# *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

2012

# *Outline*

- Next-gen sequencing: Background
- The details
  - Library construction
  - Sequencing
  - Data analysis

# *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
- Life Technologies SOLiD
  - http://media.invitrogen.com.edgesuite.net/ab/applications-technologies/solid/SOLiD_video_final.html

# *NGS enabling technologies*

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

# *The Details: Categories*

- Library Construction
- Sequencing
- Data Analysis

# Instruments: Roche Workflow

## Library Preparations – How to start

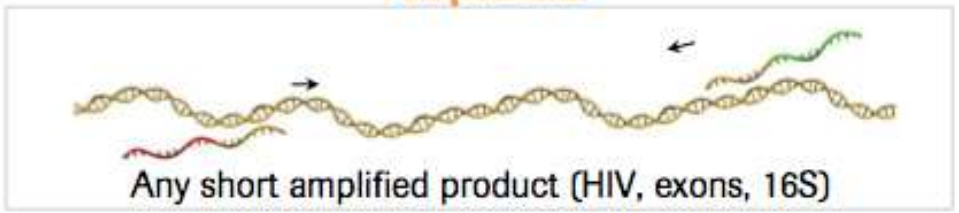
*Easy to use strategies for every sample type*

Roche

**Blunt-end Ligation**

ncRNA
Ancient DNA
Short ESTs
Short DNA fragments

**Amplicons**

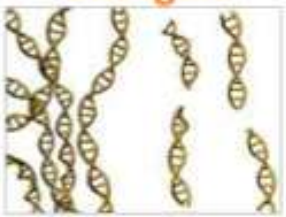


Any short amplified product (HIV, exons, 16S)

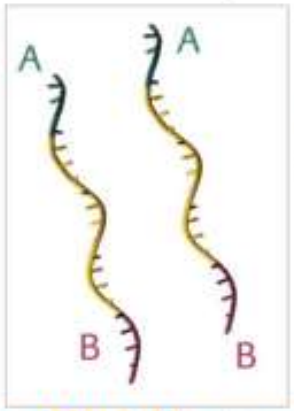
*4 h**

*0 h**

**Shotgun**



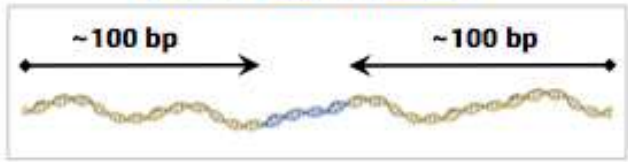
*4.5 h**




ssDNA Library

*14 h**

**Paired End Reads**




Whole genomes
Metagenomes
BACs
Long Range PCR
Full length transcripts

De novo assembly
Structural Variations

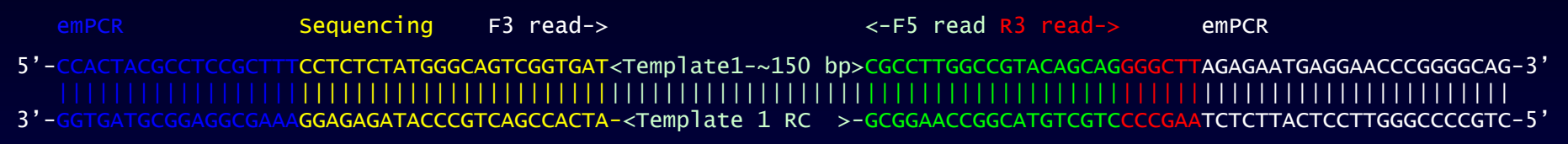
www.roche-applied-science.com

454 SEQUENCING

# *Read Types vs Library Types*

- Clear terms:
  - Fragment library
  - Mate-paired library
  - Paired-end read
- Ambiguous terms:
  - Paired-end library
  - Mate-paired read

# *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

# SE vs PE

**Table 3: Alignment statistics for Illumina PE and frag sequencing libraries**

| | Illumina Frag | Illumina PE |
|---|---|---|
| Total reads aligned | 33,524,973 | 37,832,835 |
| Total data aligned (Gbp) | 2.51 | 2.84 |
| Reads on target (%) | 67.62 | 78.0 |
| Duplicate reads (%) | 30.97 | 8.3 |
| Mean coverage (X)* | 24 | 52 |
| Median coverage (X)* | 20 | 40 |
| Targets hit (%) | 99.36 | 99.6 |
| Bases ≥1× Coverage (%) | 96.39 | 98.9 |
| Bases ≥10× Coverage (%)* | 71.33 | 90.8 |
| Bases ≥20× Coverage (%)* | 51.23 | 76.9 |

*Calculated after duplicate read removal.

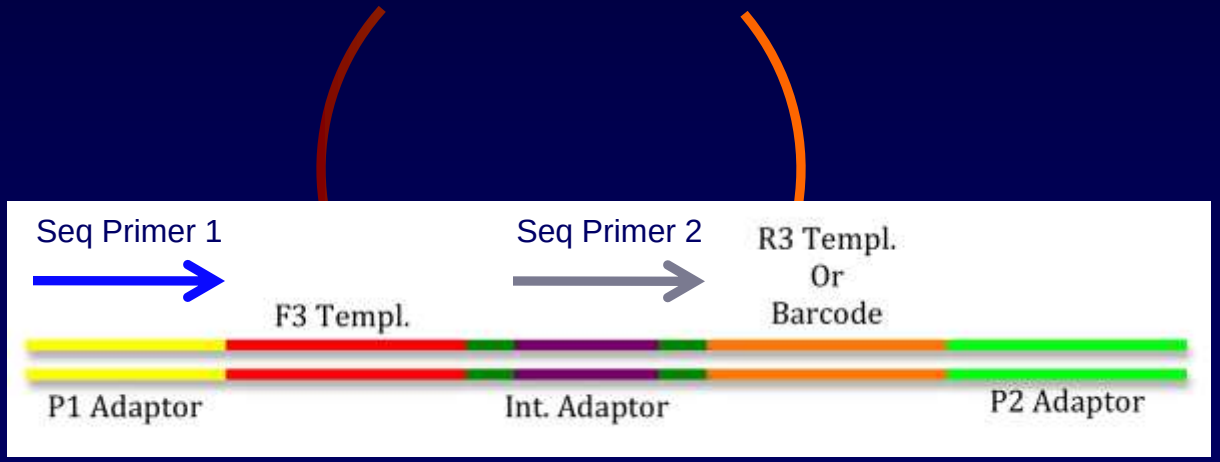
From: Bainbridge *et al*. Genome Biology 2010, 11:R62

# 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 |

# Mate-Pair Library Construction



# *Mate-Pair Library Construction*

- Shearing – size, size distribution
- Ligation biases (x4)
- Digestion – length, distribution
- Final gel cut

# *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
- Life Technologies SOLiD
  - http://media.invitrogen.com.edgesuite.net/ab/applications-technologies/solid/SOLiD_video_final.html

