



Genomic Data Access Through BLAST

<https://blast.ncbi.nlm.nih.gov>

Accessing genomic sequence data through BLAST, on the web or using standalone tools

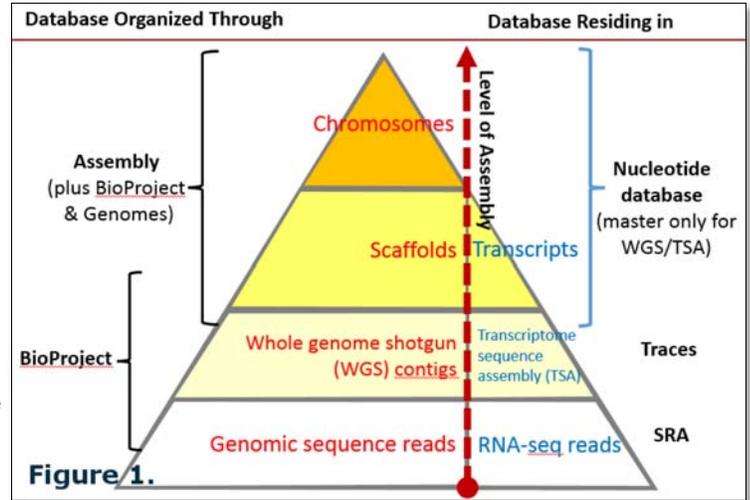
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Introduction

Advances in next generation sequencing technology (NGS) have led to the availability of genomic sequence data for an increasing large number of organisms. BLAST searching against these datasets, particularly the annotated assemblies based on raw sequence reads, can provide significant insight into the biology of these biomedically, agriculturally and ecologically important species. However, assembly and annotation from raw sequence reads is a complex process. The best sequence data availability will vary from organism to organism and may require different access strategies. In this document, we will go over the organization of genomic sequence data available from NCBI, and ways to locate the best genomic dataset for the organism of interest for sequence alignment purposes, through both the BLAST homepage at <https://blast.ncbi.nlm.nih.gov/>, and through other standalone tools provided by NCBI.

Workflow and Organization for Nucleotide Sequence Data

From an NGS-centric point of view, we can use a pyramid to represent the organization of available nucleotide sequence data based on volume, degree of assembly and information density, as well as the databases they reside in vs databases through which related records are organized (Figure 1). The figure separates nucleotide sequences into genomic and transcripts entries (left and right), and sorts them by their level of assembly (bottom to top). We should use the records at the top of the pyramid with the highest level of assembly and annotation are the most useful. It is better to access the the databases through the organizational databases such as Assembly or BioProject (left half) since these databases organize and connect related nucleotide records, such as individual chromosomes for a specific organism.



BioProject and Assembly Entries with Genomic Data

A BioProject database record provides a summary of a specific research project and lists all data available from the project. The result below (A) is from searching with “Barley[orgn] AND bioproject_assembly[filter]” (B). Click the title to open a record (C) for more details. The Project Data table lists available nucleotide sequences, with the number linked to actual records (wgs contigs in this case, D). The right hand column lists related records in other NCBI databases (insert, E). For project with only raw sequence reads, the link will point to Sequence Read Archive (SRA, F).

Project Types: Primary submission (4)

Data Types: Other (2)

Project Data: Nucleotide (6), Assembly (6), SRA (2)

Scope: Monoisolate (6)

Organism Groups: Plants (6)

Search fields: Choose ...

Display Settings: Summary, 20 per page, Sorted by Default order

A Search results: Items: 6

B Barley[orgn] AND bioproject_assembly[filter]

- Barley BAC Assemblies
1. In frame of the IBSC sequenced BACs from Barley, that constitute the MTP of barley.
Project data type: Other
Scope: Monoisolate
IBSC
Accession: PRJEB13020 ID: 340103
- Whole genome shotgun survey sequencing of Barley genotype B...
assembly of gene-space
Project data type: Other
Scope: Monoisolate
IPK-Gatersleben
Accession: PRJEB3038 ID: 313283
- Hordeum vulgare subsp. vulgare strain:cultivar Bowman
Whole Genome Shotgun Sequence assembly of Barley cv. Bo...
Organism: Hordeum vulgare subsp. vulgare
Taxonomy: Hordeum vulgare subsp. vulgare (domesticated barley)

C

Hordeum vulgare subsp. vulgare strain:cultivar Bowman (domesticated barley)
Accession: PRJEB88 ID: 179053
<https://www.ncbi.nlm.nih.gov/bioproject/179053>
Whole Genome Shotgun Sequence assembly of Barley cv. Bowman
Barley (Hordeum vulgare L.) is amongst the oldest domesticated crop plants and remains one of the world's most important crop species. More...

