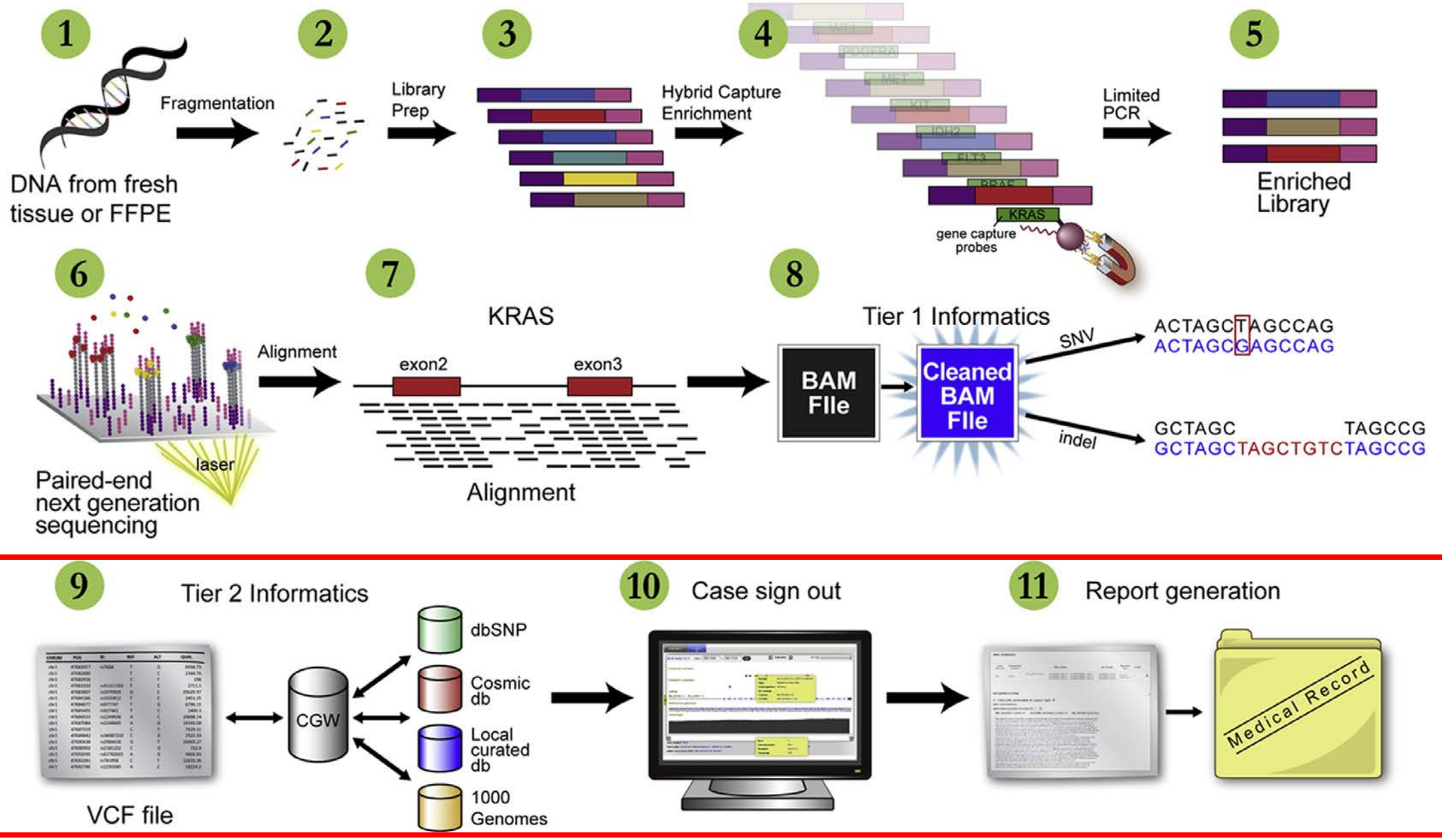
Fundamentals of Next-Generation Sequencing part 2: Analysis and Annotation

Frank C. Kuo, MD, PhD
Brigham and Women's Hospital
Boston, MA

Sept 7, 2017 SHEAHP 2017 Chicago



JMD 2014 16:89-105

Massively parallel sequencing (Next-Generation Sequencing)

