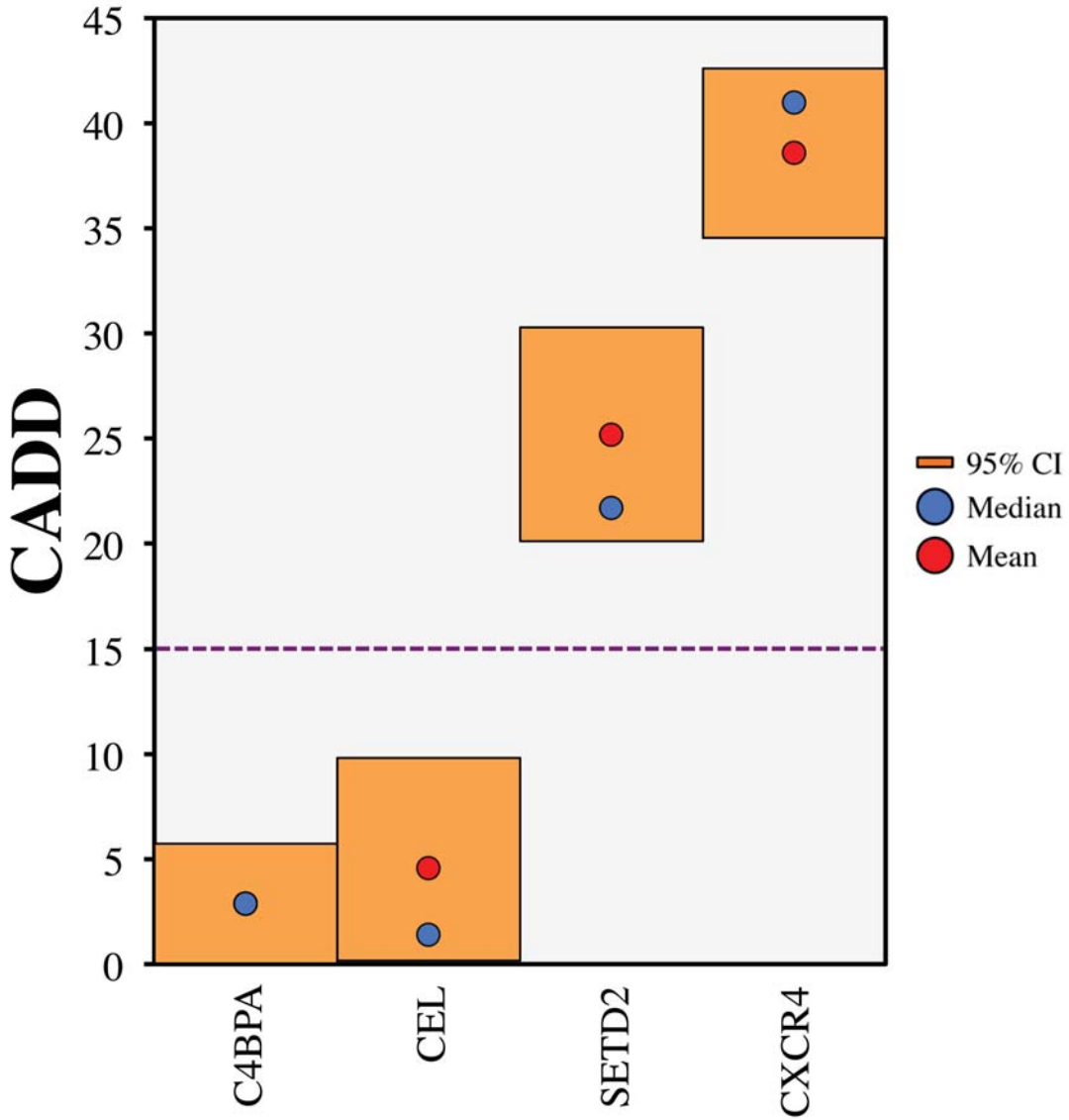


Highly damaging variant → strong phenotypic impact

Highly damaged gene → weak phenotypic impact

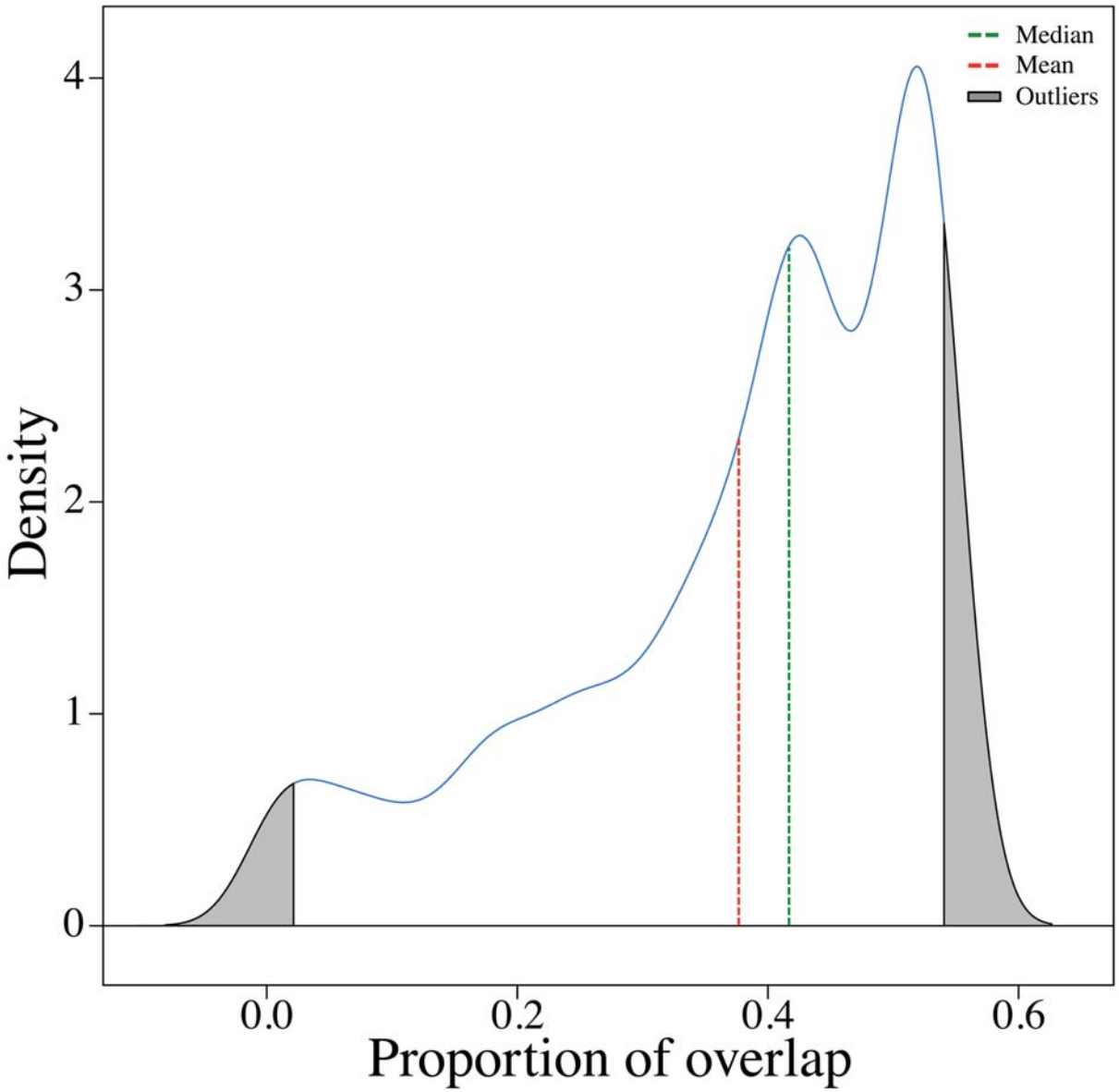
Itan, Y. and Casanova, JL. (2015) Can the impact of human genetic variations be predicted? *PNAS*, 112(37):11426-7.

Current variant impact prediction methods use a fixed benign/damaging cutoff for all human genes



