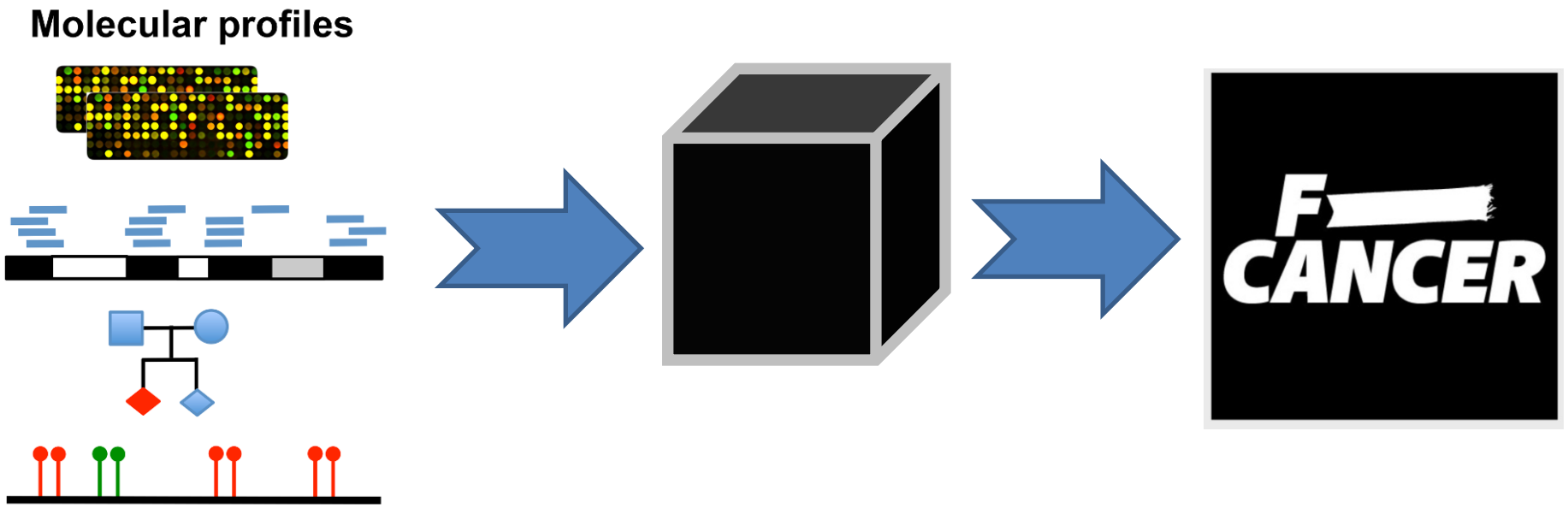
The ICGC-TCGA DREAM Somatic Mutation Calling Challenge: Preliminary Results

May 12, 2014

Dr. Paul C. Boutros

Principal Investigator, Informatics & Biocomputing
Ontario Institute for Cancer Research

General Plan for Data-Analysis

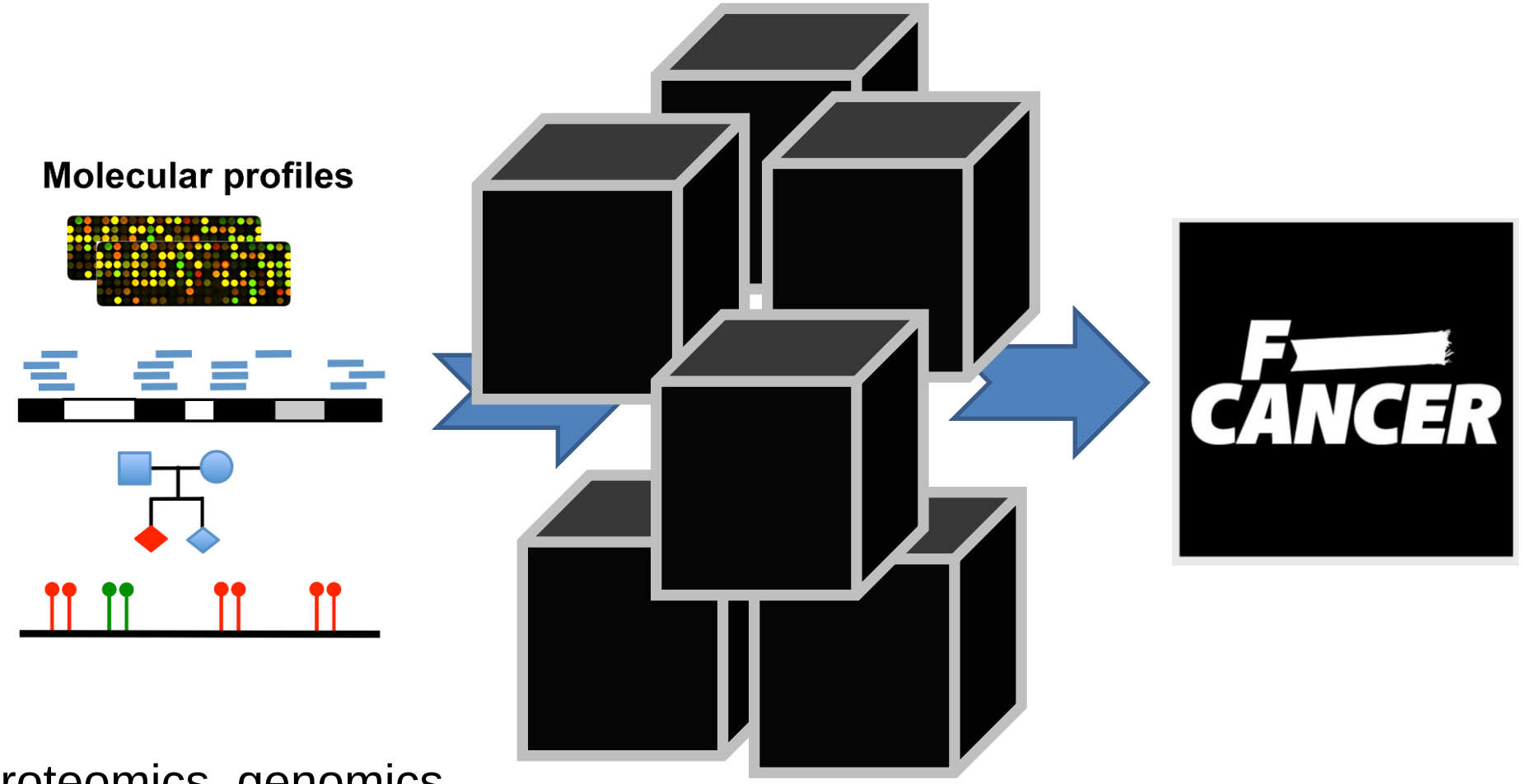


Proteomics, genomics, metabolomics…

Data-Analysis

Results

Different Analysis; Same Conclusions?

