GSA

GSA Submission

***Alias**

Submission name of the GSA. This field is used when the record does not yet have an accession and needs to be referenced by other objects.

***Data Released**

Select Release on specified date or give release data in correct format (yyyy-MM-dd).

Experiments

Meta Information

***Platfom**

The sequencing platform and instrument model

Instrument Model
454 GS 20
454 GS FLX
454 GS FLX Titanium
454 GS FLX+
454 GS Junior
AB 310 Genetic Analyzer
AB 3130 Genetic Analyzer
AB 3130xL Genetic Analyzer
AB 3500 Genetic Analyzer
AB 3500xL Genetic Analyzer
AB 3730 Genetic Analyzer
AB 3730xL Genetic Analyzer
AB 5500 Genetic Analyzer
AB 5500xl Genetic Analyzer
AB SOLiD 3 Plus System

AB SOLiD 4 System
AB SOLiD 4hq System
AB SOLiD PI System
AB SOLiD System 1.0
AB SOLiD System 2.0
AB SOLiD System 3.0
Complete Genomics
Helicos HeliScope
Illumina Genome Analyzer
Illumina Genome Analyzer II
Illumina Genome Analyzer IIx
Illumina HiScanSQ
Illumina HiSeq 1000
Illumina HiSeq 1500
Illumina HiSeq 2000
Illumina HiSeq 2500
Illumina MiSeq
Ion Torrent PGM
Ion Torrent Proton
PacBio RS
Illumina Nextseq 500
Illumina HiSeq 3000
Illumina HiSeq 4000
Illumina HiSeq X-10
Agilent Infinity 1290 UHPLC - 6550 QTOF MS
454 GS 20
454 GS FLX
454 GS FLX Titanium
454 GS FLX+

454 GS Junior
AB 310 Genetic Analyzer
AB 3130 Genetic Analyzer
AB 3130xL Genetic Analyzer
AB 3500 Genetic Analyzer
AB 3500xL Genetic Analyzer

*Alias

Submission name of the experiment. This field is used when the record does not yet have an accession and needs to be referenced by other objects.

*Title

Short text that can be used to call out experiment records in searches or in displays.

*Project accession

Link data to BioProject that describes the research.

* Sample accession

Enter a BioSample or GSA Sample Accession. BioSample accessions have ‘SAMN’ prefix. A BioSample describes the biological source material for your sequence library preparation.

* Library Construction/Experiment design

Enter the details about your experimental design and molecular strategies including hybrid selection and affinity capture reagents; any detail that distinguishes your experiment from other similar experiments.

Library

The library descriptor specifies the origin of the material being sequenced and any treatments that the material might have undergone that affect the sequencing result.

Library name

The submitter's name for this library.

*Strategy

Sequencing technique intended for this library.

Strategy	***Sequencing strategy used in the experiment***
WGA	Random sequencing of the whole genome following non-PCR

	amplification
WGS	Random sequencing of the whole genome
WXS	Random sequencing of exonic regions selected from the genome
RNA-Seq	Random sequencing of whole transcriptome
miRNA-Seq	Random sequencing of small miRNAs
Tn-Seq	Sequencing from transposon insertion sites
WCS	Random sequencing of a whole chromosome or other replicon isolated from a genome
CLONE	Genomic clone based (hierarchical) sequencing
POOLCLONE	Shotgun of pooled clones (usually BACs and Fosmids)
AMPLICON	Sequencing of overlapping or distinct PCR or RT-PCR products
CLONEEND	Clone end (5', 3', or both) sequencing
FINISHING	Sequencing intended to finish (close) gaps in existing coverage
ChIP-Seq	Direct sequencing of chromatin immunoprecipitates
MNase-Seq	Direct sequencing following MNase digestion
DNase-Hypersensitivity	Sequencing of hypersensitive sites, or segments of open chromatin that are more readily cleaved by DNaseI
Bisulfite-Seq	Sequencing following treatment of DNA with bisulfite to convert cytosine residues to uracil depending on methylation status
EST	Single pass sequencing of cDNA templates
FL-cDNA	Full-length sequencing of cDNA templates
CTS	Concatenated Tag Sequencing
MRE-Seq	Methylation-Sensitive Restriction Enzyme Sequencing strategy
MeDIP-Seq	Methylated DNA Immunoprecipitation Sequencing strategy
MBD-Seq	Direct sequencing of methylated fractions sequencing strategy
OTHER	Library strategy not listed (please include additional info in the “design description”)

* Source

The library source specifies the type of source material that is being sequenced.

Source	***Type of genetic source material sequenced***
GENOMIC	Genomic DNA (includes PCR products from genomic DNA)
TRANSCRIPTOMIC	Transcription products or non-genomic DNA (EST, cDNA, RT-PCR, screened libraries)
METATRANSCRIPTOMIC	Transcription products from community targets
METAGENOMIC	Mixed material from metagenome
SYNTHETIC	Synthetic DNA
VIRAL RNA	Viral RNA
OTHER	Other, unspecified, or unknown library source material (please include additional info in the “design description”)

*Selection

Whether any method was used to select and/or enrich the material being sequenced.