GATGCTGATGCTGACGTAC

GATGCTGACGTACAAGTTT

‘

‘

‘

‘

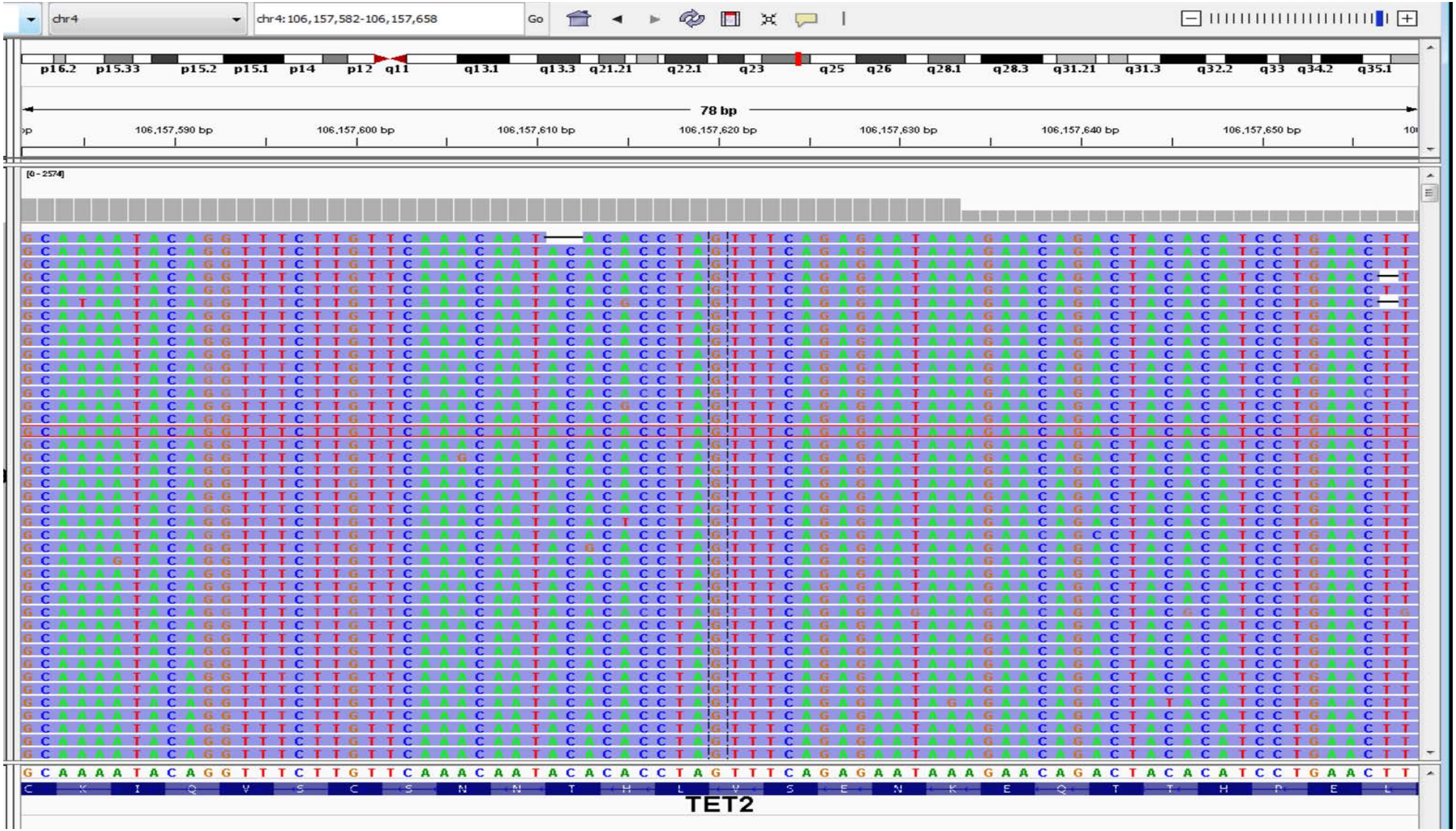
‘

‘

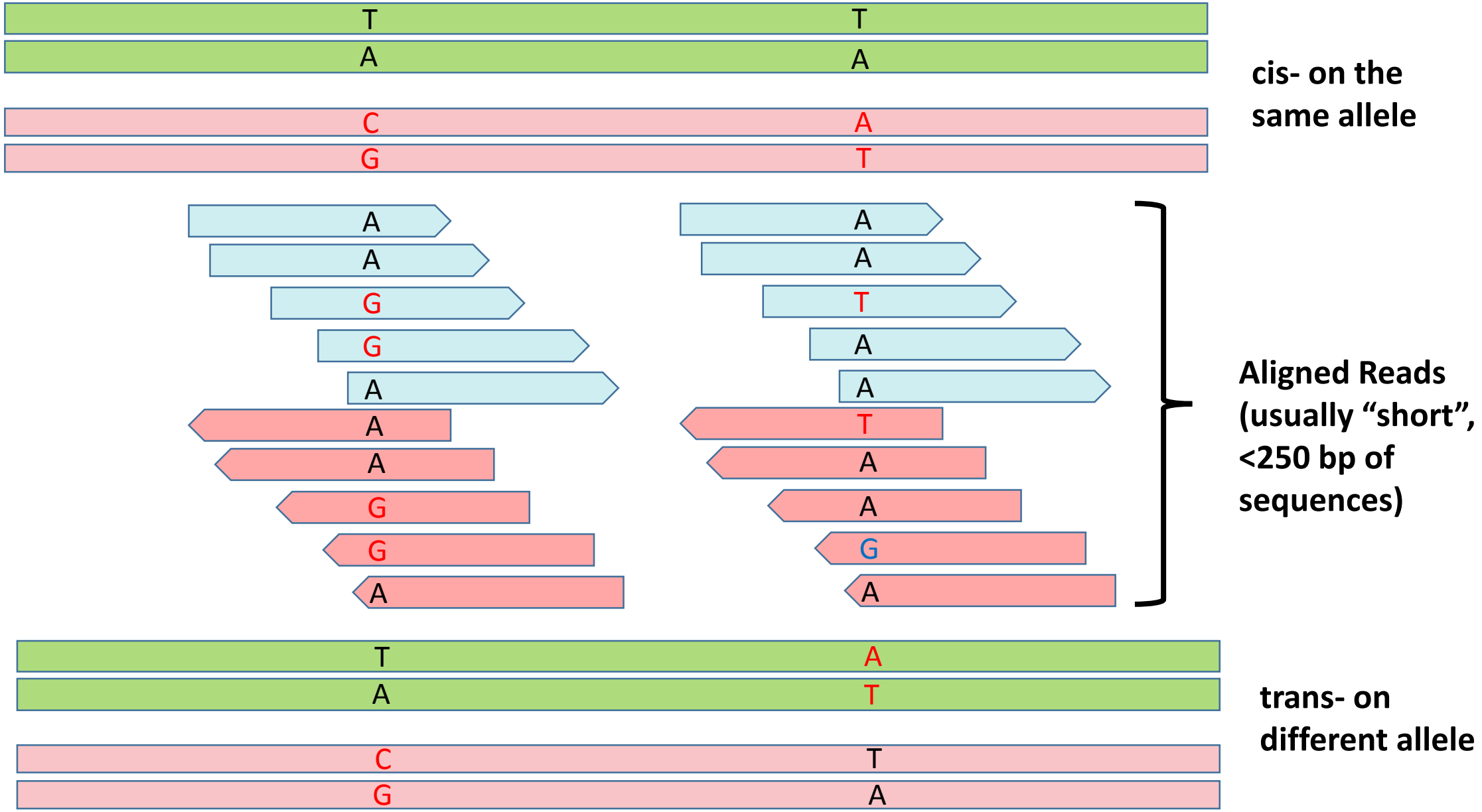
GATGCTGGACGTACCGGAT

ACGTACAAGTAAACGTAAG

Align the millions of “reads” to a reference genome



Identification of sequences that do not match with the reference genome



Variant Call Format (VCF) file - header

##fileformat=VCFv4.1
##FORMAT=<ID=GQX,Number=1,Type=Integer,Description="Minimum of {Genotype quality assuming variant position,Genotype quality assuming non-variant position}">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allele Depth">
##FORMAT=<ID=VF,Number=1,Type=Float,Description="Variant Frequency">
##FORMAT=<ID=NL,Number=1,Type=Integer,Description="Applied BaseCall Noise Level">
##FORMAT=<ID=SB,Number=1,Type=Float,Description="StrandBias Score">
##INFO=<ID=TI,Number=.,Type=String,Description="Transcript ID">
##INFO=<ID=GI,Number=.,Type=String,Description="Gene ID">
##INFO=<ID=EXON,Number=0,Type=Flag,Description="Exon Region">
##INFO=<ID=FC,Number=.,Type=String,Description="Functional Consequence">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FILTER=<ID=LowVariantFreq,Description="Low variant frequency < 0.01">
##FILTER=<ID=LowGQ,Description="GQ below < 30.00">
##FILTER=<ID=R8,Description="IndelRepeatLength is greater than 8">
##FILTER=<ID=LowDP,Description="Low coverage (DP tag), therefore no genotype called">
##FILTER=<ID=SB,Description="Variant strand bias too high">
##annotator=MARS
##fileDate=20160504
##source=CallSomaticVariants 3.5.2.1
##CallSomaticVariants_cmdline=" -B D:\Illumina\MiSeqAnalysis\160503_M03875_0078_000000000-AP7K2\Data\Intensities\BaseCalls\Alignment -g [C:\Illumina\MiSeq Reporter\Genomes\Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFASTA,] -f 0.01 -fo False -b 20 -q 100 -c 10 -s 0.5 -a 20 -F 20 -gVCF True -i true -r D:\Illumina\MiSeqAnalysis\160503_M03875_0078_000000000-AP7K2"
##reference=C:\Illumina\MiSeq Reporter\Genomes\Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFASTA\genome.fa
##contig=<ID=chr1,length=249250621>
##contig=<ID=chr2,length=243199373>
##contig=<ID=chr3,length=198022430>
##contig=<ID=chr4,length=191154276>
##contig=<ID=chr5,length=180915260>
##contig=<ID=chr6,length=171115067>

Variant Call Format (VCF) file - content

chr1	65301238	rs2274945	A	G	100	PASS	DP=426;TI=NM_002227;GI=JAK1;FC=Silent	0/1:100:400,26:0.0610:20:-100.0000:100
chr1	65306996	.	GT	G	44	PASS	DP=755;TI=NM_002227;GI=JAK1;FC=Frameshift;EXON	0/1:44:734,21:0.0278:20:-100.0000:44
chr1	65310489	rs2230588	T	C	100	PASS	DP=134;TI=NM_002227;GI=JAK1;FC=Synonymous_P733P;EXON	0/1:100:69,65:0.4851:20:-100.0000:100
chr1	65311214	rs3737139	G	C	32	PASS	DP=37;TI=NM_002227;GI=JAK1;FC=Synonymous_A699A;EXON	0/1:32:33,4:0.1081:20:-100.0000:32
chr1	65316675	rs310216	G	A	100	PASS	DP=305;TI=NM_002227;GI=JAK1;FC=Silent	0/1:100:186,119:0.3902:20:-100.0000:100
chr1	65321250	rs2230586	G	A	100	PASS	DP=1041;TI=NM_002227;GI=JAK1;FC=Synonymous_I530I;EXON	0/1:100:969,70:0.0672:20:-100.0000:100
chr1	65321388	rs310229	G	A	100	PASS	DP=198;TI=NM_002227;GI=JAK1;FC=Silent	0/1:100:99,99:0.5000:20:-100.0000:100
chr1	65321409	rs2274948	T	C	75	PASS	DP=201;TI=NM_002227;GI=JAK1;FC=Silent	0/1:75:186,14:0.0697:20:-100.0000:75
chr1	65321419	rs11579283	G	A	77	PASS	DP=195;TI=NM_002227;GI=JAK1;FC=Silent	0/1:77:181,14:0.0718:20:-100.0000:77
chr1	115251293	.	GA	G	53	PASS	DP=312;TI=NM_002524;GI=NRAS;FC=Noncoding	0/1:53:298,14:0.0449:20:-100.0000:53
chr1	115252270	.	T	C	55	PASS	DP=479;TI=NM_002524;GI=NRAS;FC=Missense_T124A;EXON	0/1:55:461,18:0.0376:20:-26.6820:55
chr1	120459251	.	G	T	59	PASS	DP=130;TI=NM_024408;GI=NOTCH2;FC=Missense_H2032N;EXON	0/1:59:120,10:0.0769:20:-100.0000:59
chr1	120468425	rs17024525	G	A	76	PASS	DP=267;TI=NM_024408;GI=NOTCH2; EXON	0/1:76:251,16:0.0599:20:-100.0000:76
chr1	155870416	rs1749409	G	A	100	PASS	DP=579;TI=NM_006912;GI=RIT1;FC=Silent	0/1:100:77,502:0.8670:20:-100.0000:100
chr1	155874765	.	GT	G	100	R8	DP=975;TI=NM_006912;GI=RIT1;FC=Noncoding	0/1:100:914,61:0.0626:20:-100.0000:100
chr1	155874765	.	G	GT	100	R8	DP=939;TI=NM_006912;GI=RIT1;FC=Noncoding	0/1:100:902,37:0.0394:20:-100.0000:100
chr1	155880391	rs867550	T	C	100	PASS	DP=810;TI=NM_006912;GI=RIT1;FC=Silent	0/1:100:560,249:0.3074:20:-100.0000:100
chr1	155880573	rs493446	C	G	100	PASS	DP=251;TI=NM_006912;GI=RIT1;FC=Silent;EXON	0/1:100:12,238:0.9482:20:-100.0000:100
chr1	155880754	.	A	C	100	PASS	DP=872;TI=NM_006912;GI=RIT1;FC=Silent	0/1:100:763,105:0.1204:20:-100.0000:100
chr1	155880760	rs5777961	C	CA	100	PASS	DP=953;TI=NM_006912;GI=RIT1;FC=Noncoding	0/1:100:135,818:0.8583:20:-100.0000:100
chr2	8178735	rs1364054	A	G	100	PASS	DP=194;TI=NR_034135;GI=C2orf46;FC=Silent	0/1:100:83,111:0.5722:20:-100.0000:100
chr2	8178798	rs62104935	C	T	100	PASS	DP=195;TI=NR_034135;GI=C2orf46;FC=Silent	0/1:100:92,103:0.5282:20:-100.0000:100

