CEES bioinformatics – yearly report for 2012

Introduction

Bioinformatics – "an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data" (Wikipedia) is growing at CEES. An increasing number of researchers are generating large 'digital' datasets that need to be analyzed using sophisticated bioinformatics tools. We use bioinformatics in a wide sense here, so that the following types of researchers are examples of bioinformaticians at CEES:

- all users of 'R'
- those doing statistical modeling
- those analyzing next generation sequencing data, be it genomics, transcriptomics, or other such data
- those working with time series data
- those extensively using the unix command-line, programming in perl, python, C, etc.

With this report we aim to provide, for the first time, an overview of the bioinformatic activities and facilities at the CEES.

People

In March 2012, a survey was undertaken among all researchers at CEES to investigate the number of people considering themselves bioinformaticians, the data types and programs they work with, and their needs. 45 people responded to the survey, a summary of which was sent around and is attached to this report.

Infrastructure

Strategic considerations

Many researchers are still able to perform their analysis on their desktop/laptop machine. However, increasingly, we need to perform tasks that are either CPU or memory intensive.

At CEES, we use a combination of self-owned servers, and CPU hours we applied for on the UiO supercomputer 'Abel' (previously called 'Titan'). This maximizes flexibility for CEES researchers in choosing the right resource for their project:

- memory-intensive applications can be run on our own servers
- CPU-intensive applications can be submitted to Abel and therefore do not take up valuable time on the servers

The servers CEES owns (see below) are attached to the Abel system. This means users can seamlessly access the same programs and disks on the self-owned servers, as well as on Abel.

For storage ('project disk space) we rent space from USIT at UiO (attached to Abel), rather than buy and administer our own. The benefit of this strategy is that we do not have to spend valuable research time on basic system administrative tasks and software installation. The HPC (high-performance computing) group of USIT is very proficient in installing programs with difficult dependencies or requirements (some of their contribution we obtain as a paid-for service, see below). Finally, backup of the data is arranged for as well.

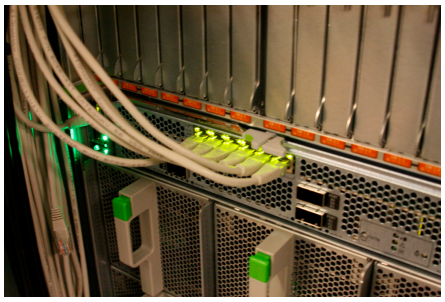
Hardware

Starting with the project to sequence and assemble the genome of Atlantic cod in 2009, CEES has invested in its own hardware for computation. These servers are hosted and maintained by the HPC group of USIT. As of December 31, 2012, the following computational infrastructure is available to the CEES:

- two high-memory servers with 24 CPUs and 128 GB of RAM, and around 1 TB disk space each ('cod1' and 'cod2', bought in 2009)
- two high-memory servers with 64 CPUs and 512GB of RAM, and around 24 TB disk space each ('cod3' and 'cod4', bought in 2011)

The following resources are rented or allocated to CEES:

- on the University computer cluster ('Titan', called 'Abel' as of September 2012) we have two allocations for CPU-intensive computations. One has been applied for by the cod group (2.5 million CPU hours), one by Anne Maria Eikeset (several 100 000 CPU hours)
- we rent 15 TB of project disk space (administrated by USIT, includes backup). The costs of this storage (5500 NOK/TB/year) are shared between users, proportionally to the used space
- we have, shared with the Norwegian Sequencing Centre, 10 TB disk space for long-term archival of data at norstore, the national Norwegian infrastructure for the management, curation and long-term archiving of digital scientific data



Investments in 2012

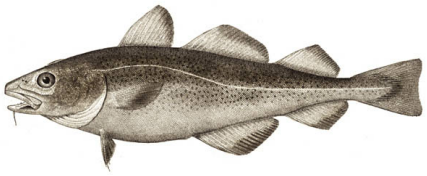
- two times 20 TB hard disk space was bought and added to cod3 and 4
- the HPC (high-performance computing) group at USIT has been 'hired' through an advanced user support allocation for projects such as installation of software packages and other extraordinary system administration tasks

Software

There are a few project specific applications available through the 'cod' servers:

- the program 'Stacks' for analysis of RAD-tag sequence data
- mysql servers for databases
- the web portal ('SMRTportal') for analysis of data from the Pacific Biosciences instrument

Administration


Atlantic cod (source: Wikipedia)

The day-to-day administration of the servers and disk space is the responsibility of USIT. However, there is still a considerable overhead for CEES staff:

- correspondence with USIT on required software, interruptions of with the servers, feedback to CEES users
- keeping an eye on the disk space allocation (when the maximum is reached, new files cannot be written anymore without warning)
- communications with the users necessary for a smooth running of the shared resource (e.g., asking users to clean up disk space)
- administration of the user base, mailing lists etc.
- instructing new users, who often are new to the field

Projects

Examples of projects requiring large computational resources and large amounts of disk space are:

- the project to generate an improved version of the Atlantic cod genome: both memory and CPU-intensive analyses, several TB of disk space
- several other genome sequencing projects
- the RAD-seq platform at CEES (SNP detection and genotyping by sequencing): computations through the Stacks software on the cod1 server
- transcriptomics analysis pipelines for species with an draft reference genome

Mailing lists

We started a mailing list, cees-bioinf@bio.uio.no (35 subscribers), for information exchange. In addition, a mailing list, cees-hpc@bio.uio.no (15 subscribers), was started for users of the computational hardware to smoothen the process of sharing the resource between us.

Meetings

The bioinformaticians at CEES had one common meeting where the survey results were discussed, and one other meeting to discuss a paper called *A Quick Guide for Developing Effective Bioinformatics Programming Skills* (http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000589).

Wiki

We started a UiO wiki (https://wiki.uio.no/mn/bio/cees-bioinf). We are collecting articles dealing with the practicalities of using the resources at CEES, tips and trick, etc. The wiki is open to the world.

