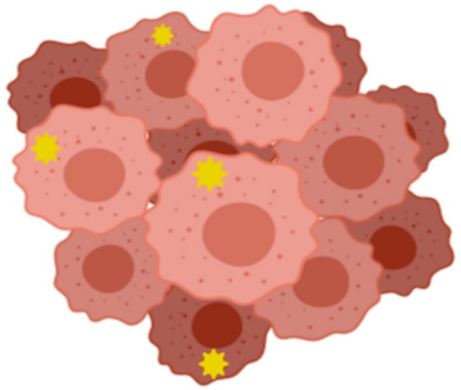
wellcome connecting science

connecting people with science

Day 3: Driver Gene Identification and Oncoplots

29th November 2023

Driver Gene Identification and Oncoplots



Presented by:

Patricia Basurto-Lozada
PhD Candidate
LIIGH-UNAM
pbasurto@liigh-unam.mx

Nyasha Chambwe
Assistant Professor
Institute of Molecular Medicine
Feinstein Institutes for Medical Research
@doc_nyasha
nchambwe@northwell.edu

Module materials and slides adapted from content developed by:

Federico Abascal, PhD
Wellcome Sanger Institute

Cancer Genome Analysis - Latin America & the Caribbean

26 November–1 December 2023

Universidad de la República, Montevideo, Uruguay

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

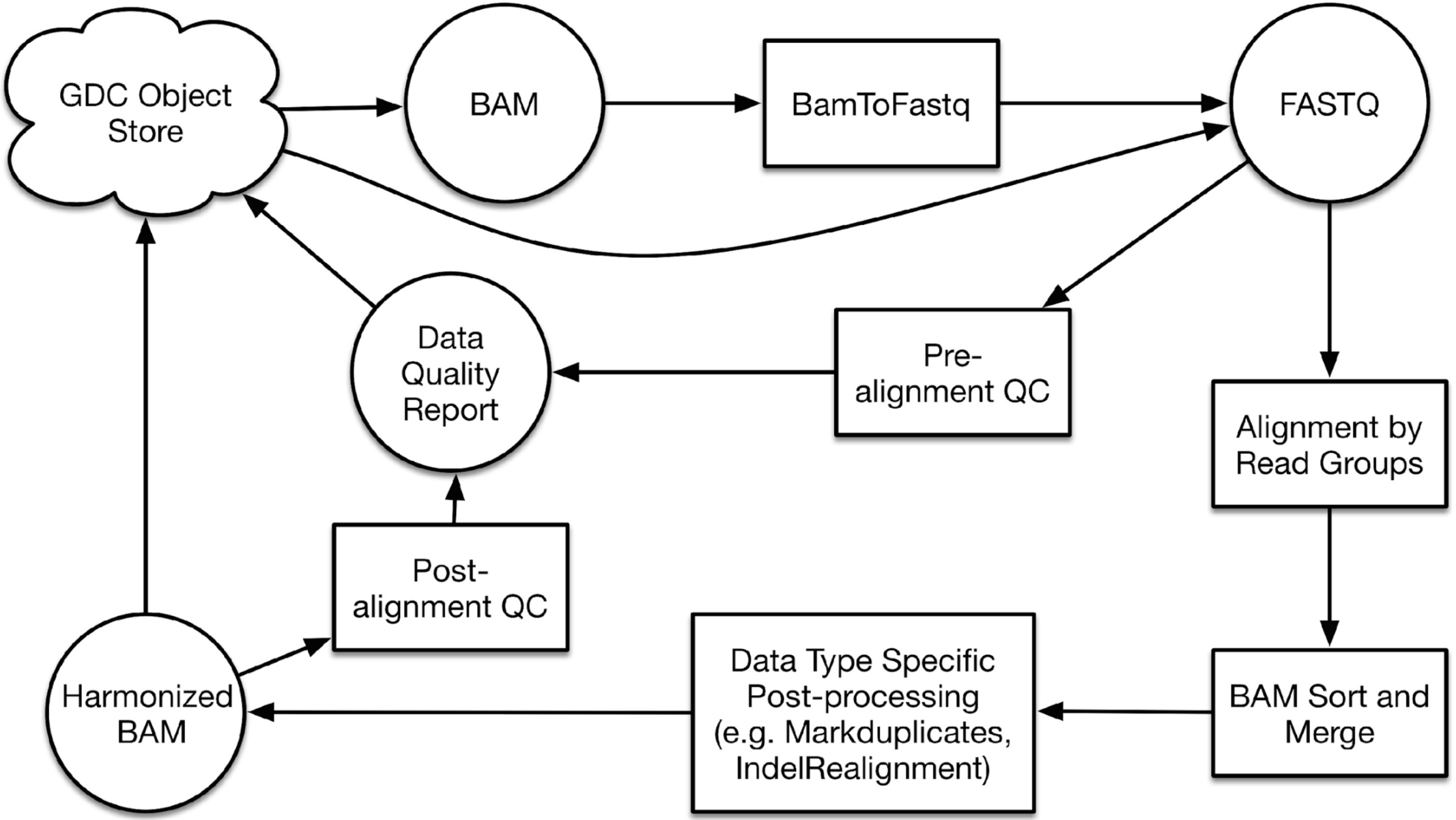
wellcome connecting science

Schedule

9:00am	Introductory Lecture: Part I
10:30am	***Coffee Break***
11:00am	Introductory Lecture: Part II
12:15pm	**Seminar: Patricia Basurto**
	"The genomic profile of acral lentiginous melanoma tumors from Mexican parents"
1:00 pm	***Lunch***
2:00 pm	Module 3 Practical: Driver detection with ***dndscv***
3:30pm	**Afternoon Coffee Break**
4:00pm	Module 3 Practical: Driver detection with ***dndscv***
	(continuation)
5:00pm	*Group Discussion, Q&A*

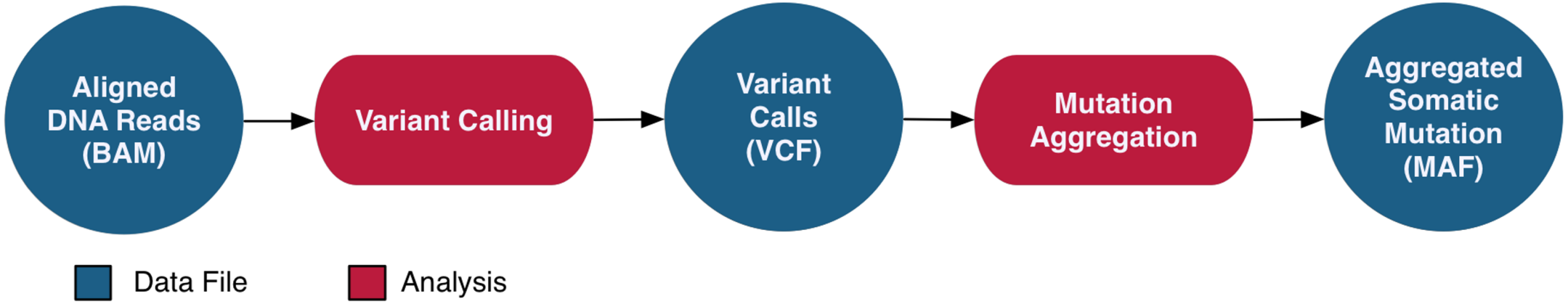
Bus to Hotel……

NCI GDC Reference Genome and Alignment Workflow



https://gdc.cancer.gov/about-data/gdc-data-processing/genomic-data-processing

NCI GDC DNA-Seq WXS Somatic Variant Analysis

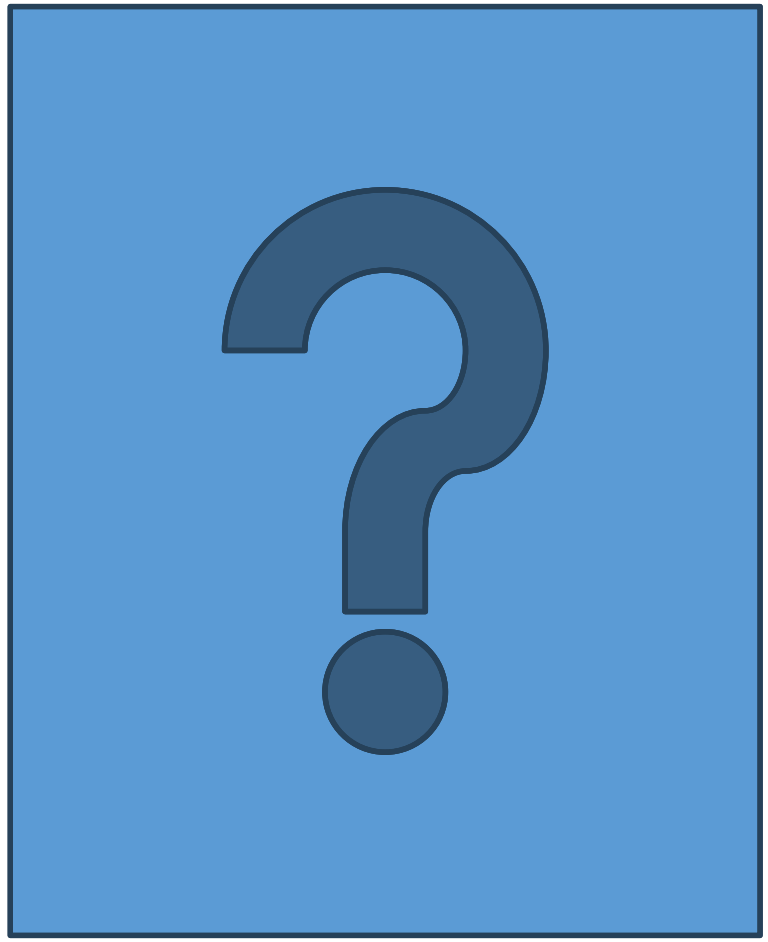


https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/#somatic-variant-calling-workflow

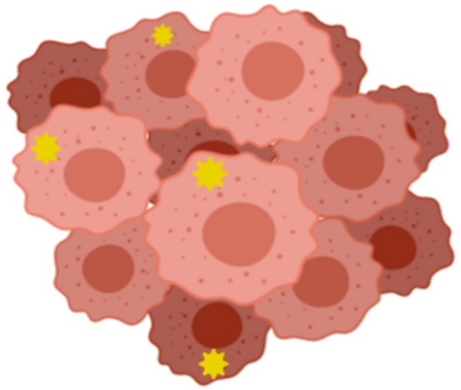
Note:

Tumor-Only Variant Calling Workflow

Tumor-Only Variant Annotation Workflow



Driver Gene Identification and Oncoplots



Presented by:

Patricia Basurto-Lozada
PhD Candidate
LIIGH-UNAM
pbasurto@liigh-unam.mx

Nyasha Chambwe
Assistant Professor
Institute of Molecular Medicine
Feinstein Institutes for Medical Research
@doc_nyasha
nchambwe@northwell.edu

Module materials and slides adapted from content developed by:

