Whole Genome Sequencing highlight & Featured Applications

Yonggang Zhao

Director, Int'l Tech-Service BU
BGI Genomics Co., Ltd.

源資生技
Tri-I Biotech,Inc.
Total Solution Provider of Systems Biology and Chemoinformatics

华大基因
BGI

Content

- BGI Overview
- BGISEQ-500 and WGS Solutions
- BGISEQ-500 other Applications

Content

源資生技 Tri-I Biotech,Inc.
Total Solution Provider of Systems Biology and Chemoinformatics

华大基因 BGI

BGI History

"China has unveiled a new Human Genome Center in Beijing that it hopes will boost its contribution to the international genome research."

- *Science* 21 Aug. 1998

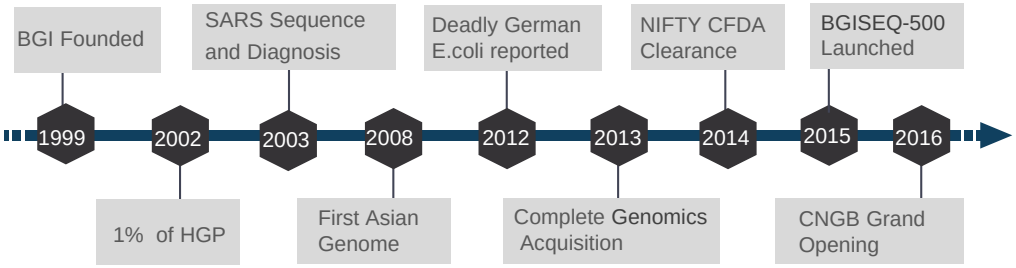
Our Mission

Gene Technology for the Benefit of Mankind

Our Vision

Improving Human Health and Leading the Era of Life Science

Key Events in BGI History



BGI Facts

Employees

+5,000 - comprising laboratory technicians, administration, sales and marketing

Samples

+1,100,000 – samples processed for research
+ 4,900,000 – samples processed for clinical applications

Publications

+1,949 – publications in leading journals across the world

Countries

+100 Countries - BGI has a global sales team and partner network covering the world.

Citations

+82,745 – times our research has been cited

Collaborators

+18,000 – collaborators BGI works with across the world

源資生技
Tri-I Biotech,Inc.
Total Solution Provider of Systems Biology and Chemoinformatics

华大基因
BGI

Content

BGI Overview

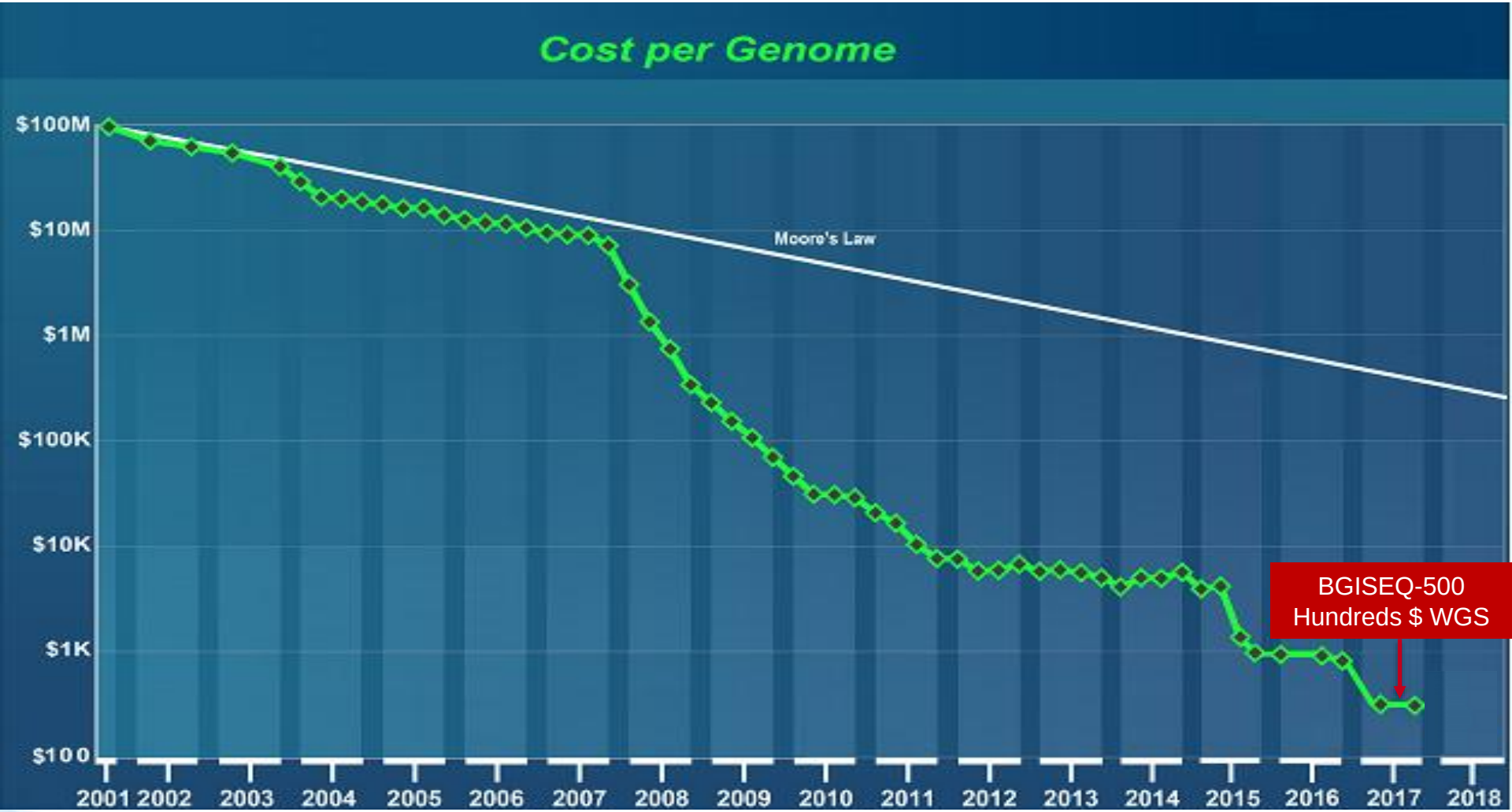
BGISEQ Based WGS Solutions

The data performance of BGISEQ-500

源資生技 Tri-I Biotech,Inc.
Total Solution Provider of Systems Biology and Chemoinformatics

