PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
SCHOOL OF MEDICINE
GRADUATE PROGRAM IN HEALTH SCIENCES

MONICA ELIZABETH DALLMANN SAUER

WHOLE GENOME SEQUENCING OF A FAMILY WITH MONOZYGOTIC TWINS DISPLAYING EARLY-ONSET LEPROSY

CURITIBA
2016

MONICA ELIZABETH DALLMANN SAUER

WHOLE GENOME SEQUENCING OF A FAMILY WITH MONOZYGOTIC TWINS DISPLAYING EARLY-ONSET LEPROSY

Thesis submitted to the Graduate Program in Health Sciences (in Portuguese, *Programa de Pós Graduação em Ciências da Saúde*, PPGCS) of the School of Medicine at Pontifícia Universidade Católica do Paraná (PUCPR) as part and in conformity with the requirements for the degree of Doctor in Health Sciences, Area of Concentration Medicine and Related Sciences.

Supervisor: Marcelo Távora Mira, Ph.D.

Co-supervisor: Christian Macagnan Probst, Ph.D.

International co-supervisor: Erwin Schurr, Ph.D.

CURITIBA

2016

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

V899c
2016

Sauer, Monica Elizabeth Dallmann
Whole genome sequencing of a family with monozygotic twins displaying early-onset leprosy / Monica Elizabeth Dallmann Sauer ; supervisor, Marcelo Távora Mira ; co-supervisors, Christian Macagnan Probst ; Erwin Schurr. – 2016.
xviii, [109] f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2016
Inclui bibliografias

1. Hanseníase. 2. Genoma. 3. Gêmeos monozigóticos.4. Ciências médicas. I. Mira, Marcelo Távora. II. Probst, Christian Macagnan. III. Schurr, Erwin. IV. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Ciências da Saúde. V. Título.

CDD 22. ed. – 610

Pontifícia Universidade Católica do Paraná
Escola de Medicina
Programa de Pós-Graduação em Ciências da Saúde - *Stricto Sensu*

PUBLIC SESSION OF THESIS EXAMINATION FROM THE GRADUATE PROGRAM IN HEALTH SCIENCES, (DOCTORATE DEGREE), PONTIFICIAL CATHOLIC UNIVERSITY OF PARANA.

On November 24, 2016 took place at the Center for Health Sciences the public session of thesis examination entitled, "**Whole genome sequencing of a family with monozygotic twins displaying early-onset leprosy**" presented by **Monica Elizabeth Dallmann Sauer**, candidate for a doctor degree in Health Science

The Board of examiners was composed for the following members:

MEMBERS OF THE BOARD	SIGNATURE
Prof. Dr. Marcelo Távora Mira (PUCPR) – President	
Prof. Dr. Salmo Raskin – (PUCPR)	
Profª. Drª. Vanessa Santos Sotomaior (PUCPR)	
Prof. Dr. Erwin Schurr (McGill University)	
Profª. Drª. Maria Luiza Petzl-Erler (UFPR)	

In accoardance with the program´s regulations, the Board of Examiners presented their evaluation, wich were the following

Prof. Dr. Marcelo Távora Mira — Evaluation: Aprovado
Prof. Dr. Salmo Raskin — Evaluation: Aprovado com Louvor
Profª. Drª. Vanessa Santos Sotomaior — Evaluation: Approved
Prof. Dr. Erwin Schurr — Evaluation: Approved
Profª. Drª. Maria Luiza Petzl-Erler — Evaluation: Approved

Parecer Final: Aprovado

Observações da Banca Examinadora:

Approved with Honors.

Prof. Dr. Marcelo Távora Mira
President of the Examining Bank

Prof. Dr. Roberto Flavio Silva Pecoits -Filho
Coordinator of PPGCS PUCPR

ACKNOWLEDGMENTS

SCIENTIFIC CREW

My main acknowledgment goes to **Dr. Marcelo Távora Mira** for accepting me as his student and allowing me to endeavour on science wonders. Thank you very much boss!

I am very grateful to **Dr. Erwin Schurr** for receiving me at his lab as a graduate research trainee. His insightful feedback and advices contributed greatly for my work's advance.

Dr. Christian Macagnan Probst had a very valuable contribution on this study by sharing his lab and guidance which allowed me to produce my first set of data. Many thanks Dr. Probst.

My research would not have been possible without the contribution of **Dr. Ana Lúcia França da Costa**. Her collaboration allowed us to contact the studied family and to accompany them since the beginning of this study. Thank you Dr. Ana!

My sincere gratitude to my colleagues and friends from the three labs that I've worked on throughout my Ph.D. (past and present; in alphabetic order):

- From Dr. Marcelo's lab (at the Multiuser Experimental Laboratory, PUCPR – Curitiba): **Alana Mazzetti, Alessandro Afornali, Alice Lichs Marssaro, Ana Paula Wasilewski da Silva, Angela Schnider Francio, Bruna Loss, Caio Cesar Silva de Castro, Elaine Doff Sotta, Geison Eduardo Cambri, Geovana Brotto Ramos, Helena Regina Salomé D'Espindula, Heloisa Salomão, Irenice Cairo, Juliane Monteiro, Laysa Toschi Martins, Liliane Machado do Nascimento, Marcelo Pilonetto, Maria Eduarda Krauss, Nathália Cristina de Oliveira Cavazzani, Priscila Uaska Sartori, Rafael Saraiva de Andrade, Renata Helena Sindeaux, Renata Iani Werneck, Roberto Gomes Tarlé, Sérgio Eduardo Fontoura da Silva** and **Wilian Corrêa de Macedo**. Thank you to all of them for their constant

support, advice, encouragement and most importantly their boundless enthusiasm. I really enjoyed spending all these years with them and I've learnt a lot with all the team work that we performed together.

- From Dr. Erwin's lab (at McGill Health Center, McGill University – Montreal): **Jeremy Manry, Marianna Orlova, Vinicius Medeiros Fava, Wilian Corrêa de Macedo** and **Yong Zhong Xu**. Their help and continuous support were priceless and I am very grateful for that.
- From Dr. Christian's lab (at Genomics and Bioinformatics laboratories, Carlos Chagas Institute – Curitiba): Thanks to **Fabiana Poerner** and **Paulo Rodrigo Claure Arauco** for providing crucial technical assistance with exome sequencing experiments.

A special thanks to the family that accepted to participate as **studied subjects** in this research.

Moreover, I would like to thank **Carlos Alves** from Immunogenetics and histocompatibility lab at *Associação Paranaense de Cultura*; **Dr. Cleber Machado de Souza** from School of Life Sciences; **Jane Fábia Domênica Pulowsky**, **Dr. Roberto Hirochi Herai** and **Dr. Vanessa Santos Sotomaior** from PPGCS at PUCPR, as well as **Dr. Aurélie Cobat** from Laboratory of Human Genetics of Infectious Diseases, Necker Hospital for Sick Children, Paris, France.

Finally, I am grateful for the financial support received from **CNPq, CAPES** and **Fundação Araucária**.

NON-SCIENTIFIC CREW

I am very grateful – and dedicate this thesis – to my family: to my lovely husband **Wilian Corrêa de Macedo;** to my sister and best friend **Pamela Yaqueline Dallmann Sauer**; and to my dear parents **Wali Sauer Greve** and **Federico Manuel Dallmann**.

I would like to thank my cousin **Diana Elizabeth Dallmann Schroetlin** and her husband **Hugo Melgar-Quiñonez** for their support during my internship in Canada. We enjoyed so much to share this experience with you guys!

My sincere gratitude to **Irenice Cairo** for her constant support during my PhD, not only while I was in Brazil, but also during my internship abroad. Thank you very much Irê!

A special thanks to my parents-in-law **Marlene Salvador Corrêa** and **João de Macedo** for their support, as well as to my aunt **Daisy Scavasin Fernandes and family** – **Gabryelle, Amanda** and **Givaldo** – for always being there for me.

Moreover, I am very grateful to my aunt **Patricia Nichols** for giving me the opportunity to initiate my academic career in Genetics.

Thank you to all my **relatives** and **friends**, near and far, for everything you've done for me. I'm so lucky to have such a wonderful group of friends and family!

ABSTRACT

Despite the availability of effective treatment, leprosy (caused by *Mycobacterium leprae*) affects over 200,000 new patients every year. Leprosy is an infectious disease with long incubation period; hence, young cases of leprosy are rare and patients with age-of-onset < 4 years represent an extreme end of the age distribution. Here, we studied monozygotic twin girls exhibiting clinical symptoms of leprosy at the age of 22 months old. The early onset of leprosy in the girls raised the possibility of a strong genetic effect controlling leprosy mechanisms. Moreover, leprosy was present in three generations of the twins' family revealing familial aggregation of cases that also supported a shared genetic component for disease susceptibility. To investigate a possible genetic contribution to leprosy in this family, we obtained genomic DNA and performed whole exome and genome sequencing in four leprosy affected individuals and one unaffected family member. Output data were processed for variant discovery using bioinformatics pipelines for next generation sequencing data analysis. On average, 26,281 exonic and splice-site variants were identified *per* subject. To narrow down the variants list, stepwise procedures of filtering and variant prioritization were applied to identify those variants that are most likely to be causal. As result, 95 variants with minor allele frequency < 30% and 37 novel functional variants were identified as candidate when assuming a recessive and dominant model, respectively. Among these, two predicted protein-damaging variants are located in *LRRK2* gene (both found in the recessive model). Interestingly, variants in this gene have previously been associated with susceptibility to leprosy phenotypes as well with Parkinson and Crohn's disease. To better understand the role of *LRRK2* variants on leprosy host genetic control, functional studies are being currently designed and performed to validate our findings in the studied family.

Keywords: Leprosy susceptibility; whole exome sequencing; whole genome sequencing; monozygotic twins; early-onset leprosy.

RESUMO

Apesar da disponibilidade de um tratamento eficaz, a hanseníase (causada pelo *Mycobacterium leprae*) afeta mais de 200.000 novos pacientes todos os anos. Esta é uma doença infecciosa com um longo período de incubação; assim, casos de hanseníase são mais frequentes em adultos e pacientes com menos de 4 anos são raros e representam um extremo da distribuição etária de casos. Aqui, nós estudamos uma família contendo um par de gêmeas monizigóticas que apresentaram sintomas clínicos de hanseníase com 22 meses de idade. O início precoce da hanseníase nas meninas levantou a possibilidade de um forte componente genético controlando a doença. Além disso, a hanseníase estava presente em três gerações da família das gêmeas revelando agregação familial de casos que também contribuem com a hipótese da presença de um forte componente genético de suscetibilidade à doença nesta família. Para investigar esta hipótese, obteve-se DNA genômico e efetuou-se sequenciamento de exoma e genoma completo em quatro indivíduos afetados e um membro não afetado da família. Os dados de sequenciamento foram processados para identificação de variants genéticas usando análise bioinformática para dados de sequenciamento de próxima geração. Em média, foram identificadas 26.281 variantes exônicas e de sítios de *splicing* por amostra. Para restringir a lista de variantes, procedimentos de filtragem e priorização de variantes foram aplicados para identificar aquelas que têm maior probabilidade de serem causais. Como resultado, 95 variantes candidatas *missense* com frequência alélica < 30% e 37 variantes funcionais novas foram identificadas assumindo-se um modelo recessivo e dominante, respectivamente. Dentre estas, duas variantes com previsão de impacto funcional em proteína foram identificados no gene *LRRK2* (ambas encontradas no modelo recessivo). Curiosamente, variantes neste gene foram previamente associadas à susceptibilidade a fenótipos da hanseníase, bem como à doença de Parkinson e à síndrome de Crohn. Para entender melhor o papel das variantes de *LRRK2* na susceptibilidade à hanseníase, estudos funcionais para validar o nosso achado na família estudada estão em andamento.

Palavras-chave: susceptibilidade à hanseníase; sequenciamento de exoma completo; sequenciamento de genoma completo; gêmeos monozigóticos; hanseníase de início precoce.

CONTRIBUTION OF AUTHORS AND FUNDING

Supervision: Dr. Marcelo Távora Mira directed the research and supervised Monica E. Dallmann Sauer's (MEDS) activities throughout the whole project. The student's activities performed at Instituto Carlos Chagas – Oswaldo Cruz Fundation (ICC/Fiocruz) were supervised by Dr. Christian Macagnan Probst. All activities conducted at McGill Health Center were supervised by Dr. Erwin Schurr.

Sample collection and DNA extraction: Blood sampling was done by Dr. Ana Lúcia França da Costa at the families' households at Teresina, Brazil. DNA extraction was performed by MEDS at Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Brazil.

Exome sequencing: MEDS performed exome-sequencing experiments at ICC, Curitiba; and exome data analysis at McGill Health Center, Montreal, Canada; and at PUCPR.

Genome sequencing: Whole genome sequencing experiments were conducted by the staff at the Rockefeller University Genomics Resource Center, New York. Complete data analysis was performed by MEDS at Research Institute of the McGill University Health Centre (RI-MUHC) and at PUCPR.

Sanger sequencing: Primer design, PCR amplifications and agarose electrophoresis were performed by MEDS at RI-MUHC. PCR products were sent to Génome Québec Innovation Centre, Montreal, where Sanger sequencing was performed by lab staff. Data analysis for the validation sequencing was performed by MEDS at RI-MUHC.

Funding: This project was funded by Universal/2011 grant from CNPq (National Counsel of Technological and Scientific Development). MEDS internship at RI-MUHC was funded by 3407/15-2 grant from CAPES (Coordination for the Improvement of Higher Level Personnel).

Publishing: Publication strategy is to combine genomic and functional data in one single, high impact manuscript. Functional experiments are ongoing.

LIST OF ABBREVIATIONS

1000G	1000 Genome Consortium Project
A	Alanine
AA	Amino acid
ABCC6	*ATP binding cassette subfamily C member 6*
ACP5	*Acid phosphatase 5, tartrate resistant*
ADAL	*Adenosine deaminase like*
AFR	African/African American
AHNAK2	*AHNAK nucleoprotein 2*
Alt	Alternative allele
AMR	Admixed American/Latin
ANK	Ankyrin domain
APC	*APC, WNT signaling pathway regulator*
Apr	April
ARM	Armadillo domain
ASTN1	*Astrotactin 1*
ATAD5	*ATPase family, AAA domain containing 5*
ATP	Adenosine triphosphate
Aug	August
B	Benign (PolyPhen-2 prediction)
BATF3	*Basic leucine zipper ATF-like transcription factor 3*
BB	Borderline-borderline leprosy
BC	Before Christ
BCG	Bacillus Calmette-Guérin
BCKDHA	*Branched chain keto acid dehydrogenase E1, alpha polypeptide*
BED	Browser Extensible Data
BL	Borderline-lepromatous leprosy
bp	Base pairs
BT	Borderline tuberculoid leprosy
BWA	Burrows-Wheeler aligner
C	Cysteine
C1orf167	*Chromosome 1 open reading frame 167*
CADD	Combined annotation–dependent depletion
CAPES	Coordination for the Improvement of Higher Education Personnel (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*)
CCDC122	*Coiled-coil domain containing 122*
CCDC141	*Coiled-coil domain containing 141*
CCDC34	*Coiled-coil domain containing 34*
CCDC88B	*Coiled-coil domain containing 88B*

CCDS	Consensus coding sequence
CCPG1	*Cell cycle progression 1*
CD	Crohn's disease
CD68	*CD68 molecule*
CDC20B	*Cell division cycle 20B*
CDH20	*Cadherin 20*
Chr	Chromosome
CIITA	*Class II major histocompatibility complex transactivator*
CNPq	National Counsel of Technological and Scientific Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*)
Conc	Concordant
COR	C-terminal of Roc domain
CP	*Ceruloplasmin*
CR1	*Complement component 3b/4b receptor 1 (Knops blood group)*
CRELD2	*Cysteine rich with EGF like domains 2*
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats/ CRISPR associated protein 9
CUBN	*Cubilin*
D	Probably damaging (PolyPhen-2 prediction)
D	Aspartic acid
dbSNP	Database of Single Nucleotide Polymorphisms
DDX31	*DEAD-box helicase 31*
Dec	December
DHX33	*DEAH-box helicase 33*
DIDO1	*Death inducer-obliterator 1*
Disc	Discordant
DMSO	Dimethyl sulfoxide
DMXL1	*Dmx like 1*
DNA	Deoxyribonucleic acid
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide
DNV	Dinucleotide variant
dsDNA	double stranded *DNA*
DZ	Dizygotic twins
E	Glutamic acid
EAS	East Asian
emPCR	Emulsion PCR
eQTL	Expression quantitative trait *loci*
ERN1	*Endoplasmic reticulum to nucleus signaling 1*
EUR	European
ExAC	Exome Aggregation Consortium
EYS	*Eyes shut homolog (Drosophila)*

F	Phenylalanine
FAT1	*FAT atypical cadherin 1*
Feb	February
Fiocruz	Oswaldo Cruz Fundation (*Fundação Oswaldo Cruz*)
FN	False negative
FP	False positive
fs	Frameshift
FSIP2	*Fibrous sheath interacting protein 2*
G	Glycine
GATA3	*GATA binding protein 3*
GATK	Genome Analysis Toolkit
Gb	Giga bases
GCN1	*GCN1, eIF2 alpha kinase activator homolog*
GDI	Gene Damage Index
gDNA	Genomic DNA
GDP	Guanosine diphosphate
GPR179	*G protein-coupled receptor 179*
GQ	Genotype Quality score
GRCh37	Genome Reference Consortium Human genome build 37
GTP	Guanosine triphosphate
GWAS	Genome-wide association study
H	Histidine
HAUS5	*HAUS augmin like complex subunit 5*
hg19	Human genome 19
HLA	Human Leukocyte Antigen
HLA-A	*Major histocompatibility complex, class I, A*
HLA-C	*Major histocompatibility complex, class I, C*
HLA-DRB1	*Major histocompatibility complex, class II, DR beta 1*
HOXA7	*Homeobox A7*
HRH4	*Histamine receptor H4*
I	Indeterminate leprosy
I	Isoleucine
ICC	Carlos Chagas Institute (*Instituto Carlos Chagas*)
IFNG	*Interferon gamma*
IFN-γ	Interferon gamma
IGV	Integrative Genome Viewer
IL12	Interleukin 12
IL12B	*Interleukin 12B*
IL-12Rβ1	Interleukin 12 receptor subunit beta 1
IL18R1	*Interleukin 18 receptor 1*
IL18RAP	*Interleukin 18 receptor accessory protein*
IL23R	Interleukin 23 receptor

IL-6	Interleukin 6
Indel	Insertion/deletion
INTU	*Inturned planar cell polarity protein*
iPSC	Induced pluripotent stem cells
IQGAP3	*IQ motif containing GTPase activating protein 3*
ISFET	Ion-Sensitive field-effect transistor
ISOC2	*Isochorismatase domain containing 2*
ISP	Ion Sphere Particles
IUPAC	International Union of Pure and Applied Chemistry
Jun	June
K	Lysine
KIAA1217	*KIAA1217*
KL	*Klotho*
KLK8	*Kallikrein related peptidase 8*
L	Leucine
LACC1	*Laccase domain containing 1*
LD	Linkage desequilibium
LEMD3	*LEM domain containing 3*
LL	Lepromatous leprosy
LPS	Lipopolysaccharides
LRP1B	*LDL receptor related protein 1B*
LRR	Leucine-rich repeat domain
LRRC25	*Leucine rich repeat containing 25*
LRRC59	*Leucine rich repeat containing 59*
LRRK1	*Leucine rich repeat kinase 1*
LRRK2	*Leucine rich repeat kinase 2*
LTA	*Lymphotoxin alpha*
LYL1	*LYL1, basic helix-loop-helix family member*
M	Methionine
M.	*Mycobacterium*
MAF	Minor allele frequency
Mar	March
MB	Multibacillary leprosy
Mb	Mega bases
MCM9	*Minichromosome maintenance 9 homologous recombination repair factor*
MCP-1	Monocyte Chemoattractant Protein-1
MDT	Multidrug therapy
MHC	Major Histocompatibility Complex
Min	Minutes
MPDU1	*Mannose-P-dolichol utilization defect 1*
MRC1	*Mannose receptor, C type 1*

MSMD	Mendelian Susceptibility to Mycobacterial Disease
MUC17	Mucin 17, cell surface associated
MZ	Monozygotic twins
N	Asparagine
NBEAL2	*Neurobeachin like 2*
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NEBL	*Nebulette*
NFAT	Nuclear factor of activated T-cells
NGS	Next-generation sequencing
NOD2	*Nucleotide binding oligomerization domain containing 2*
NPHP4	*Nephrocystin 4*
NSFL1C	*NSFL1 cofactor*
NUP153	*Nucleoporin 153*
OTOP1	*Otopetrin 1*
P	Possibly damaging (PolyPhen-2 prediction)
P	Proline
PACRG	*PARK2 coregulated*
PARK2	*Parkin RBR E3 ubiquitin protein ligase*
PB	Paucibacillary leprosy
PCR	Polymerase chain reaction
PD	Parkinson disease
PDCD11	*Programmed cell death 11*
PKHD1	*Polycystic kidney and hepatic disease 1 (autosomal recessive)*
PLD2	*Phospholipase D2*
PolyPhen-2	Polymorphism phenotyping version 2
PPGCS	Graduate Program in Health Sciences (*Programa de Pós-Graduação em Ciências da Saúde*)
PPP4R2	*Protein phosphatase 4 regulatory subunit 2*
PRDM15	*PR/SET domain 15*
Prep	Preparation
PRIMPOL	*Primase and DNA directed polymerase*
PRKACB	*Protein kinase cAMP-activated catalytic subunit beta*
PUCPR	Pontifical Catholic University of Paraná (*Pontifícia Universidade Católica do Paraná*)
Q	Quality Phred score
Q	Glutamine
qPCR	Quantitative PCR
R	Arginine
R1	Read 1
R2	Read 2
RAB32	*RAB32, member RAS oncogene family*

RBBP6	*RB binding protein 6, ubiquitin ligase*
RBL2	*RB transcriptional corepressor like 2*
Ref	Reference allele
Rev	Revision
RGAG1	*Retrotransposon gag domain containing 1*
RI-MUHC	Research Institute of the McGill University Health Centre
RIPK2	*Receptor interacting serine/threonine kinase 2*
RNA	Ribonucleic acid
RNF39	*Ring finger protein 39*
RNH1	*Ribonuclease/angiogenin inhibitor 1*
Roc	Ras of complex proteins domain
ROS1	*ROS proto-oncogene 1, receptor tyrosine kinase*
S	Serine
S.	*Salmonella*
SALL4	*Spalt like transcription factor 4*
SAS	South Asian
SCAF1	*SR-related CTD associated factor 1*
Sec	Seconds
Sept	September
SGK223	*Homolog of rat pragma of Rnd2*
SIFT	Sorting intolerant from tolerant
SLC11A1	*Solute carrier family 11 member 1*
SLC17A9	*Solute carrier family 17 member 9*
SLC22A24	*Solute carrier family 22 member 24*
SLC25A25	*Solute carrier family 25 member 25*
SMPD3	*Sphingomyelin phosphodiesterase 3*
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variants
SOCS1	*Suppressor of cytokine signaling 1*
SOD2	*Superoxide dismutase 2, mitochondrial*
SORBS2	*Sorbin and SH3 domain containing 2*
SOS2	*SOS Ras/Rho guanine nucleotide exchange factor 2*
SOWAHB	*Sosondowah ankyrin repeat domain family member B*
STAB2	*Stabilin 2*
SUV39H1	*Suppressor of variegation 3-9 homolog 1*
SYNPO2	*Synaptopodin 2*
T	Threonine
T1R	Leprosy Type 1 Reaction
T2R	Leprosy Type 2 Reaction
TB	Tuberculosis
TCF20	*Transcription factor 20*
Th1	T-helper 1

Th2	T-helper 2
TINAG	*Tubulointerstitial nephritis antigen*
TLR1	*Toll like receptor 1*
TLR2	*Toll like receptor 2*
TLR4	*Toll like receptor 4*
TLR7	*Toll like receptor 7*
TMAP	Torrent Mapping Alignment Program for Ion Torrent
TNF	*Tumor necrosis factor*
TNFSF15	*Tumor necrosis factor superfamily member 15*
TNXB	*Tenascin XB*
TP53	*Tumor protein p53*
TREML2	*Triggering receptor expressed on myeloid cells like 2*
TRIP4	*Thyroid hormone receptor interactor 4*
TS	Torrent Suite
TT	Tuberculoid leprosy
TVC	Torrent variant caller
UCSC	University of California, Santa Cruz
UFPI	Federal University of Piauí (*Universidade Federal do Piauí*)
UGT3A1	*UDP glycosyltransferase family 3 member A1*
UTR	Untranslated region
v	Version
V	Valine
VCF	Variant Call Format
VEGA	Vertebrate Genome Annotation
VQSR	Variant Quality Score Recalibration
W	Tryptophan
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization
X	Stop-gain (AA change)
X	Times/folds (Read depth of coverage)
Y	Tyrosine
Yrs	Years
ZIM2	*Zinc finger imprinted 2*
ZNF185	*Zinc finger protein 185 (LIM domain)*
ZNF469	*Zinc finger protein 469*
ZNF480	*Zinc finger protein 480*
ZNF678	*Zinc finger protein 678*
ZZEF1	*Zinc finger ZZ-type and EF-hand domain containing 1*

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

1 INTRODUCTION

1.1 AN OVERVIEW OF LEPROSY

1.1.1 Historical overview and epidemiology

Leprosy is an infectious disease present throughout the history of mankind. A skeleton found in India with lesions characteristic of the disease, dated 2000 B.C., represents the oldest documented skeletal evidence for leprosy (1). It has been proposed that leprosy was originated in Africa and spread around the world following the routes of human migration (reviewed in (2)). For a long time, leprosy was believed to be result of a punishment from God, and patients were stigmatized as "unclean" (reviewed in (3)). For this reason, during the Middle Ages – a time when the prevalence of the disease reached its peak in Europe – carriers of the disease were forced to use characteristic clothing and, in some places, to carry a bell to indicate their arrival (reviewed in (4)). Only in 1873, Norwegian physician Gerhard Henrik Armauer Hansen identified leprosy's causative agent (5), the *Mycobacterium leprae* bacilli. Even after the discovery of a biological cause, leprosy patients were still stigmatized and marginalized. For a long time, affected individuals were forced to leave their cities and move out to colonies exclusively created for them, a strategy used to prevent disease spreading. In Brazil, the compulsory isolation of disease carriers was established in 1923 and isolation colonies, "*leprosários*", were created (reviewed in (2)).

In the 1940s, dapsone, the first effective antibiotic against *M. leprae*, was synthesized. Due to treatment effectiveness, Brazilian compulsory isolation law was officially abolished in 1963. However, in the 1960s, several cases of treatment-resistant bacilli were detected worldwide (reviewed in (2)). At the same time, two effective drugs against *M. leprae* – rifampicin and clofazimine – were synthesized. In 1981, the World Health Organization (WHO) implemented a multidrug therapy regimen

(MDT), composed of these three drugs (see section 1.1.3), which is worldwide offered for free by the WHO since 1995 (6). In 1986, the WHO presented the first proposal to eliminate leprosy as a public health problem by the year 2000. For that, in 1991, WHO created a resolution that established the goal for leprosy elimination: reduction of disease prevalence rate to less than one case per 10,000 individuals (7). As a result of global efforts and the effectiveness of MDT, there was a drastic reduction in leprosy prevalence and 98 countries reached the elimination goal by the year 2000 (6). However, elimination of leprosy as a public health problem has not been achieved in some endemic countries, in particular at a subnational level (8). Moreover, the transmission continues to occur and the number of new leprosy cases has been stable over the past eight years (9,10). Since 2000, WHO has organized four additional campaings for leprosy control worldwide, focused on early recognition of leprosy cases and prevention of permanent disabilities (8,11). Despite the progress in reducing the number of leprosy cases, it is still likely that the elimination will not be achieved at sub-national levels in some countries in near future.

According to the last WHO report, leprosy global prevalence for the first quarter of 2015 was 174,608 cases (8). Second worldwide in number of cases – after India – and first in the Americas, Brazil's prevalence is 23,995 leprosy cases, corresponding to 13.7% and 85.8% of global and American cases respectively (8). Brazilian prevalence rate – as for December 31th, 2015 – is 1.01 case *per* 10,000 habitants (12). However the disease is distributed unevenly across the country, with prevalence rates ranging from 0.1 in Rio Grande do Sul state to 7.75 in Mato Grosso state (13). As for leprosy incidence, a total of 210,758 new cases were reported worldwide in 2015 (8). In three countries – India, Brazil and Indonesia –, more than ten thousand new leprosy cases are reported each year (**Figure 1**). The new cases detected in these countries correspond to 81% of the global incidence (8). In Brazil, 26,395 new cases were reported to WHO in 2015 (8).

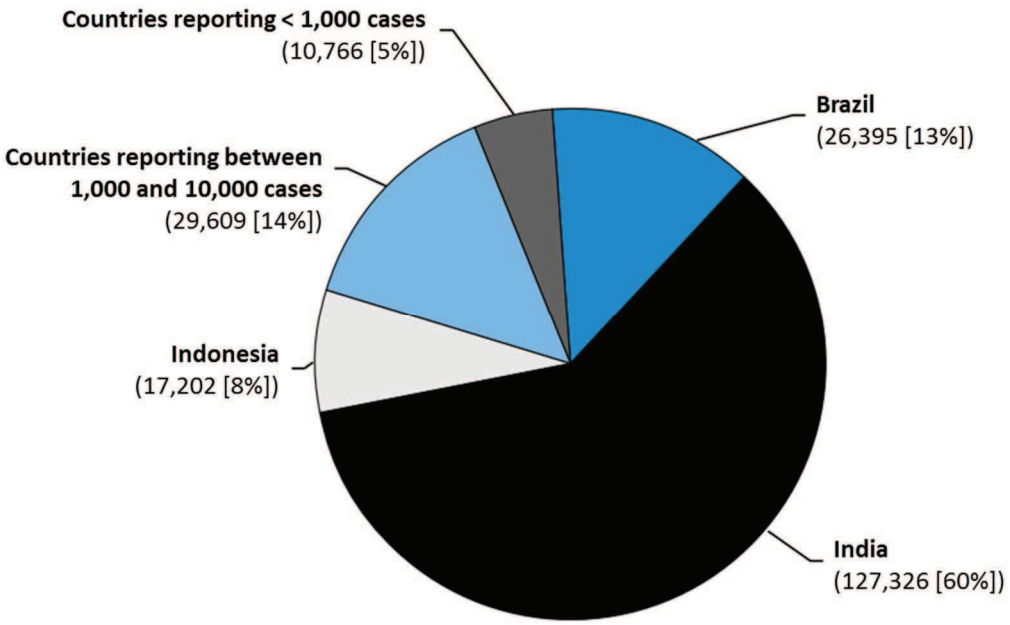


Figure 1. Distribution of leprosy new cases reported to WHO in 2015.
Source: Adapted from WHO, 2016 (8).

1.1.2 Etiological agent

M. leprae, leprosy etiological agent, is an obligate intracellular bacillus, non-culturable *in vitro*, that has tropism for Schwann cells on the peripheral nervous system and skin macrophages (reviewed in (14)). It is a straight or slightly curved bacillus, acid-resistant, which features red Ziehl-Neelsen staining and can be found isolated or grouped in globi (reviewed in (14,15)). Reproduction of *M. leprae* occurs by binary fission and its duplication time ranges from 12 to 14 days, making it the human pathogen with the longest known required time for duplication (reviewed in (15)). Another notable feature of *M. leprae* is its optimum temperature for survival and proliferation, which is between 30°C to 35°C. For this reason, cooler tissues – such as skin and peripheral nerves – are preferred targets for infection (reviewed in (2,15)).

In 2001, *M. leprae*'s 3.27 Mb genome was sequenced and compared with *Mycobacterium tuberculosis*' 4.4 Mb genome (16). While 90.8% of *M. tuberculosis'* genome comprises protein-coding genes with a total of 3,924 genes, *M. leprae* contains only 1,604 coding genes corresponding to 49.5% of its genome (16). This

indicates that *M. leprae* may have lost more than 2,000 genes after diverging from a common ancestor to *M. tuberculosis* (16). Moreover, *M. leprae*'s genome contains 1,116 pseudogenes (inactive reading frames with functional counterparts in other mycobacteria), while *M. tuberculosis* contains only 6 pseudogenes (16). Therefore, *M. leprae* is an extreme case of reductive evolution with elimination of many important metabolic pathways, maintaining a minimal gene set required to survive as an obligate intracellular parasite. Interestingly, genome comparisons of different *M. leprae* isolates obtained in different geographical regions and belonging to different periods of history has demonstrated low DNA variability, with a 99.995% genome identity among the evaluated strains (17).

1.1.3 Clinical forms and treatment

The disease manifests itself primarily through dermato-neurological signs and symptoms: skin lesions with decrease or loss of sensitivity and involvement of nerves with neural thickening (reviewed in (15)). Due to decreased sensitivity, unnoticed injury or burns in affected regions may lead to wounds and ulcers and the occurrence of infections with potential to lead to physical disabilities that may even develop into permanent deformities (reviewed in (2)). Leprosy diagnosis is based on the observation of one or more of the following signs: i) hypopigmented or erythematous skin lesion(s) with definite loss of sensation; ii) thickened peripheral nerve with loss of sensation; and iii) positive skin smear for acid-fast bacilli ((18), reviewed in (15)). The delay in diagnosis can have significant negative consequences, such as increasing the risk of permanent nerve damage and disease transmission (reviewed in (19)).

According to classic definition of Ridley and Jopling (20), clinical forms of leprosy are distributed in a spectrum with two extreme poles and three intermediate forms (**Figure 2**). Patients from the tuberculoid (TT) pole have well defined lesions, absence of bacilli in skin and nerves and predominantly Th1 immune response (cell-mediated immune response). Patients in the opposite, lepromatous (LL) pole, have multiple lesions, presence of bacilli in the skin and nerves and a predominance of Th2

type of immune response (humoral). Between the two poles, intermediate forms are defined as borderline tuberculoid (BT), borderline-borderline (BB) and borderline-lepromatous (BL). From BT to BL, there is a progressive reduction of cell-mediated immune response accompanied by an increase in the number of bacilli and skin and peripheral nerves lesions (**Figure 2**). Individuals who do not fit into this spectrum are classified as presenting the Indeterminate form (I) (20).

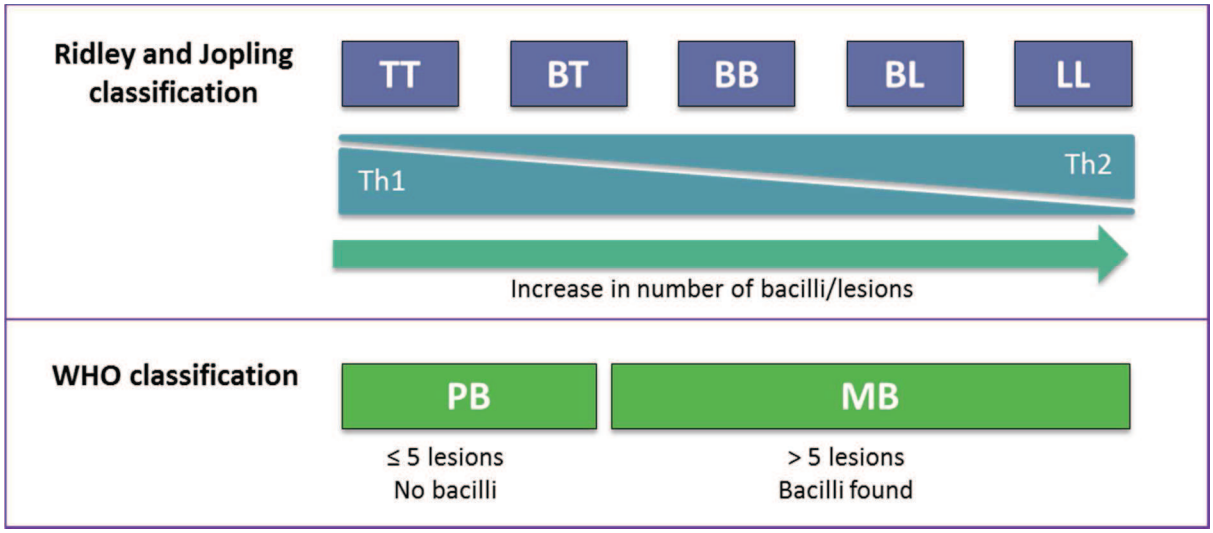


Figure 2. Leprosy clinical subtypes according to Ridley & Jopling and WHO classifications. BB: Borderline-borderline; BL: Borderline lepromatous; BT: Borderline tuberculoid; LL: Lepromatous leprosy; MB: Multibacillary leprosy; PB: Paucibacillary leprosy; Th1: T-helper 1; Th2: T-helper 2; TT: Tuberculoid leprosy; WHO: World Health Organization.
Source: Adapted from Fava *et al.* (21) and Sauer *et al.* (22).

A simplified classification system has been developed by WHO for operational purposes and therapeutic guidance (23). The WHO leprosy classification protocol distributes the disease into paucibacillary (PB) and multibacillary (MB) leprosy, which correspond approximately to TT and BT for the former and BB, BL and LL forms for the latter (**Figure 2**). This classification is based on the detection of bacilli – when available – and the number of lesions. Thus, individuals have PB leprosy when presenting negative smear and up to five lesions, whilst patients have MB leprosy when presenting smear-positive and/or more than five lesions (**Figure 2**) (23). Moreover, leprosy patients may develop severe nerve damage and pain as a result of the reactional states or leprosy reactions. These are sudden and intense inflammatory processes that may manifest itself along the course of leprosy, during and even years

after the completion of treatment (reviewed in (14)). These reactional states are classified as type 1 reaction (T1R), which commonly affect borderline patients, or type 2 reaction (T2R), which affects lepromatous patients. Comprehensive reviews of leprosy reactional states can be found in (14,21).

As for leprosy treatment, WHO's standard MDT consists of a combination of rifampicin, clofazimine and dapsone for MB leprosy patients for a period of 12 months and rifampicin and dapsone for PB leprosy patients for six months (Table 1) (18,24). Rifampicin is the main therapeutic agent, therefore it is present in both treatment regimens with supervised administration (18,24). For pediatric leprosy of patients under 15 years old, doses are set according to the weight and age of the patient group (**Table 1**) (18).

Table 1. Leprosy multidrug therapy.

Age	Leprosy clinical form	Multidrug therapy (MDT)			
		Drug	Supervised administration	Self-administration	Duration
Children > 6 years old	MB	Rifampicin	150-300 mg/month		12 month
		Clofazimine	100 mg/month	100 mg/week	
		Dapsone	25 mg/month	25 mg/day	
	PB	Rifampicin	150-300 mg/month		6 month
		Dapsone	25 mg/month	25 mg/day	
Children from 6 to 14 years old	MB	Rifampicin	300-450 mg/month		12 month
		Clofazimine	150-200 mg/month	150 mg/ week	
		Dapsone	50-100 mg/month	50-100 mg/day	
	PB	Rifampicin	300-450 mg/month		6 month
		Dapsone	50-100 mg/month	50-100 mg/day	
Adult (≥ 15 years old)	MB	Rifampicin	600 mg/month		12 month
		Clofazimine	300 mg/month	50 mg/day	
		Dapsone	100 mg/month	100 mg/day	
	PB	Rifampicin	600 mg/month		6 month
		Dapsone	100 mg/month	100 mg/day	

MB: Multibacillary leprosy; PB: paucibacillary leprosy.

Source: Adapted from WHO (24) and *Brasil, Ministério da Saúde* (18).

1.1.4 Transmission and incubation period

The exact mechanism of transmission of leprosy remains unclear; however, it has been hypothesized that leprosy is transmitted through nasal/mouth droplets from untreated MB patients to susceptible individuals in close and prolonged contact (reviewed in (2)). Despite humans being regarded as the main reservoir for *M. leprae*, interspecies transmission between human and armadillo (*Dasypus novemcinctus*) has also been suggested (25): Truman, *et al.* sequenced the *M. leprae* genome obtained from an infected armadillo and from lesions of a human leprosy patient, both from Southern United States. They found that both species were infected by the same *M. leprae* strain. Interestingly, a recent study from Neumann *et al.* investigated the hypothesis of leprosy transmission by mosquitoes (*Aedes aegypti* and *Culex quinquefasciatus*) or kissing bugs (*Rhodnius prolixus*) (26). As a result, they observed that none of the mosquito species presented potential to transmit leprosy. On the other hand, *M. leprae* remained alive during 20 days in the kissing bug digestive tract and even in their fezzes. In fact, the authors have demonstrated that viable *M. leprae* obtained from the kissing bug feces still has potential to infect, indicating that leprosy could also be transmitted by this insect in tropical regions.

Still a major challenge in leprosy research is estimation of the incubation period, as define by the interval between exposure to the pathogen and the clinical manifestation of the disease. Much of the existing evidence is based on leprosy diagnosed in individuals living in non-endemic areas whose exposure can be inferred through a history of contact or previous residence in an endemic area (reviewed in (27)). What these studies show is that the incubation period varies considerably, ranging from months to 30 years, and that it generally appears to be longer for those with LL leprosy than with TT subtype. However, a common feature for both leprosy subtypes is that most estimates suggest incubation periods of several years: the mean incubation period is estimated to be 4 years for tuberculoid and 10 years for lepromatous leprosy (reviewed in (27,28)). In agreement with that, it is observed that – even in endemic areas – leprosy cases in patients younger than four years of age are very rare (29,30). The vast majority of cases are detected in adulthood, whilst the

age group that is most commonly affected among children younger than 15 years old is between 10 and 14 years of age (reviewed in (31)).

1.2 HUMAN GENETIC SUSCEPTIBILITY TO INFECTIOUS DISEASES

A common observation in human infectious diseases is that contact with the pathogen is necessary but not enough for an individual to become infected and develop clinical disease. Host factors, including genetic background, have a crucial role on the outcome of microbial exposure (reviewed in (32)). A complete review of genetic risk factors predisposing to infection is beyond the scope of this introduction; comprehensive reviews of this topic have been published by Alcaïs *et al.* (2009) (32) and Chapman and Hill (2012) (33). Yet, selected empirical evidence and studies supporting the concept – with focus on mycobacterial infections – are presented below.

In mycobacterial diseases, a dramatic example that demonstrates the inherent spectrum of susceptibility to infection is a tragic event known as the Lübeck disaster. In that case – in the pre-antibiotic 1930s –, 251 newborns were accidentally vaccinated with Bacillus Calmette–Guérin (BCG) vaccine contaminated with a virulent *M. tuberculosis* strain (reviewed in (34)). As a consequence, 228 infants developed clinical disease and 72 died from tuberculosis (TB) within a year of inoculation. Overall, 68% of those who had developed clinical TB recovered spontaneously, indicating natural resistance to TB. Also, analysis of the available data indicates that different vaccine batches were contaminated with different amounts of *M. tuberculosis.* It was observed that the infection dose have an important impact on the outcome of such exposure, since increase mortality was attributed to increased dose of *M. tuberculosis*. However, children who have been inoculated with similar amounts of *M. tuberculosis* displayed a broad spectrum of clinical symptoms ranging from total absence of clinical disease to death. This suggests that host-related factors, such as genetic background, may play an important role in innate resistance to this infectious disease (reviewed in (34)).

Some of the most compelling evidence that human genetics does indeed determine the occurrence of infection comes from the identification of Mendelian forms of susceptibility or resistance to infectious diseases (reviewed in (35,36)). For example, Mendelian Susceptibility to Mycobacterial Disease (MSMD) is a disorder characterized by selective predisposition to clinical disease caused by weakly virulent mycobacteria species such as attenuated live *M. bovis* from the BGC vaccine and nontuberculous environmental mycobacteria. This condition manifests during childhood and is caused by rare mutations in genes encoding proteins of the IL12-IFN-γ pathway that are transmitted following a recessive model (reviewed in (36)). Other infectious diseases – with the exception of salmonellosis – rarely occur in these patients (reviewed in (36)). Intriguingly, inborn errors of immunity can lead to monogenic predisposition not only to multiple infectious diseases but also to a single type of infection (reviewed in (32,35,37)). A proof of concept of the existence of such traits was the identification of monogenic predisposition to TB in patients with IL-12Rβ1 deficiency, a genetic cause of MSMD (38,39). Severe TB characterizes these cases in the absence of prior infection by weakly virulent mycobacteria. Indeed, an increasing number of disorders in which a single gene – with variable penetrance – confers predisposition to a single infectious agent have been identified in children, adolescents, and even young adults (reviewed in (32,36)).

1.3 HUMAN GENETIC SUSCEPTIBILITY TO LEPROSY

The observations that i) after exposure to *M. leprae,* a majority of people will never develop clinical symptoms of leprosy; ii) the bacteria has low genetic variability; and iii) the disease has a wide range of clinical phenotypes which is dependent on host immune response; strongly suggest that most of the disease variability, including susceptibility to leprosy *per se*, is dependent of the genetic background of the host (40). Today, it is widely accepted the notion that different sets of genes modify host susceptibility to leprosy in three different stages, namely: i) the control of infection *per se*, that is the disease irrespective of clinical subtype, ii) the definition of different clinical forms of the disease and iii) the risk of developing leprosy reactions (reviewed

in (21,22,41–44). Observational studies in twins have revealed increased concordance of leprosy phenotypes in monozygotic (MZ) as compared with dizygotic twins (DZ) (reviewed in (41)). In 1966, a study of 35 pairs of twins (65.7% MZ) – recruited in three regions of India – identified a concordance rate of 82.6% in MZ twins vs 16.7% in DZ twins, for leprosy *per se* (45). Later, in 1973, Chakravartti and Vogel increased this population sample to 102 pairs of twins (MZ 60.8%) and detected leprosy concordance rate of 59.7% in MZ vs 20% in DZ twins (46). When leprosy clinical forms were evaluated, again the authors observed a higher concordance in MZ twins (51.6%) compared with DZ twins (15%) (46). Moreover, complex segregation analysis has clearly shown the presence of a genetic component controlling leprosy susceptibility (47–49). Even though the best-fit inheritance model is not consensus across studies, they consistently detected the existence of a major gene(s) controlling leprosy susceptibility (47–49). In addition, linkage and association studies involving several genomic regions and candidate genes, as well as genome-wide studies, have resulted in the description of several common genetic variants associated with leprosy (reviewed in (21,22,41–44)). A limited selection of findings from genome-wide linkage and association studies – candidate-free approaches – as well as follow-up studies is presented next.

The first leprosy genome-wide linkage study, conducted by Siddiqui *et al.* in 2001, resulted in evidence of linkage between leprosy and chromosome region 10p13 in a population sample of Indian families composed by 95% of PB affected individuals (50). A subsequent genomic scan performed by Mira *et al.* identified evidence linking this same region and PB leprosy in a familial sample from Vietnam (51). Based on these findings, Alter *et al.* performed an association study for three single nucleotide polymorphisms (SNPs) in *MRC1* gene – located in this chromosomal region – and susceptibility to leprosy and its clinical forms (52). The study was conducted in Vietnamese and Brazilian population samples. Besides being a positional candidate, *MRC1* is also a functional candidate since the protein encoded by this gene is a receptor that recognizes pathogen-associated molecular patterns. Surprisingly, results revealed association between *MRC1* variants and leprosy *per se* and MB leprosy, but not PB. Moreover, in 2014, the same group conducted a gene-centered high-density association scan of the underlying interval for susceptibility to the disease and its clinical forms (53). In total, 39 genes located in the 10p13 region were tested for

association in two independent family-based population samples from Vietnam. As a result, the authors identified two independent association signals in *CUBN* and *NEBL* genes. Again, the evidence of association was statistically significant between MB leprosy and both genes, and not with the PB clinical form. In 2016, Medeiros *et al.* tested variants in *GATA3*, a gene located in an interval of 6.5 Mb from the linkage peak on chromosome 10p13, for association with leprosy *per se* and its clinical forms (54). Seven tag SNPs have been selected covering the gene, and a stepwise association study in two case-control population samples from Brazil has been performed. One *GATA3* SNP was associated with leprosy *per se* in both population samples. Taking together, these association studies have identified three 10p13 and one neighbouring genes involved in the control of leprosy susceptibility. However, these findings still do not explain the PB linkage peak at 10p13; one hypothesis is that common variants (which are tested in association studies) may not be enough to explain the linkage peak detected for the 10p13 region and PB leprosy. To test this hypothesis, studies involving rare variants are necessary in order to better understand the relationship of this genomic region and PB leprosy in these populations – as well as validation studies in independent population samples (51,53).

