The University of Texas at Austin, Genomic Sequencing and Analysis Facility

or

for short

The Good, Bad, and Ugly of Next-Gen Sequencing

Scott Hunicke-Smith

2014

Outline

- Next-gen sequencing
  - Enabling technologies
  - Methods
  - Data Analysis
- Applications

NGS enabling technologies

- Clonal amplification (Exception: SMS)
  - Two methods: emulsion PCR (454, SOLiD), bridge amplification (Illumina)
- Sequencing by synthesis
- Massive parallelism

How they work videos

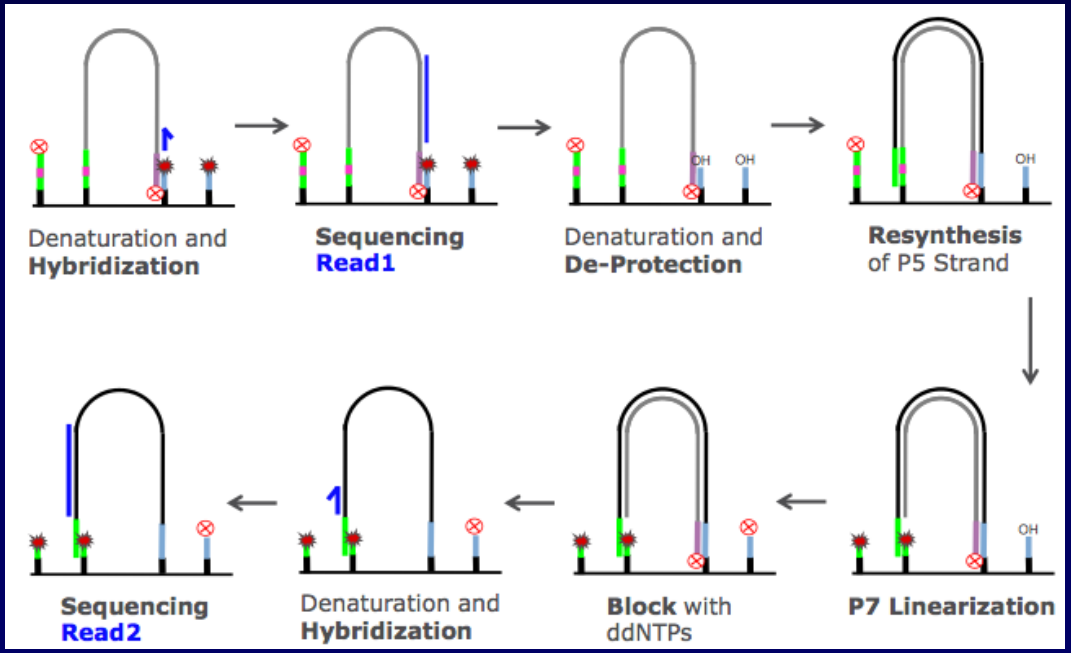
- Roche/454
  - http://454.com/products-solutions/multimedia-presentations.asp
- Illumina (Solexa) Genome Analyzer
  - http://www.youtube.com/watch?v=77r5p8IBwJk
- Pacific Biosciences
  - http://www.youtube.com/watch?v=NHCJ8PtYCFc

NGS enabling technologies

- Clonal amplification (Exception: SMS)
  - Two methods: emulsion PCR (454, SOLiD), bridge amplification (Illumina)
- Sequencing by synthesis
- Massive parallelism

Technologies employed

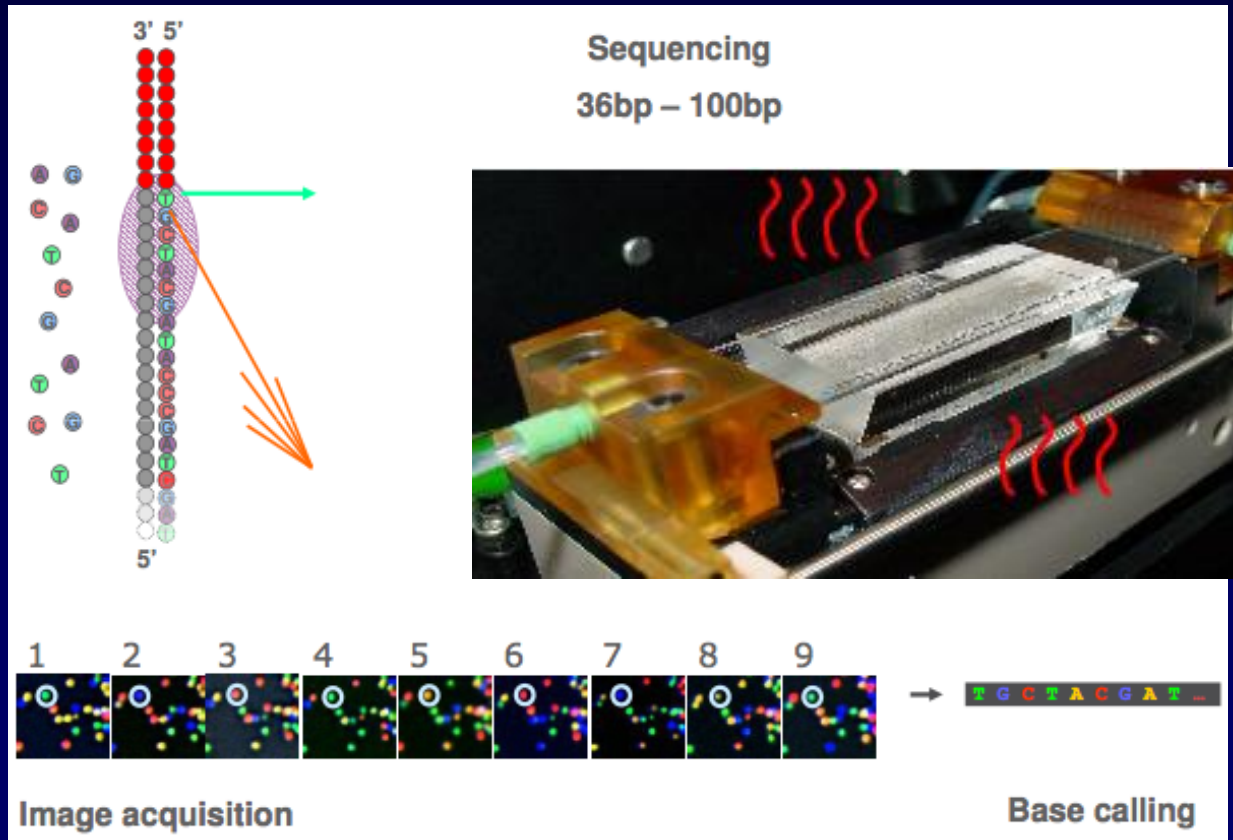
- Clonal amplification
  - Fully automated
  - Hardest part: [DNA]



Technologies employed

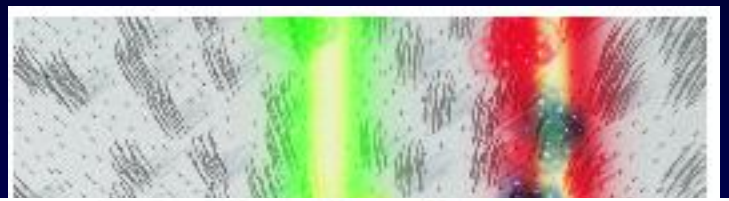
- Sequencing by synthesis
  - Four labeled, blocked dNTPs

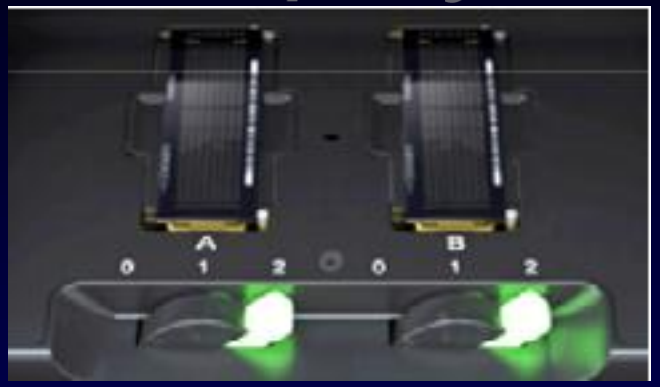
Algorithm:
1. Add dNTP & polymerase
2. React
3. Wash
4. Image
5. Unblock & cleave dye
6. Repeat

