Annual Danish Bioinformatics Conference

August 25-26, 2016

PROGRAMME

POSTER ABSTRACTS
LIST OF PARTICIPANTS
ORGANISATION
PRACTICAL INFORMATION
CONTENT

Programme ..............................................3-7
Abstracts - talks .................................7-  29
Abstracts - posters ..............................30-102
Abstracts – ELIXIR DK booth.......103-105
Participants ..........................................106-109
Organisation ...........................................110

Practical information  .................111-112
  • Contact
  • Emergency
  • Taxi
  • Dinner
  • Hotels

WIFI
Network: guest
Password: 5000

#elixirdk
Annual Danish Bioinformatics Conference

August 25-26, 2016 - Odense, Denmark

Sessions

Computational Proteomics • Computerome & ELIXIR Denmark Industrial View on Danish Bioinformatics • Machine Learning in Bioinformatics Network Biology in Disease • Non-coding RNA Bioinformatics Population Genetics & Evolution • Transcriptomics & Transcriptional regulation

Confirmed Speakers

- Anaïs Baudot, CNRS, FR
- Søren Brunak, DTU, UCPH, RegionH, DK
- Jürgen Cox, MPG; DE
- Rute Fonseca, UCPH, DK
- Jon Ison, DTU, DK
- Anders Krogh, UCPH, DK
- Peter Løngreen, DTU, DK
- Lennart Martens, UGent, BE
- Anton Petrov, EMBL-EBI, UK
- Mikkel H. Schierup, AU, DK
- Bruce Shapiro, NIH/NCI, USA
- Martin Simonsen, QIAGEN, DK
- Robert Young, UoE, UK

Programme Committee

- Robin Andersson, UCPH, DK
- Søren Basenbacher, AU, DK
- Kirstine G. Belling, UCPH, DK
- Vivi Gregersen, AU, DK
- Ida Moltke, UCPH, DK
- Kasper Munch, AU, DK
- Bent Petersen, DTU DK
- Veit Schwämmle, SDU, DK
- Stefan Seemann, UCPH, DK

Contact

Myhanh Nguyen <myng@bio.dtu.dk>

Venue

DOK5000 Conference Centre,
Havnegade 18, 5000 Odense

Registration

http://www.elixir-denmark.org
- Fee, students: none
- Fee, non-students: 1.200 DKK, excl. VAT

Deadlines

- Registration: Aug. 22, 2016
### Annual Danish Bioinformatics Conference 2016

**Programme**

**Day 1**

<table>
<thead>
<tr>
<th>Thursday</th>
<th>Session</th>
<th>Chair</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30 – 10:00</td>
<td>Registration</td>
<td></td>
</tr>
<tr>
<td>10:00 – 10:20</td>
<td>Welcome</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>View on Danish Bioinformatics.</em> Søren Brunak, UCPH, DTU, RegionH;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anders Krogh, UCPH</td>
<td></td>
</tr>
<tr>
<td>10:20 – 10:30</td>
<td>CBioVikings</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Bringing the next generation of Bioinformaticians together.</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alexander Hauser, CBioVikings</td>
<td></td>
</tr>
<tr>
<td>10:30 – 12:00</td>
<td>Network Biology in Disease</td>
<td></td>
</tr>
<tr>
<td>10:30-11:00</td>
<td><em>The Perseus computational platform for comprehensive analysis of (pro)omics data.</em></td>
<td>Kirstine Belling</td>
</tr>
<tr>
<td>11:00-11:30</td>
<td>Jürgen Cox, MPG, DE (keynote speaker)</td>
<td>Bent Petersen</td>
</tr>
<tr>
<td>11:30-11:45</td>
<td><em>Modeling and mining networks to study diseases.</em> Anaïs Baudot, CNRS, FR (keynote speaker)</td>
<td></td>
</tr>
<tr>
<td>11:45-12:00</td>
<td><em>Systems BioMedicine using high-throughput screening - From RNAi, miRNA and drug screens to targeting signaling pathways.</em> Jan Baumbach, SDU, DK</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Unsupervised learning of complex and time-series OMICS data together with biological networks.</em> Richard Röttger, SDU, DK</td>
<td></td>
</tr>
<tr>
<td>12:00 – 12:45</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>12:45 – 14:15</td>
<td>Transcriptomics and Transcriptional Regulation</td>
<td></td>
</tr>
<tr>
<td>12:45-13:15</td>
<td><em>Bidirectional transcription marks accessible chromatin and is not specific to enhancers.</em> Robert Young, UoE, UK (keynote speaker)</td>
<td>Robin Andersson</td>
</tr>
<tr>
<td>13:30-13:45</td>
<td><em>Characterizing age-dependent regulatory variation in the human frontal lobe region.</em> Maria Dalby, UCPH, DK</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td>Chair</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>13:45-14:00</td>
<td><strong>Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters.</strong> Yun Chen, UCPH, DK</td>
<td>Robin Andersson</td>
</tr>
<tr>
<td>14:00-14:15</td>
<td><strong>The Landscape of Isoform Switching in human cancer.</strong> Kristoffer Vitting Seerup, UCPH, DK</td>
<td>Ida Moltke</td>
</tr>
<tr>
<td>16:00 - 16:15</td>
<td><strong>Break</strong></td>
<td></td>
</tr>
<tr>
<td>14:30 - 16:00</td>
<td><strong>Population genetics and evolution</strong></td>
<td></td>
</tr>
<tr>
<td>14:30-15:00</td>
<td><strong>Hummingbird diversification: a tale of introgression.</strong> Rute R. da Fonseca, UCPH, DK (keynote speaker)</td>
<td></td>
</tr>
<tr>
<td>15:00-15:15</td>
<td><strong>Nationwide genomic study in Denmark reveals remarkable population homogeneity.</strong> Georgios Athanasiadis, AU, DK</td>
<td>Søren Besenbacher</td>
</tr>
<tr>
<td>15:15-15:30</td>
<td><strong>Reconstructing ancient pathogens - discovery of Yersinia pestis in Eurasia 5,000 years ago.</strong> Simon Rasmussen, CBS, DTU, DK</td>
<td>Kasper Munch</td>
</tr>
<tr>
<td>15:30-15:45</td>
<td><strong>Improved D-statistic for low-coverage data.</strong> Samuele Soraggi, UCPH, DK</td>
<td></td>
</tr>
<tr>
<td>15:45-16:00</td>
<td><strong>Inference of distribution of fitness effects: impact of beneficial mutations.</strong> Paula Tataru, AU, DK</td>
<td></td>
</tr>
<tr>
<td>16:00 - 16:15</td>
<td><strong>Break</strong></td>
<td></td>
</tr>
<tr>
<td>16:15 - 17:15</td>
<td><strong>POSTER SESSION</strong></td>
<td></td>
</tr>
<tr>
<td>17:15 - 18:45</td>
<td><strong>Non-coding RNA Bioinformatics</strong></td>
<td></td>
</tr>
<tr>
<td>17:15-17:45</td>
<td><strong>RNAcentral and Rfam: Tools for understanding the RNA universe.</strong> Anton Petrov, EMBL-EBI, UK (keynote speaker)</td>
<td>Stefan Seemann</td>
</tr>
<tr>
<td>17:45-18:00</td>
<td><strong>Large-scale prediction of RNA-RNA interactions and siRNA off-targets.</strong> Ferhat Alkan. UCPH, DK</td>
<td></td>
</tr>
<tr>
<td>18:00-18:45</td>
<td><strong>Conditional Activation of RNA Nanoconstructs: Computational Design and Experimental Verification.</strong> Bruce Shapiro, NIH/NCI, US (keynote speaker)</td>
<td></td>
</tr>
<tr>
<td>19:30</td>
<td><strong>Dinner at Musikhuset Posten, Østre Stationsvej 35, 5000 Odense C</strong></td>
<td></td>
</tr>
<tr>
<td>Friday</td>
<td>Session</td>
<td>Chair</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>9:00 – 10:30</td>
<td><strong>Machine Learning in Bioinformatics</strong></td>
<td></td>
</tr>
<tr>
<td>9:00-9:30</td>
<td>• <em>Multi-layered latent variable models for learning genome-to-phenome maps</em>. Luc Janss, AU, DK (keynote speaker)</td>
<td></td>
</tr>
<tr>
<td>9:30-9:45</td>
<td>• <em>Identification of known and novel recurrent viral sequences in data from multiple patients and multiple cancers</em>. Jose MG Izarzugaza, DTU, DK</td>
<td></td>
</tr>
<tr>
<td>9:45-10:00</td>
<td>• <em>Tandem mass spectrometry peptide fragment ion prediction by Hidden Markov Models</em>.</td>
<td>Vivi Gregersen</td>
</tr>
<tr>
<td></td>
<td>Jan Christian Refsgaard, UCPH, DK</td>
<td></td>
</tr>
<tr>
<td>10:00-10:15</td>
<td>• <em>Chromosomal transcriptional initiation walks reflect three dimensional chromatin organization</em>. Sarah Rennie, UCPH, DK</td>
<td></td>
</tr>
<tr>
<td>10:15-10:30</td>
<td>• <em>Cataloging the landscape of RNA bioinformatics tools</em>. Anne Wenzel. UCPH, DK</td>
<td></td>
</tr>
<tr>
<td>10:30 - 10:45</td>
<td><strong>Break</strong></td>
<td></td>
</tr>
<tr>
<td>10:45 - 11:45</td>
<td><strong>ELIXIR Denmark</strong></td>
<td>Peter Løngreen</td>
</tr>
<tr>
<td>10:45-11:25</td>
<td>• <em>bio.tools: ELIXIR Life Science Software Registry &amp; Demo of Software</em>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jon Ison &amp; Emil Rydza, DTU, DK</td>
<td></td>
</tr>
<tr>
<td>11:25-11:45</td>
<td>• <em>Activities at the genomedk HPC cluster</em>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mikkel H. Schierup, AU, DK</td>
<td></td>
</tr>
<tr>
<td>11:45 – 12:30</td>
<td><strong>POSTER SESSION</strong></td>
<td></td>
</tr>
<tr>
<td>12:30 – 13:15</td>
<td><strong>Lunch</strong></td>
<td></td>
</tr>
</tbody>
</table>
Day 2 - continued

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Chair</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:15 – 14:15</td>
<td><strong>Industrial View on Danish Bioinformatics</strong></td>
<td></td>
</tr>
<tr>
<td>13:15-13:35</td>
<td>• Rapid detection of drug-resistance – a successful industry-academic cooperation. Martin Simonsen, QIAGEN, DK (keynote speaker)</td>
<td></td>
</tr>
<tr>
<td>13:55-14:15</td>
<td>• Providing NGS data analysis as a service – XploreRNA. Jesper Culmsee Tholstrup, Exiqon (keynote speaker)</td>
<td></td>
</tr>
<tr>
<td>14:15 – 14:30</td>
<td><strong>Break</strong></td>
<td></td>
</tr>
<tr>
<td>14:30 – 15:45</td>
<td><strong>Computational Proteomics</strong></td>
<td></td>
</tr>
<tr>
<td>14:30-15:00</td>
<td>• Teaching our proteomics tools new trick: innovative algorithms from old data. Lennart Martens, UGent, BE (keynote speaker)</td>
<td></td>
</tr>
<tr>
<td>15:00-15:15</td>
<td>• Proteogenomic analyses for discovery of bispecific chimeric antigen receptor targets. Lars Rønn Olsen, CBS, DTU, DK</td>
<td></td>
</tr>
<tr>
<td>15:15-15:30</td>
<td>• Employing complementary ions for deconvolution of mixture tandem mass spectra. Vladimir Gorshkov, SDU, DK</td>
<td>Veit Schwämmle</td>
</tr>
<tr>
<td>15:30-15:45</td>
<td>• Scientific workflows for combining MS and MS/MS data and improving mass measurement accuracy in proteomics. Arzu Tugce Guler, LUMC, NL</td>
<td></td>
</tr>
<tr>
<td>15:45 – 16:00</td>
<td><strong>Meeting closure &amp; Announcement of best poster</strong></td>
<td>Søren Brunak</td>
</tr>
</tbody>
</table>
Systems BioMedicine using high-throughput screening – From RNAi, miRNA and drug screens to targeting signaling pathways

Markus List, Jan Baumbach

Computational Biology and Applied Algorithms, Max Planck Institute for Informatics, Saarbrücken, Germany
Computational Biology group, Institute of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

Abstract:

High-throughput screening (HTS) is an indispensable tool for drug (target) discovery that currently lacks user-friendly software tools for the robust identification of putative hits from HTS experiments and for the interpretation of these findings in the context of systems biology. We developed HiTSeekR as a one-stop solution for chemical compound screens, siRNA knock-down and CRISPR/CAS9 knock-out screens, as well as microRNA inhibitor and - mimics screens. We will present three use cases that demonstrate how using HiTSeekR one may fully exploit HTS screening data in quite heterogeneous contexts to generate novel hypotheses for follow-up experiments: (1) a genome-wide RNAi screen to uncover modulators of TNF-alpha, (2) a combined siRNA and miRNA mimics screen on vorinostat resistance and (3) a small compound screen on KRAS synthetic lethality. HiTSeekR is the first approach to close the gap between raw data processing, network enrichment and wet lab target generation for various HTS screen types.

Web: http://hitseekr.compbio.sdu.dk

Papers:
Unsupervised learning of complex and time-series OMICS data together with biological networks

Christian Wiwie¹, Jan Baumbach¹,²,³, and Richard Röttger¹

¹Department of Mathematics and Computer Science, University of Southern Denmark
{wiwiec,jaumbac,roettger}@imada.sdu.dk
²Computational Systems Biology, Max Planck Institute for Informatics
³Cluster of Excellence for Multimodal Computing and Interaction, Saarland University

Abstract
In the last decade we were able to witness a literal explosion of the amount, wealth and quality of available biological data. Researchers in various fields in biomedicine increasingly face challenges which can be regarded as Big Data problems. Here, unsupervised learning is very often a first step of an in-depth analysis used for instance for feature reduction to analyzing the inherent structure of the data fueling analysis pipelines from disease sub-typing to network enrichment to mention just a few.

The practitioner faces a multitude of challenges in practice: How to find suitable tools, how to identify optimal parameters, how to pre- and post-process the data, and the like? Furthermore, more complex data demand also more sophisticated computational methods. Popular example comprise the analysis of time-series data or the combination of various types of data, e.g., OMICs data with biological networks.

We have tackled some of these challenges and present with ClustEval[1], a fully automated clustering framework, and TiCoNE, a time-series clustering tool with integrated network enrichment, two tools enabling highly sophisticated analyses for complex biological data. We will demonstrate the power of both tools by analyzing time-series gene expression data of influenza infected patients. We first identify groups of genes behaving consistently over time during an infection and afterwards identify highly enriched parts of the protein-protein interaction network of these consistent genes. Both tools are freely available for download at http://clusteval.sdu.dk and http://ticone.compbio.sdu.dk.

References


Malte Thodberg¹,², Ajuna Azad¹,², Axel Thieffry¹,², Jette Lange¹,², Mette Boyd¹,², Olaf Nielsen³, Albin Sandelin¹,²

Affiliations:
¹Sandelin Group, Bioinformatics Centre, Department of Biology, University of Copenhagen
²Basic Research and Innovation Centre, University of Copenhagen
³Cell Cycle and Genome Stability Group, Department of Biology, University of Copenhagen

Abstract:
A large number of studies have shown that alternative promoter usage facilitates diversity and flexibility in gene expression and thus complexity in the final transcript- and proteome. Still, the fine details of this regulatory mechanism remain unknown. We performed Cap Analysis of Gene Expression (CAGE) on 15 samples of *S. pombe*, a common unicellular model organism, to investigate promoter usage under various stress conditions: nitrogen starvation, heat stress, oxidative stress, and change of growth medium.

The resulting atlas of almost 14,000 promoters shows a consistent bias in the position of transcription start sites compared to reference annotation from PomBase as well as extensive usage of multiple promoters by single genes. This includes cases where alternative promoters are preferred under specific types of stress, so-called alternative promoter shifts.

This promoter atlas, along with differential expression information and alternative promoters, the first of its kind, will be made publically available through the PomBase online database to facilitate future research into *S. pombe* transcriptional regulation and promoter architecture.
Characterizing age-dependent regulatory variation in the human frontal lobe region

Maria Dalby\textsuperscript{a}, Cornelis Blauwendraat\textsuperscript{b}, Sarah Rennie\textsuperscript{a}, Margherita Francescatto\textsuperscript{b}, Patrizia Rizzu\textsuperscript{b}, Peter Heutink\textsuperscript{b}, Robin Andersson\textsuperscript{a}

\textsuperscript{a}The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{b}German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany

Abstract:
Age is a major risk factor for most common neurodegenerative diseases, such as Alzheimer’s disease, Parkinson’s disease, and cognitive impairment. While a considerable effort has been made to associate gene expression changes with neurodegenerative disease state and etiology, the general impact of aging on human transcription and transcriptional regulation remains to a large degree unexplored. To this end, we have characterized gene expression levels and regulatory activities over the healthy human aging process, with a particular focus on the frontal lobe brain region.

We systematically characterized the frontal lobe transcriptome using Cap Analysis of Gene Expression (CAGE) on total RNA isolated from post mortem frontal lobe samples of 144 healthy individuals with an age span from 2 to 95 years. From this data, we determined genome-wide activities of frontal lobe gene promoters and transcriptional enhancers and the effect of aging on transcriptional and regulatory variability.

Via detailed characterization and clustering of the major age related trends we identified a conservative set of gene promoters (4.3%) demonstrating significant expression changes across lifespan. The identified set recapitulates known genes associated with senescence in non-human species. Similarly, we identified a fraction (2.4%) of enhancers active in frontal lobe displaying significant expression association with chronological age. Our detailed map of regulatory variation over the human aging process will enable assessments of the impact of regulatory genetic variants on age-related transcriptional programs as well as a focused analysis of genetic variants associated with major neurodegenerative diseases.
Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters

Yun Chen¹,², Athma A. Pai³, Jan Herudek⁴, Michal Lubas²,⁴, Nicola Meola⁴, Aino I. Järvelin⁵,#, Robin Andersson¹, Vicent Pelechano⁵,€, Lars M. Steinmetz⁵,⁶,⁷, Torben Heick Jensen⁴*, Albin Sandelin¹,²*

¹The Bioinformatics Centre, Department of Biology, University of Copenhagen, Denmark.
²Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Denmark.
³Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.
⁴Centre for mRNP Biogenesis and Metabolism, Department of Molecular Biology and Genetics, Aarhus University, Denmark.
⁵European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany
⁶Stanford Genome Technology Center, Palo Alto, 94304 CA, USA
⁷Department of Genetics, Stanford University School of Medicine, Stanford, 94305 CA, USA # Present address: Department of Biochemistry, Oxford University, United Kingdom.
€ Present address: SciLifeLab, Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Sweden.

Corresponding authors: Torben Heick Jensen (thj@mbg.au.dk) and Albin Sandelin (albin@binf.ku.dk)

Abstract:
Mammalian transcriptomes are complex and formed by extensive promoter activity. In addition, gene promoters are largely divergent and initiate transcription of reverse-oriented promoter upstream transcripts (PROMPTs). Although PROMPTs are commonly terminated early, influenced by polyadenylation (pA) sites, promoters often cluster so that the divergent activity of one might impact another. Here, we find that the distance between promoters strongly correlates with the expression, stability and length of their associated PROMPTs. Adjacent promoters driving divergent mRNA transcription support PROMPT formation, but due to pA site constraints, these transcripts tend to spread into the neighboring mRNA on the same strand. This mechanism to derive new alternative mRNA transcription start sites (TSSs) is also evident at closely spaced promoters supporting convergent mRNA transcription. We suggest that basic building blocks of divergently transcribed core promoter pairs, in combination with the wealth of TSSs in mammalian genomes, provides a framework with which evolution shapes transcriptomes.
The Landscape of Isoform Switching in human cancer

Authors:
Kristoffer Vitting-Seerup*, Albin Sandelin

*author to present

Affiliations:
Section for Computational and RNA Biology (SCARB), Department of Biology, University of Copenhagen, Denmark.
Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Denmark.

Abstract:
Isoform switching, which referees to the differential usage of different gene-isoforms in different conditions, are with a few exceptions largely overlooked in cancer biology. This lack of knowledge probably occurs in part because it is difficult to find and predict the functional impact of such switches and in part because the extend of isoform switching in cancers is not known.

To solve these problems we developed IsoformSwitchAnalyzeR, an easy to use R package which enables statistical identification, annotation and visualisation of isoform switches. We used IsoformSwitchAnalyzeR to identify isoform switches 12 cancer types covering almost 6000 cancer patients from The Cancer Genome Atlas (TCGA). We find that isoform switches are extremely common: across the 12 solid cancer types more than 4000 (34% of all) multi-transcript genes display differential isoform usage in at least one cancer type. In 2500 of these genes (20% of all genes) the changes have easily predicted functional consequences such as domain loss, domain switch or loss of coding potential. The genes with isoform switching are not random, but are highly enriched for genes in cell signalling, adhesion and cancer signatures. Many for the found isoform switches are furthermore pan-cancer events and we both identify known isoform switches in new cancer types as well as describe novel pan-cancer isoform switches.

At The Annual Danish Bioinformatics Conference I will mainly focus on communicating the results from the pan-cancer analysis of isoform switching.
Nationwide genomic study in Denmark reveals remarkable population homogeneity

Georgios Athanasiadis

*Bioinformatics Research Centre, Aarhus University, Denmark

Abstract:
Denmark’s genetic history has never been studied in detail. In this work, we analysed genetic and anthropometrical data from ~800 Danish students as part of an outreach activity promoting genomic literacy in secondary education. DNA analysis revealed remarkable homogeneity of the Danish population after discounting contributions from recent immigration. This homogeneity was reflected in PCA and AMOVA, but also in more sophisticated LD-based methods for estimating admixture. Notwithstanding Denmark’s homogeneity, we observed a clear signal of Polish admixture in the East of the country, coinciding with historical Polish settlements in the region before the Middle Ages. In addition, Denmark has a substantially smaller effective population size compared to Sweden and Norway, possibly reflecting further lack of strong population structure. None of these three Scandinavian countries seems to have suffered a depression due to the Black Death in the Middle Ages. Finally, we used the students’ genetic data to predict their adult height after training a novel prediction algorithm on public summary statistics from large GWAS. We validated our prediction using the students’ self-reported height and found that we could predict height with a remarkable ~64% accuracy.
Reconstructing ancient pathogens – discovery of *Yersinia pestis* in Eurasia 5,000 Years Ago

Simon Rasmussen¹, Morten Erik Allentoft², Kasper Nielsen¹, Ludovic Orlando², Martin Sikora², Karl-Göran Sjögren³, Anders Gorm Pedersen¹, Mikkel Schubert³, Alex Van Dam¹, Christian Møløn Outzen Kapel⁴, Henrik Bjørn Nielsen¹, Søren Brunak¹,⁵, Pavel Avetisyan⁶, Andrey Epimakhov⁷, Mikhail Viktorovich Khalyapin⁸, Artak Gnuni⁹, Aivar Kriska¹⁰, Irena Lasak¹¹, Mait Metspalu¹², Vyacheslav Moiseyev¹³, Andrei Gromov¹³, Dalia Pokutta³, Lehti Saag¹², Liivi Varul¹², Levon Yepiskoposyan¹⁴, Thomas Sicheritz-Pontén¹, Robert Foley¹⁵, Marta Mirazón Lahr¹⁵, Rasmus Nielsen¹⁶, Kristian Kristiansen³, Eske Willerslev²,¹⁷

¹Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, 2800 Kongens Lyngby, Denmark.
²Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark.
³Department of Historical Studies, University of Gothenburg, 405 30 Gothenburg, Sweden.
⁴Section for Organismal Biology, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark.
⁵Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen, Denmark.
⁶Division of Armenology and Social Sciences, Institute of Archaeology and Ethnography, National Academy of Sciences, 0025 Yerevan, Republic of Armenia.
⁷Institute of History and Archaeology RAS (South Ural Department), South Ural State University, 454080 Chelyabinsk, Russia.
⁸Orenburg Museum of Fine Arts, 460000 Orenburg, Russia.
⁹Department of Archaeology and Ethnography, Yerevan State University, 0025 Yerevan, Republic of Armenia.
¹⁰Department of Archaeology, University of Tartu, 51003 Tartu, Estonia.
¹¹Institute of Archaeology, University of Tartu, 50-139 Tartu, Estonia.
¹²Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, 199034 St Petersburg, Russia.
¹³Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, 0014 Yerevan, Armenia.
¹⁴Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge, CB2 1QH United Kingdom.
¹⁵Center for Theoretical Evolutionary Genetics, University of California, Berkeley, California 94720-3140, USA.
¹⁶Department of Zoology, University of Cambridge, Downing St, Cambridge, CB2 3EJ United Kingdom.

Abstract:
The bacteria *Yersinia pestis* is the etiological agent of plague and has caused human pandemics with millions of deaths in historic times. However, studying infectious organisms from samples that are isolated in the present time have the disadvantage that it is difficult to identify the order and timing of the key evolutionary events. By studying ancient samples, it is possible to travel back through time and directly identify the order of these events. Here, we report the oldest direct evidence of *Yersinia pestis* identified by ancient DNA in human teeth from Asia and Europe dating from 2,800 to 5,000 years ago. By sequencing the genomes, we find that these ancient plague strains are basal to all known *Yersinia pestis*. We find the origins of the *Yersinia pestis* lineage to be at least two times older than previous estimates. We also identify a temporal sequence of genetic changes that lead to increased virulence and the emergence of the bubonic plague. Our results show that plague infection was endemic in the human populations of Eurasia at least 3,000 years before any historical recordings of pandemics. Our findings open the possibility of identifying other blood-borne pathogens directly from human remains.
Improved D-Statistic for Low-Coverage Data

Author(s): Samuele Soraggi[1], Carsten Wiuf[2], Anders Albrechtsen[3]
Affiliation(s): [1,2] Copenhagen University, Department of Mathematics, Copenhagen, Denmark. [3] Copenhagen University, The Bioinformatics Centre, Copenhagen, Denmark.

Abstract:
Next Generation Sequencing provides a massive quantity of data, and it is widely analyzed in population genetics research. Since many NGS datasets are sequenced at low coverage, SNP and genotype calling have high uncertainty. This might affect statistical methods that make use of genotypes. A commonly used tool for detecting ancient admixture events is the D-statistic.
The D-statistic makes use of the patterns between alleles in four different groups of individuals in order to show the direction of an ancient or modern gene flow or to assess the correctness of a phylogeny of four populations in the configuration (H4(H3(H2,H1))), where H4 is an outgroup. For low-depth sequencing the D-statistic is highly susceptible to errors deriving from the SNP and genotype calling. For low-depth sequencing where genotype calling is not possible the method relies on sampling one allele from a single individual from each group at each site to evaluate ABBA and BABA patterns. This sampling procedure ignores much of the information in the data and only works for one individual from each group.
Moreover, the D-statistic does not allow to infer information on more complex phylogenies, for example having multiple admixture events or more than four groups of individuals. We have implemented in ANGSD a version of the D-statistic that does not require genotype calling but is still able to utilize all the information and allows for multiple individuals for each of the four groups. Using both simulations and real data we evaluate the method's power and false positive rate using different topologies. For example ( African ( CEU ( Han Chinese, Native Americans) ) ) is used to show a wrong phylogeny and the evidence of admixture from CEU to Native Americans.
We show that the weighted D-statistic has more power than the previous method of sampling only one base per site and is still powerful with the use of low-coverage data. For medium and high depth data the method has similar power to that of perfectly inferred genotypes. In addition the methods can be use even in the presence of high error rates which makes it especially appropriate for ancient DNA.
Inference of distribution of fitness effects: impact of beneficial mutations

Paula Tataru¹, Maéva Mollion¹, Sylvain Glemin² and Thomas Bataillon¹
¹Bioinformatics Research Centre, Aarhus University, Aarhus, 8000, Denmark
²Evolutionary Biology Centre, Uppsala University, Uppsala, 752 36, Sweden

Abstract:
New mutations are the ultimate source of heritable variation, and their fitness effects determine the possible evolutionary trajectories a population can follow. The relative frequencies of the fitness effects of new mutations are best specified through a distribution of fitness effects (DFE) that spans deleterious, neutral and beneficial mutations. As such, the DFE is key to several aspects of the evolution of a population and inference of DFE from patterns of polymorphism and divergence has been a longstanding goal of evolutionary genetics. The large amount of datasets currently available enables us to learn and compare the DFE in numerous species. Understanding what determines variation in DFE from species to species can foster a deeper understanding of the forces underlying the process of evolution.

