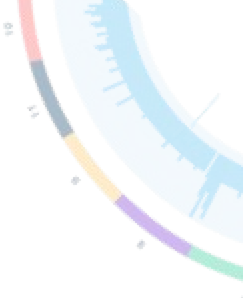


Embracing Heterogeneity:

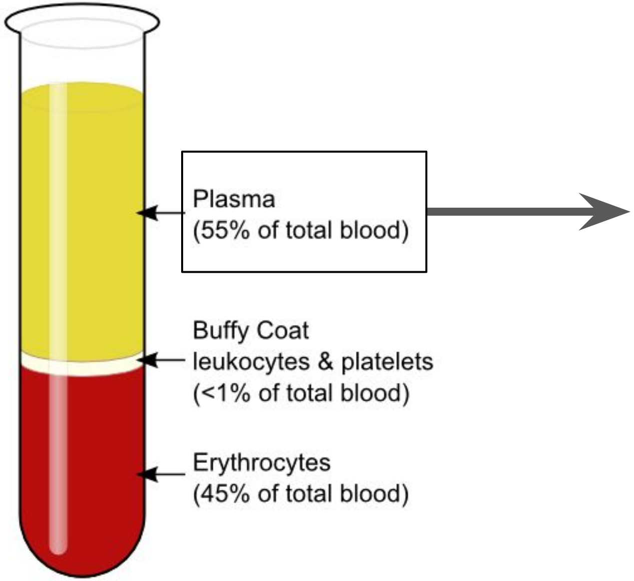
The Freenome, Information, and Early Disease Detection

Imran S. Haque, PhD

Cancer Crosslinks
25 Oct 2017

@ImranSHaque
@freenome

The liquid in “liquid biopsy”



The “freenome”: small molecules, macromolecules: proteins, RNAs, **circulating cell-free DNA (cfDNA).**

cfDNA arises from many different tissues in the body. Have known for >20 years that some amount of cfDNA in cancer patients comes from the tumor: **circulating tumor DNA (ctDNA)**.

Today’s great hope: using **ctDNA** to predict treatment, response, and existence of cancer.

Image credit: KnuteKnudsen, Wikipedia

My claim:

Measurements of circulating tumor DNA (ctDNA) will not solve the clinical problem of detecting early stage cancers.

but integrating multi-analyte signals beyond ctDNA using modern machine learning will.

freenome

About Me

Chief Scientific Officer at Freenome

Formerly VP scientific affairs at Counsyl: early tech dev and research in medical and population genetics, cancer genetics, assay development.

PhD CS Stanford: large-scale machine learning for drug discovery.

About Freenome

Early-stage startup based in South San Francisco, working on **early diagnosis** and **early intervention** in cancer.

Our technical vision:

The cell-free "Freenome" has significant information beyond sequence variation. Decoding this information will require a joint effort in assay development, computational biology, and machine learning, with significant advances to be made in all three.

35+ People at Freenome

$77M Total financing: Andreessen Horowitz, Google Ventures, Polaris, Founders Fund, and others

20+ Clinical partners: Multi-cancer groups sharing cohorts for sample collection + collabs for publication

6 Paid pharma partners: Including Fortune 500 biopharma clients

Outline

I

How do tumor fraction, cfDNA concentration, and heterogeneity come together to define the limitations and costs of mutation-detection liquid biopsy?

II

Where can we find biological opportunities to go beyond variant calling in the development of cancer early detection?

III

What are the statistical opportunities that may enable us to avoid the historical challenges of biomarker discovery?

I: Statistical Limitations of Liquid Biopsy

Detecting cancer by flipping coins

With a fair coin, I **expect** one heads if I flip twice.

But if I'm unlucky, I may need more. With >95% probability, I will see at least one if I flip 8 times:

If my coin is unfair, I might need to flip many more times! If p=0.10, I need to flip thirty times to see at least one heads with >95% conf!

