This version of the deck has notes like this in Times New Roman on some slides to explain what's going on for those who couldn't see it live!

Making Hay of Needles

Imran S. Haque
@imranshaque

16 Oct 2018
30th Anniversary AACR Special Conference --
Convergence: Artificial Intelligence, Big Data, and Prediction in Cancer
Newport, RI; #AACRBioSci18

There were a lot of talks here about successes of machine learning in biology. This will be a talk about the opposite: the ways in which current applications of machine learning and statistics **fail** when applied to biomarker discovery, and what to do about it: how we take a needle in a haystack and, by applying methods incorrectly, turn it into more hay.

Disclaimers

I’ve specialized from macromolecules on down, so this will come from a genomic/proteomic/molecular standpoint. I’d love to learn more from those of you who know about cytology and imaging.

All opinions here are my own and do not represent those of any past, present, future, or subjunctive employers.

Also, my opinions change a lot, so they may not even represent my own past or future opinions, except insofar as they were in my head when I wrote these slides.

Biomarker Discovery at a Crossroads

Mechanistic / target-driven

Sample acquisition is super expensive, let's use the fewest samples possible at each step.

1. Identify tumor signal or biological mechanism in cases.
2. Sequentially move to retrospective and prospective follow-up.
3. Pray that performance generalizes.

Empirical / data-driven

There are too many unknown unknowns in biology for us to form useful hypotheses upfront; let the data speak for itself.

1. Collect lots of data
2. ???
3. Success?

Reimagining the role of mechanism as providing constraints (answers) rather than hypotheses (questions) can help us bridge the data and understanding gap for empirical discovery.

Mechanistic Discovery

Case Study: ctDNA

Why Mechanistic Discovery?

- We think we know what's going on in cancer, and want to leverage that understanding.

Example: ctDNA

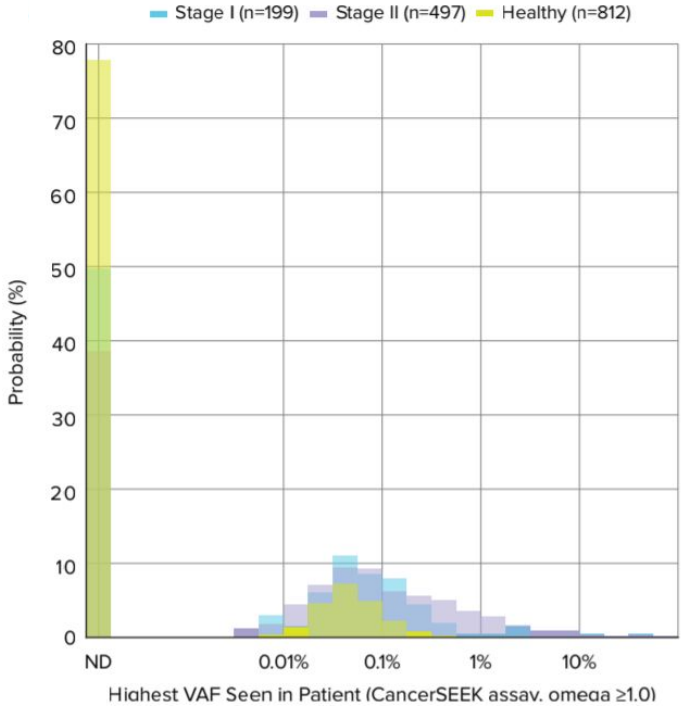
- Tumors have mutations, most of which normal cells shouldn't.
- Lots of cells shed DNA into the blood; so do tumors.
- Even if rare, perhaps we could specifically pick up tumor-derived mutations from patients with cancer.

Note: the fundamental reasoning here hinges on ctDNA being a highly **specific** biomarker because it is solely **tumor-derived**.

The critical parameter: tumor fraction

Tumor fraction: what fraction of the cfDNA actually comes from the tumor?

- Can be estimated by examining allele frequency of detected somatic mutations.
- Associated with stage: later stage usually means higher tumor fraction.
- TEC-Seq: ~50% of stage I and ~30% of stage II cancer patients have TF <0.1%
- CancerSEEK: ~50% of stage I and ~40% of stage II patients have TF < ~0.05%

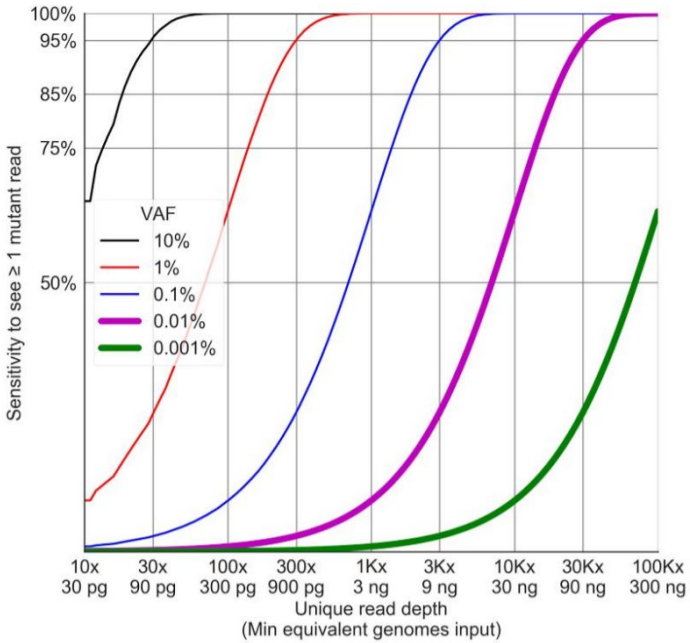


Note that the Cohen et al assay sequenced to higher unique coverage than TEC-Seq (Phallen), allowing them to assess lower VAFs than possible in Phallen, but that this also showed somatic heterogeneity - nonzero VAF in healthy individuals.

Haque IS and Elemento OE, *bioRxiv* 2017
Phallen J et al. *Sci Transl Med* 2017.
Haque IS et al. AACR 2018

Detection of rare events

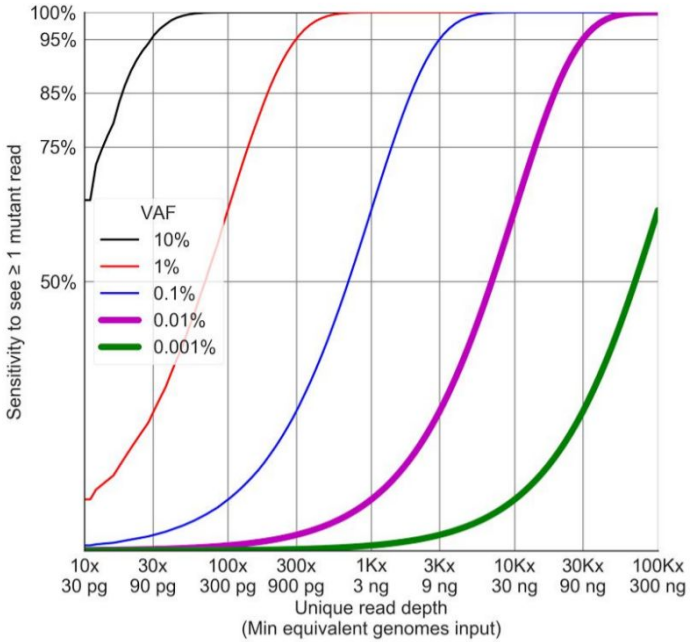
- The rarer the event, the more independent trials you have to sample in order to have high confidence of seeing it.



