An integrated Asian human SNV and indel benchmark combining multiple sequencing methods

Chuanfeng Huang[1,] *, Libin Shao[2,] *, Shoufang Qu[1,] *, Junhua Rao[3,] *, Tao Cheng[4], Zhisheng Cao[5], Sanyang Liu[6], Jie Hu[2], Xinming Liang[3], Ling Shang[4], Yangyi Chen[7], Zhikun Liang[8], Jiezhong Zhang[6], Peipei Chen[5], Donghong Luo[7], Anna Zhu[8], Ting Yu[1], Wenxin Zhang[1], Guangyi Fan[2,9,10], Fang Chen[3, †], Jie Huang[1, †]

1.National Institutes for food and drug Control (NIFDC), No.2, Tiantan Xili Dongcheng District, Beijing 10050, P. R. China
2.BGI-Qingdao, BGI-Shenzhen, Qingdao, Shandong, 266555, P. R. China
3.MGI, BGI-Shenzhen, Shenzhen, Guangdong,518083, P. R. China
4.BerryGenomics Co., Ltd. Building #5, 4 Science Park Road, ZGC Life Science Park, Beijing,102200, P. R. China
5.Tianjin Novogene Bioinformatic Technology Co., Ltd. Entrepreneurial Headquarters Base B07-B09, Wuqing Development Zone, Tianjin, 301700, P. R. China
6.Annoroad Gene Technology, Building B1, Yard 88, kechuang 6 Rd, Beijing Economic-Technological Development Area, Beijing,102200, P. R. China
7.CapitalBio Genomics Co., Ltd., Building 11, GuanTai Biotechnology Cooperation Incubation Center, No.1, Taoyuan Road, Songshan Lake Hi-Tech Industrial Development Zone, Dongguan, Guangdong,523808, P.R. China
8.Guangzhou Daruia Biotechnology Co. Ltd., 5 buildings No. 11 Nanxiang Third Road, Science City, Luogang District, Guangzhou, Guangdong, 510663, P.R. China
9.BGI-Shenzhen, Shenzhen, Guangdong,518083, P.R. China
10. China National GeneBank, BGI-Shenzhen, Shenzhen Guangdong,518120, P.R. China

*These authors contributed equally to this work.
†Correspondence authors: Fang Chen(fangchen@genomics.cn), Jie Huang (jhuang5522@126.com)

Abstract

Precision medicine of human requires an accurate and complete reference variant benchmark for different populations. A human standard cell line of NA12878 provides a good reference for part of the human populations, but it is still lack of a fine reference standard sample and variant benchmark for the Asians. Here, we constructed a stabilized cell line of a Chinese Han volunteer. We received about 4.16T clean data of the sample using eight sequencing strategies in different laboratories, including two BGI regular NGS platforms, three Illumina regular NGS platforms, two linked-read libraries, and PacBio CCS model. The sequencing depth and reference coverage of eight sequencing strategies have reached the saturation. We detected small variants of SNPs and Indels using the eight data sets and obtained eight variant sets by performing a series of strictly quality control. Finally, we got 3.35M SNPs and 349K indels supported by all of sequencing data, which could be considered as a high confidence standard small variant sets for the studies. Besides, we also detected 5,913 high quality SNPs located in the high homologous regions supported by both linked-reads and CCS data benefited by their long-range information, while these regions are recalcitrant to regular NGS data due to the limited mappability and read length. We compared the later SNPs against the public databases and 969 sites of them were novel SNPs, indicating these SNPs provide a vital complement for the variant database. Moreover, we also phased more than 99% heterozygous SNPs also supported by linked-reads and CCS data. This work provided an integrated Asians SNV and indel benchmark for the further basic studies and precision medicine.

Keywords

Reference standard NGS Linked-read CCS Benchmark

Introduction

Thousands of human genomes are now available and whole genome sequencing (WGS) is likely to become a routine part of medical care in many countries. WGS data allows the identification of genetic changes associated with disease and paves the way for precision medicine, medical care customized according to the genetic make-up of a patient [1]. Diseases are often associated with particular single nucleotide variants (SNVs), or insertion or deletion events (indels) [2, 3]. In order to fully capitalize on the vast genome data generated, reference genomes are required to allow genome comparisons and benchmarking of new sequencing technologies and analysis methods. The current human reference genome (NA12878) is from a Caucasian from the U.S. state of Utah. Significant insights have been gained from NA12878, but it is appreciated that reference genomes from additional populations are needed [4]. Several Asian genomes are now available from individuals of Chinese [5] ,Korean [6] and Pakistani [7] descent. However, the majority of these studied used next-generation sequencing (NGS) platforms to generate short reads and could not resolve SNVs and indels located in complex regions .For example, targeted DNA-HiSeq [8] identified 1,281 SNVs in 193 genes in the Asian reference sample YH that could not be detected in the original study [5].The 193 genes are associated with hereditary diseases with a higher incidence in the Chinese population, a clear example of the need for high quality reference genomes in addition to NA12878 [7]. It is now apparent that a combination of long read, short read, and linked-read sequencing is required to fully characterize human reference genomes[9]. Herein, we generated an Asian SNV and indel benchmark genome by combining diverse short and long read sequencing platforms, an approach which could balance the systematic sequencing bias of different platforms.

Results

Sequencing and quality control

To develop a represented Asian high-quality genotype call sets, we recruited a Han Chinese volunteer from Beijing City (Research ethics ID: XHEC-C-2019-086, HJ). We sequenced this individual using five frequently-used NGS short-read sequencing platforms (BGISEQ-500, MGISEQ-2000, NextSeq-CN500, NextSeq550Dx and NovaSeq6000; three technical replicates), single tube long fragment read (stLFR) sequencing[10], 10X Genomics Chromium linked-read sequencing[11], PacBio single molecule real-time circular consensus sequencing (SMRT CCS) long-read sequencing[12] and Oxford Nanopore MinION sequencing [13]. After processing (**Figure 1**), we generated 3.12Tb high quality general NGS data for HJ totally. This included an average coverage of 86.58× from 2×100bp reads on two BGISEQ-500 sequencers and 60.07× from 150bp reads on three Illumina sequencers. We obtained 250.78 Gb (~51.97×) stLFR data with a molecular length of 117,499 bp, 277.60 Gb (~84.7×) 10X Genomics Chromium data with a molecular length of 191,294bp, and 77.23 Gb (~24.4×) PacBio CCS data with mean read length of 12.09 kb. For the general NGS data, 99.88% of raw reads could be mapped to the human reference genome (hs37d5) with coverage of 99.92% and 85.75% of mapped reads were uniquely mapped reads. For the stLFR data and 10X Genomics Chromium data, we aligned 99.35% and 99.71% of them against the reference genome with 98.86% and 98.90% coverage, respectively. For the CCS reads, it could be unambiguously mapped to reference genome at 93.18% coverage (**Figure S1, Table S1**).

