Genomic data in MSK-Chord

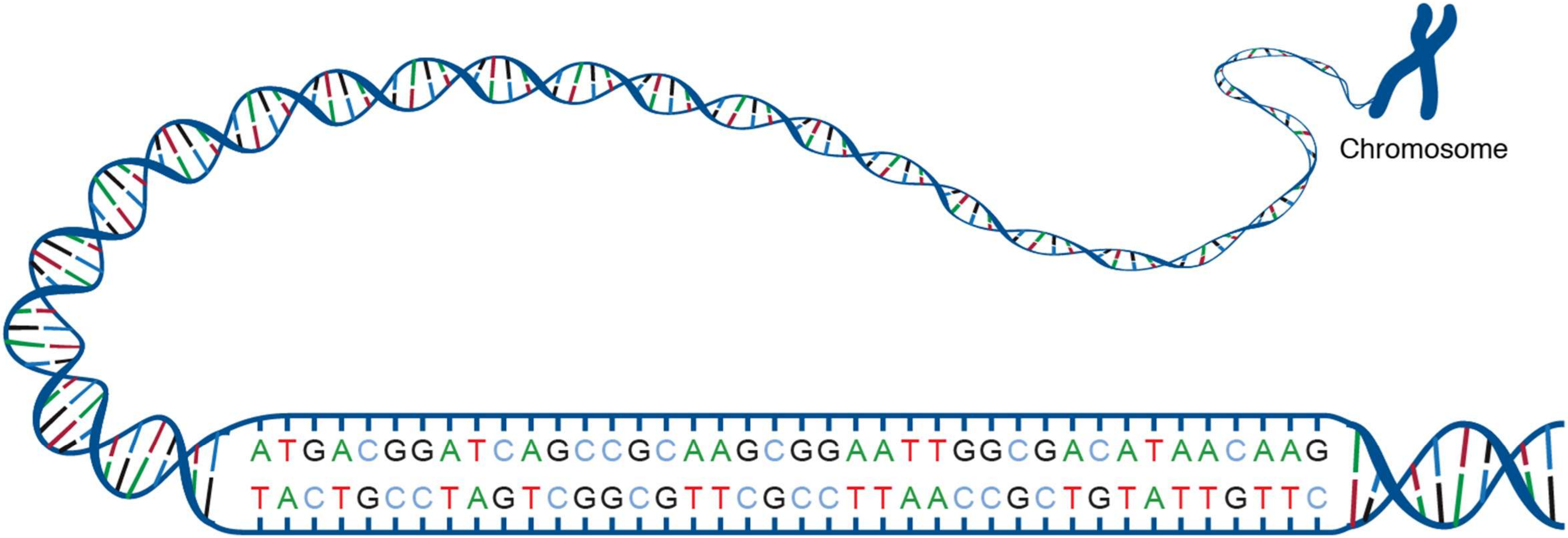
Goals today

- Sequencing: reading DNA*

- Copy-number variation
- Somatic mutations
- Structural variants
- Nuts and bolts: which genes were tested?

* from a biostatistician, for biostatisticians

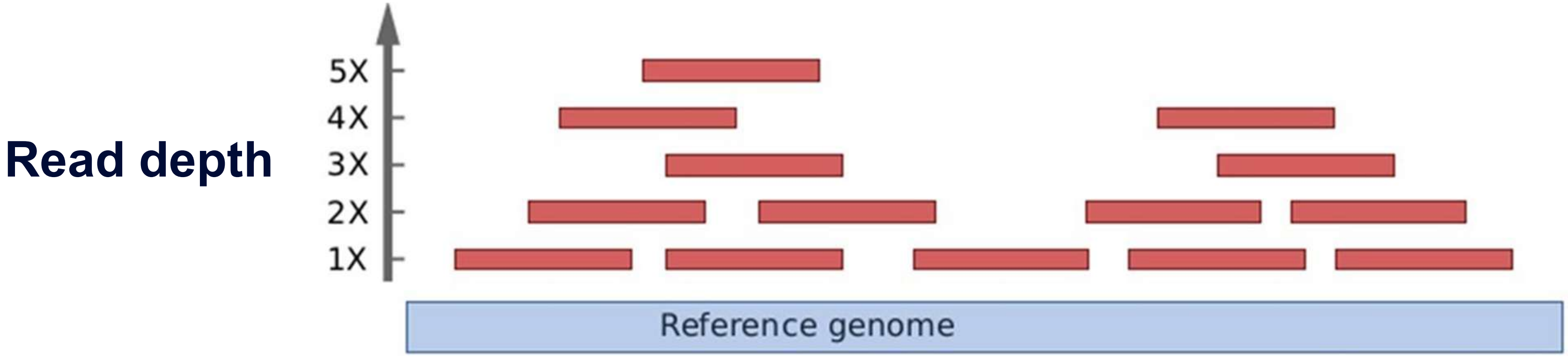
DNA


Chromosome
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATTGTTC

DNA is a long string of nucleotides: (A, C, G, T).
Our genome has 3B nucleotides.

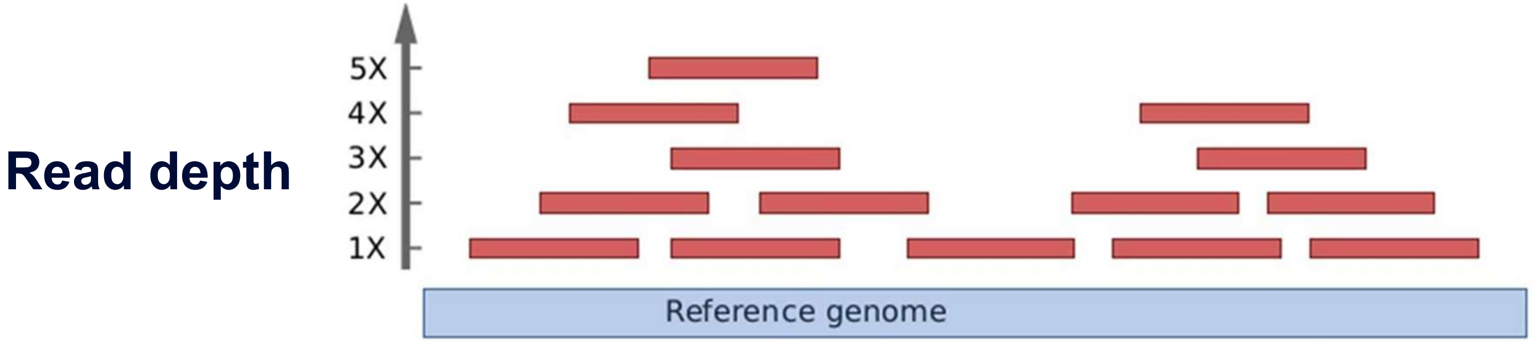
Sequencing

- We want to read (sequence) all 3B nucleotides in our genome.
- We have no way to do this in one pass - we **can** read short sequences. (~150 nucleotides at a time using Illumina, as in MSK-Chord.)
- We amplify DNA (cut into short fragments and copy), sequence each read (AGTCAAA...), and align the reads to a reference genome.



Sequencing

- We amplify DNA (make copies) and chop them into short strings. These are read (AGTCAAA...) and aligned to a reference genome:



- Once aligned, we know where in the genome each read came from. We can detect mutations, count copies, and find rearrangements.

- Now: reading the whole genome is expensive. Instead, MSK-Chord targets a panel of 500 cancer-related genes.

Some thoughts

- Sequencing is crucial for biomedical research and ubiquitous.
  A basic understanding is helpful.

- Reading sequences is an imperfect biological process.

- Aligning to the reference genome is an imperfect algorithm, which makes many assumptions.

  - Some reads may not align anywhere. We throw them away.
    Are they errors?
    Are they real signal?

Today: three types of genomic alterations

- **Copy number variation**: deleted or amplified genes
- **Somatic mutations**: point mutations, small indels (insertion/deletion)
- **Structural variants**: fusions, rearrangements

Goals today

- ~~Sequencing: reading DNA*~~

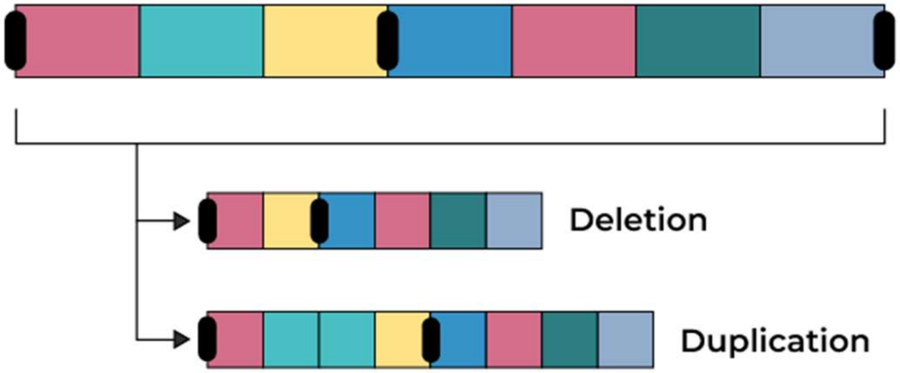
- **Copy-number variation**
- Somatic mutations
- Structural variants
- Nuts and bolts: which genes were tested?

* from a biostatistician, for biostatisticians

What is copy number variation?

Humans are diploid (2 copies of most genes). Sometimes in cancer, chromosomes chunks are deleted or duplicated.