Courses

Continuing the success of previous such courses (Unix and perl), the following internal courses were organized in 2012:

- the local IT department at the Institute was recruited to give an introductory course in using the unix command line for CEES researchers; 10 people attended (May 2012)
- Karin Lagesen gave a course in python programming for 12 people (June 2012)
- on initiative from CEES bioinformaticians, Software Carpentry (http://software-carpentry.org/) gave a so-called Bootcamp at UiO (September 2012). From the website: 'Software Carpentry helps researchers be more productive by teaching them basic computing skills. We run boot camps at dozens of sites around the world, and also provide open access material online for self-paced instruction. The benefits are more reliable results and higher productivity: a day a week is common, and a ten-fold improvement isn't rare.' The bootcamp was open for all researches and around 15 people attended, including three CEES-bioinformaticians

Outlook 2013

Bioinformatics at CEES is undoubtedly going to grow in 2013. The following can already be said regarding the next year:

- the two oldest servers (cod1 and cod2) expire in August 2013. We are looking into possibilities for, and costs associated with, extending the service contract for these servers for another year. Losing these servers would mean a significant reduction in computational resources at CEES
- we intend to apply for one common CEES allocation of CPU hours on Abel
- we are already running low on empty disk space and will have to increase with several TB to accommodate new datasets
- two of us (Karin Lagesen and Lex Nederbragt) will give a new Software Carpentry Bootcamp at UiO in July. In order for preparing for this event, we intend to try out teaching the material to CEES bioinformaticians during spring 2013
- the AquaGenome project will kick-off in 2013, and is poised to generate huge amounts of data, requiring significant computational resources for the analyses of these

The Pacific Biosciences RS, source of new sequencing data to be analysed using HPC rsources

Blindern, March 14th 2013
Lex Nederbragt, with help from many others.

CEES-bioinformatics survey 2012

Survey conducted in March 2012

Sent out to CEES-all list

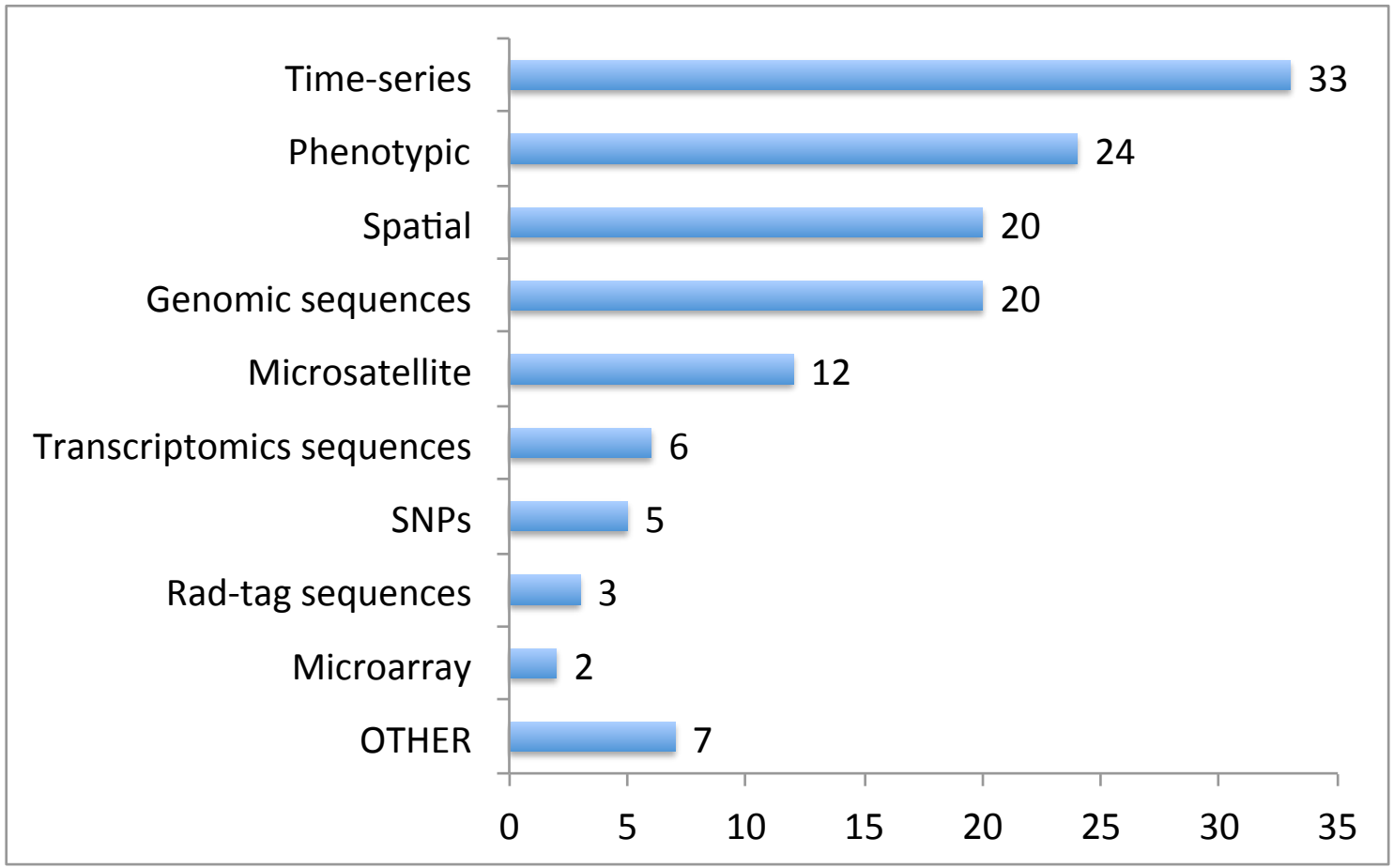
48 responses

Thanks to all who responded!

