Bioinformatics Workshop

ALLBIO BIOINFORMATICS

"Evaluation of Test Cases"

12 - 14 December 2012
Milan, Italy

Bioinformatics Workshop "Evaluation of Test Cases"

12 – 14 December 2012
Milan, Italy

Programme

12 December 2012

- 13:00 – 14:00 Registration and Lunch
- 14:00 – 14:30 Welcome by the chair Erik Bongcom-Rudloff and the local organizer Andreas Gisel
- 14:30 – 18:00 Presentation of the 15 selected test cases – each 10 plus some questions

1. **FAIRE-seq data analyser** – Amin Omidbakhshfard (University of Potsdam, Germany)
2. **Identification of large structural variations** – TBD
3. **Discovery of gene regulatory networks** – Nooshin Omranian (Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany)
4. **Primer design for ChIP-PCR studies** – Prashant Garapati (Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany)
5. **Integrative Omics analysis workflow** Karl Göran Andersson (Swedish University for Agricultural Sciences, Uppsala, Sweden)
6. **Bacterial Comparative Genomics** – Laurent Falquet Swiss Institute of Bioinformatics, Lausanne, Switzerland)
7. **Processing an un-annotated genome** – TBD
8. **Discover non-coding RNA in yeast** – TBD
9. **Pathway analysis of a poorly annotated but sequenced plant genome** – Erik Alexandersson (Swedish University for Agricultural Sciences, Uppsala, Sweden)
10. **Gene prediction modelling for plants** – TBD
11. **Genome assembly of heterozygous diploid data** –TBD
12. **Functional annotation of the potato genome** – Erik Alexandersson (Swedish University for Agricultural Sciences, Uppsala, Sweden)
13. **Transcriptome and small RNA comparison of four phenotypically distinct cv "Primitivo" grapevine clones** – Annalisa Giampetruzzi (Institute of Plant Virology CNR, Bari, Italy)
14. **Improved cell factories** – Mikko Arvas (VTT Technical Research Centre of Finland, Espoo, Finland)
15. **Genetics, Genomics and Evolution of prolific breeds of domestic sheep (Ovis aries)** – Juha Kantanen (MTT Agrifood Research Finland, Jokioinen, Finland)

- 16:20 Coffee Break
- 18:00 – 20:00 Guided visit at Cenacolo di Leonardo

13 December 2012

➡ 09:00 – 11:30 Topic wise parallel sessions to discuss the test cases

We will divide the work in 4 parallel groups, one for each of the following topics:

- data formats / standardization and workflows
- scalability of tools (multi genome comparison/data compression)
- non model organisms and multiploidy
- visualization and training
-

The 4 groups will analyze (according to the topic) each test case to identify

- the difficulties or the current limitations affecting the test case
- the existing tools required (with or without adaptation)
- the missing tools required
- keep minutes of discussion
- prepare a summary for discussion

➡ 10:30 Coffee Break

➡ 11:30 – 13:00 Reporting of the different sessions and general discussions

Each group will report in 15min the results of their session

- is there a common trait in several test cases?
- are there common tools required or missing?
- can we split test cases according to operations?
- can we group test cases?

➡ 13:00 – 14:00 Lunch

➡ 14:00 – 15:00 Presentations of relevant data analysis tools – speaker to be defined

- Oren Tzfadia (Weizmann Institute of Science, Tel Aviv, Israel) – The MORPH algorithm: ranking candidate genes for membership in Arabidopsis thaliana and Solanum lycopersicum pathways
- Dimitris Vlachakis (Biomedical Research Foundation, Academy of Athens, Greece) – Recipes to make the most out of your wetlab data in less than 10 minutes.
- Jacques van Helden (Université Libre de Bruxelles, Bruxelles, Belgium) – A quick tour of Regulatory Sequence Analysis Tools (RSAT) and Network Analysis Tools (NeAT)
- Tatyana Goldberg (Technische Universität, München, Germany) – Loctree2 predicts localization for all domains of life
- Domenica D'Elia (Institute for Biomedical Tachnology, Bari, Italy) and Gianvito Pio (University of Bari, Italy) – HOCCLU2: a data mining tool for easily handling interactions data and discovering regulatory networks
- Bruno Contreras Moreira (CSCI, Zaragoza, Spain) – Introducing get_homologues and footprintDB
- Petri Klemelä (IT Center for Science, Helsinki, Finland) – Chipster: user-friendly analysis software for high-throughput data
- Paul Kersey (European Bioinformatics Institute, Hinxton, GB) – Ensembl beyond humans: genomic resources for animals, plants and microbes
- Mikko Arvas (VTT Technical Research Centre of Finland) – Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species

- Marco Masseroli (Politecnico di Milano, Milan, Italy) – Genomic and Proteomic Knowledge Base (GPKB) & Bio Search Computing (Bio-SeCo)

➡ 15:00 – 18:00 Parallell session for test cases specific discussions in light of the results of the topics-wise discussions

We will divide again the work in 3 parallel sessions. Each session will analyze 5 test cases (random distribution).

For each test case we need:

- identification of tools
- a SWOT analysis (Strength, Weaknesses, Opportunities and Threats)
- a preliminary work plan for a follow-up activity
- propose other collaborators for the follow-up activity
- keep minutes of discussion

➡ 16:20 Coffee Break
➡ 20:00 Dinner at the Hotel

14 December 2012

➡ 09:00 – 11:00 Cross-presentation of the results from the test case discussions

Selection of test cases (or operations) in 3 categories and discussion of the work plan

- relatively easy, workflow exists or can be adapted within weeks
- more difficult, some tools are missing, workflows are partial, need months to adapt
- very difficult, scalability problem, tools completely missing, requires new developments (projects over years)

➡ 10:30 Coffee Break
➡ 11:00 – 13:00 Organization of the follow-up events to approach the solutions of some test cases

- a) create work groups to prepare the hack-a-thon events
- b) decide number of events and divide test cases or operations, plan a soon solution of one test case
- c) collect emails and skype accounts
- d) decide upon video meetings schedule and chairperson
- e) keep minutes of discussion and decisions