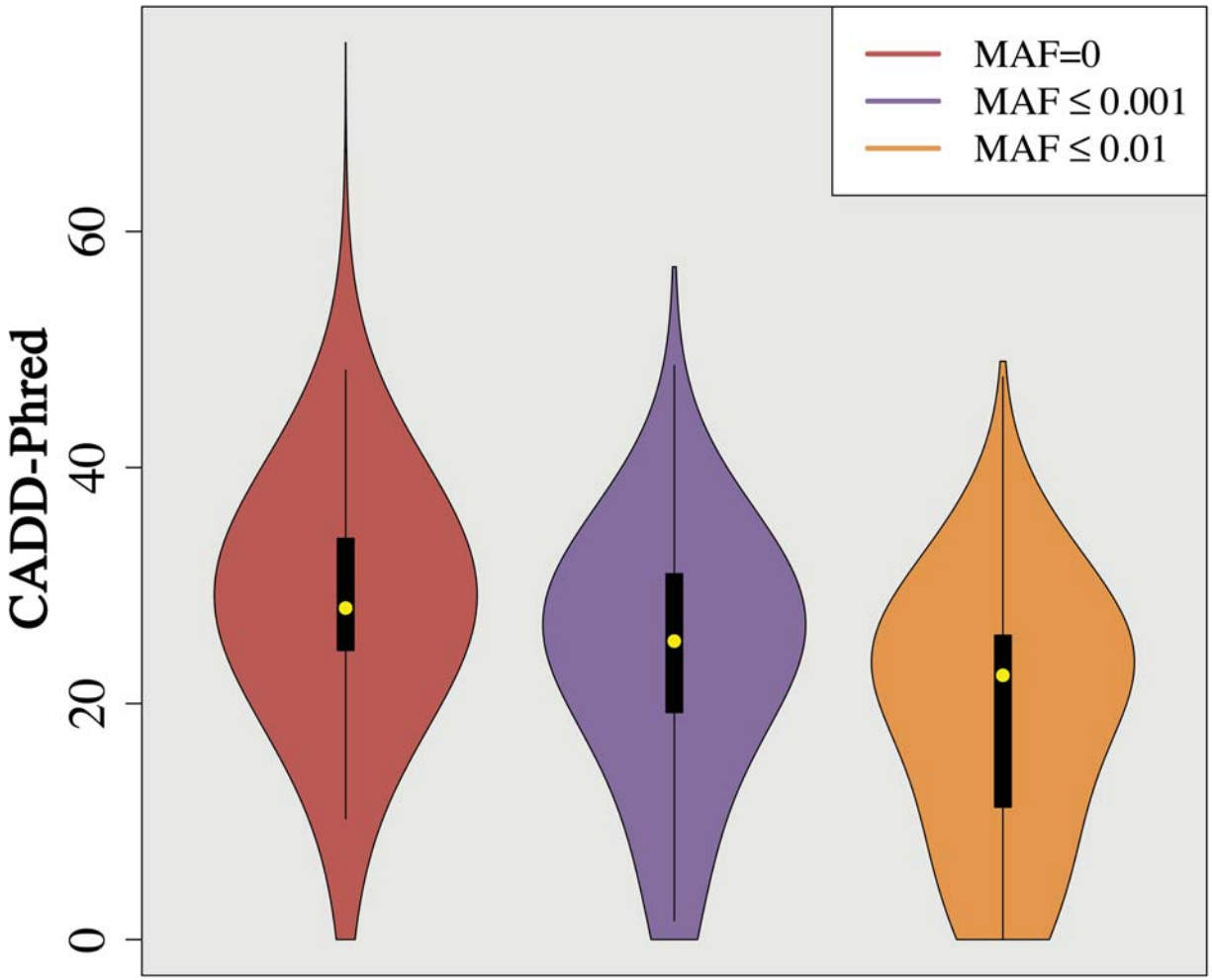
- False negative (FN, a true disease-causing allele predicted as benign) rates of new disease-causing alleles ≥36% when using CADD, PolyPhen-2 and SIFT with a fixed cutoff for all human genes
- Current methods cannot be safely used for hard filtering in NGS data (i.e. removing variants predicted to be benign)

