# MegaBOLT Supplementary Information

## 1. Data used in brochure:

### a) DNBSEQ PCR PE100 94Gbp:

https://ftp.cngb.org/pub/CNSA/data1/CNP0000059/CNS0001678/CNX0001989/CNR0002144/MGISEQ-2000.WGS.PE100_1.fq.gz
https://ftp.cngb.org/pub/CNSA/data1/CNP0000059/CNS0001678/CNX0001989/CNR0002144/MGISEQ-2000.WGS.PE100_2.fq.gz

### b) DNBSEQ PCR PE150 138Gbp:

https://ftp.cngb.org/pub/CNSA/data1/CNP0000059/CNS0001678/CNX0023580/CNR0028192/MGISEQ-2000.WGS.PCR-1.read1.fq.gz
https://ftp.cngb.org/pub/CNSA/data1/CNP0000059/CNS0001678/CNX0023580/CNR0028192/MGISEQ-2000.WGS.PCR-1.read2.fq.gz

### c) DNBSEQ PCR-Free PE150 126Gbp:

https://ftp.cngb.org/pub/CNSA/data1/CNP0000059/CNS0001678/CNX0023582/CNR0028194/MGISEQ-2000.WGS.PCR-Free-1.read1.fq.gz
https://ftp.cngb.org/pub/CNSA/data1/CNP0000059/CNS0001678/CNX0023582/CNR0028194/MGISEQ-2000.WGS.PCR-Free-1.read2.fq.gz

### d) DNBSEQ MGIEasy Exome V5 PE150 11Gbp（extract 11Gbp from 22Gbp after download）:

https://ftp.cngb.org/pub/CNSA/data2/CNP0000662/CNS0110241/CNX0094950/CNR0117174/MGIEasy_V5_Universal_Library_3.1.fq.gz
https://ftp.cngb.org/pub/CNSA/data2/CNP0000662/CNS0110241/CNX0094950/CNR0117174/MGIEasy_V5_Universal_Library_3.2.fq.gz

## 2. References used in MegaBOLT (all data are downloaded from GATK resource bundle and NCBI):

### a) hg19.fa, dbsnp_151 and knownSites:

ftp://ftp.broadinstitute.org/bundle/hg19/
ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/GATK/All_20180423.vcf.gz

### b) hs37d5.fa, dbsnp_138 and knownSites:

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/b37/

**c) hg38.fa, dbsnp_151 and knownSites:**

ftp://ftp.broadinstitute.org/bundle/hg38/
ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/VCF/GATK/All_20180418.vcf.gz

# 3. Scripts used in brochure (Please download the package "MegaBOLT Scripts for Brochure" from the website):

## 3.1 Pre-process WES data and simulate Somatic data:

**a) Extract WES data:**

data/Extract_WES.sh

**b) Simulate Somatic WGS tumor data (Generated by bamsurgeon):**

Generate_Somatic_WGS_data/run.sh

**c) Simulate Somatic WES tumor data (Generated by bamsurgeon):**

Generate_Somatic_WES_data/run.sh

## 3.2 Performance test on single task and multiple tasks:

**a) Performance on Germline WGS**

Data: DNBSEQ PCR PE100 94Gbp
Reference: hg19
Scripts:
Germline_WGS/Germline_WGS_MegaBOLT/run.sh
Germline_WGS/Germline_WGS_MegaBOLT_FULL/run.sh

**b) Performance on Germline WES**

Data: DNBSEQ MGIEasy Exome V5 PE150 11Gbp
Reference: hg19
Scripts:
Germline_WES/Germline_WES_MegaBOLT/run.sh
Germline_WES/Germline_WES_MegaBOLT_FULL/run.sh

**c) Performance on Somatic WGS**

Data:
Normal: DNBSEQ PCR-Free PE150 126Gbp
Tumor: Generate from "DNBSEQ PCR-Free PE150 126Gbp" by bamsurgeon
Reference: hg19
Scripts:
Somatic_WGS/Somatic_WGS_MegaBOLT/run.sh

**d) Performance on Somatic WES**

Data:
Normal: DNBSEQ MGIEasy Exome V4 PE100 26Gbp
Tumor: Generate from “DNBSEQ MGIEasy Exome V4 PE100 26Gbp” by bamsurgeon
Reference: hg19
Scripts:
Somatic_WES/Somatic_WES_MegaBOLT/run.sh

### e) Performance on Multiple tasks WGS

Data: DNBSEQ PCR PE100 94Gbp
Reference: hg19
Scripts:
Scheduler_WGS/Scheduler_WGS_Single/run.sh
Scheduler_WGS/Scheduler_WGS_Multiple/run.sh
Scheduler_WGS/Scheduler_WGS_Single_FULL/run.sh
Scheduler_WGS/Scheduler_WGS_Multiple_FULL/run.sh

### f) Performance on Multiple tasks WES

Data: DNBSEQ MGIEasy Exome V5 PE150 11Gbp
Reference: hg19
Scripts:
Scheduler_WES/WES_Single/run.sh
Scheduler_WES/WES_Multiple/run.sh
Scheduler_WES/WES_Single_FULL/run.sh
Scheduler_WES/WES_Multiple_FULL/run.sh

## 3.3 Accuracy test on WGS and WES:

### a) Accuracy test on WGS data

Data:
DNBSEQ PCR PE100 94Gbp
DNBSEQ PCR PE150 138Gbp
DNBSEQ PCR-Free PE150 126Gbp
Reference: hs37d5
Scripts for DNBSEQ PCR PE100 94Gbp:
Accuracy_WGS/DNBSEQ_PCR_PE100_94Gbp/MegaBOLT/run.sh
Accuracy_WGS/DNBSEQ_PCR_PE100_94Gbp/MegaBOLT-DV/run.sh
Scripts for DNBSEQ PCR PE150 138Gbp:
Accuracy_WGS/DNBSEQ_PCR_PE150_138Gbp/MegaBOLT/run.sh
Accuracy_WGS/DNBSEQ_PCR_PE150_138Gbp/MegaBOLT-DV/run.sh
Scripts for DNBSEQ PCR-Free PE150 126Gbp:
Accuracy_WGS/DNBSEQ_PCR-Free_PE150_126Gbp/MegaBOLT/run.sh
Accuracy_WGS/DNBSEQ_PCR-Free_PE150_126Gbp/MegaBOLT-DV/run.sh

### b) Accuracy test on WES data

Data: DNBSEQ MGIEasy Exome V5 PE150 11Gbp
Reference: hs37d5
Scripts:
Accuracy_WES/MegaBOLT/run.sh
Accuracy_WES/MegaBOLT-DV/run.sh