Center for Biomedical Informatics and Information Technology (CBIIT)
NCI/Shady Grove Campus

- We are not endorsing, promoting any specific tool during this training.

- We may recommend open source tools based on the literature and our experience.

Outline of Exome Sequencing Training

Overview

- Quick introduction of CGBG Team..
- Introduction to exome sequencing analysis workflow.
- Challenges in Whole Exome Sequencing.
- Exome sequencing pipelines.

Hands On Tutorial

- Get a Galaxy account
- Upload dataset and pre-alignment QC
- Alignment with BWA and post-alignment Summary
- Build pileup for variant call and variant détection and variant annotation
- Visualize alignments and variants on IGV

NATURE PROTOCOLS | PERSPECTIVE

DNA sequencing technologies: 2006–2016

Elaine R Mardis

Nature Protocols **12**, 213–218 (2017) | doi:10.1038/nprot.2016.182

Received 17 August 2016 | Accepted 20 October 2016 | Published online 05 January 2017

Abstract

Abstract • Introduction • References • Acknowledgments • Author information

Recent advances in the field of genomics have largely been due to the ability to sequence DNA at increasing throughput and decreasing cost. DNA sequencing was first introduced in 1977, and next-generation sequencing technologies have been available only during the past decade, but the diverse experiments and corresponding analyses facilitated by these techniques have transformed biological and biomedical research. Here, I review developments in DNA sequencing technologies over the past 10 years and look to the future for further applications.

Seminars in Oncology

February 2016 Volume 43, Issue 1, Pages 36–48

The promise of omics-based approaches to cancer prevention

Daoud Meerzaman, Barbara K. Dunn, Maxwell Lee, Qingrong Chen, Chunhua Yan, Sharon Ross

DOI: http://dx.doi.org/10.1053/j.seminoncol.2015.09.004

Abstract

Cancer is a complex category of diseases caused in large part by genetic or genomic, transcriptomic, and epigenetic or epigenomic alterations in affected cells and the surrounding microenvironment. Carcinogenesis reflects the clonal expansion of cells that progressively acquire these genetic and epigenetic alterations—changes that, in turn, lead to modifications at the RNA level. Gradually advancing technology and most recently, the advent of next-generation sequencing (NGS), combined with bioinformatics analytic tools, have revolutionized our ability to interrogate cancer cells. The ultimate goal is to apply these high-throughput technologies to the various aspects of clinical cancer care: cancer-risk assessment, diagnosis, as well as target identification for treatment and prevention. In this article, we emphasize how the knowledge gained through large-scale omics-oriented approaches, with a focus on variations at the level of nucleic acids, can inform the

Computational Genomics & Bioinformatics Branch (GCBB)

Computational Genomics & Bioinformatics Branch

Provide state-of-art data analysis on a variety of multifaceted cancer-related research projects including proteogenomic, AI, and clinical high-throughput data.

Computational Genomics & AI

- NGS and Integrated analysis of proteogenomic data for Intramural NCI and moonshot projects (APOLLO, CIMAC-CIDC and SeQC2).
- Multimodal AI and ML approaches for patient outcome prediction. Deep learning algorithms, CNNs.

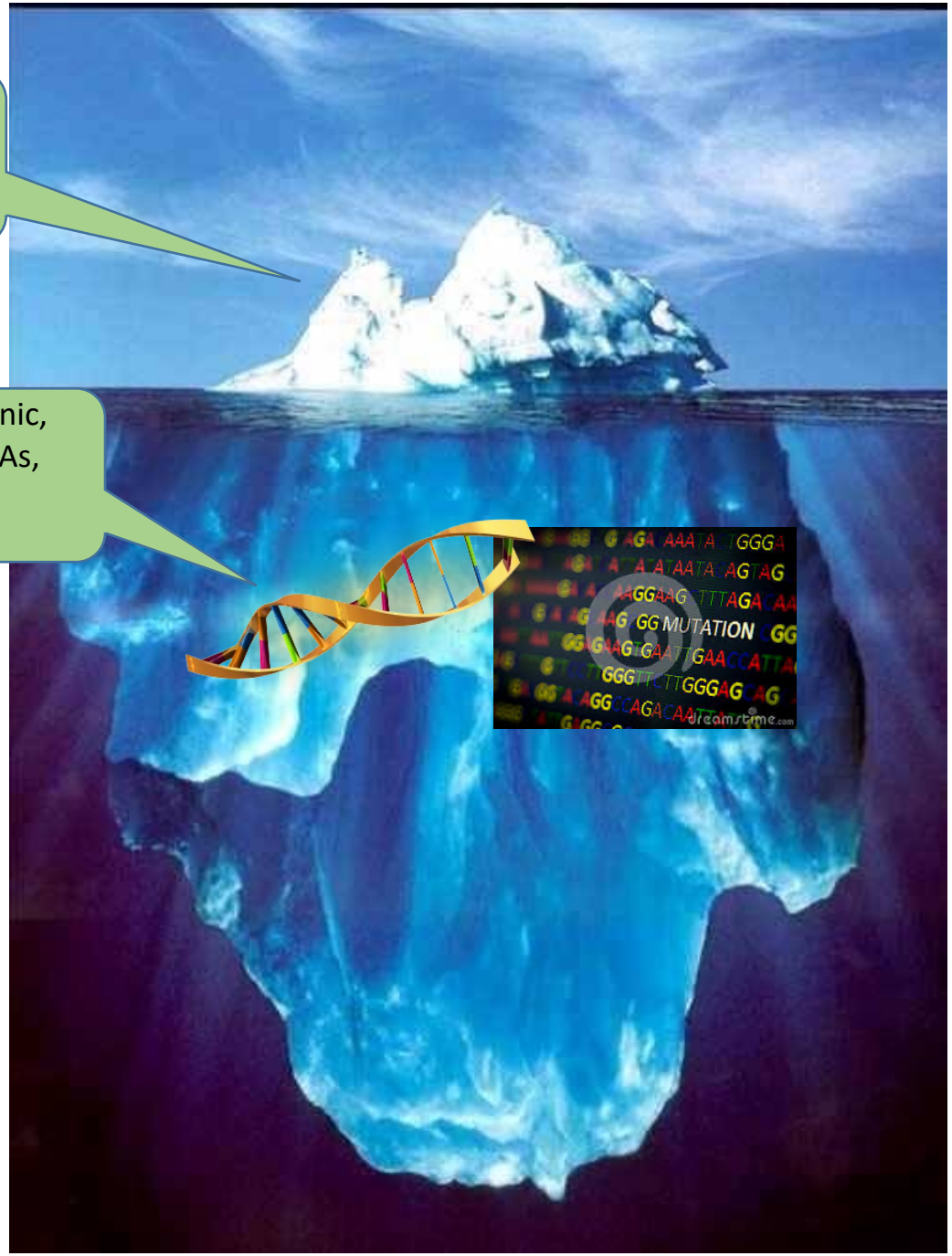
Scientific & Tools Consulting and training

- Generate and implement a web-based workflow in CRDC (CGC) and OmicCircos and (MOGSA)
- Widely used tools for NGS analyses and Visualization tools
- ITCR Scientific software
- CBIIT Summer Internship Program

NIH NATIONAL CANCER INSTITUTE

Exome Sequencing (protein coding)

Introns, intergenic, non-coding RNAs, 5’-3TUR and Promoters



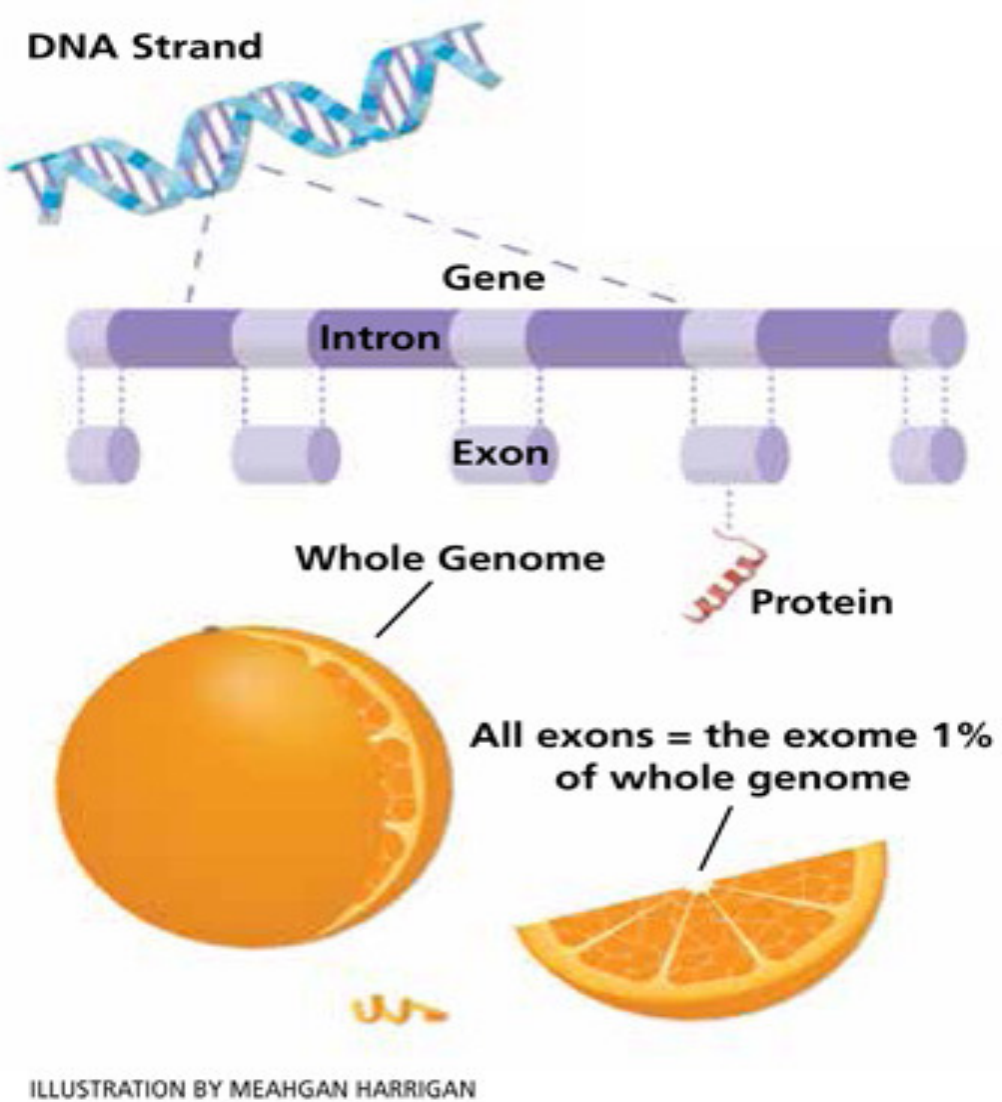
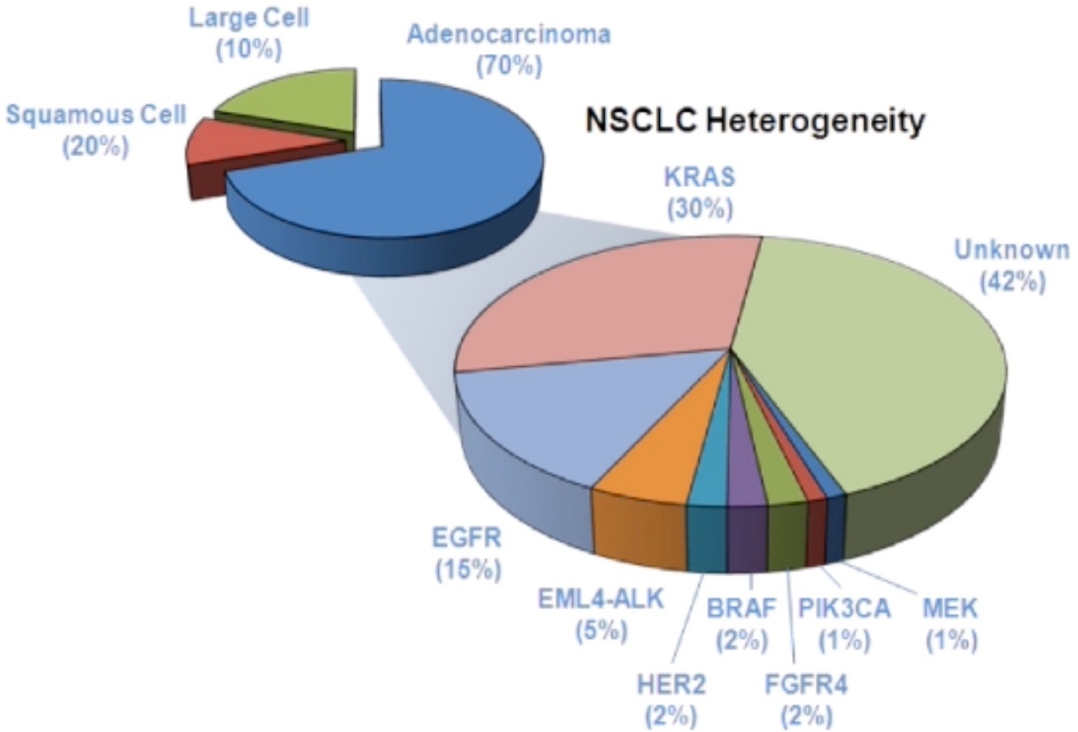


