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!

The organizers asked me to give a "provocative" talk that would force the audience to consider how they might need to think differently in order to really make a difference in cancer discovery and therapy.

See the accompanying blog post at
https://ihaque.org/posts/2019/05/07/barriers-to-entry-barriers-to-validation

We Are Legion

(yes, the title is a reference to the character from Mass Effect 2.)

Statistics and Generalization from Cells to Populations

Imran S. Haque, PhD

29 Apr 2019
3rd CRUK International Symposium on Oesophageal Cancer
London, UK

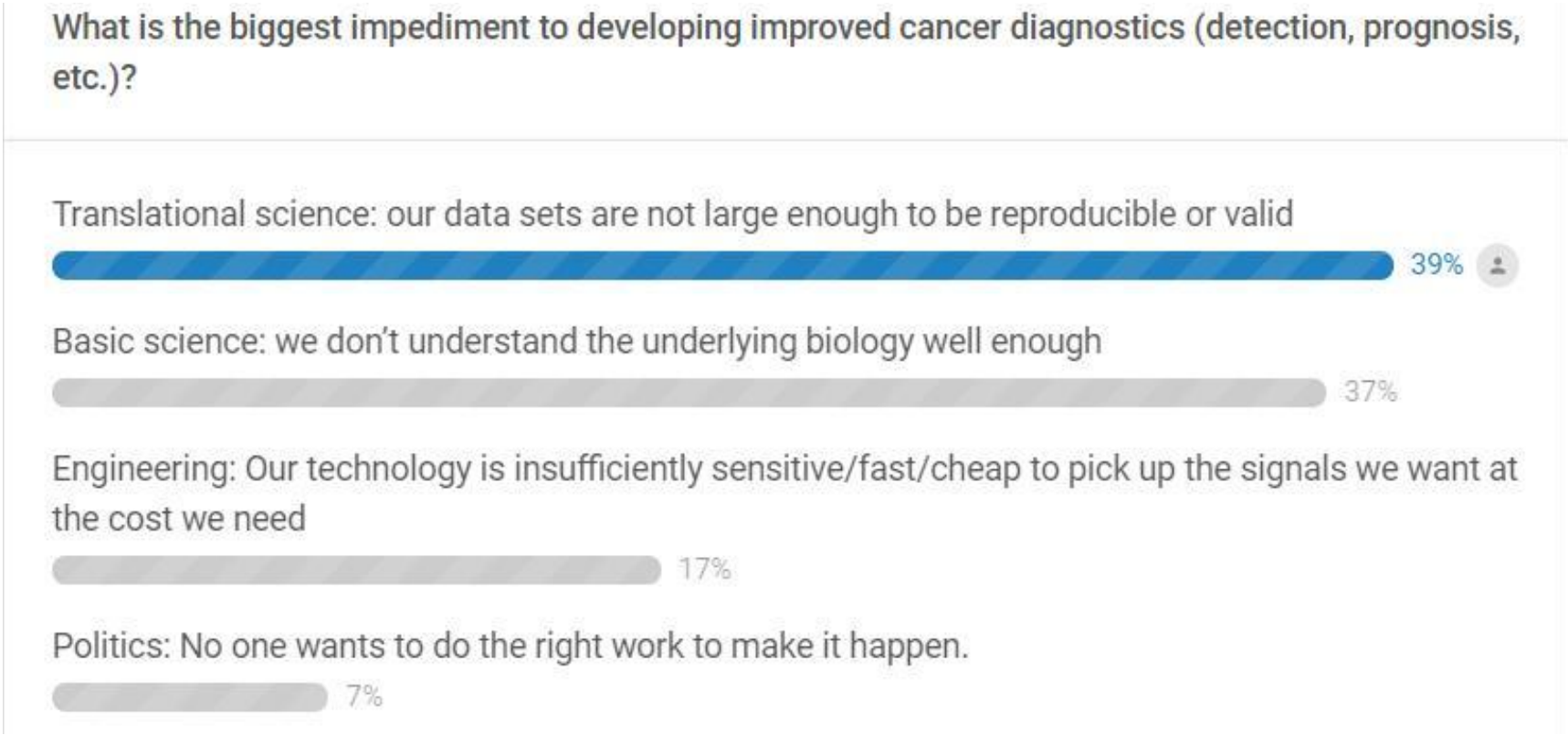
Twitter: @imranshaque
https://ihaque.org
ish@ihaque.org

Slido

What is the biggest impediment to developing improved cancer diagnostics (detection, prognosis, etc.)?

1. Basic science: we don't understand the underlying biology well enough.
2. Translational science: our data sets are not large enough to be reproducible or valid.
3. Engineering: Our technology is insufficiently sensitive/fast/cheap to pick up the signals we want at the cost we need.
4. Politics: No one wants to do the right work to make it happen.

Slido Poll Results



The Two Cultures of Biological Modeling

Mechanistic / deductive / target-driven

Sample acquisition is super expensive, let's use the fewest samples possible at each step by deriving from known mechanisms.

1. Identify/validate mechanism in cases alone.
2. Sequentially move to more rigorous controlled cohorts (e.g., retrospective and prospective follow-up).
3. Hope that model performance generalizes.

Empirical / inductive / 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!

The Two Cultures of Translation

Academic / discovery-driven

What's interesting is finding something interesting.

1. Identify interesting biological finding using method of choice.
2. Publish finding, write subsequent grant.
3. Use funding from (2) to repeat.

Indsutrial / scale-driven

What's interesting is to take something interesting and make it boringly reliable.

1. Take preliminary finding from literature, expand and stabilize with internal dev.
2. Scale up, bring to market.
3. Use profits from (2) to repeat.

Considering both the scientific challenges of discovery as well as the political challenges of organization and funding will be critical to next-generation cancer diagnostics and therapeutics.

Mechanistic Modeling and 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

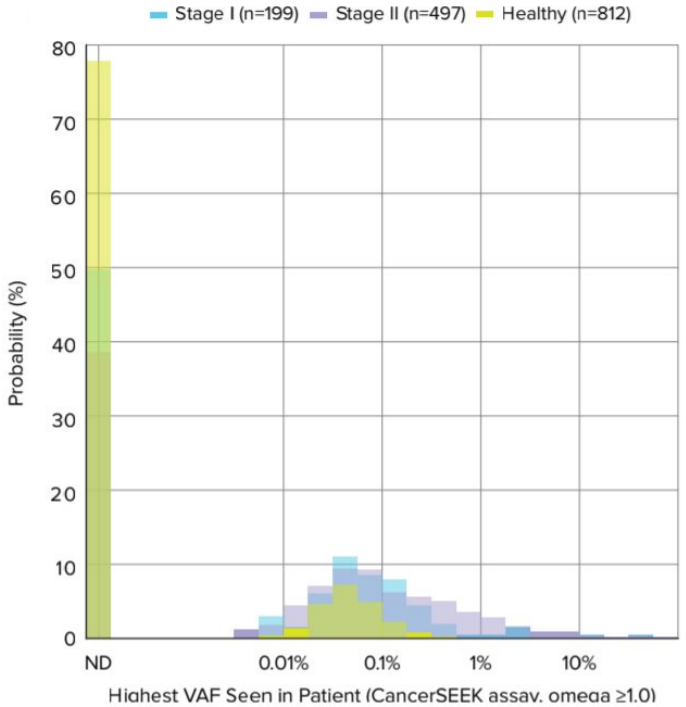
- Tumors have mutations, most of which normal cells shouldn't.
- Lots of cells shed DNA into the blood; so do tumors.
- Even if rare, perhaps we could specifically pick up tumor-derived mutations from patients with cancer.

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%

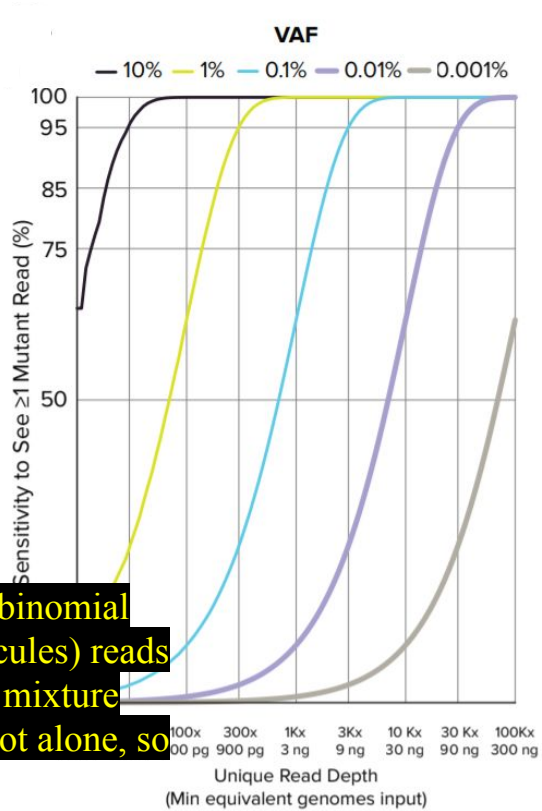


Note that the Cohen et al assay sequenced to higher unique coverage than TEC-Seq (Phallen), allowing them to assess lower VAFs than possible in Phallen, but that this also showed somatic heterogeneity - nonzero VAF in healthy individuals.

