Capstone Two Project: Genotyping SNP classification

Carsten Bruckner

https://github.com/cabruck/DataScienceCapstoneTwo

# Overview

In the interest of time, using this document to capture data cleaning steps, rather than implement all in Jupyter Notebook

Two primary input files:

* \git\_repositories\DataScienceCapstoneTwo\raw\_data\Output\_allps.zip\Output\_allps\genotype-inliers\filtered\Ps.performance.txt
* \git\_repositories\DataScienceCapstoneTwo\raw\_data\Output\_allps.zip\Output\_allps\CC\_ignoreNC\_probeset\_CC.txt

The output file from this procedure:

* \git\_repositories\DataScienceCapstoneTwo\data\data\_cleaning\_step1.zip\data\_cleaning\_step1.txt

The input “Ps.performance.txt” file contains standard metrics routinely generated by automated SNP quality control. Each row is summary statistics for one SNP (“probeset\_id” is the unique identifier). rows can be grouped into different categories of markers:

* standard diploid probesets that measure exactly 2 alleles. This is large majority of observations in dataset.
* multiallelic probesets: measure more than 2 alleles.
* special chromosome probesets: statistics computed only only a subset of samples (only female for X, only male for Y),
* or only measure mitochondrial DNA, which has different probeset properties
* other probesets that can report more than standard 3 genotype calls (haploid and zero copy number calls)

Many of the measured features of the probesets are specific to these special categories of markers, which means that there is high missingness of data in many features that are only computed for some categories of markers. Ther,efore this file needs to be cleaned up.

The input “CC\_ignoreNC\_probeset\_CC.txt” file contains observed accuracy metrics for a subset of probesets for which accuracy can be computed. The surrogate measurements of accuracy are various concordance metrics, of which the primary metric of interest is “CC” (overall concordance or agreement. for the probeset\_id vs an independent technology prediction of the genotype calls). This file has some rows where CC is missing, and has other rows where the concordance calculation is based on limited data (“n” samples < 50, out of over 250 possible samples measured). Therefore this input file must also be cleaned.

The “CC” metric is the metric we’re trying to predict with a machine learning model, and so it must be joined with metrics in other table. Therefore, after cleaning both input tables (limited to deleting rows and columns at this time), an inner join is done on both tables.

Some derived columns are also computed to simplify selection of rows to filter, and also as potential new features to be used for modeling.

The processing steps described herein, and resulting visualizations, were performed with the TIBCO Spotfire Analysis desktop application. I haven’t paid extra to manage data in cloud, and to support interactive access of visualizations and filtering by another user.

# Steps to clean the Ps.performance.txt file

Strategy of probeset rows to keep: diploid biallelic probesets where metrics computed on all samples.

Filter on these columns to retain these values:

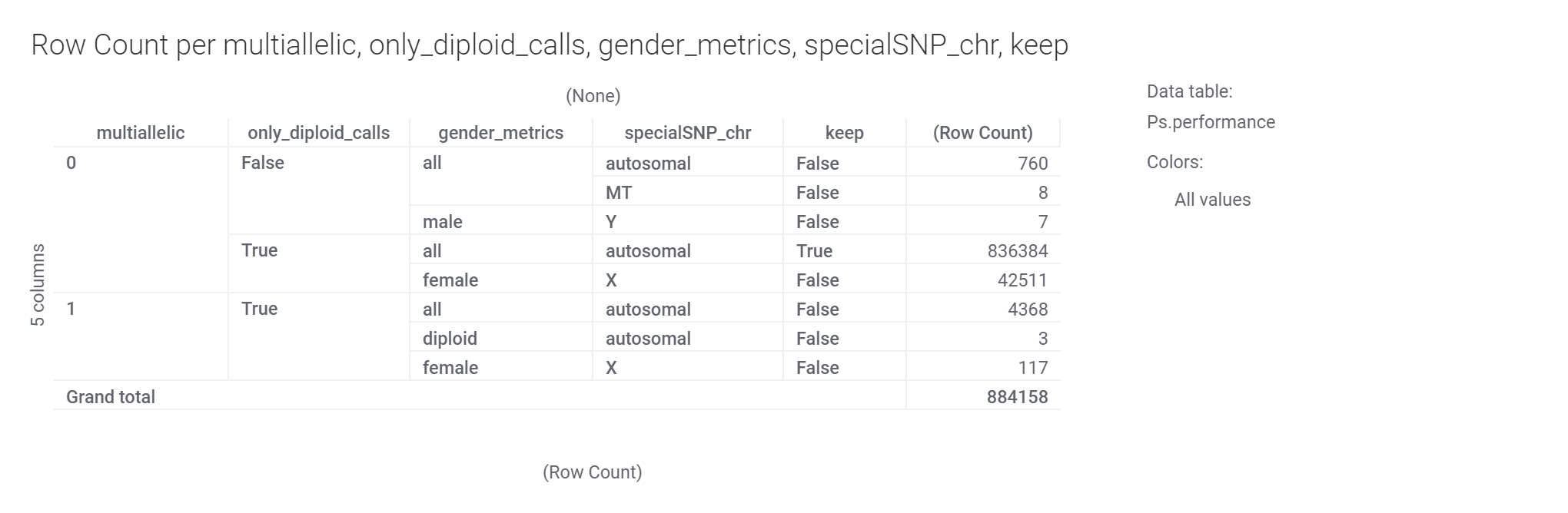
* multiallelic = 0
* gender\_metrics = "all" (all samples are used to compute feature values)
* only\_diploid\_calls = true (derived feature, which excludes remaining non-diploid calling probesets)

only\_diploid\_calls: true if sum of haploid calls is 0:

