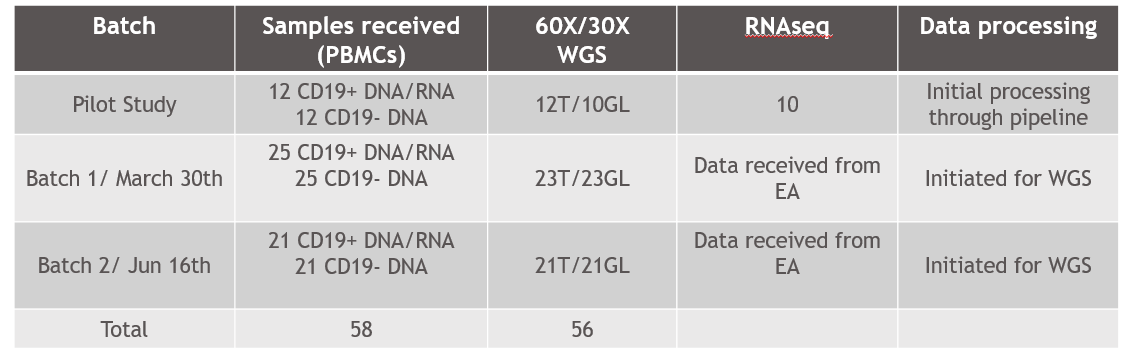
CLL del17p (BMS – Preeti)

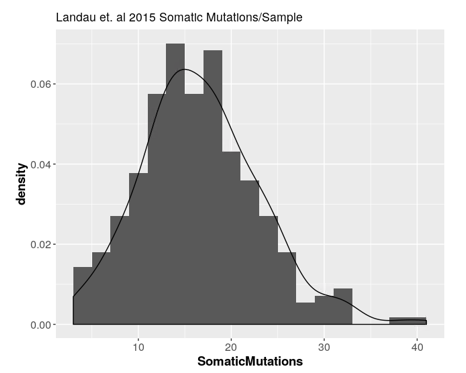
November 30, 2021

# Brief Overview

BMS, in collaboration with the Mayo Clinic, has collected 58 Chronic Lymphocyte Leukemia tumor samples. They have been whole genome sequenced (60x tumor 30x normal) and RNA sequenced. This cohort is all positive for a high-risk prognostic mutation: a large deletion of chr17p, which includes TP53.



The goals of this project are to:

1. Validate the chr17p deletion status in this cohort
2. Identify TP53 mutation status of each sample (deletion/unmutated, deletion/silent, deletion/missense, deletion/nonsense)
3. Identify the spectrum of small somatic mutations in this cohort
   1. Expectation for general CLL: 16 mutations (median)
   2. 
4. Identify the spectrum of large CNVs (in addition to chr17p) in this cohort
5. Assess whether TP53 mutation status, burden of mutation in known CLL driver genes, or any large copy number changes, associate with overall survival from treatment

# Location of data

All data is currently located on the BMS server at 172.25.130.233. The data can be transferred, or access to this server can be granted. This data consists of:

1. 44 dbnsfp-annotated tumor/normal vcf files of the form
   1. <sid>-T-<sid>-N.vcf.gz
2. 39 dbnsfp-annotated tumor-only vcf files of the form
   1. <sid>-T-null.vcf.gz
   2. There is a 39-sample overlap between (1) and (2)
3. 44 battenberg CNV calls of the form
   1. <sid>-T\_subclones.txt
   2. There is a 44-sample overlap between (1) and (3)
4. 44 manta indel calls of the form
   1. <sid>-T/results/variants/candidateSmallIndels.vcf.gz
   2. There is a 44-sample overlap between (1) and (4)
5. Clinical information for 58 samples, with IDs matching in the <patient> column
   1. There is a 44-sample overlap between (1) and (5)
6. 10 dbnsfp-annotated cd19+/cd19- vcf files of the form
   1. <sid>\_CD19-<sid>\_CD19-.vcf.gz
   2. There is a 10-sample overlap between (5) and (6)
   3. There is a 0-sample overlap between (1) and (5)
7. 9 battenberg CNV calls of the form
   1. <sid>\_CD19\_subclones.txt
   2. There is a 9-sample overlap between (5) and (6) [2596 is missing]
   3. There is a 0-sample overlap between (1) and (5)
8. 10 manta indel calls of the form
   1. <sid>\_CD19/results/variants/candidateSmallIndels.vcf.gz
   2. There is a 10-sample overlap between (5) and (6)
   3. There is a 0-sample overlap between (1) and (5)

# Required analysis

Initial QC

* Ensure Battenberg calls for 17p are all deletions
* Histogram of somatic mutations – are we in line with ~6-25 mutations per sample
* Ensure Battenberg CNV calls are consistent with FISH data
  + chr6, chr11q, chr12tri, chr13q, chr14 – fish data all given as binary for “some abnormality” as z\_fish6, z\_fish11q, z\_fish12tri, z\_fish13q, and z\_fish14
* Remove samples with high contamination of normal (1696, possibly 3053+1823)
* Remove samples with estimated ploidy > 4 or estimated cellularity < 0.5 (Battenberg)
* Tabulate cohort clinical demographics (sex, age, treatment, fish abnormalities, OS) after filtering

CNV analysis

* Square off Battenberg calls into (sample x CNV) matrices and annotate with chromosomal band
* Produce a GISTIC plot of DUP/DEL prevalence for each chromosome
* Perform longitudinal test for overall survival (use survminer) versus duplication or deletion status

TP53 analysis

* Identify the “status” of TP53 in each sample (copy number + missense or nonsense burden)
* Ensure consistency with “P53\_result” column in metadata (abnormal corresponding to some kind of mutational burden, or weird copy number)
* Perform longitudinal test for overall survival versus TP53 status

Somatic mutation analysis

* Merge somatic mutation VCFs into mutational counts and tumor allele frequency by gene (I.e., for sample 3053, FAM160B harbors a missense mutation at TAF=83%)
  + Multiple mutations – take most severe, then most frequent
* Produce plot of gene mutational prevalence and clonality for the top-most mutated genes in del17p cohort
* Perform longitudinal test for overall survival versus each gene burden of high-impact variation
  + If specific mutations appear to recur, you can test these as well