This plot shows the theoretical estimate, based on the binomial distribution, of how many **independent** (unique molecules) reads you'd need to detect one mutant read, as a function of mixture proportion. But ~no one has understood it from this plot alone, so let's talk analogies...

Haque IS and Elemento OE, *bioRxiv* 2017

Imagine that we would like a tour of Willy Wonka's factory. We know golden tickets are rare, so we stage a raid on a warehouse and get cases upon cases of candy bars. It's likely that somewhere in here, there will be a golden ticket!

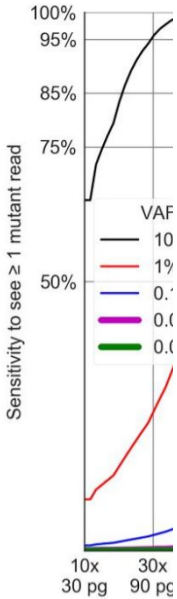


Haque IS and Elemento OE, *bioRxiv* 2017

ents

pendent

o have

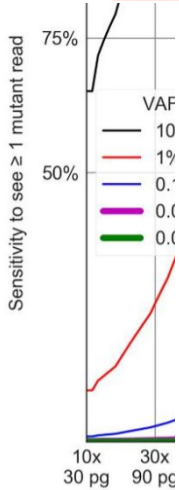


But we don't want to look through every single bar to find the ticket, so we summon Magneto for help.

Detection o ents

pendent

o have

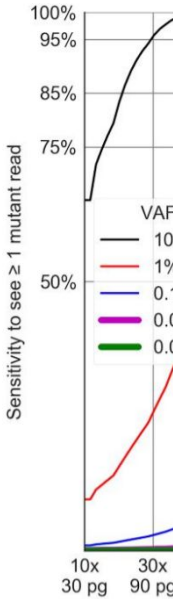


Magneto uses his power over metal to summon the one bar with a golden ticket. This is like enrichment in sequencing: we use hybrid capture or PCR to only pick out the fragments we care about (those in particular regions).

ents

pendent

o have

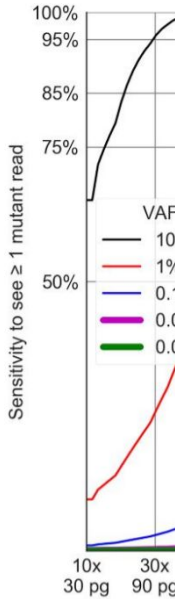


But if instead of a warehouse, we picked up all the chocolate at a single convenience store, it's likely that in that much smaller supply, there wouldn't be even one golden ticket.

ents

pendent

o have

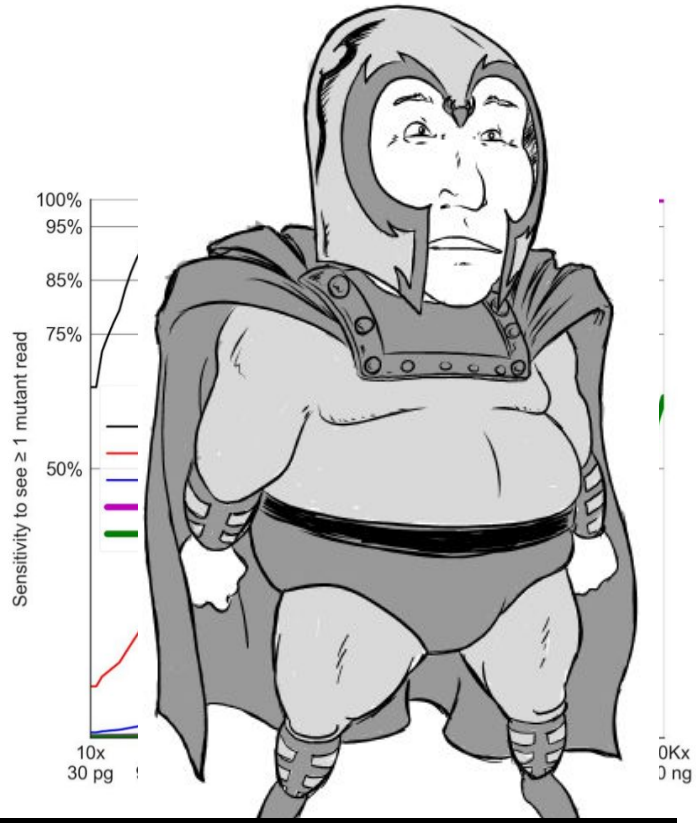


Since there's no ticket, there's nothing Magneto can do! In sequencing terms, if you don't collect enough blood to even have a single mutant molecule, no mutation enrichment strategy or background depletion strategy could help you.

