가톨릭대학교
바이오 데이터 엔지니어 양성사업

Principles of targeted sequencing analysis

(Mutation and Copy number alternations)

Terms

Mutation
permanent inheritable change in DNA base sequence

Variants/Mutation/Polymorphism
Variant: broadest term
Polymorphism : common (>1%) variants in populations
Mutation: rare (<1%) variants in populations

Driver mutation
Genetic change that give selective advantage to clones during cancer development

Passenger mutations
Genetic change that do not confer any selective advantage in cancer development

Germline mutations
Mutations in germ cells that are inheritied

Somatic mutations
Mutations in non-germ cells that are acquired as oppsed to inheritied

Terms

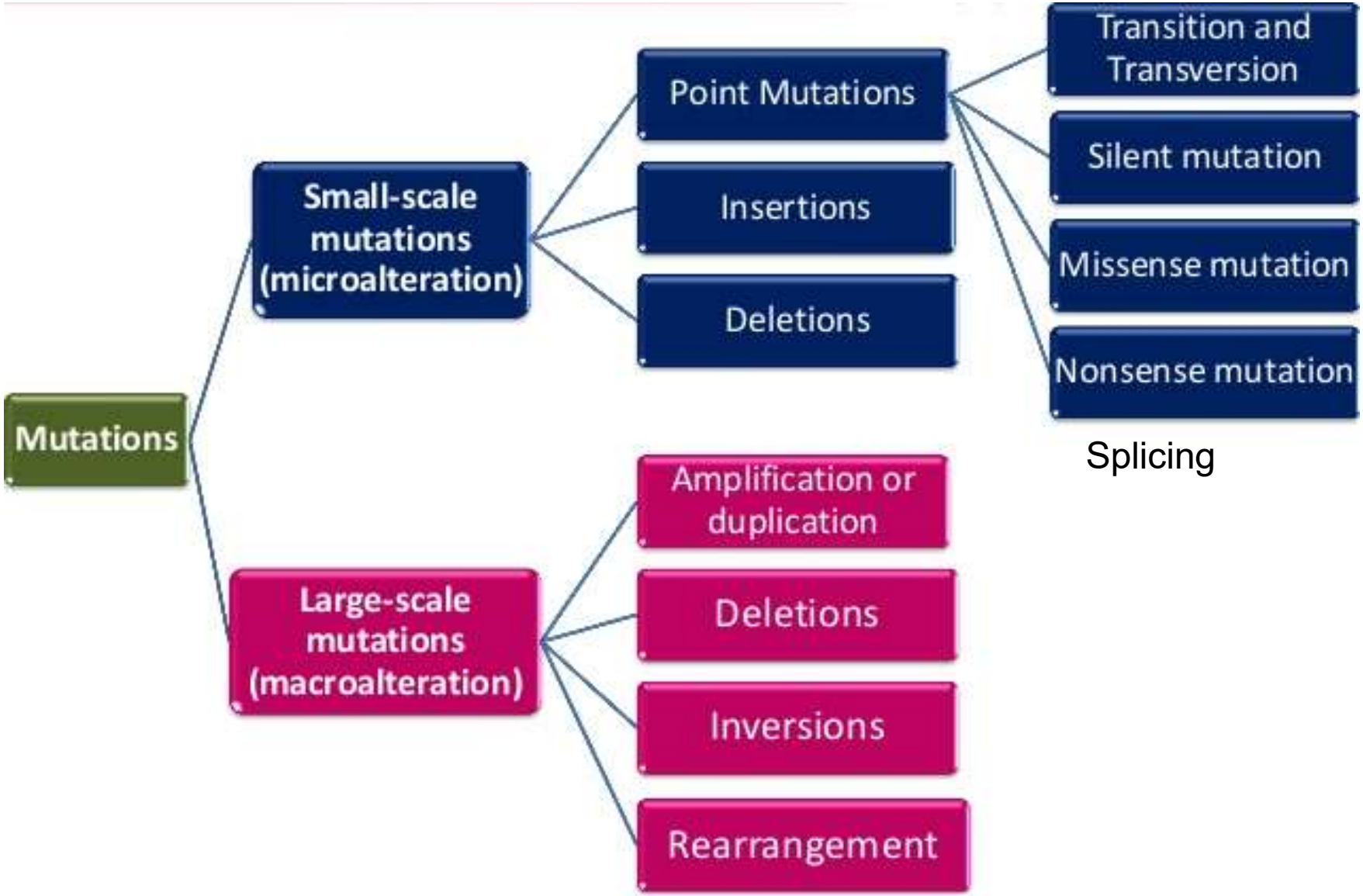
Base substitutions / insertion and deletions / structrual variations
Base substitutions: A type of mutation in which one base is replaced by another
Insertions and deletions (Indels): A type of mutation that arises from the insertion or deletion of one or more nucleotides
Structural variations: Large-scale genomic changes (typically >1kb) such as deletions, tandem duplications, amplifications, inversions and translocations.

Copy number variation/alternations
: Phenomenon in which sections of the genome (>1kb) are repeated and the number of repeats in the genome varies between individuals
: Difference between variants/alternations similar to polymorphism/mutations
→ further classified as amplification (> 3 copies), gain, neutral (2), loss, deletion (0)

Loss of heterozygosity (LOH)
Loss of heterozygosity is defined as the loss of one parent's contribution to the cell and can be caused by deletion, or others (not equal to loss)

Allelic imbalance (AI)
Phenomenon where the two alleles of a given gene are differentially expressed

Classification of mutations



Principle of mutation

Rare event == Likely to be more pathogenic
Polymorphism: different between populations (ExAC, gnomAD east asian)

(Procto)oncogene: with hotspot (ex BRAF V600E residue)
Tumor suppressor gene: without hotspot (ex TP53)
Two-hit hypothesis (The Knudson hypothesis): two hits required for cancer in TSG

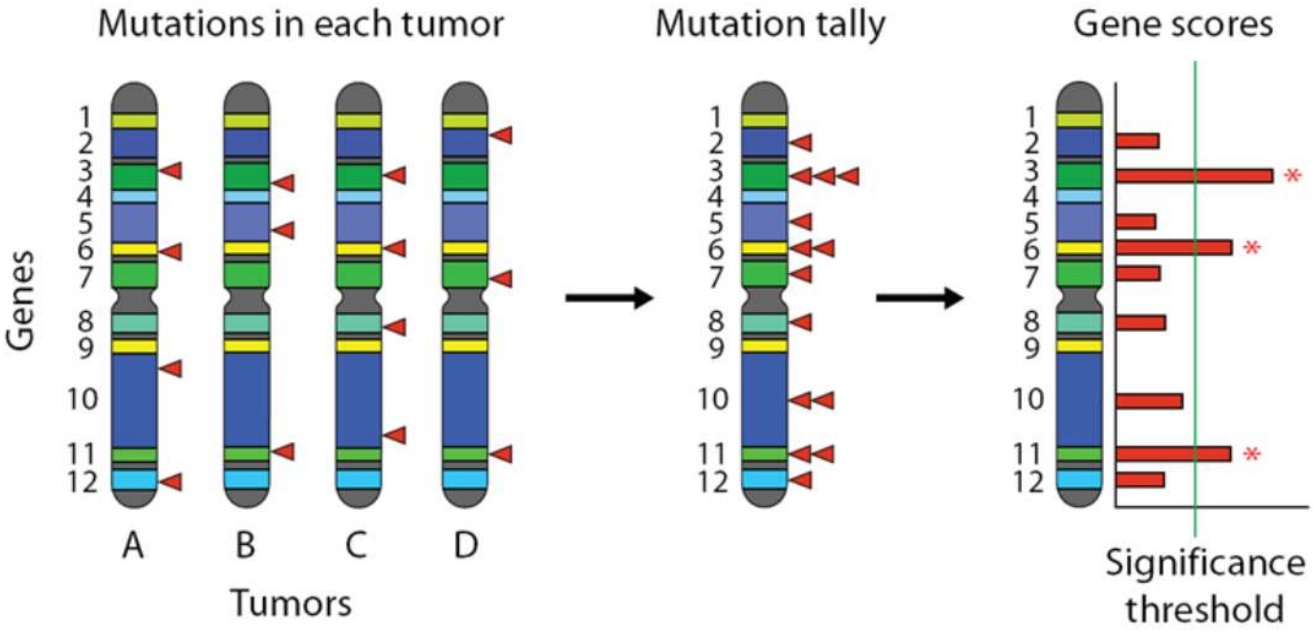
Possible normal cell contamination in tumor (more likely)
Possible tumor cell contamination in normal (less likely)
Normal sample is in needed for genotype, which could be revealed in low depth

