ASSOCIATION ANALYSIS FOR BIG DATA RELATED TO RHEUMATOID ARTHRITIS BASED ON HAPLOTYPE BLOCK PARTITIONING AND SINGLE NUCLEOTIDE POLYMORPHISMS

By

Mohamed Nagy Saad Mohamed Elziftawy

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Biomedical Engineering and Systems

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

ASSOCIATION ANALYSIS FOR BIG DATA RELATED TO RHEUMATOID ARTHRITIS BASED ON HAPLOTYPE BLOCK PARTITIONING AND SINGLE NUCLEOTIDE POLYMORPHISMS

By
Mohamed Nagy Saad Mohamed Elziftawy

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Biomedical Engineering and Systems

Under the Supervision of

Prof. Dr. Ayman M. Eldeib

……..

Professor
Biomedical Engineering and Systems Department
Faculty of Engineering, Cairo University

Prof. Dr. Olfat G. Shaker

……..

Professor
Medical Biochemistry and Molecular Biology Department
Faculty of Medicine, Cairo University

Assoc. Prof. Dr. Mai S. Mabrouk

……

Associate Professor
Biomedical Engineering Department
Faculty of Engineering, Misr University
for Science and Technology

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

ASSOCIATION ANALYSIS FOR BIG DATA RELATED TO RHEUMATOID ARTHRITIS BASED ON HAPLOTYPE BLOCK PARTITIONING AND SINGLE NUCLEOTIDE POLYMORPHISMS

By
Mohamed Nagy Saad Mohamed Elziftawy

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Biomedical Engineering and Systems

Approved by the
Examining Committee

_____

Prof. Dr. Mohamed W. T. Fakhr, External Examiner
Computer Engineering Department, AAST

_____

Prof. Dr. Ahmed H. Kandil, Internal Examiner

_____

Prof. Dr. Ayman M. Eldeib, Thesis Main Advisor

_____

Prof. Dr. Olfat G. Shaker, Member
Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Cairo University

_____

Assoc. Prof. Dr. Mai S. Mabrouk, Member
Biomedical Engineering Department, MUST

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

Engineer's Name: Mohamed Nagy Saad Mohamed
Date of Birth: 20/03/1983
Nationality: Egyptian
E-mail: nagybioeng@yahoo.com
Phone: 01111606852
Address: 11 Ali Elshakh st.,Shoubra,Cairo

Registration Date: 01/03/2012
Awarding Date: …./…./ 2017
Degree: Doctor of Philosophy
Department: Biomedical Engineering and Systems

Supervisors:

Prof. Ayman M. Eldeib
Prof. Olfat G. Shaker
Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Cairo University
Assoc. Prof. Mai S. Mabrouk
Biomedical Engineering Department, MUST

Examiners:

Prof. Mohamed W. T. Fakhr (External examiner)
Computer Engineering Department, AAST
Prof. Ahmed H. Kandil (Internal examiner)
Prof. Ayman M. Eldeib (Thesis main advisor)
Prof. Olfat G. Shaker (Member)
Assoc. Prof. Mai S. Mabrouk (Member)

Title of Thesis:

Association Analysis for Big Data Related to Rheumatoid Arthritis Based on Haplotype Block Partitioning and Single Nucleotide Polymorphisms

Key Words:
Genetic Association Study; Haplotype Block; Linkage Disequilibrium; Rheumatoid Arthritis; Single Nucleotide Polymorphism

Summary:

Genetics of autoimmune diseases represent a growing domain with surpassing biomarker results with rapid progress. Rheumatoid arthritis (RA) is an autoimmune disease which has a significant socio-economic impact. The exact cause of RA is unknown, but it is thought to have both a genetic and an environmental bases. This thesis is concerned with the methods of identifying the genetic biomarkers of RA. Most of the researchers in the field of identifying RA biomarkers use single nucleotide polymorphism (SNP) approaches to express the significance of their results. Although, haplotype block methods are expected to play a complementary role in the future of that field. The used datasets belong to Egyptian population and North American population. Selection of the method used for the association discovery has a large impact on the resulted biomarkers. Finally, individual SNP approach and haplotype block methods should be applied side by side to discover valuable RA biomarkers.