ents

pendent

o have

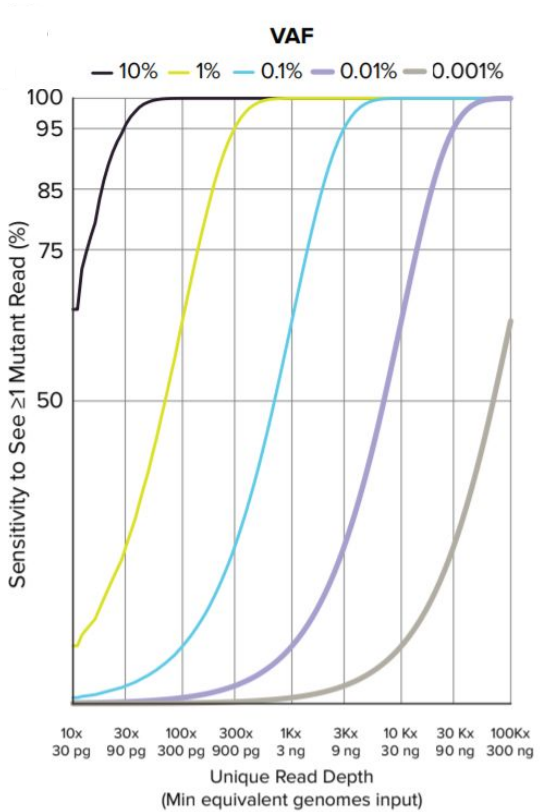


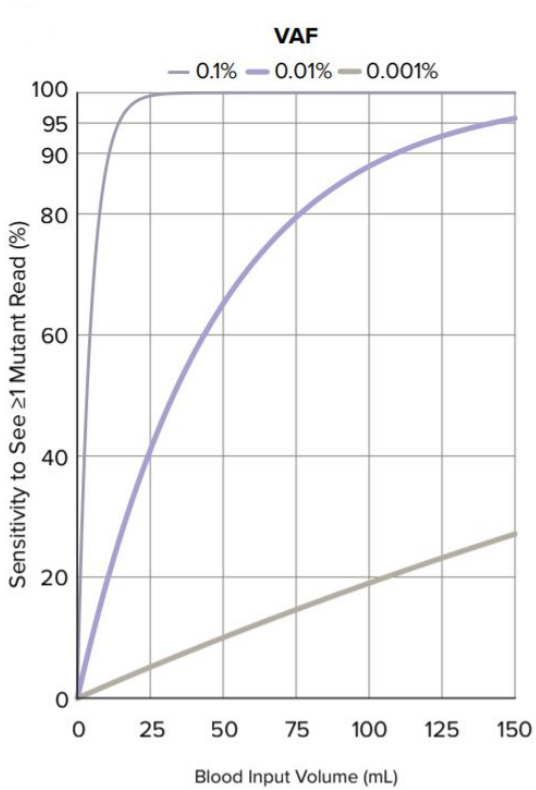
And then there would be nothing left to do but to eat all the candy and get fat (or waste a lot of money on sequencing).

https://sbcarnival.wordpress.com/2012/01/29/fat-magneto/

Detection of rare events

- The rarer the mutation, the more independent molecules you have to sample in order to have high confidence of seeing it.
- To have a test with <5% failure rate, we can only count on 2.3 ng cfDNA/mL plasma = ~770 genomes/mL
- At (really really high) efficiency of 50%, need ~80mL blood draw to detect 1 molecule at 0.01% VAF - and you probably want more than 1.

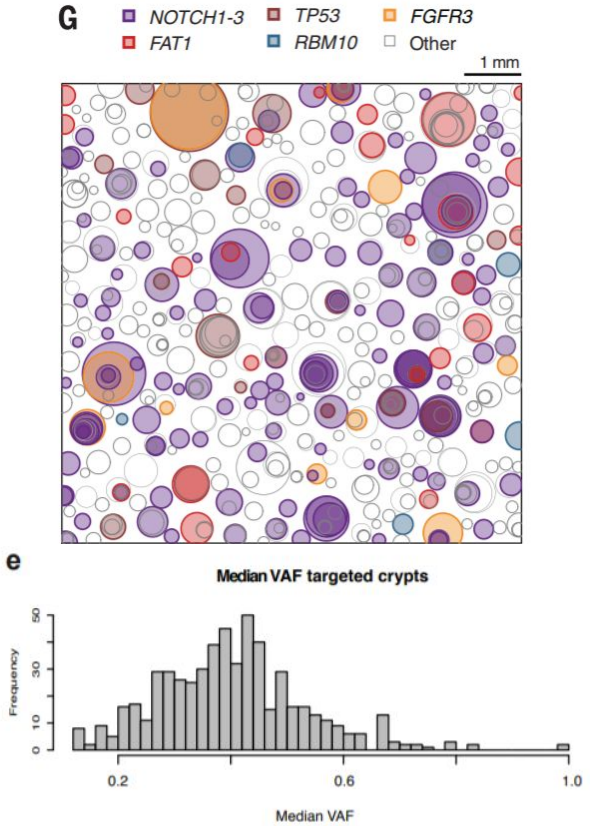




Haque IS and Elemento OE, *bioRxiv* 2017
Haque IS et al. AACR 2018

Multi-site assays and somatic heterogeneity

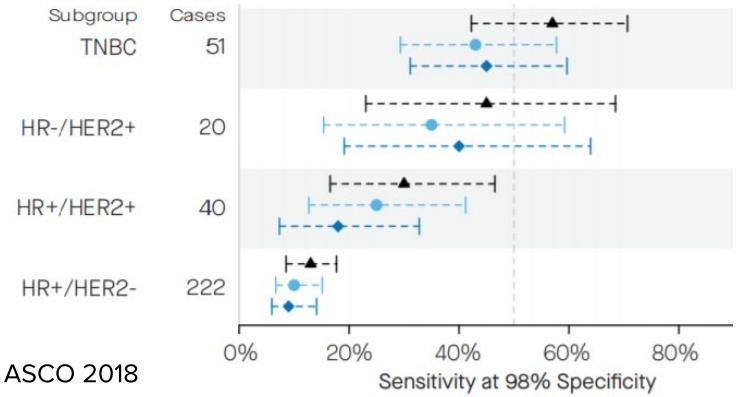
- Looking at multiple sites could help: if independent, VAFs add (10 sites @ 0.01% ~ 1 site @ 0.1%).
- Somatic heterogeneity appears to be the natural state of even healthy tissues, with age dependence. 1% of healthy colon crypts carry cancer driver mutations.
- Too narrow: need too much blood
  Too broad: compromised specificity



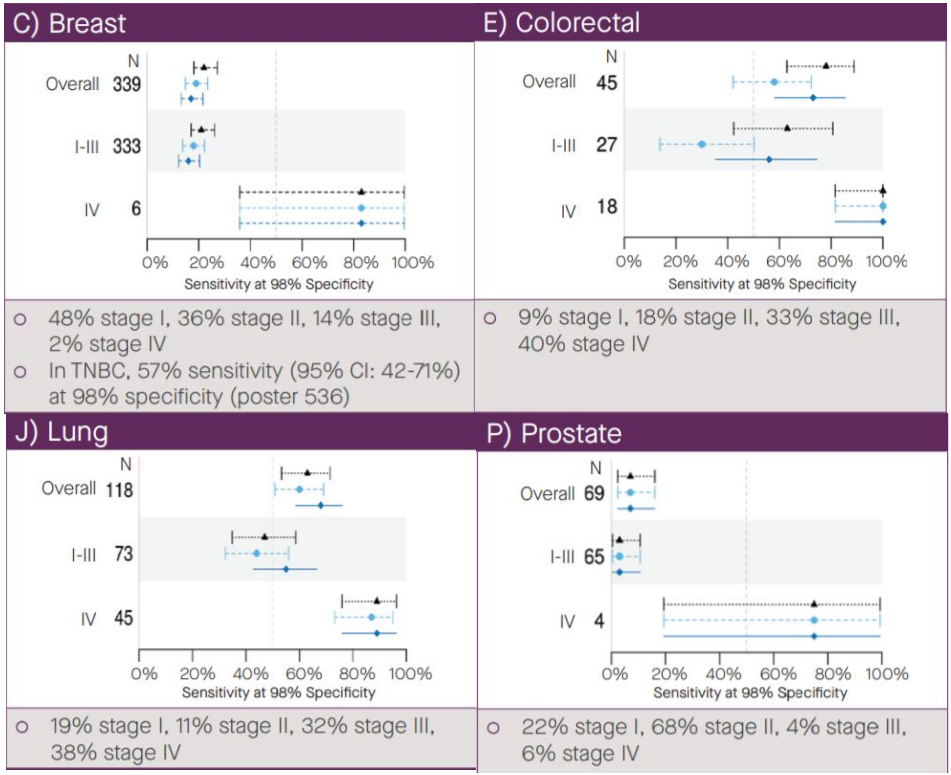
Martincorena I et al *Science* 2015
Lee-Six H et al, *bioRxiv* 2018