In addition to the linkage peak in the 10p13 region, the genomic scan from Mira *et al.* also found a linkage peak at chromosome 6q25-q27 for leprosy *per se* (51). In a subsequent study, the same group held a fine mapping association study of the region in two population samples from Vietnam and Brazil (55). This analysis resulted in the identification of 17 SNPs associated with leprosy susceptibility in the Vietnamese population, 15 of them located in and around the promoter region shared by two genes: *PARK2* (a well-known Early Onset Parkinson disease related gene) and *PACRG*. In the same study, these results were validated in a separate set of unrelated individuals from Brazil. Two subsequent validation studies performed in an Indian (56) and a Chinese population (57) did not detected significant association between SNPs in *PARK2/PACRG* and leprosy susceptibility. Later, two independent studies conducted by Alter *et al.* (58) and Chopra *et al.* (59) performed fine mapping association analysis of *PARK2/PACRG* regulatory region in independent population samples; both studies confirmed association and revealed that differences in linkage disequilibrium patterns across different ethnicities may explain the heterogeneity of association between this *locus* and leprosy in previous studies. Moreover, Alter *et al.* also demonstrated that

PARK2/PACRG association with the disease is dependent on the age-at-diagnosis: a more pronounced genetic effect is found in early-onset patients (58). Curiously, *PARK2/PACRG* leprosy polymorphisms have been also described in association with typhoid and paratyphoid fever in an Indonesian population (caused by infection with *Salmonella typhi* and *S. paratyphi* respectively) (60). This finding suggests that the *PARK2/PACRG* genetic effect would not be specific to infection with *M. leprae*, but related to host responses against intracellular parasites. Recently, a functional study by Manzanillo *et al.* strengthened this hypothesis by showing that Parkin – the protein encoded by *PARK2* gene – plays a role in the pathway that leads to the degradation of intracellular pathogens by lysosomes (61). The study showed that Parkin operates controlling infection by different intracellular pathogens – such as mycobacteria, salmonella and listeria – in different hosts – such as mouse and *Drosophila melanogaster*. In addition, it has been shown by de Léséleuc *et al.* that abrogation of *PARK2* in macrophages and Schwann cells affects their ability to produce IL-6 and MCP-1 – two key pro-inflammatory cytokines – in response to mycobacteria and lipopolysaccharides (LPS) (62). Besides *PARK2/PACRG* association, a recent study conducted by Ramos *et al.* found a new gene located at chromosome 6q25-27 – called *SOD2* gene – as a risk factor for leprosy susceptibility in two independent Brazilian population samples (63). Indeed, *SOD2* expression was shown to be downregulated in human acute monocytic leukemia cell lineage THP-1 after stimulation with live *M. leprae* (64).

In addition to the evidence linking chromosome 6q25-q27 and leprosy *per se*, the genome-wide linkage study performed by Mira *et al.* has detected a second leprosy *per se* linkage signal at chromosomal region 6p21 (51). This region has been also linked to leprosy susceptibility in a Brazilian population sample (65). Chromosome 6p21 harbours the Major Histocompatibility Complex (MHC) – in humans, known as the Human Leukocyte Antigen (HLA) –, a cluster of highly polymorphic genes organized in three classes and with crucial role in immune response regulation (reviewed in (66)). In fact, several studies have already reported the involvement of HLA alleles and haplotypes as important genetic factors controlling susceptibility to leprosy, in particular for *HLA-DRB1* located in HLA class II (reviewed in (67)). To further explore the region underlying the linkage peak at 6p21, Alcaïs *et al.* performed a stepwise association scan of a 10.4 Mb region that encompass 224 annotated genes

located within and centromeric to HLA class II and class III regions (68). As a result, the authors identified a functional SNP in the HLA class III gene *LTA* as a risk factor in leprosy susceptibility in ethnically distinct populations. Interestingly, it seems that the *LTA* genetic effect on leprosy risk is age dependent, since evidence for association was as clearer as the age-at-diagnosis of cases decreased. Then, to identify additional genetic risk factors for leprosy in the 6p21 chromosomal region, the same group performed a high-resolution association scan of 1.9 Mb underlying the HLA complex (69). The association study was conducted in a Vietnamese population, followed by stepwise replication in an independent sample from Vietnam and from North India. The authors identified eight intergenic HLA class I region SNPs as novel genetic risk factors for leprosy *per se* and their results strongly implicate the *HLA-C* gene in leprosy susceptibility. In addition to the above-mention genes in HLA region, there is cumulative evidence that class III gene *TNFA* is also involved in the immune response against leprosy. Specifically, a promoter variant -308 of *TNF* has been extensively studied in leprosy, with controversial results (reviewed in (42)). To better understand the effect of the *TNFA* -308 variant on the disease, Cardoso *et al.* conducted a large association study involving four population samples and more than 2,500 individuals, followed by a meta-analysis (70). As a result, association between the promoter variant of *TNF* and leprosy has been confirmed – interestingly, the effect was stronger in the Brazilian samples.

In 2009, Zhang *et al.* published the first genome-wide association study (GWAS) of leprosy, in which a total of 491,883 markers scattered across the genome were tested for association with leprosy in a case-control Chinese population sample (71). From these, 93 SNPs were significantly associated and reanalyzed in three independent Chinese population samples. The authors identified 15 SNPs located in five *loci – HLA-DR-DQ, RIPK2, TNFSF15, NOD2* and *CCDC122-LACC1* – associated with the disease and one SNP in *LRRK2* gene presented a trend of association with leprosy. Later, Barrington *et al.* performed an association study between *NOD2* gene and leprosy *per se* and leprosy reactions in a population sample from Nepal (72). The authors validated *NOD2* association with leprosy, as well as showed that variants in this gene were also associated with susceptibility to leprosy reactions. In 2012, Grant *et al.* genotyped the 16 SNPs from the GWAS in a family-based Vietnamese population sample (73). As a result, association of SNPs located in *HLA-DR-DQ,*

RIPK2, NOD2 and *CCDC122-LACC1* were validated as risk factors for leprosy susceptibility in this population. Interestingly, the same research group stratified the Vietnamese population sample by T1R status and found that variants located in both genes that were not associated to leprosy *per se* by Grant *et al* – *TNFSF15* and *LRRK2* – are actually associated with T1R in this population (74,75). In addition, two association studies conducted by Wong *et al.,* in population samples from India and West Africa, only validated two *loci* from the GWAS: *HLA-DR-DQ* and *CCDC122-LACC1* (76,77). A recent study by Sales-Marques *et al.* conducted a stepwise association study of leprosy *per se* and the non-HLA genes that were significantly associated in the GWAS (*RIPK2, TNFSF15, NOD2* and *CCDC122-LACC1*) in five independent Brazilian population samples (78). Initially, 36 SNPs were genotyped, capturing the complete information of the five genes, in a family-based population sample from the Prata Village – an isolated, leprosy hyper endemic population located in the Brazilian Amazon. Two SNPs located in *NOD2* and *CCDC122-LACC1* were associated with the disease and were subsequently replicated in three independent Brazilian case-control population samples (78).

In 2011, the same Chinese group who published the first leprosy GWAS released the results of an expanded analysis performed by combining their first data set with additional control subjects (79). In this study, two additional genes were identified associated with leprosy: *IL23R* and *RAB32*. Later, they expanded the Chinese population sample even more, reaching a total 8,313 cases and 16,017 controls (80). In addition to confirming all *loci* identified in the two previous GWAS (71,79), this later study identified *BATF3*, *CCDC88B* and *CIITA-SOCS1* as new leprosy susceptibility genes.

Complemetary to these findings, several studies have reported other regions and genes as candidates involved in the control of leprosy susceptibility. For example, chromosome regions 2p14 (81), 17q22 (65), 20p12-13 (65,82), and genes such as *IFNG* (83,84), *IL10* (85–87), *IL12B* (88,89), *IL18RAP-IL18R1* (88), *SLC11A1* (90,91) and toll-like receptor genes – including *TLR1, TLR2* and *TLR4* (reviewed in (21,41,43)) – have been linked or associated with leprosy *per se*, clinical subtype or leprosy reactions. Understanding the genetic and molecular basis of control of host's susceptibility to leprosy is crucial to progress on the understanding of its pathogenesis.

1.4 MISSING HERITABILITY

Over the past recent years, genetic studies on common complex traits have been focusing on the identification of common variants that could explain predisposition to disease. In this setting, numerous GWAS have been published and, as result, several common variants were described to be associated with complex diseases. However as the numbers of GWAS raised, it became clear that part of the genetic effect controlling disease susceptibility was missing for several complex traits (reviewed in (92,93)). Therefore, common genetic variability is unlikely to explain the entire genetic predisposition to disease, giving rise to the term “missing heritability” referring to the heritable component of a disease not captured by association studies, including GWAS (reviewed in (92,94)).

Additional contributions in the genetic control of complex traits with missing heritability will depend on alternative research approaches and strategies. An interesting hypothesis to be tested is one that argues that variants too rare (MAF < 1%) to be detected by GWAS may explain – at least partially – some of the missing heritability (reviewed in (93,94)). According to this hypothesis, rare variants with relatively large effects on risk may contribute substantially to the genetic control of common complex diseases. Moreover, identifying more refined phenotypes or endophenotypes would provide a more tractable target in GWA studies than broad disease phenotype. An alternative approach has been to identify and study individuals or families with cases that are exceptionally severe or deviate otherwise from the typical disease cases (extreme cases), which could harbour single or oligo gene effects and therefore be compatible with a Mendelian hypothesis (reviewed in (95)).

In the context of infectious diseases, a model presented by Alcaïs *et al.* suggests that the genetic architecture of infectious diseases is a continuous spectrum ranging from single-gene variations predisposing individuals to infectious disease in childhood (specific or non-specific infections) to polygenic factors for complex diseases in adults (96). In leprosy, the vast majority of cases are detected in adulthood and young cases are rare (29,30). Patients with age-of-onset below four years represent the extreme end of the age distribution of cases and may be considered as and extreme leprosy endophenotype. Hence, early-onset leprosy cases could be

investigated under a monogenic model for identifying rare or low frequency causal variants with strong effect that could parcially contribute to the understanding of the missing heritability in this disease. By definition, rare cases cannot be studied by classic population-based genetic-epidemiological studies. Thus, different research approaches are needed. With the advent of Next-generation sequencing (NGS) technologies, candidate genes, large genomic regions or even whole genomes of a small number of affected individuals can be assessed for the identification of high impact and possibly disease-causing variants (reviewed in (97,98)).

The aforementioned approach has been successfully applied to determine the genetic basis of rare disorders, much of them Mendelian, through the study of a small number of affected individuals (reviewed in (95)). In this scenario, an interesting question would be whether the same strategy could be applied to the identification of disease-causing variants possibly contributing to the risk of occurrence of a complex disease, such as common infections (reviewed in (95)). In this much more complex context, leprosy has been considered as an excellent model to the study of genetic susceptibility to common infectious diseases (reviewed in (99)). The *M. leprae* is widely known because of its limited diversity between strains of different locations (17); this near clonal characteristic, together with the observation of a wide range of leprosy clinical phenotypes, strongly suggest that most of the disease variability, including susceptibility to disease *per se*, is dependent on the genetic background of the host. It is reasonable to believe that innovative approaches based on NGS technology could help to unravel much of the "missing heritability" observed in leprosy and other infectious diseases.

1.5 WHOLE EXOME/GENOME SEQUENCING

Next-generation – also known as second generation – sequencing technologies allow massively parallel sequencing of DNA and RNA and their use can be directed to the identification of common and rare variants by sequencing candidate genes, large genomic regions or even whole genomes (reviewed in (97,98)). Today, there are

several NGS platforms commercially available (reviewed in (97), (98)). In our study, we applied two NGS platforms – Ion Proton™ (Thermo Fisher Scientific) and HiSeq® 2500 (Illumina). These sequencers make use of different proprietary sequencing chemistry and base detection method; however, they share general NGS DNA sample processing steps (reviewed in (97)). For both methods, the initial step is to prepare libraries of DNA fragments ligated to universal oligonucleotide adapters. For that, genomic DNA (gDNA) is randomly fragmented in products with a length range that depend on the platform's targeted read length and the used chemistry (reviewed in (97)). Then, platform-specific adaptors are binded to both ends of each DNA fragment in order to ensure uniform PCR amplification of all molecules using a single pair of primers that are complementary to the adaptors' sequences. Optionally, adaptors containing specific DNA sequences called barcodes can be used to each sample's library to allow multiplexing in subsequent steps (reviewed in (97), (98)).

Methods for targeted enrichment can be coupled to massive parallel sequencing in order to sequence only a subset of the genome (reviewed in (97)). These capture methods can be applied to analyze genomic regions encompassing all coding regions of known protein-coding genes (defined as exome) (reviewed in (100,101)). In case of exome sequencing, an additional step of target enrichment is necessary after DNA library preparation and it is based on enrichment by hybridization capture (reviewed in (97,100)). The aim of this approach is to separate DNA fragments that contain the target sequences from the remaining DNA fragments by hybridization with biotinylated oligonucleotide baits (probes) that are complementary to the exome targets. After DNA libraries are incubated with these probes, magnetic streptavidin beads are added so that the biotin binds to streptavidin. This allows, by applying a magnetic field, to pull-down the bound libraries and wash out DNA fragments that remained free in the solution. Exome-enriched libraries are then eluded and used in template preparation step (reviewed in (97,100)).

For parallel sequencing, each molecule present in the library need to be spatially separated, attached to a solid surface or support and clonally amplified prior to sequencing. For that, two different *in vitro* template preparation methods are used for Ion Proton™ and HiSeq® 2500 platforms (reviewed in (97)). In Ion Proton™ workflow, the templates are prepared by emulsion PCR (emPCR): each DNA fragment – together with a primer-coated bead and PCR reagents – is isolated in independent

aqueous micro-reactors surrounded by an oil phase (reviewed in (97), (102)). These primers are complementary to the adaptors sequences so that, after the emPCR, all clonal amplicons are physically attached to the beads. These template-positive beads are called Ion Sphere Particles (ISP). On the other hand, template preparation in HiSeq® platforms is performed by bridge-PCR (reviewed in (97), (98)). For that, the DNA library and PCR reagents are dispensed onto a solid surface – a special slide known as "flow cell" – that habors primers complementary to Illumina adapters. During amplification, each single-strand DNA fragment binds to the primers attached to the slide and is copied by DNA polymerase. The newly copied DNA fragment is now physically attached to the flow cell by its 3' end. In the next PCR cycle, the free end of the attached single strand template can randomily hybridise to an immediately adjacent primer, thus forming a "bridge-like" structure. Based on the later, a new round of PCR reaction takes place, and the cycle bridge formation - PCR is repeated several times, producing thousands of clonal copies that form a 'cluster'. The clusters need to be spatially separated from each other in order to produce an unambiguous, monoclonal signal by the sequencer. Noteworthy, for a cluster or bead to be able to generate a read, each entity must be composted by monoclonal amplicons originating by a single DNA fragment (reviewed in (97), (98)).

The next step is the DNA sequencing, which for both platforms relies on a NGS method known as sequencing-by-synthesis: data acquisition is performed while DNA polymerases copy all molecules in a given cluster or ISP and each nucleotide that is incorporated simultaneously in all fragments will yield a single signal. After all molecules have been copied, the collection of signals gathered for a given cluster or ISP will be recorded as one read. However, each sequencer applies different methods for detecting and recording the addition of these nucleotides (reviewed in (97), (98)). In Ion Proton™, the ISPs – as well as sequencing primers and polymerase – are loaded to a sequencing chip (103). The chip contains millions of microwells on its surface, each one with a diameter sufficient to accommodate only one bead. Under the microwells, there is a sensor plate (ISFET: Ion-Sensitive field-effect transistor) sensitive to pH changes ((103), reviewed in (102)). Sequencing occurs when the chip is flooded with a solution containing an unmodified deoxinucleotide (A, T, G, or C), which is added in a sequential order (flow order). At each flow, if the nucleotide injected into the system is complementary to the template, the polymerase incorporates it to

the nascent DNA strand. When a phosphodiester bond is formed, the release of a hydrogen ion leads to a pH change inside the microwell. According to the flow order and the magnitude of the pH change at each flow (known as “flow signal”), Ion Proton™ sequencer builds the read for a given microwell ((103), reviewed in (97,102)). In HiSeq® 2500 platform, sequencing takes place on a number of lanes of a flow cell, which habors the template clusters. Four fluorescently labeled nucleotides are incorporated to the reaction simultaneously – each one with a different fluorescence color – together with sequencing primers and DNA polymerase. Once a nucleotide is added to the nascent DNA by the polymerase, the nucleotide’s fluorophore occupies the 3’-OH preventing the addition of subsequent nucleotides. Following incorporation, the remaining unincorporated nucleotides are washed away and an imaging step is performed to record cluster-specific fluorescence to determine the identity of the incorporated nucleotide in each cluster. Then, the bond between nucleotide and fluorescent dye is cleaved and an additional washing step is performed before the cycle is repeated (reviewed in (97), (98)).

NGS platforms can sequence one or both ends of the same DNA molecule, which creates single-end and pair-end reads respectively (**Figure 3-A**) (reviewed in (97)). Data quality (Quality Phred score, Q) of each base in a read is reported based on the logarithmic Phred scale (reviewed in (104)). For example, if a base is identified with Q20, it means 1% probability of misidentification (99% identification accuracy), if it is called with Q30, the probability of error is 0.1% (99.9% identification accuracy) and so on. In NGS data analysis, bioinformatics pipelines are then followed in order to process the reads and identify variants (reviewed in (104)). Initially, the reads are aligned/mapped to a reference sequence, which aims to reconstruct the original sequence from which the reads were generated (**Figure 3-B**). The number of times each base is sequenced in independent events (non-PCR-duplicates reads) defines the depth of coverage (**Figure 3-B**) (reviewed in (104,105)). Finally, after reads alignment, it is possible to perform identification of single nucleotide variants (SNVs) and small insertions and deletions (indels), a process term as variant calling (**Figure 3-B**) (reviewed in (104)). Several parameters of quality control are considered in this step in order to obtain high quality variant detection (106). However, platform-specific false positive rate can be high in NGS sequencing, in particular for indel calling (reviewed in (107)). Hence, a common practice in genomic laboratories is to confirm

the findings by Sanger sequencing, which is considered gold-standard method regarding sequencing accuracy (reviewed in (107)). Sanger sequencing is based on the utilization of fluorescently labeled dideoxynucleotides acting as chain terminators during sequencing reaction, followed by capillary electrophoresis and this method is reviewed in (97,108).

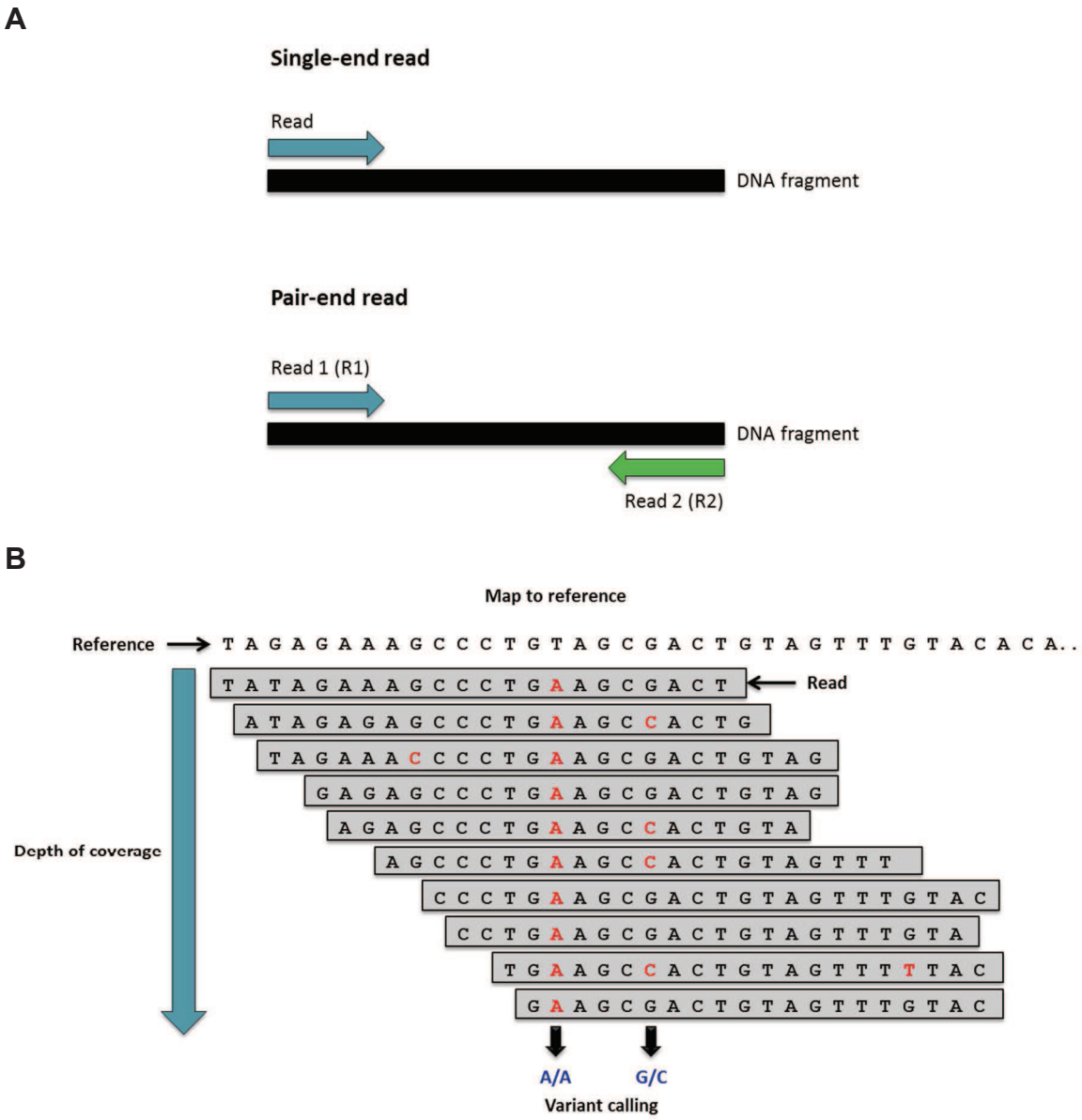


Figure 3. Schematic representation of read sequencing and analysis.
A) DNA library sequencing in single-end and pair-end reads. Arrows indicate sequencing direction. **B)** Read alignment/mapping to reference sequence and variant calling in NGS data for one sample. Here, the identification of two SNVs are shown, where in the first example (left) the individual is homozygous for the alternative allele, while in the second example (right) the individual is heterozygous. For both variants, the depth of coverage of their *loci* is 10X.
Source: A) Adapted from Morey, *et al.* (97). **B)** Adapted from Pavlopoulos *et al.* (105).

Once the variants are identified, they are annotated according to genomic and functional databases, for: function (i.e. non-coding, exonic, splicing or intronic); gene; classification (i.e. synonymous, missense, nonsense or frameshift indel); amino acid change; minor allele frequency (MAF); NCBI dbSNP reference number (rsID) (109) and functional prediction scores, etc. Regarding to MAF, it can be searched in several population samples from public databases such as the 1000 Genome Consortium Project (1000G) (110) and Exome Aggregation Consortium (ExAC) (111). 1000G reconstructed the genomes of 2,504 individuals from around the world using a combination of low-coverage WGS, deep WES and dense microarray genotyping (110), while ExAC database compiled WES data from 60,706 individuals from different ethnicity (111).

In addition, computational tools can be used to predict the biochemical impact of variants and help prioritize candidate variants and genes that are most likely to be deleterious for the protein structure and function (reviewed in (112)). Understanding how these tools calculate each prediction score is beyond the scope of this introduction, but a short presentation of some computational prediction programs are present as follow.

As variant-level approaches, SIFT (Sorting intolerant from tolerant) (113) and PolyPhen-2 (Polymorphism phenotyping version 2) (114) are widely used tools for *in silico* prediction of possible impact of amino acid substitution (missense variants) on the protein stability and function. The first one is based on protein sequence conservation, while the second is based not only on sequence conservation but also on biochemical properties of amino acids (112–114). Comparison of both methods indicated that PolyPhen-2 predictions are more accurate than SIFT's results (114,115). In PolyPhen-2, the prediction score ranges from 0 to 1 – where the higher the score, the most damaging is the variant for the protein structure and function (116). While these tools are restricted to missense SNVs, CADD (Combined Annotation–Dependent Depletion) can be applied for scoring the deleteriousness of SNVs and short indels in coding and non-coding regions of human genome (117). CADD integrates multiple genome annotations – including conservation metrics, functional genomic data, transcript information; and protein-level scores – into a single score. The higher the score, the more likely is the variant to have a deleterious effect. Based on the rank of each variant relative to all possible substitutions (including observed

and simulated variants), CADD scores are scaled in a Phred-like score ranging from 1 to 99 – where a scaled CADD of 10 compresses the top 10% of variants with highest score, CADD-20 the top 1%, CADD-30 the top 0.1% and so on (117,118). A remarkable study that experimentally tested these prediction tools was conducted by Miosge *et al.* (115). They performed *in vivo* tests of 30 missense mutations in 23 genes with known immunological phenotype in knockout mice as well as *in vitro* experiments to test the impact of 2,314 possible missense variants in human *TP53* gene and compared these results to the scores from computational prediction softwares including SIFT, PolyPhen-2 and CADD. The result from *in vivo* and *in vitro* experiments showed that the prediction softwares generate a low rate of false negative (FN) but a high rate of false positive (FP) (115). Despite the high FP rate, integrating variant annotation prediction such as CADD has successfully facilitated the identification of causative variants in WES data from patients with rare conditions (119,120).

Besides variant-level approaches, computational tools can be used to prioritize candidate variants based on properties of the genes where they are located. For example, a gene-level metric called Gene Damage Index (GDI) was recently developed by Itan, *et al.* (121). This tool was designed based on the observation that 58% of rare variants in the protein-coding exome of general populations are located in only 2% of the genes (121). Thus, these genes are less likely to cause monogenic diseases. However, variants in these genes may pass as false positive candidates when allele frequency and variant impact are considered during filtering steps in WES studies. So, the idea behind this tool is to identify – and filter out – FP variants in genes that accumulate high impact variants in the general population and are unlikely to be disease causing. GDI score was calculated based on gene variations of populations from 1000G database and the CADD score for calculating impact (121,122). Finally, the authors tested this approach with WES data from 84 patients with Primary Immunodeficiencies, which are rare disorders that impair host defense mechanisms and result in predisposition to multiple infectious diseases. As result, it was demonstrated that GDI was highly effective for detecting FP variants of highly mutated genes in these patients (121).

2 RATIONALE AND OBJECTIVES

Since leprosy is an infectious disease with a significant genetic component, it is reasonable to assume that a better understanding of the molecular basis of mechanisms controlling susceptibility to leprosy phenotypes will provide critical new insights into the disease pathogenesis. Past efforts to unravel the exact nature of these genetic mechanisms have resulted in the description of several common variants associated with the disease. However, these findings cannot explain the totality of the large genetic effect reported in twin studies, and innovative research approaches are needed. In order to contribute to the efforts in this direction, we used whole genome and exome sequencing by NGS, allied to whole exome analysis, to seek for both rare and common variants in a very particular small pedigree containing multiple leprosy-affected individuals, including a unique pair of MZ twin girls who developed childhood tuberculoid leprosy during their first two years of life. Clinical features of the disease, such as number and distribution of lesions were strikingly similar on both girls. The concordant, extreme early-onset leprosy observed in the twin pair combined with the presence of the disease across three generations suggests a strong genetic effect controlling leprosy mechanisms in the pedigree. Thus, we hypothesized that there is a leprosy genetic component following a Mendelian (monogenic) trait in the studied family that led to extreme early-onset leprosy in the MZ twin girls. To investigate this hypothesis, the objectives of this study are:

1. To describe a complete set of genomic variants present in the studied family through combined whole exome and whole genome sequencing using two different second generation sequencing technologies;
2. To develop and apply different custom filtering approaches and models to determine co-segregation of coding variants with leprosy in the studied family;
3. To perform *in silico* prediction of functional impact of the candidate variants identified.

3 CASE REPORT – THE PIAUÍ FAMILY

The pedigree of the studied family is presented in **Figure 4;** leprosy affected individuals are shown in black, unaffected members of the family are shown in white and unknown phenotype is indicated in grey. A description of the case is present next.

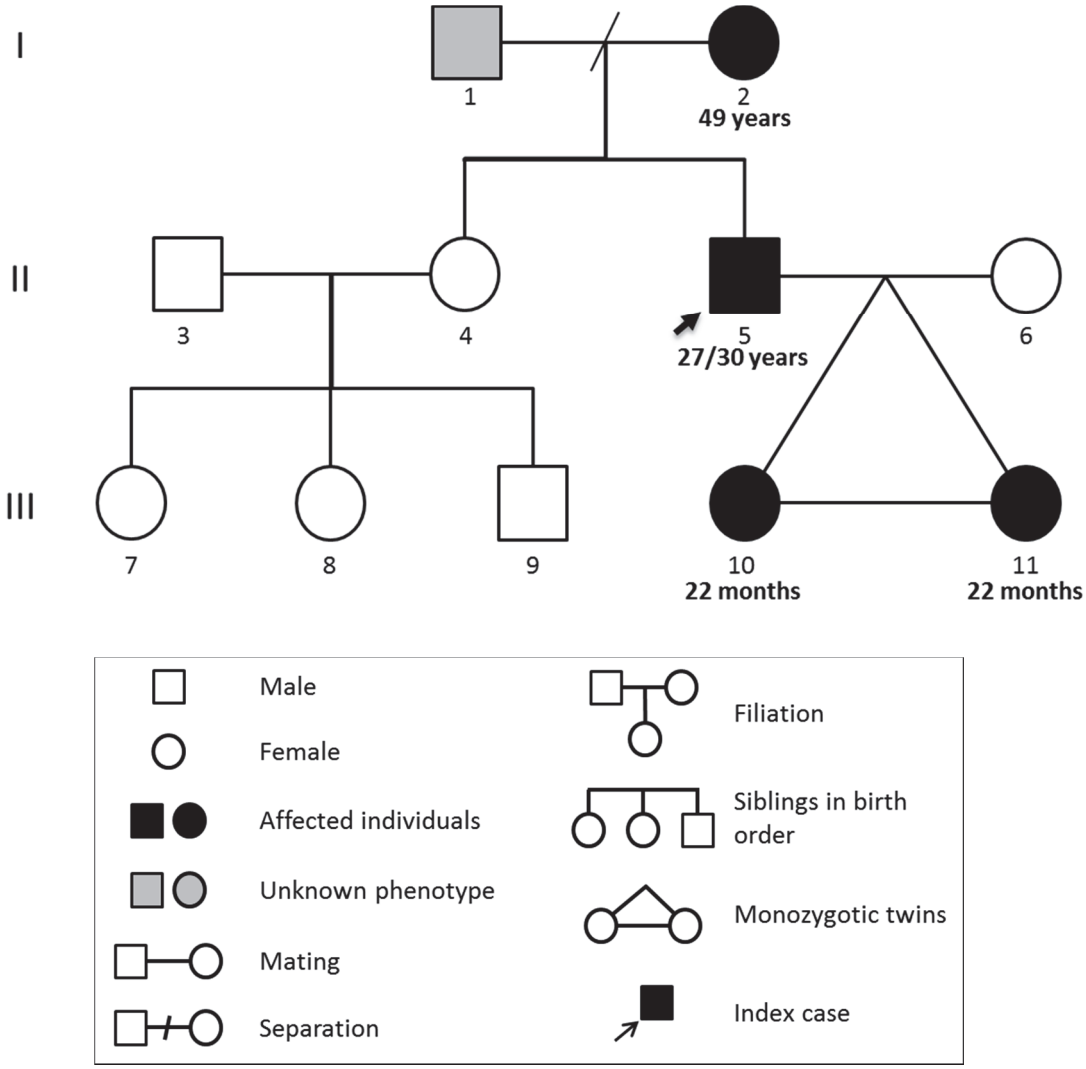


Figure 4. Pedigree of the studied family.
Each generation of the family is identified by roman numerals on the left and each individual is numerated. Age-at-diagnosis of leprosy patients is shown in bold. Pedigree symbols are shown in the box.

In June 2008, a 27 year-old man (II-5) was diagnosed with LL leprosy at *Maria Imaculada* Center, a referral hospital on leprosy located in Teresina, capital city of Piauí, northeast of Brazil, and initiated the MDT-MB treatment as implemented by WHO (18). Two months later, he returned to the hospital to receive his supervised dose of medication, accompanied by his 29 year-old wife (II-6) and twin daughters (III-10 and III-11) of 22 months of age. Surveillance of household contacts is part of Brazilian strategies for leprosy control (123). Hence, wife and daughters were submitted to contact examination. The wife did not present any skin or neural abnormality compatible with leprosy. The two girls caught the attention of the doctor for both presenting five small, well-delimitated nodules in their faces, arms and legs. Upon examination by two specialist dermatologists with large experience with leprosy, both girls were diagnosed with TT leprosy. Strikingly, the distribution of the lesions was remarkably similar on both girls. Neither biopsy nor bacilloscopy were performed. The girls were also exanimated by a panel of leprosy specialists from the Brazilian Ministry of Health that were visiting the Center, who confirmed the diagnosis. Based on clinical diagnosis, treatment for PB leprosy was initiated. The twins' treatment coursed without intercurrences and they were reported cured. In parallel, it was reported that the twin's father (II-5) abandoned the treatment before the expected period. In February 2011, the paternal grandmother (I-2) of the twin girls was diagnosed with leprosy (I subtype) at the age 49 years old and was treated with MDT-PB. Moreover, in June 2011, the index case attended Getúlio Vargas Hospital – another referral centre on leprosy in Teresina – where he was diagnosed with MB leprosy again, due to the presence of several skin lesions and positive bacilloscopy. Again, MDT-MB treatment was iniciated and fully conducted according to protocol, and the patient was reported as cured. Since then, the whole family is being followed-up by the medical team from Getúlio Vargas Hospital. To date, neither of the ex-leprosy patients developed new clinical symptoms of leprosy nor leprosy reactions. Clinical data from the studied family are presented in **Table 2**.

Table 2. Clinical characteristics in the studied family.

ID*	Name*	Sex	Date of birth	Leprosy status	Date of diagnosis¶	Leprosy subtype¶	
						Ridley & Jopling	WHO
I-1	Grandfather	M	Unknown	Unknown			
I-2	Grandmother	F	Mar, 1961	Affected	Feb, 2011	I	PB
II-3	Uncle	M	Unknown	Unaffected			
II-4	Aunt	F	Dec, 1979	Unaffected			
II-5	Father	M	Mar, 1981	Affected	Jun, 2008	LL	MB
					Jun, 2011	LL	MB
II-6	Mother	F	1979**	Unaffected			
III-7	Cousin1	F	Mar, 1999	Unaffected			
III-8	Cousin2	F	Feb, 2004	Unaffected			
III-9	Cousin3	M	Apr, 2005	Unaffected			
III-10	Twin1	F	Sept, 2006	Affected	Aug, 2008	TT	PB
III-11	Twin2	F	Sept, 2006	Affected	Aug, 2008	TT	PB

* "ID" refers to the identification of each individual in the family pedigree from Figure 4, while "Name" refers to each sample identification used in this thesis.

** Unknown month of birth.

¶ Data from medical reports.

Apr: April; Aug: August; Dec: December; F: female; Feb: February; I: indeterminate; ID: identification; Jun: June; LL: lepromatous leprosy; M: male; Mar: March; MB: multibacillary leprosy; MDT: multidrug therapy; PB: paucibacillary leprosy; Sept: September; TT: tuberculoid leprosy.

The grandfather (I-1) of the twin sisters lost contact with the family before this study was initiated, thus no clinical or demographic data is available for this individual. The remaining members of the family live in the same city, but in three different households: i) the affected grandmother (I-2) lives alone; ii) the twin's unaffected uncle (II-3), aunt (II-4) and cousins (III-7, III-8 and III-9) live in the same household and iii) the unaffected mother (II-6), the affected father (II-5) and twin girls (III-10 and III-11) live together in a third household. Therefore, the unaffected mother (II-6) has been in prolongued contact with three leprosy-affected individuals as a household contact. On the other hand, exposure to *M. leprae* of the family members that live in the second household is not clear. Therefore, only the mother was included in the variant filtering approaches as an unaffected control (See section 5.5). All individuals in the family were BCG vaccinated. To date, none of the family members developed other mycobacterial infectious disease besides leprosy.

4 EXPERIMENTAL STRATEGY

The experimental approach pursued in this study consisted of sequencing the exome of family members searching candidate causal variants of the genetic predisposition to leprosy in the Piauí family (presented in chapter 3). For that, whole exome sequencing of four leprosy affected individuals – the twins (III-10, III-11), their father (II-5) and grandmother (I-2) – and one unaffected family member (the twins' mother, II-6) was performed. Exome was captured using the Targetseq Exome kit (Thermo Fisher Scientific) (124), which targets every exon of approximately 21,500 protein-coding genes and nearly 8,000 non-coding RNA (ncRNA) genes based on annotation of consensus coding sequence (CCDS) project (125) and RefSeq database (126). To cover the target regions, this in-solution array contains more than 2 million oligonucleotide probes ranging from 60 to 100 bp that tile 52.7 Mb of target regions including the exome and flanking areas. Then, the exon-enriched DNA libraries were sequenced by 200 bp single-end reads on Ion Proton™ Sequencer (Thermo Fisher Scientific) using the Ion PI™ Chip v2 (Thermo Fisher Scientific). Sequence data analysis was conducted using a pipeline for variant discovery with Torrent Suite (TS) software v5.0 available on GitHub (127) and variant annotation with wANNOVAR (128).

To identify those variants that are most likely to be causal, a custom stepwise procedure of filtering was developed and applied. Variants were filtered assuming both recessive and dominant traits and modeled based on the age-at-diagnosis of the affected members of the family. In addition to the segregation model, filtering steps were based on variant's location within protein coding genes (i.e. coding or splice-site), type (i.e. missense, nonsense or frameshift) and frequency in public databases. Once candidate variants have been identified, variant-level and gene-level metrics based on computational prediction were used to prioritize the variants that are most likely to have an impact on the protein function. Variant-level metrics were defined by using PolyPhen-2 (114) and CADD score (117). As gene-level metric, GDI was used in order to identify probable false positive candidate variants (121). As quality control, a selected number of variants identified in WES were validated by Sanger sequencing.

Even though WES analysis of the family allowed the identification of several candidate variants, a considerable fraction of the target exonic regions was insufficiently covered for variant identification and artifacts/missgenotype were detected. Consequently, WGS from the same samples (I-2, II-5, II-6, III-10 and III-11) was performed to improve data acquisition. Moreover, the aunt's sample (II-4) was included to increase accuracy of multisample variant calling from WGS data. Genome sequencing was performed on HiSeq® 2500 platform (Illumina) with 150 pair-end reads. The output data was analyzed following the best practices protocol created by the Genome Analysis Toolkit (GATK) development team at Broad Institute (129,106,130). Once SNVs and short indels were called and annotated, variant filtering and prioritization were performed – following the same procedure used in WES analysis – in order to identified candidate variants in coding regions. Finally, filtering results from WES and WGS data were compared.

A flowchart of the experimental approach followed in this study is presented in **Figure 5**. Each step is thoroughly described in "Methods" (section 5).

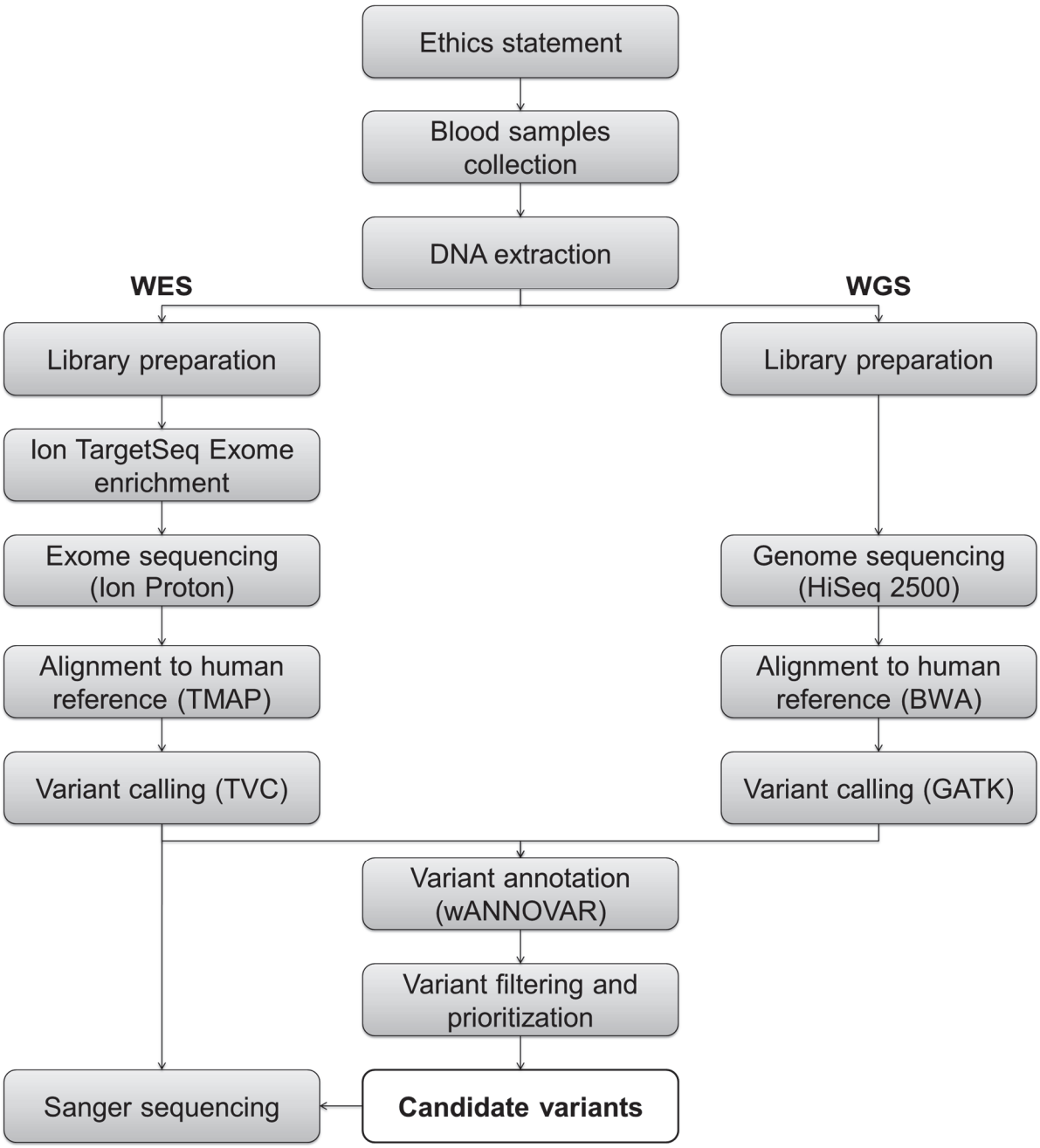


Figure 5. Flowchart of experimental approach.
BWA: Burrows-Wheeler aligner; GATK: Genome Analysis Toolkit; TMAP: Torrent Mapping Alignment Program for Ion Torrent; TVC: Torrent variant caller; WES: whole exome sequencing; WGS: whole genome sequencing.

5 METHODS

5.1 ETHICS STATEMENT

The Research Ethics Committees from the Pontifical Catholic University of Paraná (PUCPR), Carlos Chagas Institute – Oswaldo Cruz Fundation (ICC/Fiocruz) and Federal Univeristy of Piauí (UFPI) approved this study (**Appendix 3**, in Portuguese). All participants – or their legal representatives – have agreed to participate and signed written informed consent (**Appendix 4**, in Portuguese).

5.2 SAMPLE COLLECTION AND DNA EXTRACTION

Peripheral blood samples (10 mL) were collected from each family member, except for the grandfather (I-1) and uncle (II-3). Then genomic DNA extraction was performed by salting-out as described by John *et al.* (131). The extracted gDNA was quantified using Qubit® dsDNA BR Assay in Qubit® 2.0 Fluorometer (Thermo Fisher Scientific). DNA samples with a 260/280 nm ration ranging from 1.5 to 2.0 were considered adequate (97). Next, gDNA working solutions were prepared to a standardized concentration of 20 ng/µL.

5.3 WHOLE EXOME SEQUENCING

5.3.1 WES in Ion Proton™ platform

Five samples were selected for exome sequencing including the four affected members of the family – the twins (III-10 and III-11), father (II-5) and grandmother (I-2) – and one unaffected individual (the mother, II-6). DNA library preparation and exome enrichment were performed following manufacturer's protocol: "Ion TargetSeq™ Exome Enrichment for the Ion Proton™ System Protocol" (132). For library preparation, 1 µg of gDNA was enzymatically fragmented and ligated to Ion-specific adapters and barcode using Ion Plus fragment library kit (Thermo Fisher Scientific). For size selection, an electrophoresis run in E-Gel SizeSelect™ 2% agarose gels using E-Gel iBase™ and E-Gel Safe Imager™ Combo Kit (Thermo Fisher Scientific) was performed using a 50 bp-ladder as reference. The DNA migration was monitored in real-time until the target DNA band – near 285 bp – reached bottom row wells, from which samples were recovered with a pipette. At the final step of library preparation, DNA libraries were amplified in a thermocycler using Platinum® PCR SuperMix High Fidelity and Library Amplification Primer Mix (from Ion Plus fragment library kit, Thermo Fisher Scientific) and the following parameters: i) initial denaturation at 95ºC for 5 min, ii) 10 cycles of denaturation at 95ºC for 15 sec, hybridization at 58ºC for 15 sec and extension at 70ºC for 1 min, and iii) final stage at 12ºC. DNA was purified using magnetic beads (Agencourt® AMPure® XP kit, Beckman Coulter) and a magnetic rack after each step of library preparation (DNA fragmentation, adapter ligation, size-select and amplification). gDNA and DNA libraries were quantified using Qubit® dsDNA HS Assay in Qubit® 2.0 Fluorometer (Thermo Fisher Scientific). To verify the adapter-ligated DNA fragments sizes, a chip-based capillary electrophoresis was performed using Agilent High Sensitivity DNA Kit in BioAnalyzer 2100 platform (Agilent), following manufacturer's instructions.

For each sample, exome was captured and enriched in two independent experiments using three pools of samples with 125 ng to 167 ng of prepared gDNA library *per* sample. Then, 500 ng of libraries pool was dried in presence of human Cot-

1 DNA® and IonTargetSeq™ blockers (from Ion TargetSeq™ Exome Kit, Thermo Fisher Scientific) using a vacuum concentrator at 60^0C. The dried samples' pool was reconstituted according to the manufacturer's protocol, heat-denatured, and mixed with biotinylated DNA probes (Ion TargetSeq™ Exome Probe Pool from Ion TargetSeq™ Exome Kit, Thermo Fisher Scientific). Hybridizations were performed at 47°C for 72 hours. Once the capture was complete, the samples were mixed with DynaBeads™ M-270 streptavidin (Thermo Fisher Scientific), incubated at 47°C for 45 min and washed with a series of stringent buffers to remove non-bonded DNA fragments (from Ion TargetSeq™ Exome Kit, Thermo Fisher Scientific). The captured fragments were PCR-amplified using the following parameters: i) initial denaturation at 95ºC for 5 min, ii) 8 cycles of denaturation at 95ºC for 15 sec, hybridization at 58ºC for 15 sec and extension at 70ºC for 1 min and iii) final stage at 4ºC. Finally, the exome-enriched libraries pool was purified with Agencourt® AMPure XP beads (Beckman Coulter) and quantified using Agilent High Sensitivity DNA Kit in BioAnalyzer 2100 platform (Agilent). Exome enrichment was performed with pools of different samples in three independent experiments. Each pool encompasses samples from: i) the twins (III-10 and III-11) and their parents (II-5 and II-6) (pool 1), ii) the twins (III-10 and III-11) and the grandmother (I-2) (pool 2), ii) the parents (II-5 and II-6) and grandmother (I-2) (pool 3).

Two hundred base pairs template was prepared by emPCR using Ion OneTouch™ (composed of the Ion OneTouch™ Instrument and Ion OneTouch™ ES), Ion PI™ Template OT2 200 Kit v2 and DynaBeads™ MyOne streptavidin C1 beads (Thermo Fisher Scientific). Template preparation was performed following manufacturer's instructions in "Ion PI™ Template OT2 200 kit v2" protocol (133). Finally, the samples were sequenced using 200 bp single-end reads sequencing on Ion Proton™ platform (Thermo Fisher Scientific) using the Ion PI™ Chip v2 (Thermo Fisher Scientific) that produces around 60 to 80 million reads for a yield of approximately 10 Gb of data. For the sequencing step, "Ion PI™ Sequencing 200 kit v2" user guide was followed (134). Each pool of exome-enriched libraries was sequenced separately. Pool 1 was sequenced in two experiments and pool 2 and 3 were sequenced in one experiment each.

5.3.2 WES data analysis: pre-processing and variant calling

Initially, quality assessment of the raw data was performed using FastQC v 0.11.4 software (135). The WES reads were aligned to the human reference GRCh37 using map4 command line implemented on TMAP for Ion Torrent in TS software (**Appendix 1**) (127). Reads shorter than 30 bp were excluded from the analysis using view command in Samtools v1.3 (136). Mapped reads were sorted according to their genomic coordinate position using SortOrder command in Picard v1.134 (137). Using the same program, PCR duplicates were flagged with MarkDuplicates and the mapped reads were merged into a sample-level BAM file with MergeSamFiles command. Quality assessment of the mapped reads was performed using QualiMap v2.1.1 (138). To visualize reads mapping to specific genome regions, Integrative Genome Viewer (IGV) v2.3 (139). BedTools v2 (140) was used to identify on-target regions that presented *per* base depth of coverage ≥ 10X or ≥ 20X. For that, GenomeCoverage command in BedTools that generated a *per* sample BED file with these regions was applied. Then, these files were intersected among all samples using IntersectBed command in Bedtools in order to identify regions with coverage above threshold in all samples.