Here, we revisit the problem of reliably estimating the DFE from site frequency spectrum (SFS) and divergence data. We developed a hierarchical probabilistic framework that extends significantly available methods, and using simulation studies, we investigate and question a widespread assumption: beneficial mutations contribute only negligibly to the SFS data. Current methods invariably assume that mutations present in the SFS are either neutral or deleterious. Consequently, positive selection is only inferred when SFS data can be complemented with divergence data. We show that this assumption is theoretically unfounded and that the DFE of beneficial mutations can often be inferred from SFS data alone. This opens the doors to the analysis of less popular species, where divergence data is often not available. Additionally, we show that when the beneficial mutations in the SFS are not modeled, the resulting DFE estimation can be substantially biased. We illustrate these points by using both our newly developed framework, but also one of the most widely used inference methods available.
Large-scale prediction of RNA-RNA interactions and siRNA off-targets

Ferhat Alkan, Anne Wenzel, Oana Palasca, Peter Kerpedjiev, Anders F. Rudebeck, Peter F. Stadler, Ivo L. Hofacker, Jan Gorodkin

Abstract:
Intermolecular interactions of non-coding RNAs form the backbone of gene regulation events and predicting these interactions is of great interest. Identification of RNA interactomes promises to prove useful in several areas, such as integrative miRNA and gene expression analysis, non-coding RNA functional studies, and interpretation of RNAi application outcomes. However, in order to fulfill this task, large-scale genome- and transcriptome-wide predictions are required. To meet the demands, we implemented an efficient method, RIsearch2, that uses suffix arrays to identify perfect complementary seed regions (including G-U wobble pairs) between given query and target sequences and extends these possible interactions on both ends using a simplified energy model. A screen of ~2600 human miRNAs on the whole repeat-masked human genome, which takes about ~6h using 16 threaded cores, exemplifies the large scale capability of RIsearch2. This is orders of magnitude faster than current other available methods. Furthermore, we constructed an efficient siRNA off-target discovery pipeline by using RIsearch2 interaction predictions in combination with accessibility and expression abundance information of binding sites. We show that this pipeline can accurately predict not only the individual off-targets of siRNAs but also their overall off-targeting potential, which in fact might further influence their repression efficiency.
Species-independent identification of known and novel recurrent genomic entities in multiple cancer patients

Jose Mg Izarzugaza

Summary
The discovery of viruses and other disease-causing pathogens from high throughput sequencing data often requires that taxonomic annotation occurs prior to association to disease. Although this bottom-up approach is effective in some cases, it fails to detect novel pathogens and remote variants not present in reference databases. We propose an alternate approach utilizes sequence clustering for the identification of nucleotide sequences that co-occur across multiple sequencing data instances. Thus, not limited to reported species.

We applied the workflow to 686 sequencing libraries from 252 different cancers and 56 controls. We used our pipeline to associate recurrent sequences to the onset of the disease but also to the use of common laboratory kits to identify common methodological or technical artifacts sourcing erroneous conclusions, as we have observed in the recent literature. We provide examples of identified inhabitants of the healthy tissue flora as well as experimental contaminants.

Scientific Justification:
The discovery of viruses and other disease-causing pathogens from high throughput sequencing data often requires that taxonomic annotation occurs prior to association to disease. Although this bottom-up approach is effective in some cases, it fails to detect novel pathogens and remote variants not present in reference databases. We propose an alternate approach utilizes sequence clustering for the identification of nucleotide sequences that co-occur across multiple sequencing data instances. Thus, not limited to reported species.

We will describe an analysis where we applied our workflow to 686 sequencing libraries from 252 different cancers and 56 controls in the context of the Cancer and Pathogens project funded by the GenomeDenmark platform, where we have used our pipeline to associate recurrent sequences to the onset of the disease with satisfactory results in the identification of known and novel oncoviruses.

A number of recent articles associating viruses and disease were rapidly disregarded as contamination stemming from the use of laboratory contaminants. We reported the erroneous identification of ATCV-1 (Kjartansdóttir et al, PNAS 2015). Concomitant to the universalisation of sequencing technologies, the identification of secondary species poses a threat to the validity of the conclusions drawn if caution is not exerted. Our method correlated recurrent sequences and common laboratory kits, deeming these observations as methodological/technical artefacts. Our intention is to make the scientific community aware of this extremely relevant issue.
Tandem mass spectrometry peptide fragment ion prediction by Hidden Markov Models

Jan Christian Refsgaard1, Christian Kelstrup1, Aybuge Altay2, Christian Grønbæk3, Jesper Velgaard Olsen1 and Lars Juhl Jensen1

1NNF Center for Protein Research, University of Copenhagen, Denmark
2Izmir University, Turkey
3University of Copenhagen, Denmark
May 1, 2016

Tandem Mass Spectrometry (MS/MS) peptide fragmentation is a complex process involving multiple competing fragmentation pathways. Accurate modeling of peptide fragmentation is essential for designing better peptide spectra match (PSM) algorithms. Here we propose a novel algorithm for inferring theoretical peptide MS/MS fragmentation spectra. The data-driven approach is made feasible by the increasing availability of MS/MS data. The varying lengths of peptides prove challenging for existing algorithms. Typical solutions include: 1) The use of a different model for each peptide length, hereby only training on a subset of the available data; an example of this approach is the Random Forest used in MS2PIP[1]. 2) Extracting a fixed number of features from the sequences; this approach is used in the Artificial Neural Network employed in PeptideArt[3] and in the boosting algorithm employed in PepNovo+[2]. We instead make use of Hidden Markov Models, which are invariant to the sequence length. There is thus no need to train multiple models or to extract a fixed number of features from variable lengths of peptides. Our preliminary benchmark shows average Pearson and Spearman correlation coefficients of 0.651 and 0.645, respectively, between observed and predicted y-ions.

References


Chromosomal transcriptional initiation walks reflect three dimensional chromatin organisation.

Sarah Rennie & Robin Andersson

The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

Abstract:
The genome is organised into non-random and highly complex structures within the nucleus, which exhibit variable stability across different types of cells. Domains of interacting initiation events fall within so-called topologically associated domains, underlying the strong positional effect of enhancers and promoters localised linearly on the DNA to specific regions of a chromosome. Such domains themselves fall into proximity on a higher order level, providing a potential mechanism for distal interactions between units separated by many mega-bases. Whilst chromatin conformation capture technologies have come some way into identifying proximal regions of DNA, both cell type coverage and resolution is still limited, thwarting the ability to experimentally detect long range proximities and computationally predict pairs of interacting transcriptional units.

We calculated simultaneous random walks across the initiation landscape of tissues over developmental stages in mouse, generating intriguing underlying latent structures reflecting the correlated positional effects of initiation events, separated from the layer of transcription attributable to a gene's regulatory program independent of its genomic position. Intersecting walks between tissues then generates high resolution maps of active co-regulation and show striking similarities to chromatin organisation as identified by Hi-C. We compare our walks with a wealth of publicly available ChIP-seq data, demonstrating clear shifts between putative domain boundaries and compartments. We then identify active topologically associated domains specific to each of nine tissues across a variety of developmental stages at high resolution, pinpointing putative enhancer-promoter interactions, both within and between domains and allowing for comparisons at an unprecedented scale. Thus, this work reveals a fascinating picture of how higher order genomic activities vary across developmental time.
Cataloging the landscape of RNA bioinformatics tools

Authors:
RNA tools and software consortium

Affiliations:
Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria
University of Freiburg, Freiburg, Germany
University of Jena, Jena, Germany
University of Leipzig, Leipzig, Germany
European Bioinformatics Institute, Hinxton, UK
International Institute of Molecular and Cell Biology, Warsaw, Poland
Paris-Sud University, Orsay, France
École polytechnique, Palaiseau, France
Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Frederiksberg C, Denmark

Abstract:
During the past decades, non-coding RNAs have gained increasing attention and RNA bioinformatics have played a pivotal role in this. Various computational methods have been developed to solve a wide range of tasks, from the analysis of RNA structure to transcriptomes. With the wealth of tools and resources available, it can be challenging to find the one best-suited for a specific research problem. To address this issue, we introduce the RNA tools registry along with an ontology that categorizes tools by their function and application domain, as well as mapping relationships between them. Around 400 tools are currently included in the registry (version 0.1) and categorized in about 50 fine-grained subclasses. For instance, around 80 tools fall under the topic “RNA secondary structure”, whose function is further subcategorized in prediction, visualization, alignment, and sequence design, among others. This exemplifies how a non-expert user can identify the relevant tools for the task at hand. To make the search easily accessible, the RNA tools registry will be imported into the ELIXIR Tools and Data Services Registry (available at http://bio.tools). Moreover, this community effort can serve as a guideline for communities from different areas to create such a tool registry.
biotoolsXSD:
description model for bioinformatics software

biotoolsXSD is a formalised XML schema (XSD) that defines a general purpose description model for bioinformatics software. It is used by the ELIXIR Tools & Data Services Registry (https://bio.tools). Bioinformaticians routinely use a large and diverse set of tools and data, and must organise, find, understand and compare – and then use and connect - the available resources. These tasks rely on consistent, and ideally machine-understandable descriptions. The need - filled by biotoolsXSD - is for a model that defines a minimum standard for the consistent description of a broad range of resources.

EDAM:
controlled vocabulary for bioinformatics

EDAM is an ontology of well established, familiar concepts that are prevalent within bioinformatics, including types of data and data identifiers, data formats, operations and topics. EDAM aims to unify semantically the bioinformatics concepts in common use, provide curators with a comprehensive controlled vocabulary that is broadly applicable, and support new and powerful search, browse and query functions. EDAM is a simple ontology - essentially a set of terms with synonyms and definitions - organised into an intuitive hierarchy for convenient use by curators, software developers and end-users.
ELIXIR:
a distributed infrastructure for biological information

ELIXIR’s mission is to construct and operate a sustainable infrastructure for biological information in Europe, supporting life science research and its translation to medicine, agriculture, bioindustries and society. The Danish ELIXIR Node focus is on addressing the growing need of the life science community to access bio-molecular databases, integrate data and use interoperable tools for data analysis and interpretation. ELIXIR Denmark leads the community-driven development of the ELIXIR Tools and Data Services Registry (bio.tools). The Registry provides essential scientific and technical information about analytical tools and data services for bioinformatics. ELIXIR DK also provides supercomputer and secure cloud infrastructure.

ELIXIR EXCELERATE:
tools interoperability & registry

ELIXIR-EXCELERATE is an EU funded project to accelerate the implementation of ELIXIR, develop and connect resources and services, and build bioinformatics capacity across Europe. The life science community has an increasing need to access bio-molecular databases, integrate data and use interoperable tools for data analysis and interpretation. The “tools interoperability & registry” Work Package 1, led by ELIXIR Denmark, focuses on biosoftware discovery and interoperability. The cornerstone is a comprehensive Tools and Data Services Registry (bio.tools), involving service monitoring, resource integration, interoperability aspects, and community centred benchmarking efforts.
bio.tools: tools & data services registry

bio.tools is a registry of bioinformatics software information, sustained by a community-driven curation effort, tailored to local needs and shared amongst a network of engaged partners. Life sciences yield huge data sets that underpin vital scientific discoveries. In support, a plethora of databases and tools are used, in technically complex and diverse forms, across a spectrum of scientific disciplines. The corpus of information for these resources is fragmented across the Web, with much redundancy, and has lacked a common information standard. The outcome is that scientists often struggle to find, understand, compare and use the best resources for the task at hand. bio.tools addresses this problem.

ComputerOME Secure Cloud: scalable secure compute environment

ComputerOME Secure Cloud is a highly specialised offering from the Danish National Supercomputer for Life Sciences. Secure Cloud eliminates problems and barriers for researchers dealing with highly sensitive data distributed across many countries by providing out of the box. Secure Cloud is powered by the infrastructure consisting of 16048 CPU cores with 92 TeraBytes of memory, connected to 3 PetaBytes of High-performance storage, and with a total peak performance of more than 483 TeraFLOPS (483 million million floating-point operations per second).
Rapid detection of drug-resistance – a successful industry-academic cooperation

Katja Einer-Jensen, Martin Simonsen.
QIAGEN Aarhus A/S, Aarhus C, Denmark.

Abstract:
At QIAGEN Aarhus we are involved in multiple international research projects within the field of bioinformatics. Through these projects we keep in touch with the scientific community and develop innovative solutions for our customers. One particular successful research project which has given rise to several interesting solutions is the EU funded PATHSEEK project.

In PATHSEEK, four partners sought to enable personalized treatment of infected patients in clinical settings based on detailed characterization of the infectious agent using Next Generation Sequencing (NGS). Efforts were focused on two challenges:
Establishment of a methodology to sequence the entire genome of pathogens extracted directly from clinical specimens.
Development of an automated bioinformatics solution.

As the result of close dialog with the partners, QIAGEN’s contribution was the development a bioinformatics analysis pipeline which enables automatic identification of the present pathogen(s), identification of pathogen specific variants known to influence drug resistance, and lastly, genotyping of the identified pathogen(s).


The solution is easy to use and delivers unified reporting and documentation in condensed views, aimed at novice bioinformatics users. The various outputs formats in addition to a simple report include overview tables and data visualization, which offers experienced users the option for a detailed review of the findings.

The decision support tool presented here has the potential to revolutionize patient management by delivering all the data required for truly personalized infectious disease treatment when compared to the currently practiced empirical treatment of MTB.
Providing NGS data analysis as a service – XploreRNA

Jesper C. Tholstrup, Asli Özen, Dhany Saputra, Man-Hung E. Tang, Jacob B. Axelsen, Yuan Mang, Iben Plate, Niels Tolstrup

(Corresponding author jet@exiqon.com)

Exiqon, like many other scientifically founded companies, is heavily reliant upon contributions from the academic world when offering new products and services. These contributions come in many shapes and sizes but they are essential and in many ways, they define the direction of research – and thus the needs of our customers.

Not too long ago Exiqon launched a new service; XploreRNA. XploreRNA is a cloud-based solution for NGS analysis and design of custom oligonucleotide products for downstream validation experiments and functional analysis.

XploreRNA uses both open source software, provided by academia, and software developed at Exiqon within bioinformatics and IT. XploreRNA is fully automated and requires no manual interaction from upload of raw sequencing data by the customer to delivery of analysis results and summary report. The complexity of the NGS analysis itself combined with the complexity of underlying logistics and distribution framework means that each component has to be very stable.

In this session, we will try to highlight some of the experiences that bioinformatics at Exiqon have gained from using open source software in a production environment. The focus will be on software related to analysis of NGS results but the concepts are easily generalized.

The goal is to highlight some aspects of open source software, which are very important when working in the industry. These topics include testing, licenses, interface, dependencies, error handling, and documentation.
Proteogenomic analyses for discovery of bi-specific chimeric antigen receptor targets

Christian Simon¹, Mike Bogetofte Barnkob², Lars Rønn Olsen²,³

¹Disease Systems Biology, Novo Nordisk Center for Protein Research, University of Copenhagen, Copenhagen, 2200, Denmark.
²MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, OX3 9DU, United Kingdom. ³Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, 2800, Denmark

Abstract:
Current clinical successes of chimeric antigen receptor (CAR) T cells have been achieved by targeting single, malignancy-specific surface molecules. In order to broaden the treatment applicability to more cancer types, progress is being made in developing bi-specific CARs that require binding to several ligands before lymphocytes are activated. Such technical advances hold the potential to increase the specificity of the therapy and thereby decrease off-target effects. Selecting good bi-specific CAR targets is a balance between specificity and sensitivity of tumor cell targeting - High specificity is required for clinical safety, and high sensitivity is required for efficient clearance of the targeted cancer type.

Well-defined cell surface protein expression profiles are needed to facilitate target selection. Defining surface protein expression profiles for CAR therapy in silico require information about expression of a large number of surface proteins on a large number of cells at different states and differentiation stages. At present, no high-throughput technique for measuring expression of large numbers of surface proteins exists. However, progress is being made with emerging technologies, such as chip cytometry and mass cytometry. In addition, surface molecule expression on individual cells has been measured at low rates using immunohistochemistry or flow cytometry for decades, and vast amounts of cell-specific expression has been published. Combined, these resources form a rich, but unstructured source of data and information.

To facilitate the definition of unique surface molecule profiles, we have collected and organized large amounts of protein expression data on human hematopoietic cells from the cytometric methods as well as the primary literature. For the latter, we employed text mining techniques for article classification and subsequently extensive manual curation. We coupled these data with analysis of large-scale transcriptomics data for the surface proteins in order to assemble a data foundation for deep characterization of cell surface profiles. The resulting database contains expression of 457 surface proteins across 279 hematopoietic cells. We then developed algorithms for data mining to define unique protein expression profiles for highly sensitive and specific CAR targeting. Ongoing efforts will expand the database to contain surface protein expression for cells in all human tissues, as well as experimental validation of potential CAR targets.
Employing complementary ions for deconvolution of mixture tandem mass spectra

Vladimir Gorshkov1, Stéphanie Yuki Kolbeck Hotta1, Thiago Verano-Braga1,2, Frank Kjeldsen1

1Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, 5230, Denmark
2Department of Physiology and Biophysics, Federal University of Minas Gerais, Belo Horizonte – MG, 31270-901, Brazil

Abstract:
Mixture fragmentation spectra produced by simultaneous fragmentation of several coisolated peptide precursors are recognized as a limitation in proteomics since they complicate the identification of peptide sequences [1]. Complementarity of b- and y-ions can be efficiently used to deconvolute mixture fragmentation spectra, significantly increasing the number of identified and quantified peptides and proteins [2]. Due to the large interest in the investigation of proteoforms and detection of unmapped changes in protein sequence, de novo sequencing methods are getting increased attention. However, de novo algorithms, due to much larger sequence search space, compared to the database search, are in general more sensitive to lower mass spectrum quality resulting from peptide precursor coisolation and thus more prone to false identifications [3].

The deconvolution approach, implemented as a node in Thermo Proteome Discoverer 2.x, matched complementary b-, y- ions to each precursor peptide mass, which allowed the creation of virtual spectra containing sequence-specific fragment ions of each coisolated peptide [2]. The effect on quantitative and qualitative performance was evaluated first with database search strategy using dimethyl labeled HeLa lysate. Four popular de novo sequencing programs (PEAKS v7.0, pNovo+ v1.3, pepNovo+ v3.1 and Novor v1.1) were used to assess the impact of mixture spectra on de novo sequencing. The approach was first tested on artificial mixture spectra with controlled abundance ratio between mixture components, including the comparison with database search method, and, consequently, was applied to real proteomics data obtained from HeLa lysate and the venom of Tityus serrulatus.

Significant improvement is observed when our approach was applied to shotgun proteomics strategy. The number of identified peptide-spectrum matches (PSMs), peptides and proteins increased, up to 70%, 40% and 20%, respectively, at the 0.01 FDR level. More PSMs and peptides were identified compared to the ion trap-based method. Enhanced identification resulted in corresponding 10%, 15% and 10% improvement in quantified PSMs, peptides, and proteins on the same raw data while the accuracy of the quantification was not affected.

Mixture spectra deconvolution was proved to be highly important for de novo sequencing. Using artificial mixture spectra with various abundance mixture ratios, we observed the severe negative effect on the true identification rate with the presence of coisolated fragments ions in the mass spectrum. The observed degradation in identification rate was more prominent, compared to database search methods, due to the larger search space tested by de novo approaches. Deconvolution resulted in equally efficient identification rates but increased an absolute number of correctly sequenced peptides (20–35% for the HeLa lysate). About 5 – 10% of correct sequences were identified only using unprocessed spectra; however, the number was lower than those obtained by our mass spectral deconvolution. Some of the missing peptide identifications could be accounted for peculiarities of candidate peptide scoring or the employed deconvolution method. Venom samples of Tityus serrulatus show only a moderate increase in the number of reported sequences. The low number of coisolations and less informative fragmentation of non-tryptic peptides might be the primary reasons for the moderate effect, the latter of which could be improved by advancing de novo sequencing programs.

Scientific workflows for combining MS and MS/MS data and improving mass measurement accuracy in proteomics

Arzu Tugce Guler1; Yassene Mohammed1; Dmitrii Travin2; Magnus Palmblad1;
1 Leiden University Medical Center, Center for Proteomics and Metabolomics, Leiden, Netherlands;
2 Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

Abstract:
Scientific workflows are important tools for automation and excellently suited for integrative mass spectrometry and proteomics data analyses. Connecting inputs and outputs of different modules is not sufficient for complete automation of an analysis. Most multi-step processes require decisions to be made by an agent and is often data-dependent. Knowledge level information processing is one approach for automating data analyses with conditional constructs in one or more steps. On a lower level, there should be actions/parameters to be done/used for the conditions defined within a particular domain. Formalized knowledge, ontologies and controlled vocabularies, make such condition-decision making feasible within scientific workflows. We here demonstrate how to combine MS and MS/MS data in order to improve mass measurement accuracy.

The workflow is created in Taverna Workbench, a scientific workflow management system. The mass spectrometry data are first converted to mzXML. One advantage is that the built-in services of Taverna works well with XML. Secondly, the Proteomics Standards Initiative has produced a controlled vocabulary (PSI-MS CV), which is used to embed metadata on how the datasets were acquired. This information is later used for selecting the correct parameters at each step. Another type of “formalized knowledge” is used for visualization. As all datasets were encoded with anatomical (tissue/cell type) origin, the spatial distribution of the protein and gene expression could be brought together and compared visually using an anatomical ontology, here for zebrafish.

The scientific workflow consists of 8 modules, each performing a specific operation within a well-defined scope. By default, the workflow reads the metadata and selects the most optimal parameters for the analysis. However, the expert user can also apply specific parameters, overriding the defaults. The first module checks all necessary inputs that will be necessary for the analysis and whether the user-specified parameters are valid. If the latter is not the case, the most optimal parameters are chosen for the data based on the metadata. The next module executes an X!Tandem database search on the MS/MS data. An alignment module then aligns MS-only and MS/MS data using a genetic algorithm to fit a piecewise function mapping accurate masses and identified peptides. The retention times are then modified accordingly in the pepXML file. PeptideProphet is then run and a probability value assigned to each peptide-spectrum match. Each MS-only mzXML file is then calibrated using identified peptides in the pepXML file as internal calibrants. Finally, each identified peptide is quantified in each sample/file. When the sample information is provided, the results can be visualized spatially to reveal biologically interesting differences in abundance levels, or unexpected correlation between gene and protein expression. Spatial information can be provided as physical coordinates or anatomical entities. For the latter, we used the Zebrafish Anatomical Ontology to align data with a novel Web-based tool for visualizing quantitative gene and/or protein expression in zebrafish and carp model systems.
Convolutional LSTM Network for analysis of iCLIP data


Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense M, Denmark.

Abstract:
Splicing of precursor mRNA (pre-mRNA) into a mature mRNA is an essential step in the formation of functional proteins in all mammalian cells. Recognition of splicing regulatory elements (SREs) by RNA-binding proteins (RBPs) is essential for correct splicing of functional mRNAs. Mutations may disrupt appropriate splicing and introduce severe pathological conditions. Evidence show that these pathological conditions may be treated by modulating the activity of SREs using splice-switching oligonucleotides (SSOs) that bind to SREs. Identifying functionally important SREs is therefore critical to developing therapies for these severe pathologies. Unfortunately, many of the SREs are as of yet poorly defined.

RNA-sequences derived from Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) experiments allows identification of in vivo protein binding sites with nucleotide resolution, but there are several sequence biases associated with this approach as well as false-positives. Motif analysis using previous methods such as HOMER and MEME can often lead to motifs that can only be found in a minority of the sequences because SREs have degenerate binding motifs that can be difficult to extract.

Deep learning has improved data processing in various fields, including biological sequence analysis and feature extraction from sequences. In this project, we aimed to create a neural network classifier that highlights putative motifs of SRE-binding proteins within RNA-sequences and use the final predictive score to estimate the affinity of the motif. We used iCLIP experiments targeting SRSF5 and hnRNPA1 was analyzed and classified by a convolutional LSTM (long short term-memory) neural network. SRSF5 is a splicing activator that binds enhancer elements and induces exon inclusion. hnRNPA1 is a splicing repressor that inhibits exon inclusion by binding splicing silencers. The network used shallow convolutional layers to identify and enhance features or important patterns within the RNA-sequences. The LSTM layers scanned the feature-enhanced RNA-sequences in a bidirectional manner and assigned contextual importance to relevant base patterns. The network accuracy was determined by classifying a restricted number of the iCLIP derived RNA-sequences previously unknown by the network. The best run achieved 87.11 % accuracy in the classification task. The network was further tested by predicting binding affinity to RNA-sequences that were previously examined in affinity purification pull down of hnRNPA1 and SRSF5 using biotinylated RNA oligonucleotides (pull-down).

Optimizing this approach to motif analysis from iCLIP experiments may provide a strong tool for identifying novel SSO targets and thereby treating diseases caused by abnormal splicing. Also, it could help define SREs and binding sites of RNA-binding proteins in heterologous contexts and help us differentiate between active and inactive SREs. This may allow us to better predict the impact of sequence variations on splicing patterns.
Clustering RNA sequences with an evolutionary conserved secondary structure using graph based motifs

Alexander Junge¹,², Milad Miladi³, Fabrizio Costa³, Stefan E. Seemann¹,², Jakob Hull Havgaard¹,², Jan Gorodkin¹,², and Rolf Backofen¹,³,⁴

¹Center for noncoding RNA in Technology and Health, University of Copenhagen, Denmark,
²Department of Veterinary Clinical and Animal Sciences, University of Copenhagen,
³Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany,
⁴Center for Biological Signalling Studies (BIOSS), University of Freiburg, Germany

*These authors contributed equally to this work. ‡To whom correspondence should be addressed: gorodkin@rth.dk or backofen@informatik.unifreiburg.de

Abstract:
A crucial step towards understanding RNA function is through secondary structure clustering. In order to identify recurrent RNA structures, it is essential to exploit their evolutionary conservation. We present RNAscClust, a clustering approach that improves upon existing strategies that do not model structure conservation. RNAscClust finds groups of paralogous RNAs that exhibit structural similarities by taking compensatory base changes into account. The input to RNAscClust is a set of multiple structural alignments of RNA sequences. Each alignment contains a paralog sequence aligned to its orthologs. After computing minimum free energy structures for each paralog that comply with base pairs conserved in its orthologs, the paralogs are clustered using a graph kernel based strategy. This approach identifies common structural features across conserved RNAs. The usage of locality sensitive hashing techniques furthermore yields a runtime complexity linear in the number of alignments. Based on a benchmark set derived from the Rfam database, we demonstrate that the accuracy of RNAscClust clearly benefits from an increasing degree of compensatory base changes in the alignments. We furthermore present initial results for clustering RNA structures conserved across vertebrates. RNAscClust is available at:
http://www.bioinf.unifreiburg.de/Software/RNAscClust/
A Multi-Tissue Transcriptome Assembly and Annotation of the Shark *Squalus Acanthias*: A Molecular Tool to Osmoregulation Research

Andres Chana-Muñoz¹, Peter A Andreasen², Christian Bendixen¹, Agnieszka Jendroszek², Frank Panitz¹

¹Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark
²Department of Molecular Biology and Genetics, Aarhus University, 10 Gustav Wieds Vej, 8000 Aarhus C, Denmark

Abstract:
The spiny dogfish shark (*Squalus acanthias*) is one of the most commonly used cartilaginous fishes in biological research, especially in the fields of nitrogen metabolism, ion transporters and osmoregulation. Nonetheless, transcriptomic data for this organism is scarce. In the present study, a multi-tissue RNA-seq experiment and a *de novo* transcriptome assembly was performed in four different *Squalus acanthias* tissues (brain, liver, kidney and ovary), being up to date the first next-generation sequencing experiment performed in this shark species.