Real world evidence: TF matters

- GRAIL has reported ctDNA data from 1785 patients (~3000x unique depth).
- Strong stage dependence: cancers with more stage I tend to perform much worse: suggests **tumor fraction is a real, fundamental limitation.**



Liu MC et al. ASCO 2018



Klein E et al. ASCO 2018

Summary: Mechanistic Discovery

Exciting prospect: mechanism-driven process that should deliver highly specific and potentially sensitive biomarkers for cancer.

New discoveries along the way that potentially constrain specificity.

Fundamental physi(ologi)cal limitations appear to constrain sensitivity.

It’s never quite as rosy as it starts.

Empirical Discovery

Case Study: circulating proteins

Why Empirical Discovery?

- We think we don't know everything that's going on in {CONDITION}, and want to be (more) hypothesis-agnostic.

Example: multi-protein biomarkers

- We think there may be various protein markers coming from the tumor or from systemic responses (e.g., immune).
- We don't know exactly how these would be perturbed; might be a combination of changes from complex/systems biology.
- We'll use statistics on large cohorts to discover these changes and learn biology.

Two case studies: PLCO and CancerSEEK.

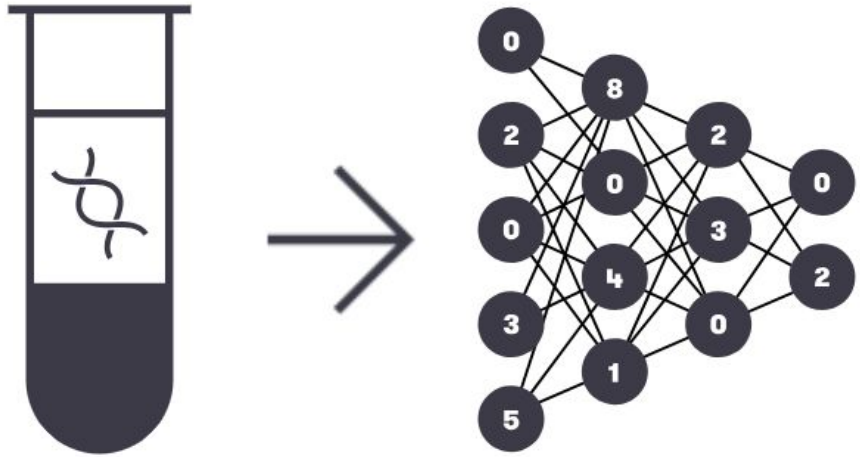
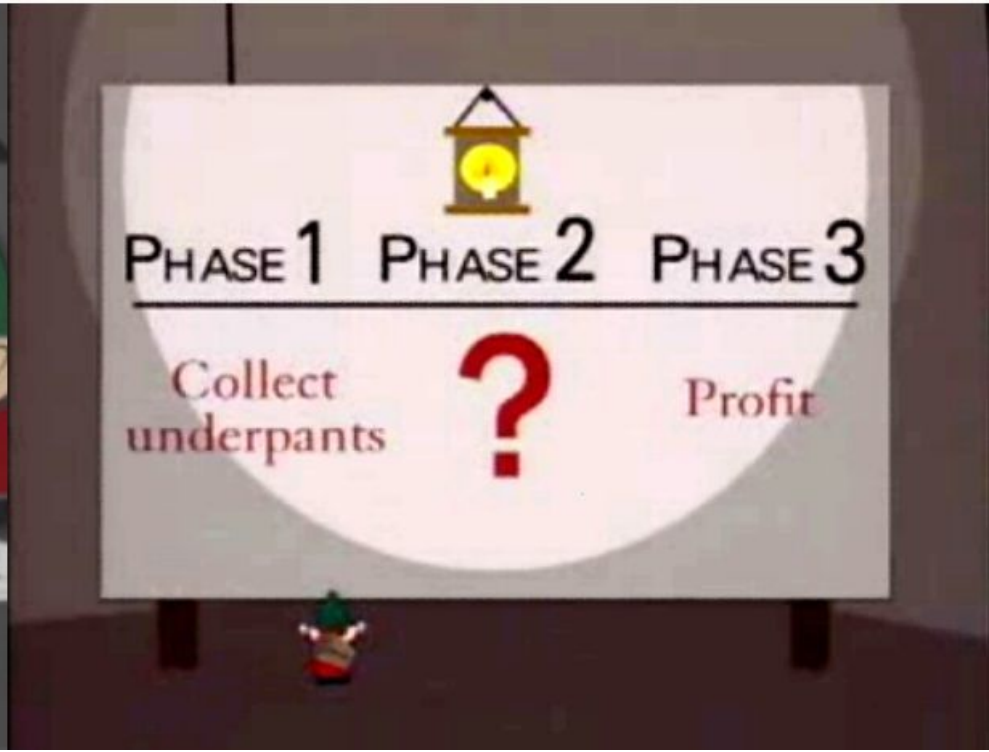


Image from freenome.com

Methodology

1. Pick a high-content assay
   (protein array, mass spec, aptamers, panel ELISA, NGS...)
2. Collect “a lot” of samples.
3. *wave hands vigorously*
4. Biomarker!

Methodology



Methodology

1. Pick a high-content assay
   (protein array, mass spec, aptamers, panel ELISA, NGS…)
2. Collect “a lot” of samples.
3. **Do MACHINE LEARNING**
4. Biomarker!

Halevy et al. *IEEE Intel Sys* 2009

Methodology

1. Pick a high-content assay
   (protein array, mass spec, aptamers, panel ELISA, NGS…)
2. Collect “a lot” of samples.
3. **Do MACHINE LEARNING**
4. Biomarker!

The Unreasonable Effectiveness of Data

“A trillion word corpus…captures even very rare aspects of human behavior. So, this corpus could serve as the basis of a complete model for certain tasks - if only we knew how to extract the model from the data.”

Halevy et al. *IEEE Intel Sys* 2009

Methodology

1. Pick a high-content assay
   (protein array, mass spec, aptamers, panel ELISA, NGS…)
2. Collect “a lot” of samples.
3. **Do MACHINE LEARNING**
4. Biomarker!

People aren’t web pages; sample processing is expensive.