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

Detection of rare events

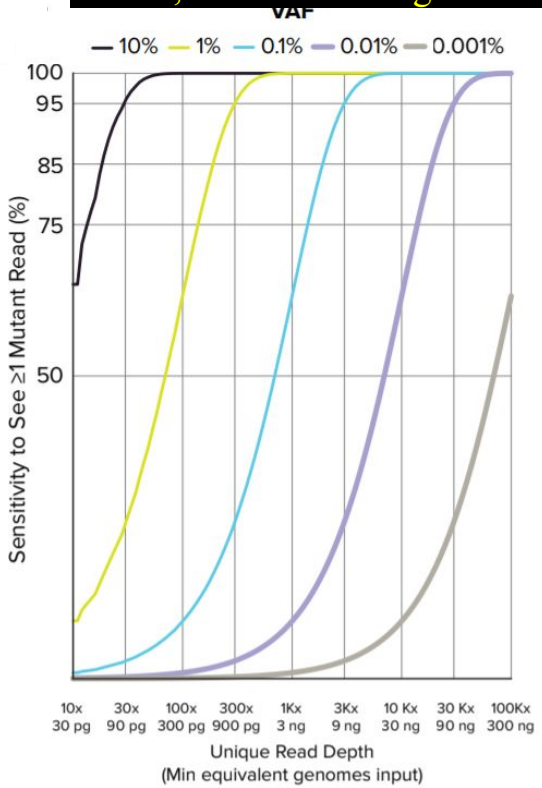
- The rarer the event, the more independent trials you have to sample in order to have high confidence of seeing it.



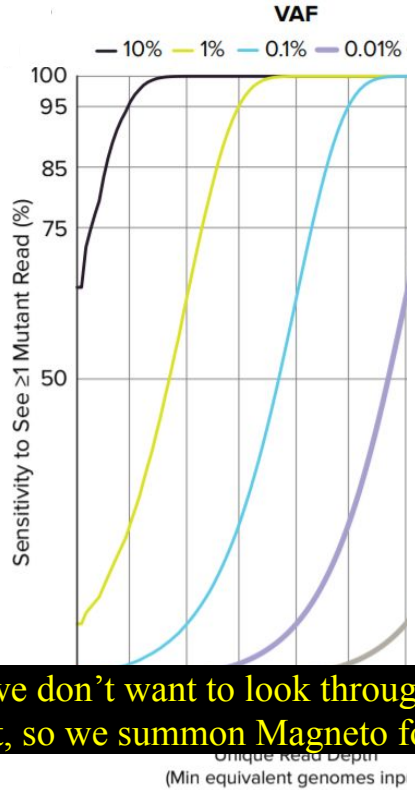
This plot shows the theoretical estimate, based on the binomial distribution, of how many **independent** (unique molecules) reads you'd need to detect one mutant read, as a function of mixture proportion. But ~no one has understood it from this plot alone, so let's talk analogies...

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

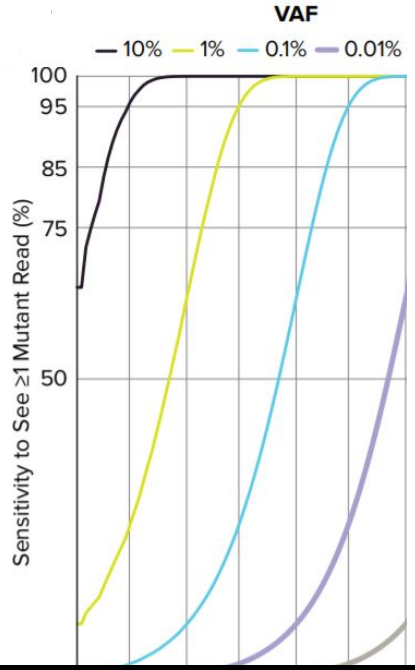


But we don't want to look through every single bar to find the ticket, so we summon Magneto for help.

Detection o ents

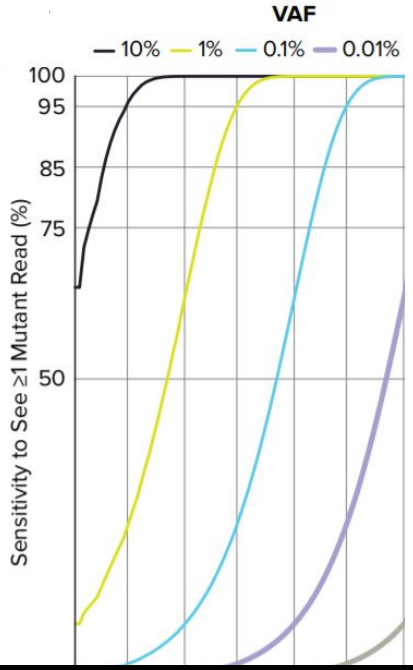
Magneto uses his power over metal to summon the one bar with a golden ticket. This is like enrichment in sequencing: we use hybrid capture or PCR to only pick out the fragments we care about (those in particular regions).

ents



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

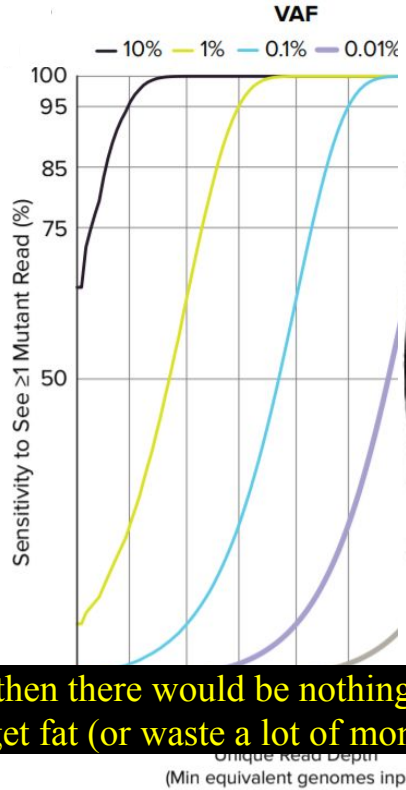


???

???

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

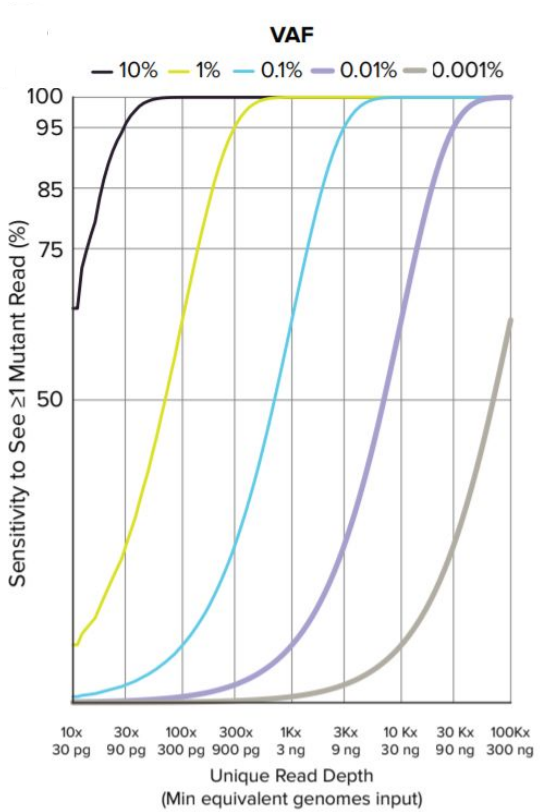


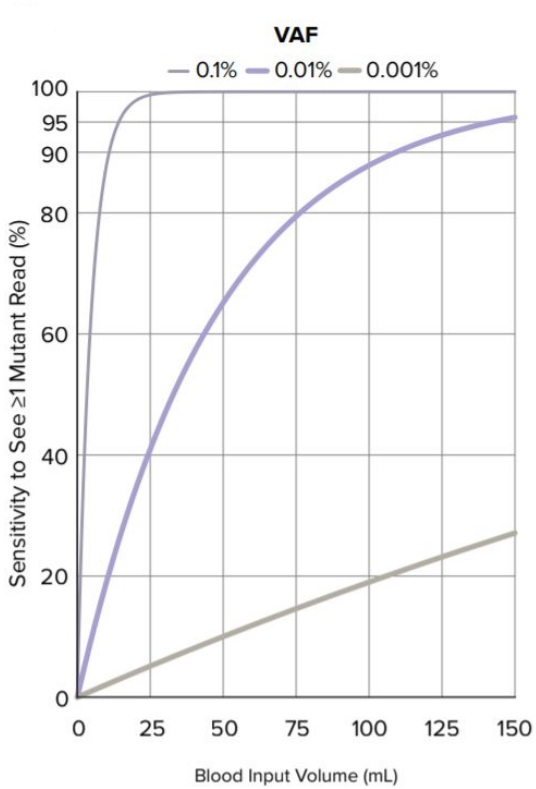
And then there would be nothing left to do but to eat all the candy and get fat (or waste a lot of money on sequencing).