Now follows a set of response summaries

Note: often multiple answers given per question

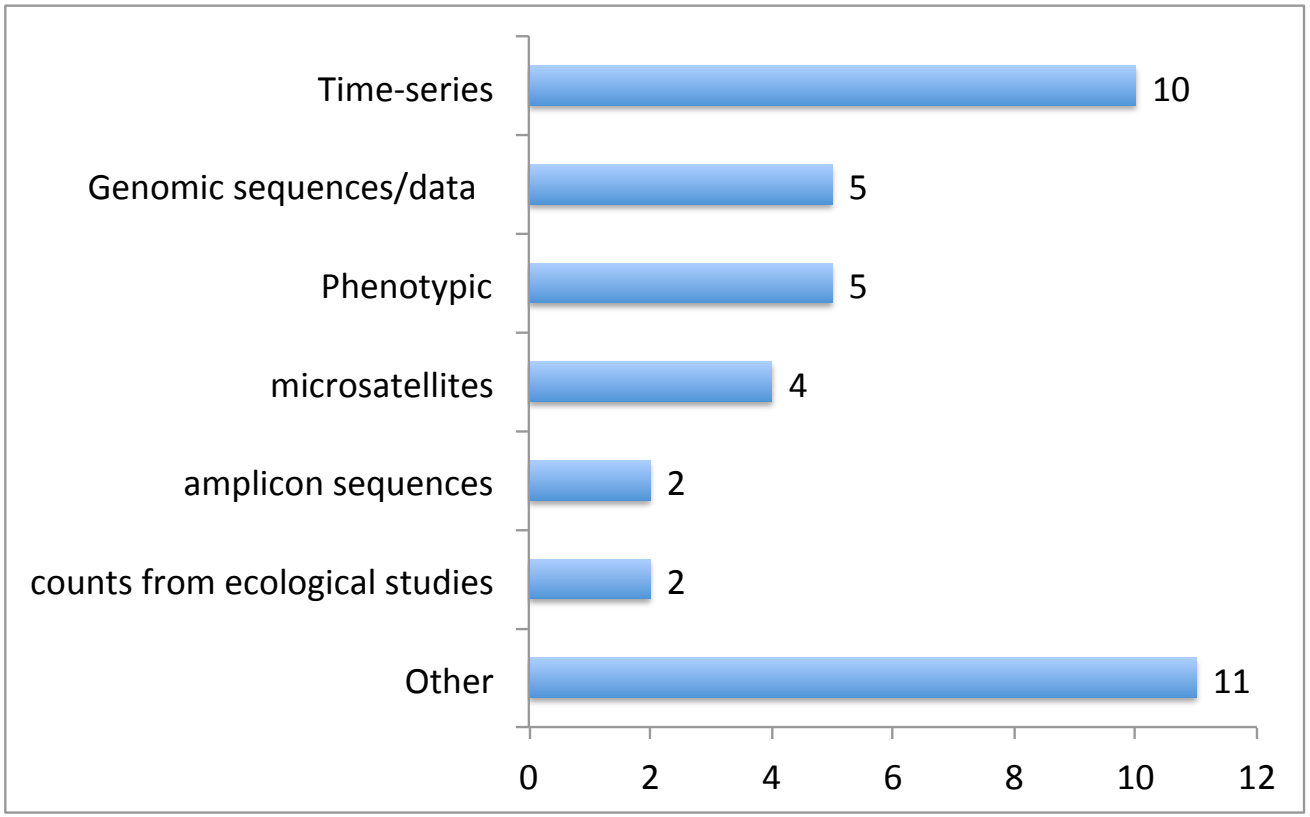
What kind of data do you use in your research?



OTHER

- spatiotemporal (space +time dim's) simulated
- georeferenced (as required by GIS)
- counts/densities of organisms
- Spectroscopy
- Demographic
- Amplicon sequences

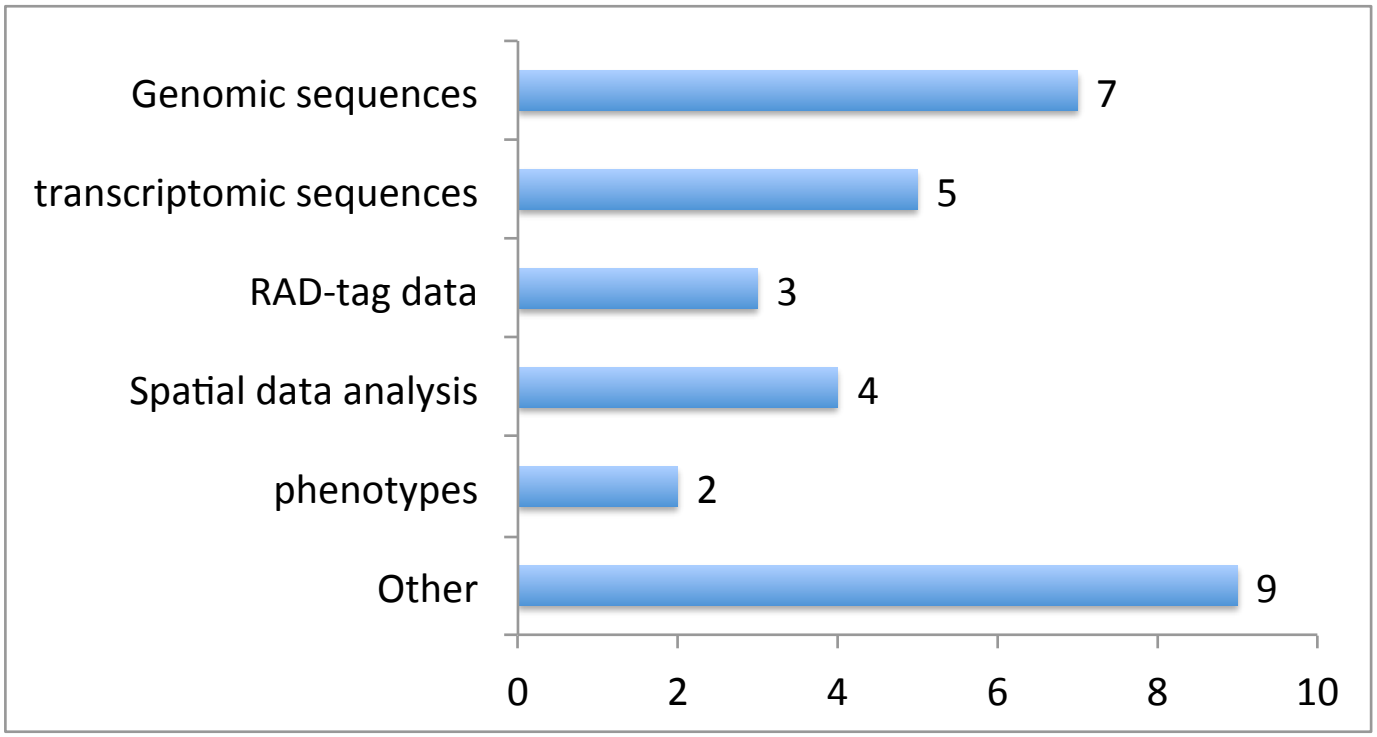
What is the one data type that you consider yourself most familiar with and that you would be willing to help others with?



