

**GWAS META-ANALYSIS: METHODOLOGY AND  
APPLICATION TO HUMAN MEIOTIC  
RECOMBINATION**

by

**Ferdouse Begum**

B. Sc. (Statistics), Jahangirnagar University, Bangladesh, 1997

M. Sc. (Statistics), Jahangirnagar University, Bangladesh, 2000

M. A. (Statistics), Ball State University, 2007

Submitted to the Graduate Faculty of

the Department of Biostatistics

the Graduate School of Public Health, in partial fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

**2013**

UNIVERSITY OF PITTSBURGH  
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

**Ferdouse Begum**

It was defended on

**July 11th, 2013**

and approved by

**Eleanor Feingold, Ph.D.**

Professor, Departments of Biostatistics & Human Genetics  
Graduate School of Public Health, University of Pittsburgh

**George Tseng, Ph.D.**

Associate Professor, Departments of Biostatistics & Human Genetics  
Graduate School of Public Health, University of Pittsburgh

**M. Ilyas Kamboh, Ph.D.**

Professor and Chairman, Department of Human Genetics  
Graduate School of Public Health, University of Pittsburgh

**Abdus Wahed, Ph.D.**

Associate Professor, Department of Biostatistics  
Graduate School of Public Health, University of Pittsburgh

**Yan Lin, Ph.D.**

Research Assistant Professor, Departments of Medicine & Biostatistics  
Graduate School of Public Health, University of Pittsburgh

Dissertation Director: **Eleanor Feingold, Ph.D.**

Professor, Departments of Biostatistics & Human Genetics  
Graduate School of Public Health, University of Pittsburgh

Copyright © by Ferdouse Begum  
**2013**

# GWAS META-ANALYSIS: METHODOLOGY AND APPLICATION TO HUMAN MEIOTIC RECOMBINATION

Ferdouse Begum, PhD

University of Pittsburgh, 2013

## ABSTRACT

Human meiotic recombination is critical to successful human reproduction and to maintaining genetic diversity. Recombination anomalies are associated with aberrant meiotic outcomes with significant consequences. One important method for studying recombination is genome-wide association studies (GWAS) of recombination phenotypes. Because such studies require nuclear or three-generation family samples that have been genotyped on GWAS chips, the number of suitable datasets is limited. The goal of this dissertation is to develop methods for increasing the available sample sizes for GWAS of recombination phenotypes.

We developed two different approaches for increasing sample size. First, we made it possible to include additional family types in the analysis. We developed methods for scoring recombination for half-sibling pedigrees and three generation pedigrees with ungenotyped individuals. Second, we developed a regionally smoothed meta-analysis method for GWAS data, which will allow the combination datasets that have been genotyped on different chips. This method will help increase available sample sizes for recombination studies, but is also applicable to all GWAS studies.

The public health significance of this work is that our developments will allow us to find new genes that control recombination and more information about already-known genes. This information can be used for improved treatment and prevention of the consequences of aberrant recombination, including infertility and births with significant chromosomal anomalies.

**Keywords:** Meiotic Recombination, Crossover, Recombination Scoring, GWAS meta-analysis, Methods of GWAS meta-analysis, Regionally smoothed meta-analysis.



## TABLE OF CONTENTS

<b>PREFACE</b> . . . . .	xxiv
<b>1.0 INTRODUCTION</b> . . . . .	1
1.1 HUMAN MEIOTIC RECOMBINATION . . . . .	1
1.2 RECOMBINATION SCORING . . . . .	3
1.3 GENOME-WIDE ASSOCIATION STUDIES (GWAS) . . . . .	3
1.4 GWAS META-ANALYSIS . . . . .	4
1.5 SPECIFIC AIMS . . . . .	5
1.5.1 Project 1: Genetics of Meiotic Recombination: Genome Wide Association Studies for Recombination Phenotypes. (Chapter Two) . . . . .	5
1.5.2 Project 2: Scoring Recombination using Complex Pedigree Structures including Half-Siblings. (Chapter Three) . . . . .	6
1.5.3 Project 3: Comprehensive Literature Review and Statistical Considerations for GWAS Meta-Analysis. (Chapter Four) . . . . .	6
1.5.4 Project 4: Regionally Smoothed Meta-Analysis for GWAS. (Chapter Five) . . . . .	6
1.6 REFERENCES . . . . .	7
<b>2.0 GENETICS OF MEIOTIC RECOMBINATION: GENOME WIDE ASSOCIATION STUDIES FOR RECOMBINATION PHENOTYPES</b> . . . . .	10
2.1 ABSTRACT . . . . .	10
2.2 INTRODUCTION . . . . .	10
2.3 METHODS . . . . .	13

2.3.1	Study Population and Samples . . . . .	13
2.3.2	Pedigrees . . . . .	13
2.3.3	Phenotypes . . . . .	13
2.3.4	Genotypes, Error Checking and Data Handling . . . . .	15
2.3.5	Genome-wide Association Studies . . . . .	15
2.4	RESULTS . . . . .	16
2.4.1	Recombination Phenotypes . . . . .	16
2.4.1.1	Average Recombination Count (ARC) . . . . .	16
2.4.1.2	Hotspot analysis: Percent of recombination occurring in hotspots (HS_PCT) . . . . .	19
2.4.1.3	Hotspot analysis: Average count of recombinants in hotspots (HS_CNT) . . . . .	20
2.4.1.4	Non-Hotspot analysis: Average count of recombinants in non-hotspot areas (NHS_CNT) . . . . .	20
2.4.1.5	Percent of recombination occurring near the motif (MOTIF)	23
2.4.2	In-depth Analysis of <i>RNF212</i> , <i>PRDM9</i> and MATP . . . . .	23
2.4.2.1	<i>PRDM9</i> gene across phenotypes . . . . .	23
2.4.2.2	<i>RNF212</i> gene across phenotypes . . . . .	27
2.4.2.3	Chromosome 17 inversion region across phenotypes . . . . .	27
2.4.2.4	Gene X Gene interaction models . . . . .	27
2.4.3	FHS Replication . . . . .	27
2.5	DISCUSSION . . . . .	31
2.6	SUPPORTING INFORMATION . . . . .	32
2.7	REFERENCES . . . . .	33
<b>3.0</b>	<b>SCORING RECOMBINATION IN COMPLEX PEDIGREE STRUC-</b> <b>TURES INCLUDING HALF-SIBLINGS . . . . .</b>	<b>36</b>
3.1	INTRODUCTION . . . . .	36
3.2	THREE-GENERATION FAMILIES WITH VARYING NUMBER OF MISS- ING GENOTYPES . . . . .	37

3.3	RECOMBINATION SCORING IN TWO-GENERATION FAMILIES WITH HALF-SIBLINGS . . . . .	39
3.4	RECOMBINATION SCORING IN THREE-GENERATION FAMILIES WITH HALF-SIBLINGS . . . . .	41
3.5	REFERENCES . . . . .	42
<b>4.0</b>	<b>COMPREHENSIVE LITERATURE REVIEW AND STATISTICAL CONSIDERATIONS FOR GWAS META-ANALYSIS . . . . .</b>	<b>43</b>
4.1	ABSTRACT . . . . .	44
4.2	INTRODUCTION . . . . .	44
4.3	GWAS META-ANALYSIS DATA AND METHODS . . . . .	45
4.4	DATABASES AND SOFTWARE . . . . .	49
4.5	LITERATURE REVIEW . . . . .	51
4.6	CASE STUDY . . . . .	53
4.7	COMPLICATIONS AND OPEN QUESTIONS . . . . .	57
4.8	CONCLUSION . . . . .	59
4.9	SUPPLEMENTARY MATERIAL . . . . .	60
4.10	ACKNOWLEDGMENTS . . . . .	60
4.11	FUNDING . . . . .	60
4.12	REFERENCES . . . . .	61
<b>5.0</b>	<b>REGIONALLY SMOOTHED META-ANALYSIS (RSM) METHODS FOR GWAS DATASETS . . . . .</b>	<b>65</b>
5.1	ABSTRACT . . . . .	65
5.2	INTRODUCTION . . . . .	65
5.3	METHODOLOGY . . . . .	68
5.3.1	First Stage . . . . .	68
5.3.2	Second Stage . . . . .	69
5.3.3	Window Type and Size . . . . .	69
5.3.4	Performance Measure of RSM Methods . . . . .	70
5.4	EXPERIMENTAL DESIGNS AND TESTING IN REAL DATA . . . . .	70

5.4.1	Introduction . . . . .	70
5.4.2	Results . . . . .	72
5.4.3	Discussion . . . . .	73
5.5	RSM SOFTWARE . . . . .	74
5.5.1	Program Description . . . . .	74
5.5.1.1	Program input files . . . . .	74
5.5.1.2	Choice of Window type . . . . .	75
5.5.1.3	Methods choice . . . . .	75
5.5.1.4	Program Workflow . . . . .	75
5.5.2	Example . . . . .	75
5.5.3	Discussion . . . . .	77
5.6	REFERENCES . . . . .	78
<b>6.0</b>	<b>CONCLUSIONS . . . . .</b>	<b>80</b>
6.1	DISSERTATION CONCLUSIONS . . . . .	80
6.2	STRENGTHS AND SHORTCOMINGS . . . . .	82
6.3	FUTURE DIRECTION . . . . .	83
<b>APPENDIX A. GENETICS OF MEIOTIC RECOMBINATION: GENOME</b>		
<b>WIDE ASSOCIATION STUDIES FOR RECOMBINATION PHENO-</b>		
<b>TYPES . . . . .</b>		<b>84</b>
A.1	LocusZoom plot of previously reported genes and SNPs in our study . . . . .	93
A.1.1	Phenotype: ARC . . . . .	93
A.1.2	Phenotype: HS_PCT . . . . .	96
A.2	LocusZoom plot of our top hits for five phenotypes . . . . .	97
A.2.1	Phenotype: ARC . . . . .	97
A.2.2	Phenotype: HS_PCT . . . . .	102
A.2.3	Phenotype: HS_CNT . . . . .	107
A.2.4	Phenotype: NHS_CNT . . . . .	112
A.2.5	Phenotype: MOTIF . . . . .	117
A.3	LocusZoom plot of our top hits in FHS data set . . . . .	121

A.3.1 Phenotype: ARC . . . . .	121
A.3.2 Phenotype: HS_PCT . . . . .	126
A.3.3 Phenotype: HS_CNT . . . . .	132
A.3.4 Phenotype: NHS_CNT . . . . .	137
A.3.5 Phenotype: MOTIF . . . . .	143
A.4 LocusZoom plot of <i>MAPT</i> gene in our meta-analysis across phenotypes . .	146
<b>APPENDIX B. SCORING RECOMBINATION IN COMPLEX PEDIGREE</b>	
<b>STRUCTURES INCLUDING HALF-SIBLINGS . . . . .</b>	150
<b>BIBLIOGRAPHY . . . . .</b>	182

## LIST OF TABLES

2.1	SNPs with lowest P-values for ARC . . . . .	18
2.2	SNPs with lowest P-values for HS_PCT . . . . .	21
2.3	SNPs with lowest P-values for HS_CNT . . . . .	22
2.4	SNPs with lowest P-values for NHS_CNT . . . . .	24
2.5	SNPs with lowest P-values for MOTIF . . . . .	25
2.6	<i>PRDM9</i> gene association across phenotypes . . . . .	26
2.7	<i>RNF212</i> gene association across phenotypes . . . . .	28
3.1	Three people Genotyped (mother and half-siblings pair) . . . . .	41
4.1	Sources of information for different methods of meta-analysis . . . . .	48
4.2	Case study results . . . . .	54
5.1	Ranks of <i>PRDM9</i> gene for HS_PCT phenotype . . . . .	72
5.2	Ranks of <i>PRDM9</i> gene for NHS_CNT phenotype . . . . .	73
5.3	Ranks of <i>RNF212</i> gene for ARC phenotype . . . . .	73
5.4	Example output file after stage one. . . . .	77
5.5	Example output file after stage two. . . . .	77

## LIST OF FIGURES

2.1	Manhattan plot of phenotype ARC (combined analysis) . . . . .	16
2.2	QQ plot of phenotype ARC (combined analysis) . . . . .	17
2.3	Manhattan plot of phenotype HS_PCT (combined analysis) . . . . .	19
2.4	QQ plot of phenotype HS_PCT (male and female combined analysis) . . .	20
2.5	<i>RNF212</i> (male) in FHS data set . . . . .	29
2.6	<i>PRDM9</i> (gender-pooled) in FHS data set . . . . .	30
3.1	(A) Pedigree structure of three-generation family; (B) Pedigree structure of two-generation nuclear family . . . . .	36
3.2	Recombination plot of chromosome 1 (Mukhopadhyay N.) . . . . .	37
3.3	Three people Genotyped (grandfather, grandmother and grandchild) . . .	38
3.4	Pedigree structure of two-generation family with half-siblings . . . . .	39
3.5	Pedigree structure of two-generation family with half-siblings and missing genotypes of fathers. . . . .	40
3.6	Pedigree structure of three-generation family with half-siblings. . . . .	41
4.1	Number of GWAS studies by year of publication. . . . .	51
4.2	Summary of GWAS meta-analysis review: (A) type of meta-analysis; (B) type of paper; (C) type of meta-analysis method; (D) software used. . . .	52
4.3	Forest plot of the selected SNPs. . . . .	56
5.1	Sample size distribution of different studies. . . . .	71
5.2	Program algorithm of RSMgwas package. . . . .	76
A.1	QQ plot ARC(female) . . . . .	85

A.2	QQ plot ARC(male) . . . . .	85
A.3	QQ plot ARC(female) . . . . .	85
A.4	QQ plot ARC(male) . . . . .	85
A.5	Manhattan plot of phenotype HS_PCT (female) . . . . .	86
A.6	Manhattan plot of phenotype HS_PCT (male) . . . . .	86
A.7	QQ plot HS_PCT(female) . . . . .	86
A.8	QQ plot HS_PCT(male) . . . . .	86
A.9	Manhattan plot of phenotype HS_CNT (combined) . . . . .	87
A.10	QQ plot HS_CNT(combined) . . . . .	87
A.11	Manhattan plot of phenotype HS_CNT (female) . . . . .	88
A.12	Manhattan plot of phenotype HS_CNT (male) . . . . .	88
A.13	QQ plot HS_CNT(female) . . . . .	88
A.14	QQ plot HS_CNT(male) . . . . .	88
A.15	Manhattan plot of phenotype NHS_CNT (combined) . . . . .	89
A.16	QQ plot NHS_CNT(combined) . . . . .	89
A.17	Manhattan plot of phenotype NHS_CNT (female) . . . . .	90
A.18	Manhattan plot of phenotype NHS_CNT (male) . . . . .	90
A.19	QQ plot NHS_CNT(female) . . . . .	90
A.20	QQ plot NHS_CNT(male) . . . . .	90
A.21	Manhattan plot of phenotype MOTIF (combined) . . . . .	91
A.22	QQ plot MOTIF(combined) . . . . .	91
A.23	Manhattan plot of phenotype MOTIF (female) . . . . .	92
A.24	Manhattan plot of phenotype MOTIF (male) . . . . .	92
A.25	QQ plot MOTIF(female) . . . . .	92
A.26	QQ plot MOTIF(male) . . . . .	92
A.27	ARC(female) . . . . .	93
A.28	ARC(female) . . . . .	93
A.29	ARC(female) . . . . .	93
A.30	ARC(female) . . . . .	93



A.31	ARC(female)	94
A.32	ARC(female)	94
A.33	ARC(female)	94
A.34	ARC(female)	94
A.35	ARC(female)	94
A.36	ARC(male)	94
A.37	ARC(male)	95
A.38	ARC(male)	95
A.39	ARC(male)	95
A.40	ARC(male)	95
A.41	ARC(combined)	95
A.42	ARC(combined)	95
A.43	ARC(combined)	96
A.44	ARC(combined)	96
A.45	HS_PCT(combined)	96
A.46	HS_PCT(combined)	96
A.47	HS_PCT(combined)	97
A.48	HS_PCT(combined)	97
A.49	ARC(combined)	97
A.50	ARC(combined)	97
A.51	ARC(combined)	98
A.52	ARC(combined)	98
A.53	ARC(combined)	98
A.54	ARC(combined)	98
A.55	ARC(combined)	98
A.56	ARC(combined)	98
A.57	ARC(combined)	99
A.58	ARC(combined)	99
A.59	ARC(female)	99

A.60	ARC(female)	99
A.61	ARC(female)	99
A.62	ARC(female)	99
A.63	ARC(combined)	100
A.64	ARC(combined)	100
A.65	ARC(female)	100
A.66	ARC(female)	100
A.67	ARC(male)	100
A.68	ARC(male)	100
A.69	ARC(male)	101
A.70	ARC(male)	101
A.71	ARC(male)	101
A.72	ARC(male)	101
A.73	ARC(male)	101
A.74	ARC(male)	101
A.75	HS_PCT(combined)	102
A.76	HS_PCT(combined)	102
A.77	HS_PCT(combined)	102
A.78	HS_PCT(combined)	102
A.79	HS_PCT(combined)	103
A.80	HS_PCT(combined)	103
A.81	HS_PCT(combined)	103
A.82	HS_PCT(combined)	103
A.83	HS_PCT(combined)	103
A.84	HS_PCT(female)	103
A.85	HS_PCT(female)	104
A.86	HS_PCT(female)	104
A.87	HS_PCT(female)	104
A.88	HS_PCT(female)	104

A.89	HS_PCT(female)	104
A.90	HS_PCT(female)	104
A.91	HS_PCT(female)	105
A.92	HS_PCT(female)	105
A.93	HS_PCT(female)	105
A.94	HS_PCT(female)	105
A.95	HS_PCT(male)	105
A.96	HS_PCT(male)	105
A.97	HS_PCT(male)	106
A.98	HS_PCT(male)	106
A.99	HS_PCT(male)	106
A.100	HS_PCT(male)	106
A.101	HS_PCT(male)	106
A.102	HS_PCT(male)	106
A.103	HS_PCT(male)	107
A.104	HS_PCT(male)	107
A.105	HS_CNT(combined)	107
A.106	HS_CNT(combined)	107
A.107	HS_CNT(combined)	108
A.108	HS_CNT(combined)	108
A.109	HS_CNT(combined)	108
A.110	HS_CNT(combined)	108
A.111	HS_CNT(combined)	108
A.112	HS_CNT(combined)	108
A.113	HS_CNT(female)	109
A.114	HS_CNT(female)	109
A.115	HS_CNT(female)	109
A.116	HS_CNT(female)	109
A.117	HS_CNT(female)	109

A.118	HS_CNT(female)	109
A.119	HS_CNT(female)	110
A.120	HS_CNT(female)	110
A.121	HS_CNT(male)	110
A.122	HS_CNT(male)	110
A.123	HS_CNT(male)	110
A.124	HS_CNT(male)	110
A.125	HS_CNT(male)	111
A.126	HS_CNT(male)	111
A.127	HS_CNT(male)	111
A.128	HS_CNT(male)	111
A.129	HS_CNT(male)	111
A.130	HS_CNT(male)	111
A.131	NHS_CNT(combined)	112
A.132	NHS_CNT(combined)	112
A.133	NHS_CNT(combined)	112
A.134	NHS_CNT(combined)	112
A.135	NHS_CNT(combined)	113
A.136	NHS_CNT(combined)	113
A.137	NHS_CNT(combined)	113
A.138	NHS_CNT(combined)	113
A.139	NHS_CNT(combined)	113
A.140	NHS_CNT(combined)	113
A.141	NHS_CNT(combined)	114
A.142	NHS_CNT(combined)	114
A.143	NHS_CNT(female)	114
A.144	NHS_CNT(female)	114
A.145	NHS_CNT(female)	114
A.146	NHS_CNT(female)	114

A.147	NHS_CNT(female)	115
A.148	NHS_CNT(female)	115
A.149	NHS_CNT(female)	115
A.150	NHS_CNT(female)	115
A.151	NHS_CNT(female)	115
A.152	NHS_CNT(male)	115
A.153	NHS_CNT(male)	116
A.154	NHS_CNT(male)	116
A.155	NHS_CNT(male)	116
A.156	NHS_CNT(male)	116
A.157	NHS_CNT(male)	116
A.158	NHS_CNT(male)	116
A.159	NHS_CNT(male)	117
A.160	NHS_CNT(male)	117
A.161	MOTIF(combined)	117
A.162	MOTIF(combined)	117
A.163	MOTIF(combined)	118
A.164	MOTIF(combined)	118
A.165	MOTIF(combined)	118
A.166	MOTIF(combined)	118
A.167	MOTIF(combined)	118
A.168	MOTIF(female)	118
A.169	MOTIF(female)	119
A.170	MOTIF(female)	119
A.171	MOTIF(female)	119
A.172	MOTIF(female)	119
A.173	MOTIF(male)	119
A.174	MOTIF(male)	119
A.175	MOTIF(male)	120

A.176	MOTIF(male)	120
A.177	MOTIF(male)	120
A.178	MOTIF(male)	120
A.179	ARC(combined)	121
A.180	ARC(combined)	121
A.181	ARC(combined)	121
A.182	ARC(combined)	121
A.183	ARC(combined)	122
A.184	ARC(combined)	122
A.185	ARC(combined)	122
A.186	ARC(combined)	122
A.187	ARC(combined)	122
A.188	ARC(combined)	122
A.189	ARC(female)	123
A.190	ARC(female)	123
A.191	ARC(female)	123
A.192	ARC(female)	123
A.193	ARC(female)	123
A.194	ARC(female)	123
A.195	ARC(female)	124
A.196	ARC(female)	124
A.197	ARC(male)	124
A.198	ARC(male)	124
A.199	ARC(male)	124
A.200	ARC(male)	124
A.201	ARC(male)	125
A.202	ARC(male)	125
A.203	ARC(male)	125
A.204	ARC(male)	125

A.205	ARC(male)	125
A.206	HS_PCT(combined)	126
A.207	HS_PCT(combined)	126
A.208	HS_PCT(combined)	126
A.209	HS_PCT(combined)	126
A.210	HS_PCT(combined)	127
A.211	HS_PCT(combined)	127
A.212	HS_PCT(combined)	127
A.213	HS_PCT(combined)	127
A.214	HS_PCT(combined)	127
A.215	HS_PCT(combined)	127
A.216	HS_PCT(female)	128
A.217	HS_PCT(female)	128
A.218	HS_PCT(female)	128
A.219	HS_PCT(female)	128
A.220	HS_PCT(female)	128
A.221	HS_PCT(female)	128
A.222	HS_PCT(female)	129
A.223	HS_PCT(female)	129
A.224	HS_PCT(female)	129
A.225	HS_PCT(female)	129
A.226	HS_PCT(female)	129
A.227	HS_PCT(male)	129
A.228	HS_PCT(male)	130
A.229	HS_PCT(male)	130
A.230	HS_PCT(male)	130
A.231	HS_PCT(male)	130
A.232	HS_PCT(male)	130
A.233	HS_PCT(male)	130

A.234	HS_PCT(male)	131
A.235	HS_PCT(male)	131
A.236	HS_PCT(male)	131
A.237	HS_PCT(male)	131
A.238	HS_CNT(combined)	132
A.239	HS_CNT(combined)	132
A.240	HS_CNT(combined)	132
A.241	HS_CNT(combined)	132
A.242	HS_CNT(combined)	133
A.243	HS_CNT(combined)	133
A.244	HS_CNT(combined)	133
A.245	HS_CNT(combined)	133
A.246	HS_CNT(female)	133
A.247	HS_CNT(female)	133
A.248	HS_CNT(female)	134
A.249	HS_CNT(female)	134
A.250	HS_CNT(female)	134
A.251	HS_CNT(female)	134
A.252	HS_CNT(female)	134
A.253	HS_CNT(female)	134
A.254	HS_CNT(female)	135
A.255	HS_CNT(female)	135
A.256	HS_CNT(female)	135
A.257	HS_CNT(male)	135
A.258	HS_CNT(male)	135
A.259	HS_CNT(male)	135
A.260	HS_CNT(male)	136
A.261	HS_CNT(male)	136
A.262	HS_CNT(male)	136



A.263	HS_CNT(male)	136
A.264	HS_CNT(male)	136
A.265	HS_CNT(male)	136
A.266	HS_CNT(male)	137
A.267	HS_CNT(male)	137
A.268	NHS_CNT(combined)	137
A.269	NHS_CNT(combined)	137
A.270	NHS_CNT(combined)	138
A.271	NHS_CNT(combined)	138
A.272	NHS_CNT(combined)	138
A.273	NHS_CNT(combined)	138
A.274	NHS_CNT(combined)	138
A.275	NHS_CNT(combined)	138
A.276	NHS_CNT(combined)	139
A.277	NHS_CNT(combined)	139
A.278	NHS_CNT(combined)	139
A.279	NHS_CNT(combined)	139
A.280	NHS_CNT(female)	139
A.281	NHS_CNT(female)	139
A.282	NHS_CNT(female)	140
A.283	NHS_CNT(female)	140
A.284	NHS_CNT(female)	140
A.285	NHS_CNT(female)	140
A.286	NHS_CNT(female)	140
A.287	NHS_CNT(female)	140
A.288	NHS_CNT(female)	141
A.289	NHS_CNT(male)	141
A.290	NHS_CNT(male)	141
A.291	NHS_CNT(male)	141

A.292	NHS_CNT(male)	141
A.293	NHS_CNT(male)	141
A.294	NHS_CNT(male)	142
A.295	NHS_CNT(male)	142
A.296	NHS_CNT(male)	142
A.297	NHS_CNT(male)	142
A.298	NHS_CNT(male)	142
A.299	MOTIF(combined)	143
A.300	MOTIF(combined)	143
A.301	MOTIF(combined)	143
A.302	MOTIF(combined)	143
A.303	MOTIF(combined)	144
A.304	MOTIF(combined)	144
A.305	MOTIF(combined)	144
A.306	MOTIF(female)	144
A.307	MOTIF(female)	144
A.308	MOTIF(female)	144
A.309	MOTIF(female)	145
A.310	MOTIF(female)	145
A.311	MOTIF(male)	145
A.312	MOTIF(male)	145
A.313	MOTIF(male)	145
A.314	MOTIF(male)	145
A.315	MOTIF(male)	146
A.316	MOTIF(male)	146
A.317	ARC(combined)	146
A.318	ARC(female)	146
A.319	ARC(male)	147
A.320	HS_PCT(combined)	147

A.321	HS_PCT(female)	147
A.322	HS_PCT(male)	147
A.323	HS_CNT(combined)	147
A.324	HS_CNT(female)	147
A.325	HS_CNT(male)	148
A.326	NHS_CNT(combined)	148
A.327	NHS_CNT(female)	148
A.328	NHS_CNT(male)	148
A.329	MOTIF(combined)	148
A.330	MOTIF(female)	148
A.331	MOTIF(male)	149

## PREFACE

This dissertation would not have been possible without the constant guidance of my advisor, committee members, and teachers from all levels of my educational career, persistent inspiration from my parents, and support from my family and friends. This dissertation not only contains the culmination of the several years of research but is also a tale of the best part of my life in graduate school.

I would like to express my deepest gratitude to my Ph.D dissertation advisor Dr. Eleanor Feingold, whose intuitive power, knowledge and wisdom impressed me overwhelmingly. She always made herself available when I needed her the most. Her inspiring words, wise and constructive criticism, and intellectual challenges motivated me to stay focused on my work.

I would like to thank my committee members Dr. Tseng for giving me the opportunity to work with him and for his encouraging words, and Dr. Wahed for his advice and suggestion regarding my dissertation in details. I am grateful to Dr. Ilyas Kamboh for his human genetics viewpoints and his suggestions beyond dissertation. I am thankful to Dr. Lin for her friendliness and open discussion from time to time.

I acknowledge the contribution of our collaborator Dr. Reshmi Chowdhury for her GWAS results for two data sets and Dr. Mary L. Marazita for providing us the data set for our research.

I am indebted to my master's thesis advisor and mentor Dr. Mir Masoom Ali, who planted the seed of statistical genetic research in me. Dr. Ali and Mrs. Ali helped me to adjust to a new culture in a new country and still inspire me to excel. I am thankful to Dr. Daniel E. Weeks for his invitation to join as a Ph.D. student in the Department of Human Genetics. His door was always open for help. I am grateful to two women Dr. Roslyn Stone

and Dr. Lisa Weissfeld for their support when I was struggling with my new motherhood. I am thankful to Dr. Ryan Minster for helping me to learn LaTeX and making the formatting of the dissertation less dreary. Thanks to Joanne Pegher for her flexible schedule and Joan Anson for various helps over the time. I would like to express my appreciation to all my teachers and friends who put faith in me and urged me to finish my dissertation. I am grateful to them for helping me stay sane in stressful times and making my life more bearable in the course of my graduate work. Special thanks to Jonathan Yabes and Yi Fan for coming to my defense and thanks Jing Zheng for inspiring me with her kind words.

My deepest love goes for my two angels Laiba and Liyana, who were born during my graduate school years. I feel guilty to steal their time and stay apart to work on this dissertation and I am deeply sorry about this. My love and appreciation goes to my husband Monir H. Sharker for his continuous support and profound understanding. Being both husband and wife Ph.D students with two children was not easy. Going through the tough times made us tougher and more understanding. This long journey together provided us richness and deeper meaning of our life.

As much I wanted this dissertation to be completed soon, my parents wanted it much more. In spite of my father's illness, he sent my mother to help me out with my children and proved that you can be equally influential in someones life even from the other side of the word. During various tough situations when I was about to give up, his encouraging words put me back to my work. When I could almost see the end of this exciting and exhausted journey, I lost my father. Though physically he is not around me yet thought and feelings of his presence around me and his blessings will always be there for me.

This dissertation is dedicated to my father, and to all members of my family.

## 1.0 INTRODUCTION

### 1.1 HUMAN MEIOTIC RECOMBINATION

Meiotic recombination plays a crucial role in human reproduction. Meiotic recombination is embedded in the process of meiosis, which starts with DNA replication followed by two cycles of cell division to ensure that the gametes (sex cells) contain half of the chromosomes as compared to the diploid cells [1, 2, 3]. Meiosis takes place in both female and male and is known as oogenesis and spermatogenesis in females and males respectively [4]. Meiosis is a two-stage process with several sub-stages in each stage. In meiosis I, there are four main sub-stages: prophase I, metaphase I, anaphase I and telophase I. Meiotic recombination, the exchange of genetic material between homologous chromosomes, takes place at the end of a prolonged sub-stage, prophase I [4]. It is a highly regulated process, which starts with controlled fragmentation of chromosomes by DNA double strand breaks (DSB) [5, 6]. DSBs repair happen in their homologs on another chromatid. DSBs are not uniformly distributed across the genome but instead have site preferences. Following DSBs, a synaptonemal complex (SC) develops and the stability of the SC increases as the double Holliday junction forms. A few of the junctions end up as recombination sites [2].

Meiotic recombination serves a number of important functions, including increased genomic diversity and ensures proper chromosomal segregation. One of the reasons that recombination is of interest is because abnormal recombination has been linked to adverse health outcomes. A decreased or increased rate of recombination is associated with improper chromosomal segregation [7, 8, 9, 10, 11]. Many studies in model organisms have also

shown this link between abnormal recombination patterns and chromosomal nondisjunction, in which chromosomes do not separate properly during different stages of meiosis [12, 13, 14]. Chromosomal nondisjunction results in chromosomal aneuploidy, or aberrant chromosome number, a common event that presents in approximately one fourth of all pregnancies [15, 16]. One third of spontaneous miscarriages are related to chromosomal aneuploidy, which makes aneuploidy one of the leading causes of pregnancy loss [15, 16]. Among live births with aneuploidy, most face profound clinical consequences including birth defects and various forms cognitive disability. Among live births with aneuploidy, the majority are trisomies, since most of the monosomies are not viable because of disruptive embryonic development [15]. Trisomies account for .3%-5% of live births [15, 17]. Reduced recombination and trisomies are also strongly associated in humans [18, 19, 20, 21, 22]. In addition, chromosomal nondisjunction afflicts women increasingly as age advances. A significant association between maternal age and location of recombination events across the genome was reported by Lamb et. al. (2005) in mothers of children with trisomy 21 [23]. But it is not obvious how advanced age is potentially associated with altered recombination patterns in general.

Recently several studies have used both genetic epidemiological and molecular methods to start to uncover the genetic determinants of human meiotic recombination. Several genes have been convincingly identified (reviewed in Chapter 2), and others have been suggested. Further investigation may lead to discovery of additional genes and to better understanding of the effects of those that have already been identified [24, 25]. Finding the genetic basis of different recombination phenotypes may identify markers associated with age-related or non-age-related aneuploidy risk, which may lead to possible interventions to lengthen womens reproductive life spans. In this dissertation I use a variety of different approaches to expand the sample sizes available for genetic epidemiological studies of meiotic recombination. I use traditional GWAS meta-analysis to combine different recombination GWAS. In addition to that I develop a recombination-scoring method using a SNP streak approach for new complex pedigree structures that will increase sample size in recombination GWAS. I also develop new method for GWAS meta-analysis that will increase sample size by incorporating more studies.

## 1.2 RECOMBINATION SCORING

Identifying the locations of the recombination events across the genome can be done either by direct laboratory approaches or by computational approaches. There are currently two available computational approaches to scoring recombination. The older method uses sparsely-placed genetic markers (no linkage disequilibrium) and applies a linkage-analysis type model to infer recombination events in three-generation families. This method is implemented in the CRI-MAP [26]. The newer method exploits the technological advance of denser SNP chips to score recombination in nuclear families with two or more children. Coop et. al. [25] and Chowdhury et. al. [27] independently implemented similar methods of scoring recombination for two-generation nuclear families with two or more children. These methods use dense SNP chips to call recombination events also known as a SNP-streak method. These methods identify informative SNPs - the SNPs for which one parent is homozygous and other parent is heterozygous - in a first step. For a pair of children, a switch from sharing the same allele to not sharing alleles inherited from a particular parent is scored as a recombination event for that parent.

Different studies have estimated different recombination phenotypes covering different aspects of recombination, such as the average number of recombinations for a parent or for a particular meiotic event (child), location of the recombination on the genome, or recombination on particular chromosomes. These studies are reviewed in detail in chapter 2.

## 1.3 GENOME-WIDE ASSOCIATION STUDIES (GWAS)

To identify the genetic basis for variation in different recombination phenotypes, several studies have performed genome-wide association studies (GWAS). Over the last decade, GWAS have become the standard tool for gene discovery in human disease research. A genome-wide association study is a hypothesis-free method for testing association between a



series of hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) and a trait of interest genome-wide. For over a decade, GWAS have been successfully used to detect genetic variants for many complex diseases. GWAS are particularly successful for relatively common variants, common complex diseases, and moderate to larger effect sizes.

GWAS analyses can be based on single or multiple SNPs or other functional units such as genes, haplotype blocks and pathways. Single-SNP GWAS analyses are most common. In GWAS analysis unrelated individuals from a population are used as the data set. The trait of interest can be either discrete or continuous in nature. Usually logistic regression models or contingency tables are used for identifying association between discrete phenotypes or traits and the variants. Different types of linear regression models are used for continuous type of phenotypes. Depending on the data, genetic model choice for the SNPs can be different too. Most of the GWAS software has the option for adjusting for other co-variates and their interactions in the model.

#### 1.4 GWAS META-ANALYSIS

Sample size in a GWAS plays an important role in detecting relatively smaller effects. Larger sample size is necessary to detect relatively smaller effect sizes with an acceptable power. But if the sample size is not sufficiently large, one might think of increasing sample size by using meta-analysis to integrate several GWAS results. Such approaches have the potential to increase the power to find the associated variants of moderate effect size.

If raw data from different GWAS studies are available, one can combine those and do mega-analysis. When genotyping of different studies has been done on different chips, then one can do meta-analysis, which is statistically equivalent to mega-analysis [28]. Standard GWAS meta-analysis methods can be divided in two major groups: combining p-values with or without weights, or combining effect sizes using a fixed effect or random effect method. Among all methods, the fixed effect method is the most widely used method in GWAS meta-

analysis. Chapter 4 is a review paper that discusses and compares GWAS meta-analysis methods in depth.

Although GWAS meta-analyses have been applied for various diseases and traits, not much methodological work has been done in this area. Some of the better methods published are not in use. Even when standard meta-analysis methods are applied carefully and optimally, there remain some unresolved statistical issues. These include handling of imputation uncertainty and inaccuracy, heterogeneity in the study cohorts, improper data cleaning, improper choice of genetic and meta-analysis models etc.

## 1.5 SPECIFIC AIMS

The objective of this dissertation is to develop methodology that can be used to increase sample sizes in GWAS studies of meiotic recombination. Project 1 is an applied project that used current methods to perform GWAS and GWAS meta-analysis for recombination. Project 2 developed methods that can be used to score recombination in additional family types. Project 3 reviewed and compared GWAS meta-analysis methods, and project 4 developed a new regionally-smoothed meta-analysis method that can be used for GWAS studies of recombination.

### 1.5.1 Project 1: Genetics of Meiotic Recombination: Genome Wide Association Studies for Recombination Phenotypes. (Chapter Two)

- (a) Score recombination in the Geneva Dental Caries Study (GDCS) data set and calculate new phenotypes.
- (b) Perform GWAS for all phenotypes for males, females, and both sexes combined.
- (c) Perform GWAS-meta analysis combining GDCS with results from the Autism Genetic Research studies (AGRE) data set.

- (d) Perform gene-based qualitative replication using the Framingham Heart Study (FHS) data set.

### **1.5.2 Project 2: Scoring Recombination using Complex Pedigree Structures including Half-Siblings. (Chapter Three)**

- (a) Develop methods for scoring recombination phenotypes in three-generation families with varying numbers of people with missing genotypes using a SNP streak method.
- (b) Develop methods for scoring recombination phenotypes for two and three-generation families with half-siblings.

### **1.5.3 Project 3: Comprehensive Literature Review and Statistical Considerations for GWAS Meta-Analysis. (Chapter Four)**

- (a) Review literature of current GWAS meta-analysis methods.
- (b) Conduct a case study to compare existing methods.
- (c) Discuss pitfalls and current challenges of GWAS meta-analysis.

### **1.5.4 Project 4: Regionally Smoothed Meta-Analysis for GWAS. (Chapter Five)**

- (a) Develop a regionally smoothed meta-analysis method for GWAS data sets genotyped on different chips or data sets with SNP sets with minimal overlap.
- (b) Apply new method to three GWAS data sets.
- (c) Compare performance of different methods.

## 1.6 REFERENCES

- [1] C. Ellermeier et al. “RNAi and heterochromatin repress centromeric meiotic recombination”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.19 (2010), pp. 8701–5.
- [2] C. O’Connor. “Meiosis, Genetic Recombination, and Sexual Reproduction”. In: *Nature Education* 1.1 (2008).
- [3] M. Petronczki, M. F. Siomos, and K. Nasmyth. “Un menage a quatre: the molecular biology of chromosome segregation in meiosis”. In: *Cell* 112.4 (2003), pp. 423–40.
- [4] I. Maayan. “”Meiosis in Humans””. In: *Embryo Project Encyclopedia* (2011).
- [5] J. Baudat F Fau Buard et al. “PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice”. In: 1095-9203 (Electronic) (2009).
- [6] H. G. Blitzblau and A. Hochwagen. “Genome-wide detection of meiotic DNA double-strand break hotspots using single-stranded DNA”. In: *Methods in molecular biology* 745 (2011), pp. 47–63.
- [7] A. F. Dernburg et al. “Meiotic recombination in *C. elegans* initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis”. In: *Cell* 94.3 (1998), pp. 387–98.
- [8] J. Hodgkin, H. R. Horvitz, and S. Brenner. “Nondisjunction Mutants of the Nematode *Caenorhabditis Elegans*”. In: *Genetics* 91.1 (1979), pp. 67–94.
- [9] A. M. Villeneuve. “A cis-acting locus that promotes crossing over between X chromosomes in *Caenorhabditis elegans*”. In: *Genetics* 136.3 (1994), pp. 887–902.
- [10] M. C. Zetka et al. “Synapsis and chiasma formation in *Caenorhabditis elegans* require HIM-3, a meiotic chromosome core component that functions in chromosome segregation”. In: *Genes development* 13.17 (1999), pp. 2258–70.
- [11] M. C. Zetka and A. M. Rose. “The meiotic behavior of an inversion in *Caenorhabditis elegans*”. In: *Genetics* 131.2 (1992), pp. 321–32.

- [12] K. E. Koehler et al. “Spontaneous X chromosome MI and MII nondisjunction events in *Drosophila melanogaster* oocytes have different recombinational histories”. In: *Nature genetics* 14.4 (1996), pp. 406–14.
- [13] M. D. Krawchuk and W. P. Wahls. “Centromere mapping functions for aneuploid meiotic products: Analysis of *rec8*, *rec10* and *rec11* mutants of the fission yeast *Schizosaccharomyces pombe*”. In: *Genetics* 153.1 (1999), pp. 49–55.
- [14] L. O. Ross, R. Maxfield, and D. Dawson. “Exchanges are not equally able to enhance meiotic chromosome segregation in yeast”. In: *Proceedings of the National Academy of Sciences of the United States of America* 93.10 (1996), pp. 4979–83.
- [15] J. C. Biancotti et al. “Human embryonic stem cells as models for aneuploid chromosomal syndromes”. In: *Stem cells* 28.9 (2010), pp. 1530–40.
- [16] T. Hassold and P. Hunt. “To err (meiotically) is human: the genesis of human aneuploidy”. In: *Nature reviews. Genetics* 2.4 (2001), pp. 280–91.
- [17] S. E. Antonarakis et al. “Chromosome 21 and down syndrome: from genomics to pathophysiology”. In: *Nature Reviews. Genetics* 5.10 (2004), pp. 725–38.
- [18] M. Bugge et al. “Non-disjunction of chromosome 18”. In: *Human molecular genetics* 7.4 (1998), pp. 661–9.
- [19] N. E. Lamb et al. “Characterization of susceptible chiasma configurations that increase the risk for maternal nondisjunction of chromosome 21”. In: *Human molecular genetics* 6.9 (1997), pp. 1391–9.
- [20] W. P. Robinson et al. “Maternal meiosis I non-disjunction of chromosome 15: dependence of the maternal age effect on level of recombination”. In: *Human molecular genetics* 7.6 (1998), pp. 1011–9.
- [21] A. R. Savage et al. “Elucidating the mechanisms of paternal non-disjunction of chromosome 21 in humans”. In: *Human molecular genetics* 7.8 (1998), pp. 1221–7.
- [22] N. S. Thomas et al. “Maternal sex chromosome non-disjunction: evidence for X chromosome-specific risk factors”. In: *Human molecular genetics* 10.3 (2001), pp. 243–50.
- [23] N. E. Lamb et al. “Association between maternal age and meiotic recombination for trisomy 21”. In: *American journal of human genetics* 76.1 (2005), pp. 91–9.

- [24] G. Coop and M. Przeworski. “An evolutionary view of human recombination”. In: *Nature reviews. Genetics* 8.1 (2007), pp. 23–34.
- [25] G. Coop et al. “High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans”. In: *Science* 319.5868 (2008), pp. 1395–8.
- [26] E. S. Lander and P. Green. “Construction of multilocus genetic linkage maps in humans”. In: *Proceedings of the National Academy of Sciences of the United States of America* 84.8 (1987), pp. 2363–7.
- [27] R. Chowdhury et al. “Genetic analysis of variation in human meiotic recombination”. In: *PLoS genetics* 5.9 (2009), e1000648.
- [28] D. Y. Lin and D. Zeng. “Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data”. In: *Genetic epidemiology* 34.1 (2010), pp. 60–6.

## 2.0 GENETICS OF MEIOTIC RECOMBINATION: GENOME WIDE ASSOCIATION STUDIES FOR RECOMBINATION PHENOTYPES

### 2.1 ABSTRACT

GWAS and molecular studies have identified two genes that clearly play a role in human meiotic recombination. *RNF212* is associated with total or average recombination, and *PRDM9* is associated with recombination in historically defined hotspots. Other genes have been suggestively identified by GWAS, but not replicated. In this study we calculate new recombination phenotypes and use new datasets to attempt to identify additional recombination genes and to further study the effects of *PRDM9* and *RNF212*.

We used three data sets totaling of 3108 two-generation families, and estimated five different recombination phenotypes using dense SNP array genotype data. We then performed gender-specific and gender-pooled GWAS analyses. We replicated previous results for *RNF212* and *PRDM9*, and by looking specifically at recombination outside of hotspots we also showed that *PRDM9* appears to have somewhat different actions in males and females. We suggested several new potential candidate SNPs/genes, including *rs12186491* (chr 5), *rs10937651* (chr 4).

### 2.2 INTRODUCTION

Meiotic recombination is critical to successful human reproduction. It is a highly regulated stable segregation process to create gametes. It is also an important mechanism for ensuring

genetic diversity. Unlike somatic recombination, meiotic recombination involves homologous DNA sequences. Meiotic recombination initiates with double-strand breaks (DSB) of DNA and repairs on the homologous DNA sequence of the homologous chromosome [1]. Too little recombination, absence of recombination, and recombination in certain high-risk locations are all associated with aberrant meiotic outcomes including chromosomal aneuploidies. Most human embryonic aneuploidies originate from maternal gametes [2, 3, 4, 5, 6, 7]. Long term meiotic arrest in the female is considered as the cause [2]. Chromosomal aneuploidies include trisomy and monosomy and can result in pregnancy loss, mental retardation and different forms of disability.

Recombination is a highly variable phenomenon. There is gender specific variability and individual level variability at every scale [8]. Recently several studies have started to uncover the genetic determinants of meiotic recombination in humans using either direct laboratory approaches or computational approaches for scoring recombination events. Different studies have focused on different aspects of recombination, such as average number of recombinants, location and frequency of the recombination in different areas on the genome and different roles in males and females.

The most commonly studied recombination phenotype is average recombination count (ARC) over multiple gametes in a single proband (parent). One gene, *RNF212*, has been conclusively shown to affect overall recombination and inter-individual variation in ARC [8, 9, 10, 11]. Kong et. al. first reported the *RNF212* gene in a GWAS study conducted in an Icelandic population and showed that it has opposite effects on male and female recombination rate. This result was later replicated by other studies [9, 10, 11, 12, 13]. Other genes that have been putatively associated with ARC are *KIAA1462* in females and *UGCG* and *NUB1* in males [8, 9]. Beside genes, an inversion on chromosome 17q21.31 is also associated with female recombination rate [8, 9, 11].

Several studies have showed that in addition to the recombination rate, the location of recombination events is also important. Abnormal recombination location has been associated with improper chromosomal segregation [14, 15]. Based on historical information as represented in patterns of linkage disequilibrium, the frequency of recombination events is



higher in some locations of the genome. These 1-2 kb areas of the genome are known as hotspots [16, 17]. Usage of hotspot areas is determined by multiple factors such as presence of a particular motif in the hotspot regions, presence of epigenetic factors and trans-acting loci [18].

*PRDM9* has been shown in several recent studies to affect recombination in hotspots. Activity of *PRDM9* varies because of allelic variation, and the genotype may affect the genome wide hotspot usage [19, 20, 21, 22, 23]. The role of *PRDM9* is not limited to human recombination hotspot usage. A recent study showed that *PRDM9* is also involved with non-exchange gene conversion [24]. All of these findings suggest that there are other determinants still to explore to understand the whole mechanism of *PRDM9* and its role in human recombination and hotspot usage. The Zinc-finger region in *PRDM9* gene tends to bind to 13-bp or 17-bp motifs. Percent of recombination near the motif is another phenotype of interest for recombination study.

In summary, two genes *RNF212* and *PRDM9* and an inversion of 17q21.31 show clear evidence of effects on human recombination, and there are some additional suggested loci. More questions remain. For example, little is known about the effects of *RNF212* and *PRDM9* on other recombination phenotypes. Interaction between them has not been tested to our knowledge. And there are probably more genes associated with other recombination phenotypes yet to be discovered.

One of the major goals of our study is to find additional recombination genes by considering phenotypes that have not previously been used in GWAS studies, and by combining additional datasets. Another goal is to further investigate the roles of the already established recombination genes. To achieve these goals, we estimated several new phenotypes such as amount of recombination in the hotspot areas and non-hotspot areas and proportion of recombination with motif overlap along with previously-studied phenotypes such as average recombination rate and percent of hotspot usage using three data sets. It is of particular interest to characterize the genetic influence on recombination in the non-hotspot areas of the genome. Such studies may help explain the effects of *PRDM9* and the preference for using hotspot areas of the genome over non-hotspot areas. In addition, given that in 40%

of human hotspots a degenerate 13-bp motif was found [25] and the *PRDM9* gene tends to bind with this motif, we were interested to explore this phenotype as well, although the motif issue is controversial.

## 2.3 METHODS

### 2.3.1 Study Population and Samples

This study included three populations: the Geneva Dental Caries Study (GDCS) [26], the Autism Genetic Resource Exchange (AGRE) [27] and the Framingham Heart Study (FHS) [28]. The GDCS and AGRE samples were genotyped on the Illumina Human610-Quad Bead-chip and FHS samples were genotyped on the Affymetrix 5.0 chip. GDCS genotype data are available at the National Center for Biotechnology Information database of Genotype and Phenotype (dbGaP). So are FHS. And AGRE is available at the science program of autism speaks database (<https://research.agre.org/>).

### 2.3.2 Pedigrees

Two-generation nuclear pedigrees with two or more children were used for this study: 171 from GDCS, 737 from AGRE, and 654 from FHS. Genotype data on each family were used to score recombination for the parents. The parents were then used as the subjects for the GWAS analyses.

### 2.3.3 Phenotypes

Recombination events in each nuclear family were called according to the method described in Chowdhury R et.al [9]. Briefly, the method is as follows. First, a set of informative markers was identified in each family. A locus is informative if one parent is homozygous and another is heterozygous. Among two or more children, one is considered as reference

child, and in a sibling pair a switch from one allele to another allele in a particular parental haplotype as we move along the chromosome indicates a recombination in that parent with heterozygous allele. For more accurate estimation, we used 5 or more consecutive markers to call a recombination event.

From the recombination data, we calculated five different recombination phenotypes: average recombination count (ARC), percent of recombination occurring in hotspots (HS\_PCT), count of recombination in hotspots (HS\_CNT), count of recombination in non-hotspot regions (NHS\_CNT) and percent of recombination occurring near the motif (MOTIF). A set of predefined historic hotspot regions was used to calculate the three phenotypes related to hotspots: HS\_PCT, HS\_CNT and NHS\_CNT. Precise definitions of these phenotypes are as follows:

#### **I Average Recombination Count (ARC)**

$$\text{ARC} = (\text{total recombination in all children of the proband} / \text{number of children})$$

#### **II Percent of Recombination occurring in Hotspots (HS\_PCT)**

$$\text{HS\_PCT} = (\text{total number of recombination overlapping hotspots in all children of the proband}) / (\text{total recombination in all children of the proband})$$

#### **III Absolute Count of Recombination occurring in Hotspots (HS\_CNT)**

$$\text{HS\_CNT} = (\text{total number of recombination overlapping hotspots in all children of the proband}) / (\text{Number of children})$$

#### **IV Absolute Count of Recombination occurring in Non-Hotspots (NHS\_CNT)**

$$\text{NHS\_CNT} = (\text{total number of recombination overlapping non-hotspots in all children of the proband}) / (\text{Number of children})$$

#### **V MOTIF**

$$\text{MOTIF} = (\text{total number of recombination with motif in all children of the proband}) / (\text{Total recombination events in all children})$$

### 2.3.4 Genotypes, Error Checking and Data Handling

For GDCS, 589,735 SNPs were released by Center for Inherited Disease Research (CIDR). Our data sets also included 520,018 SNPs for AGRE and 388,060 SNPs for FHS. To ensure the quality, an extensive data cleaning was performed for the data sets. Full details of data cleaning steps for GDCS can be found in Geneva consortium website (<http://www.genevastudy.org/>). Briefly, measures of identity-by-descent were used to verify relationships, SNP intensities of X and Y-chromosomes were used to verify gender, and principal component analysis (PCA) was used to examine genetic ancestry. Two thresholds used in the analysis are a Hardy-Weinberg disequilibrium cut-off of 0.0001 and minimum minor allele frequency cut-off of 0.02 for all SNPs.

### 2.3.5 Genome-wide Association Studies

To identify genes or SNPs associated with different aspects of recombination, we conducted three genome-wide association studies for each phenotype; we conducted separate male and female analyses as well as performing a combined analysis. We used PLINK (29) to conduct all GWAS using an additive genetic model. All of our phenotypes are continuous; so we used the linear regression option in PLINK for the association tests. As per significance level of association studies, we used the threshold with p-values less than  $e^{-08}$  as significant and p-value between  $e^{-06}$  to  $e^{-08}$  as suggestive regions or SNPs.

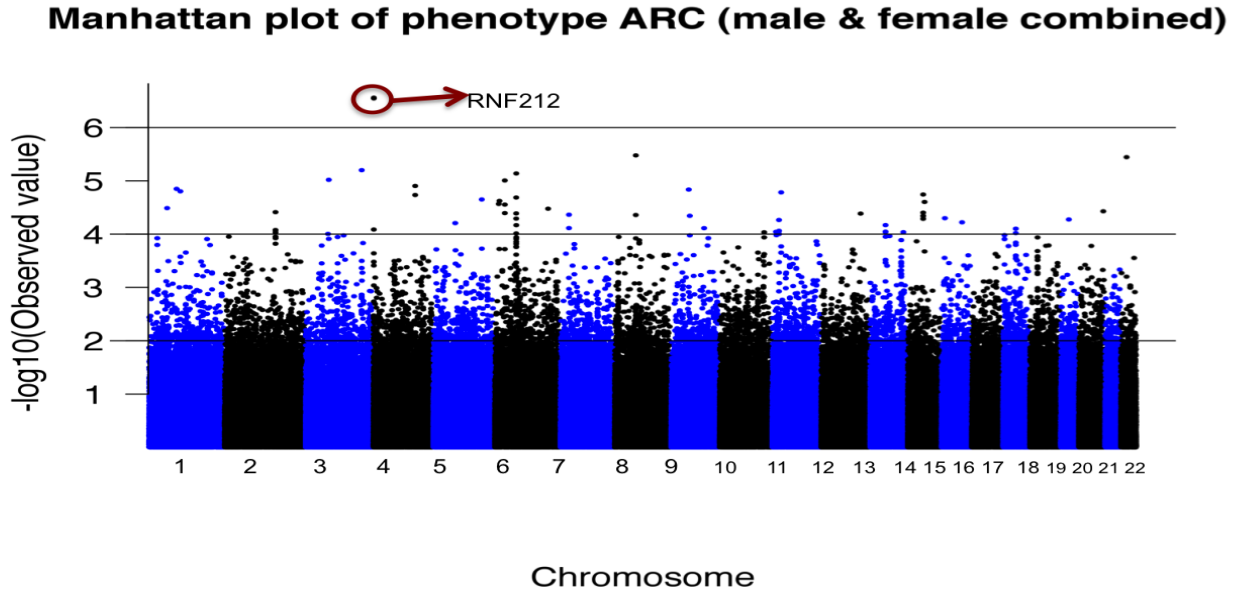
We combined the AGRE and GDCS GWAS using GWAS meta-analysis instead of combining all three data sets. The AGRE and GDCS data sets were genotyped on the same platform (Illumina 610 chip) and the FHS data set was genotyped on the Affymetrix 5.0 chip. Because of this platform difference, the SNP sets are very different with minimal overlap. We used fixed effects meta-analysis to combine the GDCS and AGRE data sets. We performed GWAS meta-analysis for each sex separately and also performed gender-pooled GWAS meta-analysis. The software METAL [29] was used to do the GWAS meta-analysis. We then used the FHS dataset for qualitative replication in regions nominated by the meta-analyses.

## 2.4 RESULTS

### 2.4.1 Recombination Phenotypes

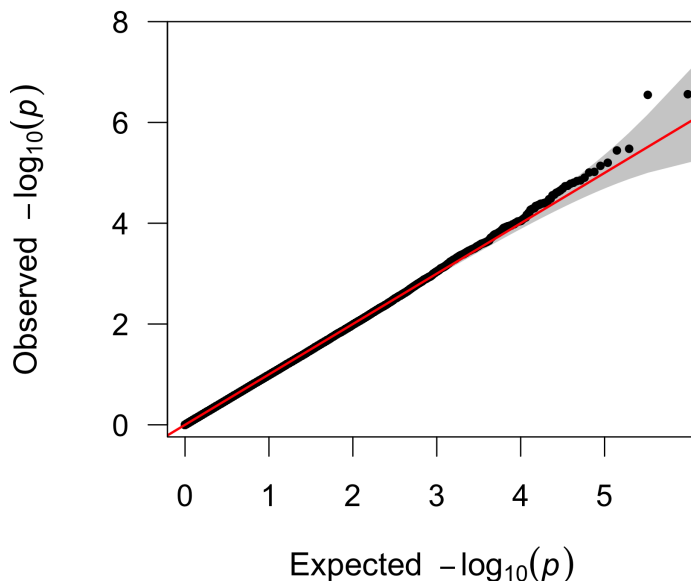
We estimated five different phenotypes based on these recombination scores. Our findings for each phenotype are presented below.

**2.4.1.1 Average Recombination Count (ARC)** The Manhattan plot of the combined meta-analysis is presented in figure 2.1 and the QQ plot of the same analysis is presented in figure 2.2. Manhattan plots of the sex-specific meta-analysis results are presented in supplementary documents in appendix A. Results of the sex-specific and combined meta-analysis of GDCS and AGRE data sets are presented in table 2.1.



**Figure 2.1:** Manhattan plot of phenotype ARC (combined analysis)

The top ten most highly associated SNPs are listed in table 2.1, which also includes nearby flanking genes in each region. The *RNF212* gene barely meets the GWAS threshold in the combined analysis. In the male only analysis, *RNF212* is the most significant gene ( $p = 1.695e^{-08}$ ) associated with average recombination count. Reviewing the effect size of *RNF212*, male and female have effect in opposite directions, which is consistent with the previous literature.



**Figure 2.2:** QQ plot of phenotype ARC (combined analysis)

Other than *RNF212*, none of the previously reported candidate genes by Chowdhury et al and Fledel-Alon et. al. meet the genome-wide association test threshold in our study. Previously reported genes for male average recombination counts are *NUB1* and *UGCG* and for female average recombination counts are *PDZK1*, *KIAA1462*, *CRHR1*, *LRRC37A*. Locus zoom plots for these gene regions are presented in appendix A. We also looked at the previously reported signals in our male and female combined analysis for this phenotype. None of the SNPs (*rs17011067*, *rs1864309*, *rs16972342*, *rs7284619*) were significantly associated with pooled-gender analysis too.

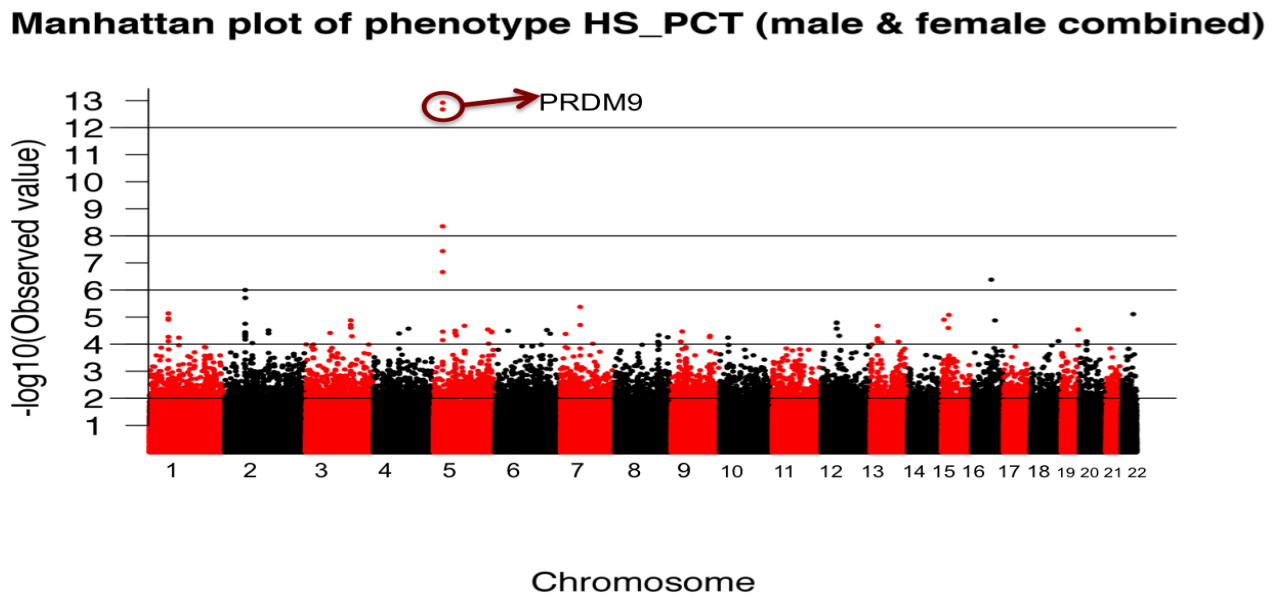
**Table 2.1:** SNPs with lowest P-values for ARC

Type	SNP	Chr	BP	P.value	Direction	Gene.list
Combined	<i>rs4974601</i>	4	1085409	2.76E-07	- - + -	<i>RNF212</i>
Combined	<i>rs444996</i>	8	40298364	3.34E-06	++++	<i>C8orf4, ZMAT4</i>
Combined	<i>rs724055</i>	22	29005922	3.60E-06	- +++	<i>LIF, OSM, GATSL3, TBC1D10A, SF3A1, CCDC157, SEC14L2, MTP18, HORMAD2</i>
Combined	<i>rs1996483</i>	3	167607627	6.31E-06	- +++	<i>chr3:167107628, 168107628</i>
Combined	<i>rs9381359</i>	6	45098602	7.28E-06	++++	<i>SUPT3H, MIR586</i>
Combined	<i>rs9311748</i>	3	60064633	9.58E-06	++++	<i>FHIT</i>
Combined	<i>rs4134943</i>	6	20591385	9.85E-06	++++	<i>E2F3, CDKAL1</i>
Combined	<i>rs11932615</i>	4	139748182	1.25E-05	- - - -	<i>SLC7A11, CCRN4L, ELF2</i>
Combined	<i>rs10493733</i>	1	83209124	1.42E-05	++++	<i>chr1:82709125, 83709125</i>
Combined	<i>rs1033753</i>	14	59463128	1.81E-05	+ - - -	<i>RTN1, C14orf135, DHRS7</i>
Female	<i>rs497793</i>	3	154948531	3.47E-07	++	<i>C3orf79, SGEF</i>
Female	<i>rs12903708</i>	15	58380596	1.13E-06	++	<i>FOXB1, ANXA2, NARG2</i>
Female	<i>rs2974754</i>	19	12922982	2.43E-06	++	<i>FARSA, DAND5, CALR, RAD23A</i>
Female	<i>rs4879584</i>	9	32402621	3.26E-06	++	<i>ACO1, DDX58</i>
Female	<i>rs9572559</i>	13	70310774	3.79E-06	- -	<i>chr13:69810775, 70810775</i>
Female	<i>rs11721955</i>	4	160266014	4.53E-06	- -	<i>C4orf45, RAPGEF2</i>
Female	<i>rs3791936</i>	2	218476666	6.40E-06	++	<i>TNS1</i>
Female	<i>rs235987</i>	16	69806846	7.58E-06	++	<i>HYDIN, FTSJD1, CALB2</i>
Male	<i>rs12645644</i>	4	1044158	7.56E-07	- -	<i>IDUA, RNF212, DGKQ, TMEM175</i>
Male	<i>rs1951371</i>	14	59425467	4.69E-06	- -	<i>RTN1</i>
Male	<i>rs1996483</i>	3	167607627	4.84E-06	++	<i>chr3:167107628, 168107628</i>
Male	<i>rs1418433</i>	6	44860545	8.68E-06	++	<i>SUPT3H, SPATS1, AARS2</i>
Male	<i>rs1035699</i>	11	19713338	9.80E-06	- -	<i>NAV2, LOC100126784</i>
Male	<i>rs10493733</i>	1	83209125	1.08E-05	++	<i>chr1:83009125, 83409125</i>
Male	<i>rs2061037</i>	11	8223204	2.08E-05	++	<i>LMO1, RIC3</i>
Male	<i>rs724055</i>	22	29005922	2.09E-05	++	<i>LIF, OSM, GATSL3, SF3A1, CCDC157, SEC14L2</i>

Column 6 of the table represents the direction of the effect size of each SNP presented in column 2 in each study. In combined analysis, studies were included in the following order (GDCS female, GDCS male, AGRE female and AGRE male). In female only analysis, first position in the direction column is for GDCS female and the 2nd position is for AGRE female and same ordering is used in male only analysis and for rest of the phenotypes.

### 2.4.1.2 Hotspot analysis: Percent of recombination occurring in hotspots (HS\_PCT)

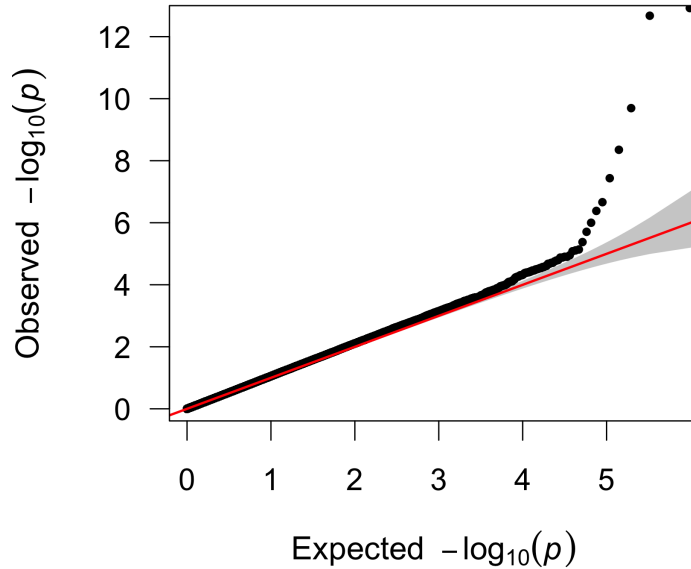
Meta-analysis for males and females can be found in appendix A. Figure 2.3 presents the Manhattan plot of meta-analysis of male and female combined along with the QQ plot in figure 2.4. The QQ plot shows no departure from the expected uniform distribution of p-values.



