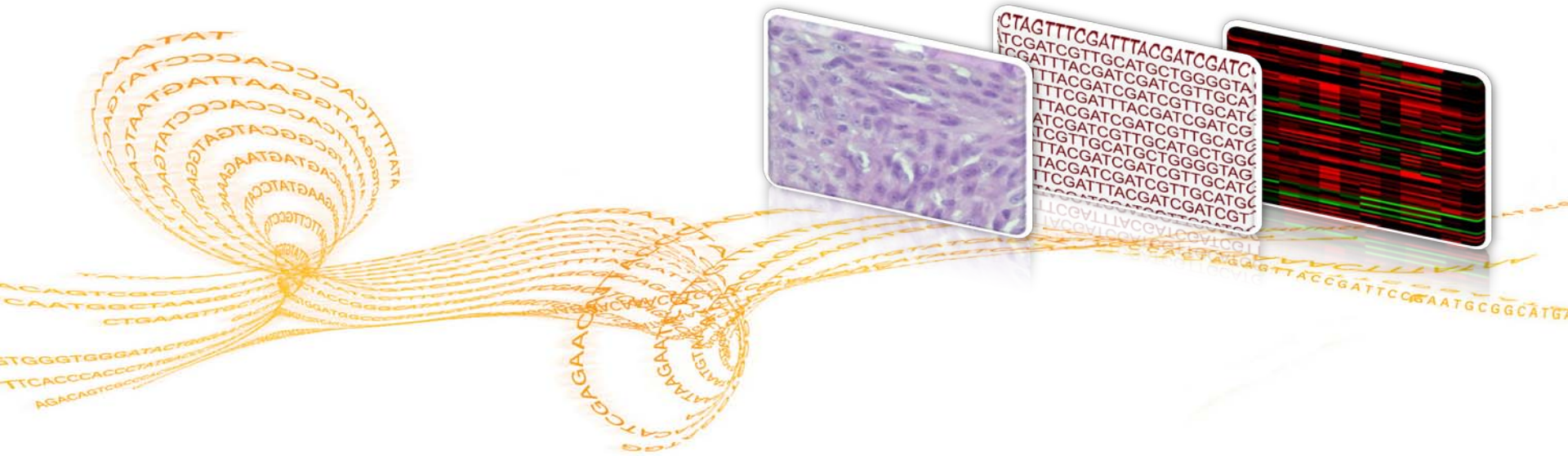


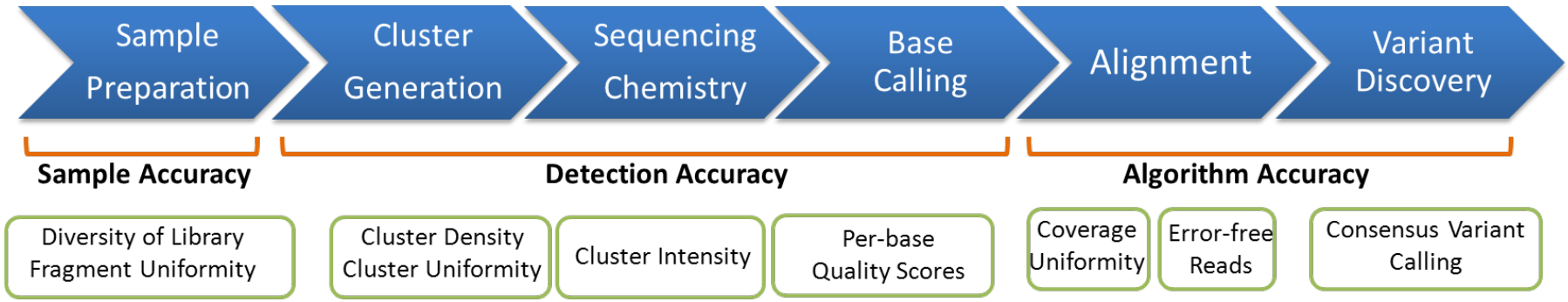
Genome Metrics

Scott D. Kahn, PhD
Chief Information Officer

illumina®

Data Quality Defined by Metrics Evaluated

- Read-level quality
  - Quality scores and consensus accuracy
  - GC content and uniformity of representation
- Coverage/gap analysis
  - Biological relevance can be a critical lens
- False positive and false negative call rates
  - SNV's, Indels, rearrangements, …
- Storage implications
  - Interplay between base quality, depth of coverage, and storage required