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.

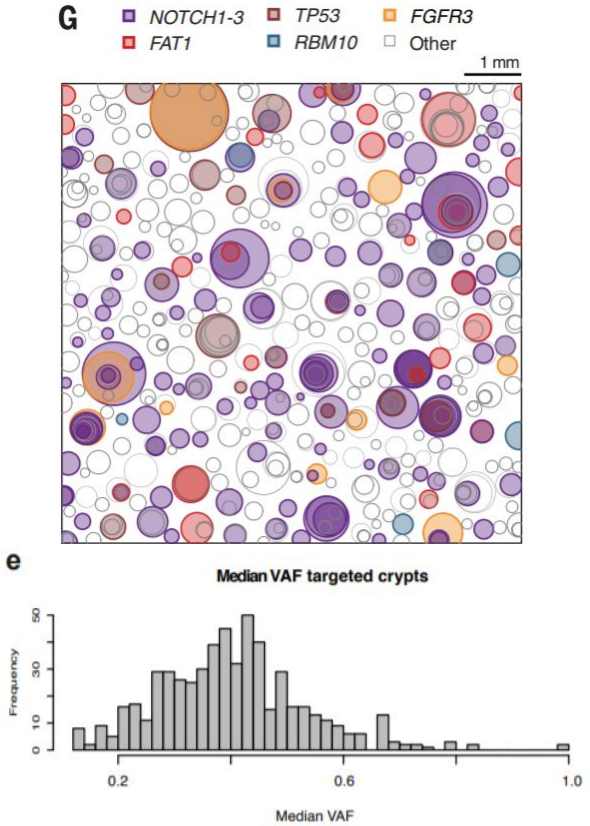




Haque IS and Elemento OE, *bioRxiv* 2017
Haque IS et al. AACR 2018

Multi-site assays and somatic heterogeneity

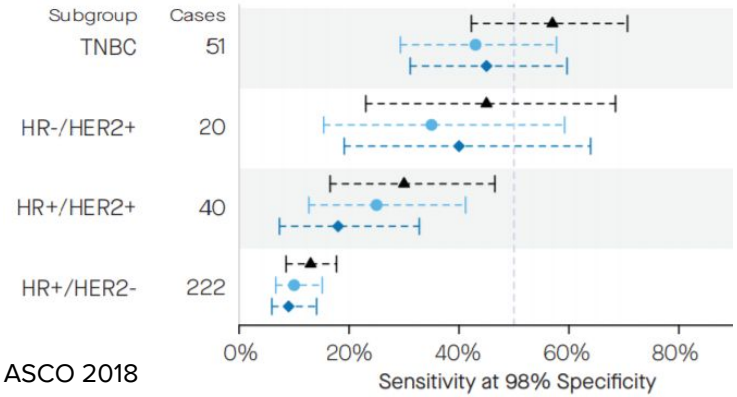
- Looking at multiple sites could help: if independent, VAFs add (10 sites @ 0.01% ~ 1 site @ 0.1%).
- Somatic heterogeneity appears to be the natural state of even healthy tissues, with age dependence. 1% of healthy colon crypts carry putative oncogenic mutations.
- Too narrow: need too much blood
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



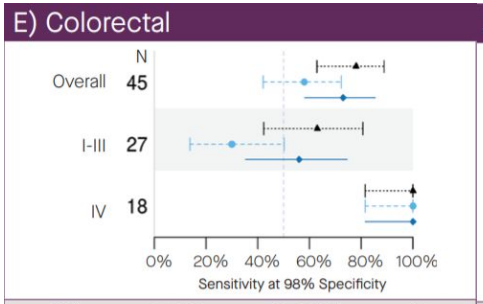
Martincorena I et al *Science* 2015
Lee-Six H et al, *bioRxiv* 2018

Real world evidence: TF matters

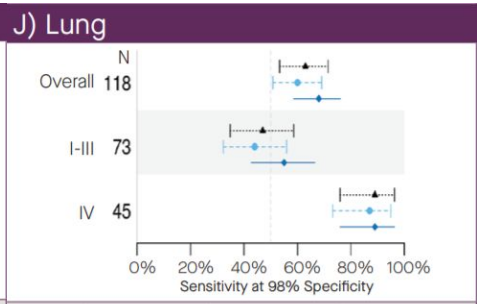
- GRAIL has reported ctDNA data from 1785 patients (~3000x unique depth).
- Strong stage dependence: cancers with more stage I tend to perform much worse: suggests **tumor fraction is a real, fundamental limitation.**



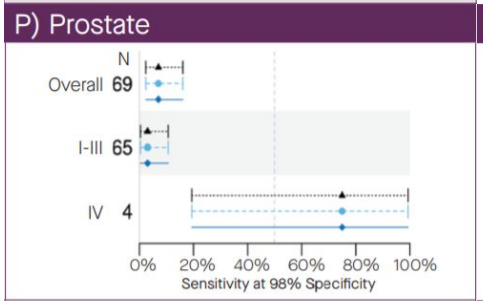
Liu MC et al. ASCO 2018



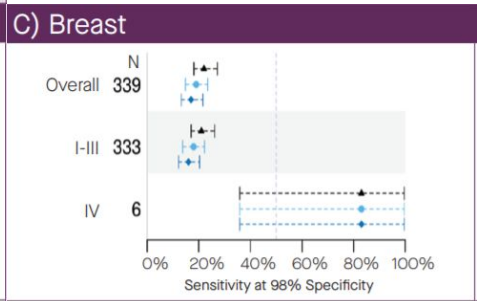
- 9% stage I, 18% stage II, 33% stage III, 40% stage IV



- 19% stage I, 11% stage II, 32% stage III, 38% stage IV



- 22% stage I, 68% stage II, 4% stage III, 6% stage IV



- 48% stage I, 36% stage II, 14% stage III, 2% stage IV
- In TNBC, 57% sensitivity (95% CI: 42-71%) at 98% specificity (poster 536)

“Mechanistic” Discovery Challenges in Therapeutics

“Amyloid-β aggregates are observed in Alzheimer’s-affected brain tissue and appear to be neurotoxic”

Note that the challenges of mechanistic discovery are not limited to diagnostics alone; it has hit therapeutics as well, with the amyloid hypothesis in Alzheimer’s disease maybe the most prominent example...

“Mechanistic” Discovery Challenges in Therapeutics

“Amyloid-β aggregates are observed in Alzheimer’s-affected brain tissue and appear to be neurotoxic”

BACE1 inhibition
- Verubecestat (PhIII term. Feb 2018)
- Lanabecestat (PhIII term. 2018)

γ-secretase inhibition
- Semagacestat (PhIII failure Aug 2010)

γ-secretase modulation
- Tarenflurbil (PhIII failure 2008)

Anti-Aβ
- Bapineuzumab (PhIII failure 2012)
- Solanezumab (PhIII failure 2018)
- Aducanumab (PhIII term. 2019)

Observational and model system data has implicated aggregates of the amyloid-beta peptide in Alzheimer’s disease pathology. However, repeated attempts to attack Alzheimer’s along different points in the pathway (beta- and gamma-secretase or amyloid-clearing antibodies) have ALL failed to show clinical benefit, bringing the mechanism-disease connection into question and burning piles of cash in the process.

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 Modeling and 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.