Around 600 million reads were generated. Using Trinity software, these reads were assembled in a per tissue manner and as a combination of the four and further annotated using Trinotate pipeline. BUSCO analysis showed almost 87% of completeness in the combined assembly. After filtering and among the combined assembly, there were 362,690 transcripts with an N50 of 1,302bp which were included in 289,515 Trinity genes. A total of 123,110 proteins longer than 60 amino acids were predicted and 78,679 and 83,164 of those had a significant (1e-05) hit in SwissProt and Uniref90 databases respectively. Additionally, 61,215 proteins aligned to at least one protein domain, 7,208 carried a signal peptide and 15,971 possessed at least one transmembrane region. Based on the annotation 81,582 transcripts were assigned to gene ontology terms and 42,078 belonged to any of the known clusters of orthologous groups (eggNOG). Transcriptome annotation revealed the presence of genes involved in urea synthesis and urea and water transport, all of them crucial in osmoregulation. More importantly, a great number of these genes identified represent new sequence data for spiny dogfish and even for the elasmobranch clade improving the current possibilities of osmoregulation research in cartilaginous fishes.

The transcriptome assemblies and the derived annotations generated in the present work will greatly improve the sequence information for *Squalus acanthias* for which a comprehensive transcriptomic characterization was limited. The new data generated will support the ongoing research in cartilaginous fishes providing a new molecular tool of great value for biological research in particular for this animal model.
Background:
Lung effusion is a common complication for patients hospitalized with pneumonia. Fast and adequate treatment is important for fast recovery and survival, and it relies solely on rapid characterization of the pathogen causing the disease. Unfortunately, the clinical microbiology practice is labor intensive and it can take days or even weeks to determine the pathogen.
In this study we aim to develop a new method to improve and accelerate the detection and characterization the infectious pathogen in patients suffering from lung effusion by the use of direct sequencing on clinical samples.

Methods:
At the hospital we collect pleural fluid by drainage from patients with pleural effusion. DNA is purified directly from the clinical samples, and several steps are used to remove the human contamination prior to sequencing. After sequencing the raw reads are trimmed and then mapped against predefined databases by the use of the MGmapper software (https://cge.cbs.dtu.dk/services/MGmapper/). Firstly the raw reads are mapped against a human genome, thereby removing any human contamination. The remaining unmapped reads are mapped against several databases containing complete and draft bacterial, fungal protozoan and viral genomes. The remaining reads that do not match anything will be mapped against the complete nucleotide database.
In order to obtain detailed characteristics of the pathogenic strain, the raw reads are mapped against databases containing virulence and resistance genes. Finally a report will be generated containing information for diagnosis to be made. This will include abundance profiles of the strains, virulence factors and resistance markers.
Cataloging the landscape of RNA bioinformatics tools

Authors:
RNA tools and software consortium

Affiliations:
Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria
University of Freiburg, Freiburg, Germany
University of Jena, Jena, Germany
University of Leipzig, Leipzig, Germany
European Bioinformatics Institute, Hinxton, UK
International Institute of Molecular and Cell Biology, Warsaw, Poland
Paris-Sud University, Orsay, France
École polytechnique, Palaiseau, France
Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Frederiksberg C, Denmark

Abstract:
During the past decades, non-coding RNAs have gained increasing attention and RNA bioinformatics have played a pivotal role in this. Various computational methods have been developed to solve a wide range of tasks, from the analysis of RNA structure to transcriptomes. With the wealth of tools and resources available, it can be challenging to find the one best-suited for a specific research problem. To address this issue, we introduce the RNA tools registry along with an ontology that categorizes tools by their function and application domain, as well as mapping relationships between them. Around 400 tools are currently included in the registry (version 0.1) and categorized in about 50 fine-grained subclasses. For instance, around 80 tools fall under the topic “RNA secondary structure”, whose function is further subcategorized in prediction, visualization, alignment, and sequence design, among others. This exemplifies how a non-expert user can identify the relevant tools for the task at hand. To make the search easily accessible, the RNA tools registry will be imported into the ELIXIR Tools and Data Services Registry (available at http://bio.tools). Moreover, this community effort can serve as a guideline for communities from different areas to create such a tool registry.
MuPeXI: A tool for prediction of neo-epitopes from tumor sequencing data

Anne-Mette Bjerregaard, Morten Nielsen¹,³, Sine R. Hadrup⁴, Zoltan Szallasi¹,², Aron C. Eklund¹

¹ Center for Biological Sequence Analysis, Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark, ² Computational Health Informatics Program (CHIP), Boston Children’s Hospital, Boston, MA 02115, Harvard Medical School, Boston, USA, ³ Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina, ⁴ Section for Immunology and Vaccinology, National Veterinary Institute, Technical University of Denmark, Copenhagen, Denmark

Abstract:
Recent successes in several types of immunotherapy have demonstrated that exploitation of a patient’s own immune system is a promising strategy to eliminate cancer. Personalization of immunotherapies such as cancer vaccines and adoptive T-cell therapy will depend on identification of patient-specific neo-epitopes that can be specifically targeted. For this reason, there is a need for a tool to analyze sequencing data, extract the tumor specific peptides and predict which are likely to be immunogenic. Here we present MuPeXI, the Mutant Peptide Extractor and Informer, a pipeline that extracts tumor mutation specific peptides and provides relevant information for peptide selection. MuPeXI takes several mutation types into account, ensuring a larger output of tumor specific peptides, including single nucleotide variations (SNVs) and indels with or without frameshift. Standard practice for immunogenicity estimations are based on theoretical considerations guided by predicted features such as binding to patients HLA class I molecules, but we are currently pursuing a qualified calibration of MuPeXI based on high-throughput T cell response experiments. Tumor specific peptide extraction with MuPeXI will enable testing of neo-epitope specific T-cell activation from a population of tumor infiltration T-cells, optimizing adoptive T-cell therapy, or facilitates the optimal combination of peptides for personalized cancer vaccines to boost a patient specific immune response.
Scientific workflows are important tools for automation and excellently suited for integrative mass spectrometry and proteomics data analyses. Connecting inputs and outputs of different modules is not sufficient for complete automation of an analysis. Most multi-step processes require decisions to be made by an agent and is often data-dependent. Knowledge level information processing is one approach for automating data analyses with conditional constructs in one or more steps. On a lower level, there should be actions/parameters to be done/used for the conditions defined within a particular domain. Formalized knowledge, ontologies and controlled vocabularies, make such condition-decision making feasible within scientific workflows. We here demonstrate how to combine MS and MS/MS data in order to improve mass measurement accuracy.

The workflow is created in Taverna Workbench, a scientific workflow management system. The mass spectrometry data are first converted to mzXML. One advantage is that the built-in services of Taverna works well with XML. Secondly, the Proteomics Standards Initiative has produced a controlled vocabulary (PSI-MS CV), which is used to embed metadata on how the datasets were acquired. This information is later used for selecting the correct parameters at each step. Another type of “formalized knowledge” is used for visualization. As all datasets were encoded with anatomical (tissue/cell type) origin, the spatial distribution of the protein and gene expression could be brought together and compared visually using an anatomical ontology, here for zebrafish.

The scientific workflow consists of 8 modules, each performing a specific operation within a well-defined scope. By default, the workflow reads the metadata and selects the most optimal parameters for the analysis. However, the expert user can also apply specific parameters, overriding the defaults. The first module checks all necessary inputs that will be necessary for the analysis and whether the user-specified parameters are valid. If the latter is not the case, the most optimal parameters are chosen for the data based on the metadata. The next module executes an X!Tandem database search on the MS/MS data. An alignment module then aligns MS-only and MS/MS data using a genetic algorithm to fit a piecewise function mapping accurate masses and identified peptides. The retention times are then modified accordingly in the pepXML file. PeptideProphet is then run and a probability value assigned to each peptide-spectrum match. Each MS-only mzXML file is then calibrated using identified peptides in the pepXML file as internal calibrants. Finally, each identified peptide is quantified in each sample/file. When the sample information is provided, the results can be visualized spatially to reveal biologically interesting differences in abundance levels, or unexpected correlation between gene and protein expression. Spatial information can be provided as physical coordinates or anatomical entities. For the latter, we used the Zebrafish Anatomical Ontology to align data with a novel Web-based tool for visualizing quantitative gene and/or protein expression in zebrafish and carp model systems.
Validation with LNA™-enhanced qPCR primers and Functional Analysis with Antisense LNA™ GapmeRs

Asli Özen, Jesper Tholstrup, Ditte Andreasen, Johnathan Lai, Niels Frandsen, Kim Barken, Niels Tolstrup (Corresponding author aoz@exiqon.com)

Abstract:
Next-generation sequencing (NGS) data analysis platforms such as XploreRNA, allows a rapid means of studying differential gene expression and provides enormous information on the genetic variants, expression levels, composition of the transcriptome, functional annotations and comparison between the given conditions. In order to achieve validity and efficiency in such a complex analytical system, well-established experimental designs principles should be applied such as; randomization, replication, control and validation [1]. Validation term is applied from the microarray expression literature as one of the most important parts of expression analysis and it also applies to RNA-seq experiments. Differentially expressed genes identified by RNA-seq can be validated using quantitative PCR (qPCR) [2, 3, 4] and it is required by many journals for publication. Although the differential expression analysis results in certain biological conclusions, researchers usually perform functional analysis using knockdown experiments to gain more detailed biological knowledge about particular targets. Therefore, Exiqon couples the RNA-seq analysis from XploreRNA with the downstream validation and functional analysis steps by offering qPCR primer sets for validation, and antisense LNA™ gapmers for potent knockdown.

In this poster we present how the researchers can use Exiqon’s online tools to obtain LNA™-enhanced qPCR assays and antisense LNA™ gapmers. Exiqon’s online oligonucleotide design tools make use of an empirically derived algorithm developed by our Bioinformatics Department. The design algorithm optimizes LNA based oligonucleotides against more than 30 parameters to identify the most effective and specific ones. Some of the important parameters we optimize against are melting temperature, GC content, self-complementarity, LNA composition and off-target analysis for achieving specificity. These design tools support all the organisms that are supported by XploreRNA platform, however, currently the off-target analysis is only performed for human, mouse and rat. Some tools are provided with advanced option steps, where the researcher is allowed to select more specific parameters and make a custom design. Design tools explained here can be accessed through http://www.exiqon.com/custom-LNA-qPCR, and http://www.exiqon.com/gapmer.

References:
Bacterial communities hitching a hike – a guide to the river system of the Red River, Disko Island, Greenland


1Center for Biosustainability, Technical University of Denmark, Hoersholm, Denmark
2Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark
3Center for Permafrost (CENPERM), University of Copenhagen, Copenhagen, Denmark
4Department of Ecology, Charles University, Prague, Czech Republic
5Chr. Hansen A/S, Hoersholm, Denmark
6Department of Environmental Science, Aarhus University, Roskilde, Denmark

Abstract:
Glacier melting and altered precipitation patterns influence Arctic freshwater and coastal ecosystems. Arctic rivers are central to Arctic water ecosystems linking glacier meltwaters and precipitation with the ocean through transport of particulate matter and microorganisms. However, the impact of different water sources on the microbial communities in Arctic rivers and estuaries remains unknown. In this study we used 16S rRNA gene amplicon sequencing to assess a small river and its estuary on Disko Island, West Greenland (69°N). We describe the bacterial community through a river into the estuary, including communities originating in a glacier and a proglacial lake. Our results show that water from the glacier and lake transports distinct communities into the river in terms of diversity and community composition. Bacteria of terrestrial origin were among the dominating OTUs in the main river, while the glacier and lake supplied the river with water containing fewer terrestrial organisms and more psychrophilic taxa were found in the dominant community supplied by the lake. At the river mouth the dominant bacterial communities from the lake and glacier were unnoticeable but became evident again further into the estuary. On average 23% of the estuary community consisted of indicator OTUs from the river. Environmental variables showed only weak correlations with community composition.
Transcription start site positioning is not affected by +1 nucleosome depletion

Axel Thieffry¹#, Babett Steglich²#, Jenna Persson ²#, Mette Boyd ¹, Jette Bornholdt¹, Punit Prasad², Olaf Nielsen³, Karl Ekwall ²*, Albin Sandelin¹*

¹ The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Denmark.
² Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden.
³ Cell Cycle and Genome Stability Group, Department of Biology, University of Copenhagen, Denmark

# Shared first authors, * Corresponding authors: Karl Ekwall (karl.ekwall@ki.se) and Albin Sandelin (albin@binf.ku.dk).

Abstract:
The processes that position transcription start sites (TSSs) in eukaryotic cells are not well understood. Because TSSs generally co-localize with the edge of the +1 nucleosome, several studies have suggested that the nucleosomal edge positions the pre-initiation complex and thereby the TSS. Here, we show that deletion of the Fun30 class chromatin remodelers Fft2 and Fft3 led to depletion of +1 nucleosome occupancy. This phenotype enables genome-wide testing of the above hypothesis. By sequencing 5' RNA ends genome-wide, we found that in the vast majority of cases, +1 nucleosome depletion had no impact on the placement of the TSS, favoring a model where TSS placement is primarily DNA sequence-driven.

Malte Thodberg$^{1,2}$, Ajuna Azad$^{1,2}$, Axel Thieffry$^{1,2}$, Jette Lange$^{1,2}$, Mette Boyd$^{1,2}$, Olaf Nielsen$^{3}$, Albin Sandelin$^{1,2}$

Affiliations:
$^1$Sandelin Group, Bioinformatics Centre, Department of Biology, University of Copenhagen
$^2$Basic Research and Innovation Centre, University of Copenhagen
$^3$Cell Cycle and Genome Stability Group, Department of Biology, University of Copenhagen

Abstract:
A large number of studies have shown that alternative promoter usage facilitates diversity and flexibility in gene expression and thus complexity in the final transcript- and proteome. Still, the fine details of this regulatory mechanism remain unknown. We performed Cap Analysis of Gene Expression (CAGE) on 15 samples of *S. pombe*, a common unicellular model organism, to investigate promoter usage under various stress conditions: nitrogen starvation, heat stress, oxidative stress, and change of growth medium.

The resulting atlas of almost 14,000 promoters shows a consistent bias in the position of transcription start sites compared to reference annotation from PomBase as well as extensive usage of multiple promoters by single genes. This includes cases where alternative promoters are preferred under specific types of stress, so-called alternative promoter shifts.

This promoter atlas, along with differential expression information and alternative promoters, the first of its kind, will be made publically available through the PomBase online database to facilitate future research into *S. pombe* transcriptional regulation and promoter architecture.
Origin and evolution of caste differentiation genes in ants

Bitao Qiu, Jacobus J. Boomsma, Guojie Zhang

Centre for Social Evolution, Department of Biology, Universitetsparken 15, University of Copenhagen, DK-2100 Copenhagen, Denmark

Abstract:
The emergence of morphologically differentiated castes in colonial insects represents a major transition in evolution that evolved independently in ants, bees and wasps. Female siblings have very similar genomes, but some develop into gynes (prospective queens) and most others into workers, a reproductive division of labor based on irreversible bifurcation of larval developmental pathways. To understand the evolution of gene expression networks associated with reproductive division of labor in ants, we investigated the genes with caste-biased expression in six ant lineages for which sequenced genomes were available and mapped their times of origin on the insect phylogeny, using the genomes of 35 insect species covering all major insect lineages. We found that both ancestral and younger, lineage-specific genes contribute to caste differentiation in a clearly stratified pattern. For example, in the attine fungus-growing ants, genes originated from ancestral Hymenoptera are bias expressed in gynes, whereas those originated from the family Formicidae are bias expressed in large workers, consistent with sterile workers being an ant-specific and thus evolutionarily derived innovation. In addition, caste-biased genes originating from the common ancestor of all insects were found to be enriched for TCA cycle functions, while caste-biased genes that originated in the common ancestor of the Hymenoptera were enriched for G protein signalling pathways. We further examined dn/ds ratios and were able to confirm that caste-biased genes evolved faster than non-biased genes.
Genetic population structure of Muskox (*Ovibos moschatus*): Serial founder effects drive populations differentiation

Charles Christian Riis Hansen  
*Section for Computational and RNA Biology, University of Copenhagen*

**Abstract:**  
As a key species of the arctic, muskoxen are of vital importance to the ecosystem, yet not much is known about their genetic structure. Using genotype likelihood based methods on GBS data from 117 individuals from Canada and Greenland, we identify six populations in their native distribution. The populations split into three main clusters (Canadian mainland, the Canadian Arctic Archipelago and Greenland). These three clusters display large genetic differentiation between each other, both in $F_{st}$ and a population tree. We report extremely low genome-wide genetic diversity in individuals, measured as heterozygosity, and also find low effective populations sizes. All of our findings fits the known phylogeography of the species and sheds new light on the most important founder events in the species history.  
Through this study we give recommendations of muskox managed ex-situ as at least two, if not three different breeding populations.
Avoiding abundance bias in the functional annotation of post-translationally-modified proteins

David Lyon¹, Christian Schölz¹, Jan C. Refsgaard¹, Chunaram Choudhary¹, Brian T. Weinert¹, Lars Juhl Jensen¹

¹Novo Nordisk Foundation Center for Protein Research, University of Copenhagen

Abstract:
Mass Spectrometry based shotgun-proteomics is inherently biased towards abundant proteins. This bias can influence GO-term enrichment analysis by showing enriched terms for abundant rather than e.g. post-translationally-modified (PTM) proteins. We have developed a method to correct for this bias and a freely accessible web-tool (https://agotool.sund.ku.dk/) to facilitate the use for the scientific community.
Investigating the evolution of X-linked ampliconic regions in Human Populations

Elise Lucotte¹, Mikkel Schierup¹
¹ Bioinformatic Research Center, Aarhus University, Aarhus 8000, Denmark

Abstract:
Although speciation is one of the most important phenomenon in evolution, the mechanisms behind hybrid incompatibilities remain mostly unknown. It is however known that the X chromosome plays a disproportionately large role in hybrid incompatibilities. One growing hypothesis, supported by recent studies, describes the emergence of selfish elements on the X chromosome as a major driver of speciation. Indeed, it would lead to a competition between the sex chromosomes for transmission to the offspring and drive rapid divergence between isolated population. However, the importance of such mechanism in primate speciation has not yet been assessed. Interestingly, large regions on the X chromosome subjected to recurrent selective sweeps have been highlighted recently in several primate species. They overlap with genomic regions depleted of Neanderthal introgression in humans, suggesting they have a role in hybrid incompatibilities. Importantly, they are associated with human ampliconic regions (ARs). ARs comprise several copies of genomic regions enriched in testis-expressed genes, and are subject to highly dynamic evolution. They are largely understudied, notably because of their repetitive nature which make them difficult to investigate. In this study, we explored the evolution of ARs by investigating copy number variation (CNV) between human populations using the Simons Genome Diversity Project, which provides read files for 260 individuals from 127 populations. We developed a method to assess CNVs using the read-depth on an artificial X chromosome composed of one repetition of each AR. Our results indicate that there are differences in copy number between human populations for ARs including genes expressed in testis. These differences are significant between several continental groups, suggesting that these regions have a highly dynamic evolution and have been subjected to recent selective pressure.
Identification of a nuclear exosome decay pathway for processed transcripts

Nicola Meola¹,#, Michal Domanski¹, 4,#, Evdoxia Karadoulama¹,², Yun Chen², Coline Gentil¹, Dennis Pultz³, Kristoffer Vitting-Seerup², Søren Lykke-Andersen¹, Jens S. Andersen³, Albin Sandelin², Torben Heick Jensen¹

¹Department of Molecular Biology and Genetics, Aarhus University, C. F. Møllers Allé 3, Bldg. 1130, DK-8000 Aarhus C, Denmark, ²The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaløesvej 5, DK-2200 Copenhagen, Denmark, ³Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark. 4Present address: Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, CH-3012 Bern, Switzerland

The RNA exosome is fundamental for the degradation of RNA in eukaryotic nuclei. Substrate targeting is facilitated by its co-factor Mtr4p/hMTR4, which links to RNA-binding protein adaptors. One such activity is the human Nuclear EXosome Targeting (NEXT) complex, composed of hMTR4, the Zn-finger protein ZCCHC8 and the RNA-binding factor RBM7. NEXT primarily targets early and unprocessed transcripts, demanding a rationale for how the nuclear exosome recognizes processed RNAs. Here, we describe the PolyA tail eXosome Targeting (PAXT) connection, comprising the hitherto uncharacterized ZFC3H1 Zn-knuckle protein as a central link between hMTR4 and the nuclear polyA binding protein PABPN1. Individual depletion of ZFC3H1 and PABPN1 results in the accumulation of common transcripts, that are generally both longer and more 3'polyadenylated than NEXT substrates. Importantly, ZFC3H1/PABPN1 and ZCCHC8/RBM7 contact hMTR4 in a mutually exclusive manner, revealing that the exosome targets nuclear transcripts of different maturation status by substituting its hMTR4-associating adaptors.
Large-scale prediction of RNA-RNA interactions and siRNA off-targets

Authors: Ferhat Alkan, Anne Wenzel, Oana Palasca, Peter Kerpedjiev, Anders F. Rudebeck, Peter F. Stadler, Ivo L. Hofacker, Jan Gorodkin

Abstract:
Intermolecular interactions of non-coding RNAs form the backbone of gene regulation events and predicting these interactions is of great interest. Identification of RNA interactomes promises to prove useful in several areas, such as integrative miRNA and gene expression analysis, non-coding RNA functional studies, and interpretation of RNAi application outcomes. However, in order to fulfill this task, large-scale genome- and transcriptome-wide predictions are required. To meet the demands, we implemented an efficient method, RIsearch2, that uses suffix arrays to identify perfect complementary seed regions (including G-U wobble pairs) between given query and target sequences and extends these possible interactions on both ends using a simplified energy model. A screen of ~2600 human miRNAs on the whole repeat-masked human genome, which takes about ~6h using 16 threaded cores, exemplifies the large scale capability of RIsearch2. This is orders of magnitude faster than current other available methods. Furthermore, we constructed an efficient siRNA off-target discovery pipeline by using RIsearch2 interaction predictions in combination with accessibility and expression abundance information of binding sites. We show that this pipeline can accurately predict not only the individual off-targets of siRNAs but also their overall off-targeting potential, which in fact might further influence their repression efficiency.
MicroWineBar –
A tool for analyzing and comparing metagenomics samples graphically

Franziska Klincke
DTU Bioinformatics, Department of Bio and Health Informatics, Technical University of Denmark

Abstract:
MicroWineBar is a tool for getting an overview of metagenomics samples. The user can generate different line/bar graphs to display the abundances of species in samples. This is also possible on higher taxonomic levels and can also be used for time series. Additionally, the tool can be used to identify species that have similar changes in abundances in many samples. The tool is especially useful for people who don’t have a background in bioinformatics. It is also easier so spot trends of microorganisms in graphs than in plain numbers.
Local structural RNA alignments with Foldalign 2.5

Havgaard JH¹, Sundfeld D², de Melo AC², Gorodkin J¹
¹Center for non-coding RNA in Technology and Health, IKVH, University of Copenhagen, Frederiksberg, Denmark
²Department of Computer Science, University of Brasilia, Brasilia, DF, Brazil
Contact: hull@rth.dk

Abstract:
Foldalign is a software tool for making local structural alignment of RNAs. Based on the sankoff algorithm, it simultaneously aligns and predicts a common secondary structure for two RNA sequences. The usage of structure information in the alignment process is essential because compensating basepairs can change the primary sequence of RNAs without changing the structure. Therefore, sequence based alignment tools, e.g. BLAST is not suited for aligning structured RNAs, neither globally or simply in local searches where the two sequences contain only small portions of similar structured RNAs with low sequence identity. Here, we present the latest version 2.5 of Foldalign, as well as preliminary results on further developments. In the latest version, we simplified the calculation of P-values for local searches making it easier and faster for the user to evaluate the predictions. Furthermore, multithreading was implemented making it possible to use current computers more efficiently. This together with a pruning heuristic, ensures that the algorithm is fast enough for interactive use. Thus, it will be demonstrated that Foldalign is capable of finding local regions of conserved structured RNAs in pairs of long transcripts which have low sequence identity. Since Foldalign can make local structural alignments, the resulting structures are not required to span the entire sequence or even the majority of the sequence.
The software and a web-server is available here: http://rth.dk/resources/foldalign
Comparison of microRNA annotation in miRBase and RNAcentral

Jamshed Pratap Singh Gill¹, Jan Gorodkin², and Stefan E Seemann²
¹The Bioinformatics Centre, Section of Computational and RNA Biology, Department Of Biology, University of Copenhagen, Biocenter, Ole Maaloes vej 5, 2200 Copenhagen N
² Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, 1870 Frederiksberg C, Denmark

Abstract:
MicroRNAs (miRNAs) play an important role in regulating post transcriptional control in most eukaryotic genomes. The primary transcripts of miRNAs are processed as hairpin precursor miRNA and then into small mature miRNAs. miRBase is a primary online repository for all microRNA sequences and annotation (1). miRBase contains annotation of precursor miRNAs and mature miRNAs. In complement to the recent RNAcentral initiative contains a comprehensive and up-to-date compilation of ncRNA database (2) including mirBASE. RNAcentral has annotation of primary miRNAs and precursor miRNAs complementing that of miRBase. In this study miRNA annotation obtained via RNAcentral and miRBase were compared in human (GRCh38) and mouse (GRCm38) to find the additional information gained through RNAcentral. We find that RNAcentral covers most of the precursor miRNAs in miRBase and adds 2668 and 2672 of primary miRNAs for both organisms respectively not overlapping precursors from sources such as RefSeq and Rfam. We show that the integrated data of RNAcentral provides a large source of novel miRNAs whose functional mature sequences have still to be discovered, e.g. through the analysis of miRNA characteristic short RNA-seq profiles.

Keywords: mature (miRNA), miRBase, RNAcentral, Precursor RNA, Primary miRNA


Evergreen; a webtool for surveillance of bacterial outbreaks

Johanne Ahrenfeldt, Jose Bellod Cisneros, Martin Thomsen, Camilla Koldbæk Lemvig and Ole Lund

1Center for Biological Sequence Analysis, Technical University of Denmark

Abstract:

Background
The cost of whole genome sequencing is falling, and therefore its use in diagnostics is increasing. We wish to make a platform for surveillance of pathogenic bacteria, where everything that is uploaded to our servers will automatically be compared to everything else in our databases, so that in case of an outbreak or an epidemic, the very closely related isolates can be identified quickly.

Methods
When an isolate is uploaded to the servers, first it will be compared to a homology reduced database of reference sequences, where the closest reference sequence will be found. Afterwards the isolate will be mapped to the found reference. Then the mapped sequence will be compared to all other uploaded sequences that previously have been found to match to the same reference. The phylogenetic distances will then be reported, together with the number of isolates, which are within the following distances 0, 1, 5, 10, 100 and 1000 SNPs.

Once a day a tree will be calculated for each of the references. And whenever a new isolate is uploaded, it will be placed in the tree.
Species-independent identification of known and novel recurrent genomic entities in multiple cancer patients

Jose Mg Izarzugaza

Summary
The discovery of viruses and other disease-causing pathogens from high throughput sequencing data often requires that taxonomic annotation occurs prior to association to disease. Although this bottom-up approach is effective in some cases, it fails to detect novel pathogens and remote variants not present in reference databases. We propose an alternate approach utilizes sequence clustering for the identification of nucleotide sequences that co-occur across multiple sequencing data instances. Thus, not limited to reported species.