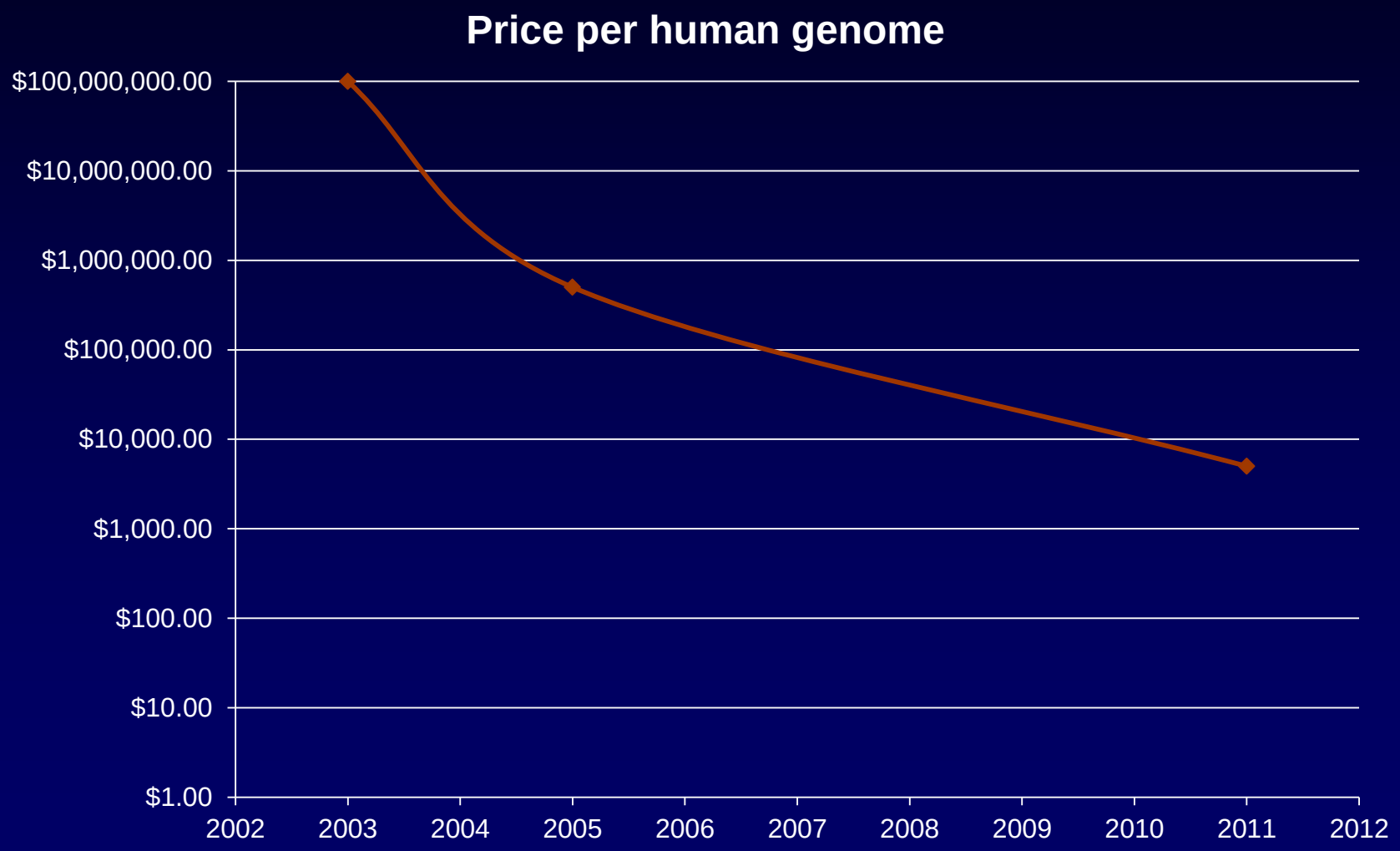
# *Essential Ideas*

- 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

# *Characteristics of SBS*

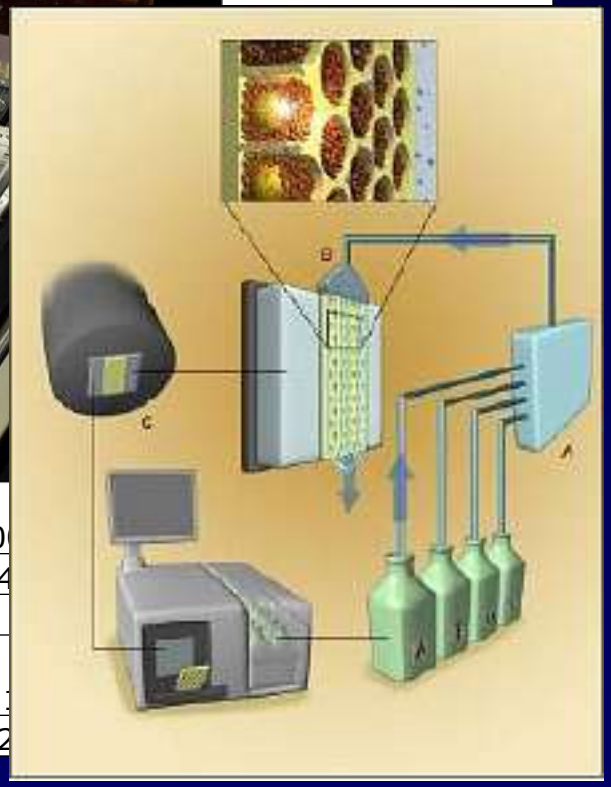
- Step-wise efficiency is <100%
  - Like inflation eating away at your savings
- This can be resolved by correcting “phasing”
  - This single software addition increased read lengths by ~10-fold
- Dominant error modalities can be predicted based on the technology
  - Fluor-term-nucleotide systems have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

# *Trajectory of Price*



# Instruments: How they work

| Step | Roche/454 | LifeTech SOLiD | Illumina HiSeq |
|---|---|---|---|
| Create ss DNA library | Shear, ligate adaptors, optional PCR | | |
| Segregate molecules | On polysty | | glass surface |
| Clonal amplification | emulsion P | | dge amplification |
| Fix colonies to seq. substrate | Deposit be picotiter p | | |
| Sequence | SBS: singl addition | | |
| Detect | Lumineser surface | | |
| | | | |
| Cost for 1 run | $ | | |
| Data from 1 run, megabase-pairs | 400 | 340 | |
| Time for 1 run, days | 1 | 14 | |
| | | | |
| Cost per megabase raw data | $17.11 | $0.1 | |
| Throughput, megabase/day | 400 | 242 | |

