RUBICON GENOMICS

breaking the sample barrier

QUALITY CONTROL AND PRE-QUALIFICATION OF NGS LIBRARIES MADE FROM CLINICAL SAMPLES

ABRF 2013 Satellite Workshop

March 2, 2013

Dr. John Langmore, CSO

langmore@rubicongenomics.com

RUBICON GENOMICS
breaking the sample barrier

Challenges for NGS of Clinical DNA Samples

Sample	Use	Quantity (ng)	Content of Useful DNA	Failure Mode
Tissue and FFPE	Cancer testing & biomarker discovery	10 – 300	1 – 50%	• Small, variable amount • Variable size (100–1000 bp) • Chemical damage (misincorp)
Biofluids-plasma, serum, urine, CSF	Cancer and prenatal testing	1 – 20	0.1 – 20%	• Small, variable amount • Small size (70 – 200 bp)
Single cells or rare events	CTC, FNA, prenatal	0.003 – 3	0.1 – 100%	• Small, variable amount • Molecular loss • High background
Functionally enriched [ChIP, meDIP, hmeDNA]	Epigenetic markers	0.001 – 10	1 – 100%	• Small, variable amount • Molecular loss • High background
Bisulfite-converted DNA	Diagnostics & biomarker discov.	1 - 100	0.1 – 100%	• Small, variable amount • Molecular loss • Chemical damage
Target capture	Genetic testing for disease	1 - 100	highly variable	• Small, variable amount • Molecular loss

RUBICON GENOMICS
breaking the sample barrier

Challenges and Possible Solutions

Challenge	Possible Solutions
Degraded DNA	Improve efficiency of internal and terminal repair Increase efficiency of ligation
Small fragment size	Improve amplification of small fragments
Very small amounts of DNA	Increase efficiency of ligation Increase fraction of useful reads
Variable input DNA amount	Increase dynamic range of library synthesis and amplification without compromise of quality
Molecular loss	Minimize sample transfers and bead binding/release
High background	Minimize sample transfers Clean reagents and reactions Increase efficiency of ligation
Chemical Damage	Improve repair reactions
Long time to result	Decrease time of library preparation Increase throughput Increase simplicity for automation

RUBICON GENOMICS
breaking the sample barrier

4

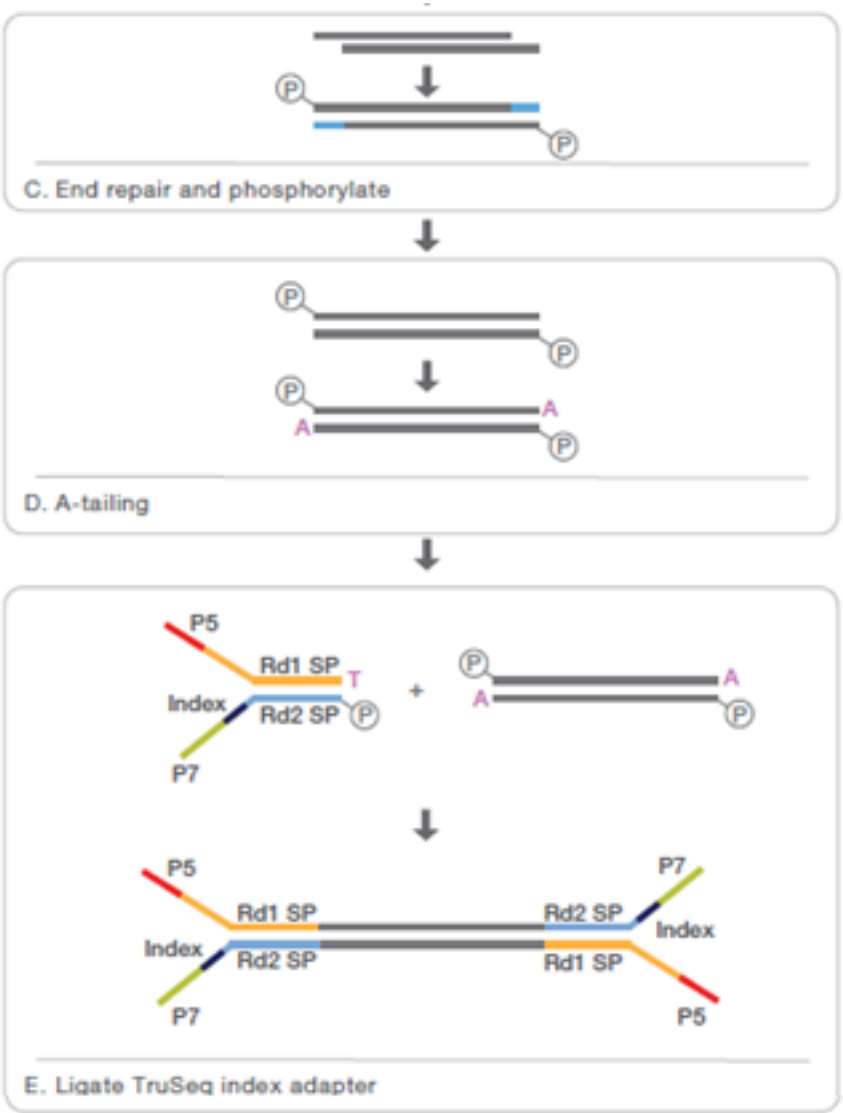
Methods of Library Production

- Illumina TruSeq "Y" Adaptor Ligation Libraries
- Rubicon "Stem-Loop" Adaptor Ligation Libraries
- Illumina Nextera "Tagmentation" Libraries
- Homebrew and Other Types of Libraries
- Legacy WGA Technologies Followed by Cleavage and Ligation

Note: all libraries require PCR amplification except if DNA input is very large (rarely the case with clinical samples). Our data indicates that PCR does not degrade performance

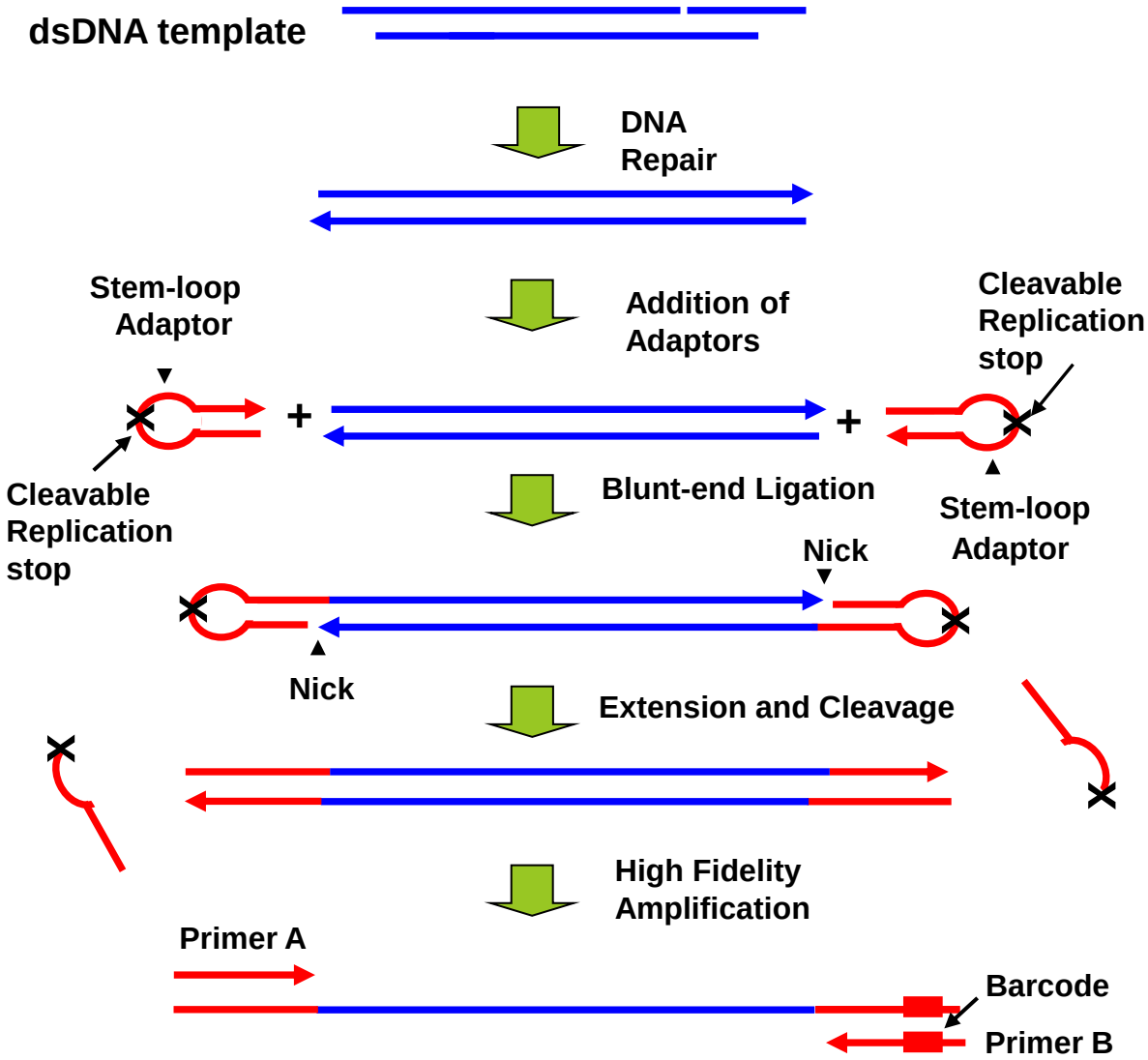
RUBICON GENOMICS
breaking the sample barrier

Solexa/Illumina "Y" Adaptors



- Characteristics of Illumina libraries
  - Depend on A-tailing to reduce adaptor dimers
  - Adaptors contain unstable secondary structure with exposed ss tails that contribute to background
  - Step-by-step chemistry requires multiple intermediate transfers and purifications

Rubicon ThruPLEX-FD Stem-Loop Adaptors



US patent 7,803,550 and international equivalents

Characteristics of ThruPLEX

- Improved repair
- Background reduced using ds adaptor no ss tails
- High-efficiency blunt-end ligation
- Adaptor-adaptor ligation reduced using blocked 5' ends
- Background reduced by destroying unused adaptors after ligation
- Time is reduced using compatible buffers and multiplexed enzymatic reactions that do not require intermediate purifications

Illumina (Nextera) "Tagmentation" Libraries

- Tagmentation
  - Rapid, simple enzymatic process to covalently break and add adaptors in one reaction
  - Advantage: Works with HMW DNA without pre-fragmentation
  - Disadvantage: Does not work with LMW DNA as found in clinical samples (e.g., FFPE, plasma, serum, urine, etc.)
  - Limited to specific amounts of input DNA
  - More sequence bias than best ligation-based libraries

Homebrew and Other Library Preps

- Characteristics of homebrew and some commercial library kits
  - Uncertain adaptor structure and chemistry
  - Some have high human background
  - Usually require multiple intermediate purifications to remove enzymes, change buffers, and remove free adaptors and primers
  - Are not optimized over range of input
- Homebrew libraries are not developed using design control, manufactured cGMP using components with FTO in diagnostics or subjected to consistent QC, and are therefore not easily compared with other labs or transitioned to CLIA or IVD applications.