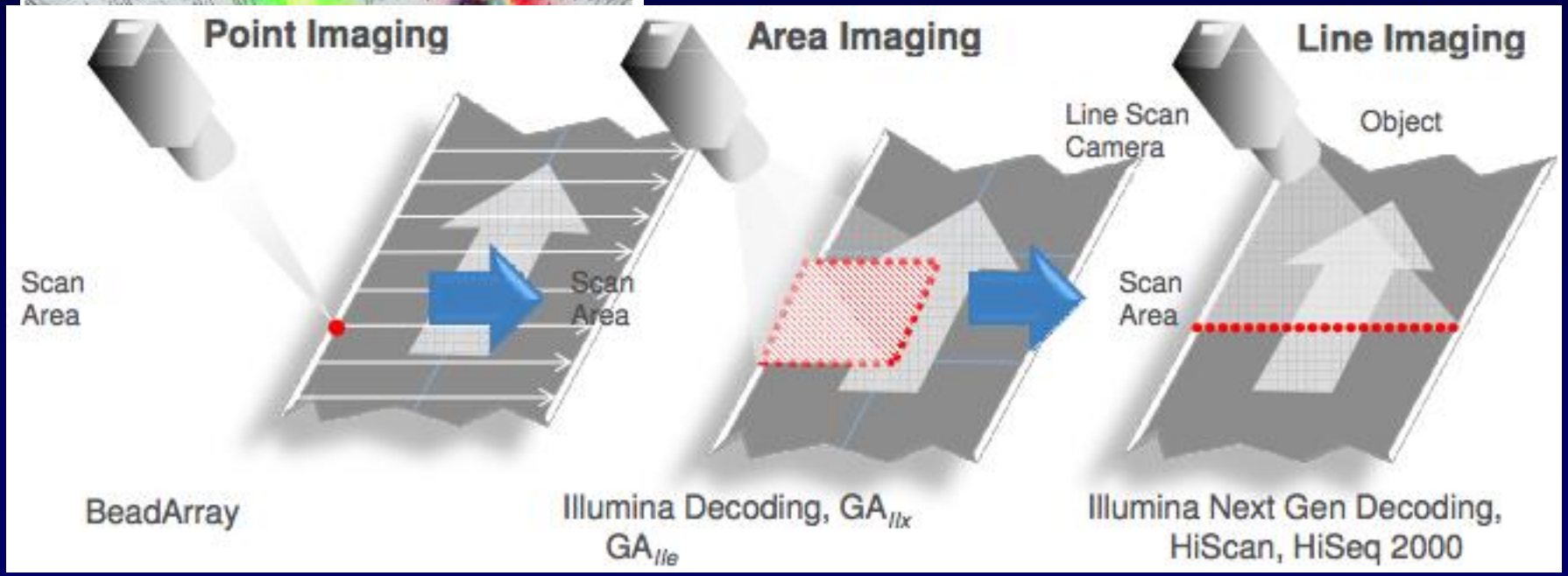


Technologies employed

- Massive parallelism







Illumina Sequencers

From Genome-Wide Discovery to Targeting Validation and Scre

Sequencing

Instrument	HiSeq X™ Ten*	HiSeq® 2500	NextSeq™ 500	MiSeqDx™	MiSeq®
Technologies	Sequencing by Synthesis (SBS) Powered by TruSeq Chemistry				
Applications	Population-Scale Whole Human Genome Sequencing	Production-Scale Genome, Exome, Transcriptome Sequencing and More	Exome, Transcriptome, Whole-Genome, Sequencing and More	FDA-Cleared *in vitro* Diagnostic System Cystic Fibrosis Screening and User-Defined Assays	Small Genome, Amplicons, Targeted Gene Panel Sequencing

* The HiSeq X Ten consists of 10 sequencing systems.

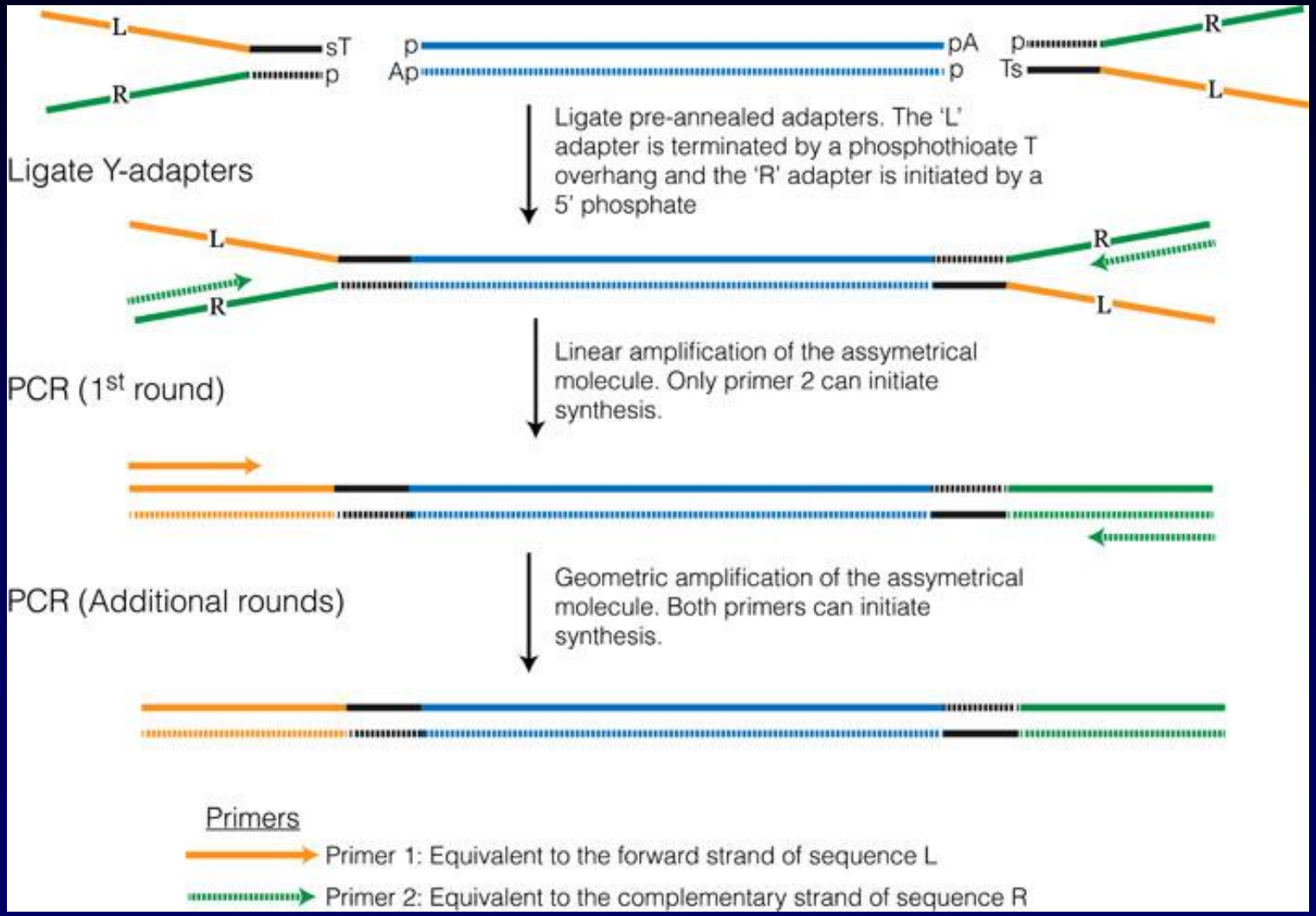
Sequencers at the UT GSAF

The Details: Categories

- Library Construction
- Sequencing
- Data Analysis

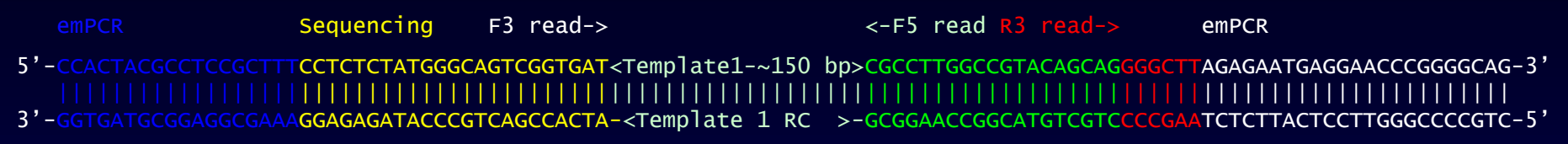
Library Construction: By Example

Clever trick: symmetric to asymmetric



From: rnaseq.uoregon.edu, "RNA-seqlopedia"

