

Electronic Thesis and Dissertation Repository

9-21-2015 12:00 AM

A Unified Framework for the Prioritization of Variants of Uncertain Significance in Hereditary Breast and Ovarian Cancer Patients

Natasha G. Caminsky
Western University

Supervisor
Dr. Peter K. Rogan
The University of Western Ontario

Graduate Program in Biochemistry
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science
© Natasha G. Caminsky 2015

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Biochemistry Commons](#), [Bioinformatics Commons](#), [Medical Genetics Commons](#), [Molecular Biology Commons](#), and the [Molecular Genetics Commons](#)

Recommended Citation

Caminsky, Natasha G., "A Unified Framework for the Prioritization of Variants of Uncertain Significance in Hereditary Breast and Ovarian Cancer Patients" (2015). *Electronic Thesis and Dissertation Repository*. 3318.

<https://ir.lib.uwo.ca/etd/3318>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

A UNIFIED FRAMEWORK FOR THE PRIORITIZATION OF VARIANTS OF
UNCERTAIN SIGNIFICANCE IN HEREDITARY BREAST AND OVARIAN
CANCER PATIENTS

(Thesis format: Integrated Article)

by

Natasha Grace Caminsky

Graduate Program in Biochemistry

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

© Natasha Grace Caminsky, 2015

Abstract

A significant proportion of hereditary breast and ovarian cancer (HBOC) patients receive uninformative genetic testing results, an issue exacerbated by the overwhelming quantity of variants of uncertain significance identified. This thesis describes a framework where, aside from protein coding changes, information theory (IT)-based sequence analysis identifies and prioritizes pathogenic variants occurring within sequence elements predicted to be recognized by proteins involved in mRNA splicing, transcription, and untranslated region binding and structure. To support the utilization of IT analysis, we established IT-based variant interpretation accuracy by performing a comprehensive review of mutations altering mRNA splicing in rare and common diseases.

Custom probes targeting 20 complete HBOC genes for sequencing in 379 *BRCA*-uninformative patients identified 47,501 unique variants and we prioritized 429 variants in both *BRCA* and non-*BRCA* genes. Our approach focuses attention on a limited set of variants from a spectrum of functional mutation types for downstream functional and co-segregation analysis.

Keywords

3' untranslated regions, binding sites, breast neoplasms, computational biology, genes, genetic testing, information theory, next-generation sequencing, non-coding, ovarian neoplasms, prioritization, RNA stability, RNA-binding protein, splicing, transcription factor binding, tumor suppressor, variants of uncertain significance

Co-Authorship Statement

Chapter 2 – Dr. Peter Rogan conceived of the study. Dr. Peter Rogan and Natasha Caminsky coordinated and structured the project. Natasha Caminsky led the literature review and data collection with the assistance of Eliseos Mucaki. All authors participated in the manuscript preparation and approved the final manuscript submission for publication.

Chapter 3 – Dr. Peter Rogan designed, coordinated, and supervised the study. Eliseos Mucaki and Edwin Dovigi performed probe design and synthesis. Natasha Caminsky and Eliseos Mucaki performed sample preparation and sequencing. Eliseos Mucaki wrote software and performed bioinformatic analysis. Natasha Caminsky, Eliseos Mucaki, and Ami Perri conducted variant analysis and prioritization. Ruipeng Lu generated the TFBS information models and Eliseos Mucaki generated the RBBS, SRF, and splicing information models. Ami Perri confirmed prioritized variants by Sanger sequencing. Dr. Matthew Halvorsen and Dr. Alain Laederach conducted the SHAPE analysis. Natasha Caminsky, Eliseos Mucaki, Ami Perri, and Dr. Peter Rogan wrote the manuscript, which has been approved by all authors.

Chapter 4 – Natasha Caminsky recruited patients, with the assistance of Abby Watts-Dickens and the guidance of Karen Panabaker. Natasha Caminsky, Eliseos Mucaki, and Yulia Maistrovski designed and synthesized the DNA probes. Natasha Caminsky, Eliseos Mucaki, and Ami Perri participated in sequencing and variant analysis. Natasha Caminsky, Eliseos Mucaki, Ami Perri, and Peter Rogan wrote the manuscript, which has been approved by all authors.

Acknowledgments

I would like to thank Dr. Peter Rogan for his guidance and direction, and for the opportunity to work on this project. Thank you to Dr. Ainsworth and Alan Stuart at the Molecular Diagnostics Lab of the LHSC for providing patient samples, the opportunity to work with the BioMek[®] Workstation, and to the staff for their assistance. I greatly appreciate the support and mentorship of Karen Panabaker, who trained me on patient recruitment and was always available if I needed to consult with her. I also owe a great amount of gratitude to Abby Watts-Dickens recruiting the first 150 patients. Thank you also to Dr. Laederach and his lab for conducting the SHAPE analyses.

Thank the Edgell and Khan labs as well as the London Regional Genomics Centre at Robarts for providing access to instruments needed for sample preparation. I am especially grateful for the support, patience, advice, and instruction of Eliseos (John) Mucaki. Thank you for previously synthesizing the first capture array and leading the sequencing runs. To Ami Perri, Dr. Knoll, Coby, Stephanie, Wahab, Ben, Li, and the remainder of the Rogan lab, thank you for many good laughs, and helpful suggestions. Finally, I would not be where I am today without the love and support of my parent, family, rowing team, and close friends, namely Maria Schrempf, Claudia Blandford, Lorne Schweitzer, Ashley Veseley, Alex McCarton, Sara Matovic, and David Wakulich.

This thesis was supported by both the London CIHR Strategic Training Program in CaRTT Summer Studentship and Award, as well as the Pamela Greenaway-Kohlmeier Translational Breast Cancer Research Unit (TBCRU) Award. This project would not have been possible without the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca). Additional funding was provided by Canadian Breast Cancer Foundation (PKR), Natural Sciences and Engineering Research Council of Canada (NSERC – PKR), Canada Research Chairs (PKR), and Canada Foundation for Innovation (PKR).

London
N.G.C.
August 2015

Roles and Responsibilities

REB approval of patient recruitment and our experimental protocol was applied for by Dr. Peter Rogan, Dr. Peter Ainsworth, Dr. Joan Knoll, Dr. Muriel Brackstone, and Karen Panabaker. Abby Watts-Dickens was a part-time paid student (January-June 2014) who led the patient recruitment process and Natasha Caminsky assisted with this task on a weekly basis for the months of June 2014-May 2015. Karen Panabaker oversaw this process. Dr. Peter Ainsworth and Alan Stuart provided the anonymous patient samples (UWO1-7) and provided assistance with the BioMek FXP[®] Automation Workstation when needed.

The design of probes targeting *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *PALB2*, and *TP53* was done by Eliseos Mucaki and were synthesized by Eliseos Mucaki and Eddie Dovigi. The initial protocol for single sample handling was developed by Eliseos Mucaki and Eddie Dovigi. Automation using the BioMek[®] FXP Automation Workstation was designed by Eliseos Mucaki and set up and optimized by Eliseos Mucaki and Natasha Caminsky. The protocol for multiplex library and sample preparation was developed by Eliseos Mucaki and Natasha Caminsky. The protocol for automated capture pull-down was developed by Eliseos Mucaki with the assistance of Natasha Caminsky and Ami Perri.

The design and synthesis of probes targeting *ATP8B1*, *EPCAM*, *MLH1*, *MRE11A*, *MSH2*, *MSH6*, *MUTYH*, *NBN*, *PMS2*, *PTEN*, *STK11*, *XRCC2* was performed by Natasha Caminsky, Eliseos Mucaki, and Yulia Maistrovski.

Sample preparation and sequencing of samples in runs 1-3 (N=20) were performed by Eliseos Mucaki and Eddie Dovigi in 2012, however due to low coverage, 7 of these samples were re-sequenced in later runs. Eliseos Mucaki participated in the preparation and sequencing of runs 4-16, Yulia Maistrovski in runs 8-11, Ami Perri in runs 13-15, and Natasha Caminsky in runs, 4-9, and 12-13.

The script for running all variants through the various bioinformatic programs that were used was written by Eliseos Mucaki. The approach for missense variant analysis was conceived by Natasha Caminsky. The software used for high-throughput *in silico* splicing analysis (the Shannon Pipeline) was developed by Ben Shirley. The ASSEDA server is proprietary software of Dr. Rogan's. The transcription factor models were made by Ruipeng Lu and the program used for transcription factor binding site analysis (Mutation Analyzer) was developed by Coby Viner. RNA-binding protein models were built by Eliseos Mucaki. The framework for variant prioritization was developed by Dr. Peter Rogan, Eliseos Mucaki, Ami Perri, and Natasha Caminsky.

Table of Contents

Abstract.....	ii
Keywords.....	ii
Co-Authorship Statement.....	iii
Acknowledgments.....	iv
Roles and Responsibilities.....	v
Table of Contents.....	vii
List of Tables.....	xii
List of Figures.....	xiv
List of Appendices.....	xvi
List of Abbreviations.....	xvii
Chapter 1.....	1
1 Introduction.....	1
1.1 Hereditary Breast and Ovarian Cancer.....	1
1.1.1 Prevalence and Risk.....	1
1.1.2 Linkage Studies and Mutation Data.....	2
1.2 Molecular Diagnostics and Variant Classification.....	3
1.2.1 Current Classification Systems.....	3
1.3 VUS and Disease.....	5
1.3.1 Consequences of Reporting Unknown Variants.....	5
1.3.2 Implications of VUS for HBOC.....	10
1.3.3 Scope of VUS.....	11
1.3.4 Prioritization of VUS.....	13
1.4 Thesis Objectives.....	14

1.5	References.....	15
Chapter 2..... 22		
2	Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis	22
2.1	Introduction.....	22
2.2	Information theory and splice site analysis.....	26
2.3	Software resources	30
2.3.1	Delila package/system	30
2.3.2	ASSA/ASSEDA.....	30
2.3.3	Shannon pipeline and Veridical	33
2.4	Natural sites	35
2.4.1	Minimum splice site information content and exceptions	41
2.4.2	Branch-point mutations.....	45
2.5	Activation of cryptic splicing.....	46
2.6	Combinatorial effects.....	51
2.7	Validation of results.....	58
2.7.1	Validation methods	58
2.7.2	Regulatory sequence variants	60
2.8	Accuracy of IT-based prediction	61
2.8.1	Predicted mutations discordant with validation results	62
2.9	Misinterpretation of variant effects.....	63
2.10	Interpretation of published variants in studies that use information analysis ...	65
2.10.1	Genotype-phenotype association	65
2.11	Polymorphisms and splicing.....	66
2.12	Inference of variant pathogenicity by IT analysis.....	67
2.13	Comparisons to other software programs	68

2.14	Other applications of information theory- based splice site analysis.....	71
2.15	Guidelines for information theory-based splicing mutation analyses.....	72
2.15.1	Report gene isoform and genomic coordinates.....	72
2.15.2	Report R_i values	73
2.15.3	Consider impact of missense and synonymous mutations on mRNA splicing.....	74
2.15.4	Experimentally validate variants.....	74
2.15.5	Report the sequence window used in the analysis	74
2.15.6	Designate genic rearrangements (insertions, deletions, duplications) with genomic coordinates	75
2.16	References	76
Chapter 3		105
3	A unifying framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer.....	105
3.1	Background.....	105
3.2	Methods.....	108
3.2.1	Design of Tiled Capture Array for HBOC Gene Panel	108
3.2.2	HBOC Samples for Oligo Capture and High-Throughput Sequencing..	109
3.2.3	Sequence Alignment and Variant Calling.....	112
3.2.4	IT-Based Variant Analysis.....	112
3.2.5	Exonic Protein-Altering Variant Analysis	117
3.2.6	Variant Classification.....	118
3.2.7	Positive control	118
3.2.8	Variant Validation.....	120
3.3	Results.....	120
3.3.1	Capture, Sequencing, and Alignment	120

3.3.2	IT-Based Variant Identification and Prioritization	121
3.3.3	Exonic Variants altering protein sequence	135
3.3.4	Variant Classification.....	138
3.3.5	Variant Verification	138
3.4	Discussion.....	142
3.4.1	Non-coding variants.....	143
3.4.2	Coding sequence changes	145
3.5	Conclusions.....	148
3.6	References.....	149
Chapter 4	165
4	Prioritizing variants in complete hereditary breast and ovarian cancer (HBOC) genes in patients lacking known <i>BRCA</i> mutations.....	165
4.1	Introduction.....	165
4.2	Methods.....	168
4.2.1	Ethics and Patient Recruitment.....	168
4.2.2	Probe Design, Sample Preparation, and Sequencing	169
4.2.3	Information Models	171
4.2.4	Variant Analysis.....	172
4.2.5	Likelihood Ratios (LRs)	173
4.3	Results.....	173
4.3.1	Variant Analysis.....	173
4.3.2	Exonic Protein-Altering Variants	184
4.3.3	Variant Prioritization	188
4.3.4	Negative Control.....	188
4.3.5	Pedigree Analysis.....	193
4.3.6	Likelihood Ratio Analyses.....	193

4.4 Discussion	196
4.5 References	201
Chapter 5	207
5 General Discussion	207
5.1 Patient Recruitment and Participation	208
5.2 Variant Prioritization in the Context of ACMG Guidelines	209
5.3 Improving Variant Prioritization and Alternative Approaches	211
5.4 Future Directions	213
5.4.1 Controls	213
5.4.2 Contribution to HBOC Literature	214
5.4.3 Patient Counseling	214
5.5 References	216
Appendices	217
Curriculum Vitae	290

List of Tables

Table 1.1. Proposed Classification System for Sequence Variants Identified by Genetic Testing.....	6
Table 1.2 ACMG Criteria for Classifying Pathogenic Variants.....	7
Table 1.3. ACGM Criteria For Classifying Benign Variants.....	8
Table 1.4. ACMG Rules for Combining Criteria to Classify Sequence Variants.....	9
Table 2.1. Splicing regulatory protein binding sites ASSEDA scans for and their associated effect on splicing.	54
Table 2.2. Concordance of splice-prediction programs to information theory-based tools for natural and cryptic sites.....	69
Table 3.1. Risk Categories for Individuals Eligible for Screening for a Genetic Susceptibility to Breast or Ovarian Cancers as determined by the Ontario Ministry of Health and Long Term-Care Referral Criteria for Genetic Counseling.....	111
Table 3.2. Prioritized Variants in the Positive Control.....	119
Table 3.3. Variants Prioritized by IT Analysis.....	123
Table 3.4. Variants Predicted by SNPfold to Affect UTR Structure.....	131
Table 3.5. Variants Resulting in Premature Protein Truncation.....	136
Table 3.6. Summary of Prioritized Variants by Gene.....	141
Table 4.1. Prioritized Variants Predicted by IT to Affect Natural and Cryptic Splicing.....	174
Table 4.2. Variants Predicted by SNPfold to Significantly Affect UTR Structure.....	182
Table 4.3. Variants Resulting in Premature Protein Truncation.....	185
Table 4.4. Comparing Counts of Prioritized Variants.....	189

Table 4.5. Distribution of Recruited Patients Among Eligibility Groups.....	191
Table 4.6. LR Values for Patients with Prioritized Truncating, Splicing, and Selected Missense Variants	194
Table 5.1. Evidence Framework from the ACGM.	210

List of Figures

Figure 2.1. Distribution of deleterious natural site variants relative to information content..	28
Figure 2.2. Sample retrieval of average change in information content (ΔR_i) with splicing mutation calculator (SMC) for published mutations	39
Figure 2.3. Frequency of Abolishing Variants in Relation to Initial Strength.....	44
Figure 2.4. Ribl used for the prediction of a variant's effect on branch-point sites	47
Figure 2.5. Outcomes of cryptic splicing mutations	48
Figure 2.6. Distribution of activated cryptic sites.....	52
Figure 3.1. Capture Probe Coverage over Sequenced Genes	110
Figure 3.2. Framework for the Identification of Potentially Pathogenic Variants.....	113
Figure 3.3. <i>BRCA1</i> Deletion Inaccurately Aligned by CASAVA	122
Figure 3.4. Predicted Isoforms and Relative Abundances as a Consequence of <i>ATM</i> splice variant c.3747-1G>A	127
Figure 3.5. Predicted Isoforms and Relative Abundances as a Consequence of <i>CHEK2</i> splice variant c.320-5T>A	128
Figure 3.6. Predicted Alteration in UTR Structure Using mfold for Variants Flagged by SNPfold.....	133
Figure 3.7. Ladder Plot Representing Variant Identification and Prioritization.....	139
Figure 4.1. Significant Genome Stabilizing Pathways, Risk, and Relevant Literature for 20 HBOC Genes	167
Figure 4.2. Distribution of Patients by Eligibility Group and Age.....	170

Figure 4.3. Predicted Isoforms and Relative Abundance as a Consequence of <i>ATM</i> natural splice variant c.6198+1G>A	177
Figure 4.4. Predicted Isoforms and Relative Abundance as a Consequence of <i>CDH1</i> cryptic splice variant c.1223C>G	179
Figure 4.5. Predicted RNA Structure Change due to Variants Flagged by SNPfold using mfold	183
Figure 4.6. Distribution of Unique Prioritized Variants by Category.....	192
Figure 4.7. Computed Likelihood Ratios for Patients with Variants of Disparate Priority..	199

List of Appendices

Appendix A: Copyright Permissions for Chapter 1 (Table 1.4)	217
Appendix B: Copyright Permission for Chapter 1 (Tables 1.5, 1.6, 1.7)	218
Appendix C: Copyright Permission for Chapter 2.....	218
Appendix D: Supplementary Bibliography	219
Appendix E: Supplementary Methods (Chapter 3).....	248
Appendix F: Ethics Approval and Amendments for Patient Recruitment.....	258
Appendix G: Letter of Invitation for 7 Genes.....	259
Appendix H: Response Card.....	262
Appendix I: Letter of Invitation for 23 Genes	264
Appendix J: Final Letter to Patients.....	268
Appendix K: Supplementary Information (Chapter 4)	270

List of Abbreviations

2°	Secondary structure
ACMG	American College of Medical Genetics and Genomics
<i>APC</i>	Adenomatous polyposis coli
ASD	ASD-Intron analysis
ASSA	Automated Splice Site Analysis
ASSEDA	Automated Splice Site and Exon Definition Analysis
<i>ATM</i>	Ataxia Telangiectasia Mutated
<i>BARD1</i>	BRCA1 Associated RING Domain 1
BC	Breast cancer
BDGP	Splice Site Prediction by Neural Network, NNSplice
<i>BEST1</i>	Bestrophin 1
BIC	Breast Cancer Information Core Database
BPS	Branch-point site
<i>BRCA1</i>	Breast Cancer 1, Early Onset
<i>BRCA2</i>	Breast Cancer 2, Early Onset
CASAVA	Consensus Assessment of Sequencing and Variation
<i>CDH1</i>	Cadherin 1, Type 1, E-Cadherin (Epithelial)
<i>CHEK2</i>	Checkpoint Kinase 2
CISBP-RNA	Catalog of Inferred Sequence Binding Preferences of RNA binding proteins
CRC	Colorectal cancer
DHPLC	Denaturing high performance liquid chromatography
DM ²	Domain Mapping of Disease Mutations
ENIGMA	Evidence-based Network for the Interpretation of Germline Mutant Alleles
<i>EPCAM</i>	Epithelial Cell Adhesion Molecule
ESE	Exonic splice enhancer
ESS	Exonic splice silencer
ExPASy	Expert Protein Analysis System
GAIix	Genome Analyzer Iix
GATK	Genome Analysis Toolkit

GeneS	GeneScan
GenS	GenScan
GM	GeneMark
GS	GeneSplicer
HBOC	Hereditary breast and ovarian cancer
HGMD	Human Gene Mutation Database
HGVS	Human Genome Variation Society
HNPCC	Hereditary non-polyposis colorectal cancer
HSF	Human Splice Finder
IARC	International Association for Research on Cancer
IGV	Integrative Genomics Viewer
Indel	Insertion/deletion
InSiGHT	International Society for Gastrointestinal Hereditary Tumours
ISE	Intronic splice enhancer
ISS	Intronic splice silencer
IT	Information theory
LHSC	London Health Sciences Centre
LOVD	Leiden Open Variant Database
LR	Likelihood ratio
MES	MaxEntScan
MGL	Molecular Genetics Laboratory
<i>MLH1</i>	MutL Homolog 1
MLPA	Multiplex ligation-dependent amplification
MMR	Mismatch repair
<i>MRE11A</i>	Meiotic Recombination 11 Homolog A
<i>MSH2</i>	MutS Homolog 2
<i>MSH6</i>	MutS Homolog 6
<i>MUTYH</i>	MutY Homolog
<i>NBN</i>	Nibrin
NG2	NetGene2
NGS	Next-generation sequencing

NMD	Nonsense-mediated decay
OC	Ovarian cancer
<i>PALB2</i>	Partner And Localizer Of BRCA2
<i>PDHA1</i>	Pyruvate dehydrogenase alpha 1
PESX	Putative Exonic Splicing Enhancers/Silencers
Phred score	Quality score as a measure of the quality of the identification of a nucleobase in automated DNA sequencing
<i>PMS2</i>	PMS2 Postmeiotic Segregation Increased 2 (S. Cerevisiae) Homolog
PTB	Polypyrimidine tract binding protein
PTC	Premature truncation of the transcript
<i>PTEN</i>	Phosphatase And Tensin Homolog
PTT	Protein Truncation Test
PWM	Position weight matrix
<i>RAD51B</i>	RAD51 Paralog B
RBBS	RNA-binding protein binding site
RBP	RNA-binding protein
RBPDB	RNA-binding protein database
R_i	Individual information
Rib1	R_i [b,1]
RR	Relative risk
$R_{sequence}$	Mean information content
S-S	Shapiro-Senepathy
S.D.	Standard deviation
SHAPE	Selective 2'-Hydroxyl Acylation analyzed by Primer Extension
SMC	Splicing Mutation Calculator
<i>SMN</i>	Survival motor neuron
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SP	Splice Predictor
SR	Serine-arginine
SRF	Splicing regulatory factor

SRFBS	Splicing regulatory factor binding site
SS	Splice site
SSF	Splice Site Finder
SSqE	Splicing Sequences Finder
<i>STK11</i>	Serine/Threonine Kinase 11
SV	SpliceView
TF	Transcription factor
TFBS	Transcription factor binding site
<i>TP53</i>	Tumor Protein P53
UTR	Untranslated region
VCF	Variant call format
VUS	Variants of uncertain significance
<i>XRCC2</i>	X-Ray Repair Cross-Complementing Protein 2
ΔR_i	Change in information

Chapter 1

1 Introduction

The prevalence and aggressive nature of familial breast (BC) and ovarian cancer (OC), combined with a large proportion of patients receiving uninformative test results, are compelling reasons for devising a means not only for identifying, but analyzing and prioritizing variants of uncertain significance (VUS). We devised a unified framework for variant analysis and propose this bioinformatic approach as a cost- and time-effective means of prioritizing variants for further analysis. This chapter outlines the current state of the hereditary breast and ovarian cancer (HBOC) literature, demonstrates the challenges and implications regarding VUS, and along with the following chapter, gives grounds for our choice of bioinformatic approach.

1.1 Hereditary Breast and Ovarian Cancer

Currently, the lifetime risk for a woman to develop BC is 1/8 and 7/50 in the case of OC¹. BC is the leading cancer affecting women, representing 25% of all cases, whereas OC ranks 11th, affecting 4%². Family history, where multiple individuals on the same side of a family present with the disease, is one of the largest risk factors for BC³ and strongly indicates a genetic association.

1.1.1 Prevalence and Risk

Approximately 5-10% of all BC cases are hereditary in nature, versus 25% for OC^{4,5}. Hereditary BC in males represents 1% of all BC cases, and these individuals have an increased risk of prostate cancer⁶. The relative risk (RR) of an individual is estimated based on the relatedness to and number of affected family members, and the age of disease onset. For example, a meta-analysis of studies having quantified BC risk concluded that patients with a 1st-degree affected family member have a 2.1 RR⁴. This increases in patients below the age of 50, or when affected family members were diagnosed before 50. In the case of OC, the RR with one affected family member is estimated at 3.1 and almost doubles if this individual is the mother (6.0)⁷.

1.1.2 Linkage Studies and Mutation Data

In the 1990s, the genes *BRCA1* and *BRCA2* were linked to families with high incidence and early onset of BC^{8,9}. These genes are thought to harbor the most number of deleterious mutations for families with multiple cases of BC and OC, as determined by Ford *et al.* (1998) in a study of all patients within the Breast Cancer Linkage Consortium (BCLC) who had at least four affected family members (N=237)¹⁰. The results of this study estimated that disease was linked to *BRCA1* and *BRCA2* in 52 and 35% of families, respectively. Findings were congruent with the model for high penetrance genes¹¹, as linkage to *BRCA1/2* increased as a function of the number of affected family members. Alternatively, families with fewer affected members showed greater linkage to non-*BRCA1/2* genes. Families with four or five affected family members showed linkage to other genes in 33% of cases, whereas those with more than six affected members were estimated at only 4% linkage.

It is important to note that in the study by Ford *et al.* (1998), the percentage of families with reported *BRCA* mutations did not agree with the linkage data: only 63% of families showing linkage to *BRCA1* were found to harbor mutations when the coding and splice junctions were sequenced. The same phenomenon was observed in families linked to *BRCA2* (roughly 38% had reported mutations). Other studies have made similar observations. For example, in a study of 706 Dutch families, pathogenic variants were identified in 7.4% of families for *BRCA1* and 3.5% for *BRCA2*¹². Also, in Spanish families with at least 3 cases of BC and/or OC, or one male BC case, deleterious variants in *BRCA1* and *BRCA2* were identified 9.6 and 8.5% of the time, respectively¹³.

Some of this variation in the proportion of pathogenic *BRCA1/2* mutations identified in these studies can be attributed to the difference in age, ethnicity, and number of affected family members for each cohort investigated. However, there are also likely unrecognized or un-identified variants in *BRCA1/2* that could account for the disparity between linkage and mutation data. Finally, with respect to families that show no linkage to the *BRCA* genes, the most likely cause of disease is still genetic in nature, as opposed

to environmental, and can be explained by moderate and low-risk susceptibility genes¹⁴⁻¹⁷.

1.2 Molecular Diagnostics and Variant Classification

When variants are detected through genetic susceptibility testing in a clinical laboratory, they are classified on the basis of their probability of pathogenicity. This allows counselors and clinicians to offer further testing recommendations, treatment options, and prevention testing within the family. This process is fairly straightforward in the case of the well-studied, high-penetrance genes, for which clear management and treatment guidelines exist¹⁸⁻²³. However, as described by Hollestelle *et al.* (2010), the risk associated with moderate and low-penetrance genes is not as easily determined, as pathogenic variants in these genes are not easily recognized²⁴. These genes do not cause disease patterns characteristic of high-penetrance, and large families within large studies (which are not always available/possible) are necessary to confidently classify an allele as pathogenic or benign. As a result, clinical recommendations for gene variants in the more recently discovered susceptibility genes have been difficult to establish. Consequently, variants identified in these genes remain in the VUS category or worse, misclassified, often resulting in patient mismanagement^{25,26}. In addition, while test centres are showing concordance in their variant identification ability²⁷, classification approaches have not been advancing at the same pace, and tend to be inconsistent between clinical laboratories^{28,29}, exacerbating the difficulties associated with the classification of VUS.

1.2.1 Current Classification Systems

Previously, variants detected through genetic susceptibility testing would fall under one of three groups: definitely pathogenic, of no clinical significance, and uncertain³⁰. According to the International Association for Research on Cancer (IARC) Unclassified Genetic Variants Working Group, this classification of variants is very conservative, such that all variants with 0.1 to 99% probability of pathogenicity are classified under the “unknown” category³¹. Variants categorized as definitely pathogenic are limited to coding variants resulting in a clear inactivating mutation (such as an insertion/deletion

[indel] causing frameshift or a nonsense mutation). Consequently, all patients with variants in this middle category are not eligible for testing of at-risk relatives and they are given no further insight to their disease.

This led to the development of a 5-tiered classification system that has been developed, expanded upon, and implemented by two key working bodies: the IARC and the American College of Medical Genetics and Genomics (ACMG).

1.2.1.1 IARC Recommendations

The IARC, which is a branch of the World Health Organization, created the Working Group with the objective of increasing communication and collaboration between large centres conducting research on genetic diseases. It was determined that it would be beneficial to introduce intermediate categories to the original 3-tiered classification system, allowing for better patient care and prevention. An “integrated” approach using multiple lines of evidence is recommended. This integration involves combining the prior probability of pathogenicity (prior knowledge of the variant and therefore information from other groups) with observed data (i.e. new information) to arrive at a final probability of pathogenicity³², otherwise known as likelihood ratio (LR). This method allows for new information to constantly be accounted for, an option that is vital in this era of constant discovery and open-access of information. Both quantitative and qualitative data can be used in this model, and can be divided into direct and indirect measures of pathogenicity:

Direct measures: co-segregation of the variant with the disease phenotype (determined by genetic testing of affected and unaffected family members), severity of cancer family history, and co-occurrence of the VUS with another clearly pathogenic variant;

Indirect measures: *in silico* assessment of sequence and structure alterations compared to evolutionary conservation, assessment of a variant’s effect on splicing, and *in vitro* functional analysis.

Following LR computation based on as many lines of evidence as possible, variants are recommended to be categorized using the classes listed in **Table 1.1**.

1.2.1.2 ACMG Recommendations

The ACMG recommendations³³ were published in March of 2015 with the goal of promoting standardized terminology and a uniform classification system between centres. The classes are similar to those described by the IARC (**Table 1.2-1.4**), in that they follow a 5-tiered system based on the likelihood of pathogenicity, however much more structure is provided when it comes to the many lines of evidence that can be used to contribute to the likelihood approximation (i.e. assigning weights and thresholds) This approach was tested by evaluating the consistency of classification between a number of different centres, and has ultimately been accepted and supported by the majority of laboratories. It is important to note that these recommendations are specifically for Mendelian disorders.

1.3 VUS and Disease

As mentioned above, while NGS has allowed for the cataloguing of numerous disease associated variants, the literature has consequently been flooded with variants lacking a clinical interpretation³⁴. Extensive efforts are being made to classify and interpret VUS, as demonstrated by the IARC Working Group and various disease-specific consortiums such as the Evidence-based Interpretation of Germline and Mutant Allele (ENIGMA) Consortium and the International Society for Gastrointestinal Hereditary Tumours (InSiGHT). In the meantime, clinicians and genetic counsellors struggle to select which variants are actionable and warrant reporting back to patients. Plon *et al.* (2011) report that physicians ordering panel testing for patients often advise inappropriately, providing counselling for a VUS that would normally apply to a clearly pathogenic variant³⁵.

1.3.1 Consequences of Reporting Unknown Variants

Studies have shown that patients who undergo genetic susceptibility testing expect a dichotomous “mutation/no mutation” result³⁶. However, as discussed above, most

Table 1.1. Proposed Classification System for Sequence Variants Identified by Genetic Testing*

Class	Description	Probability of being Pathogenic
5	Definitely Pathogenic	> 0.99
4	Likely Pathogenic	0.95-0.99
3	Uncertain	0.05-0.949
2	Likely Not Pathogenic or of Little Clinical Significance	0.001-0.049
1	Not Pathogenic or of No Clinical Significance	< 0.001

*Plon, S. E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**, 1282–1291 (2008). Copy of license agreement for Table re-use is provided from John/Wiley & Sons, Inc. (see **Appendix A**).

Table 1.2 ACMG Criteria for Classifying Pathogenic Variants*

Evidence of pathogenicity	Category
Very strong	<p>PVS1 null variant (nonsense, frameshift, canonical ± 1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease</p> <p>Caveats:</p> <ul style="list-style-type: none"> • Beware of genes where LOF is not a known disease mechanism (e.g., <i>GFAP</i>, <i>MYH7</i>) • Use caution interpreting LOF variants at the extreme 3' end of a gene • Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact • Use caution in the presence of multiple transcripts
Strong	<p>PS1 Same amino acid change as a previously established pathogenic variant regardless of nucleotide change</p> <p>Example: Val→Leu caused by either G>C or G>T in the same codon</p> <p>Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level</p> <p>PS2 De novo (<u>both</u> maternity and paternity confirmed) in a patient with the disease and no family history</p> <p>Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo transfer, and so on, can contribute to nonmaternity.</p> <p>PS3 Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product</p> <p>Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical diagnostic laboratory setting are considered the most well established.</p> <p>PS4 The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls</p> <p>Note 1: Relative risk or OR, as obtained from case-control studies, is >5.0, and the confidence interval around the estimate of relative risk or OR does not include 1.0. See the article for detailed guidance.</p> <p>Note 2: In instances of very rare variants where case-control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence.</p>
Moderate	<p>PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation</p> <p>PM2 Absent from controls (or at extremely low frequency if recessive) (Table 6) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium</p> <p>Caveat: Population data for insertions/deletions may be poorly called by next-generation sequencing.</p> <p>PM3 For recessive disorders, detected in trans with a pathogenic variant</p> <p>Note: This requires testing of parents (or offspring) to determine phase.</p> <p>PM4 Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants</p> <p>PM5 Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before</p> <p>Example: Arg156His is pathogenic; now you observe Arg156Cys</p> <p>Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level.</p> <p>PM6 Assumed de novo, but without confirmation of paternity and maternity</p>
Supporting	<p>PP1 cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease</p> <p>Note: May be used as stronger evidence with increasing segregation data</p> <p>PP2 Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease</p> <p>PP3 Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)</p> <p>Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm should not be counted as an independent criterion. PP3 can be used only once in any evaluation of a variant.</p> <p>PP4 Patient's phenotype or family history is highly specific for a disease with a single genetic etiology</p> <p>PP5 Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation</p>

LOF, loss of function; OR, odds ratio.

*Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015). Copy of license agreement for Table re-use is provided from Nature Publishing Group (see **Appendix B**).

Table 1.3. ACGM Criteria For Classifying Benign Variants*

Evidence of benign impact	Category
Stand-alone	BA1 Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
Strong	BS1 Allele frequency is greater than expected for disorder (see Table 6)
	BS2 Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age
	BS3 Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing
	BS4 Lack of segregation in affected members of a family Caveat: The presence of phenocopies for common phenotypes (i.e., cancer, epilepsy) can mimic lack of segregation among affected individuals. Also, families may have more than one pathogenic variant contributing to an autosomal dominant disorder, further confounding an apparent lack of segregation.
Supporting	BP1 Missense variant in a gene for which primarily truncating variants are known to cause disease
	BP2 Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern
	BP3 In-frame deletions/insertions in a repetitive region without a known function
	BP4 Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.) Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm cannot be counted as an independent criterion. BP4 can be used only once in any evaluation of a variant.
	BP5 Variant found in a case with an alternate molecular basis for disease
	BP6 Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation
	BP7 A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

*Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015). Copy of license agreement for Table re-use is provided from Nature Publishing Group (see **Appendix B**).

Table 1.4. ACMG Rules for Combining Criteria to Classify Sequence Variants[§]

Pathogenic	<ul style="list-style-type: none"> (i) 1 Very strong (PVS1) <i>AND</i> <li style="padding-left: 20px;">(a) ≥ 1 Strong (PS1–PS4) <i>OR</i> <li style="padding-left: 20px;">(b) ≥ 2 Moderate (PM1–PM6) <i>OR</i> <li style="padding-left: 20px;">(c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) <i>OR</i> <li style="padding-left: 20px;">(d) ≥ 2 Supporting (PP1–PP5) (ii) ≥ 2 Strong (PS1–PS4) <i>OR</i> (iii) 1 Strong (PS1–PS4) <i>AND</i> <li style="padding-left: 20px;">(a) ≥ 3 Moderate (PM1–PM6) <i>OR</i> <li style="padding-left: 20px;">(b) 2 Moderate (PM1–PM6) <i>AND</i> ≥ 2 Supporting (PP1–PP5) <i>OR</i> <li style="padding-left: 20px;">(c) 1 Moderate (PM1–PM6) <i>AND</i> ≥ 4 supporting (PP1–PP5)
Likely pathogenic	<ul style="list-style-type: none"> (i) 1 Very strong (PVS1) <i>AND</i> 1 moderate (PM1–PM6) <i>OR</i> (ii) 1 Strong (PS1–PS4) <i>AND</i> 1–2 moderate (PM1–PM6) <i>OR</i> (iii) 1 Strong (PS1–PS4) <i>AND</i> ≥ 2 supporting (PP1–PP5) <i>OR</i> (iv) ≥ 3 Moderate (PM1–PM6) <i>OR</i> (v) 2 Moderate (PM1–PM6) <i>AND</i> ≥ 2 supporting (PP1–PP5) <i>OR</i> (vi) 1 Moderate (PM1–PM6) <i>AND</i> ≥ 4 supporting (PP1–PP5)
Benign	<ul style="list-style-type: none"> (i) 1 Stand-alone (BA1) <i>OR</i> (ii) ≥ 2 Strong (BS1–BS4)
Likely benign	<ul style="list-style-type: none"> (i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) <i>OR</i> (ii) ≥ 2 Supporting (BP1–BP7)
Uncertain significance	<ul style="list-style-type: none"> (i) Other criteria shown above are not met <i>OR</i> (ii) the criteria for benign and pathogenic are contradictory

[§]Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015). Copy of license agreement for Table re-use is provided from Nature Publishing Group (see **Appendix B**).

variants are classified as unknown and approximately 90% of patients receive an “uncertain” result³⁷. Consequently, these patients are given no further insight to their disease, and no opportunity to perform genetic prevention screening in family members, despite a clear indication of a genetic basis for the disease.

Two types of uncertain results exist: uncertain negative and uncertain positive³⁷. The former represents cases where no mutation was identified in the investigated regions. Uncertain positive results encompass variants that were identified in the coding region, but their pathogenicity and association with disease was uncertain (“unknown” variant classification category mentioned previously). Studies have also shown that disclosing information regarding an uncertain variant is more likely to cause distress and these results are often misinterpreted. For example, patients have been documented to undergo unnecessary prophylactic surgery following the reporting of an unknown variant³⁸. A better understanding of these novel, highly prevalent variants is therefore necessary for the improvement of screening, risk assessment, prevention, and genetic counseling.

1.3.2 Implications of VUS for HBOC

Prior to the advent of sequencing technologies, which made the investigation of areas outside the coding regions more affordable and realistic, VUS within the coding regions of *BRCA1/2* represented a significant proportion of identified variants and currently account for 20% of cases where an uncertain result is reported (32 and 53% of *BRCA1* and *BRCA2* mutations, respectively)¹². The remaining 80% of unknown cases result from no pathogenic mutation detected in the tested regions, or “uncertain positive” results. This large proportion is thought to be composed of: variants in untested genes, variants in untested regions of *BRCA1/2* (non-coding regions)^{30,39}, and unrecognized deleterious *BRCA1/2* variants that are detected and falsely classified as non-pathogenic (false negative) due to an inability to accurately assess pathogenicity¹². Consequently, VUS in *BRCA1/2* greatly outnumber known deleterious mutations (71.8% of variants listed in the BIC database are either pending classification or of unknown clinical significance)⁴⁰.

1.3.3 Scope of VUS

Classes of pathogenic variants are limited to clear truncating mutations (indels, nonsense, and variants within the dinucleotides of natural SSs), and missense mutations for which RR and/or functional analysis have been performed^{41,42}. However, coding regions of the genome represent only a small fraction of the genetic code (1.5%), and therefore any pathogenic variants within the non-coding regions and intergenic regions are ignored⁴³⁻⁴⁹. Coding regions are also known to harbor binding sites that play roles in splicing, which are also likely to be overlooked.

1.3.3.1 VUS in Non-coding Regions

The significance of investigating VUS and non-coding regions is widespread; a meta-analysis of single nucleotide polymorphisms (SNPs) associated with 22 common diseases found 39% to occur in intergenic regions⁵⁰. Furthermore, 80.4% of *MUTYH* variants, associated with hereditary non-polyposis colorectal cancer (HNPCC), are of unknown significance and 22% of β^0 -thalassemia single nucleotide variants (SNVs) were located within intergenic regions^{51,52}.

Non-coding regions often harbor sequences bound by regulatory factors involved in splicing, transcription, or untranslated region (UTR) stability⁵³⁻⁵⁶. The impact of a single nucleotide change in a recognition site can range from insignificant to complete abolition. Interpretation of these variants is therefore complex and computational methods are required for their analysis⁵⁷. In the case of variants affecting splicing, Horvath, *et al.* (2013) used RNA sequencing of triple-negative, non-triple negative, and HER2-positive BC tumor samples to identify novel variants affecting splicing and validated their data using functional assays⁵⁸. (Triple-negative breast cancer refers to the absence of estrogen receptor (ER), progesterone receptor (PR), and hormone epidermal growth factor receptor 2 (HER2) expression by the tumor, such that the tumor will not respond to receptor-targeted treatment. Non-triple negative breast cancer is when one, two, or three of these receptors is positive [i.e. expressed by the tumor].) In another study of unclassified variants occurring in Spanish patients with Lynch Syndrome, splicing mutations accounted for an estimated 15% of cases⁵⁹.

As for the implication of transcription factor binding sites (TFBSs) affected by variants, Guo and Jamison (2005) showed that SNPs found within the promoter and upstream regions of genes are not evenly distributed. Rather, they are concentrated in regions closer to the transcriptional start site and a higher proportion fall within predicted TFBSs than non-binding sites⁶⁰.

Variants within the UTR can affect both structure and stability of the transcript. Binding proteins responsible for the regulation of these processes, as well as cellular localization and translation, are also commonly contained within the UTR. Some variants appear to have an effect on a combination of these functions, as demonstrated by Zeng *et al.* (2014) where a 3'UTR variant that altered transcript structure, thereby decreasing stability and allowing for the creation of a GAPDH binding site⁶¹. Another example is the *SMAR1* gene from the growth arrest pathway, which is stabilized by prostaglandin binding to its 5' stem loop structure. BC- derived cell lines have been shown to have a 5'UTR variant leading to the abolition of this stem loop and low levels of the *SMAR1* tumor suppressor⁶².

Finally, regions associated with microRNAs, histone binding, and other epigenetic events can affect transcript levels and thus protein expression, however these topics are beyond the scope of this thesis.

1.3.3.2 VUS in Coding Regions

Pathogenic variants identified in *BRCA1/2* coding regions are attributed to nonsense variants, as well as indel and splicing variants resulting in frameshift, prematurely truncating the mRNA transcript and downstream degradation through nonsense-mediated decay^{63,64}. Predicting the functional significance of such mutations is straightforward, whereas VUS in these regions (synonymous, missense, short in-frame indels and splice variants outside the canonical dinucleotide donor and acceptor sites) are more difficult to interpret as they confer either a negligible alteration or none at all to the downstream protein sequence⁶⁴. *In silico* prediction tools exist to determine alterations in conserved sequences in the case of missense variants, but these predictions provide no experimental

proof, would be time-consuming to validate, and thus problematic for risk evaluation in the clinical lab⁶⁵⁻⁶⁷. The coding regions may also harbor cryptic SSs or regions bound by regulatory factors involved in splicing and transcription factor (TF) binding, which are similarly difficult to interpret and unrealistic for a clinical lab to include in their routine testing.

1.3.4 Prioritization of VUS

Despite efforts to classify variants based on recommended guidelines, the majority remain as VUS. Lack of pedigree information, under-reporting of rare variants, and an inability to functionally validate every VUS are major limiting factors to variant classification. We therefore see a need to efficiently and effectively evaluate variants in a way that allows one to rank, or prioritize, variants for further investigation, while remaining confident that deleterious variants are not overlooked.

1.3.4.1 Information Theory-based Analysis

Ex vivo transfection assays developed to determine the pathogenicity of VUS predicted (using *in silico* tools) to lead to splicing aberrations have been successful in identifying pathogenic mutations from VUS^{68,69}. Information theory (IT)-based analysis of splicing variants has proven to be robust and accurate in analyzing SS variants (both of natural and cryptic sites) as well as splicing regulatory factor binding sites (SRFBSs), and can distinguish polymorphisms from these variants in both rare and common diseases⁷⁰. However, IT can be applied to any sequence recognized and bound by another factor⁷¹, such as with TFBSs and RNA-binding protein binding sites (RBBSs). IT is described in further detail in a recently published literature review of information-theoretical analysis applied to mRNA splicing mutations and disease (*Chapter 2*)⁷⁰.

IT measures nucleotide sequence conservation, and cannot provide information on effects of variants on mRNA secondary (2°) structure, nor can it accurately predict effects of amino acid sequence changes. Other *in silico* methods address these deficiencies. For example, Halvorsen *et al.* (2010) introduce an algorithm called SNPfold, which computes the potential effect of a SNV on mRNA 2° structure⁴⁵. Predictions made by SNPfold can

be tested by the SHAPE assay (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension)⁷², which provides evidence for sequence variants that lead to structural changes in mRNA by detection of covalent adducts in mRNA.

1.4 Thesis Objectives

The inability to functionally validate the overwhelming amount of VUS generated by panel and NGS sequencing has created a need for process that quickly and accurately bridges the variant identification and classification processes. *We hypothesize that IT-based analysis can be applied in a unified framework for the interpretation and prioritization of gene variants in non-coding and coding regions.* In order to investigate this, our objectives were to:

1. Design custom probes for in-solution hybridization targeting the complete coding, non-coding, and up- and downstream regions of a panel of HBOC susceptibility genes;
2. Recruit a cohort of high-risk *BRCA*-uninformative patients to participate in this study;
3. Evaluate the ability of IT-based models to predict and prioritize potential non-coding sequence mutations in SS, TFBS, and RBBS for a cohort of anonymous high-risk *BRCA*-uninformative HBOC patients. Then, apply the model to the cohort of recruited patients and supplement the interpretation with pedigree analysis.

1.5 References

1. Howlader, N. *et al.* Cancer Statistics Review, 1975-2011 - SEER Statistics. (2014). at <http://seer.cancer.gov/csr/1975_2011/>
2. Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* **127**, 2893–2917 (2010).
3. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* **358**, 1389–1399 (2001).
4. Pharoah, P. D., Day, N. E., Duffy, S., Easton, D. F. & Ponder, B. A. Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int. J. Cancer J. Int. Cancer* **71**, 800–809 (1997).
5. Walsh, T. *et al.* Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18032–18037 (2011).
6. Plevová, P. & Hladíková, A. [Genetic counselling in male carriers of BRCA1 and BRCA2 gene mutations]. *Klin. Onkol. Cas. České Slov. Onkol. Společnosti* **25 Suppl**, S67–73 (2012).
7. Stratton, J. F., Pharoah, P., Smith, S. K., Easton, D. & Ponder, B. A. A systematic review and meta-analysis of family history and risk of ovarian cancer. *Br. J. Obstet. Gynaecol.* **105**, 493–499 (1998).
8. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
9. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**, 2088–2090 (1994).
10. Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **62**, 676–689 (1998).
11. Claus, E. B., Risch, N. & Thompson, W. D. Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am. J. Hum. Genet.* **48**, 232–242 (1991).

12. Vink, G. R., van Asperen, C. J., Devilee, P., Breuning, M. H. & Bakker, E. Unclassified variants in disease-causing genes: nonuniformity of genetic testing and counselling, a proposal for guidelines. *Eur. J. Hum. Genet. EJHG* **13**, 525–527 (2005).
13. Díez, O. *et al.* Analysis of BRCA1 and BRCA2 genes in Spanish breast/ovarian cancer patients: A high proportion of mutations unique to Spain and evidence of founder effects. *Hum. Mutat.* **22**, 301–312 (2003).
14. Ahlbom, A. *et al.* Cancer in twins: genetic and nongenetic familial risk factors. *J. Natl. Cancer Inst.* **89**, 287–293 (1997).
15. Antoniou, A. C. & Easton, D. F. Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898–5905 (2006).
16. Mack, T. M., Hamilton, A. S., Press, M. F., Diep, A. & Rappaport, E. B. Heritable breast cancer in twins. *Br. J. Cancer* **87**, 294–300 (2002).
17. Peto, J. Breast cancer susceptibility—A new look at an old model. *Cancer Cell* **1**, 411–412 (2002).
18. American College of Obstetricians and Gynecologists, ACOG Committee on Practice Bulletins--Gynecology, ACOG Committee on Genetics & Society of Gynecologic Oncologists. ACOG Practice Bulletin No. 103: Hereditary breast and ovarian cancer syndrome. *Obstet. Gynecol.* **113**, 957–966 (2009).
19. Kluijdt, I. *et al.* Familial gastric cancer: guidelines for diagnosis, treatment and periodic surveillance. *Fam. Cancer* **11**, 363–369 (2012).
20. Recently updated NCCN Clinical Practice Guidelines in Oncology™ (NCCN Guidelines®): Genetic/Familial High-Risk Assessment: Colorectal. Version 2.2014. at <http://www.nccn.org/professionals/physician_gls/recently_updated.asp>
21. NHSBSP No 74: Protocols for the surveillance of women at higher risk of developing breast cancer. at <<http://www.cancerscreening.nhs.uk/breastscreen/publications/nhsbsp74.html>>
22. Vasen, H. F. A. *et al.* Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer). *J. Med. Genet.* **44**, 353–362 (2007).

23. Shulman, L. P. Hereditary Breast and Ovarian Cancer (HBOC): Clinical Features and Counseling for BRCA1 and BRCA2, Lynch Syndrome, Cowden Syndrome, and Li-Fraumeni Syndrome. *Obstet. Gynecol. Clin. North Am.* **37**, 109–133 (2010).
24. Hollestelle, A., Wasielewski, M., Martens, J. W. & Schutte, M. Discovering moderate-risk breast cancer susceptibility genes. *Curr. Opin. Genet. Dev.* **20**, 268–276 (2010).
25. Wideroff, L. *et al.* Hereditary breast/ovarian and colorectal cancer genetics knowledge in a national sample of US physicians. *J. Med. Genet.* **42**, 749–755 (2005).
26. Pal, T. *et al.* A statewide survey of practitioners to assess knowledge and clinical practices regarding hereditary breast and ovarian cancer. *Genet. Test. Mol. Biomark.* **17**, 367–375 (2013).
27. Schroeder, C. *et al.* HBOC multi-gene panel testing: comparison of two sequencing centers. *Breast Cancer Res. Treat.* **152**, 129–136 (2015).
28. Vail, P. J. *et al.* Comparison of locus-specific databases for BRCA1 and BRCA2 variants reveals disparity in variant classification within and among databases. *J. Community Genet.* 1–9 (2015). doi:10.1007/s12687-015-0220-x
29. Yorzcyk, A., Robinson, L. S. & Ross, T. S. Use of panel tests in place of single gene tests in the cancer genetics clinic. *Clin. Genet.* [**Epub ahead of print**], (2014).
30. Plon, S. E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**, 1282–1291 (2008).
31. Goldgar, D. E. *et al.* Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am. J. Hum. Genet.* **75**, 535–544 (2004).
32. Goldgar, D. E. *et al.* Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Hum. Mutat.* **29**, 1265–1272 (2008).
33. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical

- Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).
34. Cassa, C. A. *et al.* Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res.* **22**, 421–428 (2012).
 35. Plon, S. E. *et al.* Genetic testing and cancer risk management recommendations by physicians for at-risk relatives. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **13**, 148–154 (2011).
 36. Press, N. A., Yasui, Y., Reynolds, S., Durfy, S. J. & Burke, W. Women's interest in genetic testing for breast cancer susceptibility may be based on unrealistic expectations. *Am. J. Med. Genet.* **99**, 99–110 (2001).
 37. Vos, J. *et al.* The counselees' view of an unclassified variant in BRCA1/2: recall, interpretation, and impact on life. *Psychooncology.* **17**, 822–830 (2008).
 38. Vos, J. *et al.* Opening the psychological black box in genetic counseling. The psychological impact of DNA testing is predicted by the counselees' perception, the medical impact by the pathogenic or uninformative BRCA1/2-result. *Psychooncology.* **21**, 29–42 (2012).
 39. Levy-Lahad, E. & Plon, S. E. Cancer. A risky business--assessing breast cancer risk. *Science* **302**, 574–575 (2003).
 40. Borg, A. *et al.* Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum. Mutat.* **31**, E1200–40 (2010).
 41. Domchek, S. & Weber, B. L. Genetic variants of uncertain significance: flies in the ointment. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **26**, 16–17 (2008).
 42. Braun, T. A. *et al.* Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. *Hum. Mol. Genet.* **22**, 5136–5145 (2013).
 43. Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends Genet. TIG* **29**, 318–327 (2013).
 44. Chatterjee, S., Berwal, S. K. & Pal, J. K. in *eLS* (John Wiley & Sons, Ltd, 2001). at <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0022408/abstract>

45. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* **6**, e1001074 (2010).
46. Araujo, P. R. *et al.* Before It Gets Started: Regulating Translation at the 5' UTR. *Comp. Funct. Genomics* **2012**, 475731 (2012).
47. Misquitta, C. M., Iyer, V. R., Werstiuk, E. S. & Grover, A. K. The role of 3'-untranslated region (3'-UTR) mediated mRNA stability in cardiovascular pathophysiology. *Mol. Cell. Biochem.* **224**, 53–67 (2001).
48. Latchman, D. S. Transcription-Factor Mutations and Disease. *N. Engl. J. Med.* **334**, 28–33 (1996).
49. Ward, A. J. & Cooper, T. A. The Pathobiology of Splicing. *J. Pathol.* **220**, 152–163 (2010).
50. Glinskii, A. B. *et al.* Identification of intergenic trans-regulatory RNAs containing a disease-linked SNP sequence and targeting cell cycle progression/differentiation pathways in multiple common human disorders. *Cell Cycle Georget. Tex* **8**, 3925–3942 (2009).
51. Out, A. A. *et al.* Leiden Open Variation Database of the MUTYH gene. *Hum. Mutat.* **31**, 1205–1215 (2010).
52. Nuinon, M. *et al.* A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. *Hum. Genet.* **127**, 303–314 (2010).
53. Ko, L. J. & Engel, J. D. DNA-binding specificities of the GATA transcription factor family. *Mol. Cell. Biol.* **13**, 4011–4022 (1993).
54. Maity, S. N. & Crombrugghe, B. de. Role of the CCAAT-binding protein CBF/NFY in transcription. *Trends Biochem. Sci.* **23**, 174–178 (1998).
55. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–577 (1997).
56. Quesne, J. L. UTRly malignant: mRNA stability and the invasive phenotype in breast cancer. *J. Pathol.* **230**, 129–131 (2013).
57. Paul, D. S., Soranzo, N. & Beck, S. Functional interpretation of non-coding sequence variation: Concepts and challenges. *BioEssays* **36**, 191–199 (2014).

58. Horvath, A. *et al.* Novel insights into breast cancer genetic variance through RNA sequencing. *Sci. Rep.* **3**, 2256 (2013).
59. Pérez-Cabornero, L. *et al.* Evaluating the effect of unclassified variants identified in MMR genes using phenotypic features, bioinformatics prediction, and RNA assays. *J. Mol. Diagn. JMD* **15**, 380–390 (2013).
60. Guo, Y. & Jamison, D. C. The distribution of SNPs in human gene regulatory regions. *BMC Genomics* **6**, 140 (2005).
61. Zeng, T. *et al.* A novel variant in the 3' UTR of human SCN1A gene from a patient with Dravet syndrome decreases mRNA stability mediated by GAPDH's binding. *Hum. Genet.* (2014). doi:10.1007/s00439-014-1422-8
62. Pavithra, L. *et al.* Stabilization of SMAR1 mRNA by PGA2 involves a stem loop structure in the 5' UTR. *Nucleic Acids Res.* **35**, 6004–6016 (2007).
63. Campos, B. *et al.* RNA analysis of eight BRCA1 and BRCA2 unclassified variants identified in breast/ovarian cancer families from Spain. *Hum. Mutat.* **22**, 337–337 (2003).
64. Vallée, M. P. *et al.* Classification of missense substitutions in the BRCA genes: a database dedicated to Ex-UVs. *Hum. Mutat.* **33**, 22–28 (2012).
65. Kavanagh, D. & Anderson, H. E. Interpretation of genetic variants of uncertain significance in atypical hemolytic uremic syndrome. *Kidney Int.* **81**, 11–13 (2012).
66. Schwartz, G. F. *et al.* Proceedings of the international consensus conference on breast cancer risk, genetics, & risk management, April, 2007. *Cancer* **113**, 2627–2637 (2008).
67. Tavtigian, S. V., Greenblatt, M. S., Lesueur, F., Byrnes, G. B. & Group, I. U. G. V. W. In silico analysis of missense substitutions using sequence-alignment based methods. *Hum. Mutat.* **29**, 1327–1336 (2008).
68. Gaildrat, P. *et al.* Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. *Methods Mol. Biol. Clifton NJ* **653**, 249–257 (2010).

69. Tournier, I. *et al.* A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum. Mutat.* **29**, 1412–1424 (2008).
70. Caminsky, N. G., Mucaki, E. J. & Rogan, P. K. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Research* **3**, 282 (2015).
71. Schneider, T. D., Stormo, G. D., Yarus, M. A. & Gold, L. Delila system tools. *Nucleic Acids Res.* **12**, 129–140 (1984).
72. Steen, K.-A., Siegfried, N. A. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by protection from exoribonuclease (RNase-detected SHAPE) for direct analysis of covalent adducts and of nucleotide flexibility in RNA. *Nat. Protoc.* **6**, 1683–1694 (2011).

Chapter 2

2 Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis

The work in this chapter is reproduced (with permission, **Appendix C**) from:

Caminsky N, Mucaki EJ and Rogan PK. Interpretation of mRNA splicing mutations in genetic disease: a review of the literature and guidelines for information-theoretical analysis. [v2; ref status: indexed, <http://f1000r.es/54y>] F1000Research 2015, 3:282. doi: 10.2688/f1000research.5654.2).

2.1 Introduction

Pre-mRNA splicing is a necessary step in the production of a functional protein product. It consists of the recognition of intron/exon boundaries, and the subsequent excision of the introns. It is important to distinguish between alternate splicing isoforms and mutant splice forms. The former consists of using different combinations of splice sites for the same gene. It is estimated to occur in over 60% of human genes, some of which will have multiple alternate isoforms^{1,2}. For example, *NF1* (neurofibromin 1 – associated with Neurofibromatosis Type 1) is reported to produce 46 splice variants³. The cell regulates this naturally occurring process through the availability of tissue specific splice factors. Alternative splicing is not generated by changes in the unspliced RNA sequence, whereas mutations that produce non-constitutive splice forms are the result of dysregulation of natural splice site recognition. Mutations can have various consequences to RNA processing, such as exon skipping, cryptic splicing, intron inclusion, leaky splicing, or less frequently, introduction of pseudo-exons into the processed mRNA. A broad range of molecular phenotypes are possible depending on the type and severity of the mutation, making it imperative to understand the consequences of splicing mutations. For the purposes of this review, we consider sequence changes in genes that affect transcript structure or abundance to be mutations, regardless of their allele frequencies. Although

spliceosomal recognition and RNA binding factors are operative in mutation-derived and normal alternative mRNA splicing events, this review is focused on aberrant sequence changes that alter constitutive splicing, and often result in clinically abnormal phenotypes.

The process of U1/U2-based mRNA splicing involves the recognition of a number of key sequence components^{4,5}, with exons defined by both intronic and exonic features^{4,6}. The intronic sequence flanking the 3' end of an intron is termed the donor site and the 5' end, the acceptor site. In typical mRNA splicing, the natural donor and acceptor splice sites of an exon span intervals of 10 and 28 bases in length, respectively. It is a common misconception that these sequences (especially the dinucleotides immediately intronic to the exon) are invariant. Although highly conserved, these sequences vary at different splice junctions within a gene as well as between genes. The particular combination of nucleotides at each position within the same splice determines its overall strength, which dictates the likelihood of recognition by the U1 and U2 spliceosomes. Indeed, the recognition of the internal exons is reliant on the strength of both natural splice sites⁷. The alteration of exonic splicing signals (described in the following paragraph) by coding variants is common (~25%)⁸, which plays a significant role in disease due to aberrant mRNA processing. The creation and loss of binding sites for these splicing factors can also result in small changes in mRNA structure and overall gene expression, and is part of the diverse tissue-specific regulatory ecosystem of the cell⁹.

In addition, binding sites for splicing regulatory elements have been shown to reside over a range of distances from the corresponding natural splice sites¹⁰; the impact of these sites appears to be related to their binding affinities to the cognate RNA binding proteins and to their distance from the proximate intron/exon boundary¹¹. Recognition sites for these regulatory proteins can reside either within introns or exons. Those within exons are commonly referred to as exonic splice enhancers or silencers (ESE or ESS, respectively), whereas the corresponding designations for intronic elements are ISE or ISS. Sequence variants affecting these protein-binding sites (or mutations in the binding proteins themselves) have been documented as contributing to aberrant splicing and pathogenic

phenotypes. We focus on variants occurring in *cis* with target genes, as opposed to those in the splicing complex (in *trans*), leading to abnormal splicing. The efficiency and specificity of splicing depends on the combination of natural splice site strengths and the binding of splicing regulatory proteins that orchestrate exon recognition¹².

Mutations that affect pre-mRNA splicing account for at least 15% of disease-causing mutations¹³ with up to 50% of all mutations described in some genes^{14,15}. Interpreting the effects that these variants have on splicing is not straightforward because natural and regulatory splice sites exhibit considerable sequence variation. Furthermore, performing *in vitro* experiments to verify the consequences of each variant is costly and time consuming, and may not be practical. *In silico* prediction methods have become essential resources for analyzing these variants. Software programs for splicing analysis use a wide variety of bioinformatic approaches. Several splice site prediction tools compare the predicted mutant sequence to a consensus sequence, based on a set of functional acceptor or donor splice sites¹⁶. A drawback of this approach is that low-frequency nucleotides present in functional splice sites are not represented, which can lead to misinterpretation and false-positive mutation predictions. One example of this was illustrated by Rogan and Schneider (1995), in which the *MSH2* (hereditary non-polyposis colon cancer) variant described by Fishel *et al.* (1993), IVS12-6T>C, was predicted to be benign, despite being located 6 nt from the natural acceptor splice junction^{17,18}. The consensus sequence fails to indicate that C and T at this position are nearly equally probable, which reclassified this transition as a polymorphism rather than a pathogenic variant. This conclusion is supported by evidence that ~10% of normal individuals without predisposition to non-polyposis colon cancer harbour this alternate allele¹⁹.

Over the last 20 years we, and others, have developed an information theory (IT)-based approach for prediction of splicing mutations, and their impact on mRNA structure and abundance. The effects of these mutations is founded on the formal relationship between IT and the second law of thermodynamics, in that the change in information ascribed to a sequence variant within a splice site is directly related to thermodynamic entropy and free energy of binding^{20,21}. A weight matrix consisting of the Shannon information (product of

the probability of each nucleotide and $-\log_2$ of its probability) at each position of the splice site is constructed. The individual information for a splice site (R_i , in bits) is defined as the dot product of this weight matrix and the unitary vector of a particular splice site sequence. The magnitude of the information content of a nucleotide within a given site is an indication of its level of conservation relative to a set of functional sites. This method retains all of the sequence variability inherent in each model of donor and acceptor splice sites. By contrast, each base in the consensus sequence has the maximum R_i value, which is actually rare in the human genome, and is generally not representative of the preponderance of natural splice sites. Prior to the introduction of IT-based approaches, consensus sequence-based methods were widely used¹⁶. The use of neural networks, trained on sequences experimentally determined to be “bound” and “unbound”, was another early approach used to predict splice sites²². However, these unbound set of sequences are known to harbour some contaminating functional sites^{23,24}, which can limit the sensitivity and specificity of these networks²⁵.

There are instances when IT does not accurately predict the consequences of a splice variant. This can often be attributed to instances involving multiple sites or multiple regulatory factors, which are not components of current splicing models. In addition, splicing regulatory proteins can share overlapping and degenerate binding sites, and may exert conflicting effects (for example, serine-arginine [SR] vs. hnRNP proteins), making *in silico* prediction less reliable and accurate in these cases²⁶. Finally, functional cryptic splicing motifs occurring deep within the introns can be challenging to identify, because they tend to be less well conserved than natural splice sites^{27,28}.

Nevertheless, a number of authors have recommended IT methods for analysis of splice site variants (N = 29; **Supplementary Table 1**; all supplementary tables are provided in the **Supplementary Content File**). In fact, this approach has been described as equivalent to using a general reference textbook as a diagnostic tool, which complemented by functional assays, may provide a complete molecular diagnosis²⁹. Most of the applications of IT for splicing mutation analysis have involved predominantly rare diseases, as well as some low frequency variants associated with more common genetic

conditions. This is because IT has been used to assess how well computed changes in binding affinity conform to levels of expression and/or patient phenotypes.

Many IT studies have focused on sequence variants in individual disorders or genes. Our synopsis of the broader implications of this work sets the stage for this compilation of peer-reviewed variants with accompanying IT analyses. We cover all publications retrieved through PubMed and Google Scholar that cite the use of IT (N = 367; **Appendix D**) before September 2014. These items include primary research articles, review articles, presentations, and theses. Of all references, 216 publications reported variants or other results or analyses pertinent to this review (**Supplementary Table 2**). In the remaining studies, analyses were either not performed, insufficient information was provided to reproduce the reported result, or authors described unrelated applications of IT-based analysis. We summarize the spectrum of variants analyzed to obtain a global perspective of splicing mutations resulting in genetic disease. We also highlight common errors that can occur in variant analysis and interpretation, and offer guidelines for optimal use of our software programs for interpretation of splicing mutations.

2.2 Information theory and splice site analysis

IT was first introduced by Claude Shannon in 1948 and is now used in a variety of disciplines to express the average number of bits (i.e. the information content) needed to communicate symbols in a message³⁰. Bits are the basic unit used in computing and can have one of two values (typically the answer to a yes/no, true/false, or +/- problem). In nucleic acid molecular biology, the symbols in the message comprise a group of related, aligned sequences, with the average number of bits in the set corresponding to the amount of information in the message. This is determined from the information content at each position in the sequence, summed over all positions³¹. The average information is depicted graphically by a sequence logo, which stacks the individual nucleotides at each position ranked by frequency, and where the height of the stack is the position-specific contribution to the average information³². If the set of sequences are functional binding sites recognized by the same factor, the individual information in each site (i.e. R_i) is related to thermodynamic entropy, and thus, to the free energy of binding²¹.

The information content of a nucleic acid binding site is related to the affinity of its interaction with proteins and other macromolecular complexes, such as the case during mRNA splicing²¹. Information theory-based position weight matrices (PWMs; R_i [b,l] - also referred to as a ribl - where b and l correspond to the nucleotide and position in the splice site, respectively) can be determined for set of known binding sites, in this case, for the purpose of calculating individual and average sequence information³¹. **Figure 2.1** shows an example of sequence logos for the canonical acceptor (or 3', recognized by the U2 spliceosome) and donor (or 5', recognized by the U1 spliceosome) splice sites, computed from the majority of constitutive sites at annotated splice junctions in the human genome³³. The information contained within the natural splice donor site is distributed between the last codon of each exon and the adjacent 7 nucleotides of intronic sequence, whereas the acceptor sites are almost entirely intronic, extending 26 nucleotides upstream from the exon boundary.

The distributions of R_i values for these sets are approximately Gaussian, with a couple of important exceptions, namely the distribution has defined upper and lower bounds²¹. The upper limit corresponds to the consensus sequence, as it is not possible to have stronger binding than an exact match to this sequence. The theoretical lower limit corresponds to $R_i = 0$ bits. An R_i value less than zero implies that energy would be required ($\Delta G > 0$ kcal/mol) for a stable binding complex to form (i.e. that the event would not occur spontaneously without an exogenous source of energy). The minimum strength site is zero bits, the equilibrium state ($\Delta G = 0$). Assuming the contacts at each position in the same binding site form independently, this approach is accurate and quantitative. Altering a nucleotide with high information (implying high prevalence and conservation at that position) will have a greater impact on binding, than if a less-well conserved base were altered. The change in information due to a mutation in a site (ΔR_i) is the difference between $R_{i,final}$ and $R_{i,initial}$ values, where $R_{i,final}$ is the information of the sequence containing the variant, and $R_{i,initial}$ the information of the reference (wild-type) sequence. The minimum fold change in binding affinity resulting from the mutation is an exponential function based on ΔR_i , or $\geq 2^{\Delta R_i}$ (Ref. 21).

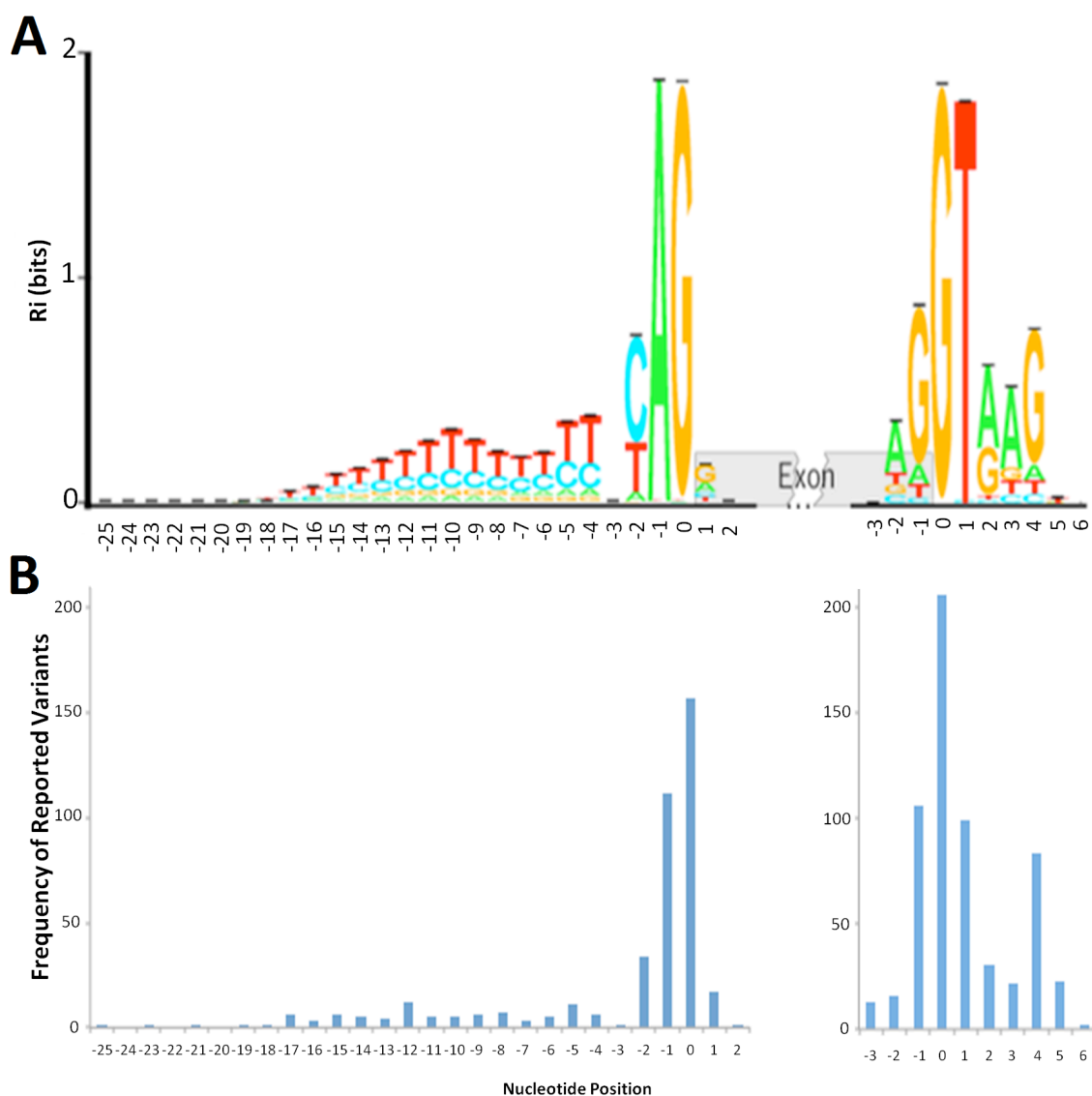


Figure 2.1. Distribution of deleterious natural site variants relative to information content

A) The sequence logo for human acceptor and donor splice sites based on the positive (+) strand of the October 2000 (hg5) genome draft. The logo shows the distribution of information contents (R_i in bits) at each position over the region of 28 nucleotides for acceptor [-25, +2] and 10 nucleotides for donor [-3, +6] from the first nucleotide of the splice junction (position 0). Nucleotide height represents its frequency at that position. The horizontal bar atop each stack indicates the standard deviation at that position. This

figure was modified from Rogan *et al.* (2003) to include splice sites in genes on both strands of the annotated human reference genome³³. **B)** The distribution of deleterious single-nucleotide variants reported at the natural acceptor (left) and donor (right) splice sites. The variants used to populate this graph (**Supplementary Table 3**) were included only if they were reported to negatively affect splicing (N = 431 for acceptors, 604 for donors). The image was aligned to the sequence logo (**A**) to illustrate potential correlation of number of splicing variants at a position to the information content at that position.

2.3 Software resources

2.3.1 Delila package/system

Information analysis was originally performed using the Delila sequence analysis system, which included a language to process nucleic acid sequences, and a library of sequence tools to retrieve and process various types of sequence data^{34,35}. Tools to measure information content of nucleic acid sequences were subsequently added to Delila³¹. Initially, models of information content of bacteriophage T7 RNA polymerase binding sites and other bacterial control systems were studied, and mRNA splice sites were subsequently developed^{31,36}. Later, tools to display binding sites as either a) sequence walkers showing individual information or b) sequence logos of average information, were incorporated into Delila^{23,32}. The former allows for the positive or negative contribution of a base at a given position to a sequence's information to be represented. The Automated Splice Site Analysis (ASSA) server introduced in 2004, and its successor, Automated Splice Site and Exon Definition Analysis server (ASSEDA), have been freely available throughout the last decade, and have been used for IT-based calculations on nucleic acid sequences for the preceding 20 years^{37,38}. Both ASSA and ASSEDA still use the Delila program suite to retrieve sequences, calculate information content, and create sequence walker representations of individual binding sites. ASSEDA is now available in a common interface with the other information theory-based tools described below at www.mutationforecaster.com.

