Applications of Computational Biology and Machine Learning to Immuno-Oncology

Non-invasive approaches for cancer detection and monitoring

Robert B Scharpf

Johns Hopkins University

Department of Oncology

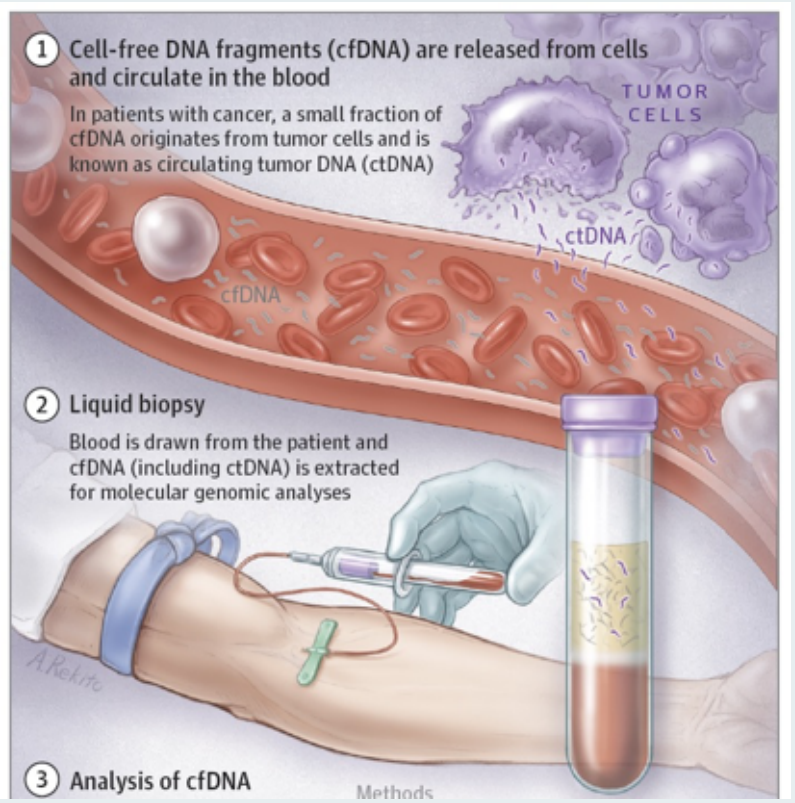
SITC Cancer Immunotherapy Winter School
Austin, TX | January 26, 2022

Disclosures

R.B.S. is a co-founder and consultant of Delfi Diagnostics, and owns Delfi Diagnostics stock subject to certain restrictions under university policy. Johns Hopkins University owns equity in Delfi Diagnostics.

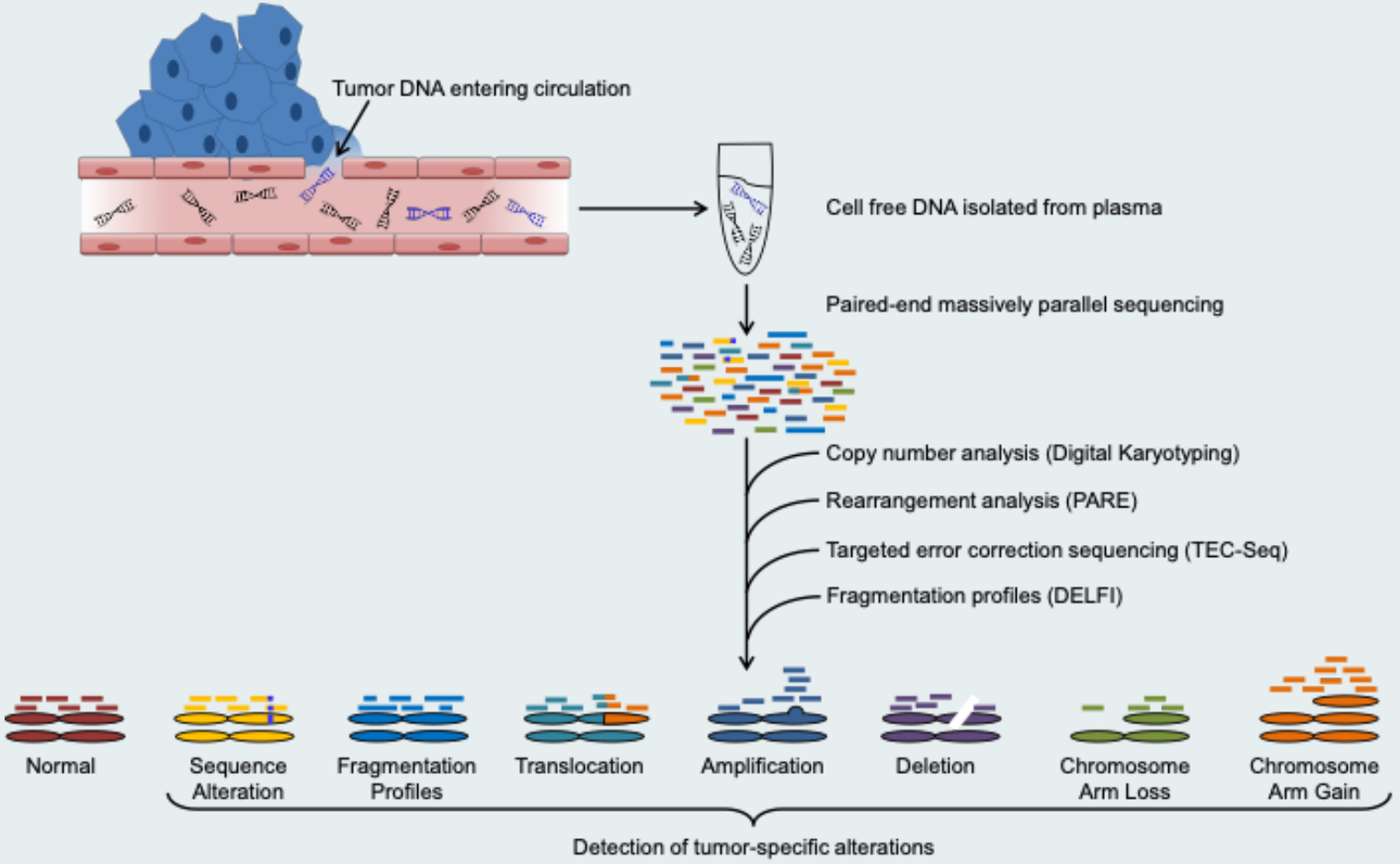
Liquid biopsies

- most cell-free DNA (cfDNA) is shed from hematopoietic cells
- for cancer patients, some cfDNA may be shed by tumor cells (ctDNA)



Husain and Velculescu, *JAMA*, 2017

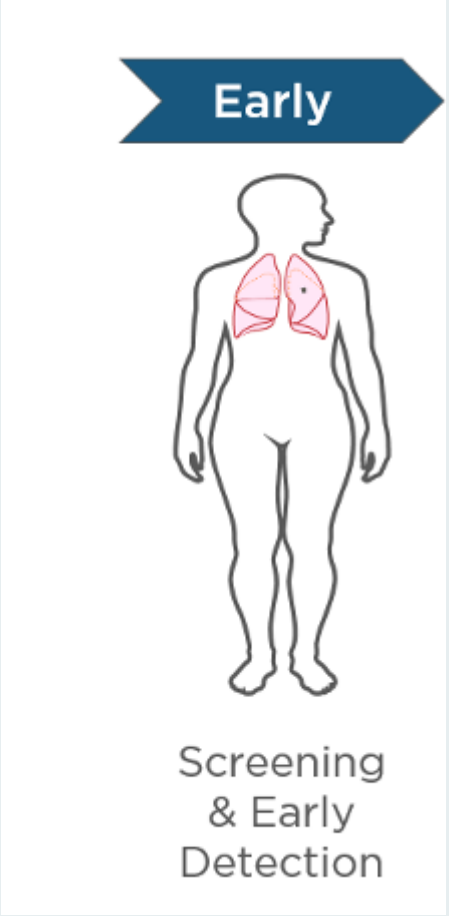
Liquid biopsy approaches



- Leary et al, *Science TM*, 2010
- Leary et al, *Science TM*, 2012
- Phallen et al, *Science TM*, 2017
- Cristiano et al, *Nature*, 2019
- Mathios et al, *Nature Comm*, 2021

Continuum of cancer care

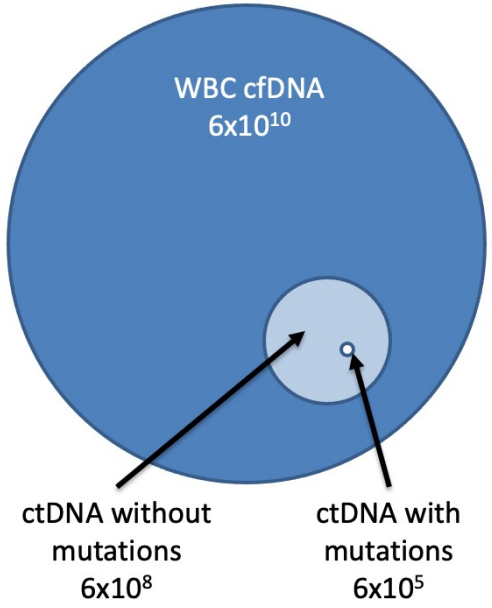
Early cancer detection and screening

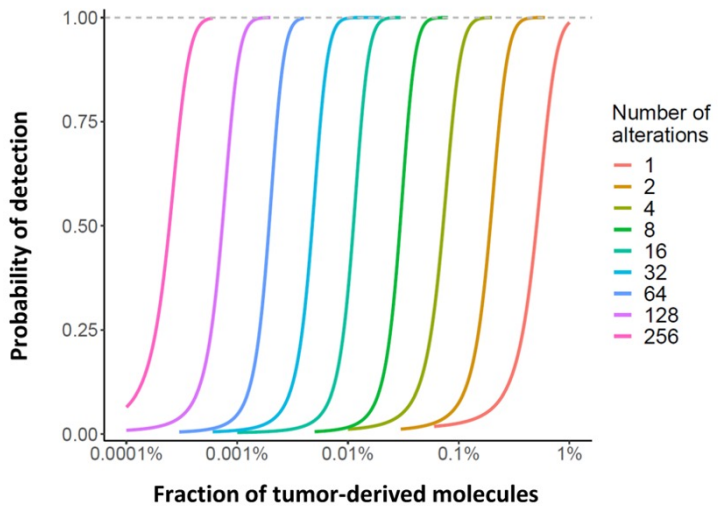


Challenges

- de novo detection required
- ctDNA/cfDNA may be very small while disease is still localized
- cancer prevalences in most screening populations are low; positive predictive value of a positive liquid biopsy test may be small