Examples of Copy Number Variation

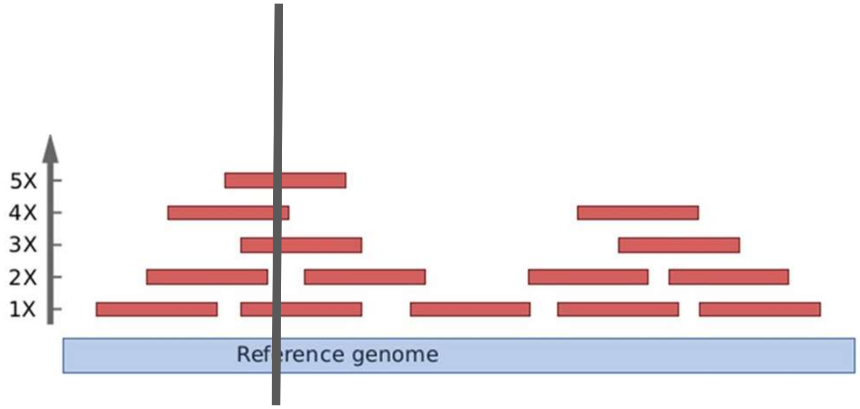


Duplicated oncogenes cause cells to reproduce uncontrollably.
Deleted tumor suppressors allow cells to reproduce uncontrollably.

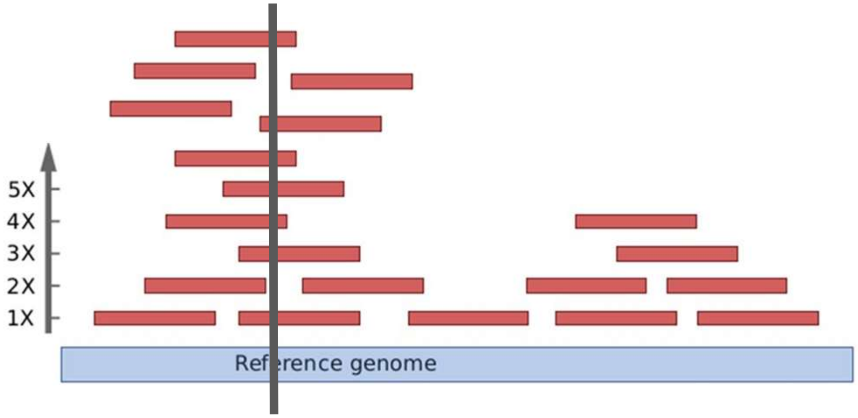
How do we quantify copy number variation?

We sequence a normal (blood) and tumor tissue.

Normal tissue



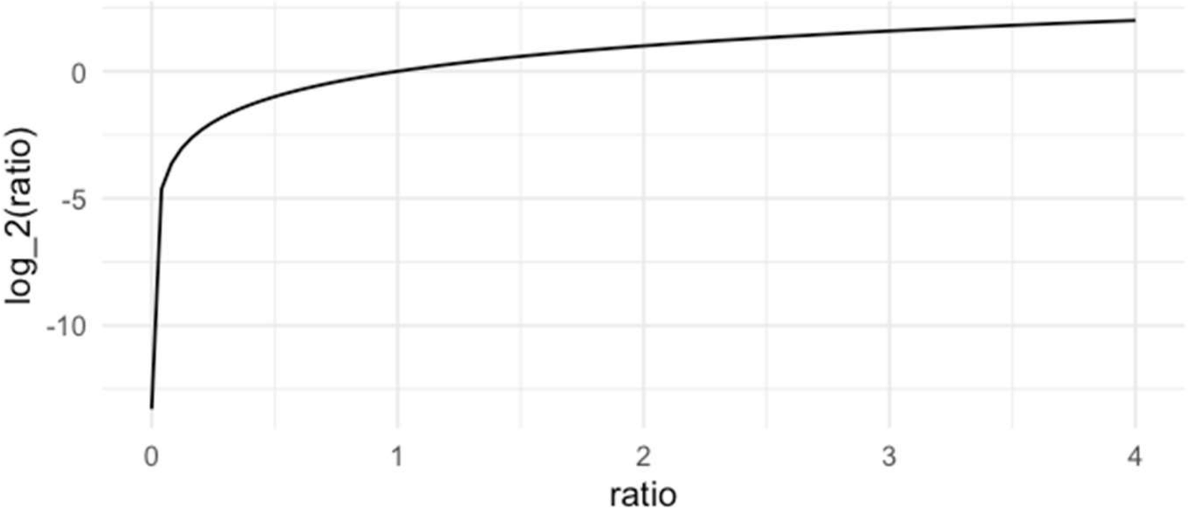
Tumor tissue



How would you distinguish between these?

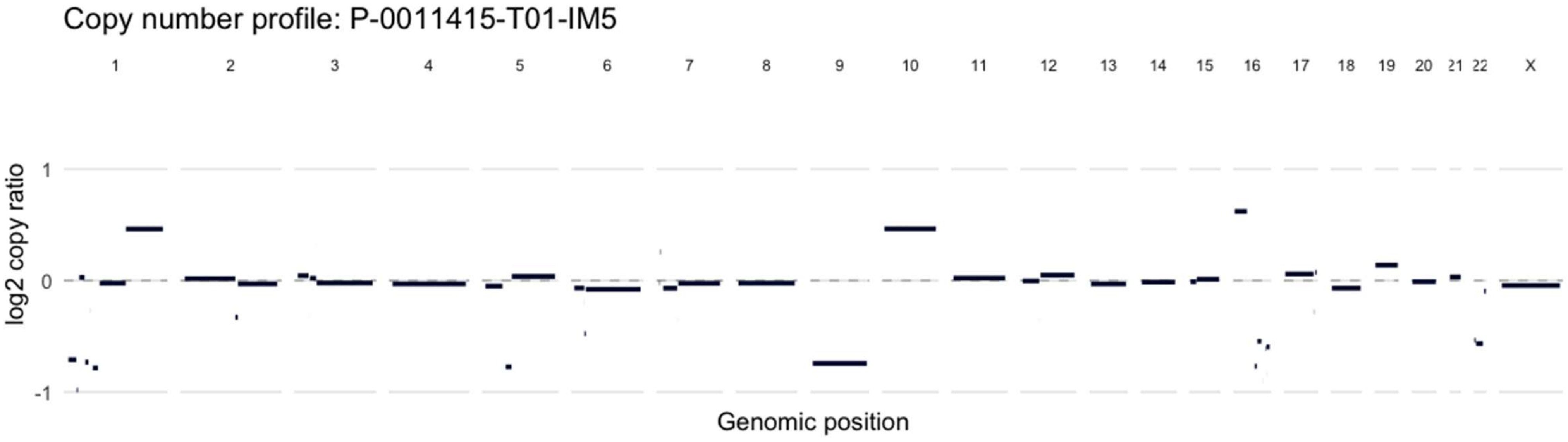
How do we quantify copy number variation?

- We use: $\log2\left(\frac{\textit{tumor depth}}{\textit{normal depth}}\right)$



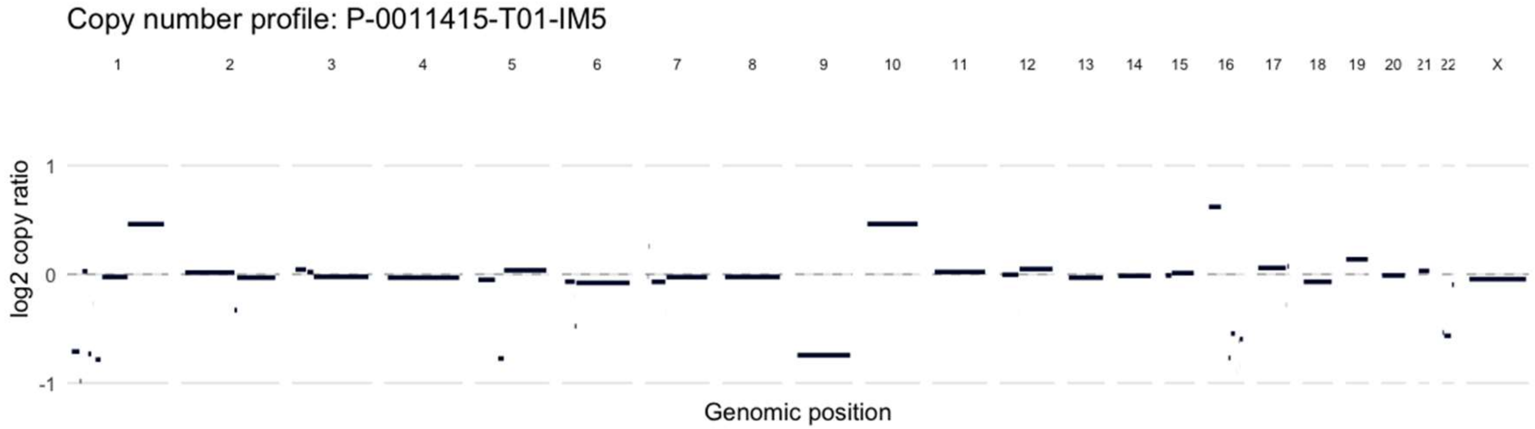
- What happens when tumor depth = normal depth?
- What happens when tumor depth > normal depth?

What copy number variation really looks like



What do you notice?

