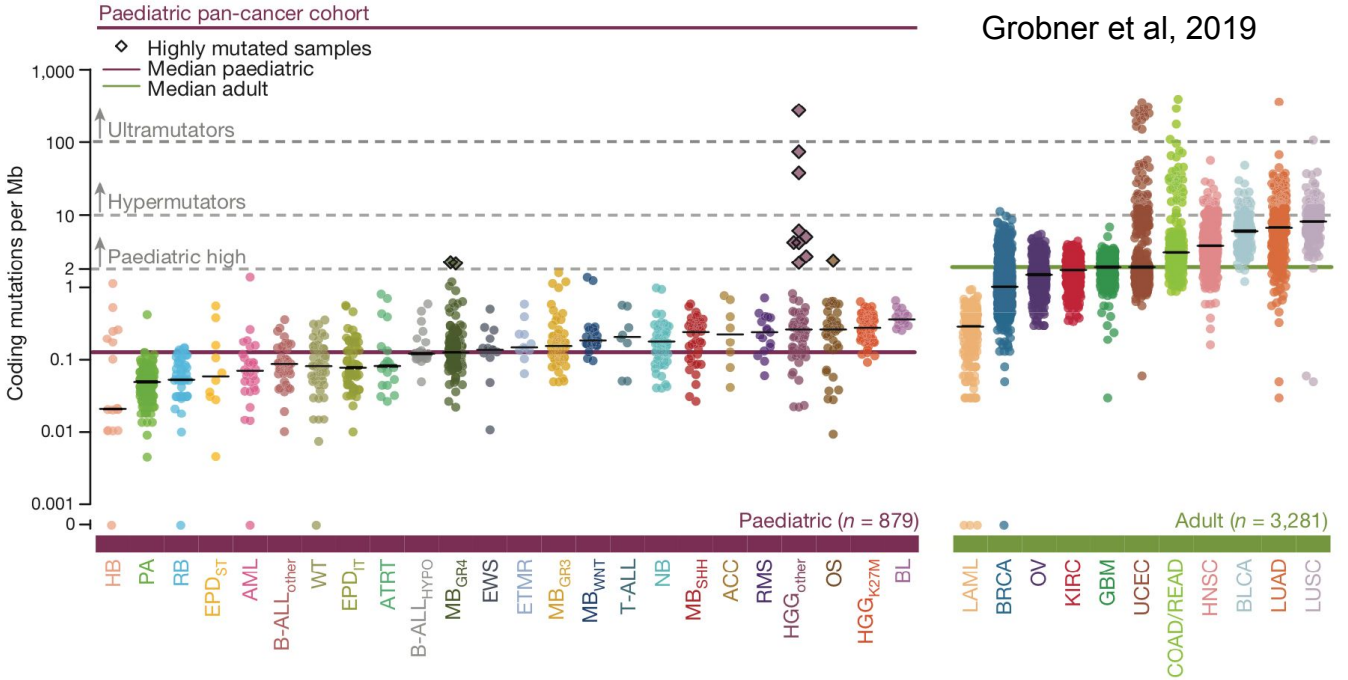
TCGA vs PBTA consensus calls

Candace Savonen

Background on this issue

https://github.com/AlexsLemonade/OpenPBTA-analysis/issues/257

Grobner et al, 2019



Methods summary: "In-house pipeline" using Samtools MuSiC-bmr

Molecular cancer types in paediatric pan-cancer cohort

- Hepatoblastoma (HB) (n = 16)
- Pilocytic astrocytoma (PA) (n = 105)
- Retinoblastoma (RB) (n = 36)
- Ependymoma supratentorial (EPD_{ST}) (n = 15)
- Acute myeloid leukaemia (AML) (n = 30)
- B-cell acute lymphoblastic leukaemia, non-hypodiploid ($B\text{-}ALL_{other}$) (n = 61)
- Wilms tumour (WT) (n = 51)
- Ependymoma infratentorial (EPD_{IT}) (n = 55)
- ATRT (n = 19)
- B-cell acute lymphoblastic leukaemia, hypodiploid ($B\text{-}ALL_{HYPO}$) (n = 20)
- Medulloblastoma Group 4 (MB_{GR4}) (n = 107)
- Ewing's sarcoma (EWS) (n = 24)
- ETMR (ETMR) (n = 11)
- Medulloblastoma Group 3 (MB_{GR3}) (n = 60)
- Medulloblastoma WNT (MB_{WNT}) (n = 21)
- T-cell acute lymphoblastic leukaemia (T-ALL) (n = 19)
- Neuroblastoma (NB) (n = 59)
- Medulloblastoma SHH (MB_{SHH}) (n = 42)
- Adrenocortical carcinoma (ACC) (n = 8)
- Rhabdomyosarcoma (RMS) (n = 21)
- High-grade glioma K27wt (HGG_{other}) (n = 67)
- Osteosarcoma (OS) (n = 42)
- High-grade glioma K27M (HGG_{K27M}) (n = 57)
- Burkitt's lymphoma (BL) (n = 15)