0 == ((If([n\_A] is null,0,[n\_A])) + (If([n\_B] is null,0,[n\_B])) + (If([n\_CN0] is null,0,[n\_CN0])))

keep = ([multiallelic]=0) And [only\_diploid\_calls] And ([gender\_metrics]="all")

Following visualization is row counts grouped by 5 columns, on raw input data. Intent is to retain only rows where keep = True.



Derived features for possible later use:

typical signal strength, the weighted average of each cluster's meanY:

meanY = ([n\_AA]\*[AA.meanY] + [n\_AB]\*[AB.meanY] + [n\_BB]\*[BB.meanY])/Sum([n\_AA],[n\_AB],[n\_BB])

... and another computed filter that finds probesets with unusual differences in allele strengths (might indicate susceptibility to making bad calls)

Hom.meanY.delta = Abs([AA.meanY] - [BB.meanY])

Most of the deleted columns named later are deleted either because they always have the same value, or because their values are almost completely Null for retained rows. The values are mostly Null because these metrics are only computed for specific types of probesets that are being deliberately removed.

Notes on miscellaneous columns to toss:

Use “OriginalCT” instead of “ConversionType” since the latter is a function of multiple related rows, unlike former (it includes problematic OtherMA category).

H.W.chisquared.statistic is only computed on probesets where n\_AA, n\_AB, n\_BB all are >=10. Related H.W.p-Value computed for all rows.

Maybe toss MMD (minimum Mahalanobis distance): only computed for probesets with three clusters. Should be correlated to FLD.

Data table 1: Ps.performance

1. Select Data > Add data...

Source: Data loaded from file

Type: Text

Location: C:\Users\carsten.bruckner\OneDrive\Documents\Springboard\DataScienceCourse\7 Capstone Two\MVPEF\Output\_allps\genotype-inliers\filtered\Ps.performance.txt

Data loaded at: 8/13/2021 5:57 AM

Data was added as a new data table

4. Data > Add calculated column...

Column name: only\_diploid\_calls

Expression: ((If([n\_A] is null,0,[n\_A])) + (If([n\_B] is null,0,[n\_B])) + (If([n\_CN0] is null,0,[n\_CN0])))=0

2. Data > Add calculated column...

Column name: keep

Expression: ([multiallelic]=0) And [only\_diploid\_calls] And ([gender\_metrics]="all")

3. Data > Add calculated column...

Column name: same conversion type

Expression: [ConversionType]=[OriginalCT]

Data table 2: Ps.performance.trimmed

1. Select Data > Add data...

Source: Data table from current analysis

Data table: Ps.performance

Update behavior: Automatic

Added transformations

Transformation name: Filter rows

Expression: [keep]=TRUE

Transformation name: Exclude columns

Excluded columns:

HomFLD\_hap

HomRO\_hap

n\_A

n\_B

n\_CN0

hemizygous

ConversionType

BestProbeset

BestandRecommended

A.meanX

A.meanY

B.meanX

B.meanY

CN0.meanX

CN0.meanY

count\_ma\_A

count\_ma\_B

count\_ma\_C

count\_ma\_D

count\_ma\_E

count\_ma\_F

FLD\_MA

MinFLD\_MA

HomFLD\_MA

HetSO\_MA

HomRO\_MA

same conversion type

nSamples

nCalls

nAllelesTested

nAllelesDetected

NHetClus

nMajorAlleles

nMinorAlleles

MAFall

MAFmax

HomCount

MajorHomCount

MinorHomCount

HetCount

HomMMA

NC.meanX

NC.meanY

maxMinorAllele

H.W.chisquared.statistic

affy\_snp\_id

multi\_snp\_id

CopyNumIssue

Data loaded at: 8/16/2021 11:59 AM

Data was added as a new data table

2. Data > Add calculated column...

Column name: meanY

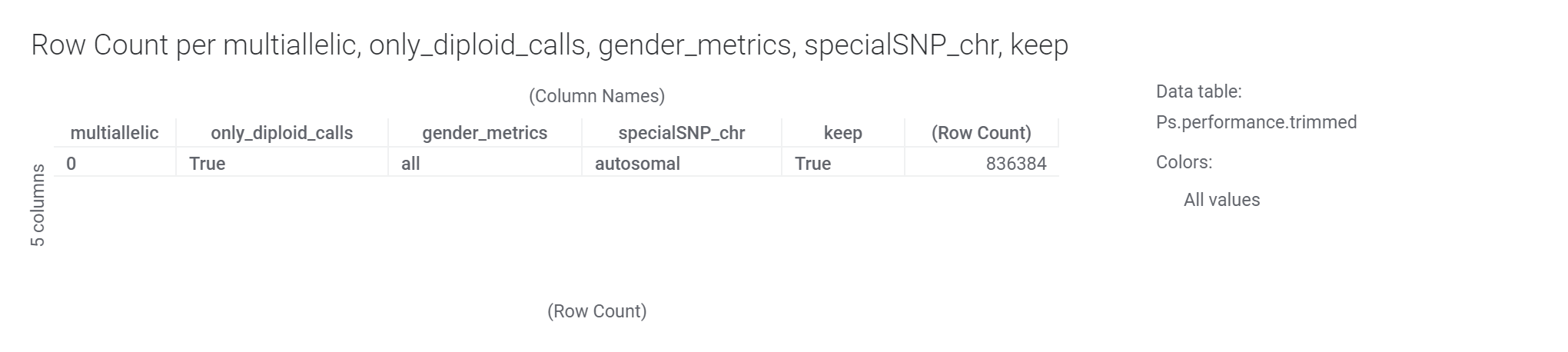
Expression: (([n\_AA] \* [AA.meanY]) + ([n\_AB] \* [AB.meanY]) + ([n\_BB] \* [BB.meanY])) / Sum([n\_AA],[n\_AB],[n\_BB])

