

WORKING WITH DATA

Karin Lagesen

karin.lagesen@vetinst.no

Overview

- Available computational resources
- Working with computer clusters
- Figuring out what to do
- Getting help

Size of data can become very big



TBs
of
HTS
data



Abel computer cluster



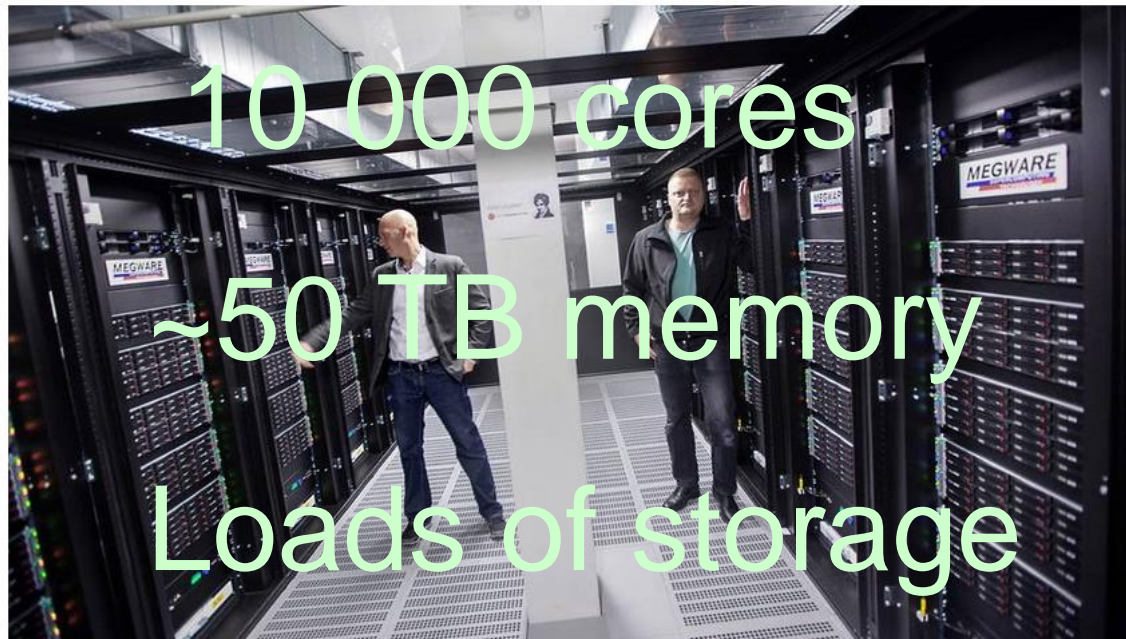
TU T2 TUTV Våre Veger Automatisering TUJobb TUEvent

Søk på tu.no

SØK

BYGG INDUSTRI IT KARRIERE KLIMA KRAFT PETROLEUM SAMFERDSEL FORSKNING

Bedrifts-IT Telekom Sikkerhet



Mer kraft: Thor Dalstedt (t.v.) fra Intel Norge har levert kjerneteknologien i den nye tungregnemaskinen "Abel" som Hans Eide, gruppeleder for vitenskapelig databehandling har driftsansvaret for ved Universitetet i Oslo. Foto: Håkon Jacobsen

SUPERDATAMASKINER I NORGE

www.sigma2.no

The screenshot shows a web browser window displaying the Sigma2 website. The browser's address bar shows the URL <https://www.sigma2.no>. The website's header features the UNINETT logo and the Sigma2 logo, followed by navigation links: Services, Access, Outreach, Support, and About Sigma2. A search bar is located on the right side of the header.

The main content area is divided into several sections:

- Access** (teal background):
 - Apply for computing time
 - Apply for data storage
 - Apply for advanced user support
 - Apply for a user account
 - Change password
 - Log on project leader-GUI
- Data archive** (white background):
 - Norstore Research Data Archive
- Live status** (white background):
 - Notur supercomputers: Live status
- News** (blue background):
 - Hexagon scheduled maintenance**: There will be a maintenance on Hexagon 20 October from 09:00. The plan is to finish by the end of the same day. [Read more »](#)
 - New allocation period**
- New e-infrastructure facilities** (white background):
 - Investment in new e-infrastructure facilities is planned in 2016.
 - [More info](#)
- Twitter** (white background):
 - #einfrastructure** (1 Oct): Heidi Laine (@heidiklaine) retweeted by Dwight Laplage. Mark Parsons from @RDA_US: Research infrastructure is not about what, but about when. #RDABin #openscience #einfrastructure
 - #einfrastructure** (1 Oct): Heidi Laine (@heidiklaine) retweeted by Dwight Laplage. Mark Parsons from @RDA_US: Research

Sigma2

- The Norwegian metacenter for computational science
- Goal: provide a modern national High Performance Computing infrastructure
- Offers HPC services to Norwegian universities, colleges, research institutes and industry
- Funded by RCN, UiO, NTNU, UiB and UiT
- Gain resources by applying for them

Sigma2 services

- High Performance Computing – NOTUR
 - Offers CPU hours on five different clusters at UiB, UiT, NTNU, Iceland Univ, and UiO (Abel)
 - Access to useful software
 - Advanced user support
 - Training
- Data Storage – NORSTORE
 - Project data storage
 - Research data archive
 - Sensitive data storage