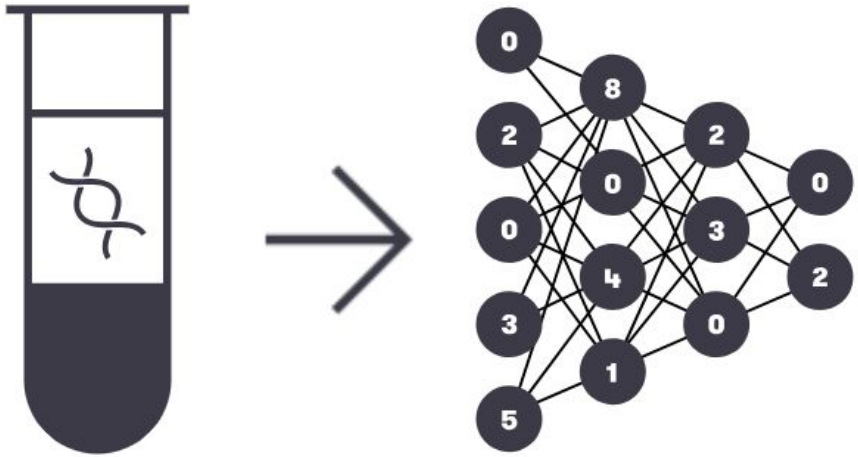
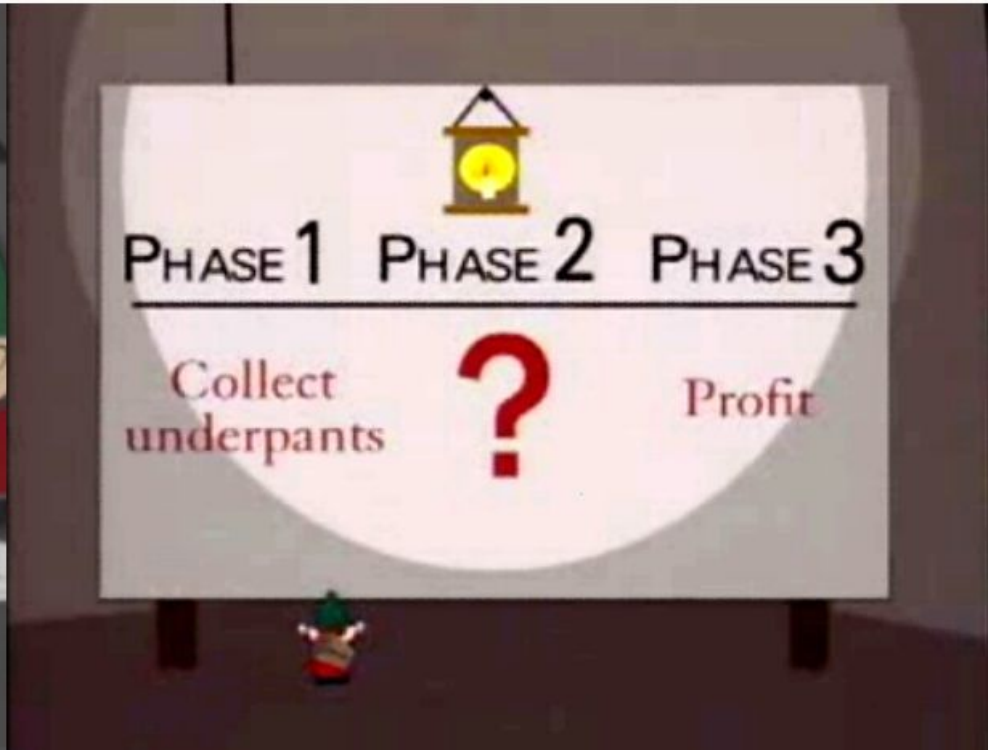


Image from freenome.com

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!

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.

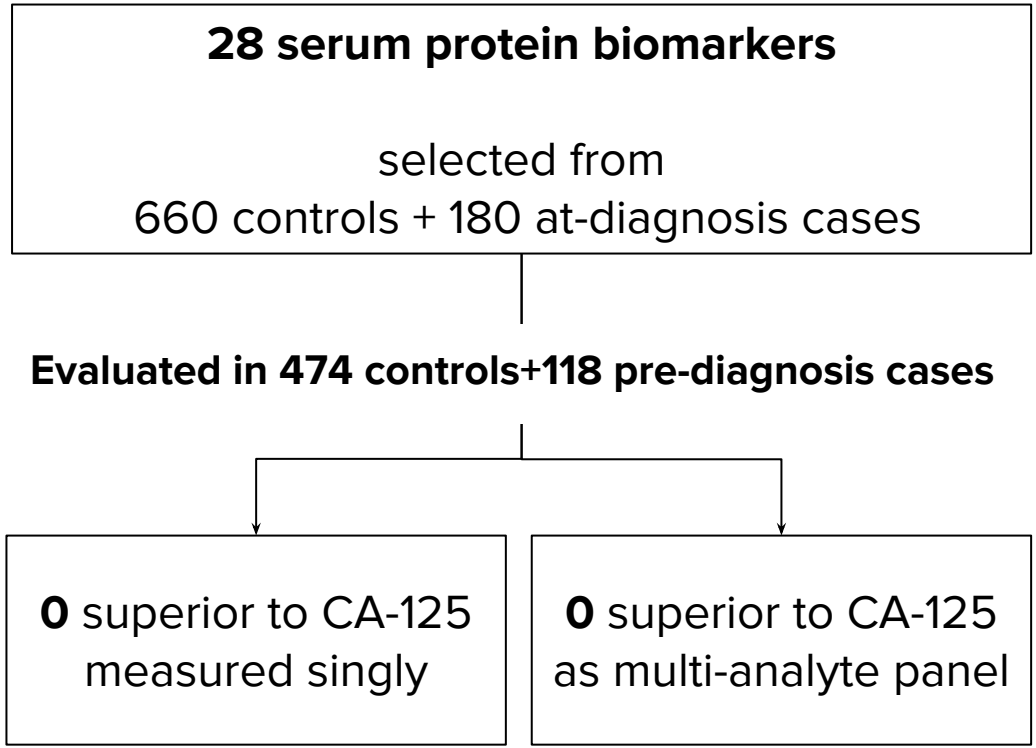
This is usually done stepwise: enriched case-control cohorts for marker discovery, sequentially larger “validation” cohorts.

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

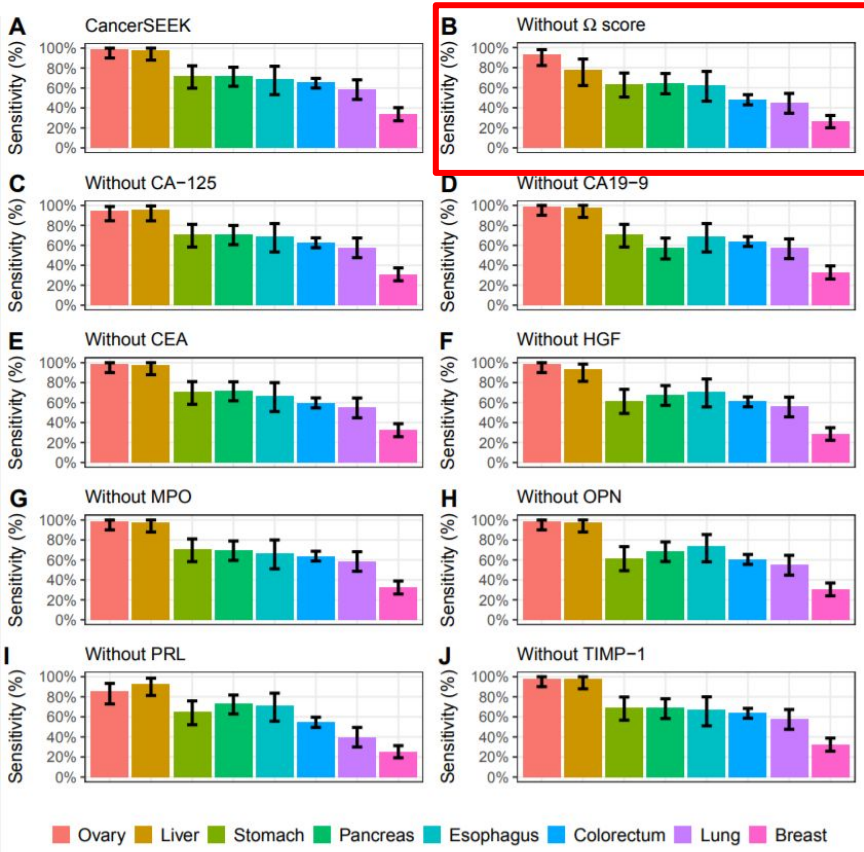
Post-diagnosis != pre-diagnosis
Screening finds different disease categories.

