# Quantifying the impact of germline gene variation on immune repertoire architecture

## Background

Antibodies and T-cell receptors, called immune receptors, are key molecules of our adaptive immune system, representing nature's most finely-tuned diagnostic and therapeutics in that they recognize and neutralize with exquisite specificity any harmful particle (antigen), such as cancer, virus and bacteria. In fact, blockbuster antibody and T-cell receptor therapies have revolutionized the treatment of human disease.

An immune receptor is a sequence of ≈300 nucleotides that dictates which antigen the receptor can target and bind. Past and current antigen encounters are written invariably into the genetic sequence information of immune receptors where it is stored as immunogenomic memory. However, inferring the target antigen from the immune receptor sequence remains one of the century problems of modern medicine, constituting the primary bottleneck towards direct human prediction, modification and de novo design of human immune repertoires. Although it is currently feasible to determine the receptor sequence of millions of different unique B- and T lymphocytes – the adaptive immune repertoire of a person – but it is much harder to predict disease risk and progression from such generated data.

A particular challenge in immune repertoire analysis is that the set of receptors present in a person reflect two partly independent processes: a genetics-driven receptor generation process (stochastic rearrangement of germline DNA building blocks called V, D and J genes to produce distinct receptor sequences across immune cells called *V(D)J recombination*) and a pathogen-driven receptor selection process (expansion of immune cells upon pathogen recognition called *clonal selection*).

While the field of repertoire-based diagnostics is currently focusing mostly on dissecting disease-associated selective pressures operating on immune receptors, accounting for the influence of individual immunogenetics has gained little attention so far. Indeed, there is increasing evidence that immune receptor germline genes vary across individuals in the form of for example single or multiple polymorphisms and gene duplications. However, both the accurate identification of such immunogenomic variation as well as its influence on repertoire architecture remains an open problem. This hinders a fundamental understanding of how immunological information is encoded into each individual's immune repertoire.

Specifically, a fundamental part of immune receptor genomics is to correctly infer the germline DNA building blocks of each receptor. Each receptor can be represented as a composition of such V, D and J gene building blocks combined with stochastic modifications (nucleotide insertions and deletions). These DNA building blocks entail individual-specific usage frequencies and allelic variation. A repertoire can thus be characterized by a combination of repertoire-wide genetic influences and receptor-specific modifications bearing the marks of antigen-driven selection. Such a representation would enable a more meaningful immunological interpretation of the repertoire architecture of a person, as well as lending itself to informative repertoire encodings for use in machine learning approaches to capture disease-associated sequence motifs (immune receptor biomarkers).

## Objectives

- Probabilistically identify germline DNA building blocks of each receptor in an immune repertoire and simultaneously infer allelic gene variants
- Quantify impact of germline variants and germline gene usage on repertoire architecture (diversity and overlap between repertoires)
- Based on a decomposition of repertoires into individual-specific DNA building blocks and receptor-specific stochastic modifications, define encodings for machine learning-based classification of simulated and human immune repertoires

## Approach

The candidate will work on algorithms for inference of allelic variants and probabilistic annotation of V, D and J-gene components of immune receptors. A major focus will be on algorithms that use such receptor annotation to decompose the total variability between repertoires into components representing allelic variability, V/D/J gene segment usage frequencies and sequence-specific selective forces operating on the repertoire. These components will subsequently be used as dimensions when encoding repertoires in vector space for machine learning-based repertoire classification approaches, allowing better distinction between primarily genotype-derived versus phenotype/environment-derived differences between sets of repertoires. By analyzing simulated and experimentally observed variability for these different components, the candidate will investigate to what degree germline variability systematically impacts repertoire diversity and overlap. Furthermore, the candidate will investigate to what degree the use of such representations affects the accuracy and interpretability of machine-learning derived discriminative models for repertoire classificaiton.

## Partners

- **Pavel Pevzner:** *Ronald R. Taylor Professor of Computer Science and Director of the NIH Center for Computational Mass Spectrometry, University of California, San Diego. http://cseweb.ucsd.edu/~ppevzner/*
  *(*394 publications, 44730 citations, h-index 88)
- **Yana Safonova:** *Postdoctoral Scholar, Qualcomm Institute, University of California, San Diego.*

  *http://bioinf.spbau.ru/members/yana-safonova*
  *(*6 publications, 99 citations, h-index 5)

- **Geir Kjetil Sandve:** *Associate professor, Department of Informatics, The Faculty of Mathematics and Natural Sciences, University of Oslo.*
  *https://www.mn.uio.no/ifi/personer/vit/geirksa*
  *(*73 publications, 1342 citations, h-index 17)
- **Victor Greiff:** *Associate professor, Department of Immunology, Faculty of Medicine, University of Oslo.*
  *https://www.med.uio.no/klinmed/english/people/aca/victogre/*
  *(*22 publications, 292 citations, h-index 8)

*(bibliometric values according to Google Scholar)*

## Reasons for UCSD-UiO collaboration

- Joint interest in immune receptors
- Joint interest in (immuno)genomics
- Overlapping computational background – e.g., sequence algorithms, patterns discovery.
- Complementary needs
  - UiO seeks better computational representations of the information inherent in immune repertoires
  - UCSD seeks to investigate further how properties of the receptor structure propagate to the repertoire and population scale

## Educational component for the PhD candidate

Depending on the background of the selected candidate, the candidate will follow a subset of the following courses at UiO/Simula:

- Autumn 2019: Statistical learning methods in Data Science (STK-IN9300).
  https://www.uio.no/studier/emner/matnat/math/STK-IN9300/index.html
- Spring 2020: Science, ethics and society (MNSES9100).
  https://www.uio.no/studier/emner/matnat/ifi/MNSES9100/index.html
- Spring 2020: Machine Learning for Image Analysis (IN5400).
  https://www.uio.no/studier/emner/matnat/ifi/IN5400/index-eng.html
- Summer 2020: Bioinformatics and statistical analysis of antibody and T-cell repertoires. Workshop as part of National research school in bioinformatics, biostatistics and systems biology (NORBIS). https://norbis.w.uib.no/activities/workshops
- Spring/autumn 2020: Communicating Scientific Research 2020.
  https://www.simula.no/education/courses/communicating-scientific-research-2019

## Monitoring progress and development

The candidate will have a supervisor group consisting of two supervisors at UCSD (Pavel Pevzner and Yana Safonova) and two supervisors at UiO (Geir Kjetil Sandve and Victor Greiff). The candidate will be situated with and part of all research group activities for the Pevzner/Sandve group when having research stay at UCSC/UiO, respectively. The PhD candidate will work on the same research project when situated at UCSC and UiO, though with a focus more towards algorithmic inference of repertoire composition when situated at UCSC and more towards analysis and classification using the repertoire composition when situated at UiO.

To ensure that all supervisors are continually kept up to date with the progress of the PhD candidate, and that all aspects of the research are consistently covered, we will have bi-weekly status meetings on Skype involving all supervisors. These will be on a weekday at around 10PM Oslo Time / 1PM San Diego time, to make them practical with regards to the time difference (Sandve and Greiff have good experience with this from a current US collaboration).