RUBICON GENOMICS
breaking the sample barrier

9

Meeting the Sample Challenge: Metrics for Success

1) Workflow metrics

2) Performance metrics

RUBICON GENOMICS
breaking the sample barrier

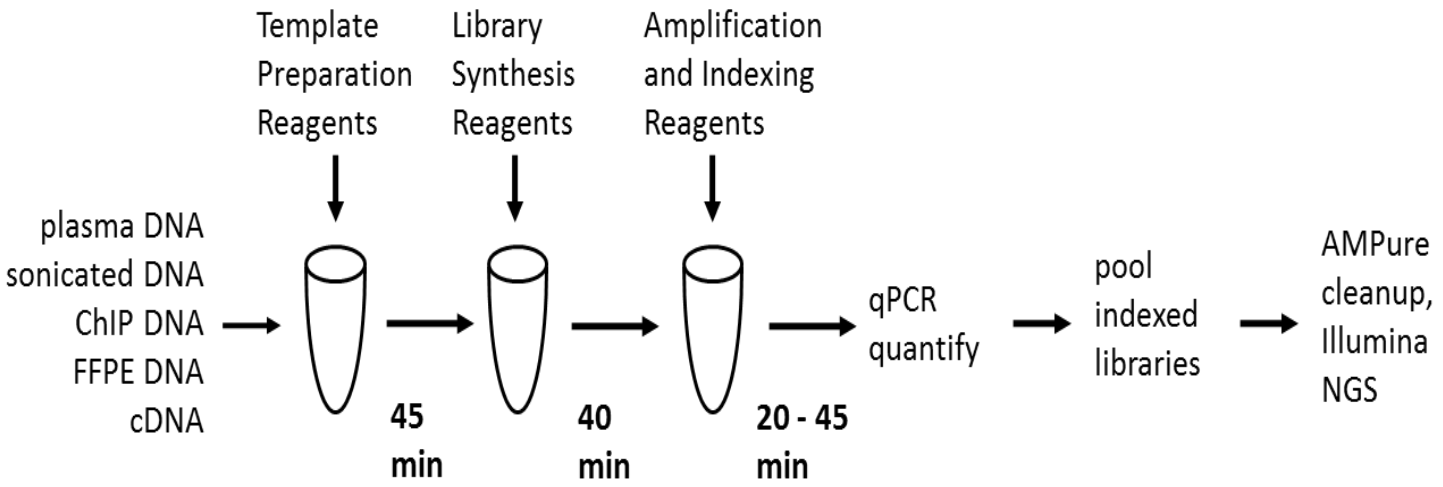
Workflow Metrics (Number Game)

1) Steps
2) Sample transfers
3) Pipetting steps
4) Reagents required (especially user-supplied)
5) Minutes or hours of hands-on time
6) Hours to result
7) Samples per day
8) Dollars per sample

RUBICON GENOMICS
breaking the sample barrier

Example: ThruPLEX-FD for DNA-seq and cDNA-seq on Illumina

- 15 – 100 X more sensitive than "Y" adaptors
- More uniform coverage than TruSeq or Nextera
- Highest gDNA dynamic range
- Significant improvements to diversity of plasma, ChIP, FFPE, and cDNA libraries
- All components included in kit. 3 tubes with enzymes, 3 buffers, water, plus 12 indexing primers (soon 96)
- Automatable 1-tube, 2-hr, 3-step protocol with no intermediate clean ups.
- High throughput (192 samples/day)
- Single-cell kits in development for Illumina and Ion Torrent



Workflow metrics

One tube with no intermediate purifications
Eliminates molecular loss and reduces cross-contamination
One technician can prepare 192 indexed samples/day
Fixed reagent concentrations and protocol from <1 pg to 100 ng

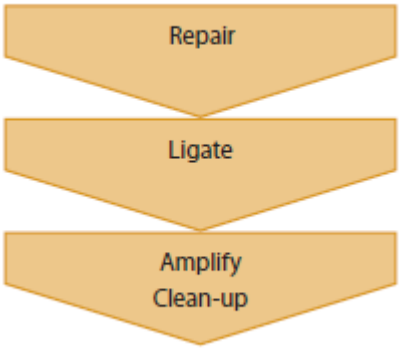
RUBICON GENOMICS
breaking the sample barrier

Workflow Comparisons Between "Y" and "Stem-Loop" Adaptor Libraries

SAMPLE INPUT

0.001 – 100 ng

ThruPLEX-FD



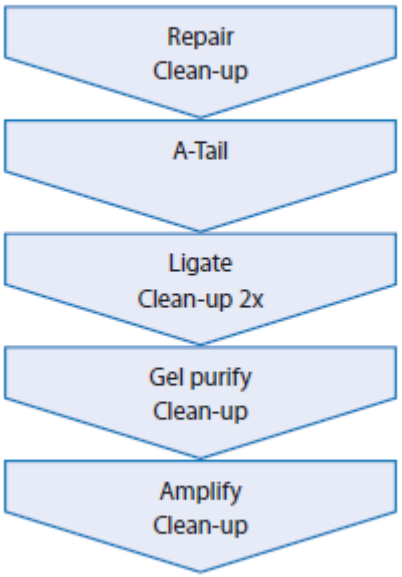
NGS-READY DNA

2 hours

Daily Output= 192/tech

10 ng - 10 µg

Illumina TruSeq™ DNA LS Sample Prep Kit*



NGS-READY DNA

7 - 11 hours

Daily Output= 12/tech

Workflow Advantages

- ✓ Faster
- ✓ Simpler
- ✓ Less hands-on time required
- ✓ No sample transfers
- ✓ Less risk of contamination
- ✓ Higher throughput
- ✓ Easily automated

Resultant Performance Improvements

- ✓ Use less quantity of samples
- ✓ Make better use of degraded samples
- ✓ Improve quality of NGS data
- ✓ Lower background
- ✓ Test more samples

RUBICON GENOMICS
breaking the sample barrier

Need for Performance Metrics

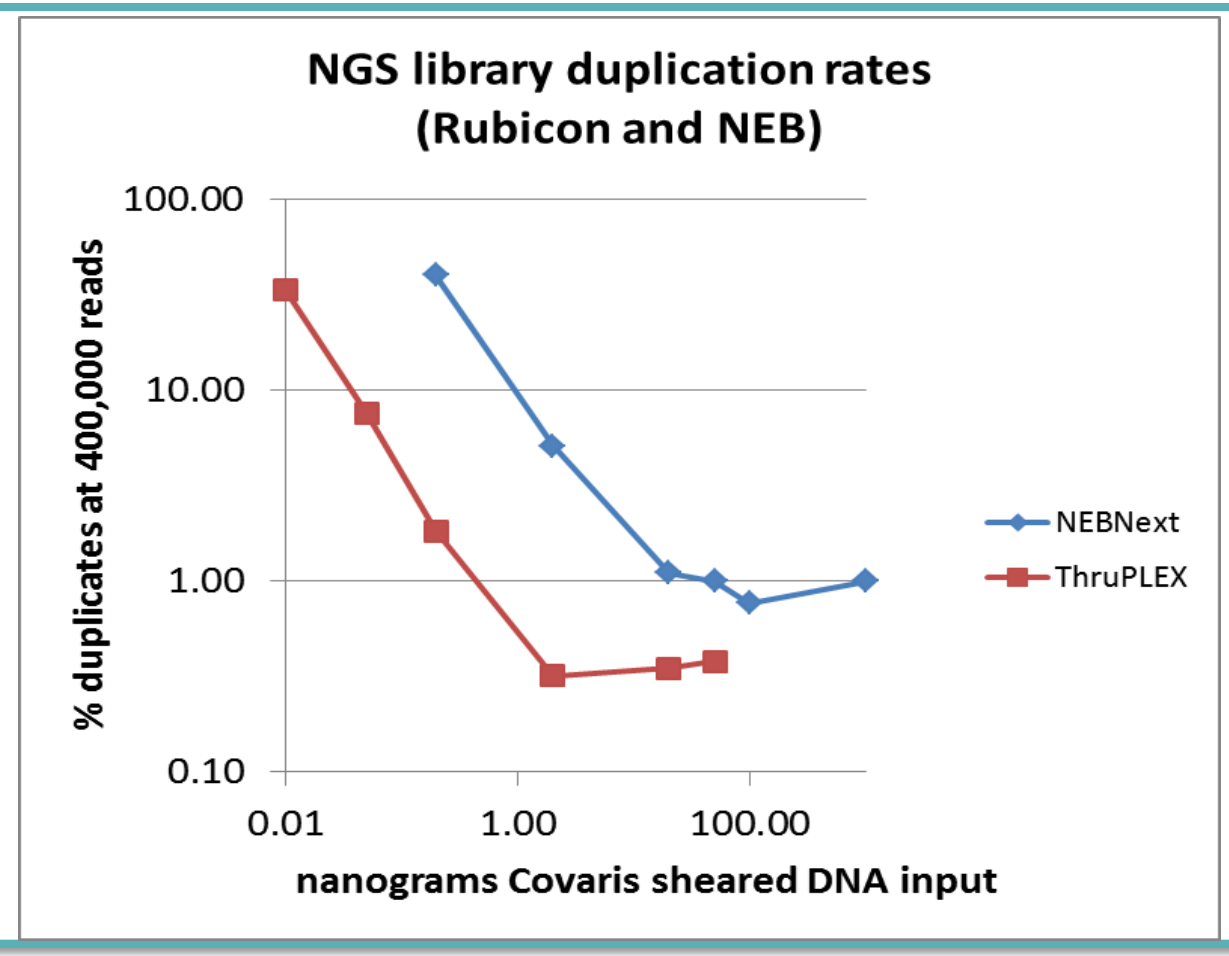
1) For kit producer
    a) Optimization of chemistry during product development
    b) Manufacturing QC

