GSA Help Document

Version 2.1, 2018

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

*** Platform**

The sequencing platform and instrument model

Platform	***Instrument Model***
LS454	454 GS
	454 GS 20
	454 GS FLX
	454 GS FLX Titanium
	454 GS FLX+
	454 GS Junior
Capillary Technologies	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
ABI Solid	AB 5500 Genetic Analyzer
	AB 5500xl Genetic Analyzer
	AB 5500x-Wl 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
BGISeq	BGISEQ-100
	BGISEQ-1000
	BGISEQ-500

CapitalBio Company	BioelectronSeq 4000
Bionano Genomics	BioNano IRYS
	BioNano SAPHYR
Complete Genomics	Complete Genomics
DAAN GENE	DA8600
Helicos BioSciences Corporation	Helicos HeliScope
HYK Genetic	HYK-PSTAR-IIA
Illumina	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 HiSeq 3000
	Illumina HiSeq 4000
	Illumina HiSeq X Ten
	Illumina MiSeq
	Illumina MiniSeq
	Illumina Nextseq 500
	Illumina Nextseq 550
	Illumina iSeq 100
	Illumina NovaSeq 5000
	Illumina NovaSeq 6000
IonTorrent	Ion Torrent PGM
	Ion Torrent Proton
Oxford Nanopore	MinION
	GridION
Berry Genomics	NextSeq CN500
PacBio SMRT	PacBio RS
	PacBio RS II
	PacBio Sequel

***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. This element is technically optional but should be used for all new records.

***Project accession**

Link data to BioProject that describes the research.

*** Sample accession**

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

*** Library Construction/Experiment design**

Choose 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. This specification is needed even if the platform does not require a library construction step per se.

Library name

The submitter's name for this library.

*** Strategy**

Sequencing technique intended for this library.

Library Strategy	***Description***
WGS	Whole genome shotgun.
WGA	Whole genome amplification.
WES	Whole exome sequencing is a genomic technique for sequencing all of the protein-coding genes in a genome (known as the exome).
WXS	Random sequencing of exonic regions selected from the genome.
RNA-Seq	Random sequencing of whole transcriptome.
miRNA-Seq	Micro RNA and other small non-coding RNA sequencing.
WCS	Whole chromosome (or other replicon) shotgun.
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.
Tn-Seq	Gene fitness determination through transposon seeding.
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.
Synthetic-Long-Read	binning and barcoding of large DNA fragments to facilitate assembly of the fragment
ATAC-seq	Assay for Transposase-Accessible Chromatin (ATAC) strategy is used to study genome-wide chromatin accessibility. alternative method to DNase-seq that uses an engineered Tn5 transposase to cleave DNA and to integrate primer DNA sequences into the cleaved genomic DNA
ChIA-PET	Direct sequencing of proximity-ligated chromatin immune-precipitates.
FAIRE-seq	Formaldehyde Assisted Isolation of Regulatory Elements
Hi-C	Chromosome Conformation Capture technique where a biotin-labeled nucleotide is incorporated at the ligation junction, enabling selective purification of chimeric DNA ligation junctions followed by deep sequencing
ncRNA-Seq	Capture of other non-coding RNA types, including post-translation modification types such as snRNA (small nuclear RNA) or snoRNA (small nucleolar RNA), or expression regulation types such as siRNA (small interfering RNA) or piRNA/piwi/RNA (piwi-interacting RNA).
RAD-Seq	Restriction Site Associated DNA Sequence
RIP-Seq	Direct sequencing of RNA immunoprecipitates (includes CLIP-Seq, HITS-CLIP and PAR-CLIP).
SELEX	Systematic Evolution of Ligands by EXponential enrichment
ssRNA-seq	strand-specific RNA sequencing
Targeted-Capture	Targeted-Capture sequencing
Tethered Chromatin Conformation Capture	Tethered Chromatin Conformation Capture sequencing
Other	Library strategy not listed.

*** 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.

***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 Information

* **Alias**

Submitter assigned name or id for the GSA submission object.

* **Run data file type**

The GSA is a raw data archive, and requires per-base quality scores for all submitted data. GSA accepts binary files such as BAM, SFF, and HDF5 formats and text formats such as FASTQ. The major data file format we accept as shown below.

Format	File suffix	Recommended	Description
Fastq format	.fastq.gz, .fq.gz .fastq.bz2, .fq.bz2	Yes	fastq files with constant read length
BAM format	.bam	Yes	Binary SAM format for use by loaders that combine alignment and sequencing data
HDF5 format	.bax.h5 .bas.h5	Yes	HDF5 is a data model, library, and file format for storing and managing data.
Reference_FASTA	.fasta.gz, .fa.gz	Yes	Reference sequence file in single fasta format used to construct SRA archive file format.

Data Blocks

Fastq format (as an example)

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 for Forward**

Please fill in the whole name of the data files (including suffix). NOTICE whitespace characters are not allowed in the file names.We only accept GZIP and BZIP2 compression formats. IN ADDITION, don't accept 7-ZIP or TAR compressed files.

* **MD5 for Forward 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.

* **File Name for Reverse**

Please fill in the whole name of the data files (including suffix). NOTICE whitespace characters are not allowed in the file names.We only accept GZIP and BZIP2 compression formats. AND don't accept 7-ZIP or TAR compressed files.

* **MD5 for Reverse 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 (as an example)

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**

    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 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.

* **Local Assembly file**

    * **Reference file name**

    The Reference's file name

    * **MD5 for reference 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.

* **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 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.