Mapping Rheumatoid Arthritis Susceptibility through Integrative Bioinformatics and Genomics

Nining Medi Sushanti¹, Wirawan Adikusuma²,³, Arief Rahman Afief¹, Anita Silas La’ah⁴, Firdayani³, Rockie Chong⁵, Zainul Amiruddin Zakaria⁷, Barkah Djaka Purwanto⁸,⁹, Rahmat Dani Satria¹⁰,¹¹, Riat El Khair¹⁰,¹¹, Abdi Wira Septama⁶, Lalu Muhammad Irham¹,⁶*¹

¹Faculty of Pharmacy, Universitas of Ahmad Dahlan, Yogyakarta, 55164, Indonesia
²Department of Pharmacy, University of Muhammadiyah Mataram, Mataram, 83127, Indonesia
³Research Center for Vaccine and Drugs, National Research and Innovation Agency (BRIN), South Tangerang, Indonesia.
⁴Taiwan International Graduate Program in Molecular Medicine, National Yang Ming Chiao Tung University and Academia Sinica, Taipei, Taiwan
⁵Department of Chemistry and Biochemistry, University of California, Los Angeles, USA
⁶Traditional Medicine, National Research and Innovation Agency (BRIN), South Tangerang, Indonesia
⁷Borneo Research on Algesia, Inflammation and Neurodegeneration (BRAIN) Group, Department of Biomedical Sciences, Faculty of Medicines and Health Sciences, University Malaysia Sabah, Malaysia
⁸Faculty of Medicine, University of Ahmad Dahlan, Yogyakarta, 55191, Indonesia
⁹PKU Muhammadiyah Bantul Hospital, Bantul, Yogyakarta, 55711, Indonesia
¹⁰Department of Clinical Pathology and Laboratory Medicine, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
¹¹Clinical Laboratory Installation, Dr. Sardjito Central General Hospital, Yogyakarta 55281, Indonesia
*Correspondence author: e-mail: lalu.irham@pharm.uad.ac.id

ARTICLE INFO

ABSTRACT

Rheumatoid arthritis (RA) is an autoimmune disease that influences several organs and tissues, especially the synovial joints, and is associated with multiple genetic and environmental factors. Numerous databases provide information on the relationship between a specific gene and the disease pathogenesis. However, it is important to further prioritize biological risk genes for downstream development and validation. This study aims to map RA-association genetic variation using genome-wide association study (GWAS) databases and prioritize influential genes in RA pathogenesis based on functional annotations. These functional annotations include missense/nonsense mutations, cis-expression quantitative trait locus (cis-eQTL), overlap knockout mouse phenotype (KMP), protein-protein interaction (PPI), molecular pathway analysis (MPA), and primary immunodeficiency (PID). 119 genetic variants mapped had a potential high risk for RA based on functional scoring. The top eight risk genes of RA are TYK2 and IFNGR2, followed by TNFRSF1A, IL12RB1 and CD40, C5, NCF2, and IL6R. These candidate genes are potential biomarkers for RA that can aid drug discovery and disease diagnosis.

This is an open access article under the CC–BY-SA license.

1. Introduction

Rheumatoid arthritis (RA) is the most common autoimmune disease in the world. It is an uncontrollable chronic inflammatory disease with various multi-system modalities followed by the proliferation of the synovial tissue that causes pain, joint erosion, and functional disorder (Garner et al., 2014). This synovial tissue proliferation can cause chronic impairments (Song & Lin, 2017). One sign of RA is persistent inflammation in the synovial joints. RA prevalence is within the age range of 40 – 60 which is between 0.5% and 1%, with greater prevalence in women than in men (Song & Lin, 2017). The main risk factors of RA are gender, family history, advanced age, exposure to

10.12928/mf.v20i1.24912 mediafarmasi@pharm.uad.ac.id

Keywords
Rheumatoid arthritis
Autoimmune
Genome
Bioinformatics
Genetic variation
silicate, and smoking cigarette (Garner et al., 2014). The primary influential determinants for RA vulnerability are genetic and environmental factors (Deane et al., 2017).

In this study, Genome-Wide Association Study (GWAS) is used to identify genetic variants that are at risk for disease pathogenesis by applying functional annotation criteria. This can also be useful in medicine through the bioinformatics-based approach (Narayan, 2016; Reay & Cairns, 2021). GWAS is a research approach used in finding the genetic variation associated with certain diseases, this involves the entire genome of several individuals. GWAS has been widely developed as a screening method not only in early-stage patients but also for people who are at risk based on family history. This research approach is useful for finding Single Nucleotide Polymorphisms (SNPs) that contribute to complex diseases such as cancer, infectious diseases (AIDS, leprosy, hepatitis), autoimmune diseases, neuropsychiatric, and various diseases (Fareed & Afzal, 2013).