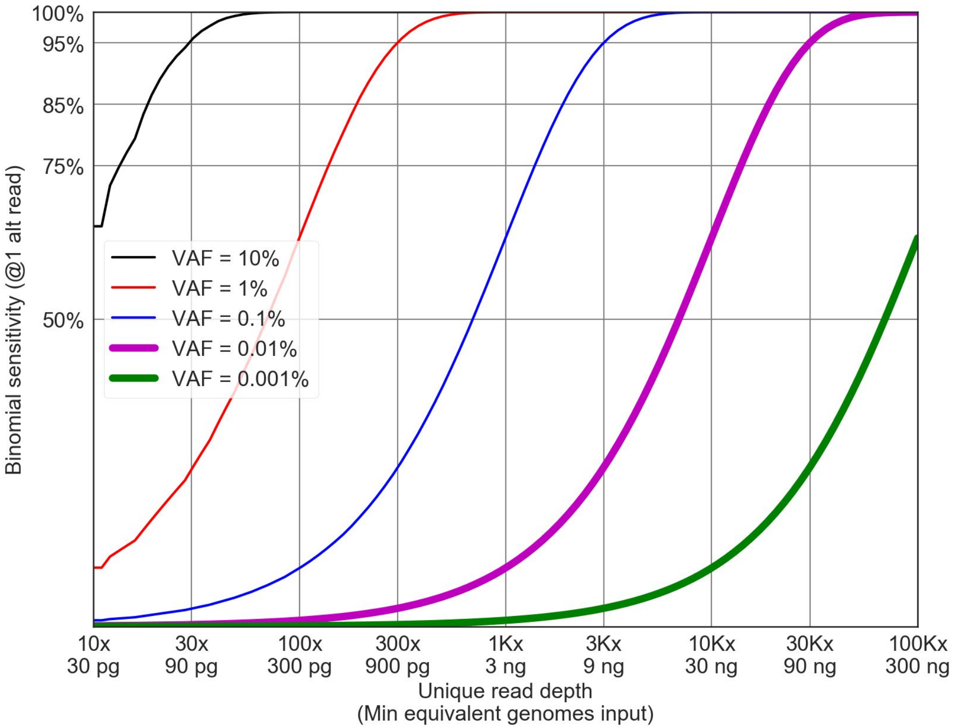
Can treat cfDNA molecules as coins: "heads" for molecules w/a tumor mutation.

Mutation detection as a binomial process

Given ***N*** unique molecules at a given site and ***p*** probability of any given molecule being tumor derived, sensitivity = the binomial probability of seeing at least one tumor derived molecule.

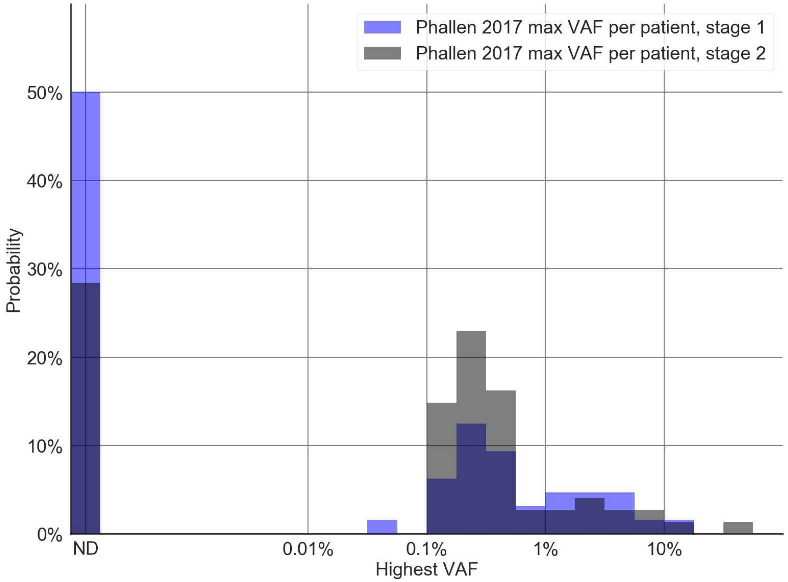
As VAF/TF drops, need **much** more depth to recover high sensitivity.

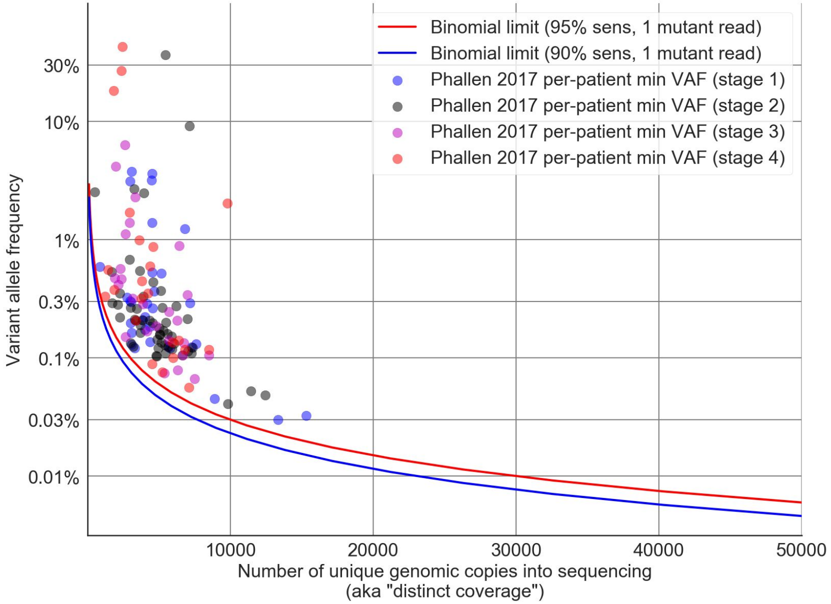
0.01% has been claimed as the VAF upper bound for early detection.



Aravanis AM et al, Cell 2017

Real world VAF is <0.1% in early-stage cancer





30-50% of early-stage patients have max VAF < 0.1%; binomial bound likely reflects what we see in experiments; the deeper you sequence, the more you find.

