Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute



Preliminary Study of Classification of Pediatric Cancers Using Gene Expression Data from the TARGET Dataset

Andrew Weisman
andrew.weisman@nih.gov

August 11, 2020

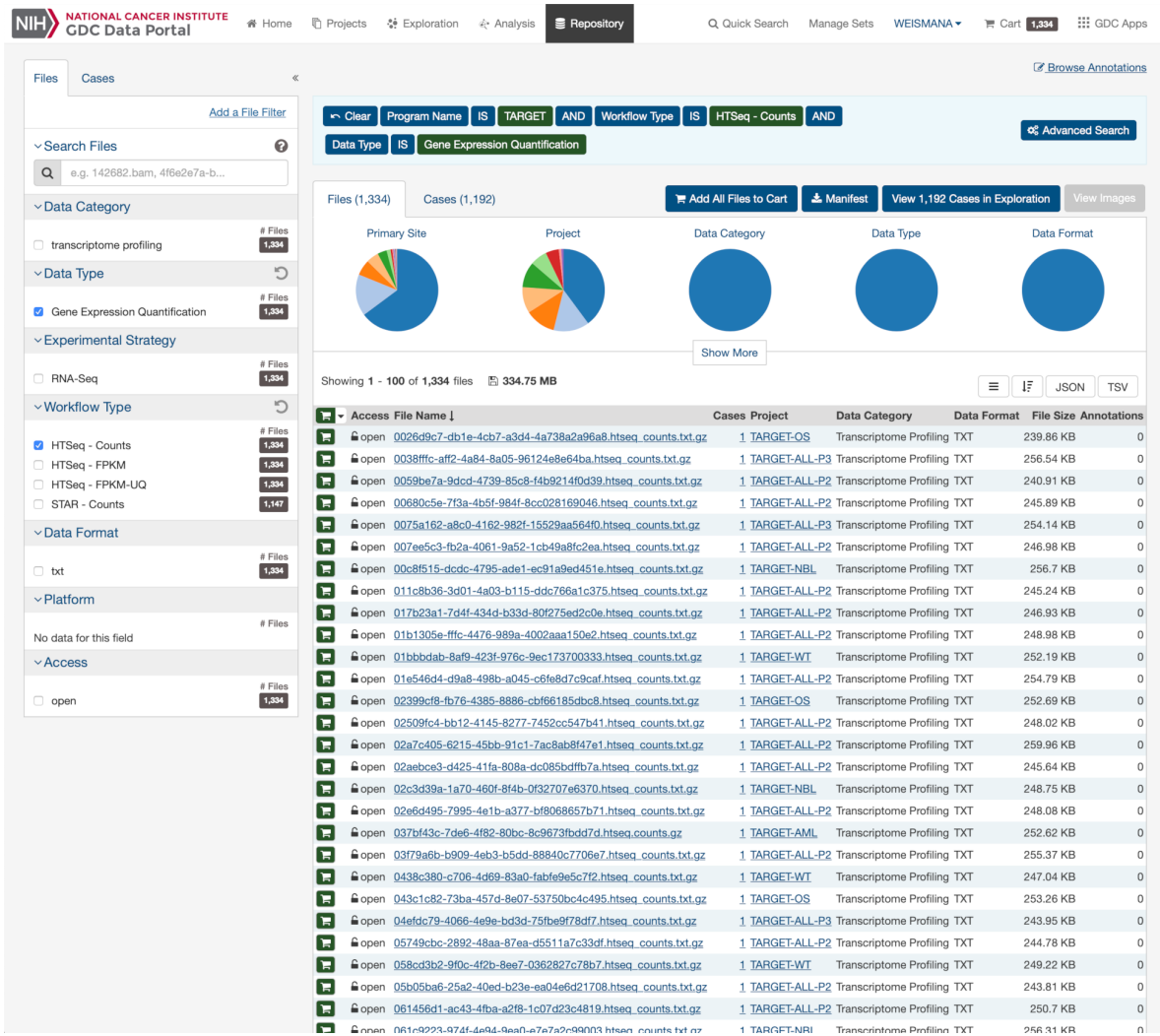
DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute
Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Overall project progression

- Initially tried using data from the TARGET project's data website https://target-data.nci.nih.gov
- Potentially helpful because there appears to be much more gene expression data there (~2x) than at the GDC Data Portal https://portal.gdc.cancer.gov
- However, we found great difficulties, especially in unifying all the gene names, IDs, accession numbers, etc., no matter the method
- Moved for the time being to using the data from the Data Portal, which is harmonized by the GDC
- Have since run initial studies using these data, described herein

Data download

- **Obtain all HTSeq-generated counts expression files in the TARGET program on Helix:**
  - Download the Manifest (blue button)
  - Use the GDC client tool (`module load gdc-client`)
- **Add all files to the cart, go to the cart, and download:**
  - Sample Sheet (blue button)
  - Metadata (blue button)
- **This yields data for 1,334 counts files and 1,192 cases (i.e., people)**