Meta-Analysis and GWAS have identified more than 100 RA loci, and GWAS is known as a database with genetic architecture also for RA (Kwon et al., 2020). GWAS is recognized as a powerful approach to map genes responsible for various diseases (Hettiarchachi & Komar, 2022). For example, earlier studies have analyzed Chinese population SNPs that affect RA by using GWAS (Ma et al., 2013). In addition, the whole genome case-control of the linkage disequilibrium (LD) mapping was conducted by utilizing SNPs and RA-associated polymorphisms in PADI4 and SLC22A4/A5 cluster was identified (Too et al., 2012). These findings showed that PADI4 is a risk gene for RA while SLC22A4/A5 has several polymorphisms associated with many autoimmune diseases. Thus, large-scale LD mapping seems effective at identifying RA polymorphisms (Ren et al., 2014).

Information on the relationship between a given gene and the pathogenesis of a disease is largely available on the GWAS database. The priority given to the most influential genes (biological risk genes) is scarcely examined. Thus, this research was conducted based on six functional annotations resulting in the prioritization of the most influential genes in RA pathogenesis. Biomarker research arises based on the need to understand the pathogenesis that causes RA (Gavrilá et al., 2016). Hence, data from GWAS could help us to identify novel RA-associated genes as biomarkers and provide new insight into managing and treating RA. The search for new biomarkers for RA plays an important role in identifying candidate genes at risk for RA.

2. Materials and Methods

2.1. RA-associated genes

Genomic variants of RA were collected from the GWAS database, and the most influential genes were prioritized based on six functional annotations (Figure 1). The variants related to RA were expanded using the criterion r² > 0.8 of HaploReg v4.1. The SNPs associated with RA were denoted as “RA-associated genes”. The genomic data were then prioritized based on six functional annotations, where genes with scores ≥ 2 were classified as “biological RA risk genes”. HaploReg v4.1 was used to dictate RA-associated SNPs (Ward & Kellis, 2016). The variants were expanded on high LD (r² > 0.8) in an Asian population, based on the 1000 Genome Project. We used the LD criterion as r² > 0.8, based on Okada et al., studies (Okada et al., 2014).

2.2. Functional annotations of RA-associated genes

Candidate genes most likely to be RA targets were prioritized based on their biological function using six functional annotations. Missense or nonsense mutations are single amino acid changes that produce different protein functions and are selected as functional annotation criteria (Zhang et al., 2012). In accordance with HaploReg v4.1, the first functional annotation is missense or nonsense mutation because it contains functional annotations from the SNPs database (DB) (Leng et al., 2020). HaploReg v4.1 was used to link cis-expression quantitative trait loci (cis-eQTL) to genetic variants in blood target tissues (Ward & Kellis, 2016). We also used WebGestalt 2019 to understand the relationship between mutant genes and phenotypes (Wang et al., 2017). Phenotype information of mouse and other mammals was derived from the ontology of the Mammalian Phenotype (MP) (Smith & Eppig, 2012). To further understand the phenotypes of mouse, we used BioMart to convert genes with human Ensembl ID into mouse Ensembl ID (Drost & Paszkowski, 2017). Genes with False Discovery Rates (FDRs) in mouse phenotypes <0.05 were considered to be significant results. RA-associated genes from biological process networks were identified using protein-protein
interactions (PPIs) (Qiu et al., 2021). In addition, we used WebGestalt 2019 to conduct reinforcement analysis and to investigate whether the genes were collected on certain functional annotations (Liao et al., 2019). Kyoto Encyclopedia of Genes and Genomes (KEGG) was used to determine what molecular pathways were gained from the list of RA-associated genes and what genes were involved. Further analysis of the biochemical pathway was carried out by WebGestalt 2019 database. The last annotation criterion was primary immunodeficiency (PID). PID refers to inborn immunity diseases and is often reported to be associated with cancer (Mortaz et al., 2016). PID attacks the immune system and causes inflammation in RA (Dimitriades & Sorensen, 2016). The hypergeometric test was employed for reinforcement analysis of these data, with the significance criterion being \( p \)-value < 0.05.

![Figure 1](image)

Fig 1. Schematic model showing how genome-based information can be integrated to map susceptibility genes for RA [Missense/nonsense; cis-expression quantitative trait locus (cis-eQTL); protein-protein interaction (PPI); molecular pathway analysis (MPA); and primary immunodeficiency (PID)]

3. Results and Discussion

3.1. Prioritized SNPs potentially cause RA

In this study we obtained 2,732 RA-associated SNPs which were dictated in the GWAS database (Table S1) and 2012 SNPs were obtained after filtering based on \( p \)-value < 10\(^{-8}\). Importantly, 396 SNPs were obtained after checking for duplicates, and further filtering based on the inclusion criterion OR >2, led to a total of 360 RA-associated SNPs (Table S2). The SNPs collected from GWAS were expanded using HaploReg V4.1 while the RA-associated genes were prioritized based on six functional annotations, namely missense/nonsense mutation, cis-eQTL, knockout mouse phenotype (KMP), protein-protein interaction (PPI), molecular pathway analysis (MPA), and primary immunodeficiency (PID).