Improving Read Quality

Reducing Density-Dependent CG Bias

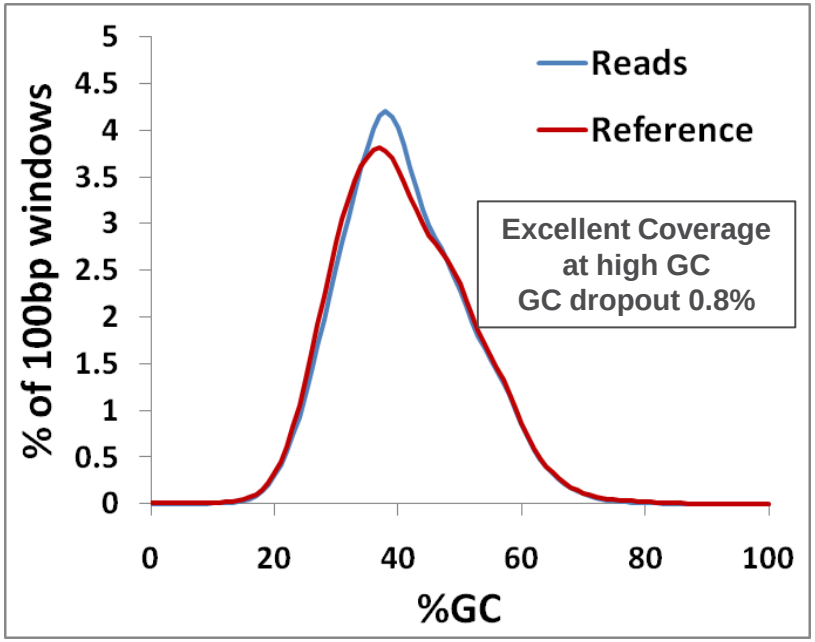
New cluster amp method equalizes growth of AT and GC rich clusters

Cluster Density

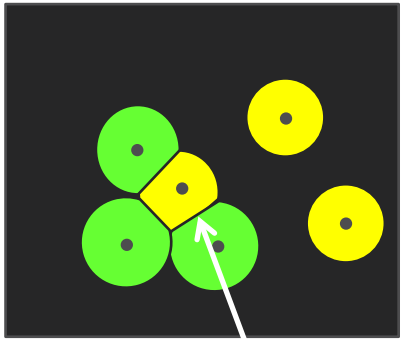
488k/mm²

TruSeq v3 cBot kit Reduces bias

825k/mm²



New Cluster Amplification High Density (966k/mm²)



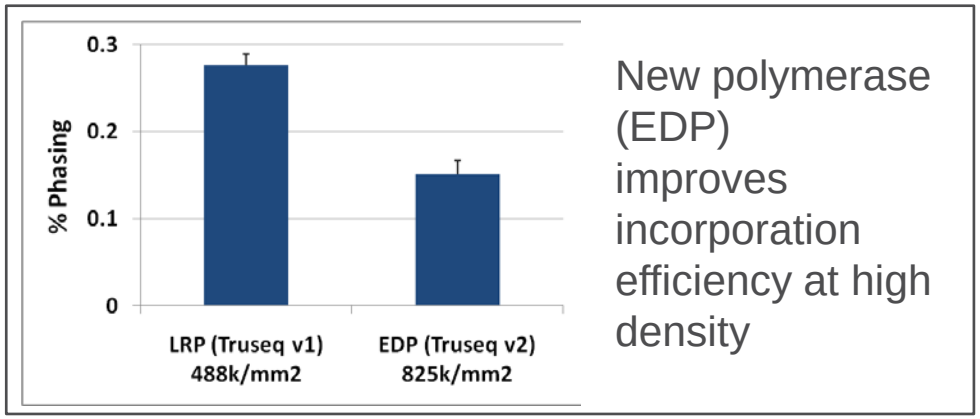
Larger, brighter GC-rich clusters are well resolved and detected

High GC

Low GC

Improving Base Accuracy

Cluster Density	Fraction Passing Filter	Imageable Area (Two FC)	Read length	Yield (Gb)	Throughput (Gb/day)	%>Q30
488k/mm^2	0.878	2949 mm^2	2x100	252.7	31.6	87%



New polymerase (EDP) improves incorporation efficiency at high density

TruSeq v2 SBS kit
Highly Accurate at High Density

825k/mm^2	0.904	4424 mm^2	2x100	652.6	60.4	84-91%*

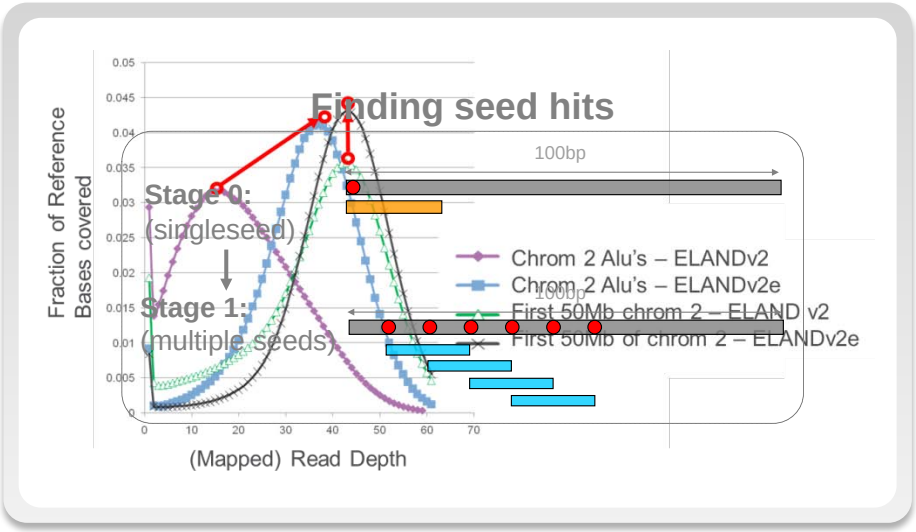
New Scan Reagent (SRE) also reduces signal decay

