Amariuta Lab

Project Statement (English)

While some human diseases are regulated by a single gene, most diseases are regulated by many, sometimes hundreds of factors, both genetic and environmental. These specific genetic factors, or nucleotides in the DNA sequence, vary between individuals. Genome-wide association studies (GWAS) have identified and quantified thousands of associations between single nucleotide polymorphisms (SNPs) and phenotypes, including diseases such as diabetes and anthropomorphic traits such as height. It has been estimated that more than 90% of these associations implicate SNPs in the noncoding genome. The noncoding genome is a vast space, accounting for approximately 98% of the 3 billion nucleotides in the human genome. For this reason, trait-associated SNPs identified in these noncoding regions are typically poorly characterized and have uncertain function. SNPs may have functional impacts on both proximal and distal DNA sequences, or regulatory elements, and may affect protein binding affinities, the accessibility or 3D conformation of the chromatin, and more. Adding to this complexity, the regulatory elements that are functionally affected by SNPs operate in cell-type-specific contexts. For these reasons, understanding regulatory variation, especially in the noncoding genome, may contribute valuable insight into human diseases and traits.

In the lab, we will computationally integrate transcriptomic, epigenomic, and genetic data, characterizing functional activity of different regulatory elements, in order to better understand mechanisms of human disease, focusing on the autoimmune disease rheumatoid arthritis. In terms of data, we will use single cell RNA sequencing of human immune cells collected from healthy (control) and diseased (case) individuals, transcription factor and protein binding data from ChIP-seq (chromatin immunoprecipitation sequencing), and human GWAS SNP-trait association values. First, we will identify transcriptomic and cellular differences between cases and controls. Specifically, we will identify which genes are differentially expressed between cases and controls and which cellular populations are differentially expanded. Second, we will make mechanistic hypotheses for these differences by integrating ChIP-seq. Specifically, we will identify which proteins are observed to bind in promoters or enhancers, identified with HiC data, of differentially expressed genes and marker genes of differentially expanded cellular populations. Third, we will integrate our findings with GWAS data to test if the SNPs located near our identified candidate regulatory genes have larger disease association values than expected by chance, validating our transcriptomic and epigenomic findings.

This project will require programming in R, or another familiar language, and statistical applications, both of which will be taught to students; no background in programming or statistics is required.

Project Statement (Russian)

Несмотря на то, что некоторые человеческие болезни регулируются каким-то одним геном, большая часть заболеваний зависит от множества факторов генетики и окружающей среды. Такими факторами являются нуклеотиды ДНК, которые различаются у разных индивидуумов. Полногеномный поиск ассоциаций (GWAS) определил тысячи однонуклеотидных полиморфизмов (SNPs), взаимосвязанных с фенотипами, включая болезни (например, диабет) и внешние признаки (например, рост). Согласно расчетам, более 90% всех SNPs приходятся на некодирующую часть генома, которая занимает примерно 98% от трех миллиардов нуклеотидов, составляющих весь человеческий геном. По этой причине связанные с фенотипами SNPs из некодирующих областей обычно мало

изучены, а их функции неизвестны. SNPs могут оказывать воздействие как на ближние, так и на отдаленные гены, или на регуляторные элементы, и влиять на аффинность связывания белков или конформацию хроматина. Более того, регуляторные элементы, зависящие от SNPs, ведут себя по-разному в зависимости от типа клеток. Поэтому изучение изменчивости, особенно в некодирующей части генома, может внести ценный вклад в понимание человеческих болезней.

В этой лаборатории мы совместим транскриптомные, эпигенетические и генетические данные, характеризующие активность различных регуляторных элементов, чтобы лучше понять механизмы болезней, и сосредоточимся на аутоиммунном заболевании – ревматоидном артрите. Мы используем секвенирование РНК одиночных клеток иммунной системы, взятых у здоровых (контрольная группа) и больных (экспериментальная) людей, информацию о связывании транскрипционных факторов и белков из ChIP-seq (секвенирование иммунопреципитации хроматина) и данные о взаимосвязи между SNPs и фенотипами из GWAS. Сначала мы установим транскриптомные и клеточные различия между контрольной и экспериментальной группами. В частности, мы определим разницу в экспрессии генов и расширении популяций клеток. После, используя данные ChIP-seq, мы выдвинем гипотезу, объясняющую эти различия. Основываясь на данных HiC, мы определим, какие белки связываются с промоторами и энхансерами различно экспрессируемых и маркерных генов в пределах клеточных популяций. Затем мы совместим результаты с данными GWAS, чтобы проверить, проявляют ли SNPs, расположенные рядом с найденными регуляторными генами, статистически значимый уровень ассоциированности с болезнями.

Работа над проектом потребует программирования на R и статистики, чему студентов обучат непосредственно в лаборатории. Никаких предварительный знаний не требуется.