Read Types vs Library Types



- Single-end (F3 read only)
  - Cheapest, highest quality
- Paired-end (F3 and F5 read)
  - Much more information content
  - Differentiates PCR duplicates
- Mate-pair (F3 and R3 read)
  - Much more information content
  - Differentiates PCR duplicates
  - Provides info on large-scale structure

Library Construction: Workflows

Mate-Pair Libraries

Vendor	Illumina	Life Tech.	Roche	DNA Mass at step
Step	**GAII(x)**	**SOLiD (V3)**	**454 (Titanium)**	**output, ug**
Shear gDNA	X	X	X	9.000
Purify	X	X	X	8.100
End-repair	X	X	X	7.290
End-tag	X	X	X	7.000
Size select	X	X	X	1.400
Purify	X	X	X	1.260
Circularize	X	X	X	0.900
Isolate	X	X	X	0.810
Nick Translate		X		
Digest or Fragment	X	X	X	0.081
Enrich	X	X	X	0.061
Purify	X	X	X	0.055
End-repair	X	X	X	0.049
A-base addition	X			
Ligation	X	X	X	0.044
Purify	X	X	X	0.040
Amplify	X	X	X	40.815
Size select	X	X	X	8.163
Purify	X	X	X	7.347
Amount Required for Sequencer Clonal Amplification:				0.0001

Question

- Which of these was NOT an enabling invention for NGS:
  A. Clonal amplification
  B. Intercalating dyes
  C. Sequencing by synthesis
  D. Massive parallelism

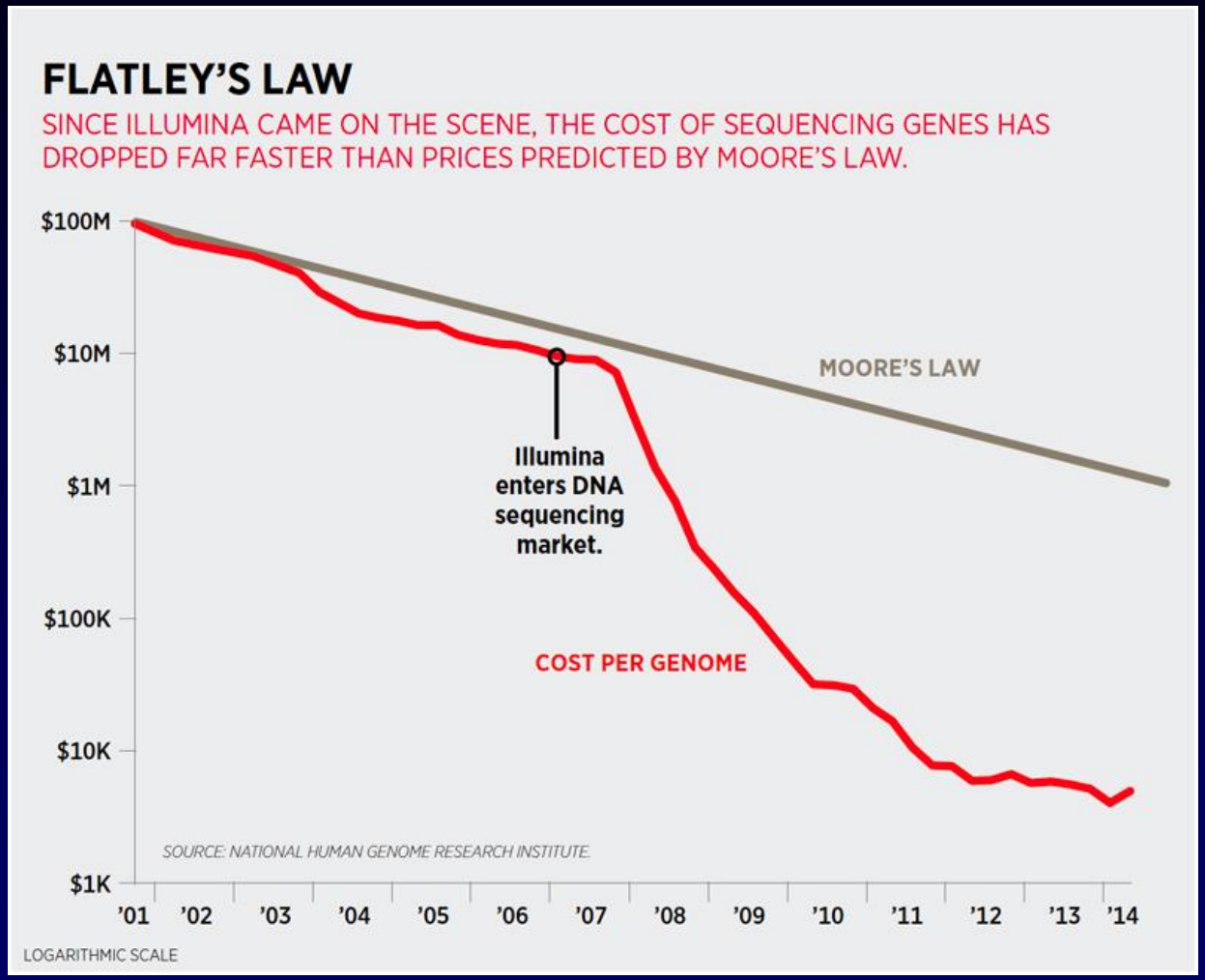
Characteristics of SBS

- Step-wise efficiency is <100%
  - Like inflation eating away at your savings
- This can be resolved by correcting “phasing”
  - This single software addition increased read lengths by ~10-fold
- Dominant error modalities can be predicted based on the technology
  - Fluor-term-nucleotide systems have ____ errors
  - Native (un-terminated) systems have ___ errors

Essential Ideas

- NGS interrogates populations, not individual clones
- Number of reads (sequences) ≅ 100x library molecules put into clonal amplification
  - MOLAR RATIOS matter!
  - Highly repeatable (from library through sequencing)
- Error rates are (very) high (compared to Sanger)
- NGS was a multi-disciplinary effort

Trajectory of Price



From "Flatley's Law: The Company Speeding A Genetic revolution", Forbes, Sept. 8, 2014

What it costs

- Examples:
  - Deep Sequencing:
    - Illumina RNA-seq: 1 sample, 40 million read-pairs: ~$500
    - Illumina *de novo*: Draft sequence ~5 megabase bacterial genome (~25 MB raw sequence): ~$150
    - Illumina human exome: $800
    - Illumina whole human genome: $5,000 (at UT), $1,000 elsewhere if you buy “by the hundreds”

What the data looks like

@M01012:85:000000000-A6FB5:1:1101:16490:1455
TGAGAGCCGCTGTAGANATGCGATCACTGGGGAAAACAGGAAAGGAGGTGAAATGCAGAGCAAGCTGTGA
+
CCCCCFCFCCCCGGGG#AAFGGGHGGHHHHHGGGHHGHHHHHHHHGHFGEFHHHHHHHHHHHHHHHHFFHH
@M01012:85:000000000-A6FB5:1:1101:14313:1461
CTCTGTTTCTTTTTTCACGTGGTTTCTCCACATGACTAGCTTAAGTTTTCTCACAGCATGGACCCTCAGG
+
AAAA@B@FFFFFGGCG3FCFGC0AA1FFHB01FFFHGGGFHFGHGHH2AGHF2ABA/FGHGGFHGC/CGF