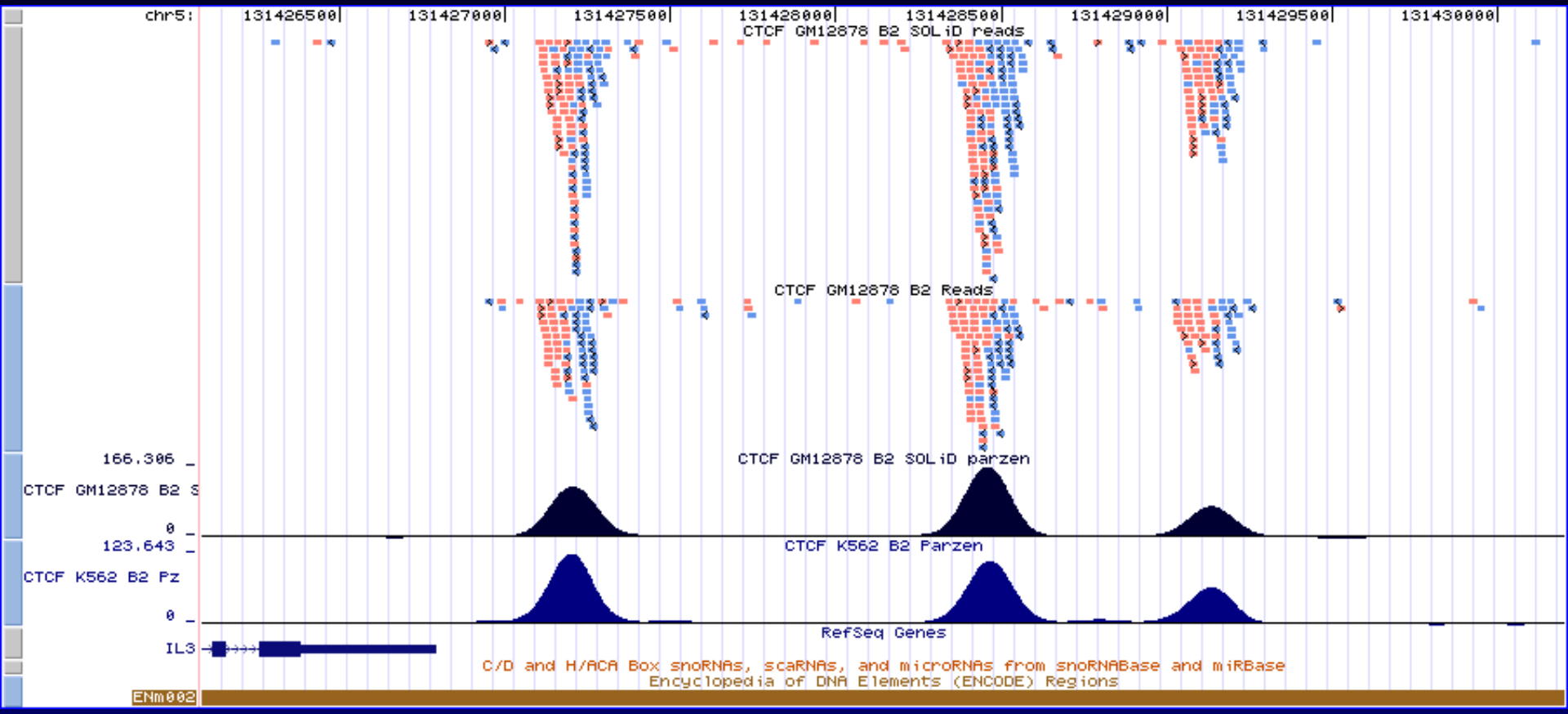


# *What it costs*

- Examples:
  - Gene expression profiling (INCLUDING array cost):
    - NimbleGen 12 samples on catalog, 72k probe, 4-plex arrays: ~$450 per sample from 1 ug total RNA or cDNA.
    - Illumina Human, Mouse, or Rat, 12 samples: ~$300 per sample from 100 ng total RNA
  - Deep Sequencing:
    - Illumina RNA-seq: 1 sample, 40 million read-pairs: $876
    - Illumina *de novo*: Draft sequence ~5 megabase bacterial genome (~25 MB raw sequence): ~$500

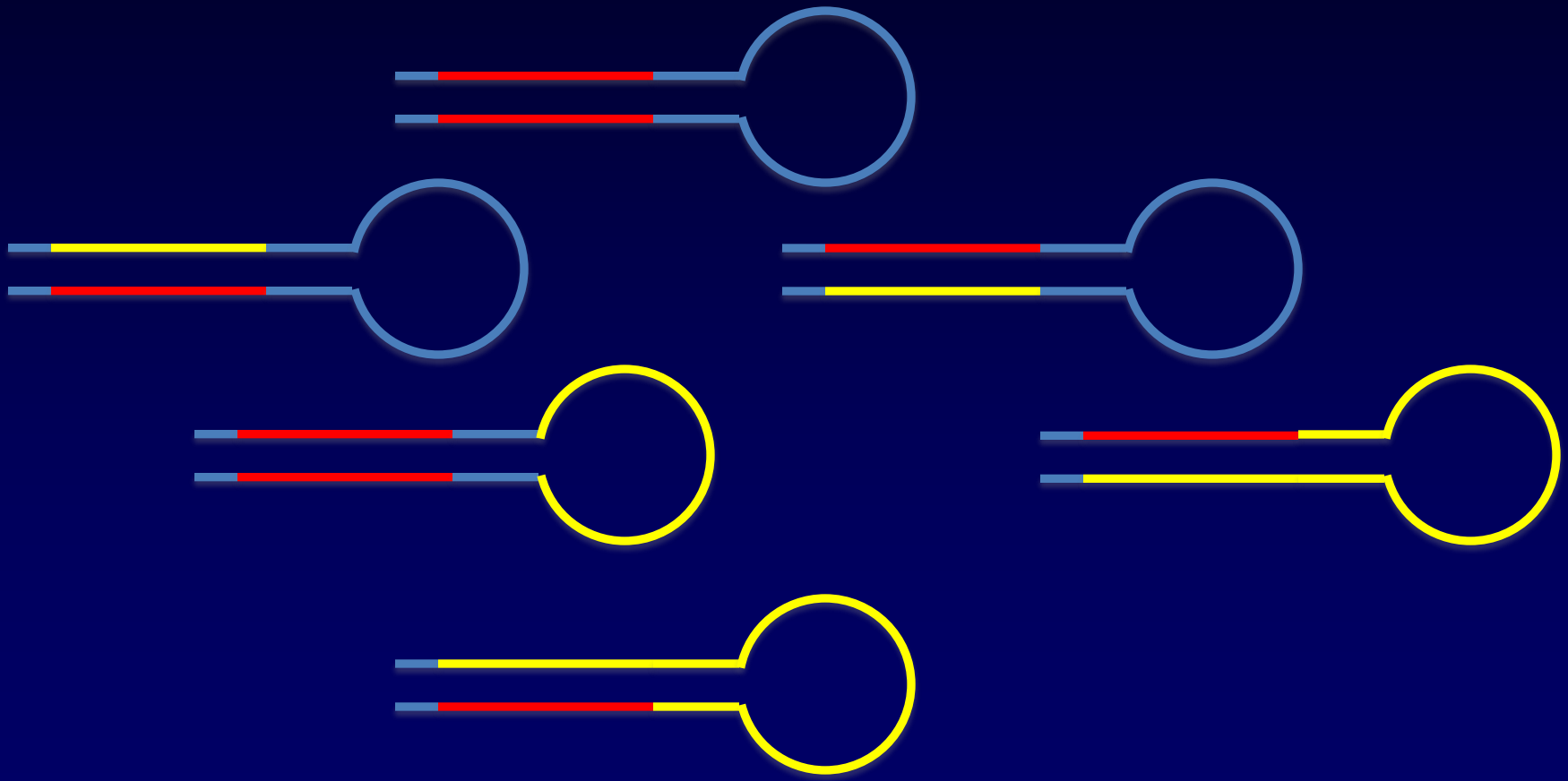
# *What, exactly, are we sequencing?*

# Good Example: ChIP-Seq



# *RNA/miRNA library*

- What's in YOUR library?



