Using crowdsourced contests to improve CMap algorithms

CROWD
INNOVATION
LAB

KRAFT FAMILY FOUNDATION

Karim Lakhani
Rinat Sergeev
Jin Paik

[topcoder]

Tim Kirchner
Andy Lamora
Jen Odess

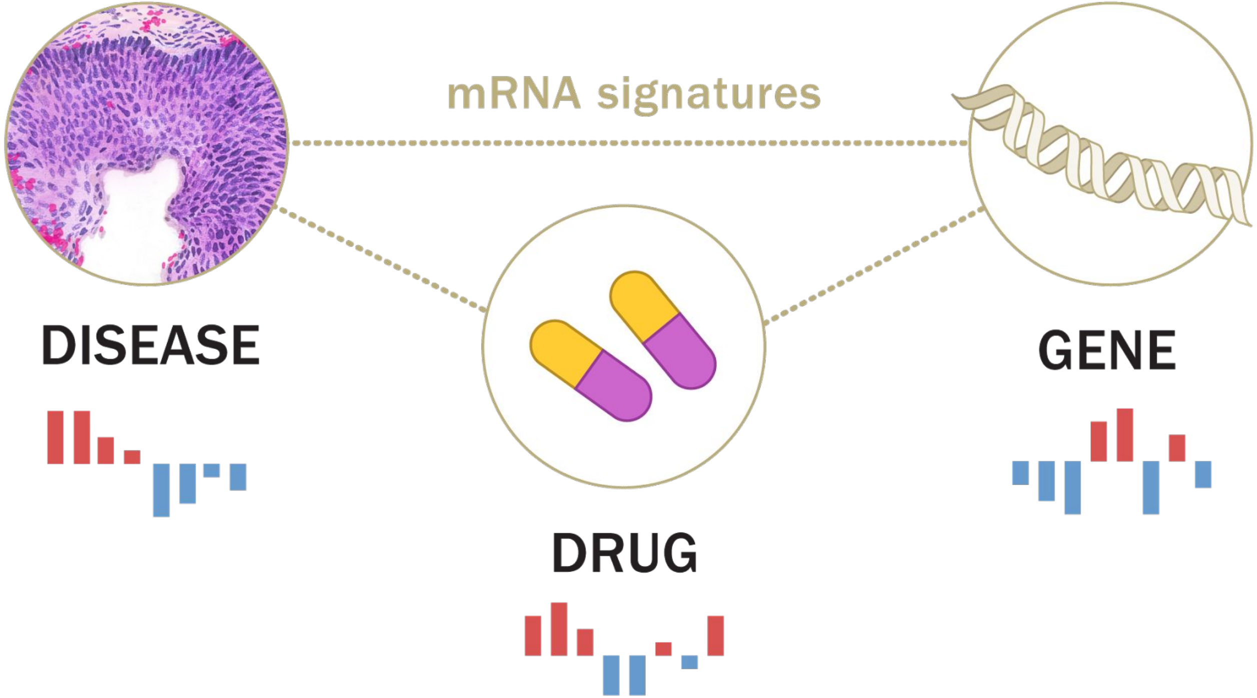
GTEx

NIH LINCS
PROGRAM

Kristin Ardlie
and team

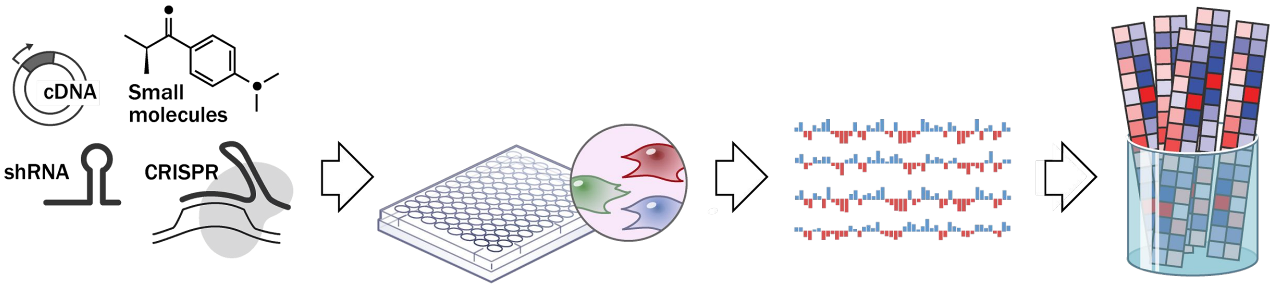
Connectivity Map (CMap) Concept

Linking disease, therapeutics and cell physiology



CMap experiment

Perturb cells, measure cellular response using a relevant molecular readout



Perturbagens