*Preliminary

illumina

Informatics Improvements

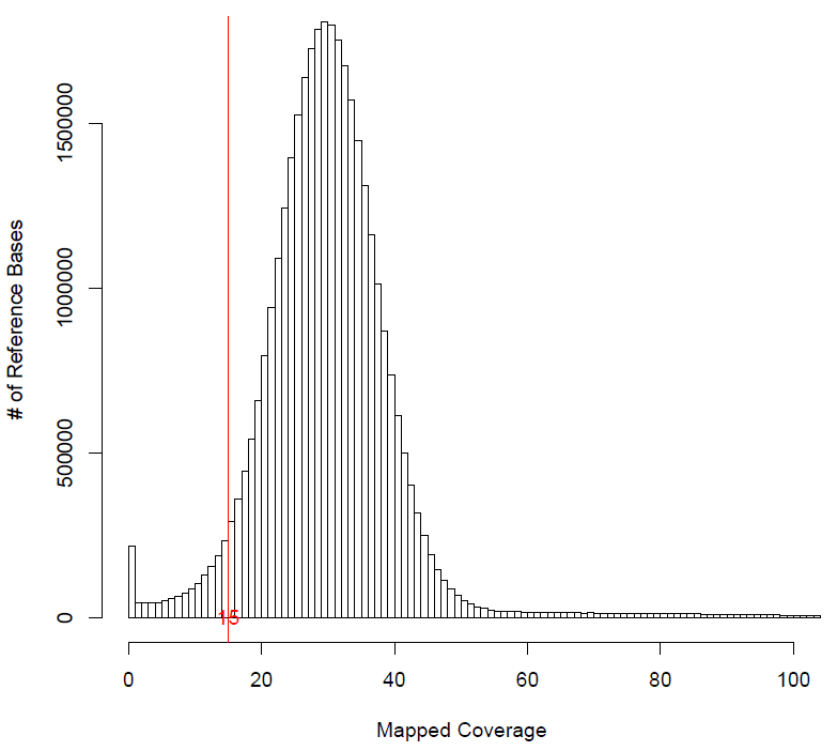
- Alignment Improvements in ELANDv2e
  - Repeat resolution using overlapping seeds
    - Improves coverage over repeat regions
  - Orphan aligner: reads that have multiple mappings, but have read partners which map uniquely, are realigned and scored
    - Increases coverage (% aligned reads: 5–7% more reads)
    - Improves indel detection



- Indel detection (up to 300 bp) and variant calling (SNPs and indels) improvements
  - SNP and indel caller changed to a probabilistic model similar to MAQ or GATK
    - Provides a predicted diploid genotype for every position in genome
  - Intersecting reads from the following two sources are realigned to generate accurate indel calls
    - Output of gapped alignment from ELANDv2e
    - Output of IndelFinder (Grouper algorithm)
- Significant reduction in processing time due to multiple optimizations
- Multiple architecture improvements described in detailed documentation (available on iCom and SEQanswers)
- CASAVA 1.8 is in early-access currently

Assessing Quality/Performance of a Genome

Tighter Distributions Reduce Data Requirements
– NA18507, Chr 21



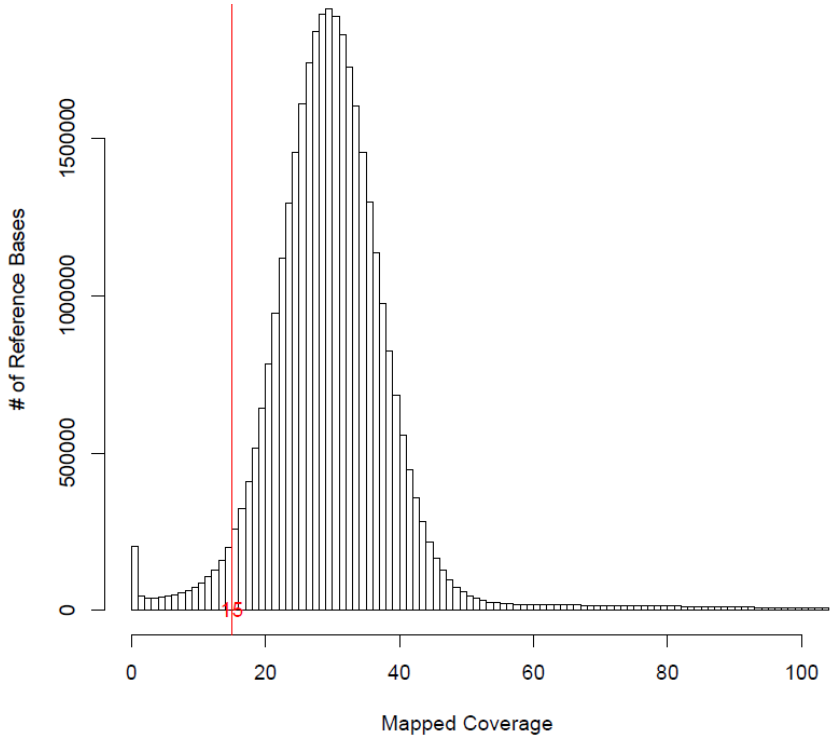
GAIIx+TruSeq

Mean mapped coverage (Chr 21) = 32.55

4.44% ref. bases at low coverage (<=15x)