Federico Abascal, PhD
Wellcome Sanger Institute

Cancer Genome Analysis - Latin America & the Caribbean

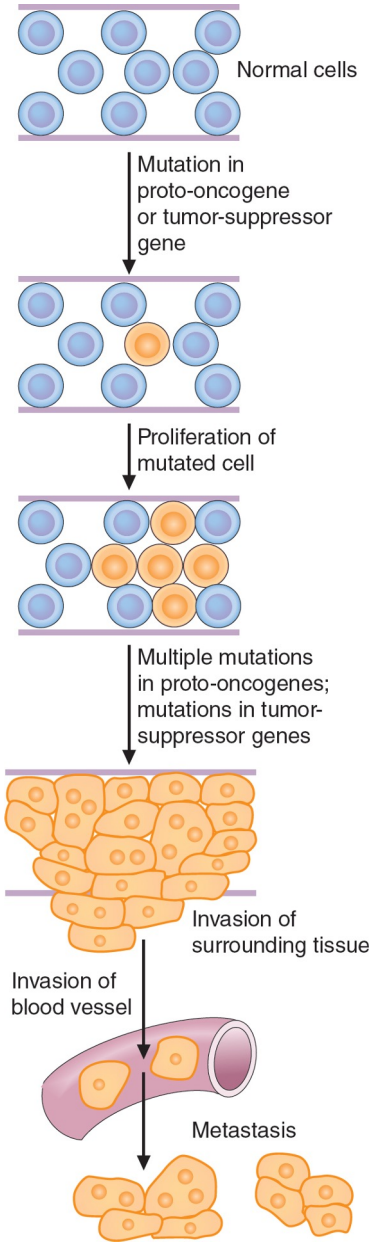
26 November–1 December 2023

Universidad de la República, Montevideo, Uruguay

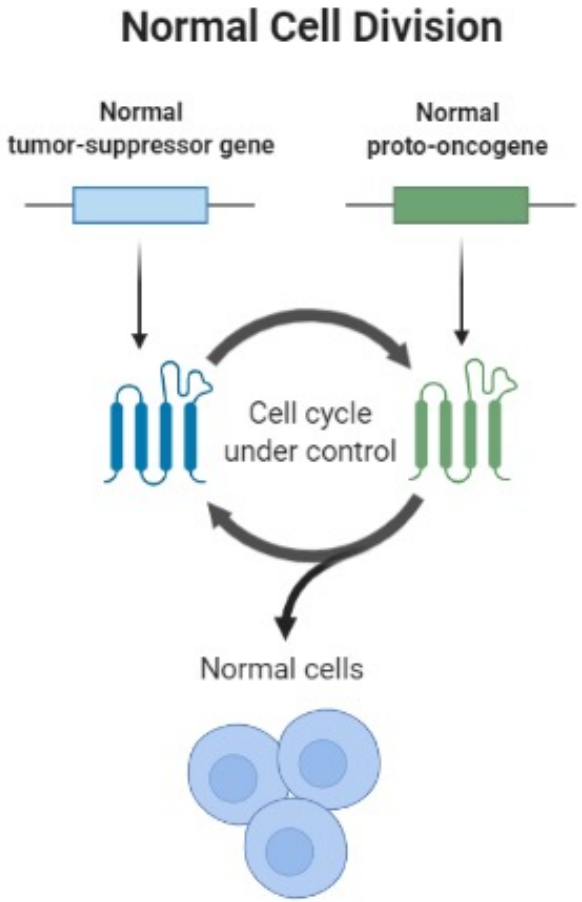
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

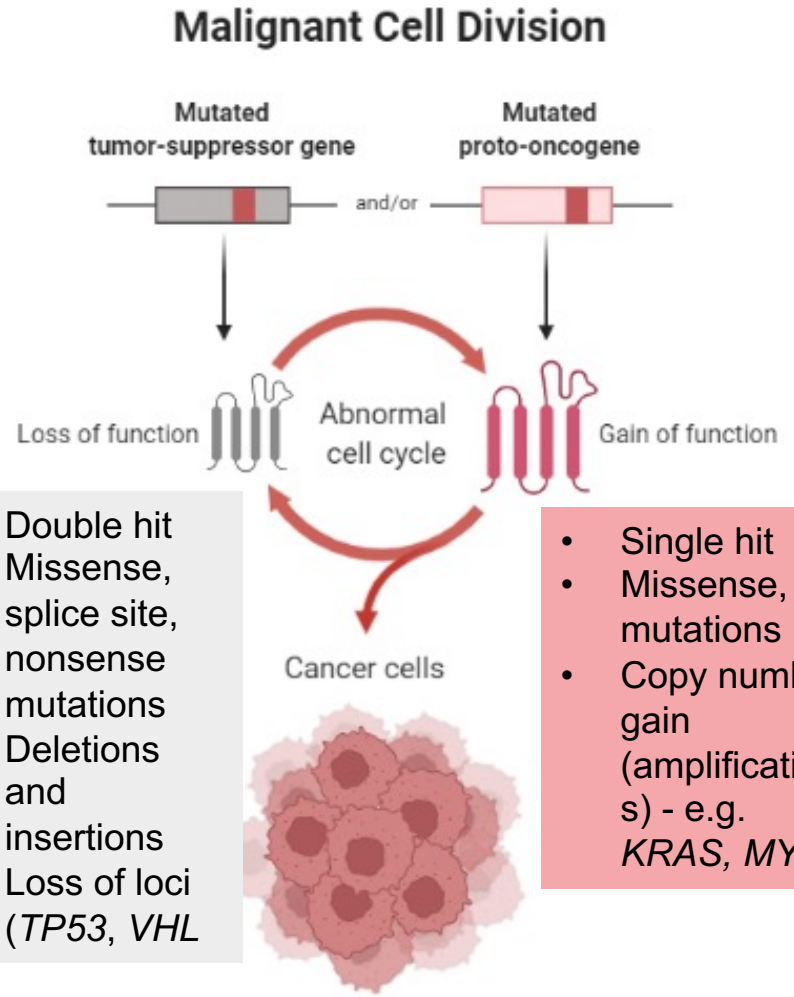
wellcome connecting science

Simple Model of Cancer Development



Concept Review: Oncogenes & Tumor Suppressors





- Double hit
- Missense, splice site, nonsense mutations
- Deletions and insertions
- Loss of loci (*TP53*, *VHL*

- Single hit
- Missense, mutations
- Copy number gain (amplifications) - e.g. *KRAS, MYC*

https://basicmedicalkey.com/the-molecular-biology-of-cancer/

What are drivers and passengers?

Drivers are *causal alterations* that enable the *hallmarks of cancer*

Substitutions and small indels – point mutations

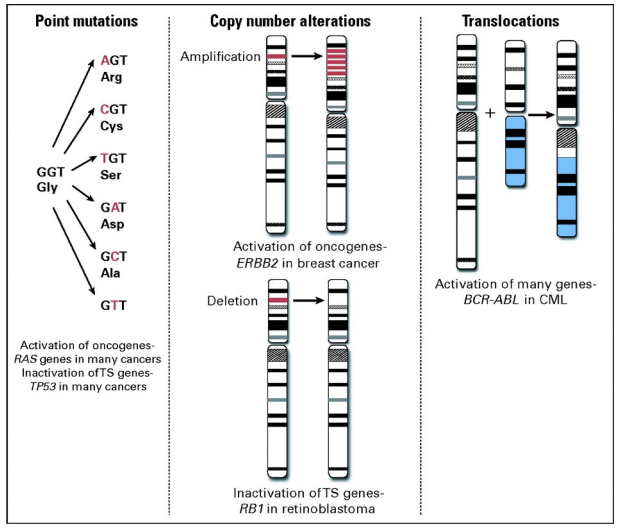
- Coding alterations: *KRAS* G12D
- Regulatory regions: *TERT* promoter

Structural rearrangements and copy number changes

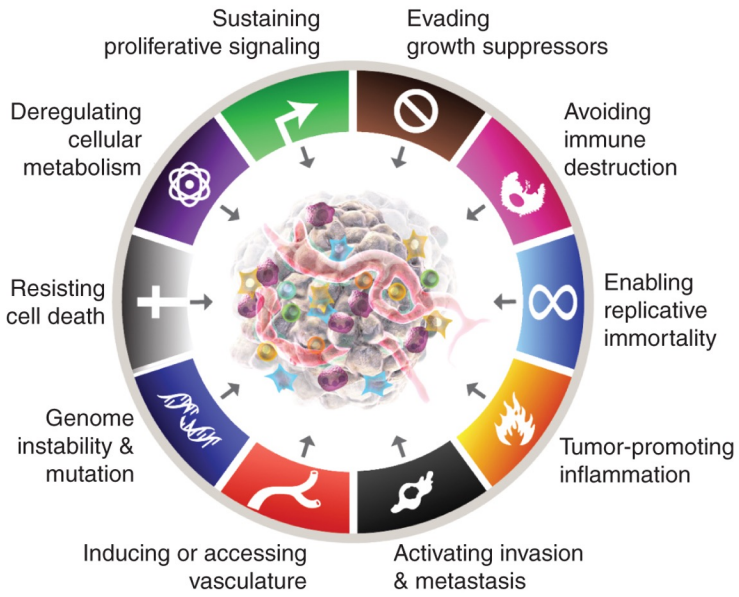
- BCR-ABL1 fusion in AML, MYC amplification, long deletions(*TP53*)