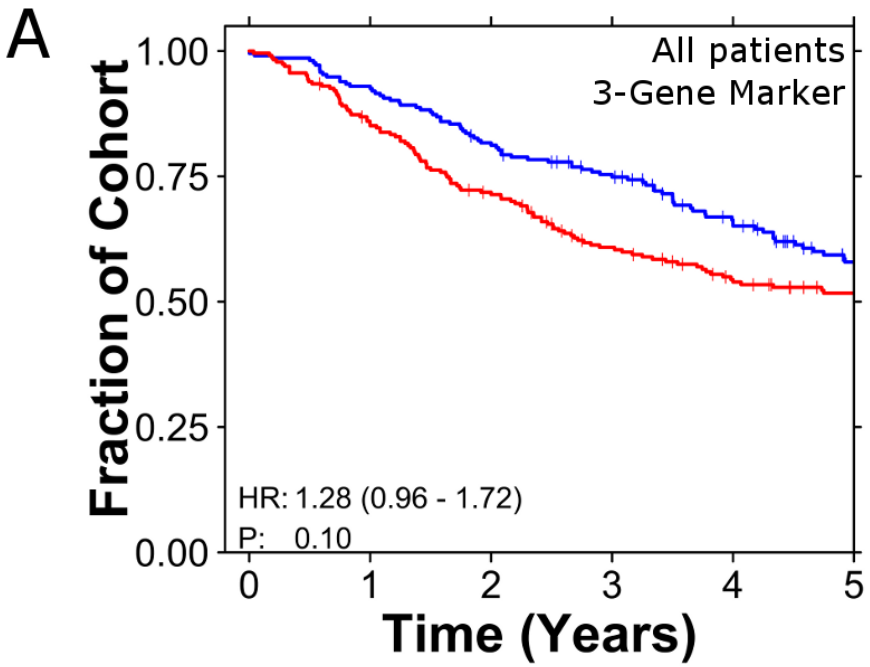


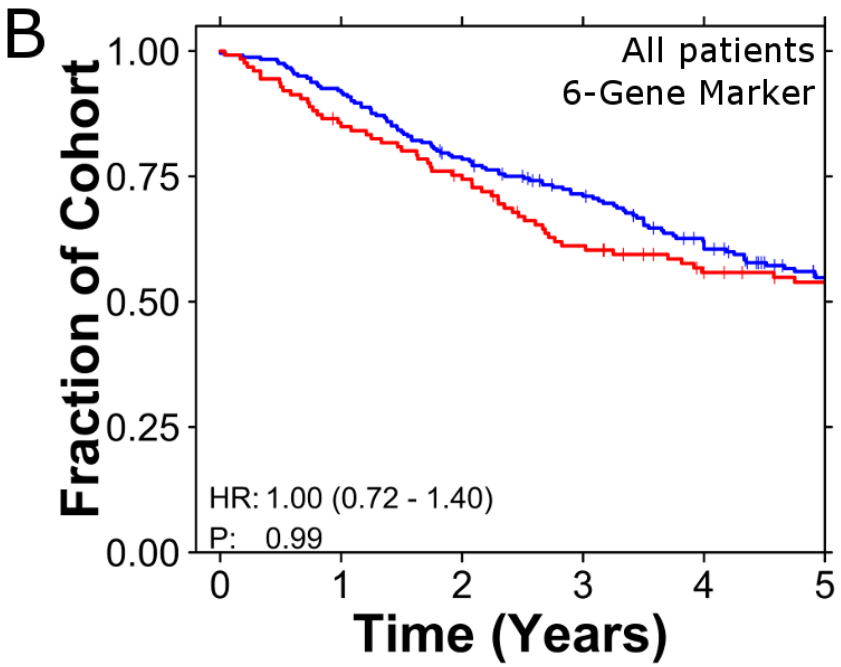
Proteomics, genomics, metabolomics…

Data-Analysis

Results

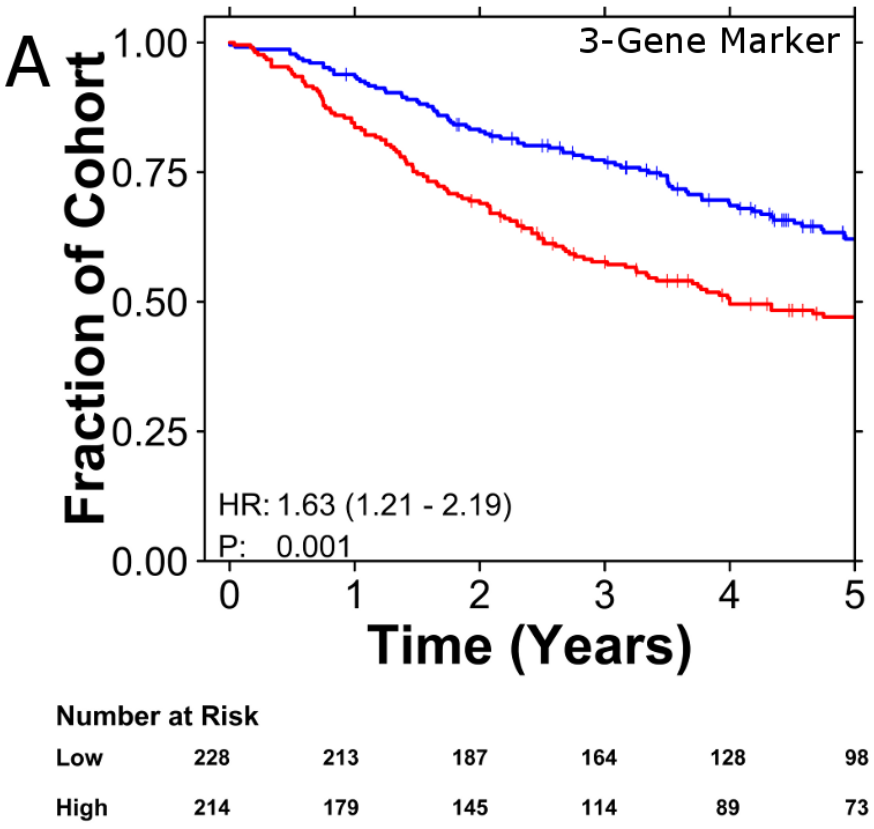
In Late 2010… A Failed Validation

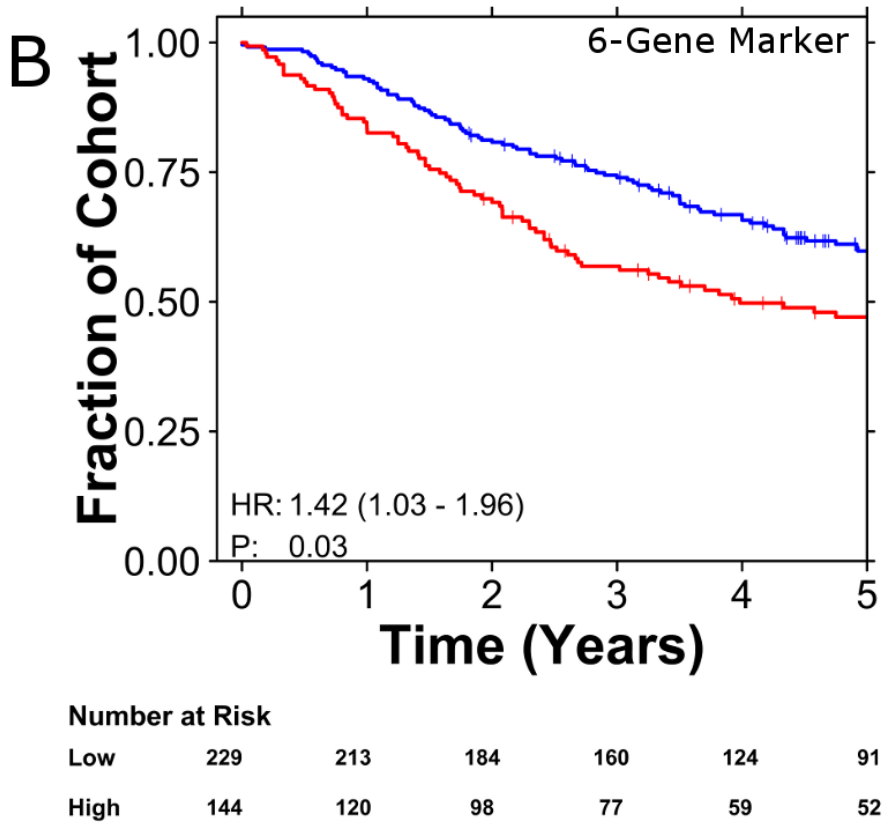




Subramanian & Simon, JNCI 2010

But When We Tried to Replicate…





Same dataset, same approach!