OTHER

- simulated
- mtDNA
- morphometric data (shape analysis)
- spatial
- Space/space-time data
- Demographic
- Climate data
- epidemiological data
- All type of sequence work/analysis of smaller datasets
- feeling pretty unsuccesfull right now in combining these, so none :-).
- I am still in learning process. I rather need a help then feeling confidant to help others.

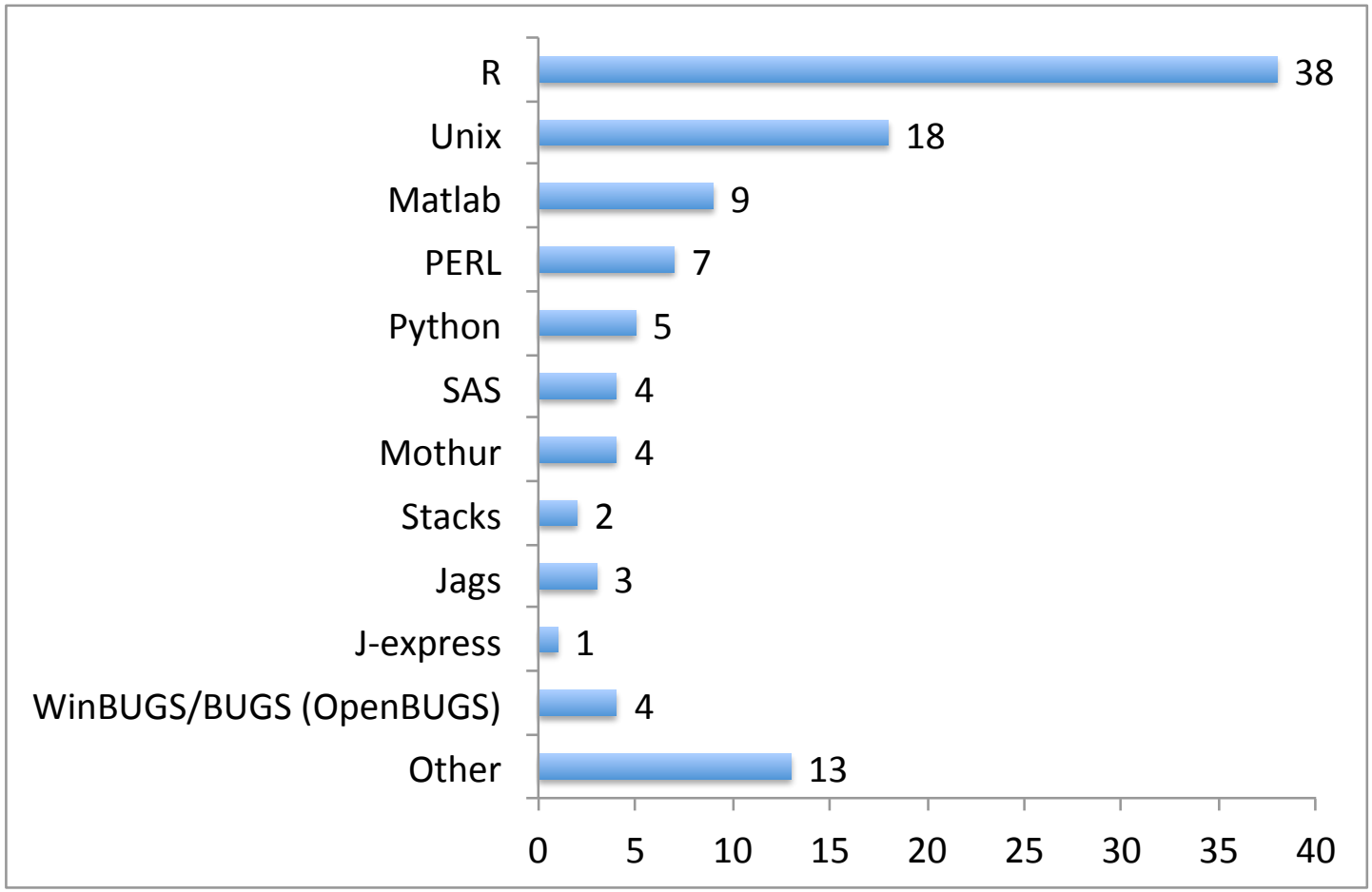
What kind of data do you most desperately need to know more about?



OTHER

- perl
- microsatellites, snps
- how to combine spatial spread with genetic changes
- count
- Handling and analysis of large datasets
- Funk gens
- Capture-mark-recapture data
- The other stuff
- I'm not desperate yet :-)

What kind of software/programming tools do you use on your data?