Variant calling was performed with Torrent variant caller (TVC) plugin from TS software, using “Germline - Proton TargetSeq - High stringency” parameter option with default settings (**Appendix 1**) (127). Identification of single nucleotide variants (SNV) and dinucleotide variants (DNV) was performed in regions with coverage ≥ 10X, while indel calling was performed only in regions with coverage ≥ 20X. Variant calling using TVC was performed for each sample separately and, as output, VCF files containing *per* sample variants were created. From VCF files, variants inside targeted regions were selected using the SelectVariants command from GATK software v3.4-0 (141). For that, Ion-TargetSeq-Exome-50Mb-hg19_revA.bed file provided by Ion Community (142) was used, which contains target regions coordinates for Ion TargetSeq™ Exome probes. Finally, the lists of checked variants from all the samples were combined in one multi-samples VCF file using CombineVariants tool in GATK.

5.3.3 Validation – Sanger sequencing

In total, 18 variants detected in WES analysis were selected for validation by Sanger sequencing (**Table 3**). From these, 10 were candidate variants that passed the variant filtering (see section 5.5) and eight were variants selected as examples of SNV, DNV and indels with heterogeneous coverage and quality or with discordant genotypes between the twin girls. Sanger sequencing was performed using DNA samples from the twins (III-10 and III-11), their parents (II-5 and II-6) and grandmother (I-2).

Primers were designed using default parameters of Primer3 (143,144). The amplicons were designed to have product size between 350 to 600 bp where the variant *locus* was at least 100 bp from the ends (**Table 3**). *In silico* PCR was performed as implemented in the UCSC genome browser in order to verify predicted specificity of the primers (145). For each individual, a PCR reaction was performed in a final volume of 25 µL *per* amplicons, as follows: 14.9 µL of water, 5 µL of DMSO, 2.5 µL of 5X PCR buffer with 25mM $MgCl_2$, 0.8 µL of each primer (10 µM), 0.4 µL of 10 mM dNTPs, 0.2 µL of 5 U/µL Taq polymerase and 1.2 µL of 20 ng/µL gDNA. Pre-set reaction conditions were: i) 94°C for 3 min followed by ii) 35 cycles at 94°C for 1 min, primers hybridization temperature (**Table 3**) for 30 sec, and 72°C for 30 sec; and iii) a final step at 72°C for 5 min followed by a cooling step at 12°C. Success of amplification was verified in 1.5% agarose gel electrophoresis. Amplicons were processed following BigDye Terminator v3.1 protocol (146) and sequenced on 3730xl DNA Analyzer platform (Thermo Fisher Scientific) using the forward primers. Electropherograms were analyzed using Lasergene's SeqMan software (DNAStar) (147).

lification and Sanger sequencing of 18 variants detected in WES.

Variation type	AA change	rsID	Primer forward	Primer reverse	Amplicon size (bp)	Hibridization Temperature
missense SNV	K576E	-	AGCTCTGGTGGTGAACTGAA	GGCTGCTGATATAGGGACCA	566	62°C
missense SNV	P170L	rs17585	TCTGGATGAGTCTGTGGGG	GTTGTCTGATGTCACACCGC	438	62°C
missense SNV	N551K	rs7308720	AGCACAGCCTACTCACACAA	CCACATCCCCACTGTCATCT	577	59°C
missense SNV	R1398H	rs7133914	GGTACTTTGATCGGTTGCTG	CACACGCACACAGACACAT	391	62°C
missense SNV	A208T	rs61755579	ACAACACCCAGCCCTGTTTA	AGTGGTGCGAGCTGAGATC	463	62°C
missense SNV	Q254K	rs9901673	CACCTGCTTCTCTCATTCCC	GTGGACAGCTGGTGAAAGAA	468	59°C
missense SNV	A229T	rs10852891	CTGGCCCGAATCTTCACTTC	AGGCTTGGCTGAATGACTGA	587	59°C
missense SNV	S284C	rs58154316	CCTGTGGAAGCGTGATCATC	GACCGACTGTGTTGTGATGG	561	59°C
missense SNV	I798L	rs6091375	AGCATCCACTCGGCATCA	GTAAAGTTCAACCCAGGCTCC	398	59°C
missense SNV	Q11L	rs179008	GCTGCTTCTACCCTCTCGAA	GCTGGGGAGATGTCTGGTAT	500	59°C
missense SNV	S144G	rs3747742	TGATCACTCAGCAGCCAGAA	GGCATGGAGGGTAGTCTGTT	559	59°C
missense SNV	A18T	rs2301721	CTTGCCCTTCCATTCTAGGC	CGGGGATGTTTTGGTCGTAG	498	59°C
missense SNV	Q1111H	rs78365431	CTGAAGAGTTTGACACATTTGGA	CAAGCGATTCTCATGCCTC	446	62°C
missense DNV	G1938E	rs386787404	TGGCCGCACAAAAGCTTATT	TCCACTTTGCCTTACAGCCA	500	59°C
frameshift deletion	P2462fs	-	CACAAAGGTTCCCATAGTGTTCA	TGTGGGGCTTTTCTGTTCTTC	580	59°C
insertion (UTR3)		rs145265135	TCCCAGCTAGTTTGAGGCAA	TCCTGTTTCCATTTATGCCCTTC	486	59°C
missense SNV	Q580R	rs61747965	GCTGGGAAGAGAAGGAGACA	GTGTCAAGTTCTCGGCTTCC	575	62°C
frameshift deletion	L2fs	rs398035013	CCTGAAGGGGAGCTTAGACC	GGCACTGAGGGAAGGCATA	496	62°C

pair; Chr: chromosome; DNV: Dinucleotide variant, Ref: reference allele; SNV: Single nucleotide variant, UTR3: 3' Untranslated region. *GRCh37/hg19.

5.4 WHOLE GENOME SEQUENCING

5.4.1 WGS in HiSeq® 2500 platform

WGS was performed for six samples including the twins (III-10 and III-11), their parents (II-5 and II-6), grandmother (I-2) and aunt (II-4). Library preparation was performed with TruSeq® DNA LT Sample Prep Kit - Set A (Illumina) following standard protocol from TruSeq® DNA Sample Preparation Guide (148). First, 1 µg of input gDNA for each sample was fragmented by ultrasound shearing using Covaris™ S2 with settings for Whole Genome Resequencing. DNA fragments were ligated with Illumina-specific adapters and then purified on a 2% agarose gel to remove unligated adapters and self-concatenated adapters, as well as select fragments with targeted size-range. For that, a gel band spanning the width of the lane and ranging in size from 400-500 bp was excised and DNA library extracted from agarose matrix. After this step, the adapter-binding libraries were PCR amplified (ten cycles) following manufacturer's instruction and quantified by qPCR. Fragment size distribution was checked using Agilent High Sensitivity DNA Kit in BioAnalyzer 2100 platform (Agilent). Finally, each library was normalized to 10 nM solution.

Template preparation and sequencing steps were done following "Sequencing in Rapid Run Mode" protocol as described in HiSeq® 2500 System Guide (149). HiSeq® Rapid Cluster Kit v2 (Illumina) was used for cluster generation (template preparation step). Finally, HiSeq® Rapid SBS Kit v2 (Illumina) was used for paired-end sequencing on HiSeq® 2500 platform (Illumina). Experiment settings were based on "Rapid run mode" set to generate paired-end 150 bp reads. With these settings, HiSeq® 2500 can sequence two independent flow cells (each with two lanes) at the same time and generates up to 180 Gb of data in a 40 hours run (149). For each run, one sample *per* flow cell was sequenced and, for each flow cell, both lanes for the same sample were used.

5.4.2 WGS data analysis: pre-processing and variant calling

WGS data was analyzed following GATK (141) best practices pipeline for alignment to reference genome and variant calling (**Appendix 1**) (129,106,130). Initially, quality assessment of the raw data was performed using FastQC software. The reads were mapped to human genome reference GRCh37 using the Burrows-Wheeler aligner "mem" algorithm (BWA-mem) v0.7.12 (150). Mapped reads were sorted according to their genomic coordinate position using SortOrder command in Picard v1.134 (137). Also with Picard tool, PCR duplicates were flagged. Next, local realignment around indels and base recalibration were performed using GATK v3.5 (141). Quality assessment of the mapped reads was performed using QualiMap tool. GATK HaplotypeCaller was used to call variants for each sample, followed by JointGenotyped for all samples together. As a result, one multiple-samples VCF file with all raw variants identified in the study was created. Next, two steps of call set refinement were performed: first, it was evaluated the likelihood of a variant being real in order to reduce the amount of false positive. For that, Variant Quality Score Recalibration (VQSR) from GATK, using default parameters (106,130) was used. Then, genotype refinement workflow from GATK was applied to filter *per* sample genotype calls that were not reliable enough for downstream analysis. For each sample, genotypes with quality score (GQ) lower than GQ20 were flagged as low quality genotype, which means that these genotypes had less than 99% chance of being correct. After applying the refinement steps, only high-quality variant calls (≥ GQ20) were used in downstream analysis.

5.5 VARIANT ANNOTATION AND FILTERING

The detected variants from WES and WGS data were annotated using wANNOVAR (as of February 2016) (128). MAFs from variants reported by 1000G and ExAC were collected from five population samples: African/African American (AFR),

Admixed American/Latin (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) (110,111). Detected variants were considered as novel when they were not previously identified in public databases and did not have a rsID. In this thesis, the term "functional variant" refers to nonsynonymous (missense and nonsense) SNV, splice-site variants and coding indels (frameshift and in frame).

Custom filtering steps were taken to help identify candidate variants (**Figure 6**). Initially, variants located within exonic (coding) and splice-site regions were selected (Filter 1). Since the goal was to identify coding, functional variants, synonymous variants were excluded (Filter 2). Next, we filtered out variants that had MAF lower or equal to thresholds (30%, 15%, 5% or novel depending on the filtering approach; see below) in each population sample, according to two public databases: 1000G and ExAC (Filter 3). Finally, we searched for variants co-segregating with disease following a Mendelian trait, modeled considering the age-at-diagnosis (Filter 4). The strategy resulted in seven filtering approaches, named "Models" and described as follows. Of note, WGS results from the twin's aunt were not included in filter 4, due to unknown exposure of this individual to *M. leprae*.

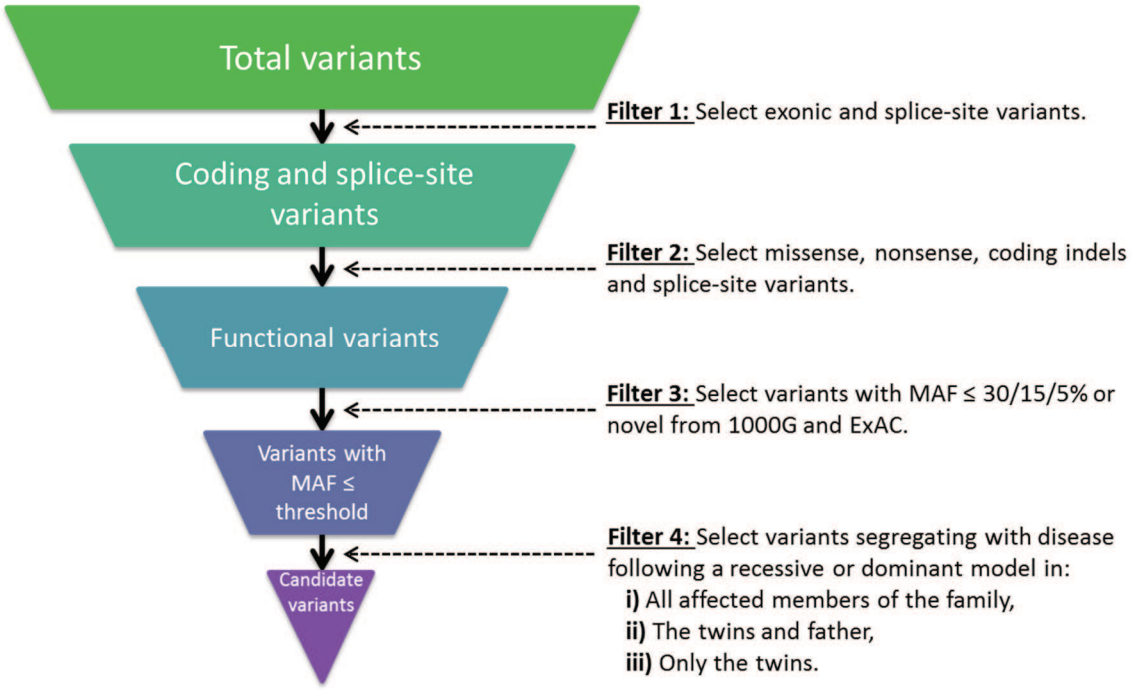


Figure 6. Variant filtering steps.
Different thresholds and criterias were used in filter 3 and 4 depending on the filtering approach that was applied (See **Figures 7** and **8**). 1000G: 1000 Genomes Consortium; ExAC: Exome Aggregation Consortium; indels: insertions/deletions; MAF: Minor allele frequency.

To investigate variants following a recessive trait, Models #1 to #4 were applied as follow:

- **Model #1**: assumes that leprosy genetic component is following a Mendelian recessive trait in all affected family members, independently of the age-at-diagnosis. For that, we searched for variants where the twins (III-10 and III-11), father (II-5) and grandmother (I-2) – but not the unaffected mother (II-6) – were homozygous for the minor allele, which frequency in public databases ≤ 30% (Model #1 in **Figure 7**).
- **Model #2**: assumes that the development of leprosy in the grandmother (age-at-diagnosis 49 years) was multifactorial, while disease genetic factor is following a Mendelian recessive trait in the younger cases (father and twins with age-at-diagnosis of 27 and <2 years, respectively). For that, we searched for homozygous minor allele presented in the twins (III-10 and III-11) and father (II-5) and absent in the grandmother (I-2) and mother (II-6) (Model #2 in **Figure 7**). In this model, we focus on variants with MAF ≤ 30% in public databases.
- **Model #3**: we searched for variants where only the twins (III-10 and III-11) were homozygous for the minor allele assuming that the genetic control to leprosy is following a Mendelian recessive trait only in the early-onset cases and a polygenic/multifactorial trait in the adulthood cases in the family (Model #3 in **Figure 7**). In this model, we selected variants with MAF ≤ 15% in all population samples from public databases.
- **Model #4**: same as Model #3 (Mendelian recessive model in early-onset cases), but here we searched for compound heterozygous in both twins (III-10 and III-11) for low frequent variants (MAF ≤ 5%) in autosomal chromosomes (Model #4 in **Figure 7**).

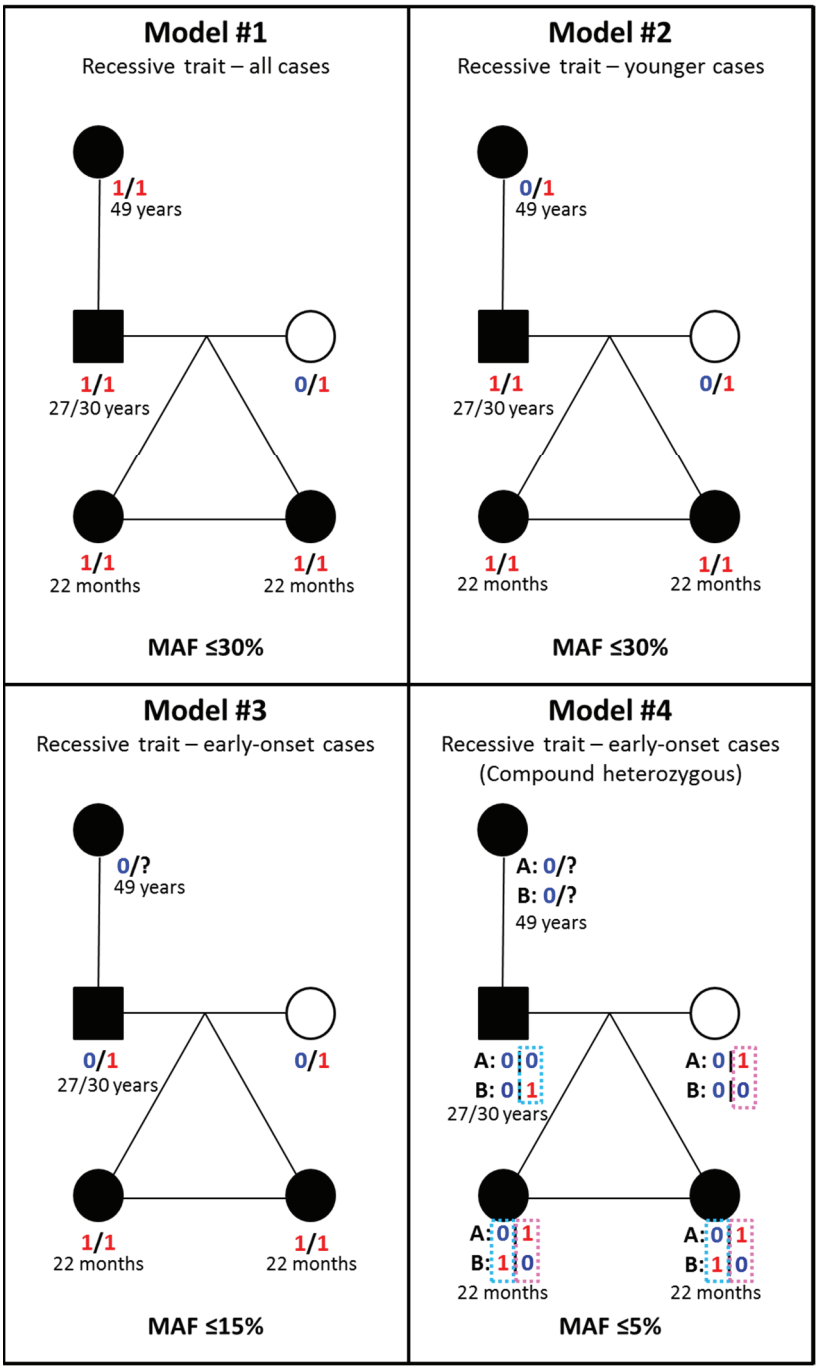


Figure 7. Filtering approaches applied to identify candidate coding functional variants segregating with disease following a recessive trait.
Number zero (in blue) represents the common allele and number one (in red) corresponds to the minor allele. Age-at-diagnosis of leprosy cases are shown in black. In Model #4, *loci* A and B are located in the same gene; haplotype transmitted from the father to the twins is inside a light blue box, while haplotype transmitted from the mother to the twins is inside a pink box.
MAF: Minor allele frequency.

For the models accounting for a dominant trait (Models #5 to #7), we focused on novel variants as follow:

- **Model #5**: assumes a Mendelian dominant trait controlling disease susceptibility in all the affected family members, independently of the age-at-diagnosis. We searched for variants in common among the twins (III-10 and III-11), father (II-5) and grandmother (I-2) and absent in the unaffected mother (II-6) (Model #5 in **Figure 8**).
- **Model #6**: assumes that only the twins (III-10 and III-11) and their father (II-5) were carriers of the causal mutation as a Mendelian dominant trait (young cases in the family) (Model #6 in **Figure 8**).
- **Model#7**: we searched for *de novo* mutations in both twins (III-10 and III-11) that could explain a leprosy genetic control following a Mendelian dominant trait only in the early-onset leprosy cases in the family (Model #7 in **Figure 8**).

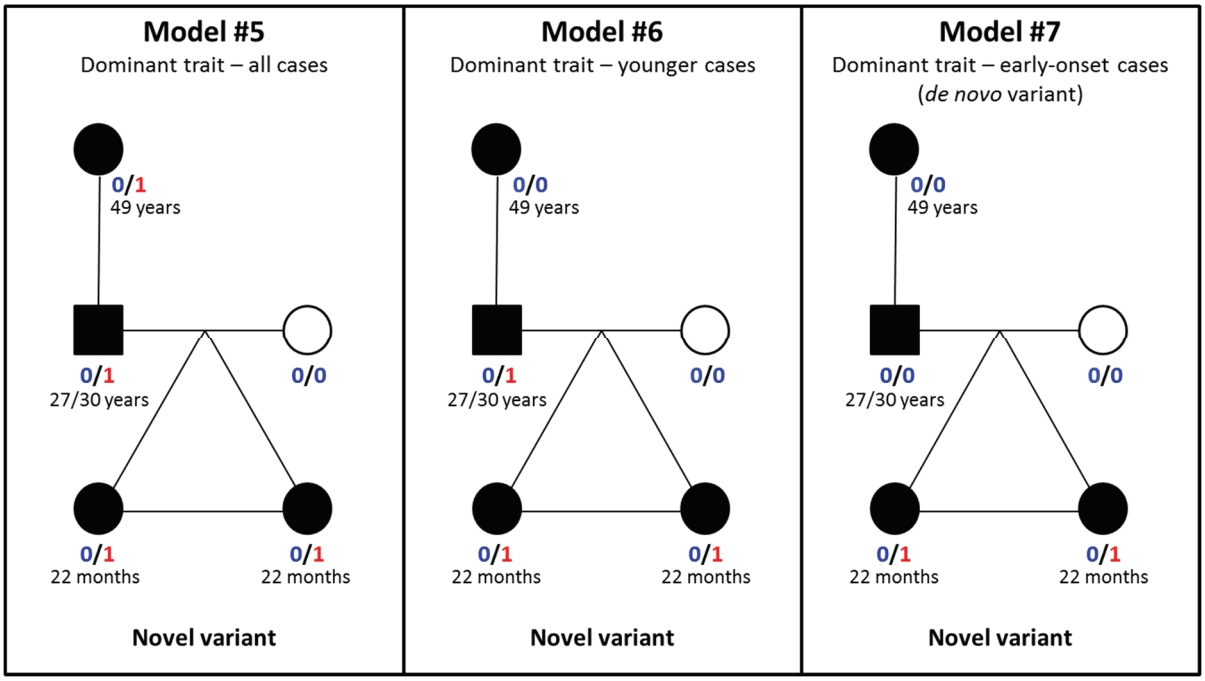


Figure 8. Filtering approaches applied to identify candidate coding functional variants segregating with disease following a dominant trait.
Number zero (in blue) represents the common allele and number one (in red) corresponds to the minor allele. Age-at-diagnosis of leprosy cases are shown in black. MAF: Minor allele frequency; yrs: years.

After whole exome screening, variants in genes previously associated with leprosy were analyzed in more detailed. For that, Huge Navigator database (151) was used for searching entries related to leprosy phenotypes up to February 2016. Variants detected in WGS from the Piauí family within genes included in the Huge Navigator search list were selected. Filtering steps were applied to identify candidate functional variants in these genes as implemented in the mentioned before models, but with a less stringent MAF threshold. For filtering approaches applied on the recessive models, we focus on variants where the allele found in the affected members of the family is the minor allele (< 50%) in all population samples from public databases. Then, in the dominant model (Models #5 to #7), we searched for variants with MAF < 10% in public databases. Finally, the genes that passed the filtering steps were searched in PubMed to confirm that they were, indeed, statistically associated to leprosy. The search term used was “leprosy” AND the gene name.

5.6 CANDIDATE VARIANTS PRIORITIZATION

To evaluate and prioritize the candidate genes, we used *in silico* bioinformatics tools to predict the variants’ impact on the protein structure and function. For that, we used PolyPhen-2 v2.2.2r398 (114) and CADD v 1.3 (117) as variant-level metrics. According to PolyPhen-2 HumVar score, missense variants were classified as i) “benign” (B, score between 0 and 0.446), ii) "possibly damaging" (P, score between 0.447 and 0.908) or iii) "probably damaging" (D, score between 0.909 and 1) (116). We focus on variants classified as possibly or probably damaging. A variant was predicted to be deleterious when presenting CADD score higher than 20, on a scale of 1–99 (152). GDI (as of February 2016) was used as gene-level metric (121). Genes presenting GDI lower than 0.958 were classified as having “Low damage”, GDI between 0.958 and 13.84 were considered “Medium damage” and genes with GDI higher than 13.84 were classified as “High damage” (121,122); we prioritized on variants located in genes with low and medium damage. In addition, involvement of candidate variants with the studied phenotype was searched in PubMed (Key words: “leprosy” AND gene name).

Linkage desequilibium (LD) estimation of selected candidate variants located in the same chromosome region was performed using Haploview software v4.2 (153), based on genotyping data from 1000G (110). Finally, when two or more variants from the same gene were detected as candidate variants, it was inferred the probable haplotypes segregation in the family. Based on these data, the probable genotypes for the grandfather (I-1) for those variants were inferred.

6 RESULTS

6.1 WHOLE EXOME SEQUENCING

6.1.1 Sequencing performance and alignment to human reference

Four WES experiments were performed with Ion Proton™ platform. In each one, nearly 84.5% of the chip's wells were loaded with templated beads. This generated 8.2 to 11 Gb of data *per* experiment (**Figure 9**). In total, it was produced 38.6 Gb of data distributed in nearly 295 million reads (**Figure 9**), where 98% (38.1 Gb) were properly identified as corresponding to one of the samples (**Table 4**). On average, 7.6 Gb of sequence data were produced *per* individual after exome sequencing, ranging from 5.4 Gb for grandmother (I-2) to 10.7 Gb for Twin2 (III-11) (**Table 4**).

Table 4. Summary of WES raw data from Ion Proton™.

Sample	Single-end reads (M)	Read length (bp)			Total data (Gb)
		Minimum	Maximum	Mean	
Grandmother (I-2)	41.9	8	362	128.8	5.4
Father (II-5)	63.0	8	362	128.35	8.1
Mother (II-6)	63.5	8	369	125.3	7.9
Twin1 (III-10)	47.8	8	361	126.07	6
Twin2 (III-11)	78.1	8	360	137.1	10.7
TOTAL	**294.3**				**38.1**

bp: base pair; Gb: Giga bases; M: Mega.

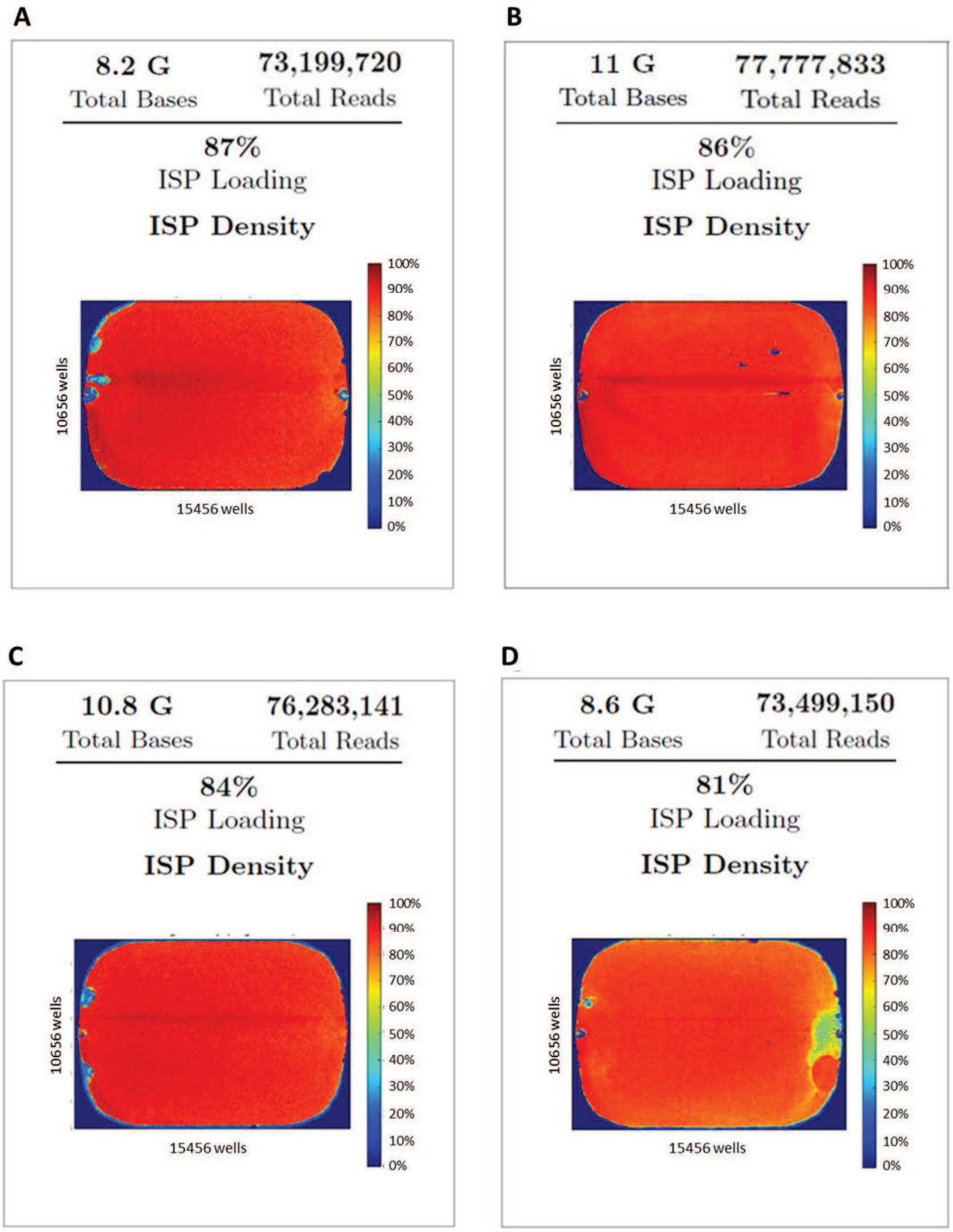


Figure 9. Sequencing chips loading with samples from pool 1 (**A** and **B**), 2 (**C**) and 3 (**D**). Colors tending towards red indicate high density of beads within the wells for a given chip area. Colors tending to blue indicate low bead density.
ISP: Ion Sphere Particles; G: giga.

The read length distribution ranged from 8 to 369 bp (**Table 4**). As expected due to the chemistry used, there is a peak of reads with length around 200bp (**Figure 10-A**). However, there is also an enrichment of shorter reads in all samples (**Figure 10-A**). On average, 71% of the reads had mean quality score higher or equal to Q20

(99% of accuracy). In all samples, 99% of reads presented a quality score between Q16 and Q26, which correspond to an accuracy of 96.5% to 99.75%, respectively (**Figure 10-B**). Finally, base-calling accuracy is maintained similar across the length of the reads, with a small decrease towards the end (**Figure 10-C**).

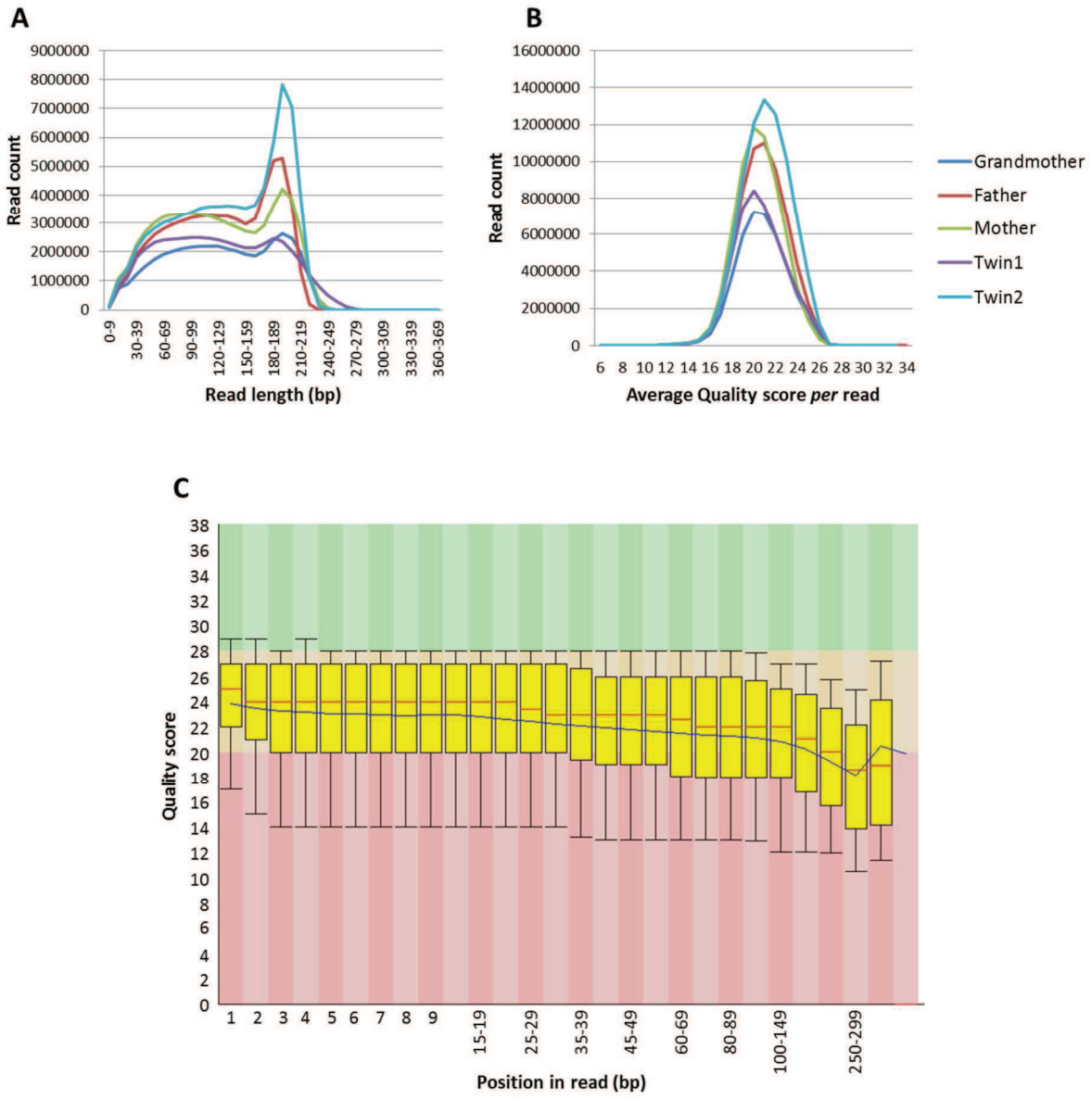


Figure 10. Raw reads length and quality.
Results from FastQC. **A)** Read length distribution **B)** *Per* read mean quality scores from the single-end reads from WES for each sample sequenced. **C)** Box plot for quality scores of the bases according to their position in the reads (Example of one representative sample data). Blue and red lines represent the mean and median quality scores respectively. The yellow box represents the inter-quartile range (25-75%). The upper and lower whiskers represent the 10% and 90% points. The graph background colors divide the quality in three groups: very good quality calls (green), calls of acceptable quality (orange), and calls of poor quality (red) (135).

From the total of 294.3 million reads generated from the five samples, 97.85% were properly aligned to the human reference genome and 71.65% to the target regions (**Table 5**). From the total reads mapped, 36.39% were PCR duplicates and were removed from the analysis. The WES data from the grandmother (I-2) and Twin2 (III-11) samples presented the lowest and highest mean coverage inside the target region, respectively: 24.5 ± 17X for grandmother (I-2) and 46.8 ± 24.9X Twin2 (III-11) (**Table 5**). Those coverage depths are sufficient for variant calling (see section 5.3.2); however, the high standard deviation in all samples indicates that the target bases were not uniformly covered (**Table 5**).

Table 5. WES data alignment to human reference genome and *per* sample mapping details.

Parameters¶	Grandmother (I-2)	Father (II-5)	Mother (II-6)	Twin1 (III-10)	Twin2 (III-11)
Mapped reads	97.77%	97.99%	97.70%	97.61%	98.04%
Unmapped reads	2.23%	2.01%	2.30%	2.39%	1.96%
On-target mapped reads	71.52%	68.68%	67.75%	72.35%	77.97%
Duplicated reads	37.27%	37.58%	34.95%	31.38%	39.21%
Mean coverage*†	24.45X	30.33X	32.17X	30.02X	46.82X
Coverage standard deviation *†	17X	19.49X	20.74X	18.58X	24.87X
Mean Mapping Quality*	55.85	56.35	55.01	55.51	57.41

X: times/folds. ¶ Results from QualiMap. * Inside target regions. † Duplicated reads are ignored.

Figure 11-A shows the fraction of target regions that presented coverage ≥ 1X to ≥ 50X in each sample. Seventy-eight percent (Grandmother, I-2) to 94% (Twin2, III-11) of targeted bases were covered at least 10X and 57% (Grandmother, I-2) to 85% (Twin2, III-11) were covered at least 20 times (**Figure 11-A**). The intersection of regions with coverage ≥ 10X and ≥ 20X among all samples is presented in **Figure 11-B**. For each coverage category, we were able to search for candidate SNVs and indels in 76% and 57% of the target regions, respectively.

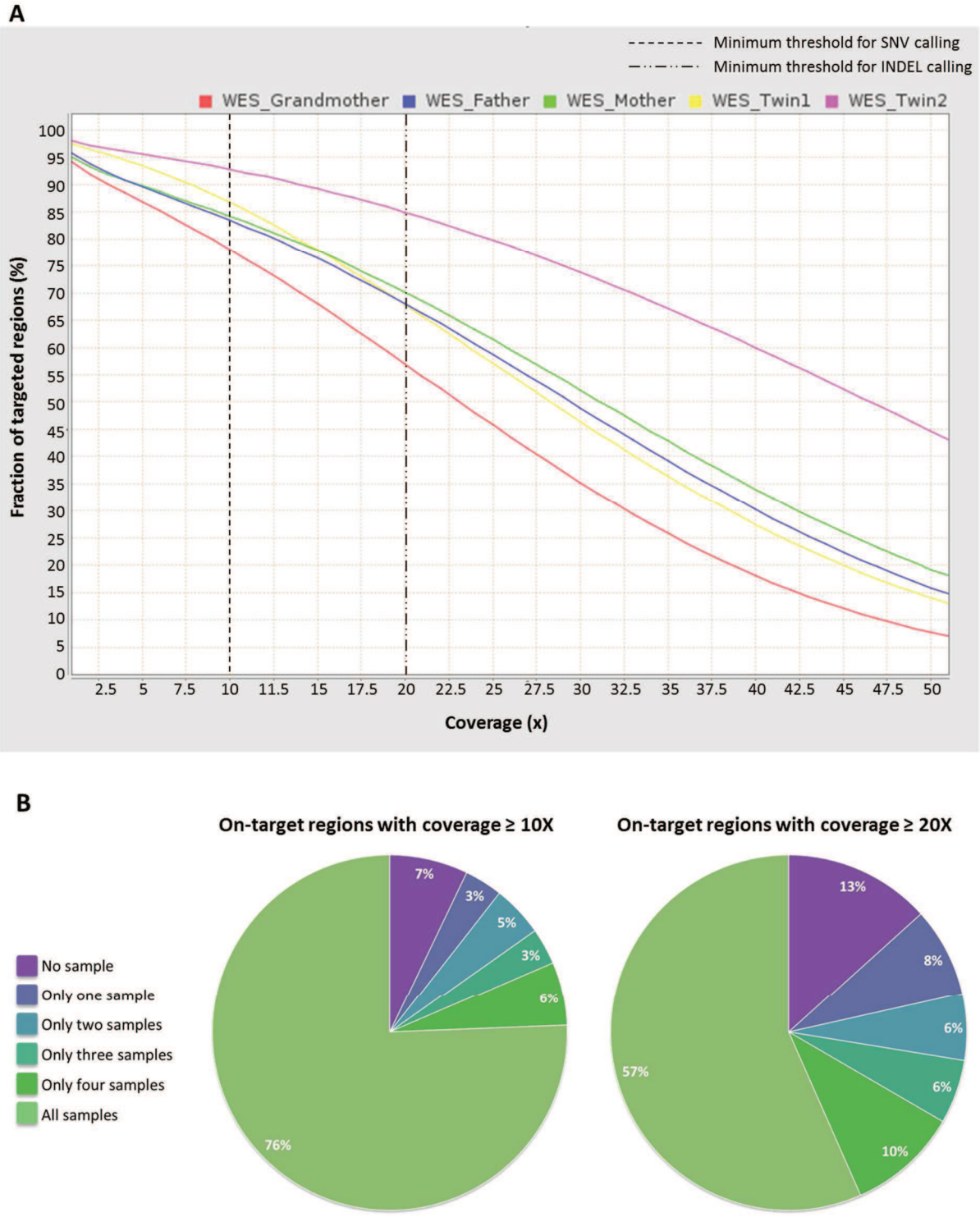


Figure 11. On-target *per* base coverage in WES data.
A) For each sample, fractions of target regions with coverage greater than or equal to 1X to 50X. (Results from QualiMap) **B**) Intersection of target regions with coverage ≥ 10X and ≥ 20X among all samples. Indel: Insertion/deletion; SNV: Single nucleotide variant; WES: Whole exome sequencing; X: times/folds.

6.1.2 Variant identification and filtering

After alignment to human reference genome and variant calling, 140,338 variants were identified, mostly non-coding variants located in target exons' flanking areas and off-target regions (74.5%) (**Table 6**). The remaining 35,784 (25.5%) variants identified were located within exonic (coding) and splice-site regions (**Table 6**). The total number of exonic/splice-site variants detected *per* sample ranged from 18,613 to 22,680 (grandmother [I-2] and Twin2 [III-10], respectively). In all samples, synonymous SNVs were the most common exonic variants, followed by missense variants (**Table 6**).

Table 6. Types of variants identified in WES data from the five samples.

Type of variant*	Number of variants	Percentage[¶]
Non-coding**	**104,554**	**74.5%**
Splice-site	**168**	**0.120%**
Exonic (Coding)	**35,616**	**25.4%**
- Synonymous	17,306	48.6%
- Missense	15,393	43.2%
- Frameshift Indel	1,723	4.8%
- Unknown[#]	839	2.4%
- In frame Indel	191	0.5%
- Nonsense	164	0.5%
TOTAL	**140,338**	**100%**

* Annotation according to wANNOVAR tool.

** Includes intergenic, intronic, ncRNA, UTR, upstream and downstream variants.

Coding loci that present different types of variants in different isoforms of the gene are classified as unknown. In this study, candidate variants with unknown classification by wANNOVAR were manually classificated.

[¶] In non-coding, splice-site and exonic variants, percentages are from total variants. In synonymous to nonsense variants, percentages are from exonic variants.

Indel: Insertion/Deletion.

6.1.2.1 Recessive model

In total, 53 candidate variants in 33 genes passed the filtering steps for the recessive models (models #1 to #4, **Table 7**). From these, one variant is homozygous in all affected family members and not in the unaffected twin's mother (II-6) (model #1), six are homozygous only in the twins (III-10 and III-11) and father (II-5) (model #2), 13 are homozygous only in both twins (III-10 and III-11) (model #3) and 33 variants in 14 genes were found as compound heterozygous in both twins (model #4, **Table 7**). Regarding the variants identified in models #1, #2 and #3, all the 20 missense variants were previously reported in databases. None of them are rare, since they have MAF higher than 1% in at least one population available in public databases (**Supplementary Table S1** in **Appendix 2**).

Table 7. Candidate variants identified in WES analysis that passed variant filtering for a recessive trait (Models #1 to #4).

Chr	Position*	Ref	Alt	Gene	Type of variant	AA Change	rsID
Model #1 (Recessive, MAF ≤ 30%, All affected)							
6	29913037	G	A	*HLA-A*	Missense	V358M	rs1137631
Model #2 (Recessive, MAF ≤ 30%, Father and twins)							
11	499120	G	A	*RNH1*	Missense	P170L	rs17585
12	40657700	C	G	*LRRK2*	Missense	N551K	rs7308720
17	7484101	C	A	*CD68*	Missense	Q254K	rs9901673
17	7490810	G	A	*MPDU1*	Missense	A229T	rs10852891
20	50406630	T	G	*SALL4*	Missense	I798L	rs6091375
X	12903659	A	T	*TLR7*	Missense	Q11L	rs179008
Model #3 (Recessive, MAF ≤ 15%, Only the twins)							
2	141242918	T	C	*LRP1B*	Missense	Q3140R	rs34488772
4	185580557	A	G	*PRIMPOL*	Missense	T82A	rs74696256
6	117710661	T	C	*ROS1*	Missense	I537M	rs28639589
12	40702911	G	A	*LRRK2*	Missense	R1398H	rs7133914
12	103988285	A	G	*STAB2*	Missense	I110V	rs17034186
14	50655307	C	T	*SOS2*	Missense	A208T	rs61755579
16	68395522	C	T	*SMPD3*	Missense	C617Y	rs71395853
17	5354204	G	C	*DHX33*	Missense	H483D	rs11653658
18	22057204	C	G	*HRH4*	Missense	S284C	rs58154316
19	11687351	C	T	*ACP5*	Missense	V148M	rs2305799

19	51503285	C	T	*KLK8*	Missense	V199I	rs16988799
21	43221797	G	C	*PRDM15*	Missense	T1376S	rs2236695
22	42607817	C	T	*TCF20*	Missense	M1165I	rs17002890
Model #4 (Recessive, Compound heterozygous, MAF ≤ 5%, Only the twins)							
1	156518421	C	T	*IQGAP3*	Missense	G649S	rs77834544
	156521798	T	A		Missense	Q513L	-
2	179702426	C	G	*CCDC141*	Missense	E1174Q	rs75153675
	179839888	G	A		Missense	A141V	rs10497529
4	128608951	G	A	*INTU*	Missense	D460N	-
	128627927	T	G		Missense	C692G	rs34311863
4	186545346	A	T	*SORBS2*	Missense	L509I	rs61736043
	186599973	C	T		Missense	R36H	rs190199282
4	187524714	C	T	*FAT1*	Missense	V3656I	rs192691397
	187530423	T	C		Missense	I3374V	rs138364727
	187540374	C	T		Missense	A2456T	rs370340394
	187628947	C	T		Missense	V679I	rs61733571
5	54468432	CTTCT	-	*CDC20B*	Frameshift deletion	R36fs	rs137940833
	54468450	T	C		Missense	D31G	rs138811807
5	118469561	G	A	*DMXL1*	Missense	V648I	rs139365266
	118485204	G	A		Missense	V1228M	rs140855219
6	51712759	T	C	*PKHD1*	Missense	T2641A	rs7766366
	51875133	G	A		Missense	R1909W	rs115338476
	51890265	T	C		Missense	E1448G	rs116809571
	51917987	G	C		Missense	P676R	rs115045643
6	54173421	T	G	*TINAG*	Missense	S25A	rs34700914
	54214618	C	T		Missense	T335M	rs139989527
9	135470281	C	T	*DDX31*	Missense	R843Q	rs306548
	135538016	C	T		Missense	E153K	rs17402080
10	105183348	T	C	*PDCD11*	Missense	V899A	rs61751511
	105201712	G	A		Missense	E1563K	-
11	62886706	C	T	*SLC22A24*	Missense	R203H	rs116063135
	62910891	C	A		Missense	V121L	rs116409312
12	120574343	G	A	*GCN1*	Missense	A2324V	rs201840533
	120574344	C	A		Missense	A2324S	-
	120613593	G	A		Missense	S333L	rs114251901
14	105414032	G	T	*AHNAK2*	Missense	L2586I	rs199905726
	105414053	T	C		Missense	M2579V	rs200965573
	105419551	C	G		Missense	G746A	rs201524595
16	16259497	G	T	*ABCC6*	Missense	L1097I	rs60707953
	16259722	G	C		Missense	Q1022E	rs57179857

*GRCh37/hg19.

AA: amino acid; Alt: Alternative allele; Chr: Chromosome; Ref: Reference allele.

6.1.2.2 Dominant model

When variant filtering was applied for the dominant model, 29 novel functional mutations were identified (models #5, #6 and #7 – **Table 8**). Eight missense SNVs were uniquely found as heterozygous in the leprosy affected members of the family (model #5). The twins (III-10 and III-11) and father (II-5) carried 18 missense mutation, one stop-gain variant and one frameshift deletion that were not identified in the grandmother (I-2) and twin's mother (II-6) (model #6). Finally, one *de novo* missense variant was identified only in the twin girls (III-10 and III-11) (model #7).

Table 8. Novel variants identified in the WES analysis that passed the variant filtering for the dominant model (Models #5, #6 and #7).

