

**EARLY POSTZYGOTIC MUTATIONS CONTRIBUTE
TO *DE NOVO* VARIATION IN A HEALTHY
MONOZYGOTIC TWIN PAIR**

A THESIS
SUBMITTED TO THE DEPARTMENT
OF MOLECULAR BIOLOGY AND GENETICS
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Gülşah Merve Dal
September, 2014

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Tayfun Özçelik (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Hilal Özdağ

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assist. Prof. Dr. Ebru Erbay

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assist. Prof. Dr. Can Alkan

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assist. Prof. Dr. Ali Güre

Approved for the Graduate School of Engineering and Science

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

EARLY POSTZYGOTIC MUTATIONS CONTRIBUTE TO *DE NOVO* VARIATION IN A HEALTHY MONOZYGOTIC TWIN PAIR

Gülşah Merve Dal

Ph.D. in Molecular Biology and Genetics

Supervisor: Prof. Dr. Tayfun Özçelik

September, 2014

Characterizing the patterns and rate of *de novo* mutations is crucial for our perception of evolution and genetic basis of human disease. Direct observation of *de novo* single nucleotide variation (SNV) rate in healthy individuals revealed a rate in a range of $0.82 - 1.70 \times 10^{-8}$ base pair per generation. However, the developmental timing of the *de novo* mutations is unknown and thus, contribution of the early post-zygotic mutations to the human *de novo* SNV rate remained unknown. In an attempt to estimate the rate of *de novo* mutations regarding the developmental timing of mutagenesis, we sequenced the whole genomes of a healthy monozygotic twin pair and their parents with a total of 170 fold coverage. We identified the *de novo* SNVs through examination of the genotypes of each individual for each of the variants in a synchronous manner. Subsequent to the Sanger sequencing based validation, we conservatively characterized a total of 32 *de novo* SNVs. Of these 23 were shared by the twin pair, 8 were specific to twin I, and 1 was specific to twin II. We estimated the overall *de novo* SNV rate of 1.31×10^{-8} for twin I and 1.01×10^{-8} for twin II. The rate of the early post-zygotic *de novo* SNVs was calculated to be 0.34×10^{-8} and 0.04×10^{-8} for twin I and twin II, respectively. These data indicate the growing importance of genome mosaicism which might be resulted from *de novo* mutations of early post-zygotic origin in disease pathogenesis.

Keywords: Mutation rate, *de novo* SNV, monozygotic twins, somatic mosaicism, early post-zygotic, next generation sequencing.

ÖZET

SAĞLIKLI BİR TEK YUMURTA İKİZ ÇİFTİNDE ERKEN POSTZİGOTİK MUTASYONLARIN *DE NOVO* MUTASTON ORANINA KATKISI

Gülşah Merve Dal

Moleküler Biyoloji ve Genetik, Doktora

Tez Yöneticisi: Prof. Dr. Tayfun Özçelik

Eylül, 2014

De novo mutasyonların özellikleri ve oranının bilinmesi hastalıkların ve evrimsel süreçlerin anlaşılabilmesi için büyük önem taşımaktadır. *De novo* tek nükleotid değişikliklerinin (SNV) doğrudan incelenmesi ile insanda bir nesilde ortaya çıkan yeni tek nükleotid değişikliği oranı $0.82 - 1.70 \times 10^{-8}$ baz çifti olarak belirlenmiştir. Ancak, bu tek nükleotid değişikliklerinin gelişimin hangi evresinde ortaya çıktığı ve dolayısıyla erken post-zigotik mutasyonların toplam *de novo* mutasyon oranına olan katkısı bilinmemektedir. Bu noktada, insanda bir nesilde ortaya çıkan yeni mutasyon oranını ortaya çıkma zamanına göre belirlemek amacıyla; anne, baba ve sağlıklı bir erkek ikiz çiftinden oluşan ailede toplam 170 X kapsamlı tüm genom dizilemesi yaptık. Bulunan her bir mutasyon için her bir bireyin genotiplerini karşılaştırmalı ve eş zamanlı olarak analiz ederek *de novo* tek nükleotid değişikliklerinin belirledik. Sanger dizilemesi ile yapılan doğrulamanın ardından toplam 32 *de novo* tek nükleotid değişikliğini karakterize ettik. Bu mutasyonlardan 23 tanesi ikizlerin her ikisinin genomunda, 8 tanesi ikizlerden birinde 1 tanesi ise diğerinde bulunmaktadır. Toplam *de novo* mutasyon oranını ikizlerden biri için 1.31×10^{-8} , diğeri için 1.01×10^{-8} olarak hesapladık. Toplam mutasyon oranı içinde, erken post-zigotik mutasyonların oranını ise ikizlerden biri için 0.34×10^{-8} ve diğeri için 0.04×10^{-8} olarak belirledik. Bulgularımız erken post-zigotik mutasyonların sebep olduğu genom mozaizminin hastalıkların anlaşılması için önem taşıdığını göstermiştir.

Anahtar Sözcükler: Mutasyon oranı, *de novo* SNV, tek yumurta ikizleri, somatik mozaizizm, erken post-zigotik, yeni nesil dizileme.

To my mother and father
Aylin and Hüseyin Dal

Acknowledgement

First and foremost I would like to express my sincere gratitude to my advisor Prof. Tayfun Özçelik. He has always been encouraging, supportive and compassionate. He provided me all facilities and his immense knowledge in the field of human genetics to complete my doctoral research successfully. I appreciate his continuous support, willingness to work on a tight schedule and constructive criticism. I consider it as a great opportunity to complete my Ph.D under his guidance. He challenged me to be a better scientist and helped me grow. I feel that I will always be grateful to him for all his teachings.

I would like to gratefully and sincerely thank Assist. Prof. Can Alkan for his continuous support, invaluable guidance, patience, and refreshing sense of humor. He has taught me the next generation sequencing data analysis and how to deal with a huge amount of data. He provided me access to his computer facilities to complete my research work. He has been always available with positive and encouraging advice despite his restricted time. I am grateful to Dr. Alkan for his generosity in terms of guidance, time and resources. I will forever be thankful to him from the depth of my heart.

I am thankful to Assist. Prof. Ebru Erbay for her contribution in my thesis following committee, providing advices and suggestions. She encouraged and supported me during my research work. I appreciate her valuable comments and advices.

I would like to thank the family participated in this study. I would like to acknowledge Dr. Bayram Yüksel, Dr. Mahmut Şamil Sağıroğlu, Bekir Ergüner and Pınar Kavak for their effort in the whole genome sequencing data production and analysis. I am also grateful to Enver Kayaaslan since he has taught me bash scripting.

I am thankful to my best friend in the lab, Füsün Doldur Ballı, for her emotional support, encouragement and valuable suggestions in my research. I will remember those times that we shared together with laugh forever.

I am grateful to Dr. Onur Emre Onat for his continuous support and impactful advice in my research. I appreciate him for his help in my experiments and data analysis. I also would like to thank Dr. Süleyman Gülsüner and Melis Atalar for their help in various forms.

I would like to thank Defne Bayık, Verda Bitirim, Gözde Güçlüler, Dilan Çelebi, Merve Mutlu, Merve Aydın, Ece Akhan, Ayça Ergül, Gurbet Karahan, Nilüfer Sayar, İnci Şimşek, Begüm Horuluoğlu and all my friends in the lab for their pleasant interactions and supportive friendship.

I am grateful to İclal Özçelik for her support during the manuscript writing and editing. I appreciate her for providing invaluable suggestions.

I appreciate the financial support from Bilkent University, Turkish State Planning Organization, Turkish Academy of Sciences, EMBO, and The Science Academy.

I would like to express my heart-felt gratitude and eternal love to Cihan Kılınç. I am grateful to him for his continuous support, encouragement and patience.

I dedicate this thesis to my mother Aylin, my father Hüseyin Kemal and my brother Taylan Eren Dal. I am indebted to them for their constant love, support, encouragement, and guidance forever.

Gülşah Merve Dal

Contents

1 Introduction.....	1
1.1 Genetic Variation in the Human Genome	1
1.1.1 Inherited mutations	2
1.1.2 <i>De novo</i> mutations	3
1.2 Patterns of the <i>De Novo</i> Mutations.....	5
1.2.1 Origins of the <i>de novo</i> mutations.....	5
1.2.2 Genome-wide distribution patterns of the <i>de novo</i> mutations.....	8
1.2.3 Sex and age specific patterns of the <i>de novo</i> mutations	9
1.2.4 Role of the <i>de novo</i> mutations in diseases	10
1.3 Rates of the <i>De Novo</i> Mutations.....	11
1.3.1 Methods for estimating <i>de novo</i> mutation rate	12
1.3.1.1 Indirect method	12

1.3.1.2 Direct method.....	13
1.3.2 <i>De novo</i> mutation rates	13
1.3.2.1 <i>De novo</i> mutation rate based on the indirect approach	13
1.3.2.2 <i>De novo</i> mutation rate based on the direct approach	14
1.3.2.2.1 <i>De novo</i> germline mutation rate.....	14
1.3.2.2.1.1 <i>De novo</i> SNV rate	14
1.3.2.2.1.2 <i>De novo</i> indel and CNV rate	15
1.3.2.2.2 Somatic mutation rate	16
1.4 Mosaicism in the Human Genome	17
1.5 Dissecting the Origins of New Mutations	18
1.6 Importance of the Incidental Findings in the Next Generation Sequencing Data	19
1.7 Aim and Strategy	20
2 Materials and Methods.....	22
2.1 Subjects of the Study	22
2.2 DNA Isolation	23
2.2.1 DNA isolation from the peripheral blood.....	23
2.2.2 DNA isolation from the buccal wash specimen	23
2.2.3 DNA isolation from the urine specimen.....	24
2.2.4 Assessment of the quality and the quantity of the DNA samples.....	24

2.3 Whole Genome Sequencing	25
2.3.1 Sample preparation	26
2.3.2 Paired-end library construction.....	26
2.3.3 Paired-end whole genome sequencing.....	26
2.4 Analysis of the Whole Genome Sequencing Data	27
2.4.1 Mapping of the sequencing reads	27
2.4.2 Discovery of the genomic variants and genotypes	27
2.4.3 Functional annotation of the variants	28
2.4.4 File formats	28
2.5 Identification of the Putative <i>De Novo</i> SNVs.....	28
2.5.1 Identification of the <i>de novo</i> SNV candidates shared by the twin pair	29
2.5.2 Identification of the twin specific <i>de novo</i> SNV candidates.....	29
2.5.3 Filtering the <i>de novo</i> SNV candidates	30
2.6 Evaluation of the Putative <i>De Novo</i> SNVs.....	30
2.6.1 Visual inspection of the sequencing data.....	30
2.6.2 Calculation of the reference allele frequency	31
2.7 Validation of the <i>De Novo</i> SNVs	31
2.7.1 Primer design	31
2.7.2 PCR amplification	32
2.7.3 Visualization of the PCR products	32

2.7.4 Sanger sequencing	34
2.7.5 Analysis of the sequencing data	34
2.8 Calculation of the <i>De Novo</i> SNV Rate	34
2.8.1 Formula.....	34
2.8.2 Calculation of the number of callable bases	35
2.8.3 Calculation of the false negative rate.....	35
2.9 Analysis of the Incidental Results	36
2.10 Enzymes, Chemicals and Reagents	37
2.10.1 Enzymes.....	37
2.10.2 Chemicals and reagents	38
2.10.3 Standard solutions and buffers	38
2.11 Web Sources	39
3 Results	40
3.1 Whole Genome Sequencing	40
3.1.1 Sample preparation and library construction	41
3.1.2 Paired-end whole genome sequencing	42
3.1.3 Whole genome sequencing data analysis: Mapping and variants calling..	42
3.2 Discovery of the Putative <i>De Novo</i> SNVs.....	47
3.2.1 Discovery of the putative <i>de novo</i> SNVs shared by the twin pair	47
3.2.2 Discovery of the twin-specific putative <i>de novo</i> SNVs	49

3.2.3 Assessment of the validity of filtering process	52
3.3 Selection of the Putative <i>De Novo</i> SNVs for Sanger Sequencing	56
3.3.1 Visual examination-based evaluation and selection	56
3.3.2 Sequencing data-based evaluation of the putative <i>de novo</i> SNVs	58
3.4 Classification of the Putative <i>De Novo</i> SNVs Regarding the Genomic Context.	60
3.5 Experimental Validation of the <i>De Novo</i> SNVs	62
3.5.1 Validation of the <i>de novo</i> SNVs shared by the twin pair	62
3.5.2 Validation of the twin-specific <i>de novo</i> SNVs	66
3.5.3 Evaluation of the Sanger sequencing results	68
3.6 Investigation of the Possible Effect of the Parental Mosaicism.....	71
3.7 Examination of the dbSNP137-Reported Variants	71
3.8 Calculation of the <i>De Novo</i> SNV Rate.....	77
3.9 Examination of the Incidental Findings	79
3.9.1 Analysis of the LoF mutations	81
3.9.1.1 LoF mutations of the twin pair	81
3.9.1.1.1 Homozygous LoF mutations of the twin pair	81
3.9.1.1.2 Heterozygous LoF mutations of the twin pair	84
3.9.1.2 LoF mutations of the parents	86
3.9.1.2.1 Homozygous LoF mutations of the parents	86
3.9.1.2.2 Heterozygous LoF mutations of the parents	88
3.9.2 Analysis of the missense mutations	90

3.9.3 Analysis of the homozygous stretches	97
4 Discussion.....	100
References	106
Appendices	119
Appendix A.....	120
Appendix B.....	125
Appendix C.....	134
Appendix D.....	145
Appendix E	167
Appendix F	176
Appendix G.....	186
Appendix H.....	189
Appendix I	195
Publications.....	199

List of Figures

1.1 Demonstration of the inherited mutations and their transmission	2
1.2 Demonstration of the timing of <i>de novo</i> mutations.....	3
1.3 Heat map and extended pedigree showing the effect of the mutations	4
1.4 Demonstration of a pre-zygotic <i>de novo</i> mutation.....	5
1.5 Demonstration of an early post-zygotic <i>de novo</i> mutation.	6
1.6 Demonstration of a late post-zygotic <i>de novo</i> mutation.....	7
2.1 DNA marker used in the study.....	24
2.2 DNA marker used for the PCR product visualization.....	33
3.1 Pedigree of the quad family.	40
3.2 Densitometry analysis of the DNA samples using agarose gel electrophoresis. .	41
3.3 Classification of the putative <i>de novo</i> SNVs regarding the genomic context.....	61

3.4 A representative Sanger sequencing electropherogram for a shared <i>de novo</i> SNV present in a heterozygous state and validated by capillary sequencing.	63
3.5 A representative Sanger sequencing electropherogram for a putative <i>de novo</i> SNV discovered to be inherited based on the capillary sequencing result	64
3.6 A representative Sanger sequencing electropherogram for a putative <i>de novo</i> SNV discovered to be not present in the genomes of the twin pair	65
3.7 A representative Sanger sequencing electropherogram for a twin I-specific <i>de novo</i> SNV	67
3.8 Homozygosity mapping analysis in the twin pair	97
B.1 IGV screenshots for the total of 290 high-confidence putative <i>de novo</i> SNVs.	125
D.1 Electropherogram for the 2 <i>de novo</i> SNVs shared by the twin pair (Chromosome1: 165599592: G/C and Chromosome1: 165599593: T/G).....	145
D.2 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome1: 245093862: G/A)	146
D.3 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome2: 136342189: G/A)	147
D.4 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome4: 145566227: C/T)	148
D.5 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome5: 11043438: T/G).....	149
D.6 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome6: 29101417: C/T)	150

D.7 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome6: 162458275: T/C).....	151
D.8 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome7: 141955453: A/G)	152
D.9 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome9: 32917739: C/T).....	153
D.10 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome10: 102527333: C/T).....	154
D.11 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome10: 128805987: G/A)	155
D.12 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome12: 3527384: T/C).....	156
D.13 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome13: 97503636: T/C).....	157
D.14 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome14: 32048285: T/A).....	158
D.15 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome15: 57953050: A/T).....	159
D.16 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome15: 80788986: G/T).....	160
D.17 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome16: 59139115: A/G)	161
D.18 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome17: 79276311: C/T).....	162

D.19 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome18: 32459844: T/C)	163
D.20 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome21: 23811291: C/T)	164
D.21 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome22: 19167292: G/A)	165
D.22 Electropherogram for the <i>de novo</i> SNV shared by the twin pair (Chromosome10: 133937827: C/T)	166
E.1 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome2: 61728351: A/T)	167
E.2 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome3: 1493383: C/T)	168
E.3 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome3: 72959480: G/A)	169
E.4 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome4: 6564000: G/A)	170
E.5 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome7: 29606885: G/A)	171
E.6 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome8: 118838053: A/G)	172
E.7 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome16: 50424085: G/A)	173
E.8 Electropherogram for the <i>de novo</i> SNV specific to Twin-I (Chromosome20: 53271335: C/T)	174

E.9 Electropherogram for the <i>de novo</i> SNV specific to Twin-II (Chromosome21: 36322701: A/G)	175
F.1 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome1: 245093862: G/A)	176
F.2 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome5: 11043438: T/G).....	177
F.3 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome6: 29101417: C/T).....	178
F.4 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome7: 141955453: A/G)	179
F.5 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome10: 128805987: G/A)	180
F.6 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome12: 3527384: T/C).....	181
F.7 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome14: 32048285: T/A).....	182

F.8 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome16: 59139115: A/G)	183
F.9 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome17: 79276311: C/T)	184
F.10 Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the <i>de novo</i> SNV shared by the twin pair (Chromosome21: 23811291: C/T)	185

List of Tables

2.1 List of the enzymes used in the experiments	37
2.2 List of the chemicals and reagents used in the experiments	38
2.3 List of the standard solutions and buffers used in the experiments	38
2.4 Web sources used in the study design and data analysis	39
3.1 Spectrophotometric measurement of the concentrations of DNA samples	42
3.2 Statistics on the genome sequence produced by the paired-end whole genome sequencing.....	44
3.3 Percentage of coverage per genome.....	45
3.4 Statistics on all variants identified through whole genome sequencing	46
3.5 Identification of the putative <i>de novo</i> SNVs shared by the twin pair and application of the filters	48
3.6 Identification of the putative <i>de novo</i> SNVs specific to twin I and application of the filters.....	50

3.7 Identification of the putative <i>de novo</i> SNVs specific to twin II and application of the filters.....	51
3.8 Identification of the control group SNVs and application of the filters.....	53
3.9 Calculation of the RAF and average depth of coverage for the control group SNVs	54
3.10 Evaluation of the efficiency of the filters in the removal of false <i>de novo</i> calls	55
3.11 Classification of the putative <i>de novo</i> SNVs based on the visual inspection of the sequence alignment data	57
3.12 Average RAF and the depth of coverage for each class of the <i>de novo</i> SNV candidates.....	59
3.13 Statistics on <i>de novo</i> SNVs tested by Sanger sequencing.....	69
3.14 Average RAF and depth of coverage for the <i>de novo</i> SNVs tested by Sanger sequencing.....	71
3.15 Discovery of the dbSNP137 reported <i>de novo</i> SNP candidates shared by the twins.....	73
3.16 Discovery of the dbSNP137 reported <i>de novo</i> SNP candidates specific to twin I	73
3.17 Discovery of the dbSNP137 reported <i>de novo</i> SNP candidates that are specific to twin II.....	74
3.18 Global minor allele frequency distribution for the <i>de novo</i> SNP candidates.....	74
3.19 Filtering of the <i>de novo</i> SNPs which have a minor allele frequency (MAF) smaller than 1% based on the depth and genotype quality	75

3.20 List of high-confidence SNPs discovered as <i>de novo</i>	76
3.21 <i>De novo</i> mutations of the twin pair and the mutation rate estimation	78
3.22 Identification of the homozygous and heterozygous LoF and missense mutations of the four individual.....	80
3.23 Novel homozygous LoF mutations of the twin pair	82
3.24 dbSNP137-reported homozygous LoF mutations of the twin pair	83
3.25 List of heterozygous LoF mutations in the twin pair located on genes related to diseases	85
3.26 List of the homozygous LoF mutations located on disease genes in the parent	87
3.27 List of the heterozygous LoF mutations located on disease genes in the parent	89
3.28 Homozygous missense variants present in the genomes of the twin pair	91
3.29 Heterozygous missense variants located on disease related genes and present in the genomes of the twin pair	92
3.30 Homozygous intervals in the twin pair	98
A.1 Primer pairs for validation of the group I and group II <i>de novo</i> SNVs	120
B.1 Index table for the Figure B.1.....	126
C.1 List of the <i>de novo</i> SNV candidates (n=159) shared by the twin pair.....	134
C.2 List of the <i>de novo</i> SNV candidates (n=83) specific to twin I	140
C.3 List of the <i>de novo</i> SNV candidates (n=48) specific to twin II.....	143
G.1 List of novel heterozygous LoF mutations present in the genomes of the twin pair	186

G.2 List of dbSNP137-reported (MAF < 0.01) heterozygous LoF mutations present in the genomes of the twin pair	188
H.1 List of novel homozygous LoF mutations present in the genome of the mother	189
H.2 List of dbSNP137-reported (MAF < 0.01) homozygous LoF mutations present in the genome of the mother	191
H.3 List of novel homozygous LoF mutations present in the genome of the father	192
H.4 List of dbSNP137-reported (MAF < 0.01) homozygous LoF mutations present in the genome of the father	193
I.1 List of novel heterozygous LoF mutations present in the genome of the mother	195
I.2 List of dbSNP137-reported heterozygous LoF mutations present in the genome of the mother	197
I.3 List of novel heterozygous LoF mutations present in the genome of the father	197

Abbreviations

APC	Adenomatous polyposis coli
AB	Allelic balance
BAM file	Binary Sequence Alignment/Map file
bcl file	Base call file
bp	Base pair
BWA	Burrows-Wheeler Aligner
CNV	Copy number variation
EtBr	Ethidium bromide
FNR	False negative rate
GATK	Genome Analysis Toolkit
Gb	Gigabase
GQ	Genotype Quality

indel	Insertions and deletions
IGV	Integrative Genomics Viewer
IRB	Institutional review boards
kbp	Kilobase pair
MQ	Mapping Quality
ng	Nano gram
NAHR	Nonallelic homologous recombination
PCR	Polymerase Chain Reaction
QUAL	Quality
RAF	Reference allele frequency
rpm	Revolutions per minute
SAM file	Sequence Alignment/Map file
SB	Strand Bias
SETBP1	Set Binding Protein 1
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
std	Standard deviation
TCR	Transcription coupled repair
Ts/Tv	Transition to transversion
VCF	Variant call format
VQSR	Variant quality score recalibration

Chapter 1

Introduction

1.1 Genetic Variation in the Human Genome

Cells store the genetic information in the form of nucleotide sequences. Duplication of the hereditary information before each cell division is an essential process. This process is carried out precisely by each cell's DNA replication and repair machinery.[1] Still, nucleotide sequences are frequently subjected to mutations because of the environmental factors, stochastic events inside the cell, and failure of the replication and repair processes.[2]

Indeed, mutations are the raw materials of the genetic diversity in nature. Some of them are deleterious. Thus, they might cause diseases. Some might be neutral or advantageous. In that case, they become fixed in a population and provide substrates for evolution.[3]

Mutations can be classified into two broad categories with respect to their inheritance patterns and occurrence time: Inherited mutations and *de novo* mutations.

1.1.1 Inherited Mutations

Inherited mutations are the nucleotide changes that are already present in all cells of the parents. These mutations are transmitted through parental germ cells during the course of fertilization and thus present in all cells of the offspring (Figure 1.1).[4]

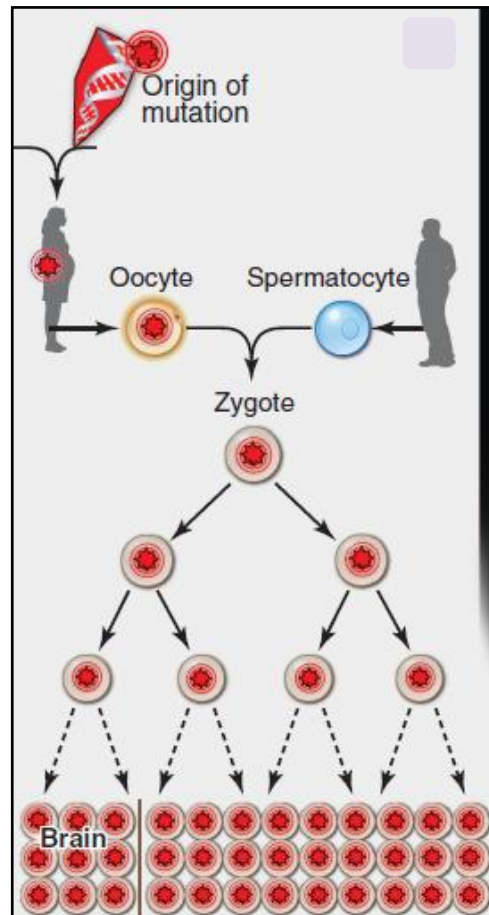


Figure 1.1: Demonstration of the inherited mutations and their transmission. The figure is taken from [A. Poduri, G. D. Evrony, X. Cai, C. A. Walsh, “Somatic mutation, genomic variation, and neurological disease,” *Science*, vol. 341, pp. 1237758-1-8, 2013]. Reprinted with permission from AAAS.

1.1.2 *De novo* Mutations

De novo mutations are the new mutations. They are present in the genome of an individual yet not present and be detectable in the constitutive DNA of the parents.[5]

According to the traditional view, they newly occur either in the parental germ cells during the gametogenesis or in the differentiated somatic cells of an individual during the postnatal development. Indeed, *de novo* mutations might occur at any time throughout the life cycle of an individual: They might appear during the early or late development of the embryo, fetal development, and postnatal growth (Figure 1.2).[4, 5, 6]

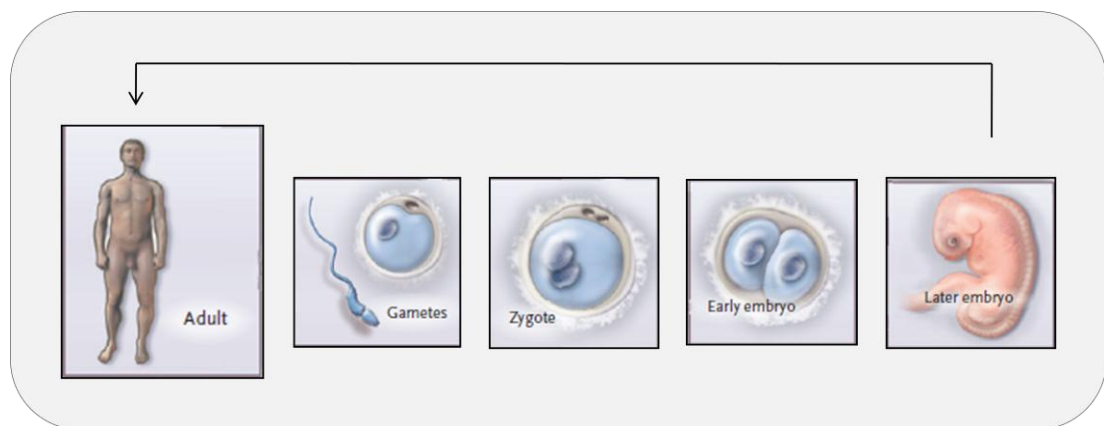


Figure 1.2: Demonstration of the timing of *de novo* mutations. Reproduced and modified with permission from (L. Vadlamudi, L. M. Dibbens, K. M. Lawrence, X. Iona, J. M. McMahon, W. Murrell, *et al.*, “Timing of De Novo Mutagenesis — A Twin Study of Sodium-Channel Mutations,” *N Engl J Med*, vol. 30, pp. 1335-40, 2010.), Copyright Massachusetts Medical Society.

De novo mutations are one of the most severe form of the rare genetic variation. They are more deleterious comparing to the mutations of ancient origin (Figure 1.3).[7, 8, 9] Their contribution to diseases might be larger than we anticipated before. Therefore, knowledge about their patterns and rates is important for our perception of the evolutionary processes as well as the genetic basis of diseases.

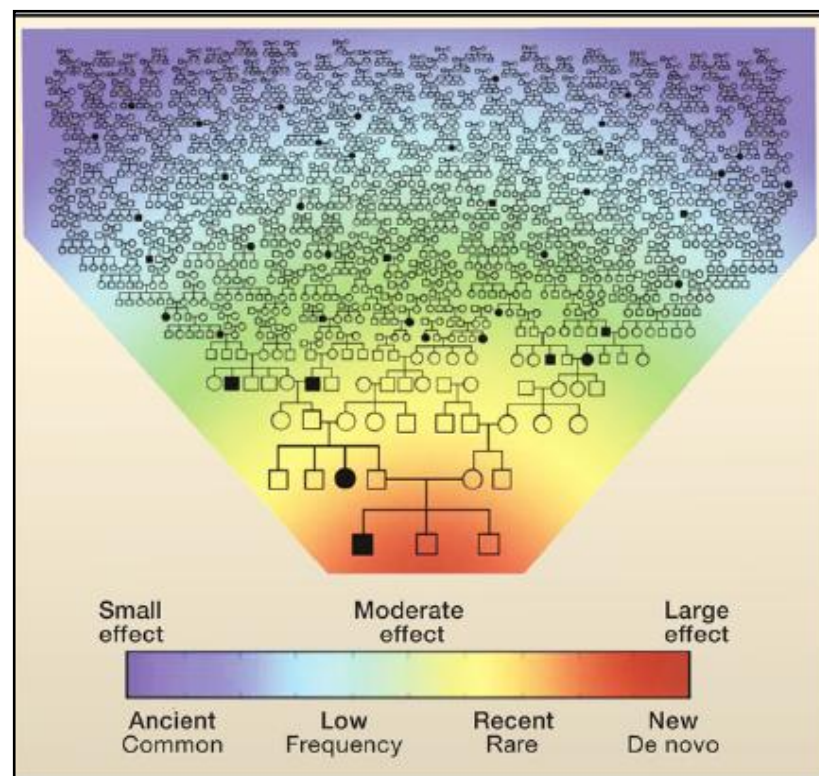


Figure 1.3: Heat map and extended pedigree showing the effect of the mutations. Copyright © 2014 Copyright Clearance Center, Inc. Reproduced from (J. R. Lupski, J. W. Belmont, E. Boerwinkle, R. A. Gibbs, “Clan genomics and the complex architecture of human disease,” *Cell*, vol. 147, pp. 32-43, 2010) with permission.

1.2 Patterns of the *De Novo* Mutations

1.2.1 Origins of the *de novo* mutations

A *de novo* mutation might occur during the gametogenesis in parents (A pre-zygotic *de novo* mutation). In that case, the mutation which is introduced during the spermatogenesis or oogenesis is not present in the parental constitutive genome yet it is transmitted and present in all cells of the offspring (Figure 1.4).[4]

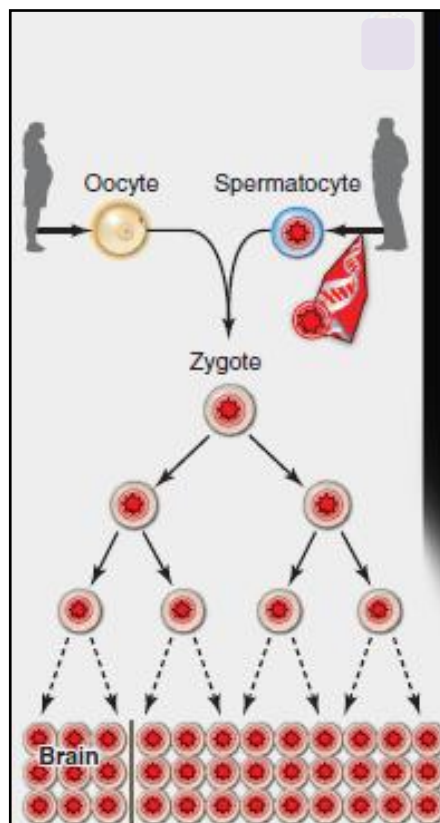


Figure 1.4: Demonstration of a pre-zygotic *de novo* mutation. The figure is taken from [A. Poduri, G. D. Evrony, X. Cai, C. A. Walsh, “Somatic mutation, genomic

variation, and neurological disease,” *Science*, vol. 341, pp. 1237758-1-8, 2013]. Reprinted with permission from AAAS.

A new mutational event can also occur in the genome of an individual after the fertilization. Herein, if the mutation occurs during the very early mitotic divisions of the zygote, only a small proportion of cells in different tissues carries the mutation (An early post-zygotic mutation) (Figure 1.5). When it occurs during the late mitotic divisions of the embryo (A late post-zygotic mutation), the mutation is present in a tissue-specific and mosaic fashion. In that case, only a small subset of cells in the tissue carries the variant (Figure 1.6).[4]

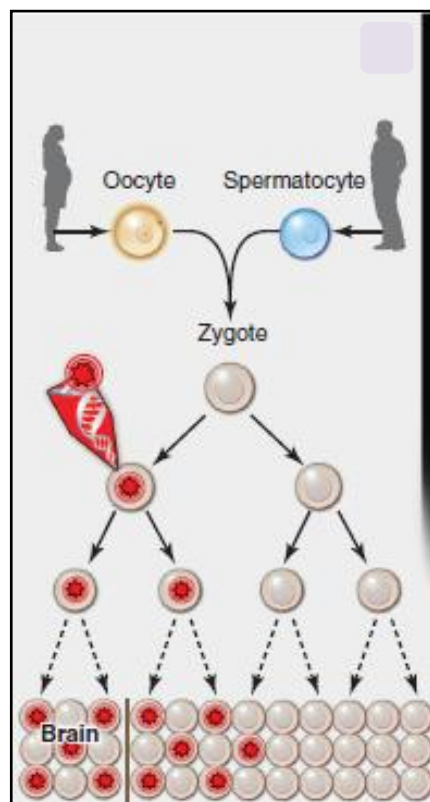


Figure 1.5: Demonstration of an early post-zygotic *de novo* mutation. The figure is taken from [A. Poduri, G. D. Evrony, X. Cai, C. A. Walsh, “Somatic mutation, genomic variation, and neurological disease,” *Science*, vol. 341, pp. 1237758-1-8, 2013]. Reprinted with permission from AAAS.

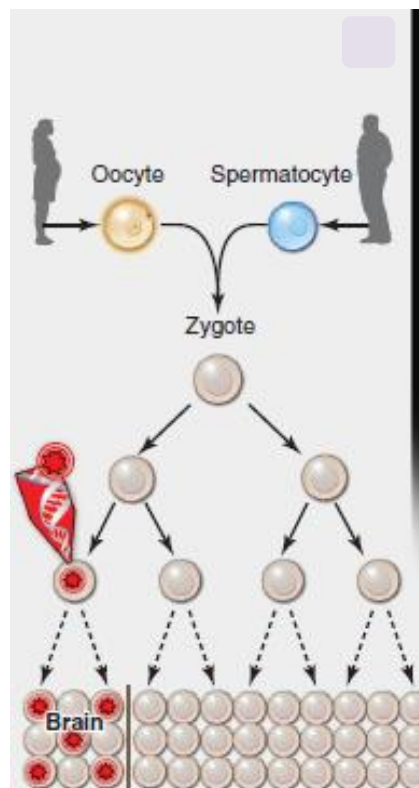


Figure 1.6: Demonstration of a late post-zygotic *de novo* mutation. The figure is taken from [A. Poduri, G. D. Evrony, X. Cai, C. A. Walsh, “Somatic mutation, genomic variation, and neurological disease,” *Science*, vol. 341, pp. 1237758-1-8, 2013]. Reprinted with permission from AAAS.

1.2.2 Genome-wide distribution patterns of the *de novo* mutations

Mutation occurs as a continuous and stochastic event. However, distribution of the new mutations throughout the genome is not random. The biases in the distribution of *de novo* mutations result from the intrinsic properties of the genome as well as the extrinsic factors.[8, 10]

De novo mutations tend to occur in the CpG dinucleotides 10 to 18 fold higher than in the non-CpG sites. The reason is that cytosine in CpG dinucleotides is subjected to methylation selectively. Frequent deamination of the 5-methylcytosine spontaneously in these sites results with the 5-methylcytosine to thymine mutation. Therefore *de novo* mutation rate in these sites increases.[11, 12, 13]

In addition to the GC content, various other patterns of the genome affect the distribution of the new mutations. For example, single base substitution density increases 53% during the replication of the DNA.[14] Transcription is another factor that induces the mutations. Herein, mutation density is high on the non-transcribed strand of the DNA. This is due to the transcription-coupled repair (TCR) machinery which corrects the mutations that occur in the transcribed strand.[15]

Moreover, it has been documented that emergence of a new mutation triggers several other nucleotide changes at the same time and these new mutations tend to cluster together both in coding and non-coding parts of the genome.[16]

Similar to the SNVs, *de novo* copy number variations (CNVs) are also distributed in a nonrandom fashion throughout the genome. There are repeats and segmental duplications in the human genome as a result of the mammalian genome evolution.[17] Herein, the frequency of large CNVs (>50 kbp) is high in the regions

where the directly oriented segmental duplications present and trigger nonallelic homologous recombination (NAHR).[18]

Non-random distribution of the new mutations results with the increased or decreased rates for different forms of mutations on the different sites of the genome.

1.2.3 Sex and age specific patterns of the *de novo* mutations

Sex and age specific patterns and thus effect of the parental origin of *de novo* mutations have been examined comprehensively. Effect of the parental origin means the tendency of *de novo* mutations to arise in the paternal or maternal germline preferentially.[19]

There is a discrepancy between males and females regarding the gametogenesis. Oogenesis is almost complete before birth. It has been calculated that twenty two cell divisions with twenty three replications are completed in the female germline. This process does not extend to the postnatal growth. Indeed, spermatogenesis is an ongoing process. During the reproductive life of man, sperm production occurs continuously, thus the number of mutations that occur during DNA replication in the male germline increases. Correspondingly, paternal bias for the *de novo* SNVs appears [20]; as fathers age increases, number of *de novo* mutations increases. This has been discovered by early genetic studies [21] and confirmed by various recent studies based on the whole genome and exome sequencing.[12,13]

Similar to the SNVs, paternal bias has also been observed for the mutations at microsatellites. Father to mother ratio was discovered to be 3.3:1.[22] Additionally, increased rate of *de novo* large chromosomal abnormalities and large CNVs that are greater than 150 kbp in the paternal germline has been documented.[23] On the other hand, trisomy of the chromosome 21 is an exception and risk for the majority of such aneuploidies increases with advanced maternal age.[24]

1.2.4 Role of the *de novo* mutations in diseases

Classical approach to understand the genetic basis of diseases has been the study of inherited mutations. However, an important proportion of the diseases including those that are common or rare as well as those that arise sporadically remained unsolved.[25]

In recent years, it has been anticipated that *de novo* mutations might contribute to pathogenesis of all diseases more significantly than we thought before. There are three rationales underlying this prospect: (a) There is an exponential human population growth and weak purifying selection [26, 27, 28], (b) *De novo* mutations are individually rare mutational events but have larger effect collectively, and (c) Selective pressures do not act on *de novo* mutations strictly, leading them to become more deleterious.[3, 25]

Involvement of the various forms of *de novo* mutations extending from the single base substitutions to the chromosomal abnormalities in both rare and common diseases has been documented. Herein, due to the feasibility of observing microscopically visible changes such as aneuploidies comparing to the other types of the mutations, roles of this class of *de novo* mutations have been determined relatively earlier. Most famous example is the *de novo* trisomy of the chromosome 21 which causes the Down syndrome.[29]

Advent of the microarray and sequencing technologies led to the examination of CNVs and SNVs in addition to the large chromosomal abnormalities. Recurrent CNVs have been involved in the pathogenesis of the malformation syndromes.[30, 31] In 2004, Vissers *et al.* identified a *de novo* CNV at chromosome 8q12 as a causal mutation for CHARGE syndrome.[32] This discovery was followed by the

identification of several other CNVs involved in the pathogenesis of several monogenic diseases.

Role of the *de novo* CNVs has also been established for the common diseases, especially for the neurodevelopmental disorders.[33] Large-scale *de novo* CNVs, those that are greater than 100 kbp in size, are present in 10% of the individuals affected with the sporadic neurodevelopmental diseases. Moreover, number of causative *de novo* CNVs is increased in individuals affected with mental retardation, intellectual disability, schizophrenia, and autism spectrum disorders.[34, 35, 36, 37]

In addition to the CNVs, *de novo* SNVs also account for an important fraction of the rare genetic diseases. Initial application of the exome sequencing to discover *de novo* mutations led to the discovery of *de novo* SNVs in *SETBP1* gene in patients affected with Schinzel-Giedion syndrome.[38] This was the first demonstration of the identification of a *de novo* mutation in a rare syndrome through exome sequencing. Subsequently, *de novo* SNVs involved in other clinical syndromes, including Kabuki syndrome, Bohring-Opitz syndrome, and KBG syndrome were identified.[39, 40, 41] Importantly, number of gene disrupting *de novo* SNVs in individuals affected with autism spectrum disorders discovered to be high comparing to the healthy controls.[42, 43, 44] These findings indicate the significant contribution of *de novo* SNVs to the disease pathogenesis in human.

1.3 Rates of the *De Novo* mutations

Human *de novo* mutation rate reflects the rate at which a new observable change in the DNA sequence occurs in each generation.[1] The studies concerned with the human mutation rate estimation trace their origin back to the 1930s.[21, 45] Since that time, several technological improvements were achieved and our knowledge about the rates for different classes of *de novo* genetic variation in human has increased.

1.3.1 Methods for estimating *de novo* mutation rate

1.3.1.1 Indirect method

Indirect estimation of the *de novo* mutation rate is the most previous approach which was pioneered by Haldane in 1935.[45] He developed a theory on the basis of the mutation-selection balance shaping the observed allele frequencies in a population. Haldane proposed that the observed allele frequencies in a population is due to the presence of a mutational pressure that balances the continuous selection acting on the deleterious alleles.[46] Based on this theory, the rate at which a new mutation arises can be estimated indirectly.

In addition to the Haldane's approach, human mutation rate might be estimated through counting the number of affected offspring whose parents are unaffected.[47] However, this kind of indirect approach may result with an underestimation, because disease gene mutations are not always result with disease phenotype.

Mutation rate per generation can also be estimated indirectly through the comparison of the nucleotide sequences between different species. This approach requires the knowledge of the divergence time and generation length per species.[48] To note that, the divergence-based measurement is based on the Kimura's neutral theory. The neutral theory implies that most of the polymorphisms and substitutions are neutral and for the neutral mutations the mutation rate is equal to the rate of evolution.[49] On the basis of this theory, rate of the new mutations can be estimated through counting the fixed differences between two closely related species. Herein, the estimated rate is not interfered with the false positives and somatic mutations. However, uncertainties about the size of the populations and divergence times might reduce the effectiveness and thus might result with an underestimation.[3]

1.3.1.2 Direct method

Direct estimation of the mutation rate is based on the examination of the nucleotide changes present in an individual's genome using whole genome and exome sequencing. It is achieved through the analysis of the sequencing data for the father-mother-offspring trios. Once the *de novo* mutations (those mutations that are present in the offspring and absent in the parents) were identified, the rate of the new mutations per generation can be calculated regarding the number of *de novo* variations and the target nucleotide size that is covered by sequencing.[3]

1.3.2 De novo mutation rates

1.3.2.1 De novo mutation rate based on the indirect approach

The most previous studies concerned with the estimation of human per generation mutation rate have relied on the homologous sequence comparisons as well as the screens of disease phenotypes, as explained above. These indirect estimations were restricted with the disease genes, pseudogenes or only specific classes of mutations (e.g. deleterious mutations) and did not survey the whole genome.

Initially, Haldane reported an indirect estimate of the human per locus mutation rate in his book, "The causes of evolution", in 1932.[50]. He proposed that in each generation at a rate of approximately 10^{-5} , new hemophilia mutations occur. Subsequent to this approximation, he addressed the mutation rate issue in a more comprehensive manner on the basis of his mutation-selection balance theory. In 1935, through considering the known frequency of the men affected with hemophilia in London (neglected in his 1932 study), he corrected his estimate of per locus mutation rate for the hemophilia gene as 2×10^{-5} .[45] These studies have been accepted as the first indirect estimates of the human per locus per generation mutation rate.

There is another approach which is based on the counting the children affected with dominant disorders when the parents are unaffected.[48, 51] This approach yielded rates ranging between 10^{-6} and 10^{-4} per locus per generation.[52]

The most reliable indirect estimates emerged from the comparison of the DNA sequences between human and a closely related species. This approach has its roots on the neutral theory. Mutation rate was estimated to be $1-2.5 \times 10^{-8}$ per locus per generation through comparing the pseudogenes and synonymous sites between human and chimpanzee.[48, 52, 53]

1.3.2.2 *De novo* mutation rate based on the direct approach

Recent improvements in the next generation sequencing technologies led to the investigation of the human genome comprehensively. Therefore, the rate at which new mutations occur has been investigated through whole genome and exome sequencing of the families. Accordingly, mutation rate per generation was documented in a large number of studies.

1.3.2.2.1 *De novo* germline mutation rate

Direct estimation of the human germline mutation rate has been achieved through sequencing the genomes of mother-father-offspring trios. Although these studies yielded the human *de novo* mutation rate per base per generation, the early post-zygotic mutations could not be distinguished from those that have a pre-zygotic origin. Therefore the reported mutation rates that are presumably accepted as germline mutation rate might reflect an overestimation.

1.3.2.2.1.1 *De novo* SNV rate

First direct estimation of the human *de novo* SNV rate is based on the analysis of Y chromosome sequences of two men separated by thirteen generations. Accordingly, Xue Y *et al.* reported a *de novo* SNV rate of 3×10^{-8} per base per generation.[54] In

2010, analysis of the whole genome sequence data of a quad family consisting of a sib-pair and their parents resulted with the estimation of the *de novo* SNV rate of approximately 1.10×10^{-8} per base per haploid genome.[55] Conrad and his colleagues sequenced the whole genomes of two parent-offspring trios and reported the *de novo* SNV rate as 0.97×10^{-8} and 1.17×10^{-8} for each trio. This study has great importance in terms of being the first direct report of the variation in mutation rates between families.[56] These initial attempts that are based on the calculation of *de novo* mutation rate per base per generation through whole genome sequencing of the nuclear families provided *de novo* SNV rates that are similar to the indirect estimations. Still, they are restricted in the scope of sequencing a single family or a specific chromosome.

In 2012, Kong A *et al* conducted a study to treat the *de novo* mutation rate issue in a more comprehensive manner through increasing the sample size. They performed deep whole genome sequencing of 78 Icelandic parent-offspring. They identified *de novo* SNVs and calculated the *de novo* SNV rate of 1.20×10^{-8} per base per generation.[12]

Rate of the *de novo* SNVs was also estimated using the advantage of the autozygous segments. Autozygous regions are homozygous regions that are inherited from a recent ancestor. Campbell CD *et al.* used autozygosity in the genomes of five parent-offspring Hutterite trios selected from a thirteen generation pedigree. Through deep whole genome sequencing of the genomes of fifteen individuals, they provided a *de novo* mutation rate estimate of 1.20×10^{-8} .[13]

1.3.2.2.1.2 *De novo* indel and CNV rate

Investigation of the genetic variations different than SNVs through whole genome sequencing is a relatively challenging process. The reason is the difficulty of mapping the short sequencing reads to the low complexity and repetitive regions where the insertions and deletions (indels) and CNVs are enriched.[3] Despite the

difficulties about the analysis of the CNVs and small indels, there are some studies that reports the mutation rate for these forms of genetic variation.

It has been reported that indels account for approximately 4% of the human spontaneous mutation.[57] Deletions are 2.3 - 4.1 times more common than the insertions. The rate of the deletions and insertions is 0.58×10^{-9} and 0.20×10^{-9} per site per generation, respectively.[58, 59] Despite the fact that indels seem to arise in lower frequencies comparing to the SNVs in the human genome, the estimated rates for both insertions and deletions may be under- or overestimation due to the difficulties in the analysis of these variants.

Initial studies concerned with the estimation of *de novo* CNV rate were restricted to the small number of loci. Based on the autosomal dominant genomic disorder data, rate of the CNVs was estimated to range between 2×10^{-5} and 1.25×10^{-4} per locus.[60] Itsara A *et al.* analyzed the *de novo* CNVs that are larger than 100 kbp in trios. They estimated the rate for large *de novo* CNVs as 1.2×10^{-2} per genome.[61] For the CNVs that span less than 100 kbp, the rate could not be calculated reliably.

1.3.2.2.2 Somatic mutation rate

Patterns and rates of somatic mutations have been investigated largely in the scope of cancer studies together with the *in vitro* cell models. The preliminary studies about the somatic mutation rate were restricted to the specific genes. For example, Iwama T indirectly estimated the rate of somatic mutations for adenomatous polyposis coli (*APC*) gene as ranging between 2×10^{-6} and 3×10^{-6} per stem cell per year.[62] Another study which was conducted later resulted with the estimation of the rate of *APC* somatic mutations as 10^{-5} per allele per year.[63] Advent of the sequencing technologies have led to the more comprehensive examination of the somatic mutations. In 2012, examination of the various tumor tissues from different cancer types through next generation sequencing resulted with the documentation of diverse

somatic mutation rates for different cancer types. Indeed, somatic mutation rate for the tumors was discovered to be average of 1.8 mutations per Mb.[64]

In previous studies, average rate of the somatic mutations in healthy individuals was estimated to be 7.7×10^{-10} per base per cell division.[59] More recently, in 2011, a study based on the whole genome sequencing of mother-father-offspring trios resulted with the estimation of non-germline *de novo* SNV rate of 2.52×10^{-7} . [56] However, this estimate is influenced by the cell-line derived mutations and did not reflect the true *de novo* somatic mutation rate. Another study reported the somatic *de novo* mutation rate for common SNPs in the genomes of healthy individuals as 1.2×10^{-7} per nucleotide through analysis of the monozygotic twins .[65] Despite the initial contributions of these studies to our knowledge about patterns and rates of somatic mutations, there is still a need for more comprehensive studies which will thoroughly assess the rate of somatic mutations including both early and late post-zygotic mutations in healthy individuals' genomes.

1.4 Mosaicism in the Human Genome

Mosaicism is the presence of more than one population of cells that have different genotypes in an individual.[66] Genome mosaicism can be classified into two broad categories: (a) germline mosaicism and (b) somatic mosaicism.[67] Germline mosaicism is the presence of the gonad in a heterogeneous state, comprised of germ cells with or without mutation. Somatic mosaicism is the presence of the heterogeneous populations of the non-germline cells in an individual.

Genome mosaicism plays an important role in the biological functions required for the normal development as well as in the pathogenesis of diseases. Hence, enlightening the mutational processes that lead to the both somatic and germline mosaicism has several biological and clinical implications.

Germline mosaicism has been found to contribute to the diseases that are inherited with an autosomal dominant inheritance pattern. In these diseases, the parents are phenotypically unaffected, leading to the wrong-anticipation of the autosomal recessive inheritance. Germline mosaic diseases are resulted from the mutations that occur during the gametogenesis in the parents and passed to the offspring.[67]

Somatic mosaicism arises via the new mutations that occur in the non-germline cells of an individual.[68] Initially, mutations occur sporadically in a precursor cell which gives rise to the different tissues. At this point, number of precursor cells that are mutated and thus the degree of mosaicism differs between different individuals. This variation is explained by the Luria-Delbruck probability distribution.[69]

Somatic mutations contribute to the pathogenesis of several diseases including Mendelian disorders, cancer and neurological diseases.[70] Moreover, there is a great evidence for the somatic mosaicism in several diseases such as Rett syndrome, Duchenne muscular dystrophy, hemophilia A, and Neurofibromatosis.[70, 71] It is important to note that, timing of the somatic mutations determines their tissue specific distribution patterns and the level of mosaicism represented in the individual's genome.[72] Hence, dissecting the origins of the new mutations that arise in the human genome is important to answer the question of “what time during the life cycle of an individual the new mutations mostly emerge?”. This will lead to the documentation of the somatic mosaicism and its level in both disease and health states.

1.5 Dissecting the Origins of New Mutations

Tracing the origins of *de novo* mutations is important. It provides insights into the timing of mutagenesis. Moreover, distinguishing the *de novo* mutations that arise mitotically from those that appear meiotically in the human genome might enable the

estimation of the real germline vs. somatic mutation rate as well as the degree of mosaicism.

It has been reported that there are two possible approaches to delineate the embryonic origins of the new mutations and thus timing of the mutagenesis.[3] First approach is based on the comparison of the genomes of monozygotic twins. Monozygotic twins are developed from a single zygote. Although monozygotic twins are expected to carry the same genetic material, genetic differences between them have been documented.[68] Herein, a *de novo* mutation which is shared by a monozygotic twin pair is anticipated as occurred during the parental gametogenesis meiotically or in the pre-twinning zygote. Otherwise, a *de novo* mutation which is present in the genome of one of the monozygotic twins is expected to arise post-zygotically during the early mitotic divisions of the embryo.[69] Presence of a post-zygotic mutation in all cells of one of the twins is interpreted as the mutation occurred most likely at the two-cell stage. If the post-zygotic mutation is present in a subset of tissues resulting with the somatic mosaicism, timing of the mutagenesis is accepted as at the four-cell stage or later.[6]

Second approach to trace the origins and timing of the new mutations is assessment of the genetic differences between the different tissues of the same individual.[72] This approach resulted with the documentation of genetic variation in different somatic tissues obtained from the post-mortem donors.[73, 74] However, these studies did not provide the rate of early post-zygotic mutations in healthy individuals, and their possible contribution to the overall mutation burden.

1.6 Importance of the Incidental Findings in the Next Generation Sequencing Data

Advent of the next generation sequencing has revolutionized the research in the field of medical genetics. During the past decades, both whole genome and exome

sequencing-based studies have identified the genes underlying several common and rare disease phenotypes and provided important clinical implications.[75, 76]

Massively parallel sequencing produces a tremendous amount of data in a cost- and time- effective manner, leading to the documentation of the genetic variation in human disease and health states. Besides the targeted and expected results, unintended results that are unrelated to the investigated problem yet have clinical importance are also present in the sequencing data [77, 78] For example, it is possible to identify all potentially pathogenic and damaging mutations in a study which is concerned with the discovery of a causal mutation for a specific disease through massively parallel sequencing. Herein, this kind of off-target (secondary) results that have clinical significance are defined as “incidental findings”.[79]

Examination of the incidental findings in the sequencing data is important for the assessment of potentially harmful mutations present in an individual’s genome. Therefore, it might enable the pre-symptomatic testing and preventive care. Additionally, these findings are important for the evaluation of the carrier status for recessive disorders as well as for the prediction of the late-onset diseases.[80] Despite the ethical issues, disclosure of the incidental findings that include the known and potential pathogenic variants derived from the sequencing studies that survey the whole genomes might have important implications for the human health.[81]

1.7 Aim and Strategy

Previous studies provided the rates of the *de novo* mutations in both healthy individuals and those that are affected with different disorders. Even in the mutation rate estimation studies that are based on the direct observation of the genetic variation in the genomes of healthy individuals through whole genome sequencing, contribution of the early post-zygotic mutations to the overall mutation burden could not be assessed comprehensively. The underlying reason is the study design which is

based on the observation of the genetic variation in parent-offspring trios. This kind of study design did not allow distinguishing the early post-zygotic mutations from those that arise during parental gametogenesis.

Despite the attempts to estimate the developmental timing of mutagenesis for some disease conditions, knowledge about the occurrence time and origin of the new mutations in the genomes of healthy individuals remained unknown. Hence, our aim is to estimate the *de novo* mutation rate in healthy individuals regarding the developmental timing of mutagenesis. At this point, we sought to determine how much of the *de novo* mutations occur meiotically during the parental gametogenesis (pre-zygotic) and how much of them arise mitotically during the initial divisions of the zygote (early post-zygotic) in the genome of a healthy individual.

To elucidate the origins of *de novo* mutations, we performed whole genome sequencing of a quad family consisting of a healthy male monozygotic twin pair and their parents. We restricted our *de novo* mutation analysis with the SNVs. Our strategy is based on the comparison of the genotypes of the twin pair and their parents for each of the SNVs concurrently to discover the *de novo* mutational events. We hypothesized that *de novo* mutations shared by the monozygotic twin pair have had parental or pre-twinning zygotic origin whereas those specific for only one of the twins have occurred in the post-twinning embryo post-zygotically. On the basis of this hypothesis, we characterized *de novo* SNVs and calculated the mutation rate regarding the timing of the new mutations.

