Genetics of human autoimmunity: From genetic information to functional insights

Kazuyoshi Ishigaki [a,b,*], Yuta Kochi [a], Kazuhiko Yamamoto [a,c]

[a] Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
[b] Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
[c] Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

ARTICLE INFO

ABSTRACT

Genome-wide association studies have identified hundreds of risk variants associated with human autoimmune diseases. Recent evidence suggests that a substantial portion of them affect gene expression in specific cell types. To obtain the functional insights of GWAS findings, comprehensive characterization of genetic variants in human genome is a key task. In parallel with GWAS, many researches in functional genomics have been conducted in the past decade, and our understandings of cell type-specific gene regulatory system have been substantially improved. In this review, we will introduce the main research topics in functional genomics, and explain their utility to understand biological mechanisms of autoimmune diseases.

1. What is genome-wide association study?

Genome-wide association study (GWAS) is a genetic study assessing the association of genetic variants with human complex diseases or quantitative traits (such as height). GWASs typically focus on single-nucleotide polymorphisms (SNPs), but other types of variants can also be assessed. Since the first successful GWAS in 2005 [1], many GWASs have been conducted, and successfully discovered thousands of variants associated with human diseases and traits. GWASs examine variants across all the genome. Therefore, it can potentially detect association signals suggesting a variety of mechanisms, and it is a promising way to study human autoimmune disorders with complex mechanisms. Unlike most of biological experiments, GWAS is a hypothesis-free study. Its unbiased nature enables us to find clues for novel biological mechanisms of target diseases.

2. Why is GWAS important?

GWAS associations are usually linked to causal mechanisms because it examines germline variants (Fig. 1). In the other types of case-control studies, there is a problem so called "reverse causality", where the findings may reflect the consequences rather than the causes. For example, C-reactive protein (CRP), a serum marker of acute inflammation, is often upregulated in rheumatoid arthritis (RA), and CRP is often used to estimate disease activity of RA [2]. However, the upregulation of CRP in RA is likely to be a consequence of inflammatory process, not a causal

mechanism of RA [3]. In contrast, GWASs of RA identified that rs2228145, a missense variant in IL6R, is associated with RA [4,5], and this suggests IL6 pathway is one of the causes, not the consequences of RA.

3. GWAS for autoimmune diseases

GWASs have discovered thousands of variants associated with human diseases (we define them as "GWAS variants"). RA is one of the most investigated diseases by GWASs, and there are around 100 loci associated with RA [4,6–8]. Multiple immunologically important genes were located around GWAS variants of RA, such as CD28, CTLA4, TYK2, IL6R and CD40. Recently, systemic lupus erythematosus (SLE) was studied by large-scale GWAS meta-analyses [9,10], and around 50 associated loci were identified. Substantial proportion of SLE GWAS variants possesses transcriptional regulators in their neighborhood, such as IRF5, IRF7, IRF8, PRDM1, and STAT4.

4. Research goals in post-GWAS era

The primary goal of GWAS is to find associated variants. In that sense, GWASs have been very successful. However, the biological mechanisms underlying GWAS variants are largely unknown. Unlike Mendelian diseases, in which a single missense or nonsense variant usually causes a disease, growing evidence show that accumulation of regulatory changes in gene expression is the main mechanism of complex diseases. In fact, while coding variants are estimated to explain only <10% of heritability in complex diseases, the majority of GWAS variants map in the non-coding regions [11]. In the 'post-GWAS' era, where we already know thousands of GWAS variants, their functional

* Corresponding author at: Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, 1-7-22, Suehiro, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan.
E-mail address: kazuyoshi.ishigaki@riken.jp (K. Ishigaki).
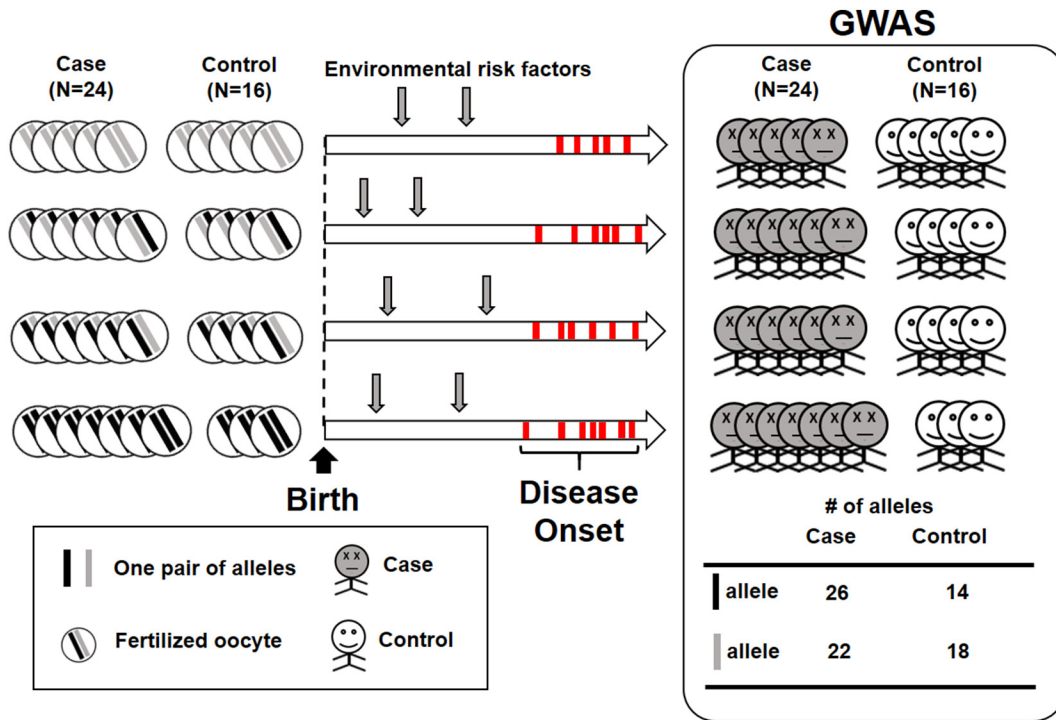
Fig. 1. GWAS variants are linked to the causal mechanisms of human diseases. We consider a situation where GWAS identified a locus associated with a disease (right box), and the black allele is the risk allele in this locus (its odds ratio is $26 \times 18/(14 \times 22) = 1.51$). Since genotypes are unchanged after fertilization, the differences in allele frequency between cases and controls are generated well before the disease onset. Therefore, black allele is a cause of the disease. Although some environmental factors might have causal effects (grey arrow), simple case-control comparison of genotype is sufficient for GWAS if we assume genetic and environmental factors are mutually independent.