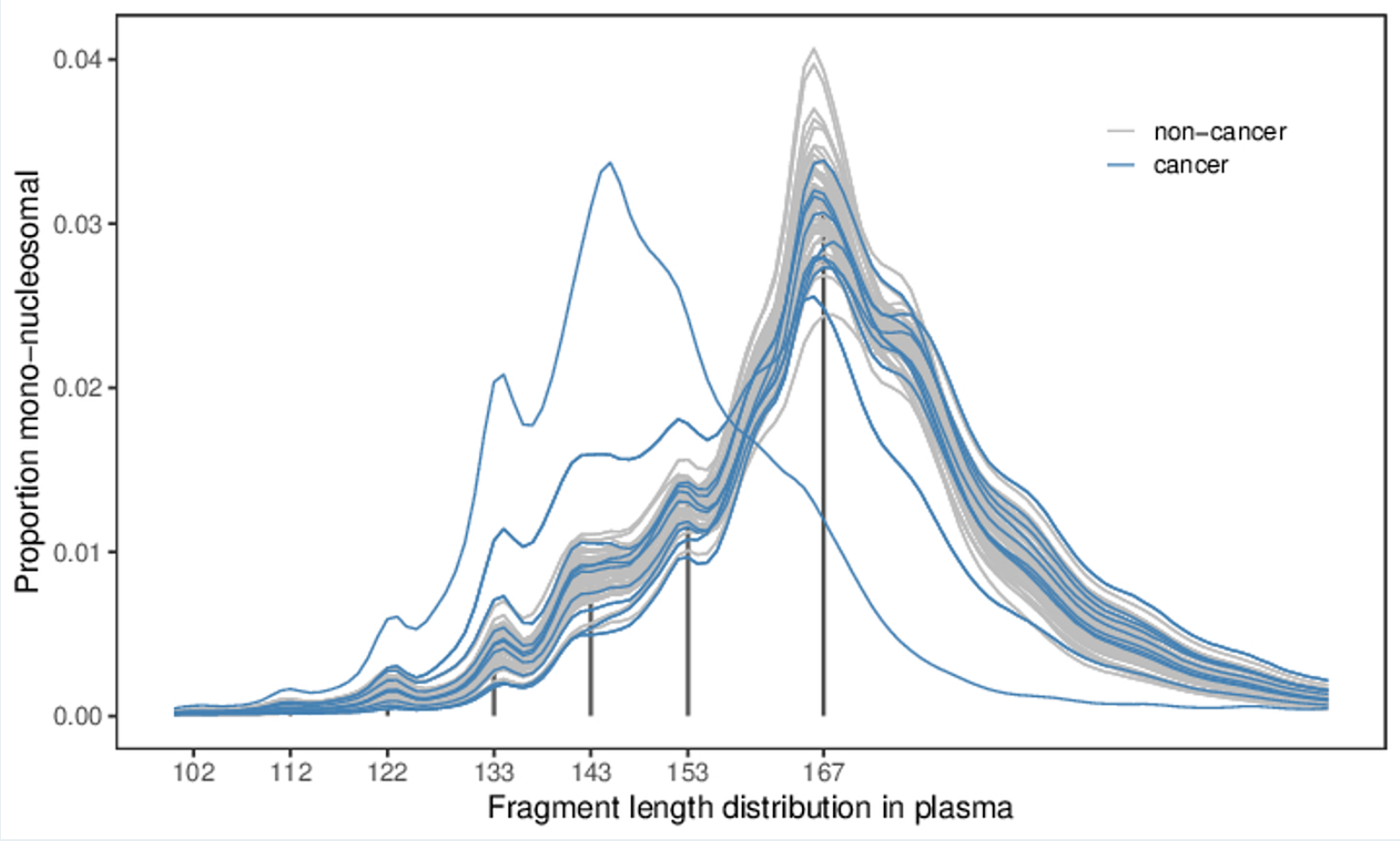
Fraction of ctDNA with mutations is small



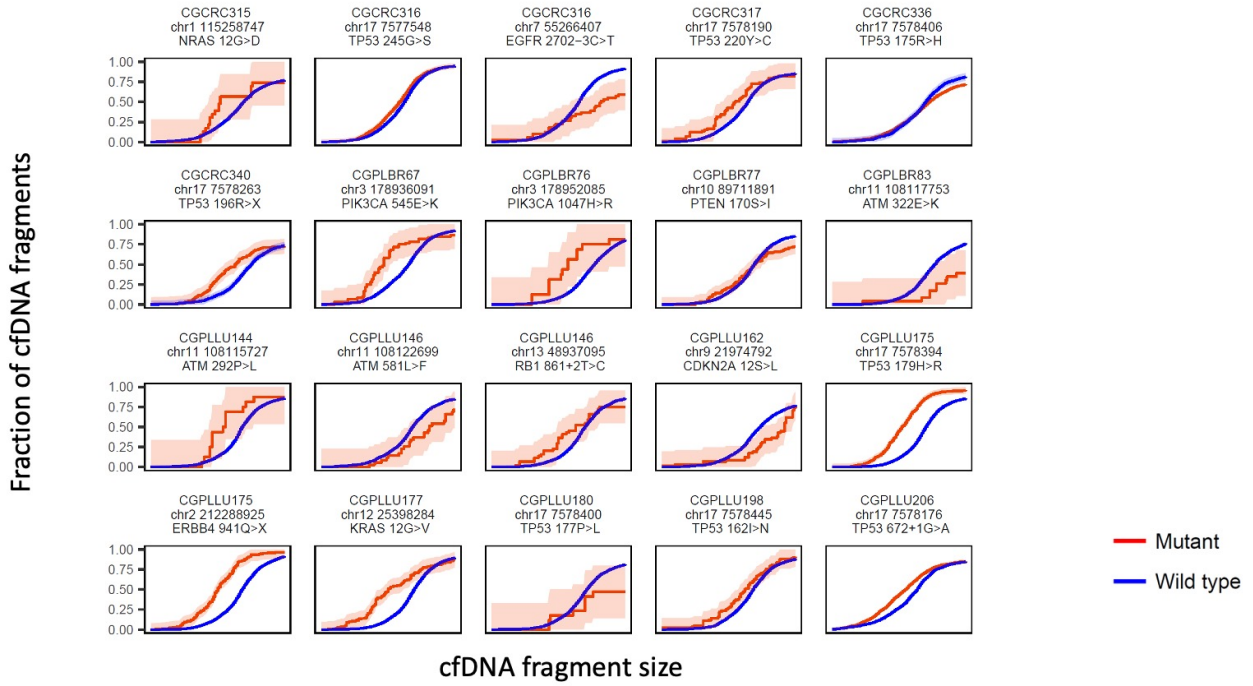


Cristiano et al., *Nature*, 2019

cfDNA fragments orginate from nucleosomes

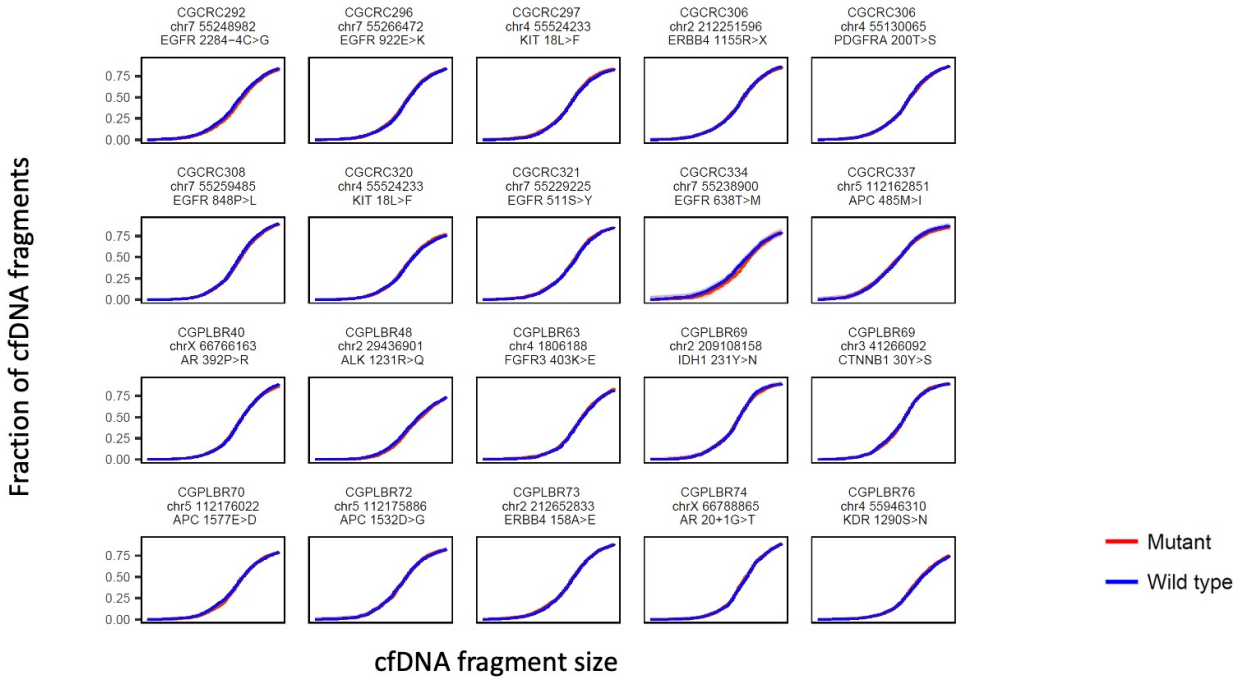


Altered fragment lengths of ctDNA



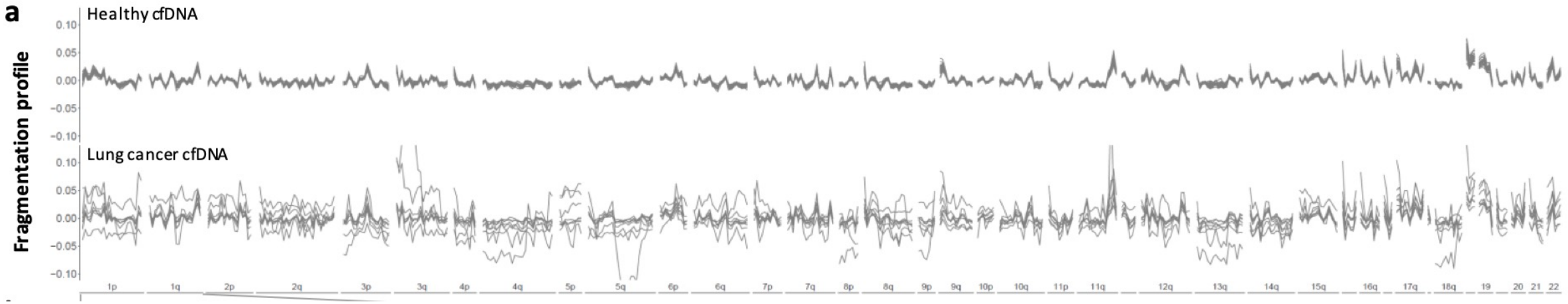
- Intra-patient comparison of fragment lengths