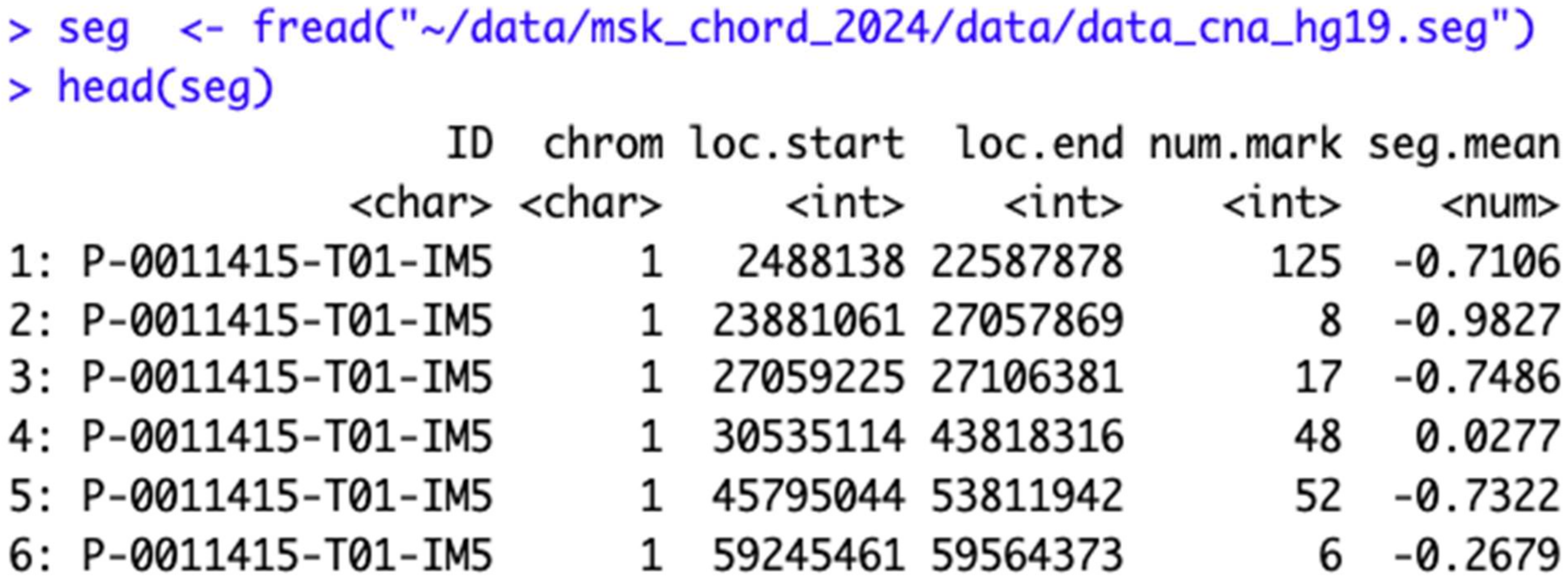
What copy number variation really looks like



The jumps away from the horizontal dashed line show us extra (or deleted) copies.

There is some variation even in normal regions of the genome.

What this data looks like in MSK-Chord



```
> seg  <- fread("~/data/msk_chord_2024/data/data_cna_hg19.seg")
> head(seg)
                    ID  chrom loc.start  loc.end num.mark seg.mean
                <char> <char>     <int>    <int>    <int>    <num>
1: P-0011415-T01-IM5      1   2488138 22587878      125  -0.7106
2: P-0011415-T01-IM5      1  23881061 27057869        8  -0.9827
3: P-0011415-T01-IM5      1  27059225 27106381       17  -0.7486
4: P-0011415-T01-IM5      1  30535114 43818316       48   0.0277
5: P-0011415-T01-IM5      1  45795044 53811942       52  -0.7322
6: P-0011415-T01-IM5      1  59245461 59564373        6  -0.2679
```

You may not want to work with this directly!
MSK-Chord uses an algorithm called GISTIC to make “calls” for copy number aberrations (broadinstitute.github.io/gistic2/).

“Calling” amplifications/deletions

The GISTIC algorithm makes “calls” about what is amplified/deleted.

Value	Meaning	Interpretation
-2	Homozygous (deep) deletion	0 copies
-1	Hemizygous (shallow) deletion	1 copy
0	Diploid / neutral	2 copies (normal)
+1	Low-level gain	3-4 copies
+2	High-level amplification	Many copies (5+)

CNV in our data: the HER2 gene

For example, HER2 is an **oncogene** (promoter):

Cancer	**-2**	**0**	**+2**
Breast	1	4692	675
Colorectal	0	5396	146
Non-Small Cell Lung	1	7697	109
Pancreatic	1	3079	26
Prostate	1	3206	4

CNV in our data: the TP53 gene

On the other hand, TP53 is a **tumor suppressor gene**:

Cancer	**-2**	**-1.5**	**0**	**+2**
Breast	45	4	5319	0
Colorectal	19	6	5516	1
Non-Small Cell Lung	31	14	7762	0
Pancreatic	3	0	3101	2
Prostate	46	12	3153	0

What do you see?

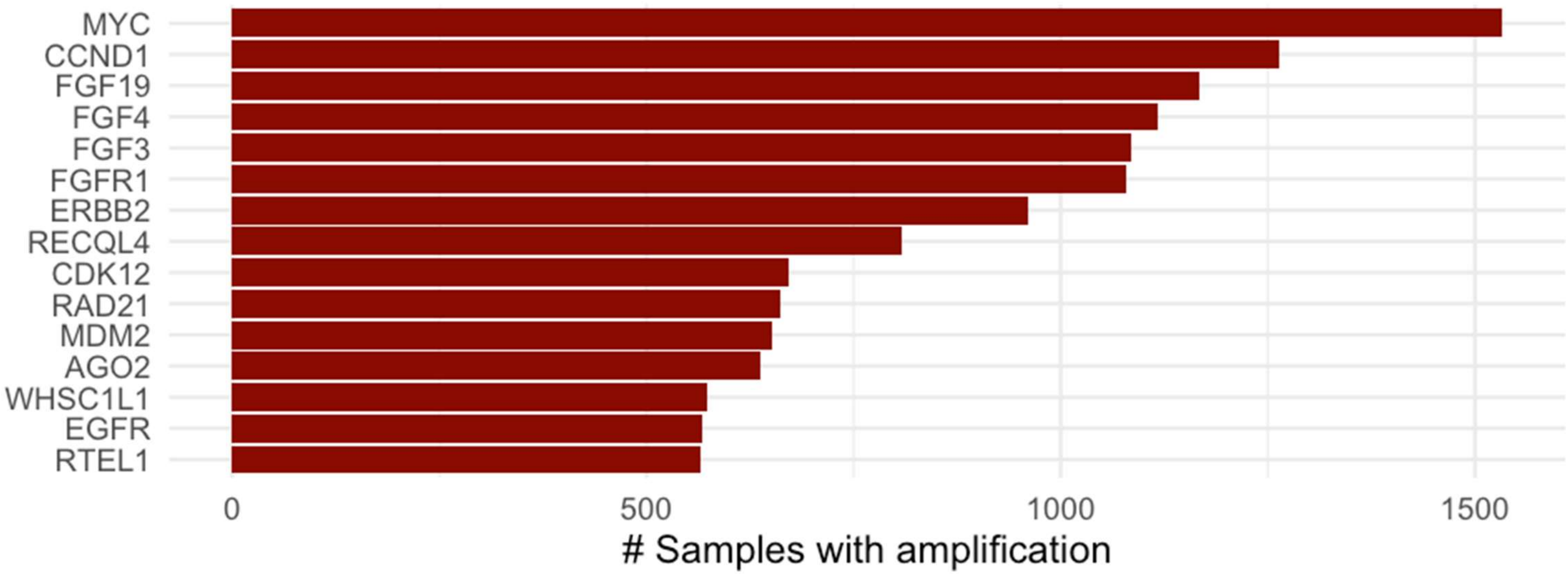
HER2

Cancer	-2	0	+2
Breast	1	4692	675
Colorectal	0	5396	146
Non-Small Cell Lung	1	7697	109
Pancreatic	1	3079	26
Prostate	1	3206	4