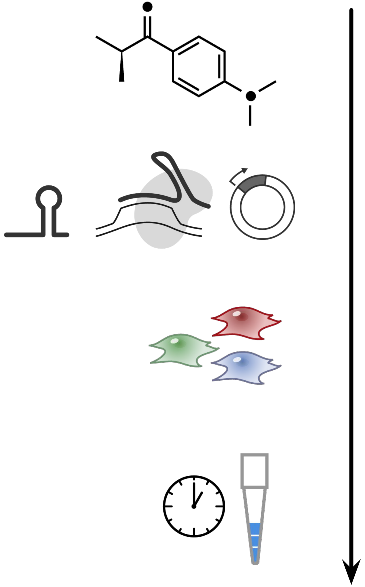
Cell types

Phenotype

- Cancer cell lines
- Immortalized
- iPSC derived
- Primary

Expanding CMap

Additional perturbagen & cellular context



More small molecule compounds

□ Drugs, tools, natural products

Genomic perturbations

□ shRNA, CRISPR, ORF, variants

Cellular context

□ Cell types, culture conditions

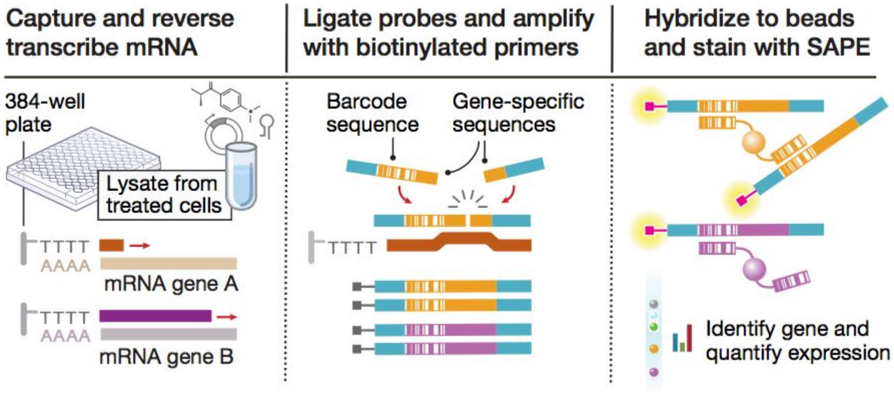
Treatment parameters

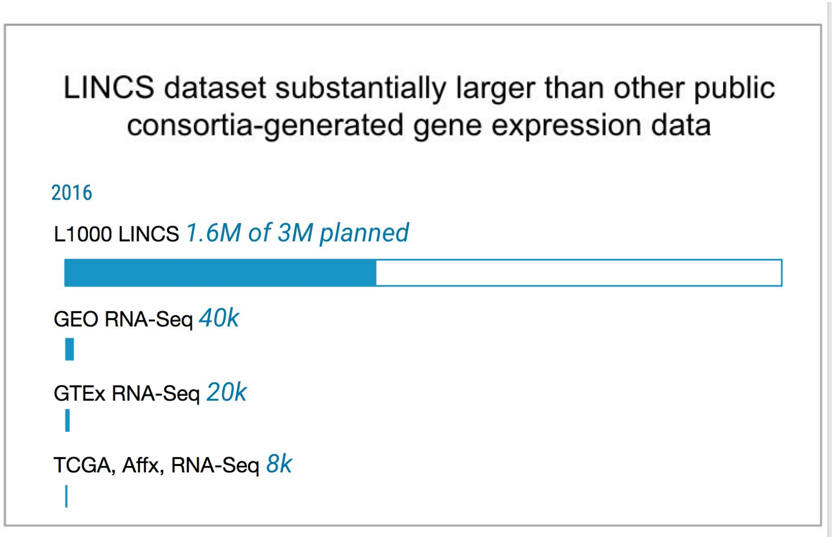
□ Concentrations, durations, combinations

