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.

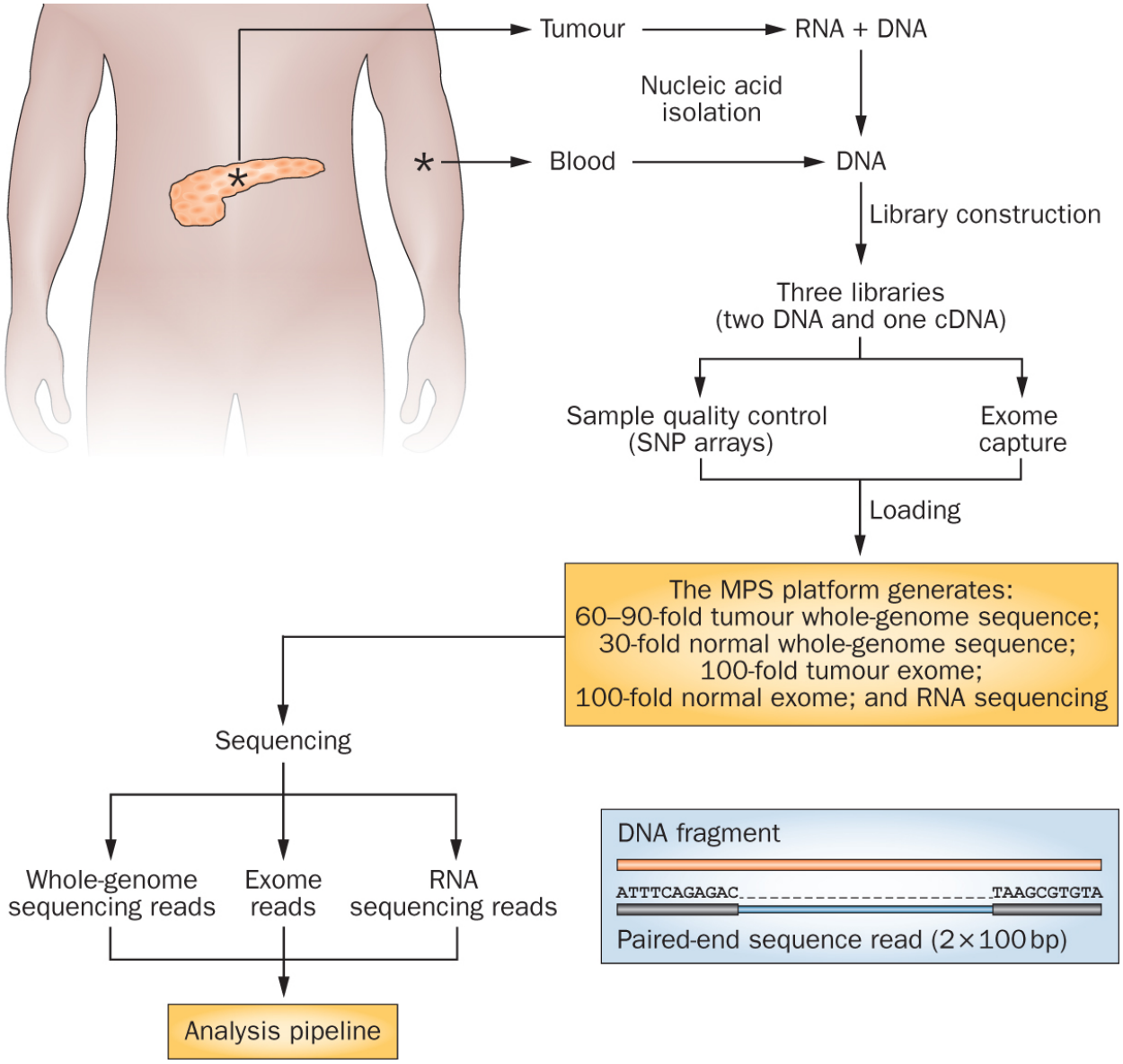
Under the following terms:

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.