2) For consumer
    a) Choosing and QC of kits for specific applications
    b) Pre-screening clinical samples before deep sequencing

RUBICON GENOMICS
breaking the sample barrier

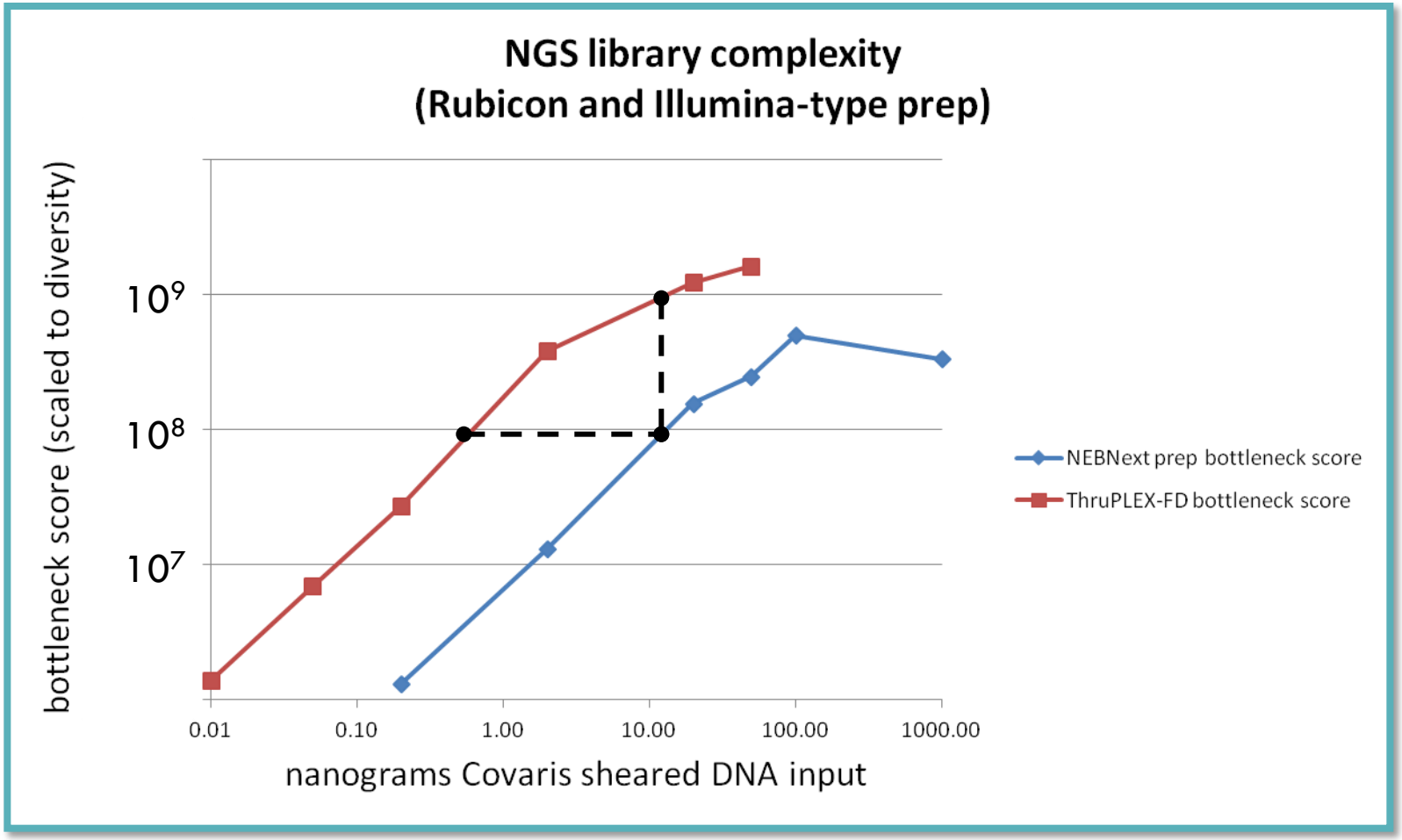
Performance Metrics for Libraries

1) Library diversity (for constant input)
2) Sensitivity to amounts of gDNA and cDNA inputs (for constant diversity)
3) GC bias (normalized GC coverage)
4) Low resolution coverage
5) Amplification background and NGS background (unmapped reads)
6) Sensitivity of first five metrics to input gDNA and number of amplification cycles
7) Lot-to-lot variation in performance
8) Compatibility of library prep with upstream and downstream applications
9) Concordance between low-input and “gold-standard” results



Metric #1: Library diversity (via read duplications)

% duplicate reads is a useful metric of randomness of libraries over a wide range of conditions and samples, however comparisons require down sampling DNA to a constant number of reads. Data from University of Michigan at 400K reads.



Metric #1: Library diversity
Metric #2: NGS sensitivity

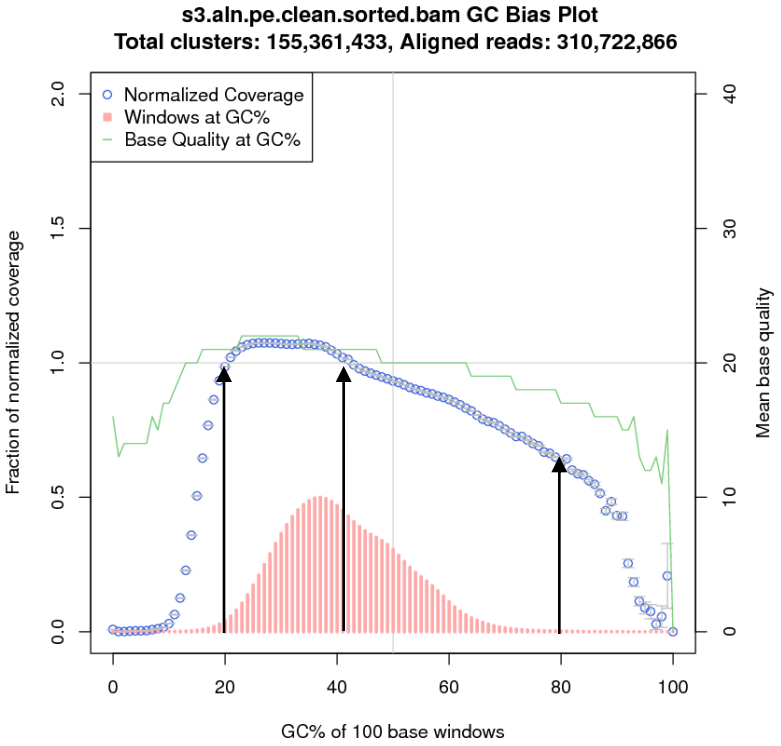
Covaris-fragmented human DNA was prepared and sequenced by a university core lab using an NEBNext prep and ThruPLEX-FD prep. Library complexity as measured by DNAnexus "bottleneck score" or diversity calculations shows ThruPLEX has ~10X higher diversity (at 10 ng input) or 30X higher sensitivity (at 10e8 diversity). These metrics can be evaluated at <100K reads

ThruPLEX-FD High Coverage (U. Washington) (High-Seq2000 with ~300M reads)

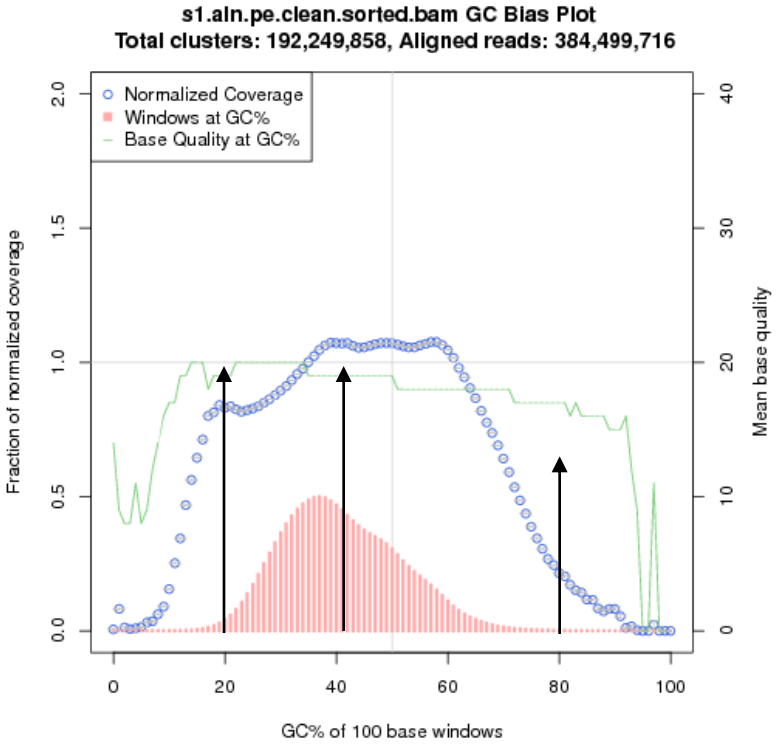
Lane	Input mass (ng) in 10uL	Insert size	Duplicate Fraction	GC Bias	Variants Called (Ti/Tv Ratio)
3	20.21	~300	0.067156	ok	1.99
4	22.21	~300	0.047533	ok	1.99
5	15.15	~300	0.074519	ok	1.98
6	32.59	~300	0.040543	ok	1.97
7	22.35	~300	0.070965	ok	1.98
8	45.64	~300	0.048136	ok	1.98

Metric #3: GC Bias at 300M Reads (U. Wash.)