Negative control

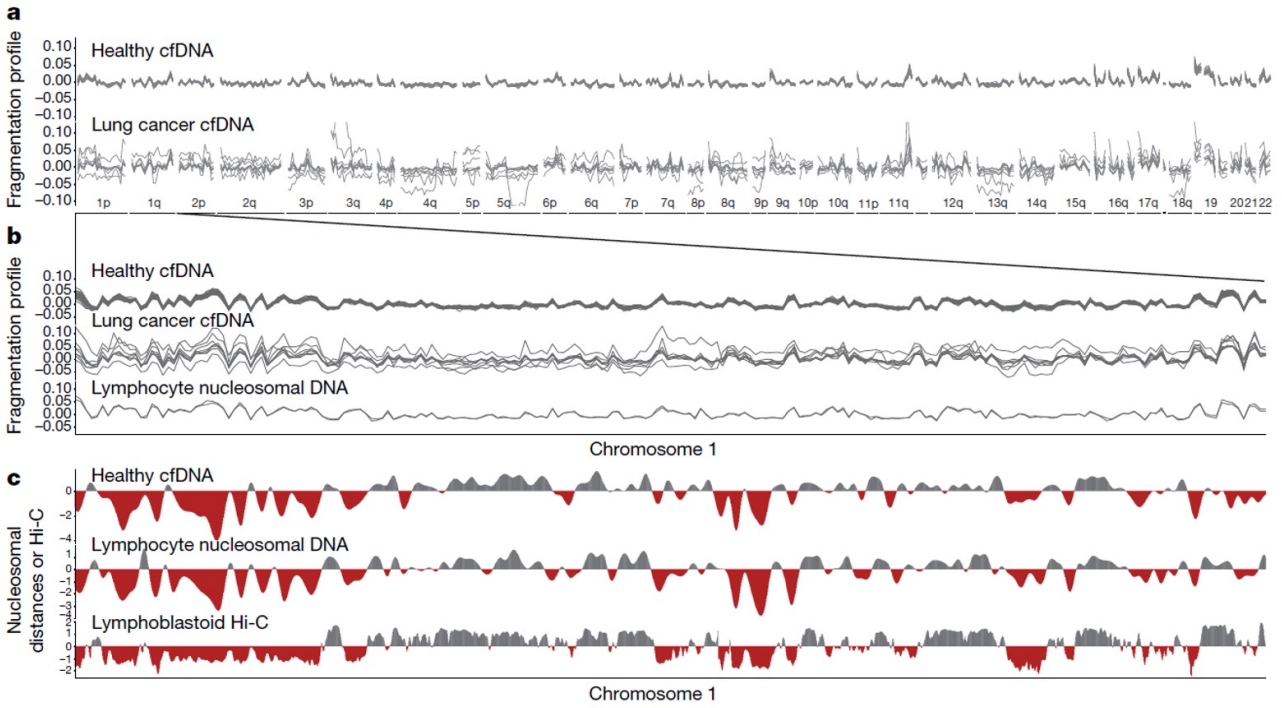


- 20 loci where germline alterations were detected

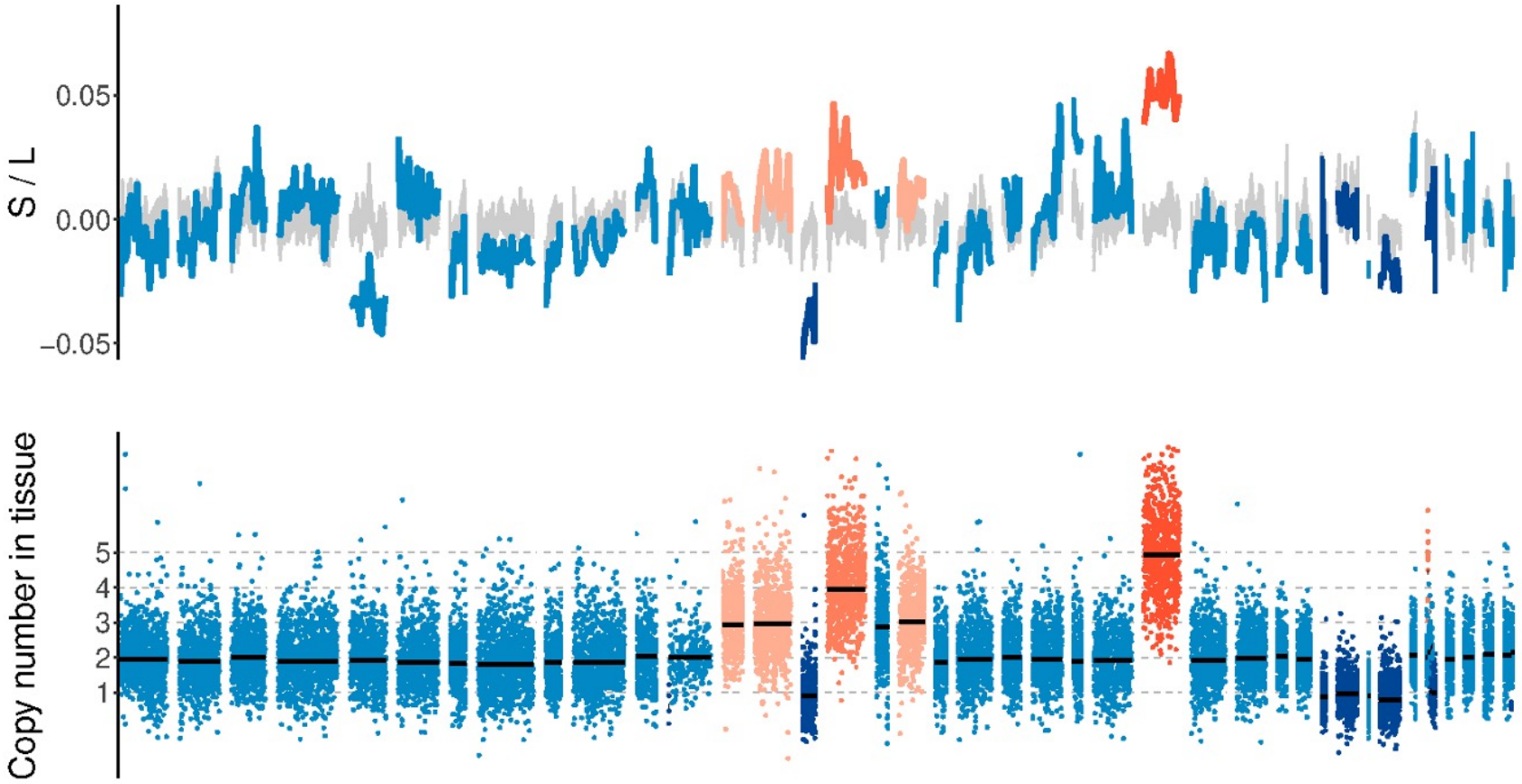
Genome-wide alterations of fragmentation patterns



A/B compartments (1 - 2x coverage)



Genomic and chromatin changes



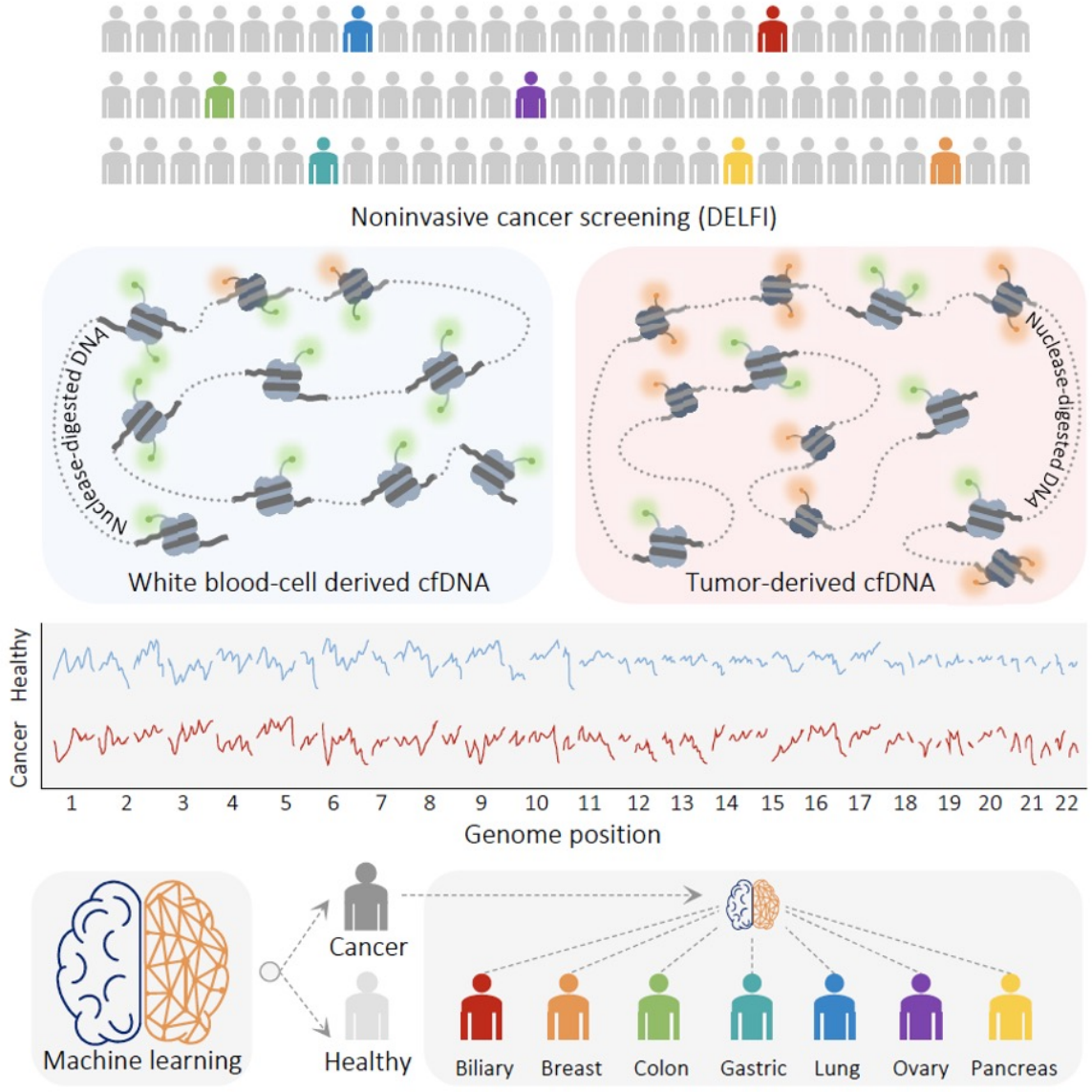
nature
International weekly journal of science

https://doi.org/10.1038/s41586-019-1272-6

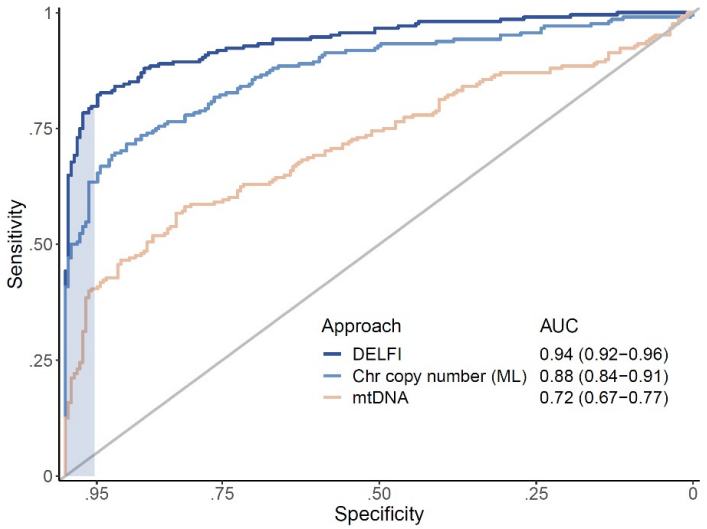
Genome–wide cell–free DNA fragmentation in patients with cancer

