SRA Handbook

Last Updated: January 14, 2016



National Center for Biotechnology Information (US), Bethesda (MD)

This documentation provides an overview and help manual for the Sequence Read Archive (SRA) at the National Center for Biotechnology Information.

Table of Contents

sing the SRA	1
Download Guide	3
Overview	3
Download with Prefetch	4
The Run Browser	5
SRA BLAST	7
Direct downloading of fasta and fastq format	9
Downloading metadata associated with SRA data files	9
Aspera Transfer Guide	13
Aspera Transfer Guide	13 13
Aspera Transfer Guide Notice Overview	13 13 13
Aspera Transfer Guide Notice Overview Aspera	13 13 13 13
Aspera Transfer Guide Notice Overview Aspera Using ascp to Download by Command Line	13 13 13 13 14
Aspera Transfer Guide Notice Overview Aspera Using ascp to Download by Command Line Using ascp to Upload by Command Line	13 13 13 13 14 15
Aspera Transfer Guide Notice Overview Aspera Using ascp to Download by Command Line Using ascp to Upload by Command Line Requirements	13 13 13 13 14 15 16

Using the SRA

SRA Handbook

Download Guide

Created: September 9, 2009; Updated: January 14, 2016.

Overview

The purpose of this document is to explain to users how to download datasets of interest and associated metadata.

Important Notes on Download Facilities

- One basic format (.sra) is provided by the SRA for all publicly available data. The SRA Toolkit is provided to allow conversion to several popular formats.
- At a minimum, users are advised to use Aspera Connect (or the equivalent command line tool, ascp) for bulk downloads, rather than HTTP or FTP. Aspera provides faster bandwidth, a higher level of flow control, user level encryption, and the ability to download trees of components.
- We most strongly recommend the use of the SRA Toolkit to download data files directly. The individual utilities are able to resolve SRA accessions and initiate downloads automatically. The 'prefetch' utility is specifically provided for researchers that wish to download SRA data using a command line utility.

Related Documents

NCBI Large Data Download Best Practices

Notices

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes.

Software Version

This guide is current to SRA Toolkit version 2.5.x. Instructions for previous versions of the SRA Toolkit may be different from those provided in this guide. We recommend that users stay current with SRA Toolkit updates to benefit from feature additions and bug fixes.

Reference Compression

Compression by reference is a sequence alignment compression process for storing data. Compression by reference stores the difference in base pairs between sequence data and the segment(s) to which it is aligned. Throughout this document you will note that the behavior and properties of reference compressed SRA data and conventional data differ significantly. Notably,

- The SRA Toolkit can output reference-compressed data as aligned sam and can perform pileup analysis.
- The SRA Toolkit requires internet connectivity in order to download reference sequences in order to process aligned SRA data.
- Only aligned data can be viewed in the NCBI Sequence Viewer.
- Aligned data cannot be filtered in the SRA Run Browser.
- Aligned data cannot currently be searched in SRA BLAST (this is actively being developed).

Download with Prefetch

The SRA Toolkit can be used to directly download SRA data files and reference sequences (see the "Reference Compression" section above). We strongly encourage users to use these methods to access SRA data as they are simple to use and they avoid many of the manual steps required by other methods (searching FTP directories, browsing and clicking, etc.).

The SRA Toolkit will have to be properly configured in order to access NCBI servers and download data. Recent versions of the Toolkit are packaged with a 'default' configuration that should work for most users. Please review the pros and cons for using the default configuration here. If the default configuration does not work for your installation, or you wish to customize aspects of file handling by the Toolkit (e.g., where downloaded files are stored locally), you will need to configure the Toolkit and then test it to confirm that it is operating as expected. Please email sra@ncbi.nlm.nih.gov if you have any problems configuring or using the Toolkit.