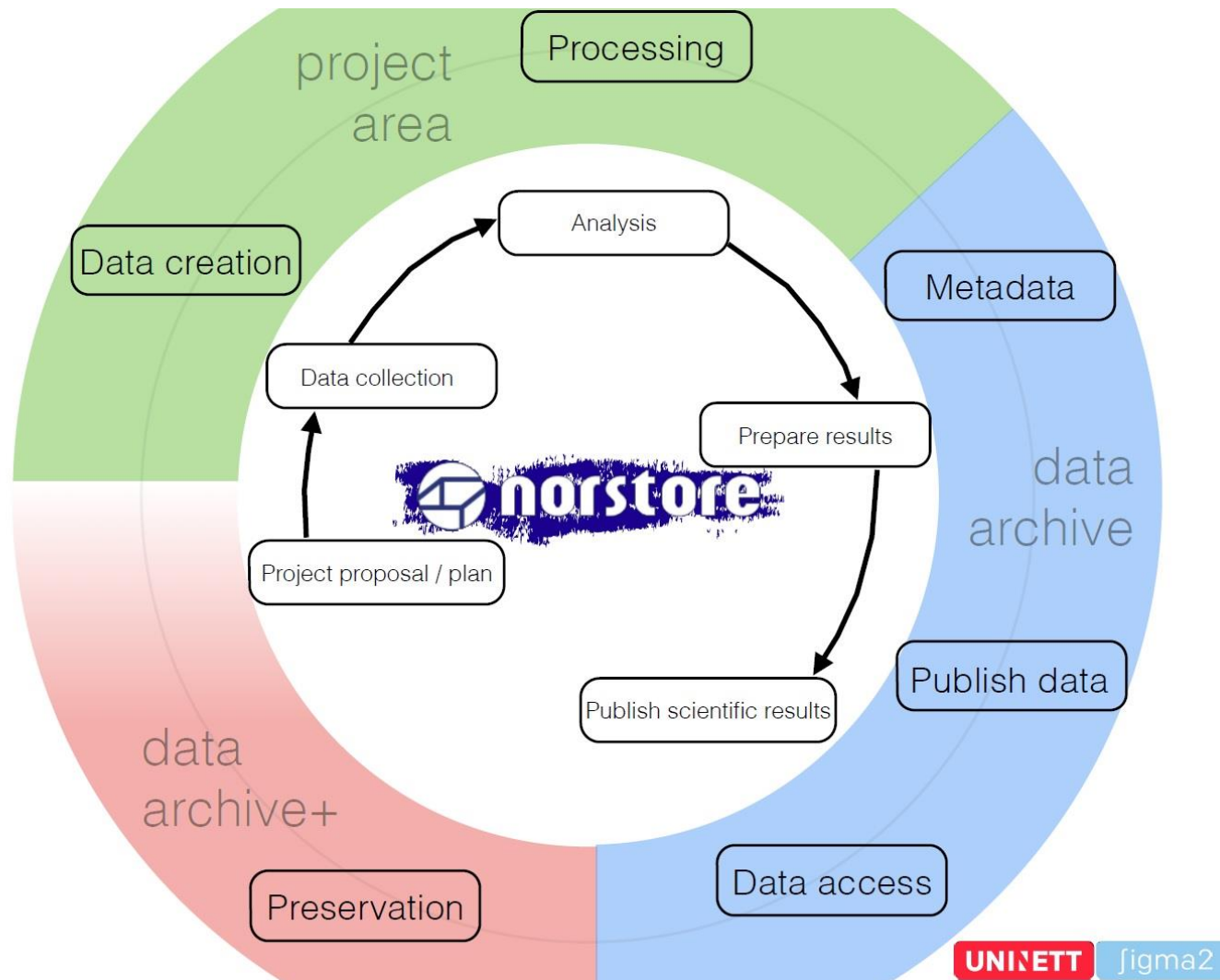
Accessing NOTUR resources

- UiO users have by default access to Abel
 - Access with normal UiO login id
 - Access to UiO CPU hours - ~10% of cluster
 - Additionally: access to freebee.abel.uio.no
 - Freebee can be used for testing purposes
- Larger scale UiO users and all others can apply to NOTUR for access
- **NO SENSITIVE (HUMAN) HTS DATA ON ABEL!**

NORSTORE – data storage

- National infrastructure for management, curation and long-term archiving of digital scientific data
- Can get storage that can be used for both
 - Active computing - /projects
 - Sensitive data
 - Archival storage
- Also, long-time archive storage with possibility for publishing research data

Norstore data storage



Applying for access

- Application deadline Feb/Aug
 - Same application form for storage and CPUs
- Application includes project description, describing what the resources will be used for
- Applications evaluated on scientific merit
- New users/projects given priority
- Main applicant must hold permanent position
- Small projects often given resources at once

The screenshot shows a web browser window with the URL <https://www.metacenter.no/mas/application/project/current/new/>. The browser's address bar and bookmarks are visible at the top. The page has a dark blue header with the Metacenter logo and navigation links. The main content area is white and contains a form titled 'Create new proposal'. The form includes a 'Project Title' field, 'Project Categories' with radio buttons for 'Notur: HPC', 'NorStore: Storage', and 'NorStore: Sensitive Storage', and a 'Permanent position' checkbox. A blue button labeled 'Create new proposal' is at the bottom right of the form. The left sidebar contains links for 'Home', 'Logout', 'Project Leader', 'Projects', 'Applications', 'User accounts', 'Information', 'Changelog', 'Help', and 'About'.

Metacenter
You are logged in as karinlag@uio.no

[Home](#)
[Logout](#)

Project Leader
[Projects](#)
[Applications](#)
[User accounts](#)

Information
[Changelog](#)
[Help](#)
[About](#)

Create new proposal

The Notur and NorStore application forms have been combined, making it possible to apply for both Notur and NorStore at the same time. When not applying for both, the pages not relevant to the application are automatically skipped.

Project Title*

The name of the project

Project Categories*

☐ Notur: HPC

☐ NorStore: Storage

☐ NorStore: Sensitive Storage

Select **Notur** to apply for CPU hours, select **NorStore** for storage resources. If you require **NorStore: Sensitive storage** select this option.