The Only Difference: Pre-Processing

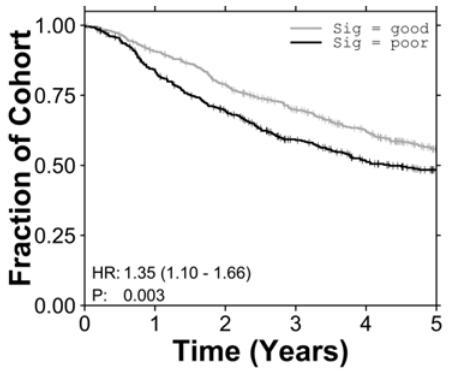
24 different pre-processing schemes; all combinations of:

Dataset handling:
- Separate
- Merged

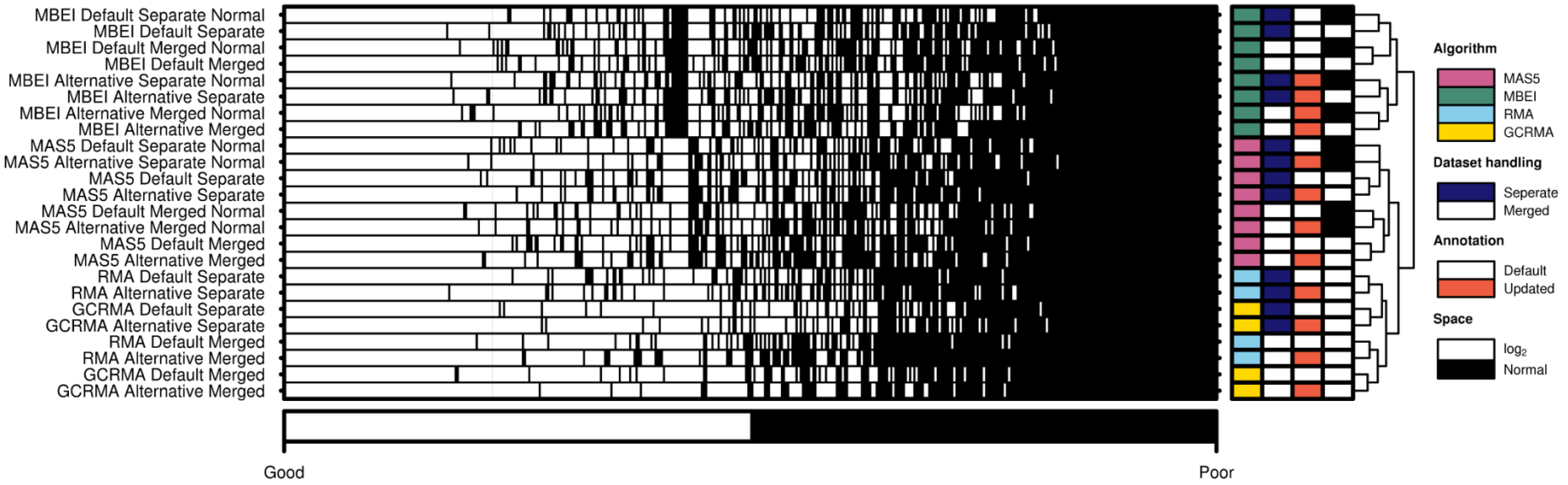
Algorithm:
- RMA
- GCRMA
- MAS5:
  - with $\log_2$-transformation
  - without $\log_2$-transformation
- MBEI:
  - with $\log_2$-transformation
  - without $\log_2$-transformation

