Next Generation Sequencing (Short Read) Technologies

Elaine R. Mardis, Ph.D., FAACR

Co-Executive Director, The Steve and Cindy Rasmussen Institute for Genomic Medicine at Nationwide Children's Hospital

The Steve and Cindy Rasmussen Nationwide Foundation Endowed Chair in Genomic Medicine

Professor of Pediatrics and Neurosurgery, The Ohio State University College of Medicine

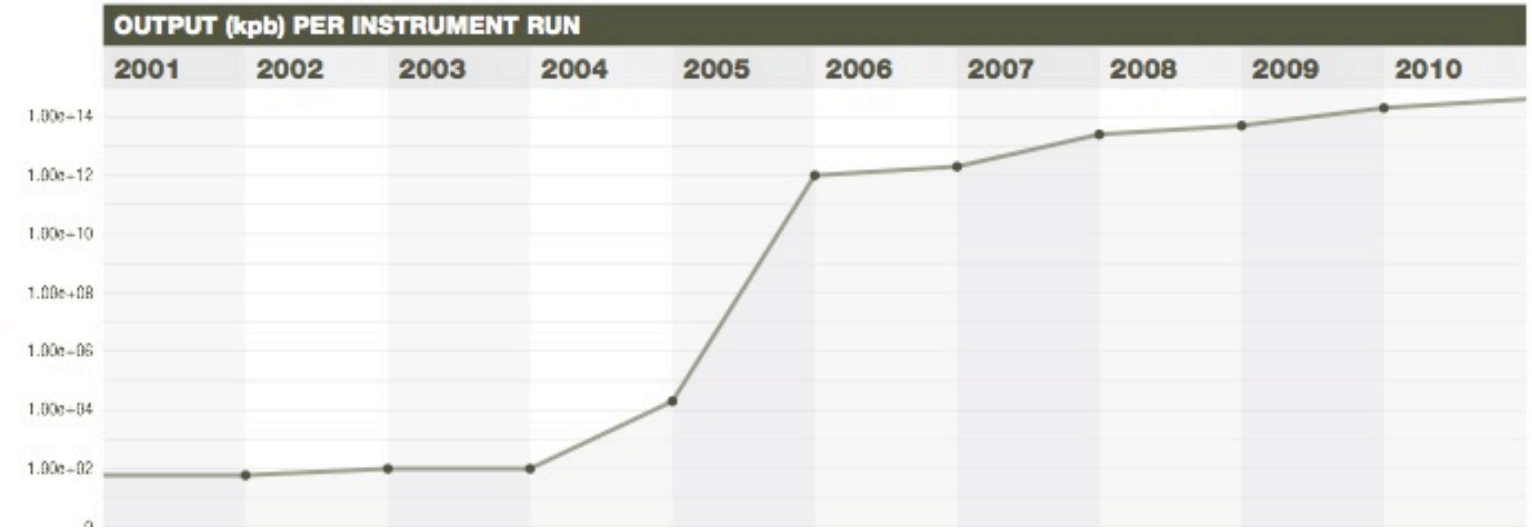
NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.℠

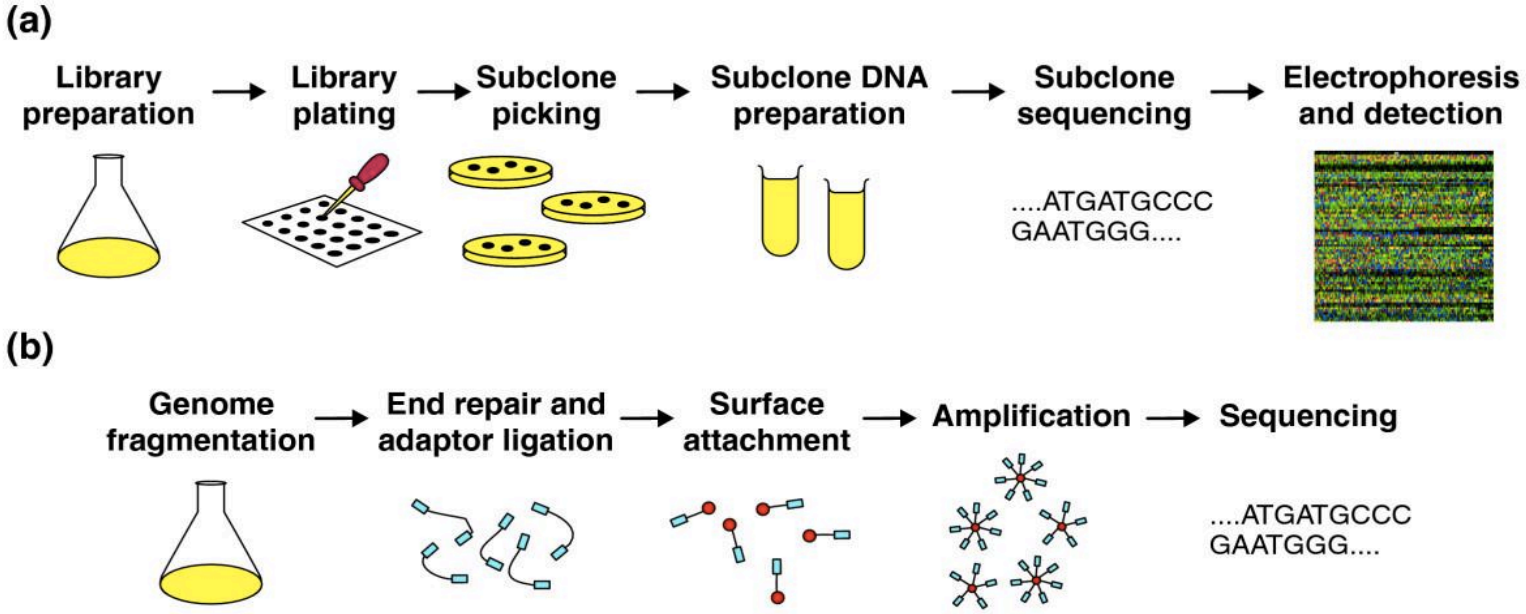
THE OHIO STATE UNIVERSITY
COLLEGE OF MEDICINE

Massively Parallel Sequencing basics

How massively parallel sequencing works

NGS has transformed biomedical inquiry





E.R. Mardis, Nature (2011) 470: 198-203, Ann. Rev. Analyt. Chem. (2013)

Next Generation Sequencing: the basics

- NGS library construction combines DNA or RNA fragments with custom synthetic DNA adapters by ligation or transposon insertion
- The resulting library fragments are amplified on a solid support (either a bead, flat surface or nanowell-covered surface) with covalently attached adapters complimentary to the library adapters
- Sequencing reactions couple nucleotide incorporation and detection in a step-wise fashion, detecting hundreds of millions to billions of sequencing reactions per instrument run = ‘‘massively parallel sequencing’’
- Shorter read lengths than capillary sequencers = requires specialized bioinformatics-based analyses

Input DNA for NGS Libraries

- Generally speaking, NGS sequencing libraries are ‘short insert’, with fragment sizes ranging from ~200-600 bp

- Starting material for NGS libraries may be derived from numerous different sources, including:
  - high molecular weight genomic DNA (cell lines, blood, fresh or frozen tissue, buccal swab/cheek scrape)
  - PCR products, including from multiplex PCR
  - low molecular weight/degraded DNA (formalin-fixed paraffin-embedded (FFPE) tissues, forensic specimens, Neanderthal bones)

- So, the first consideration is the method of DNA isolation needed, and then evaluating the quality/intactness and yield of DNA that results

Evaluating Input DNA Quality and Quantity

DNA Quality/Intactness:

- cheapest = agarose gel and ethidium staining
- easiest = precast gel (Flash)
- large sample numbers: Agilent BioAnalyzer or TapeStation

DNA Quantity:

- NanoDrop (with caveats)
- Qubit
- PicoGreen + 96 well plate reader

If high mw and concentration, fragmentation is next

If low mw and concentration, challenges and adjustments will be required!

Fragmenting High MW Genomic DNA

Mechanical methods:

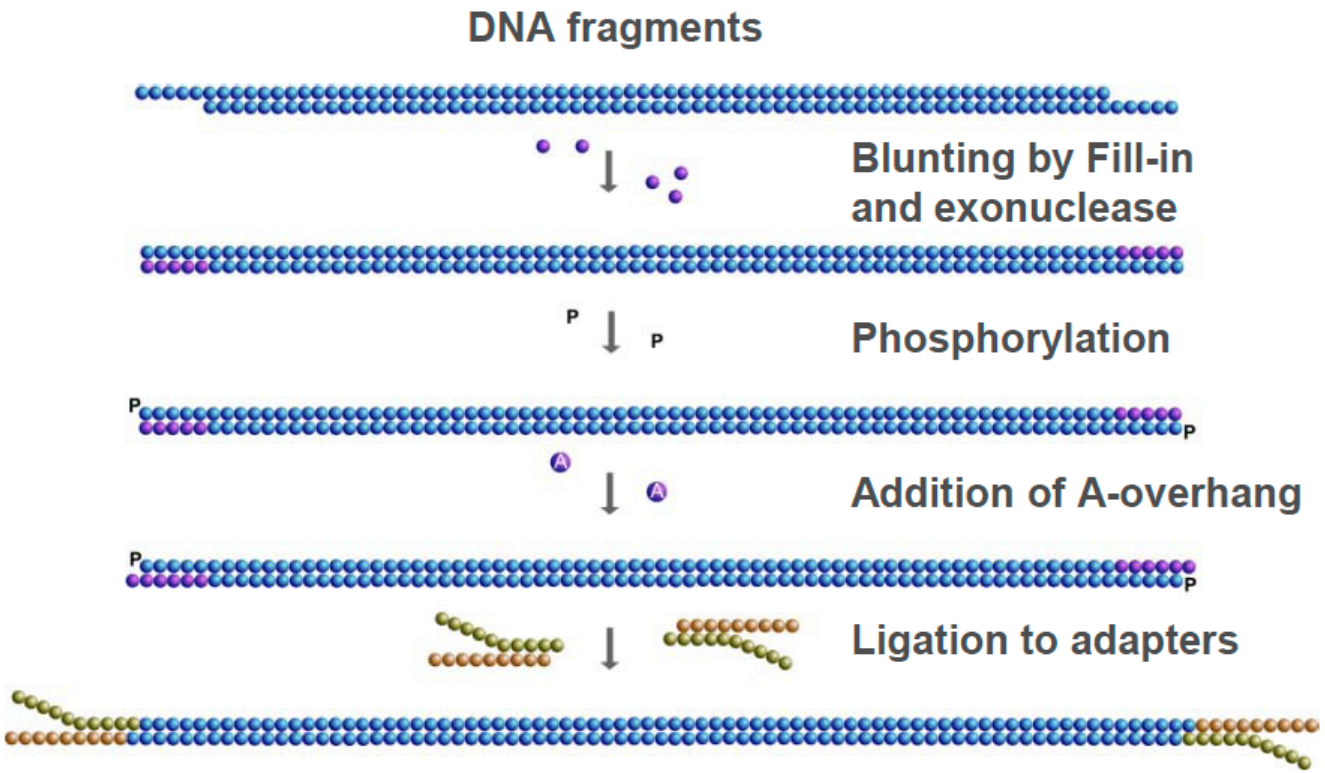
- syringe/needle high pressure shearing device
- HydroShear device
- Covaris ultrasonicator or G-tube device

Enzymatic methods:

- restriction enzyme(s) cleavage
- Zinc treatment
- proprietary nuclease cocktail / “fragmentase”

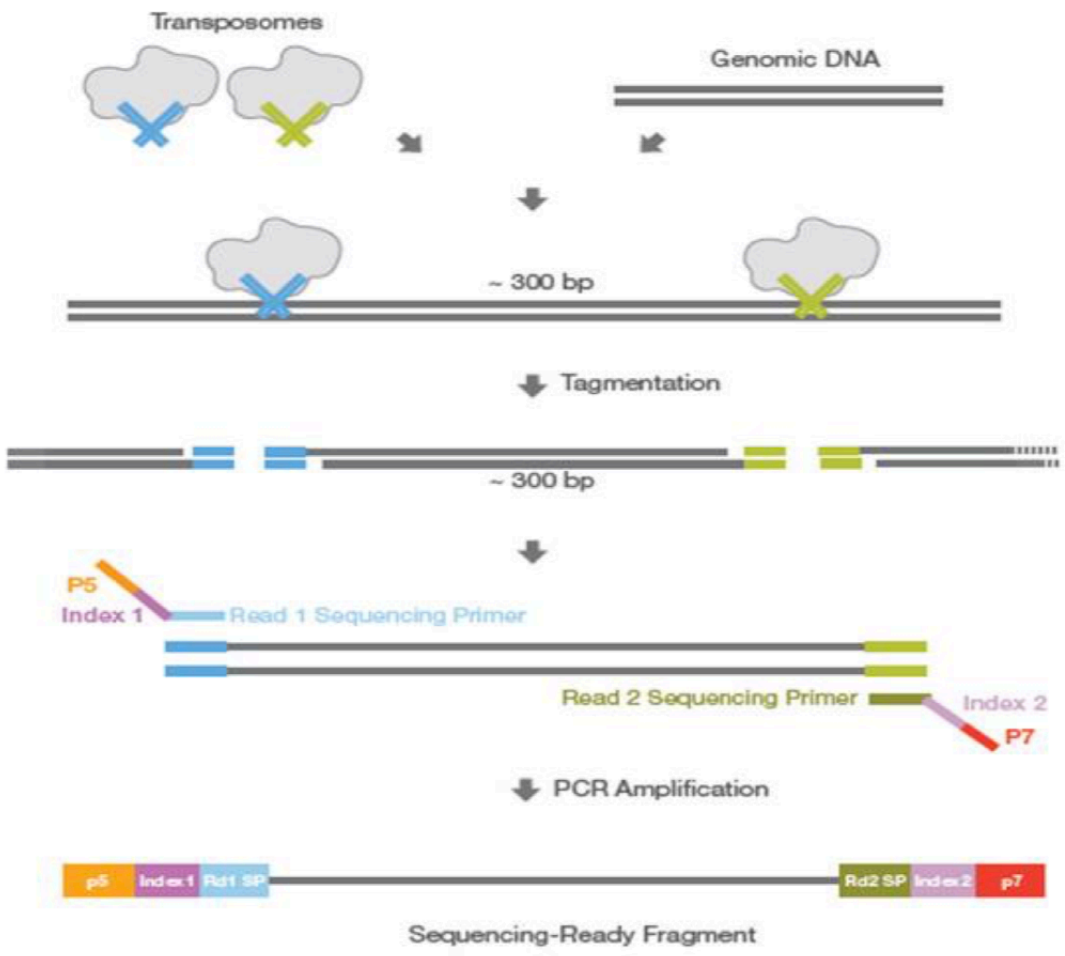
Initially, evaluate one or more post-shearing samples on gel for size range optimization

Adapter Ligation-based Library Construction



- Shear or enzymatically treat high molecular weight DNA to fragment, OR derive DNA fragments from other preparatory methods
- Enzymatic treatments to blunt ends
- Ligate synthetic DNA adapters (with indexes), PCR amplify
- Quantitate library
- Proceed to WGS, or isolate exome or specific panel by hybrid capture

Transposon-based Library Construction



PCR-related Problems in NGS

PCR is an effective vehicle for amplifying DNA, however…

In NGS library construction, PCR can introduce preferential amplification ("jackpotting") of certain fragments

- Duplicate reads with exact start/stop alignments
- Need to "de-duplicate" after alignment and keep only one pair
- Low input DNA amounts favor jackpotting due to lack of complexity in the fragment population