In addition to our main target (*de novo* mutation rate estimation), we also intended to evaluate the incidental results present in the genomes of the twin pair and their parents. Herein, we performed additional analysis concerned with the examination of the genomes of the four individuals to document the disease predisposing and the potentially pathogenic variants.

Chapter 2

Materials and Methods

2.1 Subjects of the Study

A four-individual family consisting of a healthy 25-year-old male monozygotic twin pair, their 59-year-old father and 49-year-old mother participated in this study. We coded the samples as 12-020 (Twin I), 12-022 (Mother), 12-023 (Father), and 12-024 (Twin II). We evaluated the zygoty of the twin pair by comparing the SNVs of each twin using “vcf-compare” module of the vcf-tools (99% similar).[82] None of the four individuals participated in this study have a Mendelian disorder. They had no exposure to chemotherapeutics. The subjects were recruited to the control group of a movement disorder study which is approved by the institutional review boards (IRB) at Bilkent, Hacettepe, Başkent, and Çukurova Universities (decisions: BEK02, 28.08.2008; TBK08/4, 22.04.2008; KA07/47, 02.04.2007; and 21/3, 08.11.2005, respectively). The participants signed the written informed consent prepared according to the guidelines of the Ministry of Health in Türkiye before the study.

2.2 DNA Isolation

2.2.1 DNA isolation from the peripheral blood

We collected the peripheral blood samples from all participants in K3-EDTA containing BD Vacutainer® Blood Collection tubes (Becton Drive, NJ, USA) using venipuncture technique. We stored the blood samples in 1.5 ml microcentrifuge tubes at -80 C°.

We isolated the genomic DNA from 200 µl peripheral blood samples of each individual using the NucleoSpin® Blood (Macherey-Nagel) kit, following the manufacturer-provided protocols. To obtain the pure DNA with high concentration for whole genome sequencing, we repeated the washing steps two times and eluted the DNA using double-distilled water instead of the elution buffer included in the kit.

2.2.2 DNA isolation from the buccal wash specimen

We obtained the buccal wash specimen from the parents (12-022 and 12-023). Prior to the collection of the mouthwash, we asked the subjects to rinse their mouth with a 5 ml regular tap water for 30 seconds and collected the samples in the 50 ml falcon tubes.

We isolated the DNA from the mouthwash samples using DNeasy Tissue Kit (QIAGEN). We first centrifuged the mouthwash samples at 3000 rpm for 5 minutes and removed the supernatant to obtain the buccal cell pellet. The proceeding steps of the DNA isolation including the incubation, washing and elution steps were performed according to the manufacturer's protocol.

2.2.3 DNA isolation from the urine specimen

We collected the urine samples of the parents (12-022 and 12-023) in the sterile containers and isolated the DNA using the DNeasy Tissue Kit (QIAGEN) according to the manufacturer-recommended protocol.

2.2.4 Assessment of the quality and the quantity of the DNA samples

We measured the quality of the DNA samples by densitometry analysis through horizontal 1% agarose gel electrophoresis. We used Mass Ruler DNA Ladder (Sigma, MO, USA) as DNA marker (Figure 2.1). We evaluated the quantity and the purity of the DNA samples, by spectrophotometric analysis using NanoDropTM ND-1000 Spectrophotometer (NanoDrop Technologies, Inc, DE, USA).

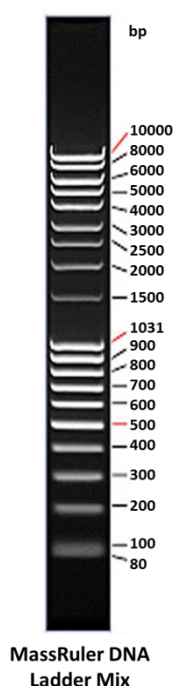


Figure 2.2: DNA marker used in the study. MassRuler DNA Ladder: 10 μ L per lane, 1% agarose gel, 1X TAE 7 V/cm, 45 minutes (Sigma, MO, USA)

2.3 Whole Genome Sequencing

Over the past few decades, there has been a remarkable revolution in sequencing technologies, leading to the capability of investigating genomes at a single nucleotide resolution level in a high throughput and cost-effective manner.[83] Following the emergence of the chain-termination method (Sanger sequencing) in 1977, and the completion of the first human genome draft sequence in 2001 [84, 85], new technologies have been developed and classified as the next generation sequencing. Through these technologies, sequencing the whole genome and exomes and generating a tremendous amount of data in an acceptable time frame has been possible.

There are several widely used platforms commercially available for next generation sequencing, including the 454 GS20 instrument (Roche Applied Science), Hiseq 2000 instrument (Illumina, Inc.), SOLiD instrument (Applied Biosystems), and Heliscope (Helicos, Inc.).[83] In this study, we used the Illumina Hiseq2000 platform to generate the whole genome sequencing data.

Illumina's sequencing technology is based on the "sequencing by synthesis" strategy. This strategy relies on the solid-phase bridge amplification of single-molecule DNA templates. In brief, single-stranded DNA fragments are attached to a flow cell through an adaptor. Subsequently, these fragments form a template for complementary strand synthesis by forming a bridge through hybridization to the complementary adaptors. Following the amplification, a large number of clusters are produced on the flow cell. The templates are sequenced in a massively parallel fashion by using reversible terminators labeled with the fluorescent colors and DNA polymerase. After the sequencing is complete, the generated sequence is determined through imaging.[86]

2.3.1 Sample preparation

For the paired-end whole genome sequencing, genomic DNA isolated from the whole blood samples of the four individuals was prepared using Illumina TruSeqTM DNA sample preparation kit according to the manufacturer's protocol.

2.3.2 Paired-end library construction

One library for each of the parents and two libraries for each of the twins were constructed from the whole blood derived DNA samples according to the Illumina-recommended protocols. For library construction, 1-3 µg of genomic DNA was fragmented by sonication. The resulting DNA fragments were end-repaired and adenosine overhangs were added. Adaptors were ligated to the end-repaired and adenosine overhang-added DNA fragments. For size selection, the resulting DNA fragments were run on 2% agarose gel and libraries of 400 bp in size were extracted from the gel using QIAGEN MinElute Gel Extraction Kit. The size selected libraries were enriched by quantitative PCR. The resulting libraries were purified for paired-end whole genome sequencing and their quality was assessed according to the Illumina's recommendations.

2.3.3 Paired-end whole genome sequencing

We performed deep whole genome sequencing using Illumina Hiseq 2000 instrument. Following the sequencing of the paired-end libraries, we performed imaging using Illumina SBS kits TruSeq V.3. We used Illumina's Real Time Analysis software V.1.13 with standard parameters for image analysis and base calling.

2.4 Analysis of the Whole Genome Sequencing Data

The common practice to analyze the whole genome sequencing data produced by the Illumina platform includes (a) conversion of the base calling data (.bcl files) into FASTQ files, (b) mapping of the FASTQ files to the reference genome, (c) formatting of the sequencing alignment mapping files for variants calling, (d) performing local indel realignment, (e) identifying the genomic variants (SNVs and indels), and (f) filtering the variants.[87]

2.4.1 Mapping of the sequencing reads

Prior to the mapping of the sequencing reads to the human reference genome, we converted the base calling data stored in the format of “.bcl files” into the FASTQ files using Illumina CASAVA V.1.8.2 software package. We mapped the paired-end sequencing reads to the NCBI Build 37 reference of the human genome using Burrows-Wheeler Aligner (BWA, version 0.6.1) with default parameters.[88] We converted the resulting Sequence Alignment/Map (SAM) files into the binary version (BAM) using SAMtools (version 0.1.18).[89] For each individual, we merged the BAM files in a single BAM file. We finally used SAMtools to mark and remove the PCR duplicates.

2.4.2 Discovery of the genomic variants and genotypes

We discovered the genomic variants including the SNVs and indels using Genome Analysis Toolkit software (GATK, version 1.6-13) with standard filtering parameters.[90] We used the RealignerTargetCreator and IndelRealigner modules of the GATK for indel realignment. This step is required to minimize the false calls due to the misalignment of the bases around indels. We identified the initial set of raw variants using multisample calling options of the GATK’s UnifiedGenotyper module. We applied the variant quality score recalibration to generate the final list of

variants of the four individuals. Finally, we performed the GATK-recommended hard filtering with “AB < 0.2 || MQ0 > 50 || SB > -0.10 || QUAL <10” parameters for low quality call-sets and generated the analysis-ready variants of the four individuals stored in a single variant call format (VCF) file.

2.4.3 Functional annotation of the variants

We performed the functional and positional annotation of the variants against human database GRCh37.68 using SnpEff.[91] SnpEff is a tool which is used for variant annotation and effect prediction. We used VCF files as input for the SnpEff and classified the variants according to their genomic context.

2.4.4 File formats

Whole genome sequencing using Illumina platform generates the data in the form of “.bcl” files. These files are converted to the FASTQ files using Illumina’s CASAVA software. FASTQ is a common file format which stores the sequencing data and the corresponding per base Phred quality scores. For each of the sequencing reads stored in the FASTQ file, there are four lines including the sequence identifier, sequence, quality score identifier and quality score. Alignment of the FASTQ files to the reference genome produces the SAM files. These files store the nucleotide alignment information and are converted to the corresponding binary version (BAM file) using SAMtools. BAM files are used as input for variants calling using GATK. The data produced by GATK is stored in VCF files. Each line in the tab-delimited VCF file contains the information for a single variant.

2.5 Identification of the Putative *De Novo* SNVs

We selected the SNVs (novel and dbSNP137 database-reported) from the multisample VCF file that includes all genomic variation (SNVs and indels) for the

four individuals (the twin pair and their parents) using ‘SelectVariants’ module of the GATK. We performed the downstream *de novo* mutation identification analysis using this SNV-only VCF file as an input.

2.5.1 Identification of the *de novo* SNV candidates shared by the twin pair

We classified the SNVs that are shared by the twin pair and not observed in the parental genomes as “group I” *de novo* SNVs. We estimated the timing of mutagenesis for this class of *de novo* SNVs as parental or post-zygotic yet prior to twinning. To identify the group I putative *de novo* SNVs, we compared the genotypes of the twin pair and their parents for each of the SNVs synchronously using SnpSift (version 1.1.3).[92] Herein, all possible genotype combinations, including those that could be assorted as genotyping errors (such as those SNVs for which one of the twins is homozygous although the variant is present in only one of the parents) were evaluated.

2.5.2 Identification of the twin specific *de novo* SNV candidates

We classified the SNVs present in one of the twins yet not present in the other twin and the parents as “group II” *de novo* SNVs. Timing of the mutagenesis for the group II *de novo* SNVs was considered as early post-zygotic in the post-twinning developmental period. To discover the group II putative *de novo* SNVs, we compared the genotypes of the twin pair and the parents for each of the SNVs concurrently using SnpSift (version 1.1.3).[92] The same strategy described for the group I *de novo* SNV candidates including the evaluation of the all genotype combinations was followed for the group II *de novo* SNVs to generate the initial list of variants.

2.5.3 Filtering the *de novo* SNV candidates

Subsequent to the identification of the group I and group II *de novo* SNV candidates, we applied a set of filters to the initial call-set progressively to distinguish the true *de novo* SNVs from the sequencing and variants calling artifacts. First, we filtered the putative *de novo* SNVs that are reported in the single nucleotide polymorphism (SNP) database dbSNP137 using BEDtools.[93] Following the selection of the novel putative *de novo* SNVs, we filtered the variants for which the genotype is missing or the read depth is smaller than three in at least one individual using a bash script. We filtered the remaining putative *de novo* SNVs regarding the genomic context using the IntersectBed tool included in the BEDtools suite. At this point, the variants intersecting the segmental duplications and simple repeats were removed. After the final filtering of *de novo* SNV candidates that are classified as genotyping errors, we selected the remaining variants that pass all filtering criteria for further analysis.

2.6 Evaluation of the Putative *De Novo* SNVs

Prior to the design of the validation experiments, we further assessed the validity of the putative *de novo* SNVs that pass the above-described filtering criteria based on the (a) visual examination of the sequence alignment data, and (b) reference allele frequency.

2.6.1 Visual inspection of the sequencing data

We visually analyzed the BAM files of the twin pair and their parents in a synchronous manner using Integrative Genomics Viewer (IGV).[94] The putative *de novo* SNVs included in the initial call-set were classified as (a) inherited, (b) no variant, (c) located on a low quality read and (d) *de novo* based on the visual analysis of the sequence alignment data. The first three groups were filtered and the variants classified as true *de novo* SNVs were included in the final call-set. Herein, the

putative *de novo* SNVs supported by at least three sequencing reads in one or both parents were classified as “inherited”. If a putative *de novo* SNV is not supported by three or more sequencing reads in the twins or is located on the “low-quality reads”, we classified it as “no variant”. We filtered the putative *de novo* SNVs as “located on a low quality read” in the presence of the high SNP and indel density on the variant supporting reads around the 40 bp flanking region. Remaining variants that pass the filtering process and thus classified as *de novo* were selected for validation by capillary sequencing.

2.6.2 Calculation of the reference allele frequency

Reference allele frequency (RAF) is the number of reference allele-supporting reads as a fraction of the total number of reads.[65] We determined the RAF threshold for the heterozygous SNVs as 0.5-0.7 in the twins and 0.95-1.0 in the parents.

2.7 Validation of the *De Novo* SNVs

2.7.1 Primer design

We designed the primer pairs (see Appendix A for the full list of primers) to test the high-confidence putative *de novo* SNVs using web-based Primer3 tool (<http://bioinfo.ut.ee/primer3-0.4.0/>).[95]

We evaluated the primer self-complementarity, GC content and the internal hairpin structures using PerlPrimer software.[96] We assessed the specificity of the primer pairs for the genomic region of interest using In Silico PCR (<http://genome.ucsc.edu/cgi-bin/hgPcr>) and BLAT (<http://genome.ucsc.edu/cgi-bin/hgBlat>) tools of the UCSC Genome Browser. We purchased the primers from Genmar Laboratories (Izmir, Turkiye) and Macrogen (Seoul Rep. of Korea).

2.7.2 PCR amplification

We amplified the targeted regions by conventional PCR using Techne TC512 thermal cycler (Bibby Scientific). PCR was performed in a total volume of 25 μ l, using 25 ng genomic DNA, One *Taq* DNA Polymerase (New England Biolabs), 5X One *Taq* standard reaction buffer, 25 mM dNTPs (Fermentas), and 10 pmol of each forward and reverse primers. We applied the following manufacturer-recommended thermal cycling conditions for the amplification: Initial denaturation at 94 °C for 30 seconds, 30-35 cycles of denaturation at 94 °C for 30 seconds, annealing at 58-62 °C for 40 seconds and elongation at 68 °C for 40 seconds and final elongation at 68 °C for 5 minutes.

For the regions that are difficult to amplify due to the complexity of the corresponding genomic region (presence of the repeats and the high GC content around the region flanking the variant of interest), we used Phusion High-Fidelity 2X Master Mix (Finnzymes, Finland) according to the manufacturer's protocol and thermal cycling conditions.

2.7.3 Visualization of the PCR products

We analyzed the PCR products through 0.8 % agarose gel electrophoresis. The agarose gel was prepared by dissolving the agarose (Basica LE, EU) in the 1X TAE buffer through heating. We added 30 ng/ml ethidium bromide (EtBr) to the gel as a fluorescence tag for visualization. We loaded the PCR products onto the agarose gel after the addition of 6X loading dye (MBI Fermentas, NY, USA).

The PCR products were run on the gel horizontally at 100-120 Volts for 20-40 minutes. We used pUCMix Marker 8 (MBI Fermentas, NY, USA) to infer the length of the amplicons (Figure 2.2). We visualized the PCR products using GelDoc imaging system (Bio-Rad, CA, USA) and captured the images by MultiAnalyst software version 1.1 (Bio-Rad, CA, USA).

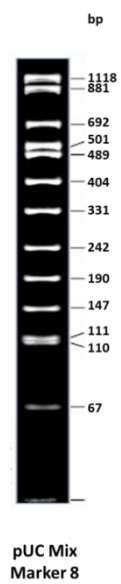


Figure 2.2: DNA marker used for the PCR product visualization. pUC Mix Marker 8: 0.5 μ g per lane, 1.7% agarose gel, 1X TvBE, 5 V/cm, 90 minutes (MBI Fermentas, NY, USA).

2.7.4 Sanger sequencing

PCR products were purified using MinEluteTM 96 UF PCR Purification Kit (Qiagen, MD, USA) prior to the Sanger based capillary sequencing. Sanger sequencing of the PCR products was performed using forward primer, reverse primer or both on ABI 3130 XL instrument (Applied Biosystems, Inc, CA, USA) by Refgen, Inc, (Ankara, Turkiye).

2.7.5 Analysis of the sequencing data

Sanger sequencing produced the data in the form of AB1 sequence trace format file. To analyze the data, we used CLC Main Workbench software package (CLCBio Inc). We aligned the Sanger sequencing data to the corresponding reference sequence. To download the reference sequence UCSC Table Browser (<https://genome.ucsc.edu/cgi-bin/hgTables?command=start>) was used.

2.8 Calculation of the *De Novo* SNV Rate

2.8.1 Formula

We estimated the *de novo* SNV rate for each of the twins based on the number of validated *de novo* SNVs, number of the callable bases and the false negative rate. The formula below was used to estimate the mutation rate:[13]

$$\mu = (N \times (1 + \text{FNR})) / L$$

In this formula, (a) μ represents the mutation rate, (b) N is the number of *de novo* SNVs validated by Sanger sequencing, (c) FNR is the false negative rate and (d) L is the number of callable bases.

2.8.2 Calculation of the number of callable bases

The number of the callable bases is represented with the “L” in the mutation rate calculation formula. We considered the bases with a minimum read depth of three and do not intersect the segmental duplications in the human genome as callable. The formula below was used to calculate the number of callable bases:

$$L = [N_{DP>3}] - [N_{SEGDUP_DP>3}]$$

$[N_{DP>3}]$ = Total number of the bases with read depth of minimum 3

$[N_{SEGDUP_DP>3}]$ = Total number of the segmental duplication intersecting bases with read depth of minimum 3

```
[NDP>3] = samtools view -b individual.PCRDuplicateRemoved.bam |  
genomeCoverageBed -ibam stdin -g b37.genome | grep genome | awk '{if($2 >= 3)  
sum += $3} END {print sum}'
```

```
[NSEGDUP_DP>3] = coverageBed -hist -abam individual.PCRDuplicateRemoved.bam -b  
~/segmental_duplications_WGAC_b37_nochr.bed | grep all | awk '{if($2 >= 3) sum  
+= $3} END {print sum}'
```

Segmental duplications file (segmental_duplications_WGAC_b37_nochr.bed) and the chromosomal length file (b37.genome) were obtained via the UCSC Table Browser (<https://genome.ucsc.edu/cgi-bin/hgTables?command=start>).

2.8.3 Calculation of the false negative rate

We calculated the false negative rate for the data according to the results of the Sanger based capillary sequencing of the putative *de novo* SNVs. The rate was calculated through dividing the number of putative *de novo* SNVs discovered to be

inherited from the parents through capillary sequencing to the total number of putative *de novo* SNVs tested by Sanger sequencing.

2.9 Analysis of the Incidental Results

We analyzed the whole genome sequencing data of the four individuals participated in this study to assess the (a) loss of function (LoF) mutations and (b) potentially damaging missense mutations. We focused on the true genetic variation through taking the advantage of the quad family sequencing. We identified all homozygous and heterozygous variants present in the genomes of the four individuals through comparative genotype analysis using SnpSift.[92] Herein, we considered the Mendelian transmission patterns for the variant selection. This analysis allowed us to select the true homozygous and heterozygous mutations and to reduce the false positives (genotyping and variants calling errors).

Subsequent to the identification of the homozygous and heterozygous mutations of the four individuals, we annotated the variants against human database (GRCh37.68 version in the SnpEff repository) using SnpEff.[91] We selected the LoF and missense mutations and filtered the remaining variants. We considered the (a) stop codon introducing (stopgain), (b) stop codon removing (stoploss), (c) splice site disrupting SNVs and (d) frameshift indels as LoF mutations.[97]

Of the LoF and missense variants, we focused on the rare novel variants and the variants that are reported in SNP database dbSNP137 with a low minor allele frequency (< 0.01). We filtered those variants that are reported in dbSNP137 with minor allele frequency of greater than 0.01 using a bash script.

We assessed the deleteriousness of the missense variants using SIFT, Polyphen2 and MutationTaster.[98, 99, 100] We evaluated the genes that carry the selected variants (both LoF and missense) in terms of being reported for a disease phenotype in OMIM database. Herein, we used the Biomart data mining tool of the Ensembl for disease gene analysis. We classified the homozygous and heterozygous

LoF and damaging missense mutations into two categories: (a) the variants located on the genes that are reported to be causal for a recessive or dominant disease in OMIM database and (b) the variants located on the genes that do not have any clinical significance. On the basis of these findings, we evaluated the carrier status of each individual for recessive diseases as well as the possible presence of late-onset dominant diseases.

2.10 Enzymes, Chemicals and Reagents

2.10.1 Enzymes

Table 2.1: List of the enzymes used in the experiments

Enzyme	Company
Proteinase K	Appligene, CA, USA
OneTaq® DNA Polymerase	New England Biolabs, UK
OneTaq® Hot Start	New England Biolabs, UK
Phusion Hot Start II High-Fidelity	Finnzymes, Finland

2.10.2 Chemicals and reagents

Table 2.2: List of the chemicals and reagents used in the experiments

Chemicals and Reagents	Company
Acetic acid	Sigma, MO, USA
Agarose	Basica LE, EU
Bromophenol blue	Sigma, MO, USA
dNTPs	Fermentas, NY, USA
EDTA	Fermentas, NY, USA
Ethanol	Merck, Germany
Ethidium bromide	Sigma, MO, USA
Tris-Base	Bio-Rad, CA, USA
Tris-HCl	Sigma, MO, USA
Xylene Cyanol	Sigma, MO, USA

2.10.3 Standard solutions and buffers

Table 2.3: List of the standard solutions and buffers used in the experiments

Solutions and Buffer	Content
Ethidium bromide:	10 mg/ml in water (stock solution) 30 ng/ml (working solution)
Agarose gel loading buffer (6X):	15% coll 0.05% bromophenol 0.05% xylene cyanol
1X TAE (Tris-acetic acid-EDTA):	40mM Tris-acetate, 2 nM EDTA pH 8.0

2.11 Web Sources

Table 2.4: Web sources used in the study design and data analysis

Description	Web address
Figure B.1	donut.cs.bilkent.edu.tr/twinstudy/
1000 Genomes	browser.1000genomes.org
NCBI dbSNPs	www.ncbi.nlm.nih.gov/projects/SNP/
UCSC Genome Browser	genome.ucsc.edu
UCSC Tables	genome.ucsc.edu/cgi-bin/hgTables?command=start
NCBI	www.ncbi.nlm.nih.gov/
BWA	bio-bwa.sourceforge.net/
BioMart	www.ensembl.org/info/data/biomart.html
SAMtools	samtools.sourceforge.net/
GATK	www.broadinstitute.org/gatk/
GATK bundle	ftp.broadinstitute.org/bundle/2.5/
HomozygosityMapper	http://www.homozygositymapper.org/
IGV	www.broadinstitute.org/igv/
Snpeff	snpeff.sourceforge.net/
Snpsift	snpeff.sourceforge.net/SnpSift.html
BEDTools	code.google.com/p/bedtools/
Primer3	bioinfo.ut.ee/primer3-0.4.0/
PerlPrimer	perlprimer.sourceforge.net/
BLAT	genome.ucsc.edu/cgi-bin/hgBlat
In-Slico PCR	genome.ucsc.edu/cgi-bin/hgPcr

Chapter 3

Results

3.1 Whole Genome Sequencing

We performed paired-end whole genome sequencing of a healthy monozygotic twin pair and their parents (Figure 3.1).

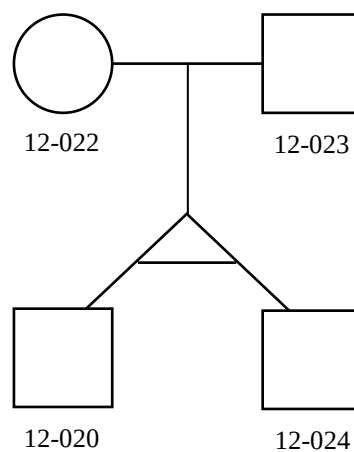


Figure 3.1: Pedigree of the quad family

3.1.1 Sample preparation and library construction

We isolated the genomic DNA from the whole blood samples of the twin pair and their parents using NucleoSpin® Blood (Macherey-Nagel) kit according to the manufacturer's protocol. We evaluated the quality and quantity of the DNA samples by densitometry analysis through 1% agarose gel electrophoresis (Figure 3.2) and spectrophotometric measurement (Table 3.1).

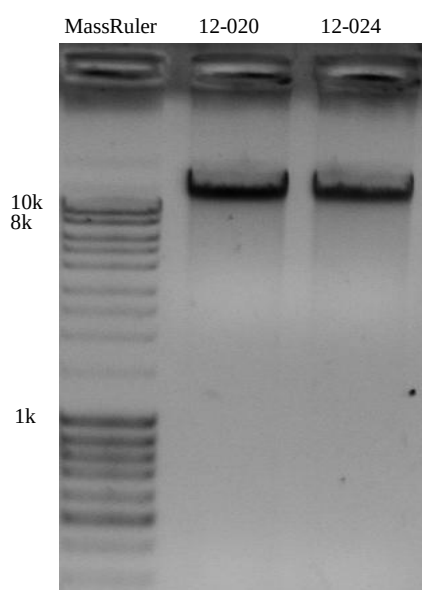


Figure 3.2: Densitometry analysis of the DNA samples using agarose gel electrophoresis. DNA samples were run horizontally on 1% agarose gel at 90V for 50 minutes. Lane 1: MassRuler DNA Ladder Mix (1 μ l); Lane 2: DNA sample of the 12-020; Lane3: DNA sample of the 12-024. Gel image was captured with BioRad Gel Doc 2000 system.

Table 3.1: Spectrophotometric measurement of the concentrations of DNA samples

Sample	Concentration (ng/μl)	260/280	260/230
Twin I	187.24	1.89	2.22
Mother	111.55	1.85	2.14
Father	108.30	1.84	2.04
Twin II	206.87	1.88	2.22

Abbreviations used in this table: ng, nano gram; μl, microliters

Whole blood derived DNA samples were prepared for the paired-end whole genome sequencing using Illumina TruSeqTM DNA sample preparation kit. We constructed one library for each of the parents and two libraries for each of the twins. Prior to the sequencing, we assessed the qualities of the libraries and performed the purification regarding the Illumina's recommendations.

3.1.2 Paired-end whole genome sequencing

We performed deep whole genome sequencing of the twin pair and their parents using Illumina Hiseq 2000 instrument. We sequenced the paired-end libraries on eight lanes for each of the twins and four lanes for each of the parents, generating the 101- and 104-bp sequencing reads.

3.1.3 Whole genome sequencing data analysis: Mapping and variants calling

We generated a total of 538 Gb genomic sequence for the four individuals with a mean coverage depth of 59.96 fold for twin I, 42.51 fold for twin II, 34.10 fold for mother, and 33.42 fold for father. We mapped the paired-end sequencing reads to the human reference genome (version GRCh37) using BWA.[88] At this point, 85.59 – 93.01 % of the reads were mapped to the reference genome successfully (Table 3.2). 99% of the whole genome was covered by at least two reads (Table 3.3).

We discovered the genomic variants (SNVs and indels) using GATK [90] according to the GATK-recommended pipeline.[87] In brief, we marked and removed the PCR duplicates using SAMtools [89] and performed local indel realignment using RealignerTargetCreator and IndelRealigner modules of the GATK. We discovered the variants using multisample calling options of the GATK's UnifiedGenotyper based on the 170 fold total coverage. Following the application of variant quality score recalibration (VQSR), we identified a total of 6,953,274 raw genomic variants for the four individuals.

We performed hard filtering to further evaluate the raw variants based on the strand bias and mapping quality and removed the very low quality calls with Phred quality score of maximum 10. After the hard filtering, we identified approximately 3,400,000 SNVs and 750,000 small indels for each of the four individuals. We calculated the transition to transversion (Ts/Tv) ratio using vcf-tools.[82] The ratio was discovered to be 2.12 per genome consistent with the expected ratio for the whole genome sequencing data (Table 3.4).

Table 3.2: Statistics on the genome sequence produced by the paired-end whole genome sequencing

Sample	Twin I	Twin II	Mother	Father
Total sequence^a	186 Gb	138 Gb	106 Gb	108 Gb
Mapped reads^b	1,667,768,752	1,154,518,105	950,095,509	930,901,596
Percentage of total reads^c	92.10	85.59	93.01	89.35
Target base coverage (%)^d	99.99	99.90	99.35	99.94
Mean coverage (fold)^e	59.96	42.51	34.10	33.42

^aLength of the sequence produced by WGS based on the total number of reads. ^bNumber of sequencing reads mapped with BWA. ^cPercentage of the mapped reads among total reads. ^dPercentage of the bases covered by at least one aligned read. ^eEffective fold coverage of the whole genome

Abbreviations used in this table: Gb, gigabase

Table 3.3: Percentage of coverage per genome

Depth (fold)	Twin I	Twin II	Mother	Father
0	0,001	0,001	0,65	0,06
≥ 2	99,930	99,857	99,27	99,87
≥ 4	99,900	99,752	99,13	99,72
≥ 10	99,590	99,133	98,07	98,53
≥ 15	99,216	97,605	95,46	95,30
≥ 20	98,419	94,153	90,32	88,83
≥ 30	94,053	80,050	68,06	64,77
≥ 40	85,258	53,518	34,58	32,07

Table 3.4: Statistics on all variants identified through whole genome sequencing

Sample	All variants		Ts/Tv	Novel Variants	
	SNVs	Indels		SNVs	Indels
Twin I	3,496,649	783,415	2.12	49,800	70,500
Twin II	3,493,010	781,583	2.12	49,182	69,927
Mother	3,468,324	754,789	2.11	50,340	64,396
Father	3,464,691	749,592	2.12	48,971	63,853

Abbreviations used in this table: SNV, single nucleotide variation; Indel, insertions and deletions; Ts, transition; Tv, transversion

3.2 Discovery of the Putative *De Novo* SNVs

We restricted our *de novo* mutation analysis with the SNVs. Our strategy to discover the *de novo* SNVs is based on the comparative analysis of the genotypes of the four individuals (twin pair and the parents) for each of the SNVs concurrently followed by a selective filtering. The discovery and filtering steps are described in the “Materials and Methods” section in detail.

3.2.1 Discovery of the putative *de novo* SNVs shared by the twin pair

We compared the genotypes of the twin pair and their parents through evaluating all possible genotype combinations for each of the SNVs synchronously using SnpSift.[92] We initially identified a total of 8,994 putative *de novo* SNVs shared by the twin pair. To discriminate the true *de novo* SNVs from the sequencing and variants calling artifacts, using a combination of sequence data analysis software and tools described in the methods section, we performed a progressive filtering based on the (a) novelty of the variants vis-à-vis dbSNP137, (b) presence of the coverage depth of minimum three for the site encompassing the variant, (c) presence of the variants on the sites that do not intersect the segmental duplications and simple repeats and (d) non-appearance of the variants as genotyping errors. Subsequent to this filtering which step-by-step removes the erroneous *de novo* SNV calls, we identified a total of 159 high-confidence putative *de novo* SNVs shared by the twin pair (Table 3.5).

Table 3.5: Identification of the putative *de novo* SNVs shared by the twin pair and application of the filters

Genotypes ^a				SNVs		Number of <i>de novo</i> SNVs after filtering		
Twin I	Twin II	Mother	Father	All ^b	Novel ^c	Read Depth ^d	Genomic Context ^e	Genotyping Error
1/1	1/1	0/0	0/0	576	105	24	0	No
0/1	0/1	0/0	0/0	3,330	662	534	159	No
1/1	0/1	0/0	0/0	86	15	2	1	Yes
0/1	1/1	0/0	0/0	200	35	4	0	Yes
1/1	1/1	0/1	0/0	822	40	20	5	Yes
1/1	1/1	0/0	0/1	708	23	11	2	Yes
1/1	1/1	0/0	1/1	1,231	64	3	1	Yes
1/1	1/1	1/1	0/0	2,041	54	8	0	Yes
Total				8,994	998	606	168	159

^aComparison of the genotypes simultaneously; 1 and 0 refer alteration and reference, respectively. ^bTotal number of *de novo* SNV candidates before applying filters. ^cSNVs reported in dbSNP137 were filtered. ^dSNVs with read depth of less than 3 were filtered. ^eSNVs intersecting segmental duplications and simple repeats were filtered

Abbreviations used in this table: SNV, single nucleotide variation

3.2.2 Discovery of the twin-specific putative *de novo* SNVs

We discovered the putative *de novo* SNVs that are specific to each of the twins using the strategy described for the discovery of *de novo* SNVs shared by the twin pair above. We first identified a total of 1,417 putative *de novo* SNVs specific to twin I and 1,852 putative *de novo* SNVs specific to twin II through comparing the genotypes of the twin pair and the parents using SnpSift [92]. After the application of the filters progressively, a total of 131 high confidence putative *de novo* SNVs were identified. Of these, 83 were specific for twin I (Table 3.6) and 48 were specific for twin II (Table 3.7).

Table 3.6: Identification of the putative *de novo* SNVs specific to twin I and application of the filters

Genotypes ^a				SNVs		Number of <i>de novo</i> SNVs after filtering		
Twin I	Twin II	Mother	Father	All ^b	Novel ^c	Read Depth ^d	Genomic Context ^e	Genotyping Error
1/1	0/0	0/0	0/0	138	36	1	1	No
0/1	0/0	0/0	0/0	919	322	299	82	No
1/1	0/0	0/1	0/0	61	9	2	0	Yes
1/1	0/1	0/1	0/0	71	8	1	0	Yes
1/1	0/0	0/0	0/1	22	6	0	0	Yes
1/1	0/1	0/0	0/1	60	4	2	1	Yes
1/1	0/1	0/0	1/1	60	7	2	0	Yes
1/1	0/1	1/1	0/0	86	6	2	1	Yes
Total				1,417	398	309	85	83

^aComparison of the genotypes simultaneously; 1 and 0 refer alteration and reference, respectively. ^bTotal number of *de novo* SNV candidates before applying filters. ^cSNVs reported in dbSNP137 were filtered. ^dSNVs with read depth of less than 3 were filtered. ^eSNVs intersecting segmental duplications and simple repeats were filtered

Abbreviations used in this table: SNV, single nucleotide variation

Table 3.7: Identification of the putative *de novo* SNVs specific to twin II and application of the filters

Genotypes ^a				SNVs		Number of <i>de novo</i> SNVs after filtering		
Twin I	Twin II	Mother	Father	All ^b	Novel ^c	Read Depth ^d	Genomic Context ^e	Genotyping Error
0/0	1/1	0/0	0/0	49	14	1	0	No
0/0	0/1	0/0	0/0	791	270	239	48	No
0/0	1/1	0/1	0/0	29	1	1	0	Yes
0/1	1/1	0/1	0/0	306	32	20	12	Yes
0/0	1/1	0/0	0/1	13	2	1	1	Yes
0/1	1/1	0/0	0/1	249	32	19	8	Yes
0/1	1/1	0/0	1/1	197	10	3	1	Yes
0/1	1/1	1/1	0/0	218	13	1	1	Yes
Total				1,852	374	285	71	48

^aComparison of the genotypes simultaneously; 1 and 0 refer alteration and reference, respectively. ^bTotal number of *de novo* SNV candidates before applying filters. ^cSNVs reported in dbSNP137 were filtered. ^dSNVs with read depth of less than 3 were filtered. ^eSNVs intersecting segmental duplications and simple repeats were filtered

Abbreviations used in this table: SNV, single nucleotide variation

3.2.3 Assessment of the validity of filtering process

We evaluated the efficacy of our filters in the removal of genotyping and variants calling errors while keeping the true *de novo* SNV candidates in the call-set. We performed this evaluation based on the “true genetic variation” that is present in the genomes of the four individuals.[101]

We identified a total of 3,437,507 SNVs that are present in the genomes of the twin pair and in both parents (Table 3.8) and classified these variants as control group SNVs. We assessed the validity of the homozygous and heterozygous control group SNVs based on the reference allele frequency (RAF). RAF calculated to be 0.01 for homozygous variants and 0.5 for heterozygous variants for each individual, suggesting that the group of SNVs selected as controls represented the true genetic variation (Table 3.9).

We calculated the ratio of “fraction of putative *de novo* SNVs removed with filters” to “fraction of control group SNVs removed with filters” ($F_{\text{pdn}}/F_{\text{control}}$) for the putative *de novo* SNVs (a) shared by the twin pair, (b) specific to twin I and specific to twin II (Table 3.10). For each group of the putative *de novo* SNVs, the filters removed more *de novo* SNV candidates than the control group SNVs. This suggested the efficiency of the filters in the removal of genotyping/variants calling errors.

Table 3.8: Identification of the control group SNVs and application of the filters

Genotypes^a				SNVs	Number of SNVs after filtering
Twin I	Twin II	Mother	Father	All^b	Read Depth^c
0/1	0/1	0/1	0/1	350,297	350,196
0/1	0/1	0/1	0/0	477,590	477,181
0/1	0/1	0/0	0/1	498,992	498,545
0/1	0/1	0/1	1/1	180,920	180,717
0/1	0/1	1/1	0/1	190,813	190,701
0/1	0/1	0/0	1/1	201,997	201,587
0/1	0/1	1/1	0/0	213,447	213,048
1/1	1/1	0/1	0/1	168,375	168,321
1/1	1/1	0/1	1/1	191,449	191,171
1/1	1/1	1/1	0/1	192,291	192,141
1/1	1/1	1/1	1/1	775,330	773,899
Total				3,441,501	3,437,507

^aComparison of the genotypes simultaneously; 1 and 0 refer alteration and reference, respectively. ^bTotal number of control group SNVs before applying filters. ^cSNVs with read depth of less than 3 were filtered.

Abbreviations used in this table: SNV, single nucleotide variation

Table 3.9: Calculation of the RAF and average depth of coverage for the control group SNVs

Class of SNVs	Twin I		Twin II		Mother		Father	
	Average RAF	Average Depth	Average RAF	Average Depth	Average RAF	Average Depth	Average RAF	Average Depth
Control group SNVs (n=3,437,507)								
Homozygous in both twins (n=1,325,532)	0.02 (0.1 std.)	57.15	0.01 (0.1 std.)	38.73	0.02 (0.1 std.)	32.94	0.02 (0.1 std.)	32.21
Heterozygous in both twins (n=2,111,975)	0.5 (0.1 std.)		0.5 (0.1 std.)		0.5 (0.1 std.)		0.5 (0.1 std.)	

Abbreviations used in this table: SNV, single nucleotide variation; RAF, reference allele frequency; std, standard deviation

Table 3.10: Evaluation of the efficiency of the filters in the removal of false *de novo* calls

Type of filter	Control group SNVs	Shared <i>de novo</i> SNVs		Twin I-specific <i>de novo</i> SNVs		Twin II-specific <i>de novo</i> SNVs	
	Percentage	Percentage	$F_{\text{pdn}}/F_{\text{control}}$	Percentage	$F_{\text{pdn}}/F_{\text{control}}$	Percentage	$F_{\text{pdn}}/F_{\text{control}}$
No Filter	3,441,501 (100%)	998 (100%)	-	398 (100%)	-	374 (100%)	-
Read Depth Filter	3,437,507 (99.88%)	606 (60.72%)	327	309 (77%)	191	285 (76.20%)	198
Genomic Context Filter	N/A	168(16.83%)	N/A	85 (21.35%)	N/A	71 (18.98%)	N/A

Abbreviations used in this table: SNV, single nucleotide variation; F_{pdn} , fraction of putative *de novo* SNVs removed by filters; F_{control} , fraction of control group SNVs removed by filters; N/A, “Genomic Context Filter” was not applied to the control group SNVs; Percentage, number and percentage of the variants remained after filtering

3.3 Selection of the Putative *De Novo* SNVs for Sanger Sequencing

We identified a total of 290 high-confidence putative *de novo* SNVs (159 shared by the twin pair, 83 specific to twin I and 48 specific to twin II) as a result of the genotype comparison and the progressive filtering. Herein, to further clean the spurious *de novo* SNV calls prior to the Sanger based capillary sequencing; we evaluated the *de novo* SNV candidates based on the visual examination of the sequencing data and the sequence coverage depth.

3.3.1 Visual examination-based evaluation and selection

We performed visual examination of the BAM files of the four individuals (the twin pair and the parents) concurrently for the total of 290 high-confidence *de novo* SNV candidates using IGV (see Appendix B).[94] We filtered 40/290 variants as “inherited”; 8/290 variants as “no variant”; and 122/290 variants as “located on a low quality read”, based on the visual inspection and the analysis of the allelic depth for each of the SNVs in each individual (Table 3.11). We selected the remaining 120 *de novo* SNV candidates (73 shared by the twin pair; 29 specific to twin I and 18 specific to twin II) as true putative *de novo* SNVs (see Appendix C for the full list of the total of 290 putative *de novo* SNVs and the predictions for each of the variants). This set of putative *de novo* SNVs were selected for validation. The criteria underlying the classification and filtering of the variants according to the visual inspection are described in the “Materials and Methods” section comprehensively.

Table 3.11: Classification of the putative *de novo* SNVs based on the visual inspection of the sequence alignment data

Class of <i>de novo</i> SNVs	Initial call-set	Inherited	No variant	Low quality read	Final call-set
Shared by the twin pair	159	26	5	55	73
Twin I-specific	83	8	3	43	29
Twin II-specific	48	6	0	24	18
Total	290	40	8	122	120

Abbreviations used in this table: SNV, single nucleotide variation

3.3.2 Sequencing data-based evaluation of the putative *de novo* SNVs

Subsequent to the visual inspection, we further assessed the validity of the putative *de novo* SNVs based on the RAF and the depth of coverage. We calculated the average RAF and coverage depth in each of the four individuals for the *de novo* SNV candidates that are shared by the twin pair (group I) and that are specific for each of the twins (group II) (Table 3.12). At this point, the RAF for the mother and father was discovered to range between 0.95 and 0.97 for each class of the putative *de novo* SNVs, supporting the absence of the variant in the parental genomes. For the putative *de novo* SNVs that are shared by the twin pair, the RAF values of 0.59 for the twin I and 0.55 for the twin II supported the heterozygous presence of the variant. We calculated the RAF of 0.68 for the twin I specific and 0.69 for the twin II specific *de novo* SNV candidates. These frequencies for the twin specific variants suggested the presence of the putative *de novo* SNVs that have an early post-zygotic origin in a mosaic fashion.

Table 3.12: Average RAF and the depth of coverage for each class of the *de novo* SNV candidates

Class of putative <i>de novo</i> SNVs	Twin I		Twin II		Mother		Father	
	Average RAF	Average Depth	Average RAF	Average Depth	Average RAF	Average Depth	Average RAF	Average Depth
<i>De novo</i> SNVs shared by the twin pair	0.59 (0.1 std.)	38.72	0.55 (0.1 std.)	24.89	0.96 (0.1 std.)	23.55	0.95 (0.1 std.)	21.01
Twin I-specific <i>de novo</i> SNVs	0.68 (0.1 std.)	36.65	0.95 (0.1 std.)	26.79	0.96 (0.1 std.)	23.24	0.95 (0.1 std.)	18.96
Twin II-specific <i>de novo</i> SNVs	0.96 (0.05 std.)	38.50	0.69 (0.1std.)	21.39	0.97 (0.04 std.)	25.50	0.95 (0.1 std.)	20.55

Abbreviations used in this table: SNV, single nucleotide variation; RAF, reference allele frequency; std, standard deviation

3.4 Classification of the Putative *De Novo* SNVs Regarding the Genomic Context

We performed functional and positional annotation of the total of 290 putative *de novo* SNVs (159 shared by the twin pair, 83 specific to twin I and 48 specific to twin II) against human database (GRCh37.68 version in the SnpEff repository) using SnpEff (see Appendix C).[91]

We classified the total of 120 putative *de novo* SNVs that were selected for Sanger based capillary sequencing as located on the (a) downstream of the genes (n=25), (b) upstream of the genes (n=1), (c) intergenic regions (n=32), (d) introns (n=58), and (e) exons (n=4) (Figure 3.3). Herein, the major proportion of the putative *de novo* SNVs was discovered to present in the non-coding regions of the genome. We did not observe any gene that is subjected to the *de novo* SNVs recurrently. The biased distribution of the putative *de novo* SNVs towards the non-coding regions was attributed to the (a) presence of the less forceful evolutionary pressure on the non-coding regions of the genome, (b) structure of the human genome (98% non-coding), and (c) our study design which is concerned with the analysis of the genomes of healthy individuals.[9]

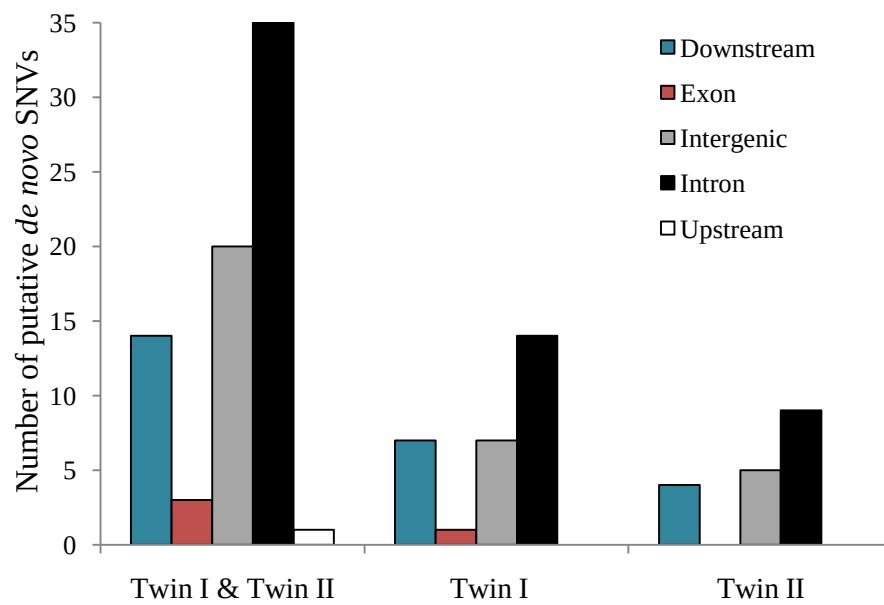


Figure 3.3: Classification of the putative *de novo* SNVs regarding the genomic context. Annotation was performed against human database GRCh37.68 using SnpEff.[91]

3.5 Experimental Validation of the *De Novo* SNVs

We sought to validate the total of 120 high-confidence *de novo* SNV candidates (73 shared, 29 twin I-specific, 18 twin II-specific) that pass our filtering criteria through Sanger sequencing. We designed the primer pairs using Primer3 (see Appendix A for the full list of primers).[95] Subsequent to the PCR amplification and the purification, we sequenced the blood-derived DNA samples of the twin pair and their parents for 83 of the 120 putative *de novo* SNVs and analyzed the sequencing data of the four individuals simultaneously using CLC Main Workbench software package (CLCBio Inc). We could not test 37 *de novo* SNV candidates due to unavailability of the primer pairs or PCR products of good quality for Sanger sequencing.

3.5.1 Validation of the *de novo* SNVs shared by the twin pair

We performed Sanger based capillary sequencing of the whole blood derived DNA samples of the twin pair and their parents to test the 49/73 high confidence putative *de novo* SNVs. 24/73 *de novo* SNV candidates could not be tested due to the limitations described above.

Of the 49 *de novo* SNV candidates that are shared by the twin pair and tested by Sanger sequencing, 23 confirmed to be true *de novo* SNV present in a heterozygous state in both twins (Figure 3.4). 8/49 discovered to be inherited from the parents (Figure 3.5) and 11/49 discovered to be not present in the genomes of the twin pair (Figure 3.6). The sequencing results were uninformative for 7/49 *de novo* SNV candidates (see Table C.1 for the Sanger sequencing results for each of the tested variants and Appendix D for the Sanger sequencing chromatograms of the 23 validated *de novo* SNVs).

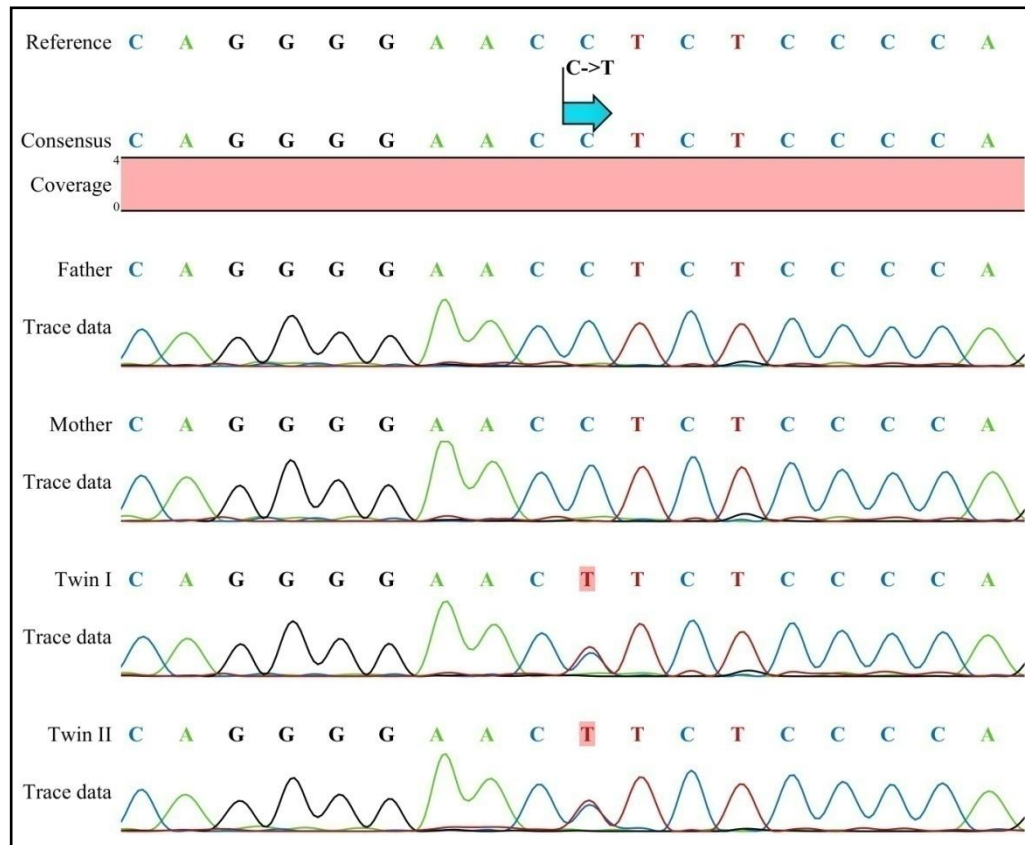


Figure 3.4: A representative Sanger sequencing electropherogram for a shared *de novo* SNV present in a heterozygous state and validated by capillary sequencing. (Chromosome 4: 145566227: C/T).

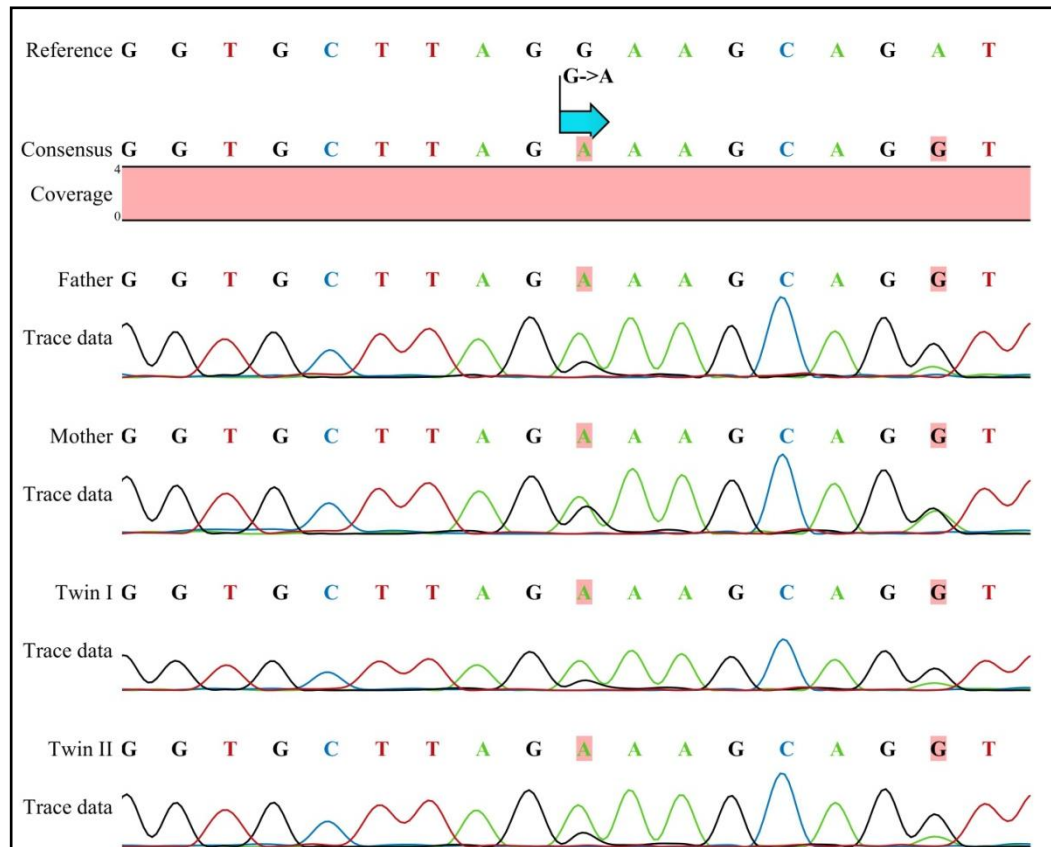


Figure 3.5: A representative Sanger sequencing electropherogram for a putative *de novo* SNV discovered to be inherited based on the capillary sequencing result (Chromosome 6: 169662202: G/A).

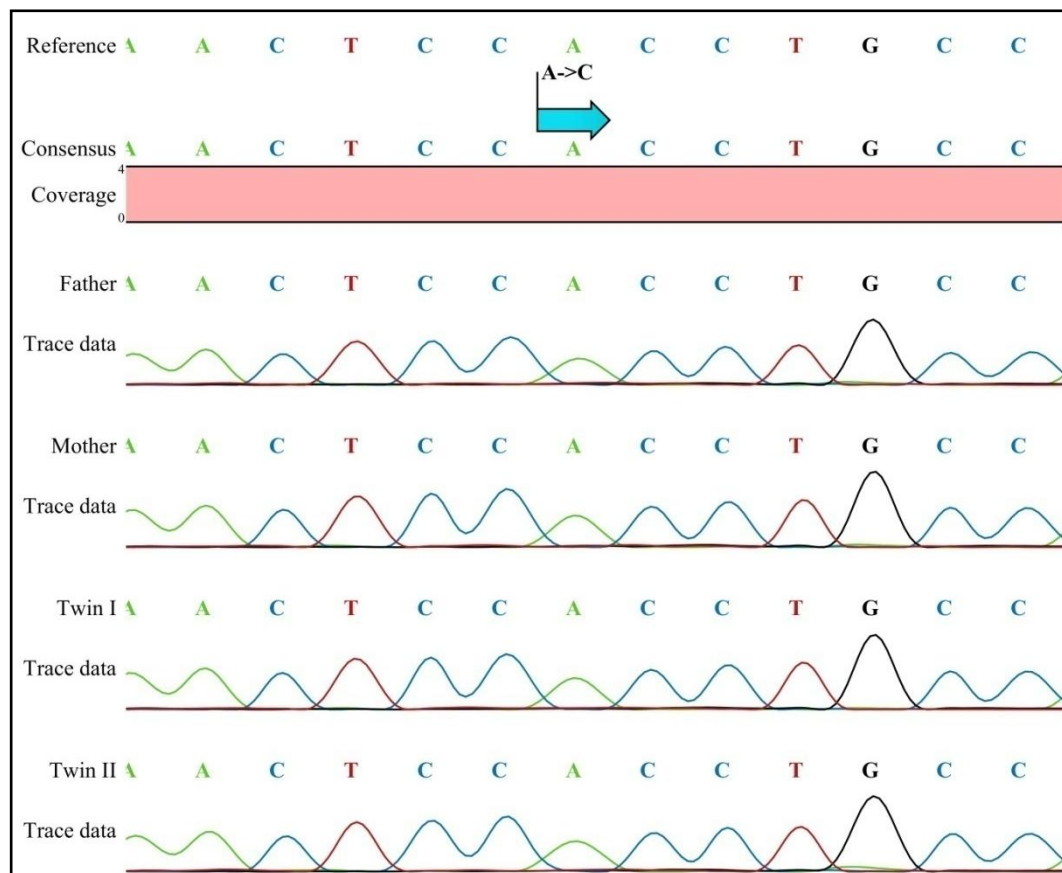


Figure 3.6: A representative Sanger sequencing electropherogram for a putative *de novo* SNV discovered to be not present in the genomes of the twin pair (Chromosome 4: 33630506: A/C).

