Logic

Data larger
Key is prioritizing
How to do this ?
Key aspect is driver v pass.

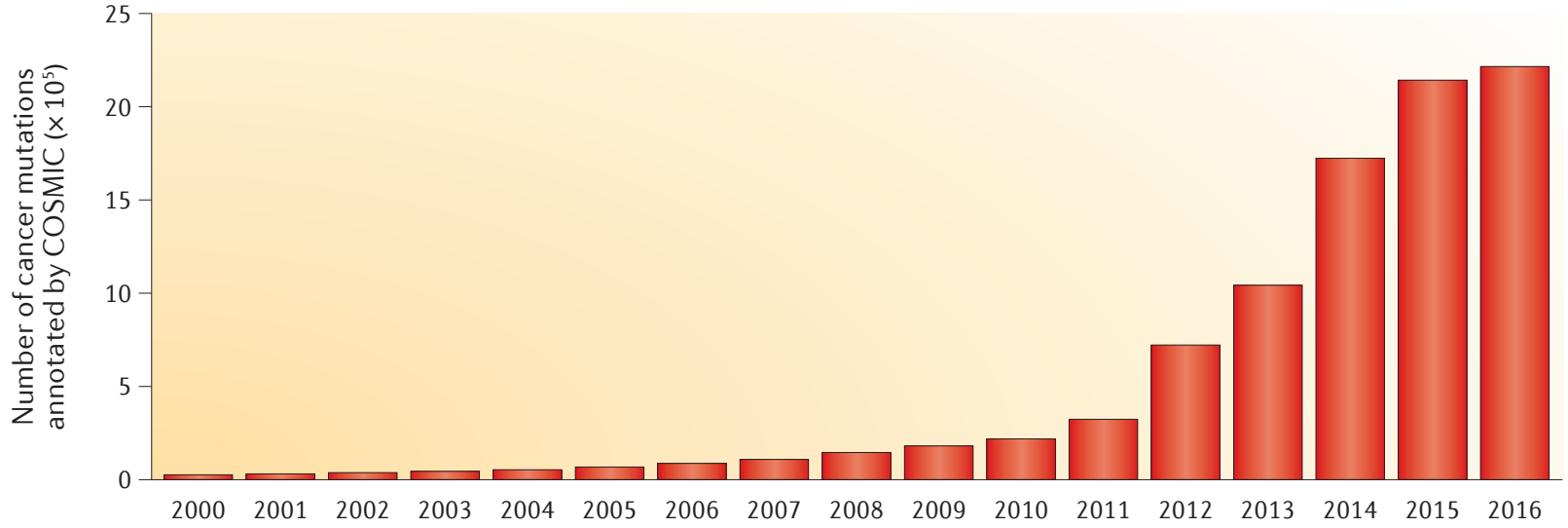
Logic

Data larger

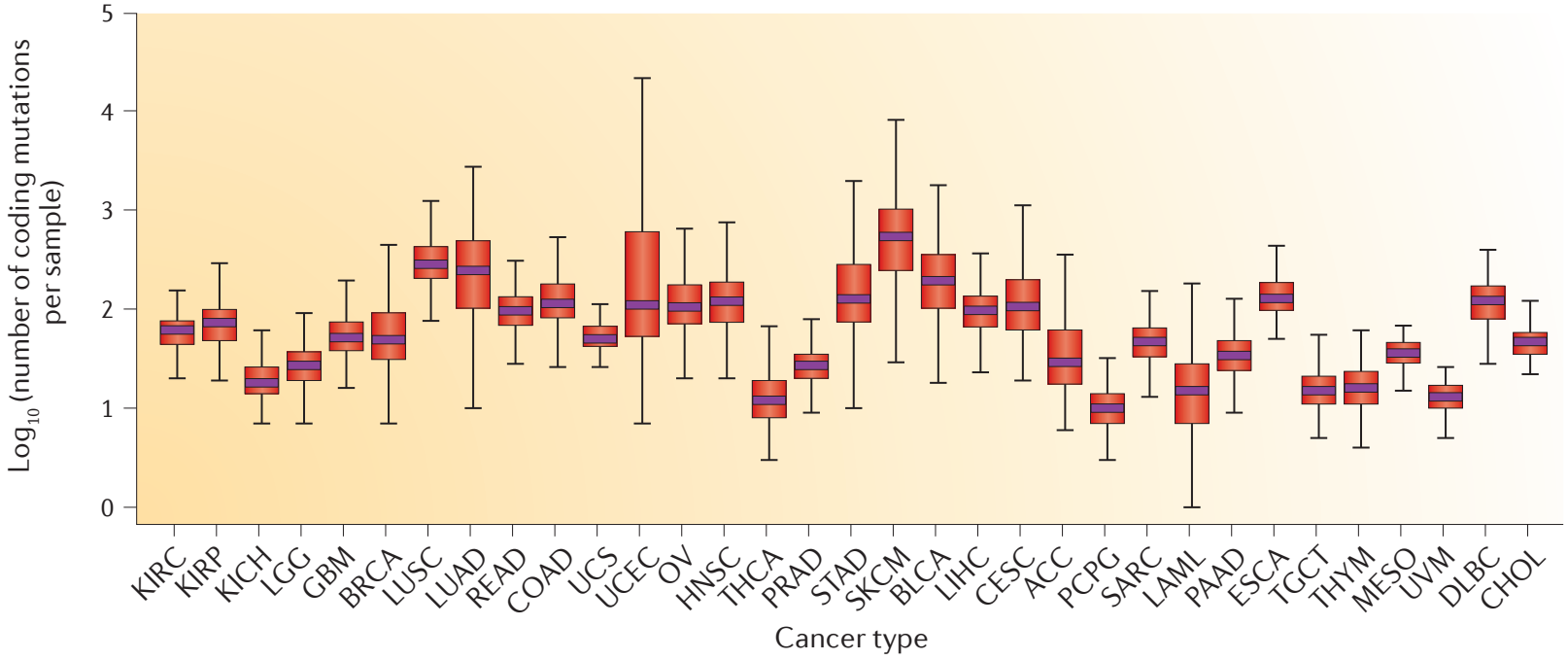
Key is prioritizing

How to do this ?

Key aspect is driver v pass.