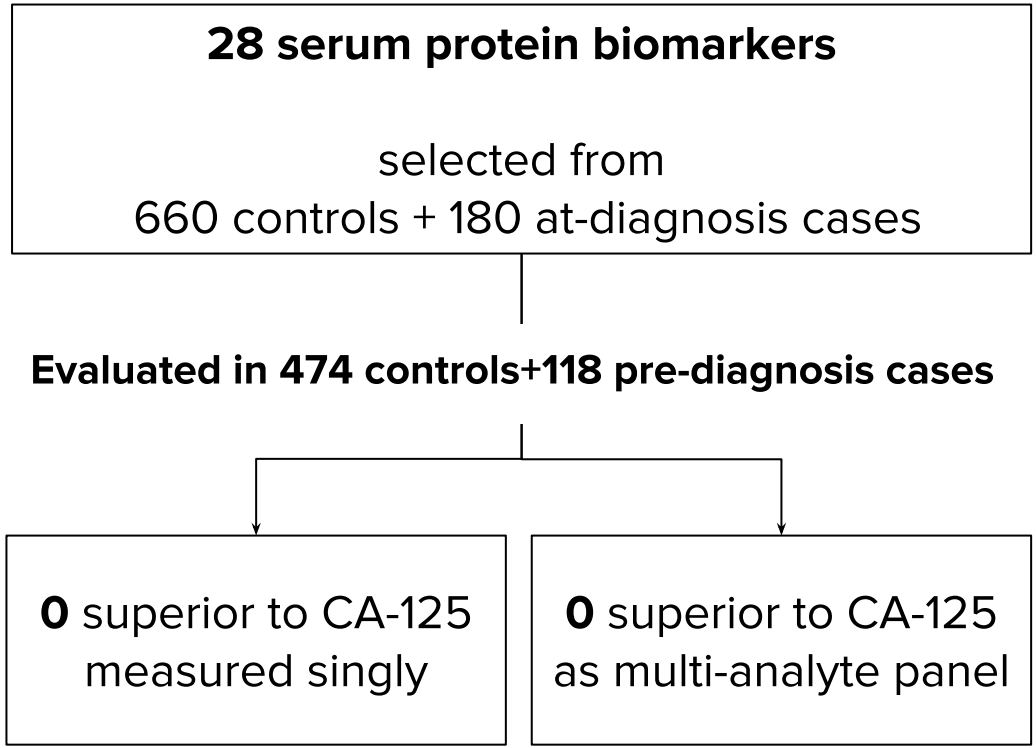
This is usually done stepwise: enriched case-control cohorts for marker discovery, sequentially larger “validation” cohorts.

The Unreasonable Effectiveness of Data

“A trillion word corpus…captures even very rare aspects of human behavior. So, this corpus could serve as the basis of a complete model for certain tasks - if only we knew how to extract the model from the data.”

Halevy et al. *IEEE Intel Sys* 2009

The PLCO experience



It is frustrating that none of the 28 ovarian cancer serum biomarkers…were shown, when evaluated singly, to have test performance characteristics that were equal, let alone superior, to CA-125 levels [in prediagnostic serum samples].

Furthermore…multianalyte…combinations of biomarkers did not improve test performance measures compared to CA-125 alone.

Mai et al. *Cancer Prev Res* 2011

Why?

Technical variability

“Markers whose assays had poor CVs also had poor performance as biomarkers”

Population variability

Systematic differences may be present between cases and controls.

Biological variability

Post-diagnosis != pre-diagnosis
Screening finds different disease categories.

Non-independence

Just because each thing (may) work alone, doesn’t mean combinations will work better.

Cramer et al. *Cancer Prev Res* 2011
Jacobs and Menon *Cancer Prev Res* 2011
Mai et al *Cancer Prev Res* 2011
Sun et al. *Nature 2018*

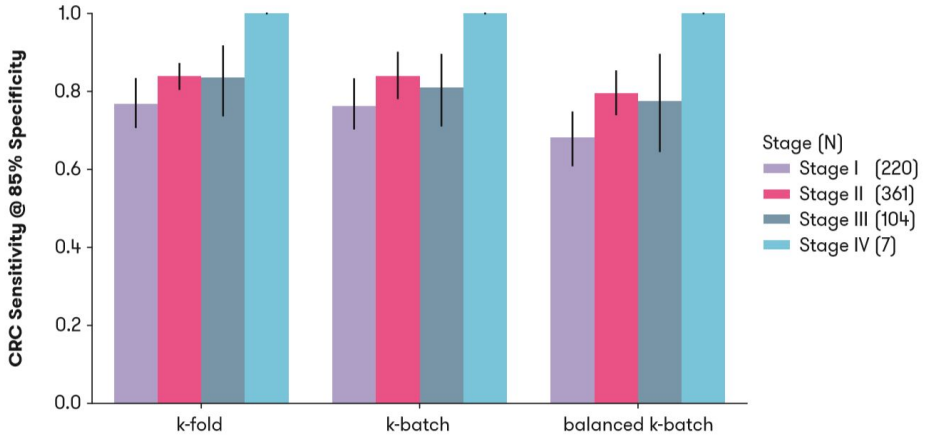
Why?

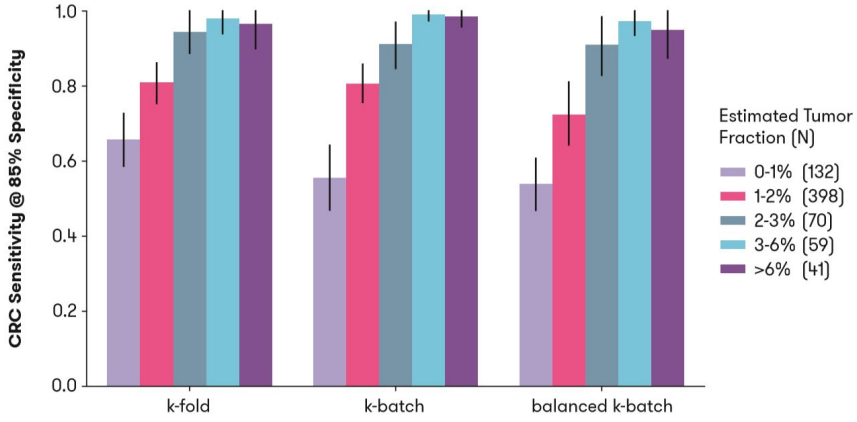
Technical variability

"Markers whose assays had poor CVs also had poor performance as biomarkers"

Data from a poster presented by Freenome at ACG 2018 compared the estimated performance of a machine learning method under three different validation schemes. The left bar is k-fold, in which "test" sets are constructed completely at random from the set of all samples. The middle stratifies these test sets by batch: any one processing batch worth of samples only shows up in training *or* in test, never both; this way, you can test for sensitivity to batch effects. The right set of bars further subsets samples by institution of origin to assess sensitivity to this variable.