Adult cancer types (TCGA)

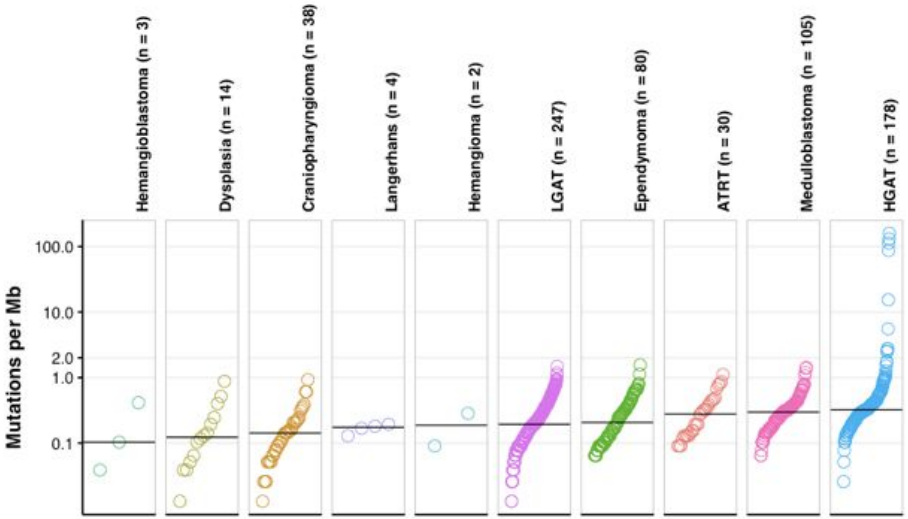
- Acute myeloid leukaemia (LAML)
- Breast adenocarcinoma (BRCA)
- Ovarian serous carcinoma (OV)
- Kidney renal clear cell carcinoma (KIRC)
- Glioblastoma (GBM)
- Uterine corpus endometrial carcinoma (UCEC)
- Colon/rectal carcinoma (COAD/READ)
- Head and neck squamous carcinoma (HNSC)
- Bladder urothelial carcinoma (BLCA)
- Lung adenocarcinoma (LUAD)
- Lung squamous cell carcinoma (LUSC)

Methods summary: Kandoth et al, 2013 "TCGA publicly available MAF files" - so probably Mutect1?

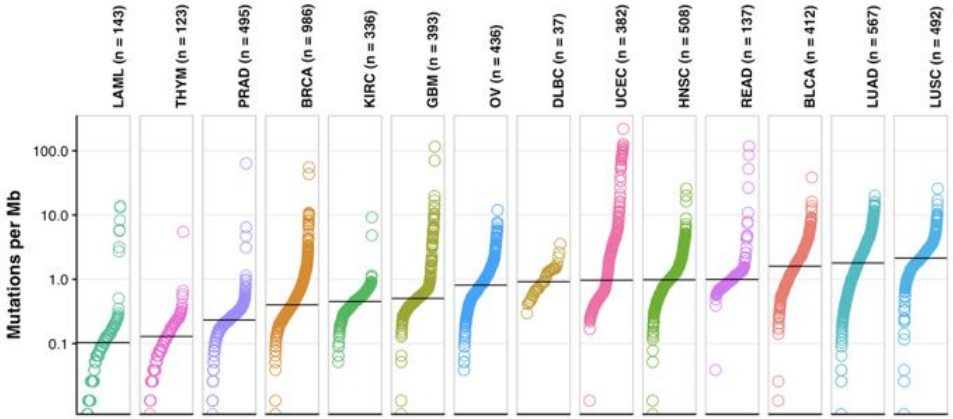
*Both have a coverage filter of > 6 - 8 reads