Annotation:
- Default
- Alternative

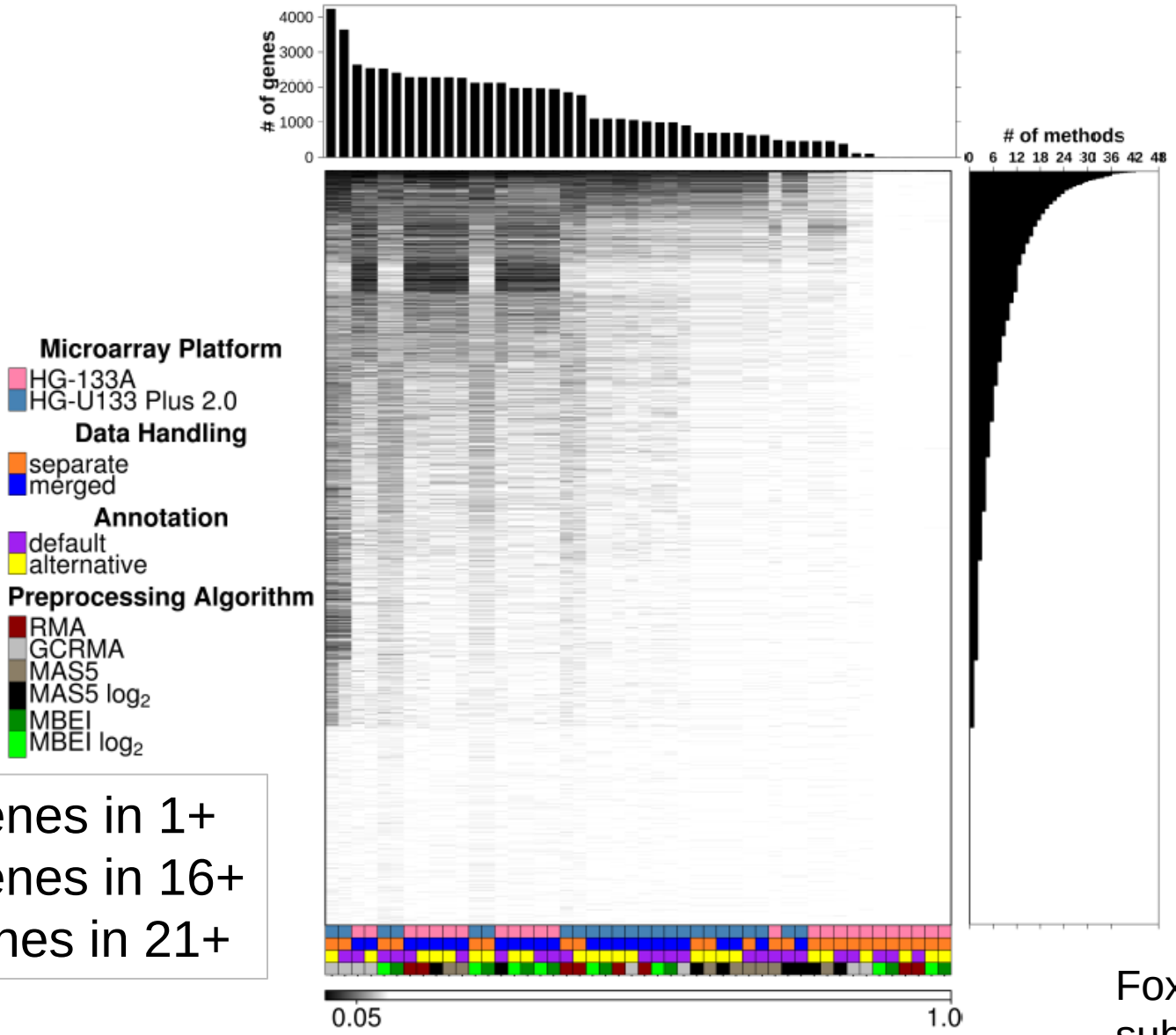
Evaluate classifier



Agreement: 151/442 Patients



This holds for all tumour-types: breast cancer



74% of genes in 1+
16% of genes in 16+
0% of genes in 21+

Fox *et al.* submitted

I Have a DREAM

- SAGE Bionetworks/DREAM
- Next contest – Genomic Calling Methods
- Collaborative between OICR, TCGA, SAGE

Dr. Adam Margolin, Dr. Josh Stuart

Our Initial Goal: Find the Best WGS Analysis Methods

The focus is solely on accuracy, not speed, computational efficiency or other considerations.

Real Tumour Data

- 10 Tumour/Normal pairs
  - sequenced to ~50x/30x
  - 5 from pancreatic tumours
  - 5 from prostate tumours

- Raw & processed data available

- All clinical information, protocols, *etc*. available

But What About Ethics Approval?

- PHI, so ICGC Data Access Coordinating Organization application needed for real data
- We have provided a template to expedite ethics approvals
- We sought and received an opinion on the Challenge from the Western IRB

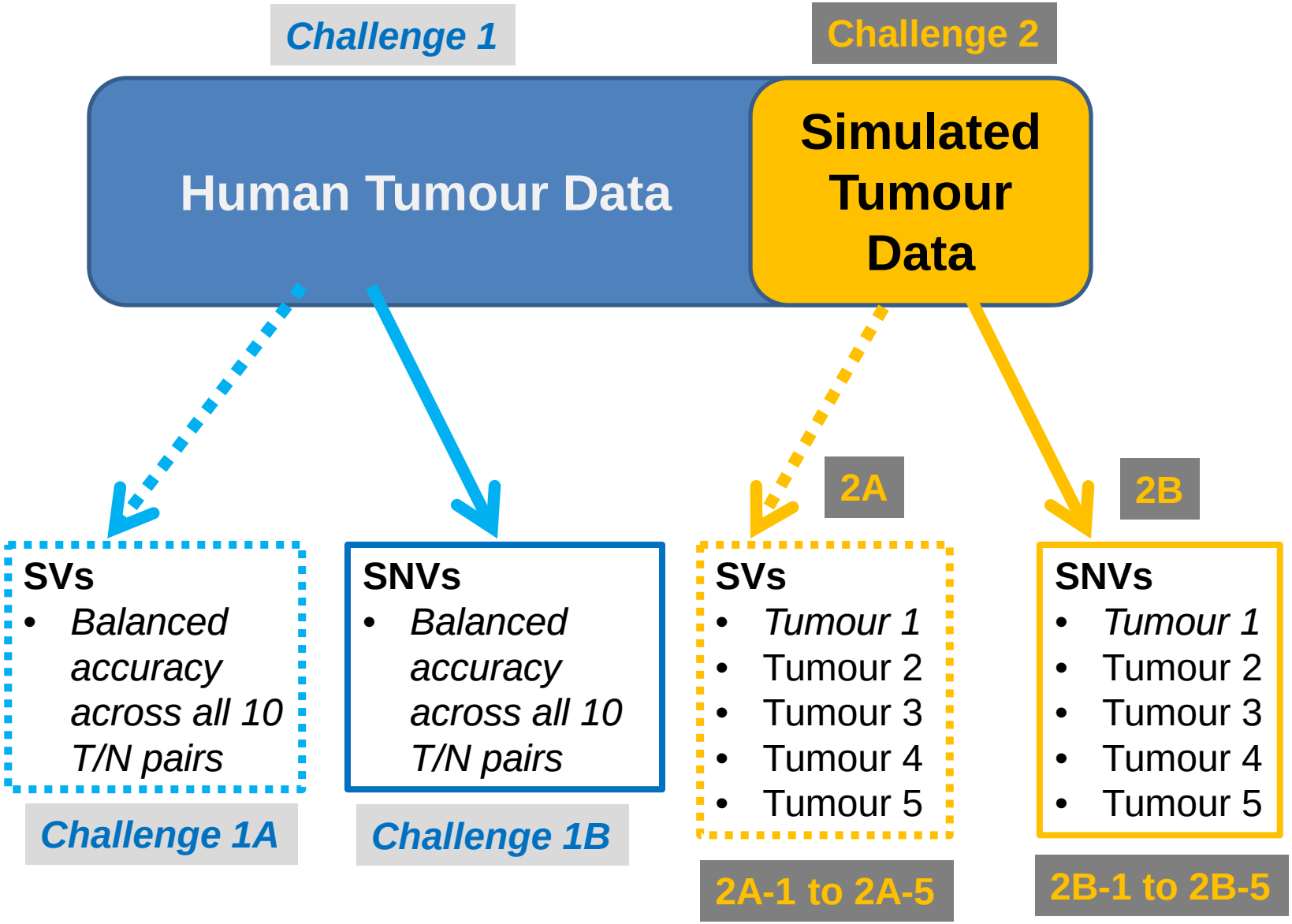
Simulated Tumour Data

- Start with a genome (cell line or germline)
- “burn in” SNVs & SVs using BAMSurgeon (Adam Ewing, UCSC)
- Take a subset of reads and introduce additional SNVs & SVs to create a tumour/normal pair
- 5 releases, **third is active now**!
- Increasing complexity, so good for “learning”

How Can You Get The Data?

- Register for the Challenge at Synapse
  - Complete an ICGC DACO Application
- Download using Annai's GeneTorrent
  - No-cost to download
- Directly access in the Google Compute Engine (Google cloud)
  - $2,000 free computing

Challenge Structure



How will the Challenge be scored?