Aligners/Mappers

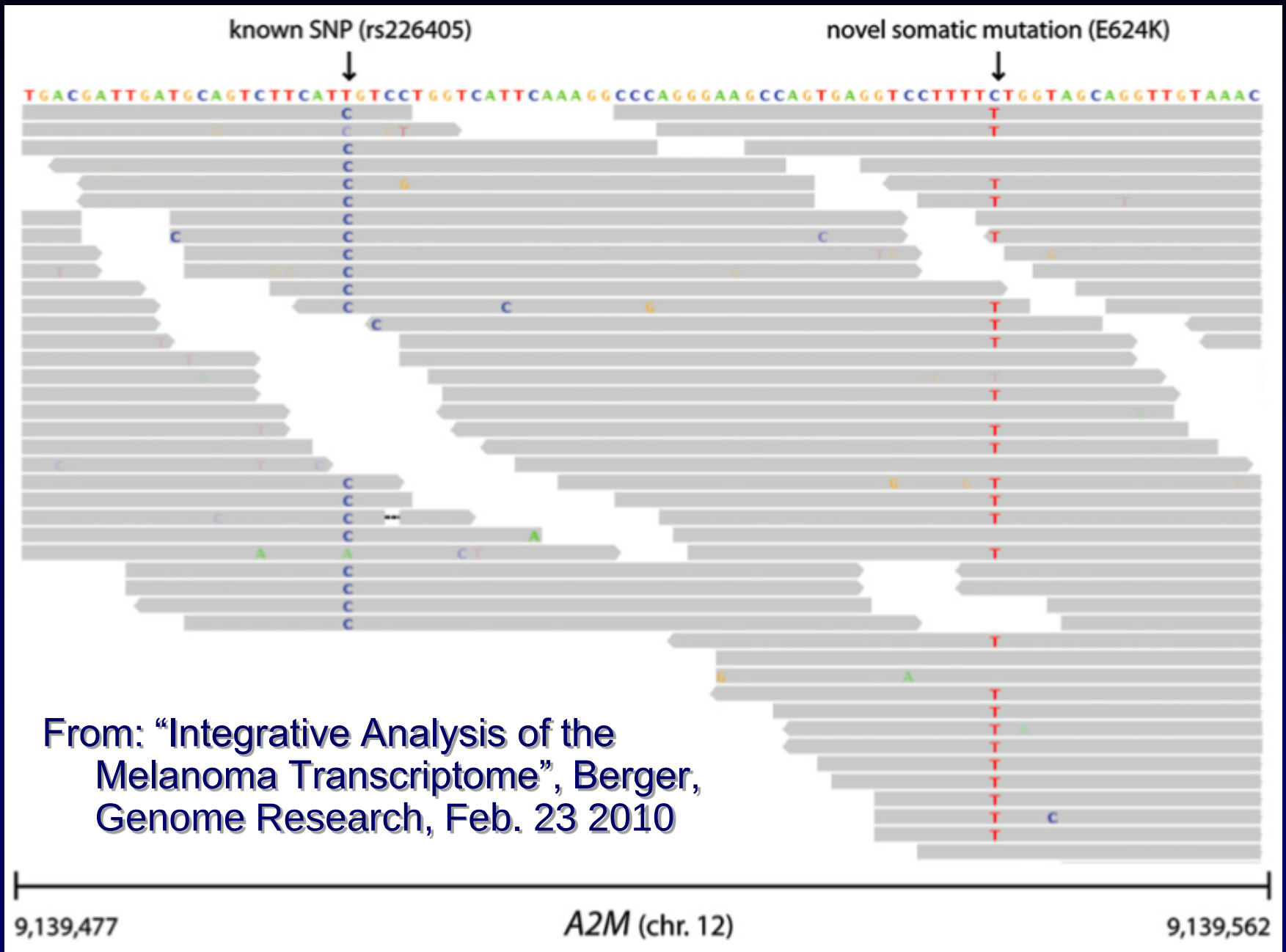
- Algorithms
  - Spaced-seed indexing
    - Hash seed words from reference or reads
  - Burrows-Wheeler transform (BWT)
- Differences
  - Speed
  - Scalability on clusters
  - Memory requirements
  - Sensitivity: esp. indels
  - Ease of use
  - Output format

Aligners/Mappers

- Differences in alignment tools:
  - Use of base quality values
  - Gapped or un-gapped
  - Multiple-hit treatment
  - Estimate of alignment quality
  - Handle paired-end & mate-pair data
  - Treatment of multiple matches
  - Read length assumptions
  - Colorspace treatment (aware vs. useful)
  - Experimental complexities:
    - Methylation (bisulfite) analysis
    - Splice junction treatment
    - Iterative variant detection

Taken from: http://www.bioconductor.org/help/course-materials/2010/CSAMA10/2010-06-14__HTS_introduction__Brixen__Bioc_course.pdf

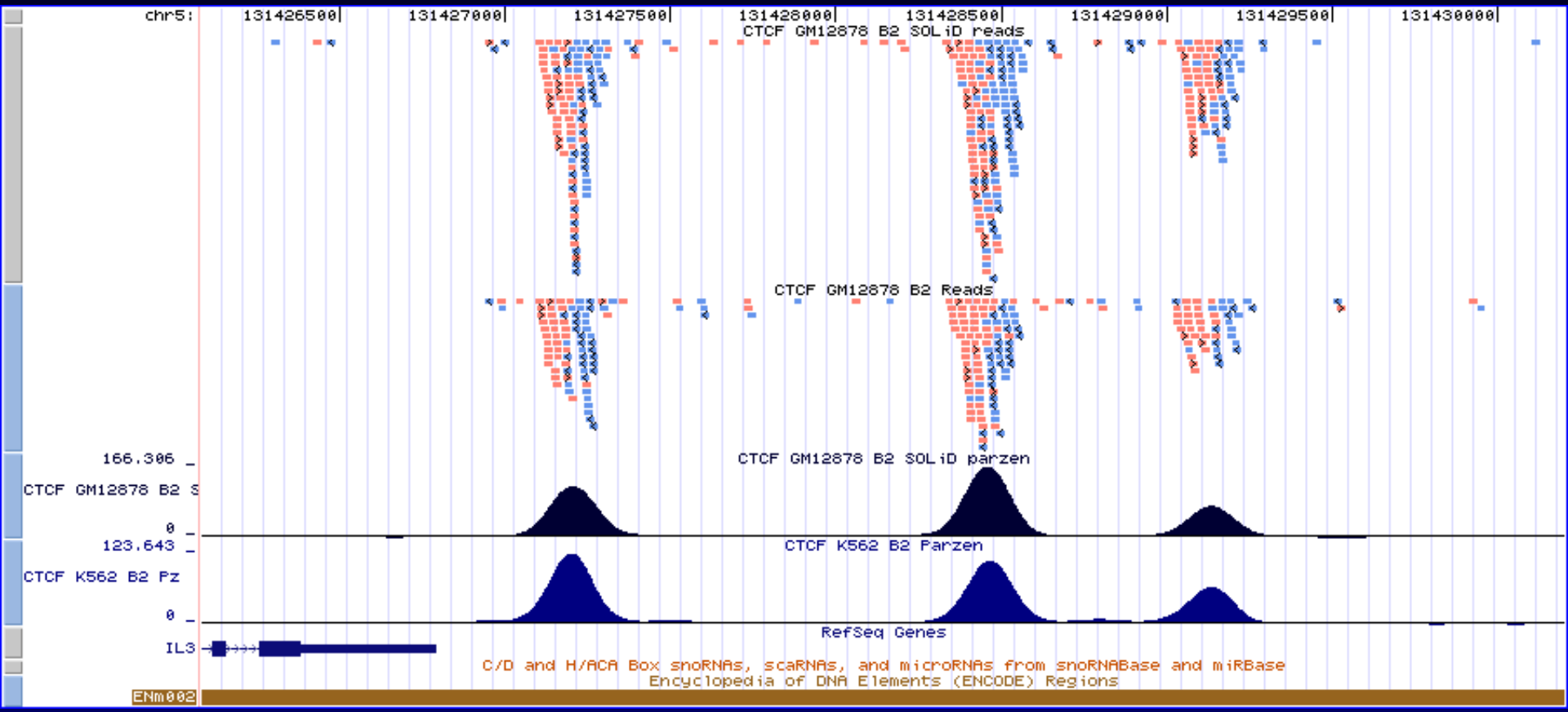
Real (applied) data



From: “Integrative Analysis of the Melanoma Transcriptome”, Berger, Genome Research, Feb. 23 2010

