Centre for computational Inference in Evolutionary Life Science – CELS 2

Research plan

High-throughput sequencing (HTS) technologies have opened for raising a new range of questions within genomic evolution. The combination of unprecedented data generation and basic analysis has been sufficient to make many new discoveries. However, much of the current analysis is limited to establishment and description of the presence of genomic variation across species and individuals. Further insights could be gained by aiming to mechanistically understand consequences and underlying causes of genomic variation, both within and across species, and at different genomic scales. While simple counting and visual browsing of genomic variants have often been sufficient for the present generation of genomic evolution studies, further insights will require consideration of more complex structural variation, larger numbers of species and individuals, as well as more complex forms of interaction.

There is currently good availability of software tools for detecting nucleotide polymorphisms and some forms of structural variants from HTS datasets derived from a given genome. However, there is little available methodology for studying the interplay between small-scale and large-scale variation, or for integrative analysis of variation across species, and/or across populations within these species. Such integrative analysis could open for a new range of evolutionary questions to be approached. The current approach to these questions involves using a linear reference genome for each species. This approach fails to provide a consistent analysis framework for the more complex situations addressed in this proposal, as the reliance on a single linear reference limits the evolutionary distance that can be incorporated into such analyses. Also, by simply annotating structural variants as locations on a linear reference, one restricts the possibilities when comparatively analyzing the structural variants, and also restricts integrative analysis using fine-scale variation. The use of graph-based representations of sequences opens up for representing large- and small-scale variations across different species in a common reference structure, enabling standardized and consistent statistical inference on a scale that is not possible through the use of linear reference genomes.

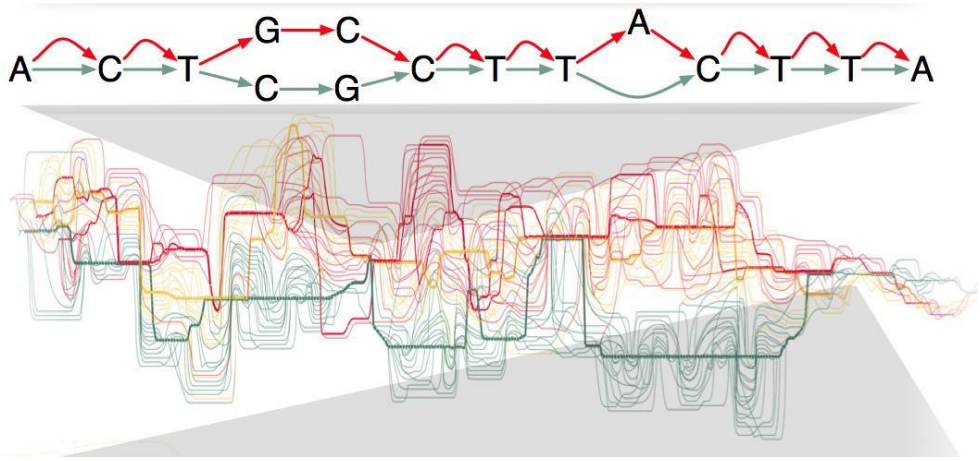

Figure 1. Instead of representing a genome as a line, newer developments may represent both the variation in a genome as well as the variation in individuals representing different populations, adaptations, phenotypes - or possibly different species. The lower part shows the variation in the human MHC. (Credit: UC Santa Cruz Genomics Institute)

The first phase of CELS allowed us to establish internationally competitive expertise in graph-based genome representations, including the development of the first ChIP-Seq peak caller for graph-based reference genomes (7). The outcome of this phase was the result of a tight integration of expertise and