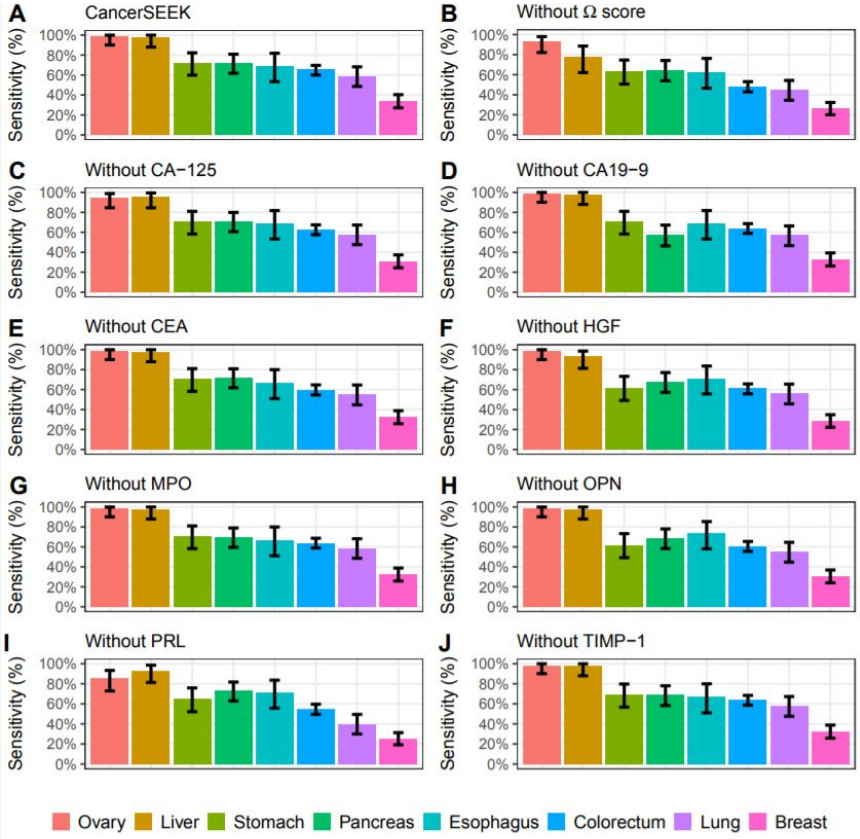
The middle and right sets of bars appear to perform worse than the left, suggesting that basic k-fold validation overestimates performance in the presence of technical biases typical in genomics.





Niehaus et al. ACG 2018

All this has happened before and will happen again



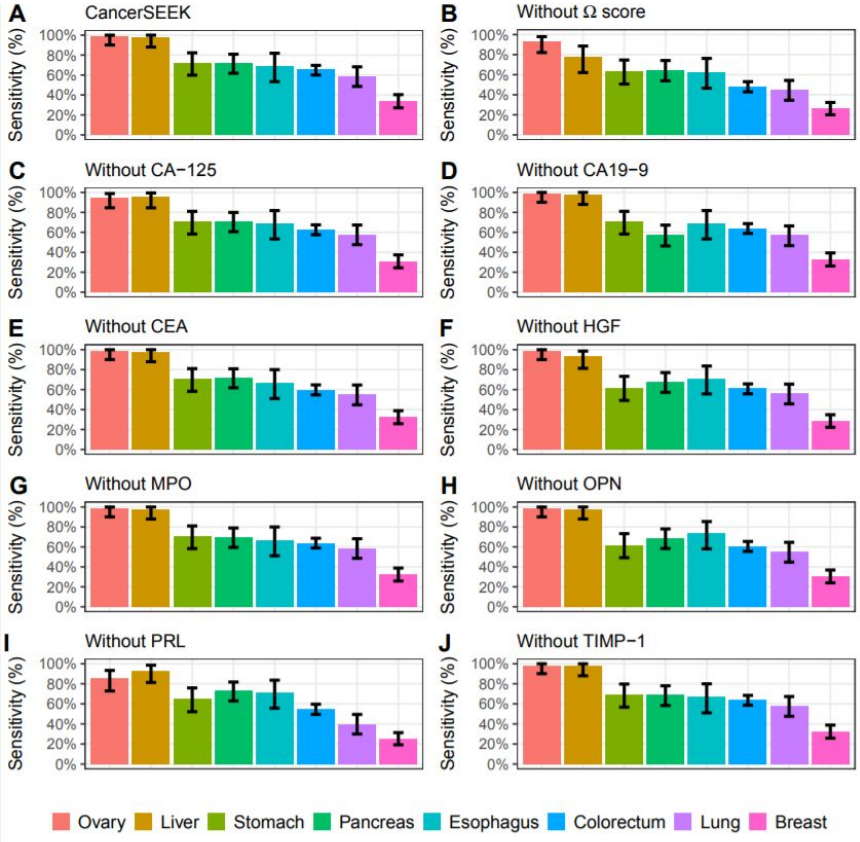
Note that panel B (from the supplementary data to Cohen 2018) suggests that removing ctDNA entirely from the CancerSEEK assay leaves assay performance largely intact: either most of the power is coming from the proteins, or all the assays are measuring similar things (are non-independent).

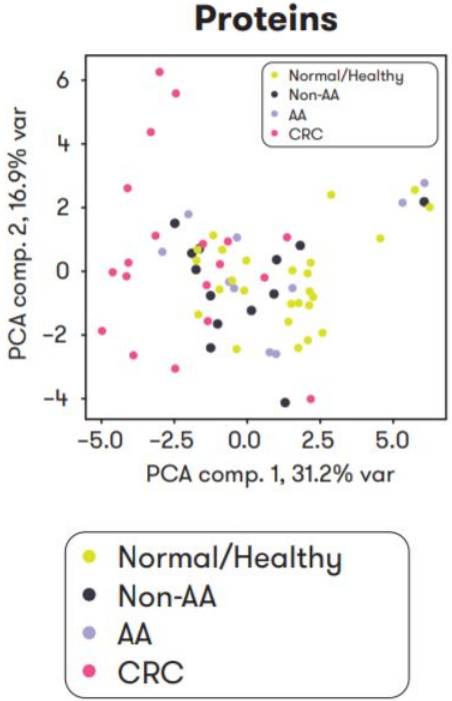
Non-independence

Just because each thing (may) work alone, doesn't mean combinations will work better.

