Key Components and Terms on the GSA Website

The Genome Sequence Archive (GSA) at the National Genomics Data Center is a repository for raw sequence data. Understanding the main terms and components will help you efficiently find and interpret datasets1[35].

Core Terms and Their Meanings

Term	Description
BioProject	An umbrella record describing the overall research project. It groups together related experiments, samples, and datasets under one project. Useful for finding all data generated from a specific study or initiative.
BioSample	A record describing the biological source material (e.g., tissue, organism, cell line) used for sequencing. Each BioSample can be linked to multiple experiments or datasets.
Experiment	Describes the sequencing library, platform, and strategy used. Each experiment is associated with a BioSample and a BioProject.
Run	The actual raw sequence data files generated from a sequencing experiment. A single experiment may have multiple runs (e.g., technical replicates or different lanes).
Accession	A unique identifier for each BioProject, BioSample, Experiment, or Run. Accession numbers help you retrieve or cite specific records.

How These Components Relate

- A **BioProject** contains one or more **BioSamples**.
- Each **BioSample** can have one or more **Experiments**.
- Each **Experiment** can have one or more **Runs** (data files).

Navigating the GSA Website

- **Search**: You can search by any accession number (BioProject, BioSample, Experiment, or Run) or by keywords related to your research interest.
- **Metadata**: Each record provides metadata describing the sample, experimental method, and data files.
- **Download**: Once you find the dataset, you can download raw data files (usually FASTQ) from the associated Run records.
- **Submission**: If submitting data, you must register BioProjects and BioSamples first, then link Experiments and Runs to these records[3].

Example Workflow for Finding Datasets

1. **Start with a BioProject**: Search for a project related to your research area.

2. **Explore BioSamples**: Within a BioProject, view the list of BioSamples (e.g., tumor, normal tissue).
3. **Check Experiments**: For each BioSample, look at the experiments to see sequencing details.
4. **Access Runs**: Download raw data from the Run records associated with the experiments.

Summary Table

Component	What it Describes	Example Accession Prefix
BioProject	Research project/study	PRJNA, CRA
BioSample	Biological source/sample	SAMN, CRA, SRS
Experiment	Library prep/sequencing details	SRX, ERX, CRX
Run	Raw sequence data file	SRR, ERR, CRR

Additional Notes

- The GSA is part of the International Nucleotide Sequence Database Collaboration (INSDC), so its structure is similar to NCBI SRA and EMBL-EBI ENA1.
- For more guidance, the GSA website provides a documentation and quick start guide3.

In summary:
Understanding these terms—BioProject, BioSample, Experiment, and Run—will help you navigate the GSA, search for datasets, and interpret the results you find. Each dataset is organized hierarchically, and accession numbers are your key to accessing specific records and data files135.

Citations:

1. https://ngdc.cncb.ac.cn/gsa/
2. https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.965908/full
3. https://ngdc.cncb.ac.cn/gsa/document/start_here.jsp
4. https://cran.r-project.org/web/packages/GSA/GSA.pdf
5. http://english.big.cas.cn/ns/es/201703/P020170315583460100220.pdf
6. https://rdrr.io/cran/GSA/man/GSA.html
7. https://www.gsa.gov/governmentwide-initiatives/open-gsa/gsa-datasets
8. https://www.gsa.gov/directives-library/gsa-information-and-data-quality-handbook-1
9. https://static.googleusercontent.com/media/www.google.com/en/support/enterprise/static/gsa/docs/admin/current/gsa_doc_set/xml_reference/xml_reference.pdf
10. https://www.biorxiv.org/content/10.1101/2025.03.31.646270v1.full.pdf

11. https://galaxyproject.org/use/galaxy-gsa/
12. https://ngdc.cncb.ac.cn
13. https://data.gov/user-guide/
14. https://submit.ncbi.nlm.nih.gov/about/bioproject-biosample/
15. https://pmc.ncbi.nlm.nih.gov/articles/PMC9039563/
16. https://ngdc.cncb.ac.cn/gsa/
17. https://www.youtube.com/watch?v=5LpXqeM0onQ
18. http://english.big.cas.cn/ns/es/201703/P020170315583460100220.pdf
19. https://www.youtube.com/watch?v=67wGBzIYQXo
20. https://www.biorxiv.org/content/10.1101/2025.03.31.646270v2.full.pdf
21. https://cran.rstudio.com/web/packages/GSA/GSA.pdf
22. https://pmc.ncbi.nlm.nih.gov/articles/PMC8728233/

Answer from Perplexity: pplx.ai/share