Challenge 1: tumour data

10 Real Tumour/Normal Pairs

- Several thousand candidates will be validated (up to 10k)
- Validation will include (at least) re-sequencing to ~300x coverage using AmpliSeq primers on an IonTorrent

Challenge 2: *in silico* data

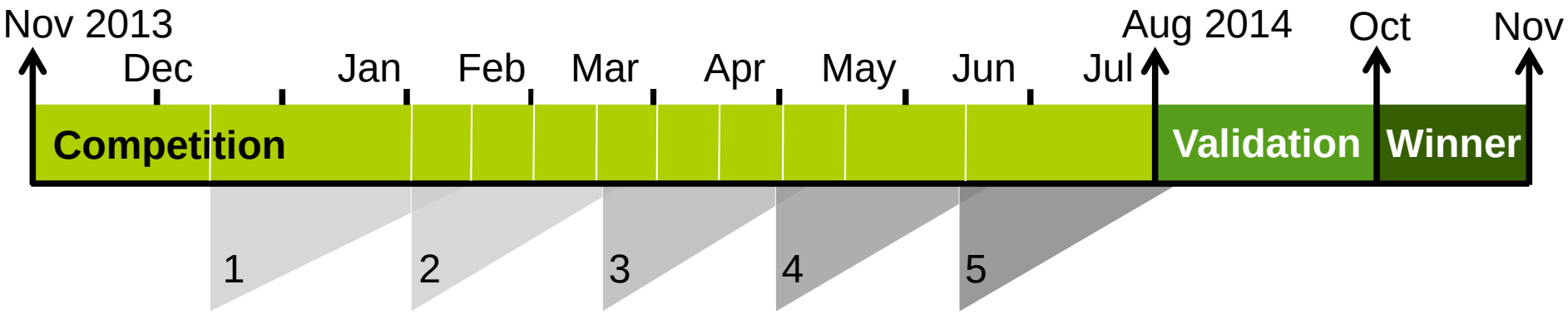
5 Synthetic Tumour/Normal Pairs

- A complete ground-truth is known for each dataset
- We will calculate sensitivity, specificity and balanced-accuracy for each genome on a held out piece of the genome

OICR
Ontario Institute for Cancer Research

UCSC

DREAM Mutation-Calling Challenge



***In silico* data:**

- 5 T/N pairs
- For “play” and dry-runs
- Releases of increasing complexity
- Rapid scoring turn-around
- BAMs (Novoalign or BWA)

Real data:

- 10 T/N pairs (50x/30x)
- Two tumour-types:
- 5 pancreatic
- 5 prostate
- Lane-level FASTQs & BAMs