Chr	Position*	Ref	Alt	Gene	Type of variant	AA Change
Model # 5 (Dominant, novel variants, All affected)						
3	73114043	C	T	*PPP4R2*	Missense	P227S
3	148905977	T	C	*CP*	Missense	K576E
4	77816977	G	C	*SOWAHB*	Missense	L676V
4	119947864	G	A	*SYNPO2*	Missense	E114K
5	35957503	C	T	*UGT3A1*	Missense	A288T
6	119147418	G	A	*MCM9*	Missense	T618I
17	62130282	T	C	*ERN1*	Missense	K704R
X	152113882	C	G	*ZNF185*	Missense	A459G
Model #6 (Dominant, novel variants, Father and twins)						
1	84650810	G	T	*PRKACB*	Stop-gain	E122X
1	156521798	T	A	*IQGAP3*	Missense	Q513L
1	176833514	T	G	*ASTN1*	Missense	D1264A
1	207785097	A	G	*CR1*	Missense	E1674G
2	186666910	C	G	*FSIP2*	Missense	H4382D
3	47037443	G	A	*NBEAL2*	Missense	V685M
4	4199651	T	A	*OTOP1*	Missense	M304L
10	105201712	G	A	*PDCD11*	Missense	E1563K
15	43627339	A	G	*ADAL*	Missense	T12A
15	55670559	C	A	*CCPG1*	Missense	G64V
15	64737241	G	A	*TRIP4*	Missense	V538I
16	24583228	C	T	*RBBP6*	Missense	P1614L
16	53515590	C	T	*RBL2*	Missense	A1031V

18	59217360	G	A	*CDH20*	Missense	D600N
19	18507073	G	A	*LRRC25*	Missense	P234L
19	41916587	C	T	*BCKDHA*	Missense	P52S
19	55964727	G	-	*ISOC2*	Frameshift deletion	P119fs
19	57301244	G	A	*ZIM2*	Missense	S158F
20	1434931	T	A	*NSFL1C*	Missense	Y155F
22	50315340	C	T	*CRELD2*	Missense	P175S
Model #7 (Dominant, novel variants, Only the twins)						
6	17625023	C	A	*NUP153*	Missense	A1346S

*GRCh37/hg19.
AA: amino acid; Alt: Alternative allele; Chr: Chromosome; Ref: Reference allele.

6.1.3 Validation – Sanger sequencing

Ten candidate variants have been selected for validation by Sanger sequencing (all six variants from model #2, three SNPs from model #3 and a novel mutation from model #5). For all of them, Sanger sequencing results were concordant with the genotypes identified in WES data analysis (variants 1 to 10 in **Table 9**). **Figures S1** to **S10** of the supplementary data (**Appendix 2**) show WES and Sanger results – as well as the electropherogram – from these candidate variants.

Moreover, eight non-candidate variants were also selected to be validated (variants 11 to 18 in **Table 9**). From these, variants 11 and 12 presented different genotypes across the twin pair (III-10 and III-11) in the WES data. When analyzing the region using IGV software, it was observed that the WES data from Twin1 (III-10) sample presented lower coverage and quality in both regions as compared to Twin2 (III-11) data (data not shown). For both variants, Sanger results were discordant with WES data only in Twin1 (III-10) (**Table 9**); discordant genotype between the twins in WES was due to incorrect genopype in the sample with lower coverage. Variant 17 was selected as an example of a variant identified in a region where all samples presented coverage near the minimum threshold for SNV. According to Sanger sequencing, the genotype of the five samples were correctly detected by WES (**Table 9**). Variant 15 was selected to be validated since it was a novel frameshift deletion identified in all samples. The variant was located in a homopolymer region and indels

in these regions are the most common type of artifact in NGS data from Ion Torrent platforms (154). As expected, the deletion was not validated (**Table 9**). Variant 16 is a known insertion that, according to IGV analysis, was expected to be correctly genotyped. However, Sanger validation showed a different genotype for the mother's sample. On the other hand, variant 18 is a known deletion not in homopolymer region, which was identified and correctly genotyped in WES data analysis (**Table 9**). Taken together, these results suggests that our WES data: i) presents correct SNV identifications even in regions with low coverage (i.e. variant 17); however, the genotypes of these variants may be incorrect in some samples in those regions (i.e. variants 11 and 12); ii) presents correct indels identification (i.e variant 18) but also artifacts or miss-genotyped indels (i.e. variants 15 and 16).

Table 9. Sanger sequencing validation for 18 selected variants identified in WES analysis.

#	Gene	AA change	rsID	Sanger sequencing validation*				
				Grandma	Father	Mother	Twin1	Twin2
1	*CP*	K576E	-	conc	conc	conc	conc	conc
2	*RNH1*	P170L	rs17585	conc	conc	conc	conc	conc
3	*LRRK2*	N551K	rs7308720	conc	conc	conc	conc	conc
4	*LRRK2*	R1398H	rs7133914	conc	conc	conc	conc	conc
5	*SOS2*	A208T	rs61755579	conc	conc	conc	conc	conc
6	*CD68*	Q254K	rs9901673	conc	conc	conc	conc	conc
7	*MPDU1*	A229T	rs10852891	conc	conc	conc	conc	conc
8	*HRH4*	S284C	rs58154316	conc	conc	conc	conc	conc
9	*SALL4*	I798L	rs6091375	conc	conc	conc	conc	conc
10	*TLR7*	Q11L	rs179008	conc	conc	conc	conc	conc
11	*TREML2*	S144G	rs3747742	conc	conc	conc	**disc**	conc
12	*HOXA7*	A18T	rs2301721	conc	conc	conc	**disc**	conc
13	*LRRK2*	Q1111H	rs78365431	conc	conc	conc	conc	conc
14	*LRRK1*	G1938E	rs386787404	conc	conc	conc	conc	conc
15	*ZZEF1*	P2462fs	-	**disc**	**disc**	**disc**	**disc**	**disc**
16	*LRRC59*	In UTR3	rs145265135	conc	conc	**disc**	conc	conc
17	*HAUS5*	Q580R	rs61747965	conc	conc	conc	conc	conc
18	*ZNF480*	L2fs	rs398035013	conc	conc	conc	conc	conc

* Concordant and discordant genotype results between WES and Sanger sequencing are shown as "conc" and "disc" respectively.

conc: concordant; disc: discordant; SNV: Single Nucleotide Variant; UTR3: 3' untranslated region.

6.2 WHOLE GENOME SEQUENCING

6.2.1 Sequencing raw data and alignment to reference

For the WGS experiment, 2.7 billions pair-end reads were obtained from the six sequenced samples, totalizing 827 Gb of data. In these pair-end reads, Read 1 (R1) and Read 2 (R2) (**Figure 3-A**) were 151 bp long each (**Table 10**). Data volume generated *per* sample ranged from 106 Gb to 159 Gb in father (II-5) and Twin2 (III-11), respectively (**Table 10**).

Table 10. Summary of WGS raw data from HiSeq® 2500.

Sample	Pair-end Reads (M)	Read length (bp)	Total data (Gb)
Grandmother (I-2)	471	2x151	142
Aunt (II-4)	417	2x151	126
Father (II-5)	350	2x151	106
Mother (II-6)	521	2x151	157
Twin1 (III-10)	438	2x151	132
Twin2 (II-11)	527	2x151	159
TOTAL	**2,724**		**823**

bp: base pair; Gb: Giga bases; M: Mega.

Figure 12-A shows an example of the quality scores obtained for a pair of sequencing data (R1 and R2). Although R2 has lower mean quality than R1 in all samples, the vast majority of reads in both groups had mean quality score higher or equal to Q20 (99% of accuracy). **Figure 12-B** presents base calling quality score throughout the reads length in R1 (left) and R2 (right). The quality score decreases toward the end of the reads in both groups, but its mean and median are maintained higher than Q20 through the reads length.

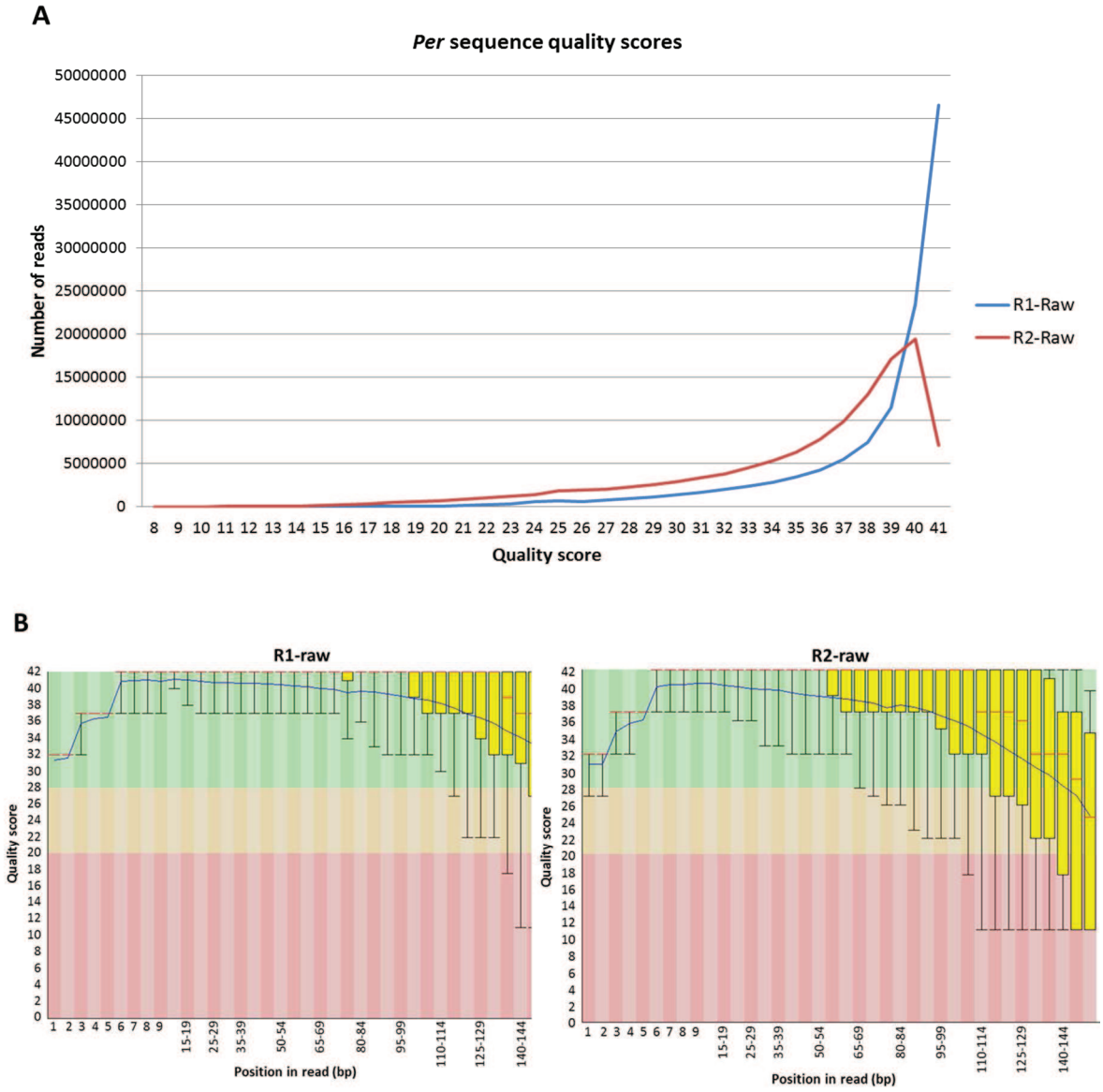


Figure 12. Raw reads length and quality.
Results from FastQC. **A)** *Per* read mean quality scores of the raw pair-end reads from WGS (example of one representative sample data). **B)** Box plot of quality scores of the bases according to their position in R1 (left) and R2 (right) reads (example of one representative sample data). Blue and red lines represent the mean and median quality scores respectively. The yellow box represents the inter-quartile range (25-75%). The upper and lower whiskers represent the 10% and 90% points. The graph background colors divide the quality in three groups: very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red) (135). Bp: base pair; R: read.

From the total reads generated, 99.18% were properly aligned to human reference genome (**Table 11**). Only 4.4% were overlapping read (where R2 overlaps to R1 from the same pair-end read). One of the overlapping reads was removed from the analysis as well as the PCR duplicates (11.62%). We were able to produce *per*

base mean coverage of 30.95 ± 11.8X in all samples, ranging from ~25 ± 10X in the father (II-5) sample to ~35 ± 13X in Twin2 (III-11) (**Table 11**). **Figure 13** shows the fraction of the genome that displayed coverage ≥ 1 to ≥ 50X in each sample. In all samples, 91.5% of the genome were covered at least 10X (**Figure 13**).

Table 11. WGS data alignment to human reference genome and *per* sample mapping details.

Parameters¶	Grandmother	Aunt	Father	Mother	Twin1	Twin2
Mapped reads	99.09%	98.85%	99.20%	99.42%	99.12%	99.41%
Unmapped reads	0.91%	1.15%	0.80%	0.58%	0.88%	0.59%
Overlapping read pairs	4.06%	5.84%	5.08%	3.78%	4.05%	3.80%
Duplicated reads	12.22%	12.20%	11.11%	12.11%	10.82%	11.26%
Mean coverage (X)*	32.03	28.39	24.87	34.63	30.33	35.44
Coverage standard deviation (X)*	12.08	10.99	10.09	12.89	11.59	13.14
Mean Mapping Quality	50.98	51	51.17	51.06	51.02	51.04

X: times/folds. ¶ Results from QualiMap. * Paired-end reads overlap and duplicated reads are ignored.

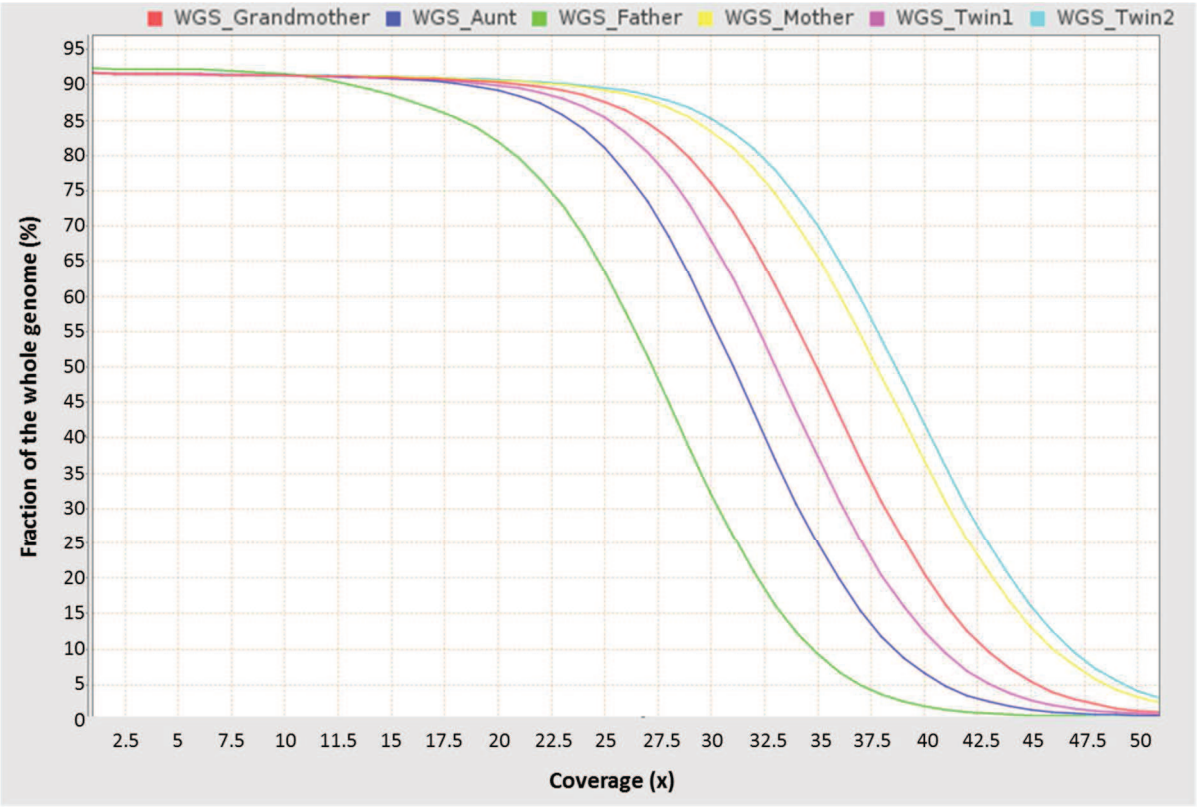


Figure 13. Fractions of the whole genome with coverage depth ≥ 1X to 50X *per* sample. WGS: whole genome sequencing; X: times/folds.

6.2.2 Variant identification and filtering

Upon alignment of the WGS reads to human reference genome, 8,375,621 variants were identified. From these, 8,330,942 (99.5%) were non-coding variants and 44,679 (0.5%) were exonic (coding) and splice-site variants (**Table 12**). On average, 26,281 exonic/splice-site variants were identified *per* sample, ranging from 25,828 in mother (II-6) to 27,560 in grandmother (I-2). As in WES, the most common exonic variants detected in all samples were synonymous SNVs and missense variants (**Table 12**).

Table 12. Types of variants identified in WGS data from the six samples.

Type of variant*	Number of variants	Percentage[¶]
Non-coding**	**8,330,942**	**99.5%**
Splice-site	**291**	**0.003%**
Exonic (Coding)	**44,388**	**0.5%**
- Synonymous	20,871	47.0%
- Missense	19,320	43.5%
- Unknown[#]	3,122	7.0%
- In frame Indel	588	1.3%
- Frameshift Indel	296	0.7%
- Nonsense	191	0.4%
TOTAL	**8,375,621**	**100%**

Indel: Insertion/Deletion. * Annotation according to wANNOVAR tool.

** Includes intergenic, intronic, ncRNA, UTR, upstream and downstream variants.

Coding *loci* that present different types of variants in different isoforms of the gene are classified as unknown. In this study, candidate variants with unknown classification by wANNOVAR were manually classificated.

[¶] In non-coding, splice-site and exonic variants, percentages are from total variants. In synonymous to nonsense variants, percentages are from exonic variants.

6.2.2.1 Recessive model

Once the variants identified in the WGS analysis were annotated, we performed the same procedure of variant filtering used in the WES data analysis (section 5.5).

Then, results for each model were compared with WES results. The candidate variants identified from the WGS data filtering assuming a recessive trait (Models #1 to #4) are shown in **Table 13**. All candidate variants identified in the WES data analysis for models #1, #2 and #3 (**Table 7**), were also detected in the WGS analysis under the same models (**Table 13**). There is no additional candidate variant in Model #1 and five candidate variants were included from the WGS data in model #2 and five in model #3 (**Table 13**). These 10 variants were also previously reported on public databases and present MAF higher than 1% in at least one population sample (**Supplementary Table S1** in **Appendix 2**). The aunt (II-4) is not homozygous for any of the 31 candidate variants identified in these recessive models.

Regarding Model #4 (compound heterozygous in the twins), 66 variants in 25 genes passed filtering steps for this model from WGS data (**Table 13**). Among these, all variants in the 15 genes identified from Model #4 in WES data (**Table 7**) are included, indicating that WGS analysis also allow us to validate WES results for this model. Moreover, variants in 10 additional genes were found as candidate in this model only from WGS data (**Table 13**).

Table 13. Candidate variants identified in the WGS analysis that passed variant filtering for the recessive model (Models #1 to #4).

Chr*	Position¶	Ref	Alt	Gene	Type of variant	AA Change	rsID
Model #1 (Recessive, MAF ≤ 30%, All affected)							
6	29913037	G	A	*HLA-A*	Missense	V358M	rs1137631
Model #2 (Recessive, MAF ≤ 30%, Father and twins)							
6	29910719	G	A	*HLA-A*	Missense	E87K	rs2230991
6	29910721	G	C	*HLA-A*	Missense	E87D	rs199474424
6	30039418	C	T	*RNF39*	Missense	A245T	rs2301752
10	24813454	G	A	*KIAA1217*	Missense	A887T	rs10828663
11	499120	G	A	*RNH1*	Missense	P170L	rs17585
12	40657700	C	G	*LRRK2*	Missense	N551K	rs7308720
17	4722876	G	A	*PLD2*	Missense	G821S	rs3764897
17	7484101	C	A	*CD68*	Missense	Q254K	rs9901673
17	7490810	G	A	*MPDU1*	Missense	A229T	rs10852891
20	50406630	T	G	*SALL4*	Missense	I798L	rs6091375
X	12903659	A	T	*TLR7*	Missense	Q11L	rs179008
Model #3 (Recessive, MAF ≤ 15%, Only the twins)							
1	227843003	G	A	*ZNF678*	Missense	C406Y	rs61740826
2	141242918	T	C	*LRP1B*	Missense	Q3140R	rs34488772

4	185580557	A	G	*PRIMPOL*	Missense	T82A	rs74696256
6	117710661	T	C	*ROS1*	Missense	I537M	rs28639589
12	40702911	G	A	*LRRK2*	Missense	R1398H	rs7133914
12	103988285	A	G	*STAB2*	Missense	I110V	rs17034186
14	50655307	C	T	*SOS2*	Missense	A208T	rs61755579
16	68395522	C	T	*SMPD3*	Missense	C617Y	rs71395853
17	5354204	G	C	*DHX33*	Missense	H483D	rs11653658
18	22057204	C	G	*HRH4*	Missense	S284C	rs58154316
19	11687195	C	T	*ACP5*	Missense	V200M	rs2229531
19	11687351	C	T	*ACP5*	Missense	V148M	rs2305799
19	13211843	C	T	*LYL1*	Missense	R48Q	rs117072928
19	51503285	C	T	*KLK8*	Missense	V199I	rs16988799
20	61512185	G	C	*DIDO1*	Missense	S1708C	rs41282984
20	61598731	C	T	*SLC17A9*	Missense	T397M	rs7271712
21	43221797	G	C	*PRDM15*	Missense	T1376S	rs2236695
22	42607817	C	T	*TCF20*	Missense	M1165I	rs17002890
Model #4 (Recessive, Compound heterozygous, MAF ≤ 5%, Only the twins)							
1	5937289	T	C	*NPHP4*	Missense	H894R	rs113097479
	5965440	T	C		Missense	T623A	rs35959882
1	11826186	G	T	*C1orf167*	Missense	A115S	rs145919329
	11826351	G	A		Missense	G170R	rs144306270
	11826421	A	G		Missense	H193R	rs187345781
	11826861	A	C		Missense	N340H	rs188115585
	11827071	C	T		Missense	R410W	rs189594838
	11827114	AC	-		Frameshift deletion	D424fs	rs141465134
	11847932	T	C		Missense	C1226R	rs116698217
1	156518421	C	T	*IQGAP3*	Missense	G649S	rs77834544
	156521798	T	A		Missense	Q513L	-
2	179702426	C	G	*CCDC141*	Missense	E1174Q	rs75153675
	179839888	G	A		Missense	A141V	rs10497529
4	128608951	G	A	*INTU*	Missense	D460N	-
	128627927	T	G		Missense	C692G	rs34311863
4	186545346	A	T	*SORBS2*	Missense	L509I	rs61736043
	186599973	C	T		Missense	R36H	rs190199282
4	187524714	C	T	*FAT1*	Missense	V3656I	rs192691397
	187530423	T	C		Missense	I3374V	rs138364727
	187540374	C	T		Missense	A2456T	rs370340394
	187628947	C	T		Missense	V679I	rs61733571
5	54468432	CTTCT	-	*CDC20B*	frameshift deletion	R36fs	rs137940833
	54468450	T	C		Missense	D31G	rs138811807
5	112043492	C	A	*APC*	Missense	S26R	rs113782655
	112102943	T	A		Missense	L103H	-
	112102945	C	G		Missense	R104G	-
5	118469561	G	A	*DMXL1*	Missense	V648I	rs139365266
	118485204	G	A		Missense	V1228M	rs140855219
6	32029369	C	T	*TNXB*	Missense	V2433I	rs200135227
	32049373	C	T		Missense	V1272M	-

	32063558	C	T		Missense	G691D	rs201146825
6	51712759	T	C	*PKHD1*	Missense	T2641A	rs7766366
	51875133	G	A		Missense	R1909W	rs115338476
	51890265	T	C		Missense	E1448G	rs116809571
	51917987	G	C		Missense	P676R	rs115045643
6	54173421	T	G	*TINAG*	Missense	S25A	rs34700914
	54214618	C	T		Missense	T335M	rs139989527
9	135470281	C	T	*DDX31*	Missense	R843Q	rs306548
	135538016	C	T		Missense	E153K	rs17402080
6	65300527	C	T	*EYS*	Missense	D1745N	rs145274061
	66005791	C	T		Missense	G663E	-
8	8176654	G	C	*SGK223*	Missense	H1077Q	rs2011560
	8235555	G	T		Missense	L122I	rs55764617
10	105183348	T	C	*PDCD11*	Missense	V899A	rs61751511
	105201712	G	A		Missense	E1563K	.
11	62886706	C	T	*SLC22A24*	Missense	R203H	rs116063135
	62910891	C	A		Missense	V121L	rs116409312
12	120574343	G	A	*GCN1*	Missense	A2324V	rs201840533
	120574344	C	A		Missense	A2324S	-
	120613593	G	A		Missense	S333L	rs114251901
13	33590898	C	T	*KL*	Missense	A107V	rs115511178
	33590971	C	A		Missense	N131K	rs79554512
14	105414032	G	T	*AHNAK2*	Missense	L2586I	rs199905726
	105414053	T	C		Missense	M2579V	rs200965573
	105419551	C	G		Missense	G746A	rs201524595
16	16259497	G	T	*ABCC6*	Missense	L1097I	rs60707953
	16259722	G	C		Missense	Q1022E	rs57179857
16	88496285	G	T	*ZNF469*	Missense	A803S	rs113484918
	88496544	C	T		Missense	A889V	rs145186655
	88504208	G	C		Missense	G3416R	rs569602115
	88504673	G	T		Missense	A3571S	rs199760004
17	29162972	G	C	*ATAD5*	Missense	A625P	rs144812489
	29187582	C	T		Missense	L1030F	rs35910070
17	36485883	C	T	*GPR179*	Missense	R1190Q	rs80172972
	36487060	C	T		Missense	A798T	rs78470373

* Candidate variants that were not previously detected in the WES analysis are highlighted in grey.
¶ GRCh37/hg19.
AA: amino acid; Alt: Alternative allele; Chr: Chromosome; Ref: Reference allele.

6.2.2.2 Dominant model

As for the recessive model, the same filtering steps used in WES analysis were applied in WGS data for the dominant model also (Models #5, #6 and #7). The 29

novel mutations already detected in WES data (**Table 8**) also passed WGS data filtering (**Table 14**). These variants were detected under the same WES model except for D600N in *CDH20* gene. In WES, this mutation was found only in the father (II-5) and twin girls (III-10 and III-11) (Model #6 – **Table 8**), while in WGS data it was also genotyped in grandmother (I-2) (Model #5 – **Table 14**). From the WGS data, eight variants were added to the list of candidate mutations for the dominant model: five were found in all affected members of the family and three were detected in the father (II-5) and twins (III-10 and III-11) (Models #5 and #6 respectively, **Table 14**). The aunt (II-4) carries six of the candidate mutations from the dominant model, which are located in *RYK, CP, SYNPO2, MCM9*, *CABLES1* and *CDH20*. These six mutations were identified under Model #5 since they were found in all affected members of the family while absent in the unaffected mother (II-6) (**Table 14**).

Table 14. Novel variants identified in the WGS analysis that passed the variant filtering for the dominant model (Models #5, #6 and #7).

Chr*	Position**	Ref	Alt	Gene	Variation	AA Change
Model # 5 (Dominant, novel variants, All affected)						
3	73114043	C	T	*PPP4R2*	Missense	P227S
3	133969413	-	GCGGCG	*RYK*¶	In frame insertion	L26insPP
3	148905977	T	C	*CP*¶	Missense	K576E
4	77816977	G	C	*SOWAHB*	Missense	L676V
4	119947864	G	A	*SYNPO2*¶	Missense	E114K
5	35957503	C	T	*UGT3A1*	Missense	A288T
6	119147418	G	A	*MCM9*¶	Missense	T618I
11	27384674	C	A	*CCDC34*	Missense	R23I
12	65563747	G	A	*LEMD3*	Missense	G124D
17	62130282	T	C	*ERN1*	Missense	K704R
18	20715728	-	GGC	*CABLES1*¶	In frame insertion	M1delinsMA
18	**59217360**	**G**	**A**	***CDH20***¶	**Missense**	**D600N**
X	48564979	C	T	*SUV39H1*	Missense	R356W
X	152113882	C	G	*ZNF185*	Missense	A459G
Model #6 (Dominant, novel variants, Father and twins)						
1	84650810	G	T	*PRKACB*	Stop-gain	E122X
1	156521798	T	A	*IQGAP3*	Missense	Q513L
1	176833514	T	G	*ASTN1*	Missense	D1264A
1	207785097	A	G	*CR1*	Missense	E1674G
2	186666910	C	G	*FSIP2*	Missense	H4382D
3	47037443	G	A	*NBEAL2*	Missense	V685M
4	4199651	T	A	*OTOP1*	Missense	M304L
9	130830773	G	A	*SLC25A25*	Missense	V59I

10	105201712	G	A	*PDCD11*	Missense	E1563K
15	43627339	A	G	*ADAL*	Missense	T12A
15	55670559	C	A	*CCPG1*	Missense	G64V
15	64737241	G	A	*TRIP4*	Missense	V538I
16	24583228	C	T	*RBBP6*	Missense	P1614L
16	53515590	C	T	*RBL2*	Missense	A1031V
19	18507073	G	A	*LRRC25*	Missense	P234L
19	41916587	C	T	*BCKDHA*	Missense	P52S
19	50156323	G	A	*SCAF1*	Missense	G893S
19	50156330	C	A	*SCAF1*	Missense	T895N
19	55964727	G	-	*ISOC2*	Frameshift deletion	P119fs
19	57301244	G	A	*ZIM2*	Missense	S158F
20	1434931	T	A	*NSFL1C*	Missense	Y155F
22	50315340	C	T	*CRELD2*	Missense	P175S
Model #7 (Dominant, novel variants, Only the twins)						
6	17625023	C	A	*NUP153*	Missense	A1346S

* Candidate variants that were not previously detected in the WES analysis are highlighted in grey. A variant filtered in different models in WES vs WGS data is shown in bold.
**GRCh37/hg19.
¶ The twin's aunt (II-4) also carries the same mutation in this gene as the affected members of the family.

6.3 PRIORITIZATION OF CANDIDATE VARIANTS

6.3.1 Recessive model

Table 15 presents computational prediction of functional impact for the variants identified in models #1, #2 and #3 together. From the 30 variants in these models, only six are predicted to be damaging for protein function according to PolyPhen-2, present high CADD scores and are located in genes medium-GDI scores (**Table 15**). These candidate variants are: rs2229531 (V200M) in *ACP5,* rs58154316 (S284C) in *HRH4,* rs7308720 (N551K) and rs7133914 (R1398H) in *LRRK2,* rs7271712 (T397M) in *SLC17A9,* and rs61740826 (C406Y) in *ZNF678* (**Tables 13** and **15**).

Table 15. Gene and variant-level metrics of functional impact prediction of candidate variants from recessive models #1, #2 and #3.

Model*	Chr	Gene	GDI**		AA Change	Scaled CADD**	PolyPhen-2**	
			Score	Damage			Score	Prediction¶
#3	20	*SLC17A9*	**5.14**	**Medium**	T397M	**28.3**	**0.619**	**P**
#2	12	*LRRK2*	**11.25**	**Medium**	N551K	**27.3**	**0.972**	**D**
#3	4	*PRIMPOL*	**6.31**	**Medium**	T82A	**25.5**	0.368	B
#3	19	*ACP5*	**5.82**	**Medium**	V200M	**24.3**	**0.731**	**P**
#3	1	*ZNF678*	**3.46**	**Medium**	C406Y	**24.1**	**0.998**	**D**
#3	22	*TCF20*	**7.94**	**Medium**	M1165I	**23.2**	0.004	B
#3	12	*LRRK2*	**11.25**	**Medium**	R1398H	**23.1**	**0.566**	**P**
#3	14	*SOS2*	**3.57**	**Medium**	A208T	**23**	0.246	B
#2	17	*CD68*	**6.44**	**Medium**	Q254K	**22.7**	0.009	B
#3	6	*ROS1*	17.08	High	I537M	**21.5**	0.051	B
#3	18	*HRH4*	**5.51**	**Medium**	S284C	**21**	**0.926**	**D**
#2	17	*MPDU1*	**8.03**	**Medium**	A229T	**20.9**	0.043	B
#2	17	*PLD2*	13.97	High	G821S	19.65	0	B
#2	11	*RNH1*	**5.85**	**Medium**	P170L	15.88	0.002	B
#3	19	*ACP5*	**5.82**	**Medium**	V148M	14.04	0.381	B
#2	10	*KIAA1217*	**8.49**	**Medium**	A887T	14.02	0.022	B
#3	17	*DHX33*	**8.49**	**Medium**	H483D	13.96	0	B
#2	6	*HLA-A*	34.46	High	E87D	13.61	0.123	B
#3	21	*PRDM15*	**5.76**	**Medium**	T1376S	13.18	0.158	B
#3	16	*SMPD3*	**3.63**	**Medium**	C617Y	12.73	0.001	B
#3	12	*STAB2*	**12.45**	**Medium**	I110V	12.53	0.006	B
#2	6	*HLA-A*	34.46	High	E87K	10.82	0.032	B
#3	19	*KLK8*	**3.58**	**Medium**	V154I	9.18	0.04	B
#1	6	*HLA-A*	34.46	High	V358M	6.95	0.114	B
#2	6	*RNF39*	**12.20**	**Medium**	A245T	6.90	0.227	B
#3	20	*DIDO1*	**4.25**	**Medium**	S1708C	5.37	0.001	B
#3	19	*LYL1*	**2.62**	**Medium**	R48Q	3.68	0	B
#2	20	*SALL4*	**11.17**	**Medium**	I798L	0.177	0.002	B
#3	2	*LRP1B*	**10.88**	**Medium**	Q3140R	0.007	0	B
#2	X	*TLR7*	**0.95**	**Medium**	Q11L	0.003	0	B

* Prioritized variants are highlight in blue: these variants present CADD > 20, PolyPhen-2 > 0.446 (if missense SNV) and are located in genes with GDI < 13.8.

** The variants are sorted according to their CADD score (from highest to lowest). GDI < 13.84 (genes with low and medium damage in general populations), CADD > 20 and PolyPhen-2 > 0.446 (possibly and probably damaging variants) are shown in bold.

¶ B: Benign, P: Possibly damaging, D: Probably damaging.

AA: amino acid; CADD Combined annotation dependent depletion; Chr: Chromosome; GDI: Gene damage index.

Under recessive Model #4, none of the candidate genes have two variants with high CADD score (**Table 16**). Therefore, none of the compound heterozygous variants were prioritized as both being deleterious.

Table 16. Gene and variant-level metrics of functional impact prediction for candidate variants from recessive model #4 (compound heterozygous).

Chr	Gene	GDI*		AA Change	Scaled CADD*	PolyPhen-2*	
		Score	Damage			Score	Prediction¶
2	*CCDC141*	**9.501**	**Medium**	A141V	**26.5**	**0.996**	**D**
				E1174Q	16.53	0.22	B
4	*FAT1*	24.105	High	V679I	**24.9**	**0.99**	**D**
				I3374V	17.7	**0.994**	**D**
				V3656I	10.87	0.004	B
				A2456T	0.008	0.002	B
6	*TINAG*	**9.413**	**Medium**	T335M	**23.7**	**0.984**	**D**
				S25A	1.827	0.232	B
10	*PDCD11*	18.504	High	V899A	**23.4**	**0.888**	**P**
				E1563K	18.4	0.075	B
5	*APC*	**4.177**	**Medium**	L103H	**22.7**	**0.497**	**P**
				R104G	18.26	0.03	B
				S26R	18.06	.	.
5	*DMXL1*	**8.934**	**Medium**	V1228M	19.54	**0.999**	**D**
				V648I	12.91	0.133	B
11	*SLC22A24*	**11.004**	**Medium**	R203H	19.33	**0.945**	**D**
				V121L	13.84	**0.941**	**D**
12	*GCN1*	-	-	A2324S	18.04	0.028	B
				S333L	14.25	0.001	B
				A2324V	14.1	0.004	B
6	*TNXB*	28.289	High	V1272M	14.43	0.102	B
				G691D	12.98	**0.959**	**D**
				V2433I	7.086	0.03	B
4	*SORBS2*	**3.175**	**Medium**	R36H	13.71	**0.934**	**D**
				L509I	11.06	**0.985**	**D**
6	*PKHD1*	**11.948**	**Medium**	T2641A	13.69	**0.994**	**D**
				R1909W	9.428	0.006	B
				P676R	5.715	0.002	B
				E1448G	3.189	0.002	B
4	*INTU*	**2.917**	**Medium**	D460N	13.6	0.217	B
				C692G	1.797	0.005	B
16	*ZNF469*	**12.25**	**Medium**	G3416R	13.36	**0.628**	**P**
				A3571S	8.346	0.036	B
				A889V	7.102	0.004	B
				A803S	4.071	0.056	B
1	*C1orf167*	-	-	A115S	13.46	-	-

				C1226R	10.2	0.004	B
				R410W	9.07	-	-
				G170R	0.066	-	-
				H193R	0.013	-	-
				N340H	-	-	-
				D424fs	-	-	-
17	*GPR179*	**9.509**	**Medium**	A798T	11.74	0.021	B
				R1190Q	0.409	0	B
1	*IQGAP3*	18.986	High	Q513L	11.41	0.036	B
				G649S	0.034	0.003	B
9	*DDX31*	**4.977**	**Medium**	E153K	11.06	0.064	B
				R843Q	6.413	0.001	B
6	*EYS*	**20.954**	**High**	G663E	10.96	0.188	B
				D1745N	10.12	0.006	B
14	*AHNAK2*	32.499	High	G746A	10.38	**0.712**	**P**
				L2586I	0.008	0.002	B
				M2579V	0.001	0	B
13	*KL*	**9.724**	**Medium**	N131K	9.19	**0.657**	**P**
				A107V	1.524	0.007	B
8	*SGK223*	25.838	High	L122I	8.771	0.07	B
				H1077Q	0.003	0.001	B
16	*ABCC6*	**9.731**	**Medium**	L1097I	8.586	**0.566**	**P**
				Q1022E	3.993	0.124	B
17	*ATAD5*	19.691	High	A625P	4.085	**0.571**	**P**
				L1030F	3.356	**0.813**	**P**
1	*NPHP4*	**6.409**	**Medium**	H894R	0.049	0.001	B
				T623A	0.009	0.001	B
5	*CDC20B*	**6.957**	**Medium**	D31G	5.451	0.009	B
				R36fs		-	-

* First, variants at each gene were sorted according to the CADD scores (from highest to lowest). Then, the genes were sorted according to the highest CADD score at each gene. GDI < 13.84 (genes with low and medium damage in general populations), CADD > 20 and PolyPhen-2 > 0.446 (possibly and probably damaging variants) are shown in bold.

¶ B: Benign, P: Possibly damaging, D: Probably damaging.

AA: amino acid; Alt: alternative allele; CADD Combined annotation dependent depletion; Chr: chromosome, GDI: Gene damage index; Ref: Reference allele.

6.3.2 Dominant model

Functional prediction scores of all candidate variants identified in the dominant model (Models #5, #6 and #7) are presented in **Table 17**. From the 37 novel variants identified in this model, seven candidate variants were prioritized, including a stop-

gain mutation in *PRKACB* (E122X), a frameshift deletion in *ISOC2* (P119fs) and five predicted-damaging missense mutations in *C1R* (E1674G), *CP* (K576E), *MCM9* (T618I), *NSFL1C* (Y155F) and *SOWAHB* (L676V) (**Tables 14** and **17**).

Table 17. Gene and variant-level metrics of functional impact prediction of candidate variants from dominant models #5, #6 and #7.

Model[¶]	Chr	Gene	GDI**		AA Change	Scaled CADD**	PolyPhen-2**	
			Score	Damage			Score	Prediction[#]
#6	1	*PRKACB*	**0.721**	**Medium**	E122X	**42**	-	-
#6	19	*ISOC2*	**1.872**	**Medium**	P119fs	**35**	-	-
#6	20	*NSFL1C*	**12.398**	**Medium**	Y155F	**30**	**0.956**	**D**
#5	6	*MCM9**	**5.14**	**Medium**	T618I	**26.7**	**0.735**	**P**
#5	3	*CP**	**4.69**	**Medium**	K576E	**26.1**	**0.731**	**P**
#7	6	*NUP153*	17.573	High	A1315S	**25.8**	**0.542**	**P**
#6	10	*PDCD11*	18.504	High	E1563K	**25.5**	0.075	B
#6	15	*TRIP4*	**1.808**	**Medium**	V538I	**24.4**	0.378	B
#6	1	*CR1*	**11.671**	**Medium**	E1674G	**24**	**0.463**	**P**
#5	4	*SOWAHB*	**10.04**	**Medium**	L676V	**24**	**0.806**	**P**
#6	16	*RBBP6*	**4.149**	**Medium**	P1614L	**23.4**	0.036	B
#5	11	*CCDC34*	**8.872**	**Medium**	R23I	**23.2**	0.146	B
#5	X	*SUV39H1*	**0.243**	**Medium**	R356W	**23.2**	0.013	B
#6	19	*BCKDHA*	-	-	P52S	**23**	0.044	B
#5	4	*SYNPO2**	15.146	High	E114K	**22.5**	0.081	B
#6	19	*ZIM2*	**3.89**	**Medium**	S158F	**22.2**	0.005	B
#6	16	*RBL2*	**6.901**	**Medium**	A1031V	**21.5**	0.001	B
#5	12	*LEMD3*	**1.854**	**Medium**	G124D	**20.7**	0.266	B
#6	19	*SCAF1*	**5.028**	**Medium**	G893S	**20.3**	0.091	B
#5	18	*CABLES1**	**8.032**	**Medium**	M1delinsMA	18.8	-	-
#6	1	*IQGAP3*	18.986	High	Q513L	18.3	0.036	B
#6	1	*ASTN1*	**4.844**	**Medium**	D1264A	14.7	0.001	B
#6	9	*SLC25A25*	**1.516**	**Medium**	V59I	13.4	0.023	B
#5	X	*ZNF185*	**6.97**	**Medium**	A459G	13.4	0.167	B
#6	3	*NBEAL2*	15.494	High	V685M	13.2	**0.482**	**P**
#6	19	*LRRC25*	**2.446**	**Medium**	P234L	12.3	0.003	B
#6	19	*SCAF1*	**5.028**	**Medium**	T895N	11.4	0.001	B
#6	15	*CCPG1*	**4.191**	**Medium**	G64V	10.4	0.01	B
#5	5	*UGT3A1*	**8.867**	**Medium**	A288T	9.8	0.055	B
#5	3	*PPP4R2*	**3.505**	**Medium**	P227S	9.6	0.019	B

#6	15	*ADAL*	**2.53**	**Medium**	T12A	8.9	0.004	B
#6	2	*FSIP2*	15.565	High	H4382D	8	-	-
#6	18	*CDH20**	**5.231**	**Medium**	D600N	7.1	0.003	B
#5	3	*RYK**	**8.995**	**Medium**	L26insPP	1.3	-	-
#5	17	*ERN1*	**4.038**	**Medium**	K704R	0.048	0.001	B
#6	22	*CRELD2*	**7.447**	**Medium**	P175S	0.006	0.008	B
#6	4	*OTOP1*	**6.377**	**Medium**	M304L	0.001	0.003	B

¶ Prioritized variants are highlight in blue: these variants present CADD > 20, PolyPhen-2 > 0.446 (if missense SNV) and are located in genes with GDI < 13.8.
* The twin's aunt (II-4) also carries the same mutation in this gene as the affected members of the family.
** The variants are sorted according to their CADD score (from highest to lowest). GDI < 13.84 (genes with low and medium damage in general populations), CADD > 20 and PolyPhen-2 > 0.446 (possibly and probably damaging variants) are shown in bold.
# B: Benign, P: Possibly damaging, D: Probably damaging.
AA: amino acid; CADD Combined annotation dependent depletion; Chr: Chromosome; GDI: Gene damage index.

6.4 VARIANTS IN LEPROSY-ASSOCIATED GENES

In addition to the filtering procedure performed in the WES/WGS data, we conducted a more detailed analysis of the coding regions from genes previously associated with leprosy. For that, we selected all exonic and splicing variants in leprosy genes identified in the studied family. Then, we applied the same filtering steps for the dominant and recessive models, except for MAF filtering for which a less stringent threshold was applied (Section 5.5). As result, seven variants passed this filtering procedure and are shown in **Table 18**. For the recessive model, there are three SNPs in *HLA-A* and two missense variants in *LRRK2* gene (**Table 18**). These five variants already passed the filtering procedure used in WES and/or WGS data since their MAF are lower than 30% in the population samples from searched databases (**Tables 7**, **13** and **S1**). Therefore, even after searching with a less stringent criteria for MAF, there were no additional candidate variants in leprosy genes following a recessive model. Both variants in *LRRK2* are predicted to be damaging by Polyphen-2 and have high CADD score (**Tables 15** and **18**). On the other hand, all missense variants in HLA-A – a highly damaged gene (high GDI) – are predicted to be benign and present low CADD scores (**Tables 15** and **18**).

For the dominant model, two variants passed the filtering steps followed for leprosy-associated genes: rs2066844 (R702W) in *NOD2* and rs569286159 (F55L) in *HLA-DRB1* (**Table 18**). The father (II-5) and twins (III-10 and III-11) carry 702W in *NOD2* (heterozygous) while the grandmother (I-2) and the unaffected mother (II-6) are homozygous for the common allele (R702). All affected members of the family carry 55L allele in *HLA-DRB1* (heterozygous). Both variants were previously present in public databases and registered under a rsID (**Tables 18** and **S1**). Here, both variants were detected because we used MAF ≤ 10% in filter 3 instead of novel variants. Consequently, since they are not novel (**Supplementary Table S1** in **Appendix 2**), none of them passed the variant filtering used in the dominant model in WES/WGS analysis. In the dominant model, only the missense variant in *NOD2* is predicted to be possibly damaging for the protein structure and function by PolyPhen-2 and has a high CADD score, while the variant in *HLA-DRB1* – a gene with high GDI score – is predicted to be benign and presents a low CADD score (**Table 18**).

es previously associated to leprosy (Recessive and dominant models).

ef	Alt	Gene	AA Change	rsID	GDI**		Scaled CADD**	PolyPhen-2**	
					Phred	Damage		Score	Prediction¶
C	G	*LRRK2*	N551K	rs7308720	**11.25**	**Medium**	**27.3**	**0.972**	**D**
G	A	*LRRK2*	R1398H	rs7133914	**11.25**	**Medium**	**23.1**	**0.566**	**P**
G	C	*HLA-A*	E87D	rs199474424	34.46	High	13.6	0.123	B
G	A	*HLA-A*	E87K	rs2230991	34.46	High	10.8	0.032	B
G	A	*HLA-A*	V358M	rs1137631	34.46	High	7.0	0.114	B
C	T	*NOD2*	R702W	rs2066844	**4.63**	**Medium**	**24.1**	**0.72**	**P**
G	C	*HLA-DRB1*	F55L	rs569286159	17.41	High	0.002	0.002	B

F's filter for leprosy associated genes than the whole exome analysis.
e: these variants present CADD > 20, PolyPhen-2 > 0.446 (if missense SNV) and are located in genes with GDI < 13.8.
their CADD score (from highest to lowest) for each trait.). GDI < 13.84 (genes with low and medium damage in general populations),
sibly and probably damaging variants) are shown in bold.
obably damaging.
ADD Combined annotation dependent depletion; Chr: Chromosome, GDI: Gene damage index; MAF: Minor allele frequency; Ref:

7 DISCUSSION

Over the past decades, intense efforts have been applied to the description of the exact nature of the genetic effect controlling leprosy susceptibility. Studies using different strategies of analysis, including GWAS, have resulted in the description of several common genetic variants associated with leprosy. However, complete understanding of the genetic mechanisms controlling host susceptibility to leprosy, despite of the large body of accumulated evidence and high quality of research already produced about the disease, is yet elusive. Aditional contributions to understand the missing heritability in leprosy will depend on alternative approaches and analysis strategies. Here, we applied advanced technology of massive, next generation sequencing to produce the complete genomic sequence of selected individuals from a Brazilian family with three generations of leprosy cases, which includes a pair of monozygotic twins who developed clinically concordant leprosy at 22 months of age. This very early clinical outcome alongside the presence of the disease in three generations of the family strongly suggest that a shared genetic component may underlie the observed enrichment of leprosy in this family. To identify this genetic component, the first aim of this study was to detect SNVs and short indels located in protein-coding regions by DNA sequencing with NGS technology. Once these variants were detected, the goal was to develop and apply custom, stepwise filtering procedures in order to identify variants that are most likely to be causal. A number of different approaches were applied to determine co-segregation of coding variants with leprosy susceptibility in the family, taking into account age-at-diagnosis and possible model of inheritance (see section 5.5). As result, several candidate variants were identified and their respective *in silico* functional prediction were used to prioritize deleterious variants.