Please note that the sensitive service (TSD), which has been developed with support from Norstore and others, is delivered by partner UiO and use will incur an administration fee which must be covered by the user. Contact tsd-contact@usit.uio.no to inquire about the costs and services offered (you may also visit: <http://www.uio.no/fjenester/it/forskning/sensitiv>).

☐ **Permanent position***

I hereby confirm that I hold a permanent position within my institution, and am [eligible to apply](#).

[Create new proposal](#)

One application form for both NOTUR and NORSTORE

Metacenter - Edit applicati...

https://www.metacenter.no/mas/application/project/V2/edit/notur/2679/project-description-and-funding/

Bookmarks Apps Save to Mendeley CS pythondev Git SWC buy vetinst Ribosomal Database dm sm galaxy BioinfBP straintracer_bg Other Bookmarks

Application

You are logged in as karinlag@uio.no

[Home](#)
[Applications](#)
[Logout](#)

- 1 Project Description and Funding
- 2 Usage and output
- 3 Notur: Software
- 4 Notur: Computing Quota
- 5 Notur: Computation characteristics
- Summary

This application is 0% done.

[Help](#)

Project Application: Testproject

This application has not yet been submitted, it was created on the 2015-10-11.
Fields marked with * are required.

1 Project Description and Funding

Please provide the necessary information regarding your proposed project below. In the second section, please provide information about the funding of the project.

<< Summary Save Next: Usage and output >>

Project title*

Testproject

Project short name

Short name, one word/acronym. It'll be used wherever the full name would be too long, for instance in pie charts

This field is required.
Project description (summary)*

This summary can be used for dissemination purposes. It should be understandable to a wide audience, from research fellows to decision makers. The summary should include objectives, a description of the problem and its scientific and/or socio-economic relevance, the main scientific methods and algorithms used, the challenges in particular with respect to computation and data management, and expected results. Please limit the text to 250 words.

This field is required.
Primary field of science.*

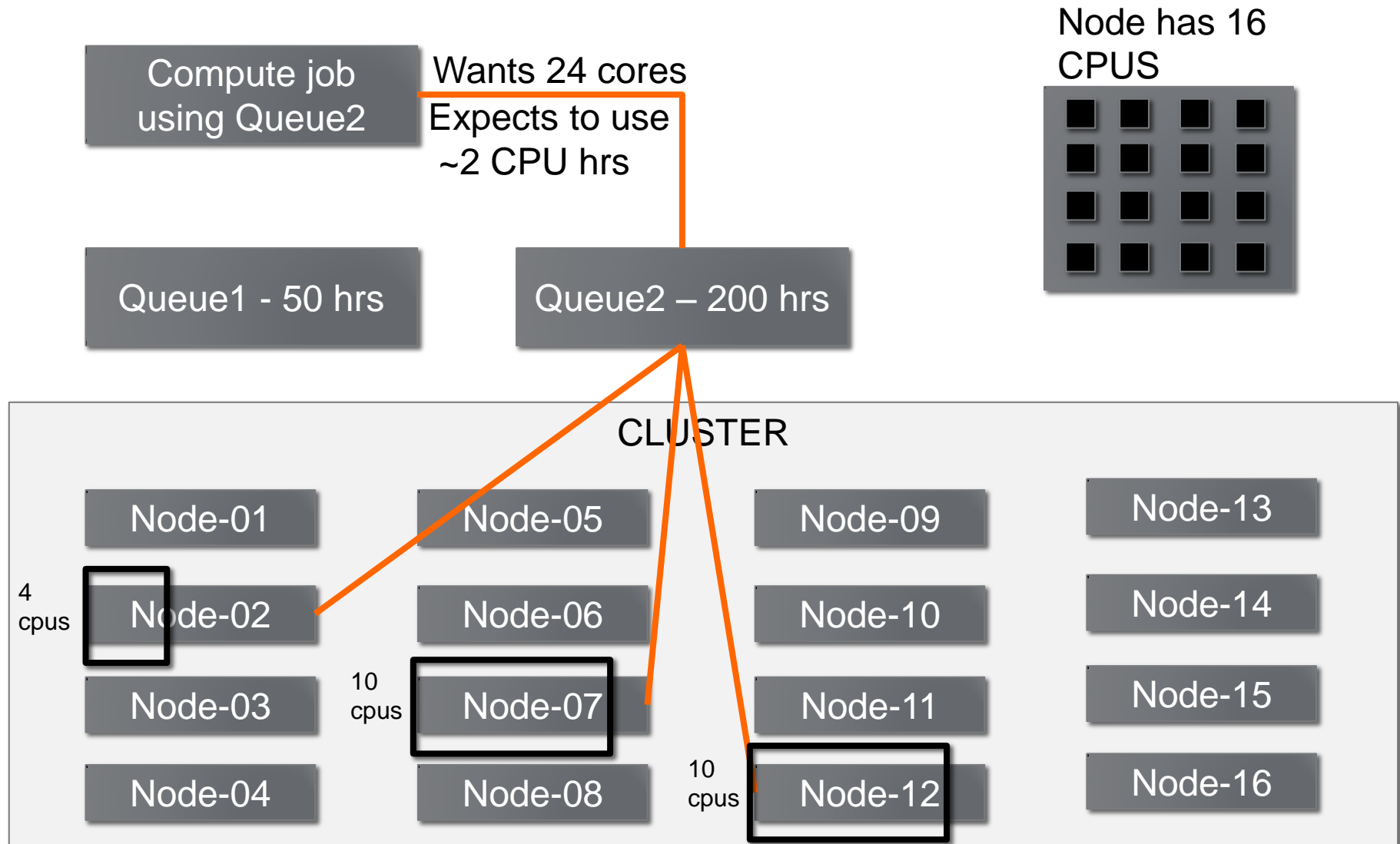
Select the option that best describes the primary field of science of your project.