# *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

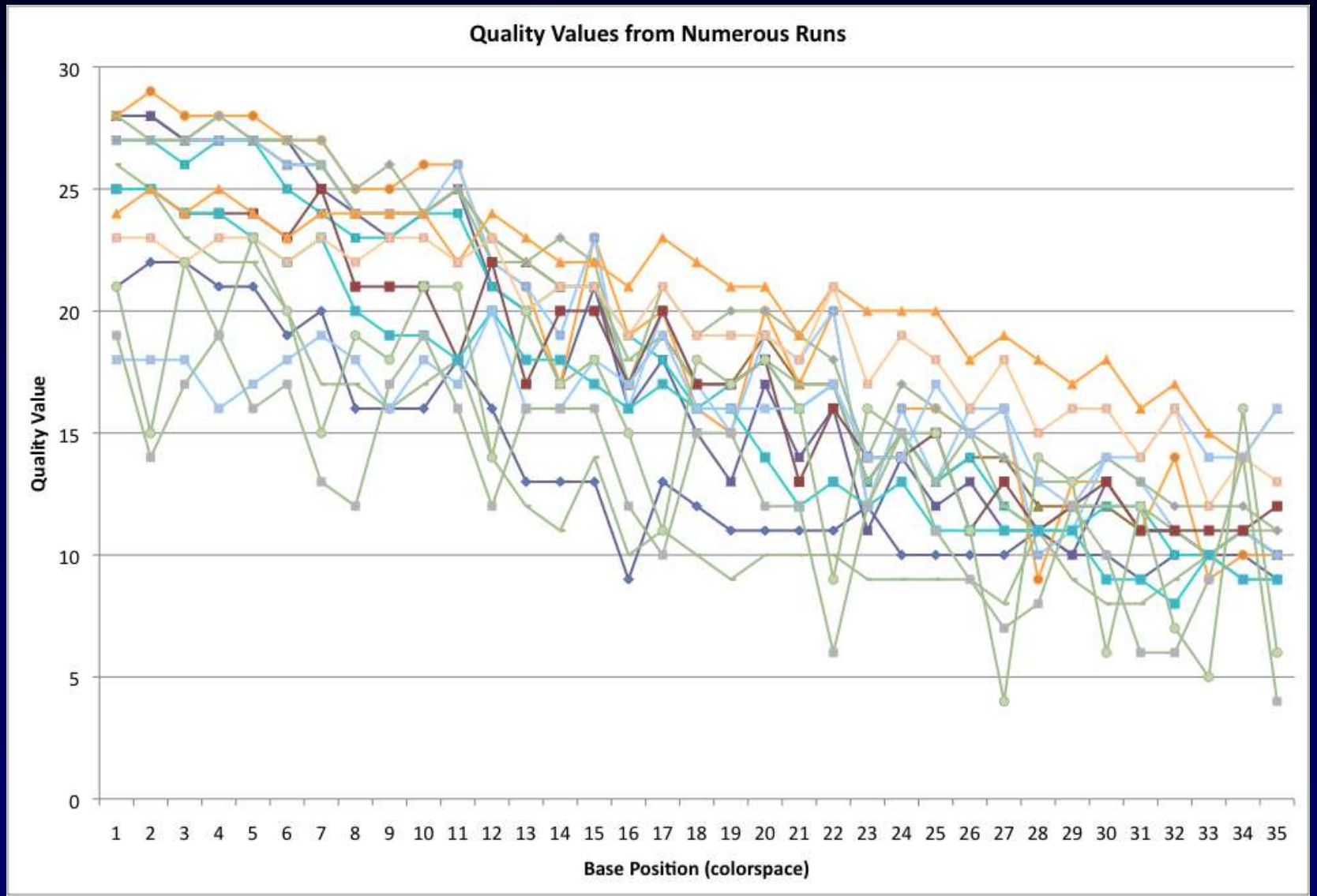
# *Sequencing*

- All instruments susceptible to:
  - Poor library quantitation leading to excessive templates (failure) or wasted space (more expensive)
  - Failures in cluster or bead generation (expensive)
  - Failures in sequencing chemistry (very expensive)
- Updates are very frequent

# *Instruments: Accuracy/Quality*

- “Error rate” - typ. to individual read
- Better: Mappable data

# Quality Values: Debated



# *Aligners/Mappers*

- Algorithms
  - Spaced-seed indexing
    - Hash seed words from reference or reads
  - Burrows-Wheeler transform (BWT)
- Differences
  - Speed
  - Scaleability 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

# *Some Comparisons*

Data courtesy Dhivya Arasappan, GSAF Bioinformatician

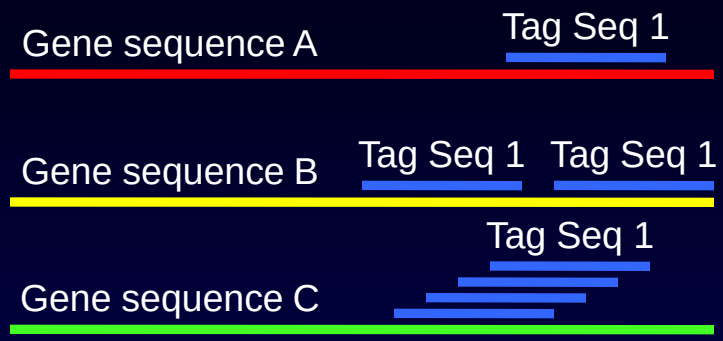
# *Informatics Pipelines: RNA-seq*

- 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)

# *Mappers – Too Many, Too Few*

| | **mapreads** | **MAQ** | **SOAP2** | **Bowtie** | **SHRiMP** | **BWA** |
|---|---|---|---|---|---|---|
| | | | | | | |
| Relative CPU time required | 100 | 10 | 1 | 1 | 10000 | 1 |
| Colorspace correction | yes | no | no | no | yes | no |
| Indels | no | no | no | no | yes | no |
| Uses base QV | yes | yes | no | yes | no | yes |
| Creates map QS | no | yes | no(?) | no | yes | yes |
| Relative memory used | | | | | | |

# *Rule Set Example*




- Basis for definition of “hit”…
- Accept all hits
- Collapse intergenic non-unique
- Select random non-unique
- Select only unique
- Apply stat model to non-unique
- Summarize by gene, exon (gene model?)

# Comparison of Short-Read Mappers & Filters

Mapping along normalized gene length – effects of post-mapping filters.

