# Unknown Primary Classification Summary

## 1. Replicating the work of tumorTracer

- The data used in the paper is available in the paper's additional files.
- The following results are replicated using their data

The following results are obtained when we train a Random Forest Model with ntree = 500 and feature set as 233 Genes with each getting marked as 1 (for amplification), -1 (for Deletion) or 0 (for No Copy Number aberration in that gene). The results are similar to that reported in the paper (except for a swap between lung and breast). OOB estimate of error rate **- 20.78%**

| | breast | endometrium | kidney | large.intestine | lung | ovary | Class Error |
|---|---|---|---|---|---|---|---|
| breast | 540 | 65 | 5 | 14 | 27 | 29 | 0.2058824 |
| endometrium | 42 | 115 | 0 | 8 | 5 | 27 | 0.4162437 |
| kidney | 6 | 1 | 209 | 13 | 8 | 3 | 0.1291667 |
| large.intestine | 17 | 18 | 2 | 330 | 17 | 5 | 0.1516710 |
| lung | 48 | 5 | 6 | 19 | 280 | 23 | 0.2650919 |
| ovary | 30 | 2 | 0 | 4 | 20 | 314 | 0.1513514 |

The following results are obtained when we train a Random Forest Model with ntree = 500 and feature set as 232 Genes with each getting marked on how many point mutation the gene has. The results are similar to that reported in the paper (except for a swap between endometrium and breast). OOB estimate of error rate **- 30.44%**

| | breast | endometrium | kidney | large.intestine | lung | ovary | Class Error |
|---|---|---|---|---|---|---|---|
| breast | 504 | 7 | 23 | 7 | 33 | 106 | 0.2588235 |
| endometrium | 28 | 120 | 0 | 26 | 10 | 13 | 0.3908629 |
| kidney | 81 | 0 | 135 | 2 | 16 | 6 | 0.4375000 |
| large.intestine | 11 | 9 | 3 | 344 | 15 | 7 | 0.1156812 |
| lung | 66 | 1 | 16 | 23 | 246 | 29 | 0.3543307 |
| ovary | 100 | 3 | 11 | 8 | 27 | 221 | 0.4027027 |

The following results are obtained when we train a Random Forest Model with ntree = 500 and feature set as the union of the above two (PM + CNV). OOB estimate of error rate **- 14.31%**

| | breast | endometrium | kidney | large.intestine | lung | ovary | Class Error |
|---|---|---|---|---|---|---|---|
| breast | 616 | 4 | 3 | 4 | 24 | 29 | 0.09411765 |
| endometrium | 22 | 135 | 0 | 8 | 3 | 29 | 0.31472081 |
| kidney | 9 | 0 | 217 | 5 | 7 | 2 | 0.09583333 |
| large.intestine | 11 | 12 | 5 | 352 | 5 | 4 | 0.09511568 |
| lung | 52 | 0 | 10 | 6 | 292 | 21 | 0.23359580 |
| ovary | 33 | 0 | 0 | 2 | 13 | 322 | 0.12972973 |

**Why do CNVs contribute so much?**

- Average number of **CNV** events per sample = **90.5**
- Average number of **somatic** mutations per sample = **5.8**

***** All files needed for the above results can be found in the following folder of the shared Google Drive **- Primary source prediction for cancer tumors/Final Report/Replicating-TumorTracer**

## 2. Is there something special about tumorTracer 232 Gene List?

The data set used by the paper is COSMIC v68 (to which we do not have access). Hence, to do an analysis using STRAND's 152 Gene List as feature set, we used COSMIC v79 data set. '

The data is taken from here - COSMIC

## COSMIC v79 Results using 231 Genes

231 because 1 gene name could not be mapped using Ensemble. This result is obtained by using only Copy Number Aberrations (A/D/N) as features and RF Model. OOB estimate of error rate **- 35.22%**

| | breast | endometrium | kidney | large.intestine | lung | ovary | Class Error |
|---|---|---|---|---|---|---|---|
| breast | 1173 | 10 | 33 | 36 | 108 | 68 | 0.1785714 |
| endometrium | 77 | 5 | 32 | 18 | 57 | 50 | 0.9790795 |
| kidney | 67 | 1 | 137 | 18 | 84 | 14 | 0.5732087 |
| large.intestine | 58 | 3 | 30 | 299 | 44 | 17 | 0.3370288 |
| lung | 118 | 3 | 41 | 29 | 632 | 65 | 0.2882883 |
| ovary | 87 | 6 | 32 | 23 | 138 | 268 | 0.5162455 |

**Why is the result significantly different from v68?**

- Average number of **CNV** events per sample in this data set (for these classes) is only **31.5**

# COSMIC v79 Results using 152 Genes

This result is obtained by using only Copy Number Aberrations (A/D/N) as features and RF Model. **OOB** estimate of  error rate - **29.56%**

| | Breast | Lung | Ovary | Class Error |
|---|---|---|---|---|
| Breast | 1111 | 227 | 92 | 0.2230769 |
| Lung | 158 | 593 | 113 | 0.3136574 |
| Ovary | 82 | 169 | 300 | 0.4555354 |

