Variation Meeting

06/08/2016

Overview of PCAWG SNV dataset (initial release)

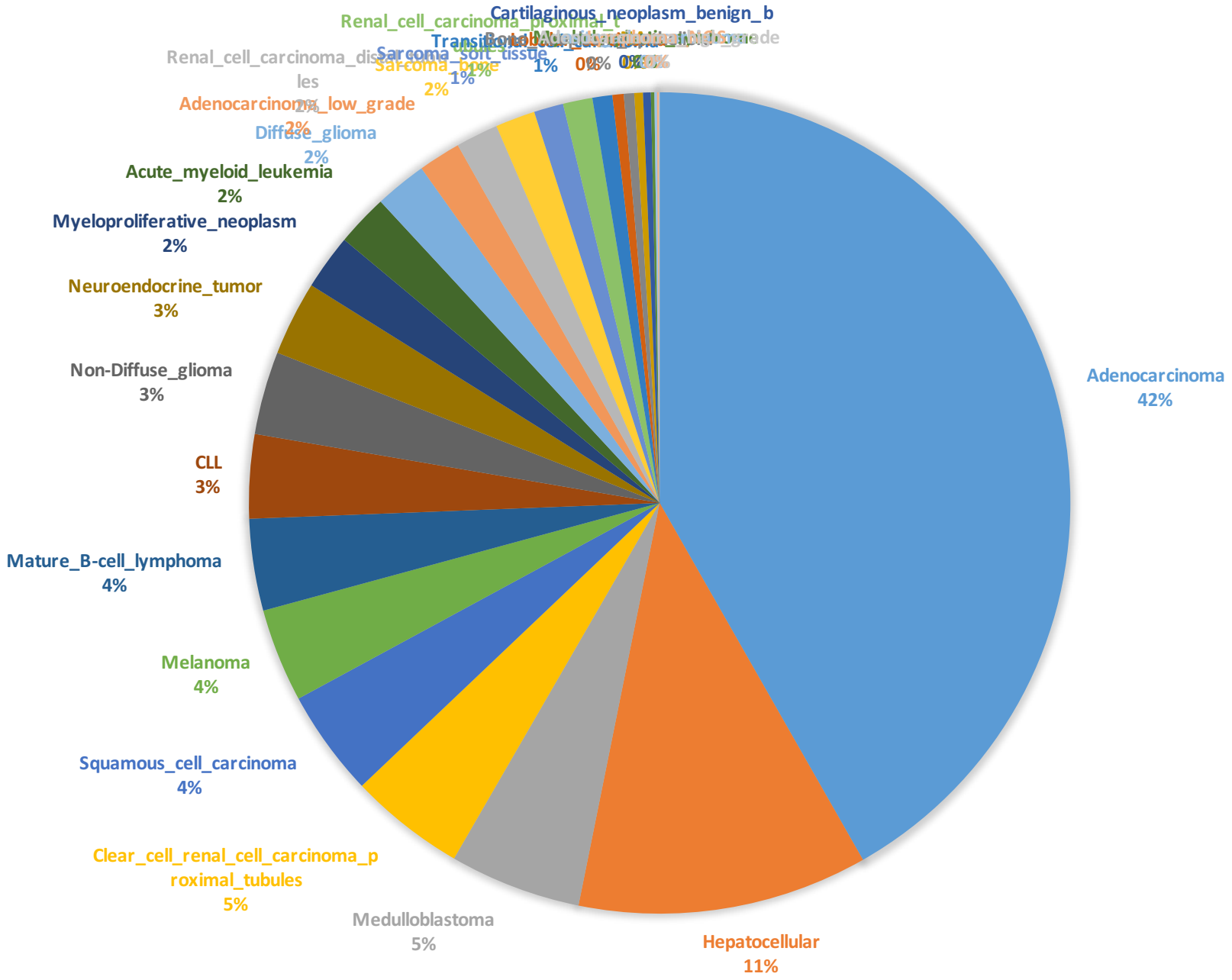
Overall 2834 Donors contributed 2961 tumor samples across multiple cancer types. Multiple samples contributed by few donors.

47 donors were black listed leading to overall 2913 tumor samples in the current release.