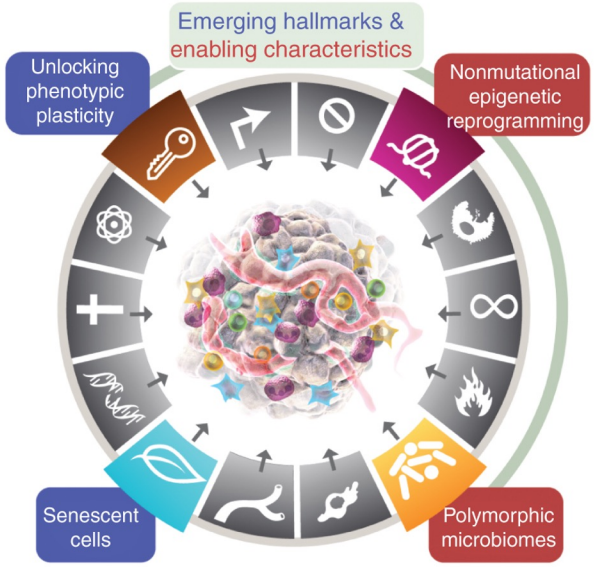
Epigenetics alterations

- *VHL expression repression through promoter hypermethylation*

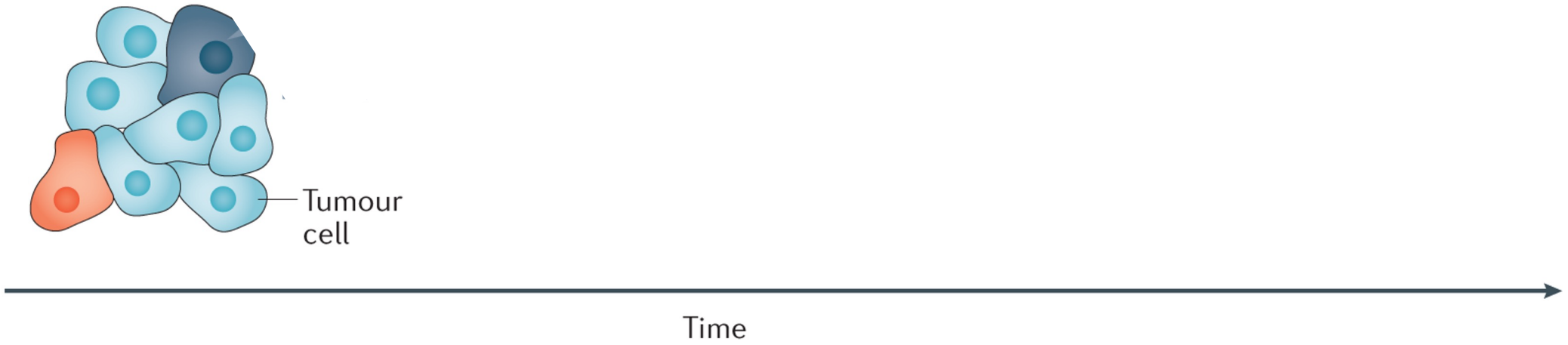


Macconaill LE, Garraway LA. J Clin Oncol. 2010

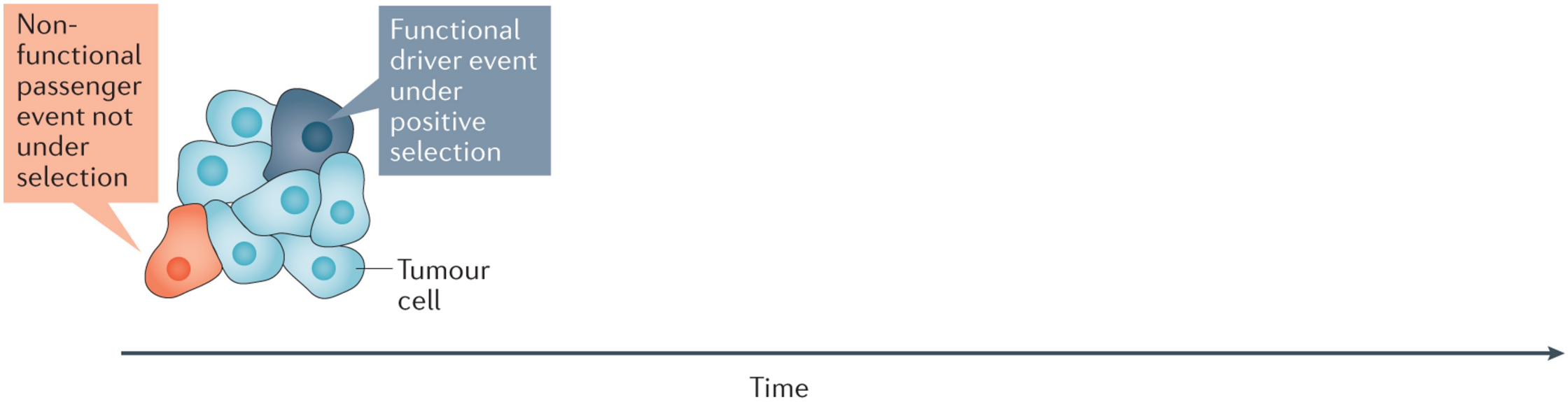




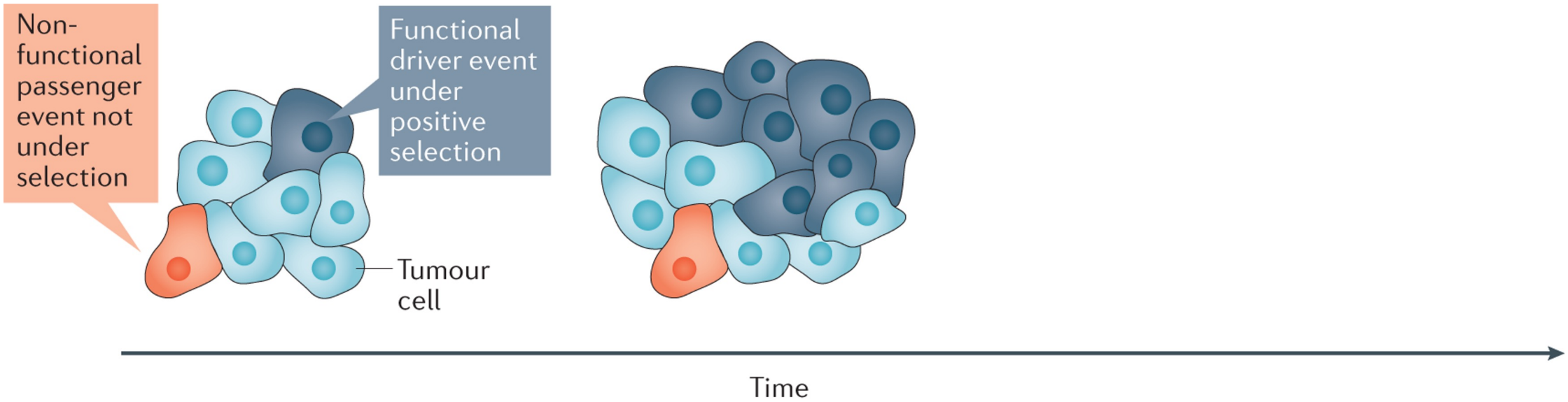
Hanahan D. Cancer Discov. 2022;12(1):31-46



Black JRM, McGranahan N. Nat Rev Cancer. 2021 Jun;21(6):379-392.

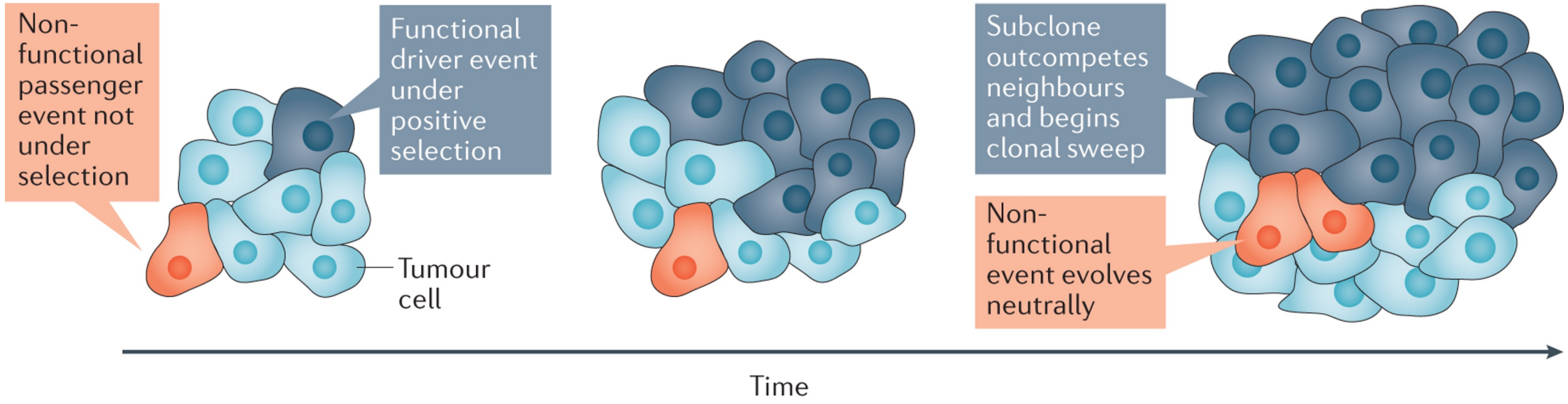


Black JRM, McGranahan N. Nat Rev Cancer. 2021 Jun;21(6):379-392.



Black JRM, McGranahan N. Nat Rev Cancer. 2021 Jun;21(6):379-392.

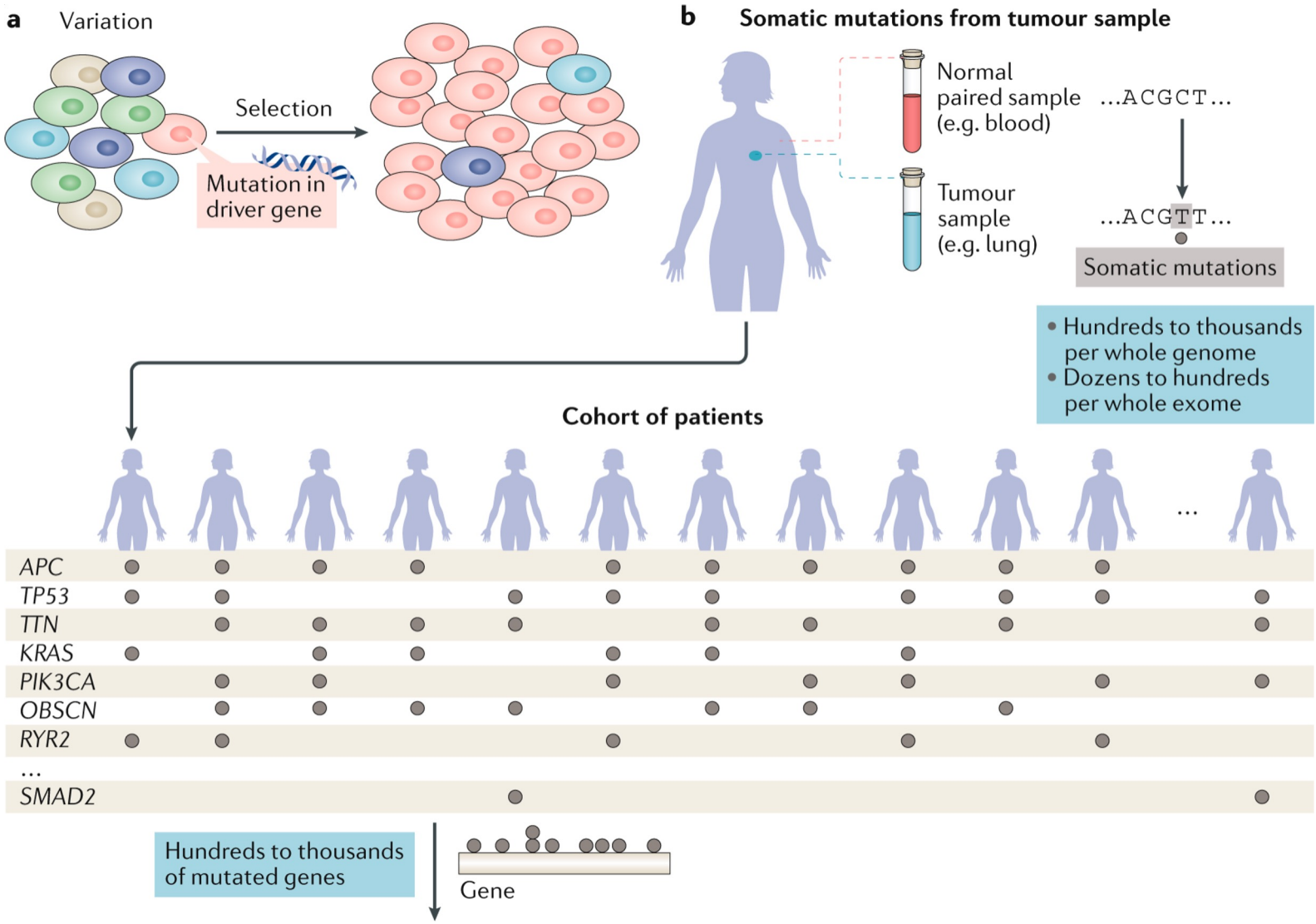
Functional and non-functional intra-tumor heterogeneity in tumor evolution



1. The increased rate of phenotypic variation in cancers compared with normal tissues means that new subclones arise and compete.
2. A minority contain a **driver event**, such as a genetic mutation or copy number alteration, that grants a **selective advantage**.
3. These subclones may grow at a faster rate than their neighbors and outcompete them in a **'selective sweep'.**

Black JRM, McGranahan N. Nat Rev Cancer. 2021 Jun;21(6):379-392.

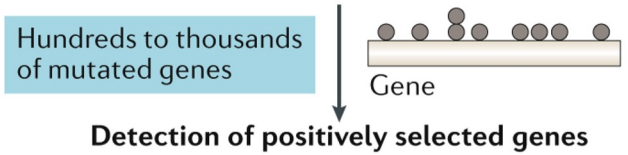
How do we find drivers in study cohorts?



Martínez-Jiménez F et. al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020 Oct;20(10):555-572.

Signals of positive selection identify driver genes

- Like finding needles in a haystack
- **Recurrence** – signature of positive selection



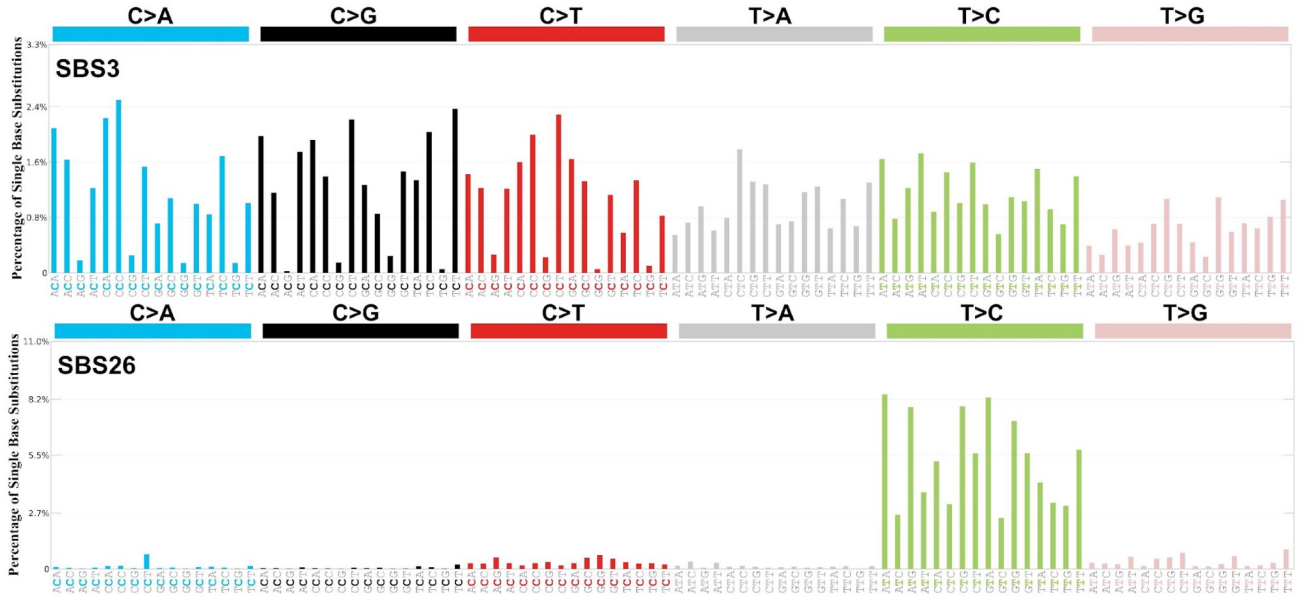
Key – proper modelling of the mutational process: Null (“expected”) model

Martínez-Jiménez F et. al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020 Oct;20(10):555-572.