IQR = 11

0.45% ref. bases of 0 read



HiSeq2000+TruSeq v3

Mean mapped coverage (Chr 21) = 32.49

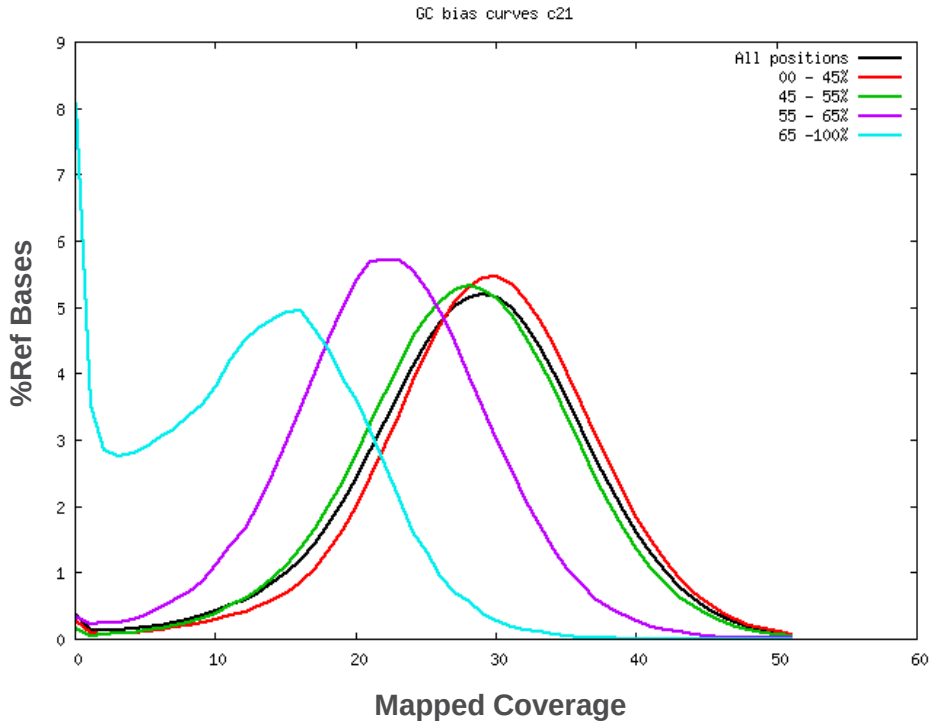
3.8% ref. bases at low coverage (<=15x)