Filtering of variants

- Raw VCF files may have thousands of variants per samples
  - Not real
    - Low quality (poor phred score, low confidence, etc.,)
    - Poor alignment
    - PCR errors
  - Real
    - Germline variant
    - Somatic variant
      - Pathogenic/likely pathogenic
      - Unknown significance
      - Passenger

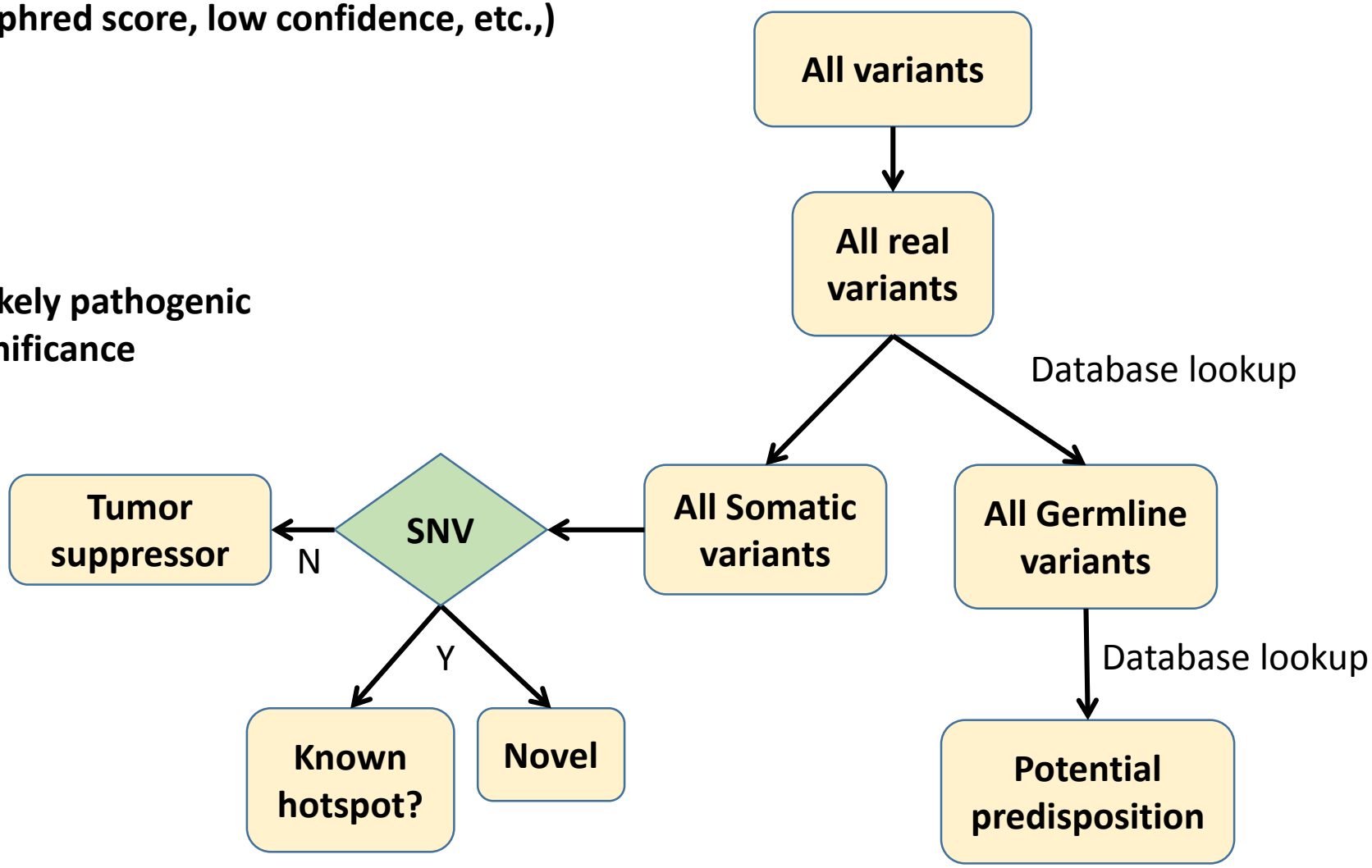
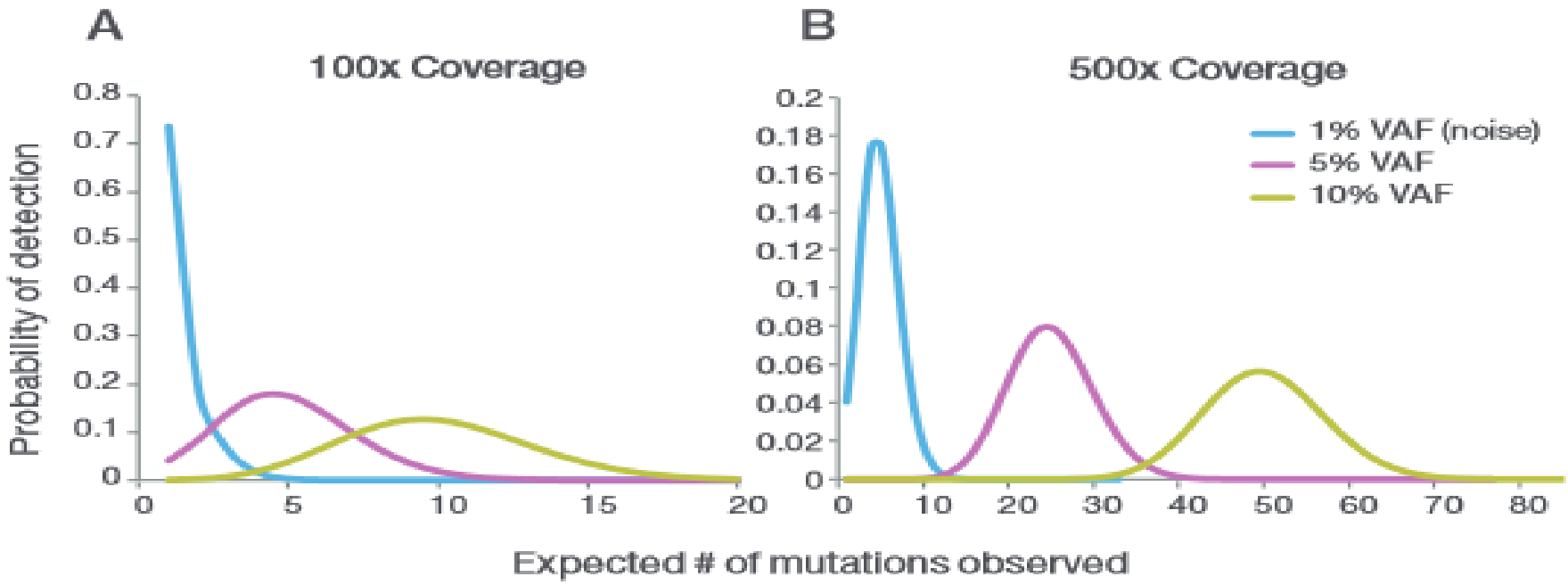