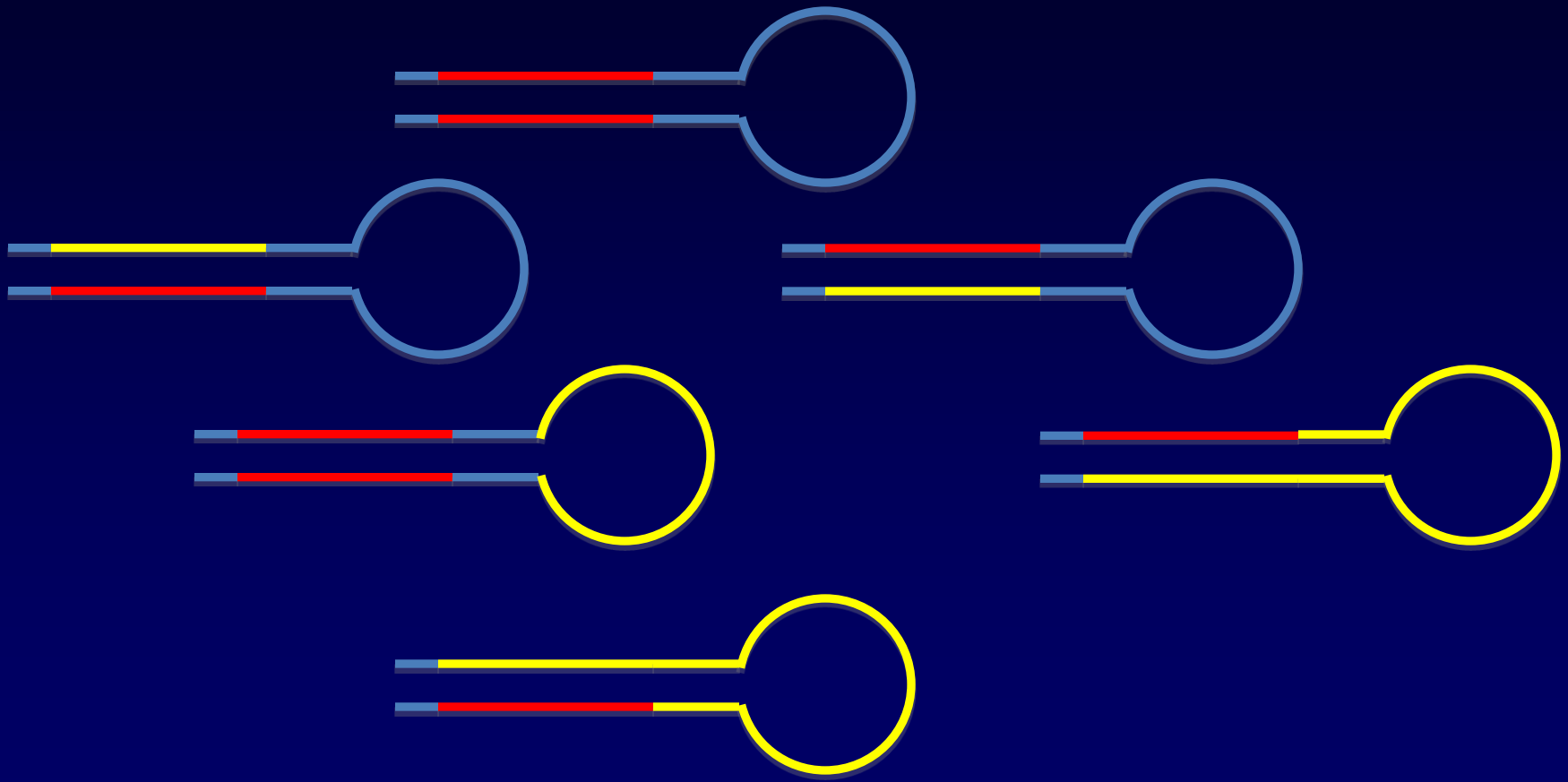
What, exactly, are we sequencing?

Good Example: ChIP-Seq

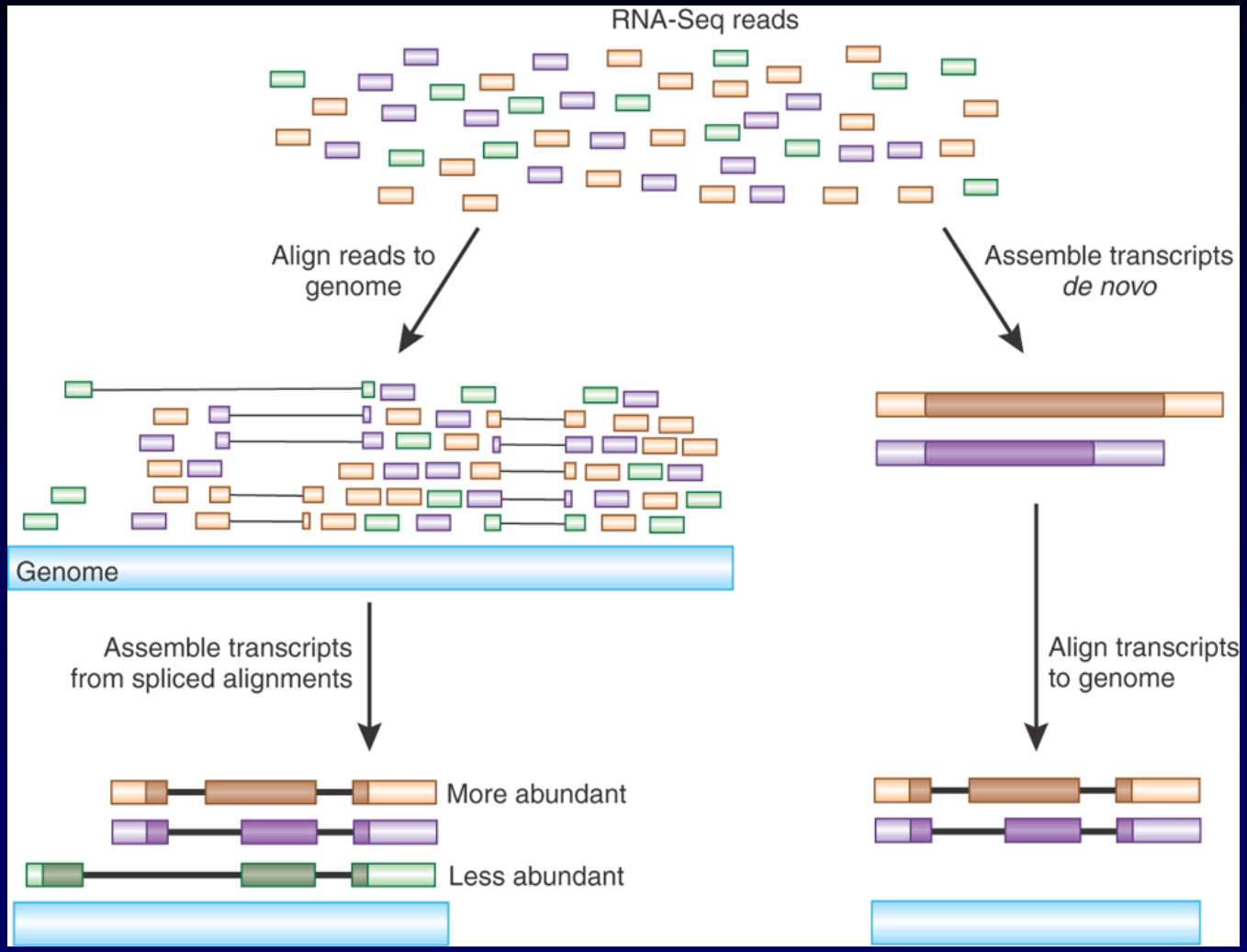


RNA/miRNA library

- What's in YOUR library?