3.5.2 Validation of the twin-specific *de novo* SNVs

We tested the putative *de novo* SNVs that are specific to each of the twins through Sanger based capillary sequencing of the whole blood derived DNA samples. Herein, we could perform Sanger sequencing for 19 of the 29 and 15 of the 18 high-confidence putative *de novo* SNVs specific to twin I and twin II, respectively (see Table C.2 and Table C.3 for the Sanger sequencing result for each of the tested variants).

Analysis of the capillary sequencing data of the twin pair and the parents synchronously resulted with the validation of 8/19 twin I-specific and 1/15 twin II-specific *de novo* SNVs (Figure 3.7 and see Appendix E for the Sanger sequencing electropherograms for the total of 9 validated twin-specific *de novo* SNVs). Of the 34 twin-specific putative *de novo* SNVs tested by capillary sequencing, 4 were discovered to be inherited and 17 were discovered to be not present in the twins. Sequencing results were uninformative for 9/34 variants.

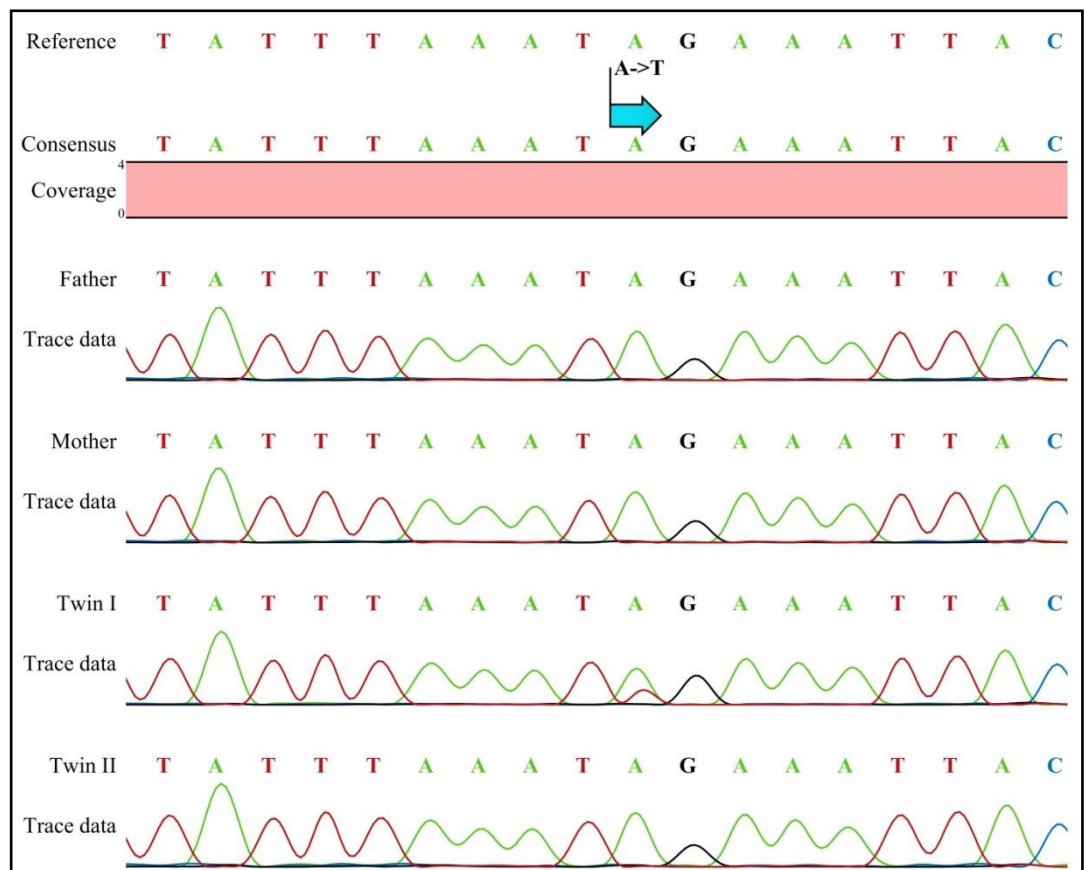


Figure 3.7: A representative Sanger sequencing electropherogram for a twin I-specific *de novo* SNV (Chromosome 2: 61728351: A/T).

3.5.3 Evaluation of the Sanger sequencing results

Subsequent to the validation of the presence of 32 *de novo* SNVs (23 shared by the twin pair, 8 specific to twin-I and 1 specific to twin II) through Sanger-based capillary sequencing, we calculated the false negative rate for the call-set (0.06 – 0.15) based on the number of putative *de novo* SNVs discovered to be inherited from the parents (Table .3.13).

We observed the twin-specific *de novo* SNVs in a mosaic state. Herein, for each of the twin-specific *de novo* SNVs, the peak for the mutant allele on the capillary sequence trace was reduced comparing to the wild-type allele, indicating the mosaic representation of the variants (see Appendix E).[67]

We calculated the RAF for each group of *de novo* SNVs validated by Sanger sequencing. For the *de novo* SNVs shared by the twin pair, average RAF was 0.53 and 0.46 for the twin I and twin II, respectively. For the twin-specific *de novo* SNVs of early post-zygotic origin, the average RAF of 0.65 for the twin I and 0.78 for the twin II were calculated. These reference allele frequencies for the twin-specific *de novo* SNVs supported the presence of the variants in a mosaic state (Table 3.14).

Table 3.13: Statistics on *de novo* SNVs tested by Sanger sequencing

Class of <i>de novo</i> SNVs	All	Primer Design Failed	PCR Failed	Inherited	No variant	Uninformative	<i>De novo</i>	FNR
Shared by the twin pair	73	8	16	8	11	7	23	0.14
Twin I-specific	29	5	5	3	5	3	8	0.15
Twin II-specific	18	2	1	1	12	1	1	0.06
Total	120	15	22	12	28	11	32	

Abbreviations used in this table: FNR, false negative rate

Table 3.14: Average RAF and depth of coverage for the *de novo* SNVs tested by Sanger sequencing

Class of <i>de novo</i> SNVs tested by Sanger sequencing	Twin I		Twin II		Mother		Father	
	Average RAF	Average Depth	Average RAF	Average Depth	Average RAF	Average Depth	Average RAF	Average Depth
<i>De novo</i> SNVs shared by the twin pair	0.53 (0.1 std.)	54.91	0.46 (0.1 std.)	34.13	0.99 (0.02 std.)	32.17	0.99 (0.0 std.)	30.47
Twin I-specific <i>de novo</i> SNVs	0.65 (0.1 std.)	50.50	1.00 (0.0 std.)	38.75	0.99 (0.01 std.)	28.63	1.00 (0.0 std.)	29.13
Twin II-specific <i>de novo</i> SNVs	1	63	0.78	41	1	40	0.96 (0.1 std.)	27
Inherited	0.65 (0.2 std.)	34.43	0.67 (0.2 std.)	20.41	0.92 (0.2 std.)	19.61	0.88 (0.1 std.)	16.86
No variant	0.81 (0.2 std.)	36.46	0.74 (0.1 std.)	24.18	0.94 (0.1 std.)	25.78	0.93 (0.1 std.)	20.92

Abbreviations used in this table: RAF, reference allele frequency; std, standard deviation

3.6 Investigation of the Possible Effect of the Parental Mosaicism

There is a probability of interpreting an SNV as *de novo* in the twin pair erroneously due to the mosaicism in the parental genomes.[67, 102, 103] To exclude this possibility, we examined the DNA from the cells of different embryonic origin, including the DNA isolated from the buccal wash specimen (ectoderm) and urine samples (endoderm) of the parents, in addition to the previously tested blood-derived DNA (mesoderm).[104]

We randomly selected 10 of the 23 *de novo* SNVs shared by the twin pair and performed capillary sequencing of the buccal wash- and urine cell-derived DNA samples of the mother and father for these variants. We did not observe the *de novo* SNVs in the endoderm- and ectoderm-derived DNA samples of the parents, and excluded the probability of parental mosaicism which may result with the misinterpretation of the data (see Appendix F for the Sanger sequencing electropherograms).

3.7 Examination of the dbSNP137-Reported Variants

We filtered the common single nucleotide polymorphisms (SNPs) present in the dbSNP137 database during the filtering process since we intended to discover the private *de novo* SNVs that are not reported previously. The mutations may occur at every single position in the genome, even at the sites that are mutated in the common population. We therefore extended our analysis to the common SNPs to prevent the underestimation of the mutation rate.

We identified a total of 3,619 dbSNP137-reported *de novo* SNP candidates (a) shared by the twin pair (n=2,728) (Table 3.15), (b) specific to twin I (n=329) (Table 3.16), and (c) specific to twin II (n=562) (Table 3.17).

We examined the minor allele frequencies (MAF) for all 3,619 *de novo* SNP candidates and discovered that the most (63.75%) of these SNPs had MAF>10% in the 1000 Genomes Project populations. We identified 65 SNVs (55 shared by the twins, 3 specific to twin I, and 7 specific to twin II) that have MAF of less than 1% (Table 3.18). For each of these 65 potentially *de novo* SNVs, we examined the depth of coverage and the genotype quality in both parents, resulting with the identification of 3 candidate *de novo* SNVs (1 shared by the twins, 1 specific to twin I, and 1 specific to twin II). The remaining 62 SNVs had either zero, or very low (<5X) coverage in either parent's genome, or had low genotyping quality (>60) (Table 3.19).

We further examined these 3 *de novo* SNV candidates based on the genotype quality, depth of coverage, genotype in the parents, and the presence of a CNV in that position (Table 3.20). We used the GenomeSTRiP [105] tool to genotype deletions in the quad family and we identified a deletion which encompasses the *de novo* SNV shared by the twin pair. We excluded the twin II specific *de novo* SNV candidate because of low coverage in the genome of twin II, as it is supported only by four reads. These analyses resulted with only one twin I specific *de novo* SNV candidate with high coverage and high genotyping quality. However, this SNV also seemed to pass our allele bias threshold (in hard filtering 0.2) with a slight difference (0.26; or 13/44). Allele bias ratio is defined as the number of reads that support the alternative allele versus the number of reads that support the reference allele. Hence, we neglected this SNV in our mutation rate estimation.

Table 3.15: Discovery of the dbSNP137 reported *de novo* SNP candidates shared by the twins

Twin I	Twin II	Mother	Father	All	Genomic context
1/1	1/1	0/0	0/0	471	177
0/1	0/1	0/0	0/0	2,668	927
1/1	0/1	0/0	0/0	71	34
0/1	1/1	0/0	0/0	165	95
1/1	1/1	0/1	0/0	782	292
1/1	1/1	0/0	0/1	685	146
1/1	1/1	0/0	1/1	1,167	299
1/1	1/1	1/1	0/0	1,987	758
Total				7,996	2,728

Table 3.16: Discovery of the dbSNP137 reported *de novo* SNP candidates specific to twin I

Twin I	Twin II	Mother	Father	All	Genomic context
1/1	0/0	0/0	0/0	102	51
0/1	0/0	0/0	0/0	597	147
1/1	0/0	0/1	0/0	52	24
1/1	0/1	0/1	0/0	63	22
1/1	0/0	0/0	0/1	16	4
1/1	0/1	0/0	0/1	56	27
1/1	0/1	0/0	1/1	53	17
1/1	0/1	1/1	0/0	80	37
Total				1,019	329

Table 3.17: Discovery of the dbSNP137 reported *de novo* SNP candidates that are specific to twin II

Twin I	Twin II	Mother	Father	All	Genomic context
0/0	1/1	0/0	0/0	35	16
0/0	0/1	0/0	0/0	521	118
0/0	1/1	0/1	0/0	28	14
0/1	1/1	0/1	0/0	274	138
0/0	1/1	0/0	0/1	11	3
0/1	1/1	0/0	0/1	217	90
0/1	1/1	0/0	1/1	187	91
0/1	1/1	1/1	0/0	205	92
Total				1,478	562

Table 3.18: Global minor allele frequency distribution for the *de novo* SNP candidates

Minor allele frequency	Number of <i>de novo</i> SNPs		
	Shared by the twins	Twin I specific	Twin II specific
0.5 - 0.1	1,794	172	342
0.1 - 0.05	227	23	50
0.05 - 0.01	212	19	39
0.01 - 0.005	25	2	4
< 0.005	30	1	3
Not reported	440	112	124

Table 3.19: Filtering of the *de novo* SNPs which have a minor allele frequency (MAF) smaller than 1% based on the depth and genotype quality

<i>De novo</i> Calls	Application of the Filters								Remaining	Del_intersect
	Mother				Father			Mother Father		
	Total	DP=0	0<DP<5	GQ<60	DP=0	0<DP<5	GQ<60	Low GQ		
Shared by the twins	55	21	4	6	12	2	9	15	1	Yes
Twin I-specific	3	0	1	1	0	0	1	1	1	No
Twin II-specific	7	1	1	4	0	0	1	4	1	No ^c

Abbreviations used in this table: DP, depth; GQ, genotype quality; Del_intersect, variant is intersecting a deletion

Table 3.20: List of high-confidence SNPs discovered as *de novo*

						Twin I	Mother	Father	Twin II	Twin I	Mother	Father	Twin II	Twin I	Mother	Father	Twin II
Chr	Pos	ID	Ref	Alt	MAF	Genotype	Genotype	Genotype	Genotype	DP^b	DP	DP	DP	GQ^c	GQ	GQ	GQ
8	20599745	rs79776133	G	C	0,00964	Hom	Ref	Het	Hom	33	22	28	21	99	63	99	57
10	105839182	rs184555413	G	A	0,00046	Het	Ref	Ref	Ref	55	42	43	36	99	99	99	99
16	70472322	rs138210796	A	G	0,00781	Het	Ref	Het	Hom	33	27	15	6	99	69	68	12

Abbreviations used in this table: Chr, chromosome; Pos, position; Ref, reference allele; Alt, alternative allele; MAF, minor allele frequency; DP, depth; GQ, genotype quality

3.8 Calculation of the *De Novo* SNV Rate

The *de novo* SNV rate for each of the twins was calculated based on the number of validated *de novo* SNVs, number of the callable bases and the false negative rate. The formula of “ $\mu = (N \times (1 + \text{FNR})) / L$ ” was used for mutation rate estimation.[13] Herein, μ is the mutation rate, N is the number of Sanger-confirmed *de novo* SNVs, FNR is the false negative rate for the call-set (Table 3.13), and L is the total number of callable bases. The bases were considered as callable if they have a read depth of minimum three and do not intersect the segmental duplications and simple repeats. We calculated 2,703,445,479 bases for twin I and 2,702,251,568 for twin II.

We estimated the rate of pre-zygotic *de novo* SNVs (shared by the twin pair) as 0.97×10^{-8} for the twin pair based on the 23 *de novo* SNVs verified by Sanger sequencing. The rate of the early post-zygotic *de novo* SNVs (specific to twins) was estimated to be 0.34×10^{-8} for the twin I and 0.04×10^{-8} for the twin II. We discovered the overall *de novo* SNV rate of 1.01×10^{-8} and 1.31×10^{-8} for twin I and twin II, respectively (Table 3.21). These data implicate the substantial contribution of early post-zygotic *de novo* SNVs to the overall mutation rate.

Table 3.21: *De novo* mutations of the twin pair and the mutation rate estimation

Sample	All SNVs		Novel SNVs		<i>De novo</i> SNVs			Validated <i>de novo</i> SNVs			Mutation Rate($\times 10^{-8}$)		
	dbSNP137	Novel	Inherited	<i>De novo</i>	Group I	Group II	Total	Group I	Group II	Total	Group I	Group II	Total
Twin I	3,446,849	49,800	48,405	1,396	159	83	242	23	8	31	0.97	0.34	1.31
Twin II	3,443,828	49,182	47,810	1,372		48	207		1	24		0.04	1.01

Abbreviations used in this table: SNV, single nucleotide variation

3.9 Examination of the Incidental Findings

We examined the incidental results in the whole genome sequencing data of the four individuals (the twin pair and their parents). In the context of the incidental result examination, we focused on the LoF mutations and potentially damaging missense mutations.

First, we selected all homozygous and heterozygous mutations present in the genomes of the twin pair and their parents through considering the Mendelian transmission patterns to eliminate the genotyping and variants calling errors. Herein, to select the true mutations, our strategy is based on the selection of the mutations through comparative analysis of the genotypes for each of the variants (SNVs and indels) using SnpSift.[92] We performed functional and positional annotation of the selected variants against human reference genome (GRCh37.68) using SnpEff.[91] We filtered the mutations that are reported in dbSNP137 with a minor allele frequency of greater than 0.01. Finally we extracted the LoF mutations and missense mutations (Table 3.22).

Table 3.22: Identification of the homozygous and heterozygous LoF and missense mutations of the four individual

Genotypes				Number of LoF and Missense Mutations									
Twin I	Twin II	Mother	Father	Novel					dbSNP137-reported (MAF < 0.01)				
				Stopgain	Stoploss	Splice site	Frameshift	Missense	Stopgain	Stoploss	Splice site	Frameshift	Missense
1/1	1/1	1/1	1/1	0	0	1	7	2	0	0	6	7	7
1/1	1/1	0/1	0/1	0	0	1	2	0	0	0	0	0	3
1/1	1/1	1/1	0/1	0	0	0	1	0	0	0	1	1	4
1/1	1/1	0/1	1/1	0	0	0	1	0	0	0	1	1	7
0/1	0/1	0/0	0/1	0	0	2	8	105	1	0	5	3	157
0/1	0/1	0/0	1/1	0	0	1	4	5	0	0	0	3	8
0/1	0/1	0/1	0/0	1	0	4	8	104	1	0	0	0	138
0/1	0/1	1/1	0/0	0	0	1	2	5	0	0	0	1	12

Abbreviations used in this table: LoF, loss of function; MAF, minor allele frequency

3.9.1 Analysis of the LoF mutations

We considered the (a) stop codon introducing (stopgain), (b) stop codon removing (stoploss), (c) splice site disrupting SNVs and (d) frameshift indels as LoF mutations.[97]

3.9.1.1 LoF mutations of the twin pair

3.9.1.1.1 Homozygous LoF mutations of the twin pair

We discovered a total of 30 LoF mutations present in the genomes of the twin pair in homozygous state. 13 of the 30 mutations were novel, not reported in dbSNP137. These 13 novel LoF mutations are located on 12 different genes. We examined the genes that carry these mutations in terms of being reported in OMIM database related to a disease phenotype. Herein, we discovered that 2 of the 13 novel LoF mutations occurred in the genes that are previously reported to be causal for 2 different diseases (Table 3.23). Of the 17 dbSNP137-reported LoF mutations (MAF <0.01), one splice site mutation (rs112378876) discovered to be located on *MTMR2* gene which is involved in the pathogenesis of the autosomal recessively inherited Charcot-Marie-Tooth disease (#601382) (Table 3.24).

Table 3.23: Novel homozygous LoF mutations of the twin pair

Chr	Position	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
1	152195729	T	-	<i>HRNR</i>	frameshift deletion	NA	-	-
5	139931628	-	GT	<i>SRA1</i>	frameshift insertion	NA	CAMPOMELIC DYSPLASIA (#114290)	AD
5	139931629	-	G	<i>SRA1</i>	frameshift insertion	NA	CAMPOMELIC DYSPLASIA (#114290)	AD
9	70900911	-	A	<i>CBWD3</i>	frameshift insertion	NA	-	-
10	116931101	-	TT	<i>ATRNL1</i>	frameshift insertion	0.96	-	-
11	76954788	-	A	<i>GDPD4</i>	frameshift insertion	0.23	-	-
12	55714406	-	A	<i>OR6C1</i>	frameshift insertion	0.34	-	-
17	6555546	-	G	<i>C17orf100</i>	frameshift insertion	1.00	-	-
17	71752500	-	T	<i>LOC100134391</i>	frameshift insertion	1.00	-	-
19	51835892	-	G	<i>VSIG10L</i>	frameshift insertion	NA	-	-
X	122336600	-	G	<i>GRIA3</i>	frameshift insertion	1.00	MENTAL RETARDATION, X-LINKED (#300699)	XR
4	6086694	-	A	<i>JAKMIP1</i>	splice site	NA	-	-
8	96047806	-	A	<i>NDUFAF6</i>	splice site	NA	-	-

Abbreviations used in this table: Chr; chromosome, Ref; reference allele, Alt; alternative allele, A.A Change; aminoacid change, AD; autosomal dominant, XR; X-linked recessive, NA; not available

Table 3.24: dbSNP137-reported homozygous LoF mutations of the twin pair

Chr	Position	dbSNP137	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
1	156565049	rs112595382	-	AC	<i>GPATCH4</i>	frameshift insertion	NA	-	-
3	129814935	rs55800015	C	-	<i>ALG1L2</i>	frameshift deletion	NA	-	-
5	156721863	rs76159126	-	C	<i>CYFIP2</i>	frameshift insertion	NA	-	-
5	175811094	rs78549596	-	GT	<i>NOP16</i>	frameshift insertion	NA	-	-
					<i>RP11-</i>				
15	31521506	rs141879900	GG	-	<i>16E12.2</i>	frameshift deletion	NA	-	-
16	89291210	rs113560259	-	GTGA	<i>ZNF778</i>	frameshift insertion	NA	-	-
22	19189003	rs78649162	-	C	<i>CLTCL1</i>	frameshift insertion	NA	-	-
X	36162684	rs201605278	-	TG	<i>CHDC2</i>	frameshift insertion	NA	-	-
X	101395780	rs200987681	-	G	<i>TCEAL6</i>	frameshift insertion	NA	-	-
3	52872001	rs113892040	TGTGTGTGCACGTGCACG	-	<i>TMEM110</i>	splice site	NA	-	-
9	73458045	rs3833697	-	A	<i>TRPM3</i>	splice site	NA	-	-
11	46342259	rs79068197	-	G	<i>CREB3L1</i>	splice site	NA	-	-
								CHARCOT-MARIE-TOOTH DISEASE, TYPE 4B1(#601382)	
11	95615662	rs112378876	TAAAA	-	<i>MTMR2</i>	splice site	NA	-	AR
12	7080212	rs17857448	T	C	<i>EMG1</i>	splice site	NA	-	-
17	61660890	rs66502552	G	-	<i>DCAF7</i>	splice site	NA	-	-
18	74208486	rs3841258	G	-	<i>FLJ44313</i>	splice site	NA	-	-
19	16268208	rs5827321	A	-	<i>HSH2D</i>	splice site	NA	-	-

Abbreviations used in this table: Chr; chromosome, Ref; reference allele, Alt; alternative allele, AR; autosomal recessive, NA; not available

3.9.1.1.2 Heterozygous LoF mutations of the twin pair

We identified a total of 45 LoF mutations present in the genomes of the twin pair in heterozygous state. Of these, 31 were novel and 14 were reported in SNP database dbSNP137 with MAF of less than 0.01 (see Appendix G for the full list of heterozygous LoF mutations of the twin pair). 3/31 novel and 2/14 dbSNP137-reported LoF mutations appeared on genes that were previously identified for autosomal dominant and recessive diseases in OMIM database (Table 3.25).

Table 3.25: List of heterozygous LoF mutations in the twin pair located on genes related to diseases

Chr	Position	dbSNP137	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
5	171765480	NA	CAAA	-	<i>SH3PXD2B</i>	frameshift deletion	NA	FRANK-TER HAAR SYNDROME (#249420)	AR
8	103250839	NA	-	G	<i>RRM2B</i>	frameshift insertion	0.06	PROGRESSIVE EXTERNAL OPHTHALMOPLÉGIA (#613077)	AD
17	57141768	NA	-	A	<i>TRIM37</i>	splice site	NA	MULIBREY NANISM (#253250)	AR
17	17697102	rs34083643	G	-	<i>RAI1</i>	frameshift deletion	NA	SMITH-MAGENIS SYNDROME (#182290)	AD
18	42456670	rs3085861	-	TCTT	<i>SETBP1</i>	frameshift insertion	NA	SCHINZEL-GIEDION MIDFACE RETRACTION SYNDROME (#269150)	AD

Abbreviations used in this table: Chr; chromosome, Ref; reference allele, Alt; alternative allele, AR; autosomal recessive, AD; autosomal dominant, NA; not available

3.9.1.2 LoF mutations of the parents

3.9.1.2.1 Homozygous LoF mutations of the parents

We identified 28 and 32 LoF mutations present in homozygous state in the genome of the mother and father, respectively (see Appendix H for the full list of homozygous LoF mutations of the parents). 12/28 mutations in the mother and 14/32 mutations in the father discovered to be novel. Of the novel LoF mutations, 2/12 and 1/14 appeared to be located on genes that are involved in autosomal dominant and X linked recessive diseases in mother and father, respectively. Of the dbSNP137-reported LoF mutations, 1 mutation in both mother and father appeared on *MTMR2* gene which is reported to be causal for Charcot-Marie-Tooth disease (#601382) (Table 3.26).

Table 3.26: List of the homozygous LoF mutations located on disease genes in the parents

Chr	Position	dbSNP137	Ref	Alt	Gene	Function	1000G	Individual	OMIM Disease	Inheritance
8	103250839	NA	-	G	<i>RRM2B</i>	frameshift insertion	0.06	Mother	PROGRESSIVE EXTERNAL OPHTHALMOPLÉGIA (#613077)	AD
11	95615662	rs112378876	TAAAA	-	<i>MTMR2</i>	splice site frameshift insertion	NA	Father&Mother	CHARCOT-MARIE-TOOTH DISEASE, TYPE 4B1(#601382)	AR
X	122336600	NA	-	G	<i>GRIA3</i>	frameshift insertion	1.00	Father&Mother	MENTAL RETARDATION, X-LINKED (#300699)	XR

Abbreviations used in this table: Chr; chromosome, Ref; reference allele, Alt; alternative allele, AR; autosomal recessive, AD; autosomal dominant, XR; X linked recessive, NA; not available

3.9.1.2.2 Heterozygous LoF mutations of the parents

We discovered a total of 45 LoF mutations present in heterozygous state in the genomes of the mother (n=20) and father (n=25) (see Appendix I for the full list of heterozygous LoF mutations of the parents). Of these, 31 (17 in the genome of mother and 14 in the genome of father) were novel. We analyzed the genes that carry all of the 45 heterozygous LoF mutations to evaluate the involvement of them in diseases. Herein, we observed that 6 of the 45 variants were located on genes that are reported to be causal for autosomal dominant and recessive diseases in OMIM (Table 3.27).

Table 3.27: List of the heterozygous LoF mutations located on disease genes in the parents

Chr	Position	dbSNP137	Ref	Alt	Gene	Function	1000G	Individual	OMIM Disease	Inheritance
5	139931628	NA	-	GT	<i>SRA1</i>	frameshift insertion	NA	Mother and Father	CAMPOMELIC DYSPLASIA (#114290)	AD
5	139931629	NA	-	G	<i>SRA1</i>	frameshift insertion	NA	Mother and Father	CAMPOMELIC DYSPLASIA (#114290)	AD
5	171765480	NA	CAAA	-	<i>SH3PXD2B</i>	frameshift deletion	NA	Mother	FRANK-TER HAAR SYNDROME (#249420)	AR
17	57141768	NA	-	A	<i>TRIM37</i>	splice site	NA	Mother	MULIBREY NANISM (#253250)	AR
17	17697102	rs34083643	G	-	<i>RAI1</i>	frameshift deletion	NA	Father	SMITH-MAGENIS SYNDROME (#182290)	AD
18	42456670	rs3085861	-	TCTT	<i>SETBP1</i>	frameshift insertion	NA	Father	SCHINZEL-GIEDION MIDFACE RETRACTION SYNDROME (#269150)	AD

Abbreviations used in this table: Chr; chromosome, Ref; reference allele, Alt; alternative allele, AR; autosomal recessive, AD; autosomal dominant, NA; not available

3.9.2 Analysis of the missense mutations

In the genomes of the twin pair, we identified a total of 23 homozygous and 534 heterozygous missense SNVs that pass our filtering criteria (novel and dbSNP137-reported with MAF of less than 0.01). None of the genes that carry the 23 homozygous missense SNVs were reported in OMIM database. We evaluated the deleteriousness of the variants based on the SIFT, Polyphen2 and MutationTaster.scores.[98, 99, 100] Herein, 7 of the 23 homozygous missense variants discovered to be damaging (Table 3.28)

Of the 534 heterozygous missense SNVs (219 novel and 315 dbSNP137-reported), 87 variants appeared on disease related genes. Despite their occurrence on the genes that are related to diseases, 22 of the 87 variants were predicted to be benign or neutral based on the SIFT, Polyphen2 and MuutationTaster (Table 3.29). Remaining 65 missense SNVs were predicted to be deleterious according to at least one of the mutation effect assessing tools.

We evaluated the deleteriousness of the remaining 447 variants that are present on the genes which do not have any clinical significance. Herein, 304 of the 447 variants predicted to be deleterious.

Table 3.28: Homozygous missense variants present in the genomes of the twin pair

Chr	Position	dbSNP137	Ref	Alt	Gene	1000G	SIFT	PP2	PP2	MutTaster	MutTaster
2	98164020	rs76158480	C	T	<i>ANKRD36B</i>	NA	.	.	.	.	.
3	195512186	rs202060675	T	C	<i>MUC4</i>	NA	.	0.234	B	.	.
5	140229086	rs251354	C	G	<i>PCDHA9</i>	NA	1.000.000	0.0	B	0.000025	D
6	31238942	rs2308590	G	T	<i>HLA-C</i>	NA	0.440000	0.0	B	.	.
6	31239622	rs9264668	C	A	<i>HLA-C</i>	NA	1.000.000	0.032	B	.	.
6	32487309	rs3200405	T	C	<i>HLA-DRB5</i>	NA	1.000.000	0.0	B	0.001085	N
6	32489856	rs77853982	C	T	<i>HLA-DRB5</i>	NA	0.300000	0.0	B	0.008476	N
6	32497961	rs1064587	T	A	<i>HLA-DRB5</i>	NA	0.000000	0.001	B	0.000040	N
6	32497962	rs701884	T	C	<i>HLA-DRB5</i>	NA	0.000000	0.0	B	0.000007	N
6	32726803	rs1049110	C	T	<i>HLA-DQB2</i>	NA	0.020000	0.666	D	.	.
9	33798017	rs855581	A	G	<i>PRSS3</i>	NA	0.430000	0.0	B	0.000224	N
10	46999604	rs3127683	A	G	<i>GPRIN2</i>	NA	0.820000	0.0	B	0.000002	N
11	61165741	rs10897158	G	C	<i>TMEM216</i>	NA	.	.	.	.	.
12	11035274	rs28607516	G	T	<i>PRH1</i>	NA	0.260000	.	.	.	.
14	74041748	rs149033118	A	G	<i>ACOT2</i>	NA	0.740000	0.001	B	0.005796	N
16	1245509	rs60593994	G	C	<i>CACNA1H</i>	0.0018	0.030000	0.998	D	0.031773	N
17	39253835	rs72625995	C	T	<i>KRTAP4-8</i>	NA	0.790000	0.0	B	0.000709	N
17	39421781	rs12938374	A	G	<i>KRTAP9-6</i>	NA	1.000.000	.	.	0.000191	D
19	4511730	rs62115190	T	C	<i>PLIN4</i>	NA	0.890000	0.0	B	.	.
21	34166190	rs74617086	A	T	<i>C21orf62</i>	NA	1.000.000	0.0	B	0.289388	N
X	153881773	rs17855367	T	C	<i>CTAG2</i>	NA	1.000.000	0.833	D	0.000034	N
3	195512042	NA	T	C	<i>MUC4</i>	NA	.	0.004	B	-0.55	N
3	195512107	NA	T	A	<i>MUC4</i>	NA	.	0.204	B	-0.55	N

Abbreviations used in this table: Chr; chromosome, Ref; reference allele, Alt; alternative allele, PP2; Polyphen2, MutTaster; MutationTaster

Table 3.29: Heterozygous missense variants located on disease related genes and present in the genomes of the twin pair

Chr	Position	dbSNP137	Ref	Alt	Gene	1000G	SIFT	PP2	PP2	MutTaster	MutTaster	OMIM Disease	Inheritance
1	22202483	rs62642528	G	A	<i>HSPG2</i>	NA	0.100000	1.0	D	0.810966	D	DYSSEGMENTAL DYSPLASIA, SILVERMAN-HANDMAKER TYPE (#224410)	AR
1	52854931	.	T	C	<i>ORC1</i>	NA	0.650000	0.074	B	0.565708	D	MEIER-GORLIN SYNDROME 1; MGORS1 (#224690)	AR
1	78407816	rs200071700	G	C	<i>NEXN</i>	NA	0.230000	0.005	B	.	.	CARDIOMYOPATHY, DILATED, 1CC (#613122)	-
1	152280900	rs74129452	T	G	<i>FLG</i>	NA	0.180000	0.0	B	0.000210	N	DERMATITIS, ATOPIC, 2 (#605803)	-
1	152281007	rs7512553	A	G	<i>FLG</i>	NA	0.560000	0.0	B	0.000005	N	DERMATITIS, ATOPIC, 2 (#605803)	-
1	152281228	rs7546186	C	G	<i>FLG</i>	NA	0.540000	0.899	P	0.000131	N	DERMATITIS, ATOPIC, 2 (#605803)	-
1	152281304	rs7512857	A	C	<i>FLG</i>	NA	0.210000	0.998	D	0.000289	N	DERMATITIS, ATOPIC, 2 (#605803)	-
1	155264432	.	A	G	<i>PKLR</i>	NA	0.610000	0.0	B	0.000626	N	PYRUVATE KINASE DEFICIENCY OF RED CELLS (#266200)	AR
1	197024952	.	G	T	<i>F13B</i>	NA	1.000.000	0.003	B	0.003300	N	FACTOR XIII, B SUBUNIT, DEFICIENCY OF (#613235)	AR
1	222826363	.	T	G	<i>MYH6</i>	NA	0.760000	0.996	D	.	.	ATRIAL SEPTAL DEFECT 3 (#614089)	-
2	1426888	.	G	A	<i>TPO</i>	NA	0.620000	0.394	B	0.001875	N	THYROID DYSHORMONOGENESIS 2A (#274500)	AR
2	72359508	rs201193616	G	A	<i>CYP26B1</i>	NA	0.000000	1.0	D	0.998712	D	RADIOHUMERAL FUSIONS WITH OTHER SKELETAL AND CRANIOFACIAL ANOMALIES (#614416)	AR
2	74593127	.	G	A	<i>DCTN1</i>	NA	0.010000	1.0	D	0.999997	D	PERRY SYNDROME (#168605) (late onset)	AD
2	179429917	.	C	T	<i>TTN</i>	NA	0.000000	0.271	B	.	.	CARDIOMYOPATHY, DILATED, 1G (#604145)	-
2	220420990	rs183329050	C	T	<i>OBSL1</i>	0.0027	0.450000	1.0	D	.	.	THREE M SYNDROME 2 (#612921)	-
2	233404771	rs148468628	G	A	<i>CHRNA</i>	0.0009	0.010000	1.0	D	0.507618	D	MULTIPLE PTERYGIUM SYNDROME, ESCOBAR VARIANT (#265000)	AR
2	234527117	.	C	T	<i>UGT1A8</i>	NA	0.090000	0.573	P	.	.	CRIGLER-NAJJAR SYNDROME, TYPE I (#218800)	AR

3	184040371	rs111924994	G	C	<i>EIF4G1</i>	0.0023	0.170000	0.001	B	0.000907	N	PARKINSON DISEASE 18 (#614251)	AD
3	186389577	rs201250387	T	C	<i>HRG</i>	NA	0.780000	0.0	B	0.004688	N	THROMBOPHILIA DUE TO HISTIDINE-RICH GLYCOPROTEIN DEFICIENCY (#613116)	-
3	189612100	.	C	T	<i>TP63</i>	NA	.	1.0	D	0.999556	D	ANKYLOBLEPHARON-ECTODERMAL DEFECTS-CLEFT LIP/PALATE (#106260)	AD
3	189838136	rs200731219	C	A	<i>LEPREL1</i>	NA	.	0.999	D	0.999975	D	MYOPIA, HIGH, WITH CATARACT AND VITREORETINAL DEGENERATION (#614292)	AR
4	2834080	rs148761331	C	T	<i>SH3BP2</i>	0.0027	0.010000	0.998	D	0.278408	N	CHERUBISM (#118400)	AD
4	52890281	rs200761715	G	A	<i>SGCB</i>	0	0.090000	1.0	D	0.999594	D	MUSCULAR DYSTROPHY, LIMB-GIRDLE, TYPE 2E (#604286)	AR
4	89013496	rs34783571	C	T	<i>ABCG2</i>	0.0023	0.300000	0.088	B	0.095330	N	BLOOD GROUP, JUNIOR SYSTEM (#614490)	-
4	114275061	.	A	T	<i>ANK2</i>	NA	.	0.725	P	.	.	CARDIAC ARRHYTHMIA, ANKYRIN-B-RELATED (#600919)	AD
4	149356870	.	C	A	<i>NR3C2</i>	NA	.	0.018	B	0.144584	N	HYPERTENSION, EARLY-ONSET, WITH SEVERE EXACERBATION (#605115)	AD
4	187179340	.	A	G	<i>KLKB1</i>	NA	0.260000	0.0	B	0.000057	N	PREKALLIKREIN DEFICIENCY (#612423)	AR
5	10262729	.	A	C	<i>CCT5</i>	NA	0.480000	0.092	B	0.999953	D	NEUROPATHY, HEREDITARY SENSORY, WITH SPASTIC PARAPLEGIA (#256840)	AR
5	13914743	rs140782270	A	T	<i>DNAH5</i>	0.0018	0.010000	0.998	D	0.964611	D	CILIARY DYSKINESIA, PRIMARY, 1 (#244400)	AR
5	125930919	.	A	G	<i>ALDH7A1</i>	NA	0.010000	0.072	B	.	.	EPILEPSY, PYRIDOXINE-DEPENDENT (#266100)	AR
6	31473513	rs2240858	C	T	<i>MICB</i>	NA	0.140000	0.002	B	0.001717	N	RHEUMATOID ARTHRITIS (#180300)	-
6	64488001	rs74636274	T	C	<i>EYS</i>	0.0041	0.350000	0.087	B	.	.	RETINITIS PIGMENTOSA 25 (#602772)	
6	144508353	rs141499372	G	A	<i>STX11</i>	0.0009	0.070000	0.892	P	0.024250	N	HEMOPHAGOCYTIC LYMPHOHISTIOCYTOSIS, FAMILIAL, 4 (#603552)	AR
6	152668337	.	G	T	<i>SYNE1</i>	NA	0.810000	0.988	D	0.986771	D	EMERY-DREIFUSS MUSCULAR DYSTROPHY 4 (#612998),	AD
7	87031444	.	C	A	<i>ABCB4</i>	NA	0.000000	0.991	D	0.999984	D	CHOLESTASIS, INTRAHEPATIC, OF PREGNANCY 3 (#614972)	AD, AR
7	107418636	.	C	T	<i>SLC26A3</i>	NA	0.020000	1.0	D	0.569841	D	DIARRHEA 1, SECRETORY CHLORIDE, CONGENITAL (#214700)	AR

7	153750131	.	G	A	<i>DPP6</i>	NA	0.480000	0.997	D	.	.	VENTRICULAR FIBRILLATION, PAROXYSMAL FAMILIAL, 2 (#612956)	-
8	1844586	rs201912073	A	G	<i>ARHGEF10</i>	0.0005	0.250000	0.004	B	0.547177	D	SLOWED NERVE CONDUCTION VELOCITY (#608236)	AD
8	96064423	rs61743028	G	A	<i>NDUFAF6</i>	0.0041	0.360000	0.01	B	0.999087	D	MITOCHONDRIAL COMPLEX I DEFICIENCY (#252010)	AR
8	141321405	rs148453804	C	T	<i>TRAPPC9</i>	NA	0.480000	0.003	B	0.000311	N	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 13 (#613192)	AR
8	144995699	.	G	A	<i>PLEC</i>	NA	0.520000	0.669	P	0.993196	D	EPIDERMOLYSIS BULLOSA SIMPLEX WITH MUSCULAR DYSTROPHY (#226670)	AR
9	406999	rs34390308	C	T	<i>DOCK8</i>	0.0014	0.230000	1.0	D	0.999437	D	HYPER-IgE RECURRENT INFECTION SYNDROME (#243700)	AR
9	13136783	rs200891478	T	C	<i>MPDZ</i>	NA	0.110000	1.0	D	0.883966	D	HYDROCEPHALUS, NONSYNDROMIC, AUTOSOMAL RECESSIVE 2 (#615219)	AR
9	35079502	rs35984312	G	A	<i>FANCG</i>	0.0014	0.030000	0.119	B	0.003211	N	FANCONI ANEMIA, COMPLEMENTATION GROUP G; FANCG (#614082)	-
9	77411729	rs143164660	C	G	<i>TRPM6</i>	0.0041	0.000000	1.0	D	0.981199	D	HYPOMAGNESEMIA 1, INTESTINAL (#602014)	AR
9	79897155	rs148656796	A	G	<i>VPS13A</i>	0.0032	0.310000	0.964	D	0.940622	D	CHOREOACANTHOCYTOSIS (#200150)	AR
9	111640962	.	T	C	<i>IKBKAP</i>	NA	0.000000	1.0	D	0.999581	D	NEUROPATHY, HEREDITARY SENSORY AND AUTONOMIC, TYPE III (#223900)	AR
9	130634139	.	G	A	<i>AK1</i>	NA	0.000000	1.0	D	0.999998	D	ADENYLATE KINASE DEFICIENCY (#612631)	AR
9	135786013	rs118203504	G	A	<i>TSC1</i>	NA	0.370000	0.976	D	.	.	FOCAL CORTICAL DYSPLASIA OF TAYLOR (#607341)	Isolated cases
10	31810670	rs148358382	G	A	<i>ZEB1</i>	NA	0.110000	1.0	D	0.836949	D	CORNEAL DYSTROPHY, FUCHS ENDOTHELIAL, 6 (#613270)	-
10	43600559	rs139790943	T	C	<i>RET</i>	NA	0.000000	0.917	P	0.346327	N	CENTRAL HYPOVENTILATION SYNDROME (#209880)	AD
10	55700727	.	G	A	<i>PCDH15</i>	NA	0.020000	0.998	D	0.999997	D	DEAFNESS, AUTOSOMAL RECESSIVE 23 (#609533), USHER SYNDROME (#601067)	AR
10	73453990	rs181255269	C	T	<i>CDH23</i>	0.0009	0.410000	0.981	D	0.824721	D	DEAFNESS, AUTOSOMAL RECESSIVE 12 (#601386)	AR
11	1253980	rs202127660	A	G	<i>MUC5B</i>	NA	0.000000	0.0	B	.	.	PULMONARY FIBROSIS, IDIOPATHIC (#178500)	AD

11	18305355	rs61755718	C	T	<i>HPS5</i>	0.0037	0.110000	0.002	B	0.006767	N	HERMANSKY-PUDLAK SYNDROME 5 (#614074)	-
11	57367469	.	G	A	<i>SERPING1</i>	NA	0.010000	0.002	B	0.005601	N	ANGIOEDEMA, HEREDITARY, TYPE I (#106100)	AD
11	117252491	rs59763167	C	G	<i>CEP164</i>	NA	0.040000	0.95	P	0.013561	N	NEPHRONOPHTHISIS 15 (#614845)	AR
11	119213319	rs145881139	C	A	<i>MFRP</i>	0.0041	.	0.744	P	.	.	MICROPTHALMIA, ISOLATED 5 (#611040)	AR
12	33031395	rs150821281	G	A	<i>PKP2</i>	0.0014	0.040000	0.007	B	0.070427	N	ARRHYTHMOGENIC RIGHT VENTRICULAR DYSPLASIA (#609040)	AD
12	49444965	.	G	A	<i>KMT2D</i>	NA	0.000000	0.0	B	.	.	KABUKI SYNDROME 1; KABUK1 (#147920)	AD
12	110784210	rs140234740	A	G	<i>ATP2A2</i>	NA	.	0.001	B	.	.	ACROKERATOSIS VERRUCIFORMIS (#101900)	AD
12	122277716	.	A	G	<i>HPD</i>	NA	0.000000	1.0	D	1.000.000	D	HAWKINSINURIA (#140350)	AD
13	20716248	.	C	G	<i>GJA3</i>	NA	0.540000	0.022	B	0.120999	N	CATARACT 14, MULTIPLE TYPES (#601885)	AD
13	23909676	rs111540787	A	C	<i>SACS</i>	0.0009	0.050000	0.999	D	0.995938	D	SPASTIC ATAXIA, CHARLEVOIX-SAGUENAY TYPE (#270550)	AR
14	20943258	.	G	A	<i>PNP</i>	NA	0.340000	0.77	P	0.978143	D	PURINE NUCLEOSIDE PHOSPHORYLASE DEFICIENCY (#613179)	AR
14	23855315	.	G	A	<i>MYH6</i>	NA	0.050000	0.003	B	0.384519	N	CARDIOMYOPATHY, DILATED (#613252)	-
14	54417527	.	G	C	<i>BMP4</i>	NA	0.170000	0.75	P	0.948209	D	MICROPTHALMIA, SYNDROMIC 6 (#607932)	AD
15	34634289	.	T	A	<i>NOP10</i>	NA	0.050000	0.928	P	0.981674	D	DYSKERATOSIS CONGENITA, AUTOSOMAL RECESSIVE, 1 (#224230)	AR
16	1263845	.	A	C	<i>CACNA1H</i>	NA	0.000000	1.0	D	0.331794	N	EPILEPSY, CHILDHOOD ABSENCE, SUSCEPTIBILITY TO, 6 (#611942)	-
16	2140294	rs148478410	C	T	<i>PKD1</i>	0.0041	0.170000	0.987	D	0.004136	N	POLYCYSTIC KIDNEY DISEASE 1 (#173900)	AD
16	3293403	rs104895094	T	C	<i>MEFV</i>	0.0032	0.150000	0.939	P	0.001830	N	FAMILIAL MEDITERRANEAN FEVER (#134610)	AD
16	74773946	rs147632811	C	T	<i>FA2H</i>	0.0005	0.670000	0.041	B	0.000283	N	SPASTIC PARAPLEGIA 35 (#602319)	AR
17	6913317	.	C	T	<i>ALOX12</i>	NA	0.010000	0.998	D	0.827348	D	COLORECTAL CANCER (#114500)	AD
17	40947281	.	C	A	<i>WNK4</i>	NA	0.450000	0.617	P	0.000970	N	PSEUDOHYPOALDOSTERONISM, TYPE IIB (#614491)	AD
17	59560453	rs140662248	C	G	<i>TBX4</i>	0	.	0.983	D	0.012565	N	SMALL PATELLA SYNDROME (#147891)	AD
17	73513281	rs148146916	C	G	<i>TSEN54</i>	0.0009	0.250000	0.753	P	0.046230	N	PONTocerebellar hypoplasia, type 2A (#277470)	AR
18	21519197	.	A	C	<i>LAMA3</i>	NA	0.510000	1.0	D	0.949775	D	EPIDERMOLYSIS BULLOSA, JUNCTIONAL,	AR

19	856130	rs17216663	C	T	<i>ELANE</i>	0.0046	0.160000	0.01	B	0.005241	N	HERLITZ TYPE (#226700)	
19	11556233	rs200168017	G	A	<i>PRKCSH</i>	0.0009	0.850000	0.839	P	0.804291	D	CYCLIC NEUTROPENIA (#162800)	AD
19	13988258	.	A	C	<i>CDHR1</i>	NA	0.520000	0.002	B	0.115886	N	POLYCYSTIC LIVER DISEASE (#174050)	AD
19	17322578	rs199695802	C	T	<i>MYO9B</i>	0.0009	0.130000	0.271	B	0.005599	N	CONE-ROD DYSTROPHY 15 (#613660)	AR
19	30193732	rs146492790	C	T	<i>C19orf12</i>	0.0005	0.080000	0.994	D	0.196292	N	CELIAC DISEASE, SUSCEPTIBILITY TO, 4 (#609753)	-
19	38968441	.	G	A	<i>RYR1</i>	NA	0.640000	0.01	B	0.010457	N	NEURODEGENERATION WITH BRAIN IRON ACCUMULATION 4 (#614298)	AR
19	50332235	rs200785685	C	T	<i>MED25</i>	NA	0.060000	1.0	D	0.998578	D	MINICORE MYOPATHY WITH EXTERNAL OPHTHALMOPLEGIA (#255320)	AR
20	32881908	rs11552695	T	C	<i>AHCY</i>	0.0027	0.070000	0.0	B	0.999552	D	CHARCOT-MARIE-TOOTH DISEASE, AXONAL, TYPE 2B2 (#605589)	AR
22	18567938	rs149153003	C	T	<i>PEX26</i>	0.0005	0.780000	0.037	B	0.033637	N	HYPERMETHIONINEMIA WITH S-ADENOSYLHOMOCYSTEINE HYDROLASE DEFICIENCY	AR
22	19165823	.	C	G	<i>SLC25A1</i>	NA	NA	NA	NA	NA	NA	PEROXISOME BIOGENESIS DISORDER 7A (ZELLWEGER) (#614872)	-
22	26860200	rs147435410	G	A	<i>HPS4</i>	0.0009	.	0.002	B	0.032402	N	COMBINED D-2- AND L-2-HYDROXYGLUTARIC ACIDURIA (#615182)	AR
X	153006093	rs201878013	A	G	<i>ABCD1</i>	NA	1.000.000	0.0	B	0.000145	N	HERMANSKY-PUDLAK SYNDROME 4 (#614073)	-
												ADRENOLEUKODYSTROPHY (#300100)	XR

Abbreviations used in this table: Chr; chromosome, Ref; reference allele, Alt; alternative allele, PP2; Polyphen2, MutTaster; Mutationtaster, AR; autosomal recessive, AD; autosomal dominant, XR; X linked recessive

3.9.3 Analysis of the homozygous stretches

We analyzed the homozygous stretches segregated within the family using HomozygosityMapper (<http://www.homozygositymapper.org/>). To identify the homozygous stretches, we introduced the twin pair as affected and the parents as healthy to the program. Herein, we identified 4 homozygous regions present in the genomes of the twin pair (Figure 3.8 and Table 3.30).

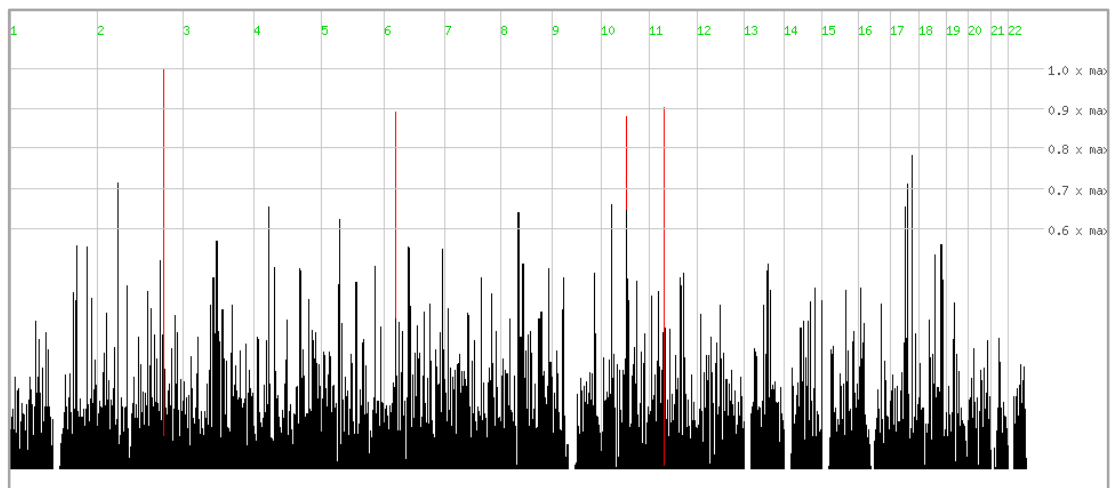


Figure 3.8: Homozygosity mapping analysis in the twin pair. Y axis is the genome wide homozygosity scores. Red lines indicates the common homozygous intervals.

Table 3.30: Homozygous intervals in the twin pair

Chr	from (bp)	to (bp)	length (bp)	from SNP	to SNP
2	186374432	187390507	1016075(1Mb)	rs186374432	rs187390507
6	30955182	31171257	216075 (0.2 Mb)	rs30955182	rs31171257
10	70199633	70786474	586841 (0.5 Mb)	rs70199633	rs70786474
11	42856020	43598428	742408 (0.7 Mb)	rs42856020	rs43598428

Abbreviations used in this table: Chr; chromosome, bp; base pair, SNP; single nucleotide polymorphism

We examined the homozygous stretches to assess the genes that are intersecting these regions. On the 1 Mb homozygous block located on chromosome 2, we identified 8 genes (*ELF2P4*, *FSIP2*, *LOC100420895*, *RPL21P32*, *RRPL23AP35*, *MED28P3*, *ZC3H15*, *DPRXP1*). Of these 6 were discovered to be pseudogenes. We analyzed the remaining 2 protein coding genes (*FSIP2* and *ZC3H15*) and observed that they did not have any clinical significance (Table 3.31).

We identified a total of 12 genes (*MUC21*, *MUC22*, *HCG22*, *C6orf15*, *PSORS1C1*, *CDSN*, *PSORS1C2*, *CCHCR1*, *TCF19*, *POU5F1*, *PSORS1C3*, *HCG27*) within the homozygous block located on chromosome 6, spanning 0.2 Mb. Of these, 9 were protein coding and 3 were non-coding RNA genes. None of these genes had a clinical significance.

A total of 16 genes discovered to be located on the homozygous block on the chromosome 10, spanning 0.5 Mb. 2 of the 16 genes had clinical importance. One of them is *STOX1* gene. Heterozygous mutations of the *STOX1* gene reported to be causal for Preeclampsia 4 (#609404). Other gene, *KIAA1279*, reported to be involved in Goldberg-Shprintzen syndrome (#609460) which is inherited in autosomal recessive fashion.

We discovered 7 genes (*LOC100533710*, *HNRNPKP3*, *API5*, *TTC17*, *CTBP2P6*, *MIR670*, *MIR670HG*) located on the 0.7 Mb homozygous stretch on chromosome 11. None of these genes involved in disease pathogenesis.

Chapter 4

Discussion

Characterizing the new mutations regarding their origins and rates has great importance in terms of providing insights into the disease pathogenesis as well as the evolutionary processes. Studies concerned with the *de novo* mutation rate estimation have largely focused on the SNVs. This is due to the limitations of current technologies in the investigation of the other types of genetic variation.[3, 25]

Direct estimation of the *de novo* SNV rate in healthy individuals which is based on the whole genome sequencing of parent-offspring trios yielded rates ranging between $0.82 - 1.70 \times 10^{-8}$ per base per generation.[12, 13, 55, 56] These studies that provided the direct *de novo* SNV mutation rates were restricted with the scope of the trio sequencing. At this point, the early post-zygotic *de novo* SNVs could not be distinguished from those which have a pre-zygotic origin. Therefore, the rate of the early post-zygotic *de novo* SNVs within the overall *de novo* SNV burden remained unknown.[3]

At this point, to elucidate the origins of *de novo* SNVs and to estimate the *de novo* SNV rate in healthy individuals regarding the developmental timing of mutagenesis, we performed deep whole genome sequencing of a healthy monozygotic twin pair together with their parents. We anticipated that *de novo* SNVs which are present in the genomes of both twins would have meiotic origin, occur during the gametogenesis pre-zygotically yet those that are specific to one of the twins could have mitotic origin, occur early during development possibly in the post-twinning embryo.[6, 65, 69, 106] On the basis of this hypothesis, we characterized twenty three *de novo* SNVs that have a pre-zygotic origin; eight early post-zygotic *de novo* SNVs in twin I, and one early post-zygotic *de novo* SNV in the genome of the twin II.