Acknowledgments

First, thanks **ALLAH**, the most merciful and most gracious, who supplied me with power and faith to work on this very difficult topic.

I would like to express my appreciation and respect towards my advisor, **Prof. Dr. Ayman Mohamed Eldeib**, for his guidance has shaped this work and has presented me valuable skills, which I will take along with me throughout my academic journey. I would like also to thank him for his understanding, patience and enthusiasm.

This gratitude extends to the member of the advisory committee, **Prof. Dr. Olfat Gamil Shaker**, who I am honored to be one of her students. She has not been only providing me an experienced guidance; but she has also been giving me support and confidence to go on. Also, she generously was the provider of the Egyptian population dataset.

I am indeed indebted to my dear friendly advisor **Assoc. Prof. Dr. Mai Said Mabrouk**, who has put the corner stone of this work. She had helped me through scientific fruitful discussions, encouragement, and endless support. Also, she played major role in making the North American Rheumatoid Arthritis Consortium (NARAC) dataset available.

I would like to acknowledge the Genetic Analysis Workshop (GAW) grant (R01 GM031575) for providing the NARAC dataset. This work is based on data that was gathered with the support of grants from the National Institutes of Health (NO1-AR-2-2263, RO1-AR-44422), and the National Arthritis Foundation. NARAC dataset couldn't be available without the support of **Prof. Dr. Jean MacCluer**; Senior Scientist Emeritus at Texas Biomedical Research Institute, and **Mrs. Vanessa Olmo**; Independent Meeting & Event Planning Professional at Business Endeavors, San Antonio, Texas Area.

I also owe great thanks to **Prof. Dr. Hamdy Elrefai Agwa**; Professor, Botany Department, Faculty of Science, Kafr Elsheikh University, for his invaluable help on the introduction of the immune system.

Special thanks go to **Eng. Waleed El-Badry**; Assistant Lecturer in Mechatronics Engineering Department at MUST, for his support in software technical issues.

My deepest gratitude goes to my parents, **Nagy** and **Thanaa**. Every single achievement in my life belongs to them and whenever they should feel proud of me, they should rather be proud of themselves.

This work could not come to light without the patience, extreme help, and encouragement of my precious wife **Doha**. With her assist, I have overcome all the obstacles that I faced in my life. No words can describe her. Really, she is a gift from **ALLAH** to me.

I would also like to thank my sister **Zeinab** for being my lifelong best friend.

Dedication

This thesis is dedicated to the soul of the queen of my heart (my daughter) **Raghad**, may **ALLAH** bless her soul. Every word in this thesis was written with true love, pure love, deep love, renewable love, and endless love to **Raghad**. Every challenging work and achievement was applied to make her expected future happier, easier, and brighter.