At hundreds of $ per profile, approach does not scale

Key Innovation - The L1000 Assay

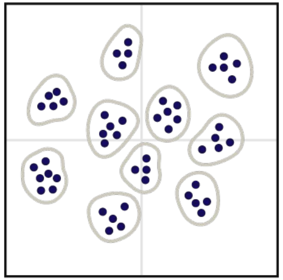
Ligation Mediated Amplification

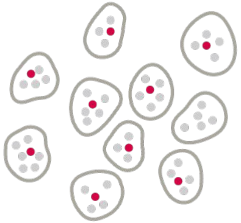

Capture and reverse transcribe mRNA
Ligate probes and amplify with biotinylated primers
Hybridize to beads and stain with SAPE
384-well plate
Lysate from treated cells
TTTT
AAAA
mRNA gene A
TTTT
AAAA
mRNA gene B
Barcode sequence
Gene-specific sequences
TTTT
Identify gene and quantify expression



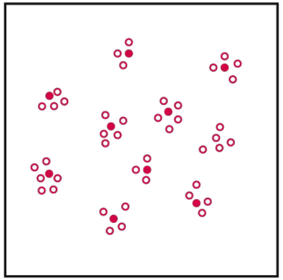
The Current CMap Inference Model

OLS linear regression





$$INF_i = w_{1i}LM_1 + w_{2i}LM_2 + \dots + w_{978i}LM_{978}$$



Landmark genes

Reduced representation of transcriptome

+

Computational inference model

Ordinary Least Squares Linear Regression

=

Genome-wide expression

A good inference model is important

- Better inference means more accurate comparisons and improved potential for discovery

Motivation

Success of CMap depends on community engagement

- Biologists & chemical biologists
  - Explore data
  - Generate and validate hypotheses
  - Develop and refine reagents to profile
- Computationalists
  - Develop and improve algorithms

HYPOTHESIS

A crowd-sourced computational challenge will engage the broader computational community and lead to impactful algorithm improvements.

Contest goals

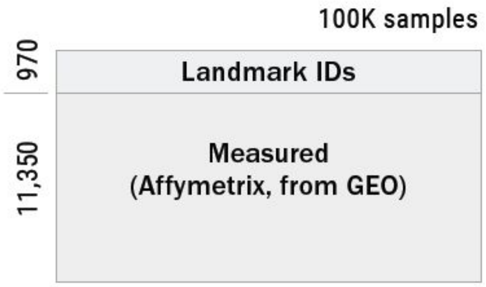
Improve inference and engage the community

- Get a better inference algorithm
- Develop an engaging and compelling computational challenge
  - Appeal to the broader computational community (not just Comp-Bio folks)
- Convert a CMap problem into a crowdsourcing problem
  - Make it understandable without deep domain knowledge
  - Make it possible to deploy winning solutions back to the dataset easily