Cohen et al. *Science* 2018
Delubac et al. AACR 2018

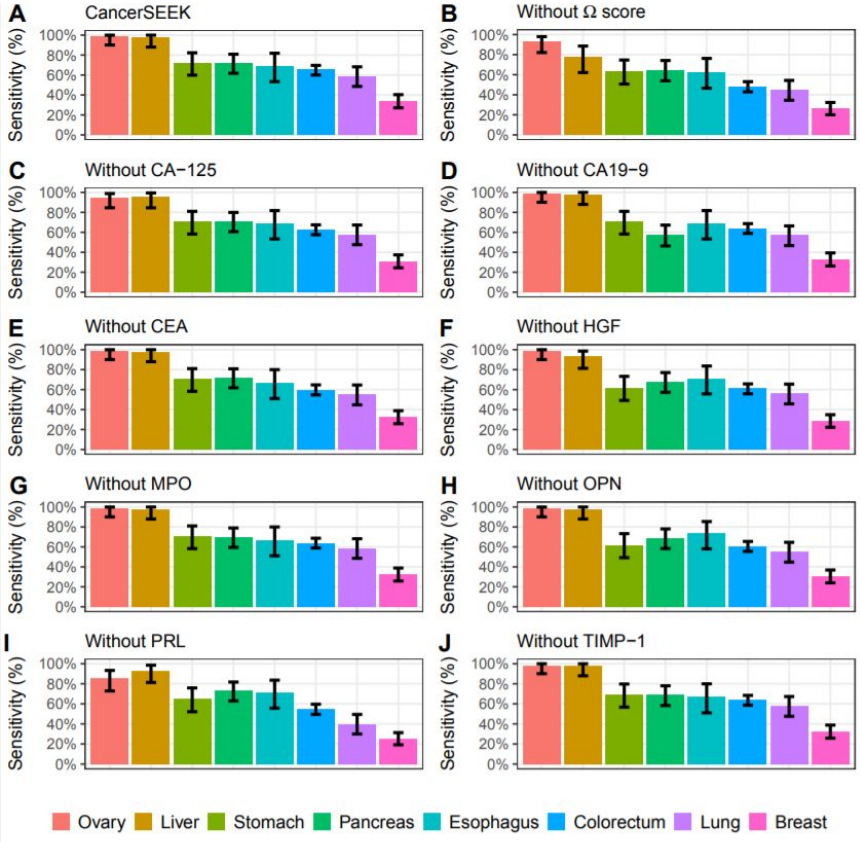
All this has happened before and will happen again

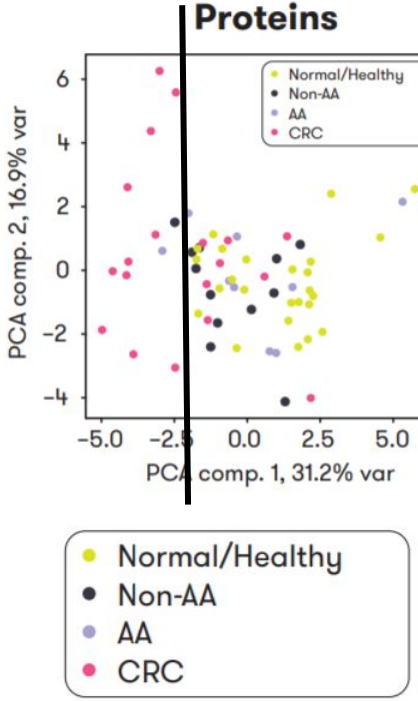




Cohen et al. *Science* 2018
Delubac et al. AACR 2018

All this has happened before and will happen again

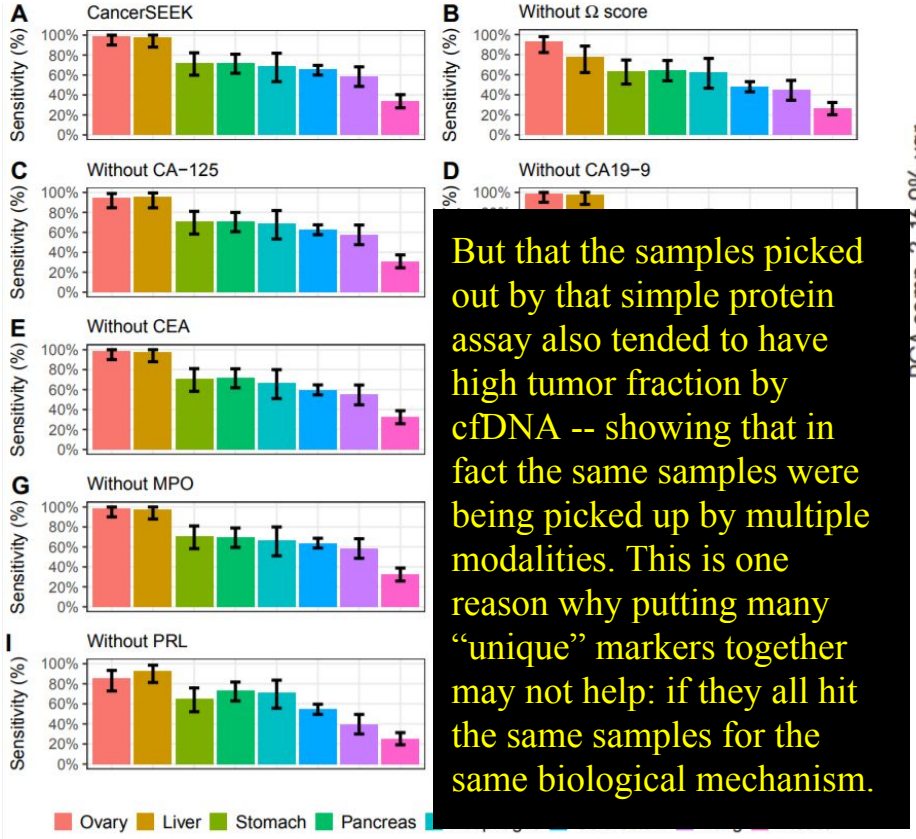




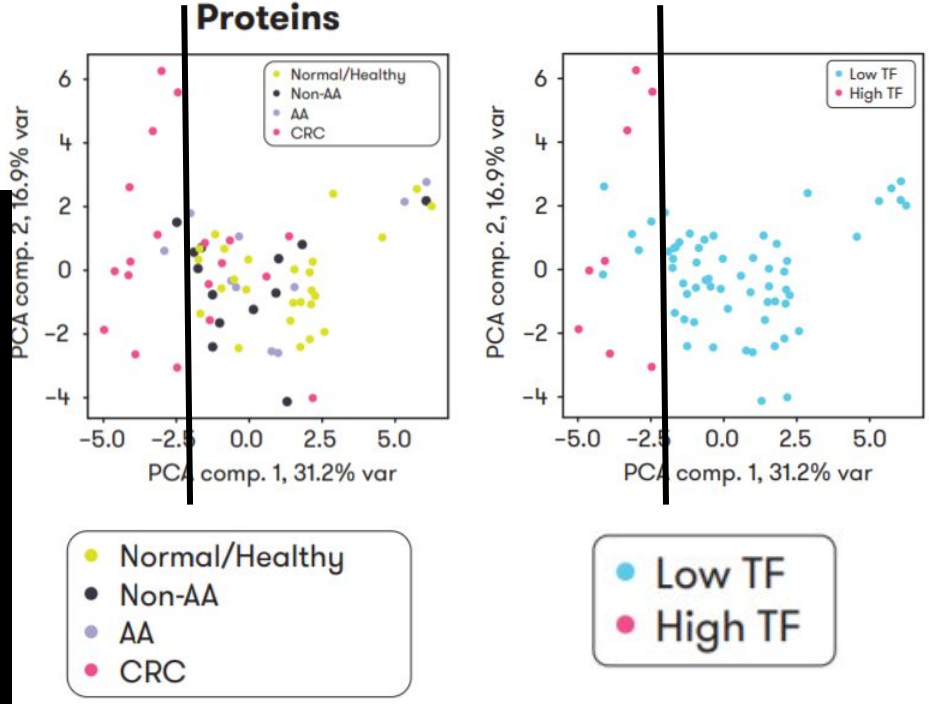
Work from another Freenome poster, presented at AACR by Delubac et al showed that in a small cohort of samples, it was possible to design a reasonable cancer detector using proteins alone...