2.3.2 ASSA/ASSEDA

To simplify mutation analysis, we built a web interface for variant analysis using Delila software as the processing backbone³⁷. Our aim was to standardize and facilitate IT-based mutation analysis by using Human Genome Variation Society (HGVS)-approved variant nomenclature (which has since become the worldwide standard), employing server-based retrieval/processing, and reporting results as concise predictions in both tabular and sequence walker display formats. Initially, ASSA results described mutations in relation to genome annotations from the first finished genome release (hg15)³⁷. While many

publications cited this version of ASSA for novel splicing mutation analysis, continued improvements have introduced more accurate reference sequences, annotations, and models (for both constitutive and regulatory splice sites) based on more comprehensive sets of binding sites. The ASSA server contained the original donor and acceptor information position weight matrices derived by manual curation of GenBank entries, murine donor and acceptor weight matrices, a subset of splicing enhancer elements (SF2/ASF, SC35 and SRp40), and the lariat branch point recognition sequence³⁶. ASSA reported the strengths of all potential sites predicted within the window selected by the user, highlighted those with the largest changes in R_i , and computed the minimum fold change in binding affinity for each mutation or polymorphism. Tabular results were colour-coded. Unaltered sites above and below the $R_{i,min}$ (described in *Minimum splice site information content and exceptions*) were highlighted grey and white, respectively. Pre-existing sites abolished by the variant (where $R_{i,final} < R_{i,min}$) were marked in red, while leaky natural sites ($R_{i,final} \geq R_{i,min}$) were highlighted in blue. Cryptic sites that were created, strengthened, or weakened were highlighted in pink, green and teal, respectively. The server parsed any mutation type described precisely by the HGVS notation, including substitutions, insertions, deletions, and combinations of these changes³⁹. Recapitulating variants described in articles before these guidelines were widely adopted proved to be highly time-consuming and error-prone²⁵. Multiple binding factors had to be analyzed simultaneously; however, results were reported independently. The analysis did not consider other factors relevant to splice site recognition, such as the resulting exon size, or potential formation of cryptically spliced exons.

ASSEDA, the successor software to ASSA, provides a new isoform-oriented type of mutation interpretation, updates the coordinate system to HG19 (GRCh37), adds current gene and single nucleotide polymorphism (SNP) annotations (dbSNP135), and provides additional ribls for other splicing regulatory sites (SRp55, TIA1, ELAVL1, hnRNP A1, hnRNP H, and PTB). All models, except those for SRp55 and hnRNP H, have been built using sequences from publicly available CLIP-seq data, and are based on a larger number of binding site sequences. They have been tested by comparing predictions to validated binding sites from published primary literature, and to any splice-altering variants found

within them³⁸. ASSEDA introduces *in silico* exon definition analysis by computing the total splicing information across an exon³⁸. Total exon information ($R_{i,total}$) is the sum of the corresponding donor and acceptor R_i values, and corrected for the gap surprisal term, which is based on the length of the potential exon formed using those sites (from RefSeq)⁴⁰. The gap surprisal function is based on the genome-wide distribution of constitutive exon lengths, also known as self-information. This term ensures that exons are computationally defined using donor and acceptor splice sites in close proximity^{40,41}.

Exons of uncommon length lead to large negative gap surprisal terms, which reduces $R_{i,total}$. When applied to predicted exons that activate a cryptic splice site, comparison of $R_{i,total}$ values can more accurately predict cryptic site use than the strength of this site alone. The gap surprisal term decreases the predicted $R_{i,total}$ value of particularly long internal exons (eg. the 3.4 kb long exon 11 of *BRCA1*; $R_{i,total} = 1.4$ bits), which tends to compensate for this effect with strong splice sites and other sequence elements that enhance natural splice site recognition and suppress internal cryptic splice sites.

The exon definition paradigm extends to the assessment of the impact of mutations in ESE/ISS elements. ASSEDA calculates $R_{i,total}$ by adding the R_i value of a regulatory splicing element to the contributions of constitutive splice sites, and applying a second gap surprisal term based on the frequency of distance from the splicing element to the nearest natural site. Currently, the effect of only a single splicing factor can be evaluated by the software at a time, although the approach itself is generalizable to multiple regulatory binding sites. If a variant causes changes in the R_i values of multiple sites, such as the simultaneous creation of both splicing enhancer and repressor elements, there will be less confidence in ASSEDA's predictions.

Two distinct sets of IT-based models for donors and acceptors are available on ASSEDA. The manually curated ribls were originally determined from 1,799 donor and 1,744 acceptor sites³⁶. We subsequently derived a set of ribl matrices from genome-wide exon annotations³³. These models were automatically curated using the criteria that enforced $R_i > 0$ for correctly annotated sites. The resultant models consisted of 108,079 acceptor and 111,772 donor splice sites, however these were not implemented on the ASSA server

until 2011³³. These genome-wide models are used in the calculation of $R_{i,total}$ values. The ΔR_i values for a single nucleotide splicing variant are similar for both sets of models. Variants having opposite predicted effects between the respective donor or acceptor ribls have not been reported. In general, the genome-wide models report slightly lower information contents, however the frequencies of nucleotides at the 5' end of the acceptor site differ significantly. This results in differences in the weights in the -4 to -20 nt region between the manually-curated and the genome-wide acceptor ribl matrix, which can significantly lower R_i values based on the genome-wide model. In the genome, thymine is more prevalent than cytosine at these positions and has a higher positive contribution to the overall R_i . This can account for up to a 1.97 bit difference between the models. Guanine nucleotides within this sequence window significantly lower the R_i values computed from the genome-wide acceptor, as well. While these differences contribute only a 0.1-0.4 bit difference to the R_i per nucleotide, the cumulative effect of multiple differences within this window can lead to significant differences between the acceptor R_i values.

2.3.3 Shannon pipeline and Veridical

High-throughput DNA sequencing is generating a deluge of novel variants in patients with genetic diseases, most of which currently have unknown significance (VUS). For example, 20% of the patients with Pelizaeus-Merzbacher disease possess VUS in the *PLP1* gene, among which are single or compound heterozygous, rare pathogenic mutations⁴². Many solutions have been proposed, however prediction of pathogenicity by bioinformatic analyses is often inaccurate⁴³. The Shannon Human Splicing Mutation Pipeline software predicts mutations at genome scale to predict which variants may alter mRNA splicing and is based on the same principles and IT models used in ASSA and ASSEDA⁴⁴. However, this software processes 5 million substitutions and/or indels in 10-15 minutes. While initially only available for the CLC-Bio Genomics platform, this software is now offered as a web service in the suite of programs available through Mutation Forecaster (www.mutationforecaster.com). Variants are batched in standard variant call format (VCF). The pipeline reports any genic variant that affects a known natural site or a cryptic site where $R_{i,initial}$ or $R_{i,final}$ are ≥ 0 bits and $\Delta R_i \geq 1.0$ bits, however

more stringent criteria for selecting variants with significant information changes can be applied.

In Shirley *et al.* (2013), all variants from the complete genomes of three cancer cell lines (A431, U2OS, U251; N = 816,275) were analyzed⁴⁴. Variants that were common ($\geq 1\%$) were removed. Variants that weakened natural sites, or strengthened cryptic sites to levels comparable to or exceeding the strength to the nearest natural site, were flagged. Variants that strengthen a natural site could have an effect on the splicing profile of a gene (i.e. reduce the frequency of exon skipping for the corresponding exon), but are less likely to cause a deleterious phenotype. The overall fraction of mutations flagged, after filtering out distant cryptic sites and small ΔR_i values, averaged 0.016%, illustrating how the software can be used for prioritizing variants. Some of the prioritized variants occurred in genes with known defective functional and biochemical pathways in these cancer cell types, i.e. cytokine signalling (in A431), DNA replication and cell cycle (in U2OS). Natural splice mutations were confirmed by expression data to a greater extent than leaky or cryptic splice site variants.

In a complete cancer cell line genome, the number of cryptic sites with altered R_i values greatly exceeds the number of affected natural splice sites. Many of these are weak decoys, which can occur throughout genes. Using the principle that novel cryptic sites that are likely to be activated must compete with the natural splice site for spliceosomal recognition, the relevant cryptic sites are restricted to those with R_i values comparable to or greater to the corresponding strength of the adjacent natural site of the same polarity²⁵. Additionally, the proximity of potential cryptic sites to the natural site should be considered in assessing whether an exon could be formed with the natural splice site of opposite polarity. Cryptic sites that are considerably weaker than the nearest natural site of the same type, or cryptic sites that would lead to unusually large exons, diminish the likelihood of cryptic site activation. Benaglio *et al.* (2014) used the Shannon Pipeline to screen 303 sequenced patients and flagged five variants that each strengthened or created a different cryptic site⁴⁵. While comparable in strength to the natural site, these were all distant (> 400 nt away) and thus, less likely to be recognized. The authors also stated that

the ΔR_i values for three of these sites were discordant with results obtained with NNSplice, a neural network based splicing prediction program. In fact, both the Shannon Pipeline and NNSplice demonstrated strengthening of these decoy cryptic splice sites.

Shirley *et al.* (2013) evaluated the predictions of the Shannon Pipeline by manually inspecting RNAseq data for each variant with significant information changes in each cell line⁴⁴. However, manual review is unfeasible for many large datasets, especially from tumors, because of the large numbers of potential somatic mutations affecting splicing in each genome. Veridical, an *in silico* method for validation of DNA sequencing variants that alter mRNA splicing, has been developed to provide high throughput, statistically-robust unbiased evaluation based on RNAseq data⁴⁶. The method has been implemented as software for analysis of potential splicing variants from large datasets and catalogues their effects. Veridical takes Shannon Pipeline output from predicted genomic variants with effects on splicing and performs a case-control analysis of corresponding expressed transcripts that cover the same genomic region, taken from normal tissues. Upon Yeo-Johnson transformation of the expressed read count distribution, parametric statistics are used to compare normal and abnormal mRNA species (exon skipping, intron inclusion and cryptic site use). Veridical is designed to be used with large data sets, as the statistical analysis gains power with an increasing numbers of control samples. A recent study of 442 breast cancer tumors from the Cancer Genome Atlas Project revealed 5,206 putative splicing mutations using the Shannon Pipeline. Veridical was then used to confirm exon skipping, leaky or cryptic splicing of 988 of these variants⁴⁷. Veridical is also available through the same interface as the above-mentioned tools (www.mutationforecaster.com).

2.4 Natural sites

The early splice site recognition literature often oversimplified the composition of the U1/U2-type 5' donor and 3' acceptor sites by presenting only consensus sequences and truncating the positions in each site^{16,48,49}. However, the conserved tracts extend well beyond the canonical GT and AG dinucleotides adjacent intron/exon junctions. Furthermore, a small, albeit significant, proportion of natural donor sites (~800, or 0.7%)

contain cytosine at position +2 in the genome. This is reflected by a corresponding small decrease in average information at this position (**Figure 2.1**). Sequences adjacent to these positions are more variable, but are nevertheless essential for the accurate recognition by the spliceosome. Specifically, the donor is defined with the three terminal nucleotides of each exon and the first seven bases of the downstream intron. Conversely, acceptor sites are represented by the first two bases of the exon and the last 26 bases of the upstream intron. Because ASSA and ASSEDA use an integer-based coordinate system, there is a zero coordinate at the first intronic base of each splice site (**Figure 2.1**), which is not used in the conventional numbering system. The coordinate ranges for the donor and acceptor site positions are therefore [-3, +6] and [-25, +2], respectively. Individual information analysis computes the R_i values over these intervals for normal and variant-containing splice sites. As discussed below, information content present in intronic intervals justifies sequencing and analyses of sequences well beyond the locations of the splice junctions themselves.

Certain variants within donor and acceptor sites are tolerated and may even have benign effects, while others have a deleterious impact on spliceosomal recognition. IT accounts for all of these possible outcomes. Unusual donor sites (with cytosine at position +2) are detected by information analysis, but could be falsely called deleterious by consensus sequence-based methods. Although the terminal position of exons contributes significantly to donor splice sites with a preference for G, a significant proportion of sites naturally possess A or U at this position, or less frequently, C.

Of the published IT-based variant analyses, single nucleotide variants (SNVs) that were reported to affect a natural splice site (multi-nucleotide and insertion/deletion variants are listed separately in **Supplementary Table 4**) were compiled and reanalyzed. After reducing this set to only those variants occurring within the intervals covered by the splice site information weight matrices described above, 1,120 SNVs were reported to affect the strengths of either natural donor or acceptor sites. A variant was considered deleterious if it was predicted to affect splicing (either leaky expression or exon skipping), or if it was experimentally shown to reduce or abolish splicing of the

corresponding exon. In instances where prediction and validation did not concur, the latter were used to determine the effect of the variant. Variants predicted to have a neutral effect but demonstrated to be deleterious in the validation study were classified as damaging. In total, 1,036 deleterious natural splice site variants were analyzed (**Supplementary Table 3**).

Sequence conservation has long been considered a surrogate measure of evolutionary constraint and, by inference, functional significance. The average information quantitates the relative conservation at each of the positions within a binding site. We compiled the mutation spectra for all mutations that significantly affected the strengths of donor and acceptor splice sites and compared these with the average information contents at each position. **Figure 2.1b** indicates, at each position of the natural acceptor and donor sites, the frequencies of variants deemed deleterious by information analysis. Interestingly, when the sequence logo is overlaid with the histogram of the corresponding mutation spectra, the relative frequencies of deleterious mutations and the average information are comparable. Indeed, these frequencies and the information contents across each type of site are strongly correlated ($r = 0.95$ for acceptors and 0.90 for donors). Our interpretation is that the susceptibility to deleterious mutation at a position is related to its overall conservation within the splice site, which reflects the contribution of that ribonucleotide to the stability of the interaction with the corresponding spliceosome. Nevertheless, there is an unstated bias in ascertainment in these mutation spectra. Variants occurring at sites with low information and/or that are benign are under represented, as they are less likely to be associated with genetic disease, and were less likely to be reported. Also, the distribution is dependent on the region sequenced by the authors of the reviewed publications; in early work, the full sequence interval containing the entire splice site was sometimes not included or unavailable for analysis.

An interactive website was created to summarize this set of SNVs. This software application renders interpretations of variant effects in a more practical, useful way than the corresponding table of supplemental data (**Supplementary Table 5**). The “Splicing Mutation Calculator” (SMC; <http://splicemc.cytognomix.com>) is a web service that

amalgamates all published results for the same type of substitution in a natural splice site, regardless of genic context. Variants that create cryptic splice sites that do not alter the strength or have a marginal effect on the natural site were excluded. We consider these cases to be sequence-specific as opposed to positional. With this program, users have the option of exploring mutation data (at present, only SNVs can be analyzed) linked to the original literature citations. SMC processes and provides literature support for the variants that occur within the defined regions spanned by natural splice sites. The user first selects the type of site (donor or acceptor), position (based on ASSEDA's integer-based system), wild-type or reference nucleotide, and the alternate substitution at that position (**Figure 2.2a**). The software tool outputs the ΔR_i and the number of variants that have been reported and analyzed to date using IT (**Figure 2.2b**). SMC provisionally classifies the reported variants based on the degree to which these predicted effects are expected to decrease spliceosomal affinity, and consequently splicing. The following criteria are empirically based on affinity changes and a summary of published phenotypes associated with these changes: "Deleterious" (if the site is weakened by more than 7.0 bits), "Probably Deleterious" (if the site is weakened such that $-4.0 \text{ bits} \geq \Delta R_i \geq -7.0 \text{ bits}$), "Leaky" (the site is weakened such that $-1.0 \text{ bits} \geq \Delta R_i \geq -4.0 \text{ bits}$), or "Benign, probable polymorphism" (if the site is weakened by less than 1.0 bits). The "Benign" variants, which are likely polymorphisms, are now catalogued (see **Supplementary Table 6** for the list of variants that were added). It is important to appreciate that the ΔR_i is a constant for a specific nucleotide change at a specific position, though the absolute strength of the splice site depends on the sequence context of the mutation. This context varies between mutations, and $R_{i,initial}$ is not the same for each case, which can result in different $R_{i,final}$ values for different mutations.

Besides published sources, the software also can predict effects of mutations by computing ΔR_i values directly. Particular substitutions that have not been reported in **Supplementary Table 5** can nonetheless be provisionally interpreted. The ΔR_i value is computed and reported from the ribl. While SMC enables rapid exploration of results for validated and novel mutations, it is, however, not a replacement for ASSEDA or the

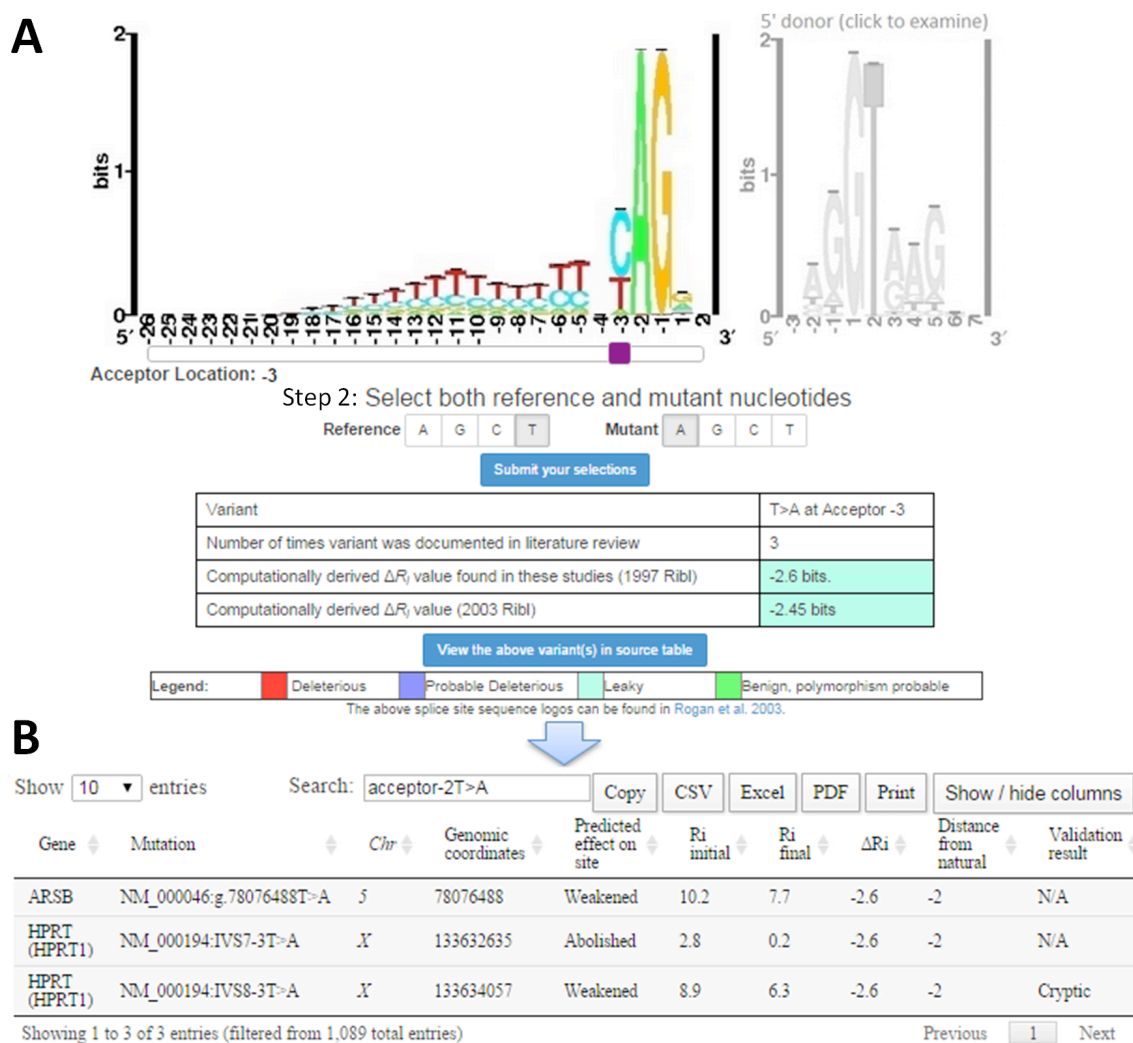


Figure 2.2. Sample retrieval of average change in information content (ΔR_i) with splicing mutation calculator (SMC) for published mutations

A) Example mutation input for SMC (T>A at the 3rd intronic position of natural acceptor). The type of splice site is selected by clicking on the corresponding sequence logo [acceptor (left) or donor (right)]. The purple slider bar appearing below the logo is used to select the position of the mutation. The reference and mutant nucleotides are then designated, and the variant is submitted to the software ('Submit your selection'). SMC outputs a table indicating the user input, the number of instances in the literature where this substitution has been analyzed using IT, and the computed ΔR_i values (in bits) using

both the old (1992; top) and new (2003; bottom) ribls. The cell color for ΔR_i values indicates the predicted severity of the inputted variant according to defined thresholds^{25,50}.

B) Tabular output detailing each instance of the selected mutation from the source table. The user may view, in a separate window, extensive details of all variants referred to in SMC output (**Supplementary Table 5**).

Shannon Pipeline, since it does not consider the sequence context, which can also influence the interpretation of deleterious, leaky, or benign variants.

2.4.1 Minimum splice site information content and exceptions

The minimum theoretical information content of a binding site, $R_{i,min}$, is zero bits, which corresponds to binding at equilibrium ($\Delta G = 0$ kcal/mol)²¹. Comparison of the R_i values of a series of inactivated and minimally active splice sites revealed the minimum strength of functional splice sites ($R_{i,min}$) to be at least 2.4 bits for the original donor and acceptor models of Stephens and Schneider (1992) (based on 103 mutations with functional validation, including 57 natural and 46 cryptic site activating mutations)²⁵. This value was redefined based on information models from a genome-wide set of donor and acceptor models (**Figure 2.1a**) to be 1.6 bits using the identical set of mutations³³. It is likely that the differences between these values are not significant and are attributable to the increased precision of the ribl using the ~50-fold larger set of sites. Weakened natural sites, with significantly reduced R_i values that remain above these thresholds, are considered to be leaky (lower affinity binding), whereas those below this threshold are found to completely abolish natural splice site recognition, resulting in either exon skipping or activation of neighbouring cryptic splice sites. However, these outcomes are not mutually exclusive, since leaky splice site mutations may also result in exon skipping and/or activate neighboring cryptic sites. Natural splice sites below these thresholds are extremely rare, and their recognition is likely enhanced through the binding of specific RNA binding proteins that promote exon definition (eg. *XPC* exon 4 acceptor and *MYBP3* exon 12 acceptor, genes involved in xeroderma pigmentosum and hypertrophic cardiomyopathy, respectively^{51,52}).

Leaky natural sites have R_i values exceeding the $R_{i,min}$ threshold, which, in theory, retain some capacity to be recognized by the spliceosome. There were 84 variants predicted to cause leaky splicing, of which 19 were shown experimentally to lead to exon skipping without any detectable residual natural splicing (**Supplementary Table 2**: #32, 120, 128/380, 195, 276.5, 355, 360, 363, 364, 365, 379, 409, 477/496/934, 573, 842, 853,

883/1589, 886, and 918). Of those, seven are donor splice site mutations at position +5 ($\Delta R_i \sim -3.5$ bits; #128/380, 195, 355, 842, 853, 883/1589, 886), four alter the first exonic nucleotide of a donor site ($\Delta R_i \sim -3.0$ bits; #276.5, 360, 379, 409), and three are donor mutations at position +4 ($\Delta R_i \sim -2.6$ bits; #120, 365, 573). The $R_{i,final}$ values of these 19 inactivated natural sites range from 2.7 bits to 8.8 bits, which suggests the possibility that the variant may also simultaneously affect other adjacent or overlapping sites that preclude recognition of the mutated natural site. Additionally, weakening of 11 of these variants activates a neighbouring cryptic splice site, in which no residual natural splicing was detected. However, changes in splice site preference due to small changes in binding affinity within exons are probably related to the processive nature of donor splice site selection⁵³.

Leaky splicing mutations are readily detected when the expressed transcript contains the causative variant or a neighbouring polymorphism. However, there are a number of practical limitations on the methods for experimental validation of leaky splicing mutations. RT-PCR alone would only be considered reliable for confirmation of homozygous mutations (and in one case, a compound heterozygote where two separate variants abolished natural splicing of the same exon), unless combined with a secondary quantitative methodology⁵⁴. Similarly, it is difficult to assess leaky splicing of heterozygotes using RNAseq data, as reduced levels of wild-type splicing are challenging to determine without adequate read coverage and controls for comparison. However, leaky splicing can be assessed by comparing the frequency of the causative allele to the normal allele in the same cell line when the variant is present within the sequenced reads⁴⁴. These are special cases however, as the variant itself must either be expressed within an exon or, if intronic, must lead to an activation of a cryptic site further into the corresponding intron.

We previously suggested that weaker splice sites are more susceptible to mutational inactivation relative to stronger sites²⁵. In the present study, all experimentally verified variants affecting natural sites (where leaky and abolished splicing could be differentiated) were analyzed (N = 98). Variants predicted to abolish splicing ($R_{i,final} <$

$R_{i,min}$ and/or $\Delta R_i < 7.0$ bits) were filtered out, as large changes in binding affinity will essentially abolish splicing, despite remaining binding strength and regardless of initial R_i value. **Figure 2.3** illustrates the frequency of abolishing by these variants relative to initial R_i value. Variants occurring at weak splice sites ($R_{i,initial} < 4.0$ bits) abolish splicing in 5 of 6 cases (where $\Delta R_i < 7.0$ bits), but are not represented as they all weaken the site below $R_{i,min}$. The remaining variant slightly weakens a site where $R_{i,initial}$ is -0.1 bits (where $\Delta R_i = 0.5$ bits), and its recognition may be supported by SR elements⁵¹. Moderate strength splice sites (5.0-11.0 bits are inactivated in 25-60% of cases), and at strong splice sites ($R_{i,initial} \geq 12.0$ bits) tend to be leaky (**Figure 2.3b**). Mutations that abolish natural sites (without cryptic splice site activation) are expected to result in a complete loss of normal splicing. However, of the 94 variants which reduced natural splice site strength below $R_{i,min}$, 11 were reported to have residual normal splicing activity (**Supplementary Table 2**: #185/750, 275, 881, 914, 1315, 1321, 1325, 1326, 1361, 1380, and 1407)^{25,44,55,56}. Two of these occurred at the G of the +1 position of the donor site (**Supplementary Table 2**: #185/750 and 1326), which is essentially invariant in functional splice sites. This suggests potential problems in IT or experimental analysis of these mutations. Surprisingly, the majority of these variants occur at the +2 position of a donor splice site and are T>G mutations, which are predicted to abolish splicing activity⁴⁴. However, the analysis of RNAseq data for these variants showed no splicing defects (**Supplementary Table 2**: #1315, 1321, 1325, 1361, 1380 and 1407). One explanation is that resultant aberrantly spliced transcripts were subjected to nonsense-mediated decay (NMD) and degraded. Another possibility is that the coverage of these splice junctions is insufficient to distinguish expression of a single allele from that same allele plus the leaky splice junction. The remaining variants differ in the position within the splice site and decrease natural site strengths to between 0.9 to 2.2 bits^{25,55}.

Theoretically, a site lacking the canonical G at +1 (donor) or -1 (acceptor) position of a natural site may exceed $R_{i,min}$. Ozaltin *et al.* (2011) and Di Leo *et al.* (2009) each assessed mutations at positions +1 or -1, which weaken natural splice sites to $R_i > R_{i,min}$, and note that these sites are predicted to be leaky^{57,58}. However, this is not the sole criterion for

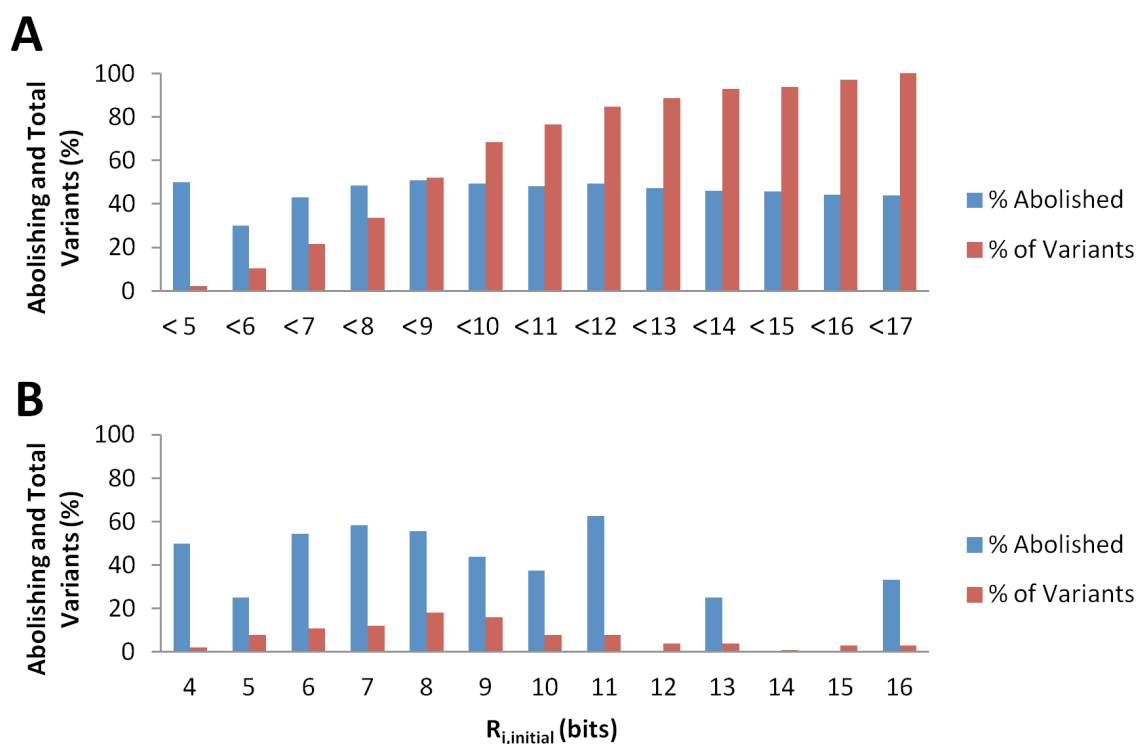


Figure 2.3. Frequency of Abolishing Variants in Relation to Initial Strength

All experimentally verified variants affecting natural sites (where leaky splicing could be assessed) were organized based on the $R_{i,initial}$ of the altered natural site ($N = 98$). Variants predicted to abolish splicing were filtered according to $R_{i,final} < R_{i,min}$; $\Delta R_i < 7$ bits. Histograms display results either cumulatively (**A**) or are binned (**B**). In panel (**A**), the X-axis represents all variants in natural splice sites, which are weaker than the indicated R_i values. In panel (**B**), the X-axis represents $R_{i,initial}$ in 1 bit intervals. The Y-axis shows the percentage of analyzed variants in that interval, as well as the percentage of those variants that were experimentally shown to abolish splicing. Variant *ZRANB3*: g.136148401A>T is an outlier mutation, with a high $R_{i,initial}$ value (20.4 \rightarrow 17.6 bits and results in leaky splicing) that was omitted to facilitate the display of the rest of the mutation distribution.

interpreting splice site mutations using IT-based methods. The overall change in binding affinity must also be considered, as both mutated sites were predicted to have only 0.4-0.5% of binding affinity of the corresponding natural splice sites^{57,58}.

2.4.2 Branch-point mutations

Although branch-point (BPS) recognition occurs independently and post-exon definition, mutations in this sequence have also been described, due to its proximity to the natural acceptor site. Following the recognition of and binding to the 5'ss (upstream donor site) by the U1 snRNP, the U2 is recruited to the 3'ss (downstream acceptor) and recognizes the BPS, resulting in the formation of the pre-spliceosome⁵⁹. Association of U2 with the BPS is essential, as it is the first energy-requiring step, allowing for the tri-snRNP complex of U4/U6 × U5 to be recruited to the BPS, which produces a catalytically active spliceosome⁶⁰. The BPS typically contains a conserved adenosine and a downstream polypyrimidine tract. It is located within 40 nt of the natural 3'ss, however there are reported cases where it can be up to 400 nt away.

Recognition of the BPS is thus a crucial step in proper splicing, and sequence variants can disrupt this event, impede lariat formation, and intron excision. The complete list of BPS variants analyzed using the ASSA and ASSEDA server can be found in **Supplementary Table 7**. The variants range in distance from 0-76 nt from the natural acceptor junction, and either weaken, abolish, or strengthen the BPS. When validation assays were performed, the prediction by the server was correct in 9/11 cases. We deemed the two other cases to be partially discordant (NM_004628.4:c.413-24A>G and NM_005902:IVS8-55A>G). ASSEDA predicted these variants to abolish the BPS, but leaky and normal splicing were observed, respectively. The predictions are partially concordant with experimental findings because ASSEDA also predicted the existence of nearby alternative BPS, which if used, could account for the observed phenotype.

Although IT-based prediction of a variant effects on BPS has been accurate, the number of validated sites used to compute the rib1 is substantially smaller (N = 20), and it is not as reliable as those used to determine R_i values of natural acceptor and donor sites.

Furthermore, these motifs are short and relatively frequent in unspliced mRNA. One possible explanation for the rarity of BPS mutations is that compensatory, alternative BPS sequences can be recognized and used. Furthermore, the weak constraint on the precision of the distance between the BPS and the 3' (acceptor) splice site (**Figure 2.4**) further enables activation of these alternative sites. These factors increase the chance that a variant will be falsely predicted to affect a BPS. For example, variants within donor splice site sequences are routinely predicted to alter strength of false BPS. This error is easily avoidable if the potential recognition sequence is filtered for the genomic context of the variant, as well as its proximity to acceptor splice sites.

2.5 Activation of cryptic splicing

It has been estimated that 1.6% of disease causing missense mutations can affect splicing and recent predictions suggest that approximately 7% of exonic variants in the general population may disrupt splicing, which includes cryptic splicing^{61,62}. The genome is replete with pseudo (or decoy) splice sites with varying degrees of similarity to natural sites that are not recognized in constitutive splicing⁶³. However, mutations that alter the strengths of either these decoys or the natural splice site of the same polarity may shift the balance of isoforms towards non-constitutive splice isoforms that predominate over or eliminate normal mRNAs (**Figure 2.5**). Mutations can create a cryptic splicing event by creating or strengthening a site in either intronic or exonic regions (**Figure 2.5**, Type 1), weaken the natural site while simultaneously altering an overlapping decoy site (**Figure 2.5**, Type 2), or exclusively weaken the natural site, leading to the activation of a pre-existing decoy site (**Figure 2.5**, Type 3). Although the contributions of cryptic splicing to genetic disease have long been recognized, IT analysis correctly predicts most, but not all, cases (**Figure 2.5**). The challenges in identifying potential cryptic sites or determining activation are attributable to our incomplete understanding of the requirements for activation⁶⁴⁻⁶⁶, which include exon length, processivity of donor site recognition, and involvement of splicing regulatory factors. A database of aberrant 3' and 5' splice sites has been compiled⁶⁶.

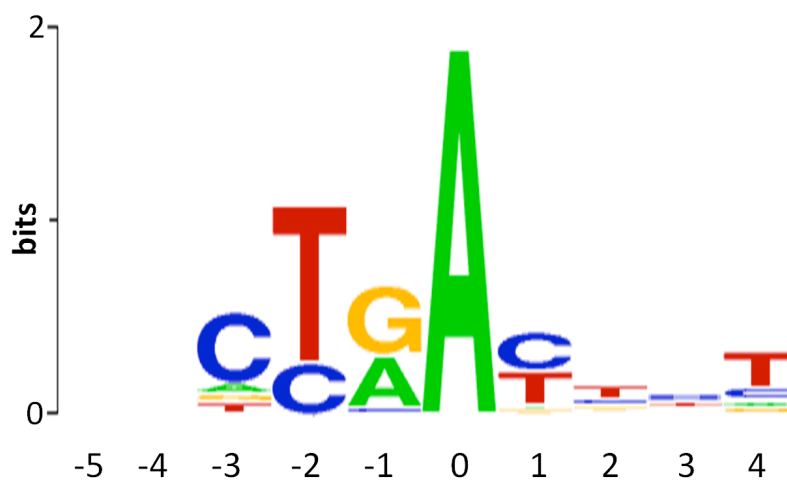


Figure 2.4. Ribli used for the prediction of a variant's effect on branch-point sites

Sequence logo for information model for the branch-point site, created using 20 annotated branch-point sequences.

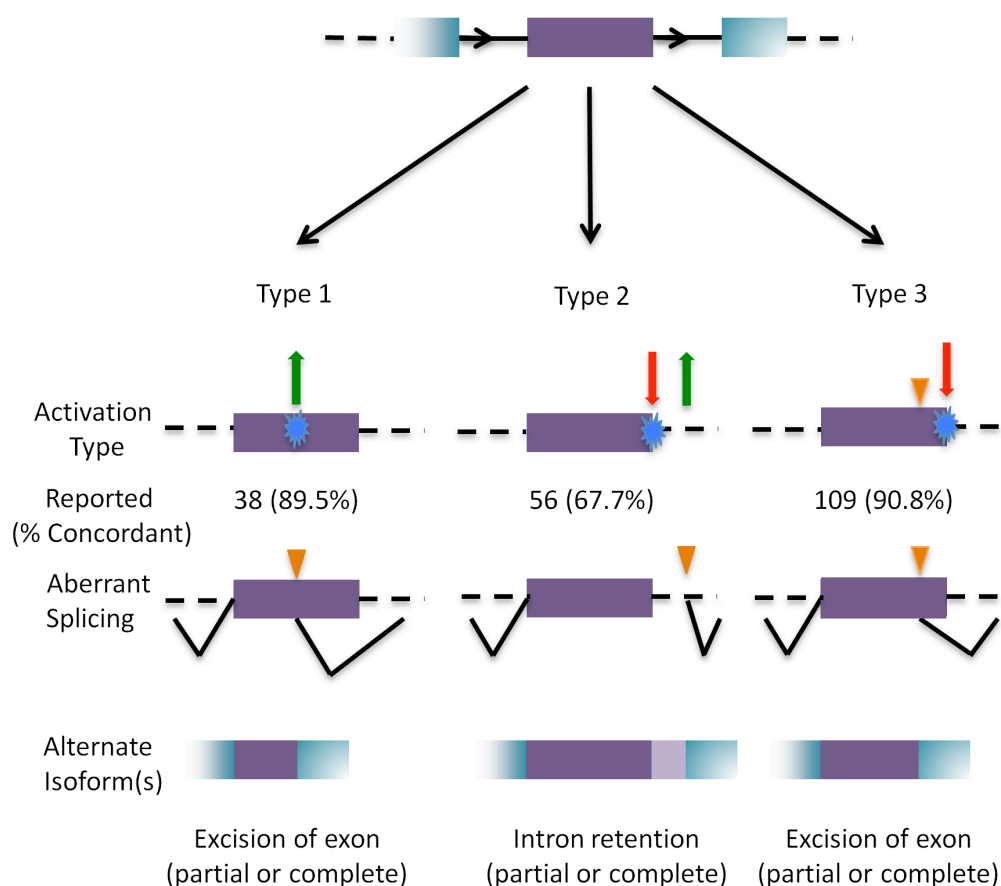


Figure 2.5. Outcomes of cryptic splicing mutations

A prototypical internal exon (in purple) with flanking exons (in blue); introns are represented by black solid, and dashed lines (top). The three types of cryptic splice site activation are then illustrated. Type 1 cryptic splice site activation (left) is caused by the activation (green arrow) of a cryptic site by strengthening a pre-existing, or by creating a novel splice site (blue). Type 2 (middle) results from the simultaneous weakening or abolition (red arrow) of the natural splice site while strengthening or creating (green arrow) a cryptic site. Type 3 (right) involves the activation of a pre-existing cryptic site due to the weakening or abolition of the natural splice site (indicated by orange triangle). The number of cases that have been reported in the literature that has been analyzed by IT for each type is indicated, with the percent accuracy in parentheses. The bottom row represents the resulting mRNA structure due to the activated cryptic splice site.

Another bioinformatic method for cryptic site recognition relies on a training set composed of cryptic sites that are known to be used⁶⁷. There are a number of drawbacks to this approach: the training set is itself not representative of all cryptic sites; and sites that are altered but unused cannot be discriminated from those that are activated (since the latter group also depends on the strength of the corresponding natural splice site). IT-based methods rank cryptic and cognate natural site strength in a way that predicts whether the site will be activated, as well as the abundance of each pair of splice isoforms. Furthermore, the structures of the prospective isoforms are presented by ASSEDA with relative quantitation of each, both prior to and post-mutation.

During our review, we noted 203 variants with experimental support for cryptic splicing (**Supplementary Tables 8-10**). Of these, 38 variants resulted in Type 1 cryptic splicing. From those, site activation (existence of the site and strength ≥ 2.4 bits²⁵) was correctly predicted by ASSEDA in 34 cases (89.5%). We identified 56 variants resulting in Type 2 splicing, 38 of which (67.7%) were accurately predicted, while the remaining 119 variants resulted in Type 3 cryptic splicing and 99 (90.8%) of the alternate splice sites matched predictions.

Prediction of Type 3 cryptic splicing was more accurate than Types 1 or 2. The criteria for concordance with experimental data were that ASSEDA predicted both the cryptic site and that the variant weakened the natural site. However, the strength of a site is not the sole determinant of whether or not a site is activated. Unlike natural sites, novel cryptic sites are not under selection to maintain binding to the spliceosome, and their genomic context is less constrained than natural splice sites. The presence of cooperative splicing enhancer or repressor elements adjacent to cryptic sites, which could influence cryptic splice site activation, is not yet predictable. Additionally, many of the reported activated cryptic sites have been confirmed using non-quantitative approaches, and therefore these may not constitute the predominant splice forms relative to constitutive exons with stronger natural sites. Finally, certain isoforms may not be detected; as aberrant transcripts are often subject to degradation and the tools used evaluate functional splicing consequences do not always have sufficient resolution to distinguish small

differences in isoform structure. All of these factors can affect the concordance of predicted cryptic site activation with experimental validation.

We also separated each sub-group of cryptic splice variants by location (intronic vs. exonic) and computed the average difference in strength between pairs of natural (post-mutation) and the activated cryptic sites. For intronic Type 1 variants, activated cryptic sites were 0.9 ± 5.3 bits stronger than the corresponding natural site ($N = 12$). There were eight Type 1 variants (4 at acceptors and 4 at donors) that were missed, because the $R_{i,final}$ value of the natural site exceeded the strength of the corresponding cryptic site by ≥ 1.0 bits (variants with $\Delta R_i < 1.0$ bits are not reliably detected experimentally). We hypothesize that these cases could be explained by concomitant change in surrounding regulatory binding site sequences. Exonic Type 1 variants were often slightly weaker than their cognate natural sites (-1.1 ± 3.8 bits; $N = 26$). Nearly all of these involved ectopic donor site activation (12 of 13), consistent with a processive mechanism for donor site recognition, which searches downstream from the acceptor splice site to the first donor site of sufficient strength to form an exon³⁸. The opposite pattern was observed with intronic Type 2 cases, in which 20 of 21 exceptions occurred at acceptor sites. On average, the activated cryptic site exceeded the strength of the cognate natural site (1.4 ± 4.6 bits; $N = 57$). Activated, exonic Type 2 acceptor cryptic sites tended to be weaker than their natural site counterparts (-2.3 ± 3.4 bits; $N = 4$). This result may be attributable a low sample size, with 2 of these mutations exhibiting natural sites that were stronger (≥ 1.0 bits) than the corresponding cryptic site (1 donor and 1 acceptor). Finally, Type 3 activated intronic cryptic sites exhibited the greatest difference between the strengths of cryptic sites and cognate natural splice sites (6.3 ± 4.9 bits; $N = 104$). This category contained the fewest number of exceptional cryptic sites, with R_i values less than those of natural sites (5 acceptors and 3 donors). This is consistent with the idea that the intronic cryptic sites are generally not under selection for adjacent functional regulatory binding sites, and, in order to be activated, are required to be substantially stronger than the natural site. Although $R_{i,final}$ values were stronger (2.1 ± 1.9 ; $N = 20$) than the natural site, exonic Type 3 cryptic splice sites did not show as great a difference in strength with a single exceptional case (of an acceptor). Despite these exceptions,

activated cryptic splice sites are generally stronger than the corresponding natural splice sites²⁵.

The distribution of activated cryptic sites relative to their natural splice site is indicated in **Figure 2.6**. Among the reported mutations, donor or acceptor cryptic sites are activated with similar frequencies (113 donors and 108 acceptors). These cryptic sites are located within both introns and exons (59.3% of cryptic donors and 38.0% of acceptors are intronic, the remainder are exonic). Cryptic sites have been confirmed to occur over a broad range of distances from corresponding natural sites, however there is a distinct preference for cryptic site activation adjacent to the acceptor intron-exon junction. These splice sites are most common at the first nucleotide downstream from the natural acceptor splice junction (**Figure 2.1**), which has particular implications for the approaches used to verify the structure of the aberrant transcript (See *Validation methods*).

2.6 Combinatorial effects

While functional natural splice sites and an intact BPS are integral for accurate and efficient splicing, other genetic elements have been shown to make essential contributions to exon definition⁶⁸. Introns will often contain more than one potential splice site recognition sequence, but nevertheless, the correct natural site is consistently selected⁶³. Differences among the strengths of potential sites, as determined by IT analysis, are a major, but not the sole, determinant of splice site utilization. The implication is that additional sequences within the gene are necessary to ensure specificity and precision of exon recognition. Studies of facultatively expressed alternative exon structures have revealed *cis*-acting sequence elements that function to enhance or repress exon recognition. These sequences cooperate with factors that recognize natural splice sites, whose sequences and relative strengths can vary considerably. Depending on their context, these elements have been referred to as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) or intronic splicing silencers (ISSs). In general, these elements serve as binding sites for *trans*-acting elements, which will either promote or impede the spliceosomal

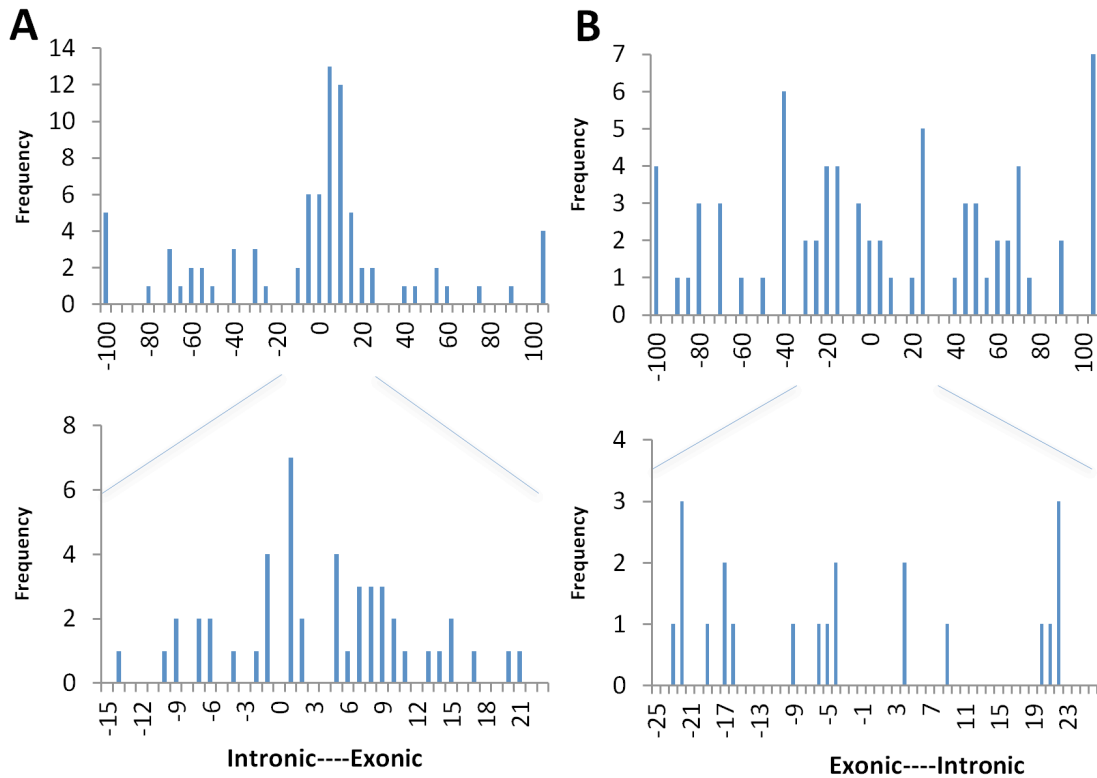


Figure 2.6. Distribution of activated cryptic sites

The frequency of validated cryptic splice acceptors (**A**) and donors (**B**) occurring at positions relative to the natural splice site. Positions are given using ASSEDA coordinates. Lower two panels magnifies the cryptic site distribution of the region circumscribing the natural splice site.

recognition of a splice site. The majority of enhancer elements will act through the recruitment of SR proteins and associate components of the U1 and U2 spliceosomes^{69,70}.

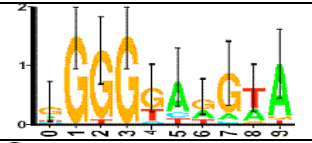
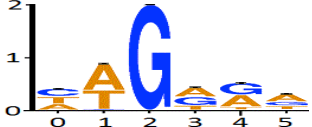
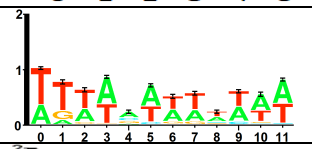
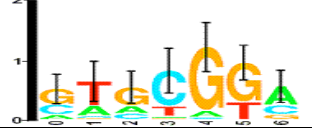



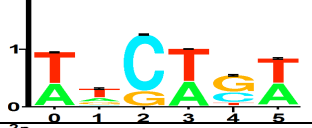
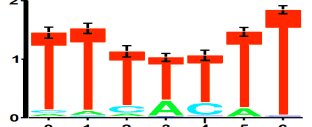
Silencers are often of the hnRNP class, which act through a diversity of mechanisms including steric hindrance, the formation of dysfunctional complexes, or blocking processiveness⁷¹⁻⁷³. To add to the complexity of splicing regulation, it has recently been shown that SR protein function is dependent on context, i.e. whether the corresponding binding site is intronic or exonic^{74,75}.

To improve accuracy of exon definition, the strengths of regulatory elements (i.e. their R_i values) have been incorporated into splicing mutation prediction. The significance of regulatory elements in disease has been demonstrated in many cases. For example, in the *NFI* gene, ESE disruption is the primary cause of exon skipping⁷⁶. Many other genes [Adenomatous polyposis coli (*APC*), survival motor neuron (*SMN*), bestrophin 1 (*BEST1*), pyruvate dehydrogenase alpha 1 (*PDHAI*)]⁷⁷⁻⁸⁰ have been proven to harbour variants that disrupt ESEs and have a confirmed impact on mRNA splicing.

Adding to the complexity, the recognition sequences for these RNA binding factors, while well defined, tend to be short, and can vary to the degree that the same sequence may contain overlapping elements of binding sites for multiple factors. However, this does not necessarily imply that such a sequence is bound with similar affinity by each factor or that it contributes to exon definition. At the same time, these sequences tend to be evolutionarily conserved and may be required for proper splicing^{81,82}.

ASSEDA optionally incorporates PWMs for regulatory binding sites for mutation analysis (**Table 2.1**) in addition to the default donor and acceptor sites. The program selects the most proximate predicted ESE/ISS to the natural splice site when calculating $R_{i,total}$. The molecular phenotype, which dictates the splice isoforms (and their relative abundance) that are predicted, accounts for both the potential effect on the natural site and the most relevant splicing regulatory site. For these regulatory binding sites, a second gap surprisal term specific to the ESE/ISS of interest is applied to the $R_{i,total}$ calculation³⁸. The gap surprisal functions for SF2/ASF and SC35 have been previously described³⁸,

Table 2.1. Splicing regulatory protein binding sites ASSEDA scans for and their associated effect on splicing.

Splicing Factor	$R_{sequence}$ (bits)	Sequence Logo	Location-dependent effect on splicing	
			Intronic	Exonic
hnRNPH1	8.9 ± 1.8		$E^{83,84} / S^{85,86}$	<u>S</u> / E^{87}
hnRNPA1 ⁱ	4.6 ± 1.5		<u>S</u> / E^{88}	<u>S</u>
TIA1	7.6 ± 3.1		<u>E</u>	N/A
SRSF6 (SRp55)	5.2 ± 1.4		E / S^{86}	<u>E</u> / S
SRSF5 (SRp40)	4.5 ± 1.5		E / S^{86}	<u>E</u> / S^{89}
SRSF2 (SC35)	4.5 ± 1.6		E / S^{90}	<u>E</u> / S^{91}
SRSF1 (SF2/ASF) ⁹²	5.8 ± 1.5		$E / S^{90,93,94}$	<u>E</u> / S
PTB ⁱⁱ	4.9 ± 1.9		<u>S</u> / E^{95}	<u>S</u>
ELAV1	9.6 ± 3.4		$S / E / N^{96,97}$	<u>S</u>

Reported dominant effect is bolded. E – Enhancer; S – Silencer; N - Neutral.

ⁱEnhancer activity by hnRNP A1 occurs at the junction⁸⁸. ⁱⁱPTB does not directly enhance splicing, but can do so indirectly by preventing the binding of splicing repressors⁹⁵.

where the most common distance of the ESE/ISS is within 10nt of the natural site. The gap surprisal penalty gradually increases with distance from the natural site. Gap surprisal distributions for ELAVL1, TIA1, and SRp55 show a similar pattern, while hnRNPA1 and PTB binding sites are strongly clustered around splice junctions. It should be feasible to include the contributions of multiple splicing regulatory binding sites of the same or different RNA binding proteins in determining $R_{i,total}$; however this capability had not yet been implemented. Currently, if multiple sites of the same type are altered, the strongest (before or after mutation) is chosen by ASSEDA software.

Although the disruption of splicing regulatory sequences can cause aberrant splicing, the interpretation of variants affecting these sites is not as straightforward. Due to their degenerate nature, short sequence, and a lack of understanding for the context of their use, altered regulatory sites should be functionally validated before being deemed pathogenic¹⁰. Using variants from a number of different studies, ASSEDA accurately predicted experimentally determined changes in binding at a splicing regulatory site 75% of the time ($N = 12$)³⁸. However, there were instances where regulatory sequences had been analyzed by IT, and considered to contribute to disease, but the results were not reproducible. For example, Kölsch *et al.* (2009) described SNPs associated with Alzheimer's Disease, one of which strengthened and created SRp40 and SRp55 sites, respectively, but were reported by authors to be abolished⁹⁸. This study did not report any evidence to support the significance of these predictions.

Functional validation of the effects of these mutations could contribute to understanding of the roles of these factors in regulating constitutive splicing. Similarly, there is still little understanding on how multiple regulatory binding sites within the same region function as a unit. Using a pull-down assay, Olsen *et al.* (2014) demonstrated how different variants affect the binding of multiple regulatory proteins²⁶. One mutation was predicted to create and strengthen multiple hnRNPA1 sites and slightly strengthen an SF2/ASF (SRSF1) site. The pull-down studies showed up-regulation of hnRNPA1 binding and a decrease in SF2/ASF binding. However, SF2/ASF binding increased when a mutation

disrupting hnRNPA1 affinity was introduced, suggesting that the strong hnRNPA1 sites outcompete the weaker SF2/ASF site.

In some instances, an alteration in regulatory splice site recognition sequence and natural splice strength were altered concomitantly, with both predicted to have similar effects splicing. Alteration of a regulatory sequence can sometimes provide a plausible explanation for discordant *in silico* prediction and experimental validation. As an example, Smaoui *et al.* (2004) analyzed a donor mutation (NM_001040667:c.1327+4A>G) in *HSF4* in a family with congenital cataracts⁵⁴. This variant was predicted to cause leaky splicing ($R_{i,final} = 5.4$ bits; $\Delta R_i = -2.6$ bits; 67.5% residual binding), however RT-PCR showed complete exon skipping. Our further analysis showed that it is predicted to also create an overlapping hnRNPA1 site ($R_{i,final} = 4.2$ bits; $\Delta R_i = 17.1$ bits). Another case involved a mutation in the *XPC* gene (NM_004628:c.2033+2T>G) that created a novel intronic cryptic site 4 nt downstream of a natural donor site⁹⁹. However, a weaker site 68 nt downstream from the natural site was activated. A possible explanation could be that activation of the cryptic site is influenced by a neighbouring hnRNPA1 site that is itself strengthened ($R_{i,final} = 5.2$ bits; $\Delta R_i = 2.2$ bits) and an SRp55 site that is significantly weakened ($R_{i,final} = 1.9$ bits; $\Delta R_i = -4.0$ bits).

The effects of changes in regulatory binding site strengths may ascribe potential functions to previous VUS. For example, Maruszak *et al.* (2009) present a *PINI* variant associated with late-onset Alzheimer's Disease (NM_006221:c.58+64C>T)¹⁰⁰. Based on IT, it is expected to abolish an intronic SC35 site, which could have either an enhancing or silencing effect (**Table 2.1**). A 2.82-fold decrease in transcript levels was demonstrated, which is concordant with previous findings reporting decreased *PINI* levels in the brains of Alzheimer's Disease. Another study described an exonic missense variant within the *ETFDH* gene in a patient with multiple acyl-CoA dehydrogenase deficiency (NM_004453:c.158A>G) that showed evidence of exon skipping. The variant was predicted to be “benign” or “tolerated” when evaluated with PolyPhen and SIFT²⁶. ASSEDA, on the other hand, predicted the creation of an hnRNPA1 site ($R_{i,final} = 5.9$ bits; $\Delta R_i = 17.1$ bits), a slightly strengthened hnRNPH site ($R_{i,final} = 4.0$ bits; $\Delta R_i = 0.2$ bits),

the abolition of an SRp40 site ($R_{i,final} = -3.3$ bits; $\Delta R_i = -6.3$ bits) and two novel, weak SF2/ASF sites ($R_{i,final} = -4.6$ bits; $\Delta R_i = 0.8$ bits and $R_{i,final} = -2.4$ bits; $\Delta R_i = 0.4$ bits)²⁶. The natural donor site was unaltered by the mutation. As indicated earlier, the mutation was confirmed experimentally to increase hnRNPH and hnRNPA1 and decrease SRp40 and SF2/ASF binding.

2.7 Validation of results

A number of early mutation studies did not perform expression analysis and relied solely on the ASSEDA or ASSA server to interpret potential mutations. This is not recommended, as there are limitations to any *in silico* predictive method, which impacts accuracy and precision of the prediction. Assuming that the impact of the mutation on expression can be detected, experimental validation of IT-based mutation analysis can reveal its limitations. We describe the various validation methods that were employed in the articles where expression data were available. Below, advantages and disadvantages of these approaches are explored, as well as how lower sensitivity validation can result in misinterpretation. Finally, we determine the accuracy of IT-based prediction, and point out some instructive, discordant cases.

2.7.1 Validation methods

The two most widely used methods for validating mutant mRNA splicing isoforms have been RT-PCR analysis of patient mRNA, and transfection of minigene constructs expressing the mutated exon into cell lines, followed by RT-PCR. These assays were, in some cases, accompanied by other techniques such as direct sequencing of cDNA, Western blotting, luciferase expression assays, or immunostaining. A number of studies used quantitative RT-PCR or real-time PCR to estimate isoform abundance. RNA or cDNA sequencing and exon expression microarrays were also used in several studies to support *in silico* predictions. Certain functional assays that we reviewed were unique to a single study, including: allelic instability, exon trapping, immunoprecipitation of splicing factors, and flow cytometry^{26,101–103}. Other indirect methods of justifying the association between a splice site variant and disease included funduscopy, loss of heterozygosity,

blood protein levels, and segregation with disease¹⁰⁴⁻¹⁰⁷. Because a variant may result in aberrant splicing but might not be accompanied by a detectable phenotypic change, we excluded the results of indirect assays of phenotype. Indirect measures of phenotype can support disease association, but do not inform about accuracy of splicing prediction.

Endpoint RT-PCR and minigene assays probe the specific variant in question, but do not reveal relative abundance of each isoform, whereas qPCR does. Neither method resolves mRNA sequence at the nucleotide level, which can fail to confirm predicted splicing mutations, especially in instances where a small number of nucleotides are retained at the constitutive splice junction¹⁰⁸. The resultant frameshifted mRNAs can cause premature truncation of the transcript (PTC), instability, and NMD, leaving no evidence of the mutated isoform (unless the cells had been treated with an NMD inhibitor). A disadvantage is that in cases where the protein is not degraded, but still impaired or dysfunctional, the result will be incorrectly categorized as benign. For example, Wessagowit *et al.* (2005) used sequencing of a *COL7A1* variant (NM_000094:c.341G>T), responsible for recessive dystrophic epidermolysis bullosa, to demonstrate a 87 nt deletion in the cDNA¹⁰⁹. The authors also performed immunostaining of the corresponding protein with a monoclonal antibody, which showed no difference between wild-type and mutant samples because the epitope was not disrupted by the deletion. Had the authors only performed the binding assay, the variant would have likely been disregarded. NMD can be a predominant cause of false-negative results when validating splice variants. When aberrant splicing causes a frameshift and PTC, translation of truncated proteins is prevented, which otherwise can have dominant negative effects or exhibit gain-of-function¹¹⁰. However, if these transcripts are degraded and only the normal allele is detectable (in the case of a heterozygote or leaky splicing), then the splicing prediction will not be supported. Interestingly, Khan *et al.* (2004) were able to show that NMD had occurred by comparing levels of total message (qPCR) between wild-type and mutant samples⁵¹. Experimental methods have been developed to stabilize transcripts with premature termination of translation, thus circumventing NMD. The use of emetine, which inhibits translation and stabilizes RNA transcripts, can increase the relative amount of aberrant transcript observed^{111,112}. However this approach

can induce a stress response within the cell and further transcription must be halted using actinomycin D. This combination was used by Bloethner *et al.* (2008) in an approach called Gene Identification by NMD Inhibition¹¹³. Similarly, the use of puromycin and cycloheximide were shown to inhibit NMD and restore predicted aberrant splice forms^{101,114}. Furthermore, certain mutations proximate to the carboxy-terminus of the coding region evade NMD^{115,116}.

2.7.2 Regulatory sequence variants

A number of assays have been developed to confirm direct effects of variants on splice site recognition, however fewer methods are available to measure effects of mutations at binding sites of splicing regulatory proteins¹¹⁷. The most reliable approach is to associate a change in splicing with a change in regulatory protein binding. A combination of electrophoretic mobility shift assay and RT-PCR were used to confirm that a predicted change in an SF2/ASF binding site caused exon skipping in the *CFTR* gene responsible for cystic fibrosis¹¹⁸. Others performed RNA affinity purification in combination with Western blotting²⁶.

Another approach tests multiple variants at the same position through minigene assays. Anczuków *et al.* (2008) observed that two variants at the same position in the highly penetrant hereditary breast/ovarian cancer gene *BRCA1* (c.3600G>T and c.3600G>C) predicted different effects on regulatory sequences, as well as different observed effects on splicing¹¹⁹. The G>T variant predicted abolishment of a SRp40 site and weakening of an SF2/ASF site by both ASSA and ESEfinder, and showed a significant reduction in the relative amount of normal transcript. The G>C variant, which did not elicit a change in splicing, was not predicted by ASSA to have a significant effect on either site (although ESEfinder predicted weakening of the SRp40 site below its default threshold). The difference in splicing efficiency could be due to the loss of binding by one or both of these regulatory proteins. This assay associates predicted changes to regulatory protein binding site strength to changes in splicing. A direct binding assay would lend key support for such predictions.

2.8 Accuracy of IT-based prediction

We previously evaluated the accuracy of IT-based prediction using a set of validated splicing mutations (85.2%; N = 61)²⁸. Other studies have also evaluated the accuracy of ASSA/ASSEDA while evaluating differences between multiple predictive programs and have shown varying levels of concordance (68.8%, N = 16; 90.1%, N = 22; 100%, N = 24)^{55,108,120}. With a comprehensive list of all published variants analyzed using IT-based methods (**Appendix D**), we perform a meta-analysis of all of these variants to minimize bias in interpretation and impact of ascertainment of specific phenotypes from individual studies. The list of variants is more extensive than any previous study examining accuracy of IT-based methods. The variants are not restricted to a single or even group of diseases, but rather cover over 150 different conditions (**Supplementary Table 2**).

In total, 905 variants were reported in 122 different publications to have been validated for their effect on splicing (1,727 total variants analyzed from 216 papers – **Supplementary Table 11**). In all cases, the authors performed information analysis; however, the validation experiments were sometimes contained in the original reports and in other cases, later studies. In a minority of mutations, the validation results were either uninformative (N = 36) or the methods did not directly imply an effect on splicing (N = 2); these mutations were therefore excluded in determining the accuracy of predictions (shaded in grey in **Supplementary Table 11**).

More specifically, in order for experimental results and predictions to be considered concordant, one or more of the following criteria had to be met:

- a. A variant predicted to abolish a splice site did abolish splicing, with no residual splicing observed. Exceptions to this were assays in which both the mutant and wild-type alleles were expressed in the same cell line or patient sample, and could not be discriminated from one another (i.e. RT-PCR);
- b. A variant predicted to be leaky exhibited residual normal splicing, with the exception of cases where a much stronger cryptic splice site was activated;

- c. A variant that strengthened the natural site and showed normal or increased levels of the wild-type isoform, consistent with it having a benign phenotype and/or being polymorphic;
- d. A variant predicted to activate a pre-existing splice site, while also reducing the natural splice site strength, was demonstrated experimentally to result in cryptic splicing, regardless of whether it was predicted it to be the predominant isoform;
- e. A variant predicted to affect a splicing regulatory protein-binding site was consistent with validation experiments explicitly assessing binding affinity and associated splicing alterations.

Cumulatively, 87.9% of variants documented by expression studies (762 of 867) that satisfied these criteria were accurately predicted by ASSEDA. A minority of papers reported variants to be “partially concordant” (3.1%; 27/867), meaning that while the cryptic site observed was predicted, it was not the most likely splice isoform relative to other expressed cryptic exons. Because this method of scoring met our criteria (see point d above), we included these in our determination.

2.8.1 Predicted mutations discordant with validation results

Limitations of both the predictive model and the validation data/methods were the primary reasons for discordance. Where information analysis predicted a neutral change or no effect, but validation showed aberrant splicing, we hypothesize that there are either unrecognized splicing regulatory protein binding sites that are weakened or abolished, or that there are underlying mechanisms that are not currently addressed by current information models^{26,38,54–56,58,100,102,103,103,118,121–131}. The validation methods used can also contribute to discordant results. We note that 41 discordant results originated from one of our own studies⁴⁴. This study used RNAseq data to validate predictions, a genome-wide approach that should be used with caution when inferring changes resulting from potential splicing mutations. Until this study was published, IT-based mutation analysis was based on single or candidate disease gene studies. RNAseq reveals all changes in transcript levels for all genes, which although potentially relevant to splicing, may not

necessarily contribute to the phenotype in question. This leads to the possibility, especially in cancer phenotypes, of bystander effects (global splicing dysregulation, natural alternative splicing) that are not directly attributable to the predicted mutations. Furthermore, because the sequence reads at splice junctions are short, and often limited in number, a relevant splicing aberration may result from a given variant, but it was not detectable. Finally, the predictions of IT can pick up variants that should alter splicing for example, of rare recessive alleles, that that may not have any disease relevance.

2.9 Misinterpretation of variant effects

While preparing this review, several variants misinterpreted with IT-based tools were noted. These variants have been re-analyzed to disseminate the correct findings and to avoid making similar errors in the analysis of newly discovered variants. **Supplementary Table 2** contains these results. The most common problems result from unfounded emphases on altered or pre-existing cryptic sites that are determined to be significantly weaker relative to the cognate natural site^{113,132–136}, and from selectively reporting a single change in the R_i value when, in fact, multiple significant changes can be detected^{52,132,137–140}. An example of the first type of error is exemplified by a variant in *CGI-58 (ABDH5)* in a patient with Dorfman-Chanarin syndrome, where the natural splice site is 9.1 bits (or ≥ 549 -fold) stronger than the reported cryptic site¹³³. Henneman *et al.* (2008) selectively reported the effect of a mutation that weakens a natural donor splice site in *APOA5* and is thought to cause hypertriglyceridemia, however only a change in the information content of an SC35 binding site was indicated¹⁴⁰.

Other common problems include incorrect declaration of small ΔR_i values as significant changes^{113,141,142}, use of incorrect $R_{i,min}$ values^{143,144}, and the computation of predicted binding strength changes on a linear scale¹⁴⁵ rather than the correct exponential function (i.e. $\leq 2^{\Delta R_i}$)²¹. Smaoui *et al.* (2004) described an 8.0 bit donor site as weak, which is actually equivalent in strength to $R_{sequence}$, the average strength³⁶. Allikmets *et al.* (1998) and Ozaltin *et al.* (2011) both described an inactivating mutation as leaky, because the weakened site remained above the $R_{i,min}$ ^{57,138}. However, the variant mutation produces a site with $< 0.7\%$ of its original binding affinity, which would substantially reduce exon

recognition and lead to exon skipping¹¹⁶. Also, cryptic sites created in the promoter regions of genes should not be considered to be splicing mutations¹⁴⁶. Variants that are predicted to create a cryptic site upstream or overlapping a natural site of the opposite polarity (i.e. cryptic donor upstream of a natural acceptor) have been reported^{135,136,147}, which would be inconsistent with established splicing mechanisms³⁸. A rare exception that could render such a site active is to the creation of a cryptic exon that occurs in conjunction with a proximate, correctly oriented, pre-existing cryptic splice site of opposite polarity^{25,33}. Insufficient numbers of examples of mutations creating cryptic exons have been reported to date for ASSEDA to accurately model predicted cryptically spliced exons by default.

Several results were generated by incorrect entry of mutations to ASSA/ASSEDA. For example, altered cryptic splice sites have been confused with natural sites^{52,141,148,149}. Additionally, ‘residual binding strength’ displayed has been misinterpreted as a percent decrease^{148,150}. Strong, pre-existing cryptic sites outside of the default sequence analysis window (54 nt circumscribing the mutation) have also been missed because the window was not expanded to include these sites¹⁵¹. Although the predicted isoform structure generated by ASSEDA will, by default, display skipping for mutated natural sites with $\Delta R_i \geq -7$ bits (or ≥ 128 -fold)¹¹⁶, smaller decreases in natural site strength of an internal exon can partially induce exon skipping. This value is adjustable, and it may be advisable to explore different thresholds depending on the particular susceptibility of a splice junction to exon skipping. Sharma *et al.* (2014) used the default threshold from ASSEDA to interpret *CFTR* mutations c.2988G>A (9.6 to 6.6 bits, natural donor site of exon 18) and c.2657+5G>A (9.1 to 5.6 bits, natural donor site of exon 16), but exon skipping was documented⁵⁵. IT analysis was not discordant for these variants, which significantly weaken the corresponding splice sites by ≥ 8 - and 11-fold, respectively, and has been shown in other genes to lead to exon skipping, leaky splicing, or both of these outcomes. Aissat *et al.* (2013) tabulated variants that were predicted to affect strengths of ESE binding sites, but did not comprehensively report all findings even though predictions by ASSA and ESEfinder were concordant (eg. *CFTR*: c.1694A>G). Alternate mutation entry methods, which do not use contextual gene name annotations, such as entry by rsID,

report predicted binding changes on both strands. And in a study of hereditary Alzheimer's disease, the abolition of SRp40 binding sites on the antisense strand was confused with binding sites for *CYP46A1*, which is transcribed from the sense strand¹⁵².

Other problems include inadvertent mislabelling of splice site type or location^{153–155}, interchange of the terms information content and change in information (R_i and ΔR_i)¹²⁶, and unclear variant interpretation (i.e. “run on into the intron”)¹⁵⁶. Moriwaki *et al.* (2009) claim ASSA did not predict a mutated natural donor site, but in fact, the site was present in our reanalysis¹⁵⁷. Published R_i values from Rogan *et al.* (1998) and von Kodolitsch *et al.* (1999) are in some instances different from current values due to updates of the reference genome sequence^{25,50}. Nevertheless, the overall predicted effect did not change, but initial and final R_i values were inconsistent. Interpretations of certain mutations could not be reproduced in some instances^{107,149,158–160}. Finally, we noted that ASSEDA can sometimes improperly parse indels entered using c. or IVS notation. Such errors have led to published false results^{71,120,161,162}.

2.10 Interpretation of published variants in studies that use information analysis

2.10.1 Genotype-phenotype association

The severity of phenotype by due to splicing mutations can be related to their effects on mRNA splicing, after careful consideration of the overall impact on mRNA levels and protein coding¹⁶³. Significant information changes (where $\Delta R_i \geq 7$ bits or where $R_i \leq 2.4$ bits) of splicing variants in hemophilia patients (*F8C* and *F9*) were shown to correspond to the severe clinical phenotypes of the disease (reduced protein activity, increased clotting time, bleeding frequency)¹³¹. The overall effect on the coding potential of the mutated transcript should be considered, as skipping events that maintain the reading frame commonly lead to milder phenotypes^{106,164,165}. Nevertheless, two variants that abolish splice site recognition in *PTPRO* in Idiopathic Nephritic Syndrome reported by Ozaltin *et al.* (2011) had similar phenotypes even though one retained the reading frame and the other caused a frameshift⁵⁷. The exon deleted by the in-frame skipping event is highly conserved⁵⁷. Exon skipping events that cause frameshifts close to the carboxy-

terminus may lead to mild phenotypes, as they avoid NMD^{116,166}. Dominant negative mutations with either $R_i > R_{i,min}$ or with modest decreases in ΔR_i , may be less likely to cause severe phenotypes, as a residual amount of the natural isoform continues to be expressed^{107,121,145,167–170}. The impact of cryptic site-activating variants on phenotype can be similarly assessed. Activated cryptic sites that shift the reading frame have been shown to be more severe clinically relative to those which maintain the same reading frame as the native gene^{106,109,171,172}.