Table of Contents

ACKNOWLEDGMENTS ..... **V**

DEDICATION ..... **VI**

TABLE OF CONTENTS ..... **VII**

NOMENCLATURE ..... **XII**

ABSTRACT ..... **XVI**

CHAPTER 1 INTRODUCTION **1**

1.1. OVERVIEW 1
1.2. PROBLEM DEFINITION 1
1.3. OBJECTIVES OF THE THESIS 2
1.4. ORGANIZATION OF THE THESIS 3

CHAPTER 2 BACKGROUND & LITERATURE REVIEW **5**

2.1. INTRODUCTION 5
2.2. IMMUNE SYSTEM 5
2.3. RHEUMATOID ARTHRITIS 6
2.4. SINGLE NUCLEOTIDE POLYMORPHISM 7
2.4.1. Types of SNPs 8
2.4.2. SNP density 9
2.4.3. Haplotype frequency 9
2.4.4. Minor allele frequency 9
2.4.5. Hardy Weinberg equilibrium 10
2.5. LINKAGE DISEQUILIBRIUM 10
2.6. GENETIC ASSOCIATION STUDY 12
2.7. BIG DATA 12
2.8. HAPLOTYPE BLOCK PROJECTS 13
2.8.1. The HapMap project 14
2.8.2. The 1000 genomes project 14
2.8.3. Haplotype block partitioning 14
2.9. RA BIOMARKERS 14
2.9.1. Biomarkers in the MHC region and Chr 6 15
2.9.2. Biomarkers outside MHC region 18
2.9.3. RA biomarkers in children 20
2.9.4. Gender specific biomarkers 21
2.9.5. Biomarkers in different populations 22
2.10. HAPLOTYPE BLOCKS VS INDIVIDUAL SNPS 23
2.11. HAPLOTYPE BLOCKS IN RA BIOMARKER DISCOVERY 24
2.12. STUDIES THAT AGREE/DISAGREE WITH OTHERS 26
2.13. SUMMARY 29

CHAPTER 3 METHODS AND MATERIALS **31**

3.1. INTRODUCTION 31
3.2. METHODS 31
3.2.1. Individual SNP approaches 31
3.2.1.1. Odds ratio 32
3.2.1.2. Pearson chi square test 32
3.2.2. Haplotype block methods 33
3.2.2.1. Four gamete test 33
3.2.2.2. Confidence interval test 35
3.2.2.3. Solid spine of linkage disequilibrium 36
..... 36
3.3. SYSTEM OVERVIEW 37
3.3.1. Egyptian dataset 37
3.3.2. NARAC dataset 39
3.4. MATERIALS 40
3.5. SUMMARY 41

CHAPTER 4 RESULTS 45

4.1. INTRODUCTION 45
4.2. EGYPTIAN DATASET RESULTS 45
4.2.1. RA susceptibility results 45
4.2.2. RA severity results 49
4.3. NARAC DATASET RESULTS 52
4.4. SUMMARY 55

CHAPTER 5 DISCUSSIONS 61

5.1. INTRODUCTION 61
5.2. EGYPTIAN DATASET DISCUSSIONS 61
5.2.1. RA susceptibility discussions 61
5.2.2. RA severity discussions 65
5.3. NARAC DATASET DISCUSSIONS 66
5.4. SUMMARY 68

CHAPTER 6 CONCLUSION AND FUTURE WORK 71

6.1. CONCLUSION 71
6.1.1. Egyptian dataset conclusions 71
6.1.1.1. RA susceptibility 71
6.1.1.2. RA severity 71
6.1.2. NARAC dataset conclusions 72
6.2. FUTURE WORK 72