Cohen et al. *Science* 2018
Delubac et al. AACR 2018

All this has happened before and will happen again



But that the samples picked out by that simple protein assay also tended to have high tumor fraction by cfDNA -- showing that in fact the same samples were being picked up by multiple modalities. This is one reason why putting many "unique" markers together may not help: if they all hit the same samples for the same biological mechanism.



Cohen et al. *Science* 2018
Delubac et al. AACR 2018

Summary: Empirical Discovery

Exciting prospect: automatic methods to combine known and unknown markers to boost their performance, without constraint of known mechanisms.

Statistical methods require more data than they appear at first, and require extreme rigor in defining the question you'd like to ask (screening is not diagnosis!)

Field hasn't done a great job internalizing the lessons of the past: cost of sample accrual remains fundamental problems that keeps getting dodged.

It's never quite as rosy as it starts.

Future Directions

Scientists are from Gryffindor, Machines are from Slytherin

Recap

Mechanism: can be relatively cheap (we think we know what we're looking for from samples and in samples, maybe, sorta). Leaves us high and dry if the mechanism just doesn't quite work (unknown biology, physical limitations, etc.).

Empiricism: Tantalizing, but unclear how to make it compatible with the economics of discovery in rare conditions (with possible exception of common-variant GWAS).

Recap

We'd love empirical discovery and machine learning to work like Hermione -- wave a magic wand and the problem goes away.

ink we know what we're looking for from

aves us high and dry if the mechanism

hysical limitations, etc.).

make it compatible with the economics

le exception of common-variant GWAS).

Recap

But instead, in biomarker discovery machine learning more often behaves like Ron.

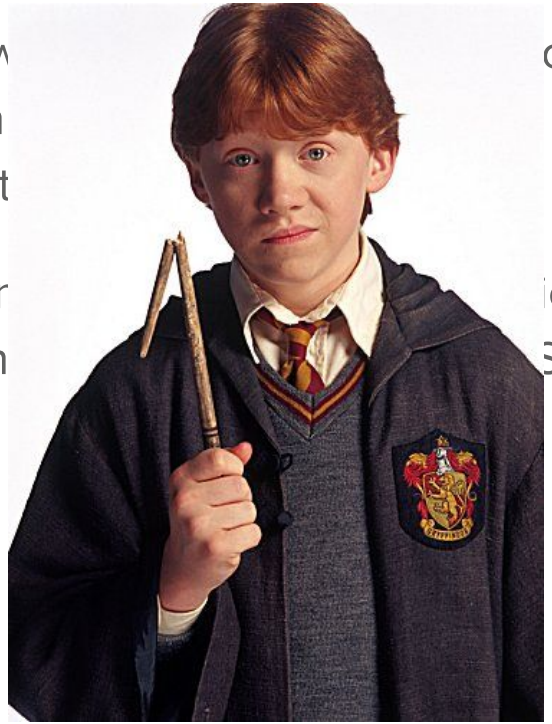
ink we know om

aves us high

hysical limit

make it cor ics

le exception S).



But even that’s not quite right. I’d argue that ML methods are more like Gilderoy Lockhart: big fakers, unless you can pin them down...

ML: The Gilderoy Lockhart of Methods

Machine learning algorithms try their hardest to be good ~~cheaters~~ fakers:

- No test set? You **will** overfit to your training set. (This is everywhere, btw.)
- Didn't stratify your cross-validation? You **will** overfit to hidden covariates.
  - Ancestry. Batch. Sample processing. Operator. Phase of the moon.
- Filtered features or hyperparameters without a second-level hold out? You know where this goes.

A useful heuristic: statistical methods will always take the easiest way out to the answer you "want": right for the wrong reasons is still right (in your small data set).

Designing For ML: Sherlock Holmes

A useful corollary: instead of designing an discovery project to work, design it to **not fail** -- specifically design around all the ways to cheat.

> *“When you have eliminated the impossible, whatever remains, however improbable, must be the truth”*

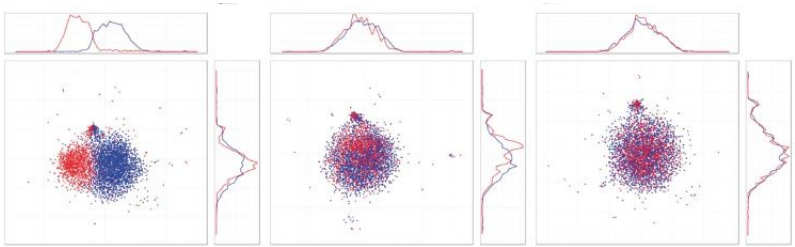
Mechanism allows us to define negative and positive controls - which provide **invariants** that you can enforce on or teach to a model.

Invariants through Mechanisms

Batch Effects

Technical replicates can give us a lot of information that shouldn't be just averaged out: bias is intrinsic.

We can train ML models to strongly reduce bias when they are given this as a constraint:



Data Augmentation

If you don't have technical replicates, maybe you know how to fake them well enough?

SOP in computer vision: crop, rotate, add noise, deform, ...

Shaham U et al. *Bioinformatics* 2017
Shaham U et al. *bioRxiv* 2018
Wang and Perez, *arXiv* 2017

Conclusions

- The conflict between mechanistic and empirical biomarker discovery boils down to tradeoffs between cost/sample acquisition and completeness.

- Physical limitations constrain the current hottest method in mechanistic discovery (ctDNA); technical and biological sampling limitations have constrained empirical methods for decades.

- Retargeting biological machine learning from end-to-end discovery to a focus on these sub-problems, by using mechanism to define invariants that can help the empirical sampling problem, may be a way to resolve the conflict.

Questions? Reach me on Twitter at @imranshaque or by email at ihaque@cs.stanford.edu.