From @tkoganti

PBTA (Mutect2)



TCGA (Mutect1)



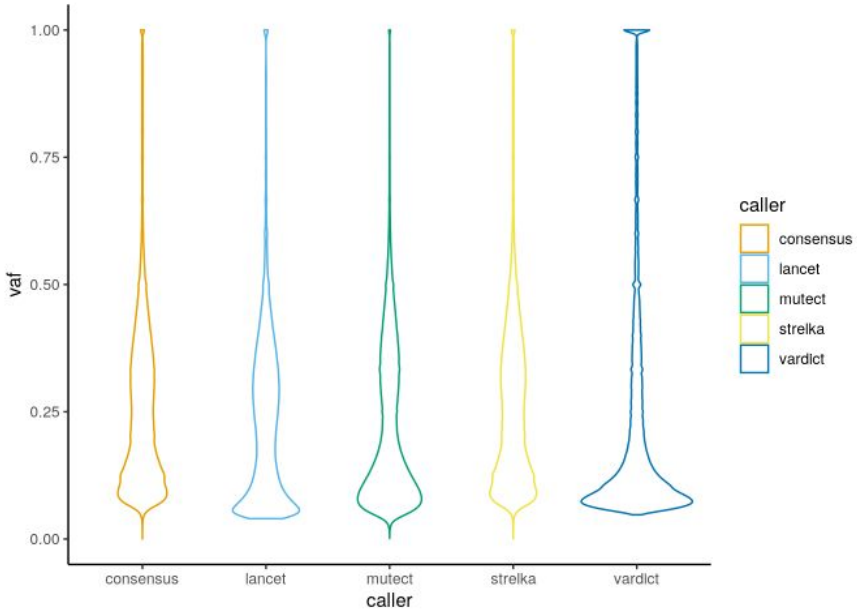

```
TMB per Mb = (# of missense + # of nonsense)*1000000/
Size of exome BED
```

Code here: https://github.com/d3b-center/scripts-/blob/master/TMB_calculation_from_MAFfiles

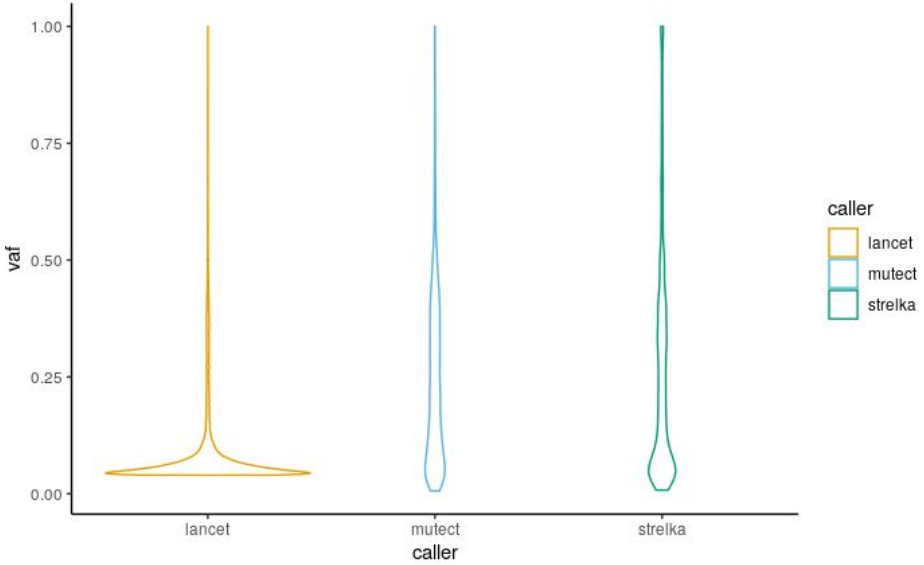
TCGA data ran through the PBTA consensus pipeline