Contest configuration

Train, predict, score, repeat

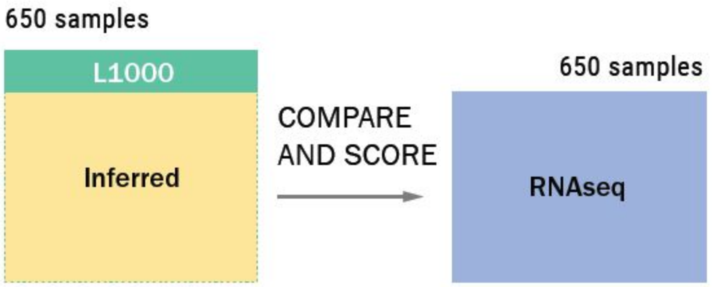
TRAINING



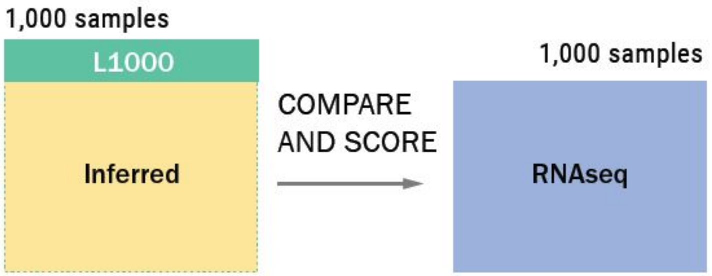
Samples from a variety of cell lines and tissue types

Samples and genes were anonymized, and no metadata were provided.

TEST SCORING (AKA 'PROVISIONAL')



HOLDOUT SCORING (AKA 'SYSTEM')



Contest Configuration

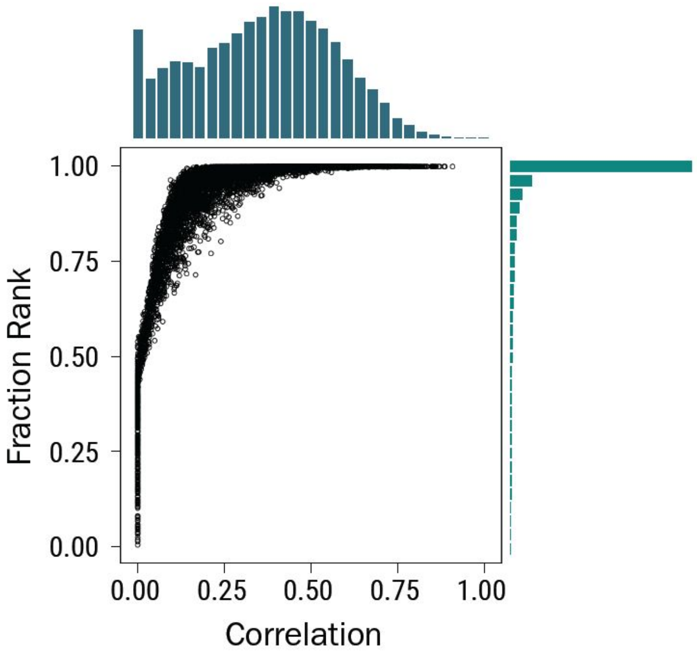
Predictions scored based on correlation with ground truth

Steps scoring algorithm performs for every gene:

- Compute correlation between inferred and measured values (self)
- Convert to rank relative to correlation between inferred gene and all other measured genes (non-self)
- Convert to fraction rank by dividing by total number of inferred genes (11,350) and subtracting from 1.