Secondary field(s) of science.

Using a compute cluster

- Large computer clusters often have queue systems
- Queue systems feed compute jobs to the computer, ensuring that it is optimally used
- Queue system used is named Slurm

Slurm and the compute cluster



Specifying slurm scheduled job

- Need to specify:
 - Estimated time
 - Queue to run in
 - # nodes
 - # cores (cpus)
 - Amount of memory
 - Program(s) to run
- Specifications saved in Slurm job script
- Use command *sbatch* to submit job to slurm

Example slurm script

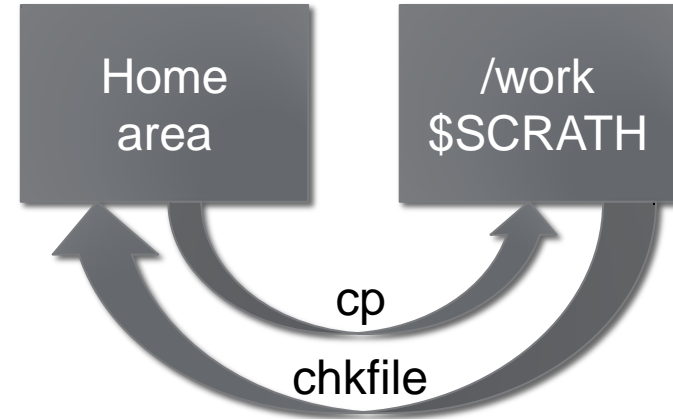
```
#!/bin/bash
#
# Job name:
# SBATCH --job-name=YourJobname
#
# Project:
# SBATCH --account=YourProject
#
# Wall clock limit:
# SBATCH --time=hh:mm:ss
#
# Number of cpus/cores
# SBATCH --ntasks=#of_cpus
#
# Max memory usage:
# SBATCH --mem-per-cpu=Size

## Set up job environment
source /cluster/bin/jobsetup

## Copy input files to the work directory:
cp MyInputFile $SCRATCH

## Make sure the results are copied back to the submit directory:
chkfile MyResultFile

## Do some work:
cd $SCRATCH
YourCommands
```



Directory created in /work with job id – directory alias is \$SCRATCH. All files local to that job are saved there. Should copy job input there to begin with

Using cluster resources

- To use more than 1 CPUs:
 - Need to ask for # CPUs in slurm script
 - `--ntasks - # CPUs requested`
 - Need to specify # CPUs in command to program
 - e.g. `fastqc -t #cpus`
 - Note: make sure `--ntasks` corresponds to # cpus used
- Note: can often run out of memory
 - Per node (16 cpus) – 61.5 Gb of RAM
 - Job will be killed if it uses more than allocated
 - May have to give task upper limit of memory, esp common with java programs

Useful commands

- `sbatch <my_slurm_script>`: put script in queue
- `squeue -u <username>`: list all jobs in queue for a specific user
- `scancel <job id>`: cancel job
- `scancel -u <username>`: cancel all jobs for user
- `cost`: how much of the allocation is used
- `dusage`: how much storage is used

Help on Abel and Slurm

- www.uio.no/tjenester/it/forskning/beregning/abel/

The screenshot shows a web browser window displaying the website for the Abel computing cluster at the University of Oslo. The browser's address bar shows the URL www.uio.no/tjenester/it/forskning/beregning/abel/. The website has a dark header with the UiO logo and navigation links. The main content area is divided into several sections:

- Tjenester og verktøy** (Services and tools): A sidebar menu with links to IT-tjenester, IT-støtte i forskning, Beregninger, and Abel beregningsklynge.
- Abel beregningsklynge**: The main heading for the cluster page.
- Hjelp og rettleiing** (Help and guidance): A section with links to the User Guide, Programvare på Abel, and the FAQ.
- Abel hendingslogg** (Abel incident log): A section titled "Maintenance finished" dated 22. sep. 2015 15:12, stating that the cluster is back in production.
- Treng du hjelp eller råd?** (Do you need help or advice?): A section with links to contact the helpdesk.
- Aktuelt** (Current): A section titled "Abel Newsletter #2, 2015" dated 12. aug. 2015 10:12, announcing CPU allocation and advanced user support deadlines.

Bioinformatics software on abel

- Lots of software available
- Different people need different kinds of software
- Solved this by packaging SW in modules
- > 400 modules available
- Useful commands:
 - *module avail*: lists all available modules
 - *module load modulename*: loads that module
 - *module list*: shows all currently loaded modules
- Use in slurm script: include module load in script

