The Elizabeth H. and James S. McDonnell III
McDonnell Genome Institute
at Washington University

Washington University in St. Louis
School of Medicine

PMBIO Module 05

Somatic. Somatic WGS and Exome Variant Analysis

Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia
Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)
29 October – 2 November, 2018
Glasgow

Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

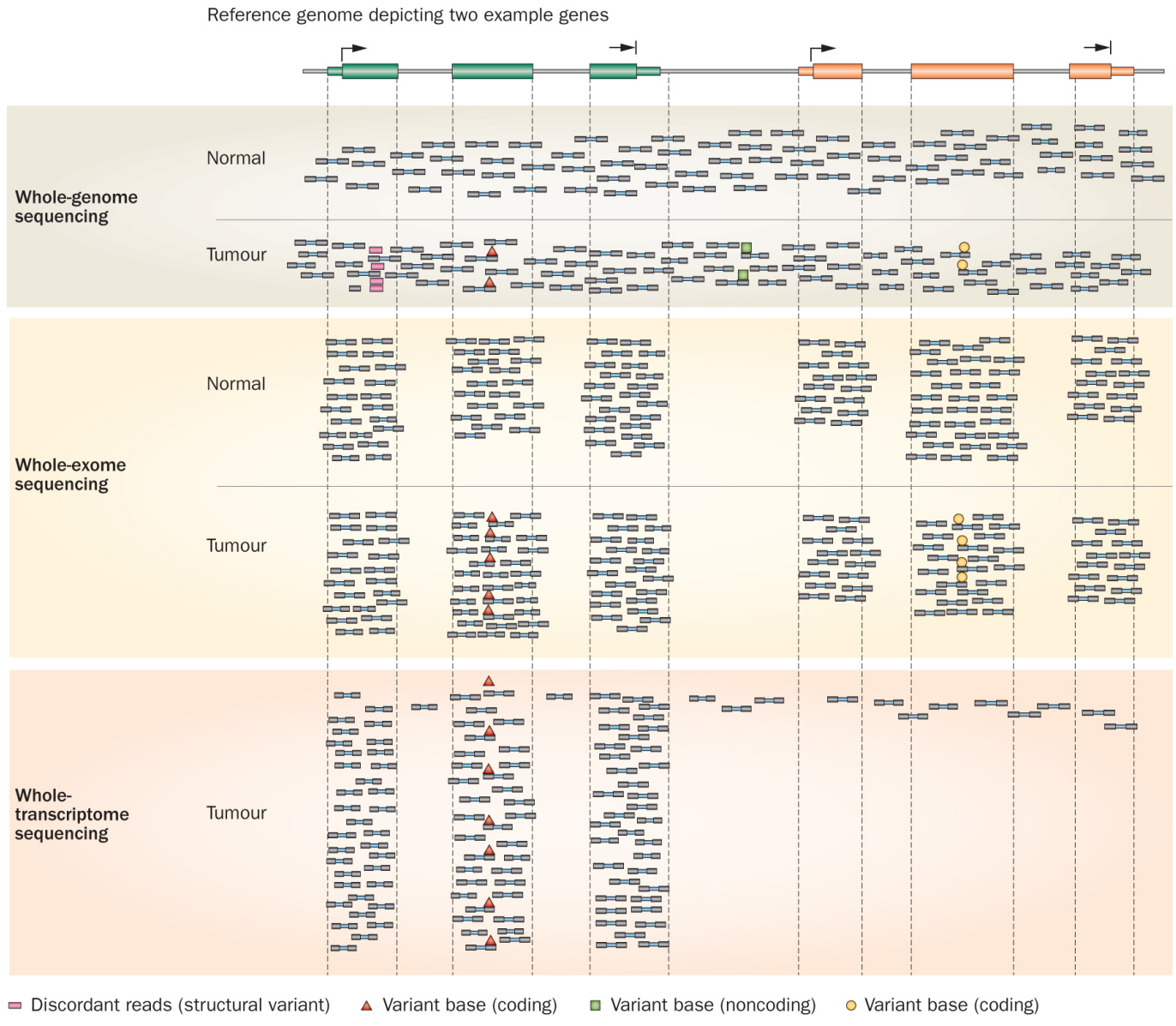
No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Learning objectives of module 05: Somatic

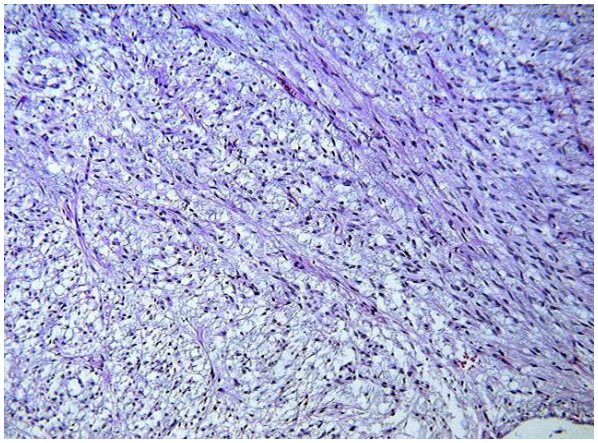
- **Key concepts**: Somatic variation, variant types (SNVs, small indels, CNVs, SVs, LOH), VCFs, variant allele fraction (VAF), purity estimation and tumor clonality

- Compare and contrast the merits of exome and WGS data for variant calling
- Compare germline and somatic variant calling strategies
- Consider the features of major variant types detected in NGS data
- Perform somatic variant calling of various types using tools specific to each type
- Understand the basic features of the VCF format
- Merge multiple VCFs into a single combined VCF
- Perform variant filtering to identify a high quality set of variants
- Annotate variants with respect to transcript annotations, population frequency, predicted function, etc.
- Manually review variants of each type to better understand how variant callers use read alignment data to identify variants

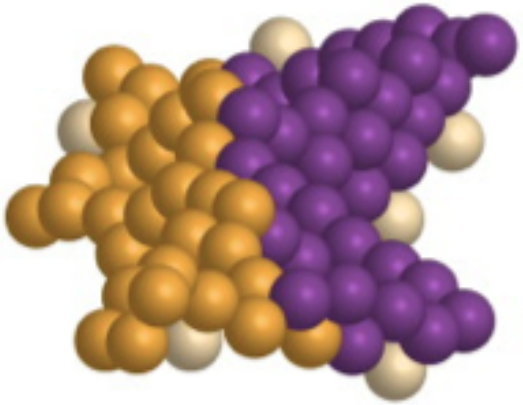
Whole genome, exome and transcriptome sequencing allows us to detect and confirm many different variant types



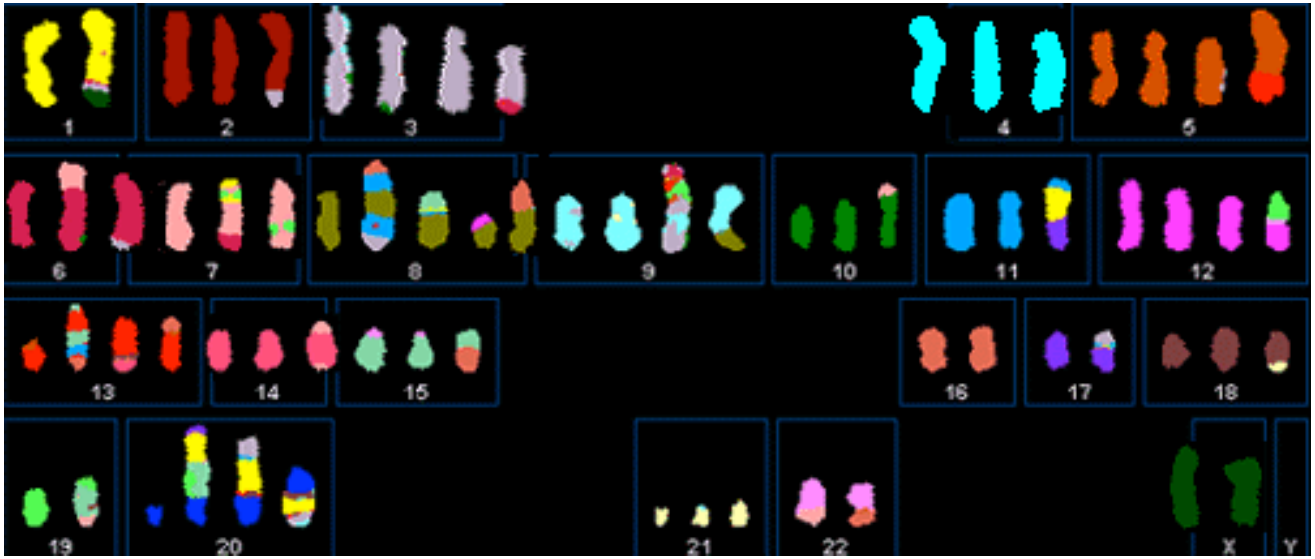
Tumors are often impure, heterogeneous, and aneuploid



Tumors are often impure (contain normal cells)