Prefetch

The 'prefetch' utility in the SRA Toolkit can be used to download SRA data and any required reference sequences in a single operation. Prefetch can use either HTTP (default) or ascp (if installed) to contact the SRA, resolve the accessions that you have specified, and then download the data. Prefetch can be used on single data files or to batch download several at a time. Below is an example prefetch command with the expected output. More information can be obtained on the prefetch documentation page and by executing 'prefetch --help'.

```
$ prefetch SRR390728
Maximum file size download limit is 20,971,520KB
2016-01-14T16:57:02 prefetch.2.5.7: 1) Downloading 'SRR390728'...
2016-01-14T16:57:02 prefetch.2.5.7: Downloading via fasp...
2016-01-14T16:57:08 prefetch.2.5.7: fasp download succeed
2016-01-14T16:57:08 prefetch.2.5.7: 1) 'SRR390728' was downloaded successfully
2016-01-14T16:57:09 prefetch.2.5.7: 'SRR390728' has 25 unresolved dependencies
2016-01-14T16:57:09 prefetch.2.5.7: 2) Downloading 'ncbi-acc:GPC_000000394.1?vdb-
ctx=refseq'...
2016-01-14T16:57:09 prefetch.2.5.7: Downloading via fasp...
2016-01-14T16:57:13 prefetch.2.5.7: fasp download succeed
2016-01-14T16:57:13 prefetch.2.5.7: 2) 'ncbi-acc:GPC_000000394.1?vdb-ctx=refseq' was
downloaded successfully
2016-01-14T16:57:13 prefetch.2.5.7: 3) Downloading 'ncbi-acc:GPC_000000395.1?vdb-
ctx=refseq'...
2016-01-14T16:57:13 prefetch.2.5.7: Downloading via fasp...
2016-01-14T16:57:15 prefetch.2.5.7: fasp download succeed
```

Note that the example file is reference-compressed and that prefetch automatically obtains the reference sequences required to extract data from the .sra file. If your Toolkit installation is not properly configured, or you elect to block the ability of the Toolkit to contact NCBI, you will then need to determine (1) if your downloaded dataset is reference-compressed, (2) if so, which references are required to access the data (see vdb-dump for an example of how to determine this), and (3) acquire the reference sequences manually here.

Other Toolkit utilities

All SRA Toolkit functions - most notably the 'dump' utilities that convert SRA data into other formats - are able to download data "on-the-fly" at runtime. This works like prefetch, as the tools will also automatically acquire all needed reference sequences. To invoke a Toolkit utility to download data as they are converted to your preferred format, simply execute the utility on an SRA accession rather than a local file. In other words, the command

Is implicitly requesting that fastq-dump download SRR390728 and its references from the SRA and then output the data in fastq format. Conversely,

\$ fastq-dump --split-files ~/Downloads/SRA/SRR390728.sra

Is instructing fastq-dump to operate on a local file that was previously downloaded from the SRA. In this case fastq-dump would still attempt to contact NCBI to obtain the references needed to convert the data to fastq (unless you have specifically configured the Toolkit to not contact NCBI).

The Run Browser

The SRA Run Browser can display sequencing and instrumentation data on a given run. Typically the Run Browser is reached as a click through from an Entrez SRA Experiment report. Users may also navigate by entering a run accession (SRR, DRR, or ERR) directly in the Run Browser.



Viewing data in the Run Browser

Reference compressed (aligned) SRA data have an "Alignment" tab. Clicking on this tab will allow you to configure the NCBI Sequence Viewer to display the data aligned to a reference sequence.

S NCB	Sit	e map All databases	Search						
🕕 Se	quen	ce Read Archi	ve						
Main B	rowse	Search Download	Submit Docum	entation	oftware	Trace Archi	ive Trace Assembly Trace Home Trace BLAST		
Studies	Sample	es Analyses Run B	rowser Provisio	nal SRA					
RNA	-Seq	(polyA+) ana	lysis of DL	BCL cel	line	HS0798	(SRR390728)		Change accession
Met	tadata	Alignment Reads	Download						
AI	ignme	ent Reads Bas	es Fract	ion					
Pri	imary	13.0 M 467.	2 Mbp 90.4	4%					
Re	eferen	ice						Range	
1							\$	1-1000000	
	omo sa What doe	piens chromosome es it do?	1, reference as:	sembly, co	mplete	sequence			
Vie	ew	scope	accession	count	in s	equence Viewer)		
		 this run 	SRR390728	1					
		same experiment	SRX079566	1					
		same sample	SRS212581	1					
		⊖same study	SRP001599	10					
		all sra		4,449					
Out	put this	s run in FASTA 🗧 f	ormat to Screen	n File					

To view the raw reads in a single Run, click on the "Reads" tab. Individual reads can be viewed and searched (see next section – note that only unaligned data can currently be searched). Various options can be applied using the "View" menu (e.g., display decimal quality scores, technical reads, etc.).