Modules...

```
[karinlag@titan ~]$ module avail
```

```
----- /usr/share/Modules/modulefiles -----
dot      module-git  module-info modules  null      use.own

----- /cluster/etc/modulefiles -----
454apps/1.1.03.24      gaussian/g09b01      ncvview/2.1.2 (default)      protest/3.2 (default)
454apps/2.0.01.02      gaussian/g09c01      netcdf/4.2.1.1 (default)     pypar/2.1.5 (default)
454apps/2.3            gaussian/g09d01 (default) netcdf.gnu/4.2.1.1 (default)  python2/2.7.3 (default)
454apps/2.5.3          gcc/4.7.2            netcdf.intel/4.2.1.1 (default) python2/2.7.6
454apps/2.6            gcc/4.8.0            netcdf.pgi/4.2.1.1 (default)  python3/3.2.3 (default)
454apps/2.7            gcc/4.8.2            newbler/1.1.03.24            python3/3.4.0
454apps/2.8 (default)  gcc/4.9.0            newbler/2.0.01.02            qiime/1.5.0 (default)
454apps/2.9            gcc/4.9.1            newbler/2.3                  qiime/1.8.0
454apps/3.0            gdal/1.9.1 (default)  newbler/2.5.3                quast/2.3 (default)
abyss/1.3.4 (default)  geneid/1.4.4 (default) newbler/2.6                  R/2.15.2
adf/2010.02b           genemark-es/2.3e     newbler/2.7                  R/2.15.2.shlib
adf/2012.01b           genemarks/19032014   newbler/2.8                  R/3.0.2.shlib
adf/2013.01 (default)  geos/3.3.5 (default)  newbler/2.9                  R/3.0.3
adf/2014.01            ghc/7.4.2 (default)  newbler/3.0 (default)        R/3.0.3.profmem
adf_gpu/2014.01        gmap/2013-09-30      nfuse/0.2.1 (default)        R/3.0.3.shlib
allpathsrlg/48777 (default) gmap/2013-11-27 (default) nltk/2.0.1 (default)        R/3.1.0
amber/12 (default)     gnu_parallel/20131022 (default) notur/0.1 (default)          R/3.1.0.profmem
amos/3.1.0 (default)   gnuplot/4.6.0 (default) novocraft/V3.02.05 (default) R/3.1.0.shlib
ampliconnoise/1.25 (default) gnuplot/4.6.3        ocaml/4.00.0 (default)       R/3.1.1 (default)
ampliconnoise/1.29     graphviz/2.28.0 (default) octave/3.6.3 (default)       R/3.1.1.gnu
aragorn/1.2.36 (default) grib_api/1.12.3      open64/5.0 (default)         R/3.1.1.profmem
asreml/2.00ah          gsl/1.15 (default)   openifs/38rlv04              R/3.1.1.shlib
```

How to figure out how to do things

- Read papers that do what you want to do – repeat their methods on your data
- Signup and use:
 - seqanswers.com
 - biostars.org
- Find blogs from people who do what you do
- Read tutorials/manuals on the software you use
- Find email lists to software
- Twitter can also be very helpful

seqanswers.com

The image shows two overlapping browser windows from the seqanswers.com website. The top window displays the forum index, and the bottom window shows a detailed view of the 'Bioinformatics' forum.

Top Window: seqanswers.com/forums/index.php

Navigation: User CP, FAQ, Community

Forum Index:

- Introductions** (3 Viewing) - New here? Stop in and introduce yourself.
- General** (4 Viewing) - Any topic/question that does not fit into the other categories.
- Core Facilities** - Dealing with customer samples, data, etc.
- Sequencing Technologies/Computational** - Platform specific questions, news, discussions.
- 454 Pyrosequencing** (1 Viewing) - Pyrosequencing in picotiter plates, etc.
- Complete Genomics** - Service-based whole human genome sequencing.
- Helicos** - True Single Molecule Sequencing.
- Illumina/Solexa** (7 Viewing) - Bridged amplification & clustering for high-throughput sequencing.
- Ion Torrent** (1 Viewing) - Integrated electronic detection of nucleic acid sequences.
- Oxford Nanopore** - Single molecule sequencing via a nanopore.
- Pacific Biosciences** (2 Viewing) - Single-molecule real-time observation of nucleic acid synthesis.
- Polonator** (1 Viewing) - "Open Source" instrument and sequencing software.

Bottom Window: seqanswers.com/forums/forumdisplay.php?f=18

Forum: Bioinformatics

Thread / Thread Starter	Last Post	Replies	Views
Sticky: ICGC-TCGA DREAM Somatic Mutation Calling Challenge khoulahan	05-28-2014 07:47 AM by karenmartin	7	40,070
Sticky: New Resources for 1000 Genomes (1 2) laura	12-05-2013 11:37 AM by laura	35	26,911
Sticky: SEQwiki den	08-10-2010 08:56 AM by scofield_gao	3	24,992
Sticky: Software packages for next gen sequence analysis (1 2 3 ... Last Page) sci_guy	12-26-2009 03:45 AM by ECO	236	409,379
How to work out coverage? James	Today 05:22 PM by albabu46	2	2,214
E. coli K12 MG1655 datasets from Illumina mido1951	Today 03:44 AM by GenoMax	5	328
read count with multicov amejag	Yesterday 10:17 PM by blanca	1	188
blast2go mapping problem khopper	Yesterday 03:43 PM by khopper	0	143
Denovo transcriptome assembly tellsparck	Yesterday 01:51 PM by Mocca	4	330
Problems with FPKM values priya	Yesterday 03:32 AM by blanca	4	253
from gene coverage to sex determination paumarc	10-09-2015 08:24 PM by doryan	1	173
GO functional annotation of sequences FelipeAd	10-09-2015 08:04 PM by Mocca	1	170
thoughts on using nanopore MinION to DNA sequence for the purpose of genome assembly? jerrybug109	10-09-2015 03:49 PM by gringer	3	545
NGS, coverage and read length appropriate for assembling a genome? (newbie here!) jerrybug109	10-09-2015 02:40 PM by maxsalm	2	234
mauveAligner & progressiveMauve tomiczeek	10-09-2015 11:20 AM by Jspano-Beriman	17	6,452
[DEXSeq] exon counts to "PSI" (exon inclusion level) yerbol	10-09-2015 08:18 AM by DerSeb	1	1,576
Calculating PSI values for rat RNA-seq tomiczeek	10-09-2015 08:13 AM by DerSeb	1	410

biostars.org

The screenshot shows the Biostars.org website interface. The top navigation bar includes links for LATEST, OPEN, RNA-SEQ, CHIP-SEQ, SNP, ASSEMBLY, TUTORIALS, TOOLS, JOBS, FORUM, PLANET, and ALL. The main header features the Biostars logo and a welcome message. A search bar is visible, with a live search dropdown showing results for 'spades'. The left sidebar displays a list of recent posts with their respective vote counts, answer counts, and view counts. The main content area shows a detailed view of a post titled 'A: Alternatives To Newbler To Assemble Ion Torrent Reads?'. The right sidebar contains sections for 'Recent Votes', 'Recent Locations', and 'Recent Awards'.