SNV and indel detection

To find the saturated sequencing depth of the different platforms, we hierarchically detected SNVs and indels by randomly extracting alignment results from the bam by picard. We found that 30× depth sequencing ensured a consistent rate of uniquely mapped reads (~99%) and number of SNVs (~3.84 M) and indels (~897 K) (**Figure 2, Figure S2**).

We also evaluated the consistency of BGI and Illumina short sequence reads generated from short-insert libraries on the same and different instruments. We found 3.26%, 95.49%, and 1.25% of SNVs could be detected, suggesting that the choice of short-read sequencing platform introduces little bias (**Figure S3**). Nevertheless, despite an adequate sequencing depth, ~33.62 Mb of the genome could not be resolved by short-read BGISEQ-500 and Illumina data (**Figure 3, Table S2**). These regions were assigned into 51,612 blocks, with an N50 of 3,942 bp, and correspond to highly homologous regions (HHRs), of which were previously reported as recalcitrant to short-read NGS sequencing[14]. Interestingly, 73.3%, 65.41% and 68.53% of these HHRs were accessible by stLFR, 10X Genomics Chromium, and PacBio SMRT CCS data, technologies which benefits from barcoding information of linked-reads or long reads (**Table S3**). We detected 3.87M, 3.47 M, and 3.80M SNPs, along with 822K, 721K, and 797K indels using stLFR, 10X Genomics Chromium, and PacBio SMRT CCS data, respectively (**Table S1**). We next wished to characterize SNVs and indels in the HJ sequencing data that could not be mapped to the Caucasian reference genome, even with long-read sequencing data. We focused on HHRs and uniquely mapped regions (UMRs).

Consistence of SNPs and indels in UMR

In the uniquely mapped regions (UMRs) 1,712,393 SNPs and 186,641 indels could be detected by all eight sequencing methods. This is less than the average number of variations (~3.72M SNPs and ~859 K indels). Unexpectedly, 10X Genomics Chromium missed ~1.63M SNPs which could be detected by all of the other methods (**Figure S4**). We, therefore, excluded this data in the further analysis, retaining ~3.35M high quality common SNPs supported by seven sequencing methods. PacBio SMRT CCS detected 234.46K specific SNPs and 240.74K specific indels; stLFR 210.45K and 223.25K; BGISEQ-NGS 11.78K and 71K; and Illumina-NGS 5.57K and 1.98K (**Figure 4, Figure 5,**). We compared the SNP quality distribution between specific SNPs and whole SNPs and found that the quality of the majority of specific SNPs were lower than whole SNPs, likely stemming from sequencing method bias Interestingly, PacBio SMRT CCS and stLFR consistently resulted in high quality variant calls (**Figure S5**).

Accessibility of SNPs and indels in HHRs

A total of 74.7K SNPs and 23.4K indels could be called by both stLFR and PacBio SMRT CCS data but not by the five short-insert library, short-read methods. These variants, nonsynonymous, were located on 129 genes and were significantly enriched for the gene ontology (GO) categories olfactory receptor activity, IgG binding, transmembrane signaling receptor activity, G protein-coupled receptor activity, molecular transducer, and signaling receptor activity pathways. We speculate that these variants are associated with immune disease in Chinese population. Among all special SNPs, 7.9% (5,913/74,717）located in HHRs, with 69 SNPs in coding regions and 19 SNPs in UTR regions. We also performed function enrichment analysis, revealing three genes significantly enriched in blood antigen-related or immune response (LILRB3, RHD, and RHCE) pathways involved into immune response diseases.

Highly homologous or repetitive regions on the genome, NGS is difficult to fully cover due to its read length, which may lead to false negative of mutations, but stLFR and CCS perform well. Complex genes are hard to be covered by NGS platforms, while linked-reads method and long reads sequences platforms do well in detecting the regions. For example, IGV shows a typical gene NBPF4, who is a member of the neuroblastoma breakpoint gene family (NBPF) which consists of dozens of recently duplicated genes primarily located in segmental duplications on human chromosome 1 (**Figure 6**).Another gene is NAIP which is part of a 500kb reverse replication on chromosome 5q13, contains at least four repeated elements and genes, and making it easy to rearrange and delete. The repeatability and complexity of the sequences also make it difficult to determine the organization of this genomic region. It is thought that this gene, modifier of spinal muscular atrophy, is a mutation in a neighboring gene SMN1. Variations detected on NAIP for NGS platform are relative small and nearly included in linked reads and long reads platforms (**Figure S6**). In addition to the genes mentioned above, there is XAGE2 (**Figure S7**), and other genes.

Haplotype phasing small variants

Human genomes are diploid, with chromosome pairs from each parent. However, most paired-end reads cannot assign variants to a particular chromosome, resulting in a combined haplotype (genotype) [15]. The popular NGS sequencing technology is all about shuffling sequences together for sequencing. After sequenced, we cannot directly distinguish which of these sequences is the parent source. It is only after phasing that we are able to make this distinction. Phasing is strongly correlated with functional interpretation of genetic variation. Therefore, due to the BGI and Illumina short sequence reads generated from short-insert libraries, we using long-range information from PacBio SMRT CCS and stLFR data to phasing, 99.63% and 99.91% of heterozygous SNPs could be phased into 19,584 and 1,262 blocks, respectively. Of these, 1.96 M were shared, with a phasing N50 of more than 11.26 Mb and 388.5k. What's more, some of chromosomes (such as Chr5 and Chr6) were almost completely phased (**Table 1**). According to the results of phasing, stLFR data performed better, so we can be assumed that the long range reads may a good choice in phasing process.

