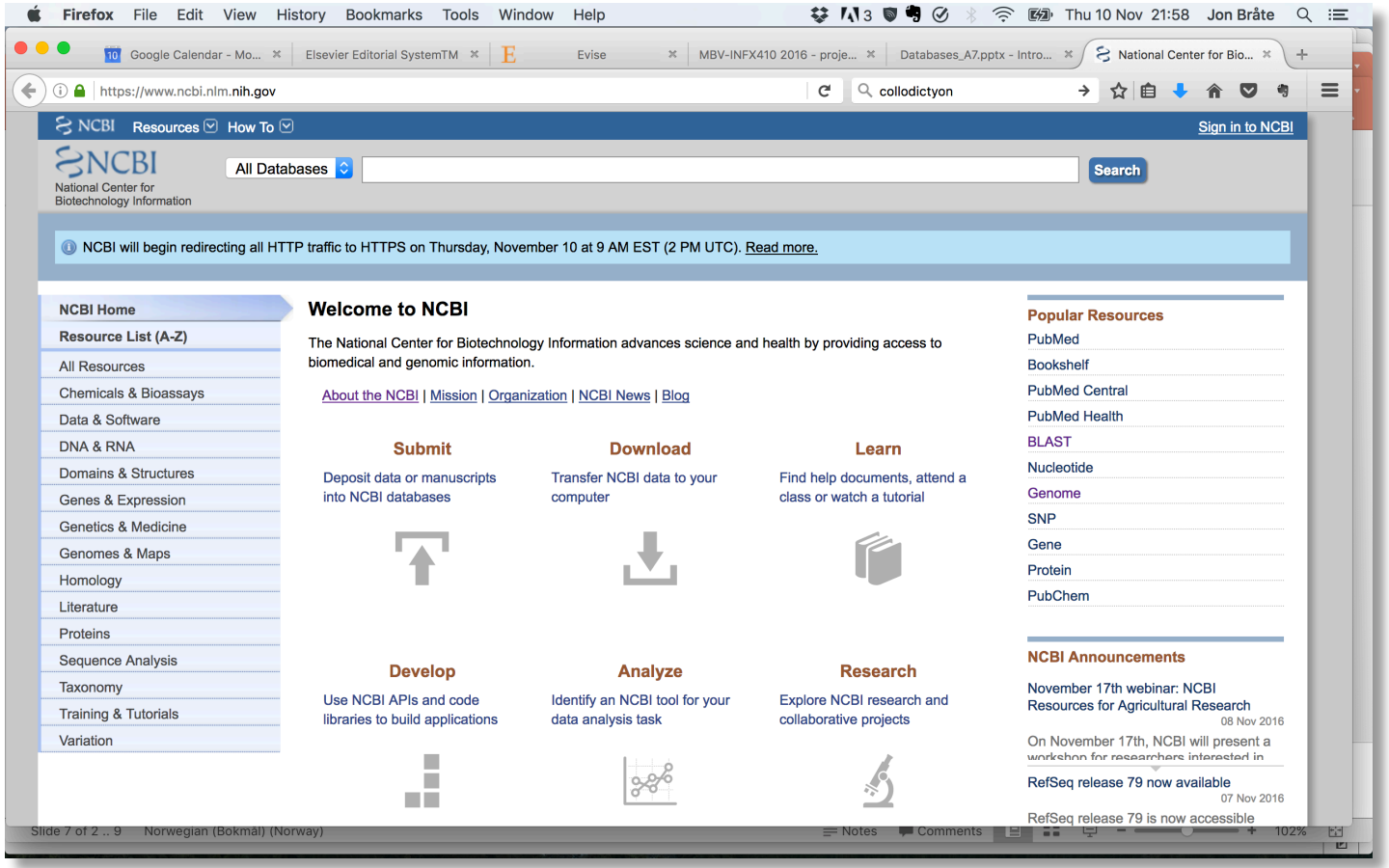
UiO : Department of Biosciences
University of Oslo