Out-of-the box ThruPLEX-FD
(20 ng input)

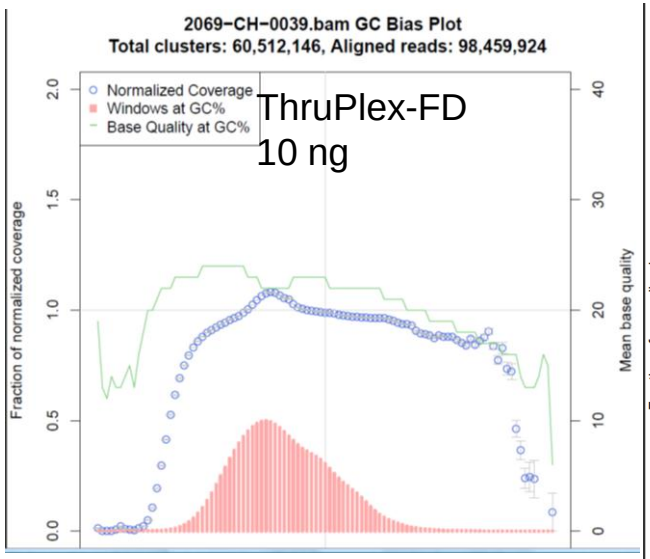


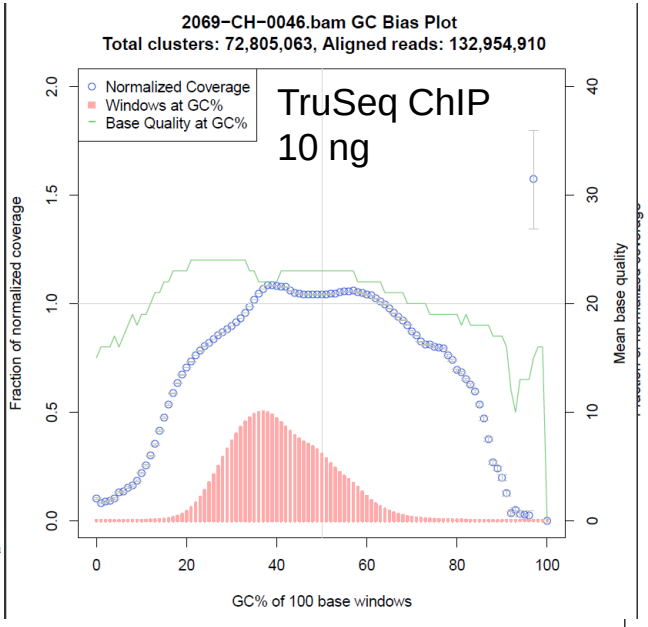
UW optimized homebrew
(1 microgram input)

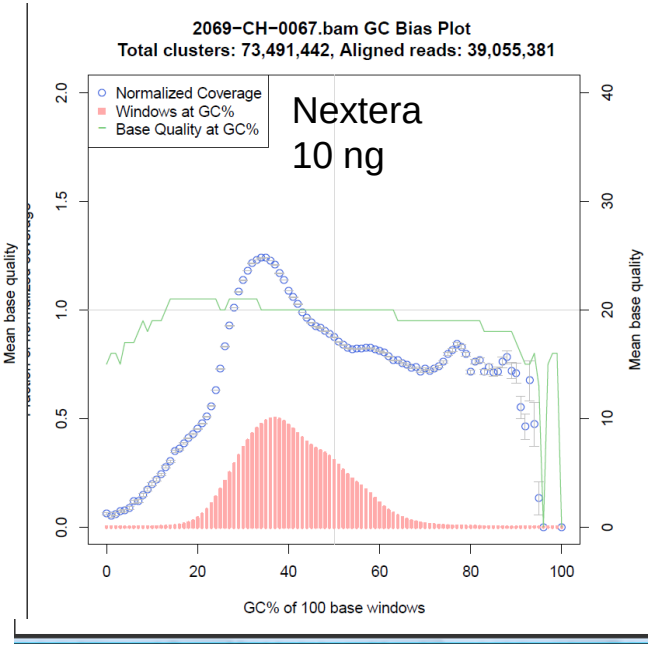


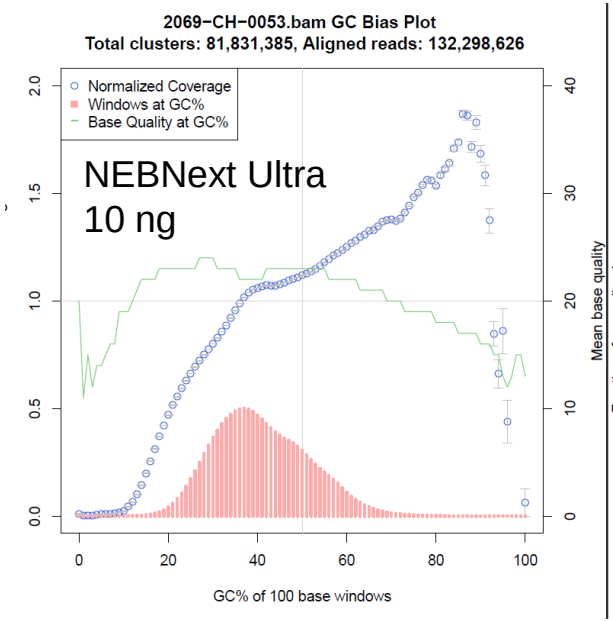
- Homebrew seems to have more narrow GC coverage
- Rubicon prep looks more uniform

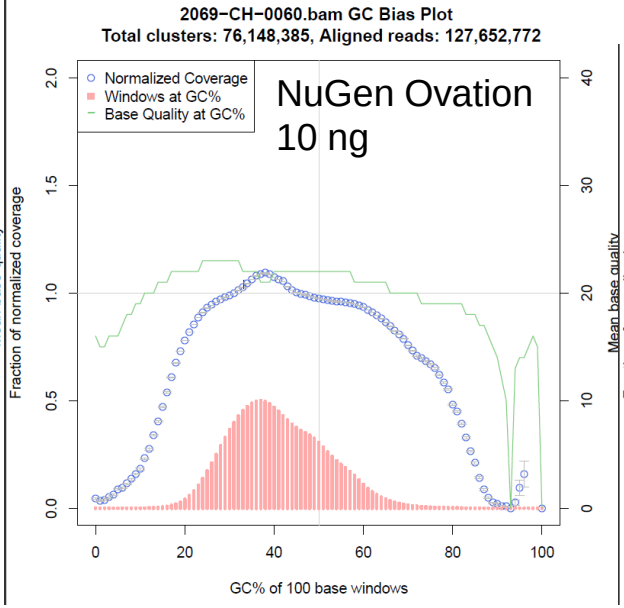
Metric#3: GC Bias at 2M Reads (HudsonAlpha)