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*

All this has happened before and will happen again



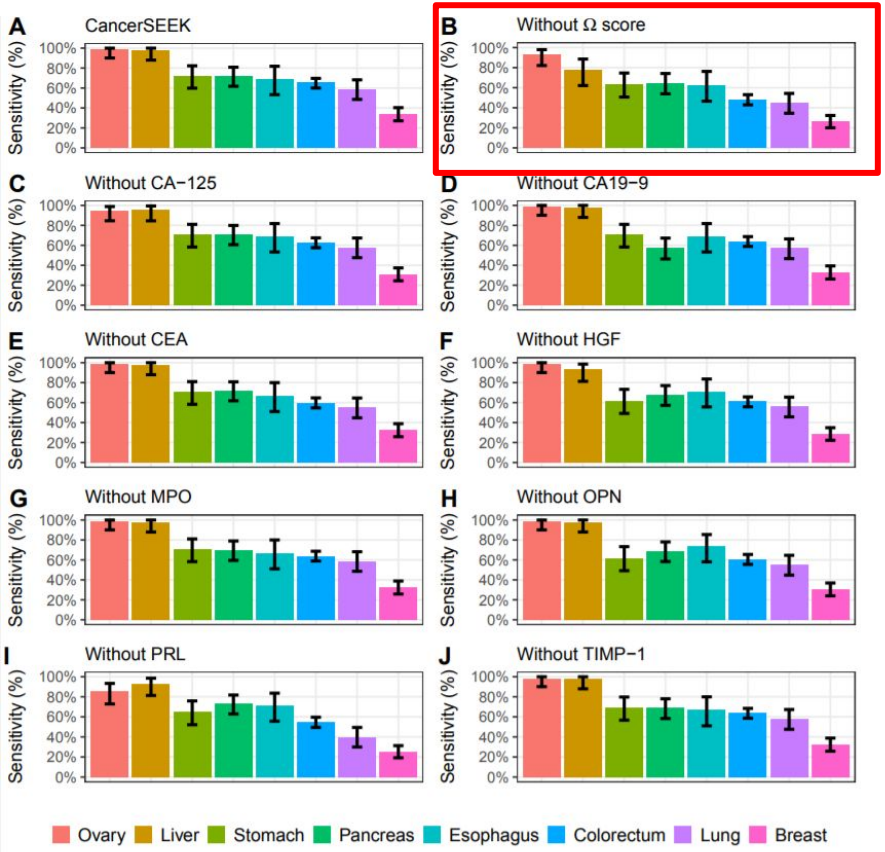
Note that panel B (from the supplementary data to Cohen 2018) suggests that removing ctDNA entirely from the CancerSEEK assay leaves assay performance largely intact: either most of the power is coming from the proteins, or all the assays are measuring similar things (are non-independent).

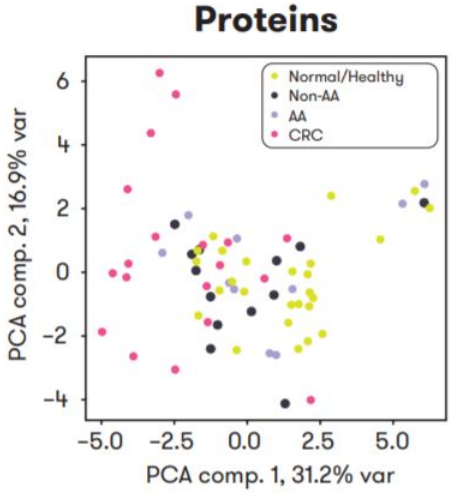
Non-independence

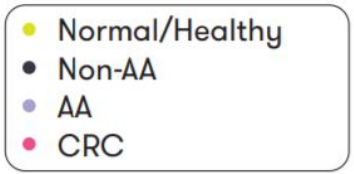
Just because each thing (may) work alone, doesn't mean combinations will work better.