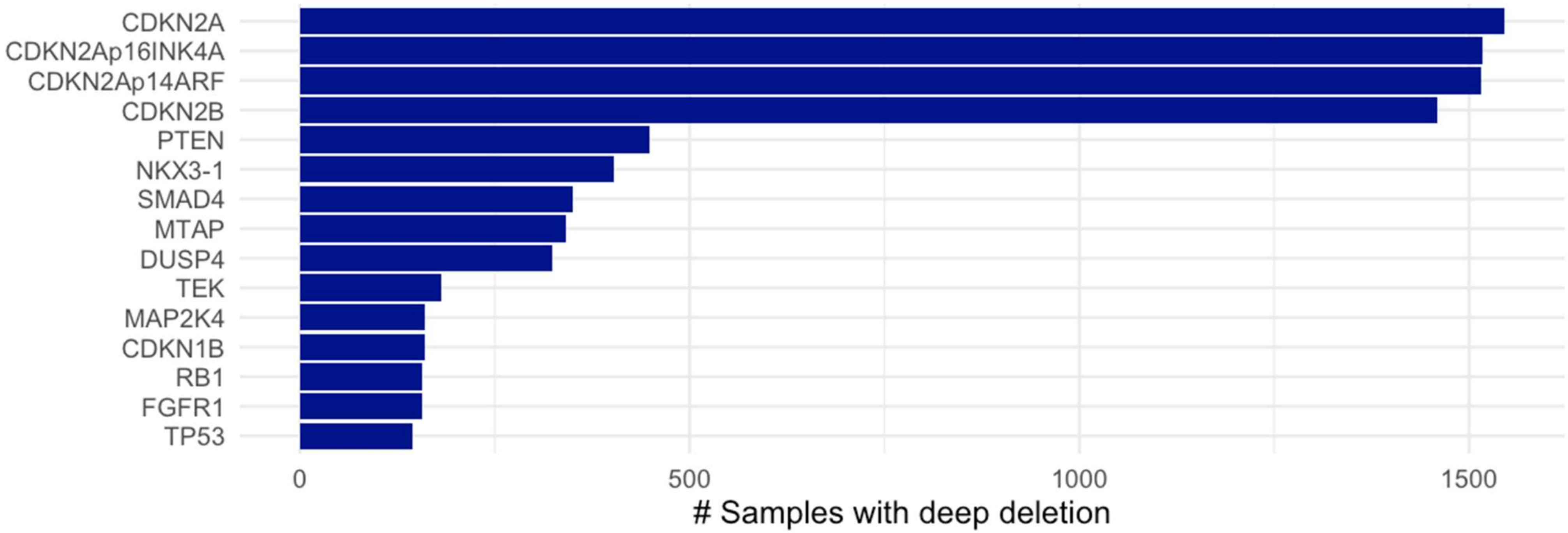
TP53

-2	-1.5	0	+2
45	4	5319	0
19	6	5516	1
31	14	7762	0
3	0	3101	2
46	12	3153	0

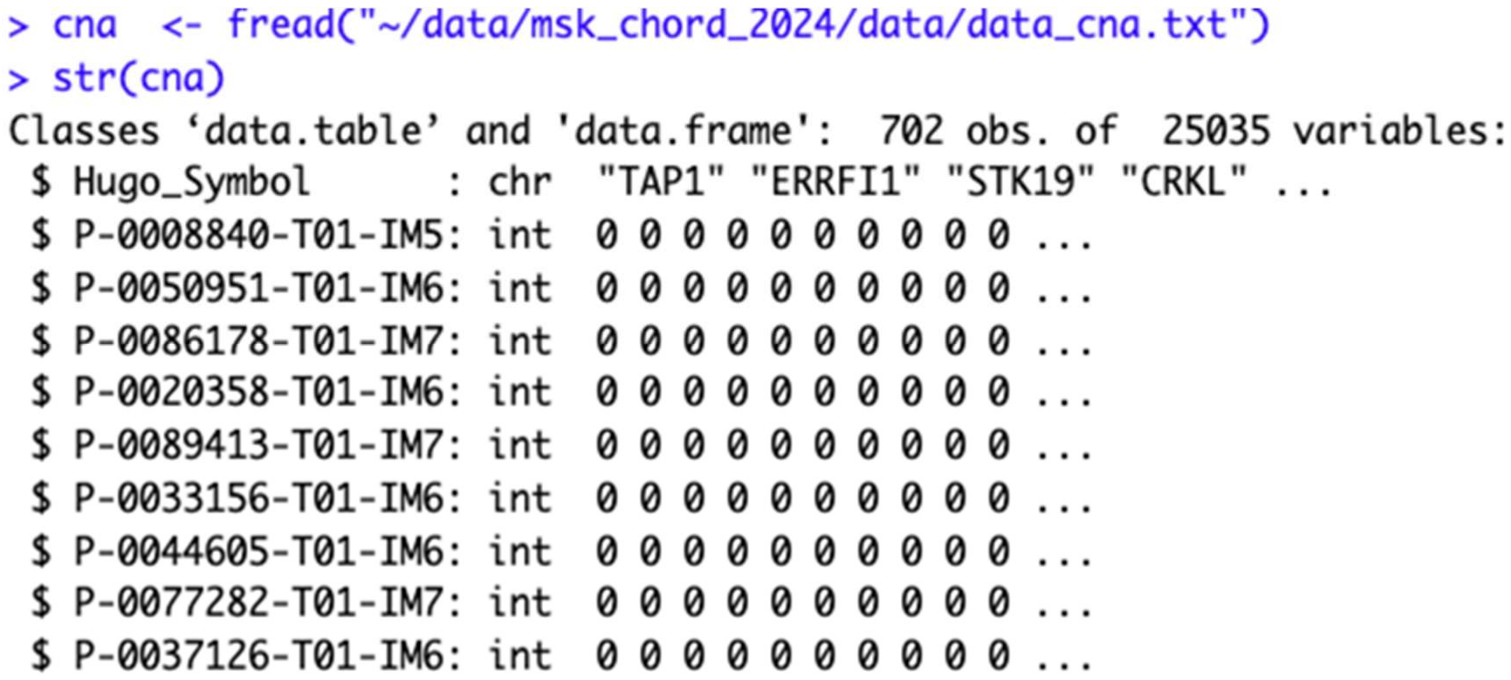
Most frequently amplified genes



Most frequently deleted genes

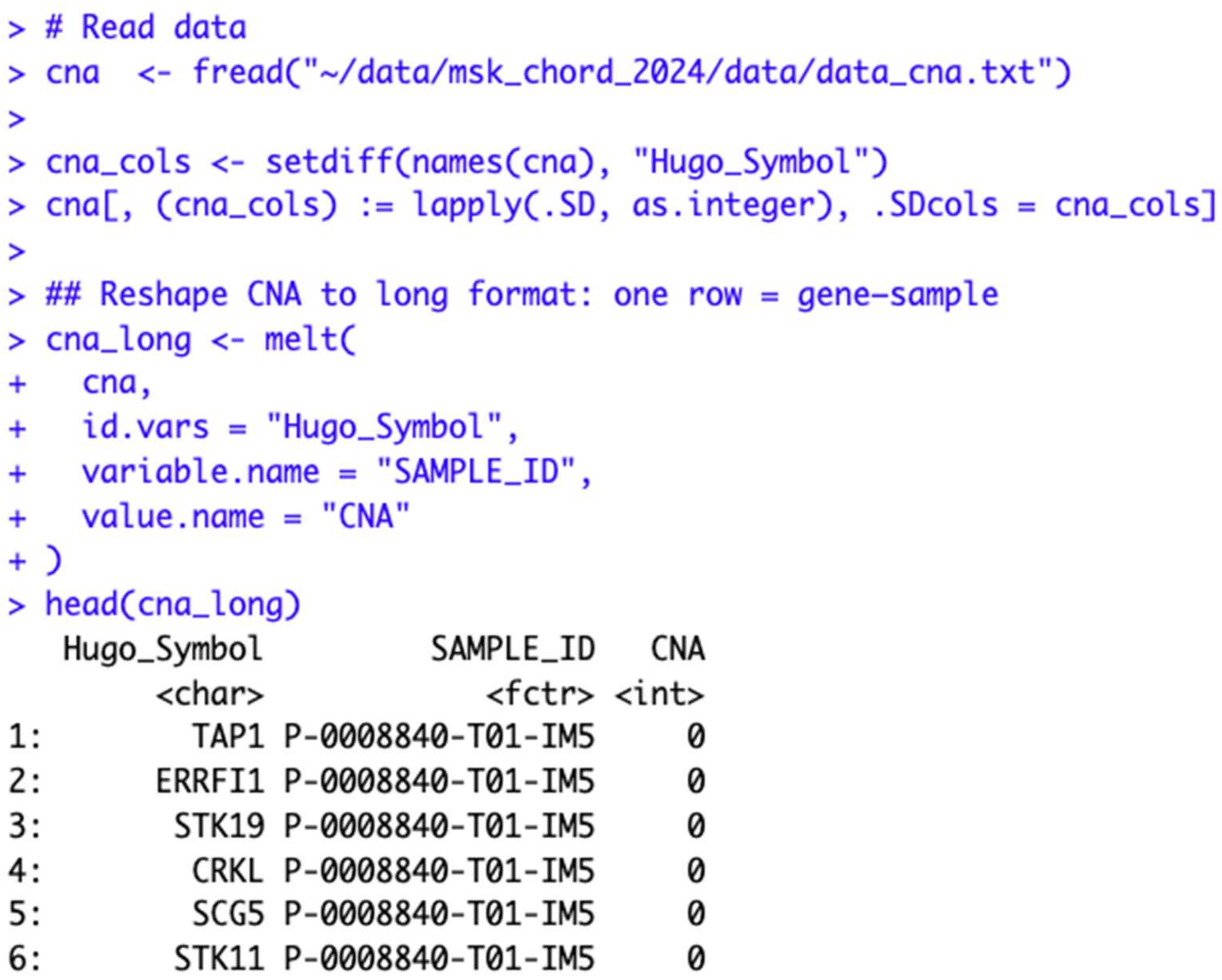


The CNA data in MSK-Chord



```
> cna  <- fread("~/data/msk_chord_2024/data/data_cna.txt")
> str(cna)
Classes ‘data.table’ and 'data.frame':  702 obs. of  25035 variables:
 $ Hugo_Symbol      : chr  "TAP1" "ERRFI1" "STK19" "CRKL" ...
 $ P-0008840-T01-IM5: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0050951-T01-IM6: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0086178-T01-IM7: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0020358-T01-IM6: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0089413-T01-IM7: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0033156-T01-IM6: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0044605-T01-IM6: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0077282-T01-IM7: int  0 0 0 0 0 0 0 0 0 0 ...
 $ P-0037126-T01-IM6: int  0 0 0 0 0 0 0 0 0 0 ...
```

702 rows (genes) by 25,035 columns (samples)
Note: some patients have multiple samples!
Also, this is hard to work with in this format!



```
> # Read data
> cna  <- fread("~/data/msk_chord_2024/data/data_cna.txt")
>
> cna_cols <- setdiff(names(cna), "Hugo_Symbol")
> cna[, (cna_cols) := lapply(.SD, as.integer), .SDcols = cna_cols]
>
> ## Reshape CNA to long format: one row = gene-sample
> cna_long <- melt(
+   cna,
+   id.vars = "Hugo_Symbol",
+   variable.name = "SAMPLE_ID",
+   value.name = "CNA"
+ )
> head(cna_long)
   Hugo_Symbol         SAMPLE_ID   CNA
        <char>            <fctr> <int>
1:        TAP1 P-0008840-T01-IM5     0
2:      ERRFI1 P-0008840-T01-IM5     0
3:       STK19 P-0008840-T01-IM5     0
4:        CRKL P-0008840-T01-IM5     0
5:        SCG5 P-0008840-T01-IM5     0
6:       STK11 P-0008840-T01-IM5     0
```

Project idea: co-occurrence

Do certain mutations tend to occur together, or exclude each other?

- Build a network of mutation/CNA co-occurrences and mutual exclusivities.

