# “Orchid: a novel management, annotation and machine learning framework for analyzing cancer mutations”

Team 2Js: Jamie Canderan and John Metzcar

# Outline

- Motivation to study cancer and applications of machine learning in cancer
    - Cancer of unknown primary (CUP)
    - Liquid biopsy
- orchid
    - Development goals
    - Structure
    - Analysis in orchid
        - Random forests
    - Workflow in orchid
- orchid in context
    - Goals achieved
    - Current applications and potential goals

# Cancer

- Approximately 1 in 3 people will be diagnosed with cancer at some point in their lives
  - We will all know some one who experiences cancer, if you do not already.
- An estimated 1.7 million new cases of cancer were diagnosed in the US in 2018 and 0.61 million will die from cancer
- Worldwide: 14 million cancer cases and 8 million cancer related deaths
- Cancer death rates have been declining since the 1990s but clearly much work remains
- Orchid is a data and machine learning framework for cancer mutation analysis
  - Looks agnostically at DNA level mutations
  - Takes annotated feature sets of cancer mutation profiles from multiple databases
  - Uses SciKit methods to learn and make predictions on mutation profiles
  - Diagnosis of cancer of unknown primary (CUP) and screening via circulating free DNA are two areas of need that the orchid system could address

https://www.cancer.gov/about-cancer/understanding/statistics
Steward B. W. and Wild, C.P. Word Cancer Report 2014. IARC
https://www.cancer.gov/types/unknown-primary/patient/unknown-primary-treatment-pdq
https://www.cancer.gov/research/areas/screening

# Cancer of unknown primary (CUP)

- Rare disease resulting in 2% of all cancer cases
- Treatment aided when tissue of origin identified
- Currently only 20-30% of the primaries are able to be identified surgically (autopsy) while 75-85% are putatively identified via biopsy.
- Treatment is often guided by location of primary tissue (breast cancer gets breast cancer treatments)
- Machine learning can predict tissue of origin based on previously learned mutational profiles → knowing the tissue of origin may aid in treatment outcomes to quickly specify known effective treatments!

Pavlidis,N. and Pentheroudakis,G. (2010) Cancer of unknown primary site: 20 questions to be answered. Ann. Oncology
Pavlidis,N. and Pentheroudakis,G. (2012) Cancer of unknown primary site. Lancet, 379, 1428–1435.
Rassy E. and Pavlidis,N. (2018) The current evidence for a biomarker-based approach in cancer of unknown primary. Cancer Treatment Reviews 67 21–28

# Cancer detection and monitoring through circulating free DNA (cfDNA) (1)

- Cancer can sometimes be treated more effectively if
  - Diagnosed early (small tumor size/less invasive extent)
  - Treatment is well monitored
- Detection comes with challenges
  - Radiological exams are expensive
  - Radiological exams create exposure to other health risks – ionizing radiation and intravenous contrasts
  - Difficult to detect minimal, early disease
- Monitoring comes with the same challenges and more
  - Time series resolution can be critical for monitoring treatment
    - Radiological changes happen slowly versus functional changes of treatment
    - Expensive, time consuming, difficult for patient, increase patient risk

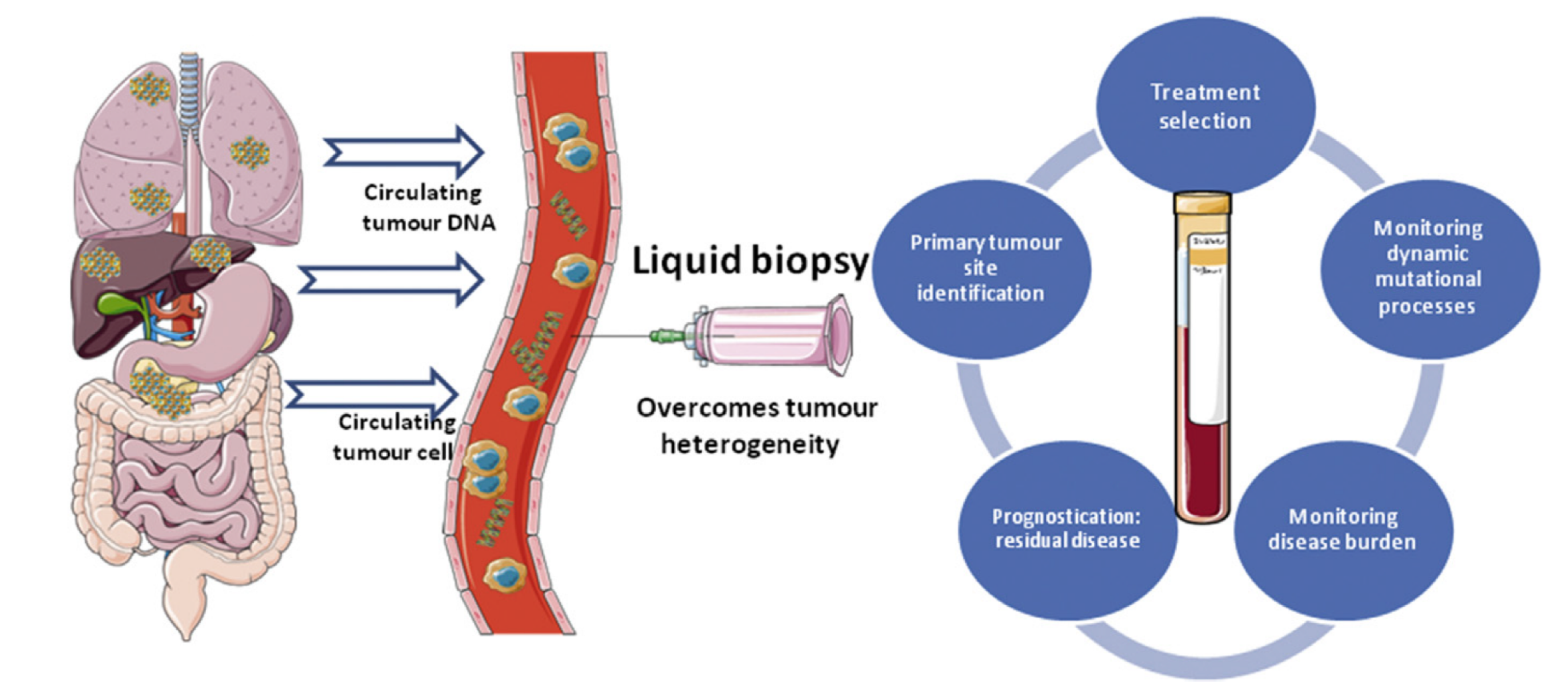
Steward B. W. and Wild, C.P. Word Cancer Report 2014. IARC
Han, X. Wang J., Sun, Y. (2016) Circulating Tumor DNA as Biomarkers for Cancer Detection

# Cancer detection and monitoring through circulating free DNA (cfDNA) (2)

- Circulating free DNA (cfDNA)
  - cfDNA is plasma (blood) derived short (<200 bp) fragments of DNA
  - Naturally present in low concentration
  - Present in higher concentration in the blood of pregnant women, cancer patients, and others
- Liquid biopsy
  - Biopsy (tissue sample) without surgery – aka blood draw
  - Machine learning can make predictions on the presence or absence of cancer based on features in cfDNA
  - Targetable mutations can be detected based on cfDNA
  - May open up diagnostics, prediction, and prognostics w/o previous issues

Snyder et al. (2016) Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. http://dx.doi.org/10.1016/j.cell.2015.11.050
E. El Rassy et al. (2018) Liquid biopsy: a new diagnostic, predictive and prognostic window in cancers of unknown primary European Journal of Cancer 105: 28-32

# Cancer detection and monitoring through circulating free DNA (cfDNA) (3)


Circulating tumour DNA
Circulating tumour cell
Liquid biopsy
Overcomes tumour heterogeneity
Treatment selection
Primary tumour site identification
Monitoring dynamic mutational processes
Prognostication: residual disease
Monitoring disease burden