IT-based tools exhibit high specificity for analysis of splicing neutral variants in breast/ovarian cancer and other disorders¹²⁰. These predictions can reduce the requirement for experimental validation of low-priority candidate mutations with minimal changes in information content^{17,25}. IT analysis has been used in numerous studies to infer neutral effects of variants^{17,37,101,113,120,123,132,133,155,160,161,173–187}. Similarly, variants that strengthen natural splice sites^{188–190} are also likely to be neutral, though these variants can increase retention of exons that are otherwise frequently alternatively spliced^{191,192}. However, binding site variants with minimal splicing information changes may still alter mRNA processing by disrupting mRNA secondary structure¹⁹³.

2.11 Polymorphisms and splicing

Early studies suggested that common polymorphic sequence variations (SNPs) at splice sites corresponded to small ΔR_i values, consistent with these changes having little impact on mRNA abundance²⁵. More recently, it has been appreciated that certain rare SNPs have significant genetic loads, can actively alter mRNA splicing profiles, and lead to non-obvious splicing phenotypes^{62,191}. Nevertheless, it is not uncommon for reports to solely analyze novel variants and ignore known SNPs^{140,160,162,194}, or limit results only to those that occur in the vicinity of natural splice sites¹⁸⁶. We find that 56.4% of common SNPs (with population frequencies $\geq 1\%$ in **Supplementary Table 2**) within natural sites significantly alter strength [12.8% abolish and 28.2% cause leaky splicing, 15.4% modestly strengthen sites ($\Delta R_i < 2.6$ bits)], and 43.6% have insignificant ΔR_i values, as expected (N = 39). The mean $R_{i,final}$ and ΔR_i values, for these natural sites are 7.9 ± 4.0 bits and -1.4 ± 3.0 bits, respectively, which suggests the effects of these polymorphisms

on splicing are nil to limited. However, polymorphisms can significantly modulate splicing, as some common SNPs are predicted to abolish natural splicing (**Supplementary Table 2**: #1291, 1296, 1431, 1435, and 1436). These include rs10190751 in *CFLAR*, which modulates the production of two short isoforms, and is associated with an increased risk of lymphoma^{191,195}, rs3892097, which alters exon inclusion in *CYP2D6*³³ and leads to a non-functional protein and altered drug metabolism¹⁹⁶, and rs1805377 in *XRCC4*³⁷, which has been associated with oral cancer susceptibility¹⁹⁷ and increased risk of gliomas¹⁹⁸. There is also experimental support for common SNPs that have been predicted to affect splicing^{25,102,111,114,118,122,153,167,191,199}. For example, experimental evidence for increased exon inclusion has been described for three of six SNPs that increase strength of natural splice sites^{191,192}. Numerous common SNPs, that were either deemed neutral or predicted to affect splicing, have not been confirmed experimentally^{17,25,28,37,50,56,98,103,135,137,148,149,152,155,168,169,181,185,187,190,200–210}.

Polymorphisms with significant information changes should be investigated, as they may not be completely benign and can have a significant impact on mRNA splicing.

2.12 Inference of variant pathogenicity by IT analysis

Recently, American College of Medical Genetics and Genomics recommendations for reporting incidental findings in sequencing have suggested that bioinformatic predictions are not sufficient to declare significance²¹¹. Preceding the publication of these guidelines, numerous peer-reviewed articles suggested variants analyzed by IT to be causative/pathogenic/disease-causing, without confirmation of the predicted splicing effect^{105,139,141,154,164,170,180,207,212–220}. Other authors have qualified the interpretation of bioinformatic results with less certain terms (i.e. ‘suggest’ and ‘likely’ pathogenic)^{114,116,178,221–225}. Leclerc *et al.* (2002) state that a predicted variant confirmed to affect splicing is likely deleterious, but could not be unequivocally shown to cause the observed phenotype¹⁶⁹. Although IT predictions can relate a sequence change to the resultant phenotype, caution should be exercised when deeming a predicted splicing variant as pathogenic in the absence of other functional evidence. The high level of concordance between IT mutation analysis and experimental findings indicates that this

approach, in conjunction with other evidence, can be used to detect splicing effects, which may be used to explain disease phenotypes.

2.13 Comparisons to other software programs

There are now over a dozen other publically available splicing prediction tools, some using strategies similar [MaxEntScan (MES)] and others, which are quite different (NNsplice) that are compared with IT^{226,227}. Vreeswijk *et al.* (2009) assessed the applicability of different splice prediction programs to diagnose *BRCAl/2* variants¹⁵¹. These authors recommended that the outcome of 3 programs was sufficient for analysis, unless all three predictions were discordant from one another (2 for false positive predictions). Despite the obvious appeal of consensus between different analytical methods, a major caveat in using polling strategies for mutation assessment is that these approaches are prone to both systematic and sampling errors⁴³.

We summarize results of 36 publications that used both IT-based prediction tools and one or more alternate prediction tool (14 for 5' and 3' splicing, six for splicing regulatory proteins) to assess mutations^{26,42,101,103,107,115,118-121,127,134,136,145,151,160,162,168,169,173,181,187,191,199,212,220,228-237}. The analysis performed by the authors allowed us to compare the similarity of predictions to those programs and IT in **Table 2.2a** and **Table 2.2b**. Those most commonly used for 5' and 3' splice sites (NNsplice, MES, NG2, HSF and SSF) were highly concordant for natural sites (85.4% for donor and 77.6% for acceptor sites; **Table 2.2a**). Discordance of acceptor predictions may be due to methodologies that do not analyze the complete acceptor site (HSF analyzes only 14 intronic nucleotides upstream of acceptor splice sites)²³⁸. Some programs (SSF, HSF) exhibit greater concordance with IT for cryptic splice site prediction (96% for donor and 76.9% for acceptor sites). The level of discordance between IT and other commonly used software programs (59.5% for donor and 60% for acceptor sites) may be attributable to the empirically-derived scoring thresholds and the validation sets used to predict mutated splice sites. Models that are typically built (or trained) using known natural splice sites may be less sensitive for differentiating true cryptic splice forms from decoys in the genome, which tend to be weaker than natural

Table 2.2a. Concordance of splice-prediction programs to information theory-based tools for natural and cryptic sites

	MES ¹	BDGP ¹	NG2 ¹	HSF	SSF ^{1,2}	SSqF ¹	GS	SV	SP	SS	GenS	ASD	GeneS	GM
Nat. Donors	42/48	37/39	24/32	23/28	25/27	15/18	6/11	9/9	5/8	2/2	1/2	1/1	1/1	-
Nat. Acc.	21/26	14/19	14/20	12/16	15/18	9/11	3/5	4/5	3/5	-	-	-	-	-
Cryp. Donors	16/24	4/8	5/10	16/17	8/8	0/7	2/2	-	-	-	0/1	0/1	-	-
Cryp. Acc.	7/13	2/3	3/4	8/11	2/2	2/2	-	-	-	-	-	-	-	0/1
Neut. Mut.	31/31	8/8	4/4	26/26	-	-	-	-	-	2/2	-	-	-	-

Table 2.2b. Concordance of splice-prediction programs to information theory-based tools for splicing regulatory proteins

	ESEfinder ^{3,4}	Rescue-ESE	Ex Skip ^{3,4}	ESEsearch	PESX
ESEs (all types)	9/15	3/4	4/14	2/3	1/1
Neut. Mut.	4/4	1/1	3/3	-	-

Concordance was assessed from the analysis of variants from 36 publications that used IT-based tools and a secondary predictive method. Each value corresponds to the number of variants that were concordant with IT-based tools versus the total number of variants for each category. ¹ – includes Vreeswijk *et al.* (2008), which may not have properly reported predicted cryptic sites, as they did not report any cryptic sites predicted by ASSA beyond the default window size (54 nt) from the mutation. ² - predictions made using the SSF-like algorithm in the Alamut splicing prediction module were combined with the SSF category (SSF is no longer supported). ³ – Aissat *et al.* (2013) contributes highly to the discordance of these programs, and may be due to improper reporting/analysis. ⁴ – Mutations predicted by alternate program to affect SR protein to which ASSEDA has no model (i.e. 9G8) were not included in statistics.

MES – MaxEntScan; BDGP – Splice Site Prediction by Neural Network, NNSplice; NG2 – NetGene2; HSF – Human Splice Finder; SSF – Splice Site Finder; SSqF – Splicing Sequences Finder; GS - GeneSplicer; SV – SpliceView; SP – Splice Predictor; SS - Shapiro-Senapathy; GenS – GenScan; ASD - ASD-Intron analysis; GeneS – GeneScan; GM – GeneMark; PESX - Putative Exonic Splicing Enhancers/Silencers.

splice sites. Tools are highly consistent when analyzing variants expected to be neutral with respect to splicing (100%; N = 71). Colombo *et al.* (2013) compared nine programs to evaluate accuracy in predicting mRNA splicing effects and reported that ASSA, along with HSF, demonstrated 100% informativeness and specificity¹²⁰.

ASSEDA has also been used to analyze RNA binding proteins that enhance or silence exon recognition (**Table 2.2b**). ESEfinder was used for 42.2% of these mutations in one or more regulatory binding sites^{239,240}. However, variants predicted by ESEfinder to have deleterious effects are discordant with some IT predictions (6 of 15; **Table 2.2b**). The discordance with ESEfinder may be associated with differences in the respective analytic methods, as several of the models (SF2/ASF, SC35, SRp40) used by ASSA and ESEfinder were created from the same source of experimental data^{91,241}. While the majority of the discordant results were cited in a single study¹¹⁴ (5/6 variants), the small size of the dataset (ranging from 28-34 sites) may artificially exacerbate differences between these results. In multiple instances, ASSA has been used to analyze SR proteins, but other programs were used to analyze 5' and 3' splice site mutations^{26,103,119}. This was surprising, since the donor and acceptor R_i values are generated by default by ASSA and ASSEDA. The advantage of performing both constitutive and regulatory splice site analysis with IT is that all results are reported on the same scale, and the strengths of all interactions, and effects of mutations are directly comparable to one another.

2.14 Other applications of information theory- based splice site analysis

The use of IT to analyze splicing is not limited to sequence variant analysis. The natural and alternative splicing of several genes have been characterized using this method^{111,202,242}. Khan *et al.* (2002) scanned all natural sites in the *XPC* gene and found a weak acceptor (-0.1 bits), and with RT-PCR found that this exon (exon 4) was skipped to a greater extent than another (exon 7), which possessed a strong acceptor, illustrating a relationship between the information content of a natural splice site and its level of alternative splicing¹¹¹. IT has also been used in genetic engineering in the design and

alteration of binding sites, and has been used in the design of constructs for transgenic animal models^{243–245}. Thus, IT-based splice site analysis can be adapted for other important molecular genetic applications.

2.15 Guidelines for information theory-based splicing mutation analyses

Our comprehensive review of the use of IT in splicing mutation analysis has led us to propose general recommendations, which we formulate as guidelines. Adoption of these guidelines should ensure the accurate and comprehensive results from IT analyses of VUS and other pathogenic variants that alter mRNA splicing.

2.15.1 Report gene isoform and genomic coordinates

When analyzing a variant with ASSEDA, the user is prompted to select an mRNA isoform (GenBank or RefSeq accession) from the gene in question. When entering the same variant (in either IVS or c. notation) for different isoforms, either the variant will parse one but not the other representation, or the variant syntax will be processed for both. In the first situation, the user is prompted to verify the position and substitution, which may elicit the realization that the incorrect isoform had been selected. However, in the case where the variant can still be parsed (despite being incorrectly entered for the isoform selected), an incorrect nucleotide may coincidentally have the same sequence, and the user may not necessarily realize that the intended position is not being analyzed. We were unable to reproduce results for several variants, because the mRNA or gene isoform was not reported. This issue could be resolved by comparing the genomic sequence in papers where the context of the mutation was included^{54,99,145,181,246–248}. Where flanking sequences were unavailable, the location of the mutation was inferred from either descriptions in the text, the corresponding R_i value of the splice site, or relative coordinate numbering^{148,249,250}. Although we attempted to reproduce all the results, this was not always possible if the specified sequence was ambiguous or the source was deprecated (GenBank accession numbers, BAC clones, etc.)^{52,101,174,181,182,210,229,234,251,252}.

We note that ASSA/ASSEDA cannot account for genes with redacted exons, where the exon numbering or sequence in the original mRNA accession has not been corrected. A well-known example is *BRCAL*, for which the constitutive isoform lacks the exon designated as number 4. IVS notation beyond this point in this gene must be reduced by one intron. Alternatively, one of the HGVS-approved methods can be used to input variants, or the variant can be designated with the genomic coordinate (g.) format. Review of ASSA/ASSEDA output (coordinates and/or the sequence walker²⁰) is a prudent approach to confirm that the correct region has been analyzed.

To eliminate ambiguity, we recommend that reported variants be accompanied by the accession number used in its analysis (consistent with HGVS notation³⁹) and the genomic coordinates with the corresponding reference genome build. The table of results from ASSEDA or Shannon pipeline output could also be included as supplementary published material. This will ensure that reported results can be reproduced and compared to other experimental or *in silico* results.

2.15.2 Report R_i values

The results generated by IT software provide $R_{i,initial}$, $R_{i,final}$, and ΔR_i for donor and acceptor sites by default, and for all other ribl matrices selected. Reporting these values along with the interpretation improves the clarity of said interpretation. Several publications have not reported R_i , and instead only the interpretation of these values^{129,142,150,214,229,253,254}. This presumes that the analysis was performed correctly, and accurately interpreted. In one instance, our reanalysis differed from the published interpretation¹⁴². Other publications provide R_i values, but were incorrectly reported, which resulted in misinterpretations^{52,126}. Simply reporting ΔR_i itself does not provide sufficient information about the context of the mutation or possible cryptic splice sites, which is necessary to fully appreciate the resultant effect on splicing^{140,247,255}. We recommend R_i values be provided for each variant analyzed. We also suggest that the specific donor and acceptor ribl used for variant analysis be indicated, because of the differences obtained using the genome-wide and original PWMs in IT analysis^{33,36}. The

distinction can also be significant, when the $R_{i,final}$ value of a mutated splice site approaches $R_{i,min}$.

2.15.3 Consider impact of missense and synonymous mutations on mRNA splicing

Missense and synonymous mutations can alter natural splicing, create cryptic sites, and alter crucial ESE and ESS binding sites²⁵⁶. IT tools have been employed to analyze exonic variants that strengthened or create exonic cryptic sites, which were also confirmed experimentally^{28,42,44,46,102,109,120,128,134,153,155,180,257,258}. Similarly, IT tools can predict potential effects on strengths of SR and hnRNP protein recognition sites^{26,121}. There is no justification for cataloguing intronic and exonic variants, but only assessing splicing effects for the intronic variants or those within natural splice sites^{123,136,177,188,210,212,216,217,250,259,260}. We recommend that IT-based analysis should evaluate all variants within a gene for potential splicing mutations.

2.15.4 Experimentally validate variants

Many studies have reported only coding changes and the results of IT (or other *in silico*) analyses without experimental validation. Our review indicated that IT-based splicing predictions are highly concordant with results for validation results (87.9%) Nevertheless, the discordant mutations support the need for robust post-prediction validation, since even a single discordant result can lead to misdiagnosis. We do not detect any consistent pattern amongst the discordant predictions to provide guidance as to which IT analyses will be erroneous. Experimental verification will mitigate incorrect interpretations of IT predictions and has been recommended by others²⁹.

2.15.5 Report the sequence window used in the analysis

ASSA/ASSEDA allows the user to alter size of sequence window range surrounding the mutation. The default window range has been set to maximize the speed of analysis, which is to some degree dictated by the number of matrices and the length of the sequence analyzed. Arbitrary abbreviation of the sequence analysis window can result in the failure to detect activated intronic or exonic cryptic sites, which can in some instances

significantly lengthen (eg. 231 and 313 nucleotide extensions, respectively^{166,171}) or shorten the corresponding natural exon. Therefore, we suggest expanding this window if one wishes to assess the possibility that long range, pre-existing cryptic splice sites may be activated.

We note that unequivocal prediction of cryptic splice site use in large exons (> 1000 nt) can be challenging due to the reliance of these gene regions on splicing enhancers, silencers, and other regulatory elements to prevent ectopic splice site use and ensure fidelity of splicing²⁶¹. In a case of familial hypobetalipoproteinaemia, Di Leo *et al.* (2007) determined a variant abolishing the natural acceptor for exon 26 of *APOB* (7572 nt long), causing the activation of a weak cryptic site 1180 nt downstream²⁶². There are several other stronger candidate cryptic splice sites that occur between the natural and cryptic splice site, but there is no evidence that any are used in the individual carrying this mutation.

2.15.6 Designate genic rearrangements (insertions, deletions, duplications) with genomic coordinates

Complex insertions and deletions in IVS or c. notation may occasionally be parsed to the wrong coordinates within a gene. Indels will parse properly when genomic coordinates are used. If IVS or c. notation is used, it is suggested that users confirm that the expected alteration of the mutation is correct by reviewing the sequence walker display generated by ASSEDA for all insertions, deletions and duplications.

2.16 References

1. Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* **11**, 889–900 (2001).
2. Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29**, 2850–2859 (2001).
3. Vandenbroucke, I., Callens, T., De Paepe, A. & Messiaen, L. Complex splicing pattern generates great diversity in human NF1 transcripts. *BMC Genomics* **3**, 13 (2002).
4. Frilander, M. J. & Steitz, J. A. Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. *Genes Dev.* **13**, 851–863 (1999).
5. Will, C. L. & Lührmann, R. Protein functions in pre-mRNA splicing. *Curr. Opin. Cell Biol.* **9**, 320–328 (1997).
6. Burge, C. B., Tuschl, T. & Sharp, P. A. in *The RNA World, 2nd Ed.: The Nature of Modern RNA Suggests a Prebiotic RNA World* **37**, 525–560 (Cold Spring Harbor Press, 1999).
7. Shepard, P. J., Choi, E.-A., Busch, A. & Hertel, K. J. Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. *Nucleic Acids Res.* **39**, 8928–8937 (2011).
8. Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D. N. & Sanford, J. R. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* **21**, 1563–1571 (2011).

9. Zhang, C. *et al.* Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.* **22**, 2550–2563 (2008).
10. Wu, Y., Zhang, Y. & Zhang, J. Distribution of exonic splicing enhancer elements in human genes. *Genomics* **86**, 329–336 (2005).
11. Graveley, B. R., Hertel, K. J. & Maniatis, T. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *EMBO J.* **17**, 6747–6756 (1998).
12. Andresen, B. S. & Krainer, A. When the genetic code is not enough - How sequence variations can affect pre-mRNA splicing and cause (complex) disease. Chapter 15. (2009). at <[http://findresearcher.sdu.dk:8080/portal/en/publications/when-the-genetic-code-is-not-enough--how-sequence-variations-can-affect-premRNA-splicing-and-cause-complex-disease-chapter-15\(a592ab10-0fee-11df-aefb-000ea68e967b\).html](http://findresearcher.sdu.dk:8080/portal/en/publications/when-the-genetic-code-is-not-enough--how-sequence-variations-can-affect-premRNA-splicing-and-cause-complex-disease-chapter-15(a592ab10-0fee-11df-aefb-000ea68e967b).html)>
13. Krawczak, M., Reiss, J. & Cooper, D. N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* **90**, 41–54 (1992).
14. Ars, E. *et al.* Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.* **9**, 237–247 (2000).
15. Teraoka, S. N. *et al.* Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. *Am. J. Hum. Genet.* **64**, 1617–1631 (1999).
16. Shapiro, M. B. & Senapathy, P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**, 7155–7174 (1987).

17. Rogan, P. K. & Schneider, T. D. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.* **6**, 74–76 (1995).
18. Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–1038 (1993).
19. Leach, F. S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**, 1215–1225 (1993).
20. Schneider, T. D. Sequence logos, machine/channel capacity, Maxwell’s demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology* **5**, 1–18 (1994).
21. Schneider, T. D. Information content of individual genetic sequences. *J. Theor. Biol.* **189**, 427–441 (1997).
22. Brunak, S., Engelbrecht, J. & Knudsen, S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220**, 49–65 (1991).
23. Schneider, T. D. Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.* **25**, 4408–4415 (1997).
24. Hengen, P. N., Bartram, S. L., Stewart, L. E. & Schneider, T. D. Information analysis of Fis binding sites. *Nucleic Acids Res.* **25**, 4994–5002 (1997).
25. Rogan, P. K., Faux, B. M. & Schneider, T. D. Information analysis of human splice site mutations. *Hum. Mutat.* **12**, 153–171 (1998).
26. Olsen, R. K. J. *et al.* The ETFDH c.158A>G variation disrupts the balanced interplay of ESE- and ESS-binding proteins thereby causing missplicing and multiple Acyl-CoA dehydrogenation deficiency. *Hum. Mutat.* **35**, 86–95 (2014).

27. Homolova, K. *et al.* The deep intronic c.903+469T>C mutation in the MTRR gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cbIE type of homocystinuria. *Hum. Mutat.* **31**, 437–444 (2010).
28. Mucaki, E. J., Ainsworth, P. & Rogan, P. K. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Hum. Mutat.* **32**, 735–742 (2011).
29. Maddalena, A. *et al.* Technical standards and guidelines: molecular genetic testing for ultra-rare disorders. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **7**, 571–583 (2005).
30. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
31. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431 (1986).
32. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
33. Rogan, P. K., Svojanovsky, S. & Leeder, J. S. Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics* **13**, 207–218 (2003).
34. Schneider, T. D., Stormo, G. D., Haemer, J. S. & Gold, L. A design for computer nucleic-acid-sequence storage, retrieval, and manipulation. *Nucleic Acids Res.* **10**, 3013–3024 (1982).
35. Schneider, T. D., Stormo, G. D., Yarus, M. A. & Gold, L. Delila system tools. *Nucleic Acids Res.* **12**, 129–140 (1984).

36. Stephens, R. M. & Schneider, T. D. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**, 1124–1136 (1992).
37. Nalla, V. K. & Rogan, P. K. Automated splicing mutation analysis by information theory. *Hum. Mutat.* **25**, 334–342 (2005).
38. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Hum. Mutat.* **34**, 557–565 (2013).
39. Dunnen, J. T. den & Antonarakis, S. E. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.* **15**, 7–12 (2000).
40. Tribus, M. *Thermostatistics and thermodynamics*. (New York : Van Nostrand, 1961). at <<http://archive.org/details/thermostaticsthe00trib>>
41. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. (John Wiley & Sons, 2006).
42. Bonnet-Dupeyron, M.-N. *et al.* PLP1 splicing abnormalities identified in Pelizaeus-Merzbacher disease and SPG2 fibroblasts are associated with different types of mutations. *Hum. Mutat.* **29**, 1028–1036 (2008).
43. Rogan, P. K. & Zou, G. Y. Best practices for evaluating mutation prediction methods. *Hum. Mutat.* **34**, 1581–1582 (2013).
44. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* **11**, 77–85 (2013).
45. Benaglio, P. *et al.* Mutational screening of splicing factor genes in cases with autosomal dominant retinitis pigmentosa. *Mol. Vis.* **20**, 843–851 (2014).

46. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3**, 8 (2014).
47. Dorman, S., Viner, C. & Rogan, P. Splicing Mutation Analysis Reveals Previously Unrecognized Pathways in Lymph Node-Invasive Breast Cancer. *Sci. Rep.* In Press (2014).
48. Green, M. R. Pre-mRNA splicing. *Annu. Rev. Genet.* **20**, 671–708 (1986).
49. Maniatis, T. & Reed, R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* **325**, 673–678 (1987).
50. von Kodolitsch, Y., Pyeritz, R. E. & Rogan, P. K. Splice-Site Mutations in Atherosclerosis Candidate Genes Relating Individual Information to Phenotype. *Circulation* **100**, 693–699 (1999).
51. Khan, S. G. *et al.* Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. *Hum. Mol. Genet.* **13**, 343–352 (2004).
52. Fei, J. Splice Site Mutation-Induced Alteration of Selective Regional Activity Correlates with the Role of a Gene in Cardiomyopathy. *J. Clin. Exp. Cardiol.* **S12:004**, (2013).
53. Robberson, B. L., Cote, G. J. & Berget, S. M. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* **10**, 84–94 (1990).
54. Smaoui, N. *et al.* A homozygous splice mutation in the HSF4 gene is associated with an autosomal recessive congenital cataract. *Invest. Ophthalmol. Vis. Sci.* **45**, 2716–2721 (2004).

55. Sharma, N. *et al.* Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. *Hum. Mutat.* **35**, 1249–1259 (2014).
56. Riveira-Munoz, E. *et al.* Evaluating PVALB as a candidate gene for SLC12A3-negative cases of Gitelman's syndrome. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **23**, 3120–3125 (2008).
57. Ozaltin, F. *et al.* Disruption of PTPRO causes childhood-onset nephrotic syndrome. *Am. J. Hum. Genet.* **89**, 139–147 (2011).
58. Di Leo, E. *et al.* Functional analysis of two novel splice site mutations of APOB gene in familial hypobetalipoproteinemia. *Mol. Genet. Metab.* **96**, 66–72 (2009).
59. Behzadnia, N. *et al.* Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes. *EMBO J.* **26**, 1737–1748 (2007).
60. Staley, J. P. & Guthrie, C. Mechanical Devices of the Spliceosome: Motors, Clocks, Springs, and Things. *Cell* **92**, 315–326 (1998).
61. Krawczak, M. *et al.* Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.* **28**, 150–158 (2007).
62. Mort, M. *et al.* MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.* **15**, R19 (2014).
63. Sun, H. & Chasin, L. A. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* **20**, 6414–6425 (2000).
64. Treisman, R., Orkin, S. H. & Maniatis, T. Specific transcription and RNA splicing defects in five cloned β -thalassaemia genes. *Nature* **302**, 591–596 (1983).

65. ElSharawy, A. *et al.* Systematic evaluation of the effect of common SNPs on pre-mRNA splicing. *Hum. Mutat.* **30**, 625–632 (2009).
66. Buratti, E., Baralle, M. & Baralle, F. E. Defective splicing, disease and therapy: searching for master checkpoints in exon definition. *Nucleic Acids Res.* **34**, 3494–3510 (2006).
67. Kapustin, Y. *et al.* Cryptic splice sites and split genes. *Nucleic Acids Res.* **39**, 5837–5844 (2011).
68. Zhang, M. Q. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3**, 698–709 (2002).
69. Fu, X. D. The superfamily of arginine/serine-rich splicing factors. *RNA N. Y. N* **1**, 663–680 (1995).
70. Graveley, B. R. Sorting out the complexity of SR protein functions. *RNA N. Y. N* **6**, 1197–1211 (2000).
71. Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C. & Black, D. L. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat. Struct. Mol. Biol.* **15**, 183–191 (2008).
72. Zheng, Z. M., Huynen, M. & Baker, C. C. A pyrimidine-rich exonic splicing suppressor binds multiple RNA splicing factors and inhibits spliceosome assembly. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14088–14093 (1998).
73. House, A. E. & Lynch, K. W. An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. *Nat. Struct. Mol. Biol.* **13**, 937–944 (2006).
74. Shen, M. & Mattox, W. Activation and repression functions of an SR splicing regulator depend on exonic versus intronic-binding position. *Nucleic Acids Res.* **40**, 428–437 (2012).

75. Erkelenz, S. *et al.* Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA* **19**, 96–102 (2013).
76. Zatkova, A. *et al.* Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of NF1. *Hum. Mutat.* **24**, 491–501 (2004).
77. Gonçalves, V. *et al.* A missense mutation in the APC tumor suppressor gene disrupts an ASF/SF2 splicing enhancer motif and causes pathogenic skipping of exon 14. *Mutat. Res.* **662**, 33–36 (2009).
78. Miyajima, H., Miyaso, H., Okumura, M., Kurisu, J. & Imaizumi, K. Identification of a cis-acting element for the regulation of SMN exon 7 splicing. *J. Biol. Chem.* **277**, 23271–23277 (2002).
79. Burgess, R. *et al.* ADVIRC is caused by distinct mutations in BEST1 that alter pre-mRNA splicing. *J. Med. Genet.* **46**, 620–625 (2009).
80. Gabut, M. *et al.* The SR protein SC35 is responsible for aberrant splicing of the E1 alpha pyruvate dehydrogenase mRNA in a case of mental retardation with lactic acidosis. *Mol. Cell. Biol.* **25**, 3286–3294 (2005).
81. Goren, A. *et al.* Overlapping splicing regulatory motifs--combinatorial effects on splicing. *Nucleic Acids Res.* **38**, 3318–3327 (2010).
82. Zahler, A. M., Damgaard, C. K., Kjems, J. & Caputi, M. SC35 and Heterogeneous Nuclear Ribonucleoprotein A/B Proteins Bind to a Juxtaposed Exonic Splicing Enhancer/Exonic Splicing Silencer Element to Regulate HIV-1 tat Exon 2 Splicing. *J. Biol. Chem.* **279**, 10077–10084 (2004).
83. Chou, M.-Y., Rooke, N., Turck, C. W. & Black, D. L. hnRNP H Is a Component of a Splicing Enhancer Complex That Activates a c-src Alternative Exon in Neuronal Cells. *Mol. Cell. Biol.* **19**, 69–77 (1999).

84. Xu, J. *et al.* A Heroin Addiction Severity-Associated Intronic Single Nucleotide Polymorphism Modulates Alternative Pre-mRNA Splicing of the μ Opioid Receptor Gene OPRM1 via hnRNPH Interactions. *J. Neurosci.* **34**, 11048–11066 (2014).
85. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* **15**, 689–701 (2014).
86. Mercado, P. A., Ayala, Y. M., Romano, M., Buratti, E. & Baralle, F. E. Depletion of TDP 43 overrides the need for exonic and intronic splicing enhancers in the human apoA-II gene. *Nucleic Acids Res.* **33**, 6000–6010 (2005).
87. Huelga, S. C. *et al.* Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.* **1**, 167–178 (2012).
88. Tavanez, J. P., Madl, T., Kooshapur, H., Sattler, M. & Valcárcel, J. hnRNP A1 Proofreads 3' Splice Site Recognition by U2AF. *Mol. Cell* **45**, 314–329 (2012).
89. Caputi, M., Freund, M., Kammler, S., Asang, C. & Schaal, H. A bidirectional SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency virus type 1 rev, env, vpu, and nef gene expression. *J. Virol.* **78**, 6517–6526 (2004).
90. Expert-Bezançon, A. *et al.* hnRNP A1 and the SR Proteins ASF/SF2 and SC35 Have Antagonistic Functions in Splicing of β -Tropomyosin Exon 6B. *J. Biol. Chem.* **279**, 38249–38259 (2004).
91. Liu, H. X., Chew, S. L., Cartegni, L., Zhang, M. Q. & Krainer, A. R. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol. Cell. Biol.* **20**, 1063–1071 (2000).
92. Pandit, S. *et al.* Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. *Mol. Cell* **50**, 223–235 (2013).

93. Han, J. *et al.* SR Proteins Induce Alternative Exon Skipping through Their Activities on the Flanking Constitutive Exons. *Mol. Cell. Biol.* **31**, 793–802 (2011).
94. Shultz, J. C. *et al.* SRSF1 Regulates the Alternative Splicing of Caspase 9 Via A Novel Intronic Splicing Enhancer Affecting the Chemotherapeutic Sensitivity of Non–Small Cell Lung Cancer Cells. *Mol. Cancer Res.* **9**, 889–900 (2011).
95. Paradis, C. *et al.* hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. *RNA N. Y. N* **13**, 1287–1300 (2007).
96. Mukherjee, N. *et al.* Integrative regulatory mapping indicates that the RNA-binding protein HuR (ELAVL1) couples pre-mRNA processing and mRNA stability. *Mol. Cell* **43**, 327–339 (2011).
97. Uren, P. J. *et al.* Genomic Analyses of the RNA-binding Protein Hu Antigen R (HuR) Identify a Complex Network of Target Genes and Novel Characteristics of Its Binding Sites. *J. Biol. Chem.* **286**, 37063–37066 (2011).
98. Kölsch, H. *et al.* CYP46A1 variants influence Alzheimer’s disease risk and brain cholesterol metabolism. *Eur. Psychiatry J. Assoc. Eur. Psychiatr.* **24**, 183–190 (2009).
99. Khan, S. G. *et al.* Xeroderma pigmentosum group C splice mutation associated with autism and hypoglycinemia. *J. Invest. Dermatol.* **111**, 791–796 (1998).
100. Maruszak, A. *et al.* PIN1 gene variants in Alzheimer’s disease. *BMC Med. Genet.* **10**, 115 (2009).
101. Caux-Moncoutier, V. *et al.* Impact of BRCA1 and BRCA2 variants on splicing: clues from an allelic imbalance study. *Eur. J. Hum. Genet. EJHG* **17**, 1471–1480 (2009).

102. Vockley, J. *et al.* Exon skipping in IVD RNA processing in isovaleric acidemia caused by point mutations in the coding region of the IVD gene. *Am. J. Hum. Genet.* **66**, 356–367 (2000).
103. Vemula, S. R. *et al.* A rare sequence variant in intron 1 of THAP1 is associated with primary dystonia. *Mol. Genet. Genomic Med.* **2**, 261–272 (2014).
104. Astuto, L. M. *et al.* Searching for evidence of DFNB2. *Am. J. Med. Genet.* **109**, 291–297 (2002).
105. López-Jiménez, E. *et al.* SDHC mutation in an elderly patient without familial antecedents. *Clin. Endocrinol. (Oxf.)* **69**, 906–910 (2008).
106. Baturina, O. A., Tupikin, A. E., Lukjanova, T. V., Sosnitskaya, S. V. & Morozov, I. V. PAH and QDPR Deficiency Associated Mutations in the Novosibirsk Region of the Russian Federation: Correlation of Mutation Type with Disease Manifestation and Severity. *J. Med. Biochem.* **33**, 333–340 (2014).
107. Dash, D. P. *et al.* Mutational screening of VSX1 in keratoconus patients from the European population. *Eye Lond. Engl.* **24**, 1085–1092 (2010).
108. Ellis, J. R., Jr, Heinrich, B., Mautner, V.-F. & Kluwe, L. Effects of splicing mutations on NF2-transcripts: transcript analysis and information theoretic predictions. *Genes. Chromosomes Cancer* **50**, 571–584 (2011).
109. Wessagowit, V., Kim, S.-C., Woong Oh, S. & McGrath, J. A. Genotype–Phenotype Correlation in Recessive Dystrophic Epidermolysis Bullosa: When Missense Doesn't Make Sense. *J. Invest. Dermatol.* **124**, 863–866 (2005).
110. Chang, Y. F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.* **76**, 51–74 (2007).
111. Khan, S. G. *et al.* The human XPC DNA repair gene: arrangement, splice site information content and influence of a single nucleotide polymorphism in a splice

- acceptor site on alternative splicing and function. *Nucleic Acids Res.* **30**, 3624–3631 (2002).
112. Goldin, E. *et al.* Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucopolipidosis IV. *Hum. Mutat.* **24**, 460–465 (2004).
113. Bloethner, S., Mould, A., Stark, M. & Hayward, N. K. Identification of ARHGEF17, DENND2D, FGFR3, and RB1 mutations in melanoma by inhibition of nonsense-mediated mRNA decay. *Genes. Chromosomes Cancer* **47**, 1076–1085 (2008).
114. Denson, J. *et al.* Screening for inter-individual splicing differences in human GSTM4 and the discovery of a single nucleotide substitution related to the tandem skipping of two exons. *Gene* **379**, 148–155 (2006).
115. Ben-Salem, S., Begum, M. A., Ali, B. R. & Al-Gazali, L. A Novel Aberrant Splice Site Mutation in RAB23 Leads to an Eight Nucleotide Deletion in the mRNA and Is Responsible for Carpenter Syndrome in a Consanguineous Emirati Family. *Mol. Syndromol.* **3**, 255–261 (2013).
116. Aggarwal, S., Jinda, W., Limwongse, C., Atchaneeyasakul, L. & Phadke, S. R. Run-on mutation in the PAX6 gene and chorioretinal degeneration in autosomal dominant aniridia. *Mol. Vis.* **17**, 1305–1309 (2011).
117. Di Giacomo, D. *et al.* Functional analysis of a large set of BRCA2 exon 7 variants highlights the predictive value of hexamer scores in detecting alterations of exonic splicing regulatory elements. *Hum. Mutat.* **34**, 1547–1557 (2013).
118. Aissat, A. *et al.* Combined computational-experimental analyses of CFTR exon strength uncover predictability of exon-skipping level. *Hum. Mutat.* **34**, 873–881 (2013).
119. Anczuków, O. *et al.* Unclassified variants identified in BRCA1 exon 11: Consequences on splicing. *Genes. Chromosomes Cancer* **47**, 418–426 (2008).

120. Colombo, M. *et al.* Comparative in vitro and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PloS One* **8**, e57173 (2013).
121. Lacroix, M. *et al.* Clinical expression and new SPINK5 splicing defects in Netherton syndrome: unmasking a frequent founder synonymous mutation and unconventional intronic mutations. *J. Invest. Dermatol.* **132**, 575–582 (2012).
122. Lamba, V. *et al.* Hepatic CYP2B6 expression: gender and ethnic differences and relationship to CYP2B6 genotype and CAR (constitutive androstane receptor) expression. *J. Pharmacol. Exp. Ther.* **307**, 906–922 (2003).
123. Lee, Y.-W. *et al.* Different spectrum of mutations of isovaleryl-CoA dehydrogenase (IVD) gene in Korean patients with isovaleric acidemia. *Mol. Genet. Metab.* **92**, 71–77 (2007).
124. Le Guédard-Méreuze, S. *et al.* Sequence contexts that determine the pathogenicity of base substitutions at position +3 of donor splice-sites. *Hum. Mutat.* **30**, 1329–1339 (2009).
125. Tournier, I. *et al.* A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum. Mutat.* **29**, 1412–1424 (2008).
126. Laššuthová, P. *et al.* Three New PLP1 Splicing Mutations Demonstrate Pathogenic and Phenotypic Diversity of Pelizaeus-Merzbacher Disease. *J. Child Neurol.* **29**, 924–931 (2013).
127. Hefferon, T. W., Broackes-Carter, F. C., Harris, A. & Cutting, G. R. Atypical 5' splice sites cause CFTR exon 9 to be vulnerable to skipping. *Am. J. Hum. Genet.* **71**, 294–303 (2002).

128. O'Neill, J. P., Rogan, P. K., Cariello, N. & Nicklas, J. A. Mutations that alter RNA splicing of the human HPRT gene: a review of the spectrum. *Mutat. Res. Mutat. Res.* **411**, 179–214 (1998).
129. Nasim, M. T. *et al.* Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. *Hum. Mutat.* **32**, 1385–1389 (2011).
130. Bocchi, L. *et al.* Multiple abnormally spliced ABCA1 mRNAs caused by a novel splice site mutation of ABCA1 gene in a patient with Tangier disease. *Clin. Chim. Acta Int. J. Clin. Chem.* **411**, 524–530 (2010).
131. von Kodolitsch, Y., Berger, J. & Rogan, P. K. Predicting severity of haemophilia A and B splicing mutations by information analysis. *Haemoph. Off. J. World Fed. Hemoph.* **12**, 258–262 (2006).
132. Hageman, G. S. *et al.* A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7227–7232 (2005).
133. Ben Selma, Z. *et al.* A novel S115G mutation of CGI-58 in a Turkish patient with Dorfman-Chanarin syndrome. *J. Invest. Dermatol.* **127**, 2273–2276 (2007).
134. Roux-Buisson, N. *et al.* Functional analysis reveals splicing mutations of the CASQ2 gene in patients with CPVT: implication for genetic counselling and clinical management. *Hum. Mutat.* **32**, 995–999 (2011).
135. Qin, S. *et al.* Systematic polymorphism analysis of the CYP2D6 gene in four different geographical Han populations in mainland China. *Genomics* **92**, 152–158 (2008).
136. Gaweda-Walerych, K. *et al.* Mitochondrial transcription factor A variants and the risk of Parkinson's disease. *Neurosci. Lett.* **469**, 24–29 (2010).

137. Fornage, M. *et al.* The soluble epoxide hydrolase gene harbors sequence variation associated with susceptibility to and protection from incident ischemic stroke. *Hum. Mol. Genet.* **14**, 2829–2837 (2005).
138. Allikmets, R. *et al.* Organization of the ABCR gene: analysis of promoter and splice junction sequences. *Gene* **215**, 111–122 (1998).
139. Simpson, M. A. *et al.* Mutations in FAM20C are associated with lethal osteosclerotic bone dysplasia (Raine syndrome), highlighting a crucial molecule in bone development. *Am. J. Hum. Genet.* **81**, 906–912 (2007).
140. Henneman, P., Schaap, F. G., Rensen, P. C. N., van Dijk, K. W. & Smelt, A. H. M. Estrogen induced hypertriglyceridemia in an apolipoprotein AV deficient patient. *J. Intern. Med.* **263**, 107–108 (2008).
141. Fong, K. *et al.* Infantile systemic hyalinosis associated with a putative splice-site mutation in the ANTXR2 gene. *Clin. Exp. Dermatol.* **37**, 635–638 (2012).
142. Douglas, D. A. *et al.* Novel mutations of epidermal growth factor receptor in localized prostate cancer. *Front. Biosci. J. Virtual Libr.* **11**, 2518–2525 (2006).
143. Gaedigk, A. *et al.* Variability of CYP2J2 expression in human fetal tissues. *J. Pharmacol. Exp. Ther.* **319**, 523–532 (2006).
144. Sabet, A. *et al.* Skin biopsies demonstrate MPZ splicing abnormalities in Charcot-Marie-Tooth neuropathy 1B. *Neurology* **67**, 1141–1146 (2006).
145. Concolino, P. *et al.* Functional analysis of two rare CYP21A2 mutations detected in Italian patients with a mildest form of congenital adrenal hyperplasia. *Clin. Endocrinol. (Oxf.)* **71**, 470–476 (2009).
146. Marras, E. *et al.* Discrepancies between in silico and in vitro data in the functional analysis of a breast cancer-associated polymorphism in the XRCC6/Ku70 gene. *Mol. Med. Rep.* **1**, 805–812 (2008).

147. Li, A. *et al.* Bietti crystalline corneoretinal dystrophy is caused by mutations in the novel gene CYP4V2. *Am. J. Hum. Genet.* **74**, 817–826 (2004).
148. Borroni, B. *et al.* Progranulin genetic variations in frontotemporal lobar degeneration: evidence for low mutation frequency in an Italian clinical series. *Neurogenetics* **9**, 197–205 (2008).
149. Kölsch, H. *et al.* RXRA gene variations influence Alzheimer's disease risk and cholesterol metabolism. *J. Cell. Mol. Med.* **13**, 589–598 (2009).
150. Jeon, G. W., Kwon, M.-J., Lee, S. J., Sin, J. B. & Ki, C.-S. Clinical and genetic analysis of a Korean patient with X-linked chondrodysplasia punctata: identification of a novel splicing mutation in the ARSE gene. *Ann. Clin. Lab. Sci.* **43**, 70–75 (2013).
151. Vreeswijk, M. P. G. *et al.* Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. *Hum. Mutat.* **30**, 107–114 (2009).
152. Kölsch, H. *et al.* Association of SORL1 gene variants with Alzheimer's disease. *Brain Res.* **1264**, 1–6 (2009).
153. Oh, S.-W., Lee, J. S., Kim, M. Y. & Kim, S.-C. COL7A1 mutational analysis in Korean patients with dystrophic epidermolysis bullosa. *Br. J. Dermatol.* **157**, 1260–1264 (2007).
154. Sanggaard, K. M. *et al.* Branchio-oto-renal syndrome: detection of EYA1 and SIX1 mutations in five out of six Danish families by combining linkage, MLPA and sequencing analyses. *Eur. J. Hum. Genet. EJHG* **15**, 1121–1131 (2007).
155. Wessagowit, V., Nalla, V. K., Rogan, P. K. & McGrath, J. A. Normal and abnormal mechanisms of gene splicing and relevance to inherited skin diseases. *J. Dermatol. Sci.* **40**, 73–84 (2005).

156. Slavotinek, A. M. *et al.* Manitoba-oculo-tricho-anal (MOTA) syndrome is caused by mutations in *FREM1*. *J. Med. Genet.* **48**, 375–382 (2011).
157. Moriwaki, K. *et al.* Deficiency of GMDS leads to escape from NK cell-mediated tumor surveillance through modulation of TRAIL signaling. *Gastroenterology* **137**, 188–198, 198.e1–2 (2009).
158. Bröer, S. *et al.* Iminoglycinuria and hyperglycinuria are discrete human phenotypes resulting from complex mutations in proline and glycine transporters. *J. Clin. Invest.* **118**, 3881–3892 (2008).
159. Kwon, M.-J. *et al.* Screening of the *SOD1*, *FUS*, *TARDBP*, *ANG*, and *OPTN* mutations in Korean patients with familial and sporadic ALS. *Neurobiol. Aging* **33**, 1017.e17–23 (2012).
160. Clark, G. R. *et al.* Development of a diagnostic genetic test for simplex and autosomal recessive retinitis pigmentosa. *Ophthalmology* **117**, 2169–2177.e3 (2010).
161. Bertolini, S. *et al.* Spectrum of mutations and phenotypic expression in patients with autosomal dominant hypercholesterolemia identified in Italy. *Atherosclerosis* **227**, 342–348 (2013).
162. Catucci, I. *et al.* *PALB2* sequencing in Italian familial breast cancer cases reveals a high-risk mutation recurrent in the province of Bergamo. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **16**, 688–694 (2014).
163. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).
164. Wang, P. *et al.* Novel mutations of the *PAX6* gene identified in Chinese patients with aniridia. *Mol. Vis.* **12**, 644–648 (2006).

165. Hamada, T. *et al.* Molecular and clinical characterization in Japanese and Korean patients with Hailey-Hailey disease: six new mutations in the ATP2C1 gene. *J. Dermatol. Sci.* **51**, 31–36 (2008).
166. Yu, H. & Patel, S. B. Recent insights into the Smith-Lemli-Opitz syndrome. *Clin. Genet.* **68**, 383–391 (2005).
167. Russcher, H. *et al.* Strategies for the characterization of disorders in cortisol sensitivity. *J. Clin. Endocrinol. Metab.* **91**, 694–701 (2006).
168. Leclerc, D., Wu, Q., Ellis, J. R., Goodyer, P. & Rozen, R. Is the SLC7A10 gene on chromosome 19 a candidate locus for cystinuria? *Mol. Genet. Metab.* **73**, 333–339 (2001).
169. Leclerc, D. *et al.* SLC7A9 mutations in all three cystinuria subtypes. *Kidney Int.* **62**, 1550–1559 (2002).
170. Marchal, A. *et al.* Un cas particulier d'épidermolyse bulleuse dystrophique. *Ann. Dermatol. Vénérologie* **138**, A168–A169 (2011).
171. Oh, K.-S. *et al.* Phenotypic heterogeneity in the XPB DNA helicase gene (ERCC3): xeroderma pigmentosum without and with Cockayne syndrome. *Hum. Mutat.* **27**, 1092–1103 (2006).
172. Lim, B. C. *et al.* Fukutin mutations in congenital muscular dystrophies with defective glycosylation of dystroglycan in Korea. *Neuromuscul. Disord. NMD* **20**, 524–530 (2010).
173. Marco, E. J. *et al.* ARHGEF9 disruption in a female patient is associated with X linked mental retardation and sensory hyperarousal. *J. Med. Genet.* **45**, 100–105 (2008).

174. Gemignani, F. *et al.* A TP53 polymorphism is associated with increased risk of colorectal cancer and with reduced levels of TP53 mRNA. *Oncogene* **23**, 1954–1956 (2003).
175. Luquin, N., Yu, B., Saunderson, R. B., Trent, R. J. & Pamphlett, R. Genetic variants in the promoter of TARDBP in sporadic amyotrophic lateral sclerosis. *Neuromuscul. Disord. NMD* **19**, 696–700 (2009).
176. Magnolo, L. *et al.* Novel mutations in SAR1B and MTTP genes in Tunisian children with chylomicron retention disease and abetalipoproteinemia. *Gene* **512**, 28–34 (2013).
177. Marr, N. *et al.* Cell-biologic and functional analyses of five new Aquaporin-2 missense mutations that cause recessive nephrogenic diabetes insipidus. *J. Am. Soc. Nephrol. JASN* **13**, 2267–2277 (2002).
178. Naiya, T. *et al.* Occurrence of GCH1 gene mutations in a group of Indian dystonia patients. *J. Neural Transm. Vienna Austria 1996* **119**, 1343–1350 (2012).
179. Fasano, T., Bocchi, L., Pisciotta, L., Bertolini, S. & Calandra, S. Denaturing high-performance liquid chromatography in the detection of ABCA1 gene mutations in familial HDL deficiency. *J. Lipid Res.* **46**, 817–822 (2005).
180. Tosetto, E. *et al.* Phenotypic and genetic heterogeneity in Dent's disease--the results of an Italian collaborative study. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **21**, 2452–2463 (2006).
181. Tosetto, E. *et al.* Novel mutations of the CLCN5 gene including a complex allele and A 5' UTR mutation in Dent disease 1. *Clin. Genet.* **76**, 413–416 (2009).
182. Tram, E. *et al.* Identification of germline alterations of the mad homology 2 domain of SMAD3 and SMAD4 from the Ontario site of the breast cancer family registry (CFR). *Breast Cancer Res. BCR* **13**, R77 (2011).

183. Xu, X. *et al.* Sequence variations of GRM6 in patients with high myopia. *Mol. Vis.* **15**, 2094–2100 (2009).
184. Pink, A. E. *et al.* Mutations in the γ -secretase genes NCSTN, PSENEN, and PSEN1 underlie rare forms of hidradenitis suppurativa (acne inversa). *J. Invest. Dermatol.* **132**, 2459–2461 (2012).
185. Chen, L. J. *et al.* Evaluation of SPARC as a candidate gene of juvenile-onset primary open-angle glaucoma by mutation and copy number analyses. *Mol. Vis.* **16**, 2016–2025 (2010).
186. Chen, L. *et al.* Genetic polymorphism analysis of CYP2C19 in Chinese Han populations from different geographic areas of mainland China. *Pharmacogenomics* **9**, 691–702 (2008).
187. Liu, J. *et al.* The association of LRP5 gene polymorphisms with ankylosing spondylitis in a Chinese Han population. *J. Rheumatol.* **38**, 2616–2618 (2011).
188. Deen, P. M. T., Dahl, N. & Caplan, M. J. The aquaporin-2 water channel in autosomal dominant primary nocturnal enuresis. *J. Urol.* **167**, 1447–1450 (2002).
189. Bonafé, L., Giunta, C., Gassner, M., Steinmann, B. & Superti-Furga, A. A cluster of autosomal recessive spondylocostal dysostosis caused by three newly identified DLL3 mutations segregating in a small village. *Clin. Genet.* **64**, 28–35 (2003).
190. Megremis, S. *et al.* Nucleotide variations in the NPHS2 gene in Greek children with steroid-resistant nephrotic syndrome. *Genet. Test. Mol. Biomark.* **13**, 249–256 (2009).
191. Rogan, P. & Mucaki, E. Population Fitness and Genetic Load of Single Nucleotide Polymorphisms Affecting mRNA splicing. *ArXiv11070716 Q-Bio* (2011). at <<http://arxiv.org/abs/1107.0716>>

192. Day, I. N. *et al.* Day INM. IDDM2 locus: 5' noncoding intron I splicing and translational efficiency effects of INS -23HphI - more than a tag for the INS promoter VNTR. (2006). at <abstracts/hgvs.org/Helsinki/Presentations/Day.ppt>
193. Taube, J. R. *et al.* PMD patient mutations reveal a long-distance intronic interaction that regulates PLP1/DM20 alternative splicing. *Hum. Mol. Genet.* **23**, 5464–5478 (2014).
194. Luquin, N., Yu, B., Trent, R. J., Morahan, J. M. & Pamphlett, R. An analysis of the entire SOD1 gene in sporadic ALS. *Neuromuscul. Disord. NMD* **18**, 545–552 (2008).
195. Ueffing, N. *et al.* A single nucleotide polymorphism determines protein isoform production of the human c-FLIP protein. *Blood* **114**, 572–579 (2009).
196. Batty, J. A. *et al.* An investigation of CYP2D6 genotype and response to metoprolol CR/XL during dose titration in patients with heart failure: a MERIT-HF substudy. *Clin. Pharmacol. Ther.* **95**, 321–330 (2014).
197. Chiu, C.-F. *et al.* A novel single nucleotide polymorphism in XRCC4 gene is associated with oral cancer susceptibility in Taiwanese patients. *Oral Oncol.* **44**, 898–902 (2008).
198. Zhao, P. *et al.* Genetic polymorphisms of DNA double-strand break repair pathway genes and glioma susceptibility. *BMC Cancer* **13**, 234 (2013).
199. Drögemüller, C., Philipp, U., Haase, B., Günzel-Apel, A.-R. & Leeb, T. A noncoding melanophilin gene (MLPH) SNP at the splice donor of exon 1 represents a candidate causal mutation for coat color dilution in dogs. *J. Hered.* **98**, 468–473 (2007).
200. Kölsch, H. *et al.* Influence of SORL1 gene variants: association with CSF amyloid-beta products in probable Alzheimer's disease. *Neurosci. Lett.* **440**, 68–71 (2008).

201. Cox, D.-G. *et al.* Haplotype of prostaglandin synthase 2/cyclooxygenase 2 is involved in the susceptibility to inflammatory bowel disease. *World J. Gastroenterol. WJG* **11**, 6003–6008 (2005).
202. Thompson, D., Easton, D. F. & Breast Cancer Linkage Consortium. Cancer Incidence in BRCA1 mutation carriers. *J. Natl. Cancer Inst.* **94**, 1358–1365 (2002).
203. Palomino-Doza, J. *et al.* Ambulatory blood pressure is associated with polymorphic variation in P2X receptor genes. *Hypertension* **52**, 980–985 (2008).
204. Xiong, Y. *et al.* A systematic genetic polymorphism analysis of the CYP2C9 gene in four different geographical Han populations in mainland China. *Genomics* **97**, 277–281 (2011).
205. Mao, M., Skogh, E., Scordo, M. G. & Dahl, M.-L. Interindividual variation in olanzapine concentration influenced by UGT1A4 L48V polymorphism in serum and upstream FMO polymorphisms in cerebrospinal fluid. *J. Clin. Psychopharmacol.* **32**, 287–289 (2012).
206. Hiller, M. *et al.* Phylogenetically widespread alternative splicing at unusual GYNGYN donors. *Genome Biol.* **7**, R65 (2006).
207. Pasvolsky, R. *et al.* A LAD-III syndrome is associated with defective expression of the Rap-1 activator CalDAG-GEFI in lymphocytes, neutrophils, and platelets. *J. Exp. Med.* **204**, 1571–1582 (2007).
208. Cartault, F. *et al.* A new XPC gene splicing mutation has lead to the highest worldwide prevalence of xeroderma pigmentosum in black Mahori patients. *DNA Repair* **10**, 577–585 (2011).
209. Wang, J. *et al.* Identification of a novel specific CYP2B6 allele in Africans causing impaired metabolism of the HIV drug efavirenz. *Pharmacogenet. Genomics* **16**, 191–198 (2006).

210. Gaedigk, A. *et al.* Identification and characterization of novel sequence variations in the cytochrome P4502D6 (CYP2D6) gene in African Americans. *Pharmacogenomics J.* **5**, 173–182 (2005).
211. Green, R. C. *et al.* ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15**, 565–574 (2013).
212. Garcia-Gonzalez, M. A. *et al.* Evaluating the clinical utility of a molecular genetic test for polycystic kidney disease. *Mol. Genet. Metab.* **92**, 160–167 (2007).
213. Leman, A. R., Pearce, D. A. & Rothberg, P. G. Gene symbol: CLN3. Disease: Juvenile neuronal ceroid lipofuscinosis (Batten disease). *Hum. Genet.* **116**, 544 (2005).
214. Keren, B. *et al.* CNS malformations in Knobloch syndrome with splice mutation in COL18A1 gene. *Am. J. Med. Genet. A.* **143A**, 1514–1518 (2007).
215. Aoyama, Y. *et al.* Molecular features of 23 patients with glycogen storage disease type III in Turkey: a novel mutation p.R1147G associated with isolated glucosidase deficiency, along with 9 AGL mutations. *J. Hum. Genet.* **54**, 681–686 (2009).
216. Kwong, A. K.-Y., Fung, C.-W., Chan, S.-Y. & Wong, V. C.-N. Identification of SCN1A and PCDH19 mutations in Chinese children with Dravet syndrome. *PloS One* **7**, e41802 (2012).
217. Li, L. *et al.* Detection of variants in 15 genes in 87 unrelated Chinese patients with Leber congenital amaurosis. *PloS One* **6**, e19458 (2011).
218. Caridi, G. *et al.* Analbuminemia Zonguldak: case report and mutational analysis. *Clin. Biochem.* **41**, 288–291 (2008).
219. Papp, J., Kovacs, M. E. & Olah, E. Germline MLH1 and MSH2 mutational spectrum including frequent large genomic aberrations in Hungarian hereditary non-

- polyposis colorectal cancer families: implications for genetic testing. *World J. Gastroenterol. WJG* **13**, 2727–2732 (2007).
220. Saeed, S. *et al.* Novel LEPR mutations in obese Pakistani children identified by PCR-based enrichment and next generation sequencing. *Obes. Silver Spring Md* **22**, 1112–1117 (2014).
221. Soran, H. *et al.* Proteinuria and severe mixed dyslipidemia associated with a novel APOAV gene mutation. *J. Clin. Lipidol.* **4**, 310–313 (2010).
222. Sznajder, Y. *et al.* A de novo SOX10 mutation causing severe type 4 Waardenburg syndrome without Hirschsprung disease. *Am. J. Med. Genet. A.* **146A**, 1038–1041 (2008).
223. Eichers, E. R. *et al.* Newfoundland rod-cone dystrophy, an early-onset retinal dystrophy, is caused by splice-junction mutations in RLBP1. *Am. J. Hum. Genet.* **70**, 955–964 (2002).
224. Dua-Awereh, M. B., Shimomura, Y., Kraemer, L., Wajid, M. & Christiano, A. M. Mutations in the desmoglein 1 gene in five Pakistani families with striate palmoplantar keratoderma. *J. Dermatol. Sci.* **53**, 192–197 (2009).
225. Hampson, G., Konrad, M. A. & Scoble, J. Familial hypomagnesaemia with hypercalciuria and nephrocalcinosis (FHHNC): compound heterozygous mutation in the claudin 16 (CLDN16) gene. *BMC Nephrol.* **9**, 12 (2008).
226. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **11**, 377–394 (2004).
227. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in Genie. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **4**, 311–323 (1997).

228. Beetz, C. *et al.* REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. *Brain J. Neurol.* **131**, 1078–1086 (2008).
229. Cruchaga, C. *et al.* Cortical atrophy and language network reorganization associated with a novel progranulin mutation. *Cereb. Cortex N. Y. N 1991* **19**, 1751–1760 (2009).
230. Martoni, E. *et al.* Identification and characterization of novel collagen VI non-canonical splicing mutations causing Ullrich congenital muscular dystrophy. *Hum. Mutat.* **30**, E662–672 (2009).
231. Naruse, H. *et al.* Determination of splice-site mutations in Lynch syndrome (hereditary non-polyposis colorectal cancer) patients using functional splicing assay. *Fam. Cancer* **8**, 509–517 (2009).
232. Pelucchi, S. *et al.* Expression of hepcidin and other iron-related genes in type 3 hemochromatosis due to a novel mutation in transferrin receptor-2. *Haematologica* **94**, 276–279 (2009).
233. Bacci, C. *et al.* Schwannomatosis associated with multiple meningiomas due to a familial SMARCB1 mutation. *Neurogenetics* **11**, 73–80 (2010).
234. Torregrossa, R. *et al.* Identification of GDNF gene sequence variations in patients with medullary sponge kidney disease. *Clin. J. Am. Soc. Nephrol. CJASN* **5**, 1205–1210 (2010).
235. Cohen, B. *et al.* A novel splice site mutation of CDHR1 in a consanguineous Israeli Christian Arab family segregating autosomal recessive cone-rod dystrophy. *Mol. Vis.* **18**, 2915–2921 (2012).
236. Fasano, T. *et al.* Lysosomal lipase deficiency: molecular characterization of eleven patients with Wolman or cholesteryl ester storage disease. *Mol. Genet. Metab.* **105**, 450–456 (2012).

237. Pernet, C. *et al.* Genitoperineal papular acantholytic dyskeratosis is allelic to Hailey-Hailey disease. *Br. J. Dermatol.* **167**, 210–212 (2012).
238. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).
239. Cartegni, L. & Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat. Genet.* **30**, 377–384 (2002).
240. Smith, P. J. *et al.* An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum. Mol. Genet.* **15**, 2490–2508 (2006).
241. Liu, H. X., Zhang, M. & Krainer, A. R. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev.* **12**, 1998–2012 (1998).
242. Lou, H. *et al.* Promoter variants in the MSMB gene associated with prostate cancer regulate MSMB/NCOA4 fusion transcripts. *Hum. Genet.* **131**, 1453–1466 (2012).
243. Cronin, C. A., Gluba, W. & Scrable, H. The lac operator-repressor system is functional in the mouse. *Genes Dev.* **15**, 1506–1517 (2001).
244. Wang, E., Dimova, N. & Cambi, F. PLP/DM20 ratio is regulated by hnRNPH and F and a novel G-rich enhancer in oligodendrocytes. *Nucleic Acids Res.* **35**, 4164–4178 (2007).
245. Schneider, T. D. & Rogan, P. K. Computational analysis of nucleic acid information defines binding sites. (1999).
246. Botta, E. *et al.* Genotype-phenotype relationships in trichothiodystrophy patients with novel splicing mutations in the XPD gene. *Hum. Mutat.* **30**, 438–445 (2009).

247. Lietman, S. A. Preimplantation genetic diagnosis for hereditary endocrine disease. *Endocr. Pract. Off. J. Am. Coll. Endocrinol. Am. Assoc. Clin. Endocrinol.* **17 Suppl 3**, 28–32 (2011).
248. Hellerud, C. *et al.* Glycerol metabolism and the determination of triglycerides--clinical, biochemical and molecular findings in six subjects. *Clin. Chem. Lab. Med. CCLM FESCC* **41**, 46–55 (2003).
249. Akiyama, M. *et al.* DNA-based prenatal diagnosis of harlequin ichthyosis and characterization of ABCA12 mutation consequences. *J. Invest. Dermatol.* **127**, 568–573 (2007).
250. Luquin, N., Yu, B., Saunderson, R. B., Trent, R. J. & Pamphlett, R. Genetic variants in the promoter of TARDBP in sporadic amyotrophic lateral sclerosis. *Neuromuscul. Disord. NMD* **19**, 696–700 (2009).
251. Koukouritaki, S. B., Poch, M. T., Cabacungan, E. T., McCarver, D. G. & Hines, R. N. Discovery of novel flavin-containing monooxygenase 3 (FMO3) single nucleotide polymorphisms and functional analysis of upstream haplotype variants. *Mol. Pharmacol.* **68**, 383–392 (2005).
252. Karaca, M. *et al.* High prevalence of cerebral venous sinus thrombosis (CVST) as presentation of cystathionine beta-synthase deficiency in childhood: molecular and clinical findings of Turkish probands. *Gene* **534**, 197–203 (2014).
253. Najah, M. *et al.* Identification of patients with abetalipoproteinemia and homozygous familial hypobetalipoproteinemia in Tunisia. *Clin. Chim. Acta Int. J. Clin. Chem.* **401**, 51–56 (2009).
254. Funghini, S. *et al.* Carbamoyl phosphate synthetase 1 deficiency in Italy: clinical and genetic findings in a heterogeneous cohort. *Gene* **493**, 228–234 (2012).

255. Lee, S.-T., Lee, J., Lee, M., Kim, J.-W. & Ki, C.-S. Clinical and genetic analysis of Korean patients with congenital insensitivity to pain with anhidrosis. *Muscle Nerve* **40**, 855–859 (2009).
256. Pagani, F. & Baralle, F. E. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat. Rev. Genet.* **5**, 389–396 (2004).
257. Wadt, K. *et al.* A cryptic BAP1 splice mutation in a family with uveal and cutaneous melanoma, and paraganglioma. *Pigment Cell Melanoma Res.* **25**, 815–818 (2012).
258. Titeux, M. *et al.* Recessive dystrophic epidermolysis bullosa caused by COL7A1 hemizygoty and a missense mutation with complex effects on splicing. *Hum. Mutat.* **27**, 291–292 (2006).
259. Hertecant, J. L. *et al.* Clinical and molecular analysis of isovaleric acidemia patients in the United Arab Emirates reveals remarkable phenotypes and four novel mutations in the IVD gene. *Eur. J. Med. Genet.* **55**, 671–676 (2012).
260. Kang, D.-H. *et al.* Identification of a novel splicing mutation in the ARSA gene in a patient with late-infantile form of metachromatic leukodystrophy. *Korean J. Lab. Med.* **30**, 516–520 (2010).
261. Bolisetty, M. T. & Beemon, K. L. Splicing of internal large exons is defined by novel cis-acting sequence elements. *Nucleic Acids Res.* **40**, 9244–9254 (2012).
262. Di Leo, E. *et al.* Abnormal apolipoprotein B pre-mRNA splicing in patients with familial hypobetalipoproteinaemia. *J. Med. Genet.* **44**, 219–224 (2007).

Chapter 3

3 A unifying framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer

The work in this chapter has been submitted for publication as:

Mucaki EJ*, Caminsky NG*, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JHM, Rogan PK. A unifying framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. *BMC Medical Genomics* (2015).

3.1 Background

Advances in NGS have enabled panels of genes, whole exomes, and even whole genomes to be sequenced for multiple individuals in parallel. These platforms have become so cost-effective and accurate that they are beginning to be adopted in clinical settings, as evidenced by recent FDA approvals^{1,2}. However, the overwhelming number of gene variants revealed in each individual has challenged interpretation of clinically significant genetic variation³⁻⁵.

After common variants, which are rarely pathogenic, are eliminated, the number of VUS in the residual set remains substantial. Assessment of pathogenicity is not trivial, considering that nearly half of the unique variants are novel, and cannot be resolved using published literature and variant databases⁶. Furthermore, loss-of-function variants (those resulting in protein truncation are most likely to be deleterious) represent a very small proportion of identified variants. The remaining variants are missense and synonymous

*EJM and NGC should be considered to be joint first authors.

variants in the exon, single nucleotide changes, or in frame insertions or deletions in intervening and intergenic regions. Functional analysis of large numbers of these variants often cannot be performed, due to lack of relevant tissues, and the cost, time, and labor required for each variant. Another problem is that *in silico* protein coding prediction tools exhibit inconsistent accuracy and are thus problematic for clinical risk evaluation⁷⁻⁹. Consequently, 90% of HBOC patients receiving genetic susceptibility testing will receive an inconclusive or uncertain result¹⁰.

One strategy to improve variant interpretation in patients is to reduce the full set of variants to a manageable list of potentially pathogenic variants. Evidence for pathogenicity of VUS in genetic disease is often limited to amino acid coding changes^{11,12}, and mutations affecting splicing, transcription activation, and mRNA stability tend to be underreported¹³⁻¹⁹. Splicing errors are estimated to represent 15% of disease-causing mutations²⁰, but may be much higher^{21,22}. The impact of a single nucleotide change in a recognition sequence can range from insignificant to complete abolition of a protein binding site. The complexity of interpretation of non-coding sequence variants benefits from computational approaches²³ and direct functional analyses²⁴⁻²⁸ that may each support evidence of pathogenicity.

Ex vivo transfection assays developed to determine the pathogenicity of VUS predicted to lead to splicing aberrations (using *in silico* tools) have been successful in identifying pathogenic sequence variants^{29,30}. IT-based analysis of splicing variants has proven to be robust and accurate at analyzing splice site (SS) variants, including splicing regulatory factor binding sites (SRFBSs), and in distinguishing them from polymorphisms in both rare and common diseases³¹. However, IT can be applied to any sequence recognized and bound by another factor³², such as with transcription factor binding sites (TFBSs) and RNA-binding protein binding sites (RBBSs). IT is used as a measure of sequence conservation and is more accurate than consensus sequences³³. The individual information (R_i) of a base is related to thermodynamic entropy, and therefore free energy of binding, and is measured on a logarithmic scale (in bits). By comparing the change in information (ΔR_i) for a nucleotide variation of a bound sequence, the resulting change in

binding affinity is $\geq 2^{\Delta Ri}$, such that a 1 bit change in information will result in at least a 2-fold change in binding affinity³⁴.

IT measures nucleotide sequence conservation and does not provide information on effects of variants on mRNA secondary (2°) structure, nor can it accurately predict effects of amino acid sequence changes. Other *in silico* methods have attempted to address these deficiencies. For example, Halvorsen *et al.* (2010) introduced an algorithm called SNPfold, which computes the potential effect of a single nucleotide variant (SNV) on mRNA 2° structure¹⁵. Predictions made by SNPfold can be tested by the SHAPE assay (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension)³⁵, which provides evidence for sequence variants that lead to structural changes in mRNA by detection of covalent adducts in mRNA.

The ramifications for better interpretation of VUS are particularly relevant for HBOC³⁶. Although linkage studies suggest approximately 85% of high-risk families have deleterious variants in *BRCA1* and *BRCA2*, less than half have identified pathogenic mutations³⁷. This implies that deleterious variants lie in untested regions of *BRCA1/2*, untested genes, or are unrecognized^{38,39}. Consequently, VUS in *BRCA1/2* greatly outnumber known deleterious mutations⁴⁰.

Here, we develop and evaluate IT-based models to predict potential non-coding sequence mutations in SSs, TFBSs, and RBBSs in 7 genes sequenced in their entirety in 102 HBOC patients who did not exhibit known *BRCA1/2* coding mutations at the time of initial testing. The genes are: *ATM* (Ataxia Telangiectasia Mutated), *BRCA1* (Breast Cancer 1, Early onset), *BRCA2* (Breast Cancer 2, Early onset), *CDH1* (Cadherin 1, Type 1, E-Cadherin), *CHEK2* (Checkpoint Kinase 2), *PALB2* (Partner and Localizer of BRCA2), and *TP53* (Tumour Protein P53), and have been reported to harbor mutations that increase HBOC risk⁴¹⁻⁶³. We apply these IT-based methods to analyze variants in the complete sequences of coding, non-coding, and up- and downstream regions of the 7 genes. In this study, we established and applied a unified IT-based framework, first filtering out common variants, then to “flag” potentially deleterious ones. Then, using

context-specific criteria and information from the published literature, we prioritized likely candidates.

3.2 Methods

3.2.1 Design of Tiled Capture Array for HBOC Gene Panel

Nucleic acid hybridization capture reagents designed from genomic sequences generally avoid repetitive sequence content to avoid cross hybridization⁶⁴. Complete gene sequences harbor numerous repetitive sequences, and an excess of denatured C₀t-1 DNA is usually added to hybridization to prevent inclusion of these sequences⁶⁵. RepeatMasker software completely masks all repetitive and low-complexity sequences⁶⁶. We increased sequence coverage in complete genes with capture probes by enriching for both single-copy and divergent repeat (> 30% divergence) regions, such that, under the correct hybridization and wash conditions, all probes hybridize only to their correct genomic locations⁶⁴. This step was incorporated into a modified version of Gnirke and colleagues' (2009) in-solution hybridization enrichment protocol, in which the majority of library preparation, pull-down, and wash steps were automated using a BioMek[®] FXP Automation Workstation (Beckman Coulter, Mississauga, Canada)⁶⁷.

Genes *ATM* (RefSeq: NM_000051.3, NP_000042.3), *BRCAl* (RefSeq: NM_007294.3, NP_009225.1), *BRCA2* (RefSeq: NM_000059.3, NP_000050.2), *CDH1* (RefSeq: NM_004360.3, NP_004351.1), *CHEK2* (RefSeq: NM_145862.2, NP_665861.1), *PALB2* (RefSeq: NM_024675.3, NP_078951.2), and *TP53* (RefSeq: NM_000546.5, NP_000537.3) were selected for capture probe design by targeting single copy or highly divergent repeat regions (spanning 10 kb up- and downstream of each gene relative to the most upstream first exon and most downstream final exon in RefSeq) using an *ab initio* approach⁶⁴. If a region was excluded by *ab initio* but lacked a conserved repeat element (i.e. divergence > 30%)⁶⁶, the region was added back into the probe-design sequence file. Probe sequences were selected using PICKY 2.2 software⁶⁸. These probes were used in solution hybridization to capture our target sequences, followed by NGS on an Illumina Genome Analyzer IIx (**Supplementary Methods – Appendix E**).