Discussion

Gene sequencing is an important part of precision medical, widely used in detection and diagnosis of various diseases, and brought potential benefits to patients. However, the NGS also some deficiencies, such as short reads, structure mutation detection, especially about the detection of HHR area will miss part of the results, thus caused by false negatives. There is currently a lack of a standard data set that represents Asian populations due to ethnic differences. In this paper, a Han Chinese adult male was recruited and 8 sequencing platforms were used to detect and compare SNV and indel. Finally, we identified a standard data set which contains 3.35M SNPs and 349K indels.

We compared and contrasted eight sequencing platforms to generate an Asian human SNV and indel benchmark. Unexpectedly, and found that 10X Genomics Chromium data did not correlate well with data generated by the other platforms, the reasons for this are not clear. However, a total of 3.35M high quality SNPs were supported by seven other methods, while linked-read stLFR and long-read PacBio SMRT CCS resolved an additional 74.7K SNPs in highly homologous regions, providing a comprehensive small variation benchmark of an Asian human. stLFR and CCS can be well supplemented and improved on the basis of NGS results.

In summary, NGS results will miss some mutations in the HHR region. By adding analysis results of stLFR and CCS platforms, standard data sets and high confidence regions that are considered relatively reliable can be obtained. This data set can be well used for further study. In order to improve the data set, it may be necessary to add samples and analysis methods for integrated analysis.

Methods

Sample collection

This study was carried out in accordance with relevant guidelines and regulations, in line with the principles of the Helsinki declaration[16] and was approved by the institutional review committee (IRB) of BGI. In this experiment, cell line genomic DNA was prepared from the National Institutes for food and drug Control (NIFDC), and it contained 10μg per tube. Used Qubit 3.0 to quantified the genomic DNA and agarose gel to make sure the genomic DNA molecular was not substantially degraded.

Library and sequencing

NGS library construction adopts the normal NGS construction process. The difference between BGISEQ-500 and Illumina platform is that the former involves rolling amplification while the latter use PCR amplification technology. In particular, the BGISEQ-500 library protocol contain three steps: including making DNA nanoballs (DNBs), loading DNBs, and sequencing. Single tube long fragment read (stLFR) library construction physically breaks the DNA into fragments of about 50Kbps, and then use Tn5 transposase for library construction, so that each identical fragment bears the same barcode[10], while 10X Genomics Chromium library construction uses microdroplets where, after the ligation step, PCR is performed and the library is ready to enter any standard next generation sequencing (NGS) workflow.

Large-insert single molecule real-time circular consensus sequencing (SMRT CCS) library preparation was conducted following the Pacific Biosciences recommended protocols[17]. In brief, a total of 60μg genomic DNA was sheared to ~20kb targeted size by using Covaris g-TUBEs (Covaris). Each shearing processed 10μg input DNA and a total of 6 shearings were performed. The sheared genomic DNA was examined by Agilent 2100 Bioanalyzer DNA12000 Chip (Agilent Technologies) for size distribution and underwent DNA damage repair/end repair, blunt-end adaptor ligation followed by exonuclease digestion.

NGS data preprocess

Data filter: SOAPnuke (version 1.5.6) was used to pre-process the 15 NGS data by removing reads with (1) adaptor contaminations, (2) more than 10% low-quality bases (quality < 10), (3) more than 10% N bases.

Mapping and variant calling: All NGS reads were mapping to the human reference genome (hs37d5) using BWA 0.71.5 [18] (an in-house Apache Hadoop version). The Genome-Analysis-ToolKit (GATK) 2.3.9-lite [19] (an in-house Apache Hadoop version) was used for variant calling from BAM files with HaplotypeCaller v2.3.9-lite .

Saturation of NGS data

Picard (version 2.18.9) was used to down-sample BAM files from 10× to the maximum depth in a 10×-step for each NGS data. Next, MegaBOLT (version 1.15) was used for variant calling from down-sampled BAM files. SNPs were hard-filtered using “QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0” and Indels were hard-filtered using “QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0”.

Uncovered region of NGS data

For each NGS data, any block with approximate read depth (DP) ≥ 5 were extracted from gVCF as a covered region. The uncovered regions of each NGS data were built by subtracting the covered regions from the human genome by BEDtools (v2.16.2). Meanwhile, the common uncovered regions of NGS data were built by subtracting the union of covered regions in all 15 NGS data from the human genome.

Linked reads Mapping

The output files (FASTQ) of the linked-read sequencing methods stLFR and 10X Genomics Chromium are similar, enabling the use of the 10X Genomics Long Ranger software after converting stLFR barcodes to a Chromium compatible format. We used SOAPnuke 1.5.6 to filter out low quality and adapter reads. Clean reads were mapped and phased using the Long Ranger 2.1.2 wgs model. Briefly, de-multiplexed FASTQ files from were de-duplicated and filtered and phased SNPs, indels were called. SNP and indel information were parsed from the final VCF file using GATK SelectVariants.

Pacbio data process

PacBio single molecule real-time circular consensus sequencing (SMRT CCS) have low base error rates, providing both highly-accurate variant calls and long-range information needed to generate haplotypes. We used the pbmm2 (version 1.0.0) alignment tool to map reads to the hs37d5 human reference genome, with the parameter --preset CCS --sample HJ –sort. GATK HaplotypeCaller was used to call SNVs and small indels. Different values of the HaplotypeCaller parameter --pcr-indel-model and VariantFiltration parameter --filter-expression were considered, setting the minimum mapping quality to 60 and using allele-specific annotations (--annotationgroup AS_StandardAnnotation) and --pcr_indel_model AGGRESSIVE. SNVs and short indels were filtered using GATK VariantFiltration with --filter_expression of AS_QD < 2.0 . Longer read lengths improve the ability to phase variants, as tools like WhatsHap demonstrate for PacBio reads [17].

Data resource access

The sequence data from this article can be found in the CNSA databases under the following accession numbers: CNP0000091.