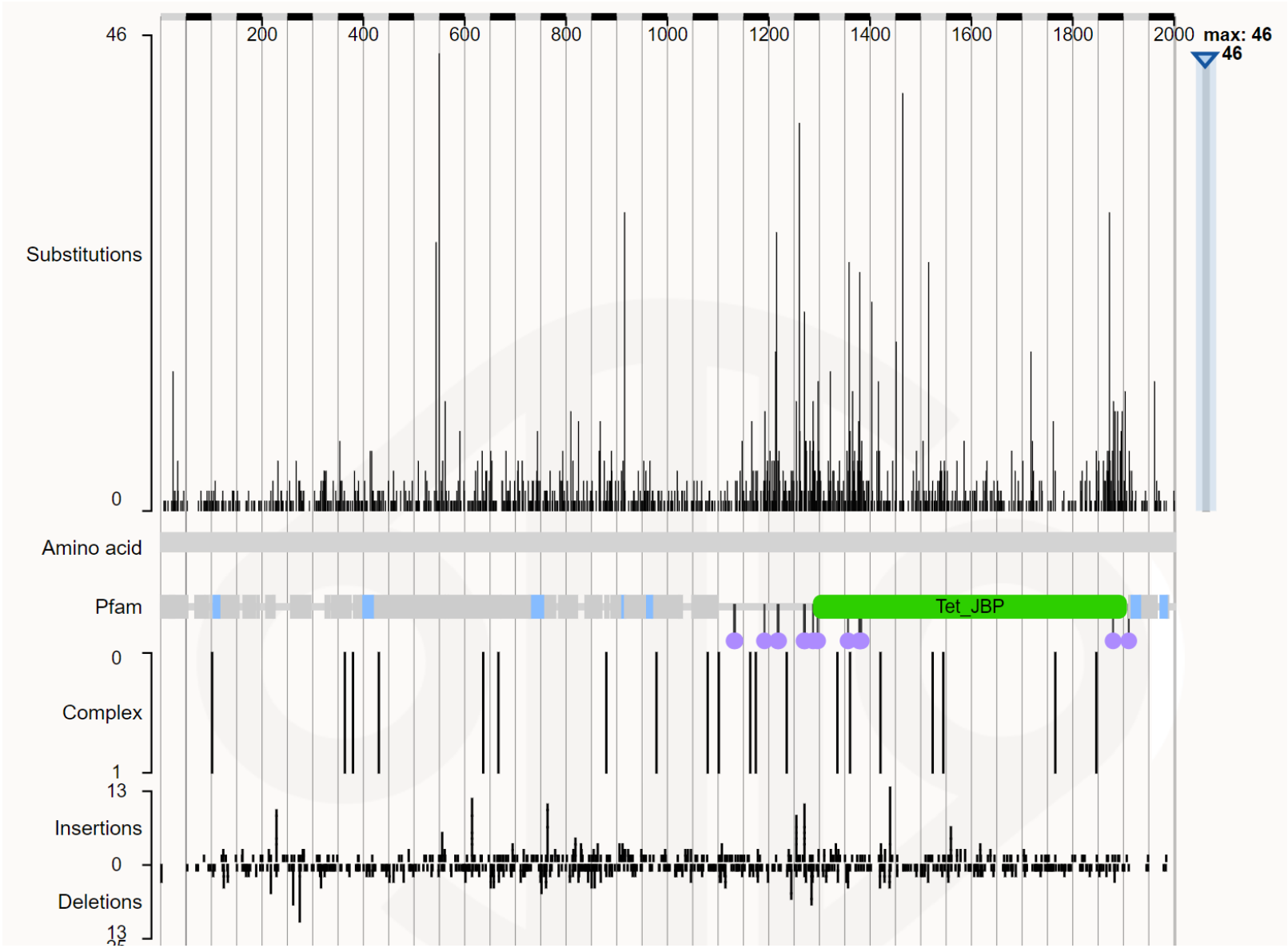
Figure 1: Impact of Coverage Depth on VAF Overlap



With only 100× coverage (panel A), there is considerable overlap between 5% and 1% VAF, inhibiting the ability to confidently call low-frequency variants below 5%. In contrast, variants below 5% frequency can be reliably called when coverage depth is increased to > 500× coverage (panel B).

TET2

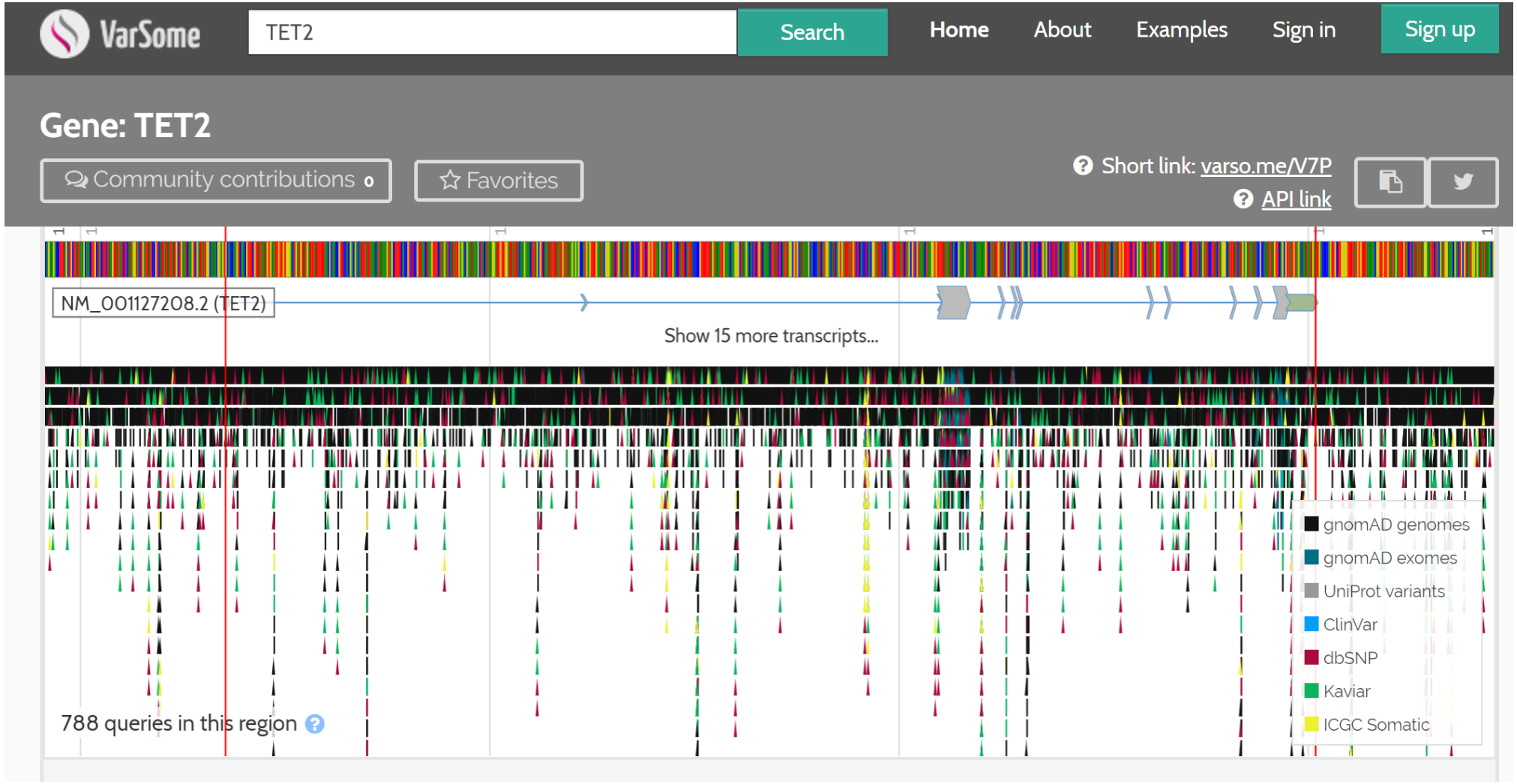
http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=TET2



http://exac.broadinstitute.org/gene/ENSG00000168769

4:106155108 G / A (rs183431550)	4	106155108	p.Gln3Gln	PASS	synonymous	6	118780	0	0.00005051
4:106155112 A / G	4	106155112	p.Arg5Gly	PASS	missense	1	119142	0	0.000008393
4:106155114 A / G	4	106155114	p.Arg5Arg	PASS	synonymous	8	119362	0	0.00006702
4:106155121 C / G (rs147112198)	4	106155121	p.His8Asp	PASS	missense	23	119872	0	0.0001919
4:106155130 G / A	4	106155130	p.Gly11Ser	PASS	missense	1	120368	0	0.000008308
4:106155131 G / T	4	106155131	p.Gly11Val	PASS	missense	1	120412	0	0.000008305
4:106155136 A / C (rs146373819)	4	106155136	p.Arg13Arg	PASS	synonymous	73	120646	0	0.0006051
4:106155167 C / G	4	106155167	p.Pro23Arg	PASS	missense	1	120988	0	0.000008265

https://varsome.com/gene/TET2



- **NGS found a mutation; then what?**
  - **Is it a germline or somatic variant?**
    - Is the tumor and germline DNA tested in pair?
    - Does its variant allele fraction help to determine whether it is somatic?
  - **It is somatic, but is it a driver or a passenger?**
    - A driver mutation is either gain-of-function (activating an oncogene) or loss-of-function (haploinsufficient or dominant-negative for a tumor suppressor gene)
    - A passenger mutation may be lost with time
  - **It is a germline SNP (single nucleotide polymorphism) that is found in tumor and germline DNA. Should I report it?**
    - Up to 10% of the cancer patients may have inherited germline variants that predispose them to get cancer. Genetic counseling may be warranted.
    - If the SNP is present in <1% of the general population, then it may be useful for identity. If the patient receives allogeneic bone marrow transplant, it may be informative for chimerism evaluation.