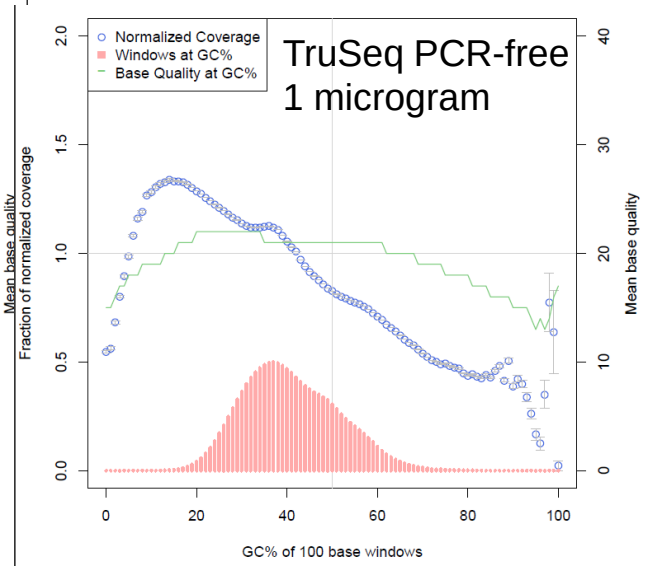




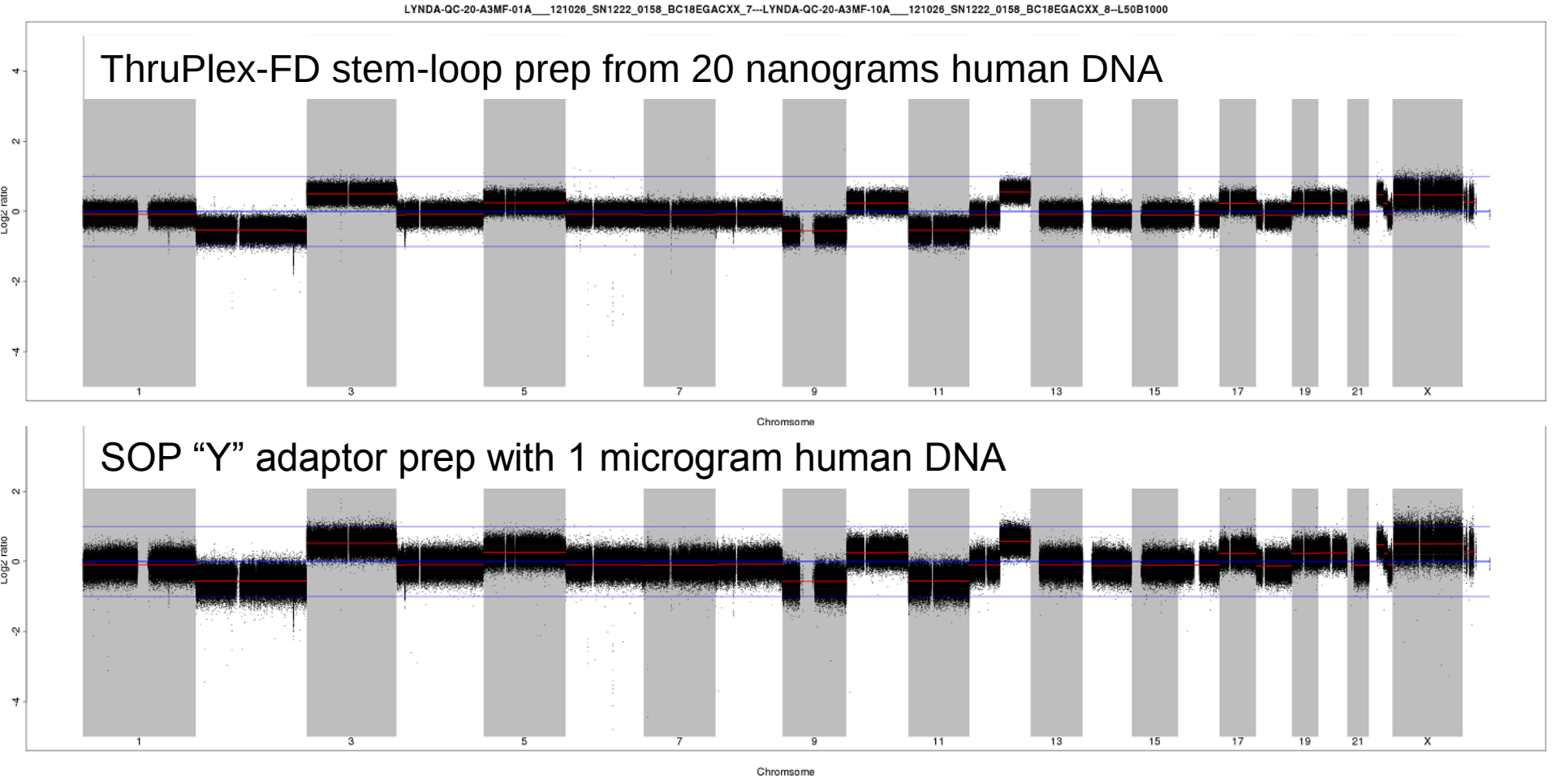


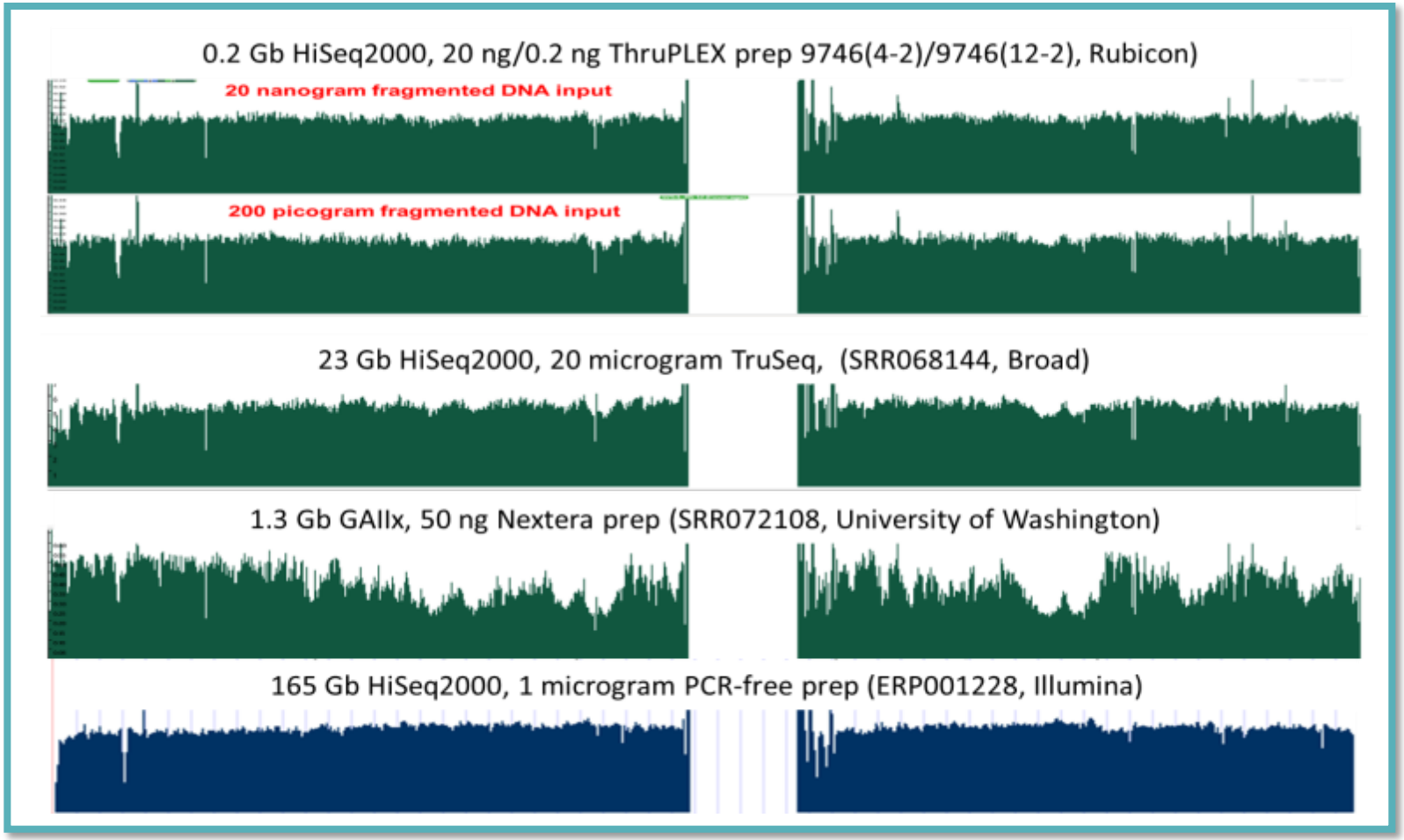




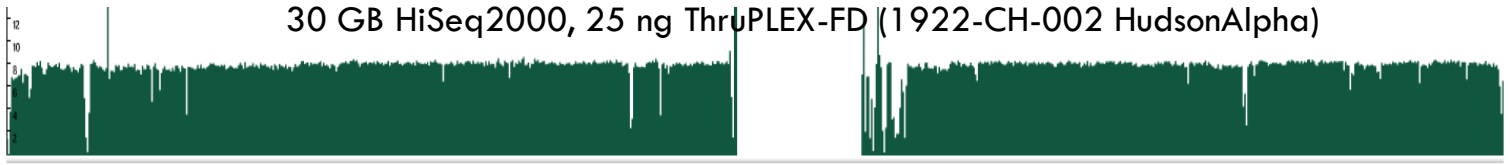


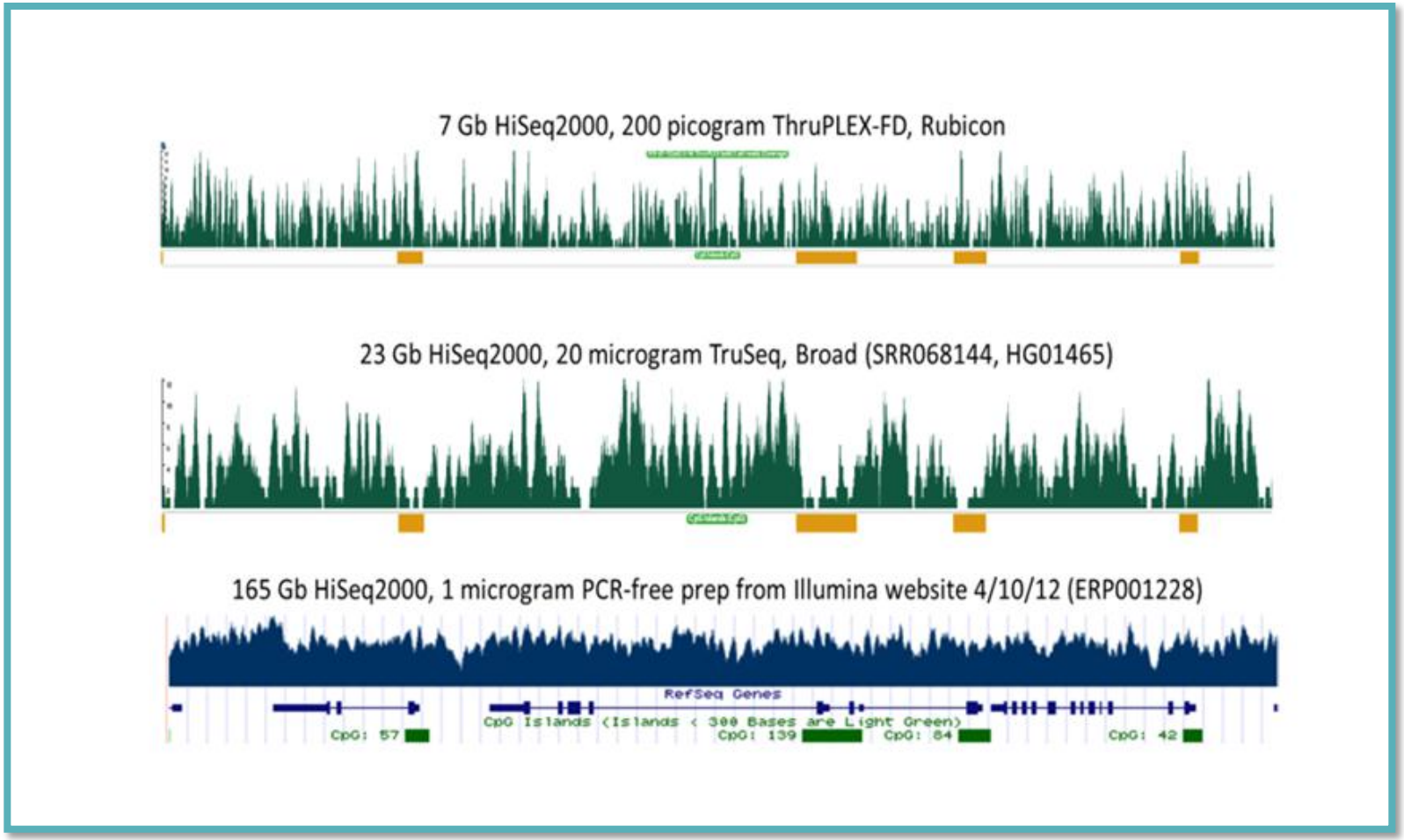
Metric#4: Whole Genome Coverage (Karyotyping with ThruPLEX and "Y" Adaptor Homebrew SOP (MD Anderson)