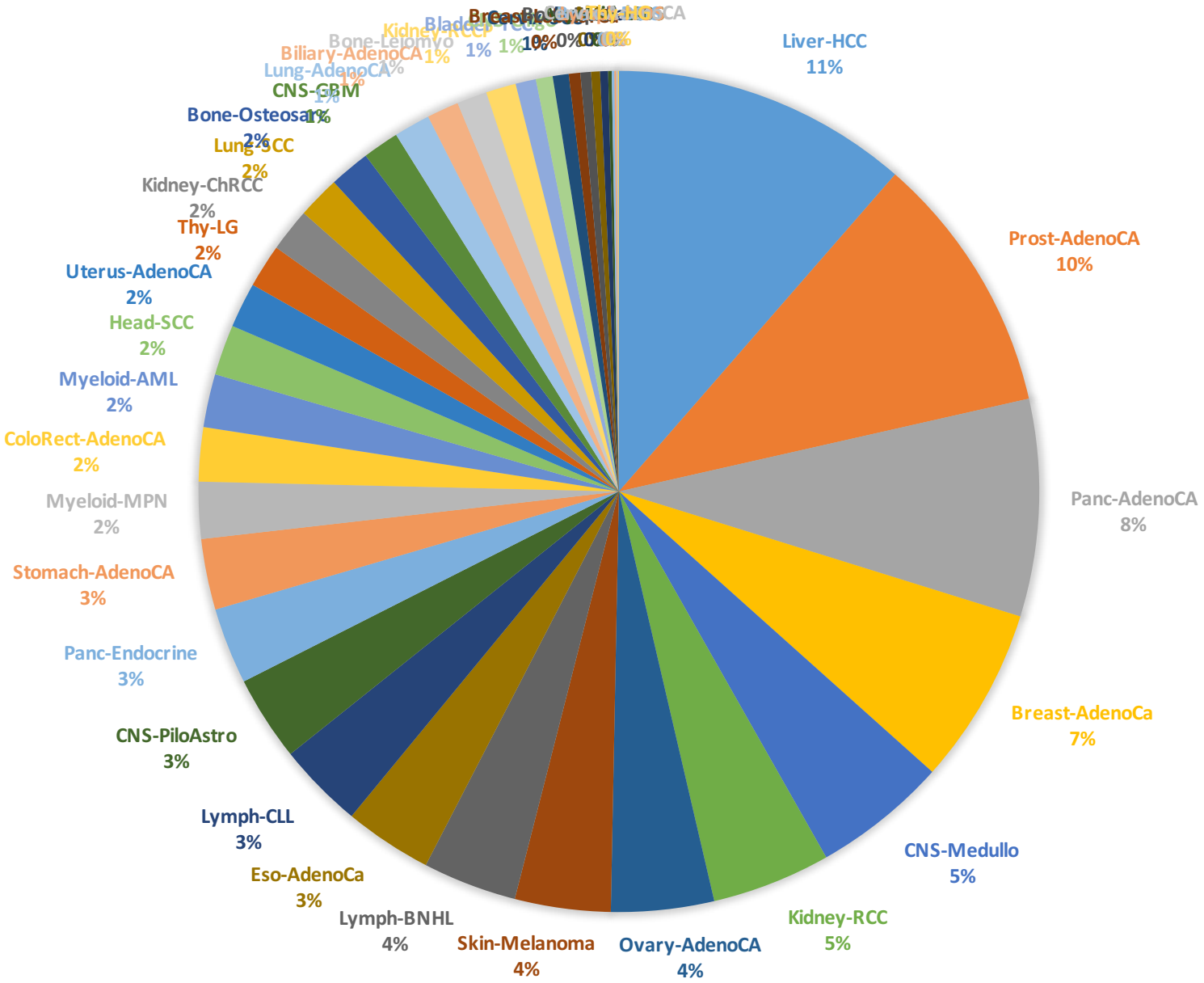
Initial release based on consensus strategy of “2+4” and were applied on OxoG-filtered SNVs.

Samples are classified based on histopathology subtypes rather than project code.

TUMOR TYPE DISTRIBUTION BASED ON SAMPLE FREQUENCY



PCAWG COHORT DISTRIBUTION BASED ON SAMPLE FREQUENCY



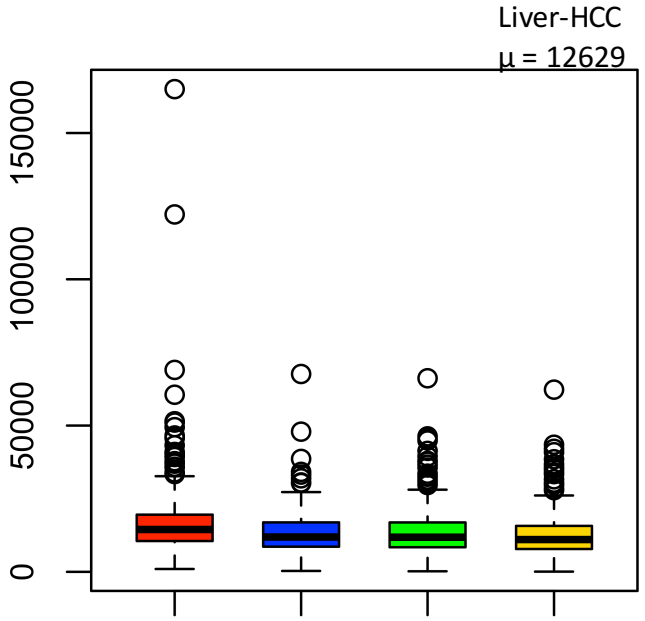
Filtering of potential false positive SNV calls