➡ 13:00 – 14:00 Lunch and end of the workshop

TEST CASE 1

Test Case Title
FAIRE-seq data analyser

Test Case Acronyme: FAIRE-seq
fasq

Test Case Class
Plants

Contact person
Bernd Mueller-Roeber, University of Potsdam

Contact
nd

Test Case Description
FAIRE is a procedure allowing the isolation of nucleosome-depleted regions from eukaryotic chromatin. FAIRE is based on the fact that different chromosomal regions are cross-linked to formaldehyde with different efficiencies, depending on the openness of the chromatin and the proteins binding to it. It has been used by researchers to identify differences in chromatin structures in different cell types (e.g. healthy vs. sick cells); FAIRE has also been employed to discover gene regulatory elements (e.g. motifs in promoters) genome-wide. Such regions of the chromatin can be straightforwardly isolated using a simple biochemical phenol-chloroform extraction procedure after formaldehyde-mediated crosslinking of histones to chromosomal DNA and random fragmentation of the DNA by sonication. The extracted genomic DNA fragments are generally free of histones and represent regions containing regulatory elements (to which transcription factors bind). Such genome segments can be analyzed by quantitative PCR, hybridization to genome tiling arrays or deep sequencing. Based on high resolution and decreasing the costs of deep sequencing, usually FAIRE-seq is using nowadays. But analysing FAIRE-seq data and finding the FAIRE peaks among the genome (which show e.g. regulatory elements) is time consuming and need different steps and is non trivial procedure.

Background knowledge
Gene expression is regulated at different levels but one of the first and most important events includes changes in chromatin structure, which can be identified using FAIRE-seq. FAIRE-seq experiments are starting to become more and more important for the identification of (long-range) gene regulatory elements; currently no tools for facile handling of such data is available.

Actors
nd

Initial state of the Test case
We established FAIRE-seq experiments on Arabidopsis thaliana as a model plant, started collaboration with Riano-Pachon´s group on data analysis.

Desired final state of the Test Case
The tool should allow an experimentalist to efficiently identify FAIRE peaks at a genome-wide scale. Also, the tool should allow to link FAIRE peaks to gene expression patterns, to discover peaks associated with transcriptional changes (stimulus-dependent, developmentally regulated, etc.)

Test Case Work Plan
Different tools can be used to align FAIRE-seq data to the genome like “bowtie” and also peak calling can be done by “MACS” or “ZINBA”.

1 Pre-processing the data. 2 Identify reads which will be aligned to the nuclear genome. 3 Identification of FAIRE peaks which will present open chromatin parts of the genome. 4 Find close genes to FAIRE peaks. 5 Doing motif analysis. 6 Identification of FAIRE peak sequences.

Discussion
Simon J.M. et al. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. 2012. Nature protocol. 7, 256-267.

TEST CASE 2

Test Case Title
Identification of large structural variations

Test Case Acronyme
NGSGAPS

Test Case Class
Plants (Animals)

Contact person
nd

Contact
nd

Test Case Description
Genome sequencing of closely related individuals has yielded valuable insights that link genome evolution to phenotypic variations. However, advancement in sequencing technology has also led to an escalation in the number of poor quality-drafted genomes assembled based on reference genomes that can have highly divergent or haplotypic regions.Especially reads on the Illumina Genome Analyzer produce millions of short reads that are rather difficult to assemble. Initiatives like the Arabidopsis 1001 genome projects has so far revealed mainly single nucleotide polymorphisms but fail to uncover large structural variations from the reference genome sequence. So far iterative mapping is a potential way to resolve the current problems, however, the time-consuming nature of this approach and requirement for manual adjustment makes it an slow process for large-scale projects.

Background knowledge
nd

Actors
nd

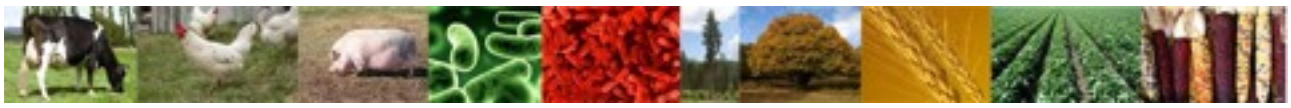
Initial state of the Test case
Currently more than 100 Arabidopsis accessions have been sequenced, so data is available in databases and currently no need for the generation of new data, unless the current data has not enough quality.

Desired final state of the Test Case
Identify the large structural variations within the genomes of the different Arabidopsis accessions with the current available sequence data. Furthermore a web-resource covering this data should become publically available.

Test Case Work Plan
Project should focus on the generation of software that first of all allows for a rapid assessment of the quality of the sequence data. New ways for mapping short reads to a reference genome should be designed, allowing for more flexible gap options, variable assembling parameters that rely on the local context of the genomic region and iterative read mapping.

Discussion

LF: identifying large structural variations with unfinished genomes is rather difficult. What is possible is to remap the paired-end or better the mate-pair reads and use that information. There are a number of tools able to do that: breakdancer, Hydra, GASV, Tigra-SV, Delly. Perhaps a general assessment of these tools should be a first step of the project. We could invite the developers of those tools. Ken Chen, Aaron Quinlan, gasv@cs.brown.edu, Tobias Rausch from EMBL would probably the best.



TEST CASE 3

Test Case Title
Discovery of gene regulatory networks

Test Case Acronyme
GRNs

Test Case Class
Plants (Animals)

Contact person
Nooshin Omranian

Contact
nd

Test Case Description
In higher plants, gene expression of the 20.000 to > 50.000 genes is regulated by 2.000 to 4.000 transcription factors (TFs). Although the functions of several TFs have been discovered, the target genes of TFs and hence the gene regulatory networks they control is unknown for the vast majority of them. Currently, more and more ChIP-Seq data are collected from Arabidopsis thaliana and very likely crops in the near future as well. Other experimental data are collected as well, from EMSA experiments, binding site selection assays, transactivation assays, inducible expression of TFs followed by microarray hybridization or RNAseq to identify genes affected by the TFs.

Background knowledge
Currently, it is very time-consuming to collect and combine the various data needed to unravel the GRNs of transcription factors. Experimentalists are collecting information from diverse sources and then try to combine the extracted information to establish the GRN.

Actors
nd

Initial state of the Test case
There are various tools available allow for coexpression analysis, motif ifentification (e.g. Patmatch), the extraction of promoters from different plant specie (e.g. phytozome), phylogenetic footprinting, cis element databases, etc. However, it is extremely time-consuming to transfer data from A to B and extract the information that is needed to establish a GRN.