We report an overall *de novo* SNV rate of 1.01×10^{-8} and 1.31×10^{-8} for each of the twins. This estimate is consistent with the previous direct *de novo* SNV rate estimations.[12, 13, 55, 56] Furthermore, we estimated the rate of the early post-zygotic *de novo* SNVs within the overall *de novo* SNV rate as 0.34×10^{-8} and 0.04×10^{-8} for twin I and twin II, respectively. Our findings indicate that the direct measurement of the early post-zygotic vs. pre-zygotic *de novo* SNV rate could be possible through whole genome sequencing of the monozygotic twins together with a progressive filtering and validation approach.

We selected a total of 120 high-confidence putative *de novo* SNVs for Sanger sequencing-based validation. We could not test 37 (24 shared, 10 twin I-specific and 3 twin II-specific) of the 120 selected putative *de novo* SNVs due to unavailability of the appropriate primer pairs or PCR products of good quality. Herein, to evaluate the degree of mutation rate underestimation due to these technical limitations, we considered the proportion of the validated variants (50%) among those that could be tested successfully. Herein, re-calculation resulted with an overall mutation rate of 2.02×10^{-8} for twin I and 1.51×10^{-8} for twin II. Moreover, rate of the early post-zygotic *de novo* SNVs increased to 0.55×10^{-8} for twin I.

The most previous studies concerned with the estimation of the somatic mutation rate has been restricted with the cancer studies as well as the *in vitro* cell models.[62, 63, 64] In healthy individuals, the average somatic mutation rate was reported to be 7.7×10^{-10} per base per cell division.[59] In 2011, Conrad and his colleagues discriminated the germline mutations and non-germline mutations and they estimated the non-germline *de novo* SNV rate as 2.52×10^{-7} . [56] Still, this rate might be an overestimation due to the probability of the interference of the rate with the cell-line derived mutations. In a very recent study, somatic *de novo* mutation rate was reported for the common single nucleotide polymorphisms (SNPs) present in the human genome through surveying the whole genomes of the monozygotic twins. However, this estimate (1.2×10^{-7} per nucleotide) might not reflect the true somatic mutation rate because the rare and private variations were neglected through focusing on only the reported variation.[65] In our study, rate of the early post-zygotic *de novo* SNVs is between the upper and lower boundaries of the previous somatic mutation rate estimations.

We observed that early post-zygotic *de novo* SNV rate varied between two monozygotic twins considerably. This is due to the characterization of 8 early post-zygotic *de novo* SNVs in one of the twins, while the number was 1 for another twin. At this point, the difference between post-zygotic mutation rates might be explained by stochastic events that affect the twin pair during the intra-uterine development.[107] Furthermore, in a paper by Steven Frank from UC Irvine, it is explained neatly that mutations that arise during the early development may follow the so called Luria-Delbruck distribution and the degree of mosaicism may vary between individuals.[72] We, still, point out that a larger sample size is required to make a firmer conclusion about the post-zygotic mutation rate.

It has been reported that the *de novo* mutations that are present in the genome of only one of the monozygotic twins might be cell line derived.[108] Since we sequenced the whole blood derived DNA samples, we excluded the possibility of overestimation of the *de novo* SNV rate due to cell line derived mutations. To note that, although our estimate is not interfered with the cell line derived mutations, it

might reflect an underestimation. The reason is the probability of observing the very low-frequency mosaic SNVs that occurred in later stages of the embryo development as mapping and variants calling errors.[109]

In our study, classification of the putative *de novo* SNVs regarding genomic context revealed a biased distribution towards non-coding regions of the genome. This is expected because (a) non-coding regions constitute the great proportion of the human genome and (b) the evolutionary pressure acting on *de novo* mutations is less stringent. We did not observe clustering of the *de novo* mutations at any region of the genome selectively. In previous studies, hotspots for *de novo* mutations in individuals affected with ASD have been documented.[44] In the context of our study, identification of the hotspots for the early post-zygotic and pre-zygotic *de novo* mutations in the genomes of the healthy individuals could be the subject of the further studies.

Previous publications on *de novo* rate in humans did not evaluate the possibility of mosaicism in parents which could be interpreted as *de novo* in the offspring.[3, 13] Although it is highly unlikely that the *de novo* mutations presented in our study have occurred in the soma of the parents, we still evaluated the possibility of mosaicism in the parents. To resolve this issue, we examined the buccal smear (ectoderm) and urine samples (endoderm) of the parents. Consequently, we observed that the parental mosaicism did not contribute to the *de novo* SNVs in the twin pair.

We observed that the early post-zygotic *de novo* SNVs was present in a mosaic state in the genomes of the twins according to the Sanger sequencing chromatograms. Moreover, capillary sequencing chromatograms showed the different levels of mosaicism for each of the post-zygotic *de novo* SNVs which might be the consequence of the timing of the mutagenesis. In our view, this finding could be critically important for the research in the field of medical genetics which is concerned with the delineation of the molecular basis of the complex diseases in human.

Contribution of the genetic heterogeneity to disease pathogenesis has been well-documented. Thus, assessment of the phenotypic, locus and allelic heterogeneity have been important for the studies concerned with the disease gene discovery.[110] Despite the success of the next generation sequencing technologies in the identification of *de novo* mutations that are causal for common neurodevelopmental diseases, there is still numerous complex phenotypes that remained unresolved.[42, 111] At this point, timing of *de novo* mutations and thus resulting somatic mosaicism might be an added level of complexity. This notion might also be supported by the involvement of the somatic mosaicism in different single gene disorders since the complex phenotypes have been anticipated to result from mutations which represent a single event of large effect.[112,113, 114]

Our findings are important in terms of being the first direct estimation of the early post-zygotic *de novo* SNV rate within the overall *de novo* SNV burden. Moreover, these results reiterate the importance of timing of mutagenesis and somatic mosaicism present in the human genome. Further work will be required to expand the scope of these studies through single cell sequencing and increasing the sample size for better understanding of the patterns and rates of early post-zygotic mutations.

Examination and documentation of the incidental findings derived from the whole genome sequencing data have an important clinical significance. [80,81] As a second part of our study, we evaluated the incidental findings present in the whole genome sequencing data of the four individuals (the twin pair and their parents). We intended to document the LoF mutations and potentially damaging missense mutations of the healthy individuals.

It has been reported that human genome contains approximately 100 LoF variants.[97] Here we identified a total of 75 LoF mutations present in the genomes of the twin pair, 48 LoF mutations in mother and 57 LoF mutations in father. 9 of the LoF mutations were located on genes that are implicated in the pathogenesis of 8 different diseases. Moreover, we identified damaging missense mutations occurred

on disease genes as well as pathogenic variants on the genes that do not have any clinical significance. Despite the possibility of reduction in power in our analysis due to the small sample size, examination of the genotypes of a quad family led us to reduce the false positive rate.

References

- [1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, “Molecular Biology of the Cell,” Fifth edition, Newyork, Garland Science, 2007.
- [2] A. Sancar, L. A. Lindsey-Boltz, K. Unsal-Kaçmaz, S. Linn, “Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints,” *Annu Rev Biochem*, vol. 73, pp. 39-85, 2004.
- [3] C. D. Campbell and E. E. Eichler, “Properties and rates of germline mutations in humans,” *Trends Genet*, vol. 29, pp. 575-84, 2013.
- [4] A. Poduri, G. D. Evrony, X. Cai, C. A. Walsh, “Somatic mutation, genomic variation, and neurological disease,” *Science*, vol. 341, pp. 1237758, 2013.
- [5] J. R. Lupski, “New mutations and intellectual function,” *Nat Genet*, vol. 42, pp. 1036-8, 2010.
- [6] L. Vadlamudi, L. M. Dibbens, K. M. Lawrence, X. Iona, J. M. McMahon, W. Murrell, *et al.*, “Timing of De Novo Mutagenesis — A Twin Study of Sodium-Channel Mutations,” *N Engl J Med*, vol. 30, pp. 1335-40, 2010.
- [7] J. R. Lupski, J. W. Belmont, E. Boerwinkle, R. A. Gibbs, “Clan genomics and the complex architecture of human disease,” *Cell*, vol. 147, pp. 32-43, 2011.

- [8] J. F. Crow, "The origins, patterns and implications of human spontaneous mutation," *Nat Rev Genet*, vol. 1, pp. 40-7, 2000.
- [9] A. Eyre-Walker and P. D. Keightley, "The distribution of fitness effects of new mutations," *Nat Rev Genet*, vol. 8, pp. 610-8, 2007.
- [10] H. Ellegren, N. G. Smith, M. T. Webster, "Mutation rate variation in the mammalian genome," *Curr Opin Genet Dev*, vol. 13, pp. 562-8, 2003.
- [11] C. Coulondre, J. H. Miller, P. J. Farabaugh, W. Gilbert, "Molecular basis of base substitution hotspots in *Escherichia coli*," *Nature*, vol. 274, pp. 775-80, 1978.
- [12] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, *et al.*, "Rate of *de novo* mutations and the importance of father's age to disease risk," *Nature*, vol. 488, pp. 471-75, 2012.
- [13] C. D. Campbell, J. X. Chong, M. Malig, A. Ko, B. L. Dumont, L. Han, *et al.*, "Estimating the human mutation rate using autozygosity in a founder population," *Nat Genet*, vol. 44, pp. 1277-80, 2012.
- [14] J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, S. R. Sunyaev, "Human mutation rate associated with DNA replication timing," *Nat Genet*, vol. 41, pp. 393-5, 2009.
- [15] P. C. Hanawalt and G. Spivak, "Transcription-coupled DNA repair: two decades of progress and surprises," *Nat Rev Mol Cell Biol*, vol. 9, pp. 958-70, 2008.
- [16] D. R. Schrider, J. N. Hourmozdi, M. W. Hahn, "Pervasive multinucleotide mutational events in eukaryotes," *Curr Biol*, vol. 21, pp. 1051-4, 2011.
- [17] J. R. Lupski, "Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits," *Trends Genet*, vol. 14, pp. 417-22, 1998.

- [18] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, *et al.*, “Origins and functional impact of copy number variation in the human genome,” *Nature*, vol. 464, pp. 704-12, 2010.
- [19] R. L. Glaser, J. P. Ramsay, I. M. Morison, “The imprinted gene and parent-of-origin effect database now includes parental origin of *de novo* mutations,” *Nucleic Acids Res*, vol. 34, pp. D29-31, 2006.
- [20] L. D. Hurst and H. Ellegren, “Sex biases in the mutation rate,” *Trends Genet*, vol. 14, pp. 446-52, 1998.
- [21] J. B. Haldane, “The mutation rate of the gene for haemophilia, and its segregation ratios in males and females,” *Ann Eugen*, vol. 13, pp. 262-71, 1947.
- [22] J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdóttir, H. Li, S. Mallick, *et al.*, “A direct characterization of human mutation based on microsatellites,” *Nat Genet*, vol. 44, pp. 1161-5, 2012.
- [23] N. S. Thomas, M. Durkie, B. Van Zyl, R. Sanford, G. Potts, S. Youings, *et al.*, “Parental and chromosomal origin of unbalanced *de novo* structural chromosome abnormalities in man,” *Hum Genet*, vol. 119, pp. 444-50, 2006.
- [24] S. I. Nagaoka, T. J. Hassold, P. A. Hunt, “Human aneuploidy: mechanisms and new insights into an age-old problem,” *Nat Rev Genet*, vol. 13, pp.493-504, 2012.
- [25] J. A. Veltman and H. G. Brunner, “*De novo* mutations in human genetic disease,” *Nat Rev Genet*, vol. 13, pp. 565-75, 2012.
- [26] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, *et al.*, “An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people,” *Science*, vol. 337, pp. 100-4, 2012.
- [27] A. Keinan and A. G. Clark, “Recent explosive human population growth has resulted in an excess of rare genetic variants,” *Science*, vol. 336, pp.740-3, 2012.

- [28] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, *et al.*, "Evolution and functional impact of rare coding variants from deep sequencing of human exomes," *Science*, vol. 337, pp. 64-9, 2012.
- [29] G. Allen, "Aetiology of Down's syndrome inferred by Waardenburg in 1932," *Nature*, vol. 250, pp. 436-7, 1974.
- [30] F. Zhang, W. Gu, M. E. Hurles, J. R. Lupski, "Copy number variation in human health, disease, and evolution," *Annu Rev Genomics Hum Genet*, vol. 10, pp. 451-81, 2009.
- [31] S. Girirajan, C. D. Campbell, and E. E. Eichler, "Human copy number variation and complex genetic disease," *Annu Rev Genet*, vol. 45, pp. 203-26, 2011.
- [32] L. E. Vissers, C. M. van Ravenswaaij, R. Admiraal, J. A. Hurst, B. B. de Vries, I. M. Janssen, *et al.*, "Mutations in a novel member of the chromodomain gene family cause CHARGE syndrome," *Nat Genet*, vol. 36, pp. 955-7, 2004.
- [33] E. H. Jr Cook and S. W. Scherer, "Copy-number variations associated with neuropsychiatric conditions," *Nature*, vol. 455, pp. 919-23, 2008.
- [34] L. E. Vissers, B. B. de Vries, and J. A. Veltman, "Genomic microarrays in mental retardation: from CNV to gene, from research to diagnosis," *J Med Genet*, vol. 47, pp. 289-97, 2010.
- [35] C. R. Marshall, A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, *et al.*, "Structural variation of chromosomes in autism spectrum disorder," *Am J Hum Genet*, vol. 82, pp. 477-88, 2008.
- [36] B. Xu, J. L. Roos, S. Levy, E. J. van Rensburg, J. A. Gogos, M. Karayiorgou, "Strong association of *de novo* copy number mutations with sporadic schizophrenia," *Nat Genet*, vol. 40, pp. 880-5, 2008.

- [37] B. B. de Vries, R. Pfundt, M. Leisink, D. A. Koolen, L. E. Vissers, I. M. Janssen, et al., "Diagnostic genome profiling in mental retardation," *Am J Hum Genet*, vol. 77, pp. 606-16, 2005.
- [38] A. Hoischen, B. W. van Bon, C. Gilissen, P. Arts, B. van Lier, M. Steehouwer, et al., "De novo mutations of *SETBP1* cause Schinzel-Giedion syndrome," *Nat Genet*, vol. 42, pp. 483-5, 2010.
- [39] S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, et al., "Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome," *Nat Genet*, vol. 42, pp. 790-3, 2010.
- [40] A. Hoischen, B. W. van Bon, B. Rodríguez-Santiago, C. Gilissen, L. E. Vissers, P. de Vries, et al., "De novo nonsense mutations in *ASXL1* cause Bohring-Opitz syndrome," *Nat Genet*, vol. 43, pp. 729-31, 2011.
- [41] A. Sirmaci, M. Spiliopoulos, F. Brancati, E. Powell, D. Duman, A. Abrams, et al., "Mutations in *ANKRD11* cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia," *Am J Hum Genet*, vol. 89, pp. 289-94, 2011.
- [42] I. Iossifov, M. Ronemus, D. Levy, Z. Wang, I. Hakker, J. Rosenbaum, et al., "De novo gene disruptions in children on the autistic spectrum," *Neuron*, vol. 74, pp. 285-99, 2012.
- [43] S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, et al., "De novo mutations revealed by whole-exome sequencing are strongly associated with autism," *Nature*, vol. 485, pp. 237-41, 2012.
- [44] J. J. Michaelson, Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin, et al., "Whole-genome sequencing in autism identifies hot spots for de novo germline mutation," *Cell*, vol. 151, pp. 1431-42, 2012.

- [45] J. B. S. Haldane, "The rate of spontaneous mutation of a human gene," *J Genet*, vol. 31, pp. 317-26, 1935.
- [46] M. W. Nachman, "Haldane and the first estimates of human mutation rate," *J Genet*, vol. 83, pp. 231-3, 2004.
- [47] D. N. Cooper and M. Krawczak, "Human gene mutation," BIOS Scientific Publishers, Oxford, 1993.
- [48] M. W. Nachman and S. L. Crowell, "Estimate of the mutation rate per nucleotide in humans," *Genetics*, vol. 156, pp. 297-304, 2000.
- [49] M. Kimura, "Evolutionary rate at the molecular level," *Nature*, vol. 217, pp. 624-6, 1968.
- [50] J. B. S. Haldane, "The causes of evolution," Longmans, Green and Co., London, 1932.
- [51] D. N. Cooper and M. Krawczak, "The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions," *Hum Genet*, vol. 85, pp. 55-74, 1990.
- [52] F. Vogel and A. G. Motulsky, "Human genetics: problems and approaches," Springer-Verlag, Berlin, 1997.
- [53] A. S. Kondrashov and J. F. Crow, "A molecular approach to estimating the human deleterious mutation rate," *Hum Mutat*, vol. 2, pp. 229-34, 1993.
- [54] Y. Xue, Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton, et al., "Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree," *Curr Biol*, vol. 19, pp. 1453-7, 2009.
- [55] J. C. Roach, G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, et al., "Analysis of genetic inheritance in a family quartet by whole-genome sequencing," *Science*, vol. 328, pp. 636-9, 2010.

- [56] D. F. Conrad, J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals *et al.*, “Variation in genome-wide mutation rates within and between human families,” *Nat Genet*, vol. 43, pp. 712-4, 2011.
- [57] A. S. Kondrashov, “Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases,” *Hum Mut*, vol. 21, pp. 12-27, 2003.
- [58] P. D. Keightley, “Rates and Fitness Consequences of New Mutations in Humans,” *Genetics*, vol.190, pp. 295-304, 2012.
- [59] M. Lynch, “Rate, molecular spectrum, and consequences of human mutation,” *Proc Natl Acad Sci U S A*, vol. 107, pp. 961-8, 2010.
- [60] J. R. Lupski, “Genomic rearrangements and sporadic disease,” *Nat Genet*, vol. 39, pp. S43-7, 2007.
- [61] A. Itsara, H. Wu, J. D. Smith, D. A. Nickerson, I. Romieu, S. J. London, and E. E. Eichler, “*De novo* rates and selection of large copy number variation,” *Genome Res*, vol. 20, pp. 1469-81, 2010.
- [62] T. Iwama, “Somatic mutation rate of the APC gene,” *Jpn J Clin Oncol*, vol. 31, pp. 185-7, 2001.
- [63] C. Hornsby, K. M. Page, I. Tomlinson, “The in vivo rate of somatic adenomatous polyposis coli mutation,” *Am J Pathol*, vol. 172, pp. 1062-8, 2008.
- [64] Z. Kan, B. S. Jaiswal, J. Stinson, V. Janakiraman, D. Bhatt, H. M. Stern, *et al.*, “Diverse somatic mutation patterns and pathway alterations in human cancers,” *Nature*, vol. 466, pp. 869-73, 2010.
- [65] R. Li, A. Montpetit, M. Rousseau, S. Y. M. Wu, C. M. T. Greenwood, T. D. Spector, *et al.*, “Somatic point mutations occurring early in development: a monozygotic twin study,” *J Med Genet*, vol. 51, pp. 28-34, 2014.
- [66] T. Strachan and A. P. Read, “Human molecular genetics,” Garland, Science, 2011.

- [67] L. G. Biesecker, N. B. Spinner, "A genomic view of mosaicism and human disease," *Nat Rev Genet*, vol. 14, pp. 307-20, 2013.
- [68] P. J. Zwijnenburg, M. Meijers Heijboer, and Boomsma, D. I., "Identical but not the same: the value of discordant monozygotic twins in genetic research," *Am. J. Med. Genet*, vol. 153, pp. 1134–49 ,2010.
- [69] E. A. Ehli, A. Abdellaoui, Y. Hu, J. J. Hottenga, M. Kattenberg, T. van Beijsterveldt, *et al.*, "De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems," *Eur J Hum Genet*, vol. 20, pp. 1037-40, 2012.
- [70] R. P. Erickson, "Somatic gene mutation and human disease other than cancer: An update," *Mutat Res*, vol. 705, pp. 96-106, 2010.
- [71] M. Topçu, C. Akyerli, A. Sayi, G. A. Törüner, S. R. Koçoğlu, M. Cimbiş, and T. Özçelik, "Somatic mosaicism for a MECP2 mutation associated with classic Rett syndrome in a boy," *Eur J Hum Genet*, vol. 10, pp. 77-81, 2002.
- [72] S. A. Frank, "Evolution in health and medicine Sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration," *Proc Natl Acad Sci U S A*, vol.107, pp. 1725-30, 2010.
- [73] M. O'Huallachain, K. J. Karczewski, S. M. Weissman, A. E. Urban, and M. P. Snyder, "Extensive genetic variation in somatic human tissues," *Proc Natl Acad Sci U S A*, vol. 109, pp. 18018-23, 2012.
- [74] J. K. Baillie, M. W. Barnett, K. R. Upton, D. J. Gerhardt, T. A. Richmond, F. De Sapio, *et al.*, "Somatic retrotransposition alters the genetic landscape of the human brain," *Nature*, vol. 479, pp. 534-7, 2011.
- [75] S. D. Kahn, "On the future of genomic data," *Science*, vol. 331, pp. 728-9, 2011.

- [76] K. M. Boycott, M. R. Vanstone, D. E. Bulman and A. E. MacKenzie, "Rare-disease genetics in the era of next-generation sequencing: discovery to translation," *Nat Rev Genet*, vol. 14, pp. 681-91, 2013.
- [77] G. Pichert, S. N. Mohammed, J. W. Ahn, C. M. Ogilvie, L. Izatt, "Unexpected findings in cancer predisposition genes detected by array comparative genomic hybridisation: what are the issues?," *J Med Genet*, vol. 48, pp. 535-9, 2011.
- [78] G. Crawford, N. Foulds, A. Fenwick, N. Hallowell and A. Lucassen, "Genetic medicine and incidental findings: it is more complicated than deciding whether to disclose or not," *Genet Med*, vol. 15, pp. 896-9, 2013.
- [79] A. N. Mayer, D. P. Dimmock, M. J. Arca, D. P. Bick, J. W. Verbsky, E. A. Worthey, *et al.*, "A timely arrival for genomic medicine," *Genet Med*, vol. 13, pp. 195-6, 2011.
- [80] J. Couzin-Frankel, "What would you do?," *Science*, vol. 331, pp. 662-5, 2011.
- [81] R. C. Green, J. S. Berg, W. W. Grody, S. S. Kalia, B. R. Korf, C. L. Martin, *et al.*, "ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing," *Genet Med*, vol. 15, pp. 565-74, 2013.
- [82] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, *et al.*, "The Variant Call Format and VCFtools," *Bioinformatics*, vol. 27, pp. 2156-8, 2011.
- [83] M. L. Metzker, "Sequencing technologies - the next generation," *Nat Rev Genet*, vol. 11, pp. 31-46, 2010.
- [84] F. Sanger, S. Nicklen, A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci. USA*, vol. 74, pp. 5463-7, 1977.
- [85] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921, 2001.

- [86] O. Morozova and M. A. Marra, “Applications of next-generation sequencing technologies in functional genomics,” *Genomics*, vol. 92, pp. 255-64, 2008.
- [87] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, et al., “The Genome Analysis Toolkit Best Practices Pipeline,” *Current Protocols in Bioinformatics*, vol. 43, pp. 11.10.1-11.10.33, 2013.
- [88] H. Li and L. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler Transform,” *Bioinformatics*, vol. 25, pp. 1754-60, 2009.
- [89] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; and 1000 Genome Project Data Processing Subgroup, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, pp. 2078-9, 2009.
- [90] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, et al., “A framework for variation discovery and genotyping using next-generation DNA sequencing data,” *Nat Genet*, vol. 43, pp. 491-8, 2011.
- [91] P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, et al., “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3,” *Fly (Austin)*, vol. 6, pp. 80-92, 2012.
- [92] P. Cingolani, V. M. Patel, M. Coon, T. Nguyen, S. J. Land, D. M. Ruden, et al., “Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift,” *Front Genet*, vol. 15, pp. 3, 2012.
- [93] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, pp. 841-2, 2010.
- [94] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, “Integrative Genomics Viewer,” *Nat Biotechnol*, vol. 29, pp. 24-6, 2011.

- [95] T. Koressaar T and M. Remm, "Enhancements and modifications of primer design program Primer3," *Bioinformatics*, vol. 23, pp. 1289-91, 2007.
- [96] O. J. Marshall, "PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR," *Bioinformatics*, vol. 20, pp. 2471-7, 2004.
- [97] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, *et al.*, "A systematic survey of loss-of-function variants in human protein-coding genes," *Science*, vol. 335, pp. 823-8, 2012.
- [98] P. Kumar, S. Henikoff, P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nat Protoc*, vol. 7, pp. 1073-81, 2009.
- [99] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, *et al.*, "A method and server for predicting damaging missense mutations," *Nat Methods*, vol. 7, pp. 248-9, 2010.
- [100] J. M. Schwarz, D. N. Cooper, M. Schuelke, D. Seelow, "MutationTaster2: mutation prediction for the deep-sequencing age," *Nat Methods*, vol. 11, pp. 361-2, 2014.
- [101] J. Reumers, P. De Rijk, H. Zhao, A. Liekens, D. Smeets, J. Cleary, *et al.*, "Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing," *Nat Biotechnol*, vol. 30, pp. 61-8, 2011.
- [102] B. T. Darras and U. Francke, "A partial deletion of the muscular dystrophy gene transmitted twice by an unaffected male," *Nature*, vol. 329, pp. 556-8, 1987.
- [103] T. Özçelik, "Uncovering the complex mysteries of mosaicism," *Nature*, vol. 417, pp. 588, 2002.
- [104] S. F. Gilbert, "Developmental Biology, 8th edition," Sinauer Associates Inc., 2003.

- [94] R. E. Handsaker, J. M. Korn, J. Nemesh, S. A. McCarroll, "Discovery and genotyping of genome structural polymorphism by sequencing on a population scale," *Nat Genet*, vol. 43, pp. 269-76, 2011.
- [106] G. M. Dal, B. Ergüner, M. S. Sağıroğlu, B. Yüksel, O. E. Onat, C. Alkan, and T. Özçelik, "Early postzygotic mutations contribute to *de novo* variation in a healthy monozygotic twin pair," *J Med Genet*, vol.51, pp. 455-9, 2014.
- [107] J. G. Hall, "Twinning: mechanisms and genetic implications," *Curr Opin Genet Dev*, vol. 6, pp. 343-7, 1996.
- [108] A. Koren, P. Polak, J. Nemesh, J. J. Michaelson, J. Sebat, S. R. Sunyaev, S. A. McCarroll, "Differential relationship of DNA replication timing to different forms of human mutation and variation," *Am J Hum Genet*, vol. 9, pp. 1033-40, 2012.
- [109] S. E. Baranzini, J. Mudge, J. C. van Velkinburgh, P. Khankhanian, I. Khrebtukova, N. A. Miller, "Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis," *Nature*, vol. 464 pp. 1351-56, 2010.
- [110] J. McClellan and M. C. King, "Genetic heterogeneity in human disease," *Cell*, vol. 141, pp. 210-17, 2010.
- [111] S. Gulsuner, T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornto, S. Casadei, *et al.*, "Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network," *Cell*, vol. 154, pp. 518-29, 2013.
- [112] B. T. Darras and U. Francke, "A partial deletion of the muscular dystrophy gene transmitted twice by an unaffected male," *Nature*, vol. 329, pp. 556-58, 1987.
- [113] C. Stern, "Genetic mosaics in animals and man. Stern C. Genetic mosaics and other essays," Harvard University Press, Cambridge, Massachusetts, pp. 27–129, 1968.

[114] J. R. Lupski, “Genome mosaicism—one human, multiple genomes,” *Science*, vol. 341, pp. 358-9, 2013.

Appendices

Appendix A

Primer List

Table A.1: Primer pairs for validation of the group I and group II *de novo* SNVs

Type	Chr	Pos	Ref	Alt	Forward	Reverse	Product Size(bp)
Group I	1	1334722	A	C	ACTCCCCTTCGGCTTCTTC	CGCTAGGGACAGACATGCTT	249
Group I	1	165599592	G	C	GATGAGGGATGTCCAGAGGA	CACTCCCGAAGTGAATGGAT	487
Group I	1	165599593	T	G	GATGAGGGATGTCCAGAGGA	CACTCCCGAAGTGAATGGAT	487
Group I	1	245093862	G	A	AGGATATGCTTGGCCTCCTT	AGGAGTGAAGGTGCGTCTGT	386
Group I	2	3720268	T	C	N/A	N/A	
Group I	2	7991048	A	C	AGGGGCTTCTGCTTAGTTCC	CACTGCCTCGCAGTCTCA	209
Group I	2	30755585	G	A	TACCTGGCCCATCAGAAAAG	TTTCCCATGATCCTTGCTTC	489
Group I	2	72375921	G	A	CCGAGCTGGGCTATTGTATC	TCGCTTTGGACTCAAATTCA	203
Group I	2	72375922	G	A	CCGAGCTGGGCTATTGTATC	TCGCTTTGGACTCAAATTCA	203

Group I	2	114789320	A	T	GGGCATCTGGAAAGATGTGT	TTTCTTCCTCCCTCCCCTT	285
Group I	2	136342189	G	A	ACCTGACCTCGGTAGCATGT	CCAAGCTTTGGAAATTGTTAGG	371
Group I	2	197151126	T	A	TGCATTACCATGAAATGTTGAA	CATAGAAGGCAGATGGCAAA	275
Group I	2	197151129	A	T	TGCATTACCATGAAATGTTGAA	CATAGAAGGCAGATGGCAAA	275
Group I	3	3168559	T	C	GCGAGAAGGGCGTAATTA	GACGTCACCAACGCACAT	229
Group I	3	11943503	C	G	ATTCTACGGGCACTTCACA	GCTCGGCTGCATATTAGCTC	656
Group I	3	50652860	A	C	GATTCTCAGGTTGGGCTCAG	GGCTCCAGACAATAGCCTCA	497
Group I	3	197614409	C	G	CCTGACCTGAAGCAGAGC	TGGGAAACTGCTGACAGAAA	595
Group I	4	701615	C	T	GACCCTGGAGCAGTCTTGTG	CATGTCATGAGCCCTTCTCA	789
Group I	4	1089608	T	C	N/A	N/A	
Group I	4	71768172	A	G	GATTCAAGACGAGCGCTACC	GGCCTGGCTACTCACAACA	240
Group I	4	145566227	C	T	TTGCAGGGATCCAGTAAACC	CAGGTGGACGGAGAGAGAAG	332
Group I	5	11043438	T	G	AGGAATTGGGCCTCTTTTGT	GCGCAAATACAACCTACCAAAGTT	359
Group I	5	172006234	G	A	GGAACCTTCATGCCTTTGCTC	AGCAAGAAAGGGCTGACTGA	195
Group I	6	29101417	C	T	TGCCCCATTTGCCTAAGAGTC	TGGGAAGTGGACAACAGTAGTTT	497
Group I	6	104212493	C	A	AGGTTGCAGTGAGCCAAGAT	GCAGTATACCAAGGGATGTGG	248
Group I	6	162458275	T	C	GCCCTTTGCAACTCTACTCC	ATGAGCACATTTGGTGATGC	182
Group I	6	169662202	G	A	ATCCGGCAAGTATCCCCAGA	TGCGTCAGAAACAGCAAATC	254
Group I	7	459309	A	G	N/A	N/A	
Group I	7	9485054	T	G	GAAGGAGGGACATTATCAACCA	AAGAAGCACAGGAGTCATGGA	534
Group I	7	70309917	A	G	TCAACTGGAATGCTATTAGTCG	GTCTGGCACAACCAGACAAA	484
Group I	7	141955453	A	G	GGCAGGTTGGCTAATTTTAC	TGCTTTGGCCTTTAATCCAG	550
Group I	7	157192929	T	C	GAGCTGCAGTAACGGGAAAG	CGTACAAGTGCCTACACGAAG	155
Group I	7	158252534	G	A	ATCCACAGTCTCCAGGATGG	GTGCAGGACAGAGCAGATGA	963
Group I	8	119959498	G	A	TATTGGTCAGGTGGGTAGGG	TTGCTAGGGTGTTGGATCGT	186
Group I	9	32917739	C	T	CTCCCTGCACAAGACCAGAC	CATCATGAGAAGGTGCCAGA	587
Group I	9	140206420	C	T	CTGCCTGATACCCTGAGACC	CAGGTAGCTGGGAGCCTTC	421
Group I	10	7400806	C	T	CTGTATGGCCATGGAAGG	TCAGGGCATACTCCATTCT	383
Group I	10	48410093	A	G	N/A	N/A	
Group I	10	102527333	C	T	TTCCCGAGACAGGACAATTC	TGTGCTGTTGCTGGAAAGTC	270
Group I	10	115563965	T	C	CTCCAAGTTTGCACATCACG	CACCCTGGAATGTTTGCTCT	506
Group I	10	128805987	G	A	GTGATGTTAGCAAGCCTGGA	TTTCCCGCAAACAATTAAGC	269

Group I	10	132099886	T	G	N/A	N/A	
Group I	10	133937827	C	T	CGGCCACATCACCTACATC	GTGGGACAGGTGACTGCAT	185
Group I	10	134392000	G	A	AGGTGAAGGTGAGCAAGACC	CACATACTCTCCTCCCCTGAA	242
Group I	11	1415375	A	C	N/A	N/A	
Group I	11	134534416	T	A	CACATGAAGGTCACCCACAC	CCCTTTCAAACCACCACTGT	413
Group I	12	3527384	T	C	CAGCACCTGCTGTTCTGAAA	AACCAAGATTGGCACTGTCC	300
Group I	13	97503636	T	C	ATTGCGGTTCTTTCCCTCCT	CTGATGGGTGTGATCCCTCT	409
Group I	13	114510560	T	C	CCCACGGAGGTGTAAGATGT	ATGGGGATGAAGTGGATCTG	664
Group I	14	32048285	T	A	CGGGCAGCTTGTTGTAATTC	CCAGCACAAATTTGAGAGCA	355
Group I	15	48937494	C	T	GGCGAGCTCTCTGGGTTTAT	GCTCCGCGTCAGATGTTC	165
Group I	15	57953050	A	T	CTGCCTTCAGGGAAGCTATG	GCAGCAGACATGATTCTGGA	448
Group I	15	80788986	G	T	CATTCAATTGCCTGTCACCAG	AGGGTGCCTCGCATAACATC	419
Group I	16	59139115	A	G	TCTCCCATAACATTGCCTCA	CACAGTTTCATGGGAAACAGAA	496
Group I	16	87715100	C	T	ATTACCCATCGCAGGAGTCA	GGGGACACCTTCTGGAATA	678
Group I	16	88497459	G	A	GGCGAGAAGAGGAAGGAAGT	GACTCTTCGGCAGTTTCAGG	350
Group I	17	79276311	C	T	GCTTACTGGGCAGTCCTACG	AGGGAGATGGGAACCAGAGT	229
Group I	17	80942321	C	T	GAGCACACAGGAGGCCATA	ACGACCCTGTAGCCTCTCC	354
Group I	18	3412160	G	A	GGAAGGGTTCAGGAAACACA	CTGGGTTTCTCGATGAGGAG	304
Group I	18	32459844	T	C	GTTCCCTCACCAACCTCTGA	CAAGGCAGGCTCTTGATAGG	350
Group I	19	523438	C	T	TACGCCAGAGCATTTCAGTG	CAGGATTTAAGCCCACCTCA	925
Group I	19	14732346	G	A	TGGCACTTTTACTCTTGTTGAGA	CAGTTAGGTAGCTTTAGGGCAAT	596
Group I	19	49842785	A	C	TCGGCCCTAAATCCCTAGAT	ACTTCCAGTGTGAGGCTGGT	487
Group I	19	53659085	C	T	CAGGGATGAATTCCTTGAA	CTGGGACTACAGGCATGAGC	476
Group I	20	14840274	T	C	ATCATGATTGCTGGTGTGGA	GCAAACCACAGACCATAGCA	489
Group I	20	50012140	C	G	TGTGACAAAGGCAAGTGCTC	CCTTTCCAGTTGCAGGCTTA	214
Group I	20	50690005	T	A	N/A	N/A	
Group I	20	62836674	T	C	AGATTCCTGGGCGGTAGATT	CCCTCCAGCTTCCTCTTCTT	472
Group I	20	62866083	T	G	N/A	N/A	
Group I	21	23811291	C	T	GCGGCTGTTGGTTAACAGAT	AAATTTACAGGGCCTTGTCCT	428
Group I	21	44743604	A	C	AGCAGGTCAGCAGTGGAAT	CTGTGGCAAAGTGAATGTGG	436
Group I	22	19167292	G	A	TTGTCGGCTAAAGCTGGTCT	GTTGTCACCACCCACCCTAC	206
Group I	X	2836434	G	A	CCTGGTGCCATTTTCCTAAA	AAGGGTTCTGACATGGATGC	228

Group II-Twin I	1	245499265	A	G	ACGCGGAGAAGGAGGAGT	TGCAGAGTGGAAACAGCAATC	388
Group II-Twin I	2	61728351	A	T	GGATGCAATTTGAGCCTTGA	TCCTTTGATTTCTCGTCCTT	490
Group II-Twin I	3	1493383	C	T	TCCATCCCTTTCTCTTCTCAA	GGAGAAGGCCAAACTAGAAAGTGTG	202
Group II-Twin I	3	51441939	A	C	CGTAGCTGCCACTCAAGACA	AAGTCCACAAAGCGGAGCTA	344
Group II-Twin I	3	72959480	G	A	CCAAGAGAAACCATGGAGGA	AGTTGCAGATGGGGAAACTG	484
Group II-Twin I	4	651520	G	A	AGATCCAGCTCATCCTGGTG	GTCCCTGCCACAGATCCAT	775
Group II-Twin I	4	6564000	G	A	CAGCAGGGAGGTAGGAAACA	AGTCCTCTTTGGCATGTGCT	261
Group II-Twin I	6	41586146	G	A	GAGCAGGGTGGACTGACTTC	CGATTCACTGGTGCTTGAGA	489
Group II-Twin I	6	134981681	T	G	CTGGCGAATTCAGTGTTCTG	TTAGGGATAGAAGCAAGGAACA	464
Group II-Twin I	7	29606885	G	A	TGGAAAGCAAATCACAGTGC	GTTCCAGTCCAGGTGTTGCT	377
Group II-Twin I	7	155510228	T	G	TCGTTGTCAATTCCACCTGA	GGGCTGAACTGTTTGCAACT	427
Group II-Twin I	7	157981209	G	C	CGGAAGGGATGTCCTAGTGC	CGGAGTGCTCTGGAAGCAG	693
Group II-Twin I	8	66686049	G	A	TTCCACATCACTGGATCAGG	GGCTGTGGGTAAGGAAATCA	499
Group II-Twin I	8	118838053	A	G	CCACCATATGCCACCTCACT	ATGCCAGGGTCAGGTTGTTA	529
Group II-Twin I	10	45869059	C	A	AACCTTCTCCACACCCTTCC	GAGCACAGAGGCTCCTTCAG	229
Group II-Twin I	12	653012	T	C	N/A	N/A	
Group II-Twin I	12	9557500	T	C	ACAGCCAGCAACTCTTCTGC	CCTGCGTTGGGTGTCTTA	443
Group II-Twin I	12	95569608	G	A	AACCAGGGTGGCTACATCAG	TTGACACCTTGATGTCTTCCAT	498
Group II-Twin I	12	131691875	G	A	ACCTGGAAGCTGTTGTGCGAG	CATACCCAATGATGCCTGTG	250
Group II-Twin I	12	133352895	G	C	GTGGACCAGGTGGCTTTG	GGGTACCTGTGTGAGGCTCT	387
Group II-Twin I	16	50424085	G	A	CCAAATCCCTGCACTGAACT	TAGTAATGCCCTGGCAGCTT	431
Group II-Twin I	18	34970824	C	T	TGATGGTGAGAGCACCTGTC	CCCTCTGCTCCATTCCATTA	398
Group II-Twin I	18	50453718	G	T	GCATTACTTG CATGAGGTCAAA	TGCACTGCCTGTAGTGGA	772
Group II-Twin I	20	53271335	C	T	TTTCAACGTCGCTGCAATAG	AGGACAGAAAGCAGCCTCAC	348
Group II-Twin I	X	2833489	C	G	N/A	N/A	
Group II-Twin I	X	7544103	T	C	N/A	N/A	
Group II-Twin I	X	144597121	C	A	N/A	N/A	
Group II-Twin I	X	148731301	C	A	ACCTTGGCCACATACTCCAG	AGACCCAGCAGACCTGAAGA	582
Group II-Twin I	Y	8094333	C	A	N/A	N/A	
Group II-Twin II	1	211950090	G	A	TCAAGAATGAGGGCTGCTTT	AGTGGGTGCCCAAAGTGTAG	287
Group II-Twin II	1	228146563	C	A	ACTGGGTTGGTAGCCCTTCT	CACTCCCTGGAATCGTGTTT	544
Group II-Twin II	2	70925457	G	T	GGCTGATGTTGGGAGGTCTA	TGAGCCCTCCTGTTCTCTTC	242

Group II-Twin II	2	241992247	C	T	CACGGATTTCTTCTGCCACT	GTTGCCGTGATACGTTTCCT	587
Group II-Twin II	6	133966995	G	T	CCCATGCTGTTTGAAGTTGA	CACATGGCCACATGTCTTCT	241
Group II-Twin II	7	158204835	C	T	N/A	N/A	
Group II-Twin II	9	14347587	G	A	GCTCAGGGCAAGCTAACAGT	TTACACTCCGCACAGTCACC	187
Group II-Twin II	11	1474891	G	T	GTCTCTTCATCGGGCAGAAC	CTCTAGCACAGCCCGAGACT	512
Group II-Twin II	12	132353193	A	C	GAGAAGGGCAAAGGGAAGTG	AGGGACCTGGGATGTGGT	223
Group II-Twin II	16	62015587	C	T	CACAGAGTCAGTTTCCCTGGAT	ATTCTCCATGCTCCCACATC	245
Group II-Twin II	21	36322701	A	G	CAGAATCACGGCTGAACTGA	AAGGCATTCCTTGTTAAGAGC	475
Group II-Twin II	21	47479749	C	T	N/A	N/A	
Group II-Twin II	X	7320881	C	G	TGGCCATAGCATTTAATCCA	GATGCCCCGAAAATTGTTAG	248
Group II-Twin II	X	71177022	T	C	ACCAAATCAGGCCACAGAAC	GAGGCAAAGTGCCAGGTAGA	213
Group II-Twin II	X	71177023	A	T	ACCAAATCAGGCCACAGAAC	GAGGCAAAGTGCCAGGTAGA	213
Group II-Twin II	X	104517151	A	G	TTCACCTGGCAAATCATTCA	AAAGTCTCCAACCAGGGACTC	164
Group II-Twin II	X	128927246	C	G	AAGAGGCCCAAGCCTAAGAC	TTTGATCCAAAGGAGGTTGG	568
Group II-Twin II	X	134054618	A	T	GACATGTTGCATTGCCAGTC	ACAGGGCAGTTTCTCTGTGG	389

Abbreviations used in this table: chr, chromosome; pos, position; ref, reference allele; alt, alternative allele; bp, base pair

Appendix B

Visual Examination of the BAM Files

Figure B.1: IGV screenshots for the total of 290 high-confidence putative *de novo* SNVs. The figure (in the format of pdf file) is available in our website <http://donut.cs.bilkent.edu.tr/twinstudy/> . The index table which shows the number of pages on the pdf file that corresponds to each of the examined variants is provided as Table B.1.

Table B.1: Index table for the Figure B.1

Group	Page	Chr	Pos	Ref	Alt	Gene	IGV prediction	Sanger	Twin I (n)		Twin II (n)		Mother (n)		Father (n)	
									Ref	Alt	Ref	Alt	Ref	Alt	Ref	Alt
Group I-Shared	1	1	1334722	A	C	<i>MRPL20</i>	<i>de novo</i>	N/A PCR product	5	4	3	3	6	0	8	0
Group I-Shared	2	1	38066423	G	A		low quality		41	21	32	13	22	6	33	9
Group I-Shared	3	1	47691626	G	C	<i>TAL1</i>	low quality		5	4	4	1	10	0	4	1
Group I-Shared	4	1	54706147	G	A	<i>C1orf191</i>	low quality		12	8	4	1	11	1	8	0
Group I-Shared	5	1	165599592	G	C		<i>de novo</i>	<i>de novo</i>	21	19	7	13	22	0	20	0
Group I-Shared	6	1	165599593	T	G		<i>de novo</i>	<i>de novo</i>	20	19	7	13	21	1	20	0
Group I-Shared	7	1	245093862	G	A		<i>de novo</i>	<i>de novo</i>	30	32	21	18	37	0	27	0
Group I-Shared	8	2	740342	T	C		low quality		13	8	11	1	13	0	5	1
Group I-Shared	9	2	3720195	T	C	<i>ALLC</i>	low quality		16	5	8	5	23	0	9	2
Group I-Shared	10	2	3720268	T	C	<i>ALLC</i>	<i>de novo</i>	N/A primer pair	20	18	12	4	30	2	15	0
Group I-Shared	11	2	7991048	A	C		<i>de novo</i>	no variant	46	14	28	17	34	4	23	3
Group I-Shared	12	2	10189061	G	A	<i>KLF11</i>	inherited		47	17	26	11	19	5	29	4
Group I-Shared	13	2	23053960	A	C		low quality		5	11	4	3	18	4	15	8
Group I-Shared	14	2	30755585	G	A	<i>LCLAT1</i>	<i>de novo</i>	inherited	12	16	21	30	6	0	31	0
Group I-Shared	15	2	72375921	G	A		<i>de novo</i>	uninformative	4	7	1	2	3	0	3	0
Group I-Shared	16	2	72375922	G	A		<i>de novo</i>	uninformative	4	7	1	2	3	0	3	0
Group I-Shared	17	2	114789320	A	T		<i>de novo</i>	N/A PCR product	23	11	16	4	32	0	12	2
Group I-Shared	18	2	136342189	G	A	<i>R3HDM1</i>	<i>de novo</i>	<i>de novo</i>	0	30	18		12	0	46	0
Group I-Shared	19	2	197151126	T	A	<i>HECW2</i>	<i>de novo</i>	no variant	27	10	27	6	19	2	40	0
Group I-Shared	20	2	197151129	A	T	<i>HECW2</i>	<i>de novo</i>	no variant	30	8	27	5	19	2	34	0
Group I-Shared	21	2	241565199	A	G	<i>GPR35</i>	low quality		12	4	5	2	18	4	4	1
Group I-Shared	22	3	3168559	T	C		<i>de novo</i>	uninformative	12	6	3	3	14	0	9	1
Group I-Shared	23	3	11943503	C	G	<i>NUP210P2</i>	<i>de novo</i>	no variant	39	12	19	3	18	0	11	0
Group I-Shared	24	3	13651179	T	G	<i>FBLN2</i>	no variant		9	5	6	3	12	1	4	0
Group I-Shared	25	3	50652860	A	C	<i>MAPKAPK3</i>	<i>de novo</i>	N/A PCR product	72	21	43	12	54	4	39	0
Group I-Shared	26	3	196189887	C	G		no variant		16	8	15	5	28	2	24	4
Group I-Shared	27	3	197614409	C	G	<i>LRCH3</i>	<i>de novo</i>	N/A PCR product	18	15	11	4	14	1	15	1
Group I-Shared	28	4	701615	C	T	<i>PCGF3</i>	<i>de novo</i>	N/A PCR product	15	6	7	5	11	1	4	1
Group I-Shared	29	4	1089608	T	C	<i>RNF212</i>	<i>de novo</i>	N/A primer pair	29	9	7	8	22	1	9	1
Group I-Shared	30	4	2062572	G	A	<i>NAT8L</i>	low quality		2	3	10	2	8	0	3	0

Group I-Shared	31	4	39264692	A	C	WDR19	low quality		10	6	5	4	6	2	2	3
Group I-Shared	32	4	71768172	A	G	MOB1B	de novo	uninformative	6	5	5	3	5	1	7	0
Group I-Shared	33	4	145566227	C	T	HHIP-AS1	de novo	de novo	23	32	13	14	25	0	27	0
Group I-Shared	34	4	190582718	G	A	RP11-462G22.1	inherited		155	34	92	20	141	21	131	0
Group I-Shared	35	4	190605292	A	T		inherited		109	18	87	17	69	9	108	0
Group I-Shared	36	5	1230092	G	T	SLC6A19	low quality		24	4	13	6	19	3	12	1
Group I-Shared	37	5	11043438	T	G	CTNND2	de novo	de novo	20	17	18	27	15	0	24	0
Group I-Shared	38	5	172006234	G	A		de novo	uninformative	35	7	15	4	19	0	13	0
Group I-Shared	39	6	29101417	C	T	OR2N1P	de novo	de novo	28	44	40	31	38	0	52	1
Group I-Shared	40	6	40598213	T	G		no variant		22	12	19	3	9	3	8	3
Group I-Shared	41	6	41472781	G	C	RP11-328M4.2	low quality+no variant		6	2	3	4	12	0	4	2
Group I-Shared	42	6	83777484	C	G	UBE3D	no variant		12	3	2	4	6	0	6	0
Group I-Shared	43	6	104212493	C	A		de novo	N/A PCR product	36	14	32	14	18	3	29	5
Group I-Shared	44	6	146092022	G	A		low quality		19	12	33	6	9	1	18	4
Group I-Shared	45	6	154769585	G	A	CNKSR3	low quality+no variant		39	14	23	13	21	5	30	8
Group I-Shared	46	6	162458275	T	C	PARK2	de novo	de novo	27	24	30	20	34	1	35	0
Group I-Shared	47	6	169662202	G	A		de novo	inherited	24	28	16	12	27	0	22	9
Group I-Shared	48	6	169662209	A	G		inherited		24	27	17	12	28	0	24	8
Group I-Shared	49	7	459309	A	G		de novo	N/A primer pair	11	3	4	5	12	0	46	0
Group I-Shared	50	7	9485054	T	G		de novo	inherited	13	17	18	26	4	1	36	0
Group I-Shared	51	7	70309917	A	G		de novo	inherited	14	15	12	12	9	0	16	2
Group I-Shared	52	7	81335206	T	G	HGF	inherited		40	13	35	25	13	4	22	4
Group I-Shared	53	7	140191996	T	C		low quality		20	11	11	12	22	2	20	0
Group I-Shared	54	7	141955453	A	G	PRSS58	de novo	de novo	32	40	18	17	28	0	33	0
Group I-Shared	55	7	157192929	T	C	DNAJB6	de novo	N/A PCR product	38	9	11	6	26	1	17	0
Group I-Shared	56	7	158037904	T	C	PTPRN2	no variant		12	4	3	1	17	0	18	0
Group I-Shared	57	7	158252534	G	A	PTPRN2	de novo	N/A PCR product	13	4	8	3	10	0	20	2
Group I-Shared	58	8	1266955	G	C		low quality		19	9	9	3	12	1	12	3
Group I-Shared	59	8	119959498	G	A	TNFRSF11B	de novo	N/A PCR product	12	10	15	4	18	1	30	0
Group I-Shared	60	8	142602829	G	A		low quality		0	##	11	4	31	0	7	2
Group I-Shared	61	8	144890487	G	A	MIR937	inherited		36	7	14	6	41	8	16	0
Group I-Shared	62	9	32504067	C	A	DDX58	inherited		33	15	12	1	25	1	16	0
Group I-Shared	63	9	32917739	C	T		de novo	de novo	41	47	12	21	37	0	31	0
Group I-Shared	64	9	88703373	T	A	GOLM1	inherited		20	7	16	4	7	0	22	2
Group I-Shared	65	9	102011855	G	A		inherited		53	21	33	13	35	10	29	0
Group I-Shared	66	9	102011909	A	G		inherited		65	11	37	8	51	7	26	0
Group I-Shared	67	9	138010720	T	G	OLFM1	low quality		33	10	12	8	16	3	17	1
Group I-Shared	68	9	139526735	C	G		low quality + no variant		14	3	9	7	18	2	13	2
Group I-Shared	69	9	140206420	C	T	EXD3	de novo	N/A PCR product	7	4	4	4	15	0	8	0

Group I-Shared	70	10	2357274	A	G		low quality		69	4	14	5	15	13	12	1
Group I-Shared	71	10	7400791	C	G	<i>SFMBT2</i>	inherited		37	11	35	8	33	0	19	2
Group I-Shared	72	10	7400796	G	A	<i>SFMBT2</i>	low quality		40	0	31	0	31	0	21	1
Group I-Shared	73	10	7400806	C	T	<i>SFMBT2</i>	<i>de novo</i>	inherited	87	9	41	5	12	24		0
Group I-Shared	74	10	48410093	A	G	<i>GDF2</i>	<i>de novo</i>	N/A primer pair	29	6	6	3	18	3	20	0
Group I-Shared	75	10	82233478	C	T	<i>TSPAN14</i>	low quality		47	15	22	13	20	8	21	4
Group I-Shared	76	10	88123119	G	T	<i>GRID1</i>	low quality		12	7	4	8	12	0	3	0
Group I-Shared	77	10	97753629	T	A	<i>CC2D2B</i>	inherited		25	4	42	11	13	0	19	2
Group I-Shared	78	10	102527333	C	T	<i>PAX2</i>	<i>de novo</i>	<i>de novo</i>	1146	##	19	19	43	0	27	0
Group I-Shared	79	10	115563965	T	C		<i>de novo</i>	uninformative	29	38	23	19	41	0	40	0
Group I-Shared	80	10	128805987	G	A	<i>DOCK1</i>	<i>de novo</i>	<i>de novo</i>	32	36	13	20	56	0	30	0
Group I-Shared	81	10	132099886	T	G		<i>de novo</i>	N/A primer pair	14	6	9	3	8	0	16	0
Group I-Shared	82	10	133937827	C	T	<i>JAKMIP3</i>	<i>de novo</i>	<i>de novo</i>	22	12	11	9	19	0	13	0
Group I-Shared	83	10	134392000	G	A	<i>INPP5A</i>	<i>de novo</i>	no variant	43	10	19	8	20	3	24	0
Group I-Shared	84	10	134837450	A	G		low quality		28	10	12	5	24	1	4	5
Group I-Shared	85	10	134996444	G	A	<i>KNDC1</i>	low quality		7	4	5	1	8	0	3	0
Group I-Shared	86	11	1415375	A	C	<i>BRSK2</i>	<i>de novo</i>	N/A primer pair	33	12	17	5	25	0	28	1
Group I-Shared	87	11	134534416	T	A		<i>de novo</i>	N/A PCR product	10	4	9	4	18	0	5	0
Group I-Shared	88	11	134605169	A	G		low quality		17	8	10	2	21	2	6	1
Group I-Shared	89	12	162687	G	A		inherited		139	35	88	28	118	0	78	11
Group I-Shared	90	12	718423	T	A	<i>NINJ2</i>	low quality		28	7	13	2	20	2	13	1
Group I-Shared	91	12	3527384	T	C	<i>PRMT8</i>	<i>de novo</i>	<i>de novo</i>	21	32	13	5	27	0	28	0
Group I-Shared	92	12	113724257	C	T	<i>TPCN1</i>	inherited		33	11	20	6	24	3	18	0
Group I-Shared	93	12	129572447	C	A	<i>TMEM132D</i>	low quality		27	7	18	6	12	0	15	3
Group I-Shared	94	12	131132579	T	C	<i>RIMBP2</i>	inherited		44	11	31	11	26	0	25	3
Group I-Shared	95	13	41929042	G	T	<i>NAA16</i>	inherited		32	20	18	13	11	12	15	11
Group I-Shared	96	13	97503636	T	C		<i>de novo</i>	<i>de novo</i>	38	30	19	33	42	0	37	0
Group I-Shared	97	13	112669460	C	T		low quality		29	7	14	3	12	0	10	0
Group I-Shared	98	13	114220087	G	C		low quality		14	6	9	2	18	1	13	1
Group I-Shared	99	13	114510530	C	T	<i>FAM70B</i>	low quality		14	6	7	2	11	0	11	0
Group I-Shared	100	13	114510547	C	G	<i>FAM70B</i>	low quality		13	7	8	4	6	0	10	0
Group I-Shared	101	13	114510560	T	C	<i>FAM70B</i>	<i>de novo</i>	N/A PCR product	14	8	6	4	5	0	6	0
Group I-Shared	102	13	114514458	T	C	<i>FAM70B</i>	low quality		9	3	2	1	3	1	5	0
Group I-Shared	103	13	114514468	G	A	<i>FAM70B</i>	low quality		7	3	4	1	4	1	6	0
Group I-Shared	104	13	114965572	C	T		inherited		7	3	11	4	8	2	10	0
Group I-Shared	105	14	32048285	T	A	<i>NUBPL</i>	<i>de novo</i>	<i>de novo</i>	39	33	25	20	45	0	29	0
Group I-Shared	106	14	105245986	A	G	<i>AKT1</i>	low quality		21	5	15	3	17	0	5	0
Group I-Shared	107	14	105345084	G	A	<i>KIAA0284</i>	inherited		20	7	19	7	15	0	15	0
Group I-Shared	108	14	105499889	A	G		low quality		13	7	6	1	12	2	6	0