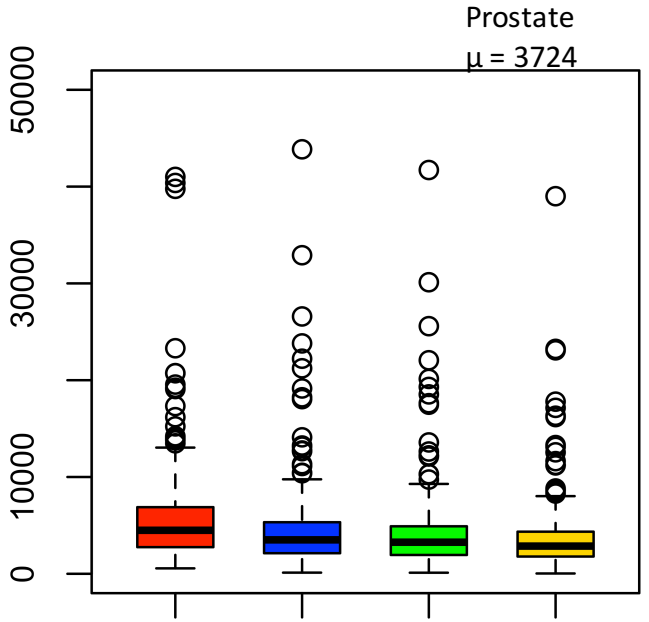
```
##fileformat=VCFv4.1
##INFO=<ID=NumCallers,Number=1,Type=Integer,Description="Number of callers that made this call">
##INFO=<ID=Callers,Number=.,Type=String,Description="Callers that made this call">
##INFO=<ID=1000genomes_AF,Number=A,Type=Float,Description="Thousand Genomes phase 3 occurance fraction if found: ALL.wgs.phase3_shapeit2_mvnc
all_integrated_v5b.20130502.sites.vcf.gz">
##INFO=<ID=1000genomes_ID,Number=1,Type=String,Description="Thousand Genomes phase 3 ID if found: ALL.wgs.phase3_shapeit2_mvncall_integrated_
v5b.20130502.sites.vcf.gz">
##INFO=<ID=VAF,Number=1,Type=Float,Description="VAF from mutect read filter if available">
##INFO=<ID=t_alt_count,Number=1,Type=Integer,Description="Tumour alt count from mutect read filter if available">
##INFO=<ID=t_ref_count,Number=1,Type=Integer,Description="Tumour alt count from mutect read filter if available">
##INFO=<ID=cosmic,Number=1,Type=String,Description="(first) cosmic ID if found, COSMICv76">
##INFO=<ID=dbsnp,Number=1,Type=String,Description="(first) dbSNP ID if found, build 147, All_20160408.vcf.gz">
##INFO=<ID=repeat_masker,Number=1,Type=String,Description="Repeat masker region if in one">
##FILTER=<ID=LOWSUPPORT,Description="Not called by enough callers in ensemble">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
1       1064990 .       G       A       255.0   .       Callers=broad,dkfz,muse,sanger;NumCallers=4;cosmic=COSN8884854;VAF=0.4231;t_alt_count
=11;t_ref_count=15
1       1497491 .       T       C       255.0   .       Callers=broad,muse;NumCallers=2;VAF=0.1429;t_alt_count=4;t_ref_count=24
1       2139875 .       G       C       255.0   .       Callers=broad,dkfz,muse,sanger;NumCallers=4;cosmic=COSN8432367;VAF=0.3214;t_alt_count
=9;t_ref_count=19
1       2938955 .       G       A       255.0   .       Callers=broad,dkfz,muse,sanger;NumCallers=4;dbsnp=rs753013139;cosmic=COSM908025;VAF=0
.4762;t_alt_count=10;t_ref_count=11
1       3548542 .       T       C       255.0   LOWSUPPORT      Callers=dkfz;NumCallers=1;dbsnp=rs2244942;cosmic=COSN19634597;1000genomes_ID=
rs2244942;1000genomes_AF=0.4347;VAF=0.2105;t_alt_count=4;t_ref_count=15
1       4090567 .       A       T       255.0   LOWSUPPORT      Callers=broad;NumCallers=1;repeat_masker=AluSq2;VAF=0.2381;t_alt_count=5;t_re
f_count=16
```

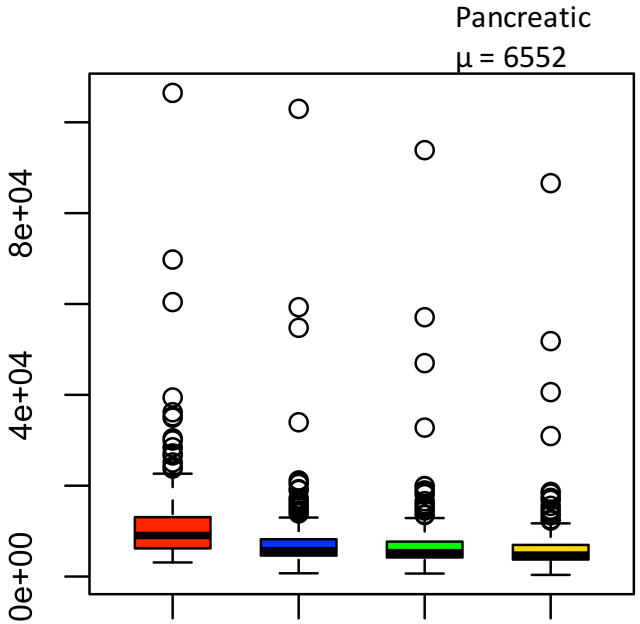
Filter1: remove SNVs marked as present in the 1000 genome/DBSNPs/Low support

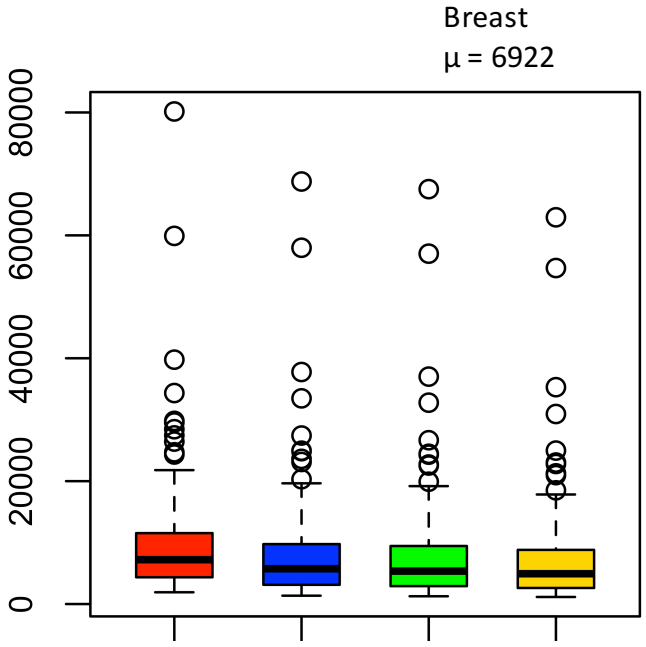
Filter2 : Apply the 1000 Genome Mask (only accounting for SNVs present in highly mapable region of the genome)