3. Data > Add calculated column...

Column name: Hom.meanY.delta

Expression: Abs([AA.meanY] - [BB.meanY])

After this filtering step to create new Ps.performance.trimmed data table, repeat the previous visualization on new data table:



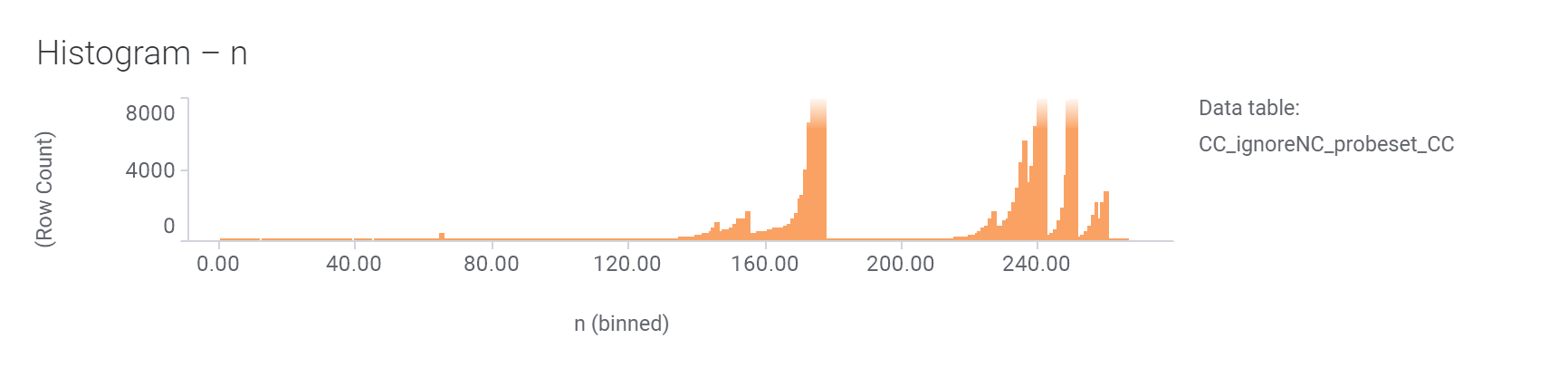
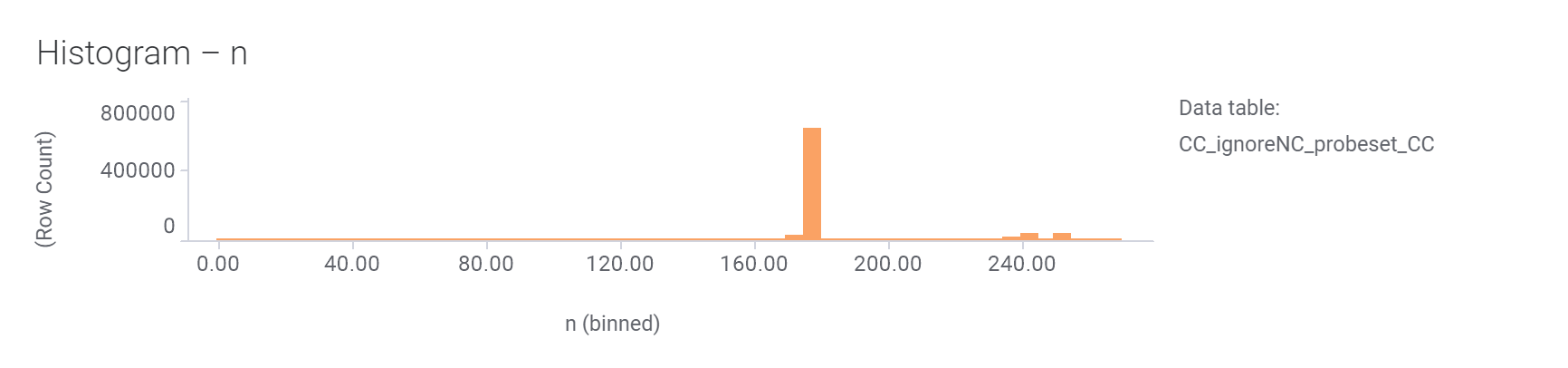
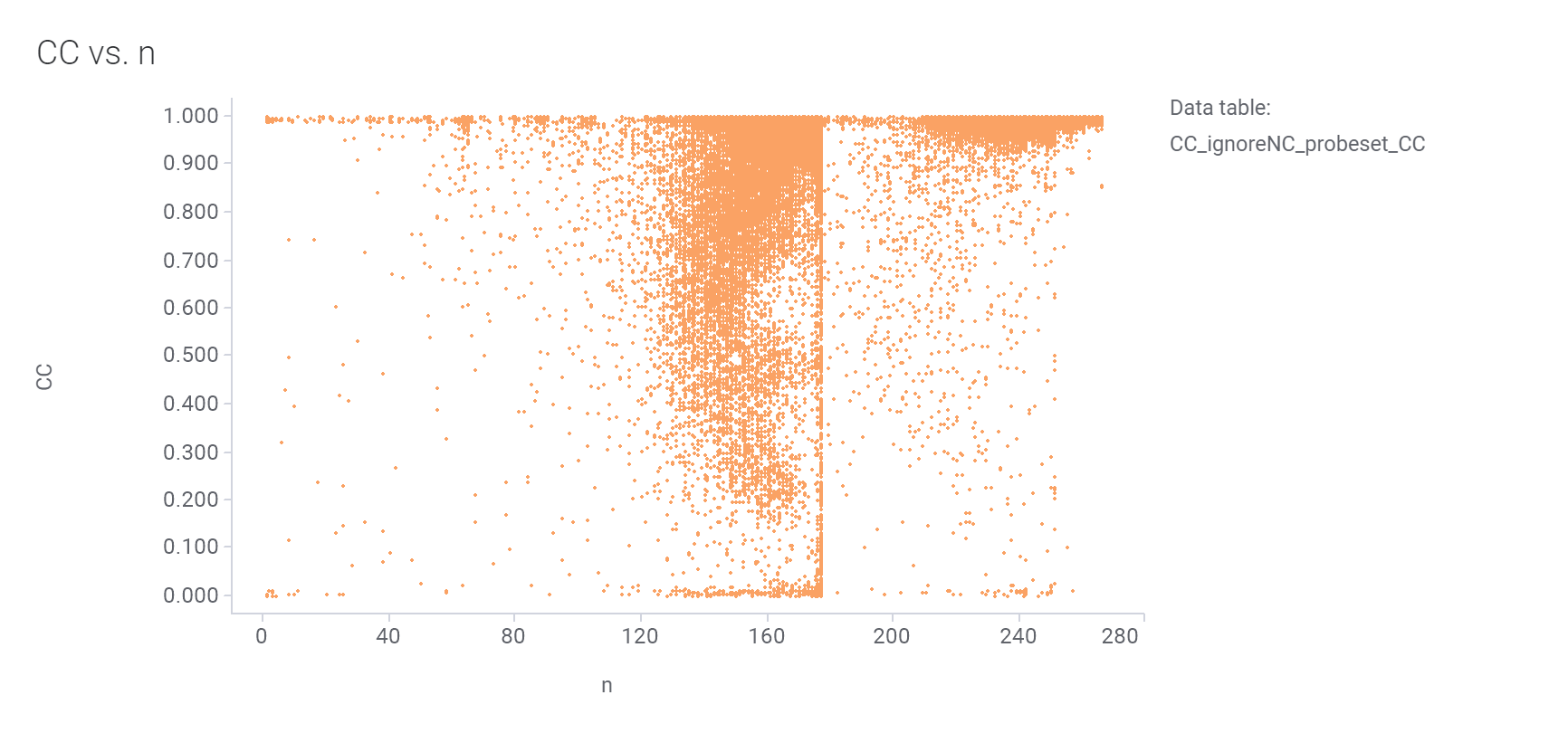
Starting table has 884158 rows, trimmed table has 836384 rows, so we retained about 95% of rows. Starting table has 86 imported columns, trimmed table has 43 columns (4 of which were derived).