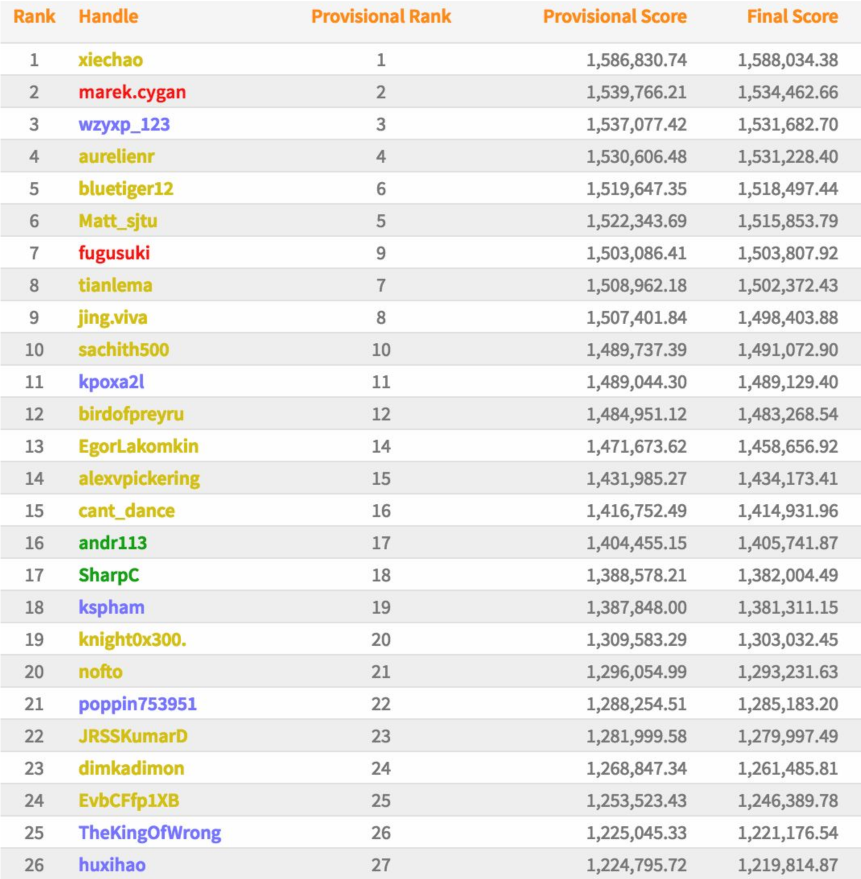
Combined Score

(Correlation + Fraction Rank) / 2



Ground Truth Data courtesy of GTEx.
Aliquots from same samples were profiled on both L1000 and RNAseq.

How am I doing?



Rank	Handle	Provisional Rank	Provisional Score	Final Score
1	xiechao	1	1,586,830.74	1,588,034.38
2	marek.cygan	2	1,539,766.21	1,534,462.66
3	wzyxp_123	3	1,537,077.42	1,531,682.70
4	aurelienr	4	1,530,606.48	1,531,228.40
5	bluetiger12	6	1,519,647.35	1,518,497.44
6	Matt_sjtu	5	1,522,343.69	1,515,853.79
7	fugusuki	9	1,503,086.41	1,503,807.92
8	tianlema	7	1,508,962.18	1,502,372.43
9	jing.viva	8	1,507,401.84	1,498,403.88
10	sachith500	10	1,489,737.39	1,491,072.90
11	kpoxa2l	11	1,489,044.30	1,489,129.40
12	birdofpreyru	12	1,484,951.12	1,483,268.54
13	EgorLakomkin	14	1,471,673.62	1,458,656.92
14	alexvpickering	15	1,431,985.27	1,434,173.41
15	cant_dance	16	1,416,752.49	1,414,931.96
16	andr113	17	1,404,455.15	1,405,741.87
17	SharpC	18	1,388,578.21	1,382,004.49
18	kspham	19	1,387,848.00	1,381,311.15
19	knight0x300.	20	1,309,583.29	1,303,032.45
20	nofto	21	1,296,054.99	1,293,231.63
21	poppin753951	22	1,288,254.51	1,285,183.20
22	JRSSKumarD	23	1,281,999.58	1,279,997.49
23	dimkadimon	24	1,268,847.34	1,261,485.81
24	EvbCFfp1XB	25	1,253,523.43	1,246,389.78
25	TheKingOfWrong	26	1,225,045.33	1,221,176.54
26	huxihao	27	1,224,795.72	1,219,814.87

How am I doing?