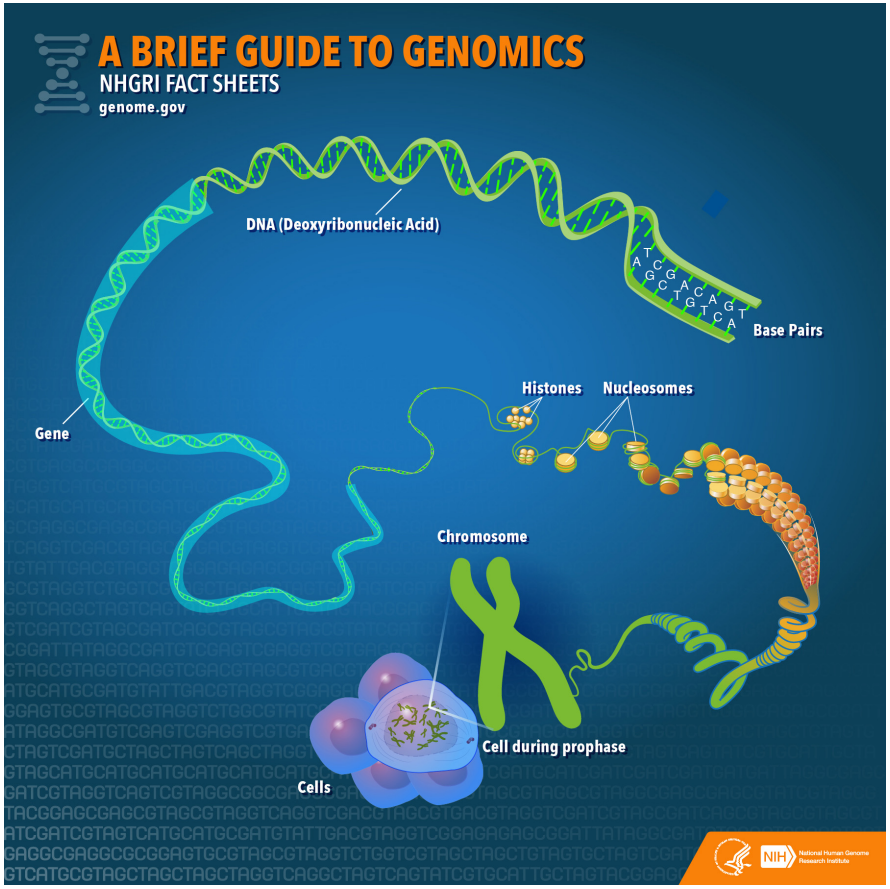
ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

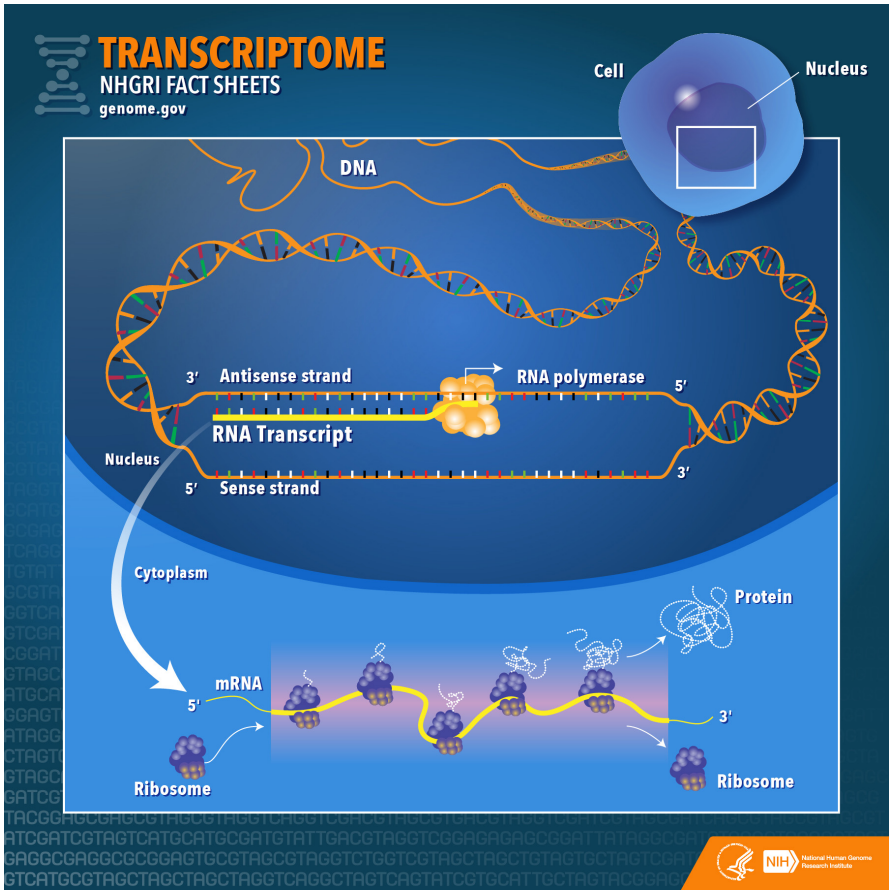
No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Cancer genomics data has exploded with rapid advances in sequencing technologies