E. El Rassy et al. (2018) Liquid biopsy: a new diagnostic, predictive and prognostic window in cancers of unknown primary European Journal of Cancer 105: 28-32

# orchid software goals

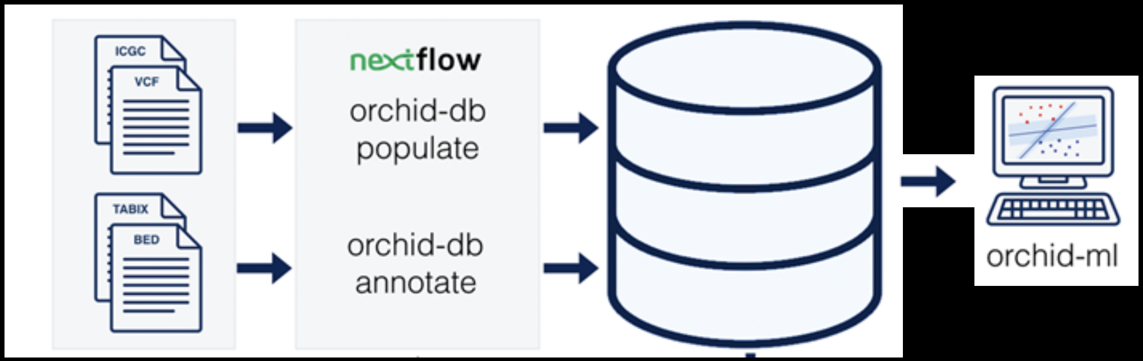
- How does one accomplish producing models with the ability to make predictions as stated above?

  → orchid!
- Orchid is a open source tumor mutation management and machine learning analysis framework
- Designed to address
  - Input from multiple databases
  - Management of millions of mutations
  - Hundreds of features
  - Any kind of annotatable mutation (coding or non-coding regions)
  - No a priori classification task

Cario and Witte. Bioinformatics, 34(6), 2018, 936–942 doi: 10.1093/bioinformatics/btx709

# Orchid Software

- orchid-db loads data from common file formats and annotates mutations into SQL database
- orchid-ml interfaces with pandas and scikit-learn to parse the database using machine learning algorithms



orchid-db populate

orchid-db annotate

orchid-ml

# orchid-db

- orchid-db is used to load data from database sources

International Cancer Genome Consortium

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

COSMIC
Catalogue Of Somatic Mutations In Cancer

- Files should be in the *vcf* format but a format convertor is available
- Downloaded data can have features annotated using *bed* or *tabix* files
  - Features can be enhancers, miRNA sites, chromatin states, promoters, etc.
- The data is processed and loaded into a MySQL database

# Data Annotation

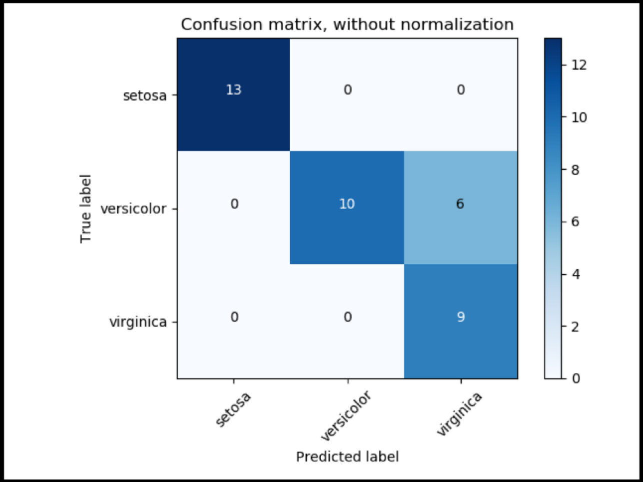
- Feature annotation: *SnpEff*
- Cancer gene network presence: *KEGG*
- Phylogenetic conservation: *phylop*
- Location within snoRNA and microRNA regions: *wgrna*
- Locations in predicted enhancers, promoters, and start sites: *segmentation*, *rfecs*, *dbsuper*, *encode*
- Locations in DNAse I sites: *dnase*
- Trinucleotide contexts, assorted composite scores (*funseq2*, *cadd*, *dann*)
- Others: *targetscans*, *remap*, *gwas*

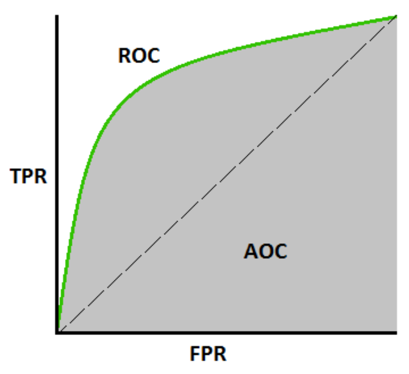
# orchid-ml

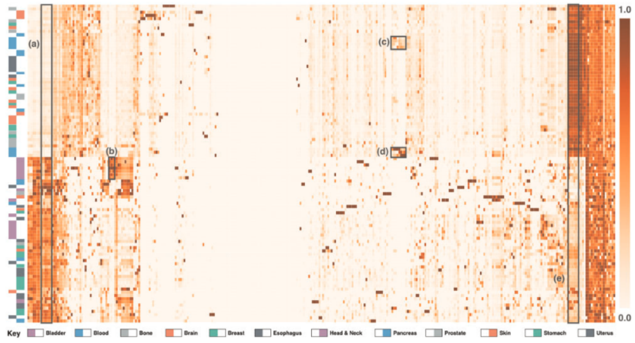
Standalone python module that can be imported into python script or notebook

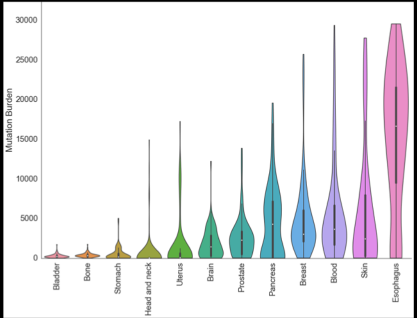
Creates MutationMatrix object used by built-in support vector machine (SVM) or random forest (RF) wrapper or by scikit-learn classifiers (PCA, k-means, others)