ILLUSTRATION BY MEAHGAN HARRIGAN

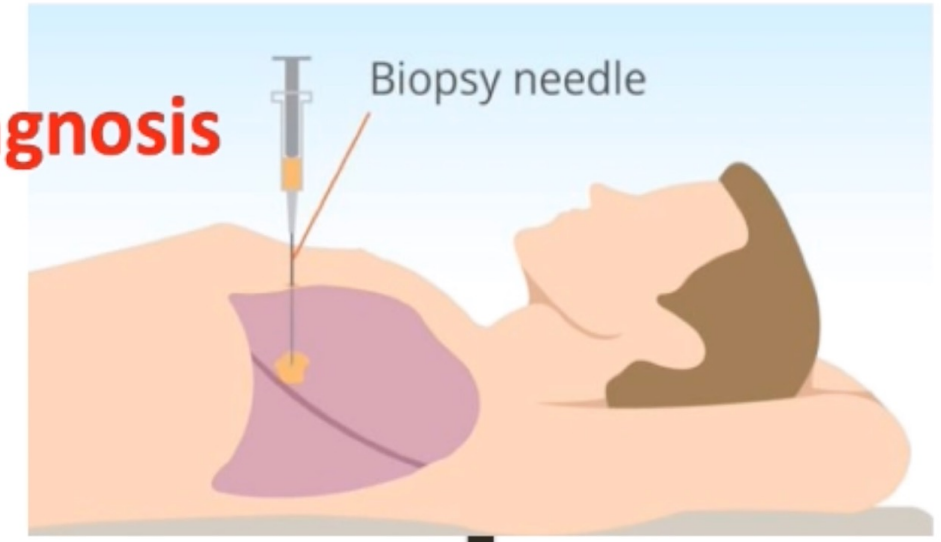
http://www.dana-farber.org

Adult cancers: genomics provides a treatment playbook

discovery



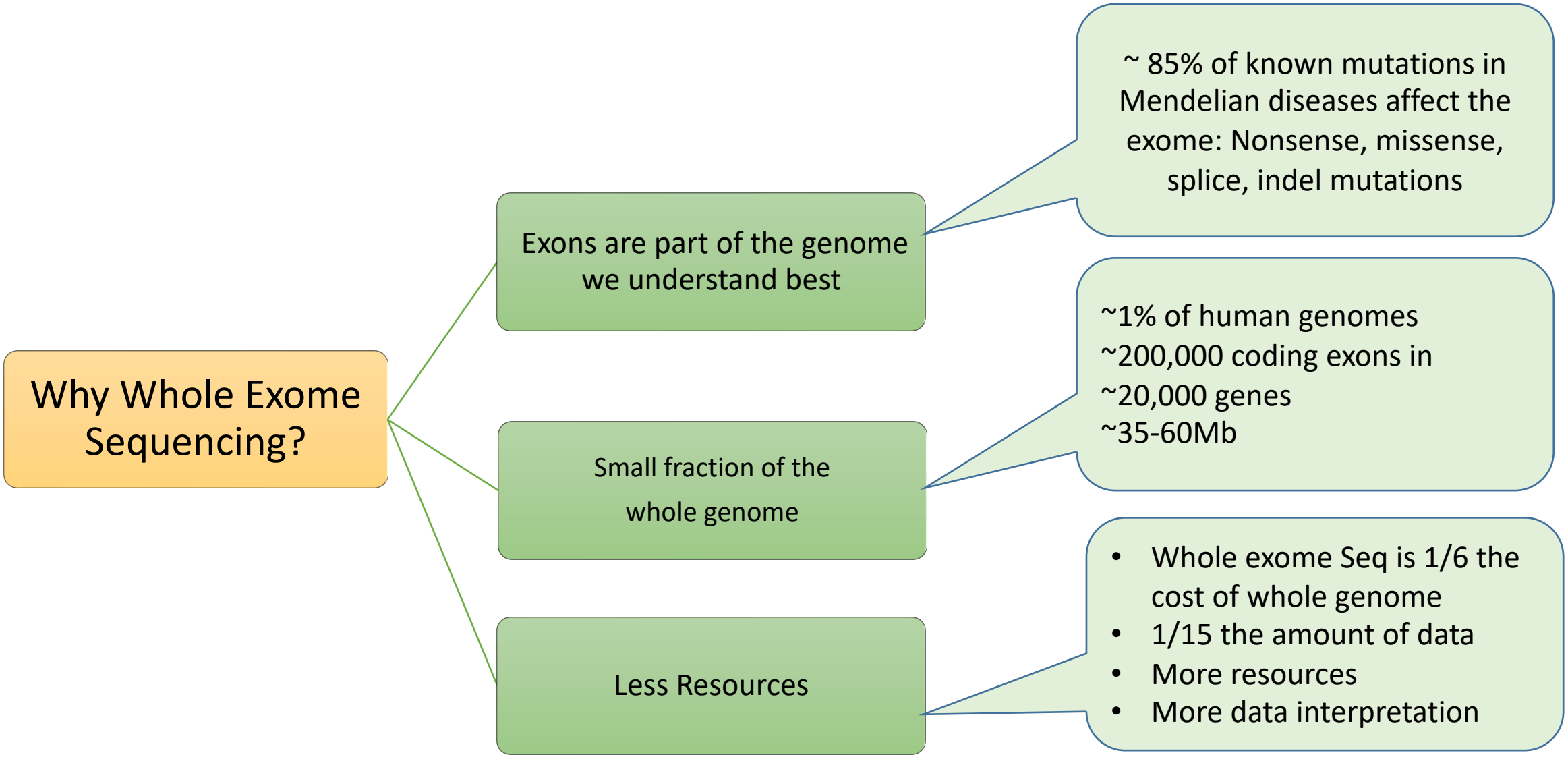
diagnosis



"Genome analysis"

EGFR mutation	EGFR & KRAS mutations	ALK or ROS fusion
RTK inhibitor	Immune therapy with or without chemotherapy	ALK/ROS inhibitor

Improving diagnostic yield in pediatric cancer precision medicine
Elaine R Mardis


Why Whole Exome Sequencing?
Exons are part of the genome we understand best
~ 85% of known mutations in Mendelian diseases affect the exome: Nonsense, missense, splice, indel mutations
Small fraction of the whole genome
~1% of human genomes
~200,000 coding exons in ~20,000 genes
~35-60Mb
Less Resources
• Whole exome Seq is 1/6 the cost of whole genome
• 1/15 the amount of data
• More resources
• More data interpretation

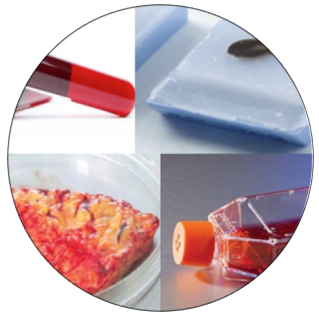
Comparing different Sequencing technologies

	Sanger	Targeted	Exome	Whole Genome
Pros	• Accurate • Cheap per exon • High turn-around	• Optimization possible • Low chance of incidental findings • Easy analysis	• No bias for genes • Standardized workflow • Re-use of performed exomes to interpret new ones	• No sequencing bias • Detect SVs and SNVs
Cons	• Low diagnostic yield for genetically heterogeneous diseases	• Design and re-design required • Different designs for different disorders	• No non-coding regions • Sequencing bias • Incidental findings	• Data analysis bottleneck • Interpretation of non-coding regions • Expensive, time-consuming

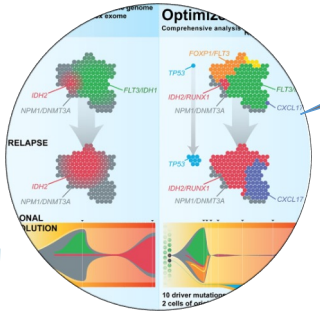
http://webinar.sciencemag.org/webinar/archive/exome-sequencing-today%E2%80%99s-lab

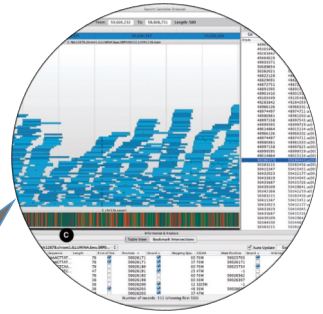
Factors Impacting Cancer Genomic Data

Bio-samples
Fresh frozen tissue,
Cell lines
FFPE tissue



Tumor purity, immune cells, fibroblasts and blood vessels.

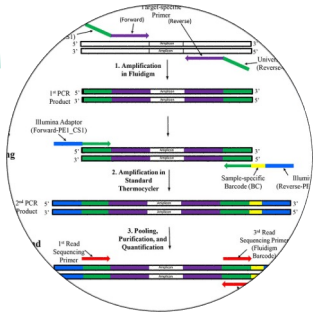


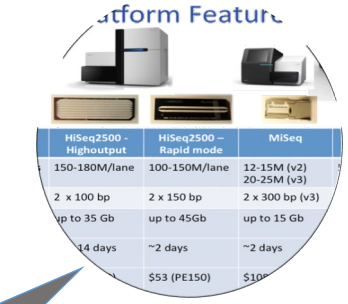


Algorithm
Reads alignment
Mapping quality score
Quality filter
Mutation allele frequency

SEQC II Project

DNA Extraction
Fragmentation
Library prep





System errors
Different platform

Different sequencing sites

Whole exome DNA sources

Tumor DNA	• Fresh frozen (FF) • Formalin-Fixed Paraffin-Embedded (FFPE) • Cell line • Xenograft
Normal tissue	• Blood • Tumor adjacent tissue