Genomic sequences from both strands were captured using overlapping oligonucleotide sequence designs covering 342,075 nt among the 7 genes (**Figure 3.1**). In total, 11,841 oligonucleotides were synthesized from the transcribed strand consisting of the complete, single copy coding, and flanking regions of *ATM* (3,513), *BRCA1* (1,587), *BRCA2* (2,386), *CDH1* (1,867), *CHEK2* (889), *PALB2* (811), and *TP53* (788). Additionally, 11,828 antisense strand oligos were synthesized (3,497 *ATM*, 1,591 *BRCA1*, 2,395 *BRCA2*, 1,860 *CDH1*, 883 *CHEK2*, 826 *PALB2*, and 776 *TP53*).

For regions lacking probe coverage (of ≥ 10 nt, N=141; 8 in *ATM*, 26 in *BRCA1*, 10 in *BRCA2*, 29 in *CDH1*, 36 in *CHEK2*, 15 in *PALB2*, and 17 in *TP53*), probes were selected based on predicted T_m s similar to other probes, limited alignment to other sequences in the transcriptome (< 10 times), and avoidance of stable, base-paired 2° structures (with *unaFOLD*)^{69,70}. The average coverage of these sequenced regions was 14.1-24.9% lower than other probe sets, indicating that capture was less efficient, though still successful.

3.2.2 HBOC Samples for Oligo Capture and High-Throughput Sequencing

Genomic DNA used in prior susceptibility testing, from 102 anonymized (only gender and age of onset were provided) patients was received from the Molecular Genetics Laboratory (MGL) at the London Health Sciences Centre in London, Ontario, Canada. Patients qualified for genetic susceptibility testing as determined by the Ontario Ministry of Health and Long-Term Care *BRCA1* and *BRCA2* genetic testing criteria⁷¹ (see **Table 3.1**). *BRCA1* and *BRCA2* were previously analyzed by Protein Truncation Test (PTT) and Multiplex Ligation-dependent Probe Amplification (MLPA). The exons of several patients (N=14) had also been Sanger sequenced. No pathogenic sequence change was found in any of these individuals. In addition, one patient with a known pathogenic *BRCA* variant was re-sequenced by NGS as a positive control.

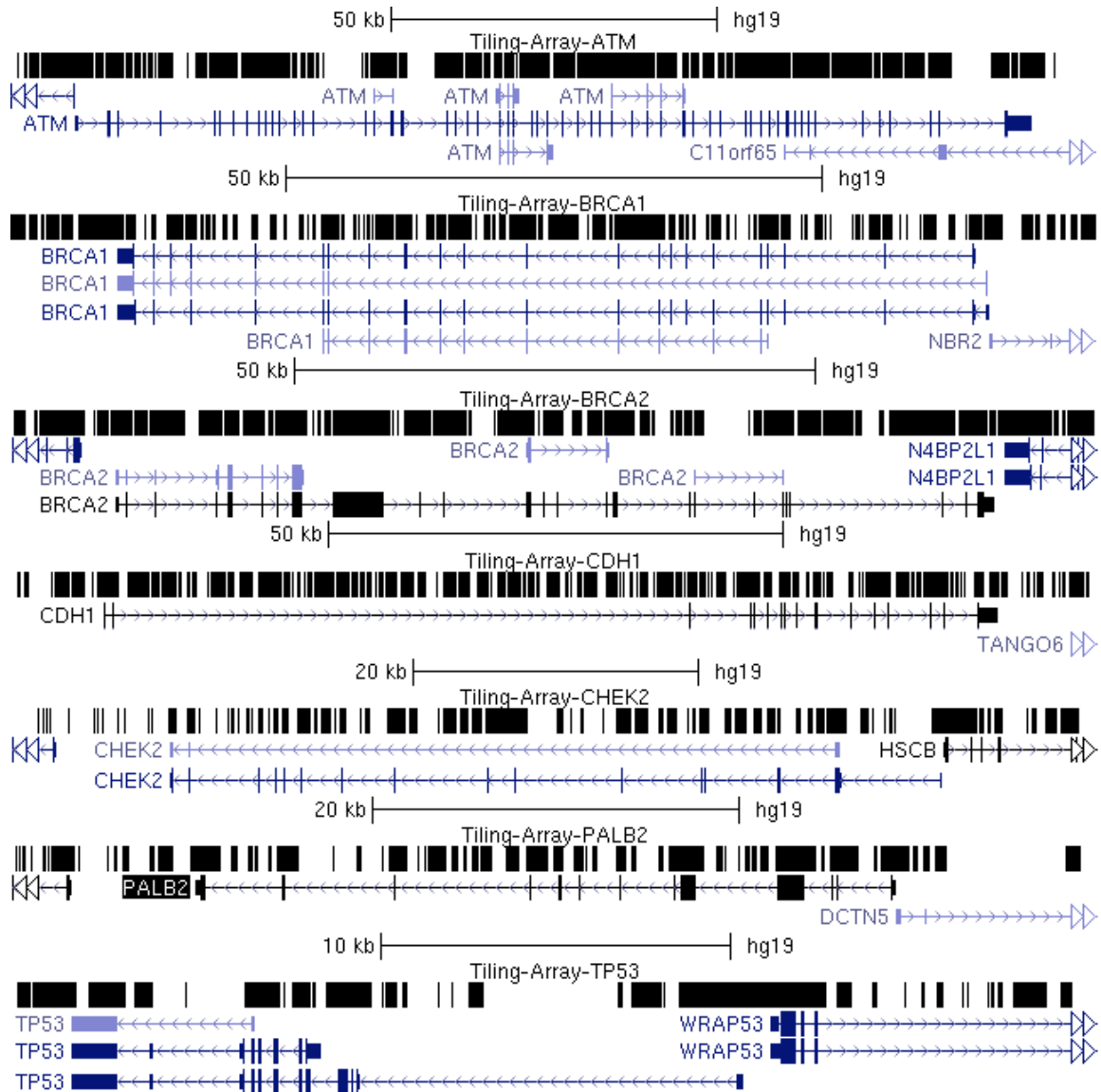


Figure 3.1. Capture Probe Coverage over Sequenced Genes

The genomic structure of the 7 genes chosen are displayed with the UCSC Genome Browser. Top row for each gene is a custom track with the “dense” visualization modality selected with black regions indicating the intervals covered by oligonucleotide capture reagent. Regions without probe coverage contain conserved repetitive sequences or correspond to paralogous sequences that are unsuitable for probe design.

Table 3.1. Risk Categories for Individuals Eligible for Screening for a Genetic Susceptibility to Breast or Ovarian Cancers as determined by the Ontario Ministry of Health and Long Term-Care Referral Criteria for Genetic Counseling

Category	Description
1	Ashkenazi Jewish and BC <50 years, or OC at any age
2	BC <35 years of age
3	Male BC
4	Invasive serous ovarian cancer at any age
5	BC <60 year, and a 1 st or 2 nd -degree relative with OC or male BC
6	BC and OC in the same individual, or bilateral BC with the first case <50
7	Two cases of OC, both <50 years, in 1 st or 2 nd -degree relatives
8	Two cases of OC, any age, in 1 st or 2 nd -degree relatives
9	Ashkenazi Jewish and BC at any age, and any family history of BC or OC
10	Three or more cases of BC or OC at any age
11	Relative of an individual with known <i>BRCA1</i> or <i>BRCA2</i> mutation
12	Ashkenazi Jewish and 1 st or 2 nd -degree relative or individual with: BC <50 years, or OC at any age, or male BC, or BC, any age, with positive family history of BC or OC.
13	A pedigree strongly suggestive of HBOC, i.e. risk of carrying a mutation for the individual being tested is >10%

Only patients from groups 5-8, and 10 were considered for this study.

3.2.3 Sequence Alignment and Variant Calling

Variant analysis involved the steps of detection, filtering, IT-based and coding sequence analysis, and prioritization (**Figure 3.2**). Sequencing data were demultiplexed and aligned to the specific chromosomes of our sequenced genes (hg19) using both CASAVA (Consensus Assessment of Sequencing and Variation; v1.8.2)⁷² and CRAC (Complex Reads Analysis and Classification; v1.3.0)⁷³ software. Alignments were prepared for variant calling using Picard⁷⁴ and variant calling was performed on both versions of the aligned sequences using the UnifiedGenotyper tool in the Genome Analysis Toolkit (GATK)⁷⁵. We used the recommended minimum phred base quality score of 30, and results were exported in variant call format (VCF; v4.1). A software program was developed to exclude variants called outside of targeted capture regions and those with quality scores < 50. Variants flagged by bioinformatic analysis (described below) were also assessed by manually inspecting the reads in the region using the Integrative Genomics Viewer (IGV; version 2.3)^{76,77} to note and eliminate obvious false positives (i.e. variant called due to polyhomonucleotide run dephasing, or PCR duplicates that were not eliminated by Picard). Finally, common variants ($\geq 1\%$ allele frequency based on dbSNP142 or > 5 individuals in our study cohort) were eliminated.

3.2.4 IT-Based Variant Analysis

All variants were analyzed using the Shannon Human Splicing Mutation Pipeline, a genome-scale variant analysis program that predicts the effects of variants on mRNA splicing^{78,79}. Variants were flagged based on criteria reported in Shirley *et al.* (2013): weakened natural site ≥ 1.0 bits, or strengthened cryptic site (within 300 nt of the nearest exon) where cryptic site strength is equivalent or greater than the nearest natural site of the same phase⁷⁸. The effects of flagged variants were further analyzed in detail using the Automated Splice Site and Exon Definition Analysis (ASSEDA) server⁸⁰.

Exonic variants and those found within 500 nt of an exon were assessed for their effects, if any, on SRFBSs⁸⁰. Sequence logos for splicing regulatory factors (SRFs) (SRSF1, SRSF2, SRSF5, SRSF6, hnRNPH, hnRNPA1, ELAVL1, TIA1, and PTB) and their

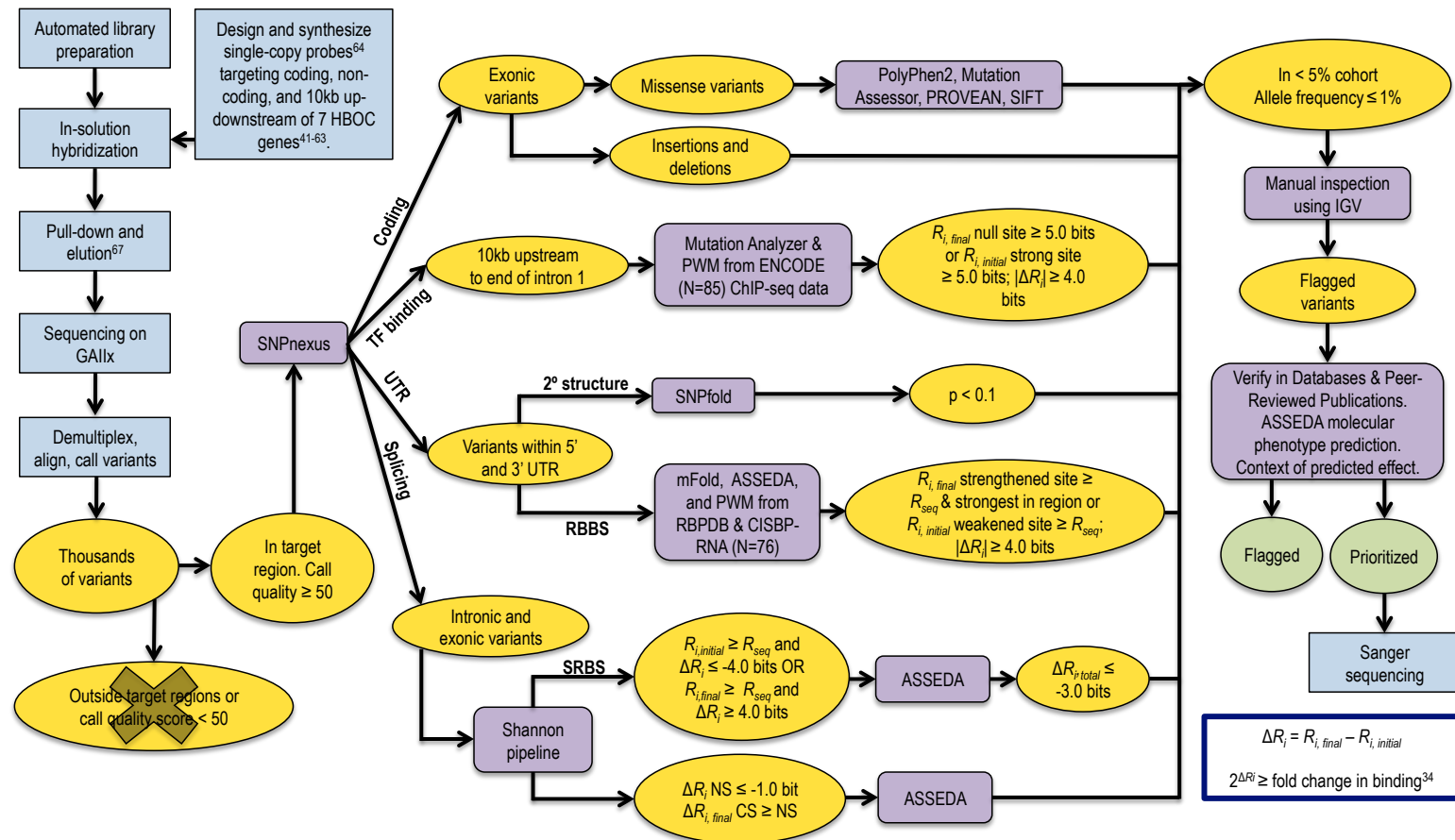


Figure 3.2. Framework for the Identification of Potentially Pathogenic Variants

Integrated laboratory processing and bioinformatic analysis procedures for comprehensive complete gene variant determination and analysis. Intermediate datasets resulting from filtering are represented in yellow and final datasets in green.

Non-bioinformatic steps, such as sample preparation are represented in blue and prediction programs in purple. Sequencing analysis yields base calls for all samples. CASAVA⁷² and CRAC⁷³ were used to align these sequencing results to HG19. GATK⁷⁵ was used to call variants from this data against GRCh37 release of the reference human genome. Variants with a quality score < 50 and/or call confidence score < 30 were eliminated along with variants falling outside of our target regions. SNPnexus¹⁰²⁻¹⁰⁴ was used to identify the genomic location of the variants. Nonsense and indels were noted and prediction tools were used to assess the potential pathogenicity of missense variants. The Shannon Pipeline⁷⁸ evaluated the effect of a variant on natural and cryptic SSs, as well as SRFBSs. ASSEDA⁸⁰ was used to predict the potential isoforms as a result of these variants. PWMs for 83 TFs were built using an information weight matrix generator based on Bipad⁹⁵. Mutation Analyzer evaluated the effect of variants found 10 kb upstream up to the first intron on protein binding. Bit thresholds (R_i values) for filtering variants on software program outputs are indicated. Variants falling within the UTR sequences were assessed using SNPfold¹⁵, and the most probable variants that alter mRNA structure ($p < 0.1$) were then processed using mfold to predict the effect on stability⁷⁰. All UTR variants were scanned with a modified version of the Shannon Pipeline, which uses PWMs computed from nucleotide frequencies for 28 RBPs in RBPDB⁹⁹ and 76 RBPs in CISBP-RNA¹⁰⁰. All variants meeting these filtering criteria were verified with IGV^{76,77}. Sanger sequencing was only performed for protein truncating, splicing, and selected missense variants.

$R_{sequence}$ values (the mean information content⁸¹) are provided in Caminsky *et al.* (2015)³¹. Because these motifs occur frequently in unspliced transcripts, only variants with large information changes were flagged, notably those with (a) ≥ 4.0 bit decrease, i.e. at least a 16-fold reduction in binding site affinity, with $R_{i,initial} \geq R_{sequence}$ for the particular factor analyzed, or (b) ≥ 4.0 bit increase in a site where $R_{i,final} \geq 0$ bits. ASSEDA was used to calculate $R_{i,total}$, with the option selected to include the given SRF in the calculation. Variants decreasing $R_{i,total}$ by < 3.0 bits (i.e. 8-fold) were predicted to potentially have benign effects on expression, and were not considered further.

Activation of pseudoexons through creating/strengthening of an intronic cryptic splice site was also assessed⁸². Changes in intronic cryptic sites, where $\Delta R_i > 1$ bit and $R_{i,final} \geq (R_{sequence} - 1 \text{ standard deviation [S.D.] of } R_{sequence})$, were identified. A pseudoexon was predicted if a pre-existing cryptic site of opposite polarity (with $R_i > [R_{sequence} - 1 \text{ S.D.}]$) and in the proper orientation for formation of exons between 10-250 nt in length was present. In addition, the minimum intronic distance between the pseudoexon and either adjacent natural exon was 100 nt. The acceptor site of the pseudoexon was also required to have a strong hnRNPA1 site located within 10 nt ($R_i \geq R_{sequence}$)⁸⁰ to ensure accurate proofreading of the exon⁸³.

Next, variants affecting the strength of SRFs were analyzed by a contextual exon definition analysis of $\Delta R_{i,total}$. The context refers to the documented splicing activity of an SRF. For example, TIA1 has been shown to be an intronic enhancer of exon definition, so only intronic sites were considered. Similarly, hnRNPA1 proofreads the 3' SS (acceptor) and inhibits exon recognition elsewhere⁸⁴. Variants that lead to redundant SRFBS changes (i.e. one site is abolished and another proximate site [≤ 2 nt] of equivalent strength is activated) were assumed to have a neutral effect on splicing. If the strength of a site bound by PTB (polypyrimidine tract binding protein) was affected, its impact on binding by other factors was analyzed, as PTB impedes binding of other factors with overlapping recognition sites, but does not directly enhance or inhibit splicing itself⁸⁵.

To determine effects of variants on transcription factor (TF) binding, we first established which TFs bound to the sequenced regions of the gene promoters (and first exons) in this study by using ChIP-seq data from 125 cell types (**Supplementary Methods**)⁸⁶. We identified 141 TFs with evidence for binding to the promoters of the genes we sequenced, including c-Myc, C/EBP β , and Sp1, shown to transcriptionally regulate *BRCA1*, *TP53*, and *ATM*, respectively⁸⁷⁻⁸⁹. Furthermore, polymorphisms in TCF7L2, known to bind enhancer regions of a wide variety of genes in a tissue-specific manner⁹⁰, have been shown to increase risk of sporadic⁹¹ and hereditary breast⁹², as well as other types of cancer^{93,94}.

IT-based models of the 141 TFs of interest were derived by entropy minimization of the DNase accessible ChIP-seq subsets⁹⁵. The details of this approach have been described in a manuscript submitted by our laboratory for publication and are not provided in this thesis. While some data sets would only yield noise or co-factor motifs (i.e. co-factors that bind via tethering, or histone modifying proteins⁹⁶), techniques such as motif masking and increasing the number of Monte Carlo cycles yielded models for 83 TFs resembling each factor's published motif. **Supplementary Table 11** (all supplementary tables are provided in the **Supplementary Content File**) contains the final list of TFs and the models we built (described below)⁹⁷.

These TFBS models (N=83) were used to scan all variants called in the promoter regions (10 kb upstream of transcriptional start site to the end of IVS1) of HBOC genes for changes in R_i ⁹⁸. Binding site changes that weaken interactions with the corresponding TF (to $R_i \leq R_{sequence}$) are likely to affect regulation of the adjacent target gene. Stringent criteria were used to prioritize the most likely variants and thus only changes to strong TFBSs ($R_{i,initial} \geq R_{sequence}$), where reduction in strength was significant ($\Delta R_i \geq 4.0$ bits), were considered. Alternatively, novel or strengthened TFBSs were also considered sources of dysregulated transcription. These sites were defined as having $R_{i,final} \geq R_{sequence}$ and as being the strongest predicted site in the corresponding genomic interval (i.e. exceeding the R_i values of adjacent sites unaltered by the variant). Variants were not

prioritized if the TF was known to a) enhance transcription and IT analysis predicted stronger binding, or b) repress transcription and IT analysis predicted weaker binding.

Two complementary strategies were used to assess the possible impact of variants within UTRs. First, SNPfold software was used to assess the effect of a variant on 2° structure of the UTR (**Supplementary Methods**)¹⁵. Variants flagged by SNPfold with the highest probability of altering stable 2° structures in mRNA (where p-value < 0.1) were prioritized. To evaluate these predictions, oligonucleotides containing complete wild-type and variant UTR sequences (**Supplementary Table 13**) were transcribed *in vitro* and followed by SHAPE analysis, a method that can confirm structural changes in mRNA³⁵.

Second, the effects of variants on the strength of RBBSs were predicted. Frequency-based, position weight matrices (PWMs) for 156 RNA-binding proteins (RBPs) were obtained from the RNA-Binding Protein DataBase (RBPDB)⁹⁹ and the Catalog of Inferred Sequence Binding Preferences of RNA binding proteins (CISBP-RNA)^{100,101}. These were used to compute information weight matrices (based on the method described by Schneider *et al.* 1984; N = 147) (see **Supplementary Methods**)³². All UTR variants were assessed using a modified version of the Shannon Pipeline⁷⁸ containing the RBPDB and CISBP-RNA models. Results were filtered to include a) variants with $|\Delta R_i| \geq 4.0$ bits, b) variants creating or strengthening sites ($R_{i,final} \geq R_{sequence}$ and the $R_{i,initial} < R_{sequence}$), and c) RBBSs not overlapping or occurring within 10 nt of a stronger, pre-existing site of another RBP.

3.2.5 Exonic Protein-Altering Variant Analysis

The predicted effects of all coding variants were assessed with SNPnexus^{102–104}, an annotation tool that can be applied to known and novel variants using up-to-date dbSNP and UCSC human genome annotations. Variants predicted to cause premature protein truncation were given higher priority than those resulting in missense (or synonymous) coding changes. Missense variants were first cross referenced with dbSNP142¹⁰⁵. Population frequencies from the Exome Variant Server¹⁰⁶ and 1000Genomes¹⁰⁷ are also provided. The predicted effects on protein conservation and function of the remaining

variants were evaluated by *in silico* tools: PolyPhen-2¹⁰⁸, Mutation Assessor (release 2)^{109,110}, and PROVEAN (v1.1.3)^{111,112}. Default settings were applied and in the case of PROVEAN, the “PROVEAN Human Genome Variants Tool” was used, which includes SIFT predictions as a part of its output. Variants predicted by all four programs to be benign were less likely to have a deleterious impact on protein activity; however this did not exclude them from mRNA splicing analysis (described above in *IT-Based Variant Analysis*). All rare and novel variants were cross-referenced with general mutation databases (ClinVar^{113,114}, Human Gene Mutation Database [HGMD]^{115,116}, Leiden Open Variant Database [LOVD]¹¹⁷⁻¹²⁴, Domain Mapping of Disease Mutations [DM²]¹²⁵, Expert Protein Analysis System [ExPASy]¹²⁶ and UniProt^{127,128}), and gene-specific databases (*BRCA1/2*: the Breast Cancer Information Core database [BIC]¹²⁹ and Evidence-based Network for the Interpretation of Germline Mutant Alleles [ENIGMA]¹³⁰; *TP53*: International Agency for Research on Cancer [IARC]¹³¹), as well as published reports to prioritize them for further workup.

3.2.6 Variant Classification

Flagged variants were prioritized if they were likely to encode a dysfunctional protein (indels, nonsense codon > 50 amino acids from the C-terminus, or abolition of a natural SS resulting in out-of-frame exon skipping) or if they exceeded established thresholds for fold changes in binding affinity based on IT (see *Methods* above). If previous studies performed functional or pedigree analyses, allowing to categorize a variant as pathogenic or benign, this superseded our analysis.

3.2.7 Positive control

We identified the *BRCA1* exon 17 nonsense variant c.5136G>A (chr17:41215907C>T; rs80357418; 2-5A)¹³² in the sample that was provided as a positive control. This was the same mutation identified by the MGL as pathogenic for this patient. We also prioritized another variant in this patient (**Table 3.2**)¹³³.

Table 3.2. Prioritized Variants in the Positive Control

Gene	mRNA Protein	rsID (dbSNP142) Allele Frequency (%) [†]	Category	Consequence	Ref
<i>BRCA1</i>	c.5136G>A p.Trp1712Ter	rs80357418	Nonsense	151 AA short	132
<i>BRCA2</i>	c.3218A>G p.Gln1073Arg	rs80358566	Missense	Listed in ClinVar as conflicting interpretations (likely benign, unknown) and in BIC as unknown clinical importance. 2 <i>in silico</i> programs called deleterious.	133
			SRFBS	Repressor action of hnRNPA1 at this site abolished (5.2 to 0.4 bits). Blocking action of PTB removed as site is abolished (5.5 to -7.5 bits) and may uncover binding sites of other SRFs.	

[†] If available. Positive control was sample 2-5A.

3.2.8 Variant Validation

Protein-truncating, prioritized splicing, and selected prioritized missense variants were verified by Sanger sequencing. Primers of PCR amplicons are indicated in **Supplementary Table 14**).

3.3 Results

3.3.1 Capture, Sequencing, and Alignment

The average coverage of capture region per individual was 90.8x (range of 53.8 to 118.2x between 32 samples) with 98.8% of the probe-covered nucleotides having ≥ 10 reads. Samples with fewer than 10 reads per nucleotide were re-sequenced and the results of both runs were combined. The combined coverage of these samples was, on average, 48.2x (± 36.2).

The consistency of both library preparation and capture protocols was improved from initial runs, which significantly impacted sequence coverage (**Supplementary Methods**). Of the 102 patients tested, 14 had been previously Sanger sequenced for *BRCA1* and *BRCA2* exons. Confirmation of previously discovered SNVs served to assess the methodological improvements introduced during NGS and ultimately, to increase confidence in variant calling. Initially, only 15 of 49 SNVs in 3 samples were detected. The detection rate of SNVs was improved to 100% as the protocol progressed. All known SNVs (N=157) were called in subsequent sequencing runs where purification steps were replaced with solid phase reversible immobilization beads and where RNA bait was transcribed the same day as capture. To minimize false positive variant calls, sequence read data was aligned using 2 different software programs, CASAVA and CRAC, and variant calling was performed for both sets of data using GATK^{72,73,75}.

GATK called 14,164 unique SNVs and 1,147 indels. Only 3,777 (15.3%) SNVs were present in both CASAVA and CRAC-alignments for at least one patient, and even fewer indel calls were concordant between both methods (N=110; 6.2%). For all other SNVs and indels, CASAVA called 6,871 and 1,566, respectively, whereas CRAC called 13,958

and 110, respectively. Some variants were counted more than once if they are called by different alignment programs in two or more patients. Intronic and intergenic variants proximate to low complexity sequences tend to generate false positive variants due to ambiguous alignment, a well known technical issue in short read sequence analysis^{134,135}, contributing to this discrepancy. For example, in **Figure 3.3**, CRAC correctly called a 19 nt deletion of *BRCA1* (rs80359876; also confirmed by Sanger sequencing) but CASAVA flagged the deleted segment as a series of false-positives. For these reasons, all variants were manually reviewed.

3.3.2 IT-Based Variant Identification and Prioritization

3.3.2.1 Natural SS Variants

The Shannon Pipeline reported 99 unique variants in natural donor or acceptor SSs. After technical and frequency filtering criteria were applied, 12 variants remained (**Supplementary Table 15**). IT analysis allowed for the prioritization of 3 variants, summarized in **Table 3.3**.

First, the novel *ATM* variant c.3747-1G>A (chr11:108154953G>A; sample number 7-4F) abolishes the natural acceptor of exon 26 (11.0 to 0.1 bits). ASSEDA reports the presence of a 5.3 bit cryptic acceptor site 13 nt downstream of the natural site, but the effect of the variant on a pre-existing cryptic site is negligible (~0.1 bits). The cryptic exon would lead to exon deletion and frameshift (**Figure 3.4A**). ASSEDA also predicts skipping of the 246 nt exon, as the $R_{i,final}$ of the natural acceptor is now below $R_{i,minimum}$ (1.6 bits), altering the reading frame. Second, the novel *ATM* c.6347+1G>T (chr11:108188249G>T; 4-1F) occurs at the natural donor of exon 44 and abolishes the 10.4 bit donor ($\Delta R_i = -18.6$ bits), resulting exclusively in exon skipping. Finally, the previously reported *CHEK2* variant, c.320-5A>T (chr22:29121360T>A; rs121908700; 4-2B)¹³⁶ weakens the natural acceptor of exon 3 (6.8 to 4.1 bits), possibly activating a cryptic acceptor (7.4 bits) 92 nt upstream of the natural acceptor (**Figure 3.5**).

Variants either strengthening (N=4) or slightly weakening ($\Delta R_i < 1.0$ bits; N=4) a natural site were not prioritized. In addition, we rejected the *ATM* variant (c.1066-6T>G;

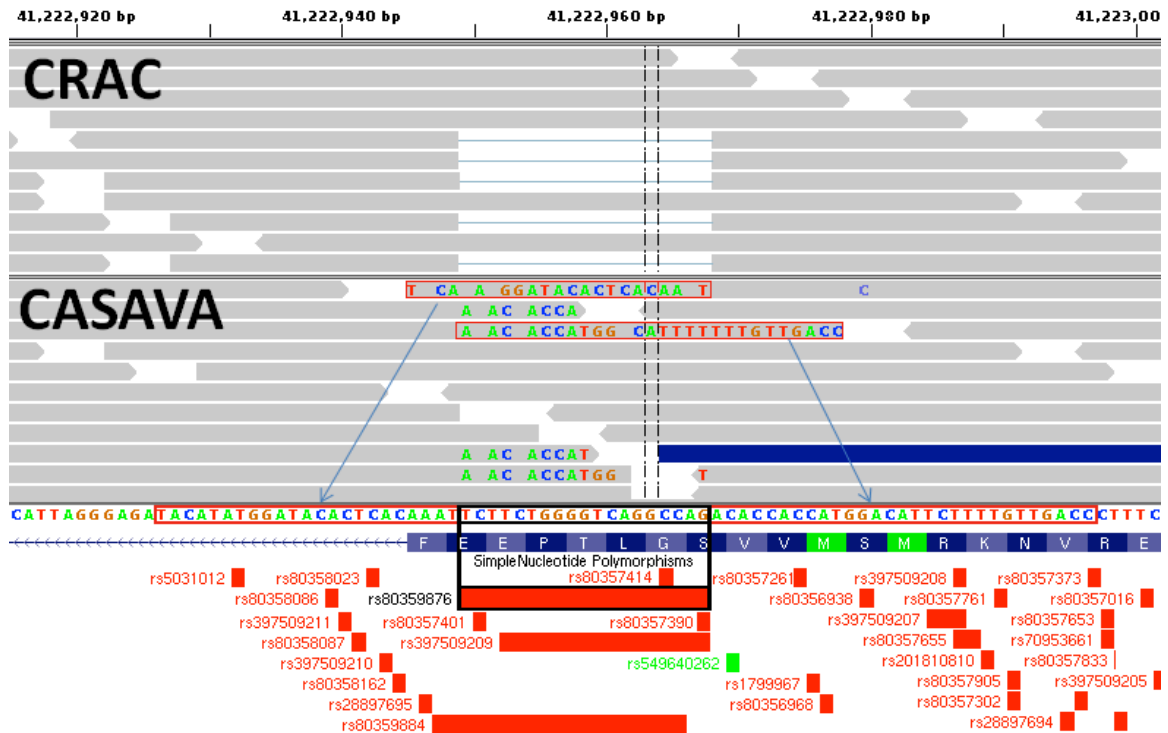


Figure 3.3. *BRCA1* Deletion Inaccurately Aligned by CASAVA

A 19 nt *BRCA1* deletion was not identified by one alignment program (CASAVA), and instead appears as a series of SNVs. The *BRCA1* deletion is still observable on IGV (middle box) as these SNVs align to two separate regions of *BRCA1* exon 15 (arrowed red boxes). This deletion leads to a frameshift starting with p.1655Ser and a premature STOP at p.1670, resulting in the loss of 193 AA.

Table 3.3. Variants Prioritized by IT Analysis

UWO ID	Gene	mRNA	rsID (dbSNP142) Allele Frequency (%)	Information Change			Consequence [‡] or Binding Factor Affected
				$R_{i,initial}$ (bits)	$R_{i,initial}$ (bits)	ΔR_i (bits)	
Abolished Natural SS							
7-4F	<i>ATM</i>	c.3747-1G>A*	Novel	11.0	0.1	-10.9	Exon skipping and use of alternative splice forms
4-1F	<i>ATM</i>	c.6347+1G>T***	Novel	10.4	-8.3	-18.6	Exon skipping
Leaky Natural SS							
4-2B	<i>CHEK2</i>	c.320-5T>A*	rs121908700 0.08	6.8	4.1	-2.7	Leaky splicing with intron inclusion
Activated Cryptic SS							
7-3E	<i>BRCA1</i>	c.548-293G>A	rs117281398 0.74	-12.1	2.6	14.7	Cryptic site not expected to be used. Total information for natural exon is stronger than cryptic exon.
7-4A	<i>BRCA2</i>	c.7618-269_7618-260del10	Novel	3.9	9.4	5.5	Cryptic site not expected to be used. Total information for natural exon is stronger than cryptic exon.

Pseudoexon formation due to activated acceptor SS

7-3F	<i>BRCA2</i>	c.8332-805G>A	Novel	-9.3	5.4	5.6	6,065/211/592
7-3D	<i>CDHI</i>	c.164-2023A>G	rs184740925 0.3	-6.6	4.3	6.5	61,236/224/1,798
5-3H	<i>CDHI</i>	c.2296-174T>A	rs565488866 0.02	7.3	8.5	5.0	1,175/50/124

Pseudoexon formation due to activated donor SS

3-6A	<i>BRCA1</i>	c.212+253G>A	rs189352191 0.08	4.1	6.7	5.2	186/63/1,250
5-2G	<i>BRCA2</i>	c.7007+2691G>A	rs367890577 0.02	4.7	7.2	7.7	2,589/103/5,272

Affected TFBSs

7-4B	<i>BRCA1</i>	c.-8895G>A	Novel	10.9	-0.2	-11.1	GATA-3 (<i>GATA3</i>)
5-3E 7-4E	<i>CDHI</i>	c.-54G>C	rs5030874 0.16	1.7	12.0	10.4	E2F-4 (<i>E2F4</i>)
5-2B	<i>PALB2</i>	c.-291C>G	rs552824227	12.1	-1.3	-13.4	GABP α (<i>GABPA</i>)

			0.1				
7-2F	<i>TP53</i>	c.-28-3132T>C	rs17882863 0.3	-6.3	10.9	17.2	RUNX3 (<i>RUNX3</i>)
4-1A	<i>TP53</i>	c.-28-1102T>C	rs113451673 0.4	5.1	12.3	7.2	E2F-4 (<i>E2F4</i>)
				8.0	12.9	4.8	Sp1 (<i>SPI</i>)
Affected RBBSs							
7-4G	<i>ATM</i>	c.-244T>A c.-744T>A c.-1929T>A c.-3515T>A	rs539948218 0.04	9.8	-19.9	-29.7	RBFOX
5-3C	<i>CDH1</i>	c.*424T>A	Novel	-20.3	9.6	29.9	SF3B4
				8.2	1.8	-6.4	CELF4
7-2E	<i>CHEK</i> ₂	c.-588G>A	rs141568342	10.9	3.7	-7.2	BX511012.1
4-3C.5-4G	<i>CHEK</i> ₂	c.-345C>T [§]	rs137853007	3.3	11.4	8.2	SF3B4
3-1A	<i>TP53</i>	c.-107T>C	rs113530090	10.5	4.5	-6.0	ELAVL1

4-1H		c.-188T>C	0.72				
4-2H		c.*1175A>C	rs78378222				
7-2F	<i>TP53</i>	c.*1376A>C	0.26	10.7	4.1	-6.6	KHDRBS1
		c.*1464A>C					

*Confirmed by Sanger sequencing; ***Ambiguous Sanger sequencing results; §Prioritized under missense and was therefore verified with Sanger sequencing. Variant was confirmed; †If available; ‡Consequences for pseudoexon formation describe how the intron is divided: “new intron A length/pseudoexon length/new exon B length.

None of the variants have been previously reported by other groups with the exception of *CHEK2* c.320-5T>A¹³⁶.

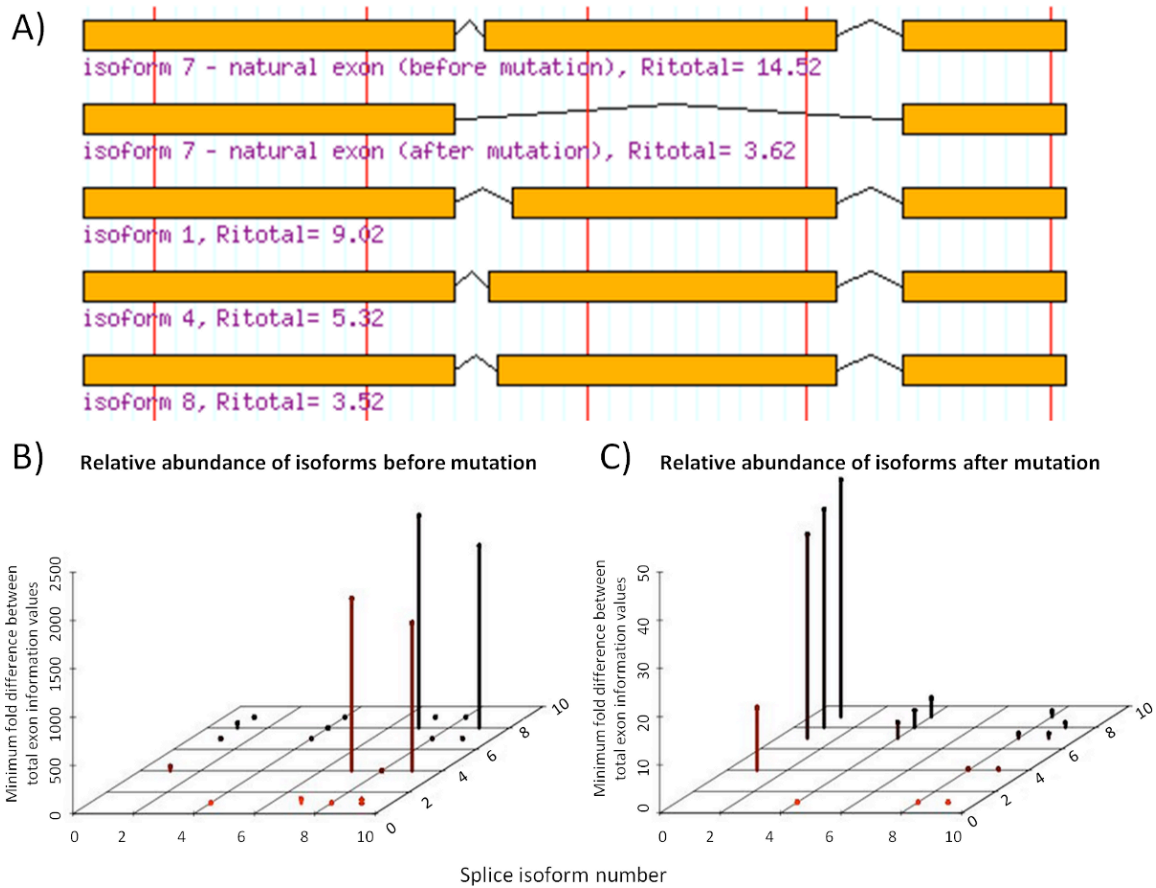


Figure 3.4. Predicted Isoforms and Relative Abundances as a Consequence of *ATM* splice variant c.3747-1G>A

Intronic *ATM* variant c.3747-1G>A abolishes (11.0 to 0.1 bits) the natural acceptor of exon 26 (total of 63 exons). **A)** ASSEDA reports the abolition of the natural exon ($R_{i,total}$ reduced from 14.5 to 3.6 bits) and predicts exon skipping as a result (isoform 7 after mutation) and/or the use of a cryptic site 13 nt downstream ($R_{i,total}$ for cryptic exon = 9.0 bits) of the natural site leading to exon deletion (isoform 1). The other isoforms use weak, alternate acceptor/donor sites leading to cryptic exons with much lower total information. **B)** Before the mutation, isoform 7 is expected to be the most abundant splice form. **C)** After the mutation, isoform 1 is predicted to become the most abundant splice form and the wild-type isoform is not expected to be expressed.

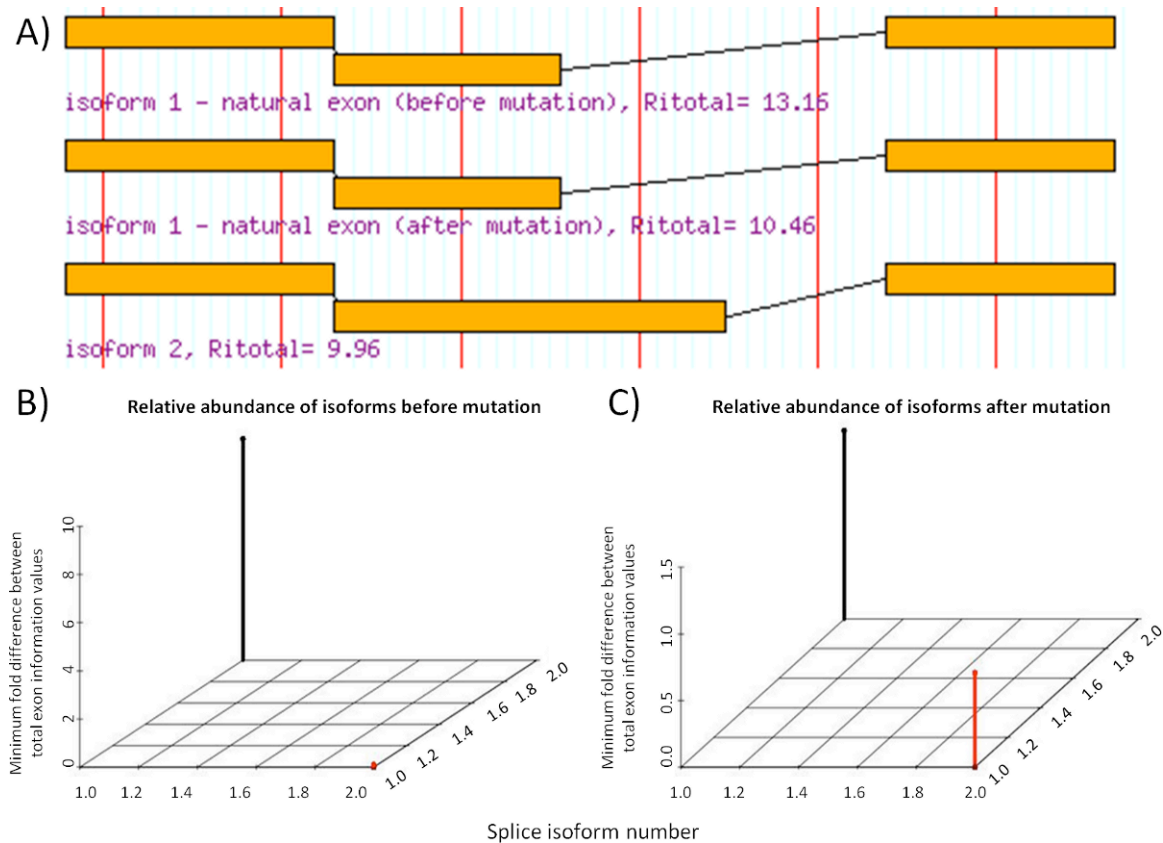


Figure 3.5. Predicted Isoforms and Relative Abundances as a Consequence of *CHEK2* splice variant c.320-5T>A

Intronic *CHEK2* variant c.320-5T>A weakens (6.8 to 4.1 bits) the natural acceptor of exon 3 (total of 15 exons). **A)** ASSEDA reports the weakening of the natural exon strength ($R_{i,total}$ reduced from 13.2 to 10.5 bits), which would result in reduced splicing of the exon otherwise known as leaky splicing. A pre-existing cryptic acceptor exists 92 nt upstream of the natural site, leading to a cryptic exon with similar strength to the mutated exon ($R_{i,total} = 10.0$ bits). This cryptic exon would contain 92 nt of the intron. **B)** Before the mutation, isoform 1 is expected to be the only isoform expressed. **C)** After the mutation, isoform 1 (wild-type) is predicted to become relatively less abundant and isoform 2 is expected to be expressed, although less abundant in relation to isoform 1.

chr11:108119654T>G; 4-1E and 7-2B), which slightly weakens the natural acceptor of exon 9 (11.0 to 8.1 bits). Although other studies have shown leaky expression as a result of this variant¹³⁷, a more recent meta-analysis concluded that this variant is not associated with increased breast cancer risk¹³⁸.

3.3.2.2 Cryptic SS Activation

Two variants produced information changes that could potentially impact cryptic splicing, but were not prioritized for the following reasons (**Table 3.3**). The first variant, novel *BRCA2* deletion c.7618-269_7618-260del10 (chr13:32931610_32931619del10; 7-4A) strengthens a cryptic acceptor site 245 nt upstream from the natural acceptor of exon 16 ($R_{i,final} = 9.4$ bits, $\Delta R_i = 5.5$ bits). Being 5.7-fold stronger than the natural site (6.9 bits), two potential cryptic isoforms were predicted, however, the exon strengths of both are weaker than the unaffected natural exon ($R_{i,total} = 6.6$ bits) and neither were prioritized. The larger gap surprisal penalties explain the differences in exon strength. The natural donor SS may still be used in conjunction with the abovementioned cryptic SS, resulting in an exon with $R_{i,total} = 3.5$ bits. Alternatively, the cryptic site and a weak donor site 180 nt upstream of the natural donor ($R_i = 0.7$ vs 1.4, cryptic and natural donors, respectively), result in an exon with $R_{i,total} = 6.5$ bits. The second variant, *BRCA1* c.548-293G>A (chr17:41249599C>T; 7-3E), creates a weak cryptic acceptor ($R_{i,final} = 2.6$ bits, $\Delta R_i = 6.2$ bits) 291 nt upstream of the natural acceptor for exon 8 ($R_i = 0.5$). Although the cryptic exon is strengthened (final $R_{i,total} = 6.9$ bits, $\Delta R_i = 14.7$ bits), ASSEDA predicts the level of expression of this exon to be negligible, as it is weaker than the natural exon ($R_{i,total} = 8.4$ bits) due to the increased length of the predicted exon (+291 nt)⁸⁰.

3.3.2.3 Pseudoexon Formation

The Shannon Pipeline initially reported 1,583 unique variants creating or strengthening intronic cryptic sites. We prioritized 5 variants, 1 of which is novel (*BRCA2* c.8332-805G>A; 7-3F), that were within 250 nt of a pre-existing complementary cryptic site and

have an hnRNPA1 site within 5 nt of the acceptor (**Table 3.3**). If used, 3 of these pseudoexons would lead to a frameshifted transcript.

3.3.2.4 SRF Binding

Variants within 500 nt of an exon junction and all exonic variants (N = 4,015) were investigated for their potential effects on affinity of sites to corresponding SRFs⁸⁰. IT analysis flagged 54 variants significantly altering the strength of at least one binding site (**Supplementary Table 16**). A careful review of the variants, the factor affected, and the position of the binding site relative to the natural SS, prioritized 36 variants (21 novel), of which 4 are in exons and 32 are in introns.

3.3.2.5 TF Binding

We assessed SNVs with models of 83 TFs experimentally shown to bind (**Supplementary Table 12**) upstream or within the first exon and intron of our sequenced genes (N=2,177). Thirteen variants expected to significantly affect TF binding were flagged (**Supplementary Table 17**). The final filtering step considered the known function of the TF in transcription, resulting in 5 prioritized variants (**Table 3.3**) in 6 patients (one variant was identified in two patients). Four of these variants have been previously reported (rs5030874, rs552824227, rs17882863, rs113451673) and one is novel (c.-8895G>A; 7-4B).

3.3.2.6 UTR Structure and Protein Binding

There were 364 unique UTR variants found by sequencing, which includes splice forms with alternate UTRs (in *BRCA1* and *TP53*). These variants were evaluated for their effects on mRNA 2° structure through SNPfold, resulting in 5 flagged variants (**Table 3.4**), all of which have been previously reported.

Analysis of three variants using mfold⁷⁰ revealed likely changes to the UTR structure (**Figure 3.6**). Two variants with possible 2° structure effects were common (*BRCA2* c.-52A>G [N=26 samples] and c.*532A>G [N=40]) and not prioritized. The 5'UTR *CDHI* variant c.-71C>G (chr16:68771248C>G; rs34033771; 7-4C) disrupts a double-stranded

Table 3.4. Variants Predicted by SNPfold to Affect UTR Structure

Class [‡]	UWO ID	Gene	mRNA	UTR position	rsID (dbSNP142) Allele Frequency (%) [†]	Rank [§]	<i>p</i> -value
F	In 26 patients	<i>BRCA2</i> [§]	c.-52A>G	5' UTR	rs206118 14.86	2/900	0.002
F	In 40 patients	<i>BRCA2</i> [§]	c.*532A>G	3' UTR	rs11571836 19.75	239/2700	0.089
P	7-4C	<i>CDHI</i> ^{**}	c.-71C>G	5' UTR	rs34033771 0.56	69/600	0.115
F	4-2E 5-4A	<i>TP53</i> [§]	c.*485C>T	3' UTR	rs4968187 5.11	169/4500	0.038
F	2-1A, 7-1B, 5-2A, 7-1D, 7-2B, 7-2F	<i>TP53</i> [§]	c.*826G>A	3' UTR	rs17884306 5.71	371/4500	0.082

7-4C						
------	--	--	--	--	--	--

[‡]F:Flagged; P:Prioritized; [§]Long Range UTR SNPfold Analysis; [¶]Local Range SNPfold Analysis; [†]If available; [§]Rank of the SNP, in terms of how much it changes the mRNA structure compared to all other possible mutations.

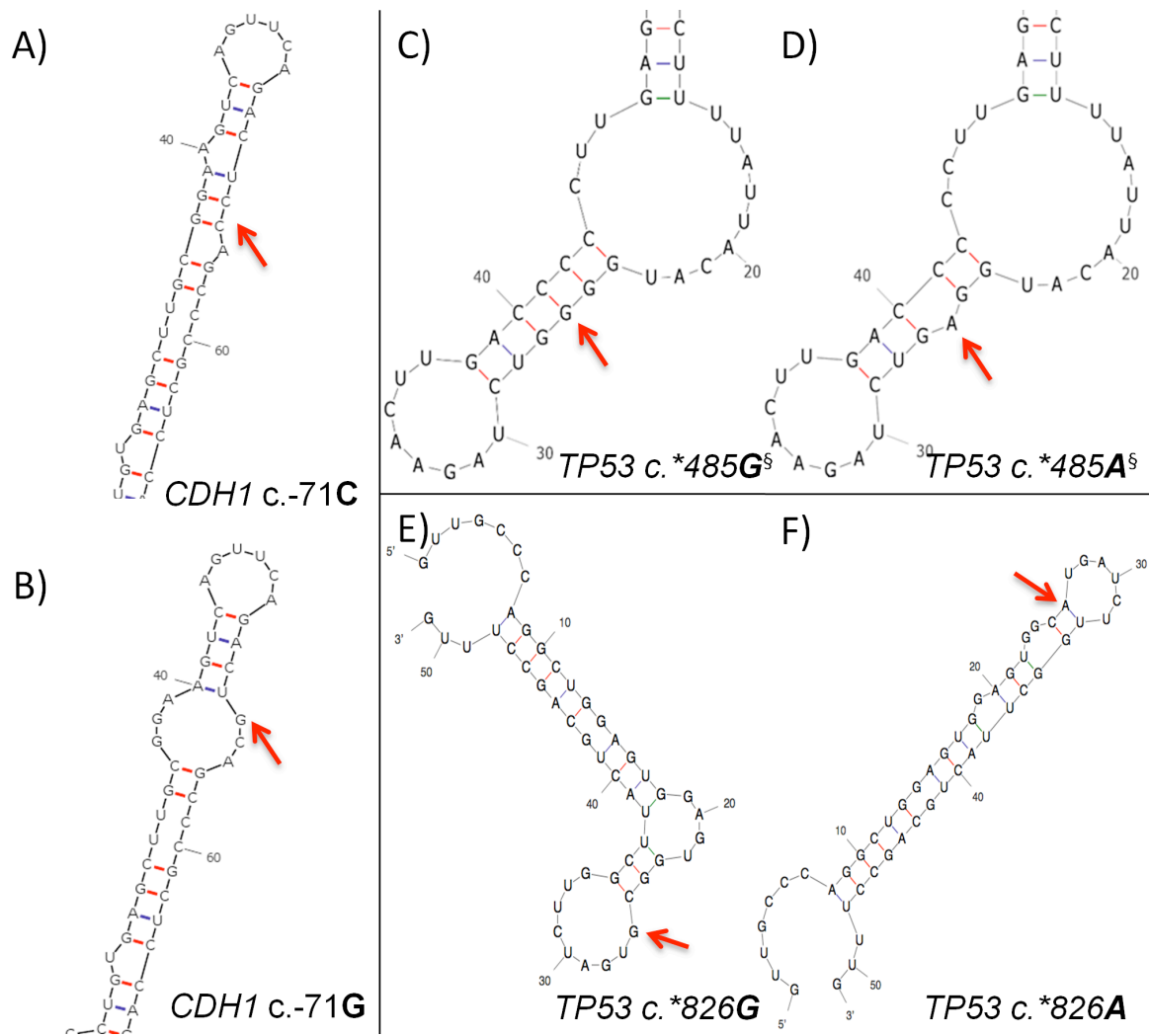


Figure 3.6. Predicted Alteration in UTR Structure Using mfold for Variants Flagged by SNPfold

Wild-type and variant structures are displayed, with the variant indicated by a red arrow. **A)** Predicted wild-type structure of *CDH1* 5'UTR surrounding c.-71. **B)** Predicted *CDH1* 5'UTR structure due to c.-71C>G variant. **C)** Predicted wild-type *TP53* 3'UTR structure surrounding c.*485. **D)** Predicted *TP53* 5'UTR structure due to c.*485G>A variant. **E)** Predicted wild-type *TP53* 3'UTR structure surrounding c.*826. **F)** Predicted *TP53* 5'UTR structure due to c.*826G>A variant. [§]SHAPE analysis revealed differences in reactivity between mutant and variant mRNAs, confirming alterations to 2° structure.

hairpin region to create a larger loop structure, thus increasing binding accessibility (**Figure 3.6A and B**). Analysis using RBPDB and CISBP-RNA-derived IT models suggests this variant affects binding by NCL by decreasing binding affinity 14-fold ($R_{i,initial} = 6.6$ bits, $\Delta R_i = -3.8$ bits) (**Supplementary Table 18**). This RBP has been shown to bind to the 5' and 3' UTR of p53 mRNA and plays a role in repressing its translation¹³⁹.

In addition, the *TP53* variant c.*485G>A (NM_000546.5: chr17:7572442C>T; rs4968187) is found at the 3'UTR and was identified in two patients (4-2E and 5-4A). *In silico* mRNA folding analysis demonstrates this variant disrupts a G/C bond of a loop in the highest ranked potential mRNA structure (**Figure 3.6C and D**). Also, SHAPE analysis shows a difference in 2° structure between the wild-type and mutant (data not shown). IT analysis with RBBS models indicated that this variant significantly increases the binding affinity of SF3B4 > 48-fold ($R_{i,final} = 11.0$ bits, $\Delta R_i = 5.6$ bits) (**Supplementary Table 18**). This RBP is one of four subunits comprising the splice factor 3B and is known to bind upstream of the branch-point sequence in pre-mRNA¹⁴⁰.

The third flagged variant also occurs in the 3'UTR of *TP53* (c.*826G>A; chr17:7572101C>T; rs17884306), and was identified in 6 patients (2-1A, 7-1B, 5-2A, 7-1D, 7-2B, 7-2F, and 7-4C). It disrupts a potential loop structure, stabilizing a double-stranded hairpin, and possibly making it less accessible (**Figure 3.6E and F**). Analysis using RBPDB-derived models suggests this variant could affect the binding of both RBFOX2 and SF3B4 (**Supplementary Table 18**). A binding site for RBFOX2, which acts as a promoter of alternative splicing by favoring the inclusion of alternative exons¹⁴¹, is created ($R_{i,final} = 9.8$ bits; $\Delta R_i = -6.5$ bits). This variant is also expected to simultaneously abolish a SF3B4 binding site ($R_{i,final} = -20.3$ bits; $\Delta R_i = -29.9$ bits).

RBPDB and CISBP-RNA-derived information model analysis of all UTR variants resulted in the prioritization of 1 novel and 5 previously-reported variants (**Table 3.4**). No patient within the cohort exhibits more than one prioritized RBBS variant.

3.3.3 Exonic Variants altering protein sequence

Exonic variants called by GATK (N=245) included insertions, deletions, nonsense, missense, and synonymous changes.

3.3.3.1 Protein-Truncating Variants

We identified 3 patients with different indels (**Table 3.5**). One was a *PALB2* insertion c.1617_1618insTT (chr16:23646249_23646250insAA; 5-3A) in exon 4, previously reported in ClinVar as pathogenic. This mutation results in a frameshift and premature translation termination by 626 residues, abolishing domain interactions with RAD51, BRCA2, and POLH¹²⁷. We also identified two known frameshift mutations in *BRCA1*: c.4964_4982del19 in exon 15 (chr17:41222949_41222967del19; rs80359876; 5-1B) and c.5266_5267insC in exon 19 (chr17:41209079_41209080insG; rs397507247; 5-3C)^{136,142}. Both are indicated as pathogenic and common in the BIC Database due to the loss of one or both C-terminal BRCT repeat domains¹²⁷. Truncation of these domains produces instability and impairs nuclear transcript localization¹⁴³, and this bipartite domain is responsible for binding phosphoproteins that are phosphorylated in response to DNA damage^{144,145}.

We also identified 4 nonsense mutations, one of which was novel in exon 4 of *PALB2* (c.1042C>T; chr16:23646825G>A; 4-4D). Another in *PALB2* has been previously reported (c.1240C>T; chr16:23646627G>A; rs180177100; 7-3A)⁴⁵. As a consequence, functional domains of PALB2 that interact with BRCA1, RAD51, BRCA2, and POLH are lost¹²⁷. Two known nonsense mutations were found in *BRCA2*, c.7558C>T in exon 15¹⁴⁶ and c.9294C>G in exon 25¹⁴⁷. The first (chr13:32930687C>T; rs80358981; 7-1G) causes the loss of the BRCA2 region that binds FANCD2, which loads BRCA2 onto damaged chromatin¹⁴⁸. The second (chr13:32968863C>G, rs80359200; 4-4A) does not occur within a known functional domain, however the transcript is likely to be degraded by nonsense mediated decay¹⁴⁹.

Table 3.5. Variants Resulting in Premature Protein Truncation

UW O ID	Gene	Exon	mRNA Protein	rsID (dbSNP142) Allele Frequency (%) [†]	ClinVar ^{abc}	Details	Ref
Insertions/Deletions							
5-1B	<i>BRCA1</i>	15 of 23	c.4964_4982del19 * p.Ser1655Tyrfs	rs80359876	6 ^a ; Pathogenic/likely pathogenic ^b ; Familial breast and breast-ovarian cancer, Hereditary cancer- predisposing syndrome ^c .	STOP at p.1670 193 AA short	-
5-3C	<i>BRCA1</i>	19 of 23	c.5266_5267insC* p.Gln1756Profs	rs397507247	13 ^a ; Pathogenic, risk factor ^b ; Familial breast, breast-ovarian, and pancreatic cancer, Hereditary cancer-predisposing syndrome ^c .	STOP at p.1788 75 AA short	136, 142
5-3A	<i>PALB2</i>	4 of 13	c.1617_1618insTT * p.Asn540Leufs	-	1 ^a ; Pathogenic ^b ; Hereditary cancer-predisposing syndrome ^c .	STOP at p.561 626 AA short	-

Stop Codons

7-1G	<i>BRCA2</i>	15 of 27	c.7558C>T** p.Arg2520Ter	rs80358981	5 ^a ; Pathogenic ^b ; Familial breast, and breast-ovarian cancer, Hereditary cancer-predisposing syndrome ^c .	899 AA short	146
4-4A	<i>BRCA2</i>	25 of 27	c.9294C>G* p.Tyr3098Ter	rs80359200	3 ^a ; Pathogenic ^b ; Familial breast and breast-ovarian cancer ^c .	321 AA short	147
7-3A	<i>PALB2</i>	4 of 13	c.1240C>T* p.Arg414Ter	rs180177100	3 ^a ; Pathogenic ^b ; Familial breast cancer, Hereditary cancer- predisposing syndrome ^c .	773 AA short	45
4-4D	<i>PALB2</i>	4 of 13	c.1042C>T* p.Gln348Ter	Novel	-	839 AA short	-

*Confirmed by Sanger sequencing; **Not confirmed by Sanger sequencing; †If available; ^aNumber of submissions; ^bClinical significance; ^cCondition(s)

3.3.3.2 Missense

GATK called 61 missense variants, of which 18 were identified in 6 patients or more and 19 had allele frequencies > 1.0% (**Supplementary Table 19**). The 40 remaining variants (15 *ATM*, 8 *BRCA1*, 9 *BRCA2*, 2 *CDH1*, 2 *CHEK2*, 3 *PALB2*, and 1 *TP53*) were assessed using a combination of gene specific databases, published classifications, and 4 *in silico* tools (**Supplementary Table 20**). We prioritized 27 variants, 2 of which were novel. None of the non-prioritized variants were predicted to be damaging by more than 2 of 4 conservation-based software programs.

3.3.4 Variant Classification

Initially, 15,311 unique variants were identified by complete gene sequencing of 7 HBOC genes. Of these, 132 were flagged after filtering, and further reduced by IT-based variant analysis and consultation of the published literature to 87 prioritized variants. **Figure 3.7** illustrates the decrease in the number of unique variants per patient at each step of our identification and prioritization process. The distribution of prioritized variants by gene is 34 in *ATM*, 13 in *BRCA1*, 11 in *BRCA2*, 8 in *CDH1*, 6 in *CHEK2*, 10 in *PALB2*, and 5 in *TP53* (**Supplementary Table 21**), which are categorized by type in **Table 3.6**.

Three prioritized variants have multiple predicted roles: *ATM* c.1538A>G in missense and SRFBS, *CHEK2* c.190G>A in missense and UTR binding, and *CHEK2* c.433C>T in missense and UTR binding. Of the 102 patients that we sequenced, 72 (70.6%) exhibited at least one prioritized variant, and some patients harbored more than one prioritized variant (N=33; 32%). **Supplementary Table 22** presents a summary of all flagged and prioritized variants for patients with at least one prioritized variant.

3.3.5 Variant Verification

We verified prioritized protein-truncating (N=7) and splicing (N=4) variants by Sanger sequencing (**Table 3.5** and **Table 3.3**, respectively). In addition, two missense variants (*BRCA2* c.7958T>C and *CHEK2* c.433C>T) were re-sequenced, since they are indicated as likely pathogenic/pathogenic in ClinVar (**Supplementary Table 20**). All protein-

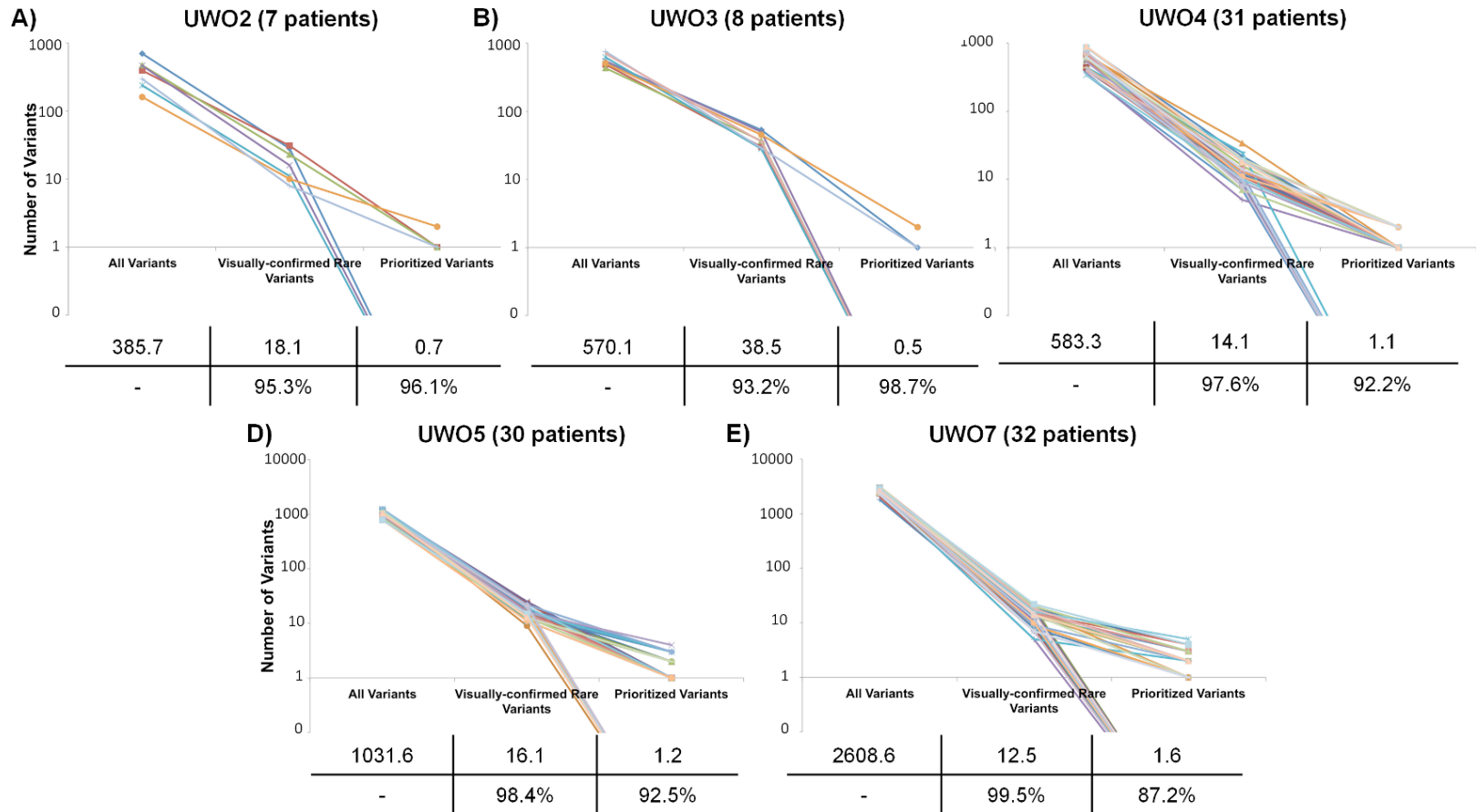


Figure 3.7. Ladder Plot Representing Variant Identification and Prioritization

Each line is representative of a different sample in each sequencing run (A-E), illustrating the number of unique variants at important steps throughout the variant prioritization process. The left-most point indicates the total number of unique variants. The second point represents the number of unique variants remaining after common (> 5 patients within cohort and/or $\geq 1.0\%$

allele frequency) and false-positive variants were removed. The right-most point represents the final number of unique. No variants were prioritized in the following patients: 2-1A, 2-5A, 2-6A, 3-2A, 3-3A, 3-4A, 3-5A, 3-8A, 4-1B, 4-2C, 4-2F, 4-3B, 4-3D, 4-4B, 4-4E, 5-1G, 5-1H, 5-3D, 5-4C, 5-4D, 5-4F, 5-4G, 5-4H, 7-1B, 7-1C, 7-1D, 7-1H, 7-2B, 7-2C, 7-2H, 7-3H, 7-4A, 7-4D, 7-4H. The average number of variants per patient at each step is indicated in a table below each plot, along with the percent reduction in variants from one step to another.

Table 3.6. Summary of Prioritized Variants by Gene

	Indel	Nonsense	Missense	Natural Splicing	Cryptic Splicing	Pseudoexon	SR Factor	TF	UTR Structure	UTR Binding	Total
<i>ATM</i>	0	0	14	2	0	0	18	0	0	1	34 [‡]
<i>BRCA1</i>	2	0	2	0	0	1	7	1	0	0	13
<i>BRCA2</i>	0	2	3	0	0	2	4	0	0	0	11
<i>CDH1</i>	0	0	2	0	0	2	1	1	1	1	8
<i>CHEK2</i>	0	0	2	1	0	0	3	0	0	2	6 [‡]
<i>PALB2</i>	1	2	3	0	0	0	3	1	0	0	10
<i>TP53</i>	0	0	1	0	0	0	0	2	0	2	5

[‡]Counts represent the number of unique variants identified (i.e. a variant is not counted twice if it appeared in multiple individuals).

Three variants were prioritized under multiple categories: *ATM* chr11:108121730A>G (missense and SRFBS), *CHEK2* chr22:29121242G>A (missense, UTR binding), and *CHEK2* chr22:29130520C>T (missense, UTR binding).

truncating variants were confirmed, with one exception (*BRCA2* c.7558C>T, no evidence for the variant was present for either strand). Two of the mRNA splicing mutations were confirmed on both strands, while the other two were confirmed on a single strand (*ATM* c.6347+1G>T and *ATM* c.1066-6T>G). Both documented pathogenic missense variants were also confirmed.

3.4 Discussion

NGS technology offers advantages in throughput and variant detection¹¹⁶, but the task of interpreting the sheer volume of variants in complete gene or genome data can be daunting. The whole genome of a Yoruban male contained approximately 4.2 million SNVs and 0.4 million structural variants¹⁵⁰. The variant density in the present study (average 948 variants per patient) was 5.3-fold lower than the same regions in HapMap sample NA12878 in Illumina Platinum Genomes Project (5,029 variants)¹⁵¹. The difference can be attributed primarily to the exclusion of polymorphisms in highly repetitive regions in our study.

Conventional coding sequence analysis, combined with an IT-based approach for regulatory and splicing-related variants, reduced the set to a manageable number of prioritized variants. Unification of non-coding analysis of diverse protein-nucleic acid interactions using the IT framework accomplishes this by applying thermodynamic-based thresholds to binding affinity changes and by selecting the most significant binding site information changes, regardless of whether the motifs of different factors overlap.

Previously, rule-based systems have been proposed for variant severity classification^{152,153}. Functional validation and risk analyses of these variants are a prerequisite to classification, but this would not be practical to accomplish without first limiting the subset of variants analyzed. With the exception of some (but not all⁸³) protein truncating variants, classification is generally not achievable by sequence analysis alone. Only a minority of variants with extreme likelihoods of pathogenic or benign phenotypes are clearly delineated because only these types of variants are considered actionable^{152,153}. The proposed classification systems preferably require functional, co-segregation, and

risk analyses to stratify patients. Nevertheless, the majority of variants are VUS, especially in the case of variants occurring beyond exon boundaries. Of the 5,713 variants listed in the BIC database, the clinical significance of 4,102 *BRCA1* and *BRCA2* variants are either unknown (1,904) or pending (2,198), while 1,535 are classified as pathogenic (Class 5)¹⁵⁴. Our results cannot be considered equivalent to validation, which might include expression assays³¹ or the use of RNA-seq data¹⁵⁵ (splicing), qRT-PCR¹⁵⁶ (transcription), SHAPE analysis (mRNA 2° structure)³⁵, and binding assays to determine functional effects of variants. Other post-transcriptional processes (eg. miRNA regulation) affected by variants have not been addressed in this study, but should also be amenable to IT-based modeling. With the proposed approach, functional prediction of variants could precede or at least inform the classification of VUS.

It is unrealistic to expect all variants to be functionally analyzed, just as it may not be feasible to assess family members for a suspected pathogenic variant detected in a proband. The prioritization procedure reduces the chance that significant variants have been overlooked. Capturing coding and non-coding regions of HBOC-related genes, combined with the framework for assessing variants balances the need to comprehensively detect all variation in a gene panel with the goal of identifying variants likely to be phenotypically relevant.

3.4.1 Non-coding variants

Variant density in non-coding regions significantly exceeded exonic variants by > 60-fold, which, in absolute terms, constituted 1.6% of the 15,311 variants. This is comparable to whole genome sequencing studies, which typically result in 3-4 million variants per individual, with < 2% occurring in protein coding regions¹⁵⁷. IT analysis prioritized 3 natural SS, 36 SRFBS, 5 TFBS, and 6 RBBS variants and 5 predicted to create pseudoexons. Two SS variants in *ATM* (c.3747-1G>A and c.6347+1G>T) were predicted to completely abolish the natural site and cause exon skipping. A *CHEK2* variant (c.320-5A>T) was predicted to result in leaky splicing.

The IT-based framework evaluates all variants on a common scale, based on bit values, the universal unit that predicts changes in binding affinity¹⁵⁸. A variant can alter the strength of one or a “set” of binding sites; the magnitude and direction of these changes is used to rank their significance. The models used to derive information weight matrices take into account the frequency of all observed bases at a given position of a binding motif, making them more accurate than consensus sequence and conservation-based approaches³¹.

IT has been widely used to analyze natural and cryptic SSs³¹, but its use in SRFBS analysis was only introduced recently⁸⁰. For this reason, we assigned conservative, minimum thresholds for reporting information changes. Although there are examples of disease-causing variants resulting in small changes in R_i ¹⁵⁹⁻¹⁶⁶, the majority of deleterious splicing mutations that have been verified functionally, produce large information changes. Among 698 experimentally deleterious variants in 117 studies, only 1.96% resulted in < 1.0 bit change³¹. For SRFBS variants, the absolute information changes for deleterious variants ranged from 0.2 - 17.1 bits (mean 4.7 ± 3.8). This first application of IT in TFBS and RBBS analysis, however, lacks a large reference set of validated mutations for the distribution of information changes associated with deleterious variants. The release of new ChIP-seq datasets will enable IT models to be derived for TFs currently unmodeled and to improve existing models¹⁶⁷.