Tumors are often genetically diverse collections of cells



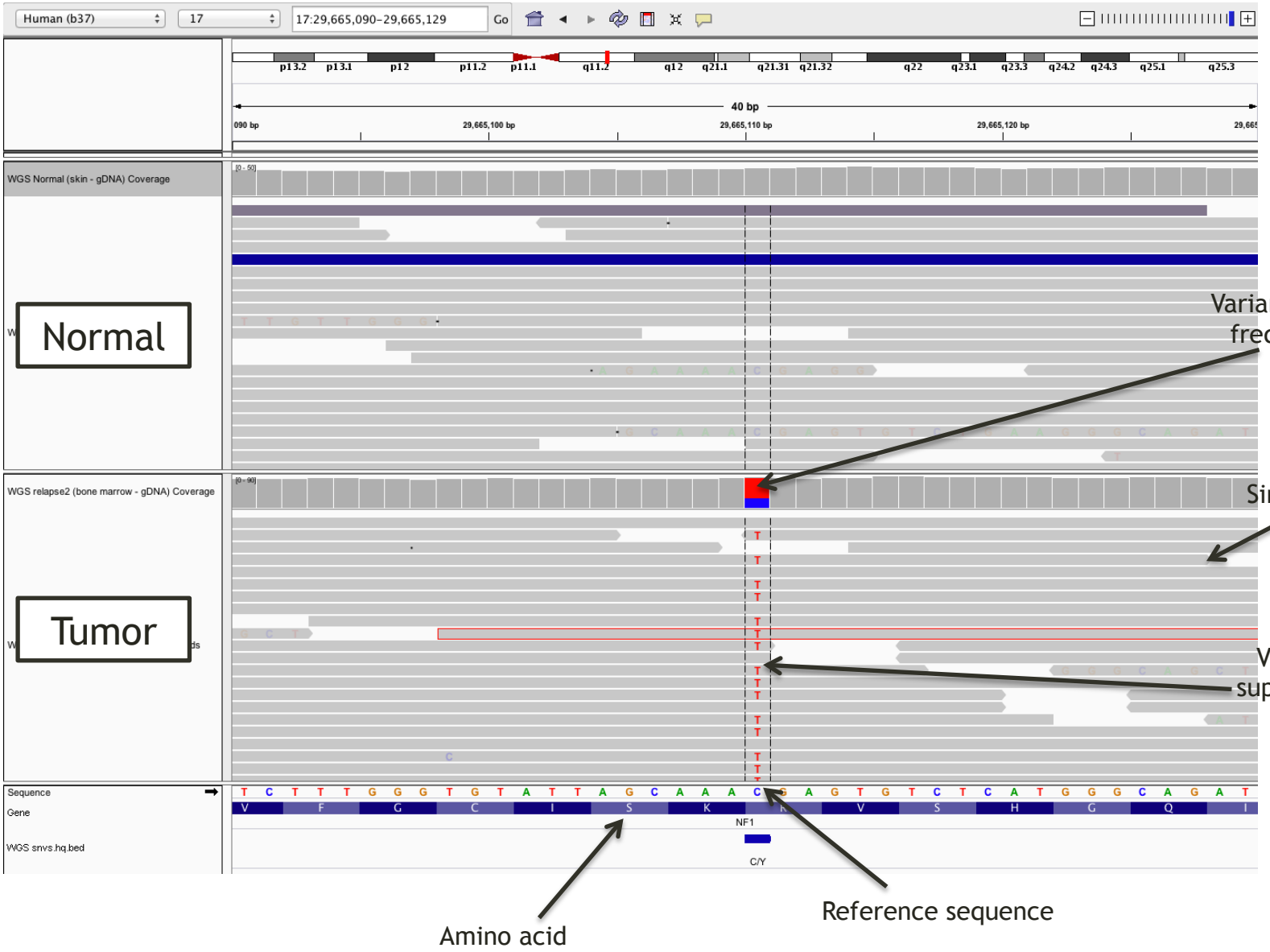
Tumors may be aneuploid

Somatic mutation vs. germline mutation vs. germline polymorphism

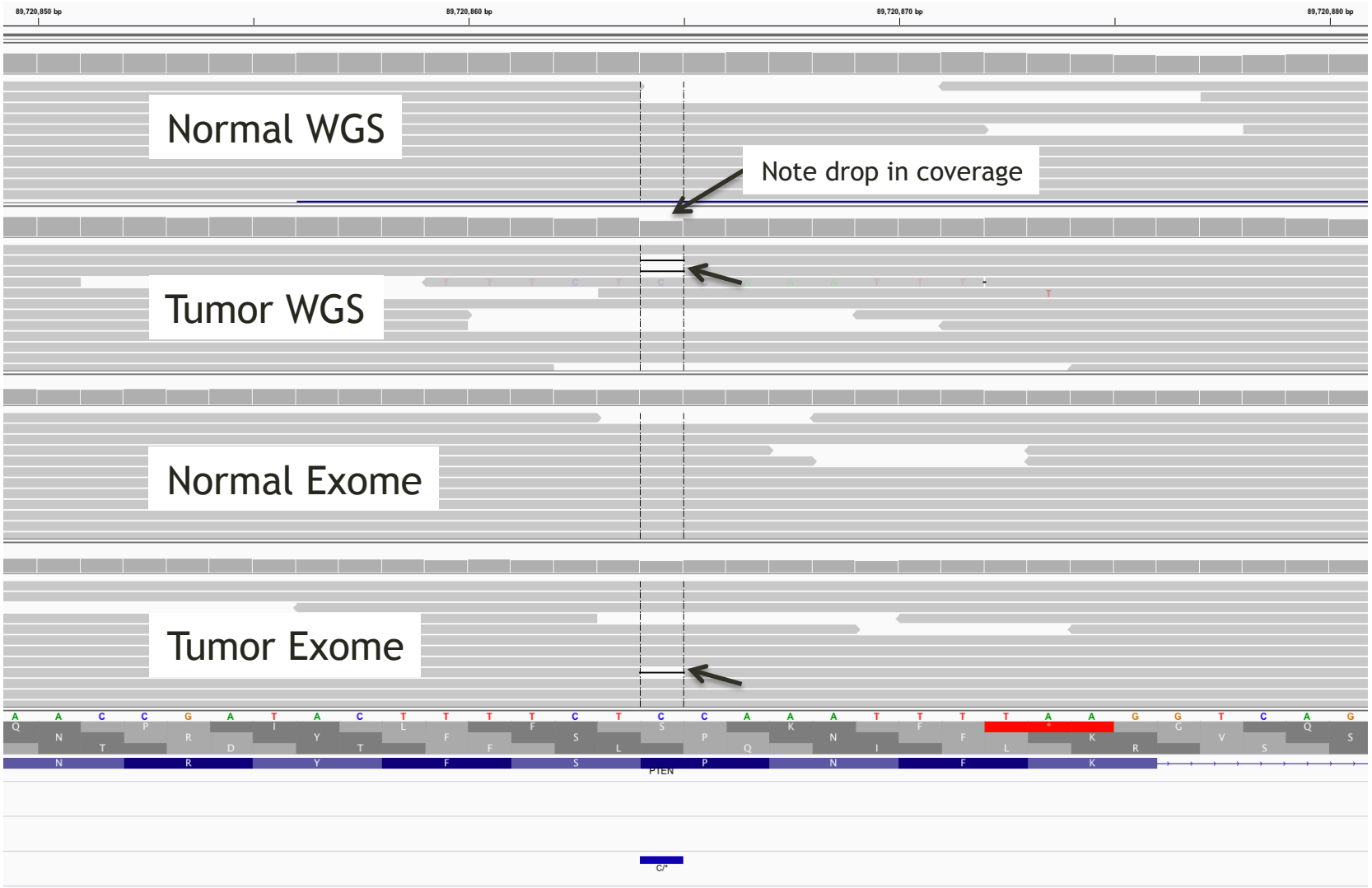
- Somatic mutations are best distinguished by adequate sequencing of a matched normal
  - Affected and unaffected family members may help to determine origin of a germline mutation

- Comparison of variants to variant databases can also help to classify variants as:
  - Germline polymorphisms
    - 1000 genomes
    - Exome sequencing project (~6,500 individuals)
    - ExAC, Exome Aggregation Consortium (~60,000 individuals)
    - gnomAD browser (123,136 WXS and 15,496 WGS)
  - Germline mutations
    - OMIM, HGMD, PharmGKB, ClinVar
    - ACMG guidelines
    - Gemini
  - Somatic mutations
    - By inference if the mutation is not a common polymorphism (often a weak inference) or is a classic hotspot mutation

Single nucleotide variants (SNVs) appear as single base alignment discrepancies from reference genome