IQR = 10

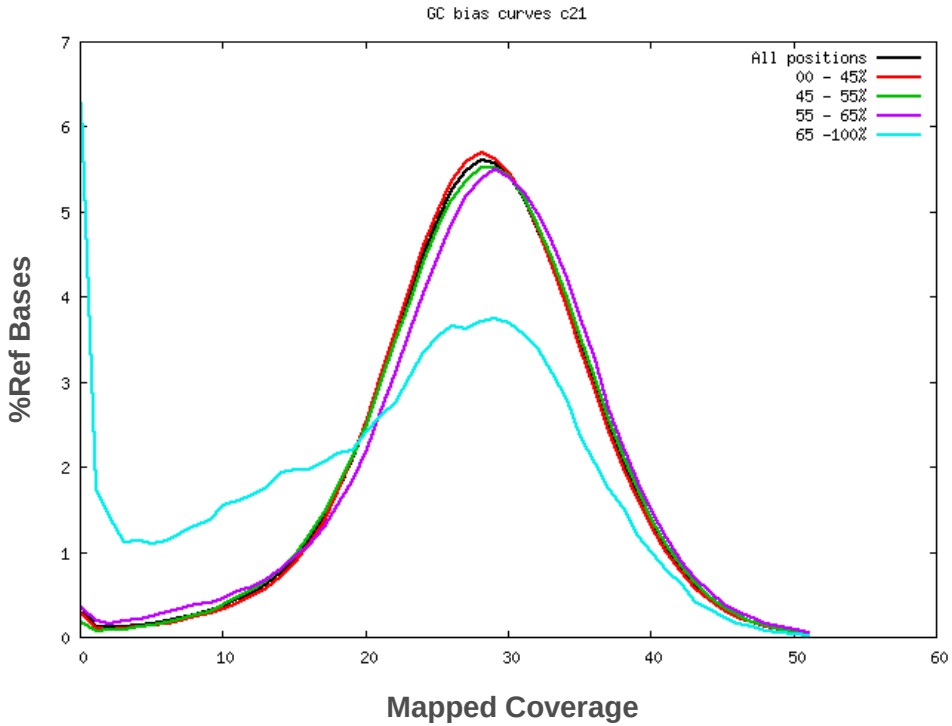
0.43% ref. bases of 0 read

1. Only uniquely mapped reads were considered (MAPQ>1)
2. 13,023,253 Ns in NCBI build 37 were excluded

Uniform (GC) Coverage Reduces Data Requirements
– NA18507, Chr 21



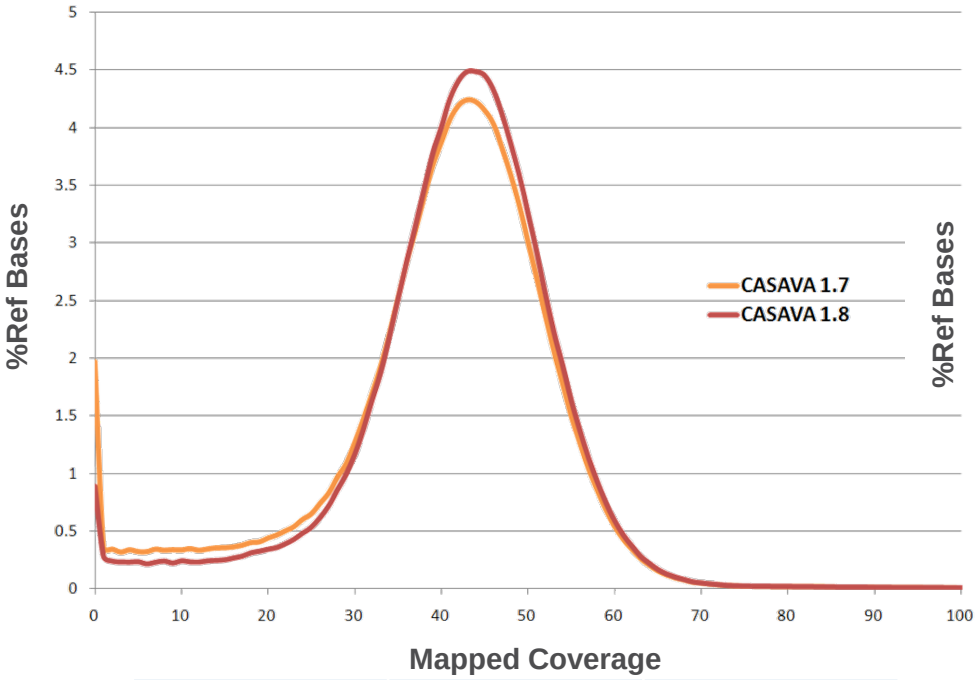
GAIIx+TruSeq



HiSeq2000+TruSeq v3

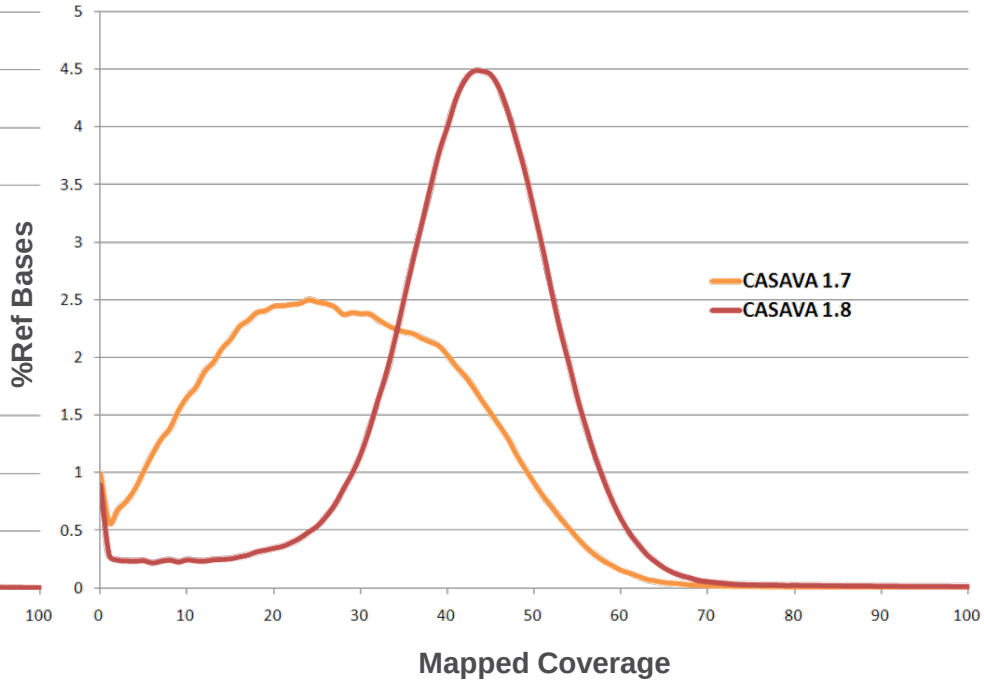
Improved Alignment Better Levers Existing Data
– NA19240, Chr 21

Long repeats (LINE: L1, L2)



	Mean mapped coverage	Bases at low coverage (≤15x)
CASAVA 1.7	41x	7.1%
CASAVA 1.8	42.7x	4.5%