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

All this has happened before and will happen again

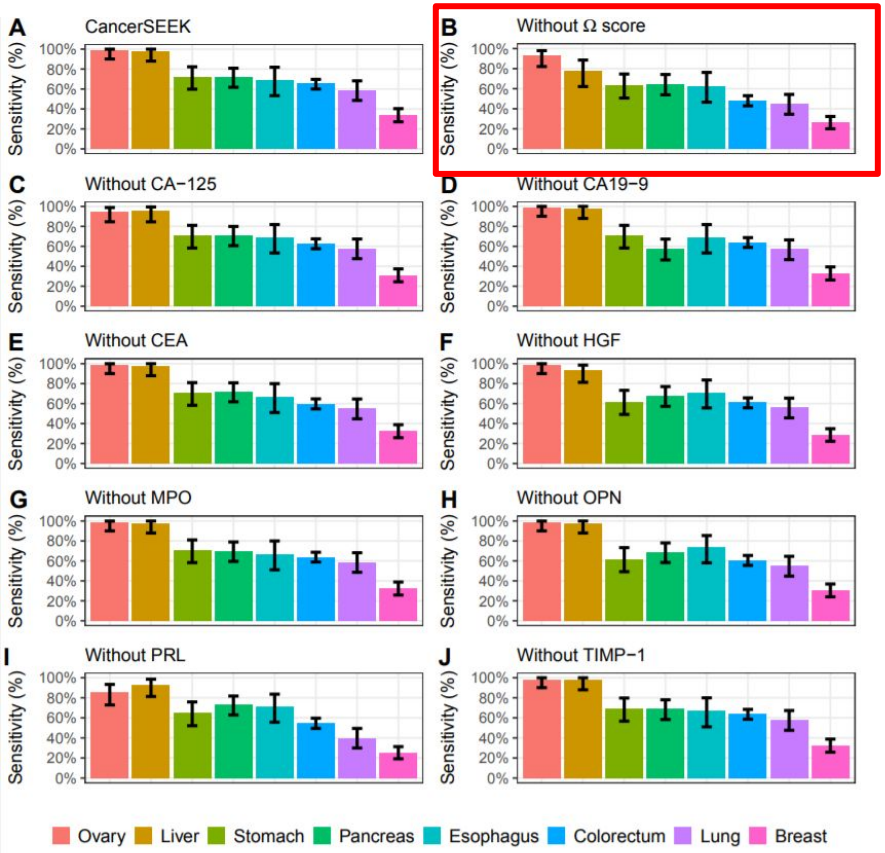


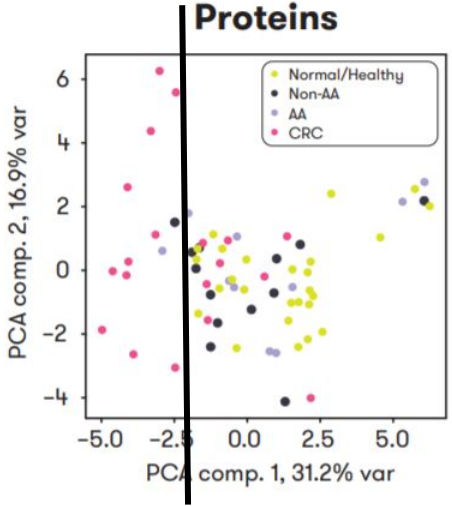




Cohen et al. *Science* 2018
Delubac et al. AACR 2018

All this has happened before and will happen again

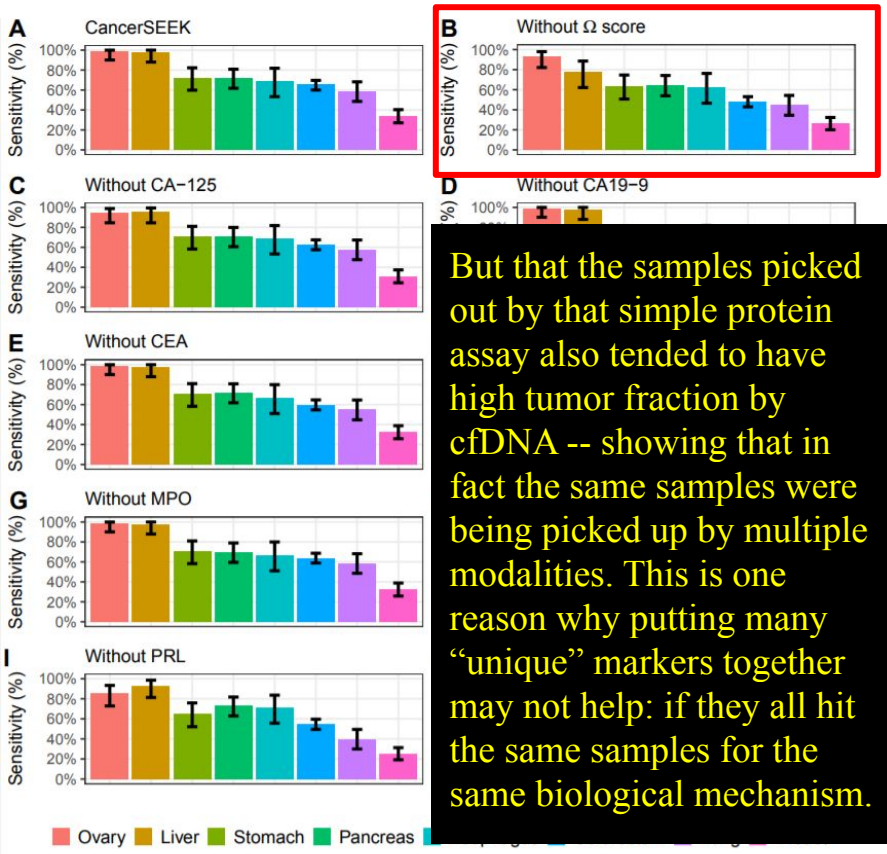




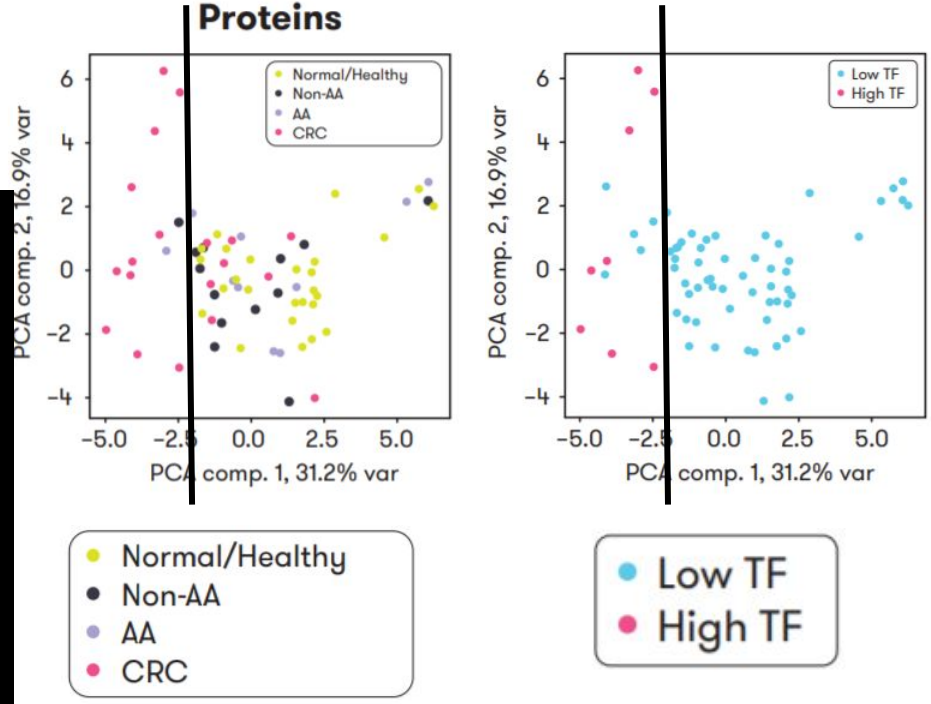
Work from another Freenome poster, presented at AACR by Delubac et al showed that in a small cohort of samples, it was possible to design a reasonable cancer detector using proteins alone...

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.

The Challenge of Translational Scale

Describing an Identity Crisis in Memes

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).

We'll change topics at this point: instead of discussing the scientific challenges that make it hard to do biological discovery, we'll talk about the social or political challenges that make it hard to execute on validation and translation.

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).

Both approaches are hacks to deal with the fundamental challenge of validation and scale: wrong answers look "right" in small data sets, but large data sets are expensive!

The Identity Crisis of Discovery/Translation

Both academic and industrial science have an identity crisis when it comes to scaling up (non-therapeutics) discovery.

	Want	Need
Academia		
Industry		

Specifically, everyone involved - whether in academia or in industry - has a divergence between what they WANT to be doing (or how they see themselves) versus what they HAVE to do on a day-to-day basis (or what they are incentivized to do).

The Identity Crisis of Discovery/Translation

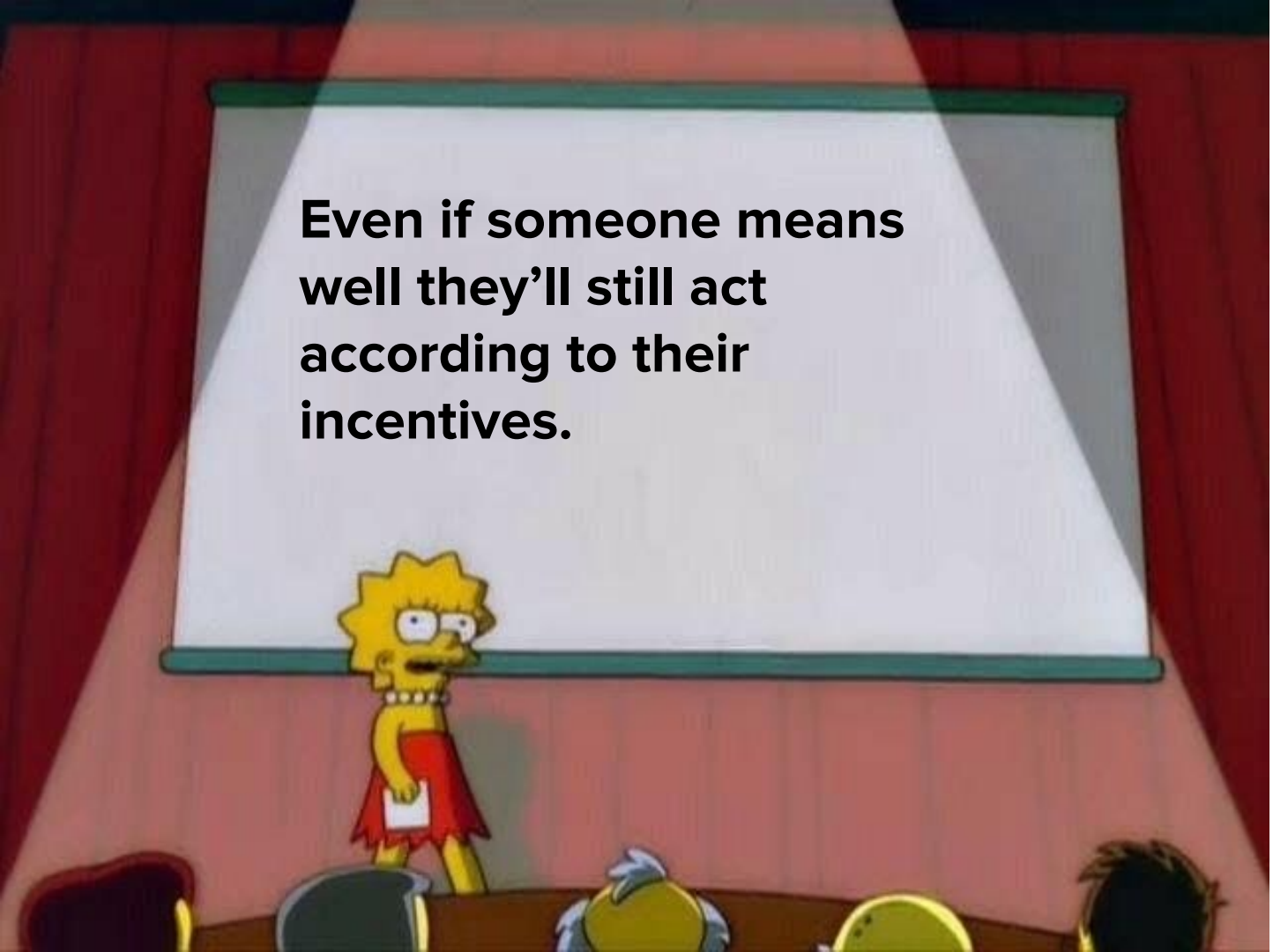
Scientists in academia and industry are not so different: (mostly) everyone is in it to make a difference in our understanding of biology and health and just tackling different aspects of the problem.

	Want	**Need**
Academia	Do cutting-edge science, making new discoveries to change the course of health.	
Industry	Do rigorous science and engineering to bring scientific advances into practice and improve health.	

The Identity Crisis of Discovery/Translation

However, in both cases, the activity required by the job is not actually perfectly aligned with the goals of the participants - and this misalignment makes it hard to actually execute the large studies we need.

	Want	**Need**
Academia	Do cutting-edge science, making new discoveries to change the course of health.	Train PhD students/postdocs, get high-IF publications, get tenure.
Industry	Do rigorous science and engineering to bring scientific advances into practice and improve health.	Make a return on investment.

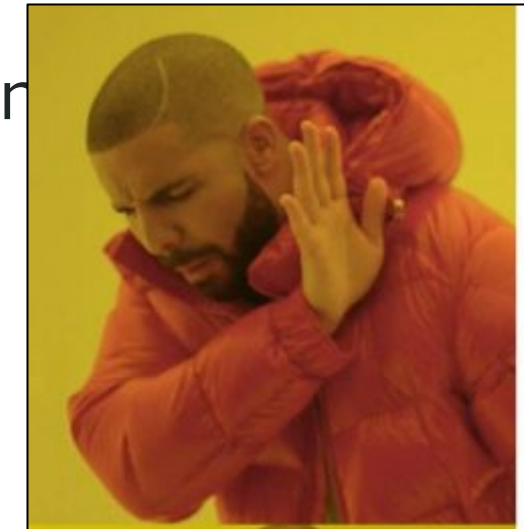


Lisa Simpson reminds us that aligning incentives and actions is important in setting policy. If we're not getting the outcomes we want (better-validated studies), maybe we can look to the existing incentive structure to understand why.

Why won’t industry do more of the really big studies?

Why won't in g studies?

Here, Drake helps us summarize a (*highly simplified*) view of industry's incentives.



- **Fundamental discovery**
- **Showing clinical utility**
- **Methods transparency**

- **Selling lots of tests for a lot of money apiece**

Why won't industry do more of the really big studies?

Industry

1. Even if you show clinical validity, it's a long road to utility, approval, reimbursement.

Clinical validity is necessary but not nearly sufficient to get a marker used in practice and changing clinical care. The scale-up and commercialization road is long and risky.