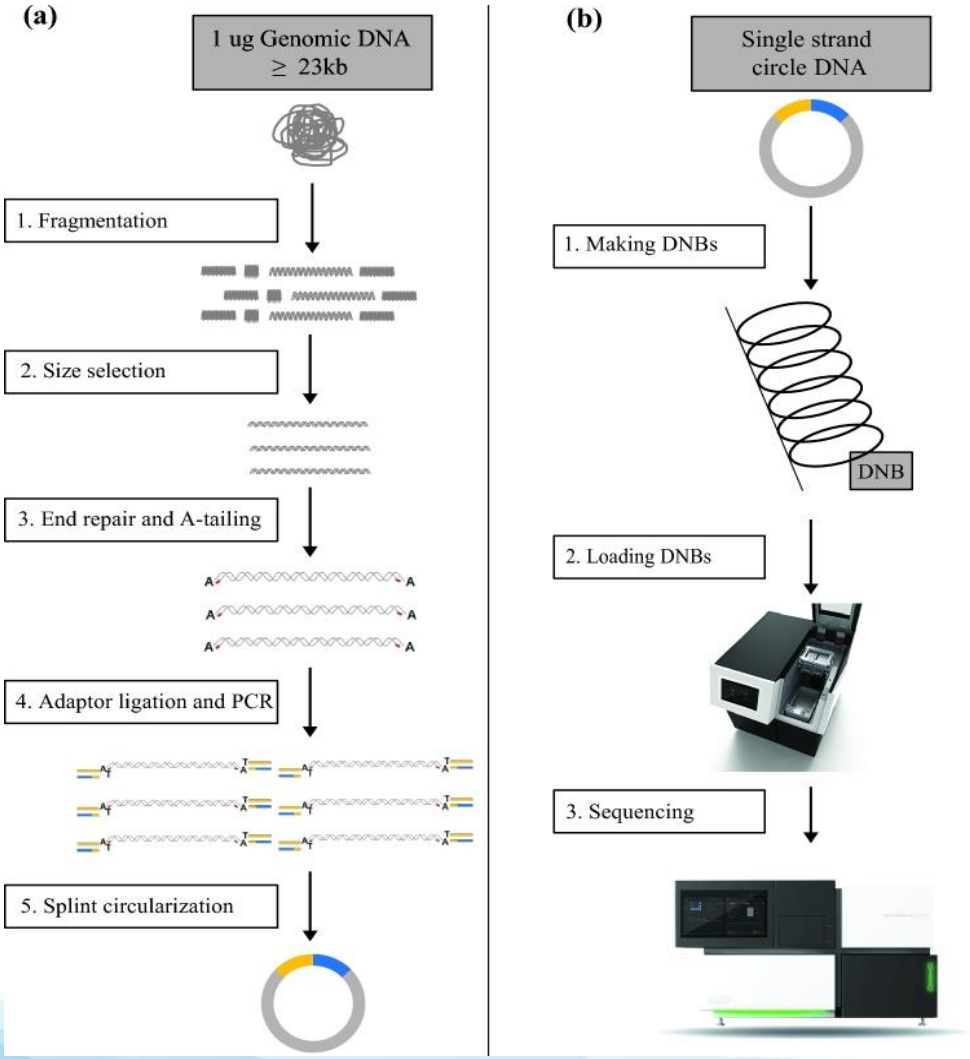
华大基因 BGI

Toward a $100 Genome



Technology enhancements enable lower sequencing costs.

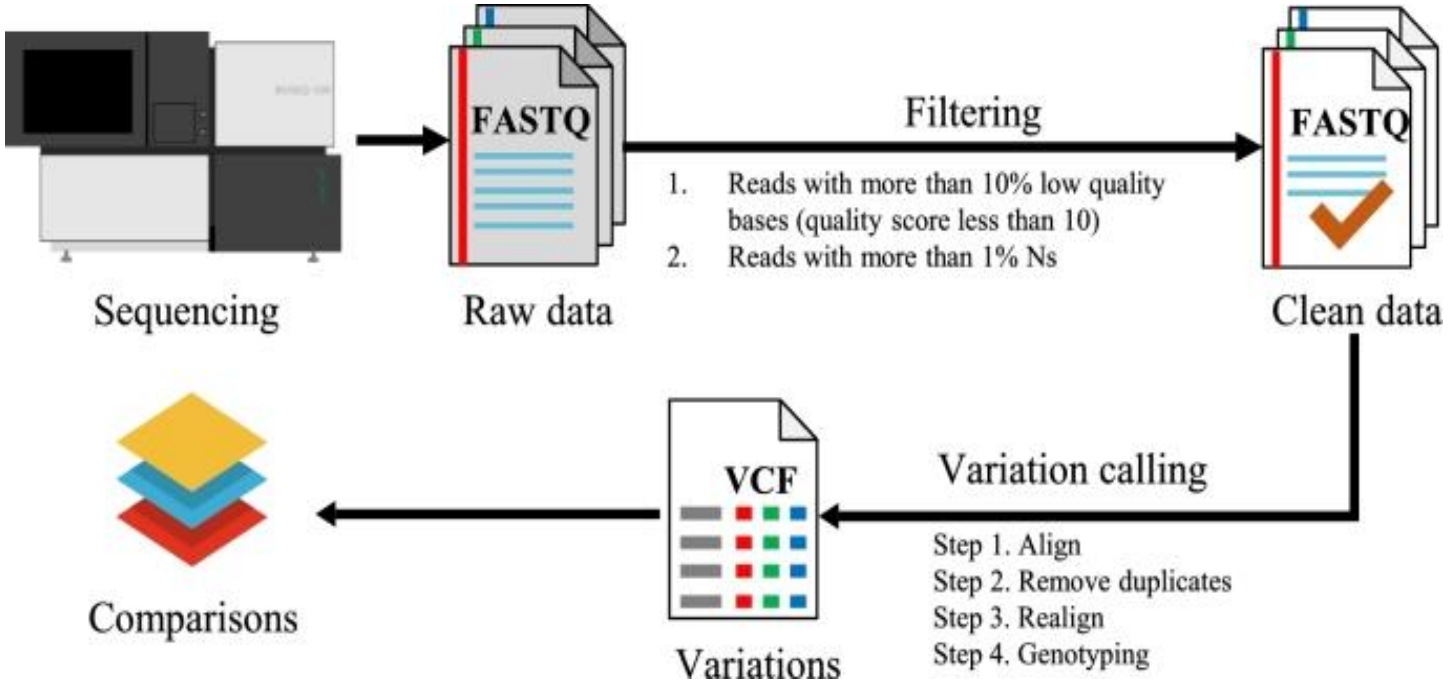
Flowchart of Library Construction and Sequencing



The library construction includes fragmentation, size selection, end repair and A-tailing, adaptor ligation, PCR amplification, and splint circularization (a).

The sequencing includes making DNBs, loading DNBs and sequencing (b).

Variation Calling



Standard data deliverable

Filter original fastq data by ***SOAPnuke***
Fastq to BAM by BWA
BAM to VCF by GATK
VCF File annotation by ***ANNODB***
CNV analysis by CNVnator
SV analysis by ***Seeksv***

BGI in-House Package *:* ***SOAPnuke***
ANNODB
Seeksv

GATK4 for Free

BGISEQ WGS PE100 Data Performance

Methods:

Standard sample Genome in a Bottle's NA12878 (GIAB) ,
3 technical repetition,
Sequenced by BGISEQ-500(PE100) and HiSeq XTen(PE150)

Results:

Table 1: The Q20 and Q30 for different platforms WGS data performance

Item	BGISEQ-500 (average)	HiSeq X Ten (average)
Raw reads	1,002,366,106	778,240,910
Raw bases (Mb)	100,237	116,996
Clean reads	1,001,630,550	732,165,210
Clean bases (Mb)	100,163	110,084
Clean data rate (%)	99.93	95.91
Clean read Q20 (%)	**95.00**	**97.01**
Clean read Q30 (%)	**89.90**	**90.47**
GC content (%)	41.71	40.94

The Q20 and Q30 of BGISEQ-500 data are comparable with HiSeq XTen platform

BGISEQ WGS PE100 Data Performance

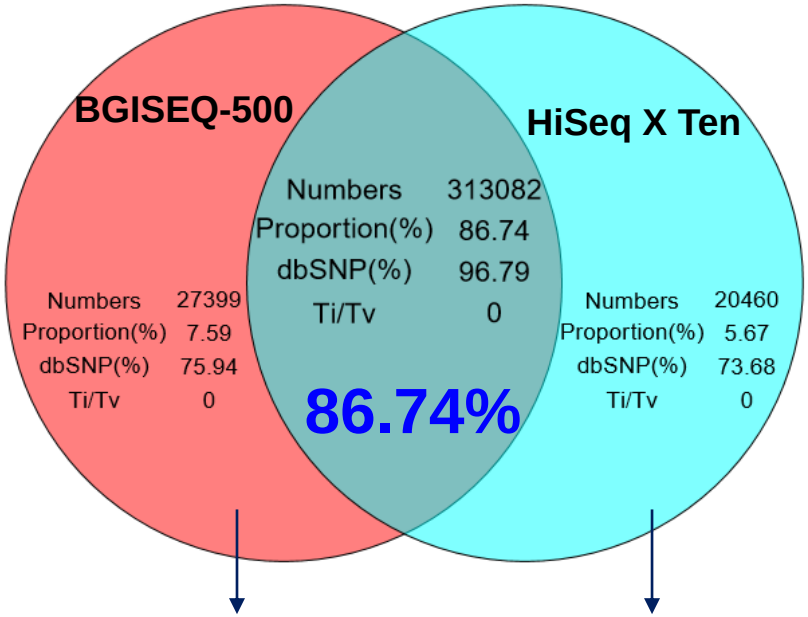
Table 2: The mapping rate and coverage statistics.