- CORNET. https://cornet.psb.ugent.be/main/precalc
- ATTED-II . http://atted.jp/top_search.shtml#CoExSearch
- CressExpress. http://cressexpress.org/
- CSB.DB http://csbdb.mpimp-golm.mpg.de/
- Patmatch in TAIR
- AGRIS, TransFac, PLACE for the identification of known cis elements in promoters of interest (e.g. of coexpressed genes).
- MEME for the identification of motifs overrepresented in a set of promoters.

Drawbacks: The main drawback is that it is extremely time-consuming to combine the different data into one single scheme that displays the GRN.

Desired final state of the Test Case
We need user-friendly tools that enables the efficient reconstruction of GRNs from experimental and computational data. For example, I want to know the most likely direct target genes of the given TF, starting out from genes that (at their expression level) respond to a change in the activity of a given TF.

Test Case Work Plan
Protocol:

1. Perform expression analysis (Affy, Agilent, RNAseq) after activation of a TF
2. Identify responding genes
3. Extract their promoter sequences
4. Identify known cis elements in the promoters of these genes
5. Identify novel motifs in the promoters of these genes
6. Check whether binding sites of the TF under analysis is present in the promoters of the responding genes; at which position, in which number? On which strand of the DNA? Etc.
7. Compare the promoters of these genes with promoters of orthologous genes from other plant species (crops!)
8. Use phylogenetic information to cross-check for cis elements / motifs.
9. Include expression data from other experiments in Arabidopsis and other species
10. Suggest a GRN in a good schematic presentation that highlights various aspects of the interaction of the TF with its downstream target genes (e.g. cis elements in the promoters to which it might bind; expression induction or repression curve of the target gene);

Discussion
LF: Possible person to invite: Chris J Needham C.Needham@leeds.ac.uk (seehttp://www.biomedcentral.com/1752-0509/3/85)

TEST CASE 4

Test Case Title
Primer design for ChIP-PCR studies

Test Case Acronyme
ChIP-qPCR

Test Case Class
Plants (Animals)

Contact person
Prashant Garapati

Contact
nd

Test Case Description
ChIP-seq and ChIP-exo studies will increasingly lead to the identification of potential target genes of transcription factors. Often, the binding of transcription factors to individual target genes under a number of different experimental conditions (treatments, developmental stages, mutants, natural variants) needs to be tested. ChIP-PCR is the method of choice for such experiments.

Background knowledge
Selecting primers suitable for qPCR experiments on promoter fragments is often tricky and the current primer design tools are not well established in that respect. In contrast to primers for qRT-PCR (for expression analysis) primers for ChIP-qPCR should anneal around the transcription factor binding site to allow detection / quantification of the genomic fragments bound to the TF.nd

Initial state of the Test case
Tools exist for primer design on coding sequences in large scale (e.g. QuantPrime), but they have not yet been adopted for large-scale primer design for ChIP-PCR experiments.

Desired final state of the Test Case
A user-friendly tool is needed for primer design for multiple (parallel) ChIP-PCR experiments.

Test Case Work Plan
Protocol:

1. Promoter sequences (1 kb, 3 kb, longer, depending on species) of genes of interest are extracted from databases; the promoters (genes) may be selected based on coexpression with a TF, based on differential expression between a wild type and a (TF) mutant, based on induced / repressed expression after activation of a TF, etc.
2. The binding sites of the TF in the promoters selected are mapped.
3. Primers are designed such that they flank the TF binding site(s) and that PCR fragments of suitable size (80 - 120 bp) are generated (suitable for quantitative PCR).
4. The tool should rank the primer pairs from most-suitable to least-suitable.
5. Primers will then be ordered and used for the experiments.

Discussion
LF: This could be an easy to solve test case, using Primer3 and some custom scripts.

TEST CASE 5

Test Case Title
Integrative Omics analysis workflow

Test Case Acronyme
CattleOmics

Test Case Class
Animals

Contact person
Patrice Humblot – SLU Sweden
Karl Göran Andersson

Contact
nd

Test Case Description
From the work planned we will produce and need to integrate and relate information from phenotypic information, genomics, transcriptomics, epigenetics in different types of tissues i.e. oocytes, granulosa cells and embryos, obtained without or after exposure to metabolic stressors (4 or 5). It is also planned to study the basic status of cells (0h) and response at different times (at least 24h maturation) after challenge with metabolic stressors. In the second project, oocytes will be produced in vivo from OPU and from slaughter house. Embryos will be produced in vitro and in vivo. Breed effects will be studied in the basic situation (absence of metabolic stressors and following challenges). It is planned to test different dosages (at least 3) per type of metabolic stressor (insulin project 1, NEFA project 2) On oocytes and granulosa cells.

Background knowledge
Dairy cow fertility has been decreasing for a long time due to genetic selection mainly orientated towards milk production. Metabolic imbalance is affecting around 70% or more of dairy cows during the period following parturition. Some cows recover from metabolic imbalance rather rapidly while others remain in negative energy balance that Induces reproductive disorders such as delayed resumption of ovarian activity and poor fertility / embryo quality. This induces longer intervals between calving and conception due mainly to conception failures and early embryonic death. There may be also differences in relation with genetic origin, in the way cows suffer from the disease and in the way reproduction is affected. The proposed projects (submitted to FORMAS in 2010 and 2012) aims to integrate phenotypic and molecular information to understand, predict and find solutions to lower the impact of metabolic stress in cows. One of the aims is to find molecular markers from oocytes, granulosa cells or embryos which may be related to different phenotypic expression and which may be differentially expressed in relation with genetic origin of cells.

Initial state of the Test case
A lot of data of different nature and different data type

- Phenotypic information (about 50 variables)
- Epigenetics and transcriptomics (50 different conditions)

A jungle of tools and data workflows, many do not interact and produce absolutely different data formats.

Desired final state of the Test Case
Defined workflows, automated if possible, with common data formats to synchronize the data integration process.