We applied the workflow to 686 sequencing libraries from 252 different cancers and 56 controls. We used our pipeline to associate recurrent sequences to the onset of the disease but also to the use of common laboratory kits to identify common methodological or technical artifacts sourcing erroneous conclusions, as we have observed in the recent literature. We provide examples of identified inhabitants of the healthy tissue flora as well as experimental contaminants.

Scientific Justification:
The discovery of viruses and other disease-causing pathogens from high throughput sequencing data often requires that taxonomic annotation occurs prior to association to disease. Although this bottom-up approach is effective in some cases, it fails to detect novel pathogens and remote variants not present in reference databases. We propose an alternate approach utilizes sequence clustering for the identification of nucleotide sequences that co-occur across multiple sequencing data instances. Thus, not limited to reported species.

We will describe an analysis where we applied our workflow to 686 sequencing libraries from 252 different cancers and 56 controls in the context of the Cancer and Pathogens project funded by the GenomeDenmark platform, where we have used our pipeline to associate recurrent sequences to the onset of the disease with satisfactory results in the identification of known and novel oncoviruses.

A number of recent articles associating viruses and disease were rapidly disregarded as contamination stemming from the use of laboratory contaminants. We reported the erroneous identification of ATCV-1 (Kjartansdóttir et al, PNAS 2015). Concomitant to the universalisation of sequencing technologies, the identification of secondary species poses a threat to the validity of the conclusions drawn if caution is not exerted. Our method correlated recurrent sequences and common laboratory kits, deeming these observations as methodological/technical artefacts. Our intention is to make the scientific community aware of this extremely relevant issue.
Identification of Known and Novel Recurrent Viral Sequences in Data from Multiple Patients and Multiple Cancers

Jens Friis-Nielsen 1, Kristín Rós Kjartansdóttir 2,†, Sarah Mollerup 2,†, Maria Asplund 2, Tobias Mourier 2, Randi Holm Jensen 2, Thomas Arn Hansen 2, Alba Rey-Iglesia 2, Stine Raith Richter 2, Ida Broman Nielsen 2, David E. Alquezar-Planas 2, Pernille V. S. Olsen 2, Lasse Vinner 2, Helena Fridholm 2, Lars Peter Nielsen 3, Eske Willerslev 2, Thomas Sicheritz-Pontén 1, Ole Lund 1, Anders Johannes Hansen 2, Jose M. G. Izarzugaza 1,† and Søren Brunak 1,4,*,†

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark; jef@cbs.dtu.dk (J.F.-N.); thomas@cbs.dtu.dk (T.S.-P.); lund@cbs.dtu.dk (O.L.); txema@cbs.dtu.dk (J.M.G.I.)
2 Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, DK-1350 Copenhagen, Denmark; kristin.kjartansdottir@snm.ku.dk (K.R.K.); sarah.mollerup@snm.ku.dk (S.M.); amasplund@snm.ku.dk (M.A.); tmourier@snm.ku.dk (T.M.); randi.jensen@snm.ku.dk (R.H.J.); thomas.hansen@snm.ku.dk (T.A.H.); zld305@alumni.ku.dk (A.R.-I.); srichter@snm.ku.dk (S.R.R.); ida.nielsen@snm.ku.dk (I.B.N.); d.e.alquezar@gmail.com (D.E.A.-P.); pvsolsen@snm.ku.dk (P.V.S.O.); lasse.vinner@snm.ku.dk (L.V.); helena.fridholm@gmail.com (H.F.); ewillerslev@snm.ku.dk (E.W.); AJHansen@snm.ku.dk (A.J.H.)
3 Department of Autoimmunology and Biomarkers, Statens Serum Institut, DK-2300 Copenhagen S, Denmark; lpn@ssi.dk
4 NNF Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark
* Correspondence: soren.brunak@cpr.ku.dk; Tel.: +45-45252477
† These authors contributed equally to this work.

Abstract: Virus discovery from high throughput sequencing data often follows a bottom-up approach where taxonomic annotation takes place prior to association to disease. Albeit effective in some cases, the approach fails to detect novel pathogens and remote variants not present in reference databases. We have developed a species independent pipeline that utilises sequence clustering for the identification of nucleotide sequences that co-occur across multiple sequencing data instances. We applied the workflow to 686 sequencing libraries from 252 cancer samples of different cancer and tissue types, 32 non-template controls, and 24 test samples. Recurrent sequences were statistically associated to biological, methodological or technical features with the aim to identify novel pathogens or plausible contaminants that may associate to a particular kit or method. We provide examples of identified inhabitants of the healthy tissue flora as well as experimental contaminants. Unmapped sequences that co-occur with high statistical significance potentially represent the unknown sequence space where novel pathogens can be identified.

Keywords: sequence clustering; taxonomic characterisation; novel sequence identification; next generation sequencing; cancer causing viruses; oncoviruses; assay contamination
1. Introduction

The International Agency for Research on Cancer (IARC) lists several biological species with carcinogenic potential in humans [1]. This list comprises a bacterium (species Helicobacter pylori), three parasitic flukes (Clonorchis sinensis, Opisthorchis viverrini and Schistosoma haematobium), and seven viruses: human papillomaviruses (HPV), human immunodeficiency virus-1 (HIV-1), Epstein-Barr virus (EBV), hepatitis B and C virus (HBV and HCV), Kaposi’s sarcoma-associated herpesvirus (KSHV), and human T-cell lymphotropic virus type 1 (HTLV-1).

With the advent and spread of low-cost sequencing technologies, many viruses were discovered in the last decade [2–8]. One interesting discovery that fuelled the search for oncoviruses was Merkel cell polyomavirus (MCPyV) found to be clonally integrated in Merkel cell carcinomas [9,10]. The computational biology community has promptly responded to the growing need for specialised algorithms and pipelines to analyse the wealth of data [9,11–25]. Table S1 summarises the main features of some of the common approaches. In spite of particularities in the implementation, these methodologies share key conceptual similarities: First, sequencing reads or assembled contigs that originate from the host are identified and discarded, a process termed computational subtraction [9,13]. When the genomes or the concentrations of foreign species are small compared to host genomes, this step eliminates a substantial fraction of the total sequencing reads. Second, the remaining non-host sequences are compared to a library of known reference sequences for taxonomic characterisation. The aforementioned methods identify species present across multiple samples, and the recurrence of a given viral entity may indicate an association to disease [10,26]. Albeit conceptually valid, this bottom-up approach is inherently limited to the pre-existence of the organism in the reference databases, whereas novel oncoviruses showing partial or no similarity to known sequences will be missed. Current efforts aiming at estimating and characterising metagenomic diversity are far from a complete mapping of the (viral) sequence-space [27]. In fact, it is common to observe that a small but significant amount of unknown sequences, the so-called dark matter [28], goes through the current analysis pipelines without proper characterisation and is discarded from further analyses [24,29,30].

Here, we propose a method capable of identifying the recurrence of sequences across related samples independently of their existence in reference databases. Our top-down approach compares samples and establishes recurrence prior to the taxonomic characterisation of the sequences. Thus, enabling the identification of both known and novel biological entities. Our method has conceptual similarities to the work of Andreotta et al. [31] where clustering of genes is used to find families that are predominantly found in pathogenic bacteria. Attending to Koch’s postulates as modified by Fredericks and Relman [32], sequences from biological entities with a causative or facilitator role would be present in diseased samples and absent in healthy controls. In addition, recent studies documented the presence of contaminating and/or artefactual sequences that source from the laboratory kits and reagents used for sample processing and library preparation [14,33–37]. If not properly addressed, these confounding observations may lead to erroneous conclusions [38,39]. Our method ascertains the statistical associations between recurrent sequences and a collection of features that describe the samples with respect to tissue, disease type, laboratory method, etc. Additionally, the presence of other known technical problems, such as cluster invasion on the sequencing flow cells [40], might be detected.

2. Materials and Methods

2.1. Ethics Statement

The study was conducted in accordance with the Declaration of Helsinki. Two ethical boards reviewed the protocol of this study: The Regional Committee on Health Research Ethics (Case No. H-2-2012-FSP2) and the National Committee on Health Research Ethics (Case No. 1304226). Because the study used only samples that were anonymised at collection both boards waived the need for informed consent in compliance with the national legislation in Denmark.
1.1. Data Sets

Two hundred and fifty-two cancer samples of 17 different types were collected from various locations in Denmark and Hungary. Cancer samples of malignant melanoma, acute myeloid leukaemia (AML), B-cell chronic lymphocytic leukaemia (B-CLL), chronic myelogenous leukaemia (CML), and T-lineage acute lymphoblastic leukaemia (T-ALL; \( n = 9 \)) were obtained from Aarhus University Hospital, Denmark. B-cell precursor acute lymphoblastic leukaemia (BCP-ALL), oropharyngeal head and neck cancer, testicular cancer, and T-ALL \( (n = 2) \) were obtained from Rigshospitalet, Denmark (Copenhagen University Hospital). Basal cell carcinoma, and mycosis fungoides (cutaneous T-cell lymphoma) were obtained from Bispebjerg Hospital (Copenhagen University Hospital). Samples of bladder cancer, breast cancer, colon cancer, as well as ascites fluid of breast cancer, colon cancer, ovarian cancer, and pancreatic cancer were obtained from the Danish Cancer Biobank, Herlev Hospital, Denmark. B-cell lymphoma cell lines were obtained from Aalborg University Hospital, Denmark. Vulva cancer samples were obtained from the National Institute of Oncology, Budapest, Hungary.

Libraries were prepared at the Center for GeoGenetics (CGG), University of Copenhagen, Denmark based on seven different methods for sample processing comprising five different enrichment methods and shotgun sequencing targeting total DNA or RNA (Table S3). The enrichment methods used in the current work were circular genome amplification, sequence capture with retrovirus probes, virion enrichment (DNA and RNA), and mRNA enrichment. Further details on sample processing and library preparation have been published elsewhere \([37,41,42]\), except for mRNA enrichment which was performed using Dynabeads mRNA direct extraction kit (Thermo Fisher Scientific, Waltham, MA, USA) followed by ScriptSeq v2 RNA-Seq Library Preparation kit as for total RNA analysis \([41]\).

Ultimately, the data set consisted of 686 DNA and RNA libraries, for which \( 2 \sim 100 \) bp paired end sequencing was performed using the Illumina HiSeq 2000 platform at BGI-Europe, Copenhagen, Denmark. The 686 sequencing libraries thus originated from 252 different cancer samples, 32 non-template controls, and 24 exogenous controls. The distribution of methods, libraries and controls for each sample type is provided in Table S2. Samples were preferably analysed with multiple methods, thus 165 out of 252 samples were analysed with more than one laboratory method (Table S3).

1.2. Constituents of the Software Pipeline and Execution Parameters

The datasets went through a sequential pipeline with modules (in order) of preprocessing, computational subtraction of host sequences, low-complexity sequence removal, sequence assembly, clustering, association to metadata features, and taxonomical annotation. Figure 1 provides a schematic representation of the pipeline used to identify recurrent sequences across related samples.

Demultiplexing was performed using a local python script to partition the reads based on exact matches in the FASTQ header lines to the multiplexed indices provided. Preprocessing of reads was performed for all datasets in parallel using AdapterRemoval \([43]\) with the following parameters \{–trimns, –trimqualities, –minquality 2, –minlength 30, –collapse, –outputcollapsed, –outputcollapsedtruncated, –singleton\}. Read ends were trimmed for low quality base calls. Reads were discarded if the length after trimming fell below 30 bp. In these cases, the other read in a pair was kept as a singleton. Overlapping paired reads from short inserts were collapsed into a single read if the overlap was longer than 11 bp, according to the default behaviour of AdapterRemoval.

Preprocessed reads were filtered if they showed homology to the human reference genome, which included decoys and alternative sequences from version GCA_000001405.15 (GRCh38) of the Genome Reference Consortium (downloaded August 20, 2014). Mapping to the human genome was done using BWA \([44]\) version 0.7.10-r789 with the MEM alignment algorithm and default parameters. All mapped reads without Sequence Alignment/Map (SAM) \([45]\) flag 4 were discarded. Single-unmapped reads from read pairs were kept. Human depleted reads were filtered for low-complexity regions using the NCBI-BLAST associated module DustMasker \([46]\) and default parameters. Reads containing low-complexity stretches of 25 bp or longer were discarded. Assembly of the remaining (non-human, high complexity) reads was performed with IDBA-UD \([47]\) and parameters \{–precorrection\}. Contigs
shorter than 200 bp were discarded. A total of 1,387,377 contigs, originating from the 686 data sets, went through the entire pipeline. Contigs ranged from 200 bp to 418,807 bp with an overall N50 of 817 bp.

Contigs from all data sets were pooled and clustered based on pairwise sequence homology using CD-HIT [48], in fast mode. We chose parameters for clustering that maximised grouping of similar sequences while minimising inclusion of unrelated sequences. We considered the following different parametrisation values: percent minimum sequence identity (-c 0.80, 0.85, 0.90, 0.95, 0.99); percent minimum alignment length based on the length of the shortest (-aS) or longest (-aL) sequence (-aS, -aL 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.90, 0.95, 0.99); global (1) or local (0) alignment mode (-G 1, 0). A full factorial combination of the aforementioned parameters resulted in 200 different settings. There were 126 successful combinations of settings listed in Table S4 from where we chose the final settings {-c 0.90–aS 0.90–G 1}

The datasets were described with a panel of 404 different binary metadata features, for example tissue or disease characteristics (Table S5). Features logically assessed whether they related to a particular dataset or not. Features describing less than five datasets were removed. Additionally, features that correlated perfectly in terms of Matthew’s correlation coefficient (MCC = -1) were merged. These filters resulted in 143 unique features (Table S5). Biological features (n = 25) defined sample type, for instance tissue or disease category. Methodological features (n = 49) described specifics for sample preparation such as extraction kits, enrichment methods, polymerases, primers, buffer, filters used, or the laboratory where the work was performed, etc. Technical features (n = 69) defined the flow cell lane identifiers and whether resequencing was performed. The distributions of datasets and samples across the features are provided in Table S5. Associations in the clustered contigs and metadata features were evaluated with Fisher’s exact test using a one-tailed alternative hypothesis (greater) and calculated in R using the function fisher.test [49].

Annotation of taxonomy was performed in two rounds. First aligning contigs with BLASTn [50] with parameters {-evalue 0.001} using default {-task megablast} to a frozen version of the NCBI
nucleotide database nt (downloaded 3 February 2015). Secondly, using BLASTx with parameters 
{-evalue 0.001} of all unmapped contig stretches to a frozen version of the NCBI non-redundant protein
database nr (downloaded 3 February 2015). The best hit by highest bit-score was kept for each contig.
The taxonomy database (downloaded 3 February 2015) was used to translate all GenBank identifiers
from hits to taxonomy identifiers. The taxonomy identifiers were then used to obtain the complete
taxonomical lineage and extract scientific names of species. The abundances of all species in each
cluster were used to calculate the species evenness index as defined by Mulder et al. [51]. Clusters
were annotated as the most abundant species in each cluster.

The software to use after the assembly step has been uploaded at https://github.com/jensfrisnielsen/sequence_recurrent. Sequence clusters that have been described in detail throughout
the manuscript have been included as supplementary files.

1. Results

1.1. Clustering Identifies Recurrent Nucleotide Sequences across Samples

Clustering performance depends on the adequate selection of parameters. We experimented
with a variety of configurations described by c0xaY0yGz where x,Y,z varied. The variables denote
minimum percentage of sequence identity x (c0x), minimum percentage of alignment length y (aY0y)
based on mode Y of shortest (aS) or longest (aL) contig in alignment, as well as using local (G0)
or global (G1) alignment mode z (Gz). For example, a configuration encoded C090aS090G1 would
represent a clustering that requires global alignments with a 90% minimum sequence identity over
90% of the length of the shortest contig. The full list of investigated parameter combinations can be
found in Table S4. We chose the parameters based on the performance of the clustering of expected
contaminant sequences from avian leukosis virus (accession id AY350569) [37] and other related avian
retroviruses (ARs) such as avian myeloblastosis virus [52]. ARs are used in the manufacture of the
reverse transcriptase FailSafe PCR enzyme (Epicentre, Madison, WI, USA) included in the utilized
ScriptSeq v2 RNA-Seq Library Preparation kit (Illumina, San Diego, CA, USA). This kit is commonly
used for preparation of RNA libraries [52]. We identified clusters containing contigs that aligned to
species of the Alpharetrovirus genus (NCBI taxa-id: 153057) according to BLASTn and BLASTx hereafter
referred to as AR clusters. All contigs in AR clusters were resolved with BLASTn and BLASTx and
two metrics were considered for AR clusters.

As the first performance metric, we computed the odds ratios (ORs) of the associations between
the presence of AR in the clusters and the use of the ScriptSeq kit. We used a 2 ^ 2 contingency table
defining the sets of libraries: AR positive and ScriptSeq positive (ARpSSp); AR positive and ScriptSeq
negative (ARpSSn); AR negative and ScriptSeq positive (ARnSSp); AR negative and ScriptSeq negative
(ARnSSn). OR is then defined as the ratio ARpSSp^ARnSSn / (ARpSSn^ARnSSp) and describes the
strength of the association between clusters and features. ORs above 1 indicate association between
the presence of the AR virus and the use of the ScriptSeq kit. ORs for all AR clusters were inspected in
different parameter settings (Figure S1). The ORs varied mostly block-wise with the parameters. The
largest differences observed were between usages of the shortest or longest sequence in alignments
with the alignment length filter. Associations from the shortest mode tended to have higher dispersion
in the range of ORs. Furthermore, one block of clustering results using global alignment mode,
alignment length based on the shortest contig, and a minimum sequence identity of 90% (c09°aSyG1),
had an overall high range of ORs as well as the highest minimum values. This suggested that the
clustering was able to reproduce the association between AR clusters and the ScriptSeq kit. In contrast,
the clustering with parameter settings c080aS030G0 had a very broad range of ORs corresponding to a
skewed clustering where some clusters had incorporated most sequences and left other clusters with
only a few contigs.

As a second performance metric we computed the species evenness [51] indices of the AR clusters
represented in Figure S1. The species evenness index is a score that derives from the Shannon’s
diversity index [53] and compares the abundance of each species within a cluster. An index of zero is assigned to clusters that are constituted uniquely by contigs mapping to a single species. Contrarily, scores closer to 1 would indicate that the cluster points to several species and that these are equally abundant. In our experiment, we favoured lower evenness indices as they indicate that clusters were able to single out species correctly. For example, parameter settings c080aL030G1 generally had a high level of species evenness (median 0.73) in clusters, suggesting an incorrect separation of species. In stark contrast, a block of parameters using global alignment mode, alignment length based on shortest sequence, 90% minimal sequence identity, and a minimum alignment length of 80/85/90/95/99% of the shortest sequence (c090aSyG1) all had a median species evenness of 0. This group of parameter settings also showed desirable performance in terms of OR, as mentioned before. Generally it seemed that global mode (G1) had better ORs than local mode (G0) when keeping other parameters fixed. Additionally keeping 90% minimal sequence identity (c090) and varying minimal length of alignment in shortest mode (aS) seemed stable in both ORs and species evenness indices indicating that these close parameter settings were generally good. We chose to proceed with a clustering based on global alignments with a 90% minimum sequence identity over 90% of the length of the shortest contig (c090aS090G1). This configuration resulted in a total of 681,858 clusters. Of these, 23,205 clusters contained contigs from at least five different data sets and represented 546,735 different contigs. The full distribution of cluster sizes can be found in Table S6.

1.1. Characterisation of the Nature of the Recurrent Sequences

The associations between the clusters and the binary metadata features were assessed using a Fisher’s one tailed exact test. There were 16,567 significant associations having \( p \)-value \( < 3.01 \times 10^{-10} \), corresponding to a 0.001 significance level when using Bonferroni’s correction for multiple testing [54]. The significant associations were arranged in 6165 unique clusters and with 73 unique features. The distribution of the significant associations showed that recurrent sequences originated from diverse sources and that individual clusters often associated to more than one feature (Figure 2). Furthermore it is evident that the clusters tend to group in their associations. Likely, these groupings represent one or more organisms. We investigated the nature of the clusters accounting only for the associated feature with the smallest \( p \)-value; hereafter described as the strongest associations. There were 50 unique features involved in all the strongest associations. The distribution of \( p \)-values for each feature is represented in Figure 3. The 6165 strongest associations were distributed according to 602 biological, 5045 methodological and 518 technical associations. These unique features were arranged in 3 biological, 24 methodological and 23 technical features. Most \( p \)-values were above \( 1 \times 10^{-24} \) and associations with lower \( p \)-values were to a few methodological features annotated as extraction kits: QIAamp DNA mini kit (f056) (Qiagen, Hilden, Germany), DNase/RNase: Promega DNase (f068) (Promega, Madison, WI, USA), and DNase/RNase: Promega DNase stop solution (f069), purification kit: RNeasy MinElute, Qiagen (f076), library build: NEBNext, New England BioLabs (f079) (New England Biolabs, Ipswitch, MA, USA), and ScriptSeq v2 RNA-Seq, Illumina (f084); the latter with a minimum \( p \)-value of \( 3.04 \times 10^{-89} \).
**Figure 2.** $p$-values of all significant associations. Rows describe features with biological features in red, methodological in green and technical in blue. There are 73 features significantly associated to one or more clusters. Columns describe all significant associations of each of the 6165 unique clusters. The cluster identifiers have been excluded to avoid cluttering.
Figure 3. Lowest p-values of clusters established by the pipeline. The p-values are arranged by feature of the strongest significant association of each of the 6165 clusters. The 50 features involved as strongest associations have been coloured by type: biological (red), methodological (green), and technical (blue). The boxes span the first and third quartiles. The dark band inside each box represents the median. The whiskers of the boxes extend to the lowest and highest values within a distance of 1.5 times the interquartile range. As can be seen, most p-values were above 1e-24, but a few methodological features have associated clusters with very low p-values, such as f056, f068, f069, f076, f079, and f084. The library preparation kit ScriptSeq v2 RNA-Seq, Illumina (f084) displays strongly associated clusters with p-values as low as 3.04e-89 that mapped as species *Avian myeloblastosis-associated virus*. Clusters that were annotated as NCBI species Parvovirus NIH/CQV were associated to laboratory-kit RNeasy MinElute, Qiagen (f076) with minimal p-value 5.48e-38. Finally, a cluster annotated as Acanthocystis turfacea chlorella virus MN0810.1 (ATCV) was associated to DNase/RNase: Promega DNase stop solution (f069) with p-value ≈ 4.19e-12.

1.1. Taxonomic Characterisation

Using BLAST and the NCBI taxonomy database a taxonomic characterisation was attempted for the 546,735 contigs in the 6165 clusters. This resulted in a taxonomical annotation of 3553 clusters using BLASTn and an additional 1630 clusters when using BLASTx. For 982 clusters, neither BLASTn nor BLASTx found significant species in the database. These clusters remained uncharacterised (Table 1). We found that almost all clusters significantly associated to biological features could be annotated (598 of 602) in contrast to non-biologically associated clusters (4584 of 5563). A total of 1524 unique species were annotated corresponding to 5183 clusters.
Table 1. Annotation of associations. The 6165 clusters were mapped using BLASTn and BLASTx. Rows describe the corresponding type of feature involved as the strongest association of each cluster.

<table>
<thead>
<tr>
<th>Feature type</th>
<th>BLASTn</th>
<th>BLASTx</th>
<th>Unmapped</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological</td>
<td>593</td>
<td>5</td>
<td>4</td>
<td>602</td>
</tr>
<tr>
<td>Methodological</td>
<td>2662</td>
<td>1515</td>
<td>868</td>
<td>5045</td>
</tr>
<tr>
<td>Technical</td>
<td>298</td>
<td>110</td>
<td>110</td>
<td>518</td>
</tr>
<tr>
<td>Total</td>
<td>3553</td>
<td>1630</td>
<td>982</td>
<td>6165</td>
</tr>
</tbody>
</table>

The Human Microbiome Project (HMP) \cite{55} defines a collection of reference genomes built from metagenomic samples and associates these to specific sites and tissues across human body sites. We used this data set of 1317 associations as a confirmation that our pipeline was able to correctly detect and taxonomically characterise recurring biologically relevant sequences. HMP provides a list of commensal organisms commonly found in the three sites that relate to our samples: the gastrointestinal tract, oral cavity and urogenital tract. We observed the strongest, significant associations between the expected organisms and biopsies from colon cancer, oral cavity cancer, and vulva cancer. The taxonomical characterisation of these clusters is described in Table 2. Seven clusters significantly associated to colon cancer biopsies describing four different organisms that inhabit the gastrointestinal tract according to HMP, and 342 clusters significantly associated to oral cavity cancer describing 13 different organisms present in the oral cavity in HMP. Finally, we also discovered a cluster significantly associated to vulva cancer annotated as species *Campylobacter ureolyticus* (p-value = 1.03e-12), an inhabitant of the urogenital tract as described by HMP.

Table 2. Taxonomical characterisation of certain biologically associated clusters. The clusters are significantly associated with lowest p-values to biological features and the species annotations are described by HMP. In cases where several clusters shared the annotated species, the lowest p-value of the associations is given \#sig: number of significant clusters.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cluster annotation (species)</th>
<th>#sig</th>
<th>p-value</th>
<th>HMP body site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon cancer biopsy</td>
<td><em>Bacteroides fragilis</em></td>
<td>2</td>
<td>2.43e-20</td>
<td>Gastrointestinal tract</td>
</tr>
<tr>
<td>Colon cancer biopsy</td>
<td><em>Faecalibacterium prausnitzii</em></td>
<td>3</td>
<td>1.60e-20</td>
<td>Gastrointestinal tract</td>
</tr>
<tr>
<td>Colon cancer biopsy</td>
<td><em>Eubacterium rectale</em></td>
<td>1</td>
<td>2.92e-17</td>
<td>Gastrointestinal tract</td>
</tr>
<tr>
<td>Colon cancer biopsy</td>
<td><em>Alistipes shahii</em></td>
<td>1</td>
<td>1.34e-13</td>
<td>Gastrointestinal tract</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Prevotella melaninogenica</em></td>
<td>292</td>
<td>1.74e-24</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Streptococcus agalactiae</em></td>
<td>2</td>
<td>4.60e-23</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Prevotella veroralis</em></td>
<td>8</td>
<td>1.73e-21</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Prevotella histicola</em></td>
<td>1</td>
<td>5.37e-16</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Streptococcus oralis</em></td>
<td>22</td>
<td>2.31e-15</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Prevotella dentalis</em></td>
<td>7</td>
<td>2.31e-15</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Porphyromonas gingivalis</em></td>
<td>2</td>
<td>4.49e-14</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Solobacterium moorei</em></td>
<td>1</td>
<td>1.34e-13</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Treponema denticola</em></td>
<td>2</td>
<td>8.26e-13</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Campylobacter rectus</em></td>
<td>1</td>
<td>2.60e-12</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Filifactor alocis</em></td>
<td>2</td>
<td>4.12e-11</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Streptococcus dysgalactiae</em></td>
<td>1</td>
<td>4.12e-11</td>
<td>Oral</td>
</tr>
<tr>
<td>Oral cavity cancer</td>
<td><em>Prevotella sp. oral taxon 306</em></td>
<td>1</td>
<td>4.85e-11</td>
<td>Oral</td>
</tr>
<tr>
<td>Vulva cancer</td>
<td><em>Campylobacter ureolyticus</em></td>
<td>1</td>
<td>1.03e-12</td>
<td>Urogenital tract</td>
</tr>
</tbody>
</table>

In the methodological associations, we correctly detected the strong known association (p-value: 3.04e-89) of avian myeloblastosis-associated virus (accession L10922.1) used in the manufacture of the ScriptSeq v2 RNA-Seq library preparation kit (f089). As the clustering parameters were evaluated with this known contaminant, this is expected. Furthermore, we annotated 19 clusters as NCBI taxonomy species Parvovirus NIH-CQV (accession KC617868.1; NCBI taxa-id 1341019), an established...
contaminant [34,35]. The associated feature with lowest \( p \)-value to the parvovirus clusters suggested a contamination from the RNeasy MinElute purification kit (f076) manufactured by Qiagen (\( p \)-value: 5.48e-38). In addition, a single cluster annotated as NCBI taxonomy species Acanthocystis turfacea Chlorella virus MN0810.1 (accession JX997174.1, taxa-id 1278272) with lowest associated \( p \)-value (\( p \)-value = 4.19e-12) to laboratory kit DNase/RNase: Promega DNase stop solution (f069). ATCV-1 was previously reported as a contaminant [36].