Live search: start typing...

spades

Recent Votes

- A: How To Translate Encode Genotype Data From Aa/Ab/Bb To Standard A,T,G,C ?
- C: Bioconductor - Error : Function found is not S4 generic
- A: How can merge RNA-Seq biological replicates ?
- A: How can merge RNA-Seq biological replicates ?
- C: Distributed Computing In Bioinformatics
- A: links back from twitter problem
- Reference Assembly - Mapping Reads To A Reference Genome

Recent Locations • All »

- Vilnius, 3 minutes ago
- Stockholm, 3 minutes ago
- Sweden, 4 minutes ago
- France, 4 minutes ago
- Germany, 7 minutes ago
- United States, 8 minutes ago
- United Kingdom, 8 minutes ago

Recent Awards • All »

- Great Question to lh3 ♦ 21k
- Commentator to lh3 ♦ 21k
- Scholar to lh3 ♦ 21k
- Teacher to lh3 ♦ 21k
- Student to William ♦ 2.2k
- Appreciated to Istvan Albert ♦♦ 45k

Limit to: all time

3 votes 1 answer 913 views **How To** **illumina**

0 votes 0 answers 61 views **Illumin** **R** **snp**

2 votes 2 answers 181 views **How c** **rna-seq**

3 votes 1 answer 835 views **Estima** **illumina**

0 votes 0 answers 33 views **News:** **news**

3 votes 1 answer 87 views **Biocon** **softwar**

1 vote 1 answer 157 views **Forum** **twitter**

0 votes 0 answers 148 views **Bioinfo** **(plants**

0 votes 0 answers 272 views **how to** **stitch** **software**

A: Alternatives To Newbler To Assemble Ion Torrent Reads?
I have been using SPAdes <http://bioinf.spbau.ru/spades> without problems. by JC

Doing genome assembly using SPAdes with mate-pair library only on public Galaxy server
Doing genome assembly using SPAdes with mate-pair library only on public Galaxy server Dear all, I need Illumina mate-pair reads (from bacterial sample) using SPAdes, and I'm thinking to use public Galaxy server install SPAdes myself. However, when I used the only public Galaxy server where I can find SPAdes, it gave After searching in SPAdes website, I concluded that the error came since the SPAdes version that I used server was 3.0.0, while the manual states that "SPAdes should not be used if only mate-pairs [...] are by sentausa

A: Smaller Assembled Genome Size Than Expected
I suggest trying spades assembler. It permits the usage of multiple kmers, merging all kmers into a by akoik063

A: No Bowtie Alignments
By default in SPAdes there is an error correction step before the assembly. If you have used the original by gyulap01

A: Assembly of whole metagenomics data
for me in this type of scenario. Also IDBA-UD and SPAdes have given consistently good results. by Mikael Huss

A: software for constructing contigs and scaffold
few out there, MIRA, Newbler, SOAPdenovo, Velvet, Spades Here is a big list: http://en.wikipedia.org/w/index.php?title=List_of_genome_assemblers by akoik063

How To Merge Contigs From Two Different Assembler :
genome assembly. I did my Denovo assembly with SPAdes(SPAdes: A New Genome Assembly Algorithm and Its Applications by HG

No Bowtie Alignments
assembled a number of contigs from Illumina reads using SPAdes. I have been trying to assess the depth of coverage by hollynns

Traffic: 501 users visited in the last hour

Reading blogs

bedtools Tutorial

Aaron Quinlan November 22, 2013

TABLE OF CONTENTS

- Synopsis
- Setup
- What are these files?
- The bedtools help
- bedtools "intersect"
- Default behavior
- Reporting the original feature in each file.
- How many base pairs of overlap were there?
- Counting the number of overlapping features.
- Find features that DO NOT overlap
- Require a minimal fraction of overlap.
- bedtools "merge"
- Input must be sorted
- Count the number of overlapping intervals.
- Merging features that are close to one another.
- bedtools "complement"
- bedtools "genomecov"
- Producing BEDGRAPH output
- Sophistication through chaining multiple bedtools
- Principal component analysis
- A Jaccard statistic for all 400 pairwise comparisons.

Synopsis

Our goal is to work through examples that demonstrate how to explore, process and manipulate genomic interval files (e.g., BED, VCF, BAM) with the [bedtools](#) software package.

Some of our analysis will be based upon the Maurano et al exploration of DnaseI hypersensitivity sites in hundreds of primary tissue types.

Maurano et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012. Vol. 33

www.sciencemag.org/content/337/6099/1190.short

This tutorial is merely meant as an introduction to whet your appetite. There are many, many more tools and options than presented here. We therefore encourage you to read the bedtools [documentation](#).

