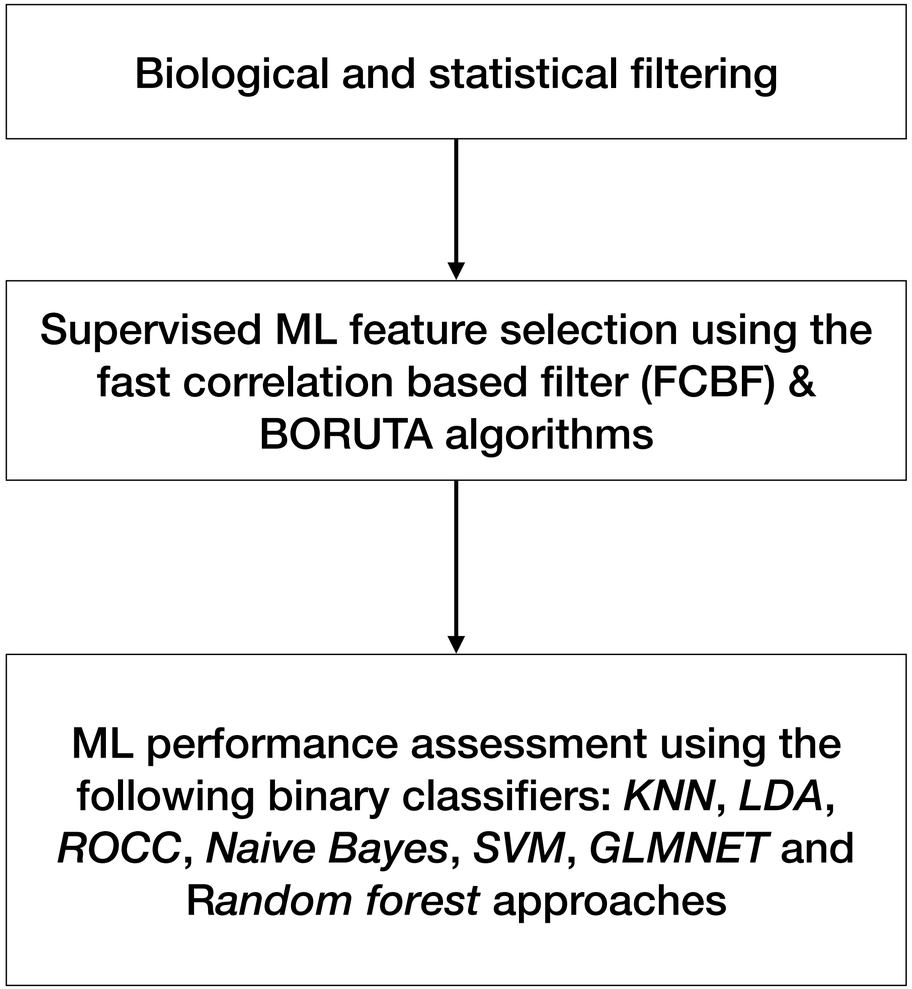
# GAW20 data analysis progress

## Data filtering

James Jafali (June 17, 2022)