https://github.com/AlexsLemonade/OpenPBTA-analysis/pull/521

PBTA

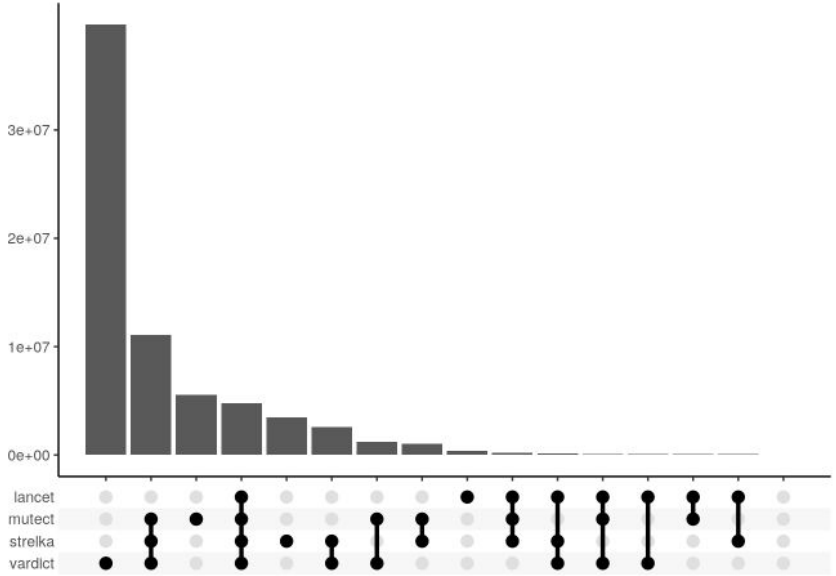


TCGA

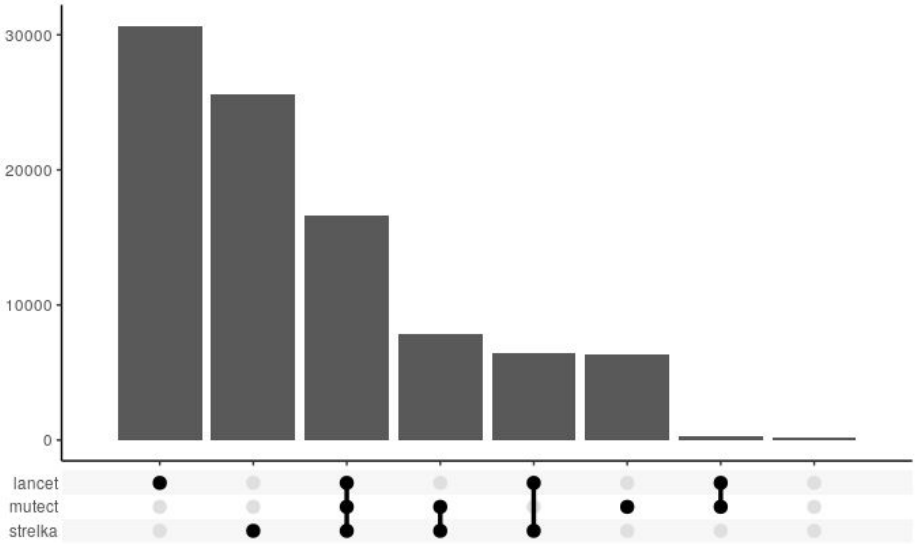


Lancet has unusually low VAF for TCGA data

PBTA



TCGA



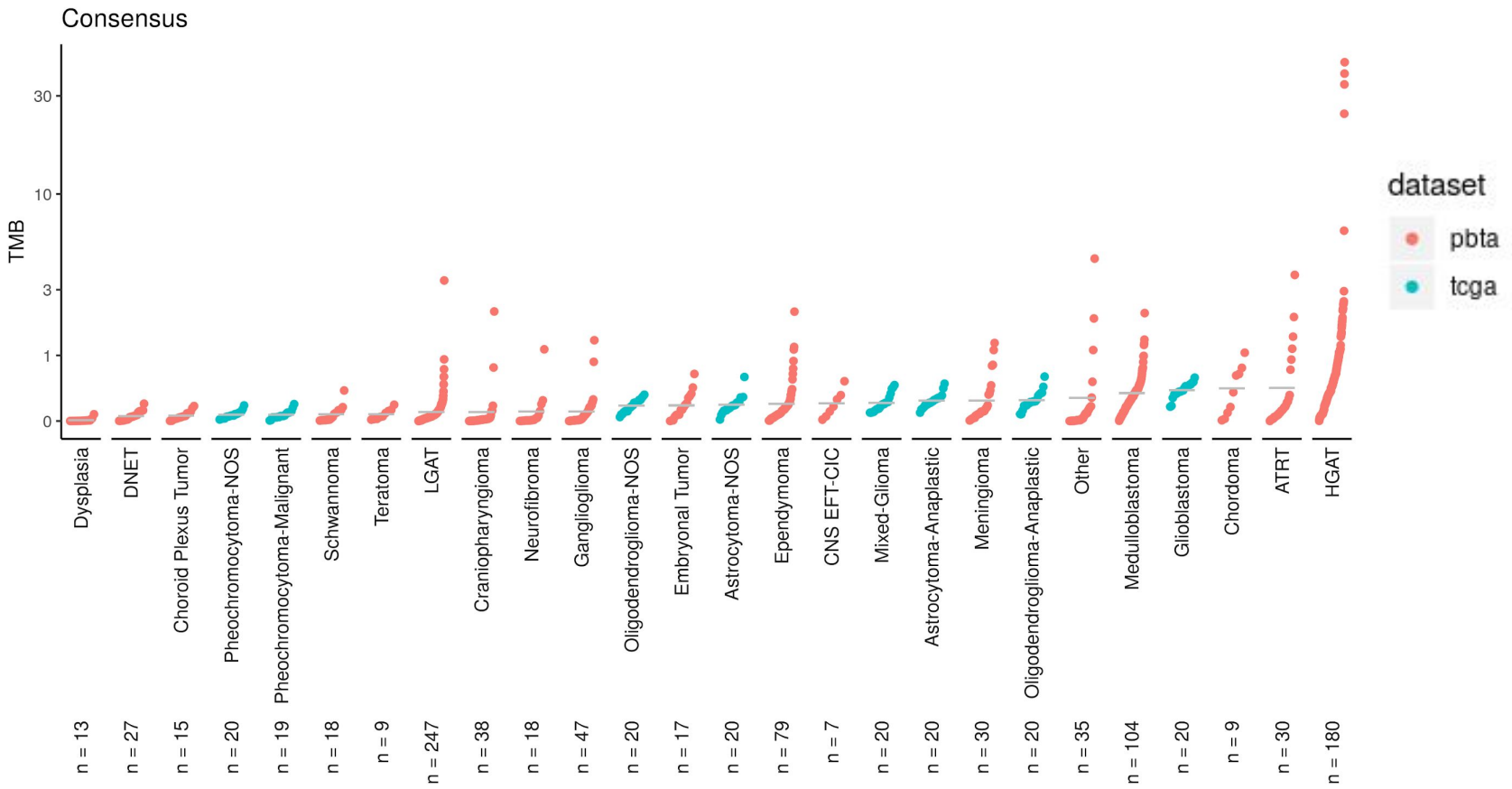