### Why is the result better than 231 Gene List?

- This might be because of leaving out difficult classes like endometrium and kidney.
- Average number of **CNV** events per sample in this data set (for these classes) is **24**

The following results were obtained on STRAND test data.

| | Breast | Lung | Ovary |
|---|---|---|---|
| Breast | 24 | 7 | 6 |
| Lung | 17 | 8 | 7 |
| Ovary | 9 | 2 | 1 |

### Why is the result not good?

- Average number of **CNV** events per sample in this data set (for these classes) is **only 6.5**

The following results were obtained when the RF model was run with a feature set as Point Mutation status of the 152 genes (0/1 showing whether the gene is mutated or not). OOB estimate of  error rate - **17.79%**

| | breast | kidney | lung | ovary | Class Error |
|---|---|---|---|---|---|
| breast | 6897 | 112 | 263 | 138 | 0.06923077 |
| kidney | 328 | 1530 | 82 | 59 | 0.23461731 |
| lung | 1976 | 146 | 18902 | 338 | 0.11515776 |
| ovary | 1577 | 158 | 842 | 495 | 0.83886719 |

This model seems to have learnt quite well. But, **Every sample on an average has only 1.14 genes (out of 152) marked to be mutated.**

### Why is the model learning well in spite of a very low variance in the data?

- This is because each class is having a marker gene that is helping the classification. The following table gives the highest marked gene's name for each class. This tells us that these four genes along with a few more (**EGFR, PIK3CA, KRAS, VHL, TP53, MED12, ESR1, PBRM1, BRAF, CTNNB1, AKT1)** is able to help in a good classification.

| Primary Cancer | No. of Samples Available |
|---|---|
| **Breast** | **7410** (**PIK3CA - 3990**, TP53 - 1863, EGFR - 42, VHL- 4) |
| **Kidney** | **1999** (**VHL - 1018**, TP53 - 116, PIK3CA - 31, EGFR - 13) |
| **Lung cancer** | **21362** (**EGFR - 12142**, TP53 - 1820, PIK3CA - 419, VHL - 8) |
| **Ovarian cancer** | **3072** (**TP53 - 1206**, PIK3CA - 388, EGFR - 23, VHL - 1) |

The following results were obtained when this model was tested against STRAND's test data

| | breast | kidney | lung | ovary |
|---|---|---|---|---|
| breast | 37 | 2 | 4 | 6 |
| kidney | 7 | 9 | 0 | 2 |
| lung | 36 | 1 | 39 | 18 |
| ovary | 7 | 1 | 0 | 3 |

**Every sample in the above STRAND's data has an average of 27 genes (out of 152) marked to be mutated.**

## COSMIC v68 Results using 54 Genes

54 Genes is the intersection between the 232 Genes from tumorTracer and Strand's 152 Gene list. This result is obtained by using only Copy Number Aberrations (A/D/N) as features and RF Model. OOB estimate of error rate **- 26.63%**

| | breast | endometrium | kidney | large.intestine | lung | ovary | Class Error |
|---|---|---|---|---|---|---|---|
| breast | 482 | 82 | 8 | 23 | 43 | 42 | 0.2911765 |
| endometrium | 38 | 119 | 0 | 10 | 8 | 22 | 0.3959391 |
| kidney | 15 | 1 | 204 | 11 | 7 | 2 | 0.1500000 |
| large.intestine | 18 | 23 | 2 | 319 | 13 | 4 | 0.1799486 |
| lung | 81 | 5 | 9 | 15 | 249 | 22 | 0.3464567 |
| ovary | 47 | 1 | 0 | 11 | 26 | 285 | 0.2297297 |

- Average number of **CNV** events per sample in this data set (for these classes) is **20.7 (out of these 54 genes)**

The following result was obtained on STRAND Test.

| | breast | lung | ovary |
|---|---|---|---|
| breast | 20 | 4 | 7 |
| endometrium | 26 | 12 | 7 |
| kidney | 0 | 0 | 0 |
| large.intestine | 2 | 1 | 0 |
| lung | 2 | 0 | 0 |
| ovary | 0 | 0 | 0 |

**Why again is the result not good?**

- Average number of **CNV** events per sample in this data set (for these classes) is **only 2.14**

In conclusion, there is nothing special about the 232 Gene list. We could get good results from STRAND's 152 Gene list if the following enhancements were made -

- ❖ Improve CNV Data collection (w.r.t variance of the CNV features).
- ❖ There is no large training data set for Somatic Features that are similar to the variance of STRAND's somatic data. (COSMIC has very less variance in its Somatic Features).

3. Comment on how large panels help with better copy number detection but detection of somatic variants might be compromised because coverage is only 50-100x. (Is this true?)
    - Looks to be true from our analysis.

4. Then a final conclusion whether 1(b), 2(b,c), 3(a,b,c) has potential or not - and if so which tissues it can be used for. what are the most significant genes that contribute to the model, whether just those genes are sufficient, whether the software outputs a confidence value or not and how to use it.
    - We should discuss this.