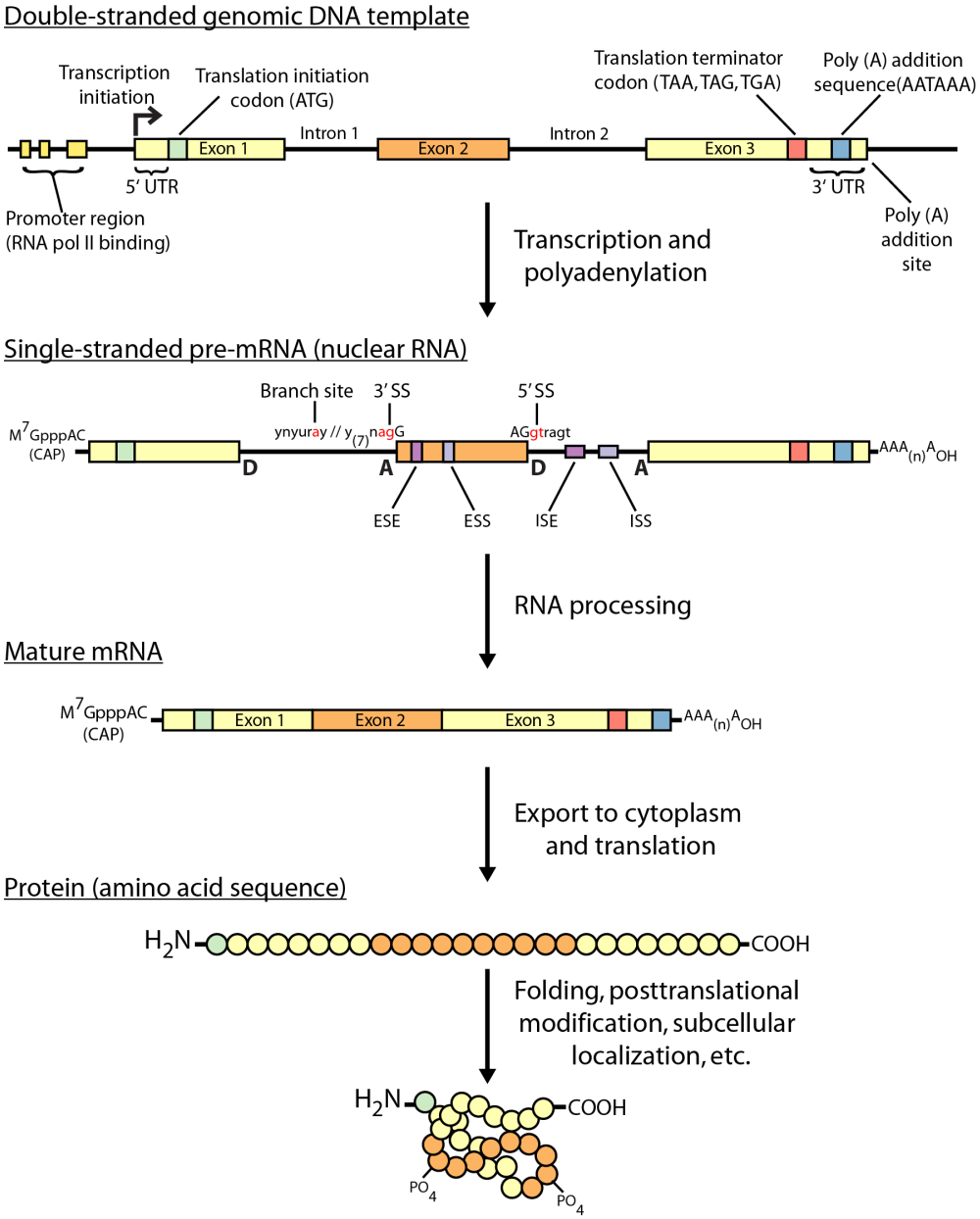


Genomes and transcriptomes

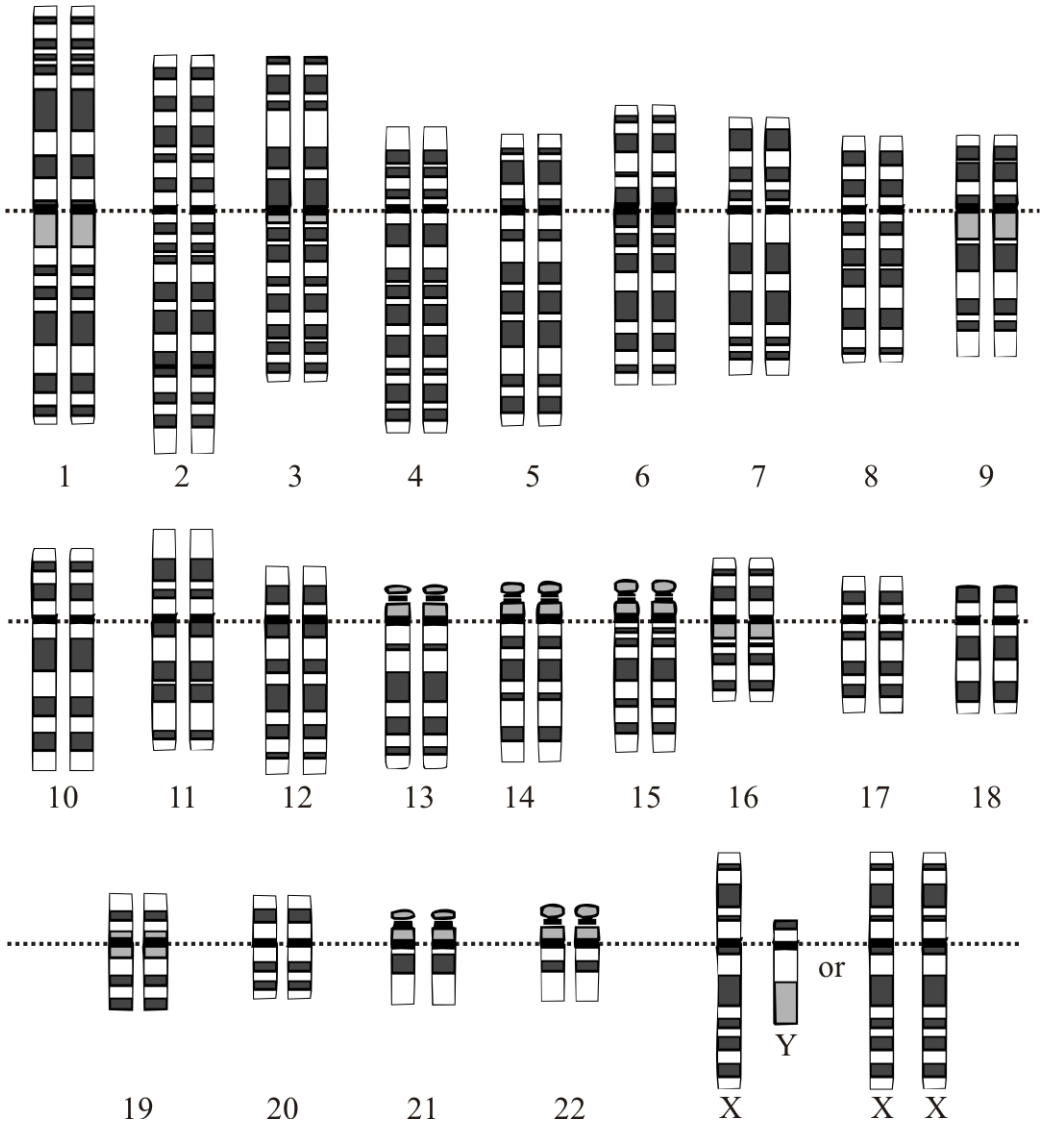




The Central Dogma

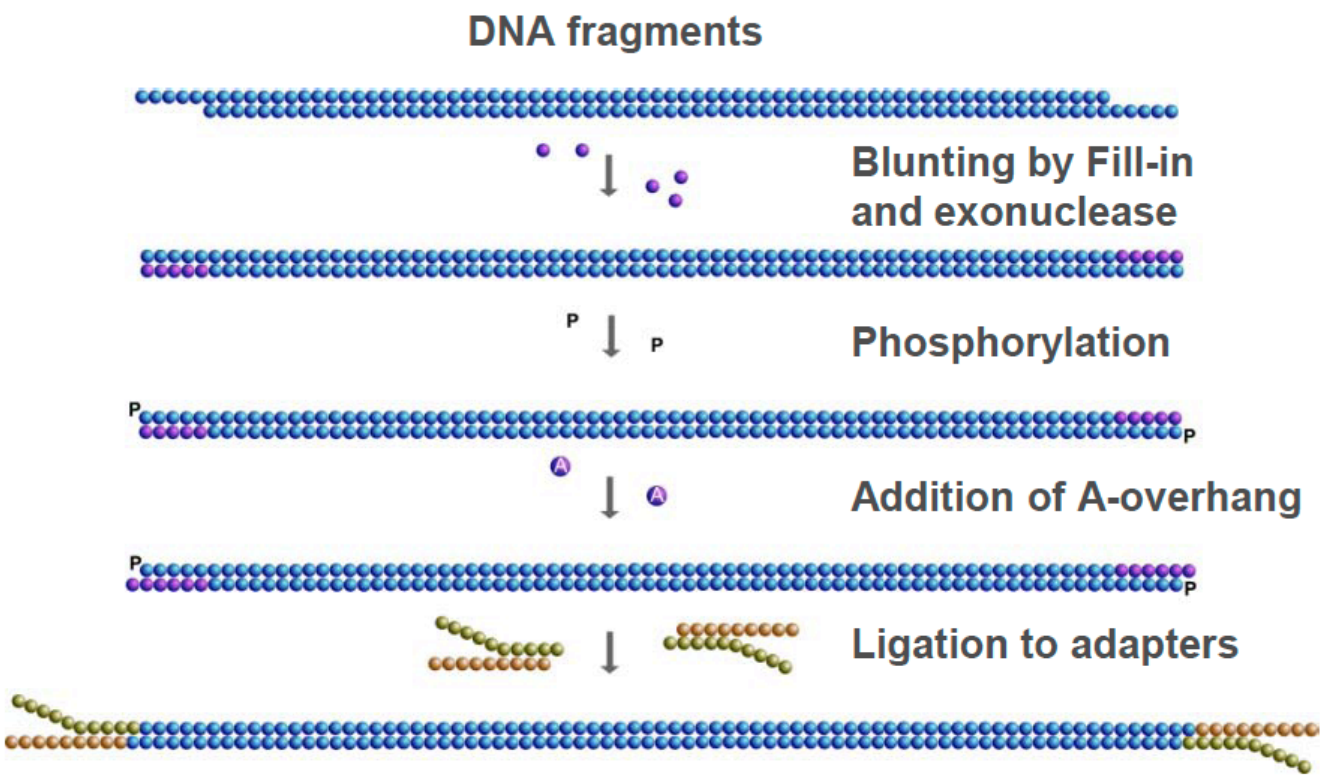


Human chromosomes (karyotype)



Data Generation

Library Construction for MPS