**Figure 2.3:** Manhattan plot of phenotype HS\_PCT (combined analysis)

For the HS\_PCT phenotype, the top signals for male only, female only and pooled-gender analysis are presented in Table 2.2. The strongest association (p.value:  $1.20e-13$ ) is with multiple SNPs in and near the *PRDM9* gene in the pooled-gender analysis. In the male and female separate analyses, *PRDM9* has a stronger association with female HS\_PCT than with male in terms of p-values. In the female only analysis, the strongest association is in *PRDM9* and it meets the genome-wide association threshold. But in male only analysis, significance of the SNP on *PRDM9* does not meet the genome-wide threshold. Other than *PRDM9*, none of the other top hits meet genome-wide association threshold. But we suggested some of the regions with significance level of in order of  $e-6$  to  $e-7$ .





**Figure 2.4:** QQ plot of phenotype HS\_PCT (male and female combined analysis)

### 2.4.1.3 Hotspot analysis: Average count of recombinants in hotspots (HS\_CNT)

Our third phenotype was the count of recombination events in hotspot areas. Manhattan plots for the phenotype HS\_CNT for pooled-gender analysis and sex-specific analysis are presented in supplementary documents. Table 2.3 shows the top 10 hits for sex-specific and combined meta-analysis. Unlike HS\_PCT, males have stronger effect of *PRDM9* on HS\_CNT than females. But in combined pooled-gender analysis, *PRDM9* is the strongest hit. Other suggestive SNPs for HS\_CNT have very minimal overlap with the suggested SNPs for HS\_PCT. Among the top hits for the male only analysis of HS\_CNT is *RNF212*.

### 2.4.1.4 Non-Hotspot analysis: Average count of recombinants in non-hotspot areas (NHS\_CNT)

In the non-hotspot analysis we looked at the average count of recombinants in non-hotspot areas of the genome. The Manhattan plots of the pooled-gender analysis and sex-specific analyses are presented in Appendix A. An unresolved concern is

**Table 2.2:** SNPs with lowest P-values for HS\_PCT

Type	SNP	Chr	BP	P.value	Direction	Gene.list
Combined	<i>rs1603084</i>	5	23567950	1.20E-13	----	<i>PRDM9</i>
Combined	<i>rs12445855</i>	16	68068843	4.16E-07	----	<i>CYB5B, MIR1538, TERF2, NFAT5</i>
Combined	<i>rs972847</i>	2	50227778	1.00E-06	++++	<i>NRXN</i>
Combined	<i>rs13232367</i>	7	43342734	4.20E-06	++++	<i>HECW1</i>
Combined	<i>rs2716140</i>	1	59244984	7.29E-06	++++	<i>LOC729467, JUN</i>
Combined	<i>rs9614870</i>	22	43448073	7.79E-06	----	<i>NCRNA00207, PRR5, ARHGAP8, PRR5, ARHGAP8</i>
Combined	<i>rs16971454</i>	15	39148429	8.32E-06	++++	<i>INO80</i>
Combined	<i>rs11710141</i>	3	132632955	1.32E-05	-+--	<i>NUDT16P1, MRPL3, NUDT16, SNORA58</i>
Combined	<i>rs12581285</i>	12	38795607	1.60E-05	----	<i>SLC2A13, LRRK2</i>
Female	<i>rs1603084</i>	5	23567950	2.54E-09	--	<i>PRDM9</i>
Female	<i>rs12445855</i>	16	68068843	4.64E-07	--	<i>CYB5B, MIR1538, NFAT5</i>
Female	<i>rs949029</i>	18	50885623	1.32E-05	++	<i>CCDC68, RAB27B, TCF4</i>
Female	<i>rs355926</i>	16	65270601	1.36E-05	--	<i>CMTM4, DYNC1LI2, CCDC79</i>
Female	<i>rs2292305</i>	15	37668113	1.58E-05	--	<i>THBS1, FSIP1</i>
Female	<i>rs6744522</i>	2	29302926	1.68E-05	--	<i>ALK, CLIP4, C2orf71</i>
Female	<i>rs7107498</i>	11	66476056	2.08E-05	++	<i>RCE1, PC, C11orf86, SYT12, RHOD</i>
Female	<i>rs10906326</i>	10	13319832	2.39E-05	++	<i>OPTN, MCM10, UCMA, , PHYH, SEPHS1</i>
Female	<i>rs972847</i>	2	50227778	2.55E-05	++	<i>NRXN1</i>
Female	<i>rs760954</i>	6	143767362	3.58E-05	--	<i>AIG1, ADAT2, PEX3, FUCA2</i>
Female	<i>rs4812661</i>	20	41008536	3.69E-05	--	<i>PTPRT</i>
Male	<i>rs10996809</i>	10	67413658	2.91E-06	++	<i>CTNNA3</i>
Male	<i>rs12958111</i>	18	71979757	6.88E-06	++	<i>ZNF516</i>
Male	<i>rs1874165</i>	5	23559104	7.59E-06	--	<i>PRDM9</i>
Male	<i>rs13378443</i>	13	92254975	8.22E-06	++	<i>GPC5, GPC6</i>
Male	<i>rs1603084</i>	5	23567950	9.79E-06	--	<i>PRDM9</i>
Male	<i>rs7581691</i>	2	13953004	1.44E-05	--	<i>chr2:13453005, 14453005</i>
Male	<i>rs10514277</i>	5	85666780	1.76E-05	--	<i>NBPF22P, COX7C</i>
Male	<i>rs2380707</i>	2	15930080	1.81E-05	++	<i>DDX1, MYCNOS, MYCN</i>
Male	<i>rs4876993</i>	9	91471222	2.36E-05	++	<i>GADD45G, LOC100129066</i>
Male	<i>rs175853</i>	6	151287006	2.53E-05	++	<i>PLEKHG1, MTHFD1L</i>
Male	<i>rs4461048</i>	15	36835522	3.08E-05	++	<i>C15orf53, C15orf54</i>

**Table 2.3:** SNPs with lowest P-values for HS\_CNT

Type	SNP	Chr	BP	P.value	Direction	Gene.list
Combined	<i>rs1874165</i>	5	23559104	3.80E-08	----	<i>PRDM9</i>
Combined	<i>rs2764928</i>	1	59195376	3.69E-07	++++	<i>JUN, LOC729467</i>
Combined	<i>rs7650855</i>	3	73602421	1.47E-06	++++	<i>PDZRN3</i>
Combined	<i>rs16863103</i>	2	15918176	1.77E-06	++++	<i>DDX1, MYCNOS, MYCN</i>
Combined	<i>rs13253524</i>	8	119294947	1.88E-06	----	<i>EXT1, SAMD12</i>
Combined	<i>rs2715252</i>	NA	107802544	4.01E-06	----	<i>chr3:107302545, 108302545</i>
Combined	<i>rs932770</i>	1	59256193	8.69E-06	++++	<i>JUN, LOC729467</i>
Combined	<i>rs7169146</i>	15	62088436	9.21E-06	----	<i>DAPK2, FAM96A, SNX1</i>
Female	<i>rs1242541</i>	14	82275789	2.87E-06	++	<i>chr14:81775790, 82775790</i>
Female	<i>rs2959776</i>	8	6415275	5.40E-06	++	<i>MCPH1, ANGPT2, AGPAT5</i>
Female	<i>rs2569491</i>	19	56276727	8.57E-06	++	<i>KLK12, KLK13, KLK14, CTU1, SIGLEC9, SIGLEC7, SIGLECP3</i>
Female	<i>rs6720182</i>	2	68848001	1.24E-05	--	<i>PROKR1, ARHGAP25, BMP10</i>
Female	<i>rs4797343</i>	18	8964854	1.47E-05	++	<i>KIAA0802, NDUFV2</i>
Female	<i>rs4881291</i>	10	4141050	1.50E-05	++	<i>KLF6</i>
Female	<i>rs749052</i>	2	232504853	2.09E-05	++	<i>MIR1471, NPPC, DIS3L2</i>
Female	<i>rs13263626</i>	8	135446671	2.16E-05	++	<i>ZFAT, ZFATAS</i>
Female	<i>rs2764928</i>	1	59195376	2.42E-05	++	<i>JUN, LOC729467</i>
Female	<i>rs6018718</i>	20	45880733	2.68E-05	--	<i>NCOA3, SULF2</i>
Female	<i>rs6985596</i>	8	135471297	2.82E-05	++	<i>ZFAT, ZFATAS</i>
Male	<i>rs10958702</i>	8	41865459	2.06E-06	--	<i>NKX6, 3, ANK1, MIR486, MYST3</i>
Male	<i>rs13378443</i>	13	92254975	3.51E-06	++	<i>GPC5, GPC6</i>
Male	<i>rs169266</i>	1	167090734	4.47E-06	++	<i>DPT, MGC4473, ATP1B1</i>
Male	<i>rs1874165</i>	5	23559104	4.62E-06	--	<i>PRDM9</i>
Male	<i>rs325702</i>	11	6216076	4.87E-06	++	<i>OR56B4, OR52B2, OR52W1, FAM160A2, PRKCDBP</i>
Male	<i>rs1558638</i>	7	158571551	5.15E-06	++	<i>LOC154822, VIPR2</i>
Male	<i>rs1347322</i>	8	103617737	5.98E-06	++	<i>UBR5, ODF1, KLF10</i>
Male	<i>rs1502800</i>	12	85000385	1.34E-05	++	<i>MGAT4C</i>
Male	<i>rs16863103</i>	2	15918176	1.35E-05	++	<i>DDX1, MYCNOS, MYCN</i>
Male	<i>rs7650855</i>	3	73602421	1.46E-05	++	<i>PDZRN3</i>
Male	<i>rs2045065</i>	4	1042487	1.73E-05	--	<i>IDUA, FGFR1, RNF212</i>

a striking deviation in QQ plot from the expected distribution. The top 10 hits of each analysis are presented in Table 2.4.

**2.4.1.5 Percent of recombination occurring near the motif (MOTIF)** As our last phenotype we looked at the percent of recombination occurring near the 13 base-pair motif. Table 2.5 listed top hits in each analysis. None of the SNPs met genome-wide threshold. But there are some suggestive SNPs in each category of analysis.

## 2.4.2 In-depth Analysis of *RNF212*, *PRDM9* and MATP

To gain insight into the roles of the two well-established recombination genes; *PRDM9* and *RNF212*, we looked at the association results across different phenotypes and also investigated the possible interaction between two genes.

**2.4.2.1 *PRDM9* gene across phenotypes** The *PRDM9* gene association results for different are presented in the following Table 2.6. For pooled-sex analysis, *PRDM9* has the smallest p-values for association with HS\_PCT followed by HS\_CNT. Both male and female effect sizes have the same direction. In NHS\_CNT, *PRDM9* has borderline genome-wide significance. *PRDM9* shows no evidence of association with the average recombination count and MOTIF phenotypes.

**Table 2.4:** SNPs with lowest P-values for NHS\_CNT

Type	SNP	Chr	BP	P.value	Direction	Gene.list
Combined	<i>rs12186491</i>	5	147573689	6.36E-08	++++	<i>SPINK5L2, SPINK6, SPINK5L3, SPINK7, SPINK9</i>
Combined	<i>rs2914263</i>	5	23488680	1.16E-07	++++	<i>PRDM9</i>
Combined	<i>rs10937651</i>	4	5596712	1.65E-07	++++	<i>STK32B, C4orf6, EVC2</i>
Combined	<i>rs7403622</i>	15	31977777	2.16E-07	++++	<i>RYR3, AVEN</i>
Combined	<i>rs11966986</i>	6	56628268	3.19E-07	++++	<i>DST</i>
Combined	<i>rs2289682</i>	4	82251226	5.05E-07	++++	<i>BMP3, PRKG2</i>
Combined	<i>rs9572544</i>	13	70230937	5.58E-07	++++	<i>chr13:69730938, 70730938</i>
Combined	<i>rs497083</i>	5	163073219	5.68E-07	----	<i>CCNG1, NUDCD2, HMMR</i>
Combined	<i>rs9510171</i>	13	21858129	8.76E-07	++++	<i>chr13:21358130, 22358130</i>
Combined	<i>rs9582126</i>	13	96778958	1.12E-06	++++	<i>MBNL2 [96278959, 97278959]</i>
Combined	<i>rs1801449</i>	15	40468490	1.27E-06	++++	<i>GANC, CAPN3, ZFP106</i>
Combined	<i>rs5998881</i>	22	32060574	1.55E-06	++++	<i>LARGE</i>
Female	<i>rs3129595</i>	13	21458281	2.28E-06	++	<i>FGF9</i>
Female	<i>rs7873463</i>	9	4211297	3.23E-06	++	<i>GLIS3</i>
Female	<i>rs2065079</i>	14	50320526	4.22E-06	++	<i>SAV1, NIN, ABHD12B, PYGL</i>
Female	<i>rs1861509</i>	2	205885994	4.65E-06	++	<i>PARD3B</i>
Female	<i>rs1571463</i>	20	54859767	5.88E-06	++	<i>TFAP2C, BMP7</i>
Female	<i>rs1603084</i>	5	23567950	6.29E-06	++	<i>PRDM9</i>
Female	<i>rs7594732</i>	2	205879950	6.54E-06	++	<i>PARD3B</i>
Female	<i>rs2539978</i>	2	63048682	7.79E-06	++	<i>EHBP1</i>
Female	<i>rs9372446</i>	6	116118619	1.10E-05	+ -	<i>FRK, LOC728402</i>
Male	<i>rs10937651</i>	4	5596712	5.16E-08	++	<i>STK32B, C4orf6, EVC2</i>
Male	<i>rs11966986</i>	6	56628268	7.41E-07	++	<i>DST</i>
Male	<i>rs6994475</i>	8	1260832	1.67E-06	++	<i>DLGAP2</i>
Male	<i>rs7900873</i>	10	14903869	2.30E-06	++	<i>CDNF, HSPA14, SUV39H2</i>
Male	<i>rs1795514</i>	12	79856997	2.56E-06	++	<i>LIN7A, MIR617, MIR618</i>
Male	<i>rs10514277</i>	5	85666780	2.62E-06	++	<i>NBPF22P</i>
Male	<i>rs4489957</i>	1	92002931	3.47E-06	++	<i>chr15:91502932, 92502932</i>
Male	<i>rs1473500</i>	6	168253864	7.93E-06	++	<i>HGC6.3, KIF25, FRMD1</i>
Male	<i>rs9845811</i>	3	35945886	9.67E-06	++	<i>ARPP21, MIR128, 2</i>
Male	<i>rs6975631</i>	7	8349419	1.08E-05	++	<i>ICA1, NXP1</i>

**Table 2.5:** SNPs with lowest P-values for MOTIF

Type	SNP	Chr	BP	P.value	Direction	Gene.list
Combined	<i>rs4331859</i>	5	179026713	7.49E-07	----	<i>RUFY1, HNRNPH1, C5orf60, CBY3, MAML1</i>
Combined	<i>rs10872388</i>	6	132417711	5.58E-06	----	<i>CTGF, MOXD1</i>
Combined	<i>rs1112898</i>	7	1784522	7.02E-06	++++	<i>ELFN1, MAD1L1</i>
Combined	<i>rs17746897</i>	4	54930941	8.73E-06	----	<i>PDGFRA, KIT</i>
Combined	<i>rs1854226</i>	13	97036242	9.01E-06	----	<i>RAP2A, IPO5</i>
Combined	<i>rs10434879</i>	6	121716035	9.66E-06	++++	<i>C6orf170, GJA1</i>
Combined	<i>rs1494651</i>	5	32976595	9.68E-06	++++	<i>NPR3, C5orf23</i>
Female	<i>rs6886928</i>	5	166916702	4.38E-06	++	<i>ODZ2</i>
Female	<i>rs6728479</i>	2	644471	8.25E-06	++	<i>TMEM18</i>
Female	<i>rs7193684</i>	16	8046983	8.38E-06	++	<i>A2BP1</i>
Female	<i>rs6055249</i>	20	7602896	9.11E-06	--	<i>HAO1</i>
Female	<i>rs6487429</i>	12	24885298	1.09E-05	++	<i>BCAT1, DAD1L</i>
Male	<i>rs1336628</i>	13	18836533	2.95E-07	+ -	<i>TUBA3C, LOC100101938, TPTE2</i>
Male	<i>rs136809</i>	22	38350821	5.88E-06	--	<i>CACNA1I</i>
Male	<i>rs10882205</i>	10	95022650	8.11E-06	++	<i>CYP26C1, CYP26A1, MYOF</i>
Male	<i>rs11049351</i>	12	9234852	1.05E-05	--	<i>PZP, LOC642846</i>
Male	<i>rs11645438</i>	16	47051221	1.18E-05	--	<i>LONP2, SIAH1, N4BP1</i>
Male	<i>rs1705665</i>	14	83448121	1.27E-05	++	<i>chr14:82948122, 83948122</i>

**Table 2.6:** *PRDM9* gene association across phenotypes

Phenotype	Type	SNP	Chr	BP	Effect	StdErr	P.value	Direction
HS_PCT	Male	<i>rs2914263</i>	5	23488680	-0.0271	0.0084	0.001355	--
		<i>rs1874165</i>	5	23559104	-0.0431	0.0096	7.59E-06	--
		<i>rs1603084</i>	5	23567950	-0.0427	0.0097	9.794e-06	--
HS_PCT	Female	<i>rs2914263</i>	5	23488680	-0.0266	0.0065	4.594e-05	--
		<i>rs1874165</i>	5	23559104	-0.044	0.0076	5.915e-09	--
		<i>rs1603084</i>	5	23567950	-0.0454	0.0076	2.54E-09	--
HS_PCT	Combined	<i>rs2914263</i>	5	23488680	-0.0268	0.0052	2.174e-07	----
		<i>rs1874165</i>	5	23559104	-0.0436	0.0059	2.115e-13	----
		<i>rs1603084</i>	5	23567950	-0.0443	0.006	1.20E-13	----
HS_CNT	Male	<i>rs2914263</i>	5	23488680	-0.1363	0.1057	0.1969	--
		<i>rs1874165</i>	5	23559104	-0.5473	0.1195	4.62E-06	--
		<i>rs1603084</i>	5	23567950	-0.5419	0.1202	6.559e-06	--
HS_CNT	Female	<i>rs2914263</i>	5	23488680	-0.1197	0.1357	0.3775	--
		<i>rs1874165</i>	5	23559104	-0.4826	0.1577	0.00221	--
		<i>rs1603084</i>	5	23567950	-0.4849	0.1590	0.002288	--
HS_CNT	Combined	<i>rs2914263</i>	5	23488680	-0.1301	0.0834	0.1187	----
		<i>rs1874165</i>	5	23559104	-0.5237	0.0952	3.80E-08	----
		<i>rs1603084</i>	5	23567950	-0.5211	0.0959	5.48e-08	----
NHS_CNT	Male	<i>rs2914263</i>	5	23488680	0.4110	0.1180	0.000498	++
		<i>rs1874165</i>	5	23559104	0.2264	0.1354	0.09446	++
		<i>rs1603084</i>	5	23567950	0.2269	0.1352	0.09336	++
NHS_CNT	Female	<i>rs2914263</i>	5	23488680	0.6277	0.1511	3.284e-05	++
		<i>rs1874165</i>	5	23559104	0.7679	0.1768	1.406e-05	++
		<i>rs1603084</i>	5	23567950	0.8056	0.1784	6.29E-06	++
NHS_CNT	Combined	<i>rs2914263</i>	5	23488680	0.4931	0.093	1.16E-07	++++
		<i>rs1874165</i>	5	23559104	0.4265	0.1075	7.247e-05	++++
		<i>rs1603084</i>	5	23567950	0.4381	0.1078	4.792e-05	++++
ARC	Male	<i>rs2914263</i>	5	23488680	0.4802	0.2495	0.05422	++
		<i>rs1874165</i>	5	23559104	-0.1448	0.2870	0.6139	--
		<i>rs1603084</i>	5	23567950	-0.1253	0.2829	0.6579	--
ARC	Female	<i>rs2914263</i>	5	23488680	0.5318	0.4582	0.2458	++
		<i>rs1874165</i>	5	23559104	0.2609	0.5365	0.6268	-+
		<i>rs1603084</i>	5	23567950	0.3037	0.5412	0.5746	-+
ARC	Combined	<i>rs2914263</i>	5	23488680	0.4920	0.2191	0.02472	++++
		<i>rs1874165</i>	5	23559104	-0.0545	0.2531	0.8294	--+-
		<i>rs1603084</i>	5	23567950	-0.0332	0.2507	0.8947	--+-
MOTIF	Male	<i>rs2914263</i>	5	23488680	-0.0108	0.0077	0.16	--
		<i>rs1874165</i>	5	23559104	-0.0027	0.0087	0.7561	+ -
		<i>rs1603084</i>	5	23567950	-0.0027	0.0088	0.7596	+ -
MOTIF	Female	<i>rs2914263</i>	5	23488680	0.0018	0.0064	0.7736	++
		<i>rs1874165</i>	5	23559104	-0.0126	0.0075	0.09211	--
		<i>rs1603084</i>	5	23567950	-0.0138	0.0076	0.06883	--
MOTIF	Combined	<i>rs2914263</i>	5	23488680	-0.0033	0.0049	0.4967	+ - + -
		<i>rs1874165</i>	5	23559104	-0.0084	0.0057	0.1387	- + - -
		<i>rs1603084</i>	5	23567950	-0.0091	0.0057	0.1141	- + - -

**2.4.2.2 *RNF212* gene across phenotypes** Table 2.7 presents the *RNF212* association p-values across. *RNF212* has the strongest association with male average recombination rate. Interestingly the *RNF212* gene also shows association with male HS\_CNT in the order of e-5. None of the other association p-values with other phenotypes are smaller than e-5.

**2.4.2.3 Chromosome 17 inversion region across phenotypes** The *MAPT* gene is on chromosome 17 in the center of the inversion region, and we looked at the association results for all the phenotypes for both sexes using that gene as the locus in LocusZoom plots. SNPs in the region appear to be associated with all phenotypes in females but not in males. Different SNPs in the region are associated with different phenotypes, though. All of the locus zoom plots for the *MAPT* gene across phenotypes are presented in appendix A.

**2.4.2.4 Gene X Gene interaction models** Looking at the detail analysis of two genes *RNF212* and *PRDM9*, we were motivated to investigate the possible interaction between these two genes. We used only GDCS data set for that purpose, since we have the genotype data available for GDCS only. We selected 2 SNPs; *rs1603084*( *PRDM9*) and *rs4974601*( *RNF212*), which showed association across phenotypes and performed regression analysis using additive and dominant model with and without interaction effects. Because of the small sample size some of the categories had very few data or no data at all. So we were unable to effectively test for interaction.

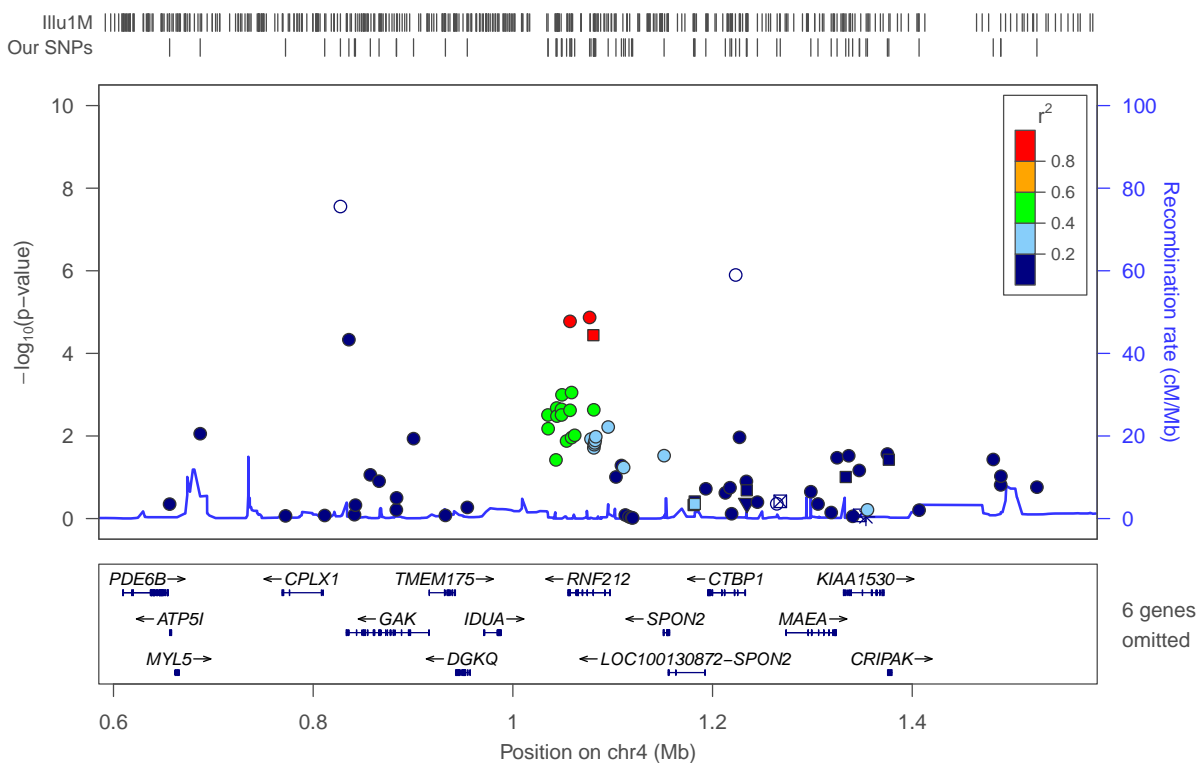
### **2.4.3 FHS Replication**

To support our findings of new genes and suggested genes for each of the phenotypes, we examined approximately 140 regions of interest in the FHS data set. We compared male only analysis with FHS male GWAS results and female only analysis with FHS female GWAS results. To compare pooled-gender analysis we combined FHS male and female using fixed effect meta-analysis and then compared. Since the FHS data set and the two other data sets have limited SNP overlap, we performed this replication analysis at the gene level.



**Table 2.7:** *RNF212* gene association across phenotypes

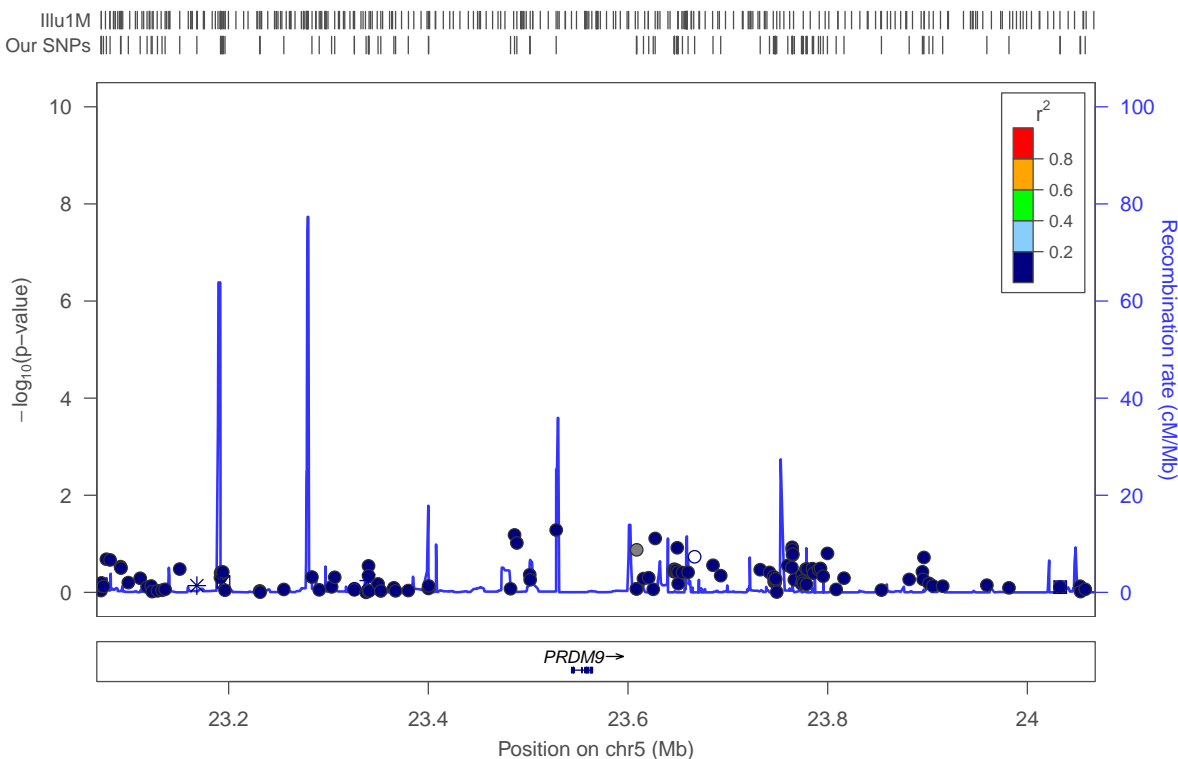
Phenotype	Type	SNP	Chr	BP	Effect	StdErr	P.value	Direction
ARC	Male	<i>rs12645644</i>	4	1044158	-0.9774	0.1976	7.56E-07	--
		<i>rs4974601</i>	4	1085409	-0.9625	0.1706	1.695e-08	--
ARC	female	<i>rs12645644</i>	4	1044158	0.5112	0.3652	0.1616	++
		<i>rs4974601</i>	4	1085409	-0.1028	0.3243	0.7512	-+
ARC	Combined	<i>rs4974601</i>	4	1085409	-0.7761	0.151	2.76E-07	--+-
		<i>rs12645644</i>	4	1044158	-0.6403	0.1738	0.0002292	+--+
HS_PCT	Male	<i>rs12645644</i>	4	1044158	-0.0183	0.0068	0.006948	--
		<i>rs4974601</i>	4	1085409	-0.0126	0.0059	0.03188	--
HS_PCT	female	<i>rs12645644</i>	4	1044158	0.0024	0.0052	0.6452	-+
		<i>rs4974601</i>	4	1085409	0.0047	0.0046	0.314	++
HS_PCT	Combined	<i>rs12645644</i>	4	1044158	-0.0053	0.0041	0.1974	--+-
		<i>rs4974601</i>	4	1085409	-0.002	0.0036	0.5894	+--+
HS_CNT	Male	<i>rs12645644</i>	4	1044158	-0.3564	0.0837	2.044e-05	--
		<i>rs4974601</i>	4	1085409	-0.2234	0.0730	0.002205	--
HS_CNT	female	<i>rs12645644</i>	4	1044158	0.0253	0.1085	0.8159	-+
		<i>rs4974601</i>	4	1085409	0.0905	0.0959	0.3453	-+
HS_CNT	Combined	<i>rs12645644</i>	4	1044158	-0.2141	0.0662	0.001233	--+-
		<i>rs4974601</i>	4	1085409	-0.1083	0.0581	0.0622	--+-
NHS_CNT	Male	<i>rs12645644</i>	4	1044158	-0.0208	0.0951	0.827	+-
		<i>rs4974601</i>	4	1085409	-0.0046	0.0827	0.9555	-+
NHS_CNT	female	<i>rs12645644</i>	4	1044158	-0.0244	0.1206	0.8395	+-
		<i>rs4974601</i>	4	1085409	0.0341	0.1081	0.7524	++
NHS_CNT	Combined	<i>rs12645644</i>	4	1044158	-0.0222	0.0747	0.7665	++--
		<i>rs4974601</i>	4	1085409	0.0097	0.0657	0.8829	+--+
MOTIF	Male	<i>rs12645644</i>	4	1044158	0.0078	0.0061	0.2055	++
		<i>rs4974601</i>	4	1085409	0.0051	0.0053	0.3428	-+
MOTIF	female	<i>rs12645644</i>	4	1044158	0.0070	0.0051	0.1667	++
		<i>rs4974601</i>	4	1085409	0.0077	0.0045	0.09058	-+
MOTIF	Combined	<i>rs12645644</i>	4	1044158	0.0073	0.0039	0.0611	++++
		<i>rs4974601</i>	4	1085409	0.0066	0.0035	0.05689	--++



This figure displays 1000kb regions around *RNF212* gene. In FHS data set *RNF212* gene is well covered. The SNPs are color-coded according to correlation (HapMap Phase II CEU) with the most significant SNP to non significant SNP, from red to dark blue presented in rectangular box in upper right corner. Known genes, with their exon, introns and orientation notes are plotted below the SNPs. HapMap recombination rates has been shown with a blue line behind the SNPs. SNP coverage in FHS data sets and Illumina one million chip is noted by tick marks above the plot.

**Figure 2.5:** *RNF212* (male) in FHS data set

We selected the top 11 SNPs from the top hit list of the fixed effect meta-analysis of GDCS and AGRE for each phenotype and made Locus Zoom plots in the FHS female data set totaling around 150 locus zoom plots. Among the top 11 SNPs in ARC phenotype, only two SNPs are genotyped in the FHS data set. Two of the SNPs are not genotyped in FHS, but there are SNPs (2) with strong (.8-1) LD and 4 with medium to high LD that are genotyped in FHS data set. But none of them showed association with significance level less than  $10^{-2}$  order. Figure 2.5 shows the locus-zoom plot of the *RNF212* gene for ARC male only analysis in FHS data set.



This figure displays the 1000kb area around *PRDM9* gene. In FHS data set there is no SNP genotyped on *PRDM9* gene.

**Figure 2.6:** *PRDM9* (gender-pooled) in FHS data set

In our HS\_PCT, HS\_CNT and NHS\_CNT phenotype, *PRDM9* gene is in the center of our interest. But in FHS data set, there is no SNP genotyped on *PRDM9* and showed in

Figure 2.6 (B). Similarly, in male only and gender-pooled analysis, there is no evidence of low p-values for our associated SNPs. Similarly; we also looked at other phenotypes. Except few of the SNPs, most of them did not show any evidence of association.

## 2.5 DISCUSSION

GDCS and AGRE meta-analysis results show that average recombination rate (ARC) has the strongest association with the *RNF212* gene. In separate male and female analyses, *RNF212* is the most significantly associated gene with the phenotype ARC among males (P.value = 7.56E-07). In females it has less effect on the phenotype and in female the effect is in opposite direction but not significant (P.value = 0.1616). Among the males of the FHS data set, *RNF212* is associated with ARC with p-value on the e(-5)th order. In depth analysis of the gene *RNF212* across phenotypes shows that along with strongest association with male average recombination count, *RNF212* is also associated with male HS\_CNT phenotype (P.value = 2.044e-05) and the direction of the effect sizes has the same direction irrespective of sex for HS\_CNT and ARC. For example SNP *rs12645644* has negative effect on ARC in male only analysis and positive effect on female only analysis and negative effect on combined analysis. Similarly in HS\_CNT analysis, this SNP has negative effect on male only analysis and positive effect on female only analysis and negative effect on combined analysis. *RNF212* might not only control the total recombination, it also regulate the recombination count in the hotspot areas in the same fashion. Further study may confirm the broader role of *RNF212*.

In summary, our finding from the three different phenotypes and from male and female separate and combined analysis about *PRDM9* gene suggesting that the females might have stronger overall regulation than the males. That is, in males who have the "high" version of *PRDM9*, both hotspot and overall combination increase. But in females, non-hotspot goes down to compensate (at least more than in males).

In male and female combined meta-analysis of GDCS and AGRE, the *PRDM9* gene showed up in top ten hit list with all three of the phenotypes. Association is strongest with phenotype HS\_PCT (1.20E-13) followed by HS\_CNT (3.80E-08) and then NHS\_CNT (1.16E-07) for male and female combined analysis. In male and female separate analysis, the association is stronger in females compared to males in HS\_PCT and NHS\_CNT. The direction of the effect is opposite between these two phenotypes in female analysis, indicating that the effect of *PRDM9* is broader than the previously suggested in regulating usage of hotspots.

Irrespective of gender, one of the SNPs *rs1874165* in *PRDM9* significantly decreases the number of recombinations in the hotspot areas and increases the number in the non-hotspot areas, which implies that this SNP has a significant effect on even distribution of recombination events in the whole genome (balancing out or a force to use both hotspot and non-hotspot areas).

In male only analysis, same SNP *rs1874165* in *PRDM9* has the effect size -0.55 with HS\_CNT and 0.23 with NHS\_CNT. Where as in females the effect sizes are -0.48 and .77 respectively indicating that there is more regulatory force on females to use non-hotspot areas.

We also identified new suggestive associations for different phenotypes, and two of the regions reached the genome-wide significance and both the regions are associated with phenotype NHS\_CNT. Male and female combined analysis for NHS\_CNT phenotype SNP *rs12186491* (chr 5) in *SPINK6* gene has p.value 6.36E-08 and for male only analysis SNP *rs10937651* (chr 4, bp 5596713, left genes *STK32B*, *C4orf6* and right gene *EVC2* and *EVC*) has p.value 5.16E-08. Qualitative gene based replication with the FHS data set replicated few of our top hits which is also previously reported in the literature. But the lack of coverage of the *PRDM9* region on the Affymetrix 5.0 chip was a serious limitation.

## 2.6 SUPPORTING INFORMATION

Additional tables, figures and discussions are presented in appendix A.

## 2.7 REFERENCES

- [1] J. Baudat F Fau Buard et al. “PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice”. In: 1095-9203 (Electronic) (2009).
- [2] T. Chiang, R. M. Schultz, and M. A. Lampson. “Meiotic origins of maternal age-related aneuploidy”. In: *Biology of reproduction* 86.1 (2012), pp. 1–7.
- [3] T. Hassold et al. “Cytogenetic and molecular studies of trisomy 13”. In: *Journal of medical genetics* 24.12 (1987), pp. 725–32.
- [4] T. J. Hassold et al. “Molecular studies of non-disjunction in trisomy 16”. In: *Journal of medical genetics* 28.3 (1991), pp. 159–62.
- [5] R. H. Martin and A. W. Rademaker. “The effect of age on the frequency of sperm chromosomal abnormalities in normal men”. In: *American journal of human genetics* 41.3 (1987), pp. 484–92.
- [6] K. M. May et al. “The parental origin of the extra X chromosome in 47,XXX females”. In: *American journal of human genetics* 46.4 (1990), pp. 754–61.
- [7] N. Takaesu et al. “Nondisjunction of chromosome 21”. In: *American journal of medical genetics. Supplement* 7 (1990), pp. 175–81.
- [8] A. Fledel-Alon et al. “Variation in human recombination rates and its genetic determinants”. In: *PLoS ONE* 6.6 (2011), e20321.
- [9] R. Chowdhury et al. “Genetic analysis of variation in human meiotic recombination”. In: *PLoS genetics* 5.9 (2009), e1000648.
- [10] G. Coop et al. “High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans”. In: *Science* 319.5868 (2008), pp. 1395–8.
- [11] A. Kong et al. “Sequence variants in the RNF212 gene associate with genome-wide recombination rate”. In: *Science* 319.5868 (2008), pp. 1398–401.
- [12] K. W. Broman et al. “Comprehensive human genetic maps: individual and sex-specific variation in recombination”. In: *American journal of human genetics* 63.3 (1998), pp. 861–9.

- [13] V. G. Cheung et al. “Polymorphic variation in human meiotic recombination”. In: *American journal of human genetics* 80.3 (2007), pp. 526–30.
- [14] V. G. Cheung, S. L. Sherman, and E. Feingold. “Genetics. Genetic control of hotspots”. In: *Science* 327.5967 (2010), pp. 791–2.
- [15] K. Kimura et al. “Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes”. In: *Genome research* 16.1 (2006), pp. 55–65.
- [16] L. Kauppi, A. J. Jeffreys, and S. Keeney. “Where the crossovers are: recombination distributions in mammals”. In: *Nature reviews. Genetics* 5.6 (2004), pp. 413–24.
- [17] M. J. Neale. “PRDM9 points the zinc finger at meiotic recombination hotspots”. In: *Genome biology* 11.2 (2010), p. 104.
- [18] Carmen Sandovici I Fau Sapienza and C. Sapienza. “PRDM9 sticks its zinc fingers into recombination hotspots and between species. LID - 37 [pii]”. In: 1757-594X (Electronic) (2010).
- [19] I. L. Berg et al. “PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans”. In: *Nature genetics* 42.10 (2010), pp. 859–63.
- [20] I. L. Berg et al. “Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.30 (2011), pp. 12378–83.
- [21] A. G. Hinch et al. “The landscape of recombination in African Americans”. In: *Nature* 476.7359 (2011), pp. 170–5.
- [22] A. Kong et al. “Fine-scale recombination rate differences between sexes, populations and individuals”. In: *Nature* 467.7319 (2010), pp. 1099–103.
- [23] L. Segurel, E. M. Leffler, and M. Przeworski. “The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans”. In: *PLoS biology* 9.12 (2011), e1001211.

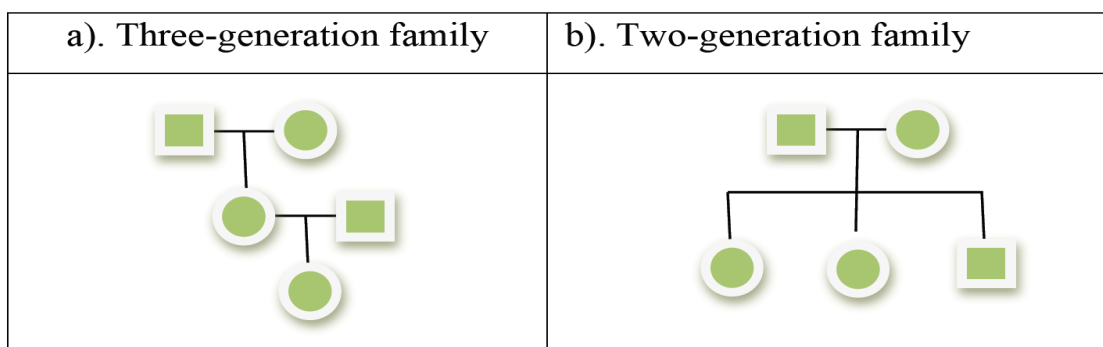
- [24] S. Sarbajna et al. “A major recombination hotspot in the XqYq pseudoautosomal region gives new insight into processing of human gene conversion events”. In: *Human molecular genetics* (2012).
- [25] S. Myers et al. “Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination”. In: *Science* 327.5967 (2010), pp. 876–9.
- [26] J. R. Shaffer et al. “Genome-wide association scan for childhood caries implicates novel genes”. In: *Journal of dental research* 90.12 (2011), pp. 1457–62.
- [27] L.A. Weiss et al. “Association between microdeletion and microduplication at 16p11.2 and autism.” In: *N Engl J Med* 358.7 (2008), pp. 737–9.
- [28] T. R. Dawber, G. F. Meadors, and Jr. Moore F. E. “Epidemiological approaches to heart disease: the Framingham Study”. In: *American journal of public health and the nation’s health* 41.3 (1951), pp. 279–81.
- [29] C. J. Willer, Y. Li, and G. R. Abecasis. “METAL: fast and efficient meta-analysis of genomewide association scans”. In: *Bioinformatics* 26.17 (2010), pp. 2190–1.



### 3.0 SCORING RECOMBINATION IN COMPLEX PEDIGREE STRUCTURES INCLUDING HALF-SIBLINGS

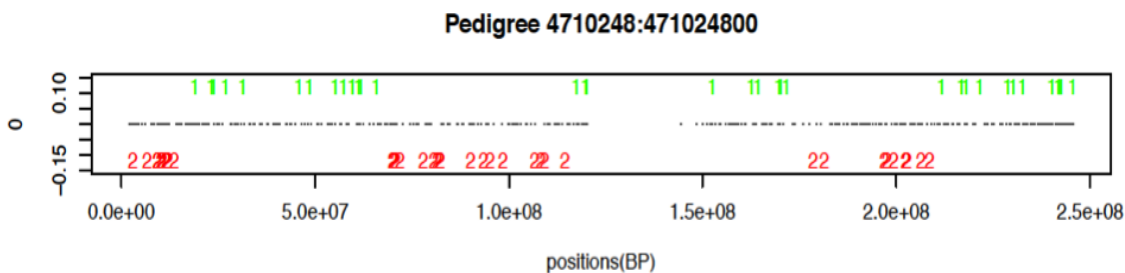
#### 3.1 INTRODUCTION

Recombination scoring methods for different family types are different. The CRIMAP [1] software scores recombination for three-generation families using likelihood methods and microsatellite markers. Coop et. al. [2] and Chowdhury et. al. [3] score recombination for two-generation nuclear families with two or more children using non-parametric SNP-streak methods with dense markers. These two pedigree structures are presented in Figure 3.1. The main objective of project 2 is to bring a broader range of pedigree types under consideration so that we can increase the sample size for recombination GWAS.



**Figure 3.1:** (A) Pedigree structure of three-generation family; (B) Pedigree structure of two-generation nuclear family

In any family type, recombination scoring starts by evaluating all SNPs, one at a time, to determine informativeness. An informative marker is defined as a marker that provides information about the grandparental source of the allele. Informativeness varies by the pedigree structure and the people genotyped in it. As we scan through the genome, depending on the pedigree structure and number of people genotyped, we find the grandparental source of the allele for a single SNP at a time. As scanning progresses and if a switch in the sources of allele (say from grandmother to grandfather) occurs, then we define this event as a recombination event. For example, Figure 3.2 is a plot of grandchild’s chromosome 1, where all 1s and 2s are presenting the source of allele from grandparent 1 and grandparent 2 respectively and all dots (middle row of plot) represent the uninformative SNPs. From this plot we can infer that there were five recombinations on this chromosome in this individual.



**Figure 3.2:** Recombination plot of chromosome 1 (Mukhopadhyay N.)

### 3.2 THREE-GENERATION FAMILIES WITH VARYING NUMBER OF MISSING GENOTYPES

The three-generation family structure shown in Figure 3.1 can be used to score recombination phenotypes in the mother as long as we have genotypes for at least one grandparent and the grandchild. We developed recombination-scoring methods for all possible informative versions of this family structure. One example is shown in Figure 3.3 where both grandparents and the grandchild genotyped, but not the parents. The other cases are shown in appendix B.

GF	GM	Mother (Possible genotype)	GC	Allele source
AA AB BB	AA	AA	AA AB	? ?
	AB	AA or AB	AA AB BB	? ? 0
	BB	AB	AA AB BB	1 ? 0
AB	AA	AA or AB	AA AB BB	? ? 1
	AB	AA or AB or BB	AA AB BB	? ? ?
	BB	AB or BB	AA AB BB	1 ? ?
BB	AA	AB	AA AB BB	0 ? 1
	AB	AB or BB	AA AB BB	0 ? ?
	BB	BB	AB BB	? ?

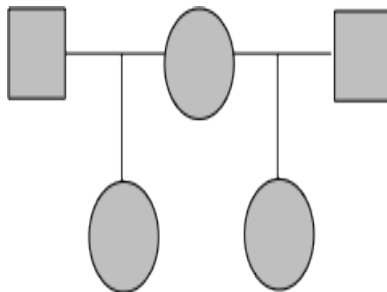
GF: Grandfather, GM: Grandmother, GC: Grandchild

**Figure 3.3:** Three people Genotyped (grandfather, grandmother and grandchild)

Figure 3.3 shows all possible combinations of genotypes for the three genotyped individuals in this pedigree and the grandparental allele scoring for each. For example, if grandfather is homozygous AA and grandmother is heterozygous AB, and if the grandchild is BB, then we can say that the source of the child's maternal B allele is the grandmother. By contrast, if the grandmother and grandfather are both AA, then no grandchild genotype will be informative. The last column of Figure 3.3 gives the information about the source of the grandparental alleles. If the allele is from grandfather, then it is denoted by 1 and if the allele is from grandmother, then it is denoted by 0 and if the source of the allele is unknown (the marker is uninformative), then it is denoted by a question mark. A switch from 0 to 1 or 1 to 0, as we move along the chromosome, is scored as a recombination event.

### 3.3 RECOMBINATION SCORING IN TWO-GENERATION FAMILIES WITH HALF-SIBLINGS

Currently, two-generation families with two or more full siblings are used to score recombination. We develop method to score recombination considering half-siblings in two generation families. A simple pedigree structure of a two-generation family with half-siblings is presented in Figure 3.4.

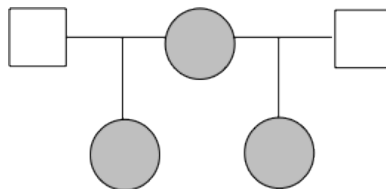


**Figure 3.4:** Pedigree structure of two-generation family with half-siblings

We worked out the scoring of grandparental alleles on all possible versions of this pedigree as listed below:

- (a) Mother, two fathers, and a pair of half-siblings are genotyped
- (b) Mother, one of the fathers, and pair of half-siblings are genotyped
- (c) Mother and pair of half-siblings are genotyped
- (d) One or two fathers and a pair of half-siblings are genotyped. (Note that, in these cases we can score recombination for the mother, but we cannot run GWAS for the mother, since her genotype is missing).
- (e) One or two people genotyped pedigrees are uninformative.

For the sake of simplicity, we show the scoring for the example of a two-generation family with three people genotyped; mother and two half-siblings in Table 3.1. The pedigree structure of this example is shown in Figure 3.5. Unlike the three-generation pedigree, in this case, each SNP does not give us information about which grandparental allele each child has. Rather, the information is whether the two children share the same grandparental alleles or have different grandparental alleles. As we move along the chromosome, a switch from same to different or vice versa indicates a recombination. (Note that, we do not know in which child the recombination occurred, but we do not need that information since it is the mother's phenotype we are trying to calculate. This is similar to the methods used in two-generation full-sibling families.)



**Figure 3.5:** Pedigree structure of two-generation family with half-siblings and missing genotypes of fathers.

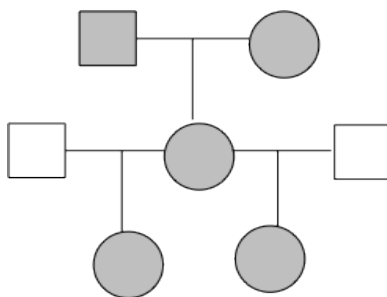
**Table 3.1:** Three people Genotyped (mother and half-siblings pair)

Mom	Half-sib1	Half-sib2	Grandparental alleles
AB	AA	AA	same
		AB	?
		BB	different
	AB	AA	?
		AB	?
		BB	?
	BB	AA	different
		AB	?
		BB	same
AA or BB			?

### 3.4 RECOMBINATION SCORING IN THREE-GENERATION FAMILIES WITH HALF-SIBLINGS

We also worked out the scoring of recombinations in three generation families with half-siblings and the pedigree structure is presented in Figure 3.6. All possible combinations allowing missing genotype in three-generation family with half-siblings are listed below:

- (a) Two grandparents, mother, and a pair of half-siblings are genotyped
- (b) One grandparent, mother, and a pair of half-siblings are genotyped.



**Figure 3.6:** Pedigree structure of three-generation family with half-siblings.

It may be possible to score recombination more efficiently by considering two half-siblings together in a three-generation family though it does not add further information about the allele switch compare to scoring two separate three-generation families with one child (grandparents, mother, and grandchild).

### 3.5 REFERENCES

- [1] E. S. Lander and P. Green. “Construction of multilocus genetic linkage maps in humans”. In: *Proceedings of the National Academy of Sciences of the United States of America* 84.8 (1987), pp. 2363–7.
- [2] G. Coop et al. “High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans”. In: *Science* 319.5868 (2008), pp. 1395–8.
- [3] R. Chowdhury et al. “Genetic analysis of variation in human meiotic recombination”. In: *PLoS genetics* 5.9 (2009), e1000648.

## 4.0 COMPREHENSIVE LITERATURE REVIEW AND STATISTICAL CONSIDERATIONS FOR GWAS META-ANALYSIS

Ferdouse Begum<sup>1</sup>, Debashis Ghosh<sup>2</sup>, George C. Tseng<sup>\*1,3</sup>, Eleanor Feingold<sup>1,3</sup>

<sup>1</sup>Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA

<sup>2</sup>Department of Statistics, Pennsylvania State University, University Park, PA, USA

<sup>3</sup>Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA

This chapter has been published in Nucleic Acids Research (NAR). This chapter is included in this dissertation with permission of Oxford University Press with the following citation:

Begum, F., Ghosh, D., Tseng, G.C. and Feingold, E. (2012) Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic acids research, 40(9): 3777-3784.



## 4.1 ABSTRACT

Over the last decade, genome-wide association studies (GWAS) have become the standard tool for gene discovery in human disease research. While debate continues about how to get the most out of these studies and on occasion about how much value these studies really provide, it is clear that many of the strongest results have come from large-scale mega-consortia and/or meta-analyses that combine data from up to dozens of studies and tens of thousands of subjects. While such analyses are becoming more and more common, statistical methods have lagged somewhat behind. There are good meta-analysis methods available, but even when they are carefully and optimally applied there remain some unresolved statistical issues. This paper systematically reviews the GWAS meta-analysis literature, highlighting methodology and software options and reviewing methods that have been used in real studies. We illustrate differences among methods using a case study. We also discuss some of the unresolved issues and potential future directions.

## 4.2 INTRODUCTION

Genome-wide Association Studies (GWAS) test for statistical association between genotype and phenotype on hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) at a time in order to find genes that contribute to human diseases or non-disease traits. Early in the GWAS era, costs were high and sample sizes were small, but with technological advances prices have come down significantly and typical sample sizes are now in the thousands. Even with those large sample sizes, discoveries have been modest for many or most phenotypes studied because typical effect sizes are quite small, and many results do not appear to replicate in subsequent studies. As a result, most GWAS publications now involve multiple datasets in order to both reduce false positives and increase statistical power to find true positives. Often these multiple datasets are analyzed individually, or some of them are only used for in-silico replication (i.e. only top markers from one dataset

are examined in the remaining datasets). There is growing recognition, however, that the most statistically robust and efficient analysis is a full-genome meta-analysis combining all studies and using all data at every marker. Meta-analysis provides optimum power to find effects that are homogeneous across cohorts, and at the same time can shed light on between-study heterogeneity [1, 2, 3, 4, 5]. Going even further, many investigators are now forming mega-consortia of a dozen or more studies for increased statistical power. Meta-analysis thus has become a routine part of GWAS, and yet there remain unresolved issues about the most powerful and robust ways to use it. This paper attempts to provide a comprehensive review of GWAS meta-analysis methods, practices, and problems, with the goal of helping both applied and methodological researchers take the necessary next steps forward. In the next section we provide an overview of GWAS meta-analysis methods, and in Section 3 we review databases and software. Section 4 summarizes the methods used in the literature, and Section 5 presents our case study. Finally, in Section 6 we discuss important open questions.

### 4.3 GWAS META-ANALYSIS DATA AND METHODS

It is fairly common for an individual investigator to perform GWAS on several different study populations and combine the results into a single report. If the genotyping is done for all studies together, data from the different populations can be directly combined (termed mega-analysis), and meta-analysis is not necessary. GWAS investigators generally turn to meta-analysis when scans are performed on different chips and/or when results from different investigators need to be combined and raw data cannot be exchanged for reasons of either confidentiality or proprietorship.

There has historically been some concern about the appropriateness of mega-analysis and even meta-analysis given the high level of heterogeneity among GWAS of the same trait. Sources of heterogeneity between studies can include different trait measurements and study designs, different ethnic groups, different environmental exposures, different genotyping chips, etc. For example, if two study populations have significantly different environmen-

tal backgrounds (say different diets in an obesity study), different genes may be relevant to the trait in the two populations (i.e. there may be gene x environment interaction). Another important source of heterogeneity is differing linkage disequilibrium patterns in different ethnic groups, so that even if the same variant is causal in both groups, the SNPs that are associated (in linkage disequilibrium) with it may differ from group to group. Recently, Lin et al. allayed some of these concerns. They showed both theoretically and by simulation that meta-analysis and mega-analysis have essentially equal statistical efficiency, and also that the efficiency of both approaches is fairly robust to between-study heterogeneity [6]. Heterogeneity remains a concern, however, and we will discuss it further throughout the paper (e.g. in the random effects model, case study and open questions).

Most GWAS meta-analysis uses relatively straightforward methods. P-values can be combined either with or without weights, or effect sizes can be combined in either fixed or random effects models. (See the companion paper on microarray meta-analysis for a more detailed exposition of the differences among these methods). Any of those methods can be applied either across all studies at once, or cumulatively as each study is added. Most GWAS meta-analysis takes a frequentist approach, but Bayesian hierarchical models can also be used, and are very well-suited to a cumulative approach [7]. Table 4.1 lists the commonly-used GWAS meta-analysis methods and the source information that is required for each. The methods are described in a bit more detail below.

The simplest GWAS meta-analysis approach is to combine p-values using Fisher’s method. The formula for the statistic is

$$X^2 = -2 \sum_{i=1}^k \log(p_i)$$

where  $p_i$  is the p-value for the  $i$ th study. Under the null hypothesis,  $X^2$  follows a chi-squared distribution with  $2k$  degrees of freedom, where  $k$  is the number of studies. A major limitation of this method is that all studies are weighted equally, which is likely to be highly suboptimal when combining GWAS studies with different sample sizes. An additional problem is that the direction of effect of each SNP is not considered, so that studies with associations in opposite directions appear to strengthen each other rather than contradicting each other.

A major improvement over Fishers method is a weighted Z-score method, in which p-values are transformed to Z-scores in a one-to-one transformation. The weighted Z-score method is more powerful and efficient than Fishers method, and allows different weights for different studies [8]. It also takes into account the direction of the effect at each SNP. The software METAL [9] implements the weighted Z-score method using the following formula:

$$Z = (\sum_i Z_i w_i) / \sqrt{(\sum_i w_i^2)},$$

where the weight  $w_i$  = square root of sample size of the  $i$ th study,  $Z_i = \phi^{-1}(1 - p_i/2) * (\text{effect direction for study } i)$ , and  $p_i$  is the p-value for the  $i$ th study. Note that the METAL paper has a typo in this formula, but we have confirmed by testing the software that the formula shown above is in fact correctly implemented in the software.

The major alternative to combining p-values and/or Z-scores is to combine effect sizes (estimates). This can be done with either a fixed effects or a random effects model. Combining effect sizes is statistically more powerful than combining Z-scores, but it requires that the trait be measured on exactly the same scale in each study, with the same units, same transformations, etc. This may be achievable in a meta-analysis of a trait with highly standardized measurements, but there are many traits for which it is unlikely to be possible, for example alcohol or tobacco use. The difference between the fixed effects and random effects models is that fixed effects meta-analysis assumes that the genetic effects are the same across the different studies. Fixed effects models provide narrower confidence intervals and significantly lower p-values for the variants than random effects models [1, 10, 2, 11, 12, 13, 14]. Both fixed effects and random effects models are briefly discussed below; details can be found in Nakaoka et. al (2009)[15].

For the fixed effects model, inverse-variance weighting is widely used, although other methods are also available. The weighted average of the effect sizes can be calculated as  $\hat{\theta}_F = (\sum_i w_i \hat{\theta}_i) / (\sum_i w_i)$  and the variance of the weighted average of the effect size is  $var(\hat{\theta}_F) = 1 / (\sum_i w_i)$ , where  $\hat{\theta}_i$  is the logarithm of the  $i$ th case-control study effect,  $w_i$  is the reciprocal of the estimated variance of the effect size for the  $i$ th case-control study.

The random effects model assumes that the mean effect (of each SNP) in each study is

**Table 4.1:** Sources of information for different methods of meta-analysis

	Fishers p	Weighted Z	Fixed effect	Random effect
p-value	X	X		
Effect size			X	X
Direction of the effect size		X		
Sample size		X		
Heterogeneity estimate				X
SE of effect size			X	X

different, with those means usually assumed to be chosen from a Gaussian distribution. The variance of that Gaussian distribution, and thus the amount of between-study heterogeneity, is estimated by the model. Thus the random effects model not only does not assume homogeneity of effect but is able to give an estimate of the degree of heterogeneity. The weight of each study incorporates the between-study variance of heterogeneity, which is expressed as  $\tau^2$ , where

$$\tau^2 = (Q - (k - 1)) / (\sum_i w_i - \sum_i w_i^2 / \sum_i w_i).$$

The weight for the random effects model is calculated as  $w_i^R = 1/(1/w_i + \hat{\tau}^2)$  and  $Q = \sum_i w_i (\theta_i - \hat{\theta}_F)^2$ , Cochran's test statistic [16] follows a chi-squared distribution with  $k - 1$  degrees of freedom under the assumption of genetic homogeneity.  $Q$  is most widely used to check the between-study heterogeneity. But  $Q$  is underpowered when the number of studies is small. To overcome this problem, there are some other statistics available, such as  $H$ ,  $R$  and  $I^2$ , defined as  $H = \sqrt{Q/(k - 1)}$ ,  $R = \sqrt{\text{var}(\hat{\theta}_R)/\text{var}(\hat{\theta}_F)}$  and  $I^2 = 100 * (Q - (k - 1))/Q$ , where  $\hat{\theta}_R$  is the genetic effect under the random effects model.  $H$ ,  $R$  and  $I^2$  have some desirable characteristics such as being scale and size invariant [10, 15]. These statistics are calculated separately for each SNP, which leads to the interesting and unsolved question of whether or how one should make a genome-wide determination of heterogeneity.

In addition to these basic methods, almost any meta-analysis method in the statistical literature can be applied to GWAS, and some of the software packages discussed below do so.

## 4.4 DATABASES AND SOFTWARE

Most GWAS meta-analyses are assembled from consortia of investigators working on similar traits, but public databases are also used. The most important GWAS database is the NIH Database of Genotype and Phenotype (dbGaP), which is the repository for both raw data and results from most NIH-funded GWAS. There are also a number of databases that contain selected results from GWAS studies, some of which are suitable for inclusion in meta-analyses of targeted regions. GWAS Central is one of the oldest such databases, which started in 1998 under a different name. On 4/27/11, it contained 708 studies. The Human Genome Epidemiology Network (HuGE Net) (<http://www.hugenet.ca>) also has a GWAS integrator webpage and contains a list of publications, hits, variants, disease and trait information etc. Like HuGE Net, The National Human Genome Research Institute (NHGRI) (<https://www.genome.gov/>) maintains a catalog of published GWAS studies [17]. Other available databases include the HKSC database with both bone mineral density (BMD) and fracture data [18] and the Millennium Genome Project (MGP) (<https://gemdbj.nibio.go.jp/dgdb/>), which has a repository of Japanese SNP(JSNP) data [19].

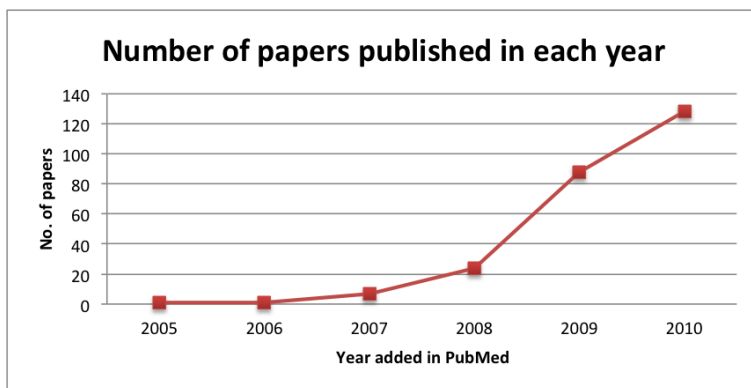
The statistical methods used for GWAS meta-analysis are very straightforward, and it is not difficult to implement them, but there are several software packages available that can make this easier and that integrate useful bioinformatics or visualization functions. The most widely used software is METAL (<http://genome.sph.umich.edu/wiki/METAL/Program>) [9]. METAL implements two strategies, a weighted Z-score method based on sample size, p-value and direction of effect in each study, and an effect-size based method weighted by the study-specific standard error. The other most commonly used package is MetABEL, which is a component of the GenABEL suite in R. MetABEL implements a fixed effects model like METAL, and results can be shown with a visualization tool. A number of other packages are also in use, including META (<http://www.stats.ox.ac.uk/~jsliu/meta.html>). GWAMA [20] has useful auxiliary features that METAL, MetABEL, and META lack. PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink/metaanal.shtml>) [21] is a free, open-source software for GWAS analysis, which also has some meta-analysis tools to do fixed

effects and random effects meta-analysis. MAGENTA (<http://www.broadinstitute.org/mpg/magenta/>) [22] can be used to test a specific hypothesis or to generate hypotheses, and it provides gene set enrichment analysis p-values and false discovery rate. Comprehensive Meta-analysis (CMA) ([www.Meta-Analysis.com](http://www.Meta-Analysis.com)) Software [23] is a commercial package to do meta-analysis which works in a spreadsheet interface and also provides forest plots, which are useful for visualizing between-study heterogeneity (see case study). Review Manager (RevMan) (<http://ims.cochrane.org/revman/about-revman-5>) [24] is another package that does meta-analysis and provides results in tabular format and graphically. It also provides different kinds of reviews including intervention reviews, diagnostic test accuracy reviews, methodology reviews and overviews of the reviews. There are several STATA modules to perform meta-analysis, such as METAN [25], HETEROGI [25] and more specifically METAGEN [26] (<http://bioinformatics.biol.uoa.gr/~pbagos/metagen/>) for genetic association studies. In R, a few other available packages for meta-analysis are Metafor (<http://www.metafor-project.org/>) [27], rmeta, and CATMAP. The Metafor package has different functions to calculate fixed, random and mixed effects along with moderator and meta-regression analysis and provides different kinds of graphical displays of results and data. Synthesis-view (<https://chgr.mc.vanderbilt.edu/synthesisview>) [28] is a visualization tool which can integrate multiple pieces of information across studies, such as p-values, effect sizes, allele frequencies etc. IGG3 [29] can integrate raw GWAS data from multiple chips and provide the input files for different imputation software, which can be used in meta-analysis later. Magi and Morris (2010) made a nice comparison of different features among a number of meta-analysis software packages [20].

One issue that is unique to GWAS meta-analysis is that SNPs may not be coded the same way in different datasets – the so-called strand issue. Opposite coding of SNPs in different studies can cause what should be similar effects to look precisely opposite. This often occurs for only a small subset of SNPs (those with minor allele frequencies near 50%) and so can be very difficult to detect. Most of the meta-analysis software packages discussed above have varying bioinformatics features to resolve this problem, including METAL, MetABEL, META, and GWAMA [20].