🗟 NCBI 🛛 Site map 🛛 All databases 🔊 Se	arch								
(III) Sequence Read Archive									
Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST									
Studies Samples Analyses Run Browser	Provisional SRA								
RNA-Seq (polyA+) analysis	of DLBCL cell line HS0798	(SRR390728)				Change accession			
Metadata Alignment Reads Downloa	ad								
< 1 1 717858 >	View:	✓ biological reads	technical reads	quality scores	advanced options				
1. SRR390728.1 SRS212581 name: 1, member: default 2. SRR390728.2 SRS212581 name: 2, member: default 3. SRR390728.3 SRS212581 name: 3, member: default 4. SRR390728.4 SRS212581 name: 4, member: default 5. SRR390728.5 SRS212581 name: 5, member: default 6. SRR390728.6 SRS212581 name: 6, member: default 7. SRR390728.8 SRS212581 name: 7, member: default 8. SRR390728.8 SRS212581 name: 8, member: default 9. SRR390728.9 SRS212581 name: 9, member: default 10. SRR390728.10 SRS212581 name: 10, member: default	Reads (separated) >gnl SRA SRR390728.1.1 1 unde CATTCTTCACGTAGTTCTCGAGCCTTC >gnl SRA SRR390728.1.2 1 unde GATGGAGAATGACTTTGACAAGCTGAC	efined (Biological, Re GTTTTCAGC efined (Biological, Re EAGAAGNTNC	verse) verse)						

The Run Browser supports IUPAC single letter nucelotide codes (data submitted in color space are presented in base space; the SRA Toolkit can be used to download and output the data in color space, if required). Quality scores are presented in the Phred scale.

Filtering and Selection

The Reads tab in the Run Browser can be used to filter and search reads according to certain regular expression pattern matching:

• Sequence substring: one of the biological reads for a spot should contain the substring. Examples: ATTGGA, ^ATTGGA, ATTGGA\$, ATGDNNAT, and ATGGA&GCGC. The strings are case insensitive, and belong to either 2NA or 4NA alphabets. String length limited to 29 characters in 4NA alphabet (includes IUPAC substitution codes) or 61 characters in 2NA alphabet (ACGT only). Search is case insensitive and strings may be combined with boolean operators & |! (AND, OR, NOT). See "SRA nucleotide search expressions" for more details.

- Name of a spot you are looking for. Example: EXWA4RL02G9Z6H
- Name of sample pool member, or "all" for all members. Example: M22_V2 will return all spots assigned to the sample pool member M22_V2 for run SRR031989.
- Spot Id. Example: 23

Please note that the filter searches across read boundaries within each spot. Thus, pattern matches within technical reads and across paired-end data boundaries will also be returned.

The filter provided in the Run Browser has limited functionality, but is quite fast if you are looking to quickly search a single Run for a defined sequence of interest. Please see the section below on SRA BLAST if you require more advanced searching or searches across multiple sequencing libraries.

Downloading Data from the Run Browser

Clicking on the "Download" tab in the Run Browser will present a selection of links that will allow you to download (1) an individual dataset (Run), (2) all datasets in a given sequencing library (Experiment), or (3) all datasets linked to a given project (Study). You are also provided with three download choices: Aspera (using the Aspera Connect plugin), HTTP (using your browser), FTP (using command line FTP or a client).

SNCBI Site map Al	ll databases 🛛 🗗	Search							
llı Sequence Read Archive									
Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST									
Studies Samples Analys	Studies Samples Analyses Run Browser Provisional SRA								
RNA-Seq (polyA	(+) analy	sis of DLBC	L cell line HS0798	(SRR39072	28)			Change accession	
Metadata Alignmen	t Reads Do	ownload							
Object	.sra								
Run SRR390728	193.6 Mb H	TTP FTP Aspera							
Experiment SRX079566	1.2 Gb H	TTP FTP Aspera							
Study SRP001599	14.0 Gb H	TTP FTP Aspera							

SRA BLAST

SRA BLAST can be used to for advanced searching of single or multiple sequencing libraries from the same or different projects. There are two ways to access SRA BLAST in order to build a "search space" from which you are attempting to pull matches to your sequence(s) of interest. Successful BLAST searches will lead you to a results / summary page that can be used to download reads of interest or be directed to the SRA Run Browser to further investigate or download the entire dataset.

Note that SRA BLAST currently has a limit of 2¹¹ reads (approximately 2 billion) per search – attempts to add more than 2¹¹ reads will result in an error and rejection of the search. Users that require more substantial search capability are advised to contact the SRA (sra@ncbi.nlm.nih.gov) to determine if other SRA BLAST tools might be of use.

Sending Entrez results to SRA BLAST

After performing an Entrez query to restrict results to datasets of interest, you may use the "Send to" feature to select datasets of interest and send them to SRA BLAST.