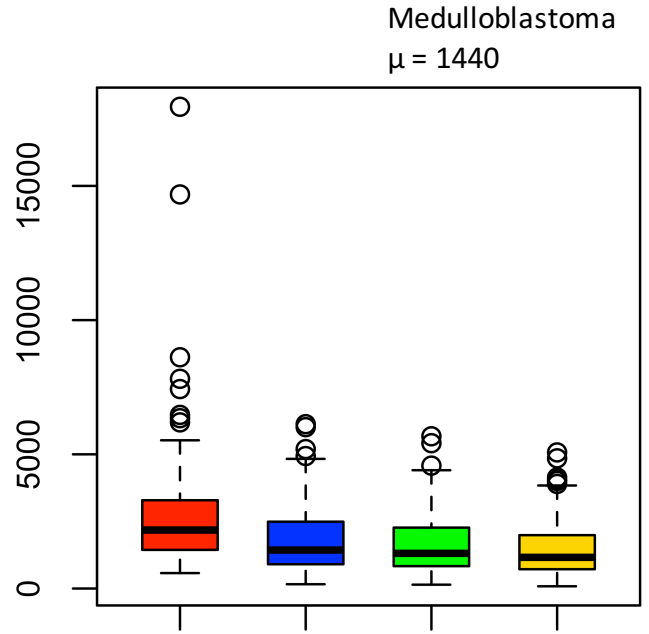
Filter3: Applied Heng Li's Mask of ignoring low complexity region of the genome (approx. 2% of the genome)