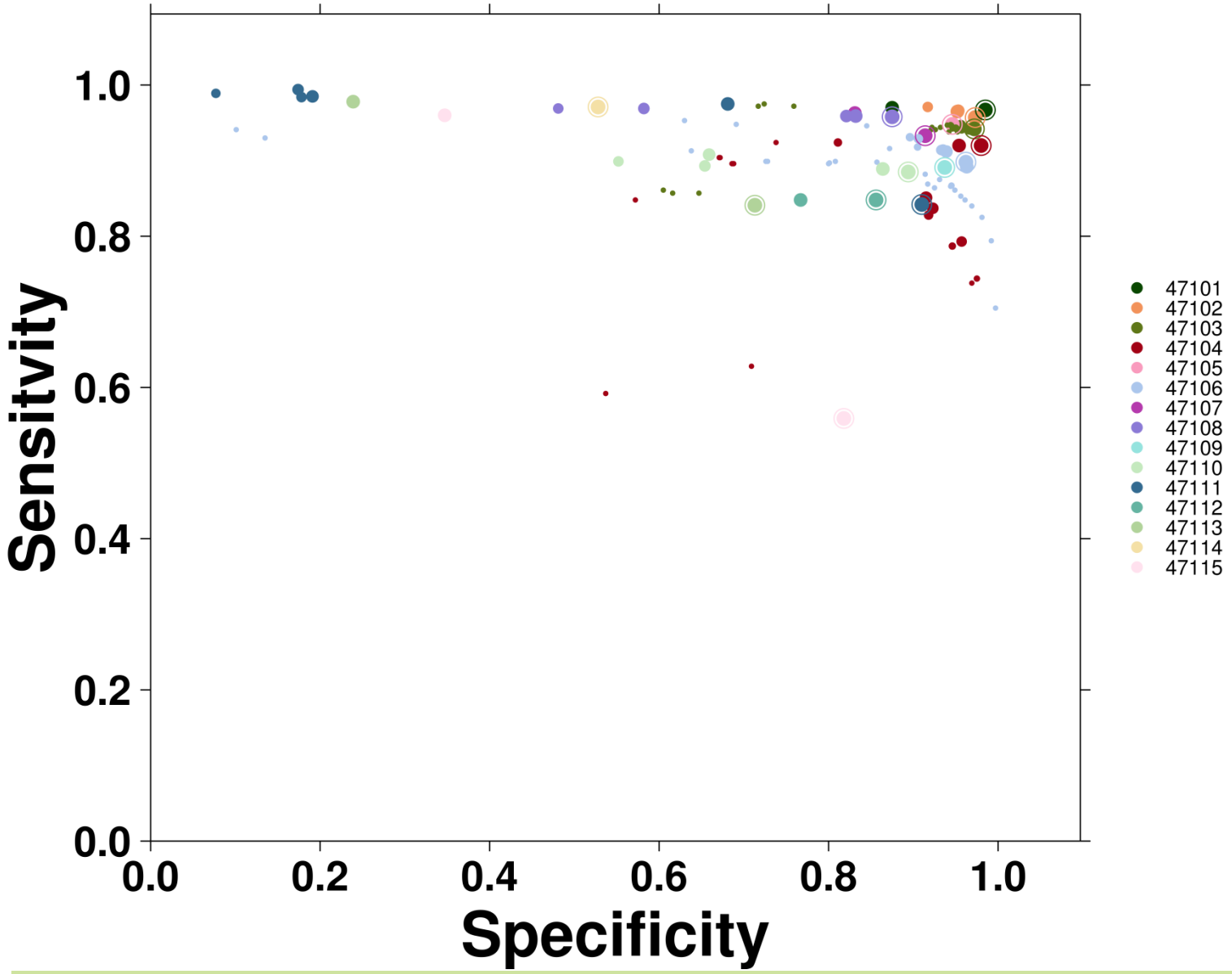
Initial Results

- So Far:
  - 268 registrants
  - 439 entries on 3 *in silico* genomes

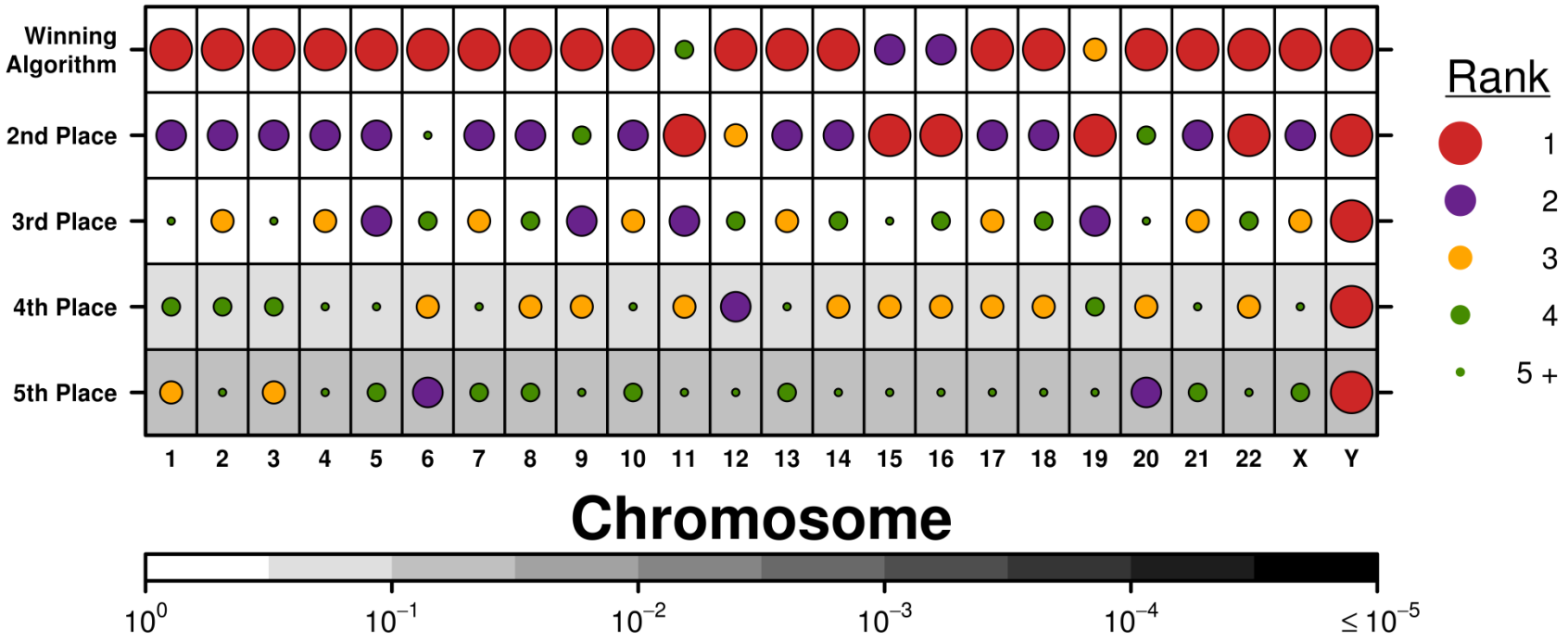
- On-going post-challenge submissions as people try to understand the failures of their algorithms (a *living* benchmark!)

- Key discussions on scoring SVs and on improving BamSurgeon (the simulator)

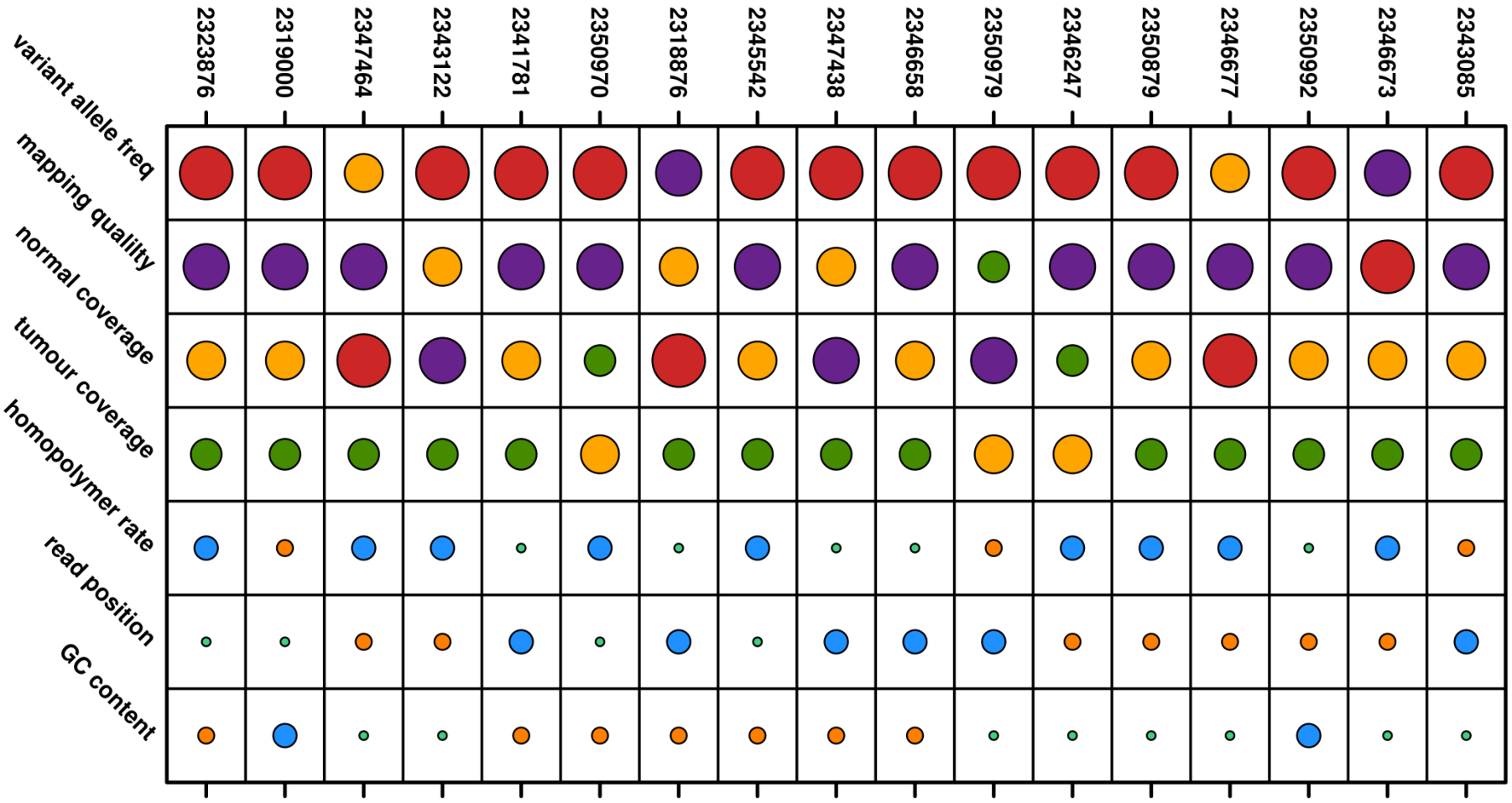
Entries Broadly Reflect a Single ROC Curve