Selection	***Method of selection or enrichment used in the Experiment***
unspecified	Library enrichment, screening, or selection is not specified (please include additional info in the “design description”)
RANDOM	Random selection by shearing or other method
PCR	Source material was selected by designed primers
RANDOM PCR	Source material was selected by randomly generated primers
RT-PCR	Source material was selected by reverse transcription PCR
HMPR	Hypo-methylated partial restriction digest
MF	Methyl Filtrated
CF-S	Cot-filtered single/low-copy genomic DNA
CF-M	Cot-filtered moderately repetitive genomic DNA
CF-H	Cot-filtered highly repetitive genomic DNA
CF-T	Cot-filtered theoretical single-copy genomic DNA
MDA	Multiple displacement amplification
MSLL	Methylation Spanning Linking Library
cDNA	complementary DNA

ChIP	Chromatin immunoprecipitation
MNase	Micrococcal Nuclease (MNase) digestion
DNAse	Deoxyribonuclease (MNase) digestion
Hybrid Selection	Selection by hybridization in array or solution
Reduced Representation	Reproducible genomic subsets, often generated by restriction fragment size selection, containing a manageable number of loci to facilitate re-sampling
Restriction Digest	DNA fractionation using restriction enzymes
5-methylcytidine antibody	Selection of methylated DNA fragments using an antibody raised against 5-methylcytosine or 5-methylcytidine (m5C)
MBD2 protein methyl-CpG binding domain	Enrichment by methyl-CpG binding domain
CAGE	Cap-analysis gene expression
RACE	Rapid Amplification of cDNA Ends
size fractionation	Physical selection of size appropriate targets
Padlock probes capture method	Circularized oligonucleotide probes
Poly-A	polyA enriched RNA-seq
other	Other library enrichment, screening, or selection process (please include additional info in the “design description”)

***Layout**

Library Layout specifies whether to expect single, Pair-end, or other configuration of reads. In the case of paired reads, information about the relative distance and orientation is specified.

- **Fragment**
- **Paired read**

***Insert size (bp)**

Fragment size for Paired reads.

Nominal size (bp)

Size of the insert for Paired reads.

Nominal standard deviation (bp)

Standard deviation of insert size (typically ~10% of Nominal Size)

Run

General info

* **Alias**

Submitter assigned name or id for the GSA submission object.

* **Run data file type Run**

The information about supported formats of the submitted sequence data. We recommend that read data is either submitted in Fastq or BAM format. Submitted data files must only contain reads from a single sample.

Format	**File suffix**	**Made available as standard Fastq**	**Notes**
Fastq format	.fastq.gz .fastq.bz2 .fq.gz .fq.bz2	Yes	
BAM format	.bam	Yes	
VCF Format	.vcf	No	
SFF Format	.sff	Yes	Spot descriptor is required.

Data Blocks

- **Fastq format**

Fastq format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

* **File Name**

We only accept GZIP and BZIP2 compression formats. Especially we don't accept 7-ZIP or TAR compressed files.

- Single reads must be submitted using a single Fastq file and can be submitted the suffix in ‘_1’, ‘_2’, etc.
- Paired reads must split and submitted using two Fastq files. The read names must

have a suffix identifying the first and second read from the pair.
For example: liver_Tumor1_male_1F.fastq.gz and liver_Tumor1_male_1R.fastq.gz, then followed by read ‘2F’, then ‘2R’, etc.

* **MD5 for file**

MD5 checksums are a 32-character alphanumeric string. For Mac and Linux system users, the native command line tools "md5sum"(Linux) and "md5"(Mac OX) can be used to generate MD5 checksums. Windows users must need to download a third-party utility.

- **BAM format**

The BAM format is an efficient method for storing and sharing data from modern, highly parallel sequencers. While primarily used for storing alignment information, BAMs can (and frequently do) store unaligned reads as well.

* **Reference Assembly Name**

* **Assembly Name or Accession**

The Reference’s assembly name or assembly accession number

* **Web URL of the Reference Assembly**

The URL of the Reference Assembly

* **File Name for bam bam**

Submitted BAM files must be readable with SAMtools. BAM file names are required to end up with the .bam suffix (e.g. 'a.bam').

* **MD5 for bam file bam**

MD5 for bam file bam

* **Local Assembly file**

***select one reference file you have uploaded/submit new reference file**

* **Reference file name**

The Reference’s file name

* **MD5 for reference file**

MD5 for reference file

* **File Name for bam**

Submitted BAM files must be readable with SAMtools. BAM file names are required to end up with the .bam suffix (e.g. 'a.bam').

* **MD5 for bam file**

MD5 for bam file bam

- **VCF format**

The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing gene sequence variations. By using the variant call format only the variations need to be stored along with a reference genome.

* **Reference Assembly Name**

*** Assembly Name or Accession**

The Reference's assembly name or assembly accession number

*** Web URL of the Reference Assembly**

The URL of the Reference Assembly

*** File Name for VCF**

VCF file names are required to end up with the .vcf suffix (e.g. 'a.vcf').

*** MD5 for VCF file VCF**

MD5 for VCF file

*** Local Assembly file**

***select one reference file you have uploaded/submit new reference file**

*** Reference file name**

The Reference's file name

*** MD5 for reference file**

MD5 for reference file

*** File Name for VCF**

VCF file names are required to end up with the .vcf suffix (e.g. 'a.vcf').

*** MD5 for VCF file**

MD5 for VCF file

- **SFF format**

Standard flowgram format (SFF) is a binary file format used to encode results of pyrosequencing from the 454 Life Sciences platform for high-throughput sequencing. SFF files can be viewed, edited and converted with DNA Baser SFF Workbench (graphic tool), or converted to FASTQ format with sff2fastq or seq_crumbs.

* **File Name**

SFF file names are required to end up with the .sff suffix (e.g. 'a.sff').

* **MD5 for file**

MD5 for SFF file bam

NOTE:

Transmitting your data files to the Gsub FTP site

Address: ftp://submit.big.ac.cn
User: Same as your Gsub Username
Password: Same as your Gsub Password