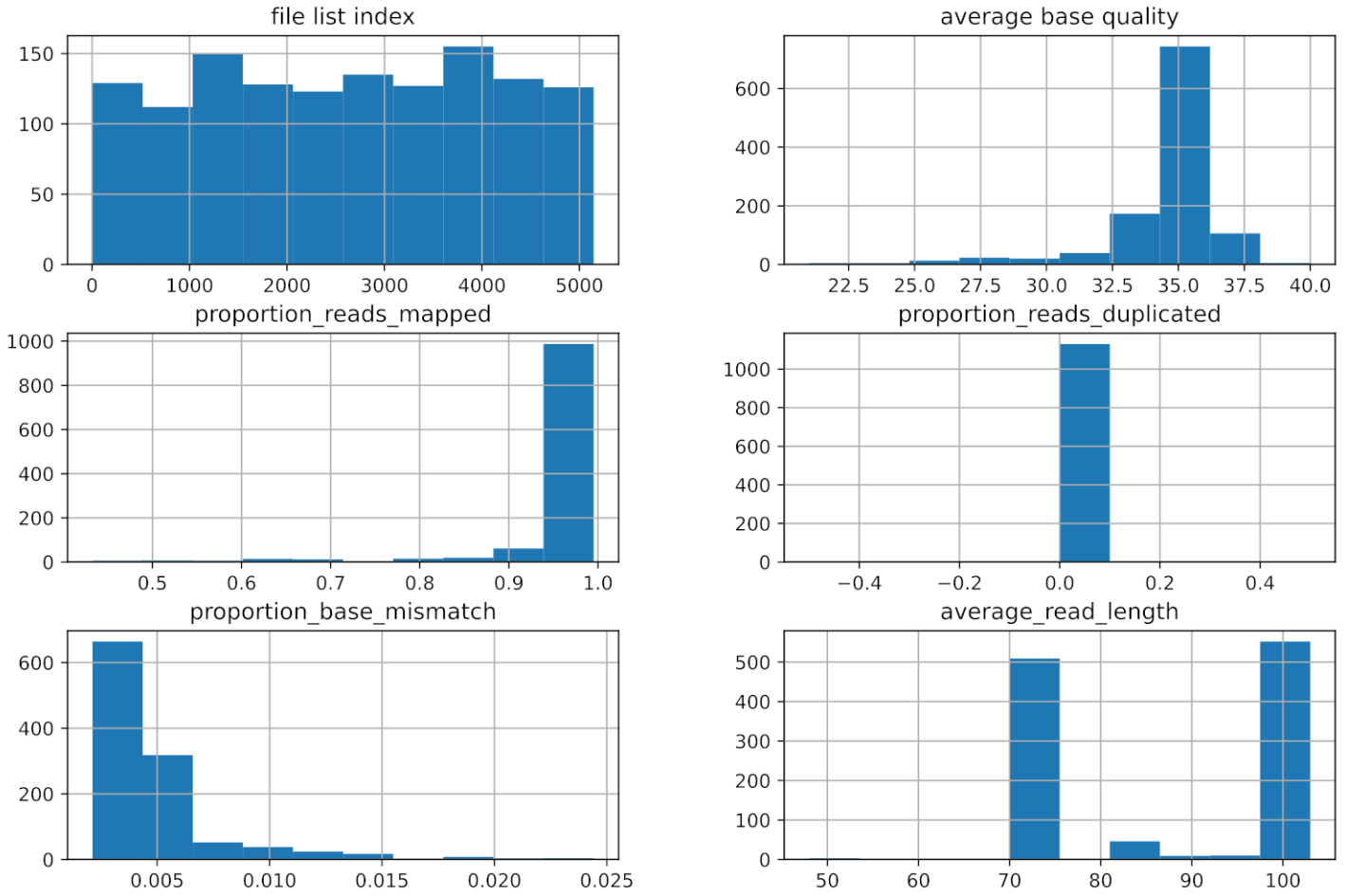


Data preprocessing

- **Read in and unify the counts files, sample sheet data, and metadata**
- **Found 1,321 unique samples in the sample sheet (instead of all 1,334)**
  - In 13 cases there are two HTSeq counts files for a single sample, i.e., multiple analyses of the same sample
  - In these cases, choose that with the best “average_base_quality” score (or choose the first in case of a tie)
- **Drop four samples that correspond to multiple cases (people), which doesn’t make sense → 1,317 samples remaining**
  - GDC said this was a bug and would be fixed in the future

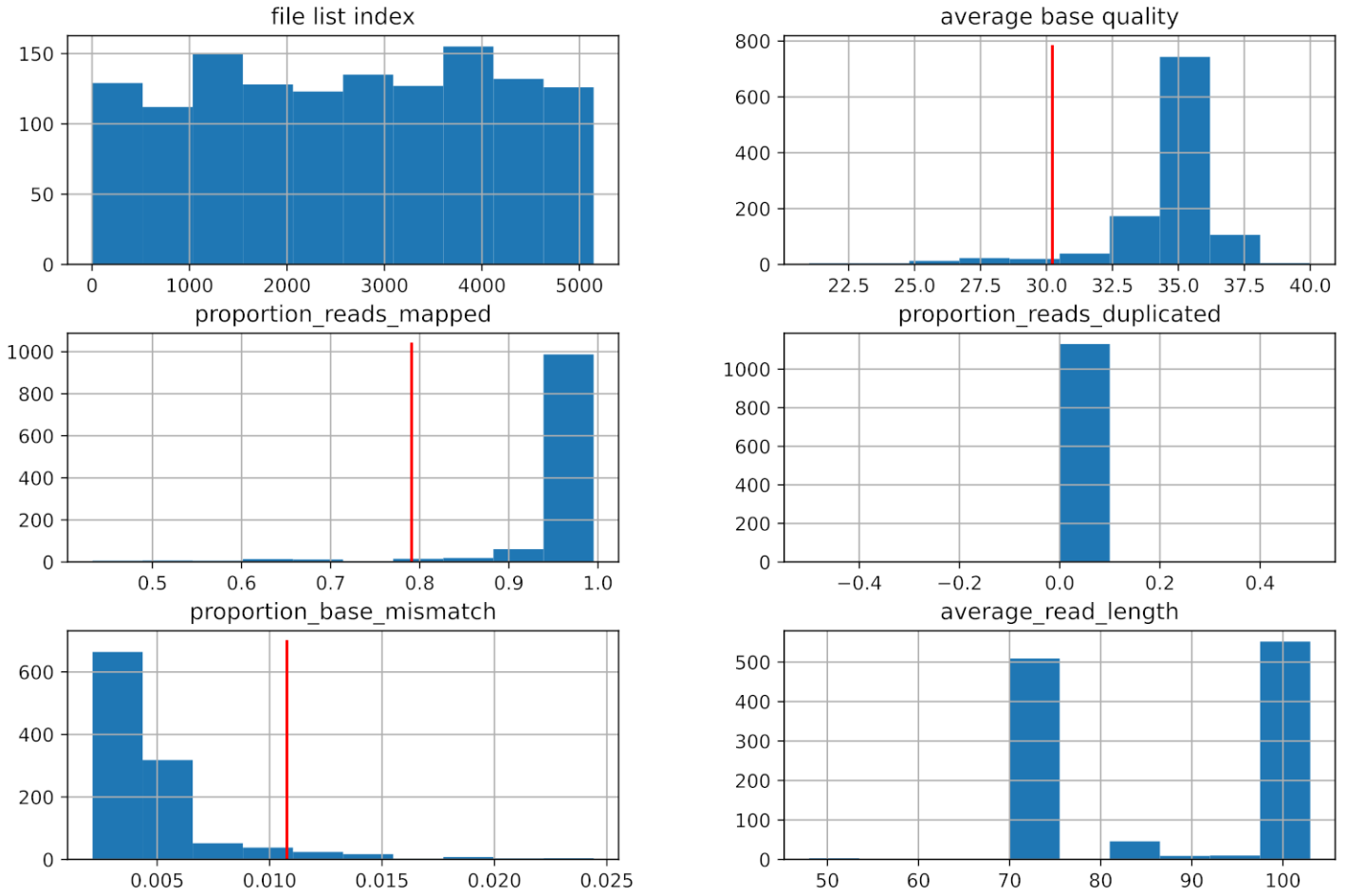
Numeric fields for each sample

- **Original:**



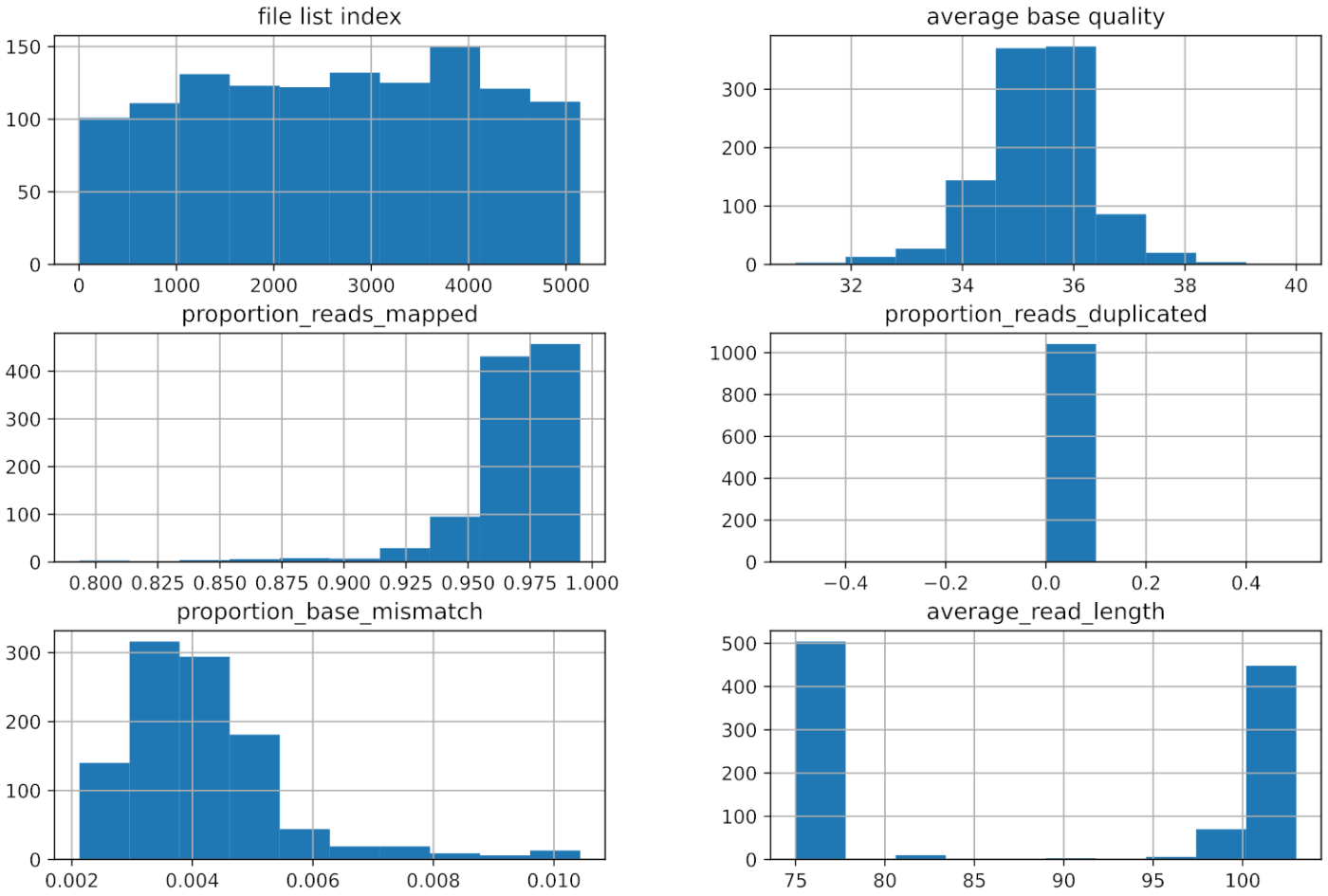
Numeric fields for each sample, ctd.