The impact prediction scores of human genes' disease-causing mutations mostly do not overlap



Only ~40% overlap → any single cutoff could not be accurate for the majority of human genes

Impact prediction scores are strongly inversely correlated with minor allele frequency (MAF)



An over-representation of private disease-causing alleles if population frequency is not considered

The mutation significance cutoff (MSC): a benign/damaging (low/high impact) cutoff specific for each human gene

Input: all HGMD/ClinVar known disease-causing Mutations in human genes

↓

Estimating CADD/ PolyPhen-2/SIFT score for each of the above

↓

Calculating 90%/95%/99% confidence interval (CI) of the above for each gene

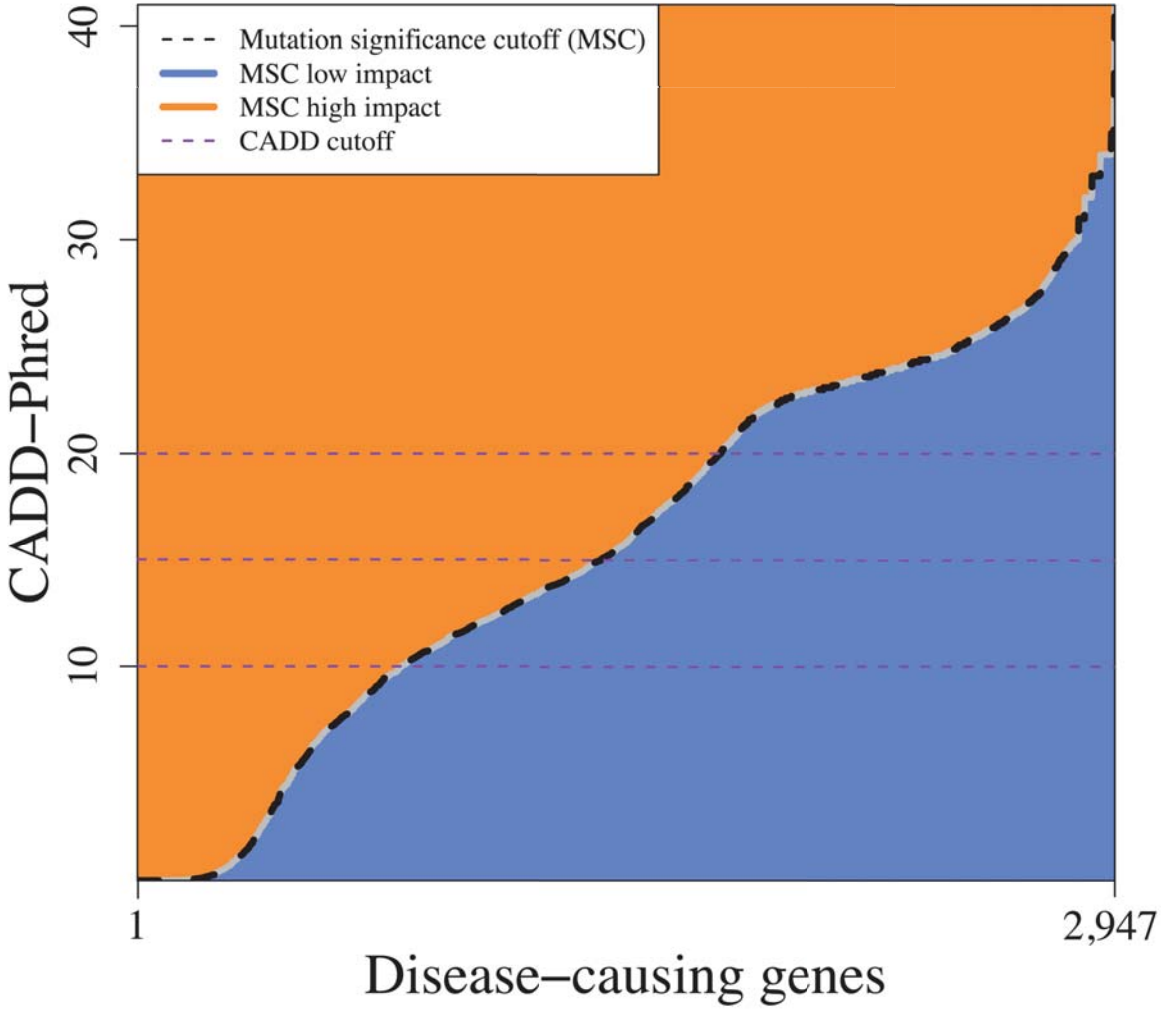
↓

MSC: the lower boundary of each CI above

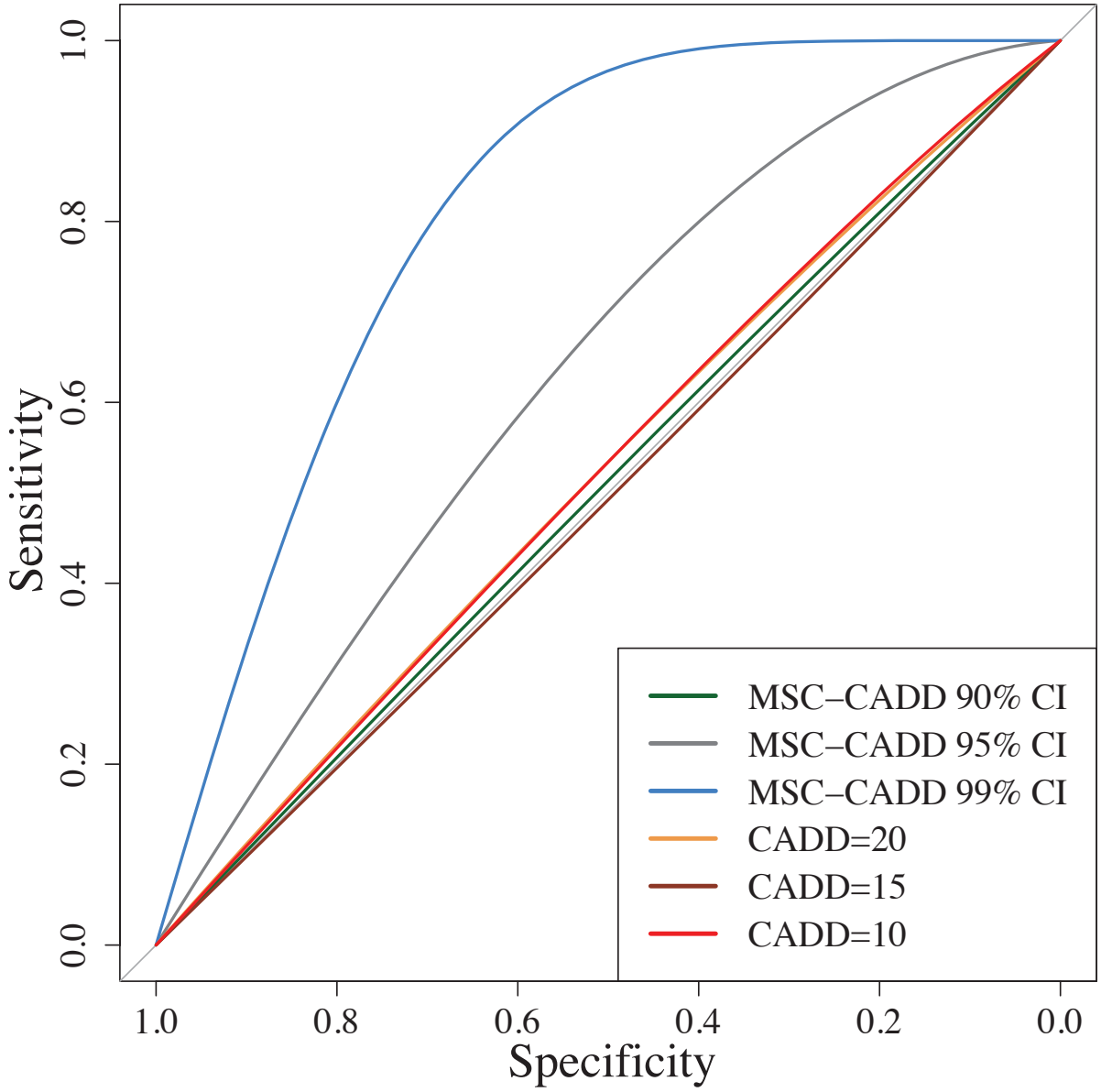
↓