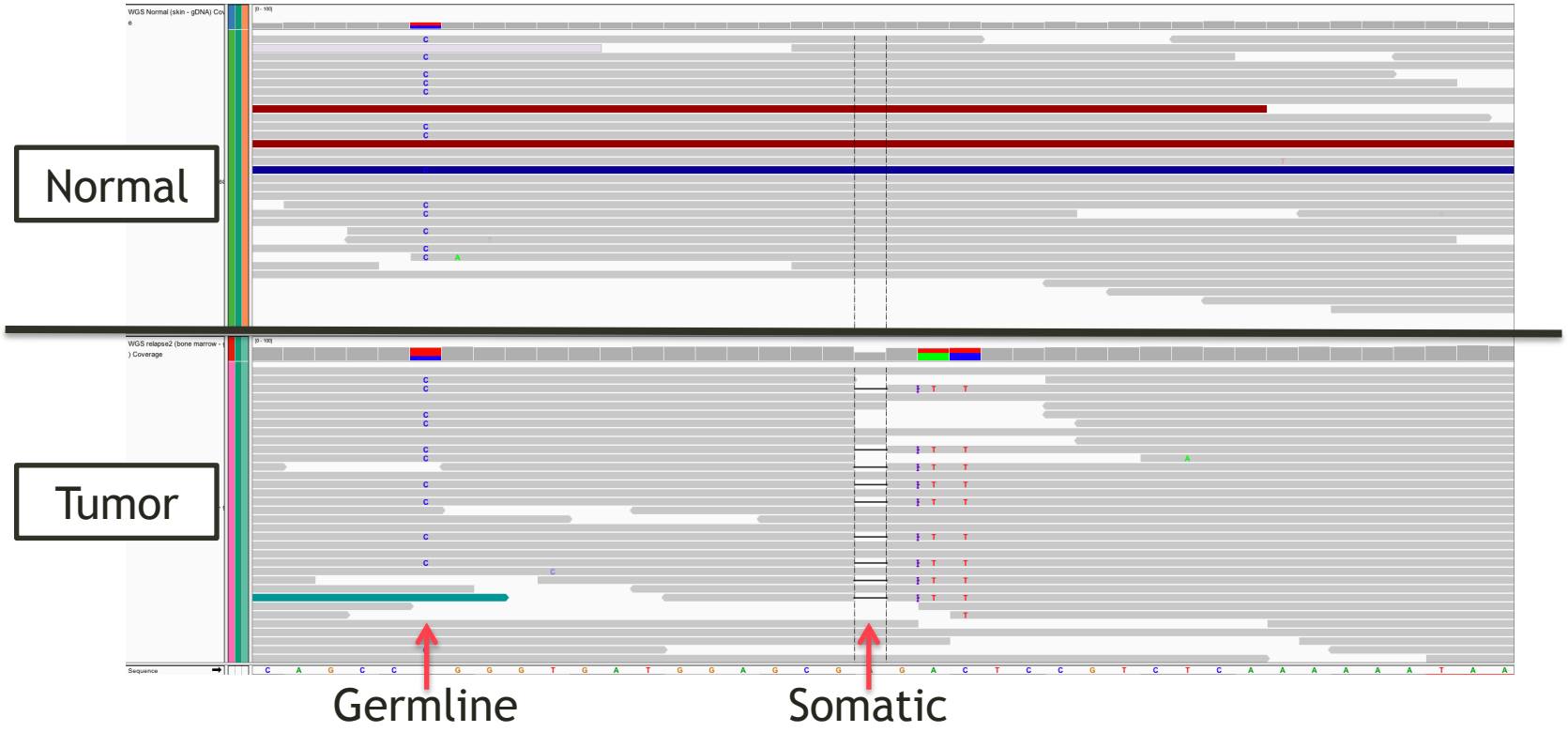


Small insertions and deletions (indels) appear as bases lost or gained in reads compared to reference



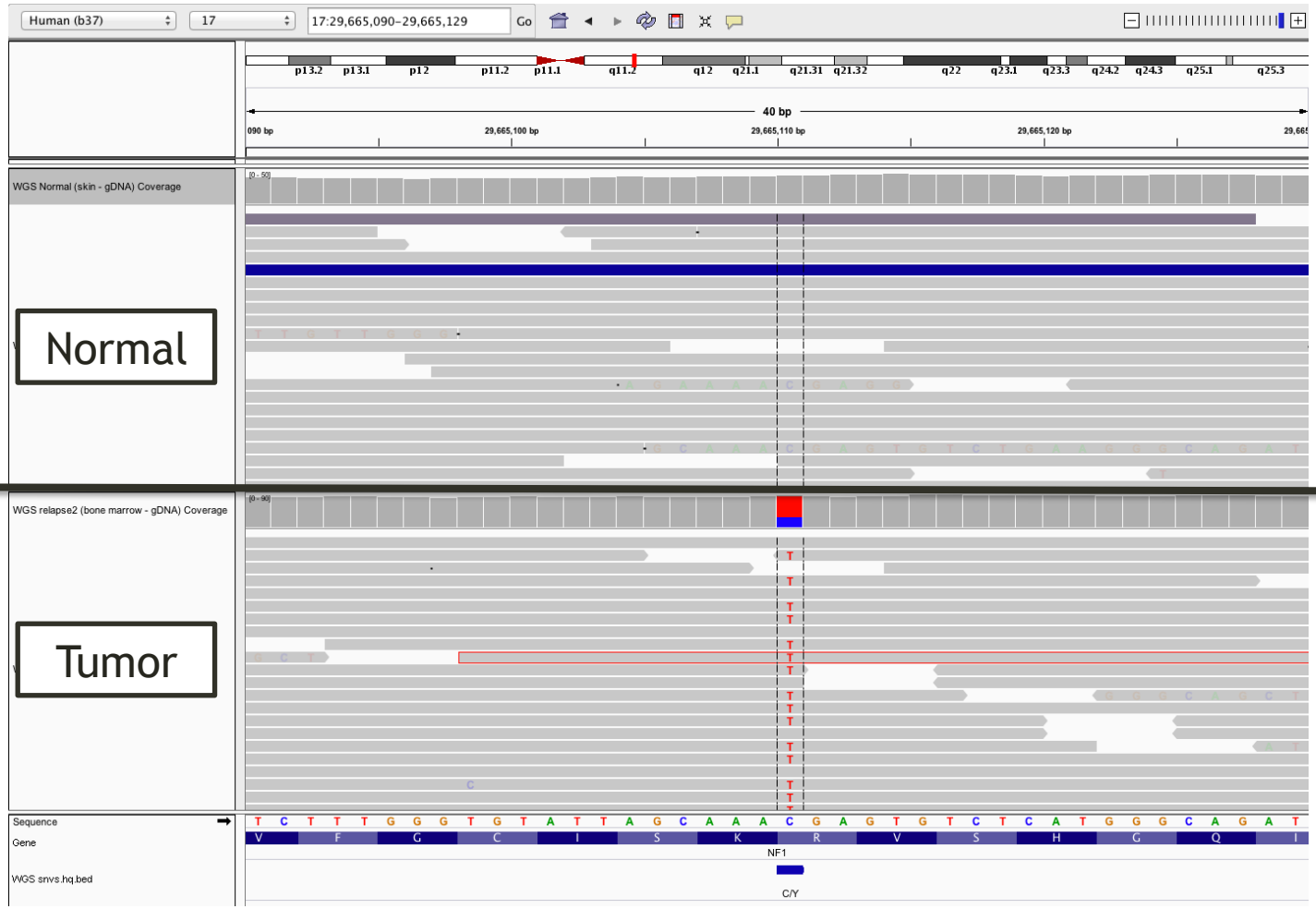
Somatic versus germline demonstration

- Germline mutations
  - Present in egg or sperm
    - All cells of affected offspring
  - Heritable
  - Cause of familial cancers
- Somatic mutations
  - Occur in non-germline tissues
    - Only tumor cells (breast, lung, blood, etc.)
  - Non-heritable
  - Cause of sporadic cancers



Variant allele frequency (VAF)