Pseudoexon activation results in disease-causing mutations¹⁶⁸, however such consequences are not customarily screened for in mRNA splicing analysis. IT analysis was used to detect variants that predict pseudoexon formation and 5 variants were prioritized. Previously, we have predicted experimentally proven pseudoexons with IT (Ref 34: Table 2, No #2 and Ref 169: Table 2, No #7)^{34,169}. Although it was not possible to confirm prioritized variants in the current study predicted to activate pseudoexons because of their low allele frequencies, common intronic variants that were predicted to form pseudoexons were analyzed. We then searched for evidence of pseudoexon activation in mapped human EST and mRNA tracks¹⁷⁰ and RNA-seq data of breast normal and tumour tissue from the Cancer Genome Atlas project¹⁷¹. One of these variants

(rs6005843) appeared to splice the human EST HY160109¹⁷² at the predicted cryptic splice site and is expressed within the pseudoexon boundaries.

Variants that were common within our population sample (i.e. occurring in > 5 individuals) and/or common in the general population (> 1.0% allele frequency) reduced the list of flagged variants substantially. This is now a commonly accepted approach for reducing candidate disease variants¹⁵², based on the principle that the disease-causing variants occur at lower population frequencies. Variants occurring in > 5 patients all either had allele frequencies above 1.0% or, as shown previously, resulted in very small ΔR_i values¹⁷³.

The genomic context of sequence changes can influence the interpretation of a particular variant³¹. For example, variants causing significant information changes may be interpreted as inconsequential if they are functionally redundant or enhancing existing binding site function (see *IT-Based Variant Analysis* for details). Our understanding of the roles and context of these cognate protein factors is incomplete, which affects confidence in interpretation of variants that alter binding. Also, certain factors with important roles in the regulation of these genes, but that do not bind DNA directly or in a sequence-specific manner, (eg. CtBP2¹⁷⁴), could not be included. Therefore, some variants may have been incorrectly excluded.

3.4.2 Coding sequence changes

We also identified 4 nonsense and 3 indels in this cohort. In one individual, a 19 nt *BRCA1* deletion in exon 15 causes a frameshift leading to a stop codon within 14 codons downstream. This variant, rs80359876, is considered clinically relevant. Interestingly, this deletion overlaps two other published deletions in this exon (rs397509209 and rs80359884). This raises the question as to whether this region of the *BRCA1* gene is a hotspot for replication errors. DNA folding analysis indicates a possible 15 nt long stem-loop spanning this interval as the most stable predicted structure (data not shown). This 15 nt structure occurs entirely within the rs80359876 and rs397509209 deletions and partially overlaps rs80359884 (13 of 15 nt of the stem loop). It is plausible that the 2°

structure of this sequence predisposes to a replication error that leads to the observed deletion.

Missense coding variants were also assessed using multiple *in silico* tools and evaluated based on allele frequency, literature references, and gene-specific databases. Of the 27 prioritized missense variants, the previously reported *CHEK2* variant c.433G>A (chr22:29121242G>A; rs137853007) stood out, as it was identified in one patient (4-3C.5-4G) and is predicted by all 4 *in silico* tools to have a damaging effect on protein function. Accordingly, Wu *et al.* (2001) demonstrated reduced *in vitro* kinase activity and phosphorylation by ATM kinase compared to the wild-type protein¹⁷⁵, presumably due to the variant's occurrence within the forkhead homology-associated domain, involved in protein-phosphoprotein interactions¹⁷⁶. Implicated in Li-Fraumeni syndrome, known to increase the risk of developing several types of cancer including breast^{177,178}, this variant is expected to result in a misfolded protein that would be targeted for degradation via the ubiquitin-proteasome pathway¹⁷⁹. Another important missense variant is c.7958T>C (chr13:32936812T>C; rs80359022; 4-4C) in exon 17 of *BRCA2*. Although classified as being of unknown clinical importance in both BIC and ClinVar, it has been classified as pathogenic based on posterior probability calculations¹⁸⁰.

It is unlikely that all prioritized variants are pathogenic in patients carrying more than one prioritized variant. Nevertheless, a polygenic model for breast cancer susceptibility, whereby multiple moderate and low-risk alleles contribute to increased risk of HBOC may also account for multiple prioritized variants^{181,182}. There was a significant fraction of patients (29.4%) in whom no variants were prioritized. This could be due to: a) the inability of the analysis to predict a variant affecting the binding sites analyzed, b) a pathogenic variant affects a function that was not analyzed or in a gene that was not sequenced, or c) the significant family history was not due to heritable, but instead to shared environmental influences.

BRCA coding variants were found in individuals who were previously screened for lesions in these genes, suggesting this NGS protocol is a more sensitive approach for detecting coding changes. However, the previous testing was predominantly based on

PTT and MLPA methods, which have lower sensitivity than sequence analysis. Nevertheless, we identified 2 *BRCA1* and 2 *BRCA2* variants predicted to encode prematurely truncated proteins. Fewer non-coding *BRCA* variants were prioritized (15.7%) than expected by linkage analysis³⁷, however this presumes at least 4 affected breast cancer diagnoses per pedigree, and, in the present study, the number of affected individuals per family was not known.

Prioritization of a variant does not equate with pathogenicity. Some prioritized variants may not increase risk, but may simply modify a primary unrecognized pathogenic mutation. A patient with a known *BRCA1* nonsense variant, used as a positive control, was also found to possess an additional prioritized variant in *BRCA2* (missense variant chr13:32911710A>G), which was flagged by PROVEAN and SIFT as damaging, as well as flagged for changing an SRFBS for abolishing a PTB site (while simultaneously abolishing an exonic hnRNPA1 site). This variant has been identified in cases of early onset prostate cancer and is considered a VUS in ClinVar¹³³. A larger cohort of patients with known pathogenic mutations would be necessary to calculate a background/basal rate of falsely flagged variants.

Other groups have attempted to develop comprehensive approaches for variant analysis, analogous to the one proposed here^{183–185}. While most employ high-throughput sequencing and classify variants, either the sequences analyzed or the types of variants assessed tend to be limited. In particular, non-coding sequences have not been sequenced or studied to the same extent, and none of these analytical approaches have adopted a common framework for mutation analysis.

Our published oligonucleotide design method⁶⁴ produced an average sequence coverage of 98.8%. The capture reagent did not overlap conserved highly repetitive regions, but included divergent repetitive sequences. Nevertheless, neighboring probes generated reads with partial overlap of repetitive intervals. As previously reported¹³⁵, we noted that false positive variant calls within intronic and intergenic regions were the most common consequence of dephasing in low complexity, pyrimidine-enriched intervals. This was not alleviated by processing data with software programs based on different alignment or

calling algorithms. Manual review of all intronic or intergenic variants became imperative. As these sequences can still affect functional binding elements detectable by IT analysis (i.e. 3' SSs and SRFBSs), it may prove essential to adopt or develop alignment software that explicitly and correctly identifies variants in these regions¹³⁵. Most variants were confirmed with Sanger sequencing (10/13), and those that could not be confirmed are not necessarily false positives. A recent study demonstrated that NGS can identify variants that Sanger sequencing cannot, and reproducing sequencing results by NGS may be worthwhile before eliminating such variants¹⁸⁶.

3.5 Conclusions

Through a comprehensive protocol based on high-throughput, IT-based and complementary coding sequence analyses, the numbers of VUS can be reduced to a manageable quantity of variants, prioritized by predicted function. Exonic variants corresponded to a small fraction of prioritized variants, illustrating the importance of sequencing non-coding regions of genes. We propose that our approach for variant flagging and prioritization is an intermediate bridge between high-throughput sequencing, variant detection, and the time-consuming process of variant classification, including pedigree analysis and functional validation.

3.6 References

1. Collins, F. S. & Hamburg, M. A. First FDA Authorization for Next-Generation Sequencer. *N. Engl. J. Med.* **369**, 2369–2371 (2013).
2. Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
3. Cassa, C. A. *et al.* Disclosing pathogenic genetic variants to research participants: Quantifying an emerging ethical responsibility. *Genome Res.* **22**, 421–428 (2012).
4. Domchek, S. M., Bradbury, A., Garber, J. E., Offit, K. & Robson, M. E. Multiplex Genetic Testing for Cancer Susceptibility: Out on the High Wire Without a Net? *J. Clin. Oncol.* **31**, 1267–1270 (2013).
5. Yorczyk, A., Robinson, L. S. & Ross, T. S. Use of panel tests in place of single gene tests in the cancer genetics clinic. *Clin. Genet.* **88**, 278–282 (2015).
6. Foley, S. B. *et al.* Use of Whole Genome Sequencing for Diagnosis and Discovery in the Cancer Genetics Clinic. *EBioMedicine* **2**, 74–81 (2015).
7. Schwartz, G. F. *et al.* Proceedings of the international consensus conference on breast cancer risk, genetics, & risk management, April, 2007. *Cancer* **113**, 2627–2637 (2008).
8. Kavanagh, D. & Anderson, H. E. Interpretation of genetic variants of uncertain significance in atypical hemolytic uremic syndrome. *Kidney Int.* **81**, 11–13 (2012).
9. Tavtigian, S. V., Greenblatt, M. S., Lesueur, F., Byrnes, G. B. & Group, I. U. G. V. W. In silico analysis of missense substitutions using sequence-alignment based methods. *Hum. Mutat.* **29**, 1327–1336 (2008).
10. Vos, J. *et al.* The counselees' view of an unclassified variant in BRCA1/2: recall, interpretation, and impact on life. *Psychooncology.* **17**, 822–830 (2008).
11. Domchek, S. & Weber, B. L. Genetic variants of uncertain significance: flies in the ointment. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **26**, 16–17 (2008).
12. Braun, T. A. *et al.* Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. *Hum. Mol. Genet.* **22**, 5136–5145 (2013).

13. Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends Genet. TIG* **29**, 318–327 (2013).
14. Chatterjee, S., Berwal, S. K. & Pal, J. K. in *eLS* (John Wiley & Sons, Ltd, 2001). at <<http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0022408/abstract>>
15. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* **6**, e1001074 (2010).
16. Misquitta, C. M., Iyer, V. R., Werstiuk, E. S. & Grover, A. K. The role of 3'-untranslated region (3'-UTR) mediated mRNA stability in cardiovascular pathophysiology. *Mol. Cell. Biochem.* **224**, 53–67 (2001).
17. Latchman, D. S. Transcription-Factor Mutations and Disease. *N. Engl. J. Med.* **334**, 28–33 (1996).
18. Ward, A. J. & Cooper, T. A. The Pathobiology of Splicing. *J. Pathol.* **220**, 152–163 (2010).
19. Araujo, P. R. *et al.* Before It Gets Started: Regulating Translation at the 5' UTR. *Comp. Funct. Genomics* **2012**, 475731 (2012).
20. Cáceres, J. F. & Kornblihtt, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet. TIG* **18**, 186–193 (2002).
21. Teraoka, S. N. *et al.* Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. *Am. J. Hum. Genet.* **64**, 1617–1631 (1999).
22. Ars, E. *et al.* Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.* **9**, 237–247 (2000).
23. Paul, D. S., Soranzo, N. & Beck, S. Functional interpretation of non-coding sequence variation: Concepts and challenges. *BioEssays* **36**, 191–199 (2014).
24. Guo, Y. & Jamison, D. C. The distribution of SNPs in human gene regulatory regions. *BMC Genomics* **6**, 140 (2005).
25. Horvath, A. *et al.* Novel insights into breast cancer genetic variance through RNA sequencing. *Sci. Rep.* **3**, 2256 (2013).
26. Pavithra, L. *et al.* Stabilization of SMAR1 mRNA by PGA2 involves a stem loop structure in the 5' UTR. *Nucleic Acids Res.* **35**, 6004–6016 (2007).

27. Pérez-Cabornero, L. *et al.* Evaluating the effect of unclassified variants identified in MMR genes using phenotypic features, bioinformatics prediction, and RNA assays. *J. Mol. Diagn. JMD* **15**, 380–390 (2013).
28. Zeng, T. *et al.* A novel variant in the 3' UTR of human SCN1A gene from a patient with Dravet syndrome decreases mRNA stability mediated by GAPDH's binding. *Hum. Genet.* **133**, 801–811 (2014).
29. Gaildrat, P. *et al.* The BRCA1 c.5434C->G (p.Pro1812Ala) variant induces a deleterious exon 23 skipping by affecting exonic splicing regulatory elements. *J. Med. Genet.* **47**, 398–403 (2010).
30. Tournier, I. *et al.* A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum. Mutat.* **29**, 1412–1424 (2008).
31. Caminsky, N. G., Mucaki, E. J. & Rogan, P. K. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Research* **3**, 282 (2015).
32. Schneider, T. D., Stormo, G. D., Yarus, M. A. & Gold, L. Delila system tools. *Nucleic Acids Res.* **12**, 129–140 (1984).
33. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
34. Rogan, P. K., Faux, B. M. & Schneider, T. D. Information analysis of human splice site mutations. *Hum. Mutat.* **12**, 153–171 (1998).
35. Steen, K.-A., Siegfried, N. A. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by protection from exoribonuclease (RNase-detected SHAPE) for direct analysis of covalent adducts and of nucleotide flexibility in RNA. *Nat. Protoc.* **6**, 1683–1694 (2011).
36. Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* **127**, 2893–2917 (2010).
37. Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **62**, 676–689 (1998).

38. Levy-Lahad, E. & Plon, S. E. Cancer. A risky business--assessing breast cancer risk. *Science* **302**, 574–575 (2003).
39. Plon, S. E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**, 1282–1291 (2008).
40. Borg, A. *et al.* Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum. Mutat.* **31**, E1200–40 (2010).
41. Adank, M. A. *et al.* CHEK2*1100delC homozygosity is associated with a high breast cancer risk in women. *J. Med. Genet.* **48**, 860–863 (2011).
42. Baloch, A. H. *et al.* Missense mutations (p.H371Y, p.D438Y) in gene CHEK2 are associated with breast cancer risk in women of Balochistan origin. *Mol. Biol. Rep.* **41**, 1103–1107 (2014).
43. Benusiglio, P. R. *et al.* CDH1 germline mutations and the hereditary diffuse gastric and lobular breast cancer syndrome: a multicentre study. *J. Med. Genet.* **50**, 486–489 (2013).
44. Brooks-Wilson, A. R. *et al.* Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria. *J. Med. Genet.* **41**, 508–517 (2004).
45. Casadei, S. *et al.* Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. *Cancer Res.* **71**, 2222–2229 (2011).
46. CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am. J. Hum. Genet.* **74**, 1175–1182 (2004).
47. Garber, J. E. & Offit, K. Hereditary cancer predisposition syndromes. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23**, 276–292 (2005).
48. Kangelaris, K. N. & Gruber, S. B. Clinical implications of founder and recurrent CDH1 mutations in hereditary diffuse gastric cancer. *JAMA* **297**, 2410–2411 (2007).

49. Kaurah, P. *et al.* Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. *JAMA* **297**, 2360–2372 (2007).
50. Kluijdt, I. *et al.* Familial gastric cancer: guidelines for diagnosis, treatment and periodic surveillance. *Fam. Cancer* **11**, 363–369 (2012).
51. Martin, A.-M. *et al.* Germline TP53 mutations in breast cancer families with multiple primary cancers: is TP53 a modifier of BRCA1? *J. Med. Genet.* **40**, e34–e34 (2003).
52. Masciari, S. *et al.* Germline E-cadherin mutations in familial lobular breast cancer. *J. Med. Genet.* **44**, 726–731 (2007).
53. Maxwell, K. N. *et al.* Prevalence of mutations in a panel of breast cancer susceptibility genes in BRCA1/2-negative patients with early-onset breast cancer. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 630–638 (2015).
54. Minion, L. E. *et al.* Hereditary predisposition to ovarian cancer, looking beyond BRCA1/BRCA2. *Gynecol. Oncol.* **137**, 86–92 (2015).
55. Olivier, M. *et al.* Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. *Cancer Res.* **63**, 6643–6650 (2003).
56. Pharoah, P. D., Guilford, P., Caldas, C. & International Gastric Cancer Linkage Consortium. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* **121**, 1348–1353 (2001).
57. Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).
58. Renwick, A. *et al.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* **38**, 873–875 (2006).
59. Sidransky, D. *et al.* Inherited p53 gene mutations in breast cancer. *Cancer Res.* **52**, 2984–2986 (1992).
60. Slater, E. P. *et al.* PALB2 mutations in European familial pancreatic cancer families. *Clin. Genet.* **78**, 490–494 (2010).
61. Thompson, D. *et al.* Cancer risks and mortality in heterozygous ATM mutation carriers. *J. Natl. Cancer Inst.* **97**, 813–822 (2005).

62. Tischkowitz, M. *et al.* Rare germline mutations in PALB2 and breast cancer risk: a population-based study. *Hum. Mutat.* **33**, 674–680 (2012).
63. Walsh, T. *et al.* Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA* **295**, 1379–1388 (2006).
64. Dorman, S. N., Shirley, B. C., Knoll, J. H. M. & Rogan, P. K. Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Res.* **41**, e81 (2013).
65. Pinkel, D. *et al.* Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 9138–9142 (1988).
66. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. (2013). at <<<http://www.repeatmasker.org>>>
67. Gnirke, A. *et al.* Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
68. Chou, H.-H., Hsia, A.-P., Mooney, D. L. & Schnable, P. S. Picky: oligo microarray design for large genomes. *Bioinforma. Oxf. Engl.* **20**, 2893–2902 (2004).
69. Markham, N. R. & Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol. Clifton NJ* **453**, 3–31 (2008).
70. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
71. Predictive Cancer Genetics Steering Committee Ontario physicians' guide to referral of patients with family history of cancer to a familial cancer genetics clinic or genetics clinic. *Ont Med Rev.* **68**, 24–30 (2001).
72. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
73. Philippe, N., Salson, M., Combes, T. & Rivals, E. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.* **14**, R30 (2013).
74. Picard. at <<http://picard.sourceforge.net/>>. Accessed June 1, 2015.

75. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
76. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
77. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
78. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* **11**, 77–85 (2013).
79. Mutation Forecaster. at <<https://www.mutationforecaster.com/index.php>>. Accessed June 1, 2015.
80. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Hum. Mutat.* **34**, 557–565 (2013).
81. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431 (1986).
82. Dhir, A. & Buratti, E. Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *FEBS J.* **277**, 841–855 (2010).
83. Peterlongo, P. *et al.* FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. *Hum. Mol. Genet.* (2015). doi:10.1093/hmg/ddv251
84. Tavanez, J. P., Madl, T., Kooshapur, H., Sattler, M. & Valcárcel, J. hnRNP A1 proofreads 3' splice site recognition by U2AF. *Mol. Cell* **45**, 314–329 (2012).
85. Paradis, C. *et al.* hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. *RNA N. Y. N* **13**, 1287–1300 (2007).
86. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

87. Boggs, K. & Reisman, D. Increased p53 transcription prior to DNA synthesis is regulated through a novel regulatory element within the p53 promoter. *Oncogene* **25**, 555–565 (2005).
88. Chen, Y. *et al.* c-Myc activates BRCA1 gene expression through distal promoter elements in breast cancer cells. *BMC Cancer* **11**, 246 (2011).
89. Gueven, N. *et al.* Site-directed mutagenesis of the ATM promoter: Consequences for response to proliferation and ionizing radiation. *Genes. Chromosomes Cancer* **38**, 157–167 (2003).
90. Frieze, S. *et al.* Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol.* **13**, R52 (2012).
91. Connor, A. E. *et al.* Associations between TCF7L2 polymorphisms and risk of breast cancer among Hispanic and non-Hispanic white women: the Breast Cancer Health Disparities Study. *Breast Cancer Res. Treat.* **136**, 593–602 (2012).
92. Burwinkel, B. *et al.* Transcription factor 7-like 2 (TCF7L2) variant is associated with familial breast cancer risk: a case-control study. *BMC Cancer* **6**, 268 (2006).
93. Chen, J., Yuan, T., Liu, M. & Chen, P. Association between TCF7L2 Gene Polymorphism and Cancer Risk: A Meta-Analysis. *PLoS ONE* **8**, e71730 (2013).
94. Purrington, K. S. *et al.* Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis* **35**, 1012–1019 (2014).
95. Bi, C. & Rogan, P. K. Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.* **32**, 4979–4991 (2004).
96. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
97. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet. TIG* **13**, 163 (1997).
98. Gadiraju, S., Vyhldal, C. A., Leeder, J. S. & Rogan, P. K. Genome-wide prediction, display and refinement of binding sites with information theory-based models. *BMC Bioinformatics* **4**, 38 (2003).

99. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* **39**, D301–8 (2011).
100. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
101. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
102. Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res.* **40**, W65–W70 (2012).
103. Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. A practical guide for the functional annotation of genetic variations using SNPnexus. *Brief. Bioinform.* **14**, 437–447 (2013).
104. Chelala, C., Khan, A. & Lemoine, N. R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655–661 (2009).
105. dbSNP. at <<http://www.ncbi.nlm.nih.gov/SNP/>>. Accessed June 1, 2015.
106. Exome Variant Server. at <<http://evs.gs.washington.edu/EVS/>>. Accessed June 1, 2015.
107. 1000Genomes. at <<http://www.1000genomes.org/>>. Accessed June 1, 2015.
108. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
109. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* **8**, R232 (2007).
110. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
111. Choi, Y. A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-locus Variants of Another Protein. in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine* 414–417 (ACM, 2012). doi:10.1145/2382936.2382989

112. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **7**, e46688 (2012).
113. ClinVar. at <<http://www.ncbi.nlm.nih.gov/clinvar/>>. Accessed June 1, 2015.
114. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* gkt1113 (2013). doi:10.1093/nar/gkt1113
115. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
116. Human Gene Mutation Database (HGMD). at <<http://hgmd.cf.ac.uk/ac/index.php>>. Accessed June 1, 2015.
117. Fokkema, I. F. A. C. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* **32**, 557–563 (2011).
118. Leiden Open Variation Database (LOVD) - Ataxia Telangiectasia Mutated (ATM). at <http://chromium.lovd.nl/LOVD2/variants.php?action=search_unique&select_db=ATM>. Accessed June 1, 2015.
119. LOVD - IARC Breast Cancer Type 1 susceptibility protein (BRCA1). at <http://brca.iarc.fr/LOVD/variants.php?action=view_unique&select_db=BRCA1>. Accessed June 1, 2015.
120. LOVD - IARC Breast Cancer Type 2 susceptibility protein (BRCA2). at <http://brca.iarc.fr/LOVD/variants.php?action=view_unique&select_db=BRCA2>. Accessed June 1, 2015.
121. LOVD - Leiden Open Variation Database Partner and localizer of BRCA2 (FANCN) (PALB2). at <https://grenada.lumc.nl/LOVD2/shared1/variants.php?action=search_unique&select_db=PALB2>. Accessed June 1, 2015.
122. LOVD - Leiden Open Variation Database tumour protein p53 (TP53). at <<http://proteomics.bio21.unimelb.edu.au/lovd/variants/TP53>>. Accessed June 1, 2015.

123. Zhejiang University Center for Genetic and Genomic Medicine (ZJU-CGGM) cadherin 1, type 1, E-cadherin (epithelial) (CDH1). at http://www.genomed.org/lovd2/variants.php?action=search_unique&select_db=CDH1>. Accessed June 1, 2015.
124. Zhejiang University Center for Genetic and Genomic Medicine (ZJU-CGGM) checkpoint kinase 2 (CHEK2). at http://www.genomed.org/lovd2/variants.php?action=search_unique&select_db=HEK2>. Accessed June 1, 2015.
125. Domain Mapping of Disease Mutations (DM2). at <http://bioinf.umbc.edu/dmdm>>. Accessed June 1, 2015.
126. Expert Protein Analysis System (ExpASy). at <http://www.expasy.org>>. Accessed June 1, 2015.
127. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
128. UniProt. at <http://uniprot.org>>. Accessed June 1, 2015.
129. Breast Cancer Information Core (BIC) Database. at <https://research.nhgri.nih.gov/projects/bic/Member/index/shtml>>. Accessed June 1, 2015.
130. Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA). at <http://enigmaconsortium.org>>. Accessed June 1, 2015.
131. International Agency for Research on Cancer (IARC) TP53 Database. at <http://p53.iarc.fr/tp53genevariations.aspx>>. Accessed June 1, 2015.
132. Ozcelik, H. *et al.* Individual and family characteristics associated with protein truncating BRCA1 and BRCA2 mutations in an Ontario population based series from the Cooperative Family Registry for Breast Cancer Studies. *J. Med. Genet.* **40**, e91 (2003).
133. Maier, C. *et al.* Subgroups of familial and aggressive prostate cancer with considerable frequencies of BRCA2 mutations. *The Prostate* **74**, 1444–1451 (2014).

134. McIver, L. J., Fondon III, J. W., Skinner, M. A. & Garner, H. R. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* **97**, 193–199 (2011).
135. Tae, H., Kim, D.-Y., McCormick, J., Settlage, R. E. & Garner, H. R. Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. *Bioinformatics* **30**, 652–659 (2014).
136. Castéra, L. *et al.* Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. *Eur. J. Hum. Genet. EJHG* **22**, 1305–1313 (2014).
137. Austen, B. *et al.* Pathogenic ATM mutations occur rarely in a subset of multiple myeloma patients. *Br. J. Haematol.* **142**, 925–933 (2008).
138. Ding, H. *et al.* Lack of association between ATM C.1066-6T > G mutation and breast cancer risk: a meta-analysis of 8,831 cases and 4,957 controls. *Breast Cancer Res. Treat.* **125**, 473–477 (2011).
139. Chen, J., Guo, K. & Kastan, M. B. Interactions of nucleolin and ribosomal protein L26 (RPL26) in translational control of human p53 mRNA. *J. Biol. Chem.* **287**, 16467–16476 (2012).
140. Champion-Arnaud, P. & Reed, R. The prespliceosome components SAP 49 and SAP 145 interact in a complex implicated in tethering U2 snRNP to the branch site. *Genes Dev.* **8**, 1974–1983 (1994).
141. Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res.* **25**, 1–13 (2015).
142. Dobričić, J. *et al.* Serbian high-risk families: extensive results on BRCA mutation spectra and frequency. *J. Hum. Genet.* **58**, 501–507 (2013).
143. Nelson, A. C. & Holt, J. T. Impact of RING and BRCT domain mutations on BRCA1 protein stability, localization and recruitment to DNA damage. *Radiat. Res.* **174**, 1–13 (2010).

144. Clark, S. L., Rodriguez, A. M., Snyder, R. R., Hankins, G. D. V. & Boehning, D. Structure-Function Of The Tumor Suppressor BRCA1. *Comput. Struct. Biotechnol. J.* **1**, (2012).
145. Leung, C. C. Y. & Glover, J. N. M. BRCT domains: easy as one, two, three. *Cell Cycle Georget. Tex* **10**, 2461–2470 (2011).
146. Håkansson, S. *et al.* Moderate frequency of BRCA1 and BRCA2 germ-line mutations in Scandinavian familial breast cancer. *Am. J. Hum. Genet.* **60**, 1068–1078 (1997).
147. Scottish/Northern Irish BRCA1/BRCA2 Consortium. BRCA1 and BRCA2 mutations in Scotland and Northern Ireland. *Br. J. Cancer* **88**, 1256–1262 (2003).
148. Hussain, S. *et al.* Direct interaction of FANCD2 with BRCA2 in DNA damage response pathways. *Hum. Mol. Genet.* **13**, 1241–1248 (2004).
149. Chang, Y. F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.* **76**, 51–74 (2007).
150. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
151. Platinum Genomes. at <<http://www.illumina.com/platinumgenomes/>>. Accessed July 31, 2015.
152. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).
153. Tavtigian, S. V., Greenblatt, M. S., Goldgar, D. E., Boffetta, P. & IARC Unclassified Genetic Variants Working Group. Assessing pathogenicity: overview of results from the IARC Unclassified Genetic Variants Working Group. *Hum. Mutat.* **29**, 1261–1264 (2008).
154. Easton, D. F. *et al.* A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am. J. Hum. Genet.* **81**, 873–883 (2007).

155. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3**, 8 (2014).
156. Carleton, K. L. Quantification of transcript levels with quantitative RT-PCR. *Methods Mol. Biol. Clifton NJ* **772**, 279–295 (2011).
157. Biesecker, L. G. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: Lessons from the ClinSeqTM project. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **14**, 393–398 (2012).
158. Schneider, T. D. Information content of individual genetic sequences. *J. Theor. Biol.* **189**, 427–441 (1997).
159. Mucaki, E. J., Ainsworth, P. & Rogan, P. K. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Hum. Mutat.* **32**, 735–742 (2011).
160. Bonnet-Dupeyron, M.-N. *et al.* PLP1 splicing abnormalities identified in Pelizaeus-Merzbacher disease and SPG2 fibroblasts are associated with different types of mutations. *Hum. Mutat.* **29**, 1028–1036 (2008).
161. Fei, J. Splice Site Mutation-Induced Alteration of Selective Regional Activity Correlates with the Role of a Gene in Cardiomyopathy. *J. Clin. Exp. Cardiol.* **S12:004**, (2013).
162. Khan, S. G. *et al.* Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. *Hum. Mol. Genet.* **13**, 343–352 (2004).
163. von Kodolitsch, Y., Berger, J. & Rogan, P. K. Predicting severity of haemophilia A and B splicing mutations by information analysis. *Haemoph. Off. J. World Fed. Hemoph.* **12**, 258–262 (2006).
164. Martoni, E. *et al.* Identification and characterization of novel collagen VI non-canonical splicing mutations causing Ullrich congenital muscular dystrophy. *Hum. Mutat.* **30**, E662–672 (2009).
165. Nasim, M. T. *et al.* Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. *Hum. Mutat.* **32**, 1385–1389 (2011).

166. Pink, A. E. *et al.* Mutations in the γ -secretase genes NCSTN, PSENEN, and PSEN1 underlie rare forms of hidradenitis suppurativa (acne inversa). *J. Invest. Dermatol.* **132**, 2459–2461 (2012).
167. Sanders, D. A., Ross-Innes, C. S., Beraldi, D., Carroll, J. S. & Balasubramanian, S. Genome-wide mapping of FOXM1 binding reveals co-binding with estrogen receptor alpha in breast cancer cells. *Genome Biol.* **14**, R6 (2013).
168. Suga, Y. *et al.* Lamellar ichthyosis with pseudoexon activation in the transglutaminase 1 gene. *J. Dermatol.* **42**, 642–645 (2015).
169. Rogan, P. K., Svojanovsky, S. & Leeder, J. S. Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics* **13**, 207–218 (2003).
170. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
171. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
172. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank: update. *Nucleic Acids Res.* **32**, D23–26 (2004).
173. Rogan, P. & Mucaki, E. Population Fitness and Genetic Load of Single Nucleotide Polymorphisms Affecting mRNA splicing. *ArXiv11070716 Q-Bio* (2011). at <<http://arxiv.org/abs/1107.0716>>
174. Di, L.-J., Fernandez, A. G., De Siervi, A., Longo, D. L. & Gardner, K. Transcriptional regulation of BRCA1 expression by a metabolic switch. *Nat. Struct. Mol. Biol.* **17**, 1406–1413 (2010).
175. Wu, X., Webster, S. R. & Chen, J. Characterization of tumor-associated Chk2 mutations. *J. Biol. Chem.* **276**, 2971–2974 (2001).
176. Durocher, D., Henckel, J., Fersht, A. R. & Jackson, S. P. The FHA domain is a modular phosphopeptide recognition motif. *Mol. Cell* **4**, 387–394 (1999).
177. Bell, D. W. *et al.* Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. *Science* **286**, 2528–2531 (1999).

178. Varley, J. M., Evans, D. G. & Birch, J. M. Li-Fraumeni syndrome--a molecular and clinical review. *Br. J. Cancer* **76**, 1–14 (1997).
179. Lee, S. B. *et al.* Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome. *Cancer Res.* **61**, 8062–8067 (2001).
180. Biswas, D. K. *et al.* NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 10137–10142 (2004).
181. Antoniou, A. C. & Easton, D. F. Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898–5905 (2006).
182. Peto, J. Breast cancer susceptibility—A new look at an old model. *Cancer Cell* **1**, 411–412 (2002).
183. Kurian, A. W. *et al.* Clinical Evaluation of a Multiple-Gene Sequencing Panel for Hereditary Cancer Risk Assessment. *J. Clin. Oncol.* **32**, 2001–2009 (2014).
184. Kassahn, K. S., Scott, H. S. & Caramins, M. C. Integrating Massively Parallel Sequencing into Diagnostic Workflows and Managing the Annotation and Clinical Interpretation Challenge. *Hum. Mutat.* **35**, 413–423 (2014).
185. Li, M.-X., Gui, H.-S., Kwan, J. S. H., Bao, S.-Y. & Sham, P. C. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.* gkr1257 (2012). doi:10.1093/nar/gkr1257
186. Kluska, A. *et al.* New recurrent BRCA1/2 mutations in Polish patients with familial breast/ovarian cancer detected by next generation sequencing. *BMC Med. Genomics* **8**, 19 (2015).

Chapter 4

4 Prioritizing variants in complete hereditary breast and ovarian cancer (HBOC) genes in patients lacking known *BRCA* mutations

The work in this chapter has been submitted for publication as:

Caminsky NG, Mucaki EJ, Perri AM, Lu R, Knoll JHM, Rogan PK. Prioritizing variants in complete hereditary breast and ovarian cancer (HBOC) genes in patients lacking known *BRCA* mutations. *Human Mutation* (2015).

Research Ethics Board (REB) approval was provided for this study (**Appendix F**) for this study. Documentation provided to patients upon invitation to the study (letters of information and response card) are provided in the Appendix (**G-J**).

4.1 Introduction

Currently, the lifetime risk for a woman to develop breast cancer (BC) is 12.3% and 1.3% in the case of ovarian cancer (OC¹). Approximately 5-10% of all BC cases are hereditary in nature, versus 25% for OC, where relative risk (RR) of BC or OC with one affected 1st degree family member is estimated at 2.1 and 3.1, respectively^{2,3}. Two highly penetrant genes, *BRCA1* and *BRCA2*, are associated with a large proportion of HBOC cases. However, the estimated rate of linkage to these genes is significantly higher than the proportion of pathogenic mutations identified in HBOC families⁴, suggesting unrecognized or unidentified variants in *BRCA1/2*.

Clinical *BRCA1/2* testing is restricted primarily to coding regions. Limitations on how variants can be interpreted, lack of functional validation, and mutations in other genes contribute to uninformative results. The heritability that is not associated with *BRCA* genes is likely due to other genetic factors rather than environmental causes, specifically moderate- and low-risk susceptibility genes⁵. Hollestelle *et al.* (2010) point out the

challenges in estimating increased risks associated with mutations in these genes, as the disease patterns are often incompletely penetrant, and require large pedigree studies to confidently assess pathogenicity⁶.

Next-generation sequencing (NGS) of gene panels for large cohorts of affected and unaffected individuals has become an increasingly popular approach to confront these challenges. Numerous HBOC gene variants have been catalogued, including cases in which RR has been determined; however the literature is also flooded with variants lacking a clinical interpretation⁷. It is not feasible to functionally evaluate the effects all of the VUS identified by NGS. Several approaches have been developed to better assess variants from exome and genome-wide NGS data^{8,9}. Nevertheless, there is an unmet need for other methods that quickly and accurately bridge variant identification and classification. To begin to address this problem, we have presented a unified IT framework to identify and prioritize variants in coding and non-coding regions of *BRCA1*, *BRCA2*, and 5 other HBOC genes (*ATM*, *CDH1*, *CHEK2*, *PALB2*, and *TP53* [Mucaki *et al.*, submitted]). This distinguishes prioritized variants from flagged or common alleles that affect regulatory protein binding and coding sequences in 70.6% of patients from a cohort of 102 *BRCA*-negative, anonymized HBOC patients.

In the present study, we have selected 13 additional genes that are being evaluated as hereditary BC loci (*BARD1* [BRCA1 Associated RING Domain 1], *EPCAM* [Epithelial Cell Adhesion Molecule], *MLH1* [MutL Homolog 1], *MRE11A* [MRE11 Meiotic Recombination 11 Homolog A], *MSH2* [MutS Homolog 2], *MSH6* [MutS Homolog 6], *MUTYH* [MutY Homolog], *NBN* [Nibrin], *PMS2* [Postmeiotic Segregation Increased 2], *PTEN* [Phosphatase And Tensin Homolog], *RAD51B* [RAD51 Paralog B], *STK11* [Serine/Threonine Kinase 11], and *XRCC2* [X-Ray Repair Complementing Defective Repair In Chinese Hamster Cells 2]¹⁰). These genes encode proteins with roles in DNA repair, surveillance, and cell cycle regulation (**Figure 4.1**; for further evidence supporting this gene set see **Supplementary Table 23**^{11,12}), and are also associated with specific disease syndromes that confer an increased risk of BC and OC, as well as many other types of cancer (**Supplementary Table 24**). High-risk genes confer > 4-times increased

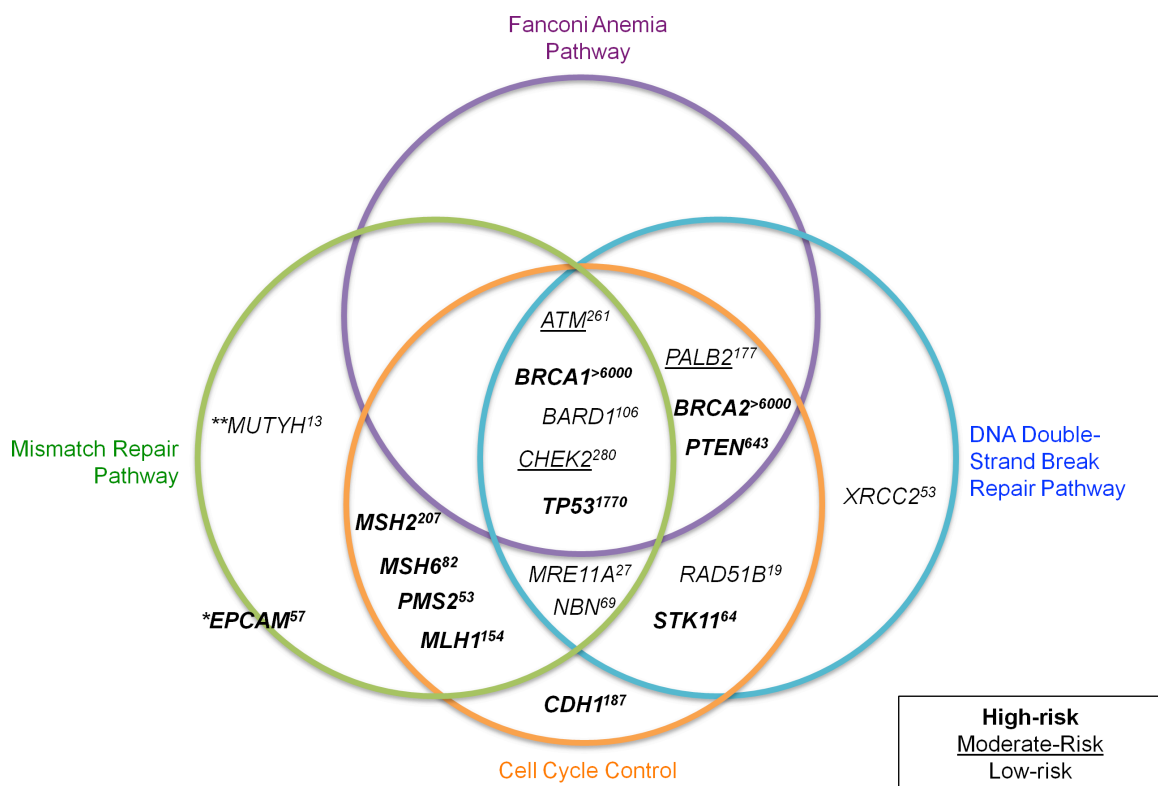


Figure 4.1. Significant Genome Stabilizing Pathways, Risk, and Relevant Literature for 20 HBOC Genes

The green, purple, and blue circles indicate sequenced genes that play important roles in the MMR, Fanconi Anemia and DNA double-strand break repair pathways, respectively. The orange circle contains genes involved in cell cycle control. Genes considered to present a high risk of breast and/or ovarian cancer when mutated are bolded, moderate-risk genes are underlined, and low-risk genes are in normal font. The estimated number of articles listing a gene's association with breast or ovarian cancer (based on a systematic search in PubMed [performed June 2015]) is indicated in superscript. ***MUTYH** is only high risk in the case of bi-allelic mutations. ****EPCAM** is not involved in any pathways, but is associated with hereditary non-polyposis colorectal cancer (HNPCC) by virtue of the fact that 3' deletions of *EPCAM* can cause epigenetic silencing of *MSH2*, causing Lynch syndrome protein. See **Supplementary Table 23** for citations and further evidence supporting this gene set.

risk of BC compared to the general population. *BRCA1* and *BRCA2* are estimated to increase risk by 20-fold¹³. Pathogenic variants in other high-risk genes, *CDH1*, *PTEN*, *STK11*, and *TP53*, are rarely seen outside of their associated syndromes, and account for < 1% of hereditary BC cases¹⁴. *EPCAM*, *MLH1*, *MSH2*, *MSH6*, and *PMS2* have also been proposed to harbor high-risk BC alleles, but the RR is still controversial¹⁴. Genes with moderate-risk alleles, *ATM*, *CHEK2*, and *PALB2*, cause between a 2- and 4-fold increased risk of BC^{11,14}. The remaining genes (*BARD1*, *MRE11A*, *MUTYH*, *NBN*, *RAD51B*, and *XRCC2*) are newly identified and currently associated with unknown risks for HBOC (**Figure 4.1**).

We report NGS of hybridization-enriched, complete genic and surrounding regions of 20 known HBOC-associated genes followed by variant analysis of 287 consented patients from Southwestern Ontario, Canada with previously uninformative HBOC susceptibility test results. We then reduced the set of gene variants in each individual by prioritizing the results of coding and IT-analyses. After applying a frequency-based filter, the IT-based framework prioritizes variants based on their predicted effect on the recognition of sequence elements involved in mRNA splicing, transcription, and untranslated region (UTR) binding, combined with UTR 2° structure and coding variant analysis. Our approach integrates disparate sources of information, including bioinformatic analyses, likelihood ratios based on familial segregation, allele frequencies, and published findings to prioritize disease-associated mutation candidates.

4.2 Methods

4.2.1 Ethics and Patient Recruitment

Recruitment and consent of human participants was approved by the University of Western Ontario Research Ethics Board (Protocol 103746). Patients were enrolled from January, 2014 through March, 2015 at London Health Sciences Centre (LHSC). Patients met the following criteria: male or female, aged between 25 and 75 years, > 10% risk of having an inherited mutation in a breast/ovarian cancer gene, diagnosed with BC and/or

OC, and previously receiving uninformative results for a known, pathogenic *BRCA1* or *BRCA2* variant in either the patient or other relatives.

The median age of onset for patients (N=287; **Figure 4.2**) with BC was 48 (N=277), and 46 for OC (N=17), and 7 were diagnosed with both BC and OC. Furthermore, 31 patients had bilateral BC (98 patients at diagnosis; 23 developed tumors on the opposite side after the initial occurrence), 1 had bilateral OC, and 13 have had recurrent BC in the same breast. There was a single case of male BC (**Supplementary Table 25**).

4.2.2 Probe Design, Sample Preparation, and Sequencing

Probes for sequence capture were designed as described in Mucaki *et al.*, (submitted) covering a total of 1,103,029 nt across the 21 sequenced genes, including negative control gene *ATP8B1* (see **Appendix K: Supplementary Information** for gene names, GenBank accession numbers, and OMIM reference numbers). This set of genes was proposed for evaluation at the Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) Consortium Meeting (2013). Other genes that have been found to be mutated in HBOC could not be included (eg. *BRIP1*, *RAD50*, *RAD51C*, *RAD51D*¹⁵⁻¹⁷).

Patient DNA extracted from peripheral blood was either obtained from the initial genetic testing at LHSC Molecular Genetics Laboratory or isolated from recent samples. NGS libraries were prepared using modifications to a published protocol¹⁸ described in Mucaki *et al.*, (submitted), except all post-capture pull-down steps were automated (**Supplementary Information**).

Library preparation and re-sequencing were repeated for samples with initial average coverage below our minimum threshold (< 30x). To ensure that the proper sample was re-sequenced, the variant call format (VCF) files from each run were compared to all others in the run using VCF-compare (<http://vcftools.sourceforge.net/>). VCF files from separate runs for the re-sequenced patient were concordant, except for minor differences in variant call rates due to differences in coverage. The aligned reads from both runs were then merged (with BAMtools; <http://sourceforge.net/projects/bamtools/>).

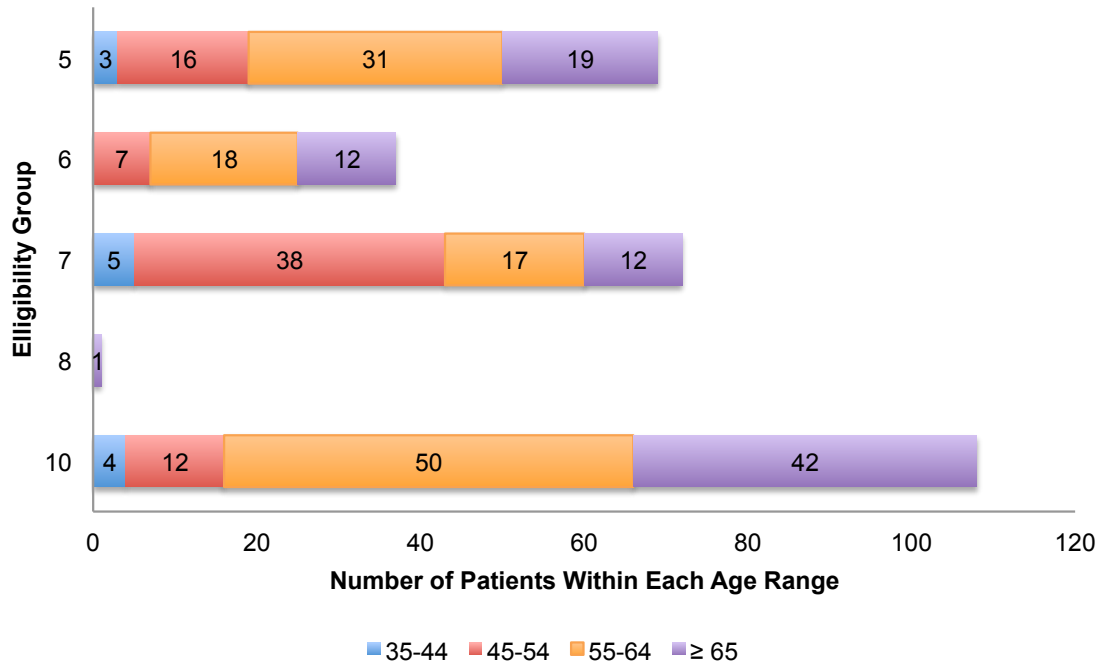


Figure 4.2. Distribution of Patients by Eligibility Group and Age

The number of patients falling within each age group (blue: 35-44 years, red: 45-54 years; orange: 55-64 years, and purple: 65 years or older) is indicated for each eligibility group.

Samples were demultiplexed and aligned using CASAVA (Consensus Assessment of Sequencing and Variation; v1.8.2¹⁹) and CRAC (Complex Reads Analysis & Classification; v1.3.0 [<http://crac.gforge.inria.fr/>]). Variants were then called using the Genome Analysis Toolkit (GATK [<https://www.broadinstitute.org/gatk/>]). Variants flagged by bioinformatic analysis (described in Mucaki *et al.*, submitted) were also assessed by manual inspection with the Integrative Genome Viewer v2.3 (IGV; <http://www.broadinstitute.org/igv/>).

4.2.3 Information Models

Models for natural splice sites (SSs) and splicing regulatory factors (SRFs) are described in Mucaki *et al.*, (2013)²⁰. These models were used to predict deleterious effects on natural splicing, the activation of cryptic SSs, and changes to binding of splicing enhancers and silencers. In addition, using a combination of cryptic site activation and hnRNPA1 site prediction, pseudoexon formation was also assessed.

We previously built models for TFBSs (N=83) using ENCODE ChIP-seq data (Mucaki *et al.*, submitted). We identified 8 additional transcription factors (TFs) with ChIP-seq evidence that they bind and regulate the additional genes; however models that passed quality control criteria could only be derived for 3 of these. Models could not be generated for the remaining TFs (N=5) as these models either consisted of noise reflecting the binding motif of an interacting TF cofactor, or the factor is a histone deacetylase, which may be detected by ChIP-seq, despite not directly binding DNA (**Supplementary Information**). **Supplementary Table 26** contains the list of TFs (N=86), and indicates which genes have experimental evidence of binding.

Information weight matrices for sequences bound by RNA-binding proteins (RBPs) from the CISBP-RNA (<http://cisbp-rna.ccbr.utoronto.ca/>) and RBPDB (<http://rbpdb.ccbr.utoronto.ca/>) databases were used to determine changes in binding affinity due to SNVs, with conservative information thresholds described in Mucaki *et al.*, (submitted). Finally, predicted changes in UTR structure due to a variant were

determined using SNPfold²¹. Significant changes in UTR structure and stability were represented using mfold (<http://unafold.rna.albany.edu/?q=mfold>).

4.2.4 Variant Analysis

Information analysis of binding site R_i changes and minimum information levels used to flag variants affecting protein binding, and variant prioritization criteria are outlined in Mucaki *et al.*, (submitted). To assess coding changes affecting predicted protein chain length or amino acid(s) composition, we used SNPnexus (<http://snp-nexus.org/>). Insertion/deletions (indels) and nonsense mutations were noted, and missense variants were further assessed with *in silico* tools, by referencing the published literature, and consulting mutation databases (listed in **Supplementary Table 27**; see Mucaki *et al.*, [submitted] for details on variant analysis).

EPCAM mutations in familial cancer are limited to 3' deletions causing epigenetic silencing of *MSH2*, and there is currently no evidence of other types of variants that alter its mRNA transcript or protein product²². Therefore, with the exception of indels, none of the variants flagged in *EPCAM* were prioritized. We chose to prioritize variants in *MUTYH* using the same framework as all other genes, despite *MUTYH* pathogenicity resulting from biallelic variants²³, because it is possible that a second *MUTYH* mutation remains unrecognized.

All protein truncating (nonsense and indels), and selected splicing and missense mutations were tested and where possible confirmed by Sanger sequencing (details in **Supplementary Table 28**).

4.2.4.1 Negative Control

Variants present in the *ATP8B1* gene were used as negative controls for our variant analysis framework. Initially, it was included in the list of prioritized HBOC genes provided by ENIGMA, but evidence for its association with HBOC is lacking in the published literature. Furthermore, it is not a known susceptibility gene for any type of cancer (mutations in *ATP8B1* cause progressive familial intrahepatic cholestasis²⁴), and is

infrequently mutated in breast tumors in several studies (for example, see Cancer Genome Atlas Network [2012]²⁵).

4.2.5 Likelihood Ratios (LRs)

Patients with prioritized coding and/or splicing variants, which we consider the most likely to be pathogenic, were selected for co-segregation analysis using an online tool that calculates the likelihood of a variant being deleterious based on pedigree information²⁶. Because the penetrance parameters cannot be altered from the settings given for *BRCA1* or *BRCA2*, the *BRCA2* option was selected for patients with prioritized variants in non-*BRCA* genes. Penetrance in *BRCA2* is known to be lower than *BRCA1* values²⁶. Current evidence suggests that mutations in non-*BRCA* genes may be less penetrant than those in the *BRCA* genes¹¹, however the penetrance of many of these variants remains unknown (**Supplementary Information**).

4.3 Results

4.3.1 Variant Analysis

We identified 38,372 unique variants among 287 patients (26,636 intronic, 7,287 intergenic, and 714 coding), on average 1,975 variants per patient, before any filtering criteria were applied.

4.3.1.1 Natural Site Variants

The Shannon Human Splicing Mutation Pipeline was used to predict the effect of the 14,458 variants that could potentially affect splicing, of which 244 reduced natural SS strength. Further stringent filtering of the natural SS based on information content changes and allele frequency resulted in 7 flagged variants (**Supplementary Table 29**).

Four of these variants were prioritized (**Table 4.1**). A novel synonymous variant in exon 2 of *RAD51B*, c.84G>A (p.Gln28=; 8-1H.9-1E), is predicted to increase exon skipping by weakening the natural splice donor ($R_{i,final} = 5.2$ bits, $\Delta R_i = -3.0$ bits). A known *ATM* variant, c.6198+1G>A (8-1D.9-1B²⁷), abolishes the natural donor SS of exon 42 ($R_{i,final} = -13.7$ bits, $\Delta R_i = -18.6$ bits). This will either lead to exon skipping or activation of a pre-

Table 4.1. Prioritized Variants Predicted by IT to Affect Natural and Cryptic Splicing

UWO ID	Gene	Variant	rsID (dbSNP142) Allele Frequency (%) [†]	Information Change			Consequence
				$R_{i,initial}$ (bits)	$R_{i,final}$ (bits)	ΔR_i (bits)	
8-1D.9-1B	<i>ATM</i>	c.6198+1G>A ^{1,2}	-	4.9	-13.7	-18.6	Abolished natural ^{a,d}
12-4E.13-5B	<i>MRE11A</i>	c.2070+2A>T*	-	7.6	-11	-18.6	Abolished natural ^{a,d}
10-4D	<i>MLH1</i>	c.306+4A>G* ³	rs267607733	8.6	6	-2.6	Weakened natural ^b
8-1H.9-1E	<i>RAD51B</i>	c.84G>A* p.Gln28=	Novel	8.2	5.2	-3	Weakened natural ^a
8-1D.9-1B	<i>BARD1</i>	c.1454C>T* p.Ala485Val	Novel	-2.7	4.4	7.1	Created cryptic ^b
15-4A	<i>BRCA1</i>	c.5074+107C>T	rs373676607	-1.3	5.7	7	Created cryptic ^{c,e}
15-3G	<i>CDH1</i>	c.1223C>G* ⁴ p.Ala408Gly	Novel	-0.6	4.3	4.9	Created cryptic ^b
10-4B	<i>RAD51B</i>	c.958-29A>T***	rs34436700 0.78	2.2	4.4	2.2	Strengthened cryptic ^c
10-5A	<i>STK11</i>	c.375-194GT>AC	rs35113943 17.61 rs117211142 0.80	7.5	8.8	1.3	Strengthened cryptic ^c
8-2F	<i>XRCC2</i>	c.122-154G>T	Novel	8.1	10	1.9	Strengthened cryptic ^c

*Confirmed by Sanger sequencing; ***Ambiguous Sanger sequencing results; †If available; ^aexon skipping; ^bexon truncation; ^cintron retention; ^duse of alternate isoform; ^ereduced expression of natural isoform; ¹Stankovic *et al.*, 1998 (Ref 27), ²Reiman *et al.*, 2011 (Ref 28), ³Tournier *et al.*, 2008 (Ref 29), ⁴Schrader *et al.*, 2011 (Ref 32).

existing cryptic site (**Figure 4.3**). An Ataxia-Telangiectasia patient with this variant exhibited low expression, protein truncation, and abolished kinase activity of ATM²⁸. *MLH1* c.306+4A>G (rs267607733; 10-4D) causes increased exon skipping (and a decrease in wild-type exon relative expression) due to the weakening ($R_{i,final} = 6.0$ bits, $\Delta R_i = -2.6$ bits) of the exon 3 natural donor. Tournier *et al.*, (2008) assessed this variant using an *ex vivo* splicing assay and observed cryptic site activation and exon 3 skipping²⁹. *MRE11A* c.2070+2A>T (12-4E.13-5B) is indicated in ClinVar as likely pathogenic and abolishes the natural donor site of exon 19 ($R_{i,final} = -11.0$ bits, $\Delta R_i = -18.6$ bits), while strengthening a cryptic site 5 nt upstream of the splice junction ($R_{i,final} = 8.1$ bits, $\Delta R_i = 0.6$ bits). Either cryptic SS activation or complete exon skipping are predicted.

The *BRCA2* variant c.68-7T>A (rs81002830; 15-6G) was not prioritized as its pathogenicity has not been proven despite evidence of its induction of (in-frame) exon skipping³⁰, and was shown to not segregate with disease in Portuguese HBOC patients³¹. This same study did not observe abnormal splicing following RT-PCR experiments. The previously discussed *ATM* variant c.1066-6T>G (rs201686625; 8-3F.9-3F) was also not prioritized (Mucaki *et al.*, [submitted]).

4.3.1.2 Activation of Cryptic Splicing

The Shannon Pipeline identified 9,480 variants that increased the strength of at least one cryptic site, of which 9 met or exceeded the defined thresholds for information change. Six of these were prioritized (**Table 4.1**). A novel *BARD1* variant in exon 6 (c.1454C>T; p.Ala485Val; 8-1D.9-1B) creates a donor SS ($R_{i,final} = 4.4$ bits, $\Delta R_i = 7.1$ bits), which would produce a 58 nt frameshifted exon if activated. The natural donor SS of exon 6, 116 nt downstream of the variant, is stronger (5.5 bits), but the ASSEDA server predicts equal levels of expression of both natural and cryptic exons. A *BRCA1* mutation 5074+107C>T (rs373676607; 15-4A) downstream of exon 16 is predicted to extend the exon by 105 nt, and be slightly more abundant than the natural exon ($R_{i,total}$ of 8.6 and 8.1 bits, respectively). *CDHI* c.1223C>G (p.Ala408Gly; 15-3G), previously reported in a *BRCA*-negative lobular BC patient with no family history of gastric cancer³², creates a

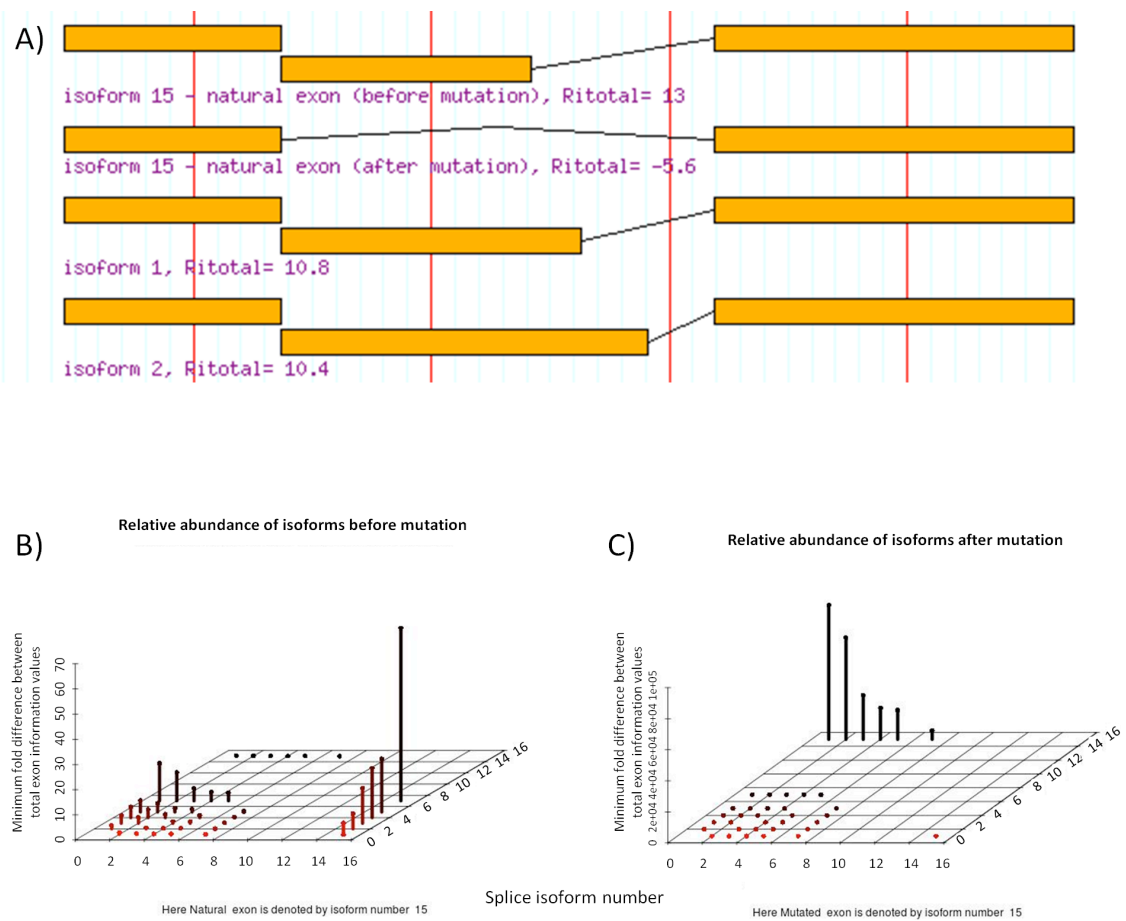


Figure 4.3. Predicted Isoforms and Relative Abundance as a Consequence of *ATM* natural splice variant c.6198+1G>A

A) Intronic *ATM* variant c.6198+1G>A abolishes the natural donor of exon 42 (4.9 -> -13.7 bits), and would either result in exon skipping (causing a frame-shift), or possibly activate a downstream cryptic site (isoform 1 maintains reading frame, isoform 2 would not). **B)** In the wild type mRNA, the natural exon (isoform 15) has the highest predicted abundance. **C)** The mutation predicts isoform 1 to be the most abundant and isoform 2 to be slightly less abundant than 1.

cryptic donor site ($R_{i,final} = 4.3$ bits, $\Delta R_i = 4.9$ bits) in exon 9, 97 nt downstream of the natural acceptor. While residual splicing of the normal exon is still expected, the cryptic is predicted to become the predominant splice form (~twice as abundant; **Figure 4.4**).

STK11 c.375-194GT>AC (rs35113943 & rs117211142; 10-5A), and the novel *XRCC2* c.122-154G>T (8-2F) both strengthen strong pre-existing cryptic sites exceeding the $R_{i,total}$ values of their respective natural exons. Finally, a known *RAD51B* variant 29 nt upstream of exon 10: c.958-29A>T (rs34436700; 10-4B) strengthens a cryptic acceptor ($R_{i,final} = 4.4$ bits, $\Delta R_i = 2.2$ bits) that, if activated, would produce a transcript retaining 21 intronic nucleotides and exon 10.

The remaining cryptic site variants (**Supplementary Table 29**) were not prioritized. The novel *BRCA2* c.7618-269_7618-260del10 (7-4A) variant is predicted to create a cryptic site with an exon having a lower $R_{i,total}$ value (5.2 bits) than the natural exon (6.6 bits). *PMS2* c.1688G>T (p.Arg563Leu; rs63750668; 15-6A, 15-3B, 15-4B) does not segregate with disease. Drost *et al.*, (2013) demonstrated that this variant does not impair DNA repair activity³³. Finally, *RAD51B* c.728A>G (p.Lys243Arg; rs34594234; 7 patients) predicts an increase in the abundance of the cryptic exon; however the natural exon remains the predominant isoform.

4.3.1.3 Pseudoexon Activation

Pseudoexons arise from creation or strengthening of an intronic cryptic SS in close proximity to another intron site of opposite polarity. Our analysis detected 623 variants with such intronic cryptic sites, of which 17 were prioritized (among 9 genes) occurring within 250 nt of a pre-existing site of opposite polarity, with an hnRNPA1 site within 5 nt of the acceptor of the predicted pseudoexon (**Supplementary Table 30**). Three are novel (*BRCA2* c.7007+824C>T, *BRCA2* c.8332-1130G>T, and *PTEN* c.802-796C>A), and the remainder were present in dbSNP. Seven of these variants (*BARD1* c.1315-168C>T, *BRCA2* c.631+271A>G, *MLH1* c.1559-1732A>T, *MRE11A* c.1783+2259A>G, *MSH6* c.260+1758G>A, *PTEN* c.79+4780C>T, and *RAD51B* c.1037-1012C>A), although rare,

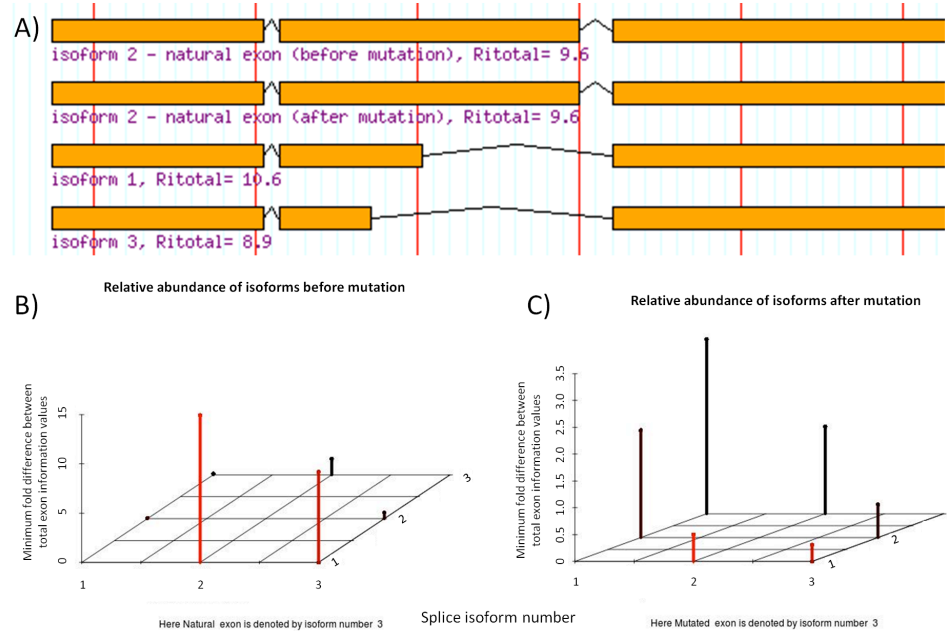


Figure 4.4. Predicted Isoforms and Relative Abundance as a Consequence of *CDH1* cryptic splice variant c.1223C>G

A) A missense variant within exon 9 of *CDH1* (c.1223C>G; p.Ala408Gly; chr16:68847301C>G; 15-3G) is novel and creates a cryptic donor (-0.6 -> 4.3 bits). If used (as predicted by ASSEDA), the resultant cryptic exon would be 97 nt shorter (isoform 1) and cause a frameshift. **B)** In the wild type mRNA, the natural exon (isoform 2) as the most abundant splice form. **C)** The mutation causes cryptic isoform 1 to be most abundant.

occur in multiple patients, and patient 12-5H has predicted pseudoexons in both *BARD1* and *RAD51B*.

4.3.1.4 SRF Binding

Variants within exons or within 500 nt of a natural SS (N=9,998) were assessed for their potential effect on SRF binding sites (SRFBSs). Initially 216 unique variants were flagged (**Supplementary Table 31**), but after considering each in the context of the SRF function and location within the gene³⁴, we prioritized 148, of which 57 are novel. Some prioritized variants affect distant SRFs that may activate cryptic sites, but were not predicted to affect natural splicing. Of the 88 suitable prioritized variants for which exon definition analysis was performed (where initial $R_{i,total}$ of the exon > SRF gap surprisal value), 55 were predicted to induce or contribute to increased exon skipping. For example, an uncommon *ATM* missense variant within exon 41, c.6067G>A (p.Gly2023Arg; rs11212587; 10-3H), strengthens an hnRNPA1 site ($R_{i,final} = 5.2$ bits, $\Delta R_i = 4.7$ bits) 30 nt from the natural donor, and is predicted to induce exon 41 skipping ($\Delta R_{i,total} = -9.5$ bits).

4.3.1.5 TF binding

To assess potential changes to TFBSs, variants occurring from 10 kb upstream of the start of transcription through the end of the first intron were analyzed by IT, flagging 88 (of 4,530 identified; **Supplementary Table 32**). Considering the gene context of each TFBS and extent of information change prioritized 36 variants. The following illustrates the rationale for highlighting these variants: *BRCAl* c.-19-433A>G (rs191197821; 15-6H) abolishes a binding site for HSF 1 ($R_{i,initial} = 5.5$ bits, $\Delta R_i = -7.8$ bits). While HSF 1 is known to be a transcriptional activator associated with poor BC prognosis³⁵, the specific effect of reduced HSF 1 binding to *BRCAl* has not been established. Similarly, *MLH1* c.-4285T>C (rs115211110; 10-6B.11-1B, 10-2C, 10-5D, 12-2G, and 13-6A) significantly weakens a C/EBP β site ($R_{i,initial} = 10.1$ bits, $\Delta R_i = -6.3$ bits), a TF that has been shown to play a role in BC development and progression³⁶. Another *MLH1* variant, c.-6585T>C (novel; 15-5H), greatly decreases the binding strength ($R_{i,initial} = 12.5$ bits, $\Delta R_i = -10.8$

bits) of the NF- κ B p65 subunit, which is activated in ER-negative breast tumors³⁷. Two prioritized variants (*PMS2* c.-9059G>C and *XRCC2* c.-163C>A) weaken PAX5 binding sites, a TF which when overexpressed can result in mammary carcinoma cells regaining epithelial cell characteristics³⁸.

4.3.1.6 Alterations to mRNA Structure

A total of 1,355 variants were identified in the 5' and 3' UTRs of the patients. Analysis of the variants, most likely to alter mRNA structure with SNPfold, flagged 3 unique variants ($p < 0.05$), in *BRCA1*, *BARD1*, and *XRCC2* (**Table 4.2**). The predicted mRNA 2° structures of the reference and variant sequences are shown in **Figure 4.5** (generated with mfold). (generated with mfold). The *BRCA1* variant occurs in the 3'UTR of all known transcript isoforms (NM_007294.3:c.*1332T>C; rs8176320; in 8-5C.9-5C, 8-6A.9-6A, 10-2G). The most likely inferred structure consisting of a short arm and a larger stem loop is destabilized when the variant nucleotide is present (**Figure 4.5A and B**). The *BARD1* variant falls within the 5' UTR of a rare isoform (XM_005246728.1:c.-53G>T; rs143914387; 8-3E.9-3E, 8-6F.9-6F, 8-4H.9-4H, 13-2E, 15-4C), and is within the coding region of a more common transcript (NM_000465.2:c.33G>T). While the top ranked isoform following mutation is similar to the wild-type structure, the second-ranked isoform ($\Delta G = +1.88\text{kcal/mol}$) is distinctly different, creating a loop in a long double-stranded structure (**Figure 4.5C and D**) The *XRCC2* variant is within its common 5' UTR (NM_005431.1:c.-76C>T; rs547538731; 15-2D) and is located 11 nt downstream from the 5' end of the mRNA. The variant nucleotide disrupts a potential GC base pair, leading to a large stem-loop that could allow access for binding of several RBPs (**Figure 4.5E and F**). The variant simultaneously strengthens a PUM2 ($R_{i,initial} = 2.8$ bits, $\Delta R_i = 4.4$ bits) and a RBM28 site ($R_{i,initial} = 4.0$, $\Delta R_i = 3.6$ bits), however there is a stronger NCL site (8.3 bits) in the area that is not affected and may compete for binding.

4.3.1.7 RBP binding

Using IT models of 76 RBBSs, 33 UTR variants were prioritized (**Supplementary Table 33**) from the initial list of 1,367 UTR variants. Interestingly, one of the three variants that destabilized the mRNA was also flagged using our RBP scan. The *BARD1* c.53A>C

Table 4.2. Variants Predicted by SNPfold to Significantly Affect UTR Structure

UWO ID	Gene	Variant	UTR Position	rsID (dbSNP142) Allele Frequency (%) [†]	Rank	<i>p</i> -value
8-3E.9-3E 8-6F.9-6F 8-4H.9-4H 13-2E 15-4C	<i>BARD1</i>	XM_005246728.1:c.- 53G>T (c.33G>T p.Gln11His)	5'UTR	rs143914387 0.04	6/600	0.01
8-5C.9-5C 8-6A.9-6A 10-2G	<i>BRCA1</i>	NM_007294.3:c.*1332T>C NM_007299.3:c.*1438T>C	3'UTR	rs8176320 0.42	13/450	0.03
15-2D	<i>XRCC2</i>	NM_005431.1:c.-76C>T	5'UTR	rs547538731 0.08	3/300	0.01

[†]If available

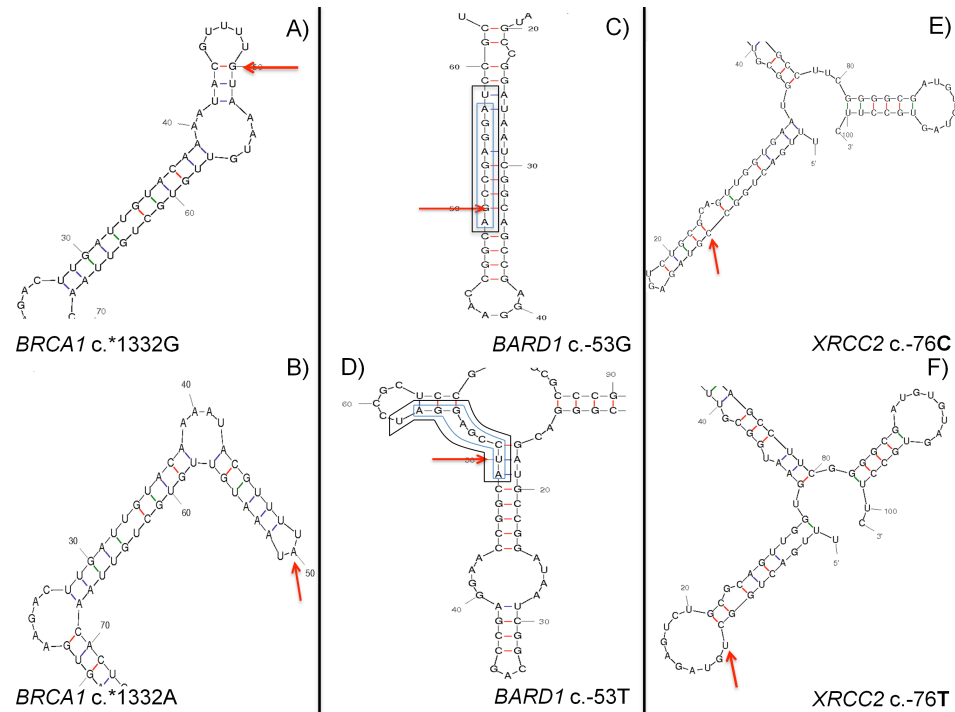


Figure 4.5. Predicted RNA Structure Change due to Variants Flagged by SNPfold using mfold

Wild-type (A, C, and E) and variant (B, D and F) structures are displayed. The variant nucleotide is marked with a red arrow. **A)** Predicted wild-type structure of *BRCA1* 3'UTR surrounding c.*1332G>A. **B)** *BRCA1* 3'UTR structure due to c.*1332A variant, extending arm length while reducing hairpin size. **C)** *BARD1* 5'UTR structure of rare isoform (XM_005246728.1:c.-53G>T). Two overlapping pre-existing RBP sites (SRSF7 [outer box] and SRSF2 [inner box]) are predicted and either could occupy this location if accessible. **D)** 2° *BARD1* 5' UTR structure of the region predicted only with sequence containing the c.-53T mutation. The primary predicted c.-53T structure is identical to wild-type (with one disrupted C-G bond leading to a 4.1 kcal/mol lower ΔG). The variant both weakens and abolishes the pre-existing SRSF7 and SRSF2 sites, respectively. **E)** *XRCC2* structure within common 5'UTR surrounding c.-76C>T variant. **F)** *XRCC2* 5'UTR structure predicted from c.-76T sequence, containing a hairpin not found in wild-type. This hairpin may allow for the binding of previously inaccessible nucleotides including the altered nucleotide.

variant weakens a predicted 8.3 bit SRSF7 site ($\Delta R_i = -3.0$ bits) while simultaneously abolishing a predicted 9.7 bit SRSF2 site ($\Delta R_i = -29.7$ bits) (**Figure 4.5C-D**).