OPEN ACCESS Freely available online

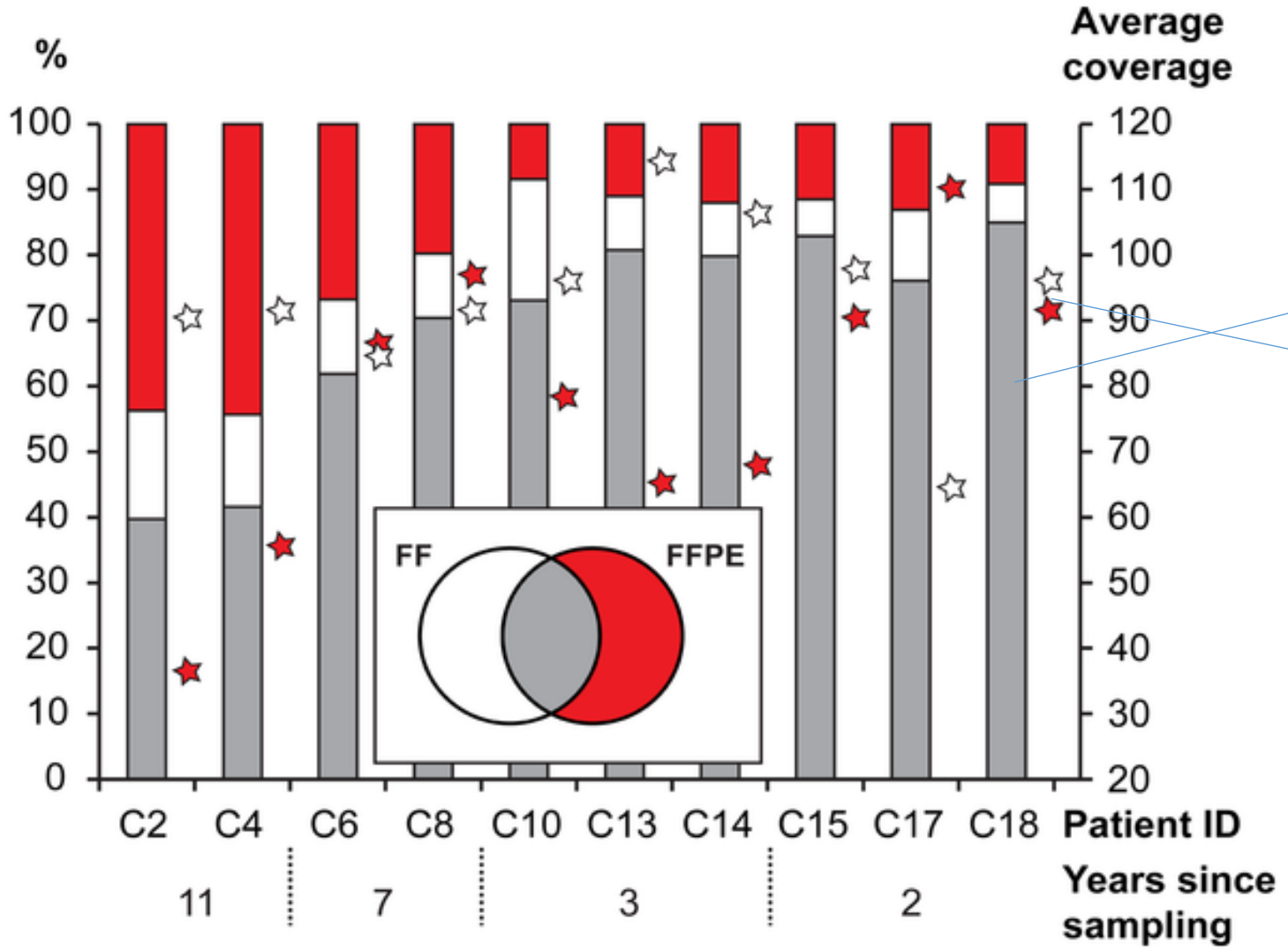
PLOS ONE

Next-Generation Sequencing of RNA and DNA Isolated from Paired Fresh-Frozen and Formalin-Fixed Paraffin-Embedded Samples of Human Cancer and Normal Tissue

Jakob Hedegaard[1]*, Kasper Thorsen[1], Mette Katrine Lund[2], Anne-Mette K. Hein[3], Stephen Jacques Hamilton-Dutoit[4], Søren Vang[1], Iver Nordentoft[1], Karin Birkenkamp-Demtröder[1], Mogens Kruhøffer[2], Henrik Hager[4], Bjarne Knudsen[3], Claus Lindbjerg Andersen[1], Karina Dalsgaard Sørensen[1], Jakob Skou Pedersen[1], Torben Falck Ørntoft[1], Lars Dyrskjøt[1]

1 Department of Molecular Medicine (MOMA), Molecular Diagnostic Laboratory, Aarhus University Hospital, Skejby, Aarhus, Denmark, **2** AROS Applied Biotechnology A/S, Science Park Skejby, Aarhus, Denmark, **3** CLC bio, Aarhus, Denmark, **4** Institute of Pathology, Aarhus University Hospital, Aarhus, Denmark

Variants detected in exome sequencing data from the paired FF/FFPE samples



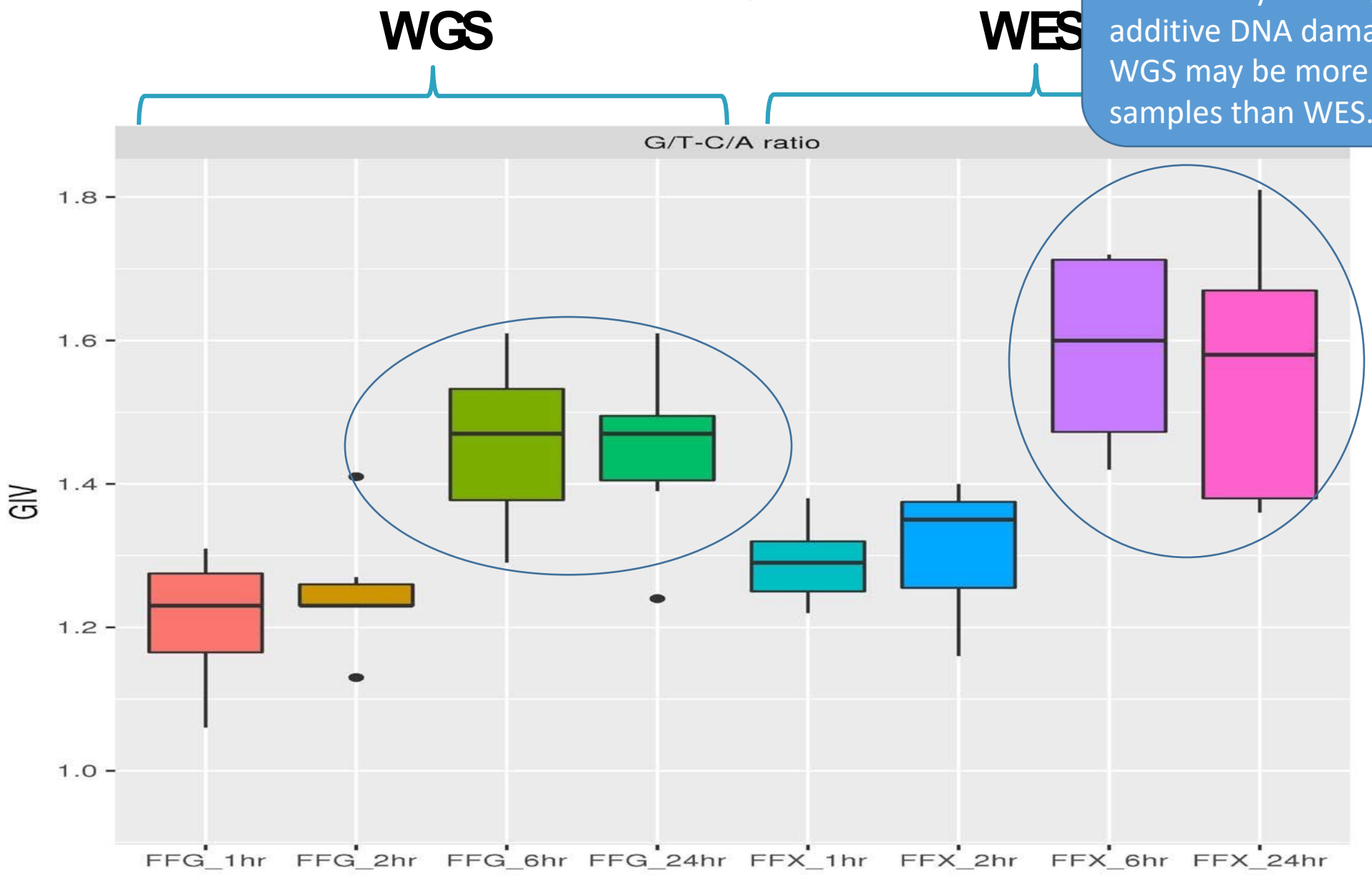
common (gray), exclusively FF (white) and exclusively FFPE (red)

average coverage stars for FF (white) and FFPE (red

Hedegaard, J. et al. (2014) PLoS ONE 9(5): e98187.

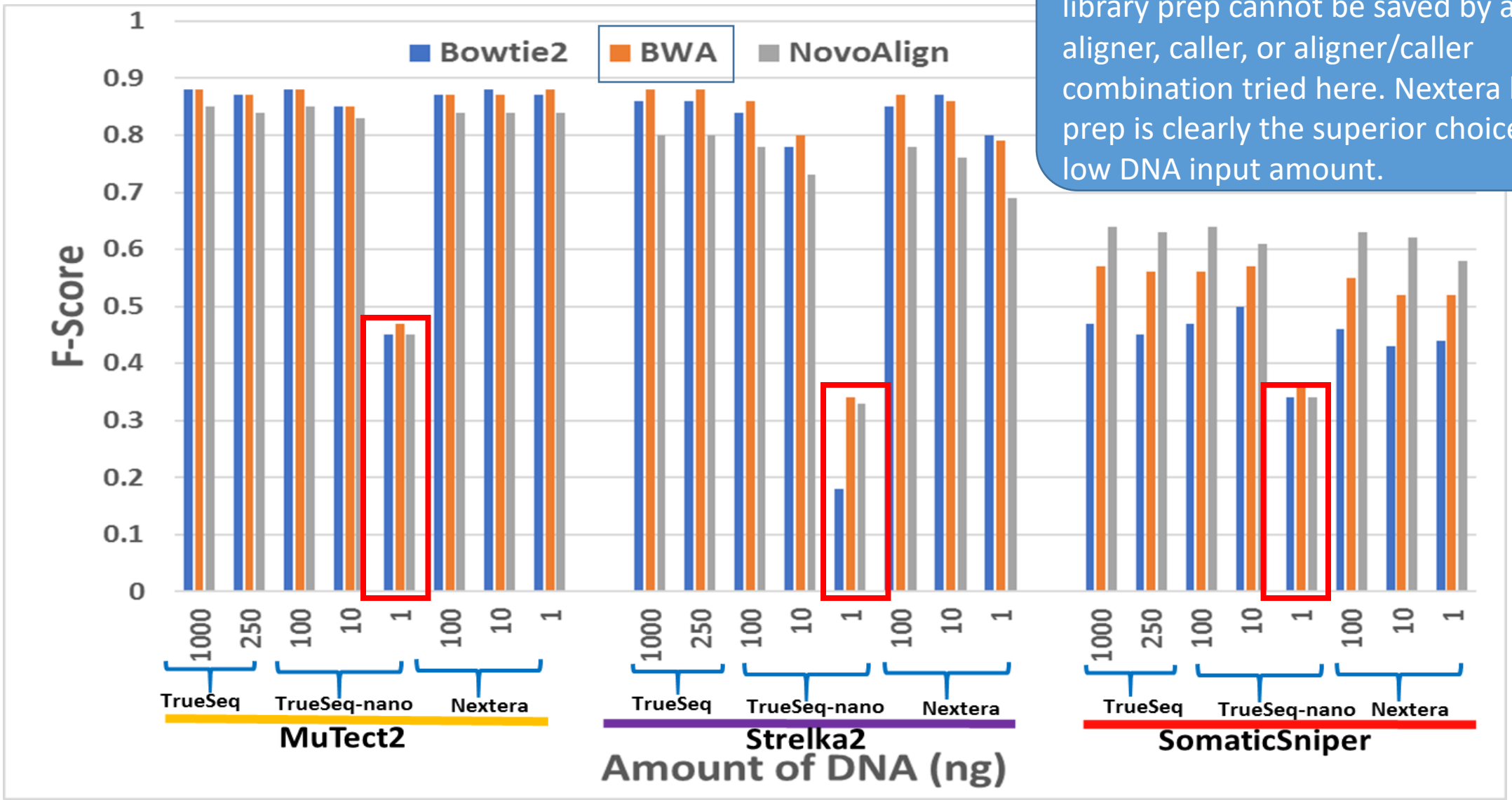
FFPE artifacts

Formaldehyde also causes the deamination of guanine. formaldehyde fixing have an additive DNA damage effect. WGS may be more suitable for FFPE samples than WES.