# Approach



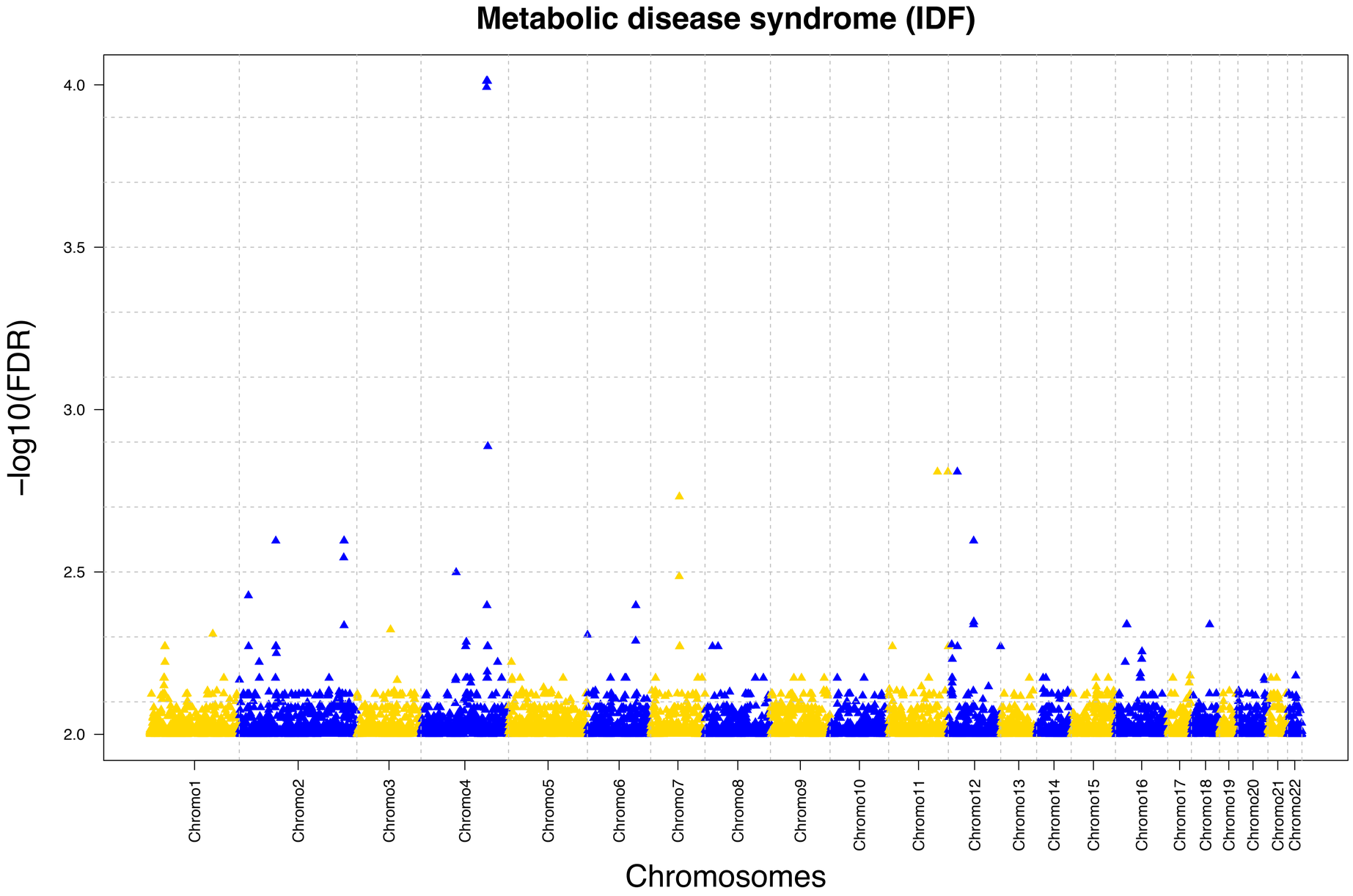

# Phenotype data

| **Total** | **N=1105.** |
|---|---|
| **Age in years** | |
| Mean (SD) | 48.224 (16.266) |
| Range | 18.000 - 87.000 |
| **Study site** | |
| Minnesota | 565 (51.1%) |
| Utah | 540 (48.9%) |
| **Smoking status** | |
| N-Miss | 1 |
| Never smoker | 780 (70.7%) |
| Past smoker | 239 (21.6%) |
| Current smoker | 85 (7.7%) |
| **HDL change (post-pre)** | |
| N-Miss | 244 |
| Mean (SD) | 2.815 (5.376) |
| Range | -23.500 - 25.000 |
| **Metabolic disease syndrom (ATP-III)** | |
| Negative | 687 (62.2%) |
| Positive | 418 (37.8%) |
| **Metabolic disease syndrom (IDF)** | |
| Negative | 647 (58.6%) |
| Positive | 458 (41.4%) |