- Shear high molecular weight DNA with sonication
- Enzymatic treatments to blunt ends
- Ligate synthetic DNA adapters (each with a DNA barcode), PCR amplify
- Quantitate library
- Proceed to WGS, or perform exome or specific gene hybrid capture

Slide courtesy of Elaine Mardis (Nationwide Children's Hospital)

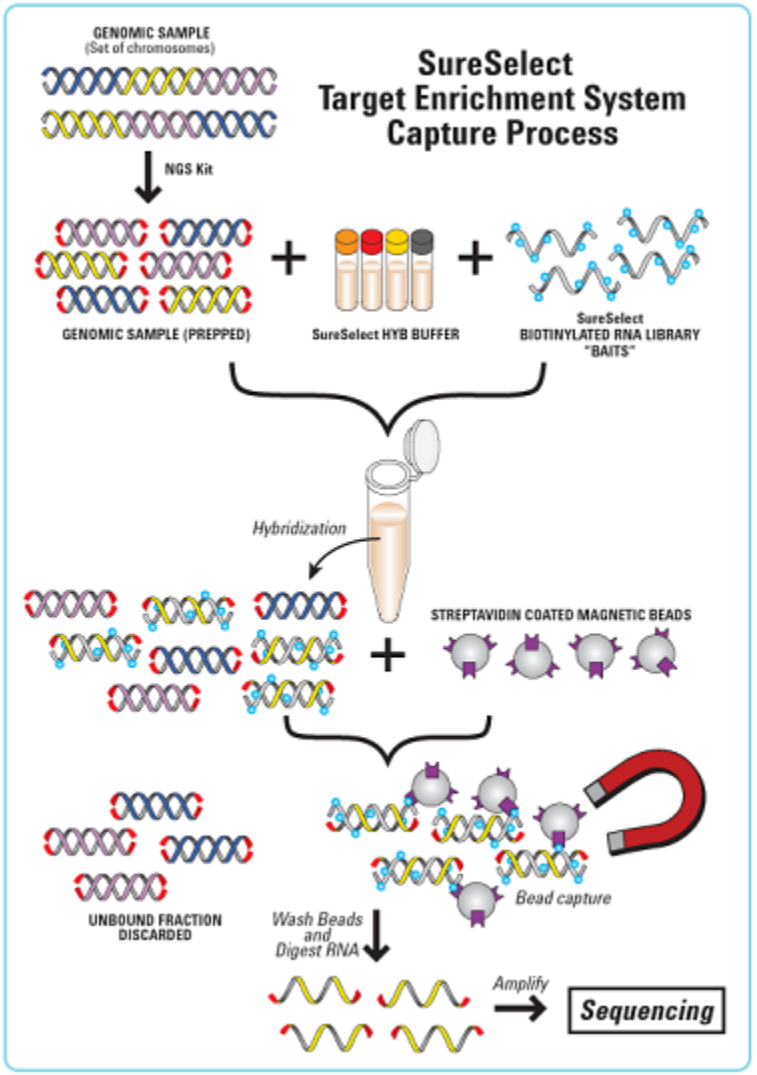
PCR-related Problems in MPS

- PCR is an effective vehicle for amplifying DNA, however…
- In MPS 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

Slide courtesy of Elaine Mardis (Nationwide Children's Hospital)