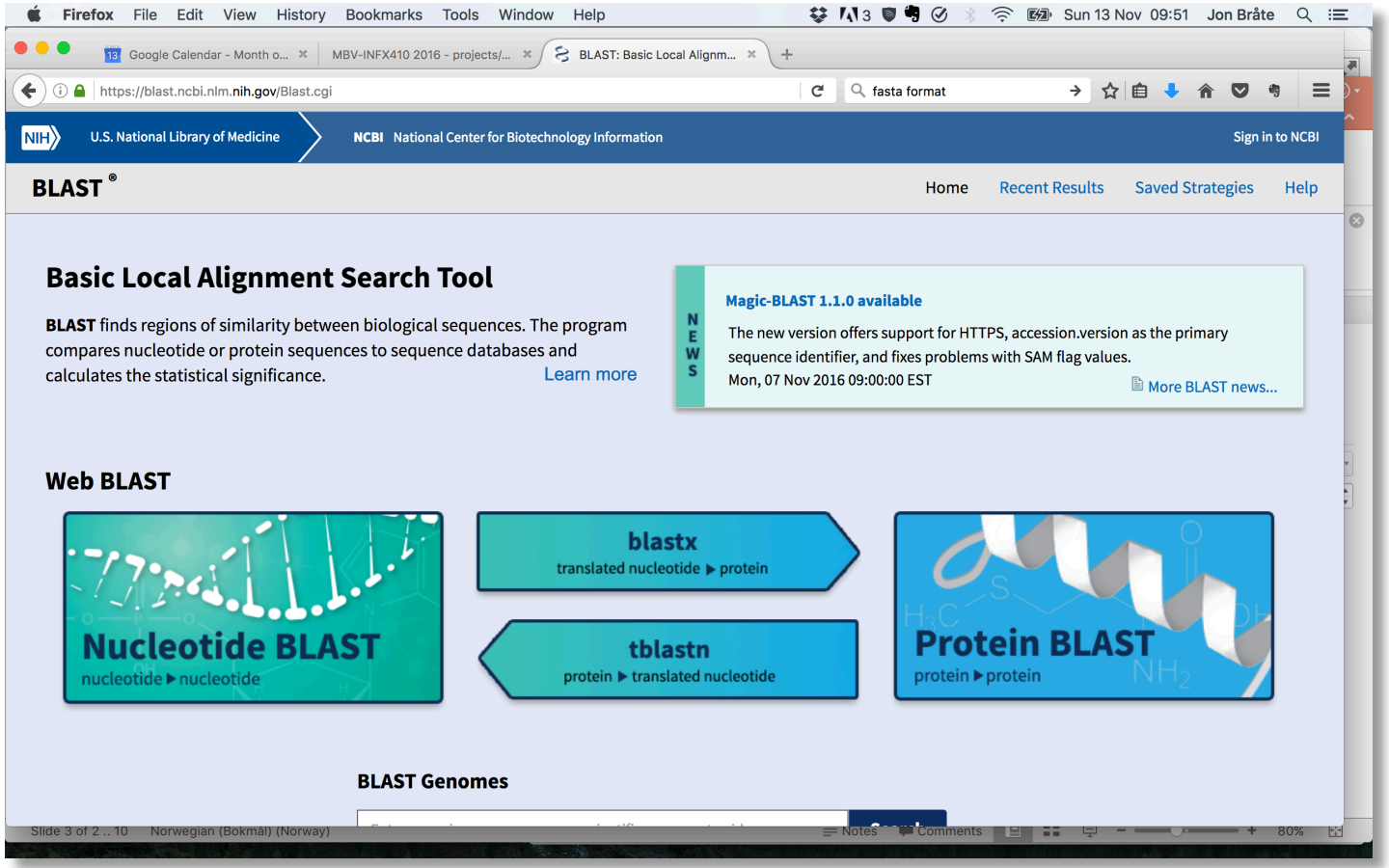
MBV4410/9410 Fall 2016

Bioinformatics for Molecular Biology

NCBI (https://www.ncbi.nlm.nih.gov/)