VAFs(Variants allele fraction): read counts supporting variants / total read counts
VAFs can be affected by tumor purity, clonal evolution, CNV status
CCF(cancer cell fraction) inferred by VAFs after adjusting purity and CNV
Clonal (all/most of tumor cells have mutation), Subclonal (Not all)

FFPE (False C>T at low VAFs), cfDNA (Very low VAFs)

Principle of mutation

Recurrent gene mutations in study or in DB== Likely driver mutation
But by not only simpe mutation counts (gene length, methylation, ...)

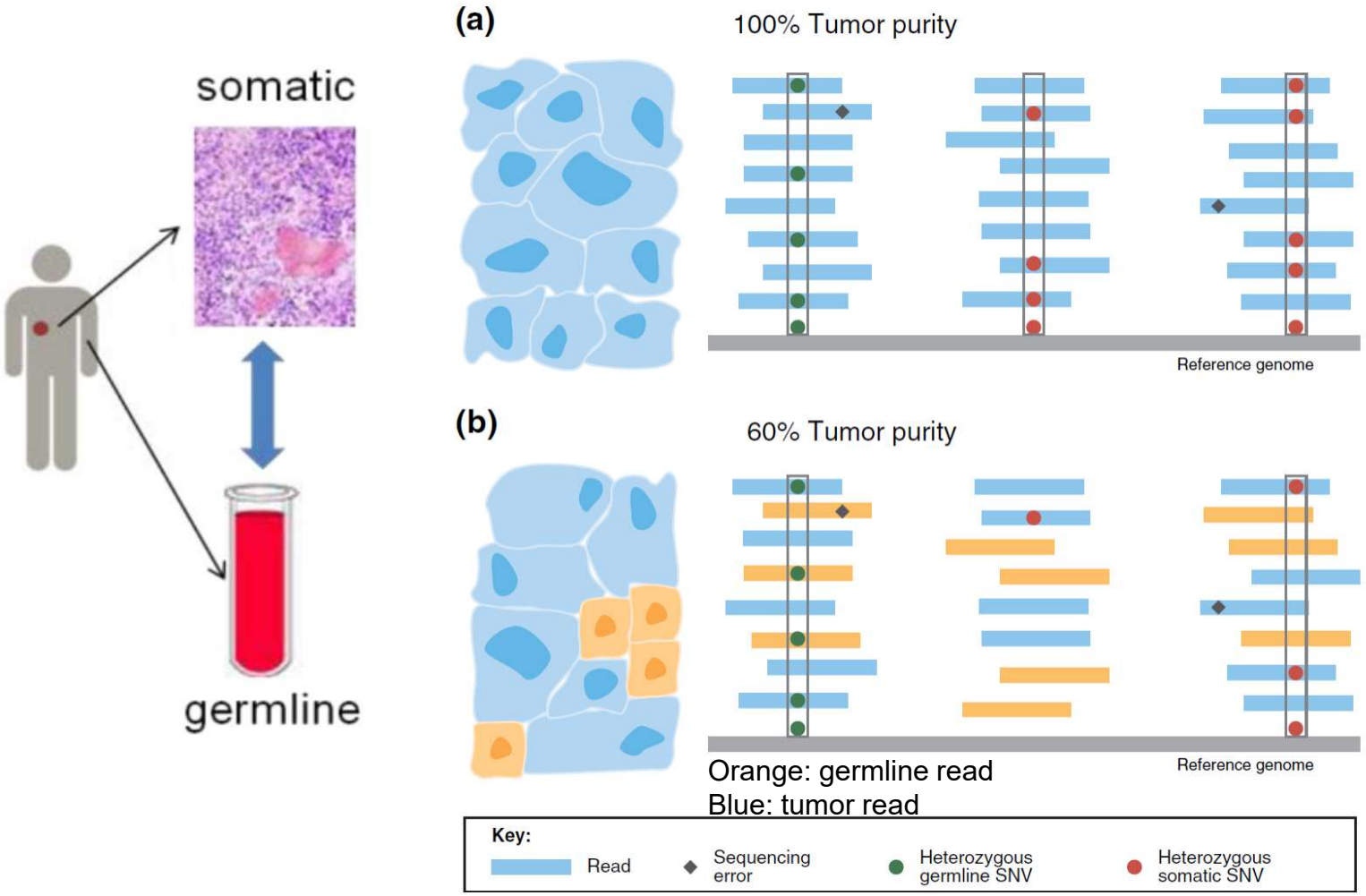


MutSig builds a model of the background mutation processes (BMR) that were at work during formation of the tumors, and it analyzes the mutations of each gene to identify genes that were mutated more often than expected by chance, given the background model.

MutSigCV (CV for 'covariate') improves the BMR estimation by pooling data from 'neighbor' genes with similar genomic properties such as DNA replication time, chromatin state (open/closed), and general level of transcription activity.

Lawrence, M. et al. Nature 2013

Principle of mutation



Tumor purity: fraction of cancer cells within sample
Read depth: read count at specific locus
VAFs: read counts supporting variants / total read counts
If tumor purity 100% and no CNAs, VAF will be 50%