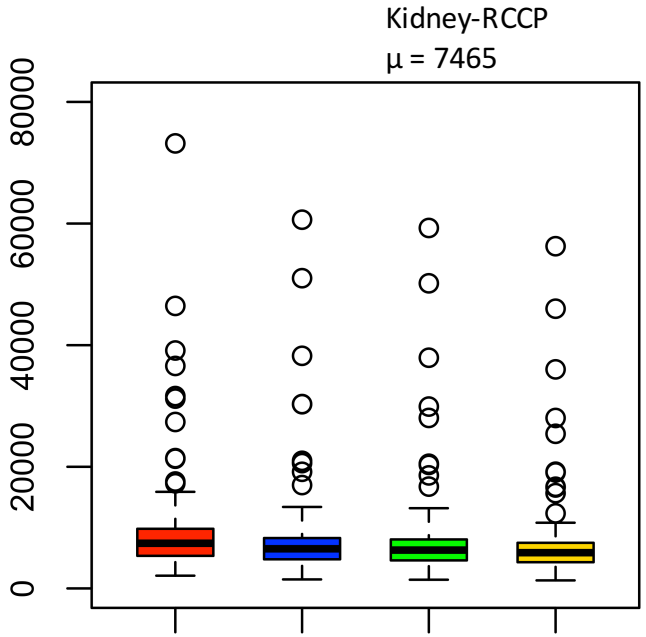

Liver-HCC
μ = 12629
0
50000
100000
150000

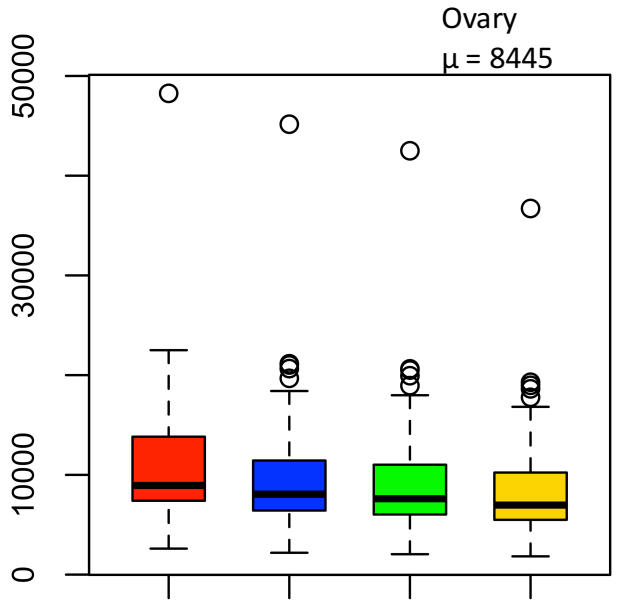

Prostate
μ = 3724
0
10000
30000
50000

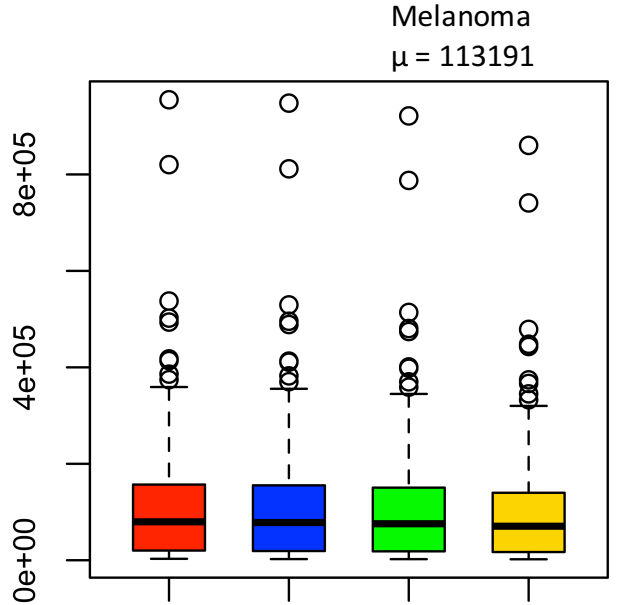

Pancreatic
μ = 6552
0e+00
4e+04
8e+04

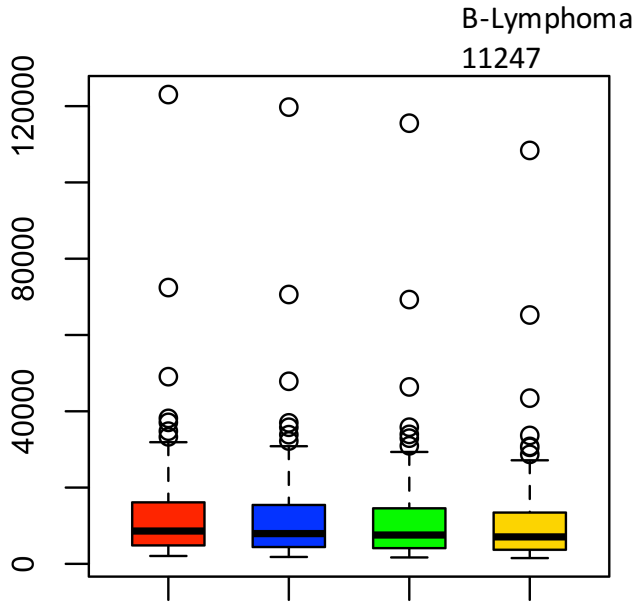

Breast
μ = 6922
0
20000
40000
60000
80000

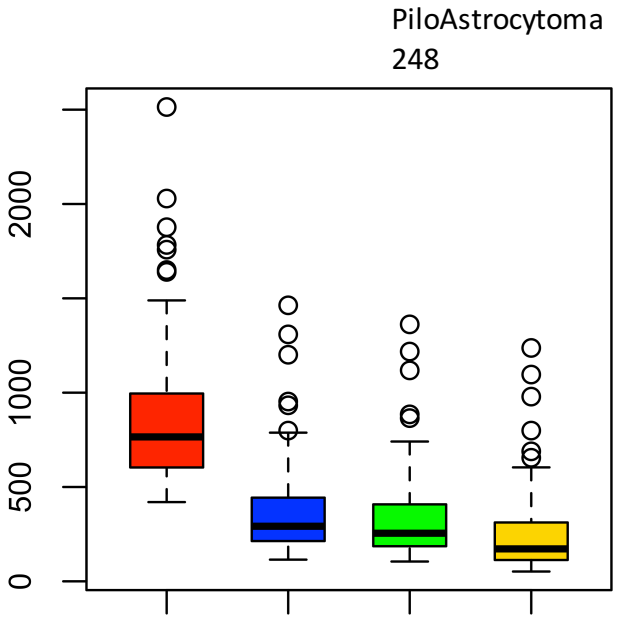

Medulloblastoma