- **NGS found a mutation; then what?**
  - **Is it a recurrent mutation in this or other cancer types?**
    - Look it up in a cancer genome database such as COSMIC, cBioportal, Intogen, CIVIC, Firebrowse, etc.,
    - A gain-of-function mutation in an oncogene usually occurs on a hotspot.
  - **Is it a truncating mutation in a presumed tumor suppressor gene?**
    - frameshift mutations caused by insertion and deletion may be novel and have never been seen before.
  - **Is it seen in normal individuals at a significant frequency (usually >1%)?**
    - Look it up in exome database of normal individuals such as ESP and EXAC
    - Many SNP (single nucleotide polymorphism) have widely different allele frequency in different ethnic groups
  - **Has it been tested in vitro?**
    - Many proteins have multiple biologic functions. It is difficult to test all the functions of a mutant allele in the appropriate context. A negative (and sometimes a positive) result should be interpreted with caution.

The Journal of Molecular Diagnostics, Vol. 19, No. 1, January 2017

the Journal of Molecular Diagnostics
jmd.amjpathol.org

SPECIAL ARTICLE

Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer

A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists

Marilyn M. Li,[*†] Michael Datto,[*‡] Eric J. Duncavage,[*§] Shashikant Kulkarni,[*¶] Neal I. Lindeman,[*‖] Somak Roy,[*,**] Apostolia M. Tsimberidou,[*††] Cindy L. Vnencak-Jones,[*‡‡] Daynna J. Wolff,[*§§] Anas Younes,[*¶¶] and Marina N. Nikiforova[*,**]

From the Interpretation of Sequence Variants in Somatic Conditions Working Group of the Clinical Practice Committee, Association for Molecular Pathology, Bethesda, Maryland; the Department of Pathology and Laboratory Medicine,† Division of Genomic Diagnostics, the Children's Hospital of Philadelphia, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania; the Duke University School of Medicine,‡ Durham, North Carolina; the Department of Pathology and Immunology,§ Washington University School of Medicine, St. Louis, Missouri; Baylor Genetics,¶ Houston, Texas; the Brigham and Women's Hospital,‖ Harvard Medical School, Boston, Massachusetts; the University of Pittsburgh Medical Center,** Pittsburgh, Pennsylvania; the Department of Investigational Cancer Therapeutics,†† University of Texas MD Anderson Cancer Center, Houston, Texas; the Department of Pathology, Microbiology and Immunology,‡‡ Vanderbilt University Medical Center, Nashville, Tennessee; the Department of Pathology and Laboratory Medicine,§§ Medical University of South Carolina, Charleston, South Carolina; and the Memorial Sloan Kettering Cancer Center,¶¶ New York, New York*

Variant identification
- depth of coverage (reads)
- variant allele fraction (VAF)

Determining whether a variant is somatic or germline
- paired germline material

cross-references to external databases
- Germline polymorphism

Determine whether a variant is pathogenic or not
- known somatic pathogenic

in silico algorithm-based predictions

Tumor purity
Tumor ploidy
Loss-of-heterozygosity (uniparental disomy)
Tumor heterogeneity
Clonal evolution

We recommend reporting germline variants with known evidence of clinical impact.

Reports should be static, and the date of issue should be clearly presented; they do not need to be automatically recalled and/or reissued when medical knowledge changes.

Table 3 Categories of Clinical and/or Experimental Evidence

Category	Therapeutic	Diagnosis	Prognosis
Level A	1. Biomarkers that predict response or resistance to FDA-approved therapies for a specific type of tumor 2. Biomarkers included in professional guidelines that predict response or resistance to therapies for a specific type of tumor	Biomarkers included in professional guidelines as diagnostic for a specific type of tumor	Biomarkers included in professional guidelines as prognostic for a specific type of tumor
Level B	Biomarkers that predict response or resistance to therapies for a specific type of tumor based on well-powered studies with consensus from experts in the field	Biomarkers of diagnostic significance for a specific type of tumor based on well-powered studies with consensus from experts in the field	Biomarkers of prognostic significance for a specific type of tumor based on well-powered studies with consensus from experts in the field
Level C	1. Biomarkers that predict response or resistance to therapies approved by the FDA or professional societies for a different type of tumor 2. Biomarkers that serve as inclusion criteria for clinical trials	Biomarkers of diagnostic significance based on the results of multiple small studies	Biomarkers of prognostic significance based on the results of multiple small studies
Level D	Biomarkers that show plausible therapeutic significance based on preclinical studies	Biomarkers that may assist disease diagnosis themselves or along with other biomarkers based on small studies or a few case reports	Biomarkers that may assist disease prognosis themselves or along with other biomarkers based on small studies or a few case reports

FDA, Food and Drug Administration.

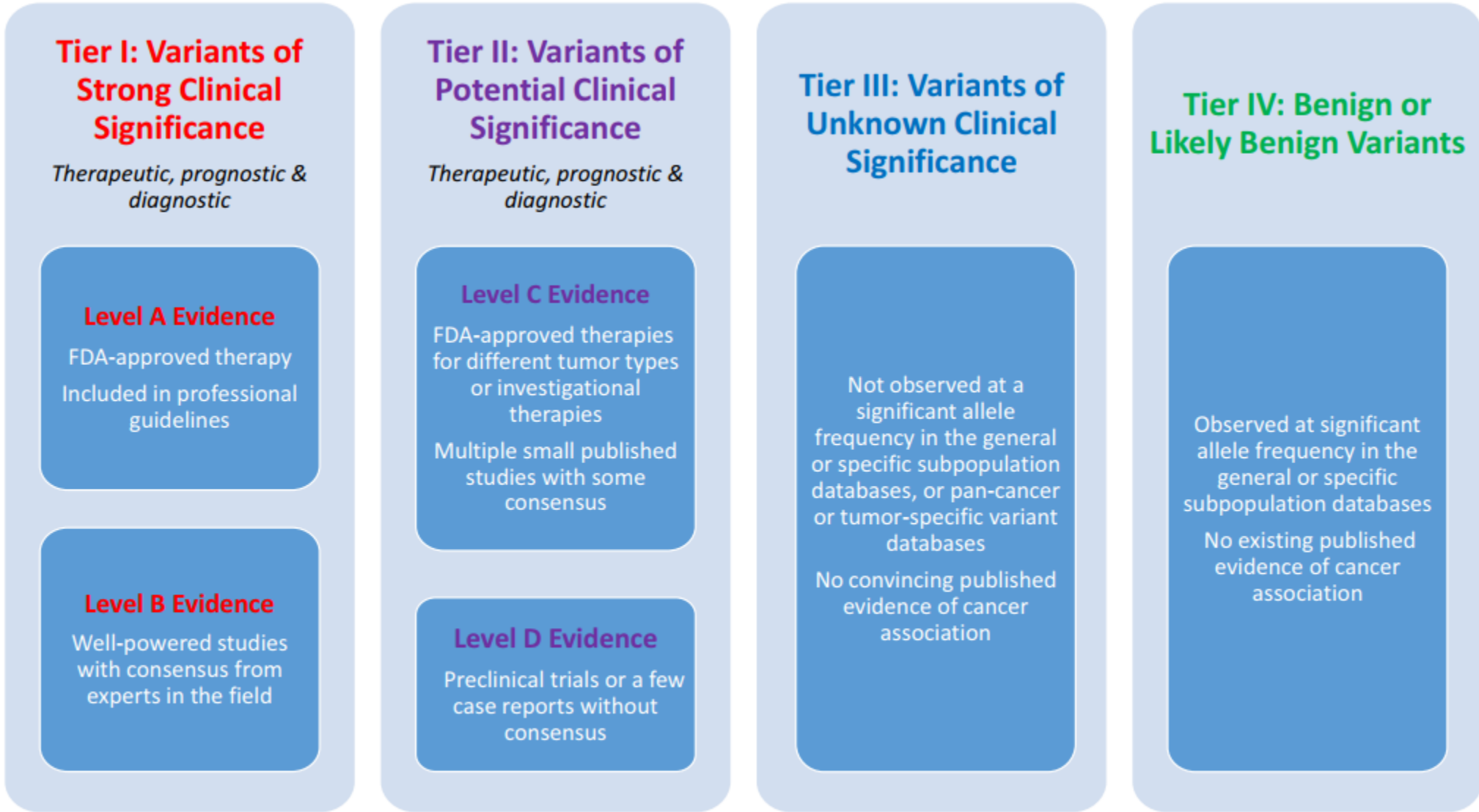


Figure 2 Evidence-based variant categorization. Somatic variants are classified into four tiers based on their level of clinical significance in cancer diagnosis, prognosis, and/or therapeutics. Variants in tier I are of strongest clinical significance, and variants in tier IV are benign or likely benign variants. FDA, Food and Drug Administration.

http://varnomen.hgvs.org/

Human Mutation
Variation, Informatics, and Disease

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY

Explore this journal >

Special Article

HGVS Recommendations for the Description of Sequence Variants: 2016 Update

Johan T. den Dunnen ✉, Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux, Timothy Smith, Stylianos E. Antonarakis, Peter E.M. Taschner,

on behalf of the Human Genome Variation Society (HGVS), the Human Variome Project (HVP), and the Human Genome Organisation (HUGO)