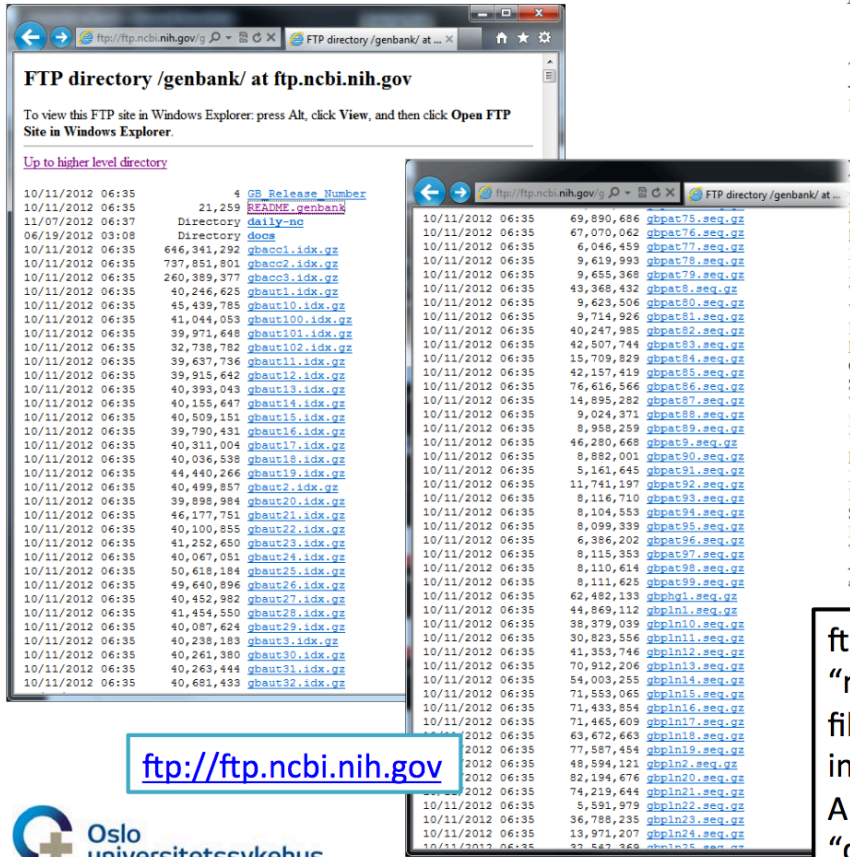


Blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi)



Jon K. Lærdahl,
Structural Bioinformatics

NCBI ftp site



Nucleic Acids Research, 2012, Vol. 40, Database issue **D49**

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 185 (8/2011)	Annual increase (%)[a]
TSA	Transcriptome shotgun data	1 874 047 448	370.1
ENV	Environmental samples	2 553 693 157	48.2
PHG	Phages	62 579 756	44.0
PAT	Patented sequences	11 154 487 762	30.9
BCT	Bacteria	6 975 597 755	30.8
INV	Invertebrates	2 535 336 197	24.5
WGS	Whole-genome shotgun data	208 315 831 132	23.1
VRL	Viruses	1 180 083 600	21.6
MAM	Other mammals	807 098 397	18.8
PLN	Plants	4 741 991 057	17.4
GSS	Genome survey sequences	20 770 772 329	12.6
SYN	Synthetic	156 218 063	9.6
VRT	Other vertebrates	2 705 250 711	6.8
EST	Expressed sequence tags	39 018 185 344	6.0
UNA	Unannotated	125 912	4.7
PRI	Primates	6 116 546 725	2.9
ROD	Rodents	4 396 957 541	2.3
HTC	High-throughput cDNA	662 320 919	0.4
STS	Sequence tagged sites	635 872 683	0.3
HTG	High-throughput genomic	24 324 068 445	0.2
TOTAL	All GenBank sequences	338 987 064 933	18.2

[a]Measured relative to Release 179 (8/2010).

ftp = file transfer protocol
"network protocol for transfer of files from one host to another on the internet"
An ftp site is very basically a "directory with files on the internet"

ftp://ftp.ncbi.nih.gov

Fasta format

Line shift
(hard return)

```
>NP_001258677.1 DNA-(apurinic or apyrimidinic site) lyase 2 isoform 2 [Homo sapiens]
MRFYRLLQIRAEALLAAGSHVIILGDLNTAHRPIDHWDAVNLECFEEDPGRKWMDSLLSNLGCQSASHVGPFIDSYRCFQPKQE
GAFTCWSAVTGARHLNYGSRLDYVLGDRTLVIDTFQASFLLPEVMGSDHCPVGAVLSVSSVPAKQCPPLCTRFLPEFAGTQLKI
LRFLVPLEQSPVLEQSTLQHNNQTRVQTCQNKAQVRSTRPQPSQVGSSRGQKNLKSYFQPSPSCPQASPDIELPSLPLMSALMT
PKTPEEKAVAKVVKGQAKTSEAKDEKELRTSFWKSVLAGPLRTPLCGGHREPCVMRTVKKPGPNLGRRFYMCARPRGPPTDPSS
RCNFFLWSRPS
```

- Header/definition line. Always begins with “>”
- Sequence on a separate line(s)
- Amino acids or nucleotide

```
>gi|330443681:c455932-455572 Saccharomyces cerevisiae S288c chromosome XII, complete
sequence
AAGAAATTTAATAATTTTGAAAATGGATTTTTTTGTTTTGGCAAGAGCATGAGAGCTTTTACTGGGCAAGAAGACAAGAGATGG
AGAGTCCAGCCGGGCCTGCGCTTAAGTGCGCGGTCTTGCTAGGCTTGTAAGTTTCTTTCTTGCTATTCCAAACGGTGAGAGATT
TCTGTGCTTTTGTTATAGGACAATTAAAACCGTTTCAATACAACACACTGTGGAGTTTTCATATCTTTGCAACTTTTTCTTTGG
GCATTCGAGCAATCGGGGCCCAGAGGTAACAAACACAAACAATTTTATCTATTCATTAAATTTTTGTCAAAAACAAGAATTTTC
GTAACTGGAAATTTTAAAATATTAA
```