Item	BGISEQ-500 (average)	HiSeq X Ten (average)
Clean reads	1,001,630,550	732,165,210
Clean bases(Mb)	100,163	110,083
Mapping rate	99.47%	96.52%
Unique rate	94.33%	85.14%
Duplicate rate	**1.77%**	**11.76%**
Mismatch rate	0.53%	0.56%
Average sequencing depth	33.02	31.57
Coverage	99.10%	98.95%
Coverage at least 4X	98.62%	98.43%
Coverage at least 10X	97.68%	97.24%
Coverage at least 20X	93.09%	91.45%

The mapping rate and coverage are comparable for BGISEQ-500 and HiSeq platforms.

BGISEQ WGS Data Performance

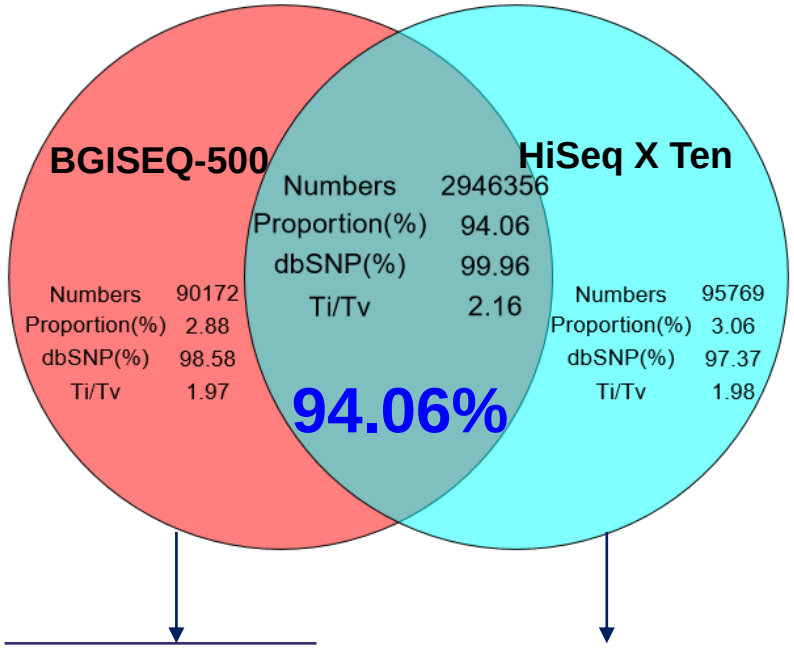
InDel Calling Consistency



TP	**22076**
FP	5270
Precision	80.73

TP	**12813**
FP	7612
Precision	62.73

SNP Calling Consistency



TP	**88890**
FP	1282
Precision	98.58

TP	**93499**
FP	2270
Precision	97.63

Compared to the Illumina HiSeq X Ten platform, the high confidence InDel calling is comparable, and high confidence SNP consistency for both platforms is about 94.06%.

*Data sets

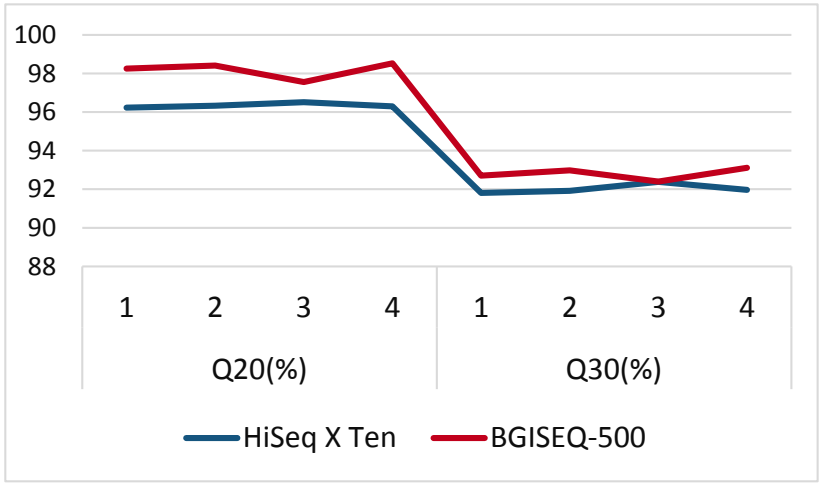
GIAB High-confidence variant calls:

High-confidence SNP, small indel, and homozygous reference calls for the Genome in a Bottle (GIAB) sample NA12878 (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.1/GRCh37/)

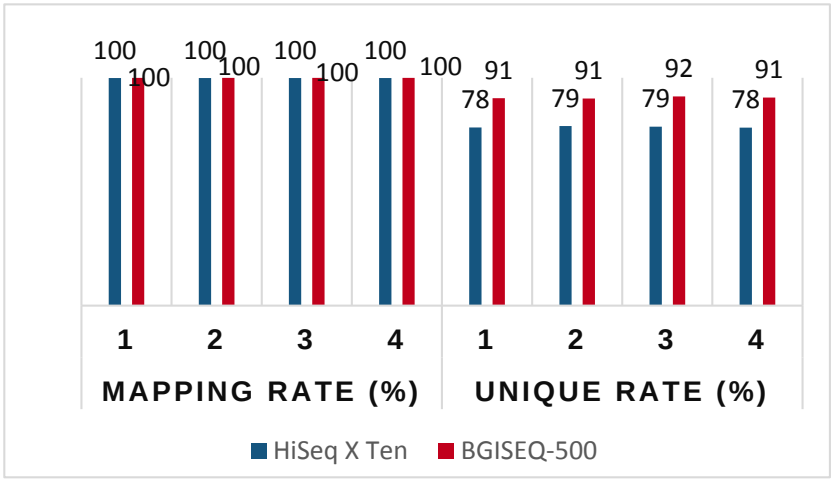
BGISEQ WGS Data Performance

Normal samples compare with Illumina HiSeq X Ten platform

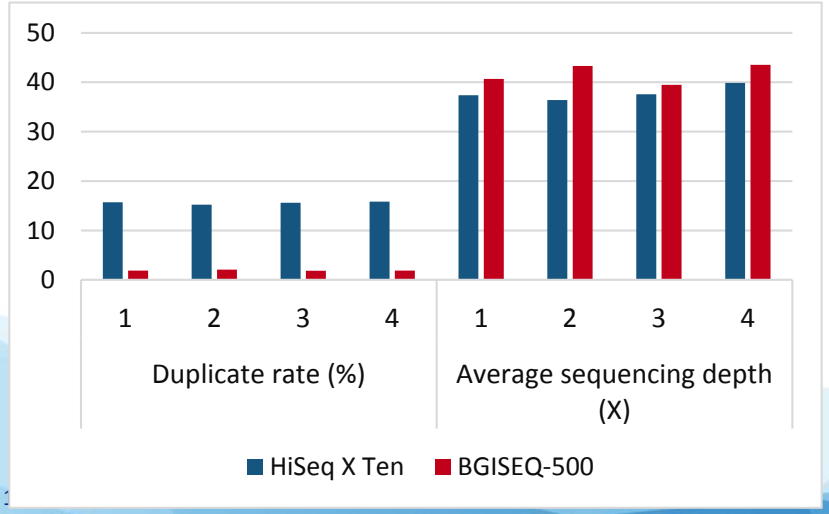
Q20&Q30



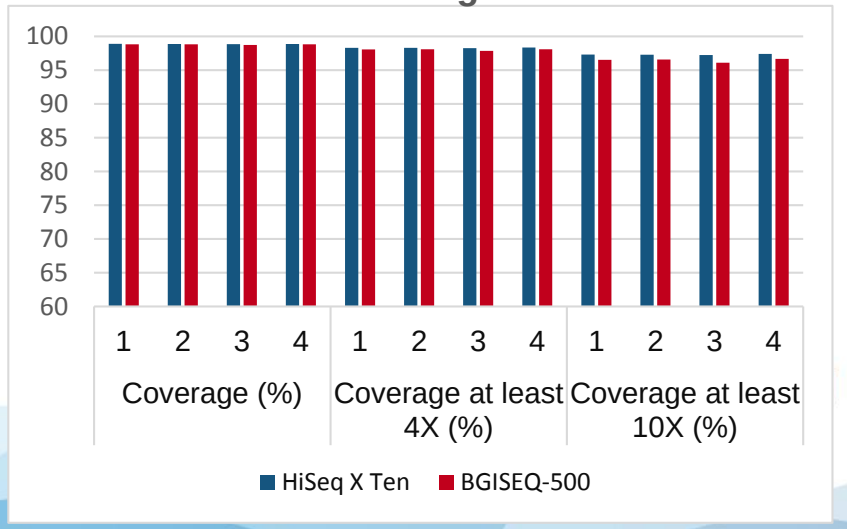
Mapping Rate



Depth and Duplicate Rate



Coverage



BGISEQ Low Pass WGS Data Performance

SNP discovery	Genotype calling accuracy	The power for genetic association studies
MAF > 0.5%: comparable for deep coverage and low coverage MAF =0.2%–0.5%: Low coverage is better. MAF <0.1% : Neither is good	High depth is better, especial for the heterozygote analysis.	With fixed sequencing effort, low-coverage sequencing of more individuals increases power compared to high-coverage sequencing of fewer individuals.