PCR also introduces false positive artifacts due to substitution errors by the polymerase

- If substitution occurs in early PCR cycles, error appears as a true variant
- If substitution occurs in later cycles, error typically is drowned out by correctly copied fragments in the cluster

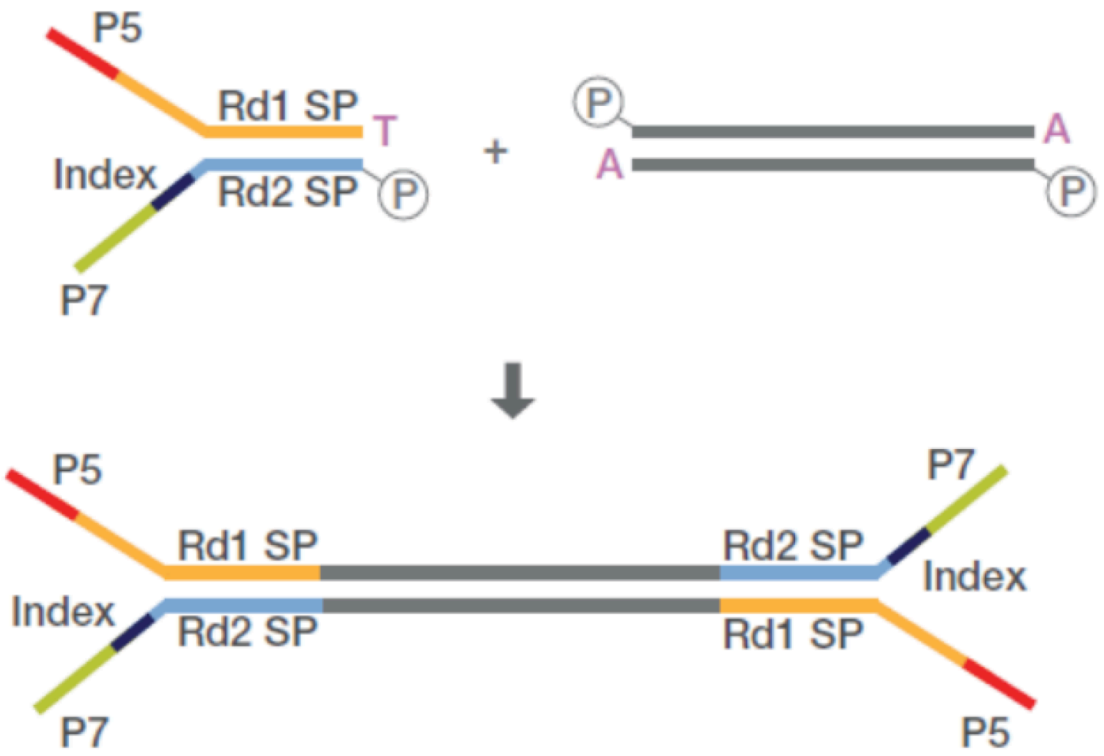
Cluster formation is a type of PCR ("bridge amplification")

- Introduces bias in amplifying high and low G+C fragments
- Reduced coverage at these loci is a result

NGS Library Multiplexing and Molecular Barcoding

- As throughput on NGS sequencers has increased, the ability to pool libraries or reaction products together is needed
- DNA barcoding schemes permit the addition of specific sequences to each library, enabling equimolar pooling (“multiplexing”)
- Post-run de-multiplexing is accomplished by software that bins reads sharing the same barcodes or barcode combinations
- As sequencing costs permit deeper sequencing of libraries, the use of unique molecular identifiers (UMIs) has become necessary to separate true mutations from sequencing “noise” during data analysis

Unique Molecular Identifiers: UMI Barcoding

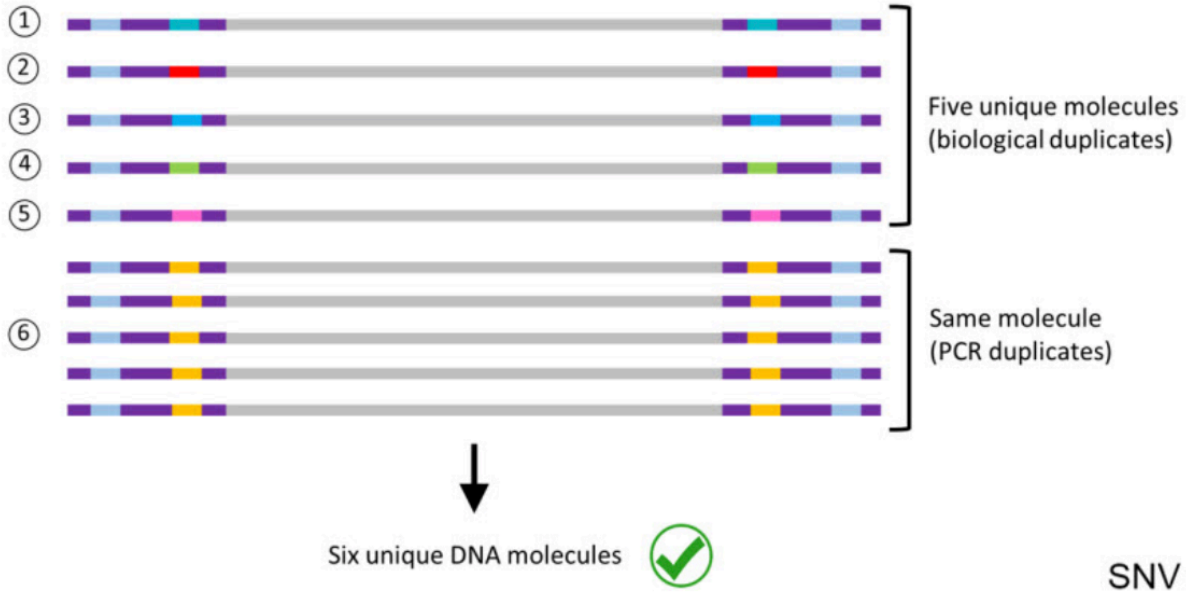


UMI in i7 arm of adapter is added during ligation

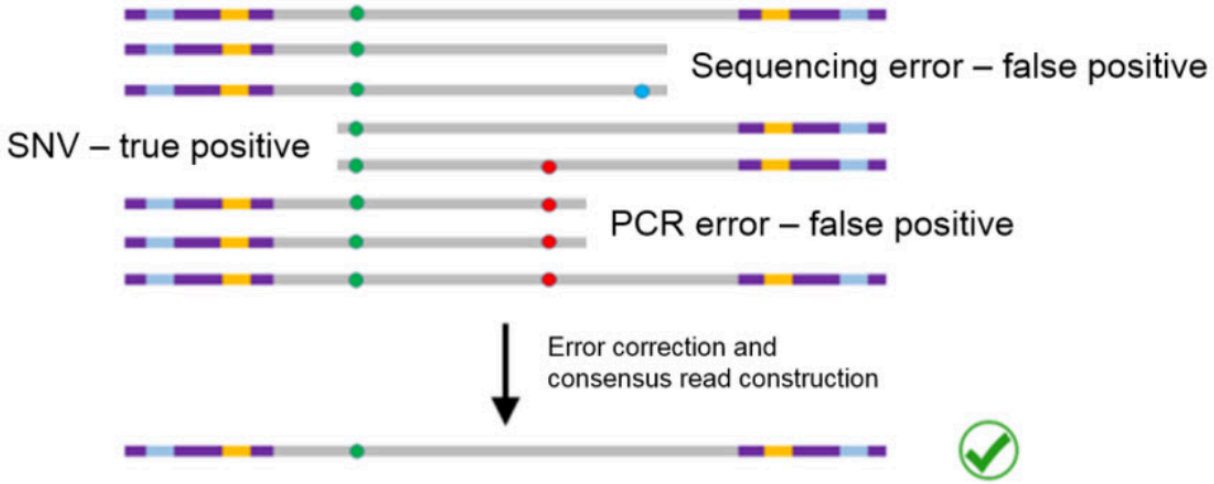
Ligation product with i7 UMI

Dual Functionality of UMIs

De-duplication



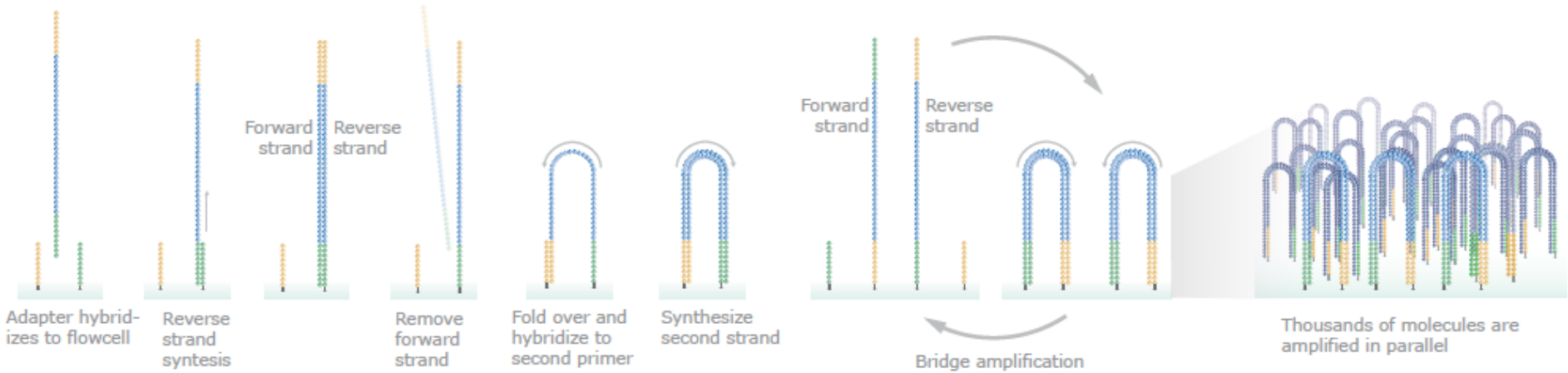
Polymerase or base-calling error elimination



Illumina Sequencing Basics

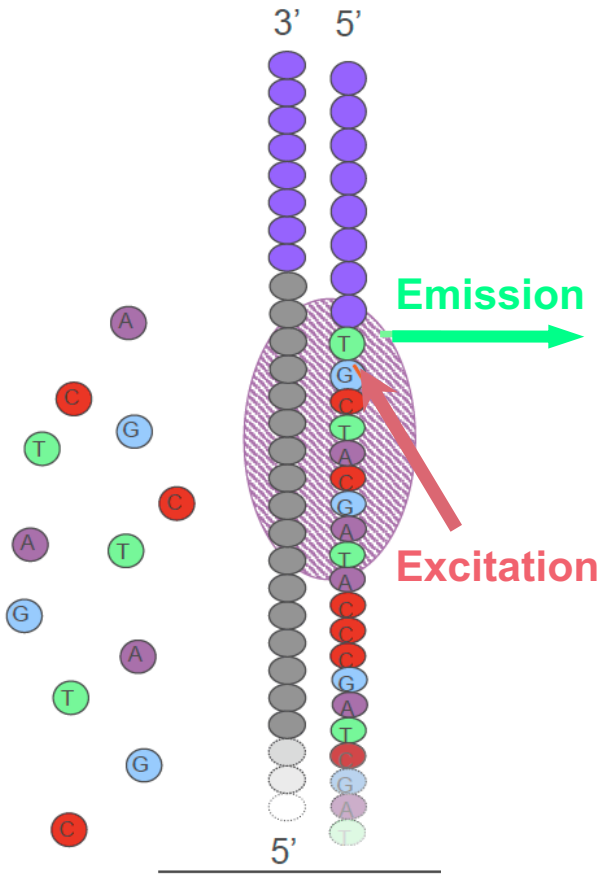
Sequencing by Synthesis

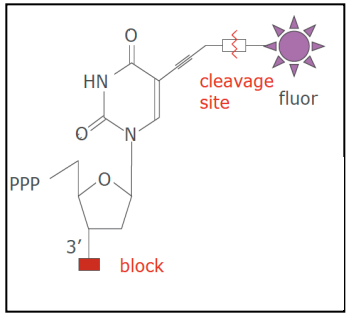
Cluster Amplification on FlowCell



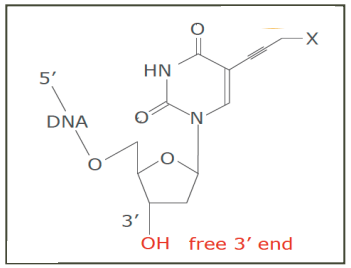
- Quantitating library fragments is an essential first step to cluster formation
- Diluted library fragments are introduced to flow cell, and are amplified *in situ* using the covalently attached complementary adapter sequences
- Cluster amplification is required to produce sufficient signal for detecting the sequencing reaction results at each nucleotide addition step = each cluster is derived from a single library fragment