Lancet has unusually high number of mutations called only by it in TCGA

TMB Calculations overview

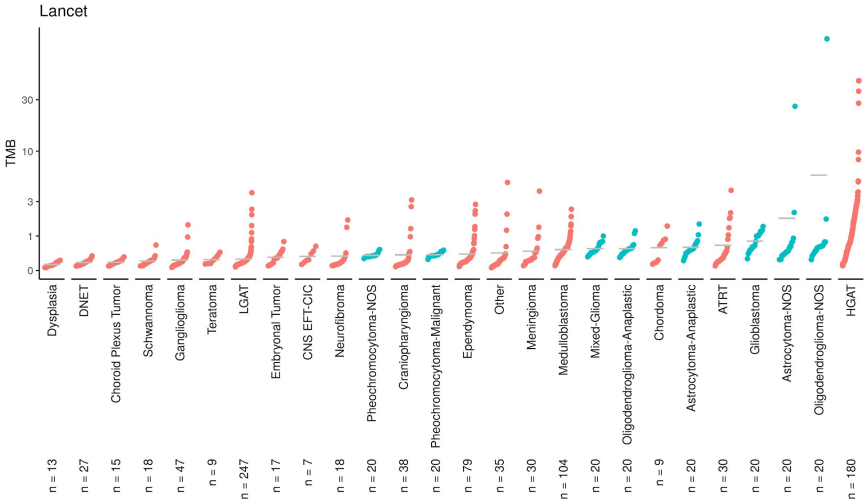
WGS_coding_only_TMB =
(total # coding sequence snvs called by all three of Strelka, Lancet, and Mutect2) /
intersection_strelka_lancet_mutect_CDS_genome_size