Mutations downloaded from COSMIC



Somatic variants in 33 cancer types

Data taken from TCGA (data from 10,489 tumor samples shown)

Yi et al, NRG (2017)

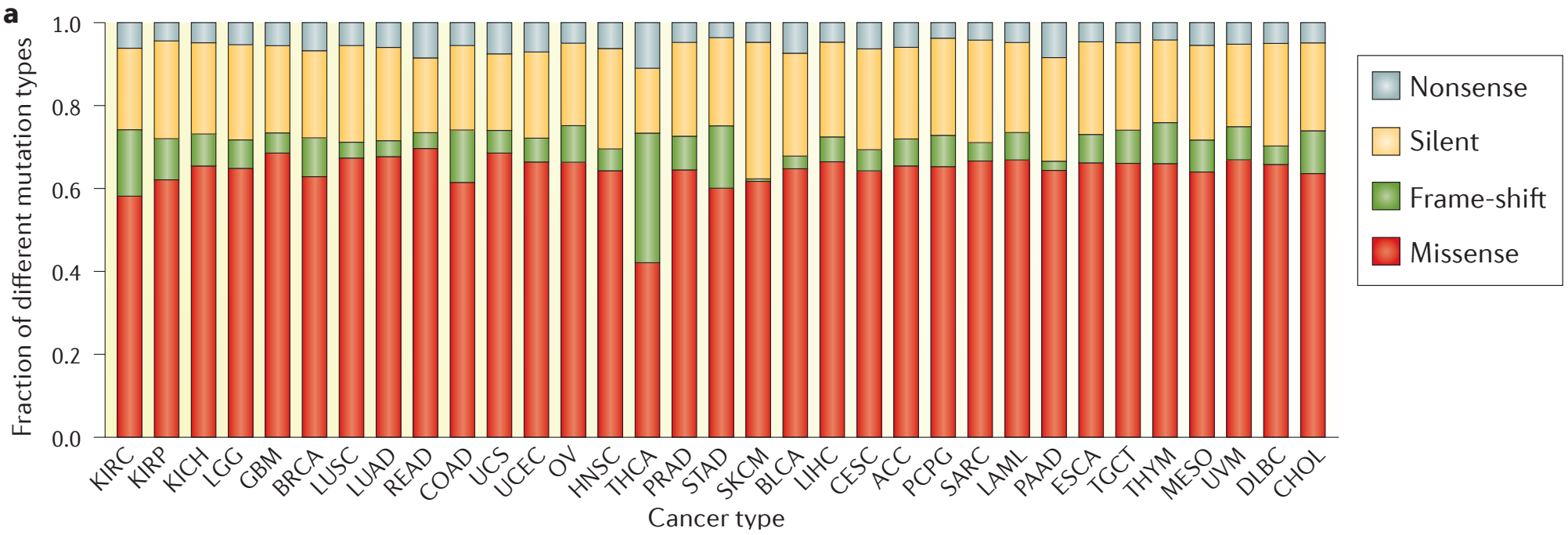
~2/3 of variants are non-synonymous SNVs

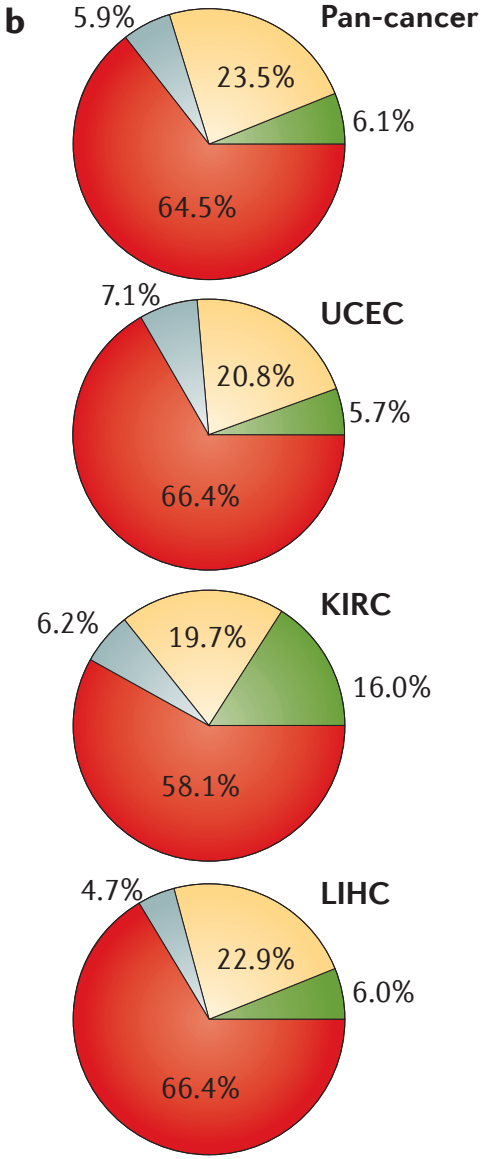
The variants reported here are not exclusive to driver events

dominated by SNVs: kidney clear-cell carcinoma, glioblastoma multiforme, hepatocellular carcinoma, acute myeloid leukaemia, colorectal carcinoma and endometrial carcinoma

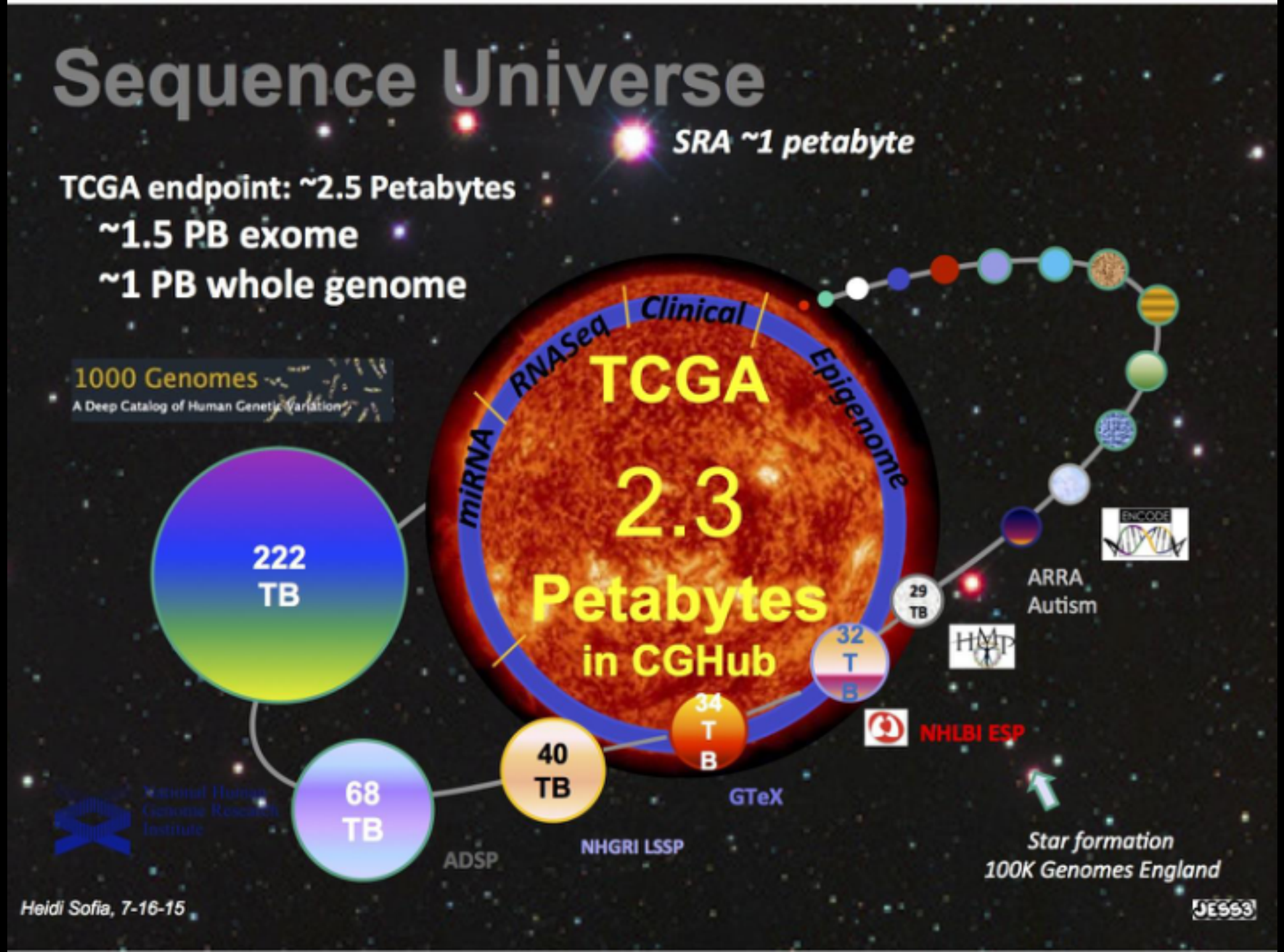
dominated by CNVs: most serous ovarian and breast carcinomas and many lung and head and neck squamous cell carcinomas