## 4.5 LITERATURE REVIEW

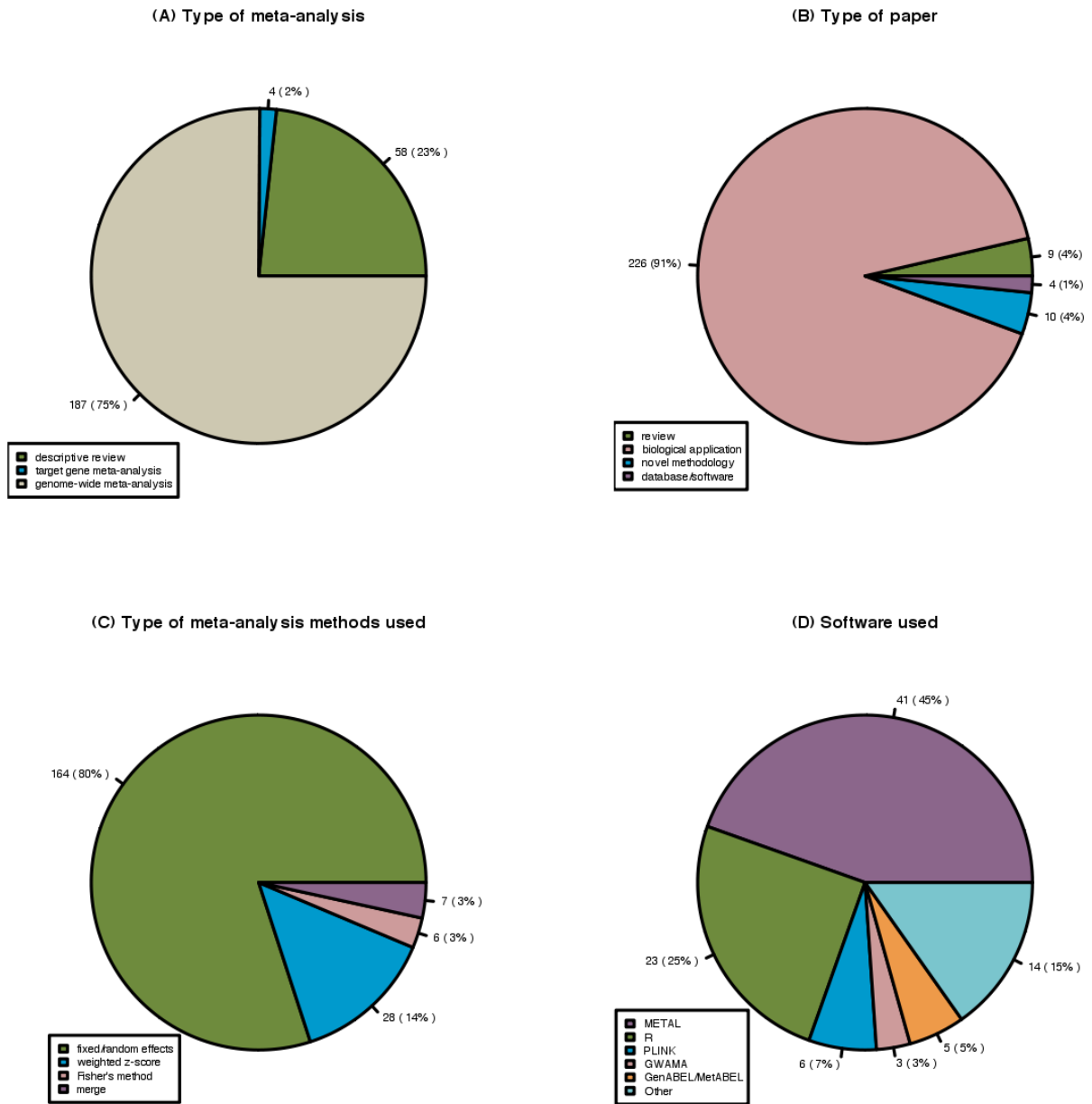
This review started with a search of GWAS meta-analysis using PubMed on 12/29/2010, which yielded 299 papers. After removing duplicates and irrelevant papers there were 249 GWAS meta-analysis papers (see complete searching and paper collection criteria in the companion paper). Figure 4.1 summarizes the number of papers by year of publication, illustrating the exponential increase between 2005 and 2010. Figure 2 summarizes the contents of the papers. One hundred and eighty-seven papers (75%) are full GWAS meta-analyses, while 58 papers (23%) are replication analyses on targeted loci (Figure 2A). Figure 2B shows that the majority of reports are biological applications (226 papers; 91%) while 10 papers (4%) are for novel methodology, 4 papers (1%) are databases and software, and 9 papers (4%) are review papers.



**Figure 4.1:** Number of GWAS studies by year of publication.

Figures 2C and 2D show the methods and software used. One hundred and sixty-four papers (80%) use fixed or random effects models, 28 (14%) combine weighted z-scores from p-values, six (3%) use Fishers method, and seven (3%) use direct data merging. For software packages, METAL (41 papers; 45%) and R packages (23 papers; 25%) are the most popular. Other software choices include PLINK (6 papers; 7%); GWAMA (3 papers; 3%); and GenABEL/MetABEL (5 papers; 5%). Detailed information of the paper list and categorization to generate Figure 2 is available in the online Supplementary Data.





**Figure 4.2:** Summary of GWAS meta-analysis review: (A) type of meta-analysis; (B) type of paper; (C) type of meta-analysis method; (D) software used.

## 4.6 CASE STUDY

In this section, we present a simple case study that demonstrates some of the differences among GWAS meta-analysis methods. Two datasets are included in this meta-analysis, which we label here as dataset 1 and dataset 2. The datasets are from different studies and different populations, but both were genotyped on the Illumina Human610-Quad Beadchip. The phenotype is total meiotic recombination across the genome, which has been of great interest in the genetics literature lately, with many new discoveries especially about the recombination hotspot gene *PRDM9*. Meiotic recombination events for both parents in nuclear families were scored according to Chowdhury et al [30]. The gene *RNF212* is well-known to be associated with recombination [30, 31, 32], so we report results for four SNPs within this gene. Because the reported associations between *RNF212* and recombination differ in males and females, we consider males and females both separately and combined in our case study, which provides an illustration of how the different meta-analysis methods behave in the presence of heterogeneity. All the methods of meta-analysis for our case study were implemented by us in R.

Table 4.2 shows the results of our case study. The first two rows give the single-study p-values for each SNP in the four datasets (dataset 1 male, dataset 1 female, dataset 2 male, dataset 2 female). These are based on standard GWAS methods using linear regression for each SNP under an additive genetic model. No multiple comparisons correction was applied. The notable result is that all p-values are highly significant in the dataset 1 males, but not in either set of females. In the dataset 2 males, two of the SNPs have p-values of .01 and two are on the order of .20. Note that the sample size in dataset 2 is much smaller than in dataset 1, so even if the effects are the same in the two datasets we would expect larger p-values in dataset 2.

When the four meta-analysis methods are used to combine the two male datasets for the first two SNPs, they all perform reasonably well, but there are clear differences. Fishers method has the lowest power (highest p-values), as would be expected because it is using equal weights for these two very different-sized datasets. The highest power is found with

**Table 4.2:** Case study results

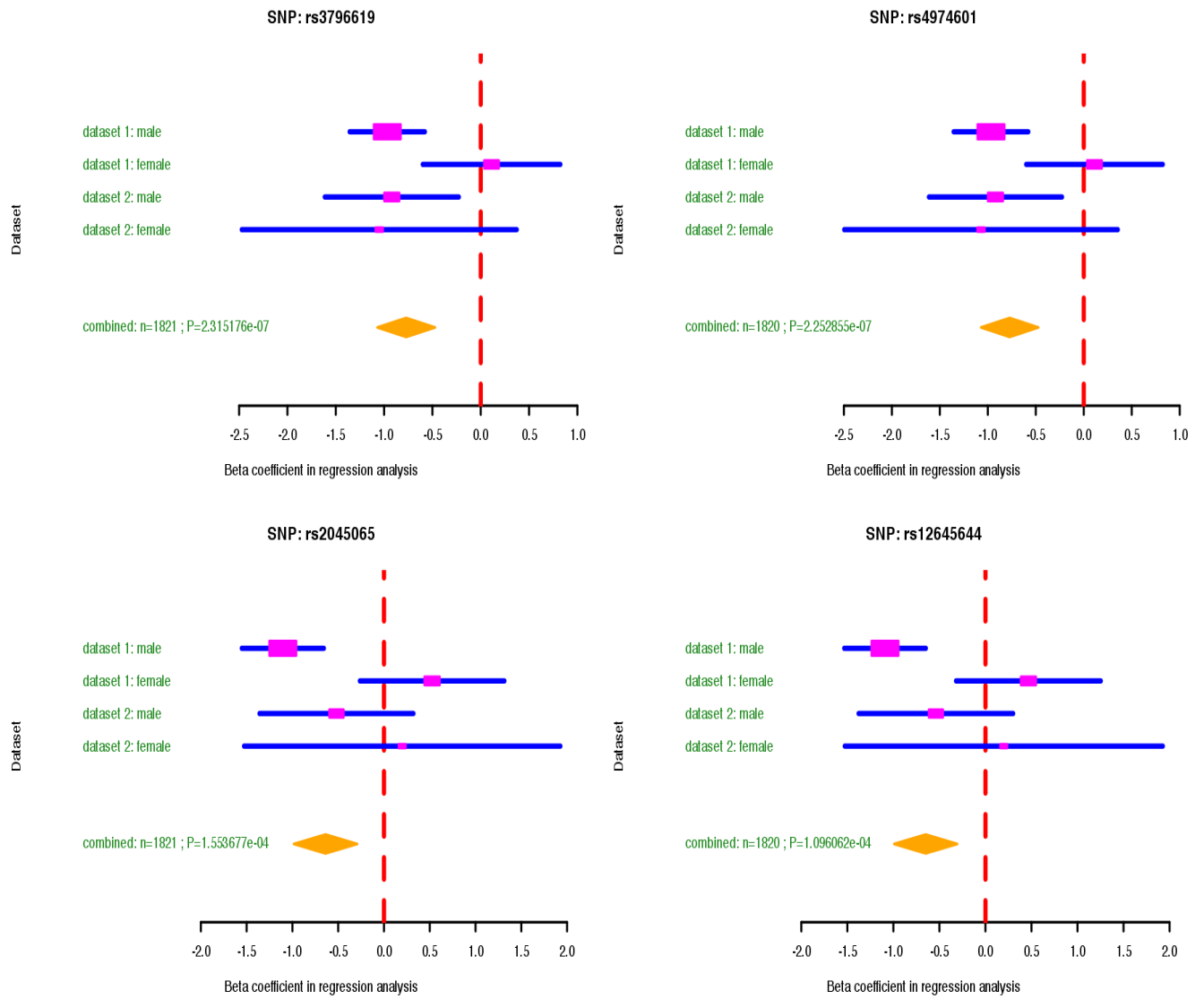
	SNPs in <i>RNF212</i>			
	<i>rs3796619</i>	<i>rs4974601</i>	<i>rs2045065</i>	<i>rs12645644</i>
<b>STUDY ANALYSIS</b>				
<b>Dataset 1, P-value</b>				
Male (n = 736)	1.4E-6	1.4E-6	1.7E-6	1.8E-6
Female (n=736)	0.76	0.76	0.19	0.25
<b>Dataset 2, P-value</b>				
Male (n = 174)	0.01	0.01	0.23	0.21
Female (n=174)	0.15	0.14	0.82	0.82
<b>META-ANALYSIS</b>				
<b>Fishers, P-value</b>				
Male	2.7E-7	2.7E-7	6.2E-6	5.9E-6
Female	0.36	0.35	0.45	0.52
Combined	2.6E-6	2.5E-6	5.7E-5	6.7E-5
<b>Weighted Z, P-value</b>				
Male	2.35E-8	2.35E-8	6.87E-7	6.34E-7
Female	0.36	0.36	0.10	0.13
Combined	1.97E-5	1.91E-5	5.96E-3	4.46E-3
<b>Fixed Effect , P-value</b>				
Male	1.7E-8	1.7E-8	7.0E-7	6.3E-7
Female	0.35	0.35	0.10	0.12
Combined	2.3E-7	2.2E-7	1.6E-4	1.1E-4
<b>Random Effect, P-value</b>				
Male	1.7E-8	1.7E-8	1.7E-1	1.5E-1
Female	0.34	0.34	0.10	0.12
Combined	3.0E-1	3.0E-1	4.5E-1	4.4E-1

both the fixed and random effects models; the similarity of these two methods for these two SNPs indicates that the fixed effects model fits well. For the third and fourth SNPs, the weighted Z-score method and the fixed effects model have better power than Fishers method. The random effects model estimates a very large random component and gives a very high p-value for the SNP. This is probably an artifact caused by fitting a random effects model to just two datasets. Based on the biology, a fixed effects model is likely to be more or less correct for this phenotype, as long as only a single sex is included in the analysis.

In combining the female datasets, all four meta-analysis methods also behave similarly, reflecting the lack of significant association.

When all four datasets (males and females) are combined, we can clearly see the effect of the heterogeneity on the different meta-analysis methods. In general the fixed effects model retains good power to detect association despite our inclusion of some studies (the females) that have little or no effect, while the random effects model completely loses power because it is fitting an incorrect model of a Gaussian random effect. That is, our male and female effects are not the same, but they are not random either - what we actually have is a mixture of two fixed-effects models. We suggest that the typical situation in a GWAS meta-analysis is likely to be similar to this - a mixture of fixed effects rather than a true random effect - and thus that the random effects model may not be the most appropriate way to deal with heterogeneity in GWAS meta-analysis. This proposition clearly deserves further study, however.

One important way to visualize heterogeneity is with a forest plot, which shows the separate estimates and their confidence intervals for each study, and also shows the combination. Figure 3 is a forest plot for all four SNPs and all four populations in the case study; the overall effect shown in the forest plots is from the fixed-effects model. The R package `rmeta` was used to generate the forest plots. These plots make it very easy to visualize some of the important features that the p-values only hint at, such as the fact that the two male populations are in fact quite consistent with each other despite the differing p-values, and the fact that the female effect is actually in the opposite direction (which is consistent with the recombination literature).



**Figure 4.3:** Forest plot of the selected SNPs.

## 4.7 COMPLICATIONS AND OPEN QUESTIONS

GWAS meta-analysis is now widely used and in general has worked well to discover genetic effects that were not uncovered in individual studies. There are, however, some remaining barriers and open methodological issues.

Genotype data cleaning: Prior to meta-analysis, it is clearly important that all datasets undergo thorough standard GWAS data cleaning, such as filtering out bad SNPs and samples using genotype call rates, tests of Hardy-Weinberg equilibrium (HWE) etc [33]. What is not entirely clear is how important it is that the data cleaning steps and standards be the same across datasets. For example, can it cause problems if different genotype call rate cutoffs are used in different datasets? This has not been systematically studied to our knowledge. In genetic association studies for targeted SNPs, there have been three ways to deal with HWE: including all studies irrespective of the HWE tests [34], doing sensitivity analysis to verify differential genetic effects in subgroups [15, 35, 36, 37], and excluding studies with statistically significant deviation from HWE [15, 38]. More recently, most large consortium meta-analyses have attempted to use consistent HWE cutoffs across studies, which is clearly the safest approach.

It is also not clear whether it is necessary or desirable to implement data cleaning steps that compare datasets to each other. The same SNP assay can behave differently on different chips, or even on the same chip in different batches, and thus it is common to scan datasets for SNPs with widely differing allele frequencies and eliminate them before combining. But if the datasets are from different ethnic groups, there will also be SNPs for which there are true differences in allele frequency. It is not clear whether there is a way to distinguish the artifacts from the real differences, and thus it is difficult to recommend an ideal cleaning strategy. Similarly, HWE testing poses issues when datasets are combined (as discussed above), but it is probably clear that HWE tests on combined datasets would be unacceptably conservative. These issues are particularly important in the situation where different studies have different phenotype distributions (or, equivalently, different case:control ratios).

Imputation: When studies are genotyped on different chips, there may be very little overlap in the SNP sets, and thus direct SNP-by-SNP meta-analysis is impossible. For example, the overlap between the Illumina 550K SNP set and the Affymetrix 500K SNP set is only about 100K or 20% of SNPs. The standard solution to this problem is to impute the genotypes of all SNPs in all samples, and a variety of good methods is available for doing so [39]. The problem this creates, which has not been carefully addressed in the literature, is that imputed genotypes have slightly higher error rates and variances than non-imputed genotypes. In general, if imputation is done carefully, the error rates are very low. Error rates can be higher, however, for areas of the genome with sparse SNP coverage or for ethnic groups that are not well represented in the dataset that is used for imputation reference (usually HapMap or 1000 genomes). As with data cleaning above, this issue can be critical if different studies have different phenotype distributions. If two studies have different case:control ratios and one is genotyped and one imputed for a particular SNP, then there is a resulting difference between case and control variances, which can cause false positive results. Conversely, if one chip has very poor coverage of a region, then imputation will create genotypes that actually convey very little information, in which case the meta-analysis can give false negative results because it is averaging in non-informative datasets. Some kind of regionally-smoothed meta-analysis may be the solution to this problem, but such methods have not been developed to our knowledge. In general, it is always advisable to check data quality or replicate results that are based predominantly on imputed data.

Choice of genetic models: In GWAS analysis, the basic association test can be based on an allele frequency comparison or on various statistical contrasts of genotype frequencies, for example an additive model, a dominant model, etc. The same model is used for each SNP, so usually something relatively robust such as the additive model is used [40]. It is most desirable in meta-analysis to use the same model in each study, but in post hoc combinations of analyses that might not always be possible. To our knowledge, no one has studied the effect of such variation in association model on meta-analysis. Clearly it causes some level of effect heterogeneity that would, at least formally, violate a fixed effects model, though it would

not fit a Gaussian random effects model either. Similar issues arise if different covariates or different methods for controlling for population stratification are used in different studies.

Between-study heterogeneity: As discussed above, between-study heterogeneity should probably be considered the norm in GWAS meta-analysis. Such heterogeneity is important to discover and report, since it can lead to important biological insights, for example differences in the genetic control of male and female recombination. The conventional wisdom in the statistical literature is that when heterogeneity is present or even likely, the random effects model is more appropriate than the fixed effects model. We suggest that this might not be the right approach for GWAS, because 1) the number of studies being combined is often not very large (leading to an imprecise heterogeneity estimate) and 2) the form of the heterogeneity typically does not fit a Gaussian random effects model. We do suggest that forest plots are an important heuristic method for discovering and understanding heterogeneity, but we also propose that further work on random or mixed-effects models that are a better fit to GWAS data might improve analyses. For example, in our recombination example we know that males and females are likely to be different, so we could fit a model that explicitly has different fixed male and female effects.

## 4.8 CONCLUSION

As the GWAS literature moves away from artificial replication and toward the more statistically optimal direct combination of all available data in a meta-analysis framework, it will be critical for investigators to understand the best methods for performing that meta-analysis. While good methods are already in use in most studies, there is room for improvement in many of the details discussed above. Many of the potential improvements are ideally addressed by planning studies in a coordinated manner from the beginning, but that is not always feasible. We still need improved methods for post hoc combinations of studies that may have significant heterogeneity in chip, study population, environmental exposures, as-



sociation tests, etc. Looking even further ahead, all of the issues addressed above will need to be re-examined for meta-analyses of SNP data derived from sequencing studies, which will undoubtedly be appearing soon in journals throughout the field.

#### **4.9 SUPPLEMENTARY MATERIAL**

Supplementary Data are available at NAR online.

#### **4.10 ACKNOWLEDGMENTS**

The authors thank Drs Vivian Cheung and Mary Marazita for the use of their data in the case study. They also thank C Song, X Wang and G Liao for collecting and printing papers.

#### **4.11 FUNDING**

National Institutes of Health (NIH) (R01MH077159, RC2HL101715 to G.C.T.); NIH (R01HD38979, R01DE14899 to E.F. and F.B.); NIH (R01GM72007 to D.G.); Huck Institute for Life Sciences (to D.G). Funding for open access charge: University of Pittsburgh.

## 4.12 REFERENCES

- [1] J. P. Higgins et al. “Measuring inconsistency in meta-analyses”. In: *BMJ* 327.7414 (2003), pp. 557–60.
- [2] J. P. Ioannidis. “Non-replication and inconsistency in the genome-wide association setting”. In: *Human heredity* 64.4 (2007), pp. 203–13.
- [3] J. R. Thompson, J. Attia, and C. Minelli. “The meta-analysis of genome-wide association studies”. In: *Briefings in bioinformatics* 12.3 (2011), pp. 259–69.
- [4] S. G. Thompson. “Why sources of heterogeneity in meta-analysis should be investigated”. In: *BMJ* 309.6965 (1994), pp. 1351–5.
- [5] R. Guerra and D. R. Goldstein. *Meta-analysis and Combining Information in Genetics and Genomics*. Mathematical and Computational Biology Series. CRC press, Taylor et al., 2010.
- [6] D. Y. Lin and D. Zeng. “Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data”. In: *Genetic epidemiology* 34.1 (2010), pp. 60–6.
- [7] E. Zeggini and J. P. Ioannidis. “Meta-analysis in genome-wide association studies”. In: *Pharmacogenomics* 10.2 (2009), pp. 191–201.
- [8] M. C. Whitlock. “Combining probability from independent tests: the weighted Z-method is superior to Fishers approach”. In: *J. Evol. Biol* 18.2005 (2005), pp. 1368–1373.
- [9] C. J. Willer, Y. Li, and G. R. Abecasis. “METAL: fast and efficient meta-analysis of genomewide association scans”. In: *Bioinformatics* 26.17 (2010), pp. 2190–1.
- [10] J. P. Higgins and S. G. Thompson. “Quantifying heterogeneity in a meta-analysis”. In: *Statistics in medicine* 21.11 (2002), pp. 1539–58.
- [11] J.P. A. Ioannidis, N. A. Patsopoulos, and E. Evangelou. “Heterogeneity in Meta-analysis of Genome-wide Association Investigations.” In: *PLOS ONE* e841.9 (2007).
- [12] J. Lau, J.P. Ioannidis, and C.H. Schmid. “Quantitive synthesis in systematic reviews.” In: *Ann Intern Med* 126 (1997), pp. 820–826.

- [13] J. Lau, J. P. Ioannidis, and C. H. Schmid. “Summing up evidence: one answer is not always enough”. In: *Lancet* 351.9096 (1998), pp. 123–7.
- [14] A. J. Sutton et al. *Methods for meta-analysis in medical research*. John Wiley and Sons, Chichester, 2000.
- [15] H. Nakaoka and I. Inoue. “Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner’s curse”. In: *Journal of human genetics* 54.11 (2009), pp. 615–23.
- [16] W. G. Cochran. “The combination of estimates from different experiments.” In: *Biometrics* 10 (1954), pp. 101–129.
- [17] S. T. Kim et al. “Prostate cancer risk-associated variants reported from genome-wide association studies: meta-analysis and their contribution to genetic Variation”. In: *The Prostate* 70.16 (2010), pp. 1729–38.
- [18] A. W. Kung et al. “Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies”. In: *American journal of human genetics* 86.2 (2010), pp. 229–39.
- [19] M. Hirakawa et al. “JSNP: a database of common gene variations in the Japanese population”. In: *Nucleic Acids Res* 30.1 (2002), pp. 158–162.
- [20] R. Magi and A. P. Morris. “GWAMA: software for genome-wide association meta-analysis”. In: *BMC bioinformatics* 11 (2010), p. 288.
- [21] S. Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. In: *American journal of human genetics* 81.3 (2007), pp. 559–75.
- [22] A. V. Segre et al. “Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits”. In: *PLoS genetics* 6.8 (2010).
- [23] H. Q. Qu et al. “In silico replication of the genome-wide association results of the Type 1 Diabetes Genetics Consortium”. In: *Human molecular genetics* 19.12 (2010), pp. 2534–8.
- [24] “Review Manager (RevMan) [Computer program]. Version 5.1. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2011.” In: ()

- [25] N. A. Patsopoulos and J. P. Ioannidis. “Susceptibility variants for rheumatoid arthritis in the TRAF1-C5 and 6q23 loci: a meta-analysis”. In: *Annals of the rheumatic diseases* 69.3 (2010), pp. 561–6.
- [26] Wu Y-w et al. “Analysis of Lingo1 variant in sporadic and familial essential tremor among Asians”. In: *Acta Neurologica Scandinavica* (2010).
- [27] Wolfgang Viechtbauer. “Conducting meta-analyses in R with the metafor package.” In: *Journal of Statistical Software* 36.3 (2010), pp. 1–48.
- [28] S. A. Pendergrass et al. “Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis”. In: *BioData mining* 3 (2010), p. 10.
- [29] M. Li et al. “ATOM: a powerful gene-based association test by combining optimally weighted markers”. In: *Bioinformatics* 25.4 (2009), pp. 497–503.
- [30] R. Chowdhury et al. “Genetic analysis of variation in human meiotic recombination”. In: *PLoS genetics* 5.9 (2009), e1000648.
- [31] A. Fledel-Alon et al. “Variation in human recombination rates and its genetic determinants”. In: *PLoS ONE* 6.6 (2011), e20321.
- [32] A. Kong et al. “Sequence variants in the RNF212 gene associate with genome-wide recombination rate”. In: *Science* 319.5868 (2008), pp. 1398–401.
- [33] C. C. Laurie et al. “Quality control and quality assurance in genotypic data for genome-wide association studies”. In: *Genetic epidemiology* 34.6 (2010), pp. 591–602.
- [34] C. Minelli et al. “How should we use information about HWE in the meta-analyses of genetic association studies?” In: *International journal of epidemiology* 37.1 (2008), pp. 136–46.
- [35] G. Salanti, S. Sanderson, and J. P. Higgins. “Obstacles and opportunities in meta-analysis of genetic association studies”. In: *Genetics in medicine : official journal of the American College of Medical Genetics* 7.1 (2005), pp. 13–20.
- [36] A. Thakkinstian et al. “A method for meta-analysis of molecular association studies”. In: *Statistics in medicine* 24.9 (2005), pp. 1291–306.

- [37] E. Zintzaras and J. Lau. “Trends in meta-analysis of genetic association studies”. In: *Journal of human genetics* 53.1 (2008), pp. 1–9.
- [38] M. R. Munafo and J. Flint. “Meta-analysis of genetic association studies”. In: *Trends in genetics : TIG* 20.9 (2004), pp. 439–44.
- [39] J. Marchini and B. Howie. “Genotype imputation for genome-wide association studies”. In: *Nature reviews. Genetics* 11.7 (2010), pp. 499–511.
- [40] C. L. Kuo and E. Feingold. “What’s the best statistic for a simple test of genetic association in a case-control study?” In: *Genetic epidemiology* 34.3 (2010), pp. 246–53.

## **5.0 REGIONALLY SMOOTHED META-ANALYSIS (RSM) METHODS FOR GWAS DATASETS**

### **5.1 ABSTRACT**

Genome-wide association studies (GWAS) are proven tools for finding disease genes. Because effect sizes are often small, many of the most successful GWAS studies have been meta-analyses that combine results from as many as 20 or more different cohorts. Often the component studies are performed on different chips with minimal SNP overlap. In some cases, raw data is not available for imputation so that only the genotyped SNP results can be used in meta-analysis. To our knowledge, there is no available method to overcome this situation. In this study we propose new methods for GWAS meta-analysis combining different GWAS with minimum SNP overlap when imputation is not an option.

This new regionally smoothed meta-analysis (RSM) method is a two-stage method. In the first stage, we divide the genome into windows and derive window-based p-values for each study. In the second stage we combine results for each window across studies. We compared several different procedures for both the first and second stages, as well as different window sizes, using test datasets from our human meiotic recombination GWAS.

### **5.2 INTRODUCTION**

Genome-wide association studies have popularly been used to map disease genes for human complex diseases. Large sample size is required to identify SNPs with moderate effect size.

Though the number of GWAS has increased exponentially over the last decade, sample sizes of most of the individual studies are not large enough to identify SNPs with small to moderate effects. GWAS investigators have addressed this problem by integrating different studies using meta-analysis to get better power. But our comprehensive review of GWAS meta-analysis papers showed that most of the studies are using a few simple meta-analysis methods and that there are a number of unresolved methodological issues in the proper application of these methods [1].

One of the methodological challenges we identified in our review was method for meta-analysis of studies that have been performed on different chips. It is very common to encounter this situation in combining GWAS studies. Because of different genotyping platforms there is often a large number of non-overlapping single nucleotide polymorphisms (SNPs) in different data sets. Even if different data sets are genotyped on the same chip but in different laboratories, differing quality control processes can also lead to non-matching SNP sets. The standard practice for GWAS meta-analysis is to impute the non-overlapping and missing SNPs in each study and perform single locus GWAS and then combine studies using any of the available standard meta-analysis methods. In some cases, however, raw data are unavailable, so that imputation cannot be performed. Moreover, if imputation is performed in regions where one or more datasets is uninformative due to few SNPs, imputation can give false negative results because meta-analysis treats the imputed data identically to the genotyped data. Thus imputation can actually be misleading if SNP sets do not match. In this study we propose a method to perform meta-analysis of GWAS with non-overlapping SNPs by doing regional smoothing.

Many previous methodological studies have proposed methods for combining results of a single GWAS across windows or groups of SNPs. The usual motivation for these methods is to increase biological significance and/or decrease the multiple testing burden. Many studies proposed different methods based on multiple SNPs or genomic regions for individual GWAS. With the technological advancement, now a day the denser SNP chips are available for analysis. Additional SNPs are adding new information and at the same time making the multiple comparison issue more critical. Availability of the denser SNP chips such as few

millions SNP chip, it is more and more imminent that single locus GWAS are not providing desirable efficiency considering power, false positives or true negatives. Too stringent cut-off points controlling the false positive rates, but at the same time it is increasing the probability of missing true positives. Multi-locus GWAS analysis often performed to examine the interactions though it is more challenging statistically, computationally and logistically [2, 3]. To trade-off between controlling false positives and missing true positives, many GWAS methods were proposed based on genomic regions instead of single SNP analysis.

The main objective of the genomic region based GWAS methods is to reduce the dimension and to incorporating information from the nearby SNPs, including LD structure information and detecting disease associated SNPs with more power. Different types of genomic region based GWAS has been proposed such as gene based methods [4, 5, 6, 7], haplotype based methods [8, 9, 10], pathway based methods [11, 12, 13, 14] and genomic region based methods [15, 16, 17]. Window based genome-wide association studies motivated us to propose regionally smoothed meta-analysis method for GWAS. In window based GWAS, different window types and sizes have been considered, such as non-overlapping fixed windows, overlapping sliding windows and overlapping variable sized sliding windows. Most of the studies aimed to capture the LD structure in the windows in finding associated disease SNPs. Depending on the principal components some studies tried to find an optimal window size. Window based GWAS methods are in used application, without an optimal solution for the window size question.

In our study, we considered varying window sizes and types as well as various methods for combining the results for the SNPs within the window. We then added an additional layer of meta-analysis to combine results for each window across studies, and we considered various methods for that combination as well. We refer to this as regionally-smoothed meta-analysis (RSM).



### 5.3 METHODOLOGY

Our proposed regionally smoothed meta-analysis (RSM) methods work on genomic intervals genome-wide and provides the ranks of the genomic intervals depending on the significance level of the association as an output. RSM methods are two-stage methods. In first stage, the whole genome is divided into fixed or sliding windows and we compute a summary of the window effect in each study individually. In the second stage, we merge the window effects across studies.

#### 5.3.1 First Stage

In the first stage we divide each chromosome into windows of a pre-decided base-pair length. The proposed statistics to use in the first stage to summarize the window effect in each study are as follows:

- (I) Fishers statistic (FS): Fishers statistic combines p-values by summing up the log scaled p-values and multiplying the sum by  $-2$ , assuming that all the p-values are independent. Under the null hypothesis, this statistic follows a chi-square distribution with two times the number of studies as the degrees of freedom. The equation of Fishers statistic is given below:

$$X^2 = -2 \sum_{i=1}^k \log(p_i) \sim \chi_{2k}^2,$$

where  $k$  is the number of studies to combine. In using this statistic to combine p-values for SNPs in a window, we know that the tests are not independent, so the asymptotic chi-square distribution does not hold. We use this statistic only as a scoring function, and do not assess statistical significance.

- (II) Minimum p (MP): Tippett's Minimum p-value [18] is defined as

$$X = \min_{1 < k < K} p_k.$$

- (III) Derived Minimum p (DMP): Tippets minimum p-value follows Beta(1;K) distribution under the null hypothesis. We used both the minimum p-value and the p-value obtained from the Beta distribution as our summary statistic.
- (IV) Mean of log of p (MLP): Arithmetic mean of the logarithm of the p-values in each window is used as a summary statistic which is given below:

$$X = \frac{1}{k} \sum_{i=1}^k \log(p_i).$$

### 5.3.2 Second Stage

After calculating the summary statistics for each window for each study, we combine different GWAS studies summary statistics for each window. In this study we used two statistics for combining studies in the second stage: Fishers statistic (FS) and Adaptively Weighted statistic (AWS). Due to limited time we did not experiment on the other statistics. FS is already described before and Adaptively Weighted statistic (AWS) is described below:

- (V) Adaptively Weighted Statistic (AWS): Adaptively Weighted Statistic [19] uses optimal weight for each study and is defined by  $X = \min_{w \in W} pU(u(w))$ , where  $w$  is the weight and  $u(w)$  is the observed statistic for the weight function  $U(w) = -\sum_{i=1}^k w_k \log(p_k)$  and  $W = \{w | w_i \in \{0, 1\}\}$ . Compare to the FS, AWS provides better power for a range of alternative hypothesis.

### 5.3.3 Window Type and Size

In the first stage of the RSM method we proposed two different types of windows. (1). Fixed Window Meta-analysis (FWM) and (2). Sliding window Meta-analysis (SWM). Statistic choice of first stage has an effect on choosing window type. Some of the statistics make none to very little change in the results for different window type. For example, if we choose MP or DMP in first stage, the change in the results should be very minimal whether we perform FWM or SWM unlike choosing FS or MLP statistic in first stage. As we already mentioned that we divide the whole genome, each chromosome at a time into some windows depending

on the base pair location. A fixed size window will slide by a smaller increment depending on the window size, which is constant throughout the genome. We experimented with different window sizes such as 50k, 75k, 100k, 300k, and 500k.

### **5.3.4 Performance Measure of RSM Methods**

In this proposed method we are not calculating the correct p-values for any statistics. Instead we use the statistics only for ranking and focus on the ranks of the regions. This approach is tailored to the most common goal of genomic analysis, which is to get the most correct ranked list of results for follow-up analysis.

Type I error: We are applying these methods as ranking procedures for prioritizing genomic regions for follow-up. As we have implemented them, they do not have correct type I error. We use the p-values only as ranking statistics, not as real p-values [20]. If one wanted to use these methods for hypothesis testing, permutation analysis would allow calculation of correct p-values.

Performance (power): Since we are applying these methods for ranking rather than for hypothesis testing, the performance needs to be judged based on the ranks each method produces. We consider a good procedure to be one that is able to put true-positive regions within the top 10 on a ranked list of regions (and the higher the better). It would also possible to calculate a positive prediction fraction [21] or other measure of ranking correctness.

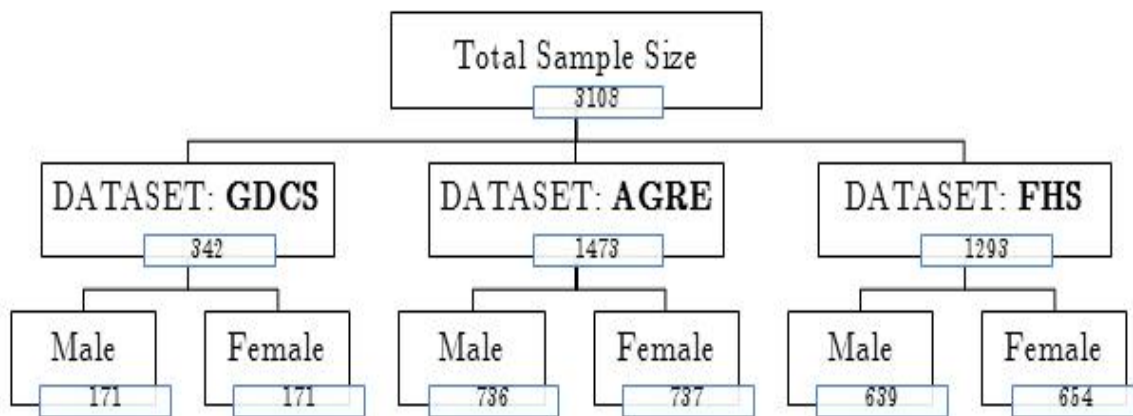
## **5.4 EXPERIMENTAL DESIGNS AND TESTING IN REAL DATA**

### **5.4.1 Introduction**

In Chapter 2, we aimed to find the genetic determinants of human meiotic recombination. We used three data sets from different populations. Two of the data sets were genotyped on the Illumina Human610-Quad Beadchip and the other one was genotyped on the 5.0 Affymetrix chip. The sample sizes of the three studies are presented in Figure 5.1. Because

of the different SNP sets on the two chips, in Chapter 2 we used two of the data sets for meta-analysis and then used the third to replicate the results. Here we include all three of the data sets in order to test our meta-analysis methods. The advantage of these datasets for testing our methods is that they contain known true positives. As described in Chapter 2, there are several well-established recombination genes, and we can use this dataset to test for our ability to find (highly-rank) those gene regions for various phenotypes. Moreover, a known problem with these datasets is that the FHS dataset, genotyped on the Affymetrix 5.0 chip, does not have any SNPs in the *PRDM9* gene, the gene that is highly associated with recombination in hotspots.

For the current experiments, we chose three of the phenotypes discussed in Chapter 2: average recombination count (ARC), hotspot percent (HS\_PCT), and non-hotspot count (NHS\_CNT). Average recombination count (ARC) of a parent is calculated as the total number of recombination events in all children divided by the number of children. Hotspot percent (HS\_PCT) is the percent of recombination events in the hotspot areas. Non-hotspot count (NHS\_CNT) is the average of the recombination event counts in non-hotspot areas. Phenotypes are described in detail in Chapter two.



**Figure 5.1:** Sample size distribution of different studies.

There are several known recombination genes associated with two of the phenotypes. We used those genes to test and compare our methods. SNPs in *RNF212* are known to be associated with total recombination, with those associations being somewhat different

in males and females. SNPs in *PRDM9* are known to be associated with recombination in historical hotspots in both males and females. That association is strongest for the phenotype percent of recombination in hotspots but can also be observed in count of recombination in hotspots and count of recombination outside of hotspots. There are also SNPs inside a chromosome 17 inversion region that are associated average recombination in females. We tested the ability of all of our methods (including variations on methods such as window size) to highly rank (and thus detect) these region/phenotype combinations.

#### 5.4.2 Results

Our first tested phenotype was HS\_PCT and we looked at the rank of the *PRDM9* gene. Performance of different methods is listed in Table 5.1. When we considered a smaller window size such as 50k, the gene was split in two windows. Ranks of the two intervals are then listed in table. Sometimes due to the positioning of the window on the gene, it may split in two or more windows even if the windows are longer. The results show that use of MP and DMP statistics in first stage and FS in the second stage performed very well irrespective of the window size. The MLP statistic performed worse with increasing window size. FS in both stages is also poorer with a bigger window size.

**Table 5.1:** Ranks of *PRDM9* gene for HS\_PCT phenotype

Stages		Window Size				
First stage	Second stage	50k	75k	100k	300k	500k
Fishers statistics (FS)	FS	1, 74	20, 6	1	6	20
Minimum p (MP)	FS	1	1	1	1	1
Derived minimum p (DMP)	FS	1, 7	1, 6	1, 58	1	1
Mean of log of p (MLP)	FS	1	3	2	8	16

Our second phenotype of interest was NHS\_CNT and we again looked at the ranks for the *PRDM9* gene. This is a more challenging test for the methods than HS\_PCT, because the effect size of the gene on this phenotype is smaller. Table 5.2 shows the ranks of the *PRDM9* gene for different window sizes and for different statistics in different stages. The result shows that DMP statistics gave lowest rank for the *PRDM9* gene. In fact, only DMP

performed well enough that the gene would be likely to be detected in a GWAS, although min P is close. Considering window size, different statistics provided lowest rank for different window sizes.

**Table 5.2:** Ranks of *PRDM9* gene for NHS\_CNT phenotype

Stages		Window Size				
First stage	Second stage	50k	75k	100k	300k	500k
Fishers statistics (FS)	FS	217	85	90	115	259
Minimum p (MP)	FS	20	13	10	4	139
Derived minimum p (DMP)	FS	8	7	1	2	56
Mean of log of P (MLP)	FS	62	89	19	104	243

Table 5.3 shows the results for the *RNF212* gene for the ARC phenotype. Again here the effect size is quite large, so all methods perform well. FS in both stages was the best-performing method, followed by MP in the first stage and FS in the second stage. Among the window sizes, 100k performed best.

**Table 5.3:** Ranks of *RNF212* gene for ARC phenotype

Stages		Window Size				
First stage	Second stage	50k	75k	100k	300k	500k
Fishers statistics (FS)	FS	1	2	1	3	3
Minimum p (MP)	FS	2	2	2	3	3
Derived minimum p (DMP)	FS	4, 6	4, 6	3	3, 4	3, 7
Mean of log of P (MLP)	FS	2, 4	3, 4	2	2	3

### 5.4.3 Discussion

In this study, we considered three phenotypes and five different window sizes. Results show that MP worked best across phenotypes even when the effect size was relatively smaller such as the association between NHS\_CNT and *PRDM9* gene. For *PRDM9* and *RNF212*, window size 100k worked best. But in general which window size will perform best will also depend on the linkage disequilibrium (LD) blocks or gene size. A window size that can capture the whole gene or LD block may yield lowest rank.

The choice of fixed window and sliding window and the choice of statistics in different stages are linked. If we choose MP or DMP in the first stage, it may not necessary to choose sliding window. Instead window size choice will be critical.

Program run time for fixed windows is faster than for sliding windows. Running fixed windows for multiple window sizes may help us to choose the optimum window size for sliding windows to refine the location on the genome.

We tested four different statistics in first stage. Among the four statistics, MP and DMP will identify the SNPs with significant p-values in few of the studies. Fisher's statistics and MLP statistic is also affected by small p-values in a subset of studies. If we want to identify SNPs with moderate effect size across studies, we need to think critically in choosing statistics in first stage.

In the second stage we show results for FS only. We investigated the scope of AWS too, but due to computational difficulty, we could not estimate p-value less than on the order of  $10^{(-7)}$ . So it became difficult to apply the ranking approach in this method. With some modification this method might be useful in second stage.

## 5.5 RSM SOFTWARE

To implement the RSM methods discussed in the previous sections, we are developing an R package named rsmGWAS. This section describes how different methods were implemented in R with the program algorithm and example.

### 5.5.1 Program Description

**5.5.1.1 Program input files** This program uses GWAS result files as input files. Currently this program is designed for PLINK [22] GWAS output files. But the program will work for any output file in the following format. Each of the GWAS result files should have the following information and all of the GWAS results files should be in the same format.

1. Chromosome Number
2. SNP names
3. Base pair location
4. Effect size of each SNP
5. SE of the effect size of each SNP
6. P-value of each SNP

Effect size and the standard error of the effect sizes of each SNP are necessary if the user wants to implement the fixed effect or random effect GWAS meta-analysis.

**5.5.1.2 Choice of Window type** Two types of windows were considered, such as fixed window and sliding window. If fixed window method is preferred, then the program needs one parameter of the window size. In the case of the sliding window method, the program requires two parameters (window size and the increment size).

**5.5.1.3 Methods choice** In this package we provide several statistic choices in both stages. The implemented methods are shown in figure 4.1 except fixed effect method and random effect method. In the first stage there are four methods; such as FS, MP, DMP or MLP and in the second stage there are two methods available (FS or AWS) to combine different studies. Fixed effect method and random effect method can also be used in first stage.

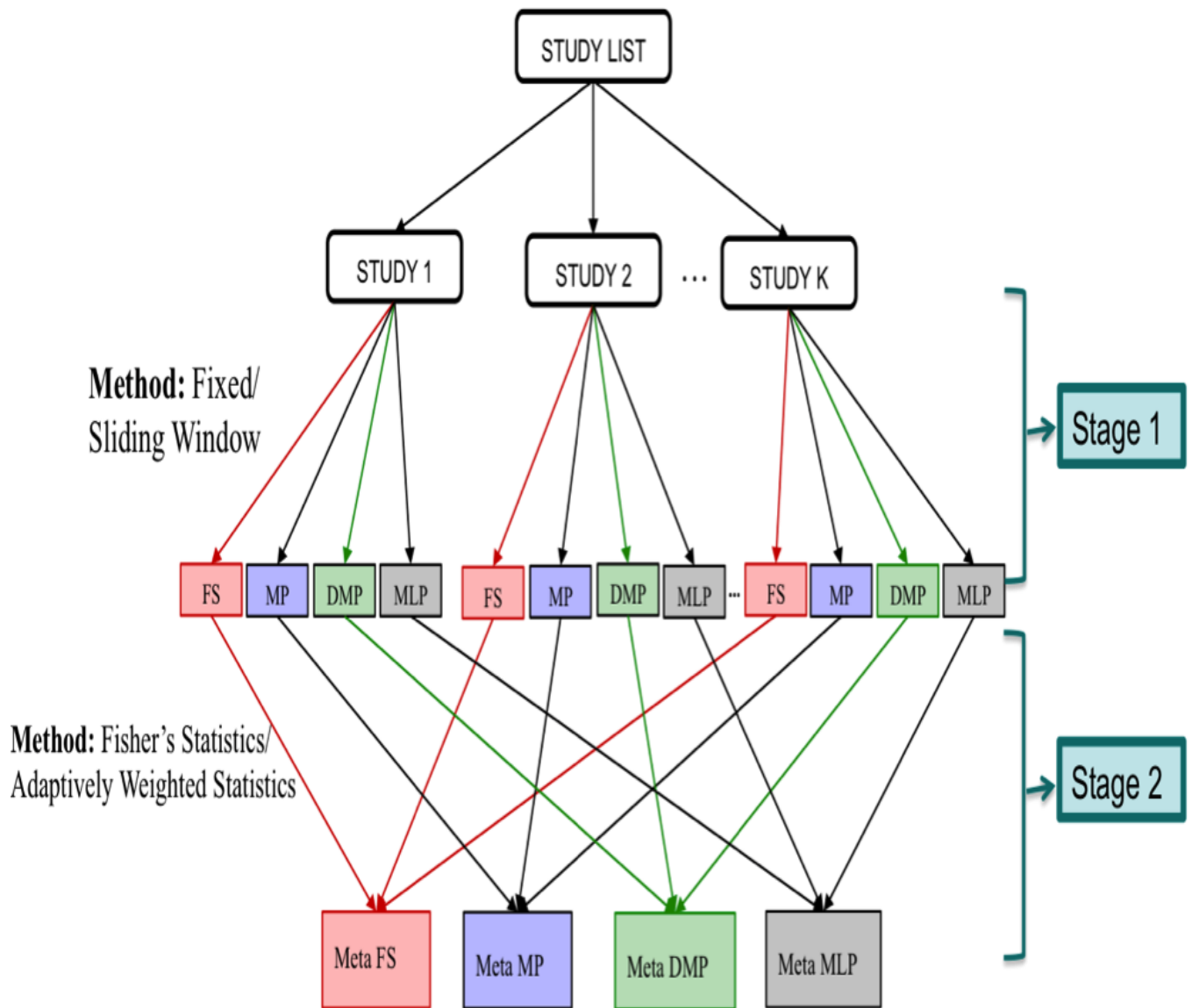
**5.5.1.4 Program Workflow** 1. Copy all the GWAS results file in one directory 2. Write a tab delimited text file listing the names of the GWAS results files 3. Source the Function 4. Call the function using different parameters

## 5.5.2 Example

Making the list file

**Example 1:** List file example





**Figure 5.2:** Program algorithm of RSMgwas package.

study1\_GWAS\_result.qassoc

study2\_GWAS\_result.qassoc

.

.

studyk\_GWAS\_result.qassoc

**Example 2:** Output file after first stage

**Table 5.4:** Example output file after stage one.

CHR	BP_range	winStart	winEnd	Fisher_P_1	.	Fisher_P_6
1	0_50000	0	50000	NA	.	0.288187997
1	1.01e+08_101050000	1.01E+08	101050000	0.038226405	.	0.974085409
1	1.02e+08_102050000	1.02E+08	102050000	0.951274052	.	0.703104411
1	1.03e+08_103050000	1.03E+08	103050000	0.246770384	.	0.65105545
1	1.04e+08_104050000	1.04E+08	104050000	NA	.	NA

**Example 3:** Output file after second stage

**Table 5.5:** Example output file after stage two.

CHR	BP_range	.	Fisher_P_6	FM_stat	df	P.value
1	0_50000	.	0.288187997	4.694759133	4	0.320074421
1	1.01e+08_101050000	.	0.974085409	16.17494896	12	0.183353853
1	1.02e+08_102050000	.	0.703104411	6.782033145	12	0.87167512
1	1.03e+08_103050000	.	0.65105545	8.011705054	12	0.784215008
1	1.04e+08_104050000	.	NA	0	0	NA

### 5.5.3 Discussion

rsmGWAS package implemented a two-stage meta analysis method using different statistics. Any number of GWAS studies can be combined with this package. Only a few of the standard statistics are implemented in this program. Many extensions such as sliding window option, fixed effect meta-analysis etc. are planned.

## 5.6 REFERENCES

- [1] F. Begum et al. “Comprehensive literature review and statistical considerations for GWAS meta-analysis”. In: *Nucleic acids research* (2012).
- [2] W. S. Bush and J. H. Moore. “Chapter 11: Genome-wide association studies”. In: *PLoS Computational Biology* 8.12 (2012), e1002822.
- [3] J. H. Moore and M. D. Ritchie. “STUDENTJAMA. The challenges of whole-genome approaches to common diseases”. In: *JAMA : the journal of the American Medical Association* 291.13 (2004), pp. 1642–3.
- [4] D. Curtis, A. E. Vine, and J. Knight. “A simple method for assessing the strength of evidence for association at the level of the whole gene”. In: *Advances and applications in bioinformatics and chemistry : AABC* 1 (2008), pp. 115–20.
- [5] D. P. Hibar et al. “Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects”. In: *NeuroImage* 56.4 (2011), pp. 1875–91.
- [6] M. Li et al. “ATOM: a powerful gene-based association test by combining optimally weighted markers”. In: *Bioinformatics* 25.4 (2009), pp. 497–503.
- [7] L. Ma, A. G. Clark, and A. Keinan. “Gene-based testing of interactions in association studies of quantitative traits”. In: *PLoS genetics* 9.2 (2013), e1003321.
- [8] A. Iliadis, D. Anastassiou, and X. Wang. “Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled DNA data”. In: *BMC genetics* 13 (2012), p. 94.
- [9] A. J. Lorenz, M. T. Hamblin, and J. L. Jannink. “Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley”. In: *PLoS ONE* 5.11 (2010), e14079.
- [10] X. Wan et al. “HapBoost: A fast Approach to Boosting Haplotype Association Analyses in Genome-Wide Association Studies”. In: *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* (2013).

- [11] M. Kanehisa and S. Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.
- [12] M. X. Li, J. S. Kwan, and P. C. Sham. “HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis”. In: *American journal of human genetics* 91.3 (2012), pp. 478–88.
- [13] K. Yu et al. “Pathway analysis by adaptive combination of P-values”. In: *Genetic epidemiology* 33.8 (2009), pp. 700–9.
- [14] F. Zhang and R. Drabier. “IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis”. In: *BMC bioinformatics* 13 Suppl 15 (2012), S7.
- [15] Y. C. Lin et al. “Using maximal segmental score in genome-wide association studies”. In: *Genetic epidemiology* 36.6 (2012), pp. 594–601.
- [16] Q. Sha, R. Tang, and S. Zhang. “Detecting susceptibility genes for rheumatoid arthritis based on a novel sliding-window approach”. In: *BMC proceedings* 3 Suppl 7 (2009), S14.
- [17] R. Tang et al. “A variable-sized sliding-window approach for genetic association studies via principal component analysis”. In: *Annals of human genetics* 73.Pt 6 (2009), pp. 631–7.
- [18] L.H.C. Tippett. *The Methods in Statistics*. 1st ed. Williams and Norgate, Ltd., 1931.
- [19] J. Li and G. C. Tseng. “An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies.” In: *The Annals of Applied Statistics* 5.2A (2011), pp. 994–1019.
- [20] M. H. Gail et al. “Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies”. In: *Biostatistics* 9.2 (2008), pp. 201–15.
- [21] M. H. Gail et al. “Probability that a two-stage genome-wide association study will detect a disease-associated snp and implications for multistage designs”. In: *Annals of human genetics* 72.Pt 6 (2008), pp. 812–20.
- [22] S. Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. In: *American journal of human genetics* 81.3 (2007), pp. 559–75.

## 6.0 CONCLUSIONS

### 6.1 DISSERTATION CONCLUSIONS

This integrated dissertation is comprised of four projects. All four projects contributed in different ways to fulfill the main objective of methods development to increase sample size of recombination studies. Projects one and two focused on meiotic recombination. Projects three and four is partly motivated by the first project but subsequently branched out into a more general methodological issue of meta-analysis for genomic data.

In this dissertation, we addressed five specific questions: (1)(a). What are the roles of known genes and genomic regions beyond known phenotypes? (1)(b). Are there additional genes associated with meiotic recombination? (2). Can we score recombination for complex pedigree structures to increase the sample size of recombination of GWAS? (3). Can existing methods of GWAS meta-analysis handle missing data without imputation? (4) How can GWAS meta-analysis be done when studies are genotyped on different chips and imputation is not an option?

In Chapter two, to assess the additional role of already known recombination genes such as *RNF212* and *PRDM9* genes and to identify additional genes associated with meiotic recombination, we scored recombination using existing methods and estimated some new recombination phenotypes (HS\_CNT, NHS\_CNT, MOTIF) and performed GWAS in two new data sets. We performed sex specific and gender pooled analysis for each phenotype and combined them using GWAS fixed effect meta-analysis. We performed a qualitative gene-based replication in a third data set. Our results showed that *RNF212* and *PRDM9*

have broader roles beyond our previous knowledge. We found that *PRDM9* has gender specific effects on recombination events in non-hotspot areas. We suggested several new candidate genes/regions for different phenotypes (Tables 2.1-2.5 in Chapter two).

In Chapter three, we developed methods for scoring recombination for three-generation families and two-generation families with half-siblings using a SNP-streak method. Our methods allow different numbers of missing persons in the pedigree.

Chapter four is a published comprehensive literature review of GWAS meta-analysis. We considered 249 papers for this review most of which are application paper (91%) and 4% novel methodology papers. In this paper we presented an overview of the most widely used GWAS meta-analysis methods. We also listed GWAS databases and available software for GWAS meta-analysis. We performed a case study in which we compared and contrasted different meta-analysis methods. To account for heterogeneous genetic effects, random effects models are popularly used. In our case, where the number of studies was not large, we showed that the choice of fixed effects is better than the random effects model. We suggested a mixture of different fixed effect models instead of the random effect models. We also discussed different unresolved issues such as data cleaning, imputation, genetic model choices, study heterogeneity etc. related to GWAS meta-analysis.

In Chapter five, we proposed a novel method for GWAS meta-analysis when data sets are genotyped on different chips with minimal overlap. Instead of single SNP meta-analysis, this method works on genome segment to overcome the missing data issue with an advantage of reduced multiple testing burden. This method is a two-stage method. In first stage we divide each chromosome into windows with a pre-decided window size, and used a summary statistic to summarize the window effect. We repeat this step for each study. In second stage, studies are combined across windows using Fishers statistic or adaptively weighted statistics. We applied these methods to meiotic recombination data and the result is promising. Among different statistics, minimum p provided best result in detecting the true positive for first stage. Our results showed that window size 100k works best for our two chosen genes. We are developing a R-package rsmGWAS.

## 6.2 STRENGTHS AND SHORTCOMINGS

The main strength of this dissertation is in achieving the goal in terms of methodology and also in terms of genetics. In terms of genetics of human meiotic recombination this dissertation provided deeper insight of some of the known genes, such as *RNF212* and *PRDM9*, and their gender-specific effects. In methodological development, the main strength of this dissertation is that we developed new methods for scoring recombination for complex pedigrees and also developed a new method for GWAS meta-analysis when samples are genotyped on different chips. Both will contribute to increasing sample size for future recombination studies. This is a well balanced research of application and method development.

Project one presented in chapter two has several strengths. We explored new aspects of recombination and considered new phenotypes. Gender-specific and gender-pooled analysis of each phenotype provided deeper insight into each phenotype. A limitation is that the different SNP set in the FHS data set, replication of our study finding was limited to gene-based qualitative analysis. Due to small sample size and non-availability of raw genetic data of two studies, the analysis of gene gene interaction was limited.

The main strength of project two on recombination scoring method for complex pedigrees is that it will increase the use of publicly available family data sets of different structures for recombination study. This newly developed method is based on a SNP-streak method and denser SNP chip data will be a plus for this method. With a denser SNP chip, the recombination scoring will be more accurate. Different complex pedigrees with varying numbers of missing persons will allow use of a wide range of family data sets including three-generation and two-generation families with half-siblings or a mixture of both.

Comprehensive literature review of GWAS meta-analysis presented in chapter four is based on a large number of GWAS meta-analysis papers. One of the important strengths of this project is that it discussed thoroughly some of the unsolved open questions in the area of GWAS meta-analysis and presented a case-study comparing different methods and showed their shortcomings.

Regionally smoothed meta-analysis (RSM), a new method for GWAS meta-analysis, has multiple strengths. We can use RSM without imputation and it reduces the burden of multiple testing issues considerably. We are still improving the method. Currently this method uses medium window sizes which may not perform well for a bigger size LD blocks or for a large gene. Choice of window type (fixed window or sliding window) is also subjective and sliding window method is computationally slow.

### 6.3 FUTURE DIRECTION

Four different projects in this dissertation can be extended along several avenues, including methodology, application and software development.

In project one, we looked at the genes regulating/controlling the hotspot usage of the recombination events. In the near future, I want to investigate the possible relationship between recombination hotspot areas and copy-number-variation hotspot areas and their genetic basis. I believe this study might answer some unanswered questions such as differential hotspot usage by sex and in parents and children.

Project two is a collaborative work in which I developed methods for scoring recombination and the collaborator did software implementation. In the future we can extend our method for next-generation sequencing data. With the use of newly developed recombination scoring methods in a three-generation family we might look at the genetic material transfer variation between male and female.

Regionally smoothed meta-analysis is a new method, which can be improved in different directions. I would like to extend my methodological work on meta-analysis to a broad range of data types such as next generation sequencing data. I would also like to investigate adapting methods to summarize rare and common variants together in first stage of the method and also would like to incorporate other sophisticated models (fixed effect, random effects, mixed models etc.) in different stages of the proposed meta-analysis method.