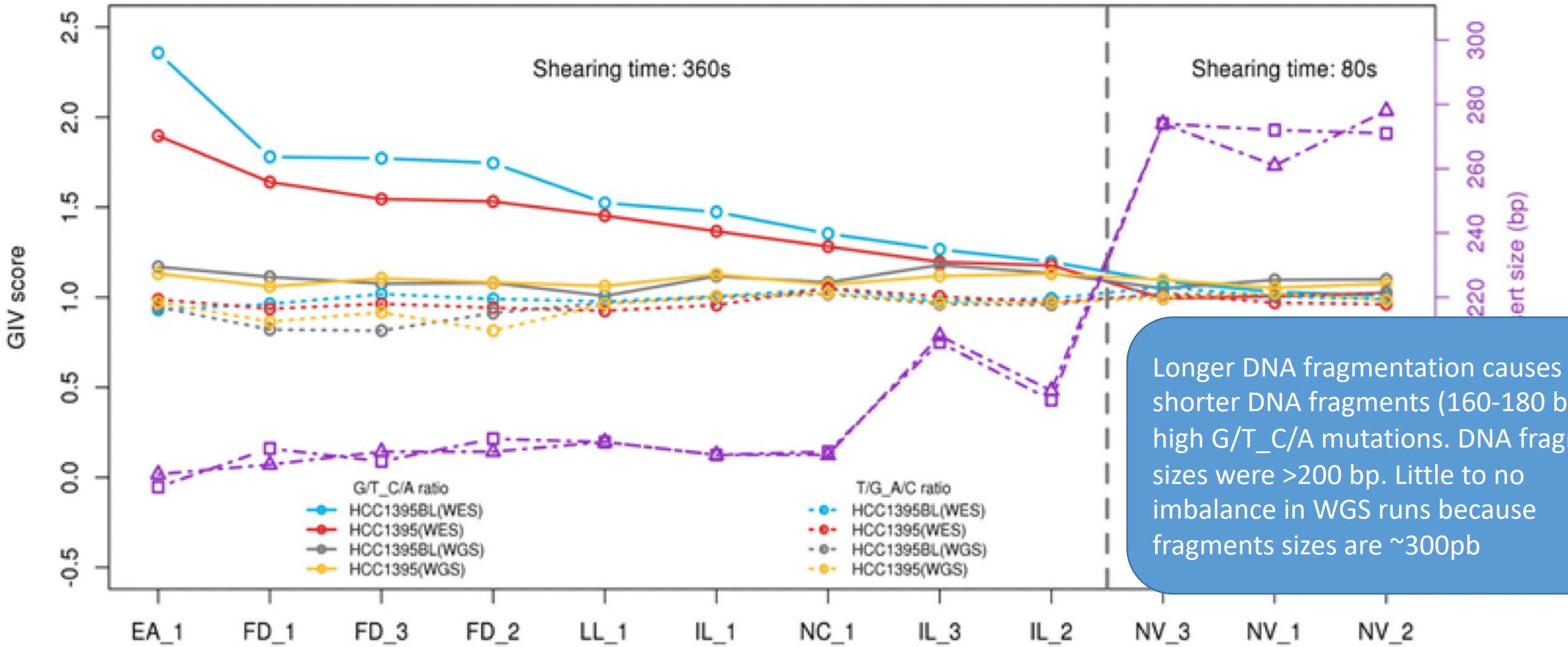


Effect of DNA inputs (How much DNA to start with)

1ng of DNA input in a TruSeq-Nano library prep cannot be saved by any aligner, caller, or aligner/caller combination tried here. Nextera library prep is clearly the superior choice for low DNA input amount.



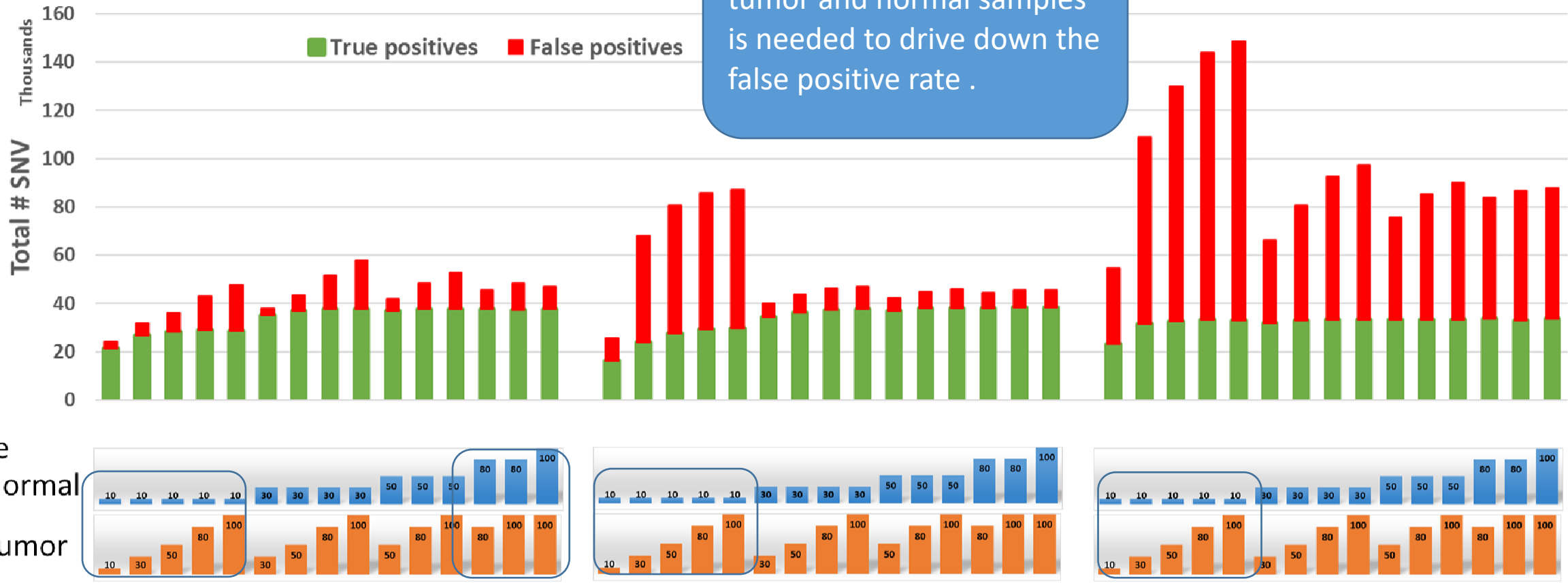
G/T_C/A mutation caused by oxidative DNA damage



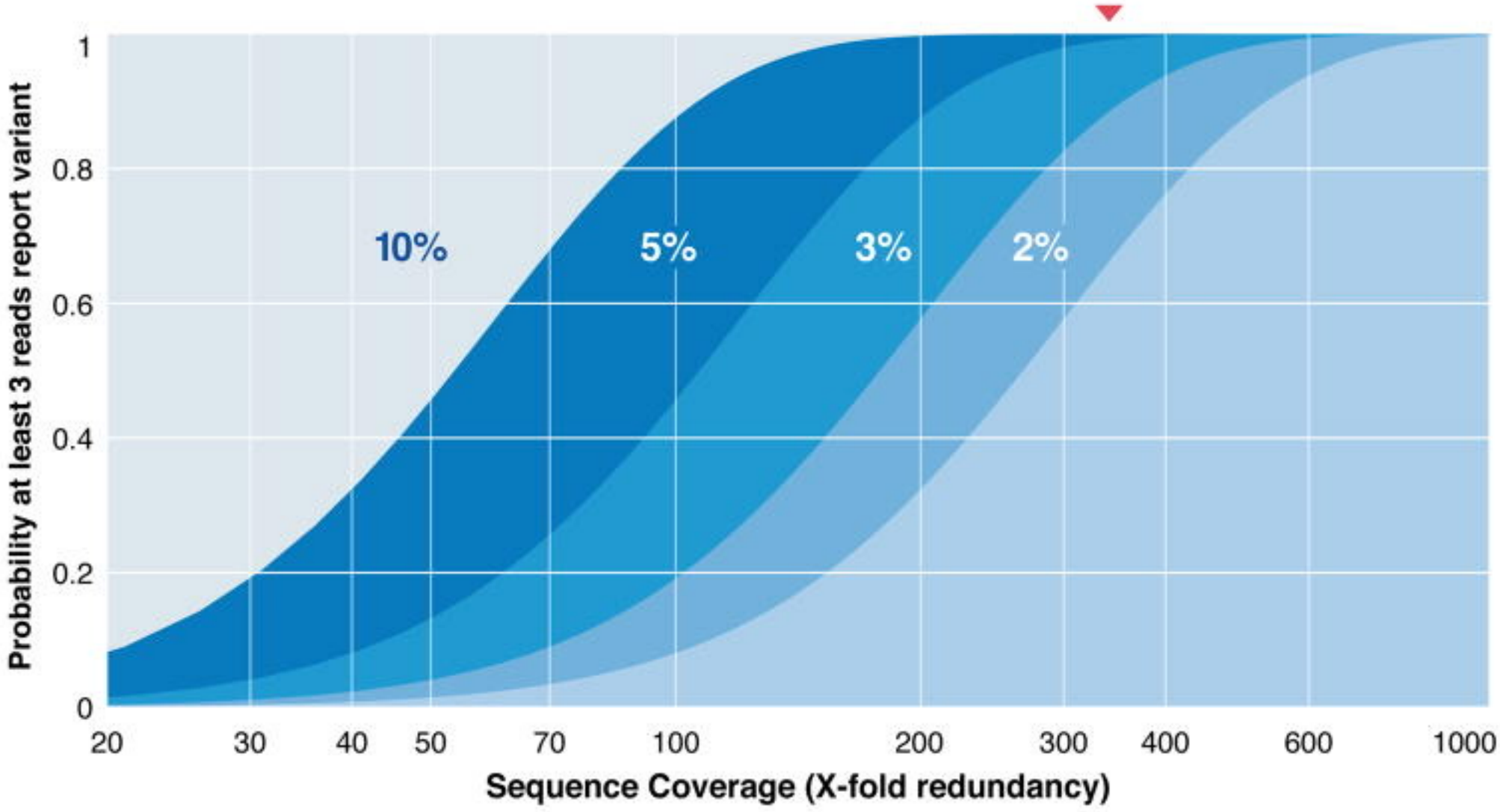
Longer DNA fragmentation causes shorter DNA fragments (160-180 b high G/T_C/A mutations. DNA frag sizes were >200 bp. Little to no imbalance in WGS runs because fragments sizes are ~300pb

reads coverages (Non-analytical factors)

>30X coverage for both tumor and normal samples is needed to drive down the false positive rate .