marek.cygan

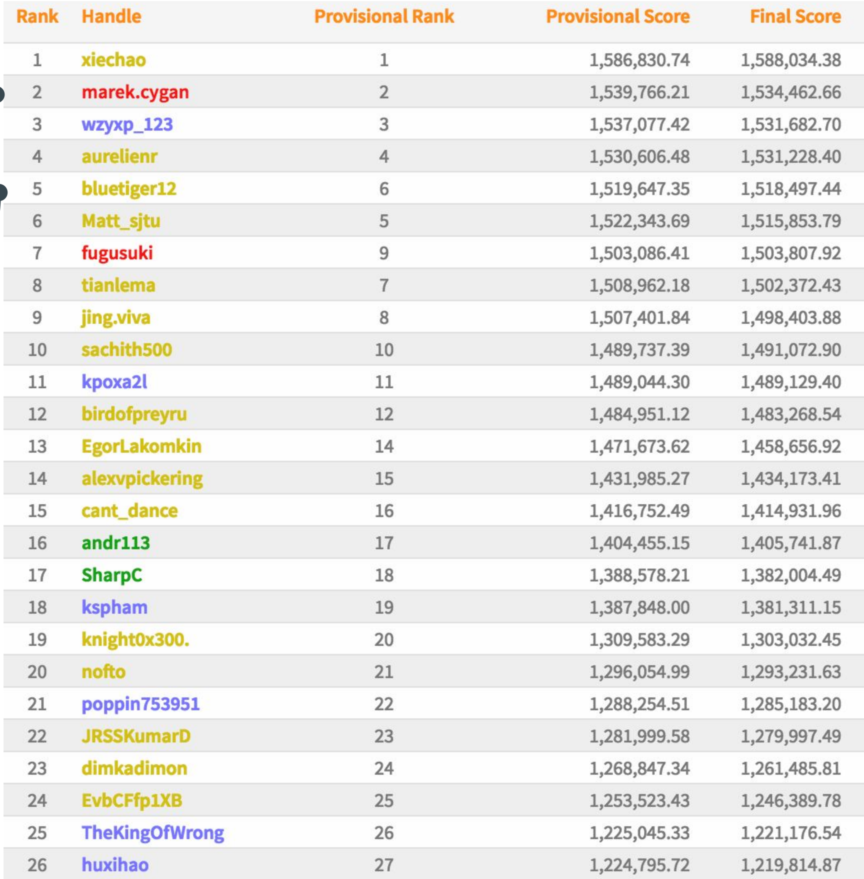
POLAND | 10 WINS
MEMBER SINCE SEPTEMBER, 2003

Assistant professor
Computer Science

bluetiger12

UNITED STATES
MEMBER SINCE JANUARY, 2013

BS in Mechanical &
Aerospace Engineering



Rank	Handle	Provisional Rank	Provisional Score	Final Score
1	xiechao	1	1,586,830.74	1,588,034.38
2	marek.cygan	2	1,539,766.21	1,534,462.66
3	wzyxp_123	3	1,537,077.42	1,531,682.70
4	aurelienr	4	1,530,606.48	1,531,228.40
5	bluetiger12	6	1,519,647.35	1,518,497.44
6	Matt_sjtu	5	1,522,343.69	1,515,853.79
7	fugusuki	9	1,503,086.41	1,503,807.92
8	tianlema	7	1,508,962.18	1,502,372.43
9	jing.viva	8	1,507,401.84	1,498,403.88
10	sachith500	10	1,489,737.39	1,491,072.90
11	kpoxa2l	11	1,489,044.30	1,489,129.40
12	birdofpreyru	12	1,484,951.12	1,483,268.54
13	EgorLakomkin	14	1,471,673.62	1,458,656.92
14	alexvpickering	15	1,431,985.27	1,434,173.41
15	cant_dance	16	1,416,752.49	1,414,931.96
16	andr113	17	1,404,455.15	1,405,741.87
17	SharpC	18	1,388,578.21	1,382,004.49
18	kspham	19	1,387,848.00	1,381,311.15
19	knight0x300.	20	1,309,583.29	1,303,032.45
20	nofto	21	1,296,054.99	1,293,231.63
21	poppin753951	22	1,288,254.51	1,285,183.20
22	JRSSKumarD	23	1,281,999.58	1,279,997.49
23	dimkadimon	24	1,268,847.34	1,261,485.81
24	EvbCFfp1XB	25	1,253,523.43	1,246,389.78
25	TheKingOfWrong	26	1,225,045.33	1,221,176.54
26	huxihao	27	1,224,795.72	1,219,814.87