Test Case Work Plan
nd

Discussion
These test cases represent an example of what could be repeated for different types of stressors (other metabolic stressors, environmental disruptors….) on oocytes and granulosa cells to study the impact of such changes on the early embryo. The same type of model could be used to study the effect of pathogens while exposing different biological material (embryos or oocytes) to study host interactions with pathogens in reproductive tissues and potential differences between breeds to infection

TEST CASE 6

Test Case Title
Bacterial Comparative Genomics

Test Case Acronyme
BacComGen

Test Case Class
Microbial

Contact person
Laurent Falquet – SIB Switzerland

Contact
nd

Test Case Description
Comparing dozens or hundreds of bacterial genomes is a challenging task. Especially if those genomes are unfinished drafts. We would like to find a tool that would compare any number of closely related genomes, identify a core set of genes in all genomes and produce per genome, a list of additional genes, a list of mutated core genes and a list of absent genes.

Background knowledge
Draft genomes are usually incomplete, the prediction of the genes and the classification is thus a difficult case. In addition when the number of draft genomes grows above 50, it renders large scale analysis probably impossible.

Initial state of the Test case
A series of draft genomes of closely related bacterial strains or species.

Desired final state of the Test Case
List of genes:

- core genes found in all genomes
- mutated core genes in each genome
- additional genes in each genome
- missing core genes in each genome

Test Case Work Plan
The participants will have to predict the genes, identify the core genome and for each genome the set of mutated, additional or missing genes. The tool should cope with hundred if not thousands of genomes…

Discussion
LF: my case, I would like to invite people involved in clustering (OrthoMCL, TGICL, etc…)

TEST CASE *7*

Test Case Title
Processing an un-annotated genomes

Test Case Acronyme
NewGenome

Test Case Class
Microbial

Contact person
Paul Bowyer, Jane Mabey Gilsenan - University of Manchester

Contact
nd

Test Case Description
In 2001, we initiated the Central Aspergillus Data Repository (CADRE). This project aims to support the international Aspergillus research community by gathering all genomic information regarding this significant genus into one resource. Using the Ensembl suite, CADRE has been publicly available since 2004. It currently facilitates the visualisation of annotation and comparative data (courtesy of Ensembl Genomes) for nine Aspergillus genomes. Currently, we are also involved in a European systems biology study of the specificity of response of the cell-mediated immune system to fungal microorganisms in order to investigate the genetic basis of susceptibility to fungal disease and elucidate molecular mechanisms of drug resistance in fungal pathogens. Our first test cases to arise from this project are nine different strains of Aspergillus fumigatus: CEA10, AF300, F17999, F18329, F18454, F20451, F21572, F21732 and F21857.

Background knowledge
Aspergillus is a genus of moulds that are found world-wide. Over 300 species of it have been recognised, a small number of which cause illness in animals (including humans): A. fumigatus, A. flavus, A. terreus and A. niger. Most humans are naturally immune and do not develop any of the diseases caused by these moulds. However, when disease does occur, it takes several forms, ranging from an allergy-type illness to life-threatening systemic infection. The severity of the infection is determined by various factors, but one of the most important is the state of the immune system of the person. The most frequently isolated of these moulds from immuno-compromised patients is A. fumigatus.

Initial state of the Test case
Setting up a pipeline to process un-annotated genomes is fraught with difficulties. In theory, given a set of related data files, a typical pipeline ties in different programs to detect repeats, to predict genes and to combine the data to provide the best gene model (see flowchart). These predicted genes can then be used to detect homologues in public databases and be embellished further. In practice, however, each file provided for a project can describe data attributed to different assembly components (sometimes with no assembly description). In addition, each piece of software produces its own variation of a standard format that cannot be used by the next program in the chain. For example, for one of our test cases (Aspergillus fumigatus CEA10), assembled sequences are provided by chromosome but other data (e.g., RNA-seq, DIP and SNP) are provided by supercontig. There is no assembly file describing the relationships between the components. Initially, this in not a problem, but it does cause difficulties further down the line when trying to improve gene models, to visualise all data within a genome viewer for further analyses and to

upload data into a related database such as CADRE. We used GeneMark-ES and Augustus for ab inito gene prediction; both required the use of different in-house programs to convert the output before it could be passed onto Evidence Modeller (EVM). The output from EVM again required conversion before being used with BLAST to find the best match within an Aspergillus database. Development of an in-house program was required to merge resultant data into gff3 format for the next step in the process. The Ensembl suite uses the gff3, agp and fasta file formats for uploading data into an ensembl database system such as CADRE. However, as pointed out earlier we do not always have consistent assembly information, therefore, we may need to re-create such files from a reference genome. A new database is initiated within CADRE with FASTA sequence files and assembly data (i.e., agp files), where appropriate. EBI software (GffDoc.pl), can then be used to upload predicted genes from gff3 files. All in all, this has become a rather bloated and time-consuming process.

Desired final state of the Test Case
To be able to use genome project data without worrying about the way in which it has been described. To enable the annotation pipeline to run without continual data reformatting.

Test Case Work Plan
As discussed in a previous section, we have written several in-house programs to try and sort out formatting issues. We would need to create a smoother process and have it less fragmented. We also need to develop gff, gtf and gff3 parsers that deal with more than just the two gene finding programs we have used. We would like to introduce GeneWise but this will probably have its own formatting issues so we would need to introduce flexibility.

Discussion
The test case will have impact for the substantial worldwide fungal genomics community

LF: a typical pipeline to build (ideal test case!?) with probably a risk of failure in the long term run
Invite people from Taverna or other such tools.

TEST CASE 8

Test Case Title
Discover non-coding RNA in yeast

Test Case Acronyme
Yeast-ncRNA

Test Case Class
Microbial

Contact person
Stefan Astrom and Feng Lin

Contact
nd

Test Case Description
We sequenced the genome of Kluyveromyces dobzhanskii, a non-model organism. As a part of this endeavor, we have also very recently finished sequencing the transcriptomes (polyA+) of these two species. When we compared the whole genome sequences between them, we found more than fifty conserved sequences longer than 200bp located in the intergenic region whose function was unknown. We call these loci conserved non-coding sequences (NCS). We deleted 10 of these NCSs in K. lactis and found that 5 of the resulting strain resulted in a phenotype different from wild type. Hence, several of the NCSs were functionally important. Further analysis showed that several NCSs defined extended 3' UTRs of the flanking genes. However, others did not and we hope to find noncoding RNAs among these elements. So what exactly are these NCSs and how many functional elements exist in the intergenic region of the whole genome? Finally, Translational frameshifting is an alternate process of translation, in which the ribosome slides back one nucleotide (-1 frameshifting) or skips one nucleotide (+1 frameshifting). We have identified a gene from K. lactis that contains a programmed frameshift of a novel type. Programmed frameshifts require a so-called slippery site and the slippery site we identified is different from the sites that were previously published.