Outlier w/fewer non-synonymous SNVs and more frame-shifts: THCA (thyroid carcinoma)





Yi et al, NRG (2017)


Sequence Universe
SRA ~1 petabyte
TCGA endpoint: ~2.5 Petabytes
~1.5 PB exome
~1 PB whole genome
1000 Genomes
A Deep Catalog of Human Genetic Variation
TCGA
2.3
Petabytes
in CGHub
miRNA
RNASeq
Clinical
Epigenome
222 TB
68 TB
40 TB
34 TB
32 TB
29 TB
ENCODE
ARRA Autism
HMP
NHLBI ESP
GTeX
NHGRI LSSP
ADSP
Star formation
100K Genomes England
Heidi Sofia, 7-16-15
JES53

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

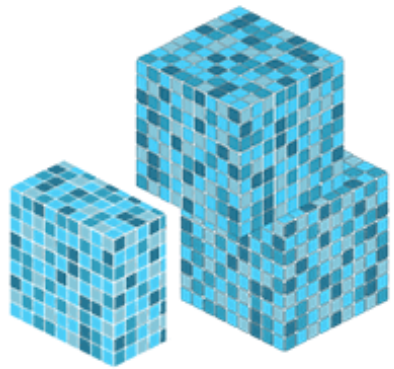
TCGA BY THE NUMBERS

TCGA produced over

2.5

PETABYTES

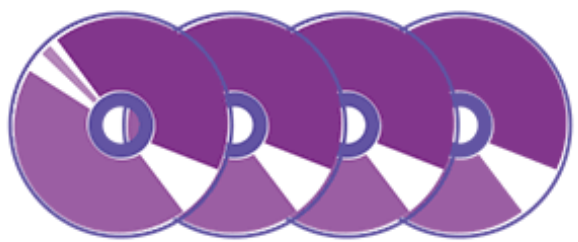
of data



To put this into perspective, **1 petabyte** of data is equal to

212,000

DVDs



TCGA data describes

33

DIFFERENT TUMOR TYPES

...including

10

RARE CANCERS

...based on paired tumor and normal tissue sets collected from

11,000

PATIENTS

...using

7

DIFFERENT DATA TYPES

cancergenome.nih.gov

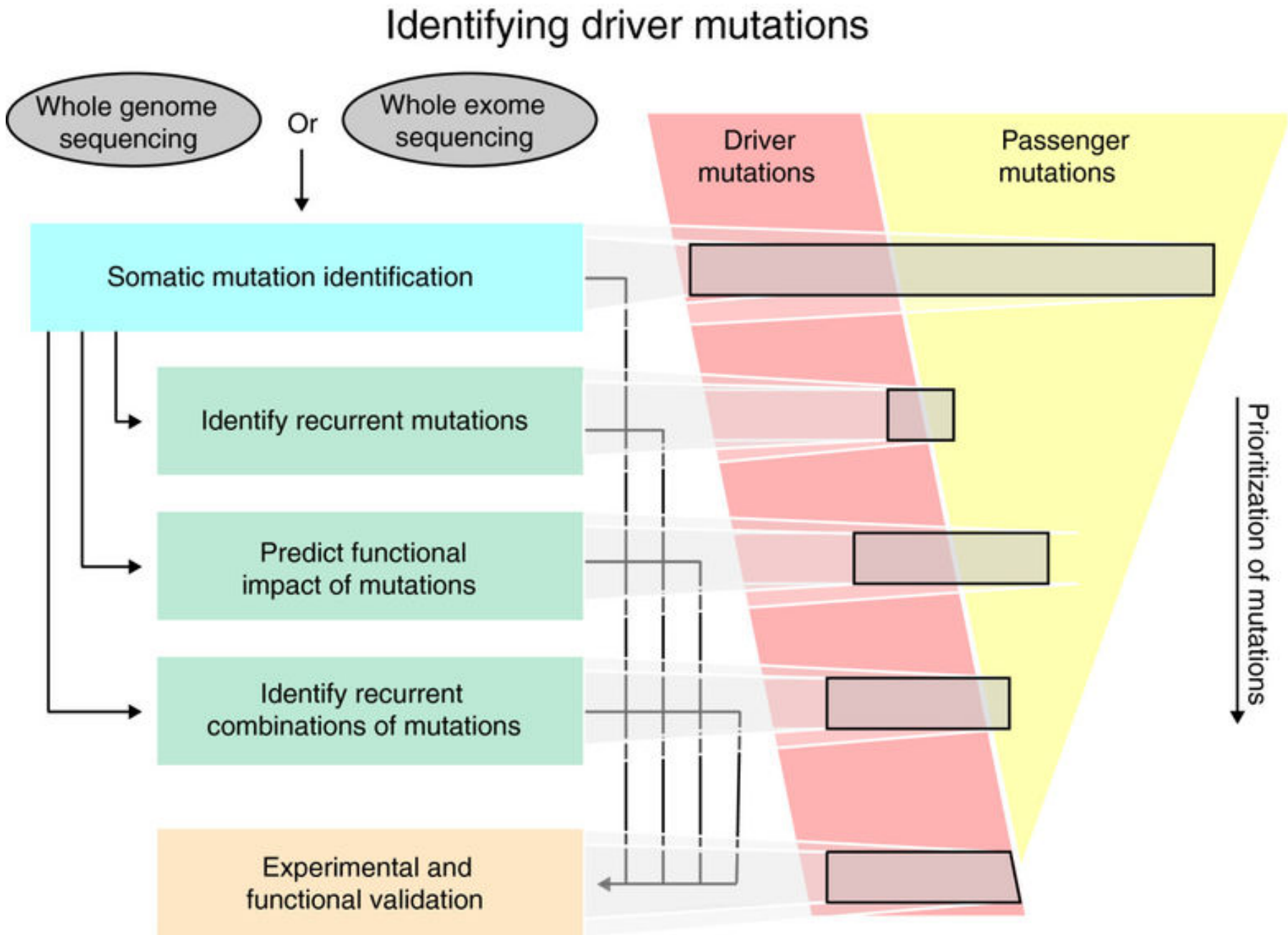
Logic

Data larger

Key is prioritizing

How to do this ?