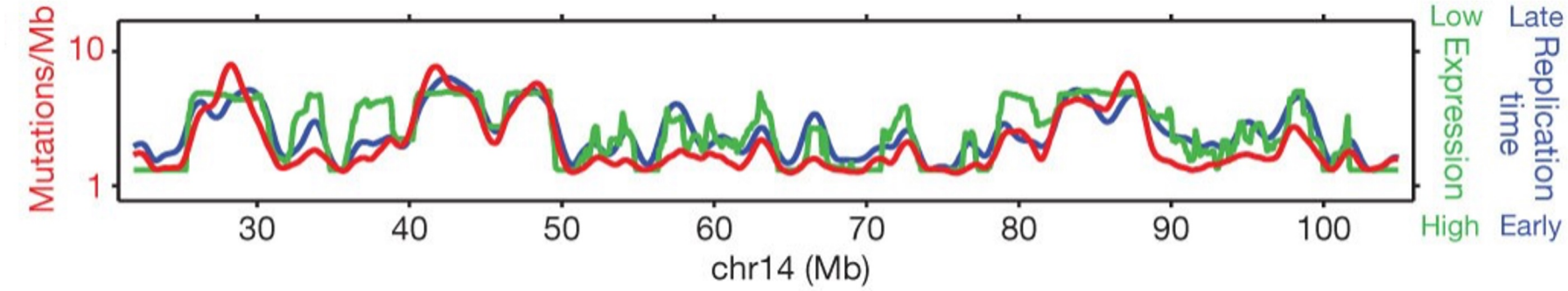
Modeling the mutational process

- Null model
- Difficult for:
  - Structural variants
    - Gistic
    - Gene fusions etc.
  - Epigenetic alterations
- Better understood for point mutations
  - Substitutions
  - Small indels

Most substitution mutational processes can be described using the tri-nucleotide context



Mutation rate varies widely across the genome and correlates with DNA replication time and expression level

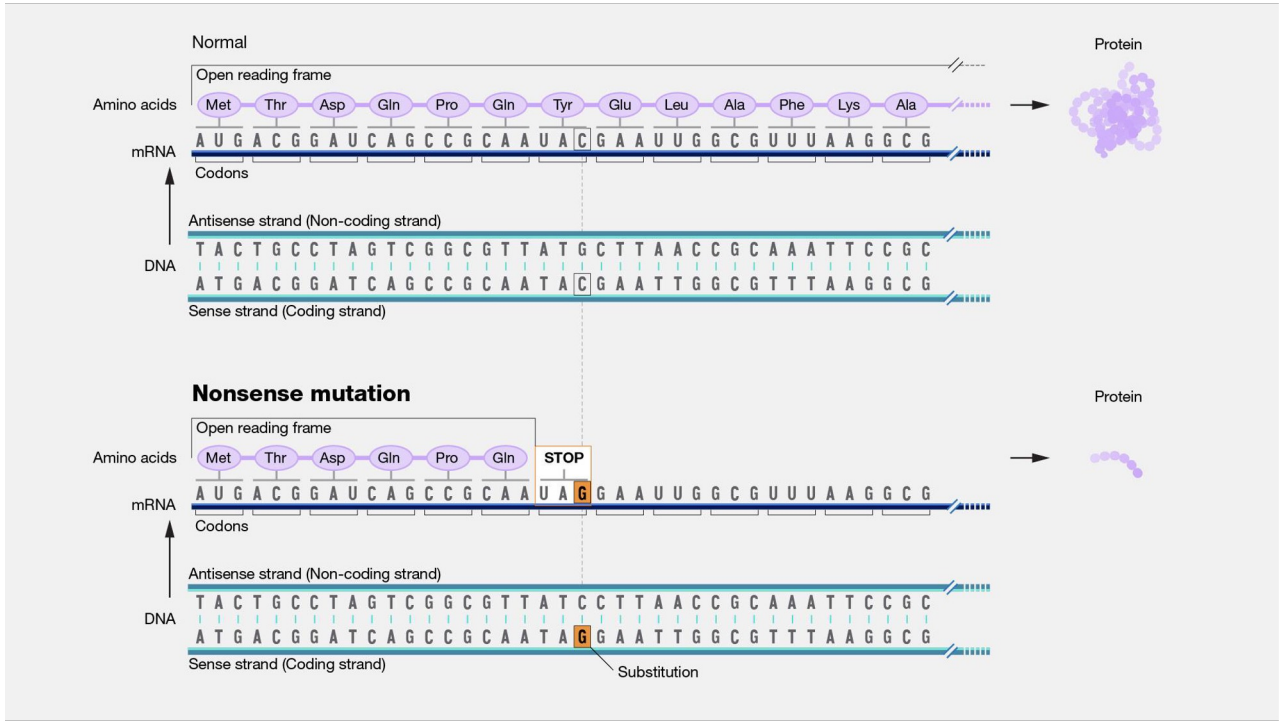


Lawrence et. al. Nature 2013

Concept Review: Types of point mutations – substitutions and indels

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Synonymous substitution, e.g. TAT > TAC (Tyr → Tyr)
Missense substitution, e.g. TAT > TGT (Tyr →Cys)
Indels: insertions/deletions, in frame vs out of frame



Nonsense substitution, e.g. TAT > TAA (Tyr → Stop*)

Identify driver genes: Estimate Selection Coefficients

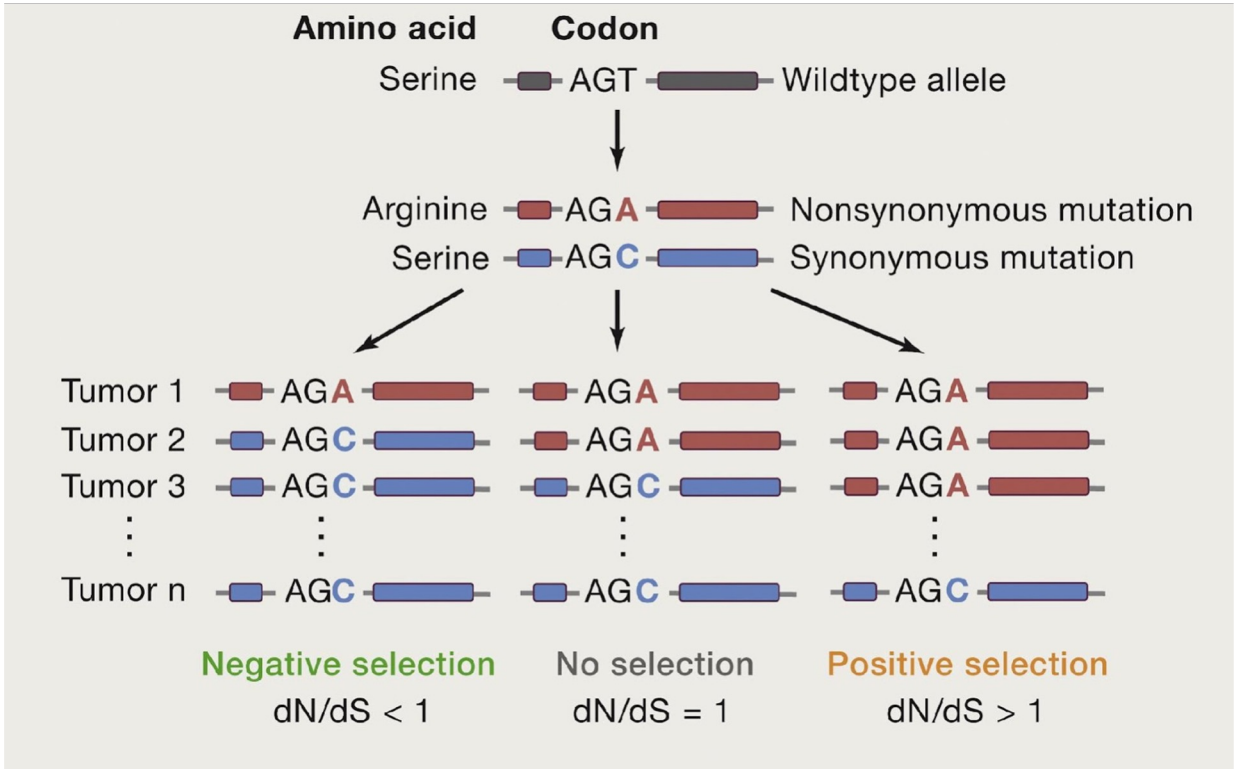
To formalize the observed versus expected test, estimate the coefficient selection ω=**dN/dS** where:

dN = number of **non-synonymous** substitutions

dS = number of **synonymous** substitutions

Non-synonymous substitutions:

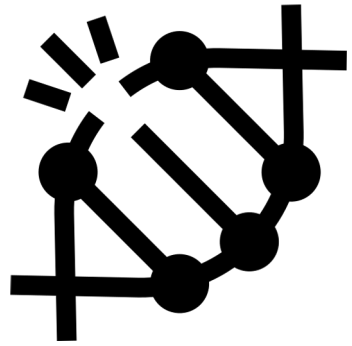
- *Missense (e.g. Leu → Pro)*
- *Nonsense (e.g. Ser → Stop)*
- *Splice sites*



Bakhoum SF, Landau DA. Cell. 2017 Nov 16;171(5):987-989.

Driver mutations

- Mutations in cancer can be classified into:
  - Driver mutations
  - Passenger mutations
- Driver mutations
  - Provide a selective advantage to the cell.
  - Promote cancer development
- Passenger mutations
  - Neutral mutations.

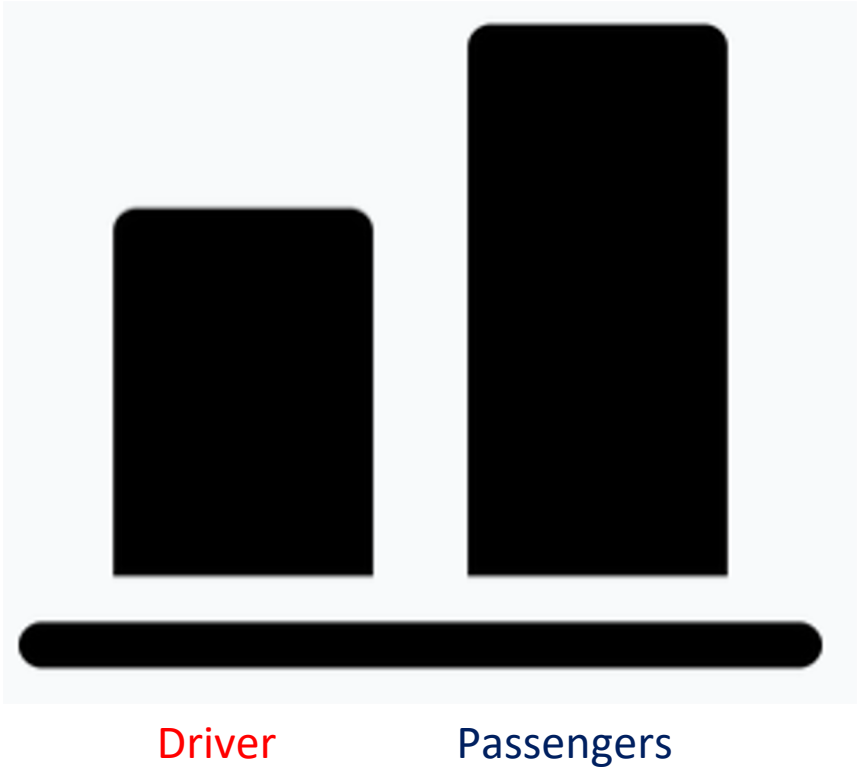


Driver genes

Genes that harbor driver mutations are called “**cancer driver genes**”.

Driver mutations

- Mutations in cancer can be classified into:
  - Driver mutations
  - Passenger mutations
- Driver mutations
  - Provide a selective advantage to the cell.
  - Promote cancer development
- Passenger mutations
  - Neutral mutations.



Oncogenes and tumor suppressor genes

- **Oncogenes**

  Genes that when altered can become “activated” (gain of function).

Missense mutations

Amplifications

- **Tumor suppressor genes**

  Genes that when altered they stop working properly (loss of function)

Missense mutations

Splice site mutations

Nonsense mutations

Deletions

Oncogenes and tumor suppressor genes

- **Oncogenes**

  Genes that when altered can become “activated” (gain of function).

KRAS

BRAF

- **Tumor suppressor genes**

  Genes that when altered they stop working properly (loss of function)

TP53

Oncogenes and tumor suppressor genes

- **Oncogenes**

  Genes that when altered c
become “activated” (gain of function).