3.2. RA risk SNPs Functional Annotations

This study was designed to prioritize the best candidate genes by using six functional annotations with a scoring system. In the functional annotation process, candidate genes with higher scores represent greater biological influences on RA pathogenesis (Figure 2). Following the six annotations, we discovered 39 SNPs of missense/nonsense mutation variants from the total 360 SNPs (Figure 3). We further examined the cis-eQTL effect using HaploReg v4.1, 101 SNPs were gained from the total 360 SNPs (Figure 3). Phenotype data was taken from the MP ontology which contains information about the phenotype of mice and other mammals. WebGestalt 2019 was applied to perform over-representation analysis (ORA) and we found 102 genes overlapping with RA risk genes (FDR <0.05). Gene Ontology (GO) annotations derived from WebGestalt 2019 were used to evaluate PPIs. Following the analysis on PPIs, 122 genes overlap with other RA risk genes (FDR < 0.05). KEGG was used to perform ORA on the molecular pathway. Thus, over 61 RA-
associated genes on KEGG pathways were identified. We also used the PID data from The International Union of Immunological Societies (IUIS) to analyze and confirm the overlapping genes. Sixteen overlapping genes were statistically significant ($p < 0.05$) from PID analysis.

We also scored the risk genes from 0 to 6 and obtained 156 genes with a score of 0. We noted that 85 genes have a score of 1, 45 genes have a score of 2, 40 genes have a score of 3, 26 genes have a score of 4, 6 genes have a score of 5, and 2 genes have a score of 6 (Figure 4). In total, we obtained 119 genes with a score of ≥ 2, that were categorized as ‘biological RA risk genes’ (Table S2). Using the aforementioned scoring criteria, the top eight RA risk genes are: tyrosine kinase 2 (TYK2), followed by interferon-gamma receptor 2 (IFNGR2). Members of the tumor necrosis factor receptor superfamily 1A (TNFRSF1A), interleukin receptor subunit 12 beta 1 (IL12RB1) and cluster of differentiation 40 (CD40), complement 5 (C5), neutrophil cytosolic factor 2 (NCF2), and interleukin receptor 6 (IL6R).

![Image](https://example.com/image.png)

**Fig. 2.** Functional annotations for rheumatoid arthritis (RA) with scores ≥ 2 based on defined scoring criteria

Tyrosine kinase 2 (TYK2) is a member of the Janus kinase (JAK) family of transforming growth factors and cytokines. The TYK2 loss analysis function reveals its importance toward immunity against infections, inflammation (automatic), and immunity (automatic). Many genome-wide association study (GWAS) in humans propose relationships between TYK2 genetic variants and many autoimmunity diseases, inflammatory diseases, and tumors. Hence, TYK2 emerges as a target of interest for therapeutic interventions (Leitner et al., 2017). The human interferon-gamma receptor is a multimer of IFN-γR1 (coded by IFNGR1) and IFN-γR2 (two chains). Interferon-gamma receptor 2 (IFN-γR2) is a gene containing a protein and in humans, this gene is coded by the IFNGR2 gene. The IFNGR2 gene codes for beta chains that are not bound to the ligands of the IFN-γR (Kong et al., 2013).

Tumor necrosis factor receptor 1A (TNFRSF1A) is a CD120a which is a member of the superfamily tumor necrosis factor alpha-binding membrane receptor. The TNFRSF1A gene instructs the making of tumor necrosis factor receptor 1 (TNFR1) protein (Chen et al., 2015). This protein is found stretching on the cell membrane, both extracellularly and intracellularly (Chen et al., 2015). Extracellularly, the TNFR1 protein enchains tumor necrosis factor (TNF). The TNF and the TNFR1 protein interaction catalyzes the latter to chain to two other TNFRI proteins, constituting trimer, a three-protein complex (Wajant & Siegmund, 2019). This formation of the trimer is important for the TNFRI protein to function. The chains of the protein of TNF and TNFRI prompt the latter to deliver a signal into the cell (Li et al., 2013). The TNFRI protein signal can stimulate inflammation or cell self- destruction (apoptosis). The signal into the cell starts a routeway activating the nuclear factor kappaB protein triggering inflammation and directing to the cytokine production, the immune system protein (Chen et al., 2015).
Fig. 3. Number of genes for each functional annotation that is prioritized for RA [Missense/nonsense; cis-expression quantitative trait locus (cis-eQTL); Knockout mice (KOmice); protein-protein interaction (PPI); Kyoto Encyclopedia of Genes and Genomes (KEGG); and primary immunodeficiency (PID)]