Key aspect is driver v pass.



Raphael BJ, et al., Genome Med. (2014)

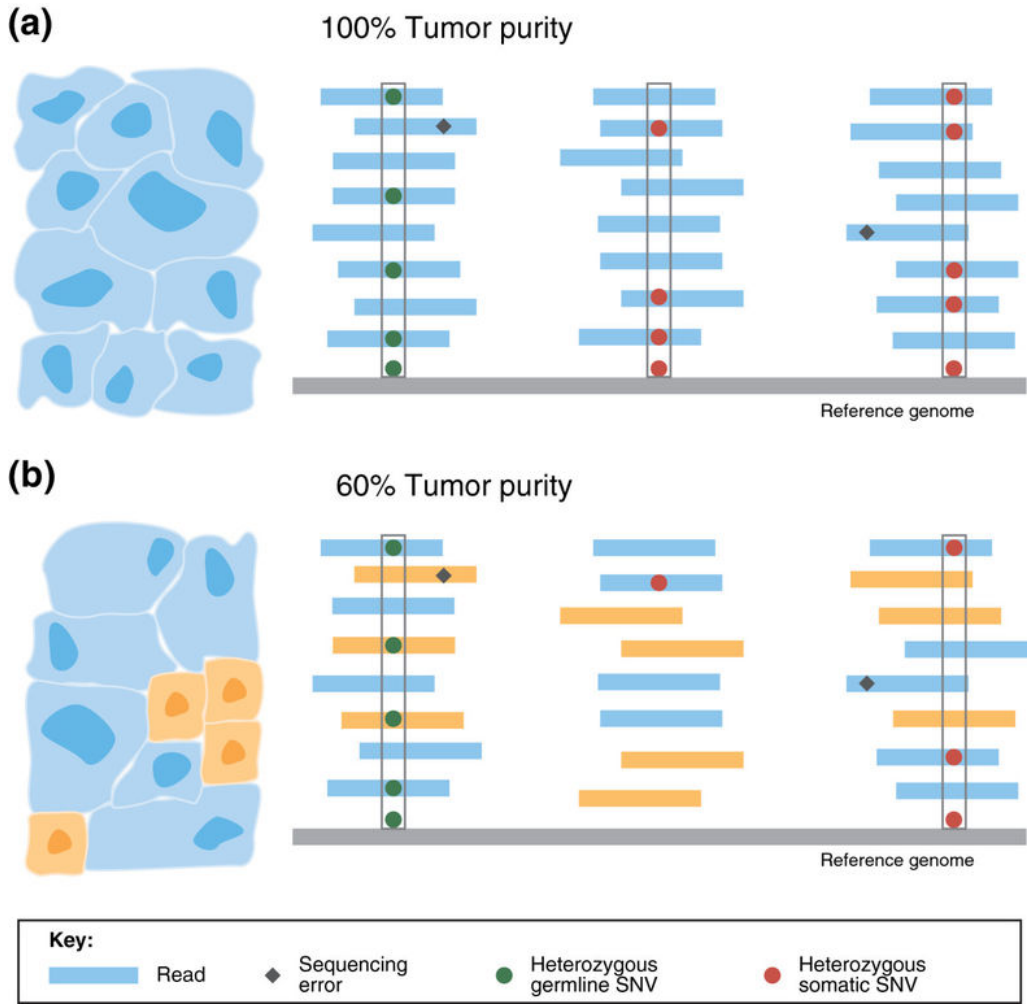
Logic

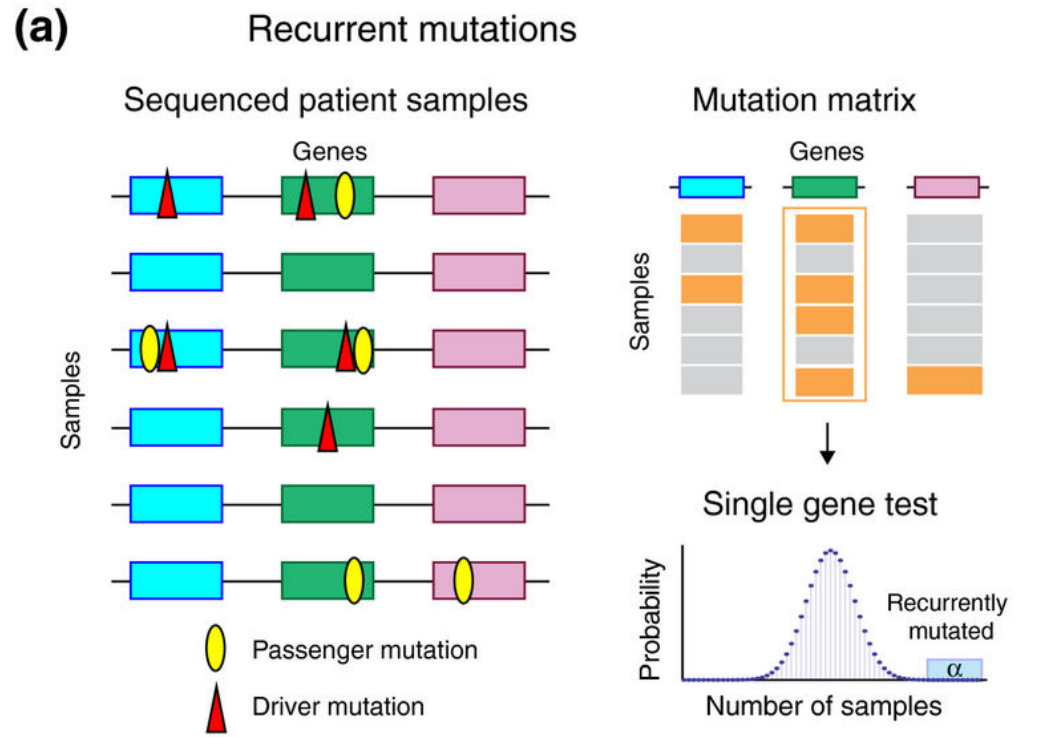
Data larger

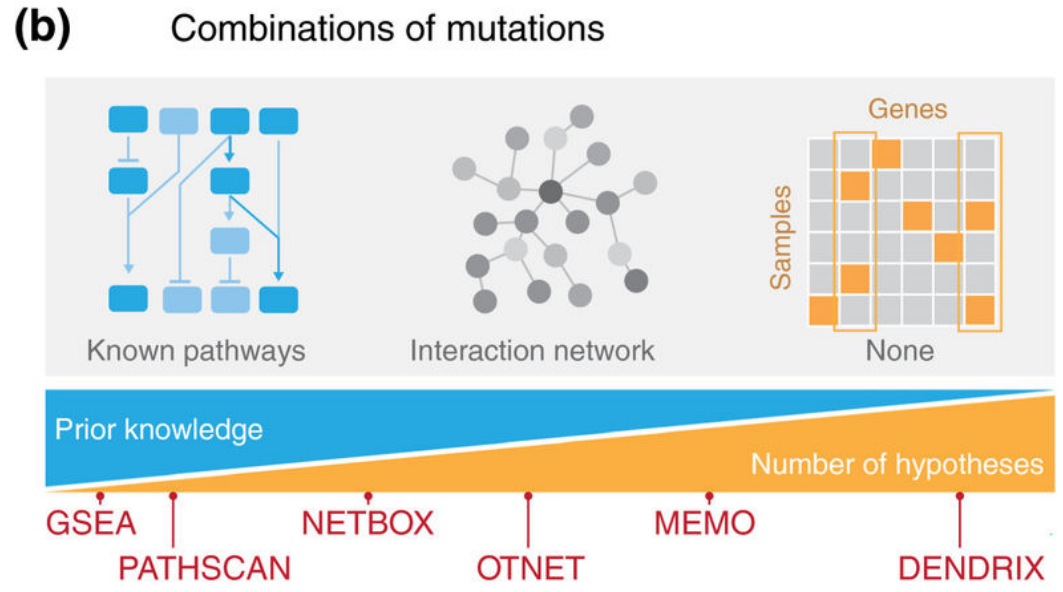
Key is prioritizing

How to do this ?

Key aspect is driver v pass.