1.1. Identification of Novel Recurrent Sequences

In addition to the sequences that were characterised in the previous step, we found 982 examples of uncharacterised clusters. The contigs in these clusters varied substantially in length ranging from a minimum of 200 bp to a maximum of 33.6 kb (N50 = 617 bp). Our approach provides the capability to discover these recurrent novel sequences, but also permits the investigation of their plausible origin. Most associations were methodological (Table 1), probably sourcing from nucleotide sequences contained in various laboratory kits (Figure 4). For instance, out of the 868 methodologically associated clusters, there were 648 associated clusters to the laboratory reagent DNase/RNase: Promega DNase stop solution (minimum \( p \)-value: 2.40e-36). Additionally, 110 recurring sequences were attributed to technical issues of the flow cell lanes (minimum \( p \)-value: 1.85e-21 in feature 383). In total, 4 unmapped clusters were associated to a biological feature, namely oral cavity cancer, with the longest contig of each cluster at 1789, 3247, 4661, and 4720 bp and with respective \( p \)-values of 1.01e-10, 1.01e-10, 1.17e-14, and 1.01e-10.

![Figure 4. Unmapped clusters. The clusters are placed by their strongest associated feature. Feature types are marked in colour as follows: biological (red), methodological (green), and technical (blue). Top: Number of clusters associated to each feature on a log-10 scaled axis. There are 648 associated clusters of feature DNase/RNase: Promega DNase stop solution (f069), and 1 associated cluster to feature Polymerases: Phusion HF, NEB (f086). Bottom: Base-pair length (bp) of all cluster representatives (longest contig of each cluster) on a log-10 scaled axis. The N50 of all unmapped cluster representatives are marked by a brown dot. The longest cluster representative is 33.6 kb with N50 = 617 bp.](image-url)
To further clarify the unresolved biologically associated sequences, we manually investigated the cluster representatives using the newest databases (December 2015) at the NCBI web-interfaces for BLASTn, BLASTx and CCD v. 3.14 (conserved domains) [56] with default parameters and an e-value <0.001 (Table 3). All cluster representatives could be explained as commensal bacteria related to the oral cavity as described by HMP. In order of increasing length, the cluster representatives were identified as: Prevotella veroralis, Prevotella veroralis, Prevotella fusca JCM 17724, and Peptostreptococcus anaerobius as the best hits with percent sequence identity: 92%, 90%, 91%, and 72%, respectively. Cluster representatives 3 and 4 contained both bacterial and phage-like conserved domains. The super family DUF4280 is of unknown function but related to bacteria and the ND2 super family is the nicotine adenine dinucleotide (NADH) dehydrogenase subunit 2 involved in electron transport. Conversely, Phage_base_V is related to the tail of phages and rve is an integrase domain that could also be explained as part of a transposon. Likely these sequences derived from less well-described parts of the microbiome.

**Table 3.** Conserved domains of unmapped biological clusters. The cluster representatives of the four unmapped biologically associated clusters were manually searched for sequence similarities and conserved domains via the NCBI web-interfaces BLASTn, BLASTx, and CCD, respectively. Cells containing a dash had no hits with an e-value < 0.001. Cluster representative: Length of the sequence. BLASTn and BLASTx: Organism name (accession)%-id/%-coverage. CCD: Domain name (accession).

<table>
<thead>
<tr>
<th>Cluster representative</th>
<th>BLASTn</th>
<th>BLASTx</th>
<th>CCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1789 bp</td>
<td>-</td>
<td>Prevotella veroralis (WP_026284690.1) 92%/42%</td>
<td>-</td>
</tr>
<tr>
<td>3246 bp</td>
<td>Prevo</td>
<td></td>
<td>DUF4280 super family (cl16620) TauE super family (cl21514)</td>
</tr>
<tr>
<td>4661 bp</td>
<td>Prev</td>
<td></td>
<td>Peptidase_M23 (pfam01551) lysozyme like super family (cl00222) DUF4280 (pfam14107) Fil_haemagg_2 (pfam13332) Phage_base_V (pfam04717)</td>
</tr>
<tr>
<td>4720 bp</td>
<td>Eubo</td>
<td></td>
<td>Acyl_transf_3 super family (cl21495) ND5 (MTH00095) rve (pfam00665) ND2 super family (cl10157)</td>
</tr>
</tbody>
</table>

1. Discussion

Usually, virus discovery in shotgun sequencing studies involves processing millions of reads in a viroinformatics pipeline. Existing tools typically offer a comprehensive taxonomical description of a single sample that is compared to the taxonomy of other samples to determine their relevance. A downside of this bottom-up methodology is that novel sequences that cannot be sufficiently well characterised in the first round are often discarded in the process. Another disadvantage is that potential contaminants will have to be controlled for in the post-processing of the data, an effort that is often omitted [38]. In the present study, we have presented a methodology to categorise recurring sequences according to experimental origin and metadata features. Additionally, using this methodology we could replicate both biological and methodological sequence associations known from the literature as well as pinpoint new unannotated recurring sequences.

In this study, we had no datasets and features of healthy biological controls. We included a comparison to published reference genomes from HMP to validate that biologically co-occurring sequences can be found with the presented methodology. In this case, we are most likely observing
normal biological inhabitants of the tissue samples, something our metadata scheme does not account for. The disease association of many of these organisms is obviously not fully known, and some of them could be related to disease features outside the cancer domain, features that we did not include in the present study.

Optimising clustering parameters for one virus family might not result in the optimal separation of other families. Here, we optimised clustering parameters to rediscover the association of sequences to a known laboratory kit. Using these clustering parameters may result in a non-optimal separation of clusters that biologically belonged together, or the reverse problem—merged clusters that reflected different biological units. Optimal separation is likely problem-specific. Different taxonomic units would require the use of different clustering parameters to separate. However, choosing taxonomy-specific parameters requires a working hypothesis of the most likely findings. Here, we focused on the general problem of associating sequences to features using a known association to guide the choice of clustering parameters.

A combination of several features may be the true foundation of particular sequences but this was not explored in this work. There may also be situations where a combination of clusters is the correct association to a particular feature. For instance, a virus that is present with a low titre may be sequenced sporadically resulting in less than full coverage and several non-coherent contigs from different viral genome regions. Each cluster may include an incomplete amount of data sets and thereby artificially show a weaker association. Merged and viewed as one, the incomplete clusters will have the correct strength of association. A grouping based on taxonomy, or a more data-driven approach that cluster sequence groups based on the associated datasets as seen in Figure 2, could be included as another iteration to properly strengthen the statistical associations. Furthermore, forming clusters only by internal sequence identity may also miss pathogenic scenarios such as an oncovirus and any necessary helper viruses that do not share homology to the oncovirus.

In the present study, we used a majority vote to assign taxonomy. There could be other ways to assign taxonomy, for instance, using a lowest common ancestor (LCA) strategy. A majority vote will likely introduce some false assignments if there are distant taxa involved in the sequence group present in nearly equal fractions. A LCA strategy can handle this but may reduce the taxonomic resolution to a level where there is no real gain of information.

After determining what the significantly co-occurring sequence groups are, more effort might resolve interesting unmapped contigs. For instance, use of more sensitive alignment algorithms, profile Hidden Markov Models (HMMs), gene predictors, artificial neural networks trained on specific signals such as viral capsid sequences [57], or PCR extraction followed by Sanger sequencing might provide the relevant clues. However, that was not within the scope of this study.

The major advantage of the top-down approach is that it works without prior knowledge of the sequences. It is not dependent on reference sequence databases to single out the promising candidates for further analysis. The top-down method can determine the relevance of unknown sequences upfront while also systematically controlling for contamination by design. Most of the annotated sequences found in this study were sequenced from cancer specimens. However, it is apparent from the association analysis that several viral sequences detected are possibly contaminants or technical artefacts. Furthermore, the unmapped clusters are retained and easily arranged by relevance according to the nature of their associated features. Having this information helps precipitate a prioritised list of sequence candidates

The quality of the associations will depend on the experimental design, sampling, available metadata, as well as the rigorousness and standardisation of both working routines and annotations. We stress the point that care must still be taken when formulating hypotheses and in the interpretation of associations.
1. Conclusions

Virus discovery using high-throughput sequencing and especially characterising clinical samples is a challenge. Many viral discovery pipelines rely on similarity to reference databases as the most compelling argument for identifying putative sequences of medical or biological importance. Although a necessary step in the analysis, it has the downside of not considering novel sequences not included in reference sets as well as not considering the origins of the discoveries. There are many examples of contamination and technical artefacts; therefore, potential discoveries should be accompanied by convincing evidence that the sequences are not instead associated with the methodology or technology in use. We suggest a different approach that has complementary advantages inherent in the design. We show that we can differentiate between biological and non-biological associations, replicate known associations and potentially add new associations of cancer-associated viruses.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/8/2/53, Table S1: Virus discovery pipelines, Table S2: Distribution of library types, Table S3: Methods in cancer samples, Table S4: Clustering parameters, Table S5: Feature descriptions, Table S6: Datasets in clusters, Figure S1: Clustering performance.

Acknowledgments: This work is supported by The Danish National Advanced Technology Foundation (The Genome Denmark platform, grant 019-2011-2).

Author Contributions: SB, JMGI, AJH and JFN conceived and designed the study; KRK and JFN compiled the metadata features; KRK, SM, MA, RHJ, TAH, ARI, SRR, IBN, DEAP, PVSO, LV and HF prepared sequencing libraries and performed laboratory experiments; JFN performed the computational analyses; JFN, JMGI, SB, OL, and TSP analysed the data; JFN, AJH, JMGI and SB wrote the manuscript; all authors discussed, read and approved the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References


© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons by Attribution (CC-BY) license ([http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/)).
Improved Epitope Prediction Utilizing Cognate Antibody Information

Kamilla K. Jensen¹, Martin C. Jespersen¹, Paolo Marcatili¹ and Morten Nielsen¹,²

¹Department of Bio and Health Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark
²Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Argentina

Abstract:
Identification of B-cell epitopes plays a critical role in development of vaccines and immunotherapy. Presently the experimental methods used to identify B-cell epitopes are both time consuming and expensive and the development of reliable tools for predicting epitopes is of considerable interest both to industry and basic research. Most existing methods relying only on antigen features for identification of B-cell epitopes, maintain relative poor performance. In a study by Sela-Culang et al. [1], it was however shown that utilizing both antibody and antigen features can improve B-cell epitope predictions. Here this approach was extended, and two different prediction pipelines were constructed, one using only the sequences and the other using the structures. This was done in order to predict the likelihood of an antibody residue binding to antigen residue. Random Forest prediction models were made using information (sequential and/or structural) from antibody-antigen crystal structures obtained from the Protein Data Bank. These predictions were improved using a singular value decomposition method and further improved using sequential smoothing. A docking protocol was constructed using these predictions to identify the most likely location of epitopes. To simulate this in the sequential pipeline a Monte Carlo was performed. Results proved that the sequential and the structural pipeline significantly outperformed state-of-the-art epitope prediction tools and both pipelines showed that utilizing information of the cognate antibody significantly enhances epitope prediction. The sequential pipeline performed 93% better than state-of-the-art and the structural pipeline performed 95% better.

References
Rapid identification of tuberculosis mutations associated with antimicrobial resistance using NGS technology

Katja Einer-Jensen¹, Amanda C Brown², Josephine Bryant³, Mette T Christiansen³, Martin Ludvigsen¹, Poul Liboriussen¹, Jens Johansen¹, Jolyon Holdstock², Daniel P Deplede³, Helena Tutill³, Jacqueline ZM Chan², Mike McAndrew², Rachel Williams³, Graham Speight², Martin Schutten⁴, Judith Breuer³

¹QIAGEN Aarhus, Denmark
²Oxford Gene Technology, United Kingdom
³University College London, United Kingdom
⁴University Medical Center Rotterdam, The Netherlands

Abstract:
Personalized treatment of infected patients is dependent on detailed characterization of the infectious agent. Recent advances in targeted- as well as whole genome sequencing allow Next Generation Sequencing (NGS) to be considered viable in clinical laboratory settings. In the EU-FP7 funded “PathSeek” project, efforts are focused on two challenges: 1) Establishment of a methodology to sequence pathogens extracted directly from clinical specimens; 2) Development of an automated bioinformatics solution, which, based on NGS data, enables identification of present pathogen(s), identification of variants known to influence drug resistance, and lastly, genotyping of the identified pathogen(s).

This presentation introduces an automated bioinformatics solutions established for the critical evaluation of NGS data from patients infected with Mycobacterium tuberculosis (Mtb). In a single workflow, raw sequence data is trimmed, mapped and realigned against the Mtb H37Rv reference sequence. Mutations known to be associated with antimicrobial resistance are subsequently automatically identified, and categorized depending on their occurrence in the resistant majority or minority of the population. In the same workflow, the consensus sequence of predefined targeted regions is reported and used for generating a phylogenetic tree showing the genetic grouping of the analyzed sample. The solution is easy to use and delivers unified reporting and documentation in condensed views, aimed at novice bioinformatics users. The various outputs formats in addition to a simple report include overview tables and data visualization, which offers experienced users the option for a detailed review of the findings.

The decision support tool presented here has the potential to revolutionize patient management by delivering all the data required for truly personalized infectious disease treatment when compared to the currently practiced empirical treatment of MTB.
kmtDNA: Extraction and de novo assembly of mitochondrial DNA from whole genome next generation sequencing data

Kosai Al-Nakeeb¹, Thomas Nordahl Petersen¹, Thomas Sicheritz-Pontén¹

¹DTU Bioinformatics, Departments of Bio and Health informatics, Technical University of Denmark, Kemitorvet, Building 208, 2800 Kgs. Lyngby, Denmark

Abstract:
Unknown mitochondrial DNA can be extracted and assembled from whole genome sequencing data mainly by aligning reads against known partial or full mitochondrial sequences of the target organism or a close relative. We have developed a method that can do this without the need of a reference sequence. This is done by separating reads of mitochondrial origin from genomic reads using the kmer frequencies of each individual read under the assumption that mitochondrial reads are more abundant in the library than genomic reads. The method then tries to de novo assemble these reads into a full or partial mitochondrial genome. We were able to assemble the mitochondrial genomes of various organisms with varying success. De novo assembly using only mitochondrial reads can be complicated by factors such as repeat regions on the mitogenome and aspects of the library design, which may result in partial mitogenomes being assembled and/or lower quality in these regions. The program is available for both Linux and Macs.
The Landscape of Isoform Switching in human cancer

Authors:
Kristoffer Vitting-Seerup*, Albin Sandelin

*author to present

Affiliations:
Section for Computational and RNA Biology (SCARB), Department of Biology, University of Copenhagen, Denmark.
Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Denmark.

Abstract:
Isoform switching, which referees to the differential usage of different gene-isoforms in different conditions, are with a few exceptions largely overlooked in cancer biology. This lack of knowledge probably occurs in part because it is difficult to find and predict the functional impact of such switches and in part because the extend of isoform switching in cancers is not known.

To solve these problems we developed IsoformSwitchAnalyzeR, an easy to use R package which enables statistical identification, annotation and visualisation of isoform switches. We used IsoformSwitchAnalyzeR to identify isoform switches 12 cancer types covering almost 6000 cancer patients from The Cancer Genome Atlas (TCGA). We find that isoform switches are extremely common: across the 12 solid cancer types more than 4000 (34% of all) multi-transcript genes display differential isoform usage in at least one cancer type. In 2500 of these genes (20% of all genes) the changes have easily predicted functional consequences such as domain loss, domain switch or loss of coding potential. The genes with isoform switching are not random, but are highly enriched for genes in cell signalling, adhesion and cancer signatures. Many for the found isoform switches are furthermore pan-cancer events and we both identify known isoform switches in new cancer types as well as describe novel pan-cancer isoform switches.

At The Annual Danish Bioinformatics Conference I will mainly focus on communicating the results from the pan-cancer analysis of isoform switching.
Proteogenomic analyses for discovery of bi-specific chimeric antigen receptor targets

Christian Simon¹, Mike Bogetofte Barnkob², Lars Rønn Olsen²,³

¹Disease Systems Biology, Novo Nordisk Center for Protein Research, University of Copenhagen, Copenhagen, 2200, Denmark.
²MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, OX3 9DU, United Kingdom. ³Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, 2800, Denmark

Abstract:
Current clinical successes of chimeric antigen receptor (CAR) T cells have been achieved by targeting single, malignancy-specific surface molecules. In order to broaden the treatment applicability to more cancer types, progress is being made in developing bi-specific CARs that require binding to several ligands before lymphocytes are activated. Such technical advances hold the potential to increase the specificity of the therapy and thereby decrease off-target effects. Selecting good bi-specific CAR targets is balance between specificity and sensitivity of tumor cell targeting - High specificity is required for clinical safety, and high sensitivity is required for efficient clearance of the targeted cancer type.

Well-defined cell surface protein expression profiles are needed to facilitate target selection. Defining surface protein expression profiles for CAR therapy in silico require information about expression of a large number of surface proteins on a large number of cells at different states and differentiation stages. At present, no high-throughput technique for measuring expression of large numbers of surface proteins exists. However, progress is being made with emerging technologies, such as chip cytometry and mass cytometry. In addition, surface molecule expression on individual cells has been measured at low rates using immunohistochemistry or flow cytometry for decades, and vast amounts of cell-specific expression has been published. Combined, these resources form a rich, but unstructured source of data and information.

To facilitate the definition of unique surface molecule profiles, we have collected and organized large amounts of protein expression data on human hematopoietic cells from the cytometric methods as well as the primary literature. For the latter, we employed text mining techniques for article classification and subsequently extensive manual curation. We coupled these data with analysis of large-scale transcriptomics data for the surface proteins in order to assemble a data foundation for deep characterization of cell surface profiles. The resulting database contains expression of 457 surface proteins across 279 hematopoietic cells. We then developed algorithms for data mining to define unique protein expression profiles for highly sensitive and specific CAR targeting. Ongoing efforts will expand the database to contain surface protein expression for cells in all human tissues, as well as experimental validation of potential CAR targets.
Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease

Lasse Folkersen1,3, Eric Fauman2, Maria Sabater-Lleal3, Rona J Strawbridge4, Mattias Frånberg3, Bengt Sennblad5, Damiano Baldassarre4,5, Fabrizio Veglia3, Steve E. Humphries6, Rainer Rauramaa7, Ulf de Faire7, Andres J. Smit8, Philippe Giralt9, Sudhir Kurli1, Elmo Mannarino12, Stefan Erroths13, Åsa Johansson13, Sofia Bosdotter Erroths14, Stefan Gustafsso15, Lars Lind13, Cecilia Lindgren16, Andrew P Morris17, Vilmantas Giedraitis16, Angela Silveira3, Anders Franco-Cereceda18, Elena Tremolli3,5, the IMPROVE study group, Ulf Gylensten13, Erik Ingelsson15,19, Søren Brunak1, Per Eriksson3, Daniel Ziemek2, Anders Hamsten3, Anders Målarstig3,20

Affiliations:
1Department of Systems Biology, Technical University of Denmark, Copenhagen, 2800, Denmark
2Pfizer Worldwide Research & Development, Cambridge, MA, 02139, USA
3Cardiovascular Medicine, Department of Medicine Solna, Karolinska Institutet, Stockholm, 17176, Sweden
4Dipartimento di Scienze Farmacologiche e Biomolecolari, Università di Milano, Milan, Italy.
5Centro Cardiologico Monzino, IRCCS, Milan Italy.
6British Heart Foundation Laboratories, University College of London, Department of Medicine, Rayne Building, London, WC1E 6BT, United Kingdom
7Foundation for Research in Health Exercise and Nutrition, Kuopio Research Institute of Exercise Medicine, Kuopio, 70100, Finland.
8Division of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, and Department of Cardiology, Karolinska University Hospital, Solna, Karolinska Institutet, Stockholm, 17176, Sweden.
9Department of Medicine, University Medical Center Groningen, Groningen, 30001, the Netherlands.
10 Assistance Publique - Hopitaux de Paris; Service Endocrinologie-Metabolisme, Groupe Hôpitalier Pitie-Salpetriere, Unités de Prévention Cardiovasculaire, Paris, 75013, France.
11Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio Campus, Kuopio, 80101, Finland.
12Internal Medicine, Angiology and Arteriosclerosis Diseases, Department of Clinical and Experimental Medicine, University of Perugia, Perugia, Italy.
13Department of Immunology, Genetics and Pathology, Science for Life Laboratory Uppsala, Uppsala University, Uppsala, 75108, Sweden
14Department of Internal Medicine, Uppsala University Hospital, 75185 Uppsala, Sweden
15Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, 75185, Sweden
16Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom
17Department of Biostatistics, University of Liverpool, Liverpool, L69 3BX, United Kingdom
18Cardiothoracic Surgery Unit, Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, 17176, Sweden
19Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA.
20Pfizer Worldwide Research and Development, Stockholm, 17176, Sweden

Abstract:
Recent advances in highly multiplexed immunoassays have allowed systematic large-scale measurement of hundreds of plasma proteins in large cohort studies. In combination with genotyping, such studies offer the prospect to 1) identify mechanisms involved with regulation of protein expression in plasma trough protein quantitative trait loci (pQTL), and 2) determine if the plasma proteins are likely to be causally implicated in disease through Mendelian Randomization (MR). To this end we are conducting a genome-wide association studies (GWAS) of 83 proteins considered relevant to cardiovascular disease (CVD), measured in 3,394 individuals with multiple CVD risk factors. We identified genome-wide significant (p<5e-8) association signals for 78 loci with 56 proteins, most of which replicated in separate validation studies (n=2,639 individuals). Using automated text mining, manual curation, and network-based methods incorporating information on expression quantitative trait loci (eQTL), we propose plausible causal mechanisms for 25 trans-acting loci. Using public GWAS data we further evaluate all 79 loci for their causal effect on coronary artery disease, and highlight potentially causal associations. Overall, a majority of the plasma proteins studied showed evidence of regulation at the genetic level. Our results enable future studies of the causal architecture of human disease, which in turn should aid discovery of new drug targets. Further, our work with meta-analysis of human pQTL studies is paving the way forward towards formation of large proteomics/genomics cohort in the Scandinavian genetics collaboration for Olink Biomarkers, hosted in the Tryggve platform for Nordic exchange of sensitive data.
Illegal drugs and bioinformatics - Classification and stability of amphetamine impurities applied in criminal investigations

Louise Stride Nielsen¹, Christian Lindholst¹, Palle Villesen²

¹Department of Forensic Medicine, Aarhus University, Palle Juul-Jensens Boulevard 99, DK-8200 Aarhus N, Denmark
²Bioinformatics Research Centre, Aarhus University, C. F. Møllers Allé 8, DK-8000 Aarhus C, Denmark

Abstract:
Amphetamine impurity profiles are extensively used to establish links between different illicit seizures. This type of forensic evaluation can be used to provide investigators and courtrooms with important forensic evidence in criminal cases. The principal challenge is, however, for the classification model to accurately distinguish between linked and non-linked seizures even though the impurity profile may change over time due to different storage conditions. In present study, stable and unstable amphetamine target compounds are evaluated in order to optimise the classification model of illicit amphetamine seizures.

Data consisted of four major amphetamine seizures and 164 amphetamine seizures originating from different seizures. The four major amphetamine seizures was repacked and exposed to different storage conditions over a period of 12 months. The impurity profiles consisting of 26 amphetamine target compounds were analysed using GC-MS. Logistic regression (binary classification) was used to fit a linear model of the pairwise amphetamine impurity distances. The best model was selected using backward and forward model selection and 3-fold cross validation and finally validated in an independent dataset. Results were compared with a standard model from the literature.

The model performance was improved 5 times using an 11 target compound combination compared to the standard model. In addition, backward and forward model selection was able to find models which were better than one million random models. The amphetamine impurity profile was less susceptible to different environmental conditions compared to the standard model but it was in general still susceptible to changes. In conclusion, unstable target compounds should be down weighted when comparing profiles and we have identified and validated a superior model that is far better than the standard model. The amphetamine profile is, however, susceptible to different environmental exposures and this should therefore be considered when amphetamine profile comparison is conducted in criminal cases.
XploreRNA: a fast, fully-automated cloud based framework for experimentally validated differential transcriptome analysis.

Man-Hung Eric Tang, Jesper C. Tholstrup, Asli Özen, Dhany Saputra, Jacob B. Axelsen, Yuan Mang, Iben Plate, Niels Tolstrup
(Corresponding author met@exiqon.com)

Abstract:
The advent of Next-Generation Sequencing technologies has allowed comprehensive analyses of patterns of gene expression and a better elucidation of the different contributions of transcript isoforms. While offering unprecedented resolution and scope, the genomic scale of RNA sequencing data has brought new challenges in the mining of scientifically meaningful biological information and the appropriate selection of relevant gene targets among thousands of candidates for follow-up experiments. We present XploreRNA, a fast, fully-automated differential transcriptome analysis framework that assists bench researchers in their selection of biologically relevant gene targets and offers them the possibility to order custom validation assays (qPCR or antisense Gapmers) for their gene transcripts of interest. XploreRNA integrates well-established NGS and functional analysis software with in-house tools such as unsupervised classification, custom design of oligonucleotides for user-selected targets. XploreRNA summarizes its findings in a detailed analysis report and allows the researcher to navigate into the transcriptome analysis results online using the Gene Wizard and order custom assays through the same web portal. XploreRNA is fast, cloud-based, available from all screens (computers, mobile devices) and does not require prior bioinformatics expertise or computational resource.
Characterizing age-dependent regulatory variation in the human frontal lobe region

Maria Dalby\textsuperscript{a}, Cornelis Blauwendraat\textsuperscript{b}, Sarah Rennie\textsuperscript{a}, Margherita Francescatto\textsuperscript{b}, Patrizia Rizzu\textsuperscript{b}, Peter Heutink\textsuperscript{b}, Robin Andersson\textsuperscript{a}

\textsuperscript{a}The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{b}German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany

Abstract:
Age is a major risk factor for most common neurodegenerative diseases, such as Alzheimer’s disease, Parkinson’s disease, and cognitive impairment. While a considerable effort has been made to associate gene expression changes with neurodegenerative disease state and etiology, the general impact of aging on human transcription and transcriptional regulation remains to a large degree unexplored. To this end, we have characterized gene expression levels and regulatory activities over the healthy human aging process, with a particular focus on the frontal lobe brain region.

We systematically characterized the frontal lobe transcriptome using Cap Analysis of Gene Expression (CAGE) on total RNA isolated from post mortem frontal lobe samples of 144 healthy individuals with an age span from 2 to 95 years. From this data, we determined genome-wide activities of frontal lobe gene promoters and transcriptional enhancers and the effect of aging on transcriptional and regulatory variability.