Table 1. Comparison of high-coverage (400 @ 30×) and low-coverage (3000 @ 4×) sequencing design given the same total sequencing effort

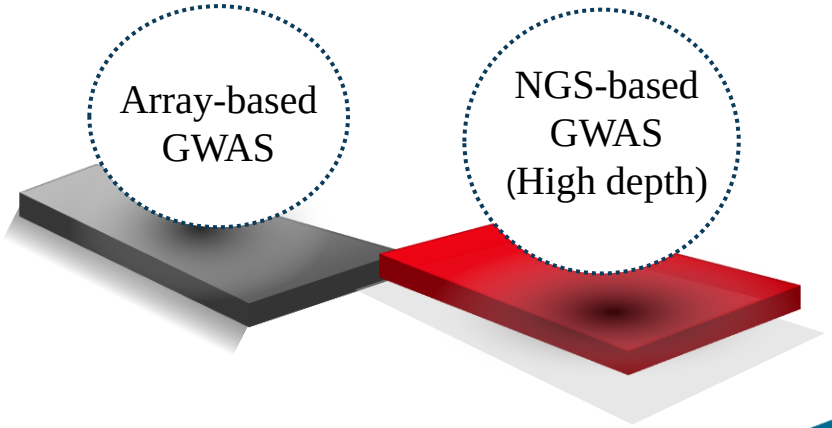
		Population MAF					
Statistic	Design	0.1%–0.2%	0.2%–0.5%	0.5%–1%	1%–2%	2%–5%	>5%
% Discovery	400@30×	65.41%	87.14%	100.00%	100.00%	100.00%	100.00%
	3000@4×	58.15%	94.39%	100.00%	100.00%	100.00%	100.00%
Overall genotypic concordance	400@30×	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	3000@4×	99.87%	99.75%	99.69%	99.75%	99.67%	99.81%
Heterozygote concordance	400@30×	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	3000@4×	82.48%	81.93%	90.39%	97.26%	98.84%	99.85%
Dosage r^2	400@30×	99.49%	99.61%	99.74%	99.81%	99.88%	99.98%
	3000@4×	63.90%	68.97%	80.21%	91.92%	95.77%	99.27%
Information content (nr^2)	400@30×	398	398	399	399	400	400
	3000@4×	1917	2069	2406	2758	2873	2978

% Discovery is the percentage of SNPs detected according to population MAF (MAF defined among 45,000 sequenced chromosomes).

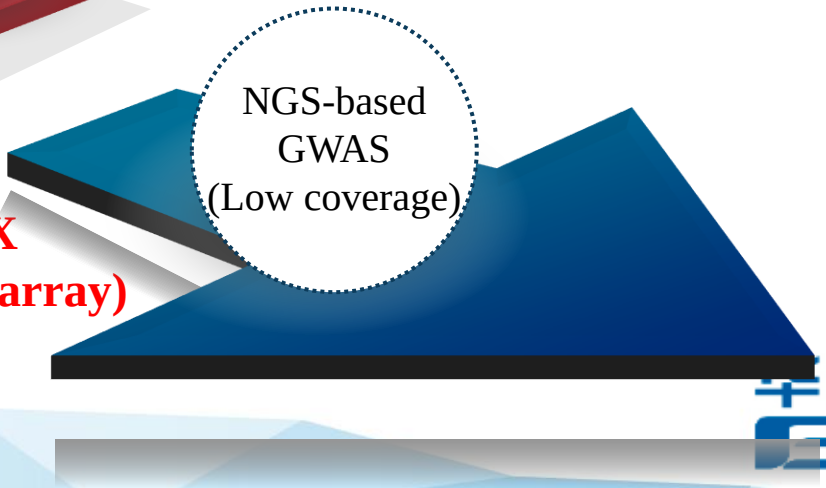
Yun Li, et al. Low coverage sequencing: Implications for the design of complex trait association studies. Genome Res. 21, 940-951 (2011)

BGISEQ Low Pass WGS Data Performance

The affordable designs for sequencing large numbers of samples will be critically important for the successful transition from GWAS to sequencing-based association studies. Given total sequencing capacity is limited, **low coverage sequencing allows much larger numbers of individuals to be studied.**



1. **The ability for SNP detection.**
2. **The accuracy for Genotype calling.**
3. **The power for genetic association studies.**



The cost for BGSIEQ WGS at 3-5X depth ＜100 USD (the cost of SNP array)

BGISEQ Low Pass WGS Data Performance

Methods:

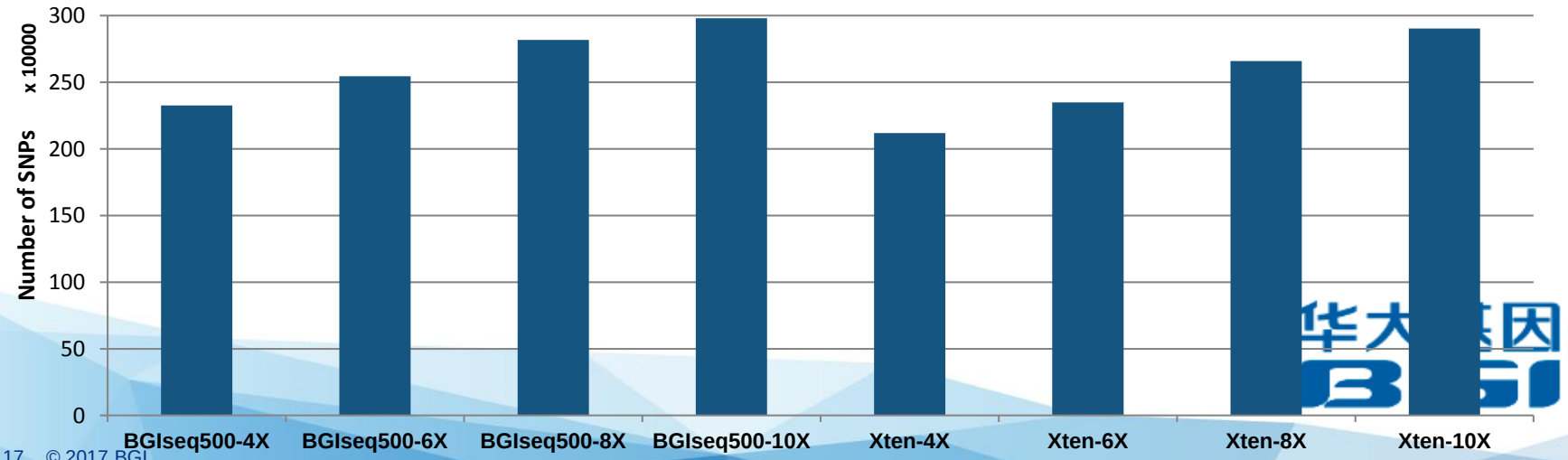
Standard sample Genome in a Bottle's NA12878 (GIAB).
Sequenced by BGISEQ-500(PE100) and HiSeq XTen (PE150) respectively.
Different coverage sequencing: 4X,6X,8X,10X

Sample	Depth	Read length
BGISeq-500 4X	4X	PE100
BGISeq-500 6X	6X	PE100
BGISeq-500 8X	8X	PE100
BGISeq-500 10X	10X	PE100
HiSeq XTen 4X	4X	PE150
HiSeq XTen 6X	6X	PE150
HiSeq Xten 8X	8X	PE150
HiSeq XTen 10X	10X	PE150