7.1 WES VERSUS WGS

Key factors in evaluation of NGS data, regardless of technology/platform, include variant calling accuracy and depth of coverage (reviewed in (104)). In addition to those parameters, DNA sequencing on Ion platforms require special handling due to limitations inherent to the technology that include a lower base calling quality compared to Illumina platforms, as well as high indels error rates, specifically on homopolymeric sequences (more than 3 nucleotide repetition) (154). Thus, to reduce false positives in Ion data, we applied Thermo Fisher Scientific's proprietary analysis software that takes into account Ion Torrent's particular data acquisition methods (127). Specifically, using Torrent Suit's TVC we leverage the use of flow order and flow signal registry to produce more accurate variant call data set (155). Applying the mentioned tools on WES data led to the identification of 35,784 exonic variants in total (**Table 6**). Indeed, comparison of this pipeline with GATK best practices (129,106) – a widely used NGS data analysis workflow mainly used for Illumina platforms – on our WES data, the former showed a lower number of indels in variant calling results, indicating TS' suitability for Ion data (data not shown).

Regarding depth of coverage, coverage flaws among the five samples were mainly due to experimental and technical reasons. However, we observed that coverage depth was not uniform throughout all exome regions for each sample, regardless of the mean coverage (**Table 5** and **Figure 11-A**). Hence, we were not able to analyze a considerable amount of target regions, with an estimated data loss of 24% on candidate SNVs and 43% on candidate indels detection in targeted sequences (**Figure 11-B**). Uneven depth of coverage is a common characteristic of target-enrichment methods; WES requires deeper sequencing than WGS in order to overcome this limitation (reviewed in (101)). Moreover, a recent study from Belkadi *et al.* has shown that WGS is more powerful to detect exonic variants than targeted WES due to greater coverage and quality uniformity as observed in six samples (156). Based on the mentioned limitations, we decided to perform WGS using the same samples used for WES plus an additional sample from the twin's aunt, conducted on Illumina HiSeq® 2500 platform. As result, more than eight million variants were

detected, 44,388 of them in exonic/splice-site regions (**Table 12**). As for depth of coverage, our WGS results were consistent to those found by Belkadi *et al.* since it was more uniformed than WES (**Table 11** and **Figure 13**). Indeed, comparing WES and WGS coverage throughout regions targeted by Ion TargetSeq™ Exome kit showed that almost 99% of those were covered at least 10X in all samples in WGS (data not shown), whilst only 78% to 84% of the same regions were covered at least 10X in WES (**Figure 11-A**). WGS data yielded more candidate variants after variant calling and filtering (**Tables 13** and **14**). An intersection analysis of the target regions from TargetSeq Exome Kit showed that all candidate variants – from dominant and recessive models – detected only in WGS experiments were located in regions targeted by the exome capture probes used in WES (data not shown). This indicates that these candidate variants are not located in off-target regions from WES but were not identified in our WES analysis due to insuficiente sequencing data or technically biased capture.

Besides the detection of additional candidate variants, WGS results also served as validation for the findings from WES analysis. Since most NGS artifacts are specific for the nucleotide detection method, variants detected by more than one sequencing platform are likely to be real. Interestingly, all candidate variants identified in WES were also detected in WGS. Hence, our WES data allowed us to properly identified most of the candidate variants. In addition to these validations, Sanger sequencing – which is the gold-standard sequencing method (reviewed in (97)) – also validated ten candidate variants. In agreement to our WGS and WES data, the Sanger sequencing results were concordant with NGS data for these candidate variants (**Table 9**).

7.2 CANDIDATE VARIANTS IN THE PIAUÍ FAMILY

We hypothesized that the genetic factor that predispose the studied family to leprosy is segregating as a monogenic recessive trait, given that this is the case for most Mendelian predisposition to infectious diseases (reviewed in (35,36)). Assuming the recessive model, a total of 95 variants in 51 genes were detected as candidates

(**Table 13**). From these, six SNPs have been predicted to be protein-damaging variants with high CADD score, located in five medium-GDI genes: *SLC17A9, LRRK2, ACP5, ZNF678* and *HRH4* (**Table 15**). Among the candidate variants with predicted impact identified in the recessive model, the two missense SNPs in *LRRK2* were of immediate interest as this gene was associated with leprosy in previous studies (71,75,157,158). *LRRK2 – Leucine Rich Repeat Kinase 2 –* is a large gene comprising 51 exons that is located at human chromosome 12q12. It encodes a multifunctional, multi-domain protein with 2,527 amino acids hosting kinase and GTPase functions surrounded by protein–protein interactions domains (**Figure 14-A**). The catalytic core is contained within a Ras of Complex proteins (Roc) domain, an adjacent C-terminal of Roc (COR) domain and a kinase domain. Protein–protein interaction domains include the N-terminal armadillo (ARM), Ankyrin (ANK), and leucine-rich repeat (LRR) domains, along with a C-terminal WD40 domain (**Figure 14-A**) (reviewed in (159)). LRRK2 normally exists as a dimer and its activity switches between an inactive GDP-bound state and an active GTP-bound state. GTP binding to the Roc domain – as well as dimerization – is required for LRRK2 kinase activity (160). *LRRK2* is a well-known Parkinson disease (PD)-related gene, since mutations in this gene are the most common causes of late-onset autosomal dominant PD and are also found in sporadic forms of PD. Most of PD risk variants are missense SNVs located especially in the GTPase and kinase domains (reviewed in (159)). Functional studies demonstrated that these variants are related to a gain-of-function in LRRK2 protein by an increase in its kinase activity or a decrease in its GTPase activity (reviewed in (161)). In PD, LRRK2 has been shown to regulate a diverse set of cellular function including vesicle transport, autophagy, cytoskeletal organization and mitochondrial effects (reviewed in (159)). Interestingly, a functional study performed by Smith *et al.* showed that LRRK2 interacts with PARK2 protein (Parkin) (162). *PARK2* gene is associated with early-onset autosomal recessive PD (reviewed in (163)), as well as leprosy susceptibility (reviewed in section 1.3 of this thesis).

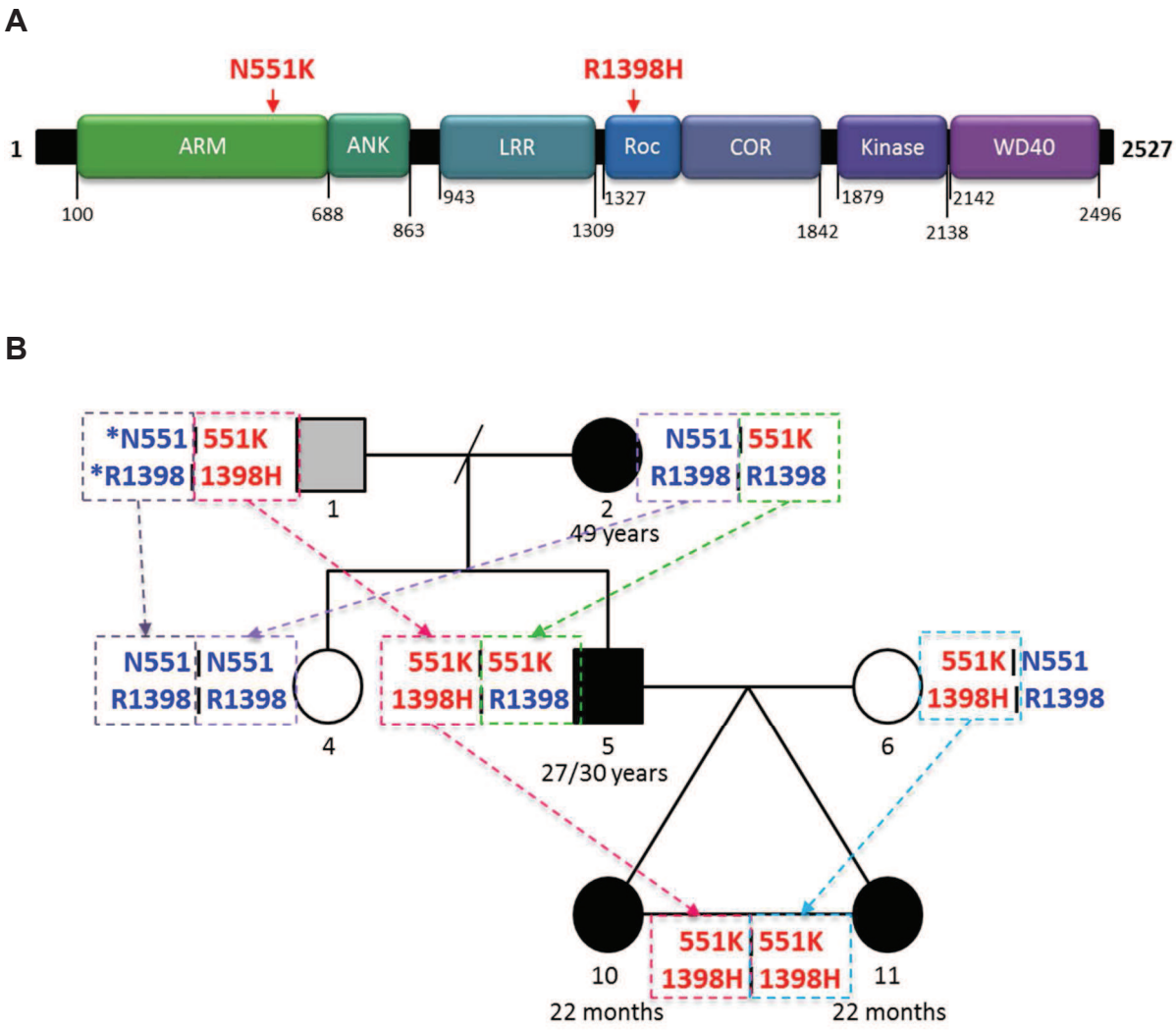


Figure 14. *LRRK2* missense variants identified in the Piauí family under the recessive model. **A)** Schematic representation of human LRRK2 protein (primary structure) and location of N551K and R1398H variants. Each colored box represents a domain, while non-domain regions are shown in black. Amino acid positions at the beginning and ending of each domain are shown below. Locations of the two missense variants identified in the studied family in the recessive model are indicated in red (above). ANK: Ankyrin repeat; ARM: Armadillo; COR: C-terminal of Roc; H: Histidine; K: Lysine; LRR: Leucine-rich repeat; N: Asparagine; R: Arginine; Roc: Ras of complex proteins - GTPase. **B)** Segregation of *LRRK2* N551K-R1398H haplotype in the Piauí family.Haplotype segregation is indicated with dashed arrows. Common alleles are shown in blue and minor alleles are shown in red. N551K and R1398H genotypes in the twin girls (III-10 and III-11), their parents (II-5 and II-6) and grandmother (I-2) were obtained from WES, WGS and Sanger sequencing results; in the aunt (II-4) they were obtained from WGS and in the twin's grandfather (I-1), the genotypes of both variants were inferred from the offspring data (they are indicated with asterisks). Each individual of the family is numerated in concordances with **Figure 4**. Age-at-diagnosis of leprosy cases are shown in black.
Source: A) Adapted from Guaitoli*, et al.* (164).

The first connection between *LRRK2* and leprosy came from the first leprosy GWAS conducted in Chinese population samples, where a trend towards association between the disease and SNP rs1873613 – located upstream from *LRRK2* gene – was found (71). Also, subgroup analysis of MB and PB leprosy subtypes found association of a *LRRK2* intronic SNP (rs1491938) and MB leprosy (71). Later, SNP rs1873613 was tested for leprosy association in two independent studies in Vietnamese and Indian population samples (73,158). No association of this SNP and leprosy was found in the Vietnamese families (73), whereas it was associated with susceptibility to leprosy *per se* and PB subtype in the Indian population (158). In addition, Wang *et al.* performed an association study of *LRRK2* and leprosy in a Chinese population sample, where five *LRRK2* SNPs (including rs1873613) were associated to susceptibility to leprosy or PB subtype (157). Interestingly, a recent study from Fava *et al.* evaluated a possible role for *LRRK2* in T1R-affected leprosy families and contrasted the results with T1R-free leprosy families from Vietnam (75). The authors identified a total of 18 T1R-specific associated variants – including rs1491938 – organized in four bins. The main SNP capturing the T1R association was the SNP rs3761863 (M2397T), a missense variant in the WD40 domain of LRRK2. Interestingly, this SNP is known to impact LRRK2 turnover – in which the protein with M2397 allele presents shorter half-life than 2397T – and was previously reported in association with Crohn's disease (CD) with the same risk allele as in T1R (M2397) (75,165). The authors went one step further and performed an eQTL analysis to investigate if SNP alleles associated with T1R were correlated with LRRK2 transcriptional levels. As result, they found nine variants belonging to the same SNP bin as M2397 that promote an increase in LRRK2 expresion in non-stimulated cells. This indicates that these eQTL SNPs counterbalanced LRRK2 shorter half-life due to the M2397 variant. However, this compensatory mechanism was abrogated following stimulation with *M. leprae* (75).

In addition to PD and leprosy, variants in *LRRK2* – including the mentioned missense variant M2397T – are also associated with CD (reviewed in (166)), an Inflammatory Bowel disease (IBD) characterized by a chronic relapsing intestinal inflammation, which is believed to be trigged – at least in some cases – by intestinal infection with mycobacterial species (reviewed in (167)). Indeed, this hypothesis was

reinforced by association studies that identified overlap of susceptibility genes for IBD (mainly CD) and leprosy, including not only *LRRK2* but also *HLA-DRB1, IL12B, IL18RAP-IL1RL1, IL23R, CCDC122-LACC1, NOD2, RIPK2* and *TNFSF15* (71,73,79,88). Functional studies have been conducted aiming to better understand *LRRK2* involvement in CD (reviewed in (168)). LRRK2 mRNA expression analysis has indicated that LRRK2 expression is enriched in human immune cells, especially in B lymphocytes, monocytes and dendritic cells (reviewed in (168)). Interestingly, Gardet *et al.* showed that when murine macrophages (RAW 264.7 cell line) were stimulated with *S. typhimurium*, a fraction of Lrrk2 protein translocate to near the pathogens (169). Moreover, the authors also demonstrated that knockdown of *Lrrk2* lead to a reduction in reactive oxygen species (ROS) production as well as an increase in bacterial survival (169). Later, Liu *et al.* have shown that mice knockout for *Lrrk2* are more susceptible to dextran sulphate sodium (DSS)-induced CD (an animal model for colonic inflammation induced by a chemical) (165). In the same study, the authors have shown that *Lrrk2* functions as negative regulator of an immune response transcription factor NFAT (nuclear factor of activated T-cells), which plays a role in immune cell maturation as well as cytokine production in macrophages and dendritic cells (165). Lack of *Lrrk2* lead to an increased production of pro-inflammatory cytokines by NFAT. Moreover, it has been recently reported that – in mouse model – *Lrrk2*, *Nod2* and *Rab2a* participate in a common axis important for the proper secretion of lysozyme by secretory epithelial cells of the small intestine, called Paneth cells, to promote barrier protection for intestinal homeostasis, which is compromised in CD (170,171).

In the Piauí family, two missense variants in *LRRK2* – rs7308720 (N551K) and rs7133914 (R1389H) – were detected in the recessive model, but following different filtering approaches. The twins (III-10 and III-11) and father (II-5) are homozygous for 551K (Model #2) and only the twins (III-10 and III-11) are homozygous for 1398H (Model #3) (**Figure 14-B** and **tables 7** and **13**). Haplotype segregation of the two variants in the family is shown in **Figure 14-B**. Interestingly, the pattern raises the question whether there is a gene-dosage effect for the number of variants on LRRK2 considering a recessive model and the age of onset of the patients. Given that the twin girls (III-10 and III-11) – who developed leprosy when they were 22 months old – are

homozygous for both variants; the father (II-5) – whose age-at-diagnosis was 27 years old – is homozygous only for one of the variants, while the grandmother (I-2) which had only one heterozygous variant, developed leprosy at the age of 49 years.

According to the 1000G database, these SNPs are common, presenting mean MAF nearly 10% in the combined populations (**Supplementary Table S1** in **Appendix 2**). However, their frequencies and LD patterns varies among population samples. In South Asian, both variants are of low frequency (MAF < 5%, expected risk-genotype frequency < 0.25%) and present strong LD (r^2 = 1), whilst in Admix American and African population samples they are more common – MAF between 10% and 15% (expected risk-genotype frequency between 1% and 2.25%) – but with lower LD (r^2 = 0.66 and 0.18, respectively) (**Supplementary Table S1** in **Appendix 2**). Both variants are predicted to be damaging according to PolyPhen-2 and present high CADD scores (**Table 15**). Interestingly, *LRRK2* variants N551K and R1398H were previously associated to PD in Asian and white population samples, where the minor alleles (551K and 1398H) confer protection against PD (172,173). Haplotype analysis showed association of PD and a three-variant protective haplotype (N551K-R1398H-K1423K) in both populations. In the Piauí family, only the twins (III-10 and III-11) are homozygous for K1423K minor allele, while both parents (II-5 and II-6) are heterozygous and the grandmother (I-2) does not present this variant (data not shown). In LRRK2, R1398H falls within the ROC domain (**Figure 14-A**) and it has been functionally demonstrated that the 1398H allele has increased GTPase activity, as well as reduced kinase activity in comparison to the reference allele R1398 (173,174). Regarding N551K, it is located in the ARM domain (**Figure 14-A**), but little is known about the impact of variants in the N-terminal of LRRK2. An interesting study performed by Waschbusch *et al.* showed that LRRK2 interacts with RAB32 via ARM domain (175). *RAB32* is a small GTPase that has been previously associated to leprosy (79). According to Waschbusch *et al.* results, overexpression of constitutively active RAB32 decreases the amount of LRRK2 in mitochondria and lysosomes containing fractions. Therefore, RAB32-dependent LRRK2 sub-cellular localization demonstrates a role for RAB32 in LRRK2 sorting and transport (175). If hypothesis that N551K variant has an impact in LRRK2-RAB32 interaction, this could lead to an alteration in LRRK2 sorting and transport. However, no functional data for N551K

impact on LRRK2 activity or on its interaction with RAB32 – or with other proteins – has been reported to date.

A second interesting candidate gene detected in the Piauí family is *HLA-A,* under the recessive model. This gene belongs to HLA class I, which molecules play a central role in the immune system by presenting peptides derived from cytosolic proteins (reviewed in (176)). Previous studies have shown that HLA-A*2 and A*11 alleles are more frequent in leprosy patients than in unaffected controls (reviewed in (177)). However, a high-density association scan of HLA class I region and leprosy susceptibility – performed in two independent Vietnamese family-based population samples and an Indian case-control population sample – have shown that HLA-C is associated with leprosy whilst there was no association between *HLA-A* SNPs and the disease (69). In our study, three missense SNPs in *HLA-A* gene passed variant filtering in the recessive model. Two of these variants, rs2230991 and rs199474424, are located at codon 87 in *HLA-A* exon 2. This exon is a polymorphic site that encodes HLA-A protein alpha 1 domain which, together with alpha 2 domain, creates the protein-binding cleft (178). Only the twin girls (III-10 and III-11) are homozygous for the minor alleles of both variants (Model #3, **Table 13**). Presence of both minor alleles in the same molecule confers a change of a glutamic acid (E) to an asparagine (N) at amino acid 87. However, E87N amino acid change due to these variants is predicted to be benign for protein function according to PolyPhen-2 and CADD scores (data not shown). The third *HLA-A* candidate SNP found in this study – rs1137631 (V358M) – is located at exon 7, which encodes the HLA-A cytoplasmatic tail (178). All affected individuals in the Piauí family were homozygous for the minor allele while unaffected mother was heterozygous (Model #1, **Tables 7** and **13**). Moreover, V358M is not predicted to be deleterious (**Table 15**). The three variants are common in all populations from 1000G and ExAC – where MAFs range from 10% to 28% (expected risk-genotype frequency ranging from 1% to 7.8%) (**Supplementary Table S1** in **Appendix 2**). Nevertheless, it was not possible to estimate LD among the three *HLA-A* SNPs because V358M was not genotyped in 1000G and rs2230991 is triallelic in this database. However, we searched for LD pattern between *HLA-A* rs199474424 and 12 SNPs located in and around *HLA-C* gene that were previously associated to leprosy in the fine mapping study mentioned before (69). As result, none of the leprosy

associated SNPs in *HLA-C* are in LD with *HLA-A* rs199474424 candidate variant (data not shown). HLA-A has high GDI score, which indicates that rare nonsynonymous variants are not unusual in this gene in general populations (**Table 15**). Taken together, computational prediction tools suggest that it is not likely that these three SNPs in *HLA-A* are causative of a Mendelian predisposition to leprosy. However, the possibility that they confer risk to the disease as a polygenic effect with other variants cannot be excluded.

As mentioned before, our central hypothesis is that the causative variant is segregating in the family as a recessive trait. Nevertheless, considering that the disease is present in three generations of the family with no skipping, it is possible that a dominant trait is involved. Based on that, the sequencing data was also analyzed for the dominant model. However, identifying candidate variants under a dominant model from WES/WGS data is more challenging than under the recessive model, since the number of FP that pass filtering approaches are usually higher (100). Therefore, a limitation of our filtering approach for the dominant model is the need to focus only on novel functional variants. In total, 37 novel mutations were identified as candidates to leprosy susceptibility under a dominant model (**Table 14**). Among the variants with high CADD score (scaled CADD >20) in medium-GDI genes, we identified one stop-gain, one frameshift deletion and five predicted damaging missense mutations. These variants were predicted to have impact on the function of the proteins encoded by seven genes: *PRKACB, ISOC2, NSFL1C, MCM9, CP, CR1* and *SOWAHB* (**Table 17**). Variants in *PRKACB, ISOC2, NSFL1C* and *CR1* were found only in the father and twins (Model #6), whilst the mutations in *MCM9, CP* and *SOWAHB* were identified in all affected members of the family (Model #5, **Table 14**). Of note, the unaffected aunt also carries the missense variants in *MCM9* and *CP*. None of the 37 candidate variants from the dominant model are located in genes previously related to leprosy susceptibility by association studies. However, interesting findings for *CR1* have been reported in early functional studies from the 90s. These studies indicated that phagocytosis of *M. leprae* by human monocytes/macrophages can be mediated by complement receptor CR1 and CR3 (179,180).

Moreover, when leprosy genes were searched in more detail for the dominant model, we identified two additional candidate missense variants in *HLA-DRB1* and

NOD2 genes. In *HLA-DRB1*, rs569286159 (F55L) was present in all affected members of the family (Model #5 with MAF < 10% as threshold; **Table 18**). In *NOD2,* missense variant rs2066844 (R702W) was found only in the twins (III-10 and III-11) and father (II-5) (Model #6 with MAF < 10% as threshold; **Table 18**). Computational prediction of functional impact of these variants indicates that only R702W in *NOD2* might be deleterious to protein function (**Table 18**). NOD2 is an intracellular microbial sensor of the innate immune system that can sense and recognize intracellular pathogens (reviewed in (181)). As mentioned before, this gene is associated to leprosy as well as CD susceptibility. In leprosy, *NOD2* gene was associated to the disease in the first Chinese leprosy GWAS and validated in several independent samples including Nepalese (72), Vietnamese (73) and Brazilian (78) populations (reviewed in section 1.3 of this thesis). The missense variant in *NOD2* which codes for R702W has been previously reported in public databases and its MAF ranges from 0% in East Asian to 5.1% in European populations (expected risk-genotypes frequency ranging from 0% to 9.9%) (**Table 18**). Interestingly, this variant is a risk factor to CD, where 702W confers risk to the disease (reviewed in (182)). Indeed, R702W – together with G908R and a frameshift deletion L1007fs – is one of *NOD2* variants most commonly found in CD patients (reviewed in (182)).

Taking all filtering approaches together, analysis of exonic regions in the Piauí leprosy family identified several candidate variants that may contribute to explain the extreme early-onset leprosy in the monozygotic twins, under different models. Literature search has helped to prioritize candidate variants in genes related to leprosy, as discussed above. The remaining variants are located in genes which function is unknown and/or with no clear involvement in leprosy or mycobacterial infections. Nevertheless, these could be causal variants in new leprosy susceptibility genes. The candidate variants in the Piauí family may have an impact in disease susceptibility itself and/or in the incubation period of leprosy in the twin girls. The role of genetic factos controlling leprosy incubation period is an exciting hypothesis, but very challenging to be clinically characterized and functionally demonstrated. Short incubation period in adults may be a hidden phenotype and, consequently, difficult to study. Moreover, we cannot exclude the possibility that the variants are involved in the disease susceptibility or incubation period via an oligo/polygenic effect together with

other(s) variant(s). Functional studies are now necessary in order to confirm and to better understand the involvement of the candidate variants in this case of leprosy in the Piauí family.

7.3 FUTURE PERSPECTIVES

Several follow-up studies can be performed based on the results and data obtained in this study. An exciting approach will be to perform a linkage analysis of the Piauí family coupled with a filtering procedure from the WGS data inside linked regions. This will allow a more detailed analysis not only of coding variants, but also of non-coding variance detected in the WGS analysis. Others interesting follow-up studies include an association study of the candidate genes identified in this work, as well as functional studies from the candidate variants. In fact, our research group is already initiatiating a new project to functionally validate the present results. The new study proposes to obtain induced pluripotent stem cells (iPSC) from biopsies of the Piauí family to be used to generate immortalized cell lines ideal for functional studies (183). Coupled with the use of iPSC, DNA edition by CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats/ CRISPR associated protein 9) will be used for the generation of cell lines with the desired genetic modifications (reviewed in (184)). This will allow us to compare cells in presence versus absence of the candidate variant(s) for a set of experiments followed *M. leprae* stimulation. Both *LRRK2* variants will be prioritized for functional validation.

8 CONCLUSION

1. Identification of exonic/splice-site variants in the Piauí family:
    a. In total, 35,784 exonic and splice-site variants were identified in WES data analysis.
    b. WGS in the Piauí family led to the identification of more than eight million variants including 44,388 exonic/splice-site variants.
2. Filtering approaches for candidate variant identification:
    a. Assuming a recessive trait, 95 exonic variants with MAF < 30% – located in 51 genes – were identified as candidate.
    b. Assuming a dominant model, a total of 37 novel exonic variants were detected as candidates to leprosy susceptibility in the Piauí family.
    c. Among all candidate genes from recessive and dominant models, *LRRK2* and *HLA-A* have been associated with leprosy in previous studies. Further analysis of known leprosy-associated genes led to the identification of two additional variants – located in *NOD2* and *HLA-DRB1* genes – as candidates in the dominant model.
3. *In silico* prediction of functional impact of candidate variants:
    a. Six candidate variants following a recessive trait are predicted to be protein-damaging: rs7271712 in *SLC17A9,* rs7308720 and rs7133914 in *LRRK2,* rs2229531 in *ACP5,* rs61740826 in *ZNF678* and rs58154316 in *HRH4.*
    b. From the total candidate variants in the dominant model, mutations in *PRKACB*, *ISOC2, NSFL1C, MCM9, CP, CR1* and *SOWAHB* genes were predicted to be damaging for the protein function. Moreover, rs2066844 in *NOD2* gene is predicted to be damaging for the protein function.

REFERENCES

1. Robbins G, Mushrif Tripathy V, Misra VN, Mohanty RK, Shinde VS, Gray KM, et al. Ancient skeletal evidence for leprosy in India (2000 B.C.). PLoS One. 2009;4(5):1–8.

2. Lastória JC, de Abreu MAMM. Leprosy: Review of the epidemiological, clinical, and etiopathogenic aspects - Part 1. An Bras Dermatol. 2014;89(2):205–18.

3. Lewis G. A Lesson from Leviticus: Leprosy. Man. 1987;22(4):593–612.

4. Petrusevski AB. History of infectious diseases development in the old and the middle ages with the emphasis on the plague and leprosy. Vojnosanit Pregl. 2013;70(7):704–8.

5. Hansen G. On the etiology of Leprosy. Br Foreign Med-Chir Rev. 1875;55:459–89.

6. WHO. Leprosy: Fact sheets [Internet]. Available from: http://www.who.int/mediacentre/factsheets/fs101/en

7. WHO. World Health Assembly (WHA) resolution to eliminate leprosy: The World Health Assembly Resolution 1991 [Internet]. Available from: http://www.who.int/lep/strategy/wha/en/

8. WHO. Global leprosy update, 2015: time for action, accountability and inclusion. Wkly Epidemiol Rec. 2016;35(91):405–20.

9. WHO. Accelerating work to overcome the global impact of neglected tropical diseases - a roadmap for implementation. Geneva: World Health Organization. 2012.

10. Blok DJ, De Vlas SJ, Richardus JH. Global elimination of leprosy by 2020: are we on track? Parasit Vectors. Parasites & Vectors; 2015;8(1):548.

11. WHO. Global leprosy strategy 2016-2020: Accelerating towards a leprosy-free world. Geneva World Heal Organ. 2016;

12. BRASIL. Ministério da Saúde. Indicadores epidemiológicos e operacionais de hanseníase, Brasil 2000-2015 [Internet]. 2016 [cited 2016 Aug 1]. Available from: http://portalsaude.saude.gov.br/images/pdf/2016/julho/07/Indicadores-epidemiol--gicos-e-operacionais-de-hansen--ase-2000-a-2015.pdf

13. BRASIL. Ministério da Saúde. Taxa de Prevalência da Hanseníase, estados, Brasil, 2015. [Internet]. 2016 [cited 2016 Aug 1]. Available from: http://portalsaude.saude.gov.br/images/pdf/2016/julho/07/Taxa-de-Preval--ncia-da-Hansen--ase--estados--Brasil--2015..pdf

14. Scollard DM, Adams LB, Gillis TP, Krahenbuhl JL, Truman RW, Williams DL. The continuing challenges of leprosy. Clin Microbiol Rev. 2006;19(2):338–81.

15. Eichelmann K, González González SE, Salas-Alanis JC, Ocampo-Candiani J. Leprosy. An update: definition, pathogenesis, classification, diagnosis, and treatment. Actas Dermosifiliogr. AEDV; 2013;104(7).

16. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, et al. Massive gene decay in the leprosy bacillus. Nature. 2001;409(6823):1007–11.

17. Monot M, Honore N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi A, et al. Comparative genomic and phylogeographic analysis of Mycobacterium leprae. Nat Genet. 2009;41(12):1282-U39.

18. BRASIL. Ministério da Saúde. Guia para o controle da hanseníase. Série A. Normas e Manuais Técnios. 2002. 2-89 p.

19. Rodrigues LC, Lockwood DNJ. Leprosy now: Epidemiology, progress, challenges, and research gaps. Lancet Infect Dis. Elsevier Ltd; 2011;11(6):464–70.

20. Ridley DS, Jopling WH. Classification of leprosy according to immunity. A five-group system. Int J Lepr other Mycobact Dis. 1966;34(3):255–73.

21. Fava V, Orlova M, Cobat A, Alcaïs A, Mira M, Schurr E. Genetics of leprosy reactions: An overview. Mem Inst Oswaldo Cruz. 2012;107(SUPPL.1):132–42.

22. Sauer MED, Salomão H, Ramos GB, D`Espindula HRS, Rodrigues RSA, Macedo WC, et al. Genetics of leprosy: Expected and unexpected developments and perspectives. Clin Dermatol. Elsevier Inc.; 2015;33(1):99–107.

23. WHO. Leprosy elimination: Classification of leprosy [Internet]. [cited 2016 Jul 1]. Available from: http://www.who.int/lep/classification/en/

24. WHO. Leprosy elimination: WHO Multidrug therapy (MDT) [Internet]. Available from: http://www.who.int/lep/mdt/en/

25. Truman RW, Singh P, Sharma R, Busso P, Rougemont J, Paniz-Mondolfi A, et al. Probable zoonotic leprosy in the southern United States. N Engl J Med. 2011;364(17):1626–33.

26. Neumann A da S, Dias F de A, Ferreira J da S, Fontes ANB, Rosa PS, Macedo RE, et al. Experimental Infection of Rhodnius prolixus (Hemiptera, Triatominae) with Mycobacterium leprae Indicates Potential for Leprosy Transmission. PLoS One. 2016;11(5):e0156037.

27. Fine PE. Leprosy: the epidemiology of a slow bacterium. Epidemiol Rev. 1982;4:161–88.

28. Britton WJ, Lockwood DNJ. Leprosy. Lancet. 2004;363(9416):1209–19.

29. Noordeen SK. The epidemiology of leprosy. In: RC H, editor. Leprosy. Edinburgh: Churchill Livingstone; 1985.

30. Pönnighaus JM, Fine PEM, Sterne JAC, Bliss L, Wilson RJ, Malema SS. Incidence Rates of Leprosy in Karonga District , Northern Malawi : Patterns by Age , Sex , BCG Status and Classification '. 1993;62(I).

31. Oliveira MBB de, Diniz LM. Leprosy among children under 15 years of age: literature review. An Bras Dermatol. 2016;91(2):196–203.

32. Alcaïs A, Abel L, Casanova J. Review series Human genetics of infectious diseases : between proof of principle and paradigm. J Clin Invest. 2009;119(9):2506–14.

33. Chapman SJ, Hill A V. Human genetic susceptibility to infectious disease. Nat Rev Genet. Nature Publishing Group; 2012;13(3):175–88.

34. Fox GJ, Orlova M, Schurr E. Tuberculosis in Newborns: The Lessons of the "Lübeck Disaster" (1929–1933). Vol. 12, PLoS Pathogens. 2016. p. 1–10.

35. Casanova J-L, Abel L. Inborn errors of immunity to infection: the rule rather than the exception. J Exp Med. 2005;202(2):197–201.

36. Casanova J-L. Severe infectious diseases of childhood as monogenic inborn errors of immunity. Proc Natl Acad Sci U S A. 2015;112(51):E7128-37.

37. Casanova J-L, Abel L. Human genetics of infectious diseases: a unified theory. EMBO J. 2007;26(4):915–22.

38. Ozbek N, Fieschi C, Yilmaz BT, de Beaucoudrey L, Demirhan B, Feinberg J, et al. Interleukin-12 receptor beta 1 chain deficiency in a child with disseminated tuberculosis. Clin Infect Dis. 2005;40(6):e55–8.

39. Caragol I, Raspall M, Fieschi C, Feinberg J, Larrosa MN, Hernandez M, et al. Clinical tuberculosis in 2 of 3 siblings with interleukin-12 receptor beta1 deficiency. Clin Infect Dis. 2003;37(2):302–6.

40. Schurr E, Alcais A, de Léséleuc L, Abel L. Genetic predisposition to leprosy: A major gene reveals novel pathways of immunity to Mycobacterium leprae. Semin Immunol. 2006;18(6):404–10.

41. Alter A, Grant A, Abel L, Alcaïs A, Schurr E. Leprosy as a genetic disease. Mamm Genome. 2011;22(1–2):19–31.

42. Cardoso CC, Pereira AC, Sales-Marques C, Moraes MO. Leprosy susceptibility: genetic variations regulate innate and adaptive immunity, and disease outcome. Future Microbiol. 2011;6(11):533–49.

43. Misch EA, Berrington WR, Vary Jr. JC, Hawn TR. Leprosy and the human genome. Microbiol Mol Biol Rev. 2010;74(4):589–620.

44. Gaschignard J, Grant AV, Thuc N Van, Orlova M, Cobat A, Huong NT, et al. Pauci- and Multibacillary Leprosy: Two Distinct, Genetically Neglected Diseases. PLoS Negl Trop Dis. 2016;10(5):e0004345.

45. Ali PM. Genetic influence in leprosy. Indian J Public Health. 1966;10(4):145–57.

46. Chakravartti M, Vogel F. A twin study on leprosy. In: Becker P, Lenz W, Vogel F, Wendt G, editors. Topics in human genetics. Vol 1. Stuttgart: Georg Thieme; 1973. p. 1–123.

47. Abel L, Lap VD, Oberti J, Van Thuc N, Van Cua V, Guilloud-Bataille M, et al. Complex segregation analysis of leprosy in southern Vietnam. Genet Epidemiol. 1995;12(1):63–82.

48. Wagener DK, Schauf V, Nelson KE, Scollard D, Brown A, Smith T. Segregation analysis of leprosy in families of northern Thailand. Genet Epidemiol. 1988;5(0741–0395):95–105.

49. Lázaro FP, Werneck RI, Mackert CCO, Cobat A, Prevedello FC, Pimentel RP, et al. A major gene controls leprosy susceptibility in a hyperendemic isolated population from north of Brazil. J Infect Dis. 2010;201:1598–605.

50. Siddiqui MR, Meisner S, Tosh K, Balakrishnan K, Ghei S, Fisher SE, et al. A major susceptibility locus for leprosy in India maps to chromosome 10p13. Nat Genet. 2001;27(4):439–41.

51. Mira MT, Alcaïs A, Van Thuc N, Thai VH, Huong NT, Ba NN, et al. Chromosome 6q25 is linked to susceptibility to leprosy in a Vietnamese population. Nat Genet. 2003;33(3):412–5.

52. Alter A, De Léséleuc L, Van Thuc N, Thai VH, Huong NT, Ba NN, et al. Genetic and functional analysis of common MRC1 exon 7 polymorphisms in leprosy susceptibility. Hum Genet. 2010;127(3):337–48.

53. Grant A V., Cobat A, Van Thuc N, Orlova M, Huong NT, Gaschignard J, et al. CUBN and NEBL common variants in the chromosome 10p13 linkage region are associated with multibacillary leprosy in Vietnam. Hum Genet. 2014;133(7):883–93.

54. Medeiros P, da Silva WL, de Oliveira Gimenez BB, Vallezi KB, Moraes MO, de Souza VNB, et al. The GATA3 gene is involved in leprosy susceptibility in Brazilian patients. Infect Genet Evol. Elsevier B.V.; 2016;39:194–200.

55. Mira MT, Alcaïs A, Nguyen VT, Moraes MO, Di Flumeri C, Vu HT, et al. Susceptibility to leprosy is associated with PARK2 and PACRG. Nature. 2004;427(6975):636–40.

56. Malhotra D, Darvishi K, Lohra M, Kumar H, Grover C, Sood S, et al. Association study of major risk single nucleotide polymorphisms in the

common regulatory region of PARK2 and PACRG genes with leprosy in an Indian population. Eur J Hum Genet. 2006;14(4):438–42.

57. Li J, Liu H, Liu J, Fu X, Yu Y, Yu G, et al. Association study of the single nucleotide polymorphisms of PARK2 and PACRG with leprosy susceptibility in Chinese population. Eur J Hum Genet. 2012;20(5):488–9.

58. Alter A, Fava VM, Huong NT, Singh M, Orlova M, Van Thuc N, et al. Linkage disequilibrium pattern and age-at-diagnosis are critical for replicating genetic associations across ethnic groups in leprosy. Hum Genet. 2013;132(1):107–16.

59. Chopra R, Ali S, Srivastava AK, Aggarwal S, Kumar B, Manvati S, et al. Mapping of PARK2 and PACRG Overlapping Regulatory Region Reveals LD Structure and Functional Variants in Association with Leprosy in Unrelated Indian Population Groups. PLoS Genet. 2013;9(7).

60. Ali S, Vollaard AM, Widjaja S, Surjadi C, Van De Vosse E, Van Dissel JT. PARK2/PACRG polymorphisms and susceptibility to typhoid and paratyphoid fever. Clin Exp Immunol. 2006;144(3):425–31.

61. Manzanillo PS, Ayres JS, Watson RO, Collins AC, Souza G, Rae CS, et al. The ubiquitin ligase parkin mediates resistance to intracellular pathogens. Nature. Nature Publishing Group; 2013;501(7468):512–6.

62. de Léséleuc L, Orlova M, Cobat A, Girard M, Huong NT, Ba NN, et al. PARK2 Mediates Interleukin 6 and Monocyte Chemoattractant Protein 1 Production by Human Macrophages. PLoS Negl Trop Dis. 2013;7(1).

63. Ramos GB, Salomão H, Francio AS, Fava VM, Werneck RI, Mira MT. Association analysis suggests SOD2 as a new leprosy susceptibility candidate gene. J Infect Dis. 2016;214:475–8.

64. Guerreiro LTA, Robottom-Ferreira AB, Ribeiro-Alves M, Toledo-Pinto TG, Rosa Brito T, Rosa PS, et al. Gene Expression Profiling Specifies Chemokine, Mitochondrial and Lipid Metabolism Signatures in Leprosy. PLoS One. 2013;8(6).

65. Miller EN, Jamieson SE, Joberty C, Fakiola M, Hudson D, Peacock CS, et al. Genome-wide scans for leprosy and tuberculosis susceptibility genes in Brazilians. Genes Immun. 2004;5(1):63–7.

66. Trowsdale J. The MHC, disease and selection. Vol. 137, Immunology Letters. 2011. p. 1–8.

67. Geluk A, Ottenhoff THM. HLA and Leprosy in the Pre and Postgenomic Eras. Vol. 67, Human Immunology. 2006. p. 439–45.

68. Alcaïs A, Alter A, Antoni G, Orlova M, Nguyen VT, Singh M, et al. Stepwise replication identifies a low-producing lymphotoxin-alpha allele as a major risk

factor for early-onset leprosy. Nat Genet. 2007;39(4):517–22.

69. Alter A, Huong NT, Singh M, Orlova M, Van Thuc N, Katoch K, et al. Human leukocyte antigen class I region single-nucleotide polymorphisms are associated with leprosy susceptibility in Vietnam and India. J Infect Dis. 2011;203(9):1274–81.

70. Cardoso CC, Pereira AC, Brito-de-Souza VN, Duraes SMB, Ribeiro-Alves M, Nery JAC, et al. TNF -308G>A single nucleotide polymorphism is associated with leprosy among Brazilians: A genetic epidemiology assessment, meta-analysis, and functional study. J Infect Dis. 2011;204(8):1256–63.

71. Zhang F-R, Huang W, Chen S-M, Sun L-D, Liu H, Li Y, et al. Genomewide association study of leprosy. N Engl J Med. 2009;361(27):2609–2618.

72. Berrington WR, Macdonald M, Khadge S, Raj B, Janer M, Hagge DA, et al. Common polymorphisms in the NOD2 gene region are associated with leprosy and its reactive states. J Infect Dis. 2011;201(9):1422–35.

73. Grant A V, Alter A, Huong NT, Orlova M, Van Thuc N, Ba NN, et al. Crohn's disease susceptibility genes are associated with leprosy in the Vietnamese population. J Infect Dis. 2012;206(11):1763–7.

74. Fava VM, Cobat A, Van Thuc N, Latini ACP, Stefani MMA, Belone AF, et al. Association of TNFSF8 regulatory variants with excessive inflammatory responses but not leprosy per se. J Infect Dis. 2015;211(6):968–77.

75. Fava VM, Manry J, Cobat A, Orlova M, Van Thuc N, Ba NN, et al. A Missense LRRK2 Variant Is a Risk Factor for Excessive Inflammatory Responses in Leprosy. PLoS Negl Trop Dis. 2016;10(2):e0004412.

76. Wong SH, Hill AVS, Vannberg FO. Genomewide association study of leprosy. N Engl J Med. 2010;362(15):1446–1448.

77. Wong SH, Gochhait S, Malhotra D, Pettersson FH, Teo YY, Khor CC, et al. Leprosy and the adaptation of human toll-like receptor 1. PLoS Pathog. 2010;6(7):1–9.

78. Sales-Marques C, Salomão H, Fava VM, Alvarado-Arnez LE, Amaral EP, Cardoso CC, et al. NOD2 and CCDC122-LACC1 genes are associated with leprosy susceptibility in Brazilians. Hum Genet. 2014;133(12):1525–32.

79. Zhang F, Liu H, Chen S, Low H, Sun L, Cui Y, et al. Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. Nat Genet. 2011;43(12):1247–51.

80. Liu H, Irwanto A, Fu X, Yu G, Yu Y, Sun Y, et al. Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. Nat Genet. 2015;47(3):267–71.

81. Yang Q, Liu H, Low HQ, Wang H, Yu Y, Fu X, et al. Chromosome 2p14 is linked to susceptibility to leprosy. PLoS One. 2012;7(1).

82. Tosh K, Meisner S, Siddiqui MR, Balakrishnan K, Ghei S, Golding M, et al. A region of chromosome 20 is linked to leprosy susceptibility in a South Indian population. J Infect Dis. 2002;186(8):1190–3.

83. Cardoso CC, Pereira AC, Brito-De-Souza VN, Dias-Baptista IM, Maniero VC, Venturini J, et al. IFNG +874 T>A single nucleotide polymorphism is associated with leprosy among Brazilians. Hum Genet. 2010;128(5):481–90.

84. Wang D, Feng JQ, Li YY, Zhang DF, Li XA, Li QW, et al. Genetic variants of the MRC1 gene and the IFNG gene are associated with leprosy in Han Chinese from Southwest China. Hum Genet. 2012;131(7):1251–60.

85. Santos AR, Suffys PN, Vanderborght PR, Moraes MO, Vieira LM, Cabello PH, et al. Role of tumor necrosis factor-alpha and interleukin-10 promoter gene polymorphisms in leprosy. J Infect Dis. 2002;186(11):1687–91.

86. Malhotra D, Darvishi K, Sood S, Sharma S, Grover C, Relhan V, et al. IL-10 promoter single nucleotide polymorphisms are significantly associated with resistance to leprosy. Hum Genet. 2005;118(2):295–300.

87. Pereira a C, Brito-de-Souza VN, Cardoso CC, Dias-Baptista IMF, Parelli FPC, Venturini J, et al. Genetic, epidemiological and biological analysis of interleukin-10 promoter single-nucleotide polymorphisms suggests a definitive role for -819C/T in leprosy susceptibility. Genes Immun. 2009;10(2):174–80.

88. Liu H, Irwanto A, Tian H, Fu X, Yu Y, Yu G, et al. Identification of IL18RAP/IL18R1 and IL12B as leprosy risk genes demonstrates shared pathogenesis between inflammation and infectious diseases. Am J Hum Genet. 2012;91(5):935–41.

89. Ali S, Srivastava AK, Chopra R, Aggarwal S, Garg VK, Bhattacharya SN, et al. IL12B SNPs and copy number variation in IL23R gene associated with susceptibility to leprosy. J Med Genet. 2013;50(1):34–42.

90. Abel L, Sánchez FO, Oberti J, Thuc N V., Hoa L Van, Lap VD, et al. Susceptibility to Leprosy is Linked to the Human NRAMP1 Gene. J Infect Dis. 1998;177(1):133–45.

91. Brochado MJF, Gatti MFC, Zago MA, Roselino AM. Association of the solute carrier family 11 member 1 gene polymorphisms with susceptibility to leprosy in a Brazilian sample. Mem Inst Oswaldo Cruz. 2016;111(2):101–5.

92. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456(7218):18–21.

93. Orlova M, Pietrantonio T Di, Schurr E. Genetics of infectious diseases: Hidden etiologies and common pathways. Clin Chem Lab Med. 2011;49(9):1427–37.

94. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. Nature Publishing Group; 2009;461(7265):747–53.

95. Cirulli ET, Goldstein DB. Uncovering the role of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. Nature Publishing Group; 2010;11(6):415–25.

96. Alcaïs A, Quintana-Murci L, Thaler DS, Schurr E, Abel L, Casanova JL. Life-threatening infectious diseases of childhood: Single-gene inborn errors of immunity? Ann N Y Acad Sci. 2010;1214(1):18–33.

97. Morey M, Fernández-Marmiesse A, Castiñeiras D, Fraga JM, Couce ML, Cocho JA. A glimpse into past, present, and future DNA sequencing. Mol Genet Metab. Elsevier Inc.; 2013;110(1–2):3–24.

98. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. Nature Publishing Group; 2009;11(1):31–46.

99. Alter A, Alcaïs A, Abel L, Schurr E. Leprosy as a genetic model for susceptibility to common infectious diseases. Hum Genet. 2008;123(3):227–35.

100. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Publ Gr. Nature Publishing Group; 2011;12(11):745–55.

101. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. G3 (Bethesda). 2015;5(August):g3.115.018564-.

102. Merriman B, Torrent I, Rothberg JM. Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis. 2012;33(23):3397–417.

103. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. Nature Publishing Group; 2011;475(7356):348–52.

104. Bao R, Huang L, Andrade J, Tan W, Kibbe W a, Jiang H, et al. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. Lib Acad. 2014;13:67–82.

105. Pavlopoulos GA, Oulas A, Iacucci E, Sifrim A, Moreau Y, Schneider R, et al. Unraveling genomic variation from next generation sequencing data. BioData Min. 2013;6(1):13.

106. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. 2013. 1-33 p.

107. Chan E. Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery. In: Komar A, editor. Single Nucleotide Polymorphisms. 2nd ed. Humana Press; 2009. p. 456.

108. Metzker ML. Emerging technologies in DNA sequencing. Genome Res. 2005;15(12):1767–76.

109. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.

110. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

111. Lek M, Karczewski KJ, Samocha KE, Banks E, Fennell T, O AH, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. Nature Publishing Group; 2016;536:285–92.

112. Itan Y, Casanova J-L. Can the impact of human genetic variations be predicted? Proc Natl Acad Sci. 2015;112(37):11426–7.

113. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–81.

114. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. Nature Publishing Group; 2010;7(4):248–9.

115. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, et al. Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci U S A. 2015;112(37):E5189-98.

116. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics. 2013. 1-41 p.

117. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. Nature Publishing Group; 2014;46(3):310–5.