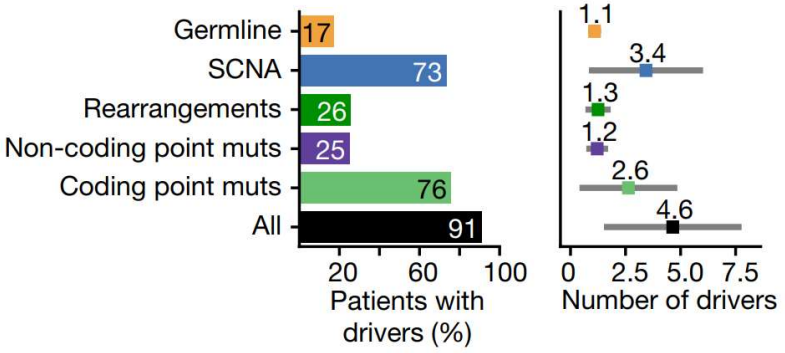
Clinical implications

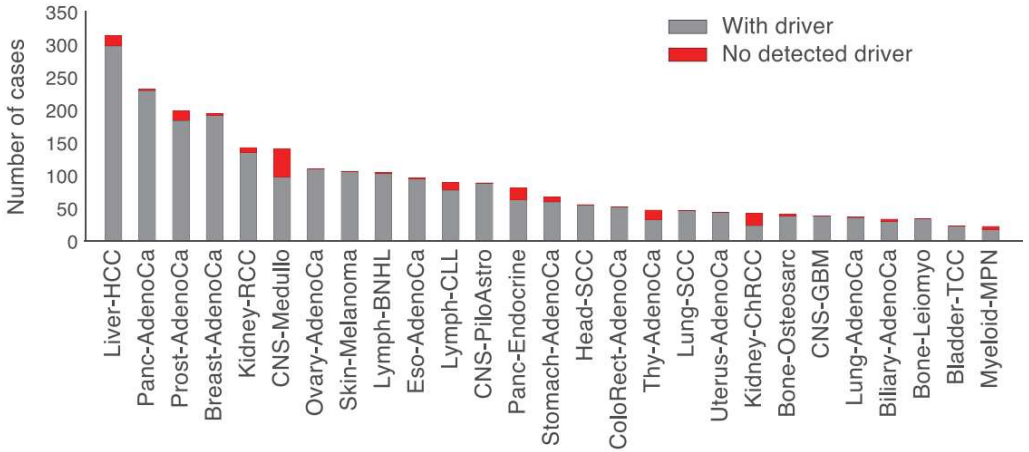
Table 1 | Genomic assays and the clinical decisions they support

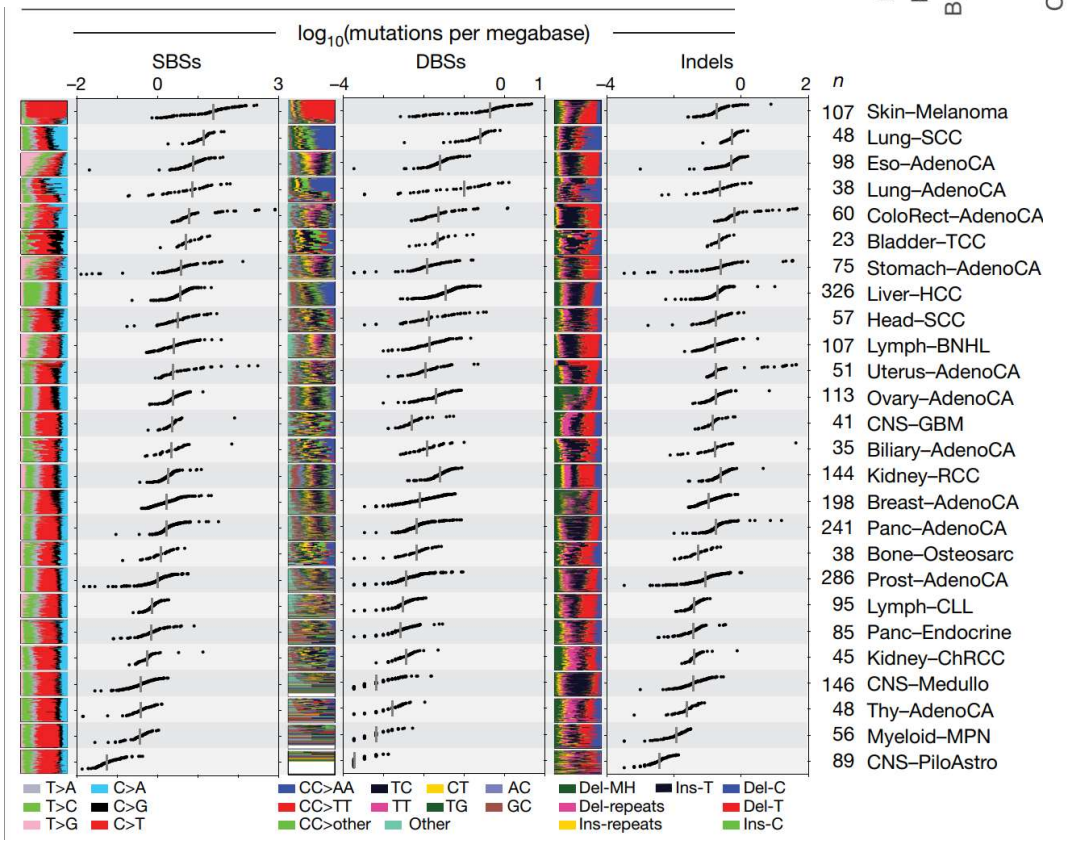
Omic feature	Time of aberration discovery	Clinical setting	Analyte	Analysis	Treatment implication	Time of implementation of first targeted therapy
BCR–ABL translocation	1973[41]	Chronic myeloid leukaemia, acute lymphoblastic leukaemia	DNA, RNA	PCR (translocation)	Imatinib, dasatinib, nilotinib	2001[42]
BCR–ABL imatinib resistance mutation	2002[119]	Chronic myeloid leukaemia	DNA	Sequencing (point mutation)	Dasatinib, nilotinib	2006[120]
PML–RAR translocation	1990[121]	Acute promyelocytic leukaemia	DNA	PCR (translocation)	All *trans* retinoic acid	1987[122]
KIT mutation	2001[123]	Gastrointestinal stromal tumour	DNA	Sequencing	Imatinib	2001[123]
HER2 amplification	1985[124]	Breast cancer, gastric cancer	Tissue	Fluorescence *in situ* hybridization, immunohistochemistry	Trastuzumab, lapatinib, pertuzumab	2001[125]
Oestrogen or progesterone receptors	1896[126]	Breast cancer	Tissue	Immunohistochemistry	Multiple hormonal-based therapies (for example, aromatase inhibitors)	1896 (oophorectomy)[126]
EGFR mutation	2004[76]	Non-small-cell lung cancer	DNA	Sequencing or PCR (point mutations)	Erlotinib, gefitinib	2003[127]
ALK fusion gene	2007[44]	Non-small-cell lung cancer	DNA	Fluorescence *in situ* hybridization (break apart)	Crizotinib	2010[43]
BRAF mutation	2001[128]	Melanoma	DNA	Sequencing or PCR (point mutations)	Vemurafenib	2011[129]
KRAS mutation	1987[130]	Colorectal cancer	DNA	Sequencing or PCR (point mutations)	Withhold cetuximab or panitumumab	2006[131]

Mutations / CNAs / Fusions : all events might be clinically applicable

Cancer genetics (PCAWG paper)