4.3.2 Exonic Protein-Altering Variants

4.3.2.1 Protein Truncating

Of the 714 identified coding variants, 6 were indels, each of which found in a single patient, and 2 preserve the reading frame. In addition, 5 nonsense variants were found in 6 different patients. All of these variants were prioritized unless otherwise stated (**Table 4.3** and **Supplementary Table 34**).

A novel insertion, c.3550_3551insA (p.Gly1184Glufs; 11-6H), identified in exon 10 of *BRCA1*, causes premature termination and loss of 676 amino acids, including both BRCT domains, and the domain responsible for binding PALB2³⁹. Two other frameshifting variants are previously reported *PALB2* deletions: c.757_758delCT (p.Leu253Ilefs; rs180177092; 10-6F), reported in a Fanconi Anemia patient⁴⁰ and c.2920_2921delAA (p.Lys974Glufs; rs180177126; 8-3A.9-3A), identified in a *BRCA*-negative HBOC patient⁴¹. These variants are predicted to cause the loss of 932 and 208 amino acids in the protein, respectively. Consequently, the WD repeat regions are lost (only 5/7 are lost in the case of the shorter deletion) along with domains important for the interaction with RAD51 and BRCA2. We found another frameshift mutation in the last exon of *BRCA2*: c.10095delCins11 (p.Ser3366Asnfs; rs276174803; 15-4E). Borg *et al.*, (2010) report this variant as a VUS in a HBOC patient due to its proximity to the carboxy terminus of the coding region⁴², however it is classified as likely benign in ClinVar, and was therefore not prioritized.

The two frame-preserving mutations are 3 nt deletions occurring in *CDH1* and *CHEK2*. The first, c.30_32delGCT (p.Leu11del; 10-4A), is novel, occurs in exon 1, and deletes a leucine from the signal peptide domain. The second variant, c.483_485delAGA (p.Glu161del; 10-1E), deletes a glutamic acid from the FHA domain, necessary for

Table 4.3. Variants Resulting in Premature Protein Truncation

UWO ID	Gene	Exon	Variant	rsID (dbSNP142) Allele Frequency (%) [†]	Details
Frameshift Insertions/Deletions					
11-6H	<i>BRCA1</i>	10 of 23	c.3550_3551insA** p.Gly1184Glufs	Novel	STOP at p.1187 676 AA short
10-6F	<i>PALB2</i>	4 of 13	c.757_758delCT* p.Leu253Ilefs	rs180177092	STOP at p.255 932 AA short
8-3A.9-3A	<i>PALB2</i>	9 of 13	c.2920_2921delAA* p.Lys974Glufs	rs180177126	STOP at p.979 208AA short
Insertions/Deletions with Conserved Reading Frame					
10-4A	<i>CDH1</i>	1 of 16	c.30_32delGCT*** p.Leu11del	Novel	Loss of 1 AA Frame and AA sequence conserved
10-1E	<i>CHEK2</i>	4 of 14	c.483_485delAGA* p.Glu161del	-	Loss of 1 AA Frame and AA sequence conserved
Stop Codons					
10-2F	<i>ATM</i>	13 of 63	c.1924G>T* p.Glu642Ter	-	2415 AA short

12-4G.13-5D	<i>ATM</i>	62 of 63	c.8977C>T* p.Arg2993Ter	-	64 AA short
8-5D.9-5D	<i>BRCA1</i>	23 of 23	c.5503C>T** p.Arg1835Ter	rs41293465	28 AA short
15-1E	<i>PALB2</i>	13 of 13	c.3549C>G* p.Tyr1183Ter	rs118203998	4 AA short

*Confirmed by Sanger sequencing; **Not confirmed through Sanger sequencing; ***Ambiguous Sanger sequencing results;

†If available; AA: amino acid

protein-protein interaction. This variant reduces CHEK2 phosphorylation in response to DNA damage, and protein instability and occurs in HBOC⁴³.

All of the nonsense mutations in the cohort (N=5) have been previously reported. *ATM*: c.1924G>T (p.Glu642Ter; 10-2F) was identified in a BC family by Goldgar *et al.*, (2011)⁴⁴ and results in a loss of 2,415 amino acids. *ATM* c.8977C>T (p.Arg2993Ter; 12-4G.13-5D) similarly truncates the protein by 64 amino acids and has been reported in cases of Ataxia-Telangiectasia^{45,46}. *BRCA1* c.5503C>T (p.Arg1835Ter; rs41293465; 8-5D.9-5D) was described by Dong *et al.*, (1998) in two HBOC patients⁴⁷ and occurs in the terminal exon. *PALB2* c.3549C>G (p.Tyr1183Ter; rs118203998; 15-1E) also occurs in the terminal exon and has been reported in BC⁴⁸. Finally, *BRCA2* c.9976A>T was identified in 2 patients (p.Lys3326Ter; rs11571833; 12-2H and 13-3C) and is likely benign due to its low odds ratio (1.01)⁴⁹.

4.3.2.2 Missense Variants

Of the 155 unique missense variants (**Supplementary Table 35**), 119 were prioritized by consulting literature and disease- and gene-specific databases. All are of unknown clinical significance and 21 have not been previously reported.

The *ATM* variant c.7271T>G (p.Val2424Gly; rs28904921; 10-1F, 12-1D) replaces a hydrophobic residue by glycine in the conserved FAT domain, in which is required for ATM activation, and confers a 9-fold increase (95% CI) in BC risk⁴⁴. Functional studies, assessing ATM kinase activity *in vitro* with TP53 as a substrate, showed that cell lines heterozygous for the mutation had less than 10% of wild-type kinase activity, such that this variant is expected to act in a dominant-negative manner⁵⁰.

CHEK2 c.433C>T (p.Arg145Trp; rs137853007; 4-3C.5-4G.14-4A) and c.470T>C (p.Ile157Thr; rs17879961; 12-2G, 15-5G) both fall within the FHA domain, which mediates ATM-dependent phosphorylation of CHEK2 and targets the protein to other binding partners such as BRCA1⁵¹. c.433C>T results in rapid degradation of the mutant protein⁵².

The *PMS2* variant c.2T>C (p.Met1Thr; 11-4H) is listed in ClinVar as pathogenic and would be expected to abrogate correct initiation of translation. This variant has not been reported in BC families, but is associated with colorectal cancer (CRC)⁵³.

4.3.3 Variant Prioritization

We prioritized an average of 18.2 variants in each gene, ranging from 7 (*XRCC2*) to 61 (*ATM*), an average of 0.41 variants/kb, and an average of 0.65 variants/patient (**Table 4.4**). *ATM* had the second greatest gene probe coverage (103,511 nt captured), the highest number of unique prioritized variants, and was among the top genes for number of prioritized variants/kb (0.59).

In total, our framework allowed for the prioritization of 346 unique variants in 246 patients, such that 85.7% of tested patients (N=287) had at least 1 prioritized variant. Most patients (84.7%) harbored fewer than 4 prioritized variants. The distribution of patients with prioritized variants did not significantly differ across the eligibility groups (**Table 4.5**). **Figure 4.6** illustrates the distribution of these variants by category, and **Supplementary Tables 36** and **37** show this information by gene and patient, respectively.

All prioritized protein-truncating (N=10), and selected splicing (N=7) and missense (N=5) variants were verified by Sanger sequencing. Of the protein-truncating variants, 4 nonsense, 1 indel with a conserved reading frame, and 2 frameshifts were confirmed (**Table 4.3**). Six splicing variants and all missense were confirmed (**Table 4.1** and **Supplementary Table 35**).

4.3.4 Negative Control

ATP8B1 was sequenced and analyzed in all patients as a negative control (**Supplementary Table 38**). We prioritized 21 *ATP8B1* variants with an average of 0.22 variants/kb and 0.57 variants/patient. This is lower than the prioritization rate for many of the documented HBOC genes, but illustrates that prioritization is a screening approach that can generate false positives.

Table 4.4. Comparing Counts of Prioritized Variants

Gene	Unique prioritized variants	Unique patients	Gene probe coverage (nt)	Prioritized variants/patient	Prioritized variants/kB
<i>ATM</i>	61	102	103511	0.60	0.59
<i>ATP8B1</i>	21	37	94793	0.57	0.22
<i>BARD1</i>	17	46	73735	0.37	0.23
<i>BRCA1</i>	19	24	52075	0.79	0.36
<i>BRCA2</i>	24	28	73332	0.86	0.33
<i>CDH1</i>	21	32	61312	0.66	0.34
<i>CHEK2</i>	12	13	28372	0.92	0.42
<i>MLH1</i>	18	25	50553	0.72	0.36
<i>MRE11A</i>	17	31	64713	0.55	0.26
<i>MSH2</i>	18	17	112437	1.06	0.16
<i>MSH6</i>	19	23	25216	0.83	0.75
<i>MUTYH</i>	8	16	21439	0.50	0.37
<i>NBN</i>	11	21	57067	0.52	0.19

<i>PALB2</i>	26	46	25319	0.57	1.03
<i>PMS2</i> *	8	15	11726	0.53	0.68
<i>PTEN</i> **	15	23	86059	0.65	0.17
<i>RAD51B</i> ***	22	47	62465	0.47	0.35
<i>STK11</i>	12	20	28373	0.60	0.42
<i>TP53</i>	11	30	23544	0.37	0.47
<i>XRCC2</i>	7	10	19942	0.70	0.35

* high homology to other regions in the genome, thus fewer probes designed within gene; ** *PTEN* has pseudogene *PTENP1*, thus fewer probes covering exonic regions; *** probes limited to 1,000 nt surrounding all exons, and 10,000 nt up- and downstream of gene

Table 4.5. Distribution of Recruited Patients Among Eligibility Groups

Eligibility Group [†]	Number of Patients within Eligibility Group	Number of Patients with Prioritized Variants
Breast cancer <60 year, and a first or second-degree relative with ovarian cancer or male breast cancer (5).	68	62
Breast and ovarian cancer in the same individual, or bilateral breast cancer with the first case <50 years (6).	37	32
Two cases of ovarian cancer, both <50 years, in first or second-degree relatives (7).	72	59
Two cases of ovarian cancer, any age, in first or second-degree relatives (8).	1	1
Three or more cases of breast or ovarian cancer at any age (10).	109	92
	287	246

The Risk Categories for Individuals Eligible for Screening for a Genetic Susceptibility to Breast or Ovarian Cancers are determined by the Ontario Ministry of Health and Long Term-Care Referral Criteria for Genetic Counseling. † Numbers in parentheses correspond to Eligibility Group designation.

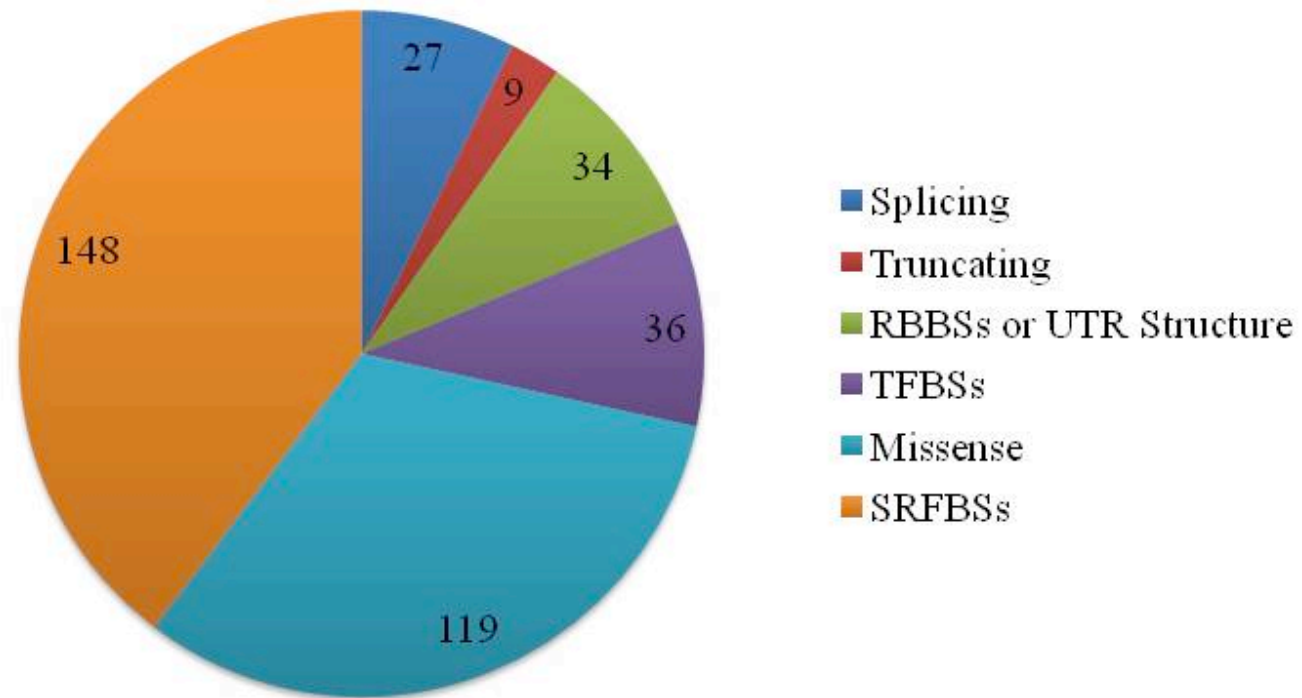


Figure 4.6. Distribution of Unique Prioritized Variants by Category

Truncating variants include indels and nonsense variants. Splicing variants include significantly weakening or abolishing natural SSs, activating cryptic SSs, or activating the formation of pseudoexons.

4.3.5 Pedigree Analysis

Pathogenic *BRC A2* variants within a region of exon 11 have been associated with a high incidence of OC. We therefore verified whether there were a high number of OC cases in the families of patients prioritized with exon 11 *BRC A2* variants (N=3). The family of 13-6H (c.4828G>A; p.Val1610Met; diagnosed with BC at 65) has 3 reported cases of BC/OC, 1 of which is OC (diagnosed at 74), 2 degrees of separation from the proband. Patient 12-1H (c.6317T>C; p.Leu2106Pro; diagnosed with BC at 52) has 3 other affected family members, 2 with OC and 1 with BC. Finally, patients 8-4B, 9-4B, 10-2A, 11-2F, and 14-3C, 15-2A (c.5199C>T; p.Ser1733=) do not have any family members with reported cases of OC.

We also selected patients with prioritized mismatch repair (MMR) variants (N=8, in 10 patients) to assess the incidence of reported CRC cases in these families. Notably, patient 8-1E (*MSH2* c.1748A>G) had 5 relatives with CRC. A similar analysis of prioritized *CDHI* variants did not reveal any patients with a family history of gastric cancer.

4.3.6 Likelihood Ratio Analyses

We carried out co-segregation analysis of 25 patients with prioritized pathogenic variants (4 nonsense, 4 frameshift, 2 in-frame deletions, 6 missense, 4 natural splicing, and 6 cryptic splicing, including 8-1D/9-1B who exhibited prioritized natural and cryptic SS variants). We compared these findings with those from patients (N=25) harboring moderate-priority variants (variants prioritized through IT analysis only) and those in whom no variants were flagged or prioritized (N=14). In instances where disease alleles could be transmitted through either founder parent, the lineage with the highest LR was reported. For patients with likely pathogenic variants, the LRs ranged from 0.00 to 70.96 (**Table 4.6** and **Supplementary Table 39**). Disease co-segregation was supported (LR > 1.0) in 18 patients, and the remainder were either neutral (LR < 1.0²⁶) or could not be analyzed either due to missing pedigree information or limited numbers of affected individuals in a family. Patient 10-6F (*PALB2*: c.757_758delCT) exhibited the highest likelihood (LR=70.96). Prioritized variants with neutral evidence include a variant that

Table 4.6. LR Values for Patients with Prioritized Truncating, Splicing, and Selected Missense Variants

Gene	Variant		Category	UWO ID	LR
	mRNA	Protein			
<i>ATM</i>	c.1924G>T	p.Glu642Ter	Nonsense	10-2F	7.46 ^{MGM} 9.61 ^{MGF}
	c.6198+1G>A	-	Natural splicing	8-1D.9-1B	1.00
	c.7271T>G	p.Val2424Gly	Missense	10-1F	1.44
				12-1D	1.96 ^P
	c.8977C>T	p.Arg2993Ter	Nonsense	12-4G.13-5D	5.30 ^P
<i>BARD1</i>	c.1454C>T	p.Ala485Val	Cryptic splicing	8-1D.9-1B	1.00
<i>BRCA1</i>	c.3550_3551insA	p.Gly1184Glufs	Frameshift indel	11-6H	3.36 ^P
	c.5503C>T	p.Arg1835Ter	Nonsense	8-5D.9-5D	41.99
<i>BRCA2</i>	c.10095delCins11	p.Ser3366Asnfs	Frameshift indel	15-4E	3.71
<i>CDHI</i>	c.30_32delGCT	p.Leu11del	Inframe deletion	10-4A	1.00
	c.1223C>G	p.Ala408Gly	Cryptic splicing	15-3G	2.14
<i>CHEK2</i>	c.470T>C	p.Ile157Thr	Missense	12-2G	2.86
				15-5G	19.44 ^P

	c.433C>T	p.Arg145Trp	Missense	4-3C.5-4G.14-4A	3.48
<i>PALB2</i>	c.3549C>G	p.Tyr1183Ter	Nonsense	15-1E	1.78
	c.757_758delCT	p.Leu253Ilefs	Frameshift indel	10-6F	70.96
	c.2920_2921delAA	p.Lys974Glufs	Frameshift indel	8-3A.9-3A	5.03
	<i>PMS2</i>	c.2T>C	p.Met1Thr	Missense	11-4H
<i>RAD51B</i>	c.84G>A	p.Gln28=	Leaky splicing	8-1H.9-1E	3.51 ^P
	c.958-29A>T	-	Cryptic splicing	10-4B	7.44 ^P
<i>STK11</i>	c.375-194GT>AC	-	Cryptic splicing	10-5A	2.67 ^M

Note: LR values in favor of neutrality are not shown. ^P paternal; ^M maternal; ^{MGF} maternal grandfather; ^{MGM} maternal grandmother.

abolishes a natural SS in *MRE11A*, c.2070+2T>A (12-4E/13-5B; LR=0.03), and an in-frame deletion c.483_485delAGA in *CHEK2* (10-1E; LR=0.00).

4.4 Discussion

Rare non-coding and/or non-truncating mutations can confer an increased risk of disease in BC⁵⁴. This study determined both coding and non-coding sequences of 20 HBOC-related genes, with the goal of discovering and prioritizing rare variants with potential effects on gene expression. This work emphasizes results from the analysis of non-coding variants, which are abundant in these genes, yet have been underrepresented in previous HBOC mutation analyses. Nevertheless, alterations to mRNA binding sites in *BRCA*, and lower risk or rare HBOC genes, have been shown to contribute to HBOC (ESEs in *ATM*⁵⁵, *BARD1*⁵⁶, and *BRCA* genes^{57,58}). We prioritized 346 unique variants that were predicted to result in 4 nonsense, 3 frameshift, 2 indels with preserved reading frame, 119 missense, 4 natural splicing, 6 cryptic splicing, 17 pseudoexon activating, 148 SRFBS, 36 TFBS, 3 UTR structure, and 31 RBBS mutations (**Supplementary Table 36**). Among these variants, 101 were novel (**Supplementary Table 40**). Compared to our initial 7-gene panel (Mucaki *et al.*, [submitted]), the inclusion of the additional genes in this study prioritized at least 1 variant in 15% additional patients (increased from 70.6 to 85.7%).

The *BRCA* genes harbor the majority of known germline pathogenic variants for HBOC families⁵⁹. However, a large proportion of the potentially pathogenic variants identified in our study were detected in *ATM*, *PALB2*, and *CHEK2*, which although of lower penetrance, were enriched because the eligibility criteria excluded known *BRCA1* and *BRCA2* carriers. *BRCA1* and *BRCA2* variants were nevertheless prioritized in some individuals. We also had expected intragenic clustering of some *BRCA* coding variants⁶⁰. For example, pathogenic variants occurring within exon 11 of *BRCA2* are known to be associated with higher rates of OC in their families⁶¹. We identified 3 variants in exon 11, however there was no evidence of OC in these families. Overall, *ATM* and *PALB2* had the highest number of prioritized variants (61 and 26, respectively). However, only 12 variants were prioritized in *CHEK2*; potentially pathogenic variants may have been

under-represented during sequence alignment as a consequence of the known paralogy with *CHEK2P2*.

Fewer *TP53*, *STK11*, and *PTEN* variants were prioritized, as pathogenic variants in these genes tend to be infrequent in patients who do not fulfill the clinical criteria for their associated syndromes (Li-Fraumeni syndrome, Peutz-Jeghers syndrome, and Cowden syndrome, respectively⁶). Although the density of prioritized variants in these genes is below average (18.2 per gene), the total number was significant (*TP53*=11, *STK11*=12, *PTEN*=15).

Certain missense variants show stronger penetrance than truncating variants, which is contrary to the notions about the severity of truncating variants. For example, the missense variant *ATM* c.7271T>G (p.Val2424Gly; 10-2F, 12-1D), has been shown to act in a dominant-negative manner and confer a higher risk of developing BC than some truncating mutations in this gene⁴⁴.

We compared the frequency of all prioritized variants in our patient cohort to the population allele frequencies (1000 Genomes Project, Phase 3) to determine if variants more common in our cohort might be suggestive of HBOC association. Three variants in at least 5 HBOC patients are present at a much lower frequency in the general population. The *NBN* c.*2129G>T (4.18% of study cohort), is considerably rarer globally (0.38% in 1000Genomes; < 0.1% in other populations). Conversely, *BARD1* c.33G>T (1.74%), has only been detected in the American and European populations (0.04%). In Southwestern Ontario, individuals are often of American or European ancestry. The allele frequency of this variant in our HBOC population may simply be enriched in a founder subset of general populations. Similarly, the *RAD51B* c.-3077G>T variant (2.09%), is rare in the European and one sub-South Asian population (0.08%). While we cannot rule out skewing of these allele frequencies due to population stratification, our findings suggest that gene expression levels could be impacted by these variants.

Co-segregation analysis is recommended by the American College of Medical Genetics and Genomics (ACMG) for variant classification⁶². Among patients with likely

pathogenic, highly penetrant mutations in our cohort, some variants had LR values consistent with causality, whereas others provided little evidence to support co-segregation among family members (**Table 4.6** and **Supplementary Table 39**). An important caveat however was that use of *BRCA2* penetrance values in non-*BRCA* genes may have resulted in underestimates of LR values.

Co-segregation analysis was also performed on patients with moderate priority variants (i.e. variants affecting binding sites) and patients with no flagged or prioritized variants (N=25 and (**Figure 4.7**). The proportion of LR values supporting neutrality and those supporting causation were comparable across all three groups of patients (**Figure 4.7**). This suggests co-segregation analysis is only useful in the context of other supporting results for assessing pathogenicity (eg. likelihood of being pathogenic or benign).

A small number of patients with a known pathogenic variant carried other prioritized variants. These were likely benign or possibly, phenotypic modifiers. For example, patient 8-5D.9-5D possessed 5 prioritized variants (1 missense, 1 SRFBS, 1 TFBS, and 2 RBBSs) in addition to a *BRCAl* nonsense mutation (c.5503C>T). While these variants may not directly contribute to causing HBOC, they may act as a risk modifier and alter expression levels⁵.

Similarly, genes lacking association with HBOC can be used as a metric for determining a false-positive rate of variant prioritization. In this study, we prioritized 21 *ATP8B1* variants among 37 of our HBOC patients (**Supplementary Table 38**) despite it having not been previously associated with any type of cancer. A variant with a deleterious effect on *ATP8B1* may lead to *ATP8B1*-related diseases, such as progressive familial intrahepatic cholestasis²⁴, but should not increase the chances of developing BC. Thus, while our framework may be effective at prioritizing variants, only genes with previous association to a disease should be included in analyses similar to the present study to minimize falsely prioritized variants.

Additional workup of prioritized non-coding and non-*BRCA* variants is particularly important, because with few exceptions⁶³, the pathogenicity of many of the genes and

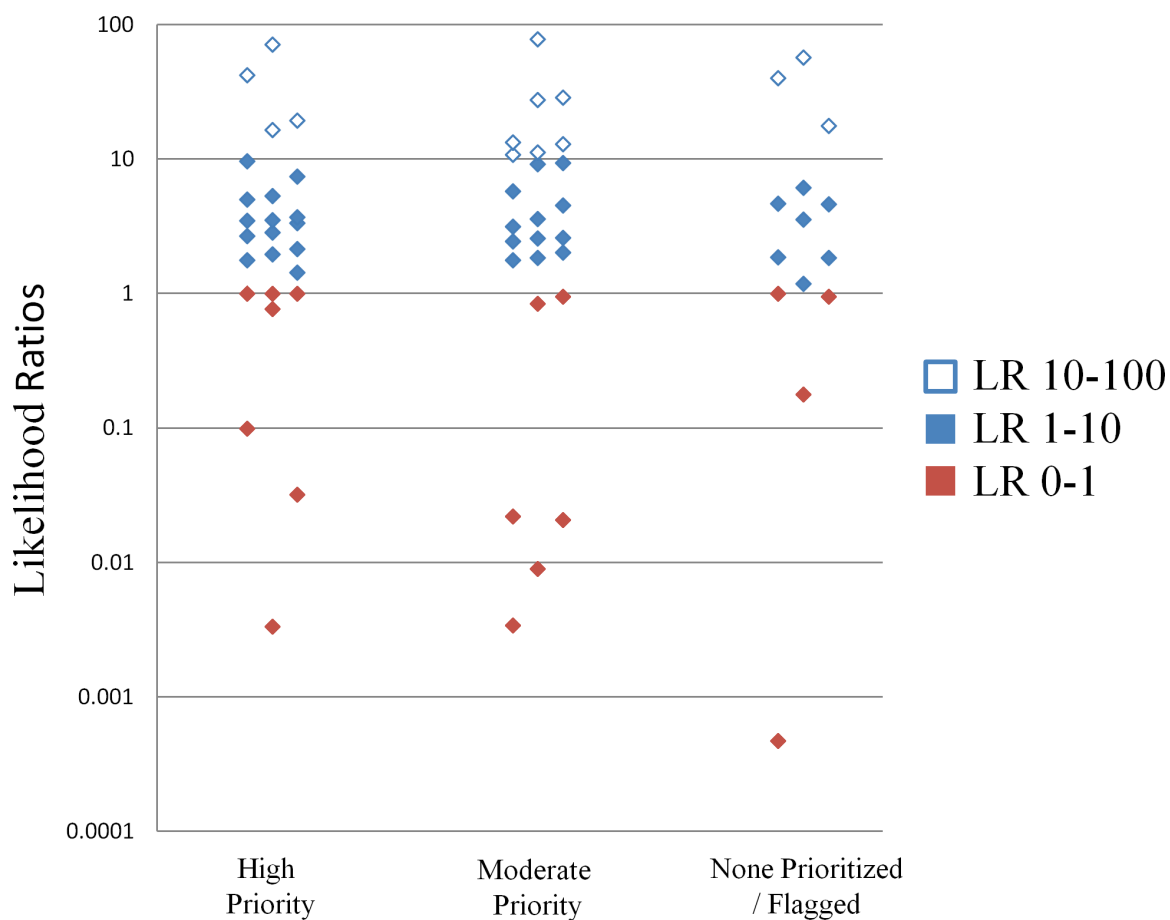


Figure 4.7. Computed Likelihood Ratios for Patients with Variants of Disparate Priority

Co-segregation analysis was performed using available pedigree information for patients with high priority variants (likely pathogenic), moderate priority variants, and individuals with no flagged or prioritized variants. The calculated likelihood ratios were graphed logarithmically. The distribution of values supporting neutrality ($LR \leq 1$) and those supporting causation ($LR > 1$) is similar between the three patient groups.

variants has not been firmly established. Furthermore, mutations in several of these genes confer risk to other types of cancer, which alters the management of these patients⁶⁴. The next step towards understanding the role these prioritized variants play in HBOC is to test family members of the proband and to carry out functional analysis. If this is not possible, then their effects on gene expression could be evaluated using assays for RNA stability and RNA localization. Protein function could be evaluated by binding site assays, protein activity, and quantitative PCR.

A significant challenge associated with VUS analysis, particularly in the case of many of these recent HBOC gene candidates, is the under-reporting of variants and thus positive findings tend to be over-represented in the literature⁶⁵. Hollestelle *et al.*, (2010) argue that a more stringent statistical standard must be applied (i.e. *P*-values of 0.01 should be used as opposed to 0.05) to under-reported variants (namely in moderate-risk alleles), because of failure to replicate pathogenic variants⁶, which we have also found⁶⁶. In the same way that we use IT-based analysis to justify prioritizing variants for further investigation, variants that are disregarded as lower priority (and that are likely not disease-causing) have been subjected to the same thresholds and criteria. Integrating this set of labeled prioritized and flagged, often rare variants, from this cohort of *BRCA*-negative HBOC patients, to findings from exome or gene panel studies of HBOC families should accelerate the classification of some VUS.

Reducing the full set of variants in a patient to a prioritized list is one approach for targeting clinically-relevant information. The ACMG recommends that known or likely pathogenic variants should be reported in 10 of the genes sequenced in this study (*BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, *MUTYH*, *PMS2*, *PTEN*, *STK11*, and *TP53*⁶⁷). However, the guidelines do not currently address interpretation of non-coding variation. Patients could be informed of prioritized VUS, which may increase patient accrual and participation⁶⁸. However, it will be critical to explain both the implications and significance of prioritization and the limitations, namely counselling patients to avoid clinical decisions based on this information⁶⁹.

4.5 References

1. Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2011, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2011/, based on November 2013 SEER data submission, posted to the SEER web site, April 2014.
2. Stratton, J. F., Pharoah, P., Smith, S. K., Easton, D. & Ponder, B. A. A systematic review and meta-analysis of family history and risk of ovarian cancer. *Br. J. Obstet. Gynaecol.* **105**, 493–499 (1998).
3. Walsh, T. *et al.* Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18032–18037 (2011).
4. Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **62**, 676–689 (1998).
5. Antoniou, A. C. & Easton, D. F. Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898–5905 (2006).
6. Hollestelle, A., Wasielewski, M., Martens, J. W. & Schutte, M. Discovering moderate-risk breast cancer susceptibility genes. *Curr. Opin. Genet. Dev.* **20**, 268–276 (2010).
7. Cassa, C. A. *et al.* Disclosing pathogenic genetic variants to research participants: Quantifying an emerging ethical responsibility. *Genome Res.* **22**, 421–428 (2012).
8. Duzkale, H. *et al.* A systematic approach to assessing the clinical significance of genetic variants. *Clin. Genet.* **84**, 453–463 (2013).
9. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
10. Minion, L. E. *et al.* Hereditary predisposition to ovarian cancer, looking beyond BRCA1/BRCA2. *Gynecol. Oncol.* **137**, 86–92 (2015).
11. Apostolou, P. & Fostira, F. Hereditary Breast Cancer: The Era of New Susceptibility Genes. *BioMed Res. Int.* **2013**, e747318 (2013).

12. Al Bakir, M. & Gabra, H. The molecular genetics of hereditary and sporadic ovarian cancer: implications for the future. *Br. Med. Bull.* **112**, 57–69 (2014).
13. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* **72**, 1117–1130 (2003).
14. Maxwell, K. N. & Domchek, S. M. Familial Breast Cancer Risk. *Curr. Breast Cancer Rep.* **5**, 170–182 (2013).
15. Heikkinen, K., Karppinen, S.-M., Soini, Y., Mäkinen, M. & Winqvist, R. Mutation screening of Mre11 complex genes: indication of RAD50 involvement in breast and ovarian cancer susceptibility. *J. Med. Genet.* **40**, e131 (2003).
16. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* **38**, 1239–1241 (2006).
17. Janatova, M. *et al.* Mutation Analysis of the RAD51C and RAD51D Genes in High-Risk Ovarian Cancer Patients and Families from the Czech Republic. *PloS One* **10**, e0127711 (2015).
18. Gnirke, A. *et al.* Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
19. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
20. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Hum. Mutat.* **34**, 557–565 (2013).
21. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* **6**, e1001074 (2010).
22. Ligtenberg, M. J. L. *et al.* Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nat. Genet.* **41**, 112–117 (2009).
23. Jones, S. *et al.* Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations. *Hum. Mol. Genet.* **11**, 2961–2967 (2002).

24. Gonzales, E., Spraul, A. & Jacquemin, E. Clinical utility gene card for: progressive familial intrahepatic cholestasis type 1. *Eur. J. Hum. Genet. EJHG* **22**, (2014).
25. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
26. Mohammadi, L. *et al.* A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example. *BMC Cancer* **9**, 211 (2009).
27. Stankovic, T. *et al.* ATM mutations and phenotypes in ataxia-telangiectasia families in the British Isles: expression of mutant ATM and the risk of leukemia, lymphoma, and breast cancer. *Am. J. Hum. Genet.* **62**, 334–345 (1998).
28. Reiman, A. *et al.* Lymphoid tumours and breast cancer in ataxia telangiectasia; substantial protective effect of residual ATM kinase activity against childhood tumours. *Br. J. Cancer* **105**, 586–591 (2011).
29. Tournier, I. *et al.* A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum. Mutat.* **29**, 1412–1424 (2008).
30. Théry, J. C. *et al.* Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *Eur. J. Hum. Genet. EJHG* **19**, 1052–1058 (2011).
31. Santos, C. *et al.* Pathogenicity Evaluation of BRCA1 and BRCA2 Unclassified Variants Identified in Portuguese Breast/Ovarian Cancer Families. *J. Mol. Diagn.* **16**, 324–334 (2014).
32. Schrader, K. A. *et al.* Germline mutations in CDH1 are infrequent in women with early-onset or familial lobular breast cancers. *J. Med. Genet.* **48**, 64–68 (2011).
33. Drost, M., Koppejan, H. & de Wind, N. Inactivation of DNA mismatch repair by variants of uncertain significance in the PMS2 gene. *Hum. Mutat.* **34**, 1477–1480 (2013).
34. Caminsky, N. G., Mucaki, E. J. & Rogan, P. K. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Research* **3**, 282 (2015).

35. Santagata, S. *et al.* High levels of nuclear heat-shock factor 1 (HSF1) are associated with poor prognosis in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18378–18383 (2011).
36. Zahnow, C. A. CCAAT/enhancer-binding protein beta: its role in breast cancer and associations with receptor tyrosine kinases. *Expert Rev. Mol. Med.* **11**, e12 (2009).
37. Biswas, D. K. *et al.* NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 10137–10142 (2004).
38. Vidal, L. J.-P. *et al.* PAX5alpha enhances the epithelial behavior of human mammary carcinoma cells. *Mol. Cancer Res. MCR* **8**, 444–456 (2010).
39. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12**, 68–78 (2012).
40. Reid, S. *et al.* Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat. Genet.* **39**, 162–164 (2007).
41. Casadei, S. *et al.* Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. *Cancer Res.* **71**, 2222–2229 (2011).
42. Borg, A. *et al.* Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum. Mutat.* **31**, E1200–40 (2010).
43. Sodha, N., Mantoni, T. S., Tavtigian, S. V., Eeles, R. & Garrett, M. D. Rare Germ Line CHEK2 Variants Identified in Breast Cancer Families Encode Proteins That Show Impaired Activation. *Cancer Res.* **66**, 8966–8970 (2006).
44. Goldgar, D. E. *et al.* Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Res. BCR* **13**, R73 (2011).
45. Li, A. & Swift, M. Mutations at the ataxia-telangiectasia locus and clinical phenotypes of A-T patients. *Am. J. Med. Genet.* **92**, 170–177 (2000).
46. Magliozzi, M. *et al.* DHPLC screening of ATM gene in Italian patients affected by ataxia-telangiectasia: fourteen novel ATM mutations. *Dis. Markers* **22**, 257–264 (2006).

47. Dong, J. *et al.* A high proportion of mutations in the BRCA1 gene in German breast/ovarian cancer families with clustering of mutations in the 3' third of the gene. *Hum. Genet.* **103**, 154–161 (1998).
48. Tischkowitz, M. *et al.* Rare germline mutations in PALB2 and breast cancer risk: a population-based study. *Hum. Mutat.* **33**, 674–680 (2012).
49. Mazoyer, S. *et al.* A polymorphic stop codon in BRCA2. *Nat. Genet.* **14**, 253–254 (1996).
50. Chenevix-Trench, G. *et al.* Dominant negative ATM mutations in breast cancer families. *J. Natl. Cancer Inst.* **94**, 205–215 (2002).
51. Li, J. *et al.* Structural and Functional Versatility of the FHA Domain in DNA-Damage Signaling by the Tumor Suppressor Kinase Chk2. *Mol. Cell* **9**, 1045–1054 (2002).
52. Lee, S. B. *et al.* Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome. *Cancer Res.* **61**, 8062–8067 (2001).
53. Senter, L. *et al.* The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. *Gastroenterology* **135**, 419–428 (2008).
54. Tavtigian, S. V. *et al.* Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am. J. Hum. Genet.* **85**, 427–446 (2009).
55. Heikkinen, K. *et al.* Association of common ATM polymorphism with bilateral breast cancer. *Int. J. Cancer* **116**, 69–72 (2005).
56. Ratajska, M. *et al.* Cancer predisposing BARD1 mutations in breast–ovarian cancer families. *Breast Cancer Res. Treat.* **131**, 89–97 (2011).
57. Gochhait, S. *et al.* Implication of BRCA2 -26G>A 5' untranslated region polymorphism in susceptibility to sporadic breast cancer and its modulation by p53 codon 72 Arg>Pro polymorphism. *Breast Cancer Res. BCR* **9**, R71 (2007).
58. Sanz, D. J. *et al.* A High Proportion of DNA Variants of BRCA1 and BRCA2 Is Associated with Aberrant Splicing in Breast/Ovarian Cancer Patients. *Clin. Cancer Res.* **16**, 1957–1967 (2010).
59. Chong, H. K. *et al.* The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. *PloS One* **9**, e97408 (2014).

60. Mucaki, E. J., Ainsworth, P. & Rogan, P. K. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Hum. Mutat.* **32**, 735–742 (2011).
61. Lubinski, J. *et al.* Cancer variation associated with the position of the mutation in the BRCA2 gene. *Fam. Cancer* **3**, 1–10 (2004).
62. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).
63. Easton, D. F. *et al.* Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372**, 2243–2257 (2015).
64. Knappskog, S. & Lønning, P. E. P53 and its molecular basis to chemoresistance in breast cancer. *Expert Opin. Ther. Targets* **16 Suppl 1**, S23–30 (2012).
65. Kraft, P. Curses--winner's and otherwise--in genetic epidemiology. *Epidemiol. Camb. Mass* **19**, 649–651; discussion 657–658 (2008).
66. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3**, 8 (2014).
67. Green, R. C. *et al.* ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15**, 565–574 (2013).
68. Murphy, J. *et al.* Public expectations for return of results from large-cohort genetic research. *Am. J. Bioeth. AJOB* **8**, 36–43 (2008).
69. Vos, J. *et al.* Opening the psychological black box in genetic counseling. The psychological impact of DNA testing is predicted by the counsees' perception, the medical impact by the pathogenic or uninformative BRCA1/2-result. *Psychooncology.* **21**, 29–42 (2012).

Chapter 5

5 General Discussion

The objective of this thesis was to show that IT-based analysis could be used uniformly for the interpretation and prioritization of variants affecting DNA and RNA binding sites. In Chapter 2, we sought to compile and summarize the complete body of literature surrounding IT-based splicing mutation analysis, covering over 300 peer-reviewed articles and nearly 2,000 individual variants. This review allowed us to illustrate the distribution of deleterious variants involving both natural and cryptic splicing. It also inspired the design of an open-access tool, called Splicing Mutation Calculator, which computes the information change for natural site variants and provides a list of published articles that have applied IT analysis to variants at the same position and nucleotide change as the user's input. The purpose of this review was not only to gain insight on the landscape of splicing variants and disease, but also to establish the accuracy of IT analysis for the interpretation of such variants. We demonstrate that 87.9% (N=867) of splicing variants that have been experimentally validated were accurately interpreted by information analysis. This gave us the grounds to seek expansion of this approach to other types of protein/nucleic acid binding events and have confidence in its ability to accurately interpret and quantify the consequence of sequence variants.

In Chapter 3, we generated information models for TFBS and RBBS, and aimed to apply IT-based analysis uniformly to sequence variants affecting splicing, TF and RBP binding, so that the consequence of all sequence changes would be evaluated using the same unit of measurement: bits of information. We applied our framework to an anonymous group of *BRCA*-negative HBOC patients. In addition, we designed custom probes targeting 7 HBOC genes (complete coding, non-coding, and 10 kb-surrounding regions) that were successfully used in solution hybridization to capture our sequences of interest prior to NGS. Our analysis, which included a comprehensive assessment of indel, nonsense, and missense variants, resulted in 70.6% (N=102) of patients having at least one prioritized variant.

In Chapter 4, we took our approach one step further by expanding our gene panel to 20 HBOC susceptibility genes, selected based on a consensus by the ENIGMA Consortium, and recruiting *BRCA*-negative HBOC patients to consensually participate in the study. In this portion of the project we optimized capture pull-down through automation and supplemented our analysis with patient pedigree information, used for co-segregation analysis. By increasing the number of genes investigated, we greatly improved the rate of variant prioritization, such that 86.1% (N=287) of patients had at least one prioritized variant.

5.1 Patient Recruitment and Participation

Between January 1st 2014 and March 25th 2015, 555 patients were invited through the Cancer Genetics Clinic at the LHSC to in this study (REB Protocol 103746). Following invitation, 21 patients were later found to be deceased, 48 declined, 4 were later deemed ineligible because their family history was not accurate, and 48 letters were returned (because the patient no longer resided at the mailing address used).

In total 292 (53%) provided verbal and written consent to participate. Of these, 10 were previously sequenced in the anonymous group of patients provided by the MGL for runs UWO2-7. To date, 287 patients have been sequenced by our lab, and the remainder will be completed at a later date, by another member of our lab.

Through personal communication with the patients, it was clear that those participating were highly motivated and interested in the study, predominantly out of concern for relatives (mainly daughters, nieces, etc.). This was demonstrated by many patients calling just after the six-month period following enrolment, wondering if their results were ready. One drawback in the design of this study was that patients were told that they could expect results within 6-8 months of enrolment. Due to delays involving sample preparation and sequencing, as well as improvements to binding models and the variant interpretation framework, patients had to wait approximately 18 months before receiving results. In general, patients were understanding, however disappointed that they would need to wait longer.

5.2 Variant Prioritization in the Context of ACMG Guidelines

Currently, the variant analysis approach presented in this thesis integrates 4 types of data (population, computational/predictive, other database, and other data) from the ACMG Framework (**Table 5.1**), and 5-6 if provided with additional DNA samples from other family members and/or given the opportunity to perform functional analysis. When considering the IT-analysis component in isolation, variants affecting protein-binding sites (SRFBSs, TFBSs, RBBSs) can only be considered to “support” pathogenicity, based on the classification criteria presented in **Table 1.5** of *Chapter 1*. This is because the definition of moderate-to-very strong support is limited to missense and truncating variants. According to ACMG guidelines, at least one “moderate” line of evidence must be met (in addition to 4 “supporting” criteria) to consider a variant as likely pathogenic and therefore unless other data is available on a given variant, IT-based analysis on its own will result in the variant remaining of uncertain significance (see **Table 1.7** of *Chapter 1*).

That being said, by developing a framework for variant prioritization, our objective was to close the gap that currently exists between variant identification and classification, as opposed to allowing direct classification as likely pathogenic/pathogenic. We do this by creating a prioritized “short list” of variants that should undergo functional (or so-segregation) analysis. Functional analysis demonstrating a deleterious effect (along with two additional supporting lines of evidence) is sufficient to classify a variant as likely pathogenic. However, with the number of VUS identified in large-scale studies (see *Chapter 3*) that additional workload associated with functional validation of variants is unacceptable for any setting, whether clinical or research. Our approach is meant to effectively streamline the functional validation process (as demonstrated by an average 13.4-fold reduction in the number of flagged variants per patient) and future work demonstrating the accuracy of IT-based variant analysis is required. Should this be the case, we anticipate that our approach would encourage functional validation of VUS by removing the risk of it being a shot in the dark.

Table 5.1. Evidence Framework from the ACGM.

	Benign			Pathogenic		
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in <i>trans</i> with a dominant variant BP2 Observed in <i>cis</i> with a pathogenic variant BP2		For recessive disorders, detected in <i>trans</i> with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

This chart organizes each of the criteria by the type of evidence as well as the strength of the criteria for a benign (left side) or pathogenic (right side) assertion. BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.⁵

⁵Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015). Copy of license agreement for Table re-use is provided from Nature Publishing Group (see **Appendix B**).

5.3 Improving Variant Prioritization and Alternative Approaches

As discussed in Chapter 4, others have presented the concept of variant prioritization. Each example varies either by the *in silico* tools or thresholds used and genic regions selected for analysis. With the approach described in this thesis, we consider our approach unique in that it the *in silico* analysis performed for protein binding sites is equivalent for all factors. More specifically, different factors can be compared to one another because the impact of a variant is assessed using the same unit of measurement, new factors can be included, and models can be continuously updated, and improves as new data becomes available. Using changes in information is also more reliable than relying on conservation data, as it computes changes related to entropy, a key concept of classical thermodynamics.

However, our approach by no means “solves” the issue of VUS, nor is it infallible at filtering out all benign variants. At the moment, variant identification using NGS technologies is leaps and bounds ahead of variant interpretation capabilities. However, it is simply a question of time before reliable and affordable high-throughput technology becomes available for the assessment of variants on sequence binding and other important cellular processes. Until this time, other approaches are necessary to bridge this gap, substantiating the development of our variant prioritization schema.

Our framework currently allows for the assessment of a broad range of consequences, from protein truncation to binding sites associated with splicing, transcription, and transcript stability. Expansion of our approach would allow for increased variant analysis robustness, the prioritization of additional variants, and more confident filtering of non-prioritized variants. Analysis could be extended to include microRNA regulation, histone binding, and promoter methylation. Information from additional databases would also contribute to a more thorough assessment of variants, especially missense and synonymous coding changes. For example, genome-wide association studies employ next-generation sequencing of large disease and control cohorts to identify variants with a

strong likelihood of being associated with disease. While these variants are not defined as disease-causing, an association would lend additional support for prioritization over rejection. Online Mendelian Inheritance in Man is another source of disease-associated variants that could be used to improve the assessment of variants within our framework.

An alternative to variant prioritization would be to emphasize the variants that have been filtered out and are not expected to have a deleterious impact on gene expression and disease. As mentioned in *Chapter 3*, studies almost exclusively highlight pathogenic variants and neglect to report or stress variants *not* expected to contribute to disease. Should this practice change, and efforts be made to catalog variants with evidence of being benign, this information could be a valuable addition to the step in variant prioritization where variants are subjected to a technical and population filter. In the interim, an additional step could be included in variant prioritization, whereby variants are subjected to a secondary round of prioritization. This second round would take into consideration variants that a) co-occur with known pathogenic variants (and can be assumed to be benign or disease-modifiers based on the idea that HBOC is a monogenic, autosomal dominant disease) or b) occur in a given percentage of patients in the study cohort. If this process eventually became automated, the user could potentially pre-set the various thresholds for the population frequency filters, and information changes filters for that matter.

An alternate approach to variant prioritization could be the integration of high-throughput functional assays. The interpretation of a variant's consequence would be key to streamlining this process, because while high throughput assays allow for an exponential decrease in experimentation time, these assays are expensive and a systematic assessment of every variant identified through NGS remains unfeasible. This functional analysis could also follow a prioritization schema, as there are many options available in terms of the type of consequence that can be probed. In the cases where expression levels are predicted to be altered due to a significant ΔR_i of a binding site, it would be sensible to first assess changes in binding affinity *in vitro*^{1,2}.

Should the results indicate a significant change in binding affinity, the next step could be more specific to the variant in question. For example, if the gene is involved in DNA damage repair, specific assays exist that allow for the quantification of DNA repair efficiency³. The advantage with all of these assays is that patient RNA is not required, as the variant oligos or proteins can be synthesized and/or expressed *in vitro*.

That being said, should fresh DNA and RNA samples be collected upon patient enrollment, there is the possibility of employing a less labor-intensive approach to high-throughput binding studies, a technique that our lab is currently developing (P.K. Rogan, personal communication). Briefly, RNA oligos for both the wild-type and variant sequences are synthesized, biotinylated, and incubated with cell or protein extracts, followed by streptavidin pulldown. Selected reaction monitoring⁴, which uses tandem mass spectroscopy⁵ to quantitatively measure the amount of a specific protein within a sample, can then be used to differentiate between RNA-protein interactions of binding sites with different strengths (i.e. before and after mutation).

Another option, also requiring patient DNA and RNA, involves a modified version of a software program developed by our lab called Veridical⁶ (an *in silico* method for the automatic validation of DNA sequencing variants that alter mRNA splicing). This program would allow for the determination of allele specific effects on transcript levels. Currently, Veridical performs a statistically valid comparison of normalized read counts between RNA-seq data from the patient and a sample lacking the sequence variant being evaluated. The modified program would count reads and compute the probability that the levels of expression of a mutant allele would differ from normal, regardless of whether a particular splice form is altered, providing direct molecular phenotypic evidence supporting or refuting an IT-based prediction.

5.4 Future Directions

5.4.1 Controls

Throughout the course of this project, we had both a positive and negative control for the purpose of our IT analysis. As a positive control, we sequenced and assessed an HBOC

patient with a known deleterious *BRCA* variant. We were also provided with a list of SNVs identified by the MGL using Sanger sequencing. We used this information to test the accuracy of our capture, sequencing, and variant calling methods. Moving forward, using a positive control for our variant interpretation framework would be desirable, however challenging, as it would require sequencing information on patients with variants known to disrupt protein/nucleic acid binding sites.

With regard to the positive control, we included in our gene panel a gene (*ATP8B1*) that is not known to be associated with BC or OC (neither sporadic, nor hereditary) in any way. Using the assumption that any variants prioritized in *ATP8B1* (aside from the previously-identified pathogenic variant) were likely not disease-causing, we wanted to have an idea of the “background” rate of false-positive prioritization. While not truly false-positive (because transcript levels may still result from a variant), these variants are not expected to cause HBOC. In the future, it would be desirable to perform this analysis on a group of normal patients that lack BC/OC family history.

5.4.2 Contribution to HBOC Literature

While it is important to develop methods for more accurate variant detection and interpretation, this information is exponentially more valuable when combined with other groups’ findings. A significant challenge associated with VUS analysis, particularly in the case of many of these recent HBOC gene candidates, is the under-reporting of variants and thus positive findings are currently over-represented in the literature⁷. As a member of the ENIGMA consortium, which is a designated international organization for the curation of HBOC mutations, our lab has committed to depositing the results of this project to the database. It is important to note that the ENIGMA database contains protections to ensure the genetic privacy of the study participants is maintained.

5.4.3 Patient Counseling

An advantage of this study is that a large proportion of the samples used came from patients who had provided informed consent to participate. These patients will be re-contacted and provided with the results of our analysis. At this point, for patients with

prioritized variants, it would be desirable to invite other family members to undergo genetic testing. Furthermore, an additional blood draw would be required of the patient, specifically for isolation of mRNA for the purpose of conducting functional analysis and validation of the IT prediction.

Depending on the gene harboring the pathogenic variant (especially if this gene is associated with a high risk of other types of cancer), the patient (and potentially their affected family members) should be provided with gene-specific counseling and preventative measures such as early screening by MRI⁸ or CT scan in families with increased risk of pancreatic cancer⁹, and risk-reducing surgery¹⁰.

In both cases of clearly pathogenic and prioritized uncertain variants, additional genotype information on family members would contribute to more accurate risk estimates through co-segregation analysis. Finally, re-contacting patients would also allow for the cataloguing of other information (eg. tumor pathology) that when combined with genotype and functional information, may lead to a more in-depth understanding of the genotype-phenotype relationship of HBOC.

5.5 References

1. Pan, Y., Duncombe, T. A., Kellenberger, C. A., Hammond, M. C. & Herr, A. E. High-throughput electrophoretic mobility shift assays for quantitative analysis of molecular binding reactions. *Anal. Chem.* **86**, 10357–10364 (2014).
2. Whitfield, T. W. *et al.* Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13**, R50 (2012).
3. Ge, J. *et al.* Micropatterned comet assay enables high throughput and sensitive DNA damage quantification. *Mutagenesis* **30**, 11–19 (2015).
4. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4**, 222 (2008).
5. Mirzaei, H. *et al.* Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 3645–3650 (2013).
6. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3**, 8 (2014).
7. Kraft, P. Curses--winner's and otherwise--in genetic epidemiology. *Epidemiol. Camb. Mass* **19**, 649–651; discussion 657–658 (2008).
8. Saslow, D. *et al.* American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. *CA. Cancer J. Clin.* **57**, 75–89 (2007).
9. Canto, M. I. *et al.* International Cancer of the Pancreas Screening (CAPS) Consortium summit on the management of patients with increased risk for familial pancreatic cancer. *Gut* **62**, 339–347 (2013).
10. Kauff, N. D. *et al.* Risk-reducing salpingo-oophorectomy in women with a BRCA1 or BRCA2 mutation. *N. Engl. J. Med.* **346**, 1609–1615 (2002).

Appendices


Appendix A: Copyright Permissions for Chapter 1 (Table 1.4)

Order Details		Billing Status: N/A
Human mutation		
Order detail ID:	67731907	Permission Status: ✔ Granted
ISSN:	1059-7794	Permission type: Republish or display content
Publication Type:	Journal	Type of use: Republish in a thesis/dissertation
Volume:		Order License Id: 3670981484240
Issue:		<input type="checkbox"/> Hide details
Start page:		Requestor type Author of requested content
Publisher:	JOHN/WILEY & SONS, INC.	Format Print, Electronic
Author/Editor:	Human Genome Variation Society	Portion chart/graph/table/figure
		Number of charts/graphs/tables /figures 1
		Title or numeric reference of the portion(s) Table 3
		Title of the article or chapter the portion is from Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results
		Editor of portion(s) N/A
		Author of portion(s) Plon et al
		Volume of serial or monograph 29
		Issue, if republishing an article from a serial 11
		Page range of portion 1282-1291
		Publication date of portion November 2008
		Rights for Main product
		Duration of use Life of current edition
		Creation of copies for the disabled no
		With minor editing privileges no
		For distribution to Worldwide
		In the following language(s) Original language of publication
		With incidental promotional use no
		Lifetime unit quantity of new product Up to 499
		Made available in the following markets education
		The requesting person/organization Natasha Caminsky
		Order reference number
		Author/Editor Natasha Caminsky
		The standard identifier MSc Thesis
		Title A unified framework for the prioritization of variants of uncertain significance in hereditary breast and ovarian cancer
		Publisher Western University
		Expected publication date Oct 2017
		Estimated size (pages) 200

Note: This item was invoiced separately through our [RightsLink service](#). [More Info](#) \$ 0.00

Appendix B: Copyright Permission for Chapter 1 (Tables 1.5, 1.6, 1.7)

Order detail ID: 67920295
Order License Id: 3683761313733
ISSN: 0028-0836
Publication Type: Journal
Volume:
Issue:
Start page:
Publisher: Nature Publishing Group

Permission Status:  **Granted**

Permission type: Republish or display content
Type of use: Republish in a thesis/dissertation

Hide details

Requestor type	Academic institution
Format	Print, Electronic
Portion	chart/graph/table/figure
Number of charts/graphs/tables/figures	4
Title or numeric reference of the portion(s)	Tables 3-5 and Figure 1
Title of the article or chapter the portion is from	N/A
Editor of portion(s)	N/A
Author of portion(s)	N/A
Volume of serial or monograph	17
Issue, if republishing an article from a serial	5
Page range of portion	412-415
Publication date of portion	May 2015
Rights for	Main product
Duration of use	Life of current and all future editions
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Made available in the following markets	education
The requesting person/organization	Natasha Caminsky
Order reference number	
Author/Editor	Natasha Caminsky
The standard identifier	MSc Thesis
Title	A unified framework for the prioritization of variants of uncertain significance in hereditary breast and ovarian cancer patients
Publisher	Western University
Expected publication date	Oct 2015
Estimated size (pages)	300

Note: This item will be invoiced or charged separately through CCC's RightsLink service. [More Info](#)

\$ 0.00

Appendix C: Copyright Permission for Chapter 2



Copyright: © 2015 Caminsky NG *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Appendix D: Supplementary Bibliography

This list contains all references that were reviewed for the purpose of this manuscript. Primary research articles, review articles, book chapters, theses, and abstracts that refer to information theory-based sequence analysis and were published up until September 2014 are included. We also reviewed the early literature describing information theory-based analysis. Certain items on this list were not referred to in the main document, because the authors did not perform information theory-based analysis, or because the article did not pertain to human disease and splicing.

1. Adachi, M. *et al.* Compound Heterozygous Mutations in the γ Subunit Gene of ENaC (1627delG and 1570-1G→A) in One Sporadic Japanese Patient with a Systemic Form of Pseudohypoaldosteronism Type 1. *J. Clin. Endocrinol. Metab.* **86**, 9–12 (2001).
2. Aggarwal, S., Jinda, W., Limwongse, C., Atchaneeyasakul, L. & Phadke, S. R. Run-on mutation in the PAX6 gene and chorioretinal degeneration in autosomal dominant aniridia. *Mol. Vis.* **17**, 1305–1309 (2011).
3. Aissat, A. *et al.* Combined computational-experimental analyses of CFTR exon strength uncover predictability of exon-skipping level. *Hum. Mutat.* **34**, 873–881 (2013).
4. Akiyama, M. *et al.* DNA-based prenatal diagnosis of harlequin ichthyosis and characterization of ABCA12 mutation consequences. *J. Invest. Dermatol.* **127**, 568–573 (2007).
5. Alcántara-Ortigoza, M. A., Belmont-Martínez, L., Vela-Amieva, M. & González-Del Angel, A. Analysis of the CTNS gene in nephropathic cystinosis Mexican patients: report of four novel mutations and identification of a false positive 57-kb deletion genotype with LDM-2/exon 4 multiplex PCR assay. *Genet. Test.* **12**, 409–414 (2008).
6. Allikmets, R. *et al.* Organization of the ABCR gene: analysis of promoter and splice junction sequences. *Gene* **215**, 111–122 (1998).
7. Anczuków, O. *et al.* Unclassified variants identified in BRCA1 exon 11: Consequences on splicing. *Genes. Chromosomes Cancer* **47**, 418–426 (2008).
8. Aoyama, Y. *et al.* Molecular features of 23 patients with glycogen storage disease type III in Turkey: a novel mutation p.R1147G associated with isolated glucosidase deficiency, along with 9 AGL mutations. *J. Hum. Genet.* **54**, 681–686 (2009).
9. Arita, K. *et al.* Unusual molecular findings in Kindler syndrome. *Br. J. Dermatol.*

- 157, 1252–1256 (2007).
10. Arnould, I. *et al.* Identifying and characterizing a five-gene cluster of ATP-binding cassette transporters mapping to human chromosome 17q24: a new subgroup within the ABCA subfamily. *GeneScreen* **1**, 157–164 (2001).
 11. Astuto, L. M. *et al.* Searching for evidence of DFNB2. *Am. J. Med. Genet.* **109**, 291–297 (2002).
 12. Atwood, C. S. METHODS FOR ASSESSING RISK OF ALZHEIMER'S DISEASE IN A PATIENT. (2013).
 13. Ayari Jeridi, H. *et al.* Genetic testing in Tunisian families with heritable retinoblastoma using a low cost approach permits accurate risk prediction in relatives and reveals incomplete penetrance in adults. *Exp. Eye Res* **124**, 48–55 (2014).
 14. Bacci, C. *et al.* Schwannomatosis associated with multiple meningiomas due to a familial SMARCB1 mutation. *Neurogenetics* **11**, 73–80 (2010).
 15. Baldelli, L. Analisi bioinformatica dei polimorfismi di alcuni geni del sistema serotoninergico. (2013).
 16. Baralle, D., Lucassen, A. & Buratti, E. Missed threads. The impact of pre-mRNA splicing defects on clinical practice. *EMBO Rep.* **10**, 810–816 (2009).
 17. Baralle, M. & Baralle, D. in *Altern. Pre-mRNA Splicing* (eds. Stamm, S., Smith, C. W. J. & Lührmann, R.) 129–138 (Wiley-VCH Verlag GmbH & Co. KGaA, 2012).
 18. Baralle, M. & Baralle, D. in *Neurofibromatosis Type 1* (eds. Upadhyaya, M. & Cooper, D. N.) 135–150 (Springer Berlin Heidelberg, 2012).
 19. Bateman, J. B. *et al.* A new betaA1-crystallin splice junction mutation in autosomal dominant cataract. *Invest. Ophthalmol. Vis. Sci.* **41**, 3278–3285 (2000).
 20. Baturina, O. A., Tupikin, A. E., Lukjanova, T. V., Sosnitskaya, S. V. & Morozov, I. V. PAH and QDPR Deficiency Associated Mutations in the Novosibirsk Region of the Russian Federation: Correlation of Mutation Type with Disease Manifestation and Severity. *jomb* **33**, 333–340 (2014).
 21. Beetz, C. *et al.* REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. *Brain J. Neurol.* **131**, 1078–1086 (2008).
 22. Ben Selma, Z. *et al.* A novel S115G mutation of CGI-58 in a Turkish patient with Dorfman-Chanarin syndrome. *J. Invest. Dermatol.* **127**, 2273–2276 (2007).
 23. Ben-Salem, S., Begum, M. A., Ali, B. R. & Al-Gazali, L. A Novel Aberrant Splice Site Mutation in RAB23 Leads to an Eight Nucleotide Deletion in the mRNA and Is Responsible for Carpenter Syndrome in a Consanguineous Emirati Family. *Mol.*

- Syndromol.* **3**, 255–261 (2013).
24. Benaglio, P. *et al.* Mutational screening of splicing factor genes in cases with autosomal dominant retinitis pigmentosa. *Mol. Vis.* **20**, 843–851 (2014).
 25. Bertola, F. *et al.* IDUA mutational profiling of a cohort of 102 European patients with mucopolysaccharidosis type I: identification and characterization of 35 novel α -L-iduronidase (IDUA) alleles. *Hum. Mutat.* **32**, E2189–2210 (2011).
 26. Bertolini, S. *et al.* Spectrum of mutations and phenotypic expression in patients with autosomal dominant hypercholesterolemia identified in Italy. *Atherosclerosis* **227**, 342–348 (2013).
 27. Bloethner, S., Mould, A., Stark, M. & Hayward, N. K. Identification of ARHGEF17, DENND2D, FGFR3, and RB1 mutations in melanoma by inhibition of nonsense-mediated mRNA decay. *Genes. Chromosomes Cancer* **47**, 1076–1085 (2008).
 28. Bocchi, L. *et al.* Multiple abnormally spliced ABCA1 mRNAs caused by a novel splice site mutation of ABCA1 gene in a patient with Tangier disease. *Clin. Chim. Acta Int. J. Clin. Chem.* **411**, 524–530 (2010).
 29. Bogaerts, V. *et al.* Genetic variability in the mitochondrial serine protease HTRA2 contributes to risk for Parkinson disease. *Hum. Mutat.* **29**, 832–840 (2008).
 30. Bonafé, L., Giunta, C., Gassner, M., Steinmann, B. & Superti-Furga, A. A cluster of autosomal recessive spondylocostal dysostosis caused by three newly identified DLL3 mutations segregating in a small village. *Clin. Genet.* **64**, 28–35 (2003).
 31. Bonnart, C. (2007). *Étude fonctionnelle de LEKTI et de sa nouvelle cible, l'élastase 2 pancréatique*. PhD. Thesis. Université Paul Sabatier: France.
 32. Bonnet-Dupeyron, M.-N. *et al.* PLP1 splicing abnormalities identified in Pelizaeus-Merzbacher disease and SPG2 fibroblasts are associated with different types of mutations. *Hum. Mutat.* **29**, 1028–1036 (2008).
 33. Borroni, B. *et al.* Progranulin genetic variations in frontotemporal lobar degeneration: evidence for low mutation frequency in an Italian clinical series. *Neurogenetics* **9**, 197–205 (2008).
 34. Botta, E. *et al.* Genotype-phenotype relationships in trichothiodystrophy patients with novel splicing mutations in the XPD gene. *Hum. Mutat.* **30**, 438–445 (2009).
 35. Brockmöller, J. & Tzvetkov, M. V. Pharmacogenetics: data, concepts and tools to improve drug discovery and drug treatment. *Eur. J. Clin. Pharmacol.* **64**, 133–157 (2008).
 36. Bröer, S. *et al.* Iminoglycinuria and hyperglycinuria are discrete human phenotypes resulting from complex mutations in proline and glycine transporters. *J. Clin.*

- Invest.* **118**, 3881–3892 (2008).
37. Buratti, E., Baralle, M. & Baralle, F. E. Defective splicing, disease and therapy: searching for master checkpoints in exon definition. *Nucleic Acids Res.* **34**, 3494–3510 (2006).
 38. Cabral, R. M. *et al.* Homozygous mutations in the 5' region of the JUP gene result in cutaneous disease but normal heart development in children. *J. Invest. Dermatol.* **130**, 1543–1550 (2010).
 39. Calandra, S., Tarugi, P. & Bertolini, S. Altered mRNA splicing in lipoprotein disorders. *Curr. Opin. Lipidol.* **22**, 93–99 (2011).
 40. Cambi, F. *et al.* Abstracts of the the American Society for Neurochemistry 37th Annual Meeting, Portland, Oregon, USA, 11-15 March 2006. *J. Neurochem.* **96** Suppl 1, 1–150 (2006).
 41. Caridi, G. *et al.* Analbuminemia Zonguldak: case report and mutational analysis. *Clin. Biochem.* **41**, 288–291 (2008).
 42. Cartault, F. *et al.* A new XPC gene splicing mutation has lead to the highest worldwide prevalence of xeroderma pigmentosum in black Mahori patients. *DNA Repair* **10**, 577–585 (2011).
 43. Castaman, G. *et al.* Deep intronic variations may cause mild hemophilia A. *J. Thromb. Haemost. JTH* **9**, 1541–1548 (2011).
 44. Castiglia, D. & Zambruno, G. Mutation mechanisms. *Dermatol. Clin.* **28**, 17–22 (2010).
 45. Catucci, I. *et al.* PALB2 sequencing in Italian familial breast cancer cases reveals a high-risk mutation recurrent in the province of Bergamo. *Genet. Med.* **16**, 688–694 (2014).
 46. Caudevilla, C. *et al.* Heterologous HIV-nef mRNA trans-splicing: a new principle how mammalian cells generate hybrid mRNA and protein molecules. *FEBS Lett.* **507**, 269–279 (2001).
 47. Caux-Moncoutier, V. *et al.* Impact of BRCA1 and BRCA2 variants on splicing: clues from an allelic imbalance study. *Eur. J. Hum. Genet. EJHG* **17**, 1471–1480 (2009).
 48. Cefalù, A. B. *et al.* Novel mutations of CETP gene in Italian subjects with hyperalphalipoproteinemia. *Atherosclerosis* **204**, 202–207 (2009).
 49. Cha E. (2008). *Computational analysis of expressed sequence tags for understanding gene regulation*. PhD. Thesis. University of Louisville: USA.
 50. Chen, L. *et al.* Genetic polymorphism analysis of CYP2C19 in Chinese Han

- populations from different geographic areas of mainland China. *Pharmacogenomics* **9**, 691–702 (2008).
51. Chen, L. J. *et al.* Evaluation of SPARC as a candidate gene of juvenile-onset primary open-angle glaucoma by mutation and copy number analyses. *Mol. Vis.* **16**, 2016–2025 (2010).
 52. Chen, Z. & Schneider, T. D. Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases. *Nucleic Acids Res.* **33**, 6172–6187 (2005).
 53. Chen, Z. *et al.* Discovery of Fur binding site clusters in *Escherichia coli* by information theory models. *Nucleic Acids Res.* **35**, 6762–6777 (2007).
 54. Cho, S. Y. *et al.* Genetic investigation of patients with undetectable peaks of growth hormone after two provocation tests. *Clin. Endocrinol. (Oxf.)* **78**, 317–320 (2013).
 55. Clark, G. R. *et al.* Development of a diagnostic genetic test for simplex and autosomal recessive retinitis pigmentosa. *Ophthalmology* **117**, 2169–2177.e3 (2010).
 56. Cleaver, J. E., Collins, C., Ellis, J. & Volik, S. Genome sequence and splice site analysis of low-fidelity DNA polymerases H and I involved in replication of damaged DNA. *Genomics* **82**, 561–570 (2003).
 57. Cohen, B. *et al.* A novel splice site mutation of CDHR1 in a consanguineous Israeli Christian Arab family segregating autosomal recessive cone-rod dystrophy. *Mol. Vis.* **18**, 2915–2921 (2012).
 58. Colobran, R. *et al.* Identification and characterization of a novel splice site mutation in the SERPING1 gene in a family with hereditary angioedema. *Clin. Immunol.* **150**, 143–148 (2014).
 59. Colombo, M. *et al.* Comparative in vitro and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PLoS ONE* **8**, e57173 (2013).
 60. Comin, M. & Antonello, M. Fast Entropic Profiler: An Information Theoretic Approach for the Discovery of Patterns in Genomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11**, 500–509 (2014).
 61. Comin, M. & Antonello, M. in *Pattern Recognit. Bioinforma.* (eds. Ngom, A., Formenti, E., Hao, J.-K., Zhao, X.-M. & Laarhoven, T. van) 277–288 (Springer Berlin Heidelberg, 2013).
 62. Concolino, P. *et al.* Functional analysis of two rare CYP21A2 mutations detected in Italian patients with a mildest form of congenital adrenal hyperplasia. *Clin. Endocrinol. (Oxf.)* **71**, 470–476 (2009).