# Data filtering summary

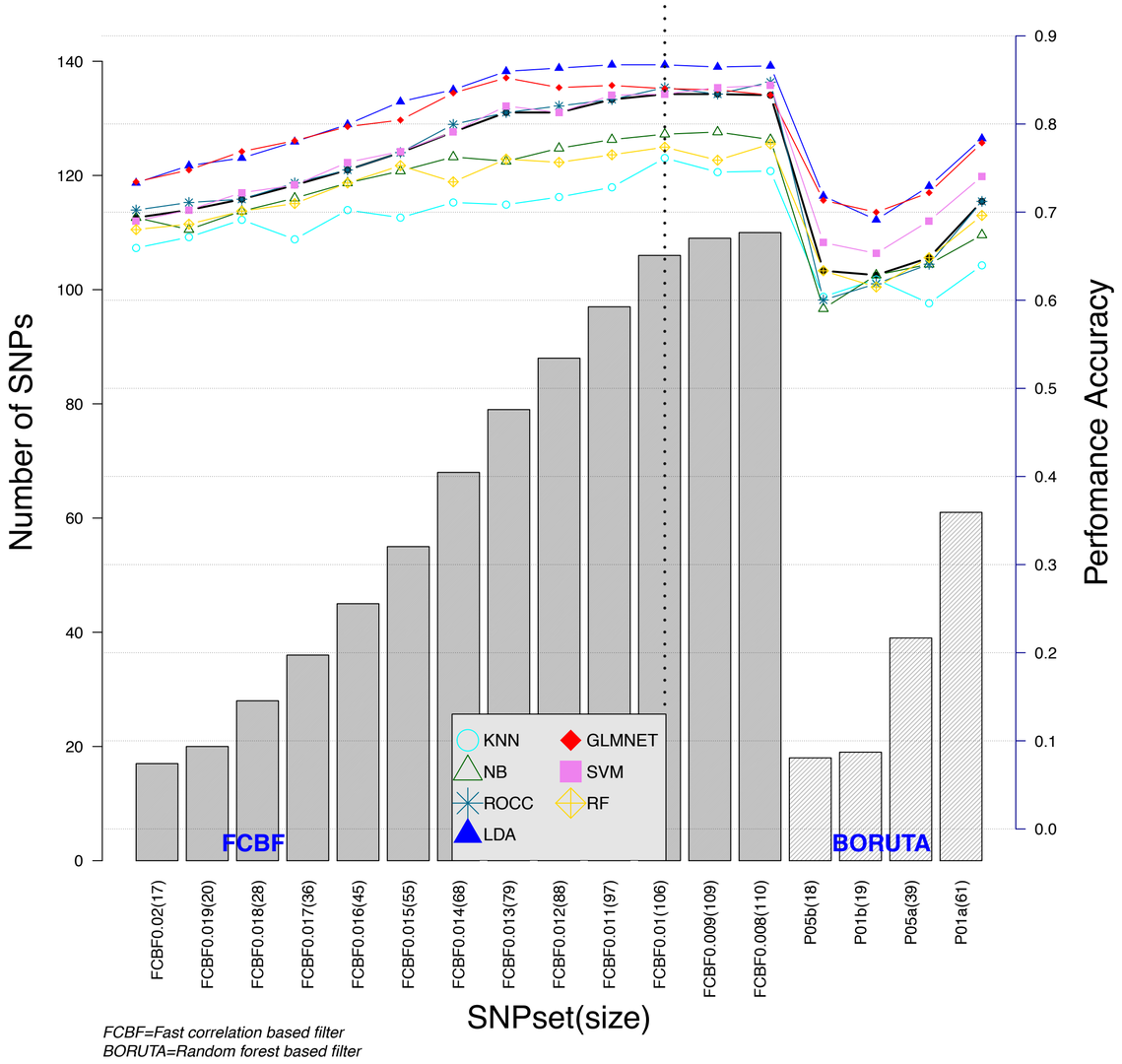
| Filter | Pass | | Fail | |
|---|---|---|---|---|
| | SNPs | Percent | SNPs | Percent |
| Total number of SNPs | 715069 | 100% | | |
| Missing values < 5% | 714966 | 99.99% | 103 | 0.01% |
| Minor allele frequency > 5% | 589679 | 82.46% | 125390 | 17.54% |
| Hardy Weinberg P-value > 1% | 694941 | 97.19% | 20128 | 2.81% |
| **Statistics association filters (P-value <0.05)** | | | | |
| Age | 42992 | 6.01% | 672077 | 93.99% |
| Smoking | 68310 | 9.55% | 646759 | 90.45% |
| Location (Minnesota Vs Utah) | 144196 | 20.17% | 570873 | 79.83% |
| HDL_change (Post-Pre) | 42078 | 5.88% | 672991 | 94.12% |
| Metabolic disease syndrome (ATP-III) | 39134 | 5.47% | 675935 | 94.53% |
| Metabolic disease syndrome (IDF) | 39720 | 5.55% | 675349 | 94.45% |
| **Statistics filters (P-value <0.01)** | | | | |
| Age | 9204 | 1.29% | 705865 | 98.71% |
| Smoking | 21618 | 3.02% | 693451 | 96.98% |
| Location (Minnesota Vs Utah) | 66694 | 9.33% | 648375 | 90.67% |
| HDL_change (Post-Pre) | 8924 | 1.25% | 706145 | 98.75% |
| Metabolic disease syndrome (ATP-III) | 8303 | 1.16% | 706766 | 98.84% |
| Metabolic disease syndrome (IDF) | 8583 | 1.2% | 706486 | 98.8% |
| **Combined Filter** (Missing values<5%, MAF<5%, HWB p-value >1% & IDF P-value<1%) | 7831 | 1.1% | 707238 | 98.9% |