Estimating MSC for the remaining human genes from 1,000 Genomes data

→ Score≥MSC → possibly damaging → keep
Score<MSC → likely benign → remove

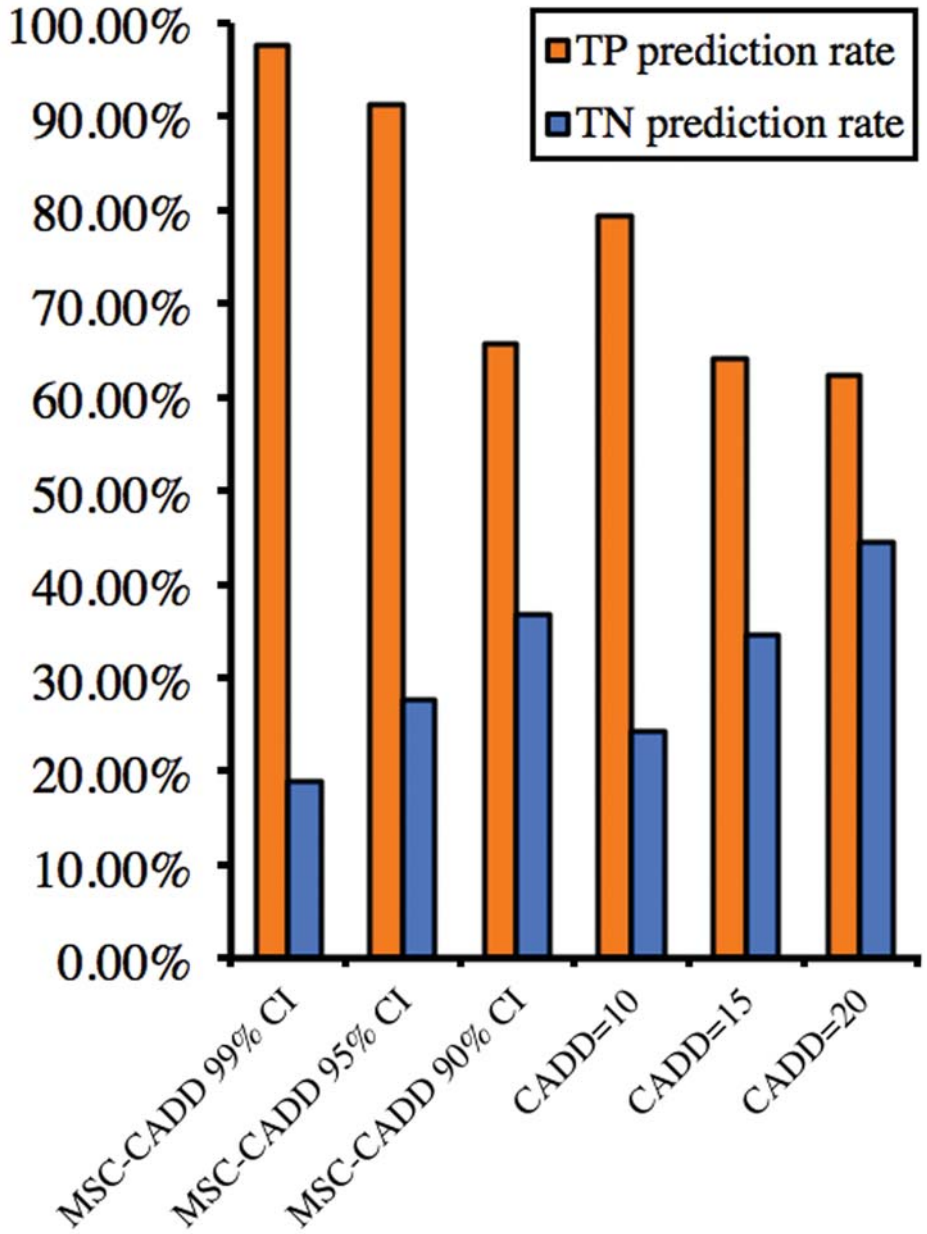


MSC performance #1



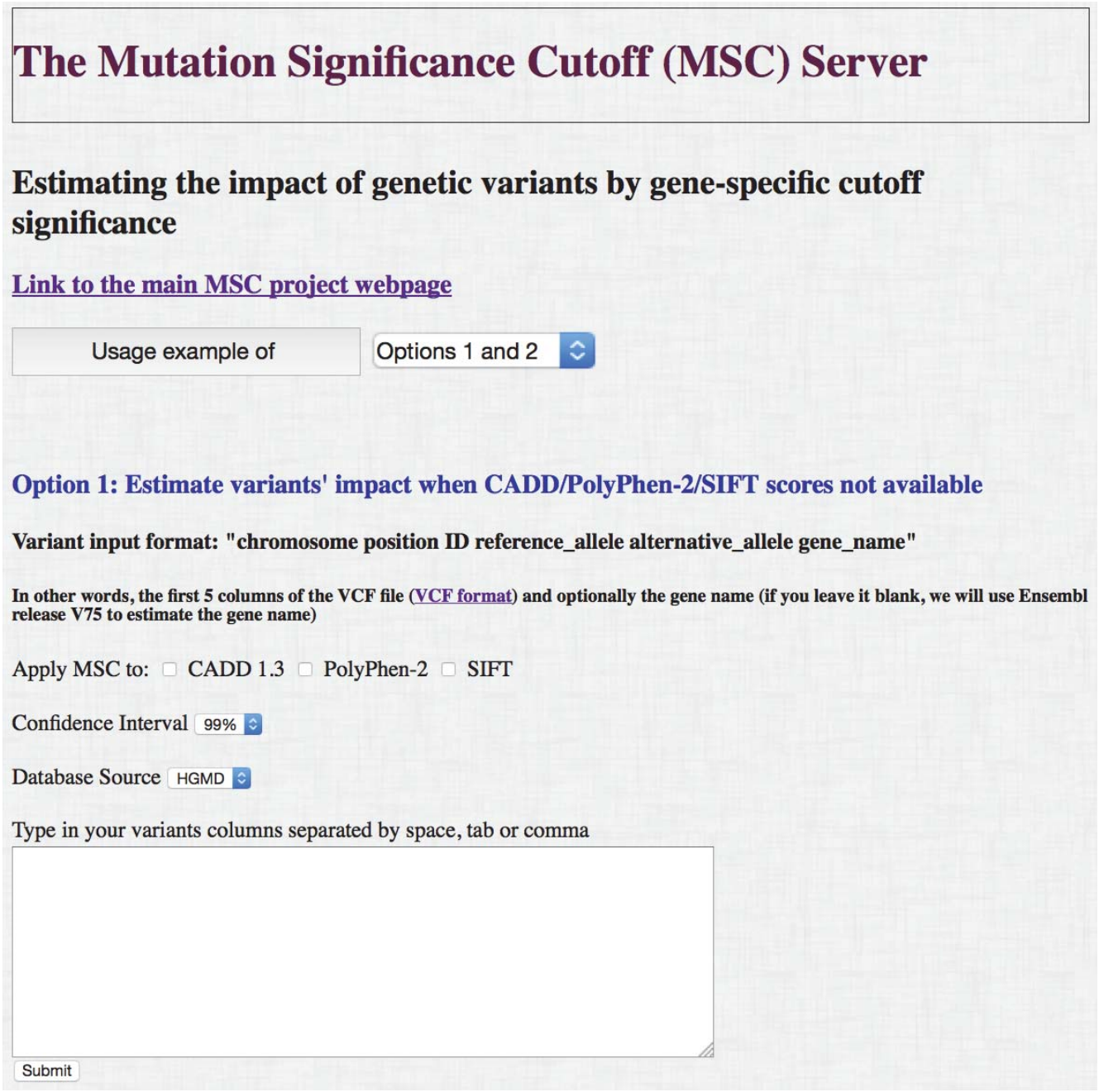
- FP validation set: private and non-disease-causing variants in patients' WES data
- TP validation set: HGMD new disease-causing alleles (not used to generate MSC)
- Best performance with CADD- and HGMD-based MSC, 99% CI