RNA-seq



From: "Advancing RNA-Seq analysis", M.C. Zody and B.J. Haas, Nature Biotechnology 28, 421–423 (2010) doi:10.1038/nbt0510-421.

RNA-seq

- Quantitation – what's in YOUR genome?
  - CAACCCCAACACCCACCGGCACACAGACCCCAACC – 99x
  - CAACCCCAACACCCACCGGCACACAGACCGGGCCC – 1x
- **You found a transcript WHERE?**
  - Jesse Gray @ Harvard:
  - ChIP-Seq data showed RNA **Pol II binding tens of KB away from any** annotated gene, in a promoter/enhancer complex
  - RNA-Seq data confirmed ~1kb transcripts arising from these binding sites

Informatics Pipelines: RNA-seq

- General workflow:
  - Pre-filter (optional)
  - Map
  - Filter
  - Summarize (e.g. by gene or exon)
  - Filter
  - Interpret
- Rule sets are required to make sense of the “unbiased” sequence data
- Rule sets can get complicated quickly
- Algorithm matters (speed, sensitivity, specificity)

Question

Which type of mathematics are you most likely to need when analyzing NGS data:

A. Calculus
B. Linear algebra
C. Statistics
D. Differential equations
E. Set theory

(Hint: it has been removed from Texas requirements for high school math)

Applications

"Good applied science in medicine, as in physics, requires a high degree of certainty about the basic facts at hand, and especially about their meaning, and we have not yet reached this point for most of medicine."

— Lewis Thomas, The Medusa and the Snail (1979)

(Thomas was Dean, Yale Med & President, Memorial Sloan Kettering)

Washington Univ: Microbiomes

- Microbiome:
  - "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space." (Wikipedia)
- Metagenomics:
  - "the study of metagenomes, genetic material recovered directly from environmental samples." (Wikipedia)

Washington Univ: Microbiomes

- NICU bacteremia watch – Phil Tarr, Barbara Warner, and George Weinstock
  - Pilot project: 632-day period
  - Every diaper is stored, all blood stored
  - Were able to find:
    - One bacteremia case identified 10 days earlier than standard clinical detection
    - Observed two cases of enterococcus – one which evolved to Daptomycin resistance & was fatal
    - Route of infection – parents, visitors, nurses, docs

Washington Univ: Microbiomes

- Fever of unknown origin
  - About 1/3 do not get a clear diagnosis from microbiology/virology
  - They have been able to identify the virus in nearly all cases tested so far
  - Typically nasopharyngeal swabs; may be blood testing (plasma)

Washington Univ: Microbiomes

- Areas of research
  - Reaching actionable results faster, cheaper, and with higher accuracy
    - “Actionable” may mean anticipating drug response
  - Tougher diseases like Kawasaki disease

MCW: General Pediatric Cases Data Analysis Pipeline

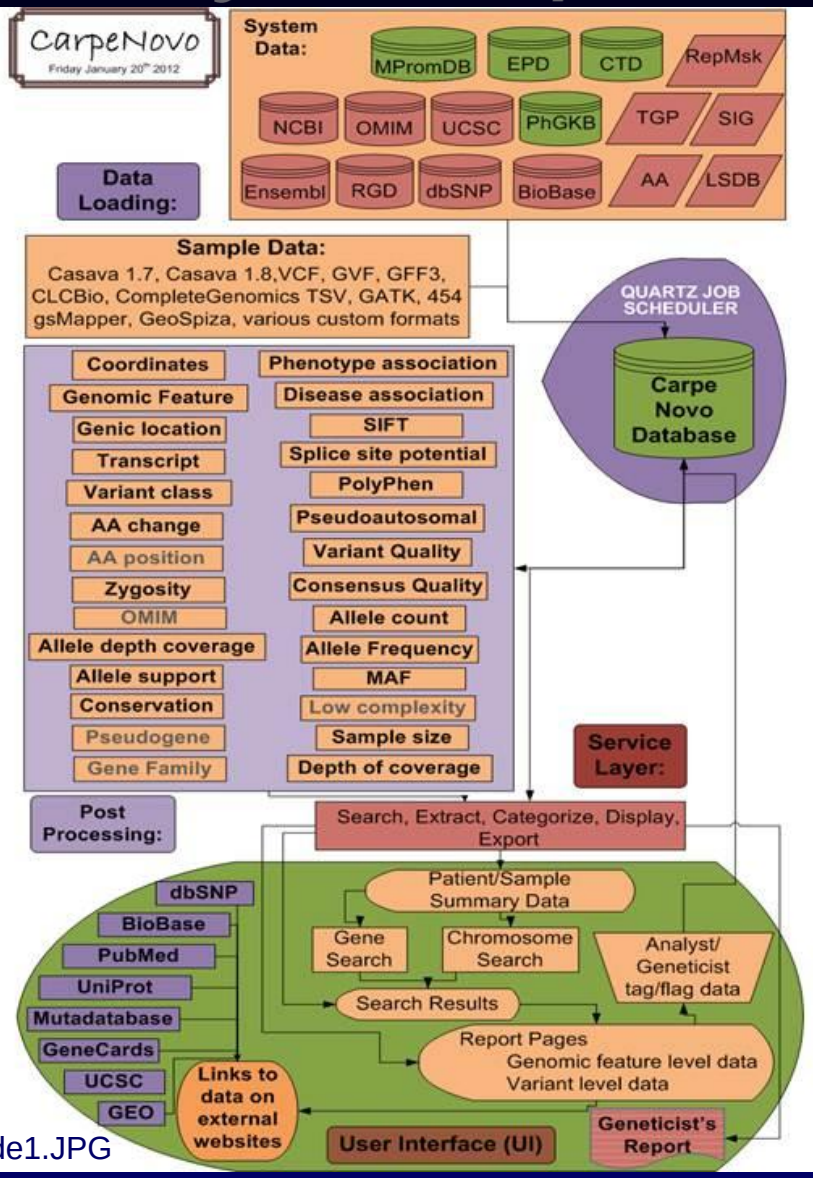
CarpeNovo – Clinical variant analysis platform

All major sequencing technologies supported

More than 100 pieces of data generated or brought in to assist in identification of disease causative/associated/drug response variations