Setup

From the Terminal, create a new directory on your Desktop called "bedtools-demo".

```
cd ~/Desktop
mkdir bedtools-demo
```

Navigate into that directory.

```
cd bedtools-demo
```

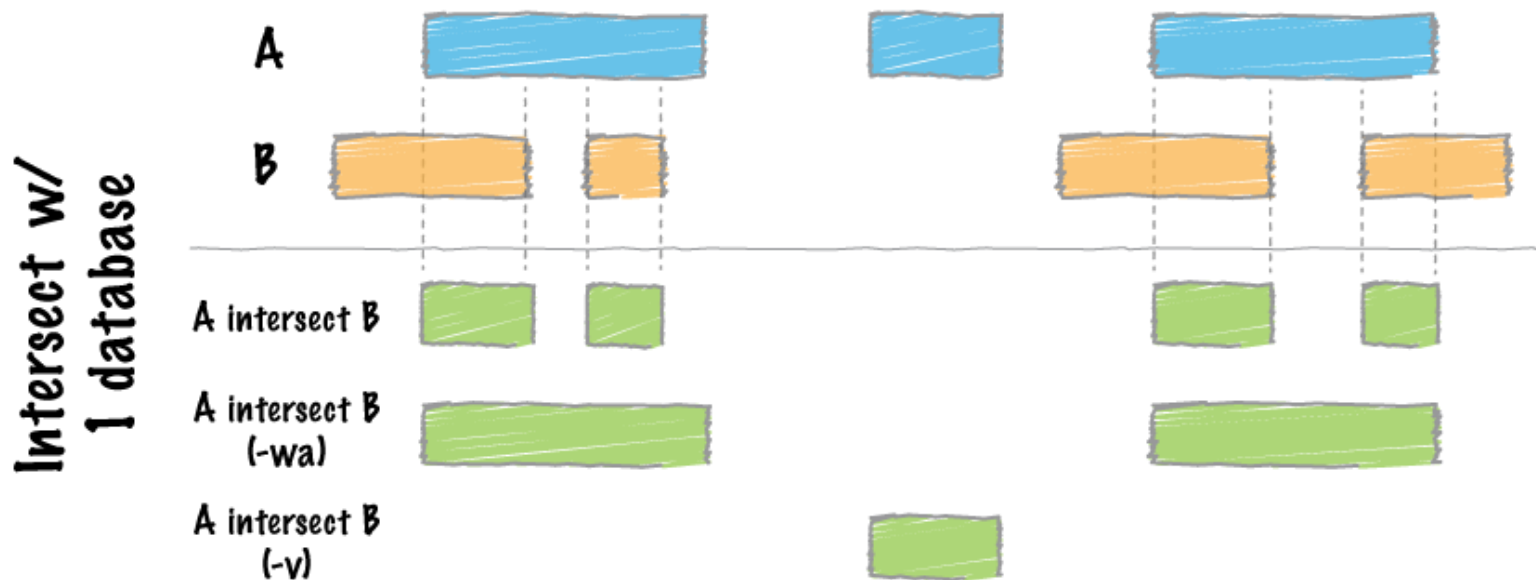
Download the sample BED files I have provided.

```
curl -O http://quinlanlab.cs.virginia.edu/cshl2013/maurano.dnaseI.tgz
curl -O http://quinlanlab.cs.virginia.edu/cshl2013/cpg.bed
curl -O http://quinlanlab.cs.virginia.edu/cshl2013/exons.bed
curl -O http://quinlanlab.cs.virginia.edu/cshl2013/gwas.bed
curl -O http://quinlanlab.cs.virginia.edu/cshl2013/genome.txt
```

Now, we need to extract all of the 20 Dnase I hypersensitivity BED files from the "tarball" named [maurano.dnaseI.tgz](#).

Connecting...