- **With cutoffs:**



Numeric fields for each sample, ctd.

- **After cutoffs applied, eliminating 89 samples → 1,228 remaining:**



Non-numeric fields for each sample

- **Columns with all unique values (1,228 of them), with example values:**
  - **counts file name:** de2b4915-95fc-45d9-b327-adde18559211.htseq_counts.txt.gz
  - **file id:** 677342f7-1f9a-4615-925c-746d2d460b49
  - **entity_submitter_id:** TARGET-10-PARTBP-09A-01R
  - **sample id:** TARGET-10-PARTBP-09A
- **Columns with uniform values:**
  - **contamination_error:** None
  - **contamination:** None
  - **state:** released
  - **platform:** Illumina

Non-numeric fields for each sample, ctd.

- **Columns with non-unique and non-uniform values:**
  - **case id** (1,099 unique values):
    - `TARGET-15-SJMPAL042946  3`
    - `TARGET-50-PAJNRL  2`
    - `TARGET-10-PARAKF  2`
    - `TARGET-15-SJMPAL041119  2`
    - `TARGET-10-PARPRW  2`
    - `...  ...`
    - `TARGET-50-PAJMKN  1`
    - `TARGET-15-SJMPAL012421  1`
    - `TARGET-10-PARNSP  1`
    - `TARGET-50-PAJLLF  1`
    - `TARGET-10-PANWEZ  1`

Non-numeric fields for each sample, ctd.

- **Columns with non-unique and non-uniform values, ctd.:**
  - **project id** (9 unique values):

TARGET-ALL-P2	**518**	Acute Lymphoblastic Leukemia – Phase II
TARGET-AML	**187**	Acute Myeloid Leukemia
TARGET-NBL	**143**	Neuroblastoma
TARGET-WT	**136**	Wilms Tumor
TARGET-ALL-P3	**135**	Acute Lymphoblastic Leukemia – Phase III (ALAL)
TARGET-RT	**69**	Rhabdoid Tumor
TARGET-OS	**24**	Osteosarcoma
TARGET-CCSK	**13**	Clear Cell Sarcoma of the Kidney
TARGET-ALL-P1	**3**	Acute Lymphoblastic Leukemia – Pilot

Non-numeric fields for each sample, ctd.

- **Columns with non-unique and non-uniform values, ctd.:**
  - **sample type** (8 unique values):

Primary Blood Derived Cancer - Bone Marrow	614
Primary Tumor	359
Recurrent Blood Derived Cancer - Bone Marrow	117
Primary Blood Derived Cancer - Peripheral Blood	109
Recurrent Tumor	12
Solid Tissue Normal	12
Recurrent Blood Derived Cancer - Peripheral Blood	4
Metastatic	1

Combine “project id” and “sample type” columns → most detailed set of labels

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

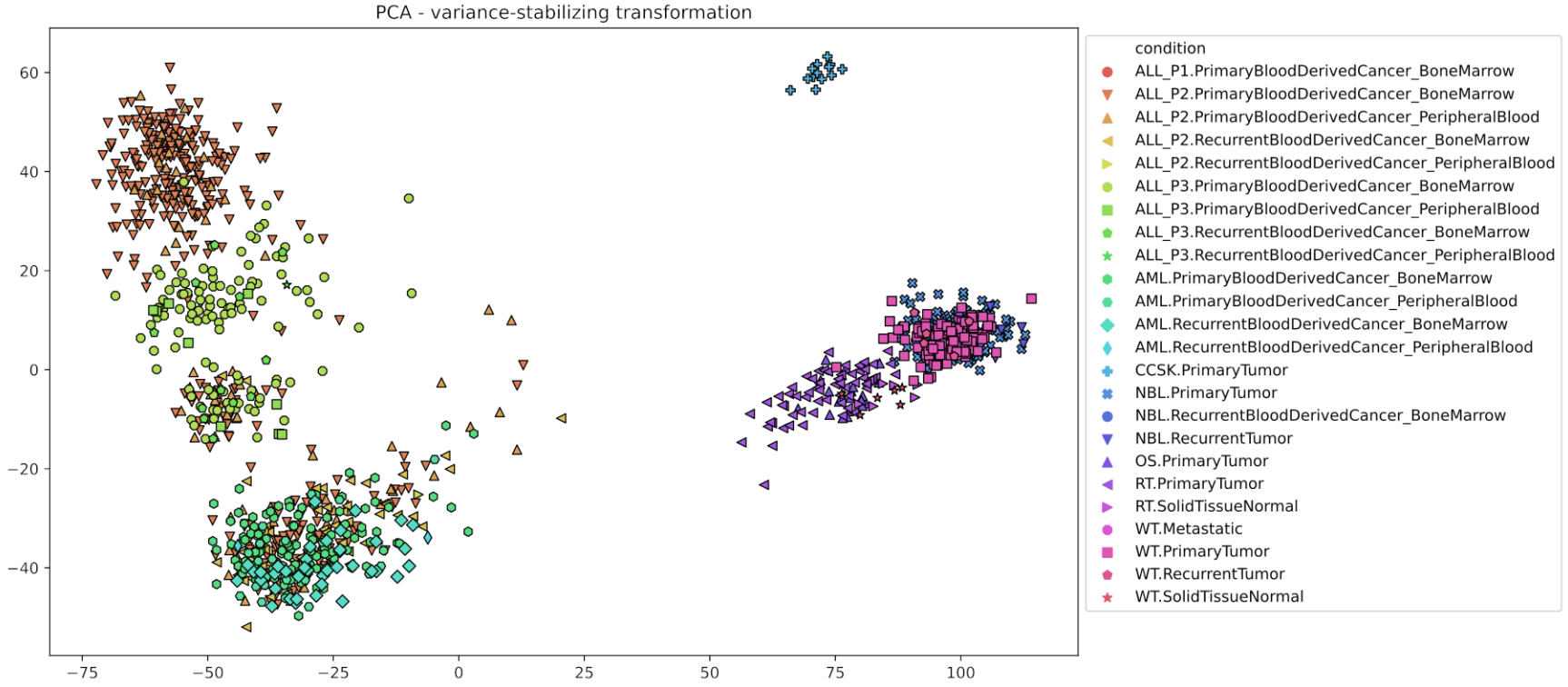
- TARGET-ALL-P2, Primary Blood Derived Cancer - Bone Marrow 379
- TARGET-NBL, Primary Tumor 135
- TARGET-WT, Primary Tumor 124
- TARGET-AML, Primary Blood Derived Cancer - Bone Marrow 119
- TARGET-ALL-P3, Primary Blood Derived Cancer - Bone Marrow 113
- TARGET-ALL-P2, Primary Blood Derived Cancer - Peripheral Blood 74
- TARGET-ALL-P2, Recurrent Blood Derived Cancer - Bone Marrow 64
- TARGET-RT, Primary Tumor 63
- TARGET-AML, Recurrent Blood Derived Cancer - Bone Marrow 40
- TARGET-AML, Primary Blood Derived Cancer - Peripheral Blood 26
- TARGET-OS, Primary Tumor 24
- TARGET-CCSK, Primary Tumor 13
- TARGET-ALL-P3, Recurrent Blood Derived Cancer - Bone Marrow 12
- TARGET-ALL-P3, Primary Blood Derived Cancer - Peripheral Blood 9
- TARGET-NBL, Recurrent Tumor 7
- TARGET-WT, Solid Tissue Normal 6
- TARGET-RT, Solid Tissue Normal 6
- TARGET-WT, Recurrent Tumor 5
- TARGET-ALL-P1, Primary Blood Derived Cancer - Bone Marrow 3
- TARGET-AML, Recurrent Blood Derived Cancer - Peripheral Blood 2
- TARGET-NBL, Recurrent Blood Derived Cancer - Bone Marrow 1
- TARGET-ALL-P3, Recurrent Blood Derived Cancer - Peripheral Blood 1
- TARGET-WT, Metastatic 1
- TARGET-ALL-P2, Recurrent Blood Derived Cancer - Peripheral Blood 1