Sequencing Chemistry for SBS

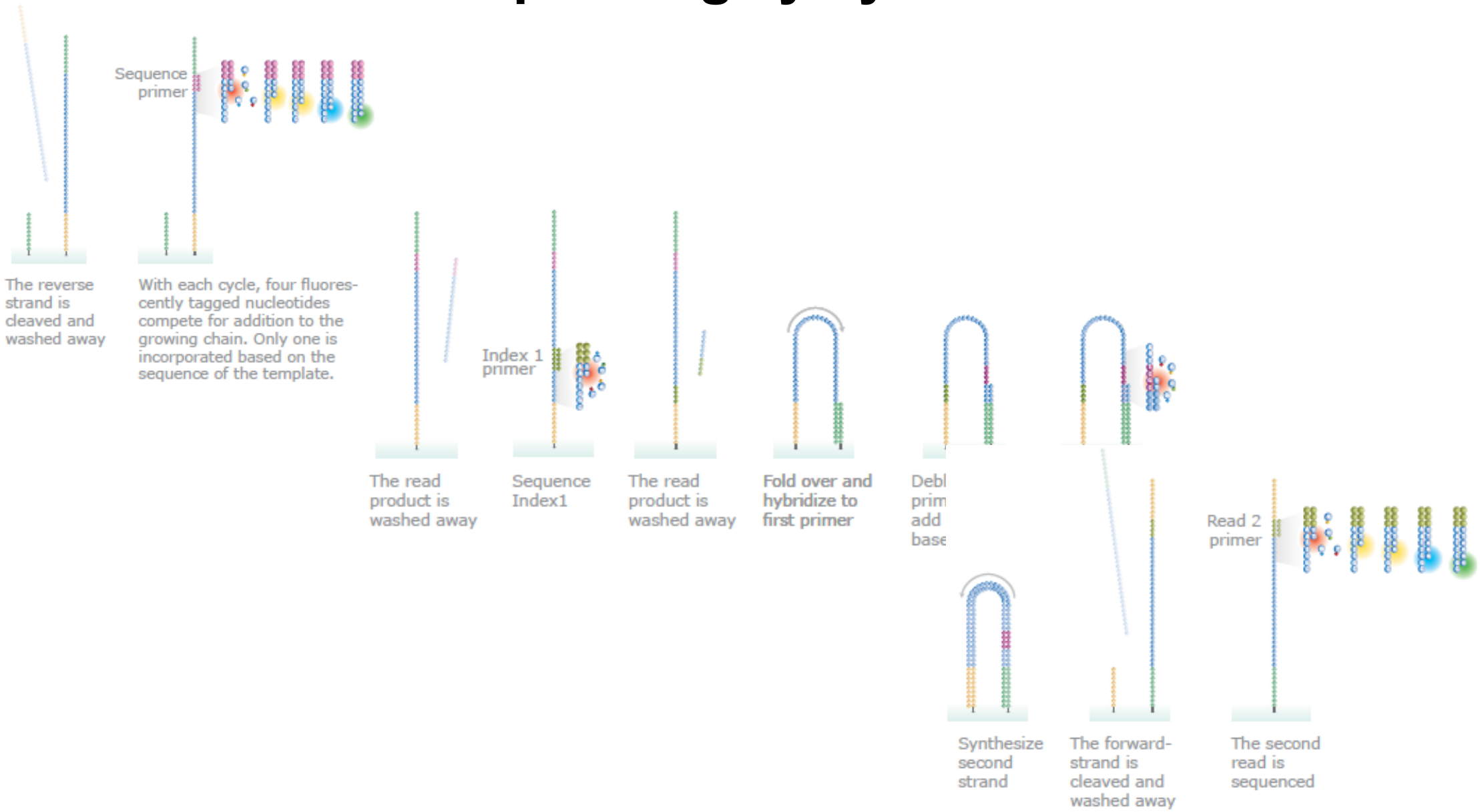




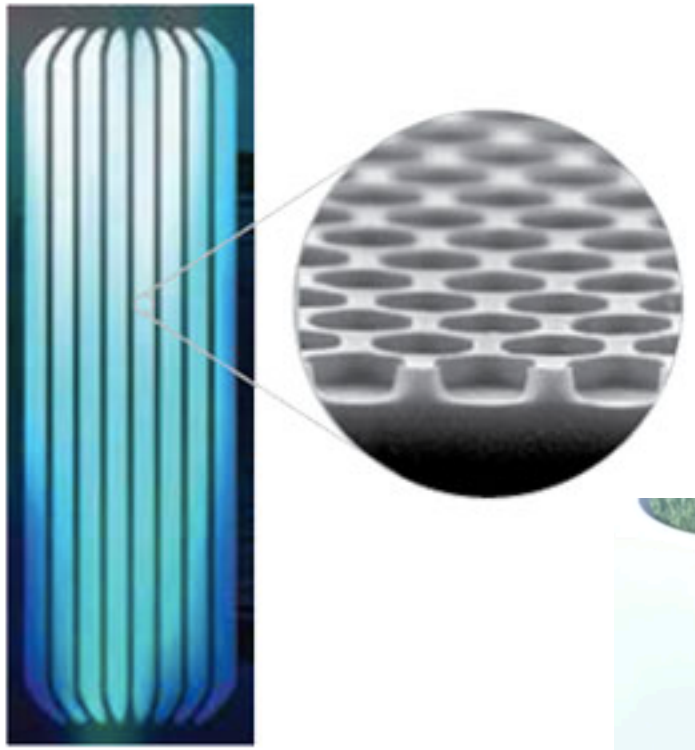
Incorporate
Detect
De-block
Cleave fluor

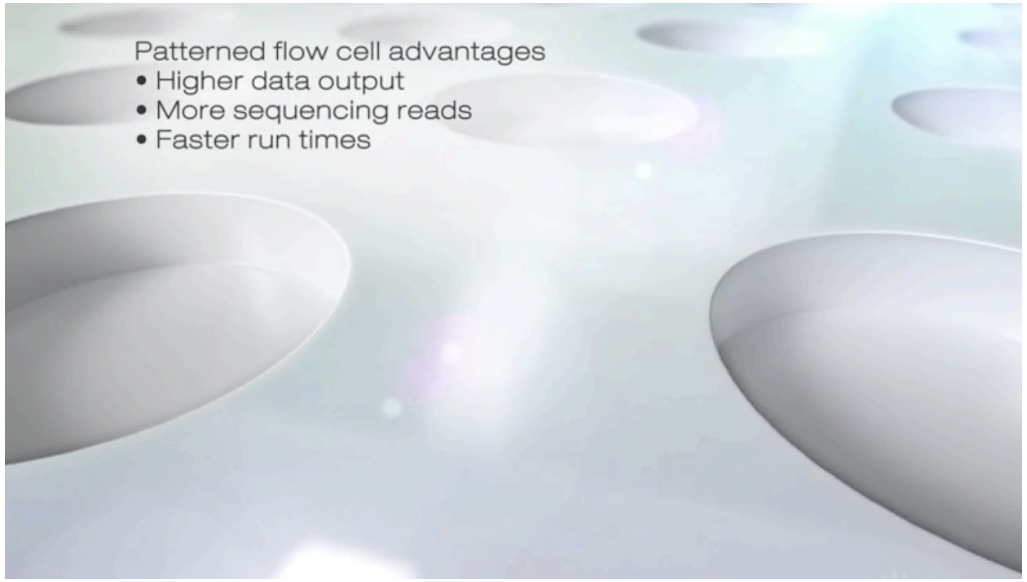


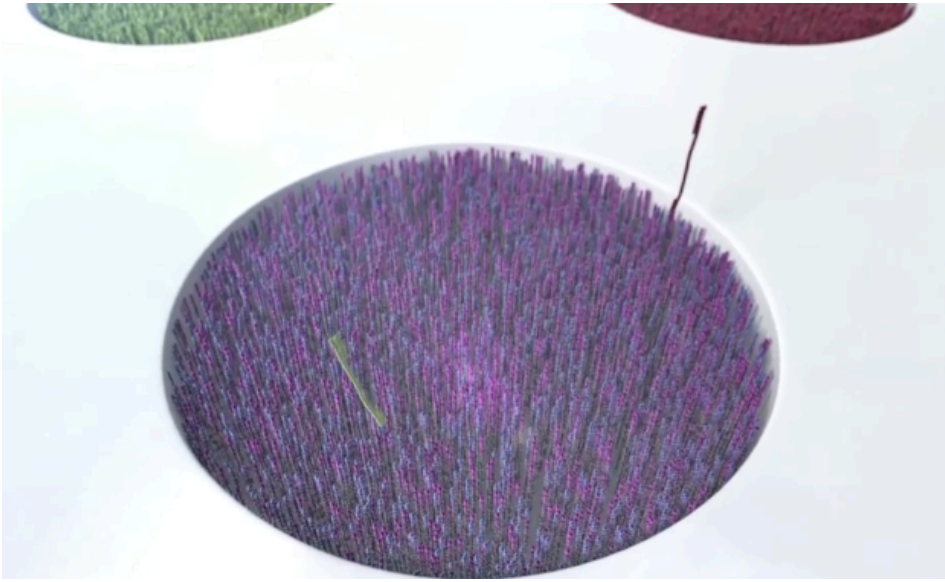
Sequencing by Synthesis



Patterned Flow Cell







Four-, Two- and One-Color Chemistry

The **MiSeq** and **HiSeq** Series Systems currently use four-channel SBS.

The **MiniSeq**, **NextSeq**, and **NovaSeq** Systems use two-channel chemistry.

The **iSeq 100** use one-channel chemistry.

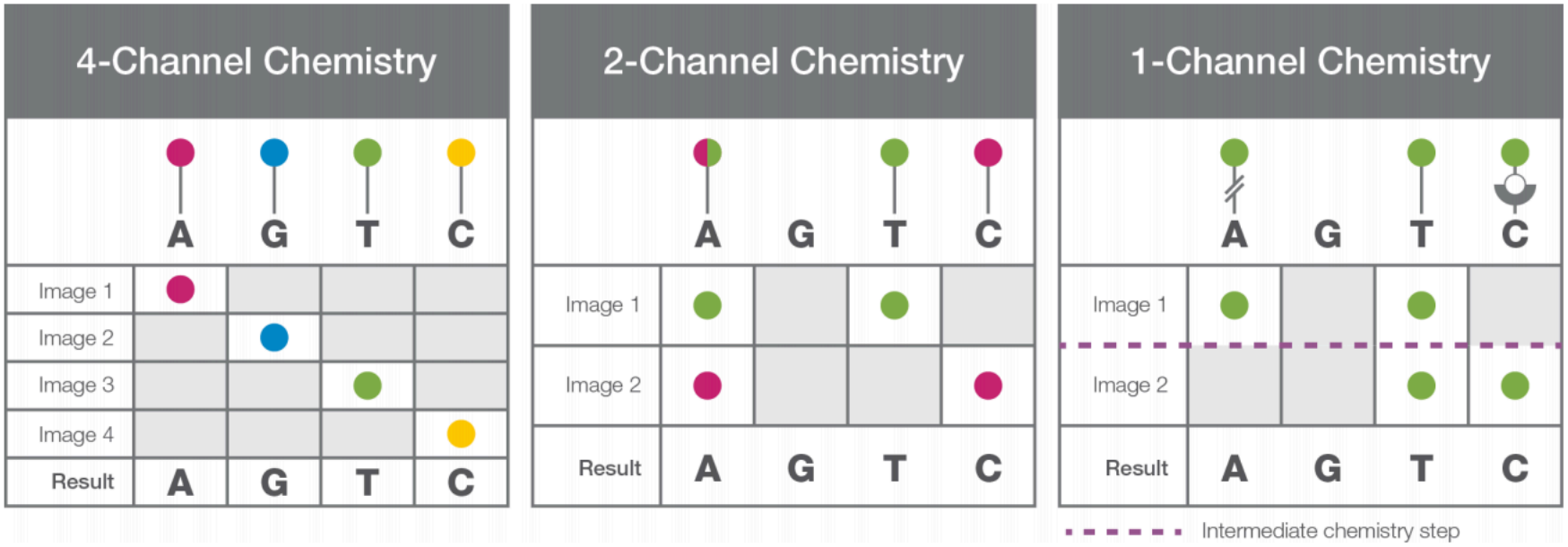


Figure 2: Four-, Two-, and One-Channel Chemistry—Four-channel chemistry uses a mixture of nucleotides labeled with four different fluorescent dyes. Two-channel chemistry uses two different fluorescent dyes, and one-channel chemistry uses only one dye. The images are processed by image analysis software to determine nucleotide identity.

Platforms

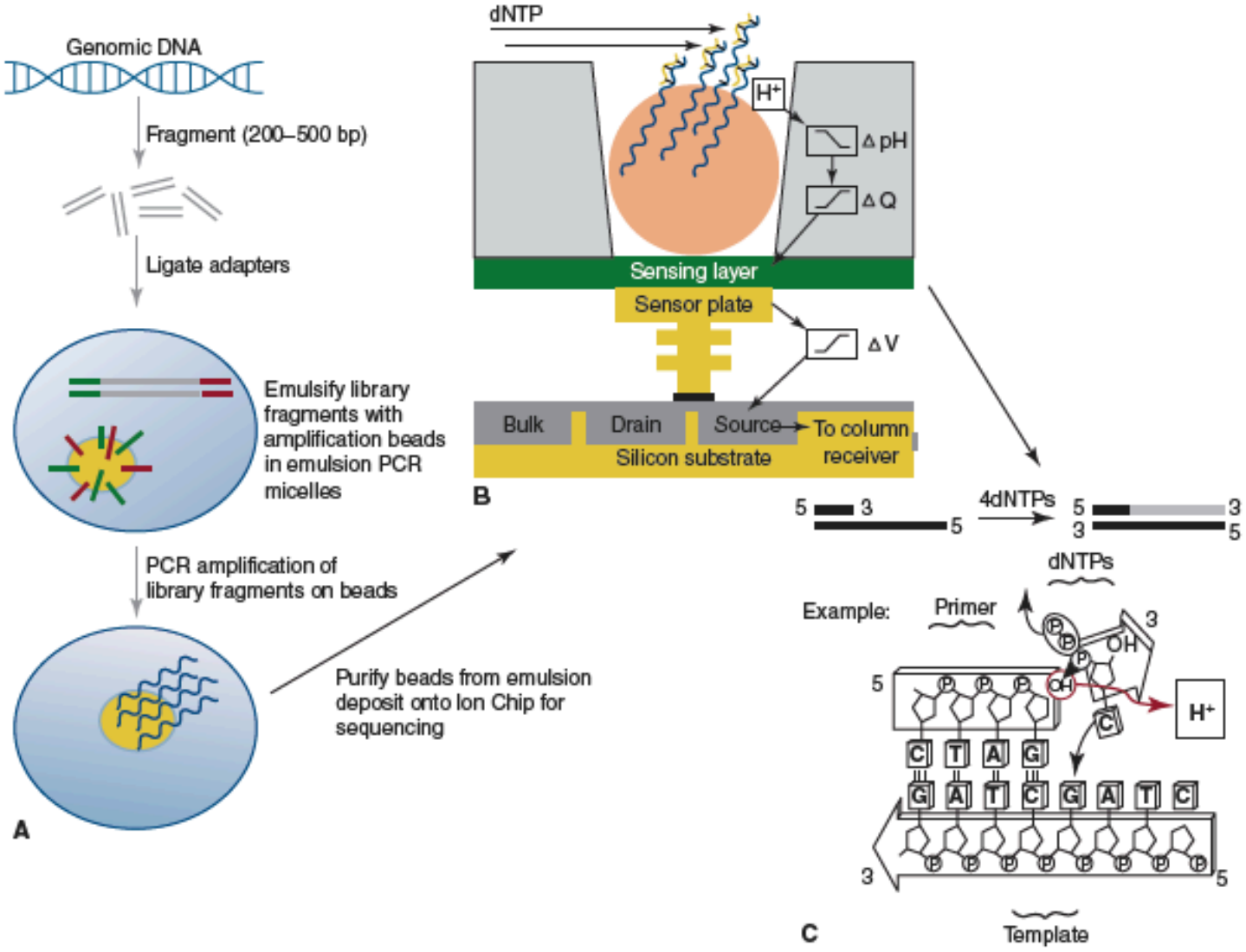
	 NextSeq*	 HiSeq 4000*	 NovaSeq 6000††*	 HiSeq X Ten†
Output Range	20-120 Gb	125-1500 Gb	167-6000 Gb	900-1800 Gb
Run Time	11-29 hr	<1-3.5 days	19-40 hr	< 3 days
Reads per Run	130-400 million	2.5-5 billion	1.4-20 billion	3-6 billion
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp
Samples per Run†	1	6-12	4-48	8-16
Relative Price per Sample†	Lower Cost	Lower Cost	Lower Cost	Lower Cost
Relative Instrument Price†	Higher Cost	Higher Cost	Higher Cost	Higher Cost

- High accuracy reads, paired end reads, range of capacity and throughput
- Longer read lengths on some platforms (MiSeq)
- Improved kits, improved software pipeline and capabilities, cloud computing in BaseSpace

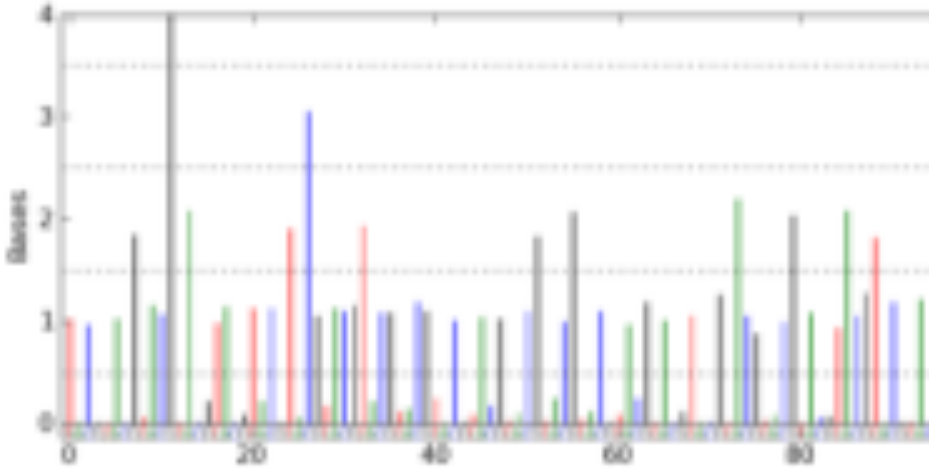
Ion Torrent Sequencing Technology

Sequencing by pH Sensing

pH Sensing of Base Incorporation

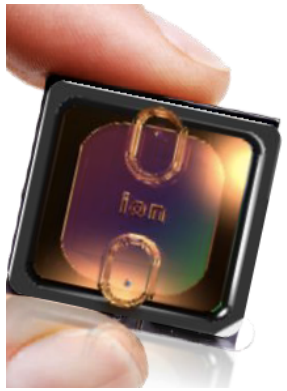


- On-bead fragment amplification by emulsion PCR
- Unlabeled native dNTPs
- Undefined read lengths