Various outputs and performance matrixes to evaluate models

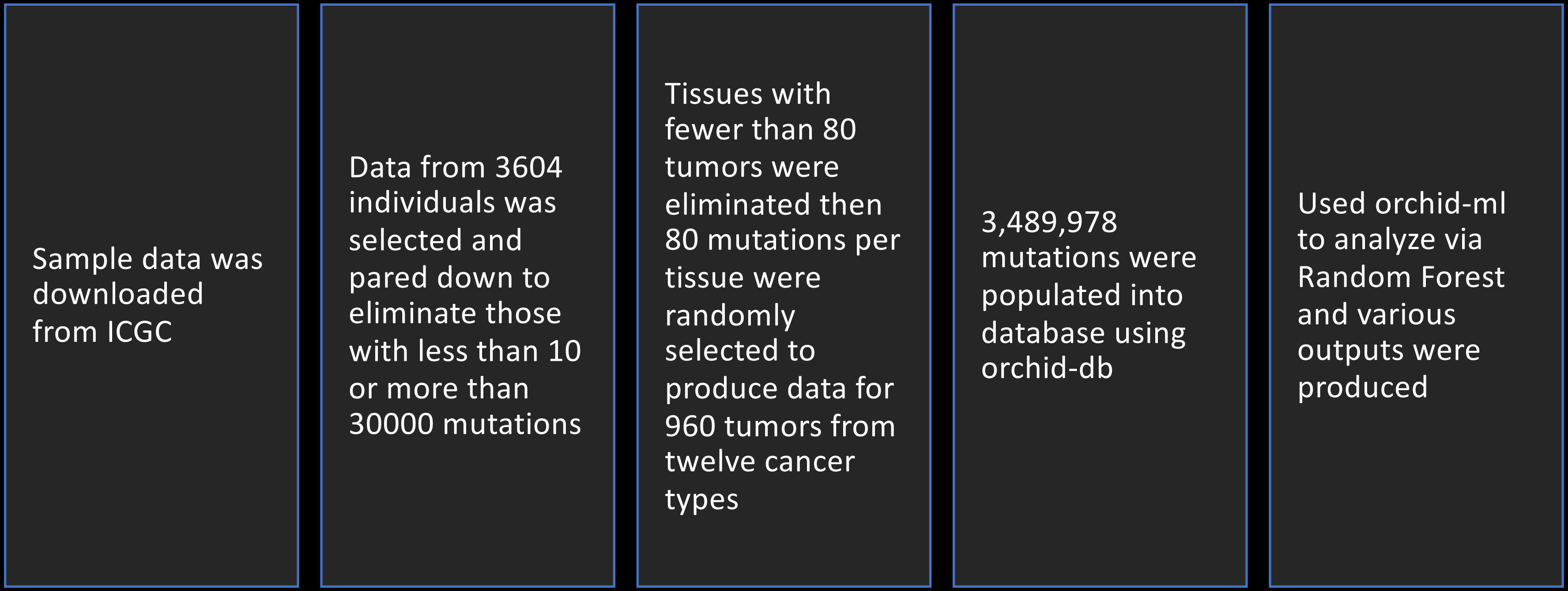








# Sample Data Analysis



Sample data was downloaded from ICGC

Data from 3604 individuals was selected and pared down to eliminate those with less than 10 or more than 30000 mutations

Tissues with fewer than 80 tumors were eliminated then 80 mutations per tissue were randomly selected to produce data for 960 tumors from twelve cancer types

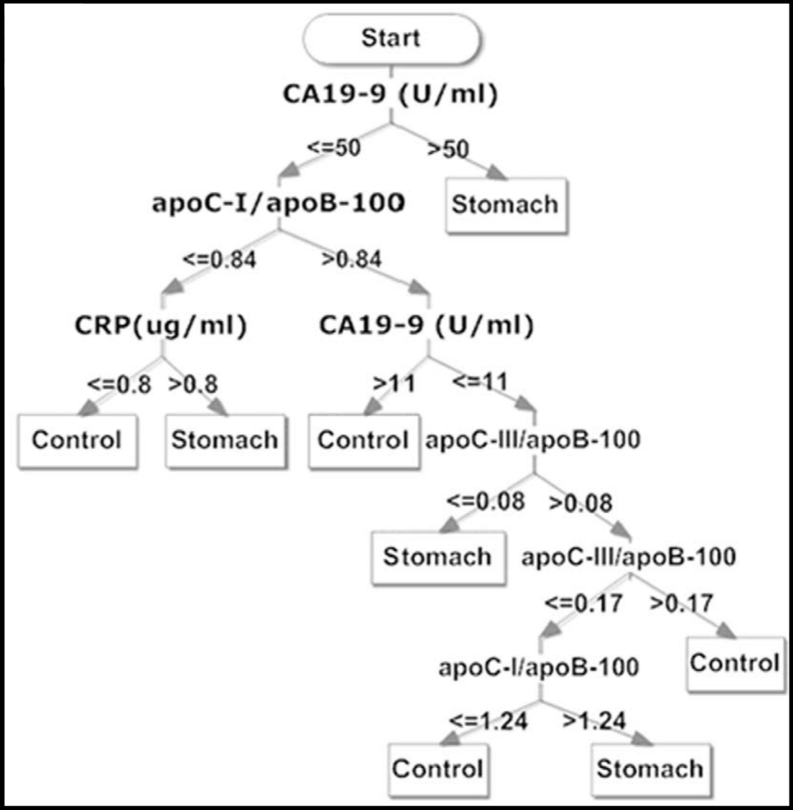
3,489,978 mutations were populated into database using orchid-db

Used orchid-ml to analyze via Random Forest and various outputs were produced

# Random Forest Classifiers: Decision Trees

- Random forests are ensembles of decision trees (get it?)
- A decision tree is at least a rooted bifurcating tree
  - Interior nodes – points of decision based on an observable
    - Ideally they homogenously split training data
  - Leaves – classifications
- Decision tree → encodes the questions required to reproducibly make a decision given a set of data

Ex: Determining the presence of gastric cancer

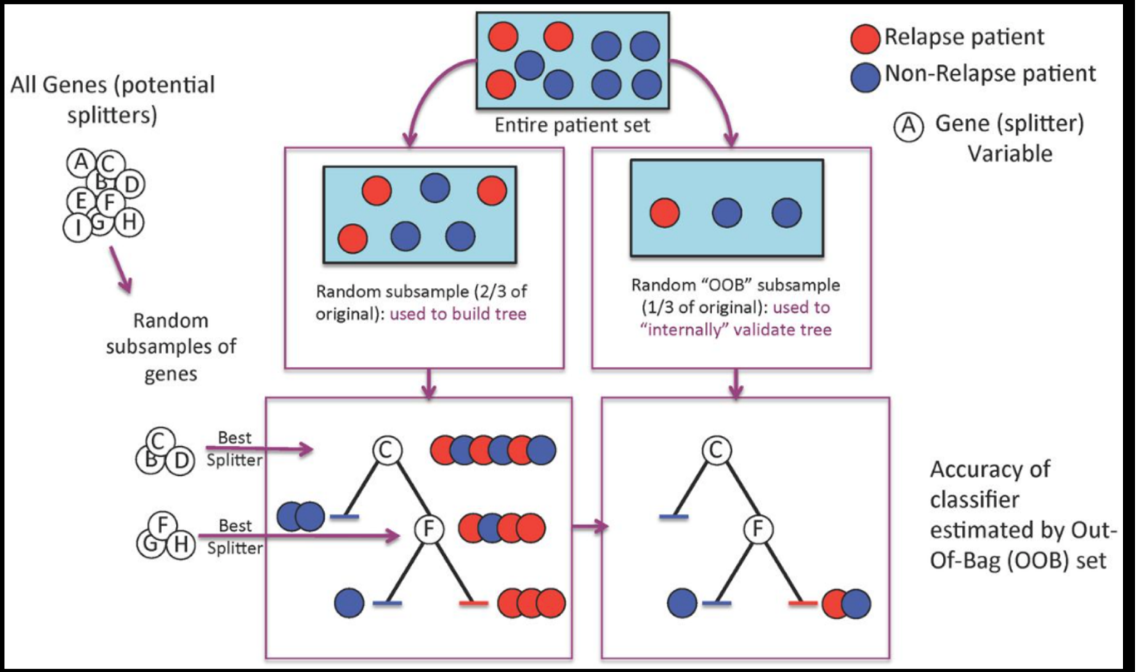


Cohen et al. 2011. PLoS ONE

# Random Forest Classifier: forest growth

- Tree growth is randomized at two levels:
  - Training – the data is randomly split into training and testing data for each tree
  - Each node (question) is trained using one variable from a random subset of available features
- Many trees are grown to produce forest
  - Number depends on the amount of features available
- Classification accuracy estimated by out-of-bag error