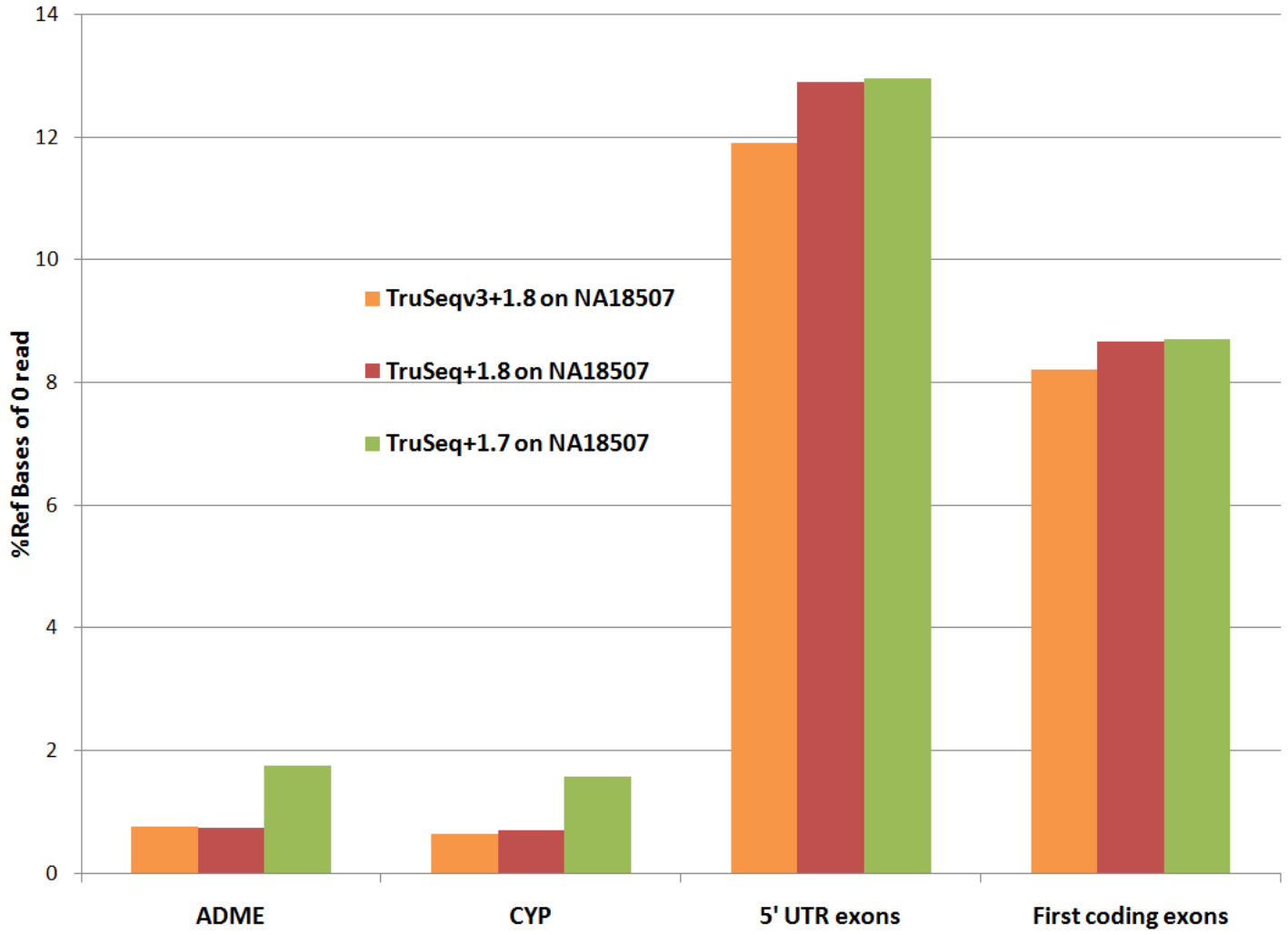
Short repeats (SINE: Alu, MIR)



	Mean mapped coverage	Bases at low coverage (≤15x)
CASAVA 1.7	28.4x	21.7%
CASAVA 1.8	38.8x	2%