MSC performance #2



- Major increase in TP prediction, minor decrease in TN prediction
- Enables hard filtering in patients' NGS variants data with low risk of FNs (2% with 99% CADD-based MSC)

The MSC server: http://lab.rockefeller.edu/casanova/MSC



MSC estimates of low/high impact predictions for variants based on:

- Variant-prediction method: CADD, PolyPhen-2, or SIFT
- Confidence interval: 90%, 95%, or 99%
- Mutations database: HGMD or ClinVar

Itan, Y., *et al.* (2016). The mutation significance cutoff: gene-level thresholds for variant predictions, *Nature Methods*, 13(2):109-10.

Back to clustering. Reminder: HSE clusters before GDI+MSC filtering

Core_gene	Patient1	Patient2	Patient3	• • •														
TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	OBSL1	TRIM55	TTN	TTN	TTN	TTN	NEB	PDLIM
MYH3	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
MYBPC2	TTN	MYBPC2	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
MYOM2	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	MYOM2	MICAL
OBSL1	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	OBSL1	TRIM55	TTN	TTN	TTN	TTN	NEB	PDLIM
TPM1	TTN	TPM1	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
NEB	TTN	TTN	TTN	TTN	TTN	NEB	TTN	TTN	NEB	TTN	MYH3	MYH1	TTN	TTN	TTN	TTN	NEB	MICAL
OR4S2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR4S2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR4S2	OR4S2	OR51B4	OR2AG2	OR52M1	OR6K
ACTN1	ACTN1	TTN	TTN	TTN	TTN	TTN	TTN	TTN	TTN	ACTN1	OBSL1	MAGI2	TTN	TTN	TTN	TTN	NEB	IQGA
OR4C15	OR4C15	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR3A1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR3A1	OR51B4	OR3A1	OR52M1	OR6K
OBSCN	TTN	TTN	OBSCN	TTN	OBSCN	OBSCN	TTN	TTN	TTN	OBSCN	ANK1	SPTA1	OBSCN	OBSCN	TTN	TTN	MYOM2	SPTBN
OR4C3	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR7A5	OR7A5	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR56B1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4B1	OR4C3	OR4B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR2AJ1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR52N2	OR4C3	OR56B1	OR52N2	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR1N2	OR4C3	OR56B1	OR1N2	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR14A2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR51D1	OR4C3	OR56B1	OR2AJ1	OR51D1	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5D13	OR4C3	OR56B1	OR2AJ1	OR14A2	OR5D13	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR6C74	OR4C3	OR56B1	OR2AJ1	OR14A2	OR6C74	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10H2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR10H2	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5P2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR2Z1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR2Z1	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10G8	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10P1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10P1	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR1Q1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR1Q1	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5M8	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR51A7	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR52W1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR52W1	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR9K2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR9K2	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4K17	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR4K17	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR10G3	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR10G3	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR5H2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR5H2	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR52N4	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4N2	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR4N2	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR4K1	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR4K1	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K
OR6K3	OR4C3	OR56B1	OR2AJ1	OR14A2	OR4C15	OR5P2	OR10G8	OR5M8	OR51A7	OR52N4	OR6K3	OR5B12	OR51M1	OR6P1	OR51B4	OR2AG2	OR52M1	OR6K