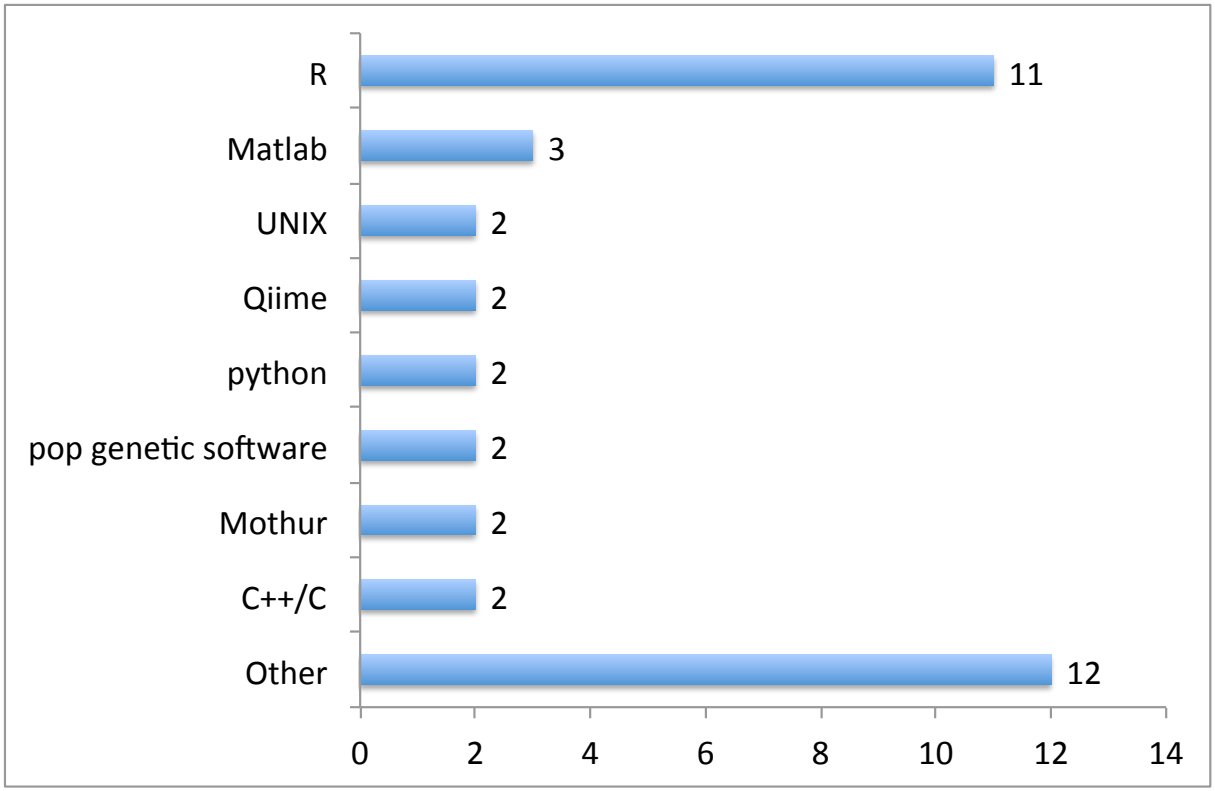


OTHER

- stat packages on microsat data
- population genetics and phylogeography dedicated softwares
- Mathematica
- PULP
- Network nc
- Mega
- JMP
- Genexel
- Fastq related and transcriptome assemblers
- Ecopath w Ecosim
- C++
- C
- anything Emiliano said :)

What is the one software/programming tool that you consider yourself most familiar with

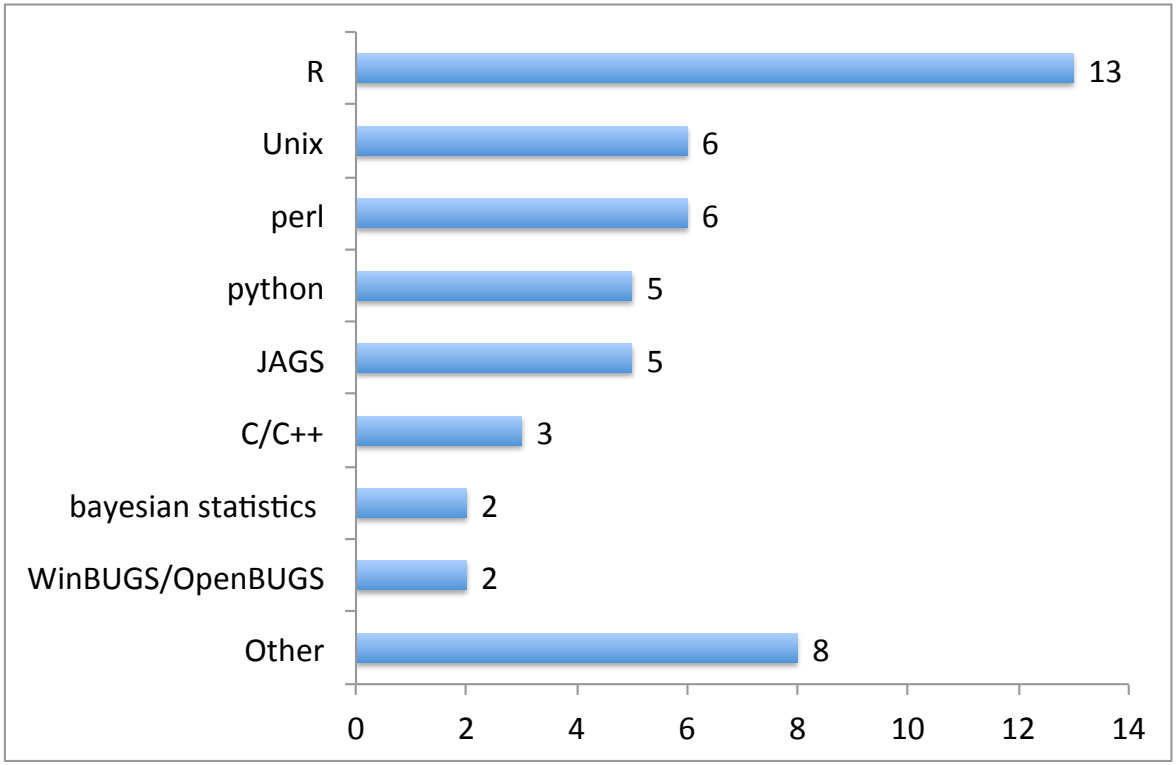
and that you would be willing to help others with?