MAPK PATHWAY
-->
Cell proliferation

KRAS

BRAF

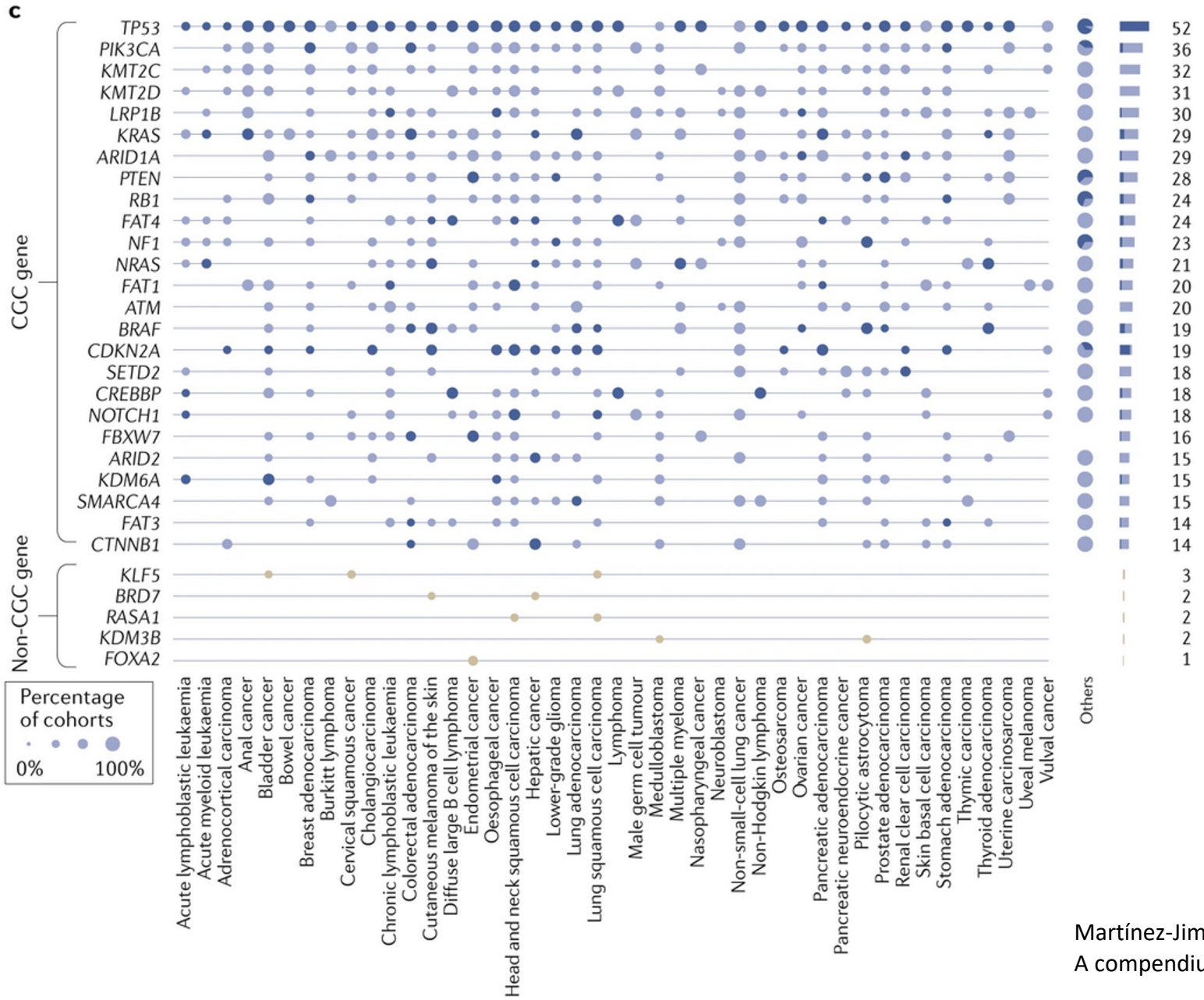
- **Tumor suppressor genes**

  Genes that when altered the
stop working properly (loss
function)

Regulation of cell division - apoptosis

TP53

Mutational driver genes

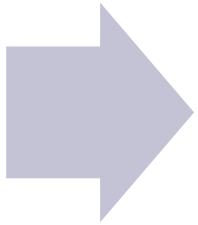


Martínez-Jiménez et al., 2020
A compendium of mutational cancer driver genes.

Research on cancer drivers

Cancer patient cohorts studies

Drivers

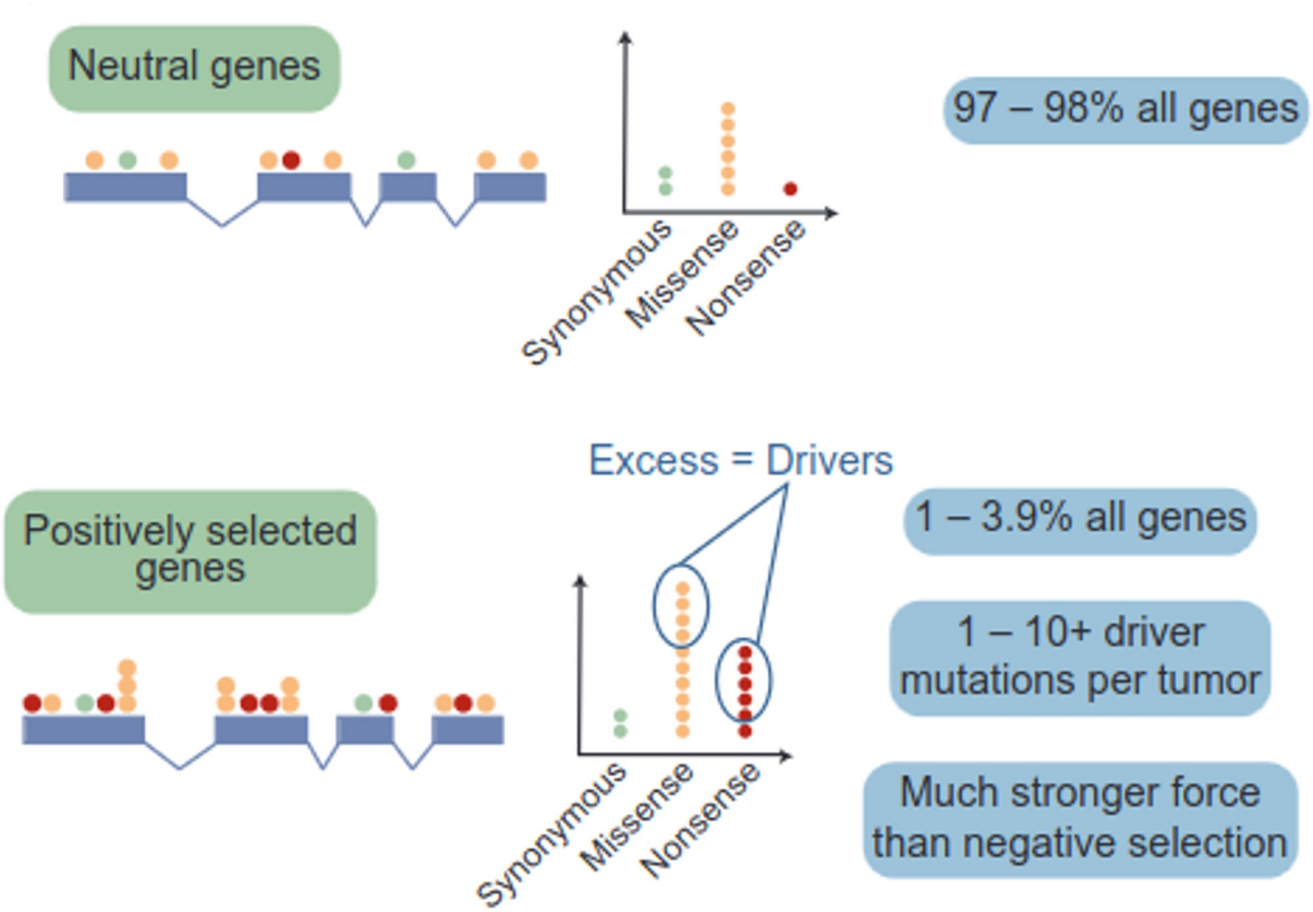


Targeted cancer therapies

Biomarkers of prognosis

Biomarkers of response to treatments

Driver genes are under positive selection in tumors



Martincorena et al., 2017

Finding drivers in cohorts

- Recurrence --> signature of positive selection
- Requires proper modelling of the mutational process.

The null (expected) model

Difficult for --> structural variants

Better for --> point mutations (substitutions and small indels)

Finding drivers in cohorts

- dN/dS --> non-synonymous / synonymous substitutions --> observed / expected
    - dN/dS < 1 Negative selection
    - dN/dS = 1 Neutral selection
    - dN/dS > 1 Positive selection

- Interpretation
    - dN/dS of 1 --> All observed non-syn are expected
    - dN/dS of 2 --> 50% of non-syn are selected
    - dN/dS of 10 --> 90% of non-syn are selected
    - dN/dS of 100 --> 99% of non_syn are selected

w = dN/dS (coef. of selection)
(w-1)/w = fraction mutations selected

dndscv R package

Identifies positively selected genes using dN/dS model

Mutational model:

Trinucleotide frequencies

Covariates for regional variation

Generates selection estimates at

- Gene level (or domains)
- Global
- Sites and codons
- Not only for cancer

https://github.com/im3sanger/dndscv

Martincorena et al, Cell 2017

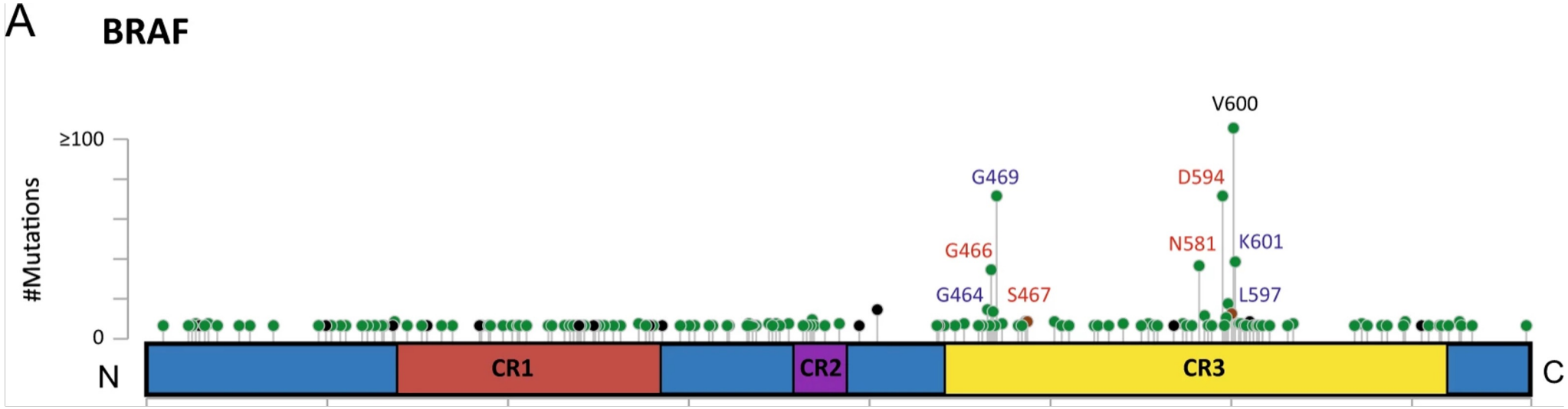
Hotspots

Recurrently mutated positions (nucleotide or amino acid).

Examples:

BRAF V600

- Melanoma
- Colorectal cancer
- Lymphoma



Danker et al., 2018

Hotspots

Recurrently mutated positions (nucleotide or amino acid).