First published: 25 March 2016 Full publication history

DOI: 10.1002/humu.22981 View/save citation

Variants detected by NGS

- SNV – (single nucleotide variant) G>T, A>C, etc.,
- Small insertions and deletions (many pipelines have difficulties reliably identify indels > 20 bp)
- Copy number variations (CNV)
- Structural variants (large inversions, insertions, deletions over several KB, chromosomal translocations)
- Variant allele fraction (VAF) – number of variant reads / number of total (reference + alternate) reads

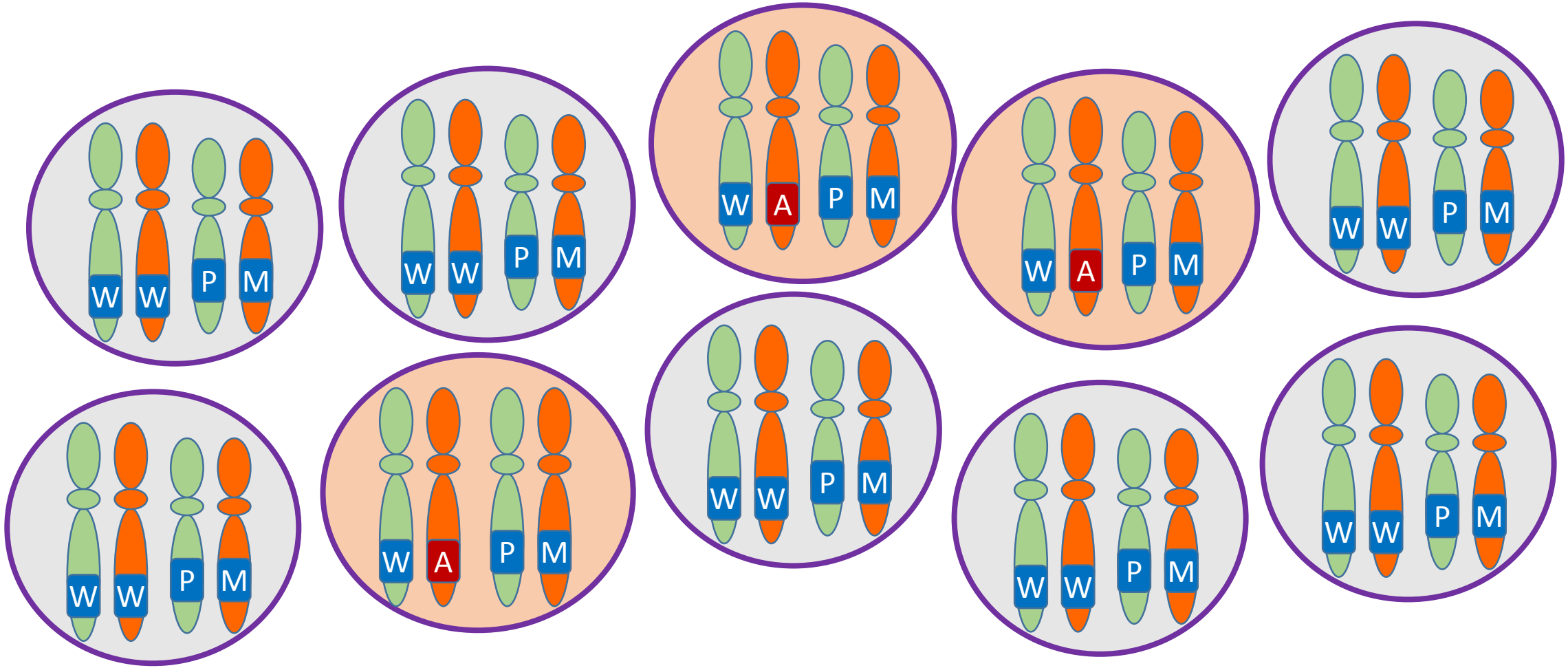
Most variants are heterozygous. A variant present in all the cells in a sample will have a VAF of 50%

A variant on chromosome X in a male patient is hemizygous. A variant present in all the cells in a sample will have a VAF of 100%

30% of cells have a heterozygous mutation

Germline SNP: VAF: P (paternal) = 50%, M (maternal) = 50%

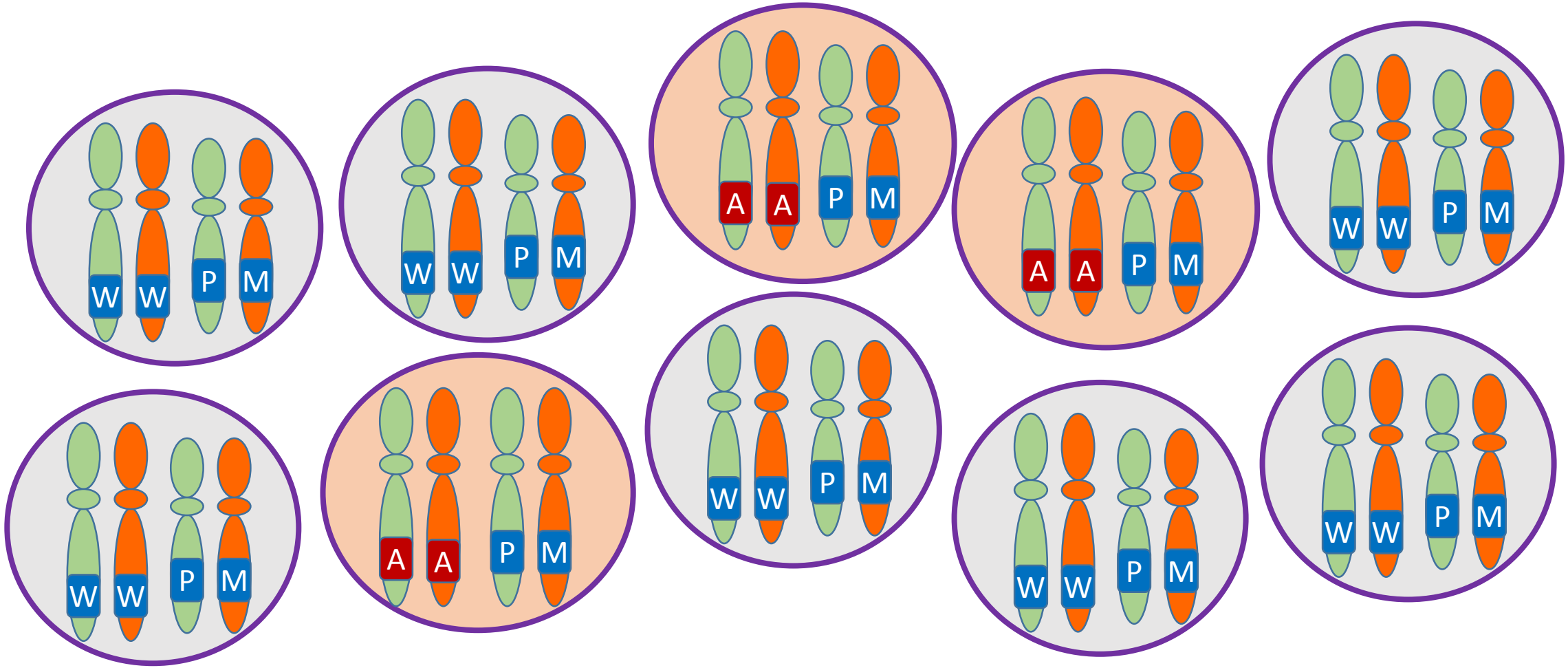
Somatic mutation: VAF: W (wild type) = 85%, A (alternate) = 15%