Via detailed characterization and clustering of the major age related trends we identified a conservative set of gene promoters (4.3%) demonstrating significant expression changes across lifespan. The identified set recapitulates known genes associated with senescence in non-human species. Similarly, we identified a fraction (2.4%) of enhancers active in frontal lobe displaying significant expression association with chronological age. Our detailed map of regulatory variation over the human aging process will enable assessments of the impact of regulatory genetic variants on age-related transcriptional programs as well as a focused analysis of genetic variants associated with major neurodegenerative diseases.
Transcriptomic changes in sensitive and tolerant ryegrass (*Lolium perenne*) facing drought

Maria de la O Leyva Pérez 1, Stephan Hentrup 1, Stephen Byrne 2, Torben Asp 1

AU Flakkebjerg. Department of molecular biology and genetics. Denmark. 
Teagasc. Oak Park Crops Research, Carlow. Ireland 
mleyva@mbg.au.dk; torben.asp@mbg.au.dk

Abstract:
In 2014 we tested 3 plants from more than 500 genotypes against drought and found some genotypes that seemed more tolerant to drought and others that seemed more susceptible. In 2015 we designed experiments to set 12 plants from 10 genotypes (5 sensitive and 5 tolerant to drought) at the desired humidity with the help of individual humidity sensors (flower power, Parrot). Then we kept 4 plants at optimal humidity as control plants. Eight plants from each genotype were less watered to get them to be at a humidity of 20% and they started to be scored every day before watering. After we got all those plants to be at 20% for a few days to adapt, we sampled those plants and stopped watering. We scored every day and sampled every second day. According to the scoring average, we selected 2 genotypes that looked more resistant to drought and 2 sensitive genotypes. We extracted RNA for 3 replicates of both control plants and drought plants at 4 time points. This made 96 samples for RNAseq. In order to compare different genotypes properly, we mapped the reads to a common reference built recently in our group (Byrne et al., 2015) using Trinity pipeline. For the tolerant genotypes, we found 1.032 and 3.608 differentially expressed (DE) genes along the drought experiment, respectively. For the sensitive cultivars, we found 2.303 and 1.291 DE genes. We extracted the DE genes that both tolerant cultivars had in common and subtracted the DE genes in common with the sensitive genotypes. As a result, we found 83 genes exclusively upregulated in both tolerant cultivars but not differentially expressed in the sensitive genotypes. We think these genes can give valuable information to identify breeding material with larger drought tolerance.

Abstract:

Adaptation to specialized diets often requires modifications at both genomic and microbiome levels. We applied a hologenomic approach to the common vampire bat (*Desmodus rotundus*), one of the only three obligate blood-feeding (sanguivorous) mammals, to study the evolution of its complex dietary adaptation. Specifically, we assembled both its high quality genome (N50=26.9 Mb) and gut microbiome metagenome and compared them against those of insectivorous, frugivorous, and carnivorous bats. Our analyses showed i) a particular common vampire bat genomic landscape regarding viral elements, ii) a dietary and phylogenetic influence in its gut microbiome taxonomic and functional profiles, and iii) that both genetic elements harbor key traits related to the nutritional (e.g. vitamins shortage) and non-nutritional challenges (e.g. blood borne pathogens and osmotic homeostasis) of sanguivory. These findings highlight the value of a holistic study of both host and microbiota when attempting to decipher adaptations underlying radical dietary lifestyles.
LAST, Open-source Software For General-purpose, Large-scale Sequence Comparison And Alignment

Martin Frith
Computational Biology Research Center, AIST, Tokyo

Abstract:
It is the only aligner that combines a traditional substitution score matrix (which models sequence divergence) with per-base uncertainty (e.g. fastq) in a rigorous way. This is useful for alignments with non-negligible divergence (e.g. cross-species alignment, ancient DNA) or unusual base frequencies (e.g. malaria, bisulfite converted DNA). It uses the statistical (pair HMM) basis of alignment to annotate the reliability of every column in an alignment.

It is the only tool that can align DNA to proteins, *allowing frameshifts*, for genome-scale data. This is useful for: annotating pseudogenes, and analyzing metagenomic DNA (where frameshifts are surprisingly common).
It can do "split alignment" of a query sequence to a genome, where it looks for a unique best match for each part of the query. It rigorously calculates the reliability (uniqueness) of each part of the alignment. This is useful for: cancer (DNA reads that cross rearrangement breakpoints), spliced RNA (where it models splice signals, intron sizes, and allows trans-splicing), and whole genome comparison (where different parts of one query chromosome match different parts of the target genome).

Using newly-optimized transition seeds, LAST found ~20,000 new alignments between the human and mouse genomes, which are missing in the standard UCSC genome alignments.
Rapid detection of drug-resistance – a successful industry-academic cooperation

Katja Einer-Jensen, Martin Simonsen.
QIAGEN Aarhus A/S, Aarhus C, Denmark.

Abstract:
At QIAGEN Aarhus we are involved in multiple international research projects within the field of bioinformatics. Through these projects we keep in touch with the scientific community and develop innovative solutions for our customers. One particular successful research project which has given rise to several interesting solutions is the EU funded PATHSEEK project.

In PATHSEEK, four partners sought to enable personalized treatment of infected patients in clinical settings based on detailed characterization of the infectious agent using Next Generation Sequencing (NGS). Efforts were focused on two challenges:
Establishment of a methodology to sequence the entire genome of pathogens extracted directly from clinical specimens.
Development of an automated bioinformatics solution.
As the result of close dialog with the partners, QIAGEN’s contribution was the development a bioinformatics analysis pipeline which enables automatic identification of the present pathogen(s), identification of pathogen specific variants known to influence drug resistance, and lastly, genotyping of the identified pathogen(s).
The solution is easy to use and delivers unified reporting and documentation in condensed views, aimed at novice bioinformatics users. The various outputs formats in addition to a simple report include overview tables and data visualization, which offers experienced users the option for a detailed review of the findings.
The decision support tool presented here has the potential to revolutionize patient management by delivering all the data required for truly personalized infectious disease treatment when compared to the currently practiced empirical treatment of MTB.
The promoter and enhancer landscape of inflammatory bowel disease: finding predictors of inflammation and disease state.

Mette Boyd1,2#, Jette Bornholdt1,2#, Morana Vitezic1,2#, Malte Thodberg1,2#, Kristoffer Vitting-Seerup1,2, Mehmet Coskun1,2,7, Yun Chen1,2, Yuan Li2,7, Anders Gorm Pedersen3, Kerstin Skovgaard4, Robin Andersson1, Thilde Bagger Terkelsen1,2, Axel Thieffry1,2, Berit Lilje1,2€, Jesper Troelsen5, Gerhard Rogler6, Pia Klausen7, Bobby Lo1,2,8, Anders Munk Pedersen8, Kim Jensen2, Jakob Benedict Seidelin7, Ole Haagen Nielsen7, Jacob Tveiten Bjerrum7*, Albin Sandelin1,2*

1The Bioinformatics Centre, Department of Biology, Ole Maaloes Vej 5, DK2200 Copenhagen N, Denmark
2Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaloes Vej 5, DK2200 Copenhagen N, Denmark
3Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, Denmark
4National Veterinary Institute, Technical University of Denmark, Bülowsvej 27, DK1870 Frederiksberg C, Denmark
5Department of Science, Systems and Models (NSM), Roskilde University, DK4000 Roskilde, Denmark
6University of Zurich
7Herlev Hospital
8Hvidovre Hospital

Abstract:
Inflammatory bowel disease is a chronic inflammatory bowel disorder. It is classified into two major entities: Ulcerative colitis (UC) and Crohn’s disease (CD). The distinction between CD and UC is critical for correct management, especially options of surgery and personalized treatment, yet the diagnosis is challenging, approximately 10% of patients are classified as ‘indeterminate colitis’. Therefore, novel biomarkers for stratifying patients and improving diagnostics are highly needed.

Here, we have applied a unique RNA sequencing technique, Cap Analysis of Gene Expression (CAGE), on intestinal biopsies from 94 patients with IBD as well as healthy controls. This provided a genome wide atlas of active transcription start sites and enhancers. We show that highly expressed immune cell related transcripts were powerful predictors of the degree of inflammation (controls vs CD & UC), whereas, low expression of epithelial cell related transcripts, including previously unidentified long non-coding RNAs and alternative TSSs of known genes, were far more powerful predictors of UC vs. CD. Using the enhancer atlas we show that enhancer transcription can distinguish the inflammatory state of patients and that similarly regulated enhancers and promoters share transcription factor binding activity. Utilizing GWAS data we show that enhancers are more enriched for the heritability of IBD than promoters. Finally, we show that putative regulatory interactions of individual enhancers and promoters can be computationally inferred.

In addition to providing insights into IBD pathogenesis and potential IBD biomarkers, this study also illustrates the general usefulness of CAGE for building disease-specific promoter and enhancer atlases.
Disease trajectories in Schizophrenia - a clustering of patients based on sequences alignment of health registry diagnoses

Morten Dybdahl Krebs,

Mental Health Services Copenhagen, Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen, Denmark; Initiative for Integrative Psychiatric Research, iPSYCH, The Lundbeck Foundation, Copenhagen, Denmark;

Abstract:

**Background** Schizophrenia is thought to be a multifactorial disorder, with great clinical heterogeneity and a relatively high heritability. The aim of this study was to use trajectories in diagnoses of psychiatric disorders to identify clusters of patients with schizophrenia. Clusters were described by a number of independent variables to indicate whether they represented distinct subgroups of patients.

**Methodology** In a registry-based cohort of patients with schizophrenia, multichannel state sequence objects were created based on the psychiatric diagnoses registered. Using sequence alignment and cluster analysis with a variety of different settings and comparing validation measures of cluster separation, a cluster solution was chosen. The cluster solution was further evaluated by chi-square tests and analyses of variance (ANOVA) using a number of independent exposure variables.

**Results** 3431 patients with schizophrenia were included in the analysis. None of the clusterings reached the validation values expected for well-separated clusters. The best performing cluster solution were found to be a hierarchical clustering using Ward’s method and 13 clusters. Clusters differed significantly by gender, age of onset and maternal age, but not by paternal age and schizophrenia in relatives.

**Conclusion** Cluster analysis based on sequence dissimilarity can be used to identify groups of patients with similar disease trajectories in schizophrenia. These groups differ on several characteristics not included in the original analysis. Further studies are necessary to reveal if between group differences are due to common risk factors, detection bias or other effects.
Comprehensive pathway comparison between human and animal models

Nadezhda T. Doncheva¹,², Oana Palasca¹,², Jan Gorodkin², Lars Juhl Jensen¹,²

¹ The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
² Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Copenhagen, Denmark

Abstract:
Nowadays, animal models are indispensable for the study of human diseases and the development of new treatment therapies. One of the main challenges thereby is to select the regulatory pathways and genes (coding as well as non-coding) in an animal model that would be useful to generate reliable hypotheses about the molecular mechanisms and processes in human and, in particular, with respect to a disease of interest. In order to transfer knowledge from one species to another, it is crucial to take into account the intrinsic differences in cellular organization and regulation between human and animal models.

Therefore, we will perform a careful and thorough analysis of the similarities and differences between human and animal pathways using publicly available data on gene expression in healthy organisms, orthology relationships, regulatory interactions of coding and non-coding genes as well as experimentally determined protein-protein interactions. Our preliminary analysis demonstrates that pathways and interaction networks for animal models are often derived only using orthology and are not organism- and tissue-specific enough. In order to compare how pathways and protein interactions vary among species and tissues, we will combine pathway data from well-known resources such as KEGG and Reactome with tissue expression data from the updated and extended version of the TISSUES database, which now covers several animal models.
Design of LNA oligonucleotides with a memetic algorithm

Niels Tolstrup, Jesper Tholstrup, Asli Özen
(Corresponding author nt@exiqon.com)

Abstract:
Background
Within the field of RNA analysis there is a need for highly specific and sensitive oligonucleotide based assays, these include assays for qPCR, inSITU, knockdown, miRNA mimics, antisense and several more. Common for design of these assays are parameters such as the melting temperature of the oligonucleotides, hybridization to other oligonucleotides present in the assay, efficient hybridization to the target and no off-target hybridization. The requirement to efficiently hybridize with the target while simultaneously avoiding hybridization to any other sequences requires a very accurate design process that optimizes all the in part conflicting design parameters simultaneously.

Methods
We present a memetic algorithm [1] for the design of highly specific and sensitive oligonucleotide based assays. The memetic algorithm combines the advantages of genetic algorithms with domain specific knowledge about the oligonucleotide design. This approach allows the design of the algorithm to focus on the scoring of different oligonucleotide properties.

Results
When applied, the algorithm will focus in on highly interesting parts of the oligonucleotide design space, thereby giving better oligonucleotide specificity and significantly faster runtime. The design algorithm has been implemented in Ruby and important structural parts of the algorithm are discussed in the poster.

Conclusion
We have found the use of memetic algorithms in the field of miRNA analysis to be highly efficient.

References:
Covariance Association Test (CVAT) Identify Genetic Markers Associated with Schizophrenia in Functionally Associated Biological Processes

Palle Duun Rohde¹,²,³, Ditte Demontis²,³,⁴, Beatriz Castro Dias Cuyabano¹, The GEMS Group^, Anders D. Børglum²,³,⁴ and Peter Sørensen¹

¹Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark; ² Centre for Integrative Sequencing, iSEQ, Aarhus University, 8000 Aarhus, Denmark; ³ The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, 8000 Aarhus, Denmark; ⁴ Department of Biomedicine, Aarhus University, 8000 Aarhus, Denmark; ^ The GEMS (Genomic Medicine for Schizophrenia) Group Ole Mors, David M. Hougaard, and Preben Bo Mortensen

Abstract:
Schizophrenia is a psychiatric disorder with large personal and social costs, and understanding the genetic etiology of this disorder is important. Such knowledge can be obtained by testing the association between a disease phenotype and individual genetic markers. However, such single marker methods have limited power to detect genetic markers with small effect sizes. Aggregating genetic markers based on biological information might increase the power to identify sets of genetic markers of etiological significance. Several set test methods, and here we propose a new set test method derived from genomic best linear unbiased prediction (GBLUP): the Covariance Association Test (CVAT). CVAT measures the covariance between the joint genetic values for the genetic markers in the feature set and the genetic values for the remaining genetic markers, outside the feature set. A feature set is a group of genetic markers defined by biological information, e.g. genes and pathways. Because CVAT was derived from GBLUP, many extensions exists. One such is to allow the genetic marker effects to be a mixture distribution. Such mixture distribution can be obtained by adding an extra genetic variance component to the GBLUP. The two genetic variance components contained either the genetic effects of rare or common genetic markers. We compared the performance of CVAT to other commonly used set test methods using a simulated population with the same genetic parameters as schizophrenia. We found that CVAT was among the top performer. When extending CVAT to utilize a mixture of SNP effects, we found an increase in power to detect the causal sets of genetic markers. Applying the methods to a Danish schizophrenia case-control data set, we found genomic evidence for association of schizophrenia with vitamin A metabolism and immunological responses, which previously have been implicated with schizophrenia based on experimental and observational studies.
Inference of distribution of fitness effects: impact of beneficial mutations

Paula Tataru¹, Maëva Mollion¹, Sylvain Glemin² and Thomas Bataillon¹
¹Bioinformatics Research Centre, Aarhus University, Aarhus, 8000, Denmark
²Evolutionary Biology Centre, Uppsala University, Uppsala, 752 36, Sweden

Abstract:
New mutations are the ultimate source of heritable variation, and their fitness effects determine the possible evolutionary trajectories a population can follow. The relative frequencies of the fitness effects of new mutations are best specified through a distribution of fitness effects (DFE) that spans deleterious, neutral and beneficial mutations. As such, the DFE is key to several aspects of the evolution of a population and inference of DFE from patterns of polymorphism and divergence has been a longstanding goal of evolutionary genetics. The large amount of datasets currently available enables us to learn and compare the DFE in numerous species. Understanding what determines variation in DFE from species to species can foster a deeper understanding of the forces underlying the process of evolution.

Here, we revisit the problem of reliably estimating the DFE from site frequency spectrum (SFS) and divergence data. We developed a hierarchical probabilistic framework that extends significantly available methods, and using simulation studies, we investigate and question a widespread assumption: beneficial mutations contribute only negligibly to the SFS data. Current methods invariably assume that mutations present in the SFS are either neutral or deleterious. Consequently, positive selection is only inferred when SFS data can be complemented with divergence data. We show that this assumption is theoretically unfounded and that the DFE of beneficial mutations can often be inferred from SFS data alone. This opens the doors to the analysis of less popular species, where divergence data is often not available. Additionally, we show that when the beneficial mutations in the SFS are not modeled, the resulting DFE estimation can be substantially biased. We illustrate these points by using both our newly developed framework, but also one of the most widely used inference methods available.
Typing and epidemiological clustering of common pathogens based on whole genome NGS data

Katja Einer-Jensen¹, Poul Liboriussen¹, Jens Johansen¹, Leif Schaueter¹, and Arne Materna¹

¹QIAGEN Aarhus, Denmark

Abstract:

Background: Next generation sequencing (NGS) data from whole pathogen genomes is frequently used for enhanced surveillance and outbreak detection of common pathogens. CLC Microbial Genomics Module 1.5 introduces functionality for molecular typing and epidemiological analysis of bacterial isolates. The module enables the user to perform a range of analyses and to take advantage of template workflows.

Methods: The isolate typing features include streamlined tools for NGS-based MLST, resistance typing as well as fast detection of genus and species. The taxonomy of the closest relative in the genome reference database is identified. Tools for epidemiological studies based on phylogenetic tree reconstruction using single nucleotide polymorphisms (SNPs) or K-mers from NGS reads or genomes enable the visualization of outbreaks sequence data in context of metadata. A new table format, acting as a database, collects typing results and associates these with metadata such as sample information, geographic origin, treatment outcome, etc.

Results: The CLC Microbial Genomics Module aims to facilitate tasks commonly carried out during outbreak investigation, such as typing or source tracking based on whole genome data by e.g. public health laboratories, government agencies, microbial pathology laboratories, and food biotechnological companies. The preconfigured workflows and simple deployment via integration into the widely used CLC Genomics Workbench and CLC Genomics Server ecosystem offer a user-friendly platform for scientists engaged in outbreak prevention and control.
Introducing the HyperProteoGenome

Praharshit Sharma*

Department of Experimental Design and Bioinformatics, Faculty of Agriculture and Biology, SGGW/, Warsaw University of Life Sciences, 166 Nowoursynowska, 02787, Warszawa, Masovian District, Poland.
*praharsit@iscb.org https://www.linkedin.com/in/praharshit

Abstract:

Background
Codon “Degeneracy”, “Usage” and “Bias” largely render the existing “Standard (Universal) Genetic Code” ambiguous, in terms of “exactGene finding”. The concept of “BIG Data Genetic Code” hypothesizes a novel relook into Genetic coding, per se [1] which can potentially improve blastx to near perfection.

Results and Discussion
Assuming a unique, 11 correspondence between the 4 nt (A| C|G|T) and the 4 colormap [2], we attempt to resolve a SuperComputational Problem (i.e.,) The HyperProteoGenome – basing on the following intuition based upon the “Principle of Computational Equivalence” [3]:
where '4' characterizes the 4nt and '20' characterizes 20aa, (nt= nucleotides, aa= aminoacids). For example, in a 1Dimensional ECA (Elementary Cellular Automata), we can Quantify with exactitude, 256 Algebraic/ Boolean rules, of which 88 are “fundamentally inequivalent” [4].

Conclusion
The conclusive remarks of this Ongoing work remain largely openended, inviting enrichingly interactive Opinions and Feedback from the broad Audience. As, \[e+1\] = Ceiling function of Napier’s constant (99.9455% approximation to the numerical solution of Equation above) is ‘3’, the currently assumed Codon Cardinality. Hence, can We expect a subsetGraphletmapping [5] of a polyhedral 4color Map w.r.t LHS to a regular 20polytope in RHS, characterized by PAM(z) – What is 'z'?

WEB References
1http://dx.doi.org/10.7490/f1000research.1111304.1
2https://en.wikipedia.org/wiki/Four_color_theorem
3http://www.wolframscience.com/nksonline/section12.4
4http://mathworld.wolfram.com/ElementaryCellularAutomaton.html
5http://science.sciencemag.org/content/353/6295/123.full
Computational Studies of Structured RNA Elements in Drosophila melanogaster

Kirsch R1,2, Seemann SE1,2, Ruzzo WL1,3,4, Cohen SM5, Stadler PF1,6,7,8,9,10, and Gorodkin J1,2

1Center for non-coding RNA in Technology and Health, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark
2Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Dyrlægevej 16, 1870 Frederiksberg C, Denmark
3Departments of Computer Science and Engineering and of Genome Sciences, University of Washington, Seattle, WA 98195, USA
4Fred Hutchinson Cancer Research Center, Seattle, WA, 98109 USA
5Department of Cellular and Molecular Medicine, University of Copenhagen, The Panum Institute, 3B Blegdamsvej, 2200 Copenhagen N, Denmark
6Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany
7Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
8Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1, 04103 Leipzig, Germany
9Department of Theoretical Chemistry, University of Vienna, Währinger Straße 17, 1090 Vienna, Austria
10Santa Fe Institute, 1399 Hyde Park Rd., NM87501 Santa Fe, USA

Abstract:
Non-coding RNAs (ncRNAs) have received increasing attention during the past decade as versatile actors in post-transcriptional regulation. Although remarkable progress has been made within the field by means of experimental as well as computational work, there is still a number of not yet characterized or entirely unknown structures, especially in organisms other than human and mouse. While a few computational screens to predict structured elements have been carried out in the model organism Drosophila melanogaster, the different approaches used vary in their preference for certain alignment features and their false discovery rates. In contrast to previous secondary structure prediction tools, CMfinder is less dependent on existing sequence-based alignments and their quality, since it exploits covariance while re-aligning given sequences to produce structure-based alignments and predictions.

Here, we apply CMfinder to the 15-way drosophilid MULTIZ alignment and find 22,000 putative structures with an estimated false discovery rate of ≤ 0.1. 12% of these are predicted in UTR exons, a similar number in coding exons, and a small number (1.5%) in non-coding exons. Almost half of the predictions lie in introns and about a quarter in intergenic regions and hence might belong to so far unknown transcripts. We find a strong enrichment of predictions within a range of existing ncRNA annotations including tRNAs, miRNAs, snoRNAs, lncRNAs and in particular snRNAs.

In the near future, we aim at using the combined information from both screens to find tissue-specific and developmental expression patterns in the fly. This way, we would like to add a level of evidence for functionality to unknown candidates and make novel statements in terms of non-coding expression in Drosophila melanogaster.
Improved D-Statistic for Low-Coverage Data

Author(s): Samuele Soraggi[1], Carsten Wiuf[2], Anders Albrechtsen[3]
Affiliation(s): [1,2] Copenhagen University, Department of Mathematics, Copenhagen, Denmark. [3] Copenhagen University, The Bioinformatics Centre, Copenhagen, Denmark.

Abstract:
Next Generation Sequencing provides a massive quantity of data, and it is widely analyzed in population genetics research. Since many NGS datasets are sequenced at low coverage, SNP and genotype calling have high uncertainty. This might affect statistical methods that make use of genotypes. A commonly used tool for detecting ancient admixture events is the D-statistic.

The D-statistic makes use of the patterns between alleles in four different groups of individuals in order to show the direction of an ancient or modern gene flow or to assess the correctness of a phylogeny of four populations in the configuration (H4(H3(H2,H1))), where H4 is an outgroup. For low-depth sequencing the D-statistic is highly susceptible to errors deriving from the SNP and genotype calling. For low-depth sequencing where genotype calling is not possible the method relies on sampling one allele from a single individual from each group at each site to evaluate ABBA and BABA patterns. This sampling procedure ignores much of the information in the data and only works for one individual from each group.

Moreover, the D-statistic does not allow to infer information on more complex phylogenies, for example having multiple admixture events or more than four groups of individuals. We have implemented in ANGSD a version of the D-statistic that does not require genotype calling but is still able to utilize all the information and allows for multiple individuals for each of the four groups. Using both simulations and real data we evaluate the method's power and false positive rate using different topologies. For example (African (CEU (Han Chinese, Native Americans))) is used to show a wrong phylogeny and the evidence of admixture from CEU to Native Americans.

We show that the weighted D-statistic has more power than the previous method of sampling only one base per site and is still powerful with the use of low-coverage data. For medium and high depth data the method has similar power to that of perfectly inferred genotypes. In addition the methods can be use even in the presence of high error rates which makes it especially appropriate for ancient DNA.
Chromosomal transcriptional initiation walks reflect three dimensional chromatin organisation.

Sarah Rennie & Robin Andersson

The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

Abstract:
The genome is organised into non-random and highly complex structures within the nucleus, which exhibit variable stability across different types of cells. Domains of interacting initiation events fall within so-called topologically associated domains, underlying the strong positional effect of enhancers and promoters localised linearly on the DNA to specific regions of a chromosome. Such domains themselves fall into proximity on a higher order level, providing a potential mechanism for distal interactions between units separated by many mega-bases. Whilst chromatin conformation capture technologies have come some way into identifying proximal regions of DNA, both cell type coverage and resolution is still limited, thwarting the ability to experimentally detect long range proximities and computationally predict pairs of interacting transcriptional units. We calculated simultaneous random walks across the initiation landscape of tissues over developmental stages in mouse, generating intriguing underlying latent structures reflecting the correlated positional effects of initiation events, separated from the layer of transcription attributable to a gene’s regulatory program independent of its genomic position. Intersecting walks between tissues then generates high resolution maps of active co-regulation and show striking similarities to chromatin organisation as identified by Hi-C. We compare our walks with a wealth of publicly available ChIP-seq data, demonstrating clear shifts between putative domain boundaries and compartments. We then identify active topologically associated domains specific to each of nine tissues across a variety of developmental stages at high resolution, pinpointing putative enhancer-promoter interactions, both within and between domains and allowing for comparisons at an unprecedented scale. Thus, this work reveals a fascinating picture of how higher order genomic activities vary across developmental time.
The dynamically developing mouse transcriptome

Sarah Rennie

Abstract:
Tissue development in mouse is a highly dynamic and coordinated process, involving specialised spatial temporal transcriptional patterns. We analysed a rich resource of transcription initiation from 134 FANTOM5 CAGE libraries covering nine tissues and fifteen time points spanning three stages in mouse development. We observe that a striking number of RNA species, including enhancer RNAs, shift dramatically in their relative transcript populations through developmental time, often suggesting tissue-specific functions. Extensive modelling of 40,082 promoters and 8,326 detected enhancers from embryonic day 12 to adult identified a wealth of developmentally dynamic signals and clustering by peak associates with stage- and tissue-specific functions, allowing us to pin-point potentially important novel transcripts. Furthermore, we observe that transcription as a whole within and between all tissues converges at a tissue specific point near to birth, suggesting that all tissues have similar requirements around birth, yet slightly different developmental timings. In all, our work reveals fascinating clues into the developmental programming of different tissues.
Functional Metagenomics, A Novel QIAGEN Bioinformatics Pipeline Enables Fast High-Quality Analysis of the Functional Content of Microbial Communities

Andreas Sand¹, Francesco Strino¹, Aske Simon Christensen¹, Martin Bundgaard¹, and Arne Materna¹

¹QIAGEN Aarhus, Denmark

Abstract:

**Background:** Detecting gene function in microbial communities based on metagenomic data is hard. Correctly measuring changes in the functional makeup between different metagenome samples is even harder. Lindgreen et al. showed that most of the benchmarked open source tools failed to correctly predict such changes at levels that are statistically significant. With our solution for microbial genomics you can more accurately detect and quantify functional elements in a sample. And the included statistical tools allow you to confidently measure statistically significant changes in function between samples. Multisample comparison is used to detect functional changes between samples and to identify samples with similar or diverging functional genomic elements. Data can be grouped and analyzed in the context of your sample-metadata. And powerful interactive visualisations can be used to discover and illustrate differences across samples or groups defined by metadata.