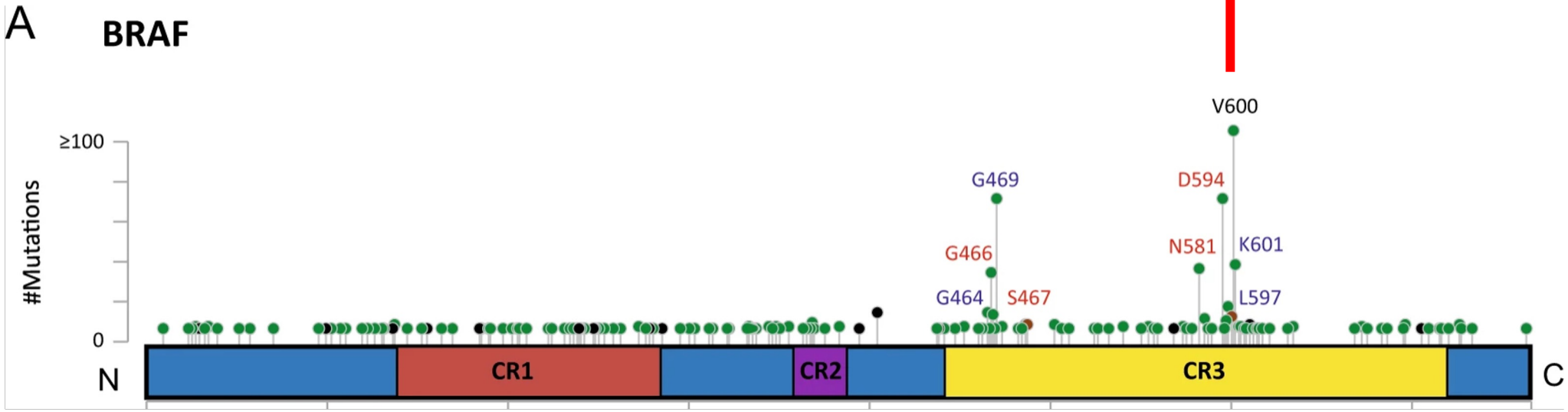
Examples:

BRAF V600

- Melanoma
- Colorectal cancer
- Lymphoma

Vemurafenib

Dabrafenib



Danker et al., 2018

Structural drivers

- Copy number gains / loses
- Gene fusions
- Rearrangements

Challenges

- Modelling the background
- Involve multiple chromosomes
- Require long read sequencing

Non-coding somatic drivers

- lncRNAs
- UTRs/promoters
- tRNAs
- small RNAs
- micro RNAs

Example:

TERT promoter

Challenges

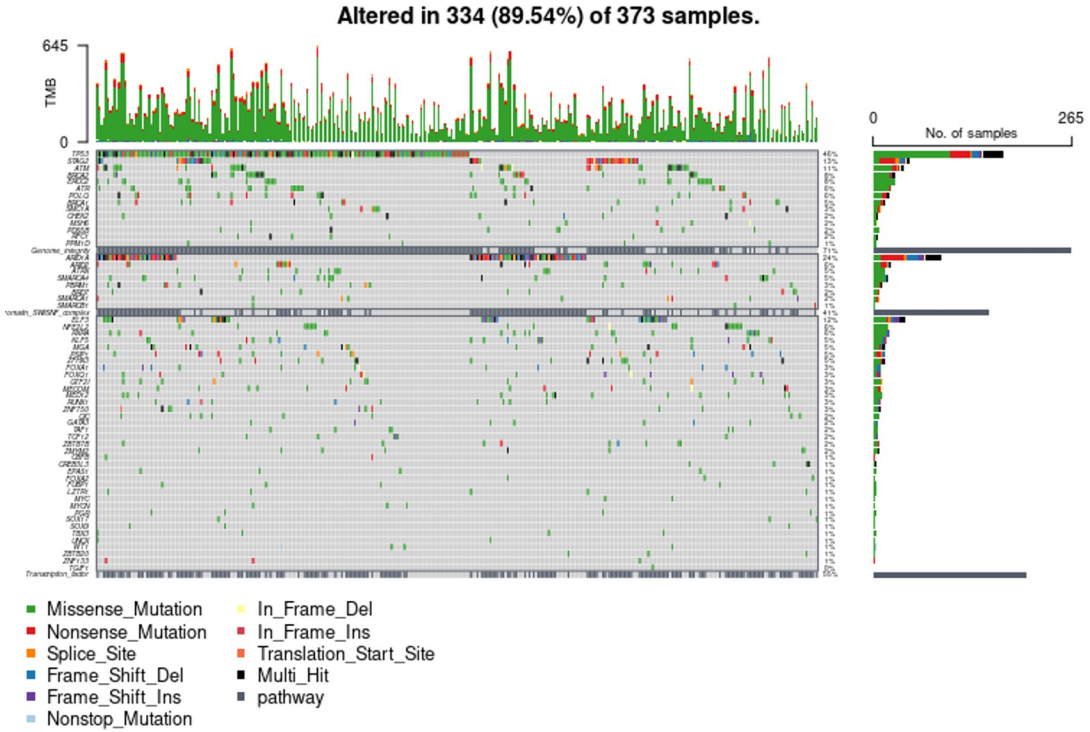
- Sequencing and mapping artefacts
- Incomplete annotation of regulatory regions
- Unknown functional effects

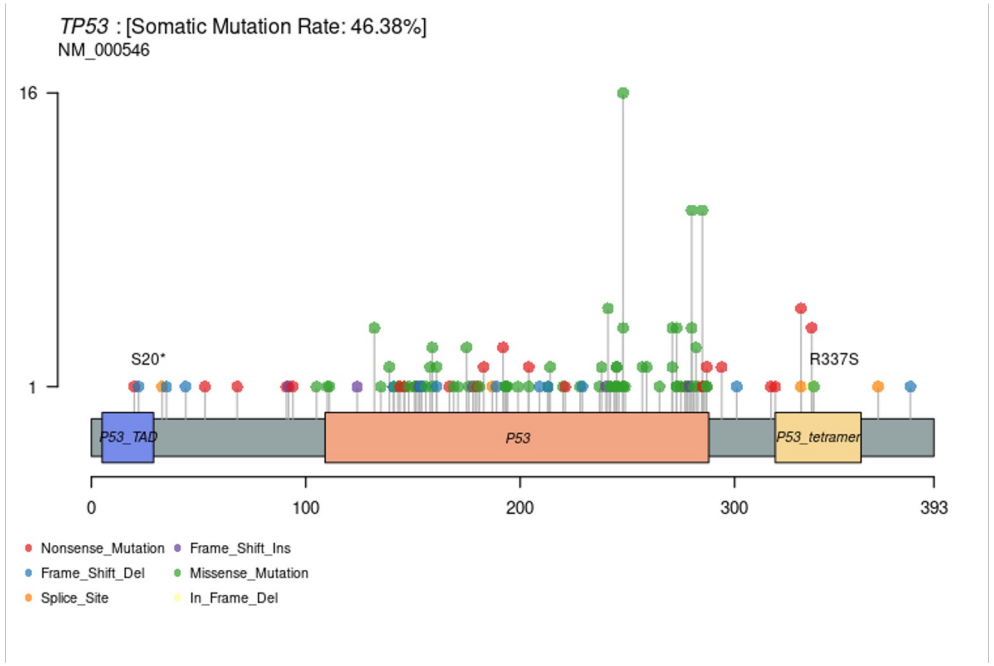
Less frequent compared to protein coding drivers (Rheinbay et at., 2020)