μ = 1440
0
5000
10000
15000

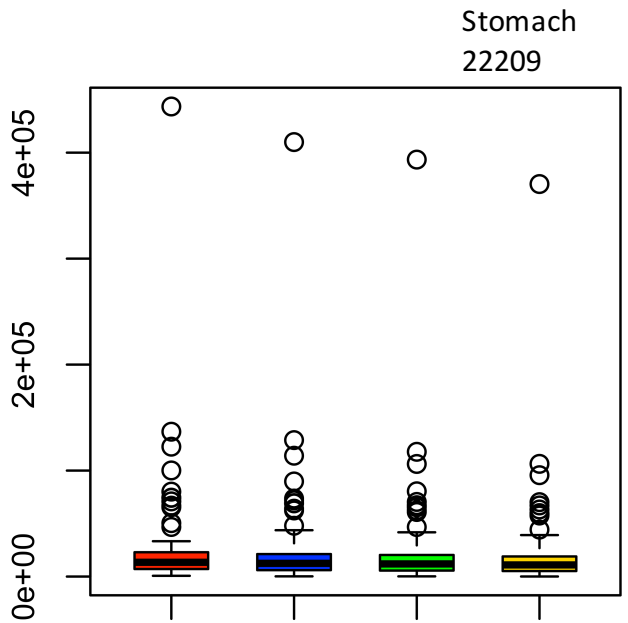

Kidney-RCCP
μ = 7465
0
20000
40000
60000
80000

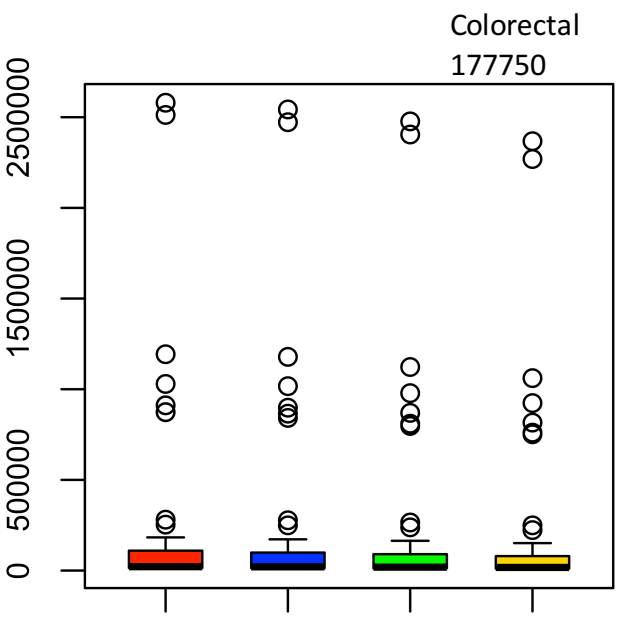

Ovary
μ = 8445
0
10000
30000
50000

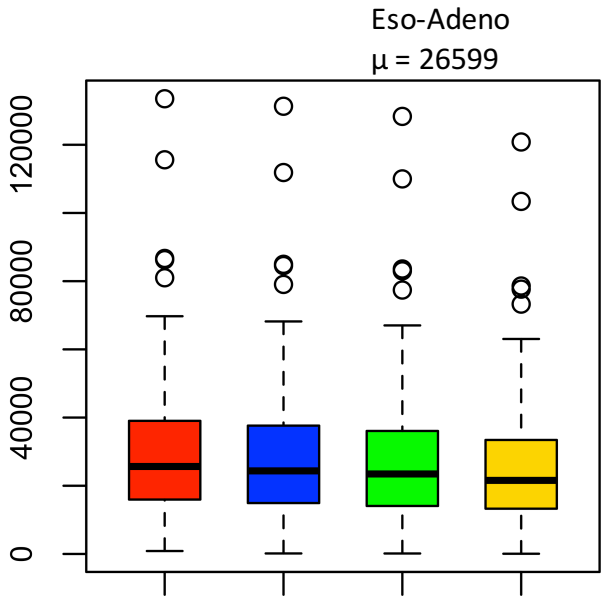

Melanoma
μ = 113191
0e+00
4e+05
8e+05

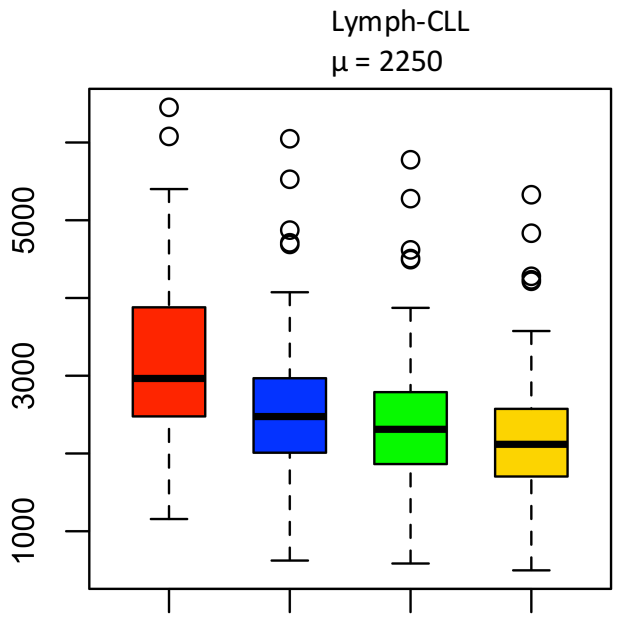

B-Lymphoma
11247
0
40000
80000
120000