interpretation, with a special focus on their gene regulation, is one of the main research goals. In the next section, we will introduce several research topics aiming this goal.

## 5. Application of functional genomics into GWAS findings

5.1. Epigenome data

In the past 10 years, epigenomic landscape of human genome was intensively studied [12–15]. GWAS variants of common complex diseases are enriched in non-coding regions, especially regulatory regions like promoters and enhancers [14,16]. Moreover, these enrichments were prominent in specific cell populations [14,16]. For examples, GWAS variants of immunological diseases were enriched in the enhancer regions of immune cells, whereas those of hyperlipidemia were enriched in the enhancer regions of liver [14,16]. These findings suggested that the altered regulation of gene expression in certain cell populations is the causal mechanism of diseases investigated by GWASs. The cell types that showed significant enrichments are consistent with our current understanding of the disease biology in most cases. Currently, this type of enrichment analysis is a widely accepted strategy to assess candidate causal cell populations, and identified plausible causal cell types or tissues: regulatory T cells (T-regs) for RA [16], pancreatic islets for fasting glucose-related traits [14], and nervous system for neuropsychiatric disorders [16]. Although very useful, the entire GWAS variants are treated as one datum, and the biological mechanism for each GWAS locus cannot be assessed in these analyses.

5.2. Expression quantitative trait loci analysis

Expression quantitative trait loci (eQTL) analysis is a genetic study assessing the association of genetic variants with gene expression levels (Fig. 2). eQTL analysis can tell us the clues for causal mechanisms of GWAS variants: i) which genes are regulated by GWAS variants, ii) in which cell types, and iii) in which directions (up or down-regulation).

These candidate causal mechanisms are very important to understand the pathophysiology of the disease investigated. Many eQTL studies focusing on immune system have been conducted so far [17–20]. However, most previous eQTL studies have utilized unfractionated peripheral blood or a few cell populations.

Recently, our group conducted a cell type-specific eQTL analysis on five immune cell populations (CD4+ T cells, CD8+ T cells, B cells, NK cells, and monocytes) from 105 healthy Japanese volunteers [21]. This is one of the largest RNA-seq based eQTL studies focusing on multiple immune cells. We successfully detected thousands of eQTL variants for each cell population, and deposited the main results of eQTL analysis to a National Bioscience Database Center (NBDC, http://humandbs. biosciencedbc.jp/en/). Candidate causal genes, their dysregulation
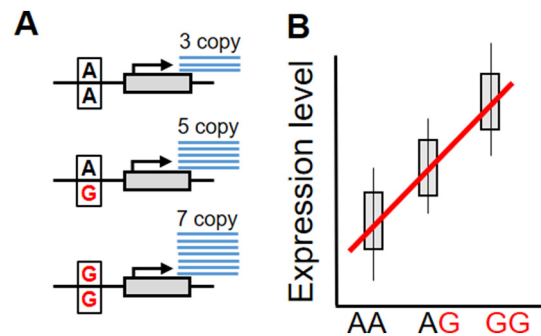
Fig. 2. eQTL analysis is a study assessing the association of genetic variants with gene expression levels. (A) Schematic illustration of eQTL variant. In this example, a eQTL variant has A and G allele, and the G allele increases the expression of a near-by gene (grey box). (B) Schematic illustration of eQTL analysis. Firstly, we sorted samples according to the genotype of eQTL variant (AA, AG, or AG). As illustrated in a boxplot, we usually observe substantial amount of variation in the expression level in each genotype group because non-genetic factors also affect gene expression level. Therefore, relatively large sample size is required when the eQTL effect size is modest. We regress the expression level on the genotype, and estimate the size and the significance of eQTL effect.