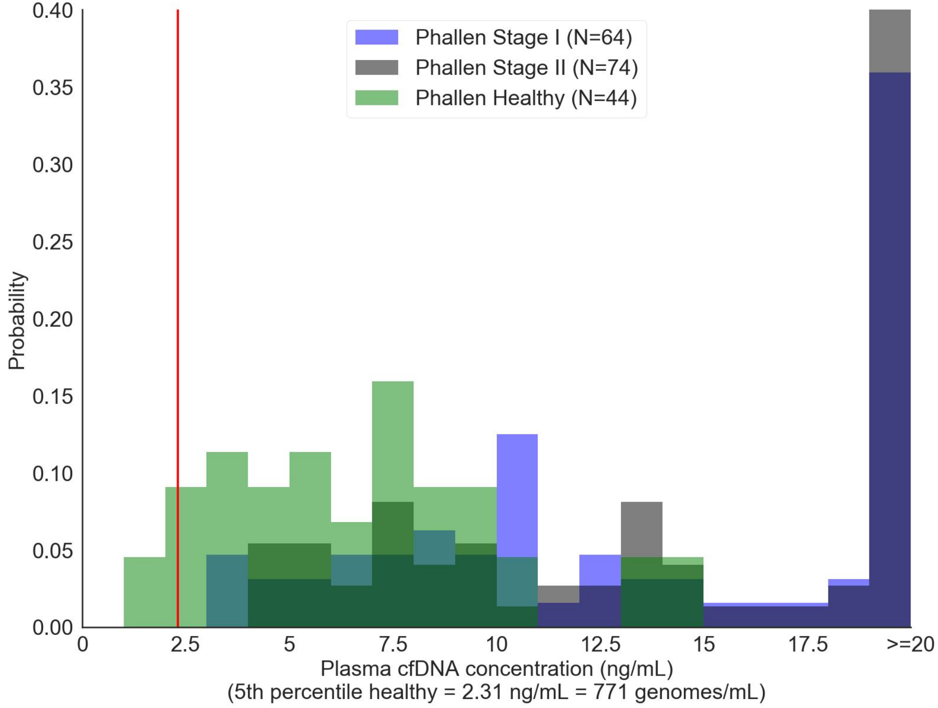
Phallen J et al. Sci Transl Med 2017

[cfDNA]: ~1 ng/mL blood

We've so far assumed that we could sequence arbitrarily deeply, but we need to have **unique** reads, which imposes input requirements.

In a production test, we care more about the tails of the distribution than the median: can't fail 50% of your samples for insufficient input!

95% of healthy indivs have >= 2.3 ng cfDNA/mL plasma = ~1.2 ng/mL blood.



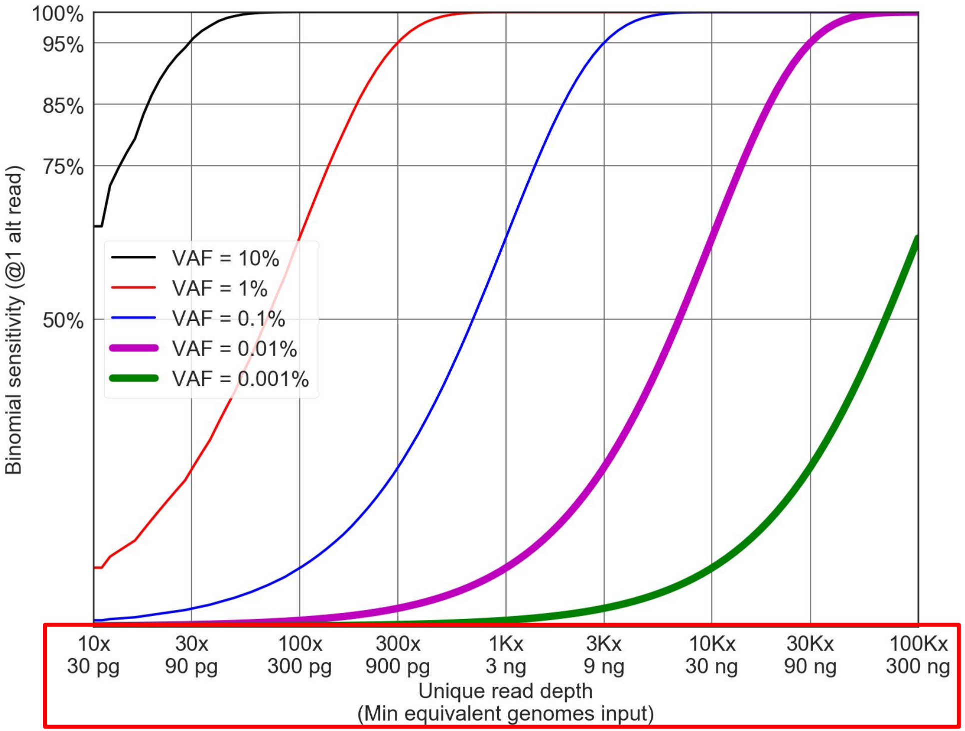
Phallen J et al Sci Transl Med 2017

ctDNA is too rare in early-stage cancer

95% of patients have >= 2.3 ng cfDNA per mL plasma = ~1.2 ng/mL blood.