REFERENCES 73

APPENDIX: ADDITIONAL NARAC DATASET RESULTS 85

List of Tables

Table 2.1: Haplotype frequencies for a pair of SNPs 11
Table 2.2: Allele frequencies relative to haplotype frequencies 11
Table 2.3: The relationship between the haplotype frequencies, allele frequencies, and (D) 11
Table 2.4: Detected SNPs associated with RA susceptibility 27
Table 2.5: Studies that agree/disagree with other studies 29
Table 3.1: The exposed group and the unexposed group in the individual SNP approaches 32
Table 3.2: The 2*2 contingency table for observed and expected values 33
Table 3.3: Example on recombination events with four haplotypes and four SNPs 35
Table 3.4: Example on SSLD 36
Table 3.5: A comparison among haplotype block definitions 36
Table 4.1: The biomarker checks of the studied SNPs 46
Table 4.2: Case-control RA susceptibility study - SNP analysis 47
Table 4.3: Case-control RA severity study – SNP analysis 50
Table 4.4: The processing time for each performed step 56
Table 4.5: Results of the individual SNP approach 57
Table 5.1: Association status for RA susceptibility of our study and previous studies . 62
Table 5.2: Association status for RA severity of our study and previous studies 65
Table 5.3: The ability of each haplotype block method in representing the significant SNPs of the individual SNP approach 68
Table A.1: Comparison among haplotype block methods results for Chr one 85
Table A.2: Comparison among haplotype block methods results for Chr two 85
Table A.3: Comparison among haplotype block methods results for Chr three 86
Table A.4: Comparison among haplotype block methods results for Chr four 86
Table A.5: Comparison among haplotype block methods results for Chr five 87
Table A.6: Comparison among haplotype block methods results for Chr six 87
Table A.7: Comparison among haplotype block methods results for Chr seven 88
Table A.8: Comparison among haplotype block methods results for Chr eight 88
Table A.9: Comparison among haplotype block methods results for Chr nine 89
Table A.10: Comparison among haplotype block methods results for Chr ten 89
Table A.11: Comparison among haplotype block methods results for Chr eleven 90
Table A.12: Comparison among haplotype block methods results for Chr twelve 90
Table A.13: Comparison among haplotype block methods results for Chr thirteen 91
Table A.14: Comparison among haplotype block methods results for Chr fourteen 91
Table A.15: Comparison among haplotype block methods results for Chr fifteen 92
Table A.16: Comparison among haplotype block methods results for Chr sixteen 92
Table A.17: Comparison among haplotype block methods results for Chr seventeen .. 93
Table A.18: Comparison among haplotype block methods results for Chr eighteen 93
Table A.19: Comparison among haplotype block methods results for Chr nineteen 94
Table A.20: Comparison among haplotype block methods results for Chr twenty 94
Table A.21: Comparison among haplotype block methods results for Chr twenty one . 95
Table A.22: Comparison among haplotype block methods results for Chr twenty two 95
Table A.23: Number of significant SNPs discovered by the three haplotype block methods compared to that of the four methods 96
Table A.24: SNP IDs that were uniquely identified by each method 97
Table A.25: Block similarity among the used haplotype block methods for the 22 Chrs 101

List of Figures

Figure 2.1: Four different types of antigens showing the binding between the epitope and the antibody [12] 6
Figure 2.2: Hand with inflammations due to RA [13] 7
Figure 2.3: Three persons' DNA with highlighted SNP in the fourth position [18] 8
Figure 2.4: Types of SNPs 9
Figure 2.5: One SNP per 30 bp detected in 2 double strand DNA sequences [25] 9
Figure 2.6: Summary of the genetic association study [41] 12
Figure 2.7: A representation of the size and complexity of big data [43] 13
Figure 2.8: More than 35 risk loci that have been previously identified as biomarkers for RA disease [50] 15
Figure 2.9: Classification of RA biomarkers depending on different categories 15
Figure 2.10: Demonstration of the MHC region [51] 16
Figure 2.11: LD structure for 12 SNPs at HLA–DRB1 and HLA– DPB1. A constructed block was shown including eight SNPs, from SNP 5 to SNP 12 [52] 17
Figure 2.12: Case-control association results at 6q23. The associated SNP (rs10499194) was about 165 kb from both TNFAIP3 and OLIG3 genes [79] 18
Figure 2.13: Manhattan plots of the GWAS meta-analysis for RA in the Japanese population [92] 21
Figure 2.14: Manhattan plot of association with RA in the European descents [99] 23
Figure 2.15: Summarization of the "Background and Literature Review" chapter 30
Figure 3.1: Block diagram of the used methods 31
Figure 3.2: Crossover results in recombination event [141] 34
Figure 3.3: The FGT is conducted between pairwise loci, and (0) and (1) denote the absence and presence, respectively, of all four gametes between locus pairs [125] 34
Figure 3.4: The CIT showing the definition of strong LD and weak LD 35
Figure 3.5: Morphological structure of SSLD at the legs of the LD chart 36
Figure 3.6: System overview for the whole study 38
Figure 3.7: A snapshot from the Egyptian dataset showing 10 controls with their corresponding 8 SNPs 39
Figure 3.8: A snapshot from the NARAC dataset showing 10 samples with their corresponding 3 SNPs 40
Figure 3.9: Flow chart of the proposed study for the Egyptian dataset 42
Figure 3.10: Summary of the proposed system for the NARAC dataset 43
Figure 4.1: Association analysis for examined SNPs with RA susceptibility. (a) Multiplicative model. (b) Dominant model. (c) Recessive model. (d) Co-dominant model. The horizontal line in each model represents the significance level of the P-value (0.01) 48
Figure 4.2: Genotype distributions in RA patients and controls 48
Figure 4.3: Pairwise LD plot for the VDR SNPs. The rs731236, rs7975232, rs1544410, and rs2228570 correspond to TaqI, ApaI, BsmI, and FokI respectively. (a) D′ values. (b) r^2 values 49
Figure 4.4: Association analysis for examined SNPs with RA severity. (a) Multiplicative model. (b) Dominant model. (c) Recessive model. (d) Co-dominant model. The horizontal line in each model represents the significance level of the P-value (0.05) 51
Figure 4.5: Genotype distributions in RA patients with and without OP 51
Figure 4.6: Pairwise LD plot for examined VDR SNPs. (a) D′ values. (b) r^2 values 52