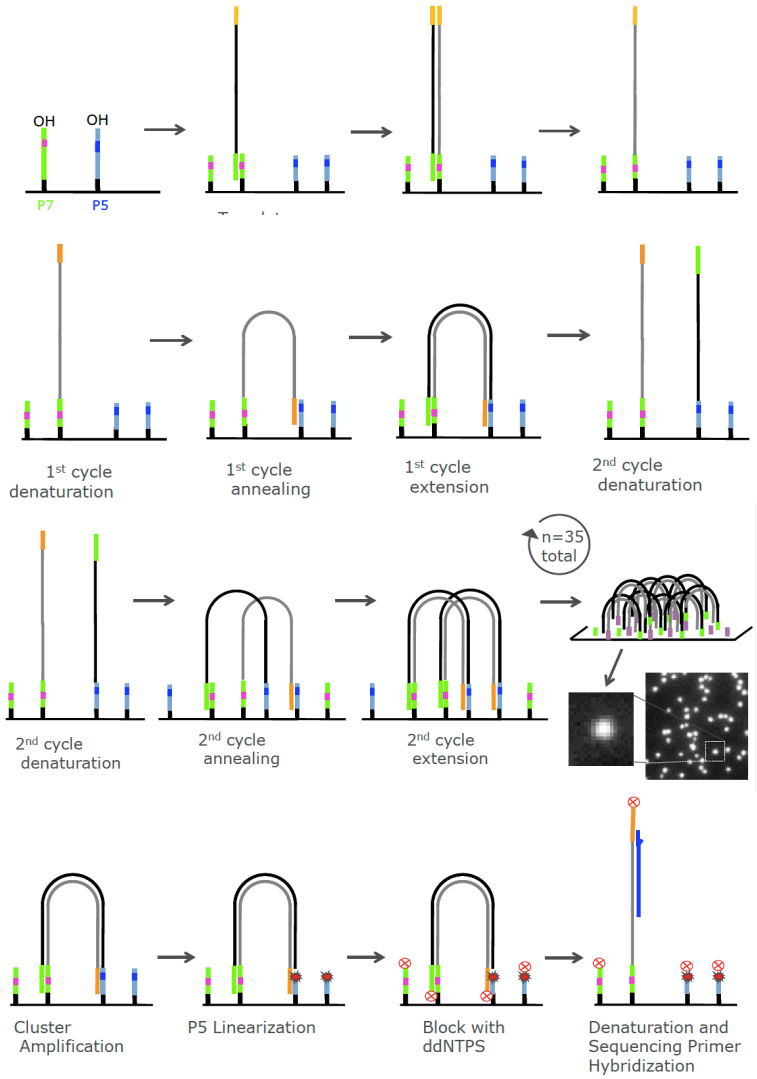
Hybrid Capture

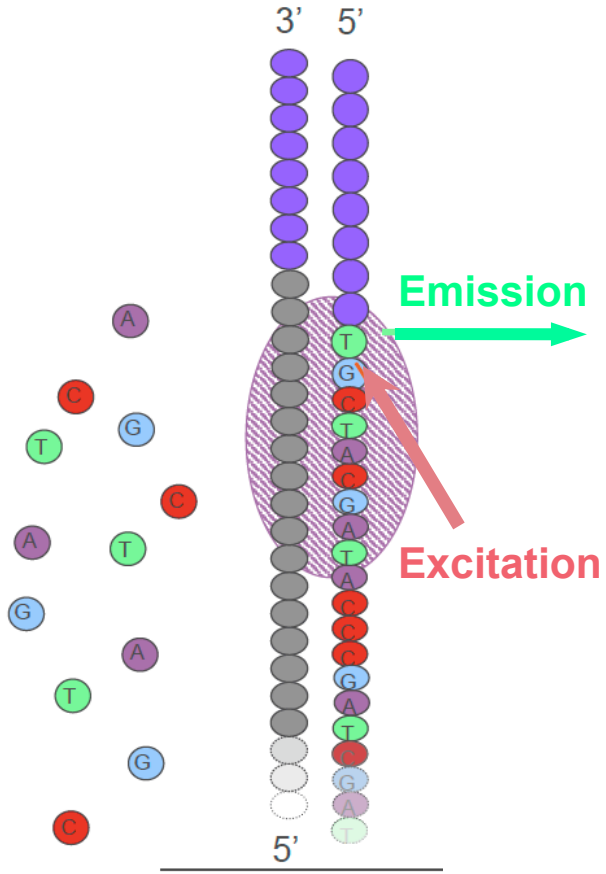
- **Hybrid capture** - fragments from a whole genome library are selected by combining with probes that correspond to most (not all) human exons or gene targets.
- The probe DNAs are biotinylated, making selection from solution with streptavidin magnetic beads an effective means of purification.
- An "**exome**" by definition, is the exons of all genes annotated in the reference genome.
- **Custom capture reagents** can be synthesized to target specific loci that may be of clinical interest.

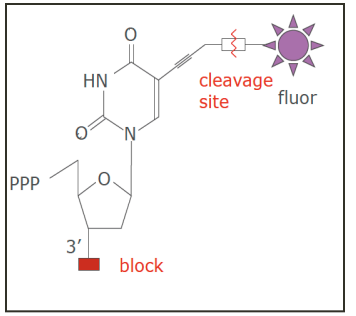


Slide courtesy of Elaine Mardis (Nationwide Children's Hospital)

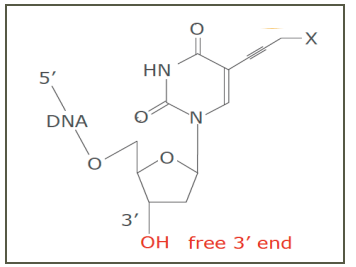
Massively Parallel Sequencing by Synthesis