Stephen Cristiano[1,2,15], Alessandro Leal[1,15], Jillian Phallen[1,15], Jacob Fiksel[1,2,15], Vilmos Adleff[1], Daniel C. Bruhm[1], Sarah Østrup Jensen[3], Jamie E. Medina[1], Carolyn Hruban[1], James R. White[1], Doreen N. Palsgrove[1], Noushin Niknafs[1], Valsamo Anagnostou[1], Patrick Forde[1], Jarushka Naidoo[1], Kristen Marrone[1], Julie Brahmer[1], Brian D. Woodward[4], Hatim Husain[4], Karlijn L. van Rooijen[5], Mai–Britt Worm Ørntoft[3], Anders Husted Madsen[6], Cornelis J. H. van de Velde[7], Marcel Verheij[8], Annemieke Cats[9], Cornelis J. A. Punt[10], Geraldine R. Vink[5], Nicole C. T. van Grieken[11], Miriam Koopman[5], Remond J. A. Fijneman[12], Julia S. Johansen[13], Hans Jørgen Nielsen[14], Gerrit A. Meijer[12], Claus Lindbjerg Andersen[3], Robert B. Scharpf[1,2]* & Victor E. Velculescu[1]*

DELFI


Noninvasive cancer screening (DELFI)
Nuclease-digested DNA
White blood-cell derived cfDNA
Nuclease-digested DNA
Tumor-derived cfDNA
Healthy
Cancer
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Genome position
Machine learning
Cancer
Healthy
Biliary
Breast
Colon
Gastric
Lung
Ovary
Pancreas

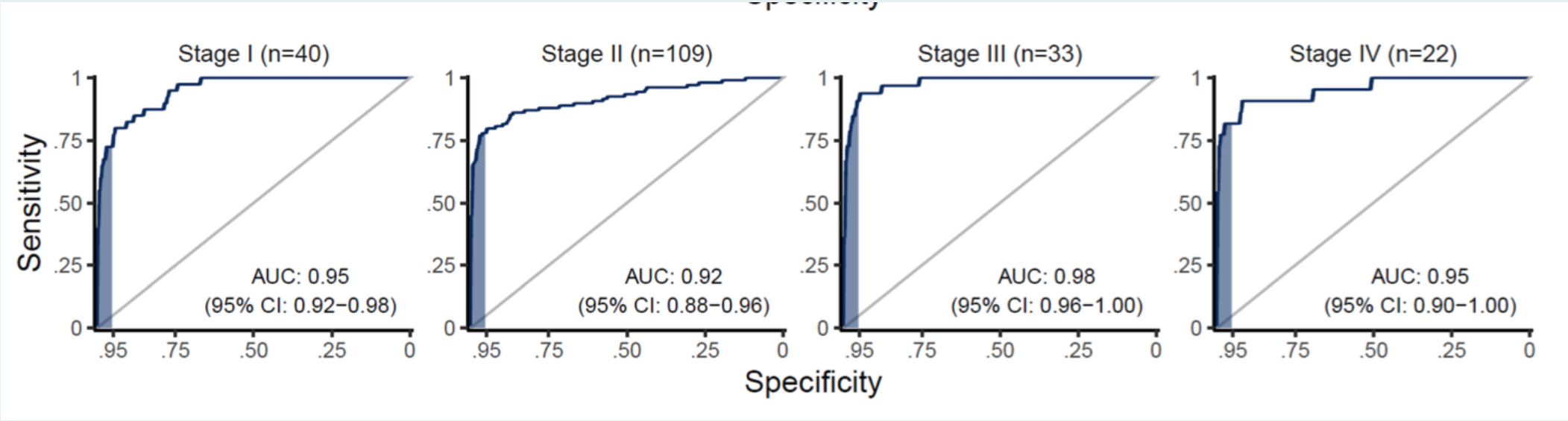
High sensitivity and specificity overall



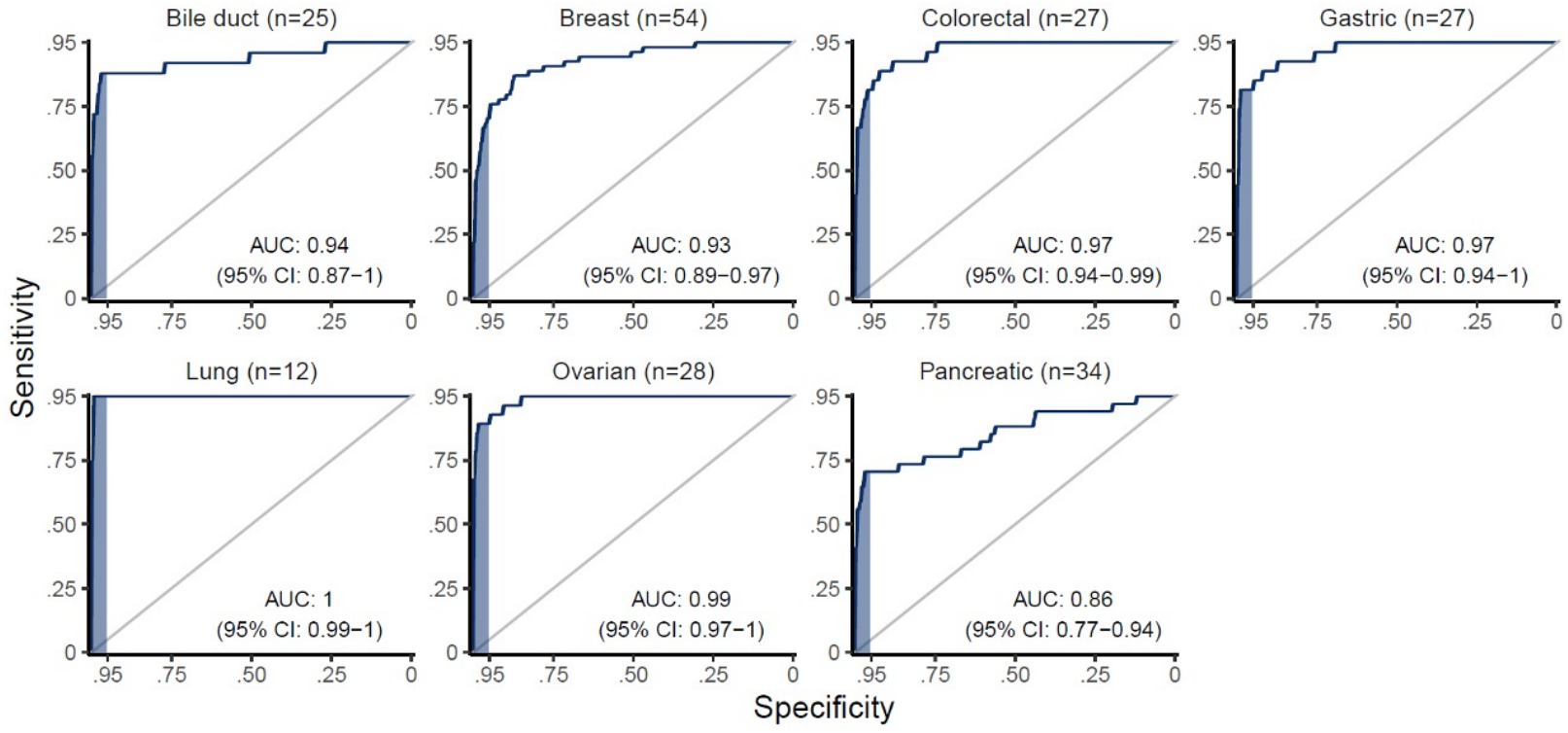
- gradient boosted model

Cristiano et al., *Nature*, 2019

By stage


Stage I (n=40)
AUC: 0.95
(95% CI: 0.92−0.98)
Stage II (n=109)
AUC: 0.92
(95% CI: 0.88−0.96)
Stage III (n=33)
AUC: 0.98
(95% CI: 0.96−1.00)
Stage IV (n=22)
AUC: 0.95
(95% CI: 0.90−1.00)
Sensitivity
Specificity
1
.75
.50
.25
0
.95
.75
.50
.25
0

By histology



Statistical challenges and rationale

Genome-wide cfDNA fragmentation data

Participants (n) :

- ~400 (~200 cancer cases and ~200 non-cancers)
  - Non-cancers were part of a screening population
  - Multiple cancer types: ovarian, lung, breast, and 5 others

cfDNA features (p) :

- ~1000 statistical summaries of cfDNA fragmention
- 39 measures of chromosome arm aneuploidy (z-scores)
- mitochondrial representation (single number)

Our primary goal is prediction

- Can we distinguish patients with cancer from individuals without cancer?
- We are willing to trade some interpretability for more sensitive and specific detection of cancer
- $n << p$

Prediction task is complex

- How do fragmentation profiles across the genome differ from fragmentation profiles we observe in non-cancer populations?
- The combination of features useful for prediction may not necessarily explain important biological pathways or provide mechanistic insight

Machine learning

> In machine learning, a model learns from examples rather than being programmed with rules

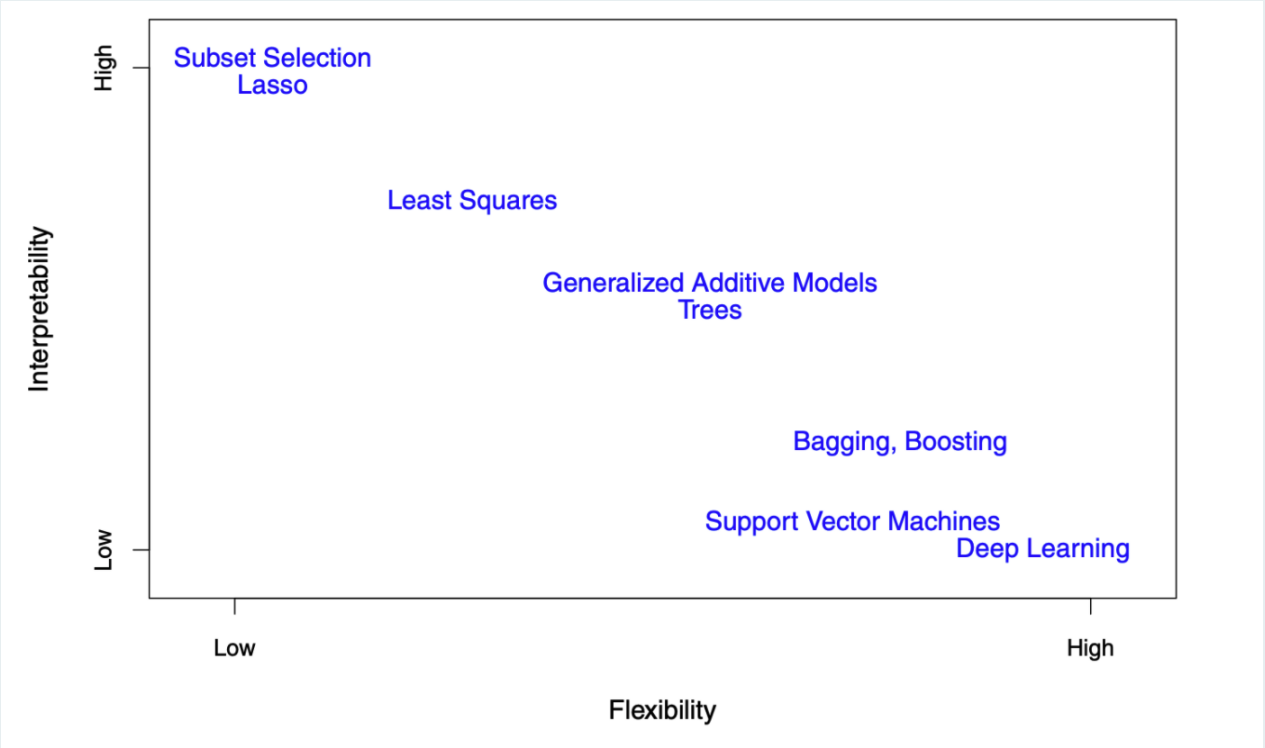
Rajkomar et al., *NEJM*, 2019

- Features: the inputs (fragmentation characteristics)
- Labels: cancer or non-cancer