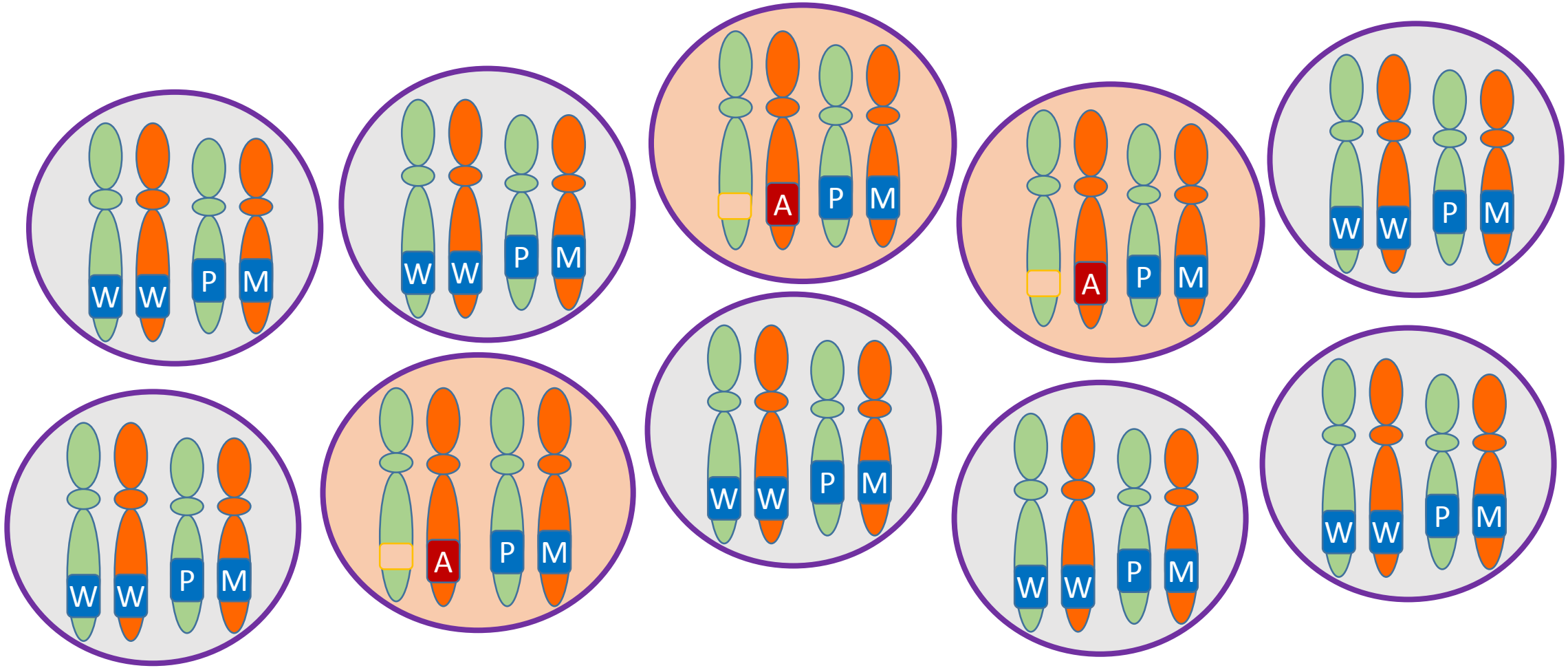
30% of cells have a mutation and loss-of-heterozygosity

Germline SNP: VAF: P (paternal) = 50%, M (maternal) = 50%

Somatic mutation: VAF: W (wild type) = 70%, A (alternate) = 30%



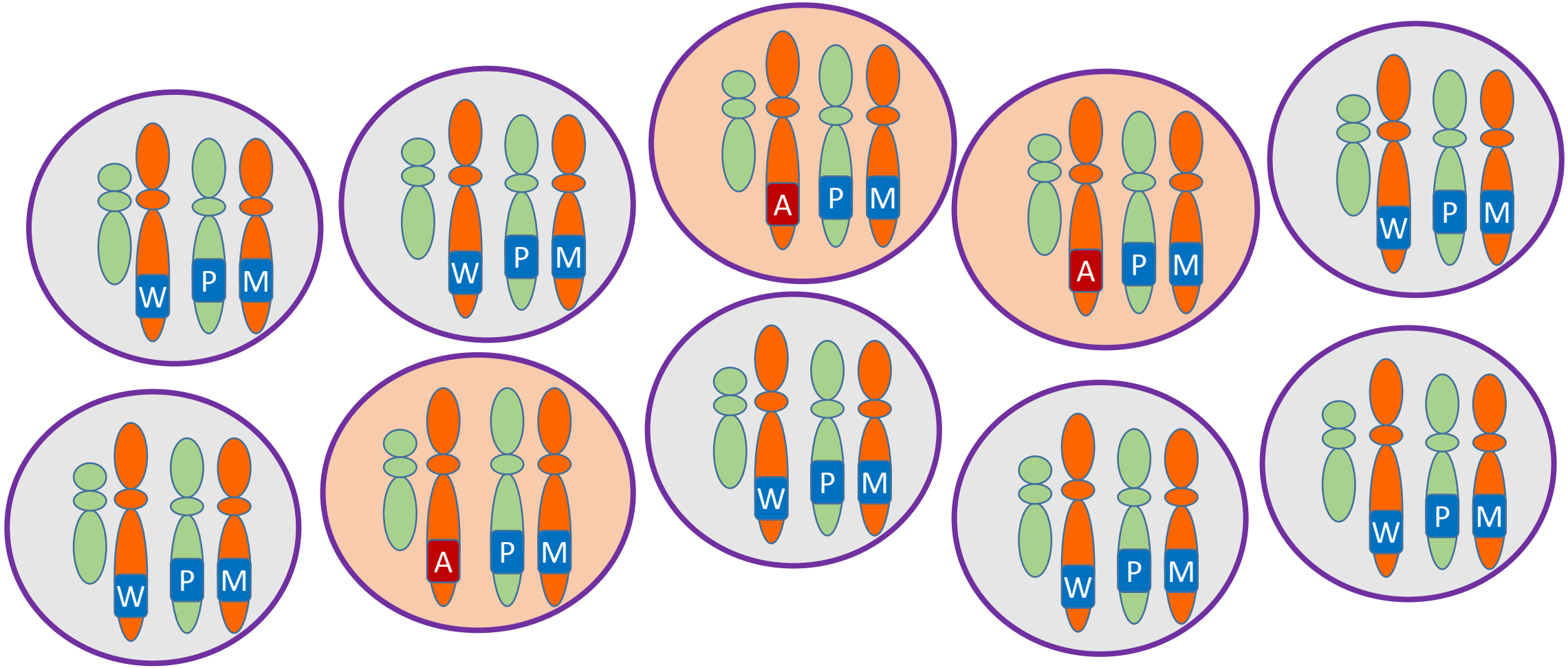
30% of cells have a heterozygous mutation + a deletion
Germline SNP: VAF: P (paternal) = 50%, M (maternal) = 50%
Somatic mutation: VAF: W (wild type) = 82%, A (alternate) = 18%