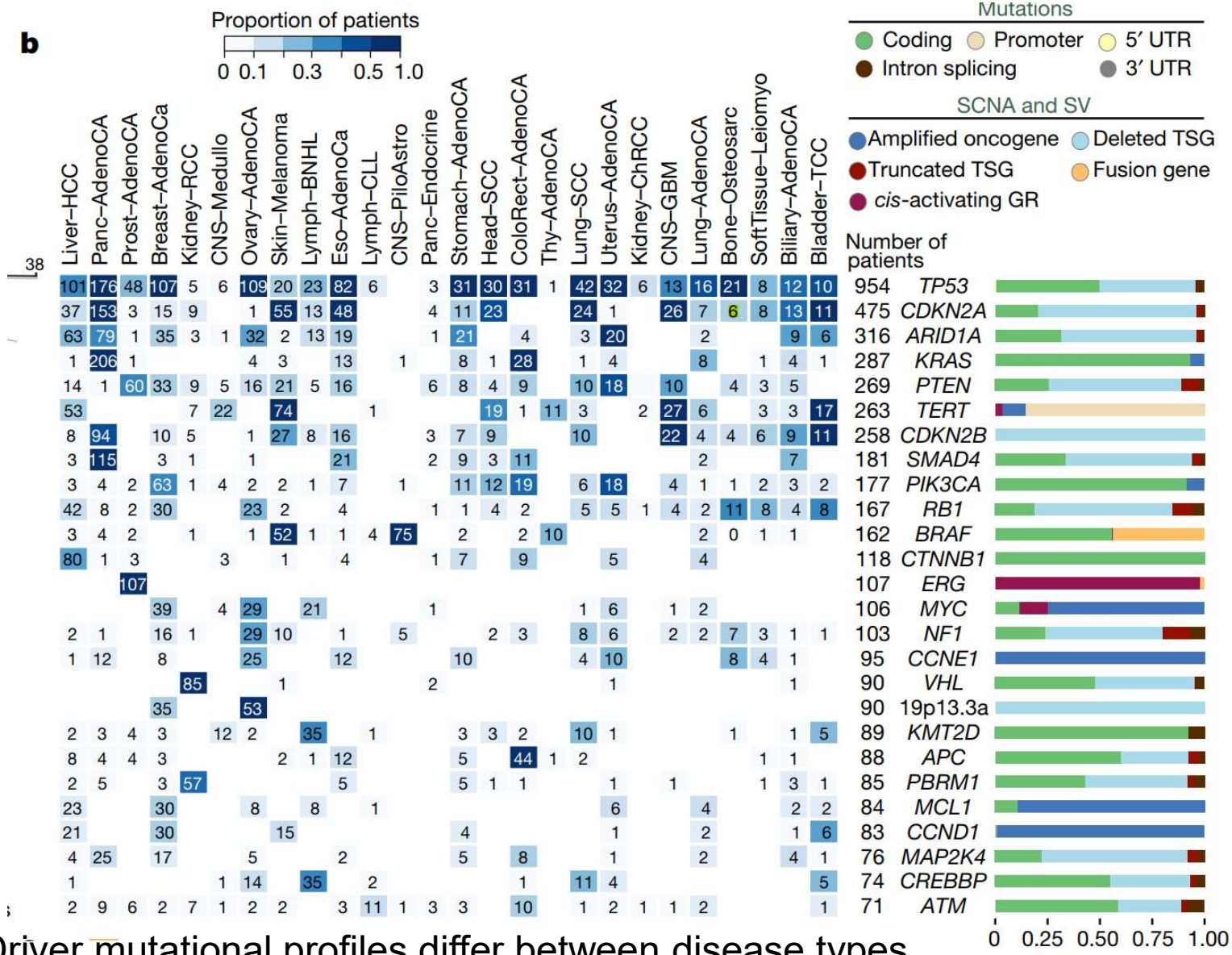




Most of cases, but not all cases have mutations in DNA

Mutational landscape differ by disease types as well as tumors of the same type

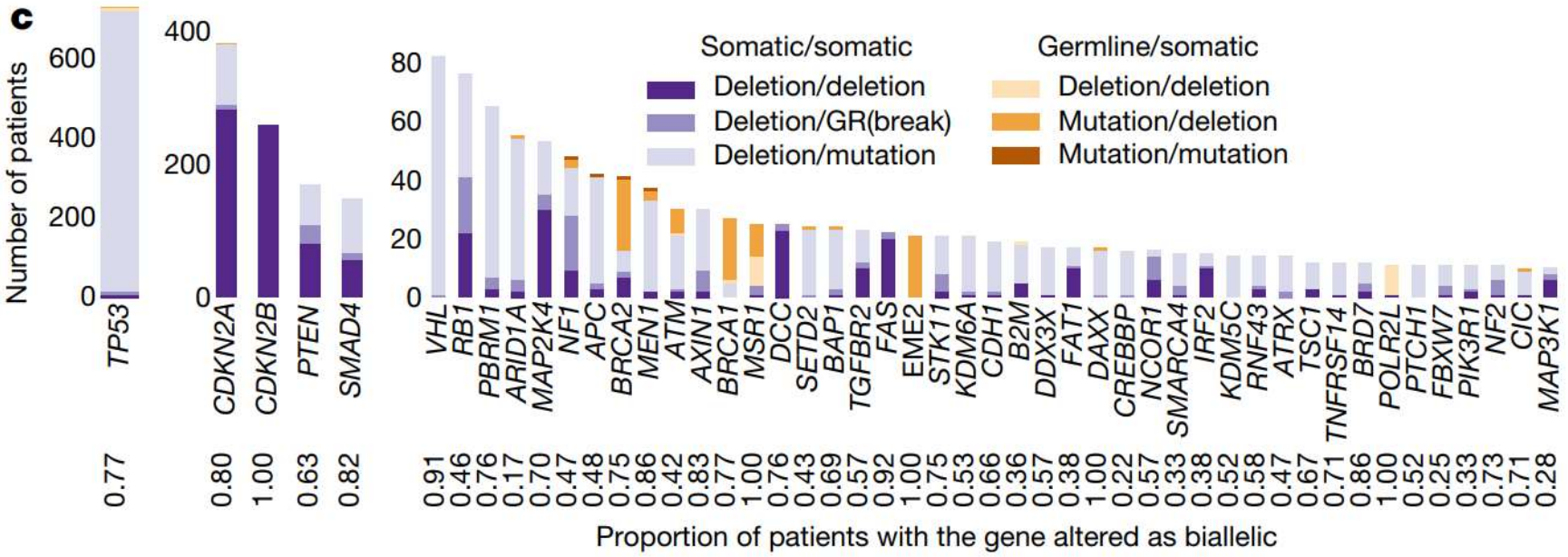
Cancer genetics (PCAWG paper)



Driver mutational profiles differ between disease types

Most events occurred in coding region (not in TERT)

Cancer genetics (PCAWG paper)



Two-hit hypothesis (The Knudson hypothesis): two hits required for cancer in TSG
TSG genes altered as billallelic in most cases

Article

Pan-cancer analysis of whole genomes

https://doi.org/10.1038/s41586-020-1969-6 The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