Tradeoff of model complexity and interpretability

n << p:

- "deep learning" approaches will not outperform simpler models
- hypothesis-driven approaches that leverage known biology and features can lead to more useful and replicable models



Friedman, Tibshirani, and Hastie, *Elements of Statistical Learning*

Mechanics

1. specify resampling-based approach for cross-validation
2. specify a machine learning architecture (logistic regression, random forest, etc)
3. train and test
4. summarize predictive performance
5. specify a *final* model

Cross-validation involves a lot of repetition

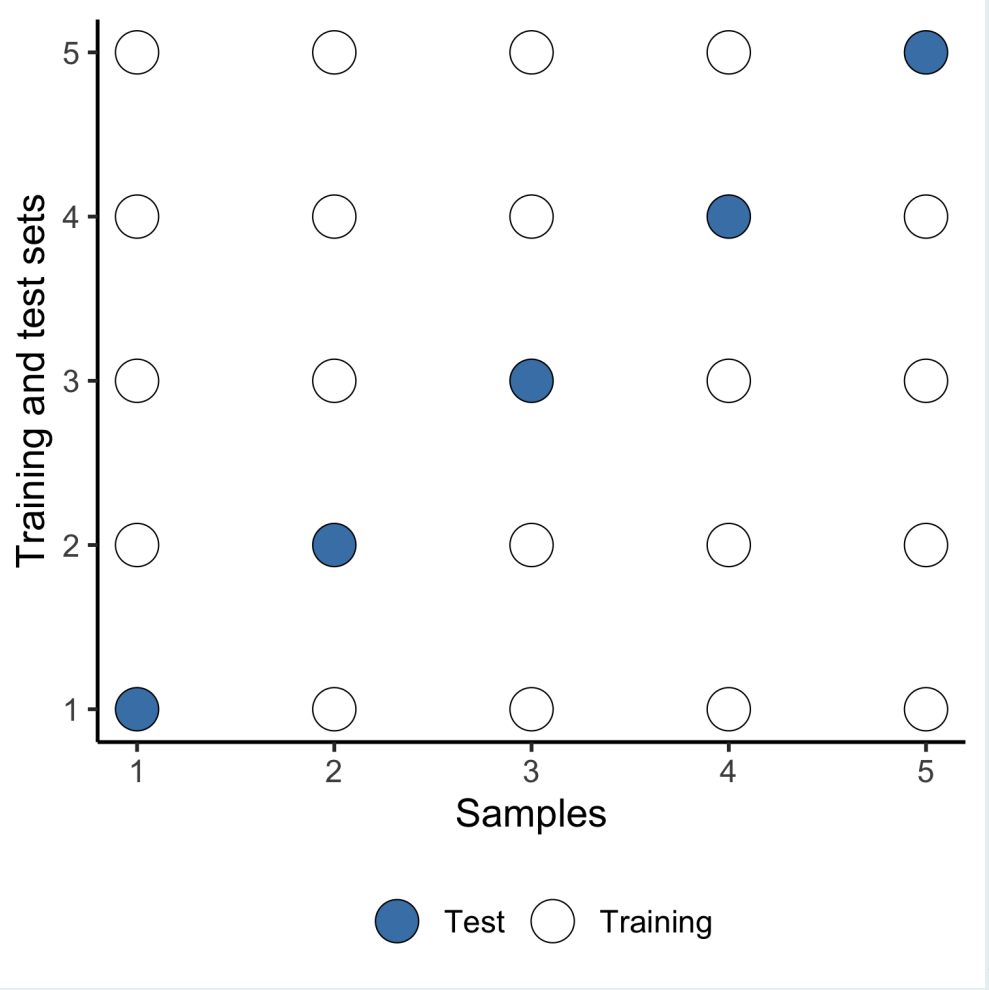
- Easy to make mistakes

Use existing infrastructure for:

- resampling (`rsample`)
- model specification (`tidymodels`, `caret`)
- tuning hyperparameters and nested cross-validation
- training and testing a model (`tidymodels`, `recipes`, `workflows`, and others)
- summarizing performance (`tidymodels`)

Resampling approach: 5-fold cross validation

```
library(tidyverse)
library(magrittr)
library(rsample)
ngenes <- 200
nsamples <- 50
x <- matrix(rnorm(ngenes*nsamples), ngenes, nsamples)
y <- factor(rbinom(nsamples, 1, 0.5))
dat <- t(x) %>%
    set_colnames(paste0("feature_", seq_len(ncol(.)))) %>%
    as_tibble() %>%
    mutate(y=y)
training.test <- vfold_cv(dat, v=5)
```



Train and test

```
library(tidymodels)
lr_mod <- logistic_reg(mixture=1,
                       penalty=0.5) %>%
    set_engine("glmnet") %>%
    set_mode("classification")
lr_workflow <- workflow() %>%
    add_model(lr_mod) %>%
    add_formula(y ~ .)
set.seed(914952)
lr_fit <- lr_workflow %>%
    fit_resamples(training.test)
```

Summarize performance

```
collect_metrics(lr_fit)
```

```
## # A tibble: 2 × 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary       0.4     5  0.0447 Preprocessor1_Model1
## 2 roc_auc  binary       0.5     5  0      Preprocessor1_Model1
```

Repeated cross-validation

- Our initial randomization of patients to the 5 folds was arbitrary
    - performance assessment could be slightly biased because of the randomization
- Repeat the process r times and average the predictions across the r repeats

```
train.test.r10 <- vfold_cv(dat,
                           v=5,
                           repeats=10)
lr_fit_rcv <- lr_workflow %>%
    fit_resamples(train.test.r10)
```

Cross-validation or split-study validation?

- Resampling based methods such as bootstrap or repeated k-fold cross-validation are nearly always your best option

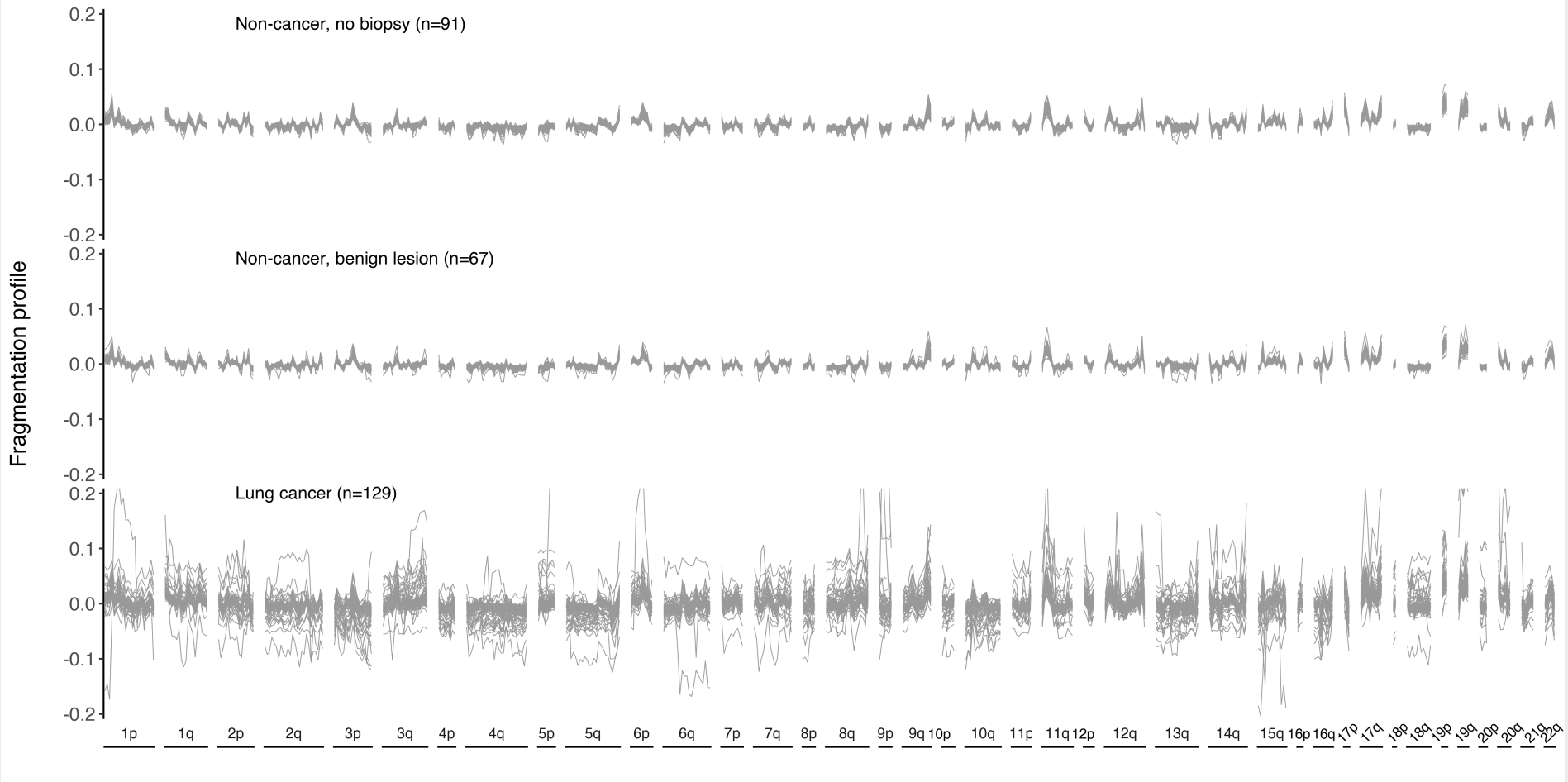
Exceptions:

- You have a large study (20,000+) and it does not bother you that you will not learn anything from 1/3 of the dataset
- Required for regulatory purposes
- Your dataset consists of multiple independent studies
  - `consider multi-study models`
  - `approaches for external validation`