research questions from the different disciplines. In a next phase, new ground can be broken through an even tighter coupling to the strong biological and evolutionary expertise at CEES, by further developing computational approaches for answering previously unapproachable questions in ongoing projects related to Atlantic cod and codfishes. At CEES, a long-standing interest in the genomics of Atlantic cod has been pursued - from both a population and a comparative genomics perspective. Published by a CEES-led team in 2011, the Atlantic cod was the first bony (teleost) fish to be genome sequenced using a pure HTS approach (12). In 2016, the team published an improved Atlantic cod genome assembly, using PacBio long-read sequencing technology (13). Besides a peculiar immune system lacking the major histocompatibility complex class II genes (12,8), Atlantic cod displays a high level of heterozygosity, as well as an unusually high frequency of simple tandem repeats (STRs) (13,14). The STRs display repeat length variation between individuals, show a substantial degree of heterozygosity within an individual (more than 22% of STRs are heterozygous within the reference genome individual), and we have shown that these STRs are highly associated with contig breakpoints (12,13). However, such breaks are to a large extent eliminated with the long PacBio reads (13). Since teleost fish represents a deeper vertebrate lineage than mammals, with the highest number of species (and diversity) known among vertebrates, comparative studies of fishes may turn out to be very rewarding - also with respect to key mammalian and human genes. Recently, we have generated more than 100 new teleost draft genome assemblies, and we have used these for comparative studies of the immune system (e.g. 8,11), hemoglobins and antifreeze proteins in relation to depth and temperature adaptations (1,2), visual pigment genes in deep sea fishes (9) and transposable elements and STRs across teleosts (10). The wealth of information associated with these genome assemblies has, however, not at all been fully exploited.

Furthermore, we are in the process of generating more complete (reference) genome assemblies of selected species. The ability to leverage novel graph-based approaches for further comparative studies are bound to open completely new avenues of inquiry. From a population genomic perspective, the CEES-team have discovered large regions (megabase-sizes) at four different linkage groups (LGs; i.e. chromosomes) in Atlantic cod that represent chromosomal inversions. The inversions are associated with migratory (Barents Sea cod) and stationary (coastal cod) behavior, and adaptations to temperature and salinity (3,4,5,6). One example is the inversion in LG1 that is discriminatory for migratory cod ("skrei") vs. coastal (stationary) cod in the Lofoten area (migratory cod being almost homozygous for one inversion variant, with coastal being almost homozygous for the other). The inversions are characterized by linkage disequilibrium (LD), implying that there is a low recombination frequency in these regions compared to the rest of the chromosome (e.g. 6). The functional importance and to some extent the underlying mechanism(s) for LD is unclear, although it is obvious that reduced recombination will be a consequence of ecotypes with inverted chromosomal regions. However, hybrids occur at variable frequencies (depending on the LG inversion, geographic location and year - from our genomic time-series). The population genomic studies are based on genome scans (SNP-chip) and full genomes of more than 1600 cod individuals - including historic samples tracing back to 1907 (prior to industrial fisheries).

In sum, building on the existing data on Atlantic cod and codfishes (plus data in production), we have a unique opportunity to address comparative and population genomics questions - by developing and using graph-based representation tools - for understanding a key species for world-wide fisheries (and potentially an emerging aquaculture species) and its codfish allies. This could best be done by endeavouring on major outstanding evolutionary questions that rely on integrative analysis of genomic and epigenomic data, at fine (population) and large (comparative; species level) scale. At the Norwegian Sequencing Centre (NSC; www.sequencing.uio.no), various sequencing systems are available locally, as is expertise for planning, generation and validation of new data. Thus, we are now ideally suited to study large-scale variation and long-range dependencies through complex haplotype structures. Also,

existing in-house expertise in the development of large-scale user-friendly software systems will be an asset for development of an appropriate computational platform that would allow biologists and computational biologists in the CEES environment to take on computational analyses at a new level of complexity, and enable us to achieve international leader status in this field.