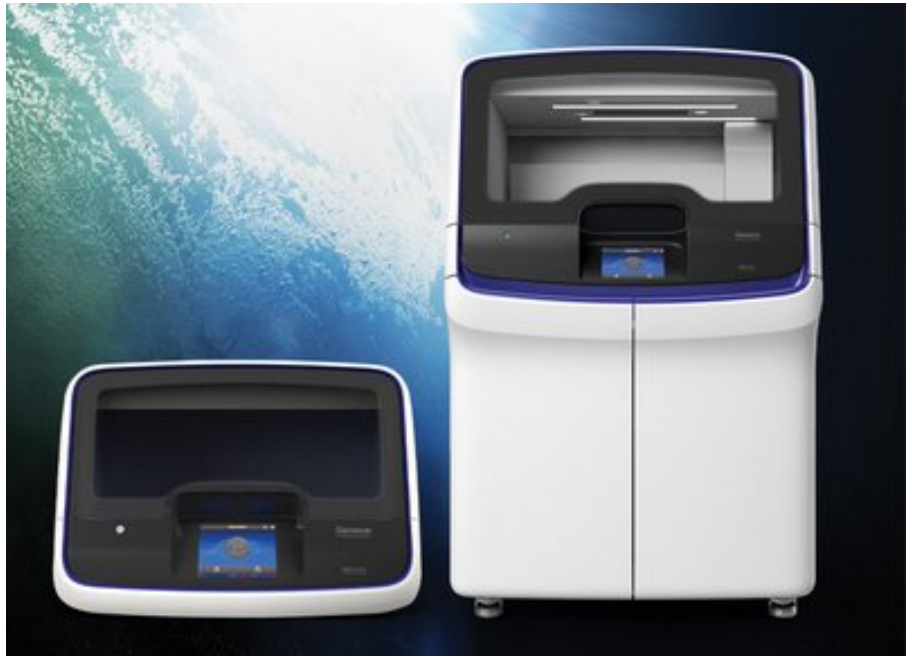


Platforms

Ion Torrent S5



Ion Chef



IonTorrent Genexus

- Low substitution error rate, in/dels problematic, no paired end reads
- Inexpensive and fast turn-around for data production
- Improved computational workflows for analysis

Sub-setting the Genome for NGS-based Assays

Exome and Gene Panels

Direct Genomic Selection

TECHNICAL REPORTS

nature genetics

BRIEF COMMUNICATIONS

Direct selection of human genomic loci by microarray hybridization

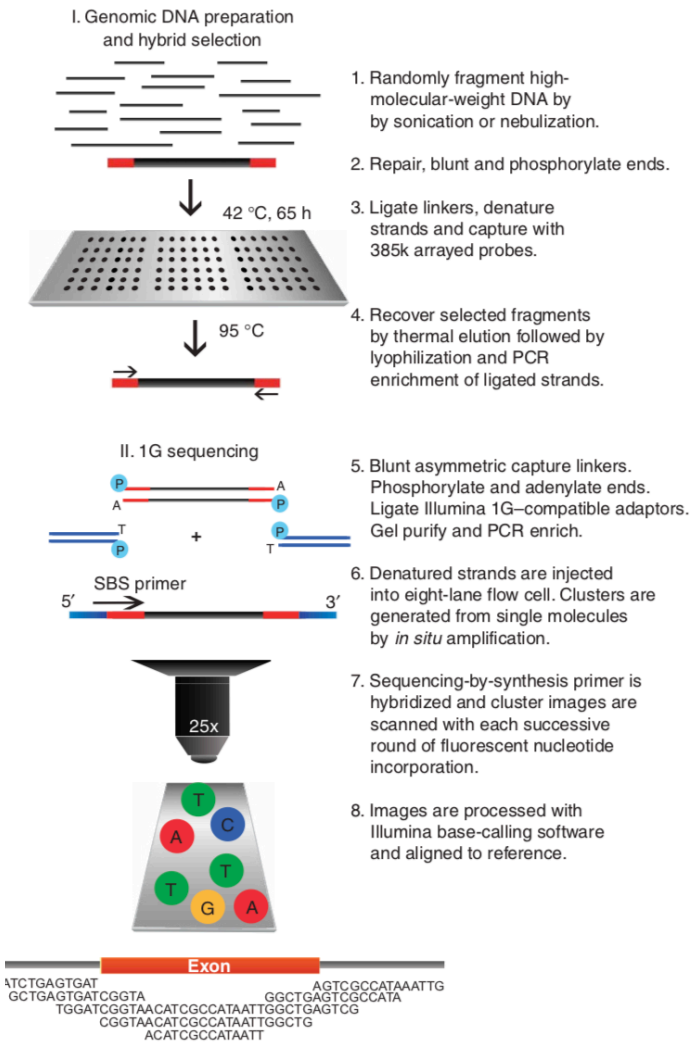
Thomas J Albert[1], Michael N Molla[1], Donna M Muzny[2], Lynne Nazareth[2], David Wheeler[2], Xingzhi Song[2], Todd A Richmond[1], Chris M Middle[1], Matthew J Rodesch[1], Charles J Packard[1], George M Weinstock[2] & Richard A Gibbs[2]

http://www.nature.com/naturemethods

We applied high-density microarrays to the enrichment of specific sequences from the human genome for high-throughput

corresponding to the targeted region[5]. The remaining sequences represent genomic repeats or other distant sequences, showing that in addition to the burden of manipulating the large BACs, the presence of repeat sequences in the selecting DNA compromized the final enrichment.

Here we report the design of custom high-density oligonucleotide microarrays (NimbleGen) to capture both dispersed short genome segments, encompassing individual gene exons, and single long segments, corresponding to entire gene loci. The 'exonic' design aimed to capture DNA representing 6,726 genomic regions (**Supplementary Table 1** online; minimum size 500 bases, ~5 Mb of total sequence) from 660 genes dispersed throughout the genome, and a second capture array series targeted areas of 200 kb, 500 kb, 1 Mb, 2 Mb and 5 Mb surrounding the human *BRCA1* gene locus. Each of these microarrays was designed with long oligonucleotide probes (>60 bases) spaced on average between



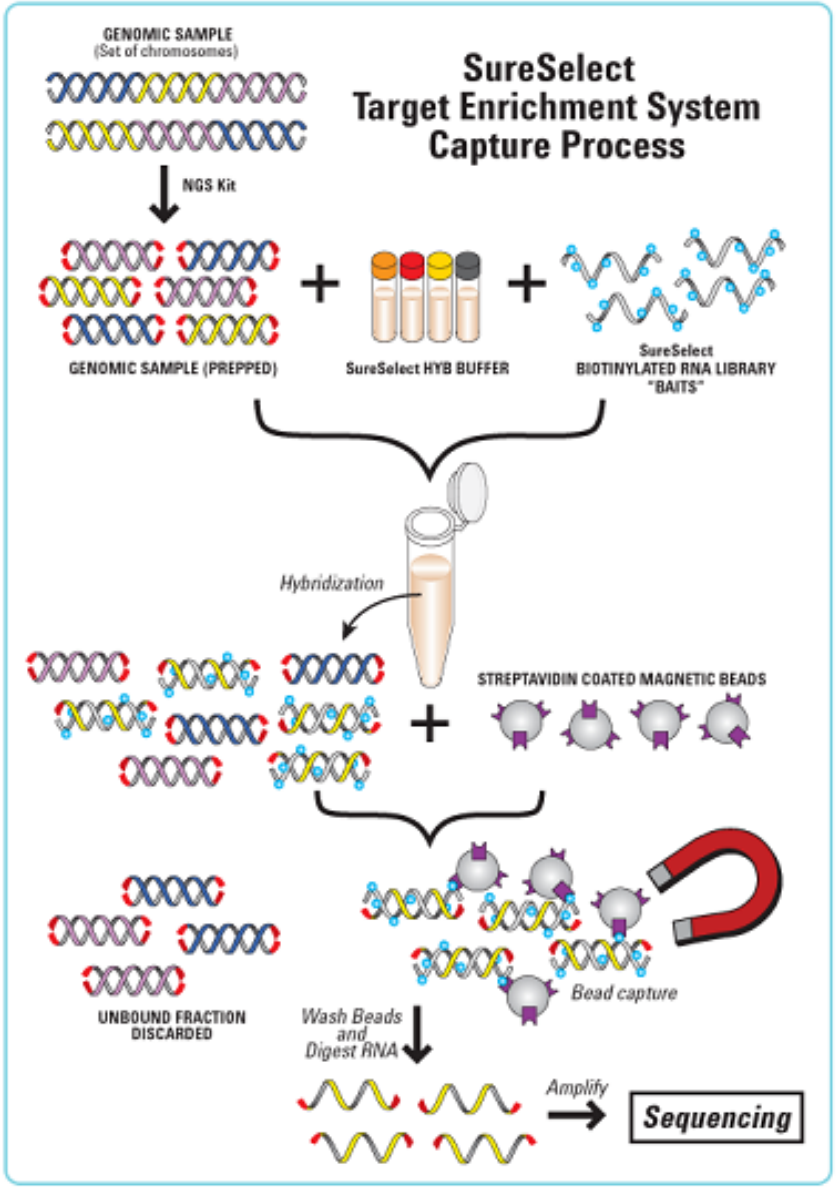
Solution-Phase Hybrid Capture

<u>Hybrid capture</u> - specific sequences from a whole genome library are selected by hybridization to probes that correspond to human exons or gene panels.

Probe DNAs are biotinylated, so hybrids can be removed from solution with streptavidin magnetic beads.

An "**exome**" by definition, is captured with probes corresponding to the exons of all genes annotated in the reference genome.

Custom capture reagents or "gene panels" can be synthesized to target specific loci that may be of clinical interest.



Hybrid Capture Terminology

Probe/Bait – Biotin-linked oligonucleotides (RNA or DNA) designed to specific genomic loci (e.g., Panels/Exomes).

Target/Pond – A fragmented DNA representation of the genome which has been prepared for NGS sequencing (platform agnostic).

Capture – The process of target enrichment by nucleic acid hybridization between Probes (RNA/DNA) and Library.

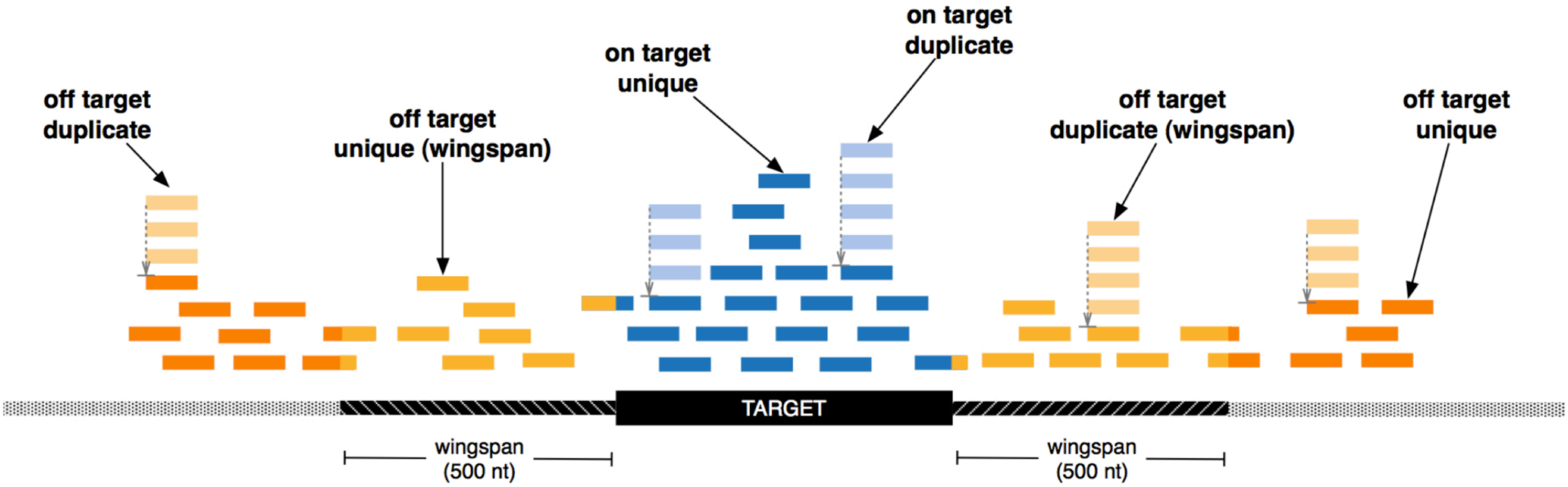
Blocking – The addition of repetitive DNA (Cot1/SS) and library-specific adapter oligonucleotides to minimize off-target capture (daisy-chaining).

Coverage – Breadth and Depth requirements to accurately call genomic variation at high specificity and sensitivity.