# Manhattan plot showing the FDR values for the filtered SNPs (m=7831)



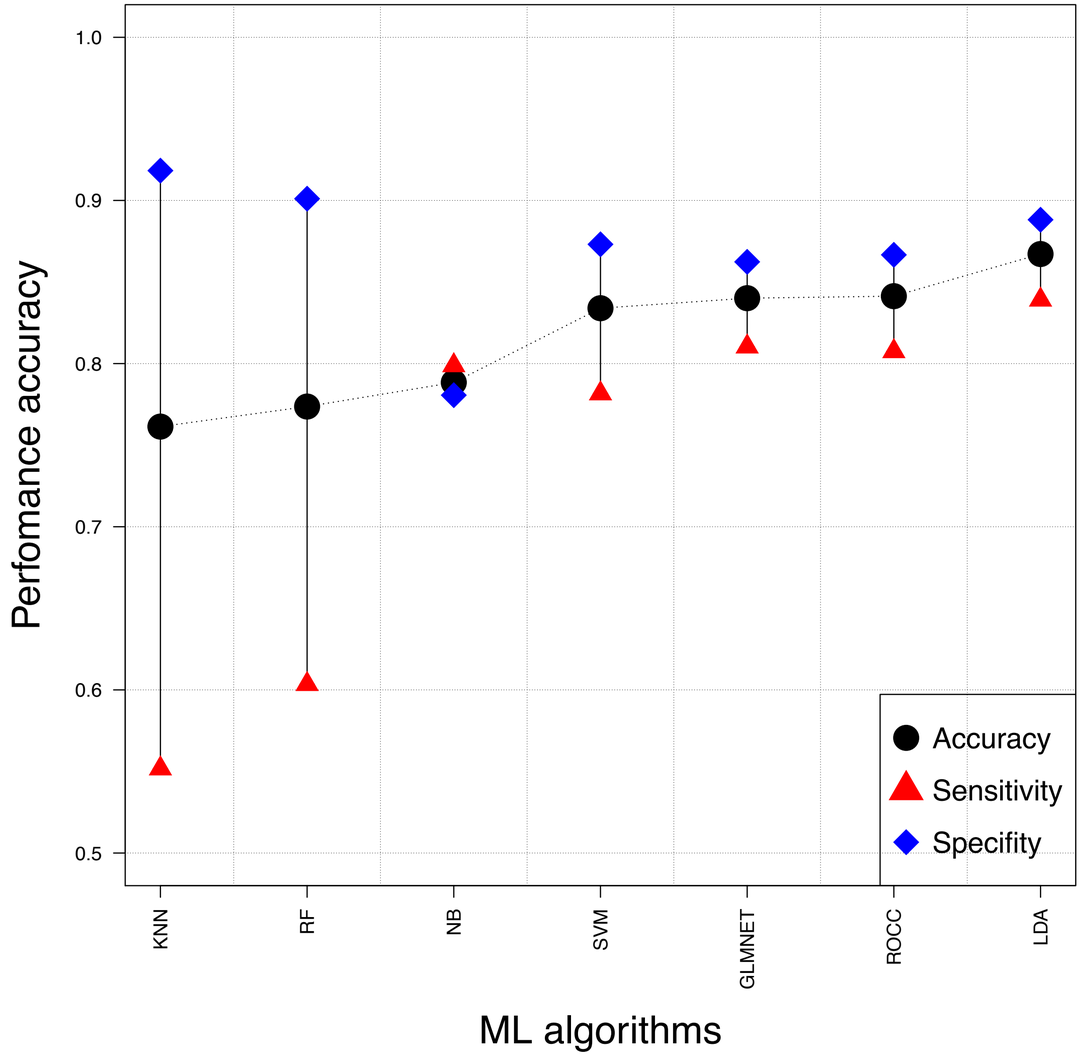
# Supervised ML analysis summary for metabolic disease syndrome using the filtered SNPs (m=7831)

**(A)**




*FCBF=Fast correlation based filter*
*BORUTA=Random forest based filter*

**(B)**




Thanks