2. Most published results will fail anyway; secondary "basic science" results don't pay the bills.

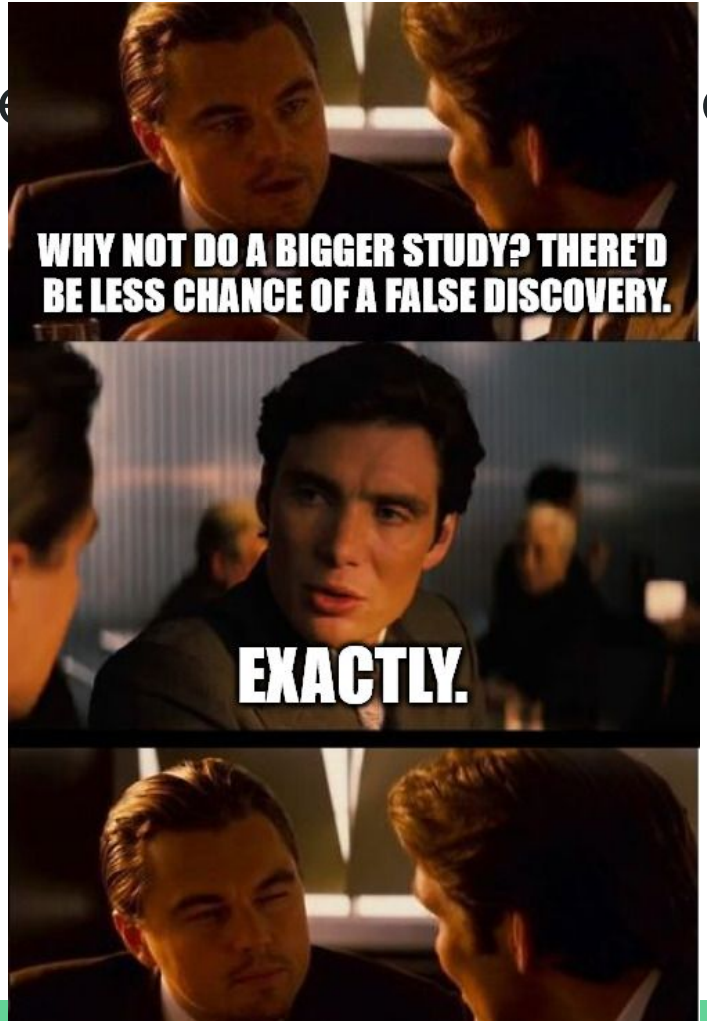
If a validation study fails, you might learn interesting biology, but *Nature* papers don't pay the bills if you're a company!

3. Huge second-mover advantage if there is no IP.

If there's no way to prevent a second-mover from immediately and cheaply copying your results after you've invested the money into a study/trial, it's very unlikely that your investors would want you to spend they money on that trial.

Why won’t academia do more of the really big studies?

Why won't acade eally big studies?



Enough about industry. A cynic speaking about academia might argue that expected impact factor is not simply maximized by sample count: in fact, it's best to neither have too few not not too many samples in your study: too few, and no high-profile journal will accept the paper. But too many, and you reduce the chances of being able to make an "exciting" but spurious discovery that you can publish.

Why won't academia do more of the really big studies?

Less cynically, there are excellent reasons why academic labs would not want to run large validation studies: they may not be compatible with the individual labs' goals of training students and postdocs, and the concentration of funding they would require would starve many other labs for research funds while elevating a small few.

1. Validation studies are huge and one-and-done.
    a. Requires specialized expertise to run; don't want to staff with trainees.
    b. Not enough content to get a PhD from one
    c. Requires huge concentration of funding and there are a lot of labs to feed.

Why won't academia do more of the really big studies?

Furthermore, although these studies may be interesting from a downstream translational perspective, they are not designed to deliver exciting science -- that would make someone's name or career - as a primary outcome, regardless of whether one is an experimental "data generating" or computational/statistical "data consuming" PI.

1. Validation studies are huge and one-and-done.
   a. Requires specialized expertise to run; don't want to staff with trainees.
   b. Not enough content to get a PhD from one
   c. Requires huge concentration of funding and there are a lot of labs to feed.

2. Basic science is at best a secondary output of these studies.

"Data Generators"	"Data Consumers"
- No new whiz-bang tech to show off here. Scale and stability are "boring".	- Validating an existing result is boring; the real cred lies in finding something new.

What reactions have we seen to these incentives?

OPEN SCIENCE (academia)

1. Open data release benefits data generators (citations, impact) and consumers (ability to specialize and scale analysis)
2. Partially addresses reproducibility challenges (independent analysis, new test sets).

(Preface: Despite my comments here on its limitations, I'm a big fan of open science and have pushed my employers towards open-access publication and sharing as much data as possible.)

Because individual academics are incentivized by publications, citations, and grant funding, the open science movement has taken shape to open up access to all data and methods used in individual studies. This benefits all kinds of labs (in priniciple), and can help with reproducibility as independent analysis is possible and data collected for one study can be repurposed as independent tests for other studies.

What reactions have we seen to these incentives?

OPEN SCIENCE (academia)

1. Open data release benefits data generators (citations, impact) and consumers (ability to specialize and scale analysis)
2. Partially addresses reproducibility challenges (independent analysis, new test sets).

INDUSTRY WILL NOT ALLOW ALL ITS DATA TO BE OPEN (OR MAKE MORE OPEN DATA) IF IT MAKES PRODUCT CLONING TRIVIAL.

However, although there's loads of data that industry collects, and indeed wouldn't mind sharing, it will not do so if doing so makes it trivial for competitors to copy results for free (or much cheaper than the original investment).

On the other side: the breakdown in biomarker IP protections has led companies to try "data moat" strategies previously seen in consumer tech: prevent competition by gaining an insurmountable lead in amount of data collected and kept private.

OPEN SCIENCE (academia)

1. Open data release benefits data generators (citations, impact) and consumers (ability to specialize and scale analysis)
2. Partially addresses reproducibility challenges (independent analysis, new test sets).

INDUSTRY WILL NOT ALLOW ALL ITS DATA TO BE OPEN (OR MAKE MORE OPEN DATA) IF IT MAKES PRODUCT CLONING TRIVIAL.

DATA MOATS (industry)

1. If biomarkers themselves can't be patented, then maybe black boxes can be validated, while keeping training data private.
2. Partially addresses IP/barrier-to-entry challenge but creates concerns around validation, trust, genericization.

Data moats create obvious concerns about verification and validation, but perhaps more interestingly create long-term economic concerns: we assume medical advances will eventually become cheap or generic (e.g. through patent expiration), but nothing guarantees a time-based "bridge" over a data moat.

OPEN SCIENCE (academia)

1. Open data release benefits data generators (citations, impact) and consumers (ability to specialize and scale analysis)
2. Partially addresses reproducibility challenges (independent analysis, new test sets).

INDUSTRY WILL NOT ALLOW ALL ITS DATA TO BE OPEN (OR MAKE MORE OPEN DATA) IF IT MAKES PRODUCT CLONING TRIVIAL.

DATA MOATS (industry)

1. If biomarkers themselves can't be patented, then maybe black boxes can be validated, while keeping training data private.
2. Partially addresses IP/barrier-to-entry challenge but creates concerns around validation, trust, genericization.

REGULATORS + PUBLIC WILL RIGHTFULLY HAVE CONCERNS ABOUT BLACK BOX MODELS AND PRIVATE DATA SETS.

What are potential realignments of the incentives?

- **Dedicated public / consortium / charity funding**: in the vein of previous large public projects (PLCO, 1000 Genomes, etc.)
    + Solves narrowly defined validation problem and creates large data set
    - Does not necessarily solve downstream utility/reimbursement questions
    - Politically challenging (to secure funds and identify study operator)

One possible direction is to set aside a separate pool of funding from the "standard" sources dedicated for these large studies - in the model of PLCO. While this might fix the funding, it doesn't necessarily fix the other incentive problems from the academic side: in particular, it's not clear that this creates useful trainee opportunities nor the "winner-take-all" nature of the funding. Similarly, on the industrial side, clinical validity is necessary but not sufficient, so this just kicks the can down the road one step.