Figure 4.7: The total no. of blocks resulted from each haplotype block method 53
Figure 4.8: The mean no. of SNPs per each block - either general or significant block for each method 53
Figure 4.9: Median block size in bp for each haplotype block method 54
Figure 4.10: No. of RA biomarkers detected by each method – all or only detected by the method 54
Figure 4.11: Manhattan plot for the associations between genotyped SNPs (Chr 6 of NARAC dataset) and RA susceptibility using individual SNP approach 58
Figure 4.12: Manhattan plot for the associations between the whole NARAC SNPs and RA susceptibility using individual SNP approach 59
Figure 5.1: The overlapping of the significant blocks resulted from the three haplotype block methods.....69

Nomenclature

A	Adenine
ACPA	Anti Citrullinated Protein Antibody
ACR	American College of Rheumatology
AFF3	AF4/FMR2 Family, Member 3
anti-CCP+	Anti-Cyclic Citrullinated Peptide Positive
ANXA3	Annexin A3
ARAP1	ArfGAP With RhoGAP Domain, Ankyrin Repeat And PH Domain 1
ARID5B	AT Rich Interactive Domain 5B
B3GNT2	UDP-GlcNAc: Beta-1,3-N-Acetylglucosaminyltransferase 2
BLK	B Lymphocyte Kinase
BMD	Bone Mineral Density
Bp	Base Pair
BRASS	Brigham RA Sequential Study
C	Cytosine
C5orf30	Chromosome 5 Open Reading Frame 30
CCL21	Chemokine (C-C Motif) Ligand 21
CCR6	Chemokine (C-C Motif) Receptor 6
CD	Celiac Disease
CD40	Cluster of Differentiation 40
CDK6	Cyclin-Dependent Protein Kinase 6
Chr	Chromosome
CIT	Confidence Interval Test
cM	Centi Morgan
CORA	Childhood Onset of RA
CSF2	Colony Stimulating Factor 2
CTLA4	Cytotoxic T-Lymphocyte Antigen 4
D	Deviation
D′	Normalized Deviation
DAS28	Disease Activity Score in 28 Joints
DC-SIGN	Dendritic Cell-Specific Intercellular Adhesion Molecule-3-Grabbing Non-Integrin

DDX6	DEAD (Asp-Glu-Ala-Asp) Box Helicase 6
DF	Degree of Freedom
DNA	Deoxyribose Nucleic Acid
DMARDs	Disease-Modifying Anti-Rheumatic Drugs
EIRA	Swedish Epidemiological Investigation of RA
F12	Coagulation Factor 12
FCGR2A	Fc Fragment of IgG, Low Affinity IIa
FGT	Four Gamete Test
FHS	Framingham Heart Study
G	Guanine
GAW	Genetic Analysis Workshop
GCI	Genomics Collaborative Initiative
GENRA	Genetics Network Rheumatology Amsterdam
GPA	Granulomatosis with Polyangiitis
GRAIL	Gene Relationships Across Implicated Loci
GSDMB	Gasdermin B
GWAS	Genome Wide Association Study
HLA	Human Leukocyte Antigen
HSPA1L	Heat Shock 70kDa Protein 1-Like
HWE	Hardy-Weinberg Equilibrium
IFNγ	Interferon Gamma
IGV	Integrative Genomics Viewer
IKZF3	IKAROS Family Zinc Finger 3
IL-1	Interleukin 1 Family
IL2RA	Interleukin 2 Receptor, Alpha
IL6ST	Interleukin 6 Signal Transducer
IRF5	Interferon Regulatory Factor 5
JRA	Juvenile RA
Kb	Kilo Base Pair
KCNB1	Potassium Voltage-Gated Channel, Shab-Related Subfamily, Member 1
KIF5A	Kinesin Family Member 5A
LD	Linkage Disequilibrium
LUMC	Leiden University Medical Center