Early detection of lung cancer

LUCAS study

Altered fragmentation profiles



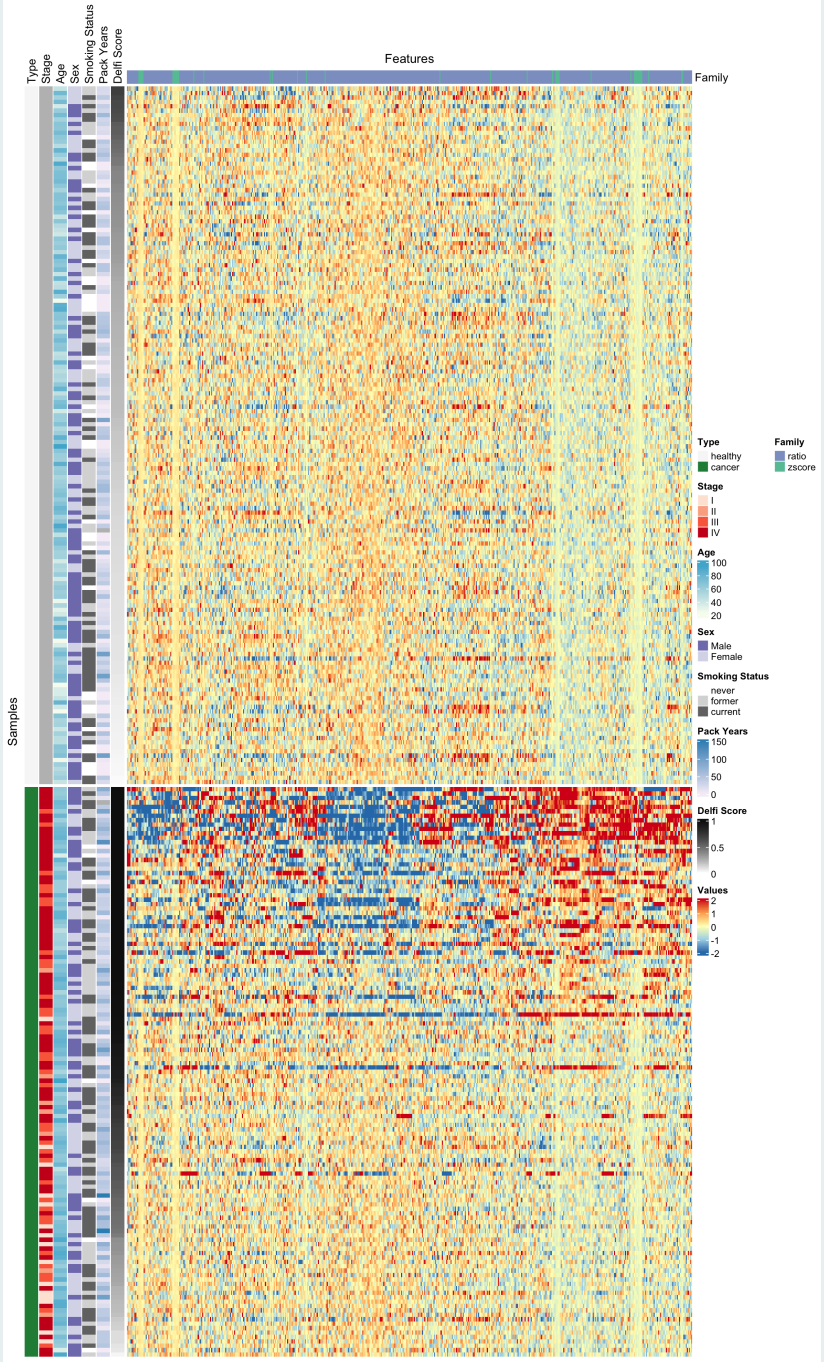
Machine learning model

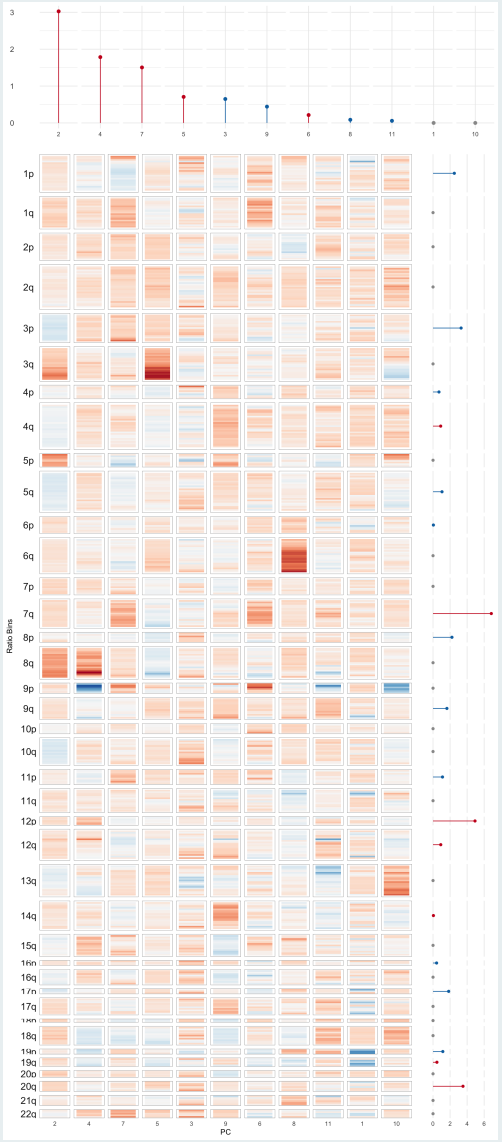
- Machine learner: penalized logistic regression

Features:

- PCA of the fragmentation profiles
- z-scores for chromosomal aneuploidy

278 x 504 matrix



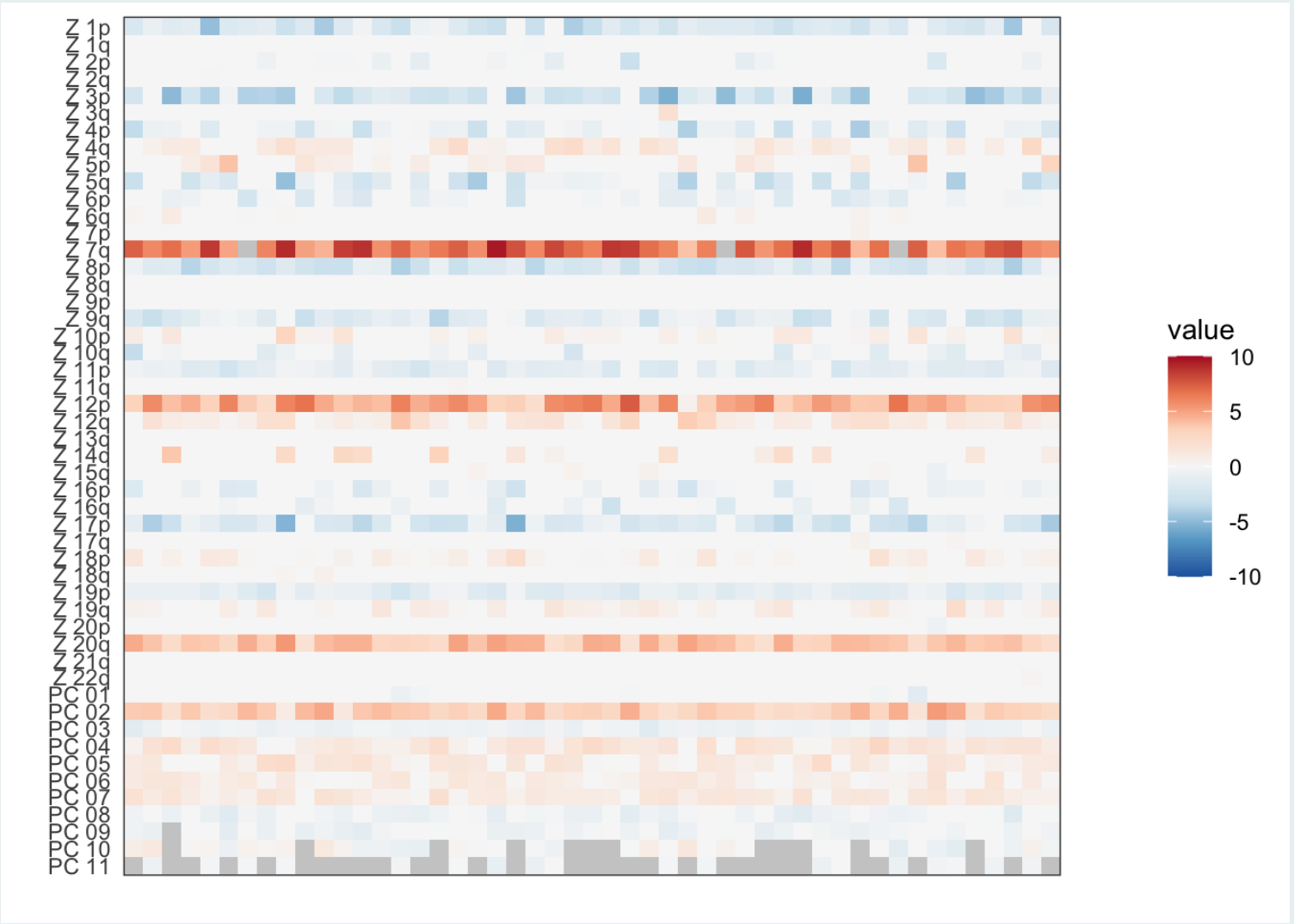


Dimensionality reduction

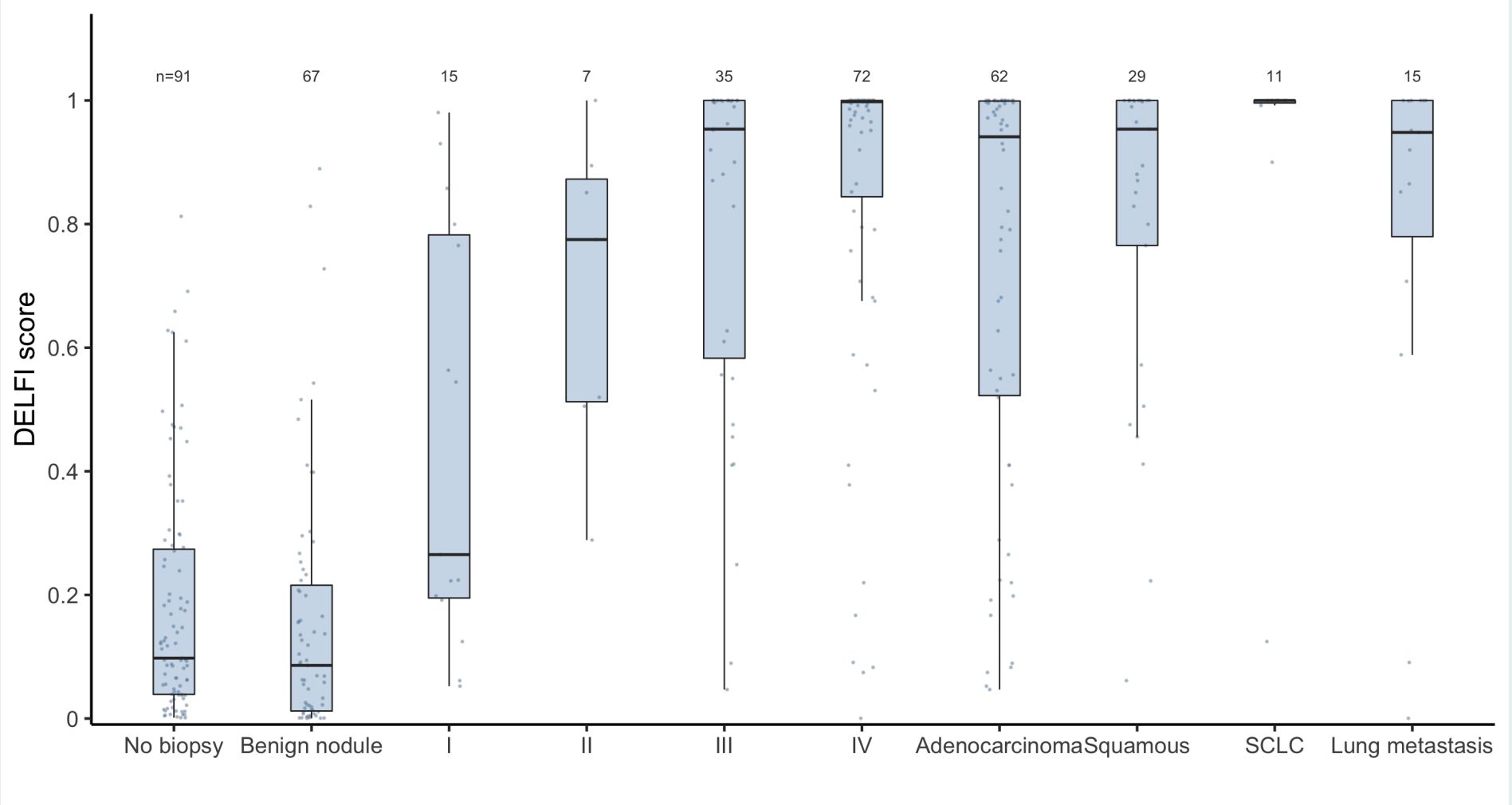
- Principal components that explain 90% of the variance across samples
    - 278 x 10 matrix
- Regression coefficients for chromosome arm-level copy number

Stochasticity of model across training sets

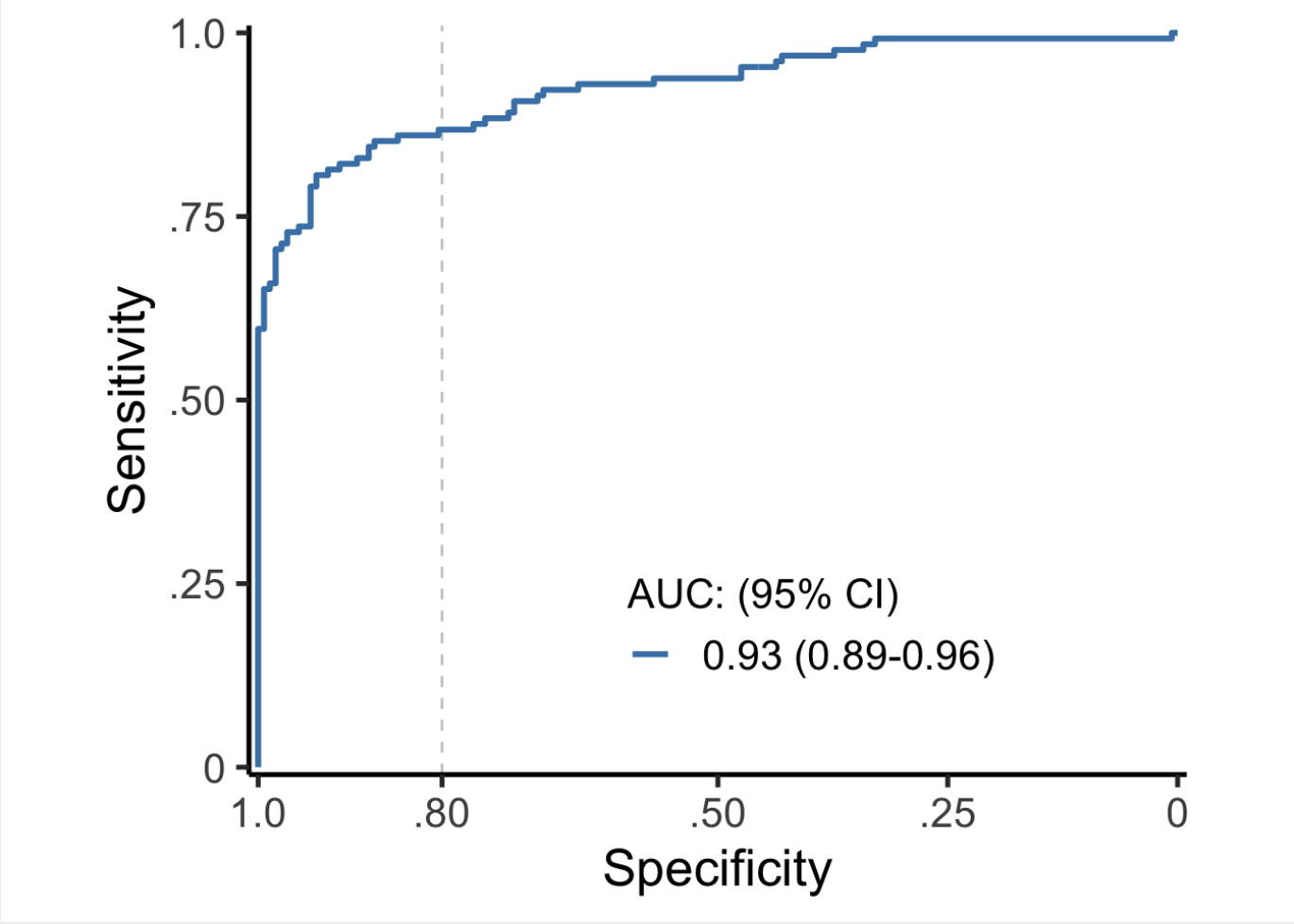
- Columns display 50 training sets from repeated cross-validation
- Rows are regression coefficients