OTHER

- microsat stats
- matlab
- clojure (lisp)
- Stacks
- phylogeography softs
- SAS
- PERL
- MorphoJ (morphometrics)
- Matlab, MatCad
- MEGAN
- JMP
- Ecopath w Ecosim

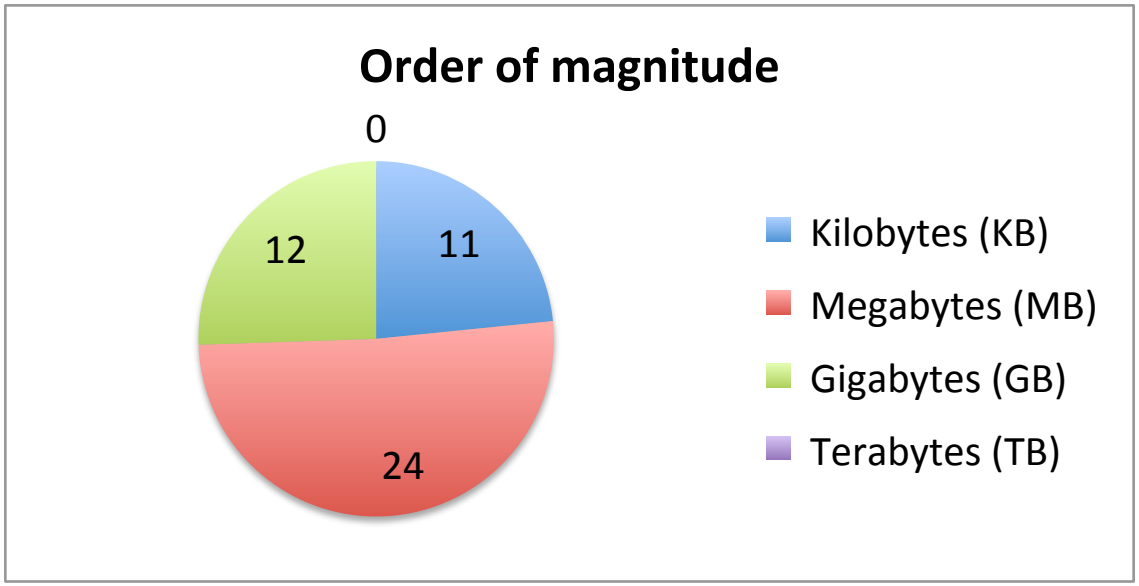
What kind of software/programming tool do you most desperately need to know more about?