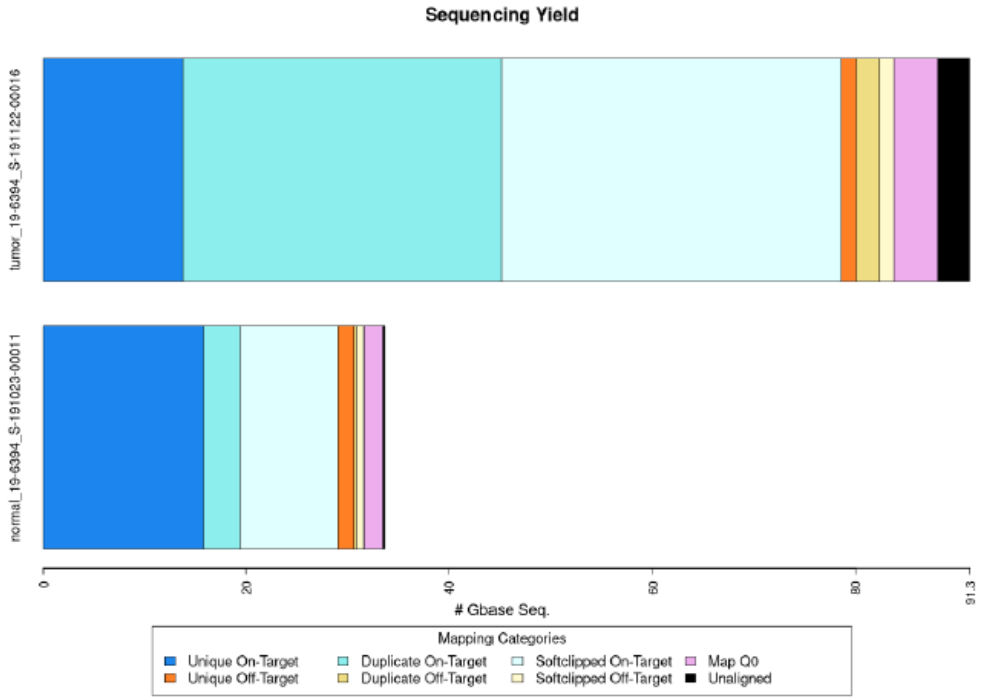
Sensitivity - Frequency

Specificity – Target

Coverage Definitions for Hybrid Capture

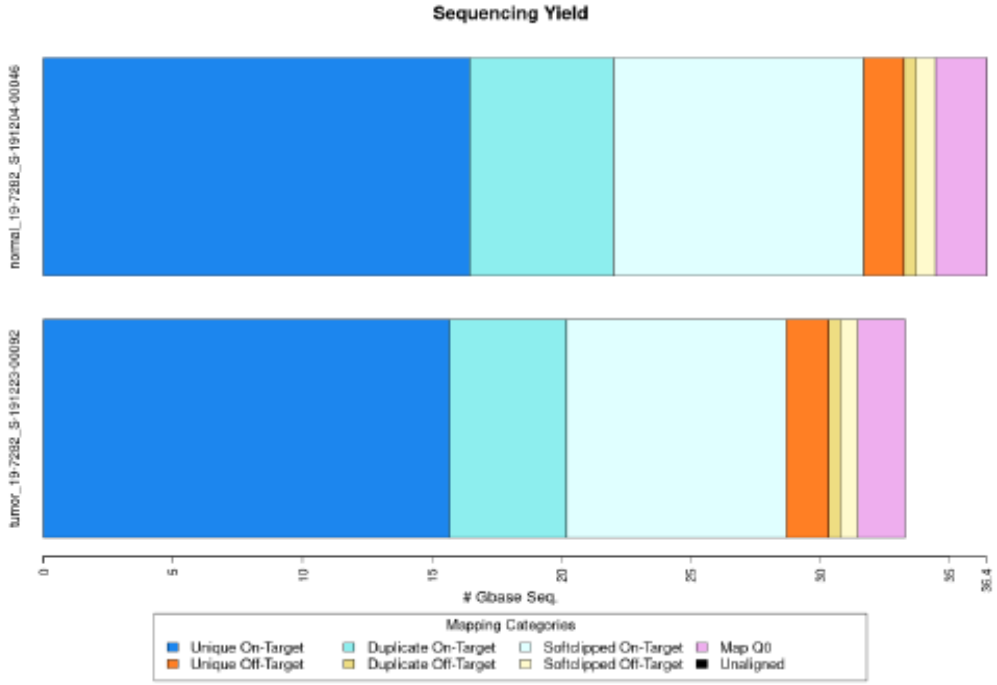


Evaluating Coverage Post-Alignment and De-dup



	normal_19-6394_S-191023-00011	tumor_19-6394_S-191122-00016
Unique On-Target	15.8 (46.9%)	13.8 (15.1%)
Duplicate On-Target	3.59 (10.7%)	31.29 (34.3%)
Softclipped On-Target	9.7 (28.8%)	33.5 (36.7%)
Unique Off-Target	1.59 (4.72%)	1.50 (1.64%)
Duplicate Off-Target	0.27 (0.801%)	2.31 (2.53%)
Softclipped Off-Target	0.703 (2.09%)	1.408 (1.54%)
MapQ0	1.86 (5.52%)	4.31 (4.72%)
Unaligned	0.214 (0.635%)	3.173 (3.48%)
Total	33.7	91.3

Table 2: Sequence allocation across mapping categories (gigabases)



	normal_19-7282_S-191204-00046	tumor_19-7282_S-191223-00092
Unique On-Target	16.5 (45.2%)	15.7 (47.1%)
Duplicate On-Target	5.54 (15.2%)	4.51 (13.5%)
Softclipped On-Target	9.68 (26.5%)	8.50 (25.5%)
Unique Off-Target	1.54 (4.22%)	1.63 (4.89%)
Duplicate Off-Target	0.486 (1.33%)	0.462 (1.39%)
Softclipped Off-Target	0.738 (2.02%)	0.666 (2.00%)
MapQ0	1.98 (5.42%)	1.83 (5.50%)
Unaligned	0.000196 (0.000537%)	0.000244 (0.000733%)
Total	36.5	33.3

Table 2: Sequence allocation across mapping categories (gigabases)

Evaluating Probe Coverage Breadth at Fixed Depth

Breadth at 20x Depth

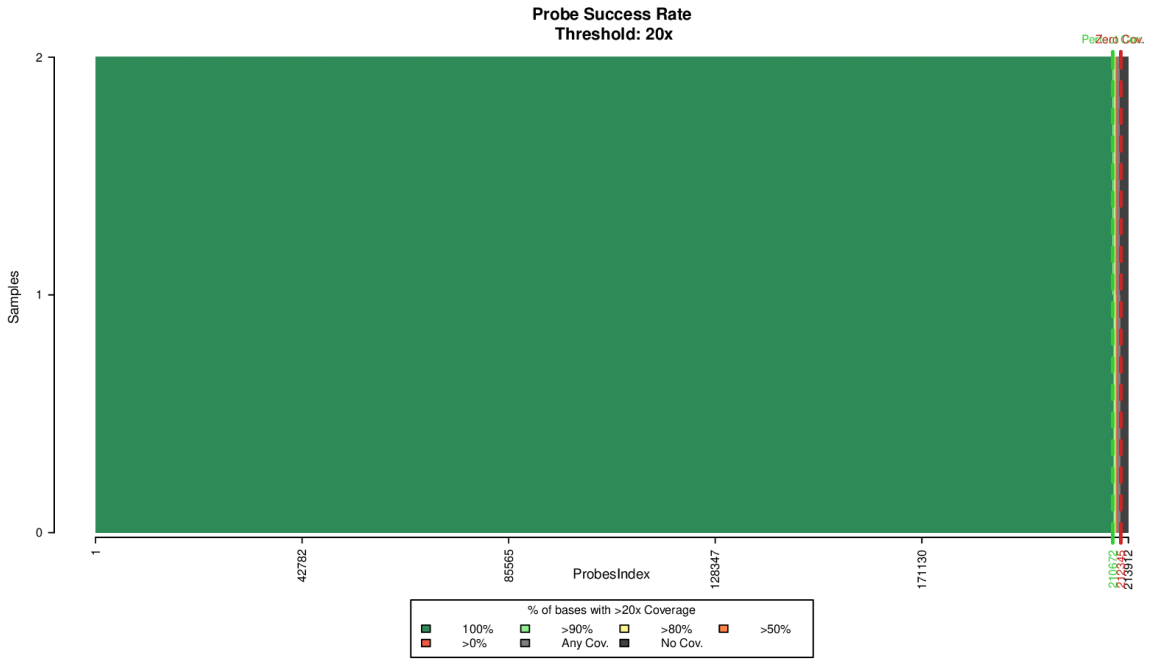


Figure 6: Probe success rate, expressed as number of samples reaching 20x coverage depth

Total Probes: 213,914

Perfect-Coverage Probes: 210,672

Zero-Coverage Probes: 1,568

Breadth at 100x Depth

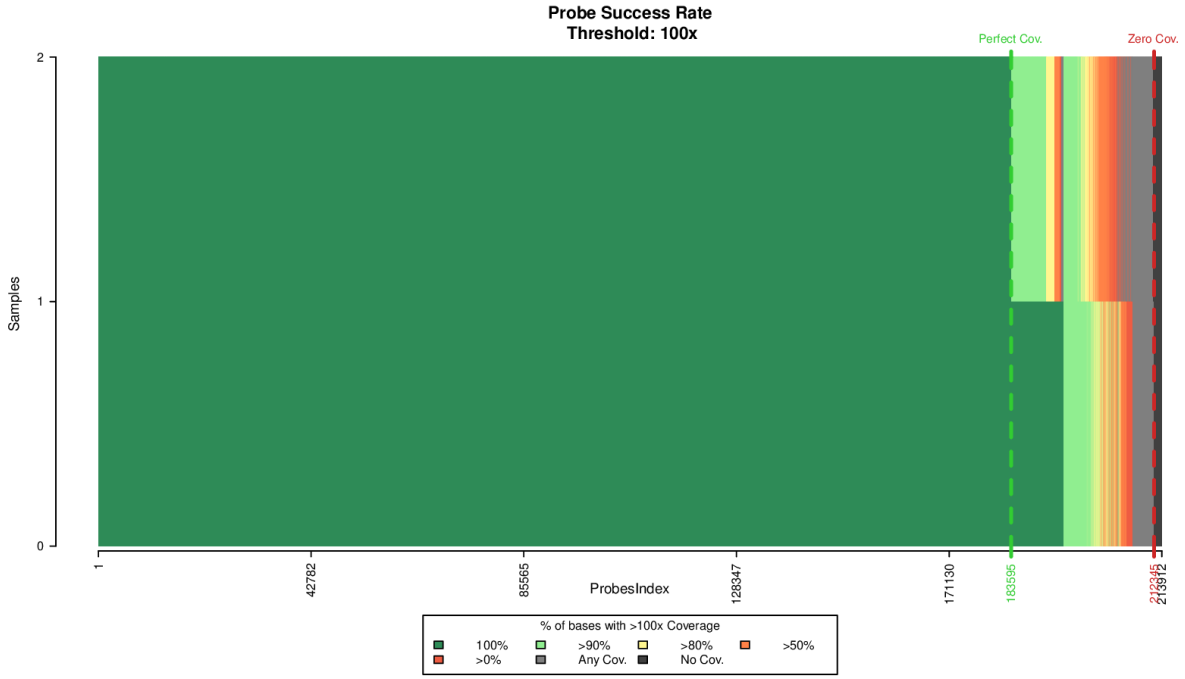


Figure 7: Probes success rate, expressed as number of samples reaching targeted coverage depth (100x)

Total Probes Regions: 213,914

Perfect-Coverage Probes: 183,595

Zero-Coverage Probes: 1,568

Probe Coverage and Outlier Analysis

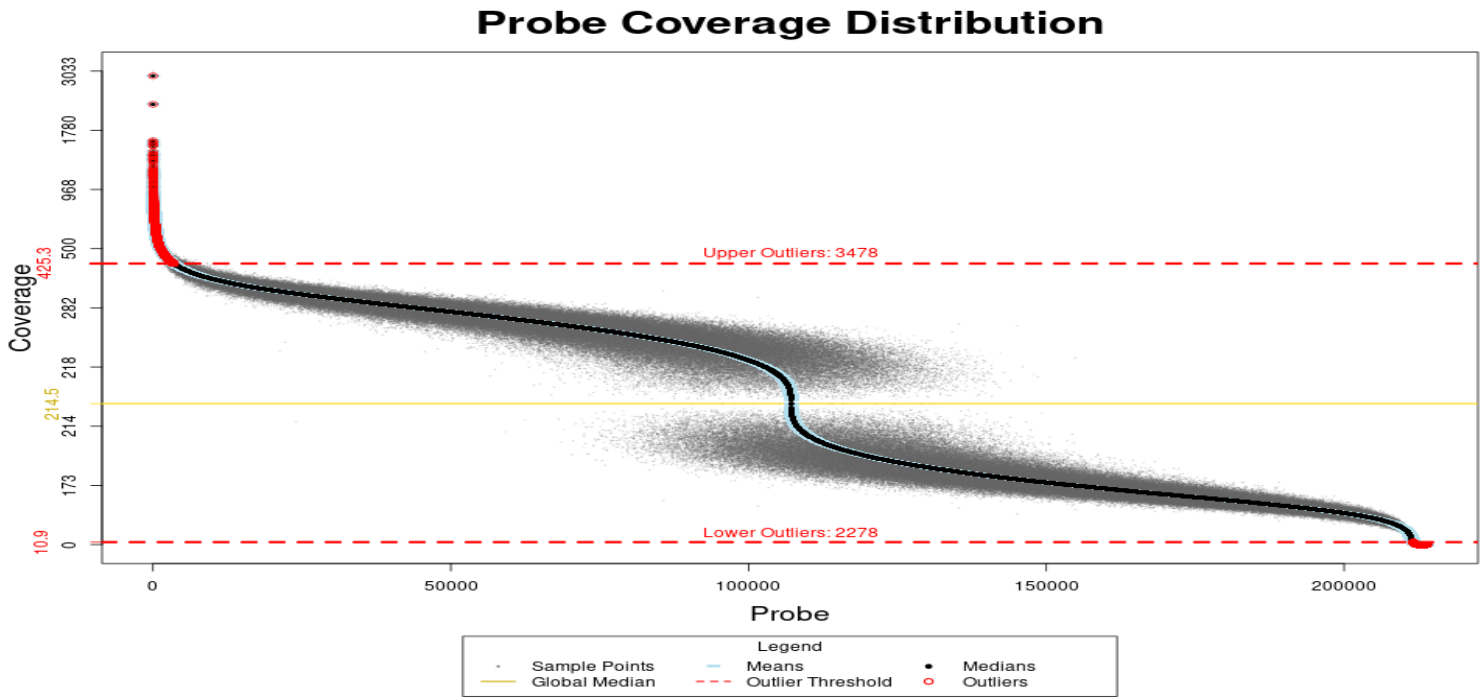


Figure 8: Curve of region performance, showing coverage rate of each and highlighting those that significantly over- or under-perform (median +/- 1.5*IQR)