High coverage is needed for low tumor



Ding, L. et al. (2014) Nat Rev Genet. 15(8):556-70

Exome Sequencing Workflow

Times can always vary at each step depending on reference lab

Tissue Input

Isolation

Fragment

+ Library Prep

Enrichment

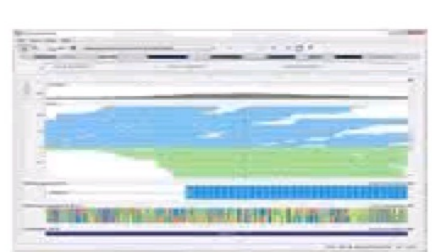
Purification

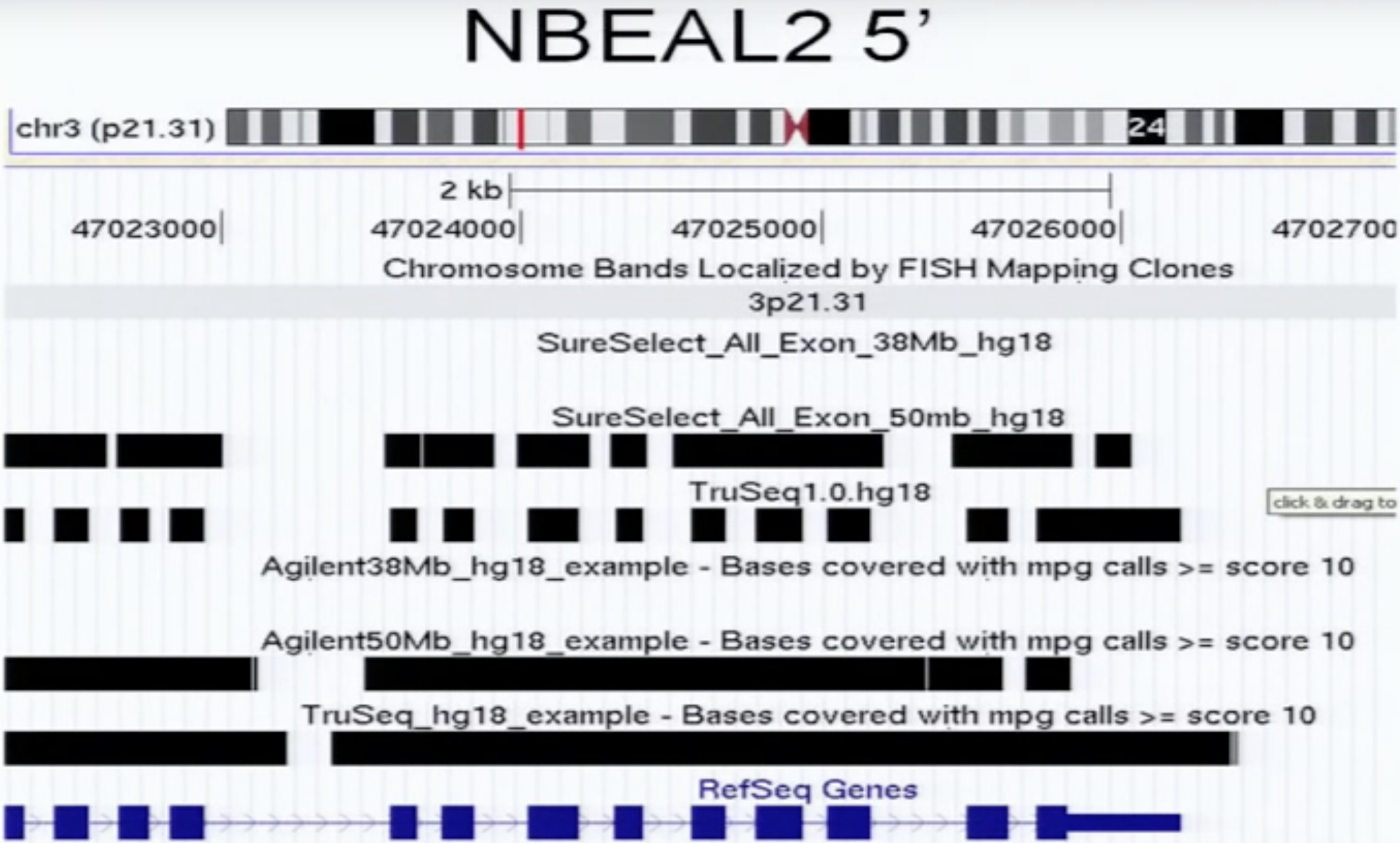
QC

+Pooling

Sequence

Analysis




NBEAL2 5'
chr3 (p21.31)
24
2 kb
47023000
47024000
47025000
47026000
4702700
Chromosome Bands Localized by FISH Mapping Clones
3p21.31
SureSelect_All_Exon_38Mb_hg18
SureSelect_All_Exon_50mb_hg18
TruSeq1.0.hg18
click & drag to
Agilent38Mb_hg18_example - Bases covered with mpg calls >= score 10
Agilent50Mb_hg18_example - Bases covered with mpg calls >= score 10
TruSeq_hg18_example - Bases covered with mpg calls >= score 10
RefSeq Genes

RESEARCH ARTICLE Open Access

Performance comparison of four exome capture systems for deep sequencing

Chandra Sekhar Reddy Chilamakuri[1,3*], Susanne Lorenz[1,3,4], Mohammed-Amin Madoui[1,4], Daniel Vodák[1,5], Jinchang Sun[1,3,4], Eivind Hovig[1,2,3,5], Ola Myklebost[1,3] and Leonardo A Meza-Zepeda[1,3,4*]

Abstract

Background: Recent developments in deep (next-generation) sequencing technologies are significantly impacting medical research. The global analysis of protein coding regions in genomes of interest by whole exome sequencing is a widely used application. Many technologies for exome capture are commercially available; here we compare the performance of four of them: NimbleGen's SeqCap EZ v3.0, Agilent's SureSelect v4.0, Illumina's TruSeq Exome, and Illumina's Nextera Exome, all applied to the same human tumor DNA sample.

Results: Each capture technology was evaluated for its coverage of different exome databases, target coverage efficiency, GC bias, sensitivity in single nucleotide variant detection, sensitivity in small indel detection, and technical reproducibility. In general, all technologies performed well; however, our data demonstrated small, but consistent differences between the four capture technologies. Illumina technologies cover more bases in coding and untranslated regions. Furthermore, whereas most of the technologies provide reduced coverage in regions with low or high GC content, the Nextera technology tends to bias towards target regions with high GC content.

Conclusions: We show key differences in performance between the four technologies. Our data should help researchers who are planning exome sequencing to select appropriate exome capture technology for their particular application.

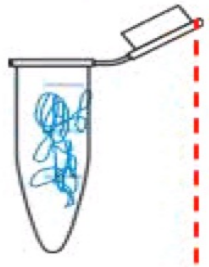
Keywords: Exome capture technology, Next-generation sequencing, Coverage efficiency, Enrichment efficiency, GC bias, Single nucleotide variant, Indel

Exome Sequencing Workflow

Times can always vary at each step depending on reference lab

Tissue Input

Isolation

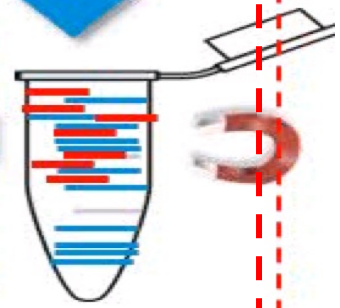


Fragment

+ Library Prep

nrichment

Purification



QC

+Pooling

Sequence

Analysis

illumina

AREAS OF INTEREST TECHNIQUES SYSTEMS PRODUCTS & SERVICES INFORMATICS SCIENTIFIC CONTENT COMPANY SUPPORT SEARCH

ALL SYSTEMS Sequencing Platforms Microarray Scanners NGS PLATFORM SELECTOR QUESTIONS

Sequencing systems for every lab

MiniSeq MiSeq NextSeq HiSeq NovaSeq HiSeq X

KEY APPLICATIONS

Production-Scale Whole-Genome Sequencing

ON THE HISEQ X FIVE SYSTEM

Population-Scale Whole-Genome Sequencing

ON THE HISEQ X TEN SYSTEM

Population power for whole-genome sequencing.

Explore HiSeq X Explore All Sequencing Platforms

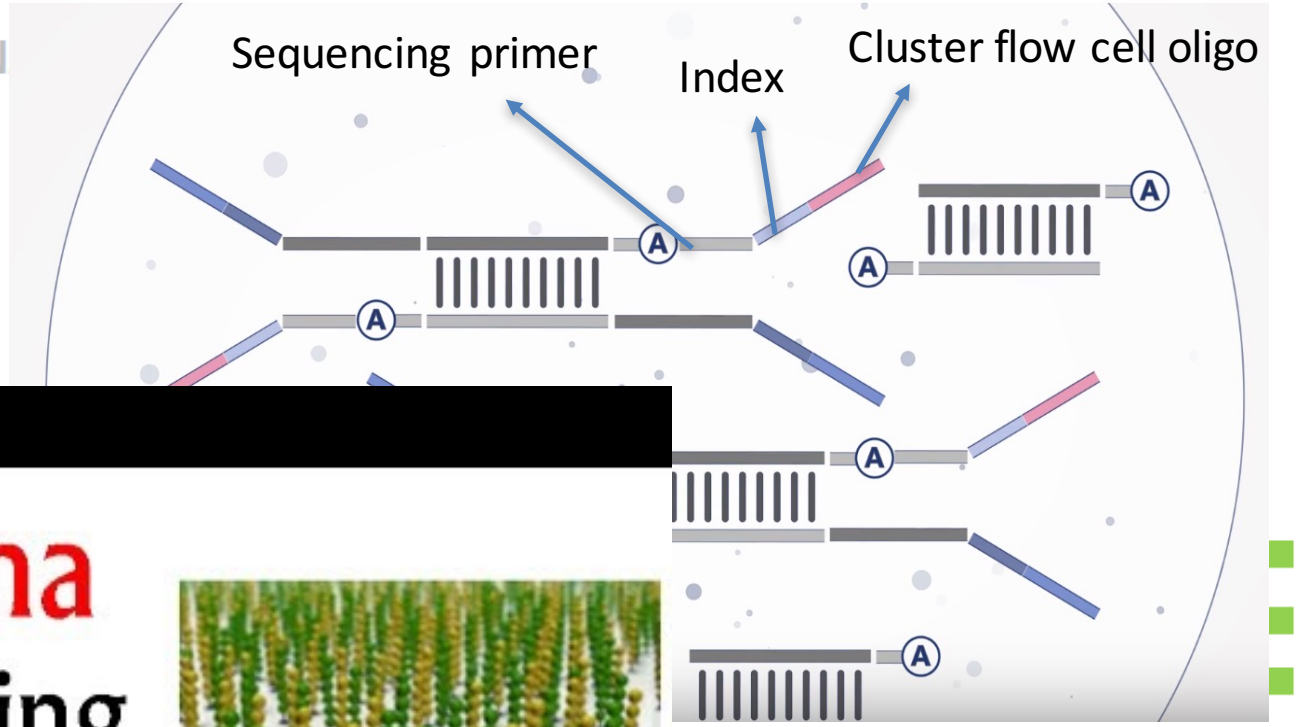
https://www.illumina.com/systems.html

CLINICAL EXOME SEQUENCING

Work flow :

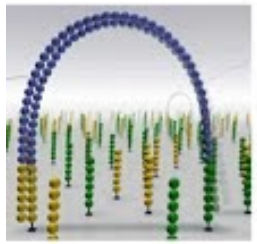
DNA (Sheared DN

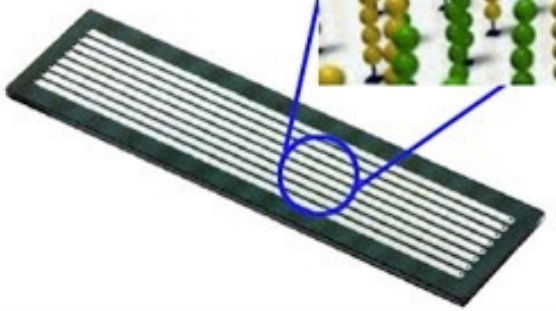
Library prep



[redacted]

Illumina sequencing





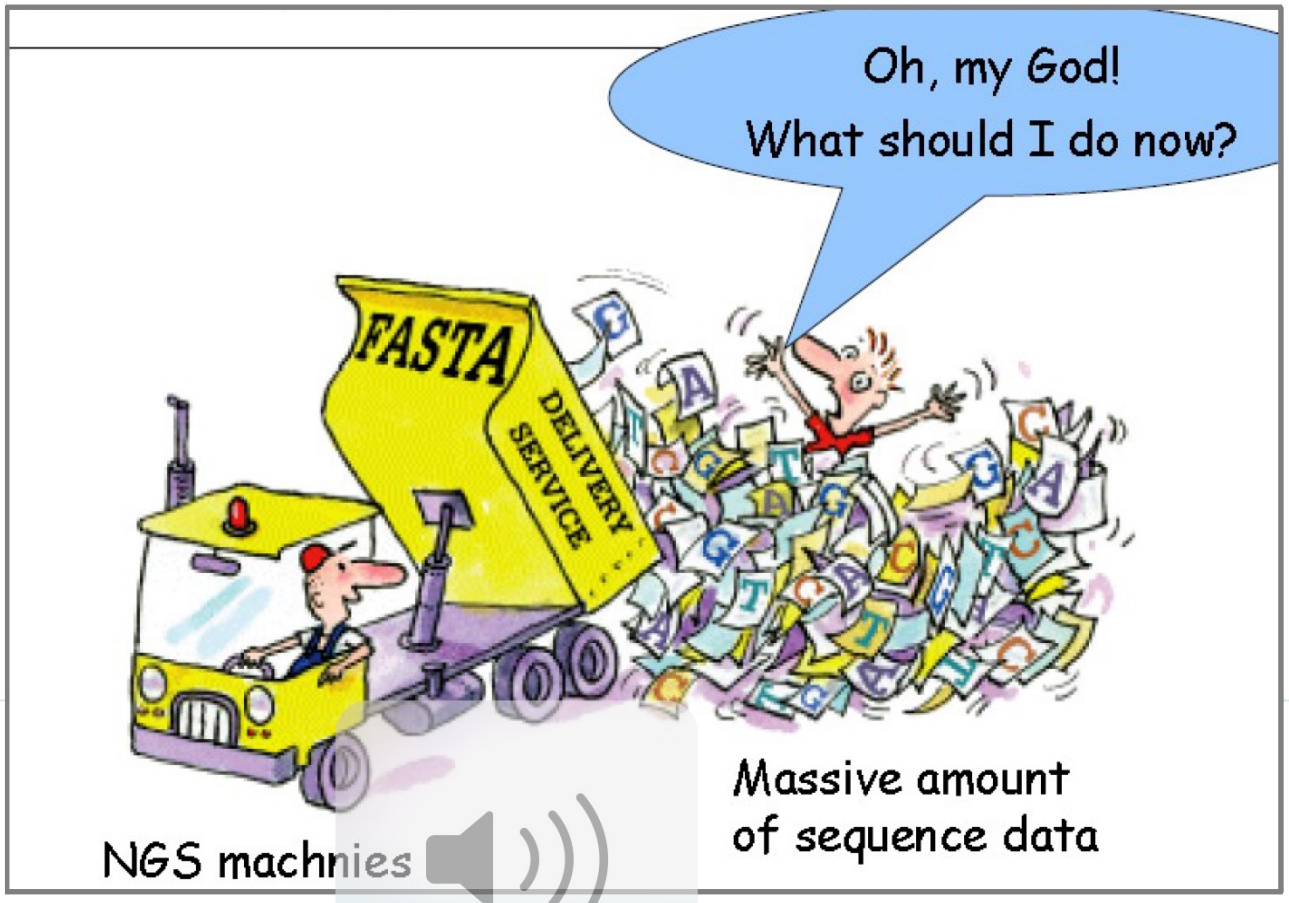
[redacted]