## APPENDIX A

### GENETICS OF MEIOTIC RECOMBINATION: GENOME WIDE ASSOCIATION STUDIES FOR RECOMBINATION PHENOTYPES

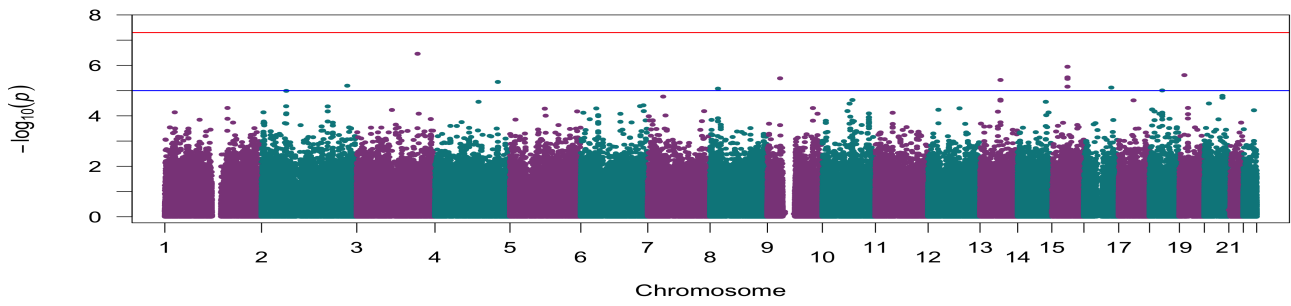


Figure A.1: QQ plot ARC(female)

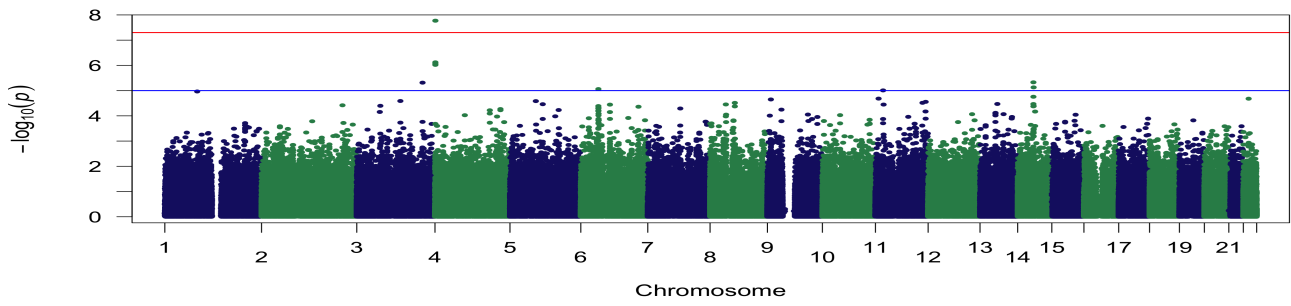


Figure A.2: QQ plot ARC(male)

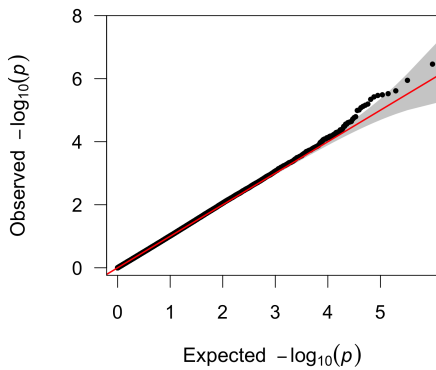


Figure A.3: QQ plot ARC(female)

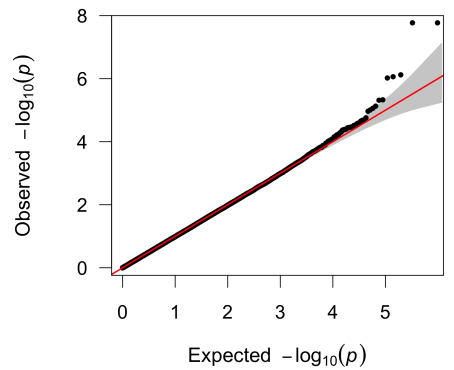
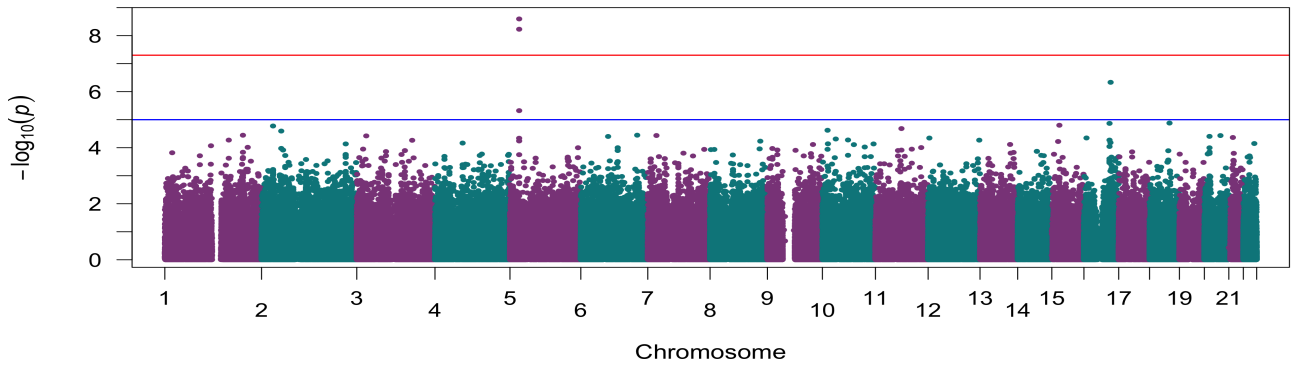
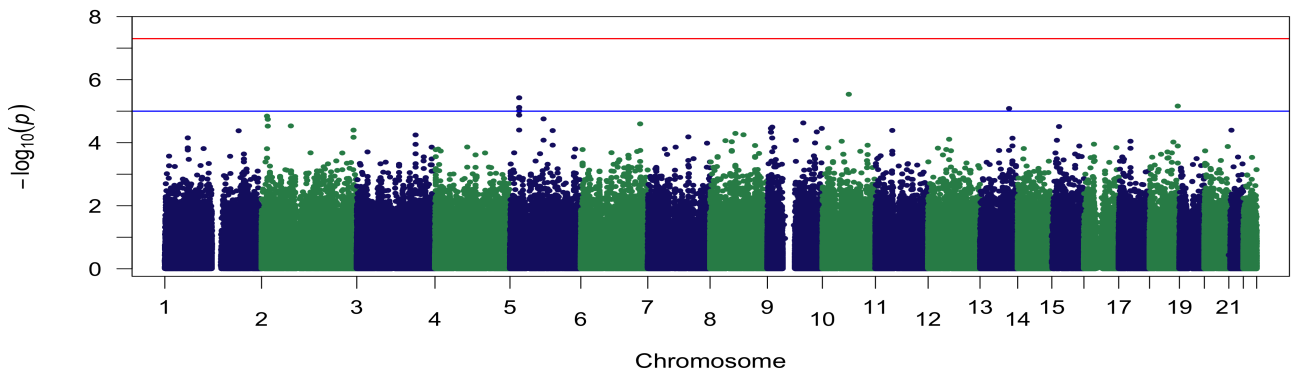


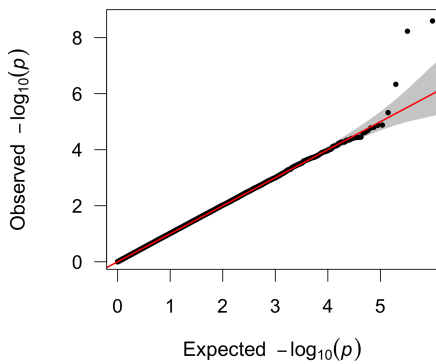
Figure A.4: QQ plot ARC(male)



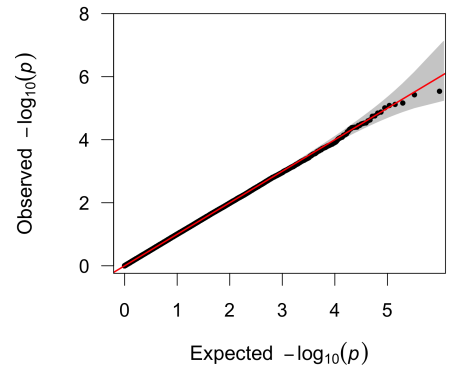
**Figure A.5:** Manhattan plot of phenotype HS\_PCT (female)



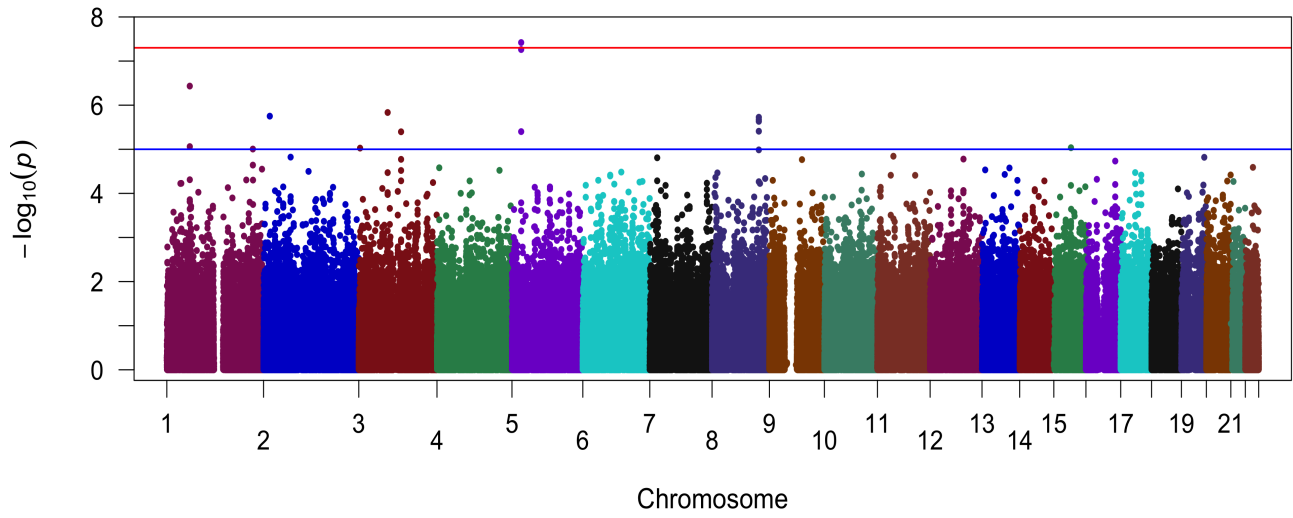
**Figure A.6:** Manhattan plot of phenotype HS\_PCT (male)



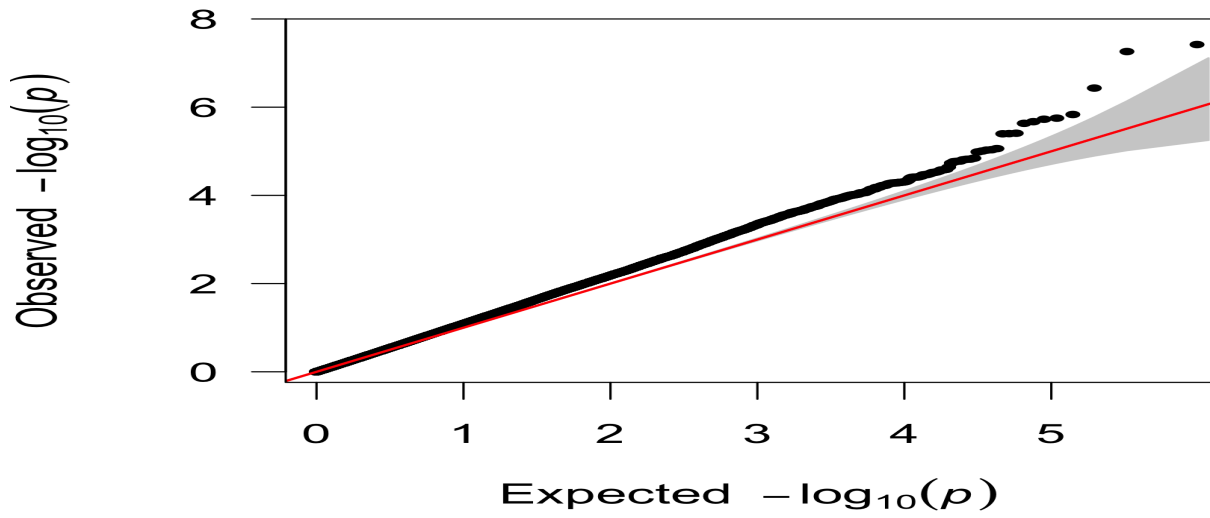
**Figure A.7:** QQ plot HS\_PCT(female)



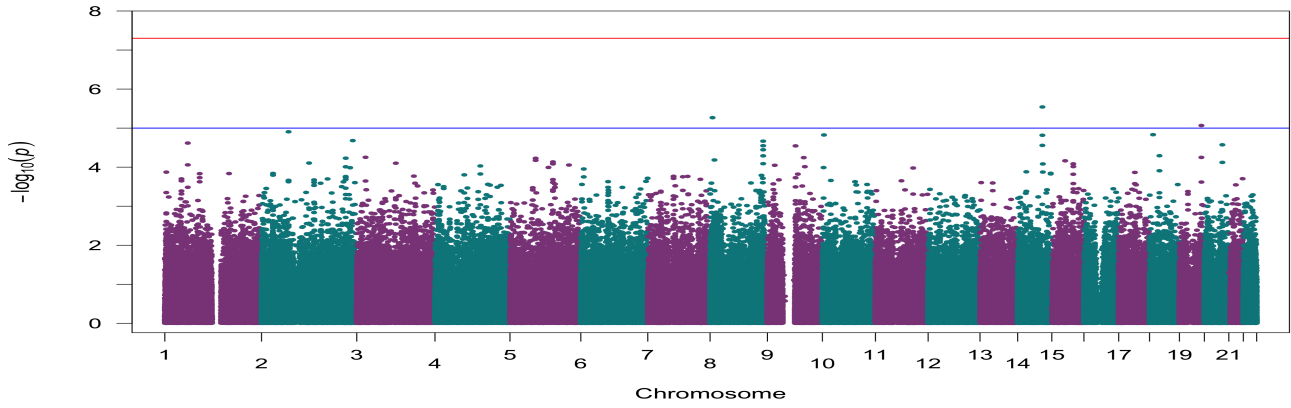
**Figure A.8:** QQ plot HS\_PCT(male)



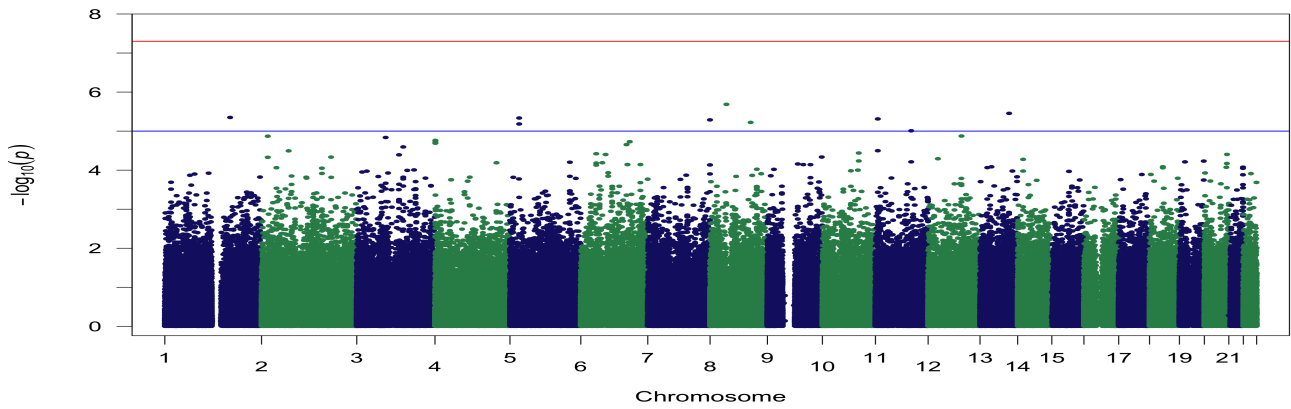
**Figure A.9:** Manhattan plot of phenotype HS\_CNT (combined)



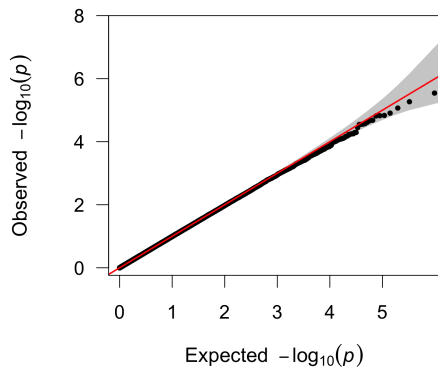
**Figure A.10:** QQ plot HS\_CNT(combined)



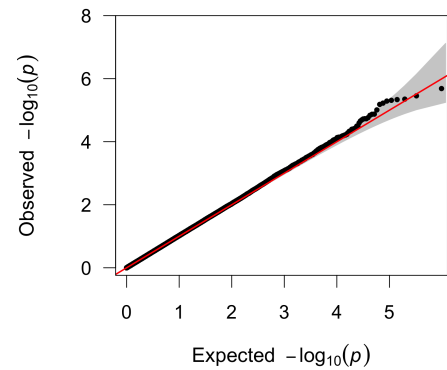
**Figure A.11:** Manhattan plot of phenotype HS\_CNT (female)



**Figure A.12:** Manhattan plot of phenotype HS\_CNT (male)



**Figure A.13:** QQ plot HS\_CNT(female)



**Figure A.14:** QQ plot HS\_CNT(male)

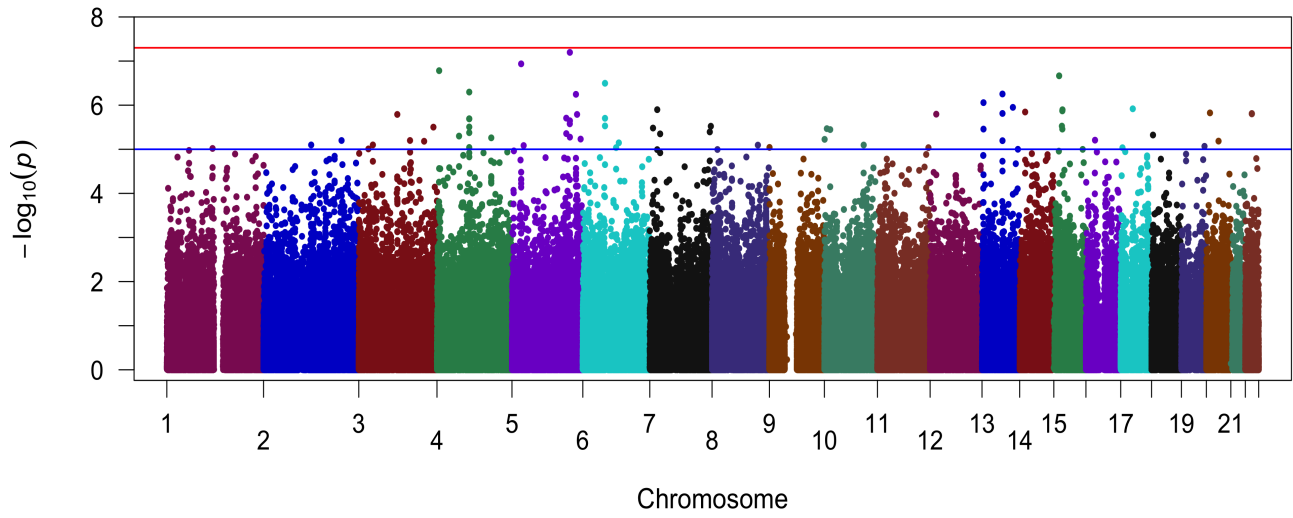


Figure A.15: Manhattan plot of phenotype NHS\_CNT (combined)

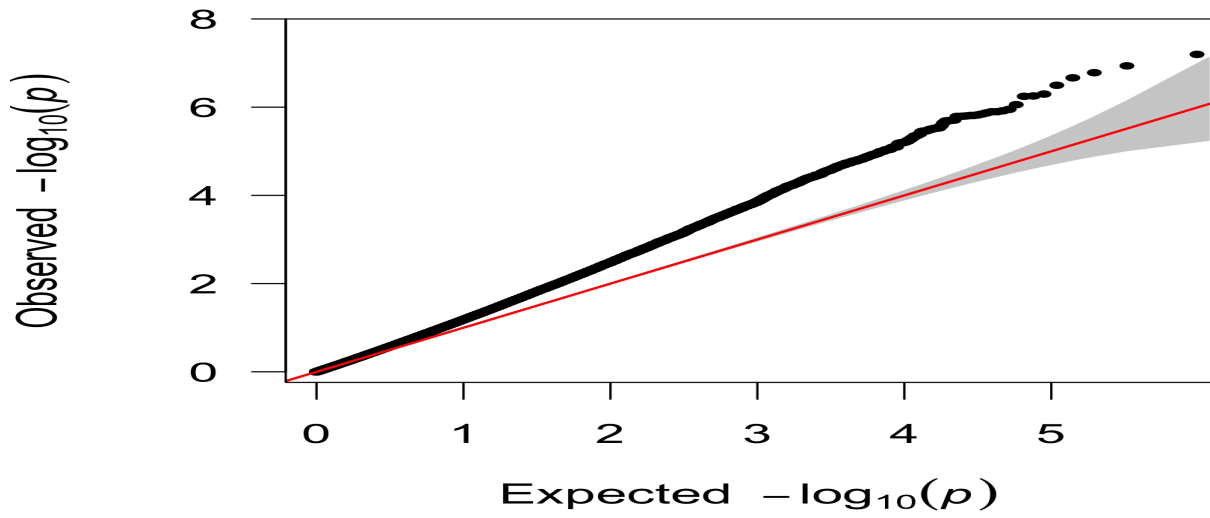


Figure A.16: QQ plot NHS\_CNT(combined)

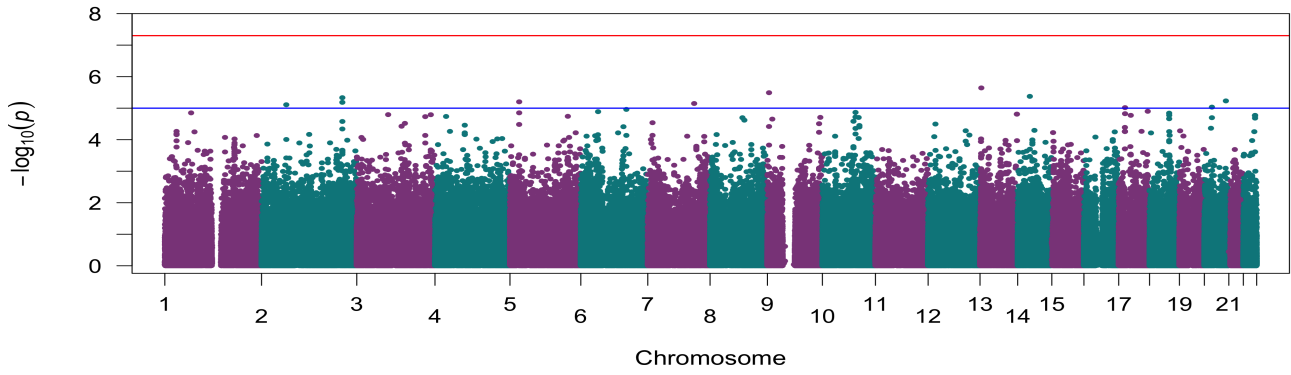


Figure A.17: Manhattan plot of phenotype NHS\_CNT (female)

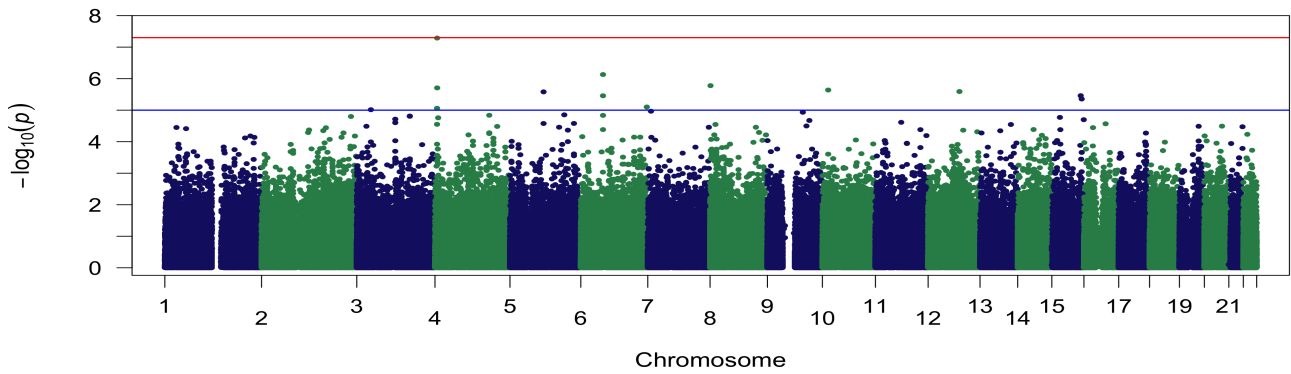


Figure A.18: Manhattan plot of phenotype NHS\_CNT (male)

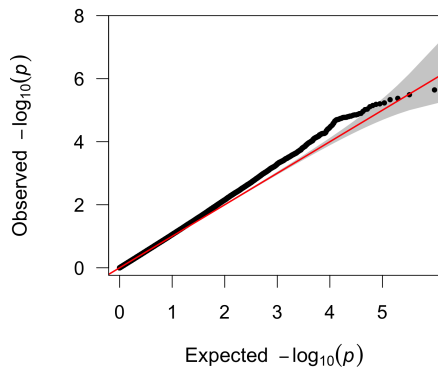


Figure A.19: QQ plot NHS\_CNT(female)

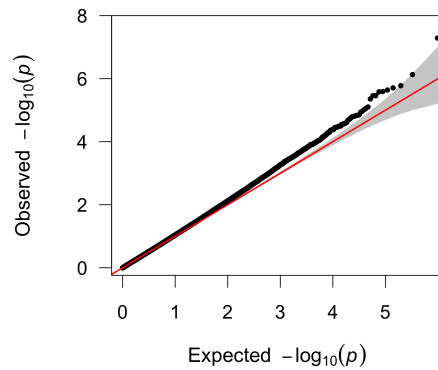


Figure A.20: QQ plot NHS\_CNT(male)

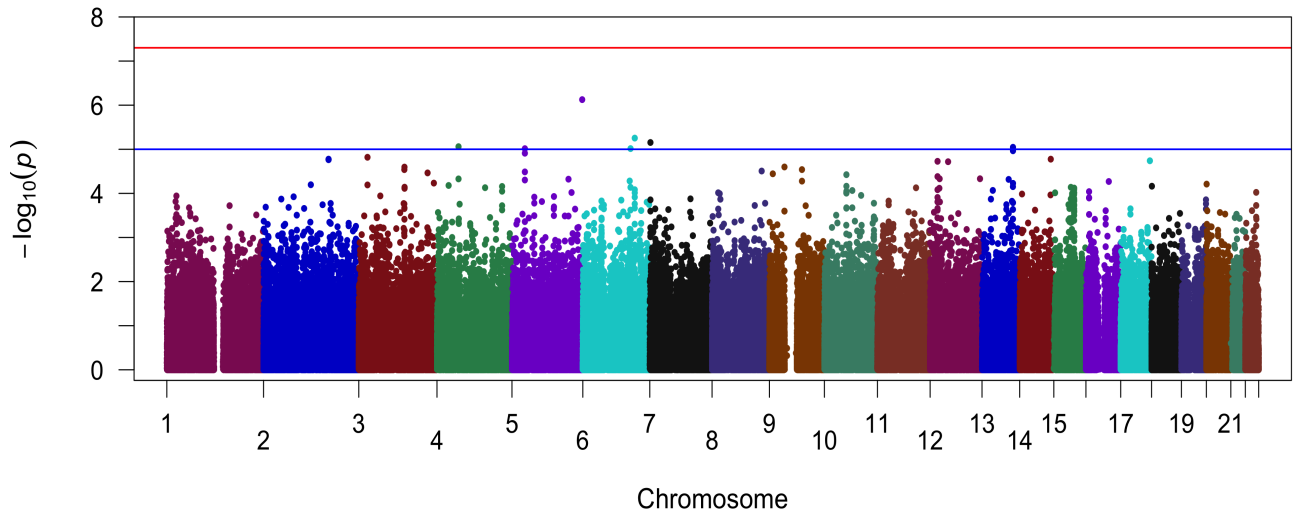


Figure A.21: Manhattan plot of phenotype MOTIF (combined)

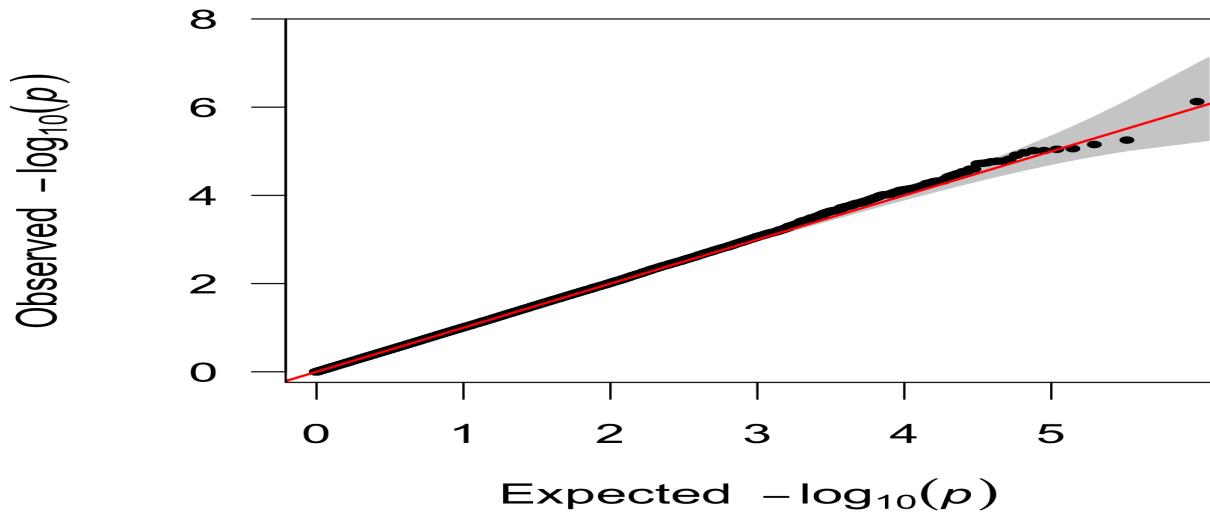


Figure A.22: QQ plot MOTIF(combined)



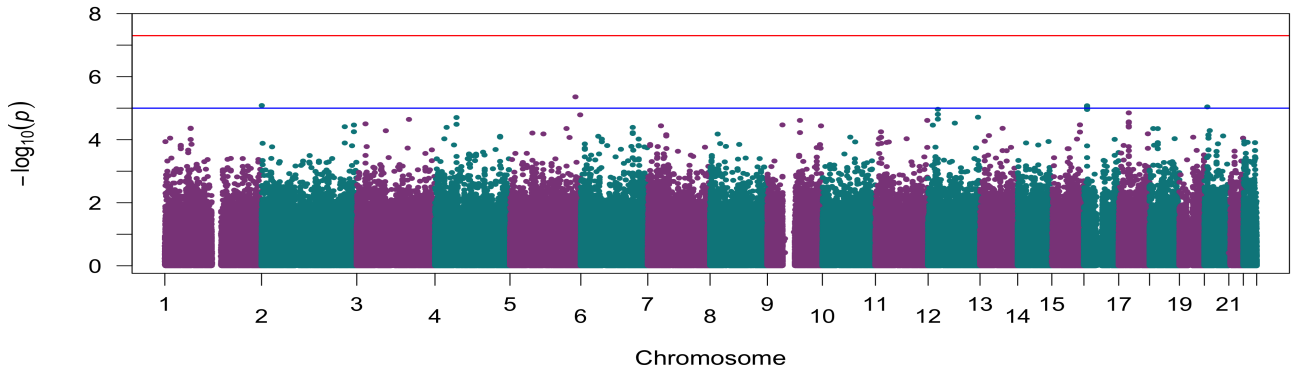


Figure A.23: Manhattan plot of phenotype MOTIF (female)

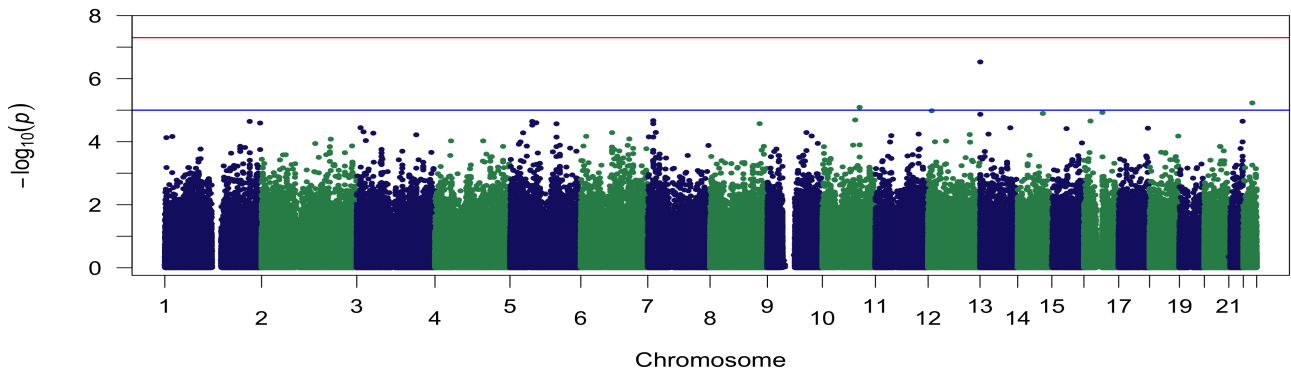


Figure A.24: Manhattan plot of phenotype MOTIF (male)

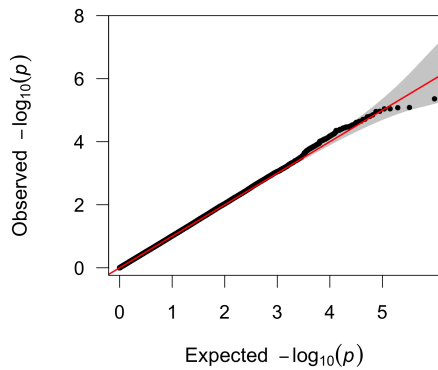


Figure A.25: QQ plot MOTIF (female)

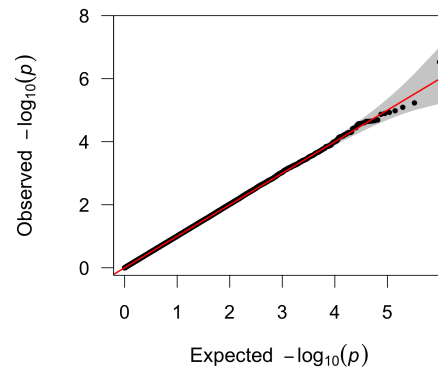


Figure A.26: QQ plot MOTIF (male)

# A.1 LOCUSZOOM PLOT OF PREVIOUSLY REPORTED GENES AND SNPS IN OUR STUDY

## A.1.1 Phenotype: ARC

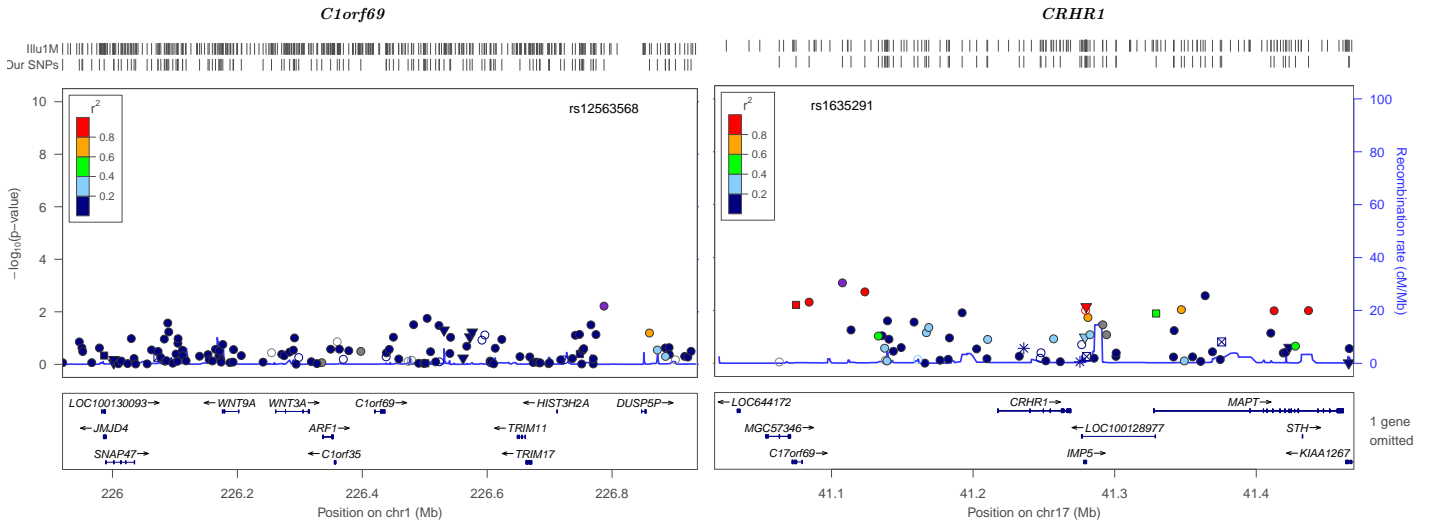


Figure A.27: ARC(female)

Figure A.28: ARC(female)

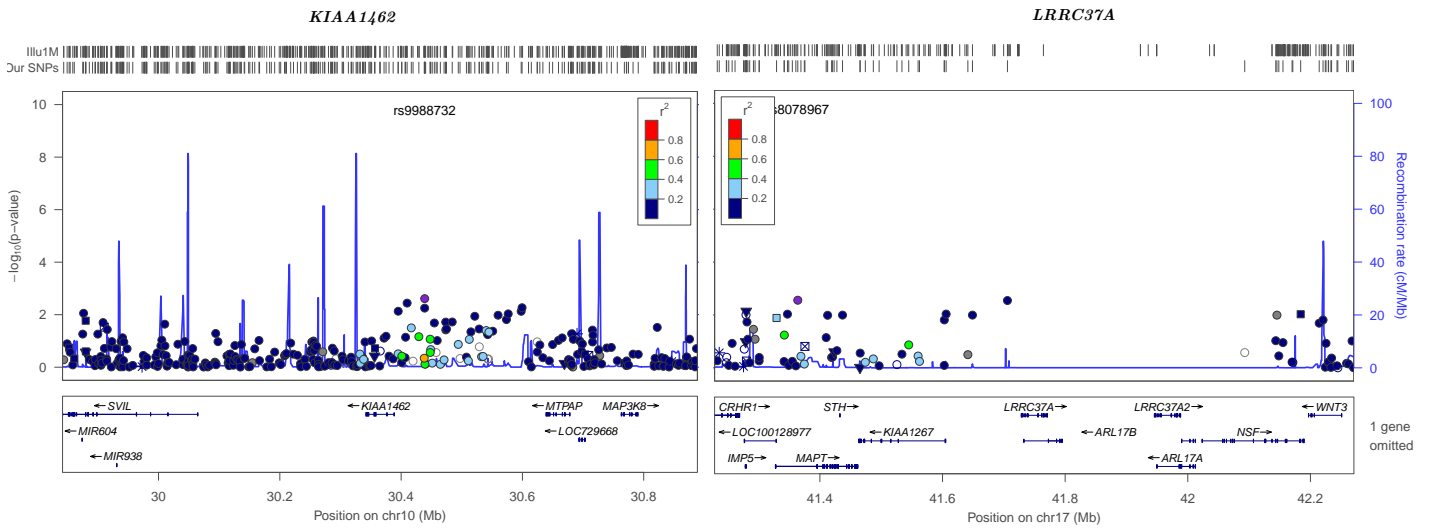


Figure A.29: ARC(female)

Figure A.30: ARC(female)

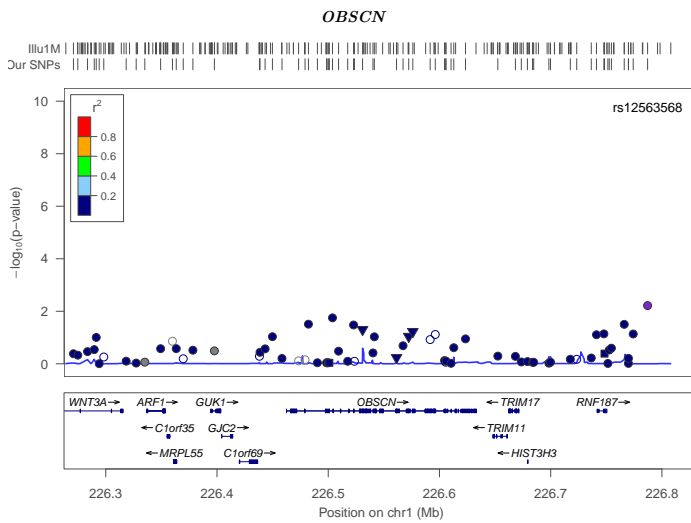


Figure A.31: ARC(female)

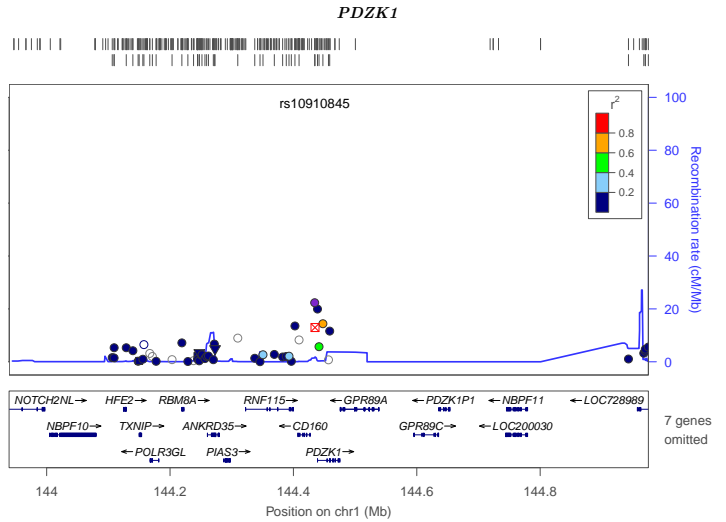


Figure A.32: ARC(female)

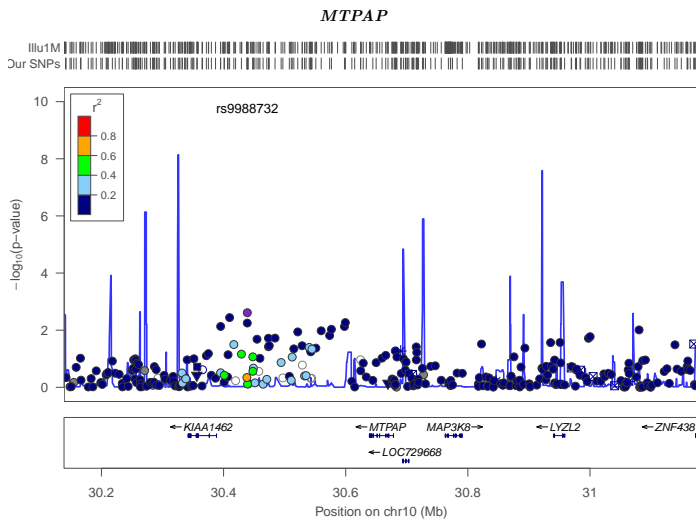


Figure A.33: ARC(female)

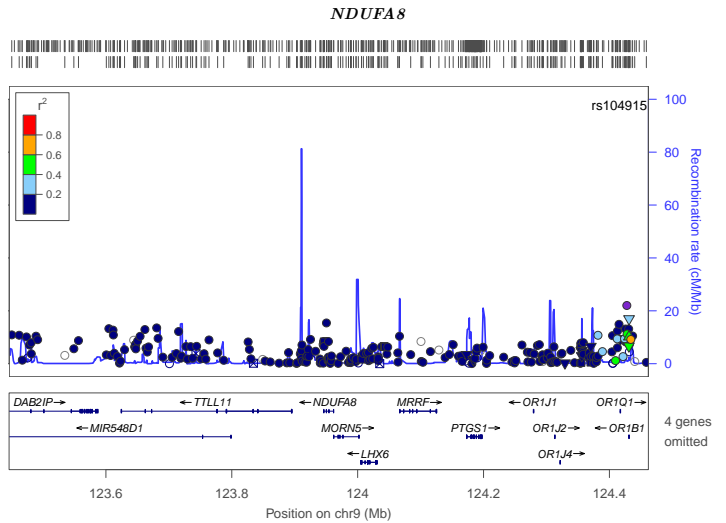


Figure A.34: ARC(female)

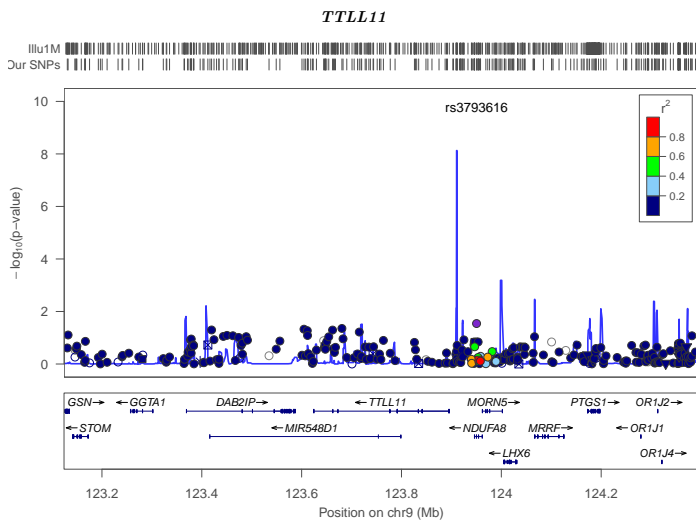


Figure A.35: ARC(female)

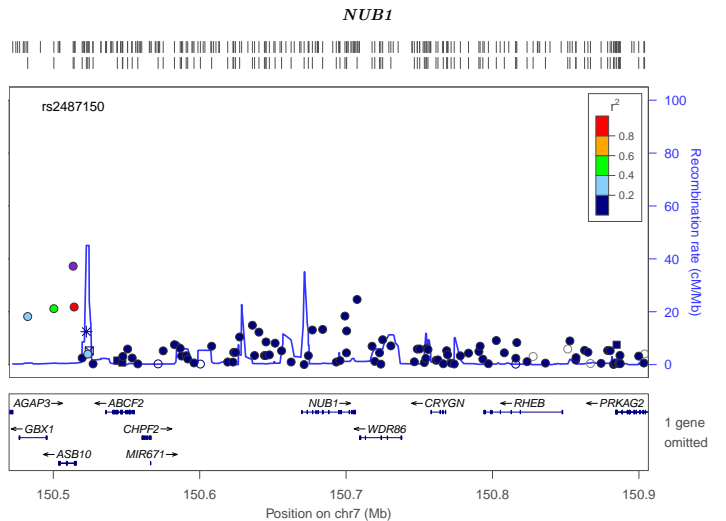


Figure A.36: ARC(male)

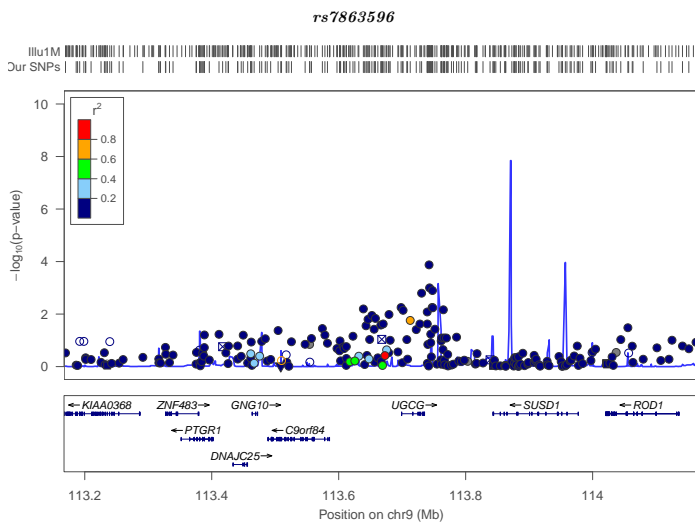


Figure A.37: ARC(male)

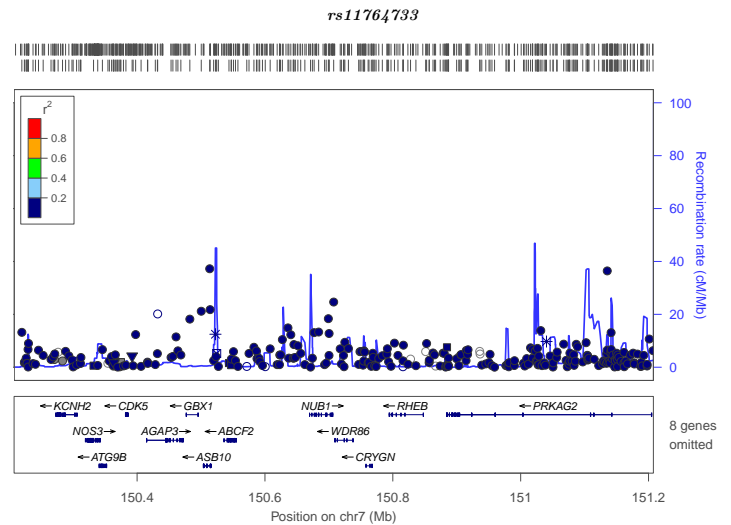


Figure A.38: ARC(male)

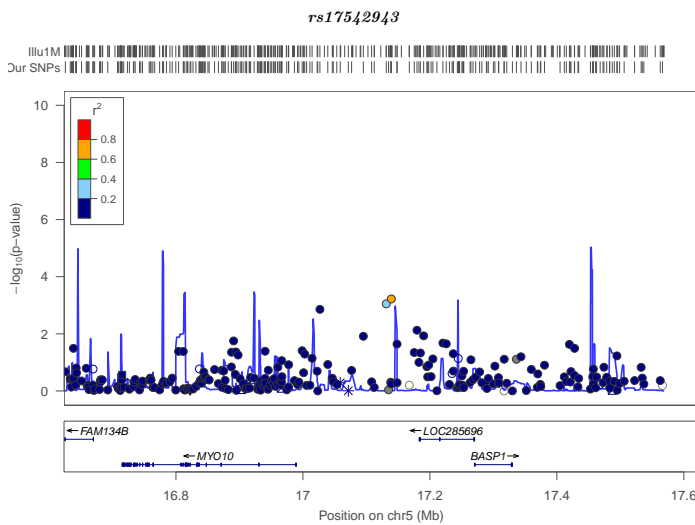


Figure A.39: ARC(male)

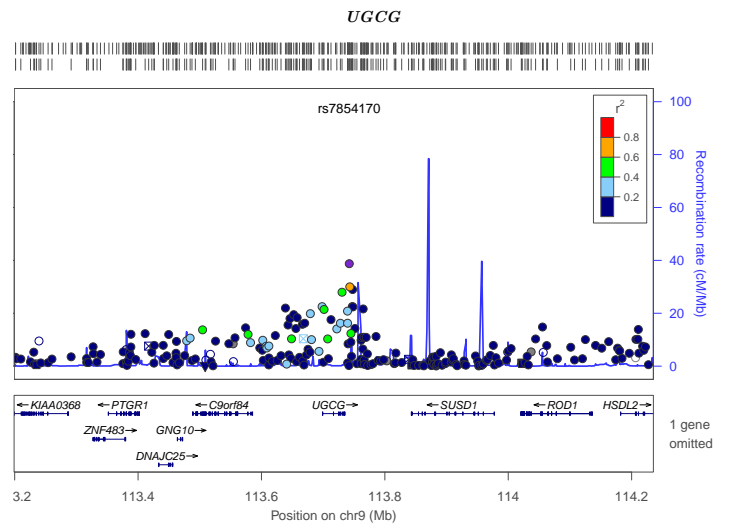


Figure A.40: ARC(male)

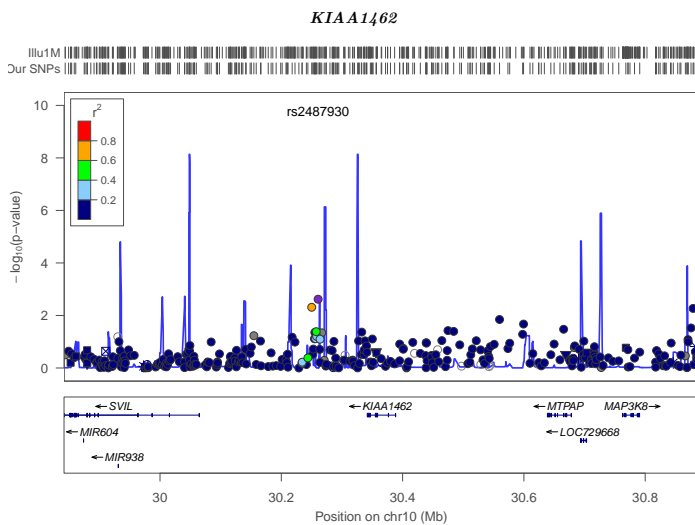


Figure A.41: ARC(combined)

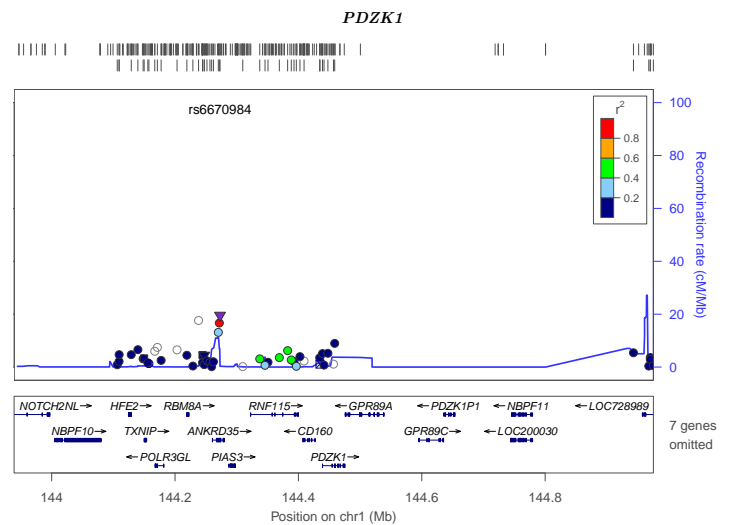


Figure A.42: ARC(combined)

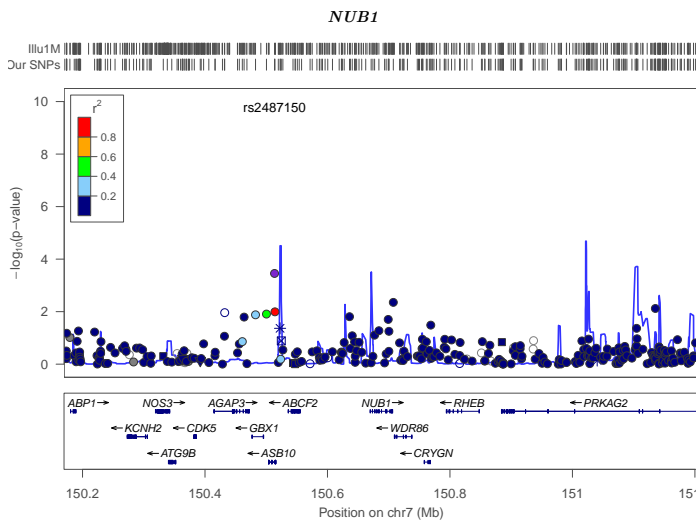


Figure A.43: ARC(combined)

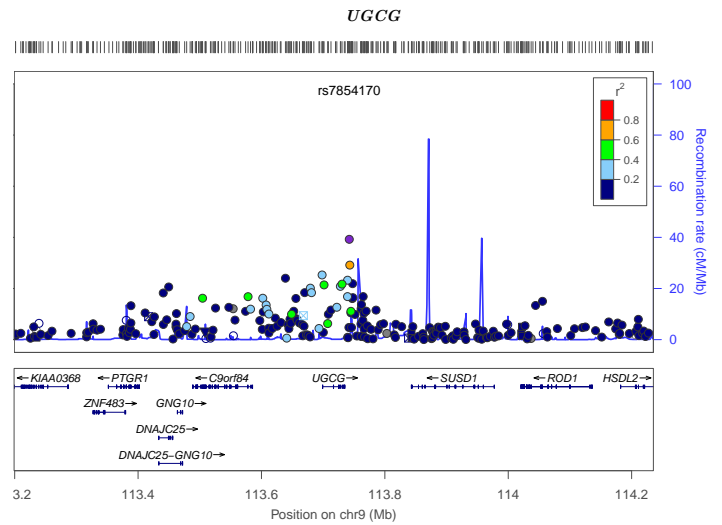


Figure A.44: ARC(combined)

### A.1.2 Phenotype: HS\_PCT

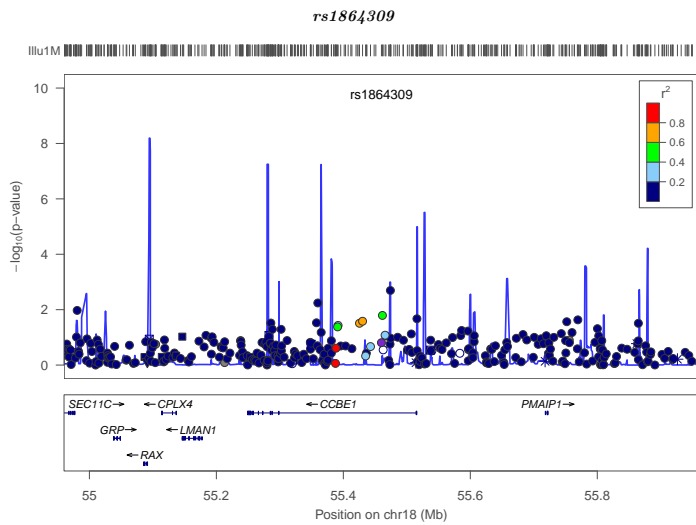


Figure A.45: HS\_PCT(combined)

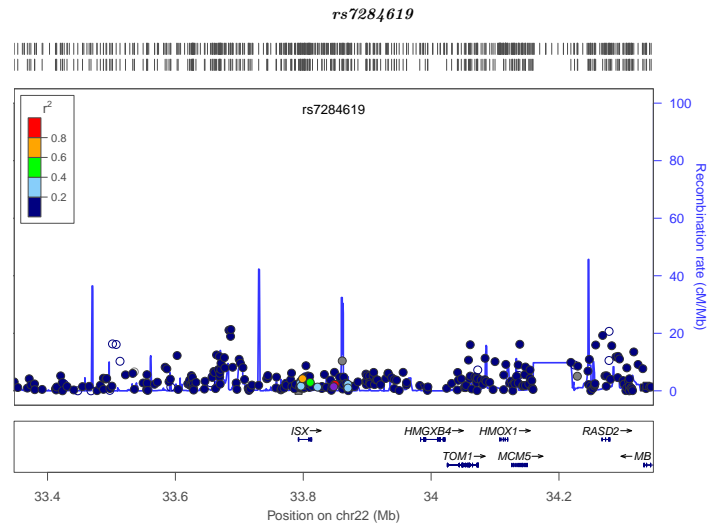


Figure A.46: HS\_PCT(combined)

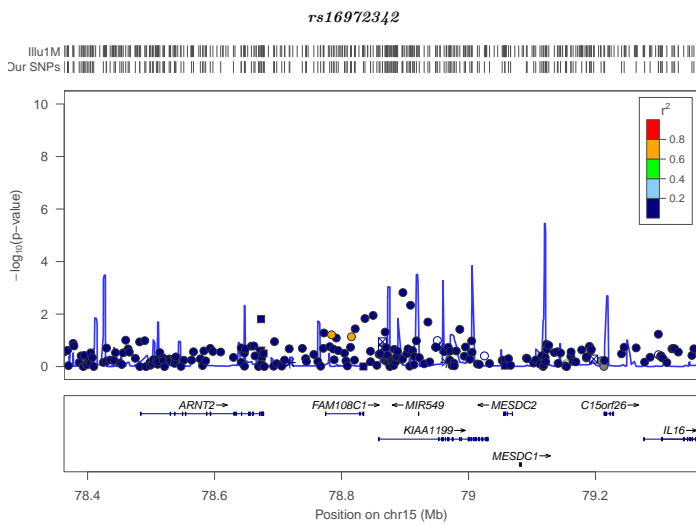


Figure A.47: HS\_PCT(combined)

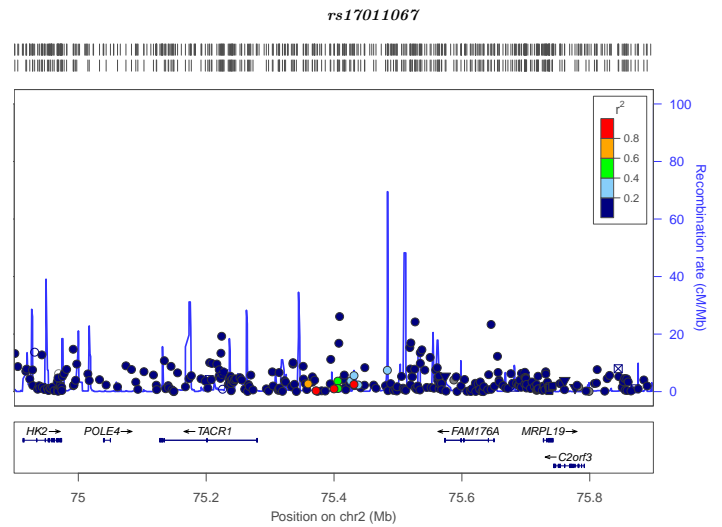


Figure A.48: HS\_PCT(combined)

## A.2 LOCUSZOOM PLOT OF OUR TOP HITS FOR FIVE PHENOTYPES

### A.2.1 Phenotype: ARC

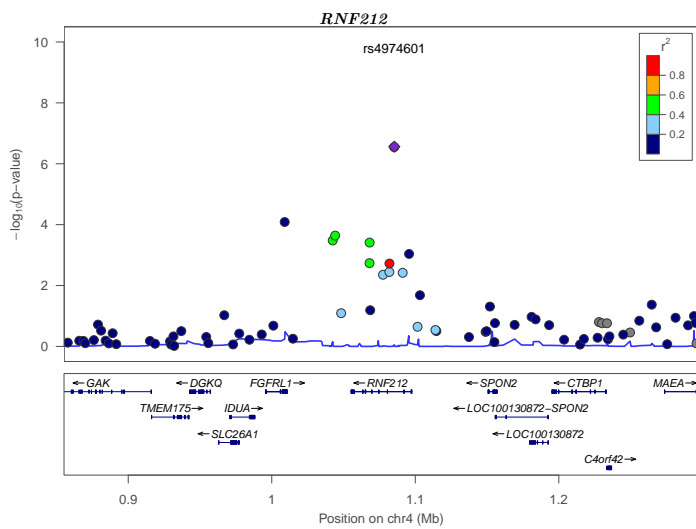


Figure A.49: ARC(combined)

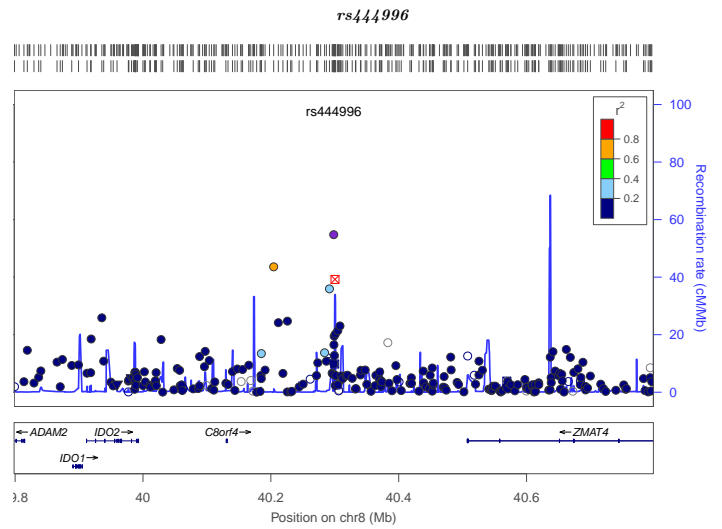


Figure A.50: ARC(combined)

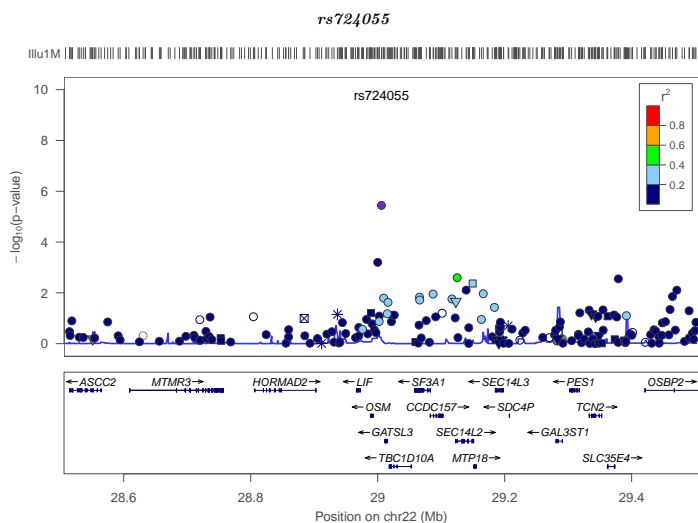


Figure A.51: ARC(combined)

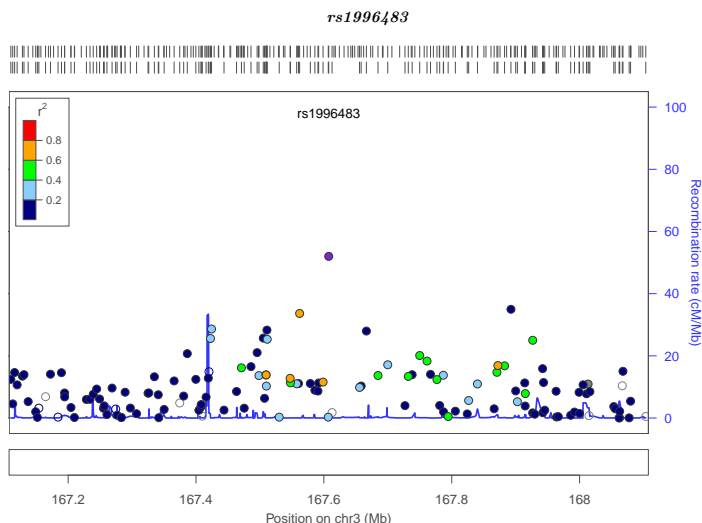


Figure A.52: ARC(combined)

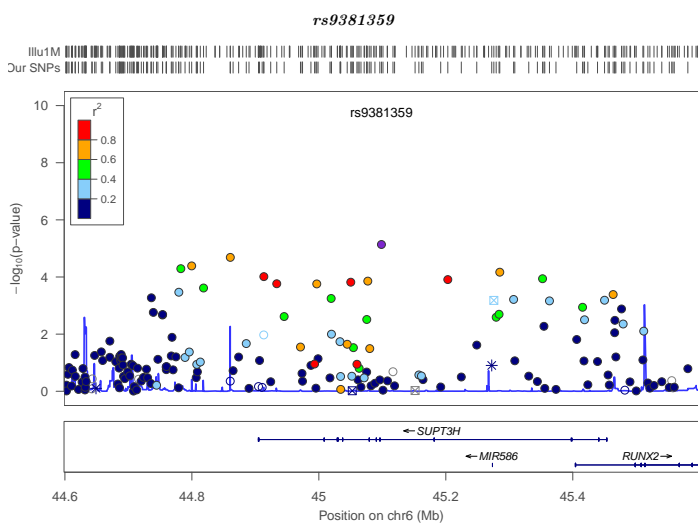


Figure A.53: ARC(combined)

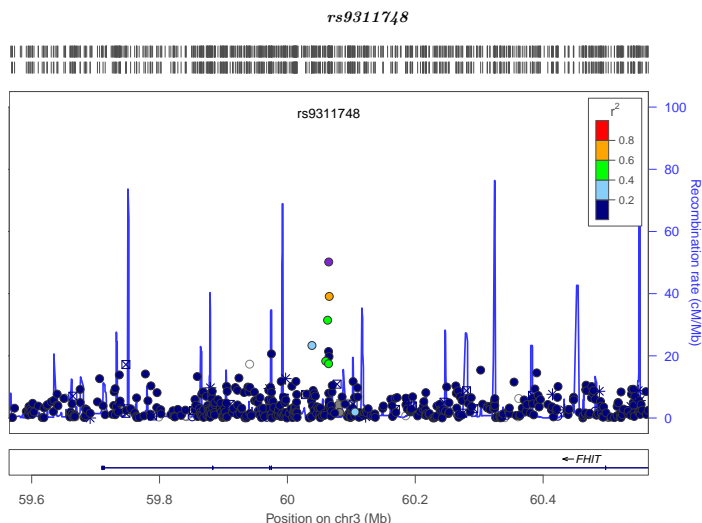


Figure A.54: ARC(combined)

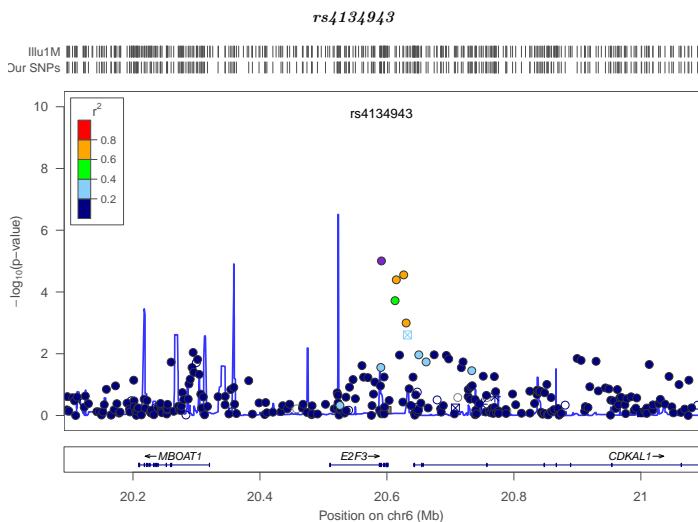


Figure A.55: ARC(combined)

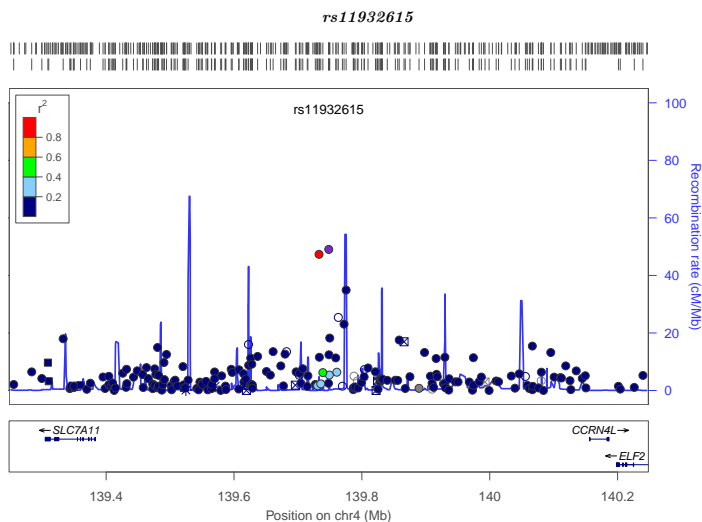


Figure A.56: ARC(combined)

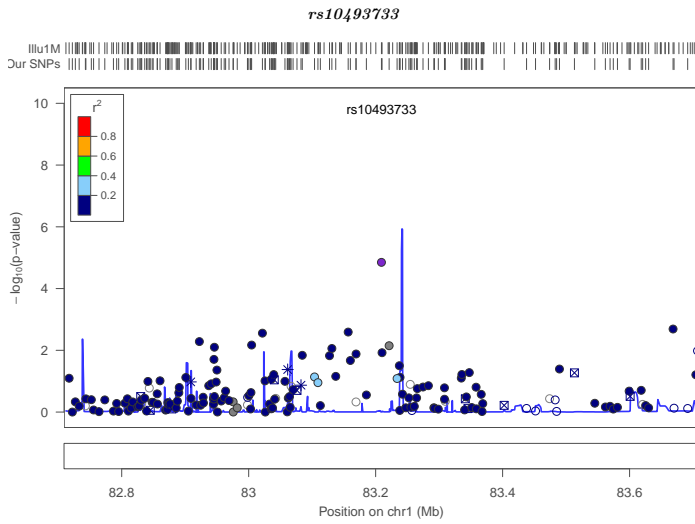


Figure A.57: ARC(combined)