Counts normalization

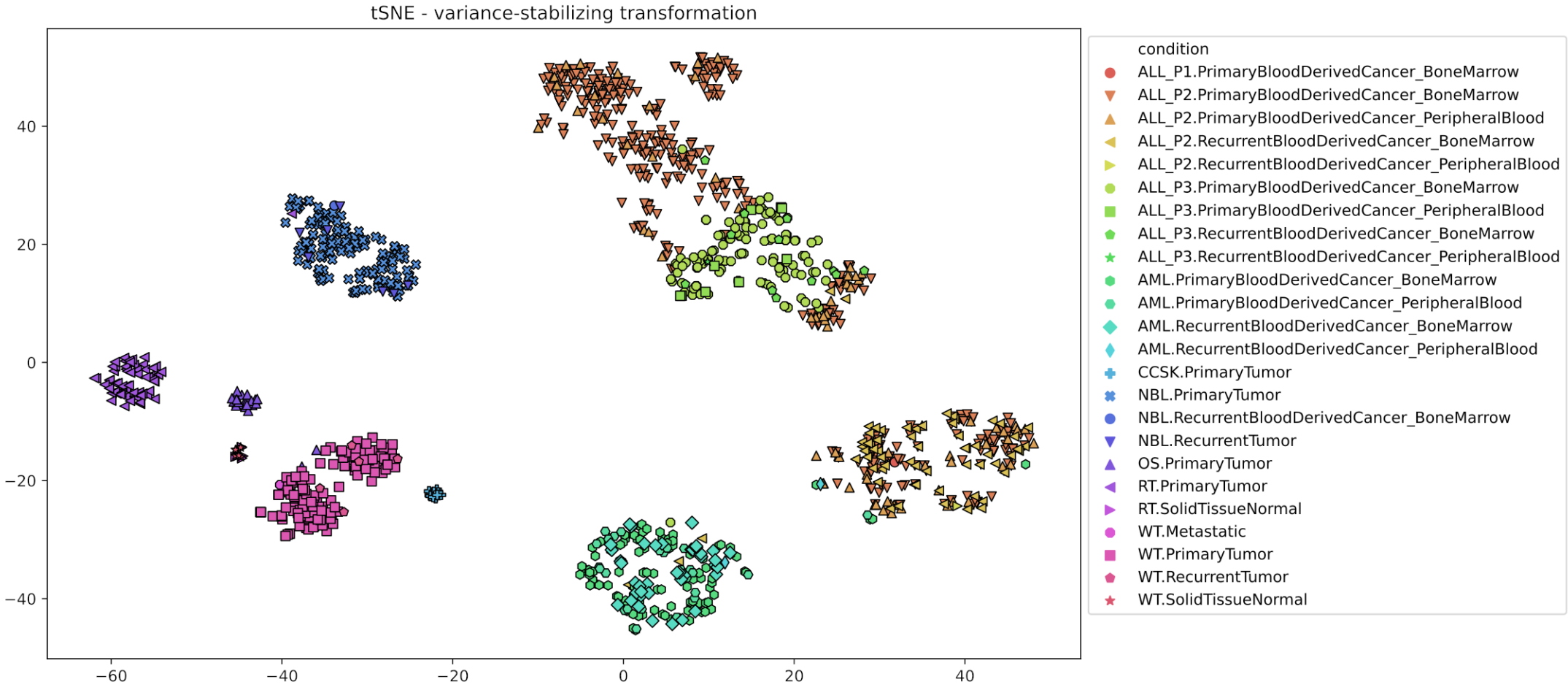
- **Most appropriate method is to perform a statistical analysis on the raw counts *C***
  - Sample packages: DESeq2, EdgeR, Limma Voom
  - We're choosing DESeq2 → popular and well-documented
- **Three main types of transformations in DESeq2 (they all transform the counts to the log-base-2 scale):**
  - **Normal transformation:** $\log_2(C^* + 1)$, where C^* represents the counts C divided by size factors (i.e., the normalized counts)
  - **Regularized log transformation:** makes count variances independent of their means using fitting to a GLM for each gene, taking library size into account
  - **Variance-stabilizing transformation (VST):** makes normalized count variances independent of their means using a monotonous mapping function based on the variance-mean dependence in a fitted negative binomial distribution, taking library size into account
    - We use this, as it is much faster and requires less memory than regularized log

Principal components analysis on the detailed dataset

- **Run VST on the counts and the detailed set of labels y_1, resulting in the data matrix X_1 consisting of VST-normalized counts**
- **Perform principal components analysis on X_1 (using y_1 as labels, and top-500-variance genes)**