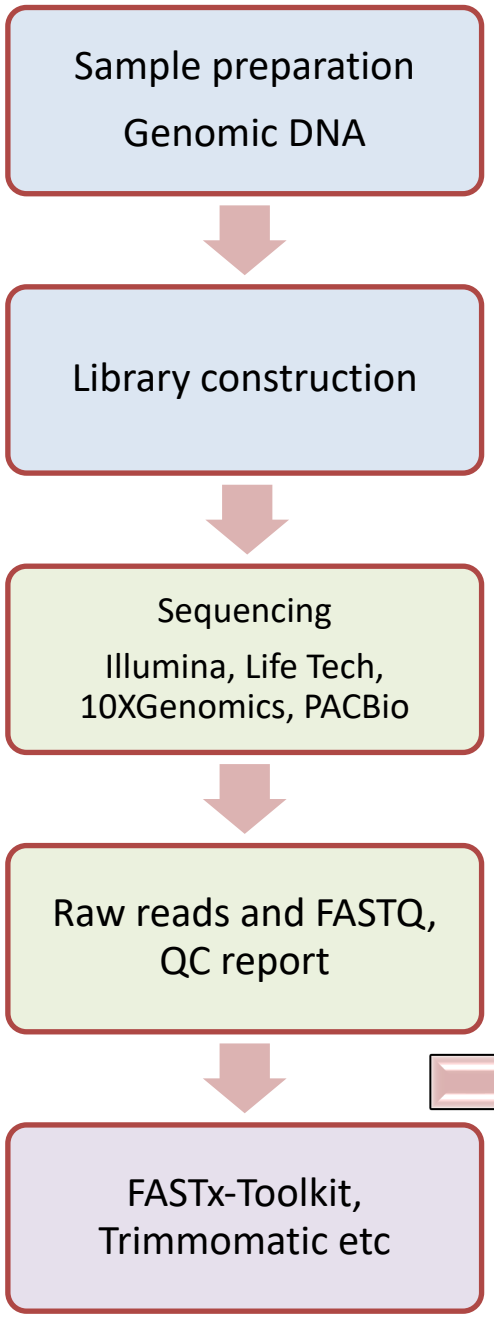
Adapters attach flow cells for cluster formation

Animation of exome sequencing

https://www.youtube.com/watch?v=womKfikWlxM

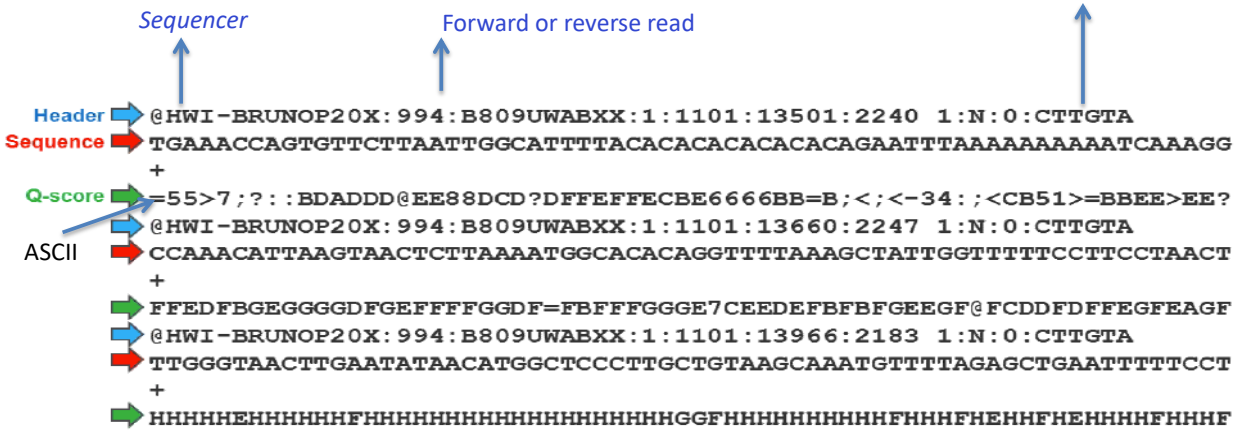
I have my sequences/images. Now what?





Quality Control and Preprocessing

FASTQ format "Raw Reads"



12 million to 3 Billion in a run

Phred Score

- Phred quality scores are defined as a property which is logarithmically related to the base-calling error probabilities

Phred qualities

Quality value	Chance it is wrong	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

- $Q = -10 \log_{10} P \quad <=> \quad P = 10^{-Q/10}$
  - Q = Phred quality score
  - P = probability of base call being incorrect

Sample preparation

Genomic DNA

Library construction

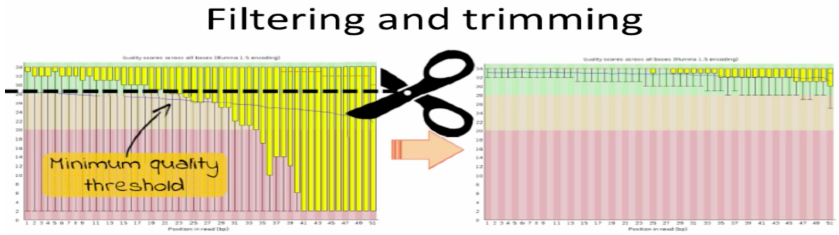
Sequencing

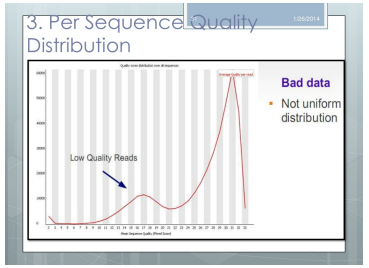
Illumina, Life Tech, 10XGenomics, PACBio

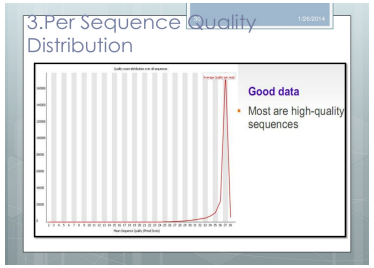
Raw reads and FASTQ, QC report

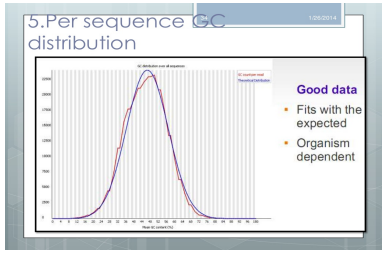
FASTx-Toolkit, Trimmomatic etc

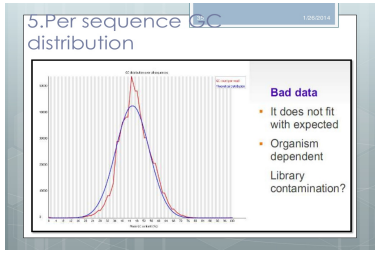
Quality Control and Preprocessing

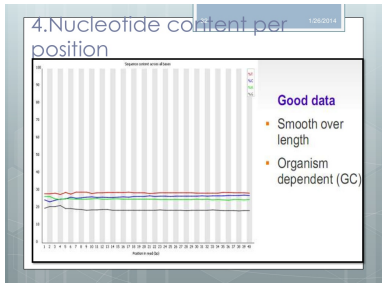


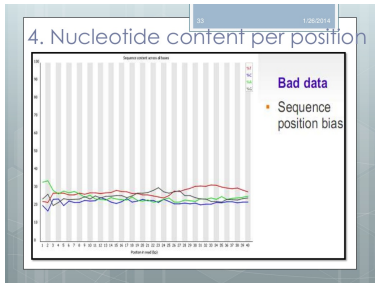




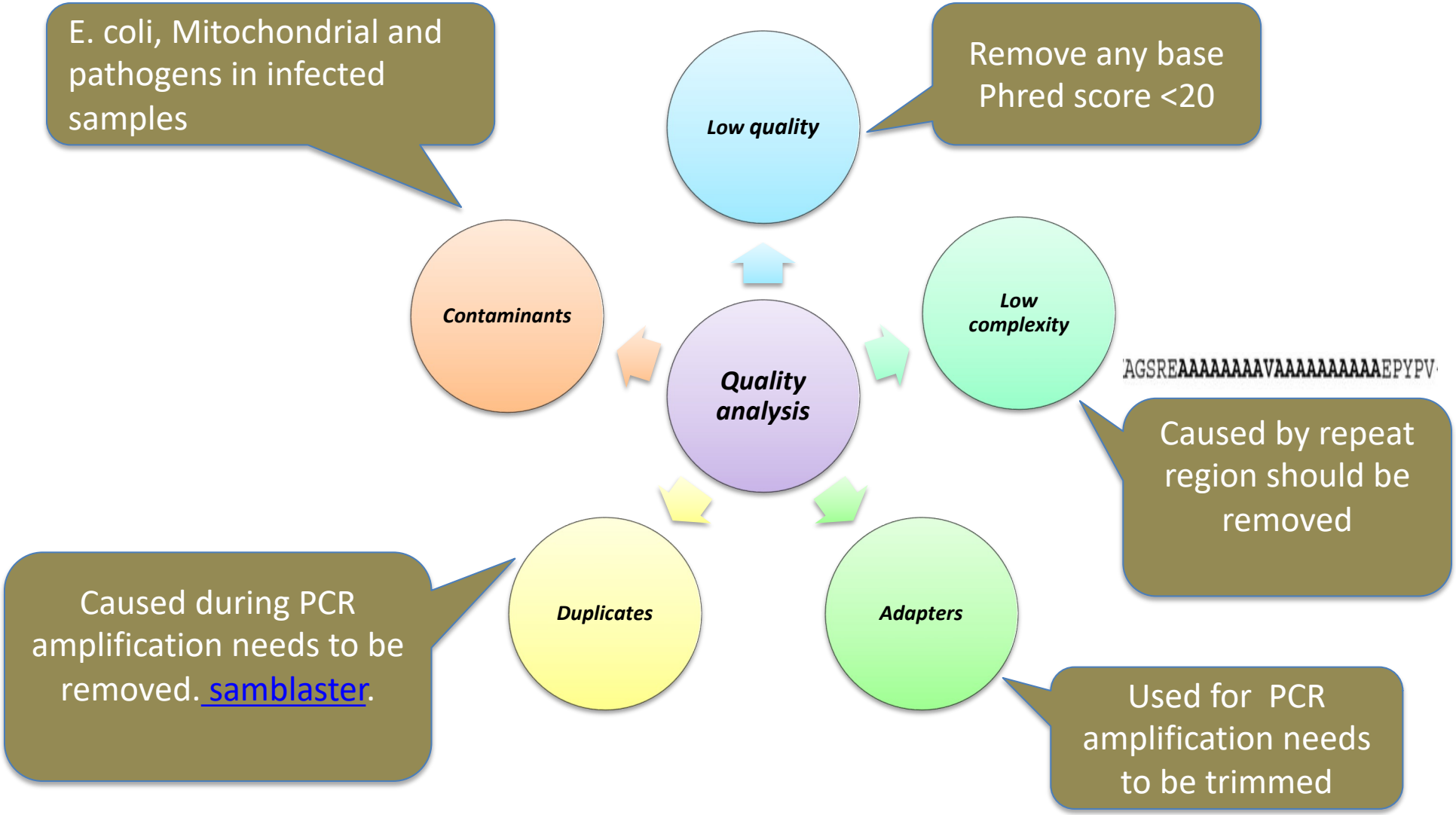








Quality Control and Preprocessing



Alignments and Mapping of raw reads

SAM/BAM Format files

SAM Format

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

BAM File

- SAM (Sequence Alignment/Map) files are typically converted to Binary Alignment/Map (BAM) format
- They are binary
  - Can NOT be opened like a text file
  - Compressed = Smaller (great for server storage)
  - Sorted and indexed = Daster
- Once you have a file in a BAM format you can delete your aligned read files
  - You can recover the FASTQ reads from the BAM format
  - Delete the aligned reads – trimmed/filtered reads
  - Do not delete the RAW reads – zip them and save them

Sample preparation

Genomic DNA

↓

Library construction

↓

Sequencing

Illumina, Life Tech, 10XGenomics, PACBio

↓

Raw reads and FASTQ, QC report

↓

Map Read

↓

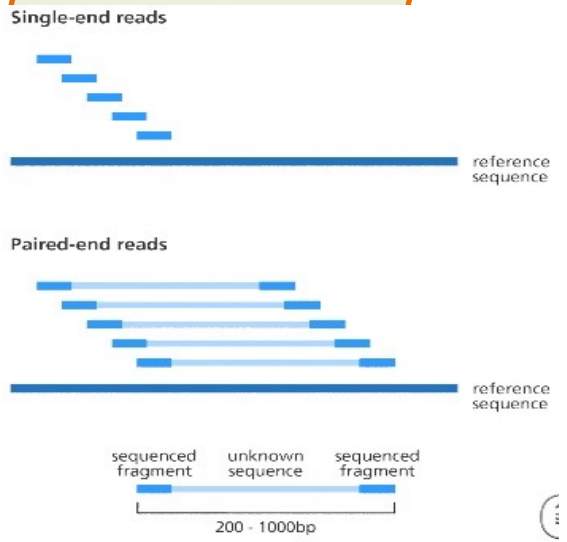
To reference genome, transcriptome, or *de novo* assembly. Human, mouse, etc.