Accession: PRJEB88
Data Type: Genome sequencing and assembly
Scope: Monoisolate
Organism: Hordeum vulgare subsp. vulgare [Taxonomy ID: 112509]
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BOP clade; Pooidae; Triticoideae; Triticeae; Hordeinae; Hordeum; Hordeum vulgare; Hordeum vulgare subsp. vulgare
Submission: Registration date: 15-Nov-2012
IPK-Gatersleben

See Genome Information for Hordeum vulgare

NAVIGATE ACROSS: 205 additional projects are related by ...

D

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	1
OTHER DATASETS	
BioSample	1

E

Related information: Assembly, BioSample, Genome, Nucleotide, Taxonomy, WGS master

F

SRA Data Details

Parameter	Value
Data volume, Gbases	59
Data volume, Mbytes	38262

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	622850
WGS master	1
OTHER DATASETS	
BioSample	1
Assembly	1

Assembly details:

Assembly	Level	WGS	BioSample	Taxonomy
GCA_000326125.1	Scaffold	CAJX000000000	SAMEA2272683	Hordeum vulgare subsp. vulgare

BioProject and Assembly Entries with Genomic Data (cont.)

An assembly database record provides summary information for a specific genomic assembly. Searching with “barley[orgn]” and filtering for the “Representatives” retrieves a single record (A). The record (B) contains a detailed description of the assembly, with chromosomal level details given in the “Global assembly definition” table (C) when available (absent for this specific assembly). The “BLAST search this assembly” link in the right hand column (insert, D) leads to a Assembly-specific BLAST search form.

Organism Search Box in the BLAST Homepage

The organism search box in the BLAST homepage (<https://blast.ncbi.nlm.nih.gov/>) streamlines the process, and allows quick access to the best genomic dataset for the input organism. It returns organism-specific nucleotide BLAST pages in the following decreasing level of assembly:

- fully annotated RefSeq chromosomal assembly (full description at <http://1.usa.gov/TphKVT>)
- Scaffold level of assembly, generally with annotation and some genomic context, but without chromosomal placement
- WGS level of assembly, with or without annotation, and without genomic context and chromosomal placement
- Nucleotide search against NT database with the input organism as limit, if all the above fail.

The example below locates the genome assembly for the Chinese hamster. It returns the organism-specific BLAST page with annotated RefSeq genome as the target database: type the organism name in the input box to see a suggested list (E), select the desired entry from the list (F), click “GO” to get to the search page (G), and click the “?” icon to view a detailed description of the selected database. The example uses the mouse mRNA of the vitamin C synthesis gene (NM_178747.1, H) to identify the hamster counterpart.

Summary Download Assemblies Send to: ▾

Links from BioProject

Filters activated: Latest, Exclude anomalous. [Clear all](#)

ASM32612v1

Organism: *Hordeum vulgare* subsp. *vulgare* (domesticated barley) **A**

Infraspecific name: Cultivar: Bowman

Submitter: IPK-Gatersleben

Date: 2012/11/15

Assembly level: Scaffold

Genome representation: full

GenBank assembly accession: GCA_000326125.1 (latest)

RefSeq assembly accession: n/a **B**

IDs: [Display](#) [Settings](#): ☑ Full Report Send to: ▾

ASM32612v1

Organism name: *Hordeum vulgare* subsp. *vulgare* (domesticated barley)

Infraspecific name: Cultivar: Bowman

BioSample: SAMEA2272683

Submitter: IPK-Gatersleben

Date: 2012/11/15

Assembly level: Contig

Genome representation: full

RefSeq category: representative genome

GenBank assembly accession: GCA_000326125.1 (latest)