SRA	SRA + human[organism] NOT sra_nuccore_alignment[Filter] Save search Advanced	Help
Show additional filters	Display Settings: 🖓 Summary 20 per page	
	Dispray Seutings, E) Summary, zu per page	
Access	Choose Destination sms [Tree]	
Controlled (6051)	Results: 1 to 20 of 6/529 Selected: 2 <- First < Prev Pac File Olipoard s (67520)	
Public (52428)	Cvisic Ebrosic Longitudinal Study	
Course	1 II LI IMMA (Illumina Schart Scharts 64 RM hases 28 AMh downloads	
DNA (50202)	Accession: SRX47934 sector and a sector and	
DNA (13786)	Send	
metagenomic (736)	Cystic Fibrosis Longitudinal Study	
metagenomic (100)	2. 1 ILLUMINA (Illumina MiSeq) run: 734,091 spots, 196.7M bases, 116.6Mb downloads	
Туре	Accession: SRX479333 THP-1 total RNA (18)	
exome (15621)	Human GM18940 whole genome (5)	
genome (10428)	GSM1335757: 487b; Homo sapiens; RNA-Seq	
	3. 1 ILLUMINA (Illumina HiSeq 2000) run: 42.5M spots, 8.6G bases, 3.4Gb downloads	
Search fields	Accession: SRX476867 Top Bioprojects	
Choose	NIH Epigenomics Roadmap Init (1845)	
	GSM1335756: 487a; Homo sapiens; RNA-Seq Production ENCODE epigenomic (1495)	
Clear all	 1 LLUMINA (Illumina HiSeq 2000) run: 21.2M spots, 4.3G bases, 1.7Gb downloads Production ENCODE Truttorian (1752) 1000 Geneme Braiset Dilution (1752) 	
Show additional filters	Accession: SKA470800 Production ENCODE transcript	
onon additional mitors	CSM1235755: 477b; Home capitons: DNA Sec. MicroRNA sequence and expres (245)	
	5 1111/11/11/11/11/11/11/11/11/11/11/11/1	
	Accession: SRX476865	

SRA-BLAST does not currently support reference compressed SRA datasets, so it is generally advised that you add the condition 'NOT sra_nuccore_alignment[Filter]' (as in the above example) to your queries to remove these datasets from the search results. Attempting to send incompatible datasets to BLAST will result in an error like the following:

Failed to convert SRX SRX079566 to SRA runs
 Invalid SRX accession(s): SRX079566

If you believe that the data you are attempting to search against should be BLAST-able, but are not, please email sra@ncbi.nlm.nih.gov for assistance and advice. After successfully sending accessions to SRA BLAST, you are then able to input your sequence(s) of interest and perform the search.

ICBI/ BLAST/ blastn s	suite	Sequence Read Archive Nucleotide BLAST	Status of the NCBI Sequence Read Archive (SRA
lastn			
Enter Query S	equence	BLASTN programs search SRA databases using a nucleotide query. 😡	Reset page Bookmark
Enter accession nu	umber(s), gi(s), or FASTA sequence(s) 🚱	Clear Query subrance (9)	
		From	
		То	
		<i>h</i>	
Or, upload file	Choose File No file chosen		
Job Title			
	Enter a descriptive title for your BLAST search 🥹		
Choose Searc	h Set		
SRA Experiment	Sequences: 127,319,242		
set (SRX)	SRX476867	+	
	SRX476866	mission) title the exigntific name as tay id. Only 20 tan suggestions will be shown	
	Liner an orvi accession (experiment, study, or sur	nnission), uue, uie solehuile name or laxile. Only zo lop suggesuons will be shown.	
Program Selec	ction		
Optimize for	 Highly similar sequences (megablast) 		
	O More dissimilar sequences (discontiguous n	negablast)	
	 Somewhat similar sequences (blastn) 		
	Choose a BLAST algorithm 😡		
DIAGT	Courts database CDA using Manablast (Onli		
BLAST	Search database SRA using Megablast (Opti	mize for highly similar sequences)	
Algorithm parame	ters Note: Parameter	values that differ from the default are highlighted in yellow and marked with \blacklozenge sign	

Building a search list in SRA BLAST directly

SRA BLAST can be accessed directly. You will then need to provide SRA Experiment (SRX, DRX, or ERX) accessions or use the autocomplete feature to help refine your search. You may enter 1 Experiment accession per

line in the search list. The '+' button can then be used to add additional sequencing libraries to the search space. Note that a running tally of the number of sequences is presented above the list of accessions. Again, there is currently a limit of approximately 2 billion sequences per individual SRA BLAST query.