What are potential realignments

- **Dedicated public / consortium / charity fun**
  public projects (PLCO, 1000 Genomes, etc.)
  + Solves narrowly defined validation pro
  - Does not necessarily solve downstrear
  - Politically challenging (to secure funds

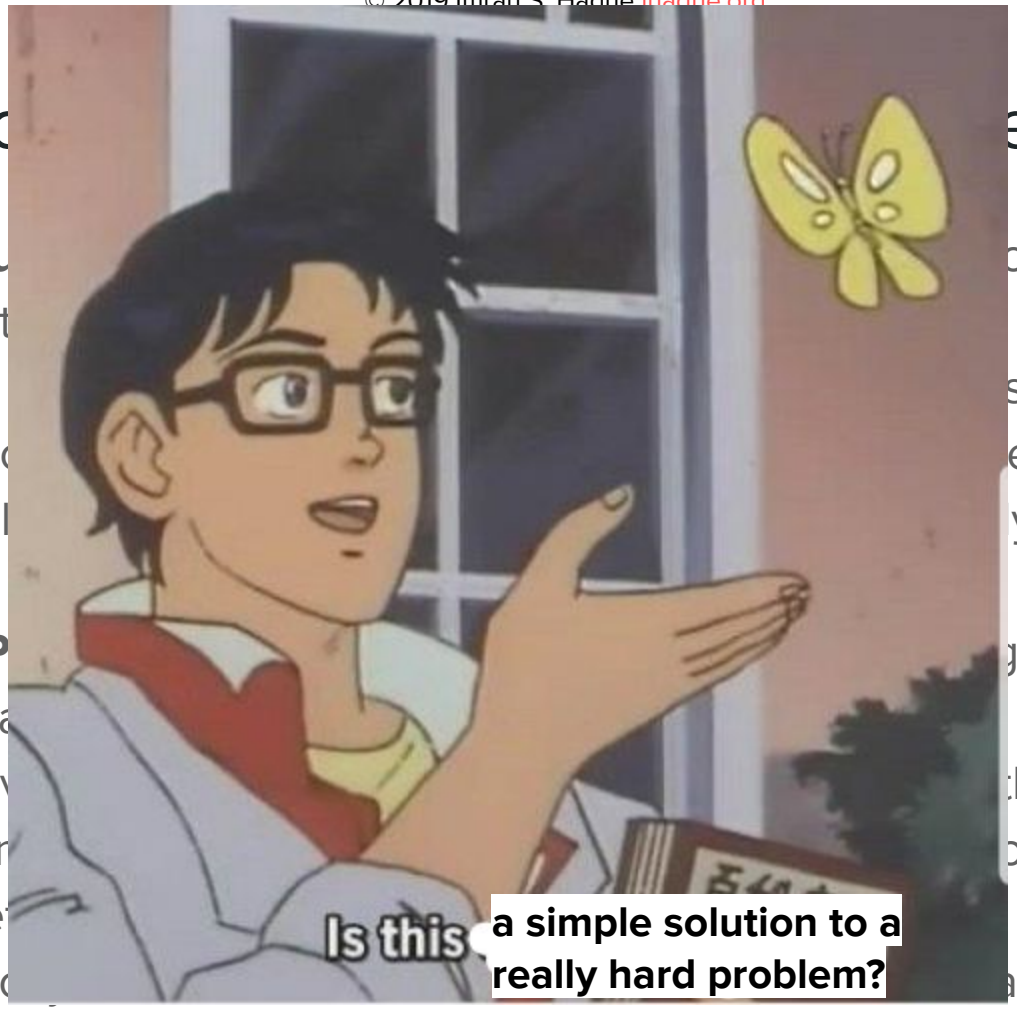
It is highly controversial, yet possibly necessary, to reconsider how we handle IP protection to shift the incentive structure. Examples like sequencing of the UK Biobank provide a potential model: pharma paid for sequencing under the condition that the funding companies would have exclusive use of the funded data for a limited time, but with the data released to the community after this embargo period. Similar models for biomarker discovery may simultaneously help jumpstart validation and launch while preserving society's interests in data and competition downstream.

- **Rethought IP protections**: bring discovery / biomarkers / diagnostics more in line with therapeutics.
  + Incentivizes private industry to spend its capital rather than public purse.
  + Some models (eg time-embargoed release) may boost eventual competition.
  - Very tricky and controversial line to strike between duration and value.

What are po entives?

- **Dedicated pu** of previous large public project
  - + Solves s large data set
  - - Does nc ement questions
  - - Political y operator)

- **Rethought IP** gnostics more in line with thera
  - + Incentiv than public purse.
  - + Some r o downstream compe
  - - Very tric ation and value.



What are po entives?

- **Dedicated pu** of previous large public project
  - + Solves s large data set
  - - Does nc ement questions
  - - Political y operator)

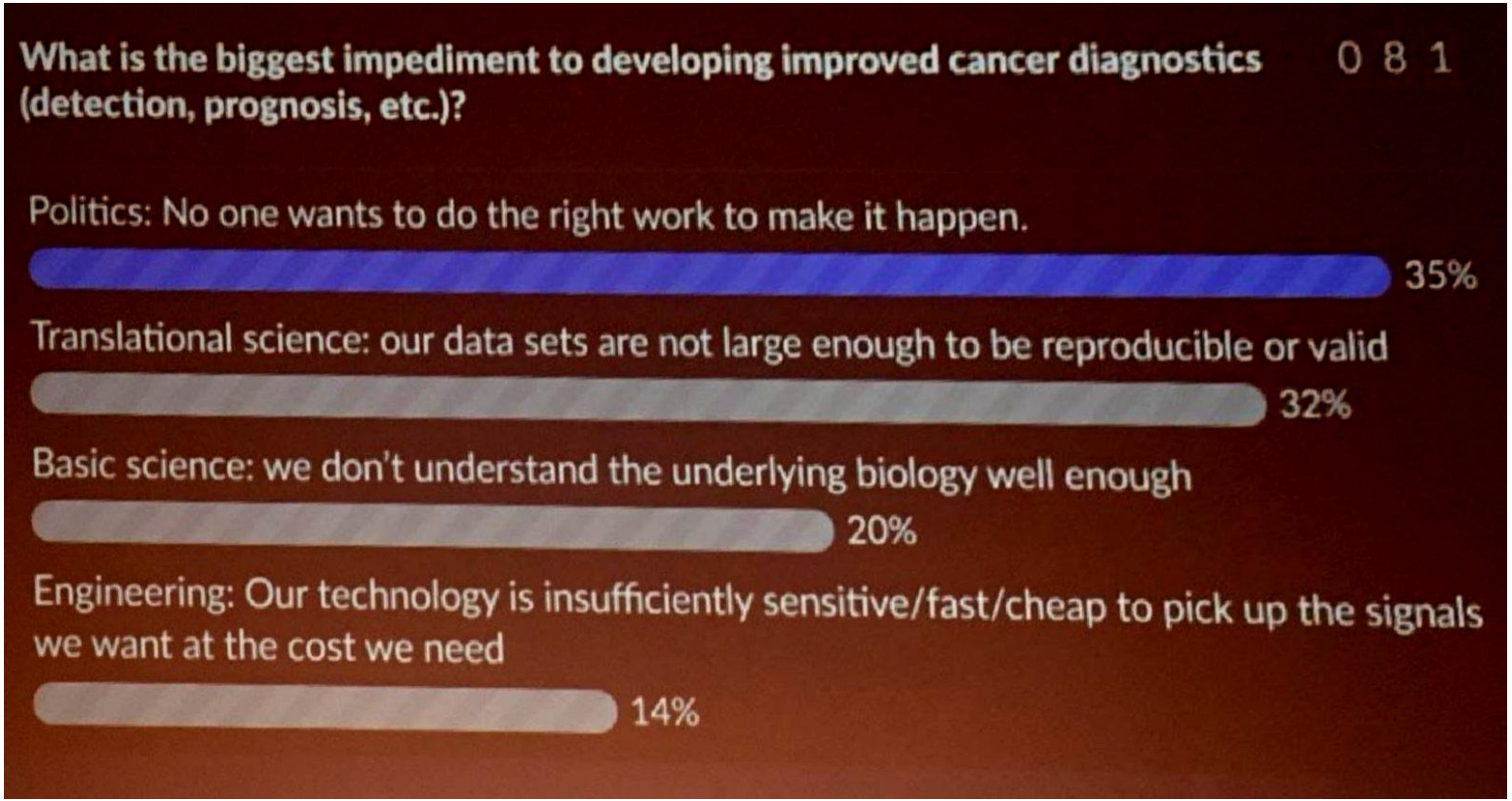
- **Rethought IP** gnostics more in line with thera
  - + Incentiv than public purse.
  - + Some r o downstream compe
  - - Very tric ation and value.

Slido Final

What is the biggest impediment to developing improved cancer diagnostics (detection, prognosis, etc.)?

1. Basic science: we don't understand the underlying biology well enough.
2. Translational science: our data sets are not large enough to be reproducible or valid.
3. Engineering: Our technology is insufficiently sensitive/fast/cheap to pick up the signals we want at the cost we need.
4. Politics: No one wants to do the right work to make it happen.

Final Slido Poll Results



Conclusions

- The conflict between mechanistic and empirical biomarker discovery boils down to tradeoffs between cost/sample acquisition and completeness, with each method suffering from physical or biological sampling limitations.

- While technical improvements may mitigate some genetic/population sampling challenges, there is ultimately no substitute for large-scale validation.

- Engagement by funding and regulatory agencies to realign the actual incentives of academic and industrial research with their stated desires will play a critical role in initiating and supporting the required work.

Send me hot takes at @imranshaque or get in touch at ish@ihaque.org!