In order to have 95% sensitivity at 0.01% MAF, we need 30,000 unique genomic equivalents = 90ng.

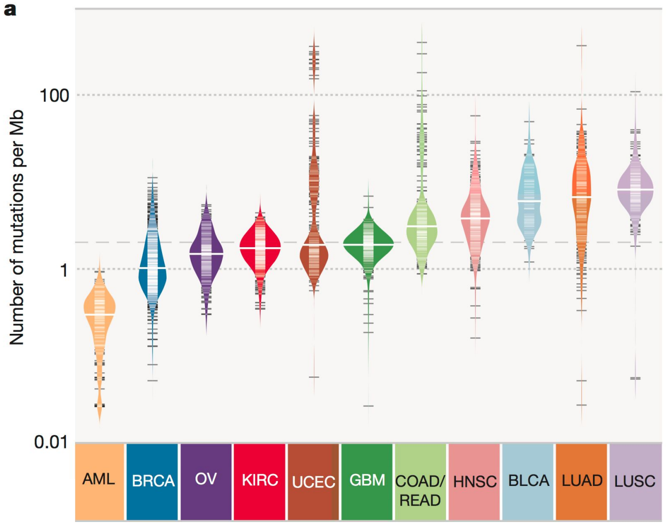
With 100% process efficiency, we'd need a 75mL blood draw; with a more realistic 25-50% efficiency, it's 150-300mL!

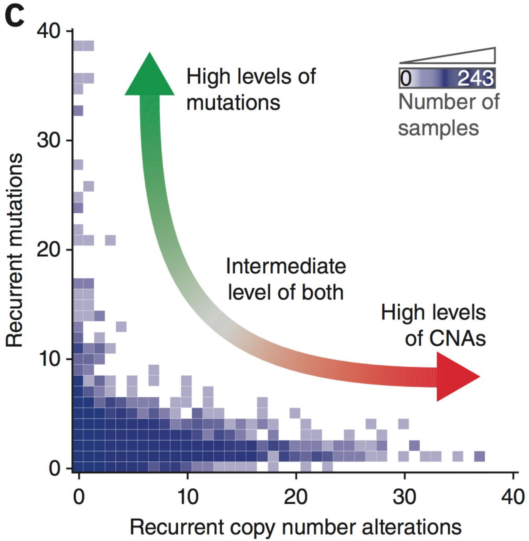


Phallen J et al Sci Transl Med 2017

Idea: Aggregate TF over a larger ROI

We assumed in the previous section that there was exactly one mutation we needed to detect from the tumor, at a particular MAF. But tumors have **many** mutations, sporadic and recurrent; what if it were OK to detect *any* of them?



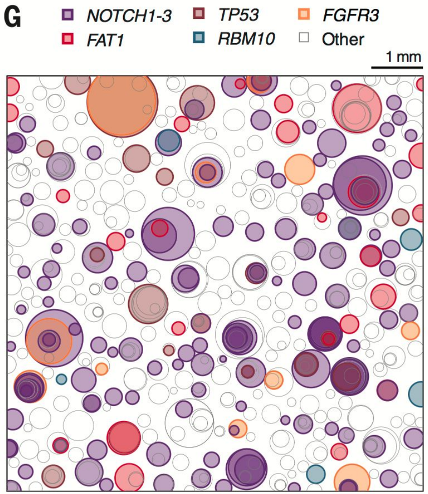


Kandoth C et al, Nature 2013
Ciriello G et al. Nat Genet 2013

Somatic heterogeneity in cancer drivers is pervasive

Deep sequencing reveals significant somatic heterogeneity present in normal tissue and in plasma of healthy individuals. There is significant overlap between driver gene mutations in cancer and somatic variants found in healthy individuals:

Age	VAF, Clonal Hematopoiesis	Fraction of population
>65yr	> 10%	10%
Avg 44yr	0.16-5.28%	16%
<50yr	>=0.1%	10%
>70yr	>=0.1%	~40%



Clonal landscape of 1 sq cm of sun-exposed eyelid skin.

This poses a filtering/PPV problem for mutation detection.