tSNE on the detailed dataset



tSNE on the detailed dataset – observations

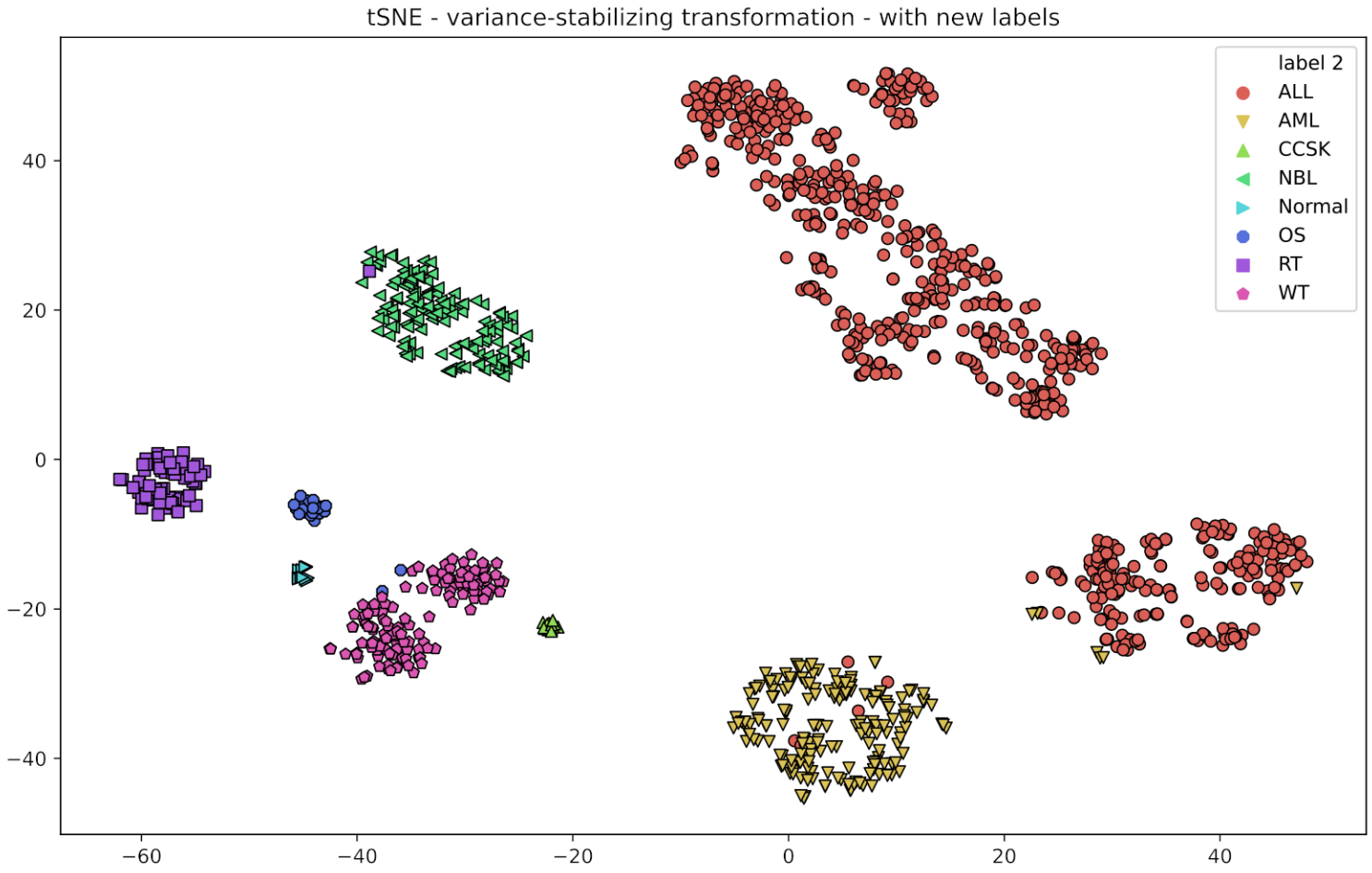
- ALL is spread out into four main clusters; in particular:
  - ALL/P2/PBDC (both /BM and /PB) is located in three of them
  - ALL/P2/RBDC (i.e., /BM) is basically isolated to one of them
  - ALL/P3 is basically in the fourth cluster
- AML has its own single cluster
- There are some ALL in an AML cluster and vice versa, and the two corresponding clusters are near each other
  - Note that even though ALL/P3 is supposed to be the "ambiguously myeloid" one (ALAL, acute leukemia of ambiguous lineage), it is primarily ALL/P2 that seems to be confused with AML
- CCSK is very tightly clustered
- NBL has its own cluster, with one RT instance that may be misclassified
- OS is tightly clustered, except for two samples in WT, which may indicate that they are misclassified
- Aside from the one possibly misclassified RT, they are all solidly clustered
- WT is clustered together
- The two normal species are very tightly clustered together
- Note the three kidney tumor types (RT, WT, CCSK), and OS and Normal, are relatively close together

Create a “less detailed” set of labels

- Combine the two normal species
- Combine all ALL projects into one
- Otherwise, use the project IDs as the labels

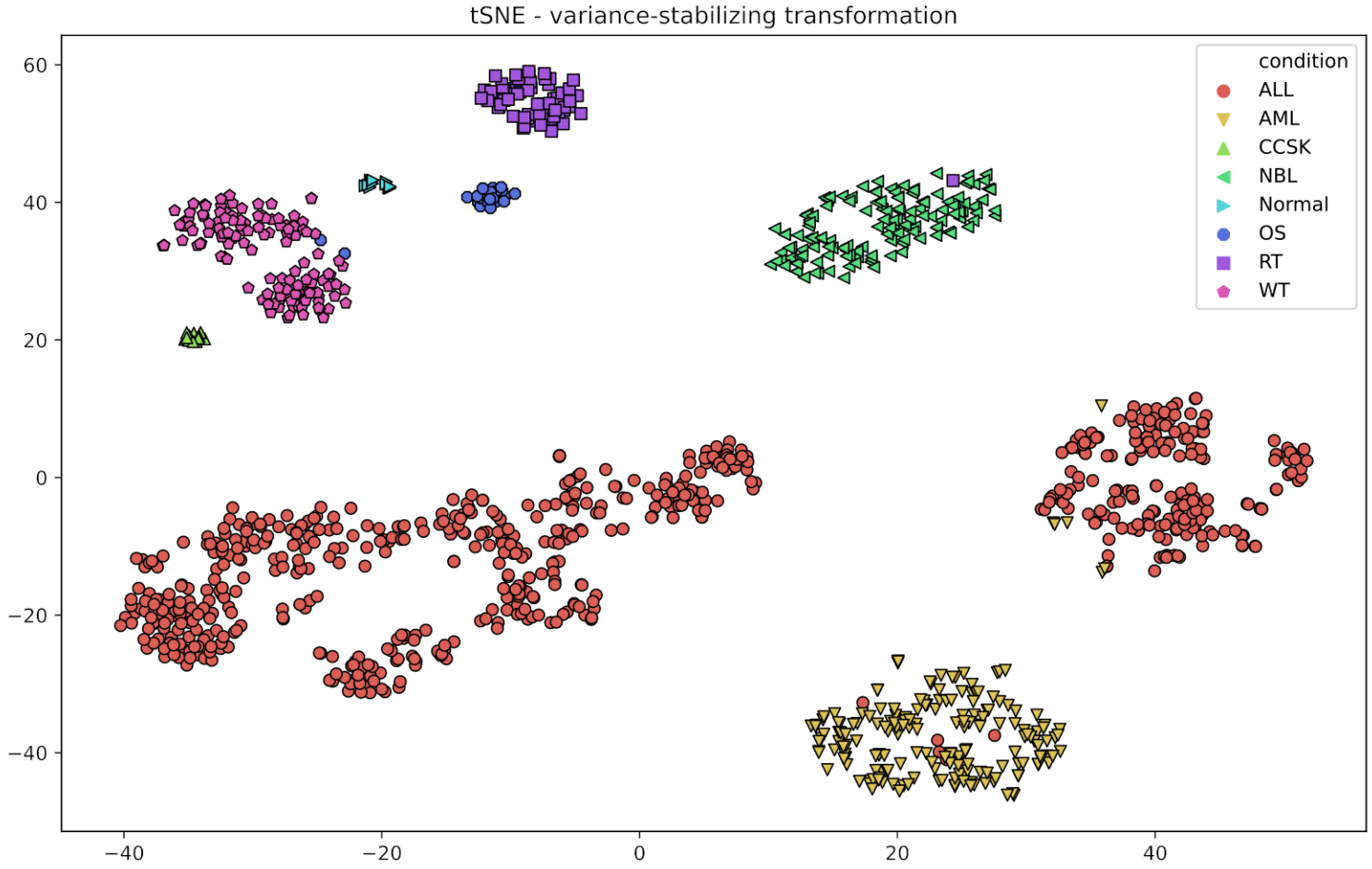
- New set of classes:
  - `ALL  656`
  - `AML  187`
  - `NBL  143`
  - `WT  130`
  - `RT  63`
  - `OS  24`
  - `CCSK  13`
  - `Normal  12`