VAF = Variant reads / Total reads

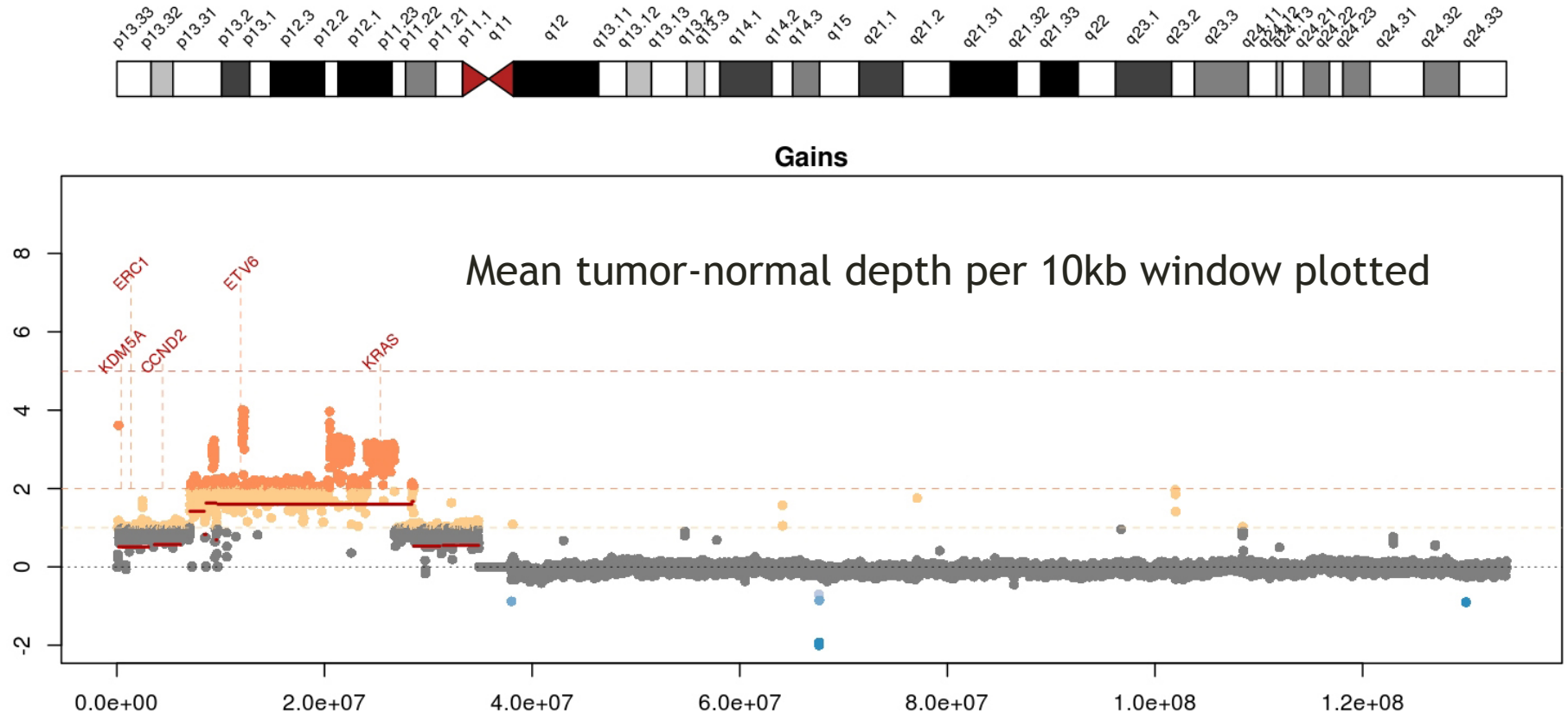


VAF = 0/20 = 0%

VAF = 14/20 = 70%

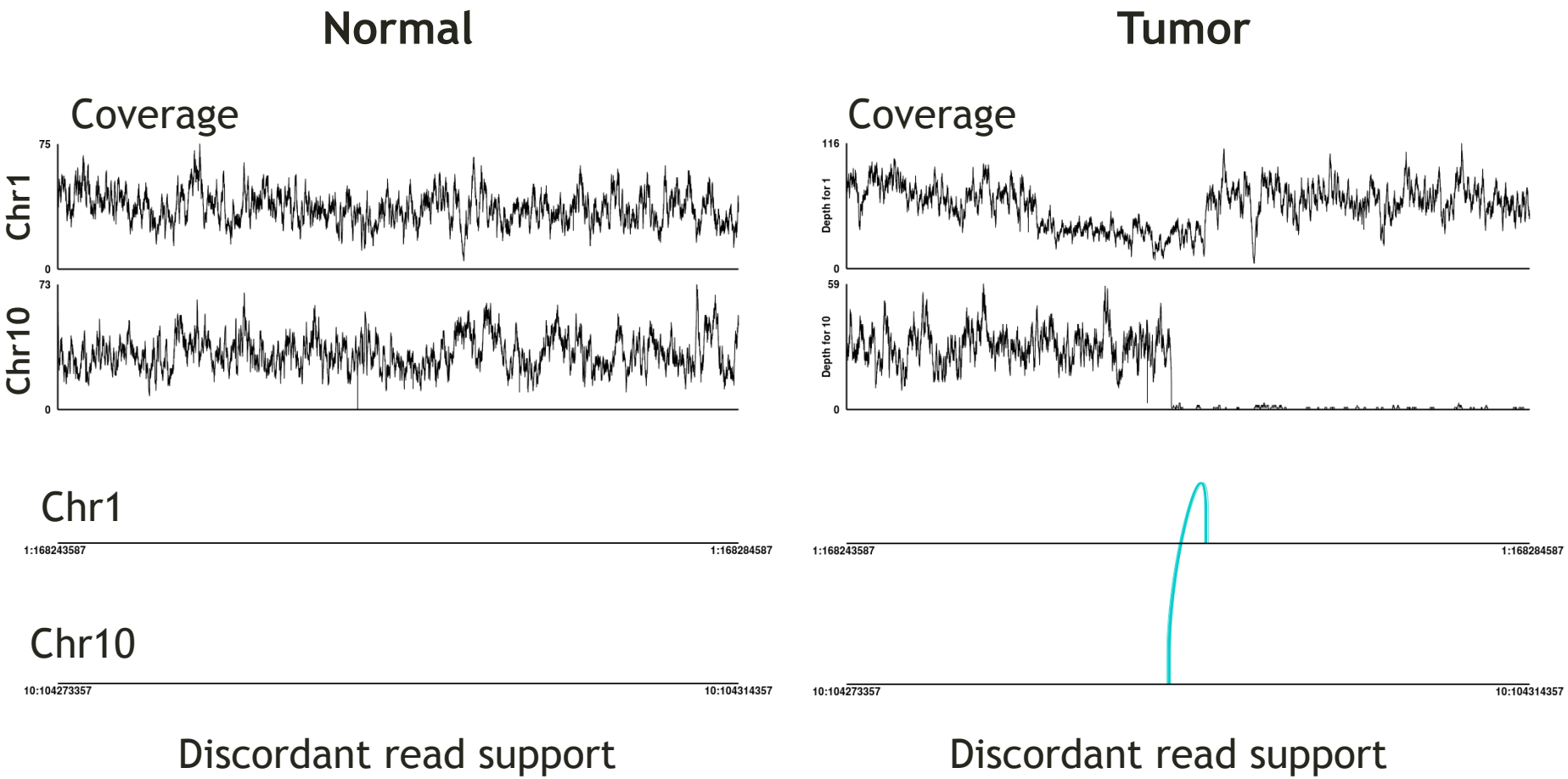
A heterozygous variant is expected to have VAF = 50%. Often not true due to sample purity, tumor heterogeneity, sampling error, alignment issues, copy number variation, etc.

Copy number variants (CNVs) appear as deviations in alignment “depth” or “coverage”



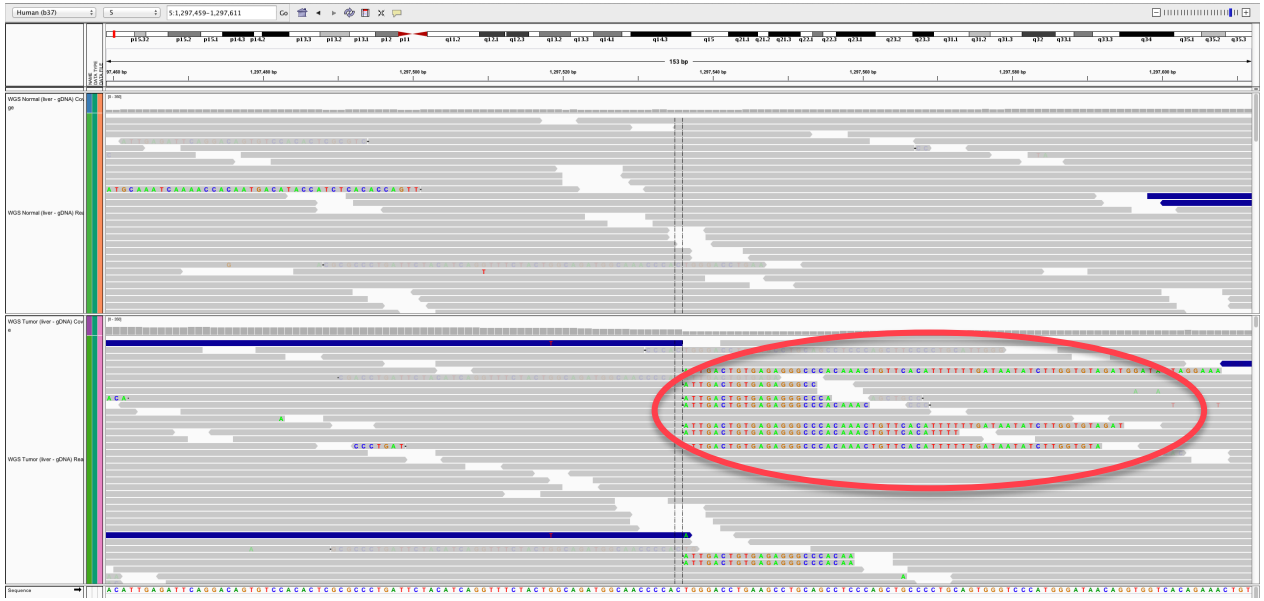
KRAS amplification in a metastatic breast cancer

Structural variants (SVs) can be identified using a combination of coverage and discordant read alignments



A Chr1-Chr10 (TBX19-SUFU) unbalanced translocation identified in an adult acute lymphocytic leukemia.