BGISEQ Low Pass WGS Data Performance

SNP Calling

Sample	BGIseq-4X	BGIseq-6X	BGIseq-8X	BGIseq-10X	Xten-4X	Xten-6X	Xten-8X	Xten-10X
Total SNPs	2325844	2545485	2817606	2979661	2118614	2349190	2658782	2903087
Fraction of SNPs in dbSNP (%)	98.8	99.68	99.67	99.67	98.66	99.64	99.65	99.59
Fraction of SNPs in 1000genomes (%)	93.55	97.92	97.94	97.96	92.52	97.75	97.88	97.72
Novel	23389	5293	5867	6236	23799	5277	5620	7528
Homozygous	1382832	1345993	1347329	1351595	1355175	1337115	1345229	1353053
Heterozygous	943012	1199492	1470277	1628066	763439	1012075	1313553	1550034
Intronic	900466	1020628	1129106	1192933	808221	937971	1061873	1157559
5' UTRs	2684	3230	3610	3837	2773	3214	3573	3838
3' UTRs	15006	17469	19185	20155	13575	15999	17950	19482
Upstream	32624	35358	39127	41359	30441	33396	37562	40749
Downstream	31896	34288	37829	39970	28984	31977	36154	39185
Intergenic	1322163	1412187	1564254	1655663	1214541	1305504	1478225	1616836
Ti/Tv	2.04	2.07	2.06	2.06	2.04	2.09	2.09	2.08

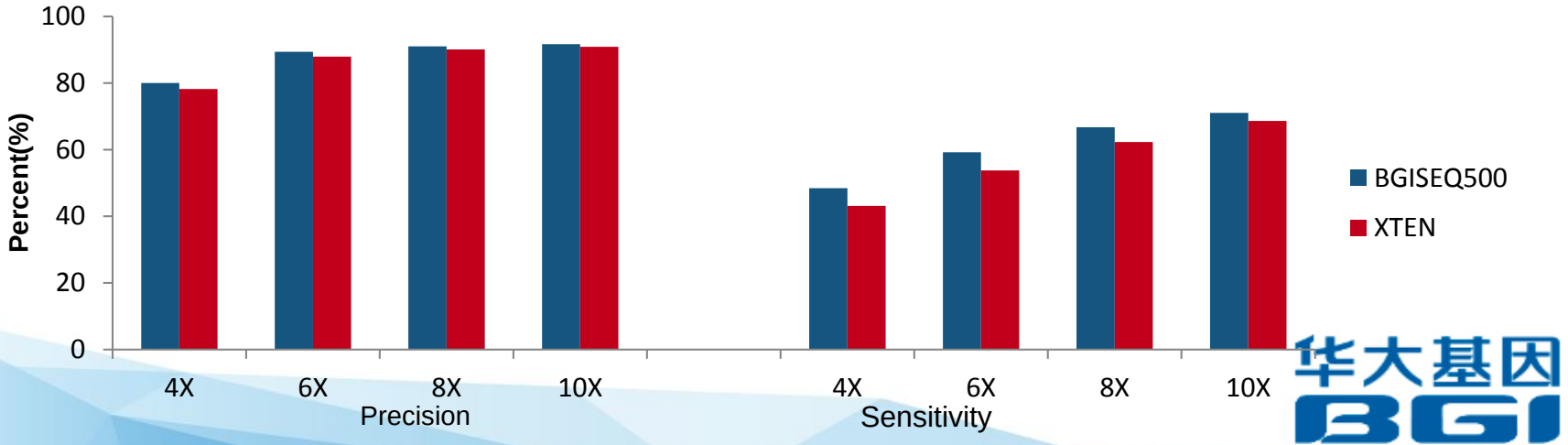


BGISEQ Low Pass WGS Data Performance

Percision & Sensitivity

Sample	True-pos-call	False-pos	False-neg	Precision	Sensitivity	F-measure
BGISeq-500 4X	1861173	464671	1981994	0.8002	0.4843	0.6034
BGISeq-500 6X	2275139	270346	1568031	0.8938	0.592	0.7122
BGISeq-500 8X	2564299	253307	1278870	0.9101	0.6672	0.77
BGISeq-500 10X	2731124	248537	1112042	0.9166	0.7106	0.8006
HiSeq XTen 4X	1657113	461501	2186061	0.7822	0.4312	0.5559
HiSeq XTen 6X	2066135	283055	1777039	0.8795	0.5376	0.6673
HiSeq Xten 8X	2394756	264026	1448414	0.9007	0.6231	0.7366
HiSeq XTen 10X	2638124	264963	1205043	0.9087	0.6864	0.7821

-- using GIAB High confidence variation set (GIAB_HC)

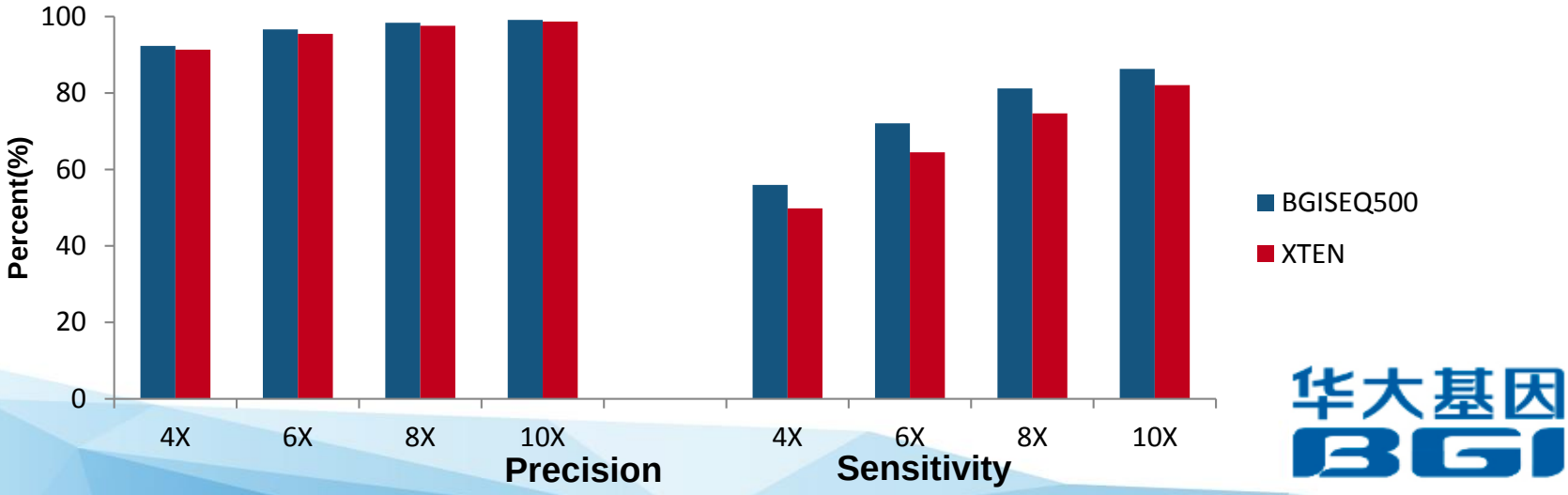


BGISEQ-500 Low Pass WGS Data Performance

Percision & Sensitivity

Sample	True-pos-call	False-pos	False-neg	Precision	Sensitivity	F-measure
BGISeq-500 4X	416945	34697	328462	0.9232	0.5594	0.6966
BGISeq-500 6X	537216	18575	208191	0.9666	0.7207	0.8257
BGISeq-500 8X	605338	9840	140069	0.9840	0.8121	0.8898
BGISeq-500 10X	643408	5719	101999	0.9912	0.8632	0.9228
HiSeq XTen 4X	371255	35223	374152	0.9133	0.4981	0.6446
HiSeq XTen 6X	480574	22770	264833	0.9548	0.6447	0.7697
HiSeq Xten 8X	556613	13680	188794	0.9760	0.7467	0.8461
HiSeq XTen 10X	611785	8163	133622	0.9868	0.8207	0.8962

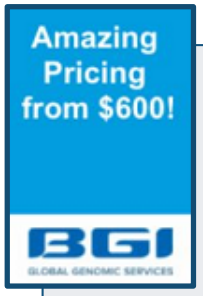
-- using Omni2.5 as reference dataset



Highlights of BGISEQ Based WGS

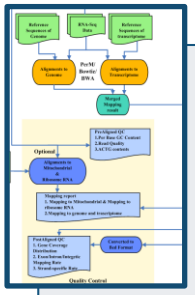
More Diversity Applications with Better Data Performance