Fig 1a: Bowtie raw output,max.100 hits per tag (No filter)

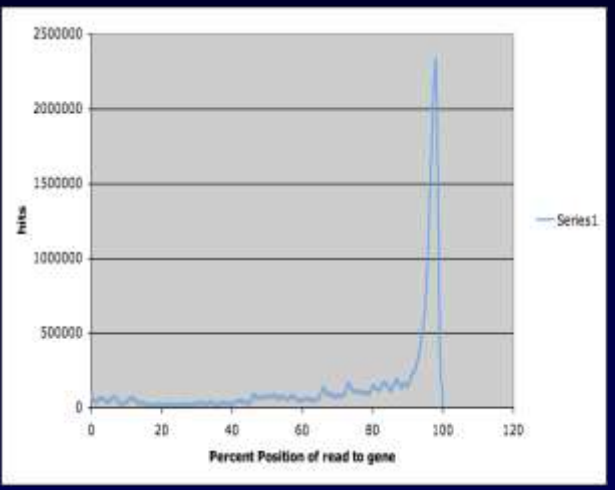



Fig 1b: Bowtie output, max.25 hits per tag, 3mis, nontiling, max. coverage of 1%

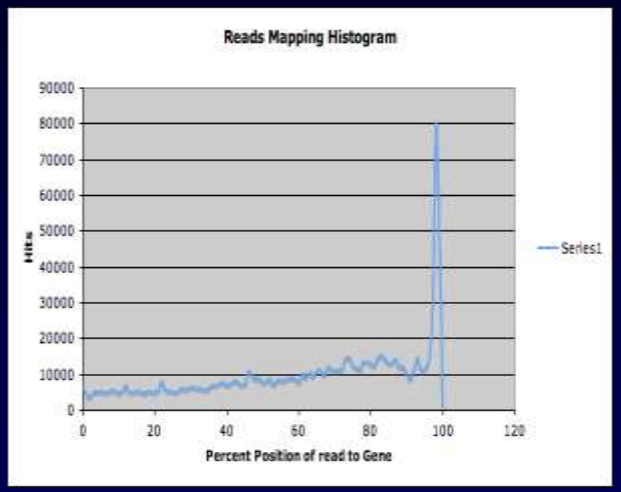



Fig 1c: Bowtie output,1 hit per tag, 3mis, nontiling, max.coverage of 1%, no polyA tails

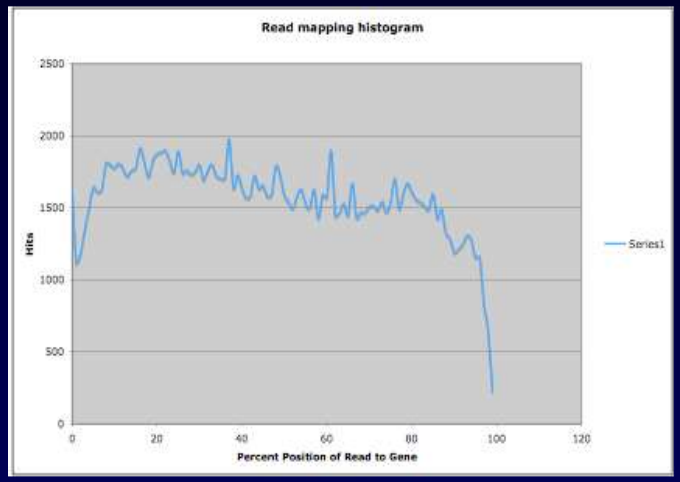



Fig 2a:SOAP2 raw output (No filter)