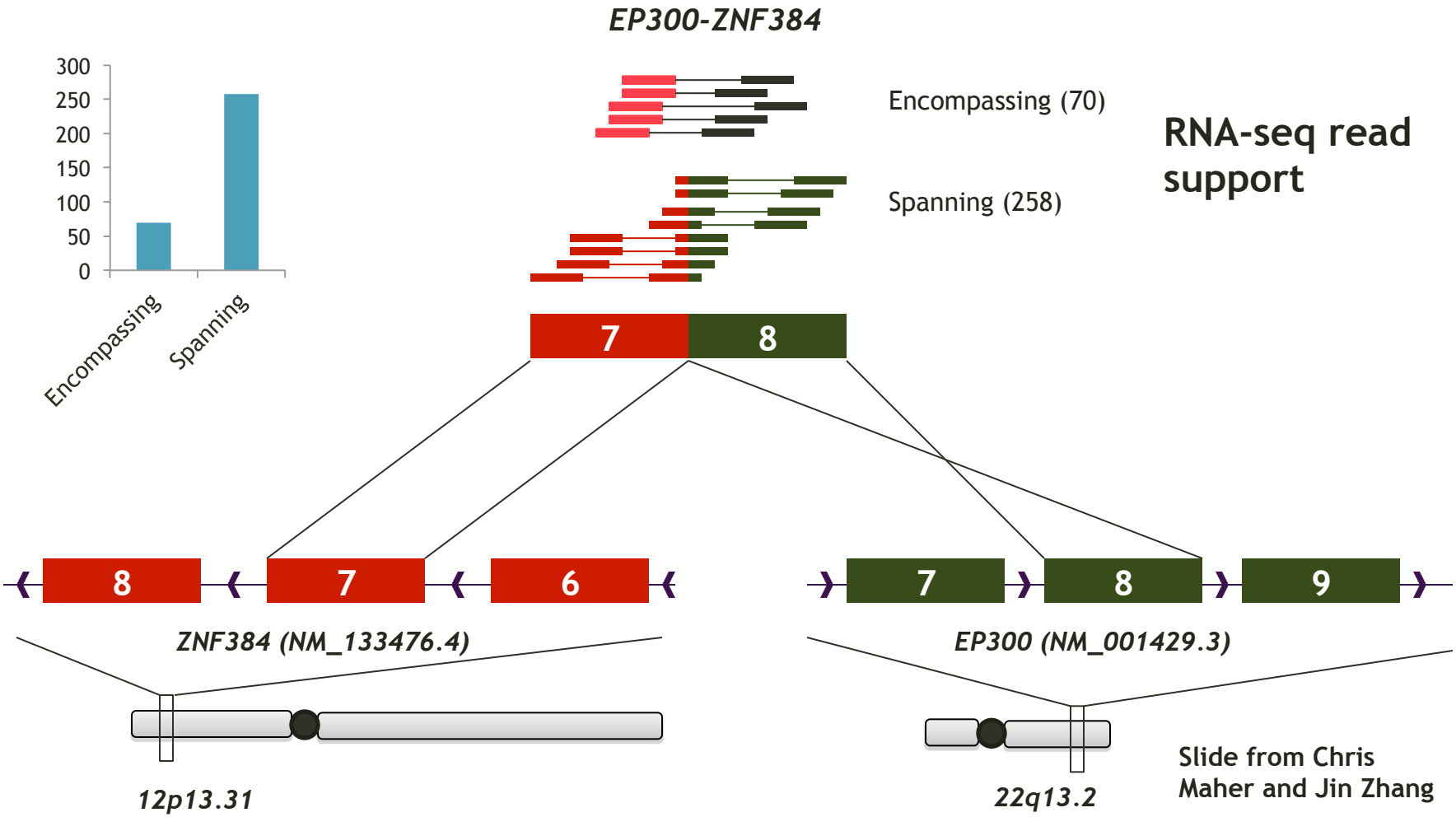
Viral integrations can be detected by analysis of WGS data (often with a custom reference)



- Many reads align directly to viral genome
- Soft-clipped reads in human reference help identify integration site

Malachi Griffith <mgriffit@genome.wustl.edu>

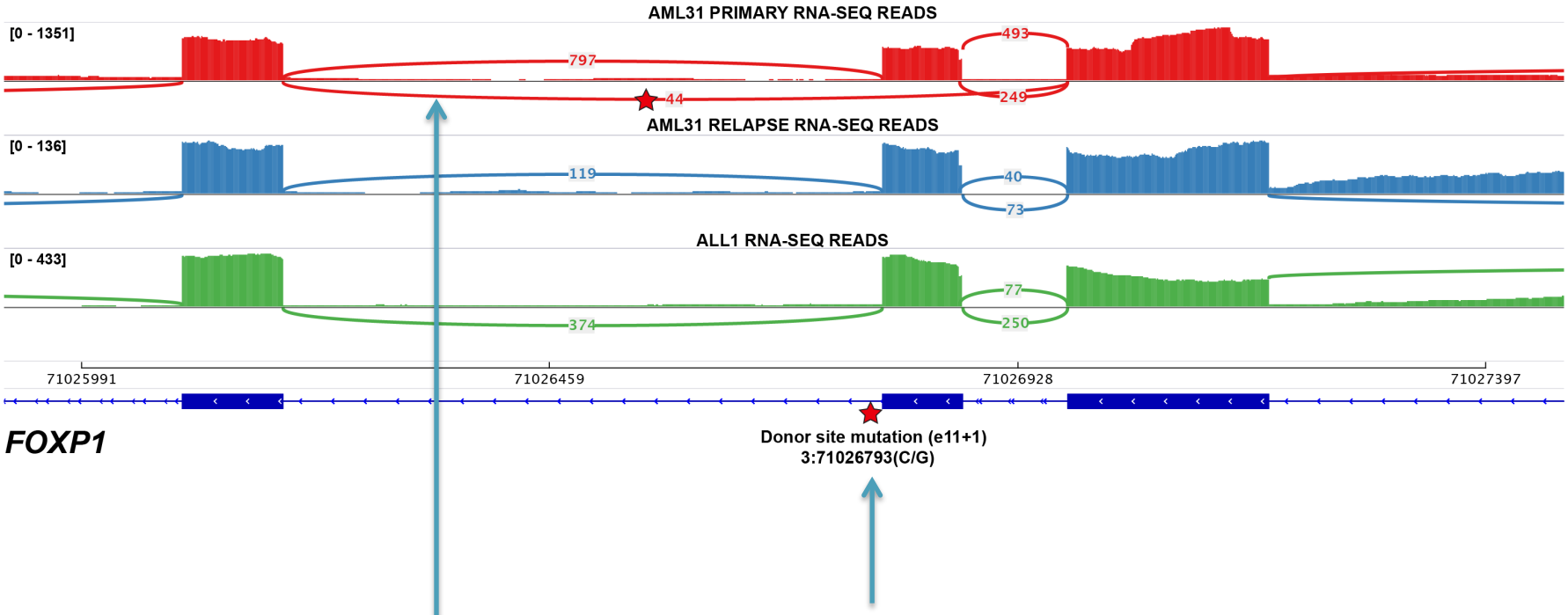
Expressed gene fusions can be identified by discordant read alignments spanning known exons from RNA-seq data



Exons 1-8 of EP300 fused to exons 7-10 of ZNF384 in head-to-tail fashion.

RNA-seq can also reveal the splicing consequence of somatic mutations detected in WGS