Genovese G et al. NEJM 2014
Phallen J et al. Sci Transl Med 2017
Razavi P et al. ASCO 2017
Martincorena I et al. Science 2015

Aggregating VAF: ↓ input, ↑ sequencing

Impact: can treat vars as indistinguishable in the binomial sampling model, so

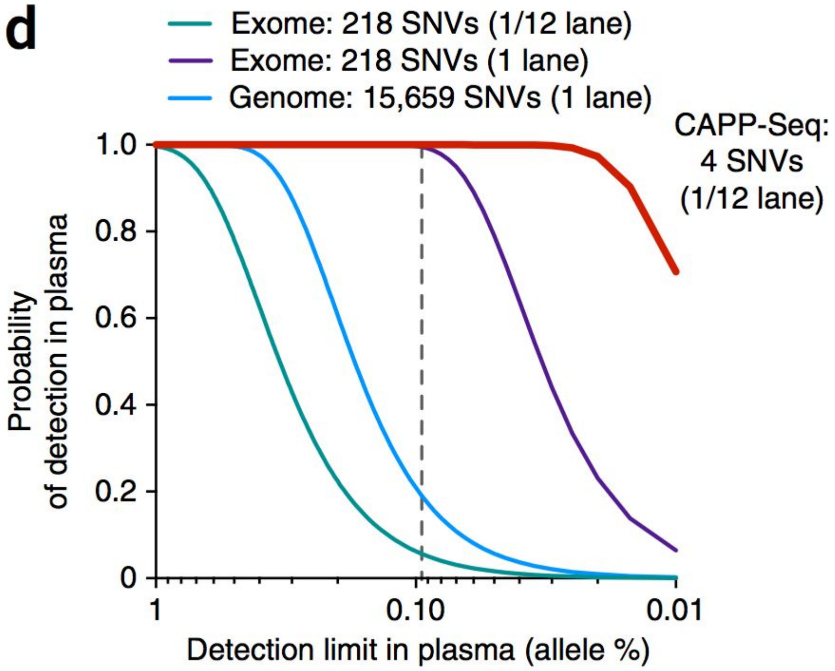
effective VAF <= Σ VAF_i

→ **Reduced input requirement**

(we use more of each genome)

But: seq bandwidth increases; reads from less-altered regions are less powerful.

→ **Increased sequencing requirement**



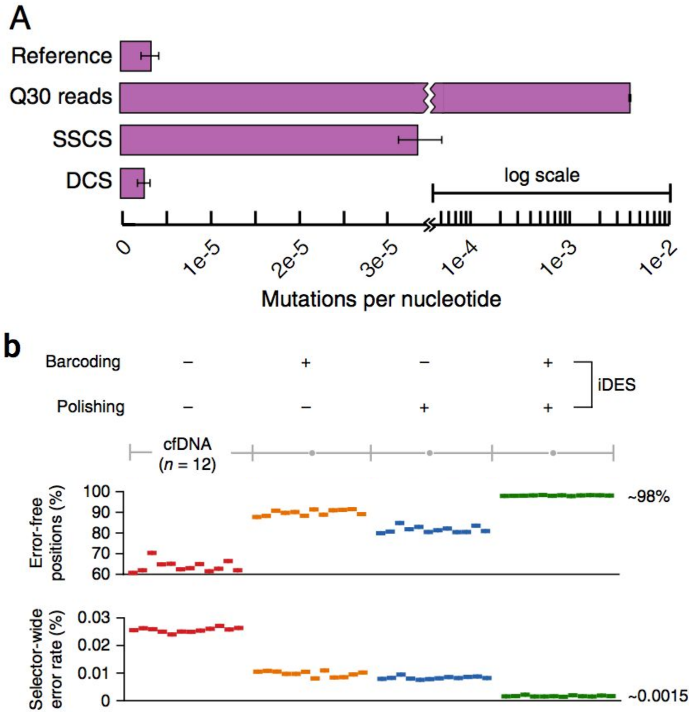
Newman AM et al Nat Med 2014

NGS error correction imposes a depth penalty

Another hitch: NGS reads have an error rate around 0.1%-0.5%, which is >= the mutation rate we want to detect.

A variety of methods (SSCS, DCS, iDES, TEC-Seq) can correct these errors through incorporation of molecular barcodes and oversampling: trading depth for error rate.

In real-world use, error correction requires 5-10x fold increase in read depth:
30,000x req'd depth = 150-300Kx raw depth



Schmitt MW et al. PNAS 2012
Newman AM et al. Nat Biotech 2016
Phallen J et al. Sci Transl Med 2017

ctDNA-based detection is clinically insufficient

Cancer types	Staging	Samples	Panel	Avg Raw Depth	Avg Eff Depth	Clin. Sens.
Breast, Lung, Colorectal, Ovarian	I / II	N=138	81kb	38,589x	6,182x	59-71%
Breast, Lung, Prostate	metastatic	N=124	2100kb	60,000x	3,000-4,000x	89%

Phallen J et al, Sci Transl Med 2017

Razavi P et al. ASCO 2017

ctDNA-based detection is clinically insufficient

Cancer types	**Staging**	**Samples**	**Panel**	**Avg Raw Depth**	**Avg Eff Depth**	**Clin. Sens.**
Breast, Lung, Colorectal, Ovarian	I / II	N=138	81kb	38,589x	6,182x	59-71%
Breast, Lung, Prostate	metastatic	N=124	2100kb	60,000x	3,000-4,000x	89%
Non-cfDNA	**Staging**	**TP+FN / Total**		**Modality**	**FN Set**	**Clin. Sens.**
Ovarian	48% I/II	42+5 / 50,078		CA-125 + ultrasound	1yr followup	89.4%
Colorectal	93% I/II/III	60+5 / 9,989		FIT + DNA	Colonoscopy	92.3%

Even large panels, with moderately high depth, on metastatic patients struggle to exceed the >90% sensitivity that existing screens achieve in early stage.