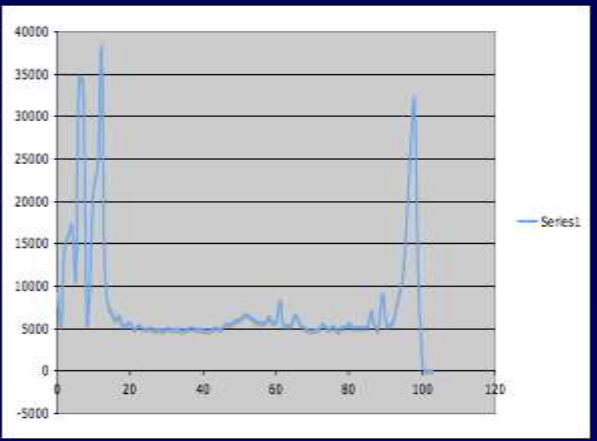



Fig 2b: SOAP2 output, 1 hit per tag, 3mis, nontiling, max. coverage of 1%

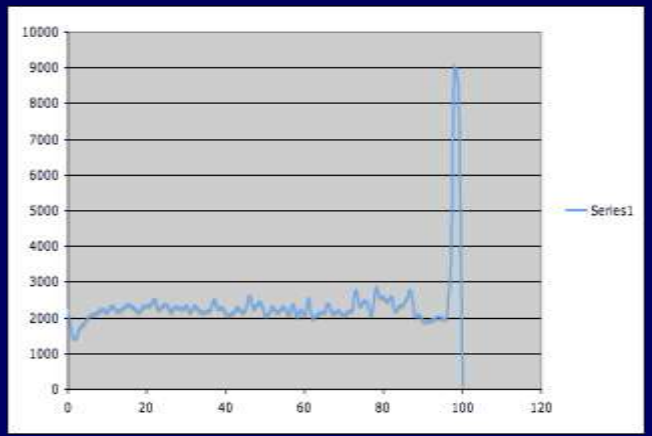
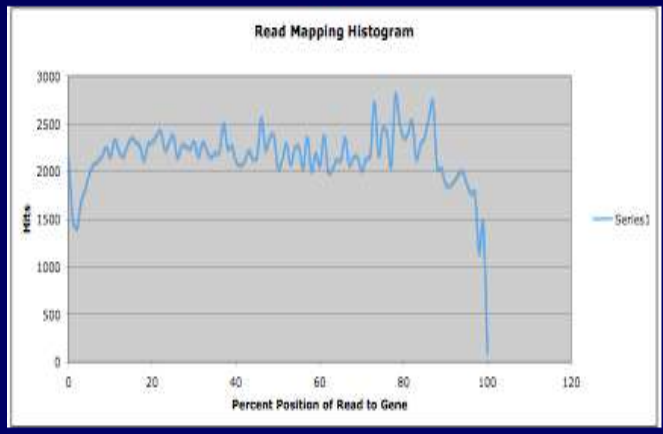
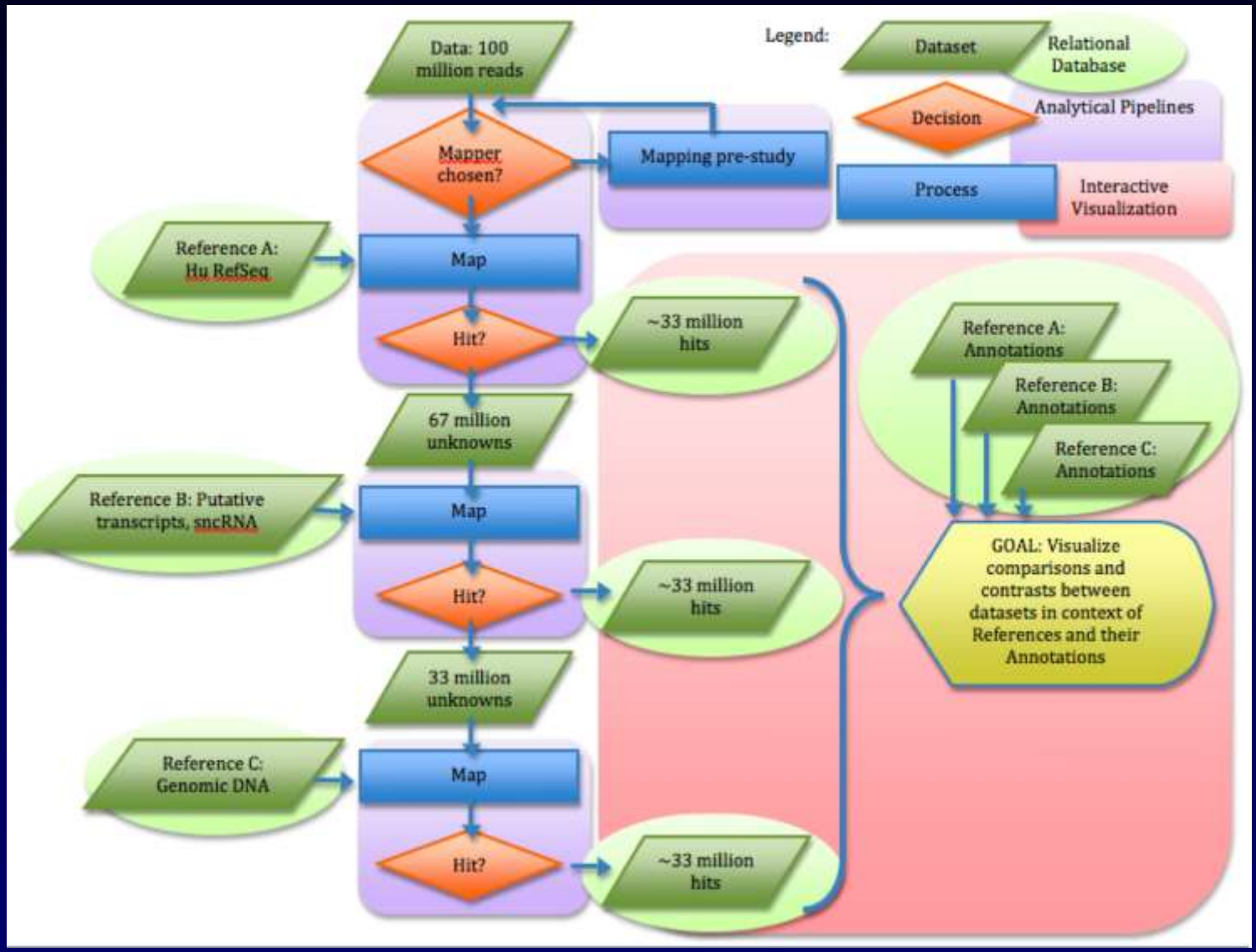


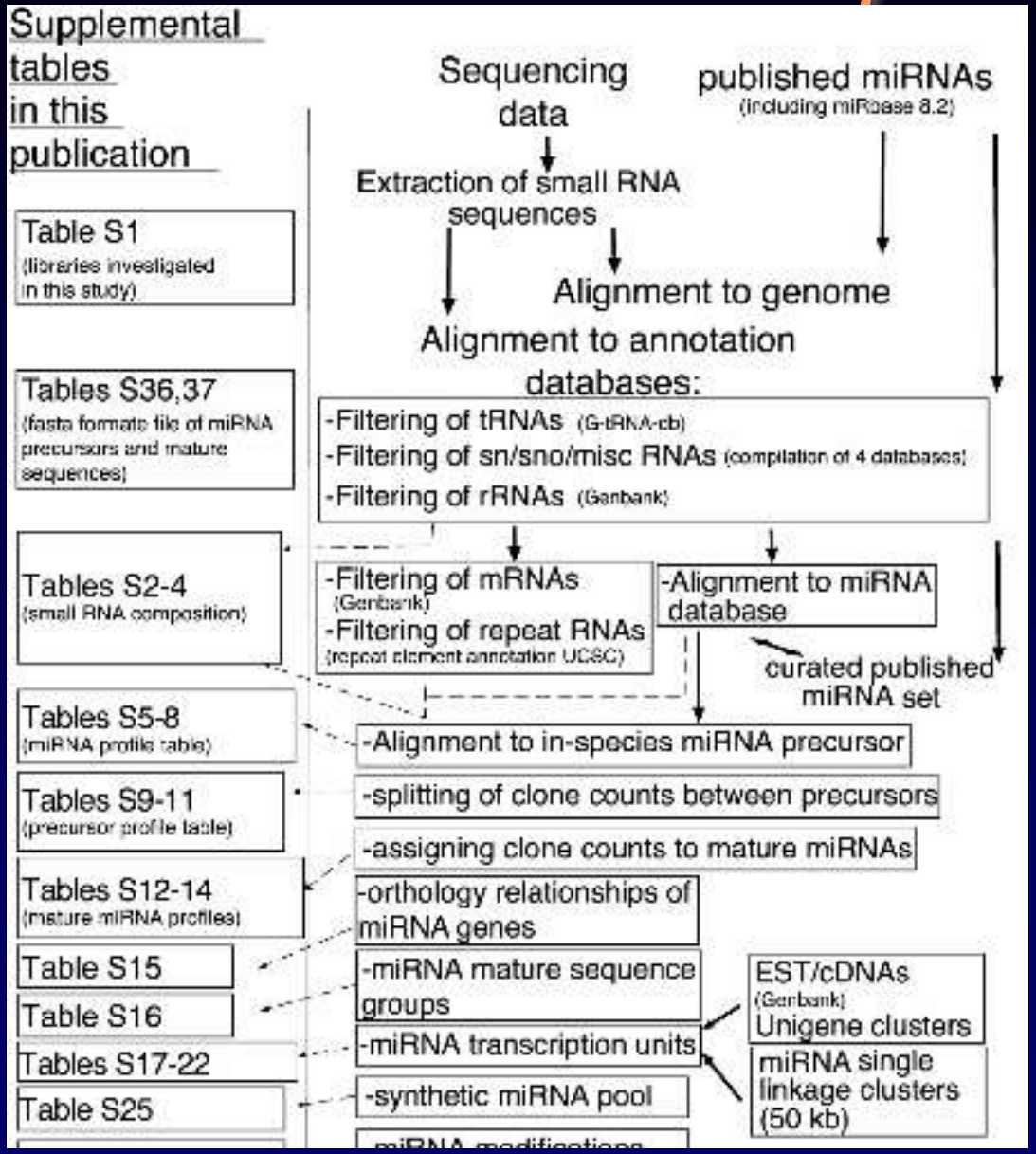

Fig 2c: SOAP2 output,1 hit per tag, 3mis, nontiling, max.coverage of 1%, no polyA tails