NCBI/ BLAST/ blastn sui	ite Sequence Read Archive N	ucleotide BLAST	Status of the NCBI Sequence Read Archive (SRA)							
blastn										
Enter Query Sec	BLASTN programs search SRA databases u	sing a nucleotide query. 😡	Reset page Bookmark							
Enter accession nun										
Enter accession nun		je 🐨								
	From									
	То									
Or, upload file	Choose File No file chosen									
JOD IITIE	Enter a descriptive title for your PLAST search									
Choose Search	Set									
SRA Experiment	Sequences: 85,017,716									
Set (SRA)	SRX47 SRX47 SRX47									
	SPX470010 Amborena inchopoda bisulprite sequencing (Amborena inchopoda taxid									
Program Selecti	C SRX470035 GSW1527146. Hord The Section Section (Citr), Mus musculus, Chin-Sec (Mus musculus)									
Optimize for	SPA470034 GSM1327149: Input DIVA (Ciri); Mus musculus; ChilP-Seq (Mus musculus; ChilP-Seq (Mus									
	SRX470035 GSW1327150. HSK9ffe5_CHIF-Seq (Ku), Mus musculus, ChiF-Seq (Mus									
	SRX470036 GSM1327151: Input DNA (Kd); Mus musculus; ChIP-Seq (Mus musculu									
	SRX470047 whole genome shotgun sequencing of Salmonella enterica subsp. salam									
	SRX470048 Whole genome shotgun sequencing of Salmonella enterica subsp. enter									
BLAST	SRX470049 Whole genome shotgun sequencing of Salmonella enterica subsp. enteri									
DLIDI	SRX470050 Whole genome shotgun sequencing of Salmonella enterica subsp. enteri									
+Algorithm parameter	SRX470051 Whole genome shotgun sequencing of Salmonella enterica subsp. enteri	abted in vellow and marked with sign								
- agenant parameter	SRX470052 Whole genome shotgun sequencing of Salmonella enterica subsp. enteri	gines in jonen and maned with v sign								
	SRX470053 Whole genome shotgun sequencing of Salmonella enterica subsp. salam									

Direct downloading of fasta and fastq format

The SRA provides a tool that can be used to download data directly in fasta or fastq format. You must provide one or more SRA Experiment (SRX, DRX, or ERX) accessions in a comma-separated list. The same filtering inputs available in the Run Browser (described above) are available here to restrict the number of returned reads. Certain reads can also be clipped to remove low quality data from the download. If more than one Run accession in the list is checked, all data will be downloaded into a single fasta or fastq file, rather than per-accession files. Note that the output format of this tool is pre-defined and cannot be adjusted at the time of download. Users with specific formatting needs (e.g., for downstream analysis) are encouraged to use the SRA Toolkit to download and convert the data (described above).

🗟 NCBI 🛛 Site ma	p All database	s 🔝 Search							
llı Sequence Read Archive									
Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST									
FASTA/FASTQ Reads Analyses Reports References									
Download	Download for Experiment SRX079566								
Accession	# of bases	# of spots		Filter					
✓ <u>SRR292241</u>	699.9 M	9.7 M		Search:	Apply				
SRR390728	516.9 M	7.2 M		What can the filter be applied to?					
				Download Format					
				🗆 filtered 📄 clipped 💿 FASTA 💿 FASTQ	Download				

Downloading metadata associated with SRA data files

SRA data files do not contain any information about the metadata (sample information, etc.) linked to the data themselves. The SRA provides a few tools to allow downloading of metadata in batch. Note that these tools differ

from the Entrez Experiment, BioSample, and BioProject reports for a given dataset and may not contain all relevant metadata.

Viewing and downloading tabular metadata with the SRA Run Selector

The SRA Run Selector can be used to view metadata from one or more projects (SRA Study accessions – SRP, DRP, or ERP) entered into the field at the top of the page. The Run Selector provides a table view of library preparation and sample attribute metadata. The table can be filtered by sample attribute(s), accessions, etc. The "Get Metadata" button can be used to download a table (.txt, tab-delimited) of all or selected metadata.