illumina

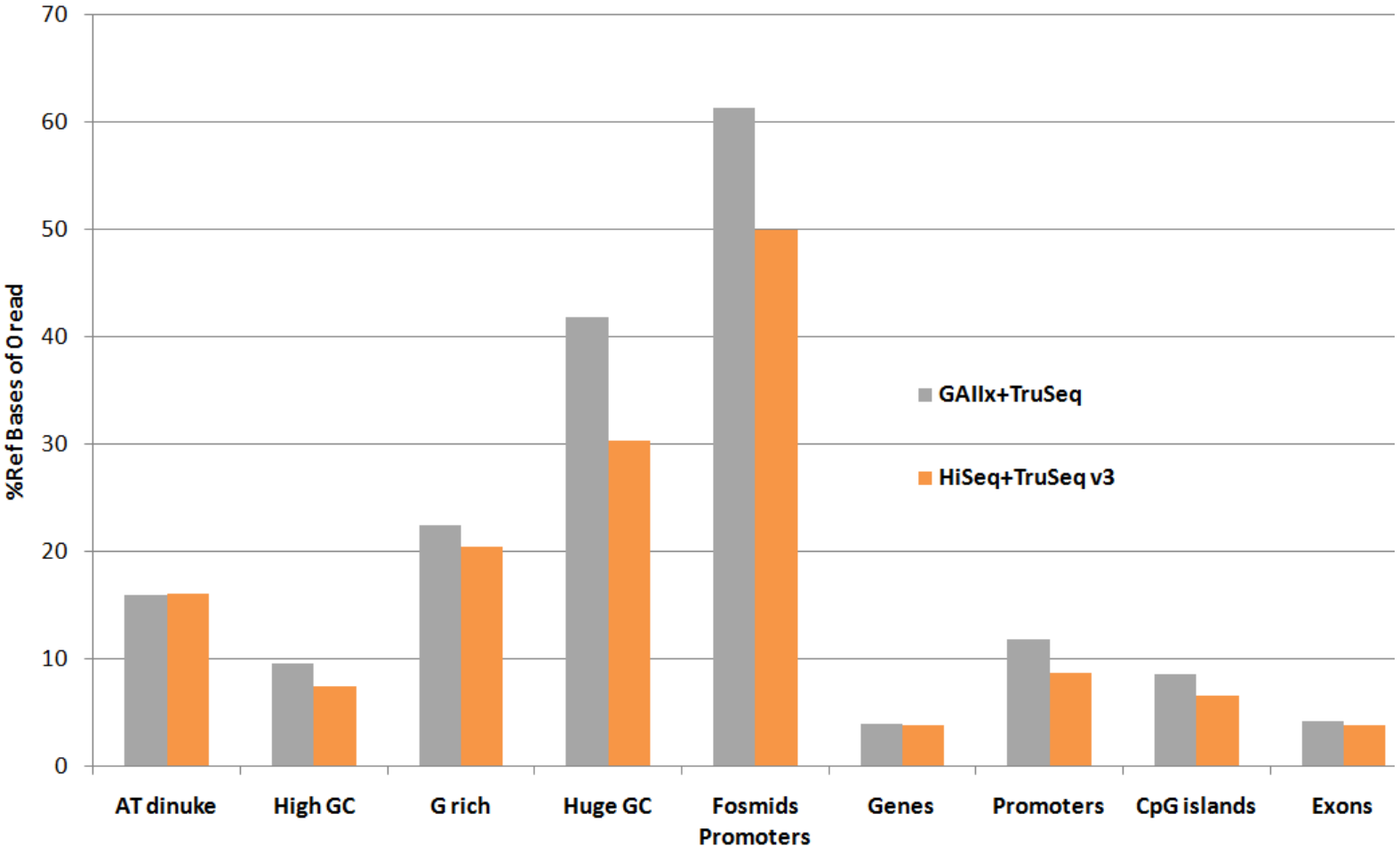
Improved Coverage Reduces Data Requirements - NA18507, whole genome



1. Core ADME gene list from http://pharmaadme.org/joomla/index.php?option=com_content&task=view&id=12&Itemid=27
2. CYP gene family from http://www.genenames.org/genefamily/cyp.php

illumina

Gaps in Challenging/Biologically Interesting Genomic Areas – NA18507, whole genome



illumina

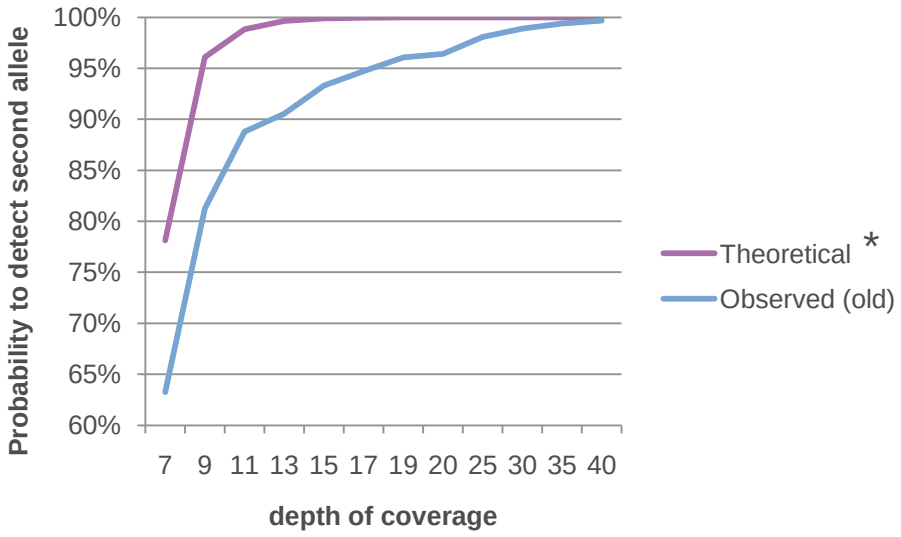
Sequencing
in a Clinical Laboratory

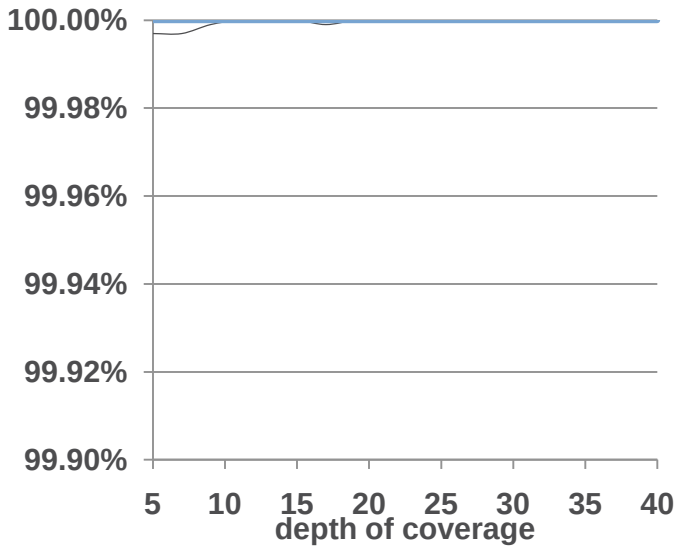
Individual Genome Sequencing in a Clinical Laboratory