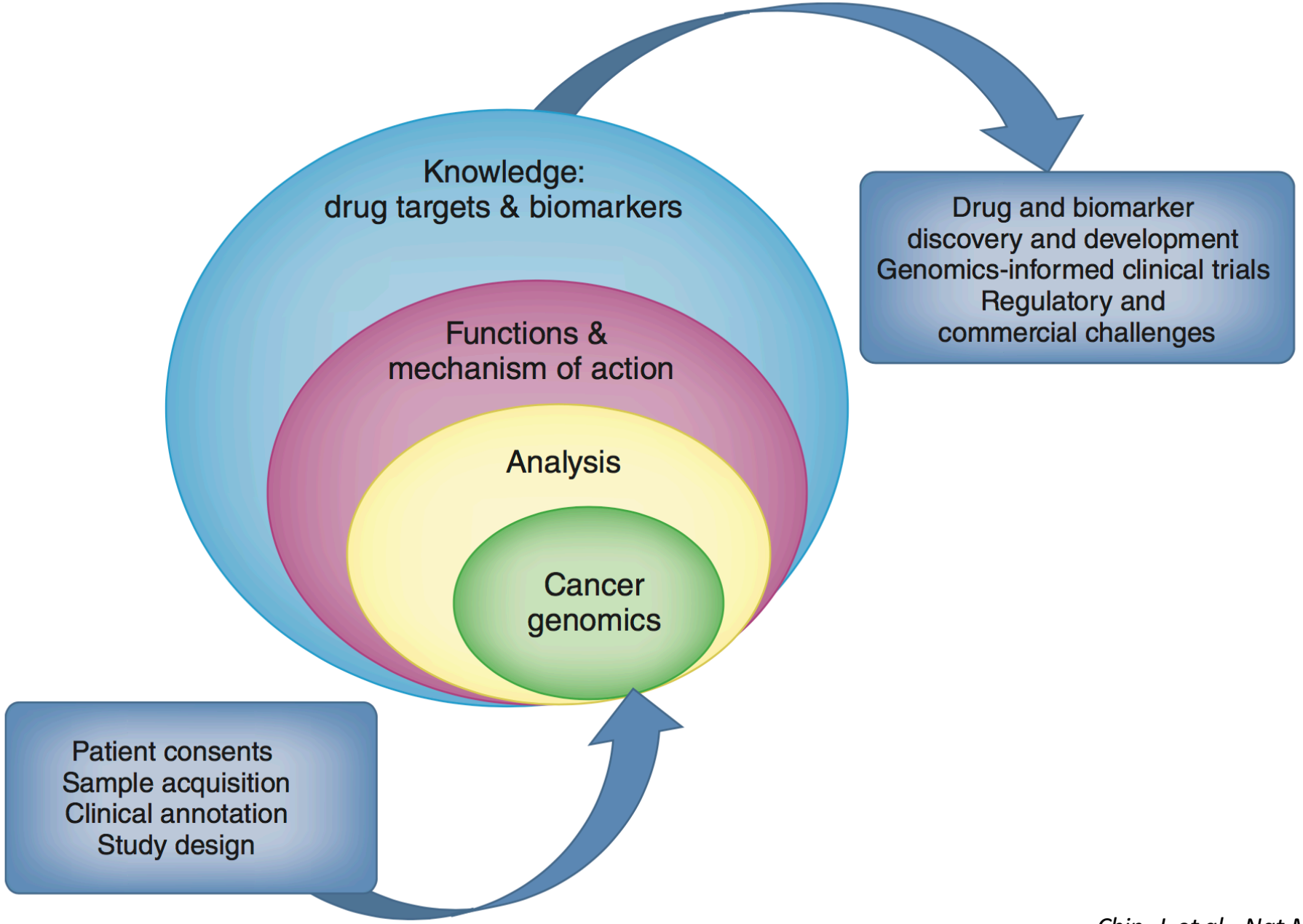






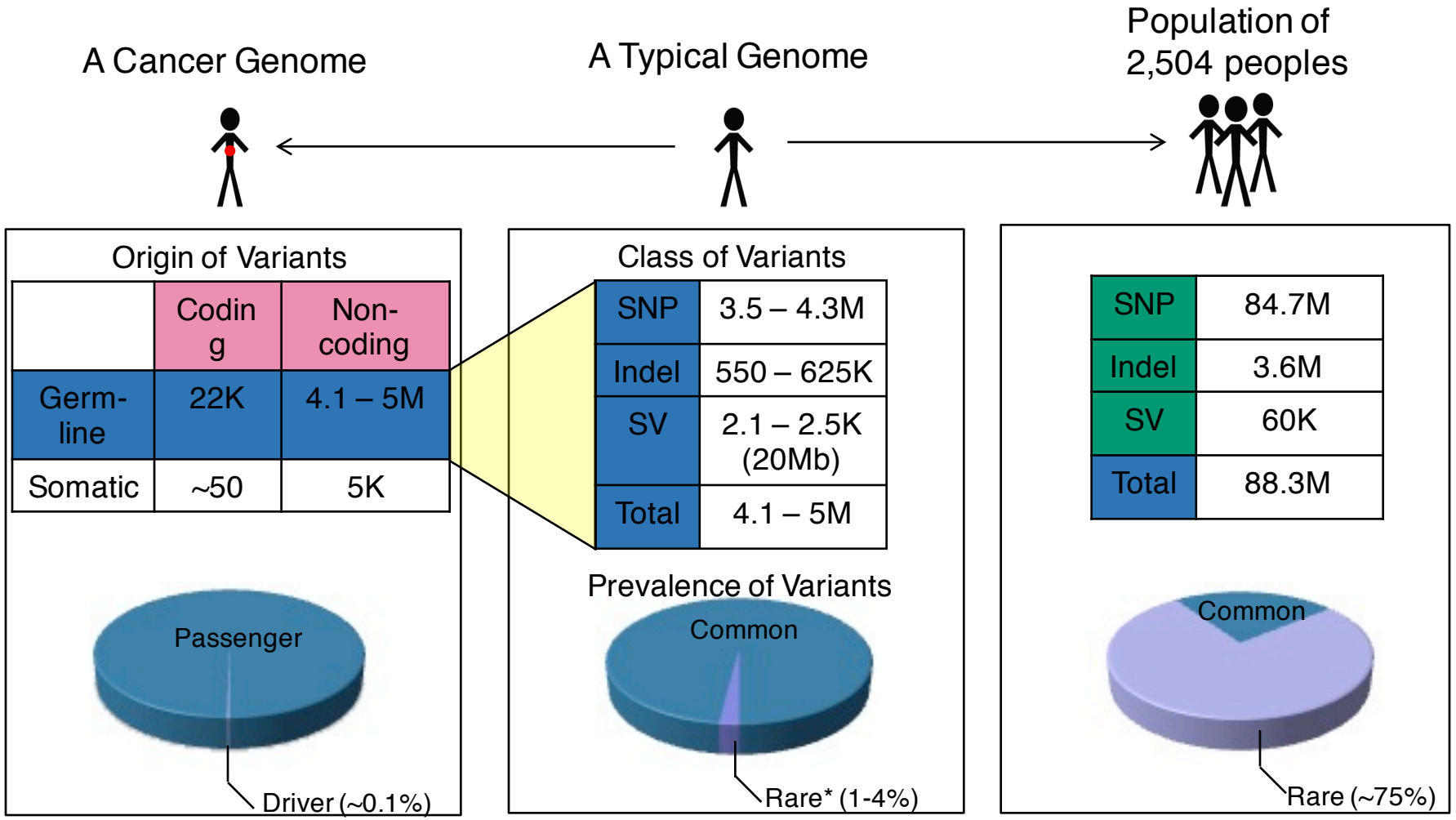
Raphael BJ, et al., Genome Med. (2014)

Misc.



Chin, L et al., Nat Med (2011)