Reading blogs

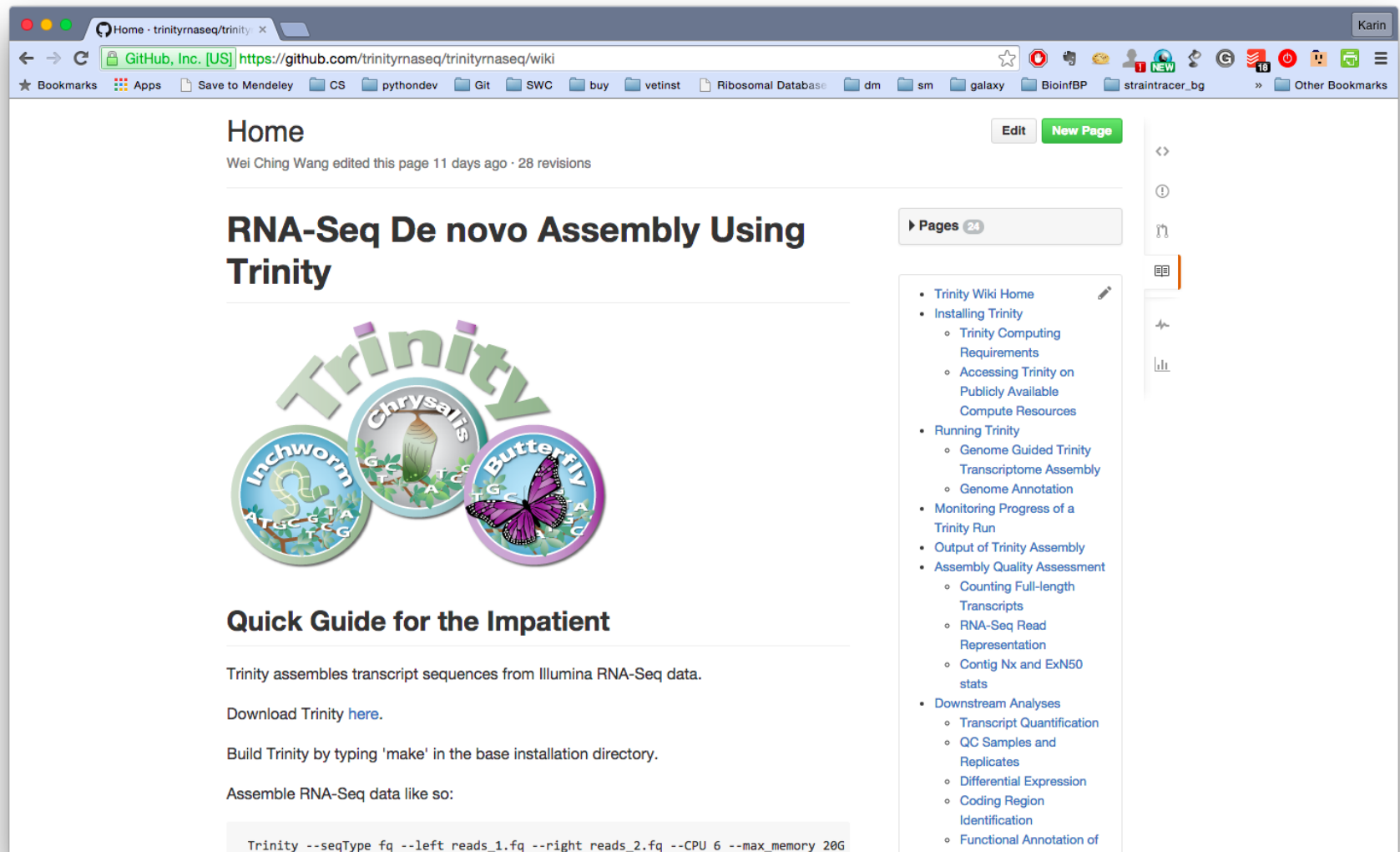


Default behavior

By default, `intersect` reports the intervals that represent overlaps between your two files. To demonstrate, let's identify all of the CpG islands that overlap exons.

```
bedtools intersect -a cpg.bed -b exons.bed | head -5
chr1 29320 29370 CpG:_116
chr1 135124 135563 CpG:_30
chr1 327790 328229 CpG:_29
chr1 327790 328229 CpG:_29
chr1 327790 328229 CpG:_29
```

Find tutorials and manuals for sw




The screenshot shows a web browser window displaying the Trinity RNA-Seq De novo Assembly Wiki page. The page title is "RNA-Seq De novo Assembly Using Trinity". The page is edited by Wei Ching Wang, 11 days ago, with 28 revisions. The page features a large graphic of the Trinity logo, which consists of three overlapping circles labeled "Inchworm", "Chrysalis", and "Butterfly", each containing a different biological illustration. Below the logo, there is a section titled "Quick Guide for the Impatient" which provides instructions on how to use Trinity. The instructions include downloading Trinity, building it, and assembling RNA-Seq data. A terminal command is shown at the bottom of the page: `Trinity --seqType fq --left reads_1.fq --right reads_2.fq --CPU 6 --max_memory 20G`. On the right side of the page, there is a sidebar with a list of pages, including "Trinity Wiki Home", "Installing Trinity", "Running Trinity", "Monitoring Progress of a Trinity Run", "Output of Trinity Assembly", "Assembly Quality Assessment", and "Downstream Analyses".

Home

Wei Ching Wang edited this page 11 days ago · 28 revisions

RNA-Seq De novo Assembly Using Trinity



Quick Guide for the Impatient

Trinity assembles transcript sequences from Illumina RNA-Seq data.

Download Trinity [here](#).

Build Trinity by typing 'make' in the base installation directory.

Assemble RNA-Seq data like so:

```
Trinity --seqType fq --left reads_1.fq --right reads_2.fq --CPU 6 --max_memory 20G
```

- Trinity Wiki Home
- Installing Trinity
 - Trinity Computing Requirements
 - Accessing Trinity on Publicly Available Compute Resources
- Running Trinity
 - Genome Guided Trinity Transcriptome Assembly
 - Genome Annotation
- Monitoring Progress of a Trinity Run
- Output of Trinity Assembly
- Assembly Quality Assessment
 - Counting Full-length Transcripts
 - RNA-Seq Read Representation
 - Contig Nx and ExN50 stats
- Downstream Analyses
 - Transcript Quantification
 - QC Samples and Replicates
 - Differential Expression
 - Coding Region Identification
 - Functional Annotation of

What to do if you get an error?

- Try with a different data set – often good to try with a smaller, well-known data set
- Change version of program if another one are available
- Google is your friend – google error message
- Check seqanswers and biostars
- Look at the webpage for the software – is your error mentioned?
- Check email list for software – is error mentioned?
- Write to software authors and/or sw email list, have they seen this before?
- Also – if on Abel/TSD: email the helpdesk

What to include in an error report

(0) What is my environment?

1. What did I do?

2. What result did I expect?

3. What result did I get?

(4) Why is this incorrect?

Error report - translated

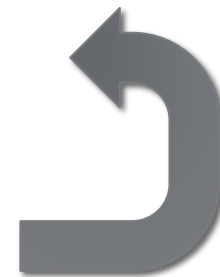
- (Shortly) explain purpose of analysis
 - Name of program, incl. version
 - Full command line, incl. all options
 - Copy-paste of error from start of program
 - For USIT: include file system location
-
- Goal: help person should be able to recreate the bug, without having to ask you more questions

Finding your way

- Many ways to the same goal
 - But: which one is “right”?
- Top trick: keep your biological question in mind
 - Software often designed to solve one particular biological problem
 - Keep your biological expectations in mind – if result breaks with that, stop and rethink
- Figure out expectations by using a “known” entity
 - either a small data set or
 - data set you already know the biological answer for

The bioinformatical process

- Working with bioinformatics *is* experimental work
- Not all software are appropriate for
 - Your biological problem
 - Your data
- Need to experiment to find the right combination of software for your problem
- Bioinformatics is an iterative process:
 - Try program on data
 - Evaluate result
 - Decide if results are appropriate



**Redo
several
times**

Log your work

- Bioinformatics is an iterative process
 - May not remember what solution you ended up with, or for that matter, why
- Keep a log with:
 - Today's date
 - What you are doing
 - Program version
 - Commandline
 - A little bit of output
 - Evaluation of output

Questions?