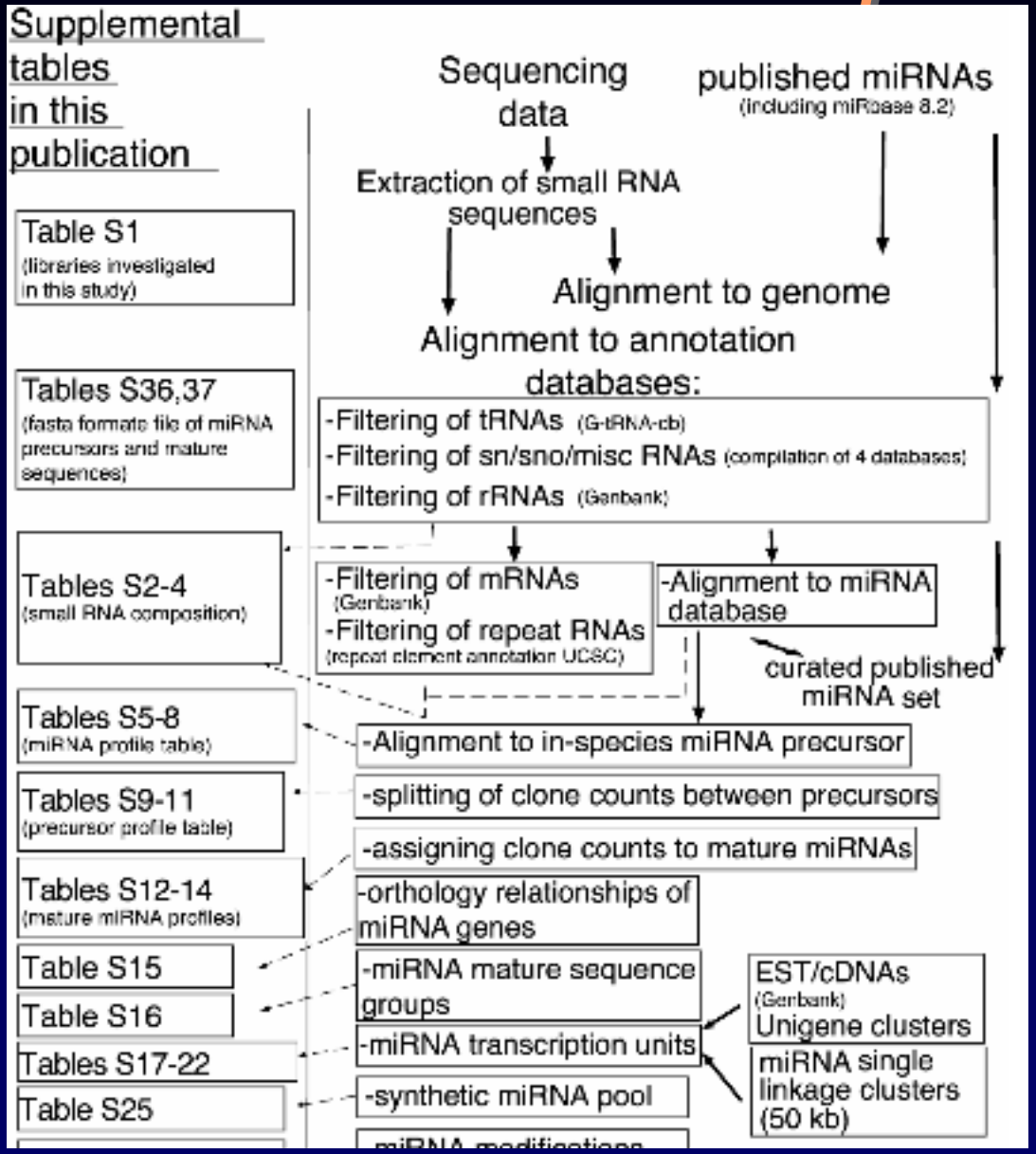
College of American Pathologists regulatory approval obtained

In use in the Agen CLIA NGS MDx lab



From: http://www.hmgc.mcw.edu/images/faculty/Slide1.JPG

Pipeline example



From: Landgraf, et. al., “A mammalian microRNA expression atlas based on small RNA library sequencing.”, Nat Biotechnol. 2007 Sep; 25(9):996-7, supplemental materials

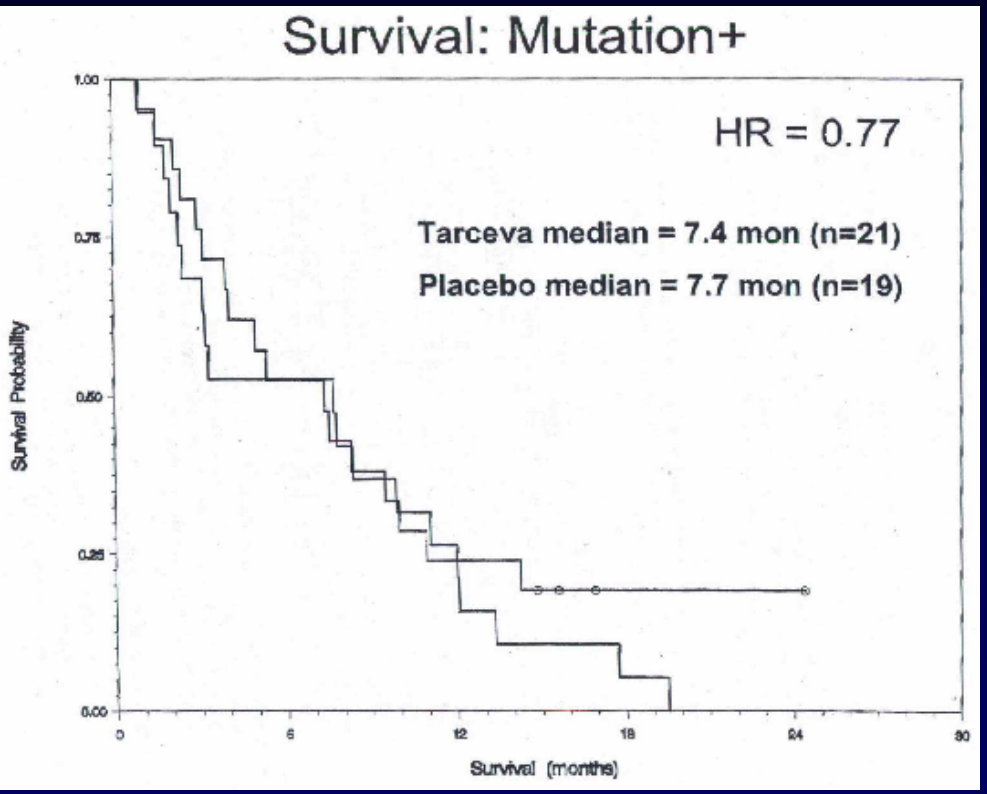
TACC: A Joy in Life

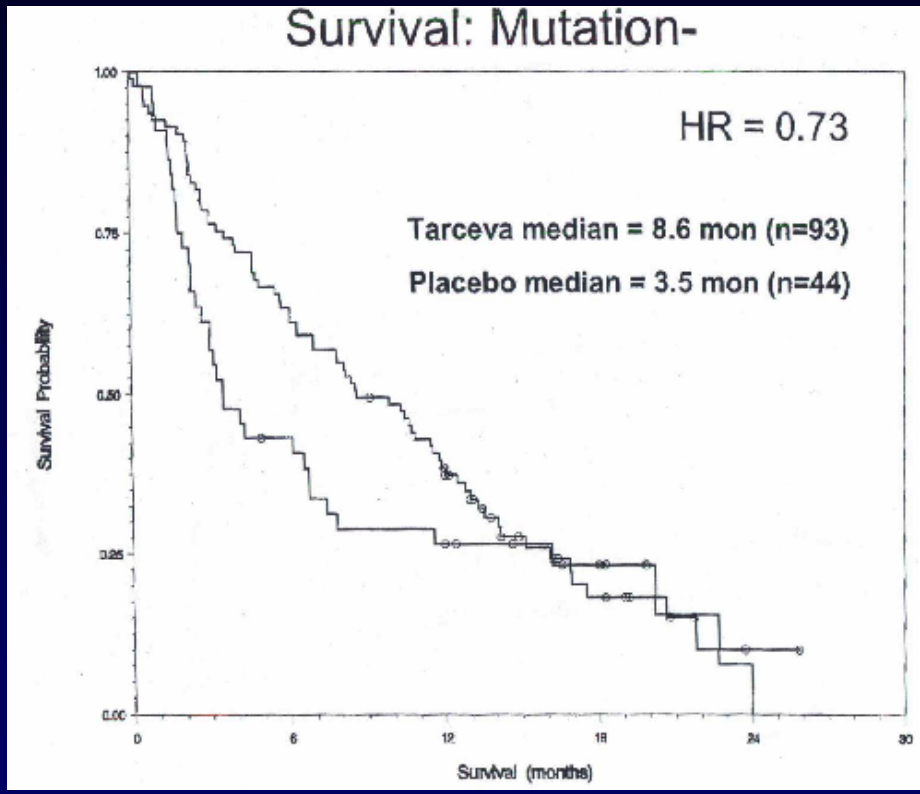
- Stampede: 492,800 processing cores, 14 PB disk space
- RANCH & CORRAL: >70 PB archive
- Typical mapping of 20e6 reads:
  - 20 hours on high-end desktop
  - 2 hours at TACC