- Comparable with HiSeq platform
- Low duplication & high unique mapping rate



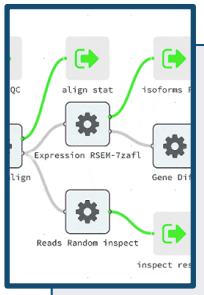
Cost-Effective

- Get the same quality data at a lower price



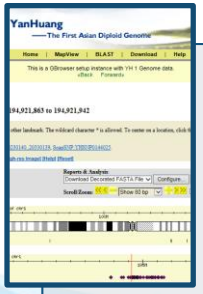
Standard Format

- Industry-recognized data format
- Other flexible data analysis



Cloud Delivery & Analysis

- BGIOnline + Edico Genome
- 100WGS analysis within 2hrs



Analysis capabilities

- BGI In-House Package and Database
- GATK 4 for free

Quality Control

- Strict QC system
- Contamination Fingerprinting via MS genotyping

华大基因
BGI

Case 1.
Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing

Publication by QIMR Berghofer/BGI accepted into *PLOS One*:

Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing

Methods

- Three patients (identified as 9869, 11202 and 11398) diagnosed with malignant pleural mesothelioma;
- BGISEQ-500 & HiSeq platform sequencing;
- Data from the BGISEQ-500 and HiSeq X Ten was analysed using the same pipeline.

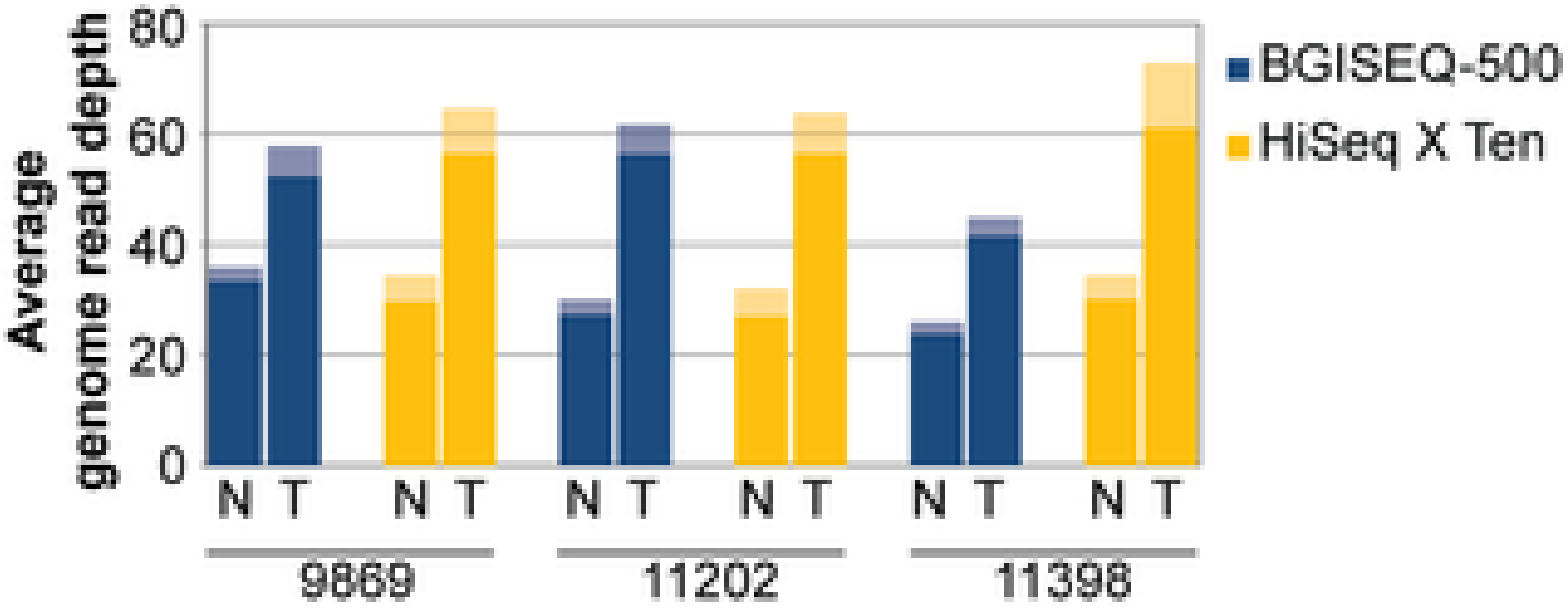


Fig 1. Average genome read depth using BGISEQ-500 and HiSeq X Ten data

Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. PLOS ONE 13(1): e0190264. https://doi.org/10.1371/journal.pone.0190264
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190264

Results

Table 1. The percent concordance of germline genotypes ascertained by SNP arrays compared to the BGISEQ-500 and HiSeq X Ten data.

Patient	SNP array vs BGISEQ-500	SNP array vs HiSeq X Ten
9869	99.797	99.789
11202	99.794	99.794
11398	99.797	99.795

https://doi.org/10.1371/journal.pone.0190264.t001

The sequence data generated on the BGISEQ-500 and the HiSeq X Ten platforms showed a >99% concordance with the genotypes obtained from the Illumina SNP arrays, indicating that both platforms were able to accurately detect common germline SNV assayed by the SNP arrays.

Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. PLOS ONE 13(1): e0190264. https://doi.org/10.1371/journal.pone.0190264
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190264

Table 2. Number of germline and somatic variants identified in three mesothelioma samples using whole genome sequencing.

		SNV				Indels			
		9869	11202	11398	All Patients	9869	11202	11398	All Patients
Germline	Identified in both platforms	3,033,980	3,146,317	3,092,543	9,272,840	193,359	190,436	185,905	569,700
		(96.8%)	(96.8%)	(96.8%)	(96.8%)	(91.7%)	(91.8%)	(92%)	(91.8%)
	HiSeq X Ten only	313,015	321,627	407,966	1,042,608	33,143	35,253	41,480	109,876
		(42.3%)	(42.3%)	(41.9%)	(42.1%)	(58.5%)	(58.4%)	(59.2%)	(58.7%)
	BGISEQ-500 only	161,128	118,336	92,050	371,514	7,025	6,931	5,789	19,745
		(4%)	(2.4%)	(4.1%)	(3.55%)	(13.8%)	(13.8%)	(11.6%)	(13.1%)
	Total	3,508,123	3,586,280	3,592,559	10,686,962	233,527	232,620	233,174	699,321
Somatic	Identified in both platforms	3,554	2,342	1,955	7,851	197	168	114	479
	HiSeq X Ten only	697	424	411	1,532	135	93	78	306
	BGISEQ-500 only	540	474	493	1,507	102	156	229	487
	Total	4,791	3,240	2,859	10,890	434	417	421	1,272