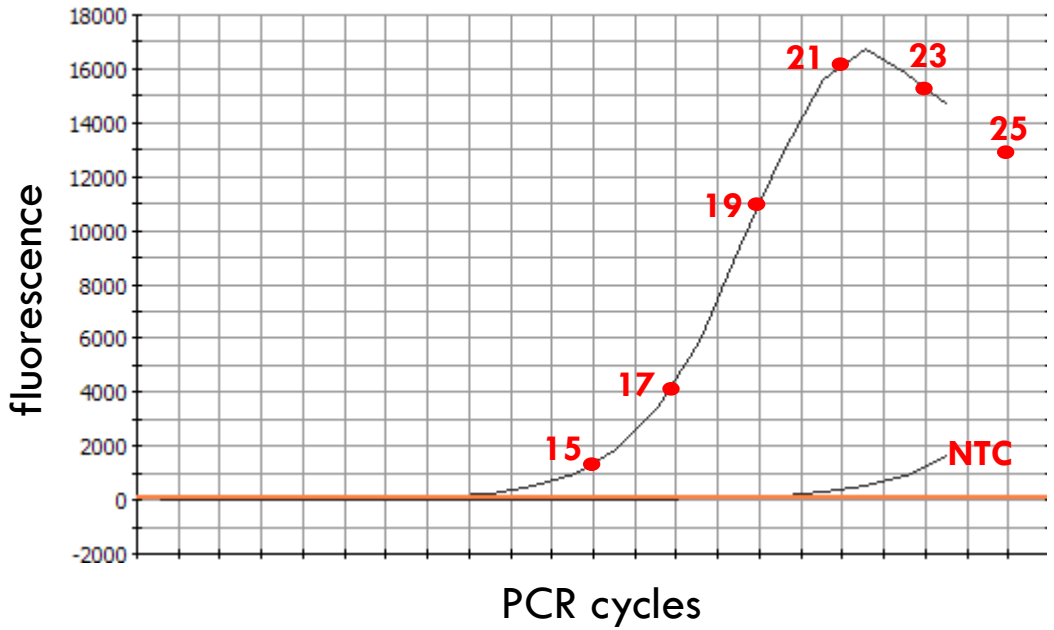
Metric #4: Low resolution uniformity of coverage across chr 1- ThruPLEX at 25 ng comparable to TruSeq PCR-free prep at 1 ug input





Metric #4: Uniformity of coverage in CpG islands

Coverage across 4 RASSF1 CpG islands for three technologies is compared. Using the Illumina PCR-free prep data as the gold standard, ThruPLEX-FD at low coverage has strong representation across CpG islands. The TruSeq prep shown has significant systematic dips in representation across CpG islands. Evaluation of RASSF1 needs ~100M reads. Evaluation of total CpG islands requires only 300K reads.



Total # Cycles	nM conc by qPCR
15 cycles	3.77
15 cycles	5.88
17 cycles	26.59
17 cycles	25.39
19 cycles	85.88
19 cycles	95.82
21 cycles	201.14
21 cycles	238.81
23 cycles	291.54
23 cycles	243.33
25 cycles	253.87
25 cycles	234.29

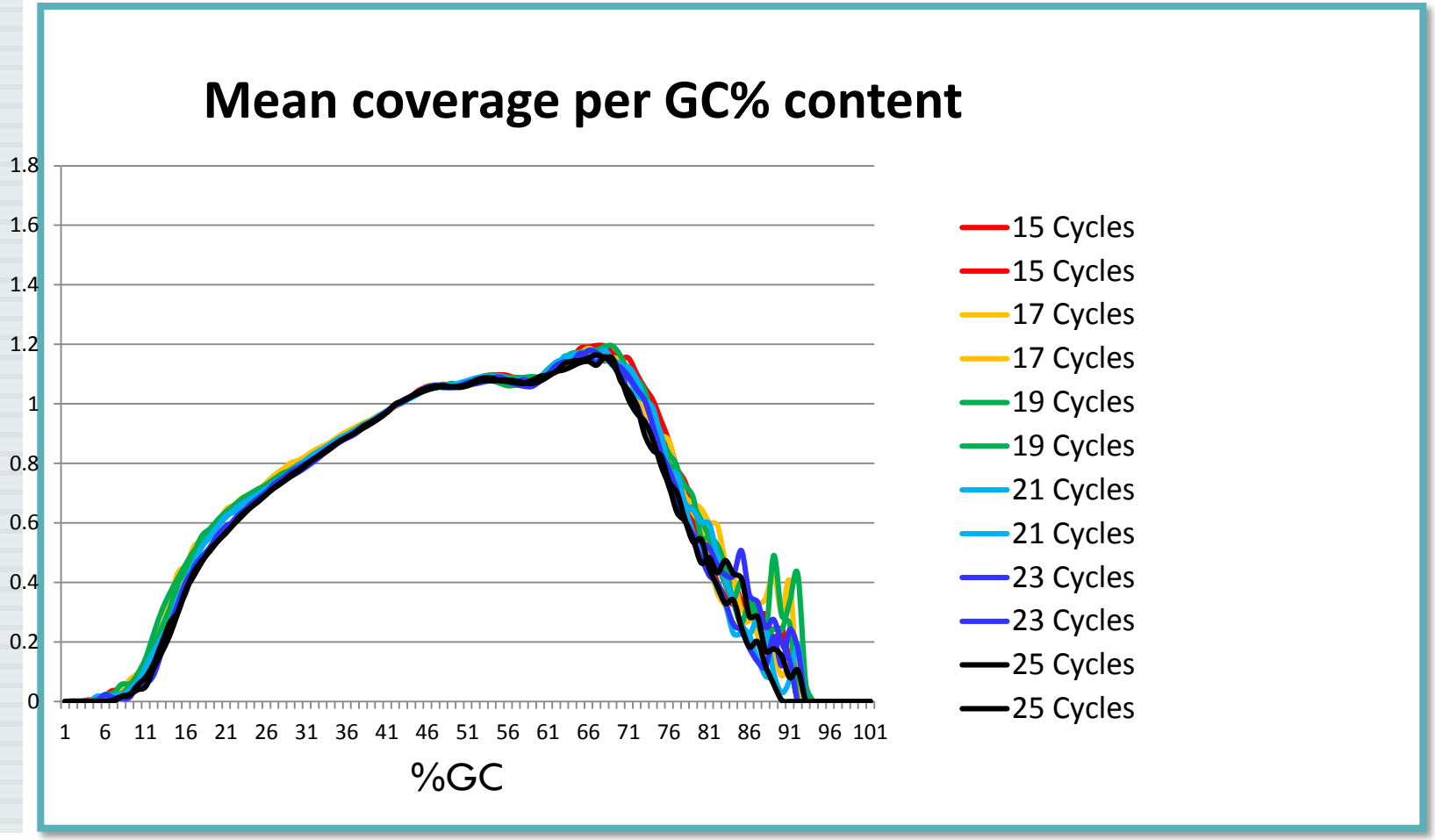
Library properties as a function of PCR cycles
Metric #5: Real-time qPCR threshold cycle and background

200 pg of Covaris-fragmented DNA was over-amplified by 10 PCR Cycles above that required for minimal sequencing, to achieve a "plateau." NTC (no template control) shows PCR background about 100X smaller than the 200 pg sample.

sample	PCR bkg	Total reads	E. coli reads	Unmapped reads	Human reads	Human contamination
15 pg E. coli	1 pg	3.2M	95.8%	4.2%	0.026%	3.9 femtograms
15 pg E. coli	2 pg	4.1M	94.5%	5.5%	0.042%	6.3 femtograms