Group I-Shared	109	15	41256583	A	C		low quality+no variant	5	5	3	3	11	1	1	2
Group I-Shared	110	15	48937494	C	T	<i>FBN1</i>	<i>de novo</i> inherited	7	4	10	6	14	0	3	1
Group I-Shared	111	15	57953050	A	T	<i>GCOM1</i>	<i>de novo</i> <i>de novo</i>	31	37	14	33	49	0	53	0
Group I-Shared	112	15	80788986	G	T	<i>ARNT2</i>	<i>de novo</i> <i>de novo</i>	26	32	10	10	50	0	36	0
Group I-Shared	113	16	10858705	T	A	<i>FAM18A</i>	low quality	19	6	4	2	24	0	14	0
Group I-Shared	114	16	22695512	T	G		inherited	123	34	57	27	75	11	64	1
Group I-Shared	115	16	59139115	A	G	<i>RP11-410D17.2</i>	<i>de novo</i> <i>de novo</i>	33	31	11	24	30	0	48	0
Group I-Shared	116	16	84085063	G	T	<i>MBTPS1</i>	low quality	65	16	27	9	43	4	36	0
Group I-Shared	117	16	87715100	C	T	<i>JPH3</i>	<i>de novo</i> no variant	15	12	4	6	15	0	12	1
Group I-Shared	118	16	88313175	A	G	<i>LA16c-444G7.1</i>	low quality	19	11	11	11	36	2	13	3
Group I-Shared	119	16	88497459	G	A	<i>ZNF469</i>	<i>de novo</i> no variant	8	4	6	4	17	0	3	1
Group I-Shared	120	16	88560062	G	A	<i>ZFPM1</i>	low quality	27	4	6	3	13	0	6	0
Group I-Shared	121	17	546943	T	C	<i>VPS53</i>	inherited	47	15	37	12	25	1	25	7
Group I-Shared	122	17	21303735	A	G	<i>KCNJ12</i>	inherited	95	36	42	11	70	9	59	0
Group I-Shared	123	17	21307296	G	C	<i>KCNJ12</i>	inherited	123	33	69	13	82	11	80	0
Group I-Shared	124	17	79276311	C	T	<i>LINC00482</i>	<i>de novo</i> <i>de novo</i>	12	22	2	5	24	0	8	0
Group I-Shared	125	17	79906331	C	T	<i>NOTUM</i>	inherited	47	13	27	3	39	5	20	0
Group I-Shared	126	17	80346681	C	G	<i>C17orf101</i>	low quality	4	4	1	2	4	0	2	2
Group I-Shared	127	17	80942321	C	T	<i>RP11-1197K16.2</i>	<i>de novo</i> N/A PCR product	37	14	21	9	24	0	21	2
Group I-Shared	128	17	81024707	C	T		inherited	12	11	4	2	19	6	10	0
Group I-Shared	129	18	3412160	G	A	<i>TGIF1</i>	<i>de novo</i> inherited	16	24	12	10	20	0	5	1
Group I-Shared	130	18	32459844	T	C	<i>DTNA</i>	<i>de novo</i> <i>de novo</i>	48	32	12	15		0	40	0
Group I-Shared	131	18	77204733	G	C	<i>NFATC1</i>	low quality	6	3	9	2	13	0	6	0
Group I-Shared	132	19	523438	C	T	<i>TPGS1</i>	<i>de novo</i> N/A PCR product	7	26	11	11	10	0	10	12
Group I-Shared	133	19	682321	T	A	<i>PRSS57</i>	inherited	16	14	12	4	14	6	23	0
Group I-Shared	134	19	11564606	A	C	<i>ELAVL3</i>	low quality + no variant	15	19	14	7	12	4	8	1
Group I-Shared	135	19	14732346	G	A	<i>EMR3</i>	<i>de novo</i> uninformative	44	12	12	4	24	1	31	0
Group I-Shared	136	19	39330862	A	G	<i>HNRNPL</i>	low quality	6	3	2	5	6	2	4	3
Group I-Shared	137	19	49842785	A	C	<i>TEAD2</i>	<i>de novo</i> no variant	6	11	7	4	11	8	9	2
Group I-Shared	138	19	50197909	T	G	<i>C19orf76</i>	low quality	31	20	15	14	17	10	17	7
Group I-Shared	139	19	53659085	C	T	<i>ZNF347</i>	<i>de novo</i> N/A PCR product	25	18	9	11	23	0	20	0
Group I-Shared	140	20	653248	C	T	<i>SCRT2</i>	low quality	12	9	4	6	14	0	12	3
Group I-Shared	141	20	14840274	T	C	<i>MACROD2</i>	<i>de novo</i> no variant	35	30	33	30	26	0	39	0
Group I-Shared	142	20	50012140	C	G	<i>NFATC2</i>	<i>de novo</i> no variant	13	5	16	4	14	1	20	1
Group I-Shared	143	20	50690005	T	A	<i>ZFP64</i>	<i>de novo</i> N/A primer pair	28	41	15	22	43	0	34	1
Group I-Shared	144	20	52184901	A	C	<i>ZNF217</i>	low quality	44	34	28	20	26	16	24	10
Group I-Shared	145	20	61607457	T	C		low quality	18	5	12	4	7	1	7	1
Group I-Shared	146	20	62836674	T	C	<i>MYT1</i>	<i>de novo</i> N/A PCR product	18	6	6	1	12	0	8	0
Group I-Shared	147	20	62866083	T	G	<i>MYT1</i>	<i>de novo</i> N/A primer pair	19	13	6	5	13	0	6	0

Group I-Shared	148	21	17686365	T	A	LINC00478	inherited		9	3	18	10	15	2	12	1
Group I-Shared	149	21	23811291	C	T		de novo	de novo	28	29	13	16	26	0	25	0
Group I-Shared	150	21	44743604	A	C	LINC00322	de novo	inherited	23	11	13	7	19	0	3	3
Group I-Shared	151	21	45206930	C	T	TMEM97P1	low quality		29	3	13	7	21	0	13	2
Group I-Shared	152	21	46304683	T	G	ITGB2	low quality		17	9	11	4	20	5	7	4
Group I-Shared	153	22	19167292	G	A	CLTCL1	de novo	de novo	24	25	13	13	25	0	13	0
Group I-Shared	154	22	19704712	T	G	SEPT5	low quality		5	5	1	2	4	2	114	13
Group I-Shared	155	22	40898746	T	C	MKL1	low quality		10	5	15	2	8	0	16	0
Group I-Shared	156	22	49761849	A	G		low quality		44	9	18	3	21	1	21	1
Group I-Shared	157	22	49840787	G	A	C22orf34	low quality		28	5	18	6	15	0	11	1
Group I-Shared	158	X	2836434	G	A	ARSD	de novo	no variant	16	7	19	7	35	0	16	2
Group I-Shared	159	X	145422663	T	C		low quality		12	6	7	2	11	0	5	0
Group II-Twin I	160	1	162730172	A	G	DDR2	inherited		36	9	46	1	17	0	29	1
Group II-Twin I	161	1	245499265	A	G	KIF26B	de novo	inherited	30	6	23	3	27	0	14	1
Group II-Twin I	162	2	61728351	A	T	XPO1	de novo	de novo	34	14	45	0	27	0	25	0
Group II-Twin I	163	2	111926337	T	G	BCL2L11	inherited		27	22	29	15	17	9	15	7
Group II-Twin I	164	2	241627105	T	G	AQP12A	low quality		9	7	3	0	9	0	7	1
Group II-Twin I	165	3	1493383	C	T		de novo	de novo	30	9	42	0	22	0	25	0
Group II-Twin I	166	3	20897127	T	C		low quality		27	12	16	2	24	1	17	3
Group II-Twin I	167	3	51090509	G	T	DOCK3	low quality		13	4	21	0	13	1	12	2
Group II-Twin I	168	3	51441939	A	C	VPRBP	de novo	no variant	34	20	28	6	293	0	17	9
Group II-Twin I	169	3	66701546	T	G		low quality		20	17	13	7	12	5	10	9
Group II-Twin I	170	3	72959480	G	A	GXYLT2	de novo	de novo	35	##	54	0	35	0	37	0
Group II-Twin I	171	3	135876070	C	T	TDGF1P6	low quality		33	8	29	1	22	0	18	2
Group II-Twin I	172	3	187442688	A	C	BCL6	low quality		29	7	16	1	30	0	29	1
Group II-Twin I	173	3	196430415	C	T	CEP19	low quality		35	11	35	1	20	1	17	1
Group II-Twin I	174	4	651520	G	A	PDE6B	de novo	N/A PCR product	11	4	7	2	13	0	8	0
Group II-Twin I	175	4	6564000	G	A	PPP2R2C	de novo	de novo	33	15	24	0	18	1	14	0
Group II-Twin I	176	4	7920679	G	A	AFAP1	low quality		55	14	34	0	26	2	22	1
Group II-Twin I	177	4	189032254	C	G	RP11-713C19.1	low quality		25	11	47	3	30	0	30	4
Group II-Twin I	178	4	189792261	C	G		low quality		1	3	46	3	42	1	69	7
Group II-Twin I	179	5	1956661	A	G	RP11-259O2.1	low quality		48	13	27	3	17	2	786	72
Group II-Twin I	180	5	54561146	A	G	DHX29	inherited		62	14	43	4	30	1	36	3
Group II-Twin I	181	5	82195939	G	A		low quality		40	16	36	10	23	6	23	8
Group II-Twin I	182	5	141576736	A	C		inherited		41	23	32	6	35	10	26	11
Group II-Twin I	183	6	16943658	A	C		low quality		37	26	32	10	32	11	17	8
Group II-Twin I	184	6	20235579	T	G	RP11-239H6.2	low quality		17	7	7	5	9	3	7	1
Group II-Twin I	185	6	41586146	G	A		de novo	inherited	47	12	30	0	37	0	24	0
Group II-Twin I	186	6	134981681	T	G		de novo	no variant	38	16	35	2	23	10	22	3

Group II-Twin I	187	7	1824582	T	C		low quality			43	11	32	1	20	1	34	3
Group II-Twin I	188	7	1824583	G	A		inherited			44	11	34	1	20	1	33	3
Group II-Twin I	189	7	29606885	G	A	<i>PRR15</i>	<i>de novo</i>	<i>de novo</i>		59	23	44	0	31	0	40	0
Group II-Twin I	190	7	40103905	T	G	<i>CDK13</i>	low quality			36	18	31	5	25	4	17	9
Group II-Twin I	191	7	76922180	A	C	<i>CCDC146</i>	low quality			39	29	27	9	19	14	20	11
Group II-Twin I	192	7	155510228	T	G	<i>RBM33</i>	<i>de novo</i>	no variant		26	14	23	1	25	3	14	4
Group II-Twin I	193	7	157981209	G	C	<i>PTPRN2</i>	<i>de novo</i>	N/A PCR product		12	5	16	0	6	0	5	0
Group II-Twin I	194	8	66686049	G	A	<i>MTFR1</i>	<i>de novo</i>	uninformative		29	9	58	0	23	0	26	0
Group II-Twin I	195	8	118838053	A	G	<i>EXT1</i>	<i>de novo</i>	<i>de novo</i>			0		0		0		0
Group II-Twin I	196	8	143671111	G	C		low quality			34	8	17	4	35	0	16	0
Group II-Twin I	197	9	140816091	C	A	<i>CACNA1B</i>	inherited			39	16	27	7	22	2	21	4
Group II-Twin I	198	10	3940738	G	A		low quality			39	9	23	2	20	1	26	2
Group II-Twin I	199	10	10872261	G	A		low quality			26	6	31	3	17	2	29	4
Group II-Twin I	200	10	45869059	C	A		<i>de novo</i>	N/A PCR product		9	6	6	3	8	0	2	1
Group II-Twin I	201	10	48410018	G	C	<i>GDF2</i>	low quality			9	8	3	0	22	2	14	0
Group II-Twin I	202	10	77154164	C	T	<i>ZNF503</i>	low quality			24	14	23	5	19	9	17	5
Group II-Twin I	203	11	65374709	T	G	<i>MAP3K11</i>	low quality			10	8	4	1	9	0	4	1
Group II-Twin I	204	12	653012	T	C	<i>B4GALNT3</i>	<i>de novo</i>	N/A primer pair		23	7	8	1	17	0	9	0
Group II-Twin I	205	12	9557500	T	C	<i>RP11-599J14.2</i>	<i>de novo</i>	no variant		5	4	4	0	13	0	9	0
Group II-Twin I	206	12	64229411	T	G		low quality			18	15	27	6	24	9	21	19
Group II-Twin I	207	12	95569608	G	A	<i>FGD6</i>	<i>de novo</i>	uninformative		43	23	32	4	30	11	26	8
Group II-Twin I	208	12	131691875	G	A	<i>RP11-638F5.1</i>	<i>de novo</i>	inherited		31	10	12	1	27	0	10	0
Group II-Twin I	209	12	133052676	T	C		low quality			27	12	12	3	25	3	10	0
Group II-Twin I	210	12	133352895	G	C	<i>GOLGA3</i>	<i>de novo</i>	N/A PCR product		33	12	24	2	25	0	13	1
Group II-Twin I	211	13	51812216	T	G	<i>FAM124A</i>	low quality			18	16	15	7	14	6	13	4
Group II-Twin I	212	13	87083684	C	A		low quality			14	5	42	0	5	0	16	1
Group II-Twin I	213	14	52997360	A	T	<i>TXNDC16</i>	low quality			9	4	28	0	10	0	10	0
Group II-Twin I	214	14	80678369	C	T	<i>DIO2</i>	low quality			38	19	31	10	26	9	23	6
Group II-Twin I	215	14	100044559	A	G	<i>CCDC85C</i>	no variant			30	11	15	3	22	3	27	3
Group II-Twin I	216	16	1970303	A	G		low quality			20	15	4	2	15	3	12	0
Group II-Twin I	217	16	10858718	A	G	<i>FAM18A</i>	low quality			19	6	10	0	26	0	11	1
Group II-Twin I	218	16	50424085	G	A	<i>RP11-21B23.2</i>	<i>de novo</i>	<i>de novo</i>		26	10	22	0	21	0	14	0
Group II-Twin I	219	16	70714811	C	G	<i>MTSS1L</i>	low quality			14	8	6	2	8	0	9	2
Group II-Twin I	220	16	77724703	G	T		low quality			34	22	18	7	22	14	14	5
Group II-Twin I	221	17	92938	C	T	<i>RPH3AL</i>	low quality			15	7	38	1	25	1	59	0
Group II-Twin I	222	17	71734490	T	C	<i>CTD-2532D12.4</i>	inherited			28	9	20	2	28	4	20	0
Group II-Twin I	223	17	73824833	G	C	<i>UNK</i>	low quality			559	0	11	1	11	6	9	6
Group II-Twin I	224	18	34970824	C	T	<i>CELF4</i>	<i>de novo</i>	uninformative		27	6	9	0	16	0	11	0
Group II-Twin I	225	18	50453718	G	T	<i>DCC</i>	<i>de novo</i>	N/A PCR product		30	6	52	2	16	1	28	1

Group II-Twin I	226	19	309882	A	G	<i>CTD-3113P16.5</i>	no variant		13	14	29	0	5	1	3	0
Group II-Twin I	227	19	33105233	T	G	<i>ANKRD27</i>	low quality		42	36	36	0	20	13	21	4
Group II-Twin I	228	20	53271335	C	T	<i>DOK5</i>	de novo	de novo	55	17	37	0	39	0	28	0
Group II-Twin I	229	20	56815552	G	C	<i>PPP4R1L</i>	low quality		10	0	4	1	6	1	2	3
Group II-Twin I	230	20	62810452	T	C	<i>MYT1</i>	inherited		23	16	12	5	31	2	22	8
Group II-Twin I	231	21	34512162	C	A		low quality		28	25	24	8	21	6	14	7
Group II-Twin I	232	X	2833489	C	G	<i>ARSD</i>	de novo	N/A primer pair	21	7	15	1	30	0	14	0
Group II-Twin I	233	X	7544103	T	C		de novo	N/A primer pair	8	7	5	0	18	1	10	0
Group II-Twin I	234	X	7544110	T	C		low quality		0	3	8	0	17	2	10	0
Group II-Twin I	235	X	10107413	T	A	<i>WWC3</i>	low quality		18	10	9	3	61	6	7	0
Group II-Twin I	236	X	68384591	T	G	<i>PJA1</i>	low quality		3	7	4	1	3	2	3	0
Group II-Twin I	237	X	116523014	T	A		low quality		13	5	30	1	29	1	23	2
Group II-Twin I	238	X	144597115	T	A		low quality		2	3	14	0	7	0	12	0
Group II-Twin I	239	X	144597121	C	A		de novo	N/A primer pair	2	3	14	0	6	0	8	0
Group II-Twin I	240	X	148731301	C	A	<i>RP5-937E21.1</i>	de novo	no variant	16	7	18	1	26	6	15	1
Group II-Twin I	241	X	150242328	A	G		no variant		15	8	11	1	23	1	5	0
Group II-Twin I	242	Y	8094333	C	A		de novo	N/A primer pair	9	5	23	0	0	0	6	0
Group II-Twin II	243	1	211950090	G	A	<i>LPGAT1</i>	de novo	no variant	55	9	28	11	25	4	24	3
Group II-Twin II	244	1	228146563	C	A		de novo	no variant	37	0	15	5	37	0	17	0
Group II-Twin II	245	2	362130	G	T		inherited		70	8	36	10	37	3	58	0
Group II-Twin II	246	2	362131	A	T		inherited		71	7	35	12	37	3	58	0
Group II-Twin II	247	2	1404408	C	A	<i>TPO</i>	low quality		38	18	36	15	28	7	23	8
Group II-Twin II	248	2	10476221	G	T	<i>HPCAL1</i>	inherited		53	9	29	11	30	1	49	0
Group II-Twin II	249	2	70925457	G	T	<i>ADD2</i>	de novo	no variant	56	0	26	5	676	0	34	0
Group II-Twin II	250	2	154061708	G	T	<i>AC079150.2</i>	low quality		31	15	19	12	19	7	18	10
Group II-Twin II	251	2	241992247	C	T	<i>SNED1</i>	de novo	uninformative	34	1	15	7	13	1	17	1
Group II-Twin II	252	4	3827696	A	C		low quality		26	0	21	7	42	4	15	0
Group II-Twin II	253	5	194385	A	C	<i>PLEKHG4B</i>	low quality		59	10	22	10	27	2	23	1
Group II-Twin II	254	5	16525233	G	A	<i>FAM134B</i>	inherited		33	6	12	6	19	4	23	0
Group II-Twin II	255	5	137885670	T	A	<i>HSPA9</i>	inherited		66	0	44	12	42	0	50	2
Group II-Twin II	256	6	57541536	G	T		inherited		155	24	108	22	94	12	100	0
Group II-Twin II	257	6	86090624	G	A		low quality		36	8	25	12	23	2	23	9
Group II-Twin II	258	6	133966995	G	T		de novo	no variant	40	5	24	9	21	1	19	4
Group II-Twin II	259	7	10338952	G	T		low quality		63	12	37	15	36	6	36	3
Group II-Twin II	260	7	56559682	A	G	<i>RBM22P3</i>	low quality		46	2	15	6	25	3	18	0
Group II-Twin II	261	7	158204835	C	T	<i>PTPRN2</i>	de novo	N/A Primer Pair	17	1	4	3	14	0	8	0
Group II-Twin II	262	7	158248408	C	T	<i>PTPRN2</i>	low quality		31	5	26	5	23	2	23	0
Group II-Twin II	263	8	142602825	C	T		low quality		23	4	12	5	34	0	8	1
Group II-Twin II	264	9	14347587	G	A	<i>NFIB</i>	de novo	inherited	26	0	8	4	15	0	9	0

Group II-Twin II	265	9	106166434	C	T	<i>RP11-436F21.1</i>	low quality		20	4	26	8	8	0	21	0
Group II-Twin II	266	10	3357139	G	A		low quality		31	2	9	6	18	0	11	1
Group II-Twin II	267	10	30518280	A	G		low quality		44	9	22	11	27	9	17	8
Group II-Twin II	268	11	1474891	G	T	<i>BRSK2</i>	<i>de novo</i>	no variant	25	0	7	4	17	0	8	0
Group II-Twin II	269	11	129307937	T	A	<i>BARX2</i>	low quality		32	7	19	10	22	2	12	2
Group II-Twin II	270	12	132353193	A	C	<i>RP11-417L19.2</i>	<i>de novo</i>	N/A PCR product	14	3	6	9	14	2	12	4
Group II-Twin II	271	13	114617416	G	A	<i>LINC00452</i>	low quality		60	5	32	8	36	3	31	0
Group II-Twin II	272	13	114619650	A	G	<i>LINC00452</i>	low quality		46	3	14	4	29	0	26	0
Group II-Twin II	273	15	76884600	G	T	<i>SCAPER</i>	low quality		27	3	27	9	20	2	42	0
Group II-Twin II	274	16	62015587	C	T	<i>CDH8</i>	<i>de novo</i>	no variant	77	4	33	9	41	0	32	3
Group II-Twin II	275	17	4782495	T	G	<i>Metazoa_SRP</i>	low quality		35	7	8	6	25	0	15	1
Group II-Twin II	276	17	79646774	T	G	<i>ARL16</i>	low quality		37	17	12	8	21	9	15	7
Group II-Twin II	277	17	81042516	T	C	<i>METRNL</i>	low quality		33	5	12	5	23	0	16	3
Group II-Twin II	278	18	18623582	G	A	<i>ROCK1</i>	low quality		18	0	44	9	13	0	26	2
Group II-Twin II	279	21	36322701	A	G	<i>RUNX1</i>	<i>de novo</i>	<i>de novo</i>	63	0	33	9	27	0	41	0
Group II-Twin II	280	21	47479749	C	T	<i>AP001476.3</i>	<i>de novo</i>	N/A Primer Pair	64	2	10	5	24	1	26	4
Group II-Twin II	281	21	47479750	C	G	<i>AP001476.3</i>	low quality		162	1	10	5	24	0	28	1
Group II-Twin II	282	X	7320881	C	G		<i>de novo</i>	no variant	32	0	15	6	32	0	22	0
Group II-Twin II	283	X	71177022	T	C	<i>NHSL2</i>	<i>de novo</i>	no variant	27	0	14	4	25	0	17	0
Group II-Twin II	284	X	71177023	A	T	<i>NHSL2</i>	<i>de novo</i>	no variant	27	0	14	4	25	0	17	0
Group II-Twin II	285	X	104517151	A	G	<i>IL1RAPL2</i>	<i>de novo</i>	no variant	24	0	12	4	37	0	19	0
Group II-Twin II	286	X	122224237	A	G		no variant		27	2	8	6	21	1	13	0
Group II-Twin II	287	X	123370010	C	T	<i>STAG2</i>	low quality		28	7	13	6	31	7	16	1
Group II-Twin II	288	X	128927246	C	G	<i>SASH3</i>	<i>de novo</i>	no variant	13	0	10	4	17	0	17	0
Group II-Twin II	289	X	134054617	A	C		no variant		27	0	9	5	32	0	18	0
Group II-Twin II	290	X	134054618	A	T		<i>de novo</i>	no variant	28	0	8	5	32	0	17	0

Abbreviations used in this table: chr, chromosome; pos, position; ref, reference allele; alt, alternative allele; n, number of reads supporting the variant; N/A, not available

Appendix C

Complete List of Putative *De Novo* SNVs

Table C.1: List of the *de novo* SNV candidates (n=159) shared by the twin pair

Chr	Pos	Ref	Alt	Genomic Context	Gene	Gene type	Twin I	Twin II	Mother	Father	IGV Prediction	Sanger
1	1334722	A	C	DOWNSTREAM	<i>MRPL20</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
1	38066423	G	A	INTERGENIC			Het	Het	Ref	Ref	low quality	
1	47691626	G	C	DOWNSTREAM			Het	Het	Ref	Ref	low quality	
1	54706147	G	A	DOWNSTREAM	<i>TAL1</i>	processed_transcript	Het	Het	Ref	Ref	low quality	<i>de novo</i>
1	165599592	G	C	INTERGENIC	<i>C1orf191</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	
1	165599593	T	G	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	
1	245093862	G	A	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	
2	740342	T	C	INTERGENIC			Het	Het	Ref	Ref	low quality	

2	3720195	T	C	INTRON	<i>ALLC</i>	protein_coding	Het	Het	Ref	Ref	low quality	
2	3720268	T	C	INTRON	<i>ALLC</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
2	7991048	A	C	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	no variant
2	10189061	G	A	DOWNSTREAM	<i>KLF11</i>	protein_coding	Het	Het	Ref	Ref	inherited	
2	23053960	A	C	INTERGENIC			Het	Het	Ref	Ref	low quality	
2	30755585	G	A	INTRON	<i>LCLAT1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	inherited
2	72375921	G	A	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	uninformative
2	72375922	G	A	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	uninformative
2	114789320	A	T	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
2	136342189	G	A	INTRON	<i>R3HDM1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
2	197151126	T	A	INTRON	<i>HECW2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
2	197151129	A	T	INTRON	<i>HECW2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
2	241565199	A	G	INTRON	<i>GPR35</i>	protein_coding	Het	Het	Ref	Ref	low quality	
3	3168559	T	C	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	uninformative
3	11943503	C	G	DOWNSTREAM	<i>NUP210P2</i>	unprocessed_pseudogene	Het	Het	Ref	Ref	<i>de novo</i>	no variant
3	13651179	T	G	INTRON	<i>FBLN2</i>	protein_coding	Het	Het	Ref	Ref	no variant	
3	50652860	A	C	DOWNSTREAM	<i>MAPKAPK3</i>	processed_transcript	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
3	196189887	C	G	INTERGENIC			Het	Het	Ref	Ref	no variant	
3	197614409	C	G	DOWNSTREAM	<i>LRCH3</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
4	701615	C	T	INTRON	<i>PCGF3</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
4	1089608	T	C	DOWNSTREAM	<i>RNF212</i>	retained_intron	Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
4	2062572	G	A	INTRON	<i>NAT8L</i>	protein_coding	Het	Het	Ref	Ref	low quality	
4	39264692	A	C	INTRON	<i>WDR19</i>	protein_coding	Het	Het	Ref	Ref	low quality	
4	71768172	A	G	EXON	<i>MOB1B</i>	processed_transcript	Het	Het	Ref	Ref	<i>de novo</i>	uninformative
4	145566227	C	T	INTRON	<i>HHIP-AS1</i>	processed_transcript	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
4	190582718	G	A	DOWNSTREAM	<i>RP11-462G22.1</i>	lincRNA	Het	Het	Ref	Ref	inherited	
4	190605292	A	T	INTERGENIC			Het	Het	Ref	Ref	inherited	
5	1230092	G	T	DOWNSTREAM	<i>SLC6A19</i>	protein_coding	Het	Het	Ref	Ref	low quality	
5	11043438	T	G	INTRON	<i>CTNND2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
5	172006234	G	A	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	uninformative
6	29101417	C	T	DOWNSTREAM	<i>OR2N1P</i>	unprocessed_pseudogene	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
6	40598213	T	G	INTERGENIC			Het	Het	Ref	Ref	no variant	
6	41472781	G	C	INTRON	<i>RP11-328M4.2</i>	antisense	Het	Het	Ref	Ref	low quality+no variant	
6	83777484	C	G	UPSTREAM	<i>UBE3D</i>	protein_coding	Het	Het	Ref	Ref	no variant	

6	104212493	C	A	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
6	146092022	G	A	INTERGENIC			Het	Het	Ref	Ref	low quality	
6	154769585	G	A	INTRON	<i>CNKS3</i>	protein_coding	Het	Het	Ref	Ref	low quality+no variant	
6	162458275	T	C	INTRON	<i>PARK2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
6	169662202	G	A	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	inherited
6	169662209	A	G	INTERGENIC			Het	Het	Ref	Ref	inherited	
7	459309	A	G	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
7	9485054	T	G	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	inherited
7	70309917	A	G	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	inherited
7	81335206	T	G	INTRON	<i>HGF</i>	protein_coding	Het	Het	Ref	Ref	inherited	
7	140191996	T	C	INTERGENIC			Het	Het	Ref	Ref	low quality	
7	141955453	A	G	SYNONYMOUS_CODING	<i>PRSS58</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
7	157192929	T	C	INTRON	<i>DNAJB6</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
7	158037904	T	C	INTRON	<i>PTPRN2</i>	protein_coding	Het	Het	Ref	Ref	no variant	
7	158252534	G	A	INTRON	<i>PTPRN2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
8	1266955	G	C	INTERGENIC			Het	Het	Ref	Ref	low quality	
8	119959498	G	A	INTRON	<i>TNFRSF11B</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
8	142602829	G	A	INTERGENIC			Het	Het	Ref	Ref	low quality	
8	144890487	G	A	DOWNSTREAM	<i>MIR937</i>	miRNA	Het	Het	Ref	Ref	inherited	
9	32504067	C	A	INTRON	<i>DDX58</i>	protein_coding	Het	Het	Ref	Ref	inherited	
9	32917739	C	T	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
9	88703373	T	A	INTRON	<i>GOLM1</i>	protein_coding	Het	Het	Ref	Ref	inherited	
9	102011855	G	A	INTERGENIC			Het	Het	Ref	Ref	inherited	
9	102011909	A	G	INTERGENIC			Het	Het	Ref	Ref	inherited	
9	138010720	T	G	INTRON	<i>OLFM1</i>	protein_coding	Het	Het	Ref	Ref	low quality	
9	139526735	C	G	INTERGENIC			Het	Het	Ref	Ref	low quality + no variant	
9	140206420	C	T	INTRON	<i>EXD3</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
10	2357274	A	G	INTERGENIC			Het	Het	Ref	Ref	low quality	
10	7400791	C	G	INTRON	<i>SFMBT2</i>	protein_coding	Het	Het	Ref	Ref	inherited	
10	7400796	G	A	INTRON	<i>SFMBT2</i>	protein_coding	Het	Het	Ref	Ref	low quality	
10	7400806	C	T	INTRON	<i>SFMBT2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	inherited
10	48410093	A	G	DOWNSTREAM	<i>GDF2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
10	82233478	C	T	INTRON	<i>SPAN14</i>	protein_coding	Het	Het	Ref	Ref	low quality	
10	88123119	G	T	INTRON	<i>GRID1</i>	protein_coding	Het	Het	Ref	Ref	low quality	
10	97753629	T	A	DOWNSTREAM	<i>CC2D2B</i>	processed_transcript	Het	Het	Ref	Ref	inherited	

10	102527333	C	T	INTRON	<i>PAX2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
10	115563965	T	C	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	uninformative
10	128805987	G	A	INTRON	<i>DOCK1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
10	132099886	T	G	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
10	133937827	C	T	INTRON	<i>JAKMIP3</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
10	134392000	G	A	INTRON	<i>INPP5A</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
10	134837450	A	G	INTERGENIC			Het	Het	Ref	Ref	low quality	
10	134996444	G	A	DOWNSTREAM	<i>KNDC1</i>	processed_transcript	Het	Het	Ref	Ref	low quality	
11	1415375	A	C	INTRON	<i>BRSK2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
11	134534416	T	A	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
11	134605169	A	G	INTERGENIC			Het	Het	Ref	Ref	low quality	
12	162687	G	A	INTERGENIC			Het	Het	Ref	Ref	inherited	
12	718423	T	A	INTRON	<i>NINJ2</i>	protein_coding	Het	Het	Ref	Ref	low quality	
12	3527384	T	C	INTRON	<i>PRMT8</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
12	113724257	C	T	INTRON	<i>TPCN1</i>	protein_coding	Het	Het	Ref	Ref	inherited	
12	129572447	C	A	INTRON	<i>TMEM132D</i>	protein_coding	Het	Het	Ref	Ref	low quality	
12	131132579	T	C	INTRON	<i>RIMBP2</i>	protein_coding	Het	Het	Ref	Ref	inherited	
13	41929042	G	T	INTRON	<i>NAA16</i>	protein_coding	Het	Het	Ref	Ref	inherited	
13	97503636	T	C	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
13	112669460	C	T	INTERGENIC			Het	Het	Ref	Ref	low quality	
13	114220087	G	C	INTERGENIC			Het	Het	Ref	Ref	low quality	
13	114510530	C	T	DOWNSTREAM	<i>FAM70B</i>	processed_transcript	Het	Het	Ref	Ref	low quality	
13	114510547	C	G	DOWNSTREAM	<i>FAM70B</i>	processed_transcript	Het	Het	Ref	Ref	low quality	
13	114510560	T	C	DOWNSTREAM	<i>FAM70B</i>	processed_transcript	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
13	114514458	T	C	INTRON	<i>FAM70B</i>	protein_coding	Het	Het	Ref	Ref	low quality	
13	114514468	G	A	INTRON	<i>FAM70B</i>	protein_coding	Het	Het	Ref	Ref	low quality	
13	114965572	C	T	INTERGENIC			Het	Het	Ref	Ref	inherited	
14	32048285	T	A	INTRON	<i>NUBPL</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
14	105245986	A	G	INTRON	<i>AKT1</i>	protein_coding	Het	Het	Ref	Ref	low quality	
14	105345084	G	A	INTRON	<i>KIAA0284</i>	protein_coding	Het	Het	Ref	Ref	inherited	
14	105499889	A	G	INTERGENIC			Het	Het	Ref	Ref	low quality	
15	41256583	A	C	INTERGENIC			Het	Het	Ref	Ref	low quality+no variant	
15	48937494	C	T	INTRON	<i>FBN1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	inherited
15	57953050	A	T	INTRON	<i>GCOM1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
15	80788986	G	T	INTRON	<i>ARNT2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>

16	10858705	T	A	DOWNSTREAM	<i>FAM18A</i>	protein_coding	Het	Het	Ref	Ref	low quality	
16	22695512	T	G	INTERGENIC			Het	Het	Ref	Ref	inherited	
16	59139115	A	G	INTRON	<i>RP11-410D17.2</i>	lincRNA	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
16	84085063	G	T	DOWNSTREAM	<i>MBTPS1</i>	protein_coding	Het	Het	Ref	Ref	low quality	
16	87715100	C	T	INTRON	<i>JPH3</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
16	88313175	A	G	INTRON	<i>LA16c-444G7.1</i>	processed_transcript	Het	Het	Ref	Ref	low quality	
16	88497459	G	A	Non_Synonymous	<i>ZNF469</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
16	88560062	G	A	DOWNSTREAM	<i>ZFPM1</i>	retained_intron	Het	Het	Ref	Ref	low quality	
17	546943	T	C	INTRON	<i>VPS53</i>	protein_coding	Het	Het	Ref	Ref	inherited	
17	21303735	A	G	INTRON	<i>KCNJ12</i>	protein_coding	Het	Het	Ref	Ref	inherited	
17	21307296	G	C	INTRON	<i>KCNJ12</i>	protein_coding	Het	Het	Ref	Ref	inherited	
17	79276311	C	T	DOWNSTREAM	<i>LINC00482</i>	lincRNA	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
17	79906331	C	T	DOWNSTREAM	<i>NOTUM</i>	protein_coding	Het	Het	Ref	Ref	inherited	
17	80346681	C	G	DOWNSTREAM	<i>C17orf101</i>	protein_coding	Het	Het	Ref	Ref	low quality	
17	80942321	C	T	DOWNSTREAM	<i>RP11-1197K16.2</i>	antisense	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
17	81024707	C	T	INTERGENIC			Het	Het	Ref	Ref	inherited	
18	3412160	G	A	UPSTREAM	<i>TGIF1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	inherited
18	32459844	T	C	INTRON	<i>DTNA</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
18	77204733	G	C	INTRON	<i>NFATC1</i>	protein_coding	Het	Het	Ref	Ref	low quality	
19	523438	C	T	DOWNSTREAM	<i>TPGS1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
19	682321	T	A	DOWNSTREAM	<i>PRSS57</i>	protein_coding	Het	Het	Ref	Ref	inherited	
19	11564606	A	C	DOWNSTREAM	<i>ELAVL3</i>	protein_coding	Het	Het	Ref	Ref	low quality + no variant	
19	14732346	G	A	INTRON	<i>EMR3</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	uninformative
19	39330862	A	G	SYNONYMOUS_CODING	<i>HNRNPL</i>	protein_coding	Het	Het	Ref	Ref	low quality	
19	49842785	A	C	DOWNSTREAM	<i>TEAD2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
19	50197909	T	G	DOWNSTREAM	<i>C19orf76</i>	protein_coding	Het	Het	Ref	Ref	low quality	
19	53659085	C	T	INTRON	<i>ZNF347</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
20	653248	C	T	INTRON	<i>SCRT2</i>	protein_coding	Het	Het	Ref	Ref	low quality	
20	14840274	T	C	INTRON	<i>MACROD2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
20	50012140	C	G	INTRON	<i>NFATC2</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	no variant
20	50690005	T	A	INTRON	<i>ZFP64</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
20	52184901	A	C	DOWNSTREAM	<i>ZNF217</i>	protein_coding	Het	Het	Ref	Ref	low quality	

20	61607457	T	C	INTERGENIC			Het	Het	Ref	Ref	low quality	
20	62836674	T	C	INTRON	<i>MYT1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
20	62866083	T	G	INTRON	<i>MYT1</i>	protein_coding	Het	Het	Ref	Ref	<i>de novo</i>	N/A primer pair
21	17686365	T	A	INTRON	<i>LINC00478</i>	processed_transcript	Het	Het	Ref	Ref	inherited	
21	23811291	C	T	INTERGENIC			Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
21	44743604	A	C	INTRON	<i>LINC00322</i>	processed_transcript	Het	Het	Ref	Ref	<i>de novo</i>	inherited
21	45206930	C	T	DOWNSTREAM	<i>TMEM97P1</i>	processed_pseudogene	Het	Het	Ref	Ref	low quality	
21	46304683	T	G	DOWNSTREAM	<i>ITGB2</i>	protein_coding	Het	Het	Ref	Ref	low quality	
22	19167292	G	A	DOWNSTREAM	<i>CLTCL1</i>	nonsense_mediated_decay	Het	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
22	19704712	T	G	INTRON	<i>SEPT5</i>	protein_coding	Het	Het	Ref	Ref	low quality	
22	40898746	T	C	INTRON	<i>MKL1</i>	protein_coding	Het	Het	Ref	Ref	low quality	
22	49761849	A	G	INTERGENIC			Het	Het	Ref	Ref	low quality	
22	49840787	G	A	INTRON	<i>C22orf34</i>	nonsense_mediated_decay	Het	Het	Ref	Ref	low quality	
X	2836434	G	A	DOWNSTREAM	<i>ARSD</i>	nonsense_mediated_decay	Het	Het	Ref	Ref	<i>de novo</i>	no variant
X	145422663	T	C	INTERGENIC			Het	Het	Ref	Ref	low quality	

Abbreviations used in this table: chr, chromosome; pos, position; ref, reference allele; alt, alternative allele; N/A, not available

Table C.2: List of the *de novo* SNV candidates (n=83) specific to twin I

Chr	Pos	Ref	Alt	Genomic Context	Gene	Gene type	Twin I	Twin II	Mother	Father	IGV Prediction	Sanger
1	162730172	A	G	INTRON	<i>DDR2</i>	protein_coding	Het	Ref	Ref	Ref	inherited	
1	245499265	A	G	INTRON	<i>KIF26B</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	inherited
2	61728351	A	T	DOWNSTREAM	<i>XPO1</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	<i>de novo</i>
2	111926337	T	G	DOWNSTREAM	<i>BCL2L11</i>	protein_coding	Het	Ref	Ref	Ref	inherited	
2	241627105	T	G	UPSTREAM	<i>AQP12A</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
3	1493383	C	T	INTERGENIC			Het	Ref	Ref	Ref	<i>de novo</i>	<i>de novo</i>
3	20897127	T	C	INTERGENIC			Het	Ref	Ref	Ref	low quality	
3	51090509	G	T	INTRON	<i>DOCK3</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
3	51441939	A	C	INTRON	<i>VPRBP</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	no variant
3	66701546	T	G	INTERGENIC			Het	Ref	Ref	Ref	low quality	
3	72959480	G	A	INTRON	<i>GXYLT2</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	<i>de novo</i>
3	135876070	C	T	DOWNSTREAM	<i>TDGF1P6</i>	processed_pseudogene	Het	Ref	Ref	Ref	low quality	
3	187442688	A	C	EXON	<i>BCL6</i>	retained_intron	Het	Ref	Ref	Ref	low quality	
3	196430415	C	T	DOWNSTREAM	<i>CEP19</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
4	651520	G	A	DOWNSTREAM	<i>PDE6B</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	N/A PCR product
4	6564000	G	A	INTRON	<i>PPP2R2C</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	<i>de novo</i>
4	7920679	G	A	INTRON	<i>AFAP1</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
4	189032254	C	G	DOWNSTREAM	<i>RP11-713C19.1</i>	processed_pseudogene	Het	Ref	Ref	Ref	low quality	
4	189792261	C	G	INTERGENIC			Het	Ref	Ref	Ref	low quality	
5	1956661	A	G	INTRON	<i>RP11-259O2.1</i>	lincRNA	Het	Ref	Ref	Ref	low quality	
5	54561146	A	G	INTRON	<i>DHX29</i>	protein_coding	Het	Ref	Ref	Ref	inherited	
5	82195939	G	A	INTERGENIC			Het	Ref	Ref	Ref	low quality	
5	141576736	A	C	INTERGENIC			Het	Ref	Ref	Ref	inherited	
6	16943658	A	C	INTERGENIC			Het	Ref	Ref	Ref	low quality	

6	20235579	T	G	INTRON	<i>RP11-239H6.2</i>	processed_transcript	Het	Ref	Ref	Ref	low quality	
6	41586146	G	A	INTERGENIC			Het	Ref	Ref	Ref	<i>de novo</i>	inherited
6	134981681	T	G	INTERGENIC			Het	Ref	Ref	Ref	<i>de novo</i>	no variant
7	1824582	T	C	INTERGENIC			Het	Ref	Ref	Ref	low quality	
7	1824583	G	A	INTERGENIC			Het	Ref	Ref	Ref	inherited	
7	29606885	G	A	DOWNSTREAM	<i>PRR15</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	<i>de novo</i>
7	40103905	T	G	INTRON	<i>CDK13</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
7	76922180	A	C	INTRON	<i>CCDC146</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
7	155510228	T	G	INTRON	<i>RBM33</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	no variant
7	157981209	G	C	INTRON	<i>PTPRN2</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	N/A PCR product
8	66686049	G	A	DOWNSTREAM	<i>MTFR1</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	uninformative
8	118838053	A	G	INTRON	<i>EXT1</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	<i>de novo</i>
8	143671111	G	C	INTERGENIC			Het	Ref	Ref	Ref	low quality	
9	140816091	C	A	INTRON	<i>CACNA1B</i>	protein_coding	Het	Ref	Ref	Ref	inherited	
10	3940738	G	A	INTERGENIC			Het	Ref	Ref	Ref	low quality	
10	10872261	G	A	INTERGENIC			Het	Ref	Ref	Ref	low quality	
10	45869059	C	A	INTERGENIC			Het	Ref	Ref	Ref	<i>de novo</i>	N/A PCR product
10	48410018	G	C	DOWNSTREAM	<i>GDF2</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
10	77154164	C	T	DOWNSTREAM	<i>ZNF503</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
11	65374709	T	G	DOWNSTREAM	<i>MAP3K11</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
12	653012	T	C	INTRON	<i>B4GALNT3</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	N/A primer pair
12	9557500	T	C	INTRON	<i>RP11-599J14.2</i>	unprocessed_pseudogene	Het	Ref	Ref	Ref	<i>de novo</i>	no variant
12	64229411	T	G	INTERGENIC			Het	Ref	Ref	Ref	low quality	
12	95569608	G	A	INTRON	<i>FGD6</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	uninformative
12	131691875	G	A	INTRON	<i>RP11-638F5.1</i>	lincRNA	Het	Ref	Ref	Ref	<i>de novo</i>	inherited
12	133052676	T	C	INTERGENIC			Het	Ref	Ref	Ref	low quality	
12	133352895	G	C	INTRON	<i>GOLGA3</i>	protein_coding	Het	Ref	Ref	Ref	<i>de novo</i>	N/A PCR product
13	51812216	T	G	INTRON	<i>FAM124A</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
13	87083684	C	A	INTERGENIC			Het	Ref	Ref	Ref	low quality	
14	52997360	A	T	INTRON	<i>TXNDC16</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
14	80678369	C	T	INTRON	<i>DIO2</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
14	100044559	A	G	INTRON	<i>CCDC85C</i>	protein_coding	Het	Ref	Ref	Ref	no variant	
16	1970303	A	G	INTERGENIC			Het	Ref	Ref	Ref	low quality	
16	10858718	A	G	DOWNSTREAM	<i>FAM18A</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
16	50424085	G	A	DOWNSTREAM	<i>RP11-21B23.2</i>	lincRNA	Het	Ref	Ref	Ref	<i>de novo</i>	<i>de novo</i>

16	70714811	C	G	INTRON	<i>MTSS1L</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
16	77724703	G	T	INTERGENIC			Het	Ref	Ref	Ref	low quality	
17	92938	C	T	DOWNSTREAM	<i>RPH3AL</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
17	71734490	T	C	INTRON	<i>CTD-2532D12.4</i>	lincRNA	Het	Ref	Ref	Ref	inherited	
17	73824833	G	C	DOWNSTREAM	<i>UNK</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
18	34970824	C	T	INTRON	<i>CELF4</i>	protein_coding	Het	Ref	Ref	Ref	de novo	uninformative
18	50453718	G	T	INTRON	<i>DCC</i>	protein_coding	Het	Ref	Ref	Ref	de novo	N/A PCR product
19	309882	A	G	DOWNSTREAM	<i>CTD-3113P16.5</i>	antisense	Het	Ref	Ref	Ref	no variant	
19	33105233	T	G	INTRON	<i>ANKRD27</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
20	53271335	C	T	DOWNSTREAM	<i>DOK5</i>	protein_coding	Het	Ref	Ref	Ref	de novo	de novo
20	56815552	G	C	DOWNSTREAM	<i>PPP4R1L</i>	processed_transcript	Het	Ref	Ref	Ref	low quality	
20	62810452	T	C	INTRON	<i>MYT1</i>	protein_coding	Het	Ref	Ref	Ref	inherited	
21	34512162	C	A	INTERGENIC			Het	Ref	Ref	Ref	low quality	
X	2833489	C	G	DOWNSTREAM	<i>ARSD</i>	retained_intron	Het	Ref	Ref	Ref	de novo	N/A primer pair
X	7544103	T	C	INTERGENIC			Het	Ref	Ref	Ref	de novo	N/A primer pair
X	7544110	T	C	INTERGENIC			Het	Ref	Ref	Ref	low quality	
X	10107413	T	A	INTRON	<i>WWC3</i>	protein_coding	Het	Ref	Ref	Ref	low quality	
X	68384591	T	G	INTRON	<i>PJA1</i>	protein_coding	Hom	Ref	Ref	Ref	low quality	
X	116523014	T	A	INTERGENIC			Het	Ref	Ref	Ref	low quality	
X	144597115	T	A	INTERGENIC			Het	Ref	Ref	Ref	low quality	
X	144597121	C	A	INTERGENIC			Hom	Ref	Ref	Ref	de novo	N/A primer pair
X	148731301	C	A	EXON	<i>RP5-937E21.1</i>	processed_pseudogene	Het	Ref	Ref	Ref	de novo	no variant
X	150242328	A	G	INTERGENIC			Het	Ref	Ref	Ref	no variant	
Y	8094333	C	A	INTERGENIC			Het	Ref	Ref	Ref	de novo	N/A primer pair

Abbreviations used in this table: chr, chromosome; pos, position; ref, reference allele; alt, alternative allele; N/A, not available

Table C.3: List of the *de novo* SNV candidates (n=48) specific to twin II

Chr	Pos	Ref	Alt	Genomic Context	Gene	Gene type	Twin I	Twin II	Mother	Father	IGV Prediction	Sanger
1	211950090	G	A	DOWNSTREAM	<i>LPGAT1</i>	processed_transcript	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
1	228146563	C	A	INTERGENIC			Ref	Het	Ref	Ref	<i>de novo</i>	no variant
2	362130	G	T	INTERGENIC			Ref	Het	Ref	Ref	inherited	
2	362131	A	T	INTERGENIC			Ref	Het	Ref	Ref	inherited	
2	1404408	C	A	INTRON	<i>TPO</i>	processed_transcript	Ref	Het	Ref	Ref	low quality	
2	10476221	G	T	INTRON	<i>HPCAL1</i>	protein_coding	Ref	Het	Ref	Ref	inherited	
2	70925457	G	T	INTRON	<i>ADD2</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
2	154061708	G	T	DOWNSTREAM	<i>AC079150.2</i>	processed_pseudogene	Ref	Het	Ref	Ref	low quality	
2	241992247	C	T	DOWNSTREAM	<i>SNED1</i>	retained_intron	Ref	Het	Ref	Ref	<i>de novo</i>	uninformative
4	3827696	A	C	INTERGENIC			Ref	Het	Ref	Ref	low quality	
5	194385	A	C	DOWNSTREAM	<i>PLEKHG4B</i>	protein_coding	Ref	Het	Ref	Ref	low quality	
5	16525233	G	A	INTRON	<i>FAM134B</i>	protein_coding	Ref	Het	Ref	Ref	inherited	
5	137885670	T	A	DOWNSTREAM	<i>HSPA9</i>	protein_coding	Ref	Het	Ref	Ref	inherited	
6	57541536	G	T	INTERGENIC			Ref	Het	Ref	Ref	inherited	
6	86090624	G	A	INTERGENIC			Ref	Het	Ref	Ref	low quality	
6	133966995	G	T	INTERGENIC			Ref	Het	Ref	Ref	<i>de novo</i>	no variant
7	10338952	G	T	INTERGENIC			Ref	Het	Ref	Ref	low quality	
7	56559682	A	G	DOWNSTREAM	<i>RBM22P3</i>	processed_pseudogene	Ref	Het	Ref	Ref	low quality	
7	158204835	C	T	INTRON	<i>PTPRN2</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	N/A Primer Pair
7	158248408	C	T	INTRON	<i>PTPRN2</i>	protein_coding	Ref	Het	Ref	Ref	low quality	
8	142602825	C	T	INTERGENIC			Ref	Het	Ref	Ref	low quality	
9	14347587	G	A	INTRON	<i>NFIB</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	inherited
9	106166434	C	T	INTRON	<i>RP11-436F21.1</i>	lincRNA	Ref	Het	Ref	Ref	low quality	

10	3357139	G	A	INTERGENIC			Ref	Het	Ref	Ref	low quality	
10	30518280	A	G	INTERGENIC			Ref	Het	Ref	Ref	low quality	
11	1474891	G	T	INTRON	<i>BRSK2</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
11	129307937	T	A	INTRON	<i>BARX2</i>	protein_coding	Ref	Het	Ref	Ref	low quality	
12	132353193	A	C	DOWNSTREAM	<i>RP11-417L19.2</i>	lincRNA	Ref	Het	Ref	Ref	<i>de novo</i>	N/A PCR product
13	114617416	G	A	INTRON	<i>LINC00452</i>	lincRNA	Ref	Het	Ref	Ref	low quality	
13	114619650	A	G	INTRON	<i>LINC00452</i>	lincRNA	Ref	Het	Ref	Ref	low quality	
15	76884600	G	T	INTRON	<i>SCAPER</i>	protein_coding	Ref	Het	Ref	Ref	low quality	
16	62015587	C	T	INTRON	<i>CDH8</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
17	4782495	T	G	DOWNSTREAM	<i>Metazoa_SRP</i>	misc_RNA	Ref	Het	Ref	Ref	low quality	
17	79646774	T	G	DOWNSTREAM	<i>ARL16</i>	protein_coding	Ref	Het	Ref	Ref	low quality	
17	81042516	T	C	INTRON	<i>METRNL</i>	protein_coding	Ref	Het	Ref	Ref	low quality	
18	18623582	G	A	DOWNSTREAM	<i>ROCK1</i>	retained_intron	Ref	Het	Ref	Ref	low quality	
21	36322701	A	G	INTRON	<i>RUNX1</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	<i>de novo</i>
21	47479749	C	T	DOWNSTREAM	<i>AP001476.3</i>	lincRNA	Ref	Het	Ref	Ref	<i>de novo</i>	N/A Primer Pair
21	47479750	C	G	DOWNSTREAM	<i>AP001476.3</i>	lincRNA	Ref	Het	Ref	Ref	low quality	
X	7320881	C	G	INTERGENIC			Ref	Het	Ref	Ref	<i>de novo</i>	no variant
X	71177022	T	C	INTRON	<i>NHSL2</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
X	71177023	A	T	INTRON	<i>NHSL2</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
X	104517151	A	G	INTRON	<i>IL1RAPL2</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
X	122224237	A	G	INTERGENIC			Ref	Het	Ref	Ref	no variant	
X	123370010	C	T	INTRON	<i>STAG2</i>	processed_transcript	Ref	Het	Ref	Ref	low quality	
X	128927246	C	G	INTRON	<i>SASH3</i>	protein_coding	Ref	Het	Ref	Ref	<i>de novo</i>	no variant
X	134054617	A	C	INTERGENIC			Ref	Het	Ref	Ref	no variant	
X	134054618	A	T	INTERGENIC			Ref	Het	Ref	Ref	<i>de novo</i>	no variant

Abbreviations used in this table: chr, chromosome; pos, position; ref, reference allele; alt, alternative allele; N/A, not available

Appendix D

Electropherograms for the Shared *De Novo* SNVs Validated through Sanger Sequencing

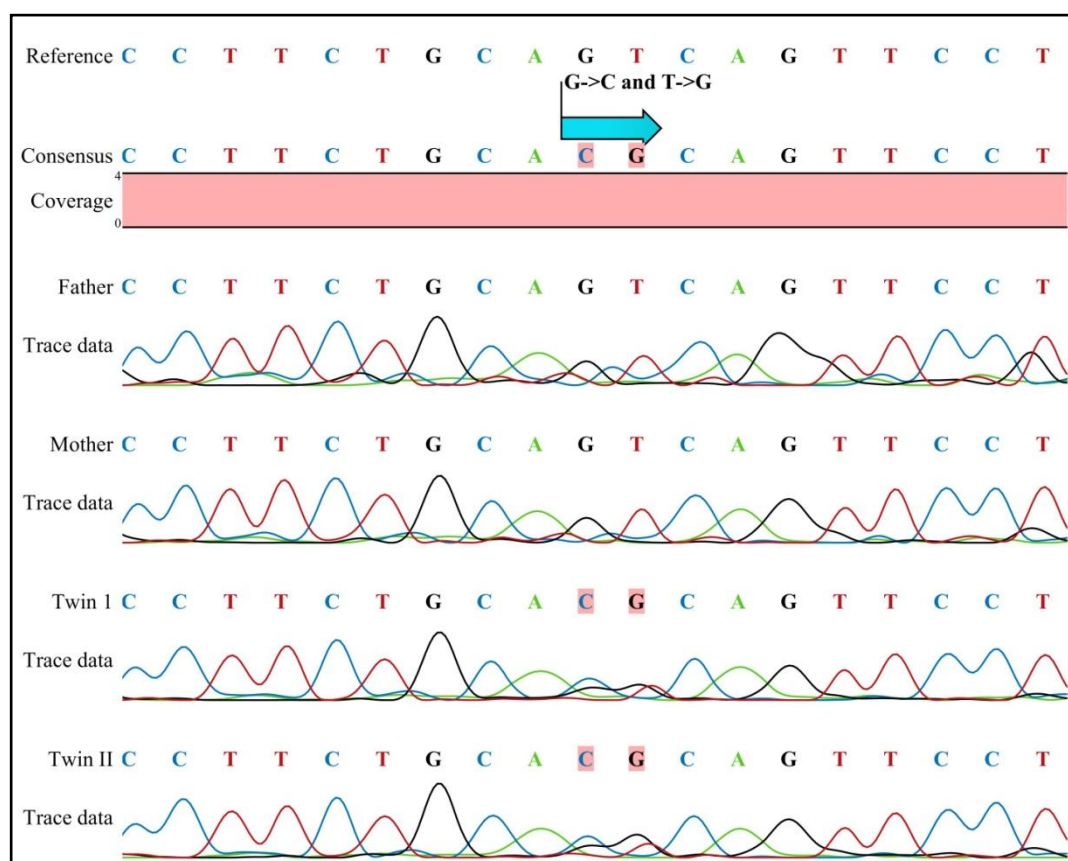


Figure D.1: Electropherogram for the 2 *de novo* SNVs shared by the twin pair (Chromosome1: 165599592: G/C and Chromosome1: 165599593: T/G).