# Steps to clean the CC\_ignoreNC\_probeset\_CC.txt file

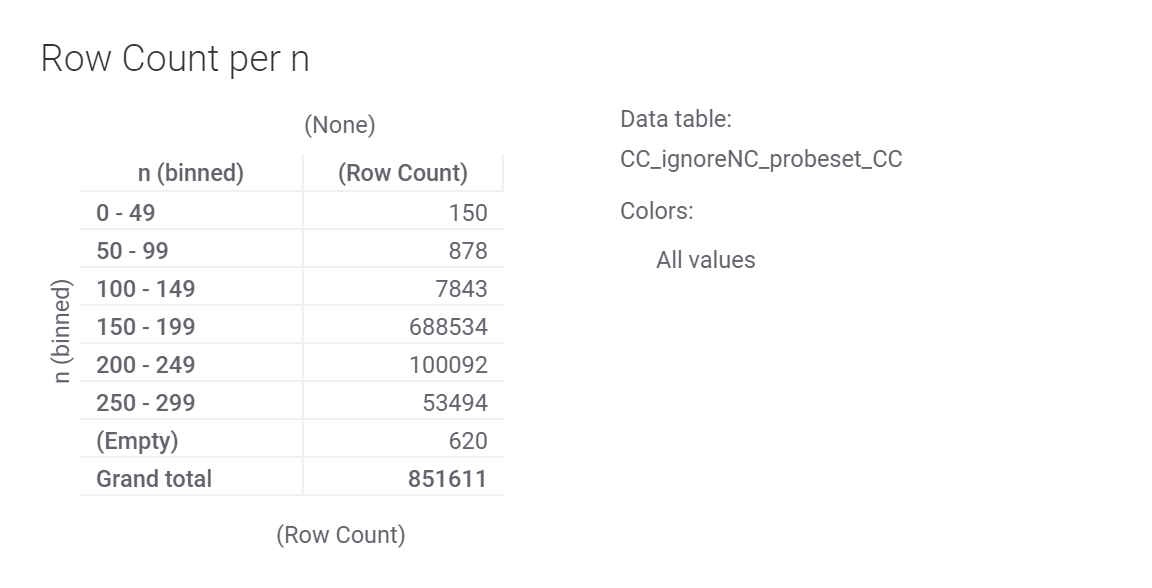
Other input file is observed Concordance (CC). This is the feature we're trying to predict by modeling of other features