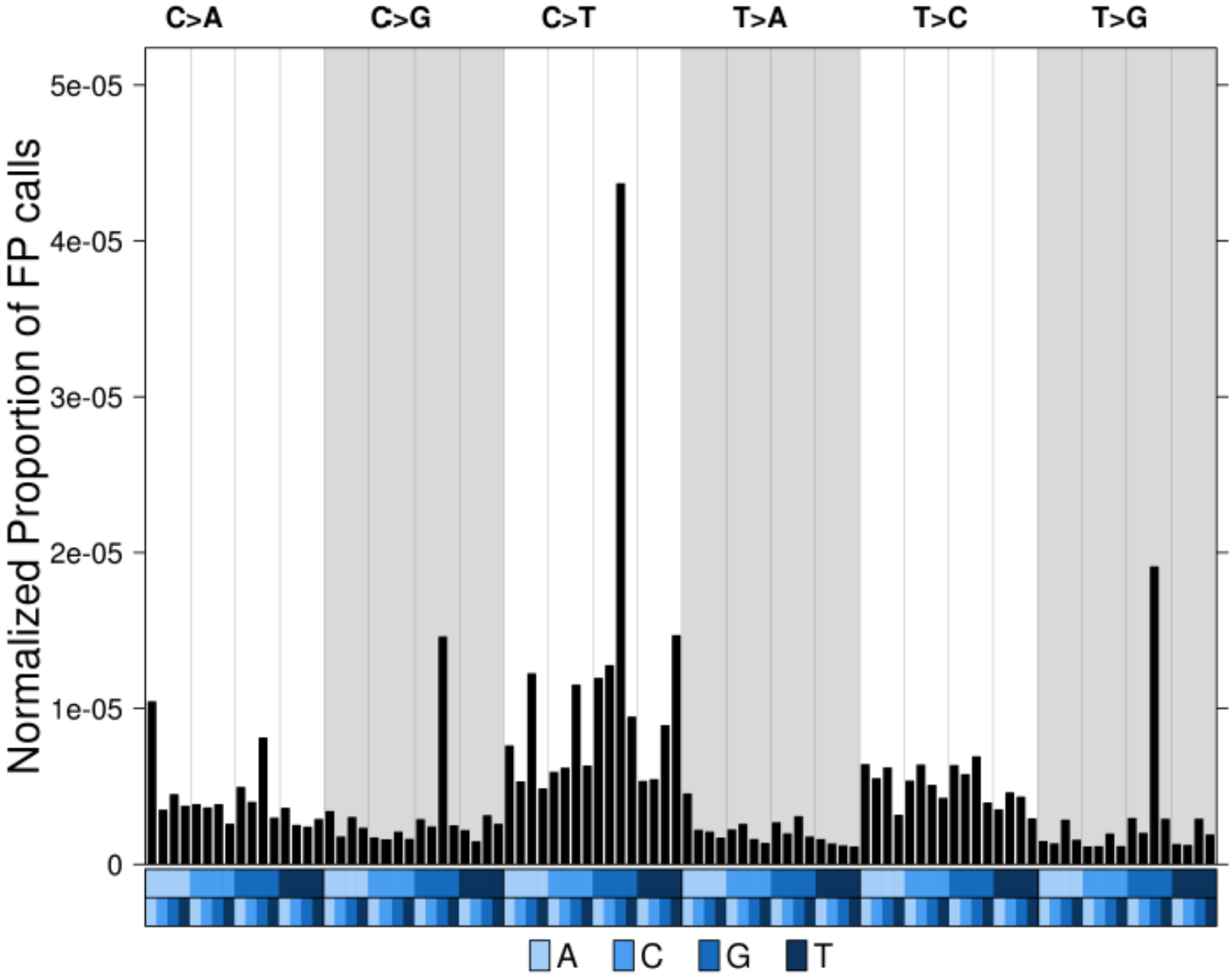
Surprisingly Large Chromosome-Bias



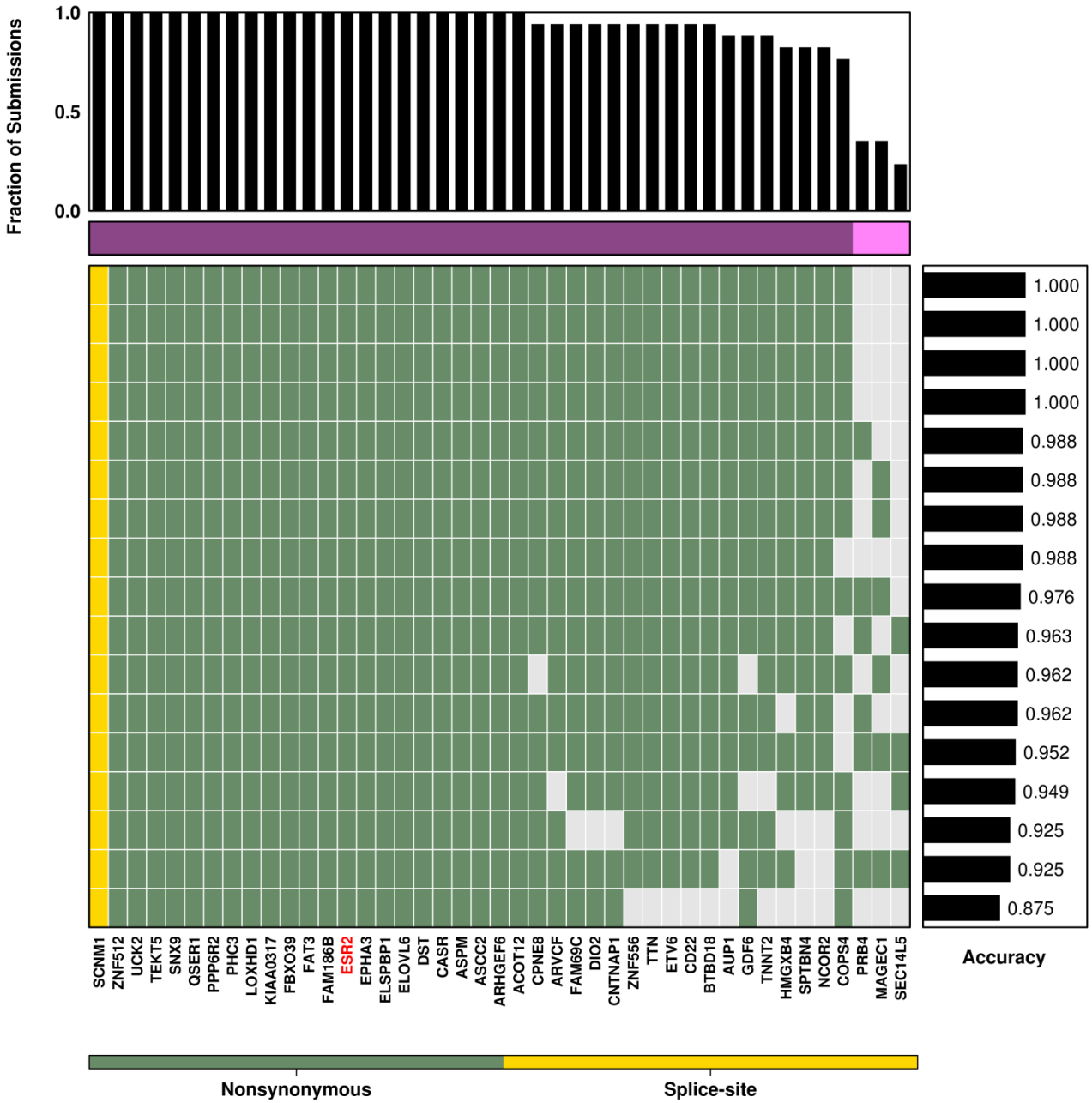
Different Determinants of Errors



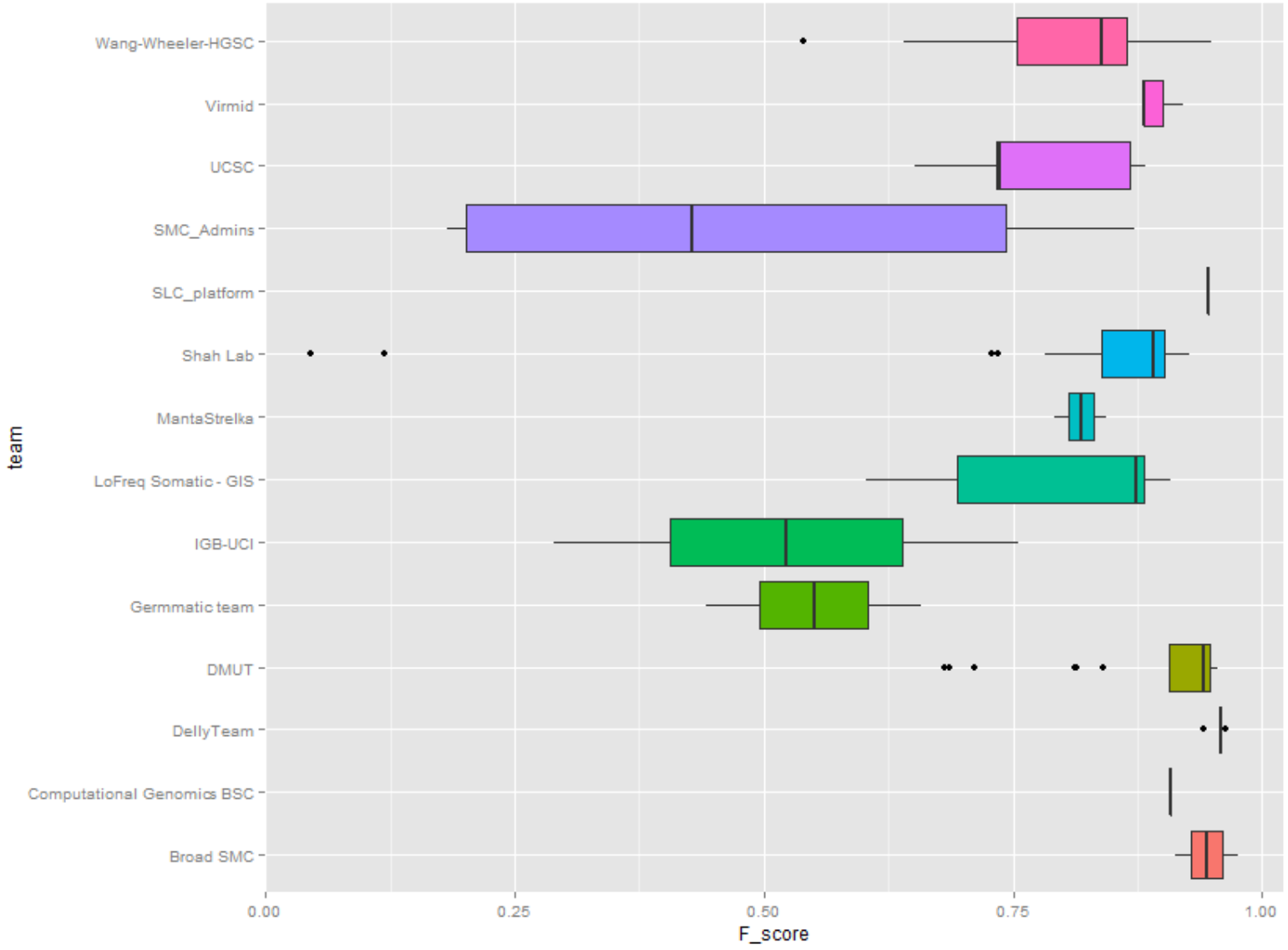
Surprisingly Strong Trinucleotide Effects



Coding Regions Had Lower Error Rates



Clearly Parameterization is Critical



In Summary: Results So Far

- Surprising trends in error-profiles:
  - Chromosomal Bias → trinucleotide bias
  - Normal coverage is “more important”?
- Identification of best methods for mutation prediction
  - SNVs: MuTect (IS1 and IS2)
  - SVs: Delly (IS1), novoBreak (IS2)
- Creation of a community for rapid algorithm-development and benchmarking for cancer NGS
- Improvement of tumour-read simulation

Pilot Surveys

Natalie Fox (grad-student, mRNA)

Dr. Maud Starmans (post-doc, mRNA)

Dr. Amin Zia (post-doc, CNAs)

Dr. Pablo Hennings-Yeomans (post-doc, GRs)

Richard de Borja (Bioinformatician)