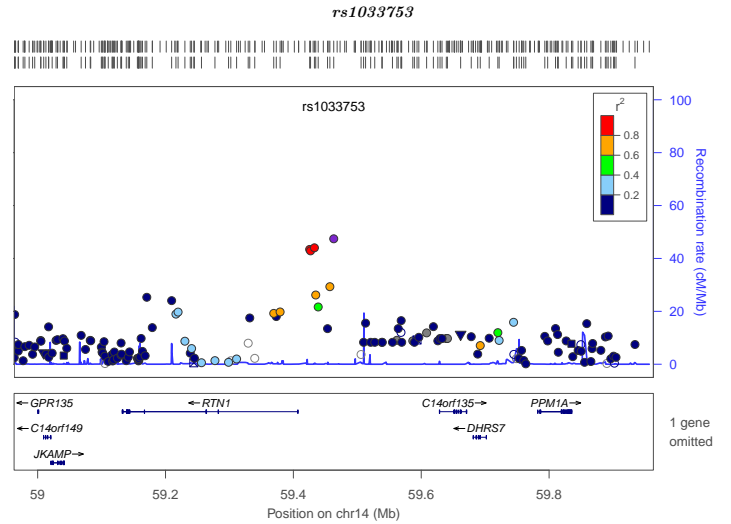


Figure A.58: ARC(combined)

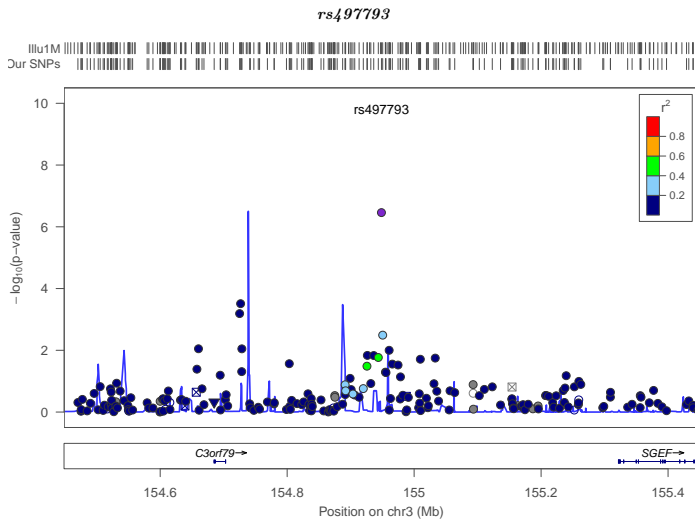


Figure A.59: ARC(female)

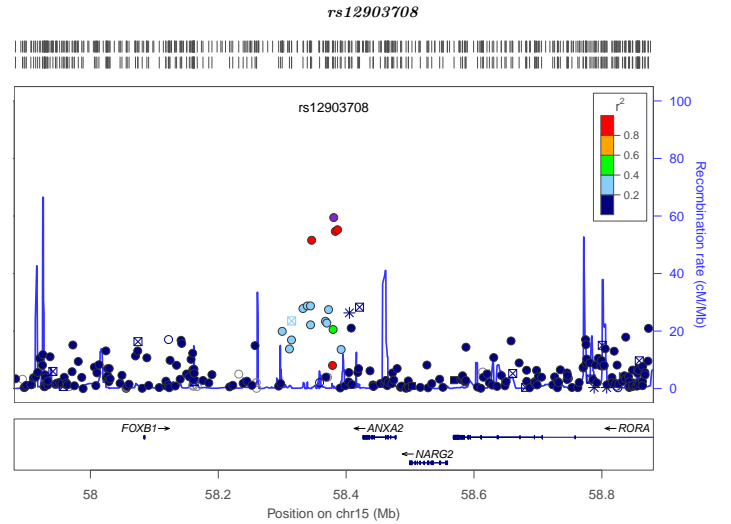


Figure A.60: ARC(female)

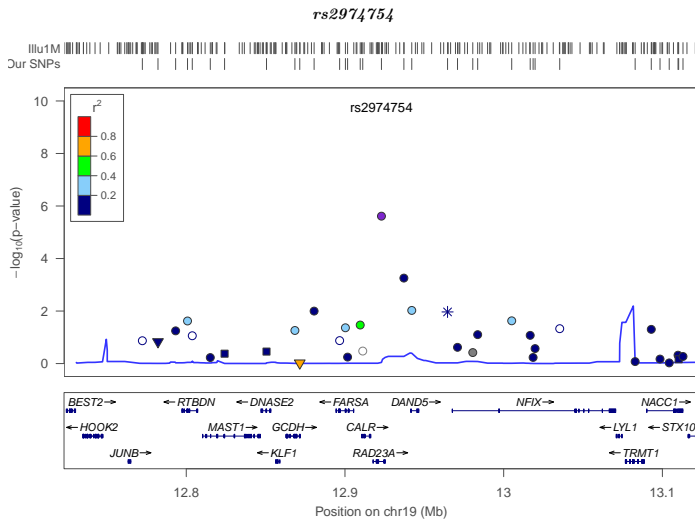


Figure A.61: ARC(female)

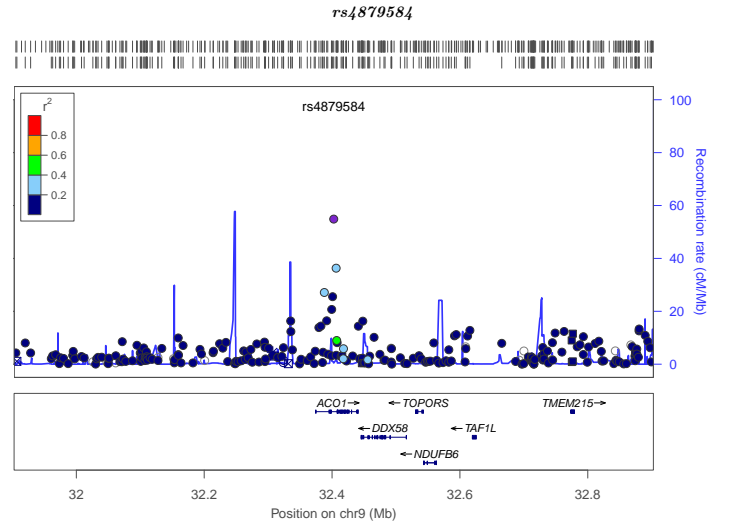


Figure A.62: ARC(female)



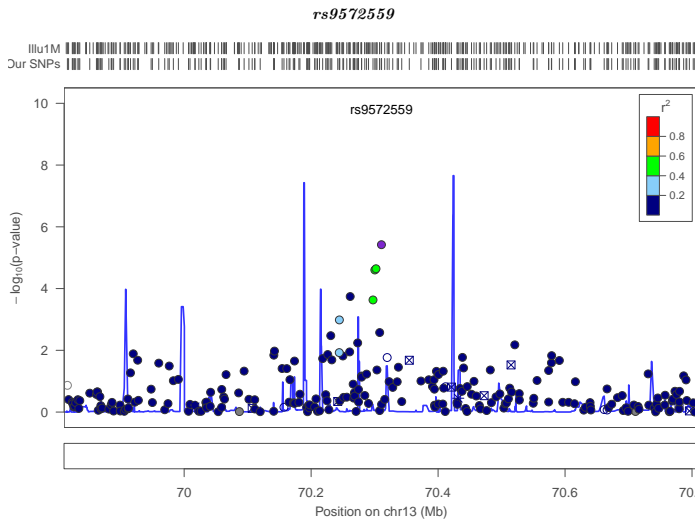


Figure A.63: ARC(combined)

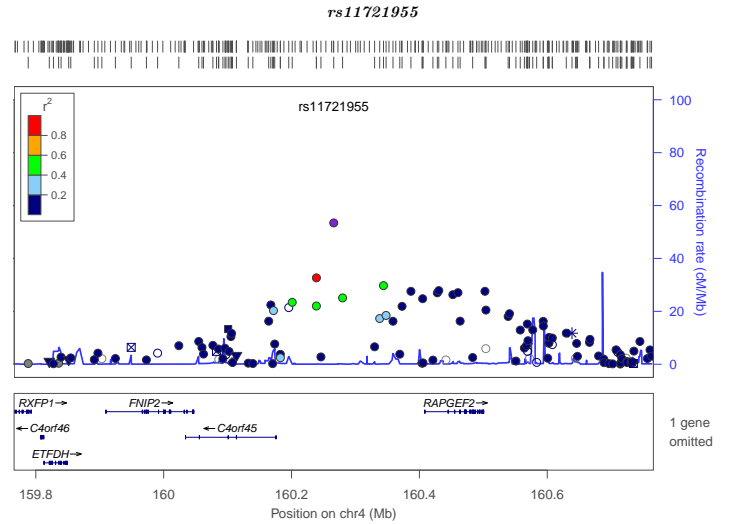


Figure A.64: ARC(combined)

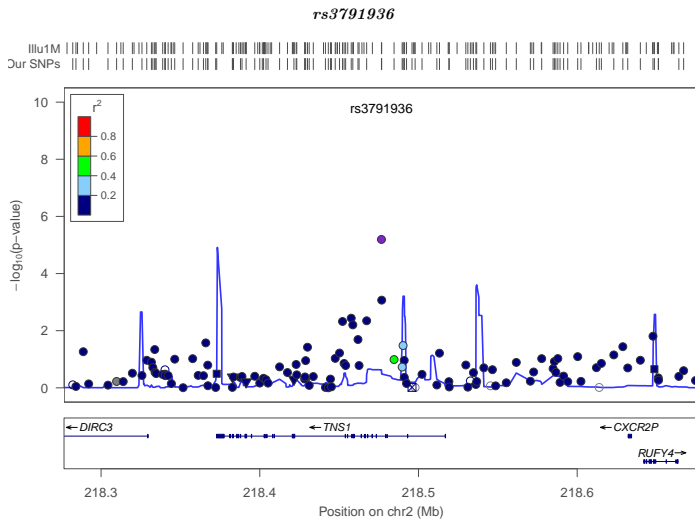


Figure A.65: ARC(female)

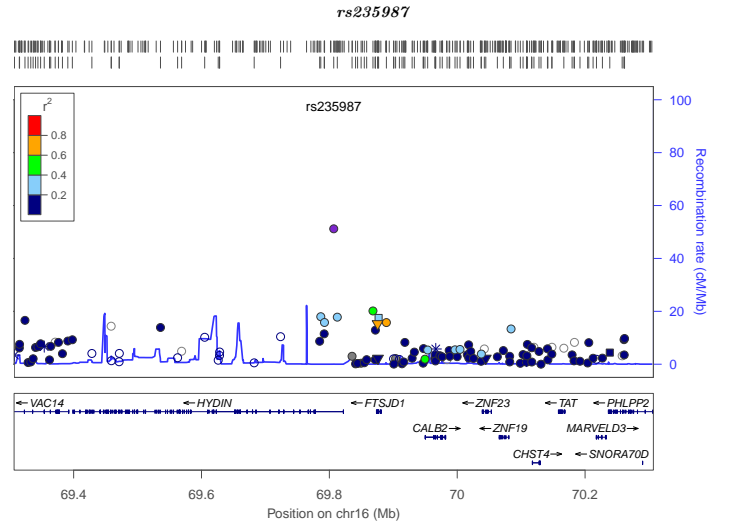


Figure A.66: ARC(female)

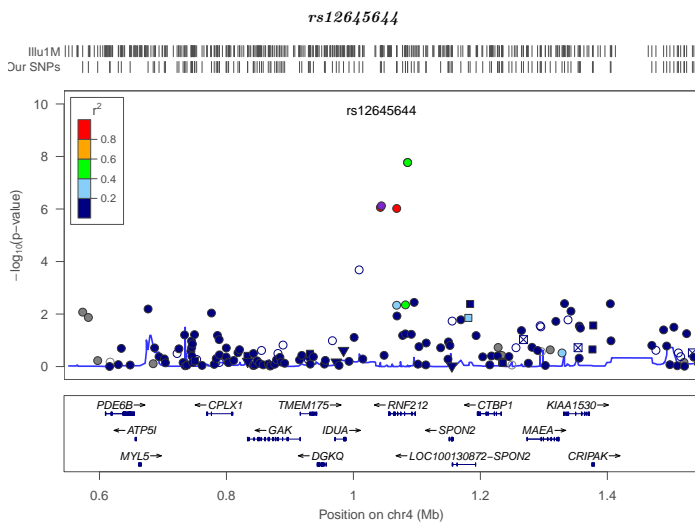


Figure A.67: ARC(male)

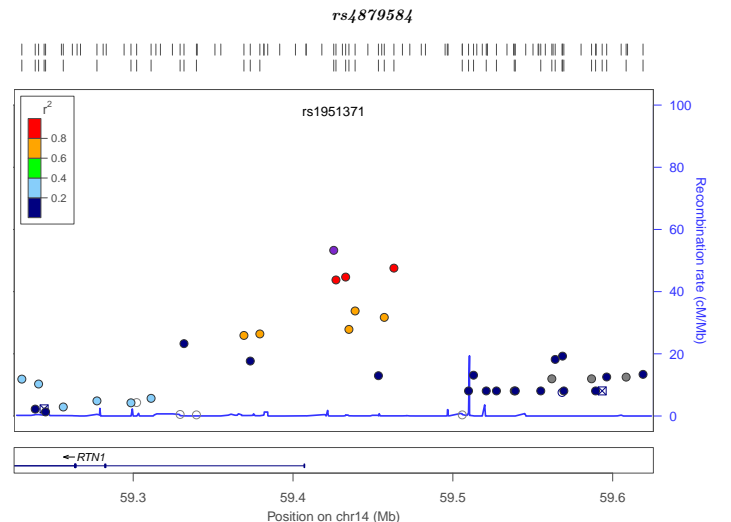


Figure A.68: ARC(male)

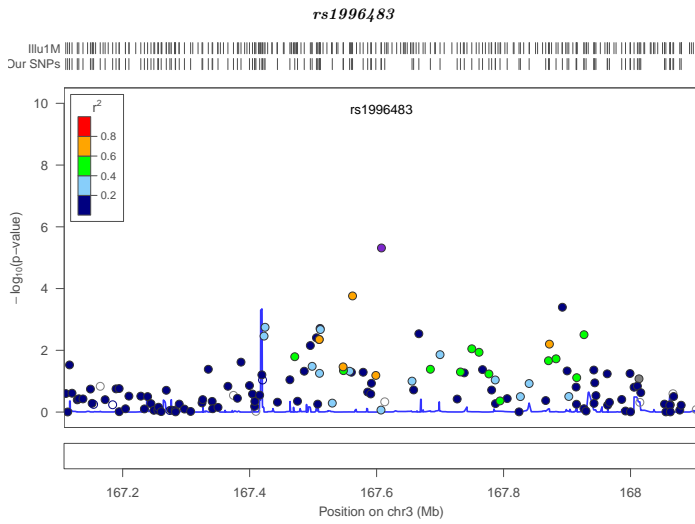


Figure A.69: ARC(male)

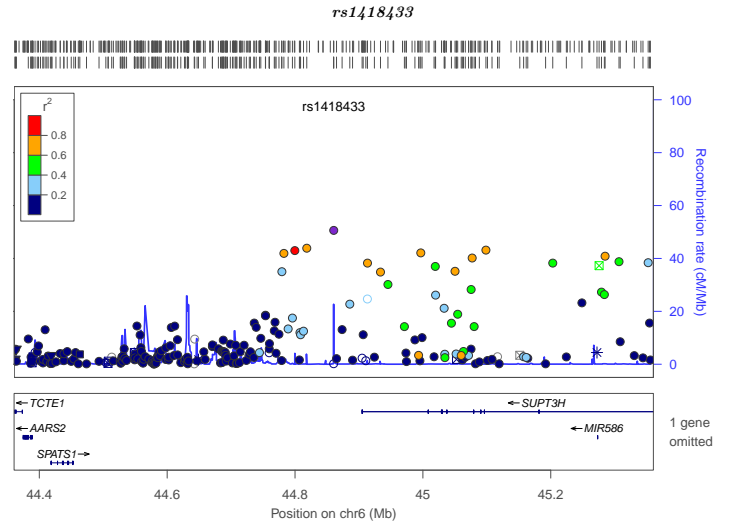


Figure A.70: ARC(male)

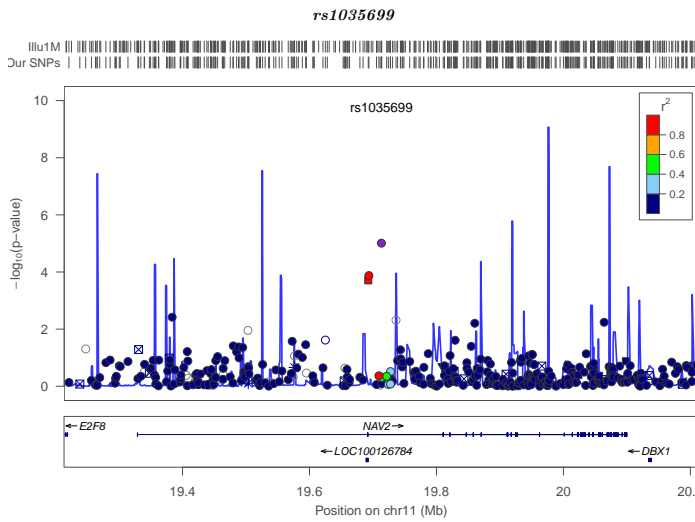


Figure A.71: ARC(male)

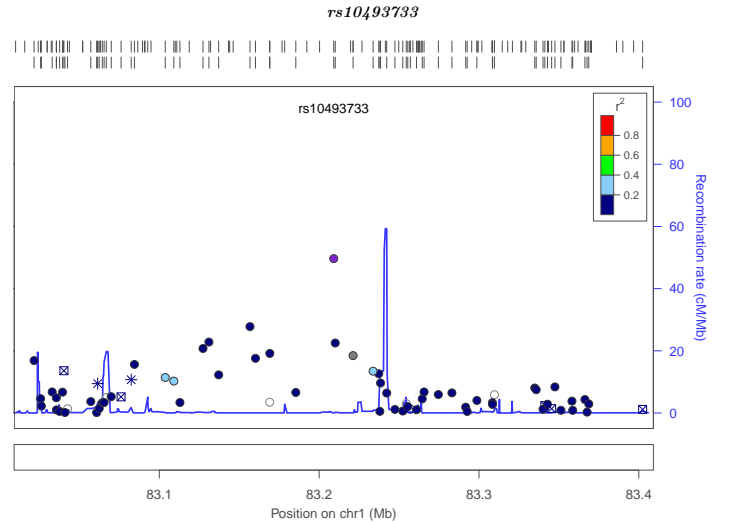


Figure A.72: ARC(male)

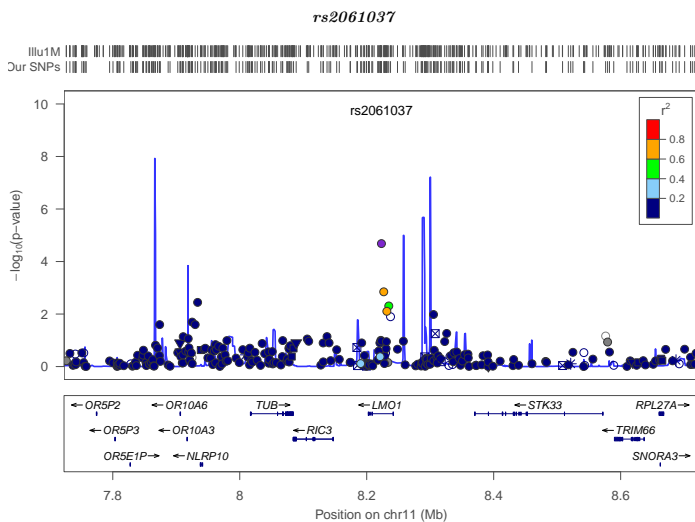


Figure A.73: ARC(male)

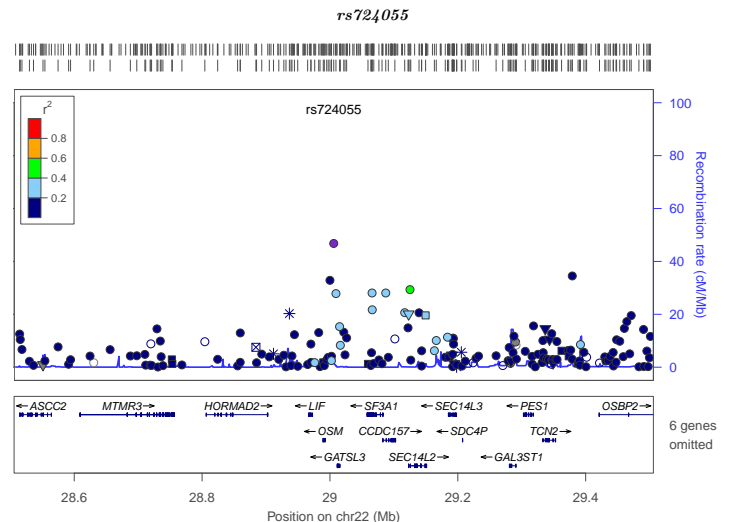


Figure A.74: ARC(male)

## A.2.2 Phenotype: HS\_PCT

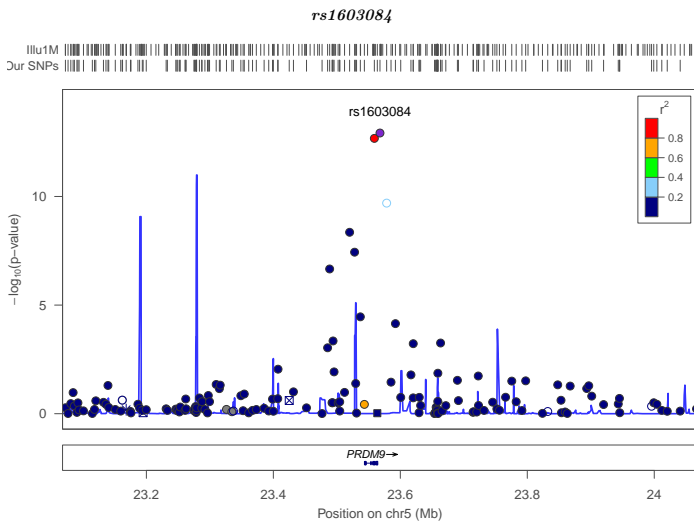


Figure A.75: HS\_PCT(combined)

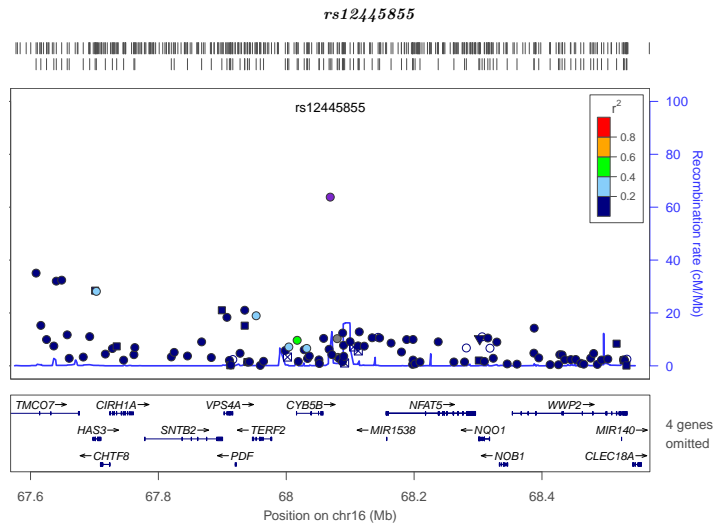


Figure A.76: HS\_PCT(combined)

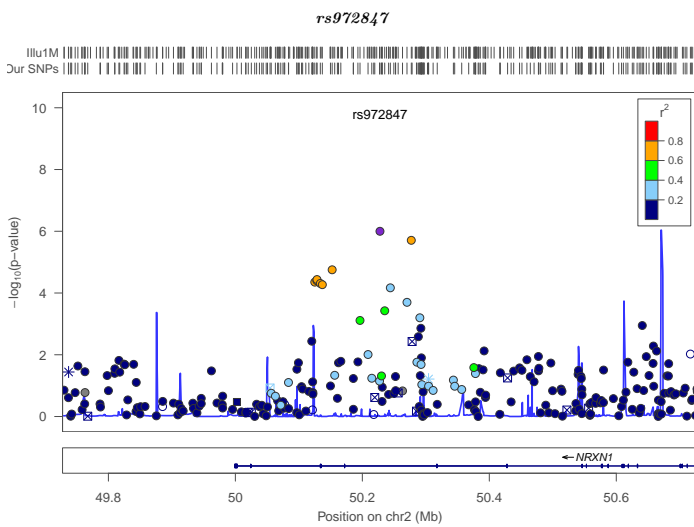


Figure A.77: HS\_PCT(combined)

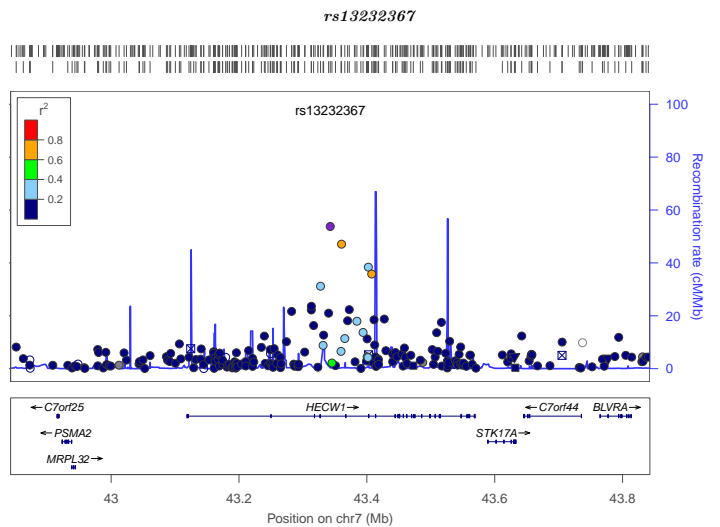


Figure A.78: HS\_PCT(combined)

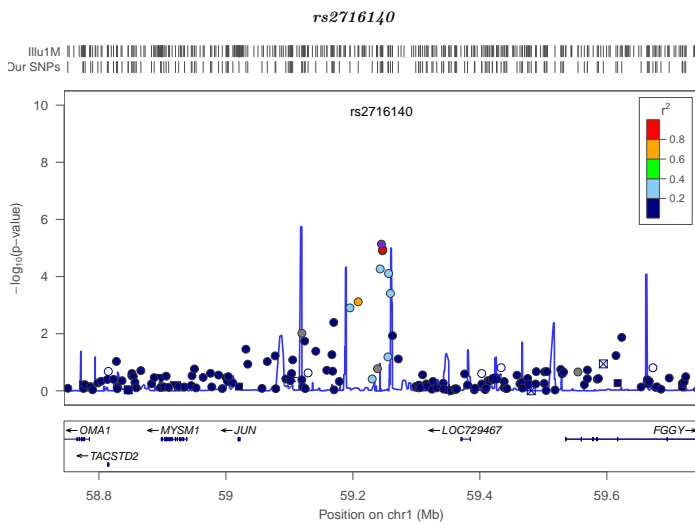


Figure A.79: HS\_PCT(combined)

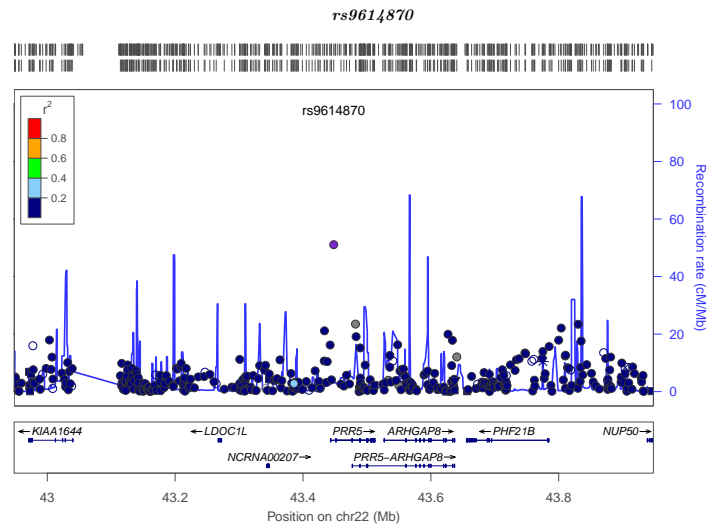


Figure A.80: HS\_PCT(combined)

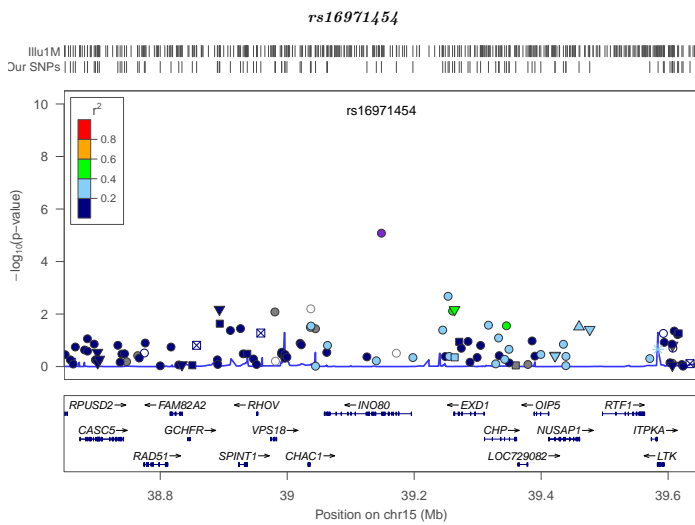


Figure A.81: HS\_PCT(combined)

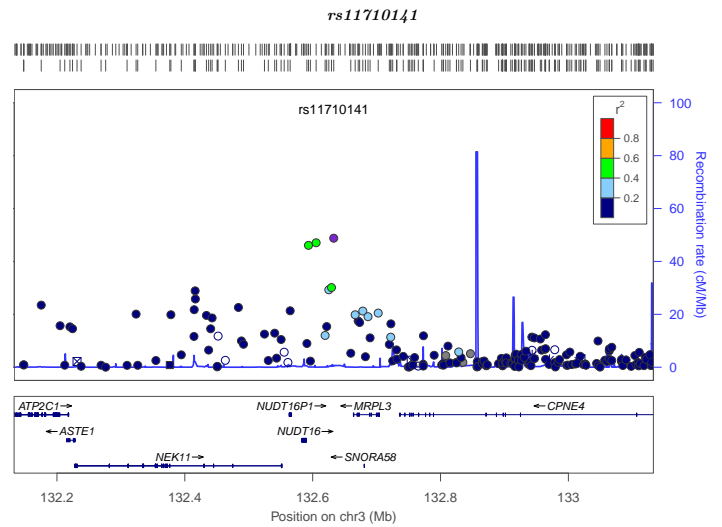


Figure A.82: HS\_PCT(combined)

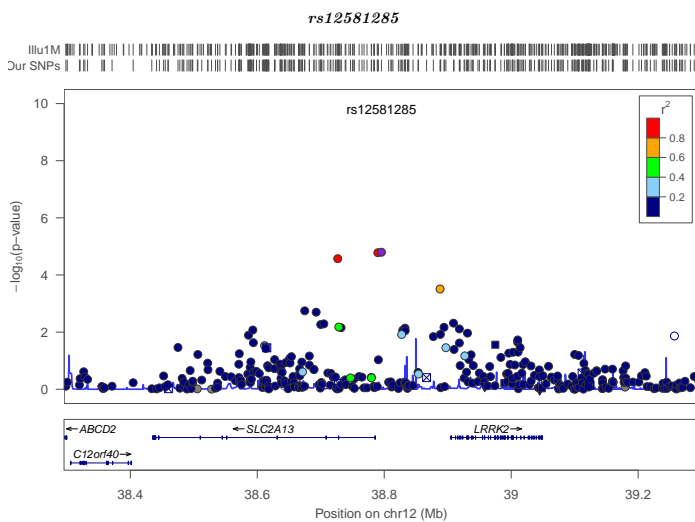


Figure A.83: HS\_PCT(combined)

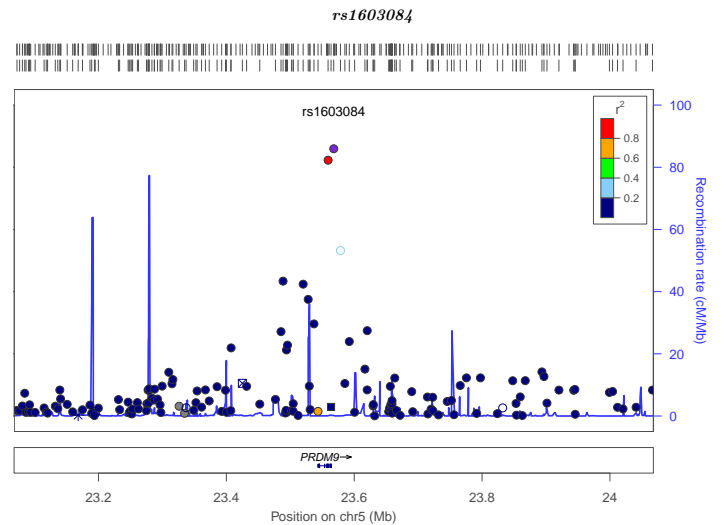


Figure A.84: HS\_PCT(female)

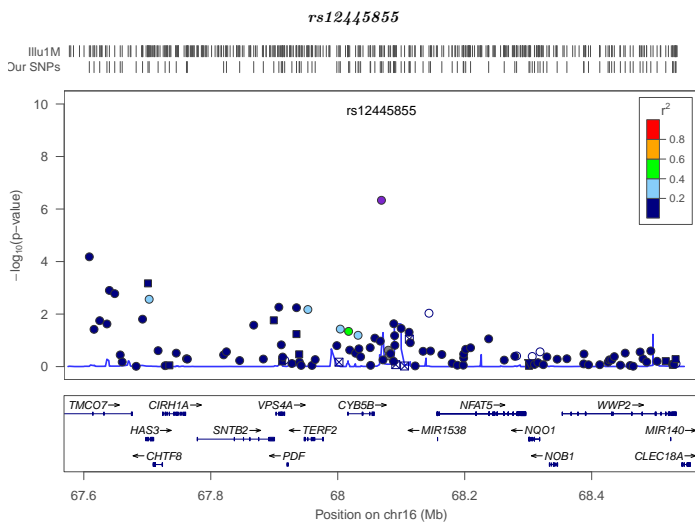


Figure A.85: HS\_PCT(female)

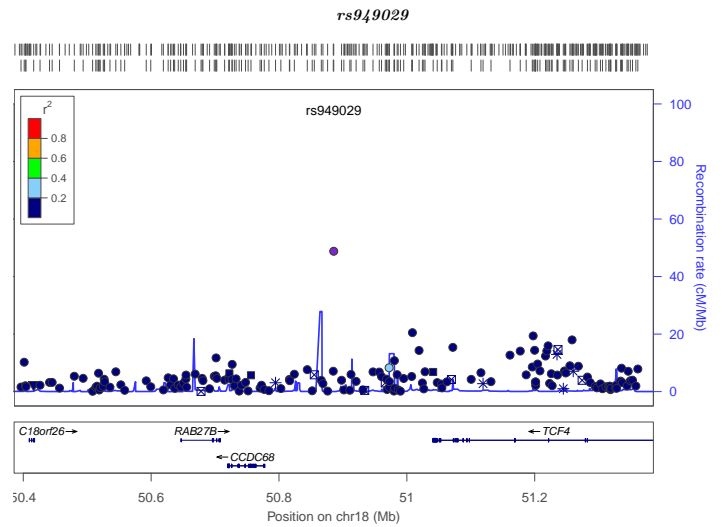


Figure A.86: HS\_PCT(female)

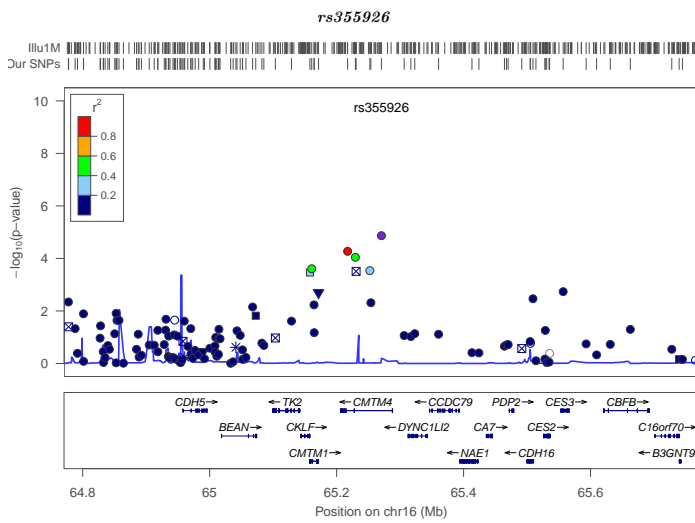


Figure A.87: HS\_PCT(female)

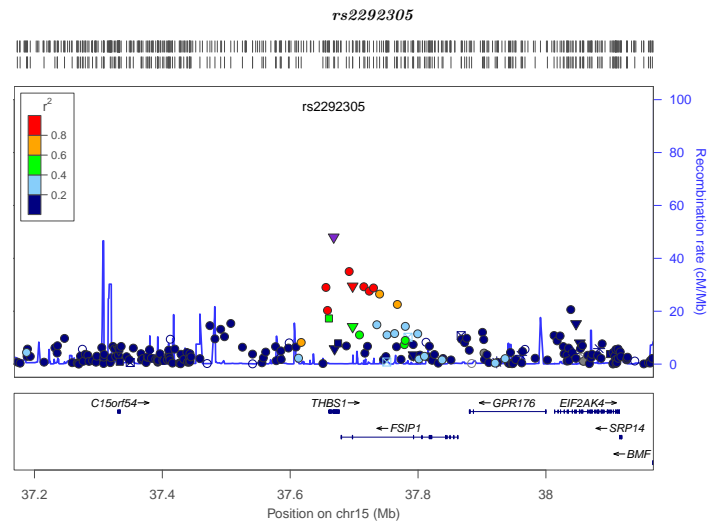


Figure A.88: HS\_PCT(female)

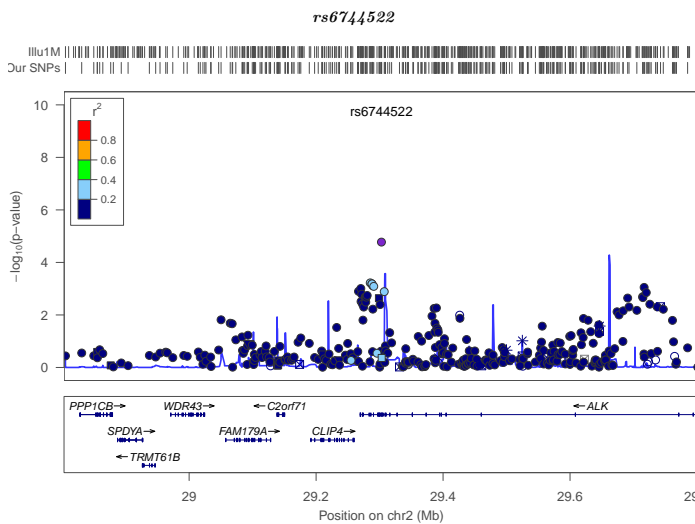


Figure A.89: HS\_PCT(female)

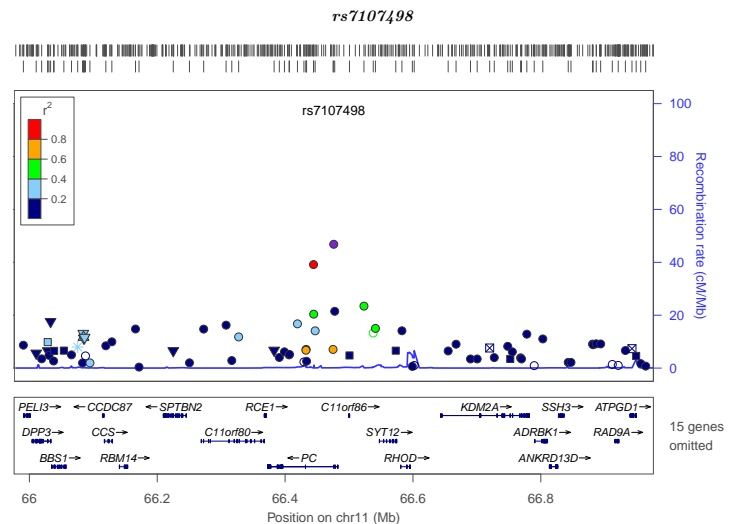


Figure A.90: HS\_PCT(female)

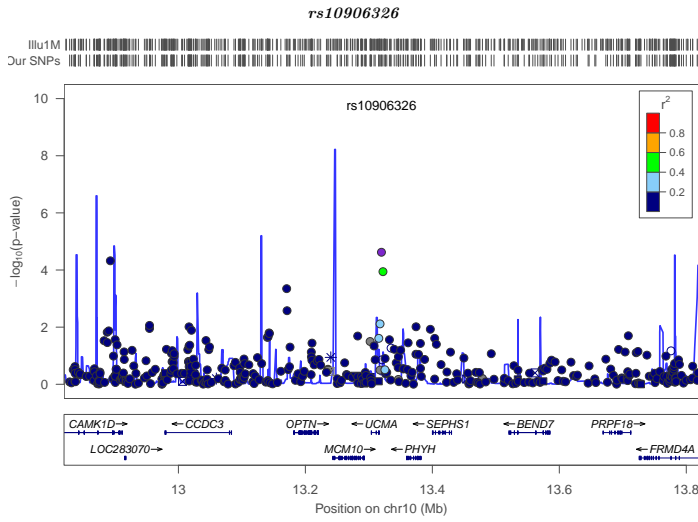


Figure A.91: HS\_PCT(female)

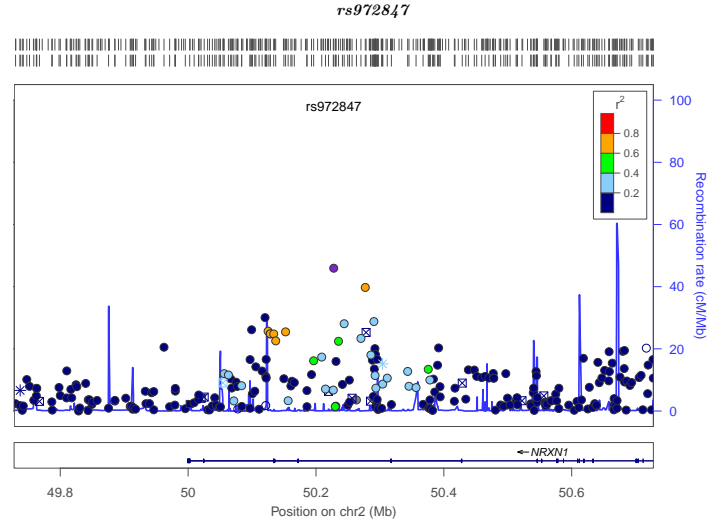


Figure A.92: HS\_PCT(female)

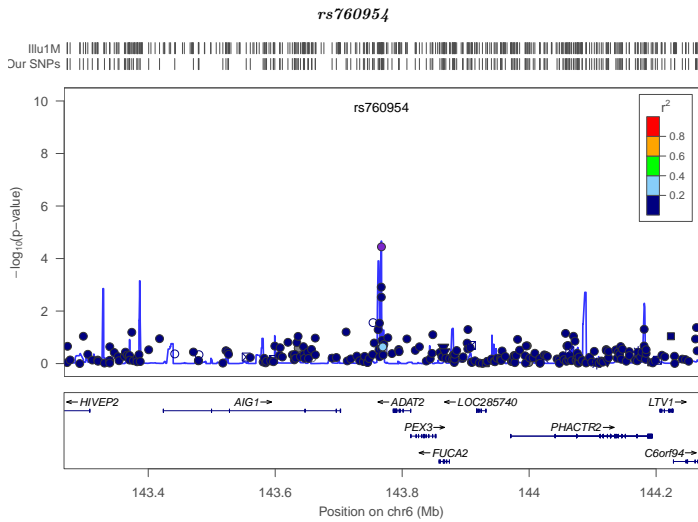


Figure A.93: HS\_PCT(female)

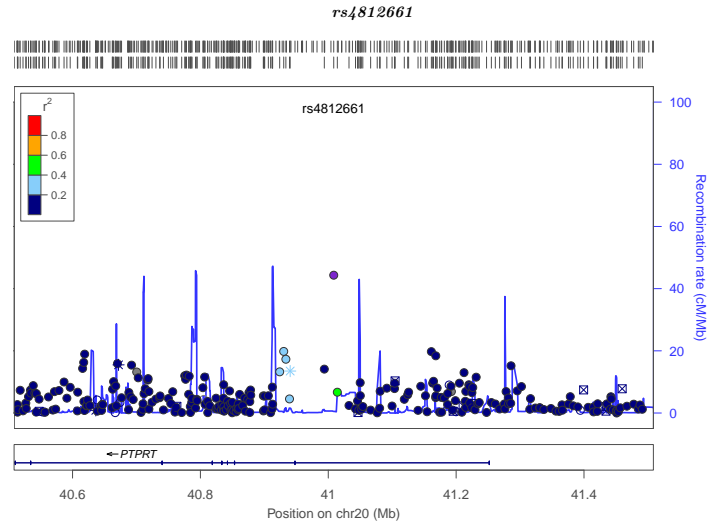


Figure A.94: HS\_PCT(female)

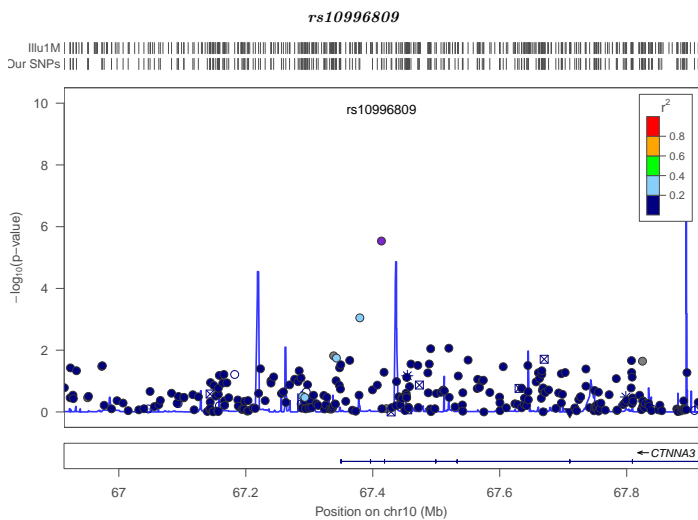


Figure A.95: HS\_PCT(male)

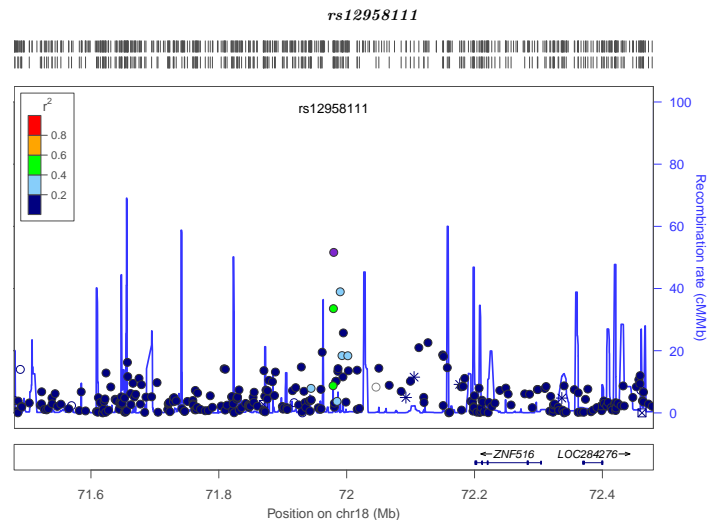


Figure A.96: HS\_PCT(male)

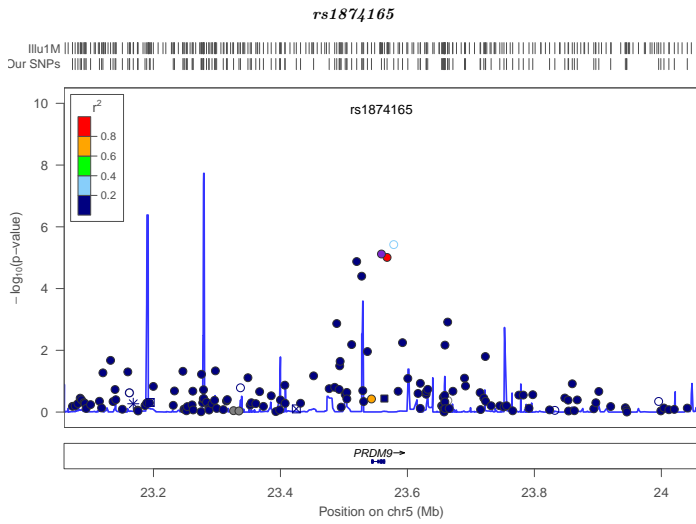


Figure A.97: HS\_PCT(male)

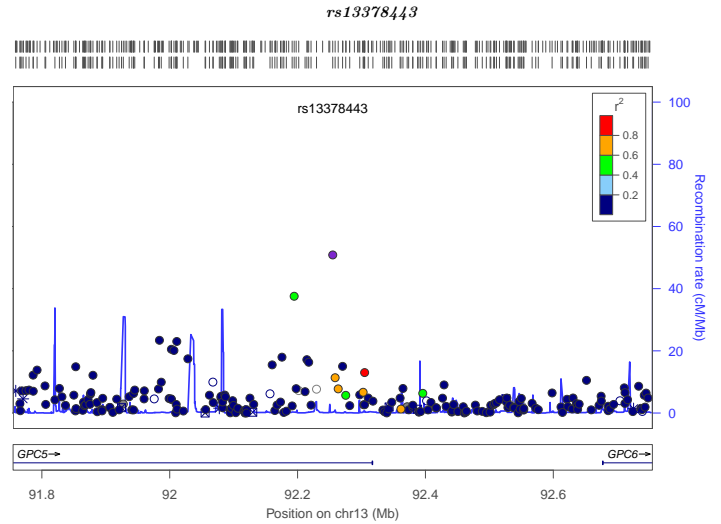


Figure A.98: HS\_PCT(male)

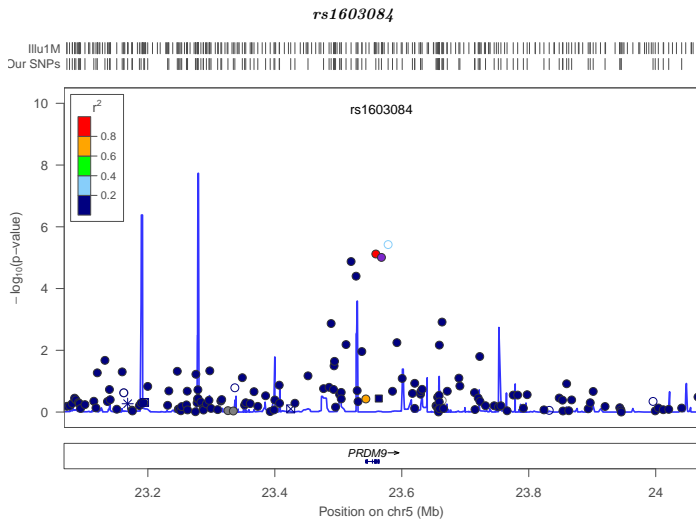


Figure A.99: HS\_PCT(male)

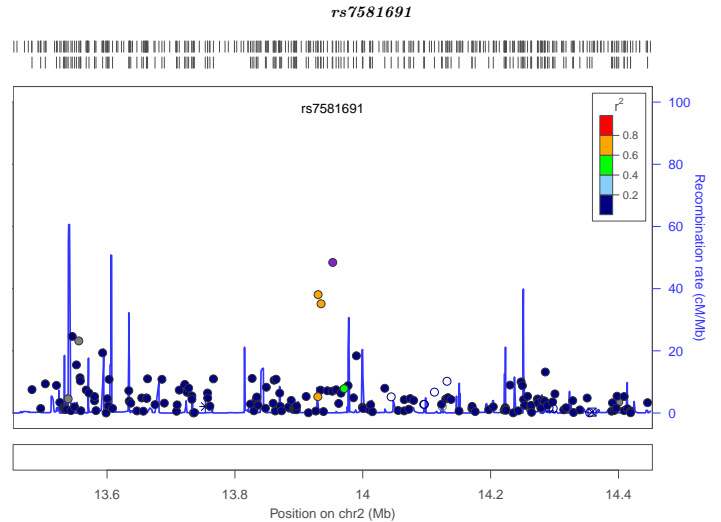


Figure A.100: HS\_PCT(male)

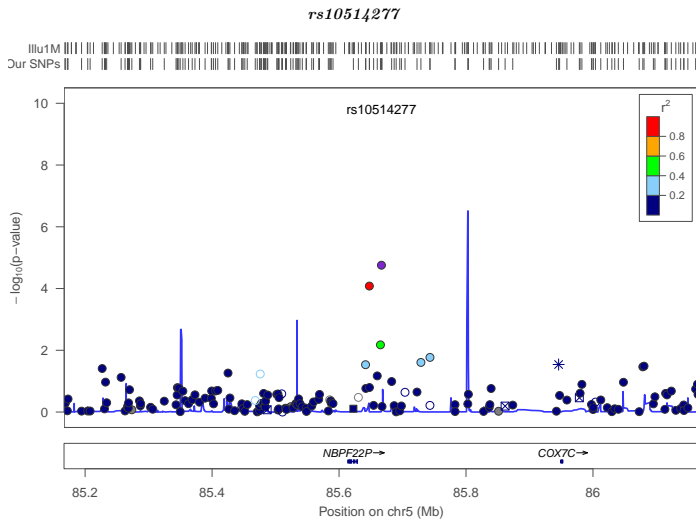


Figure A.101: HS\_PCT(male)

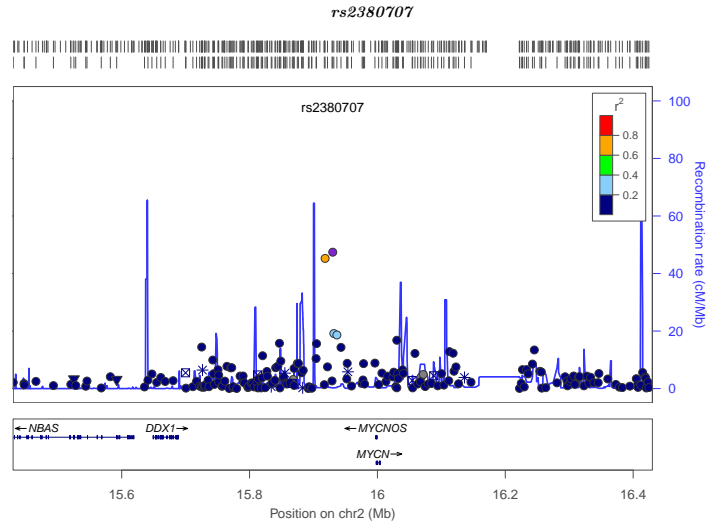


Figure A.102: HS\_PCT(male)

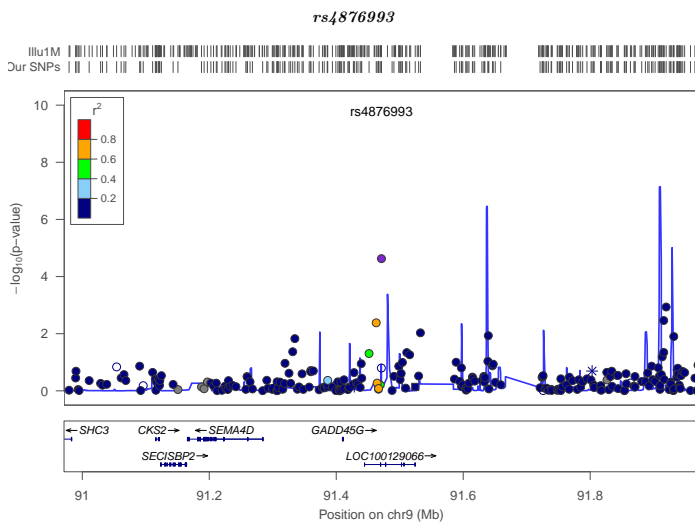


Figure A.103: HS\_PCT(male)

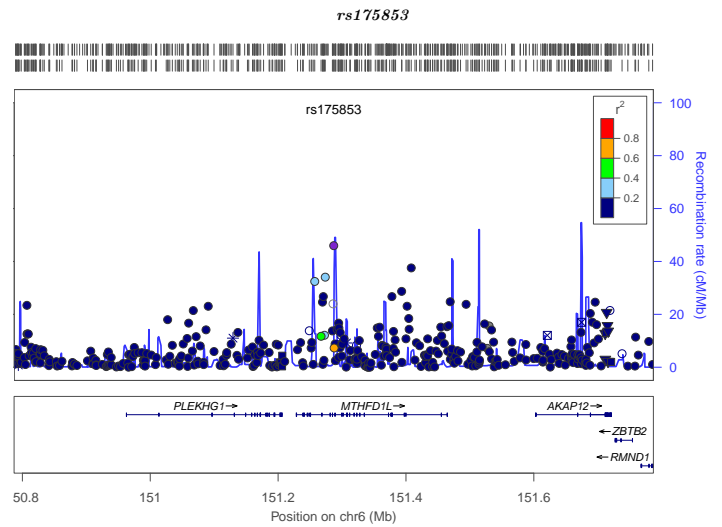


Figure A.104: HS\_PCT(male)

### A.2.3 Phenotype: HS\_CNT

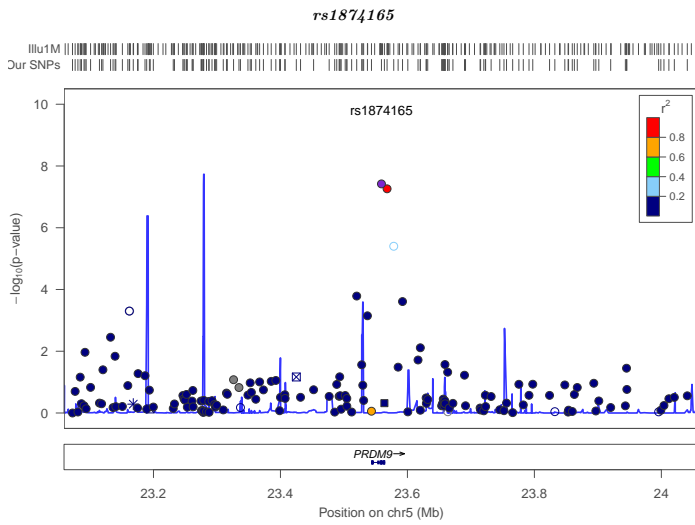


Figure A.105: HS\_CNT(combined)

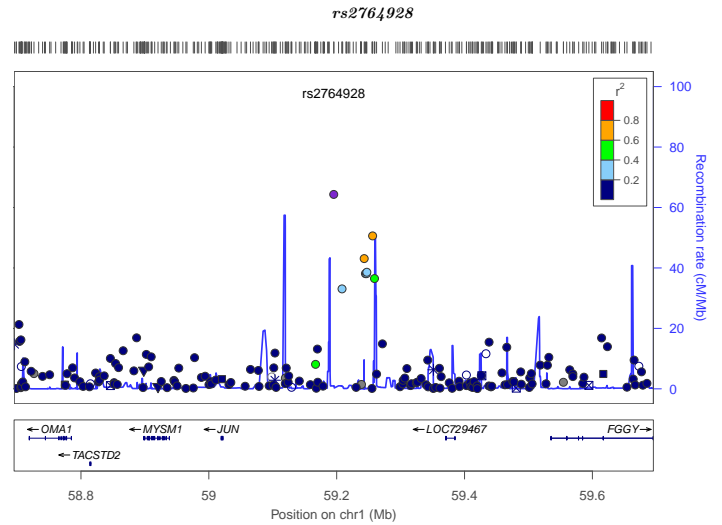


Figure A.106: HS\_CNT(combined)



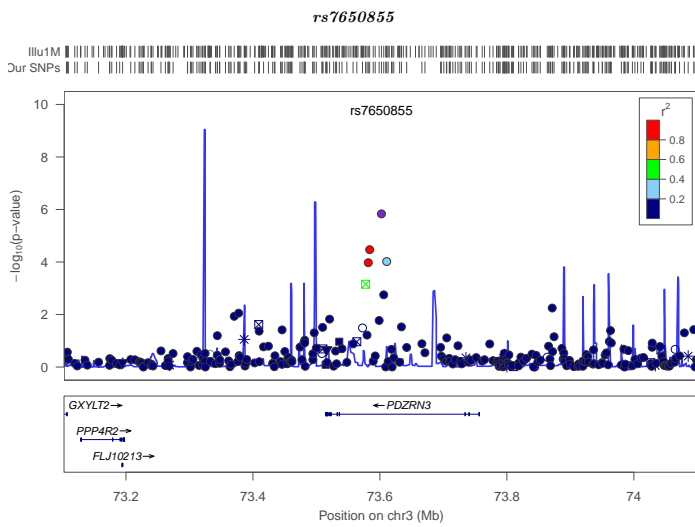


Figure A.107: HS\_CNT(combined)

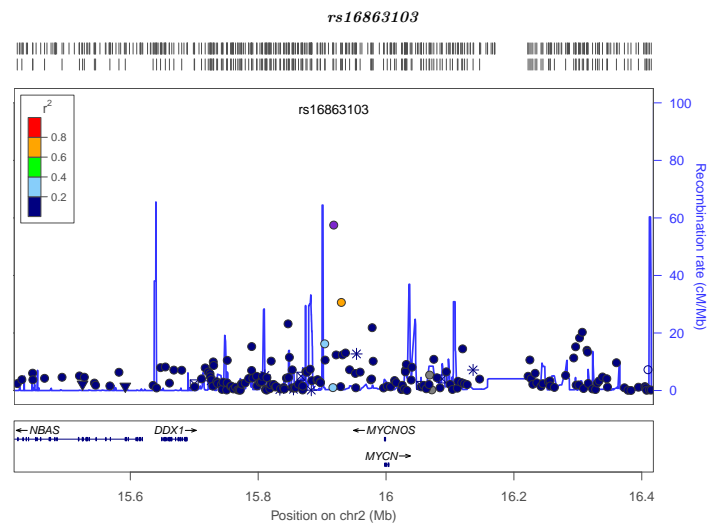


Figure A.108: HS\_CNT(combined)

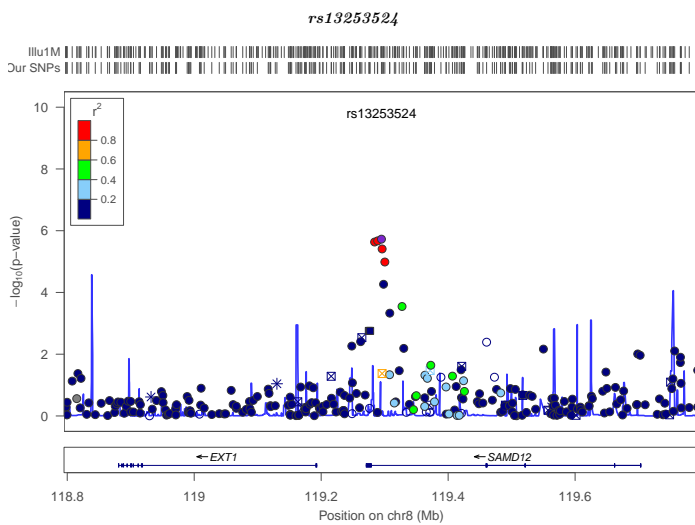


Figure A.109: HS\_CNT(combined)

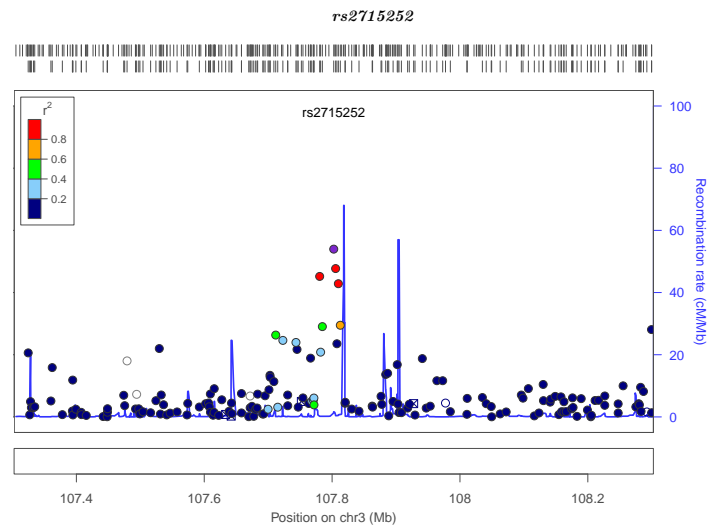


Figure A.110: HS\_CNT(combined)

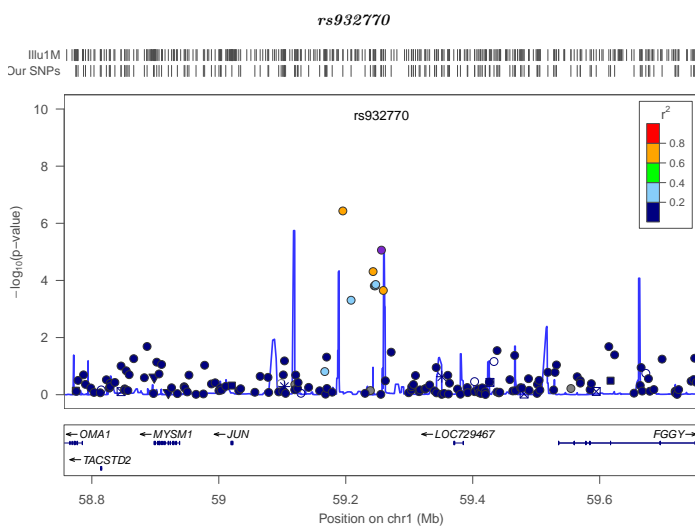


Figure A.111: HS\_CNT(combined)