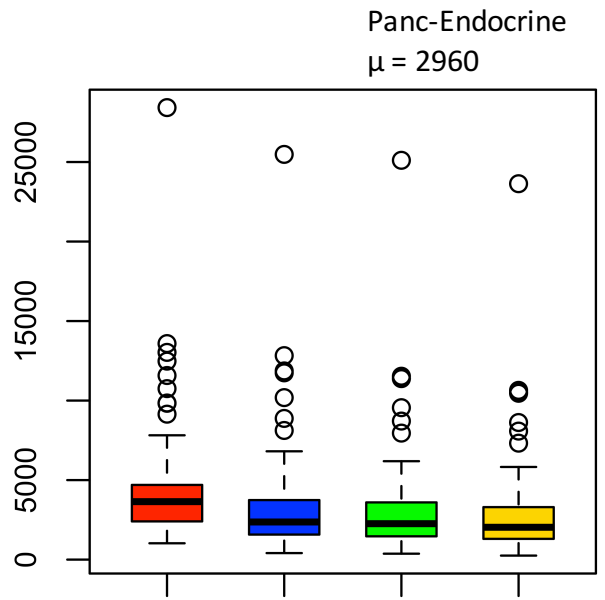

PiloAstrocytoma
248
0
500
1000
2000

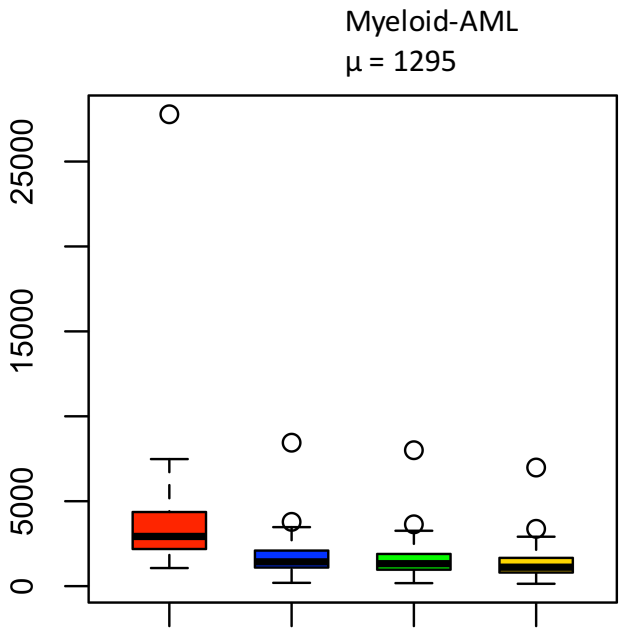

Stomach
22209
0e+00
2e+05
4e+05

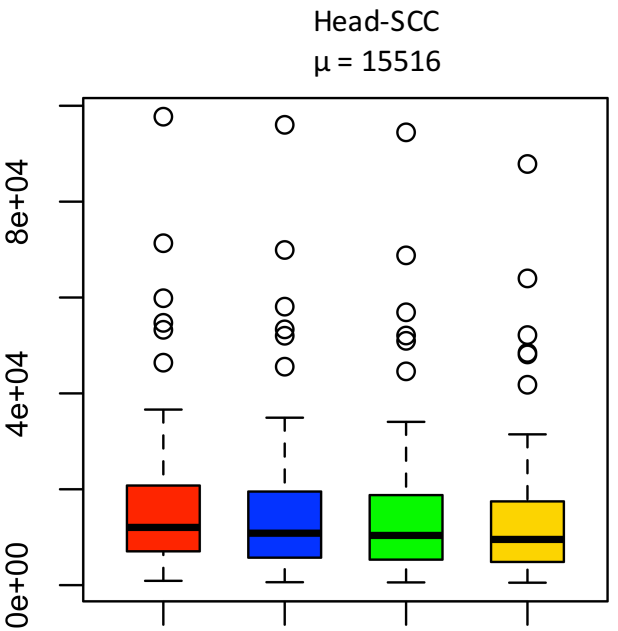

Colorectal
177750
0
500000
1500000
2500000

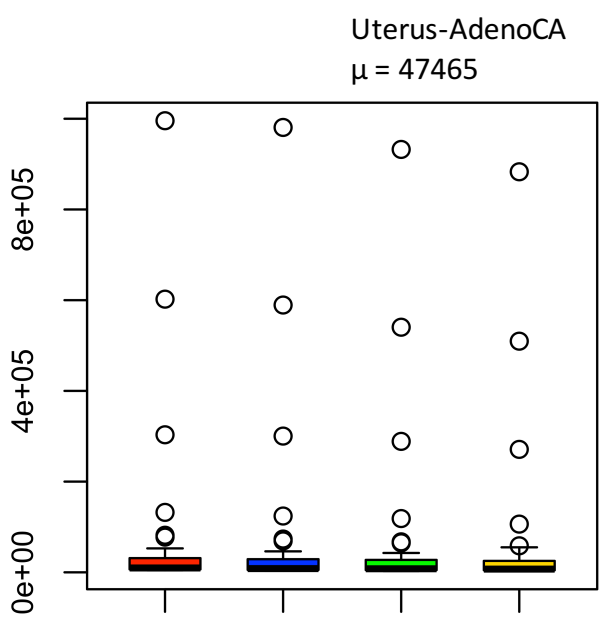

Eso-Adeno
μ = 26599
0
40000
80000
120000


Lymph-CLL
μ = 2250
1000
3000
5000

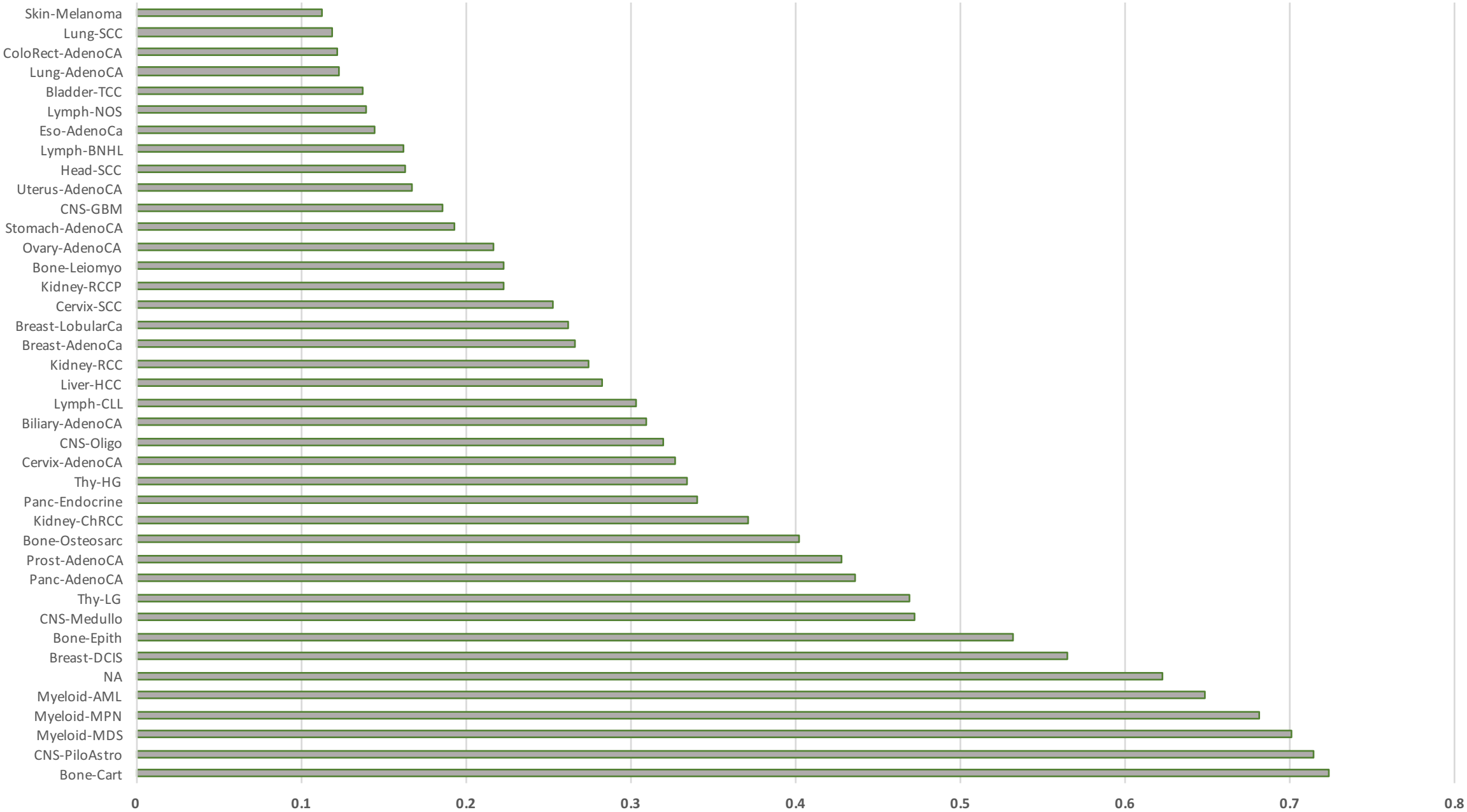