The aim of CELS2 is - by focusing on the achievements in graph representation - to take on the research with significant bioinformatics challenges within comparative and population genomics. By focusing on development of graph-based representations for investigation of small-scale and large-scale genomic variations (including structural variants, copy number variations (CNVs), recombinations and simple tandem repeat variation; STRs), we will be in a position to address evolutionary questions about Atlantic cod and codfishes. Although we build on current activities at CEES (fish evolutionary genomics) and IFI (graph-based representations), CELS2 represents a significant departure from current activities in the SRI group, allowing the involved environments to take on an entirely new array of research questions. To ensure critical mass for CELS2, we have chosen to define both positions with a cross-disciplinary focus, rather than tying one position to each of the environments. We have thus defined two projects with a partly overlapping focus, where we ensure that each of the projects entails the following two crucial characteristics: 1) they will permit a new class of evolutionary biology questions to be investigated, and 2) they rely on advanced computational approaches. Moreover, the results obtained in CELS2 will have implications and be useful far outside the field of fish biology.

Project plan for position "A". Analytical approaches to connect evolutionary events at large and small scales

The aim of project A is to develop novel analytical approaches for the combined investigation of large-scale and small-scale genomic variation. Previous research at CEES has uncovered inversions in linkage groups 1, 2, 7 and 12 of the Atlantic cod genome that show differential associations between cod populations and ecotypes (6). The candidate will work together with researchers at both CEES and IFI to draw on their various expertise in evolution, structural variation, cod biology, software development, data representation and statistical genome analysis. The candidate will develop novel analytical methodology that facilitates the study of small-scale events, like micro-recombinations, simple tandem repeats and nucleotide polymorphisms, and the larger-scale variation represented by the inversions in linkage groups (LG) 1, 2, 7 and 12, across populations and ecotypes. The candidate will exploit recent developments related to graph-based representations, representing genomic data on different populations and ecotypes as distinct paths of a unified graph. The unified representation allows integrative statistical analysis of small and large scale events associated with these populations and ecotypes, opening up for investigating these variations with respect to evolutionary processes (selection, mutation and drift).

Project A will utilize the Atlantic cod reference genome generated using PacBio long reads. Population genomics data from various cod individuals (sequenced at around 10x coverage) representing adaptations to different environments will be mapped against the reference genome. As an initial stage, we will assess the resulting graph structures using current graph-based methodologies like vg (https://github.com/vgteam/vg) applied with such low coverage. A main limitation of low coverage is that considerable parts of the genome (particularly heterozygous regions) cannot be resolved with high certainty. We will relate to this challenge from two angles: a) develop approaches for representing individual genomes in graph structures that quantifies uncertainty along the genome and allows for representing an individual genome by interspersed paths through a reference (paper 1), and b) develop approaches for making probabilistic inference on the paths of individual genomes by borrowing power from closely related individuals, *i.e.* from the same population (paper 2). To assess the accuracy of uncertainty and probability estimates, we will make use of a smaller subset of individuals that are already sequenced at between 20-40x coverage In addition to the Atlantic cod reference genome, which is

generated from a Barents sea (skrei) individual, CEES has also generated a coastal cod genome assembly. We will based on these linear references construct a pan-cod graph reference that will include inversion patterns on LG 1, 2 and 7, as well as CNVs with respect to MHC I and other multi-copy immune genes, for large populations of sequenced codfishes. We aim for a paper on methodology development (paper 3) and another on biological insights gained from a comparative analysis of these variants, in light of their phenotypic variation (paper 4). Finally, we will address how graph-based approaches are able to handle the large amount (thousands and more) of individuals typically used in population studies. We will investigate ways to deal with this challenge computationally, through optimized genome representations that are succinct in memory and disk, and that are well suited for hardware-accelerated computation (paper 5).

Project plan for position "B". Comparative genomics using graph-based reference genomes

Researchers at CEES have sequenced and assembled 26 different codfish species (8), four of which have been selected for generating more complete genome assemblies. In this project, the candidate will use the genomic data for the different species to create a graph structure where the large and small scale variations between species can be investigated. By integrating gene annotations into the graph, the candidate can investigate segmental duplications leading to gene copy number variations. Further, several gene families are greatly expanded in codfishes, such as MCHI, nod-like receptor genes and, to some lesser extent, hemoglobins and antifreeze proteins (1,2,8,12,14). While in mammals many high copy-number gene families are present as tandem repeat units on the same chromosome, in codfishes several of these gene families are in codfishes distributed on different chromosomes. Representing these gene families in a graph-structure will facilitate addressing how these gene families spread across genomes as results of evolutionary processes.