**Results:** Lindgreen et al. published a comprehensive, independent evaluation of 14 different whole metagenome analysis toolkits in Nature Scientific Reports in January 2016. We here compare our solution to five toolkits out of the fourteen that allow functional metagenome analysis using the test data published by Lindgreen et al. Statistical comparison (Edge test performed in CLC Genomics Workbench) of pairwise differential abundance of the individual functional elements predicted in the two test communities detects a statistically significant difference for all of the three functional elements that were analyzed in the paper: photosynthesis, nitrogen fixation and pathogenesis (all p-values < 0.01). Fold-changes predicted using our tools capture the expected overall pattern of functional changes and estimate the actual fold-change with higher precision than any other tool in all three functional roles.
Patterns in Protein Sequences

Steffen B. Petersen¹ and Jakob Bohr²:

¹: Email: Steffen1357@gmail.com, Research: https://www.researchgate.net/profile/Steffen_Petersen3
University: http://www.hst.aau.dk/research-groups-centres/biomedicine/medical-photronics/
Snail Mail: Medical Photonics Group, Department of Health Science and Technology, Aalborg University, 9220 Aalborg, Denmark.
²: Email: jabo@nanotech.dtu.dk
Snail mail: DTU Nanotech, Department of Micro and Nanotechnology, Technical University of Denmark, Building 345 East, Ørsteds Plads, DK 2800 Kongens Lyngby, Denmark.

Abstract:

30,000 protein sequences were downloaded from PIR. For each protein sequence the sequence location of every amino acid type was recorded. The gap between any two consecutive members of a particular amino acid type was cumulated for all the protein sequences. An identical dataset is collected for the randomized protein sequences, and subtracted from the first data. Surprising details is revealed by this simple method, indicating ancestral origins of the uncovered patterns as well as novel insight into the different properties of amino acid pairs that are broadly perceived as homologous. In the example below is shown the cumulative gap curve for ILE (top) and for LEU (bottom). LEU and Ile are widely assumed to be extremely homologous, differing only in the location of a methyl group. Yet the two curves document that ILE does not populate short repeats of length 1-10, whereas LEU clearly does.

If sequences contain many tandem repeats, then Fourier or Wavelet analysis will be a potential method for extracting such information, provided that insertion and deletions have not introduced phase shifts. However if only 1 or a few repeats are left intact by molecular evolution, such methods are likely to fail. The method proposed here will be more sensitive, since it is isolating the gap between two consecutive putative repeats.

Scale-Free Behaviour of Amino Acid Pair Interactions in Folded Proteins
Steffen B. Petersen1,2,3*, Maria Teresa Neves-Petersen1,4*, Svend B. Henriksen5., Rasmus J. Mortensen5,6., Henrik M. Geertz-Hansen7.
PLOSONE July 2012 | Volume 7 | Issue 7 | e41322

Hyperdimensional Analysis of Amino Acid Pair Distributions in Proteins
Svend B. Henriksen1, Rasmus J. Mortensen1., Henrik M. Geertz-Hansen1., Maria Teresa NevesPetersen2,3*, Omar Arnason1, Jo’n So’ring1, Steffen B. Petersen4, December 2011 | Volume 6 | Issue 12 | e256385.
Shining New Light on the Nature and Origin of Amino Acid Sequences and Protein Structures

Steffen B. Petersen

Email: Steffen1357@gmail.com, Research: https://www.researchgate.net/profile/Steffen_Petersen3
University: http://www.hst.aau.dk/research-groups-centres/biomedicine/medical-photonics/
Snail Mail: Medical Photonics Group, Department of Health Science and Technology, Aalborg University, 9220 Aalborg, Denmark.

Abstract:

For a number of years our focus has been on understanding the complex interplay between photons and biological material, in particular proteins. We now know that we can break disulfide bridges by exciting an aromatic amino acid (TRP or TYR). This triggers ionization of the aromatic electronic system. The electron migrates through the protein – probably maneuvering in the electrical field inside the protein. If the electron encounters a disulfide bridge, the disulfide bridge is reduced. This is the underlying mechanism for breakage of the disulfide bridge. The mechanism works up to a distance of at least 14 Å, thus for all practical purposes, if the protein contains both an aromatic residue as well as a disulfide bridge, it is likely that it will respond to UV illumination with opening of the SS bond. Smaller proteins and polypeptides are richly decorated with SS bonds – probably in order to stabilize the 3D structure, that is too small to have a substantial hydrophobic core. Larger proteins typically have less SS bonds per 100 amino acids. However, some receptor proteins are excessively decorated with SS bonds. This is the case for EGFR, the epidermal growth factor receptor. We have shown that this receptor is exceedingly sensitive to UV illumination, and as a consequence it loses its ability to bind its ligand EGF. EGFR is also a key target for many chemotherapeutics, because blocking the EGFR receptor also inhibits cancer growth– and our new insight indicates that we can block the same receptor with UV light.

Underlying all of the above is the 3D protein structure, which in turn is coded for by the amino acid sequence. Surprisingly we still do not know how to predict the 3D structure from the amino acid sequence alone. Sometimes we can use homology to proteins with similar amino acid sequences, to argue for 3D structural similarity. We decided to investigate both amino acid sequences and the packing of amino acid pairs in protein 3D structures from a radically new point of view. We will here present some hitherto unknown features of amino acid sequences as well as for the packing of amino acid pairs into folded protein structures. We will present some observations that indicate that underlying patterns are shared by all amino acid sequences, and that the packing of amino acid pairs into 3D folded structures are scale free, and follows the same organizational patterns as the world wide web.

Modulating the Structure of EGFR with UV Light: New Possibilities in Cancer Therapy
Manuel Correia1., Viruthachalam Thiagarajan2,3, Isabel Coutinho2, Gnana Prakash Gajula2, Steffen B. Petersen4,5, Maria Teresa Neves-Petersen2, PLOS ONE 2014, Volume 9 | Issue 11 | e111617
Scale-Free Behaviour of Amino Acid Pair Interactions in Folded Proteins
Steffen B. Petersen1,2,3*, Maria Teresa Neves-Petersen1,4*, Svend B. Henriksen5., Rasmus J. Mortensen5,6., Henrik M. Geertz-Hansen7. PLOS ONE July 2012 | Volume 7 | Issue 7 | e41322
Hyperdimensional Analysis of Amino Acid Pair Distributions in Proteins
Svend B. Henriksen1, Rasmus J. Mortensen1., Henrik M. Geertz-Hansen1., Maria Teresa NevesPetersen2,3*, Omar Arnason1, Jo `n So `ring1, Steffen B. Petersen4, December 2011 | Volume 6 | Issue 12 | e256385.
Disease Trajectory Comorbidity Browser, a tool for exploring, filtering and analyzing Danish disease trajectories.

Troels Siggaard, Anders Boeck Jensen, Søren Brunak

Abstract:
Bioinformatics tools have a profound impact on research and the translational ramifications of bioinformatics research. The Disease Trajectory Comorbidity Browser (DTCB) is a new tool with a focus on disseminating the positive effects of research in the area of co- and multimorbidities. We present a tool with novel disease trajectory research data for basic researchers, clinicians and software developers interested in the co-morbid landscape of diseases. Data from the population-wide Danish National Patient Register has carefully been aggregated and processed to make this DTCB tool possible.

DTCB is a bioinformatics web tool prototype in active development, with more features and data to come for future iterations of the application.

The DTCB web application will be demoed at the poster session as a live interactive demonstration, with the possibility of trying out the prototype.
Therapeutic Drug Quantification Using Targeted Mass Spectrometry

Michael Kruse Meyer¹,², Marlene Andersen¹, Troels Vindbæk Stausbo², Tue Bjerg Bennike²,³, Svend Birkelund², Grethe Neumann Andersen¹ & Allan Stensballe²

1) Division of Rheumatology/Center for Clinical Research, North Denmark Regional Hospital, Hjørring, DK.
2) Laboratory for Medical Mass Spectrometry, Aalborg University, Aalborg, DK.
3) Harvard Medical School, Department of Pathology, Boston Children’s Hospital, USA.

Abstract:

Background:
Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disease which aggressively degrades the joint cartilage ultimately destroying the joints. The concept of personalized medicine has been of interest since the beginning of this century (Leffers et al. 2011) and has become necessary in the field of rheumatology because not all patients show reduction in disease when treated with bDMARDs or when switching between the biologic agents. Therefore, development of methods predicting and monitoring the agents are of interest in order to treat effectively (Jong et al. 2014). During the past six years Division of Rheumatology, North Denmark Regional Hospital, has made an effort to reduce dosage and interval between dosage attempting to reduce medicine cost (Meyer, et. al. submitted). Before introduction of personal medicine, the treatment was based on an empirically decision without no biochemistry evaluation, i.e. drug quantification. Therefore, Division of Rheumatology has implemented a strategy which yield to improve patient care by quantify biological drug concentrations in the patients towards archiving an effective dosage. In this study, we applied a multiplexed targeted proteomics approach to quantify the CLTA4 inhibitor, CRP, and the compliment system to evaluate the effect on the innate immune system, when the adaptive immune system is inhibited.

Methods:
A total number of eleven patients diagnosed with RA in stable treatment with the costimulatory CTLA4 inhibition therapy (abatacept), are included and sampled during a period of one treatment interval. Additionally, a total number seven patients on the same drug is included and sampled prior to treatment and used as control, and decoy group. A standard ELISA preparation method described by (Willrich et. al. 2015) was optimized and prepared for analysis on a nanoLC-MS/MS instrument. Afterwards, the assay was further optimized for a parallel reaction monitoring (PRM) method used to carry out drug identification and absolute quantification. Finally, the assay was expanded to measure on the complement system, C-reactive protein, and an indexed retention time standard was included to increase cross-lab reproducibility.

Results:
Based on a pure standard, the assay was further in vitro spike-in optimized, and applied to the patients on stable costimulatory CTLA4 inhibition therapy. The results first show an interpersonal variation between patient drug concentration. Interestingly, the results indicate that several compliment proteins are less abundant with high patient drug concentration, and increases during decreased drug concentration. The study results need for further investigation and analysis which potentially could include patients who do not respond to bDMARDs. Participants in the patient group have all measurable drug concentrations.

Conclusion:
The above mentioned method development outlines the opportunity by using targeted mass spectrometry (MS) to quantify serum drug concentrations in patients with RA in stable treatment by incorporating the individually biochemistry profile of the patient prior treatment. As such, this targeted MS approach has the potential to improve both RA treatment and incorporating the
LSTM and convolutional neural networks to predict peptide binding to MHC II molecules

Vanessa Isabell Jurtz\textsuperscript{1}, Casper Kaae Sønderby\textsuperscript{2}, Søren Kaae Sønderby\textsuperscript{2}, Ole Winther\textsuperscript{2,3}, Morten Nielsen\textsuperscript{1}

\textsuperscript{1}DTU Bioinformatics, Technical University of Denmark  
\textsuperscript{2}Bioinformatics Centre, Department of Biology, University of Copenhagen  
\textsuperscript{3}DTU Compute, Technical University of Denmark

Abstract:
MHC II molecules are an essential part of the adaptive immune system and involved in the detection of extracellular pathogens. Predicting the interaction of peptides with MHC II molecules is important for the reliable identification T cell epitopes and in this way could advance the development of novel vaccines.

Here we show that the combination of traditional feed forward neural networks with long short-term memory (LSTM) and convolutional neural networks leads to improved binding predictions.
Fine mapping of a quantitative trait locus

Vivi R. Gregersen1, Lone B. Madsen1, Nina A. Poulsen2, Bo Thomsen1, Frank Panitz1, Mikka S. Hansen3, Albert J. Buitenhuis1, Lars-Erik Holm1, Lotte B. Larsen2 and Christian Bendixen1

1.Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark 2.Department of Food Science, Aarhus University, 8830 Tjele, Denmark 3 Arla Foods amba, Viby J, Denmark

Abstract:
A major locus was identified in relation to cow milk coagulation properties that are very important in cheese production. The locus encompassed the casein cluster located on chromosome 6, and the three most significant SNPs were situated within and upstream CSN3, coding for κ-casein. This protein is very important in relation to cheese production as it is situated on the surface of the casein micelles helping to keep the structure in balance. When chymosin, an enzyme, is added to the milk it cleaves the surface protein and the micelles become unstable and start to aggregate, which is the initial process in formation of cheese.

In order to identify the causal SNP and assess its effect, we conducted a large number of bioinformatics analyses. A workflow starting with a haplotype analysis and whole genome sequencing of selected animals led to a number of candidate SNPs. These were located in the CSN3 3’UTR and promoter region. Functional assays designed for the promoter SNPs were found not to affect the relative expression, but when investigating three haplotypes based on the 3’UTR SNPs, clear differences in relative expression of CSN3 was identified. Only one SNP differed between two of the haplotypes indicating causality. In silico mining did, however, not identify any potential binding targets for miRNA and a target protection assay was designed giving the same result leaving uncertainty about the regulatory mechanism.

The direct effect on milk coagulation was investigated by sampling milk of homozygote animals of the three haplotypes. Milk samples were collected for the rheological analysis and estimation of micelle size. In addition, milk samples were collected for RNA sequencing of mammary epithelia cells. The RNA-seq analysis confirmed the 3’UTR functional analysis results. Besides, overall differential expression analyses did not identify any other differences between the groups of homozygotes. The milk coagulation properties acted as expected with poor milk coagulation properties for the lowly expressed CSN3 haplotype, whereas no difference was seen between the other two haplotypes. The micelle size was found to be negatively correlated to the relative expression of CSN3 and as this could be directly linked to the milk coagulation properties we have not only identified the causal SNP, we are also able to explain the biological impact of this mutation, that high CSN3 expression result in smaller micelles and this results in better micelle aggregation properties.
Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters

Yun Chen¹,², Athma A. Pai³, Jan Herudek⁴, Michal Lubas²,⁴, Nicola Meola⁴, Aino I. Järvelin⁵,#, Robin Andersson¹, Vicent Pelechano⁵,€, Lars M. Steinmetz⁵,⁶,⁷, Torben Heick Jensen⁴*, Albin Sandelin¹,²*

¹The Bioinformatics Centre, Department of Biology, University of Copenhagen, Denmark.
²Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Denmark.
³Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.
⁴Centre for mRNP Biogenesis and Metabolism, Department of Molecular Biology and Genetics, Aarhus University, Denmark.
⁵European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany
⁶Stanford Genome Technology Center, Palo Alto, 94304 CA, USA
⁷Department of Genetics, Stanford University School of Medicine, Stanford, 94305 CA, USA
# Present address: Department of Biochemistry, Oxford University, United Kingdom.
€ Present address: SciLifeLab, Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Sweden.

Corresponding authors: Torben Heick Jensen (thj@mbg.au.dk) and Albin Sandelin (albin@binf.ku.dk)

Abstract:
Mammalian transcriptomes are complex and formed by extensive promoter activity. In addition, gene promoters are largely divergent and initiate transcription of reverse-oriented promoter upstream transcripts (PROMPTs). Although PROMPTs are commonly terminated early, influenced by polyadenylation (pA) sites, promoters often cluster so that the divergent activity of one might impact another. Here, we find that the distance between promoters strongly correlates with the expression, stability and length of their associated PROMPTs. Adjacent promoters driving divergent mRNA transcription support PROMPT formation, but due to pA site constraints, these transcripts tend to spread into the neighboring mRNA on the same strand. This mechanism to derive new alternative mRNA transcription start sites (TSSs) is also evident at closely spaced promoters supporting convergent mRNA transcription. We suggest that basic building blocks of divergently transcribed core promoter pairs, in combination with the wealth of TSSs in mammalian genomes, provides a framework with which evolution shapes transcriptomes.
An automatic end-to-end solution for disease-causing variant detection in rare and hereditary diseases with a high case solve rate and a much reduced false positive rate

A. Joecker¹, S. Shah², M. Hadjisavas², D. Richards², B. Oester¹, G. Eley³, B. Solomon³, J.G. Vockley³
¹ QIAGEN Aarhus, Aarhus, Denmark;
²QIAGEN Silicon Valley, Redwood City, USA;
³Inova Translational Medicine Institute, VA, USA

Abstract:
Background: Identification of causal variants in hereditary diseases can be both challenging and time consuming. Often a lot of time and resources are wasted on identifying a disease causing variant from a TRIO and trying to validate variants, which are actually not disease causing or artefacts. Furthermore, sometimes no causal variant can be identified at all in the patient. QIAGEN Bioinformatics' hereditary disease solution delivers increased sensitivity for identifying causal variants, while reducing significantly the list of candidate variants for follow-up. In addition, it is very easy to run as data analysis and interpretation steps are embedded in a streamlined end-to-end workflow with optimized parameter settings. In this study we show the first results from a benchmarking study with six whole genome trios and one whole exome trio.

Results: Six whole genome were analysed and one whole exome TRIO using QIAGEN's new hereditary disease solution. We show that for diseases caused by de novo variants, the complete workflow results in the identification of the disease causing variant without any additional candidates. This is achieved using default parameters and providing phenotype information to the filter cascade. On dominant inherited diseases we were able to reduce the number of candidates for follow-up by more than 99%. In addition, we are allowing the easy validation of the candidates and an optimization of the filter cascade. As a result less time and resources have to be spent on additional validation steps.
ABSTRACTS
ELIXIR DK booth

bio.tools : ELIXIR Life Science Software Registry

- by Jon Ison on behalf of ELIXIR DK and partners

biotoolsXSD:
description model for bioinformatics software

biotoolsXSD is a formalised XML schema (XSD) that defines a general purpose description model for bioinformatics software. It is used by the ELIXIR Tools & Data Services Registry (https://bio.tools). Bioinformaticians routinely use a large and diverse set of tools and data, and must organise, find, understand and compare – and then use and connect - the available resources. These tasks rely on consistent, and ideally machine-understandable descriptions. The need - filled by biotoolsXSD - is for a model that defines a minimum standard for the consistent description of a broad range of resources.

EDAM:
controlled vocabulary for bioinformatics

EDAM is an ontology of well established, familiar concepts that are prevalent within bioinformatics, including types of data and data identifiers, data formats, operations and topics. EDAM aims to unify semantically the bioinformatics concepts in common use, provide curators with a comprehensive controlled vocabulary that is broadly applicable, and support new and powerful search, browse and query functions. EDAM is a simple ontology - essentially a set of terms with synonyms and definitions - organised into an intuitive hierarchy for convenient use by curators, software developers and end-users.
ELIXIR:

a distributed infrastructure for biological information

ELIXIR’s mission is to construct and operate a sustainable infrastructure for biological information in Europe, supporting life science research and its translation to medicine, agriculture, bioindustries and society. The Danish ELIXIR Node focus is on addressing the growing need of the life science community to access bio-molecular databases, integrate data and use interoperable tools for data analysis and interpretation. ELIXIR Denmark leads the community-driven development of the ELIXIR Tools and Data Services Registry (bio.tools). The Registry provides essential scientific and technical information about analytical tools and data services for bioinformatics. ELIXIR DK also provides supercomputer and secure cloud infrastructure.

ELIXIR EXCELERATE:

tools interoperability & registry

ELIXIR-EXCELERATE is an EU funded project to accelerate the implementation of ELIXIR, develop and connect resources and services, and build bioinformatics capacity across Europe. The life science community has an increasing need to access bio-molecular databases, integrate data and use interoperable tools for data analysis and interpretation. The “tools interoperability & registry” Work Package 1, led by ELIXIR Denmark, focuses on biosoftware discovery and interoperability. The cornerstone is a comprehensive Tools and Data Services Registry (bio.tools), involving service monitoring, resource integration, interoperability aspects, and community centred benchmarking efforts.
bio.tools: tools & data services registry

bio.tools is a registry of bioinformatics software information, sustained by a community-driven curation effort, tailored to local needs and shared amongst a network of engaged partners. Life sciences yield huge data sets that underpin vital scientific discoveries. In support, a plethora of databases and tools are used, in technically complex and diverse forms, across a spectrum of scientific disciplines. The corpus of information for these resources is fragmented across the Web, with much redundancy, and has lacked a common information standard. The outcome is that scientists often struggle to find, understand, compare and use the best resources for the task at hand. bio.tools addresses this problem.