Use the new set of labels on the previous tSNE plot



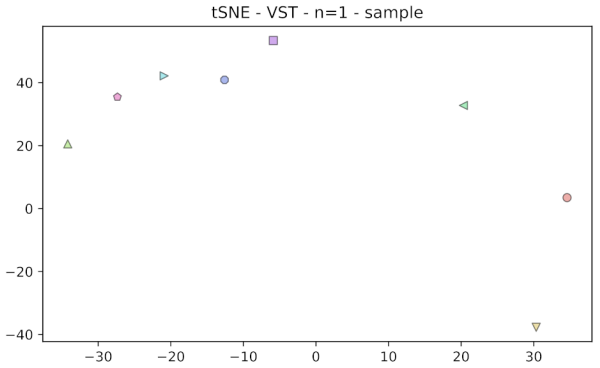
tSNE on the less-detailed dataset

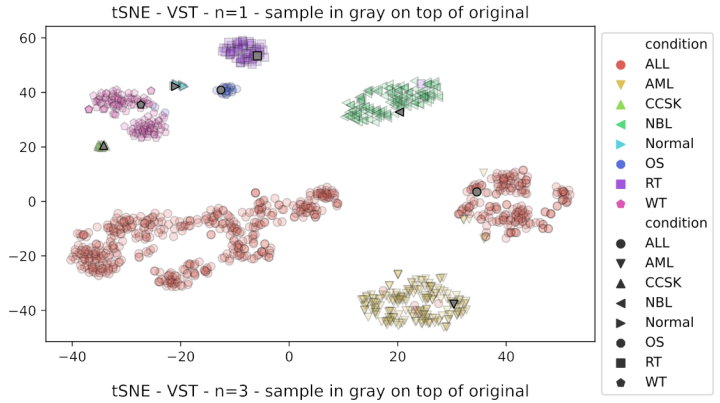
- **Must re-run VST on the raw counts using the new, less-detailed set of labels y_2, resulting in X_2**



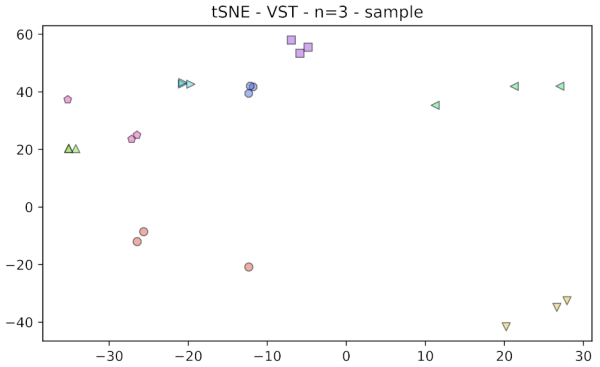
Sample with replacement from each class

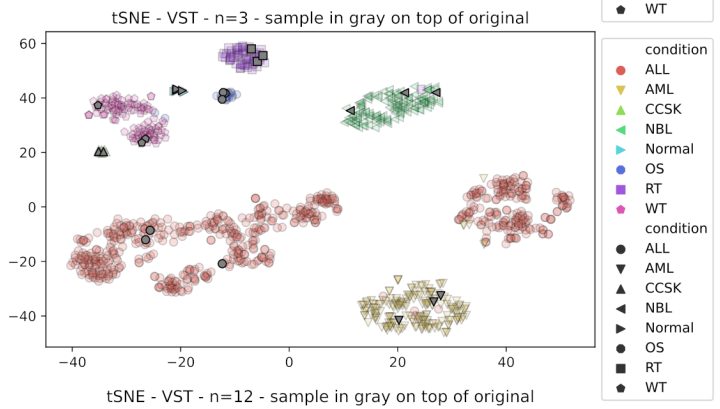
N=1



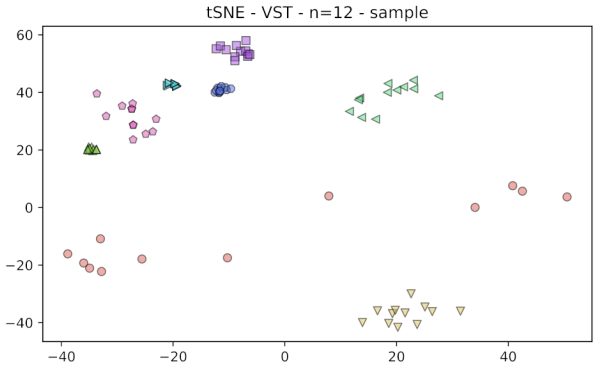


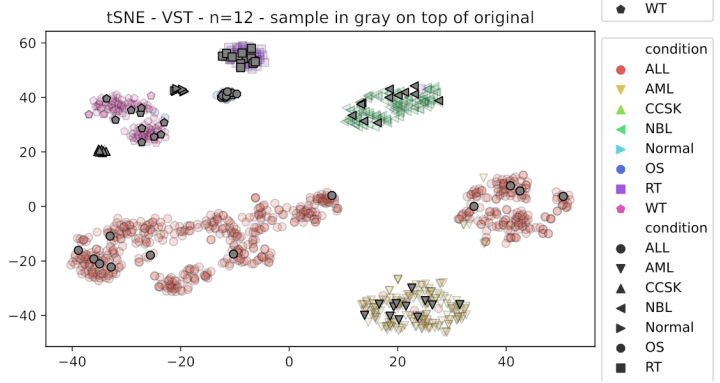
N=3





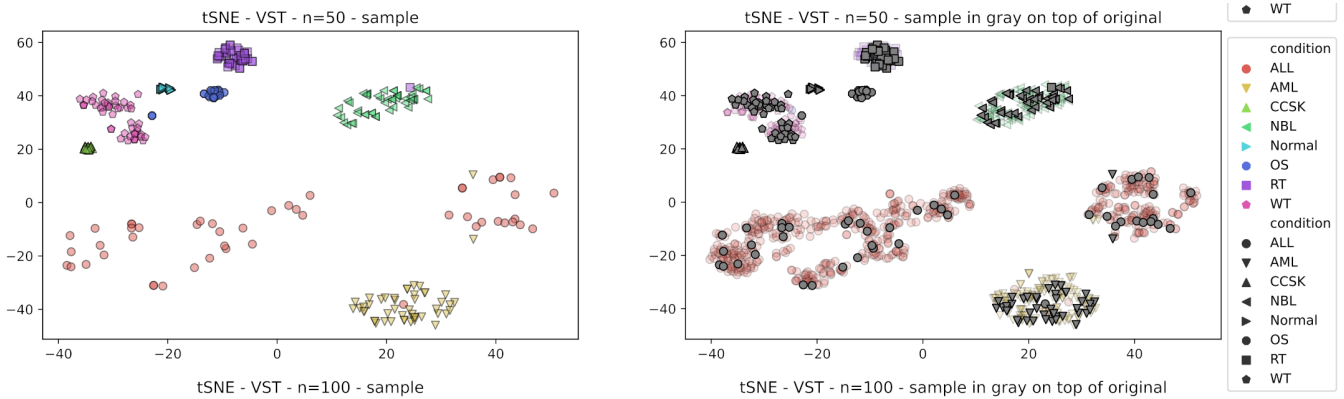
N=12



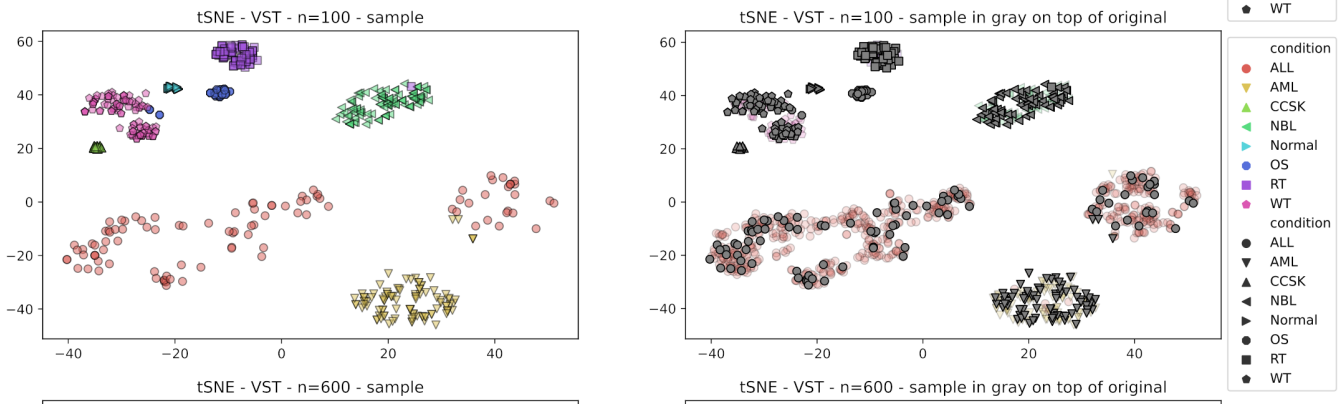


Sample with replacement from each class, ctd.