- Apply network clustering to identify “modules” of co-altered genes

- Do modules correspond to known pathways? Predict outcomes?

Goals today

- ~~Sequencing: reading DNA*~~

- ~~Copy-number variation~~
- **Somatic mutations**
- Structural variants
- Nuts and bolts: which genes were tested?

* from a biostatistician, for biostatisticians

What is a somatic mutation?

- A somatic mutation is a DNA alteration (single nucleotide change, insertion, or deletion) in tumor cells and absent from normal cells.

  Normal cell: AAA**T**CGATA
  Tumor cell: AAA**G**CGATA

- We find them by comparing the DNA in tumor vs normal tissue. For example, in one sample in MSK-Chord:

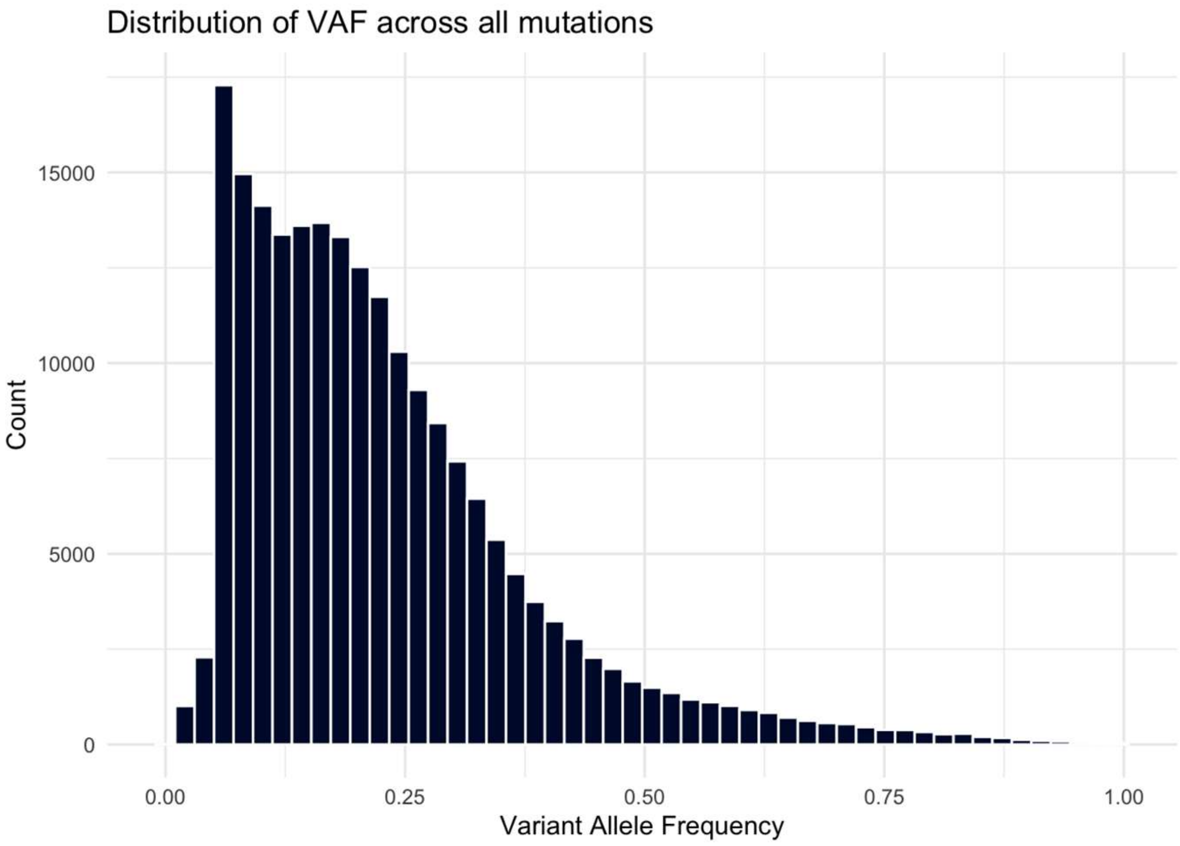
	# reference	**# mutated**
tumor	319	288
normal	281	0

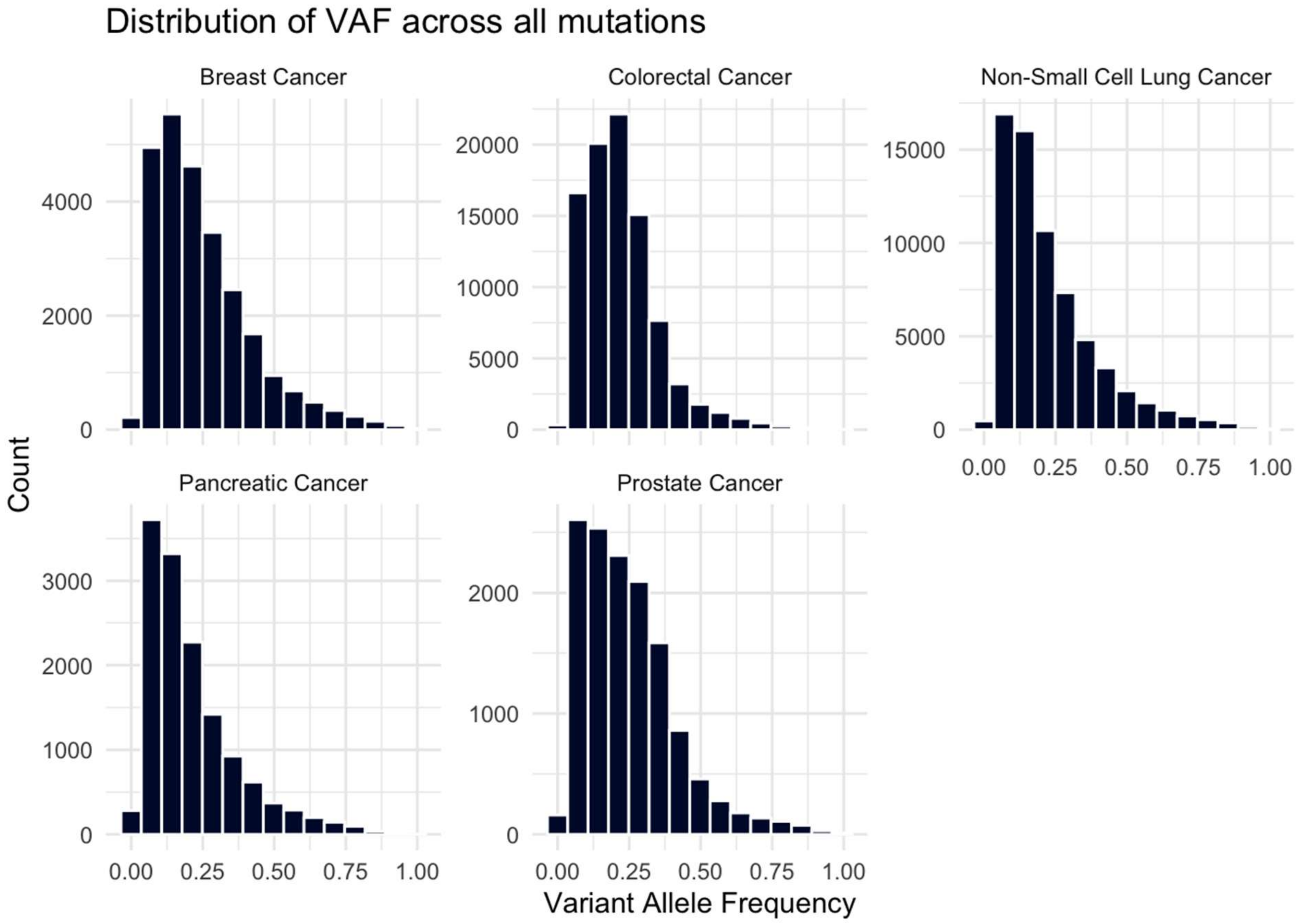
How do we think about somatic mutations?

	# reference	# mutated
tumor reads	319	288
normal reads	281	0

- The **Variant Allele Frequency** (VAF) is the fraction of tumor reads with the mutation.
- In this example, it is about 47%: $\frac{288}{319 + 288}$

What is the distribution of VAF in MSK-Chord?




Distribution of VAF across all mutations
Breast Cancer
Colorectal Cancer
Non-Small Cell Lung Cancer
Pancreatic Cancer
Prostate Cancer
Count
Variant Allele Frequency
0
2000
4000
0
5000
10000
15000
20000
0
5000
10000
15000
0
1000
2000
3000
0
1000
2000
0.00
0.25
0.50
0.75
1.00

What is the distribution of VAF in MSK-Chord?