Phallen J et al, Sci Transl Med 2017
Razavi P et al. ASCO 2017
Menon U et al. Lancet Oncol 2009
Imperiale TF et al. NEJM 2014

Summary: early detection by mutation detection

Desiderata and assumptions:

- <=5% of samples fail test for insufficient input
- 95% sensitivity to detect a cancer-derived allele
- 50% process efficiency tube->sequencer; 5x oversampling for error corr.
- 100% on-target rate for capture
- “$1000 genome” sequencing costs: $1000 / 30x3Gbp
- Other reagents, labor, etc. cost $0.

	VAF 95% sens	Corrected depth	Raw depth	Input vol. (blood)	*TEC-Seq* 58 genes **81kb**	*GRAIL* 508 genes **2,000Kb**	*WES* ~20k genes **50,000Kb**
Tumor LB	0.1%	3,000x	15,000x	15mL	$14	$340	$8,300
Early detection	0.01%	30,000x	150,000x	150mL	$140	$3400	$83,000

II: ...now what?

(Biological opportunities)

Concentration vs quantity

The fundamental limitation affecting ctDNA for early detection is quantity: at 0.01% concentration, there aren’t enough copies of the genome present per mL to be detectable.

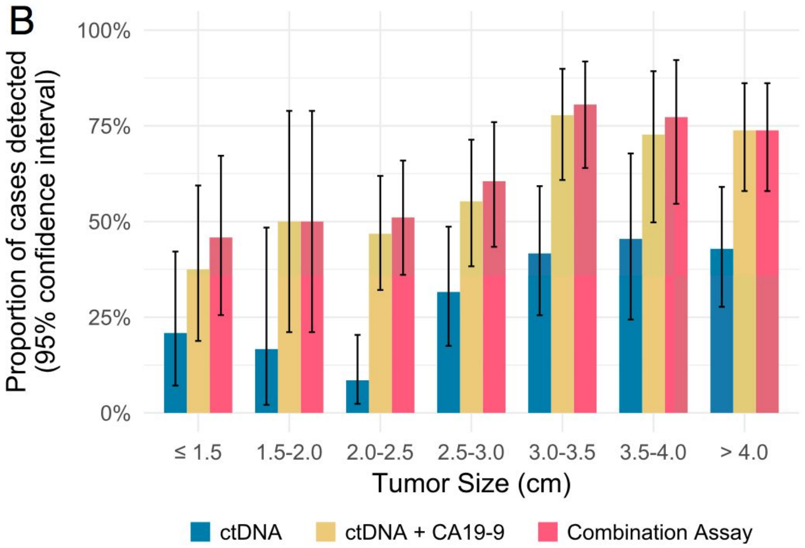
Will highlight two options today:

1. Find tumor-derived material with **count** > 0 even at concentration <0.01%
2. Find informative material in the non-tumor-derived 99.99%

Tumor-derived macromolecules

DNA only has CN~=2 in cells. Other macromolecules have much higher copy number, so they may have >0 count even in small, low-concentration samples.

Combining ctDNA (*KRAS* 2-codon) assay with quantification of 1-4 proteins boosted sensitivity in early stage and small pancreatic tumors, while maintaining specificity (1/182 FPR).



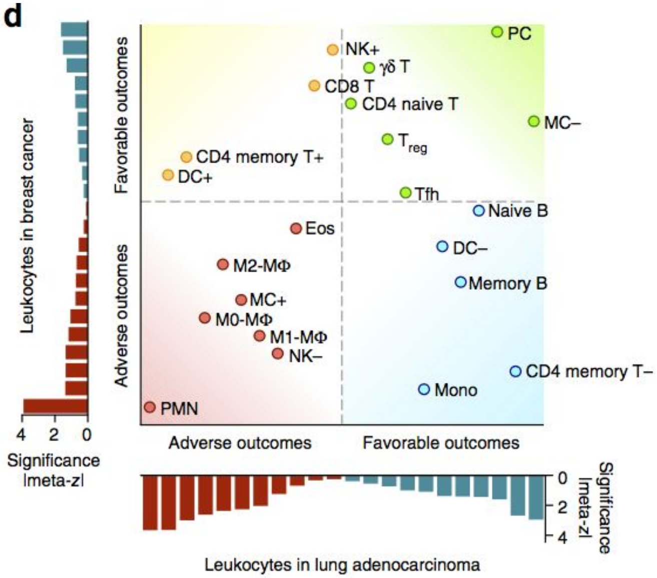
Cohen JD et al. PNAS AOP Sep 2017

Non-tumor-derived material

Immune surveillance is involved in carcinogenesis and early clearing of cancer.

Macromolecules: differential cytokine and antibody production are observed in cancer vs healthy individuals.