MAF	Minor Allele Frequency
Mb	Mega Base Pair
MCP-1	Monocyte Chemotactic Protein-1
MHC	Major Histocompatibility Complex
MMEL1	Membrane Metallo-Endopeptidase-Like 1
mRNA	Messenger Ribonucleic Acid
MTHFR	Methylenetetrahydrofolate Reductase
MyEIRA	Malaysian Epidemiological Investigation of RA
NARAC	North American Rheumatoid Arthritis Consortium
NF-kB	Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells
NFKBIE	NF-kB Inhibitor, Epsilon
NHS	Nurses Health Study
NICRAP	National Inception Cohort of RA Patients
NK	Natural Killers
NKAPL	NF-kB Activating Protein-Like
ObsHET	Observed Heterozygosity
OLIG3	Oligodendrocyte Transcription Factor 3
OP	Osteoporosis
OR	Odds Ratio
ORMDL3	ORM1-Like 3
PADI4	Peptidyl-Arginine Deiminase Type 4
PDE2A	Phosphodiesterase 2A
PHF19	PHD Finger Protein 19
PIP4K2C	Phosphatidylinositol-5-Phosphate 4-Kinase, Type II, Gamma
PLD4	Phospholipase D Family, Member 4
PRDM1	PR Domain Zinc Finger Protein 1
PredHET	Predicted Heterozygosity
PRKCQ	Protein Kinase C Theta Type
PTPN22	Protein Tyrosine Phosphatase Non-Receptor Type 22
PTPRC	Protein Tyrosine Phosphatase Receptor Type C
PXK	PX Domain Containing Serine/Threonine Kinase
r^2	Correlation Coefficient
RA	Rheumatoid Arthritis
RAG1	Recombination Activating Gene 1

RBPJ	Recombination Signal Binding Protein For Immunoglobulin Kappa J Region
REL	Reticuloendotheliosis
RF	Rheumatoid Factor
SD	Standard Deviation
SE	Shared Epitope
SH2B3	SH2B Adaptor Protein 3
SLE	Systemic Lupus Erythematosus
SNP	Single Nucleotide Polymorphism
SONORA	Study of New Onset RA
SPRED2	Sprouty-Related, EVH1 Domain Containing 2
SSLD	Solid Spine of LD
STAT4	Signal Transducer And Activator of Transcription 4
T	Thymine
TAGAP	T Cell Activation RhoGTPase Activating Protein
TGFβ1	Transforming Growth Factor Beta 1
Th1	T Helper 1 Cells
TNFα	Tumor Necrosis Factor Alpha
TNFB	Tumor Necrosis Factor Beta
TNFAIP3	TNF Alpha-Induced Protein 3
TNFR	TNF Receptor
TNFRSF14	TNFR Superfamily Member 14
TRAF1-C5	TNF Receptor-Associated Factor 1 – Complement Component 5
UBASH3A	Ubiquitin Associated And SH3 Domain Containing A
UBE2L3	Ubiquitin-Conjugating Enzyme E2L3
UKRAG	United Kingdom RA Genetics
USNHGRI	US National Human Genome Research Institute
VDR	Vitamin D Receptor
WRDDB	Wichita Rheumatic Disease Data Bank
WTCCC	Wellcome Trust Case Control Consortium
ZNF	Zinc Finger Protein
χ^2	Pearson Chi Square Test