Why does VAF vary so much? Two main reasons:

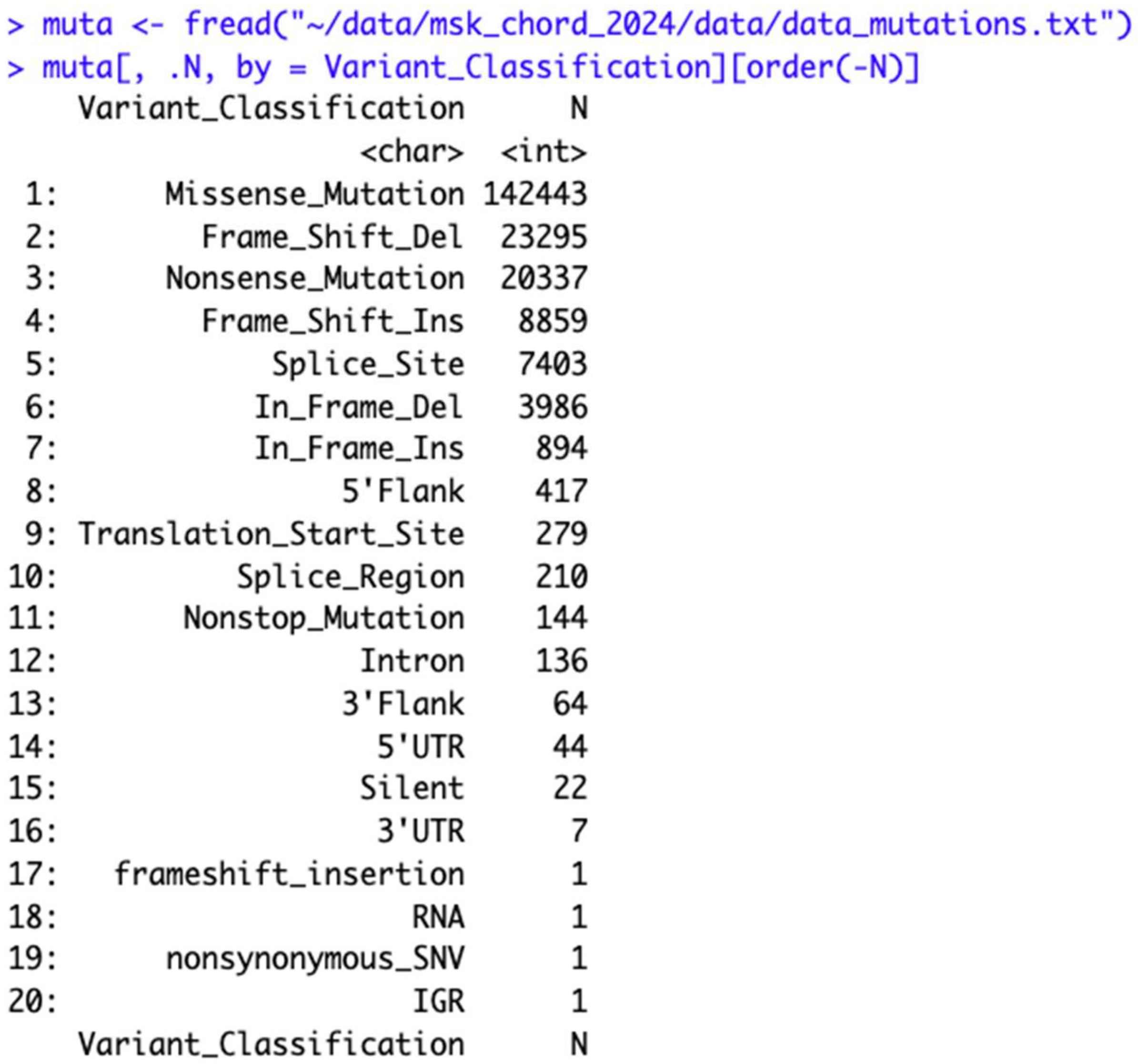
1. **Tumor purity**: The sample is a mix of tumor and normal cells. If only 50% of cells are tumor cells, even a mutation present in all tumor cells will have VAF ~ 0.25 (half the cells × half the chromosomes).

2. **Clonality**: Not all tumor cells have the same mutations. Early “founder” mutations are in all tumor cells (clonal), while later mutations may only be in a subset (subclonal).

What are the different types of mutations?

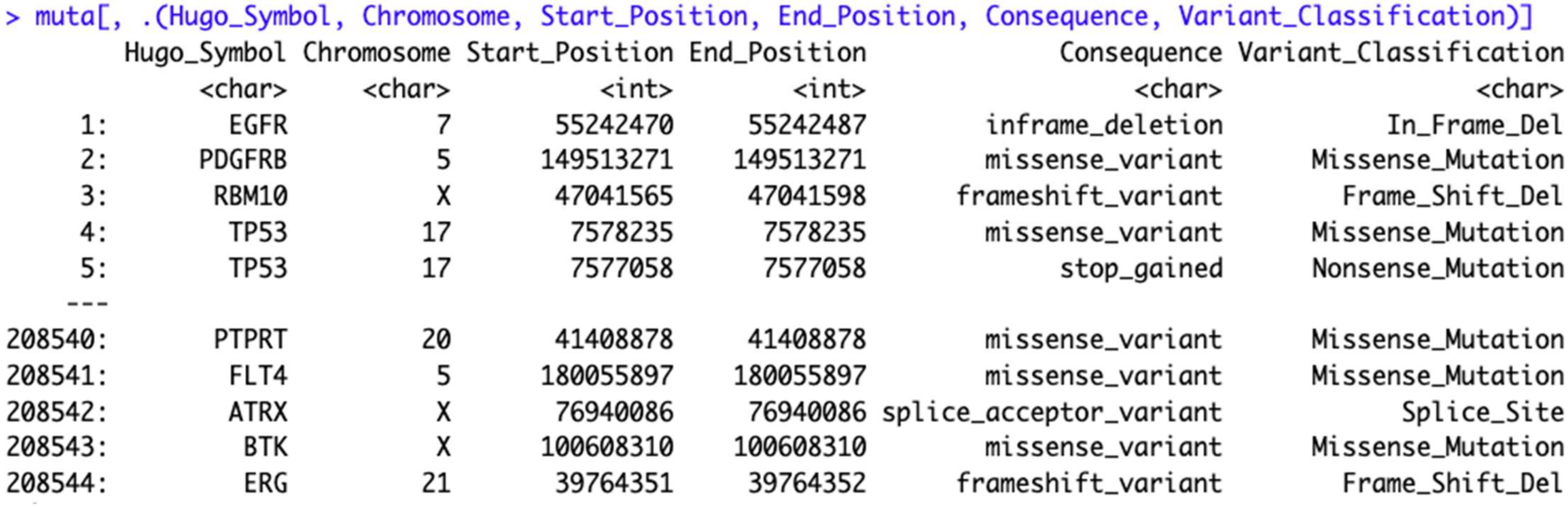
Type	Description	Effect
Missense	Amino acid changed	May alter function
Nonsense	Early STOP codon	Truncated protein
Frame shift	Reading frame shifted	Garbled protein
Splice site	Splicing disrupted	Abnormal mRNA

What are the different types of mutations?



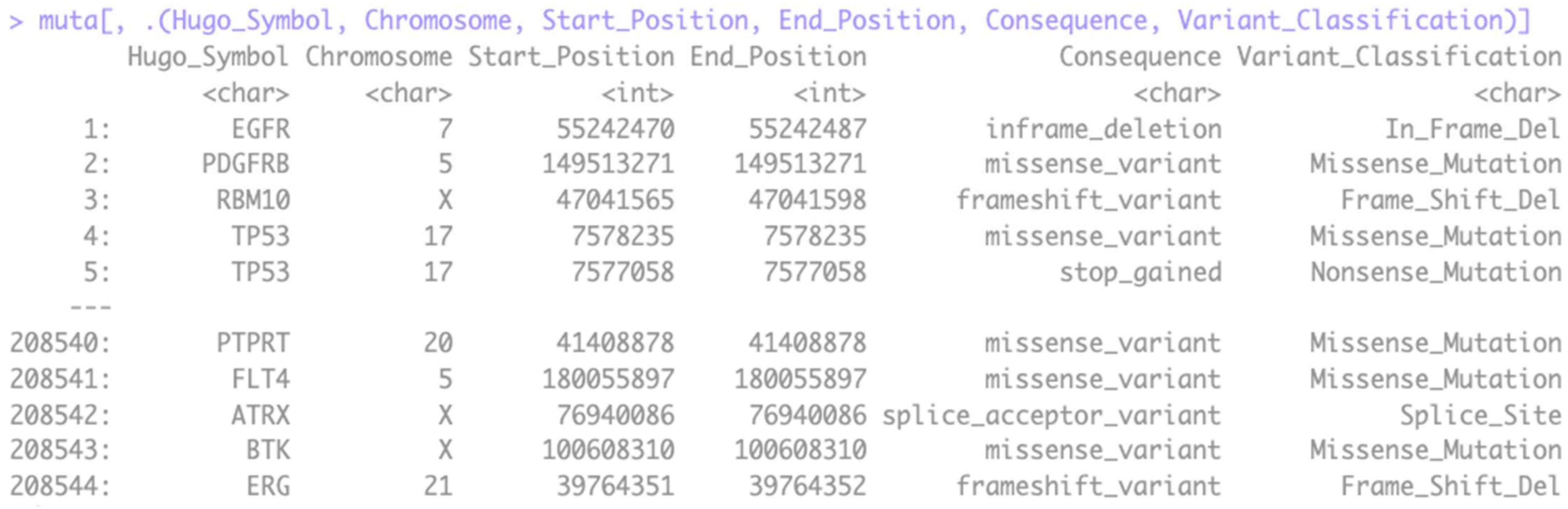
```
> muta <- fread("~/data/msk_chord_2024/data/data_mutations.txt")
> muta[, .N, by = Variant_Classification][order(-N)]
      Variant_Classification      N
                      <char>  <int>
 1:        Missense_Mutation 142443
 2:          Frame_Shift_Del  23295
 3:        Nonsense_Mutation  20337
 4:          Frame_Shift_Ins   8859
 5:              Splice_Site   7403
 6:             In_Frame_Del   3986
 7:             In_Frame_Ins    894
 8:                  5'Flank    417
 9:   Translation_Start_Site    279
10:            Splice_Region    210
11:         Nonstop_Mutation    144
12:                   Intron    136
13:                  3'Flank     64
14:                    5'UTR     44
15:                   Silent     22
16:                    3'UTR      7
17:     frameshift_insertion      1
18:                      RNA      1
19:        nonsynonymous_SNV      1
20:                      IGR      1
      Variant_Classification      N
```

What else is in this file?