Cytology: Different immune cell populations are observed in different tumor types, with an effect on cancer prognosis.

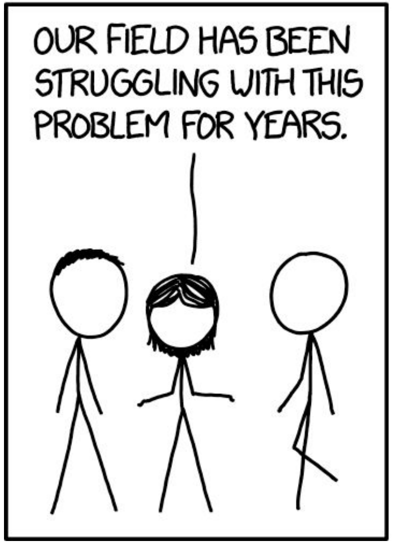


Hanash SM, Pitteri SJ, Faca VM. Nature 2008.
Gentles AJ et al. Nat Med 2015.

Integrative analysis for early detection

ctDNA alone provides neither the sensitivity nor specificity to make a sufficient early detection assay.

But there is good literature evidence to suggest that combining approaches can boost the performance of a detection protocol to useful levels.

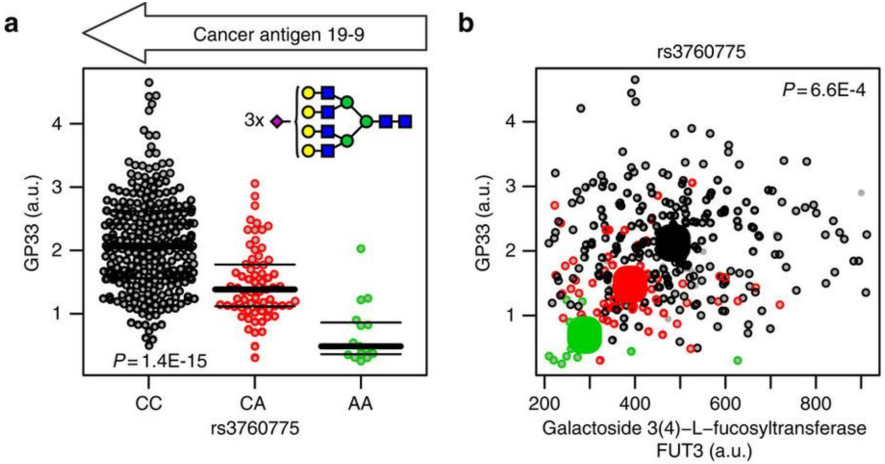


xkcd #1831

Biomarker discovery has not been easy

This is highly multiparametric and relies on statistical recovery of signals from a high dimensional space. The field is littered with failures from confounders: stress response, genotype, circadian rhythms, collection conditions (anesthesia), etc.

Example: Genotype strongly affects levels of circulating plasma proteins, including those used as biomarkers:



Cohen JD et al. PNAS AOP Sep 2017

Suhre K et al. Nat Comm 2017.

III: ...so really, now what?

(Statistical opportunities)

High dimensional modeling: beyond *n* vs *p*

The fundamental limitation of statistical modeling is one of information flow:

1. How many bits of (conditional) information are present in our source data?
2. How many bits needed to discover the label-conditional structure in our data?
3. How many bits needed to classify instances in our target classes?

Older biomarker discovery techniques constrained by (1) and (2): enough apparent information in the samples to classify them, but not enough to factor out the hidden confounders.

Mo’ data, mo’ problems

New methods offer the potential to secure much larger amounts of information per sample, but simply acquiring **more** data is not a panacea:

- Early GWAS: failure to account for ancestry confounding
- Early proteomics biomarkers: failure to account for sample acquisition biases

The great challenge is (2): building models with the appropriate factorization to separate relevant confounders. Structure learning requires **much more** data than just predicting labels directly!

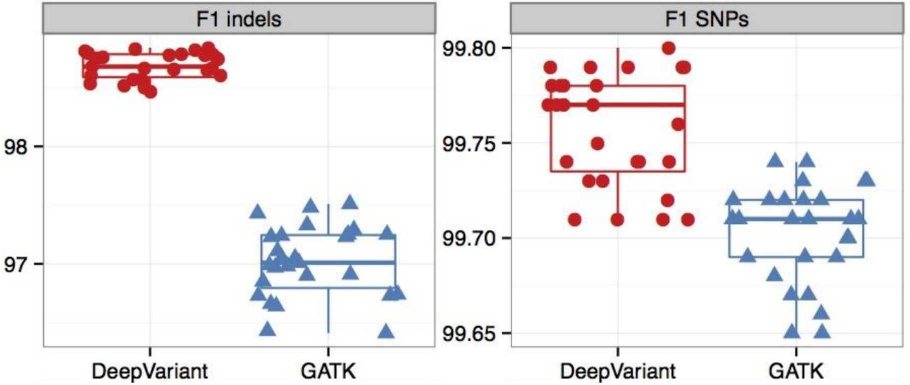
Typical practice: specify confounders up front, and train filters or linear models to remove biases explicitly.