63. Covaciu, C. *et al.* A founder synonymous COL7A1 mutation in three Danish families with dominant dystrophic epidermolysis bullosa pruriginosa identifies exonic regulatory sequences required for exon 87 splicing. *Br. J. Dermatol.* **165**, 678–682 (2011).
64. Cox, D.-G. *et al.* Haplotype of prostaglandin synthase 2/cyclooxygenase 2 is involved in the susceptibility to inflammatory bowel disease. *World J. Gastroenterol. WJG* **11**, 6003–6008 (2005).
65. Cronin, C. A., Gluba, W. & Scrable, H. The lac operator-repressor system is functional in the mouse. *Genes Dev.* **15**, 1506–1517 (2001).
66. Cruchaga, C. *et al.* Cortical atrophy and language network reorganization associated with a novel progranulin mutation. *Cereb. Cortex N. Y. N 1991* **19**, 1751–1760 (2009).
67. Dash, D. P. *et al.* Mutational screening of VSX1 in keratoconus patients from the European population. *Eye Lond. Engl.* **24**, 1085–1092 (2010).
68. Day IN, Kralovicova J, Gaunt TR, *et al.*: **IDDM2 locus: 5' noncoding intron I splicing and translational efficiency effects of INS -23HphI - more than a tag for the INS promoter VNTR.** HUGO's 11th Human Genome Meeting (HGM2006), Helsinki Finland. 2006. 2011
69. Deen, P. M. T., Dahl, N. & Caplan, M. J. The aquaporin-2 water channel in autosomal dominant primary nocturnal enuresis. *J. Urol.* **167**, 1447–1450 (2002).
70. Denecke, J., Kranz, C., Kemming, D., Koch, H.-G. & Marquardt, T. An activated 5' cryptic splice site in the human ALG3 gene generates a premature termination codon insensitive to nonsense-mediated mRNA decay in a new case of congenital disorder of glycosylation type Id (CDG-Id). *Hum. Mutat.* **23**, 477–486 (2004).
71. Denson, J. *et al.* Screening for inter-individual splicing differences in human GSTM4 and the discovery of a single nucleotide substitution related to the tandem skipping of two exons. *Gene* **379**, 148–155 (2006).
72. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).
73. Desmet, F.-O., Hamroun, D., Collod-Bérout, G. & Beroud, C. Bioinformatics identification of splice site signals and prediction of mutation effects. *Res. Adv. Nucleic Acid Res. KeralaGlobal Res. Netw.* 1–14 (2010).
74. Di Leo, E. *et al.* A point mutation in the lariat branch point of intron 6 of NPC1 as the cause of abnormal pre-mRNA splicing in Niemann-Pick type C disease. *Hum. Mutat.* **24**, 440 (2004).
75. Di Leo, E. *et al.* Abnormal apolipoprotein B pre-mRNA splicing in patients with

- familial hypobetalipoproteinaemia. *J. Med. Genet.* **44**, 219–224 (2007).
76. Di Leo, E. *et al.* Functional analysis of two novel splice site mutations of APOB gene in familial hypobetalipoproteinemia. *Mol. Genet. Metab.* **96**, 66–72 (2009).
77. Dinakarbandian, D., Raheja, V., Mehta, S., Schuetz, E. G. & Rogan, P. K. Tandem machine learning for the identification of genes regulated by transcription factors. *BMC Bioinformatics* **6**, 204 (2005).
78. Douglas, D. A. *et al.* Novel mutations of epidermal growth factor receptor in localized prostate cancer. *Front. Biosci. J. Virtual Libr.* **11**, 2518–2525 (2006).
79. Drera, B. *et al.* Branch point and donor splice-site COL7A1 mutations in mild recessive dystrophic epidermolysis bullosa. *Br. J. Dermatol.* **161**, 464–467 (2009).
80. Drögemüller, C., Philipp, U., Haase, B., Günzel-Apel, A.-R. & Leeb, T. A noncoding melanophilin gene (MLPH) SNP at the splice donor of exon 1 represents a candidate causal mutation for coat color dilution in dogs. *J. Hered.* **98**, 468–473 (2007).
81. Dua-Awereh, M. B., Shimomura, Y., Kraemer, L., Wajid, M. & Christiano, A. M. Mutations in the desmoglein 1 gene in five Pakistani families with striate palmoplantar keratoderma. *J. Dermatol. Sci.* **53**, 192–197 (2009).
82. Dunn, D. M. *et al.* Common variant of human NEDD4L activates a cryptic splice site to form a frameshifted transcript. *J. Hum. Genet.* **47**, 0665–0676 (2002).
83. Dutrannoy, V. *et al.* Clinical variability and novel mutations in the NHEJ1 gene in patients with a Nijmegen breakage syndrome-like phenotype. *Hum. Mutat.* **31**, 1059–1068 (2010).
84. Eckl, K.-M. *et al.* Molecular analysis of 250 patients with autosomal recessive congenital ichthyosis: evidence for mutation hotspots in ALOXE3 and allelic heterogeneity in ALOX12B. *J. Invest. Dermatol.* **129**, 1421–1428 (2009).
85. Eichers, E. R. *et al.* Newfoundland rod-cone dystrophy, an early-onset retinal dystrophy, is caused by splice-junction mutations in RLBP1. *Am. J. Hum. Genet.* **70**, 955–964 (2002).
86. Ellard, S., Patrinos, G. P. & Oetting, W. S. Clinical applications of next-generation sequencing: the 2013 human genome variation society scientific meeting. *Hum. Mutat.* **34**, 1583–1587 (2013).
87. Ellis, J. R., Jr, Heinrich, B., Mautner, V.-F. & Kluwe, L. Effects of splicing mutations on NF2-transcripts: transcript analysis and information theoretic predictions. *Genes. Chromosomes Cancer* **50**, 571–584 (2011).
88. ElSharawy, A. *et al.* Systematic evaluation of the effect of common SNPs on pre-

- mRNA splicing. *Hum. Mutat.* **30**, 625–632 (2009).
89. Emmert, S., Schneider, T. D., Khan, S. G. & Kraemer, K. H. The human XPG gene: gene architecture, alternative splicing and single nucleotide polymorphisms. *Nucleic Acids Res.* **29**, 1443–1452 (2001).
 90. Fahey, M. E. & Higgins, D. G. Gene expression, intron density, and splice site strength in *Drosophila* and *Caenorhabditis*. *J. Mol. Evol.* **65**, 349–357 (2007).
 91. Fang, S. *et al.* Novel GPR143 mutations and clinical characteristics in six Chinese families with X-linked ocular albinism. *Mol. Vis.* **14**, 1974–1982 (2008).
 92. Fasano, T. *et al.* Lysosomal lipase deficiency: molecular characterization of eleven patients with Wolman or cholesteryl ester storage disease. *Mol. Genet. Metab.* **105**, 450–456 (2012).
 93. Fasano, T. *et al.* Novel mutations of ABCA1 transporter in patients with Tangier disease and familial HDL deficiency. *Mol. Genet. Metab.* **107**, 534–541 (2012).
 94. Fasano, T., Bocchi, L., Pisciotta, L., Bertolini, S. & Calandra, S. Denaturing high-performance liquid chromatography in the detection of ABCA1 gene mutations in familial HDL deficiency. *J. Lipid Res.* **46**, 817–822 (2005).
 95. Fattal-Valevski, A. *et al.* Variable expression of a novel PLP1 mutation in members of a family with Pelizaeus-Merzbacher disease. *J. Child Neurol.* **24**, 618–624 (2009).
 96. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).
 97. Faz, D. B. *et al.* Bases genéticas de la conducta. (Editorial UOC, 2008).
 98. Fei, J. Splice Site Mutation-Induced Alteration of Selective Regional Activity Correlates with the Role of a Gene in Cardiomyopathy. *J Clin Exp Cardiol* **01**, (2013).
 99. Ferlini, A., Neri, M. & Gualandi, F. The medical genetics of dystrophinopathies: molecular genetic diagnosis and its impact on clinical practice. *Neuromuscul. Disord. NMD* **23**, 4–14 (2013).
 100. Fernandes, F., Freitas, A. T., Almeida, J. S. & Vinga, S. Entropic Profiler - detection of conservation in genomes using information theory. *BMC Res. Notes* **2**, 72 (2009).
 101. Fong, K. *et al.* Infantile systemic hyalinosis associated with a putative splice-site mutation in the ANTXR2 gene. *Clin. Exp. Dermatol.* **37**, 635–638 (2012).
 102. Foretová, L., Navrátilová, M. & Machácková, E. [Limitations of genetic testing in oncology]. *Klin. Onkol. Cas. České Slov. Onkol. Společnosti* **22 Suppl**, S65–68

- (2009).
103. Fornage, M. *et al.* The soluble epoxide hydrolase gene harbors sequence variation associated with susceptibility to and protection from incident ischemic stroke. *Hum. Mol. Genet.* **14**, 2829–2837 (2005).
 104. Funghini, S. *et al.* Carbamoyl phosphate synthetase 1 deficiency in Italy: clinical and genetic findings in a heterogeneous cohort. *Gene* **493**, 228–234 (2012).
 105. Gadiraju, S., Vyhlidal, C. A., Leeder, J. S. & Rogan, P. K. Genome-wide prediction, display and refinement of binding sites with information theory-based models. *BMC Bioinformatics* **4**, 38 (2003).
 106. Gaedigk, A. & Leeder, J. S. Reply*. *Clin. Pharmacol. Ther.* **80**, 558–560 (2006).
 107. Gaedigk, A. *et al.* Identification and characterization of novel sequence variations in the cytochrome P4502D6 (CYP2D6) gene in African Americans. *Pharmacogenomics J.* **5**, 173–182 (2005).
 108. Gaedigk, A. *et al.* Variability of CYP2J2 expression in human fetal tissues. *J. Pharmacol. Exp. Ther.* **319**, 523–532 (2006).
 109. Gaedigk, A., Gaedigk, R. & Leeder, J. S. CYP2D7 splice variants in human liver and brain: does CYP2D7 encode functional protein? *Biochem. Biophys. Res. Commun.* **336**, 1241–1250 (2005).
 110. Gaedigk, A., Ndjountché, L., Leeder, J. S. & Bradford, L. D. Limited association of the 2988g > a single nucleotide polymorphism with CYP2D641 in black subjects. *Clin. Pharmacol. Ther.* **77**, 228–230; author reply 230–231 (2005).
 111. Gallagher, C. (2009). *Development of an automated identification system for nanocrystal encoded microspheres in flow cytometry*. PhD. Thesis. Cranfield University: UK.
 112. Gao, S., Zhang, N., Zhang, L., Duan, G.-Y. & Zhang, T. [The human variome project and its progress]. *Yi Chuan Hered. Zhongguo Yi Chuan Xue Hui Bian Ji* **32**, 1105–1113 (2010).
 113. Garcia-Blanco, M. A. Alternative splicing: therapeutic target and tool. *Prog. Mol. Subcell. Biol.* **44**, 47–64 (2006).
 114. Garcia-Gonzalez, M. A. *et al.* Evaluating the clinical utility of a molecular genetic test for polycystic kidney disease. *Mol. Genet. Metab.* **92**, 160–167 (2007).
 115. Gaweda-Walerych, K. *et al.* Mitochondrial transcription factor A variants and the risk of Parkinson's disease. *Neurosci. Lett.* **469**, 24–29 (2010).
 116. Gemignani, F. *et al.* A TP53 polymorphism is associated with increased risk of

- colorectal cancer and with reduced levels of TP53 mRNA. *Oncogene* **23**, 1954–1956 (2003).
117. Gerykova-Bujalkova, M., Krivulcik, T. & Bartosova, Z. Novel approaches in evaluation of pathogenicity of single-base exonic germline changes involving the mismatch repair genes MLH1 and MSH2 in diagnostics of Lynch syndrome. *Neoplasma* **55**, 463–471 (2008).
118. Gibbons, W. J., Jr, Yan, Q., Li, R., Li, X. & Guan, M.-X. Genomic organization, expression, and subcellular localization of mouse mitochondrial seryl-tRNA synthetase. *Biochem. Biophys. Res. Commun.* **317**, 774–778 (2004).
119. Godefroid, N. *et al.* A novel splicing mutation in SLC12A3 associated with Gitelman syndrome and idiopathic intracranial hypertension. *Am. J. Kidney Dis. Off. J. Natl. Kidney Found.* **48**, e73–79 (2006).
120. Goldin, E. *et al.* Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucopolidosis IV. *Hum. Mutat.* **24**, 460–465 (2004).
121. Gozukara, E. M. *et al.* A stop codon in xeroderma pigmentosum group C families in Turkey and Italy: molecular genetic evidence for a common ancestor. *J. Invest. Dermatol.* **117**, 197–204 (2001).
122. Gruber, F. X., Hjorth-Hansen, H., Mikkola, I., Stenke, L. & Johansen, T. A novel Bcr-Abl splice isoform is associated with the L248V mutation in CML patients with acquired resistance to imatinib. *Leukemia* **20**, 2057–2060 (2006).
123. Hageman, G. S. *et al.* A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7227–7232 (2005).
124. Hamada, T. *et al.* Molecular and clinical characterization in Japanese and Korean patients with Hailey-Hailey disease: six new mutations in the ATP2C1 gene. *J. Dermatol. Sci.* **51**, 31–36 (2008).
125. Hampson, G., Konrad, M. A. & Scoble, J. Familial hypomagnesaemia with hypercalciuria and nephrocalcinosis (FHHNC): compound heterozygous mutation in the claudin 16 (CLDN16) gene. *BMC Nephrol.* **9**, 12 (2008).
126. Hartmann, L., Theiss, S., Niederacher, D. & Schaal, H. Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases? *Front. Biosci. J. Virtual Libr.* **13**, 3252–3272 (2008).
127. Hefferon, T. W., Broackes-Carter, F. C., Harris, A. & Cutting, G. R. Atypical 5' splice sites cause CFTR exon 9 to be vulnerable to skipping. *Am. J. Hum. Genet.* **71**, 294–303 (2002).
128. Hellerud, C. *et al.* Clinical heterogeneity and molecular findings in five Polish patients with glycerol kinase deficiency: investigation of two splice site mutations

- with computerized splice junction analysis and Xp21 gene-specific mRNA analysis. *Mol. Genet. Metab.* **79**, 149–159 (2003).
129. Hellerud, C. *et al.* Glycerol metabolism and the determination of triglycerides--clinical, biochemical and molecular findings in six subjects. *Clin. Chem. Lab. Med. CCLM FESCC* **41**, 46–55 (2003).
 130. Henriksen, A. M., Tümer, Z., Tommerup, N., Tranebjaerg, L. & Larsen, L. A. Identification of a novel EYA1 splice-site mutation in a Danish branchio-oto-renal syndrome family. *Genet. Test.* **8**, 404–406 (2004).
 131. Hengen, P. N., Lyakhov, I. G., Stewart, L. E. & Schneider, T. D. Molecular flip-flops formed by overlapping Fis sites. *Nucleic Acids Res.* **31**, 6663–6673 (2003).
 132. Henneman, P., Schaap, F. G., Rensen, P. C. N., van Dijk, K. W. & Smelt, A. H. M. Estrogen induced hypertriglyceridemia in an apolipoprotein AV deficient patient. *J. Intern. Med.* **263**, 107–108 (2008).
 133. Hertecant, J. L. *et al.* Clinical and molecular analysis of isovaleric acidemia patients in the United Arab Emirates reveals remarkable phenotypes and four novel mutations in the IVD gene. *Eur. J. Med. Genet.* **55**, 671–676 (2012).
 134. Hickford, J. G. H., Zhou, H., Slow, S. & Fang, Q. Diversity of the ovine DQA2 gene. *J. Anim. Sci.* **82**, 1553–1563 (2004).
 135. Hiller, M. *et al.* Phylogenetically widespread alternative splicing at unusual GYNGYN donors. *Genome Biol.* **7**, R65 (2006).
 136. Hines, R. N., Koukouritaki, S. B., Poch, M. T. & Stephens, M. C. Regulatory polymorphisms and their contribution to interindividual differences in the expression of enzymes influencing drug and toxicant disposition. *Drug Metab. Rev.* **40**, 263–301 (2008).
 137. Hobson, G. M. *et al.* Splice-site contribution in alternative splicing of PLP1 and DM20: molecular studies in oligodendrocytes. *Hum. Mutat.* **27**, 69–77 (2006).
 138. Houdayer, C. *et al.* Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum. Mutat.* **29**, 975–982 (2008).
 139. Houdayer, C. *et al.* Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum. Mutat.* **33**, 1228–1238 (2012).
 140. Houdayer, C. In silico prediction of splice-affecting nucleotide variants. *Methods Mol. Biol. Clifton NJ* **760**, 269–281 (2011).
 141. Hube, F. *et al.* Alternative splicing of the first intron of the steroid receptor RNA activator (SRA) participates in the generation of coding and noncoding RNA

- isoforms in breast cancer cell lines. *DNA Cell Biol.* **25**, 418–428 (2006).
142. Inui, H. *et al.* Xeroderma pigmentosum-variant patients from America, Europe, and Asia. *J. Invest. Dermatol.* **128**, 2055–2068 (2008).
 143. Ishitsuka, Y., Furuta, J., Miyashita, T. & Otsuka, F. Splicing aberration in naevoid basal cell carcinoma syndrome. *Acta Derm. Venereol.* **92**, 619–620 (2012).
 144. Jeon, G. W., Kwon, M.-J., Lee, S. J., Sin, J. B. & Ki, C.-S. Clinical and genetic analysis of a Korean patient with X-linked chondrodysplasia punctata: identification of a novel splicing mutation in the ARSE gene. *Ann. Clin. Lab. Sci.* **43**, 70–75 (2013).
 145. Jian, X., Boerwinkle, E. & Liu, X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet. Med.* **16**, 497–503 (2014).
 146. Jimenez, N. L. *et al.* Targeted ‘next-generation’ sequencing in anophthalmia and microphthalmia patients confirms SOX2, OTX2 and FOXE3 mutations. *BMC Med. Genet.* **12**, 172 (2011).
 147. Johnson, D. S. (2010). *Study of a possible genetic cause of CHARGE association*. MD Thesis. University of Glasgow: UK.
 148. Johnson, A. D. Single-nucleotide polymorphism bioinformatics: a comprehensive review of resources. *Circ. Cardiovasc. Genet.* **2**, 530–536 (2009).
 149. Kahn, A. B. *et al.* Ontogenomic study of the relationship between number of gene splice variants and GO categorization. *Bioinforma. Oxf. Engl.* **26**, 1945–1949 (2010).
 150. Kang, D.-H. *et al.* Identification of a novel splicing mutation in the ARSA gene in a patient with late-infantile form of metachromatic leukodystrophy. *Korean J. Lab. Med.* **30**, 516–520 (2010).
 151. Kannabiran, C. *et al.* Autosomal dominant zonular cataract with sutural opacities is associated with a splice mutation in the betaA3/A1-crystallin gene. *Mol. Vis.* **4**, 21 (1998).
 152. Kaput, J. *et al.* Planning the human variome project: the Spain report. *Hum. Mutat.* **30**, 496–510 (2009).
 153. Karaca, M. *et al.* High prevalence of cerebral venous sinus thrombosis (CVST) as presentation of cystathionine beta-synthase deficiency in childhood: molecular and clinical findings of Turkish probands. *Gene* **534**, 197–203 (2014).
 154. Keren, B. *et al.* CNS malformations in Knobloch syndrome with splice mutation in COL18A1 gene. *Am. J. Med. Genet. A.* **143A**, 1514–1518 (2007).
 155. Kern, J. S. (2005). *The molecular basis of dystrophic epidermolysis bullosa* :

mutation detection and study of clinical, biochemical and molecular findings in 29 patients. MD Thesis. University of Freiburg: Germany.

156. Khan, S. G. *et al.* The human XPC DNA repair gene: arrangement, splice site information content and influence of a single nucleotide polymorphism in a splice acceptor site on alternative splicing and function. *Nucleic Acids Res.* **30**, 3624–3631 (2002).
157. Khan, S. G. *et al.* Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. *Hum. Mol. Genet.* **13**, 343–352 (2004).
158. Khan, S. G. *et al.* Xeroderma pigmentosum group C splice mutation associated with autism and hypoglycinemia. *J. Invest. Dermatol.* **111**, 791–796 (1998).
159. Kim, G., Ko, J. & Yoo, H. A novel intronic point mutation of CPS1 gene in a korean family with CPS1 deficiency.
160. Kodolitsch, Y. von, Pyeritz, R. E. & Rogan, P. K. Splice-Site Mutations in Atherosclerosis Candidate Genes Relating Individual Information to Phenotype. *Circulation* **100**, 693–699 (1999).
161. Kölsch, H. *et al.* Association of SORL1 gene variants with Alzheimer's disease. *Brain Res.* **1264**, 1–6 (2009).
162. Kölsch, H. *et al.* CYP46A1 variants influence Alzheimer's disease risk and brain cholesterol metabolism. *Eur. Psychiatry J. Assoc. Eur. Psychiatr.* **24**, 183–190 (2009).
163. Kölsch, H. *et al.* Influence of SORL1 gene variants: association with CSF amyloid-beta products in probable Alzheimer's disease. *Neurosci. Lett.* **440**, 68–71 (2008).
164. Kölsch, H. *et al.* RXRA gene variations influence Alzheimer's disease risk and cholesterol metabolism. *J. Cell. Mol. Med.* **13**, 589–598 (2009).
165. Koukouritaki, S. B., Poch, M. T., Cabacungan, E. T., McCarver, D. G. & Hines, R. N. Discovery of novel flavin-containing monooxygenase 3 (FMO3) single nucleotide polymorphisms and functional analysis of upstream haplotype variants. *Mol. Pharmacol.* **68**, 383–392 (2005).
166. Královicová, J., Lei, H. & Vorechovský, I. Phenotypic consequences of branch point substitutions. *Hum. Mutat.* **27**, 803–813 (2006).
167. Krawczak, M. *et al.* Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.* **28**, 150–158 (2007).

168. Kwon, M.-J. *et al.* Screening of the SOD1, FUS, TARDBP, ANG, and OPTN mutations in Korean patients with familial and sporadic ALS. *Neurobiol. Aging* **33**, 1017.e17–23 (2012).
169. Kwong, A. K.-Y., Fung, C.-W., Chan, S.-Y. & Wong, V. C.-N. Identification of SCN1A and PCDH19 mutations in Chinese children with Dravet syndrome. *PloS One* **7**, e41802 (2012).
170. Lacroix, M. *et al.* Clinical expression and new SPINK5 splicing defects in Netherton syndrome: unmasking a frequent founder synonymous mutation and unconventional intronic mutations. *J. Invest. Dermatol.* **132**, 575–582 (2012).
171. Lamba, V. *et al.* Hepatic CYP2B6 expression: gender and ethnic differences and relationship to CYP2B6 genotype and CAR (constitutive androstane receptor) expression. *J. Pharmacol. Exp. Ther.* **307**, 906–922 (2003).
172. Laššuthová, P. *et al.* Three New PLP1 Splicing Mutations Demonstrate Pathogenic and Phenotypic Diversity of Pelizaeus-Merzbacher Disease. *J. Child Neurol.* **29**, 924–931 (2013).
173. Le Guédard-Méreuze, S. *et al.* Sequence contexts that determine the pathogenicity of base substitutions at position +3 of donor splice-sites. *Hum. Mutat.* **30**, 1329–1339 (2009).
174. Lebel, K. Génétique moléculaire du glaucome. WDR36: un gène modificateur potential pour la sévérité du glaucome. (2008).
175. Leclerc, D. *et al.* SLC7A9 mutations in all three cystinuria subtypes. *Kidney Int.* **62**, 1550–1559 (2002).
176. Leclerc, D., Wu, Q., Ellis, J. R., Goodyer, P. & Rozen, R. Is the SLC7A10 gene on chromosome 19 a candidate locus for cystinuria? *Mol. Genet. Metab.* **73**, 333–339 (2001).
177. Lee, P. H. (2009). *Prioritizing SNPs for disease-gene association studies: Algorithms and systems*. PhD Thesis. Queen's University: Canada.
178. Lee, P. P. W. *et al.* Clinical and molecular characteristics of 35 Chinese children with Wiskott-Aldrich syndrome. *J. Clin. Immunol.* **29**, 490–500 (2009).
179. Lee, S. H., Kim, S.-E., Noh, E. B., Oh, S.-W. & Kim, S.-C. Novel deletion mutation (c.3717del5) in COL7A1 in a patient with recessive dystrophic epidermolysis bullosa. *J. Dermatol.* **40**, 59–61 (2013).
180. Lee, S.-T., Lee, J., Lee, M., Kim, J.-W. & Ki, C.-S. Clinical and genetic analysis of Korean patients with congenital insensitivity to pain with anhidrosis. *Muscle Nerve* **40**, 855–859 (2009).
181. Lee, Y.-W. *et al.* Different spectrum of mutations of isovaleryl-CoA

- dehydrogenase (IVD) gene in Korean patients with isovaleric acidemia. *Mol. Genet. Metab.* **92**, 71–77 (2007).
182. Lehtokari, V.-L. *et al.* Nemaline myopathy caused by mutations in the nebulin gene may present as a distal myopathy. *Neuromuscul. Disord. NMD* **21**, 556–562 (2011).
183. Leman, A. R., Pearce, D. A. & Rothberg, P. G. Gene symbol: CLN3. Disease: Juvenile neuronal ceroid lipofuscinosis (Batten disease). *Hum. Genet.* **116**, 544 (2005).
184. Leverenz, J. B. *et al.* A novel progranulin mutation associated with variable clinical presentation and tau, TDP43 and alpha-synuclein pathology. *Brain J. Neurol.* **130**, 1360–1374 (2007).
185. Li, A. *et al.* Bietti crystalline corneoretinal dystrophy is caused by mutations in the novel gene CYP4V2. *Am. J. Hum. Genet.* **74**, 817–826 (2004).
186. Li, L. *et al.* Confirmation and refinement of an autosomal dominant congenital motor nystagmus locus in chromosome 1q31.3-q32.1. *J. Hum. Genet.* **57**, 756–759 (2012).
187. Li, L. *et al.* Detection of variants in 15 genes in 87 unrelated Chinese patients with Leber congenital amaurosis. *PloS One* **6**, e19458 (2011).
188. Li, R., Li, X., Yan, Q., Qin Mo, J. & Guan, M.-X. Identification and characterization of mouse MTO1 gene related to mitochondrial tRNA modification. *Biochim. Biophys. Acta* **1629**, 53–59 (2003).
189. Li, X. & Guan, M.-X. A Human Mitochondrial GTP Binding Protein Related to tRNA Modification May Modulate Phenotypic Expression of the Deafness-Associated Mitochondrial 12S rRNA Mutation. *Mol. Cell. Biol.* **22**, 7701–7711 (2002).
190. Li, X. & Guan, M.-X. Identification and characterization of mouse GTPBP3 gene encoding a mitochondrial GTP-binding protein involved in tRNA modification. *Biochem. Biophys. Res. Commun.* **312**, 747–754 (2003).
191. Li, X., Li, R., Lin, X. & Guan, M.-X. Isolation and characterization of the putative nuclear modifier gene MTO1 involved in the pathogenesis of deafness-associated mitochondrial 12 S rRNA A1555G mutation. *J. Biol. Chem.* **277**, 27256–27264 (2002).
192. Li, X., Zhang, L. S. & Guan, M.-X. Cloning and characterization of mouse mTERF encoding a mitochondrial transcriptional termination factor. *Biochem. Biophys. Res. Commun.* **326**, 505–510 (2005).
193. Lietman, S. A. Preimplantation genetic diagnosis for hereditary endocrine disease. *Endocr. Pract. Off. J. Am. Coll. Endocrinol. Am. Assoc. Clin. Endocrinol.*

- 17 Suppl 3**, 28–32 (2011).
194. Lietman, S. A., Goldfarb, J., Desai, N. & Levine, M. A. Preimplantation genetic diagnosis for severe albright hereditary osteodystrophy. *J. Clin. Endocrinol. Metab.* **93**, 901–904 (2008).
 195. Lim, B. C. *et al.* Fukutin mutations in congenital muscular dystrophies with defective glycosylation of dystroglycan in Korea. *Neuromuscul. Disord. NMD* **20**, 524–530 (2010).
 196. Lim, B. C. *et al.* SCN1A mutational analysis in Korean patients with Dravet syndrome. *Seizure J. Br. Epilepsy Assoc.* **20**, 789–794 (2011).
 197. Lin, Z., Wang, G., Demello, D. E. & Floros, J. An alternatively spliced surfactant protein B mRNA in normal human lung: disease implication. *Biochem. J.* **343**, 145–149 (1999).
 198. Liu, J. *et al.* The association of LRP5 gene polymorphisms with ankylosing spondylitis in a Chinese Han population. *J. Rheumatol.* **38**, 2616–2618 (2011).
 199. Liu, Z., Venkatesh, S. S. & Maley, C. C. Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples. *BMC Genomics* **9**, 509 (2008).
 200. Locke, G., Haberman, D., Johnson, S. M. & Morozov, A. V. Global remodeling of nucleosome positions in *C. elegans*. *BMC Genomics* **14**, 284 (2013).
 201. Longpre, K. M. *et al.* Seasonal variation of urinary microRNA expression in male goats (*Capra hircus*) as assessed by next generation sequencing. *Gen. Comp. Endocrinol.* **199**, 1–15 (2014).
 202. López-Jiménez, E. *et al.* SDHC mutation in an elderly patient without familial antecedents. *Clin. Endocrinol. (Oxf.)* **69**, 906–910 (2008).
 203. Lou, H. *et al.* Promoter variants in the MSMB gene associated with prostate cancer regulate MSMB/NCOA4 fusion transcripts. *Hum. Genet.* **131**, 1453–1466 (2012).
 204. Luquin, N., Yu, B., Trent, R. J., Morahan, J. M. & Pamphlett, R. An analysis of the entire SOD1 gene in sporadic ALS. *Neuromuscul. Disord. NMD* **18**, 545–552 (2008).
 205. Luquin, N., Yu, B., Saunderson, R. B., Trent, R. J. & Pamphlett, R. Genetic variants in the promoter of TARDBP in sporadic amyotrophic lateral sclerosis. *Neuromuscul. Disord.* **19**, 696–700 (2009).
 206. Mackay, D. S. *et al.* Novel mutations in MERTK associated with childhood onset rod-cone dystrophy. *Mol. Vis.* **16**, 369–377 (2010).

207. Maddalena, A. *et al.* Technical standards and guidelines: molecular genetic testing for ultra-rare disorders. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **7**, 571–583 (2005).
208. Magnolo, L. *et al.* Novel mutations in SAR1B and MTTP genes in Tunisian children with chylomicron retention disease and abetalipoproteinemia. *Gene* **512**, 28–34 (2013).
209. Malueka, R. G. *et al.* Categorization of 77 dystrophin exons into 5 groups by a decision tree using indexes of splicing regulatory factors as decision markers. *BMC Genet.* **13**, 23 (2012).
210. Mao, M. S. (2012). *Clinical Pharmacogenetics of Olanzapine: with Focus on FMO Gene Polymorphisms*. PhD Thesis. Uppsala University: Sweden..
211. Mao, M., Skogh, E., Scordo, M. G. & Dahl, M.-L. Interindividual variation in olanzapine concentration influenced by UGT1A4 L48V polymorphism in serum and upstream FMO polymorphisms in cerebrospinal fluid. *J. Clin. Psychopharmacol.* **32**, 287–289 (2012).
212. Marchal, A. *et al.* Un cas particulier d'épidermolyse bulleuse dystrophique. *Ann. Dermatol. Vénérologie* **138**, A168–A169 (2011).
213. Marco, E. J. *et al.* American Neurological Association 131st Annual Meeting October 8-11, 2006 Chicago, Illinois. *Ann. Neurol.* **60**, 625–645 (2006).
214. Marco, E. J. *et al.* ARHGEF9 disruption in a female patient is associated with X linked mental retardation and sensory hyperarousal. *BMJ Case Rep.* Epub Jul 2 (2009).
215. Marco, E. J. *et al.* ARHGEF9 disruption in a female patient is associated with X linked mental retardation and sensory hyperarousal. *J. Med. Genet.* **45**, 100–105 (2008).
216. Marr, N. *et al.* Cell-biologic and functional analyses of five new Aquaporin-2 missense mutations that cause recessive nephrogenic diabetes insipidus. *J. Am. Soc. Nephrol. JASN* **13**, 2267–2277 (2002).
217. Marras, E. *et al.* Discrepancies between in silico and in vitro data in the functional analysis of a breast cancer-associated polymorphism in the XRCC6/Ku70 gene. *Mol Med Rep* **1**, 805–812 (2008).
218. Martoni, E. *et al.* Identification and characterization of novel collagen VI non-canonical splicing mutations causing Ullrich congenital muscular dystrophy. *Hum. Mutat.* **30**, E662–672 (2009).
219. Maruszak, A. *et al.* PIN1 gene variants in Alzheimer's disease. *BMC Med. Genet.* **10**, 115 (2009).

220. May, E. The Emergence of biological coding theory as a mathematical framework for modeling, monitoring, and modulating biomolecular systems. in *43rd Annu. Conf. Inf. Sci. Syst. 2009 CISS 2009* 865–869 (2009).
221. McGrory, J. & Cole, W. G. Alternative splicing of exon 37 of FBN1 deletes part of an ‘eight-cysteine’ domain resulting in the Marfan syndrome. *Clin. Genet.* **55**, 118–121 (1999).
222. Megremis, S. *et al.* Nucleotide variations in the NPHS2 gene in Greek children with steroid-resistant nephrotic syndrome. *Genet. Test. Mol. Biomark.* **13**, 249–256 (2009).
223. Milone, M. *et al.* Myasthenic syndrome due to defects in rapsyn Clinical and molecular findings in 39 patients. *Neurology* **73**, 228–235 (2009).
224. Mintchev, N., Zamba-Papanicolaou, E., Kleopa, K. A. & Christodoulou, K. A novel ALS2 splice-site mutation in a Cypriot juvenile-onset primary lateral sclerosis family. *Neurology* **72**, 28–32 (2009).
225. Moriwaki, K. *et al.* Deficiency of GMDS leads to escape from NK cell-mediated tumor surveillance through modulation of TRAIL signaling. *Gastroenterology* **137**, 188–198, 198.e1–2 (2009).
226. Mucaki, E. J., Ainsworth, P. & Rogan, P. K. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Hum. Mutat.* **32**, 735–742 (2011).
227. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. *Hum. Mutat.* **34**, 557–565 (2013).
228. Mukhopadhyay, A. *et al.* Erosive vitreoretinopathy and wagner disease are caused by intronic mutations in CSPG2/Versican that result in an imbalance of splice variants. *Invest. Ophthalmol. Vis. Sci.* **47**, 3565–3572 (2006).
229. Mullins, R. F. *et al.* Autosomal recessive retinitis pigmentosa due to ABCA4 mutations: clinical, pathologic, and molecular characterization. *Invest. Ophthalmol. Vis. Sci.* **53**, 1883–1894 (2012).
230. Murphy, L. C. & Leygue, E. The role of estrogen receptor- β in breast cancer. *Semin. Reprod. Med.* **30**, 5–13 (2012).
231. Naiya, T. *et al.* Occurrence of GCH1 gene mutations in a group of Indian dystonia patients. *J. Neural Transm. Vienna Austria 1996* **119**, 1343–1350 (2012).
232. Najah, M. *et al.* Identification of patients with abetalipoproteinemia and homozygous familial hypobetalipoproteinemia in Tunisia. *Clin. Chim. Acta Int. J. Clin. Chem.* **401**, 51–56 (2009).

233. Nakano, T. & Suda, T. *NSF Workshop on Molecular Communication/Biological Communications Technology*. 49 (University of Virginia, Irvine. National Science Foundation, 2008).
234. Nalla, V. K. & Rogan, P. K. Automated splicing mutation analysis by information theory. *Hum. Mutat.* **25**, 334–342 (2005).
235. Naruse, H. *et al.* Determination of splice-site mutations in Lynch syndrome (hereditary non-polyposis colorectal cancer) patients using functional splicing assay. *Fam. Cancer* **8**, 509–517 (2009).
236. Nasim, M. T. *et al.* Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. *Hum. Mutat.* **32**, 1385–1389 (2011).
237. Nyiraneza, C. *et al.* Distinctive patterns of p53 protein expression and microsatellite instability in human colorectal cancer. *Hum. Pathol.* **42**, 1897–1910 (2011).
238. O’Neill, J. P., Rogan, P. K., Cariello, N. & Nicklas, J. A. Mutations that alter RNA splicing of the human HPRT gene: a review of the spectrum. *Mutat. Res. Mutat. Res.* **411**, 179–214 (1998).
239. Oetting, W. S. & Tabone, T. The 2004 Human Genome Variation Society scientific meeting. *Hum. Mutat.* **26**, 160–163 (2005).
240. Oh, K.-S. *et al.* Phenotypic heterogeneity in the XPB DNA helicase gene (ERCC3): xeroderma pigmentosum without and with Cockayne syndrome. *Hum. Mutat.* **27**, 1092–1103 (2006).
241. Oh, S.-W., Lee, J. S., Kim, M. Y. & Kim, S.-C. COL7A1 mutational analysis in Korean patients with dystrophic epidermolysis bullosa. *Br. J. Dermatol.* **157**, 1260–1264 (2007).
242. Oh, S.-W., Lee, J. S., Kim, M. Y. & Kim, S.-C. Novel keratin 5 mutations in epidermolysis bullosa simplex: cases with unusual genotype-phenotype correlation. *J. Dermatol. Sci.* **48**, 229–232 (2007).
243. Ohe, K. (2013). *Intronic and Exonic Nucleotide Variations that Affect RNA Splicing in Humans*. <
<https://www.iconceptpress.com/download/paper/12051922401407.pdf>>
244. Okubo, M. *et al.* A novel APOA5 splicing mutation IVS2+1g>a in a Japanese chylomicronemia patient. *Atherosclerosis* **207**, 24–25 (2009).
245. Olsen, R. K. J. *et al.* The ETFDH c.158A>G variation disrupts the balanced interplay of ESE- and ESS-binding proteins thereby causing missplicing and multiple Acyl-CoA dehydrogenation deficiency. *Hum. Mutat.* **35**, 86–95 (2014).
246. Ozaltin, F. *et al.* Disruption of PTPRO causes childhood-onset nephrotic

- syndrome. *Am. J. Hum. Genet.* **89**, 139–147 (2011).
247. Palomino Doza, J. *et al.* Low-frequency intermediate penetrance variants in the ROCK1 gene predispose to Tetralogy of Fallot. *BMC Genet.* **14**, 57 (2013).
248. Palomino-Doza, J. *et al.* Ambulatory blood pressure is associated with polymorphic variation in P2X receptor genes. *Hypertension* **52**, 980–985 (2008).
249. Papi, L. *et al.* A PALB2 germline mutation associated with hereditary breast cancer in Italy. *Fam. Cancer* **9**, 181–185 (2010).
250. Papp, J., Kovacs, M. E. & Olah, E. Germline MLH1 and MSH2 mutational spectrum including frequent large genomic aberrations in Hungarian hereditary non-polyposis colorectal cancer families: implications for genetic testing. *World J. Gastroenterol. WJG* **13**, 2727–2732 (2007).
251. Pasmooij, A.M. (2006). *Genotyping of unusual phenotypes of epidermolysis bullosa*. PhD Thesis. University of Groningen: Netherlands.
252. Pasmooij, A. M. G., Pas, H. H., Bolling, M. C. & Jonkman, M. F. Revertant mosaicism in junctional epidermolysis bullosa due to multiple correcting second-site mutations in LAMB3. *J. Clin. Invest.* **117**, 1240–1248 (2007).
253. Pasvolsky, R. *et al.* A LAD-III syndrome is associated with defective expression of the Rap-1 activator CalDAG-GEFI in lymphocytes, neutrophils, and platelets. *J. Exp. Med.* **204**, 1571–1582 (2007).
254. Pelucchi, S. *et al.* Expression of hepcidin and other iron-related genes in type 3 hemochromatosis due to a novel mutation in transferrin receptor-2. *Haematologica* **94**, 276–279 (2009).
255. Pérez, B. *et al.* Propionic acidemia: identification of twenty-four novel mutations in Europe and North America. *Mol. Genet. Metab.* **78**, 59–67 (2003).
256. Pernet, C. *et al.* Genitoperineal papular acantholytic dyskeratosis is allelic to Hailey-Hailey disease. *Br. J. Dermatol.* **167**, 210–212 (2012).
257. Philips, A. V. & Cooper, T. A. RNA processing and human disease. *Cell. Mol. Life Sci. CMLS* **57**, 235–249 (2000).
258. Pink, A. E. *et al.* Mutations in the γ -secretase genes NCSTN, PSENEN, and PSEN1 underlie rare forms of hidradenitis suppurativa (acne inversa). *J. Invest. Dermatol.* **132**, 2459–2461 (2012).
259. Piva, F., Giulietti, M., Nardi, B., Bellantuono, C. & Principato, G. An improved in silico selection of phenotype affecting polymorphisms in SLC6A4, HTR1A and HTR2A genes. *Hum. Psychopharmacol.* **25**, 153–161 (2010).
260. Priore Oliva, C. *et al.* A novel sequence variant in APOA5 gene found in

- patients with severe hypertriglyceridemia. *Atherosclerosis* **188**, 215–217 (2006).
261. Qadah, T., Finlayson, J. & Ghassemifar, R. In vitro characterization of the α -thalassemia point mutation HBA2:c.95+1G>A [IVS-I-1(G>A) (α 2)]. *Hemoglobin* **36**, 38–46 (2012).
262. Qin, S. *et al.* Systematic polymorphism analysis of the CYP2D6 gene in four different geographical Han populations in mainland China. *Genomics* **92**, 152–158 (2008).
263. Rady, P. L., Penzien, J., Vargas, T., Tyring, S. K. & Matalon, R. Novel splice site mutation of aspartoacylase gene in a Turkish patient with Canavan disease. *Eur. J. Paediatr. Neurol.* **4**, 27–30 (2000).
264. Rajkumar, S. (2012). *Genetic and biochemical modulation of Superoxide Dismutase SOD in human senile cataracts*. PhD Thesis. Manipal University: India.
265. Rajkumar, S. *et al.* Exploration of Molecular Factors Impairing Superoxide Dismutase Isoforms Activity in Human Senile Cataractous Lenses. *Invest. Ophthalmol. Vis. Sci.* **54**, 6224–6233 (2013).
266. Remaley, A. T. *et al.* Human ATP-binding cassette transporter 1 (ABC1): Genomic organization and identification of the genetic defect in the original Tangier disease kindred. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 12685–12690 (1999).
267. Rhyne, J., Mantaring, M. M., Gardner, D. F. & Miller, M. Multiple splice defects in ABCA1 cause low HDL-C in a family with hypoalphalipoproteinemia and premature coronary disease. *BMC Med. Genet.* **10**, 1 (2009).
268. Riveira-Munoz, E. *et al.* Evaluating PVALB as a candidate gene for SLC12A3-negative cases of Gitelman's syndrome. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **23**, 3120–3125 (2008).
269. Riveira-Munoz, E. *et al.* Transcriptional and functional analyses of SLC12A3 mutations: new clues for the pathogenesis of Gitelman syndrome. *J. Am. Soc. Nephrol. JASN* **18**, 1271–1283 (2007).
270. Roca, X., Krainer, A. R. & Eperon, I. C. Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes Dev.* **27**, 129–144 (2013).
271. Rogan, P. & Mucaki, E. Population Fitness and Genetic Load of Single Nucleotide Polymorphisms Affecting mRNA splicing. *ArXiv11070716 Q-Bio* (2011).
272. Rogan, P. K. & Schneider, T. D. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.* **6**, 74–76 (1995).
273. Rogan, P. K., Faux, B. M. & Schneider, T. D. Information analysis of human

- splice site mutations. *Hum. Mutat.* **12**, 153–171 (1998).
274. Rogan, P. K., Svojanovsky, S. & Leeder, J. S. Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics* **13**, 207–218 (2003).
275. Rossi, P. I. A. *et al.* The metabotropic glutamate receptor 1, GRM1: evaluation as a candidate gene for inherited forms of cerebellar ataxia. *J. Neurol.* **257**, 598–602 (2010).
276. Roux-Buisson, N. *et al.* Functional analysis reveals splicing mutations of the CASQ2 gene in patients with CPVT: implication for genetic counselling and clinical management. *Hum. Mutat.* **32**, 995–999 (2011).
277. Russcher, H. (Henk). (2006). *Glucocorticoid Receptor Variants Modulate the Sensitivity to Cortisol*. PhD Thesis. Erasmus University Rotterdam: Netherlands.
278. Russcher, H. *et al.* Strategies for the characterization of disorders in cortisol sensitivity. *J. Clin. Endocrinol. Metab.* **91**, 694–701 (2006).
279. Sabet, A. *et al.* Skin biopsies demonstrate MPZ splicing abnormalities in Charcot-Marie-Tooth neuropathy 1B. *Neurology* **67**, 1141–1146 (2006).
280. Saeed, S. *et al.* Novel LEPR mutations in obese Pakistani children identified by PCR-based enrichment and next generation sequencing. *Obesity (Silver Spring)* **22**, 1112–1117 (2014).
281. Sahashi, K. *et al.* In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites. *Nucleic Acids Res.* **35**, 5995–6003 (2007).
282. Sanggaard, K. M. *et al.* Branchio-oto-renal syndrome: detection of EYA1 and SIX1 mutations in five out of six Danish families by combining linkage, MLPA and sequencing analyses. *Eur. J. Hum. Genet. EJHG* **15**, 1121–1131 (2007).
283. Schneider, T. D. A brief review of molecular information theory. *Nano Commun. Netw.* **1**, 173–180 (2010).
284. Schneider, T. D. Consensus Sequence Zen. *Appl. Bioinformatics* **1**, 111–119 (2002).
285. Schneider, T. D. in *Entropy Meas. Maximum Entropy Princ. Emerg. Appl.* (ed. Karmeshu, P.) 229–237 (Springer Berlin Heidelberg, 2003).
286. Schneider, T. D. Information content of individual genetic sequences. *J. Theor. Biol.* **189**, 427–441 (1997).
287. Schneider, T. D. Measuring molecular information. *J. Theor. Biol.* **201**, 87–92 (1999).

288. Schneider, T. D. Twenty Years of Delila and Molecular Information Theory: The Altenberg-Austin Workshop in Theoretical Biology Biological Information, Beyond Metaphor: Causality, Explanation, and Unification Altenberg, Austria, 11-14 July 2002. *Biol. Theory* **1**, 250–260 (2006).
289. Schneider, T. S. Claude Shannon: biologist (information theory used in biology). *IEEE Eng. Med. Biol. Mag.* **25**, 30–33 (2006).
290. Schneider, T. D. & Rogan, P. K. Computational analysis of nucleic acid information defines binding sites. (1999).
291. Schönfelder, E.-M. *et al.* Mutations in Uroplakin IIIA are a rare cause of renal hypodysplasia in humans. *Am. J. Kidney Dis. Off. J. Natl. Kidney Found.* **47**, 1004–1012 (2006).
292. Schwaderer, P. *et al.* Clinical course and NPHS2 analysis in patients with late steroid-resistant nephrotic syndrome. *Pediatr. Nephrol. Berl. Ger.* **23**, 251–256 (2008).
293. Schwartz, J. M. (2013). *MutationTaster: ein web-basiertes Computerprogramm zur Bewertung des Krankheitspotentials von DNA-Mutationen*. PhD Thesis. Freie Universität Berlin: Germany.
294. Shankar S. (2005). *Modifier genes in the photypic manifestation of primary disease-causing mutations*. PhD Thesis. University of Iowa: USA.
295. Sharma, N. *et al.* Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. *Hum. Mutat.* **35**, 1249–1259 (2014)
296. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* **11**, 77–85 (2013).
297. Shirley, B. Interpretation, Stratification and Validation of Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences. *Univ. West. Ont. - Electron. Thesis Diss. Repos.* (2013). at <<http://ir.lib.uwo.ca/etd/1199>>
298. Shultzaberger, R. K. & Schneider, T. D. Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Res.* **27**, 882–887 (1999).
299. Shultzaberger, R. K. *et al.* Correlation between binding rate constants and individual information of E. coli Fis binding sites. *Nucleic Acids Res.* **35**, 5275–5283 (2007).
300. Shultzaberger, R. K., Bucheimer, R. E., Rudd, K. E. & Schneider, T. D. Anatomy of Escherichia coli ribosome binding sites. *J. Mol. Biol.* **313**, 215–228

- (2001).
301. Shultzaberger, R. K., Chen, Z., Lewis, K. A. & Schneider, T. D. Anatomy of Escherichia coli sigma70 promoters. *Nucleic Acids Res.* **35**, 771–788 (2007).
 302. Simpson, M. A. *et al.* Mutations in FAM20C are associated with lethal osteosclerotic bone dysplasia (Raine syndrome), highlighting a crucial molecule in bone development. *Am. J. Hum. Genet.* **81**, 906–912 (2007).
 303. Skandalis, A. & Uribe, E. A survey of splice variants of the human hypoxanthine phosphoribosyl transferase and DNA polymerase beta genes: products of alternative or aberrant splicing? *Nucleic Acids Res.* **32**, 6557–6564 (2004).
 304. Skipper, L. *et al.* Analysis of LRRK2 functional domains in nondominant Parkinson disease. *Neurology* **65**, 1319–1321 (2005).
 305. Slavotinek, A. M. *et al.* Manitoba-oculo-tricho-anal (MOTA) syndrome is caused by mutations in FREM1. *J. Med. Genet.* **48**, 375–382 (2011).
 306. Slavotinek, A. M. *et al.* VAX1 mutation associated with microphthalmia, corpus callosum agenesis, and orofacial clefting: the first description of a VAX1 phenotype in humans. *Hum. Mutat.* **33**, 364–368 (2012).
 307. Smaoui, N. *et al.* A homozygous splice mutation in the HSF4 gene is associated with an autosomal recessive congenital cataract. *Invest. Ophthalmol. Vis. Sci.* **45**, 2716–2721 (2004).
 308. Smit P. (2006). *Factors determining glucocorticoid sensitivity in man*. PhD Thesis. Erasmus University: Netherlands.
 309. Smith, A. C. *et al.* Mutations in the enzyme glutathione peroxidase 4 cause Sedaghatian-type spondylometaphyseal dysplasia. *J. Med. Genet.* **51**, 470–474 (2014).
 310. Song, H. R. *et al.* PHEX gene mutations and genotype-phenotype analysis of Korean patients with hypophosphatemic rickets. *J. Korean Med. Sci.* **22**, 981–986 (2007).
 311. Soran, H. *et al.* Proteinuria and severe mixed dyslipidemia associated with a novel APOAV gene mutation. *J. Clin. Lipidol.* **4**, 310–313 (2010).
 312. Spurdle, A. B. *et al.* Prediction and assessment of splicing alterations: implications for clinical testing. *Hum. Mutat.* **29**, 1304–1313 (2008).
 313. Stasia, M., Bordigoni, P., Martel, C. & Morel, F. A novel and unusual case of chronic granulomatous disease in a child with a homozygous 36-bp deletion in the CYBA gene (A220) leading to the activation of a cryptic splice site in intron 4.

- Hum. Genet.* **110**, 444–450 (2002).
314. Stockley, T. L. *et al.* A recurrent EYA1 mutation causing alternative RNA splicing in branchio-oto-renal syndrome: implications for molecular diagnostics and disease mechanism. *Am. J. Med. Genet. A.* **149A**, 322–327 (2009).
315. Svojanovsky, S. R., Schneider, T. D. & Rogan, P. K. Redundant designations of BRCA1 intron 11 splicing mutation; c. 4216-2A>G; IVS11-2A>G; L78833, 37698, A>G. *Hum. Mutat.* **16**, 264 (2000).
316. Sznajder, Y. *et al.* A de novo SOX10 mutation causing severe type 4 Waardenburg syndrome without Hirschsprung disease. *Am. J. Med. Genet. A.* **146A**, 1038–1041 (2008).
317. Tahsin, T., Mucaki, E. J. & Rogan, P. K. Information Theory-based exon definition analysis of mRNA splicing mutations. in 75–79 (2011).
318. Tartaglia-Polcini, A. *et al.* SPINK5, the defective gene in netherton syndrome, encodes multiple LEKTI isoforms derived from alternative pre-mRNA processing. *J. Invest. Dermatol.* **126**, 315–324 (2006).
319. Taube, J. R. *et al.* PMD patient mutations reveal a long-distance intronic interaction that regulates PLP1/DM20 alternative splicing. *Hum. Mol. Genet.* **23**, 5464–5478 (2014).
320. Tazi, J., Durand, S. & Jeanteur, P. The spliceosome: a novel multi-faceted target for therapy. *Trends Biochem. Sci.* **30**, 469–478 (2005).
321. Thompson, T. E., Rogan, P. K., Risinger, J. I. & Taylor, J. A. Splice variants but not mutations of DNA polymerase beta are common in bladder cancer. *Cancer Res.* **62**, 3251–3256 (2002).
322. Titeux, M. *et al.* Recessive dystrophic epidermolysis bullosa caused by COL7A1 hemizyosity and a missense mutation with complex effects on splicing. *Hum. Mutat.* **27**, 291–292 (2006).
323. Torregrossa, R. *et al.* Identification of GDNF gene sequence variations in patients with medullary sponge kidney disease. *Clin. J. Am. Soc. Nephrol. CJASN* **5**, 1205–1210 (2010).
324. Tosetto, E. *et al.* Novel mutations of the CLCN5 gene including a complex allele and a 5' UTR mutation in Dent disease 1. *Clin. Genet.* **76**, 413–416 (2009).
325. Tosetto, E. *et al.* Phenotypic and genetic heterogeneity in Dent's disease--the results of an Italian collaborative study. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **21**, 2452–2463 (2006).
326. Tournier, I. *et al.* A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum. Mutat.* **29**,

- 1412–1424 (2008).
327. Tram, E. *et al.* Identification of germline alterations of the mad homology 2 domain of SMAD3 and SMAD4 from the Ontario site of the breast cancer family registry (CFR). *Breast Cancer Res. BCR* **13**, R77 (2011).
328. Tsai, K.-N., Chen, G.-W. & Chen, C. Y.-C. A novel algorithm for identification of activated cryptic 5' splice sites. *J. Biomol. Struct. Dyn.* **29**, 1089–1099 (2012).
329. Tsai, K.-N. & Wang, D. Identification of activated cryptic 5' splice sites using structure profiles and odds measure. *Nucleic Acids Res.* **40**, e73 (2012).
330. Tunca, B. *et al.* Analysis of mismatch repair gene mutations in Turkish HNPCC patients. *Fam. Cancer* **9**, 365–376 (2010).
331. Tuohy, T. M. F. & Burt, R. W. in *Hered. Colorectal Cancer* (eds. Rodriguez-Bigas, M. A., Cutait, R., Lynch, P. M., Tomlinson, I. & Vasen, H. F. A.) 253–267 (Springer US, 2010). at <http://link.springer.com.proxy1.lib.uwo.ca/chapter/10.1007/978-1-4419-6603-2_14>
332. Vemula, S. R. *et al.* A rare sequence variant in intron 1 of THAP1 is associated with primary dystonia. *Mol Genet Genomic Med* **2**, 261–272 (2014).
333. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3**, 8 (2014).
334. Vinga, S. Information theory applications for biological sequence analysis. *Brief. Bioinform.* **15**, 376–389 (2014).
335. Vockley, J. *et al.* Exon skipping in IVD RNA processing in isovaleric acidemia caused by point mutations in the coding region of the IVD gene. *Am. J. Hum. Genet.* **66**, 356–367 (2000).
336. Von Kodolitsch, Y., Berger, J. & Rogan, P. K. Predicting severity of haemophilia A and B splicing mutations by information analysis. *Haemophilia* **12**, 258–262 (2006).
337. Von Kodolitsch, Y., Nienaber, C. A., Fliegner, M. & Rogan, P. K. Spleißstellenmutationen und Atherosklerose: Mechanismen und Modelle der Prädiktion. *Z. Für Kardiologie* **90**, 87–95 (2001).
338. Vorechovský, I. Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.* **34**, 4630–4641 (2006).
339. Vreeswijk, M. P. G. *et al.* Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. *Hum.*

- Mutat.* **30**, 107–114 (2009).
340. Vyhldal, C. A., Rogan, P. K. & Leeder, J. S. Development and refinement of pregnane X receptor (PXR) DNA binding site model using information theory: insights into PXR-mediated gene regulation. *J. Biol. Chem.* **279**, 46779–46786 (2004).
341. Wadt, K. *et al.* A cryptic BAP1 splice mutation in a family with uveal and cutaneous melanoma, and paraganglioma. *Pigment Cell Melanoma Res.* **25**, 815–818 (2012).
342. Wan, L. *et al.* Identification of eight novel mutations of the acid alpha-glucosidase gene causing the infantile or juvenile form of glycogen storage disease type II. *J. Neurol.* **255**, 831–838 (2008).
343. Wang, E., Dimova, N. & Cambi, F. PLP/DM20 ratio is regulated by hnRNPH and F and a novel G-rich enhancer in oligodendrocytes. *Nucleic Acids Res.* **35**, 4164–4178 (2007).
344. Wang, J. *et al.* Identification of a novel specific CYP2B6 allele in Africans causing impaired metabolism of the HIV drug efavirenz. *Pharmacogenet. Genomics* **16**, 191–198 (2006).
345. Wang, P. *et al.* Novel mutations of the PAX6 gene identified in Chinese patients with aniridia. *Mol. Vis.* **12**, 644–648 (2006).
346. Watnick, T. J., Garcia-Gonzalez, M., Germino, G. G. & Jones, J. G. Pkd mutations and evaluation of same. (2008).
347. Weiss, R. B., Lalouel, J.-M., Pankow, J. & Dunn, D. M. Variants of NEDD4L associated with hypertension and viral budding. US20050277123 A1 (2005).
348. Wessagowit, V. & McGrath, J. A. Clinical and molecular significance of splice site mutations in the plakophilin 1 gene in patients with ectodermal dysplasia-skin fragility syndrome. *Acta Derm. Venereol.* **85**, 386–388 (2005).
349. Wessagowit, V., Kim, S.-C., Woong Oh, S. & McGrath, J. A. Genotype–Phenotype Correlation in Recessive Dystrophic Epidermolysis Bullosa: When Missense Doesn’t Make Sense. *J. Invest. Dermatol.* **124**, 863–866 (2005).
350. Wessagowit, V., Nalla, V. K., Rogan, P. K. & McGrath, J. A. Normal and abnormal mechanisms of gene splicing and relevance to inherited skin diseases. *J. Dermatol. Sci.* **40**, 73–84 (2005).
351. Wong, T. *et al.* Potential of fibroblast cell therapy for recessive dystrophic epidermolysis bullosa. *J. Invest. Dermatol.* **128**, 2179–2189 (2008).
352. Wu, J. Y., Yuan, L. & Havlioglu, N. in *Encycl. Mol. Cell Biol. Mol. Med.*

(Wiley-VCH Verlag GmbH & Co. KGaA, 2006).

353. Wu, Y., Zhang, Y. & Zhang, J. Distribution of exonic splicing enhancer elements in human genes. *Genomics* **86**, 329–336 (2005).
354. Xiong, Y. *et al.* A systematic genetic polymorphism analysis of the CYP2C9 gene in four different geographical Han populations in mainland China. *Genomics* **97**, 277–281 (2011).
355. Xu, X. *et al.* Sequence variations of GRM6 in patients with high myopia. *Mol. Vis.* **15**, 2094–2100 (2009).
356. Yan, Q. & Guan, M.-X. Identification and characterization of mouse TRMU gene encoding the mitochondrial 5-methylaminomethyl-2-thiouridylate-methyltransferase. *Biochim. Biophys. Acta* **1676**, 119–126 (2004).
357. Yan, Q. *et al.* Human TRMU encoding the mitochondrial 5-methylaminomethyl-2-thiouridylate-methyltransferase is a putative nuclear modifier gene for the phenotypic expression of the deafness-associated 12S rRNA mutations. *Biochem. Biophys. Res. Commun.* **342**, 1130–1136 (2006).
358. Yang, M. Novel method for discerning the action of selection during evolution. *J. Biomed. Sci. Eng.* **03**, 109–113 (2010).
359. Yu, B. In Silico Interpretation of the Splicing Code and Estimating the Abundance of Expressed mRNA Isoforms. *Hum. Mutat.* **34**, v–v (2013).
360. Yu, B. Role of in silico tools in gene discovery. *Mol. Biotechnol.* **41**, 296–306 (2009).
361. Yu, H. & Patel, S. B. Recent insights into the Smith-Lemli-Opitz syndrome. *Clin. Genet.* **68**, 383–391 (2005).
362. Zaffanello, M. *et al.* Type IV Bartter syndrome: report of two new cases. *Pediatr. Nephrol. Berl. Ger.* **21**, 766–770 (2006).
363. Zampieri, S. *et al.* Splicing mutations in glycogen-storage disease type II: evaluation of the full spectrum of mutations and their relation to patients' phenotypes. *Eur. J. Hum. Genet. EJHG* **19**, 422–431 (2011).
364. Zhang, X. H.-F., Heller, K. A., Hefter, I., Leslie, C. S. & Chasin, L. A. Sequence Information for the Splicing of Human Pre-mRNA Identified by Support Vector Machine Classification. *Genome Res.* **13**, 2637–2650 (2003).
365. Zhang, K., Nowak, I., Rushlow, D., Gallie, B. L. & Lohmann, D. R. Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression. *Hum. Mutat.* **29**, 475–484 (2008).
366. Zhang, Q. *et al.* A variant form of Oguchi disease mapped to 13q34 associated

with partial deletion of GRK1 gene. *Mol. Vis.* **11**, 977–985 (2005).

367. Zhao, L. *et al.* Two novel FBN1 mutations associated with ectopia lentis and marfanoid habitus in two Chinese families. *Mol. Vis.* **15**, 826–832 (2009).

Appendix E: Supplementary Methods (Chapter 3)

Design of Tiled Hybridization Capture Reagent for BRCA Gene Panel

Probe sequences within single copy intervals¹ in *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *PALB2*, and *TP53* were selected using PICKY 2.2 software²; settings set to 65°C T_m, 30-70% GC content, 5 probes per sequence, 20 nt maximum overlap, all other settings default. PICKY will only report a maximum of 5 oligos per sequence analyzed. Therefore, gene sequences were split into 100 nt segments, overlapping by 50 nt. Probes were designed for both the forward and reverse strand of each gene. These overlapping and opposite-stranded sequences were run through PICKY separately; as the program would remove identical probes selected from two sequence segments (including probes which are reverse compliments of each other). This method helps result in significant probe overlap, which leads to a more efficient capture. If regions were lacking probes due to high/low %GC, the process was repeated with expanded GC settings (20% minimum or 80% maximum) and probes over those regions were added to the initial probe file. A Perl script entitled “Amalgamated-Post-Picky-Program” was written to perform 4 tasks: 1) MPI-BLAT (through Shared Hierarchical Academic Research Computing Network or SHARCNET³) each PICKY-selected oligo to the transcriptome and eliminate any which match elsewhere (< 1 hour on 16 nodes), 2) Eliminate redundant probes (i.e. removal of a smaller probe overlapped by a larger probe on the same strand), 3) Reduce highly overlapped regions by eliminating one of two near identical probes that differ by a 1 nt shift, and 4) Generate a BED genome browser track of the accepted oligos for visual evaluation of the coverage of the generated tiling array oligos.

Generating, Cleaving, and Purifying Tiled BRCA Microarray Oligos

Primer binding sites were added to each end of the designed capture oligos (5' ATCGCACCAGCGTGTN₃₆₋₇₀CACTGCGGCTCCTCA). The selected sequences were then synthesized onto two cleavable 12K microarray chips using a Combimatrix B3 CustomArray Synthesizer (CustomArray, Inc., Bothell, WA) in our laboratory, requiring approximately 48 hours. Forward and reverse strand oligos were placed on separate chips

to avoid cross-hybridization between complementary strands, which could reduce capture efficiency.

The cleavable microarrays were treated with concentrated (14.5N) ammonium hydroxide at 65°C for 4 hours. This served to break the sulfonyl-amidite bond linking the oligonucleotide to the microarray. The base was cooled, transferred into a microcentrifuge tube, and placed into a speed-vac for 1 hour at 65°C. The resulting pellet was resuspended into 100uL of 1x TE buffer, and was then purified using a MicroSPIN DNA column (#11814419001, Roche, Indianapolis, IN). Purified oligos were then amplified by conventional PCR (25 cycles) using Kapa HiFi DNA Polymerase (#KK2602, KapaBiosystems, Wilmington, MA) (forward 5-CTGGGAATCGCACCAGCGTGT-3; reverse 5-CGTGGATGAGGAGCCGCAGTG-3).

The PCR product was purified using a Qiagen MinElute PCR Purification Kit (#28006, Qiagen, Valencia, CA) and then amplified again (25 additional cycles) using a forward primer with an SP6 promoter-binding site (5-CATACGATTTAGGTGACACTATAGAAATCGCACCAGCGTGT-3). Biotin-labelled RNA bait was generated from this product using a MAXIscript SP6 *in vitro* transcription kit (#AM1310, Ambion, Carlsbad, CA) with a UTP to biotin-16-UTP (#11388908910, Roche) ratio of 4 to 1.

Sample Preparation, Library Preparation, and Oligo Capture for Sequencing

Genomic DNA (gDNA) was previously extracted from whole blood using the MagNA Pure Compact Nucleic Acid Isolation kit I (#03730964001, Roche) and stored in 1x low TE buffer. Samples with inadequate available gDNA (< 3 µg) were whole genome amplified using the Illustra GenomiPhi V2 DNA Amplification Kit (#25-6600-30, GE Healthcare, Little Chalfont, UK). The gDNA was diluted to 100 ng/µL in a volume of 51 µL for S220 Focused-ultrasonicator (Covaris, Woburn, USA) shearing (150-300 nt fragments generated with the following settings: Time 120 sec, Duty cycle 10%, Intensity 5, and Cycles per burst 200).

The sheared samples were prepared using KAPA Biosystems Standard (KK8200, Kapa Biosystems) and High Throughput (KK8234, Kapa Biosystems) Library Preparation kits, following the manufacturer's protocol. Standard Illumina paired-end and multiplex adapter oligos and primers (sequences provided by Illumina) were purchased from IDT (Coralville, IA). Adapters were hybridized together by mixing two oligos to a final concentration of 100 μ M each, heating to 95°C for 5 minutes on a thermocycler (Mastercycler pro, Eppendorf) and gradually cooling 0.1°C/sec to 4°C. Samples being treated were initially purified using Qiagen MinElute PCR purification kits and Sigma GenElute gel purification kits (#NA1111, Sigma, St. Louis, MO). Sample loss was greatly decreased by switching to DNA-binding Agencourt Ampure XP beads (#A63880, Beckman Coulter, Brea, CA), using the protocol described in Fisher *et al.* (2011), which allows the re-use of beads by the rebinding of DNA using a 20% polyethylene glycol (PEG) in a 2.5M NaCl solution and avoids gel extraction and column purification steps⁴. The switch to Ampure beads allowed for the automation of sample preparation using a Beckman Coulter BioMek FX workstation, increasing sample throughput to 32 samples processed simultaneously (end-repair, A-tailing, and adaptor ligation steps were performed on the BioMek FX along with all wash steps; PCR steps were performed separately in a thermocycler). The amplified library samples were then reduced to a volume of 6.8 μ L using a SpeedVac concentrator prior to genomic capture.

Genes of interest were enriched with Tiled *BRCA* RNA bait, following a modified version of the hybridization selection protocol from Gnirke *et al.* (2009)⁵. Modifications which increased coverage include: a) an increase in sample quantity from 0.5 μ g to half of the library prepared per capture (1 to 2 μ g depending on sample prep yield), b) an RNA bait increase from 0.5 to 1.5 μ g and, c) an increase in the quantity of M-280 streptavidin Dynabeads (#1205D, Invitrogen, Carlsbad, CA) from 50 μ L to 75 μ L to account for this increase in RNA bait concentration. Genomic sequences in each library sample were captured in two separate solutions, one for each strand of RNA bait (which were pooled at the end of the procedure). The capture was then incubated at 65°C for 66-72 hours, in a thermocycler with the heated lid set to 80°C (to help reduce volume loss due to evaporation). Forward and reverse strand capture reactions for the same sample

were then purified with streptavidin beads and mixed on a nutator (Adams Nutator, #1105, Clay Adams, Franklin Lakes, NJ) for 30 minutes at room temperature. Use of freshly transcribed RNA bait also greatly improved coverage values. These improvements enabled multiplexed sequencing (4 samples per lane), where the index was added during the post-hybridization amplification using standard Illumina Multiplex PCR Primers (#2, 4, 6, 7; chosen for index dissimilarity). After post-hybridization amplification, remaining primers were removed by Ampure bead purification. DNA samples were then quantified using qPCR following the protocol outlined by KAPA Library Quantification Kit for Illumina Platform (#KK4824, KAPA Biosystems). Samples were then pooled (if multiplexing), and treated to standard Illumina paired-end sequencing on a Genome Analyzer Iix. Read length was increased from 36x36 (sequencing batch 1 and 2) to 50x50 (batch 3) and finally to 70x7x70 (batch 4, 5 and 6) to maximize overall coverage.

Position Weight Matrix Generator (PoWeMaGen)

We have developed an information weight matrix generator designed to automatically create position weight matrices (PWMs). These PWMs are used to accurately predict and localize transcription factor binding sites (TFBSs) from genome-scale epigenetic data (i.e. ChIP-seq). The PoWeMaGen software engine, in its current implementation, is set to run exclusively on SHARCNET. The script, however, can easily be converted to run on different Linux-based platforms. The engine, outlined in **Figure SM1** below, has three primary sections: Preliminary file processing, ChIP-seq data filtering, and the execution of the Bipad.

The PoMaWeGen model analyzer program acts as an intermediary in the model generation pipeline, beginning by processing of input, output, and run-time-dependent files. This input includes the epigenetic data, and if selected, the DNase hypersensitivity data set (track: EncodeRegDnaseClusteredV3). The epigenetic data was filtered by intersection using intersectBed, from the BEDtools package⁶. In this study, available ChIP-seq and CLIP-seq (Cross-Linking ImmunoPrecipitation) data was intersected with DNase I hypersensitivity tracks for only TFs.

Model building is based on Bipad⁷, an algorithm we previously published to minimize entropy across a set of unaligned sites. Bipad is run with biologically-inspired parameters, i.e. whether the site is known to be homogenous or bipartite, if bipartite, allowing for defined range of gap lengths separating half sites. Each sequence may, but is not required, to contain one binding site. This program employs the sq framework to run jobs in parallel. The parts of this program that run in parallel perform the vast majority of computation, and are perfectly parallel. This program spawns a job for each model that it generates. Only one job is spawned if UII is not employed (i.e. if confined to a specific motif length). Each spawned job runs a single instance of Bipad, which comprises the vast majority of the computational load for this step.

Epigenetic input is commonly provided as interval data, but Bipad requires input as sequence. PoWeMaGen converts the interval data to sequence using the reference genome (Encyclopedia of DNA Elements [ENCODE]⁸) and the BEDtools program `fastaFromBed`. Then a custom java program called `Fasta2Bipad.java` converts this fasta file to the format required by Bipad. Finally, the engine executes Bipad, with user-selected arguments which specify, among other things, whether the model is homogenous or bipartite, use ZOOPS (Zero or One Occurrence Per Sequence) or OOPS (One Occurrence Per Sequence), whether to consider one or both reading strands, estimated motif length, and the number of Monte Carlo cycles required.

In instances where the length of the site has not been well defined in published data, the Unit Information Increment Index heuristic, which maximizes the information density across the binding site as a function of binding site length⁹ was used to compute the length. This algorithm chooses a range of motif lengths, considers DNA helical periodicity of 10.6 base pairs, and varies these lengths based upon the result of UII computations.

Selecting TFs for Model Building

TF ChIP-seq data (track: `wgEncodeRegTfbsClusteredV2`) and DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE were downloaded and intersected using the

Galaxy Browser¹⁰⁻¹², to select for DNA regions accessible to TFs for binding. We then extracted genomic intervals from 10 kb upstream of the transcriptional start site, up to the end of intron 1, for each of the 7 genes. We identified 146 TFs with evidence for binding to the promoters of the genes in our panel.

Generating SNPfold Input

A script was written to retrieve 100 nt of mRNA sequence up- and downstream of the variant of interest while altering the sequence based on any other variants found in the region during sequencing (phase was ignored). This was used as the input data for SNPfold. Only transcribed sequence was included (if variant was < 100 nt from start/end of transcription). For long-range effects, the variants were evaluated against the entire UTR they resided in (using the reference sequence; patient-specific variants were not included).

Generation of RNA Binding Protein Models from RBPDB and CISBP-RNA

PWMs for 156 RBPs were downloaded from the RBPDB and CISBP-RNA¹³⁻¹⁵. PWMs containing frequencies were converted into information weight matrices (N = 147). Binding sites for factors with low expression in normal breast tissues (N= 59) based on the GTEx database¹⁶ RPKM < 10 were removed. We also eliminated binding sites for factors with highly variable expression among the 57 available breast tissue samples (where median RPKM < 3 standard deviations; N=11). All UTR variants were then analyzed with models for the remaining RBPs (N=76).

Models for the following factors were used to determine variant effect on RBP binding sites: A2BP1, ANKHD1, ANKRD17, BRUNOL4, BRUNOL5, BRUNOL6, BX511012.1, CIRBP, CSDA, DAZAP1, EIF4B, ELAVL1, FMR1, FUS, FXR1, FXR2, G3BP2, HNRNPA1, HNRNPA2B1, HNRNPA3, HNRNPC, HNRNPF, HNRNPH1, hnRNPK, HNRNPL, hnRNPLL, HNRNPR, HNRPDL, KHDRBS1, KHSRP, MATR3, MBNL1, NCL, NONO, NOVA2, PABPC1, PABPC4, PABPN1, PCBP1, PCBP2, PCBP4, PSPC1, PTBP1, PTBP2, PUM2, RALY, RBFOX2, RBM28, RBM42, RBM5, RBM6, RBM8A, RBMS1, RBMX, ROD1, SAMD4B, SART3, SF3B4, SFPQ,

SNRNP70, SNRPA, SNRPB2, SRSF1, SRSF10, SRSF2, SRSF4, SRSF6, SRSF7, SRSF9, SYNCRIP, TAF15, TARDBP, TIA1, U2AF1, U2AF2, ZCRB1, ZNF638.

Models were not built for the following models because only blank files were provided by CISBP-RNA (N=9): ELAVL4, FUBP1, G3BP1, HNRNPAB, KHSRP, NOVA1, PUM1, PUM2, SRSF5, TRA2B, ZRANB2 (KHSRP and PUM2 models present in RBPDB, therefore not included in count).

Models were not built for the following factors due to variable expression between the 57 available breast tissue samples (where median RPKM < 3 standard deviations) (N=11): ACO1, MBNL2, PABPC1L, RBM3, RBM38, SRSF3, YBX1, YTHDC1, ZFP36, ZFP36L1, ZFP36L2.