118. Combined Annotation Dependent Depletion (CADD) [Internet]. University of Washington and Hudson-Alpha Institute for Biotechnology. [cited 2016 Mar 1]. Available from: http://cadd.gs.washington.edu/info

119. Nomura A, Tada H, Teramoto R, Konno T, Hodatsu A, Won HH, et al. Whole exome sequencing combined with integrated variant annotation prediction identifies a causative myosin essential light chain variant in hypertrophic cardiomyopathy. J Cardiol. Japanese College of Cardiology; 2016;67(2):133-9.

120. Tada H, Kawashiri M aki, Nohara A, Saito R, Tanaka Y, Nomura A, et al. Whole exome sequencing combined with integrated variant annotation prediction identifies asymptomatic tangier disease with compound heterozygous mutations in ABCA1 gene. Atherosclerosis. Elsevier Ltd; 2015;240(2):324–9.

121. Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Vélez M, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci U S A. 2015;112(44):13615–20.

122. Human Gene Damage Index (GDI) [Internet]. The Rockefeller University. [cited 2016 Jan 1]. Available from: http://lab.rockefeller.edu/casanova/GDI

123. Ministério da Saúde, "Vigilância em saúde: situação epidemiológica da hanseníase no Brasil" [Internet]. 2008 [cited 2016 Jun 1]. Available from: http://bvsms.saude.gov.br/bvs/publicacoes/vigilancia_saude_situacao_hanseniase.pdf

124. Product Bulletin: exome sequencing using the Ion Proton™ System [Internet]. Thermo Fisher Scientific. [cited 2016 Jan 1]. Available from: http://tools.thermofisher.com/content/sfs/brochures/Proton-Exome-Product-Bulletin.pdf

125. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009;19(7):1316–23.

126. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: An update on mammalian reference sequences. Nucleic Acids Res. 2014;42(D1):756–63.

127. Torrent Suite [Internet]. Thermo Fisher Scientific. [cited 2015 Jul 1]. Available from: https://github.com/iontorrent/TS

128. Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. 2012;433–7.

129. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.

130. GATK Best Practices [Internet]. Broad Institute. [cited 2015 Jul 1]. Available from: https://www.broadinstitute.org/gatk/guide/best-practices.php

131. John SWM, Weitzner G, Rozen R, Scriver CR. A rapid procedure for extracting genomic DNA from leukocytes. Nucleic Acids Res. 1991;19(2):408.

132. Ion TargetSeq ™ Exome Enrichment for the Ion Proton ™ System User Guide [Internet]. Thermo Fisher Scientific. Available from:

https://tools.thermofisher.com/content/sfs/manuals/MAN0006730_TargetSeqExomeEnrich_IonProton_UG.pdf

133. Ion PI™ Template OT2 200 Kit v2 User Guide [Internet]. Vol. MAN0007624, Thermo Fisher Scientific. Available from: http://hwwgenotyping.ksu.edu/protocols/MAN0007624_Ion PI Template OT2 200 Kit v2_UG_Rev2_01July2013.pdf

134. Ion PI™ Sequencing 200 Kit v2 User Guide [Internet]. Vol. MAN0007961, Thermo Fisher Scientific. Available from: https://tools.thermofisher.com/content/sfs/manuals/MAN0007961_Ion_PI_Sequencing_Kit_v2_UG.pdf

135. Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. Babraham Bioinformatics. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

136. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

137. Picard tools [Internet]. Broad Institute. Available from: http://broadinstitute.github.io/picard/

138. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2015;32(2):292–4.

139. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.

140. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

141. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

142. Ion Community [Internet]. Thermo Fisher Scientific. Available from: https://ioncommunity.thermofisher.com/welcome

143. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007;23(10):1289–91.

144. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):1–12.

145. Kent J. UCSC In-Silico PCR [Internet]. University of California Santa Cruz. Available from: https://genome.ucsc.edu/cgi-bin/hgPcr

146. BigDye™ Terminator v3.1 Cycle Sequencing Kit User Guide [Internet]. Thermo Fisher Scientific. p. 1–9. Available from: https://tools.thermofisher.com/content/sfs/manuals/cms_081527.pdf

147. Burland TG. DNASTAR's Lasergene Sequence Analysis Software. Methods Mol Biol. 2000;132:71–91.

148. TruSeq DNA Sample Preparation Guide [Internet]. Illumina. 2012. Available from: http://support.illumina.com/content/dam/illumina-support/documents/myillumina/f5f619d3-2c4c-489b-80a3-e0414baa4e89/truseq_dna_sampleprep_guide_15026486_c.pdf

149. HiSeq ® 2500 System Guide [Internet]. Illumina. 2015. Available from: http://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseq2500/hiseq-2500-system-guide-15035786-01.pdf

150. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95.

151. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. Nat Genet. 2008;40(2):124–5.

152. Itan Y, Shang L, Boisson B, Ciancanelli MJ, Markle JG, Martinez-barricarte R, et al. The mutation significance cutoff: gene- level thresholds for variant predictions. Nat Publ Gr. Nature Publishing Group; 2016;13(2):109–10.

153. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263–5.

154. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30(5):434–9.

155. Torrent Variant Detection Algorithms [Internet]. Thermo Fisher Scientific. 2015. Available from: http://129.130.90.13/ion-docs/White-Paper---Torrent-Variant-Detection-Algorithms_15007750.html

156. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015;112(17):5473–8.

157. Wang D, Xu L, Lv L, Su L-Y, Fan Y, Zhang D-F, et al. Association of the LRRK2 genetic polymorphisms with leprosy in Han Chinese from Southwest China. Genes Immun. 2014;16(2):112–9.

158. Marcinek P, Jha AN, Shinde V, Sundaramoorthy A, Rajkumar R, Suryadevara NC, et al. LRRK2 and RIPK2 Variants in the NOD 2-Mediated Signaling Pathway Are Associated with Susceptibility to Mycobacterium leprae in Indian Populations. PLoS One. 2013;8(8).

159. Wallings R, Manzoni C, Bandopadhyay R. Cellular processes associated with LRRK2 function and dysfunction. FEBS J. 2015;282(15):2806–26.

160. Ito G, Okai T, Fujino G, Takeda K, Ichijo H, Katada T, et al. GTP binding is essential to the protein kinase activity of LRRK2, a causative gene product for familial Parkinson's disease. Biochemistry. 2007;46(5):1380–8.

161. Cookson MR. LRRK2 Pathways Leading to Neurodegeneration. Curr Neurol Neurosci Rep. 2015;15(7):564.

162. Smith WW, Pei Z, Jiang H, Moore DJ, Liang Y, West AB, et al. Leucine-rich repeat kinase 2 (LRRK2) interacts with parkin, and mutant LRRK2 induces neuronal degeneration. Proc Natl Acad Sci U S A. 2005;102(51):18676–81.

163. Corti O, Lesage S, Brice A. What genetics tells us about the causes and mechanisms of Parkinson's disease. Physiol Rev. 2011;91(4):1161–218.

164. Guaitoli G, Raimondi F, Gilsbach BK, Gómez-Llorente Y, Deyaert E, Renzi F, et al. Structural model of the dimeric Parkinson's protein LRRK2 reveals a compact architecture involving distant interdomain contacts. Proc Natl Acad Sci. 2016;201523708.

165. Liu Z, Lee J, Krummey S, Lu W, Cai H, Lenardo MJ. The kinase LRRK2 is a regulator of the transcription factor NFAT that modulates the severity of inflammatory bowel disease. Nat Immunol. 2011;12(11):1063–70.

166. Lewis P a, Manzoni C. LRRK2 and human disease: a complicated question or a question of complexes? Sci Signal. 2012;5(207):pe2.

167. Behr M a, Schurr E. Mycobacteria in Crohn's disease: a persistent hypothesis. Inflamm Bowel Dis. 2006;12(10):1000–4.

168. Liu Z, Lenardo MJ. The role of LRRK2 in inflammatory bowel disease. Cell Res. Nature Publishing Group; 2012;22:1092–4.

169. Gardet A, Benita Y, Li C, Sands BE, Ballester I, Stevens C, et al. LRRK2 is involved in the IFN-gamma response and host response to pathogens. J Immunol. 2010;185(9):5577–85.

170. Zhang Q, Pan Y, Yan R, Zeng B, Wang H, Zhang X, et al. Commensal bacteria direct selective cargo sorting to promote symbiosis. Nat Immunol. 2015;16(9):918–26.

171. Rocha JDB, Schlossmacher MG, Philpott DJ. LRRK2 and Nod2 promote lysozyme sorting in Paneth cells. Nat Publ Gr. Nature Publishing Group; 2015;16(9):898–900.

172. Ross OA, Soto-Ortolaza AI, Heckman MG, Aasly JO, Abahuni N, Annesi G, et al. Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: A case-control study. Lancet Neurol. 2011;10(10):898–908.

173. Tan EK, Peng R, Teo YY, Tan LC, Angeles D, Ho P, et al. Multiple LRRK2 variants modulate risk of Parkinson disease: A Chinese multicenter study. Hum Mutat. 2010;31(5):561–8.

174. Nixon-Abell J, Berwick DC, Grannó S, Spain VA, Blackstone C, Harvey K. Protective LRRK2 R1398H Variant Enhances GTPase and Wnt Signaling Activity. Front Mol Neurosci. 2016;9(March):18.

175. Waschbüsch D, Michels H, Strassheim S, Ossendorf E, Kessler D, Gloeckner CJ, et al. LRRK2 transport is regulated by its novel interacting partner Rab32. PLoS One. 2014;9(10).

176. Neefjes J, Jongsma ML, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. Nature Publishing Group; 2011;11(12):823–36.

177. Jarduli LR, Sell AM, Reis PG, Sippert EÂ, Ayo CM, Mazini PS, et al. Role of HLA, KIR, MICA, and cytokines genes in leprosy. Biomed Res Int. 2013;2013.

178. Bodmer WF. The HLA system: structure and function. J Clin Pathol. 1987;40:948–58.

179. Schlesinger LS, Horwitz MA. Phagocytosis of leprosy bacilli is mediated by complement receptors CR1 and CR3 on human monocytes and complement component C3 in serum. J Clin Invest. 1990;85(4):1304–14.

180. Schlesinger LS, Horwitz M a. Phagocytosis of Mycobacterium leprae by human monocyte-derived macrophages is mediated by complement receptors CR1 (CD35), CR3 (CD11b/CD18), and CR4 (CD11c/CD18) and IFN-gamma activation inhibits complement receptor function and phagocytosis of this bacte. J Immunol. 1991;147(6):1983–94.

181. Borzutzky A, Fried A, Chou J, Bonilla FA, Kim S, Dedeoglu F. NOD2-associated diseases: Bridging innate immunity and autoinflammation. Clin Immunol. Elsevier Inc.; 2010;134(3):251–61.

182. Naser SA, Arce M, Khaja A, Fernandez M, Naser N, Elwasila S, et al. Role of ATG16L, NOD2 and IL23R in Crohn's disease pathogenesis. World J Gastroenterol. 2012;18(5):412–24.

183. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell. 2007;131(5):861–72.

184. Charpentier E, Marraffini LA. Harnessing CRISPR-Cas9 immunity for genetic engineering. Curr Opin Microbiol. Elsevier Ltd; 2014;19(1):114–9.

APPENDIX 1 – COMMAND LINES USED FROM ALIGNMENT TO VARIANT CALLING STEPS

WES data processing workflow

The reads were mapped to GRCh37 (hg19) decoy reference using TMAP map4 command from Torrent Suit 5.0 (TS) software creating one alignment file *per* sample run.

```
$ tmap map4 \
-f human_g1k_v37_decoy.fasta \
-r WES_RawData_${sample}_${library}.bam \
-s WES_aligned_${sample}_${library}.bam \
-v -Y -u -o 2
```

Mapped reads were sorted according to their genomic coordinate position using Picard. SortSam algorithm.

```
$ java -jar picard.jar SortSam \
INPUT= WES_aligned_${sample}_${library}.bam \
OUTPUT=WES_sorted_${sample}_${lane}.bam \
SORT_ORDER=coordinate
```

All sorted files for each individual sample run were merged into a unique BAM file *per* sample with MergeSamFiles algorithm.

```
$ java -jar picard.jar MergeSamFiles \
INPUT=WES_sorted_${sample}_Library1.bam \
INPUT=WES_sorted_${sample}_Library2.bam \
INPUT=WES_sorted_${sample}_Library3.bam \
INPUT=WES_sorted_${sample}_Library4.bam \
OUTPUT=WES_merged_$sample.bam \
ASSUME_SORTED=true
```

PCR duplicates were marked with Picard MarkDuplicates.

```
$ java -jar picard.jar MarkDuplicates \
INPUT=WES_merged_$sample.bam \
OUTPUT=WES_dedup_$sample.bam \
METRICS_FILE=metrics_WES_dedup_$sample.txt \
ASSUME_SORTED=true
```

Reads shorter than 30 bp were excluded from the analysis. Next, the reads were indexed with Picard BuildBamIndex.

```
$ samtools view -h WES_dedup_$sample.bam | awk 'length($10) >= 30 || $1 ~
/^@/' | samtools view -bS - > WES_dedup_30pb_$sample.bam

$ java -jar picard.jar BuildBamIndex \
INPUT=WES_dedup_30pb_$sample.bam \
OUTPUT=WES_dedup_30pb_$sample.bam.bai
```

The variant calling was performed with Torrent variant caller (TVC) plugin from TS software, using “Germline - Proton TargetSeq - High stringency” parameter option with default settings.

```
$ tvc \
--input-bam WES_dedup_30pb_$sample.bam \
--reference human_g1k_v37_decoy.fasta \
--output-vcf WES_Variants_$sample.vcf \
--parameters-file targetseq_germline_highstringency_p1_parameters.json
```

The lists of checked variants from all the samples were combined in one multi-samples VCF file using CombineVariants tool in GATK.

```
$ java -jar GenomeAnalysisTK.jar \
-T CombineVariants \
-R human_g1k_v37_decoy.fasta \
--variant WES_Variants_Grandmother.vcf \
--variant WES_Variants_Father.vcf \
--variant WES_Variants_Mother.vcf \
--variant WES_Variants_Twin1.vcf \
--variant WES_Variants_Twin2.vcf \
-o WES_Variants_AllSamples.vcf \
-genotypeMergeOptions UNIQUIFY
```

VCF file containing only variants located inside target regions (from Ion TargetSeq™ Exome probes) was created using GATK SelectVariants.

```
$ java -jar GenomeAnalysisTK.jar \
-T SelectVariants \
-R human_g1k_v37_decoy.fasta \
-V WES_Variants_AllSamples.vcf \
-o OnTarget_WES_Variants_Allsamples.vcf \
-L Ion-TargetSeq-Exome-50Mb-hg19_revA.bed
```

WGS data processing workflow

The reads were mapped to GRCh37 (hg19) decoy reference using the BWA mem, creating one alignment file (SAM) *per* sample run.

```
$ bwa mem human_g1k_v37_decoy.fasta RawData_${sample}_${lane}.R1.fastq.gz
WGS_RawData_${sample}_${lane}.R2.fastq.gz | gzip -3 >
WGS_aligned_${sample}_${lane}.sam.gz
```

Mapped reads were sorted according to their genomic coordinate position using Picard SortSam.

```
$ java -jar picard.jar SortSam \
INPUT=WGS_aligned_${sample}_${lane}.sam.gz \
OUTPUT=WGS_sorted_${sample}_${lane}.bam \
SORT_ORDER=coordinate
```

Next, library BAMs from the same sample were merged with Picard MergeSamFiles.

```
$ java -jar picard.jar MergeSamFiles \
INPUT=WGS_sorted_${sample}_L1.bam \
INPUT=WGS_sorted_${sample}_L2.bam \
INPUT=WGS_sorted_${sample}_L3.bam \
INPUT=WGS_sorted_${sample}_L4.bam \
OUTPUT=WGS_merged_$sample.bam \
ASSUME_SORTED=true
```

PCR duplicates were marked with Picard MarkDuplicates.

```
$ java -jar picard.jar MarkDuplicates \
INPUT=WGS_merged_$sample.bam \
OUTPUT=WGS_dedup_$sample.bam \
METRICS_FILE=metrics_WGS_dedup_$sample.txt \
ASSUME_SORTED=true
```

The alignment files (BAM) were then indexed and assigned read groups that are necessary for GATK.

```
$ java -jar picard.jar AddOrReplaceReadGroups \
INPUT=WGS_dedup_$sample.bam \
OUTPUT=WGS_AddGroup_dedup_$sample.bam \
RGLB=$Barcode \
RGSM=$sample \
RGPL=illumina \
```

```
RGPU=$Run \
CREATE_INDEX=true
```

Local realignment around indels was performed with GATK RealignerTargetCreator and IndelRealigner.

```
$ java -jar GenomeAnalysisTK.jar \
-T RealignerTargetCreator \
-R human_g1k_v37_decoy.fasta \
-I WGS_AddGroup_dedup_$sample.bam \
-known Mills_and_1000G_gold_standard.indels.b37.vcf \
-known 1000G_phase1.indels.b37.vcf \
-o target_localrealignment_WGS_$sample.list \
--filter_mismatching_base_and_quals

$ java -jar GenomeAnalysisTK.jar \
-T IndelRealigner \
-R human_g1k_v37_decoy.fasta \
-I WGS_AddGroup_dedup_$sample.bam \
-targetIntervals target_localrealignment_WGS_$sample.list \
-known Mills_and_1000G_gold_standard.indels.b37.vcf \
-known 1000G_phase1.indels.b37.vcf \
-o WGS_realigned_$sample.bam \
--filter_mismatching_base_and_quals
```

Base quality scores were then recalibrated using GATK BaseRecalibrator, AnalyzeCovariates and PrintReads commands.

```
$ java -jar GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R human_g1k_v37_decoy.fasta \
-I WGS_realigned_$sample.bam \
-knownSites Mills_and_1000G_gold_standard.indels.b37.vcf \
-knownSites 1000G_phase1.indels.b37.vcf \
-knownSites dbsnp_138.b37.vcf \
-o WGS_Table_recal_$sample.table

$ java -jar GenomeAnalysisTK.jar
-T BaseRecalibrator \
-R human_g1k_v37_decoy.fasta \
-I WGS_realigned_$sample.bam \
-knownSites Mills_and_1000G_gold_standard.indels.b37.vcf \
-knownSites 1000G_phase1.indels.b37.vcf \
-knownSites dbsnp_138.b37.vcf \
-BQSR WGS_Table_recal_$sample.table \
-o WGS_Table_Pos-recal_$sample.table \

$ java -jar GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R human_g1k_v37_decoy.fasta \
-before WGS_Table_recal_$sample.table \
-after WGS_Table_Pos-recal_$sample.table \
-plots recalibration_plots_WGS_$sample.pdf \
```

```
$ java -jar GenomeAnalysisTK.jar \
-T PrintReads \
-R human_g1k_v37_decoy.fasta \
-I WGS_realigned_$sample.bam \
-BQSR WGS_Table_recal_$sample.table \
-o WGS_recal_$sample.bam
```

GATK HaplotypeCaller was used to call variants for each sample, followed by JointGenotyped for all samples together. These steps were performed *per* chromosome and output files were merged using GATK CatVariants.

```
$ java -jar GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R human_g1k_v37_decoy.fasta \
-I WGS_recal_$sample.bam \
-L $chr \
--dbsnp dbsnp_138.b37.vcf \
--emitRefConfidence GVCF \
-o raw_HC_WGS_${sample}_Chr$chr.g.vcf

$ java -jar GenomeAnalysisTK.jar \
-T GenotypeGVCFs \
-R human_g1k_v37_decoy.fasta \
--variant raw_HC_WGS_Grandmother_Chr$chr.g.vcf \
--variant raw_HC_WGS_Aunt_Chr$chr.g.vcf \
--variant raw_HC_WGS_Father_Chr$chr.g.vcf \
--variant raw_HC_WGS_Mother_Chr$chr.g.vcf \
--variant raw_HC_WGS_Twin1_Chr$chr.g.vcf \
--variant raw_HC_WGS_Twin2_Chr$chr.g.vcf \
-o GVCFs_WGS_AllSamples_Chr$chr.vcf

$ java -cp GenomeAnalysisTK.jar org.broadinstitute.gatk.tools.CatVariants \
-R human_g1k_v37_decoy.fasta \
-V GVCFs_WGS_AllSamples_Chr1.vcf -V GVCFs_WGS_AllSamples_Chr2.vcf
-V GVCFs_WGS_AllSamples_Chr3.vcf -V GVCFs_WGS_AllSamples_Chr4.vcf
-V GVCFs_WGS_AllSamples_Chr5.vcf -V GVCFs_WGS_AllSamples_Chr6.vcf
-V GVCFs_WGS_AllSamples_Chr7.vcf -V GVCFs_WGS_AllSamples_Chr8.vcf
-V GVCFs_WGS_AllSamples_Chr9.vcf -V GVCFs_WGS_AllSamples_Chr10.vcf
-V GVCFs_WGS_AllSamples_Chr11.vcf -V GVCFs_WGS_AllSamples_Chr12.vcf
-V GVCFs_WGS_AllSamples_Chr13.vcf -V GVCFs_WGS_AllSamples_Chr14.vcf
-V GVCFs_WGS_AllSamples_Chr15.vcf -V GVCFs_WGS_AllSamples_Chr16.vcf
-V GVCFs_WGS_AllSamples_Chr17.vcf -V GVCFs_WGS_AllSamples_Chr18.vcf
-V GVCFs_WGS_AllSamples_Chr19.vcf -V GVCFs_WGS_AllSamples_Chr20.vcf
-V GVCFs_WGS_AllSamples_Chr21.vcf -V GVCFs_WGS_AllSamples_Chr22.vcf
-V GVCFs_WGS_AllSamples_ChrX.vcf \
-out GVCFs_WGS_AllSamples_AllChr.vcf \
--assumeSorted \
```

GATK Variant Quality Score Recalibration (VQSR), with parameters as default, was used to evaluate the likelihood of a variant being real in order to reduce the amount of false positive.

```
$ java -jar GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R human_g1k_v37_decoy.fasta \
-input GVCFs_WGS_AllSamples_AllChr.vcf \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap_3.3.b37.vcf \
-resource:omni,known=false,training=true,truth=true,prior=12.0
1000G_omni2.5.b37.vcf \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
1000G_phase1.snps.high_confidence.b37.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_138.b37.vcf \
-an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an ReadPosRankSum \
-mode SNP \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
-recalFile WGS_AllSamples_recalibrate_SNP.recal \
-tranchesFile WGS_AllSamples_recalibrate_SNP.tranches \
-rscriptFile WGS_AllSamples_recalibrate_SNP_plots.R

$ java -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R human_g1k_v37_decoy.fasta \
-input GVCFs_WGS_AllSamples_AllChr.vcf \
-mode SNP \
--ts_filter_level 99.9 \
-recalFile WGS_AllSamples_recalibrate_SNP.recal \
-tranchesFile WGS_AllSamples_recalibrate_SNP.tranches \
-o WGS_AllSamples_AllChr_recalibrated_snps_raw_indels.vcf

$ java -jar GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R human_g1k_v37_decoy.fasta \
-input WGS_AllSamples_AllChr_recalibrated_snps_raw_indels.vcf \
-resource:mills,known=false,training=true,truth=true,prior=12.0
Mills_and_1000G_gold_standard.indels.b37.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_138.b37.vcf \
-an QD -an DP -an FS -an SOR -an MQRankSum -an ReadPosRankSum \
-mode INDEL \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
--maxGaussians 4 \
-recalFile WGS_AllSamples_recalibrate_INDEL.recal \
-tranchesFile WGS_AllSamples_recalibrate_INDEL.tranches \
-rscriptFile WGS_AllSamples_recalibrate_INDEL_plots.R

$ java -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R human_g1k_v37_decoy.fasta \
-input WGS_AllSamples_AllChr_recalibrated_snps_raw_indels.vcf \
-mode INDEL \
--ts_filter_level 99.0 \
-recalFile WGS_AllSamples_recalibrate_INDEL.recal \
-tranchesFile WGS_AllSamples_recalibrate_INDEL.tranches \
-o WGS_AllSamples_AllChr_recalibrated_variants.vcf
```

Then, genotype refinement workflow from GATK was applied to filter *per* sample genotype calls that were not reliable enough for downstream analysis. For each sample, genotypes with quality score (GQ) lower than GQ20 were flagged as low quality genotype.

```
$ java -jar GenomeAnalysisTK.jar \
-T CalculateGenotypePosteriors \
-R human_g1k_v37_decoy.fasta \
--supporting 1000G_phase3_v4_20130502.sites.vcf \
-ped pedigree_for_GATK_Father_Mother_Twin2.txt \
--pedigreeValidationType SILENT \
-V WGS_AllSamples_AllChr_recalibrated_variants.vcf \
-o WGS_AllSamples_AllChr_recalibrated_variants.postCGP.vcf


$ java -jar GenomeAnalysisTK.jar \
-T VariantFiltration \
-R human_g1k_v37_decoy.fasta \
-V WGS_AllSamples_AllChr_recalibrated_variants.postCGP.vcf \
-G_filter "GQ < 20.0" \
-G_filterName lowGQ \
```

APPENDIX 2 – SUPPLEMENTARY DATA

Sanger sequencing was performed in exonic regions comprising ten candidate variants identified in WES. *Per* sample variants' genotypes from WES and Sanger data are shown in **figures S1** to **S10**, as well as Sanger electropherogram.

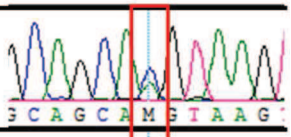
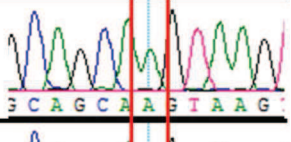
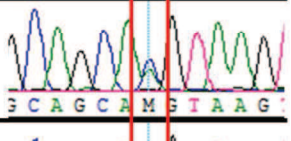
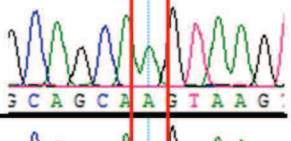
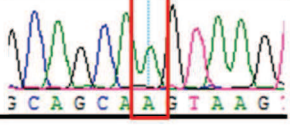
	Ion Proton	Sanger	Electropherogram
Grandmother	C/A	C/A	
Father	A/A	A/A	
Mother	C/A	C/A	
Twin1	A/A	A/A	
Twin2	A/A	A/A	

Figure S1. Comparison between WES and Sanger sequencing results for rs9901673 (Q254K) in *CD68* gene. The electropherogram refers to nucleotides from the complementary strand. Red rectangle correnponds to variant nucleotide peak. M: A or C (IUPAC nucleotide code).

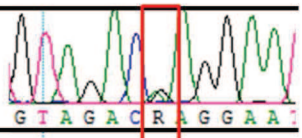
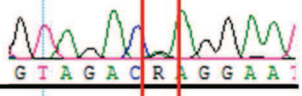
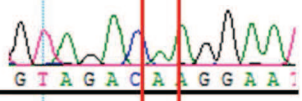
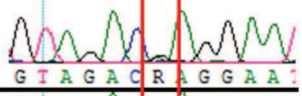
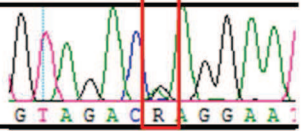
	Ion Proton	Sanger	Electropherogram
Grandmother	A/G	A/G	
Father	A/G	A/G	
Mother	A/A	A/A	
Twin1	A/G	A/G	
Twin2	A/G	A/G	

Figure S2. Comparison between WES and Sanger sequencing results for a novel missense variant (K576E) in *CP* gene. Red rectangle correnponds to variant nucleotide peak. R: A or G (IUPAC nucleotide code).

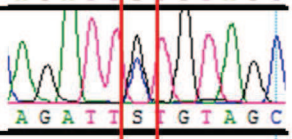
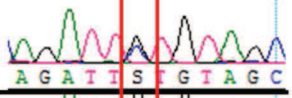
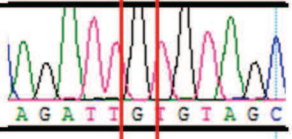
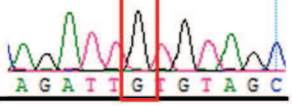
	Ion Proton	Sanger	Electropherogram
Grandmother	C/G	C/G	
Father	C/G	C/G	
Mother	C/G	C/G	
Twin1	G/G	G/G	
Twin2	G/G	G/G	

Figure S3. Comparison between WES and Sanger sequencing results for rs58154316 (S284C) in *HRH4* gene. Red rectangle correnponds to variant nucleotide peak. S: G or C (IUPAC nucleotide code).

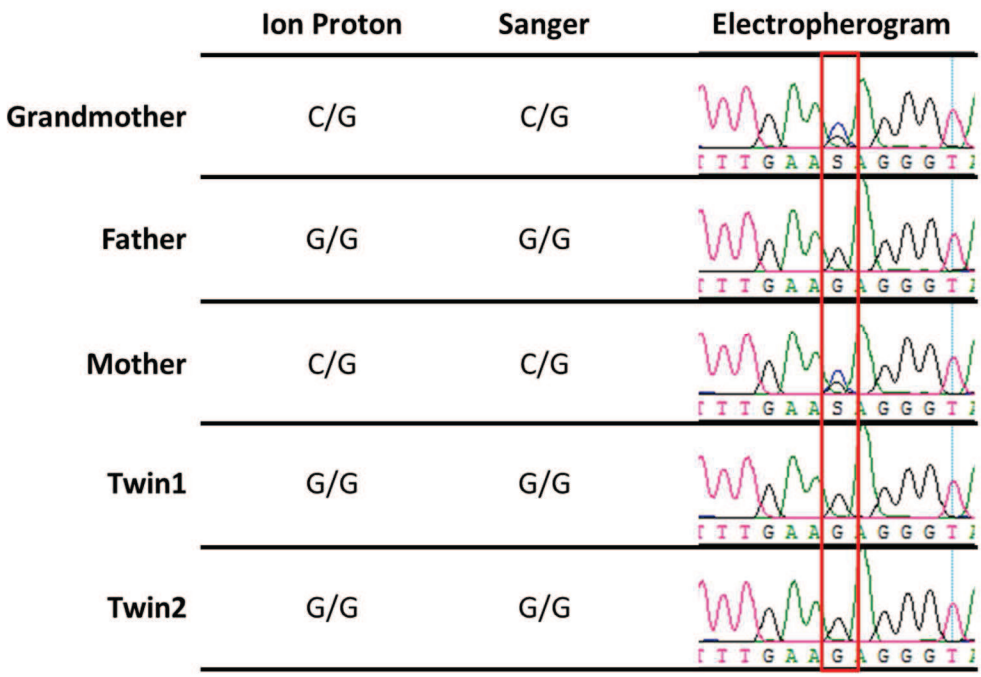


Figure S4. Comparison between WES and Sanger sequencing results for rs7308720 (N551K) in *LRRK2* gene and comparison with WES results. Red rectangle correnponds to variant nucleotide peak. S: G or C (IUPAC nucleotide code).

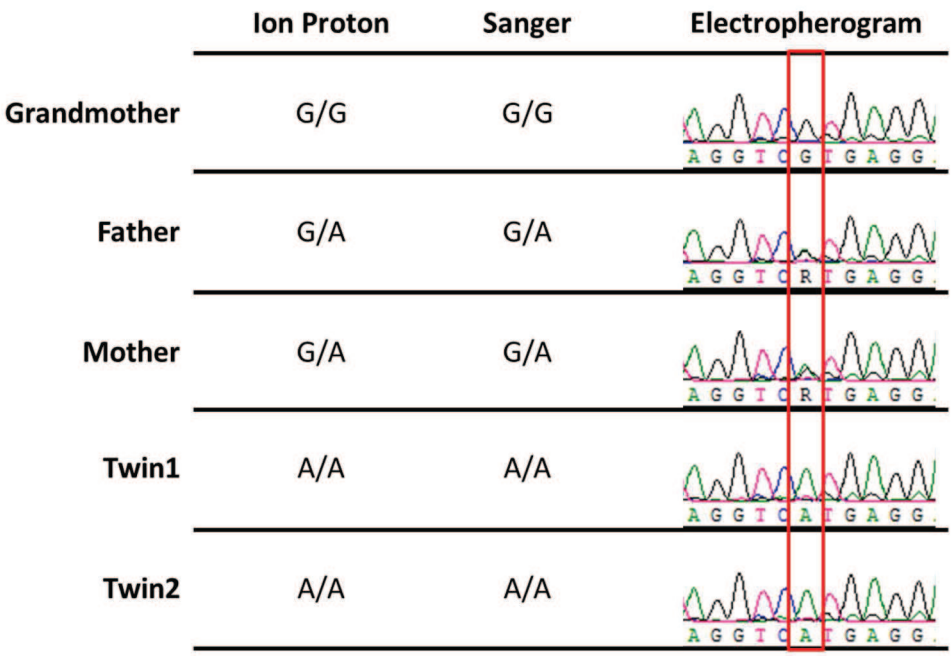


Figure S5. Comparison between WES and Sanger sequencing results for rs7133914 (R1398H) in *LRRK2* gene. Red rectangle correnponds to variant nucleotide peak. R: A or G (IUPAC nucleotide code).

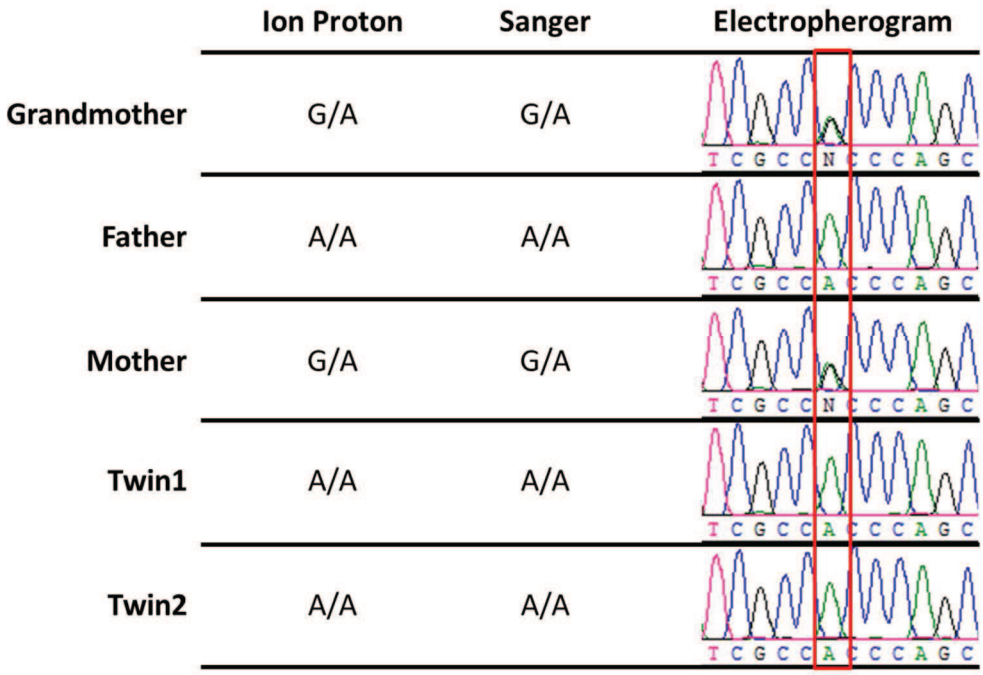


Figure S6. Comparison between WES and Sanger sequencing results for rs10852891 (A229T) in *MPDU1* gene. Red rectangle correnponds to variant nucleotide peak. R: A or G (IUPAC nucleotide code).

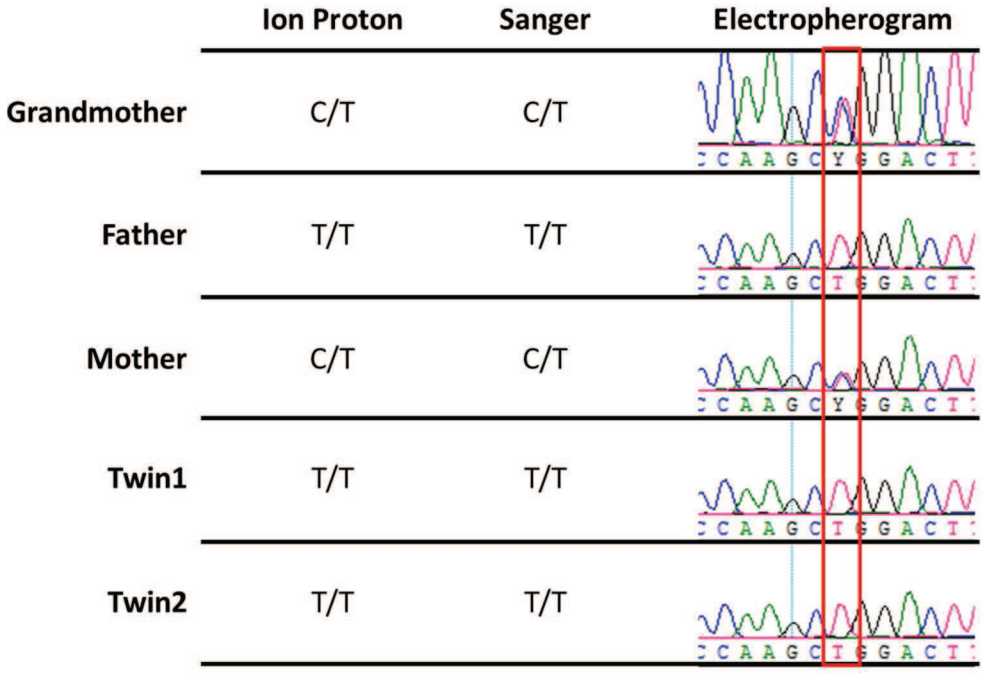


Figure S7. Comparison between WES and Sanger sequencing results for rs17585 (P170L) in *RNH1* gene. Red rectangle correnponds to variant nucleotide peak. Y: C or T (IUPAC nucleotide code).

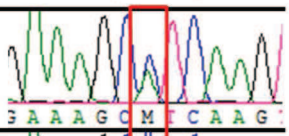
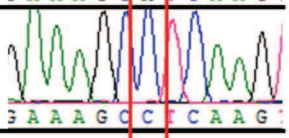
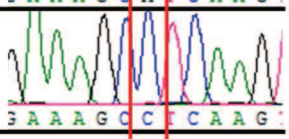
	Ion Proton	Sanger	Electropherogram
Grandmother	A/C	A/C	
Father	C/C	C/C	
Mother	A/C	A/C	
Twin1	C/C	C/C	
Twin2	C/C	C/C	

Figure S8. Comparison between WES and Sanger sequencing results for rs6091375 (I798L) in *SALL4* gene. Red rectangle correnponds to variant nucleotide peak. M: A or C (IUPAC nucleotide code).

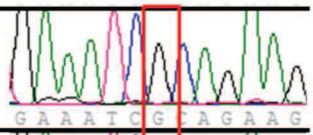
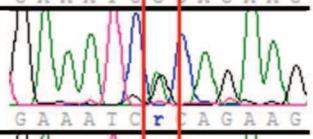
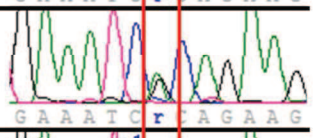
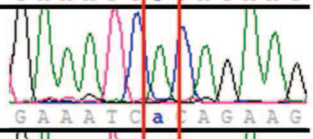
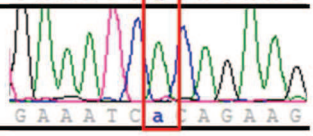
	Ion Proton	Sanger	Electropherogram
Grandmother	G/G	G/G	
Father	G/A	G/A	
Mother	G/A	G/A	
Twin1	A/A	A/A	
Twin2	A/A	A/A	

Figure S9. Comparison between WES and Sanger sequencing results for rs61755579 (A208T) in *SOS2* gene. Red rectangle correnponds to variant nucleotide peak. R: A or G (IUPAC nucleotide code).

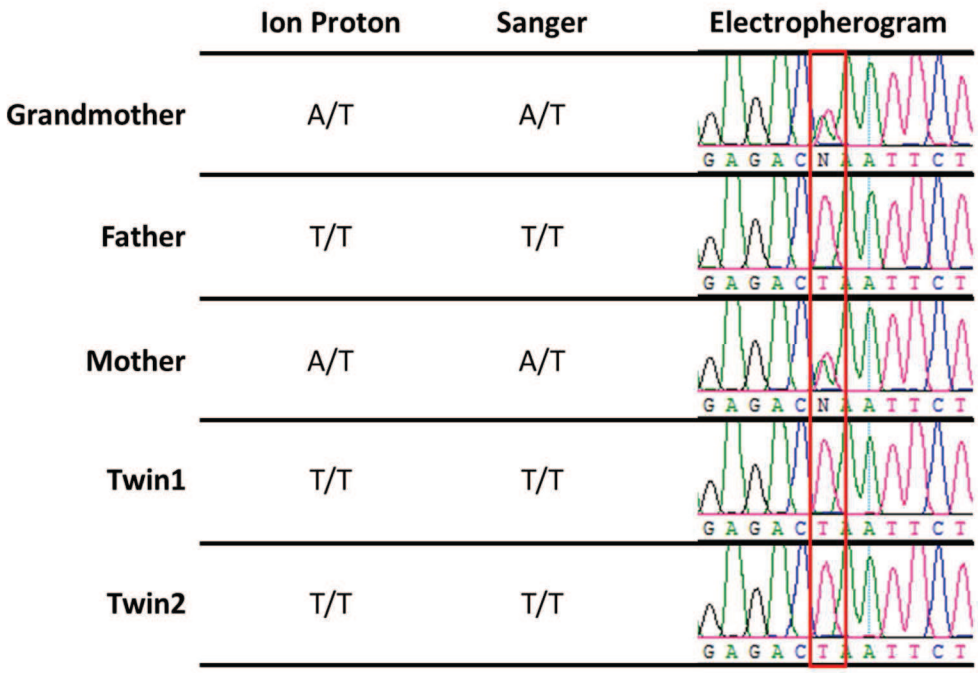


Figure S10. Comparison between WES and Sanger sequencing results for rs179008 (Q11L) in *TLR7* gene. Red rectangle correnponds to variant nucleotide peak. R: A or G (IUPAC nucleotide code).

Table S1 shows MAFs of known candidate variants in two public databases: 1000G and ExAC. The frequencies are shown for five population samples from both databases as well as a mean MAF of all samples together per database.

(MAF) of candidate variants previously reported in different population samples from two public databases:) and Exome Aggregation Consortium (ExAC). ALL refers to mean MAF from all population samples together. can/African American (AFR), AdMixed American/Latin (AMR), East Asian (EAS), European (EUR) and South wn in bold.

	ALL		AFR		AMR		EAS		EUR*		SAS	
	1000G	ExAC	1000G	ExAC	1000G	ExAC	1000G	ExAC	1000G	ExAC	1000G	ExAC
1	-	15.7%	-	**22.6%**	-	13.7%	-	15.0%	-	15.1%	-	17.7%
ns)												
3	16.0%	15.5%	17.0%	**18.3%**	12.0%	9.0%	16.0%	16.1%	16.0%	16.8%	17.0%	15.0%
1	18.0%	15.7%	**27.0%**	-	13.0%	-	13.0%	-	10.0%	-	22.0%	-
24	18.0%	15.6%	**28.0%**	25.5%	14.0%	18.7%	14.0%	16.7%	11.0%	12.4%	21.0%	22.1%
3	19.0%	22.6%	21.0%	23.5%	21.0%	15.5%	11.0%	10.8%	**29.0%**	27.6%	11.0%	14.0%
0	9.9%	8.6%	13.0%	14.1%	**16.0%**	15.9%	10.0%	10.8%	6.4%	7.6%	4.9%	4.0%
1	16.0%	15.4%	16.0%	**18.0%**	12.0%	9.0%	16.0%	16.0%	16.0%	16.8%	17.0%	15.1%
7	18.0%	16.0%	**30.0%**	28.8%	17.0%	15.1%	11.0%	9.5%	17.0%	15.4%	13.0%	15.2%
2	3.5%	12.3%	3.2%	8.7%	3.7%	-	3.6%	11.5%	4.6%	**23.6%**	2.6%	5.9%
	8.7%	11.4%	7.6%	9.4%	8.8%	7.0%	0.3%	0.2%	**14.0%**	13.7%	14.0%	12.1%
5	6.6%	5.1%	16.0%	**16.1%**	4.9%	3.1%	-	0.1%	6.5%	5.3%	1.6%	2.3%
	12.0%	18.0%	9.1%	13.3%	16.0%	17.8%	-	0.02%	18.0%	**22.0%**	3.9%	9.1%
)												
1	7.1%	9.7%	0.8%	2.7%	6.8%	5.9%	10.0%	8.6%	**12.0%**	11.3%	8.3%	9.1%
9	8.1%	10.3%	3.1%	4.0%	7.5%	6.2%	10.0%	8.6%	**13.0%**	12.1%	8.4%	9.2%
8	4.7%	5.6%	7.1%	7.8%	3.9%	3.2%	0.2%	0.4%	**8.7%**	7.0%	2.4%	2.3%

4	5.5%	6.4%	**11.0%**	10.5%	2.2%	3.2%	2.0%	2.8%	6.0%	7.1%	3.8%	5.7%
6	3.7%	1.1%	**13.0%**	12.0%	0.7%	0.4%	-	-	0.1%	0.01%	-	0.01%
9	5.0%	4.6%	**11.0%**	10.5%	3.7%	3.1%	-	-	5.5%	5.4%	1.7%	1.8%
2	3.9%	4.8%	**8.4%**	7.9%	1.4%	2.0%	0.1%	0.2%	5.0%	5.4%	2.6%	2.9%
4	10.0%	8.4%	**15.0%**	14.1%	13.0%	14.1%	10.0%	10.2%	6.6%	7.6%	4.9%	4.1%
28	2.1%	8.9%	0.4%	2.2%	4.2%	9.5%	**-**	**-**	6.4%	**11.7%**	0.9%	3.2%
5	6.1%	3.8%	0.1%	0.2%	7.6%	11.0%	**13.0%**	12.7%	1.0%	0.9%	11.0%	8.8%
6	2.8%	0.9%	**9.8%**	9.7%	1.1%	0.4%	**-**	**-**	0.1%	0.1%	-	0.01%
9	7.2%	4.6%	10.0%	8.1%	8.4%	**12.7%**	12.0%	12.5%	2.6%	2.5%	2.8%	1.8%
2	4.2%	2.7%	**9.2%**	8.2%	6.6%	2.7%	**-**	**-**	3.8%	2.9%	0.3%	0.6%
3	2.3%	3.9%	2.9%	3.1%	4.2%	2.3%	**-**	**-**	3.6%	**5.2%**	1.0%	1.7%
9	0.7%	1.9%	-	0.5%	0.9%	0.7%	0.1%	0.2%	2.2%	**2.7%**	0.7%	1.0%
6	4.1%	1.2%	**15.0%**	12.9%	0.6%	0.4%	**-**	**-**	-	0.02%	-	0.01%
0	3.2%	1.0%	**12.0%**	10.2%	-	0.4%	**-**	**-**	-	0.1%	-	0.01%
6	1.1%	2.1%	0.2%	0.4%	1.0%	1.0%	**-**	**-**	1.7%	2.5%	3.1%	**3.6%**
erozygous - Only the twins)												
3	0.8%	0.3%	2.4%	**2.5%**	0.6%	0.2%	-	0%	0.2%	0.1%	-	0.02%
7	0.2%	0.1%	0.7%	**0.8%**	-	0%	-	0%	-	0.003%	-	0%
26	-	0.1%	-	**0.3%**	-	0.02%	-	0%	-	0.2%	-	0.01%
73	0.1%	0.1%	**0.3%**	0.1%	**0.3%**	0.04%	-	0%	-	0.2%	-	0%
95	1.4%	2.5%	0.2%	0.6%	1.9%	1.6%	-	0.1%	**3.4%**	3.3%	2.2%	2.6%
55	0.7%	0.2%	2.4%	**2.9%**	0.3%	1.1%	-	0%	-	0%	-	0%
	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-

89	0.02%	0.01%	0.1%	**0.2%**	-	0.01%	-	0%	-	0%	-	0%
0	0.8%	0.2%	**2.9%**	2.4%	-	0.2%	-	0%	-	0%	-	0%
29	0.5%	0.2%	0.8%	1.1%	-	0%	**1.2%**	0.5%	-	0%	0.1%	0.1%
70	0.5%	0.2%	0.9%	1.1%	-	0%	**1.2%**	0.5%	-	0%	0.1%	0.1%
81	0.9%	0.2%	**3.3%**	2.7%	0.3%	0%	-	0%	-	0%	-	0%
85	0.4%	-	**1.4%**	-	-	-	-	-	0.1%	-	-	-
38	0.9%	0.8%	3.2%	**3.3%**	0.3%	0%	-	0%	-	0%	-	0%
34	0.3%	0.3%	1.1%	**1.2%**	-	0%	-	0%	-	0%	-	0%
17	0.4%	0.2%	0.9%	**1.3%**	-	0%	0.7%	0%	-	0.03%	0.1%	0.8%
5	0.04%	0.01%	**0.2%**	0.1%	-	0.01%	-	0%	-	0%	-	0%
9	1.7%	2.6%	-	0.7%	**4.9%**	3.2%	-	0%	2.9%	3.5%	2.2%	2.5%
33	0.8%	0.3%	3.0%	**3.5%**	0.3%	0.1%	-	0%	-	0.01%	-	0.01%
07	0.6%	1.2%	0.2%	0.3%	1.0%	0.5%	-	0.02%	1.2%	**1.5%**	0.9%	1.4%
	1.4%	0.4%	**5.0%**	3.8%	0.1%	0.2%	-	0%	0.1%	0.01%	-	0.01%
0	0.6%	1.2%	0.1%	0.3%	1.4%	0.8%	-	0%	1.4%	**1.7%**	0.5%	0.8%
66	0.4%	0.2%	-	0.03%	-	0.1%	-	0%	-	0.1%	**1.8%**	0.9%
19	0.1%	0.03%	**0.5%**	0.3%	-	0.03%	-	0%	-	0.001%	-	0%
61	0.04%	0.02%	**0.2%**	0.1%	-	0%	-	0%	-	0%	-	0%
	-	-	-	-	-	-	-	-	-	-	-	-
97	0.2%	0.1%	0.5%	**0.6%**	0.3%	0.02%	-	0%	-	0.003%	-	0.01%
27	0.02%	0.01%	**0.1%**	**0.1%**	-	0.01%	-	0%	-	0.002%	-	0.01%
94	-	**0.001%**	-	0%	-	0%	-	0%	-	**0.001%**	-	0%
1	1.0%	0.3%	**3.7%**	3.2%	0.1%	0.1%	-	0%	0.1%	0.01%	-	0.01%
33	-	0.02%	-	0%	-	**0.2%**	-	0%	-	0.02%	-	0%

	-	-	-	-	-	-	-	-	-	-	-	-
01	0.6%	0.1%	**2.3%**	1.6%	0.1%	0.1%	-	0.01%	-	0.002%	-	0.01%
2	1.1%	0.3%	**4.0%**	3.8%	0.1%	0.1%	-	0%	-	0.003%	-	0.1%
3	1.1%	0.4%	3.9%	**5.0%**	0.3%	0.1%	-	0%	-	0.01%	-	0%
	-	0.001%	-	0%	-	0%	-	0%	-	0%	-	**0.01%**
3	0.8%	0.3%	**2.7%**	2.0%	0.6%	0.4%	-	0%	0.2%	0.2%	-	0.02%
4	0.6%	0.2%	2.1%	**2.2%**	0.1%	0.1%	-	0%	-	0.005%	-	0.01%
	-	-	-	-	-	-	-	-	-	-	-	-
78	0.5%	0.2%	1.5%	**2.5%**	0.4%	0.1%	-	0%	-	0.003%	-	0%
2	0.6%	0.1%	**2.1%**	1.6%	0.1%	0.1%	-	0%	-	0.005%	-	0%
79	0.2%	0.1%	0.8%	**1.6%**	-	0.1%	-	0%	-	0%	-	0%
2	0.1%	0.04%	**0.4%**	**0.4%**	0.1%	0.04%	-	0%	-	0%	-	0.01%
1	0.4%	0.9%	0.1%	0.3%	0.7%	0.6%	-	0%	1.0%	**1.2%**	0.3%	0.6%
	-	-	-	-	-	-	-	-	-	-	-	-
6	0.9%	0.3%	**3.4%**	2.5%	0.1%	0.2%	-	0%	-	0.02%	-	0%
76	0.6%	0.2%	**2.0%**	1.8%	0.3%	0.1%	-	0.01%	-	0.01%	-	0.01%
71	0.7%	0.2%	**2.5%**	2.1%	0.3%	0.1%	-	0%	-	0.01%	-	0%
43	0.8%	0.2%	**2.9%**	2.4%	0.1%	0.1%	-	0%	-	0.01%	-	0%
0	0.9%	0.2%	**3.3%**	2.8%	0.1%	0.2%	-	0%	-	0.005%	-	0.01%
7	1.1%	0.3%	**3.9%**	**3.9%**	0.1%	0.1%	-	0%	-	0.003%	-	0%
35	1.2%	0.4%	**4.4%**	3.8%	-	0%	-	0.2%	-	0.02%	-	0.1%
12	1.0%	0.3%	**2.8%**	2.3%	0.3%	0.1%	-	0.2%	-	0%	1.2%	0.9%
3	1.1%	0.4%	**4.2%**	3.9%	0.3%	0.2%	-	0%	-	0.02%	-	0.01%
82	0.1%	0.04%	0.3%	**0.4%**	-	0%	-	0%	-	0%	-	0%

4	0.9%	0.3%	**3.0%**	2.8%	0.4%	0.1%	-	0%	-	0.02%	-	0.01%
27	-	0.01%	-	**0.1%**	-	0.02%	-	0%	-	0.005%	-	0.01%
27	0.6%	0.3%	**2.1%**	1.7%	0.3%	0.3%	-	0%	-	0.2%	-	0.01%
	-	0.005%	-	0%	-	**0.02%**	-	0%	-	0.01%	-	0%
25	-	0.01%	-	**0.1%**	-	0%	-	0%	-	0%	-	0%
18	1.3%	0.4%	**4.8%**	4.6%	0.1%	0.4%	-	0%	-	0%	-	0%
55	1.2%	0.2%	**4.4%**	3.8%	0.1%	0.5%	-	0%	-	0%	-	0%
15	0.2%	0.03%	0.5%	0.3%	0.3%	**1.2%**	-	0%	-	0%	-	0%
04	0.5%	0.1%	**2.0%**	1.5%	-	0%	-	0%	-	0%	-	0%
sociated genes												
2	-	2.7%	-	2.1%	-	2.1%	-	0.65%	-	**3.4%**	-	0.7%
4	1.4%	2.3%	0.2%	0.8%	2.5%	1.8%	-	-	**5.1%**	3.5%	0.1%	0.04%

Mixed American/Latin, EAS: East Asian, EUR: European*, SAS: South Asian, 1000G: 1000 Genomes Consortium, ExAC: Exome
respond only to Non-Finnish European, while 1000G's EUR data includes Finnish population.

