Meeting report of the Winter School organised within the ESF Research Networking Programme ConGenOmics:

“Bioinformatics of Adaptation Genomics: Adaptation Genomics in the Realm of Next Generation Sequencing”

Application Reference Nr.: 5915

1. Summary

Date: March 1st – 7th 2015

Duration: 7 days

Venue: Alexander & Gerbi Hotel, Weggis - Switzerland

European Science Foundation (ESF) Budget: 30’000 Euro

Adaptation to a Changing Environment (ACE) Budget: 12’000 CHF (~ 11’500 Euro)

Participants: 33 participants

Instructors: 5 instructors

Host: Prof. Dr. Alex Widmer, ETH Zürich - Switzerland

Organising Committee: Dr. Simone Fior, ETH Zürich - Switzerland
Dr. Martin C. Fischer, ETH Zürich - Switzerland
Alexandra Jansen van Rensburg, University of Zürich - Switzerland
Dr. Stefan Zoller, ETH Zürich - Switzerland

The rapid technological advance combined with ever decreasing costs has made genome-scale sequencing projects feasible for most evolutionary biology and conservation genetics labs. This provides exciting opportunities to study adaptation, particularly in non-model organisms, where an unprecedented amount of genetic data is available for analysing genotype-phenotype-environment associations. There has been rapid development over recent years in bioinformatic tools required for processing and analysing next-generation sequencing (NGS) data. However, the fast growing number of available analytical tools, and their technical complexity, makes it difficult for researchers to understand the tools available, and decide which are the best for their specific research questions and data. The objectives of this Winter School were to provide researchers with the opportunity to understand the technical rationale underlying the available methods, and to acquire knowledge regarding best practices in generating and analysing NGS data. Within this remit, five experienced researchers each taught for one day on their field of expertise. Each day included lectures and computer practicals, with opportunity for discussion throughout the day, as well as in a social context in the evenings.

The winter school took place in March 2015 in the scenic town of Weggis, Switzerland. The workshop addressed primarily evolutionary and conservation biologists with a need to gain a deeper understanding of the rationale and statistics underpinning commonly used analytical packages used to infer adaptation from genomic data. To facilitate discussion and active participation, preference was given to applicants with experience in genomic data analysis, particularly in the Unix environment. Prospective participants submitted a one page statement of interest and a brief CV via email. From the 126 applications received, the organising committee selected 20 participants representing 14 countries, with 45% female participants. An additional 6 participants joined the workshop through their affiliation with the ACE Initiative (ETH Zurich), which co-funded the school.

2. Description of Scientific Content and Discussions

The rapid development of NGS technologies and their application to adaptation genomics research holds great promise to increase our understanding of genotype-phenotype-environment interactions. The pioneering findings of recent years have inspired the launch of genome-scale projects in an ever increasing number of organisms. A great effort has been made to develop software that can handle the large datasets that typically characterise ecological genomics studies, while still providing robust analytical platforms for supporting their application in non-model organisms. However, due to the sheer number and complexity of available packages, it is often difficult for investigators to assess the potential and limitations of alternative methods, and determine which are best suited to their dataset and question. This Winter School aimed at providing participants an opportunity to gain insight into the rationale behind some of the established analytical pipelines in adaptation genomics, and acquire knowledge on the best practices in experimental design, data generation and analyses.

The first two days of the Winter School focussed on methods for extracting SNP data from various forms of genomic data (Part 1), followed by three days focussing on analytical methods for inferring evolutionary signal from these data (Part 2). To achieve the workshops objective of understanding the technical underpinnings of the various analytical methods, the instructors were chosen as experts in their field, but were young scientists and software developers with current working knowledge of the methods and pipelines in question. The participants were exposed to state of the art methods and examples of their application to real data during computer-based training sessions.

The workshop was aimed particularly at researchers with some experience in bioinformatics and NGS data analysis who wanted to improve their understanding of the technical underpinnings of the available methods. From the large applicant pool of 126, 20 participants were chosen. Priority was given to applicants from the contributing member organisations of the ESF ConGenOmics programme. To facilitate discussion and interaction in the workshop, participants with diverse backgrounds in evolutionary and conservation biology, as well as with specialized knowledge in different study systems, were chosen. The participants represented 14 countries, and a balanced gender ratio (47% female). The workshop was co-funded by the ACE initiative of ETH Zurich. As such, six additional participants joined the workshop through their ACE affiliation. Together with the five organisers and two senior scientists from the Congenomics network, the participants were in total 33.