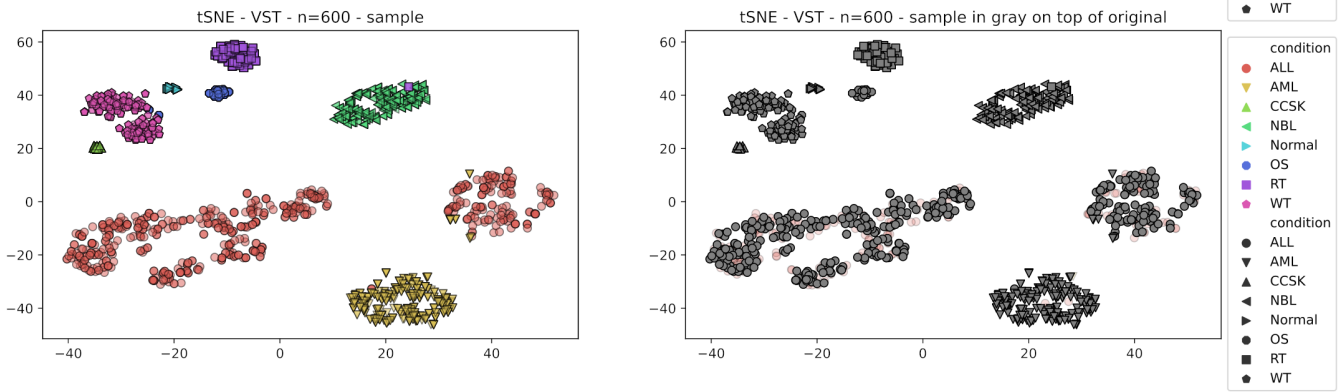
N=50



N=100

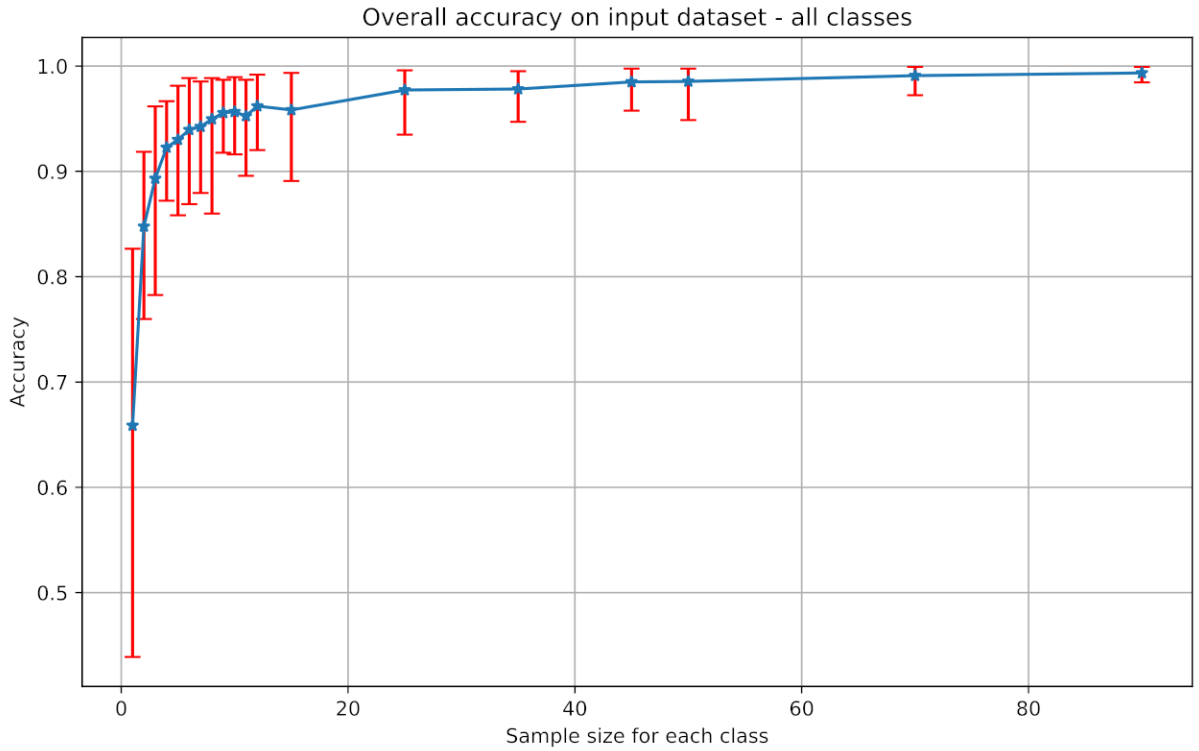


N=600



Learn from these balanced datasets of various sizes

- **For various sample sizes (*x*-axis), train a random forest classification model and compute the accuracy on the entire, imbalanced 1,228-sample dataset**
- **Run 30 different models for each differently-sized dataset**