A. Sub-clonal somatic splicing event in *FOXP1* observed in the primary tumor but cleared in relapse

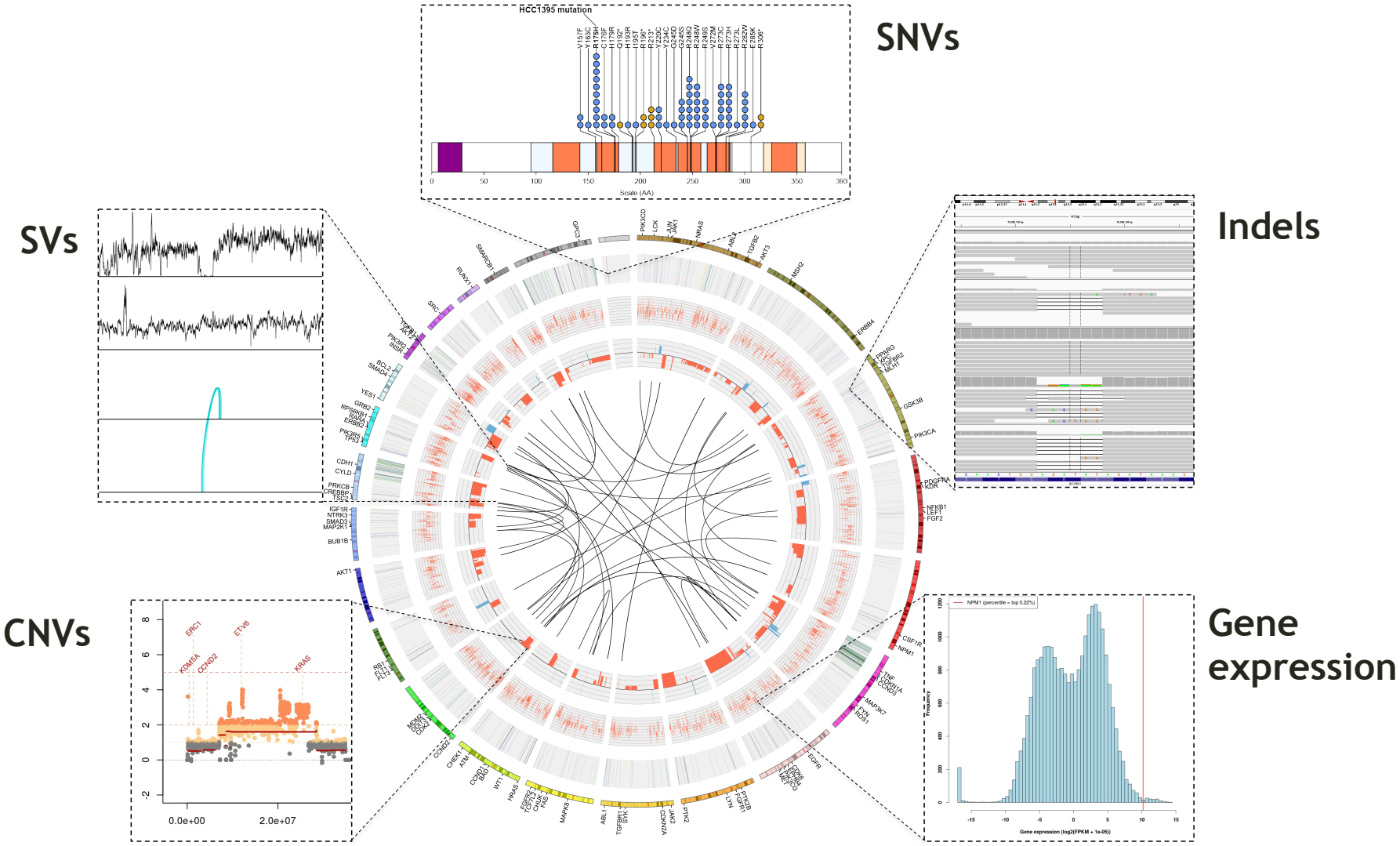


Splice site mutation

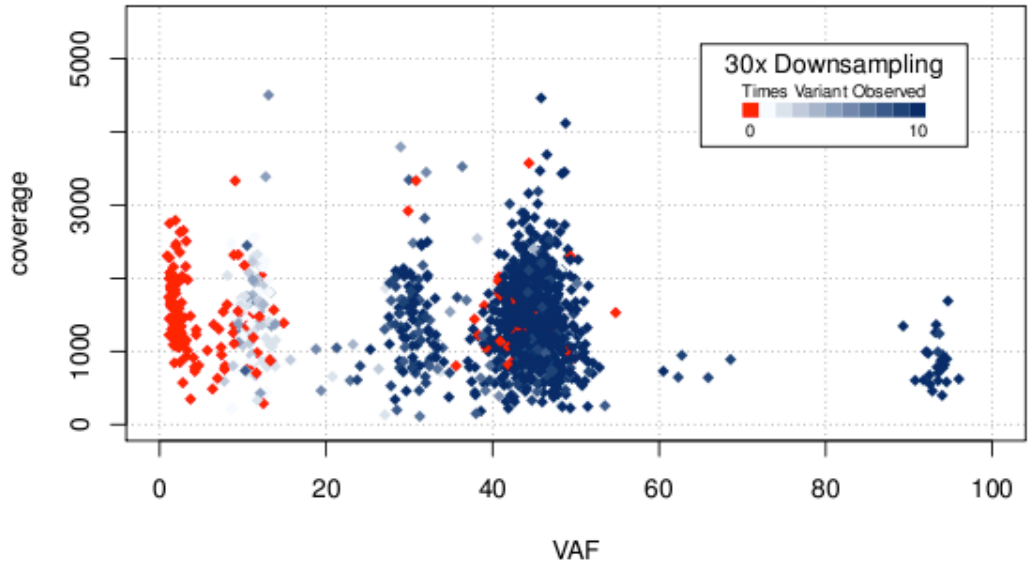
Read alignments support the skipping of affected exon

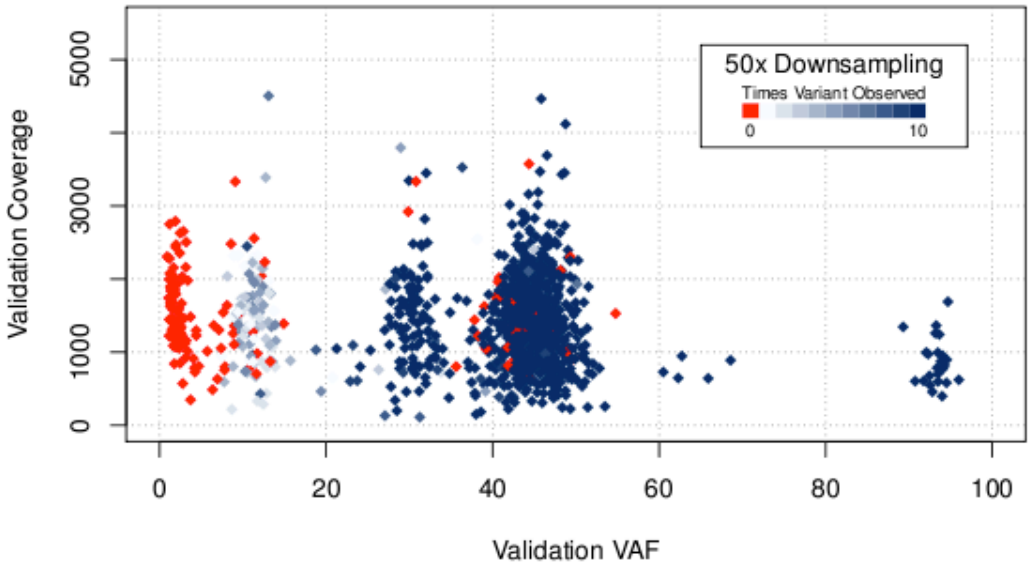
We created 'RegTools' to help characterize these kinds of events

Tumor genome analysis will typically reveal dozens to thousands of alterations of multiple types