Acknowledgments

This work was supported by Grants from National Key Research and Development Program of China（2017YFC0906501, 2019B020226001）. We thank Dr. Xin Liu for his guidance on the project and technical assistance.

Competing Interests

Competing interest statement: The author denies that he has any intention to obtain any financial interests.

Figure 1. Overview of variation calling pipeline. The major steps included data filtering, alignment, variation calling, and integrated analysis.

Figure 2. Saturation analysis. The relationship between SNPs(A)/indels(B) and depth, with the X axis for sequencing depth and the Y axis for the number of SNPs/indels detected.

Figure 3. Uncovered region by NGS in each sequencing platform.

Figure 4. Density maps of sequencing platforms SNP and indel variations. From inside to outside circles are BGISEQ-NGS, Illumina-NGS,stLFR and Pacbio CCS respectively, Window =1000000bp,Inside and outside are SNP and indel.

Figure 5. Consistency analysis: BGI regular NGS platforms, Illumina regular NGS platforms, two linked-read libraries, and PacBio CCS mode SNP(A) and indel(B) consistency analysis.

Figure 6. Depth and coverage of NBPF4 gene in HHRs.

Table 1. Haplotype phasing small variants.

References

1. Ashley EA. Towards precision medicine. Nature Reviews Genetics. 2016;17(9):507.

2. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Human molecular genetics. 2010;19(R2):R131-R6.

3. Ramensky V, Bork P, Sunyaev S. Human non‐synonymous SNPs: server and survey. Nucleic acids research. 2002;30(17):3894-900.

4. Telenti A, Pierce LC, Biggs WH, Di Iulio J, Wong EH, Fabani MM, et al. Deep sequencing of 10,000 human genomes. Proceedings of the National Academy of Sciences. 2016;113(42):11901-6.

5. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, et al. The diploid genome sequence of an Asian individual. Nature. 2008;456(7218):60.

6. Cho YS, Kim H, Kim H-M, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nature communications. 2016;7:13637.

7. Azim MK, Yang C, Yan Z, Choudhary MI, Khan A, Sun X, et al. Complete genome sequencing and variant analysis of a Pakistani individual. Journal of human genetics. 2013;58(9):622.

8. Wei X, Ju X, Yi X, Zhu Q, Qu N, Liu T, et al. Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing. PloS one. 2011;6(12):e29500.

9. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Human molecular genetics. 2018;27(R2):R234-R41.

10. Wang O, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome research. 2019;29(5):798-808.

11. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature biotechnology. 2016;34(3):303.

12. Larsen PA, Heilman AM, Yoder AD. The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. BMC genomics. 2014;15(1):720.

13. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. Genomics, proteomics & bioinformatics. 2016;14(5):265-79.

14. Mandelker D, Schmidt RJ, Ankala A, Gibson KM, Bowser M, Sharma H, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genetics in Medicine. 2016;18(12):1282.

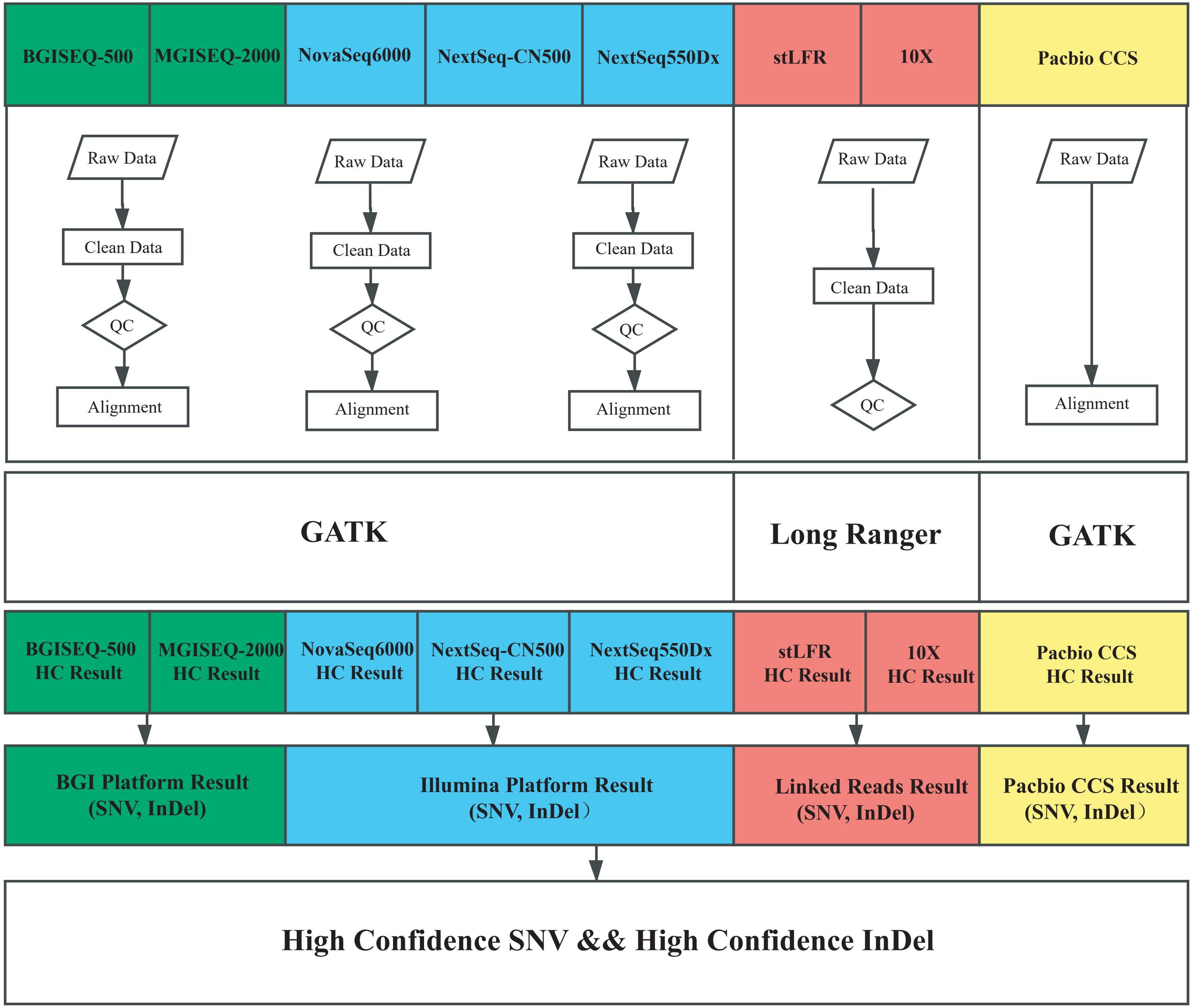
15. Menelaou A, Marchini J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. Bioinformatics. 2012;29(1):84-91.