GenBank vs. RefSeq

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

http://www.ncbi.nlm.nih.gov/books/NBK21105/#ch1.Appendix_GenBank_RefSeq_TPA_and_UniP

Firefox File Edit View History Bookmarks Tools Window Help Thu 10 Nov 22:03 Jon Bråte

Google Calendar - Mo... | Elsevier Editorial SystemTM | Evise | MBV-INFX410 2016 - proje... | Databases_A7.pptx - Intro... | UniProt

www.uniprot.org

UniProt

UniProtKB Advanced Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional informati

Basket 5

UniProtKB
UniProt Knowledgebase

Swiss-Prot (552,884)
Manually annotated and reviewed.

TrEMBL (70,656,157)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data
Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2016_10
N-acyl amino acids: a new treatment for obesity? | Cross-references to DisGeNET and OpenTargets | Change of the cross-references to Phos...

UniProt release 2016_09
Ki-67: the great leap from simple marker to functional actor | Change of RDF representation of the cross-

News archive

Getting started

Text search
Our basic text search allows you to search all the resources available

UniProt data

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

Protein spotlight

Seeing Red
October 2016

The smallest of things can cause devastation. A word can hurt deeply. Distraction can create an accident. A spark can set off a fire. In the world of

www.uniprot.org/uniprot

Slide 7 of 2 .. 10 Norwegian (Bokmål) (Norway) Notes Comments 102%

Firefox File Edit View History Bookmarks Tools Window Help Thu 10 Nov 22:07 Jon Bråte

Google Calendar - Mo... | Elsevier Editorial SystemTM | Evise | MBV-INFX410 2016 - proje... | Databases_A7.pptx - Intro... | Ensembl genome brow...

www.ensembl.org/index.html

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login/Register

Search all species...

Search: All species for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Favourite genomes

Human GRCh38.p7

Human GRCh37

Mouse GRCm38.p4

Zebrafish GRCz10

Edit favourites

All genomes

Still using Human GRCh37?

Go to GRCh37

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

Compare genes across species

What's New in Ensembl Release 86 (October 2016)

- Mouse Strains
- Chicken new assembly and gene set
- Macaque new assembly and genebuild
- Mouse lemur new assembly and genebuild
- Zebrafish: update to Ensembl-Havana merged gene set

Full details | All web updates, by release | More news on our blog

- 28 Oct 2016: Ensembl genomes 33 is out!
- 17 Oct 2016: Ensembl helpdesk maintenance
- 05 Oct 2016: Ensembl 86 has been released!

Go to Ensembl blog

Tweets by @ensembl

Ensembl @ensembl

Cluster analysis of p53 binding site sequences reveals 3 subsets with different functions #UsingEnsembl @cinjournal

Firefox File Edit View History Bookmarks Tools Window Help Thu 10 Nov 22:09 Jon Bråte

Google Calendar ... | Elsevier Editorial Syst... | Evise | MBV-INFX410 2016 - ... | Databases_A7.pptx - ... | Ensembl genome... | HGNC database ...

www.genenames.org | HGNC

HGNC

HUGO Gene Nomenclature Committee

Search everything | Search symbols, keywords or IDs

Use * to search with a root symbol (eg ZNF*)

Home | Downloads | Gene Families | Tools | Useful links | About | Newsletters | Contact Us | Help | VGNC | Request Symbol

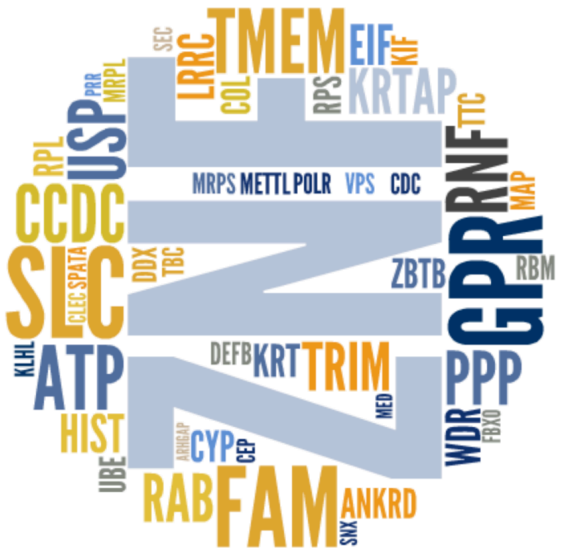
HGNC is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication.

genenames.org is a curated online repository of HGNC-approved gene nomenclature, gene families and associated resources including links to genomic, proteomic and phenotypic information.