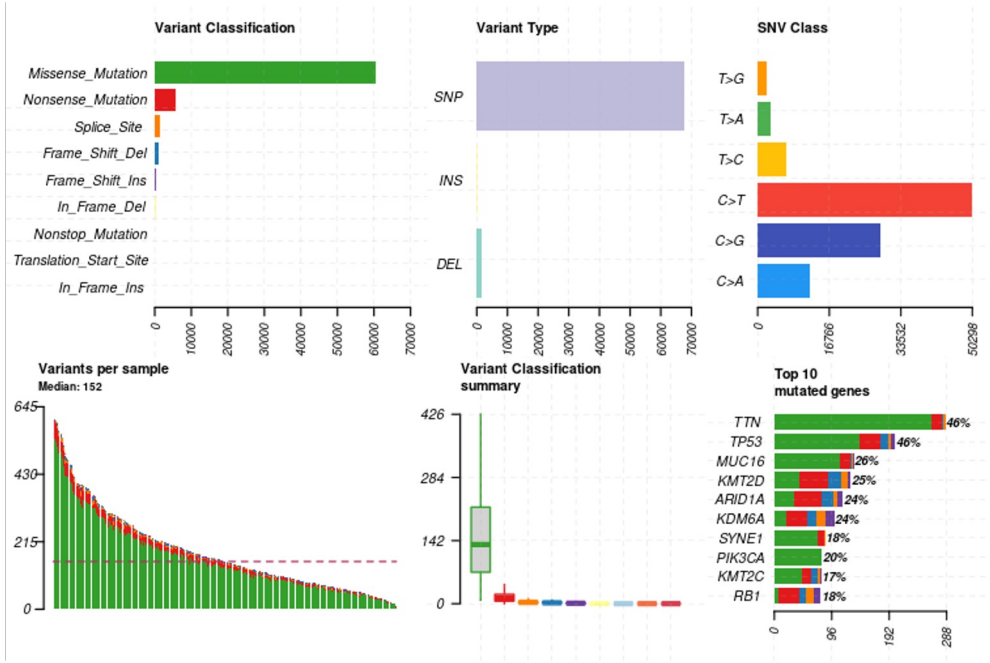
Oncoplots

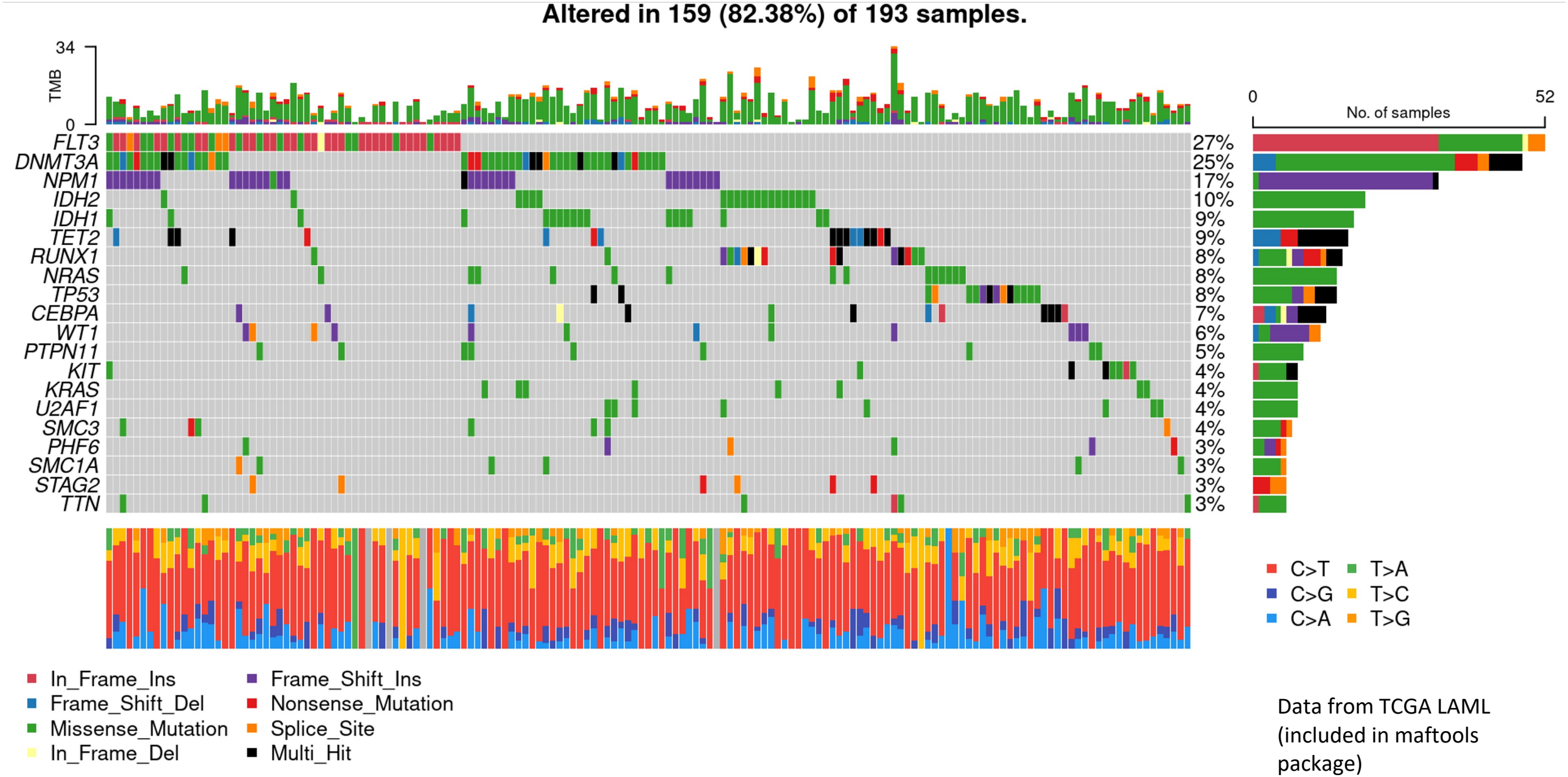
maftools

- R package
- Work with data stored in MAF
- Different type of visualizations



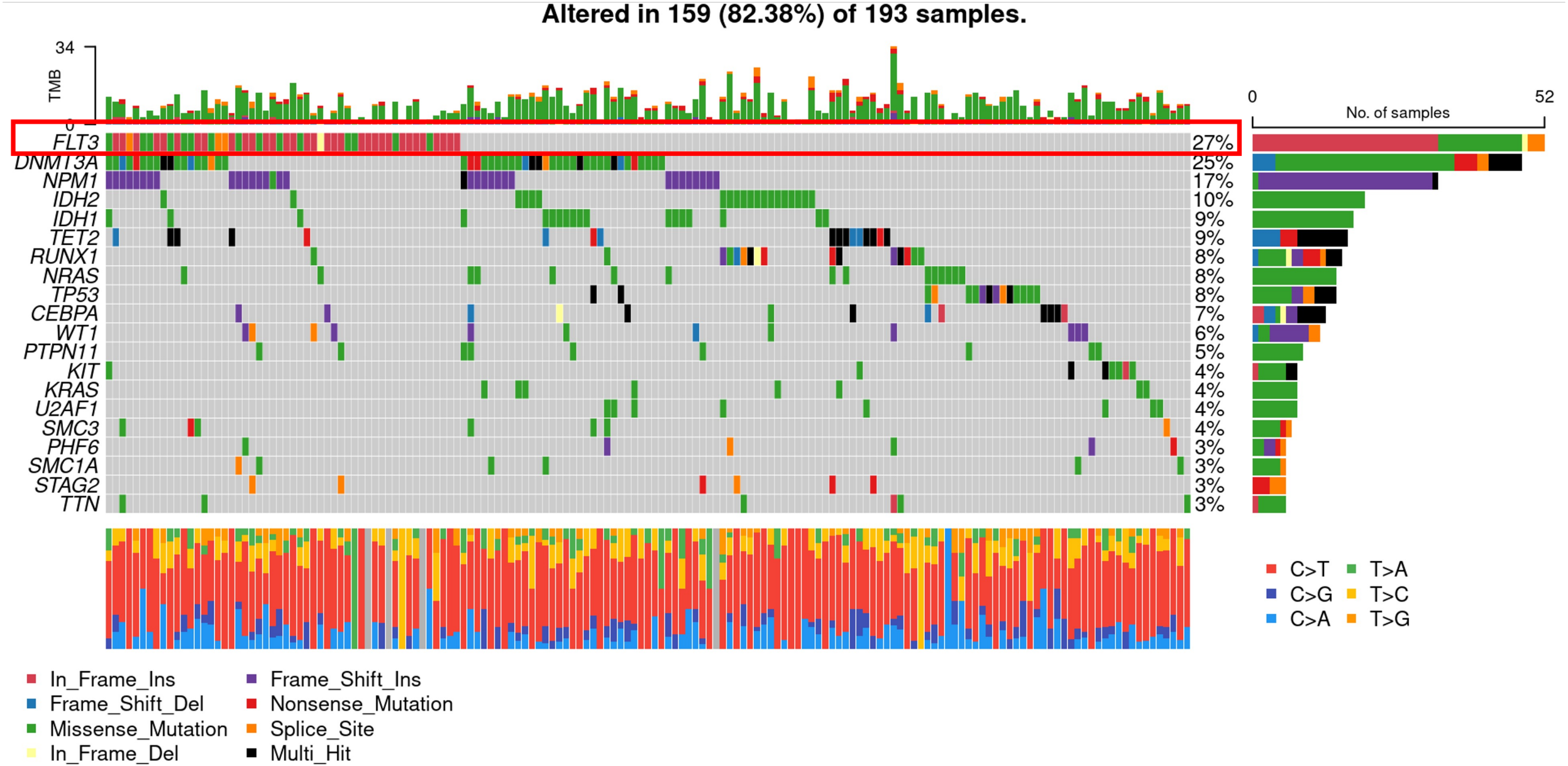


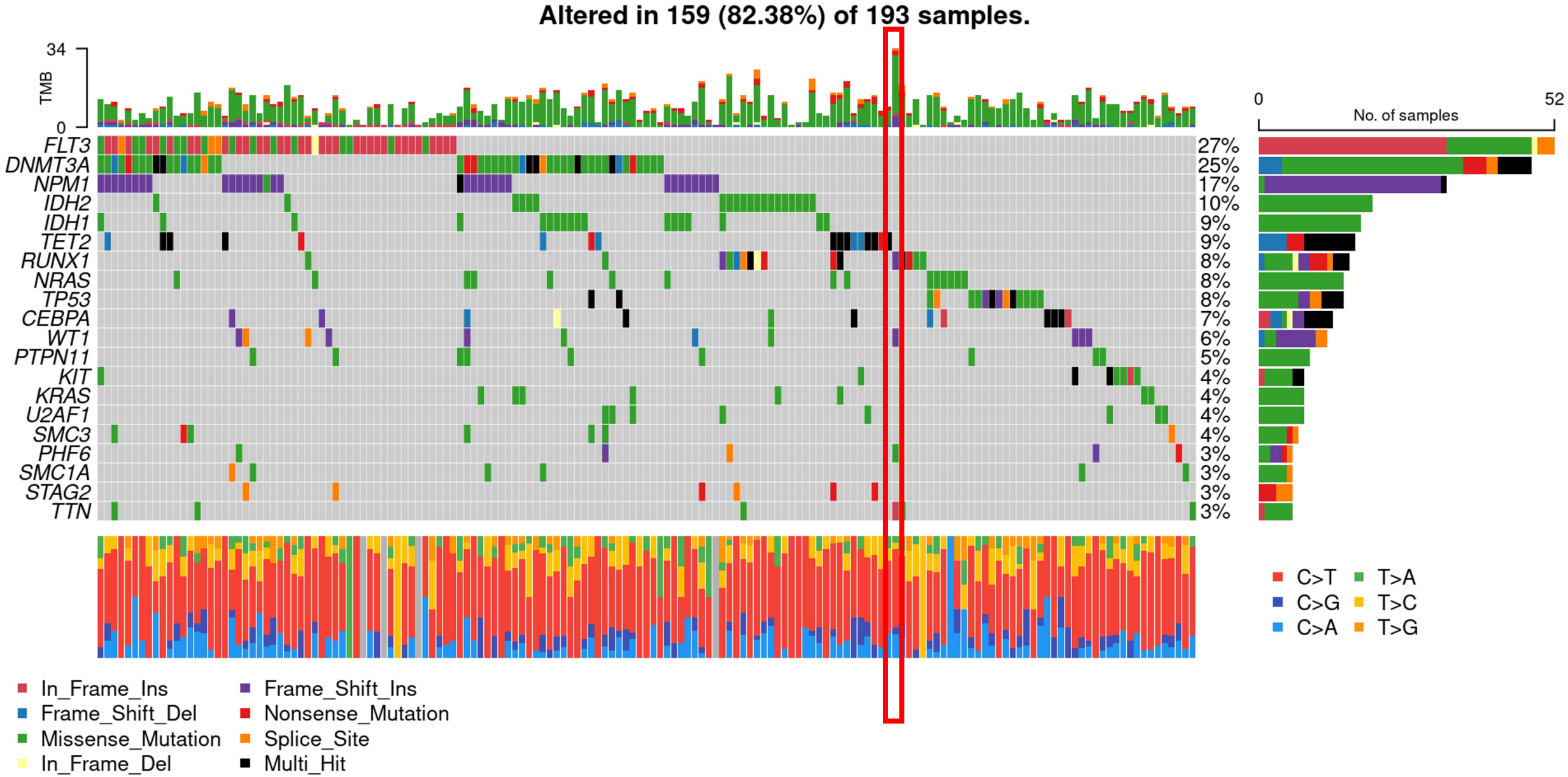




Data from TCGA LAML (included in maftools package)

Mayaconda, A, 2022


Altered in 159 (82.38%) of 193 samples.
TMB
34
0
FLT3 27%
DNMT3A 25%
NPM1 17%
IDH2 10%
IDH1 9%
TET2 9%
RUNX1 8%
NRAS 8%
TP53 8%
CEBPA 7%
WT1 6%
PTPN11 5%
KIT 4%
KRAS 4%
U2AF1 4%
SMC3 4%
PHF6 3%
SMC1A 3%
STAG2 3%
TTN 3%
0
52
No. of samples
C>T
C>G
C>A
T>A
T>C
T>G
In_Frame_Ins
Frame_Shift_Del
Missense_Mutation
In_Frame_Del
Frame_Shift_Ins
Nonsense_Mutation
Splice_Site
Multi_Hit


Altered in 159 (82.38%) of 193 samples.
TMB
34
0
FLT3 27%
DNMT3A 25%
NPM1 17%
IDH2 10%
IDH1 9%
TET2 9%
RUNX1 8%
NRAS 8%
TP53 8%
CEBPA 7%
WT1 6%
PTPN11 5%
KIT 4%
KRAS 4%
U2AF1 4%
SMC3 4%
PHF6 3%
SMC1A 3%
STAG2 3%
TTN 3%
0
52
No. of samples
C>T
C>G
C>A
T>A
T>C
T>G
In_Frame_Ins
Frame_Shift_Del
Missense_Mutation
In_Frame_Del
Frame_Shift_Ins
Nonsense_Mutation
Splice_Site
Multi_Hit

Practical - dndscv

Input file

Five column table of mutations

1) sampleID
2) chromosome
3) position
4) reference
5) mutant

```
##   sampleID chr      pos ref mut
## 1 Sample_1   1   871244   G   C
## 2 Sample_1   1  6648841   C   G
## 3 Sample_1   1 17557072   G   A
## 4 Sample_1   1 22838492   G   C
## 5 Sample_1   1 27097733   G   A
## 6 Sample_1   1 27333206   G   A
```

Outputs

- List of objects in R
    - Usually the most important object is the neutrality test results (**dndsout$sel**). This table has information on the number of substitutions for each class, dN/dS ratios and global p and q values.

```
sel_cv = dndsout$sel_cv
print(head(sel_cv), digits = 3)
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind wmis_cv wnon_cv wspl_cv
## 18057      TP53     1    43     5     4     5  113.66   221.8   221.8
## 12977    PIK3CA     3    34     0     0     3   29.75     0.0     0.0
## 9225       KRAS     1    21     0     0     0  125.98     0.0     0.0
## 18924       VHL     3     9     1     0     4   24.75    38.3    38.3
## 1296        APC     2     8    10     0     6    2.82    31.2    31.2
## 1465     ARID1A     1     7    10     0     3    3.30    53.5    53.5
##       wind_cv  pmis_cv ptrunc_cv pallsubs_cv  pind_cv qmis_cv qtrunc_cv
## 18057   138.8 0.00e+00  1.11e-16    0.00e+00 1.73e-09 0.00000  2.23e-12
## 12977    30.7 0.00e+00  5.68e-01    0.00e+00 2.12e-04 0.00000  9.47e-01
## 9225      0.0 0.00e+00  8.44e-01    0.00e+00 1.00e+00 0.00000  9.47e-01
## 18924   204.5 7.46e-09  2.23e-02    7.97e-09 1.32e-08 0.00003  9.47e-01
## 1296     23.1 4.96e-02  2.81e-10    2.12e-09 1.81e-06 0.77064  1.41e-06
## 1465     14.4 3.64e-02  2.96e-12    2.38e-11 1.84e-03 0.77064  2.97e-08
##       qallsubs_cv pglobal_cv qglobal_cv
## 18057    0.00e+00   0.00e+00   0.00e+00
## 12977    0.00e+00   0.00e+00   0.00e+00
## 9225     0.00e+00   0.00e+00   0.00e+00
## 18924    2.00e-05   4.00e-15   2.01e-11
## 1296     6.07e-06   1.31e-13   5.26e-10
## 1465     1.20e-07   1.39e-12   4.67e-09
```