Source file: CC\_ignoreNC\_probeset\_CC.txt

Concordance is computed over n samples (CC = 1 is best, 100% agreement with reference data). The larger the n, the more reliable the concordance calculation as a predictor of probeset performance. Here’s a scatterplot of CC vs n, followed by a histogram of count of n, and a zoomed-in histogram:



Most CC measurements are computed over more than 160 samples (n > 160). Let’s remove the small fraction of probesets where the CC predictor is less reliable (n<50, 150 probesets), or not even available (620 probesets):



Data cleaning: some CC values are empty. Exclude these probeset rows.

Some probesets (150) have <50 samples used to measure concordance, so these CC values are not as reliable. Exclude these probeset rows.

Actually, both row filtering steps can be accomodated by single filter: keep rows where n >= 50. Empty CC rows are also empty n rows, and are removed.

Data Table 3: Predictor features from different table:

“CC\_ignoreNC\_probeset\_CC”:

1. Select Data > Add data...

Source: Data loaded from file

Type: Text

Location: C:\Users\carsten.bruckner\OneDrive\Documents\Springboard\DataScienceCourse\7 Capstone Two\MVPEF\Output\_allps\CC\_ignoreNC\_probeset\_CC.txt

Data loaded at: 8/16/2021 10:04 AM

Data was added as a new data table

2. Data > Add calculated column...

Column name: CC Binned

Expression: BinBySpecificLimits([CC],0.899,0.949,0.959,0.969,0.979,0.989,1)

Data Table 4: filtered predictor features

"CC\_ignoreNC\_probeset\_CC.trimmed"

1. Select Data > Add data...

Source: Data table from current analysis

Data table: CC\_ignoreNC\_probeset\_CC

Update behavior: Automatic

Added transformations

Transformation name: Change column names

Columns to rename:

n

n\_het

Expression: [%C]&"\_CC"

Transformation name: Filter rows

Expression: [n\_CC]>=50

Transformation name: Exclude columns

Excluded columns:

CC\_major\_hom

CC\_minor\_hom

n\_major\_hom

n\_minor\_hom

MAC

n\_refmaj\_maj

n\_refmaj\_het

n\_refmaj\_min

n\_refmaj\_nc

n\_refhet\_maj

n\_refhet\_het

n\_refhet\_min

n\_refhet\_nc

n\_refmin\_maj

n\_refmin\_het

n\_refmin\_min

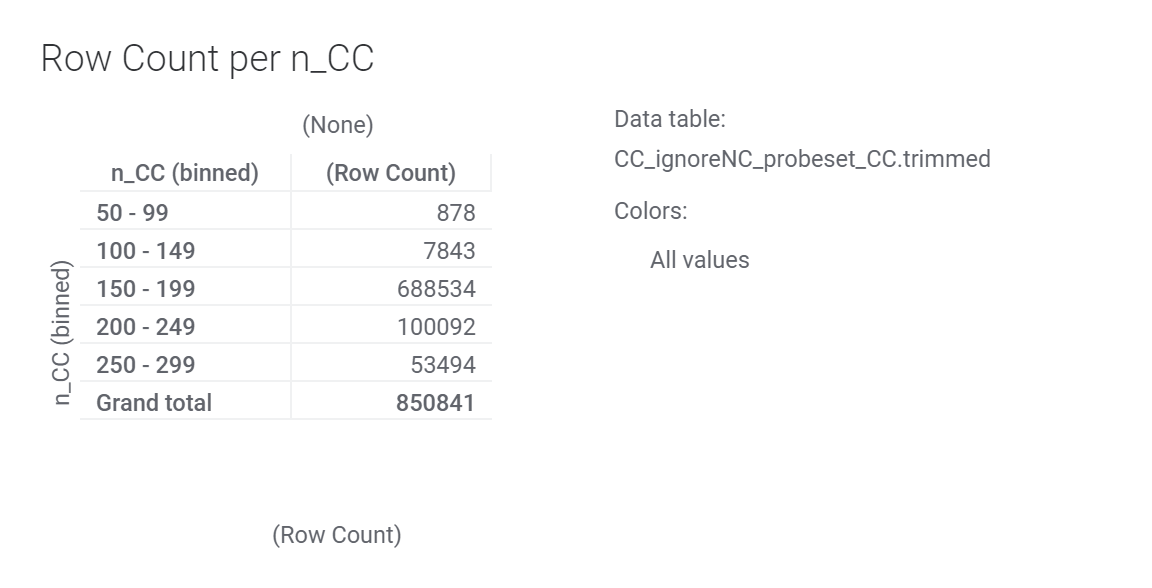
n\_refmin\_nc

minor\_allele

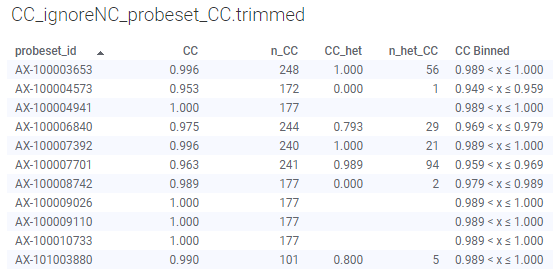
Data loaded at: 8/16/2021 11:44 AM

Data was added as a new data table

The following table shows row counts by “n” bin (renamed to n\_CC) after keeping only rows where n has a value and is at least 50. Compare to previous graphic.



There are now 850841 rows, vs 851611 rows in original CC file, retaining almost all rows. The named columns that are excluded in trimmed table are detail metrics useful for understanding the kinds of discordances, and will not be used here. Screenshot of retained columns, some of which might be used later:



Notes for downstream feature engineering:

* As a predicted feature, due to skew in CC, maybe convert CC to two or three categorical bins, like:
  + if binary categories
    - "Suspect" = "no" if CC >=99
    - Suspect" = "yes" if not "no"
  + If ternary:
    - "Suspect" = "no" if CC=100
    - "Suspect" ="maybe" if not "no" and CC >=99
    - "Suspect" ="yes" if not "no" and not "maybe"

# Join files, and remove bookkeeping columns

Data Table 5: inner join of independent and predictor features from both .trimmed tables

"inner\_join"

1. Select Data > Add data...

Source: Data table from current analysis

Data table: CC\_ignoreNC\_probeset\_CC.trimmed

Update behavior: Automatic

Data loaded at: 8/16/2021 11:48 AM

Data was added as a new data table

2. Select Data > Add data...

Source: Data table from current analysis

Data table: Ps.performance.trimmed

Update behavior: Automatic

Data loaded at: 8/16/2021 11:59 AM

Data was added as new columns in data table 'inner\_join'

Matching behavior: Tries to match the specified columns when data is loaded

Matched columns: probeset\_id – probeset\_id

Added columns:

CR

FLD

HomFLD

HetSO

HomRO

nMinorAllele

Nclus

n\_AA

n\_AB

n\_BB

n\_NC

specialSNP\_chr

gender\_metrics

HomHet

AA.meanX

AA.meanY

AA.varX

AA.varY

AB.meanX

AB.meanY

AB.varX

AB.varY

BB.meanX

BB.meanY

BB.varX

BB.varY

AA.varX.Z

AA.varY.Z

AB.varX.Z

AB.varY.Z

BB.varX.Z

BB.varY.Z

MMD

MinorAlleleFrequency

H.W.p-Value

multiallelic

OriginalCT

ordered\_alleles

keep

only\_diploid\_calls

meanY

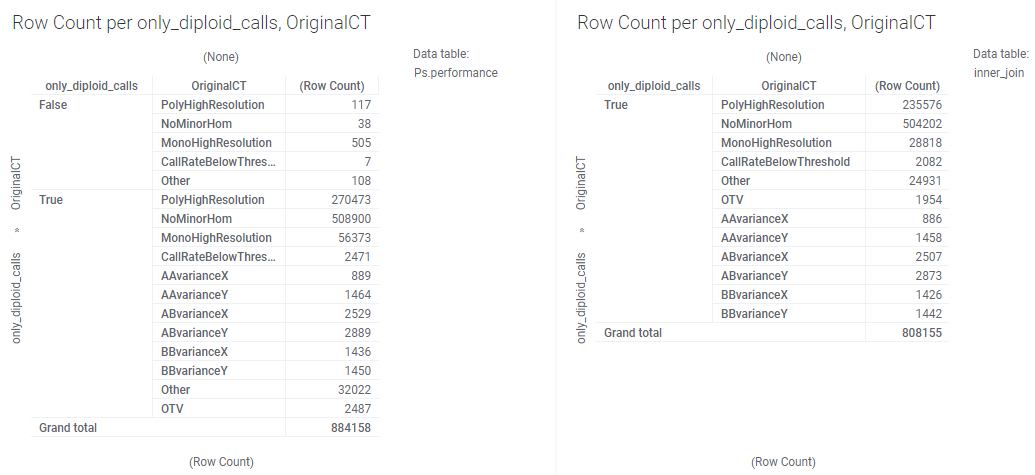
Hom.meanY.delta

Ignored columns: (None)