ComputerOME Secure Cloud: scalable secure compute environment

ComputerOME Secure Cloud is a highly specialised offering from the Danish National Supercomputer for Life Sciences. Secure Cloud eliminates problems and barriers for researchers dealing with highly sensitive data distributed across many countries by providing out of the box. Secure Cloud is powered by the infrastructure consisting of 16048 CPU cores with 92 TeraBytes of memory, connected to 3 PetaBytes of High-performance storage, and with a total peak performance of more than 483 TeraFLOPS (483 million million floating-point operations per second).
<table>
<thead>
<tr>
<th>No.</th>
<th>First Name</th>
<th>Surname</th>
<th>Email</th>
<th>Institution</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ajuna</td>
<td>Azad</td>
<td><a href="mailto:ajuna.azad@sund.ku.dk">ajuna.azad@sund.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>University of Copenhagen</td>
</tr>
<tr>
<td>2</td>
<td>Albin</td>
<td>Sandelin</td>
<td><a href="mailto:albin@binf.ku.dk">albin@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Professor</td>
</tr>
<tr>
<td>3</td>
<td>Alejandro</td>
<td>Aguayo Orozco</td>
<td><a href="mailto:alejandro.aguayo@pr.ku.dk">alejandro.aguayo@pr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Phd Student</td>
</tr>
<tr>
<td>4</td>
<td>Alexander</td>
<td>Gronning</td>
<td><a href="mailto:algro11@student.sdu.dk">algro11@student.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Student</td>
</tr>
<tr>
<td>5</td>
<td>Alexander</td>
<td>Hauser</td>
<td><a href="mailto:alexander.hauser@gmail.com">alexander.hauser@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD Fellow</td>
</tr>
<tr>
<td>6</td>
<td>Alexander</td>
<td>Junge</td>
<td><a href="mailto:ajunge@rth.dk">ajunge@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>7</td>
<td>Alfred</td>
<td>Ferrer Florensa</td>
<td><a href="mailto:f.f.alfred@gmail.com">f.f.alfred@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
</tr>
<tr>
<td>8</td>
<td>Ali</td>
<td>Altintas</td>
<td><a href="mailto:ali@rth.dk">ali@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Postdoc</td>
</tr>
<tr>
<td>9</td>
<td>Ali</td>
<td>Syed</td>
<td><a href="mailto:alisyed@csb.dtu.dk">alisyed@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>HPC</td>
</tr>
<tr>
<td>10</td>
<td>Alicja</td>
<td>Paciorek</td>
<td><a href="mailto:paciorek@mbg.au.dk">paciorek@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Associate Professor</td>
</tr>
<tr>
<td>11</td>
<td>Anders</td>
<td>Krog</td>
<td><a href="mailto:krog@binf.ku.dk">krog@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Professor</td>
</tr>
<tr>
<td>12</td>
<td>Anja</td>
<td>PSoule woman</td>
<td><a href="mailto:anja@mbg.au.dk">anja@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Postdoc</td>
</tr>
<tr>
<td>13</td>
<td>Andrea</td>
<td>Brugger</td>
<td><a href="mailto:andrea@mbg.au.dk">andrea@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
</tr>
<tr>
<td>14</td>
<td>Andrew</td>
<td>Campbell</td>
<td><a href="mailto:andrew.campbell@hpe.com">andrew.campbell@hpe.com</a></td>
<td>Hewlett Packard Enterprise</td>
<td>Business Development</td>
</tr>
<tr>
<td>15</td>
<td>Anja</td>
<td>Pen</td>
<td><a href="mailto:anja.pen@mbg.au.dk">anja.pen@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Ph.D. Student</td>
</tr>
<tr>
<td>16</td>
<td>Anna Maria</td>
<td>Malte</td>
<td><a href="mailto:anna.malte@mbg.au.dk">anna.malte@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
</tr>
<tr>
<td>17</td>
<td>Anna Maria</td>
<td>Malte</td>
<td><a href="mailto:anna.malte@mbg.au.dk">anna.malte@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
</tr>
<tr>
<td>18</td>
<td>Anne</td>
<td>Wenzel</td>
<td><a href="mailto:anja@rth.dk">anja@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Phd student</td>
</tr>
<tr>
<td>19</td>
<td>Annelaura</td>
<td>Bach</td>
<td><a href="mailto:annelaura.bach@pr.ku.dk">annelaura.bach@pr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>20</td>
<td>Anne-Mette</td>
<td>Bjerregaard</td>
<td><a href="mailto:amb@mbg.au.dk">amb@mbg.au.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>PhD Student</td>
</tr>
<tr>
<td>21</td>
<td>Anton</td>
<td>Petrov</td>
<td><a href="mailto:petrov@ebi.ac.uk">petrov@ebi.ac.uk</a></td>
<td>EMBL-EBI</td>
<td>Bioinformatician</td>
</tr>
<tr>
<td>22</td>
<td>Arzu Tugce</td>
<td>Guler</td>
<td><a href="mailto:a.t.guler@lumc.nl">a.t.guler@lumc.nl</a></td>
<td>Center for Proteomics and Metabolomics, Leiden</td>
<td>PhD student in bioinformatics</td>
</tr>
<tr>
<td>23</td>
<td>Asli</td>
<td>Özen</td>
<td><a href="mailto:aoz@exiqon.com">aoz@exiqon.com</a></td>
<td>Exiqon</td>
<td>Scientist</td>
</tr>
<tr>
<td>24</td>
<td>Aviaja</td>
<td>Hauptmann</td>
<td><a href="mailto:aviaja@mbg.au.dk">aviaja@mbg.au.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Ph.D. Student</td>
</tr>
<tr>
<td>25</td>
<td>Axel</td>
<td>Thieffry</td>
<td><a href="mailto:axel.thieffry@bio.ku.dk">axel.thieffry@bio.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD fellow</td>
</tr>
<tr>
<td>26</td>
<td>Bernt</td>
<td>Guldbrandt</td>
<td><a href="mailto:bernt.guldbrandt@mbg.au.dk">bernt.guldbrandt@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Professor</td>
</tr>
<tr>
<td>27</td>
<td>Bitao</td>
<td>Qiu</td>
<td><a href="mailto:fkc259@alumni.ku.dk">fkc259@alumni.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>28</td>
<td>Bo</td>
<td>Thomsen</td>
<td><a href="mailto:bo.thomsen@mbg.au.dk">bo.thomsen@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Associate Professor</td>
</tr>
<tr>
<td>29</td>
<td>Bruce</td>
<td>Shapiro</td>
<td><a href="mailto:shapirbr@n.ni.gov">shapirbr@n.ni.gov</a></td>
<td>National Cancer institute/NIH</td>
<td>Senior investigator</td>
</tr>
<tr>
<td>30</td>
<td>Cecilia</td>
<td>Engel Thomas</td>
<td><a href="mailto:cecilia.thomas@pr.ku.dk">cecilia.thomas@pr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD Student</td>
</tr>
<tr>
<td>31</td>
<td>Charles</td>
<td>Hansen</td>
<td><a href="mailto:ccrhansen@gmail.com">ccrhansen@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
</tr>
<tr>
<td>32</td>
<td>Christian</td>
<td>Anthon</td>
<td><a href="mailto:anthon@rth.dk">anthon@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Systems administrator</td>
</tr>
<tr>
<td>33</td>
<td>Christian</td>
<td>Grønbæk</td>
<td><a href="mailto:christian.gronbaek@gmail.com">christian.gronbaek@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>post doc</td>
</tr>
<tr>
<td>34</td>
<td>Christian</td>
<td>Simon</td>
<td><a href="mailto:christian.simon@pr.ku.dk">christian.simon@pr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Phd Student</td>
</tr>
<tr>
<td>35</td>
<td>Christian</td>
<td>Wiwe</td>
<td><a href="mailto:wiwe@imada.sdu.dk">wiwe@imada.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>PhD student</td>
</tr>
<tr>
<td>36</td>
<td>Christina</td>
<td>Rasmussen</td>
<td><a href="mailto:chr113@student.sdu.dk">chr113@student.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>bachelor student</td>
</tr>
<tr>
<td>37</td>
<td>Damilova</td>
<td>Babatunde</td>
<td><a href="mailto:myetreade247@gmail.com">myetreade247@gmail.com</a></td>
<td>OLABISI ONABANJO UNIVERSITY</td>
<td>STUDENT (MASTERS)</td>
</tr>
<tr>
<td>38</td>
<td>Dan</td>
<td>Søndergaard</td>
<td><a href="mailto:dan1990@gmail.com">dan1990@gmail.com</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
</tr>
<tr>
<td>39</td>
<td>David</td>
<td>Lyon</td>
<td><a href="mailto:david.lyon@pr.ku.dk">david.lyon@pr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PostDoc</td>
</tr>
<tr>
<td>40</td>
<td>David</td>
<td>Westergaard</td>
<td><a href="mailto:david.westergaard@pr.ku.dk">david.westergaard@pr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>41</td>
<td>Dennis</td>
<td>Pultz</td>
<td><a href="mailto:dp@mbg.sdu.dk">dp@mbg.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Post Doc</td>
</tr>
<tr>
<td>42</td>
<td>Derya</td>
<td>Aslan</td>
<td><a href="mailto:derya.aslan@gmail.com">derya.aslan@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
</tr>
<tr>
<td>43</td>
<td>Dhanay</td>
<td>Saputra</td>
<td><a href="mailto:dsa@exiqon.com">dsa@exiqon.com</a></td>
<td>Exiqon</td>
<td>Scientist</td>
</tr>
<tr>
<td>44</td>
<td>Diogo</td>
<td>Marinho</td>
<td><a href="mailto:diogoma@mbg.sdu.dk">diogoma@mbg.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>PhD student</td>
</tr>
<tr>
<td>45</td>
<td>Elise</td>
<td>Lucotte</td>
<td><a href="mailto:elucotte@biroc.au.dk">elucotte@biroc.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>postdoctoral fellow</td>
</tr>
<tr>
<td>46</td>
<td>Emil</td>
<td>Rydza</td>
<td><a href="mailto:emil.rydza@pr.ku.dk">emil.rydza@pr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Computer Specialist</td>
</tr>
<tr>
<td>47</td>
<td>Evdoxia</td>
<td>Karadoulama</td>
<td><a href="mailto:evka@mbg.sdu.dk">evka@mbg.sdu.dk</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
</tr>
<tr>
<td>48</td>
<td>Federico</td>
<td>De Masi</td>
<td><a href="mailto:federico.de.masi@gmail.com">federico.de.masi@gmail.com</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Senior Researcher</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Ferhat</td>
<td>Alkan</td>
<td><a href="mailto:ferro@rth.dk">ferro@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD Fellow</td>
</tr>
<tr>
<td>51</td>
<td>Francesco</td>
<td>Russo</td>
<td><a href="mailto:russof85@gmail.com">russof85@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>52</td>
<td>Frank</td>
<td>Panitz</td>
<td><a href="mailto:frank.panitz@mbg.au.dk">frank.panitz@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Associate Professor</td>
</tr>
<tr>
<td>53</td>
<td>Franziska</td>
<td>Klincke</td>
<td><a href="mailto:klincke@csb.dtu.dk">klincke@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>PhD student</td>
</tr>
<tr>
<td>54</td>
<td>Georgios</td>
<td>Athanasiadis</td>
<td><a href="mailto:athanasiadis@bic.au.dk">athanasiadis@bic.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Postdoc</td>
</tr>
<tr>
<td>55</td>
<td>Gustav</td>
<td>Ahlberg</td>
<td><a href="mailto:ahlberg.gustav@gmail.com">ahlberg.gustav@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD Student</td>
</tr>
<tr>
<td>56</td>
<td>Hans</td>
<td>Ienasescu</td>
<td><a href="mailto:ienasescu@bio.ku.dk">ienasescu@bio.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Research Assistant</td>
</tr>
<tr>
<td>57</td>
<td>Henriette</td>
<td>Husum Bak Jørgensen</td>
<td><a href="mailto:hhu@bio.ku.dk">hhu@bio.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Research Manager</td>
</tr>
<tr>
<td>58</td>
<td>Henrik</td>
<td>Nielsen</td>
<td><a href="mailto:hnielsen@csb.dtu.dk">hnielsen@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Associate professor</td>
</tr>
<tr>
<td>59</td>
<td>Hiroyuki</td>
<td>Prada</td>
<td><a href="mailto:xsh723@alumni.ku.dk">xsh723@alumni.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
</tr>
<tr>
<td>60</td>
<td>Isabella</td>
<td>Friis Jørgensen</td>
<td><a href="mailto:isabella.jorgensen@cprr.ku.dk">isabella.jorgensen@cprr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>61</td>
<td>Jakob Hull</td>
<td>Havnagard</td>
<td><a href="mailto:hull@rth.dk">hull@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>RTH</td>
</tr>
<tr>
<td>62</td>
<td>Jamshed Pratap Gill</td>
<td><a href="mailto:Jamshed21@gmail.com">Jamshed21@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>Jan</td>
<td>Baumbach</td>
<td><a href="mailto:jan.baumbach@imada.sdu.dk">jan.baumbach@imada.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Associate Professor</td>
</tr>
<tr>
<td>64</td>
<td>Jan</td>
<td>Gorodkin</td>
<td><a href="mailto:gorodkin@rth.dk">gorodkin@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Professor</td>
</tr>
<tr>
<td>65</td>
<td>Jan Christian</td>
<td>Refsgaard</td>
<td><a href="mailto:jancedefgaard@gmail.com">jancedefgaard@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Bioinformatician</td>
</tr>
<tr>
<td>66</td>
<td>Jesper</td>
<td>Tholstrup</td>
<td><a href="mailto:jte@exiqon.com">jte@exiqon.com</a></td>
<td>Exiqon</td>
<td>Scientist</td>
</tr>
<tr>
<td>67</td>
<td>Jette Bornholdt</td>
<td>Lange</td>
<td><a href="mailto:bornholdt@binf.ku.dk">bornholdt@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Assistant professor</td>
</tr>
<tr>
<td>68</td>
<td>Jim</td>
<td>Samuelsen</td>
<td><a href="mailto:jmsm@novonordisk.com">jmsm@novonordisk.com</a></td>
<td>Novo Nordisk</td>
<td>Senior Scientist</td>
</tr>
<tr>
<td>69</td>
<td>Jinjie</td>
<td>Duan</td>
<td><a href="mailto:jduan@bic.au.dk">jduan@bic.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
</tr>
<tr>
<td>70</td>
<td>Jonas</td>
<td>Berglund</td>
<td><a href="mailto:jonas.berglund@bic.au.dk">jonas.berglund@bic.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>postdoc</td>
</tr>
<tr>
<td>71</td>
<td>Jonas</td>
<td>Meisner</td>
<td><a href="mailto:khv475@alumni.ku.dk">khv475@alumni.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>MSc Student</td>
</tr>
<tr>
<td>72</td>
<td>Jonas Andreas</td>
<td>Sibbesen</td>
<td><a href="mailto:jasi@binf.ku.dk">jasi@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>73</td>
<td>Jose Alejandro</td>
<td>Romero Herrer</td>
<td><a href="mailto:jose.romero@cprr.ku.dk">jose.romero@cprr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Phd student</td>
</tr>
<tr>
<td>74</td>
<td>Jose Mg</td>
<td>Izarzugaza</td>
<td><a href="mailto:txema@csb.dtu.dk">txema@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Associate Professor</td>
</tr>
<tr>
<td>75</td>
<td>Joseph</td>
<td>Blair</td>
<td><a href="mailto:josephblair2000@hotmail.com">josephblair2000@hotmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
</tr>
<tr>
<td>76</td>
<td>Kamilla Kjærgaard</td>
<td>Jensen</td>
<td><a href="mailto:kamjen@csb.dtu.dk">kamjen@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>PhD</td>
</tr>
<tr>
<td>77</td>
<td>Karen</td>
<td>Flyvbjerg</td>
<td><a href="mailto:kaffy10@student.sdu.dk">kaffy10@student.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Master student</td>
</tr>
<tr>
<td>78</td>
<td>Kasper</td>
<td>Munch</td>
<td><a href="mailto:kaspermunch@bic.au.dk">kaspermunch@bic.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Associate professor</td>
</tr>
<tr>
<td>79</td>
<td>Katja</td>
<td>Einer-Jensen</td>
<td><a href="mailto:Katja.einerj@qiagen.com">Katja.einerj@qiagen.com</a></td>
<td>QIAGEN Aarhus A/S</td>
<td>Senior Bioinformatician</td>
</tr>
<tr>
<td>80</td>
<td>Katrine Højholt Iversen</td>
<td><a href="mailto:katrine@csb.dtu.dk">katrine@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Kosai</td>
<td>Al-Nakeeb</td>
<td><a href="mailto:kosai@csb.dtu.dk">kosai@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>PhD Student</td>
</tr>
<tr>
<td>82</td>
<td>Kristian</td>
<td>Hanghoj</td>
<td><a href="mailto:kristianhanghoj@gmail.com">kristianhanghoj@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
</tr>
<tr>
<td>83</td>
<td>Kristoffer</td>
<td>Rigbolt</td>
<td><a href="mailto:kri@gtb.r.dk">kri@gtb.r.dk</a></td>
<td>Gubra</td>
<td>Bioinformatician</td>
</tr>
<tr>
<td>84</td>
<td>Kristoffer Vitting-Seerup</td>
<td><a href="mailto:kristoffer.vittingseerup@bio.dtu.dk">kristoffer.vittingseerup@bio.dtu.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD Student</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>Lars</td>
<td>Rønn Olsen</td>
<td><a href="mailto:lro@csb.dtu.dk">lro@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Assistant Professor</td>
</tr>
<tr>
<td>86</td>
<td>Lasse</td>
<td>Folkestad</td>
<td><a href="mailto:lasfol@csb.dtu.dk">lasfol@csb.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Assistant professor</td>
</tr>
<tr>
<td>87</td>
<td>Lea Merch</td>
<td>Harder</td>
<td><a href="mailto:mnh@bmb.sdu.dk">mnh@bmb.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Post Doc</td>
</tr>
<tr>
<td>88</td>
<td>Leo</td>
<td>Dworkin</td>
<td><a href="mailto:leodworkin@gmail.com">leodworkin@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Research Assistant</td>
</tr>
<tr>
<td>89</td>
<td>Leonor</td>
<td>Rib</td>
<td><a href="mailto:maleonor.rp@gmail.com">maleonor.rp@gmail.com</a></td>
<td>University of Lausanne</td>
<td>PhD</td>
</tr>
<tr>
<td>90</td>
<td>Lisa</td>
<td>Cantwell</td>
<td><a href="mailto:lissacantwell@gmail.com">lissacantwell@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>MSc student</td>
</tr>
<tr>
<td>91</td>
<td>Lise Lolle</td>
<td>Christensen</td>
<td><a href="mailto:loll@bmb.sdu.dk">loll@bmb.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Phd student</td>
</tr>
<tr>
<td>92</td>
<td>Louise</td>
<td>Stride Nielsen</td>
<td><a href="mailto:lsn@forens.au.dk">lsn@forens.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
</tr>
<tr>
<td>93</td>
<td>Lucia</td>
<td>Robles</td>
<td><a href="mailto:luciarobles@bmb.sdu.dk">luciarobles@bmb.sdu.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
</tr>
<tr>
<td>94</td>
<td>Lukasz</td>
<td>Berger</td>
<td><a href="mailto:lukbe@du.dk">lukbe@du.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Scientific Programmer</td>
</tr>
<tr>
<td>95</td>
<td>Maja</td>
<td>Dembic</td>
<td><a href="mailto:majad@bmb.sdu.dk">majad@bmb.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Postdoc</td>
</tr>
<tr>
<td>96</td>
<td>Malte</td>
<td>Thodberg</td>
<td><a href="mailto:malte.thodberg@bio.ku.dk">malte.thodberg@bio.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD Student</td>
</tr>
<tr>
<td>No.</td>
<td>Name</td>
<td>Email</td>
<td>Affiliation</td>
<td>Role</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------</td>
<td>-------------------------</td>
<td>--------------------------------------------------</td>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>Man-Hung Tang</td>
<td><a href="mailto:met@exiqon.com">met@exiqon.com</a></td>
<td>Exiqon</td>
<td>Scientist</td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>Maria Dalby</td>
<td><a href="mailto:maria@binf.ku.dk">maria@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>Maria De La O Leyva Pérez</td>
<td><a href="mailto:mleyva@mbg.au.dk">mleyva@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>postdoc</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>Maria Izabel Alves</td>
<td><a href="mailto:izabelcavassim@gmail.com">izabelcavassim@gmail.com</a></td>
<td>AU - Aarhus University</td>
<td>Master Student</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Marie Lisandra</td>
<td><a href="mailto:lisandracady@gmail.com">lisandracady@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>Martin Frith</td>
<td><a href="mailto:mcfrrith@edu.k.u-tokyo.ac.jp">mcfrrith@edu.k.u-tokyo.ac.jp</a></td>
<td>AIST &amp; University of Tokyo</td>
<td>Professor</td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>Martin Simonsen</td>
<td><a href="mailto:martinsimonsen@qiagen.com">martinsimonsen@qiagen.com</a></td>
<td>QIAGEN Aarhus A/S</td>
<td>Director Bioinformatics</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>Martin Closter</td>
<td><a href="mailto:martincj@cbs.dtu.dk">martincj@cbs.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>PhD</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>Michael Ben Ezra</td>
<td><a href="mailto:mbeneza@gmail.com">mbeneza@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Master Student</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>Michal Lubas</td>
<td><a href="mailto:michal.lubas@bric.ku.dk">michal.lubas@bric.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Postdoctoral researcher</td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>Mohamed Mounir</td>
<td><a href="mailto:mounir.mohamed92@gmail.com">mounir.mohamed92@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>Monica Hannani</td>
<td><a href="mailto:monicahannani@gmail.com">monicahannani@gmail.com</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Student</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>Morana Vitezic</td>
<td><a href="mailto:mvitezic@binf.ku.dk">mvitezic@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>postdoc</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>Morten Lindow</td>
<td><a href="mailto:morten.lindow@roche.com">morten.lindow@roche.com</a></td>
<td>Roche Innovation Center Copenhagen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>Morten Nielsen</td>
<td><a href="mailto:mniel@ctu.dtu.dk">mniel@ctu.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Professor</td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>Morten Olesen</td>
<td><a href="mailto:morten.salling.olesen@gmail.com">morten.salling.olesen@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Associate Professor, PhD, Head of Laboratory</td>
<td></td>
</tr>
<tr>
<td>113</td>
<td>Morten Dybdahl Krebs</td>
<td><a href="mailto:morten.dybdahl.krebs@regionh.dk">morten.dybdahl.krebs@regionh.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Research Year Student</td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>Myhanh Nguyen</td>
<td><a href="mailto:myng@dtu.dk">myng@dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Project coordinator</td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>Nadezhda Doncheva</td>
<td><a href="mailto:nadezhda.doncheva@cpu.ku.dk">nadezhda.doncheva@cpu.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PostDoc</td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>Nazli Deniz Yalniz</td>
<td><a href="mailto:nazliyalniz@hotmail.com">nazliyalniz@hotmail.com</a></td>
<td>Istanbul University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>117</td>
<td>Nick Clark</td>
<td><a href="mailto:nicholas.anders.clark@gmail.com">nicholas.anders.clark@gmail.com</a></td>
<td>Infosys</td>
<td>Data Scientist</td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>Nicolas Alcaraz</td>
<td><a href="mailto:nialcaraz@imada.sdu.dk">nialcaraz@imada.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>Niels Tolstrup</td>
<td><a href="mailto:mt@exiqon.com">mt@exiqon.com</a></td>
<td>Exiqon</td>
<td>Manager</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>Nikos Sidiropoulos</td>
<td><a href="mailto:nikos@finsenlab.dk">nikos@finsenlab.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>Palle Villesen</td>
<td><a href="mailto:palle@bric.au.dk">palle@bric.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Associate professor, BIRC</td>
<td></td>
</tr>
<tr>
<td>122</td>
<td>Palle Duun Rohde</td>
<td><a href="mailto:palle.d.rohde@mbg.au.dk">palle.d.rohde@mbg.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>Paolo Marcatilli</td>
<td><a href="mailto:pmar@ctu.dtu.dk">pmar@ctu.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Assistant Professor</td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>Paula Tataru</td>
<td><a href="mailto:paula@ctu.dtu.dk">paula@ctu.dtu.dk</a></td>
<td>AU - Aarhus University</td>
<td>Postdoc</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>Peter Longgreen</td>
<td><a href="mailto:peterl@ctu.dtu.dk">peterl@ctu.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>COO</td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>Pingping Jiang</td>
<td><a href="mailto:pipi@sund.ku.dk">pipi@sund.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Academic Staff</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>Po-Shun Chuang</td>
<td><a href="mailto:ps.chuang@oist.jp">ps.chuang@oist.jp</a></td>
<td>Okinawa Institute of Science and technology</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>Poul Liboriussen</td>
<td><a href="mailto:Poul.Liboriussen@qiagen.com">Poul.Liboriussen@qiagen.com</a></td>
<td>QIAGEN Aarhus A/S</td>
<td>Software Developer</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>Praharshit Sharma</td>
<td><a href="mailto:praharshit@iscb.org">praharshit@iscb.org</a></td>
<td>SGGW/ Warsaw University of Life Sciences</td>
<td>EU-Erasmus Mundus exchange scholar</td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>Rebecca Kirsch</td>
<td><a href="mailto:rebecca@th.dk">rebecca@th.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>131</td>
<td>Richard Röttger</td>
<td><a href="mailto:roettger@imada.sdu.dk">roettger@imada.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Assistant Professor</td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>Richard Sprenger</td>
<td><a href="mailto:richards@bmb.sdu.dk">richards@bmb.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td></td>
<td></td>
</tr>
<tr>
<td>133</td>
<td>Rikke Linemann Nielsen</td>
<td><a href="mailto:rni@ctu.dtu.dk">rni@ctu.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Research assistant</td>
<td></td>
</tr>
<tr>
<td>134</td>
<td>Robin Andersson</td>
<td><a href="mailto:robin@binf.ku.dk">robin@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Assistant professor</td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>Rudolfs Berzins</td>
<td><a href="mailto:rudolfs.berzins.bio@gmail.com">rudolfs.berzins.bio@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Masters Student</td>
<td></td>
</tr>
<tr>
<td>136</td>
<td>Samuele Soraggi</td>
<td><a href="mailto:samuele@math.ku.dk">samuele@math.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>ph.d. student</td>
<td></td>
</tr>
<tr>
<td>137</td>
<td>Sangzi Chen</td>
<td><a href="mailto:sangzic@umich.edu">sangzic@umich.edu</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
<td></td>
</tr>
<tr>
<td>138</td>
<td>Sarah Rennie</td>
<td><a href="mailto:sarah@binf.ku.dk">sarah@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Postdoc</td>
<td></td>
</tr>
<tr>
<td>139</td>
<td>Signe Tang Karlsen</td>
<td><a href="mailto:jw732@alumni.ku.dk">jw732@alumni.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>Simon Rasmussen</td>
<td><a href="mailto:simon@ctu.dtu.dk">simon@ctu.dtu.dk</a></td>
<td>DTU - Technical University of Copenhagen</td>
<td>Associate Professor</td>
<td></td>
</tr>
<tr>
<td>141</td>
<td>Simon Horksøe Rasmussen</td>
<td><a href="mailto:simras@gmail.com">simras@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD Student</td>
<td></td>
</tr>
<tr>
<td>142</td>
<td>Simon Tilgreen Mortensen</td>
<td><a href="mailto:simor11@student.sdu.dk">simor11@student.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Student</td>
<td></td>
</tr>
<tr>
<td>143</td>
<td>Siyang Liu</td>
<td><a href="mailto:siyang.liu@bio.ku.dk">siyang.liu@bio.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>Sofie Jakobsen</td>
<td><a href="mailto:sofie.jakobsen@qiagen.com">sofie.jakobsen@qiagen.com</a></td>
<td>QIAGEN Aarhus A/S</td>
<td>Bioinformatics Scientist</td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>Name</td>
<td>Email</td>
<td>Institution</td>
<td>Position</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>------------------</td>
<td>------------------------------</td>
<td>--------------------------------------</td>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>145</td>
<td>Sonia Maria</td>
<td><a href="mailto:piedrahita@yahoo.com">piedrahita@yahoo.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>MSc in Bioinformatics Student</td>
<td></td>
</tr>
<tr>
<td>146</td>
<td>Stefan Seemann</td>
<td><a href="mailto:seemann@rth.dk">seemann@rth.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Assistant Professor</td>
<td></td>
</tr>
<tr>
<td>147</td>
<td>Steffen Petersen</td>
<td><a href="mailto:steffen1357@gmail.com">steffen1357@gmail.com</a></td>
<td>AAU - Aalborg University</td>
<td>Professor</td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>Søren Besenbacher</td>
<td><a href="mailto:besenbacher@clin.au.dk">besenbacher@clin.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Lektor</td>
<td></td>
</tr>
<tr>
<td>149</td>
<td>Søren Brunak</td>
<td><a href="mailto:soren.brunak@cpr.ku.dk">soren.brunak@cpr.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Professor of Disease Systems Biology</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>Tammi Vesth</td>
<td><a href="mailto:tCVE@bio.dtu.dk">tCVE@bio.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Post Doc</td>
<td></td>
</tr>
<tr>
<td>151</td>
<td>Thanh Thien Nguyen</td>
<td><a href="mailto:s123343@student.dtu.dk">s123343@student.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Master student</td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>Thomas Doktor</td>
<td><a href="mailto:thomaskd@bmb.sdu.dk">thomaskd@bmb.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Academic staff/Postdoc</td>
<td></td>
</tr>
<tr>
<td>153</td>
<td>Thomas Sicheritz-Ponten</td>
<td><a href="mailto:thomas@CBS.dtu.dk">thomas@CBS.dtu.dk</a></td>
<td>DTU - Technical University of Denmark</td>
<td>Professor</td>
<td></td>
</tr>
<tr>
<td>154</td>
<td>Thager Krogh</td>
<td><a href="mailto:tkrogh@sund.sdu.dk">tkrogh@sund.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Ph.D. student</td>
<td></td>
</tr>
<tr>
<td>155</td>
<td>Troels Siggaard</td>
<td><a href="mailto:troels.siggaard@CPR.ku.dk">troels.siggaard@CPR.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Scientific Programmer</td>
<td></td>
</tr>
<tr>
<td>156</td>
<td>Troels Stausbo</td>
<td><a href="mailto:tstaus12@student.aau.dk">tstaus12@student.aau.dk</a></td>
<td>AAU - Aalborg University</td>
<td>Master Student</td>
<td></td>
</tr>
<tr>
<td>157</td>
<td>Tsonko Tsonkov</td>
<td><a href="mailto:tsonko.tsonkov@sund.ku.dk">tsonko.tsonkov@sund.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD-student</td>
<td></td>
</tr>
<tr>
<td>158</td>
<td>Ukpor Friday Akogor</td>
<td><a href="mailto:engrfrifun@gmail.com">engrfrifun@gmail.com</a></td>
<td>BLUE VISTA SERVICES LIMITED</td>
<td>ICT EXPERT</td>
<td></td>
</tr>
<tr>
<td>159</td>
<td>Vanessa Jurtz</td>
<td><a href="mailto:vanessa.jurtz@gmail.com">vanessa.jurtz@gmail.com</a></td>
<td>DTU - Technical University of Denmark</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>Veit Schwammle</td>
<td><a href="mailto:veits@bmb.sdu.dk">veits@bmb.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>Assistant professor</td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>Vivi Raundahl</td>
<td><a href="mailto:vivi.gregersen@MBQ.au.dk">vivi.gregersen@MBQ.au.dk</a></td>
<td>AU - Aarhus University</td>
<td>Aarhus University</td>
<td></td>
</tr>
<tr>
<td>162</td>
<td>Vladimir Gorshkov</td>
<td><a href="mailto:vgor@bmb.sdu.dk">vgor@bmb.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>PostDoc</td>
<td></td>
</tr>
<tr>
<td>163</td>
<td>Wojciech Michalak</td>
<td><a href="mailto:wojciechm@BMB.sdu.dk">wojciechm@BMB.sdu.dk</a></td>
<td>SDU - University of Southern Denmark</td>
<td>MSc in Computational Biomedicine student</td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>Xiaolong Yuan</td>
<td><a href="mailto:xiaolong.yuan@sund.ku.dk">xiaolong.yuan@sund.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Section for Comparative Pediatrics and Nutrition, University of Copenhagen</td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>Xiaoyu Pan</td>
<td><a href="mailto:xiaoyu.pan@sund.ku.dk">xiaoyu.pan@sund.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>PhD student</td>
<td></td>
</tr>
<tr>
<td>166</td>
<td>Yun Chen</td>
<td><a href="mailto:yuc@binf.ku.dk">yuc@binf.ku.dk</a></td>
<td>UCPH - University of Copenhagen</td>
<td>post doc</td>
<td></td>
</tr>
<tr>
<td>167</td>
<td>Zarathustra Goertzel</td>
<td><a href="mailto:zariuq@gmail.com">zariuq@gmail.com</a></td>
<td>UCPH - University of Copenhagen</td>
<td>Student</td>
<td></td>
</tr>
<tr>
<td>168</td>
<td>Zsuzsanna Etches</td>
<td><a href="mailto:zsuzsanna.Etches@qiagen.com">zsuzsanna.Etches@qiagen.com</a></td>
<td>QIAGEN Aarhus A/S</td>
<td>Senior Bioinformatics Scientist</td>
<td></td>
</tr>
</tbody>
</table>
Programme Committee

- Robin **Andersson**, UCPH, DK
- Søren **Basenbacher**, AU, DK
- Kirstine G. **Belling**, UCPH, DK
- Vivi **Gregersen**, AU, DK
- Ida **Moltke**, UCPH, DK
- Kasper **Munch**, AU, DK
- Bent **Petersen**, DTU, DK
- Veit **Schwämmle**, SDU, DK
- Stefan **Seemann**, UCPH, DK

Management Committee

- Peter **Løngreen**, DTU, DK
- Myhanh **Nguyen**, DTU, DK
- Maria Maddalena **Sperotto**, DTU, DK

Student Trainee

- Ricarda **Lohmann**, DTU, DK
PRACTICAL INFORMATION

Contact
Myhanh Nguyen
Cellular phone: +45 - 9351 1288
Email: myng@dtu.dk

Emergency
Ambulance number: 112

Taxi
Danataxa: +45 70252525

Dinner
Musikhuset Posten
Østre Stationsvej 35
DK-5000 Odense C

Hotels

- **HOTEL ANSGAR**
  Østre Stationsvej 32
  DK-5000 Odense C
  Phone +45 66119693
  Fax +45 66119675
  ansgar@millinghotels.dk

- **CABINN**
  Østre Stationsvej 9
  DK-5000 Odense C
  Phone +45 63 14 57 00
  Fax. +45 63 14 57 01
  odense@cabinn.com

- **CITY HOTEL ODENSE**
  Hans Mules Gade 5
  DK-5000 Odense C
  Phone +45 66 12 12 58
  Fax +45 66 12 93 64
  mail@cityhotelodense.eu

- **RADISSON BLUE HC**
  Claus Bergs Gade 7
  DK-5000 Odense C
  Phone +45 66 14 78 00
  Fax +45 66 14 78 90
  hcandersen@radissonblu.com

- **HOTEL WINDSOR**
  Vindegade 45
  DK-5000 Odense C
  Phone +45 66120652
  Fax +45 65910023
  windsor@millinghotels.dk
Venue
DOK5000 Conference Centre,
Havnegade 18, 5000 Odense

Route map
Conference center 🍇
Train Station 🟢