Ex. Determining likelihood of recurrence in breast cancer

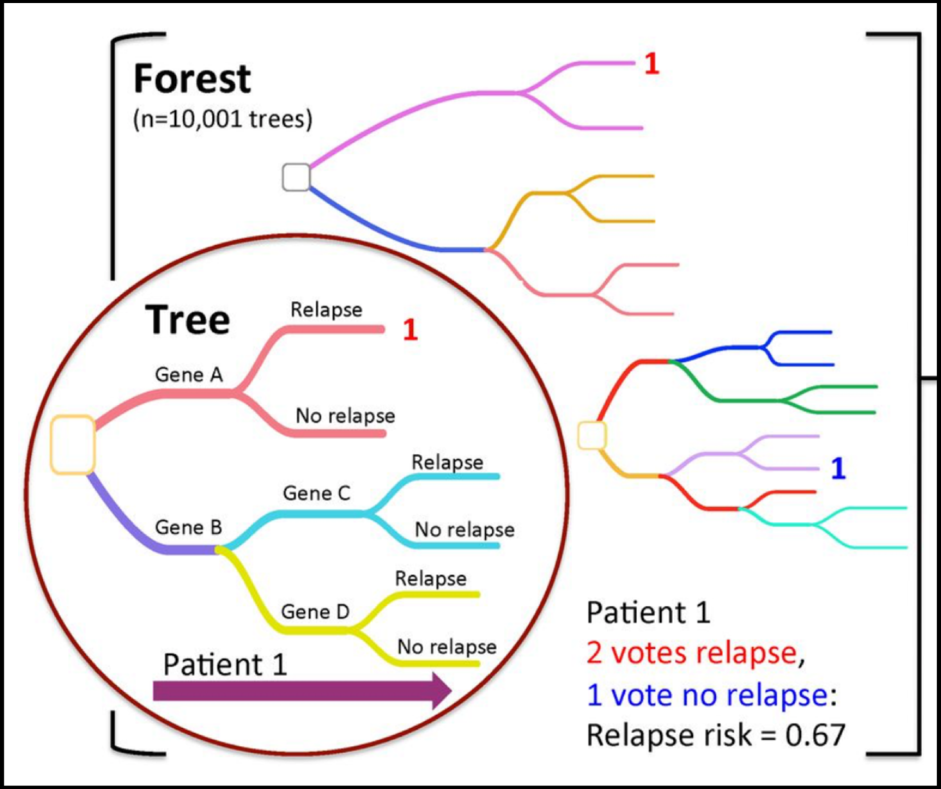

All Genes (potential splitters)
A
C
B
D
E
F
I
G
H
Random subsamples of genes
Entire patient set
● Relapse patient
● Non-Relapse patient
A Gene (splitter) Variable
Random subsample (2/3 of original): used to build tree
Random "OOB" subsample (1/3 of original): used to "internally" validate tree
C
B
D
Best Splitter
F
G
H
Best Splitter
C
F
C
F
Accuracy of classifier estimated by Out-Of-Bag (OOB) set

What are decision trees? Nat Biotechnol. 2008 Sep; 26(9): 1011–1013.
https://www.biostars.org/p/86981/
Griffith et al. (2013) Genome Medicine 5:92. A robust prognostic signature for hormone-positive node-negative breast cancer.

# Random Forest Classifiers: Ensemble function and feature importance

- Each tree gets one vote on data
- The class with the highest number of votes determines how the forest classifies the data
- Frequency of feature use as a node variable can be used to rank features

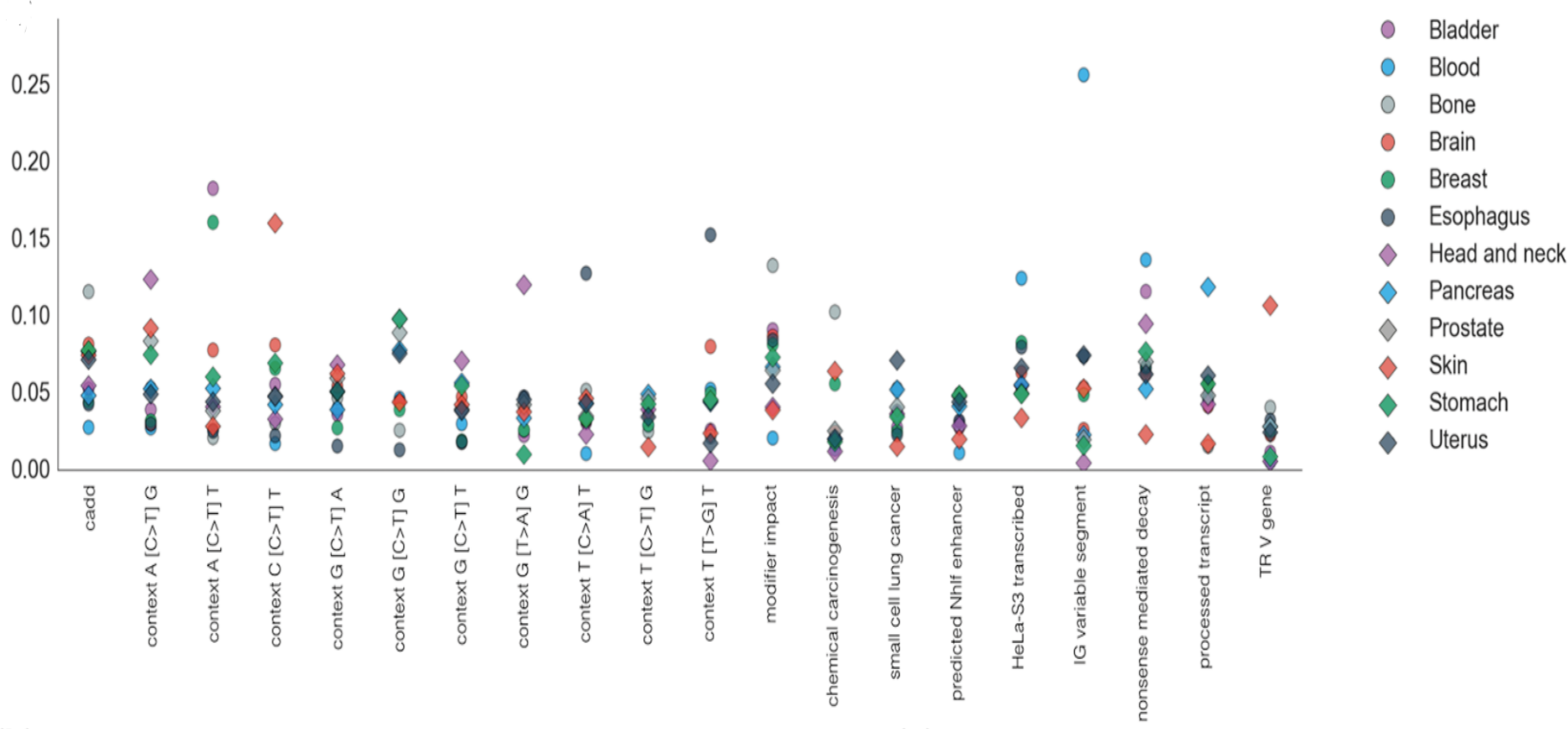
Ex. Determining likelihood of recurrence in breast cancer