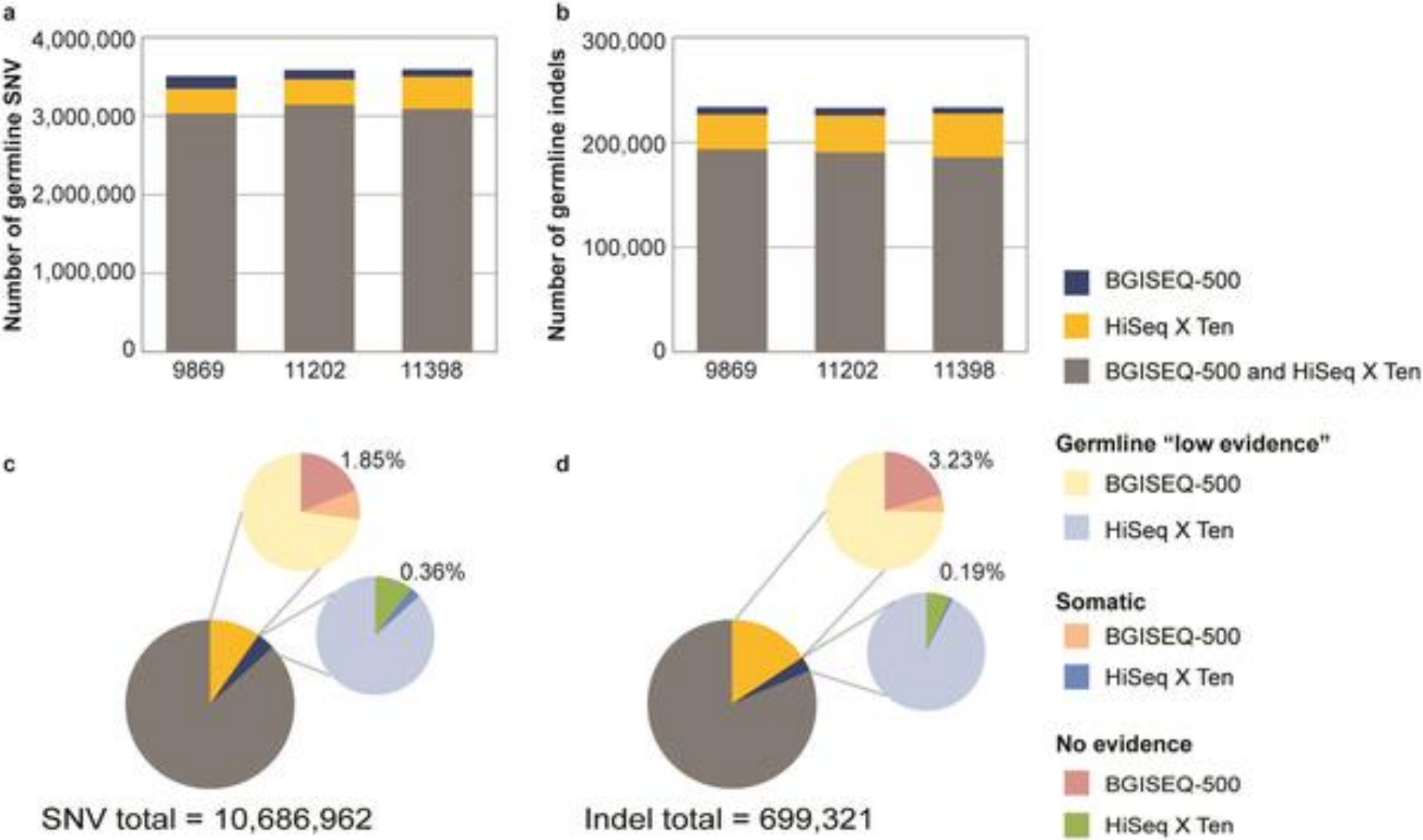
https://doi.org/10.1371/journal.pone.0190264.t002

PLOS ONE | TENTH ANNIVERSARY

Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. PLOS ONE 13(1): e0190264. https://doi.org/10.1371/journal.pone.0190264
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190264

华大基因 BGI

Fig 2. Germline variants identified in three mesothelioma samples (patients: 9869, 11202 and 11398) using BGISEQ-500 and HiSeq X Ten data.



Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. PLOS ONE 13(1): e0190264. https://doi.org/10.1371/journal.pone.0190264

Content

BGI Overview

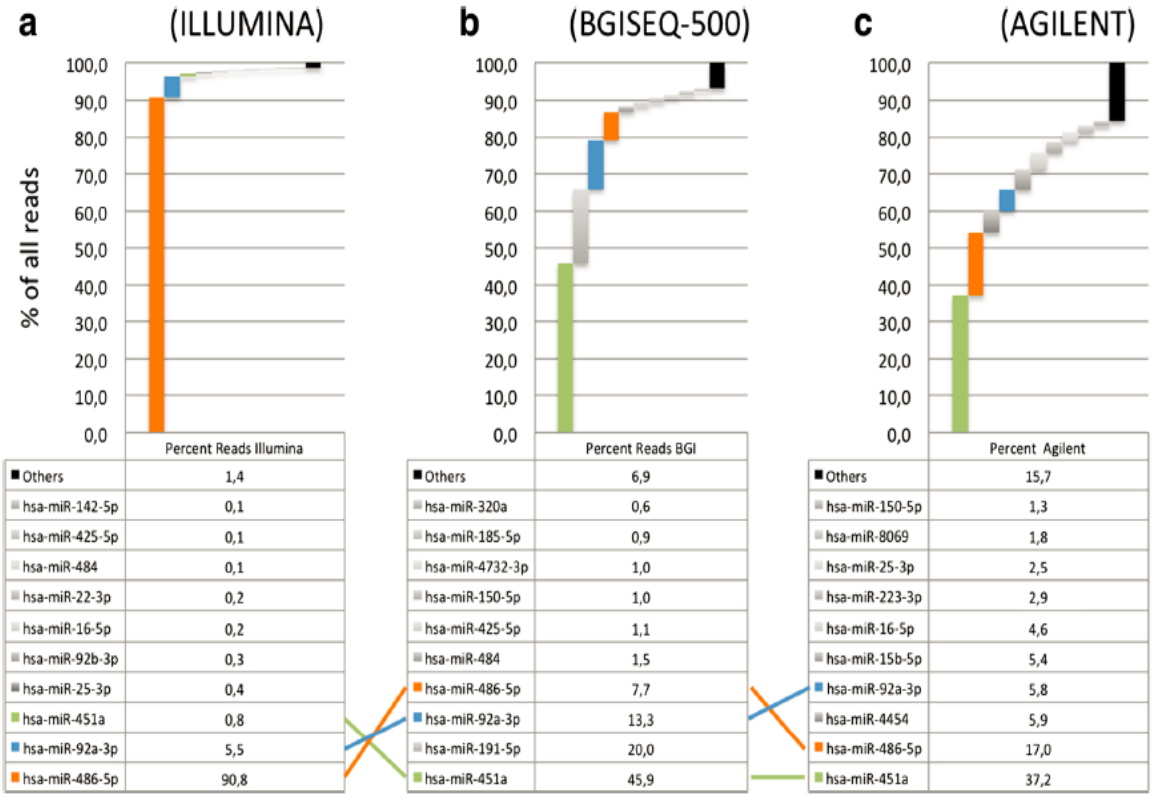
BGISEQ Based WGS Solutions

More Applications for BGISEQ Platform

The Case of BGISEQ Based Small RNA Sequencing

RESEARCH Open Access

cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs



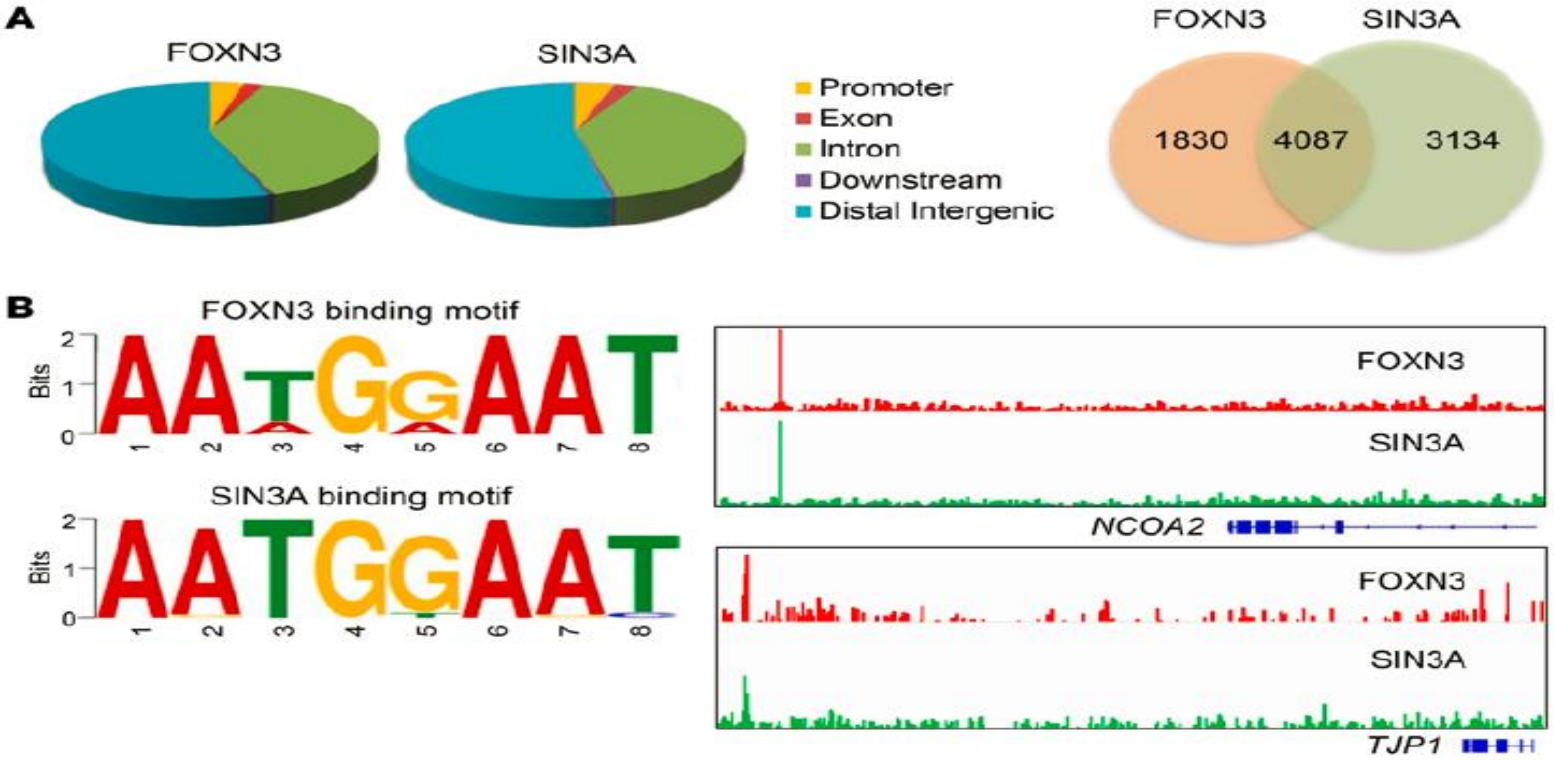
Percent Reads Illumina	
Others	1,4
hsa-miR-142-5p	0,1
hsa-miR-425-5p	0,1
hsa-miR-484	0,1
hsa-miR-22-3p	0,2
hsa-miR-16-5p	0,2
hsa-miR-92b-3p	0,3
hsa-miR-25-3p	0,4
hsa-miR-451a	0,8
hsa-miR-92a-3p	5,5
hsa-miR-486-5p	90,8