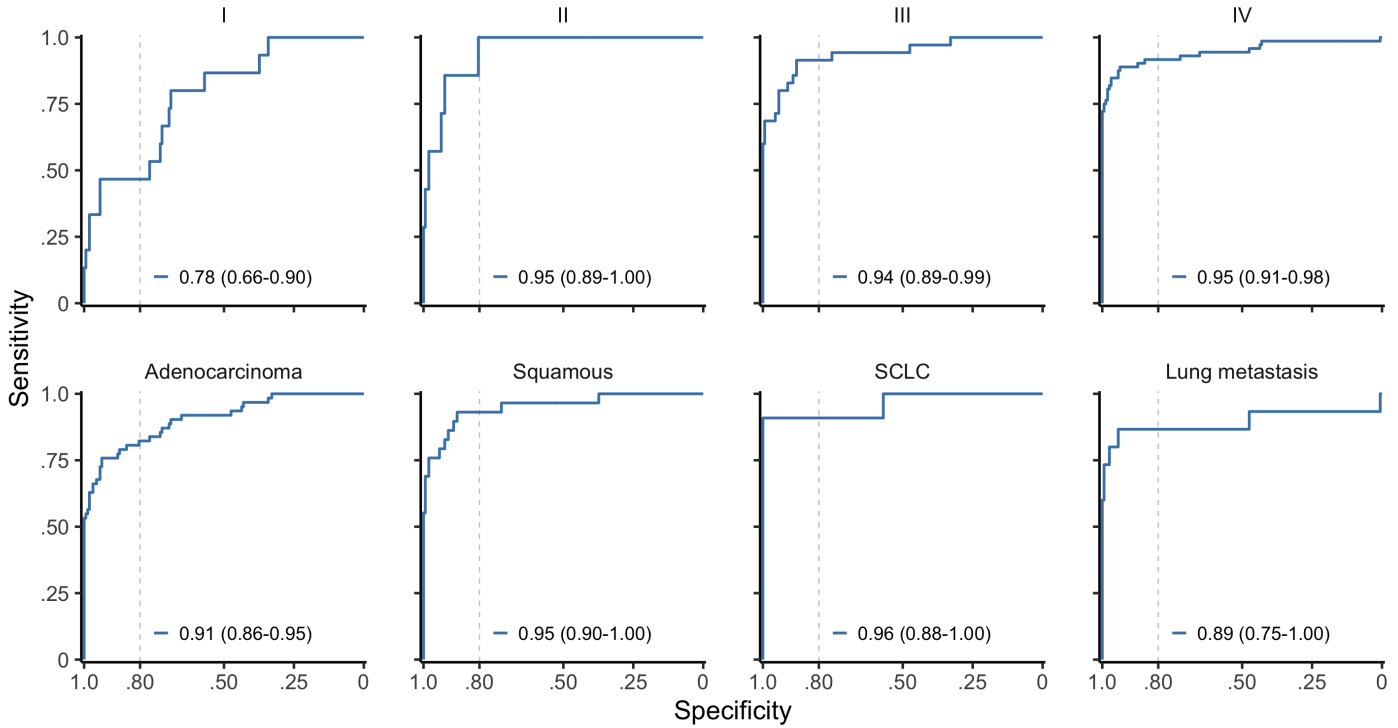
Scores by stage and histology



Internally cross-validated performance



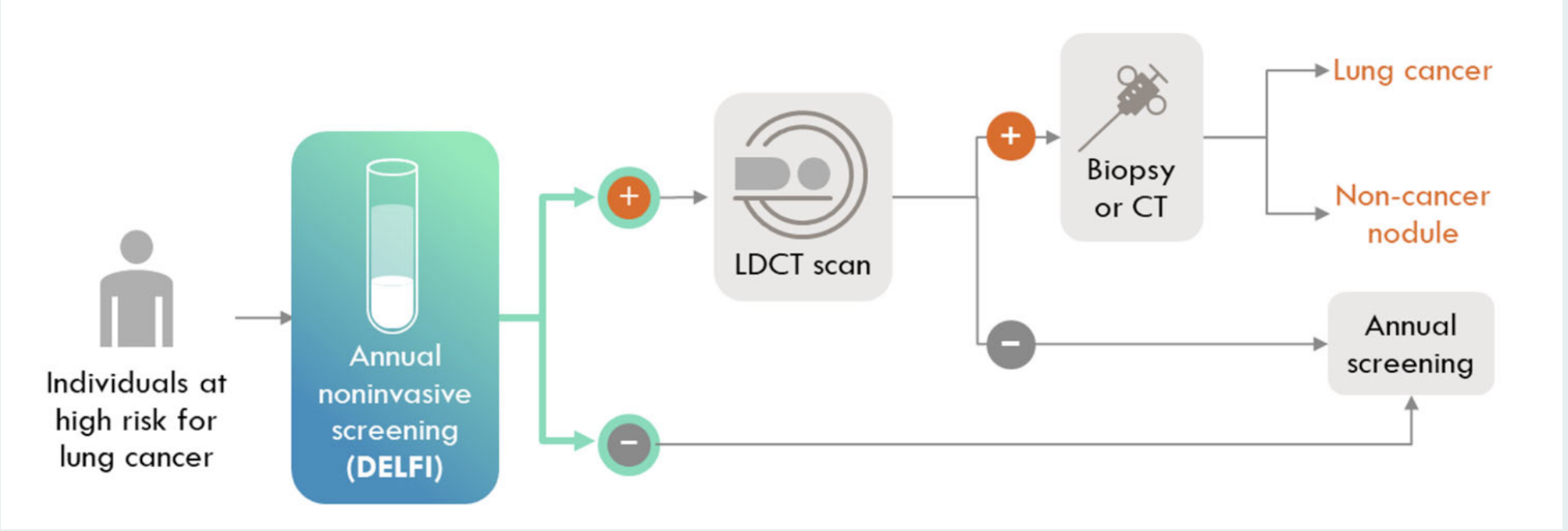
Internally cross-validated performance



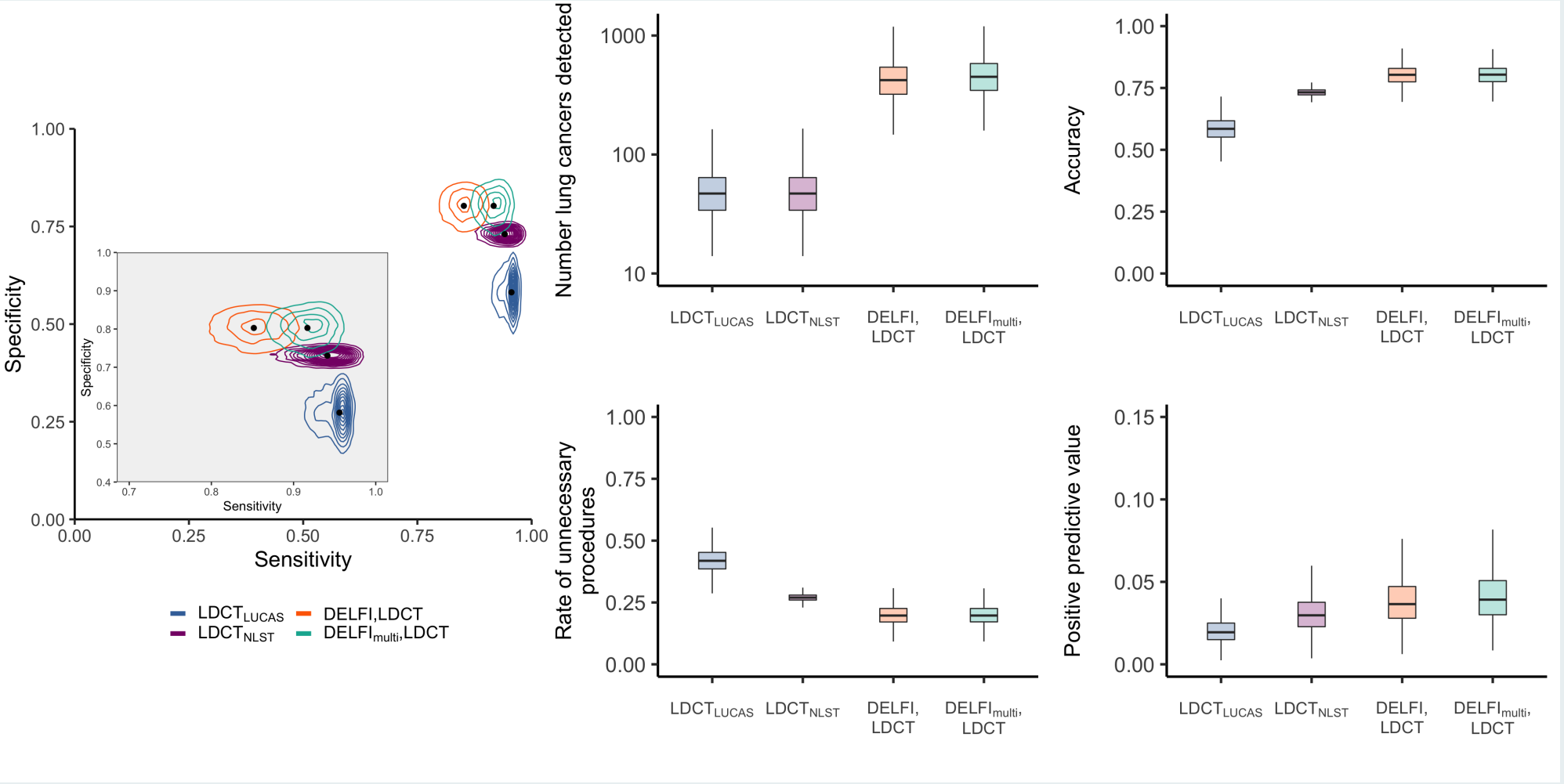
External validation

- Fixed model
  and cutoff at
  80% specificity

Sequential screening



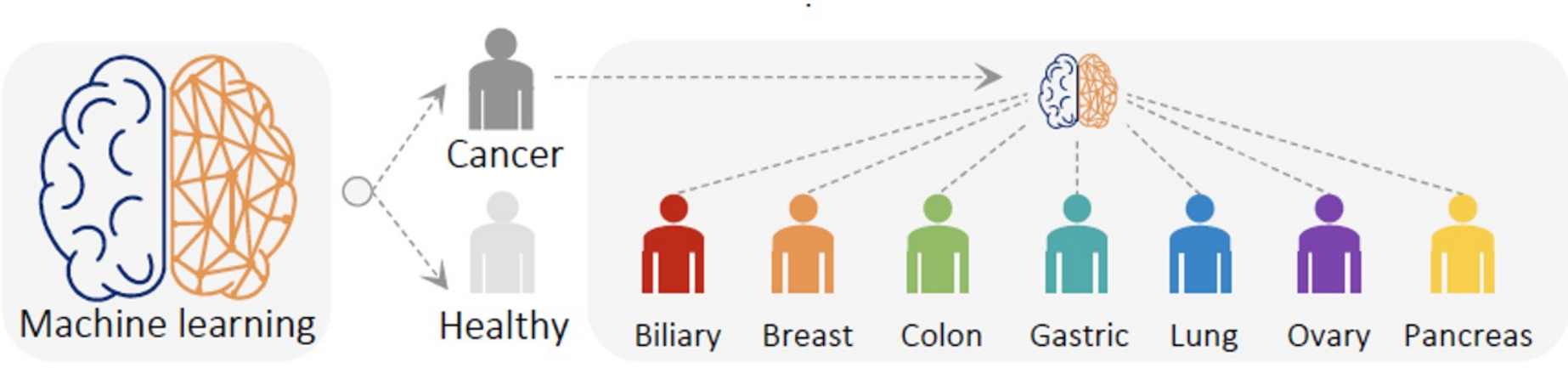
Sequential screening



Fragmentation profiles differ between cancer types

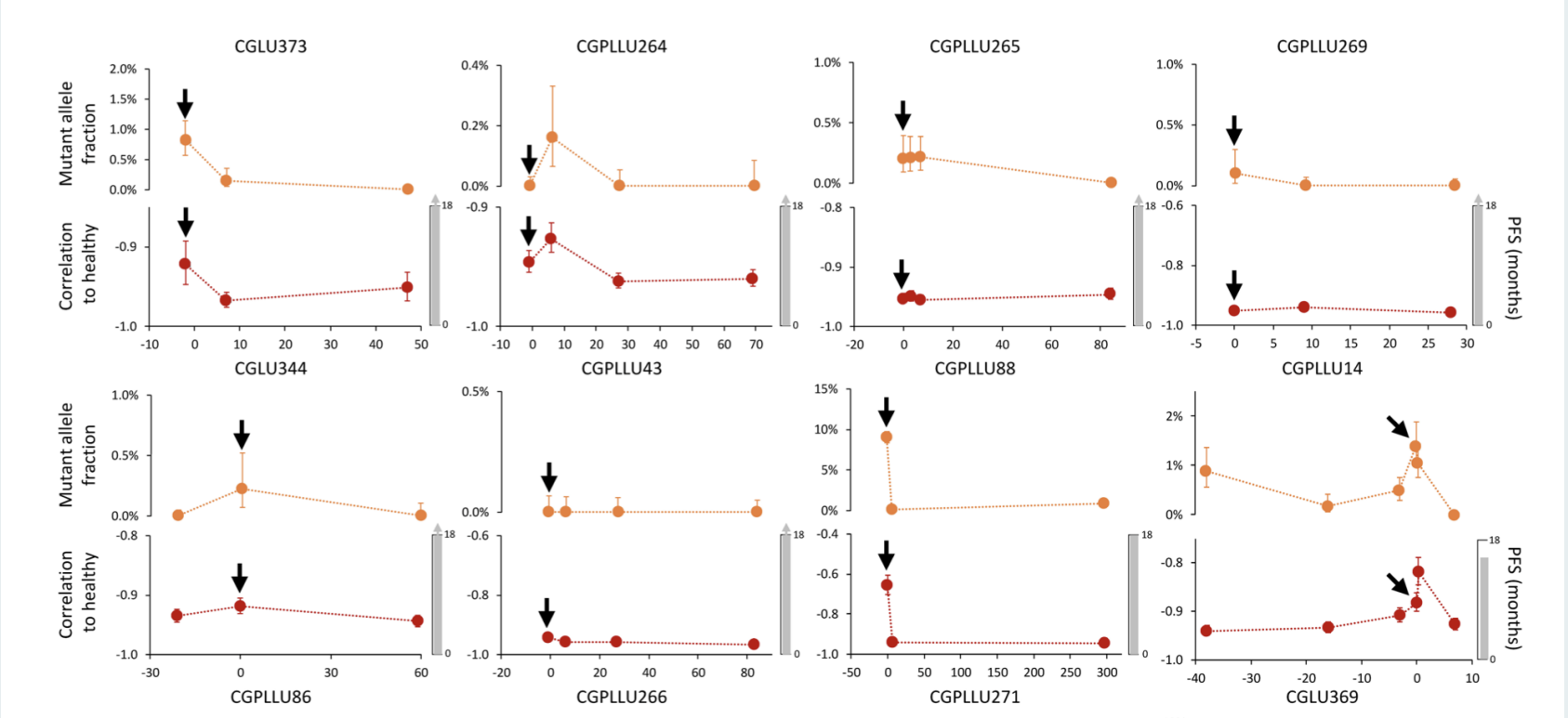
Cristiano et al., *Nature*, 2019

Tissue of origin



Cristiano et al., *Nature*, 2019

Disease monitoring



Cristiano et al., *Nature*, 2019

Machine learning for noninvasive detection and monitoring of cancer

- Genome-wide cfDNA fragmentation profiles reflect abnormal packaging and genomic content of cancer genomes
- Liquid biopsies can detect cancer recurrence early and provide opportunities for intervention
- With enough examples, gradient boosted models / random forests can do a reasonable job at feature selection and prediction
- Leverage known biology when possible especially for $n << p$
    - fragments derived from nucleosomes, aneuploidy
    - coverage at transcription start sites
- Multi-study models, ensembling machine learners, and approaches for external cross-validation can lead to more replicable classifiers

Acknowledgements

- Vilmos Adleff
- Akshaya Annapragada
- Jacob Carey
- Stephen Cristiano
- Jacob Fiksel
- Alessandro Leal
- Dimitrios Mathios
- Noushin Niknafs
- Jillian Phallen
- Victor Velculescu