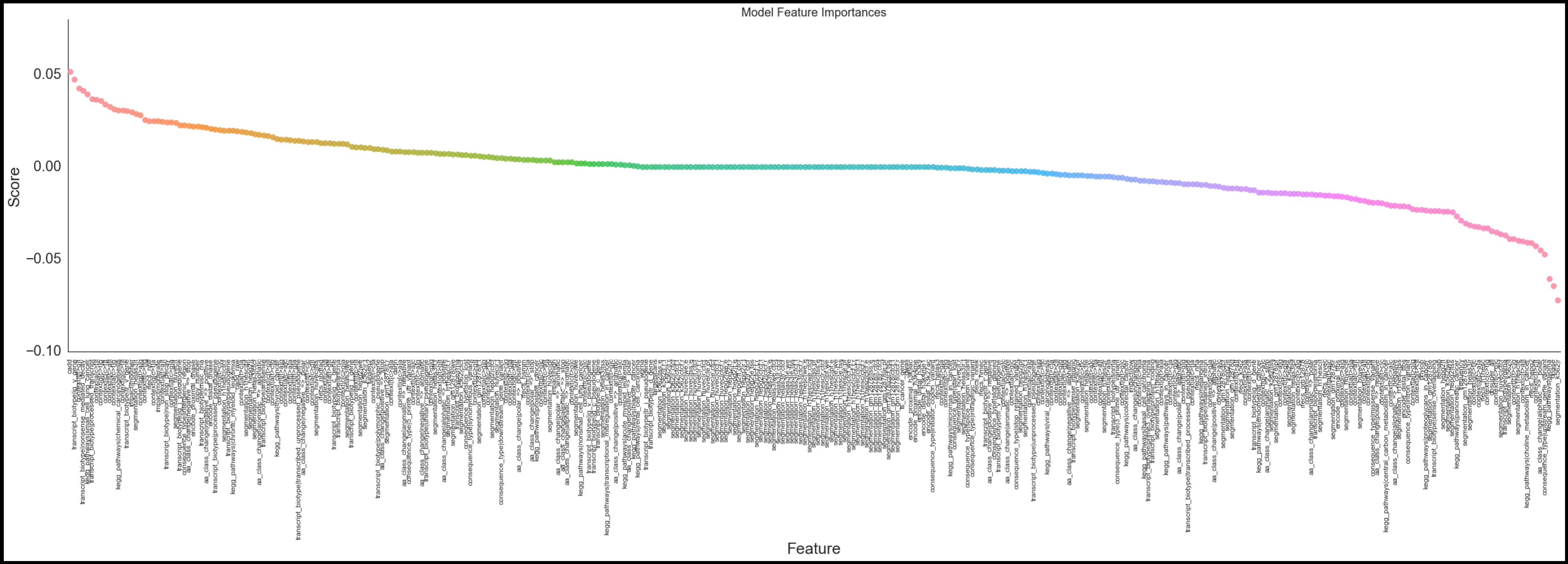
https://www.biostars.org/p/86981/
Griffith et al. (2013) Genome Medicine 5:92. A robust prognostic signature for hormone-positive node-negative breast cancer

# Displaying Model Output

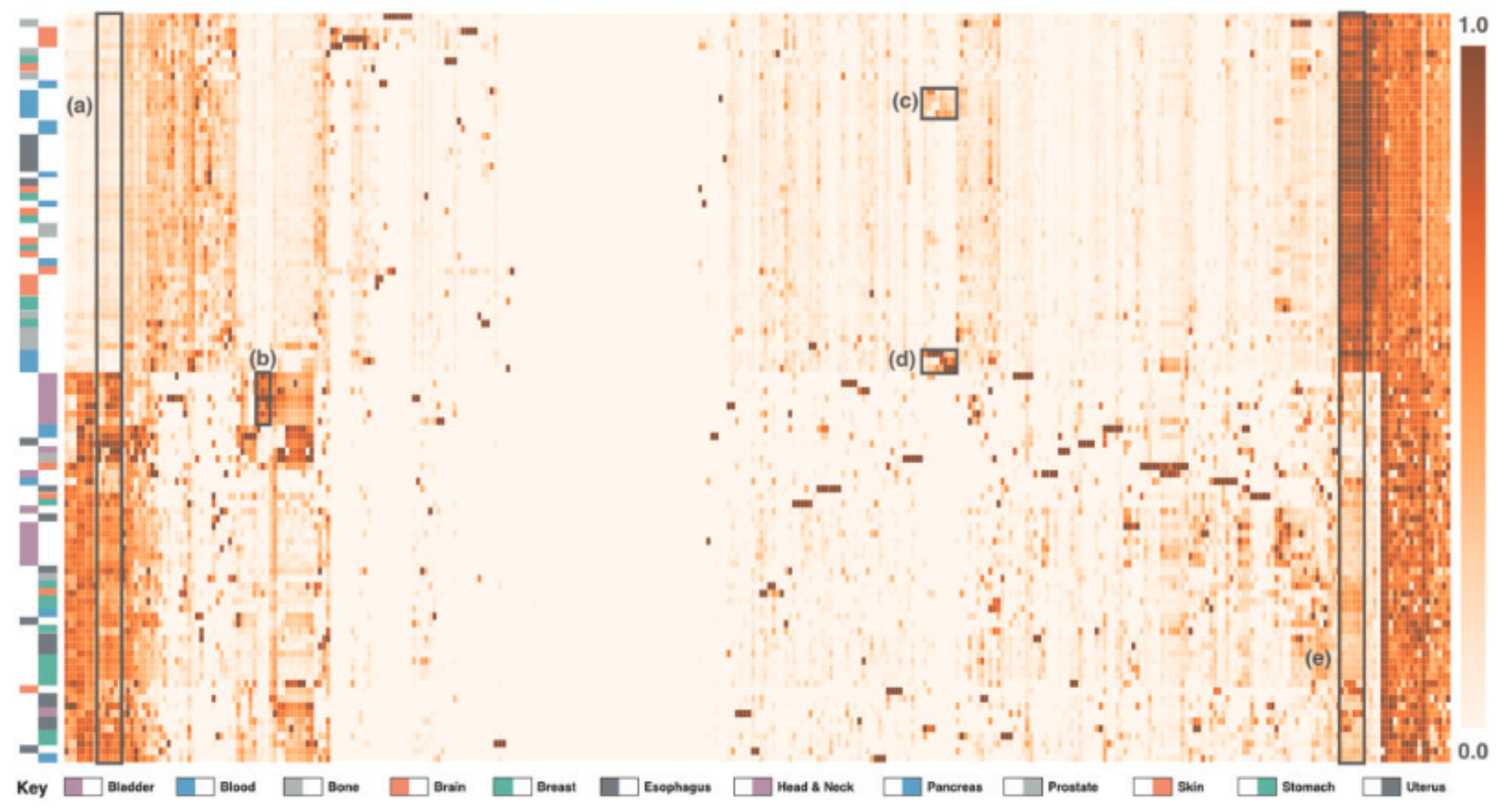
- orchid-ml supports multiple ways to display useful information
  - Feature selection graphs to show most important features to determine a cancer type
  - Feature score graph to display rankings of importance of each feature
  - Dendrogram + heat map clustering to show clusters and distances between clusters
  - Violin plots to show differences in mutation numbers between cancer types