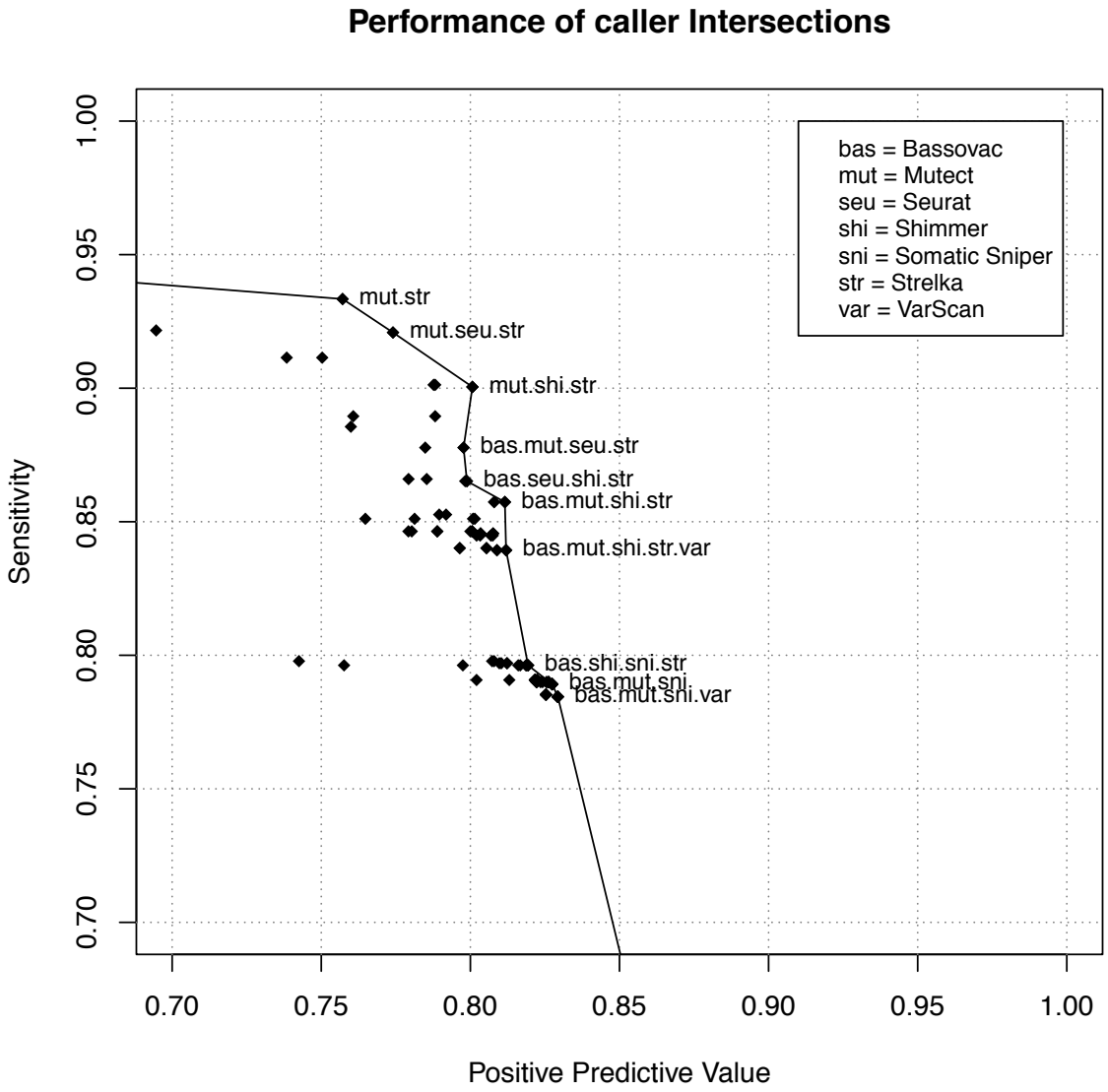


Depth is critical to overall sensitivity of somatic variant discovery

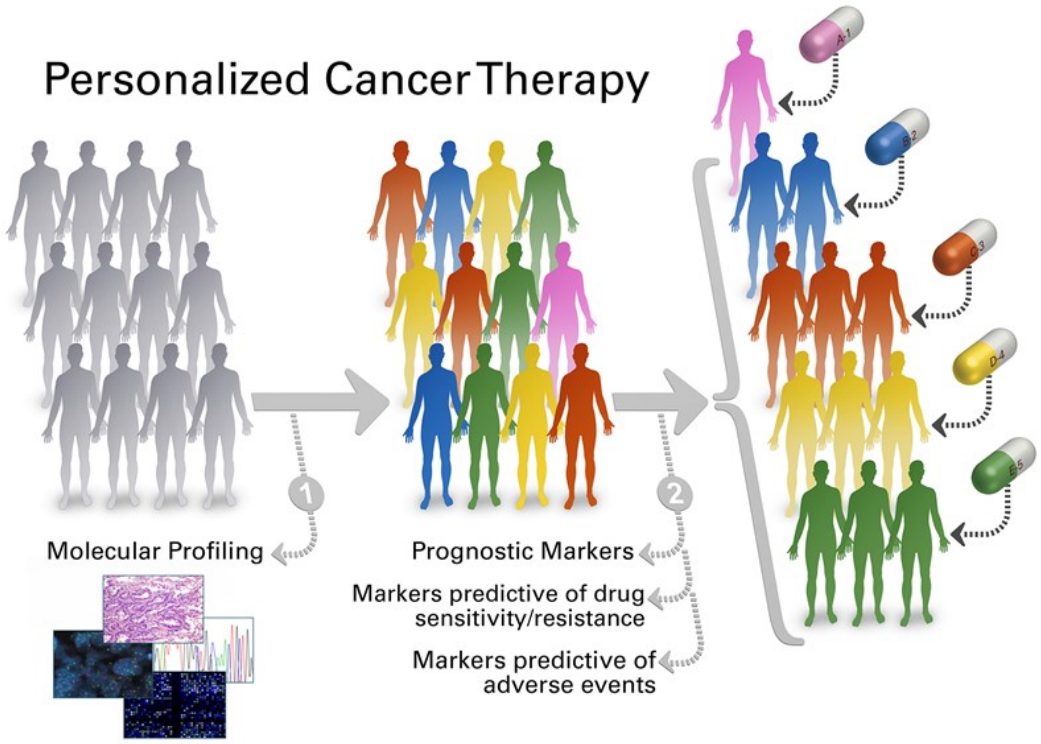




Use of multiple variant callers can improve sensitivity and accuracy



One goal of somatic variant detection is to enable precision medicine targeting of driver mutations



1980s: Development of Targeted Therapies

2000: Human Genome Sequencing Project

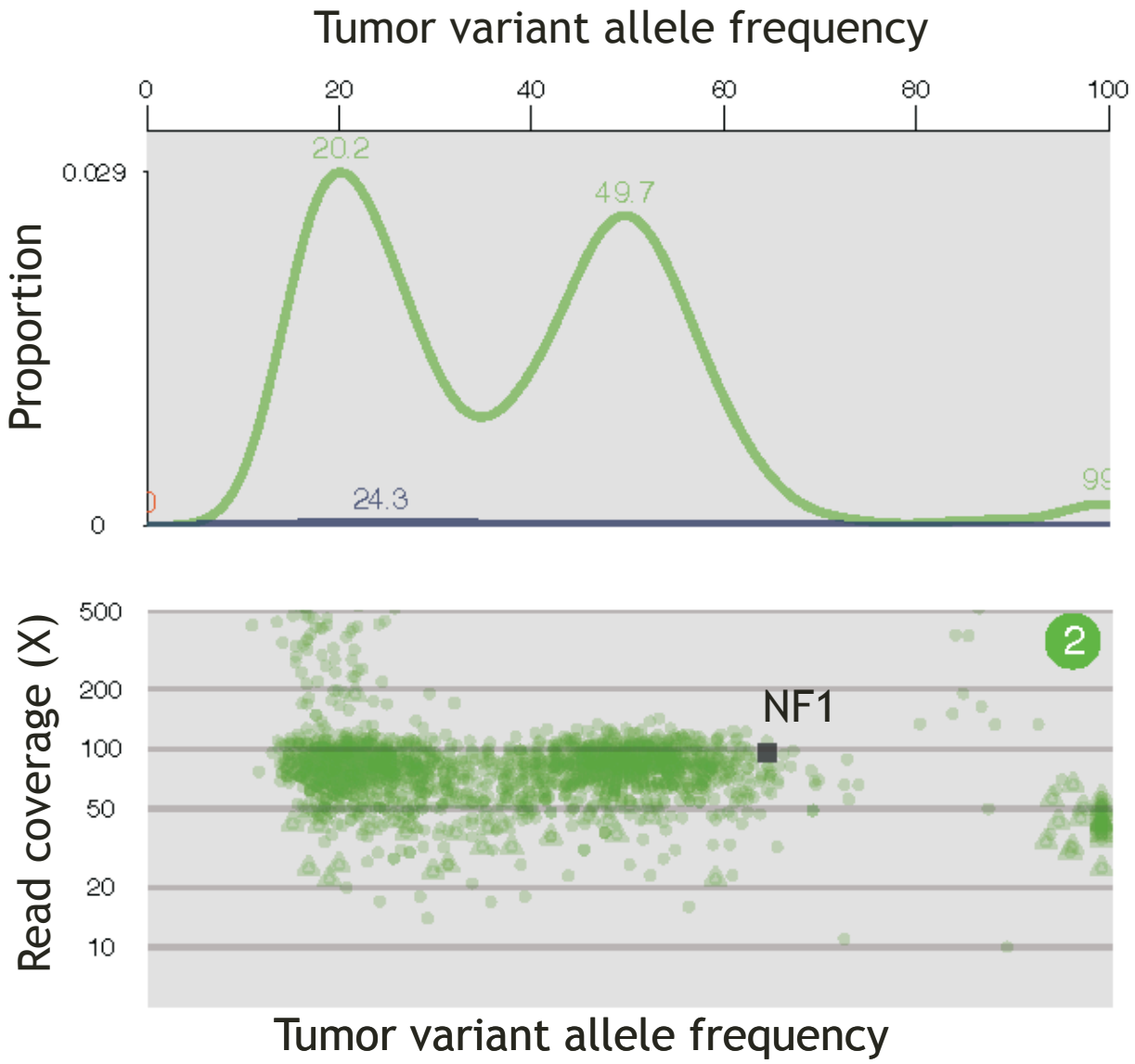
2001: FDA Approval for Imatinib in BCR-ABL1 CML

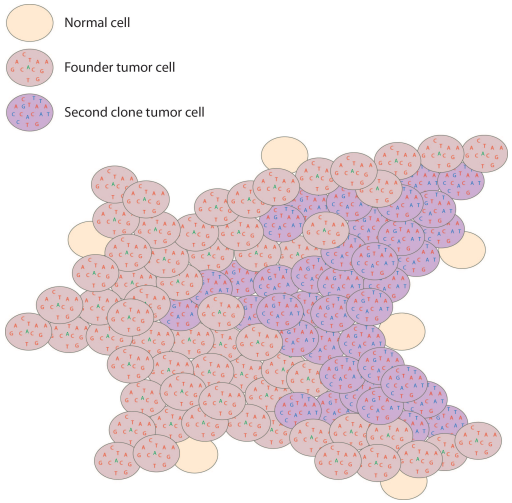
2014: FDA Approval for BRCA Testing

2017: FDA Approval of Keytruda for MSI tumors

BRAF -> V600E -> Melanoma -> Predictive -> Vemurafenib
ERBB2 -> Amplification -> Breast -> Predictive -> Traztuzumab
EGFR -> L858R -> Lung -> Predictive -> Erlotinib
ALK -> Fusions -> Lung -> Predictive -> Crizotinib
EWSR1-FLI1 -> Fusions -> Ewing Sarcoma -> Diagnostic
DNAJB1-PRKACA -> Fusions -> flHCC -> Diagnostic
VHL -> Loss of function mutations -> Kidney -> Predisposing
… an increasingly long tail of rare but clinically relevant variants