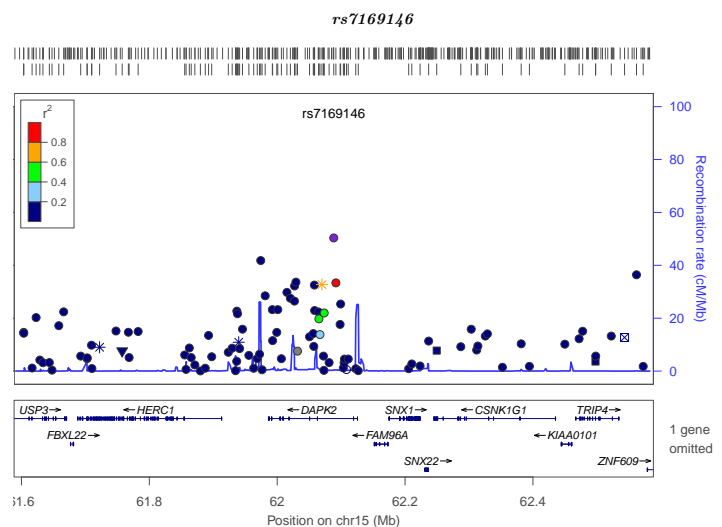


Figure A.112: HS\_CNT(combined)

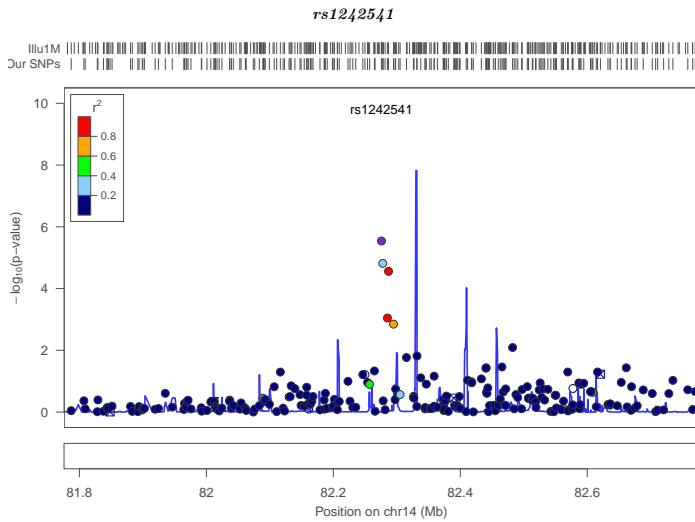


Figure A.113: HS\_CNT(female)

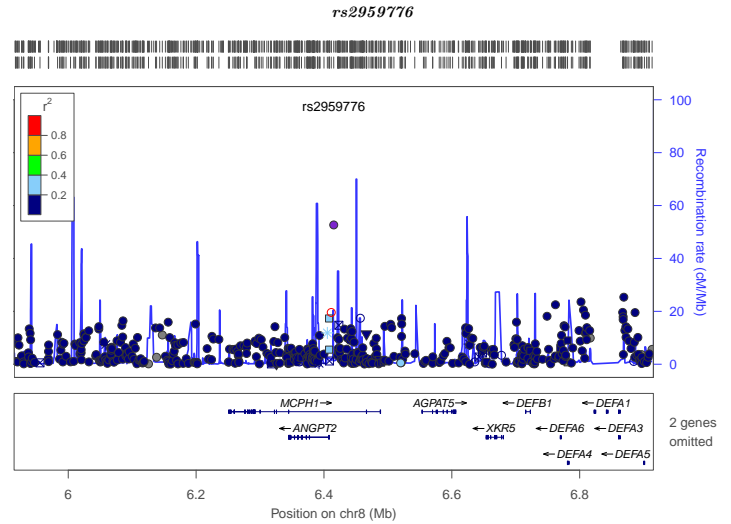


Figure A.114: HS\_CNT(female)

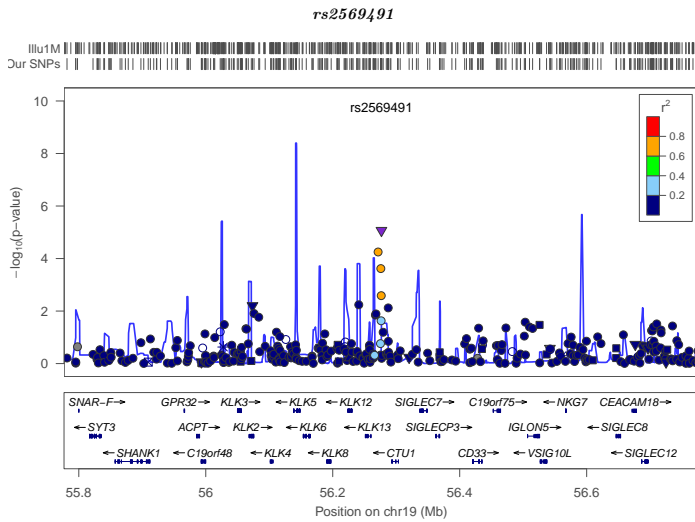


Figure A.115: HS\_CNT(female)

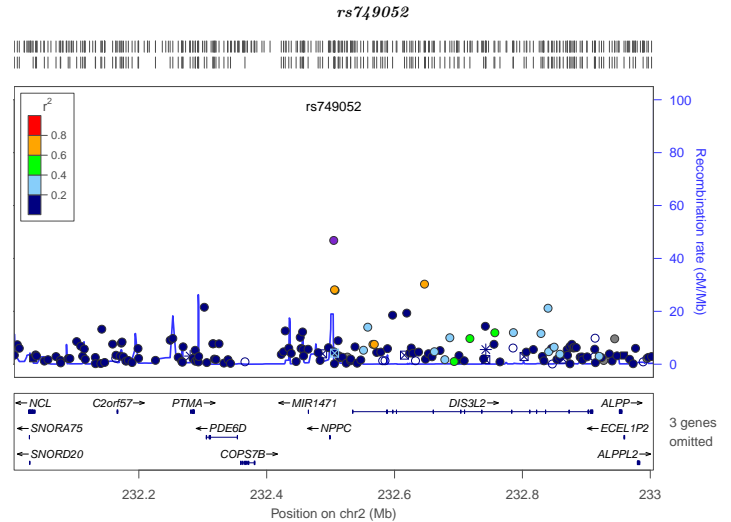


Figure A.116: HS\_CNT(female)

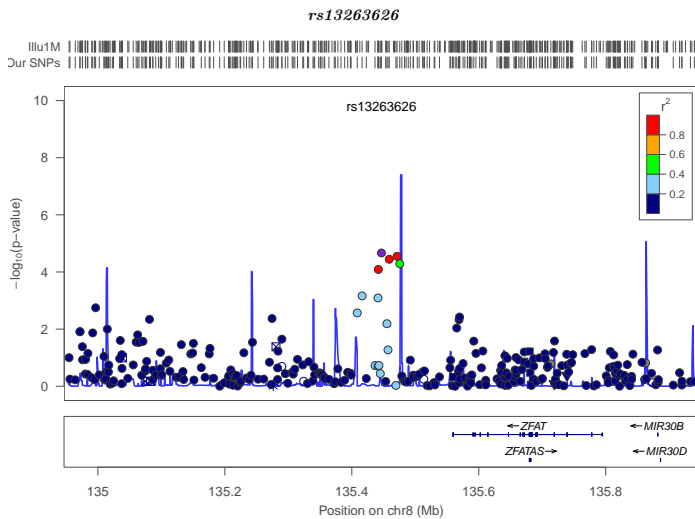


Figure A.117: HS\_CNT(female)

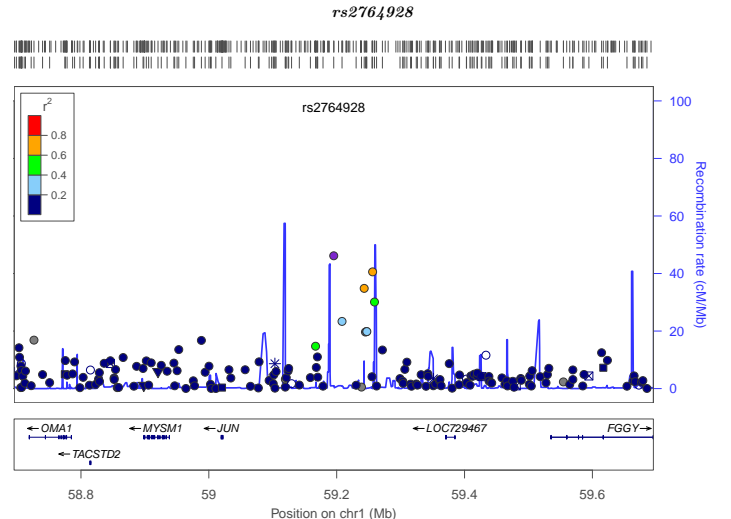


Figure A.118: HS\_CNT(female)

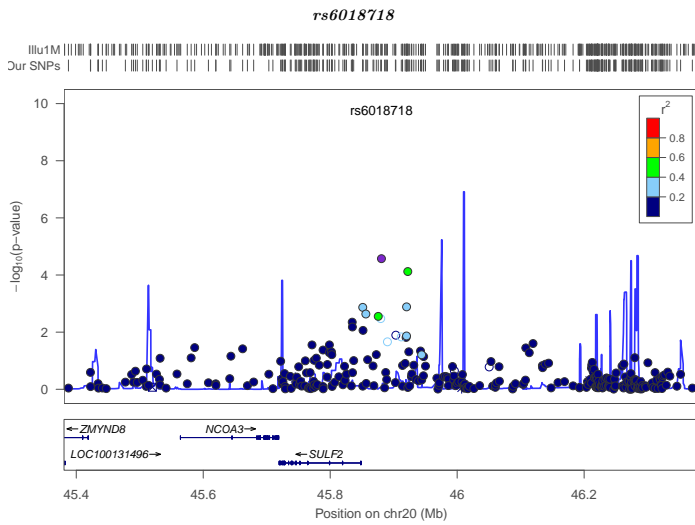


Figure A.119: HS\_CNT(female)

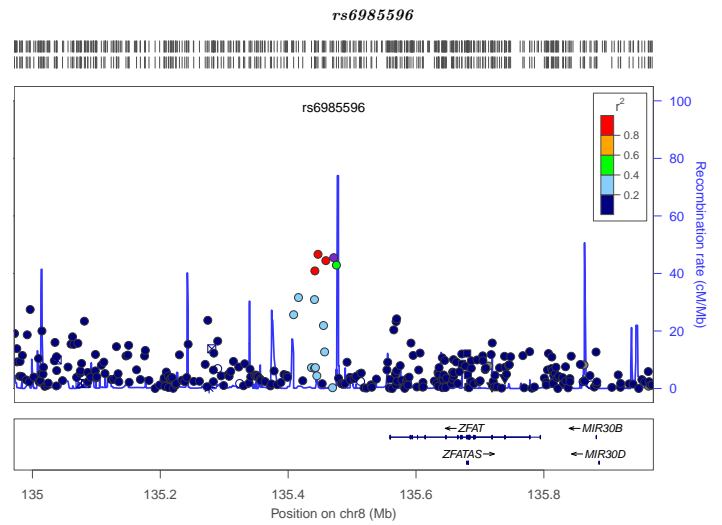


Figure A.120: HS\_CNT(female)

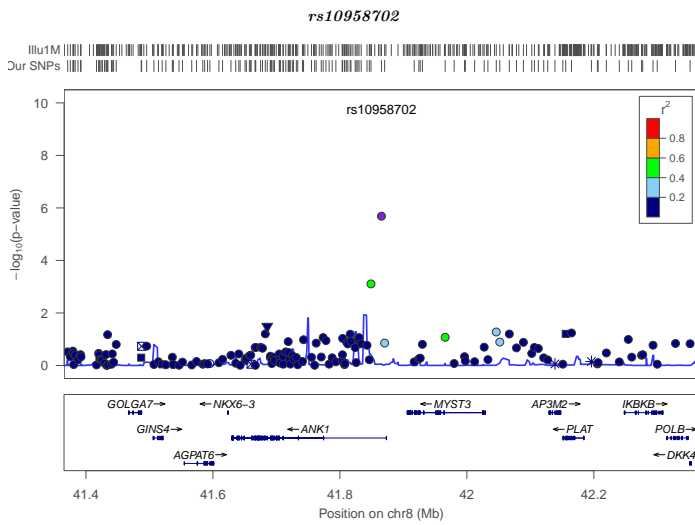


Figure A.121: HS\_CNT(male)

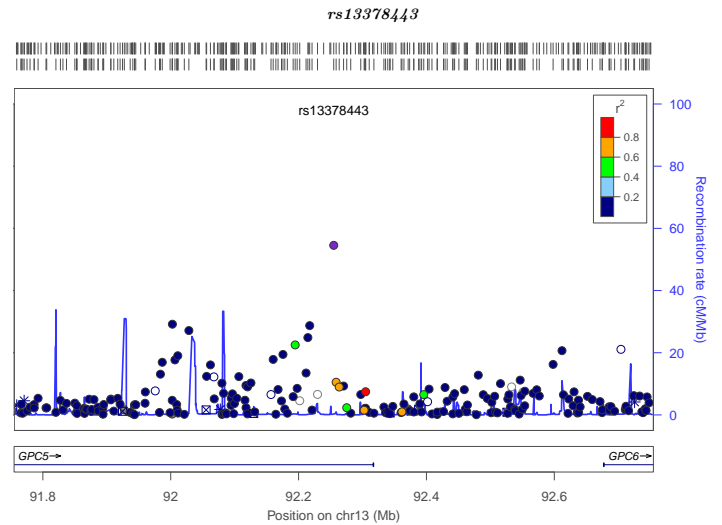


Figure A.122: HS\_CNT(male)

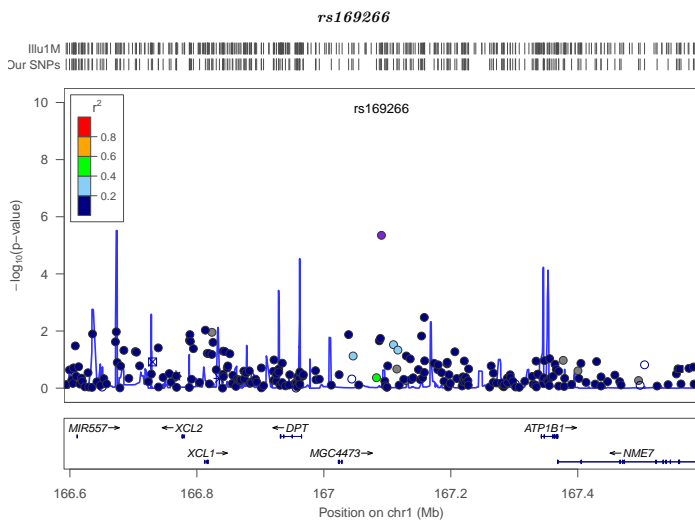


Figure A.123: HS\_CNT(male)

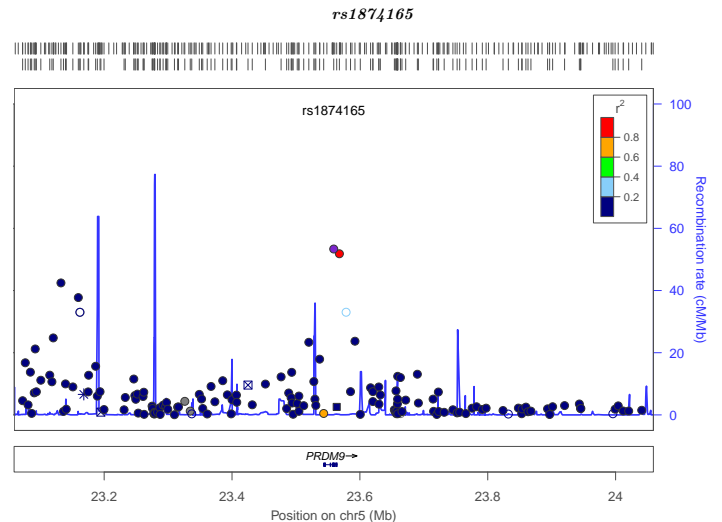


Figure A.124: HS\_CNT(male)

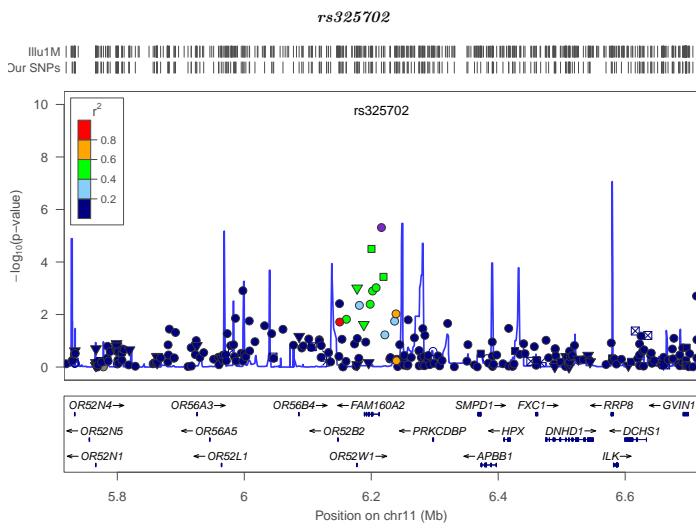


Figure A.125: HS\_CNT(male)

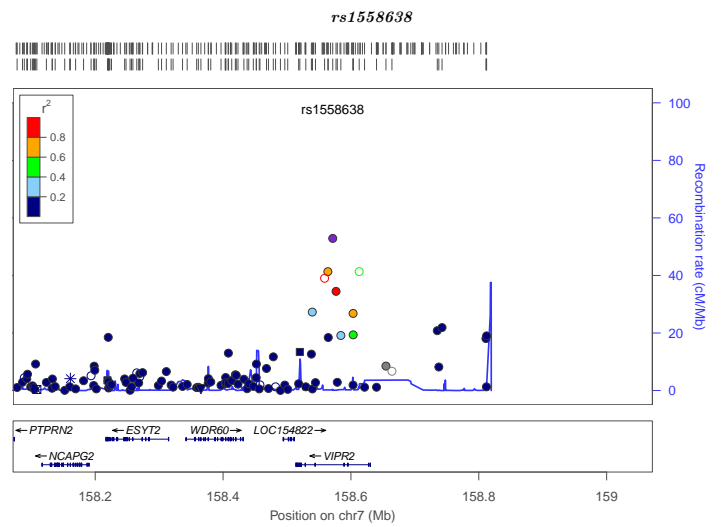


Figure A.126: HS\_CNT(male)

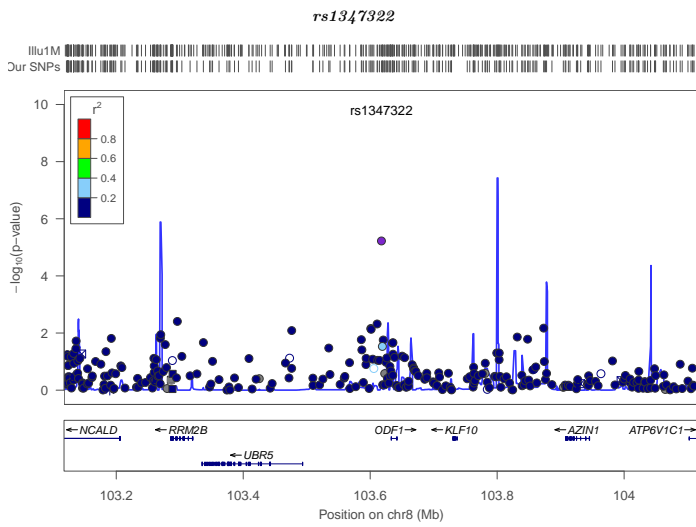


Figure A.127: HS\_CNT(male)

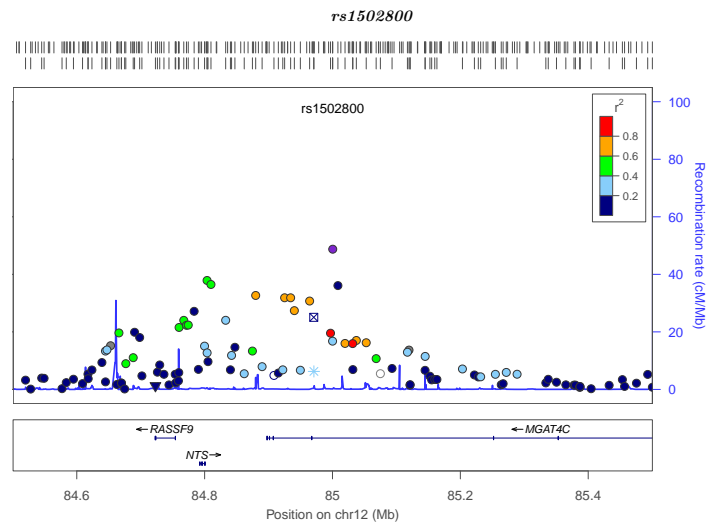


Figure A.128: HS\_CNT(male)

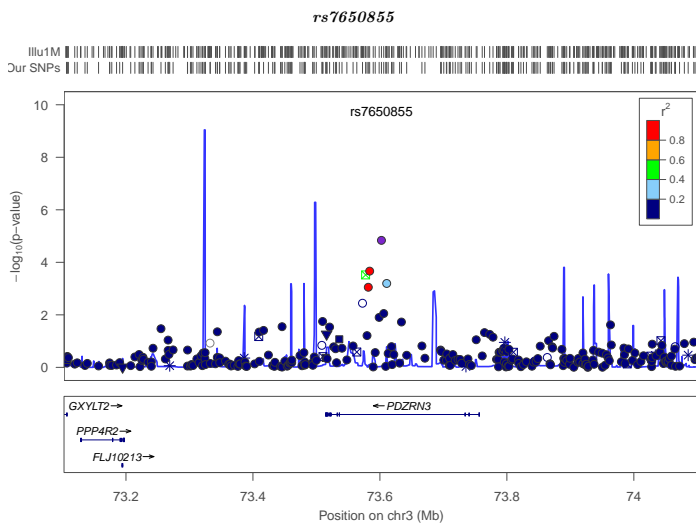


Figure A.129: HS\_CNT(male)

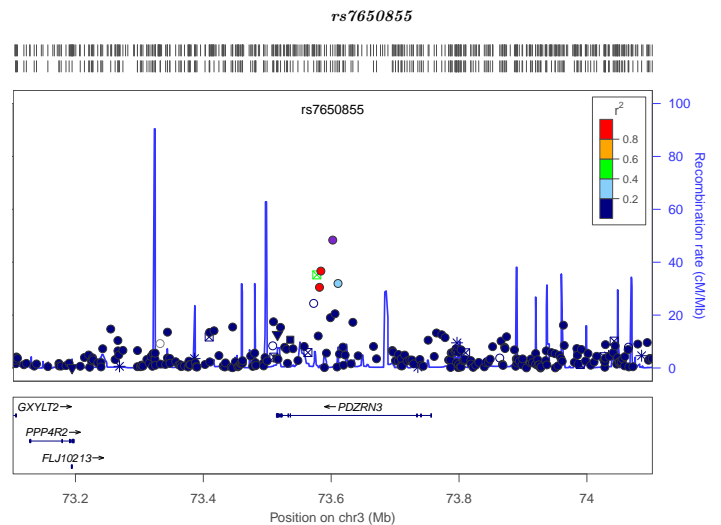


Figure A.130: HS\_CNT(male)

## A.2.4 Phenotype: NHS\_CNT

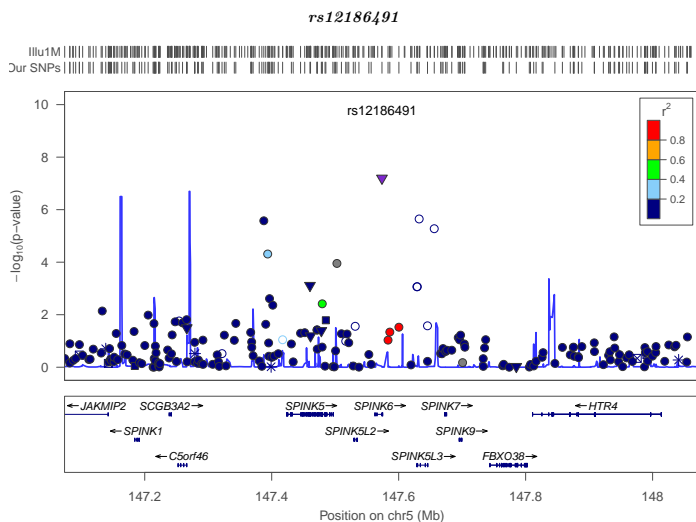


Figure A.131: NHS\_CNT(combined)

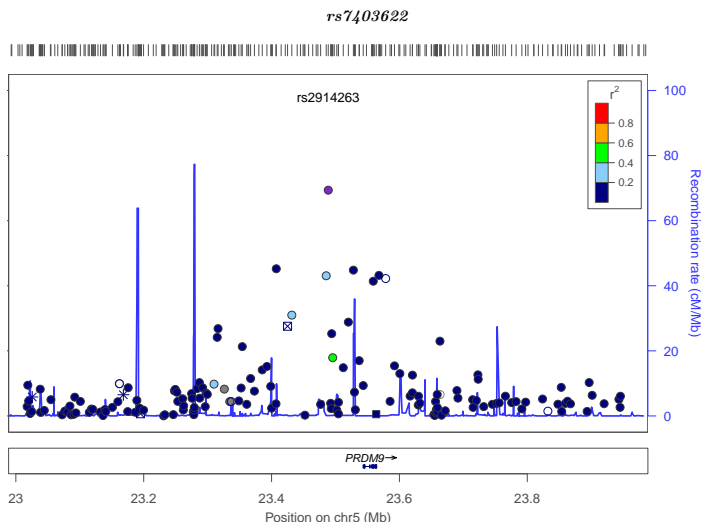


Figure A.132: NHS\_CNT(combined)

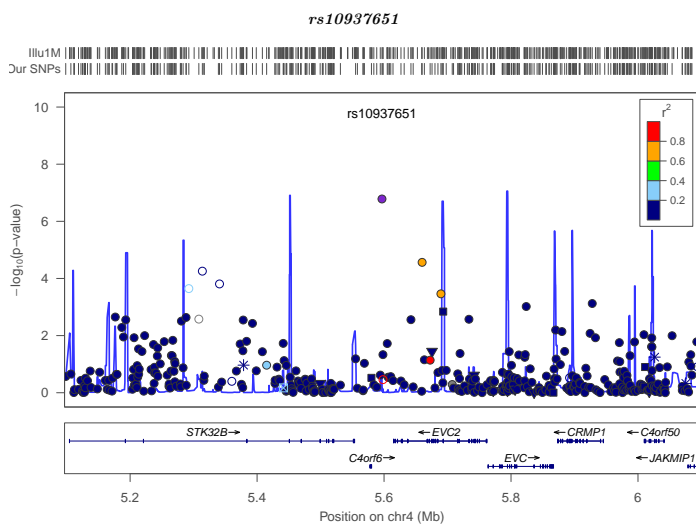


Figure A.133: NHS\_CNT(combined)

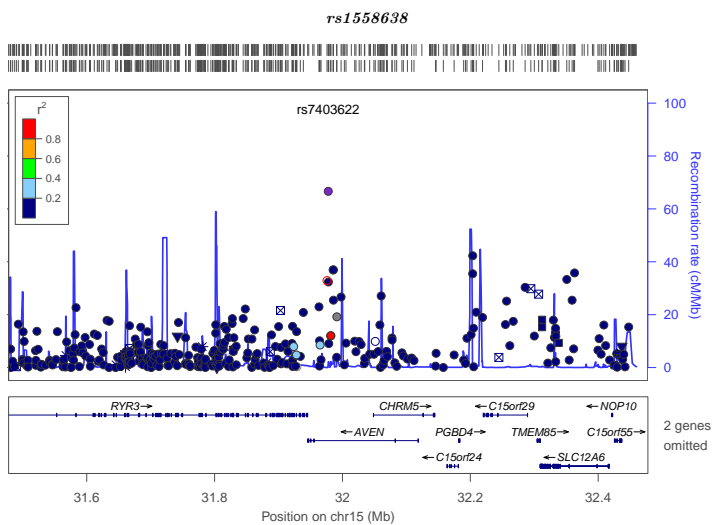


Figure A.134: NHS\_CNT(combined)

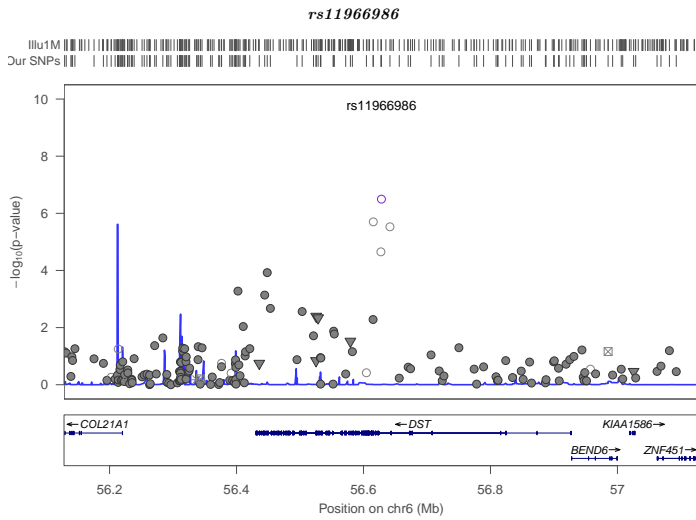


Figure A.135: NHS\_CNT(combined)

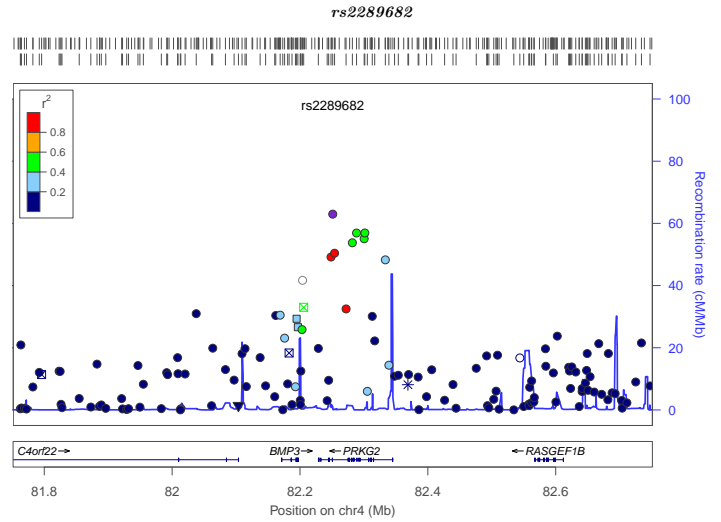


Figure A.136: NHS\_CNT(combined)

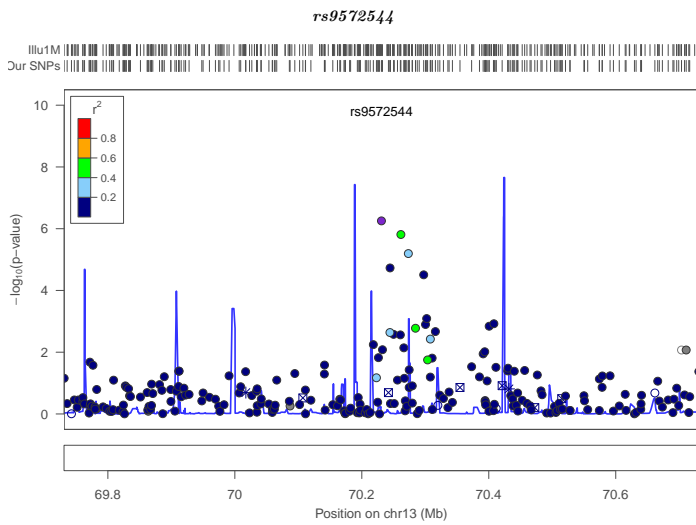


Figure A.137: NHS\_CNT(combined)

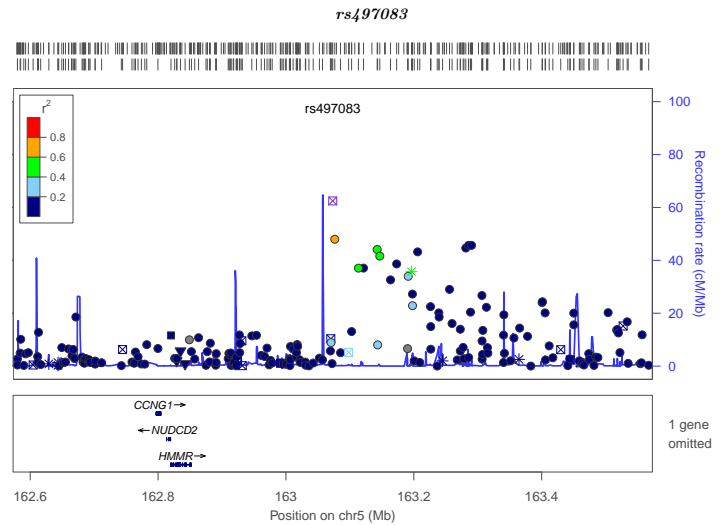


Figure A.138: NHS\_CNT(combined)

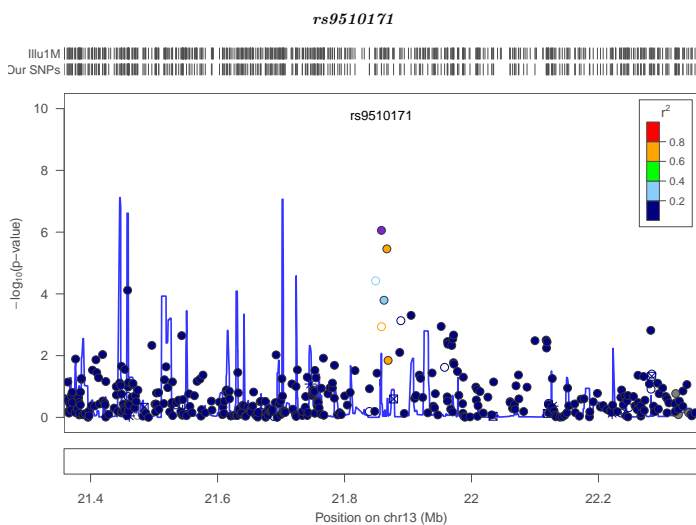


Figure A.139: NHS\_CNT(combined)

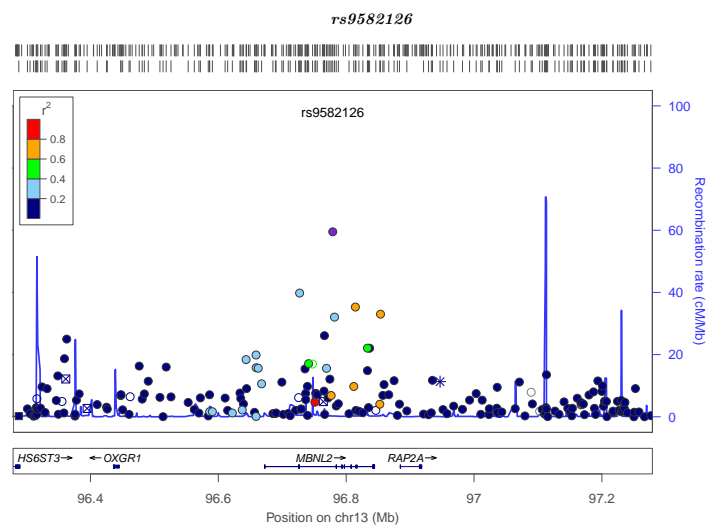


Figure A.140: NHS\_CNT(combined)

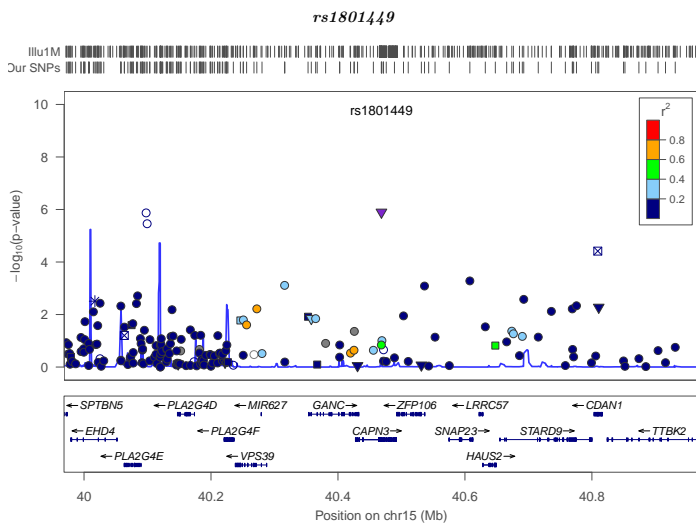


Figure A.141: NHS\_CNT(combined)

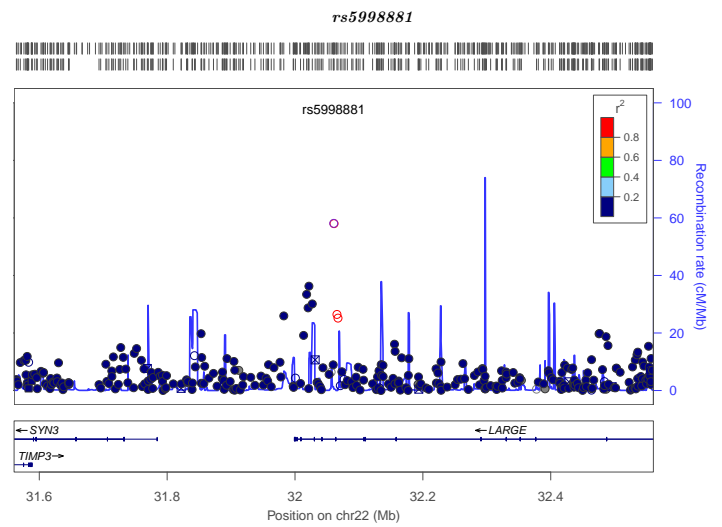


Figure A.142: NHS\_CNT(combined)

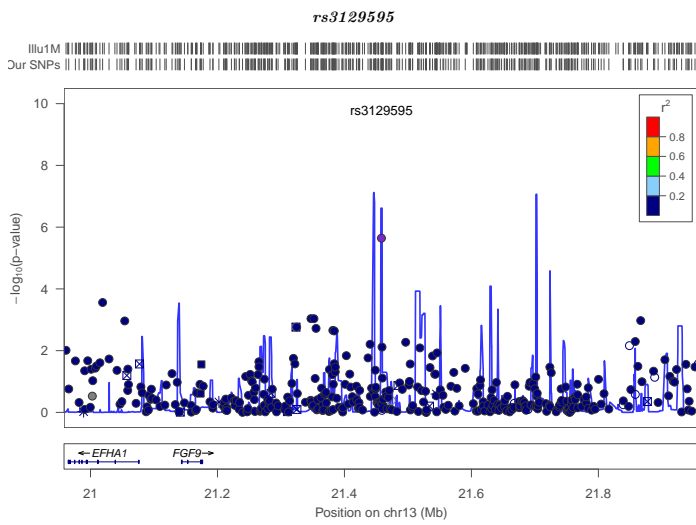


Figure A.143: NHS\_CNT(female)

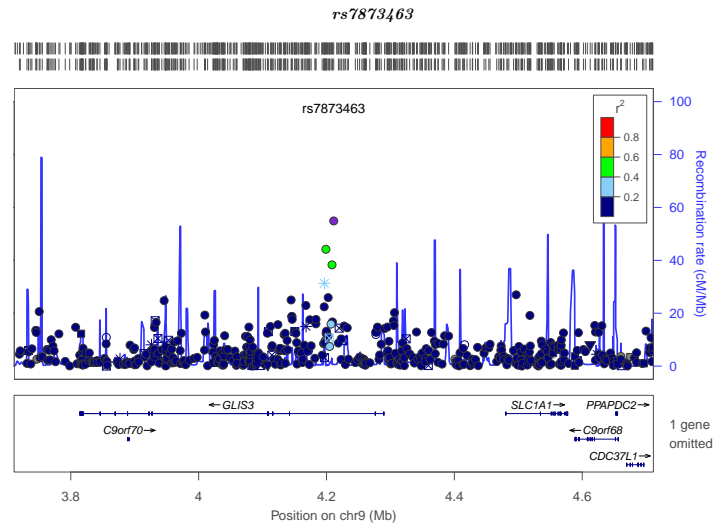


Figure A.144: NHS\_CNT(female)

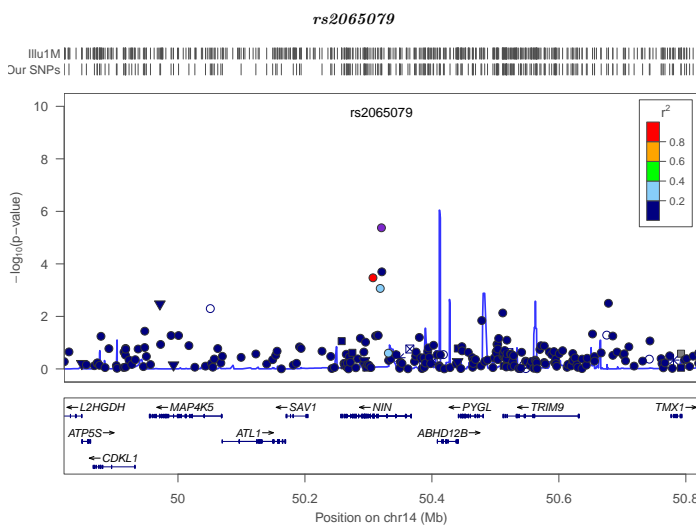


Figure A.145: NHS\_CNT(female)

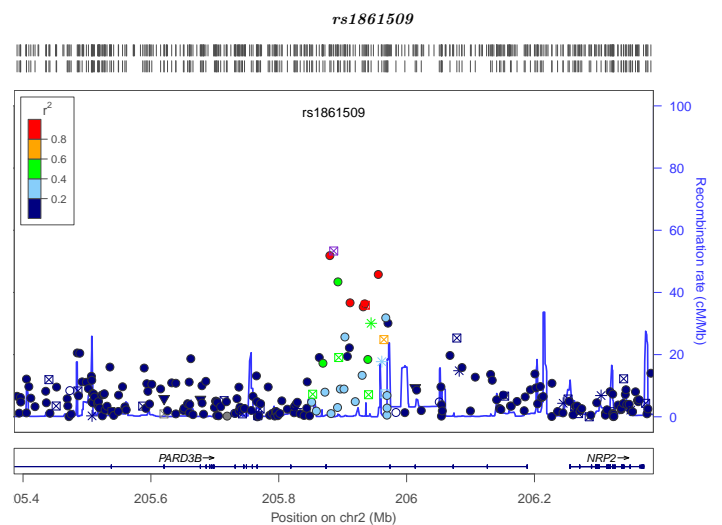


Figure A.146: NHS\_CNT(female)

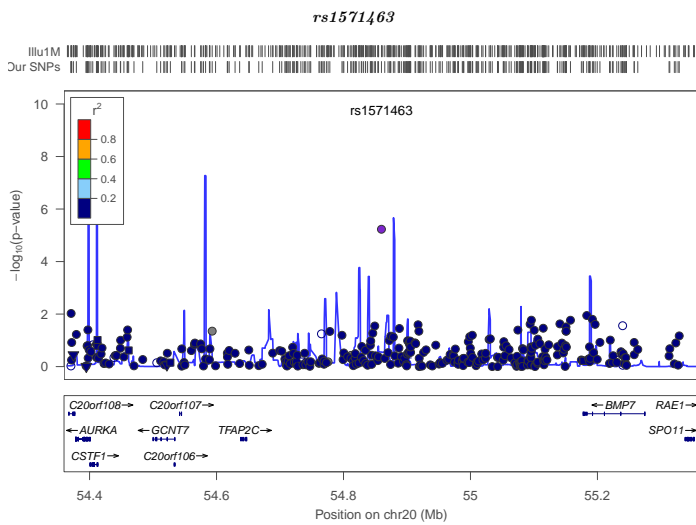


Figure A.147: NHS\_CNT(female)

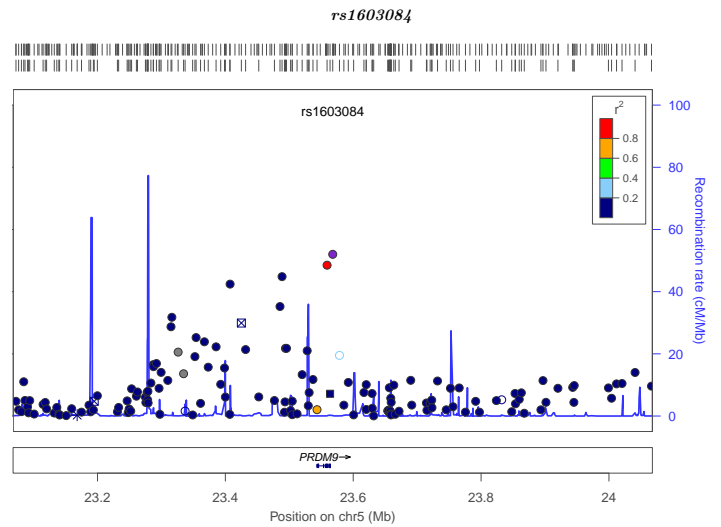


Figure A.148: NHS\_CNT(female)

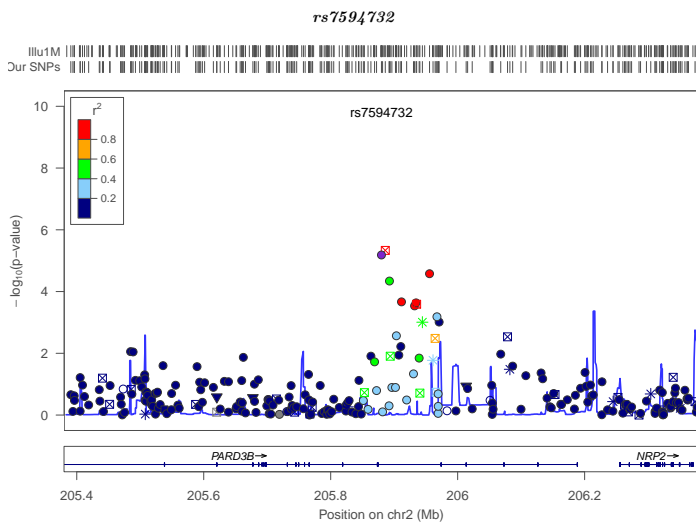


Figure A.149: NHS\_CNT(female)

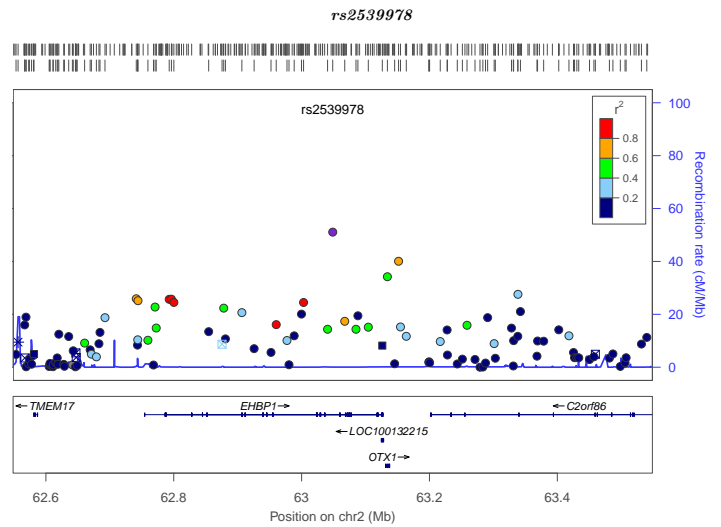


Figure A.150: NHS\_CNT(female)

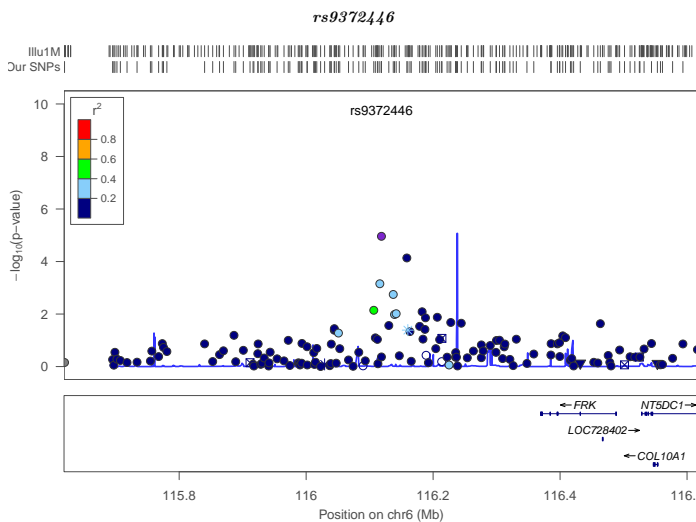


Figure A.151: NHS\_CNT(female)

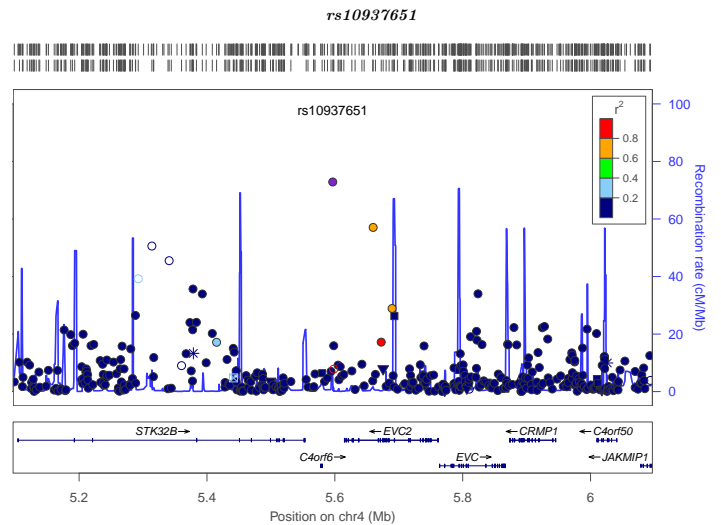


Figure A.152: NHS\_CNT(male)



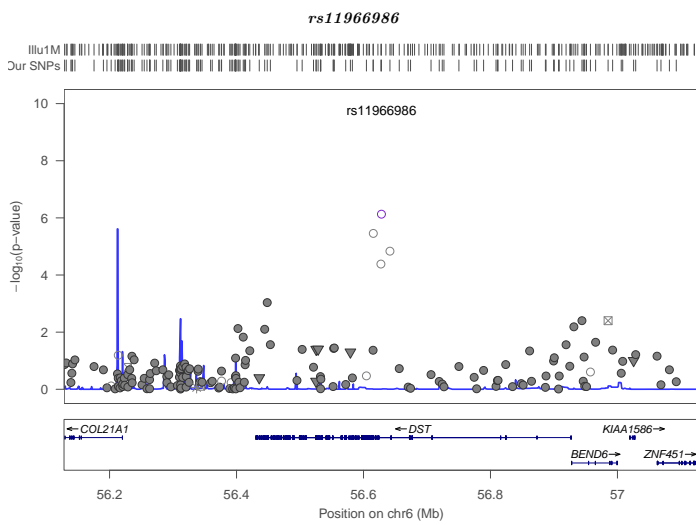


Figure A.153: NHS\_CNT(male)

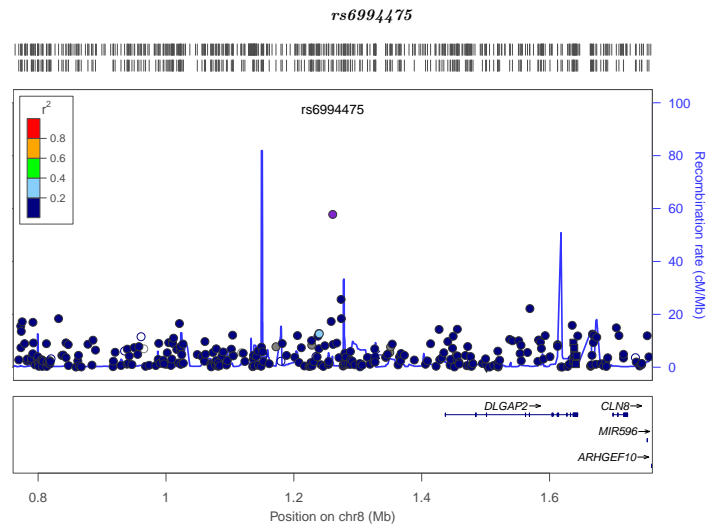


Figure A.154: NHS\_CNT(male)

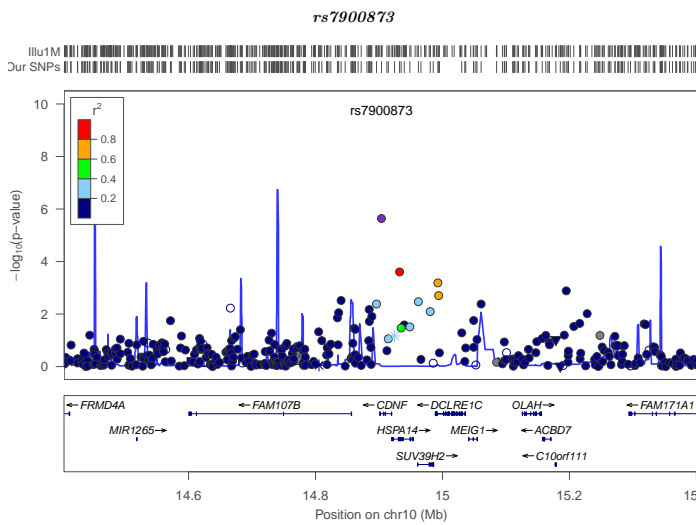


Figure A.155: NHS\_CNT(male)

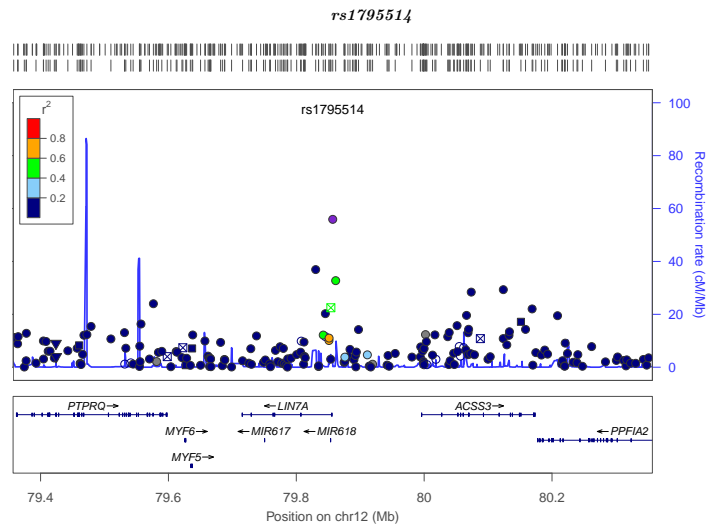


Figure A.156: NHS\_CNT(male)

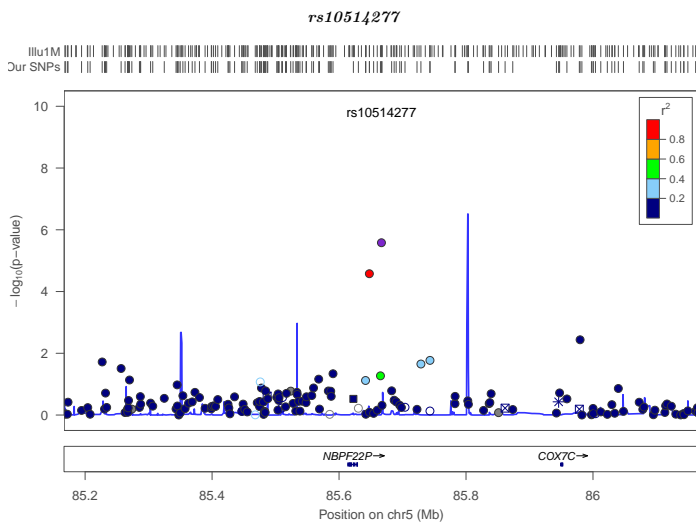


Figure A.157: NHS\_CNT(male)

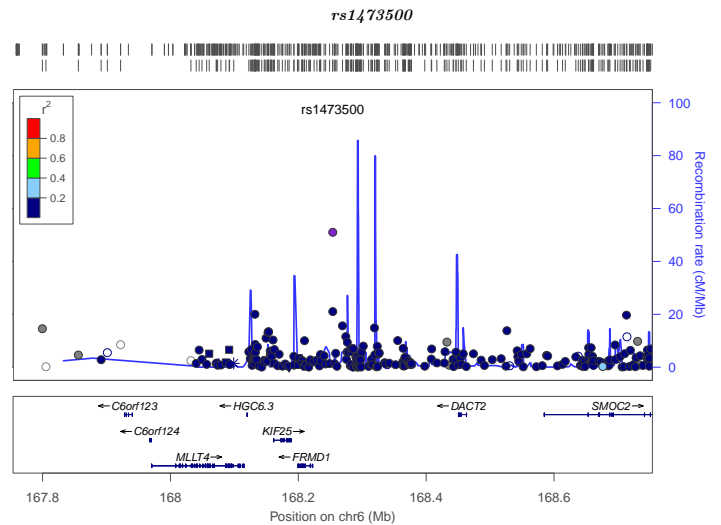


Figure A.158: NHS\_CNT(male)

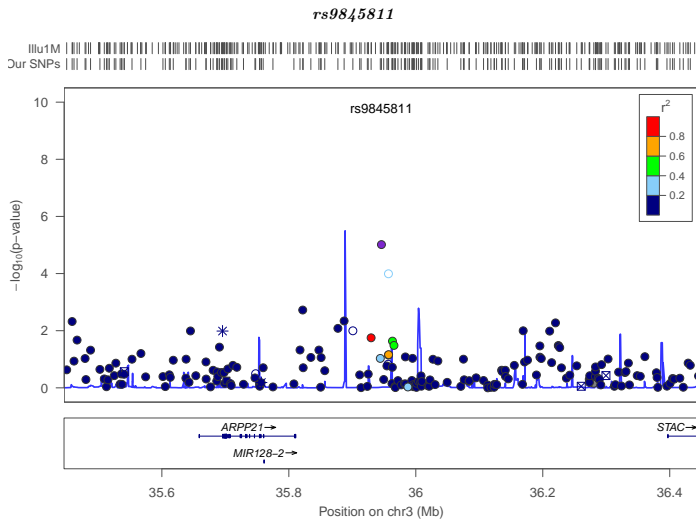


Figure A.159: NHS\_CNT(male)

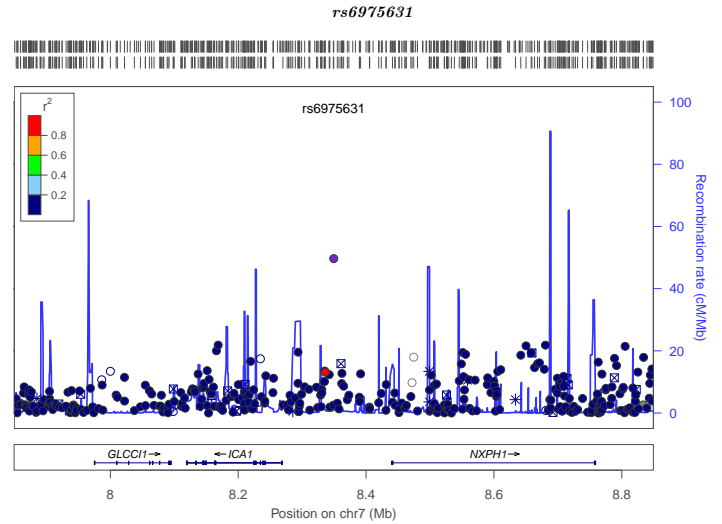


Figure A.160: NHS\_CNT(male)

## A.2.5 Phenotype: MOTIF

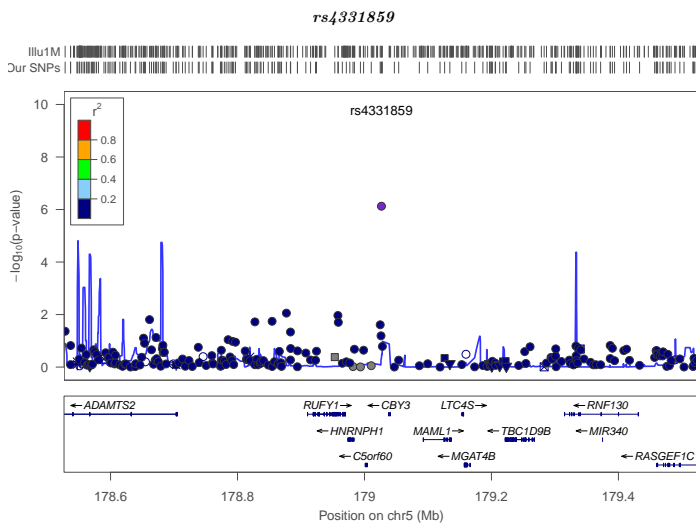


Figure A.161: MOTIF(combined)

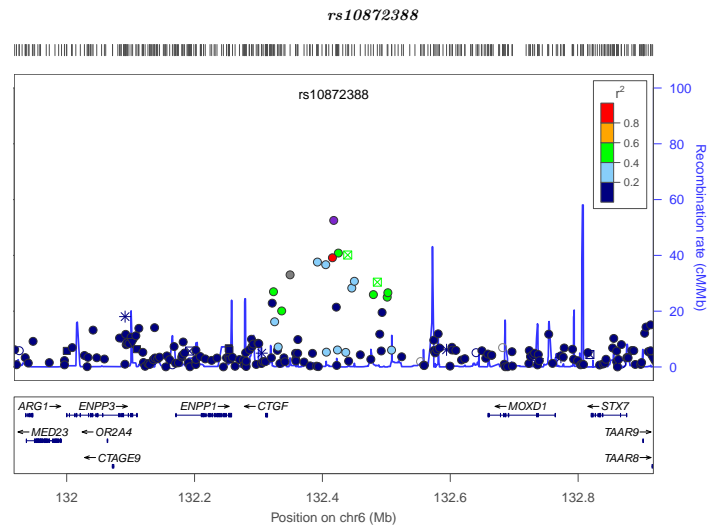


Figure A.162: MOTIF(combined)

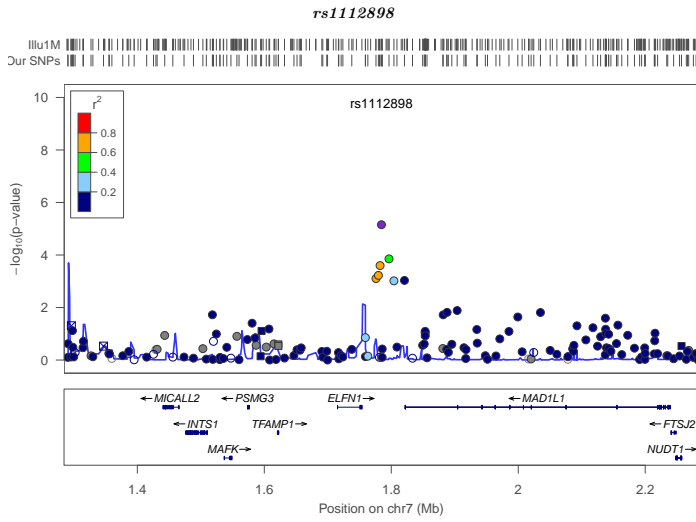


Figure A.163: MOTIF(combined)

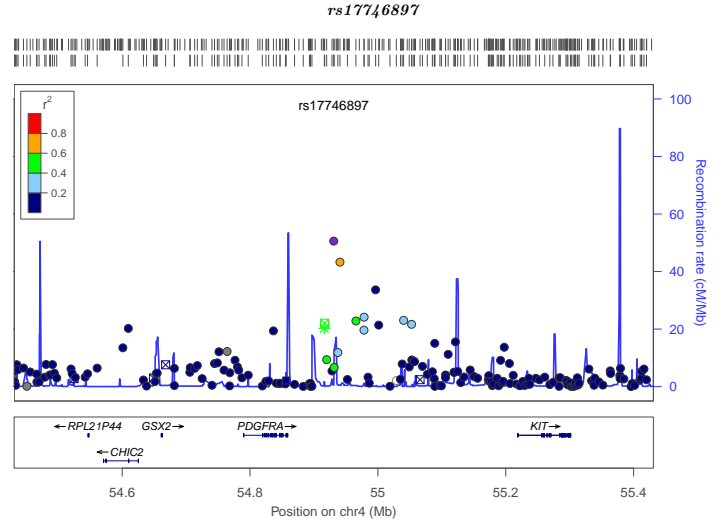


Figure A.164: MOTIF(combined)

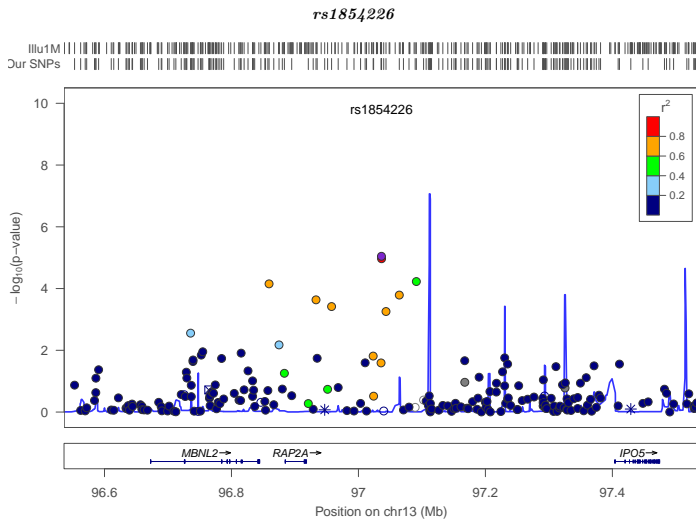


Figure A.165: MOTIF(combined)