Robert Denroche (Bioinformatician)

Challenge Organizing Team

Sage/DREAM Organizers

- Gustavo Stolovitzky
- Stephen Friend
- Adam Margolin
- Thea Norman
- Christine Suver
- Christopher Bare
- Kristen Dang
- Bruce Hoff
- Mike Kellen
- Yin Hu

External Organizers

- Paul C. Boutros (OICR)
- Josh Stuart (UCSC)
- Lincoln Stein (OICR)
- Kyle Ellrott (UCSC)
- Adam Ewing (UCSC)
- ***Katie Houlahan (OICR)***
- Cristian Caloian (OICR)
- Takafumi Yamaguchi (OICR)
- Andre Masella (OICR)

Data Contributors

Funding/Sponsoring/Publication Partners Include:

DO YOU WANT TO CHANGE THE ANALYSIS TENS OF THOUSANDS OF CANCER GENOMES? CAN YOU SET NEW STANDARDS?

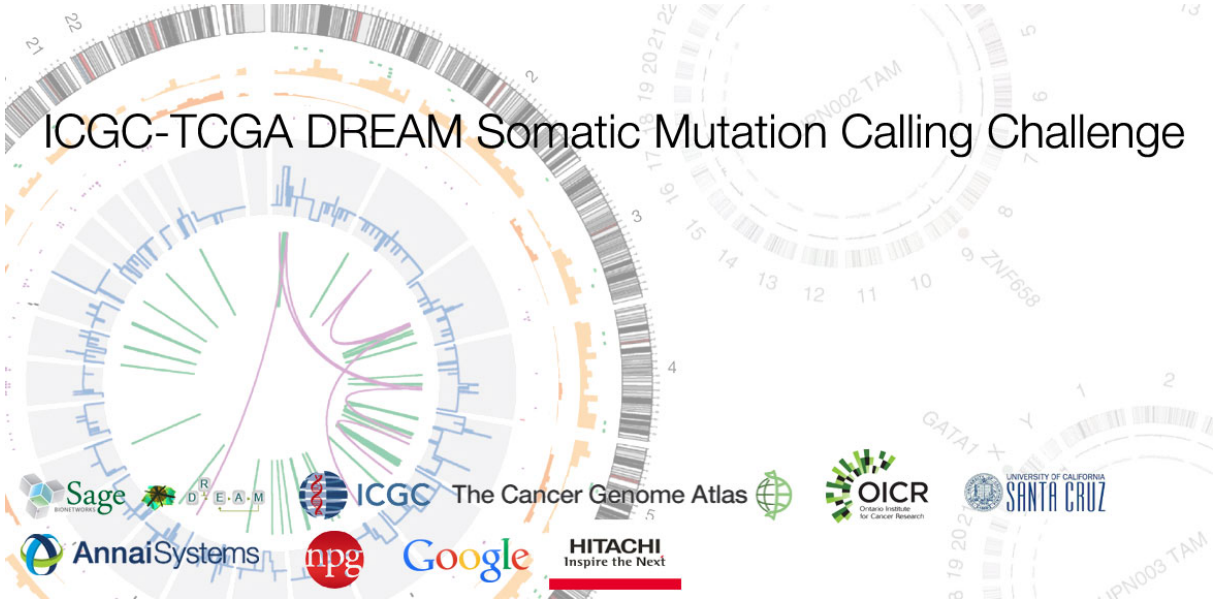
The ICGC-TCGA DREAM Somatic Mutation Calling Challenge

Registration open: NOW!

in silico data available: NOW!

Real data available: NOW!

Deadline #3: May 17!



SMC Challenge Website: https://www.synapse.org/ #!Challenges:DREAM