Further reading (Nature, 2020)
수천cases, pan-cancer, WGS, ICGC/PCAWG

Sequencing of DNA

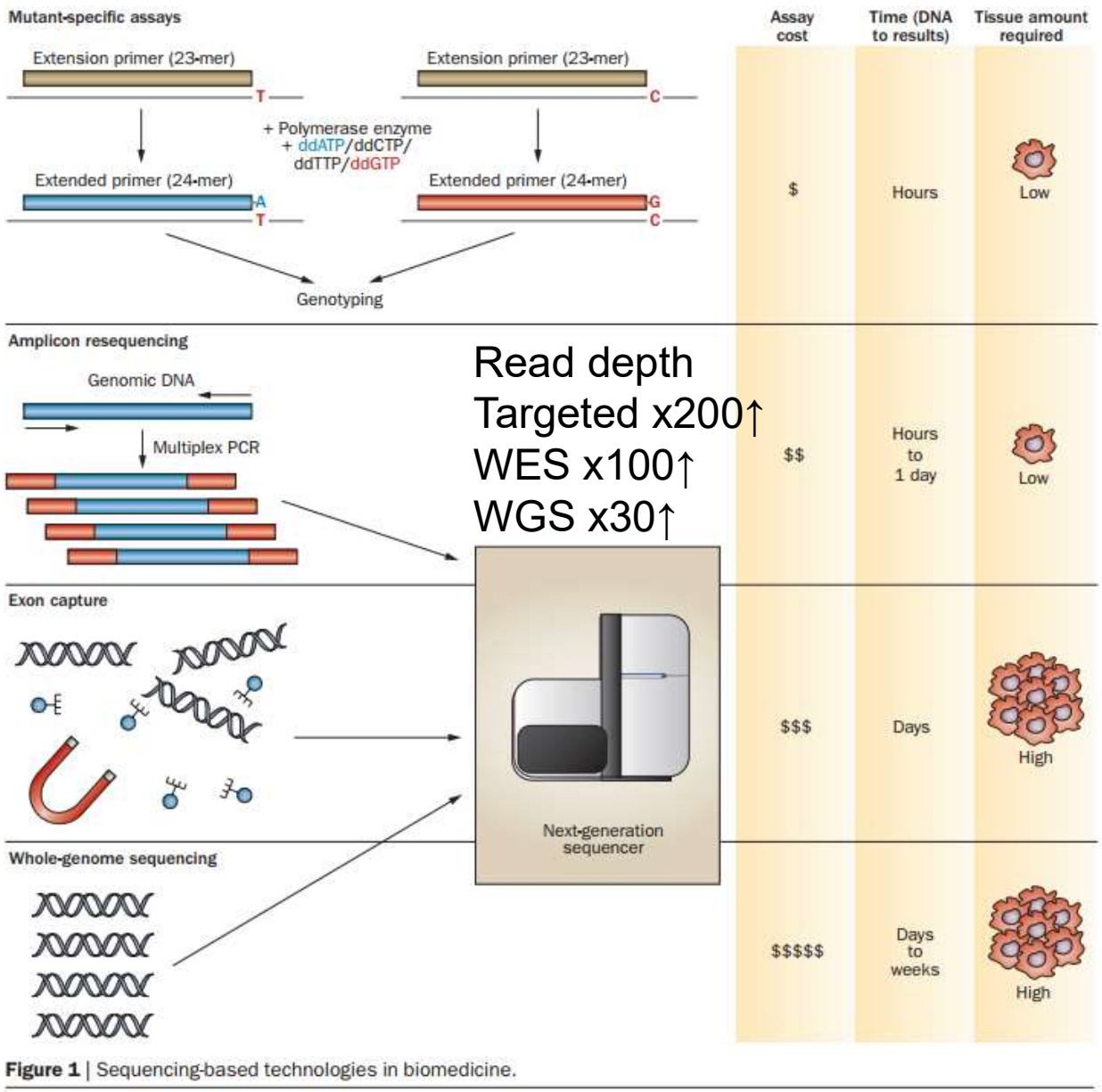


Figure 1 | Sequencing-based technologies in biomedicine.

Sanger sequencing
(capillary sequencing)
VAF>10%

Read depth
Targeted x200↑
WES x100↑
WGS x30↑

Ampli-seq
(Thermo Fischer)

Hyb-seq
(Agilent, Celemics)

Illumina

Targeted/WES/WGS

	Targeted	WES	WGS
Application	Mutation, CNAs (MSI, TMB)	Mutation, CNAs, Signature, MSI, TMB Purity, Telomere	WES applications + SVs (chromothripsis, translocation, fusion)
Mutation count	수~수십	수십~수천	수천~수십만
Target region	Various Less than 3Mb	3-6Mb	1Gb~3Gb
Sequencing	Various Less than 3Gb	About 10Gb	About 30Gb
Indication	cfDNA, small FFPE, Validation	Research standard Frozen/FFPE	Future standard Frozen Difficult to analysis and interpret
Kit	Ion AmpliSeq Celemics Kit SureSelectXT Target	Sureselect Exon V6	TruSeq DNA Nano
Cost	30만전후	80만전후	250만전후

Panel sequencing

Panel?
Which gene, mutation/CNAs, Hotspots/coding sequence

Pros
1. Focused on Clinically applicable information

2. Accuracy due to high read depth
3. High sensitivity detecting low VAF mutations

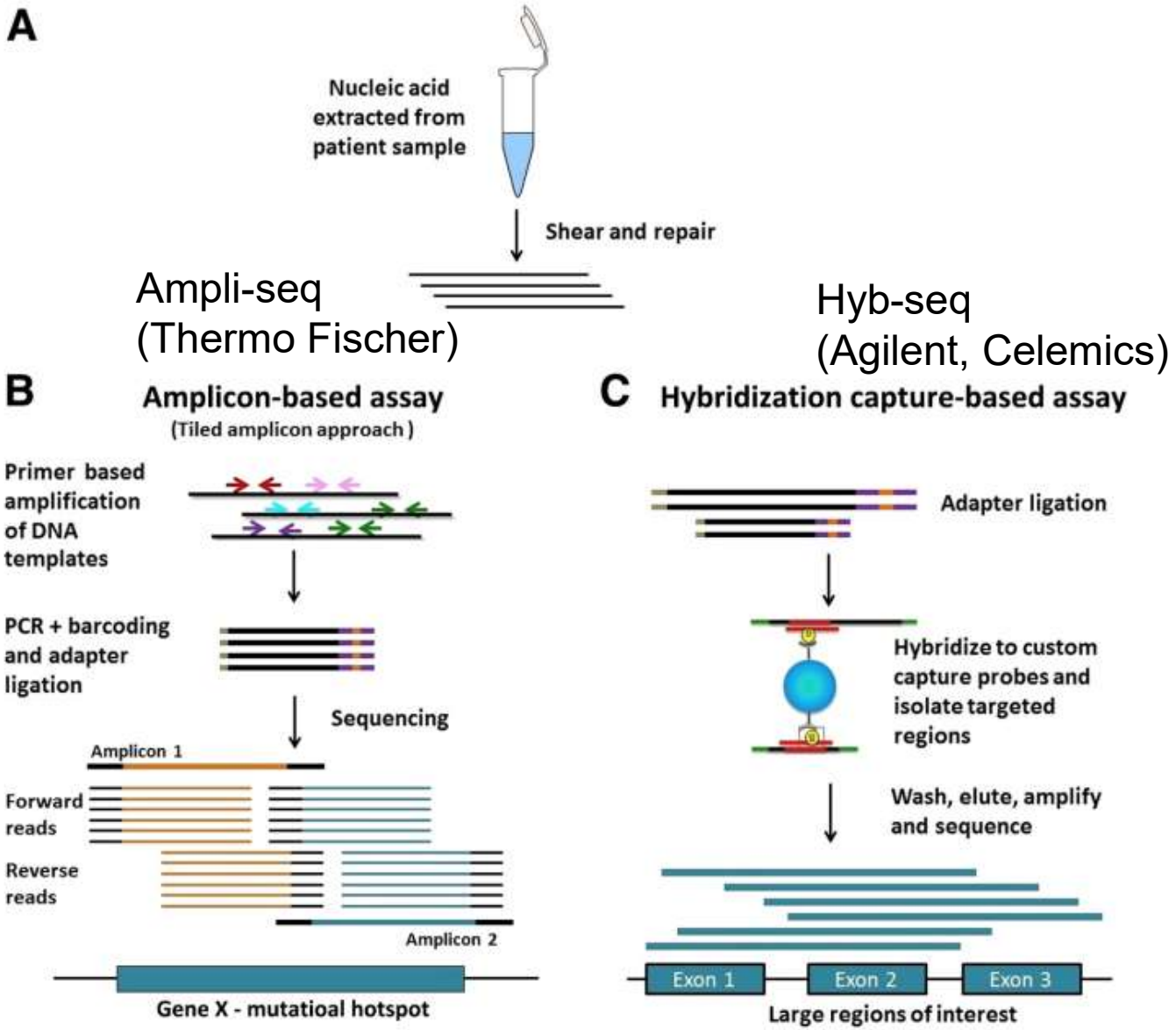
4. Small amount DNA available (10ng)
5. FFPE/Liquid biopsy available

6. Low cost
7. Faster Sequencing and Anlaysis

Panel sequencing only available (small sample, small FFPE, cfDNA,...)