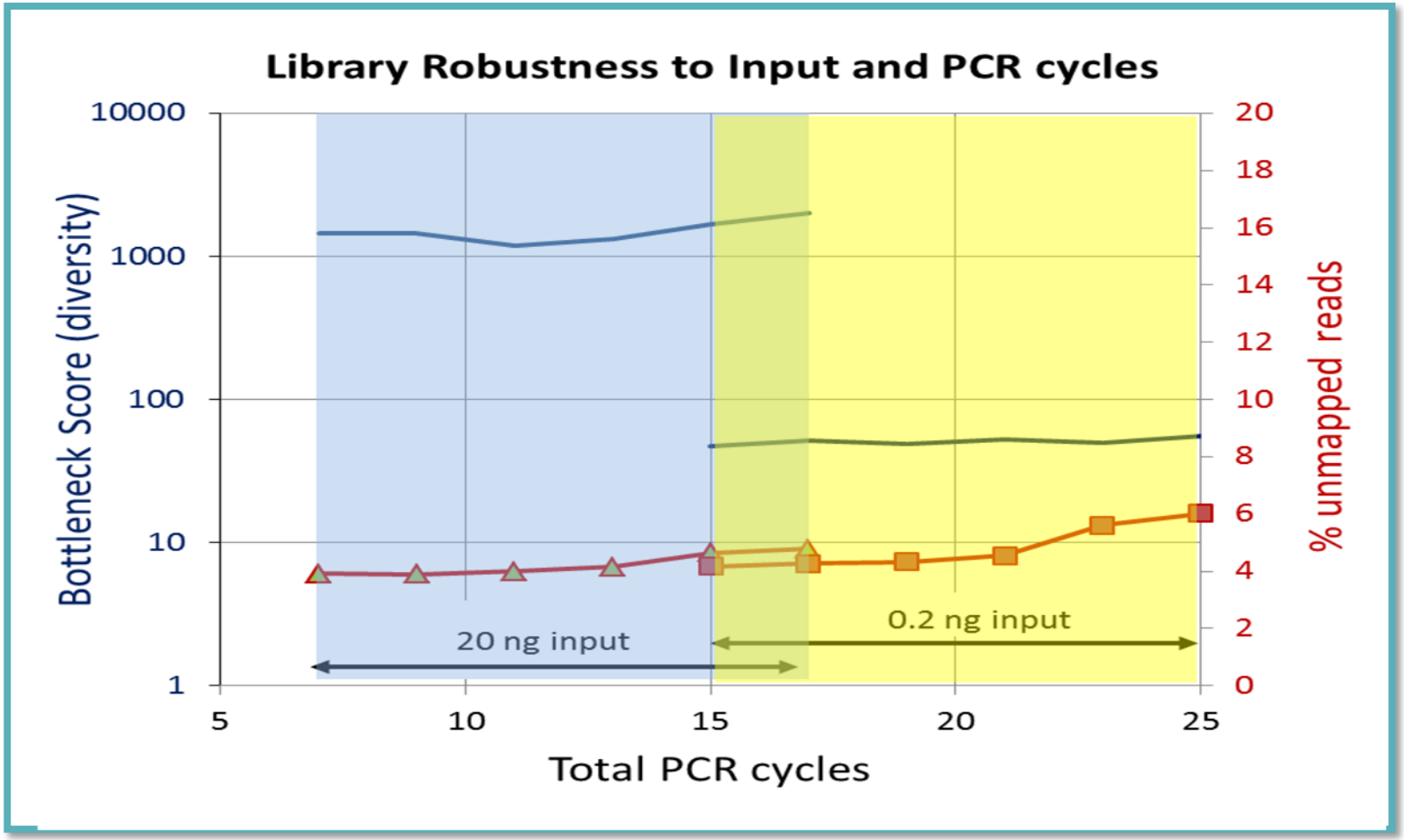
Metric #5, Background in ThruPLEX-FD

ThruPLEX-FD library preps have less than 0.001 GE of human contamination



Metric #6: Sensitivity of GC representation to PCR cycles

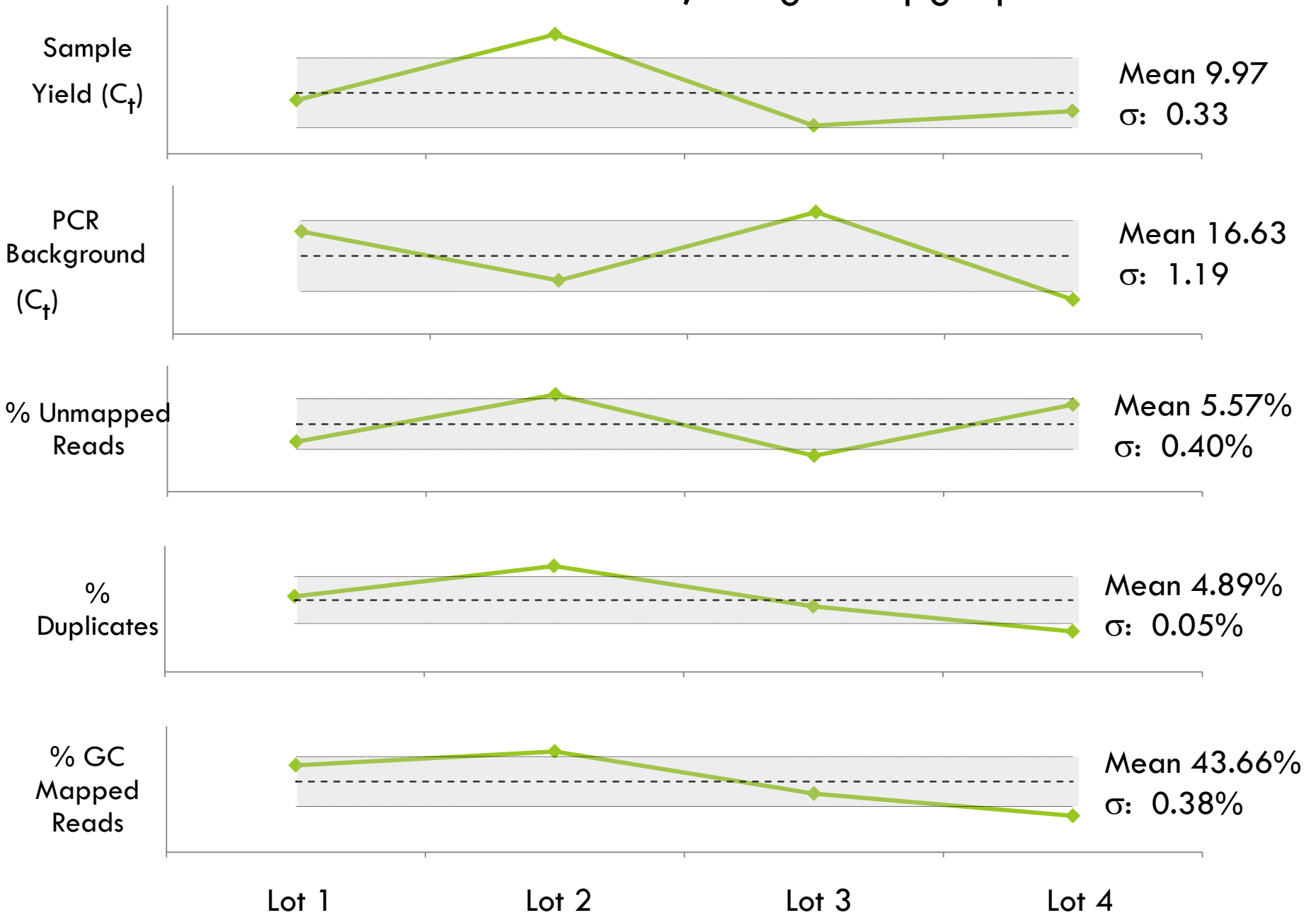
200 pg of Covaris-fragmented DNA was over-amplified by 10 PCR Cycles above that required for minimal sequencing, without serious changes in GC represention. GC-representation calculated by DNAnexus



Metric #6: Sensitivity of library diversity to PCR cycles
Metric #6: Sensitivity of unmapped reads to PCR cycles

20 ng and 0.2 ng of Covaris-fragmented DNA were made into libraries and amplified 10 PCR cycles more than required for sequencing without adverse effects on library diversity or number of unmapped reads. Samples over 100X range of input can be prepared using constant 15 – 17 cycles of PCR, without compromising performance.

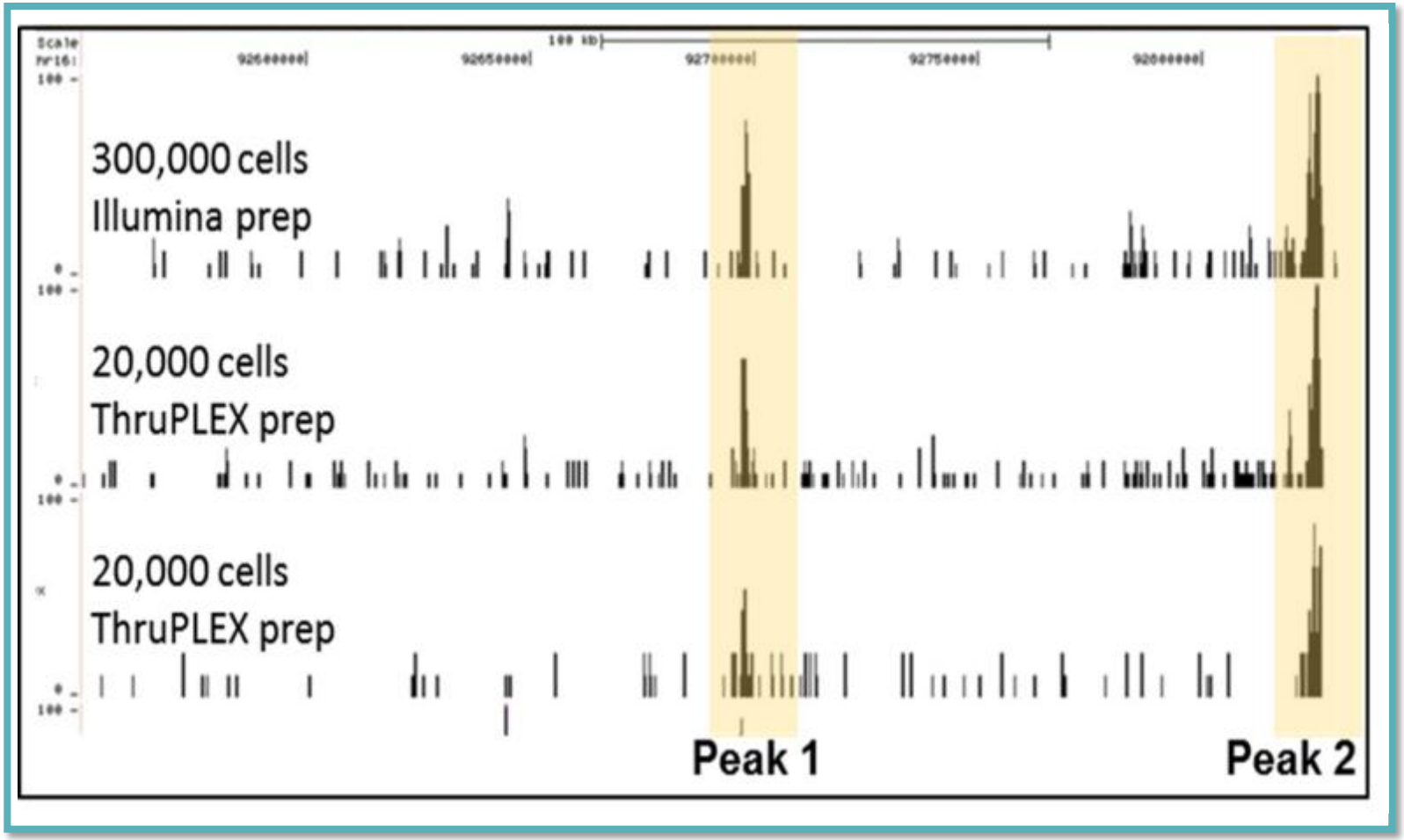
Metric#7: Lot-to-Lot ThruPLEX-FD QC, Using 200 pg Input



2x100 bp/kinome/3M	ThruPLEX 10ng input	standard XT2 1 microgram
Reads in:	3,000,470	3,000,248
Percent duplicate reads:	**14.45%**	**0.36%**
Number of reads in targeted regions:	979,195	1,216,113
Percentage reads in targeted regions:	46.97%	44.50%
Percentage reads in regions +/- 100bp:	55.29%	52.55%
Percentage of targeted bases covered by		
...at least 1 read:	98.23%	98.08%
...at least 5 reads:	93.18%	92.31%
...at least 10 reads:	84.43%	83.33%
...at least 20 reads:	64.35%	66.14%

Metric #8, 9: Compatibility and concordance of SureSelect kinome enrichment using low ThruPLEX-FD compared to unamplified input

10 ng of Covaris-sheared DNA was synthesized into a ThruPLEX-FD library and amplified before enrichment using SureSelect. The fraction of reads on target was the same for the ThruPLEX-FD sample as the unamplified sample.