S	NCBI	SRA RUN	Sele	ctor	Chan	ge Study srpo	001599	Change						
0	Common	attributes												
	SRA Stud	v BioProje	ct ana	lyte type	gap	accession	is tumor	study name			Assay Type	Center Name	Platform	Conse
	CDD00150	0 pbc00023			phc0(00225	1	NCI Cancor Conomo C	haractorization	a Initiativo (CCCI)	DNA-Soc	BCCACSC		public
	3KF00139	9 phs00023		`	prisor	00233	1	NCI Cancel Genome C		Tinuauve (CGCI)	КМА-Зеч	DECAGGE	ILLOPIINA	public
		Runs	BioSam	nples I	MBytes	MBases		Pe	ermaLink					
	Total	20		10	13,363	22,364	Get Meta	adata						
	Filtered	20		10	13,363	22,364	+	- x						
	Selected	0		0	0	0	Show	w selected	letadata					
	Selected	U		U	U	U	U onor		letauata					
				Sample								Library		
	Run	BioSar	nple	Name	S	RA Sample	DSMZ	body_site	cell_line	sex	study_subject_	id Name	MBases	MBytes
	CDD2022		000070	UCOCOF		00212500	ACC 47	nlaunal offician	DOULL 2	mala	DOULL 2	LICOCOF	205	220
	SRR2922	40 SAMINU	0630373	H50685	5	R5212580	ALC 47	fluid	DOHH-2	male	DOHH-2	H50685	395	239
	SRR2922	41 SAMNO	0630374	HS0798	S	RS212581	ACC 539	peritoneal cavity	DB	male	DB	HS0798	667	958
								fluid						
	SRR2922	42 SAMNO	0630375	HS0841	S	RS212582	ACC 32	pleural effusion	Karpas 422	female	Karpas 422	HS0841	718	447
	SRR2922	43 SAMNO	0630376	HS0842	S	RS212583	ACC 583	lymph node	NU-DHL-1	male	NU-DHL-1	HS0842	681	526
0	SRR2922	44 SAMNO	0630377	HS0900	S	RS212584	ACC 572	peritoneal cavity	SU-DHL-6	male	SU-DHL-6	HS0900	766	691
_								fluid						
	SRR2922	45 SAMNO	0630378	HS0901	S	RS212585	ACC 575	pleural effusion	WSU-DLCL2	male	WSU-DLCL2	HS0901	896	1,053
0	CDD0000		0630370	HS1163	0	DC212586	ACC 722	fluid	OCI-LV1	male		HS1163	1 775	1 386
0	CDD2022	47 CAMNO	0030379	LC1103	3	00212500	ACC 699	poriphoral blood	OCLIVZ	male	OCLUX7	HS1103	2,140	1,500
0	CDD2022	47 SAMINU	0030300	101102	5	00212507	ACC 000	periprieral blood	OCI-LIT	fomale	OCI-LIV	HS1102	2,140	1,005
	SKR2922	40 SAMINU	0030301	HC2047	3	R5212500	ACC-520	Done marrow	OCI-Ly19-KI	Terridie	OCI-Ly19-KI	HS1105	1,019	1,205
	SKR2922	49 SAMINU	0630382	H52047	5	R5212589	ACC 579	cerebrospinal fluid	NU-DUL-1	male	NU-DUL-1	HS2047	3,462	1,454
	SRR3907	26 SAMNU	0630382	HS2047	S	RS212589	ACC 579	cerebrospinal fluid	NU-DUL-1	male	NU-DUL-1	HS2047	3,394	821
	SRR3907	27 SAMNO	0630373	HS0685	S	RS212580	ACC 47	pleural effusion	DOHH-2	male	DOHH-2	HS0685	342	137
	SRR3907	28 SAMNO	0630374	HS0798	S	RS212581	ACC 539	peritoneal cavity	DB	male	DB	HS0798	492	184
0					-			fluid						
	SRR3907	29 SAMNO	0630375	HS0841	S	RS212582	ACC 32	pleural effusion	Karpas 422	female	Karpas 422	HS0841	602	143
	0000007			1100040	-	00040500	100 500	fluid				11000.40		100
	SRR3907	30 SAMNO	0630376	HS0842	S	RS212583	ACC 583	lymph node	NU-DHL-1	male	NU-DHL-1	HS0842	442	180
	SRR3907	31 SAMNO	0630377	HS0900	S	RS212584	ACC 5/2	fluid	SU-DHL-6	male	SU-DHL-6	HS0900	501	185
	SRR3907	32 SAMNO	0630378	HS0901	S	RS212585	ACC 575	pleural effusion	WSU-DLCL2	male	WSU-DLCL2	HS0901	557	238
0								fluid						
	SRR3907	33 SAMNO	0630379	HS1163	S	RS212586	ACC 722	bone marrow	OCI-LY1	male	OCI-LY1	HS1163	1,021	634
	SRR3907	34 SAMNO	0630380	HS1182	S	RS212587	ACC 688	peripheral blood	OCI-LY7	male	OCI-LY7	HS1182	917	613
	SRR3907	35 SAMNO	0630381	HS1183	S	RS212588	ACC-528	bone marrow	OCI-Ly19-R1	female	OCI-Ly19-R1	HS1183	977	606

Command line access to metadata with the SRA Run Info CGI

Users can access the SRA Run Info CGI either through a browser or using a command line tool like wget.

wget -0 <file_name.csv> 'http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi? save=efetch&db=sra&rettype=runinfo&term=<query>'

As a parallel to the above example in the Run Selector,

```
wget -0 ./SRP001599_info.csv 'http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?
save=efetch&db=sra&rettype=runinfo&term= SRP001599'
```