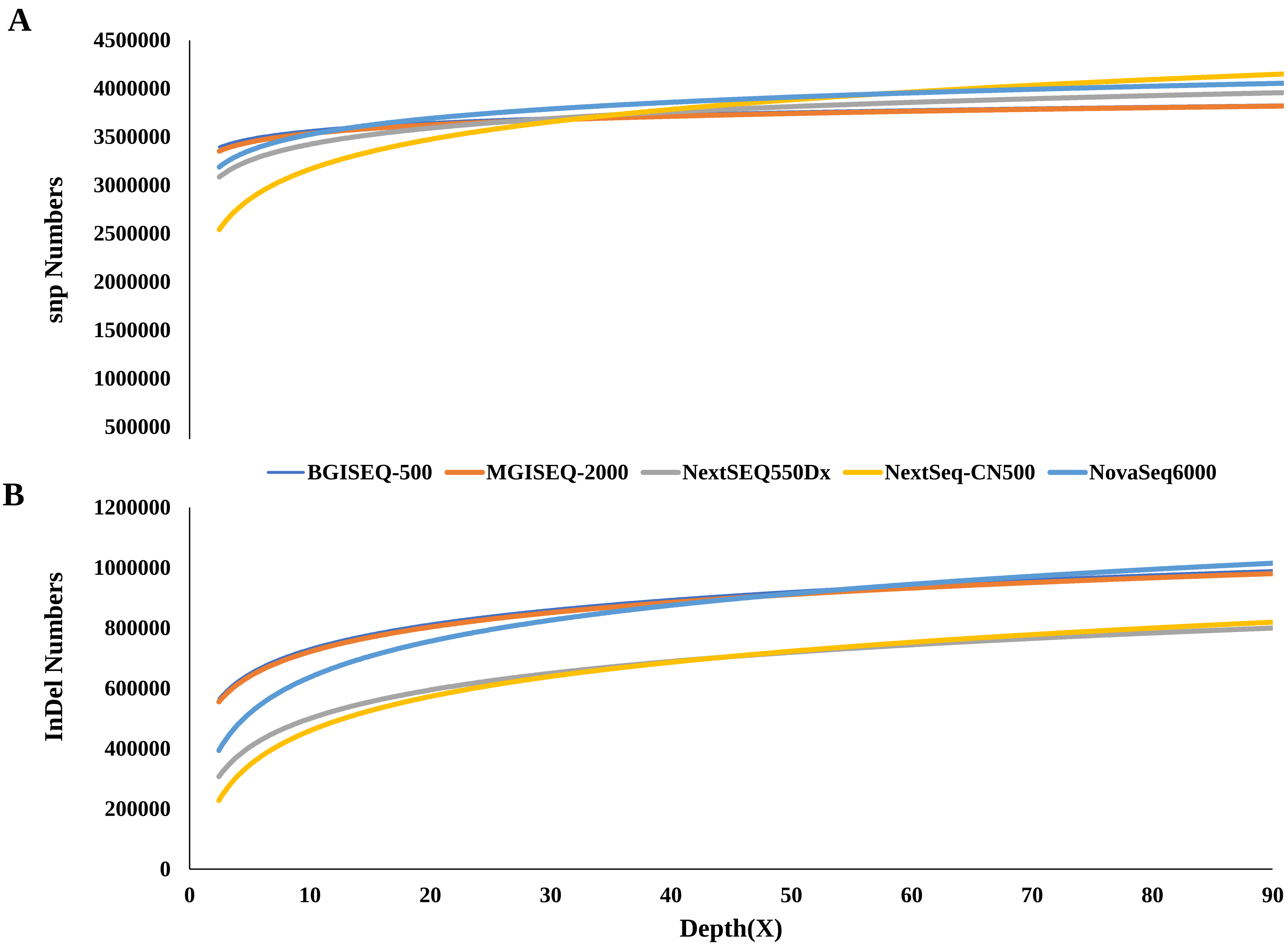
16. Association GAotWM. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. The Journal of the American College of Dentists. 2014;81(3):14.