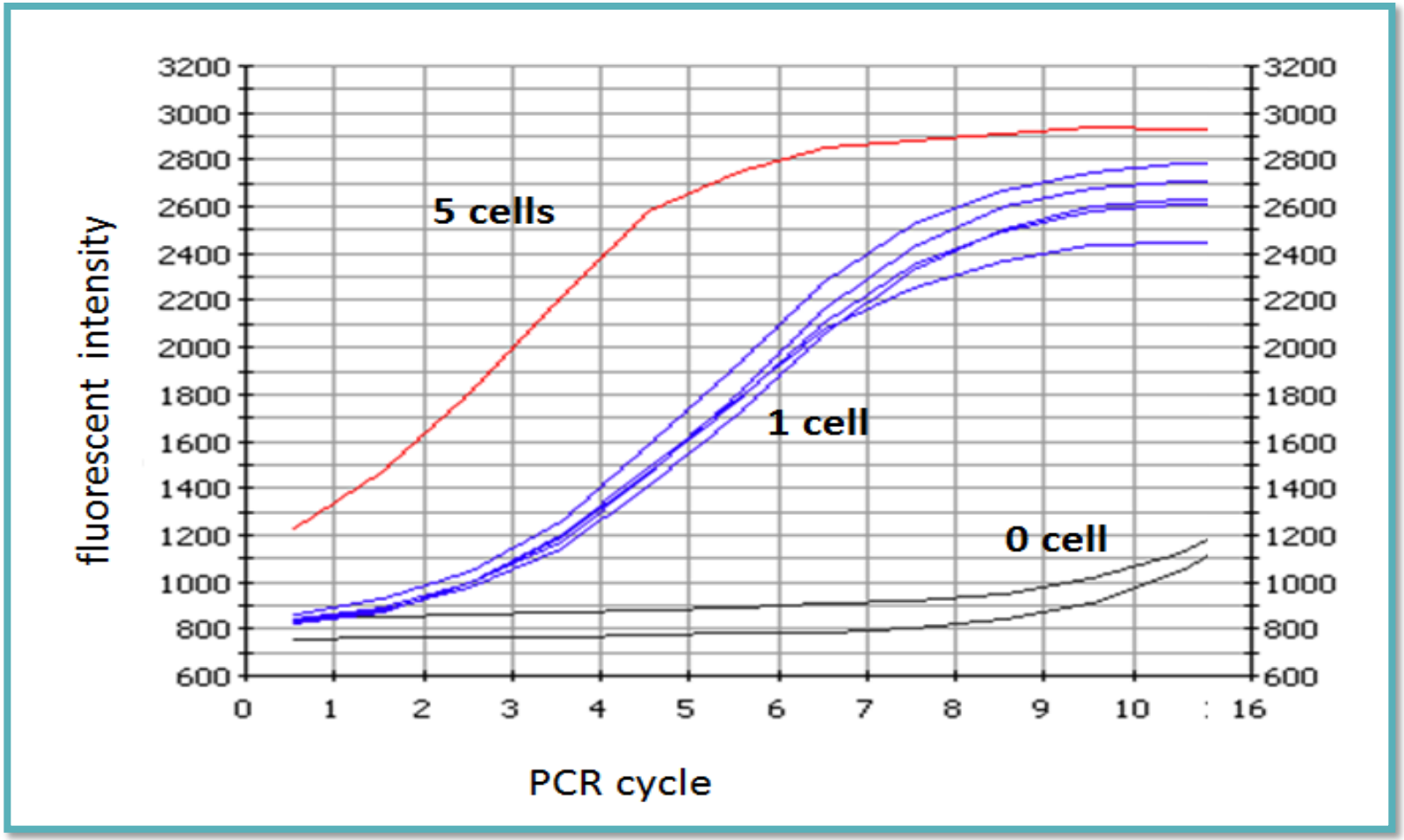
Metric #9,10: Compatibility and concordance of low input ChIP peaks and gold standard of high input of same sample using TruSeq-Chip

ThruPLEX-FD peaks from 50 - 200 pg ChIP DNA input were >92% concordant with peaks from 10 ng ChIP DNA from 300,000 cells precipitated with the same antibodies and prepped with Illumina ChIP-seq kit. However, TruSeq has lower % duplicates, because there are fewer reads on target (a case of pseudo-diversity). Data provided by Baylor Medical College.

RUBICON GENOMICS
breaking the sample barrier

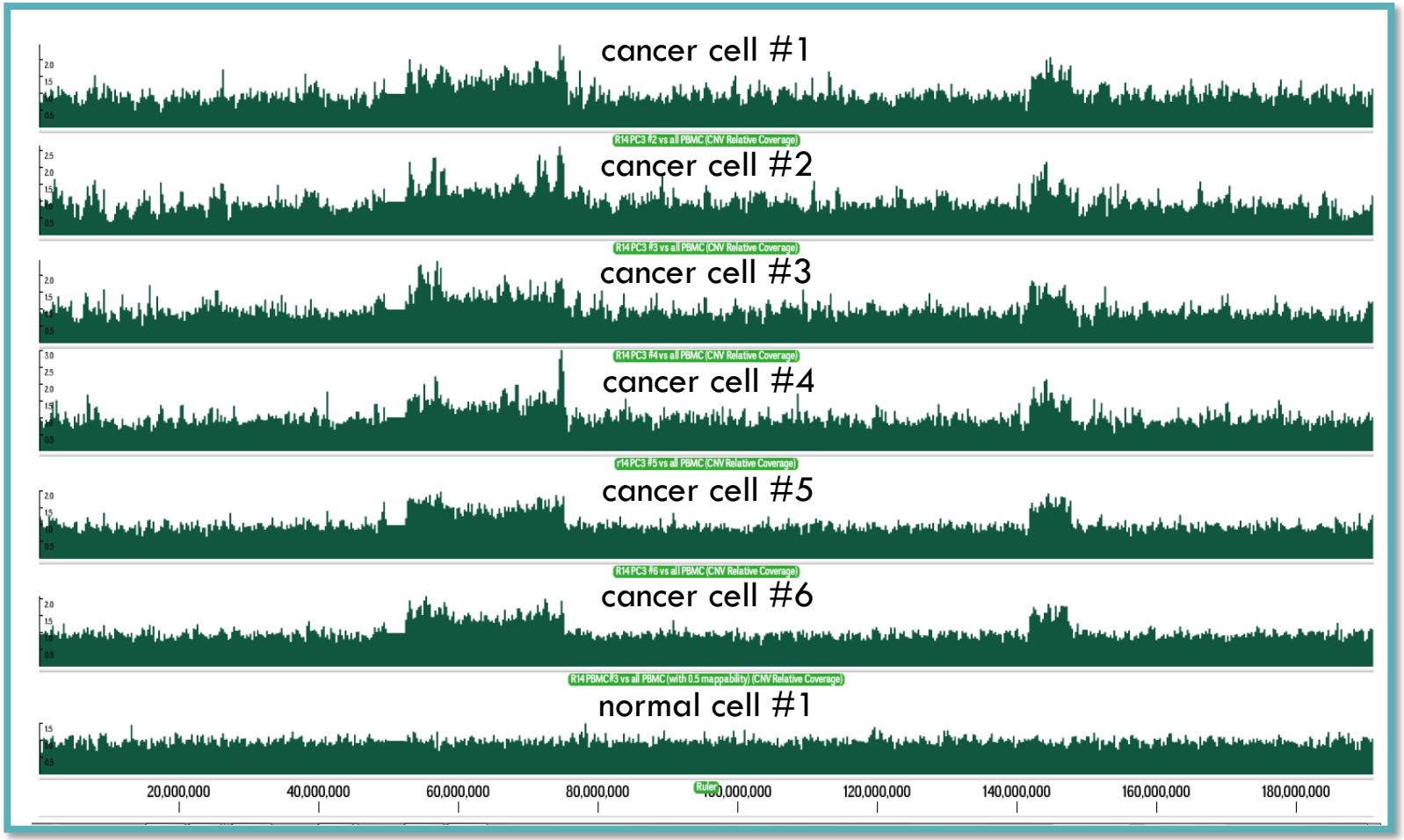
New Rubicon NGS Library Products in Development

- PicoPLEX-scD single-cell prep for CNV [IVF,CTC]
- ThruPLEX-bfD for biofluids [prenatal and cancer MDx]
- ThruPLEX-FD for high sensitivity RNA-seq using cDNA libraries (e.g., Clontech SMARTer) [MDx]



Real-Time PicoPLEX shows high linearity and reproducibility with low background

Human immortalized cancer cells were sorted into groups of 8 wells having 0, 0, 1, 1, 1, 1, 1 and 5 cells. PicoPLEX WGA was monitored in real time to demonstrate linearity and reproducibility of library synthesis and amplification. qPCR shows that non-specific background is less than 50 fg.



PicoPLEX Reproducible Determination of Copy Number Variations in Chr4 of Single Cancer Cells Using PicoPLEX-scD

Single PC3 cells were flow sorted into a 96-well plate. Six cancer and 6 normal cell samples were prepared in 12 wells with PicoPLEX-scD and sequenced on HiSeq. Triploid regions were reproducibly determined in all 6 PC3 samples at 10 M total genomic reads. Normal cells did not show copy variations.

Noninvasive Whole-Genome Sequencing of a Human Fetus
Sci Transl Med **4**, 137ra76 (2012);

Jacob O. Kitzman,[1*] Matthew W. Snyder,[1] Mario Ventura,[1,2] Alexandra P. Lewis,[1] Ruolan Qiu,[1] LaVone E. Simmons,[3] Hilary S. Gammill,[3,4] Craig E. Rubens,[5,6] Donna A. Santillan,[7] Jeffrey C. Murray,[8] Holly K. Tabor,[5,9] Michael J. Bamshad,[1,5] Evan E. Eichler,[1,10] Jay Shendure[1*]

Analysis of cell-free fetal DNA in maternal plasma holds promise for the development of noninvasive prenatal genetic diagnostics. Previous studies have been restricted to detection of fetal trisomies, to specific paternally inherited mutations, or to genotyping common polymorphisms using material obtained invasively, for example, through chorionic villus sampling. Here, we combine genome sequencing of two parents, genome-wide maternal haplotyping, and deep sequencing of maternal plasma DNA to noninvasively determine the genome sequence of a human fetus at 18.5 weeks of gestation. Inheritance was predicted at 2.8×10^6 parental heterozygous sites with 98.1% accuracy. Furthermore, 39 of 44 de novo point mutations in the fetal genome were detected, albeit with limited specificity. Subsampling these data and analyzing a second family trio by the same approach in- dicate that parental haplotype blocks of ~300 kilo–base pairs combined with shallow sequencing of maternal plasma DNA is sufficient to substantially determine the inherited complement of a fetal genome. However, ultradeep sequencing of maternal plasma DNA is necessary for the practical detection of fetal de novo mutations genome-wide. Although technical and analytical challenges remain, we anticipate that noninvasive analysis of inherited variation and de novo mutations in fetal genomes will facilitate prenatal diagnosis of both recessive and dominant Mendelian disorders.

ThruPLEX Used in Milestone Non-Invasive Fetal WGS

First non-invasive prenatal WGS was enabled by high complexity, accuracy, and coverage of ThruPLEX-FD preps of maternal plasma.

RUBICON GENOMICS
breaking the sample barrier

QC Dashboard Projects (DNAnexus and Maverix)