Panc-Endocrine
μ = 2960
0
5000
15000
25000


Myeloid-AML
μ = 1295
0
5000
15000
25000

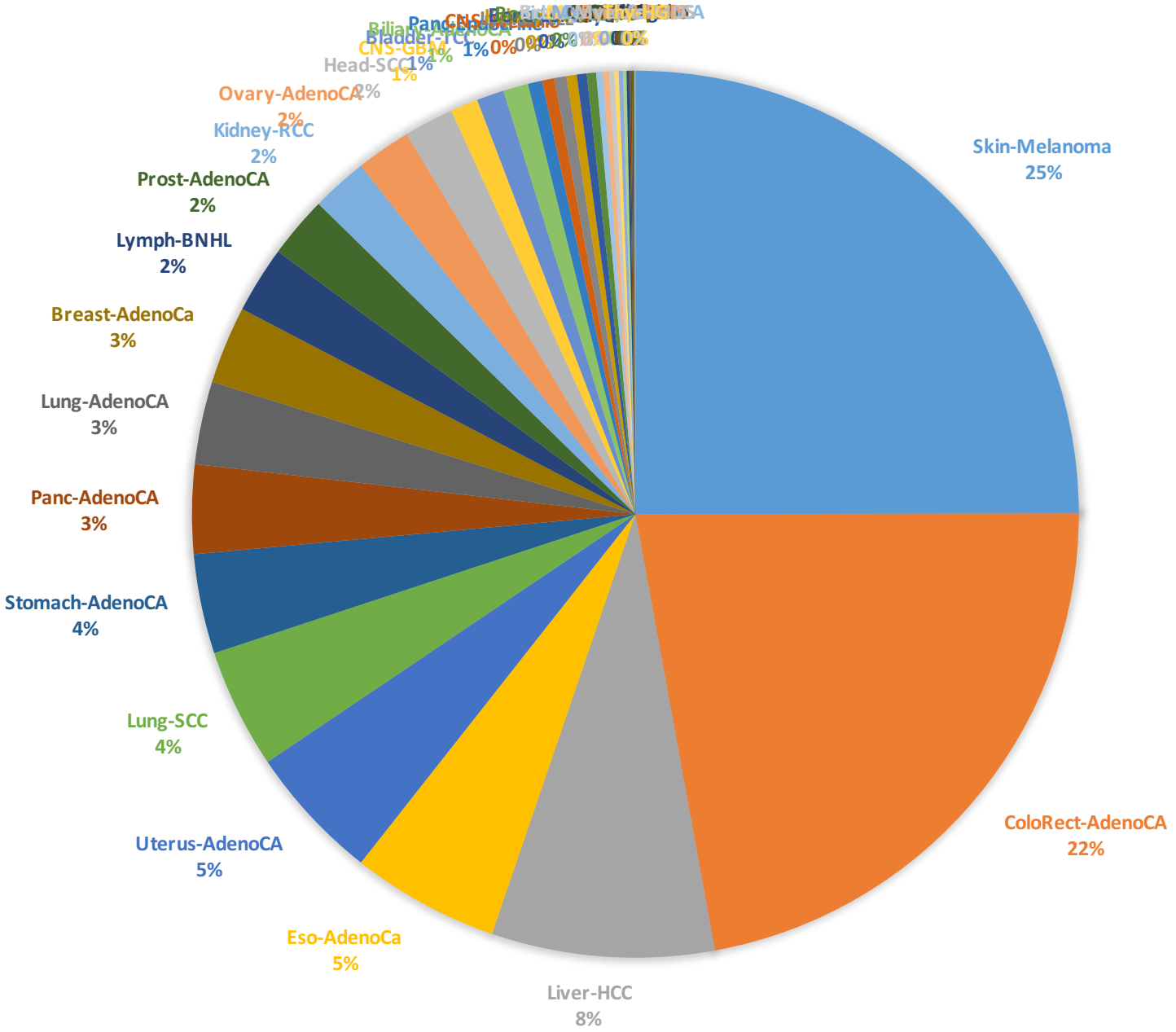

Head-SCC
μ = 15516
0e+00
4e+04
8e+04


Uterus-AdenoCA
μ = 47465
0e+00
4e+05
8e+05

Fraction of SNPs removed upon applying filtering strategy



TOTAL SNP CONTRBUTIONS OF DIFFERENT PCAWG COHORTS



PER SAMPLE SNP CONTRIBUTIONS OF DIFFERENT PCAWG COHORTS