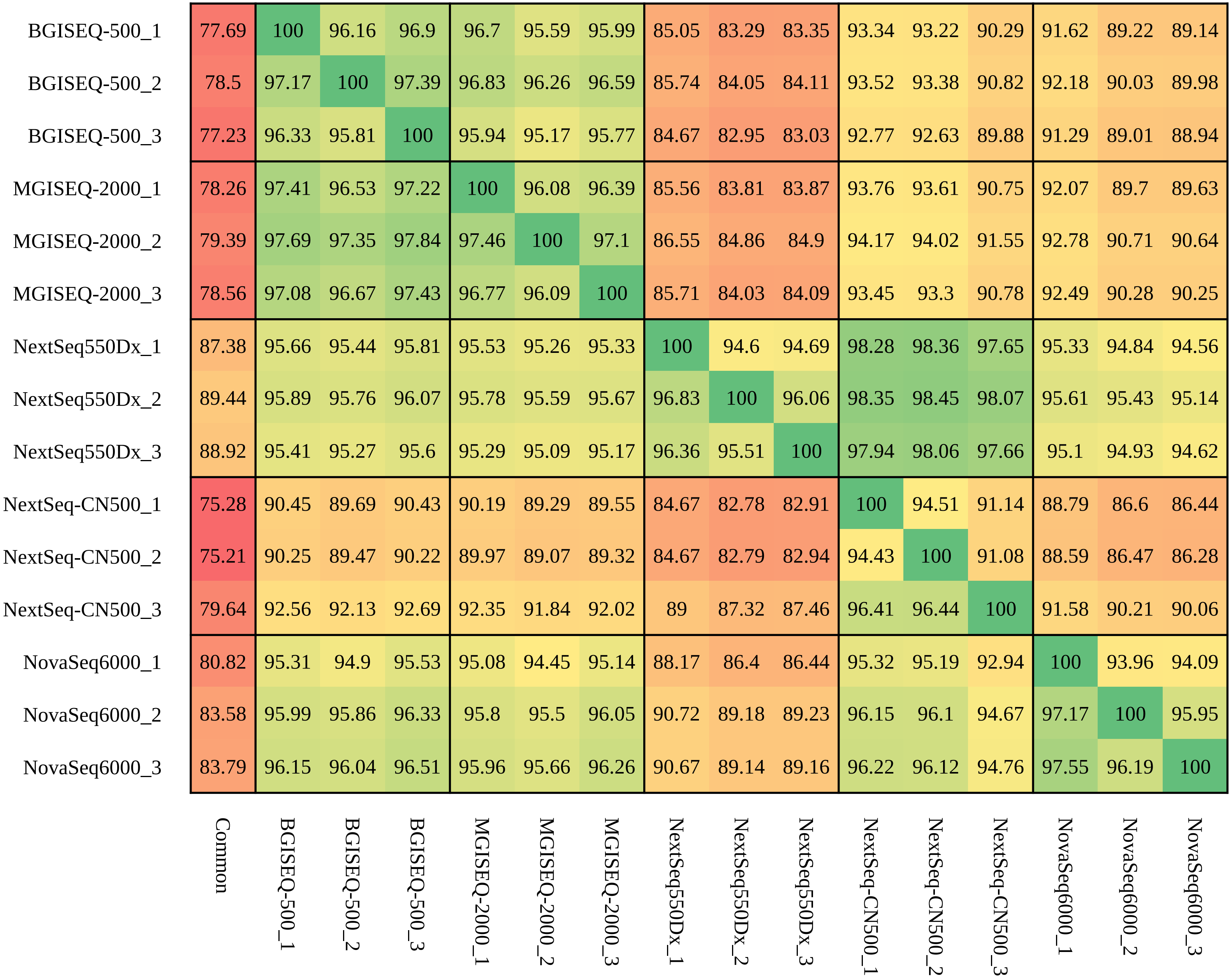
17. Westbrook CJ, Karl JA, Wiseman RW, Mate S, Koroleva G, Garcia K, et al. No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing. Human Immunology. 2015;76(12):891-6.

18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

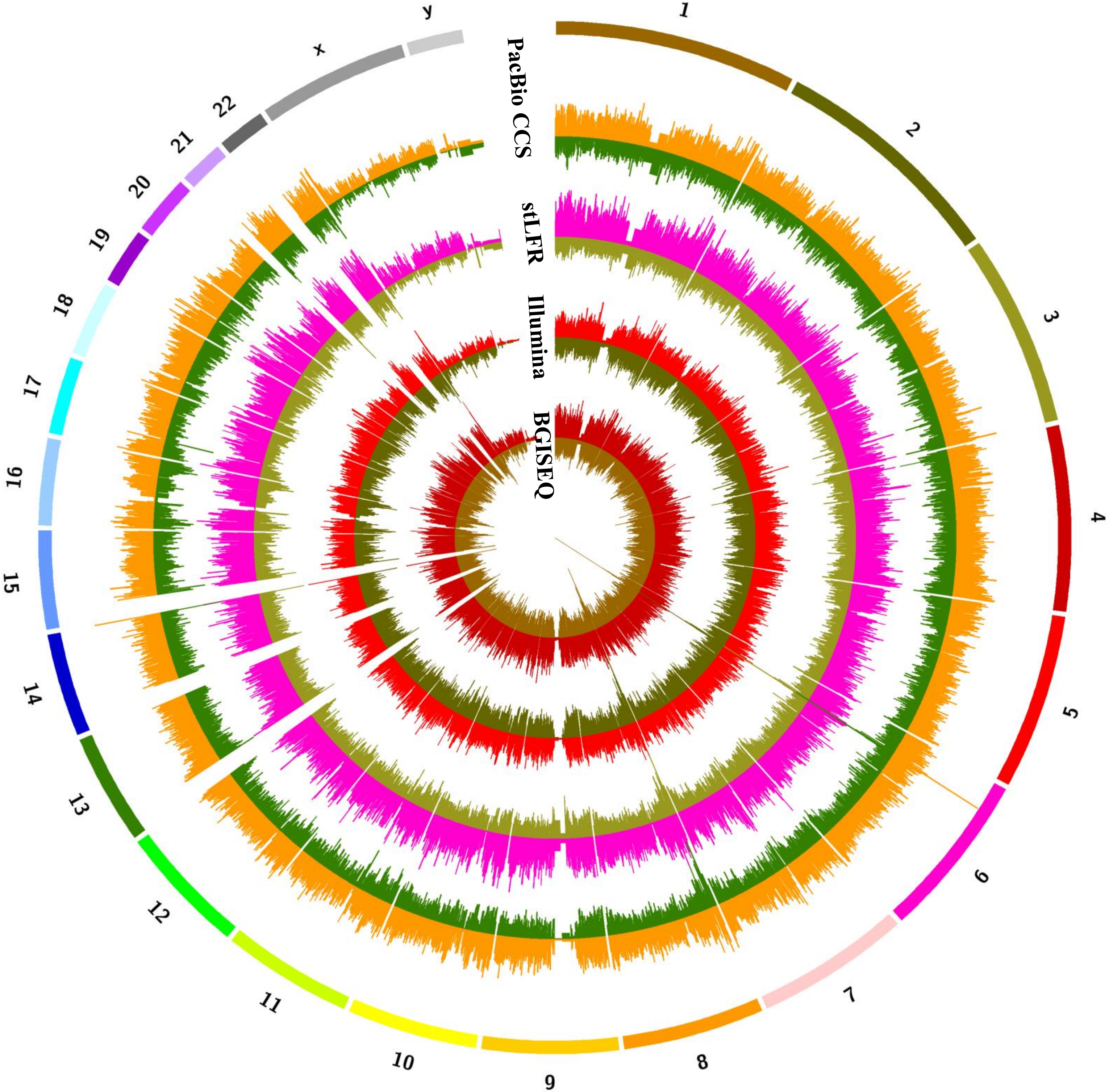
19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.