RefSeq assembly accession: n/a

RefSeq assembly and GenBank assembly identical: n/a

WGS Project: CAJX01

IDs: 513038 [UID] 513038 [GenBank]

[History \(Show revision history\)](#)

Global statistics

Total sequence length	1,779,486,241
Total assembly gap length	16,772,123
Number of contigs	2,961,602
Contig N50	1,647
Contig L50	235,981
Total number of chromosomes and plasmids	0

[Assembly Definition](#) [Assembly Statistics](#)

Global assembly definition **C** [Download the full sequence report](#)

The primary assembly unit does not have any assembled chromosomes or linkage groups. Please download the full sequence report for information on the scaffolds.

Access the data

[Download the GenBank assembly](#)

[BLAST search the assembly](#) **D**

[Download the full sequence report](#)

[Download the statistics report](#)

BLAST Assembled Genomes

Find Genomic BLAST pages:

chinese ha **E** GO

Chinese hamster (taxid:10029)

Chinese hamsters (taxid:10029)

Chinese hawthorn (taxid:510735)

Chinese hairy crab (taxid:95602)

Chinese habu (taxid:103944)

Chinese hare (taxid:112022)

F

Find Genomic BLAST pages:

Chinese hamster (taxid:10029) GO **G**

Enter organism common name, scientific name, or tax id.

Enter Query Sequence **H**

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange ▾

NM_178747.1

Or, upload file Choose File No file chosen ?

Job Title

Enter a descriptive title for your BLAST search ?

Choose Search Set

Database Genome (CriGri_1.0 reference, Annotation Release 101) (109152 sequences) ▾

Title: *Cricetulus griseus* CriGri_1.0 [GCF_000223135.1] chromosomes plus unplaced and unlocalized scaffolds (reference assembly in Annotation Release 101)

Description: The reference assembly set of RefSeq genomic top-level sequences (chromosomes, unplaced and unlocalized scaffolds) in a specific annotation run

Molecule Type: Genomic

Update date: 2014/05/07

Number of sequences: 109152

An Example Web BLAST Search Result

Searching for the vitamin C synthesis gene, using the mouse mRNA as a query and discontinuous megablast to optimize cross-species comparison, finds good matches in the hamster genome. The graphical overview indicates that all the query is covered by high score matches. The vertical lines indicate exon boundaries (A). Using the “Formatting options” (B), we can adjust the alignment view format and add translation of coding sequence. We can also sort the segments of alignment by the query start (C) to see them in exon order.

Web Access: Pros & Cons

BLAST access of genomic sequence data through the web interface has advantages and limitations (Table 1). For genomic BLAST searches requiring high throughput, customization, or workflow integration, alternative approaches may be better, which includes the standalone BLAST+ package, the vdbbased BLAST programs from the NCBI sratoolkit, as well as the cloud implementation. Table 2 summarizes some of the characteristics for blast+ and vdb-based BLAST tools.

Table 1. Characteristics of Genomic Data Access Through Web BLAST Interface

Pros	Cons
<ul style="list-style-type: none"> FAST: distributed computation (splitd) GUI: Interactive graphical user interface Extensive links to: <ul style="list-style-type: none"> * NCBI records * Tools (TreeView, Taxonomy Report) Visually informative format Coding Sequence Translation Graphical rendering through SV 	<ul style="list-style-type: none"> Browser bottleneck Limited capacity from CPU time restriction Difficult to <ul style="list-style-type: none"> * automate * incorporate into other workflow Limited search customization Limited custom data access

Table 2. Alternative Approaches for Genomic Data Access

Standalone blast+	Vdb-based blast
<ul style="list-style-type: none"> Client-server access through “-remote” using database and computation power at NCBI Local database access (downloaded from NCBI or formatted locally) Multi-threaded Search once and format multiple times through “-outfmt 11” and blast_formatter 	<ul style="list-style-type: none"> For sequence data stored in vdb format (WGS, TSA & SRA) Locally stored or on-demand download from prefetch Multi-threaded <p>[both alternatives are command line only with no graphical user interface]</p>