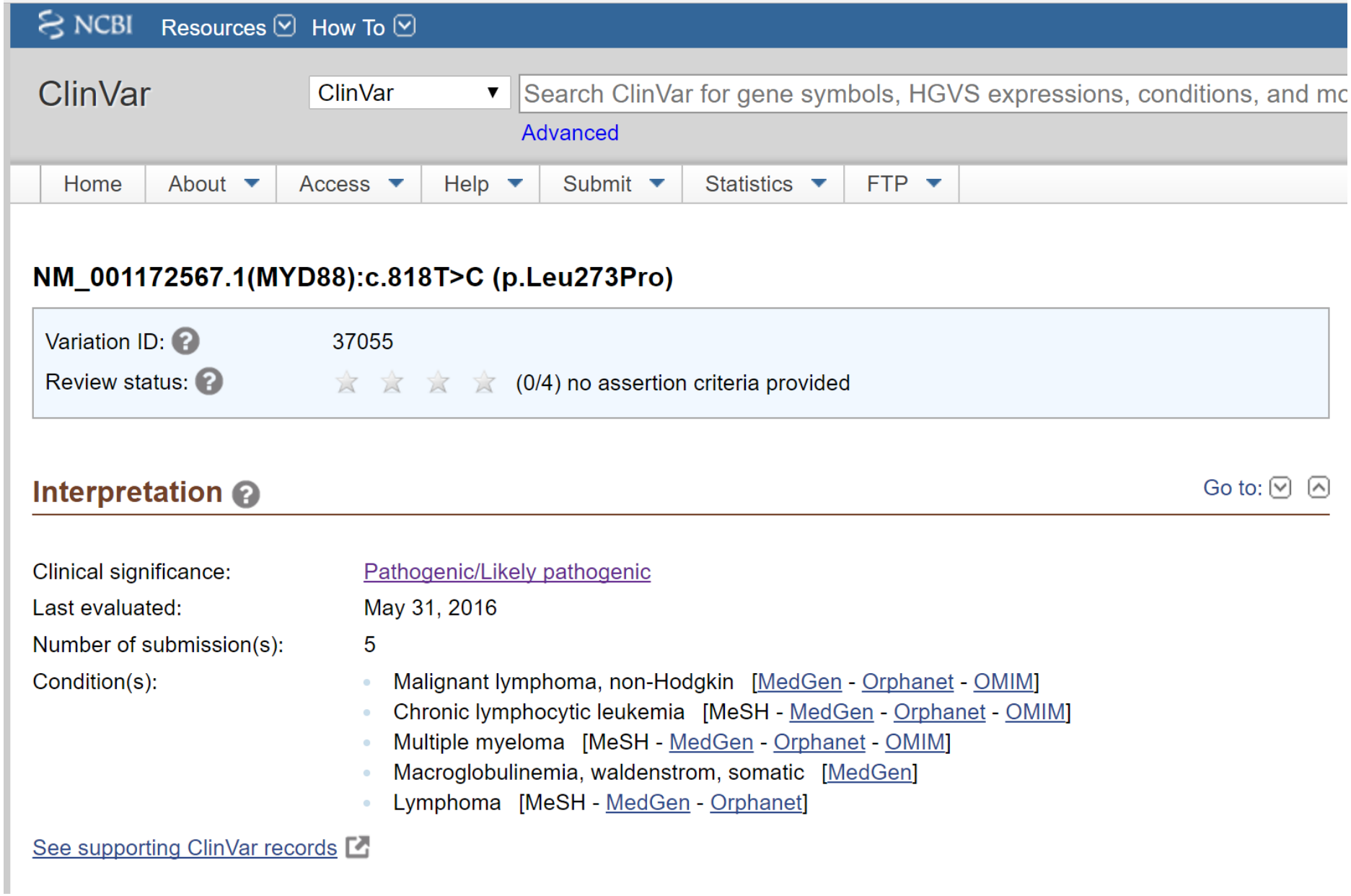
30% of cells have a hemizygous mutation

Germline SNP: VAF: P (paternal) = 50%, M (maternal) = 50%

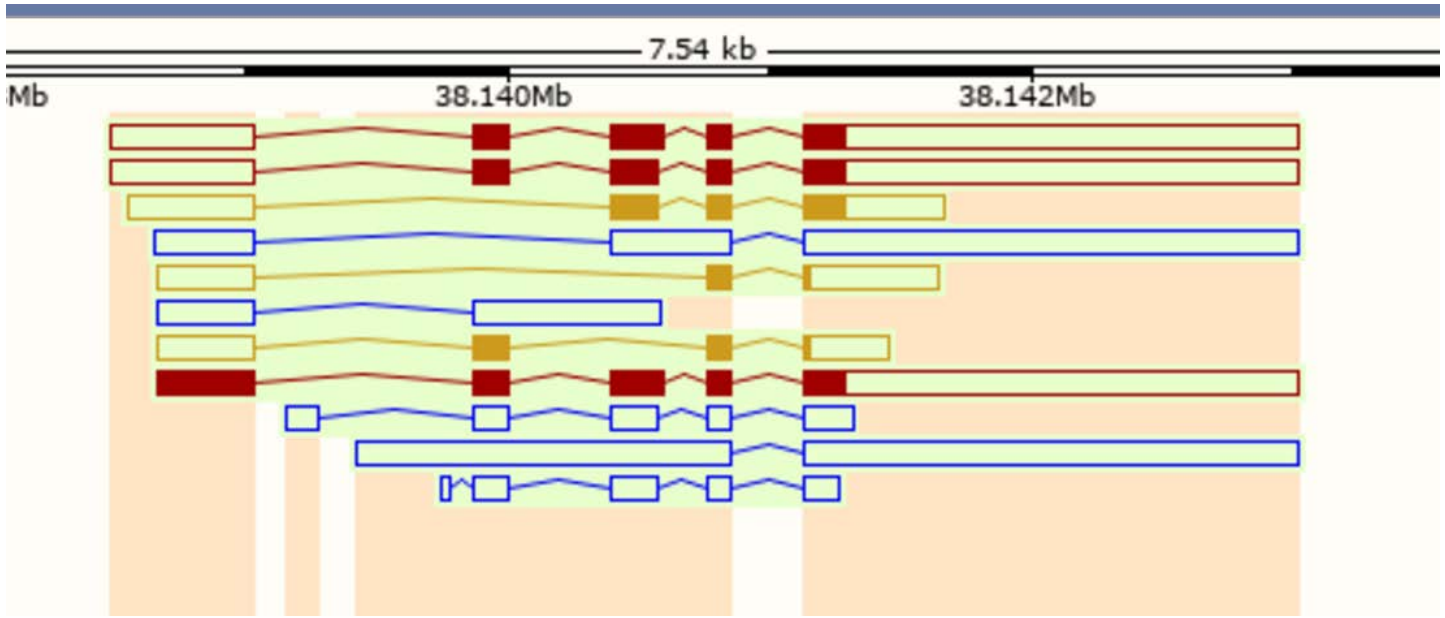
Somatic mutation: VAF: W (wild type) = 70%, A (alternate) = 30%



Why do we need the reference transcript?



NM_001172567.1(MYD88):c.818T>C (p.Leu273Pro)



Allele ID: 45735

Variant type: single nucleotide variant

Cytogenetic location: 3p22

Genomic location:
- Chr3: 38141150 (on Assembly GRCh38)
- Chr3: 38182641 (on Assembly GRCh37)

Protein change: L265P, L273P, *160R

HGVS:
- NG_016964.1:g.7673T>C
- NM_001172566.1:c.478T>C
- NM_001172567.1:c.818T>C
- NP_001166037.1:p.Ter160Arg
- NP_001166038.1:p.Leu273Pro
- NP_002459.2:p.L265P
- NC_000003.12:g.38141150T>C (GRCh38)
- LRG_157t1:c.818T>C
- NC_000003.11:g.38182641T>C (GRCh37)
- LRG_157p1:p.Leu273Pro
- LRG_157:g.7673T>C

...less

Note: Note that rs38182641, from OMIM 602170.0004, is incorrect.

Links:
- OMIM: 602170.0004
- dbSNP: 387907272

NCBI 1000 Genomes Browser: rs387907272

Molecular consequence:
- NM_001172566.1:c.478T>C: stop lost SO:0001578
- NM_001172567.1:c.818T>C: missense variant SO:0001583

Allele frequency:
- GMAF 0.00020 (C)
- ExAC 0.00010 (C)

Why do we need both c. and p.?

***BRAF* p.V600E**

c.1799T>A GTG (V) -> GAG (E) >95%

c.1799_1780TGdelinsAA GTG (V) -> GAA (E) <5%

So two tumors both stains positive for BRAF V600E specific antibody by IHC does not guarantee that they have exactly the same mutation (or they are clonally related!)

Why do we need variant allele fraction (VAF) and coverage?

A 70-year-old anemic male found to have DNMT3A
NM_175629.2:c.2645G>A p.R882H in the peripheral blood.

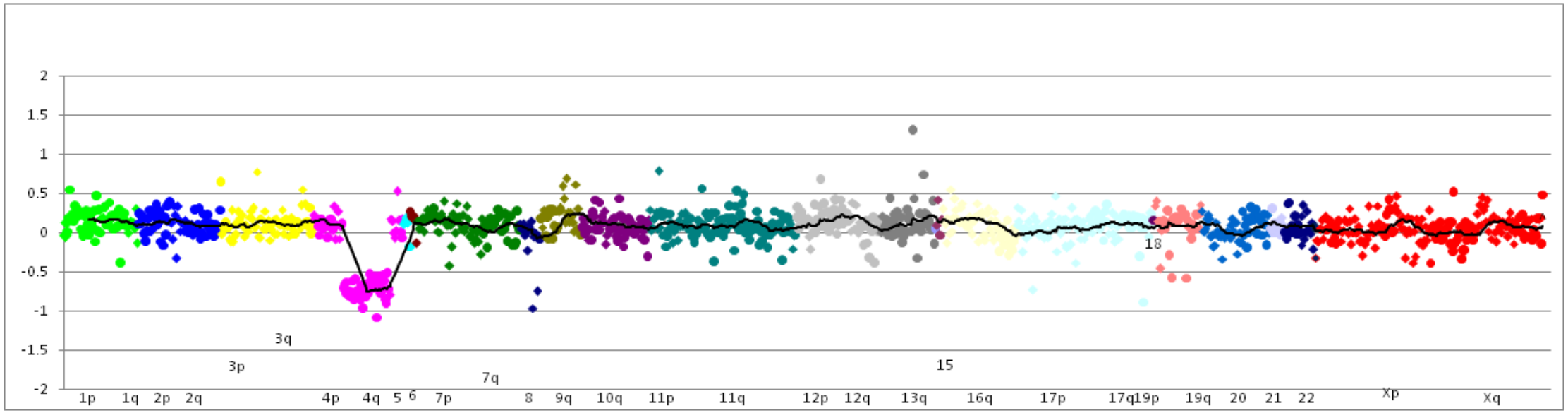
VAF = 1% total reads = 1000 -> ICUS
VAF = 10% total reads = 1000 -> CCUS or MDS
VAF = 10% total reads = 10 -> repeat testing

Apparent increase of VAF >>50% may be caused by

Copy-number neutral Loss-of-heterozygosity (uniparental disomy): the sequence on the wild type allele is replaced by the mutant allele.

Deletion (loss) of the wild type allele (Copy number variant [CNV])

- ***DNMT3A*** **NM_175629 c.2101T>A p.F701I - in 44.0% of 554 reads**
- ***TET2*** **NM_001127208 c.480_481insA p.V160fs* - in 75.2% of 612 reads**
- ***JAK2*** **NM_004972.3:c.1849G>T p.V617F – in 97.1% of 725 reads**
- ***NRAS*** **NM_002524.4:c.35G>A p.G12D – in 6.8% of 915 reads**



Summary

- Standardize! Follow the guideline and recommendation
- Use the full HGVS nomenclature
- Variant allele fraction and coverage are important
- To be (pathogenic) or not to be, that is the question…