Background knowledge
K. dobzhanskii is closely related species to Kluyveromyces lactis. K. lactis, also called milk yeast, must be considered a model as many groups world-wide work with this organism. The K. lactis genome is relatively well annotated. The general idea of our work is to establish K. lactis and K. dobzhanskii as a pair of organisms that can be compared to learn more about how yeast genomes evolved. As an example conserved regulatory sequences can be identified using these two genomes (i.e. phylogenetic footprinting). Similar work has been performed with S. cerevisiae and close relatives (i.e. S. bayanus), but since the Kluyveromyces yeasts are separated by >100×106 years of evolution from the Saccharomyces yeasts, we argue that novel and interesting discoveries can be made.

Initial state of the Test case
To address these issues we hope to obtain bioinformatics help. One issue is to separate 5' and 3' UTRs from distinct intergenic transcripts. Next, distinct intergenic transcripts must be filtered by BLAST- and pFAM-searches to exclude short protein coding transcripts that undoubtedly exist among them. We also need help to estimate the quality of the RNA seq data. Part of the data has been assembled back to the genome sequences. We noticed there are some expressed sequences in the intergenic region at variable expression level. We need bioinformatic methods to do the

entire genome analysis for predicting non-coding RNA or other functional elements. We are also interested in the difference of expression pattern between the two genomes in general.

Further we would like to analyze this novel slippery site with respect to the presence genes and if slippage can generate a novel gene product.

Desired final state of the Test Case
nd

Test Case Work Plan
nd

Discussion
LF: complexe case mixing different questions, it needs to be more clearly defined or split into several individual test cases.

TEST CASE 9

Test Case Title
Pathway analysis of a poorly annotated but sequenced plant genome

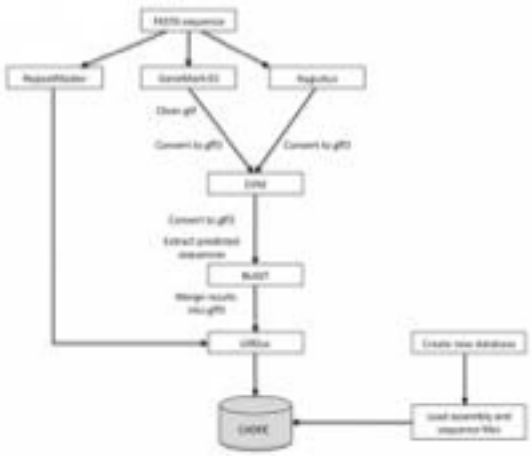
Test Case Acronyme
PlantPathX

Test Case Class
Plants

Contact person
Erik Andreasson - SLU Sweden

Contact
nd

Test Case Description
In order to be able to efficiently analyze large scale '-omics data' (e.g. from microarrays, RNA-seq or quantitative proteomics) and put them in biological context a division of genes into functional categories, such as pathways, is very efficient. For non-model organisms, however, little functional information for individual genes or proteins exists. Instead, researchers have to rely on information derived from sequence identity to better studied model organisms. With the drop in cost for sequencing organisms, the genome sequences of more and more 'non-model' are known and subsequently '-omics' data can efficiently be generated as well. These large-scale '-omics'-data can be very informative and should ultimately be used together with sequence identity in the annotation effort of the sequenced non-model species. Efficient method/workflows to derive information by sequence identities to model organisms combined with efficient use of existent "-omics" data for the species studied would be desirable for visualisation and analysis. We are studying potato, which genome was sequenced last year. Closely, related to this species is the tomato genome which is expected to be released in 2012.



Background knowledge
When exploring a non-model organism with a sequenced genome, comparison of gene families (e.g. presence or absence and putative number of homologs) is an initial, but crucial step giving a first overview. For plant species PLAZA db provides a good platform for this. Tools for visualisation, incorporation of '-omics' data and network analysis exist in Cytoscape. However, for a biologist a best-practice concept and/or training in using these tools would be helpful. Also an integrated approach could be useful.

Initial state of the Test case
GSII Illumina reads exist for 3 different potato cultivars. In addition, microarray data and quantitative proteomics data from various states exist. Samples are from leaves. Currently, the OrthoMCL clusters generated in conjunction to the publication of the potato genome (Nature 475, 189–195) are used for gene family analysis together with BLAST. A MapMan binning file has been generated based on the potato genome and is publicly available. Visualisation of '-omics' data has been done in a commercial software, QluCore, but this does not handle multiple data types simultaneously well and does not visualise functional pathways.

Desired final state of the Test Case
A good way to easily compare similarities and differences of putative pathways between species and cultivars, and at the same time visualise and explore multiple data types that can be linked to genes and gene products (transcripts and proteins); "multi-omics".

Test Case Work Plan
nd

Discussion
This type of analysis would be useful for all non-model organisms with a sequenced genome – a category that is rapidly expanding due to lower sequencing costs.

LF: a test case similar to TC12, but for larger plant genomes, with the ability to add and crossvalidate *omics information. Huge amount of work required and still a research field. I would not try to do everything, may be ask them to specify the most important tool they need.

TEST CASE 10

Test Case Title
Gene prediction modelling for plants

Test Case Acronyme
GPM4P

Test Case Class
Plants

Contact person
Erik Andreasson and Hans Ronne - SLU Sweden

Contact
nd

Test Case Description
There is no established workflow for gene modelling. We would like to test whether an optimal method to create gene models can be found for plant species. One possibility is to try different pipelines in parallel and verify and benchmark them against available organism-specific proteomics data. The idea would be to try some different plant species to see whether the same workflow suits them equally well. Arabidopsis, rice and Physcomitrella could be used as plant species because of their evolutionary divergence and availability of proteomic data. Potatoes would be of special interest for one of the actors (Erik Andreasson's group).