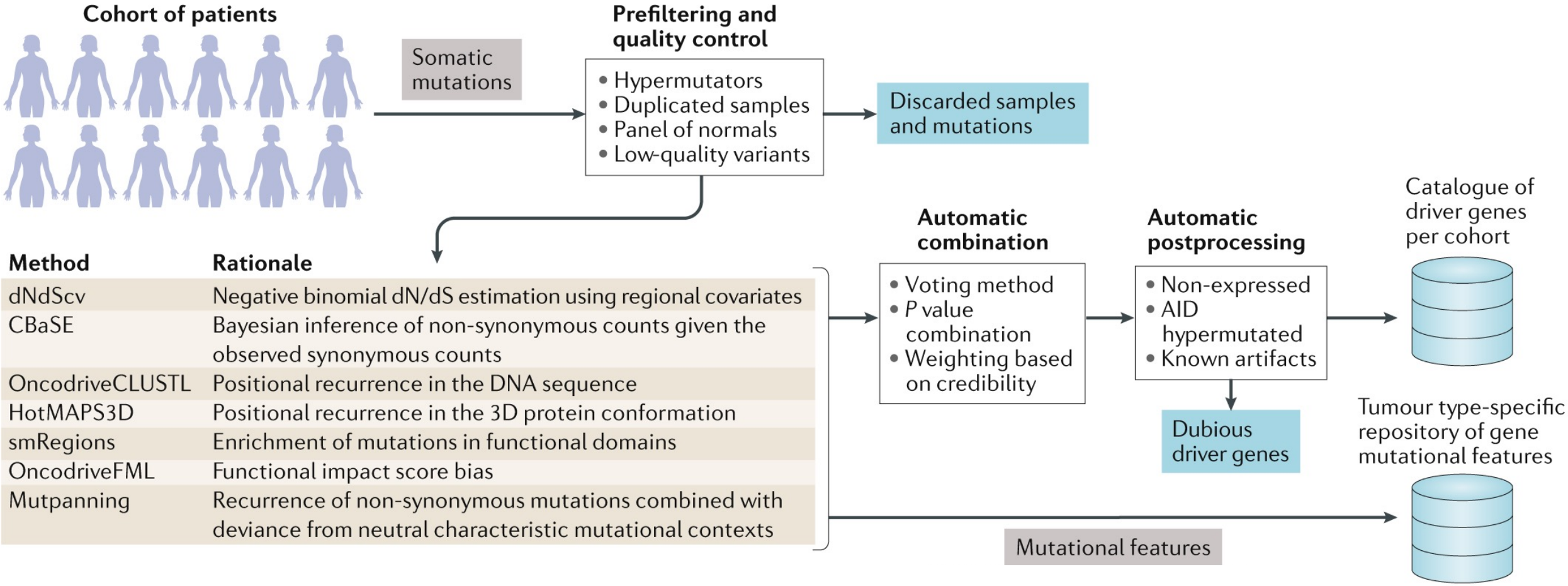
Take Home Messages

- Oncogenes and tumor suppressors behave very differently – no clear cut rules of thumb in Biology
- Passengers >>>>>>>> drivers (hypermutators are particularly problematic)
- Recurrence = positive selection (obs > exp)
- Most drivers are protein coding (+ Tert) - ~ 1% of the genome
- Not all non-synonymous mutations are drivers (dN/dS)
- Structural and epigenetic alterations can be drivers too

Cancer Driver Callers NOT EXHAUSTIVE	Basis/Rationale	Reference
ddscnv	Negative binomial dN/dS estimation using regional covariates	Martincorena, I. et al. **Universal Patterns of Selection in Cancer and Somatic Tissues.** Cell 171, 1029-1041.e21 (2017). doi: 10.1016/j.cell.2017.09.042
2020+		Collin J. Tokheim, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin. **Evaluating the evaluation of cancer driver genes.** PNAS 2016 ; published ahead of print November 22, 2016, doi:10.1073/pnas.1616440113
CBaSE	Bayesian inference of non-synonymous counts given the observed synonymous counts	Weghorn D, Sunyaev S. **Bayesian inference of negative and positive selection in human cancers**. Nat Genet. 2017 Dec;49(12):1785-1788. doi: 10.1038/ng.3987. Epub 2017 Nov 6. PMID: 29106416.
HotSpot3D	3D proximity tool can be used to identify mutation hotspots from linear protein sequence and correlate the hotspots with known or potentially interacting domains, mutations, or drugs.	Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, Wyczalkowski MA, Liang WW, Zhang Q, McLellan MD, Sun SQ, Tripathi P, Lou C, Ye K, Mashl RJ, Wallis J, Wendl MC, Chen F, Ding L. **Protein-structure-guided discovery of functional mutations across 19 cancer types.** Nat Genet. 2016 Aug;48(8):827-37. doi: 10.1038/ng.3586. Epub 2016 Jun 13. Erratum in: Nat Genet. 2017 Jul 27;49(8):1286. PMID: 27294619; PMCID: PMC5315576.
HotMAPS3D	Positional recurrence in the 3D protein conformation	Tokheim C, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer research. 2016a;76:3719–3731. doi: 10.1158/0008-5472.CAN-15-3190
Mutpanning	recurrence of non-synonymous mutations combined with deviance from neutral characteristic mutational contexts	Dietlein, F., Weghorn, D., Taylor-Weiner, A. et al**. Identification of cancer driver genes based on nucleotide context.** Nat Genet (2020). https://doi.org/10.1038/s41588-019-0572-y
OncodriveCLUSTL	a sequence-based clustering method to identify significant clustering signals in nucleotide sequence	Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. Bioinformatics. 2019 Nov 1;35(22):4788-4790. doi: 10.1093/bioinformatics/btz501. Erratum in: Bioinformatics. 2019 Dec 15;35(24):5396. PMID: 31228182; PMCID: PMC6853674.
smRegions	Enrichment of mutations in functional domains	Francisco Martínez-Jiménez, et al**. Disruption of ubiquitin mediated proteolysis is a widespread mechanism of tumorigenesi**s. bioRxiv 2019. doi: https://doi.org/10.1101/507764
OncodriveFML	Functional impact score bias	Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. **OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations**. Genome Biol. 2016 Jun 16;17(1):128. doi: 10.1186/s13059-016-0994-0. PMID: 27311963; PMCID: PMC4910259.

Consensus Driver Identification Pipeline: IntOGen

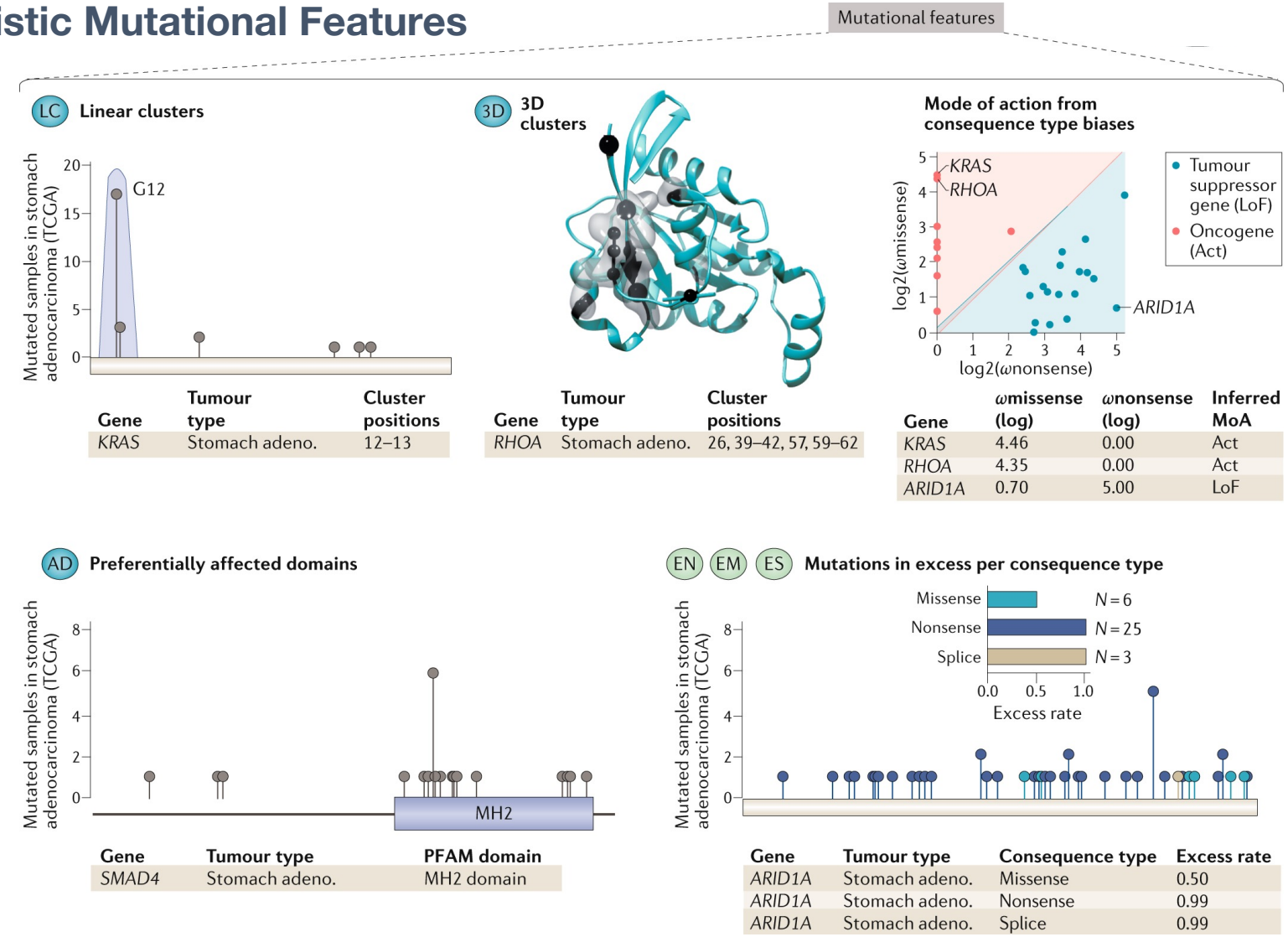
Schematic representation of the Integrative OncoGenomics (IntOGen) pipeline



Martínez-Jiménez F et. al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020 Oct;20(10):555-572.

Consensus Driver Identification Pipeline: IntOGen

Characteristic Mutational Features



Gene	Tumour type	Cluster positions
KRAS	Stomach adeno.	12–13

Gene	Tumour type	Cluster positions
RHOA	Stomach adeno.	26, 39–42, 57, 59–62

Gene	ωmissense (log)	ωnonsense (log)	Inferred MoA
KRAS	4.46	0.00	Act
RHOA	4.35	0.00	Act
ARID1A	0.70	5.00	LoF

Gene	Tumour type	PFAM domain
SMAD4	Stomach adeno.	MH2 domain

Gene	Tumour type	Consequence type	Excess rate
ARID1A	Stomach adeno.	Missense	0.50
ARID1A	Stomach adeno.	Nonsense	0.99
ARID1A	Stomach adeno.	Splice	0.99

Martínez-Jiménez F et. al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020 Oct;20(10):555-572.

References

1. Nowell PC. **The clonal evolution of tumor cell populations.** Science. 1976 Oct 1;194(4260):23-8. doi: 10.1126/science.959840. PMID: 959840. *The first paper to describe the evolution dynamics in cancer.*

2. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, Gonzalez-Perez A, Lopez-Bigas N. **A compendium of mutational cancer driver genes.** Nat Rev Cancer. 2020 Oct;20(10):555-572. doi: 10.1038/s41568-020-0290-x. Epub 2020 Aug 10. PMID: 32778778.

3. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. **Universal Patterns of Selection in Cancer and Somatic Tissues.** Cell. 2018 Jun 14;173(7):1823. doi: 10.1016/j.cell.2018.06.001. Erratum for: Cell. 2017 Nov 16;171(5):1029-1041.e21. PMID: 29906452; PMCID: PMC6005233.

4. Rheinbay E,……; PCAWG Drivers and Functional Interpretation Working Group; PCAWG Structural Variation Working Group; Weischenfeldt J, Beroukhim R, Martincorena I, Pedersen JS, Getz G; PCAWG Consortium. **Analyses of non-coding somatic drivers in 2,658 cancer whole genomes.** Nature. 2020 Feb;578(7793):102-111. doi: 10.1038/s41586-020-1965-x. Epub 2020 Feb 5. Erratum in: Nature. 2023 Feb;614(7948):E40. PMID: 32025015; PMCID: PMC7054214.

5. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. **IntOGen-mutations identifies cancer drivers across tumor types.** Nat Methods. 2013 Nov;10(11):1081-2. doi: 10.1038/nmeth.2642. Epub 2013 Sep 15. PMID: 24037244; PMCID: PMC5758042.

6. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, de Torres C, Dienstmann R, Gonzalez-Perez A, Lopez-Bigas N. **Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations**. Genome Med. 2018 Mar 28;10(1):25. doi: 10.1186/s13073-018-0531-8. PMID: 29592813; PMCID: PMC5875005.

7. Bailey MH,……. MC3 Working Group…….. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018 Apr 5;173(2):371-385.e18. doi: 10.1016/j.cell.2018.02.060. Erratum in: Cell. 2018 Aug 9;174(4):1034-1035. PMID: 29625053; PMCID: PMC6029450.