Search our catalogue of more than 39,000 symbol reports using our improved search engine (see Search help), search lists of symbols using our Multi-symbol checker and identify possible orthologs using our HCOP tool.

Download our ready-made data files from our Statistics and Downloads page, create your own datasets using either our Custom Downloads tool or BioMart service, or write a script/program utilising our REST service.

Submit your gene symbol and name proposals to us to be accredited with HGNC approved nomenclature for use in publications, databases and presentations.



Firefox File Edit View History Bookmarks Tools Window Help Thu 10 Nov 22:17 Jon Bråte

Google Calendar - Mo... | Elsevier Editorial SystemTM | Evise | MBV-INFX410 2016 - proje... | Databases_A7.pptx - Intro... | 7hpf_2013_47 - SRA - ...

https://www.ncbi.nlm.nih.gov/sra/SRX2226548[accn] ncbi sra

NCBI Resources How To Sign in to NCBI

SRA SRA Search
Advanced Help

NCBI will begin redirecting all HTTP traffic to HTTPS on Thursday, November 10 at 9 AM EST (2 PM UTC). Read more.

Full Send to:

SRX2226548: 7hpf_2013_47
1 ILLUMINA (Illumina HiSeq 2000) run: 80.2M spots, 17.3G bases, 10.2Gb downloads

Design: PolyA selected RNA from developmental time-course

Submitted by: Stowers Institute for Medical Research

Study: Mnemiopsis leidyi Transcriptome or Gene expression
PRJNA344880 • SRP090909 • All experiments • All runs
show Abstract

Sample: Developmental time-course from 1 hour post-fertilization to 9 hours post-fertilization
SAMN05851461 • SRS1730600 • All experiments • All runs
Organism: Mnemiopsis leidyi

Library:
Name: ML_2013_47
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: PAIRED

Runs: 1 run, 80.2M spots, 17.3G bases, 10.2Gb

Run	# of Spots	# of Bases	Size	Published
SRR4374769	80,159,921	17.3G	10.2Gb	2016-10-11

Related information
BioProject
BioSample
Taxonomy

Recent activity
Turn Off Clear
Mnemiopsis AND ("biomol rna"[Properties]) (75) SRA
Mnemiopsis (78) SRA
Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), transcript varia Nucleotide
N-glycosylase/DNA lyase isoform 1a [Homo sapiens] Protein
RefSeq Protein Links for Gene (Select 4968) (16) Protein
See more...

Slide 10 of 2 .. 13 English (United States) Notes Comments 102%

Firefox File Edit View History Bookmarks Tools Window Help Thu 10 Nov 22:17 Jon Bråte

Google Calendar - Mo... | Elsevier Editorial SystemTM | Evise | MBV-INFX410 2016 - proje... | Databases_A7.pptx - Intro... | 7hpf_2013_47 - SRA - ...

https://www.ncbi.nlm.nih.gov/sra/SRX2226548[accn]

NCBI Resources How To Sign in to NCBI

SRA SRA Search
Advanced Help

NCBI will begin redirecting all HTTP traffic to HTTPS on Thursday, November 10 at 9 AM EST (2 PM UTC). Read more.

Full Send to:

SRX2226548: 7hpf_2013_47
1 ILLUMINA (Illumina HiSeq 2000) run: 80.2M spots, 17.3G bases, 10.2Gb downloads

Design: PolyA selected RNA from developmental time-course

Submitted by: Stowers Institute for Medical Research

Study: Mnemiopsis leidyi Transcriptome or Gene expression
PRJNA344880 • SRP090909 • All experiments • All runs
show Abstract

Sample: Developmental time-course from 1 hour post-fertilization to 9 hours post-f
SAMN05851461 • SRS1730600 • All experiments • All runs
Organism: Mnemiopsis leidyi

Library:
Name: ML_2013_47
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: PAIRED

Runs: 1 run, 80.2M spots, 17.3G bases, 10.2Gb

Run	# of Spots	# of Bases	Size	Published
SRR4374769	80,159,921	17.3G	10.2Gb	2016-10-11

Related information
BioProject
BioSample

1. ssh

```
-bash-4.1$ clear
-bash-4.1$ ./fastq-dump SRR4374769
```

Slide 10 of 2 .. 13 English (United States) Notes Comments 102%