↓

To create an Sequence/Alignment Map (SAM)/Binary Alignment Map (BAM) files

Algorithms

Bwa-Mem, Novoalign, SAMtools



SAM/BAM Files Format

- SAM (Sequence Alignment/Map) format
  - Single unified format for storing read alignments to a reference genome
- BAM (Binary Alignment/Map) format
  - Binary equivalent of SAM
  - Advantages
    - Supports indexing
    - Compact size

Variant Calling Algorithms

Computational and Structural Biotechnology Journal 16 (2018) 15–24

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data

Chang Xu

Life Science Research and Foundation, Qiagen Sciences, Inc., 6951 Executive Way, Frederick, Maryland 21703, USA

ARTICLE INFO

Article history:
Received 8 September 2017
Received in revised form 20 January 2018
Accepted 28 January 2018
Available online 6 February 2018

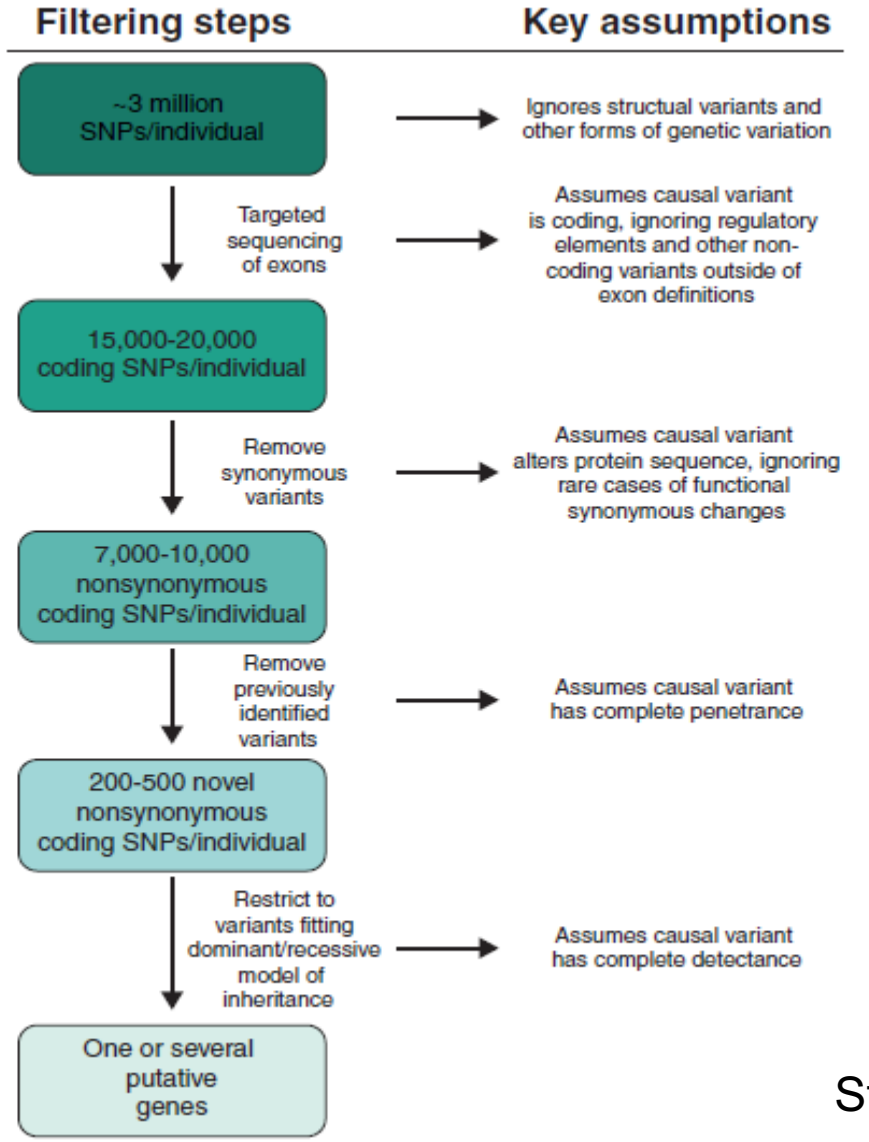
Keywords:
Variant calling
Somatic mutation
Unique molecular identifier
Low-frequency mutation
Benchmarking

ABSTRACT

Detection of somatic mutations holds great potential in cancer treatment and has been a very active r
field in the past few years, especially since the breakthrough of the next-generation sequencing tech
A collection of variant calling pipelines have been developed with different underlying models, filter
data requirements, and targeted applications. This review aims to enumerate these unique feature
state-of-the-art variant callers, in the hope to provide a practical guide for selecting the appropriate p
for specific applications. We will focus on the detection of somatic single nucleotide variants, rangi
traditional variant callers based on whole genome or exome sequencing of paired tumor-normal s
to recent low-frequency variant callers designed for targeted sequencing protocols with unique mo
identifiers. The variant callers have been extensively benchmarked with inconsistent performances
these studies. We will review the reference materials, datasets, and performance metrics that hav
used in the benchmarking studies. In the end, we will discuss emerging trends and future direction
variant calling algorithms.

Variant caller	Type of variant	Single-sample mode	Type of core algorithm
BAYSIC [48]	SNV	No	Machine learning (ensemble caller)
CaVEMan [34]	SNV	No	Joint genotype analysis
deepSNV [38]	SNV	No	Allele frequency analysis
EBCall [37]	SNV, indel	No	Allele frequency analysis
FaSD-somatic [31]	SNV	Yes	Joint genotype analysis
FreeBayes [44]	SNV, indel	Yes	Haplotype analysis
HapMuC [42]	SNV, indel	Yes	Haplotype analysis
JointSNVMix2 [30]	SNV	No	Joint genotype analysis
LocHap [43]	SNV, indel	No	Haplotype analysis
LoFreq [36]	SNV, indel	Yes	Allele frequency analysis
LoLoPicker [39]	SNV	No	Allele frequency analysis
MutationSeq [45]	SNV	No	Machine learning
MuSE [40]	SNV	No	Markov chain model
MuTect [35]	SNV	Yes	Allele frequency analysis
SAMtools [8]	SNV, indel	Yes	Joint genotype analysis
Platypus [41]	SNV, indel, SV	Yes	Haplotype analysis
qSNP [24]	SNV	No	Heuristic threshold
RADIA [26]	SNV	No	Heuristic threshold
Seurat [33]	SNV, indel, SV	No	Joint genotype analysis
Shimmer [25]	SNV, indel	No	Heuristic threshold
SNooPer [47]	SNV, indel	Yes	Machine learning
SNVSniffer [32]	SNV, indel	Yes	Joint genotype analysis
SOAPsnv [27]	SNV	No	Heuristic threshold
SomaticSeq [46]	SNV	No	Machine learning (ensemble caller)
SomaticSniper [28]	SNV	No	Joint genotype analysis
Strelka [17]	SNV, indel	No	Allele frequency analysis
TVC [97]	SNV, indel, SV	Yes	Ion Torrent specific

Identifying causal variants: filtering



Stitziel, Genom Biol 2011

ARTICLE

OPEN

doi:10.1038/nature19057

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P. Birnbaum[2], Jack A. Kosmicki[1,2,6,13], Laramie E. Duncan[1,2,6], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2], Emma Pierce-Hoffman[1,2], Joanne Berghout[14,15], David N. Cooper[16], Nicole Deflaux[17], Mark DePristo[18], Ron Do[19,20,21,22], Jason Flannick[2,23], Menachem Fromer[1,6,19,20,24], Laura Gauthier[18], Jackie Goldstein[1,2,6], Namrata Gupta[2], Daniel Howrigan[1,2,6], Adam Kiezun[18], Mitja I. Kurki[2,25], Ami Levy Moonshine[18], Pradeep Natarajan[2,26,27,28], Lorena Orozco[29], Gina M. Peloso[2,27,28], Ryan Poplin[18], Manuel A. Rivas[2], Valentin Ruano-Rubio[18], Samuel A. Rose[6], Douglas M. Ruderfer[19,20,24], Khalid Shakir[18], Peter D. Stenson[16], Christine Stevens[2], Brett P. Thomas[1,2], Grace Tiao[18], Maria T. Tusie-Luna[30], Ben Weisburd[2], Hong-Hee Won[31], Dongmei Yu[6,25,27,32], David M. Altshuler[2,33], Diego Ardissino[34], Michael Boehnke[35], John Danesh[36], Stacey Donnelly[2], Roberto Elosua[37], Jose C. Florez[2,26,27], Stacey B. Gabriel[2], Gad Getz[18,26,38], Stephen J. Glatt[39,40,41], Christina M. Hultman[42], Sekar Kathiresan[2,26,27,28], Markku Laakso[43], Steven McCarroll[6,8], Mark I. McCarthy[44,45,46], Dermot McGovern[47], Ruth McPherson[48], Benjamin M. Neale[1,2,6], Aarno Palotie[1,2,5,49], Shaun M. Purcell[19,20,24], Danish Saleheen[50,51,52], Jeremiah M. Scharf[2,6,25,27,32], Pamela Sklar[19,20,24,53,54], Patrick F. Sullivan[55,56], Jaakko Tuomilehto[57], Ming T. Tsuang[58], Hugh C. Watkins[44,59], James G. Wilson[60], Mark J. Daly[1,2,6], Daniel G. MacArthur[1,2] & Exome Aggregation Consortium†