Fig. 4. Gene frequency distribution across defined functional scores for RA, showing 156 genes with a score of 0.85, genes with a score of 1, and 119 (45 + 40 + 26 + 6 + 2) genes with a total score of ≥ 2 are classified as RA risk genes

Interleukin-12 receptor beta-1 subunit (IL12RB1) functions as an interleukin receptor that binds to interleukin-12 with low affinity which is involved in IL12 transduction (Floss et al., 2016). IL12RB2 makes a functional high-affinity receptor for IL12 (Chen et al., 2015). IL23R forms the receptor of interleukin-23 functioning in IL23 signal transduction via JAK/STAT signaling cascade activation (Pastor-Fernández et al., 2020). D40 is a member of the tumor necrosis factor (TNF) superfamily found in antigen-presenting cells, including dendritic cells (DC), B cells, monocytes, and macrophages, which play an important role in activating the immune system (Laha et al., 2021). The expression expands to various non-hematopoietic cells, including nerve cells, epithelial cells, endothelial cells, and fibroblasts (Karnell et al., 2019). Typical genetic variants at the TRAF1-C5 locus on the chromosome correlate with an increased risk of anti-ccp-positive RA (Huang et al., 2019).

Genetic variants rs7021206 and rs3761847 at the TRAF1-C5 locus are associated with RA in the Han Chinese population, indicating that TRAF1-C5 may play an important role in the development of RA thereby reducing the pathogenic role of TRAF1-C5 in ethnic variations (van Steenbergen et al., 2015). NCF2 is carried to the membrane of the cell to be combined with other components to build aNOX system that is active through microbial stimulation. In other studies, the polymorphism NCF2 rs10911363 was linked to SLE risk in Sweden and US populations. The variant of NCF2 rs789181 was discovered to have a relationship with mild RA in men (T.-P. Zhang et al., 2020). IL6 is an important cytokine in mediating inflammation and RA systemic overview, including synovitis, fatigue, anemia, anorexia, and osteoporosis (Narazaki et al., 2017).
In this study, we used the GWAS database to obtain the genomic variants for RA. The GWAS database is easily accessible and effective because it is diverse and structured. The GWAS method effectively identifies genetic loci associated with the pathogenesis of a disease, including prioritizing candidate loci, exploring disease mechanisms, predicting disease risk, identifying new drug targets, and gathering information about the desired trait or population (Caliskan et al., 2021). However, the GWAS database has drawbacks in that detailed phenotypic descriptions of the study's ethnic background are not provided by many GWAS papers, and even nowadays, there is no standard way to make an ethnicity report (Hart & Kranzler, 2015). The prospect for further research is to develop a wider use of functional annotation criteria to get more results to determine genetic variations that affect RA pathogenesis and identify drug candidates that target biological genes for RA risk. The impact of this research is to find out the genetic variation that influences the pathogenesis of RA and gain an understanding by utilizing genetic variation that can be used for drug development by using a bioinformatics approach.

4. Conclusion

Our research showed that genomic information is highly important for mapping RA genomic variants based on defined functional annotations from this study. Two candidate genes, TYK2 and IFNGR2, were obtained based on defined scoring criteria, and we propose these genes as potential RA biomarkers to be prioritized for downstream clinical development and validation.

Author Contributions: Nining Medi Sushanti, Lalu Muhammad Irham, and Lalu Muhammad Irham conceived and designed the study. Nining Medi Sushanti and Lalu Muhammad Irham performed the computational analysis. Nining Medi Sushanti wrote the manuscript. Nining Medi Sushanti, Lalu Muhammad Irham, Wirawan Adikusuma, Anita Silas La’ah, Arief Rahman Afief, Zainul Amiruddin Zakaria, Firdayani, Rahmat Dani Satria, Riat El Khair, Abdi Wira Septama and Rockie Chong, revised the manuscript. Lalu Muhammad Irham supervised and coordinated this study. All authors have read and approved the manuscript and have made significant contributions to this study.

Funding

No funding was provided for this research.

Competing Interests

The authors have not declared a conflict of interest for this article.

Acknowledgment

I would like to thanks to my alma mater Universitas Ahmad Dahlan as always support my study as well as my school SMK Semesta Bumiayu for giving me opportunity to accomplish my master degree.

References


Dimitriades, V. R., & Sorensen, R. (2016). Rheumatologic manifestations of primary


Sushanti et al. (Mapping Rheumatoid Arthritis...