Background knowledge
Recently a Nature article was published (Nat Rev vol 13:329-342) discussing various pipelines for gene prediction. This article is a good start to discuss these issues and have good suggestions of possible pipelines.

An observation from RNA-seq data related to this question is that incorrect splicing and splicing variants are common in plants. A better understanding on possibilities and drawbacks with different gene prediction workflows would help to address also this issue.

Initial state of the Test case
Public sequencing and proteomic data is available for Arabidopsis, rice and Physcomitrella. Erik Andreasson's group has both RNA-seq and proteomics data from potato. Various tools are used for gene prediction, e.g. Augustus, Genemark.

Desired final state of the Test Case
A benchmark of different pipelines for plant gene prediction, if possible, establishing a "golden standard" for gene prediction in plants.

Test Case Work Plan
We can provide help with sequences (RNA-seq) and finding suitable, public proteomics data. For potato proteomics, data that could be used for benchmarking exists in-house.

Discussion
This comparison of workflows could be very useful for the plant community and establish whereas a "golden standard" is possible or whether individual pipeline solutions should be found for each plant species.

LF: test case similar to TC14, a pipeline to benchmark gene prediction could be designed. Perhaps a contest similart to nGASP: http://wiki.wormbase.org/index.php/NGASP Invite people who organised such a contest.

TEST CASE 11

Test Case Title
Genome assembly of heterozygous diploid data

Test Case Acronyme
HETERO

Test Case Class
Animals (Plants)

Contact person
Jacob Höglund, Biao Wang – EcolGen; Ignas Bunikis, Hossein Jorjani, Magnus Alm Rosenblad, Eva Hellmén – University of Uppsala

Contact
nd

Test Case Description
Many NGS projects will not have access to data from inbred lines. This will cause problems in the assembly. How can this be solved?

Background knowledge
Current technology generate short reads, or low coverage. Extent and nature of DNA duplications is unknown. Which NGS approach is best?

Initial state of the Test case
nd

Desired final state of the Test Case
nd

Test Case Work Plan
nd

Discussion
LF: this is a research topic, especially for multiploid genomes. At the moment no solution exist. We could invite people how worked on similar projects or specialised in phasing.

TEST CASE 12

Test Case Title
Functional annotation of the potato genome

Test Case Acronyme
PotGen

Test Case Class
Plants

Contact person
Erik Alexandersson

Contact
nd

Test Case Description
The potato genome assembly is available through the Potato Genome Sequencing Consortium (v. 3;http://potatogenomics.plantbiology.msu.edu/index.html). Gene predictions were done ab initio with parameters trained for A. thaliana and also based on sequence similarity with four other plant genomes (ref). Functional annotation of predicted genes were done by identifying orthologous and paralogous gene families in 12 sequenced plant species by OrthoMCL (ref). A MapMan mapping file based on sequence identity to Arabidopsis is also available.

Background knowledge
On the 11 of July last year the sequence of the potato (Solanum tuberosum) genome (850 Mbp) was published (ref). During the sequencing project the potato genome consortium run into several problems due to sequence heterogeneity and eventually genome assembly could only be successfully done based on a homozygous doubled-monoploid potato clone (S. tuberosum group Phureja). The genome structure of this clone differs greatly from the cultivars that are commonly studied, i.e. crop potato cultivars grown for food or as starch for industrial use.

Currently, the genes predicted in the sequenced genome are relatively poorly functionally annotated and, in addition, the predicted genes and annotations from the reference genome need to be transferred to cultivars studied. The goal is that the genome sequence will advance our understanding of molecular processes in the potato and ultimately facilitate advances in breeding.

Initial state of the Test case
Currently no public Gene Ontology (GO) annotation is available for the potato genome. We have set up Blast2Go (ref) in order to retrieve GO terms from the Arabidopsis TAIR resource, but have encountered problems with the MySQL settings. Ideally, we would like to create a system where we retrieve functional gene information from several plant genomes based on sequence identity. The heterogeneity and tetraploidy of the potato genome also need to be taken into account and a robust, unified system for naming gene variants and alleles should be established. Importantly, gene variants and alleles related to the same gene needs to be linked and annotated appropriately, as should related transcript and protein variants.

Desired final state of the Test Case
An annotation effort is needed to make the potato genome more useful. In addition, genes should be linked in putative pathways based on sequence identity to other plants species. The heterogeneity and tetraploid nature need to be considered carefully.

Test Case Work Plan
nd

Discussion
LF: interesting case probably for Paul Kersey and ENSEMBL plant.

TEST CASE 13

Test Case Title
Transcriptome and small RNA comparison of four phenotypically distinct cv "Primitivo" grapevine clones

Test Case Acronyme
TraPrimi

Test Case Class
Plants

Contact person
Annalisa Giampetruzzi Istituto di Virologia Vegetale del CNR, Bari;

Contact
nd

Test Case Description
We selected four different grapevine clones of the grapevine cv. Primitivo. The four clones are phenotypically distinct with respect to yield and morphology of the bunches. Particularly, this latter trait, depending mainly on the size of the berries, directly influences the technological characteristics of grape/must (ratio skin/pulp and than antocians and polyphenols contents) and compactness of bunches directly linked with the different susceptibility to fungal agents of bunch rots. The goal of the study is to identify molecular pathway linked to these phenotypic traits by analyzing small RNA and mRNA libraries from the tissues of these four distinct grapevine clones.

Background knowledge
nd

Initial state of the Test case
nd

Desired final state of the Test Case
nd

Test Case Work Plan
nd

Discussion
LF: Yet another full research project. The should really split it into smaller tasks, that could be analysed separately. e.g., a) transcriptome assembly or remapping onto reference genome? b) counting gene expression (both for mRNA and small RNA) c) compare several gene expression vs each other or vs a reference

TEST CASE 14

Test Case Title
Improved cell factories

Test Case Acronyme: FAIRE-seq
CELLFAC

Test Case Class
Microbial and Plants

Contact person
Dr Mikko Arvas, VTT Technical Research Centre of Finland

Contact
nd

Test Case Description
We carry out basic and applied research of eukaryotic cell factories in order to understand production of and better produce metabolites and proteins required by bioeconomy. For example we study production of monomers for novel plastics by yeast, production of cellulolytic enzymes for biofuel production by filamentous fungi and production of pharmaceuticals by plants. In order to discover novel enzymes and novel metabolic pathways and to learn to improve the performance of eukaryotic cell factories we study genomes of bacteria, fungi and plants. We use proprietary and public genome sequences, however producing standard basic annotation i.e. genes and protein families etc. for eukaryotic sequences is major impediment for our work. A public resource from which to retrieve and with which to produce such annotation would be gravely needed. Secondly, in order to understand the metabolism of cell factories we need to explore and model it. The standard tool of biological data analysis and modeling, R, however lacks high quality interfaces and programming libraries for this task.