Join method: Inner join

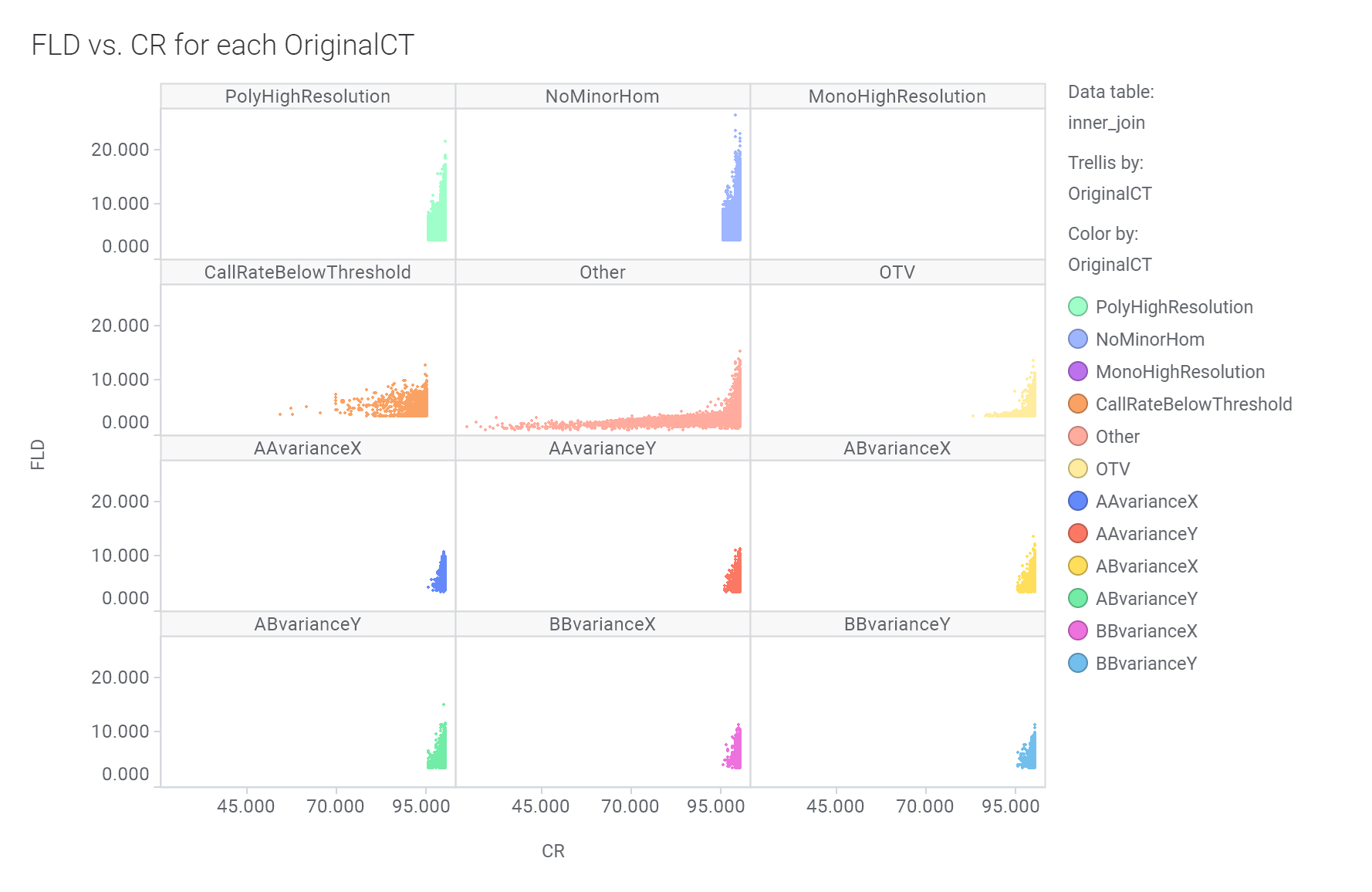
Treat empty values as equal: No

The following visualization counts the probesets in original Ps.performance.txt and in final joined file, grouped by whether diploid probeset and automate Original Conversion Type (OriginalCT). Roughly 90% of original rows are still available for modeling. Some probesets in Ps.performance\_filtered table were not available in CC table, so inner join removed them:



“OriginalCT” is an automated categorization into recommended vs non-recommended probesets for end users to use in downstream steps. The only categories recommended for downstream use are “PolyHighResolution”, “NoMinorHom”, and “MonoHighResolution”. These category assignments have no knowledge of predicted concordance to another technology. As a reminder, the goal of this project is to see whether we can develop a model that predicts “good vs bad” concordance better than purely binning by “recommended vs non-recommended” categories.

The following visualization highlights the sparseness in the remaining table. Two valuable features that predict performance are Call Rate (CR, the % of samples returning a genotype call), and FLD (Fisher’s Linear Discriminant, a measure of cluster separation, which makes it easier to make a clear call).



The FLD metric is not computed for MonoHighResolution markers, which is why that pane of the trellis plot is empty. The various “variance” conversion types are only computed for probeset rows where CR >=95%. Apparently this supplemental computation is only performed on markers that might otherwise pass QC (appear in top three panes), and flags additional probesets into one of these 6 non-recommended conversion types.

Low FLD probesets are likely to have reduced call rate and lower calling accuracy.