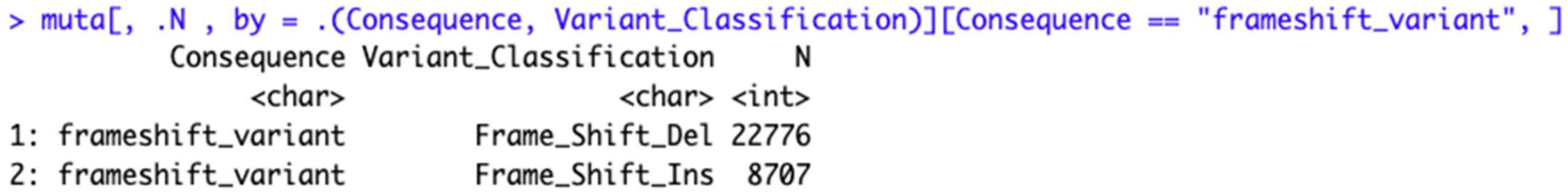


```
> muta[, .(Hugo_Symbol, Chromosome, Start_Position, End_Position, Consequence, Variant_Classification)]
        Hugo_Symbol Chromosome Start_Position End_Position            Consequence Variant_Classification
             <char>     <char>          <int>        <int>                 <char>                 <char>
     1:        EGFR          7       55242470     55242487       inframe_deletion           In_Frame_Del
     2:      PDGFRB          5      149513271    149513271       missense_variant      Missense_Mutation
     3:       RBM10          X       47041565     47041598     frameshift_variant        Frame_Shift_Del
     4:        TP53         17        7578235      7578235       missense_variant      Missense_Mutation
     5:        TP53         17        7577058      7577058            stop_gained      Nonsense_Mutation
    ---
208540:       PTPRT         20       41408878     41408878       missense_variant      Missense_Mutation
208541:        FLT4          5      180055897    180055897       missense_variant      Missense_Mutation
208542:        ATRX          X       76940086     76940086 splice_acceptor_variant            Splice_Site
208543:         BTK          X      100608310    100608310       missense_variant      Missense_Mutation
208544:         ERG         21       39764351     39764352     frameshift_variant        Frame_Shift_Del
```

What else is in this file?

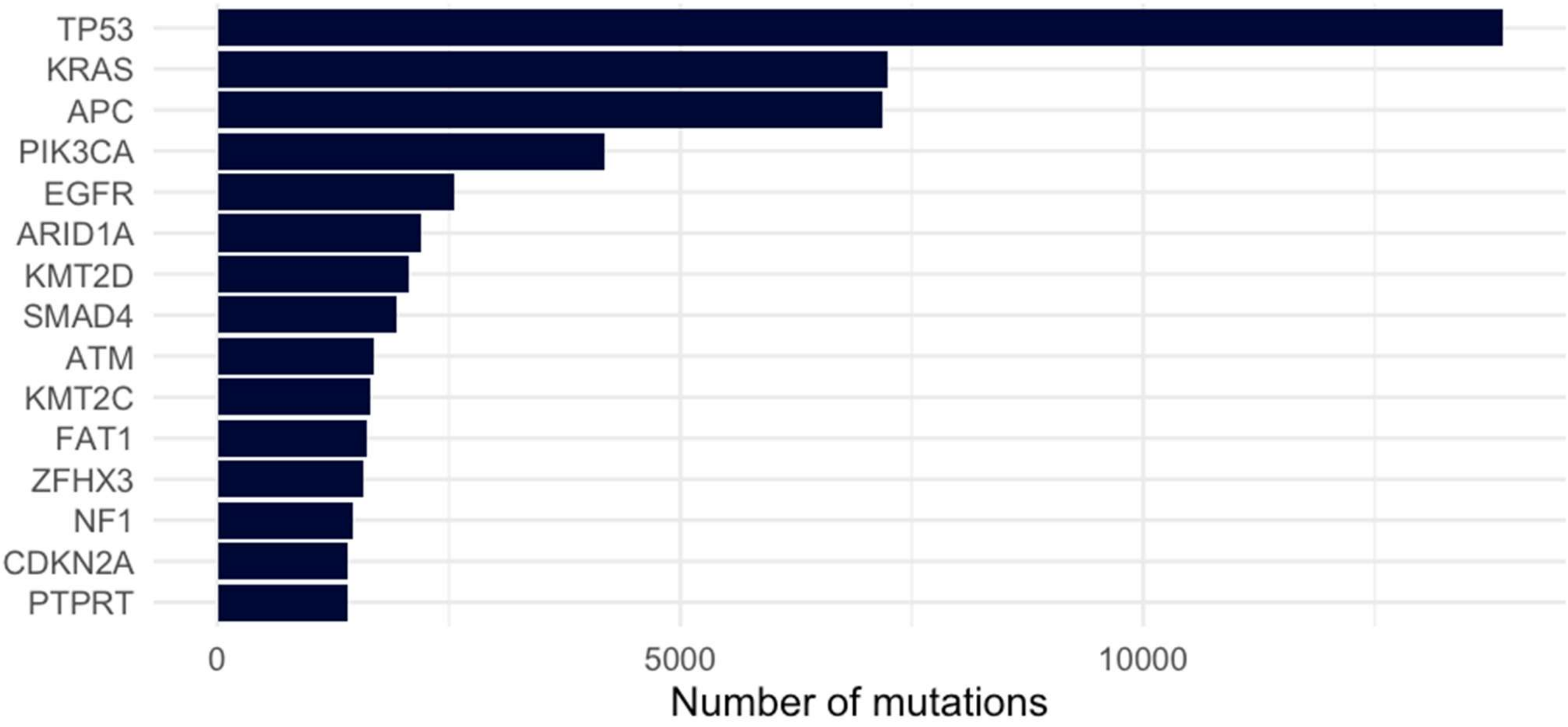


```
> muta[, .(Hugo_Symbol, Chromosome, Start_Position, End_Position, Consequence, Variant_Classification)]
        Hugo_Symbol Chromosome Start_Position End_Position             Consequence Variant_Classification
             <char>     <char>          <int>        <int>                  <char>                 <char>
     1:        EGFR          7       55242470     55242487        inframe_deletion           In_Frame_Del
     2:      PDGFRB          5      149513271    149513271        missense_variant      Missense_Mutation
     3:       RBM10          X       47041565     47041598      frameshift_variant         Frame_Shift_Del
     4:        TP53         17        7578235      7578235        missense_variant      Missense_Mutation
     5:        TP53         17        7577058      7577058             stop_gained      Nonsense_Mutation
    ---
208540:       PTPRT         20       41408878     41408878        missense_variant      Missense_Mutation
208541:        FLT4          5      180055897    180055897        missense_variant      Missense_Mutation
208542:        ATRX          X       76940086     76940086 splice_acceptor_variant            Splice_Site
208543:         BTK          X      100608310    100608310        missense_variant      Missense_Mutation
208544:         ERG         21       39764351     39764352      frameshift_variant         Frame_Shift_Del
```



```
> muta[, .N , by = .(Consequence, Variant_Classification)][Consequence == "frameshift_variant", ]
          Consequence Variant_Classification     N
               <char>                 <char> <int>
1: frameshift_variant        Frame_Shift_Del 22776
2: frameshift_variant        Frame_Shift_Ins  8707
```

Top mutated genes



Project idea: clonal evolution from VAF

Can you infer tumor subpopulations from VAF distributions?

- Cluster mutations by VAF within each patient (using e.g. mixture models).
- Estimate number of clones per patient.
- Does clonal complexity predict worse outcomes?

Goals today

- ~~Sequencing: reading DNA*~~

- ~~Copy-number variation~~
- ~~Somatic mutations~~
- **Structural variants**
- Nuts and bolts: which genes were tested?

* from a biostatistician, for biostatisticians

What is a structural variation?

- Structural variants (SVs) are large-scale rearrangements of the genome.

- There are four categories of structural variants:
  - **Deletion**: large segment removed
  - **Inversion**: segment flipped in orientation
  - **Translocation**: segment moved between chromosomes
  - **Duplication**: segment copied

For example

- **Fusion** is a type of translocation where two genes are joined together.

- Some fusions are targetable with drugs:
  - EML4::ALK *(lung cancer)* Crizotinib, Alectinib
  - BCR::ABL1 *CML (leukemia)* Imatinib (Gleevec)

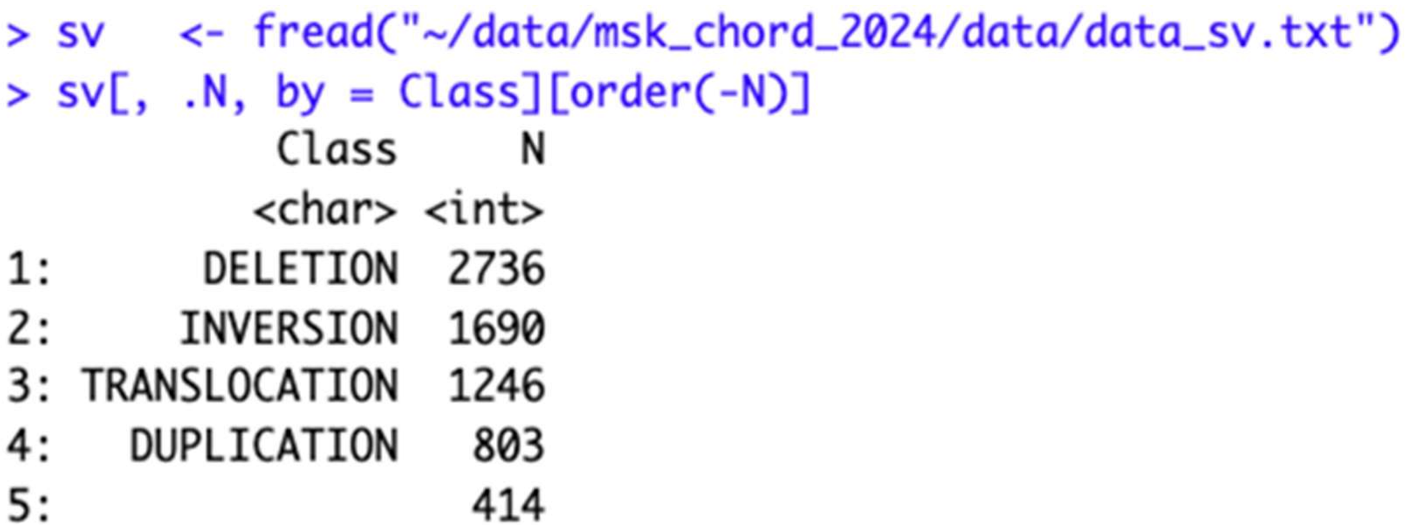
Normal: **═════[EML4]═════ ═════[ALK]═════**