Computational support was provided through the Genetic Diversity Centre (GDC), ETH Zürich, which is associated with the ACE initiative. To support the computationally intensive requirements of the workshop, the necessary bioinformatic programs and exercises were hosted on the GDC server. All participants and instructors were granted access to the server through a fast wireless network (250 Mbit/sec) that was specifically installed for this purpose at the conference centre. Based on positive feedback, particularly from the instructors, this arrangement contributed significantly to the success of the exercise sessions.

Part 1: Extracting SNP data from NGS reads

There has been much discussion in the literature about biases and considerations intrinsic to NGS data, from experimental design, through different wet lab library preparation techniques, to considerations when filtering reads and calling SNPs. Understanding the details of these is crucial for inferring which biases may be intrinsic to different data sets, and ultimately for correctly interpreting downstream analyses of these data

On the first day, Dr. Jonathan Puritz gave an overview of the different restriction associated DNA (RAD) methods currently available. He discussed how the fundamental differences between these should be considered during experimental design. This was followed by an in depth discussion of a pipeline for producing SNPs from RADseq data. The focus was particularly on the dDocent pipeline developed by Dr. Puritz for analysing ddRAD data, with emphasis on how this compares to alternative tools. Exercises in the afternoon used ddocent.FB (which implements FreeBayes) to call SNPs from an example dataset, as well as comparative analyses using pyRAD.

This brought us to the second day, where Mr. Erik Garrison, the developer of FreeBayes, gave an overview of the theory of SNP and indel calling from NGS data. There was a particular focus on the differences between algorithms which call SNPs only, and those that are able to robustly identify insertions and deletions. Haplotype and variant detection methods were introduced, and the way these are implemented in FreeBayes was discussed. During the afternoon exercises and discussion, Mr Garrison walked us through the best practices to call SNPs from a sample dataset and he thoroughly explained the numerous options of FreeBayes.

Discussions during both these sessions revolved around best practices in experimental design, and the trade-offs in filtering options and variant detection methods for obtaining the optimal final dataset. The fact that both the developer of dDocent, which implements FreeBayes to call SNPs, and that of Freebayes itself were jointly present allowed a very fruitful interaction in the class to further delve into the algorithms of both programs. This ideal combination was greatly appreciated by the audience.

Part 2: Inference of evolutionary signal from SNP data

Genome-wide markers provide us with information about an organism at a previously unprecedented depth. This is particularly pertinent in the case of non-model organisms, where it is now possible to generate a large dataset of markers, across divergent populations or taxa, to approach a comprehensive survey of the variation in a genome. However, it is important to carefully consider which analyses are best suited to the particular dataset and question. The second part of the workshop focussed on three topics:

1. Detection of signatures of selection
2. Inferring complex demographic histories from genomic data
3. Finding environmental and phenotypic associations with genotypes

Dr. Sam Yeaman gave an overview of population genetics theory of the migration-drift-selection balance from single locus models. This theory can be expanded to multiple loci and genome scale data. Together with considerations of linkage between loci, this theory can be used to make predictions of the expected genomic architecture of adaptation. The theory was followed by a discussion of how to use models to make predictions about the expected genomic signature under different biological scenarios. The afternoon session focussed on specific analyses that can be used to detect signals of selection, with a discussion of the limitations of the different methods. The exercise in the afternoon walked through the process of calculating genome-wide F_{ST} and exploring the patterns in the data.

In the search for signals of adaptation, it is important to consider demographic history of the studied population, as this may confound the findings of the outlier or association analyses. Prof. Dr. Daniel Wegmann gave a comprehensive introduction to the coalescent, and Bayesian methods for modelling demographic history. An overview was given of how to infer complex demographic histories from frequency spectra using composite likelihood approaches. This was followed by an introduction to Approximate Bayesian Computation (ABC), with an example of its use in estimating complex demographic history using a case study. During the afternoon practical session, participants were guided through the process of using ABC, followed by the use of

FastSimCoal. There was a lot of discussion during the exercise session about interpretation of the outputs from these programs.