Background knowledge
-

Actors
nd

Initial state of the Test case
Genome sequences produced by various tools, oligonucleotide microarray data

Desired final state of the Test Case
-SBML support in R -interface to STRING, KEGG, Metacyc and SEED in R -easy way (like microbesonline) to get the following info for non-model organisms from databases: chromosome/ scaffold sequences, gene sequences and chromosomal coordinates, protein sequences, InterPro functional predictions, protein clustering for getting started in ortholog/paralog searches

Test Case Work Plan
-

Discussion
-

TEST CASE 15

Test Case Title
Genetics, Genomics and Evolution of prolific breeds of domestic sheep (Ovis aries)

Test Case Acronyme
SOMICS

Test Case Class
Livestock

Contact person
Juha Kantanen

Contact
juha.kantanen@mtt.fi

Test Case Description
The test case applies state-of-art tools of modern genomic research, animal feeding experiment and animal physiology studies to unfold the genetics and genomic architecture of prolific sheep breeds as well as the environmental and nutritional effect on the prolificacy of sheep. The final goal of the project is to promote sustainable use of genetic resources of prolific sheep breeds through the use of genomic data.

Background knowledge
The highly prolific breeds of domestic sheep (Ovis aries) are valuable genetic resources for global sheep industry. E.g. the native Finnsheep is well-known for its prolific traits and has been imported into 40 countries to develop new breeds and to improve fertility of local sheep breeds. Female prolificacy traits of sheep are critical factors controlling biological and financial performance of sheep production systems. In our study we are interested in the high ovulation rate and litter size. An access of modern high-throughput genomic tools has led to an increased understanding of the genetic architecture and nature of these complex traits in domestic animals, including sheep.

In domestic sheep, a large range in litter size (1-8 offspring) has been observed among- and within- breeds. This is contrast to several other domestic species whose females have generally either 1-2 (e.g. cattle and goat) or ≥ 4 offspring (e.g. dog and pig). Studies have indicated that the ovulation rate and litter size can be genetically regulated either by a set of different genes with each having a small effect or alternatively, by the action of single genes with major effect, namely the fecundity genes. In addition, nutrition is one of the environmental factors that affect reproduction in sheep.

Our aim is to investigate genetic basis of the highly prolific domestic sheep breeds and their utilisation. By using genome-wide SNP genotyping, candidate gene (e.g. FecB, GDF9, BMP-15, BMPR-1B) sequencing, transcriptome sequencing, animal nutrition experimentation, animal physiology studies, bioinformatics, and population genetics tools, the project will elucidate the specific contributions of genetic and nutritional factors in shaping the phenotypic variability in prolific sheep. In practise, information will enable breeding plans to be developed that maximize the benefits of increased prolificacy in the sheep breeds and their crosses.

The sheep breeds under our focus are also globally important genetic resources being worthy of genetic and genomic characterization.

Actors
PhD, Professor Meng-Hua Li, MTT, Jokioinen, Finland and Chinese Academy of Sciences, Beijing, PR of China PhD, Professor Juha Kantanen, MTT, Jokioinen, Finland PhD, Pricipal Research Scientist Jaana Peippo, MTT, Jokioinen, Finland MSc, Bioinformatics Scientist Kisun Pokharel, MTT, Jokioinen, Finland PhD, Bioinformatics Scientist Daniel Fischer, MTT, Jokioinen, Finland PhD, Researcher Mervi Honkatukia, MTT, Jokioinen, Finland Vet Med Johanna Rautiainen, ProAgria Pirkanmaa, Tampere, Finland International Sheep Genomics Consortium (http://www.sheephapmap.org)

Initial state of the Test case

1. RNA sequencing and micro-RNA profiling (Illumina HiSeq2000); approximately 18 individuals in 2013
2. Whole-genome SNP genotyping (Illumina ovine 700K SNP BeadChip); >1,000 individuals for population genomics analysis and GWAS in 2013
3. Whole-genome sequences (Illumina HiSeq2000); “few individuals” selected from the feeding experiment in 2014

In addition, there are available Sheep HapMap (http://www.sheephapmap.org) whole-genome sequence data. In total, 75 genomes of sheep have been sequenced (Illumina and Roche 454 approach). Currently the sheep genome assembly v3.1 has been released by NCBI and the ISGC (International Sheep Genomics Consortium).

Desired final state of the Test Case
The objectives of the SOMICS-project are:

1. By applying genome-wide association studies, we aim to identify the SNP variations and the candidate genes associated with different fertility phenotypes of sheep and to detect signatures of selective sweeps.
2. By further sequencing candidate regions and whole-genome sequencing of few sheep samples with different phenotypes, we aim to get a detailed and whole picture of functionally relevant candidate regions and the whole genomic profile of few sheep.
3. By using the statistical and bioinformatics methods in conducting the global analyses of the multiple data sets including the molecular, nutritional and environmental data for the sheep breeds, we aim to uncover accurate gene interaction networks and pathways. We further compare the genomics of sheep with other mammals such as mice and human.
4. By designing a feeding experiment for diffrent sheep breeds and by investigating transcriptome and micro-RNA sequences of the breeds, we aim to evaluate the effect of nutrition on gene expression patterns and to infer putative candidate genes for the prolificacy of sheep.

Discussions

Topic wise parallel sessions to discuss the test cases

We will divide the work in 4 parallel groups, one for each of the following topics:

- data formats / standardization and workflows
- scalability of tools (multi genome comparison/data compression)
- non model organisms and multiploidy
- visualization and training

The 4 groups will analyze (according to the topic) each test case to identify

- the difficulties or the current limitations affecting the test case
- the existing tools required (with or without adaptation)
- the missing tools required
- keep minutes of discussion
- prepare a summary for discussion

Each group will report in 15min the results of their session

- is there a common trait in several test cases?