Will return essentially the same information. Note that the CGI returns data in a comma-separated value (.csv) format, rather than the tab-delimited format of the Run Selector. The last component, <query>, can contain any set of Entrez parameters. Users may refine a search using Entrez and then copy over the search terms to a script for batch downloading. As an example, the search string

"Homo sapiens"[Organism] AND "cancer"[All Fields] AND "cluster_public"[prop] AND "strategy wgs"[Properties]

Will return these results in an Entrez search of the SRA. The equivalent Run Info CGI search would be

```
wget -0 ./query_results.csv 'http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?
save=efetch&db=sra&rettype=runinfo&term="Homo sapiens"[Organism] AND "cancer"[All Fields]
AND "cluster_public"[prop] AND "strategy wgs"[Properties]'
```

Note that Entrez groups by Experiment accession, but that the CGI does not. It is, therefore, to be expected that the Run Info CGI will return a longer list of results than Entrez, but will still contain the same datasets.

SRA Handbook

Aspera Transfer Guide

Created: May 11, 2009; Updated: April 16, 2014.

Notice

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes.

Overview

This document provides instructions on the use and installation of Aspera Connect for high throughput file transfer with NCBI. As the sizes of the datasets have increased, we have found that the traditional methods of *ftp* or *http* do not have the performance characteristics needed to support this load of data.

Requirements for large scale data transfer over the internet include high bandwidth, auto checksum, recursive copy, and security based on strong keys. NCBI has chosen to use a product from Aspera, Inc (Emeryville, CA) because of improved data transfer characteristics. FTP and HTTP access will continue to be available and are the default options for users without Aspera installed. Instructions are provided below for investigators to use this data transfer technology. NCBI also is open to using additional products with the appropriate performance characteristics.

Scope

This document is intended for users transferring large data files to and from NCBI. It applies to the Sequence Read Archive (SRA), dbGaP, and other archives where aspera download is enabled.

Aspera

Aspera Connect

Aspera Connect is software that allows download and upload via a web plugin for popular browsers on machines running Linux, Windows, and Macintosh. The software also includes a command line tool (ascp) that allows scripted data transfer. The software client is free for users exchanging data with NCBI.

Download and install Aspera Connect software from: http://downloads.asperasoft.com/connect2/

The website's download button will default to the detected operating system of the user's computer. To download for a different OS, click the link to 'See all installers'.

Please note the Requirements and consult with your network administrator to ensure transfers with aspera will not be blocked.

Aspera can be installed for individual users. However users of shared machine may want to have the software installed for all users by a system administrator.

The fasp Protocol

The FASP protocol from Aspera (www.asperasoft.com) uses UDP, eliminating the latency issues seen with TCP, and provides bandwidth up to 5 gigabit per second (Gbps) to transfer data. It has a restart capability if data transfer is interrupted midstream and is well behaved, so if there is other data traffic on your network

connections, it will back off in order to avoid starving other protocols. We have seen effective throughput up to 800 megabits per second (Mbps) to a single site.

Downloading Data with Aspera Connect Browser Plugin

Once the plugin has been installed in your browser, you may download files or entire directories from NCBI using Aspera. Example: In your browser window, go to

http://www.ncbi.nlm.nih.gov/public/?/ftp/sra/sra-instant/reads/ByRun/sra/SRR/SRR292/SRR292241

Click 'SRR292241.sra' to begin saving the data. You will be prompted to select where the file is to be saved. For example:

🕘 Save As					x
🕞 🗢 🖡 🕨 Do	wnloads	▼ 4 j	Search Download	s	٩
Organize 🔻 New	v folder			!≡ ▼ (0
Favorites Favorites Desktop Downloads Example Recent Places Libraries Documents Music Pictures Videos	▲ Name	No items match you	Date modified r search.	Туре	
					P.
File <u>n</u> ame: Save as <u>t</u> ype:	SRR1103214.sra sra files (*.sra)				•
Hide Folders			Save	Cancel)

You can download full directories or a single file at a time. The Aspera Connect plugin works with Chrome, Internet Explorer (IE), Safari, and FireFox web browsers. In some cases Aspera Connect may create a popup window to get a confirmation for file transfer and this popup window can be hidden behind your current web browser.

Using ascp to Download by Command Line

The command line program *ascp* is a utility delivered along with the Aspera Connect product.

```
ascp -i <asperaweb_id_dsa.openssh with path> -k1 -Tr -l100m
anonftp@ftp.ncbi.nlm.nih.gov:/<files to transfer> <local destination>
```

• -i <asperaweb_id_dsa.openssh with path> = fully qualified path & file name where

this public key file is located. This file is part of Aspera Connect distribution and is usually located in the 'etc' subdirectory.