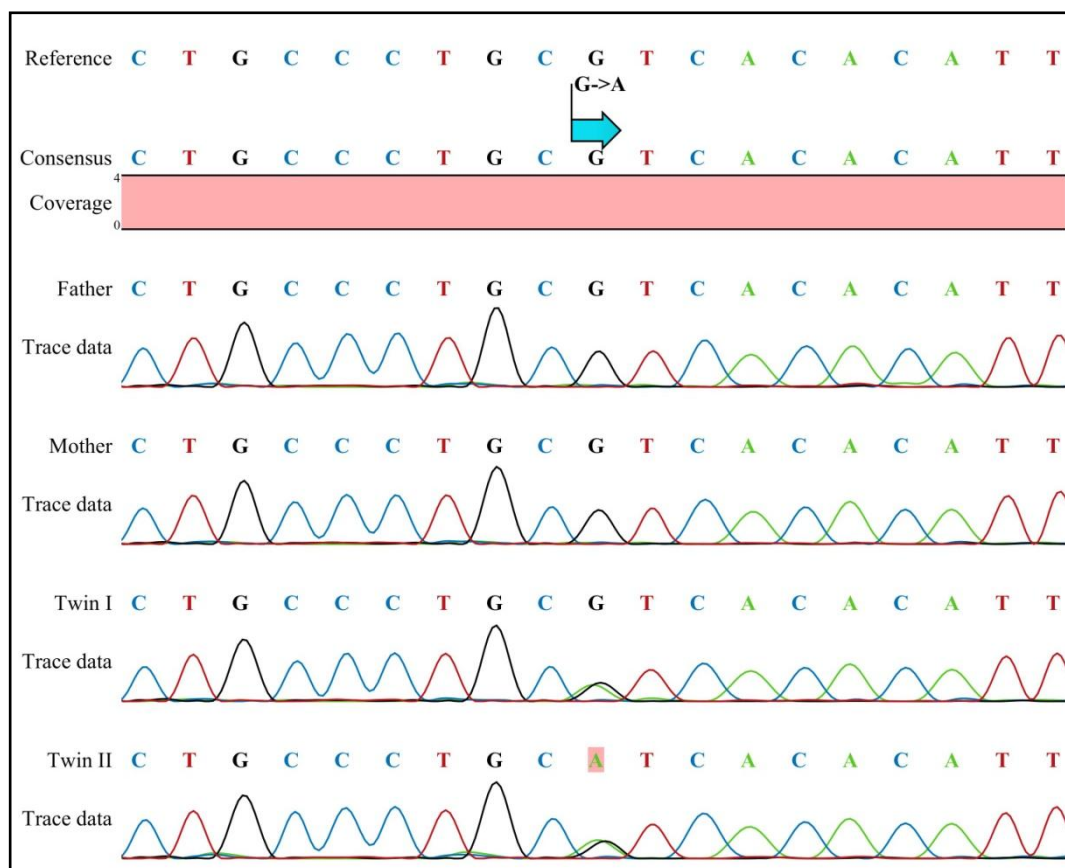


Figure D.2: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome1: 245093862: G/A).

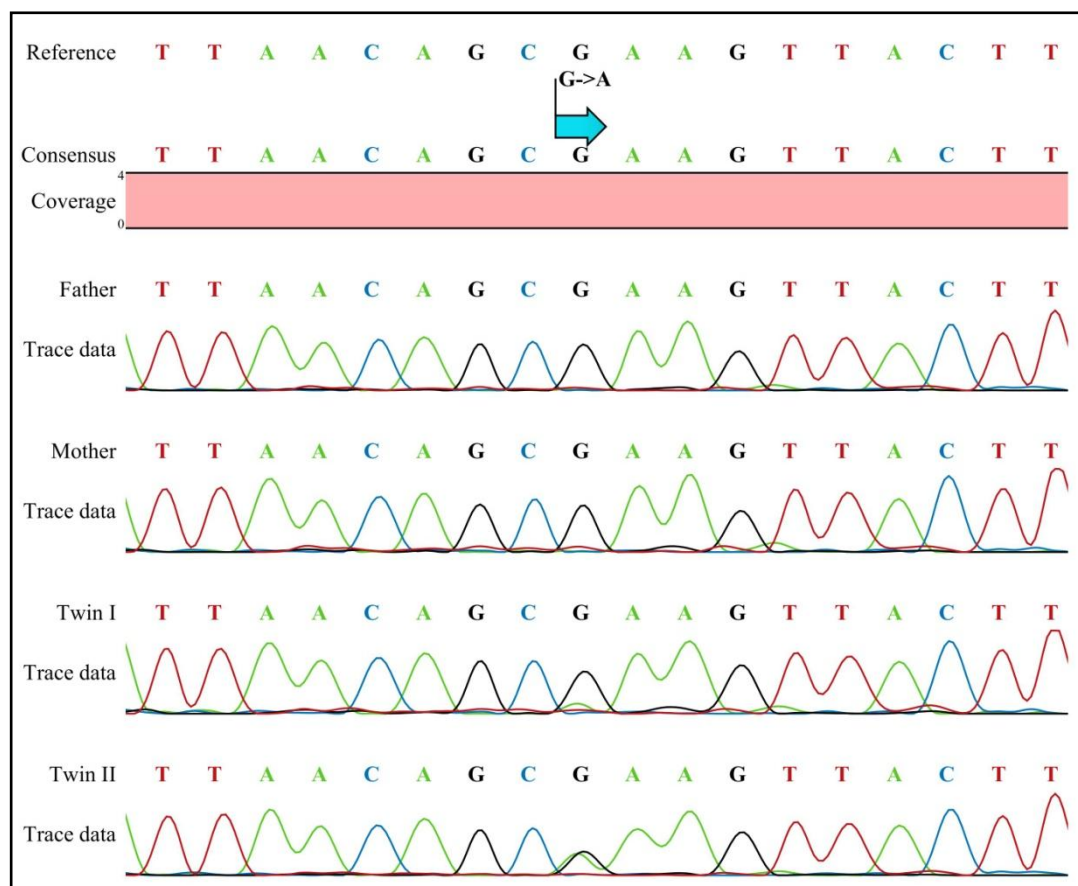


Figure D.3: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome2: 136342189: G/A).

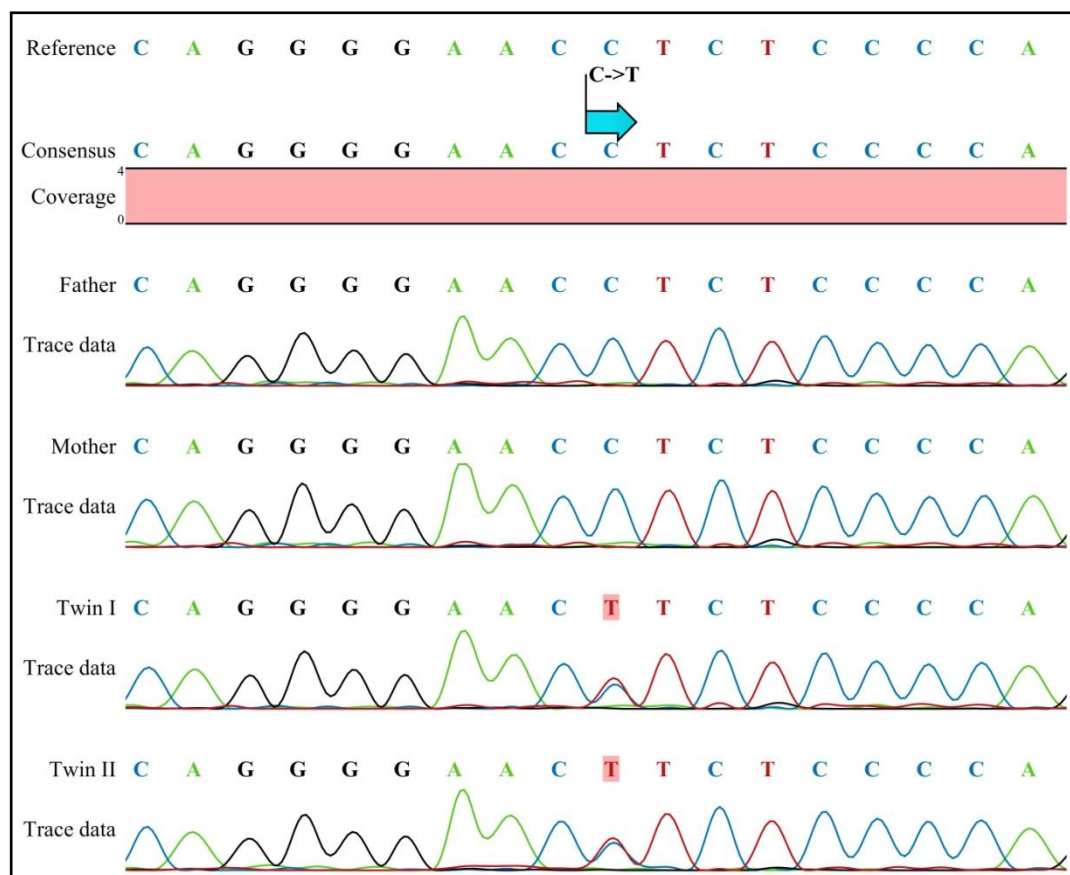


Figure D.4: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome4: 145566227: C/T).

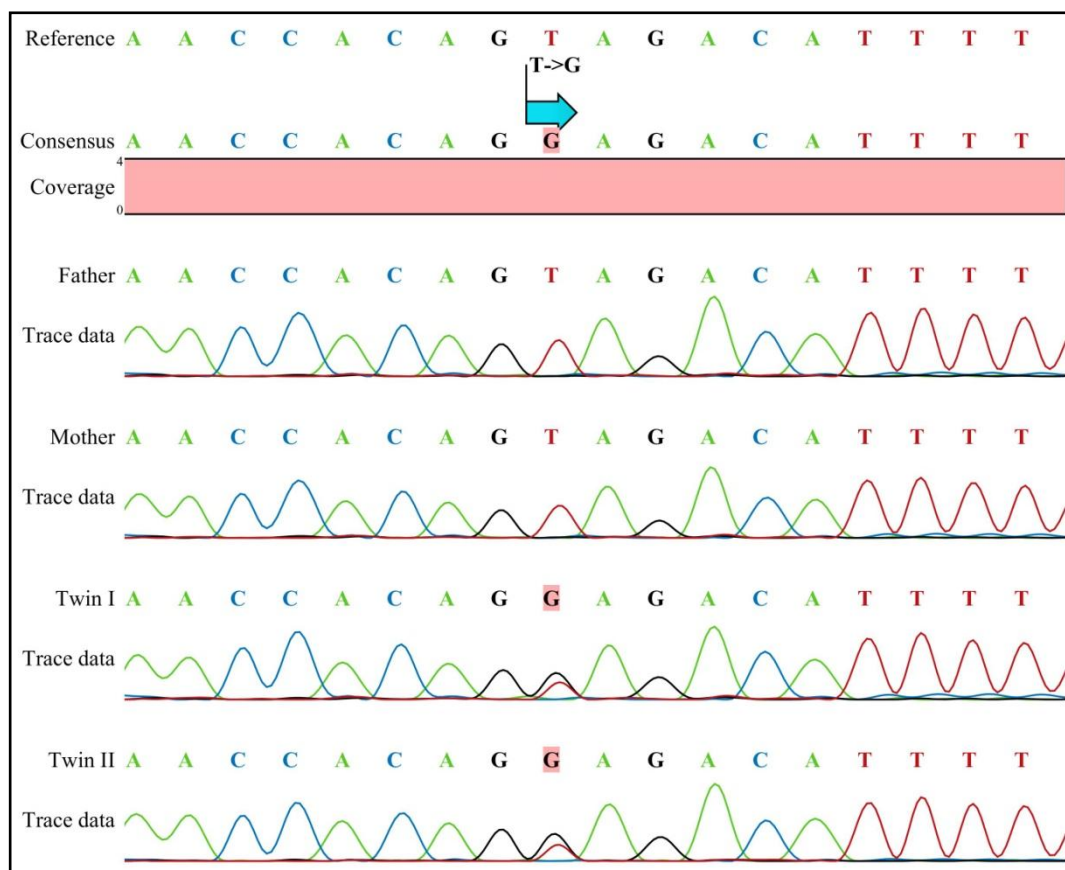


Figure D.5: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome5: 11043438: T/G).

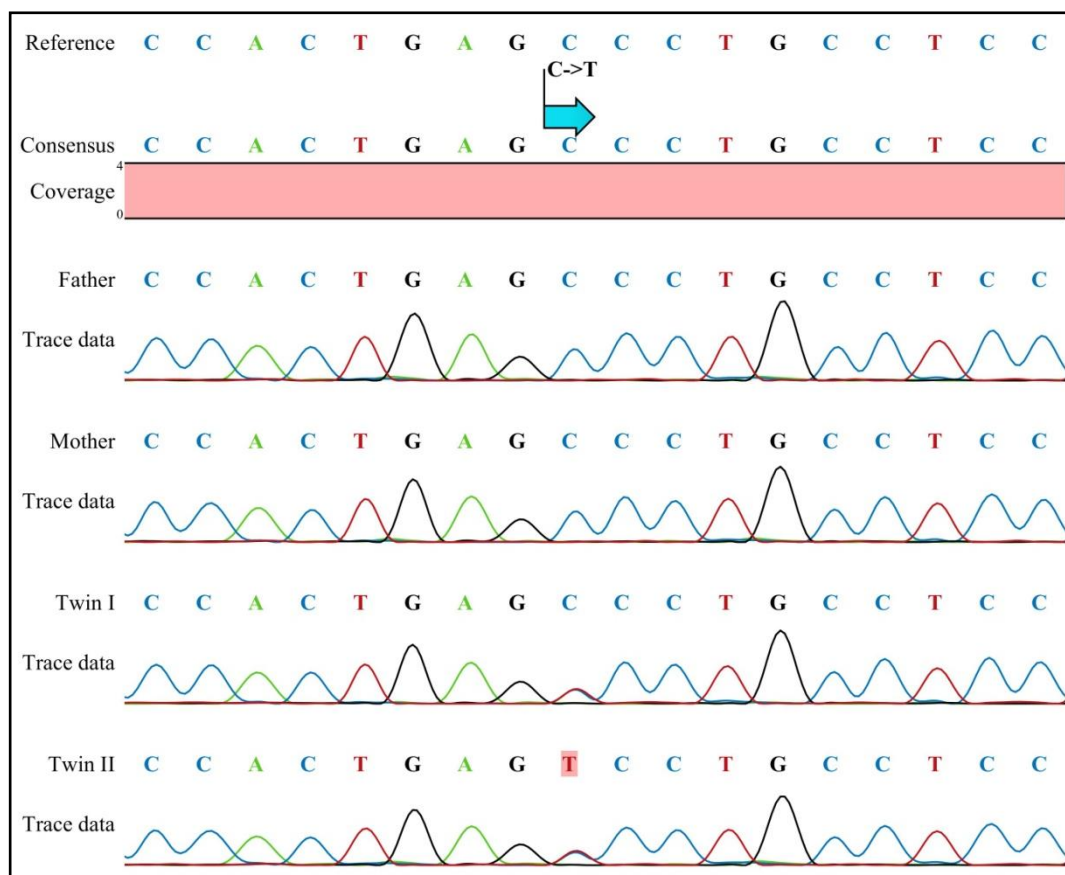


Figure D.6: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome6: 29101417: C/T).

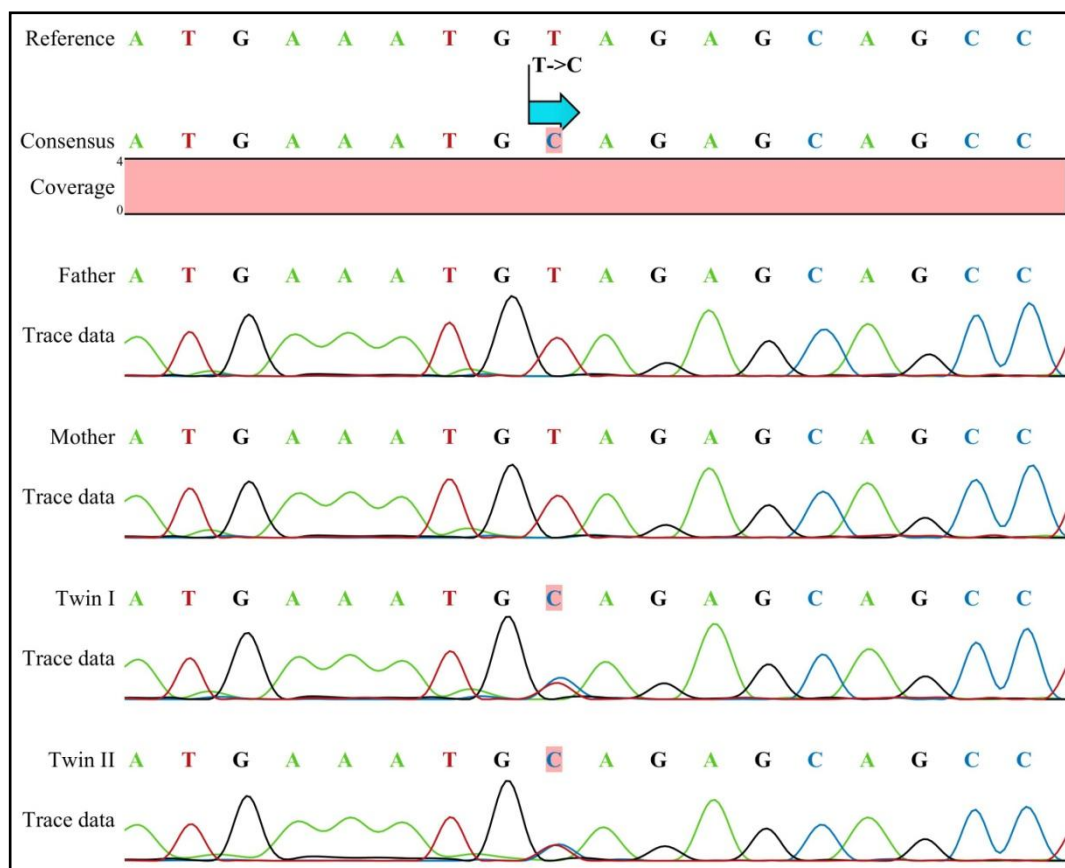


Figure D.7: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome6: 162458275: T/C).

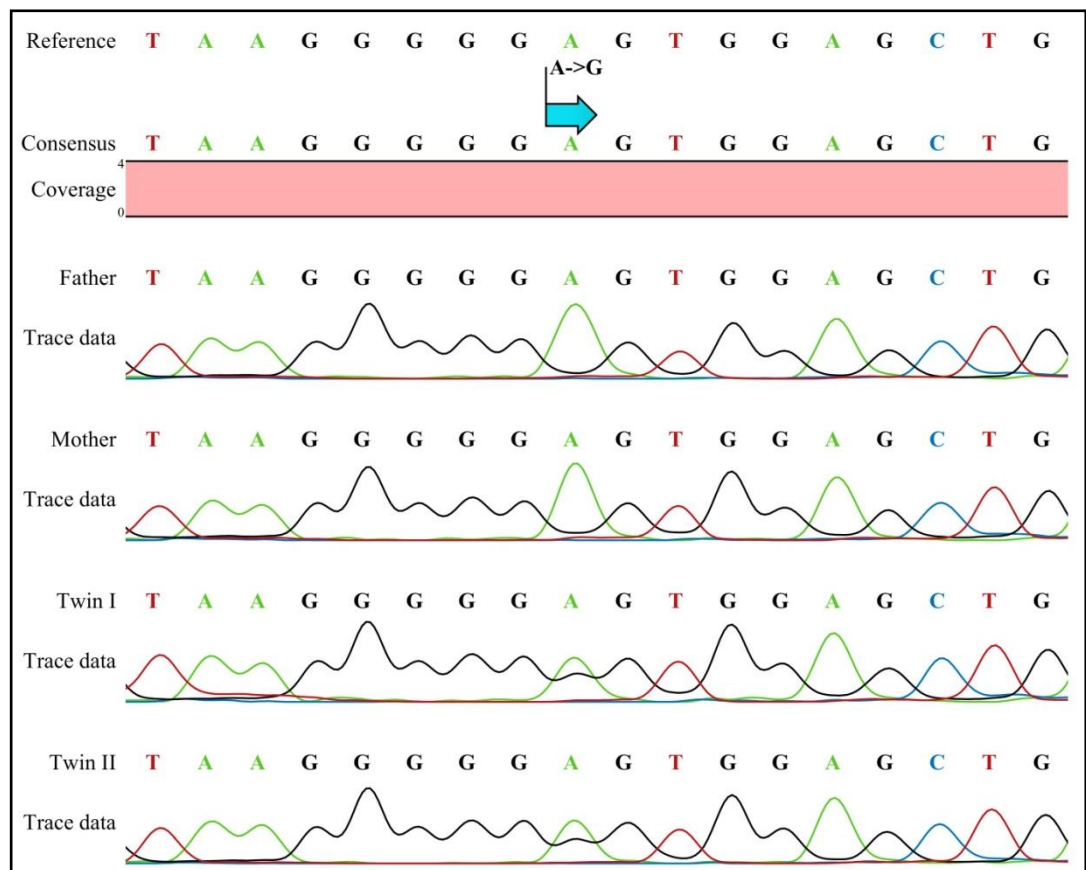


Figure D.8: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome7: 141955453: A/G).

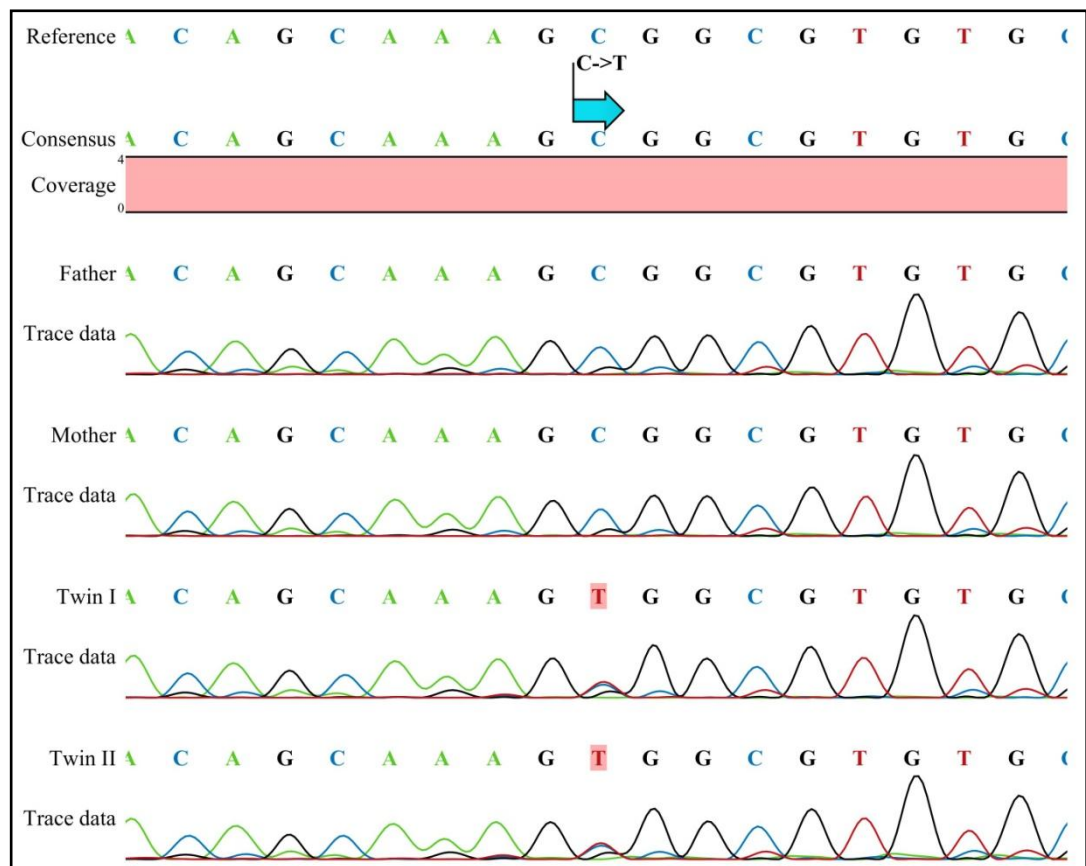


Figure D.9: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome9: 32917739: C/T).

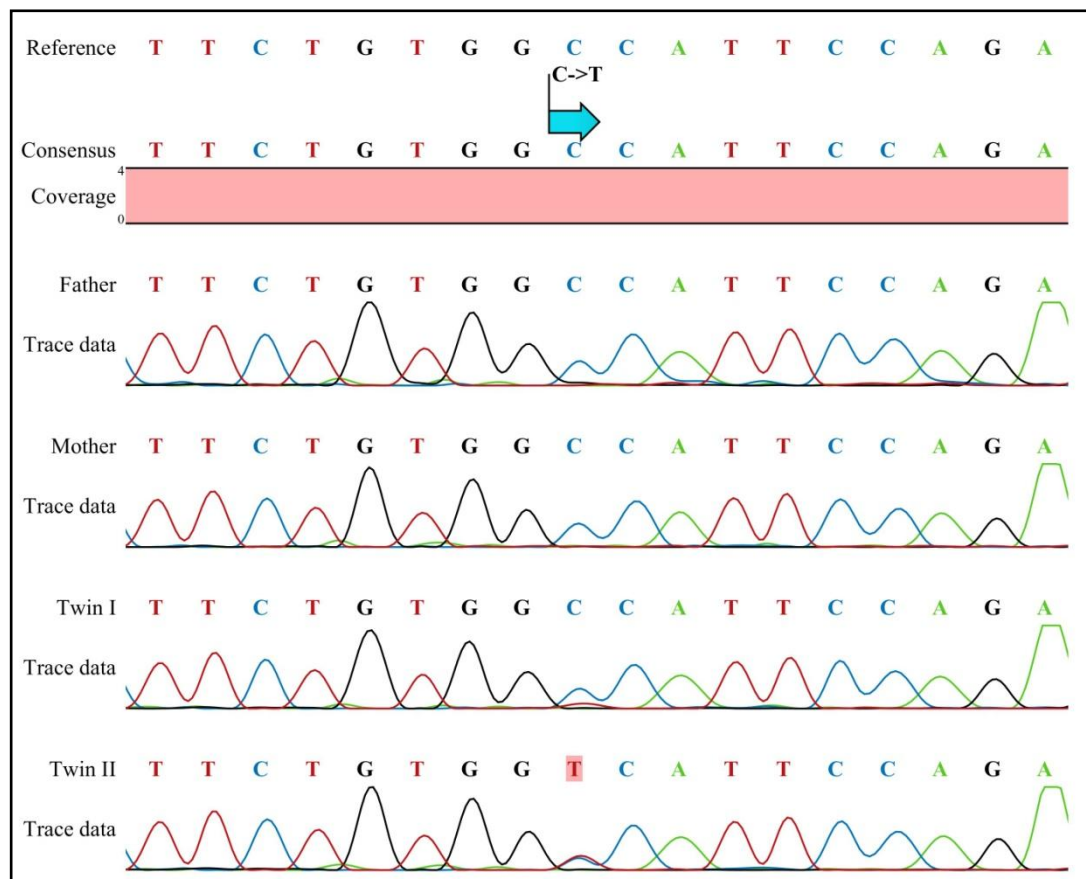


Figure D.10: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome10: 102527333: C/T).

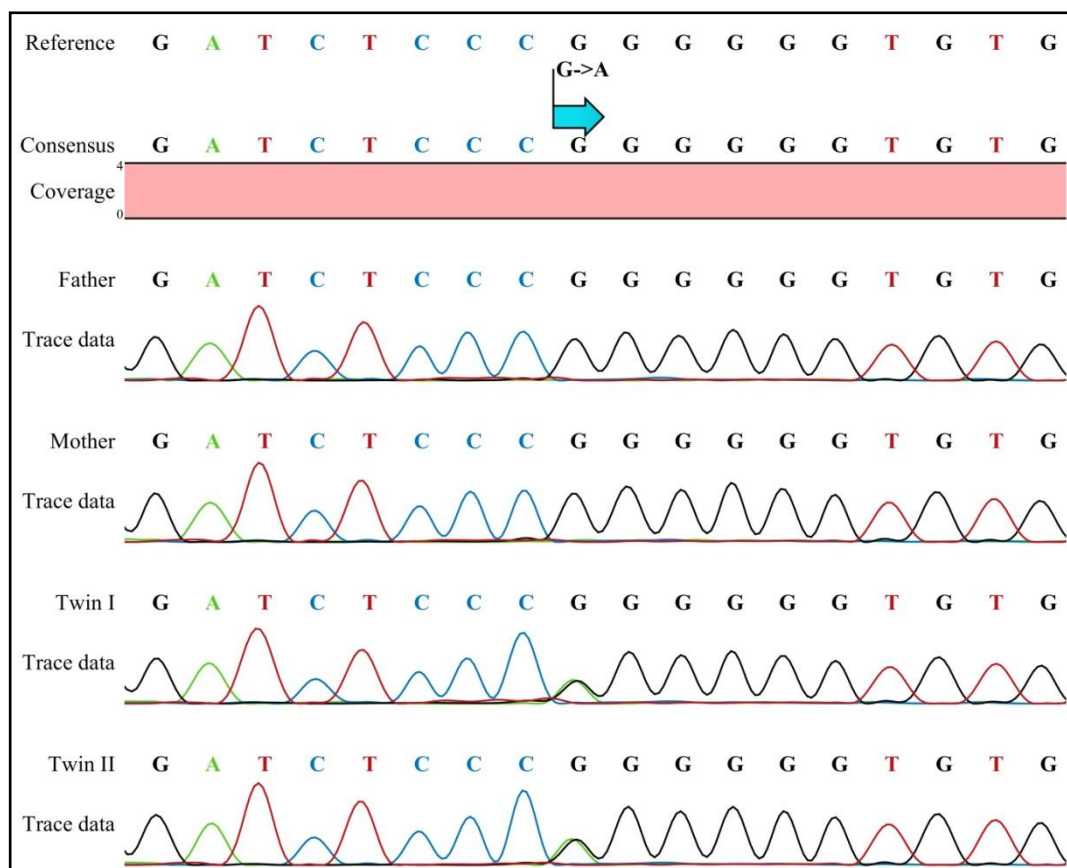


Figure D.11: Electropherogram for the *de novo* SNV shared by the twin pair
(Chromosome10: 128805987: G/A).

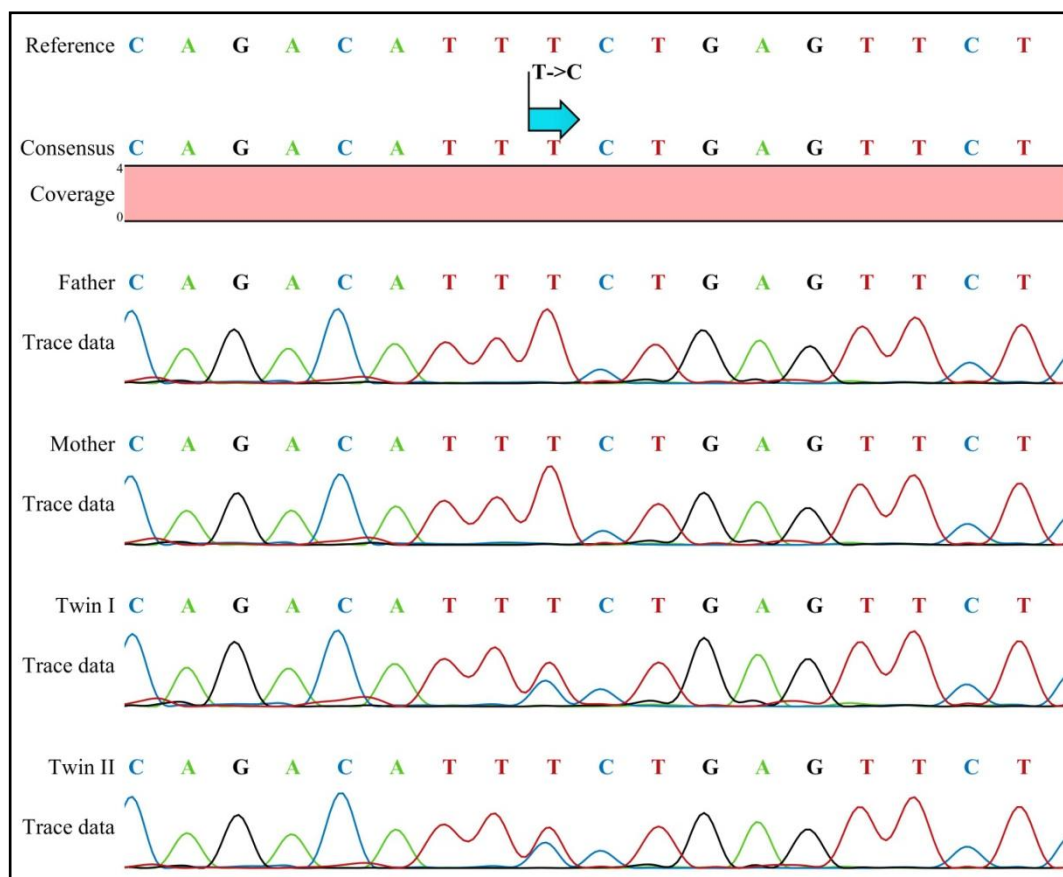


Figure D.12: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome12: 3527384: T/C).

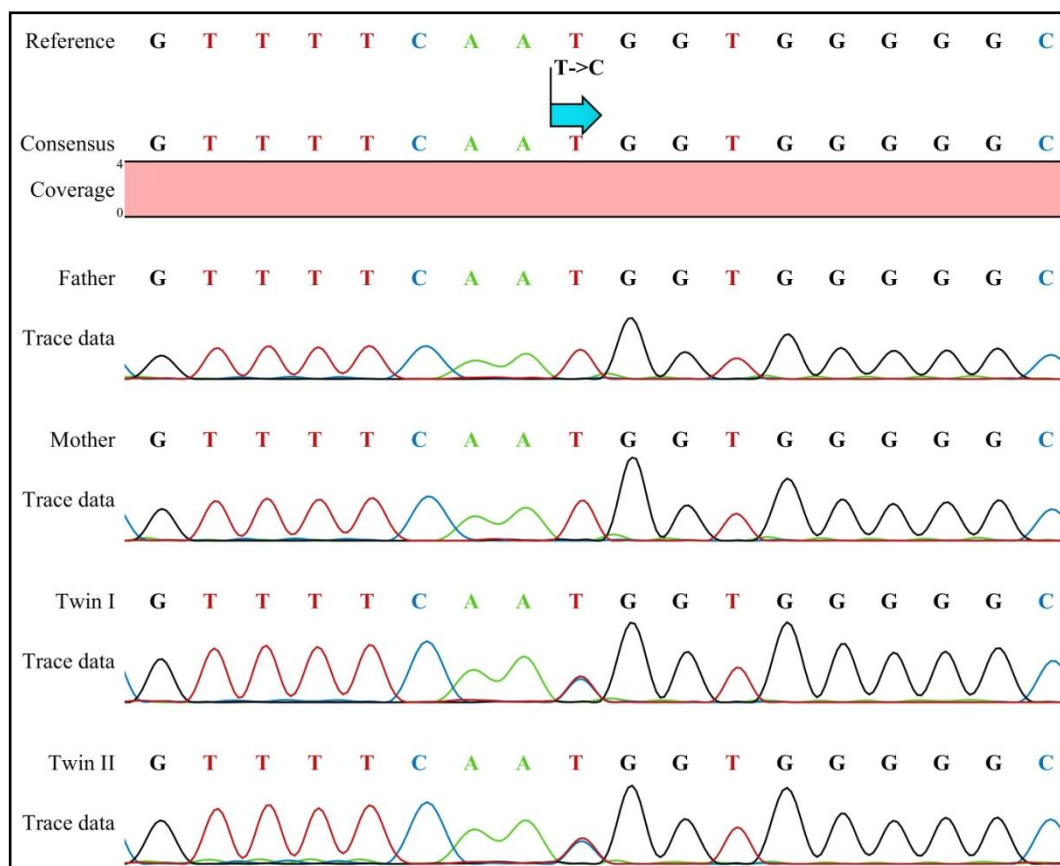


Figure D.13: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome13: 97503636: T/C).

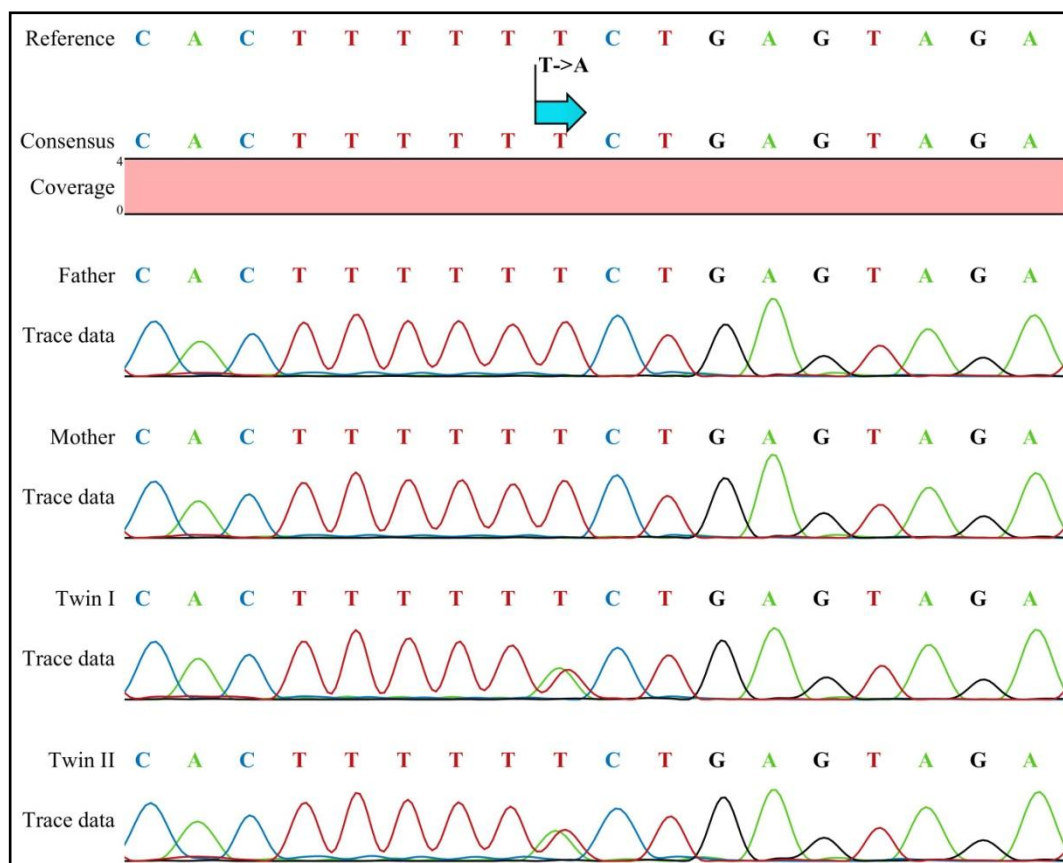


Figure D.14: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome14: 32048285: T/A).

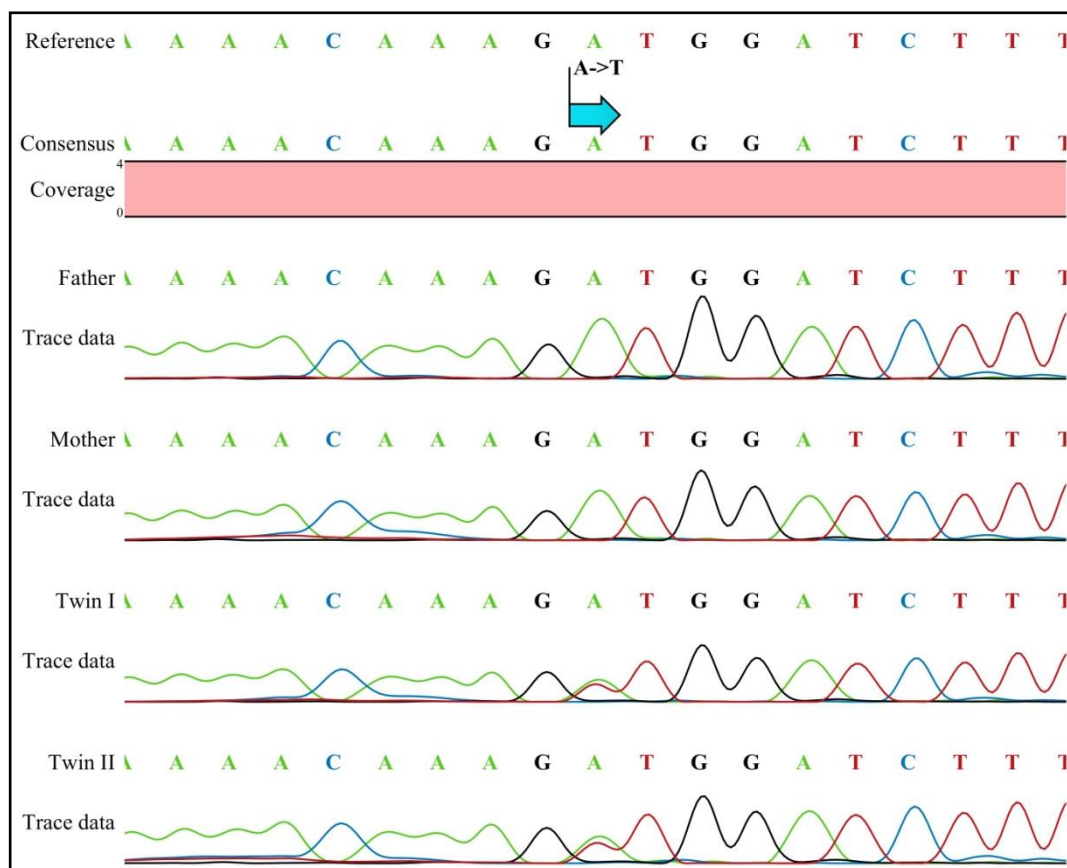


Figure D.15: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome15: 57953050: A/T).

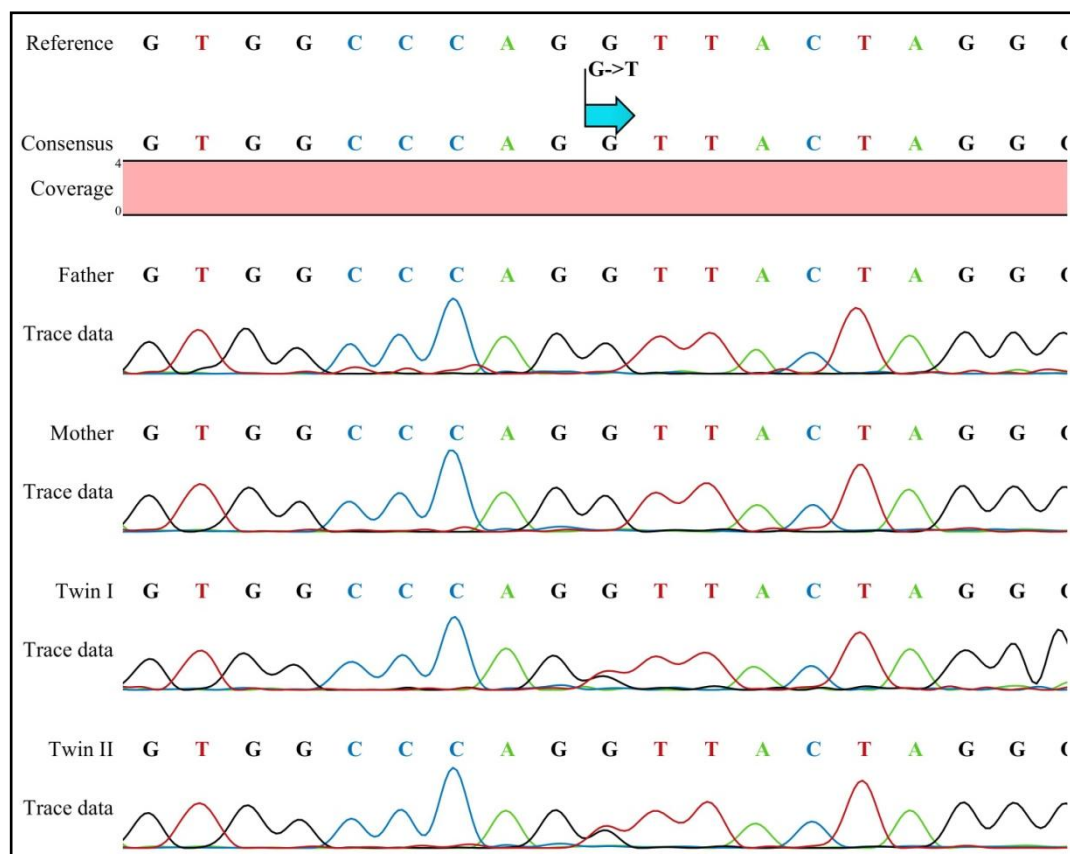


Figure D.16: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome15: 80788986: G/T).

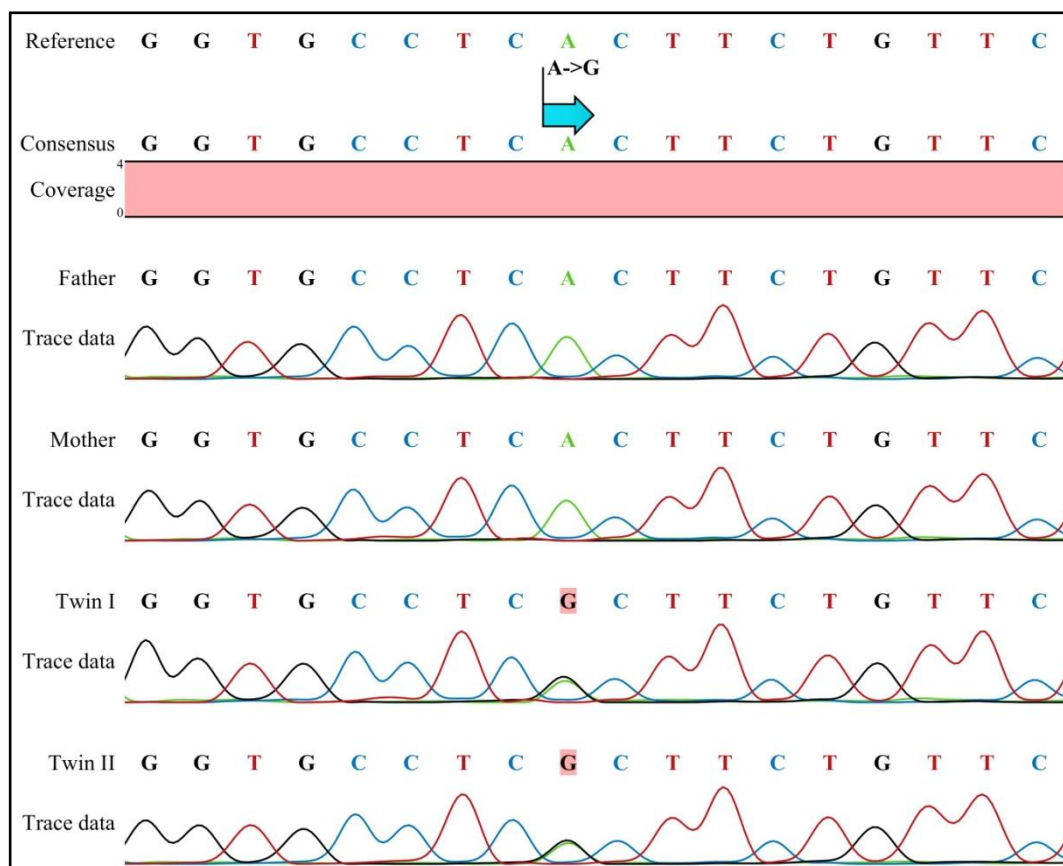


Figure D.17: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome16: 59139115: A/G).

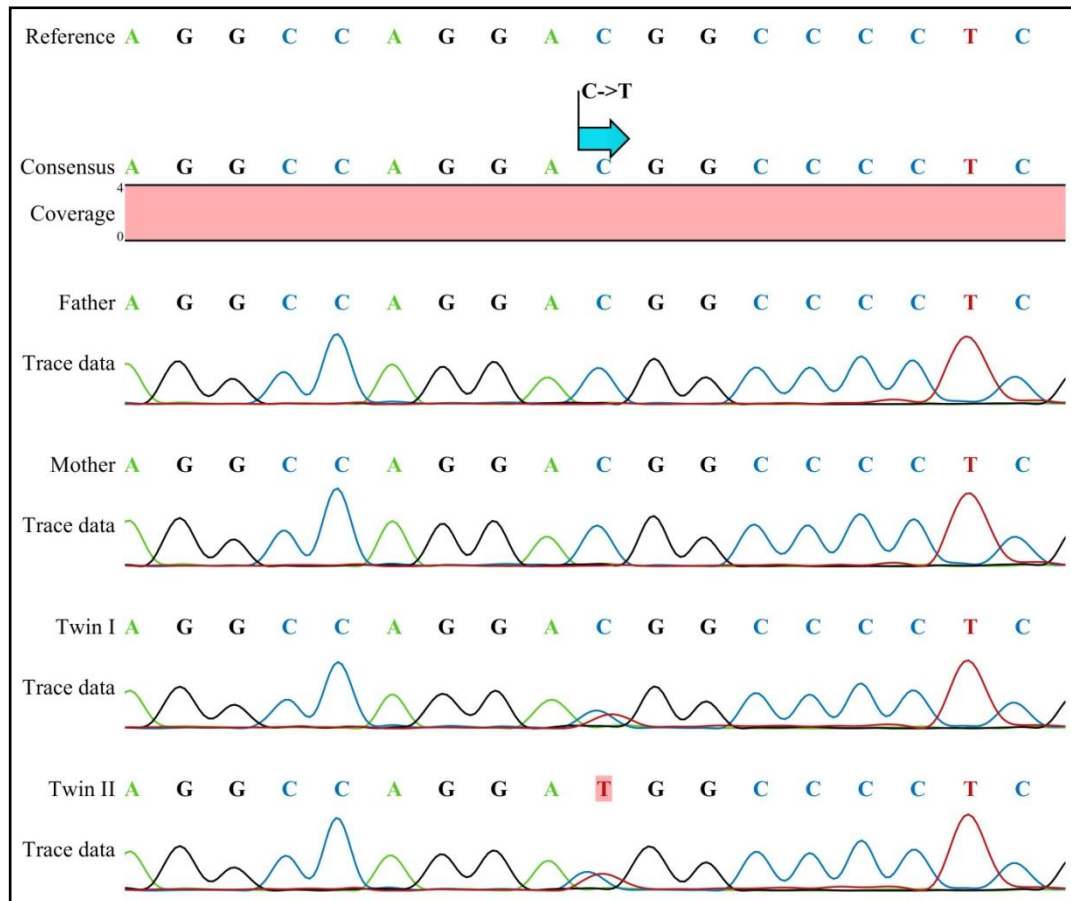


Figure D.18: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome17: 79276311: C/T).

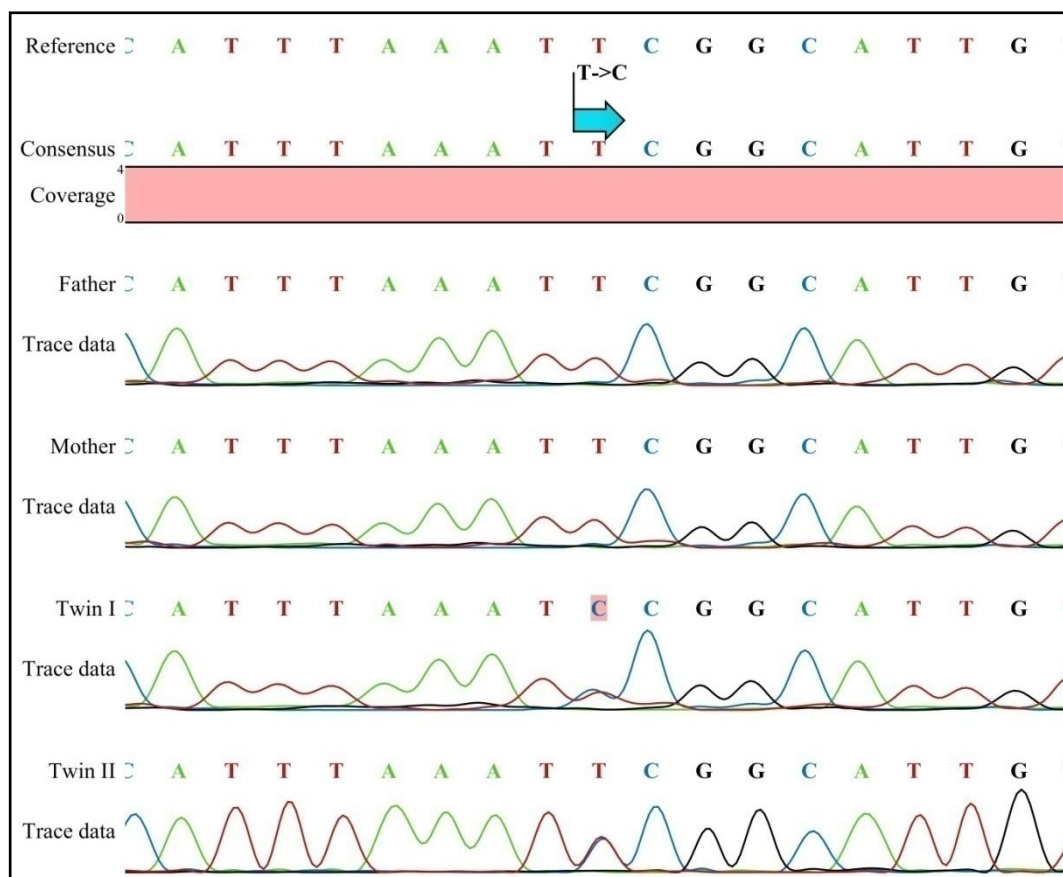


Figure D.19: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome18: 32459844: T/C).

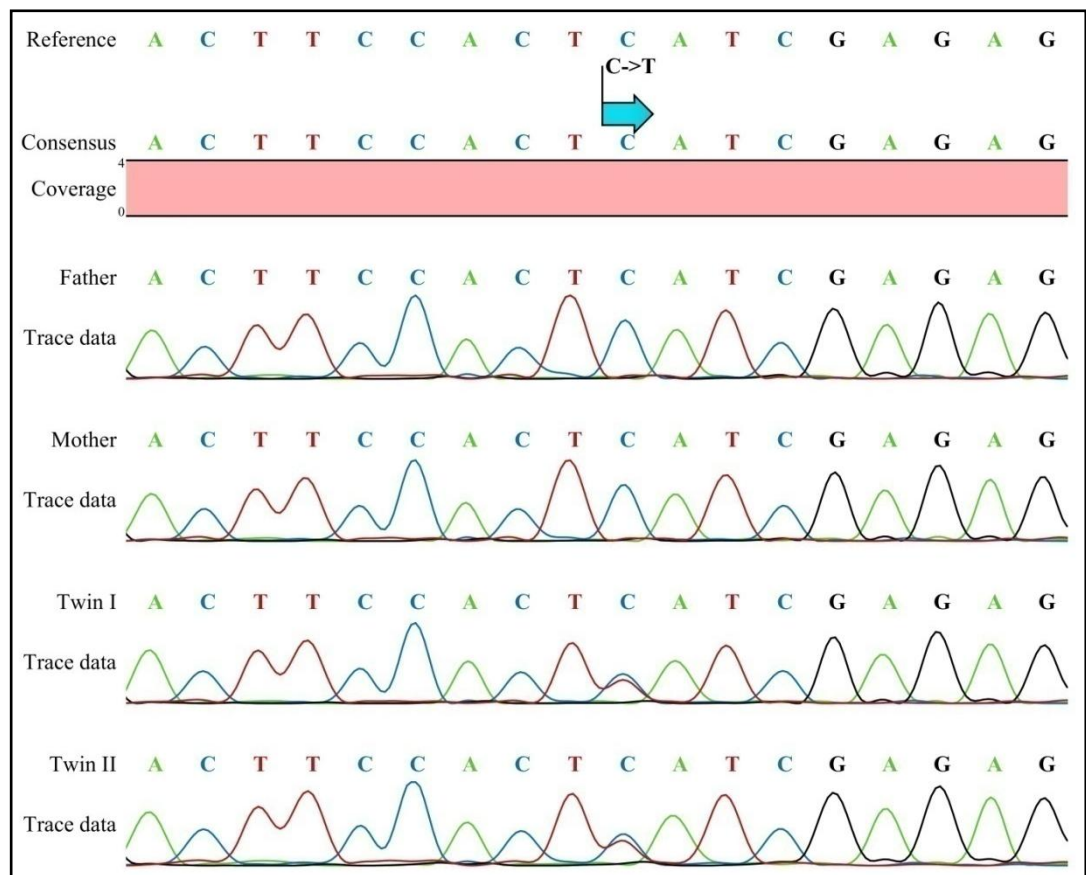


Figure D.20: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome21: 23811291: C/T).