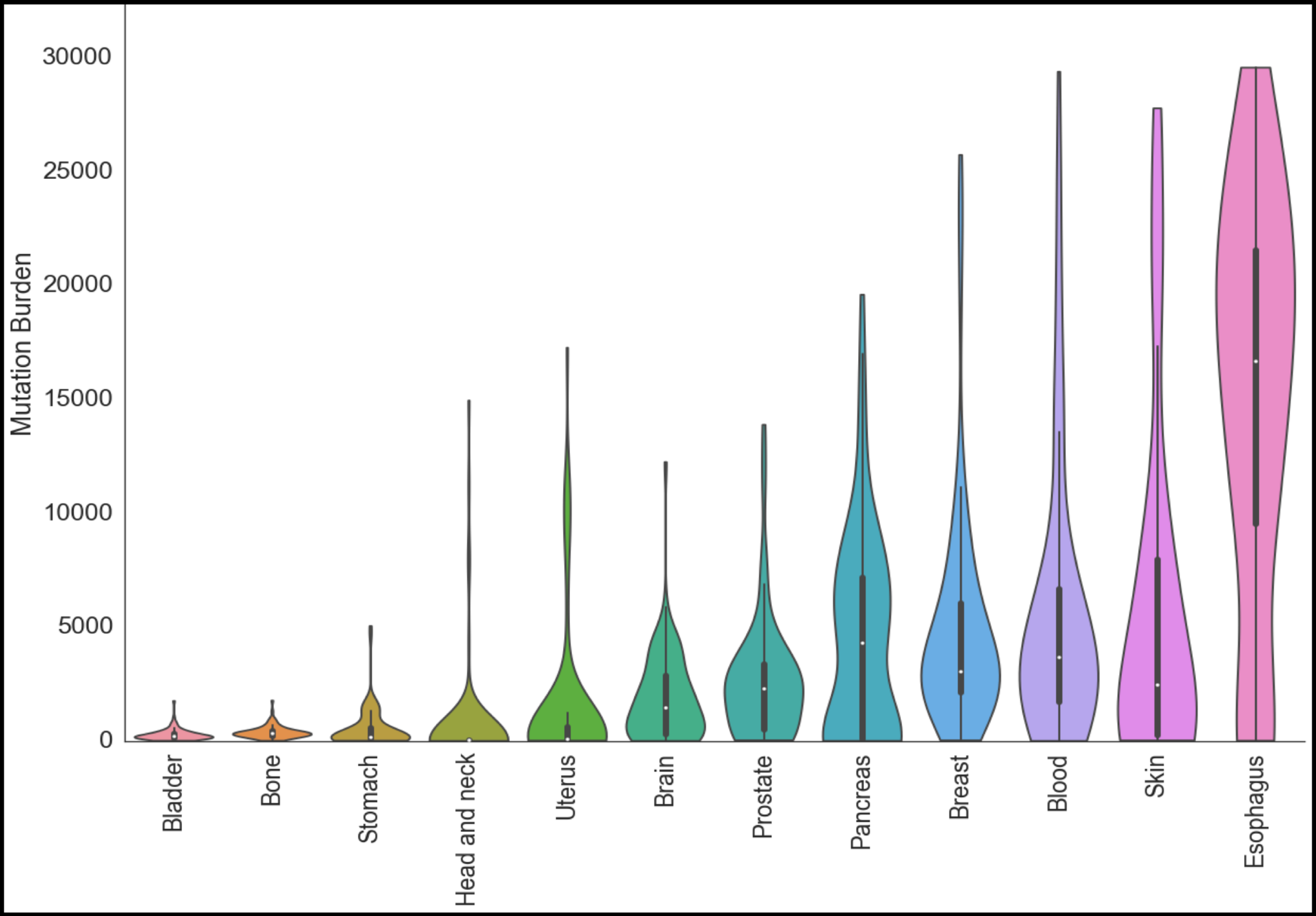
Orchid-ml's select_features() function showing the 20 most important features used to determine cancer types