Cons
mutation/CNAs only, outside panel unknown

Panel sequencing



J Mol Diagn. 2017 May;19(3):341-365

Panel sequencing

Sample source	PCR based = Hybrid capture based (Tissue, FFPE, liquid biopsy)	
Input DNA	**PCR based** > Hybrid capture based	
Panel size	PCR based < **Hybrid capture based**	
Data generation time	**PCR based** > Hybrid capture based	
Homopolymer analysis	PCR based < **Hybrid capture based**	
Application	PCR based < **Hybrid capture based** (Mutation, CNA) (Mutation, CNA, **MSI test**)	
	More prone to amplification error (?)	High off-target (Large CNAs) (but more cose)

Low sample input amount = More amplification, high duplication rate = low read depth

95 genes

Mutation	Mutation + CNA	CNA
NRAS, DDR2, IDH1, ROS1, SMO, GNAQ, RET, HRAS, MAP2K1, IDH2, GNA11, MAP2K2, JAK3, MPL, DNMT3A, XPO1, NFE2L2, SF3B1, VHL, MYD88, RHOA, CSF1R, NPM1, EZH2, ABL1, NOTCH1, PLEKHS1, WDR74, SDHD, PTPN11, POLE, FLT3, TSHR, B2M, SPOP, SRC, GNAS, U2AF1, MAPK1, ARAF, MED12,	MTOR, JAK1, ALK, ERBB4, RAF1, CTNNB1, PIK3CA, FGFR3, PDGFRA, KIT, FGFR4, ESR1, EGFR, MET, BRAF, FGFR1, JAK2, FGFR2, KRAS, ERBB3, CDK4, AKT1, ERBB2, AR, MDM4, MLH1, KDR, TERT, RAC1, AURKA	MYCN, CDK6, MCL1, MDM2, NKX2-1, CCNE1, CDC274, PDCD1LG2, RPPH1, HS6ST3

CDS + CNA

ARID1A, FBXW7, APC, CCND1, MYC, CDKN2A, PTEN, BRCA2, RB1, CDH1, TP53, BRCA1, SMAD4, STK11

* 비유전성 고형암 필수 유전자

Mutation annotation and prioritazation

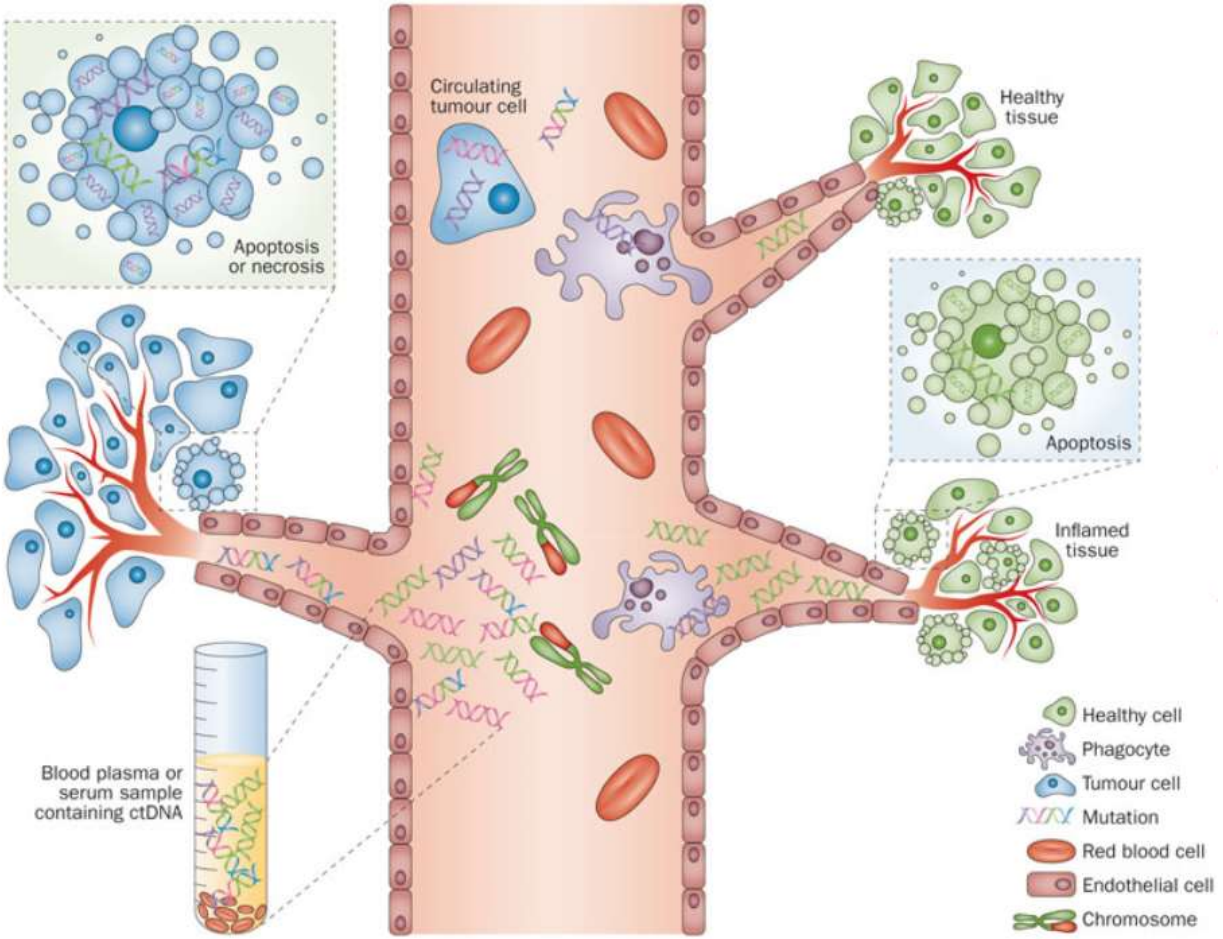
Step1: Classify mutations into germline and somatic mutations

Step2: Annotate **somatic mutations** (add information)
using webannovar http://wannovar.wglab.org/

Step3: Filter less likely somatic mutations (Filter subject to research objective)
Population minor allele frequency > 0.1% (ExAC, gnomAD)
Read counts < 20 (Normal or tumor)
VAFs < 2-5%

Step4: Prioritize somatic mutations (to find putative driver mutations)
COSMIC, cbioportal, oncokb, Polyphen-2, ClinVar,…