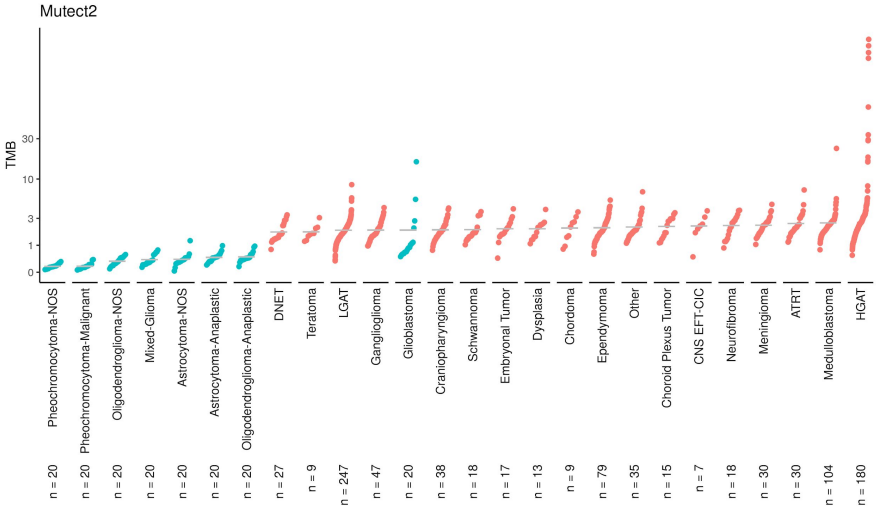
WXS_coding_only_TMB =
(total # coding sequence snvs called by all three of Strelka, Lancet, and Mutect2) /
intersection_wxs_CDS_genome_size

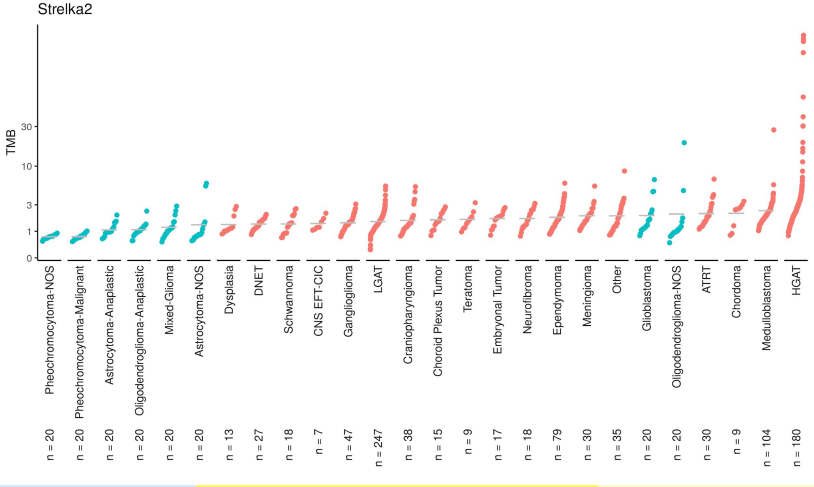
TMB calculated based on consensus of Lancet, Strelka2, Mutect2



Do coding TMB comparisons change with different callers?