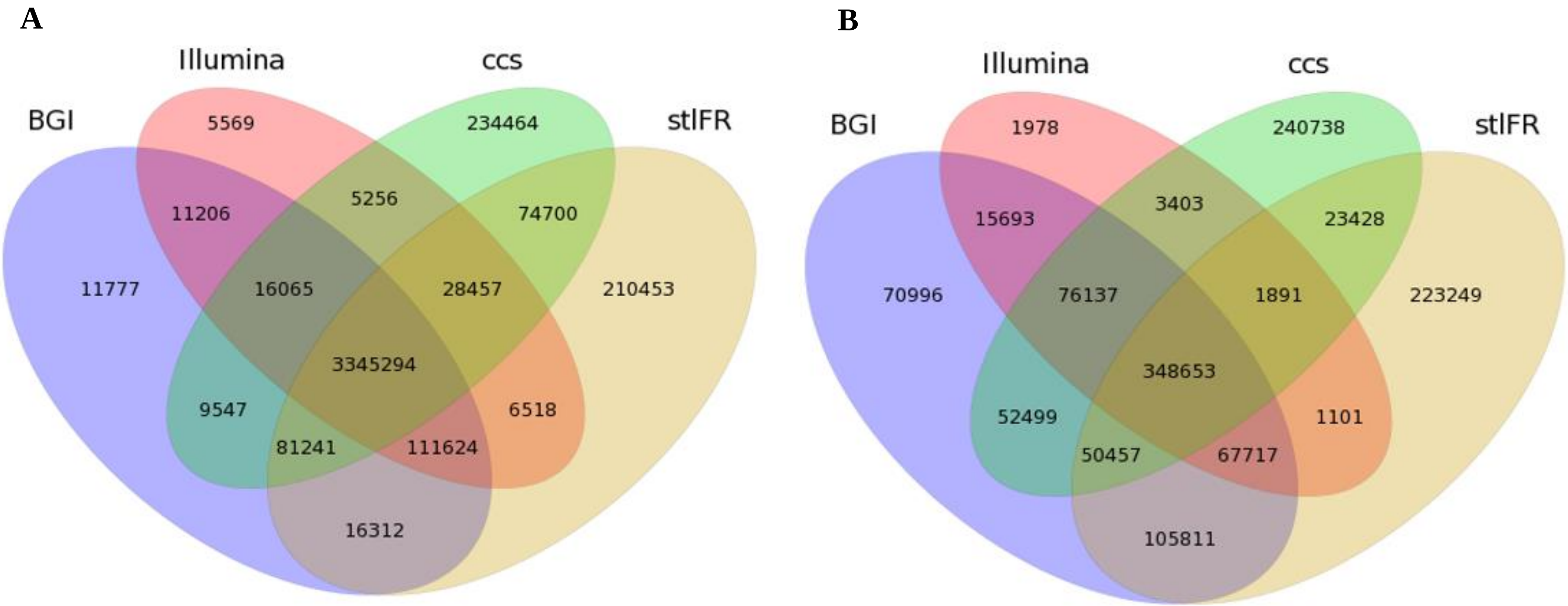

BGISEQ-500
MGISEQ-2000
NovaSeq6000
NextSeq-CN500
NextSeq550Dx
stLFR
10X
Pacbio CCS
Raw Data
Clean Data
QC
Alignment
Raw Data
Clean Data
QC
Alignment
Raw Data
Clean Data
QC
Alignment
Raw Data
Clean Data
QC
Raw Data
Alignment
GATK
Long Ranger
GATK
BGISEQ-500 HC Result
MGISEQ-2000 HC Result
NovaSeq6000 HC Result
NextSeq-CN500 HC Result
NextSeq550Dx HC Result
stLFR HC Result
10X HC Result
Pacbio CCS HC Result
BGI Platform Result (SNV, InDel)
Illumina Platform Result (SNV, InDel）
Linked Reads Result (SNV, InDel)
Pacbio CCS Result (SNV, InDel）
High Confidence SNV && High Confidence InDel


A
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
snp Numbers
BGISEQ-500
MGISEQ-2000
NextSEQ550Dx
NextSeq-CN500
NovaSeq6000
B
1200000
1000000
800000
600000
400000
200000
0
InDel Numbers
0
10
20
30
40
50
60
70
80
90
Depth(X)