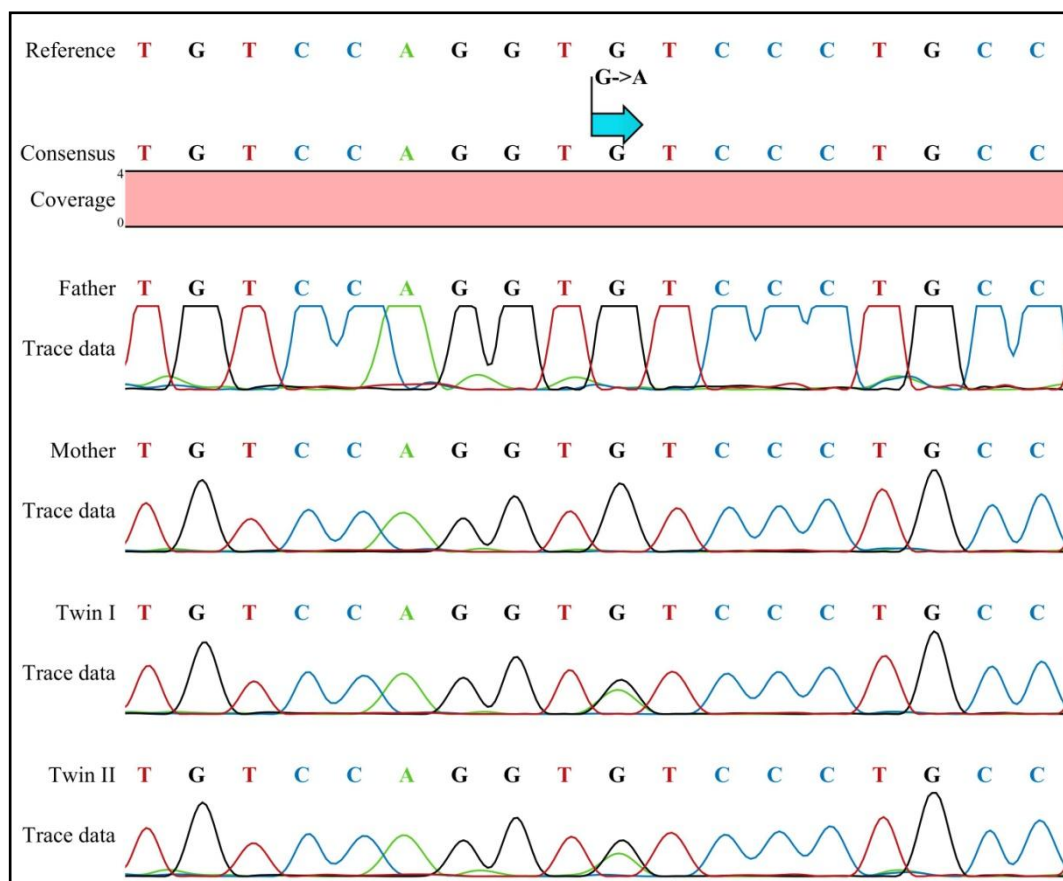


Figure D.21: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome22: 19167292: G/A).

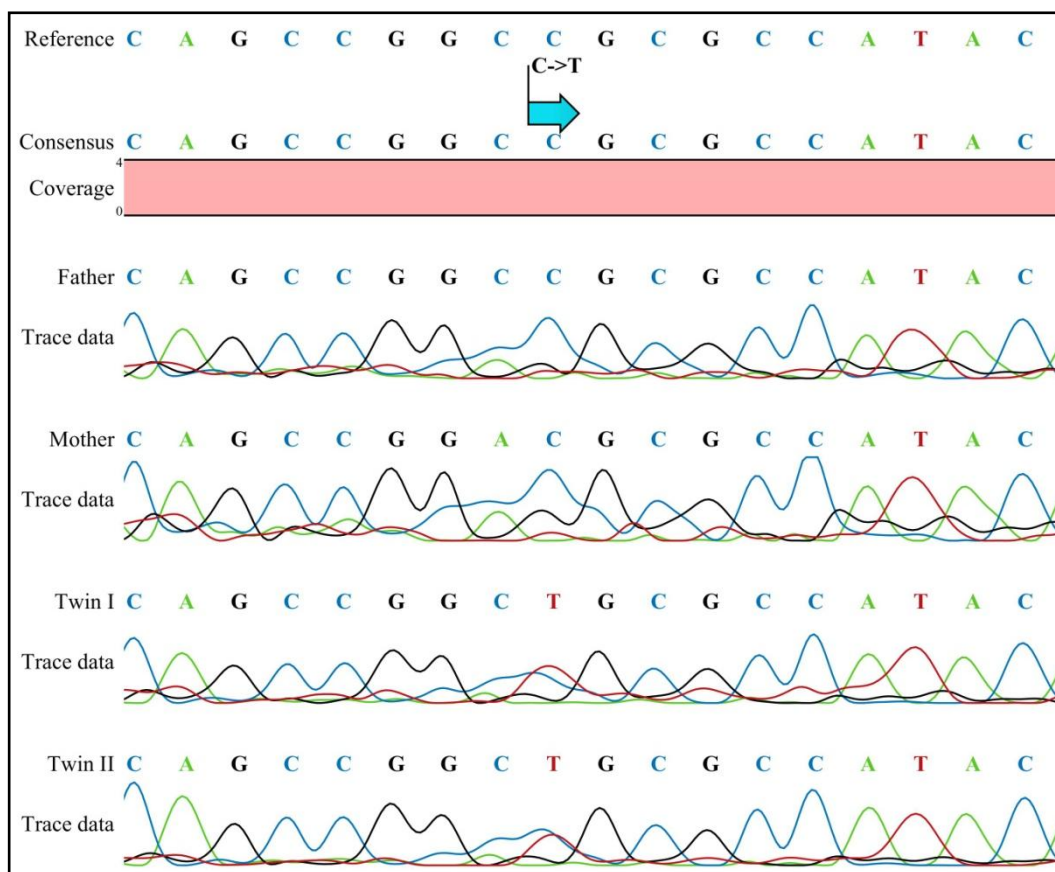


Figure D.22: Electropherogram for the *de novo* SNV shared by the twin pair (Chromosome10: 133937827: C/T).

Appendix E

Electropherograms for the Twin-Specific *De Novo* SNVs Validated through Sanger Sequencing

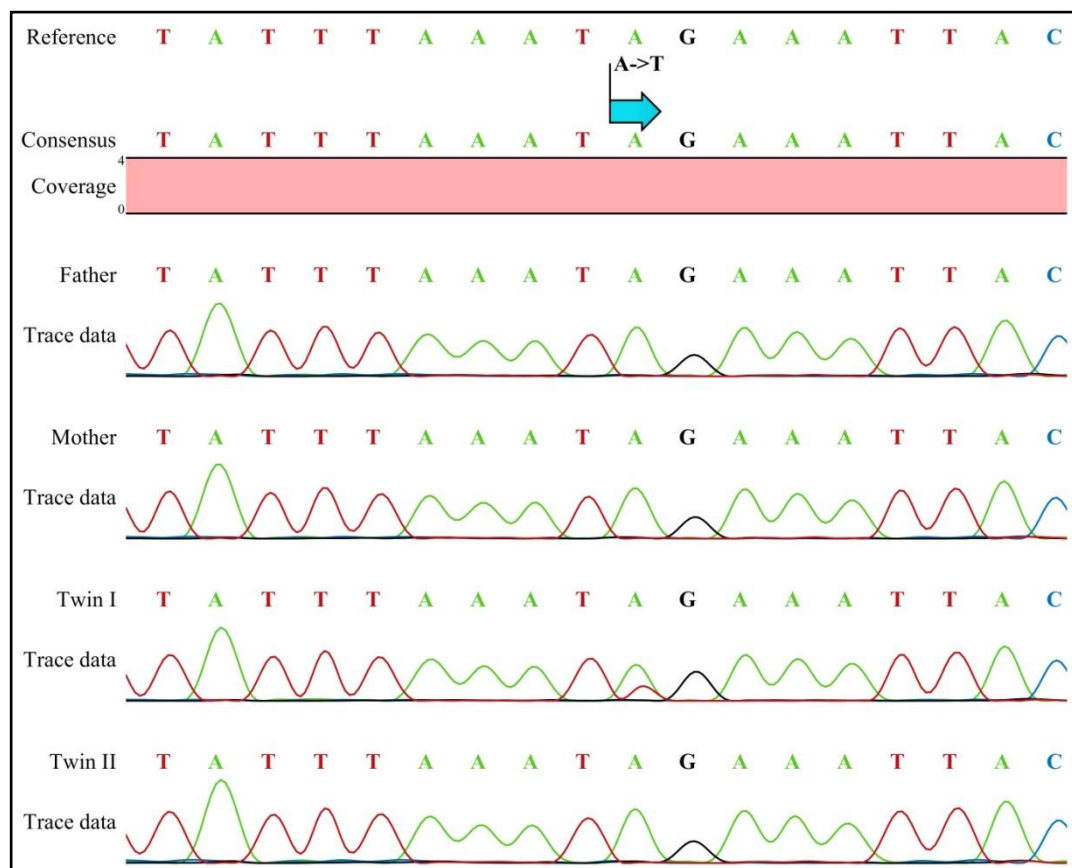


Figure E.1: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome2: 61728351: A/T).

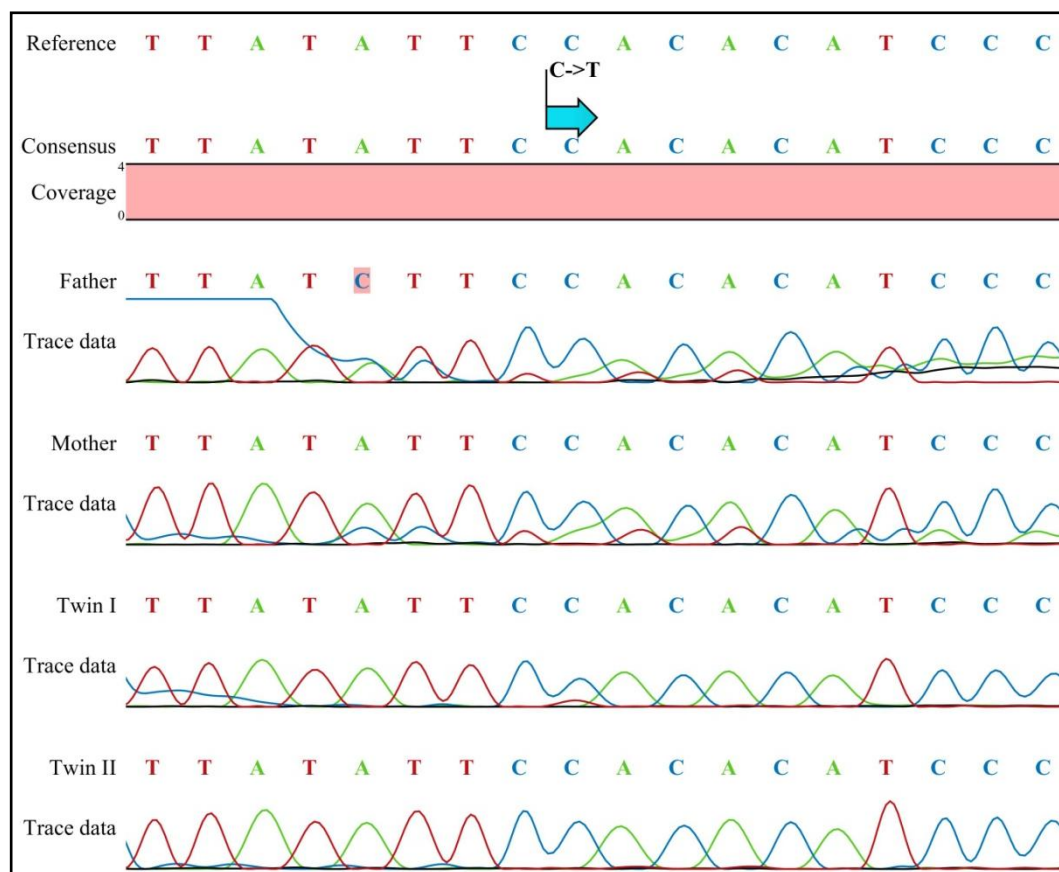


Figure E.2: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome3: 1493383: C/T).

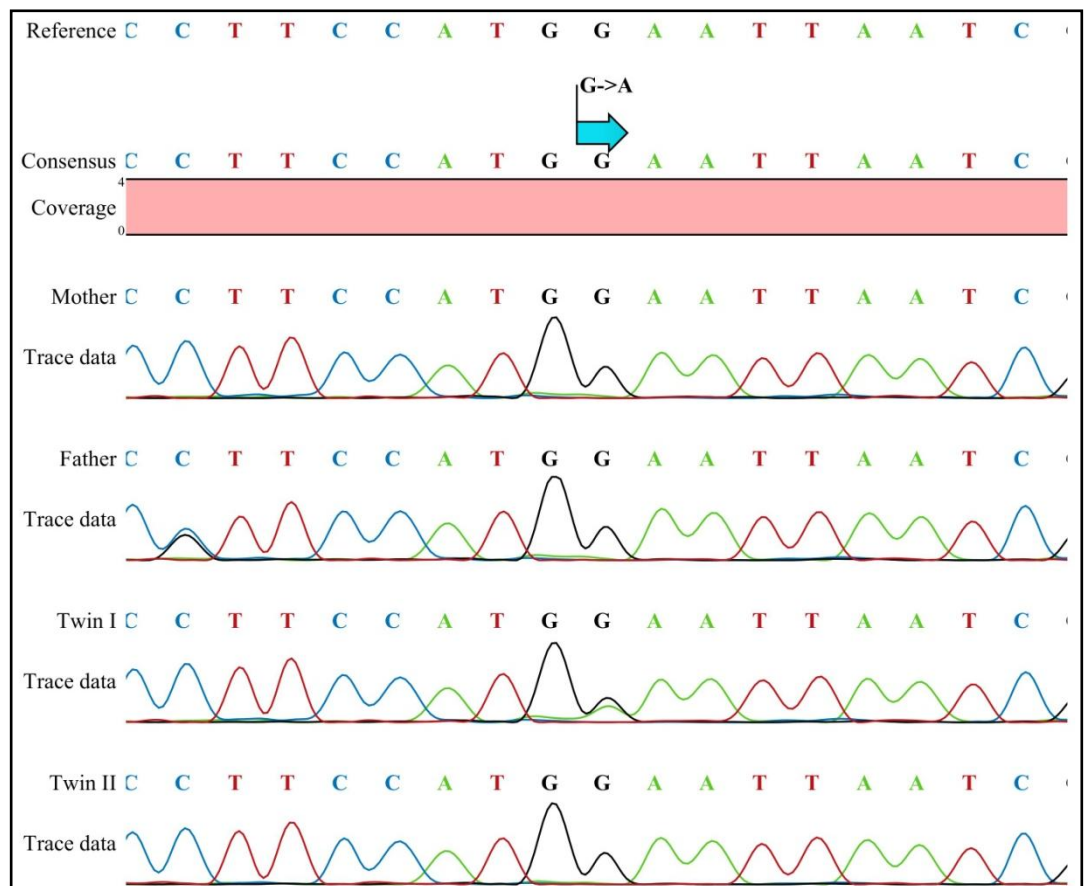


Figure E.3: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome3: 72959480: G/A).

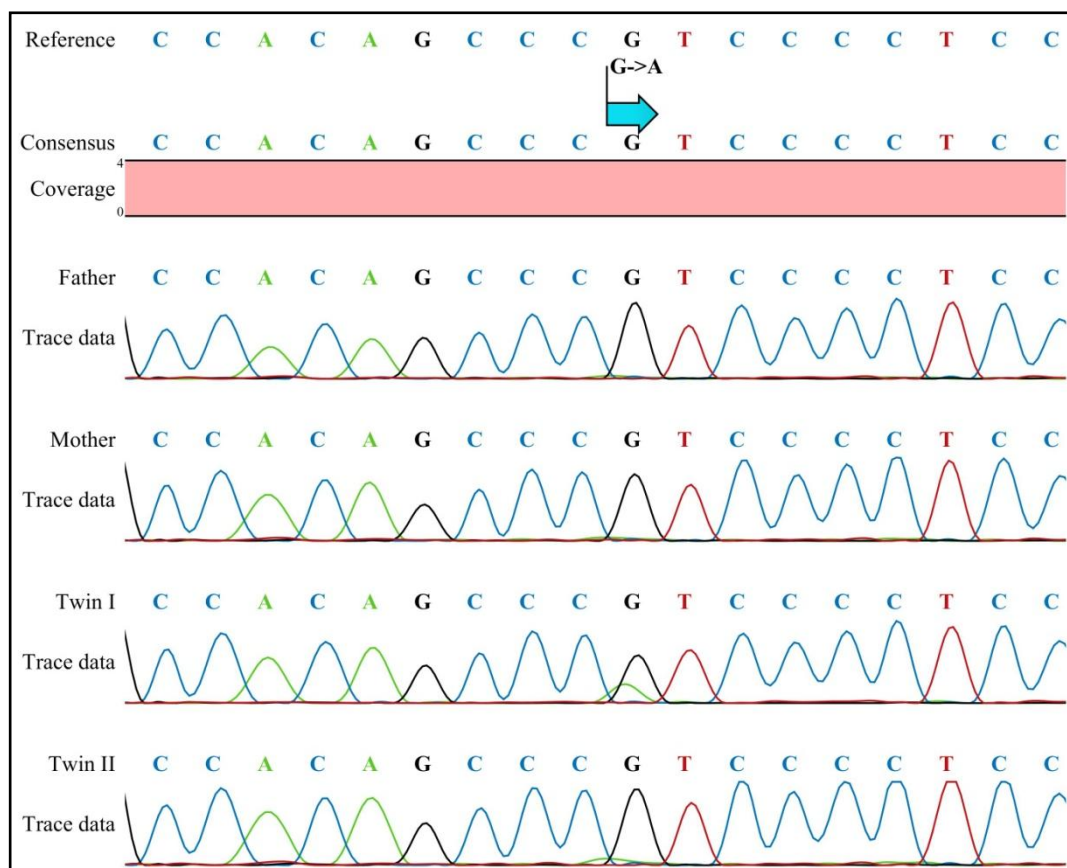


Figure E.4: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome4: 6564000: G/A).

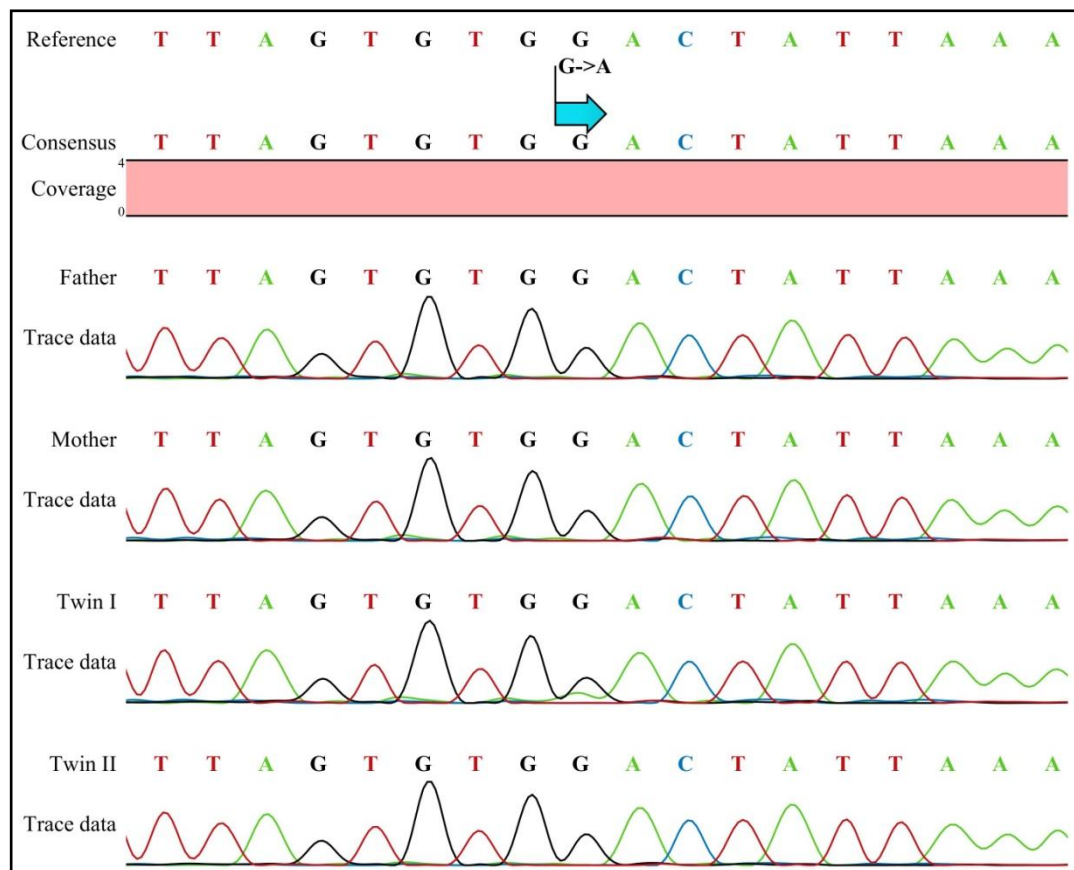


Figure E.5: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome7: 29606885: G/A).

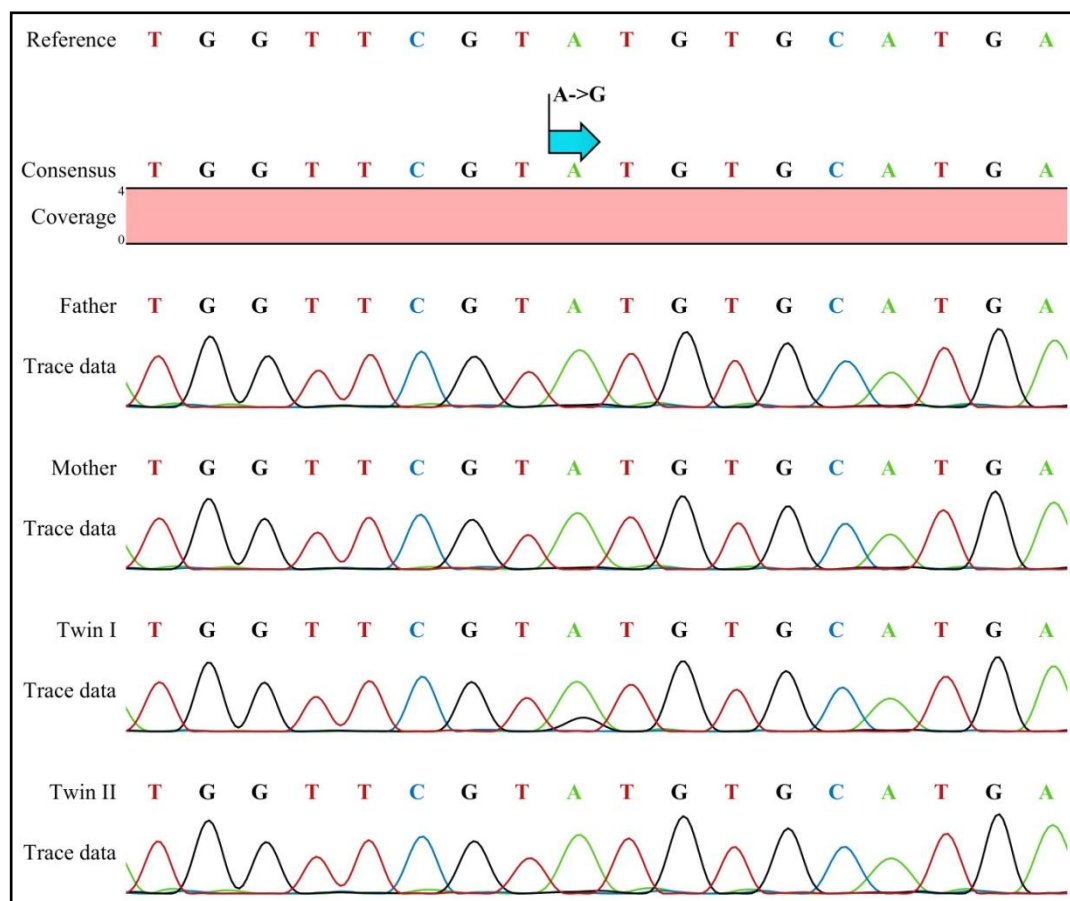


Figure E.6: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome8: 118838053: A/G).

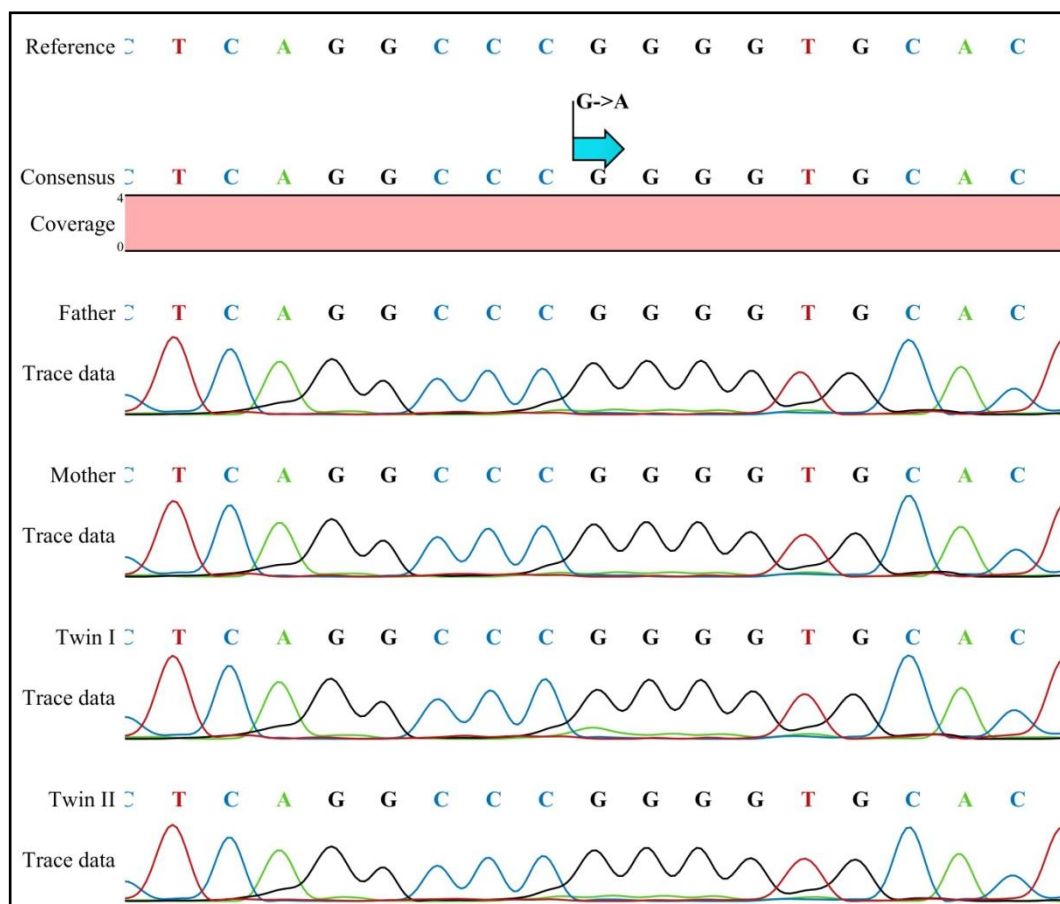


Figure E.7: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome16: 50424085: G/A).

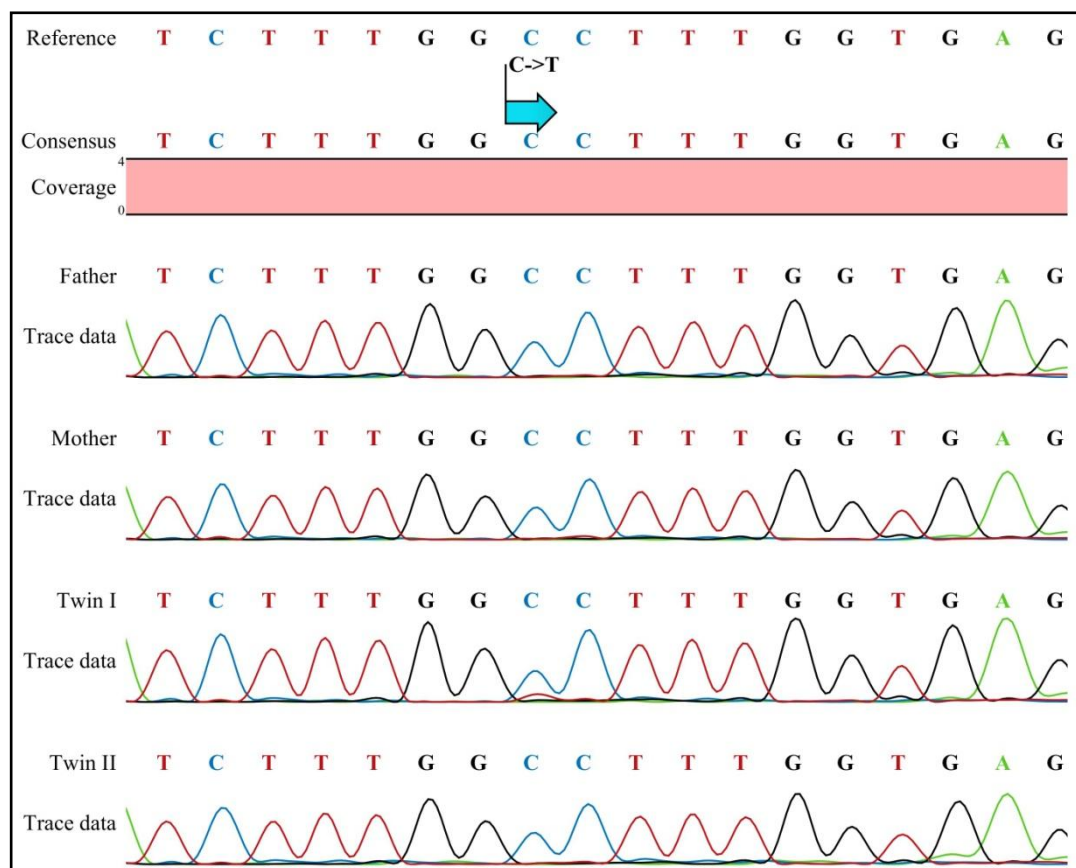


Figure E.8: Electropherogram for the *de novo* SNV specific to Twin-I (Chromosome20: 53271335: C/T).

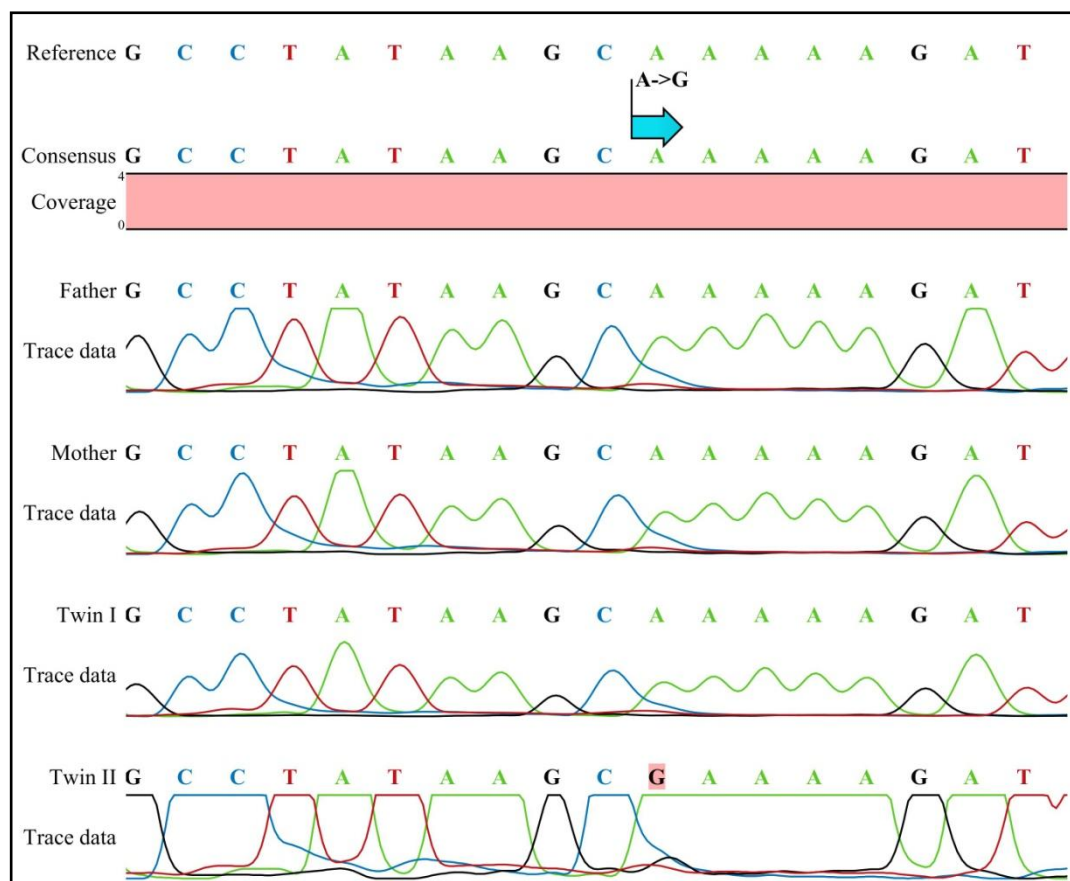


Figure E.9: Electropherogram for the *de novo* SNV specific to Twin-II (Chromosome21: 36322701: A/G).

Appendix F

Electropherograms for the *De Novo* SNVs Tested in the Parental Endoderm- and Ectoderm- -Derived DNA Samples

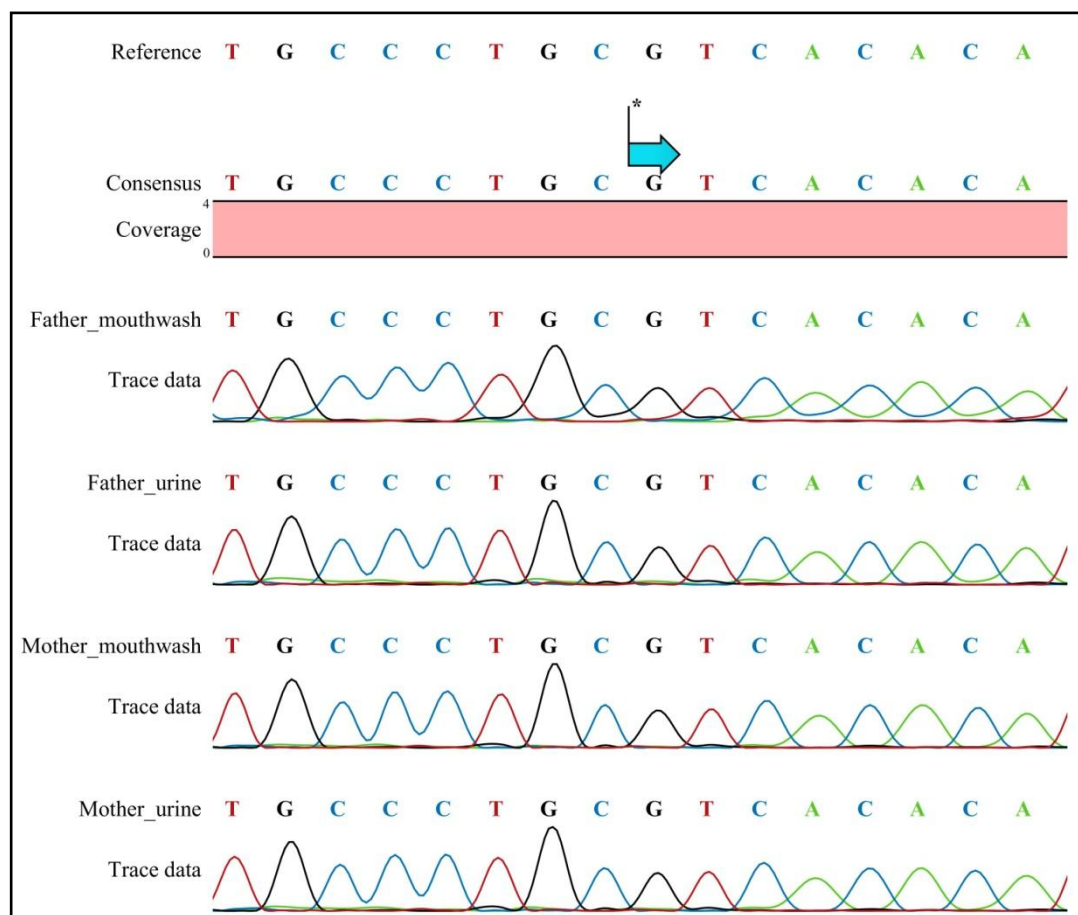


Figure F.1: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome1: 245093862: G/A)

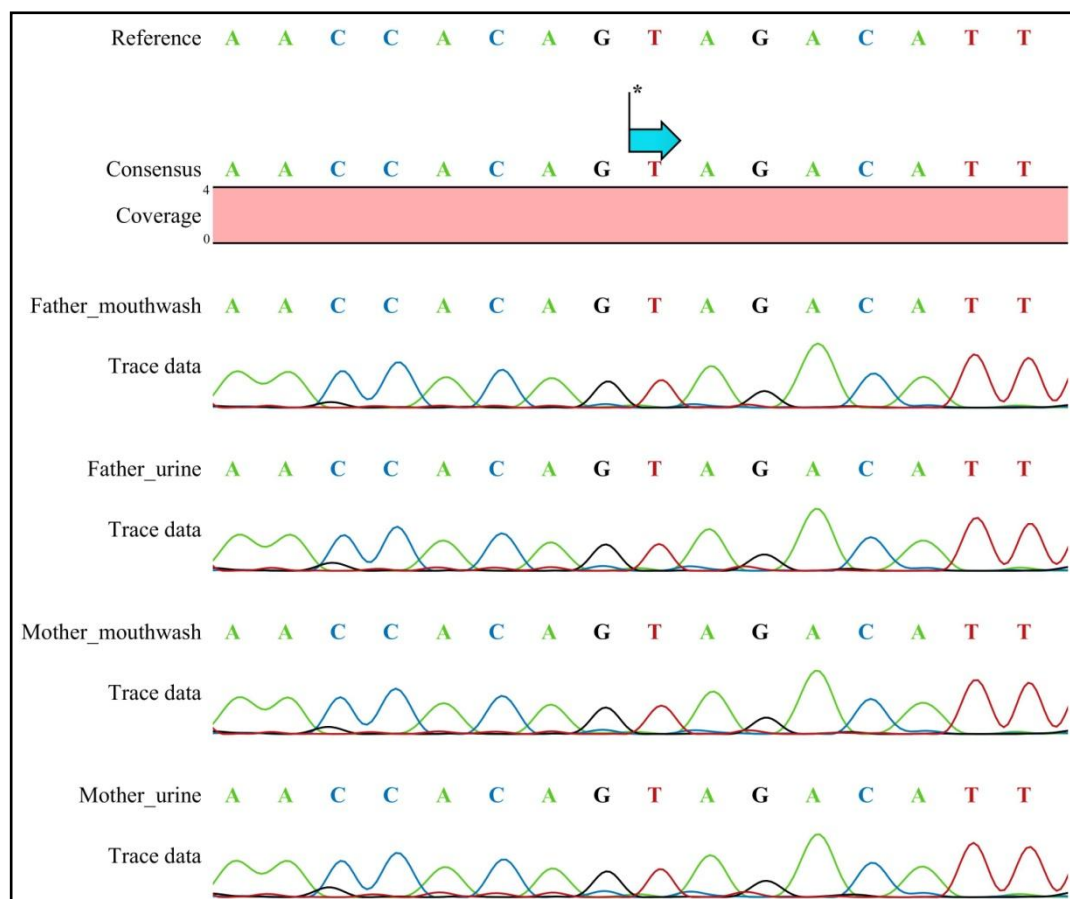


Figure F.2: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome5: 11043438: T/G).

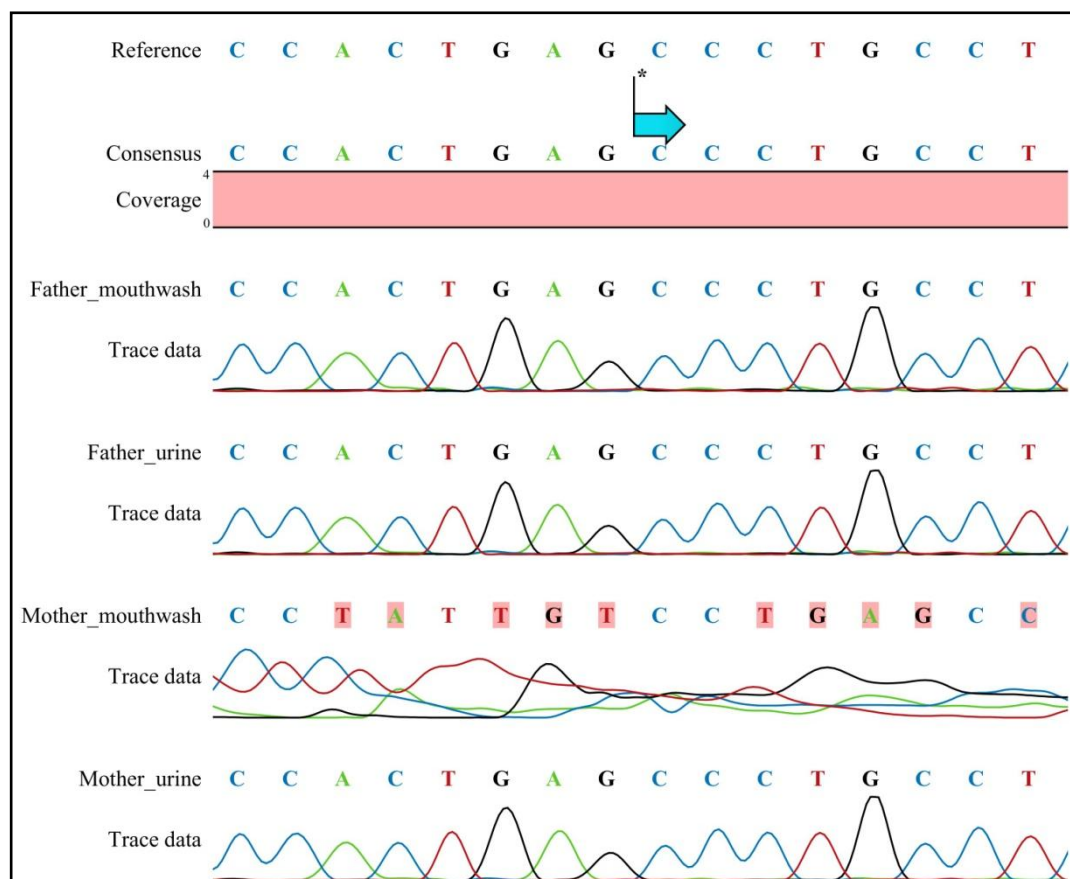


Figure F.3: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome6: 29101417: C/T).

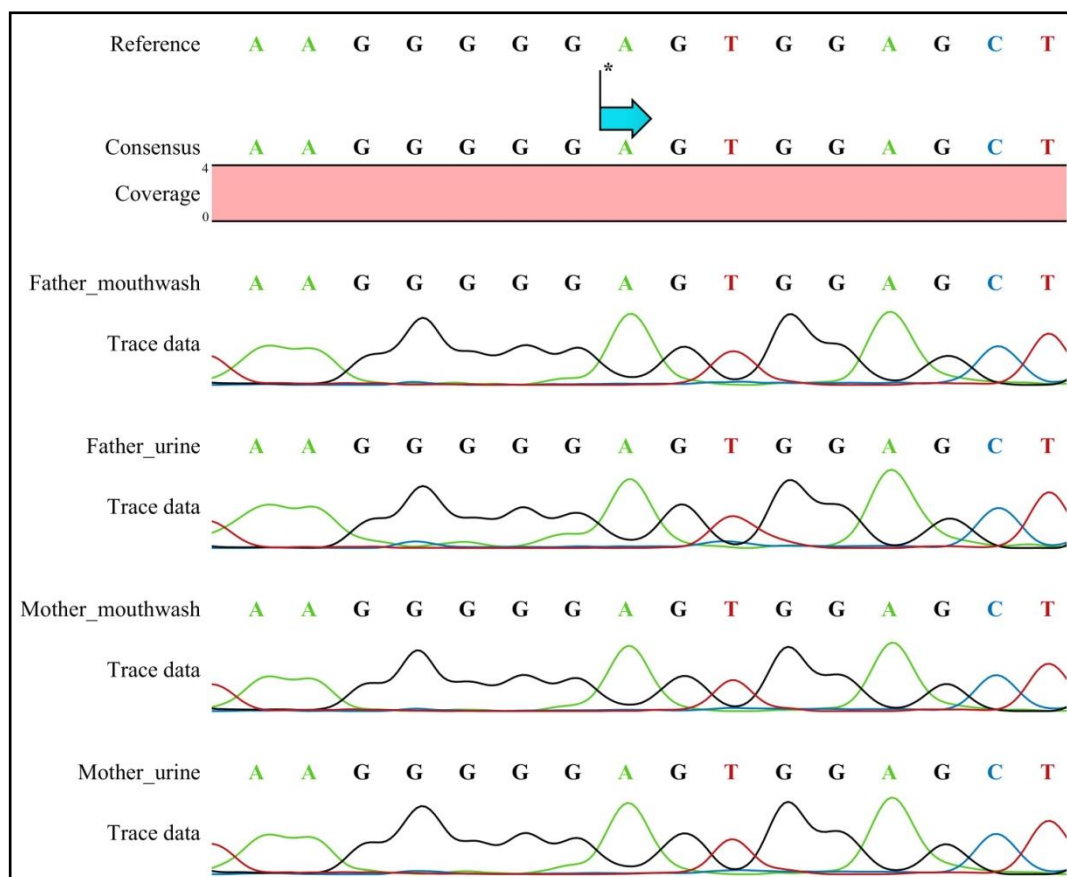


Figure F.4: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome7: 141955453: A/G).

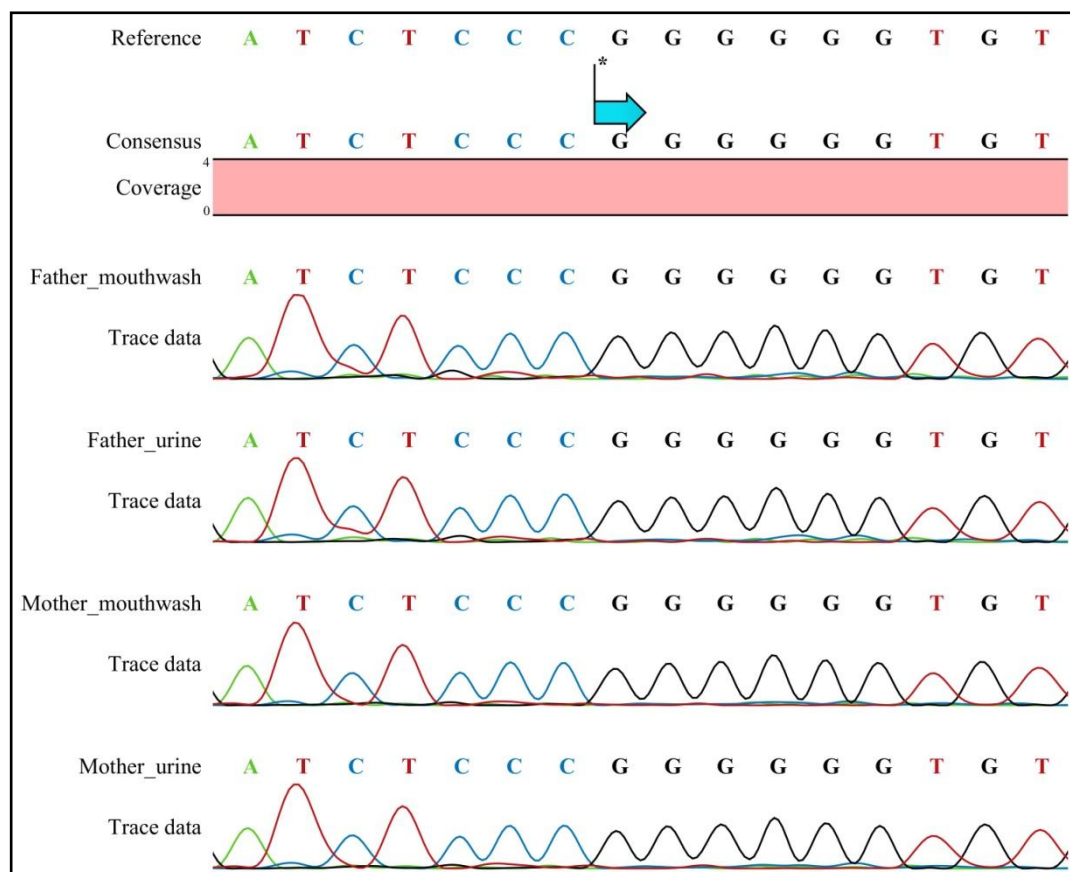


Figure F.5: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome10: 128805987: G/A).

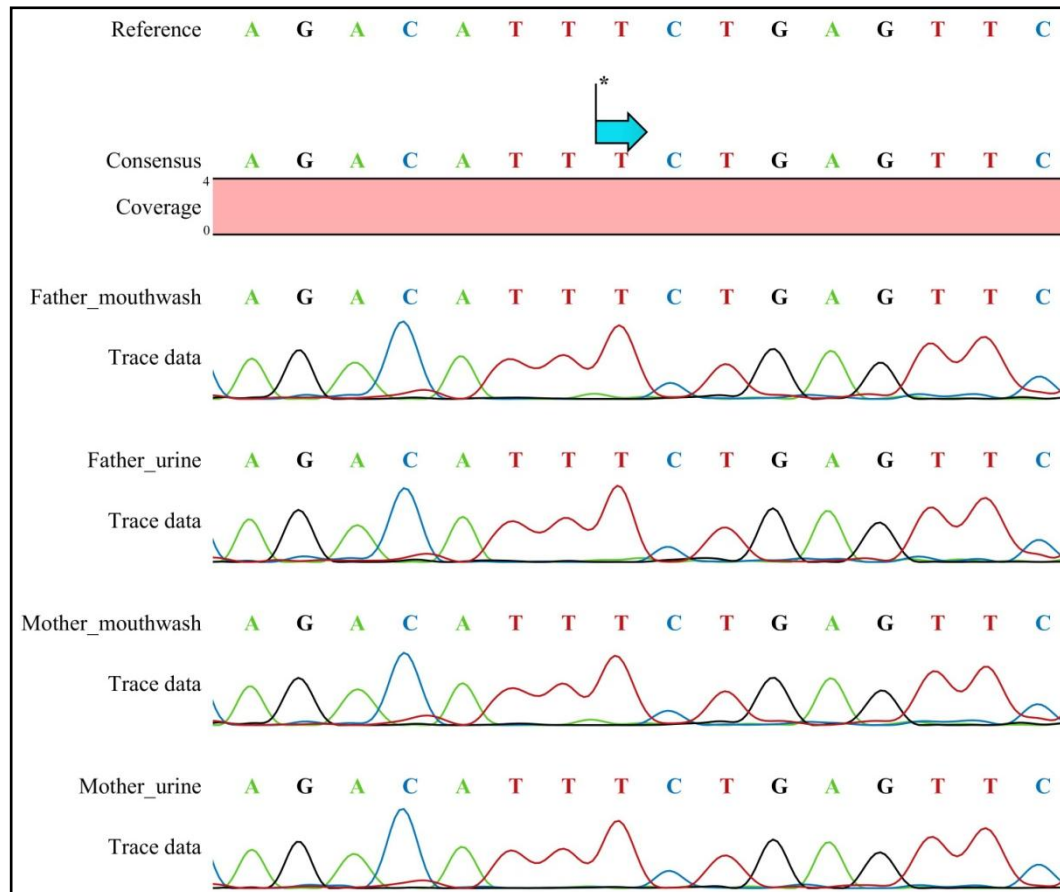


Figure F.6: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome12: 3527384: T/C).

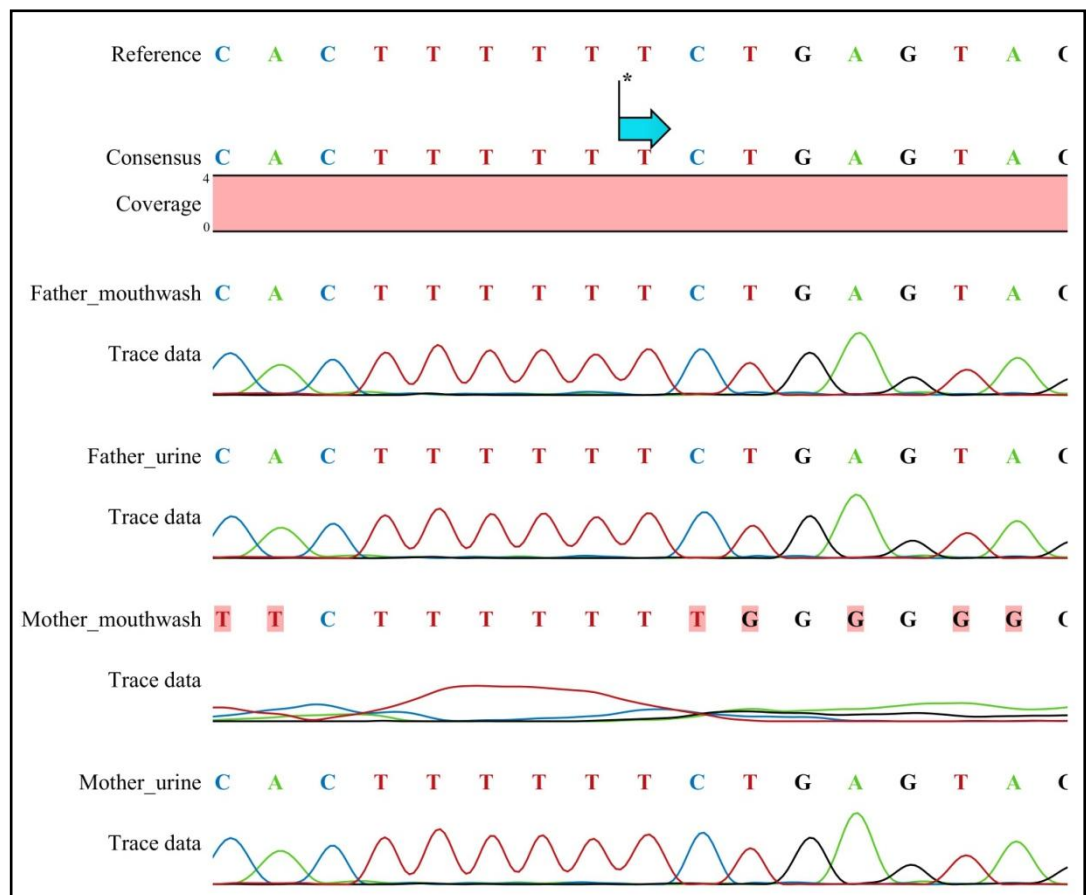


Figure F.7: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome14: 32048285: T/A).