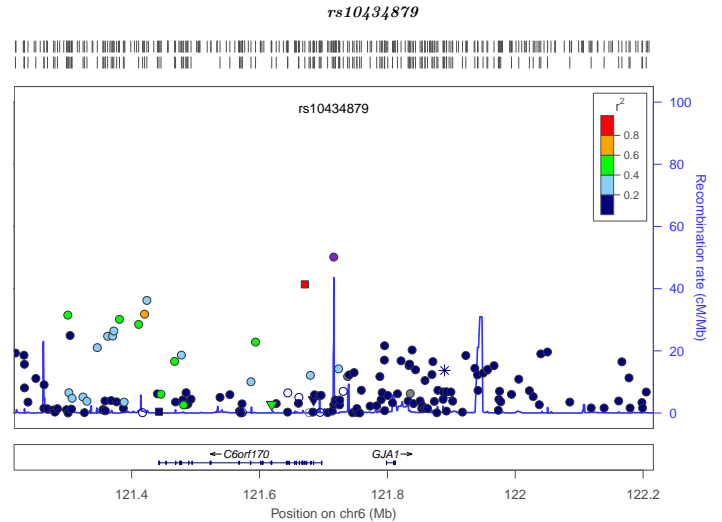


Figure A.166: MOTIF(combined)

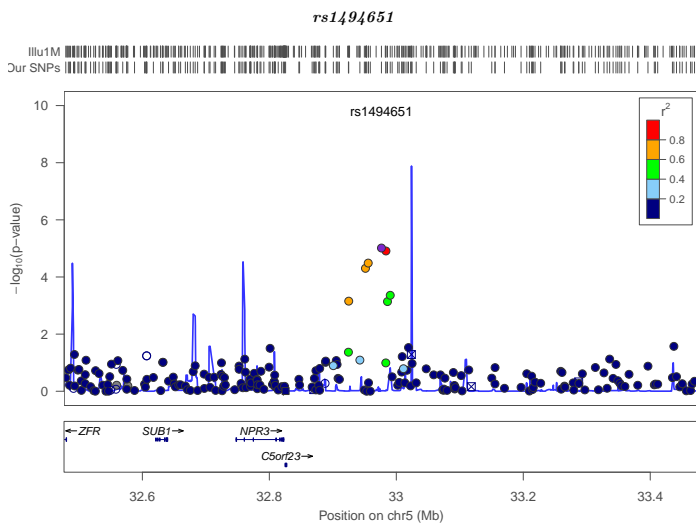


Figure A.167: MOTIF(combined)

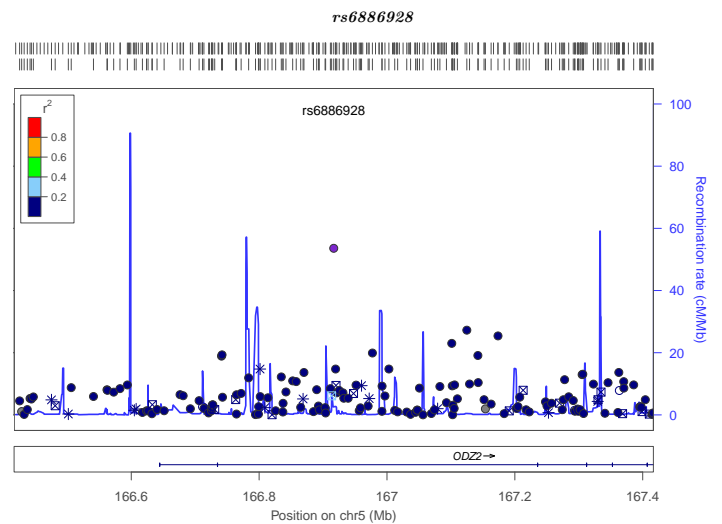


Figure A.168: MOTIF(female)

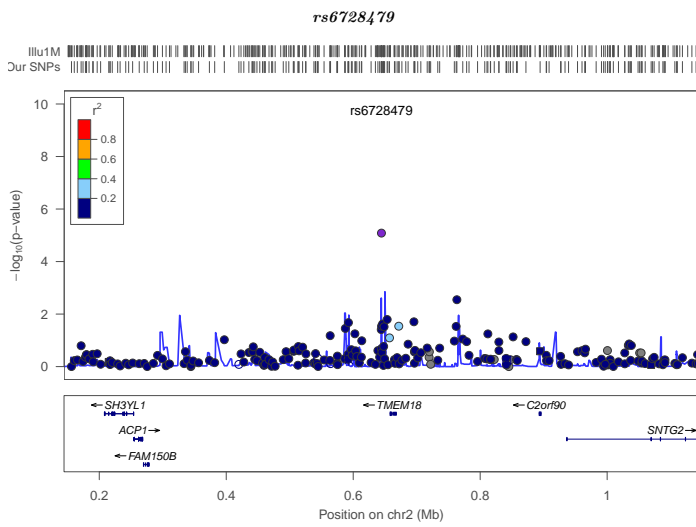


Figure A.169: MOTIF(female)

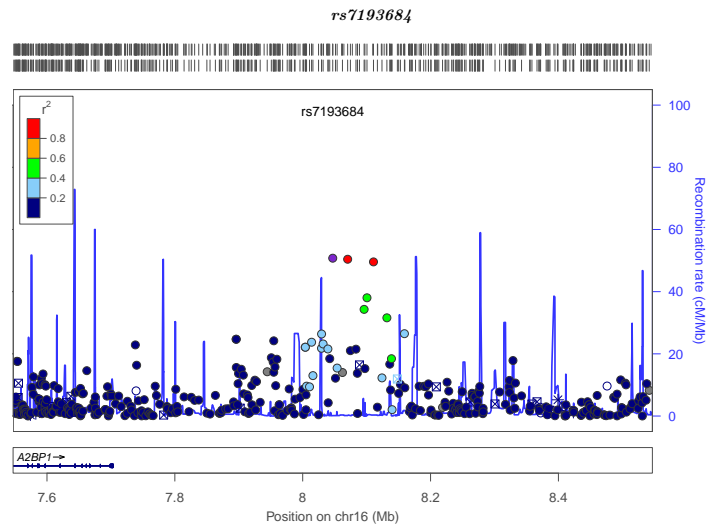


Figure A.170: MOTIF(female)

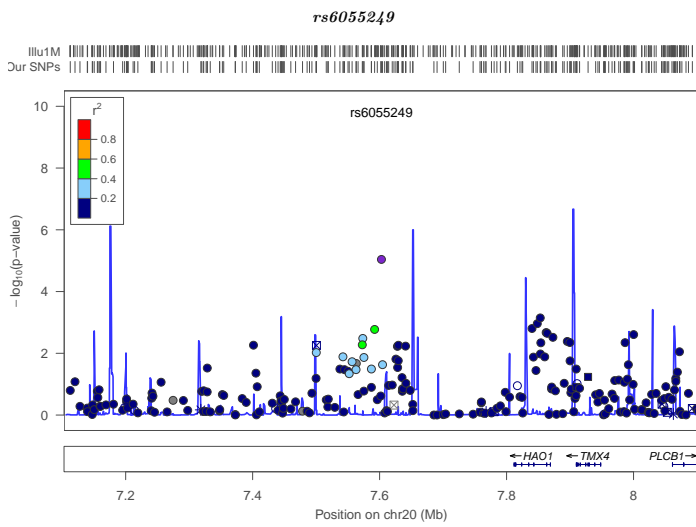


Figure A.171: MOTIF(female)

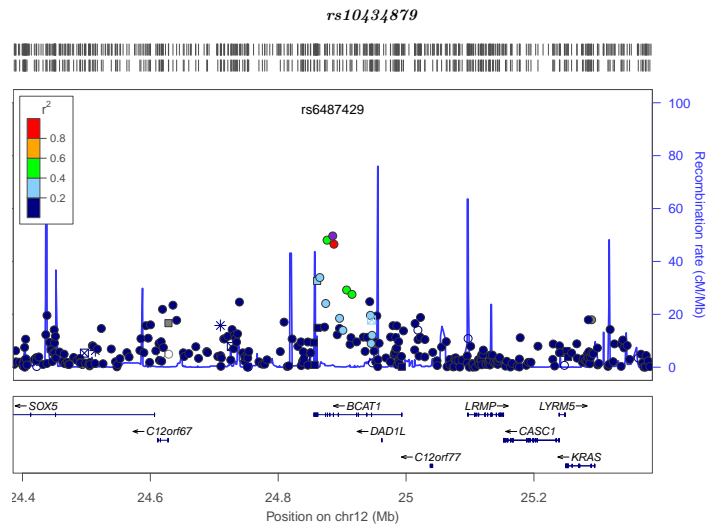


Figure A.172: MOTIF(female)

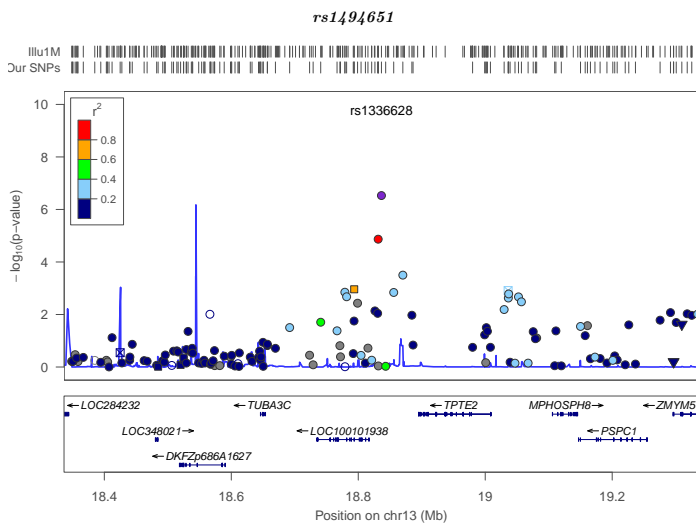


Figure A.173: MOTIF(male)

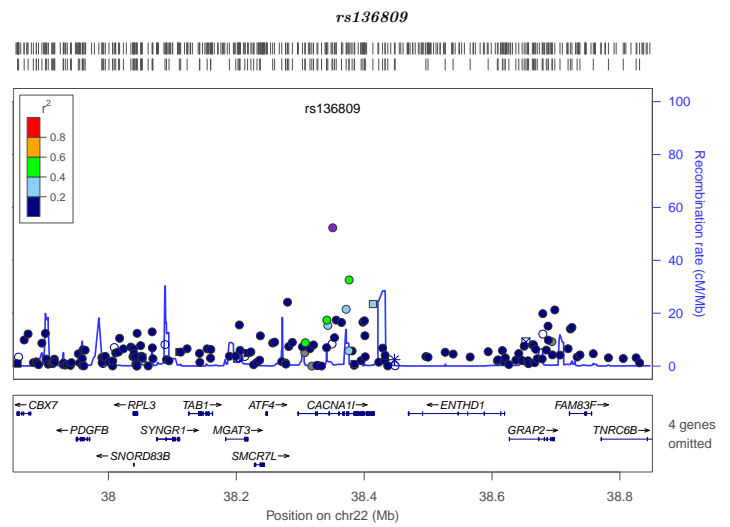


Figure A.174: MOTIF(male)

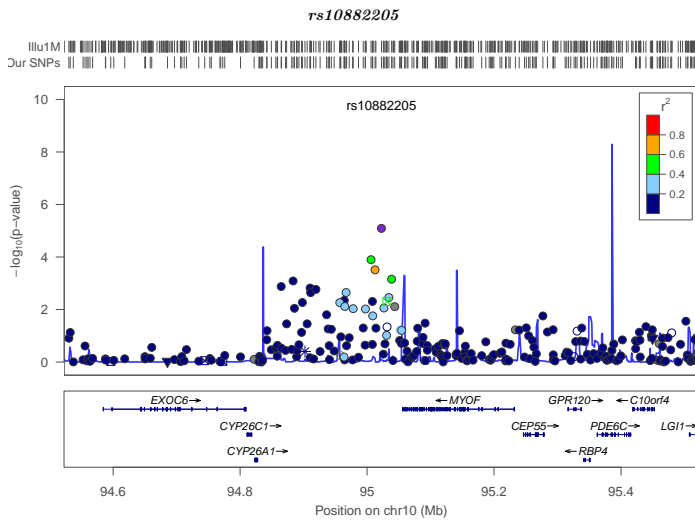


Figure A.175: MOTIF(male)

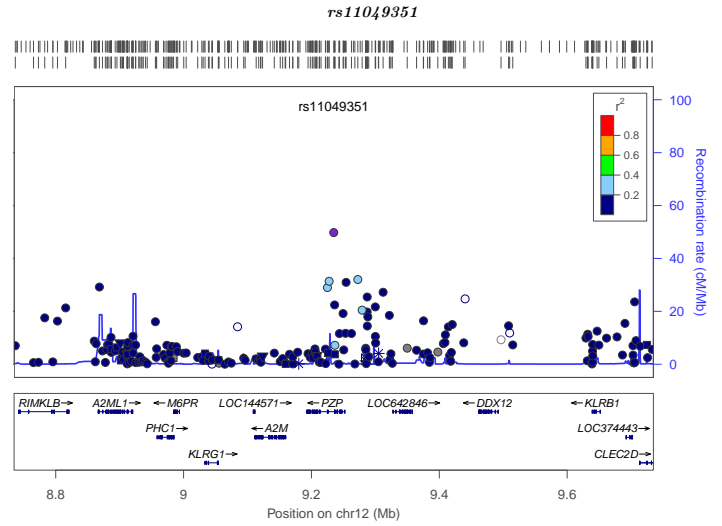


Figure A.176: MOTIF(male)

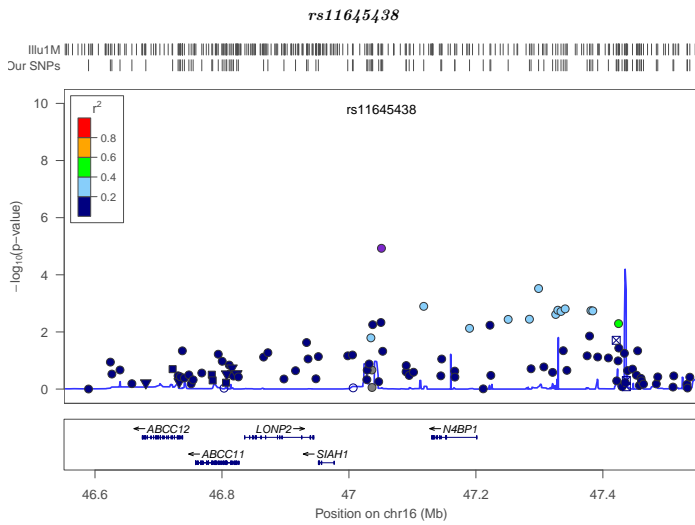


Figure A.177: MOTIF(male)

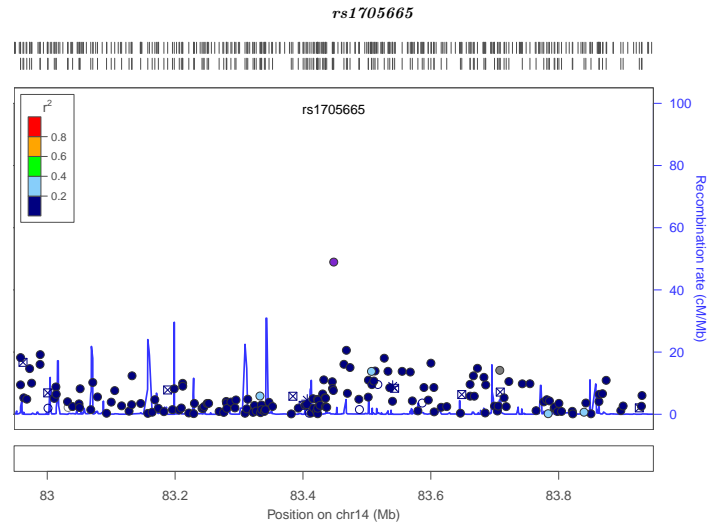
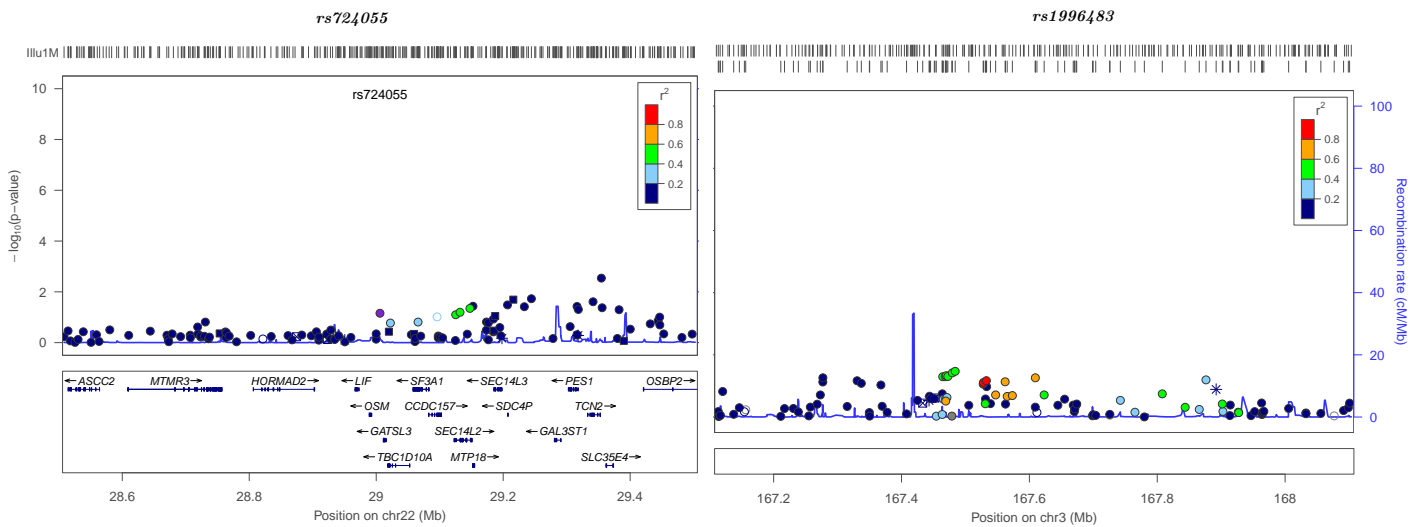
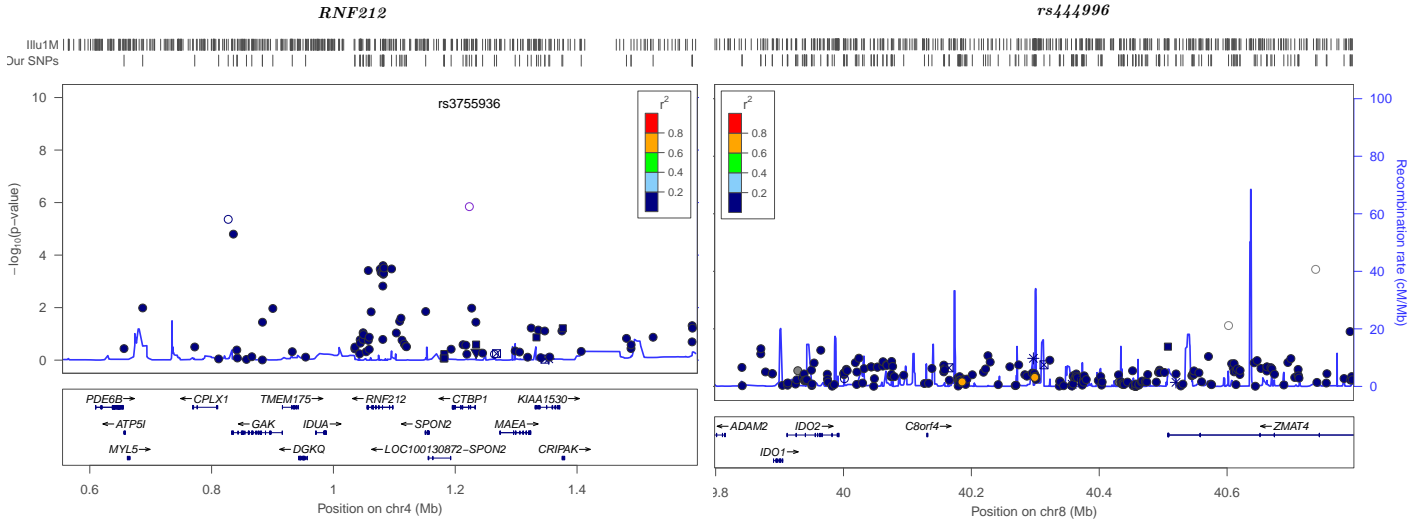
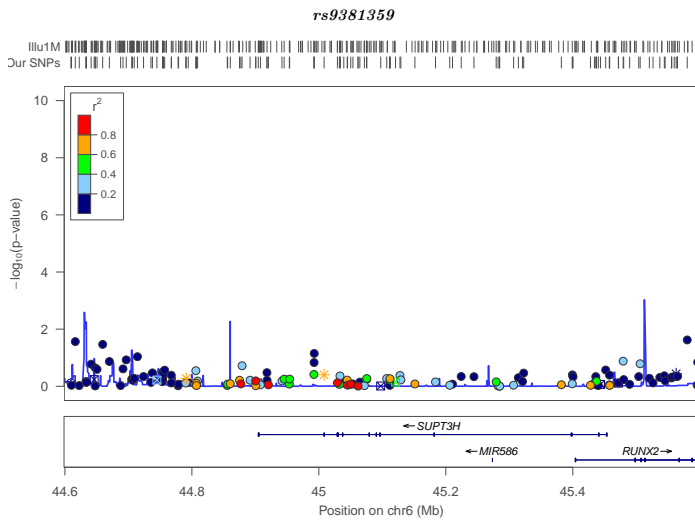


Figure A.178: MOTIF(male)

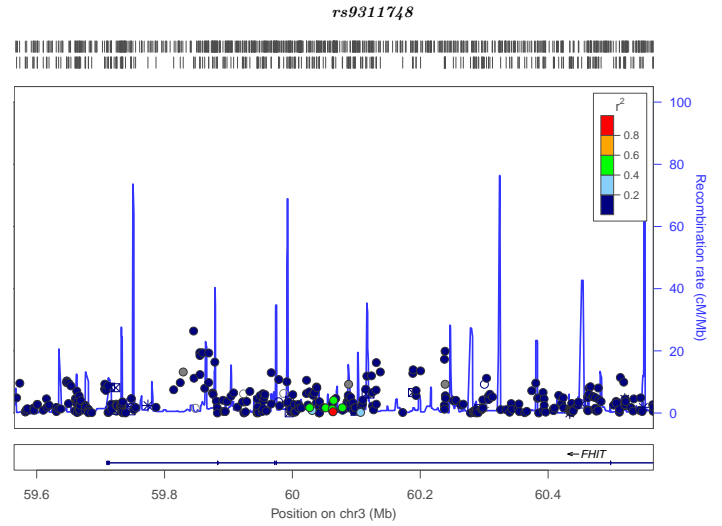
## A.3 LOCUSZOOM PLOT OF OUR TOP HITS IN FHS DATA SET

### A.3.1 Phenotype: ARC

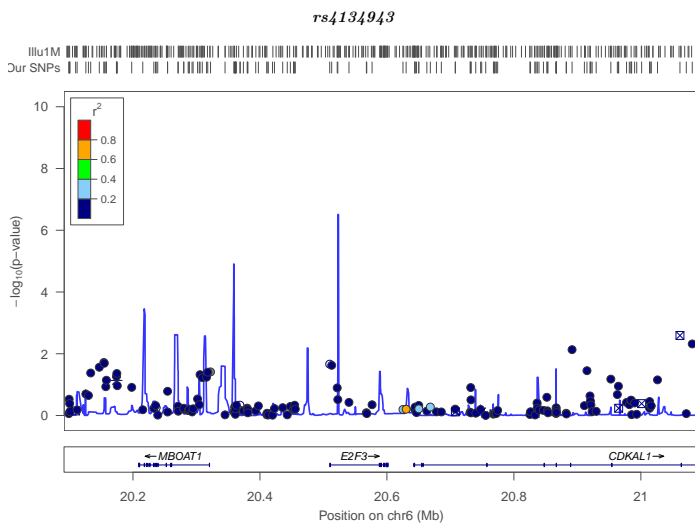




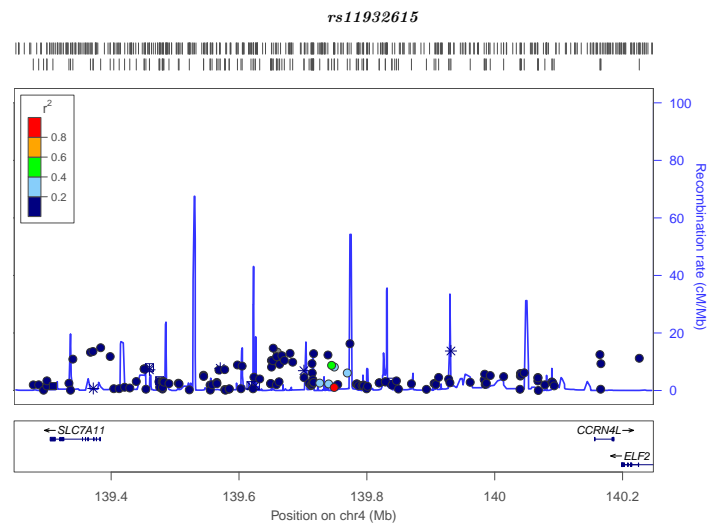
**Figure A.183: ARC(combined)**



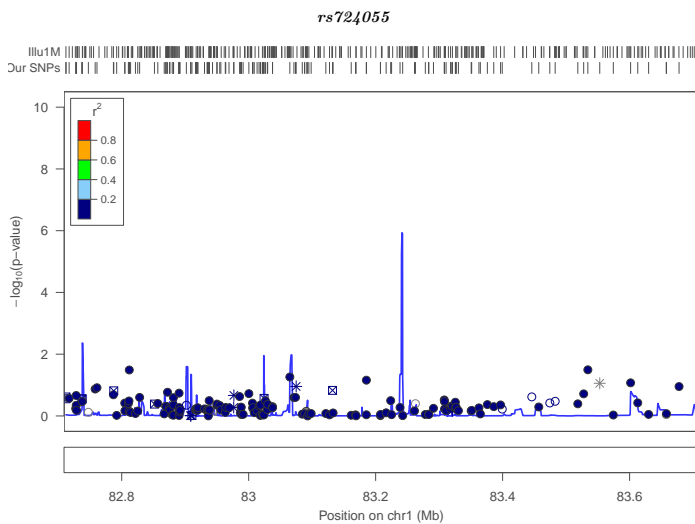
**Figure A.184: ARC(combined)**



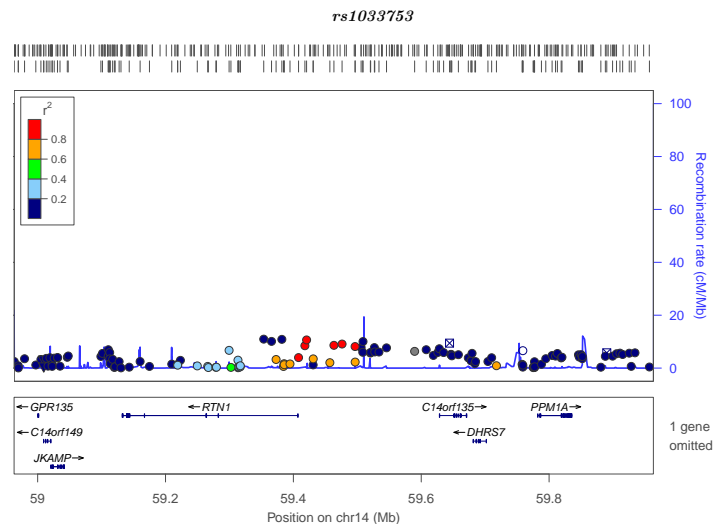
**Figure A.185: ARC(combined)**



**Figure A.186: ARC(combined)**



**Figure A.187: ARC(combined)**



**Figure A.188: ARC(combined)**

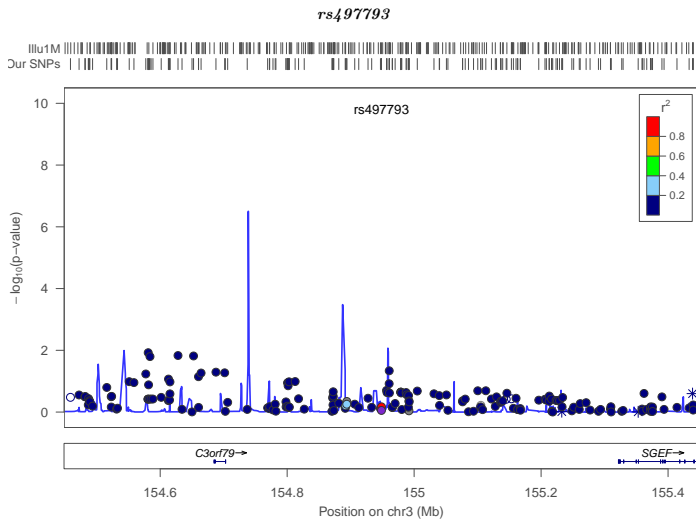


Figure A.189: ARC(female)

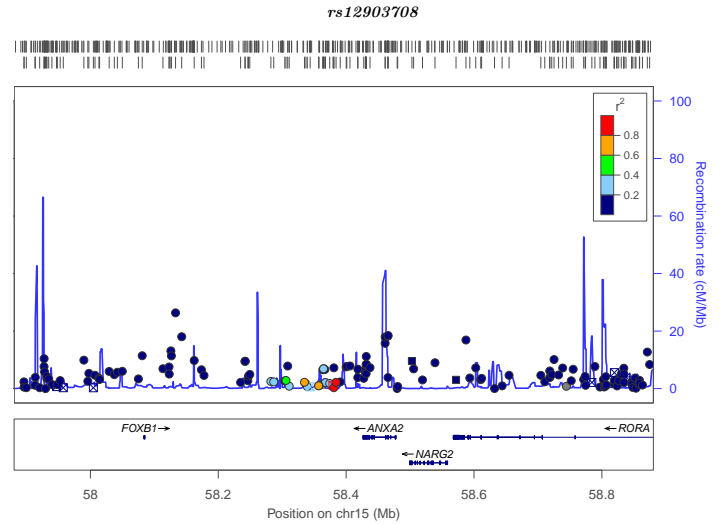


Figure A.190: ARC(female)

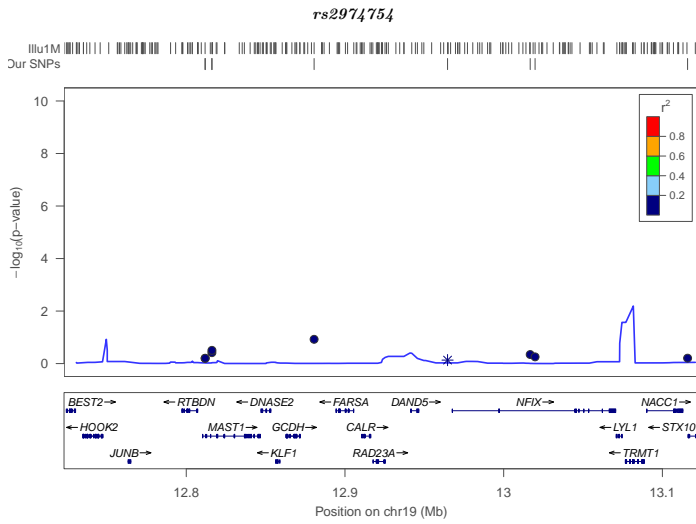


Figure A.191: ARC(female)

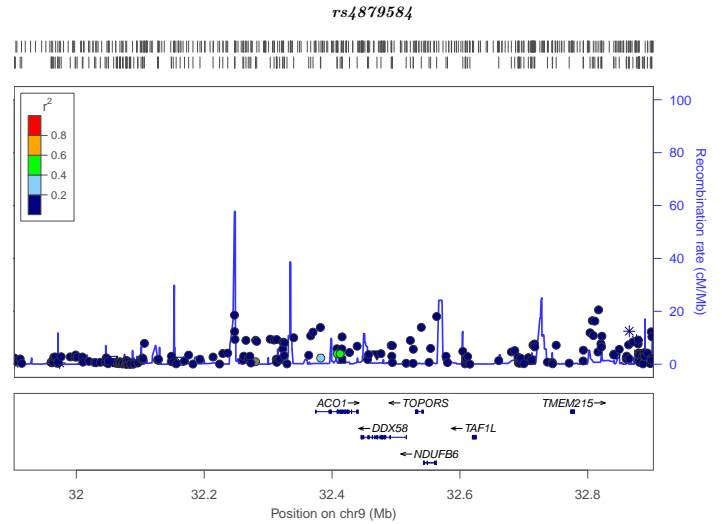


Figure A.192: ARC(female)

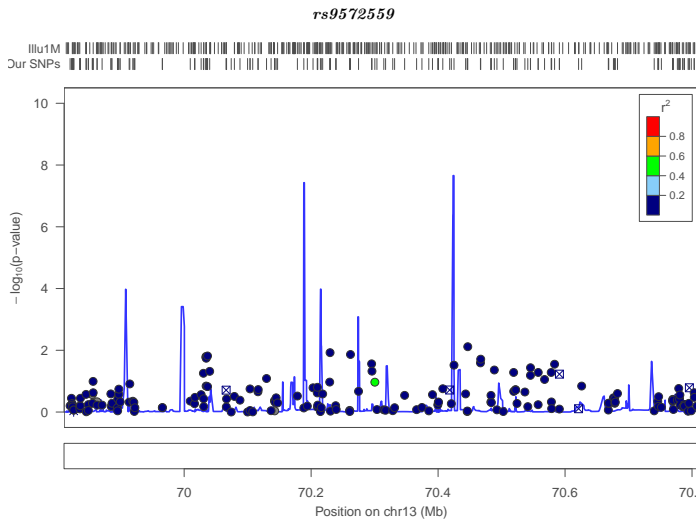


Figure A.193: ARC(female)

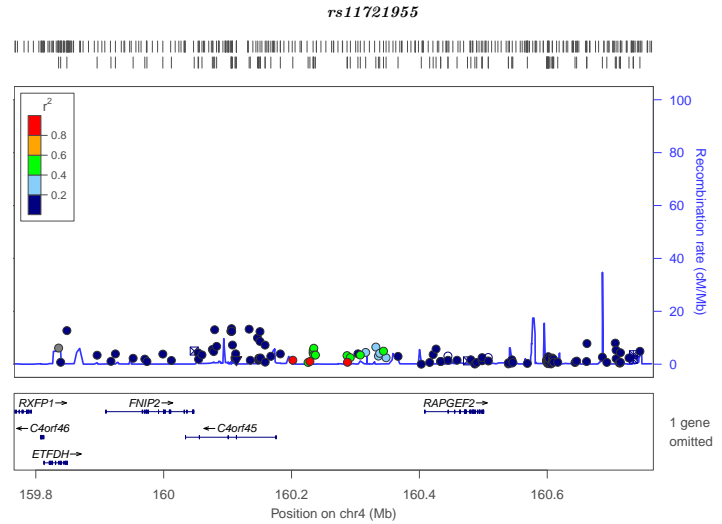


Figure A.194: ARC(female)



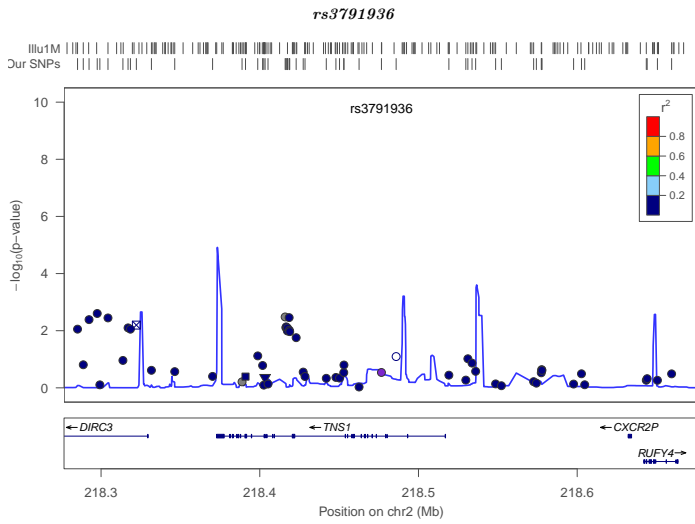


Figure A.195: ARC(female)

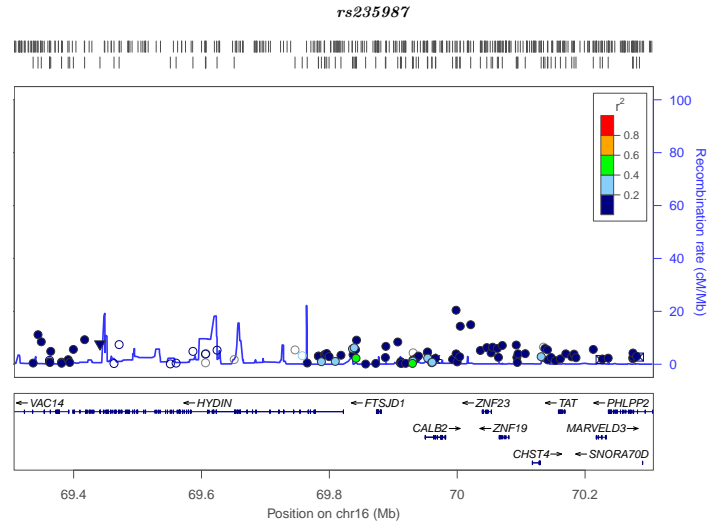


Figure A.196: ARC(female)

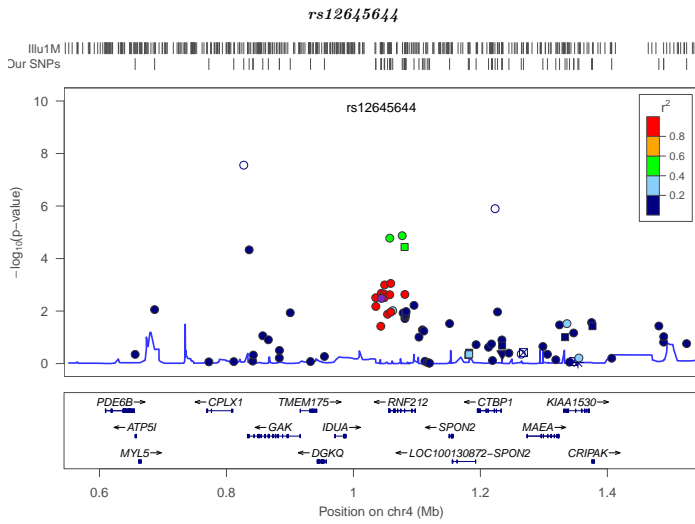


Figure A.197: ARC(male)

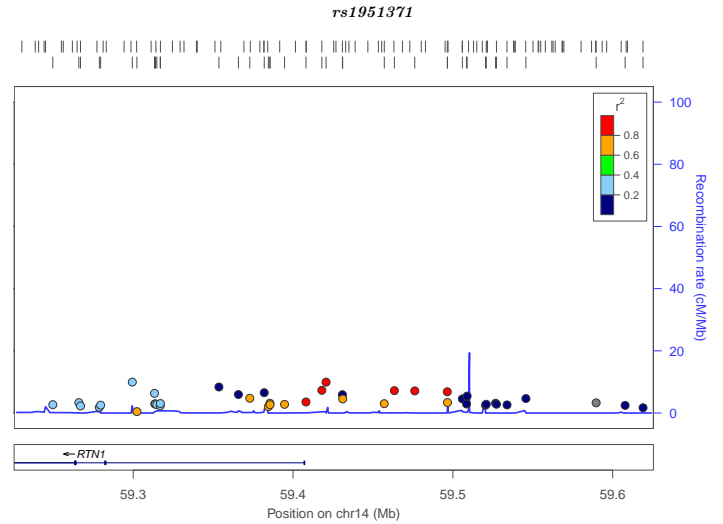


Figure A.198: ARC(male)

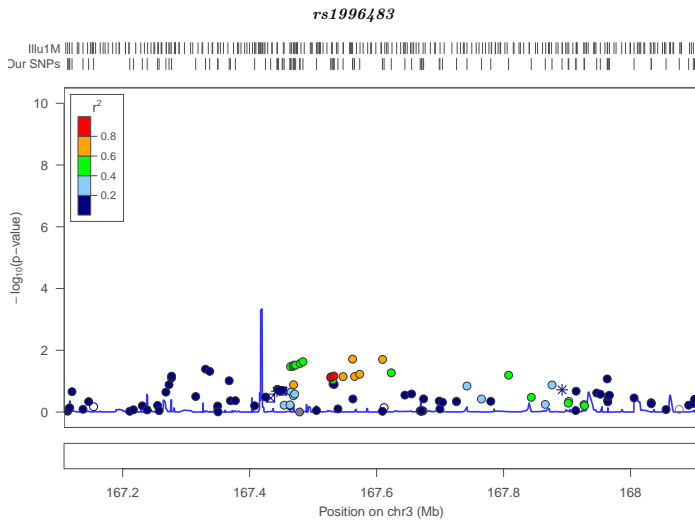


Figure A.199: ARC(male)

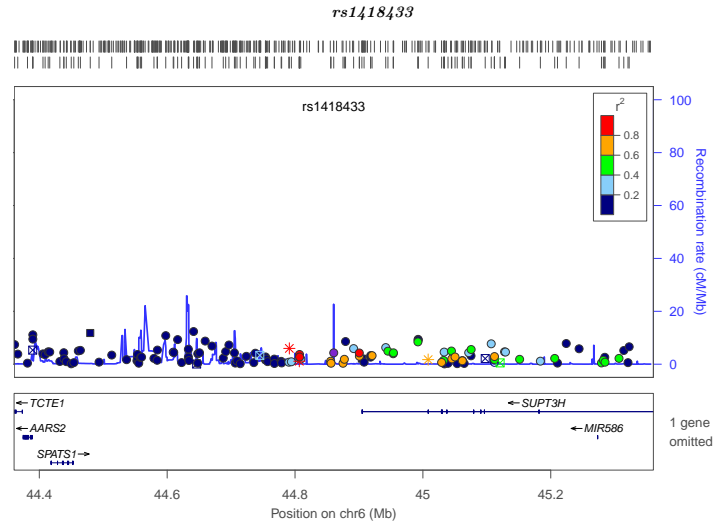


Figure A.200: ARC(male)

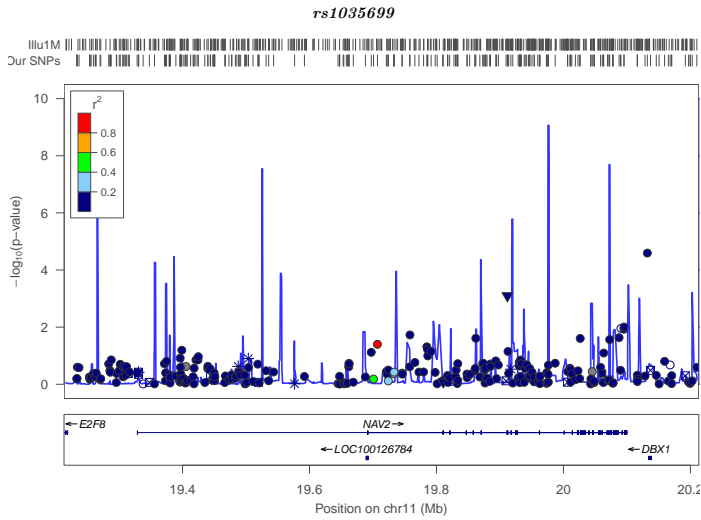


Figure A.201: ARC(male)

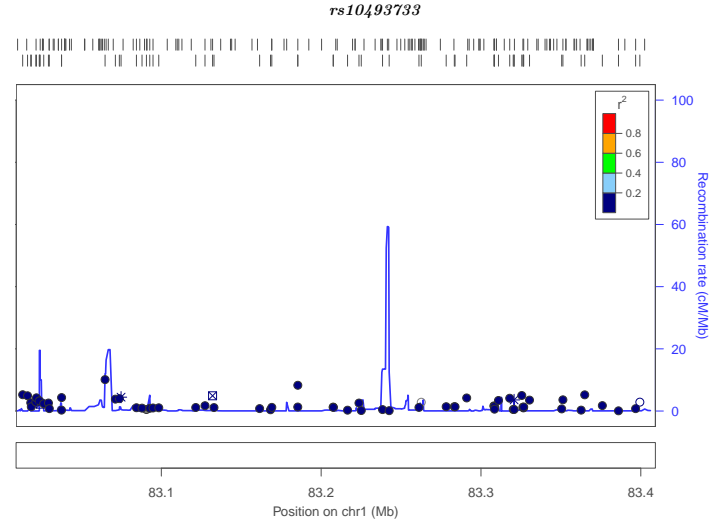


Figure A.202: ARC(male)

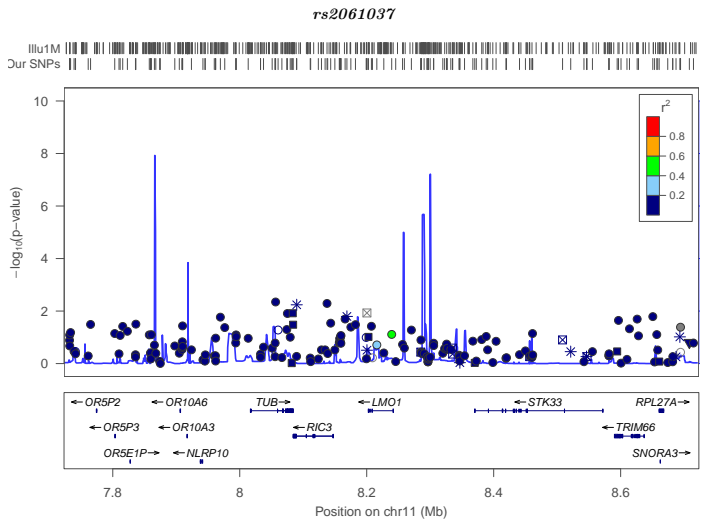


Figure A.203: ARC(male)

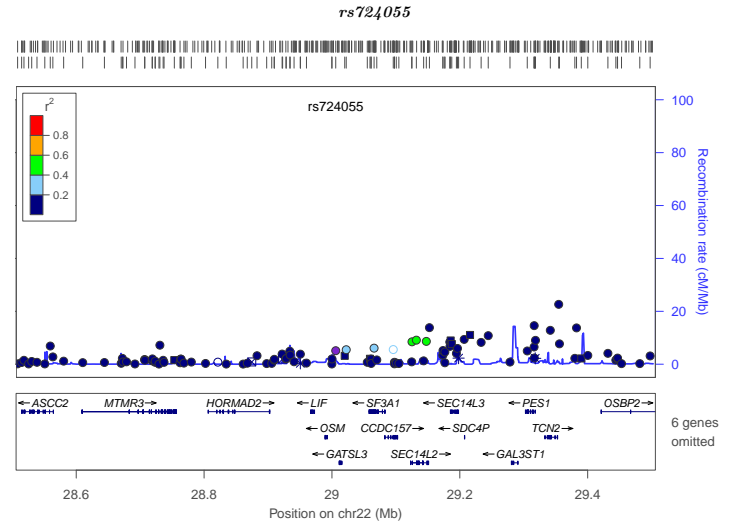


Figure A.204: ARC(male)

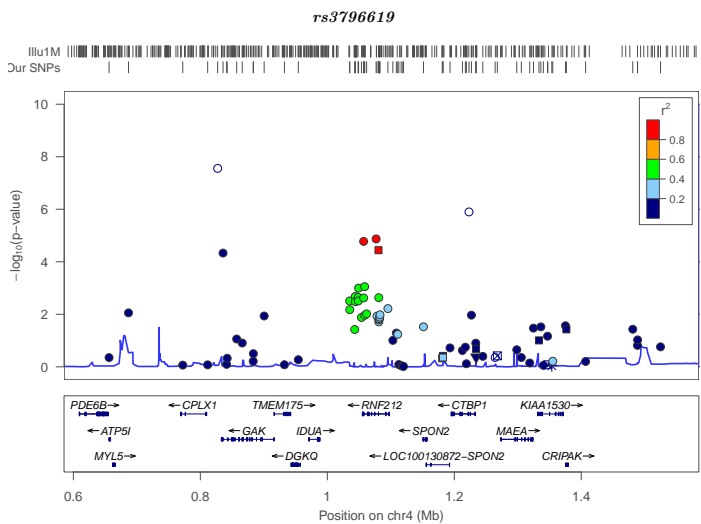


Figure A.205: ARC(male)

### A.3.2 Phenotype: HS\_PCT

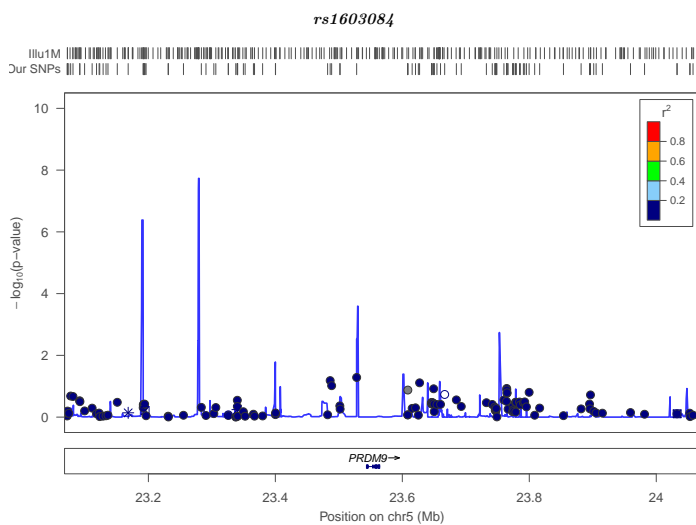


Figure A.206: HS\_PCT(combined)

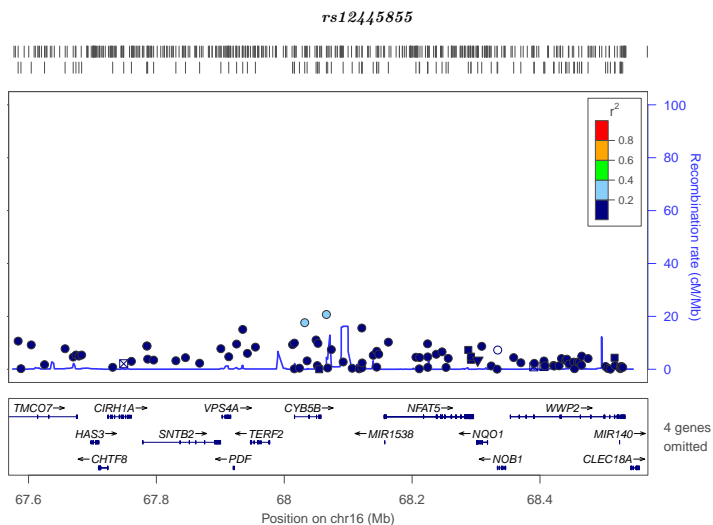


Figure A.207: HS\_PCT(combined)

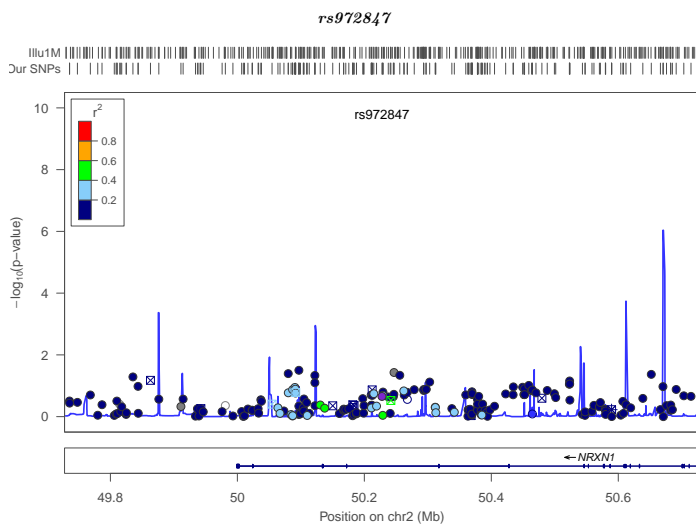


Figure A.208: HS\_PCT(combined)

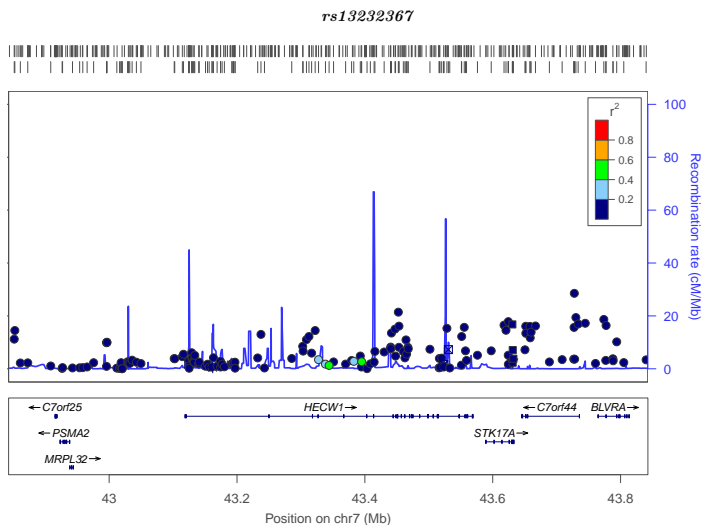


Figure A.209: HS\_PCT(combined)

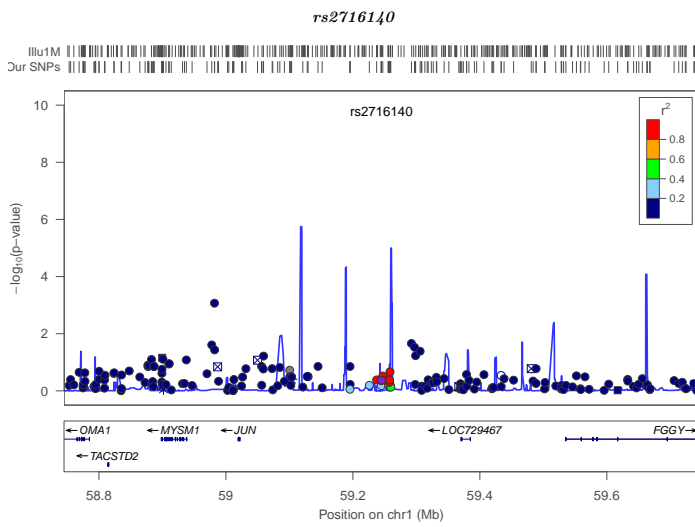


Figure A.210: HS\_PCT(combined)

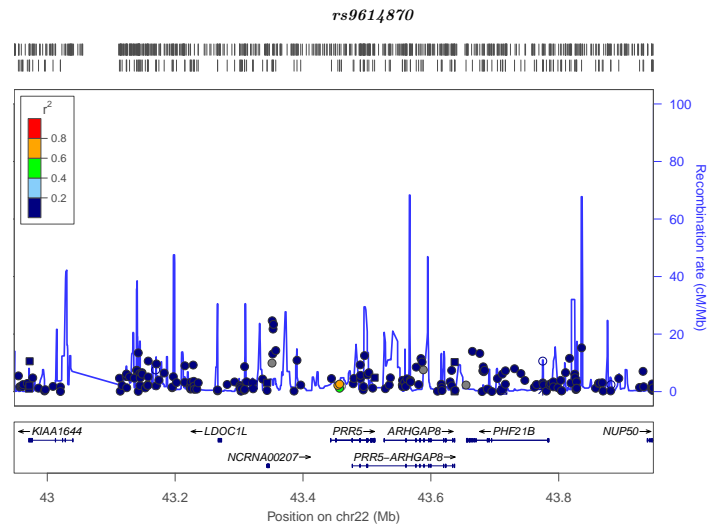


Figure A.211: HS\_PCT(combined)

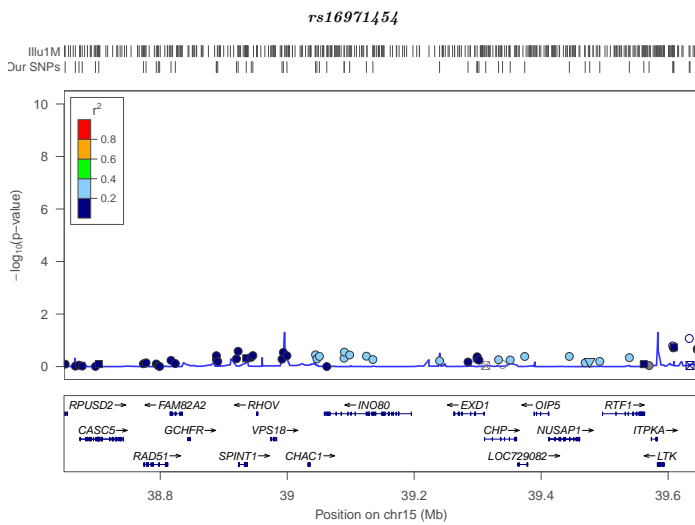


Figure A.212: HS\_PCT(combined)

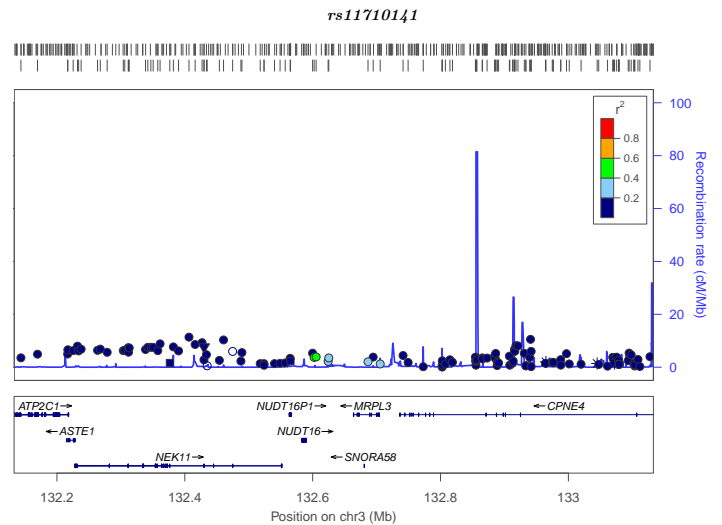


Figure A.213: HS\_PCT(combined)

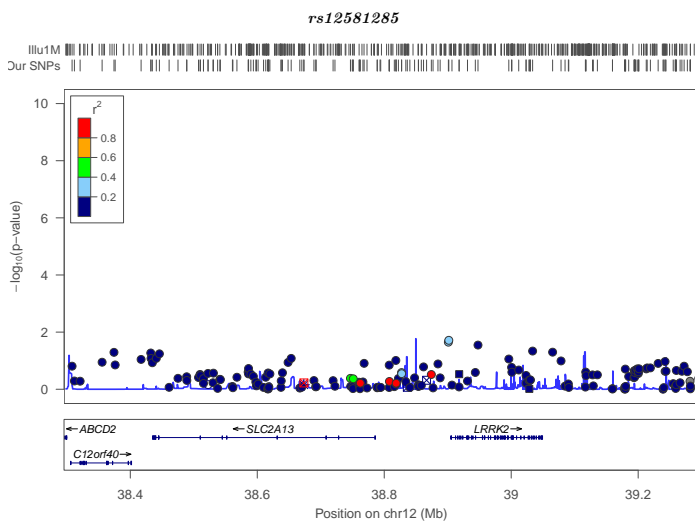


Figure A.214: HS\_PCT(combined)

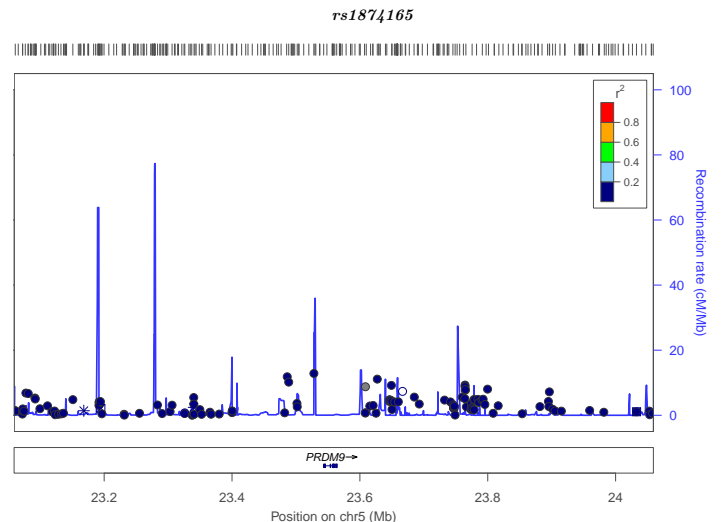


Figure A.215: HS\_PCT(combined)

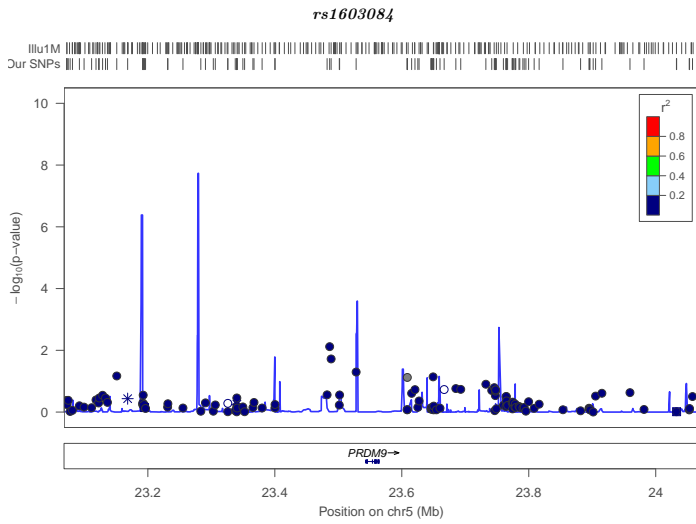


Figure A.216: HS\_PCT(female)

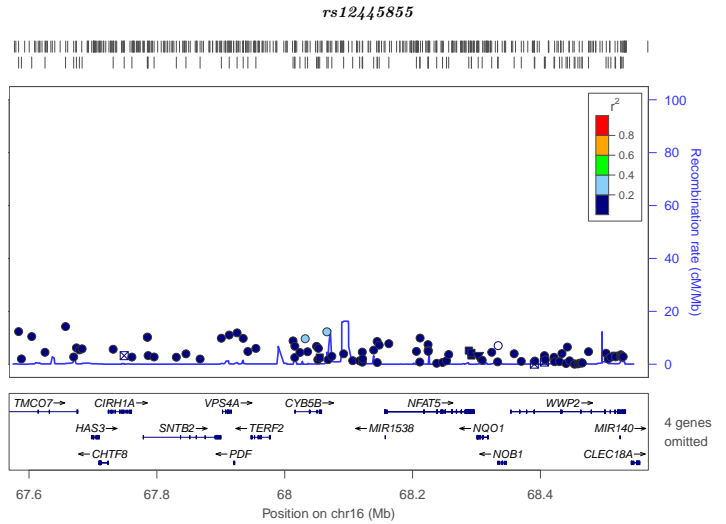


Figure A.217: HS\_PCT(female)

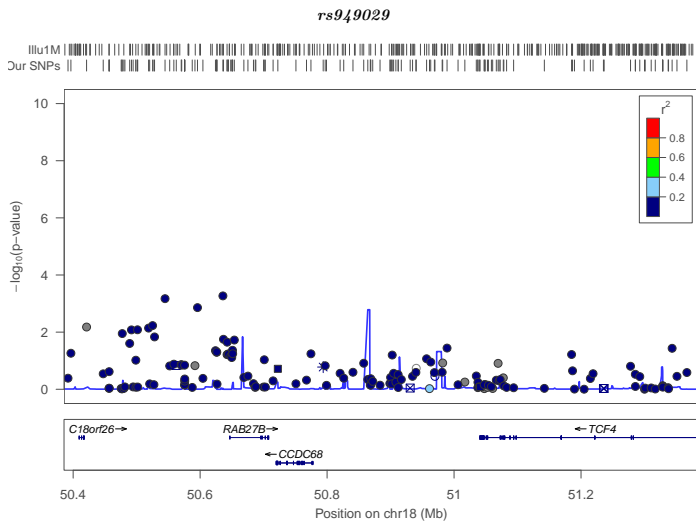


Figure A.218: HS\_PCT(female)

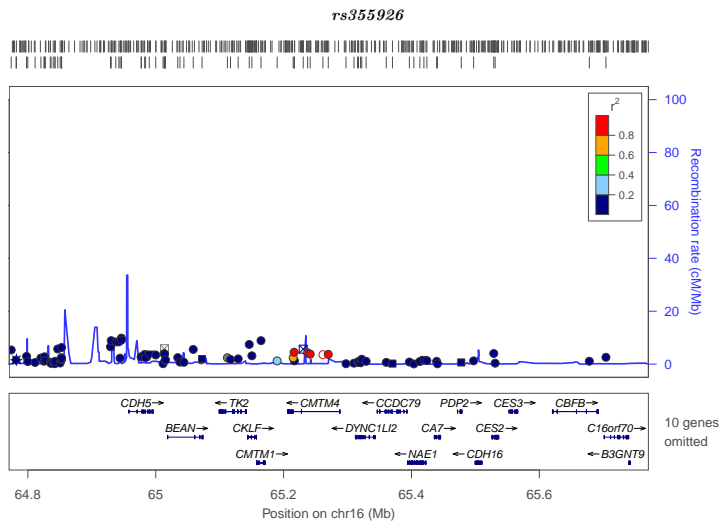


Figure A.219: HS\_PCT(female)