Models were not built for the following factors due to low or extremely variable expression in normal breast tissue (N=59): A1CF, CELF3, CNOT4, CPEB2, CPEB3, CPEB4, EIF2S1, ELAVL2, ELAVL3, ENOX1, ENOX2, ESRP1, ESRP2, HNRNPA1L2, HNRNPCL1, HNRNPH2, IGF2BP1, IGF2BP2, IGF2BP3, KHDRBS2, KHDRBS3, LIN28A, LIN28B, MBNL3, MEX3B, MEX3C, MEX3D, MSI1, MSI2, PABPC3, PABPC5, PABPN1L, PCBP3, PPRC1, QKI, RBFOX3, RBM24, RBM4, RBM41, RBM45, RBM46, RBM47, RBM4B, RBMS2, RBMS3, RBMXL1, RBMXL2, RBMXL3, RBMY1A1, RBMY1B, RBMY1D, RBMY1E, RBMY1F, RBMY1J, SAMD4A, SRSF12, STAR, YBX2, ZC3H10, ZFP36.

Supplementary Methods References

1. Dorman, S. N., Shirley, B. C., Knoll, J. H. M. & Rogan, P. K. Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Res.* 41, e81 (2013).
2. Chou, H.-H., Hsia, A.-P., Mooney, D. L. & Schnable, P. S. Picky: oligo microarray design for large genomes. *Bioinforma. Oxf. Engl.* 20, 2893–2902 (2004).

3. Shared Hierarchical Academic Research Computing Network (SHARCNET). at <https://www.sharcnet.ca/my/front/>. Accessed June 1, 2015.
4. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* 12, R1–2011–12–1–r1. Epub 2011 Jan 4 (2011).
5. Gnirke, A. *et al.* Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. *Nat. Biotechnol.* 27, 182–189 (2009).
6. Quinlan, A. R. *et al.* Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* 20, 623–635 (2010).
7. Bi, C. & Rogan, P. K. Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.* 32, 4979–4991 (2004).
8. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
9. Bi, C. & Rogan, P. K. Determining Thresholds for Binding Site Sequence Models Using Information Theory. (2005).
10. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* Chapter 19, Unit 19.10.1–21 (2010).
11. Giardine, B. *et al.* Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* 15, 1451–1455 (2005).
12. Goecks, J., Nekrutenko, A., Taylor, J. & \$author.lastName, \$author firstName. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11, R86 (2010).
13. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* 39, D301–8 (2011).

14. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443 (2014).
15. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 172–177 (2013).
16. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348, 648–660 (2015).

Workflow Diagram

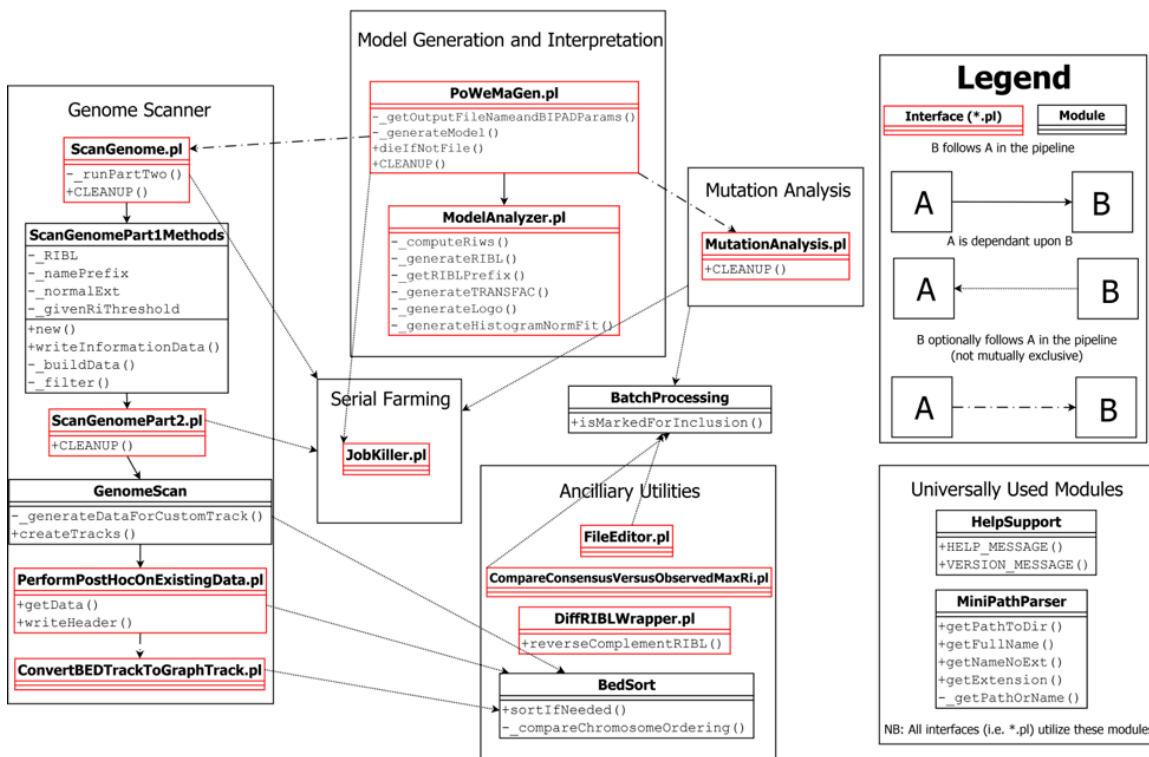


Figure SM1. Diagram of the PWM Generator

UML Diagram of the distributed processing system. The modules shown call the Bipad and Shannon pipeline C++ engines to perform entropy minimization and genome scanning. Solid arrows indicate embedded capabilities called by the preceding software module; discontinuous arrows indicate separate programs with compatible input/outputs. Ancillary utilities are used for post-processing of output from main programs to compare different models and sort results from GenomeScan output.

Appendix F: Ethics Approval and Amendments for Patient Recruitment



Research Ethics

Use of Human Participants - Ethics Approval Notice

Principal Investigator:
 File Number:103746
 Review Level:Delegated
 Approved Local Adult Participants:650
 Approved Local Minor Participants:0
 Protocol Title:Sequencing and functional analysis of genetic variants of unknown significance in women with inherited breast and/or ovarian cancer
 Department & Institution:Science/Biochemistry,Western University
 Sponsor:Canadian Breast Cancer Foundation

Ethics Approval Date:September 25, 2013 Expiry Date:March 31, 2018
 Documents Reviewed & Approved & Documents Received for Information:

Document Name	Comments	Version Date
Letter of Information & Consent	Consent form	2013/04/07
Letter of Information	Information letter	2013/04/07
	Power analysis. Justification of sample size	2013/04/07
Western University Protocol		2013/04/18
Recommendations Form	Recommendations	2013/07/11
Revised Western University Protocol	REB Protocol exported and revised - September 11, 2013	2013/09/11
Letter of Information & Consent	Revised combined letter of information and consent form	2013/09/11
Response to Board Recommendations	Response to questions from REB on original submission.	2013/09/11
Other	Response Card to be sent to eligible study participants	2013/09/11

This is to notify you that The University of Western Ontario Research Ethics Board for Health Sciences Research Involving Human Subjects (HSREB) which is organized and operates according to the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans and the Health Canada/ICH Good Clinical Practice Practices: Consolidated Guidelines; and the applicable laws and regulations of Ontario has reviewed and granted approval to the above referenced revision(s) or amendment(s) on the approval date noted above. The membership of this REB also complies with the membership requirements for REB's as defined in Division 5 of the Food and Drug Regulations.

The ethics approval for this study shall remain valid until the expiry date noted above assuming timely and acceptable responses to the HSREB's periodic requests for surveillance and monitoring information. If you require an updated approval notice prior to that time you must request it using the University of Western Ontario Updated Approval Request Form.

Members of the HSREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the HSREB.

The Chair of the HSREB is [redacted] The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Ethics Officer to Contact for Further Information

--	--	--

This is an official document. Please retain the original in your files.

Appendix G: Letter of Invitation for 7 Genes



LETTER OF INFORMATION AND CONSENT FORM
Sequencing and functional analysis of genetic variants of unknown significance in women with inherited breast and/or ovarian cancer study.

Page 1 of 4

PRINCIPAL INVESTIGATOR:

LOCAL CO-INVESTIGATORS:

PURPOSE OF RESEARCH:

Multiple cases of cancer can occur in families due to environmental and inherited factors. In this study, investigators are looking for an inherited factor (otherwise known as a gene), that has a change or “*mutation*” in it that might be associated with having a higher risk of cancer. The discovery of genes and the methods used in laboratories to search for gene mutations have evolved significantly over the years, and this study involves the application of *Next Generation Sequencing* to search for mutations in seven breast cancer-related genes.

You are being invited to participate in this study because your personal and/or family history of cancer is suggestive of having an inherited breast/ovarian cancer gene mutation. Although you previously participated in BRCA1/2 gene testing, specifically, no mutation was identified using the testing methods available at that time. This study is offering an improved method of genetic analysis for BRCA1 and BRCA2, and includes testing of five additional breast cancer-related genes. It is estimated that approximately 650 women and men will be eligible to participate in this study through the London Health Sciences Centre (LHSC) Cancer Genetics clinic.

ELIGIBILITY:

Women or men, age 25 to 75, who have a >10% risk of having an inherited breast/ovarian cancer gene mutation as determined by family history review and LHSC-specific BRCA1/2 mutation detection rates are eligible to participate in this study. Participants must be English-speaking, and sufficiently competent to provide informed consent.

PROCEDURE:

If you agree to participate in this study, then you will be requested to contact the study Research Assistant at _____, who will schedule an appointment with you to first review and

Page 2 of 4

update your personal and family history of cancer, explain the research study and facilitate your provision of informed consent, and arrange for a sample of your blood (15 ml or 3 teaspoons) to be drawn. You will have every opportunity to ask questions prior to signing this consent form. Should you require further information related to your personal and/or family cancer risks and available cancer screening, arrangements can be made for you to speak or meet with a Genetic Counsellor at a later time. It is estimated that it will take up to 30 minutes of your time to review your personal/family history of cancer and provide informed consent. It is estimated that it will take approximately 6 months for your research test results to be complete.

At the conclusion of your genetic analysis, you will be contacted by telephone to review your test results. Only participants whose research test results reveal clinical importance (i.e. are believed to be responsible for your personal and family history of cancer) will be invited to return to the Genetics clinic for further genetic counselling and clinical validation testing. If you decide not to receive your research test results from this study, you will not be contacted again in regards to your personal involvement in this study. In the event that you are unavailable at the time when your research test results are complete, you will be asked to provide an additional contact name (next of kin) and telephone number so that the research team can ensure that your genetic research results are communicated.

Your participation in this study is voluntary. If you are interested in participating in this study, please contact the Research Assistant at 519-685-8500 ext.74996, to schedule an appointment. If you are not interested in participating in this study, please complete the enclosed Response Card that indicates your wishes to not be involved, and return it and all of the enclosed items to us in the envelope provided. You may refuse to participate, refuse to answer any questions, or withdraw from this study at any time with no effect on your current or future care (i.e. choice of, or access to, cancer treatment or cancer screening). In addition, your participation in this study may be terminated at any time by the research team, with or without your consent.

PRIVACY & CONFIDENTIALITY:

Your personal contact information, date of birth, personal and/or family history of cancer, and genetic test results will be kept strictly confidential. All participants will be entered into a clinical research database that is located on the hospital network, and is therefore password-protected and behind the hospital firewall. Your signed consent form and any other hard copies of your personal information (i.e. pathology reports, family pedigree) will be kept in a locked filing cabinet in the Senior Genetic Counsellor, Ms. Karen Panabaker's office. Your blood sample will not be used for commercial purposes. Your sample will be stored in specified storage facility in the investigator's laboratory. Laboratory safeguards to protect your privacy and confidentiality include removal of identifiers from the specimen samples and replacement with numeric codes.

Any information generated by this study will not become part of your personal medical record, unless you provide consent to learn more about potentially important research findings. All study information will be available to the study research team only, and will not be released to any other party, except upon your expressed written consent. Representatives of the Western University Health Sciences Research Ethics Board may contact you or require access to your study-related records to monitor the

Page 3 of 4

conduct of the research. Representatives of this research study's granting agency, the Canadian Breast Cancer Foundation, may inspect the research records at any time. Furthermore, if the results of this study are published, your name will not be used. At the completion of this study, by January 1, 2018, all data generated from this study will be deleted or shredded.

BENEFITS & RISKS:

Knowledge gained from this study may or may not benefit you directly, however, information that is obtained from this research study might directly benefit members of your family and could possibly be helpful for others diagnosed with this condition in the future. Knowledge gained from this study may help the future development of earlier diagnostic tests, new forms of treatment and may help improve genetic counselling for hereditary cancer.

Having a blood sample drawn is a safe and minimally invasive procedure. Common side effects include temporary discomfort with the needle stick, minimal and temporary bleeding, or bruising. We acknowledge that questions or discussions about breast or ovarian cancer family history can, for some individuals, increase a person's anxiety concerning the development of the disease. You will have access to a Genetic Counsellor at all times, and if needed, referrals can be made to social work or other psychosocial support.

REQUEST FOR MORE INFORMATION:

If you have any questions or concerns about this study, you may contact the local principal investigator, _____, or any of the other local co-investigators. If you have any questions about your rights as a research participant or the conduct of this study, you may contact

You will not be compensated for your participation in this research study. You do not waive any legal rights by signing this information and consent form.

**HEREDITARY CANCER GENE IDENTIFICATION
CONSENT FORM**

I have read the letter of information, have had the nature of the study explained to me, and I agree to participate. All questions have been answered to my satisfaction.

At the conclusion of my genetic analysis, please contact me at either of the following numbers:

Home number: _____

Cell number: _____

If I am not available in the future when my research test results become available, please contact:

Full name: _____

Relationship to me: _____

Telephone contact number: (h) _____ (c) _____

I have received a signed copy of this consent form for my own records.

Subject's Signature	Printed name	Date
---------------------	--------------	------

Signature of Person Obtaining Consent	Printed name/Study Role	Date
--	-------------------------	------

Appendix H: Response Card

*Sequencing and functional analysis of genetic variants of unknown significance
in women with inherited breast and/or ovarian cancer study.*

RESPONSE CARD

I am NOT interested in participating in the above mentioned study.

Printed Name

Signature

Date

Appendix I: Letter of Invitation for 23 Genes



LETTER OF INFORMATION AND CONSENT FORM
Sequencing and functional analysis of genetic variants of unknown significance in women with inherited breast and/or ovarian cancer study.

Page 1 of 4

PRINCIPAL INVESTIGATOR:

LOCAL CO-INVESTIGATORS:

PURPOSE OF RESEARCH:

Multiple cases of cancer can occur in families due to environmental and inherited factors. In this study, investigators are looking for an inherited factor (otherwise known as a gene), that has a change or “*mutation*” in it that might be associated with having a higher risk of cancer. The discovery of genes and the methods used in laboratories to search for gene mutations have evolved significantly over the years, and this study involves the application of *Next Generation Sequencing* to search for mutations in twenty-three breast cancer-related genes.

You are being invited to participate in this study because your personal and/or family history of cancer is suggestive of having an inherited breast/ovarian cancer gene mutation. Although you previously participated in BRCA1/2 gene testing, specifically, no mutation was identified using the testing methods available at that time. This study is offering an improved method of genetic analysis for BRCA1 and BRCA2, and includes testing of twenty-one additional breast cancer-related genes. It is estimated that approximately 650 women and men will be eligible to participate in this study through the London Health Sciences Centre (LHSC) Cancer Genetics clinic.

ELIGIBILITY:

Women or men, age 25 to 75, who have a >10% risk of having an inherited breast/ovarian cancer gene mutation as determined by family history review and LHSC-specific BRCA1/2 mutation detection rates are eligible to participate in this study. Participants must be English-speaking, and sufficiently competent to provide informed consent.

PROCEDURE:

If you agree to participate in this study, then you will be requested to contact the study Research Assistant at _____, who will update your personal and family history of cancer, explain the

Page 2 of 4

research study and facilitate your provision of informed consent. In many instances, banked DNA will likely be available in the laboratory from your previous genetic analyses, to be used for this study. If sufficient DNA is not available, arrangements will be made for you to have a new sample of your blood drawn (15 ml or 3 teaspoons). You will have every opportunity to ask questions prior to signing this consent form. Should you require further information related to your personal and/or family cancer risks and available cancer screening, arrangements can be made for you to speak or meet with a Genetic Counsellor at a later time. It is estimated that it will take up to 30 minutes of your time to review your personal/family history of cancer and provide informed consent. It is estimated that it will take approximately 6 months for your research test results to be complete.

At the conclusion of your genetic analysis, you will be contacted by telephone to review your test results. Only participants whose research test results reveal clinical importance (i.e. are believed to be responsible for your personal and family history of cancer) will be invited to return to the Genetics clinic for further genetic counselling and clinical validation testing. If you decide not to receive your research test results from this study, you will not be contacted again in regards to your personal involvement in this study. In the event that you are unavailable at the time when your research test results are complete, you will be asked to provide an additional contact name (next of kin) and telephone number so that the research team can ensure that your genetic research results are communicated.

Your participation in this study is voluntary. If you are interested in participating in this study, please contact the Research Assistant at _____, to schedule an appointment. If you are not interested in participating in this study, please complete the enclosed Response Card that indicates your wishes to not be involved, and return it and all of the enclosed items to us in the envelope provided. You may refuse to participate, refuse to answer any questions, or withdraw from this study at any time with no effect on your current or future care (i.e. choice of, or access to, cancer treatment or cancer screening). In addition, your participation in this study may be terminated at any time by the research team, with or without your consent.

PRIVACY & CONFIDENTIALITY:

Your personal contact information, date of birth, personal and/or family history of cancer, and genetic test results will be kept strictly confidential. All participants will be entered into a clinical research database that is located on the hospital network, and is therefore password-protected and behind the hospital firewall. Your signed consent form and any other hard copies of your personal information (i.e. pathology reports, family pedigree) will be kept in a locked filing cabinet in the Senior Genetic Counsellor, Ms. _____'s office. Your blood sample will not be used for commercial purposes. Your sample will be stored in specified storage facility in the investigator's laboratory. Laboratory safeguards to protect your privacy and confidentiality include removal of identifiers from the specimen samples and replacement with numeric codes.

Any information generated by this study will not become part of your personal medical record, unless you provide consent to learn more about potentially important research findings. All study information will be available to the study research team only, and will not be released to any other party, except upon your expressed written consent. Representatives of the Western University Health Sciences

Page 3 of 4

Research Ethics Board may contact you or require access to your study-related records to monitor the conduct of the research. Representatives of this research study's granting agency, the Canadian Breast Cancer Foundation, may inspect the research records at any time. Furthermore, if the results of this study are published, your name will not be used. At the completion of this study, by January 1, 2018, all data generated from this study will be deleted or shredded.

BENEFITS & RISKS:

Knowledge gained from this study may or may not benefit you directly, however, information that is obtained from this research study might directly benefit members of your family and could possibly be helpful for others diagnosed with this condition in the future. Knowledge gained from this study may help the future development of earlier diagnostic tests, new forms of treatment and may help improve genetic counselling for hereditary cancer.

Having a blood sample drawn is a safe and minimally invasive procedure. Common side effects include temporary discomfort with the needle stick, minimal and temporary bleeding, or bruising. We acknowledge that questions or discussions about breast or ovarian cancer family history can, for some individuals, increase a person's anxiety concerning the development of the disease. You will have access to a Genetic Counsellor at all times, and if needed, referrals can be made to social work or other psychosocial support.

REQUEST FOR MORE INFORMATION:

If you have any questions or concerns about this study, you may contact the local principal investigator, _____, or any of the other local co-investigators. If you have any questions about your rights as a research participant or the conduct of this study, you may contact

You will not be compensated for your participation in this research study. You do not waive any legal rights by signing this information and consent form.

Page 4 of 4

**HEREDITARY CANCER GENE IDENTIFICATION
CONSENT FORM**

I have read the letter of information, have had the nature of the study explained to me, and I agree to participate. All questions have been answered to my satisfaction.

At the conclusion of my genetic analysis, please contact me at either of the following numbers:

Home number: _____

Cell number: _____

If I am not available in the future when my research test results become available, please contact:

Full name: _____

Relationship to me: _____

Telephone contact number: (h) _____ (c) _____

I have received a signed copy of this consent form for my own records.

Subject's Signature Printed name Date

Signature of Printed name/Study Role Date
Person Obtaining Consent

Appendix J: Final Letter to Patients



FINAL LETTER OF CONTACT

Sequencing and functional analysis of genetic variants of unknown significance in women with inherited breast and/or ovarian cancer study.

Page 1 of 2

PRINCIPAL INVESTIGATOR:

LOCAL CO-INVESTIGATORS:

As you may recall, you were previously sent a *Letter of Information and Consent Form*, inviting you to participate in a research study for patients with a personal and family history of breast and/or ovarian cancer who have received inconclusive results following previous genetic susceptibility testing.

We are re-contacting you because you have neither declined (by returning a completed Response Card), nor returned a signed consent form indicating your wish to participate in this study. We would like to inform you that the study **will close to new enrollment on March 25th, 2015**. Unless we receive your written consent form (contained on following page) by this date, we will no longer be able to enroll you in this study.

Please note that if you already gave verbal consent to participate, this is not sufficient for you to be enrolled – the written consent form is absolutely required. If you are unsure whether or not you have given verbal consent only, or if you would like to learn more about this study before making your decision, please contact the Research Assistant for this study at _____ who will help clarify the status of your participation in the study.

REQUEST FOR MORE INFORMATION:

If you have any questions or concerns about this study, you may also contact the local principal investigator, _____, or any of the other local co-investigators. If you have any questions about your rights as a research participant or the conduct of this study, you may contact

You will not be compensated for your participation in this research study. You do not waive any legal rights by signing this information and consent form.

Page 2 of 2

**HEREDITARY CANCER GENE IDENTIFICATION
CONSENT FORM**

I have read the letter of information, have had the nature of the study explained to me, and I agree to participate. All questions have been answered to my satisfaction.

At the conclusion of my genetic analysis, please contact me at either of the following numbers:

Home number: _____

Cell number: _____

If I am not available in the future when my research test results become available, please contact:

Full name: _____

Relationship to me: _____

Telephone contact number: (h) _____ (c) _____

I have received a signed copy of this consent form for my own records.

Subject's Signature Printed name Date

Signature of Printed name/Study Role Date
Person Obtaining Consent

Initials _____

Appendix K: Supplementary Information (Chapter 4)

Probe Design

Probes for 14 genes were designed as described in Mucaki *et al.*, (submitted): 22,191 sense strand oligonucleotides were synthesized, consisting of 2,828 *ATP8B1* (OMIM 602397; NM_005603.4, NP_005594.1), 2,332 *BARD1* (OMIM 601593; NM_000465.3, NP_000456.2), 786 *EPCAM* (OMIM 185535; NM_002354.2, NP_002345.2), 1,580 *MLH1* (OMIM 120436; NM_000249.3, NP_000240.1), 1,986 *MRE11A* (OMIM 600814; NM_005591.3, NP_005582.1), 3,355 *MSH2* (OMIM 609309; NM_000251.2, NP_000242.1), 723 *MSH6* (OMIM 600678; NM_000179.2, NP_000170.1), 626 *MUTYH* (OMIM 604933; NM_012222.2, NP_036354.1), 1,794 *NBN* (OMIM 602667; NM_002485.4, NP_002476.2), 206 *PMS2* (OMIM 600259; NM_000535.5, NP_000526.1), 2,599 *PTEN* (OMIM 601728; NM_000314.4, NP_000305.3), 1,764 *RAD51B* (OMIM 602948; NM_002877.5, NP_002868), 845 *STK11* (OMIM 602216; NM_000455.4, NP_000446.1), 568 *XRCC2* (OMIM 600375; NM_005431.1, NP_005422.1), and repeated probes from our original tiled capture array for *BRCA1* (50) and *BRCA2* (149) where improved coverage was desired. From the antisense strand, 21,967 oligos (2,757 *ATP8B1*, 2,260 *BARD1*, 800 *EPCAM*, 1,546 *MLH1*, 1,948 *MRE11A*, 3,254 *MSH2*, 743 *MSH6*, 619 *MUTYH*, 1,804 *NBN*, 293 *PMS2*, 2,608 *PTEN*, 1,701 *RAD51B*, 842 *STK11*, 585 *XRCC2*, and repeats of 44 *BRCA1* and 163 *BRCA2* probes) were synthesized on two additional custom microarray chips. Genic probes for *RAD51B* were limited to 1,000 nt surrounding each exon as it contains two > 100 kb introns, which would have taken an exorbitant amount of space on our capture arrays (31,342 *RAD51B* probes were initially designed).

However, the above-mentioned probes left repeat-free gaps. For repeat-free gaps of ≥ 10 nt (N=71), probes were designed manually for these regions (21 *ATP8B1*, 14 *BARD1*, 8 *EPCAM*, 3 *MLH1*, 3 *MRE11A*, 4 *MSH2*, 1 *MSH6*, 4 *NBN*, 6 *PMS2*, 1 *PTEN*, 4 *RAD51B*, and 2 *STK11*). *PMS2* had entire exons uncovered due to its homology to other regions in the genome. Therefore, this gene had the most handpicked probes (33 probes for 6 gaps).

When comparing the coverage of these “hand-picked” regions to those designed by Picky, we found they had an average of 6.9% lower coverage (up to a 22% deviation from average in one patient) than the average of all the other probes we designed (both those described above, and in Mucaki *et al.*, submitted). This result is comparable to the difference seen between PICKY- and manually-designed probes from Mucaki *et al.* (submitted).

Probes for *ATM* (OMIM 607585; NM_000051.3, NP_000042.3), *BRCA1* (OMIM 113705; NM_007294.3, NP_009225.1), *BRCA2* (OMIM 600185; NM_000059.3, NP_000050.2), *CDH1* (OMIM 192090; NM_004360.3, NP_004351.1), *CHEK2* (OMIM 604373; NM_145862.2, NP_665861.1), *PALB2* (OMIM 610355; NM_024675.3, NP_078951.2), and *TP53* (OMIM 191170; NM_000546.5, NP_000537.3) were identical to those described in Mucaki *et al.*, (submitted). The combination of all sets of tiled capture array oligos cover a total of 1,103,029 nt across the 21 genes.

Sample Preparation – Automation of Capture Pulldown

Post-capture samples were transferred to a standard 96-well plate after pull-down. Post pull-down bead wash steps were increased (twice with Wash 1 buffer, four times with Wash 2 buffer) as volumes were reduced to 190 μ L due to the reduced maximum volume. Elution and neutralization buffers were also reduced by 20% (80 μ L and 112 μ L, respectively). These changes did not seem to have any appreciable effect on capture efficiency. Samples were multiplex sequenced (48/run) on an Illumina Genome Analyzer IIx (GAIIx) using the standard Illumina indexing PCR primers (#2, 4, 5, 6, 7 and 12; chosen to maximize dissimilarity).

TFBS Information Models

As described in Mucaki *et al.*, (submitted), we used ChIP-seq data from ENCODE (ENCODE Project Consortium 2012) to identify TFs that bind our sequenced genes. In addition to the TFs for which we previously built models, we identified 8 additional factors due to the expansion of our target genes: TFIIB150 [*BDP1*], HD8 [*HDAC8*], Ikaros [*IKZF1*], PBX3 [*PBX3*], PGC-1- α [*PPARGC1A*], RPC155 [*POLR3A*], BRG1

[*SMARCA4*], and ZNF274 [*ZNF274*]. We successfully built models for TFIIB150 [*BDP1*], PBX3 [*PBX3*], and ZNF274 [*ZNF274*].

A model was not built for HD8 because it is a histone deacetylase. It is common for histone modifying proteins to result in ChIP-seq data, despite not binding DNA directly (Wang *et al.*, 2012). A model for BRG1 could not be generated, as the model building software would only reveal models for two of its known interacting partners, AP-1 and GATA-1 (Xu *et al.*, 2006; Baron *et al.*, 2007). Attempts to mask these sequences, in order to reveal the BRG1 motif, only resulted in noise models. Similarly, the computed models for the three remaining TFs (Ikaros, PGC-1- α , and RPC155) were noise models, differing completely from the expected motif, and thus could not be used.

Pedigree Analysis

Prior to analysis, pedigrees were de-identified while retaining sex, age, disease status and age of onset. We generated .ped files based on the family pedigrees provided by the LHSC. Very large pedigrees were truncated to contain the most immediate and informative individuals possible, due to restrictions in the software with respect to pedigree size (pedigrees with more than 32 individuals tends to cause the program to fail). If a patient's age was unknown, the average age for that generation was assigned (as recommended by Mohammadi *et al.*, 2009). If the age of onset for a family member with breast and/or ovarian cancer was not indicated, the mean age of diagnosis (based on Supplementary Table 2 of Mohammadi *et al.*, 2009) for breast and ovarian cancer in non-carriers were used (72 for breast cancer and 85 for ovarian cancer). Family members below the age of 18 were not included. In families where breast/ovarian cancer occurred through two different lineages, both options were assessed separately.

Supplementary Bibliography (applies to all supplementary tables, figures, and files)

Akiyama Y, Sato H, Yamada T, Nagasaki H, Tsuchiya A, Abe R, Yuasa Y. 1997. Germ-line mutation of the hMSH6/GTBP gene in an atypical hereditary nonpolyposis colorectal cancer kindred. *Cancer Res.* 57: 3920–3923.

Alessi DR, Sakamoto K, Bayascas JR. 2006. LKB1-dependent signaling pathways. *Annu. Rev. Biochem.* 75: 137–163.

Baron S, Escande A, Alb erola G, Bystricky K, Balaguer P, Richard-Foy H. 2007. Estrogen receptor alpha and the activating protein-1 complex cooperate during insulin-like growth factor-I-induced transcriptional activation of the pS2/TFF1 gene. *J. Biol. Chem.* 282: 11732–11741.

Bartek J, Lukas J. 2003. Chk1 and Chk2 kinases in checkpoint control and cancer. *Cancer Cell* 3: 421–429.

Becker-Catania SG, Chen G, Hwang MJ, Wang Z, Sun X, Sanal O, Bernatowska-Matuszkiewicz E, Chessa L, Lee EY, Gatti RA. 2000. Ataxia-telangiectasia: phenotype/genotype studies of ATM protein expression, mutations, and radiosensitivity. *Mol. Genet. Metab.* 70: 122–133.

Beck NE, Tomlinson IP, Homfray T, Hodgson SV, Harocopos CJ, Bodmer WF. 1997. Genetic testing is important in families with a history suggestive of hereditary non-polyposis colorectal cancer even if the Amsterdam criteria are not fulfilled. *Br. J. Surg.* 84: 233–237.

Blanco A, Hoya M de la, Osorio A, Diez O, Miramar MD, Infante M, Martinez-Bouzas C, Torres A, Lasa A, Llorca G, Brunet J, Graña B, et al., 2013. Analysis of PALB2 gene in BRCA1/BRCA2 negative Spanish hereditary breast/ovarian cancer families with pancreatic cancer cases. *PloS One* 8: e67538.

Borg A, Haile RW, Malone KE, Capanu M, Diep A, Torngren T, Teraoka S, Begg CB, Thomas DC, Concannon P, Møller M, Bernstein L, et al., 2010. Characterization of

BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum. Mutat.* 31: E1200–40.

Bouwman P, Gulden H van der, Heijden I van der, Drost R, Klijn CN, Prasetyanti P, Pieterse M, Wientjens E, Seibler J, Hogervorst FBL, Jonkers J. 2013. A high-throughput functional complementation assay for classification of BRCA1 missense variants. *Cancer Discov.* 3: 1142–1155.

Bronner CE, Baker SM, Morrison PT, Warren G, Smith LG, Lescoe MK, Kane M, Earabino C, Lipford J, Lindblom A. 1994. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* 368: 258–261.

Brown KD, Rathi A, Kamath R, Beardsley DI, Zhan Q, Mannino JL, Baskaran R. 2003. The mismatch repair system is required for S-phase checkpoint activation. *Nat. Genet.* 33: 80–84.

Carney EF, Srinivasan V, Moss PA, Taylor AM. 2012. Classical Ataxia Telangiectasia Patients Have a Congenitally Aged Immune System with High Expression of CD95. *J. Immunol.* 189: 261–268.

Casadei S, Norquist BM, Walsh T, Stray S, Mandell JB, Lee MK, Stamatoyannopoulos JA, King M-C. 2011. Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. *Cancer Res.* 71: 2222–2229.

Castillejo A, Vargas G, Castillejo MI, Navarro M, Barberá VM, González S, Hernández-Illán E, Brunet J, Ramón y Cajal T, Balmaña J, Oltra S, Iglesias S, et al., 2014. Prevalence of germline MUTYH mutations among Lynch-like syndrome patients. *Eur. J. Cancer Oxf. Engl.* 1990 50: 2241–2250.

Catucci I, Peterlongo P, Ciceri S, Colombo M, Pasquini G, Barile M, Bonanni B, Verderio P, Pizzamiglio S, Foglia C, Falanga A, Marchetti M, et al., 2014. PALB2 sequencing in Italian familial breast cancer cases reveals a high-risk mutation recurrent in the province of Bergamo. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 16: 688–694.

Cavalieri S, Funaro A, Porcedda P, Turinetto V, Migone N, Gatti RA, Brusco A. 2006. ATM mutations in Italian families with ataxia telangiectasia include two distinct large genomic deletions. *Hum. Mutat.* 27: 1061.

Chenevix-Trench G, Healey S, Lakhani S, Waring P, Cummings M, Brinkworth R, Deffenbaugh AM, Burbidge LA, Pruss D, Judkins T, Scholl T, Bekessy A, et al., 2006. Genetic and Histopathologic Evaluation of BRCA1 and BRCA2 DNA Sequence Variants of Unknown Clinical Significance. *Cancer Res.* 66: 2019–2027.

Chenevix-Trench G, Spurdle AB, Gatei M, Kelly H, Marsh A, Chen X, Donn K, Cummings M, Nyholt D, Jenkins MA, Scott C, Pupo GM, et al., 2002. Dominant negative ATM mutations in breast cancer families. *J. Natl. Cancer Inst.* 94: 205–215.

Chen W, Yurong S, Liansheng N. 2008. Breast cancer low-penetrance allele 1100delC in the CHEK2 gene: not present in the Chinese familial breast cancer population. *Adv. Ther.* 25: 496–501.

Cyr JL, Heinen CD. 2008. Hereditary cancer-associated missense mutations in hMSH6 uncouple ATP hydrolysis from DNA mismatch binding. *J. Biol. Chem.* 283: 31641–31648.

Deng C-X. 2006. BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic Acids Res.* 34: 1416–1426.

Desrichard A, Bidet Y, Uhrhammer N, Bignon Y-J. 2011. CHEK2 contribution to hereditary breast cancer in non-BRCA families. *Breast Cancer Res. BCR* 13: R119.

Devlin LA, Graham CA, Price JH, Morrison PJ. 2008. Germline MSH6 mutations are more prevalent in endometrial cancer patient cohorts than hereditary non polyposis colorectal cancer cohorts. *Ulster Med. J.* 77: 25–30.

Dong J, Chang-Claude J, Wu Y, Schumacher V, Debatin I, Tonin P, Royer-Pokora B. 1998. A high proportion of mutations in the BRCA1 gene in German breast/ovarian cancer families with clustering of mutations in the 3' third of the gene. *Hum. Genet.* 103: 154–161.

Dörk T, Bendix R, Bremer M, Rades D, Klöpper K, Nicke M, Skawran B, Hector A, Yamini P, Steinmann D, Weise S, Stuhmann M, et al., 2001. Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients. *Cancer Res.* 61: 7608–7615.

Drost M, Koppejan H, Wind N de. 2013. Inactivation of DNA mismatch repair by variants of uncertain significance in the PMS2 gene. *Hum. Mutat.* 34: 1477–1480.

Drost M, Zonneveld JBM, Hees S van, Rasmussen LJ, Hofstra RMW, Wind N de. 2012. A rapid and cell-free assay to test the activity of lynch syndrome-associated MSH2 and MSH6 missense variants. *Hum. Mutat.* 33: 488–494.

Durocher F, Shattuck-Eidens D, McClure M, Labrie F, Skolnick MH, Goldgar DE, Simard J. 1996. Comparison of BRCA1 polymorphisms, rare sequence variants and/or missense mutations in unaffected and breast/ovarian cancer populations. *Hum. Mol. Genet.* 5: 835–842.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.

Fletcher O, Johnson N, Santos Silva I dos, Orr N, Ashworth A, Nevanlinna H, Heikkinen T, Aittomäki K, Blomqvist C, Burwinkel B, Bartram CR, Meindl A, et al., 2010. Missense variants in ATM in 26,101 breast cancer cases and 29,842 controls. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* 19: 2143–2151.

Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, et al., 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* 62: 676–689.

Giardiello FM, Brensinger JD, Tersmette AC, Goodman SN, Petersen GM, Booker SV, Cruz-Correa M, Offerhaus JA. 2000. Very high risk of cancer in familial Peutz-Jeghers syndrome. *Gastroenterology* 119: 1447–1453.

Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro ANA, Tavtigian SV, Couch FJ, Breast Cancer Information Core (BIC) Steering Committee. 2004. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am. J. Hum. Genet.* 75: 535–544.

Golmard L, Caux-Moncoutier V, Davy G, Ageeli E Al, Poirot B, Tirapo C, Michaux D, Barbaroux C, Enghien CD d', Nicolas A, Castéra L, Sastre-Garau X, et al., 2013. Germline mutation in the RAD51B gene confers predisposition to breast cancer. *BMC Cancer* 13: 484.

Guarinos C, Castillejo A, Barberá V-M, Pérez-Carbonell L, Sánchez-Heras A-B, Segura Á, Guillén-Ponce C, Martínez-Cantó A, Castillejo M-I, Egoavil C-M, Jover R, Payá A, et al., 2010. EPCAM Germ Line Deletions as Causes of Lynch Syndrome in Spanish Patients. *J. Mol. Diagn. JMD* 12: 765–770.

Gudmundsdottir K, Ashworth A. 2006. The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* 25: 5864–5874.

Gutiérrez-Enríquez S, Fernet M, Dörk T, Bremer M, Lauge A, Stoppa-Lyonnet D, Moullan N, Angèle S, Hall J. 2004. Functional consequences of ATM sequence variants for chromosomal radiosensitivity. *Genes. Chromosomes Cancer* 40: 109–119.

Hegde M, Blazo M, Chong B, Prior T, Richards C. 2005. Assay validation for identification of hereditary nonpolyposis colon cancer-causing mutations in mismatch repair genes MLH1, MSH2, and MSH6. *J. Mol. Diagn. JMD* 7: 525–534.

Heikkinen K, Rapakko K, Karppinen S-M, Erkko H, Nieminen P, Winqvist R. 2005. Association of common ATM polymorphism with bilateral breast cancer. *Int. J. Cancer* 116: 69–72.

Hellebrand H, Sutter C, Honisch E, Gross E, Wappenschmidt B, Schem C, Deissler H, Ditsch N, Gress V, Kiechle M, Bartram CR, Schmutzler RK, et al., 2011. Germline mutations in the PALB2 gene are population specific and occur with low frequencies in familial breast cancer. *Hum. Mutat.* 32: E2176–2188.

Hisada M, Garber JE, Fung CY, Fraumeni JF, Li FP. 1998. Multiple primary cancers in families with Li-Fraumeni syndrome. *J. Natl. Cancer Inst.* 90: 606–611.

Howlett NG, Taniguchi T, Olson S, Cox B, Waisfisz Q, Die-Smulders C De, Persky N, Grompe M, Joenje H, Pals G, Ikeda H, Fox EA, et al., 2002. Biallelic inactivation of BRCA2 in Fanconi anemia. *Science* 297: 606–609.

Ikonen T, Matikainen M, Mononen N, Hyytinen ER, Helin HJ, Tammela TL, Pukkala E, Schleutker J, Kallioniemi OP, Koivisto PA. 2001. Association of E-cadherin germ-line alterations with prostate cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 7: 3465–3471.

Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, Teer JK, Mullikin JC, Biesecker LG. 2012. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *Am. J. Hum. Genet.* 91: 97–108.

Jonsson B-A, Bergh A, Stattin P, Emanuelsson M, Grönberg H. 2002. Germline mutations in E-cadherin do not explain association of hereditary prostate cancer, gastric cancer and breast cancer. *Int. J. Cancer J. Int. Cancer* 98: 838–843.

Keller G, Vogelsang H, Becker I, Plaschke S, Ott K, Suriano G, Mateus AR, Seruca R, Biedermann K, Huntsman D, Döring C, Holinski-Feder E, et al., 2004. Germline mutations of the E-cadherin(CDH1) and TP53 genes, rather than of RUNX3 and HPP1, contribute to genetic predisposition in German gastric cancer patients. *J. Med. Genet.* 41: e89.

Kolodner RD, Marsischky GT. 1999. Eukaryotic DNA mismatch repair. *Curr. Opin. Genet. Dev.* 9: 89–96.

Larson GP, Zhang G, Ding S, Foldenauer K, Udar N, Gatti RA, Neuberg D, Lunetta KL, Ruckdeschel JC, Longmate J, Flanagan S, Krontiris TG. 1997. An allelic variant at the ATM locus is implicated in breast cancer susceptibility. *Genet. Test.* 1: 165–170.

Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomäki P, Sistonen P, Aaltonen LA, Nyström-Lahti M. 1993. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* 75: 1215–1225.

Lee PS, Fang J, Jessop L, Myers T, Raj P, Hu N, Wang C, Taylor PR, Wang J, Khan J, Jasin M, Chanock SJ. 2014. RAD51B Activity and Cell Cycle Regulation in Response to DNA Damage in Breast Cancer Cell Lines. *Breast Cancer Basic Clin. Res.* 8: 135–144.

Lee SB, Kim SH, Bell DW, Wahrer DC, Schiripo TA, Jorczak MM, Sgroi DC, Garber JE, Li FP, Nichols KE, Varley JM, Godwin AK, et al., 2001. Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome. *Cancer Res.* 61: 8062–8067.

Li A, Swift M. 2000. Mutations at the ataxia-telangiectasia locus and clinical phenotypes of A-T patients. *Am. J. Med. Genet.* 92: 170–177.

Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, Zheng Z, Bose S, Call KM, Tsou HC, Peacocke M, Eng C, Parsons R. 1997. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat. Genet.* 16: 64–67.

Ligtenberg MJL, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TYH, Bodmer D, Hoenselaar E, Hendriks-Cornelissen SJB, Tsui WY, Kong CK, et al., 2009. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nat. Genet.* 41: 112–117.

Li M, Zhang P. 2009. The function of APC/CCdh1 in cell cycle and beyond. *Cell Div.* 4: 2.

Liu C, Srihari S, Cao K-AL, Chenevix-Trench G, Simpson PT, Ragan MA, Khanna KK. 2014. A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. *Nucleic Acids Res.* 42: 6106–6127.

Luo Y, Lin F-T, Lin W-C. 2004. ATM-mediated stabilization of hMutL DNA mismatch repair proteins augments p53 activation during DNA damage. *Mol. Cell. Biol.* 24: 6430–6444.

Lynch HT, Lynch JF, Snyder CL, Riegert-Johnson D. 2011. EPCAM deletions, Lynch syndrome, and cancer risk. *Lancet Oncol.* 12: 5–6.

Lynch HT, Smyrk TC, Watson P, Lanspa SJ, Lynch JF, Lynch PM, Cavalieri RJ, Boland CR. 1993. Genetics, natural history, tumor spectrum, and pathology of hereditary nonpolyposis colorectal cancer: an updated review. *Gastroenterology* 104: 1535–1549.

Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. 2015. Milestones of Lynch syndrome: 1895-2015. *Nat. Rev. Cancer* 15: 181–194.

Magliozzi M, Piane M, Torrente I, Sinibaldi L, Rizzo G, Savio C, Lulli P, Luca A De, Dallapiccola B, Chessa L. 2006. DHPLC screening of ATM gene in Italian patients affected by ataxia-telangiectasia: fourteen novel ATM mutations. *Dis. Markers* 22: 257–264.

Marinovic-Terzic I, Yoshioka-Yamashita A, Shimodaira H, Avdievich E, Hunton IC, Kolodner RD, Edelman W, Wang JYJ. 2008. Apoptotic function of human PMS2 compromised by the nonsynonymous single-nucleotide polymorphic variant R20Q. *Proc. Natl. Acad. Sci. U. S. A.* 105: 13993–13998.

Masciari S, Dillon DA, Rath M, Robson M, Weitzel JN, Balmana J, Gruber SB, Ford JM, Euhus D, Lebensohn A, Telli M, Pochebit SM, et al., 2012. Breast cancer phenotype in women with TP53 germline mutations: a Li-Fraumeni syndrome consortium effort. *Breast Cancer Res. Treat.* 133: 1125–1130.

Mauget-Faysse M, Vuillaume M, Quaranta M, Moullan N, Angèle S, Friesen MD, Hall J. 2003. Idiopathic and radiation-induced ocular telangiectasia: the involvement of the ATM gene. *Invest. Ophthalmol. Vis. Sci.* 44: 3257–3262.

Mazoyer S, Dunning AM, Serova O, Dearden J, Puget N, Healey CS, Gayther SA, Mangion J, Stratton MR, Lynch HT, Goldgar DE, Ponder BA, et al., 1996. A polymorphic stop codon in BRCA2. *Nat. Genet.* 14: 253–254.

McKinnon PJ. 2004. ATM and ataxia telangiectasia. *EMBO Rep.* 5: 772–776.

Menon V, Povirk L. 2014. Involvement of p53 in the repair of DNA double strand breaks: multifaceted Roles of p53 in homologous recombination repair (HRR) and non-homologous end joining (NHEJ). *Subcell. Biochem.* 85: 321–336.

Menzel T, Nähse-Kumpf V, Kousholt AN, Klein DK, Lund-Andersen C, Lees M, Johansen JV, Syljuåsen RG, Sørensen CS. 2011. A genetic screen identifies BRCA2 and PALB2 as key regulators of G2 checkpoint maintenance. *EMBO Rep.* 12: 705–712.

Miyaki M, Konishi M, Tanaka K, Kikuchi-Yanoshita R, Muraoka M, Yasuno M, Igari T, Koike M, Chiba M, Mori T. 1997. Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nat. Genet.* 17: 271–272.

Mohammadi L, Vreeswijk MP, Oldenburg R, Ouweland A van den, Oosterwijk JC, Hout AH van der, Hoogerbrugge N, Ligtenberg M, Ausems MG, Luijt RB van der, Dommering CJ, Gille JJ, et al., 2009. A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example. *BMC Cancer* 9: 211.

Mosor M, Ziółkowska I, Pernak-Schwarz M, Januszkiewicz-Lewandowska D, Nowak J. 2006. Association of the heterozygous germline I171V mutation of the NBS1 gene with childhood acute lymphoblastic leukemia. *Leukemia* 20: 1454–1456.

Newman B, Austin MA, Lee M, King MC. 1988. Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc. Natl. Acad. Sci. U. S. A.* 85: 3044–3048.

Nicolaides NC, Papadopoulos N, Liu B, Wei YF, Carter KC, Ruben SM, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM. 1994. Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature* 371: 75–80.

Novak DJ, Chen LQ, Ghadirian P, Hamel N, Zhang P, Rossiny V, Cardinal G, Robidoux A, Tonin PN, Rousseau F, Narod SA, Foulkes WD. 2008. Identification of a novel CHEK2 variant and assessment of its contribution to the risk of breast cancer in French Canadian women. *BMC Cancer* 8: 239.

Olson E, Nievera CJ, Liu E, Lee AY-L, Chen L, Wu X. 2007. The Mre11 complex mediates the S-phase checkpoint through an interaction with replication protein A. *Mol. Cell. Biol.* 27: 6053–6067.

Paglia LL, Laugé A, Weber J, Champ J, Cavaciuti E, Russo A, Viovy J-L, Stoppa-Lyonnet D. 2010. ATM germline mutations in women with familial breast cancer and a relative with haematological malignancy. *Breast Cancer Res. Treat.* 119: 443–452.

Panguluri RC, Brody LC, Modali R, Utley K, Adams-Campbell L, Day AA, Whitfield-Broome C, Dunston GM. 1999. BRCA1 mutations in African Americans. *Hum. Genet.* 105: 28–31.

Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD. 1994. Mutation of a mutL homolog in hereditary colon cancer. *Science* 263: 1625–1629.

Paramio JM, Navarro M, Segrelles C, Gómez-Casero E, Jorcano JL. 1999. PTEN tumour suppressor is linked to the cell cycle control through the retinoblastoma protein. *Oncogene* 18: 7462–7468.

Petereit DG. 2013. Prevalence of ATM sequence variants in Northern plains American Indian cancer patients. *Radiat. Oncol.* 3: 318.

Petrini JHJ, Stracker TH. 2003. The cellular response to DNA double-strand breaks: defining the sensors and mediators. *Trends Cell Biol.* 13: 458–462.

Pizarro JG, Folch J, Torre AV de la, Junyent F, Verdaguer E, Jordan J, Pallas M, Camins A. 2010. ATM is involved in cell-cycle control through the regulation of retinoblastoma protein phosphorylation. *J. Cell. Biochem.* 110: 210–218.

Quesnel S, Verselis S, Portwine C, Garber J, White M, Feunteun J, Malkin D, Li FP. 1999. p53 compound heterozygosity in a severely affected child with Li-Fraumeni syndrome. *Oncogene* 18: 3970–3978.

Reid S, Schindler D, Hanenberg H, Barker K, Hanks S, Kalb R, Neveling K, Kelly P, Seal S, Freund M, Wurm M, Batish SD, et al., 2007. Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat. Genet.* 39: 162–164.

Reiman A, Srinivasan V, Barone G, Last JJ, Wootton LL, Davies EG, Verhagen MM, Willemsen MA, Weemaes CM, Byrd PJ, Izatt L, Easton DF, et al., 2011. Lymphoid tumours and breast cancer in ataxia telangiectasia; substantial protective effect of residual ATM kinase activity against childhood tumours. *Br. J. Cancer* 105: 586–591.

Resta N, Pierannunzio D, Lenato GM, Stella A, Capocaccia R, Bagnulo R, Lastella P, Susca FC, Bozzao C, Loconte DC, Sabbà C, Urso E, et al., 2013. Cancer risk associated with STK11/LKB1 germline mutations in Peutz-Jeghers syndrome patients: results of an Italian multicenter study. *Dig. Liver Dis. Off. J. Ital. Soc. Gastroenterol. Ital. Assoc. Study Liver* 45: 606–611.

Rossi BM, Lopes A, Oliveira Ferreira F, Nakagawa WT, Napoli Ferreira CC, Casali Da Rocha JC, Simpson CC, Simpson AJG. 2002. hMLH1 and hMSH2 gene mutation in Brazilian families with suspected hereditary nonpolyposis colorectal cancer. *Ann. Surg. Oncol.* 9: 555–561.

Rudd MF, Webb EL, Matakidou A, Sellick GS, Williams RD, Bridle H, Eisen T, Houlston RS, GELCAPS Consortium. 2006. Variants in the GH-IGF axis confer susceptibility to lung cancer. *Genome Res.* 16: 693–701.

Sandoval N, Platzer M, Rosenthal A, Dörk T, Bendix R, Skawran B, Stuhmann M, Wegner RD, Sperling K, Banin S, Shiloh Y, Baumer A, et al., 1999. Characterization of ATM gene mutations in 66 ataxia telangiectasia families. *Hum. Mol. Genet.* 8: 69–79.

Sanjosé S de, Léoné M, Bérez V, Izquierdo A, Font R, Brunet JM, Louat T, Vilardell L, Borrás J, Viladiu P, Bosch FX, Lenoir GM, et al., 2003. Prevalence of BRCA1 and BRCA2 germline mutations in young breast cancer patients: a population-based study. *Int. J. Cancer J. Int. Cancer* 106: 588–593.

Santos C, Peixoto A, Rocha P, Pinto P, Bizarro S, Pinheiro M, Pinto C, Henrique R, Teixeira MR. 2014. Pathogenicity Evaluation of BRCA1 and BRCA2 Unclassified Variants Identified in Portuguese Breast/Ovarian Cancer Families. *J. Mol. Diagn.* 16: 324–334.

Sanz DJ, Acedo A, Infante M, Durán M, Pérez-Cabornero L, Esteban-Cardena E, Lastra E, Pagani F, Miner C, Velasco EA. 2010. A High Proportion of DNA Variants of BRCA1 and BRCA2 Is Associated with Aberrant Splicing in Breast/Ovarian Cancer Patients. *Clin. Cancer Res.* 16: 1957–1967.

Sarrió D, Moreno-Bueno G, Hardisson D, Sánchez-Estévez C, Guo M, Herman JG, Gamallo C, Esteller M, Palacios J. 2003. Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability. *Int. J. Cancer* 106: 208–215.

Sauer MK, Andrulis IL. 2005. Identification and characterization of missense alterations in the BRCA1 associated RING domain (BARD1) gene in breast and ovarian cancer. *J. Med. Genet.* 42: 633–638.

Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, Lindblom A, Lagerstedt K, Thibodeau SN, Lindor NM, Young J, Winship I, et al., 2008. The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. *Gastroenterology* 135: 419–428.

Shamseldin HE, Elfaki M, Alkuraya FS. 2012. Exome sequencing reveals a novel Fanconi group defined by XRCC2 mutation. *J. Med. Genet.* 49: 184–186.

Shaw PH. 1996. The role of p53 in cell cycle regulation. *Pathol. Res. Pract.* 192: 669–675.

Shimodaira H, Yoshioka-Yamashita A, Kolodner RD, Wang JYJ. 2003. Interaction of mismatch repair protein PMS2 and the p53-related transcription factor p73 in apoptosis response to cisplatin. *Proc. Natl. Acad. Sci. U. S. A.* 100: 2420–2425.

Sinilnikova OM, Egan KM, Quinn JL, Boutrand L, Lenoir GM, Stoppa-Lyonnet D, Desjardins L, Levy C, Goldgar D, Gragoudas ES. 1999. Germline brca2 sequence variants in patients with ocular melanoma. *Int. J. Cancer J. Int. Cancer* 82: 325–328.

Slupska MM, Luther WM, Chiang J-H, Yang H, Miller JH. 1999. Functional Expression of hMYH, a Human Homolog of the Escherichia coli MutY Protein. *J. Bacteriol.* 181: 6210–6213.

Sommer SS, Jiang Z, Feng J, Buzin CH, Zheng J, Longmate J, Jung M, Moulds J, Dritschilo A. 2003. ATM missense mutations are frequent in patients with breast cancer. *Cancer Genet. Cytogenet.* 145: 115–120.

Song MS, Salmena L, Pandolfi PP. 2012. The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.* 13: 283–296.

Spurdle AB, Lakhani SR, Healey S, Parry S, Silva LM Da, Brinkworth R, Hopper JL, Brown MA, Babikyan D, Chenevix-Trench G, Tavtigian SV, Goldgar DE, et al., 2008. Clinical classification of BRCA1 and BRCA2 DNA sequence variants: the value of cytokeratin profiles and evolutionary analysis--a report from the kConFab Investigators. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 26: 1657–1663.

Stankovic T, Kidd AM, Sutcliffe A, McGuire GM, Robinson P, Weber P, Bedenham T, Bradwell AR, Easton DF, Lennox GG, Haites N, Byrd PJ, et al., 1998. ATM mutations and phenotypes in ataxia-telangiectasia families in the British Isles: expression of mutant ATM and the risk of leukemia, lymphoma, and breast cancer. *Am. J. Hum. Genet.* 62: 334–345.

Steffen J, Varon R, Mosor M, Maneva G, Maurer M, Stumm M, Nowakowska D, Rubach M, Kosakowska E, Ruka W, Nowecki Z, Rutkowski P, et al., 2004. Increased cancer risk of heterozygotes with NBS1 germline mutations in Poland. *Int. J. Cancer J. Int. Cancer* 111: 67–71.

Stewart GS, Maser RS, Stankovic T, Bressan DA, Kaplan MI, Jaspers NGJ, Raams A, Byrd PJ, Petrini JHJ, Taylor AMR. 1999. The DNA Double-Strand Break Repair Gene

hMRE11 Is Mutated in Individuals with an Ataxia-Telangiectasia-like Disorder. *Cell* 99: 577–587.

Suchy J, Kurzawski G, Jakubowska K, Rać ME, Safranow K, Kładny J, Rzepka-Górska I, Chosia M, Czeszyńska B, Oszurek O, Scott RJ, Lubiński J. 2006. Frequency and nature of hMSH6 germline mutations in Polish patients with colorectal, endometrial and ovarian cancers. *Clin. Genet.* 70: 68–70.

Sudo T, Ota Y, Kotani S, Nakao M, Takami Y, Takeda S, Saya H. 2001. Activation of Cdh1-dependent APC is required for G1 cell cycle arrest and DNA damage-induced G2 checkpoint in vertebrate cells. *EMBO J.* 20: 6499–6508.

Suwaki N, Klare K, Tarsounas M. 2011. RAD51 paralogs: roles in DNA damage signalling, recombinational repair and tumorigenesis. *Semin. Cell Dev. Biol.* 22: 898–905.

Takahashi M, Shimodaira H, Andreutti-Zaugg C, Iggo R, Kolodner RD, Ishioka C. 2007. Functional analysis of human MLH1 variants using yeast and in vitro mismatch repair assays. *Cancer Res.* 67: 4595–4604.

Tan M-H, Mester J, Peterson C, Yang Y, Chen J-L, Rybicki LA, Milas K, Pederson H, Remzi B, Orloff MS, Eng C. 2011. A clinical scoring system for selection of patients for PTEN mutation testing is proposed on the basis of a prospective study of 3042 probands. *Am. J. Hum. Genet.* 88: 42–56.

Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Calvez-Kelm F Le, Lesueur F, Byrnes GB, Chuang S-C, Forey N, Feuchtinger C, Gioia L, et al., 2009. Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am. J. Hum. Genet.* 85: 427–446.

Taylor CF, Charlton RS, Burn J, Sheridan E, Taylor GR. 2003. Genomic deletions in MSH2 or MLH1 are a frequent cause of hereditary non-polyposis colorectal cancer: identification of novel and recurrent deletions by MLPA. *Hum. Mutat.* 22: 428–433.

Tikhomirova OS, Grudinina NA, Golubkov VI, Mandel'shtam MI, Vasil'ev BV. 2007. [Novel BRCA1 gene mutations in breast cancer patients from St. Petersburg]. *Genetika* 43: 1263–1268.

Tischkowitz M, Capanu M, Sabbaghian N, Li L, Liang X, Vallée MP, Tavtigian SV, Concannon P, Foulkes WD, Bernstein L, WECARE Study Collaborative Group, Bernstein JL, et al., 2012. Rare germline mutations in PALB2 and breast cancer risk: a population-based study. *Hum. Mutat.* 33: 674–680.

Tischkowitz MD, Sabbaghian N, Hamel N, Borgida A, Rosner C, Taherian N, Srivastava A, Holter S, Rothenmund H, Ghadirian P, Foulkes WD, Gallinger S. 2009. Analysis of the gene coding for the BRCA2-interacting protein PALB2 in familial and sporadic pancreatic cancer. *Gastroenterology* 137: 1183–1186.

Tommiska J, Jansen L, Kilpivaara O, Edvardsen H, Kristensen V, Tamminen A, Aittomäki K, Blomqvist C, Børresen-Dale A-L, Nevanlinna H. 2006. ATM variants and cancer risk in breast cancer patients from Southern Finland. *BMC Cancer* 6: 209.

Tournier I, Vezain M, Martins A, Charbonnier F, Baert-Desurmont S, Olschwang S, Wang Q, Buisine MP, Soret J, Tazi J, Frébourg T, Tosi M. 2008. A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum. Mutat.* 29: 1412–1424.

Tutlewska K, Lubinski J, Kurzawski G. 2013. Germline deletions in the EPCAM gene as a cause of Lynch syndrome – literature review. *Hered. Cancer Clin. Pract.* 11: 9.

Ui A, Ogiwara H, Nakajima S, Kanno S, Watanabe R, Harata M, Okayama H, Harris CC, Yokota J, Yasui A, Kohno T. 2014. Possible involvement of LKB1-AMPK signaling in non-homologous end joining. *Oncogene* 33: 1640–1648.

Vandenberg CJ, Gergely F, Ong CY, Pace P, Mallery DL, Hiom K, Patel KJ. 2003. BRCA1-independent ubiquitination of FANCD2. *Mol. Cell* 12: 247–254.

Varon R, Vissinga C, Platzer M, Cerosaletti KM, Chrzanowska KH, Saar K, Beckmann G, Seemanová E, Cooper PR, Nowak NJ, Stumm M, Weemaes CM, et al., 1998. Nibrin,

a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome. *Cell* 93: 467–476.

Wagner A, Barrows A, Wijnen JT, Klift H van der, Franken PF, Verkuijlen P, Nakagawa H, Geugien M, Jaghmohan-Changur S, Breukel C, Meijers-Heijboer H, Morreau H, et al., 2003. Molecular analysis of hereditary nonpolyposis colorectal cancer in the United States: high mutation detection rate among clinically selected families and characterization of an American founder genomic deletion of the MSH2 gene. *Am. J. Hum. Genet.* 72: 1088–1100.

Walsh T, King M-C. 2007. Ten Genes for Inherited Breast Cancer. *Cancer Cell* 11: 103–105.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, et al., 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22: 1798–1812.

Wang Q, Zhang H, Guerrette S, Chen J, Mazurek A, Wilson T, Slupianek A, Skorski T, Fishel R, Greene MI. 2001. Adenosine nucleotide modulates the physical interaction between hMSH2 and BRCA1. *Oncogene* 20: 4640–4649.

Westermarck UK, Reyngold M, Olshen AB, Baer R, Jasin M, Moynahan ME. 2003. BARD1 participates with BRCA1 in homology-directed repair of chromosome breaks. *Mol. Cell. Biol.* 23: 7926–7936.

Wong MW, Nordfors C, Mossman D, Pecenpetelovska G, Avery-Kiejda KA, Talseth-Palmer B, Bowden NA, Scott RJ. 2011. BRIP1, PALB2, and RAD51C mutation analysis reveals their relative importance as genetic susceptibility factors for breast cancer. *Breast Cancer Res. Treat.* 127: 853–859.

Xu Z, Meng X, Cai Y, Koury MJ, Brandt SJ. 2006. Recruitment of the SWI/SNF protein Brg1 by a multiprotein complex effects transcriptional repression in murine erythroid progenitors. *Biochem. J.* 399: 297–304.

Yang G, Scherer SJ, Shell SS, Yang K, Kim M, Lipkin M, Kucherlapati R, Kolodner RD, Edelman W. 2004. Dominant effects of an Msh6 missense mutation on DNA repair and cancer susceptibility. *Cancer Cell* 6: 139–150.

Yoshida K, Miki Y. 2004. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci.* 95: 866–871.

Yuan ZQ, Gottlieb B, Beitel LK, Wong N, Gordon PH, Wang Q, Puisieux A, Foulkes WD, Trifiro M. 2002. Polymorphisms and HNPCC: PMS2-MLH1 protein interactions diminished by single nucleotide polymorphisms. *Hum. Mutat.* 19: 108–113.

Zhang F, Ma J, Wu J, Ye L, Cai H, Xia B, Yu X. 2009. PALB2 links BRCA1 and BRCA2 in the DNA-damage response. *Curr. Biol.* CB 19: 524–529.

Zhang H, Richards B, Wilson T, Lloyd M, Cranston A, Thorburn A, Fishel R, Meuth M. 1999. Apoptosis induced by overexpression of hMSH2 or hMLH1. *Cancer Res.* 59: 3021–3027.

Zhang J, Willers H, Feng Z, Ghosh JC, Kim S, Weaver DT, Chung JH, Powell SN, Xia F. 2004. Chk2 phosphorylation of BRCA1 regulates DNA double-strand break repair. *Mol. Cell. Biol.* 24: 708–718.

Curriculum Vitae

Name: Natasha Grace Caminsky

Post-secondary Education and Degrees: Dawson College
Montreal, Quebec, Canada
2008-2010 DEC (Diplôme d'Éducation Collégial – First Choice Health Science)

The University of Western Ontario
London, Ontario, Canada
2010-2014 BSc. (HSp Genetics, Major Physiology)

The University of Western Ontario
London, Ontario, Canada
2014-2015 MSc. (Biochemistry)

Honors and Awards:

09/2010 Western Scholarship of Distinction (\$1,000), Western University

09/2010 OCE Connections Program, Ontario Centres for Excellence (\$3,230).
Dorman, S.N., Shirley, B.C., Patrick, J.C., Caminsky, N.G., Knoll, J.H., Rogan, P.K. "DNA probe design for FISH and microarray analysis"

09/2011 James G. Farmer Award (\$500), Western University

09/2012 Gordon Risk Bursary (\$1,000), Western University

2010-2014 Scholar's Athlete Award

05/2014 The London CIHR Strategic Training Program in Cancer Research and Technology Transfer (CaRTT) Summer Studentship (\$6,000)

06/2014 Hon. G. Howard Ferguson Award, Western University

09/2014 CIHR Strategic Training Program in Cancer Research and Technology Transfer Trainee Award (\$18,100)

09/2014 Translational Breast Cancer Research Unit Trainee Award (\$17,000)

11/2014 Rhodes Scholarship Finalist (Québec Region)

Related Work Experience:

Position: Laboratory Assistant

Institution: Sir Mortimer B. David Jewish General Hospital

Dates: 2008-2010

Position: Biochemistry Undergraduate Summer Research Program

Institution: Schulich School of Medicine and Dentistry, The University of Western Ontario

Dates: 2011

Position: Research Assistant

Institution: The University of Western Ontario

Dates: 2012-2013

Publications:

Peer-Reviewed Articles

1. Rogan PK, Li Y, Wickramasinghe A, Subasinghe A, **Caminsky N**, Khan W, Samarabandu J, Wilkins R, Flegal F, Knoll JH. Automating dicentric chromosome detection from cytogenetic biodosimetry data. *Radiation Protection Dosimetry* 2014; 159:95-104.
2. **Caminsky NG**, Mucaki EJ and Rogan PK. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis [v2; ref status: indexed, <http://f1000r.es/54y>] *F1000 Research* 2015, 3:282.
3. Eliseos JM, Caminsky NG, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JHM, Rogan PK. A unified framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. Submitted to *BMC Medical Genomics* on August 11, 2015.
4. Eliseos JM, Caminsky NG, Perri AM, Knoll JH, Rogan PK. Prioritizing variants in complete Hereditary Breast and Ovarian Cancer (HBOC) genes in patients lacking known BRCA mutations. In preparation for submission to *Human Mutation*.

Oral/Poster Presentations and Other Scientific Contributions

1. Dorman, S.N., Shirley, B.C., **Caminsky, N.G.**, Rogan, P.K. “Developing single copy probes for fluorescence in-situ hybridization and array comparative genomic hybridization microarray design” Presented at the *6th Annual Canadian Student Conference on Biomedical Computing and Engineering* (London, Canada) May 2011.
2. Dorman, S.N., **Caminsky, N.G.**, Shirley, B.C., Khan, W.A., Guo, L., Knoll, J.H.M., Rogan P.K. “Development of single copy FISH probes to detect chromosomal abnormalities in small tumour suppressor and oncogenes” Presented at the *CIHR – Strategic Training Program in Cancer Research & Technology Transfer and the Department of Oncology Research & Education Day* (London, Canada) June 2011.
3. Dorman, S.N., Shirley, B.C., **Caminsky, N.G.**, Mucaki, E.J., Khan, W.A., Guo, L., Knoll, J.H.M., Rogan, P.K. “Next generation genomic microarrays and custom FISH probes for molecular cytogenetic analysis designed by ab initio sequence analysis” Presented at the *12th International Congress of Human Genetics/ASHG 61st Annual Meeting* (Montreal, Canada) October 2011 and the *London Health Research Day* (London, Canada) April 2012.
4. Dorman, S.N., Shirley, B.C., **Caminsky, N.G.**, Knoll, J.H.M., Rogan, P.K. “Expanding probe sequence repertoire and reproducibility in FISH and aCGH” Presented at the *Medical Genetics Rounds, London Health Sciences Centre* (London, Canada) 2012.
5. Dorman, S.N., Shirley, B.C., **Caminsky, N.G.**, Knoll, J.H.M., Rogan, P.K. “Expanding probe sequence repertoire and reproducibility in human genomic hybridization” Presented at the *Department of Biochemistry Graduate Student Seminars, University of Western Ontario* (London, Canada) 2012.
6. Dorman, S.N., Shirley, B.C., **Caminsky, N.G.**, Knoll, J.H.M., Rogan, P.K. “Next generation genomic microarrays and custom FISH probes for molecular cytogenetic analysis designed by ab initio sequence analysis” Presented at the *Great Lakes Chromosomes Conference* (Toronto, Canada) May 2012.
7. Dorman, S.N., Shirley, B.C., **Caminsky, N.G.**, Knoll, J.H.M., Rogan, P.K. “Next generation genomic microarrays and custom FISH probes for molecular cytogenetic analysis designed using ab initio sequences” Presented at the *CIHR – Strategic Training Program in Cancer Research & Technology Transfer and the Department of Oncology Research & Education Day* (London, Canada) June 2012.
8. Rogan, P.K., Li, Y., Wickramasinghe, A., Subasinghe, A., **Caminsky, N.**, Khan, W., Samarabandu, J., Knoll, J.H., Wilkins, R., Flegal, F. “Automating dicentric chromosome detection from cytogenetic biodosimetry data” Manuscript submitted to the *EPRBioDose International Conference* (Leiden, Netherlands) March 2013.
9. Rogan, P.K., Li, Y., Wickramasinghe, A., Subasinghe, A., **Caminsky, N.**, Khan, W., Samarabandu, J., Knoll, J.H., Wilkins, R., Flegal, F. “Automated Dicentric Chromosome

- Identifier: Software for high throughput determination of exposures in a mass casualty radiation event” Presented at the *EPR 2013/Dart-Dose CMCR Meeting* (Dartmouth, USA) June 2013.
10. **Caminsky, N.G.**, Mucaki, E.J., Shirley, Coby Viner, C.S., Dovigi, E., Knoll, J.M., Ainsworth, P., Rogan, P.K. “Identification, Prediction, and Prioritization of Non-Coding Variants of Uncertain Significance in Heritable Breast Cancer Using Multiplex Approach for Sample Preparation and Sequencing” Presented at the *Undergraduate Thesis Poster Presentation Session and Undergraduate Thesis Oral Presentations* (London, Canada) August 2013.
 11. E.J. Mucaki, **N. Caminsky**, E. Dovigi, A. Stuart, C. Viner, B. Shirley, J.H. Knoll, P. Ainsworth, P.K. Rogan. “Identification, Prediction, and Prioritization of Non-Coding Variants of Uncertain Significance in Heritable Breast/Ovarian Cancer” Presented at *London Health Research Day* (London, Canada) April 2014.
 12. Eliseos Mucaki, **Natasha Caminsky**, Alan Stuart, Coby Viner, Ben Shirley, Joan Knoll, Peter Ainsworth, Peter Rogan. “*Identification, Prediction, and Prioritization of Non-Coding Variants of Uncertain Significance in Heritable Breast/Ovarian Cancer*” Proffered paper presented at the *Fifth International Symposium on Hereditary Breast and Ovarian Cancer, BRCA: Twenty Years of Advances* (Montreal, Canada) May 2014.
 13. **Caminsky NG**, Mucaki EJ, Maistrovski Y, Viner C, Shirley BC, Stuart A, Ainsworth P, Knoll JH, Rogan PK. “*Identification and Prioritization of Variants of Uncertain Significance in Hereditary Breast and Ovarian Cancer*” Presented at the *CIHR – Strategic Training Program in Cancer Research & Technology Transfer and the Department of Oncology Research & Education Day* (London, Canada) June 2014.
 14. **Caminsky, Natasha**. “The 5th International Symposium on Hereditary Breast and Ovarian Cancer” Published as a blog post on the website of *The Breast Cancer Society of Canada* July 7, 2014. <http://www.bpsc.ca/blog/category/research/>
 15. **N.G. Caminsky**, E.J. Mucaki, P.K. Rogan. “Predicting splicing mutations by information theory-based analysis in rare and common diseases: performance and best practices” Presented at the *64th Annual Meeting of the American Society of Human Genetics* (San Diego, USA) October 2014.
 16. *Rogan Lab - mRNA splicing in genetic diseases*. (2015). at https://www.youtube.com/watch?v=_IbFaPObcXI and https://www.youtube.com/watch?v=GRyHf_3PexU.
 17. **Caminsky NG**, Mucaki EJ, Lu R, Perri A, Rogan PK. “A Unified Framework for the Identification and Prioritization of Coding and Non-Coding Variants in Heritable Breast Cancer (HBOC)” Poster presentation at *London Health Research Day* (London, Canada) April 2015.

18. **Caminsky NG**, Mucaki EJ, Lu R, Perri A, Rogan PK. “A Unified Framework for the Identification and Prioritization of Coding and Non-Coding Variants in Heritable Breast Cancer (HBOC)” Oral Presentation at the *CIHR – Strategic Training Program in Cancer Research & Technology Transfer and the Department of Oncology Research & Education Day* (London, Canada) June 2015. Granted 1 of 2 Oral Presentation Awards (8 presentations total, 198 abstracts).
19. **Caminsky NG**, Mucaki EJ, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JHM, Rogan PK. “Seeking the “Missing Heritability” in High-Risk Hereditary Breast and Ovarian Cancer (HBOC) Patients By Prioritizing Coding and Non-Coding Variants in 21 Genes” Accepted for the *2015 Canadian Cancer Research Conference* (Montreal, Canada) November 2015.

Relevant Volunteer/Community Service and Involvement

Organization: Cancer Genetics Clinic, London Health Sciences Centre (ON)
 Role: Sponsored Student Position, Volunteer
 Dates of Service: 2013-2014

Organization: Canadian Cancer Society
 Role: Speaker at Ignite Cancer London
 Dates of Service: March 25th, 2014

Organization: Breast Cancer Society of Canada
 Role: Mother’s Day Walk Volunteer
 Dates of Service: May 10th, 2015