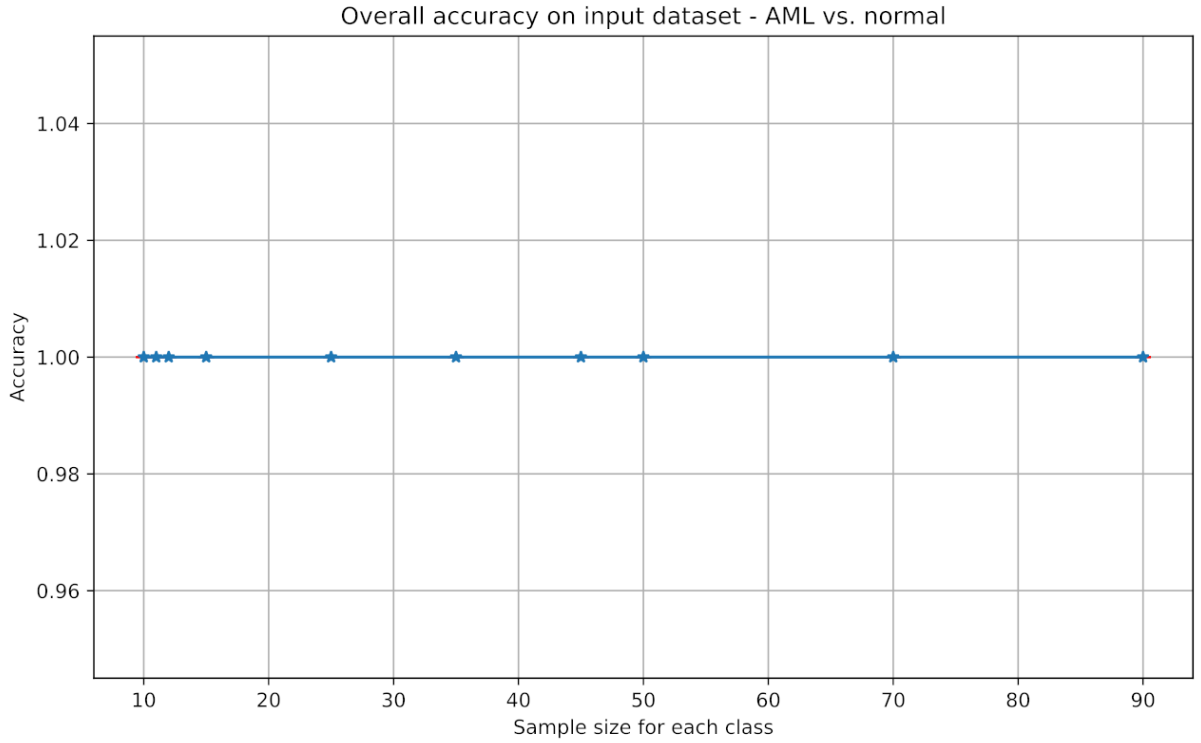


Roughly determine the most important genes

- **Running a crude feature importance study using the same random forest models, we find these most important genes for separating the eight classes (with score):**
  - 0.000953: UNC homeobox
  - 0.000903: WASP family member 3
  - 0.000896: matrix metallopeptidase 13
  - 0.000874: MyoD family inhibitor
  - 0.000806: DENN domain containing 11
  - 0.000805: small integral membrane protein 2
  - 0.000766: SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1
  - 0.000752: LY6/PLAUR domain containing 1
  - 0.000746: FZD10 antisense divergent transcript

Run the same methodology for AML vs. Normal

- **For various sample sizes, train a random forest classification model and compute the accuracy on the imbalanced dataset containing just AML and Normal samples**
- **Run 30 different models for each differently-sized dataset**

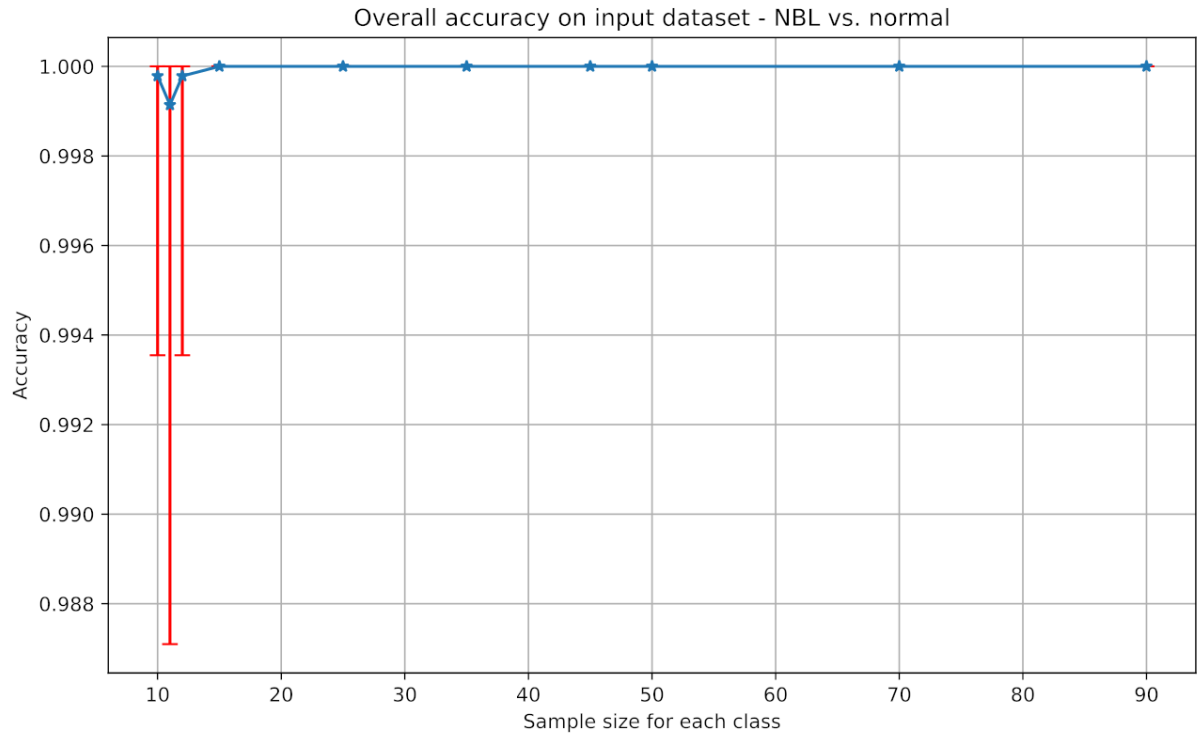


Roughly determine the most important genes

- **Running a crude feature importance study using the same random forest models, we find these most important genes for separating AML and Normal (with score):**
  - 0.000433: hyaluronidase 2
  - 0.000433: olfactory receptor family 7 subfamily E member 14 pseudogene
  - 0.000433: ERGIC and golgi 2
  - 0.000433: PDZ domain containing ring finger 3
  - 0.000433: FA complementation group B
  - 0.000433: laminin subunit gamma 2
  - 0.000433: LLGL scribble cell polarity complex component 2
  - 0.000400: POU class 3 homeobox 4
  - 0.000400: agrin
  - 0.000400: secretin receptor
  - 0.000400: adhesion G protein-coupled receptor F1
  - 0.000400: LOC339192
  - 0.000400: ADAM metallopeptidase with thrombospondin type 1 motif 9
  - 0.000400: plakophilin 4
  - 0.000400: solute carrier family 4 member 11
  - 0.000400: claudin 8
  - 0.000400: ADPGK antisense RNA 1
  - 0.000400: nucleolar protein 3
    - **Paper:** **“Expression of ARC [synonym] (apoptosis repressor with caspase recruitment domain), an antiapoptotic protein, is strongly prognostic in AML.”**

Run the same methodology for NBL vs. Normal

- **For various sample sizes, train a random forest classification model and compute the accuracy on the imbalanced dataset containing just NBL and Normal samples**
- **Run 30 different models for each differently-sized dataset**



Roughly determine the most important genes

- **Running a crude feature importance study using the same random forest models, we find these most important genes for separating NBL and Normal (with score):**
  - 0.000533: sulfatase modifying factor 1
  - 0.000533: keratinocyte differentiation factor 1
  - 0.000533: proline rich and Gla domain 1
  - 0.000533: paired like homeobox 2B
    - **Paper:** "Pleiotropic effect of common PHOX2B variants in Hirschsprung disease and neuroblastoma."
  - 0.000500: receptor accessory protein 2
  - 0.000500: methionyl-tRNA synthetase 1
  - 0.000500: DnaJ heat shock protein family (Hsp40) member C7
  - 0.000500: Nedd4 family interacting protein 2
  - 0.000500: tight junction protein 3
  - 0.000500: basigin (Ok blood group)
  - 0.000500: heterogeneous nuclear ribonucleoprotein K pseudogene 4

Conclusions

- Using even a small number of samples for training default random forest models, we obtain excellent predictability of cancer type
- The genes that are found to best discriminate are often associated with tumors, carcinoma, cell proliferation, etc.
- As a sanity check of AML vs. Normal and NBL vs. Normal, papers are easily found indicating that the differentiating genes are known indicators of these specific cancer types
- This gives us more confidence in our multiclassification model as well and indicates that we are getting the right answers for the right reasons

Potential next steps

- **Improve upon feature importance analysis:**
  - The random forest models used only default settings; they can likely be improved, particularly for determination of feature importance
  - Study the feature importance results more closely, perhaps even comparing with differential expression analyses
- **Run the same analysis on the more detailed design (i.e., X_1 and y_1) to try to learn what may differentiate between sample types**
  - In this case, the plotting of confusion matrices would likely be illuminating

Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute



Presentation Title

Presenter
Title, Affiliation

Date 1, 2018

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Secondary Slide Title

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

- **Bullet point**
- **Bullet point**
- **Bullet point**

Section Title

Presenter