# Data Analysis Workflow: RNA-Seq



# 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

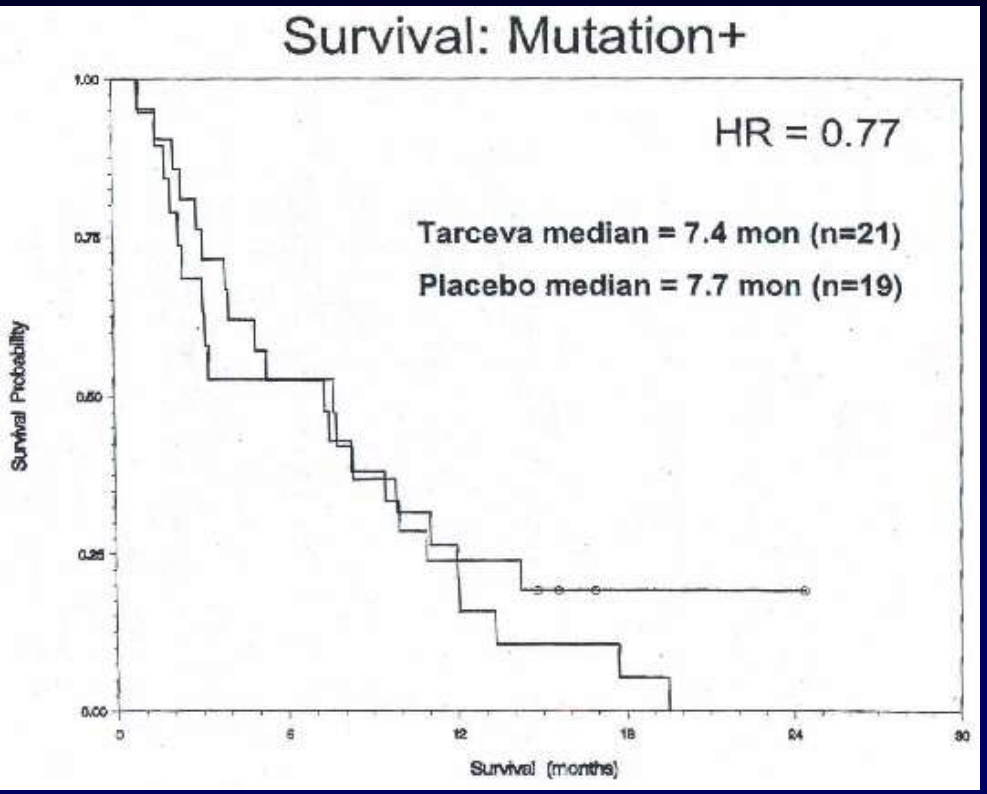
# *TACC: A Joy in Life*

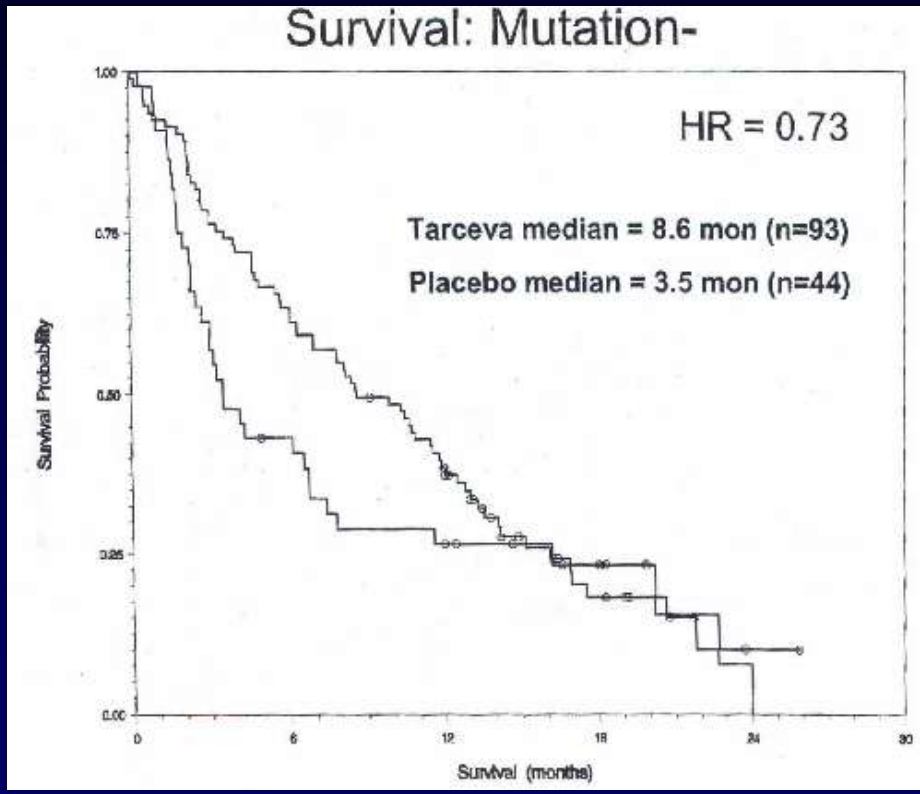
- RANGER: 63,000 processing cores, 1.73 PB shared disk
- LONESTAR: 5,840 processing cores, 103 TB local disk
- RANCH & CORRAL: 3.7 PB archive
- Typical mapping of 20e6 reads:
  - 20 hours on high-end desktop
  - 2 hours at TACC