Because some of the columns in inner\_join table are used only for some visualizations and QC of filtering operations, the following are excluded from export to cleaned data file:

CC Binned

specialSNP\_chr

gender\_metrics

multiallelic

keep

only\_diploid\_calls

# Summary stastics for table at end of this procedure

Columns with Count <808155 are sparse.

| Column | Count | UniqueCount | Min | Median | Avg | Max | StdDev |
| --- | --- | --- | --- | --- | --- | --- | --- |
| probeset\_id | 808155 | 808155 | AX-100003653 |  |  | AX-98295645 |  |
| CC | 808155 | 5563 | 0 | 1 | 0.993 | 1 | 0.054 |
| CC\_het | 764300 | 1175 | 0 | 1 | 0.977 | 1 | 0.13 |
| n\_CC | 808155 | 217 | 50 | 177 | 187.79 | 266 | 26.95 |
| n\_het\_CC | 764300 | 149 | 1 | 6 | 16.14 | 152 | 24.71 |
| CR | 808155 | 186 | 23.913 | 100 | 99.431 | 100 | 2.229 |
| FLD | 777314 | 86758 | 1.072 | 6.782 | 6.861 | 26.374 | 1.583 |
| HomFLD | 258881 | 109465 | 2.546 | 14.656 | 14.895 | 45.06 | 3.25 |
| HetSO | 777343 | 16299 | -2.496 | 0.325 | 0.318 | 2.402 | 0.192 |
| HomRO | 808126 | 43673 | -2.344 | 1.933 | 1.976 | 6.401 | 0.687 |
| nMinorAllele | 808155 | 277 | 0 | 9 | 29.17 | 276 | 49.62 |
| Nclus | 808155 | 3 | 1 | 2 | 2.28 | 3 | 0.53 |
| n\_AA | 808155 | 277 | 0 | 2 | 99.72 | 276 | 122.08 |
| n\_AB | 808155 | 251 | 0 | 9 | 21.85 | 276 | 30.34 |
| n\_BB | 808155 | 277 | 0 | 237 | 152.86 | 276 | 124.66 |
| n\_NC | 808155 | 186 | 0 | 0 | 1.57 | 210 | 6.15 |
| HomHet | 808155 | 2 | 0 | 1 | 0.64 | 1 | 0.48 |
| AA.meanX | 808155 | 42299 | -2.635 | 2.389 | 2.394 | 6.54 | 0.614 |
| AA.meanY | 808155 | 37484 | 7.246 | 10.142 | 10.16 | 14.395 | 0.598 |
| AA.varX | 808155 | 6381 | 0.03 | 0.03 | 0.165 | 1.026 | 0.155 |
| AA.varY | 808155 | 4565 | 0.03 | 0.03 | 0.115 | 1.401 | 0.099 |
| AB.meanX | 808155 | 31431 | -4.045 | 0.33 | 0.302 | 4.651 | 0.425 |
| AB.meanY | 808155 | 39838 | 7.469 | 10.501 | 10.486 | 14.641 | 0.594 |
| AB.varX | 808155 | 6242 | 0.03 | 0.157 | 0.152 | 1.01 | 0.108 |
| AB.varY | 808155 | 6152 | 0.03 | 0.134 | 0.134 | 1.748 | 0.097 |
| BB.meanX | 808155 | 41427 | -8.06 | -1.793 | -1.84 | 2.344 | 0.625 |
| BB.meanY | 808155 | 40363 | 7.27 | 10.142 | 10.146 | 14.002 | 0.677 |
| BB.varX | 808155 | 5770 | 0.03 | 0.26 | 0.211 | 1.199 | 0.14 |
| BB.varY | 808155 | 5153 | 0.03 | 0.191 | 0.166 | 1.054 | 0.11 |
| AA.varX.Z | 246168 | 5782 | -3.263 | -0.716 | -1.129 | 7.998 | 1.751 |
| AA.varY.Z | 246168 | 4131 | -2.971 | -0.721 | -1.029 | 20.298 | 1.629 |
| AB.varX.Z | 246168 | 4763 | -2.667 | -0.199 | -0.079 | 11.958 | 1.083 |
| AB.varY.Z | 246168 | 4654 | -2.333 | -0.209 | -0.069 | 15.347 | 1.061 |
| BB.varX.Z | 246168 | 5151 | -3.868 | -0.467 | -0.969 | 9.866 | 1.887 |
| BB.varY.Z | 246168 | 4550 | -3.036 | -0.504 | -0.761 | 11.694 | 1.575 |
| MMD | 258516 | 179352 | 7.642 | 43.869 | 44.214 | 129.087 | 9.141 |
| MinorAlleleFrequency | 808155 | 4216 | 0 | 0.017 | 0.053 | 0.5 | 0.091 |
| H.W.p-Value | 808155 | 34370 | 0 | 1 | 0.764204547 | 1 | 0.37719077 |
| OriginalCT | 808155 | 12 | PolyHighResolution |  |  | BBvarianceY |  |
| ordered\_alleles | 808155 | 855 | -/A |  |  | T/TCTT |  |
| meanY | 808155 | 779856 | 7.246 | 10.134 | 10.147 | 14.384 | 0.69 |
| Hom.meanY.delta | 808155 | 19601 | 0 | 0.183 | 0.216 | 4.389 | 0.174 |

Known issues:

If n\_AA = 0, might need to reset all AA.\* metrics to Null, and same idea with n\_AB=0, n\_BB=0. The AA.\* metrics currently report default values if there are no observations of AA cluster in dataset.