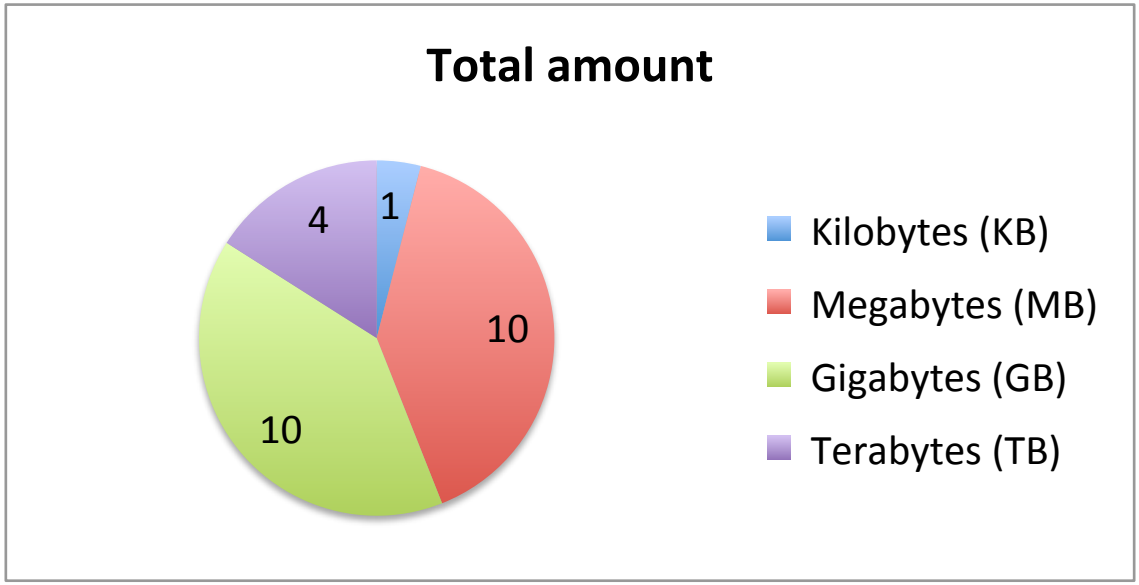


OTHER

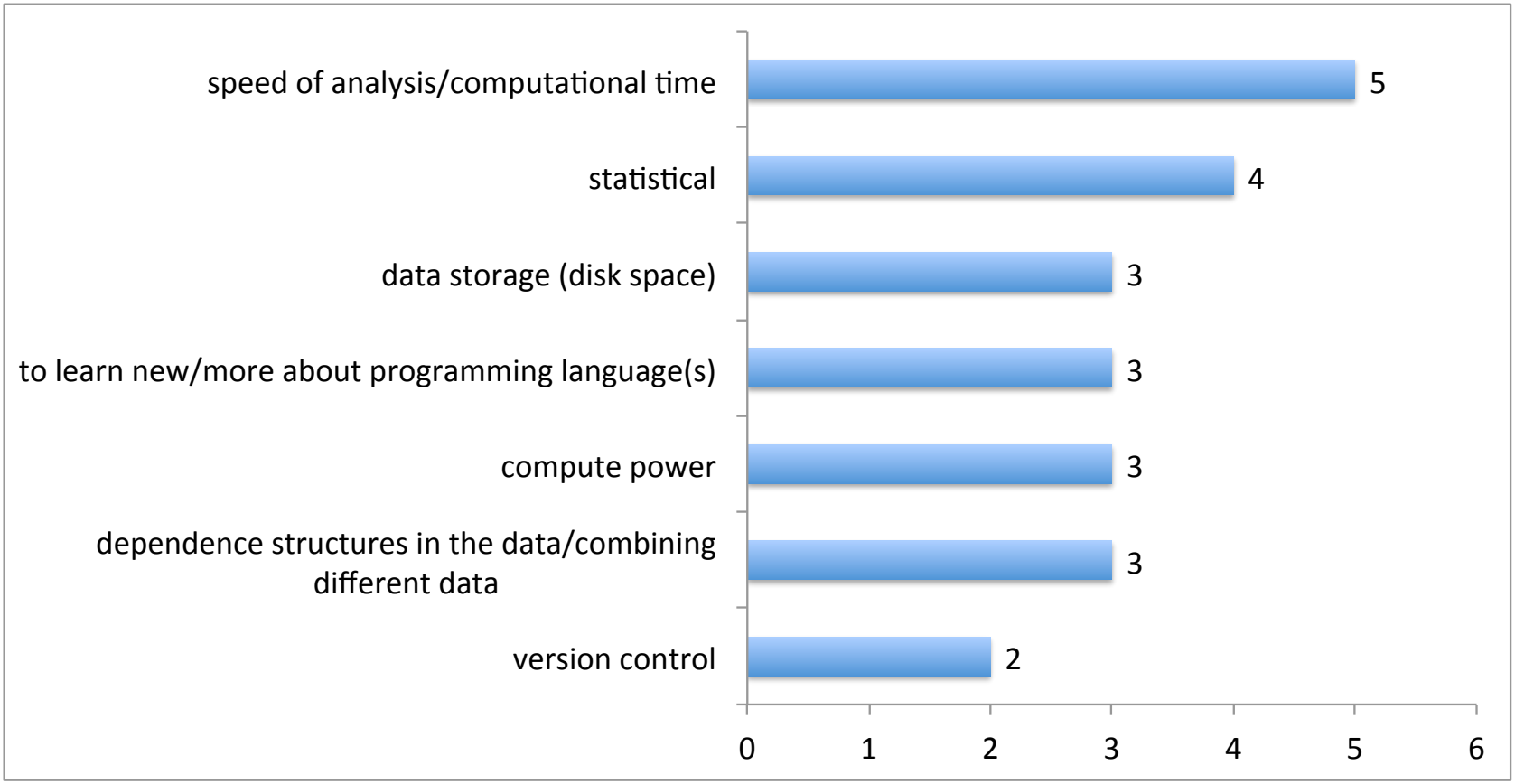
- Unsure
- statistical/genetical software
- software for SNPs and genomic sequences
- version control
- Transcriptome assemblers (de novo)
- GRASS
- Not entirely desperate, but seems I should learn how to access a computer cluster (e.g. Titan)
- ARC GIS

Data sizes

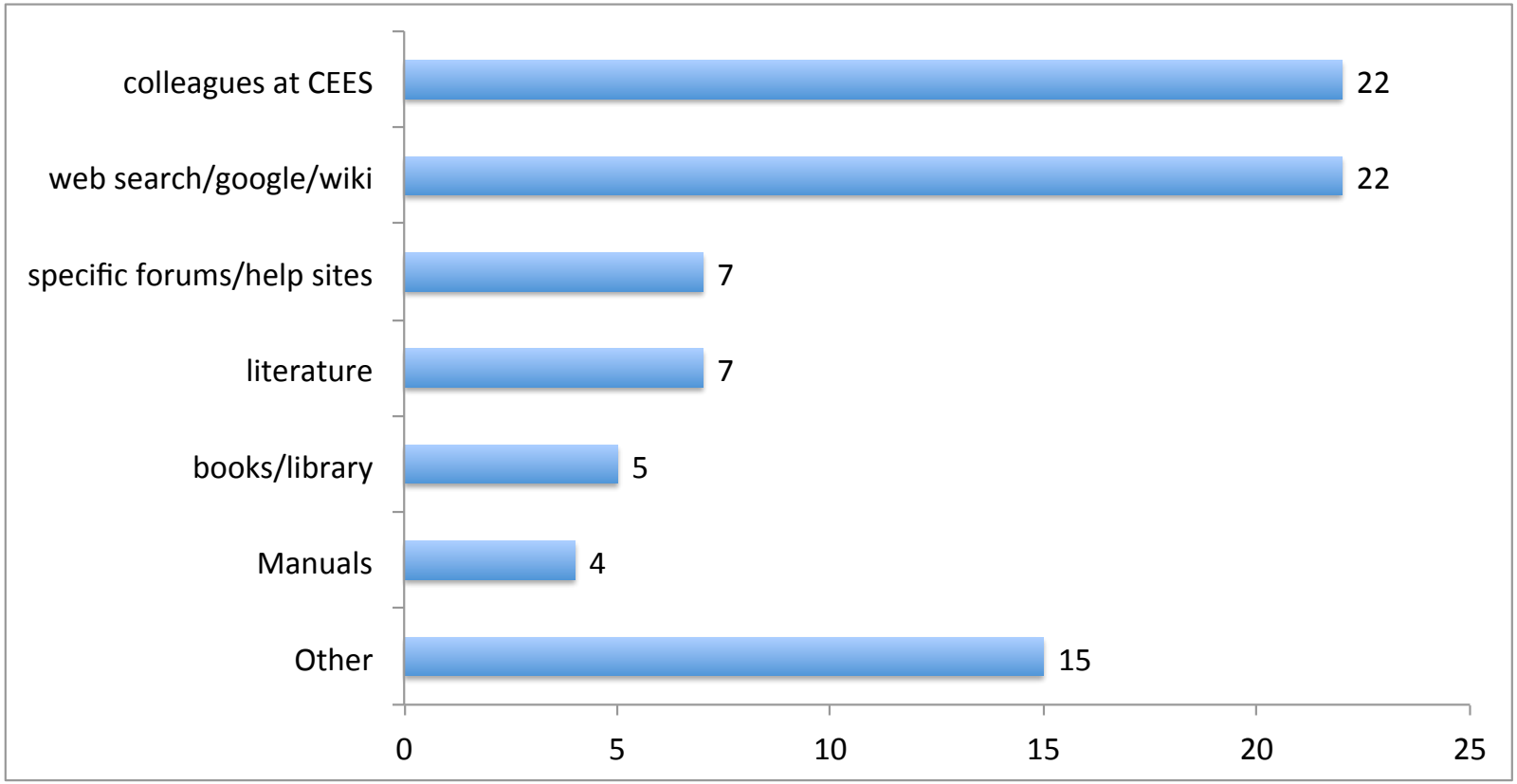

Order of magnitude
0
11
12
24
Kilobytes (KB)
Megabytes (MB)
Gigabytes (GB)
Terabytes (TB)


Total amount
1
10
10
4
Kilobytes (KB)
Megabytes (MB)
Gigabytes (GB)
Terabytes (TB)

What is the main difficulty you face when working with your data?