Human Genetic Variation

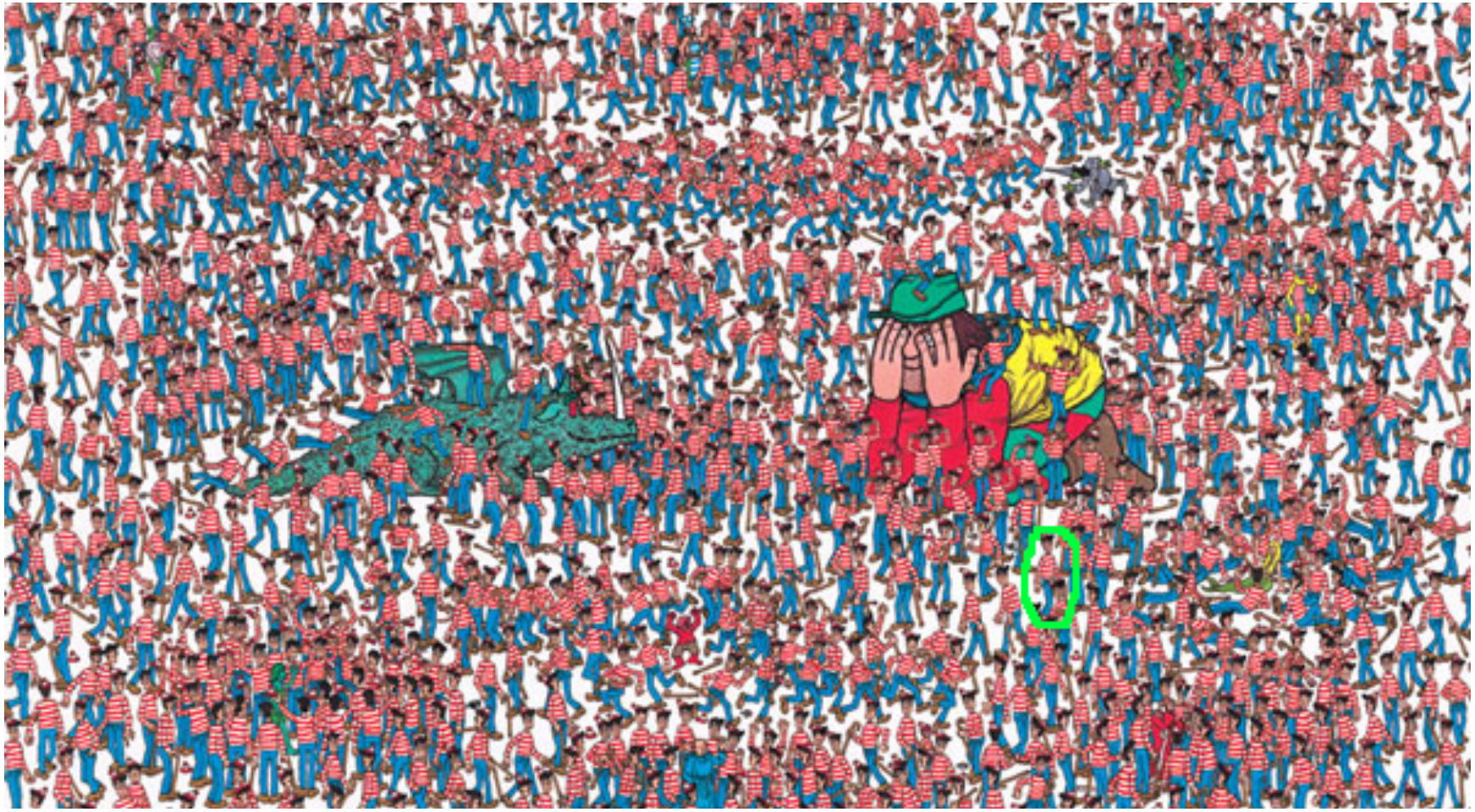


* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

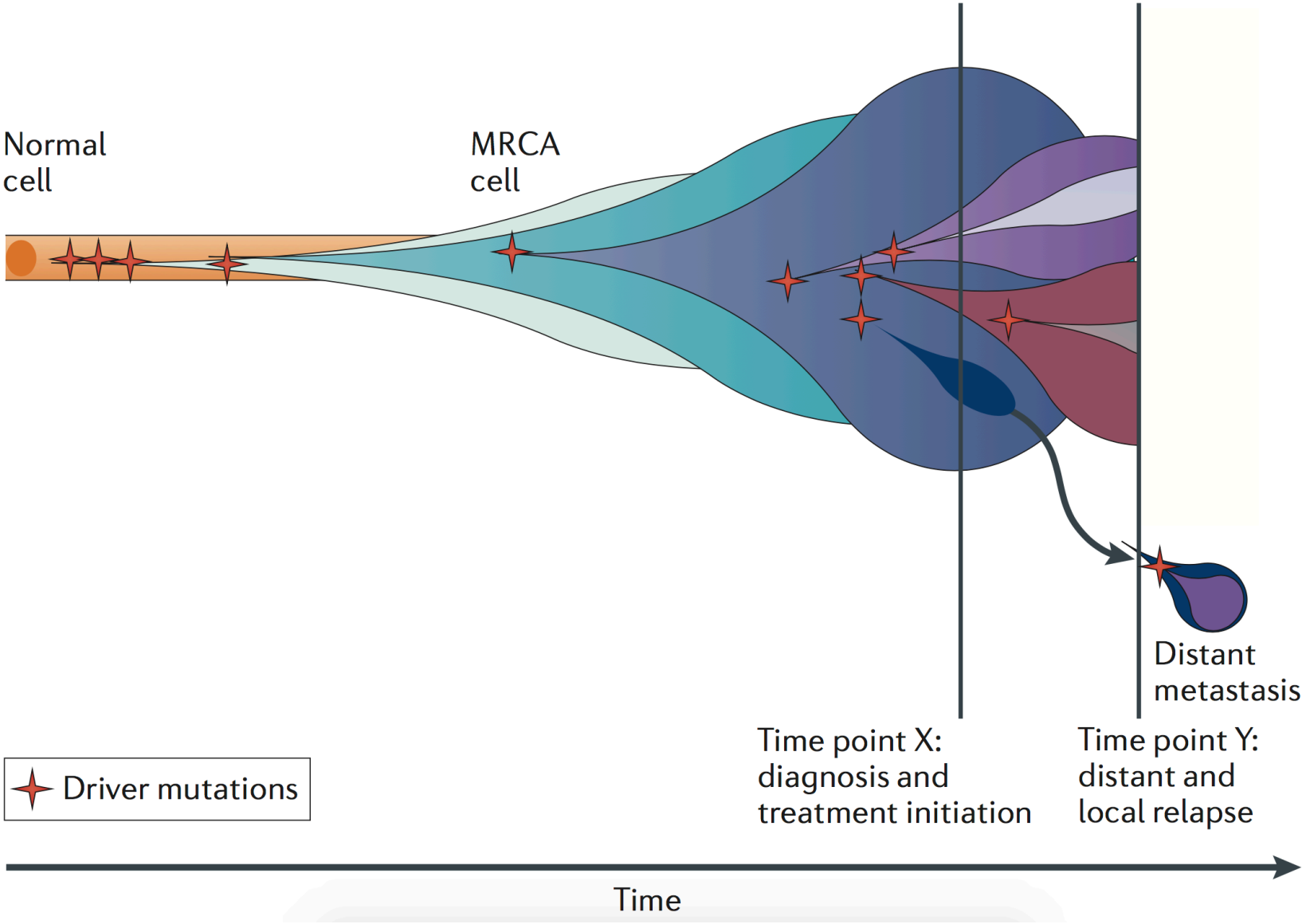
The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Cancer drivers: Significance & identification

(Finding the key mutations in ~3M Germline variants &
~5K Somatic Variants in a Tumor Sample)