# *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*



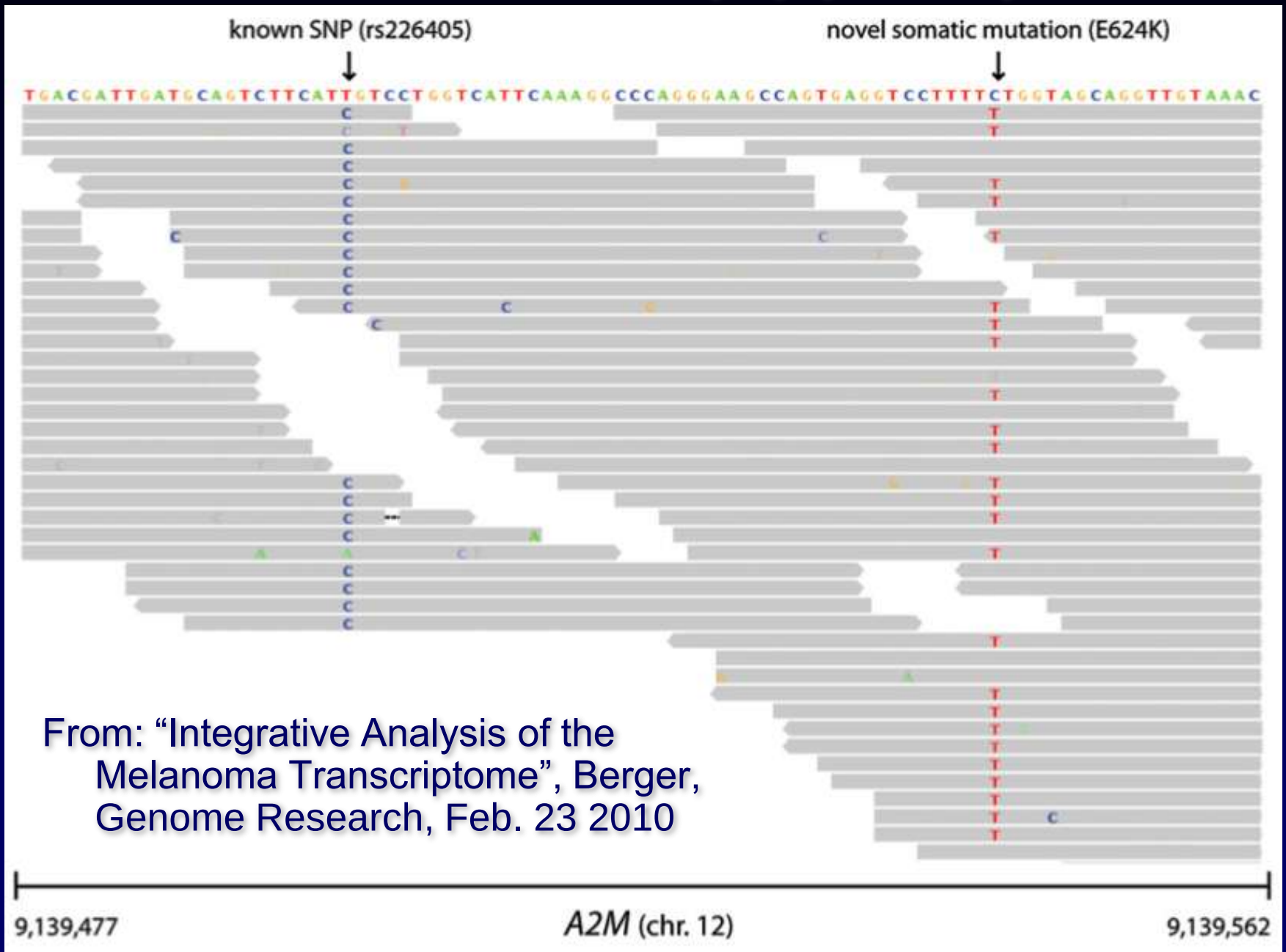




- EGFR mutation improves survival, but nullifies effect of treatment

# *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

# Real (applied) data




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

# *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.

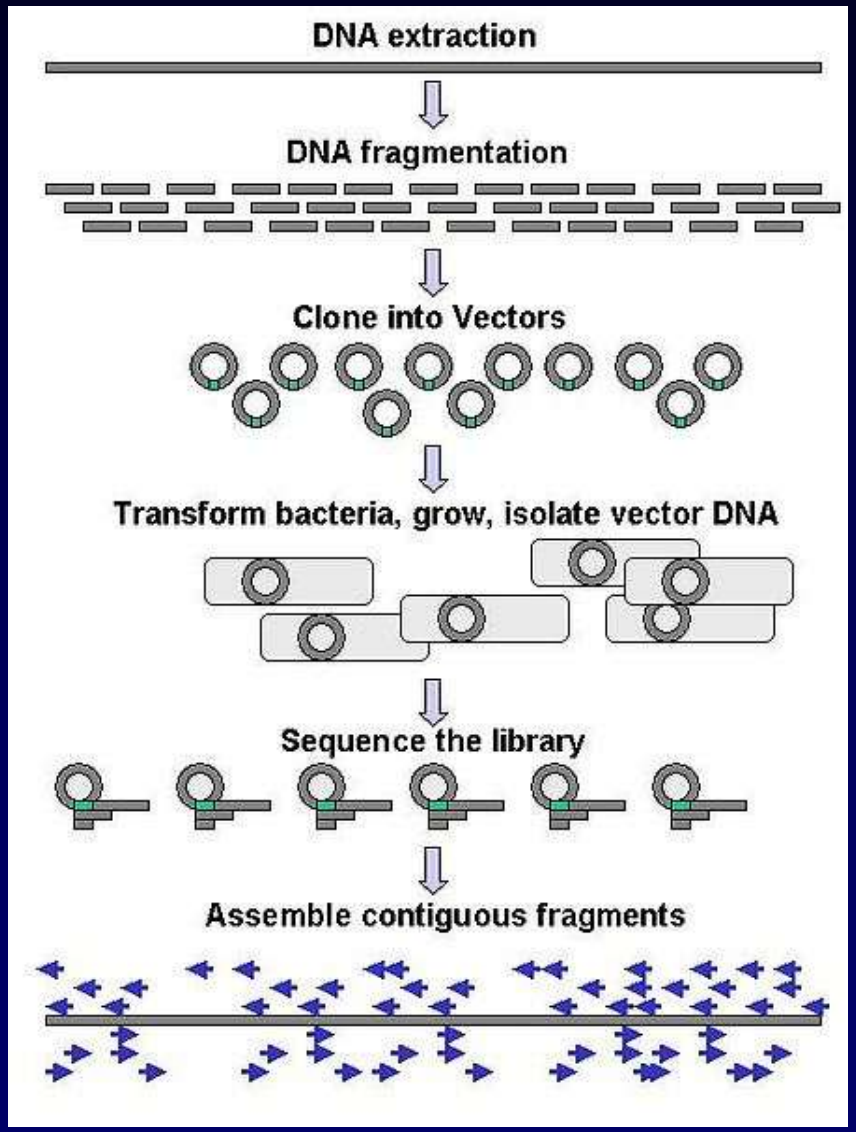
# *Epidemiology*

- Metagenomics to be specific:
  - Key point: survey of microbial communities by culture is biased; survey by sequencing is completely unbiased
  - Can thus survey any biological milieu:
    - Individual: sinus, skin, gut, etc. either singly or in aggregate
    - Survey water supply, environmental samples
    - Corporate: survey raw sewage streams at sentinel locations to monitor outbreaks

# *Key Take-Home Concepts*

- Access to DNA and RNA-based information will be trivial in the next 10 years
- Understanding of genome information will be take a lot longer
- Consider bioinformatics in your curriculum and your career

# Conventional Sequencing




From “DNA Sequencing” - Wikipedia

# *From the UT GSAF*

Scott Hunicke-Smith, Ph.D. – Director
Dhivya Arasappan, M.S. – Bioinformatician
Jessica Wheeler – Lab Manager
Melanie Weiler – RA
Jillian DeBlanc – RA
Heather Deidrick – RA
Yvonne Murray – Administrator
Gabriella Huerta – RA
Terry Heckmann – RA
Margaret Lutz – RA

# *Preliminaries*

- All the world's a sequence…
  - *De novo* sequencing
  - Re-sequencing: SNP discovery, genotyping, rearrangements, targeted resequencing, etc.
  - Regulatory elements: ChIP-Seq
  - Methylation
  - Small RNA discovery & quantification
  - mRNA quantification: RNA-Seq
- Combination data:
  - mRNA -> cDNA -> nextgen =
    - Gene expression
    - Splice variants
    - SNPs