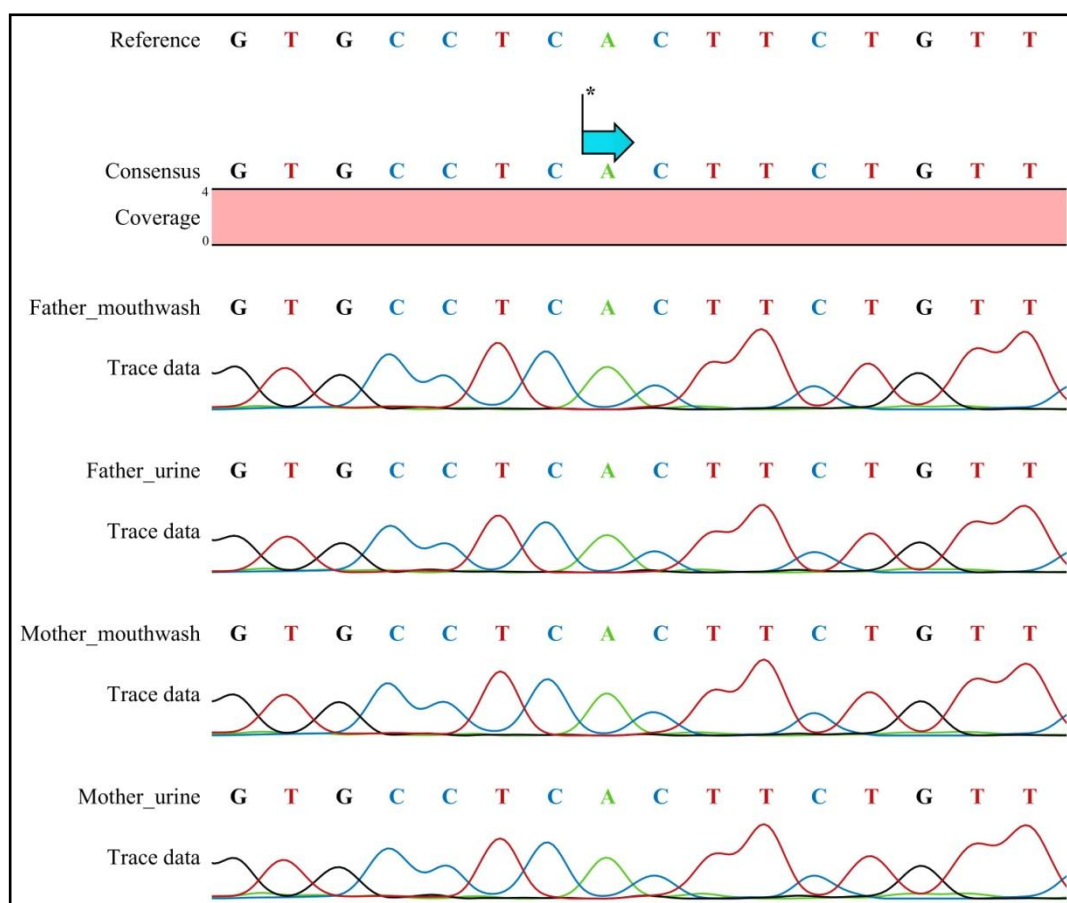


Figure F.8: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome16: 59139115: A/G).

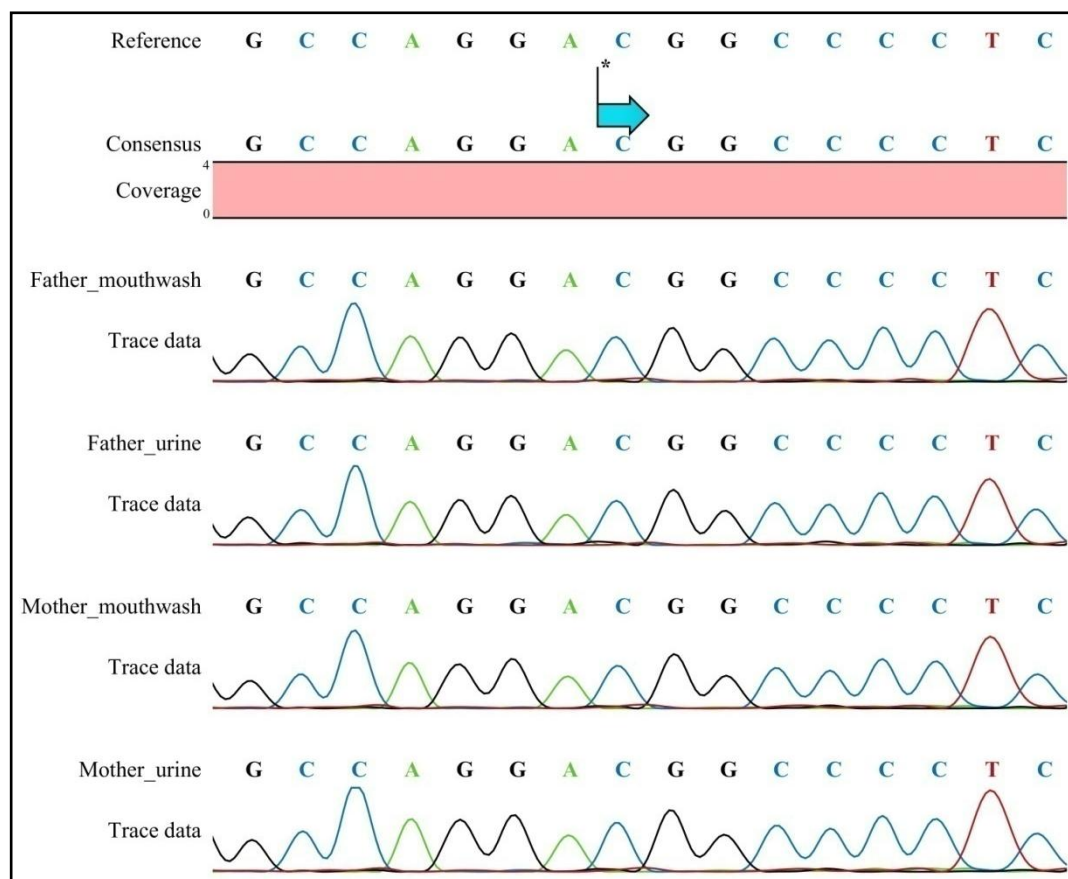


Figure F.9: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome17: 79276311: C/T).

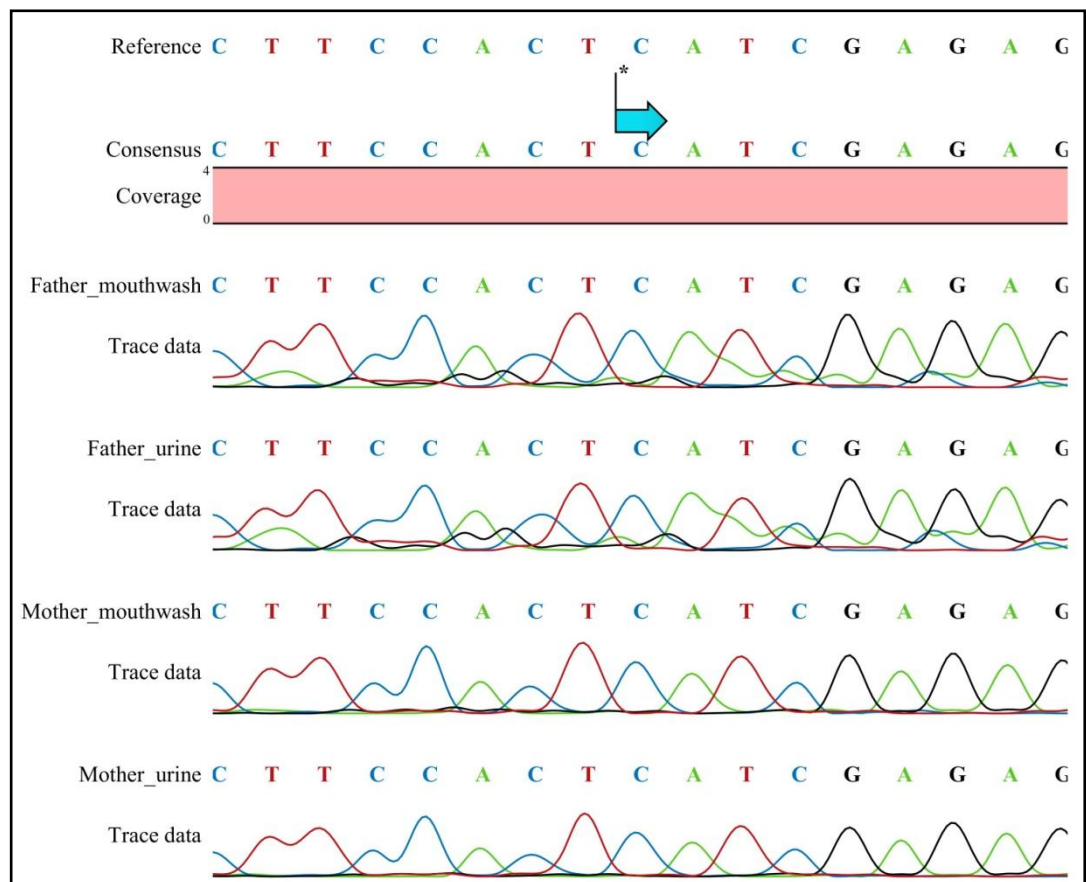


Figure F.10: Electropherogram of the Sanger sequencing of the parental mouthwash and urine sample-derived DNA for the *de novo* SNV shared by the twin pair (Chromosome21: 23811291: C/T).

Appendix G

Heterozygous LoF Mutations Present in the Genomes of the Twin Pair

Table G.1: List of novel heterozygous LoF mutations present in the genomes of the twin pair

Chr	Position	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
1	17085995	-	C	<i>MST1L</i>	frameshift insertion	NA	-	-
1	54605318	-	GG	<i>CDCP2</i>	frameshift insertion	NA	-	-
1	112020739	A	-	<i>C1orf162</i>	frameshift deletion	NA	-	-
3	44763789	AGTG	-	<i>ZNF502</i>	frameshift deletion	NA	-	-
3	53910109	-	G	<i>ACTR8</i>	splice site	NA	-	-
3	54952510	CT	-	<i>LRTM1</i>	frameshift deletion	NA	-	-
3	121207510	AATAGTA	-	<i>POLQ</i>	frameshift deletion	NA	-	-
4	54327211	-	A	<i>LNK1</i>	splice site	NA	-	-
5	171765480	CAAA	-	<i>SH3PXD2B</i>	frameshift deletion	NA	FRANK-TER HAAR SYNDROME (#249420)	AR
6	90577707	G	-	<i>CASP8AP2</i>	frameshift deletion	NA	-	-

6	90577712	CTTTGCCCAGACATGGA	-	<i>CASP8AP2</i>	frameshift deletion	NA	-	-
7	81695842	-	A	<i>CACNA2D1</i>	splice site	NA	-	-
8	90796369	-	A	<i>RIPK2</i>	splice site	NA	-	-
8	103250839	-	G	<i>RRM2B</i>	frameshift insertion	0.06	PROGRESSIVE EXTERNAL OPHTHALMOPLEGIA (#613077)	AD
9	40774307	CTCC	-	<i>ZNF658</i>	frameshift deletion	NA	-	-
9	125273385	-	T	<i>OR1J2</i>	frameshift insertion	0.06	-	-
10	27702260	-	T	<i>PTCHD3</i>	frameshift insertion	NA	-	-
10	37425563	C	T	<i>ANKRD30A</i>	stopgain SNV	NA	-	-
11	1016471	T	-	<i>MUC6</i>	frameshift deletion	NA	-	-
11	4592708	-	AG	<i>C11orf40</i>	frameshift insertion	0.62	-	-
11	10711918	-	A	<i>MRVI1</i>	splice site	NA	-	-
12	48596875	-	A	<i>OR10AD1</i>	frameshift insertion	0.21	-	-
13	31287978	-	GTGT	<i>ALOX5AP</i>	frameshift insertion	NA	STROKE, ISCHEMIC susceptibility (#601367)	-
13	78173524	G	A	<i>SCEL</i>	splice site	NA	-	-
16	1823390	G	-	<i>EME2</i>	frameshift deletion	NA	-	-
16	55844924	-	A	<i>CES1</i>	splice site	NA	-	-
17	46882287	CTTTCCAAAGTCGGGCA	-	<i>TTLL6</i>	frameshift deletion	NA	-	-
17	57141768	-	A	<i>TRIM37</i>	splice site	NA	MULIBREY NANISM (#253250)	AR
19	20045260	AT	-	<i>ZNF93</i>	frameshift deletion	NA	-	-
19	35612163	CC	-	<i>FXVD3</i>	frameshift deletion	NA	-	-
19	53046581	-	AA	<i>ZNF808</i>	frameshift insertion	NA	-	-

Table G.2: List of dbSNP137-reported (MAF < 0.01) heterozygous LoF mutations present in the genomes of the twin pair

Chr	Position	Ref	Alt	Gene	Function	1000G	dbSNP137	OMIM Disease	Inheritance
6	109589014	T	A	<i>LOC100996634</i>	stopgain SNV	0.0005	rs191936197	-	-
12	53167416	C	A	<i>KRT76</i>	stopgain SNV	0.0037	rs149868801	-	-
3	112253058	-	A	<i>ATG3</i>	frameshift insertion	NA	rs139705694	-	-
6	31380161	-	CT	<i>MICA</i>	frameshift insertion	NA	rs41293539	-	-
10	51768675	AA	-	<i>AGAP6</i>	frameshift deletion	NA	rs141217862	-	-
16	450140	-	AG	<i>NME4</i>	frameshift insertion	NA	rs35963490	-	-
16	58577316	A	-	<i>CNOT1</i>	frameshift deletion	NA	rs5817153	-	-
17	17697102	G	-	<i>RAI1</i>	frameshift deletion	NA	rs34083643	SMITH-MAGENIS SYNDROME (#182290)	AD
18	42456670	-	TCTT	<i>SETBP1</i>	frameshift insertion	NA	rs3085861	SCHINZEL-GIEDION MIDFACE RETRACTION SYNDROME (#269150)	AD
5	132085115	G	C	<i>CCNI2</i>	splice site	0.0005	rs140882152	-	-
10	46965887	-	G	<i>SYT15</i>	splice site	NA	rs112965082	-	-
10	50507279	G	A	<i>C10orf71</i>	splice site	0.0009	rs150519054	-	-
12	56500502	-	A	<i>PA2G4</i>	splice site	NA	rs34728522	-	-
21	47707039	-	AAAAAA	<i>YBEY</i>	splice site	NA	rs71318058	-	-

Appendix H

Homozygous LoF Mutations Present in the Genomes of the Parents

Table H.1: List of novel homozygous LoF mutations present in the genome of the mother

Chr	Position	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
1	152195729	T	-	<i>HRNR</i>	frameshift deletion	NA	-	-
10	116931101	-	TT	<i>ATRNL1</i>	frameshift insertion	0.96	-	-
11	76954788	-	A	<i>GDPD4</i>	frameshift insertion	0.23	-	-
12	55714406	-	A	<i>OR6C1</i>	frameshift insertion	0.34	-	-
17	6555546	-	G	<i>C17orf100</i>	frameshift insertion	1.00	-	-
17	71752500	-	T	<i>LOC100134391</i>	frameshift insertion	1.00	-	-
19	51835892	-	G	<i>VSIG10L</i>	frameshift insertion	NA	-	-
X	122336600	-	G	<i>GRIA3</i>	frameshift insertion	1.00	MENTAL RETARDATION, X-LINKED (#300699)	XR

8	103250839	-	G	<i>RRM2B</i>	frameshift insertion	0.06	PROGRESSIVE EXTERNAL OPHTHALMOPLÉGIA (#613077)	AD
9	40774307	CTCC	-	<i>ZNF658</i>	frameshift deletion	NA		-
4	6086694	-	A	<i>JAKMIP1</i>	splice site	NA		-
16	55844924	-	A	<i>CES1</i>	splice site	NA		-

Table H.2: List of dbSNP137-reported (MAF < 0.01) homozygous LoF mutations present in the genome of the mother

Chr	Position	Ref	Alt	Gene	Function	1000G	snp137	OMIM Disease	Inheritance
1	156565049	-	AC	<i>GPATCH4</i>	frameshift insertion	NA	rs112595382	-	-
3	129814935	C	-	<i>ALG1L2</i>	frameshift deletion	NA	rs55800015	-	-
5	156721863	-	C	<i>CYFIP2</i>	frameshift insertion	NA	rs76159126	-	-
5	175811094	-	GT	<i>NOP16</i>	frameshift insertion	NA	rs78549596	-	-
10	51768675	AA	-	<i>AGAP6</i>	frameshift deletion	NA	rs141217862	-	-
15	31521506	GG	-	<i>RP11-16E12.2</i>	frameshift deletion	NA	rs141879900	-	-
16	89291210	-	GTGA	<i>ZNF778</i>	frameshift insertion	NA	rs113560259	-	-
22	19189003	-	C	<i>CLTCL1</i>	frameshift insertion	NA	rs78649162	-	-
X	101395780	-	G	<i>TCEAL6</i>	frameshift insertion	NA	rs200987681	-	-
3	52872001	TGTGTGTGCACGTGCACG	-	<i>TMEM110</i>	splice site	NA	rs113892040	-	-
9	73458045	-	A	<i>TRPM3</i>	splice site	NA	rs3833697	-	-
11	46342259	-	G	<i>CREB3L1</i>	splice site	NA	rs79068197	-	-
11	95615662	TAAAA	-	<i>MTMR2</i>	splice site	NA	rs112378876	CHARCOT-MARIE-TOOTH DISEASE, TYPE 4B1(#601382)	AR
12	7080212	T	C	<i>EMG1</i>	splice site	NA	rs17857448		-
17	61660890	G	-	<i>DCAF7</i>	splice site	NA	rs66502552		-
19	16268208	A	-	<i>HSH2D</i>	splice site	NA	rs5827321		-

Table H.3: List of novel homozygous LoF mutations present in the genome of the father

Chr	Position	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
1	17085995	-	C	<i>MST1L</i>	frameshift insertion	NA	-	-
1	54605318	-	GG	<i>CDCP2</i>	frameshift insertion	NA	-	-
1	152195729	T	-	<i>HRNR</i>	frameshift deletion	NA	-	-
3	53910109	-	G	<i>ACTR8</i>	splice site	NA	-	-
4	6086694	-	A	<i>JAKMIP1</i>	splice site	NA	-	-
9	70900911	-	A	<i>CBWD3</i>	frameshift insertion	NA	-	-
9	125273385	-	T	<i>OR1J2</i>	frameshift insertion	0.06	-	-
10	27702260	-	T	<i>PTCHD3</i>	frameshift insertion	NA	-	-
10	116931101	-	TT	<i>ATRNL1</i>	frameshift insertion	0.96	-	-
12	55714406	-	A	<i>OR6C1</i>	frameshift insertion	0.34	-	-
17	6555546	-	G	<i>C17orf100</i>	frameshift insertion	1.00	-	-
17	71752500	-	T	<i>LOC100134391</i>	frameshift insertion	1.00	-	-
19	51835892	-	G	<i>VSIG10L</i>	frameshift insertion	NA	-	-
X	122336600	-	G	<i>GRIA3</i>	frameshift insertion	1.00	MENTAL RETARDATION, X-LINKED (#300699)	XR

Table H.4: List of dbSNP137-reported (MAF < 0.01) homozygous LoF mutations present in the genome of the father

Chr	Position	Ref	Alt	Gene	Function	1000G	snp137	OMIM Disease	Inheritance
1	156565049	-	AC	<i>GPATCH4</i>	frameshift insertion	NA	rs112595382	-	-
5	156721863	-	C	<i>CYFIP2</i>	frameshift insertion	NA	rs76159126	-	-
5	175811094	-	GT	<i>NOP16</i>	frameshift insertion	NA	rs78549596	-	-
15	31521506	GG	-	<i>RP11-16E12.2</i>	frameshift deletion	NA	rs141879900	-	-
16	89291210	-	GTGA	<i>ZNF778</i>	frameshift insertion	NA	rs113560259	-	-
22	19189003	-	C	<i>CLTCL1</i>	frameshift insertion	NA	rs78649162	-	-
X	101395780	-	G	<i>TCEAL6</i>	frameshift insertion	NA	rs200987681	-	-
3	52872001	TGTGTGTGCACGTGCACG	-	<i>TMEM110</i>	splice site	NA	rs113892040	-	-
11	46342259	-	G	<i>CREB3L1</i>	splice site	NA	rs79068197	-	-
11	95615662	TAAAA	-	<i>MTMR2</i>	splice site	NA	rs112378876	CHARCOT-MARIE-TOOTH DISEASE (#601382)	AR
12	7080212	T	C	<i>EMG1</i>	splice site	NA	rs17857448	-	-
17	61660890	G	-	<i>DCAF7</i>	splice site	NA	rs66502552	-	-
19	16268208	A	-	<i>HSH2D</i>	splice site	NA	rs5827321	-	-
X	36162684	-	TG	<i>CHDC2</i>	frameshift insertion	NA	rs201605278	-	-
18	74208486	G	-	<i>FLJ44313</i>	splice site	NA	rs3841258	-	-
3	112253058	-	A	<i>ATG3</i>	frameshift insertion	NA	rs139705694	-	-
16	450140	-	AG	<i>NME4</i>	frameshift insertion	NA	rs35963490	-	-
16	58577316	A	-	<i>CNOT1</i>	frameshift deletion	NA	rs5817153	-	-

Appendix I

Heterozygous LoF Mutations Present in the Genomes of the Parents

Table I.1: List of novel heterozygous LoF mutations present in the genome of the mother

Chr	Position	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
3	121207510	AATAGTA	-	<i>POLQ</i>	frameshift deletion	NA	-	-
5	139931628	-	GT	<i>SRA1</i>	frameshift insertion	NA	CAMPOMELIC DYSPLASIA (#114290)	AD
5	139931629	-	G	<i>SRA1</i>	frameshift insertion	NA	CAMPOMELIC DYSPLASIA (#114290)	AD
5	171765480	CAAA	-	<i>SH3PXD2B</i>	frameshift deletion	NA	FRANK-TER HAAR SYNDROME (#249420)	AR
6	90577707	G	-	<i>CASP8AP2</i>	frameshift deletion	NA	-	-
6	90577712	CTTTGCCCAGACATGGA	-	<i>CASP8AP2</i>	frameshift deletion	NA	-	-
8	90796369	-	A	<i>RIPK2</i>	splice site	NA	-	-
8	96047806	-	A	<i>NDUFAF6</i>	splice site	NA	-	-

9	70900911	-	A	<i>CBWD3</i>	frameshift insertion	NA	-	-
10	37425563	C	T	<i>ANKRD30A</i>	stopgain SNV	NA		
11	10711918	-	A	<i>MRVI1</i>	splice site	NA	-	-
13	31287978	-	GTGT	<i>ALOX5AP</i>	frameshift insertion	NA	STROKE, ISCHEMIC susceptibility (#601367)	-
13	78173524	G	A	<i>SCEL</i>	splice site	NA	-	-
16	1823390	G	-	<i>EME2</i>	frameshift deletion	NA	-	-
17	57141768	-	A	<i>TRIM37</i>	splice site	NA	MULIBREY NANISM (#253250)	AR
19	20045260	AT	-	<i>ZNF93</i>	frameshift deletion	NA	-	-
19	35612163	CC	-	<i>FXYD3</i>	frameshift deletion	NA	-	-

Table I.2: List of dbSNP137-reported heterozygous LoF mutations present in the genome of the mother

Chr	Position	Ref	Alt	Gene	Function	1000G	dbSNP137	OMIM Disease	Inheritance
6	109589014	T	A	<i>LOC100996634</i>	stopgain SNV	0.0005	rs191936197	-	-
X	36162684	-	TG	<i>CHDC2</i>	frameshift insertion	NA	rs201605278	-	-
18	74208486	G	-	<i>FLJ44313</i>	splice site	NA	rs3841258	-	-

Table I.3: List of novel heterozygous LoF mutations present in the genome of the father

Chr	Position	Ref	Alt	Gene	Function	1000G	OMIM Disease	Inheritance
5	139931628	-	GT	<i>SRA1</i>	frameshift insertion	NA	CAMPOMELIC DYSPLASIA (#114290)	AD
5	139931629	-	G	<i>SRA1</i>	frameshift insertion	NA	CAMPOMELIC DYSPLASIA (#114290)	AD
8	96047806	-	A	<i>NDUFAF6</i>	splice site	NA	-	-
11	76954788	-	A	<i>GDPD4</i>	frameshift insertion	0.23	-	-
1	112020739	A	-	<i>C1orf162</i>	frameshift deletion	NA	-	-
3	44763789	AGTG	-	<i>ZNF502</i>	frameshift deletion	NA	-	-
3	54952510	CT	-	<i>LRTM1</i>	frameshift deletion	NA	-	-
11	1016471	T	-	<i>MUC6</i>	frameshift deletion	NA	-	-
11	4592708	-	AG	<i>C11orf40</i>	frameshift insertion	0.62	-	-

12	48596875	-	A	<i>OR10AD1</i>	frameshift insertion	0.21	-	-
17	46882287	CTTTCCAAAGTCGGGCA	-	<i>TTLL6</i>	frameshift deletion	NA	-	-
19	53046581	-	AA	<i>ZNF808</i>	frameshift insertion	NA	-	-
4	54327211	-	A	<i>LNK1</i>	splice site	NA	-	-
7	81695842	-	A	<i>CACNA2D1</i>	splice site	NA	-	-

Table I.4: List of dbSNP137-reported heterozygous LoF mutations present in the genome of the father

Chr	Position	Ref	Alt	Gene	Function	1000G	dbSNP137	OMIM Disease	Inheritance
3	129814935	C	-	<i>ALG1L2</i>	frameshift deletion	NA	rs55800015	-	-
9	73458045	-	A	<i>TRPM3</i>	splice site	NA	rs3833697	-	-
12	53167416	C	A	<i>KRT76</i>	stopgain SNV	0.0037	rs149868801	-	-
6	31380161	-	CT	<i>MICA</i>	frameshift insertion	NA	rs41293539	-	-
17	17697102	G	-	<i>RAI1</i>	frameshift deletion	NA	rs34083643	SMITH-MAGENIS SYNDROME (#182290) SCHINZEL-GIEDION MIDFACE RETRACTION SYNDROME (#269150)	AD
18	42456670	-	TCTT	<i>SETBP1</i>	frameshift insertion	NA	rs3085861	(#269150)	AD
5	132085115	G	C	<i>CCNI2</i>	NA	0.0005	rs140882152	-	-
10	46965887	-	G	<i>SYT15</i>	NA	NA	rs112965082	-	-
10	50507279	G	A	<i>C10orf71</i>	NA	0.0009	rs150519054	-	-
12	56500502	-	A	<i>PA2G4</i>	NA	NA	rs34728522	-	-
21	47707039	-	AAAAAA	<i>YBEY</i>	NA	NA	rs71318058	-	-

Publications

Cover image



Cover credit: Detecting post-zygotic SNVs in monozygotic twins, from Dal *et al.* pg 455.

ORIGINAL ARTICLE

Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair

Gülşah M Dal,¹ Bekir Ergüner,² Mahmut S Sağiroğlu,² Bayram Yüksel,³ Onur Emre Onat,¹ Can Alkan,⁴ Tayfun Özçelik^{1,5}

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/jmedgenet-2013-102197>).

¹Department of Molecular Biology and Genetics, Faculty of Science, Bilkent University, Ankara, Turkey

²TÜBİTAK-BİLGEM-UEKAE, Gebze, Kocaeli, Turkey

³TÜBİTAK-MAM-GEPI, Gebze, Kocaeli, Turkey

⁴Faculty of Engineering, Department of Computer Engineering, Bilkent University, Ankara, Turkey

⁵Institute of Materials Science and Nanotechnology (UNAM), Bilkent University, Ankara, Turkey

Correspondence to

Dr Tayfun Özçelik, Faculty of Science, Department of Molecular Biology and Genetics, Bilkent University, Ankara 06800, Turkey; tozcelik@bilkent.edu.tr

Received 9 December 2013

Revised 30 March 2014

Accepted 2 April 2014

ABSTRACT

Background Human de novo single-nucleotide variation (SNV) rate is estimated to range between $0.82\text{--}1.70 \times 10^{-8}$ mutations per base per generation. However, contribution of early postzygotic mutations to the overall human de novo SNV rate is unknown.

Methods We performed deep whole-genome sequencing (more than 30-fold coverage per individual) of the whole-blood-derived DNA samples of a healthy monozygotic twin pair and their parents. We examined the genotypes of each individual simultaneously for each of the SNVs and discovered de novo SNVs regarding the timing of mutagenesis. Putative de novo SNVs were validated using Sanger-based capillary sequencing.

Results We conservatively characterised 23 de novo SNVs shared by the twin pair, 8 de novo SNVs specific to twin I and 1 de novo SNV specific to twin II. Based on the number of de novo SNVs validated by Sanger sequencing and the number of callable bases of each twin, we calculated the overall de novo SNV rate of 1.31×10^{-8} and 1.01×10^{-8} for twin I and twin II, respectively. Of these, rates of the early postzygotic de novo SNVs were estimated to be 0.34×10^{-8} for twin I and 0.04×10^{-8} for twin II.

Conclusions Early postzygotic mutations constitute a substantial proportion of de novo mutations in humans. Therefore, genome mosaicism resulting from early mitotic events during embryogenesis is common and could substantially contribute to the development of diseases.

INTRODUCTION

Mutations provide the means of genetic diversity on which natural selection operates. Characterising the patterns and rate of de novo mutations is crucial to the understanding of mutational processes that may lead to evolution and human disease.¹ Recent studies based on the direct observation of de novo single-nucleotide variation (SNV) in parent–offspring trios revealed a rate that spans a wide range ($0.82\text{--}1.70 \times 10^{-8}$ mutations per base per generation).^{2,3} However, the proportion of pre-zygotic versus early postzygotic mutations within the de novo mutation rate is not yet known.

In an attempt to further elucidate the origins of de novo mutations, we performed deep whole-genome sequencing of a healthy monozygotic twin pair and their parents (figure 1A). Our working hypothesis is that those de novo mutations observed in only one twin would be of mitotic origin, occur early during development possibly in the post-twinning embryo and be detectable in whole-blood-derived constitutive DNA (figure 1B).

We indeed observed, following a conservative characterisation and validation, 23 shared, 8 twin I-specific and 1 twin II-specific de novo SNVs (table 1). These results implicate that a substantial proportion of de novo SNVs in the genome of an individual could be of early postzygotic origin.

STUDY SAMPLES AND METHODS

Study samples

A quad family consisting of a male 25-year-old healthy monozygotic twin pair and their parents (59 and 49 years old) participated in this study. They were recruited to the control group of a study on movement disorders, which was approved by the institutional review boards (IRB) at Bilkent, Hacettepe, Başkent and Çukurova Universities (decisions: BEK02, 28.08.2008; TBK08/4, 22.04.2008; KA07/47, 02.04.2007; and 21/3, 08.11.2005, respectively). Zygosity of the twin pair was assessed by comparing the SNVs of each twin (99.8% similar). A Mendelian disorder has not been documented in the family. The participants had no exposure to chemotherapeutics. Written informed consent prepared according to the guidelines of the Ministry of Health in Turkey was attained from all individuals prior to the initiation of this study. Blood, buccal smear and urine samples were obtained from all participants. DNA isolation from whole blood of each individual was done using NucleoSpin Blood (Macherey-Nagel) kit, according to the manufacturer's protocol. Buccal smear and urine cell-derived DNA were isolated using DNeasy Tissue Kit (QIAGEN).

Whole-genome sequencing

Whole-genome sequencing data were generated from blood DNA using Illumina HiSeq 2000 instrument. DNA samples were prepared for whole-genome sequencing using Illumina TruSeq DNA sample preparation kit according to the manufacturer's recommended protocols. For each of the parent single library and for each of the twins two libraries were prepared. For library preparation, 1–3 µg of genomic DNA was fragmented through sonication. The resulting DNA fragments were end-repaired and adapter ligation was done. End-repaired and adapter-ligated DNA fragments were run on 2% agarose gel for size selection. Following the isolation of the fragments of approximately 400 bp in size from the gel using QIAGEN MinElute Gel Extraction Kit, size-selected libraries were subjected to PCR amplification. Enriched libraries were purified and their

To cite: Dal GM, Ergüner B, Sağiroğlu MS, et al. *J Med Genet* Published Online First: [please include Day Month Year] doi:10.1136/jmedgenet-2013-102197

Somatic mosaicism

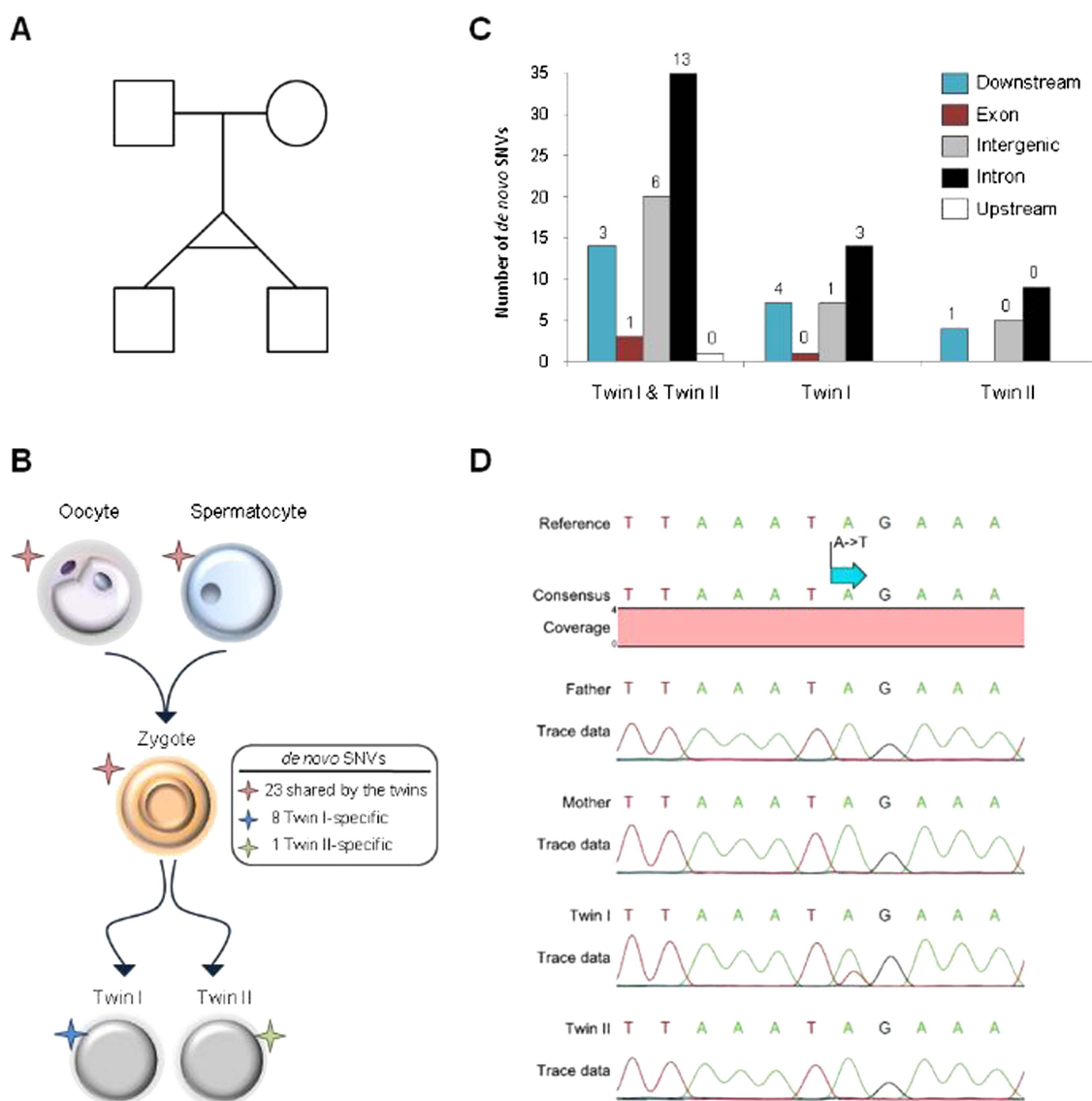


Figure 1 Discovery and characterisation of de novo single-nucleotide variations (SNVs) in a monozygotic twin pair. (A) Pedigree of the whole-genome sequenced quad family. (B) Prezygotic versus postzygotic de novo SNVs found in twins I and II. (C) Annotation of the putative de novo SNVs against human database GRCh37.68 using SnpEff.¹³ twin I and twin II: de novo SNVs shared between the twin pair. Twin I, twin II: twin I-specific and twin II-specific de novo SNVs, respectively. The number of validated SNVs is shown above each bar. The only validated exonic SNV shared by the twins is a synonymous mutation in *PRSS58*. (D) Sanger sequencing chromatogram for a twin I-specific intronic de novo SNV (*XPO1* gene, chr2: 61728351, A>T).

Table 1 De novo mutations of the twin pair and the mutation rate estimation

Sample	All SNVs*		Novel SNVs†		De novo SNVs‡			Validated de novo SNVs§			Mutation rate ($\times 10^{-8}$)¶		
	dbSNP137	Novel	Inherited	De novo**	Group I††	Group II‡‡	Total	Group I	Group II	Total	Group I	Group II	Total
Twin I	3 446 849	49 800	48 405	1396	159	83	242	23	8	31	0.97	0.34	1.31
Twin II	3 443 828	49 182	47 810	1372		48	207		1	24		0.04	1.01

*Number of all SNVs based on the variants calling through GATK.

†Number of SNVs that are not reported dbSNP137.

‡Number of de novo SNVs after applying filters (see online supplementary tables S5–S7).

§De novo SNVs validated through Sanger sequencing.

¶Mutation rate calculated based on the validated de novo SNVs and callable bases of each twin.

**Number of novel SNVs detected as de novo before applying filters (see online supplementary tables S5–S7).

††De novo SNVs classified as occurred in parental gametes or before separation of the zygote into two embryos.

‡‡De novo SNVs classified as occurred after separation of the zygote into two embryos but very early during development.

GATK, Genome Analysis Toolkit; SNV, single-nucleotide variation.

quality was controlled with respect to the recommendations provided by Illumina. Resulting libraries were sequenced on eight lanes for each twin and four lanes for each parent using an Illumina HiSeq 2000 instrument, employing 2×101 –104 cycles of incorporation. Imaging was performed using Illumina SBS kits TruSeq V3. Image analysis and base calling were done using Illumina's Real Time Analysis software V1.13 with default parameters.

Analysis of the whole-genome sequencing data

We converted base calling data into FASTQ files by using Illumina's CASAVA V1.8.2 software package. The paired-end reads were aligned to the NCBI Build 37 version (GRCh37) of the human reference genome using Burrows-Wheeler Aligner (BWA)⁴ V0.6.1 with standard parameters. BAM files were generated from the resulting alignment files and merged into a single BAM file using SAMtools⁵ V0.1.18. We also used SAMtools to mark and remove the PCR duplicates found in the resulting BAM file. SNVs and insertions and/or deletions (indels) were discovered using Genome Analysis Toolkit (GATK)⁶ software V1.6–13 with standard filtering parameters. In order to minimise spurious SNV calls due to misalignment of bases in regions around indels, the reads were locally realigned using RealignerTargetCreator and IndelRealigner tools of GATK. UnifiedGenotyper tool of GATK was used to generate an initial list of raw variants. Finally, variant recalibration with the variant quality score recalibration was applied to the SNVs/indels in order to generate the final list of variants of four individuals in a single VCF file. GATK-recommended hard filtering for very low quality call-sets was applied with 'AB <0.2 || MQ0 >50 || SB >−0.10 || QUAL <10' parameters.

Discovery of the putative de novo SNVs shared by the twins (group I)

We used the following approaches to discover putative de novo SNVs and to classify them according to the timing of mutagenesis. We selectively extracted SNVs (both novel and reported in dbSNP137) using GATK's 'SelectVariants' module. We compared the genotypes of the parents and the twin pair for each of the SNVs concurrently using SnpSift⁷ V1.1.3. We evaluated all possible genotype combinations, including those that could be classified as genotyping errors (such as those SNVs in a homozygous state in one of the twins yet not present in one of the parents). Using BEDtools,⁸ we first filtered the SNVs reported in SNP database dbSNP137. Novel putative de novo SNVs were removed when they were in non-variant sites or in sites for which the read depth is less than 3 in each individual simultaneously. We filtered the variants intersecting segmental duplications and/or simple repeats in the genome using the intersectBed tool within the BEDtools⁸ suite. Finally, we excluded de novo SNV candidates classified as genotyping errors. Remaining de novo SNV candidates were selected for further analysis. We considered the timing of mutagenesis for the shared de novo SNVs as parental or post-zygotic yet prior to twinning.

Discovery of twin-specific putative de novo SNVs (group II)

This class of de novo SNVs was discovered using the above-mentioned strategy including the comparison of the genotypes of the mother, father and the twin pair for all SNVs simultaneously. SNVs observed in one of the twins and not observed in the other twin as well as both parents were considered as candidate de novo SNVs. Same set of filters described above was applied to the resulting candidate de novo SNVs to generate a list of high-confidence putative twin-specific de novo SNVs. We

considered the timing of mutagenesis for the twin-specific de novo SNVs as early postzygotic after the twinning event.

Evaluation of the putative de novo SNVs based on visual examination

We used Integrative Genomics Viewer (IGV)⁹ to examine the Sequence Alignment/Map (BAM) files to further evaluate the putative de novo SNVs. We classified the variants in our initial call-set as (a) inherited, (b) no variant, (c) located on a low-quality read and (d) de novo. In the presence of more than three reads supporting the variant in one or both parents, the variant was classified as inherited and filtered. When the SNV or indel density was high around the region flanking 40 bp the putative de novo SNVs, we classified the sequencing reads supporting the variants as 'low-quality reads' and filtered them. Additionally, the putative de novo SNVs that are not supported by at least three reads and located on the 'low-quality reads' were classified as 'no variant', and thus filtered. We included the remaining putative de novo SNVs in our final call-set and selected them for validation by Sanger-based capillary sequencing (see online supplementary table S1).

Validation of de novo SNVs with Sanger sequencing

We selected all 120 high-confidence putative de novo SNVs for validation. Of these, 37 could not be tested due to unavailability of appropriate primers or PCR products of sufficient quality for capillary sequencing. We designed primer pairs using Primer3¹⁰ and PerlPrimer¹¹ (see online supplementary table S2). PCR was performed using 25 ng genomic DNA, One Taq DNA Polymerase (New England Biolabs), 25 mM dNTPs (Fermentas), 5X One Taq standard reaction buffer and 10 pmol of each forward and reverse primers. Thermocycling conditions were as recommended by the manufacturer (94°C for 30 s; 30–35 cycles of 94°C for 30 s, 58–60°C for 40 s and 68°C for 40 s; and 68°C for 5 min). PCR products were analysed through 0.8% agarose gel electrophoresis. Following the purification of PCR products, Sanger sequencing was performed using forward primer, reverse primer or both. CLC Main Workbench software package (CLCBio Inc) was used to analyse Sanger sequencing data.

Calculation of the reference allele frequency

We used reference allele frequency (RAF) to evaluate the putative de novo SNVs. RAF is the total number of reads supporting the reference allele as a fraction of total number of reads supporting both reference and alternative allele.¹² For the putative de novo SNVs that are present in a heterozygous state, RAF threshold was determined as 0.5–0.7 for the twins and 0.95–1.0 for the parents.

RESULTS

Mapping of the whole-genome sequencing data and discovery of the sequence variants

We sequenced the whole genomes of the monozygotic twin pair and their parents using the Illumina HiSeq2000 platform and generated a total of 538 Gb genomic sequence with a mean coverage depth of more than 30-fold per individual (see online supplementary table S3). Using BWA,⁴ we mapped the paired-end sequencing reads to the human reference genome (GRCh37) and discovered genomic variants using the GATK.⁶

Herein, we discovered genomic variants using the multisample calling options of UnifiedGenotyper (a total of 170-fold coverage). We identified approximately 3.4 million SNVs and 750 000 small insertions and deletions (indels) per genome (see online supplementary table S4).

Identification of the putative de novo SNVs

Examination of the genotypes of the twin pair and their parents synchronously for each of the SNVs identified an initial set of putative de novo SNVs. To distinguish true de novo SNVs from sequencing and variant calling artefacts, we progressively applied the following set of filters: (a) novelty vis-à-vis dbSNP137, (b) having a coverage depth of at least three for the site encompassing the variant in each individual, (c) presence outside of the segmental duplications and/or repetitive DNA and (d) genotyping error (see 'Methods'). With this filtering strategy, we identified a total of 290 high-confidence putative de novo SNVs. Of these, 159 were shared by the twin pair (group I); and 83 were specific for twin I and 48 for twin II (group II) (see online supplementary tables S5–S7).

Selection of the high-confidence putative de novo SNVs for validation

Subsequent to the identification of the initial call-set, we visually examined the BAM files for all of the 290 high-confidence putative de novo SNVs using IGV.⁹ We also considered the allelic depth for each SNV in each individual. This resulted in the filtering of 86/159 shared, 54/83 twin I-specific and 30/48 twin II-specific de novo SNVs (see 'Methods' and online supplementary figure S1 and tables S8–S11).

We selected the remaining de novo SNV candidates for Sanger sequencing (73 shared, 29 twin I-specific and 18 twin II-specific) and also classified them by genomic context, which revealed a biased distribution towards non-coding regions of the genome (figure 1C).¹³ This is expected because of two reasons: (a) non-coding regions constitute most of the genome, and (b) they are subjected to less stringent evolutionary pressure.

We calculated the average RAF and coverage depth values for all of the putative de novo SNVs (a) shared by the twin pair, (b) specific to twin I and (c) specific to twin II (see 'Methods' and online supplementary table S12). As expected, average RAF is between 0.95 and 0.97 for the parents. For the twin pair, average RAF for shared de novo SNVs is 0.59 for twin I and 0.55 for twin II, supporting the heterozygous presence of the variants. And for the twin-specific de novo SNV candidates, RAF values are consistent with mosaic presence (0.68–0.69) in the twin that carries the variant.

Experimental validation of de novo SNV candidates

Of the 73 shared de novo SNV candidates, 49 could be tested using capillary sequencing, and 23 confirmed to be de novo SNV in heterozygous state (see online supplementary table S8). We considered the timing of mutagenesis as parental or pre-twinning zygote for these 23 shared de novo SNVs.

Of the 29 twin I-specific and 18 twin II-specific high-confidence putative de novo SNVs (group II), 19 and 15, respectively, could be analysed by capillary sequencing; and 8/19 for twin I and 1/15 for twin II were confirmed (see online supplementary tables S9 and S10).

In order to control the validity of the filtering process that targeted the putative de novo SNVs classified as genotyping error, we selected all 34 for capillary sequencing. Of these, 19 could be tested, and none confirmed as a de novo SNV (see online supplementary table S13).

We confirmed the exclusive representation of twin-specific de novo SNVs by simultaneous Sanger sequencing of the whole family. The mutant allele for these twin-specific de novo SNVs was clearly visible along with the wild-type allele in the capillary chromatograms (figure 1D).

We calculated one more time the average RAF and coverage depth values for the confirmed de novo SNVs as described above (see online supplementary table S14). Average RAF is 0.96–1.00 for the parents. For the twin pair, it is 0.53 (twin I) and 0.46 (twin II) for the shared; and 0.65 (twin I) and 0.78 (twin II) for the twin-specific de novo SNVs. These results are consistent with the Sanger sequencing chromatograms, which indicate the mosaic presence of twin-specific de novo SNVs.

Mosaic representation of a mutation in a parent could be interpreted as a de novo mutation in the twins.^{14–16} We therefore extended our de novo mutation validation in blood (mesoderm) to the analysis of DNA samples isolated from the buccal (ectoderm) and the urine (endoderm) cells of the parents. We did not observe the de novo SNVs in either cell type, which suggests that parental mosaicism does not contribute to the shared de novo SNVs in the twins (see online supplementary figure S2).

Calculation of the mutation rate

Mutation rate was calculated based on the number of validated de novo SNVs and callable bases. Bases with a minimum read depth of 3 and do not intersect segmental duplications were considered as callable. We calculated 2 703 445 479 bases for twin I and 2 702 251 568 bases for twin II. We used the formula below to calculate the mutation rates³:

$$\mu = (N \times (1 + \text{FNR}) / L$$

At this point, N represents the number of de novo SNVs validated by capillary sequencing, FNR is the false-negative rate (see online supplementary table S15) and L is the total number of callable bases.

We calculated a rate of 0.97×10^{-8} bp per generation for de novo SNVs shared by the twin pair. For twin-specific de novo SNVs, we calculated the rate of 0.34×10^{-8} bp per generation for twin I and 0.04×10^{-8} bp per generation for twin II, respectively.

DISCUSSION

The rate of de novo SNVs in healthy individuals has been investigated through whole-genome sequencing of mother–father–offspring trios and yielded rates ranging between 0.82 – 1.70×10^{-8} per base pair.¹⁷ Among these overall rates, the early postzygotic de novo SNVs have not been distinguished from those which have parental origin. Here we present overall de novo SNV rate of 1.01 and 1.31×10^{-8} in a healthy monozygotic twin pair consistent with previous direct estimations. We also observe that parental mosaicism does not contribute to the shared de novo SNVs in the twins. Furthermore, we report the contribution of early postzygotic mutations to the average de novo SNV burden with a rate of 0.04 – 0.34×10^{-8} . Our deep whole-genome sequencing-based approach combined with effective filtering and validation enabled the direct measurement of early postzygotic versus prezygotic de novo SNV rate.

The very first studies concerned with the estimation of somatic mutation rate were restricted to coding or non-coding regulatory regions and mainly focused on disease phenotypes.^{18–19} The average somatic mutation rate has been reported to be 7.7×10^{-10} .²⁰ A more recent study that surveyed the whole genome estimated the somatic mutation rate for common SNPs in healthy monozygotic twins to be 1.2×10^{-7} per base pair.¹² Conrad and his colleagues distinguished the germline and non-germline (somatic and cell line derived) mutations through whole-genome sequencing of a CEU trio and provided a rate of 2.52×10^{-7} per nucleotide for non-germline mutations. This estimate, which is greater than the estimate presented in this study, is probably influenced by mutagenicity and age of the cell culture.²¹

De novo mutations that are not shared by monozygotic twins could be cell line derived.²² We excluded this possibility by directly sequencing whole-blood-derived DNA. We, however, point out that the rate of postzygotic mutations is likely to be underestimated. This is because mutations that occur later in embryo development will be observed as very low-frequency mosaic SNVs that will be regarded as sequencing errors by mapping and variant calling algorithms.²³

High levels of allelic, locus and phenotypic heterogeneity have important implications for gene discovery in complex human diseases.²⁴ Genomic microarrays and next-generation sequencing technologies are now enabling researchers to dissect the molecular basis of complex phenotypes that arise from de novo mutations.²⁵ This has been demonstrated in common neurodevelopmental diseases such as schizophrenia²⁶ and autistic-spectrum disorders,²⁷ pointing to a monogenic basis of disease with the mutation representing a single event of large effect. However, even after the spectacular success of these modern genetic studies, a majority of cases remain unsolved. An added level of complexity, consistent with this 'de novo model', could be the timing of mutational events whereby early postzygotic de novo mutations could be critically important. This is supported by the observations that somatic mosaicism has been well documented in Mendelian phenotypes,^{14 28 29} including monozygotic twins discordant for a given disorder,³⁰ and by demonstration of extensive genetic variation in human tissues³¹ including brain.³² It will be crucial to expand genome sequencing studies to the next level of tissue or even cell type-specific interrogation to better delineate causal mutations especially in sporadic forms of Mendelian disorders.

Acknowledgements We are grateful to the family that participated in this study. We are also grateful to Pinar Kavak (TÜBİTAK-BİLGEM-UEKAE) for her help in bioinformatics analysis. This project was funded by the Turkish State Planning Organization (to TÜBİTAK-MAM Advanced Genomics and Bioinformatics Center), the Turkish Academy of Sciences (to TÖ), a TÜBİTAK grant (112E135 to CA) and an EMBO grant (IG-2521 to CA). CA also acknowledges support from The Science Academy, Turkey, under the BAGEP program.

Contributors TÖ conceived the study. GMD, TÖ and CA planned the experiments and wrote the manuscript. MSS and BY generated whole-genome sequencing data. GMD and CA analysed the data. GMD, BE and CA performed bioinformatics analyses. GMD and OEO completed experimental validation.

Funding Turkish State Planning Organization, The Scientific and Technological Research Council of Turkey (TÜBİTAK) and Turkish Academy of Sciences (TÜBA).

Competing interests None.

Patient consent Obtained.

Ethics approval Bilkent, Hacettepe, Başkent, and Çukurova Universities Institutional Review Boards.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

- Haldane JBS. The rate of spontaneous mutation of a human gene. *J Genet* 1935;31:317–26.
- Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* 2012;488:471–5.
- Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, Vives L, O'Roak BJ, Sudmant PH, Shendure J, Abney M, Ober C, Eichler EE. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* 2012;44:1277–80.
- Li H, Durbin L. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–60.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–9.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AV, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–8.
- Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet* 2012;15:3.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–6.
- Koreasaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 2007;23:1289–91.
- Marshall OJ. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics* 2004;20:2471–7.
- Li R, Montpetit A, Rousseau M, Wu SY, Greenwood CM, Spector TD, Pollak M, Polychronakos C, Richards JB. Somatic point mutations occurring early in development: a monozygotic twin study. *J Med Genet* 2014;51:28–34.
- Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w¹¹¹⁸; iso-2; iso-3. *Fly (Austin)* 2012;6:80–92.
- Darras BT, Francke U. A partial deletion of the muscular dystrophy gene transmitted twice by an unaffected male. *Nature* 1987;329:556–8.
- Biesecker LG, Spinner NB. A genomic view of mosaicism in human disease. *Nat Rev Genet* 2013;14:307–20.
- Ozcelik T. Uncovering the complex mysteries of mosaicism. *Nature* 2002;417:588.
- Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. *Trends Genet* 2013;29:575–84.
- Araten DJ, Golde DW, Zhang RH, Thaler HT, Gargiulo L, Notaro R, Luzzatto L. A quantitative measurement of the human somatic mutation rate. *Cancer Res* 2005;65:8111–17.
- Iwama T. Somatic mutation rate of the APC gene. *Jpn J Clin Oncol* 2001;31:185–7.
- Lynch M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* 2010;107:961–8.
- Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurler ME, Awadalla P; 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. *Nat Genet* 2011;43:712–14.
- Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* 2012;91:1033–40.
- Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebukova I, Miller NA, Zhang L, Farmer AD, Bell CJ, Kim RW, May GD, Woodward JE, Caillier SJ, McElroy JP, Gomez R, Pando MJ, Clendenen LE, Ganusova EE, Schilkey FD, Ramaraj T, Khan OA, Huntley JJ, Luo S, Kwok PY, Wu TD, Schroth GP, Oksenberg JR, Hauser SL, Kingsmore SF. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* 2010;464:1351–6.
- McClellan J, King MC. Genetic heterogeneity in human disease. *Cell* 2010;141:210–17.
- Veltman JA, Brunner HG. *De novo* mutations in human genetic disease. *Nat Rev Genet* 2012;13:565–75.
- Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S, Rippey C, Shahin H, Consortium on the Genetics of Schizophrenia (COGS); PAARTNERS Study Group; Nimgaonkar VL, Go RC, Savage RM, Swerdlow NR, Gur RE, Braff DL, King MC, McClellan JM. Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 2013;154:518–29.
- Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, Kendall J, Grabowska E, Ma B, Marks S, Rodgers L, Stepansky A, Troge J, Andrews P, Bekrisky M, Pradhan K, Ghiban E, Kramer M, Parla J, Demeter R, Fulton LL, Fulton RS, Magrini VJ, Ye K, Darnell JC, Darnell RB, Mardis ER, Wilson RK, Schatz MC, McCombie WR, Wigler M. *De novo* gene disruptions in children on the autistic spectrum. *Neuron* 2012;74:285–99.
- Stern C. Genetic mosaics in animals and man. In: Stern C, eds. *Genetic mosaics and other essays*. Cambridge, Massachusetts: Harvard University Press, 1968:27–129.
- Lupski JR. Genome mosaicism—one human, multiple genomes. *Science* 2013;341:358–9.
- Vadlamudi L, Dibbens LM, Lawrence KM, Iona X, McMahon JM, Murrell W, Mackay-Sim A, Scheffer IE, Berkovic SF. Timing of *de novo* mutagenesis—a twin study of sodium-channel mutations. *N Engl J Med* 2010;363:1335–40.
- O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP. Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci USA* 2012;109:18018–23.
- Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzo P, Smith S, Fell M, Talbot RT, Gustinich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 2011;479:534–7.