## Non-Coding RNA group

The Non-Coding RNA group at OUS/UiO focuses on studying evolution in different systems. We use Experimental, Computational and Theoretical approaches to do this and have several research projects available for masters’ students.

## Human Cytomegalovirus and Breast Cancer

Breast cancer is considered the most common malignancy and a leading cause of death in women worldwide. A key challenge is the limited treatment options and variation in response. This is commonly associated with tumour heterogeneity. However, our preliminary results demonstrate a broad presence of human cytomegalovirus (HCMV) in different types of breast cancer patients in Norway, and increased expression of CMV proteins seems to be associated with more aggressive breast cancer phenotypes such as triple-negative breast cancer3. We are characterizing the role of HCMV in promoting breast cancer by using a HCMV-infected breast cancer organoid model in combination with single-cell sequencing. We will test our findings using clinical samples from breast cancer patients to understand the role of HCMV in disease progression. Ultimately, a panel of inhibitors targeting the identified mechanisms, as well as anti-CMV drugs, will be evaluated for the effects of inhibiting cancer growth in breast cancer organoids. With this approach, we will identify novel treatment targets for CMV-associated breast cancer and provide proof-of-concept results for the potential of anti-CMV treatments as an add-on precise therapy for breast cancer in subsequent clinical trials.

* We are looking for both experimental and computational students to participate in this work.

### Using Evolutionary Therapy to Develop Optimised Triple Negative Breast Cancer Treatments

Investigation of the molecular mechanisms associated with triple negative breast cancer (TNBC) have failed to identify clinically established targets for therapy. Also, due the intrinsic heterogeneity of the tumour, drug resistance to chemotherapy commonly occurs and TNBC patients have a relatively poor outcome compared to other breast cancer patients. We are applying a strategy called “Evolutionary Therapy” (ET) to develop optimal drug treatment regimens for TNBC patients and to translate our findings to a clinical context. In this approach we investigate the evolutionary processes associated with drug treatment. Using this method, we use evolutionary game theory (EGT) to “play off” drug tolerant cells against each other to impede development of drug resistance. To our knowledge, we are the only group in Norway working on this approach.

* We are looking for experimental students to participate in this work.

## The Non-Coding Genome, Population Variation and Disease

Genome Wide Association Studies (GWAS) seek to identify Single Nucleotide Polymorphisms (SNPs) associated with a particular disease or trait. However, studies are heavily biased towards European populations and focus on the 2% of the genome that correspond to protein coding regions. More recently there have been studies such as the 1000 genomes project, the 100 000 Han Chinese Project and H3Africa that seek to more uniformly represent global populations. However, it remains a challenge to interpret SNPs that are specific to a particular population, but which occur outside protein coding regions. We have been developing computational and theoretical tools to aid interpretation of the non-coding genome. These projects are as follows:

## microRNAs (miRNAs) and isomiRs (characterization, classification and variation)

miRNAs are short non-coding RNAs that play a central role in gene regulation, and their dysregulation has been associated with a broad range of diseases8. While interesting from a research standpoint, they are also highly relevant from a clinical perspective as biomarkers or therapeutic agents9,10. There are already several miRNAs that are in a pre-clinical testing phase. For example, in breast cancer the miR-200 family has tumour suppressive function and the use of miRNA mimics to enhance tumour suppression is being tested11. Similarly, miR-10b has been identified as an onco-miR and anti-miRs are being tested for their ability to suppress miRNA expression10.

However, rather than existing as a single well-defined entity, miRNAs consist of several similar but slightly different sequences, or ***isomiRs***. The diagnostic potential of isomiRs in the study of cancers is well established. It has been shown that isomiR distributions in cancers can be used to classify different breast cancer subtypes (luminal A versus luminal B)13, and distinguish ethnicity in triple negative breast cancer14. However, these studies have only used the isomiRs distribution to classify the cancer types, for example, they haven’t looked in detail at what changes in the isomiR distribution are associated with a specific cancer. We have developed software tools to investigate small RNA Next Generation Sequencing datasets to examine how isomiR populations change between healthy and patient cohorts. We have two projects

* How do isomiR populations change among different cancers and patient cohorts? This will involve using existing software to investigate isomiR changes in publicly available Small RNA Sequencing datasets.
* How can we develop a metric that can characterise observed isomiR changes in Small RNA Sequencing datasets? This will involve further developing existing software tools and testing novel metrics to quantify isomiR changes

### The Non-Coding Genome and Population Variation

GWAS have been effective in identifying disease associated loci, but a significant fraction of genetic factors remains undiscovered. Also, the focus of GWAS remains on the interpretation of variants located within the 2% of the genome containing coding regions but many studies have found associations between non-coding genome and human disease. We are performing characterization of non-coding features to clarify how they vary by different populations and the functional consequences (for example, how the variation affects disease risk)

* We have computational projects to develop tools to analyze, visualize and interpret the data.

### non-coding genome variation: developing novel ways to characterize regulatory changes

Living systems function by the complex co-ordination of many connected parts. This connectivity exists across different scales, from food web networks in eco-systems17, down to molecular interactions within cells. Investigating this connectivity can yield important insights into system mechanisms. We are interested in investigating the collective repressive effect of many miRNAs on one mRNA in MMRNs. We have found that miRNAs only use a fraction of their repressive ability on strong interactions, the bulk of repressive effects are the result of many weak interactions20. Thus, we are investigating the changes in the connectivity or topology of miRNA-mRNA regulatory networks can be used a measure of disease risk or progression.

We have both computational and theoretical projects to develop (i) metrics to characterise changes in network topology and (ii) a model to associate disease risk with network stability

## Metagenomics

### Metagenomics

We are working with several groups to investigate changes in bacterial populations in different microbial environments. This involves 16s rRNA sequencing followed by downstream analysis.

* We have a project to identify novel viral pathogens in 16s RNA metagenomic read data using statistical learning techniques.

## Text Mining

### Scientific publishing is the standard way of publishing research findings. However, in addition to the scientific work, every publication is accompanied by a rich selection of metadata (such as country of origin, year of publication and publishing journal). By integrating this data with other information, such as journal impact factor and number of citations, it is possible to build up detailed pictures about scientific publishing. For example, which research areas are more likely to published in high impact journals or attract funding? We are using this information to investigate scientific publishing in two ways

### Text mining: Publication bias

We are investigating the potential bias that exists in scientific publishing within the context of infectious disease outbreaks. For example, until recently, Ebola virus outbreaks were limited to West Africa. However, in 2014 the first cases of the disease were recorded in Europe and the US. This was accompanied by a rapid increase in publications in high profile journals by researchers in both the US and Europe. This is part of a more general trend in which introduction of a disease in these regions appears to associate with publication in higher impact factor journals.

### Text Mining: Covid research and societal impact

One of the defining characteristics of the Covid pandemic is the volume of information that has been generated. There has been unparalleled research output (There are currently ~430 000 related PubMed publications as of 18th May 2021, and more than 5 300 000 NGS datasets). At the same time, the pandemic has seen the emergence of citizen science, such as the Reddit forum **r/COVID19** which has ~356 000 members discussing Covid-19 related research. Taken together, these data resources represent a unique opportunity to understand how scientific research is perceived by the general public. For example, what are common factors among publications that generate the most controversy? How has public opinion change over the course of the pandemic (such as the perception of Long Covid) ?

* For both projects we will use data science, statistical learning and text mining approaches to integrate and standardize the data, and to identify patterns within the data. We are also working with social scientists to interpret the data.