Fused: **═════[EML4═══ALK]═════**

Targetable fusions

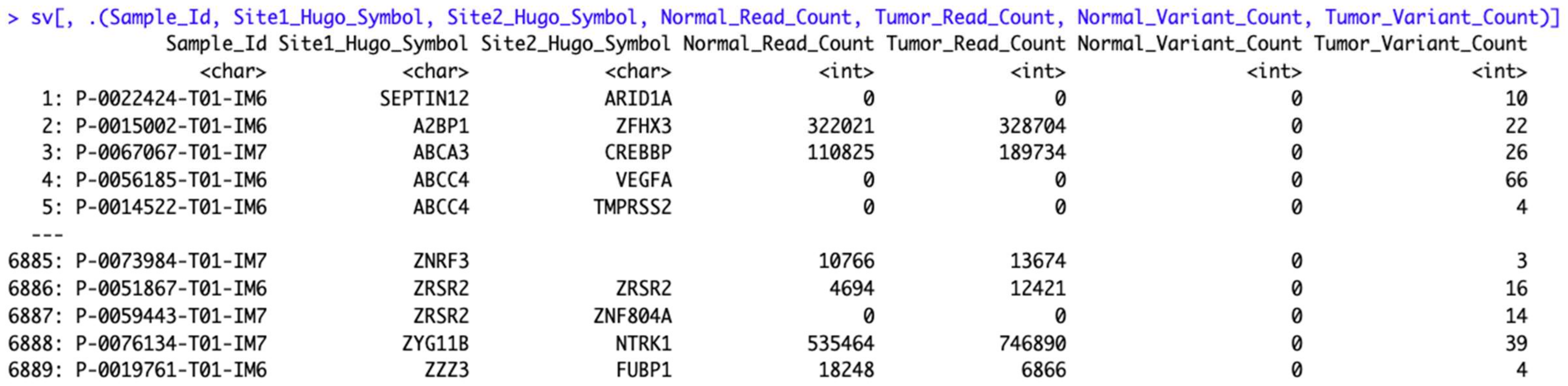
Fusion	Cancer	Treatment
EML4::ALK	Lung	Crizotinib, Alectinib
BCR::ABL1	CML	Imatinib (Gleevec)
TMPRSS2::ERG	Prostate	Diagnostic marker
FGFR fusions	Multiple	FGFR inhibitors

The SV data in MSK-Chord



```
> sv   <- fread("~/data/msk_chord_2024/data/data_sv.txt")
> sv[, .N, by = Class][order(-N)]
          Class     N
         <char> <int>
1:     DELETION  2736
2:    INVERSION  1690
3: TRANSLOCATION 1246
4:  DUPLICATION   803
5:                414
```

The SV data in MSK-Chord



```
> sv[, .(Sample_Id, Site1_Hugo_Symbol, Site2_Hugo_Symbol, Normal_Read_Count, Tumor_Read_Count, Normal_Variant_Count, Tumor_Variant_Count)]
                Sample_Id Site1_Hugo_Symbol Site2_Hugo_Symbol Normal_Read_Count Tumor_Read_Count Normal_Variant_Count Tumor_Variant_Count
                   <char>            <char>            <char>             <int>            <int>                <int>               <int>
   1: P-0022424-T01-IM6          SEPTIN12            ARID1A                 0                0                    0                  10
   2: P-0015002-T01-IM6             A2BP1             ZFHX3            322021           328704                    0                  22
   3: P-0067067-T01-IM7             ABCA3            CREBBP            110825           189734                    0                  26
   4: P-0056185-T01-IM6             ABCC4             VEGFA                 0                0                    0                  66
   5: P-0014522-T01-IM6             ABCC4           TMPRSS2                 0                0                    0                   4
  ---
6885: P-0073984-T01-IM7             ZNRF3                               10766            13674                    0                   3
6886: P-0051867-T01-IM6             ZRSR2             ZRSR2              4694            12421                    0                  16
6887: P-0059443-T01-IM7             ZRSR2           ZNF804A                 0                0                    0                  14
6888: P-0076134-T01-IM7            ZYG11B             NTRK1            535464           746890                    0                  39
6889: P-0019761-T01-IM6              ZZZ3             FUBP1             18248             6866                    0                   4
```

Goals today

- ~~Sequencing: reading DNA*~~
- ~~Copy-number variation~~
- ~~Somatic mutations~~
- ~~Structural variants~~
- **Nuts and bolts: which genes were tested?**

* from a biostatistician, for biostatisticians

Which genes were tested?

MSK-Chord uses **targeted panels** (~500 genes, not whole genome).

Not every patient received the same panel:

Panel	N
IMPACT468	12891
IMPACT505	7155
IMPACT410	3973
IMPACT341	1019
IMPACT-HEME-400	2

“No mutation” is different from “Not tested”

For example, CALR is in IMPACT410+ but not IMPACT341.

For IMPACT341 samples, we cannot say whether they have a CALR mutation.

Project idea: Predicting progression-free survival

Can baseline genomics predict time to progression?

- Build survival models (Cox, random survival forests) using e.g. mutation, CNA, and clinical features.

- Compare simple and complex models.

Summary

Now we have an overview of the genomic data in MSK-Chord:

- Copy-number variation
- Somatic mutations
- Structural variants

What questions do you have?

In case of extra time….

Promises and Perils of Observational Data

MSK-Chord is observational.

MSK-Chord is **not** a clinical trial:

- Patients were sequenced as part of routine clinical care.
- No randomization to treatment.
- No protocol-defined follow-up.
- Patients entered the cohort at different times and stages of disease.

This creates both **opportunities** and **challenges**.

Overview

1. **Promises of large observational cohorts**
2. Perils of large observational cohorts

What are some promises of observational data?

Sample size and power

MSK-Chord has ~25,000 patients, which is much larger than most clinical trials.

We can:

- Detect rare genomic alterations
- Study rare cancer subtypes
- Identify small effect sizes
- Do subgroup analyses with adequate power

Real world patients

Clinical trials have strict eligibility criteria:

- Good performance status
- No major comorbidities
- Often younger patients

MSK-Chord reflects the rich diversity of patients seen in the clinic.

Longitudinal data

MSK-Chord includes rich long-term timeline data:

- Treatment sequences over time
- Progression events
- Tumor site changes
- Lab values

This enables questions about **treatment sequencing** and **disease trajectory** that trials typically can't answer.

Hypothesis generation

Observational data is excellent for generating hypotheses:

- Discover unexpected associations
- Identify candidate biomarkers
- Find patterns across cancer types

These hypotheses can then be tested in prospective studies.

Overview

1. Promises of large observational cohorts

2. **Perils of large observational cohorts**

What are some perils of observational data?

Selection bias

Who gets sequenced at MSK?

- Patients who can travel to a major cancer center
- Patients with insurance/resources
- Patients healthy enough to undergo biopsy
- Patients whose tumors are accessible

This is not a random sample of cancer patients.

Survivorship bias

We only observe patients who survived long enough to be included.

Example: "Metastatic patients in MSK-Chord have median survival of X months."

- But patients who died before sequencing are not in the dataset
- We're conditioning on survival to a certain point

This can make prognosis look better than it truly is.

Confounding

Treatments are not randomized. Doctors choose treatments based on:

- Patient health status
- Tumor characteristics
- Prior treatments
- Patient preferences

These same factors also affect outcomes. This is called confounding.

EXAMPLE: Confounding

Suppose patients on Drug A survive longer than patients on Drug B.

What are some possible explanations?

EXAMPLE: Confounding

Suppose patients on Drug A survive longer than patients on Drug B.

Possible explanations:

1. Drug A is better *(causal)*
2. Healthier patients get Drug A *(confounding)*
3. Drug A is given to patients with better-prognosis tumors *(confounding)*

To distinguish between these, we must (carefully!) use techniques from causal inference.

Immortal time bias

Immortal time:

Time during which the outcome cannot occur, often due to study design.

Example: "Patients who received genomic sequencing survived longer."

- But you have to survive long enough to get sequenced!
- Time before sequencing is "immortal": death would exclude you.

This artificially inflates survival in the sequenced group.

EXAMPLE: Immortal time bias

MSK-Chord patients were sequenced at different points in their disease:

- Some at diagnosis
- Some after progression
- Some after multiple treatments

Comparing patient outcomes without accounting for **when they entered** the cohort is dangerous.

Missing data

Not all data is collected consistently:

- Different sequencing panels (IMPACT341 vs IMPACT505)
- Incomplete treatment records
- Loss to follow-up
- Missing progression dates

Multiple testing

Doing hypothesis testing? Adjust for multiple hypothesis testing!

With 500 genes and multiple endpoints, the multiple testing burden is severe.

- 500 genes × 5 cancer types × 3 outcomes = 7,500 tests
- At $\alpha = 0.05$, expect 375 false positives by chance

Require rigorous correction (Bonferroni, FDR) and validation.