HSE clusters after applying GDI+MSC filtering

Core gene	Patient1	Patient2	Patient3	Patient4	Patient5	Patient6	Patient7	Patient8	Patient9	Patient10	Patient11	Patient12	Patient13	Patient14	Patient15	Patient16	Patient17	Patient18	Patient1
TLR3	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP		CCDC47	IRAK4	TLR3		
TICAM1	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP		CCDC47	IRAK4	TLR3		
UNC93B1	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	IKBKB	TRAF5			TLR3	TRAIP	
TBK1	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP		CCDC47	IRAK4	TLR3		
TRAF3	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP		CCDC47	IRAK4	TLR3		
OPTN	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP		CCDC47	IRAK4	TLR3		
CCDC47	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	IKBKB	TRAF5	CCDC47		TLR3	TRAIP	
TRAF5	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP	TRAF5	CCDC47	IRAK4	TLR3		
TLR8	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	IKBKB	TRAF5			TLR3	TRAIP	TLR8
CEP152					SSI8	CEP78	SYNJ2		NINL					SYNJ2	NINL		TUBGCP2		SSI8
FARSA	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	IKBKB	TRAF5			TLR3	TRAIP	
SYNJ2					SSI8	GTSE1	SYNJ2		NINL					SYNJ2	NINL		TUBGCP2		SSI8
RIPK3	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3			CCDC47		TLR3	RIPK3	
ITGB4					ITGB4	ITGB4							CTNND1		ITGB4	ITGA8			
ZBP1	TBK1	TBK1	TRAF3	TLR3	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3			CCDC47		TLR3	RIPK3	
NINL					SSI8	CEP78			NINL						NINL		TUBGCP2		SSI8
ITGA8					ITGB4	ITGB4							CTNND1		ITGB4	ITGA8			
TUBGCP2					SSI8	CEP78			NINL						NINL		TUBGCP2		SSI8
CEP78					SSI8	CEP78			NINL						NINL		TUBGCP2		SSI8
LAMC1					ITGB4	ITGB4							CTNND1		ITGB4	ITGA8			
ITGA5					ITGB4	ITGB4							CTNND1		ITGB4	ITGA8			
ITGAD					ITGB4	ITGB4							CTNND1		ITGB4	ITGA8			
PTPN13							PTPN13		BRD7						PTPN13				PTPN13
CEP63					SSI8	CEP78			NINL						NINL		TUBGCP2		SSI8
NEIL1	RPA2	CHAF1A	CHAF1A		RFC5							WRN				RAD1			MSH6
TBKBP1	TBK1	TBK1	TRAF3	TBKBP1	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TBKBP1	TLR3					TLR3		
KCNQ5			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
FXR2	TBK1	TBK1	TRAF3	TBKBP1	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TBKBP1	TLR3					TLR3		
KCNC4			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
GTSE1					SSI8	GTSE1	SYNJ2							SYNJ2			TUBGCP2		SSI8
KCNA6			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
KCNC2			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
KCNQ1			KCNA6			KCNC4	KCNB1	KCNQ5			KCNQ1					KCNG2			
KCNS1			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
KCNF1			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
KCNB1			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
KCNG2			KCNA6			KCNC4	KCNB1	KCNQ5			KCNC2					KCNG2			
F10					ITGB4	ITGB4			FHL2		ITGB5		ITGA7		ITGB4	ITGA8			
ITGAM					ITGB4	ITGB4			FHL2		ITGB5		ITGA7		ITGB4	ITGA8			
DUSP1					SSI8	GTSE1	SYNJ2		NINL				DUSP1	SYNJ2	NINL		TUBGCP2		SSI8
TIRAP	TBK1	TBK1	TRAF3	TLR3	UNC93B1	PELI3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP	TRAF5	CCDC47	IRAK4	TLR3	TRAIP	
IRAK4	TBK1	TBK1	TRAF3	TLR3	UNC93B1	PELI3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP	TRAF5	CCDC47	IRAK4	TLR3	TRAIP	
PELI3	TBK1	TBK1	TRAF3	TLR3	UNC93B1	PELI3	TLR3	TICAM1	UNC93B1	TLR3	TLR3	TLR3	TIRAP	TRAF5	CCDC47	IRAK4	TLR3	TRAIP	
TRAIP	TBK1	TBK1	TRAF3	TLR3		TLR3	TLR3	TICAM1	SP1	TLR3	MAP3K5	BIRC3	IKBKB	TRAF5			TLR3	TRAIP	
BCL11A		CHAF1A	CHAF1A		BCL11A				SP1		CEBPB					BCL11A	INGI		
CLCA4	CLCA4								CLCA4										
IKBKE	TBK1	IKBKE	TRAF3	TBKBP1	UNC93B1	TLR3	TLR3	TICAM1	UNC93B1	TLR3	TBKBP1	TLR3					TLR3		
BIRC3	TBK1	TBK1	TRAF3	TLR3		TLR3	TLR3	TICAM1		TLR3	MAP3K5	BIRC3	IKBKB	TRAF5			TLR3	RIPK3	
BRD7							PTPN13		BRD7						PTPN13				PTPN13
RPA2	RPA2	CHAF1A	CHAF1A		RFC5							WRN				RAD1			MSH6
RPA4	RPA2	CHAF1A	CHAF1A		RFC5							WRN				RAD1			MSH6
TRIM37	TBK1	TBK1	TRAF3	TLR3		TLR3	TLR3	TICAM1	SP1	TLR3	MAP3K5	BIRC3	IKBKB	TRAF5			TLR3		TRIM37

Known HSE-causing gene
Plausible HSE modifier gene
Plausible HSE-causing gene
Possible novel HSE pathway