In Project B, a first step is to develop a genome-wide comparison of the Atlantic cod and haddock reference genomes using graph representation. Both genomes are planned in the near future to be improved by additional long-range approaches, such as 10X Genomics, HiC, and if possible, Bionano (in collaboration with the Wellcome Trust Sanger Centre, UK), providing chromosome-length scaffolding. Construction of this cross-species reference will be undertaken in collaboration with current bioinformatics researchers at CEES. In ongoing projects, we are in the process of generating reference genomes of polar cod (*Boreogadus saida*) and burbot (*Lota lota*). Polar cod is extremely adapted to freezing temperatures (living in the Arctic pack ice - tolerating temperatures below zero degrees Celsius), and the burbot is the only codfish species that has adapted to a pure freshwater environment. While polar cod is quite closely related to Atlantic cod (and haddock), the burbot belongs to the family Lotidae (ling and cusk), that is much less related to the cod family (Gadidae). Thus, developing a comparative analysis between Atlantic cod, polar cod and burbot is likely to be highly rewarding (but challenging) - with respect to genomic (and chromosomal) rearrangements, CNVs, TEs and STRs (paper 2). To achieve this, a major challenge will be to develop graph-based methodology for simultaneously handling the broad range of scales at which genomic variation occurs. This should enable representation and analysis of detailed variation, while allowing the fine scale variation to be ignored in the graph when it is computationally required to analyze large-scale variation across many species (paper 3). For unbiased and effective analyses of multiple types of evolutionary patterns, we will develop powerful visualization approaches of graph representations, enabling efficient browsing of variants and other genome annotations across scales and across species (paper 4). To follow up on insights made from a visual exploration phase, we will develop a first statistical methodology for analyzing genomic annotations on a graph reference, providing both descriptive statistical explorative analysis and inferential analysis (for statistical testing of hypotheses formed based on exploration). For this, we will leverage our published data representation format for coordinates and intervals in graph reference

genomes, and combine this with insights gained from a decade of statistical genome analysis experience through the Genomic HyperBrowser project (paper 5).

1. Baalsrud HT, et al. (2017) Evolution of hemoglobin genes in codfishes driven by ocean depth adaptation. *Scient Rep* 7:7956

2. Baalsrud HT, et al. (2018) *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol Biol Evol*, 35(3): 593-606

3. Barth JMI, et al. (2017) Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Mol Ecol* 26:4452-4466

4. Barth JMI, et al. (2018) Disentangling structural and behavioral barriers in a sea of connectivity. *Mol Ecol* (acc. p. revision)

5. Berg, P.R., et al. (2015). Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L). *Genome Biology and Evolution* 7 (6): 1644-1663

6. Berg PR, et al. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scient Rep* 6: 23246

7. Grytten I, et al. (2018) Graph Peak Caller: calling ChIP-Seq Peaks on Graph-based Reference Genomes (preprint) bioRxiv 286823; doi: https://doi.org/10.1101/286823

8. Malmstrøm M, et al. (2016) Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics* 48 (10): 1204-1210

9. Musilova Z, et al. (2018) Rod opsin-based colour vision in deep-sea fishes BioRxiv https://doi.org/10.1101/424895

10.Reinar WB, et al. (2018) Genomic repeat landscape across the teleost fish lineage. (subm)

11.Solbakken MH, et al. (2017) Linking habitat and past paleoclimate events to evolution of the teleost innate immune system. *Proc Roy Soc B* 284: 2610.2810

12.Star, B., et al. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 477: 207-210.

13.Tørresen OK, et al. (2017) An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* 18:95

14. Tørresen OK, et al. (2018) Genomic architecture of haddock (*Melanogrammus aeglefinus*) shows expansions of innate immune genes and short tandem repeats. *BMC Genomics* **19**:240