APPENDIX 3 – RESEARCH ETHICS BOARD APPROVAL LETTERS

(IN PORTUGUESE)

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Análise comparativa de sequência de genoma completo de um pedigree contendo gêmeas monozigóticas concordantes para hanseníase

Pesquisador: Marcelo Távora Mira

Área Temática: Área 1. Genética Humana.
(Trata-se de pesquisa envolvendo genética humana não contemplada acima.);

Versão: 2

CAAE: 08152712.6.0000.0020

Instituição Proponente: Pontifícia Universidade Católica do Parana - PUCPR

DADOS DO PARECER

Número do Parecer: 169.382

Data da Relatoria: 05/12/2012

Apresentação do Projeto:

Nas últimas décadas, intensos esforços têm sido aplicados na identificação da exata natureza do componente genético controlando susceptibilidade
à hanseníase. Estudos utilizando diferentes estratégias de análise, incluindo scan genômicos de associação, resultaram na descrição de numerosas
variantes genéticas comuns associadas à hanseníase. Porém, estas variantes não explicam o forte efeito genético observado em estudos de
gêmeos e análises de segregação complexa. Uma possível explicação para esta "herança oculta" é a existência de variantes raras exercendo forte
impacto sobre fenótipos mendelianos que, combinadas, se manifestariam em doenças complexas comuns. Nossa proposta é de utilizar tecnologia
de sequenciamento de próxima geração para produzir a sequência completa do genoma de indivíduos selecionados de um pedigree contendo um
par de gêmeas monozigóticas que apresentam fenótipos concordantes de hanseníase. As duas crianças desenvolveram hanseníase antes de
atingirem os dois anos de idade; curiosamente, a doença manifestou-se com semelhança incomum em ambas as irmãs, sugerindo fortemente uma
característica mendeliana. O sequenciamento será realizado na plataforma de sequenciamento de

próxima geração ABI SOLiD 4. As sequências
obtidas serão analisadas seguindo protocolos desenvolvidos para detectar variantes raras, possivelmente causadoras do fenótipo de doença. Como
resultado, nós esperamos descrever, pela primeira vez, um caso de hanseníase sob controle mendeliano.

Objetivo da Pesquisa:
O objetivo principal deste projeto é aplicar, pela primeira vez, poderosas ferramentas de sequenciamento de genoma completo para avançar na
compreensão da natureza complexa do componente genético que controla a suscetibilidade humana à hanseníase

Avaliação dos Riscos e Benefícios:
Os pesquisadores destacaram que os riscos físicos decorrentes deste estudo são muito pequenos e limitados ao procedimento de coleta de sangue, em que pode haver desconforto temporário devido à picada da agulha. Não haverá constituição de biobanco derivado da pesquisa e todos os cuidados para a diminuição dos eventuais constrangimentos serão tomados

Comentários e Considerações sobre a Pesquisa:
Trata-se de pesquisa de extrema relevância, com grande clareza argumentativa, teórica e em seu delineamento.

Considerações sobre os Termos de apresentação obrigatória:
Todos os documentos foram apresentados satisfatoriamente

Recomendações:
Não há

Conclusões ou Pendências e Lista de Inadequações:
Não Há

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

Considerações Finais a critério do CEP:
Lembramos aos senhores pesquisadores que, no cumprimento da Resolução 196/96, o Comitê de Ética em Pesquisa (CEP) deverá receber relatórios anuais sobre o andamento do estudo, bem como a qualquer tempo e a critério do pesquisador nos casos de relevância, além do envio dos relatos de eventos adversos, para conhecimento deste Comitê. Salientamos ainda, a necessidade de relatório completo ao final do estudo.

Eventuais modificações ou emendas ao protocolo devem ser apresentadas ao CEPPUCPR de forma clara e sucinta, identificando a parte do protocolo a ser modificado e as suas justificativas. Se a pesquisa, ou parte dela for realizada em outras instituições, cabe ao pesquisador não iniciá-la antes de receber a autorização formal para a sua
realização. O documento que autoriza o início da pesquisa deve ser carimbado e assinado pelo responsável da instituição e deve ser mantido em poder do pesquisador responsável, podendo ser requerido por este CEP em qualquer tempo

CURITIBA, 12 de Dezembro de 2012

Assinador por:
NAIM AKEL FILHO
(Coordenador)

PARECER CONSUBSTANCIADO DO CEP

Elaborado pela Instituição Coparticipante

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Análise comparativa de sequência de genoma completo de um pedigree contendo gêmeas monozigóticas concordantes para hanseníase

Pesquisador: Marcelo Távora Mira

Área Temática: Área 1. Genética Humana.
(Trata-se de pesquisa envolvendo genética humana não contemplada acima.);

Versão: 2

CAAE: 08152712.6.0000.0020

Instituição Proponente: Pontifícia Universidade Católica do Parana - PUCPR

Patrocinador Principal: Financiamento Próprio
Conselho Nacional de Desenvolvimento Científico e Tecnologico ((CNPq))

DADOS DO PARECER

Número do Parecer: 205.054

Data da Relatoria: 25/02/2013

Apresentação do Projeto:

Nas últimas décadas, intensos esforços têm sido aplicados na identificação da exata natureza do componente genético controlando susceptibilidade à hanseníase. Estudos utilizando diferentes estratégias de análise, incluindo scan genômicos de associação, resultaram na descrição de numerosas variantes genéticas comuns associadas à hanseníase. Porém, estas variantes não explicam o forte efeito genético observado em estudos de

gêmeos e análises de segregação complexa. Uma possível explicação para esta "herança oculta" é a existência de variantes raras exercendo forte impacto sobre fenótipos mendelianos que, combinadas, se manifestariam em doenças complexas comuns.

A proposta dos pesquisadores é de utilizar tecnologia de sequenciamento de próxima geração para produzir a sequência completa do genoma de indivíduos selecionados de um pedigree contendo um par de gêmeas monozigóticas que apresentam fenótipos concordantes de hanseníase. As duas crianças desenvolveram hanseníase antes de atingirem os dois anos de idade; curiosamente, a doença manifestou-se com semelhança incomum em ambas as irmãs, sugerindo fortemente uma característica mendeliana.

O sequenciamento será realizado na plataforma de sequenciamento de próxima geração ABI SOLiD 4. As sequências obtidas serão analisadas seguindo protocolos desenvolvidos para detectar variantes raras, possivelmente causadoras do fenótipo de doença. Como resultado, espera-se descrever, pela primeira vez, um caso de hanseníase sob controle mendeliano.
Esta é uma proposta de aplicar tecnologia de sequenciamento de próxima geração para produzir informação de sequência do genoma completo de um caso extremo de hanseníase e seus pais e avó.

O objetivo é descrever o primeiro caso de hanseníase em humanos causado por uma ou mais variantes raras, seguindo um modelo de herança mendeliana. A proposta é produzir dados de sequência do genoma completo, ao invés do exoma completo, para ter acesso não só à informação das regiões transcritas e traduzidas do genoma, mas também dos introns, regiões promotoras, regiões UTRs, sítios de splicing alternativo e DNA não-codificado. Inicialmente, será realizada uma análise de exoma completo, visando identificar variações de base única (SNP) e pequenas inserções e deleções com um possível papel causal. Se nenhuma variante exômica rara for encontrada, a análise será expandida para o genoma completo, seguindo um racional hierárquico de busca em sequências com maior probabilidade de gerarem impacto funcional.

Desfecho Primário:
Identificação de uma mutação rara em homozigose que possa explicar a ocorrência de hanseníase extrema nas gêmeas. A descrição da natureza exata da variação - o gene envolvido e sua função biológica - irá contribuir para o esclarecimento do mecanismo que controla a susceptibilidade à hanseníase e, talvez, outras doenças infecciosas.

Objetivo da Pesquisa:
Hipótese:
A susceptibilidade humana à hanseníase per se e às suas formas clínicas é parcialmente controlado por determinantes genéticos. Apesar dos avanços recentes, a exata extensão do componente genético, o número de genes envolvidos, a localização e identidade desses genes, as variantes genéticas funcionais associadas com os fenótipos de hanseníase e o mecanismo biológico subjacente a essas associações ainda são desconhecidos.

Objetivo Primário:
O objetivo principal deste projeto é aplicar, pela primeira vez, poderosas ferramentas de sequenciamento de genoma completo para avançar na compreensão da natureza complexa do componente genético que controla a suscetibilidade humana à hanseníase.

Objetivo Secundário:
Os objetivos específicos são:
(i) Descrever e obter amostras de DNA de membros de um pedigree apresentando um caso raro de hanseníase de início extremamente precoce em gêmeas monozigóticas de dois anos de idade exibindo semelhança clínica notável;
(ii) produzir dados de sequenciamento em massa do genoma completo de uma das gêmeas e de outros membros informativos do pedigree;
(iii) pesquisar, utilizando ferramentas modernas de análise, todas as sequências contidas nos genomas estudados, visando identificar variantes raras que possam ser candidatas a causadoras de um caso de hanseníase explicável sob um modelo de herança genética.

Avaliação dos Riscos e Benefícios:
Riscos:
Os riscos físicos para a saúde de participação deste estudo são muito pequenos e limitados ao procedimento de coleta de sangue. Durante a coleta de sangue, a pessoa poderá sentir um desconforto temporário devido à picada da agulha. A coleta de sangue poderá resultar em uma pequena lesão que quase sempre cura-se sozinha. Em raros casos, pode ocorrer infecção localizada. Se o participante desenvolver infecção localizada
devido ao procedimento de coleta de sangue, o tratamento será providenciado pela equipe médica envolvida no estudo, sem custo para o paciente.

Não haverá constituição de biobanco derivado da pesquisa e todos os cuidados para a diminuição dos eventuais constrangimentos serão tomados. Será explicada a natureza exata dos experimentos realizados e a forma correta de se interpretar, sob a luz do conhecimento científico atual, o significado das variações genéticas encontradas.
Serão enfatizados também os potencias benefícios da identificação dos fatores genéticos que levam à predisposição à hanseníase para a prevenção, tratamento e controle da doença. Os participantes serão alertados sobre os possíveis riscos dos testes genéticos para variantes que predispões a doenças.

Benefícios:
Nos últimos anos, a tecnologia de sequenciamento de próxima geração emergiu como uma ferramenta genômica revolucionária, capaz de produzir dados de sequencias de DNA numa velocidade sem precedentes, permitindo assim potenciais avanços científicos previamente

inimagináveis. Nesse contexto, o grupo apresenta uma proposta inovadora de utilizar esta tecnologia no estudo de fatores de risco genéticos de susceptibilidade à hanseníase, uma doença infecciosa comum e ainda um problema de saúde pública no Brasil. Os pesquisadores acreditam que a análise do sequenciamento de genoma completo deste pedigree contribuirá significativamente no entendimento do background genético da susceptibilidade não só a esta doença, mas também a outras doenças infecciosas. Em relação aos benefícios diretos aos sujeitos da pesquisa, os pesquisadores propõem oferecer aos participantes acesso à informação produzida na forma de aconselhamento genético, oferecido por médico geneticista experiente, a todos que assim o desejarem, conforme descrito no Termo de Consentimento Livre e Esclarecido (TCLE). Assim, qualquer voluntário participante no estudo que solicitar ou aceitar receber aconselhamento genético, a qualquer momento, será informado sobre os aspectos gerais da descoberta de variantes genéticas que predispõem a doenças em geral e à hanseníase em particular. Será dada ênfase à natureza comum da variabilidade genética e seu papel como fator definidor dos aspectos positivos e negativos da individualidade.

Comentários e Considerações sobre a Pesquisa:
O projeto está suficientemente claro em seus propósitos, é de extrema relevância, e está devidamente fundamentado.

O projeto já foi aprovado pelo Comitê de Ética em Pesquisa da PUCPR e se enquadra no grupo II.

De acordo com o pesquisador as amostras serão destruídas assim que a pesquisa for finalizada. Se for necessário nova amostra para estudos adicionais, isso terá como condição uma nova avaliação e aprovação do projeto de pesquisa pelo Comitê de Ética pertinente, novo TCLE e nova amostra de sangue. Para este estudo não será autorizado o armazenamento do material em biobanco.

Considerações sobre os Termos de apresentação obrigatória:
Foram apresentados:

-Folha de rosto;
-Projeto de pesquisa;
-Cronograma atualizado;
-Orçamento da pesquisa;
-Termo de consentimento livre e esclarecido;
-Carta resposta dos pesquisadores em respostas às geradas pelo CEP PUCPR;
-Carta final de aprovação do Comitê de Ética em Pesquisa da PUCPR.

Recomendações:

Apresentar relatórios parciais (anuais) e relatório final do projeto de pesquisa é responsabilidade indelegável do pesquisador principal.

Qualquer modificação ou emenda ao projeto de pesquisa em pauta deve ser submetida à apreciação do CEP Fiocruz/IOC.

Conclusões ou Pendências e Lista de Inadequações:

Diante do exposto, o Comitê de Ética em Pesquisa do Instituto Oswaldo Cruz (CEP FIOCRUZ/IOC), de acordo com as atribuições definidas na Resolução CNS 196/96, manifesta-se pela aprovação do projeto de pesquisa proposto.

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

O sujeito de pesquisa ou seu representante, quando for o caso, deverá rubricar todas as folhas do Termo de Consentimento Livre e Esclarecido-TCLE apondo sua assinatura na última página do referido Termo.

O pesquisador responsável deverá da mesma forma, rubricar todas as folhas do Termo de Consentimento Livre e Esclarecido- TCLE apondo sua assinatura na última página do referido Termo.

RIO DE JANEIRO, 25 de Fevereiro de 2013

Assinador por:
José Henrique da Silva Pilotto
(Coordenador)

PARECER CONSUBSTANCIADO DO CEP

Elaborado pela Instituição Coparticipante

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Análise comparativa de sequência de genoma completo de um pedigree contendo gêmeas monozigóticas concordantes para hanseníase

Pesquisador: Marcelo Távora Mira

Área Temática: Genética Humana:
(Trata-se de pesquisa envolvendo Genética Humana que não necessita de análise ética por parte da CONEP;);

Versão: 2

CAAE: 08152712.6.3001.5214

Instituição Proponente: Pontifícia Universidade Católica do Parana - PUCPR

Patrocinador Principal: Financiamento Próprio
CONS NAC DE DESENVOLVIMENTO CIENTIFICO E TECNOLOGICO

DADOS DO PARECER

Número do Parecer: 657.779

Data da Relatoria: 26/11/2013

Apresentação do Projeto:

Segundo o pesquisador trata-se de"[...] é uma proposta de aplicar tecnologia de sequenciamento de próxima geração para produzir informação de sequência do genoma completo de um caso extremo de hanseníase e seus pais e avó. O objetivo é descrever o primeiro caso de hanseníase em humanos causado por uma ou mais variantes raras, seguindo um modelo de herança mendeliana [...]". Segundo ele, "estudos utilizando diferentes estratégias de análise, incluindo scan genômicos de associação, resultaram na descrição de numerosas

variantes genéticas comuns associadas à hanseníase. Porém, estas variantes não explicam o forte efeito genético observado em estudos de gêmeos e análises de segregação complexa. Uma possível explicação para esta "herança oculta" é a existência de variantes raras exercendo forte impacto sobre fenótipos mendelianos que, combinadas, se manifestariam em doenças complexas comuns. Nossa proposta é de utilizar tecnologia sequenciamento de próxima geração para produzir a sequencia completa do genoma de indivíduos selecionados de um pedigree contendo um par de gêmeas monozigóticas que apresentam fenótipos concordantes de hanseníase. As duas crianças desenvolveram hanseníase antes de atingirem os dois anos de idade; curiosamente, a

doença manifestou-se com semelhança incomum em ambas as irmãs, sugerindo fortemente uma característica mendeliana. O sequenciamento será realizado na plataforma de sequenciamento de próxima geração ABI SOLiD 4. As sequências obtidas serão analisadas seguindo protocolos desenvolvidos para detectar variantes raras, possivelmente causadoras do fenótipo de doença. Como resultado, nós esperamos descrever, pela primeira vez, um caso de hanseníase sob controle mendeliano.

Objetivo da Pesquisa:

Objetivo Primário:

O objetivo principal deste projeto é aplicar, pela primeira vez, poderosas ferramentas de sequenciamento de genoma completo para avançar na compreensão da natureza complexa do componente genético que controla a suscetibilidade humana à hanseníase.

Objetivo Secundário:

Os objetivos específicos são:

(i) Descrever e obter amostras de DNA de membros de um pedigree apresentando um caso raro de hanseníase de início extremamente precoce em gêmeas monozigóticas de dois anos de idade exibindo semelhança clínica notável;

(ii) produzir dados de sequenciamento em massa do genoma completo de uma das gêmeas e de outros membros informativos do pedigree;

(iii) pesquisar, utilizando ferramentas modernas de análise, todas as sequências contidas nos genomas estudados, visando identificar variantes raras que possam ser candidatas a causadoras de um caso de hanseníase explicável sob um modelo de herança genética.

Avaliação dos Riscos e Benefícios:

"Os riscos físicos para a saúde de participação deste estudo são muito pequenos e limitados ao procedimento de coleta de sangue. Durante a coleta de sangue, a pessoa poderá sentir um desconforto temporário devido à picada da agulha. A coleta de sangue poderá resultar em uma pequena lesão que quase sempre cura-se sozinha. Em raros casos, pode ocorrer infecção localizada. Se o participante desenvolver infecção localizada devido ao procedimento de coleta de sangue, o tratamento será providenciado pela equipe médica envolvida no estudo, sem custo para o paciente.

Além dos riscos à saúde, é importante considerar que há o risco de perda de privacidade, inerente a qualquer projeto de pesquisa de natureza semelhante. Para minimizar esse risco, vários procedimentos serão implementados. Especificamente, todos os dados serão mantidos, em todo momento, em arquivos de computadores com acesso restrito. Da mesma forma, todas as amostras de DNA serão armazenadas em freezer a -20°C com acesso restrito, sempre localizado no Laboratório Experimental Multiusuário da PUCPR e sob responsabilidade do professor investigador principal deste estudo. Acesso a qualquer informação ou amostra de DNA envolvida no estudo só será concedida mediante autorização do investigador principal. O compartilhamento de dados experimentais com os cientistas de diferentes instituições estará condicionado à garantia de que nenhuma identificação seja fornecida. Durante o recrutamento, todos os indivíduos serão informados sobre os riscos de perda de privacidade e as medidas adotadas para minimizar esses riscos."

Benefícios:
"Acreditamos que a análise do sequenciamento de genoma completo deste pedigree contribuirá significativamente no entendimento do background genético da susceptibilidade não só a esta doença, mas também a outras doenças infecciosas. Em relação aos benefícios diretos aos sujeitos da pesquisa, neste projeto propomos oferecer aos participantes acesso à informação produzida na forma de aconselhamento genético, oferecido por médico geneticista experiente, a todos que assim o desejarem, conforme descrito no Termo de Consentimento Livre e Esclarecido (TCLE)."

Comentários e Considerações sobre a Pesquisa:
A hanseníase é uma doença infectocontagiosa que acomete milhares de indivíduos no Brasil e no mundo.Os primeiros casos de hanseníase remontam à antiguidade identificada com o designativo de lepra. É uma doença que adquiriu um caráter estigmatizante. Os indivíduos acometidos dessa enfermidade, desde a antiguidade, sofrem preconceito na sociedade. Desta forma, a pesquisa proposta poderá contribuir significativamente no que diz respeito a origem e evolução da doença relativa a indivíduos gêmeos monozigóticos.

A pesquisa foi submetida à análise dos Comitês de Ética em Pesquisa das instituições parceiras - Pontifícia Universidade Católica do Paraná (PUC-PR) e Instituto Oswaldo Cruz (FIOCRUZ - RJ) nos quais obteve parecer favorável.

A pesquisa apresenta financiameto do Conselho Nacional de Desenvolvimento Científico e

Tecnológico - CNPQ através do edital Universal. A equipe de pesquisadores é oriunda de outras instituições como Universidade Federal do Piauí e de da Université de Paris Renée Descartes.

Com relação à metodologia de pesquisa os Comités citados elaboraram pareceres consubstanciados que avaliaram de maneira positiva o desenvolvimento da pesquisa.

Considerações sobre os Termos de apresentação obrigatória:
O projeto de pesquisa apresentou aos CEPs da PUC-RJ e da FIOCRUZ-RJ a documentação integral. Contudo, ao reapresentarem o projeto no CEP da UFPI nota-se a omissão Carta de apresentação encaminhada ao atual Coordenador com as devidas assinaturas dos pesquisadores participantes.

O TCLE apresenta-se bem escrito, numa linguagem acessível, relata os riscos físicos, contudo não inclui os riscos emocionais ou de constrangimento. Não se pode esquecer que esta enfermidade é vista de forma preconceituosa pela sociedade em geral.

Recomendações:
Incluir o CPF dos pesquisadores
Encaminhar a carta de apresentação ao atual Coordenador do CEP, com a assinatura dos pesquisadores
Adequar cronograma

Incluir os CPF dos pesquisadores

Conclusões ou Pendências e Lista de Inadequações:
O projeto se encontra apto para aprovação.

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

Considerações Finais a critério do CEP:

O CEP/UFPI vem, mui respeitosamente, se desculpar com o proponente do projeto, Prof. Marcelo Távora Mira, dada a demora na apreciação de seu projeto, o que se deveu à desativação do referido comitê, à época de submissão do projeto, bem como ao posterior acúmulo de demanda resultante, ao longo de todo ano de 2013.

TERESINA, 22 de Maio de 2014

Assinado por:
Alcione Corrêa Alves
(Coordenador)

APPENDIX 4 – INFORMED CONSENTS

(IN PORTUGUESE)

"Análise comparativa de sequência de exoma completo de um pedigree contendo gêmeas monozigóticas concordantes para hanseníase"

A Pontifícia Universidade Católica do Paraná (PUCPR) estará realizando um estudo científico sobre hanseníase, e a sua filha ou filho está sendo convidada/o a participar do estudo. Nele, pretende-se entender melhor porque algumas pessoas pegam hanseníase e outras não, mesmo às vezes estando muito próximas ou sendo parentes. Para isso, estudaremos a genética das pessoas que aceitarem participar, sequenciando seu genoma completo, ou seja: olharemos com muito detalhe o código que carregamos desde que nascemos dentro de nossas células, que herdamos de nossos pais, e que define muitas das nossas características individuais. Se bem sucedido o projeto irá aumentar o entendimento da doença, e poderá ajudar a melhorar, no futuro, a forma como os médicos a combatem.

Este estudo será coordenado pelo Dr. Marcelo Távora Mira, professor titular, pesquisador em atividade no Programa de Pós-Graduação em Ciências da Saúde, da Escola de Medicina, e coordenador do Laboratório Experimental Multiusuário (LEM) da PUCPR, e terá a maioria de seus experimentos realizados sob supervisão do Dr. Christian Macagnan Probst, pesquisador adjunto do Instituto Carlos Chagas (ICC), pelo farmacêutico Wilian Corrêa de Macedo, mestrando do Programa de Pós-Graduação em Genética da Universidade Federal do Paraná e pela farmacêutica Monica Elizabeth Dallmann Sauer, mestranda do Programa de Pós-Graduação em Ciências da Saúde da PUCPR.

1) Procedimentos

Se você concordar que sua filha/filho participe deste estudo, ela/ele será submetida à coleta de uma amostra de sangue. Serão coletados 5,0mL (um tubo pequeno) de sangue do antebraço dela/dele, de maneira idêntica àquela realizada nos laboratórios de análise clínicas para exame de sangue. O sangue será utilizado para extração de uma substância chamada "ácido desoxirribonucléico", ou simplesmente, DNA. O DNA extraído será armazenado e estudado por uma técnica chamada "sequenciamento", que nos permitirá descrever o DNA em grande detalhe. Em seguida, estudaremos essa descrição do DNA, tentando buscar nela uma explicação para a ocorrência ou não da hanseníase. Informações contidas no

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

prontuário médico da sua filha/filho também poderão ser lidas pelos pesquisadores e utilizadas no estudo.

2) Armazenamento

Se assinar este termo de consentimento, você não está autorizando a estocagem de amostra de DNA da sua filha/filho depois deste estudo terminar. Isto significa que a amostra será armazenada pelo tempo necessário para a finalização desta pesquisa e depois será destruída. Se for necessário utilizar nova amostra dela/dele para estudos adicionais de susceptibilidade à hanseníase isso terá como condição uma nova avaliação e aprovação do projeto de pesquisa pelo Comitê de Ética pertinente, novo termo de consentimento livre e esclarecido e nova amostra de sangue; ou seja: se for necessário, nós voltaremos a lhe pedir autorização para o uso de nova amostra dela para outros estudos futuros da hanseníase se o comitê de ética da PUCPR aprovar.

3) Local de estudo

Todas as atividades de campo relacionadas com acesso aos dados clínicos e coleta de sangue da sua filha/filho, assim como o acompanhamento, durante o tratamento da doença, serão realizadas pelo(a) seu(sua) médico(a). As análises da amostra de DNA, relacionadas com a pesquisa, serão realizadas parte nos laboratórios do ICC, e parte no LEM da PUCPR, ambos em Curitiba, Paraná. Estas parcerias entre a PUCPR e outras instituições de pesquisa irão aumentar a chance de sucesso do estudo.

4) Riscos/Desconfortos

Os riscos físicos para a saúde de participação deste estudo são muito pequenos e limitados ao procedimento de coleta de sangue. Durante a coleta de sangue, a sua filha/filho poderá sentir um desconforto temporário devido à picada da agulha. A coleta de sangue poderá resultar em uma pequena lesão que quase sempre cura-se sozinha. Em raros casos, pode ocorrer infecção localizada.

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

5) Tratamento e compensação de danos

Se a sua filha/filho desenvolver infecção localizada devido ao procedimento de coleta de sangue, o tratamento será providenciado pela equipe médica envolvida no estudo. O custo deste tratamento será totalmente coberto pelo projeto, se for o caso.

6) Alternativas

Você tem total liberdade para decidir ou não pela participação da sua filha/filho neste estudo. Caso você decida que ela não participe, ou desista da participação dela no estudo a qualquer momento, esta decisão não irá interferir de nenhuma forma em qualquer procedimento médico para diagnóstico ou tratamento de hanseníase ou qualquer outra doença que você e sua família possam necessitar no futuro.

7) Custos para os participantes

No caso de você decidir que a sua filha/filho participe do estudo, vocês não terão nenhum custo. Custos com testes laboratoriais e análises das amostras para propósito de pesquisa serão cobertos pelo estudo.

8) Benefícios

Esta pesquisa não irá resultar em uma mudança imediata na forma como a hanseníase é diagnosticada e tratada pelos médicos hoje. No entanto, esperamos que nossos resultados mudem muito, para melhor, o diagnóstico e o tratamento da hanseníase no futuro. É impossível prever quanto tempo vai levar para que essas mudanças aconteçam.

9) Reembolso/pagamento

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

Você ou sua filha/filho não serão pagos por participar deste estudo.

10) Exclusividade do uso do material genético

As amostras de DNA serão utilizadas apenas para pesquisa de suscetibilidade à hanseníase. Nenhuma outra doença será estudada. Todos os resultados obtidos no estudo, após análise do conjunto completo de dados, serão publicados em artigo científico, porém sem identificação das pessoas que concordaram em doar seu material para participar. Nós não sabemos nem controlamos a maneira como estes dados publicados serão usados por outros investigadores. Importante no caso deste estudo: como iremos fazer uma análise muito detalhada do DNA da sua filha/filho, é possível – e até mesmo provável – que sejam encontradas mutações associadas a outras doenças. Isso não significa que ela terá estas doenças; o significado destes achados ainda está em estudo pelos cientistas e médicos, portanto, não é possível colocar o que encontrarmos no DNA dela em um resultado de exame, por exemplo. Caso você queira saber detalhes do que foi encontrado, e o que isso significa, você terá uma consulta marcada com um médico especialista em genética, que lhe passará esta informação, da maneira correta e consagrada pela medicina, através de um procedimento conhecido como "aconselhamento genético". Importante lembrar que, apesar de que informações associadas a outras doenças possam ser descobertas, somente aquelas que possam ajudar a entender a hanseníase serão utilizadas no estudo.

11) Confidencialidade dos dados

A participação em projetos de pesquisa deste tipo pode resultar em perda de privacidade. Além disso, existe a possibilidade de que, no futuro, informações sobre uma pessoa ser ou não suscetível a ter uma doença sejam mal usadas para que outras pessoas, patrões ou empresas, as vejam de forma negativa. Entretanto, procedimentos serão adotados pelos responsáveis por este estudo no intuito de proteger a confidencialidade das informações que você fornecer e as informações produzidas pelo projeto. NENHUMA IDENTIFICAÇÃO PESSOAL SERÁ TORNADA PÚBLICA. As informações serão codificadas e mantidas em um local reservado o tempo todo. Somente os pesquisadores envolvidos neste estudo terão acesso às informações. Após o término deste estudo, as informações serão transcritas para arquivos de computador, mantidos em local restrito com acesso permitido apenas aos mesmos pesquisadores. Os dados deste estudo poderão ser discutidos com

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

pesquisadores de outras instituições, mas nenhuma identificação pessoal será fornecida.

12) Acesso à informação e dados para contato

Nós garantimos assistência durante toda a pesquisa, bem como seu livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que você queira saber antes, durante e depois da participação da sua filha/filho. Você receberá uma cópia deste consentimento para mantê-lo consigo. A qualquer momento, se tiver qualquer dúvida sobre a participação dela/dele neste estudo, você poderá utilizar os seguintes meios de contato com o pesquisador responsável e demais pesquisadores envolvidos:

Dr. Marcelo Távora Mira
Telefone: (41) 3271-2030
E-mail: m.mira@pucpr.br

Dr. Christian Macagnan Probst
Telefone: (41) 3316-3236
E-mail: cprobst@tecpar.br

Farm. Wilian Corrêa de Macedo
Telefone: (41) 9814-5623
E-mail: wilian.macedo@hotmail.com

Farm. Monica E. Dallmann Sauer
Telefone: (41) 9639-8327
E-mail: monica.sauer@pucpr.br

Em caso de reclamações ou qualquer tipo de denuncia sobre este estudo, você pode ligar para o Comitê de Ética em Pesquisa da PUCPR (CEP PUCPR), no telefone (41) 3271-2292, ou mandar um e-mail para nep@pucpr.br

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

A PARTICIPAÇÃO NA PESQUISA É VOLUNTÁRIA

Eu _____ **(nome do tutor),** _____ **(idade),** _____ **(RG),** tendo sido orientado quanto ao teor deste termo de consentimento e tendo compreendido a natureza e o objetivo do estudo, autorizo a participação de _____ **(nome da filha/filho),** _____ **(idade dela/dele)** _____ **(RG)** na referida pesquisa. **Porém, eu não autorizo a estocagem da amostra de DNA dela pelo tempo que esta durar**. Isto significa que a amostra será armazenada pelo tempo necessário para a finalização desta pesquisa e depois será destruída.

Eu entendo que tenho o direito de não concordar com a participação da minha filha/filho ou mesmo de retirá-la do estudo em qualquer momento que queira; sem riscos para o tratamento médico dela, ou de familiares. Estou ciente de que a sua privacidade será respeitada, ou seja que toda informação pessoal será mantida em sigilo.

_____

Assinatura do tutor/responsável

_____

Dr. Marcelo Távora Mira

Coordenador do projeto

_____

Data, hora e local

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

"Análise comparativa de sequência de exoma completo de um pedigree contendo gêmeas monozigóticas concordantes para hanseníase"

A Pontifícia Universidade Católica do Paraná (PUCPR) estará realizando um estudo científico sobre hanseníase, e você esta sendo convidado a participar do estudo. Nele, pretende-se entender melhor porque algumas pessoas pegam hanseníase e outras não, mesmo às vezes estando muito próximas ou sendo parentes. Para isso, estudaremos a genética das pessoas que aceitarem participar, sequenciando seu genoma completo, ou seja: olharemos com muito detalhe o código que carregamos desde que nascemos dentro de nossas células, que herdamos de nossos pais, e que define muitas das nossas características individuais. Se bem sucedido o projeto irá aumentar o entendimento da doença, e poderá ajudar a melhorar, no futuro, a forma como os médicos a combatem.

Este estudo será coordenado pelo Dr. Marcelo Távora Mira, professor titular, pesquisador em atividade no Programa de Pós-Graduação em Ciências da Saúde, da Escola de Medicina,e coordenador do Laboratório Experimental Multiusuário (LEM) da PUCPR, e terá a maioria de seus experimentos realizados sob supervisão do Dr. Christian Macagnan Probst, pesquisador adjunto do Instituto Carlos Chagas (ICC), pelo farmacêutico Wilian Corrêa de Macedo, mestrando do Programa de Pós-Graduação em Genética da Universidade Federal do Paraná e pela farmacêutica Monica Elizabeth Dallmann Sauer, mestranda do Programa de Pós-Graduação em Ciências da Saúde da PUCPR.

1) Procedimentos

Se você concordar em participar deste estudo, será submetido à coleta de uma amostra de sangue. Serão coletados 5,0mL (um tubo pequeno) de sangue de seu antebraço, de maneira idêntica àquela realizada nos laboratórios de análise clínicas para exame de sangue. O sangue será utilizado para extração de uma substância chamada "ácido desoxirribonucléico", ou simplesmente, DNA. O DNA extraído será armazenado e estudado por uma técnica chamada "sequenciamento", que nos permitirá descrever seu DNA em grande detalhe. Em seguida, estudaremos essa descrição do seu DNA, tentando buscar nela uma explicação para a ocorrência ou não da hanseníase. Informações contidas no seu prontuário médico também poderão ser lidas pelos pesquisadores e utilizadas no estudo.

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

2) Armazenamento

Se assinar este termo de consentimento, você não está autorizando a estocagem da sua amostra de DNA depois deste estudo terminar. Isto significa que a amostra será armazenada pelo tempo necessário para a finalização desta pesquisa e depois será destruída. Se for necessário utilizar nova amostra sua para estudos adicionais de susceptibilidade à hanseníase isso terá como condição uma nova avaliação e aprovação do projeto de pesquisa pelo Comitê de Ética pertinente, novo termo de consentimento livre e esclarecido e nova amostra de sangue; ou seja: se for necessário, nós voltaremos a lhe pedir autorização para o uso de nova amostra para outros estudos futuros da hanseníase se o comitê de ética da PUCPR aprovar.

3) Local de estudo

Todas as atividades de campo relacionadas com acesso a seus dados clínicos e coleta de sangue, assim como o seu acompanhamento, durante o tratamento da doença, serão realizadas pelo(a) seu(sua) médico(a). As análises da sua amostra de DNA, relacionadas com a pesquisa, serão realizadas parte nos laboratórios do ICC, e parte no LEM da PUCPR, ambos em Curitiba, Paraná. Estas parcerias entre a PUCPR e outras instituições de pesquisa irão aumentar a chance de sucesso do estudo.

4) Riscos/Desconfortos

Os riscos físicos para a saúde de participação deste estudo são muito pequenos e limitados ao procedimento de coleta de sangue. Durante a coleta de sangue, você poderá sentir um desconforto temporário devido à picada da agulha. A coleta de sangue poderá resultar em uma pequena lesão que quase sempre cura-se sozinha. Em raros casos, pode ocorrer infecção localizada.

5) Tratamento e compensação de danos

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

Se você desenvolver infecção localizada devido ao procedimento de coleta de sangue, o tratamento será providenciado pela equipe médica envolvida no estudo. O custo deste tratamento será totalmente coberto pelo projeto, se for o caso.

6) Alternativas

Você tem total liberdade para decidir ou não pela participação neste estudo. Caso você decida não participar, ou desista de participar do estudo a qualquer momento, esta decisão não irá interferir de nenhuma forma em qualquer procedimento médico para diagnóstico ou tratamento de hanseníase ou qualquer outra doença que você e sua família possam necessitar no futuro.

7) Custos para os participantes

No caso de você decidir participar do estudo, você não terá nenhum custo. Custos com testes laboratoriais e análises de suas amostras para propósito de pesquisa serão cobertos pelo estudo.

8) Benefícios

Esta pesquisa não irá resultar em uma mudança imediata na forma como a hanseníase é diagnosticada e tratada pelos médicos hoje. No entanto, esperamos que nossos resultados mudem muito, para melhor, o diagnóstico e o tratamento da hanseníase no futuro. É impossível prever quanto tempo vai levar para que essas mudanças aconteçam.

9) Reembolso/pagamento

Você não será pago por participar deste estudo.

10) Exclusividade do uso do material genético

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

As amostras de DNA serão utilizadas apenas para pesquisa de suscetibilidade à hanseníase. Nenhuma outra doença será estudada. Todos os resultados obtidos no estudo, após análise do conjunto completo de dados, serão publicados em artigo científico, porém sem identificação das pessoas que concordaram em doar seu material para participar. Nós não sabemos nem controlamos a maneira como estes dados publicados serão usados por outros investigadores. Importante no caso deste estudo: como iremos fazer uma análise muito detalhada do seu DNA, é possível – e até mesmo provável – que sejam encontradas mutações associadas outras doenças. Isso não significa que você terá estas doenças; o significado destes achados ainda está em estudo pelos cientistas e médicos, portanto, não é possível colocar o que encontrarmos em seu DNA em um resultado de exame, por exemplo. Caso você queira saber detalhes do que foi encontrado, e o que isso significa, você terá uma consulta marcada com um médico especialista em genética, que lhe passará esta informação, da maneira correta e consagrada pela medicina, através de um procedimento conhecido como "aconselhamento genético". Importante lembrar que, apesar de que informações associadas a outras doenças possam ser descobertas, somente aquelas que possam ajudar a entender a hanseníase serão utilizadas no estudo.

11) Confidencialidade dos dados

A participação em projetos de pesquisa deste tipo pode resultar em perda de privacidade. Além disso, existe a possibilidade de que, no futuro, informações sobre uma pessoa ser ou não suscetível a ter uma doença sejam mal usadas para que outras pessoas, patrões ou empresas, as vejam de forma negativa. Entretanto, procedimentos serão adotados pelos responsáveis por este estudo no intuito de proteger a confidencialidade das informações que você fornecer e as informações produzidas pelo projeto. NENHUMA IDENTIFICAÇÃO PESSOAL SERÁ TORNADA PÚBLICA. As informações serão codificadas e mantidas em um local reservado o tempo todo. Somente os pesquisadores envolvidos neste estudo terão acesso às informações. Após o término deste estudo, as informações serão transcritas para arquivos de computador, mantidos em local restrito com acesso permitido apenas aos mesmos pesquisadores. Os dados deste estudo poderão ser discutidos com pesquisadores de outras instituições, mas nenhuma identificação pessoal será fornecida.

12) Acesso à informação e dados para contato

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

Nós garantimos assistência durante toda a pesquisa, bem como seu livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que você queira saber antes, durante e depois da sua participação. Você receberá uma cópia deste consentimento para mantê-lo consigo. A qualquer momento, se tiver qualquer dúvida sobre a sua participação neste estudo, você poderá utilizar os seguintes meios de contato com o pesquisador responsável e demais pesquisadores envolvidos:

Dr. Marcelo Távora Mira
Telefone: (41)3271-2030
E-mail: m.mira@pucpr.br

Dr. Christian Macagnan Probst
Telefone: (41)3316-3236
E-mail: cprobst@tecpar.br

Farm. Wilian Corrêa de Macedo
Telefone: (41)9814-5623
E-mail: wilian.macedo@hotmail.com

Farm. Monica E. Dallmann Sauer
Telefone: (41)9639-8327
E-mail: monica.sauer@pucpr.br

Em caso de reclamações ou qualquer tipo de denuncia sobre este estudo, você pode ligar para o Comitê de Ética em Pesquisa da PUCPR (CEP PUCPR), no telefone (41)3271-2292, ou mandar um e-mail para nep@pucpr.br

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

A PARTICIPAÇÃO NA PESQUISA É VOLUNTÁRIA

Eu _____ **(nome),** _____ **(idade),** _____ **(RG),** tendo sido orientado quanto ao teor deste termo de consentimento e tendo compreendido a natureza e o objetivo do estudo, concordo em participar na referida pesquisa. **Porém, eu não autorizo a estocagem da minha amostra de DNA pelo tempo que durar**. Isto significa que a amostra será armazenada pelo tempo necessário para a finalização desta pesquisa e depois será destruída.

Eu entendo que tenho o direito de não concordar em participar ou mesmo de me retirar do estudo em qualquer momento que queira; sem riscos para meu tratamento médico, ou de meus familiares. Estou ciente de que a minha privacidade será respeitada, ou seja que toda informação pessoal será mantida em sigilo.

_____

Assinatura do voluntário

_____

Dr. Marcelo Távora Mira

Coordenador do projeto

_____

Data, hora e local

RÚBRICA DO SUJEITO DE PESQUISA

RÚBRICA DO PESQUISADOR

APPENDIX 5 – REVIEW ARTICLE AND LICENCE

Clinics in Dermatology (2015) **33**, 99–107

ELSEVIER

Clinics in Dermatology

Genetics of leprosy: Expected and unexpected developments and perspectives

Monica E.D. Sauer, BPharm, PhD candidate[a,1], Heloisa Salomão, BPharm, MS[a,1], Geovana B. Ramos, BPharm, PhD candidate[a], Helena R.S. D`Espindula, BPharm, MS candidate[a], Rafael S.A. Rodrigues, BPharm, PhD candidate[a], Wilian C. Macedo, BPharm, MS[a], Renata H.M. Sindeaux, PhD[b], Marcelo T. Mira, PhD[a,b,*]

[a]*Group for Advanced Molecular Investigation, Graduate Program in Health Sciences, School of Medicine, Pontifical Catholic University of Paraná, Curitiba, Paraná, Brazil*
[b]*School of Health and Biological Sciences, Pontifical Catholic University of Paraná, Curitiba, Paraná, Brazil*

Abstract A solid body of evidence produced over decades of intense research supports the hypothesis that leprosy phenotypes are largely dependent on the genetic characteristics of the host. The early evidence of a major gene effect controlling susceptibility to leprosy came from studies of familial aggregation, twins, and Complex Segregation Analysis. Later, linkage and association analysis, first applied to the investigation of candidate genes and chromosomal regions and more recently, to genome-wide scans, have revealed several leukocyte antigen complex and nonleukocyte antigen complex gene variants as risk factors for leprosy phenotypes such as disease per se, its clinical forms and leprosy reactions. In addition, powerful, hypothesis-free strategies such as Genome-Wide Association Studies have led to an exciting, unexpected development: Leprosy susceptibility genes seem to be shared with Crohn's and Parkinson's diseases. Today, a major challenge is to find the exact variants causing the biological effect underlying the genetic associations. New technologies, such as Next Generation Sequencing that allows, for the first time, the cost and time-effective sequencing of a complete human genome, hold the promise to reveal such variants. Strategies can be developed to study the functional effect of these variants in the context of infection, hopefully leading to the development of new targets for leprosy treatment and prevention.