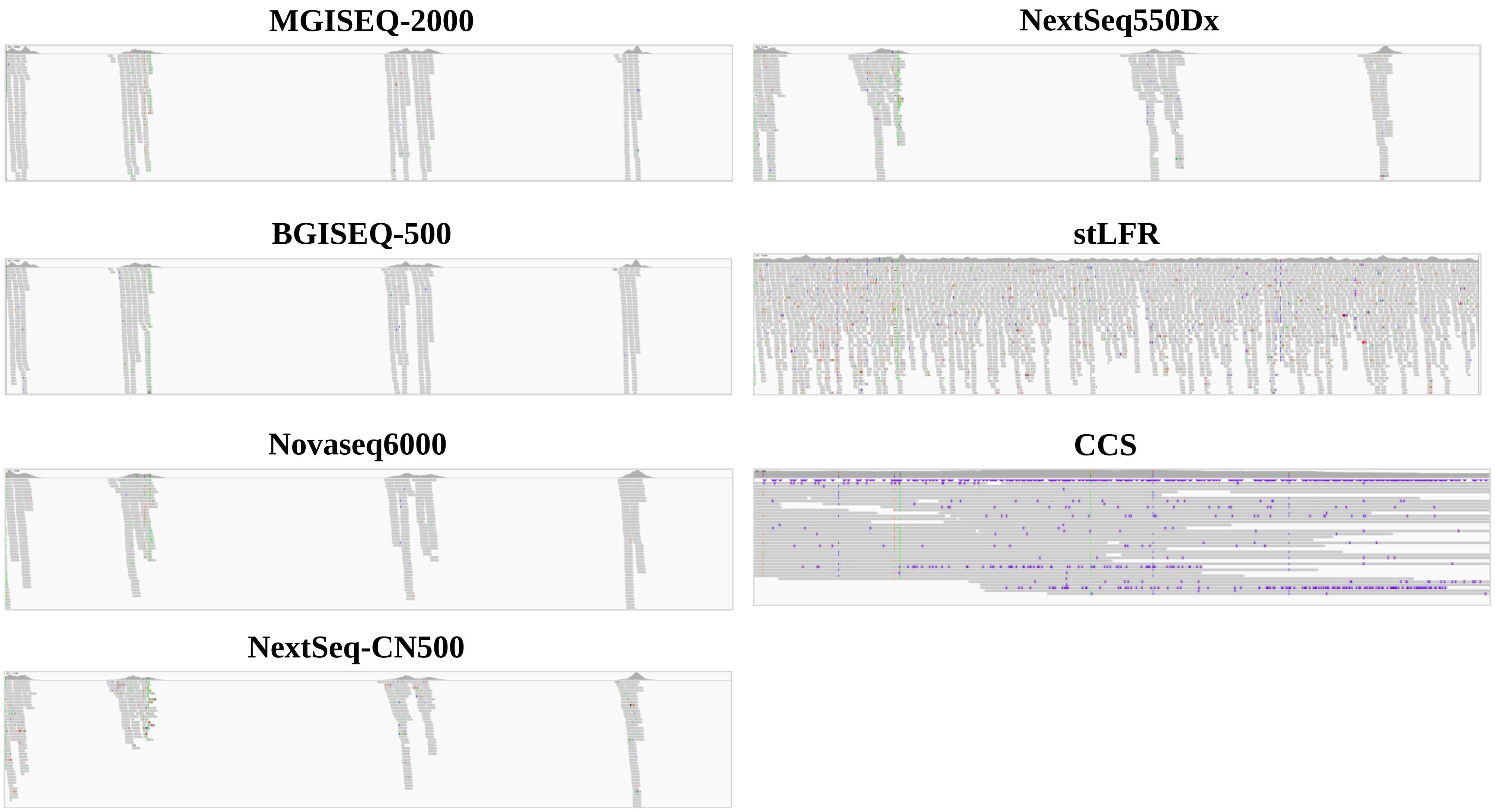


	Common	BGISEQ-500_1	BGISEQ-500_2	BGISEQ-500_3	MGISEQ-2000_1	MGISEQ-2000_2	MGISEQ-2000_3	NextSeq550Dx_1	NextSeq550Dx_2	NextSeq550Dx_3	NextSeq-CN500_1	NextSeq-CN500_2	NextSeq-CN500_3	NovaSeq6000_1	NovaSeq6000_2	NovaSeq6000_3
BGISEQ-500_1	77.69	100	96.16	96.9	96.7	95.59	95.99	85.05	83.29	83.35	93.34	93.22	90.29	91.62	89.22	89.14
BGISEQ-500_2	78.5	97.17	100	97.39	96.83	96.26	96.59	85.74	84.05	84.11	93.52	93.38	90.82	92.18	90.03	89.98
BGISEQ-500_3	77.23	96.33	95.81	100	95.94	95.17	95.77	84.67	82.95	83.03	92.77	92.63	89.88	91.29	89.01	88.94
MGISEQ-2000_1	78.26	97.41	96.53	97.22	100	96.08	96.39	85.56	83.81	83.87	93.76	93.61	90.75	92.07	89.7	89.63
MGISEQ-2000_2	79.39	97.69	97.35	97.84	97.46	100	97.1	86.55	84.86	84.9	94.17	94.02	91.55	92.78	90.71	90.64
MGISEQ-2000_3	78.56	97.08	96.67	97.43	96.77	96.09	100	85.71	84.03	84.09	93.45	93.3	90.78	92.49	90.28	90.25
NextSeq550Dx_1	87.38	95.66	95.44	95.81	95.53	95.26	95.33	100	94.6	94.69	98.28	98.36	97.65	95.33	94.84	94.56
NextSeq550Dx_2	89.44	95.89	95.76	96.07	95.78	95.59	95.67	96.83	100	96.06	98.35	98.45	98.07	95.61	95.43	95.14
NextSeq550Dx_3	88.92	95.41	95.27	95.6	95.29	95.09	95.17	96.36	95.51	100	97.94	98.06	97.66	95.1	94.93	94.62
NextSeq-CN500_1	75.28	90.45	89.69	90.43	90.19	89.29	89.55	84.67	82.78	82.91	100	94.51	91.14	88.79	86.6	86.44
NextSeq-CN500_2	75.21	90.25	89.47	90.22	89.97	89.07	89.32	84.67	82.79	82.94	94.43	100	91.08	88.59	86.47	86.28
NextSeq-CN500_3	79.64	92.56	92.13	92.69	92.35	91.84	92.02	89	87.32	87.46	96.41	96.44	100	91.58	90.21	90.06
NovaSeq6000_1	80.82	95.31	94.9	95.53	95.08	94.45	95.14	88.17	86.4	86.44	95.32	95.19	92.94	100	93.96	94.09
NovaSeq6000_2	83.58	95.99	95.86	96.33	95.8	95.5	96.05	90.72	89.18	89.23	96.15	96.1	94.67	97.17	100	95.95
NovaSeq6000_3	83.79	96.15	96.04	96.51	95.96	95.66	96.26	90.67	89.14	89.16	96.22	96.12	94.76	97.55	96.19	100


PacBio CCS
stLFR
Illumina
BGISEQ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
x
y


A
Illumina
ccs
BGI
stIFR
5569
234464
5256
11206
74700
11777
16065
28457
210453
3345294
9547
6518
81241
111624
16312
B
Illumina
ccs
BGI
stIFR
1978
240738
3403
15693
23428
70996
76137
1891
223249
348653
52499
1101
50457
67717
105811

NBPF4 chr1:108,918,492-108,931,992



MGISEQ-2000

NextSeq550Dx

BGISEQ-500

stLFR

Novaseq6000

CCS

NextSeq-CN500