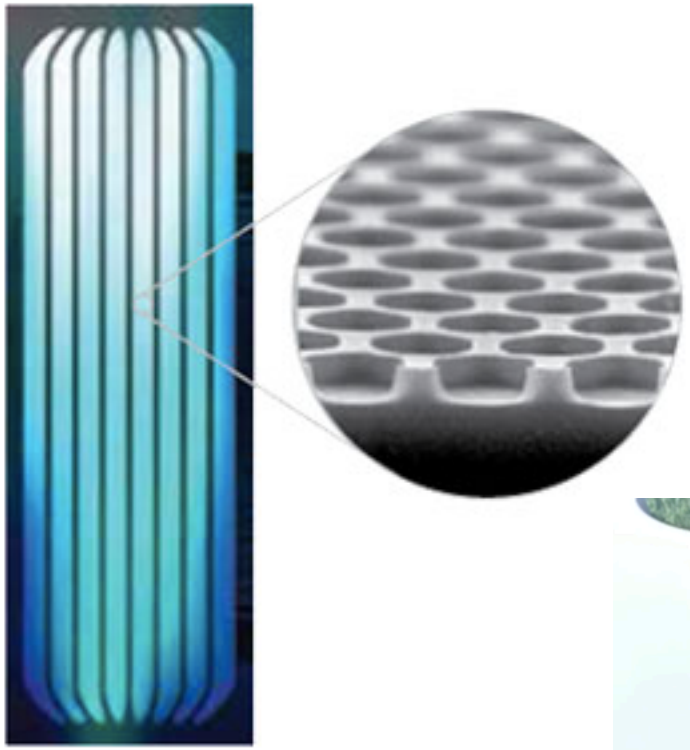


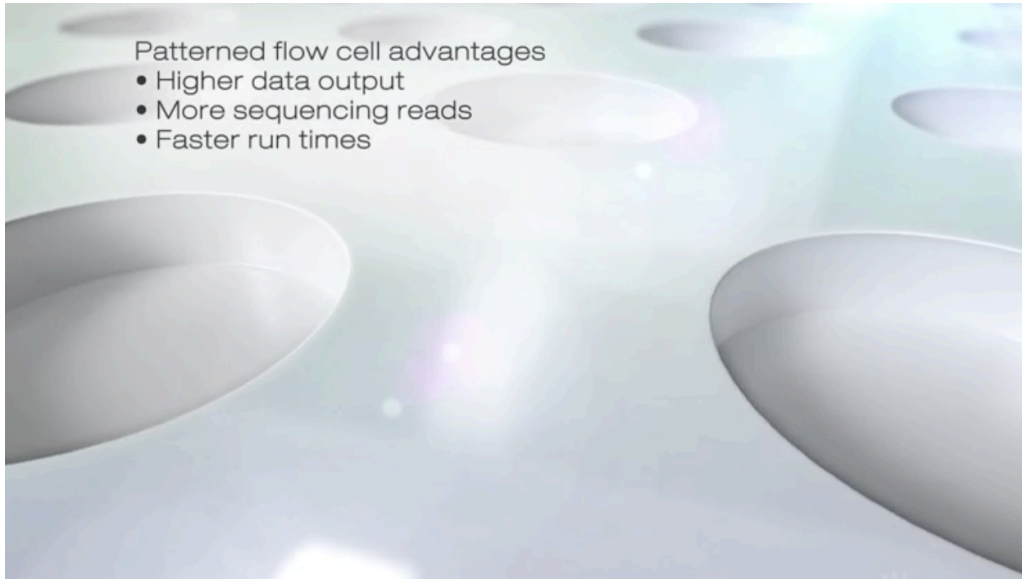
Incorporate
Detect
De-block
Cleave fluor

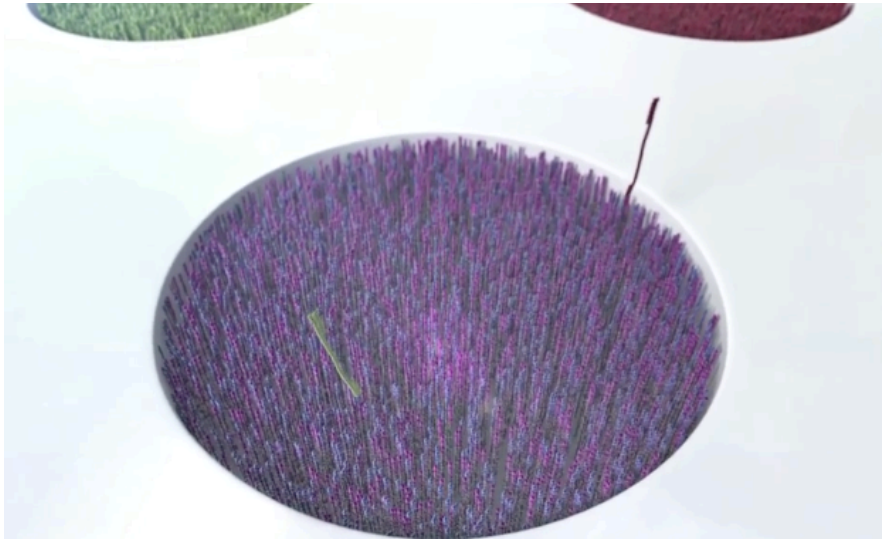


Slide courtesy of Elaine Mardis (Nationwide Children's Hospital)