Genetics of host susceptibility to infectious diseases

The burden of infectious diseases has been massive throughout history of mankind. Infections have been responsible for a strong selective pressure; yet, some of them are, still today, major public health problems. Recent advances, such as the development of vaccines and antibiotics, combined with a general increase of the education and socio-economical level of human populations led to an increase of life expectancy, but not to eradication of infectious diseases.[1] To understand this scenario, it is necessary to consider a very complex interplay between environmental (microbial and nonmicrobial) and human

* Corresponding author. Tel.: +55 (41) 3271-2030; fax: +55 (41) 3271-1657.
E-mail address: m.mira@pucpr.br (M.T. Mira).
[1] These authors share first authorship.

http://dx.doi.org/10.1016/j.clindermatol.2014.10.001
0738-081X/

(genetic and nongenetic) factors that determines immunity to infection or its clinical outcome.[1]

A classic feature of human infections is that only a proportion of exposed individuals develop clinical disease.[2] Accumulating evidence suggests that host genetic factors play a particularly important role in controlling susceptibility to these diseases.[3] Some of the most compelling evidence that human genetics does indeed determine the occurrence of infection comes from primary immunodeficiencies (PIDs). The PIDs are typically monogenic (Mendelian) disorders that impair host defense mechanisms and result in predisposition to multiple infectious diseases. The PIDs are responsible for more than 200 known clinical syndromes, at least 100 of which presenting a well-defined molecular genetic basis. Examples of PIDs include mutations in genes encoding proteins of the IL-12/23-IFN-γ pathway associated with the Mendelian Susceptibility to Mycobacterial Diseases syndrome; complement defects associated with *Neisseria sp* invasive disease; X-linked lymphoproliferative disease associated with Epstein-Barr virus infection; mutations in genes *EVER1* or *EVER2* associated with Epidermodysplasia *verruciformis*; and apolipoprotein L-1 deficiency, associated with trypanosomiasis.[1,4–6]

The profound influence of the genetic make-up of the host over resistance to infection has been investigated in several models. Studies in mice, based on reverse (gene-targeted knock-out and knock-in mutations) and forward genetics (natural mutation and random mutagenesis), have provided important insights into the mechanisms controlling infection and immunity in human, natural conditions.[7] In human genetics, epidemiologic studies of adopted individuals showed that predisposition to infection were largely inherited, paradoxically, even more than diseases associated with less obvious environmental risk factors, such as cancer.[1,8] Studies comparing concordance rates between monozygotic and dizygotic twins have provided powerful evidence for the existence of a host genetic background controlling susceptibility to different infectious diseases.[1] Finally, several genes have been associated with diseases such as AIDS (*HLA*, *MICA*, *PSORS1 C3*, *KIR*, and *CCR5*); hepatitis B (genes *HLA*); tuberculosis (*MBL, VDR, NRAMP1,* genes *HLA*); malaria (*HBB, SCO1*, *DDC*); and meningococcal disease (*CFH*).[5,9,10]

In this context, leprosy presents as a very good model for the study of genetic predisposition to infection: *Mycobacterium leprae*, the etiologic agent, is known for its limited diversity between strains of different locations[11,12]; this near clonal characteristic, together with the observation of a wide range of leprosy clinical phenotypes, strongly suggests that most of the disease variability, including susceptibility to leprosy per se, is dependent upon the genetic background of the host.[13]

Genetics of leprosy

Today, it is widely accepted that the exposure to *M leprae* is necessary but not sufficient to trigger the outcome of the disease, and different sets of genes modify host susceptibility to leprosy in three different stages, namely: (i) the control of infection per se, that is, the disease regardless of its clinical form manifestation; (ii) after the infection is established, the definition of different clinical forms of the disease; and (iii) the risk of developing leprosy reactions (Figure 1).

Observational studies indicate the presence of a familial component to susceptibility to the leprosy,[14] as well as increased concordance rates of disease per se and its clinical forms among monozygotic compared with dizygotic twins.[15,16] In addition, several Complex Segregation Analyses have been performed for leprosy in different populations,[17–19] aiming to identify the best-fit model of inheritance of the phenotype, given a collection of pedigrees. All of these studies supported a polygenic model of inheritance that includes a major gene effect.

Taken together, these results indicate the existence of a strong genetic component controlling susceptibility to leprosy; however, these observational designs do not provide any information about the exact nature of the genetic factors involved, that is, the identity and number of genes, as well as the genetic variants of these genes, causative of the leprosy phenotypes. For that, molecular studies are necessary, and a vast number of these studies have been conducted over the past decades. As a result, several candidate chromosomal regions and genes have been described, such as the MHC/HLA-liked genes of class I and II, *TNFA, LTA, MICA, MICB,* as well as non-HLA genes, such as *CCDC122*, *IFNG, IL10, IL12 B, IL23 R, KIR, LACC1 (*formerly *C13 orf31), LTA4 H, LRRK2, MRC1, NOD2, PARK2/PACRG, RIPK2, SLC11A1 (*formerly *NRAMP1), TAP, TLR, TNFSF15,* and *VDR*. Among these, a few candidates have been consistently replicated by independent studies and/or successfully investigated by functional studies. A brief summary of selected genes is presented in Table 1 and expanded next.

Major histocompatibility complex genes

The major histocompatibility complex (MHC), in humans known as the leukocyte antigen complex (HLA), is a cluster of highly polymorphic genes contained within a 3.6 megabase (Mb) interval located on chromosome 6p21. Most of these genes encode for proteins that are essential players in the processing and binding of antigenic peptides during the immune response. The HLA region is organized in 3 classes: HLA class l contains subclasses HLA-A, −B and -C, which present antigenic peptides of intra-cellular origin to CD8+ T cells; HLA class II includes subclasses HLA-DR, −DQ, −DM, and -DP, that primarily bind peptides of extra-cellular origin and present them to CD4+T cells, and HLA class III contains genes coding for cytokines, such as tumor necrosis factor alpha (TNFA) and lymphotoxin alpha (LTA), for enzymes involved in steroid synthesis, for heat-shock proteins and for other intermediates of the immune response mechanisms.[20]

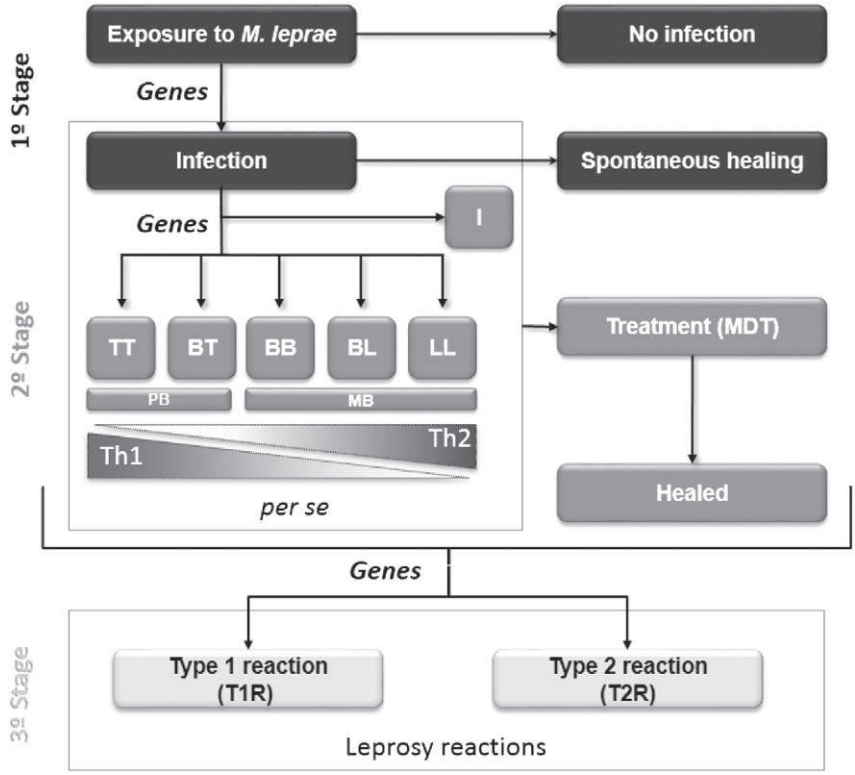


Fig. 1 A schematic stage model of genetic susceptibility to leprosy. The exposure to the mycobacteria does not always result in infection. In an initial stage, a first group of genes confers susceptibility to infection (leprosy per se). Among the individuals who develop the disease, a second group of genes determines the type of host immune response and subsequent leprosy subtype. Finally, a third group of genes confers the risk of developing leprosy reactions. TT: Tuberculoid-tuberculoid, BT, Borderline-tuberculoid; BB, Borderline-borderline; BL, Borderline-lepromatous; LL, Lepromatous-lepromatous; I, Indeterminate; PB, Paucibacillary; MB, Multibacillary; Th1, T-helper 1; Th2, T-helper 2; MDT, Multidrug therapy.

The crucial role of the HLA region in immune response regulation makes it the most exhaustively studied genomic candidate region in infection.[21] It seems reasonable to assume a particularly important role for HLA genes in leprosy, given that clinical manifestation of the disease depends on the type of immune response shown by the host. The exchanges between Th1 and Th2 types of immune response may be partially controlled by a mechanism of antigen presentation involving HLA molecules.[22] A large number of linkage and association studies reported the involvement of HLA alleles and haplotypes as important genetic factors controlling susceptibility to leprosy. In particular, HLA-DR alleles have been consistently associated with leprosy[23]: Several studies reported an association of the HLA-DR2 alleles HLA-DRB1*04, DRB1*10 DRB1*12, DRB1*15, and DRB1*16 with susceptibility or resistance to leprosy in populations around the world.[24–30] In addition, HLA-DR3 alleles were also found to be associated with leprosy susceptibility in two different populations.[31,32]

HLA class I has been also intensively studied in leprosy, and HLA-A*2, A*11, B*40, and C*7 are some examples of alleles detected more often among leprosy cases compared with unaffected controls.[33] A recent study performed a high-resolution association scan of a 1.9 Mb region in the HLA complex in a Vietnamese population, followed by step-wise replication in an independent sample from Vietnam and a sample from North India. The result was the identification of eight intergenic HLA class I region single nucleotide polymorphisms (SNPs) as novel genetic risk factors for leprosy per se, particularly implicating the HLA-C in leprosy susceptibility.[34]

There is cumulative evidence that class III genes *TNFA* and *LTA* are involved in the immune response against leprosy.[35] *TNFA* encodes TNF-α, a proinflammatory and immunostimulatory cytokine that belongs to the TNF superfamily. This cytokine is involved in the regulation of a wide spectrum of biological processes, including the modulation of innate and adaptive immune responses. It is mainly secreted by macrophages, and functions as an important modulator of antigen presentation, through class II molecules, and cytokine production necessary for effective leukocyte response. In leprosy, a large body of functional experimental data indicates that TNF-α plays a central role by mediating the protective response to *M leprae* invasion. Genetic studies have consistently indicated that *TNFA* variants can influence leprosy phenotypes.[36,37] Recently, a large association study involving 4 population samples and more than 2500 individuals, followed by a meta-analysis, confirmed association between a promoter variant of *TNFA* and leprosy, interestingly, the effect seems to be restricted to the Brazilian samples.[37]

Table 1 Selected leprosy associated genes replicated in at least two distinct population samples

Official symbol	Official full name	Function
HLA-DRB1	*Major histocompatibility complex, class II, DRB1*	Heterodimer anchored in the membrane that present antigenic peptides of extra-cellular origin to CD8+ T cells
TNFA	*Tumor necrosis factor alpha*	Proinflammatory cytokine
LTA	*Lymphotoxin alpha*	Proinflammatory cytokine
IL10	*Interleukin 10*	Anti-inflammatory cytokine
PARK2	*Parkinson protein 2, E3 ubiquitin protein ligase*	E3 ubiquitin ligase
NOD2	*Nucleotide-binding oligomerization domain containing 2*	Cytoplasmic receptor that plays a role in the immune response to intracellular bacterial lipopolysaccharides
RIPK2	*Receptor-interacting serine-threonine kinase 2*	Member of the receptor-interacting protein (RIP) family of serine/threonine protein kinases
LACC1	*Laccase (multicopper oxidoreductase) domain containing 1*	Unknown
CCDC122	*Coiled-coil domain containing 122*	Unknown

LTA is also a member of the TNF superfamily, but compared with TNF-α, much less is known about its function.[35] This cytokine is produced by lymphocytes and forms heterotrimers with lymphotoxin-beta, which anchor lymphotoxin-alpha to the cell surface. LTA signaling has proven important in controlling infection by intracellular pathogens, including *Mycobacterium tuberculosis* and *M leprae* among others.[38,39] Interestingly, it seems that, in leprosy, the *LTA* genetic effect is more pronounced in patients with early onset of disease.[40] A recent study using knockout mice showed that the combination of TNF and LTA are necessary to the formation and maintenance of granulomas in response to *M leprae*: LTA seems to regulate granuloma formation whereas TNF-α is responsible for its integrity.[35,39]

Finally, variants of additional HLA-linked genes, such MICA (MHC class I polypeptide-related sequence A), MICB (MHC class I polypeptide-related sequence B), and TAP (Transporter 1, ATP-binding cassette, sub-family B [MDR/TAP]), have also been described in association with leprosy phenotypes in different populations.[22]

Of note, the interpretation of genetic association studies involving the HLA complex requires caution, due to the close proximity and very high gene density typical of this locus: Once positive association is detected between a leprosy phenotype and an MHC/HLA marker, the challenge is to determine if the causative gene is the one being investigated of another in close proximity to the marker used, a phenomenon called linkage disequilibrium.[22]

Non-HLA genes

Interleukin 10

The *Interleukin 10* (*IL10*) gene is located at the 1q31-q32 chromosomal region and encodes for the anti-inflammatory cytokine IL-10, secreted by cells of the monocyte/macrophage lineage and T-cell subsets such as Type 1 Tr, regulatory T, and T-helper 17.[35,41] The IL-10 exerts its anti-inflammatory actions by blocking the production of proinflammatory cytokines by macrophages and their ability to serve as antigen-presenting or costimulatory cells.[41] More specifically, this cytokine inhibits the production of IL-1, IL-6, and TNF-α in LPS- and IFN-γ-activated macrophages.[36,41–43]

High levels of IL-10 are observed in multibacillary/lepromatous leprosy patients compared with paucibacillary/tuberculoid patients and a low TNF-α/IL-10 ratio is correlated to disease progression.[35,42] Genetic epidemiology studies have been consistently reporting association between leprosy and SNPs located at the *IL10* gene.[44–49] The exact reason why *IL10* polymorphisms are associated with leprosy is yet to be cleared. As this cytokine suppress the production of inflammatory mediator and boosts the development of Th2 immunity,[50] it is plausible that these polymorphisms somehow change *IL10* expression, directing the patient towards one of the poles of the leprosy clinical spectrum.

PARK2/PACRG (Parkinson protein 2, E3 ubiquitin protein ligase/Parkin co-regulated gene)

A genome-wide linkage analysis conducted in a Vietnamese population mapped a leprosy susceptibility *locus* to chromosome 6q25-27, an effect distributed along the entire leprosy clinical spectrum.[51] In a follow-up study, the same group performed a systematic association scan of the candidate region and found 17 SNPs associated with leprosy susceptibility,[52] 15 of them located in and around the promoter region shared by two genes: *PARK2* and *PACRG*. These results were validated in a separate set of unrelated individuals from Brazil. Later, an independent case–control study found significant association between leprosy and *PARK2/PACRG* SNPs in an Indian population sample; however, the signal did not resist correction for multiple testing.[53] A study conducted in a geographically isolated Croatian community with a well-documented history

of leprosy showed that the protective alleles of two *PARK2* SNPs associated with the disease in Vietnam and Brazil were enriched in this population,[54] suggesting positive selection. More recently, a study involving a Vietnamese and an Indian population sample confirmed the *PARK2/PACRG* effect and revealed that age at diagnosis and differences in linkage disequilibrium patterns across different ethnicities are important for the correct interpretation of these association results.[55] Curiously, association has been also reported between the *PARK2/PACRG* leprosy polymorphisms and typhoid and paratyphoid fever in an Indonesian population.[56] The finding raised the hypothesis that the *PARK2/PACRG* genetic effect would not be specific to infection with *M leprae*, but related to host responses against intracellular parasites.

The *PARK2* gene encodes Parkin, an E3 ubiquitin ligase involved with the ubiquitin-proteasome complex that mediates the targeting of protein substrates for proteasomal degradation.[13] Replicated association between leprosy and *PARK2/PACRG* variants revealed a new ubiquitin-dependent pathway of immunity to infection with *M leprae*, an idea supported by a functional study which demonstrated that proteasome function is important for *M leprae*-induced apoptosis.[57]

The Genome Wide Association Studies genes

Genome Wide Association Studies (GWAS) are a powerful study design based on extensive coverage of the entire genome with hundreds of thousands of makers, genotyped in one single experiment, that capture the vast majority of common variants in the genome sequence.[58,59] The genotyping data are then used in association testing that, if performed in samples large enough to achieve an adequate statistical power, allows for the identification of very small genetic effects, without the need of a previous hypothesis. The first GWAS on leprosy included 491,883 SNPs genotyped in 706 cases and 1225 controls from Eastern China. A total of 93 SNPs showed association with leprosy at the GWAS significance level; these SNPs were then tested in three independent replication sets totaling 3254 patients and 5955 controls from Eastern and Southern China. As a result, 15 SNPs distributed in six genes, *CCDC122*, *LACC1, NOD2, TNFSF15*, *RIPK2,* and the *HLA-DR–DQ*, were consistently associated with leprosy across all samples. In addition, a trend toward association was detected between leprosy and one SNP of *LRRK2*.[60] In 2011, the same group published an expanded GWAS by combining their first data set with additional control subjects–two additional genes were identified associated with leprosy: *IL23R* and *RAB32*.[61] Later, another study identified a relative increase in *IL23R* gene copy number significantly associated with paucibacillary leprosy.[62]

Given the nature of the GWAS, association studies involving a tremendous number of tests performed on a single experiment, therefore under strong inflation of type I error (false positive), these results, although exciting, must be validated by replication and/or by functional independent studies.[63]

Association between leprosy and *HLA-DR-DQ* was replicated in an Indian population[64] and the *LACC1* and *CCDC122* signals were replicated in an Indian and African population.[65] More recently, a family-based replication study conducted in 474 Vietnamese leprosy families re-tested all 16 SNPs associated with leprosy in the Chinese original GWAS; six of them, located at *CCDC12*, *LACC1, NOD2, RIPK2,* and the *HLA-DR-DQ* genes were replicated.[66] Association between *NOD2* and leprosy has also been replicated in Nepal.[67]

Several of the proteins encoded by these genes are involved in microbial sensing and in the early immune and inflammatory responses.[63] *NOD2* is located on chromosome 16q12 and encodes an intracellular receptor that recognizes a muramyl dipeptide component of the bacterial wall. After the interaction with activated NOD2 molecules, RIPK2 undergoes poly-ubiquitination mediated by an E3 ubiquitin ligase and promotes the activation of the TGFβ-activated kinase 1 (TAK1) complex. The activated TAK1 complex, again via poly-ubiquitination of a mediator, leads to degradation of transcriptional regulator nuclear factor κB (NF-κB) repressor IκB, releasing NF-κB to promote the transcription of pro-inflammatory genes, one of them, *TNFSF15*.[63,68–70] A functional study reinforced the importance of the NOD2 cascade in leprosy, by demonstrating that after NOD2 from monocytes interact with *M leprae's* muramyl dipeptide, a distinct interleukin-32-dependent induction of innate immune responses takes place, leading to the differentiation of monocytes into dendritic cells.[71] These antigen-presenting cells are competent to define the adaptive immune response in leprosy.[72–74]

Genetics of leprosy reactions

LRs are sudden and intense inflammatory processes that affect individuals at all stages of the disease, from diagnosis, during treatment and even in the post-cure. The pathophysiologic mechanisms involved are still widely unknown. Leprosy reactions are classified as type 1 (T1R or reversal reaction), which commonly affects patients at the tuberculoid side of the clinical spectrum; or type 2 (T2R, or *Erythema Nodosum Leprosum*), which affects mainly patients from the lepromatous pole of the disease.[75–79] Only recently, human genetic epidemiology tools have been applied to the investigation of the molecular mechanisms controlling susceptibility to this extreme leprosy phenotype, as recently reviewed by Fava and cols.[77,80]

The first evidence of association between LR and genetic polymorphisms came from studies involving Toll-like receptor (TLR) genes. An investigation involving a Nepalese population sample revealed polymorphisms on *TLR1* and *TLR2* associated with higher risk for T1R.[81,82] A functional polymorphism of *TLR1*, which causes a substitution of asparagine to serine (N248S), was found associated with susceptibility to leprosy reactions in a Bangladeshi

population sample.[83] Interestingly, a recent study detected association between the same N248S polymorphism and susceptibility to leprosy in a Brazilian population sample.[84] The TLRs are transmembrane proteins that play a critical role in the inflammatory response to microbial pathogens.[64,85] *TLR1* is located on chromosome 9q33.1 and its protein forms a heterodimer with TLR2 or TLR6 and mediates the recognition of several mycobacterial motifs: The heterodimer TLR1/TLR2 is involved in the recognition of *M leprae*,[86] and TLR1/TLR6 seems to be related to *M leprae* persistence in Schwann cells.[87]

A prospective study of a Brazilian population sample resulted in strong evidence implicating variants of the *IL6* gene with susceptibility to T2R. Upon diagnosis, leprosy patients were monitored for at least 1 year for the occurrence of LR. Patients who developed T1R or T2R within the follow-up period were included in the group of cases, and leprosy patients who did not develop reactions were used to compose the control group. Cases of T1R and T2R were matched with controls by clinical form of leprosy and compared for the allele frequencies of markers physically covering the entire *IL6* gene. No association was observed between the *IL6* markers and T1R. Two independent signals of association with T2R were detected; one of them was captured by SNP rs1800795, a variant with known impact over *IL6* expression. These results support an important role of *IL6* in the development of T2R.[77]

Additional studies on genetics of LR have implicated variants of *NOD2* and *VDR* as risk factors for the occurrence of T2R and T1R, respectively.[67,88] Finally, a study in a Brazilian sample demonstrated an association between a SNP of *SCL11A1* with leprosy reactions.[89] These findings are yet to be replicated.

Leprosy, Crohn's, and Parkinson's diseases: a common genetic background?

In the past recent years, interesting findings concerning the genetic control of complex diseases have been observed: Some disorders, apparently unrelated, share genetic risk factors. In this context, the identification of leprosy susceptibility genes revealed an unexpected overlap with inflammatory bowel conditions and Parkinson's disease.[9]

Inflammatory bowel disease (IBD) is characterized by a chronic, relapsing intestinal inflammation. The two major types of IBD are Crohn's disease (CD) and ulcerative colitis (UC). The identification of *Mycobacterium avium* subspecies *paratuberculosis* RNA in mucosal samples of CD and UC patients suggested that the development of the disease, at least in some individuals, might be triggered by mycobacterial infection.[90,91] This hypothesis have been supported by recent genetic studies revealing CD susceptibility genes that encode proteins responsible for recognition of bacterial structures and/or are present in immunologic pathways.[9,92] Some of the these CD susceptibility loci are shared with leprosy,[63] as clearly exposed by the Chinese leprosy GWAS: Five of the newly described leprosy susceptibility genes have been previously associated with CD: *TNFSF15, NOD2, LACC1, LRRK2,* and *IL23R*.[60,61] Motivated by this unexpected finding, the same group later performed a systematic, comprehensive association study testing all previously described IBD loci as leprosy susceptibility candidates in a large Chinese leprosy sample, again, two IBD loci were associated with leprosy: *IL18RAP/IL18R1,* and *IL12B*.[93] These outstanding results corroborates the hypothesis that IBD, mainly CD, and leprosy share genetic risk factors and sum additional evidence supporting the role of an infectious agent participating in the initial events leading to CD manifestation.

One additional intriguing finding of the Chinese study is that the in silico analysis that places *NOD2*, *RIPK2,* and *TNFSF15* in the same pathway also included *PARK2* and *LRRK2*, genes encoding proteins that directly interact.[60] Strikingly, *PARK2* and *LRRK2* are well known Parkinson's disease (PD) susceptibility genes. Of particular interest, *LRRK*2, shown to be expressed in macrophage and monocytes,[94] harbors variants that have been associated with CD[95] and Parkinson's disease, as well as leprosy.

Based on this results, one can speculate that because NOD2/RIPK2 initiates a signaling process that involves an ubiquination process through TRAF6 (TNF receptor-associated factor 6), an E3 ubiquitin ligase, it is possible that parkin, also an E3, plays a role in this process; in addition, LRRK2 is thought to regulate the ligase activity of PARK2[96]; therefore, it may also take part in the signaling control. These observations suggest the existence of a partially shared genetic control of susceptibility to an infectious disease, an inflammatory disease and a neurodegenerative disorder. The complete elucidation of the cross talk between those susceptibility genes is a difficult, yet tremendously exciting task.

Future perspectives

Classic genetic studies on susceptibility to leprosy have been focusing on the identification of common variants that could explain predisposition to disease and, as a result, several common variants were described to be associated with leprosy phenotypes. The assumption made is that a set of these variants in one or several genes of a biochemical pathway would act together to contribute to a clinical outcome. These findings cannot explain the totality of the large genetic effect observed in descriptive genetic epidemiology studies. Interesting, this scenario remain true for a number of complex traits.[97]

With the popularization of the GWAS, it has become increasingly clear that a large part of the genetic effect controlling disease susceptibility was missing[98]: With rare

exceptions, more than 90-95% of the heritable component of a disease has been left unexplained after extensive GWAS on several complex diseases, giving rise to the term "missing heritability".[98,99] The idea behind these "missing effect" is that common genetic variability is unlikely to explain the entire genetic predisposition to disease.[98,100,101]

As a result of intense debate, a new scenario of not one, but two major hypothesis has risen, both aiming to offer a better understanding on how different classes of genetic variations can account for a specific outcome: The "Common Disease-Common Variant (CDCV)" and the "Common Disease–Multiple Rare Variant (CDMRV)" hypothesis. The first one argues that common variants with small effect are responsible for the genetic susceptibility to common diseases; the CDMRV hypothesis defends the idea that multiple rare alleles of large effect, explains the genetic susceptibility to common diseases.[100,101]

In fact, the idea that rare variants are behind of human susceptibility to common diseases is not new. A remarkable example of the impact of a rare variant over disease phenotypes has been provided by Altare et al. in 1998[102]: By applying molecular biology tools a genetic analysis on the study of a single young girl presenting generalized, atypical *M. bovis*-BCG infection. The authors found a homozygous, 4.4 kb-long deletion in the gene *IL12B* that impaired IL12-dependent, IFN-γ mediated response against a non-virulent mycobacteria. The finding, that clearly implicates the IL-12/IFN-γ axis as critical for the control of the immune response against mycobacterial infection, represents strong support to the idea that a rare structural, homozygous variant can underlie the mechanism controlling host susceptibility to infection. This study and others provided rising evidence that rare variations are important pieces of the puzzle of human phenotypic variation.[103] Identifying these rare variants without previous indication of their possible location used to be a daunting task, given (i) the need to study rare, extreme cases of disease; and (ii) the limited throughput and the high costs of classic, Sanger-based methods for genome sequencing.

The CDMRV hypothesis gained momentum with the very recent advent of next generation genetic analysis platforms capable of sequencing massive segments of the human genome, whole exomes or even genomes, in short time frames and for a reasonable cost. By reducing the time and cost limitations to a minimum challenge, these platforms of next-generation sequencing (NGS) technologies, also known as massively parallel DNA sequencing, are ideal tools to be used on the search of such rare variations.[100,104]

Conclusions

The NGS technology has the potential to revolutionize our understanding on how genes or genomic regions are involved in the pathogenesis of human diseases.[105] The use of NGS can be directed to the identification of causative disease mutation by resequencing the whole genome (or exome) of a small number of affected individuals, typically displaying extreme phenotypes of the disease. The approach has been successfully applied to determine the genetic basis of rare disorders, much of them Mendelian, through the study of a small number of affected individuals. In this scenario, an interesting question would be whether the same strategy could be applied to the identification of rare mutations possibly contributing to the risk of occurrence of a complex disease, such as common infections. In this much more complex context, leprosy has been considered as an excellent model to the study of genetic susceptibility to common infectious diseases.[106] It is reasonable to believe that innovative approaches based on NGS technology could help to unravel much of the "missing heritability" observed in leprosy and other infectious diseases. Also, classic experimental design such as linkage analysis can be coupled to these approaches to increase its power.

References

1. Casanova JL, Abel L. Inborn errors of immunity to infection: The rule rather than the exception. *J Exp Med.* 2005;202:197-201.
2. Chapman SJ, Hill AV. Human genetic susceptibility to infectious disease. *Nat Rev Genet.* 2012;13:175-188.
3. Alcais A, Abel L, Casanova JL. *Hum Genet* of infectious diseases: Between proof of principle and paradigm. *J Clin Ivest.* 2009;119:2506-2514.
4. Picard C, Casanova JL, Abel L. Mendelian traits that confer predisposition or resistance to specific infections in humans. *Curr Opin Immunol.* 2006;18:383-390.
5. Abel L, Casanova JL. *Hum Genet* of infectious diseases: Fundamental insights from clinical studies. *Semin Immunol.* 2006;18:327-329.
6. Quintana-Murci L, Alcais A, Abel L, et al. Immunology in nature: Clinical, epidemiologic and evolutionary genetics of infectious diseases. *Nat Immunol.* 2007;8:1165-1171.
7. Casanova JL, Abel L. The human model: A genetic dissection of immunity to infection in natural conditions. *Nat Rev Immunol.* 2004;4:55-66.
8. Sorensen TI, Nielsen GG, Andersen PK, et al. Genetic and environmental influences on premature death in adult adoptees. *N Engl J Med.* 1988;318:727-732.
9. Orlova M, Di Pietrantonio T, Schurr E. Genetics of infectious diseases: Hidden etiologies and common pathways. *Clin Chem Lab Med.* 2011;49: 1427-1437.
10. Bellamy R. Genetic susceptibility to tuberculosis. *Clin Chest Med.* 2005;26:233-246. [vi].
11. Monot M, Honore N, Garnier T, et al. Comparative genomic and phylogeographic analysis of Mycobacterium leprae. *Nat Genet.* 2009;41: 1282-1289.
12. Truman RW, Singh P, Sharma R, et al. Probable zoonotic leprosy in the southern United States. *N Engl J Med.* 2011;364:1626-1633.
13. Schurr E, Alcais A, de Leseleuc L, et al. Genetic predisposition to leprosy: A major gene reveals novel pathways of immunity to Mycobacterium leprae. *Semin Immunol.* 2006;18:404-410.
14. Shields ED, Russell DA, Pericak-Vance MA. Genetic epidemiology of the susceptibility to leprosy. *J Clin Ivest.* 1987;79:1139-1143.
15. Mohamed AP, Ramanujam K. Leprosy in twins. *Int J Lepr.* 1966;34:405-407.
16. Chakravartti MR, Vogel F. A twin study on leprosy. Georg Thieme Verlag. 19731-123.
17. Abel L, Vu DL, Oberti J, et al. Complex segregation analysis of leprosy in southern Vietnam. *Genet Epidemiol.* 1995;12:63-82.
18. Abel L, Demenais F. Detection of major genes for susceptibility to leprosy and its subtypes in a Caribbean island: Desirade island. *Am J Hum Genet.* 1988;42:256-266.

19. Lazaro FP, Werneck RI, Mackert CC, et al. A major gene controls leprosy susceptibility in a hyperendemic isolated population from north of Brazil. *J Infect Dis.* 2010;201:1598-1605.
20. Prado Montes de Oca E. Human Polymorphisms as Clinical Predictors in Leprosy. *J Trop Med.* 2011, Article ID 923943, 14 pages (1-14) doi: 10.1155/2011/923943.
21. Trowsdale J. The MHC, disease and selection. *Immunol Lett.* 2011;137:1-8.
22. Mira MT. Genetic host resistance and susceptibility to leprosy. *Microbes Infect.* 2006;8:1124-1131.
23. Fava VM, Mira MT. Genetics of leprosy. In: Nunzi E, Massone C, eds. *Leprosy: A Practical Guide.* Springer; 2012. p. 19-26.
24. Schauf V, Ryan S, Scollard D, et al. Leprosy associated with HLA-DR2 and DQw1 in the population of northern Thailand. *Tissue Antigens.* 1985;26:243-247.
25. Soebono H, Giphart MJ, Schreuder GM, et al. Associations between HLA-DRB1 alleles and leprosy in an Indonesian population. *Int J Lepr Other Mycobact Dis.* 1997;65:190-196.
26. Tosh K, Ravikumar M, Bell JT, et al. Variation in MICA and MICB genes and enhanced susceptibility to paucibacillary leprosy in South India. *Hum Mol Genet.* 2006;15:2880-2887.
27. Motta PM, Cech N, Fontan C, et al. Role of HLA-DR and HLA-DQ alleles in multibacillary leprosy and paucibacillary leprosy in the province of Chaco (Argentina). *Enferm Infecc Microbiol Clin.* 2007;25:627-631.
28. Vanderborght PR, Pacheco AG, Moraes ME, et al. HLA-DRB1*04 and DRB1*10 are associated with resistance and susceptibility, respectively, in Brazilian and Vietnamese leprosy patients. *Genes Immun.* 2007;8:320-324.
29. da Silva SA, Mazini PS, Reis PG, et al. HLA-DR and HLA-DQ alleles in patients from the south of Brazil: Markers for leprosy susceptibility and resistance. *BMC Infect Dis.* 2009;9:134.
30. Zhang F, Liu H, Chen S, et al. Evidence for an association of HLA-DRB1*15 and DRB1*09 with leprosy and the impact of DRB1*09 on disease onset in a Chinese Han population. *BMC Med Genet.* 2009;10:133.
31. van Eden W, de Vries RR, D'Amaro J, et al. HLA-DR-associated genetic control of the type of leprosy in a population from surinam. *Hum Immunol.* 1982;4:343-350.
32. Gorodezky C, Flores J, Arevalo N, et al. Tuberculoid leprosy in Mexicans is associated with HLA-DR3. *Lepr Rev.* 1987;58:401-406.
33. Shankarkumar U. HLA associations in leprosy patients from Mumbai, India. *Lepr Rev.* 2004;75:79-85.
34. Alter A, Huong NT, Singh M, et al. Human leukocyte antigen class I region single-nucleotide polymorphisms are associated with leprosy susceptibility in Vietnam and India. *J Infect Dis.* 2011;203:1274-1281.
35. Cardoso CC, Pereira AC, de Sales Marques C, et al. Leprosy susceptibility: Genetic variations regulate innate and adaptive immunity, and disease outcome. *Future Microbiol.* 2011;6:533-549.
36. Misch EA, Berrington WR, Vary Jr JC, et al. Leprosy and the human genome. *Microbiol Mol Biol Rev.* 2010;74:589-620.
37. Cardoso CC, Pereira AC, Brito-de-Souza VN, et al. TNF −308 G>A single nucleotide polymorphism is associated with leprosy among Brazilians: A genetic epidemiology assessment, meta-analysis, and functional study. *J Infect Dis.* 2011;204:1256-1263.
38. Roach DR, Briscoe H, Saunders B, et al. Secreted lymphotoxin-alpha is essential for the control of an intracellular bacterial infection. *J Exp Med.* 2001;193:239-246.
39. Hagge DA, Saunders BM, Ebenezer GJ, et al. Lymphotoxin-alpha and TNF have essential but independent roles in the evolution of the granulomatous response in experimental leprosy. *Am J Pathol.* 2009;174:1379-1389.
40. Alcais A, Alter A, Antoni G, et al. Stepwise replication identifies a low-producing lymphotoxin-alpha allele as a major risk factor for early-onset leprosy. *Nat Genet.* 2007;39:517-522.
41. Hsieh CS, Heimberger AB, Gold JS, et al. Differential regulation of T helper phenotype development by interleukins 4 and 10 in an alpha beta T-cell-receptor transgenic system. *Proc Natl Acad Sci U S A.* 1992;89:6065-6069.
42. Ishida H, Hastings R, Thompson-Snipes L, et al. Modified immunologic status of anti-IL-10 treated mice. *Cell Immunol.* 1993;148:371-384.
43. Kuhn R, Lohler J, Rennick D, et al. Interleukin-10-deficient mice develop chronic enterocolitis. *Cell.* 1993;75:263-274.
44. Santos AR, Suffys PN, Vanderborght PR, et al. Role of tumor necrosis factor-alpha and interleukin-10 promoter gene polymorphisms in leprosy. *J Infect Dis.* 2002;186:1687-1691.
45. Moraes MO, Pacheco AG, Schonkeren JJ, et al. Interleukin-10 promoter single-nucleotide polymorphisms as markers for disease susceptibility and disease severity in leprosy. *Genes Immun.* 2004;5:592-595.
46. Fitness J, Floyd S, Warndorff DK, et al. Large-scale candidate gene study of leprosy susceptibility in the Karonga district of northern Malawi. *Am J Trop Med Hyg.* 2004;71:330-340.
47. Malhotra D, Darvishi K, Sood S, et al. IL-10 promoter single nucleotide polymorphisms are significantly associated with resistance to leprosy. *Hum Genet.* 2005;118:295-300.
48. Pereira AC, Brito-de-Souza VN, Cardoso CC, et al. Genetic, epidemiologic and biological analysis of interleukin-10 promoter single-nucleotide polymorphisms suggests a definitive role for −819 C/T in leprosy susceptibility. *Genes Immun.* 2009;10:174-180.
49. Franceschi DS, Mazini PS, Rudnick CC, et al. Influence of TNF and IL10 gene polymorphisms in the immunopathogenesis of leprosy in the south of Brazil. *Int J Infect Dis.* 2009;13:493-498.
50. Kang TJ, Yeum CE, Kim BC, et al. Differential production of interleukin-10 and interleukin-12 in mononuclear cells from leprosy patients with a Toll-like receptor 2 mutation. *Immunology.* 2004;112:674-680.
51. Mira MT, Alcais A, Van Thuc N, et al. Chromosome 6 q25 is linked to susceptibility to leprosy in a Vietnamese population. *Nat Genet.* 2003;33:412-415.
52. Mira MT, Alcais A, Nguyen VT, et al. Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature.* 2004;427:636-640.
53. Malhotra D, Darvishi K, Lohra M, et al. Association study of major risk single nucleotide polymorphisms in the common regulatory region of PARK2 and PACRG genes with leprosy in an Indian population. *EJHG.* 2006;14:438-442.
54. Bakija-Konsuo A, Mulic R, Boraska V, et al. Leprosy epidemics during history increased protective allele frequency of PARK2/PACRG genes in the population of the Mljet Island, Croatia. *Eur J Med Genet.* 2011;54:e548-e552.
55. Alter A, Fava VM, Huong NT, et al. Linkage disequilibrium pattern and age-at-diagnosis are critical for replicating genetic associations across ethnic groups in leprosy. *Hum Genet.* 2013;132:107-116.
56. Ali S, Vollaard AM, Widjaja S, et al. PARK2/PACRG polymorphisms and susceptibility to typhoid and paratyphoid fever. *Clin Exp Immunol.* 2006;144:425-431.
57. Fulco TO, Lopes UG, Sarno EN, et al. The proteasome function is required for Mycobacterium leprae-induced apoptosis and cytokine secretion. *Immunol Lett.* 2007;110:82-85.
58. Manry J, Quintana-Murci L. A genome-wide perspective of human diversity and its implications in infectious disease. *Cold Spring Harb Perspect Med.* 2013;3:a012450.
59. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inform.* 2012;10:117-122.
60. Zhang FR, Huang W, Chen SM, et al. Genomewide association study of leprosy. *N Engl J Med.* 2009;361:2609-2618.
61. Zhang F, Liu H, Chen S, et al. Identification of two new loci at IL23 R and RAB32 that influence susceptibility to leprosy. *Nat Genet.* 2011;43:1247-1251.
62. Ali S, Srivastava AK, Chopra R, et al. IL12 B SNPs and copy number variation in IL23 R gene associated with susceptibility to leprosy. *J Med Genet.* 2013;50:34-42.
63. Schurr E, Gros P. A common genetic fingerprint in leprosy and Crohn's disease? *N Engl J Med.* 2009;361:2666-2668.

64. Wong SH, Gochhait S, Malhotra D, et al. Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog.* 2010;6:e1000979.
65. Wong SH, Hill AV, Vannberg FO, et al. Genomewide association study of leprosy. *N Engl J Med.* 2010;362:1446-1447. [author reply 1447–1448].
66. Grant AV, Alter A, Huong NT, et al. Crohn's disease susceptibility genes are associated with leprosy in the Vietnamese population. *J Infect Dis.* 2012;206:1763-1767.
67. Berrington WR, Macdonald M, Khadge S, et al. Common polymorphisms in the NOD2 gene region are associated with leprosy and its reactive states. *J Infect Dis.* 2010;201:1422-1435.
68. Hitotsumatsu O, Ahmad RC, Tavares R, et al. The ubiquitin-editing enzyme A20 restricts nucleotide-binding oligomerization domain containing 2-triggered signals. *Immunity.* 2008;28:381-390.
69. Kanneganti TD, Lamkanfi M, Nunez G. Intracellular NOD-like receptors in host defense and disease. *Immunity.* 2007;27:549-559.
70. Hasegawa M, Fujimoto Y, Lucas PC, et al. A critical role of RICK/RIP2 polyubiquitination in Nod-induced NF-kappaB activation. *EMBO J.* 2008;27:373-383.
71. Schenk M, Krutzik SR, Sieling PA, et al. NOD2 triggers an interleukin-32-dependent human dendritic cell program in leprosy. *Nat Med.* 2012;18:555-563.
72. Krutzik SR, Tan B, Li H, et al. TLR activation triggers the rapid differentiation of monocytes into macrophages and dendritic cells. *Nat Med.* 2005;11:653-660.
73. Sieling PA, Chatterjee D, Porcelli SA, et al. CD1-restricted T cell recognition of microbial lipoglycan antigens. *Science.* 1995;269:227-230.
74. Sieling PA, Jullien D, Dahlem M, et al. CD1 expression by dendritic cells in human leprosy lesions: Correlation with effective host immunity. *J Immunol.* 1999;162:1851-1858.
75. Alter A, Grant A, Abel L, et al. Leprosy as a genetic disease. *Mamm Genome.* 2011;22:19-31.
76. Sampaio LH, Stefani MM, Oliveira RM, et al. Immunologically reactive M. leprae antigens with relevance to diagnosis and vaccine development. *BMC Infect Dis.* 2011;11:26.
77. Sousa AL, Fava VM, Sampaio LH, et al. Genetic and immunologic evidence implicates interleukin 6 as a susceptibility gene for leprosy type 2 reaction. *J Infect Dis.* 2012;205:1417-1424.
78. Kahawita IP, Walker SL, Lockwood DNJ. Leprosy type 1 reactions and erythema nodosum leprosum. *An Bras Dermatol.* 2008;83:75-82.
79. Britton WJ, Lockwood DN. Leprosy. *Lancet.* 2004;363:1209-1219.
80. Fava V, Orlova M, Cobat A, et al. Genetics of leprosy reactions: An overview. *Mem Inst Oswaldo Cruz.* 2012;107(Suppl 1):132-142.
81. Misch EA, Hawn TR. Toll-like receptor polymorphisms and susceptibility to human disease. *Clin Sci (Lond).* 2008;114:347-360.
82. Bochud PY, Sinsimer D, Aderem A, et al. Polymorphisms in Toll-like receptor 4 (TLR4) are associated with protection against leprosy. *Eur J Clin Microbiol Infect Dis.* 2009;28:1055-1065.
83. Schuring RP, Hamann L, Faber WR, et al. Polymorphism N248 S in the human Toll-like receptor 1 gene is related to leprosy and leprosy reactions. *J Infect Dis.* 2009;199:1816-1819.
84. de Sales Marques C, Brito-de-Souza VN, Albuquerque Guerreiro LT, et al. Toll-like receptor 1 (TLR1) N248 S single nucleotide polymorphism is associated with leprosy risk and regulates immune activation during mycobacterial infection. *J Infect Dis.* 2013;208:120-129.
85. Bochud PY, Hawn TR, Siddiqui MR, et al. Toll-like receptor 2 (TLR2) polymorphisms are associated with reversal reaction in leprosy. *J Infect Dis.* 2008;197:253-261.
86. Krutzik SR, Ochoa MT, Sieling PA, et al. Activation and regulation of Toll-like receptors 2 and 1 in human leprosy. *Nat Med.* 2003;9:525-532.
87. Mattos KA, Oliveira VG, D'Avila H, et al. TLR6-driven lipid droplets in Mycobacterium leprae-infected Schwann cells: Immunoinflammatory platforms associated with bacterial persistence. *J Immunol.* 2011;187:2548-2558.
88. Sapkota BR, Macdonald M, Berrington WR, et al. Association of TNF, MBL, and VDR polymorphisms with leprosy phenotypes. *Hum Immunol.* 2010;71:992-998.
89. Teixeira MA, Silva NL, Ramos AL, et al. NRAMP1 gene polymorphisms in individuals with leprosy reactions attended at two reference centers in Recife, northeastern Brazil. *Rev Soc Bras Med Trop.* 2010;43:281-286.
90. Greenstein RJ. Is Crohn's disease caused by a mycobacterium? Comparisons with leprosy, tuberculosis, and Johne's disease. *Lancet Infect Dis.* 2003;3:507-514.
91. Jeyanathan M, Boutros-Tadros O, Radhi J, et al. Visualization of Mycobacterium avium in Crohn's tissue by oil-immersion microscopy. *Microbes Infect.* 2007;9:1567-1573.
92. Behr MA, Schurr E. Mycobacteria in Crohn's disease: A persistent hypothesis. *Inflamm Bowel Dis.* 2006;12:1000-1004.
93. Liu H, Irwanto A, Tian H, et al. Identification of IL18 RAP/IL18 R1 and IL12 B as leprosy risk genes demonstrates shared pathogenesis between inflammation and infectious diseases. *Am J Hum Genet.* 2012;91:935-941.
94. Thevenet J, Pescini Gobert R, Hooft van Huijsduijnen R, et al. Regulation of LRRK2 expression points to a functional role in human monocyte maturation. *PLoS One.* 2011;6:e21519.
95. Umeno J, Asano K, Matsushita T, et al. Meta-analysis of published studies identified eight additional common susceptibility loci for Crohn's disease and ulcerative colitis. *Inflamm Bowel Dis.* 2011;17:2407-2415.
96. Smith WW, Pei Z, Jiang H, et al. Leucine-rich repeat kinase 2 (LRRK2) interacts with parkin, and mutant LRRK2 induces neuronal degeneration. *Proc Natl Acad Sci U S A.* 2005;102:18676-18681.
97. Frazer KA, Murray SS, Schork NJ, et al. Human genetic variation and its contribution to complex traits. *Nat Rev Genet.* 2009;10:241-251.
98. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461:747-753.
99. Maher B. Personal genomes: The case of the missing heritability. *Nature.* 2008;456:18-21.
100. Zhang J, Chiodini R, Badr A, et al. The impact of next-generation sequencing on genomics. *J Genet Genomics.* 2011;38:95-109.
101. Schork NJ, Murray SS, Frazer KA, et al. Common versus rare allele hypotheses for complex diseases. *Curr Opin Genet Dev.* 2009;19:212-219.
102. Altare F, Lammas D, Revy P, et al. Inherited interleukin 12 deficiency in a child with bacille Calmette-Guerin and Salmonella enteritidis disseminated infection. *J Clin Ivest.* 1998;102:2035-2040.
103. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 2001;69:124-137.
104. Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet.* 2010;11:31-46.
105. Novelli G, Predazzi IM, Mango R, et al. Role of genomics in cardiovascular medicine. *World J Cardiol.* 2010;2:428-436.
106. Alter A, Alcais A, Abel L, et al. Leprosy as a genetic model for susceptibility to common infectious diseases. *Hum Genet.* 2008;123:227-235.

Title: Genetics of leprosy: Expected and unexpected developments and perspectives

Author: Monica E.D. Sauer,Heloisa Salomão,Geovana B. Ramos,Helena R.S. D`Espindula,Rafael S.A. Rodrigues,Wilian C. Macedo,Renata H.M. Sindeaux,Marcelo T. Mira

Publication: Clinics in Dermatology

Publisher: Elsevier

Date: January–February 2015

Order Completed

Thank you for your order.

This Agreement between Monica Dallmann Sauer ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

Printable details.

License Number	3982820904096
License date	Nov 05, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Clinics in Dermatology
Licensed Content Title	Genetics of leprosy: Expected and unexpected developments and perspectives
Licensed Content Author	Monica E.D. Sauer,Heloisa Salomão,Geovana B. Ramos,Helena R.S. D`Espindula,Rafael S.A. Rodrigues,Wilian C. Macedo,Renata H.M. Sindeaux,Marcelo T. Mira
Licensed Content Date	January–February 2015
Licensed Content Volume	33
Licensed Content Issue	1
Licensed Content Pages	9
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	WHOLE GENOME SEQUENCING OF A FAMILY WITH MONOZYGOTIC TWINS DISPLAYING EARLY-ONSET LEPROSY
Expected completion date	Nov 2016
Estimated size (number of pages)	136
Elsevier VAT number	GB 494 6272 12

Requestor Location	Monica Dallmann Sauer Rua Atilio Borio 161 Curitiba, 80050250 Brazil Attn: Monica Dallmann Sauer
Total	0.00 USD

ORDER MORE **CLOSE WINDOW**