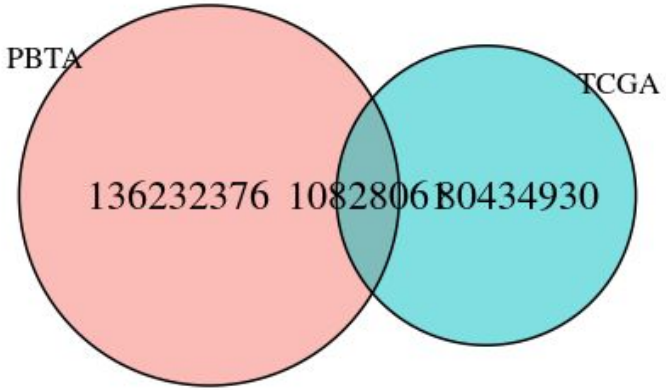






TCGA’s WXS data

WXS Target Region Differences between TCGA and PBTA



Ratio of PBTA overlapped: 0.074

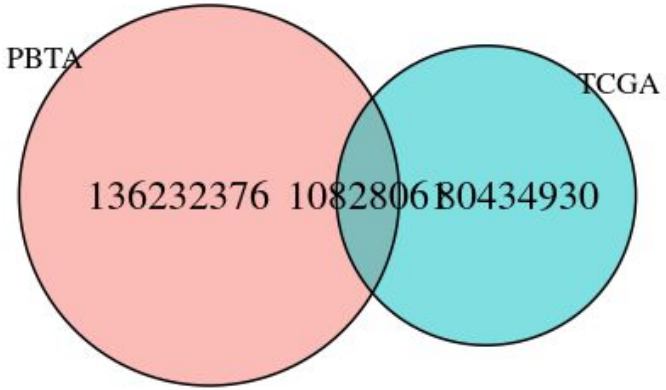
Ratio of TCGA overlapped: 0.119

TCGA Target BED file from MC3:

https://github.com/AlexsLemonade/OpenPBTA-analysis/pull/521#issuecomment-583053607

Base pair overlap using GenomicRanges::intersect

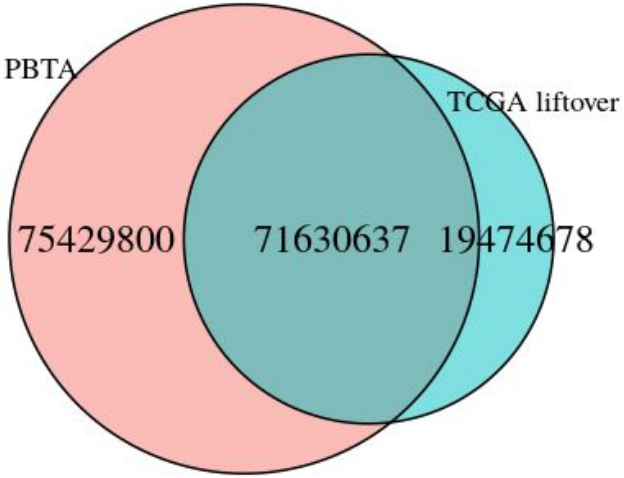
WXS Target Region Differences between TCGA and PBTA



Ratio of PBTA overlapped: 0.074

Ratio of TCGA overlapped: 0.119

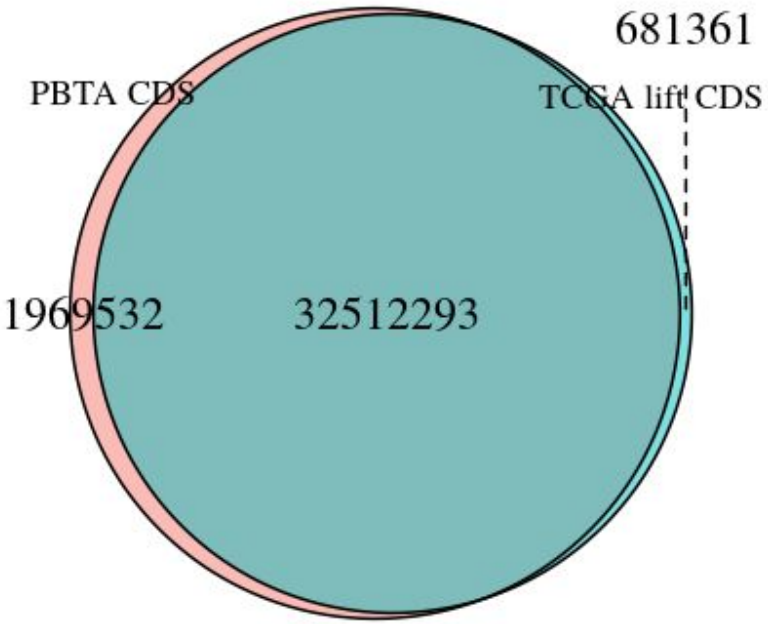
Liftover of the Target BED for TCGA using:
https://genome.ucsc.edu/cgi-bin/hgLiftOver



Ratio of PBTA overlapped: 0.487

Ratio of TCGA liftover overlapped: 0.786

What about the coding regions? - what we use for TMB?




```
Ratio of PBTA CDS overlapped: 0.943
Ratio of TCGA lift CDS overlapped: 0.979
```

Summary:

- Do we need a bigger *n* for TCGA data?
- Lancet's local assembly may lead to problems for TCGA.
- WXS Target regions: how are they incorporated into the calls?
- Immune activated diseases have higher TMBs (Chalmers et al, 2017)