This brought us to the last day, which hosted a teaching session jointly prepared by Prof. Dr. Andrew Eckert and Dr. Chris Friedline, and practically given by the latter. This focussed on associations between genotypes, phenotypes, and the environment. An overview was given of the genetic basis of local adaptation, touching on the different methods that can be used to detect local adaptation. The rest of the session was spent explaining methods in association mapping, using experimental examples. During the afternoon exercise a detailed Python pipeline implementing association analyses using the program Bayenv was provided for detecting local adaptation. Participants were guided in a step-wise manner through a sophisticated platform developed by Dr. Friedline to allow parallel execution of the commands and discussion of the graphical outputs.

In all these sessions, emphasis was placed on interpretation of the results. The questions and discussion throughout the sessions ranged from better understanding the theoretical predictions, to in depth interpretation of the results from the different analytical methods. The sessions were highly interactive, with the speakers frequently pausing to provide explanations whenever questions arose.

The exercises and course material have all been made available to the participants for future use (http://www.adaptation.ethz.ch/education/winter-school-2015/teaching-resources.html). Due to the technical detail of the course, much of what was learnt will be best consolidated while working through the exercises with own data.

3. Assessment of the results and future impact

The Winter School provided an in depth discussion of the bioinformatics needed to process and analyse NGS data to infer candidate adaptive loci using a variety of state-of-the-art analytical pipelines. This is pertinent for a wide range of evolutionary biology studies, including conservation genomics and the timely debate on how to detect adaptive variation in practical cases of conservation concern.

Although many introductory bioinformatics courses are available, there is a need for advanced understanding of the algorithms behind the programs, in order to foster cautious usage and better interpretation of the outputs. This workshop provided a unique opportunity for participants to learn more about the underpinnings of these analyses from experienced instructors with hands-on bioinformatics experience. The pros and cons in these analytical methods were discussed and best practices were shown on empirical datasets, with much of the data and methods as yet unpublished. As judged from the positive feedback obtained from both the participants and the instructors, everyone benefitted from the interactions and the instructions during this week. Apart from the learning opportunity, the workshop provided an excellent opportunity for establishing personal contacts and potential future collaborations, particularly between the European groups.

The school was very focussed on the rationale behind the methods used during analysis of NGS data. Much of the discussion and questions from students dealt with understanding the assumptions of these methods, and how this relates to their data, as well as interpretation of the outputs. Since many of the steps made during analysis involve subjective decisions about the optimal analytical parameters, it is imperative that practitioners understand the trade-offs involved. Overall the school addressed many of the challenges one faces during NGS data analysis, providing a comprehensive outline of the tools currently available as well as yet at the developmental stage.

Based on positive feedback from participants we believe that the workshop addressed many of the challenges faced by them. The instructors also expressed a very positive experience at the workshop, in terms of organisation, the participants' interest and eagerness to learn, and the subjects covered during the week. They have all expressed their willingness to participate in similar future events.

4. Annexes

A: Programme of the meeting

B: Full list of all the instructors and participants

C: Photograph of the school participants at Weggis

Annex 4A: Programme of the meeting

Sunday March 1st

Arrival and welcome

10:00 - 18:00 Arrival and registration of participants – Alexander & Gerbi Hotel

19:00 Welcome and introductory talk

20:00 Opening dinner

EXTRACTING SNP DATA FROM NGS READS

Monday March 2nd

1: Rad SNPs from RADseq - The bioinformatics of reduced representation libraries

Dr. Jonathan Puritz (Harte Research Institute, Texas A&M Corpus Christi, USA)

Session 1 - An overview of the many different RAD techniques, RAD experimental design, and minimizing bias in the laboratory

Session 2 - Pipelines, software packages, and scripting - how to generate SNPs from RADseq data

Session 3 - Skimming SNPs from the top: techniques for filtering RADseq SNPs

Tuesday March 3rd

2: Whole Genome and Exome Datasets

Mr. Erik Garrison (Wellcome Trust Sanger Institute & University of Cambridge, UK)

Session 1 - Overview and theory of SNP and indel calling from NGS data; approaches, pitfalls and best practise