Illumina Patterned Flow Cell







Slide courtesy of Elaine Mardis (Nationwide Children's Hospital)

Platforms: Illumina

	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

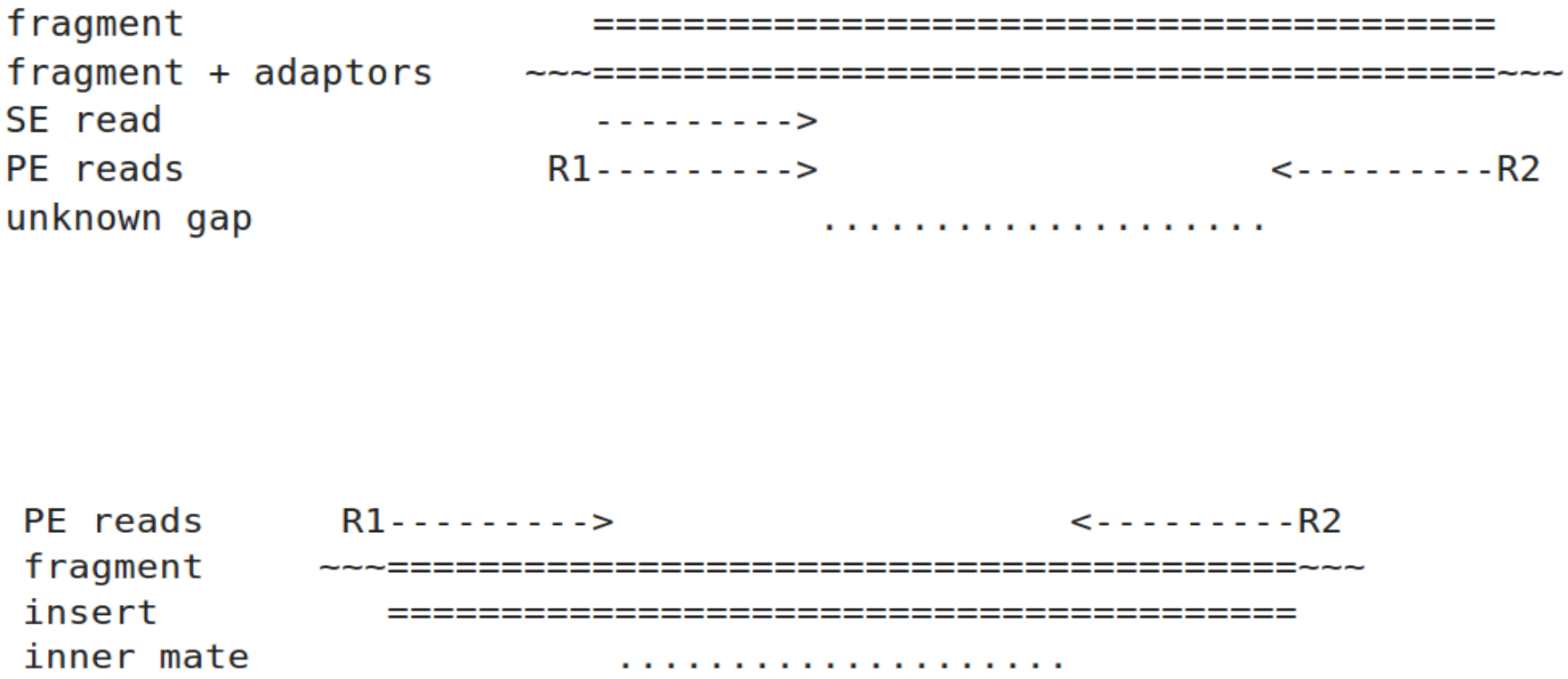
- High accuracy, range of capacity and throughput
- Longer read lengths on some platforms (MiSeq)
- Improved kits, improved software pipeline and capabilities, cloud computing in BaseSpace

Slide courtesy of Elaine Mardis (Nationwide Children's Hospital)

Our reference genome

- All reference files were obtained from the 1000 genomes project
  - The GRCh38 build of the human genome is used
  - This is the latest version of the human reference
- For the tutorial, two chromosomes are used (chr. 6 and chr. 17)
  - The reason for this is to reduce run time for the tutorial
  - Performing this analysis on the complete genome reference would require only minor modification of the commands
    - Would also require more storage, compute resources, and time

Insert size terminology explained



http://thegenomefactory.blogspot.ca/2013/08/paired-end-read-confusion-library.html