Mean Coverage: 221.8

Median Coverage: 214.5

High Cutoff: 425.3

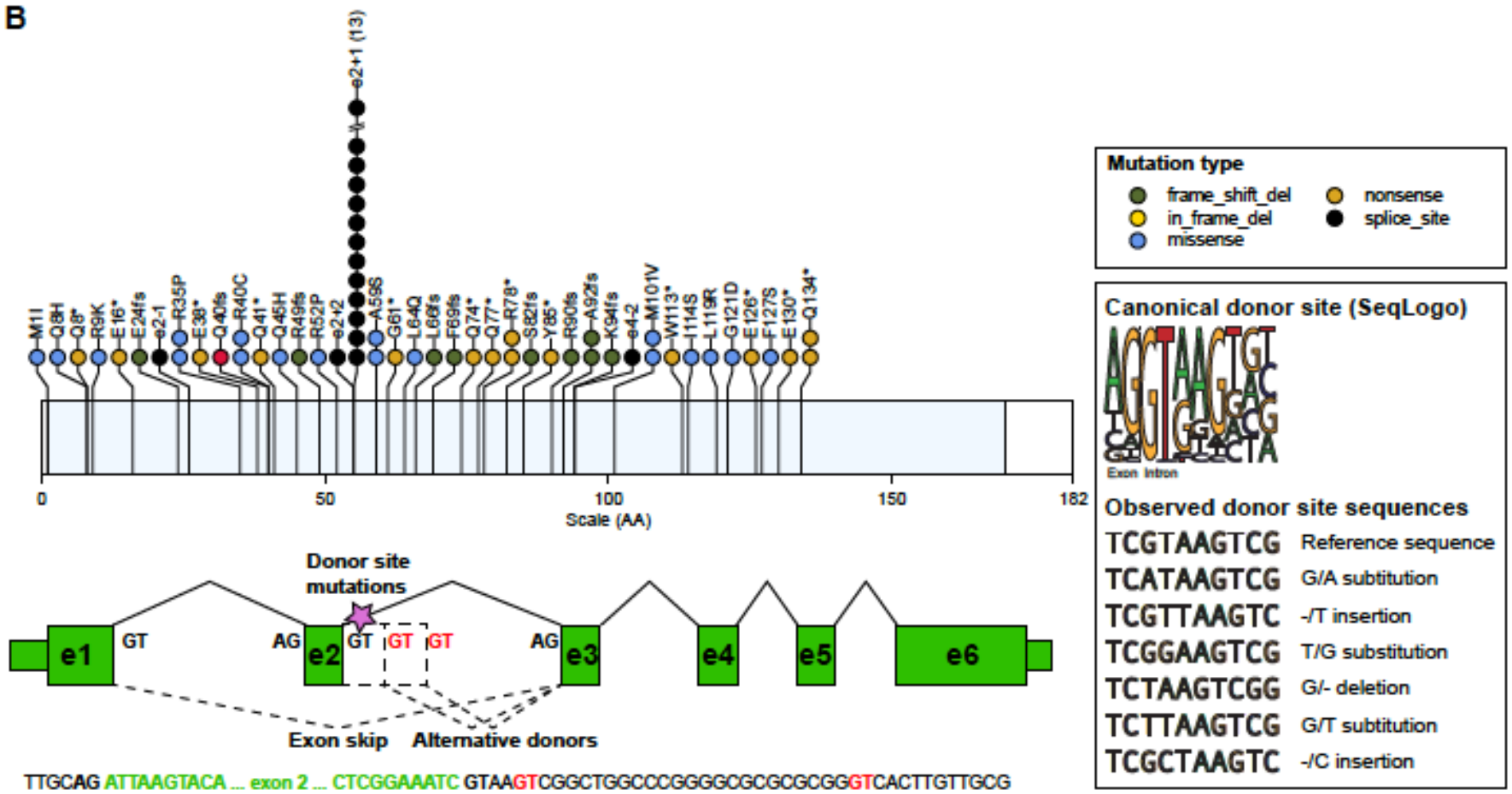
Low Cutoff: 10.9

Total Outliers: 5,755

High Outliers: 3,478

Low Outliers: 2,278

Hybrid Probe Placement Matters: Missed Splice Site Mutation



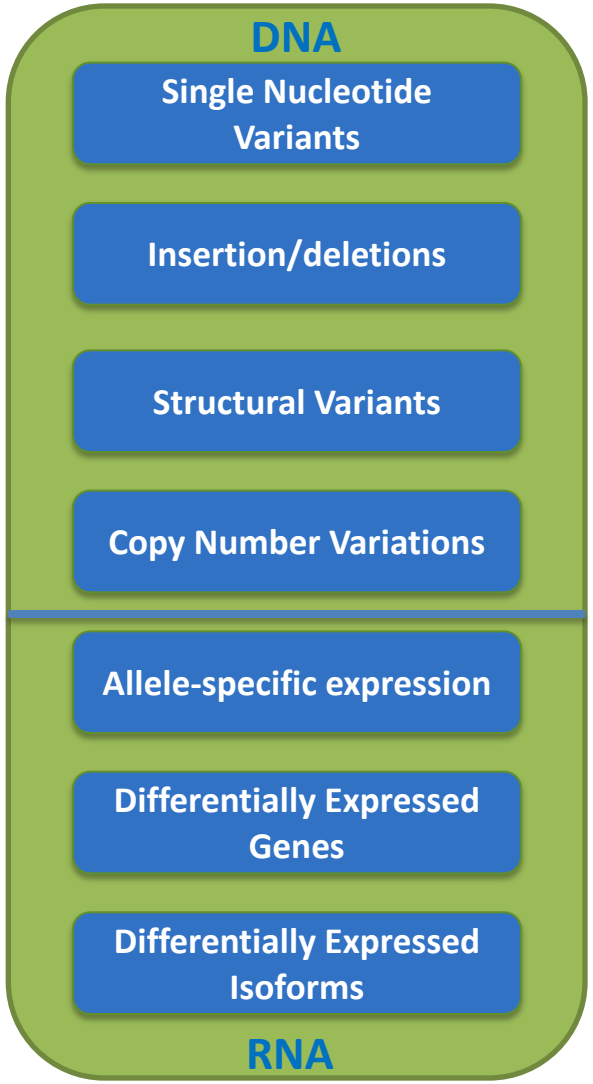
O.L. Griffith et al., Nat. Commun. 2018

Post Data Generation Analyses

Computational approaches for NGS read data

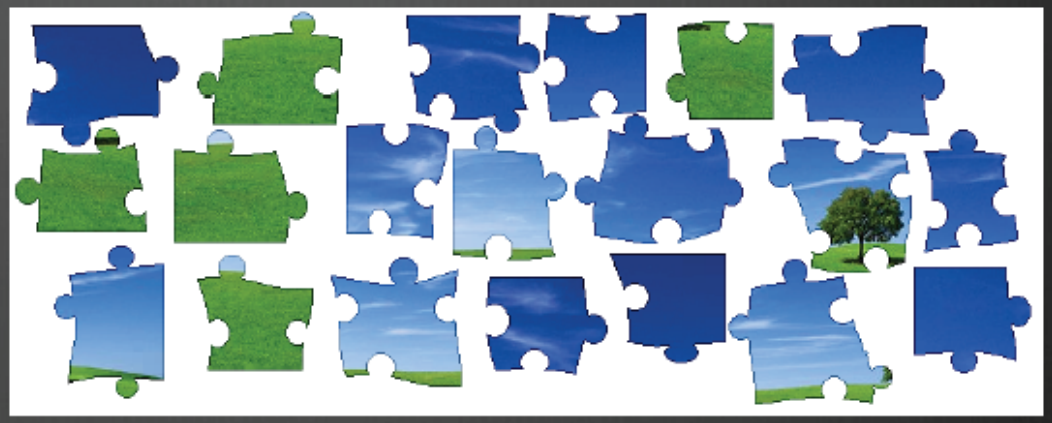
The Human Genome enables NGS Analysis

- The human genome reference sequence is the keystone for identifying variation in NGS sequencing read data
- **Alignment** of NGS reads to the human reference sequence is the first step to identify variation of all types: we align the reads, then identify variants in comparison to the human reference genome
- By overlaying the locations of genes onto the sequence variants identified, we can interpret the changes to the encoded protein(s)
- Functions of some proteins are known, and the impact of specific variants is sometimes understood, but this knowledge is incomplete at best



Short Read Alignment…

Is like a jigsaw puzzle…



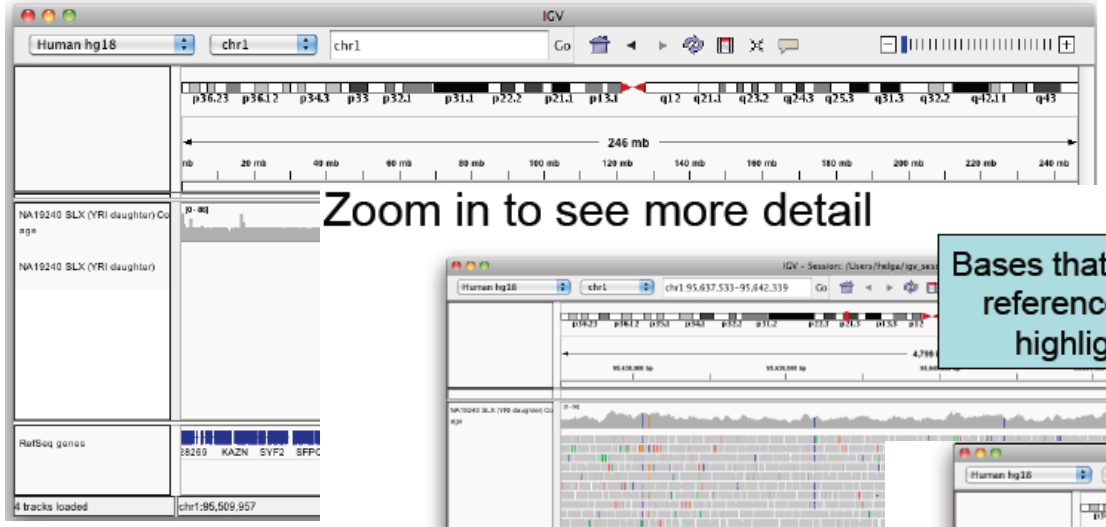
…where they give you the cover on the box

Data visualization

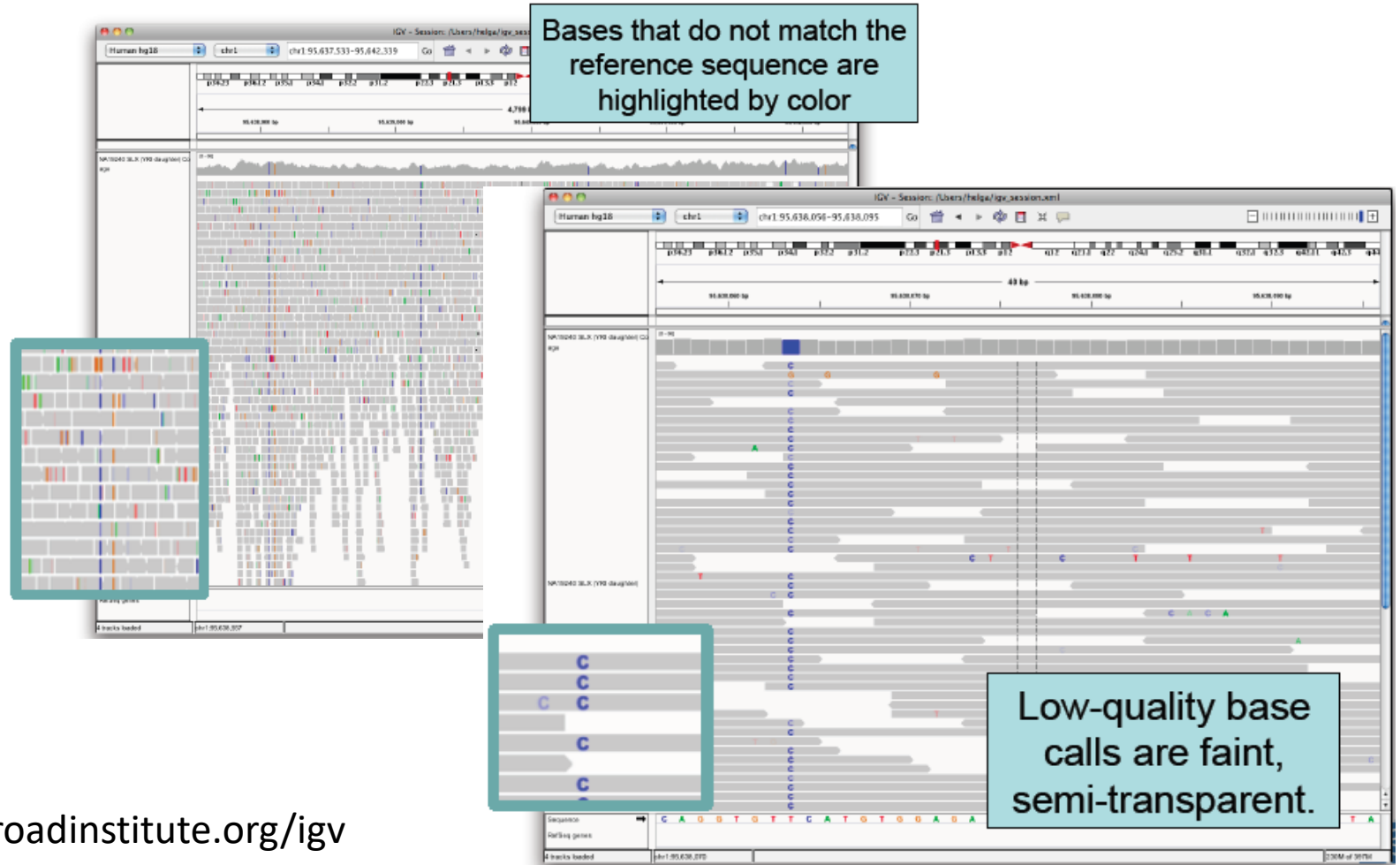
Examining NGS data

Integrated Genomics Viewer (IGV)