Session 2 - Case studies in variant calling: germlines, populations, somatic variation, pools, polyploids, haplotypes, and structural variants

Session 3 - Hands-on real data

INFERENCE OF EVOLUTIONARY SIGNAL FROM SNP DATA

Wednesday March 4th

3: Detection of signatures of selection

Dr. Sam Yeaman (University of British Columbia, Canada)

Session 1 - Theoretical predictions about the genomics of local adaptation

Session 2 - Empirical approaches to detecting signatures of local adaptation in genomic data: from single-locus statistics to genome-level assessment

Session 3 - Case studies

Thursday March 5th

4: Inferring complex demographies from Genomic data

Prof. Dr. Daniel Wegmann (University of Fribourg, Switzerland)

Session 1 - Estimating diversity from NGS data: the problem of filtering and some solutions.

Session 2 - Inferring complex demographies from Frequency spectra using composite likelihood approaches.

Session 3 - Inferring complex demographies using Approximate Bayesian Computation.

Friday March 6th

5: Environmental and Phenotypic Associations

Prof. Dr. Andrew Eckert & Dr. Chris Friedline, presented by Chris Friedline (Virginia Commonwealth University, USA)

Session 1 – Evolutionary genetics of associations among genotypes, phenotypes, and environments

Session 2 – Empirical methods with which to detect environmental and phenotypic associations: single and multiple locus methodologies

Session 3 – Analysis of single nucleotide polymorphism datasets for the detection of associations in non-model species20:00 Workshop

Closing dinner (Swiss speciality)

Saturday March 7th

Departure

Annex 4B: Full list of all instructors and participants

Name	Surname	Institute	Country
Instructors			
Jonathan	Puritz	Harte Research Institute	USA
Erik	Garrison	Wellcome Trust Sanger Institute	UK
Sam	Yeaman	University of British Columbia	Canada
Daniel	Wegmann	University of Fribourg	Switzerland
Chris	Friedline	Virginia Commonwealth University	USA
Congenomics Paricipants			
Tim	Bray	Palacky University	Czech Republic
Anurag	Chaturvedi	KU Leuven	Belgium
Chris	Clarkson	Liverpool School of Tropical Medicine	UK
Kalina	Davies	Queen Mary University of London	UK
Andrew	Foote	Uppsala University	Sweden
Inês	Fragata	University of Lisbon	Portugal
Paolo	Franchini	University of Konstanz	Germany
Peter	Frandsen	University of Copenhagen	Denmark
Tim	Janicke	Centre d'Ecologie Fonctionnelle et Evolutive	France
Deborah	Leigh	University of Zurich	Switzerland
Gareth	Linsmith	University of Ghent/Fondazione Edmund Mach	Belgium/Italy
María Lucena	Pérez	Doñana Biological Station (EBD-CSIC)	Spain

Michael	Matschiner	University of Oslo	Norway
Tiina	Mattila	University of Oulu	Finland
Eduardo	Moreno González	Max Planck Institute for Developmental Biology	Germany
Nina	Overgaard Therkildsen	Stanford University	USA
Aritz	Ruiz-Gonzalez	University of the Basque Country	Spain
Anna	Tigano	Queen's University	Canada
Vera	Warmuth	Uppsala University	Sweden
Kristine	Westergaard	Norwegian Institute for Nature Research	Norway
Joop	Ouborg	Radboud University	The Netherlands
Philippine	Vergeer	Wageningen University	The Netherlands

ACE Paricipants

Jessica	Stephenson	ETH Zurich	Switzerland
Ursina	Messmer	ETH Zurich	Switzerland
Robert	Duenner	ETH Zurich	Switzerland
Mortiz	Luerig	ETH Zurich	Switzerland
Simon	Crameri	ETH Zurich	Switzerland
James	Buckley	ETH Zurich	Switzerland

Organisers

Alex	Widmer	ETH Zurich	Switzerland
Simone	Fior	ETH Zurich	Switzerland

Martin	Fischer	ETH Zurich	Switzerland
Alexandra	JvRensburg	University of Zurich	Switzerland
Stefan	Zoller	ETH Zurich	Switzerland

ANNEX 4C: Photograph of the school participants at Weggis