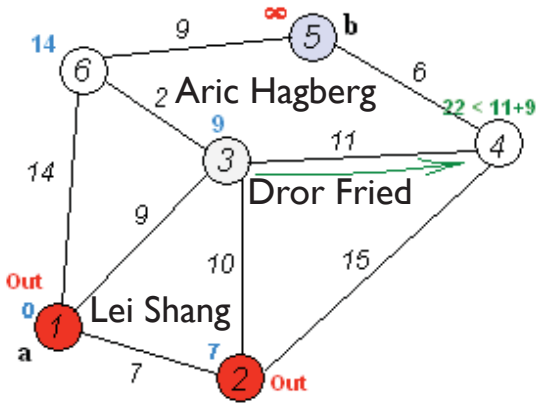
$P<10^{-9}$ for TLR3 pathway genes enrichment in top 1% clusters genes

Thank you all

Jean-Laurent Casanova

Laurent Abel

Lluis Quintana-Murci




Janet Markle


Alexandre Bolze


Jill De Jong


Emmanuelle Jouanguy


Fabien Lafaille


Stephanie Boisson-Dupuis

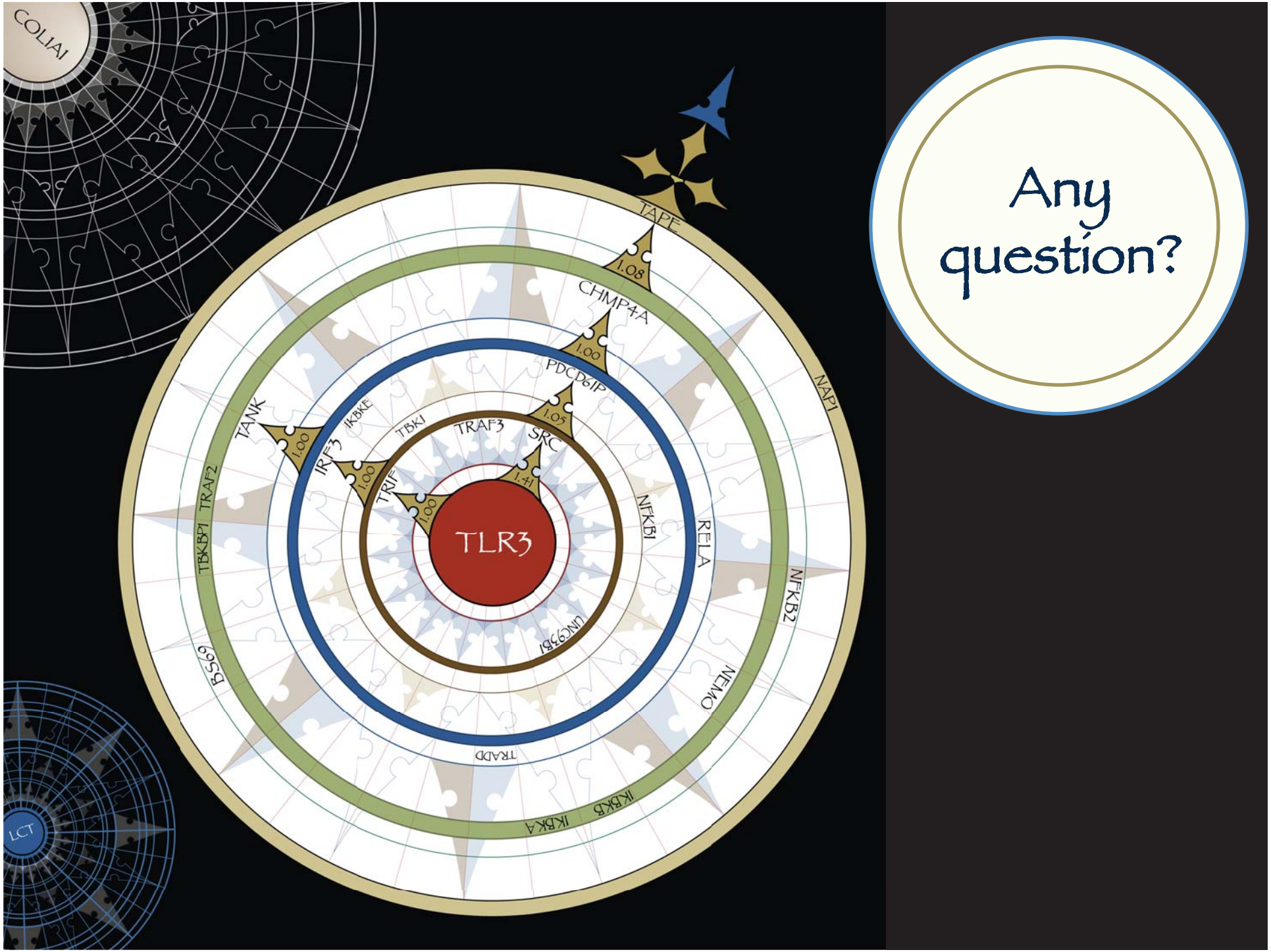

Ruben Martinez Barricarte


Marcela Moncada Velez

SCIENCE FOR THE BENEFIT OF HUMANITY

HHMI
HOWARD HUGHES MEDICAL INSTITUTE

CTSA Clinical & Translational Science Awards ®


COLIA1
LCT
TAPE
1.08
CHMP4A
1.00
PDCD6IP
1.05
SRC
1.41
UNC93B1
TRAF3
TBK1
IKBKE
1.00
IRF3
1.00
TRIF
1.00
TANK
1.00
TLR3
NAP1
NFKB2
RELA
NFKBI
TRADD
IKBKB
IKBKA
TBKBP1
TRAF2
BCL10
Any question?