Medical Examples

- Gleevec targeting BCL/ABL
  - First CML, then GIST
  - “Too” specific… and $32,000/year
  - See also: Herceptin, Avastin, Cetuximab…
- Warfarin
  - CYP 450 enzymes have regulators too…
- Irinotecan: UGT1A1
  - Irinotecan is converted by an enzyme into its active metabolite SN-38, which is in turn inactivated by the enzyme UGT1A1 by glucuronidation.
  - # The most common polymorphism is a variation in the number of TA repeats in the TATA box region of the UGT1A1 gene. The presence of seven TA repeats (UGT1A1*28) instead of the normal six TA repeats (UGT1A1*1) reduces gene expression and results in impaired metabolism. This variant allele is common in many populations, and occurs in 38.7% of Caucasians, 16% of Asians and 42.6% of Africans.1,2
  - # Studies have shown that impaired metabolism in patients who are homozygous for the UGT1A1*28 allele results in severe, dose-limiting toxicity during irinotecan therapy. These findings led to a recent update in the irinotecan label to include dosing recommendations based on the presence of a UGT1A1*28 allele.3.
  - From: http://www.twt.com/clinical/ivd/ugt1a1.html

Tarceva: EGFR



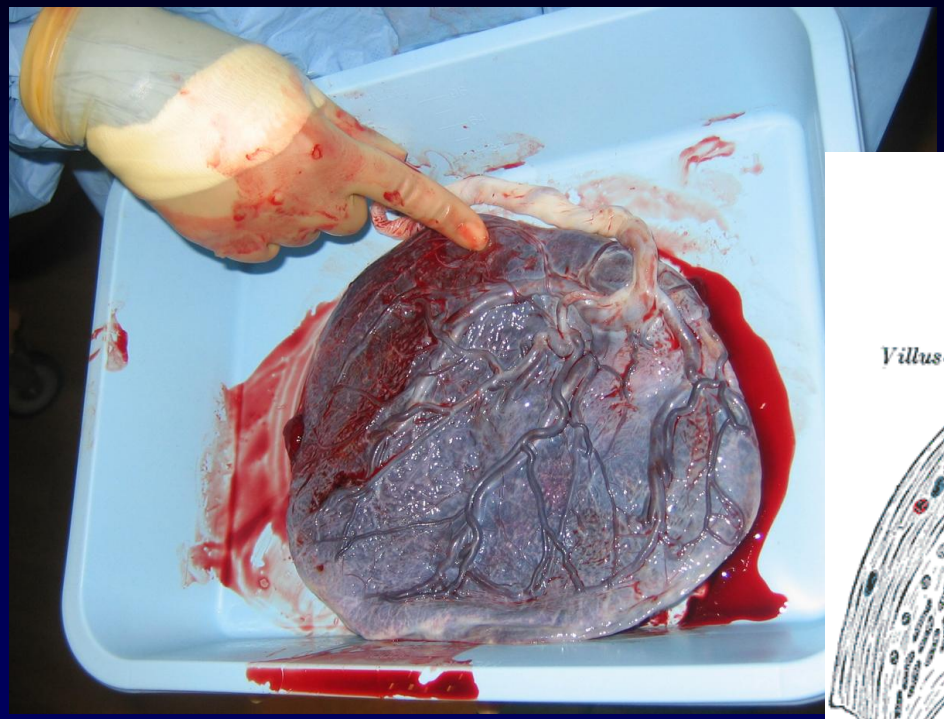


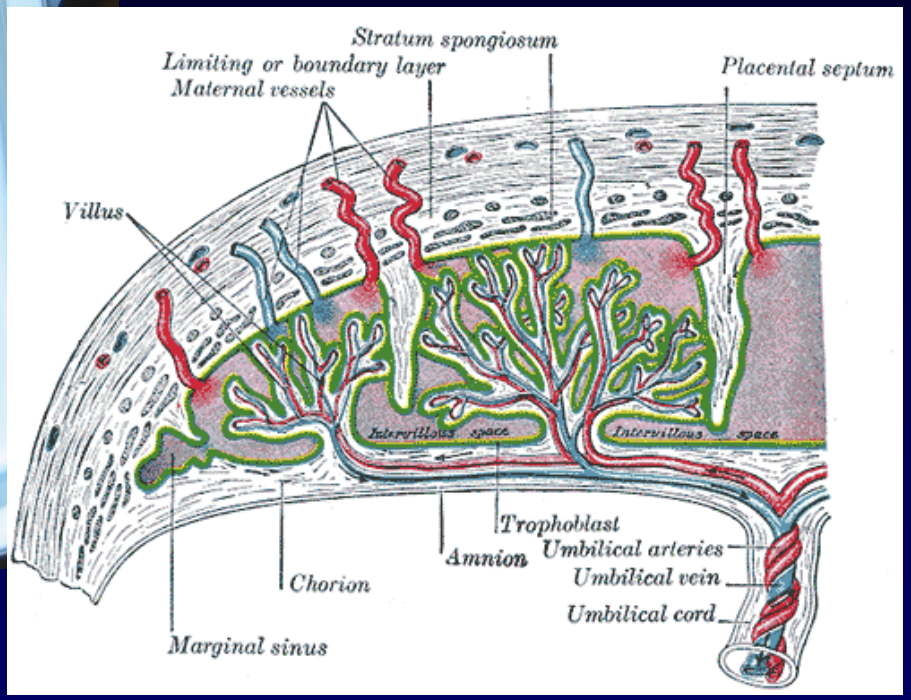
- EGFR mutation improves survival, but nullifies effect of treatment

Wackier side of genetics: Chimerism

Tetragametic Chimerism

Wackier side of genetics: Chimerism





Microchimerism

From the UT GSAF

Scott Hunicke-Smith, Ph.D. – Director
Jessica Wheeler – Lab Manager
Tony Hwang – PostDoc
Mani Singh – UGRA
Yen-Chia Ting - UGRA

Heather Deidrick – RA
Yvonne Murray – Administrator
Gabriella Huerta – RA
Terry Heckmann – RA
Matt Barnette– RA

The future of cancer treatment

- Researchers at St. Jude's and Dana Farber both predict sequencing of all incoming cancer patients in the next 2-3 years
- Applications will be:
  - Predicting tumor response (pt stratification)
  - Characterizing resistance to anticancer agents (this is the challenge in most metastatic solid tumors) and
  - Profiling the full spectrum of informative genetic/molecular alterations

Personalized cancer detection

- Personalized Analysis of Rearranged Ends (PARE) – Leary @ Johns Hopkins
- Do one mate-pair sequence analysis of the primary tumor
- Identify transpositions/gene fusions/etc. that are specific to that patient's tumor
- Use as a detection target for recurrence at least, or as a drug target
- Science Translational Medicine, 24 Feb. 2010

Pharmacogenomics & the FDA

- 13,000 drugs on-market
- 1,200 were reviewed for PGx labels
- 121 have them, and 1 in 4 outpatients use them

- Measurements and Main Results. Pharmacogenomic biomarkers were defined, FDA-approved drug labels containing this information were identified, and utilization of these drugs was determined. Of 1200 drug labels reviewed for the years 1945–2005, 121 drug labels contained pharmacogenomic information based on a key word search and follow-up screening. Of those, 69 labels referred to human genomic biomarkers, and 52 referred to microbial genomic biomarkers. Of the labels referring to human biomarkers, 43 (62%) pertained to polymorphisms in cytochrome P450 (CYP) enzyme metabolism, with CYP2D6 being most common. Of 36.1 million patients whose prescriptions were processed by a large pharmacy benefits manager in 2006, about 8.8 million (24.3%) received one or more drugs with human genomic biomarker information in the drug label.

- Conclusion. Nearly one fourth of all outpatients received one or more drugs that have pharmacogenomic information in the label for that drug. The incorporation and appropriate use of pharmacogenomic information in drug labels should be tested for its ability to improve drug use and safety in the United States.
- From: Lesko et. Al., “Pharmacogenomic Biomarker Information in Drug Labels Approved by the United States Food and Drug Administration: Prevalence of Related Drug Use”, Pharmacotherapy, Volume: 28 | Issue: 8 , August 2008.

Essential Ideas

- NGS interrogates populations, not individual clones
- Number of reads (sequences) ≅ 100x library molecules put into clonal amplification
  - MOLAR RATIOS matter!
  - Highly repeatable (from library through sequencing)
- Error rates are (very) high (compared to Sanger)
- NGS was a multi-disciplinary effort