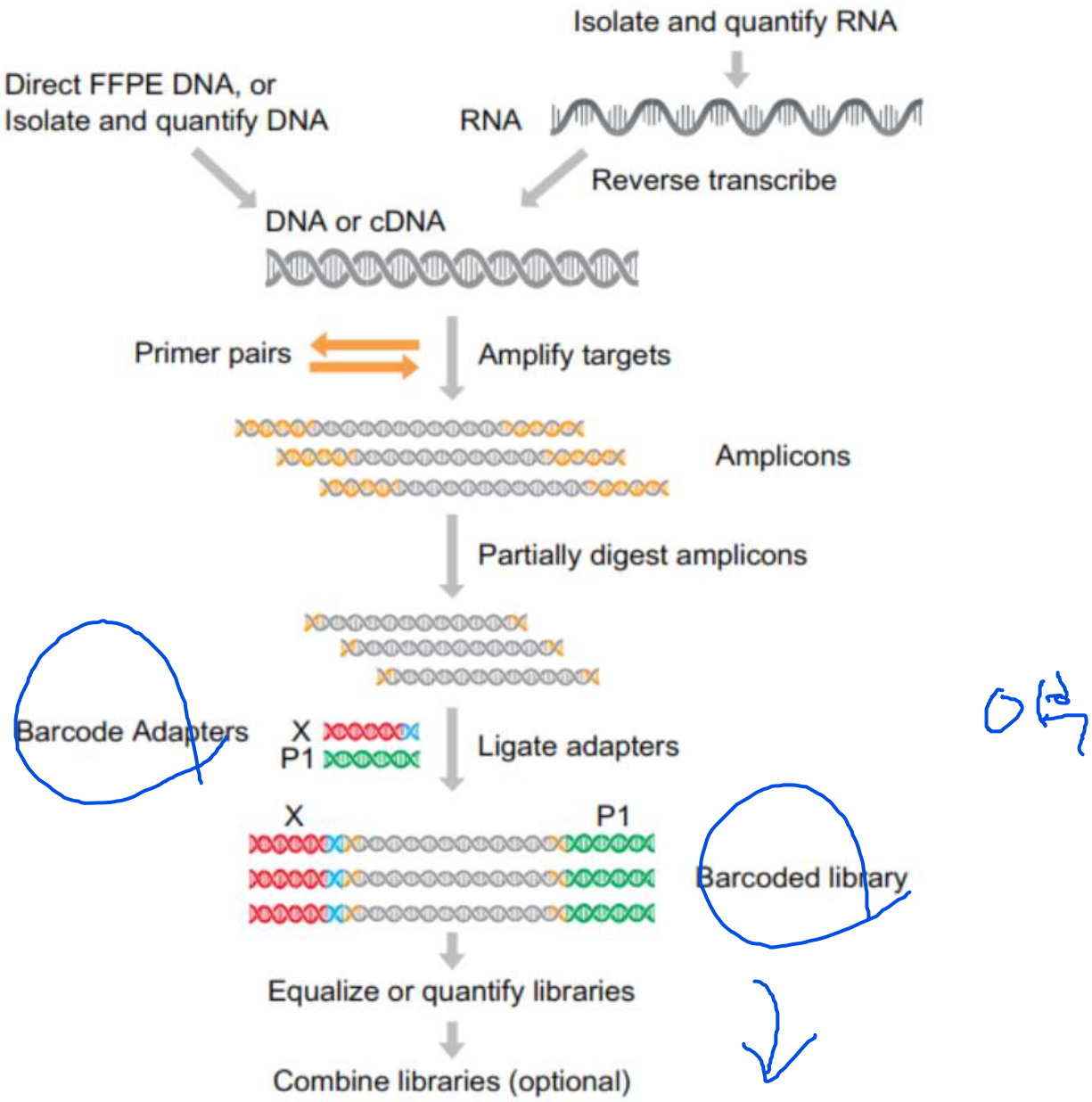
cfDNA/ctDNA



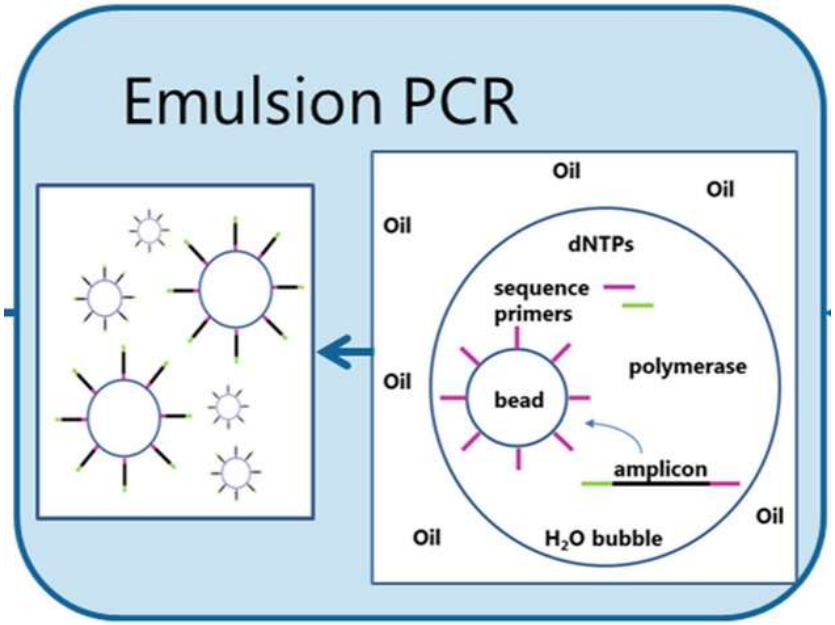
cfDNA
=cell-free DNA
: Degraded DNA fragments released to the blood plasma
: Including ctDNA

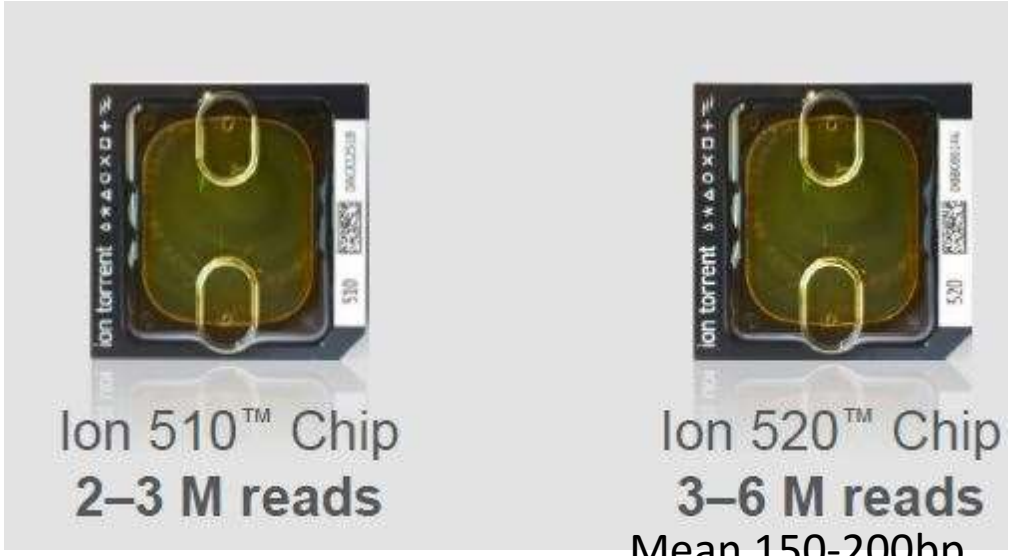
ctDNA
=circulating tumor DNA
:Tumor-derived fragmented DNA in the bloodstream