Percent Reads BGI	
Others	6,9
hsa-miR-320a	0,6
hsa-miR-185-5p	0,9
hsa-miR-4732-3p	1,0
hsa-miR-150-5p	1,0
hsa-miR-425-5p	1,1
hsa-miR-484	1,5
hsa-miR-486-5p	7,7
hsa-miR-92a-3p	13,3
hsa-miR-191-5p	20,0
hsa-miR-451a	45,9

Percent Agilent	
Others	15,7
hsa-miR-150-5p	1,3
hsa-miR-8069	1,8
hsa-miR-25-3p	2,5
hsa-miR-223-3p	2,9
hsa-miR-16-5p	4,6
hsa-miR-15b-5p	5,4
hsa-miR-92a-3p	5,8
hsa-miR-4454	5,9
hsa-miR-486-5p	17,0
hsa-miR-451a	37,2

Expression distribution of the 10 miRNAs with the highest detection in the blood RNA on the HiSeq system (a), BGISEQ-500 (b), and microarray system (c). It showed more even coverage of lower abundant miRNA species on BGISEQ-500 system, which facilitates the discovery of new miRNAs.

The Case of BGISEQ Based ChIP-Seq



FOXN3- and SIN3A-associated DNAs were sequenced using BGISEQ-500. The experiments identified a total of 4,087 genes targeted by the FOXN3-SIN3A complex (A, right). Significantly, FOXN3 and SIN3A had very similar binding motifs, and FOXN3 and SIN3A exhibited similar peak locations on the representative target genes (B), supporting the physical interaction and functional connection between FOXN3 and SIN3A.

JCI The Journal of Clinical Investigation

J Clin Invest. 2017;127(9):3421–3440. doi:10.1172/JCI94233.

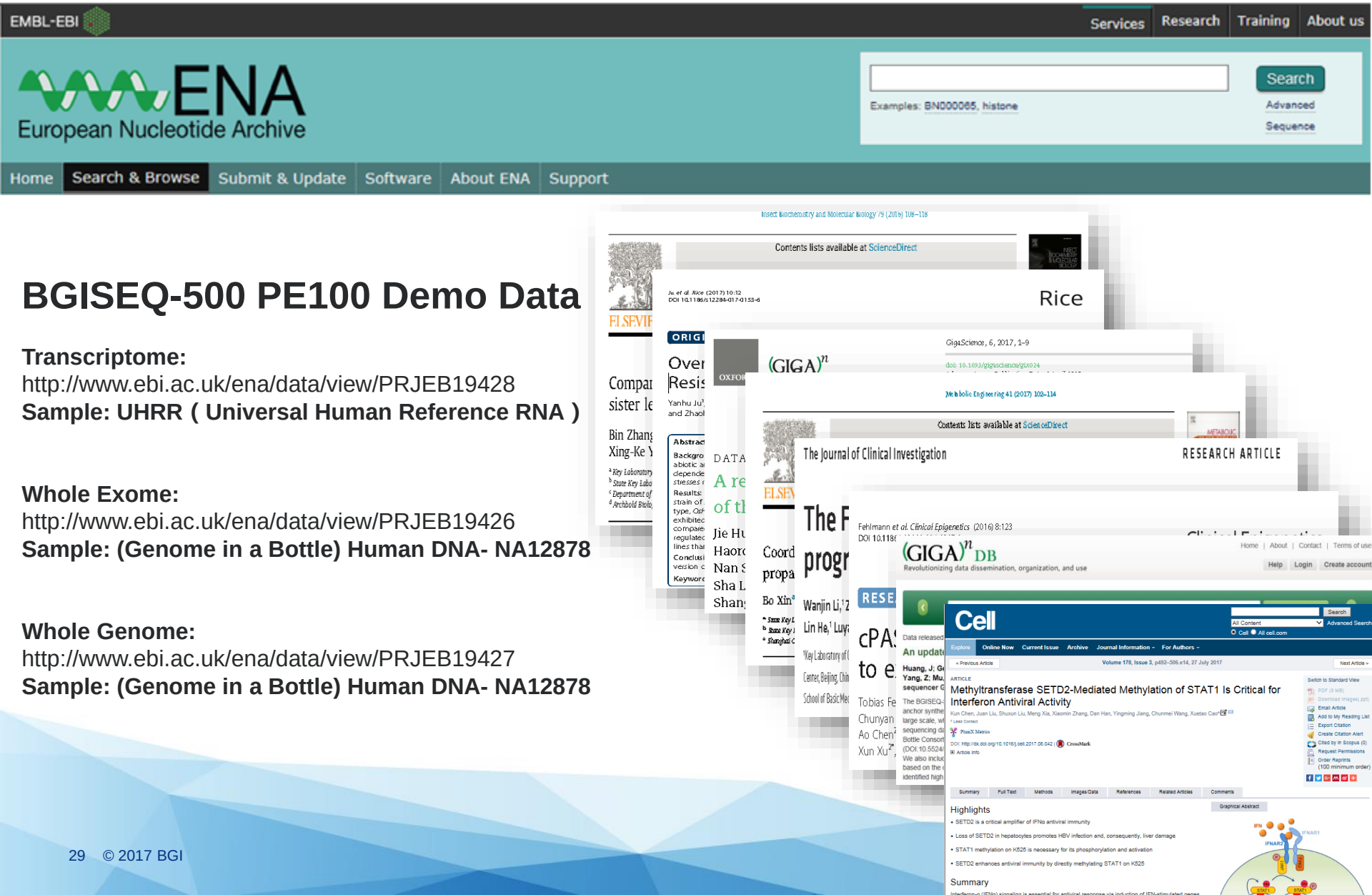
BGISEQ Platform *Demo* Data

BGISEQ-500 PE100 Demo Data

Transcriptome:
http://www.ebi.ac.uk/ena/data/view/PRJEB19428
Sample: UHRR (Universal Human Reference RNA)

Whole Exome:
http://www.ebi.ac.uk/ena/data/view/PRJEB19426
Sample: (Genome in a Bottle) Human DNA- NA12878

Whole Genome:
http://www.ebi.ac.uk/ena/data/view/PRJEB19427
Sample: (Genome in a Bottle) Human DNA- NA12878



THANK YOU

Yonggang (Jason) Zhao

Tech-service BU Director

zhaoyonggang@bgi.com

源資生技
Tri-I Biotech,Inc.
Total Solution Provider of Systems Biology and Chemoinformatics

info@tri-ibiotech.com.tw
02-26432031