Displaying Features and Their Associated Importance Score



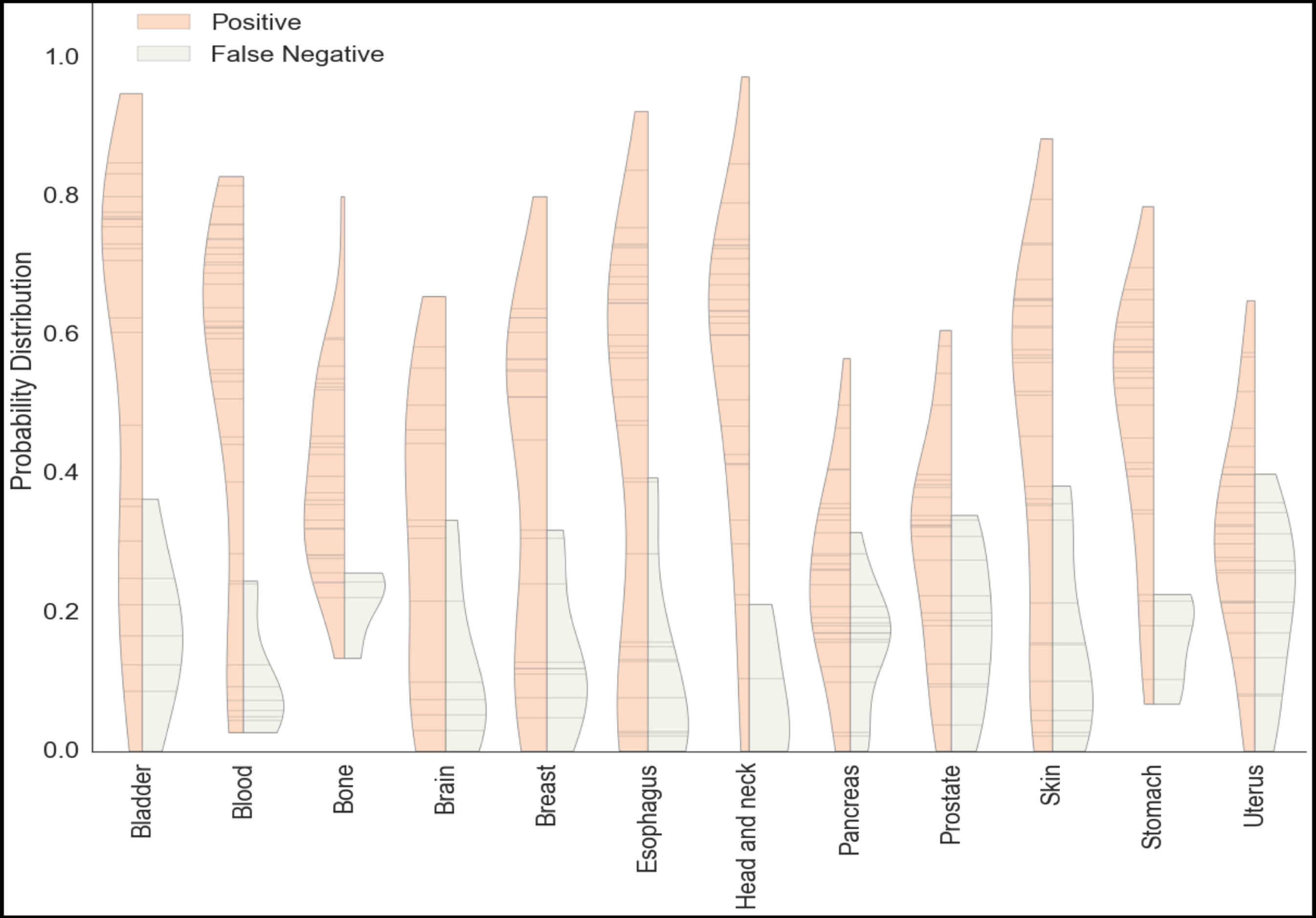
Dendrogram/Heat Map



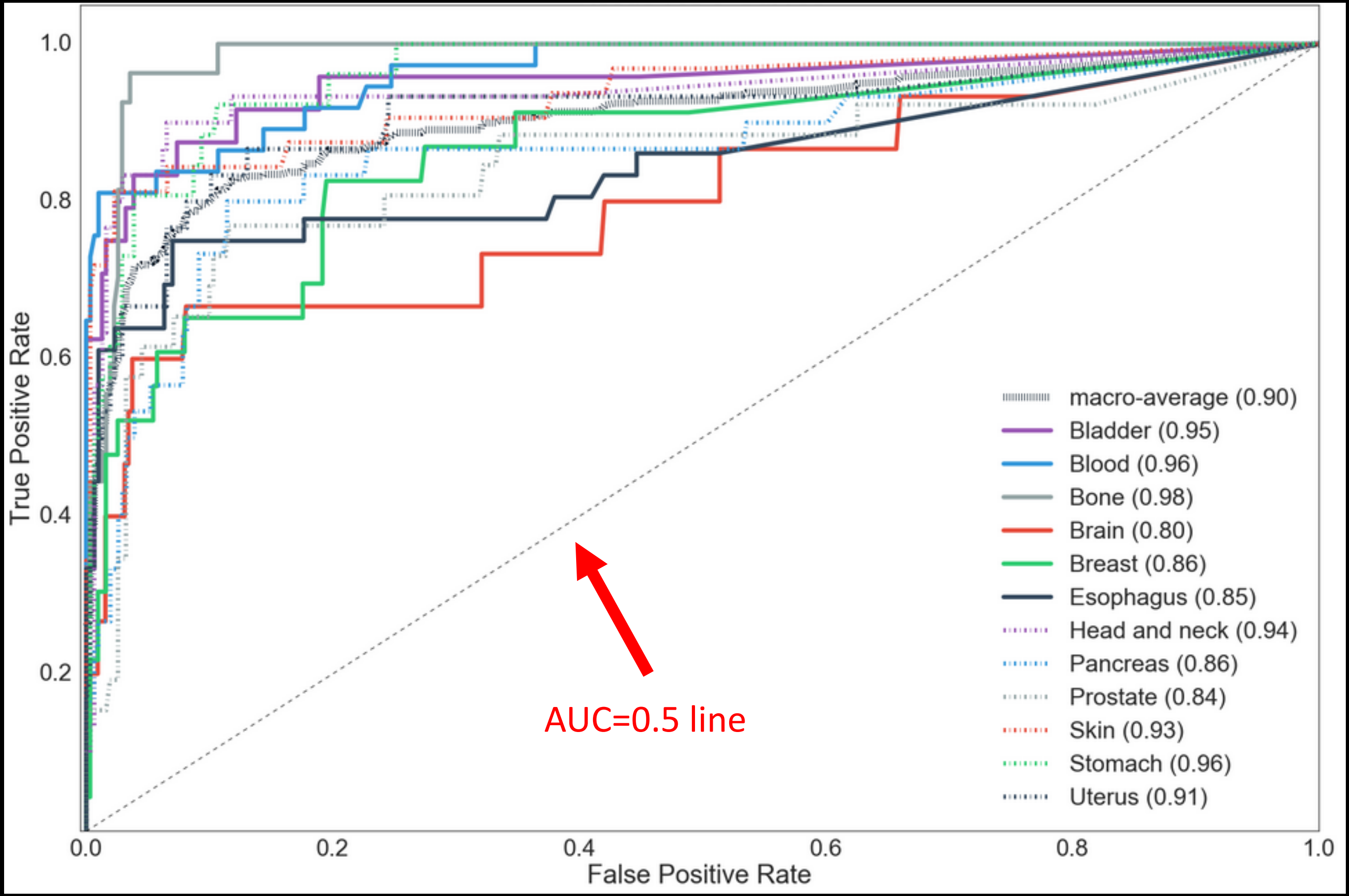
Mutation Burden by Tissue

# Assessing Model Performance

- orchid-ml supports multiple methods to assess the performance of the generated model
  - Split violin plots showing true positive vs. false positive rate
  - Receiver Operating Characteristic (ROC) curves where true positive rate is plotted as a function of false positive rate
  - Confusion matrices demonstrating predicted tissue type versus true tissue type (includes true positives, false positives, and false negatives)