Ampliseq (1): Library preperation



Ampliseq (2): Templating and sequencing

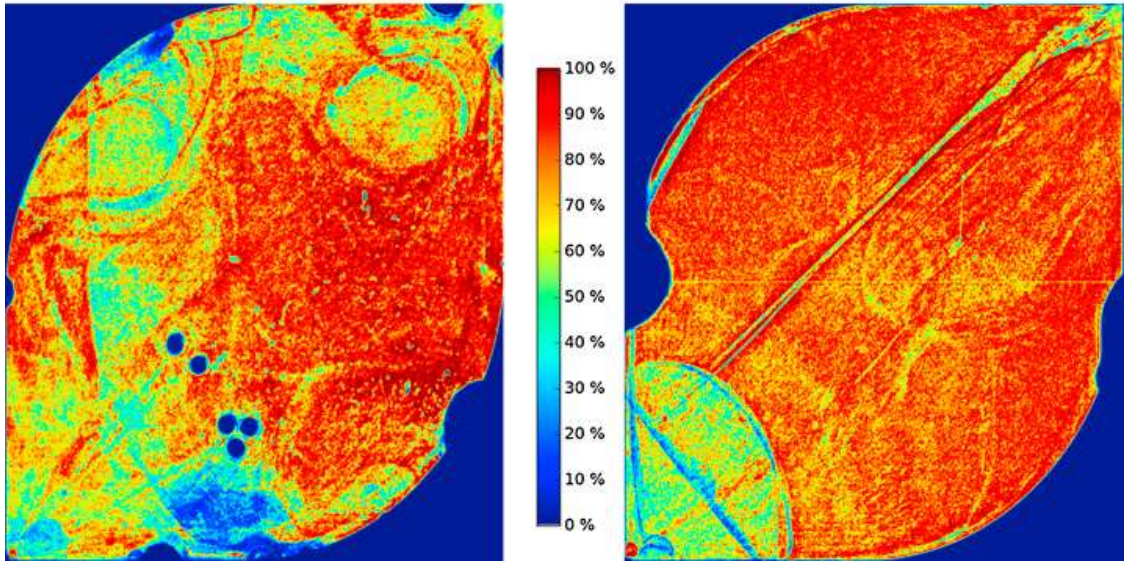




Mean 150-200bp

Total: ~1Gb

Manual loading

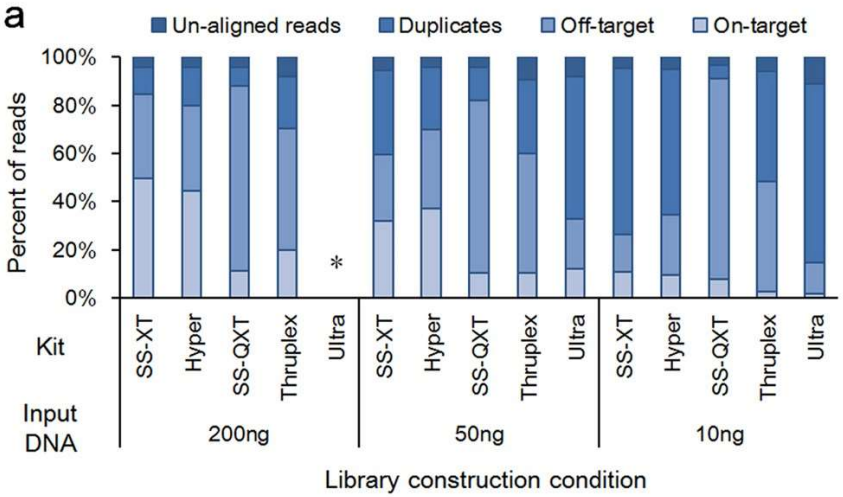


Loading with Ion Chef™

Key point for successful DNA sequencing study

DNA amount, Quality (DIN), Purity, Sample condition (FFPE, Thawing)

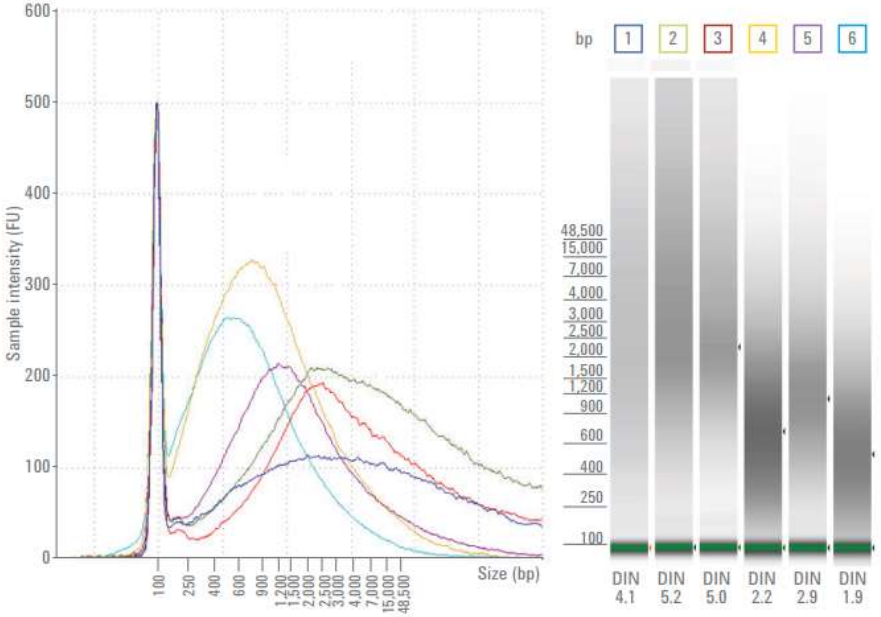
Park WY, Sci Rep

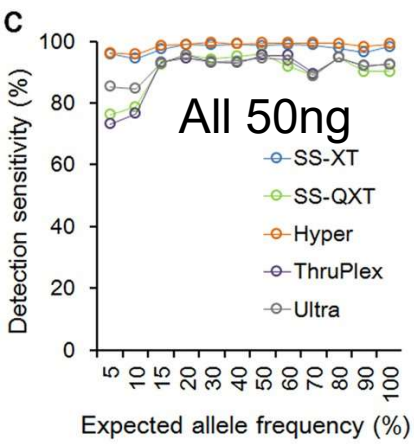


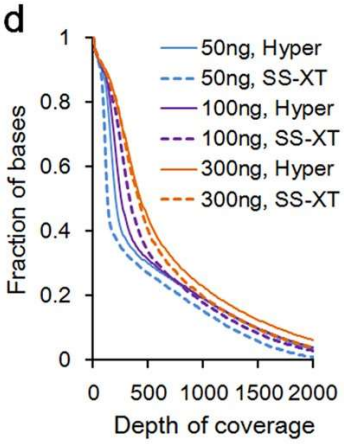
Agilent Tapestation DIN for FFPE

Table 1. The DIN and the sequencing quality criteria obtained for the six samples shown in Figure 3.

Sample	DIN	On-target rate % > 70 %	10x Coverage rate % > 90 %	Deduplication rate %
1	4.1	79.7	98.6	66.0
2	5.2	78.0	99.0	74.2
3	5.0	80.7	98.1	87.1
4	2.2	23.8	94.0	65.0
5	2.9	51.5	97.7	95.7
6	1.9	47.4	96.6	97.1







Same amount, different depth

False negative (esp. in low VAF) (DIN,conta) or false positive (Amp error)