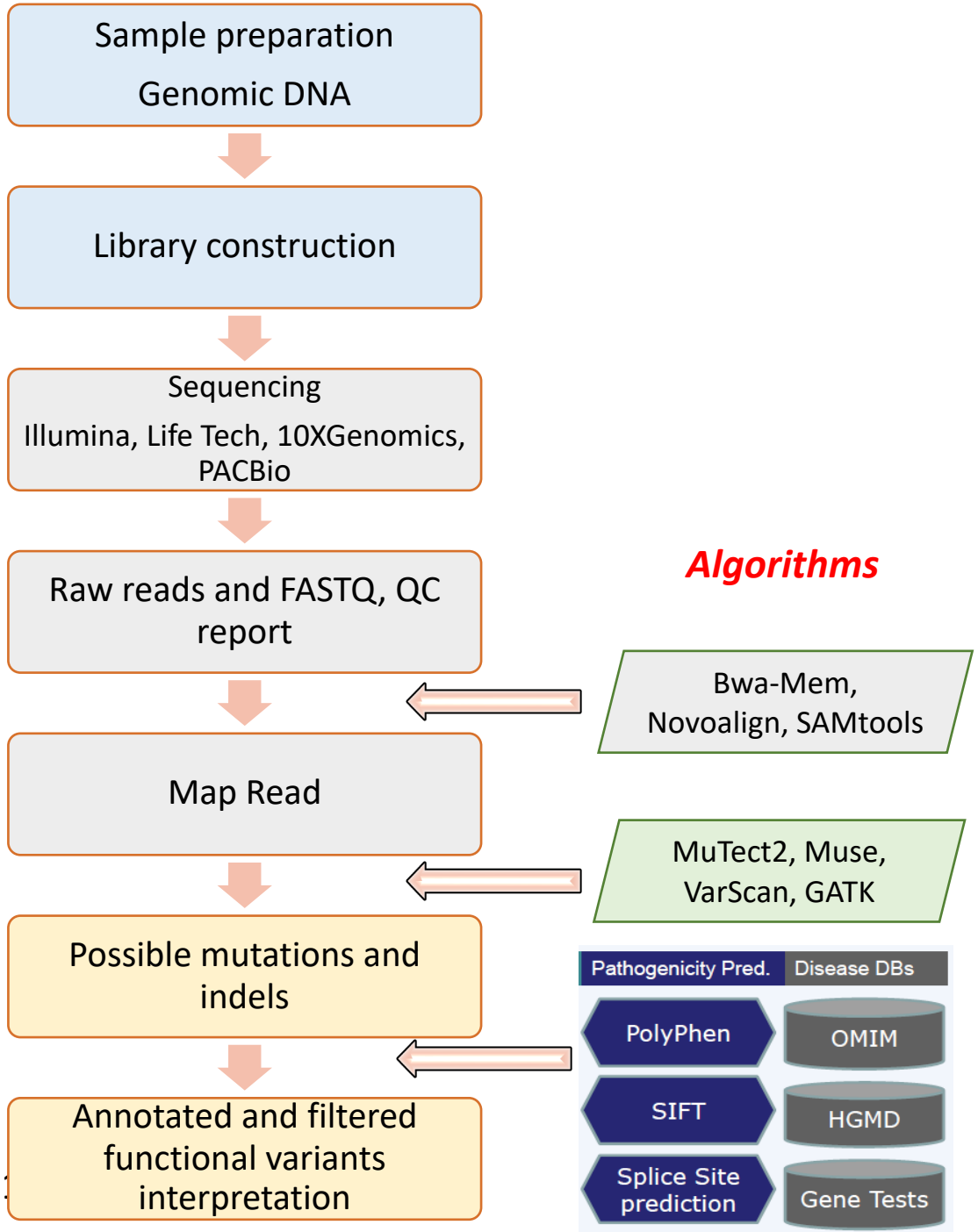
Variants Annotation and Interpretation

Level I
Annotation and analysis of individual genetic alterations

Level II
Population-based analysis of genetic alterations and identification of significant alterations, genes, pathways and networks

Example tools

SNPeff	VEP	MuSiC	MutSig
ANNOVAR	SIFT	Oncodrive	TieDIE
PolyPhen2	CHASM	HotNet	PathScan
MutationAssessor		Dendrix	MEMo
ActiveDriver		PARADIGM	



Ding L et al. (2014) Nat Rev Genet. 1

Variant Annotation

- Gene/intergenic annotation of the variants
- Exonic, intronic, 3'UTR, 5'UTR, promoter region, conserved transcription
- factor binding sites, conserved intergenic region, and active
- transcription region identified using ChIP-Seq experiments.
- Silent, missense, non-sense, frame-shift, in-frame annotation of genic variants.
- Comparison with OncoMD somatic mutation and germline collection.
- Comparison with OMIM, SNPedia and other relevant databases
- Comparison with common SNP databases: 1000-genomes project,
- dbSNP 135, and personal genome variants.

http://www.scigenom.com/whitepapers/exomesequencingandanalysis.pdf

ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

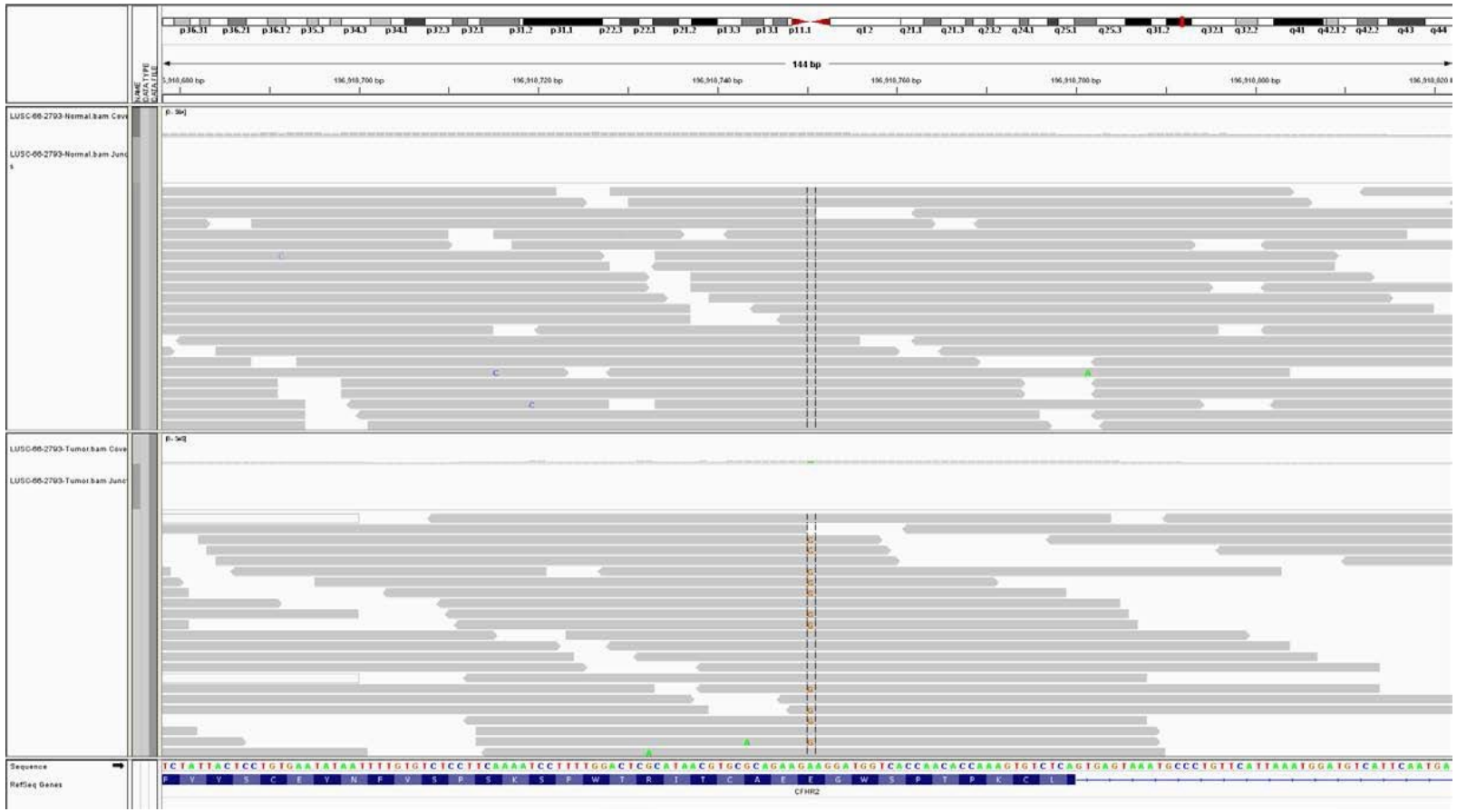
Disclaimer: These ACMG Standards and Guidelines were developed primarily as an educational resource for clinical laboratory geneticists to help them provide quality clinical laboratory services. Adherence to these standards and guidelines is voluntary and does not necessarily assure a successful medical outcome. These Standards and Guidelines should not be considered inclusive of all proper procedures and tests or exclusive of other procedures and tests that are reasonably directed to obtaining the same results. In determining the propriety of any specific procedure or test, the clinical laboratory geneticist should apply his or her own professional judgment to the specific circumstances presented by the individual patient or specimen. Clinical laboratory geneticists are encouraged to document in the patient's record the rationale for the use of a particular procedure or test, whether or not it is in conformance with these Standards and Guidelines. They also are advised to take notice of the date any particular guideline was adopted and to consider other relevant medical and scientific information that becomes available after that date. It also would be prudent to consider whether intellectual property interests may restrict the performance of certain tests and other procedures.

False Negativity/Positivity

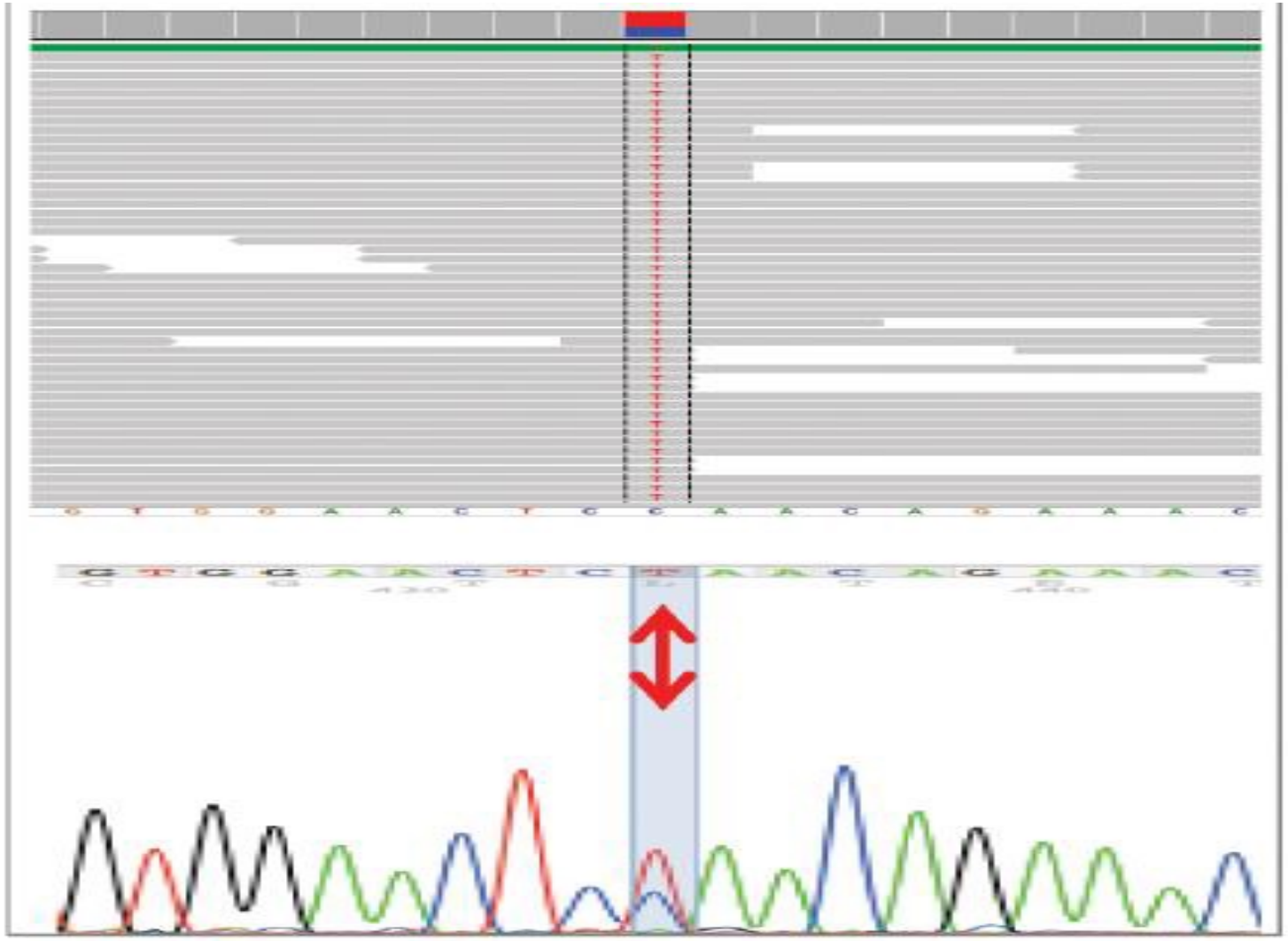
- Most false negatives are due to lack of coverage

- False positives are due to multiple reasons, including:
  - Variant is only called on one strand
  - Variant is only called at the end of the read
  - Coverage of the matched normal at that locus is poor
  - Gene has a pseudogene/paralog and the reads are mis-mapped
  - High sensitivity variant calling algorithms have elevated false positive rates to achieve detection of subclonal variants and low false negative rates

Mutation calls from paired samples



Variant Calling and Validation



Differences to the reference

Reference: C
Sample: T

Gene Set Enrichment Analysis

- Examine whether multiple genes in pre-defined sets have more mutations than expected

Ding, Nat Rev Genet. 2014