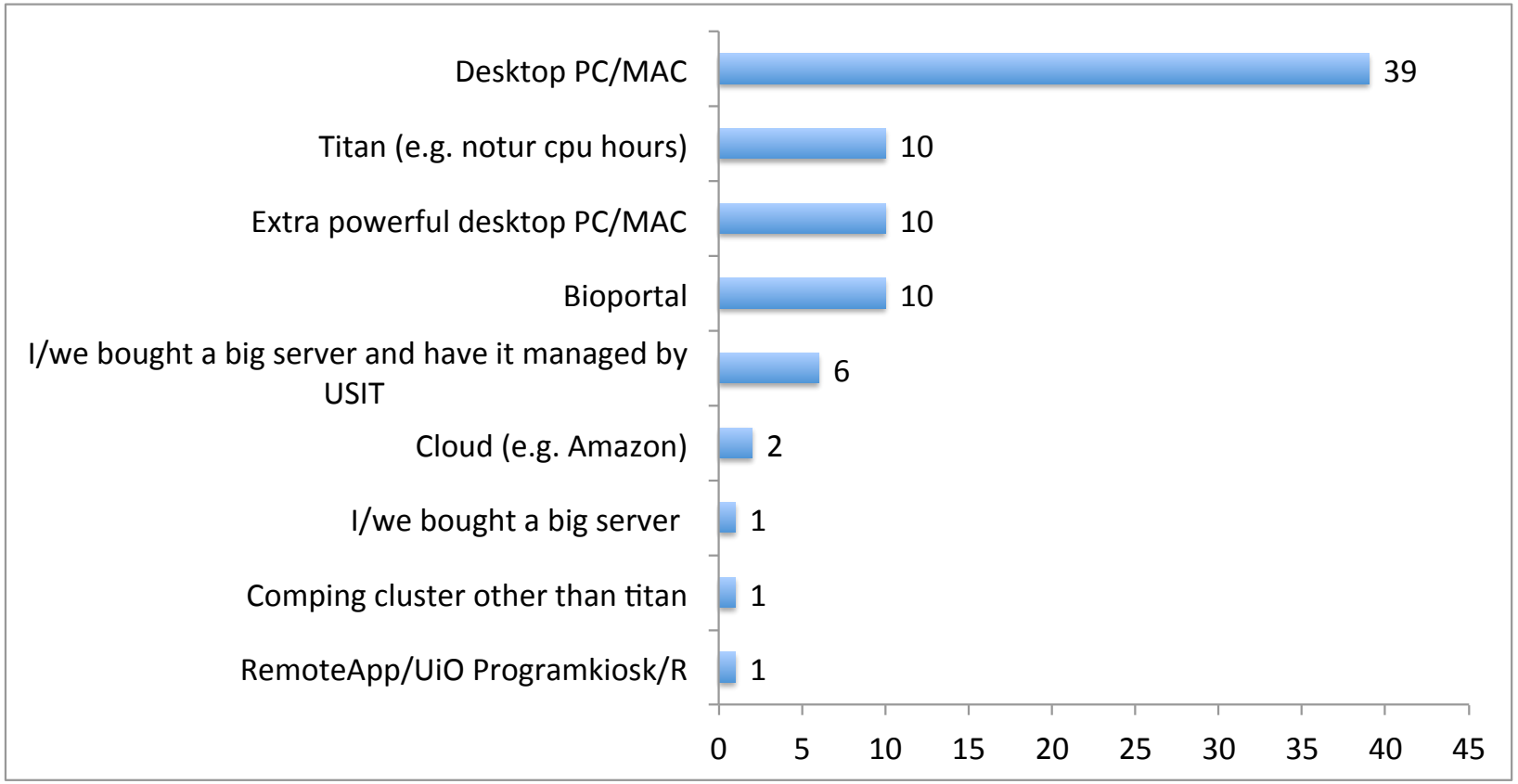
Where do you usually go to find answers on how to work with your data?



OTHER

- To Titan-guys.
- Supervisor
- Statisticians
- R-packages
- R-help
- People I know from elsewhere who use these tools
- Have no data yet
- just walking around and getting new ideas
- e-mail exchange with authors of programs.
- develop own software
- courses
- competent people I know
- but there is lots of program out there....
- Yes

What kind of resources do you use for working with the above data and tools?



Conclusions

- Time series data is the most widely used type, followed by phenotypical and spatial data
- There is a big need in training on handling sequencing data (genomic/transcriptomic)
- R is by far the most-used programming language. It is also the tool most people need to know more about
- There is also a need for training in UNIX and programming languages (perl/python)

Conclusions

- Most people work with data in the megabyte range, although several report having datasets that in total reach the terabyte range
- Computational time is a big bottleneck, as well statistics
- For help, CEES researchers mostly use each other, and the internet

Recommendations

- There is a need for training on aspects of bioinformatics such as R, UNIX, handling of sequencing data, programming

- There is enough knowledge at CEES to consider doing such training with 'local' teachers

Recommendations

- The R users at CEES could start an R user group to see how they can benefit from each other's knowledge

- There is a need for speeding up analysis, e.g. by parallelization, code optimization, and increased computational resources

Thanks

To all CEES members that responded

Karin Lagesen & Lex Nederbragt, March/April 2012