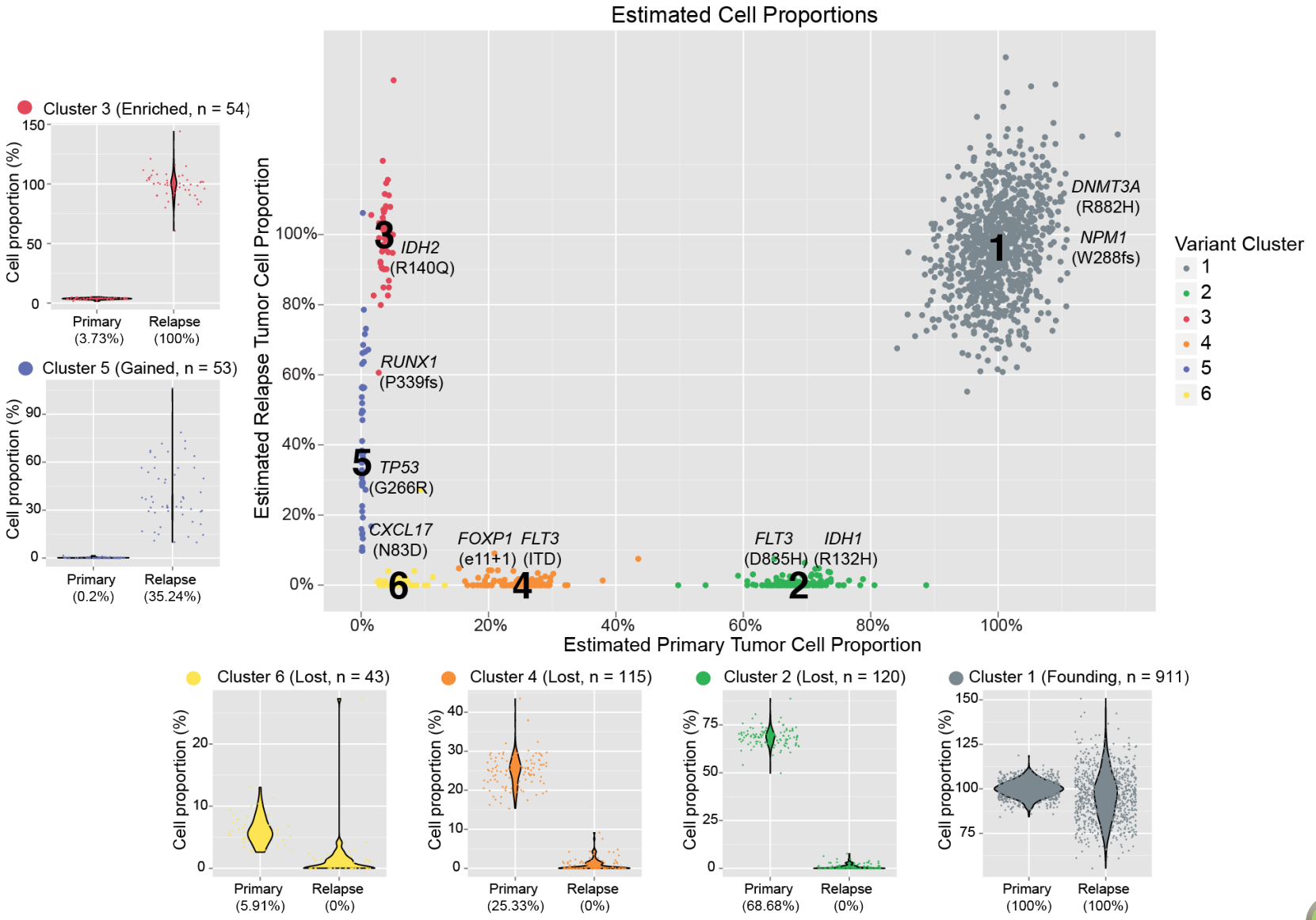
Dominant clone vs. sub-clonal (and driver vs. passenger)



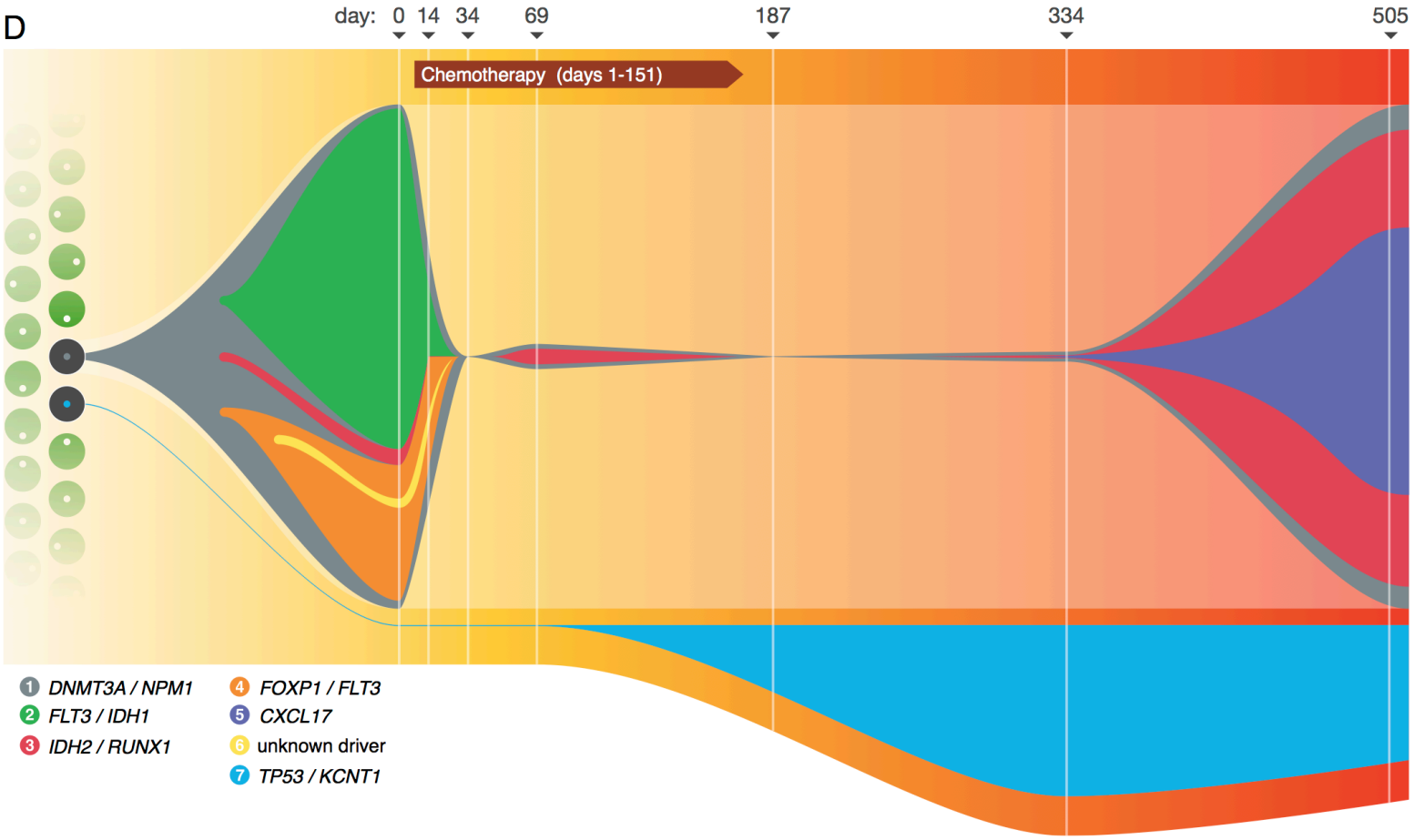


WGS data

Profiling over time can expose complex clonal evolution



Dramatic shifts can be observed



Most variant results are now represented in VCF format. Calling variants is complex but interpreting is harder

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	H_TU-GTB15-3685	H_TU-GTB15-M1501867
1	1026106	.	G	T	.	PASS	NT=ref;QSS=18;QSS_NT=18;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1216591	.	G	A	.	PASS	NT=ref;QSS=120;QSS_NT=108;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1249123	.	G	T	.	PASS	NT=ref;QSS=16;QSS_NT=16;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1262394	.	G	T	.	PASS	NT=ref;QSS=34;QSS_NT=34;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1326886	.	C	T	.	PASS	NT=ref;QSS=199;QSS_NT=157;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1391597	.	T	C	.	PASS	NT=ref;QSS=32;QSS_NT=32;SGT=TT->CT;TQSS=2;TQSS_NT=2	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1904481	.	G	T	.	PASS	NT=ref;QSS=24;QSS_NT=24;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1912142	.	G	T	.	PASS	NT=ref;QSS=33;QSS_NT=33;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1919717	.	G	A	.	PASS	NT=ref;QSS=17;QSS_NT=17;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	2319028	.	C	T	.	PASS	NT=ref;QSS=76;QSS_NT=76;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	2333646	.	G	T	.	PASS	NT=ref;QSS=26;QSS_NT=26;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3328555	.	G	T	.	PASS	NT=ref;QSS=20;QSS_NT=20;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3350384	.	G	A	.	PASS	NT=ref;QSS=33;QSS_NT=33;SGT=GG->AG;TQSS=2;TQSS_NT=2	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3388456	.	C	T	.	PASS	NT=ref;QSS=55;QSS_NT=55;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3662615	.	G	T	.	PASS	NT=ref;QSS=18;QSS_NT=18;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3774072	.	G	T	.	PASS	NT=ref;QSS=21;QSS_NT=21;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6021727	.	G	A	.	PASS	NT=ref;QSS=16;QSS_NT=16;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6271112	.	G	T	.	PASS	NT=ref;QSS=52;QSS_NT=52;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6278217	.	G	T	.	PASS	NT=ref;QSS=30;QSS_NT=30;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6609812	.	G	A	.	PASS	NT=ref;QSS=74;QSS_NT=74;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	9338624	.	G	A	.	PASS	NT=ref;QSS=15;QSS_NT=15;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	10678477	.	G	T	.	PASS	NT=ref;QSS=26;QSS_NT=26;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	10720178	.	G	T	.	PASS	NT=ref;QSS=33;QSS_NT=33;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11140620	.	A	C	.	PASS	NT=ref;QSS=20;QSS_NT=20;SGT=AA->AC;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11194363	.	G	T	.	PASS	NT=ref;QSS=19;QSS_NT=19;SGT=GG->GT;TQSS=2;TQSS_NT=2	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11294450	.	C	T	.	PASS	NT=ref;QSS=35;QSS_NT=35;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11561899	.	G	A	.	PASS	NT=ref;QSS=32;QSS_NT=32;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11595041	.	G	A	.	PASS	NT=ref;QSS=137;QSS_NT=105;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11735264	.	G	T	.	PASS	NT=ref;QSS=170;QSS_NT=122;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11852226	.	G	T	.	PASS	NT=ref;QSS=39;QSS_NT=39;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	11855448	.	G	A	.	PASS	NT=ref;QSS=32;QSS_NT=32;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		
1	12198424	.	G	A	.	PASS	NT=ref;QSS=24;QSS_NT=24;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:		

Details of the VCF file format: hts-specs, VCF-v4.2.pdf