- Illumina established a protocol for sequencing whole genomes
  - Using Professional and Regulatory agency guidelines
  - Under advisement from an external Ethics Advisory Board
  - Geneticists, Clinicians, Ethicists and Attorneys participated in design of protocol
- Our protocol and validations were evaluated externally
  - CLIA certified for high complexity testing
  - CAP accredited
- Elements of the protocol
  - Process workflow and safety checks
  - Technical validation of the assay
  - Providing a human genome deliverable

Assessment of Accuracy

- 14 samples of known Factor V genotype sequenced at >6 million-fold depth
- 100% accuracy in genotype calls
- Sub-sampled 100,000 times at various folds of coverage:





Sensitivity
At average 30 fold coverage:
100% for hom./98.89% for het

Specificity
At average 30 fold coverage:
100%

* $P(x, p, N)=\sum_{K=x}^{N-X} \frac{N!}{(X!)(N-X)!}\ p^X q^{(N-X)}$

Concluding Remarks

Data Management Burden

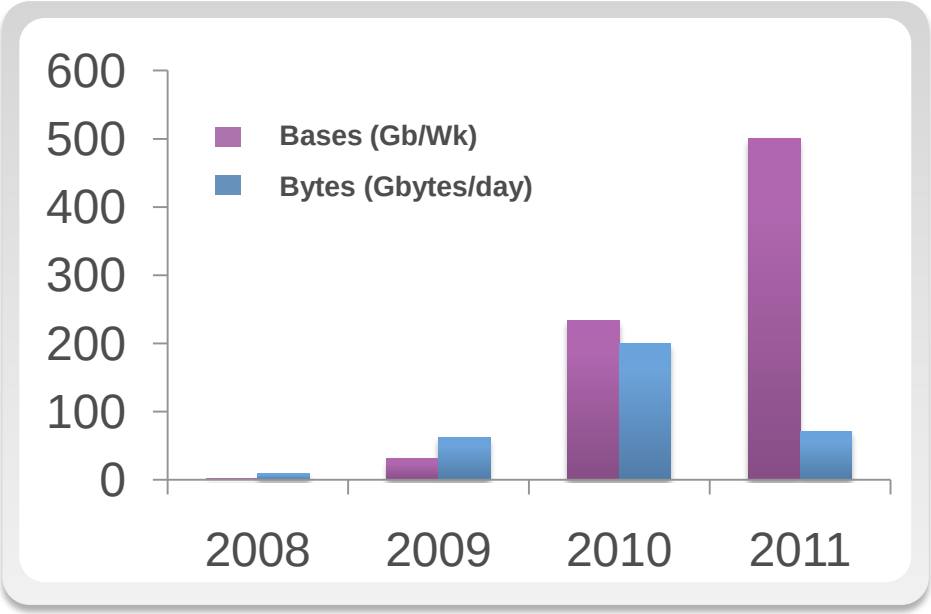
ONGOING REDUCTION THE DATA FOOTPRINT

STREAMLINING DATA ANALYSIS

INCREASING COMPUTATIONAL PERFORMANCE

Mapping	Bytes / Base	Data Size 30X Genome
Pipeline (2008)	30	3,000 GB
CASAVA (2009)	14	1,400 GB
CASAVA (2010)	6	600 GB
Summarized (2011+)	.1	~10 GB

CASAVA 1.8 uses 1 byte/base



illumina

Structural Variants and FP/FN Rates

Accuracy Metric	Sensitivity			Specificity
Type	**Simulated data**	**dbSNP[1]**	**1000 Genomes[2]**	**Simulated data**
Small variant				
SNV		97%		
Insertion	>90%	63%	>60%	
Deletion	80%		>30%	
Structural variant				
Large insertion	80%	n/v	2%	>90%
Large deletion	>80%	n/v	30%	>90%
Inversion	50%	n/v	n/v	>90%
Tandem duplication	40-45%	n/v		>95%
Translocation		n/v	n/v	
Copy number variant				
CNV gain	74%	n/v		100%
CNV loss	33%	n/v	>15%	77%

1. dbSNP: For SNVs, A subset of 245,469 SNPs were identified as being present in dbSNP130 and further validated by a capillary sequencing study in NA19240 and at least one other Yoruban individual (of NA18506, NA18507 and NA18508) reported in Nature 2008 453(7191):56-64.
   For indels, a subset of 31,665 indels were identified and validated in the same study.
2. 1000 Genomes: For insertions, a set of 562 validated SVs were extracted from the Pilot 2 study of 1000 Genome project.
3. For deletions, a set of 9695 validated deletions were extracted from the Pilot 2 study of 1000 Genome project.
   For CNV loss, a set of 480 CNVs were identified by SOLiD sequencing on NA19240 in the subproject led by Applied Bioscience as part of 1000 Genomes project, and were further validated by at least one low-throughput assay (Mills et al. Nature 2011 470:59-65)