Whole chromosome view

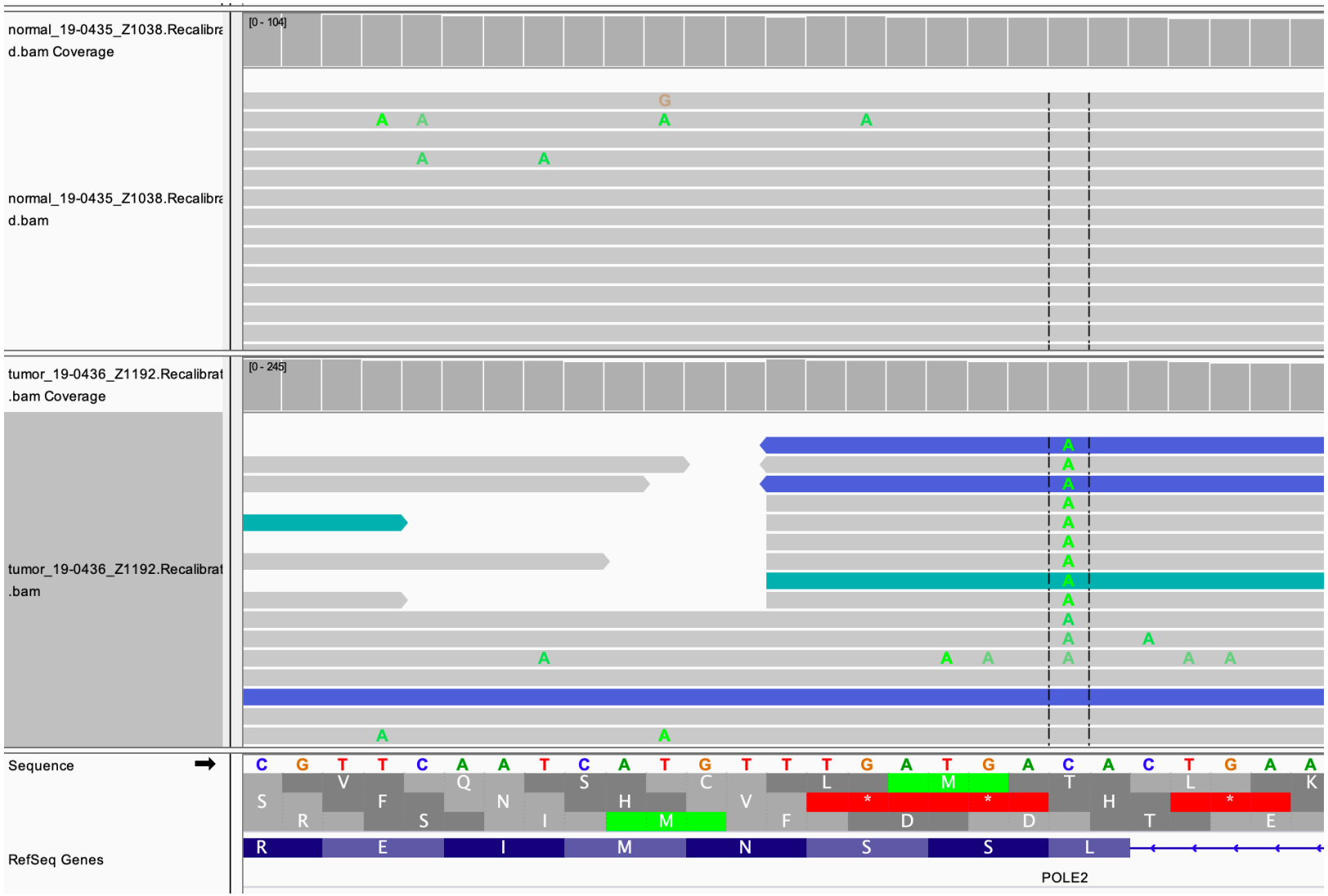


Zoom in to see more detail

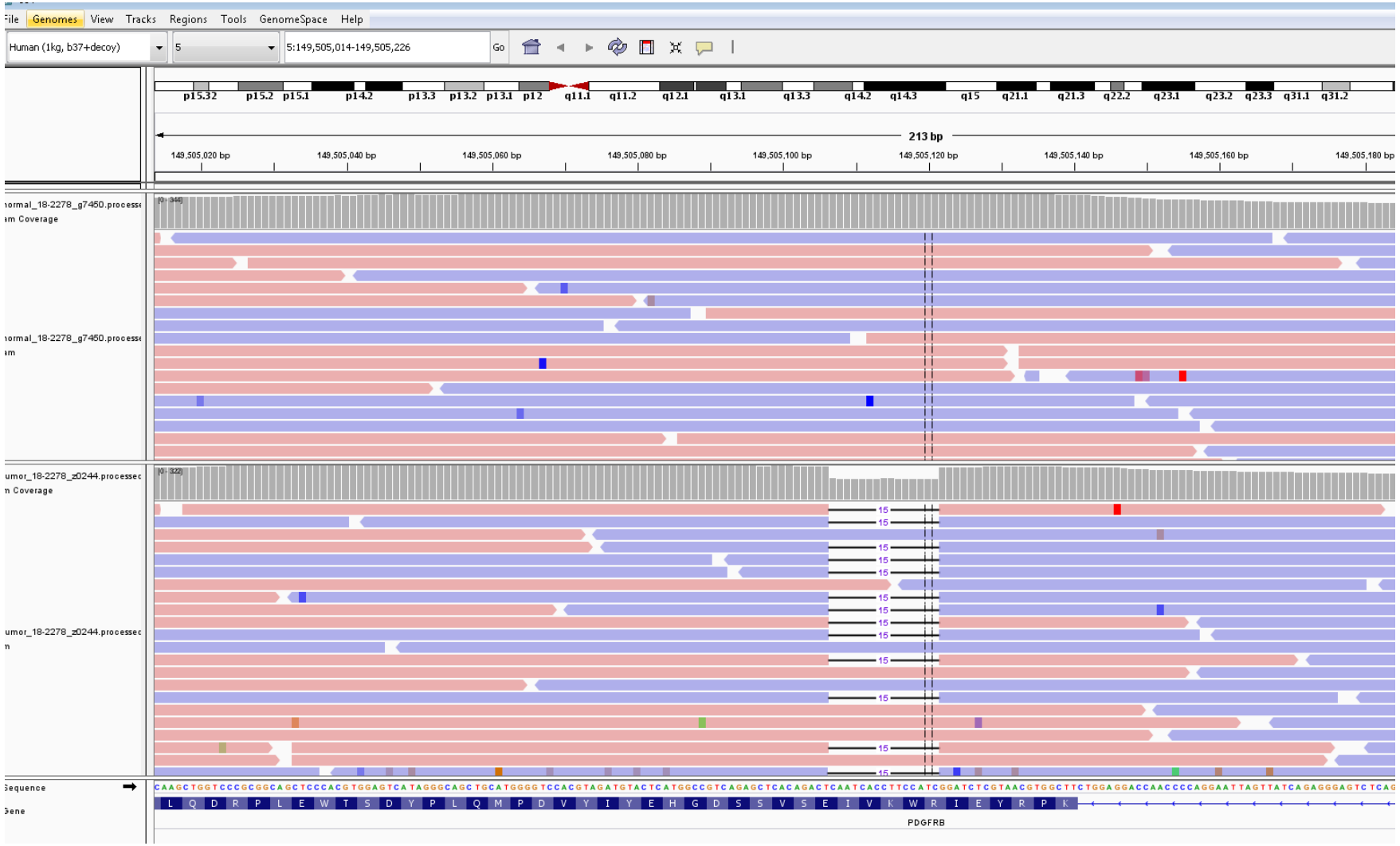


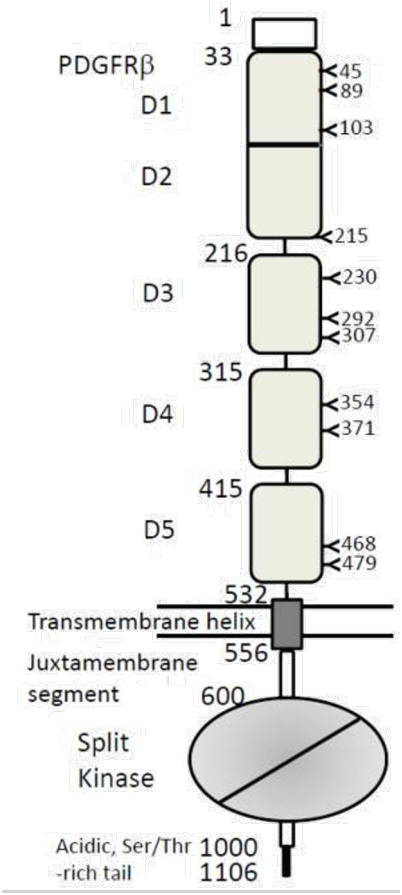
http://www.broadinstitute.org/igv

IGV Visualization of Tumor vs. Normal : Somatic SNV

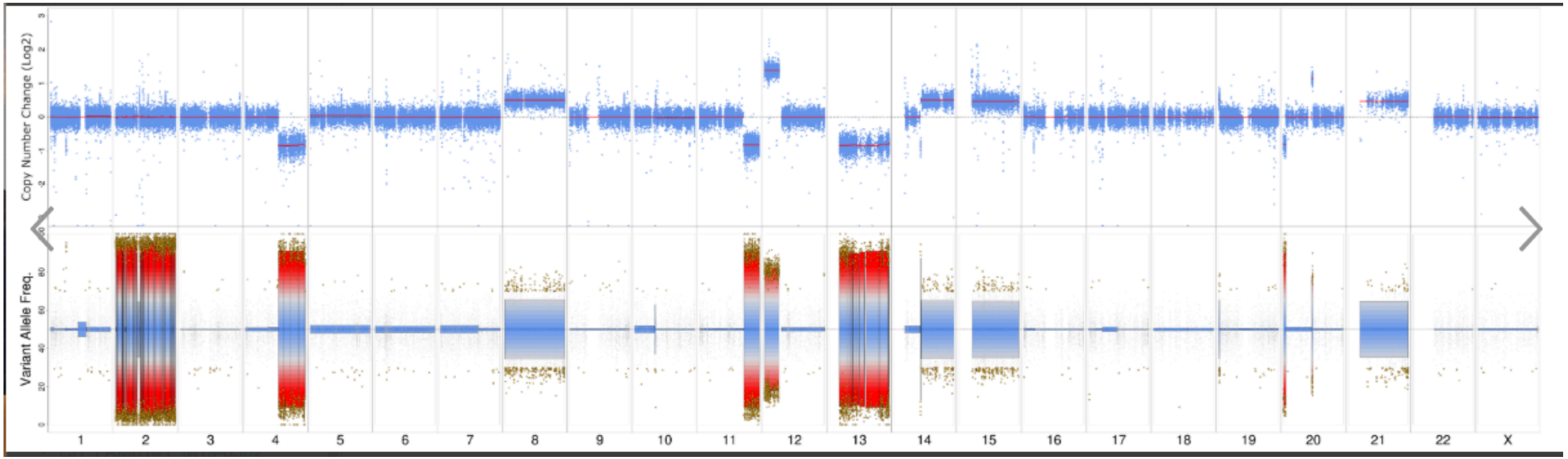


IGV: Somatic Insertion/Deletion

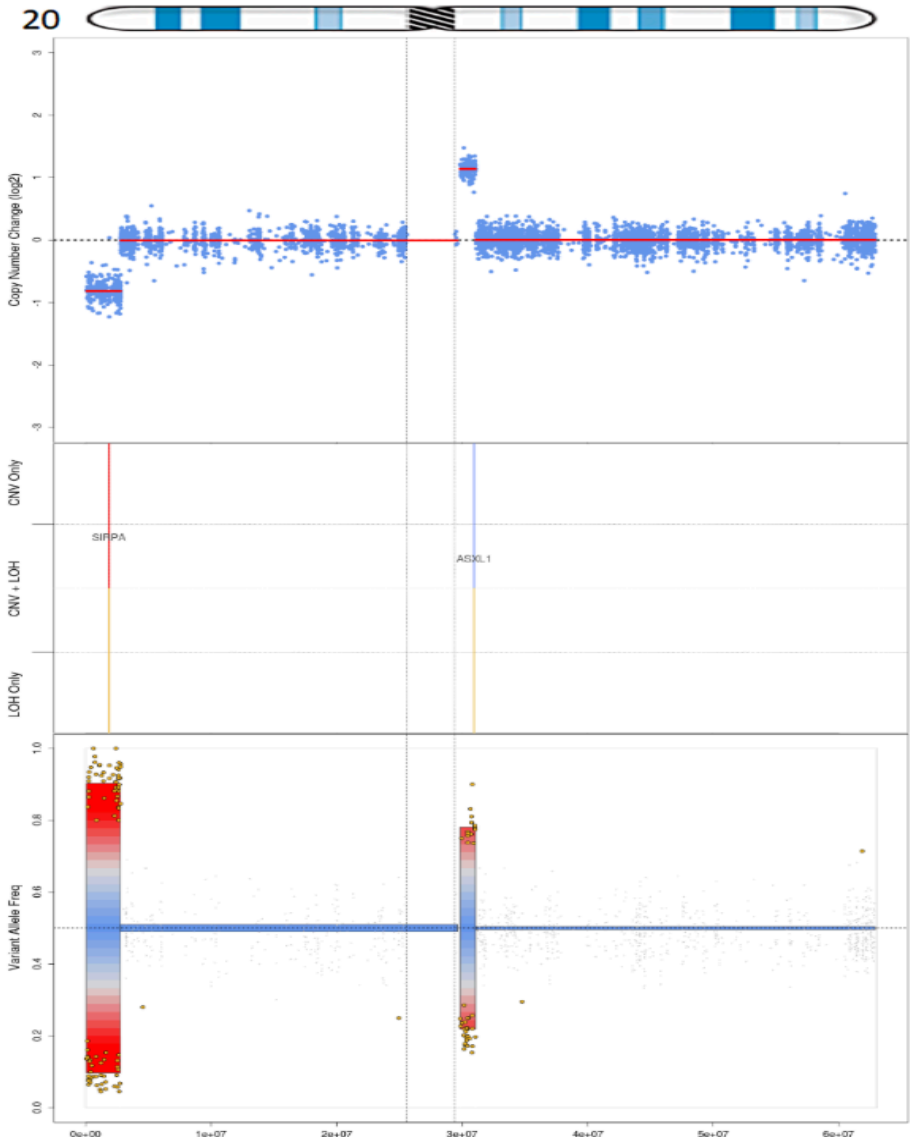


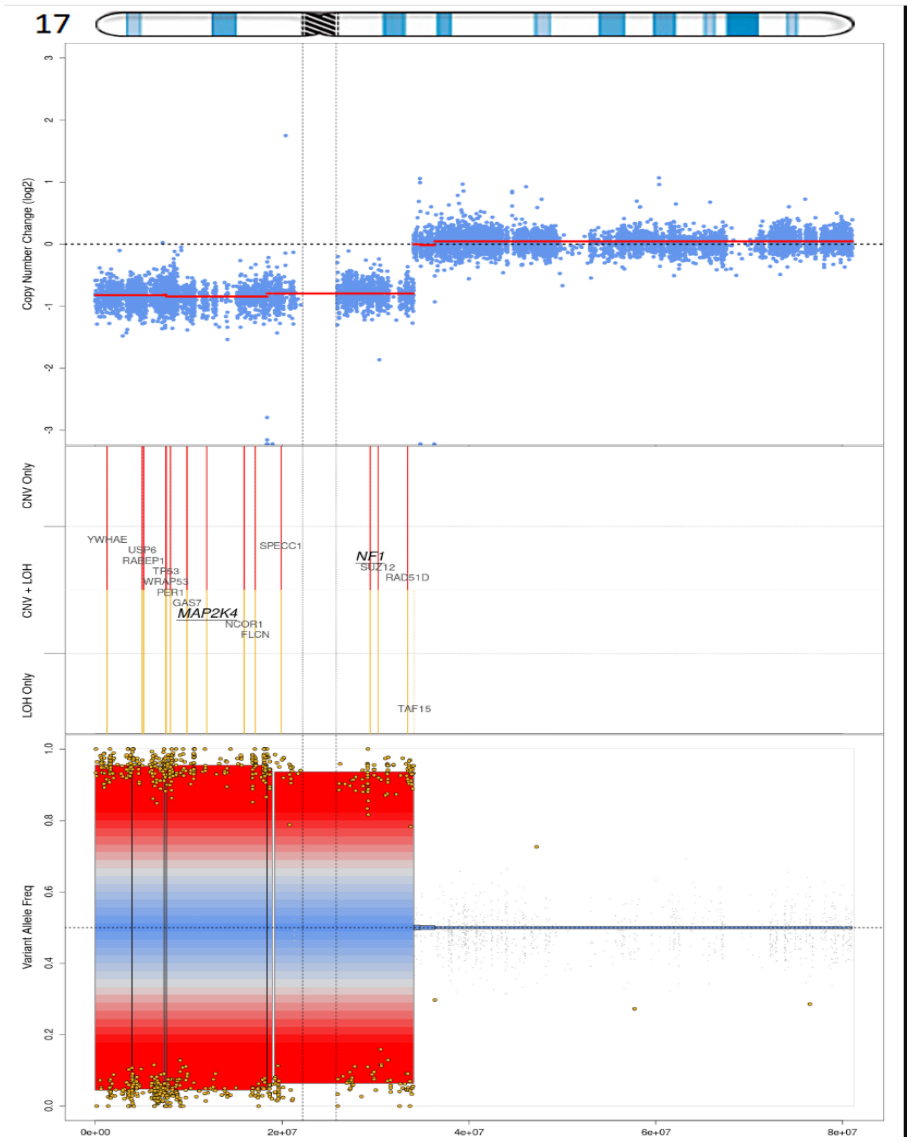


Copy Number Alterations Genome-wide



Copy Number Alterations: Arm-level and Focal





RNAseq

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.℠

THE OHIO STATE UNIVERSITY
COLLEGE OF MEDICINE

RNAseq: Why

- Increasing emphasis on sequencing the transcriptome, combined with multiple types of computational analysis
- Fusion detection, exon skipping, allele-specific silencing, epigenetic links to driver gene amplification, immune deconvolution (in cancer)
- Multiple sizes of RNAs are being identified and linked causally to disease and other mechanisms of gene regulation (miRNAs, lncRNAs)
- Identifying differential gene expression, alternative splicing, RNA editing
- Providing a gene set for organisms without a reference genome

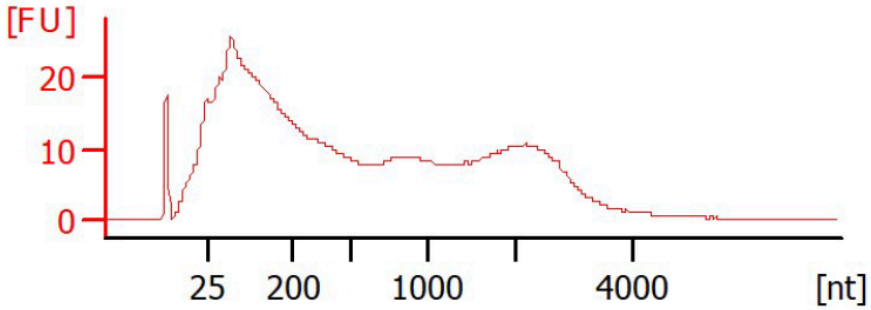
RNAseq "Gotchas"

- RNA is sensitive to degradation from heat and FFPE treatment
- Genomic DNA can co-isolate with RNA and cause multiple problems
- Isolating mRNA decreases the total amount of RNA into library prep
- Strandedness is important
- Short reads make detecting certain RNA-specific attributes or aberrations difficult (exon skipping, etc.)