- -T to disable encryption
- -k 1 enables resume of partial transfers
- -r recursive copy
- -l (maximum bandwidth of request, try 100M and go up from there)

Experiment with transfers starting at 100 Mbps and working up to 400 Mbps. Select the bandwidth setting that gives good performance with unattended operation.

- <files(s) to transfer> = names of files to transfer (including path)
- <local destination path> = location to store the downloaded data

Windows Executable Location

The *ascp* program for Microsoft Windows is located by default in "*C*:*Program Files**Aspera**Aspera* Connect*bin**ascp.exe*"

OS X Executable Location

The ascp Mac program location is /Applications/Aspera Connect.app/Contents/Resources/ascp

Linux Executable Location

The ascp Linux program location is /opt/aspera/bin/ascp

Additional information is available at the Aspera Web site: http://downloads.asperasoft.com/documentation/

Using ascp to Upload by Command Line

In order to use the Aspera upload service you will need to use a **private** SSH key, individual users can contact us at sra@ncbi.nlm.nih.gov to request an Aspera private key.

Upload Command

ascp -i <private key file> -T -l 100m <file(s) to transfer> asp-****@upload.ncbi.nlm.nih.gov:<destination directory>

- -i < private key file > = fully qualified path & file name of the private SSH key
- -T to disable encryption
- -k 1 enables resume of partial transfers
- -l (maximum bandwidth of request, try 100M and go up from there)

Experiment with transfers starting at 100 Mbps and working up to 400 Mbps. Select the bandwidth setting that gives good performance with unattended operation.

- <files(s) to transfer> = names of files to transfer (including path)
- <destination directory> = deposit location of the uploaded data (typically either 'test' or 'incoming')

For password protected private keys, it is possible to run *ascp* in an autonomous, unattended manner that does not require repeated login. The environmental variable ASPERA_SCP_PASS can be used to store the private key path for a scripted series of bulk uploads.

Key Pairs

SSH keys are used for establishing secure connections to remote computers.

Submitters using a dedicated center account can find instructions for generating a key pair or converting PuTTY format private keys to OpenSSH format in this guide.

```
http://www.ncbi.nlm.nih.gov/books/NBK180157/
```

Requirements

Firewall Requirements

Your local firewall must permit UDP data transfer in both directions on ports 33001-33009 for the following IP ranges:

130.14.*.*

165.112.*.*

The firewall must also allow ssh traffic outbound to NCBI.

Troubleshooting

Here are some example commands demonstrating a test download.

Mac OS X:

```
ascp -T -1640M -i "/Applications/Aspera Connect.app/Contents/Resources/
asperaweb_id_dsa.openssh" anonftp@ftp.ncbi.nlm.nih.gov:1GB /tmp/
```

Linux:

```
ascp -T -1640M -i /opt/aspera/etc/asperaweb_id_dsa.openssh
anonftp@ftp.ncbi.nlm.nih.gov:1GB /tmp/
```

MS Windows:

```
C:\TEMP>"C:\Program Files (x86)\Aspera\Aspera Connect\bin\ascp.exe" -T -1640M -
i "C:\Program Files (x86)\Aspera\Aspera Connect\etc\asperaweb_id_dsa.openssh " anon
ftp@ftp.ncbi.nlm.nih.gov:1GB C:\Temp\
```

For additional assistance, please contact the NCBI Help desk at info@ncbi.nlm.nih.gov

When you are about to contact the NCBI Help desk please provide them some basic information like operating system, version of aspera connect, type of disk storage used for transferring files and the type of network connection your organization has to the internet.

If you have a Linux or MacOS X operating system you may run these commands and show us their output:

```
curl -o /dev/null ftp://ftp.ncbi.nlm.nih.gov/1GB
curl -o /dev/null http://www.ncbi.nlm.nih.gov/staff/beloslyu/large.tar
traceroute ftp.ncbi.nlm.nih.gov
```

First two commands download a 1GB file from NCBI using ftp and http protocols, the content is dumped to /dev/null. The third command will let us see the latency in your internet connection and possible congestions on the way to NCBI.

Another possibility is to make some test downloads from Aspera's demo server, for Linux the command line is:

env ASPERA_SCP_PASS=demoaspera ascp -L- -T -l100m aspera@demo.asperasoft.com:aspera-testdir-large/1GB /tmp/

Aspera Connect is a commercial product and program specific support is available from the manufacturer at http://asperasoft.com/support/

The currently up-to-date documentation for ascp can be found at http://downloads.asperasoft.com/en/ documentation/8