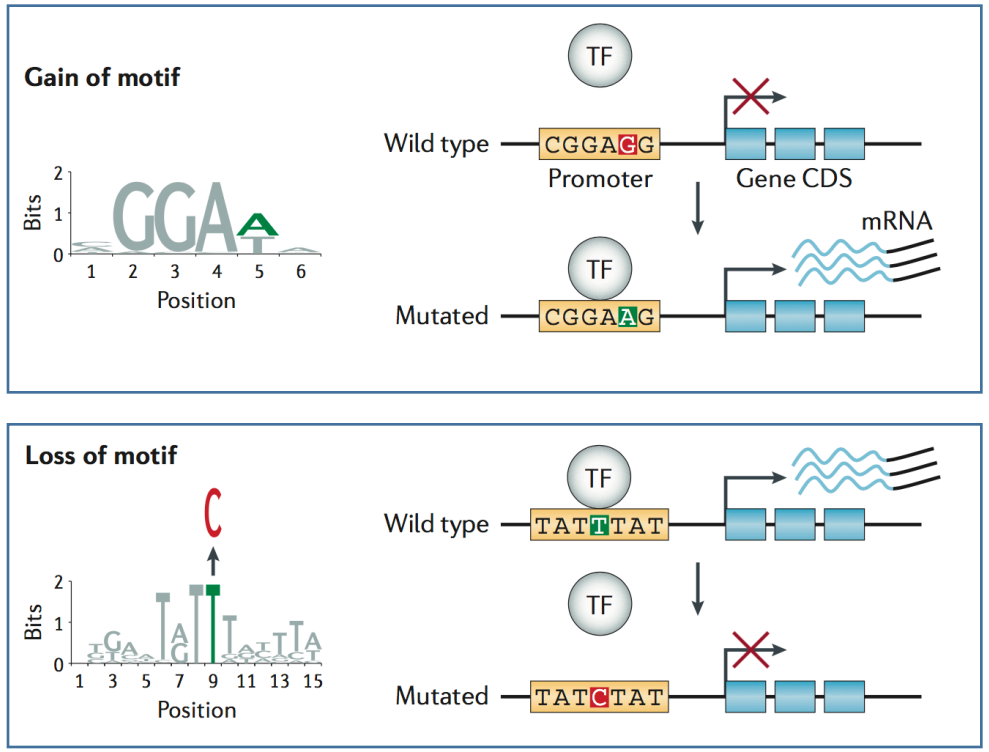


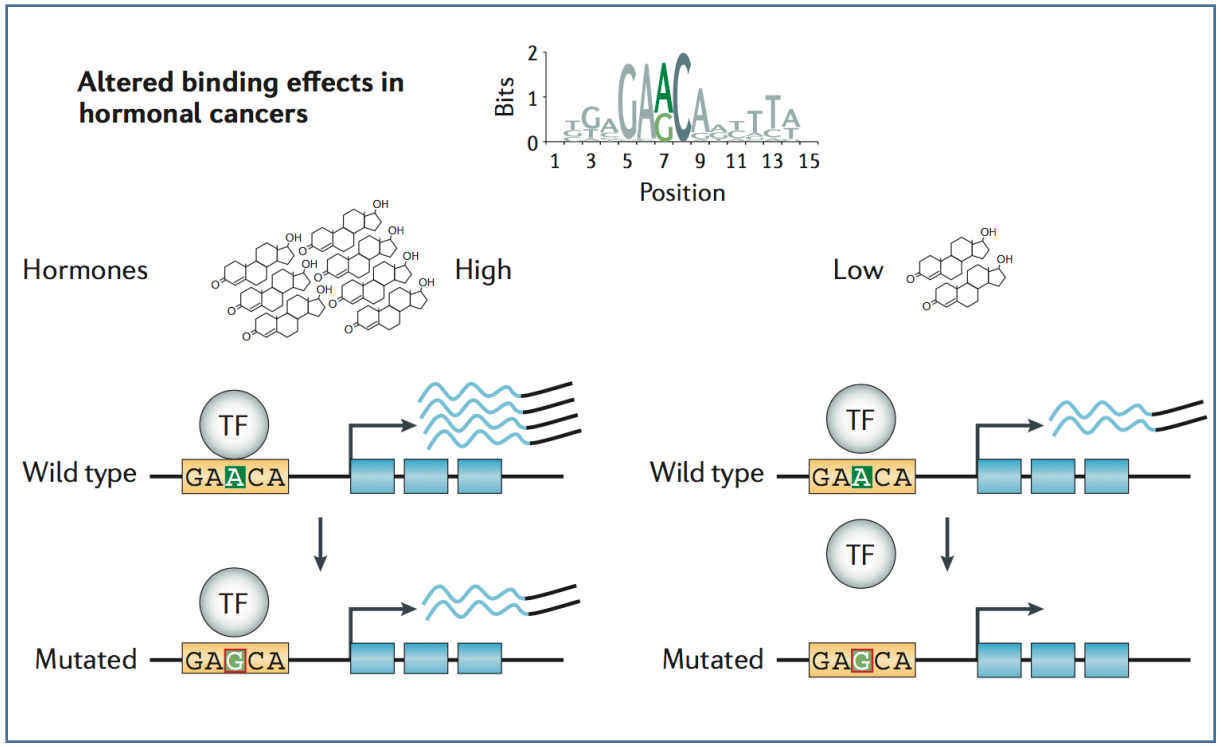
Cancer drivers: Significance & identification



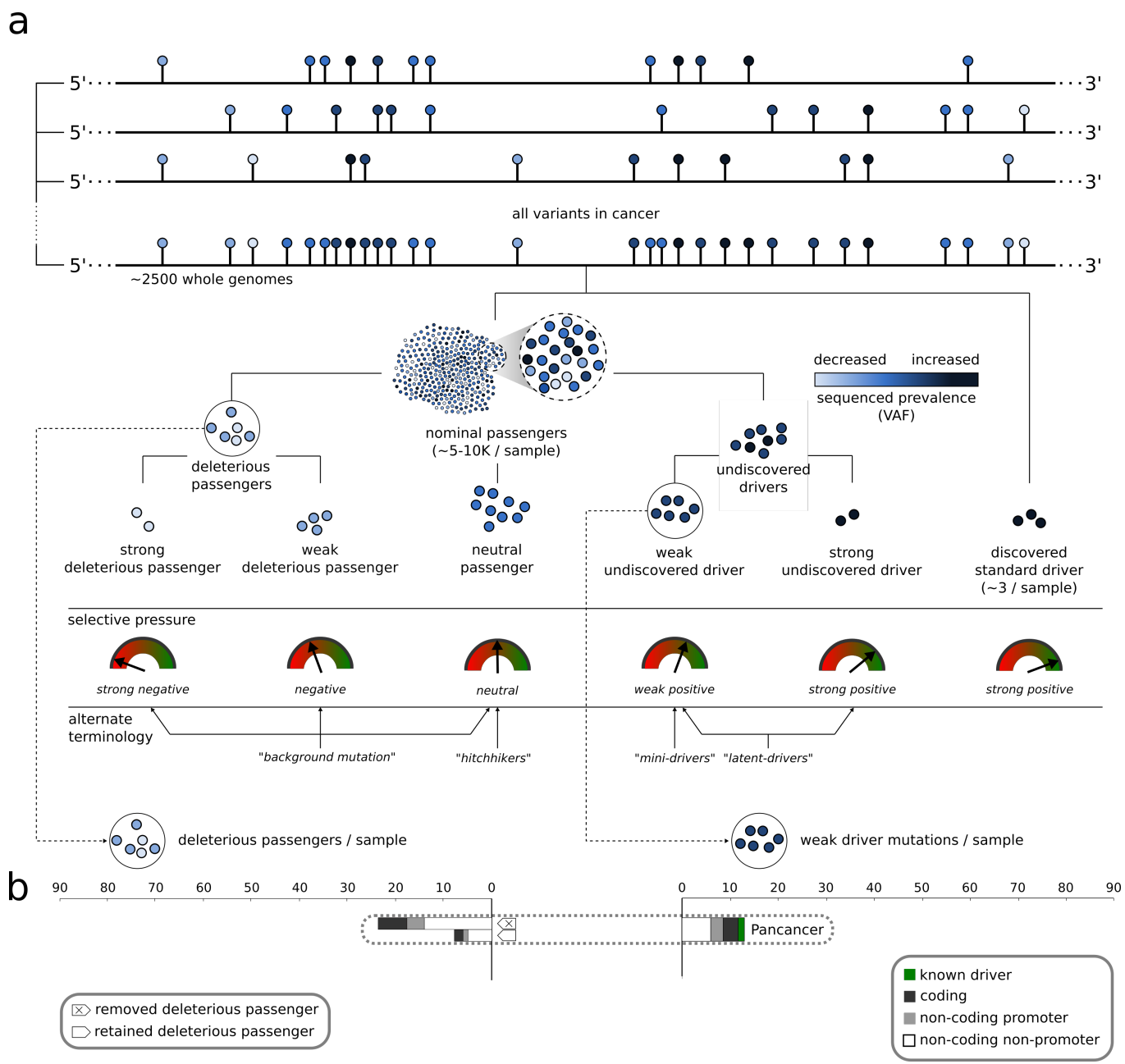
Yates et al, NRG (2012)

Cancer drivers: Significance & identification: functional annotations





Khurana et al, NRG (2016)


a
5'
3'
all variants in cancer
~2500 whole genomes
decreased
increased
sequenced prevalence (VAF)
nominal passengers (~5-10K / sample)
deleterious passengers
undiscovered drivers
strong deleterious passenger
weak deleterious passenger
neutral passenger
weak undiscovered driver
strong undiscovered driver
discovered standard driver (~3 / sample)
selective pressure
strong negative
negative
neutral
weak positive
strong positive
strong positive
alternate terminology
"background mutation"
"hitchhikers"
"mini-drivers"
"latent-drivers"
deleterious passengers / sample
weak driver mutations / sample
b
90 80 70 60 50 40 30 20 10 0
0 10 20 30 40 50 60 70 80 90
Pancancer
removed deleterious passenger
retained deleterious passenger
known driver
coding
non-coding promoter
non-coding non-promoter