QC of RNA is Essential

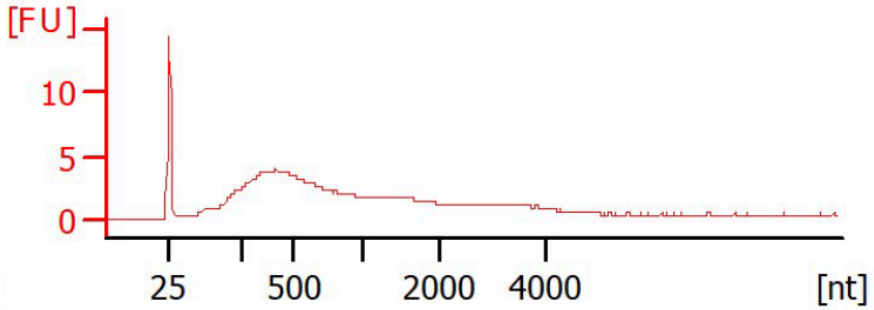
Exp5_B-15_6266_Core Punch

RIN: 3.10 ; DV200 50-70%



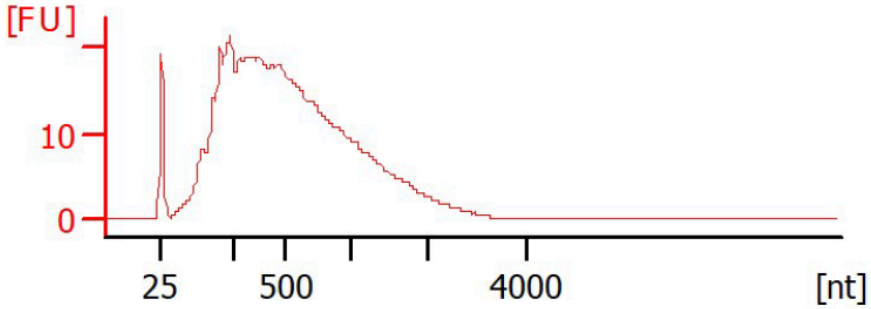
Exp5_B-16_12632_Core Punch

RIN: 2.20 ; DV200 > 70%



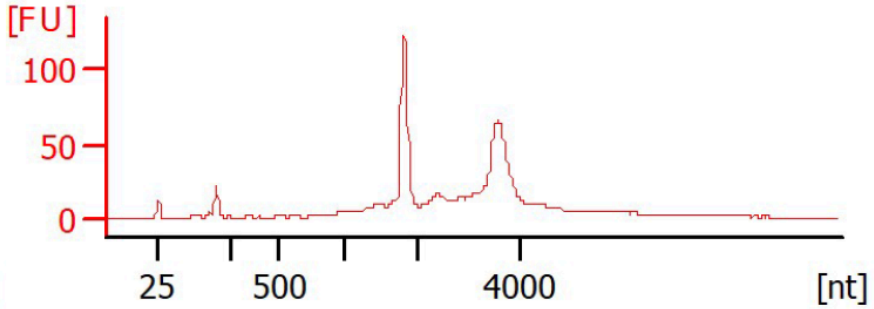
Exp5_Thymus_8866_M_No Oil_1Scroll

RIN:2 ; DV200 > 70%



Sample 11

RIN: 8.40



RNAseq Overview

rRNA-depletion or poly(A) selection

↓

cDNA synthesis

↓

Sequencer-specific library prep

↓

Library concentration normalization/sequencing

rRNA Depletion using RiboZero Reagent

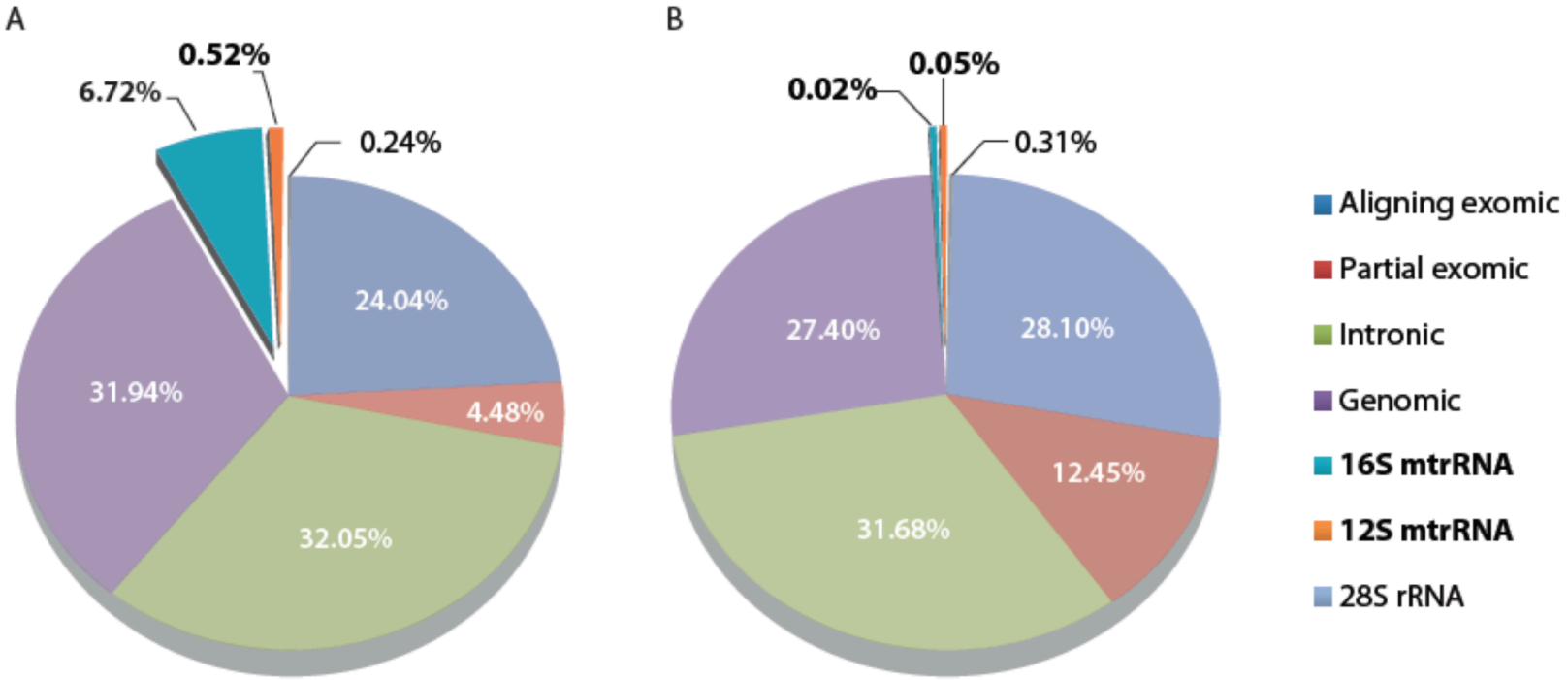


Figure 1. Profiles of RNA-seq libraries prepared after treatment with the Ribo-Zero™ (A) and Ribo-Zero Gold (B) Kits. Total RNA from MCF-7 cells was treated with either the standard Ribo-Zero Kit or the Ribo-Zero Gold Kit, and RNA-Seq libraries were prepared using the ScriptSeq™ Kit. Libraries were sequenced on Illumina® GAII and HiSeq 2000 sequencers. Data courtesy Vladimir Benes and Jonathon Blake, EMBL GeneCore, Heidelberg, Germany.

NEBNext Ultra II Directional

Probe Hybridization

RNase H Digestion

DNase I Digestion

SPRI Select Clean up

RNA Fragmentation
(0, 5, 10, & 15 minutes)

First Strand cDNA Synthesis

Second Strand cDNA Synthesis

End repair/ dA-tailing

Adapter Ligation

SPRI Select

PCR Enrichment of Libraries

TruSeq Stranded Total RNA

RiboZero RNA Depletion

SPRI Select Clean up

RNA Fragmentation
(4, 6, 7, & 8 minutes)

First Strand cDNA Synthesis

Second Strand cDNA Synthesis

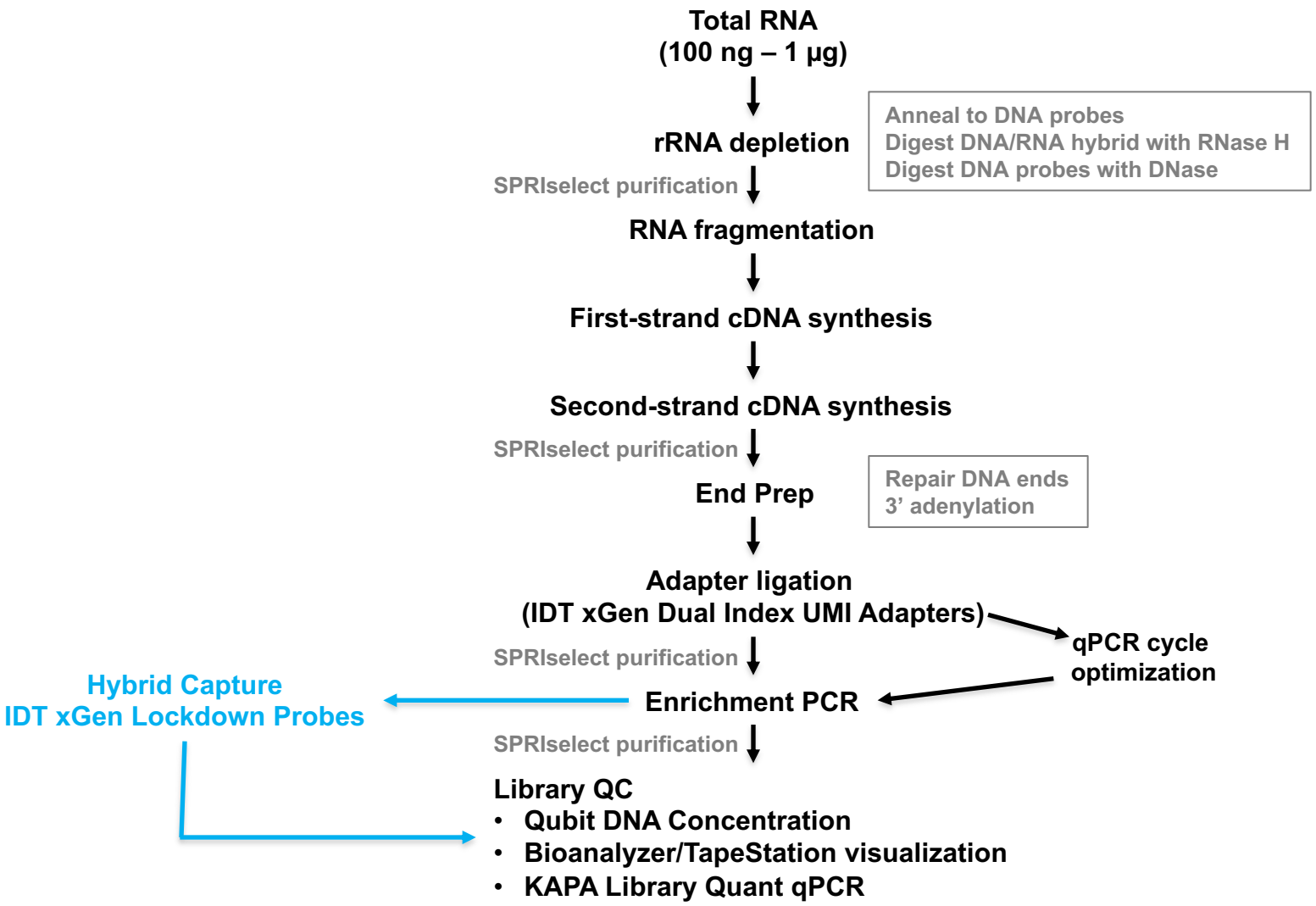
Adenylate 3’ Ends

Adapter Ligation

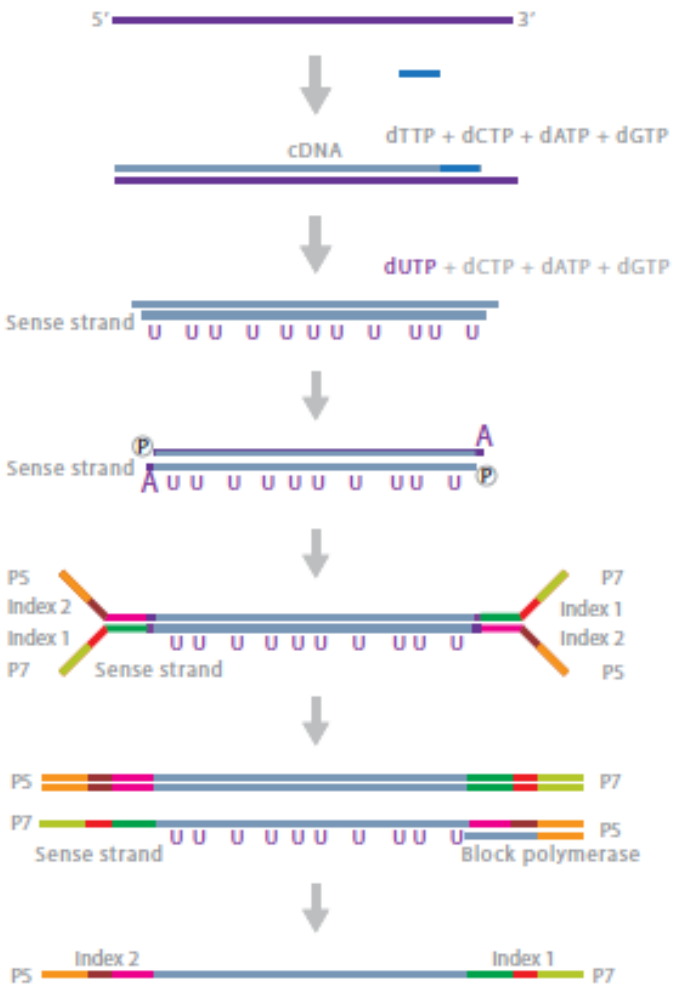
SPRI Select

PCR Enrichment of Libraries

NEBNext Ultra II Directional



RNA "stranded" Sequencing



RNA
Random primer

Create cDNA

Create second strand cDNA

End repair
Phosphorylate
A-overhang

Adaptor ligation

Denature and amplify

Product ready for cluster generation

Conclusions

- Widespread use of NGS platforms for biological research has changed the scale of biomedical inquiry and discovery
- A variety of platforms and approaches are available for implementation of NGS, each with strengths and weaknesses.
- Due to the size and complexity of the resulting data sets, validated approaches to assure quality of the data, along with analytical pipelines having appropriate sensitivity, specificity and reproducibility are required.
- Similarly, data visualization interfaces are needed to permit genome-wide and locus-specific evaluation of variants and underlying data support.