A simple example: variant calling

Variant callers traditionally built with a variety of custom-trained filters: sequence context, allele balance, strand bias, etc.

DeepVariant: automatically trained convolutional neural net on NA12878 platinum genome outperformed hand-tuned GATK HaplotypeCaller

Modern methods have potential for automatic discovery of conditional structure in complicated genomics data.



Poplin R et al. bioRxiv 092890 (2016).

When adding even mo’ problems can help

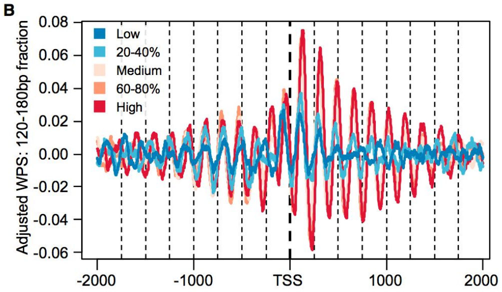
Variant calling is unique: enormous amounts of training/validation data, relatively simple conditional dependence structure.

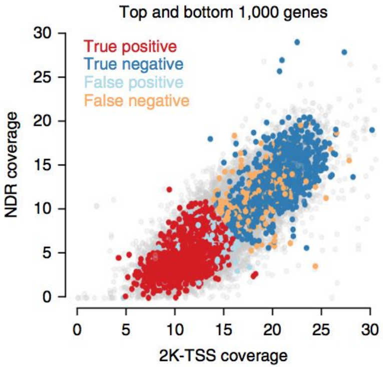
Biomarker discovery involves complex confounders, small N, large data per sample, and relatively few labels. Fundamental challenge: we are constrained by the information content in our **label set**, not our feature set.

Idea: use external data sets and problems to provide the information for learning structure; save label bits for our own problem: **multi-task** and **transfer learning.**

Vision: transfer learning for biomarker discovery

Prior limitation was a lack of total label data to train structure in statistical models. Newer methods allow us to integrate large-scale external data sets with high-dimensional, multi-parametric data extracted from newer assay technology.





FANTOM
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

TCGA

The 10,000 Immunomes

ENCODE

GTEx

Example: expression and genotype data from cfDNA.

Ulz P et al. Nat Genet 2016
Snyder MW et al. Cell 2016

Conclusions

Summary

1. Variant-calling based early detection has serious limitations on sensitivity:
   a. Possibly insufficient cfDNA for it to work at all.
   b. Multiple OOM in sequencing cost away from commercial viability.
   c. Significant specificity challenges from somatic heterogeneity.

2. Markers beyond ctDNA have potential but pose daunting statistical challenges.

3. Advances in multi-task and transfer learning may offer a route to statistically robust biomarker development.

freenome

A vision for the next 5 years

1. Large, publicly-available reference compendia of cell-, tissue-, and population-specific genomic information enable large-scale machine learning of the causal structure behind cancer genesis, maintenance, and clearance.

2. ML-derived methods enable point-in-time population screening with greater accuracy and adherence for conditions with existing screens, and meet the unmet need in unsolved conditions.

3. Repeat longitudinal testing in the general population enables the calibration of tests to individual variability, minimizing overtreatment and informing discovery of fundamental biology.

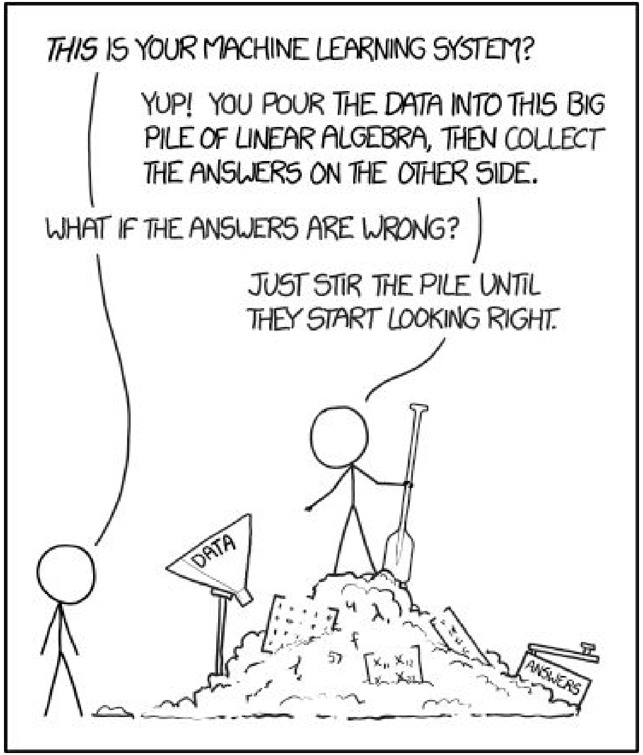
xkcd #1838

</talk>

ihaque@freenome.com

@ImranSHaque

@freenome


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

We're hiring: freenome.com/careers