- are there common tools required or missing?

- can we split test cases according to operations?

- can we group test cases?

We will divide again the work in 3 parallel sessions. Each session will analyze 5 test cases (random distribution).

For each test case we need:

- identification of tools
- a SWOT analysis (Strength, Weaknesses, Opportunities and Threats)
- a preliminary work plan for a follow-up activity
- propose other collaborators for the follow-up activity
- keep minutes of discussion

Selection of test cases (or operations) in 3 categories and discussion of the work plan

- relatively easy, workflow exists or can be adapted within weeks

- more difficult, some tools are missing, workflows are partial, need months to adapt

- very difficult, scalability problem, tools completely missing, requires new developments (projects over years)

Organization of the follow-up events to approach the solutions of some test cases

- create work groups to prepare the hack-a-thon events

- decide number of events and divide test cases or operations, plan a soon solution of one test case

- collect emails and skype accounts

- decide upon video meetings schedule and chairperson

- keep minutes of discussion and decisions

Participants

Name	Surname	Institute	E-mail	Info
Mohammad Amin	Omidbakhshfard	University of Potsdam	aminomid@uni-potsdam.de	Test Case 1
Nooshin	Omranian	Max Planck Institute of Molecular Plant Physiology, Potsdam	omranian@mpimp-golm.mpg.de	Test Case 3
Prashanth	Garapati	Max Planck Institute of Molecular Plant Physiology, Potsdam	Garapati@mpimp-golm.mpg.de	Test Case 4
Karl Göran	Andersson	Swedish University for Agricultural Sciences, Uppsala	Goran.Andersson@slu.se	Test Case 5
Laurent	Falquet	Swiss Institute of Bioinformatics, Lausanne	Laurent.Falquet@isb-sib.ch	Test Case 6 AllBio
Annalisa	Giampetruzzi	Institute of Plant Virology CNR, Bari	annalisa79@hotmail.it	Test Case 13
Mikko	Arvas	VTT Technical Research Centre of Finland	Mikko.Arvas@vtt.fi	Test Case 14
Juha	Kantanen	MTT Agrifood Research Finland	juha.kantanen@mtt.fi	Test Case 15
Erik	Alexandersson	Swedish University for Agricultural Sciences, Alnarp	Erik.Alexandersson@slu.se	Test Cases 9,12
Ashfaq	Ali	Swedish University for Agricultural Sciences, Alnarp	ashfaq.ali@slu.se	

Veselin	Baev	University of Plovdiv	vebaev@gmail.com	
Endre	Barta	University of Debrecen	barta.endre@unideb.hu	
Kajsa Pernilla	Bjerling	University of Uppsala	pernilla.bjerling@imbim.uu.se	
Erik	Bongcam Rudloff	Swedish University for Agricultural Sciences, Uppsala	erikbong@mac.com	AllBio
Giovanni Nicola	Bubici	Institute of Plant Virology CNR, Bari	g.bubici@ba.ivv.cnr.it	
Domenico	Catalano	Institute of Plant Genetics CNR, Bari	domenico.catalano@igv.cnr.it	
Bruno	Contreras Moreira	CSCI, Zaragoza	bcontreras@eead.csic.es	
Domenica	D'Elia	Institute for Biomedical Technologies CNR, Bari	domenica.delia@ba.itb.cnr.it	
Agnieszka Danuta	Danek	Silesian University of Technology, Poland	agnieszka.danek@gmail.com	
Tomas	Di Domenico	University of Padova	tdido@bio.unipd.it	
Andreas	Gisel	Institute for Biomedical Technologies CNR, Bari	andreas.gisel@ba.itb.cnr.it	AllBio
Tatyana	Goldberg	Technische Universität, München	goldberg@rostlab.org	
Thomas James	Hardcastle	University of Cambridge	tjh48@cam.ac.uk	

Paul Julian	Kersey	EBI, Hinxton	pkersey@ebi.ac.uk	AllBio
Petri Juho Antero	Klemelä	IT Center for Science, Helsinki	petri.klemela@csc.fi	
Sofia	Kossida	Biomedical Research Foundation, Academy of Athens	skossida@bioacademy.gr	
Ida	Lager	Swedish University for Agricultural Sciences, Alnarp	Ida.Lager@slu.se	
Vito Flavio	Licciulli	Institute for Biomedical Technologies CNR, Bari	flavio.licciulli@ba.itb.cnr.it	
Shahid	Manzoor	Swedish University for Agricultural Sciences, Uppsala	Shahid.Manzoor@slu.se	
Tiffanie Yael	Maoz	Weizmann Institute of Science, Tel Aviv	Yael.Maoz@weizmann.ac.il	
Giovanni	Minervini	University of Padova	giovanni.minervini@bio.unipd.it	
Intawat	Nookaew	Chalmers University of Technology, Göteborg	intawat@chalmers.se	
Jaana Maarit Helena	Peippo	MTT Agrifood Research Finland	jaana.peippo@mtt.fi	
Gianvito	Pio	University of Bari	gianvito.pio@uniba.it	
Alberto	Policriti	Università di Udine	policriti@gmail.com	

Estelle	Proux-Wera	Swedish University for Agricultural Sciences, Alnarp	estelle.wera@slu.se	
Gregoire	Rossier	Swiss Institute of Bioinformatics, Lausanne	gregoire.rossier@isb-sib.ch	
Shumaila	Sayyab	Swedish University for Agricultural Sciences, Uppsala	shsa0001@stud.slu.se	
Axel	Thieffry	Swedish University for Agricultural Sciences, Alnarp	axel.thieffry@gmail.com	
Angelica	Tulipano	Institute for Biomedical Technologies CNR, Bari	angelica.tulipano@ba.itb.cnr.it	
Oren	Tzfadia	Weizmann Institute of Science, Tel Aviv	oren.tzfadia@weizmann.ac.il	
Jacques Hippolyte H	van Helden	Université Libre de Bruxelles	Jacques.van.Helden@ulb.ac.be	AllBio
Dimitrios	Vlachakis	Biomedical Research Foundation, Academy of Athens	dvlachakis@bioacademy.gr	
Gerrit	Vriend	Radboud University Nijmegen Medical Centre	vriend@cmbi.ru.nl	AllBio