wzyxp_123

CHINA | 1 WINS
MEMBER SINCE OCTOBER, 2010

PhD student
Machine learning

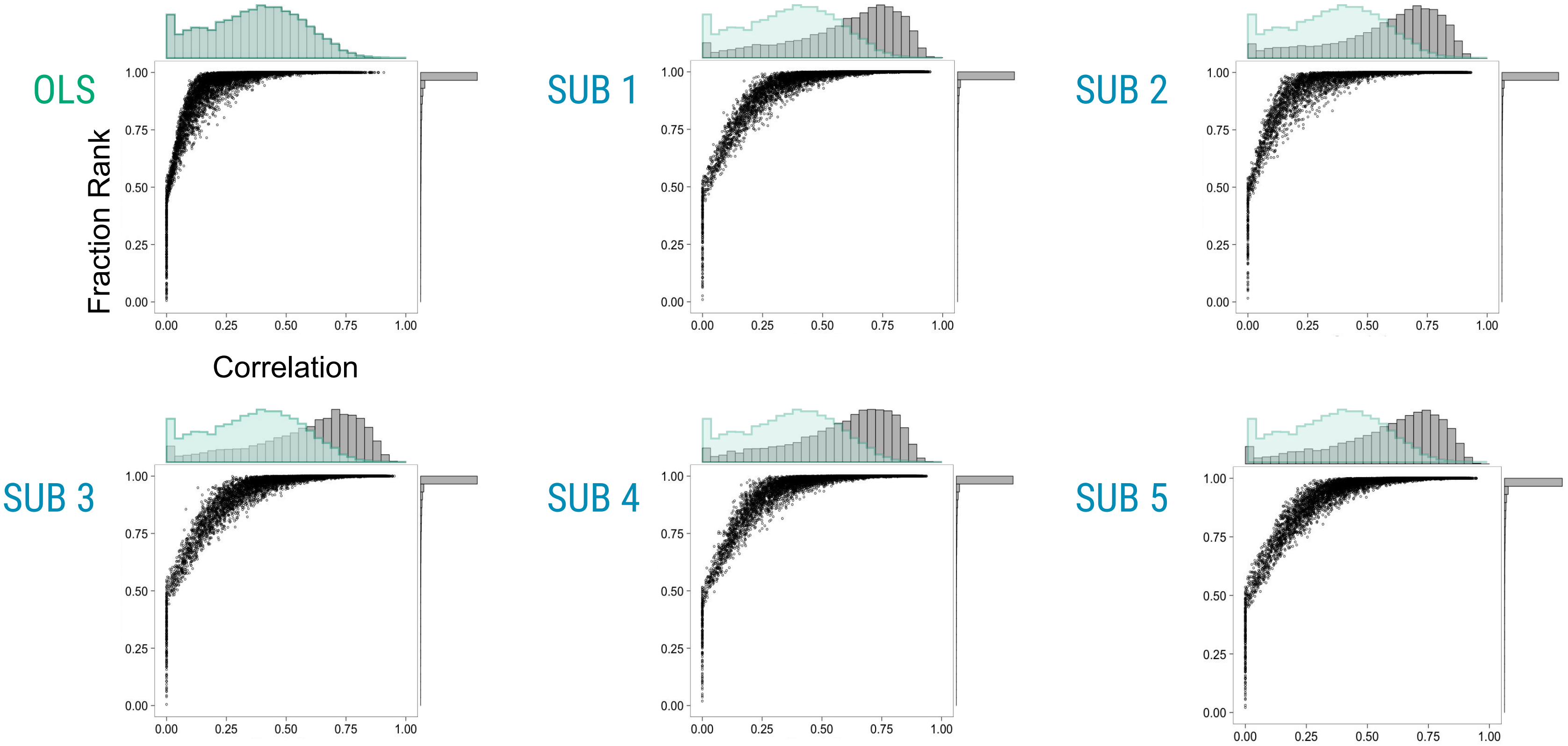
sachith500

SRI LANKA | 1 WINS
MEMBER SINCE AUGUST, 2011

PhD student
Machine learning

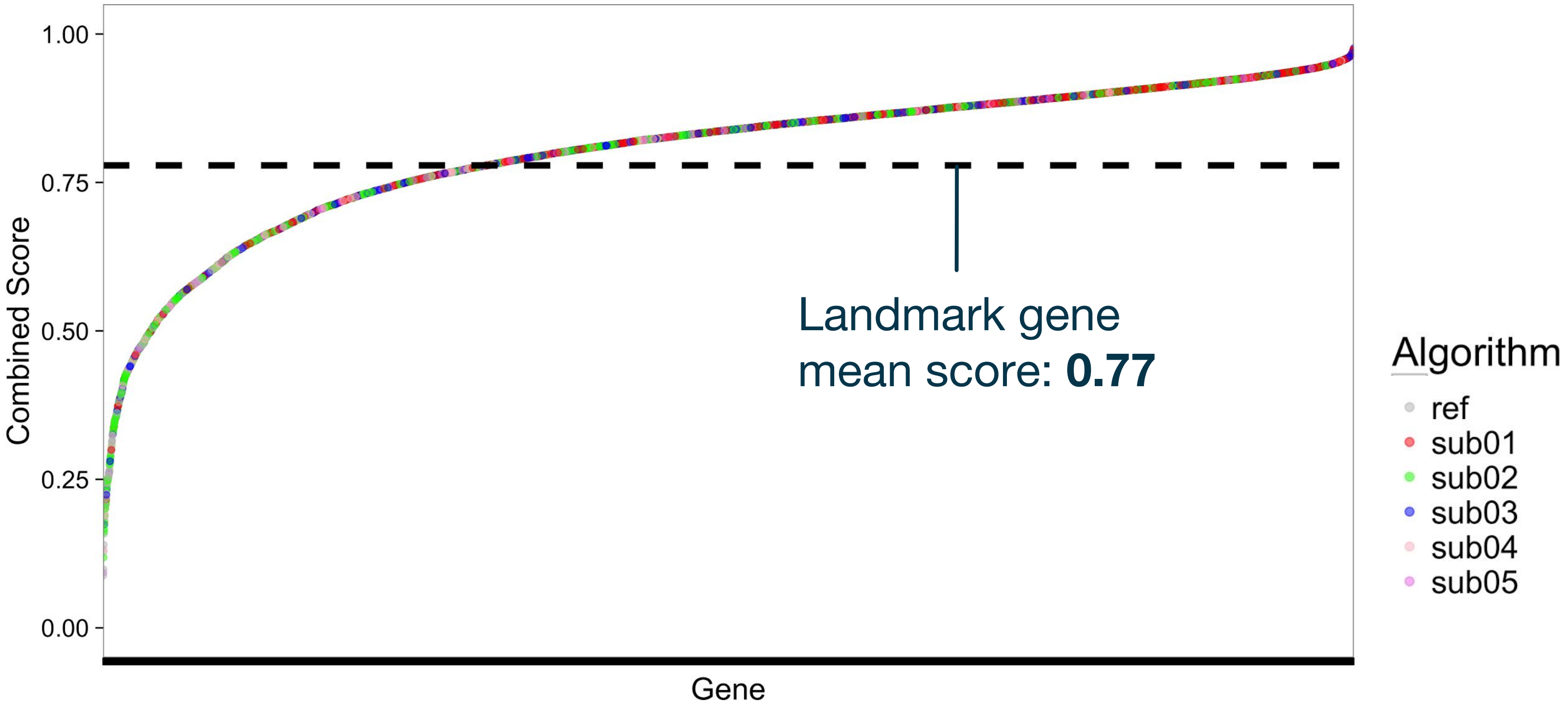
Contestants' models show marked improvements

Correlations improve, ranks remain high



Combining models provides further improvement

69% of genes are inferred with the same accuracy as if they were directly measured



Landmark gene mean score: **0.77**

Competitor demographics

Many new competitors engaged

468 registrants

88 competitors (made at least one submission)

50 new competitors (CMap Gene Inference was their first challenge)

1,116 submissions (average of 13.3 submissions per competitor)

5 of top ten contestants had little or no previous computational biology experience