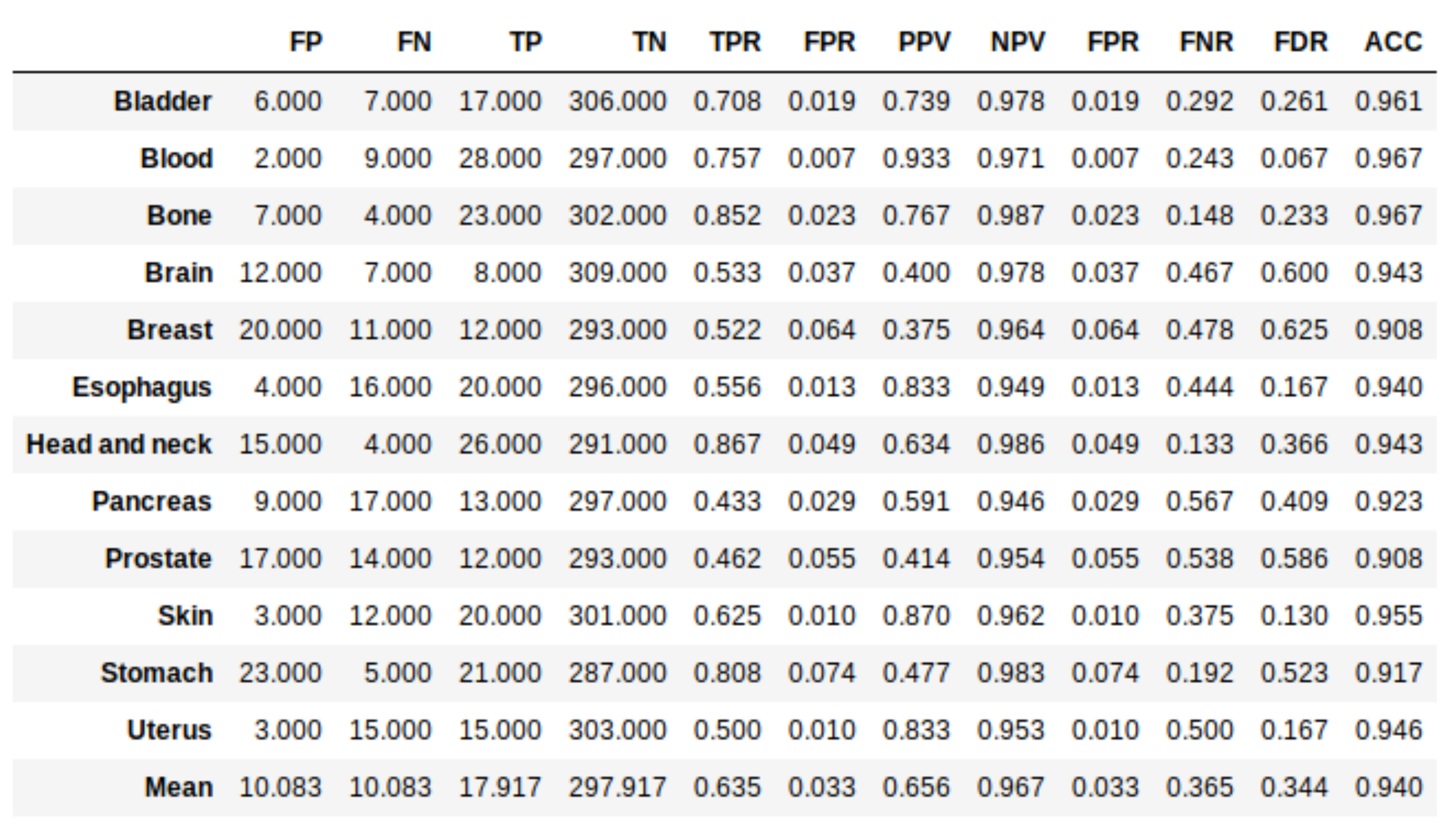


Split Violin Plot



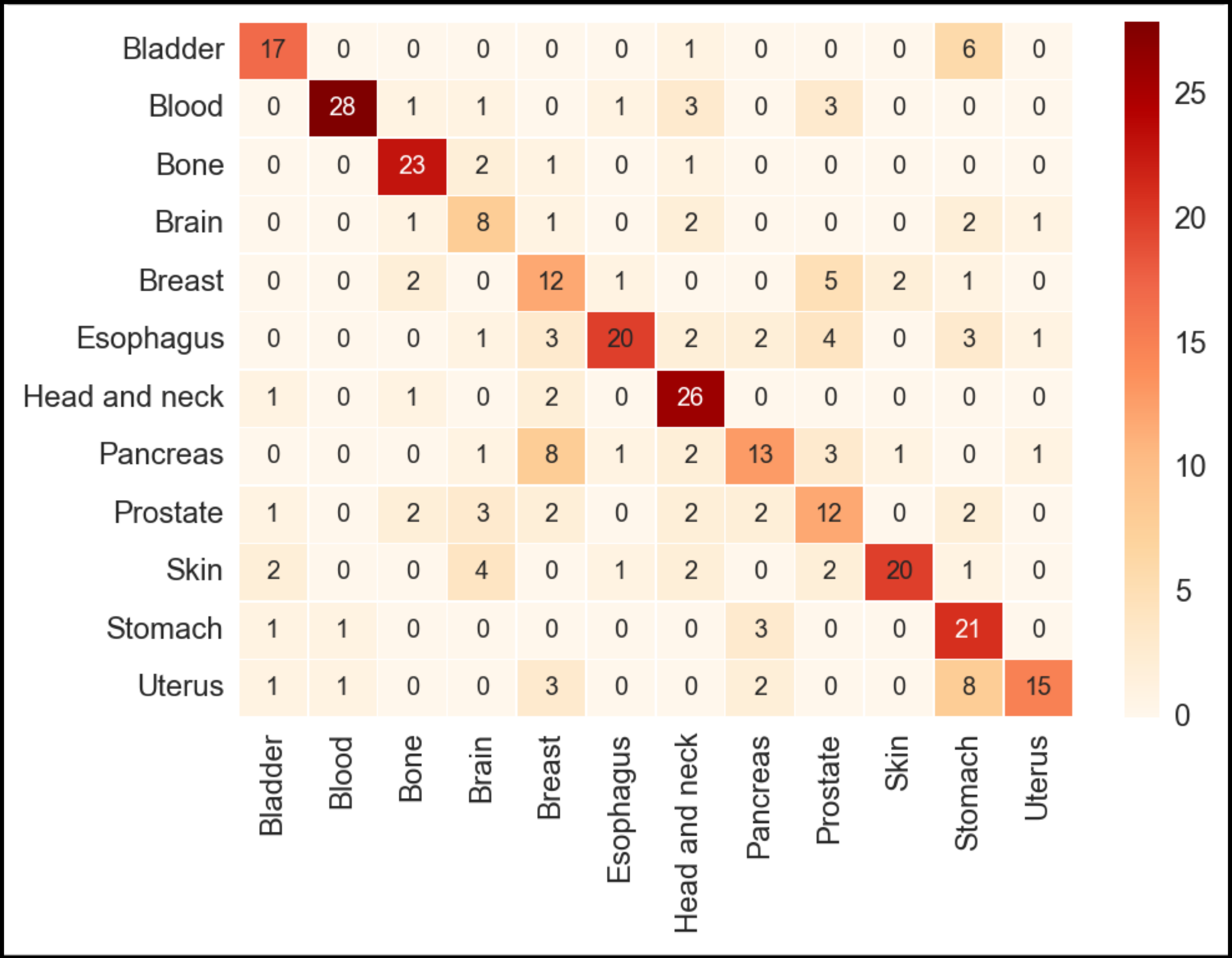
AUC=0.5 line

Sample tissue ROC curves over 5-fold cross validation +- 1 standard deviation



| | FP | FN | TP | TN | TPR | FPR | PPV | NPV | FPR | FNR | FDR | ACC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bladder** | 6.000 | 7.000 | 17.000 | 306.000 | 0.708 | 0.019 | 0.739 | 0.978 | 0.019 | 0.292 | 0.261 | 0.961 |
| **Blood** | 2.000 | 9.000 | 28.000 | 297.000 | 0.757 | 0.007 | 0.933 | 0.971 | 0.007 | 0.243 | 0.067 | 0.967 |
| **Bone** | 7.000 | 4.000 | 23.000 | 302.000 | 0.852 | 0.023 | 0.767 | 0.987 | 0.023 | 0.148 | 0.233 | 0.967 |
| **Brain** | 12.000 | 7.000 | 8.000 | 309.000 | 0.533 | 0.037 | 0.400 | 0.978 | 0.037 | 0.467 | 0.600 | 0.943 |
| **Breast** | 20.000 | 11.000 | 12.000 | 293.000 | 0.522 | 0.064 | 0.375 | 0.964 | 0.064 | 0.478 | 0.625 | 0.908 |
| **Esophagus** | 4.000 | 16.000 | 20.000 | 296.000 | 0.556 | 0.013 | 0.833 | 0.949 | 0.013 | 0.444 | 0.167 | 0.940 |
| **Head and neck** | 15.000 | 4.000 | 26.000 | 291.000 | 0.867 | 0.049 | 0.634 | 0.986 | 0.049 | 0.133 | 0.366 | 0.943 |
| **Pancreas** | 9.000 | 17.000 | 13.000 | 297.000 | 0.433 | 0.029 | 0.591 | 0.946 | 0.029 | 0.567 | 0.409 | 0.923 |
| **Prostate** | 17.000 | 14.000 | 12.000 | 293.000 | 0.462 | 0.055 | 0.414 | 0.954 | 0.055 | 0.538 | 0.586 | 0.908 |
| **Skin** | 3.000 | 12.000 | 20.000 | 301.000 | 0.625 | 0.010 | 0.870 | 0.962 | 0.010 | 0.375 | 0.130 | 0.955 |
| **Stomach** | 23.000 | 5.000 | 21.000 | 287.000 | 0.808 | 0.074 | 0.477 | 0.983 | 0.074 | 0.192 | 0.523 | 0.917 |
| **Uterus** | 3.000 | 15.000 | 15.000 | 303.000 | 0.500 | 0.010 | 0.833 | 0.953 | 0.010 | 0.500 | 0.167 | 0.946 |
| **Mean** | 10.083 | 10.083 | 17.917 | 297.917 | 0.635 | 0.033 | 0.656 | 0.967 | 0.033 | 0.365 | 0.344 | 0.940 |

Confusion Matrix Table



| | Bladder | Blood | Bone | Brain | Breast | Esophagus | Head and neck | Pancreas | Prostate | Skin | Stomach | Uterus |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bladder | 17 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 6 | 0 |
| Blood | 0 | 28 | 1 | 1 | 0 | 1 | 3 | 0 | 3 | 0 | 0 | 0 |
| Bone | 0 | 0 | 23 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Brain | 0 | 0 | 1 | 8 | 1 | 0 | 2 | 0 | 0 | 0 | 2 | 1 |
| Breast | 0 | 0 | 2 | 0 | 12 | 1 | 0 | 0 | 5 | 2 | 1 | 0 |
| Esophagus | 0 | 0 | 0 | 1 | 3 | 20 | 2 | 2 | 4 | 0 | 3 | 1 |
| Head and neck | 1 | 0 | 1 | 0 | 2 | 0 | 26 | 0 | 0 | 0 | 0 | 0 |
| Pancreas | 0 | 0 | 0 | 1 | 8 | 1 | 2 | 13 | 3 | 1 | 0 | 1 |
| Prostate | 1 | 0 | 2 | 3 | 2 | 0 | 2 | 2 | 12 | 0 | 2 | 0 |
| Skin | 2 | 0 | 0 | 4 | 0 | 1 | 2 | 0 | 2 | 20 | 1 | 0 |
| Stomach | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 21 | 0 |
| Uterus | 1 | 1 | 0 | 0 | 3 | 0 | 0 | 2 | 0 | 0 | 8 | 15 |

Confusion Matrix with False Positives as Rows and False Negatives as Columns

# Software Impressions

Easy to install

Some quirks: requires python 2, Java 8/9 (latest is 11)

Convenient but not novel

Largely a framework to combine multiple other pieces of software more than does something new

Relies significantly on scikit (orchid-ml is largely just SQL commands with scikit algorithms)

Can be difficult to find files in correct format

Relies on conversion tool or requesting access for vcf files

Use of nextflow for parallel processing is useful

This can speed up processing of large datasets by orchid-db

Name choice could be better

Orchid is an incredibly common term

# Discussion/conclusion

- Orchid achieves design goals - use case in article
  - Does handle hundreds of features and millions of features
  - Did not demonstrate usefulness across multiple data bases
- It is limited to genomic evaluation
  - Epigenetic, RNA-seq, transfusions etc are unavailable
- Demonstrates one use case
  - No known use cases outside of this paper (only cited in reviews)
- Long-term - authors are looking at precision medicine
  - Guessing cfDNA for cancer detection/progression, mutation determination for targeted therapies, and liquid biopsy treatment monitoring
  - https://avail.bio/about/