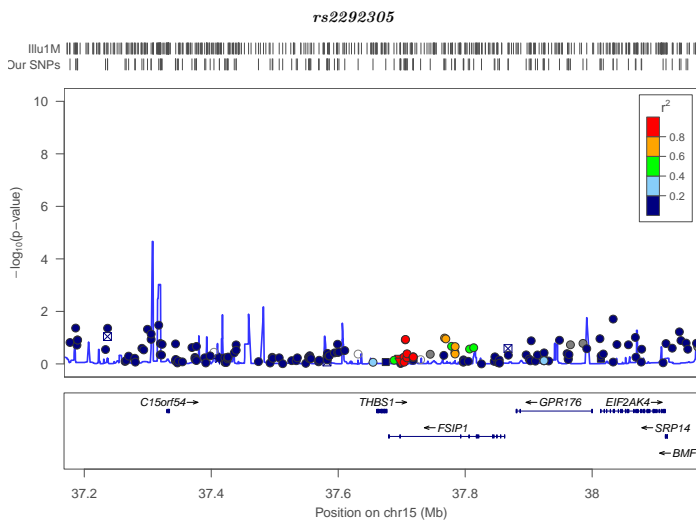


Figure A.220: HS\_PCT(female)

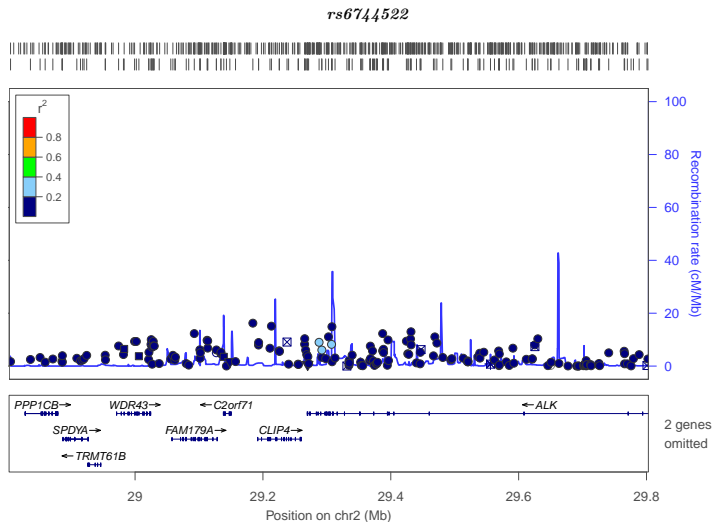


Figure A.221: HS\_PCT(female)

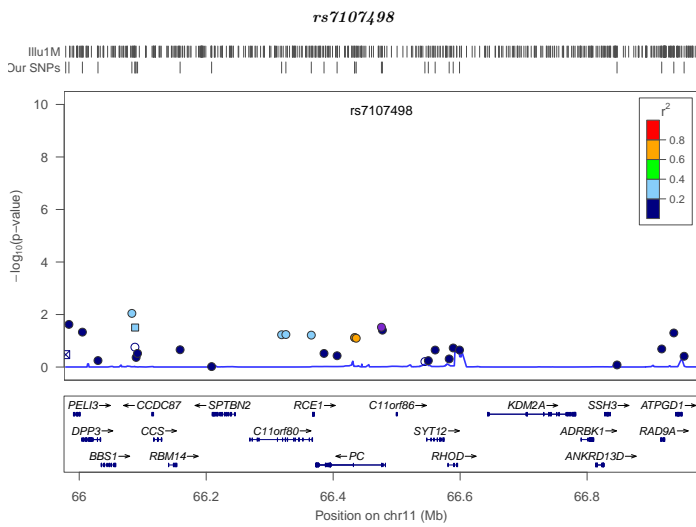


Figure A.222: HS\_PCT(female)

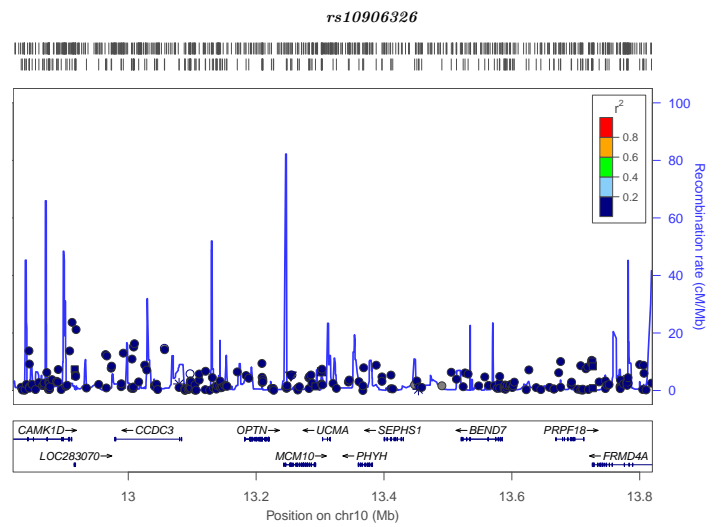


Figure A.223: HS\_PCT(female)

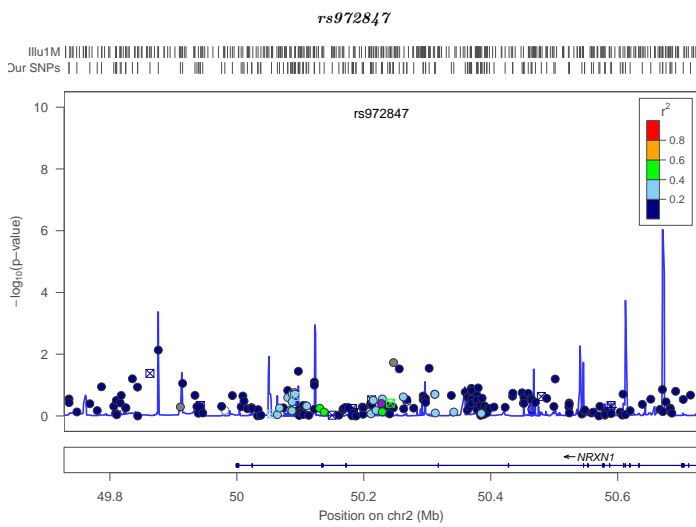


Figure A.224: HS\_PCT(female)

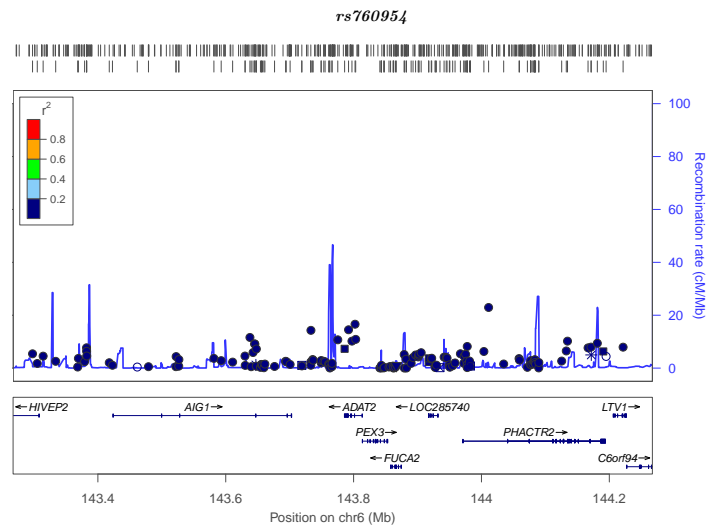


Figure A.225: HS\_PCT(female)

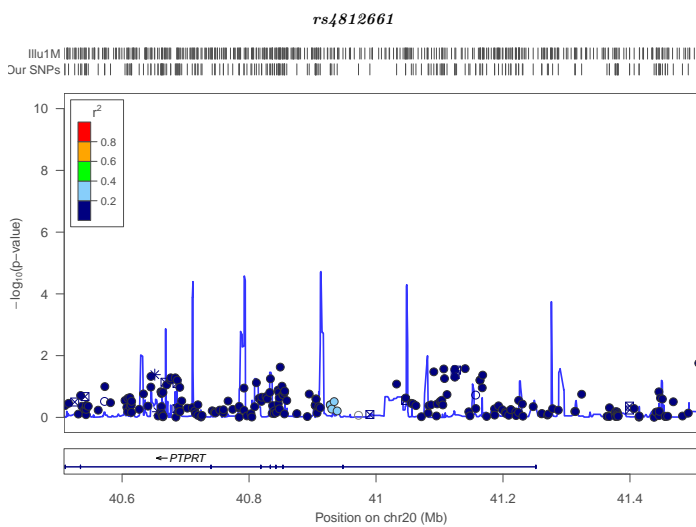


Figure A.226: HS\_PCT(female)

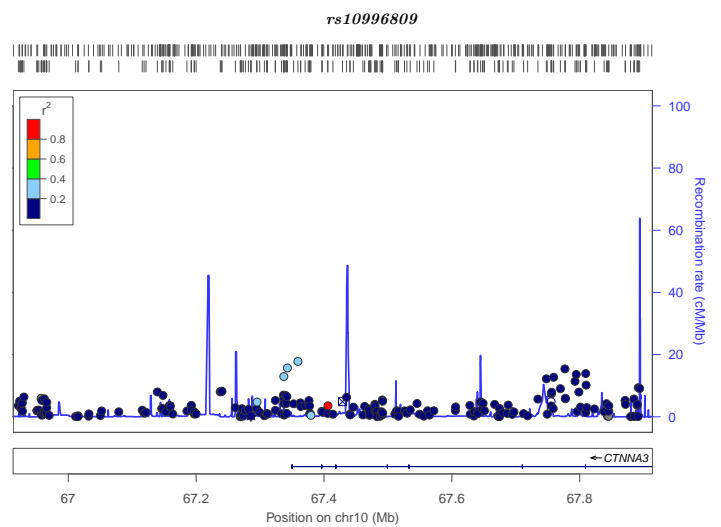


Figure A.227: HS\_PCT(male)

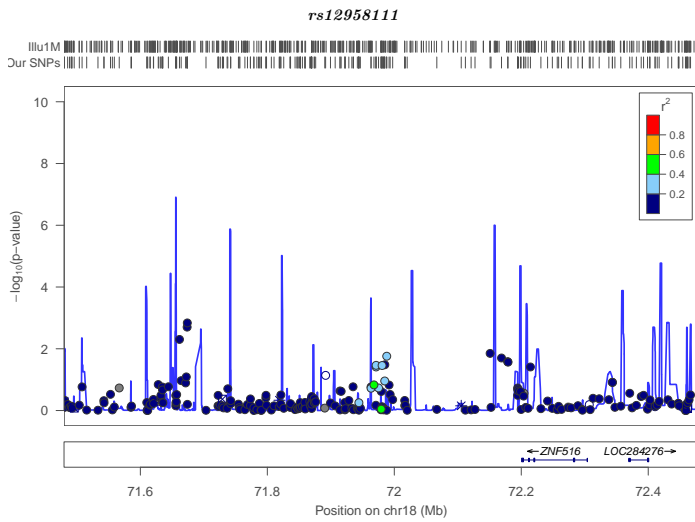


Figure A.228: HS\_PCT(male)

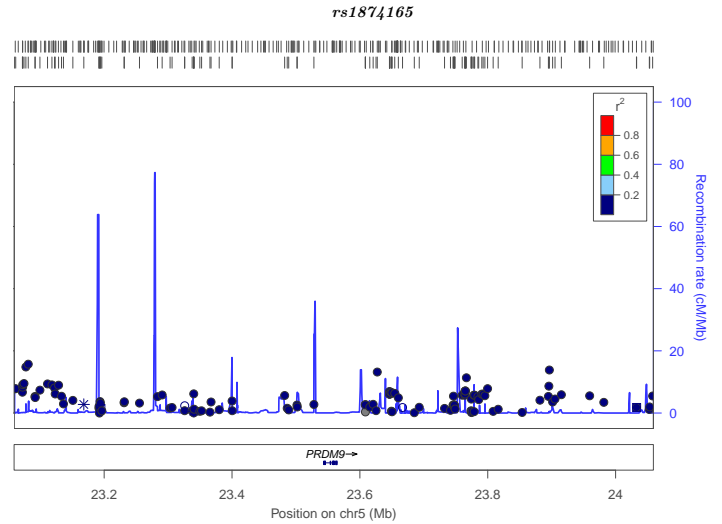


Figure A.229: HS\_PCT(male)

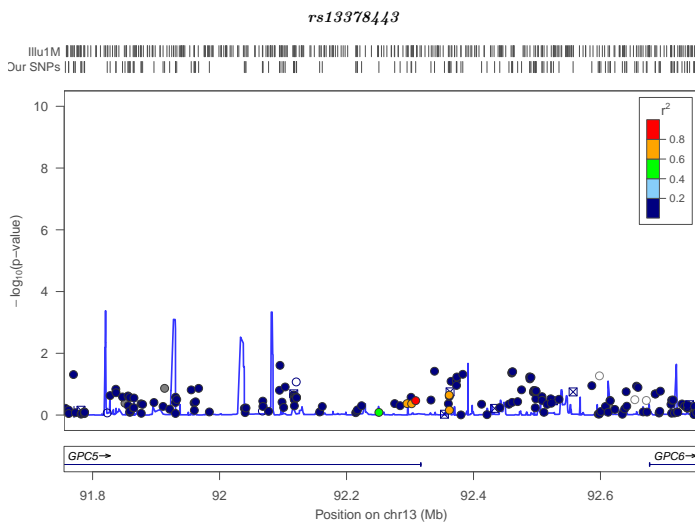


Figure A.230: HS\_PCT(male)

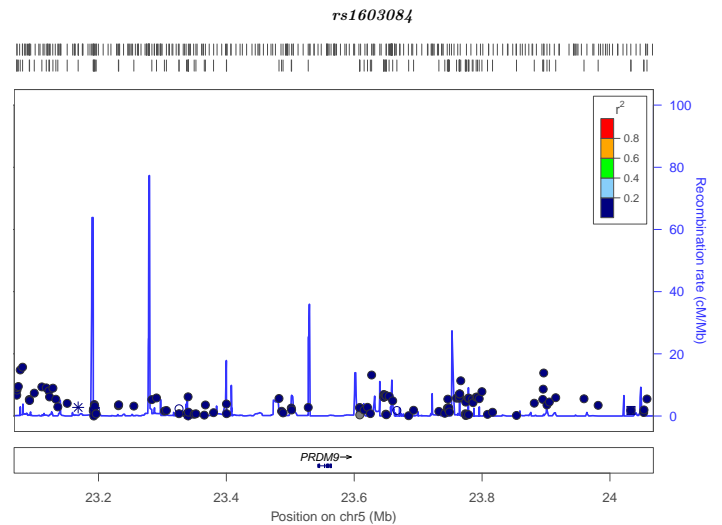


Figure A.231: HS\_PCT(male)

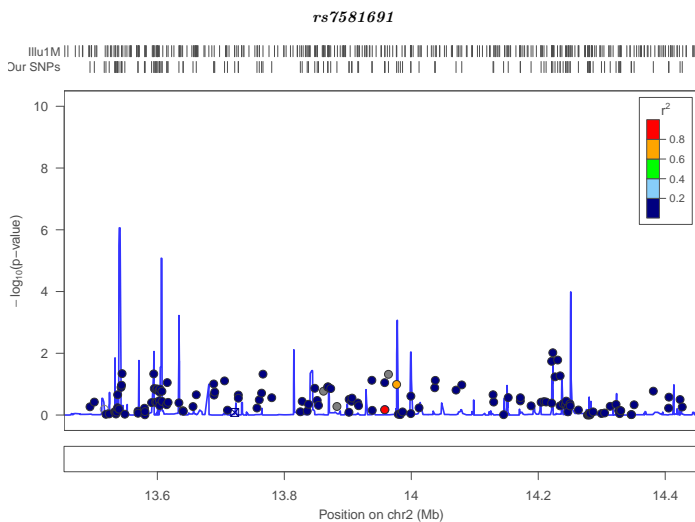


Figure A.232: HS\_PCT(male)

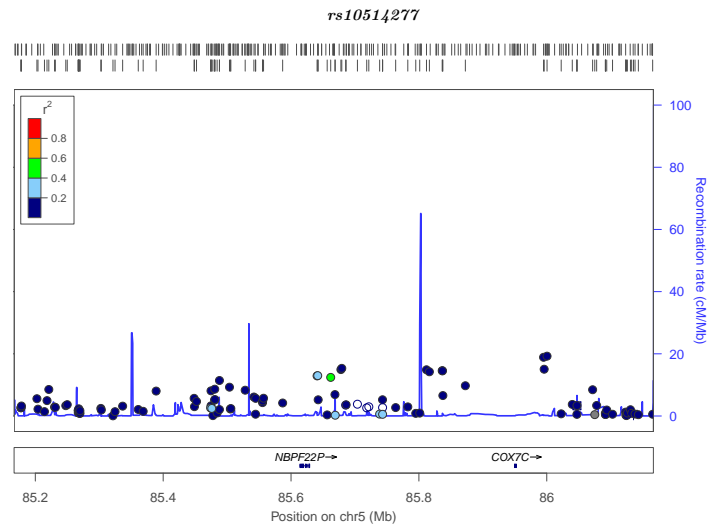
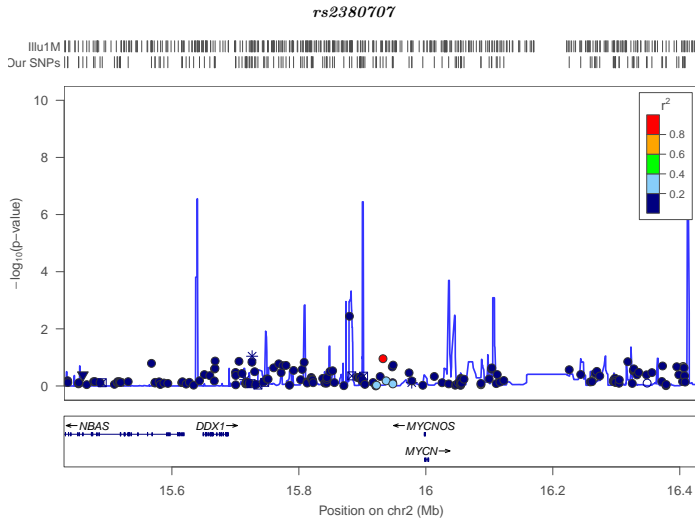
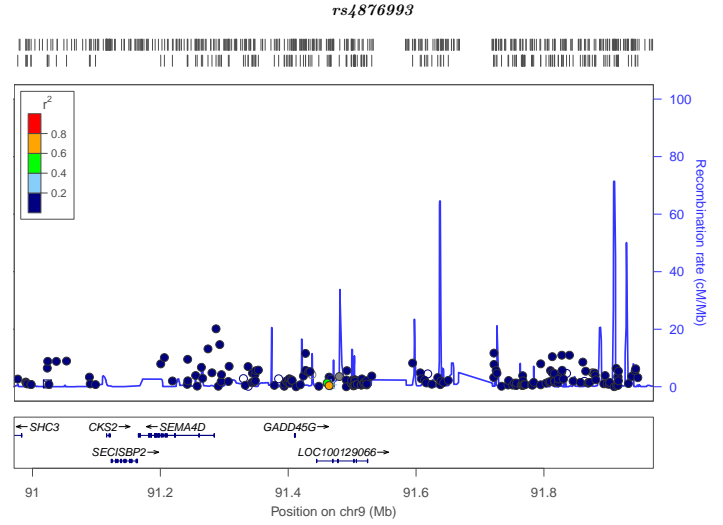


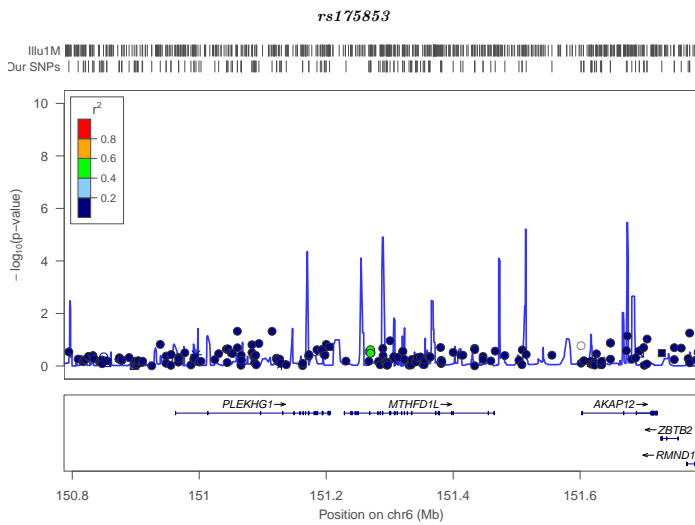
Figure A.233: HS\_PCT(male)



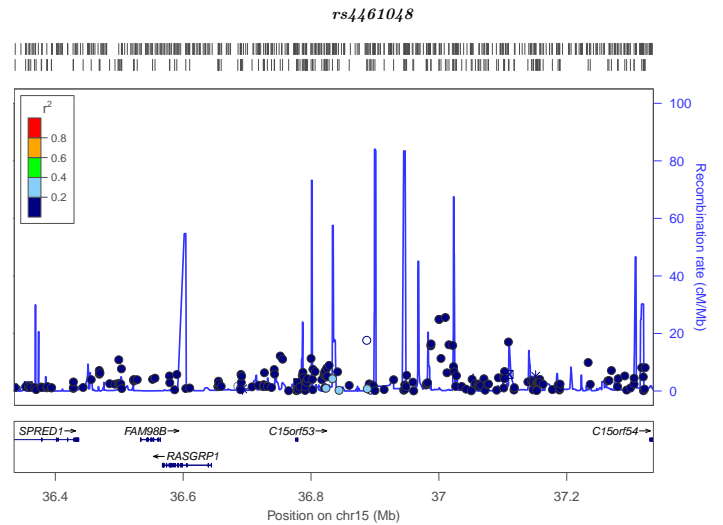
**Figure A.234: HS\_PCT(male)**



**Figure A.235: HS\_PCT(male)**



**Figure A.236: HS\_PCT(male)**



**Figure A.237: HS\_PCT(male)**



### A.3.3 Phenotype: HS\_CNT

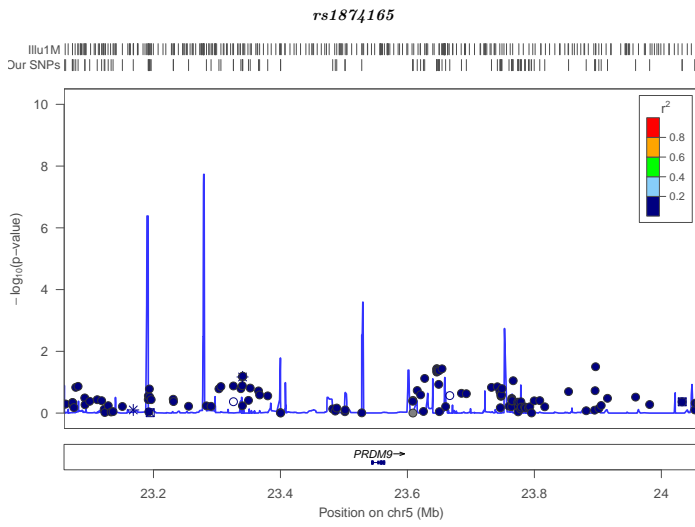


Figure A.238: HS\_CNT(combined)

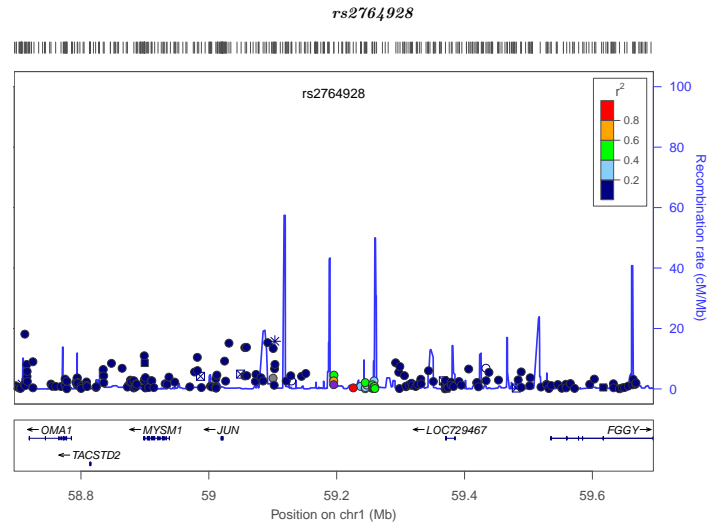


Figure A.239: HS\_CNT(combined)

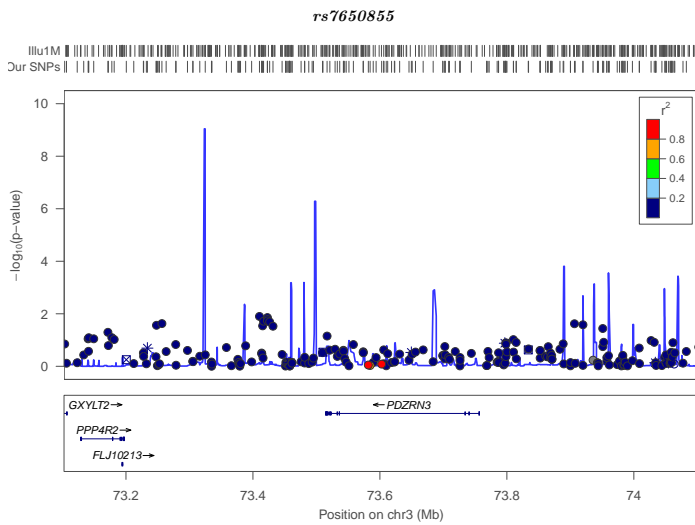


Figure A.240: HS\_CNT(combined)

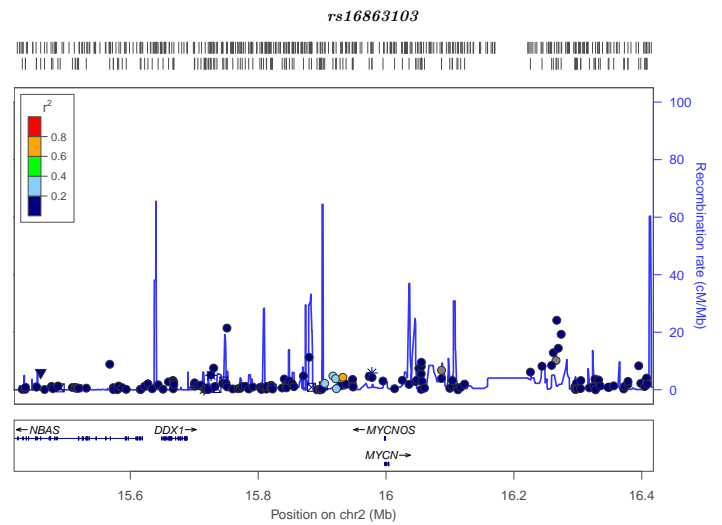


Figure A.241: HS\_CNT(combined)

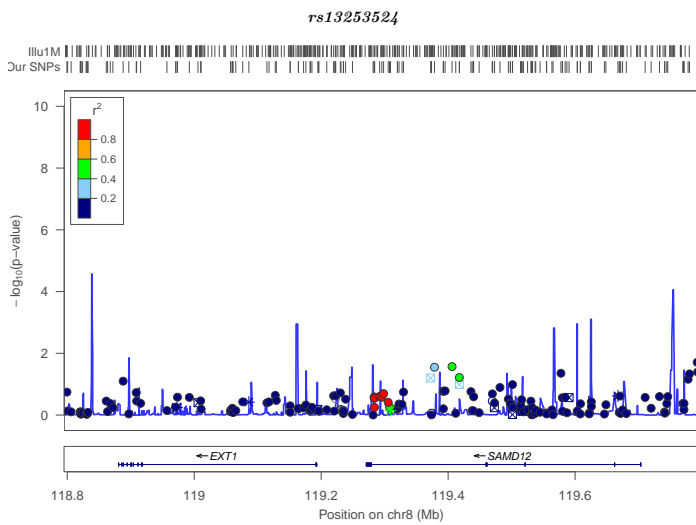


Figure A.242: HS\_CNT(combined)

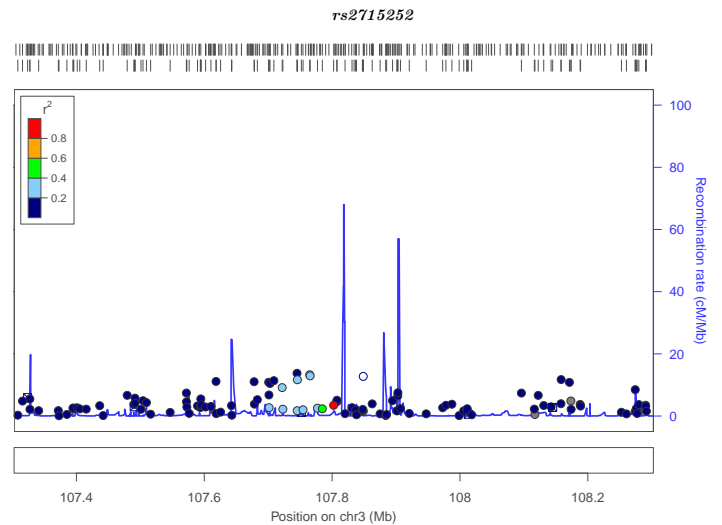


Figure A.243: HS\_CNT(combined)

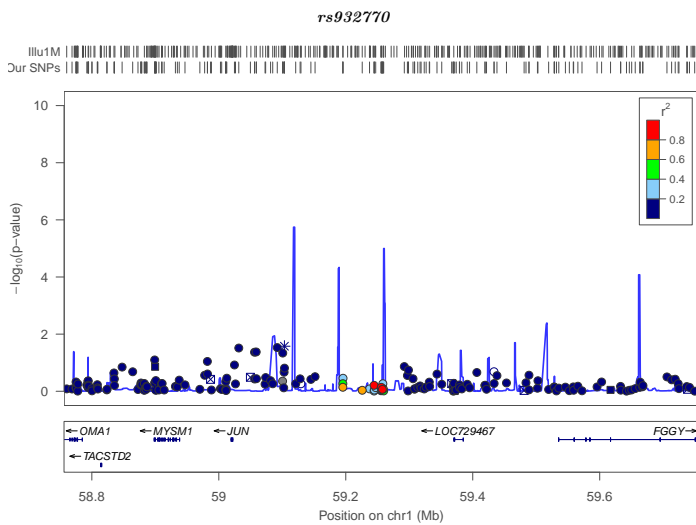


Figure A.244: HS\_CNT(combined)

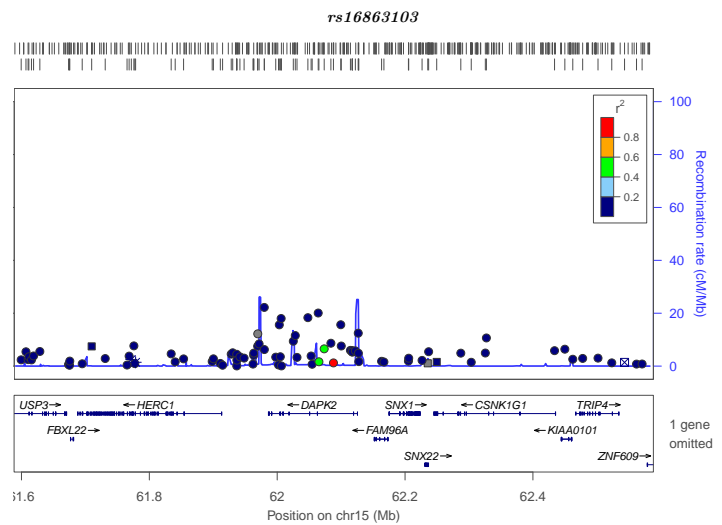


Figure A.245: HS\_CNT(combined)

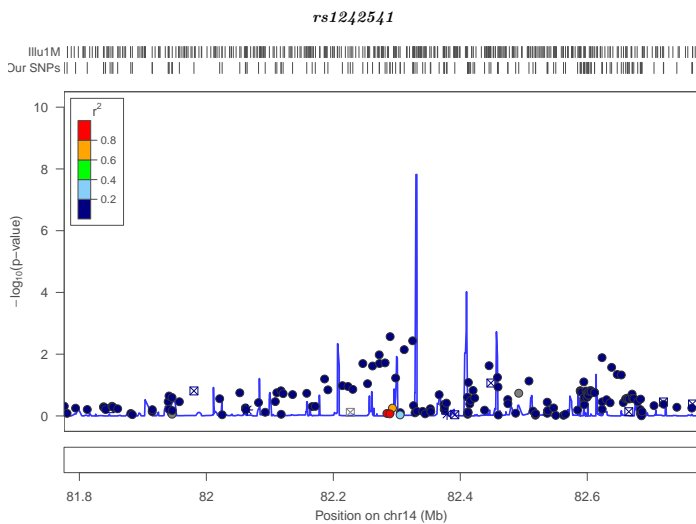


Figure A.246: HS\_CNT(female)

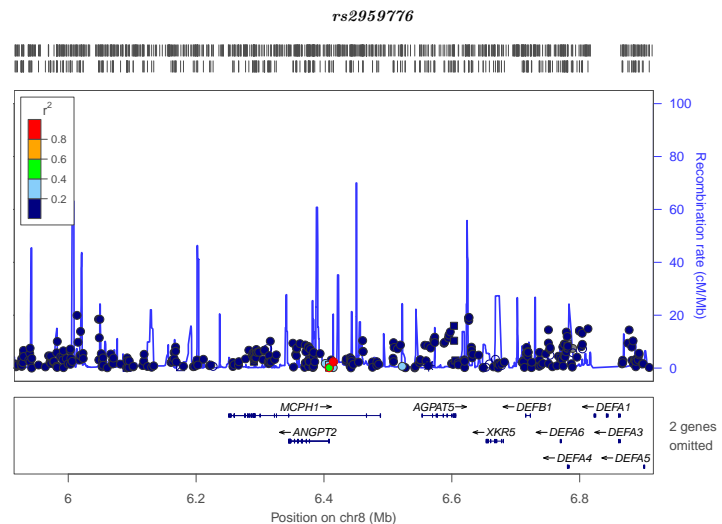


Figure A.247: HS\_CNT(female)

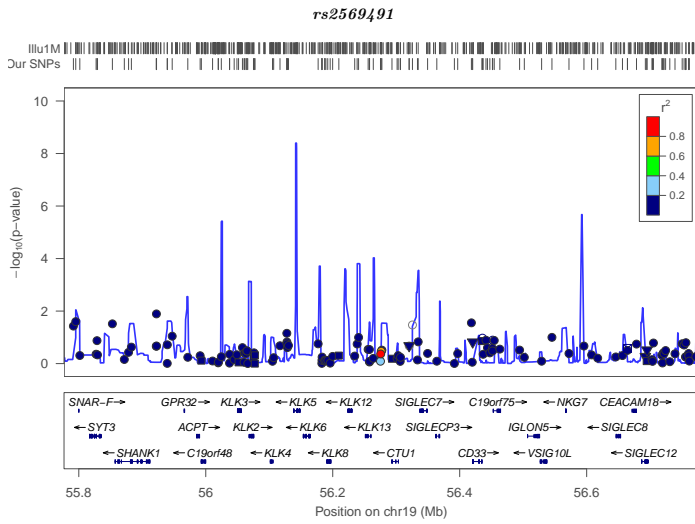


Figure A.248: HS\_CNT(female)

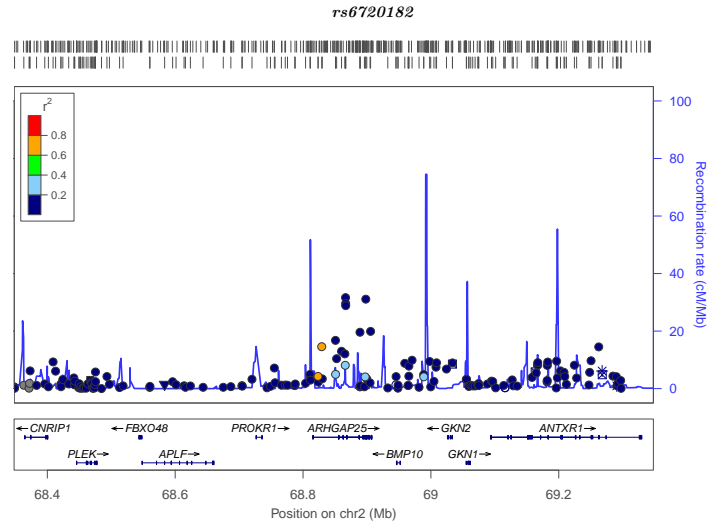


Figure A.249: HS\_CNT(female)

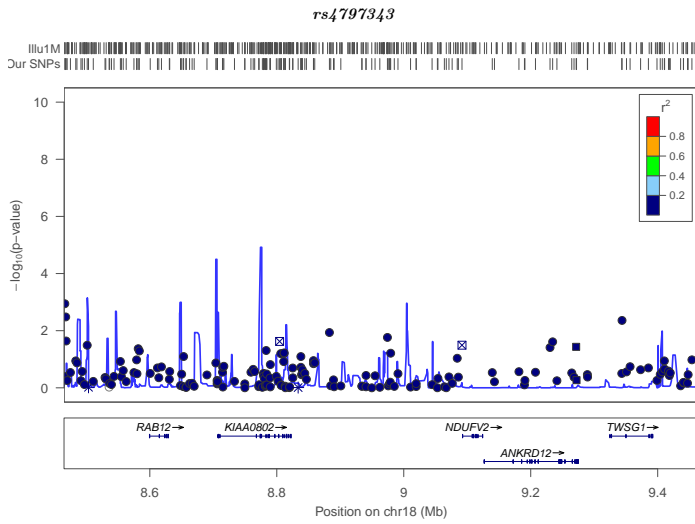


Figure A.250: HS\_CNT(female)

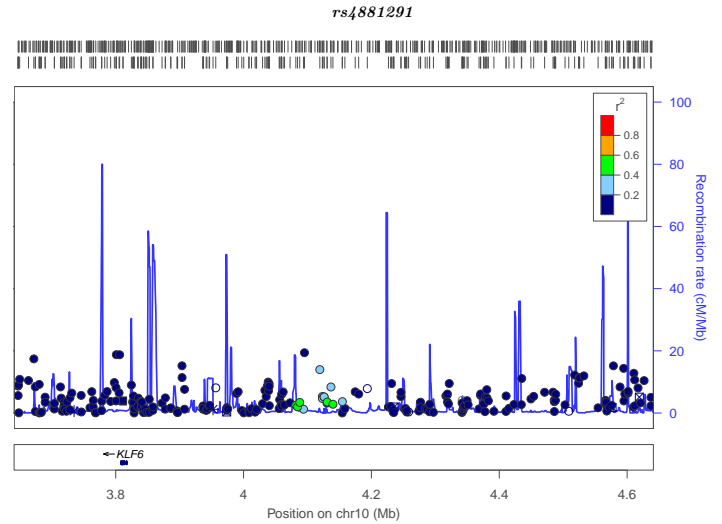


Figure A.251: HS\_CNT(female)

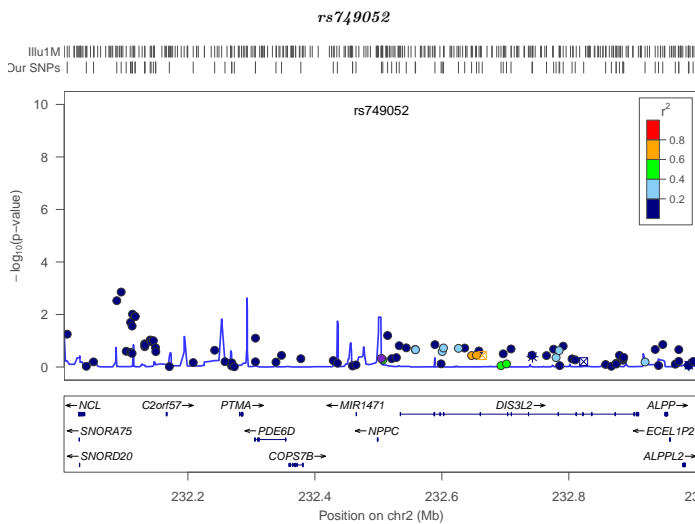


Figure A.252: HS\_CNT(female)

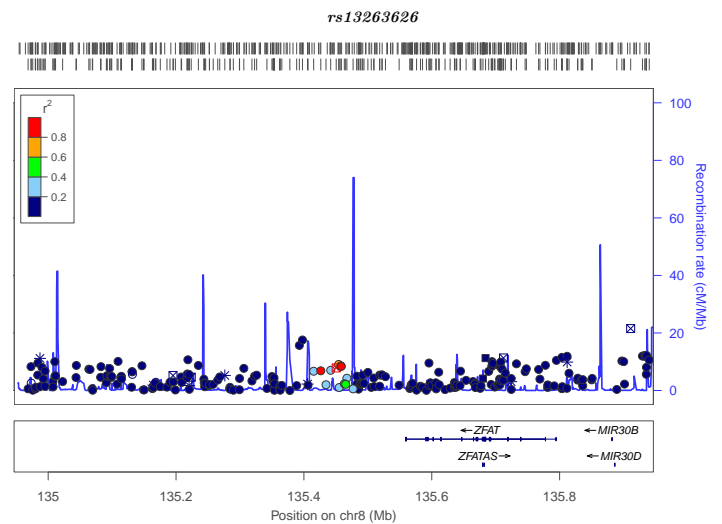


Figure A.253: HS\_CNT(female)

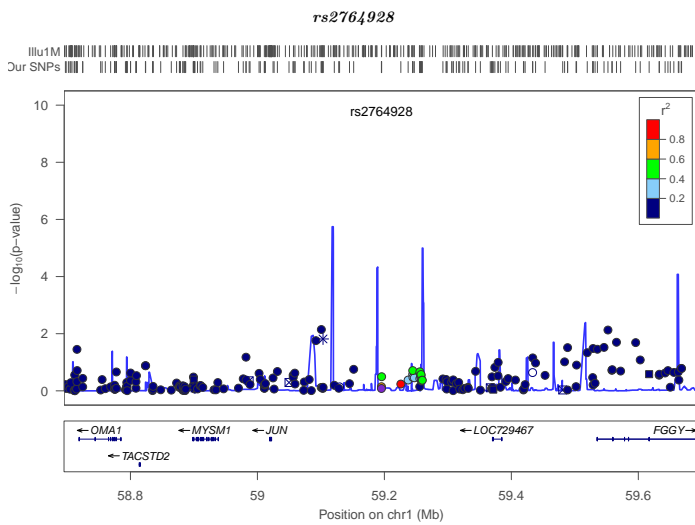


Figure A.254: HS\_CNT(female)

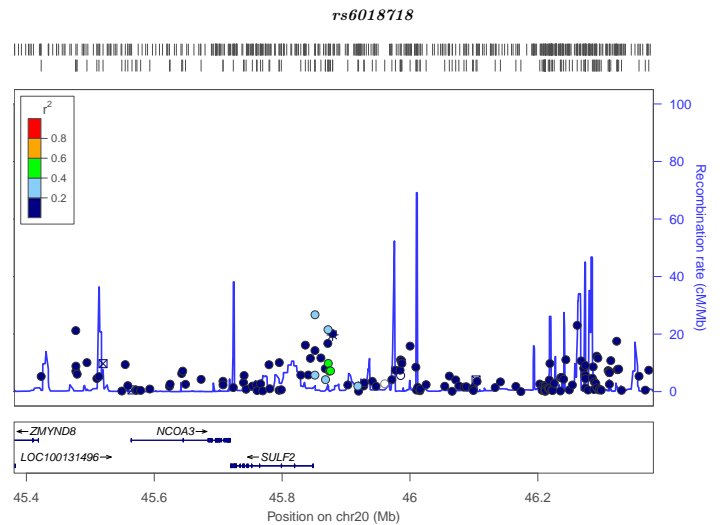


Figure A.255: HS\_CNT(female)

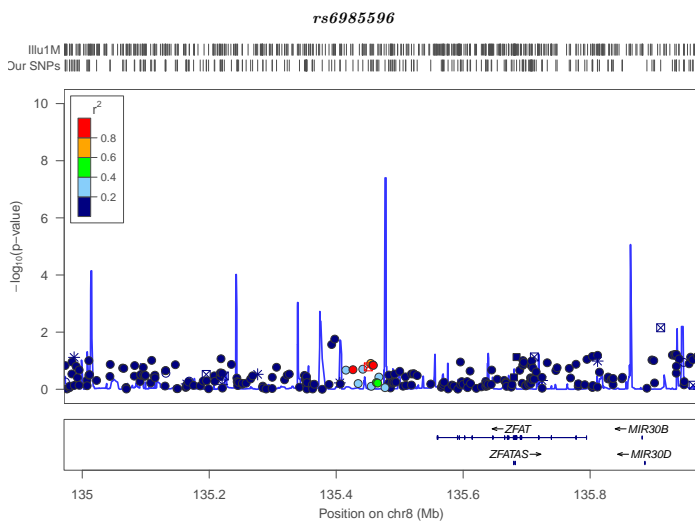


Figure A.256: HS\_CNT(female)

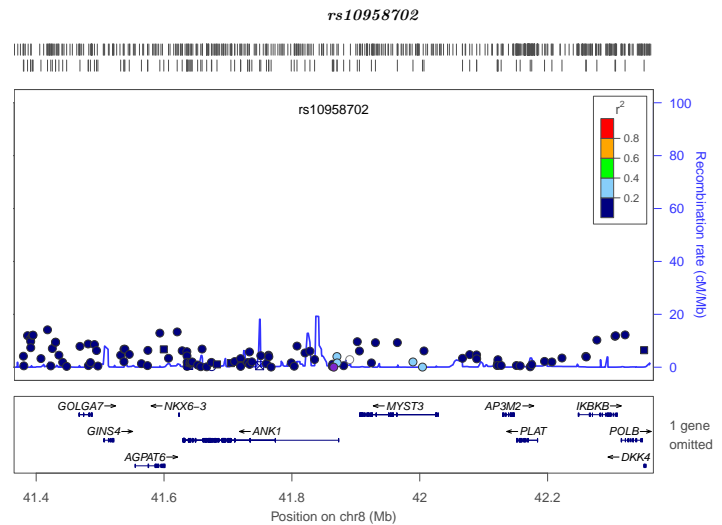


Figure A.257: HS\_CNT(male)

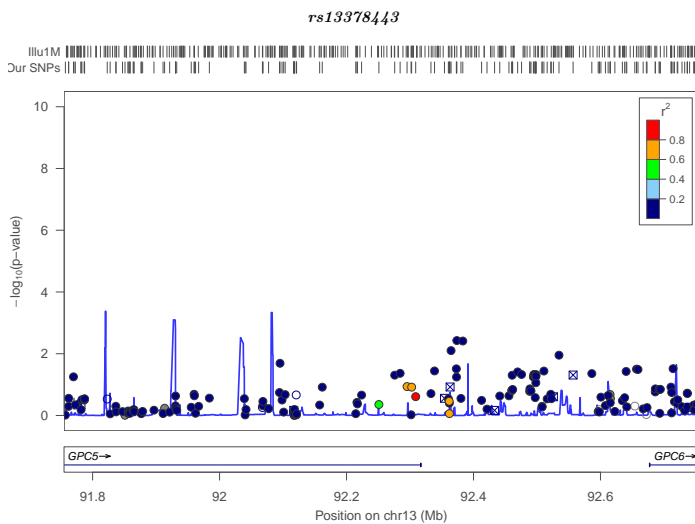


Figure A.258: HS\_CNT(male)

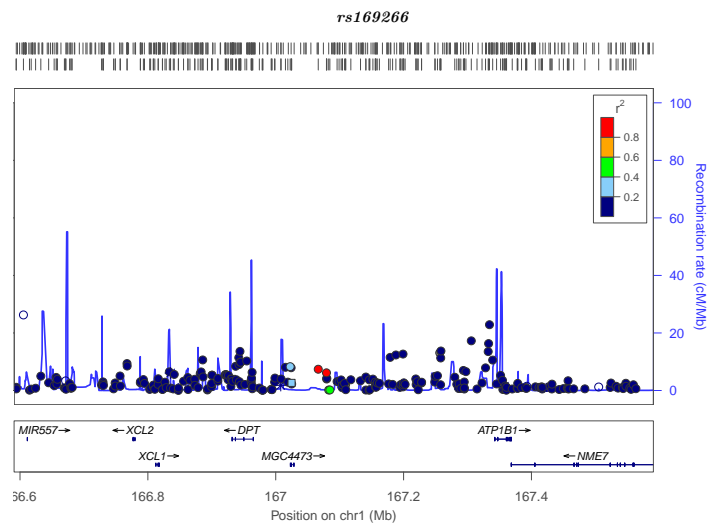


Figure A.259: HS\_CNT(male)

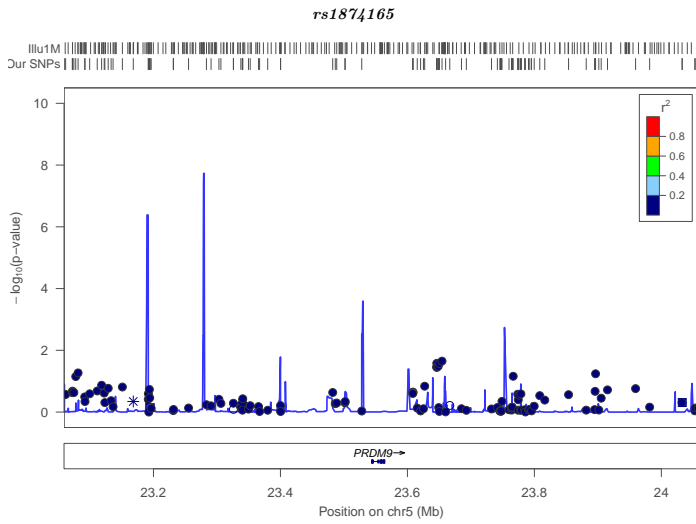


Figure A.260: HS\_CNT(male)

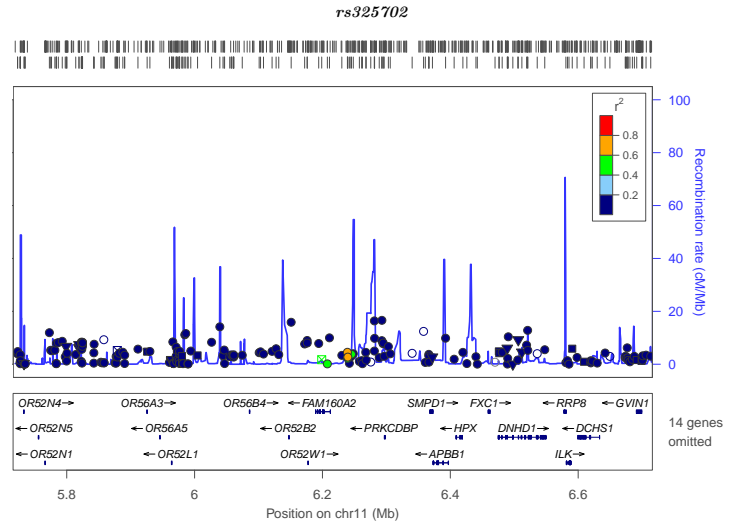


Figure A.261: HS\_CNT(male)

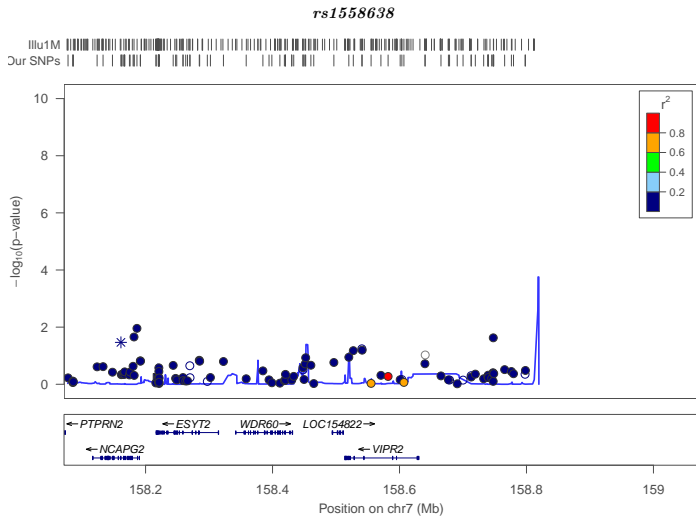


Figure A.262: HS\_CNT(male)

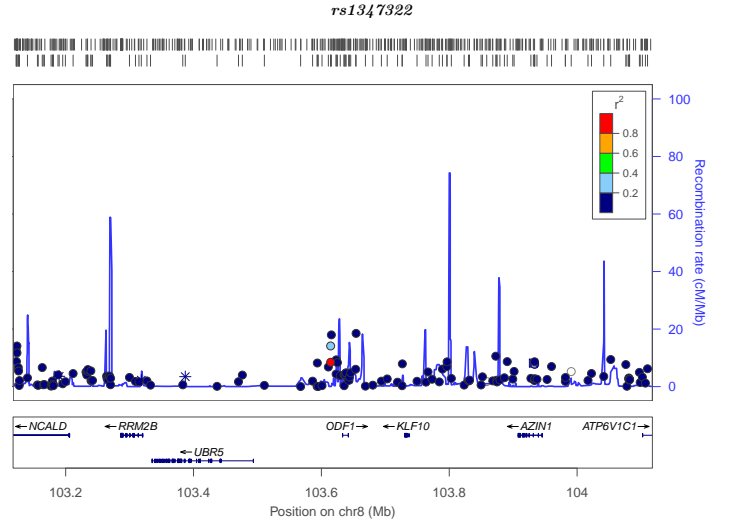


Figure A.263: HS\_CNT(male)

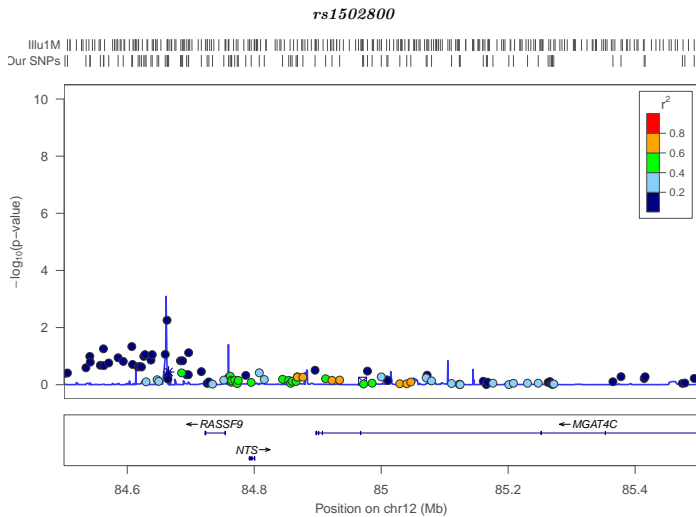


Figure A.264: HS\_CNT(male)

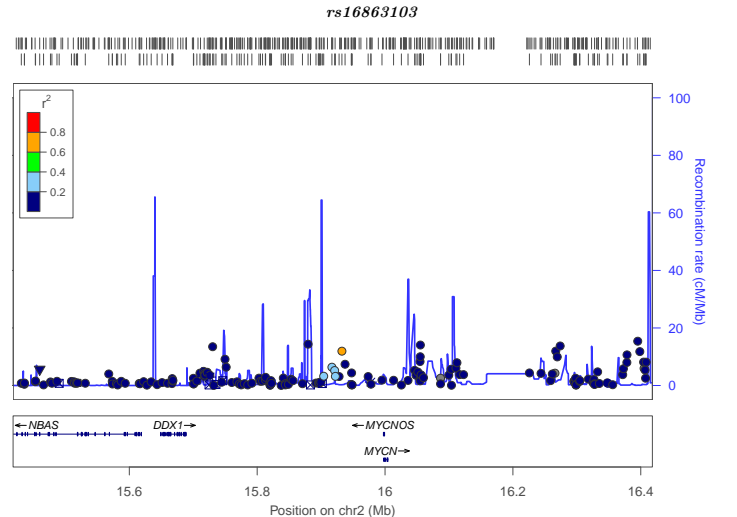


Figure A.265: HS\_CNT(male)

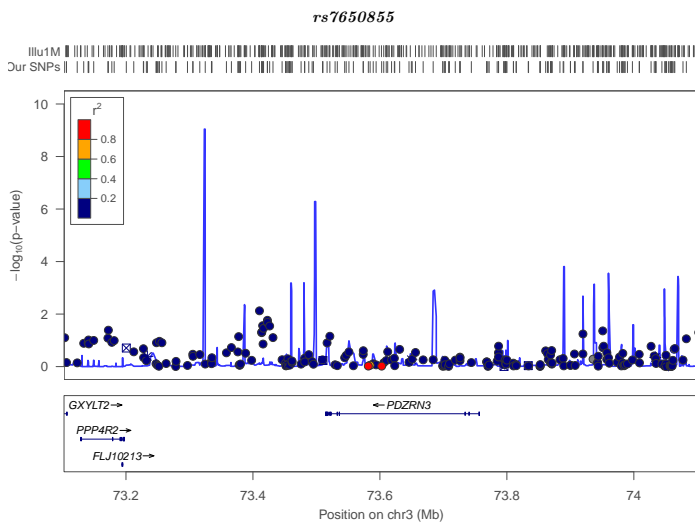


Figure A.266: HS\_CNT(male)

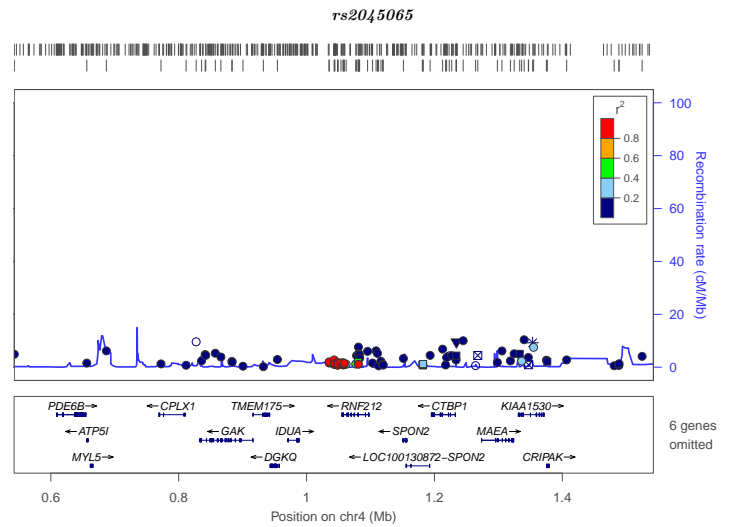


Figure A.267: HS\_CNT(male)

### A.3.4 Phenotype: NHS\_CNT

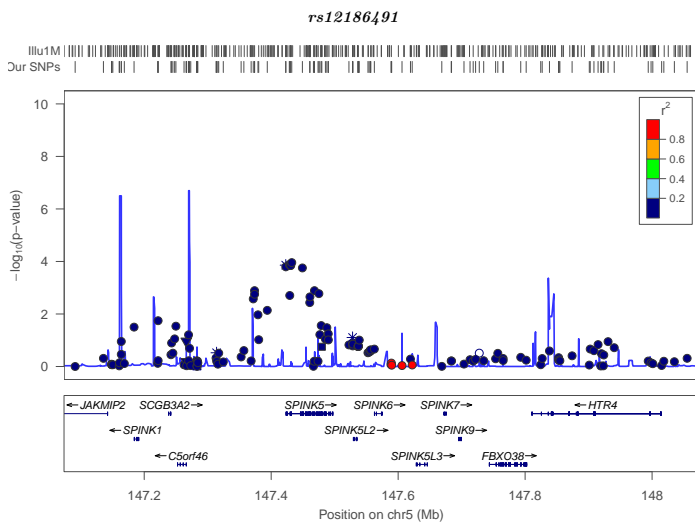


Figure A.268: NHS\_CNT(combined)

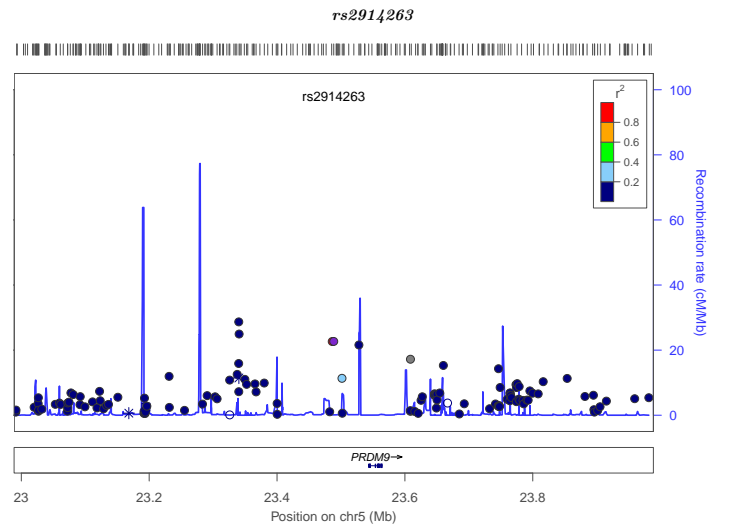


Figure A.269: NHS\_CNT(combined)

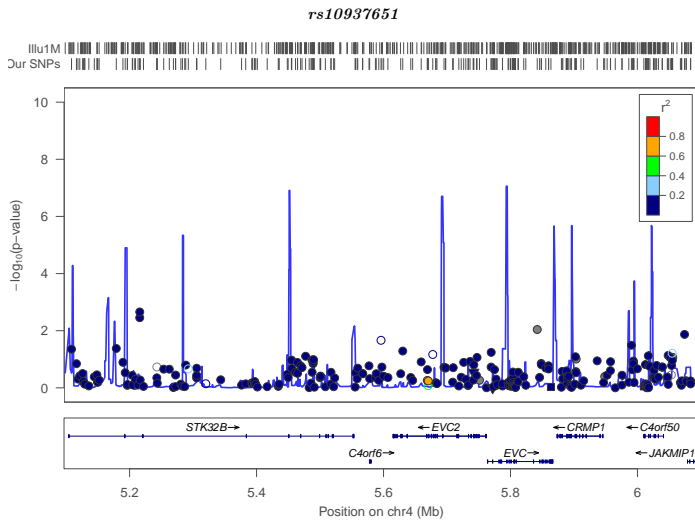


Figure A.270: NHS\_CNT(combined)

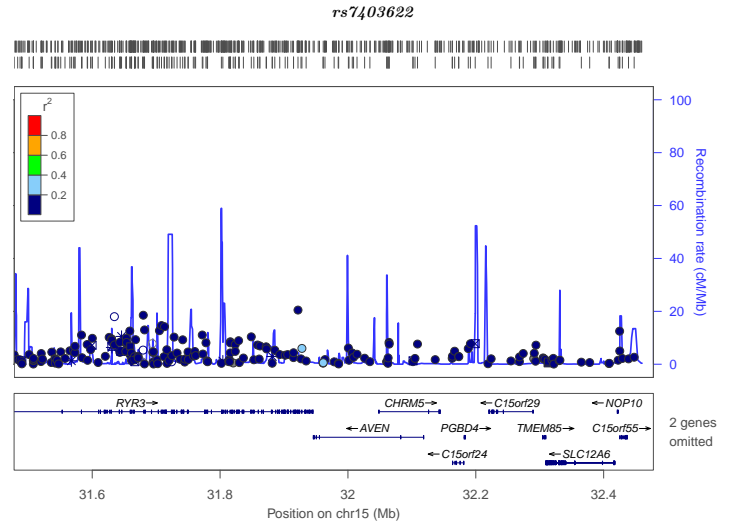


Figure A.271: NHS\_CNT(combined)

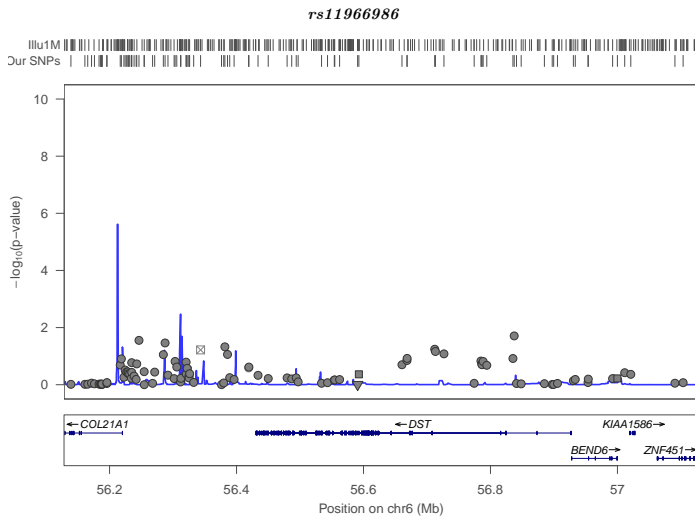


Figure A.272: NHS\_CNT(combined)

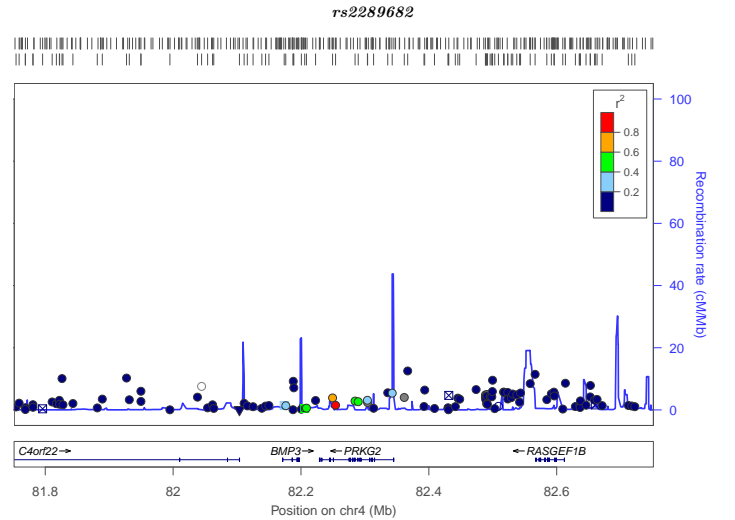


Figure A.273: NHS\_CNT(combined)

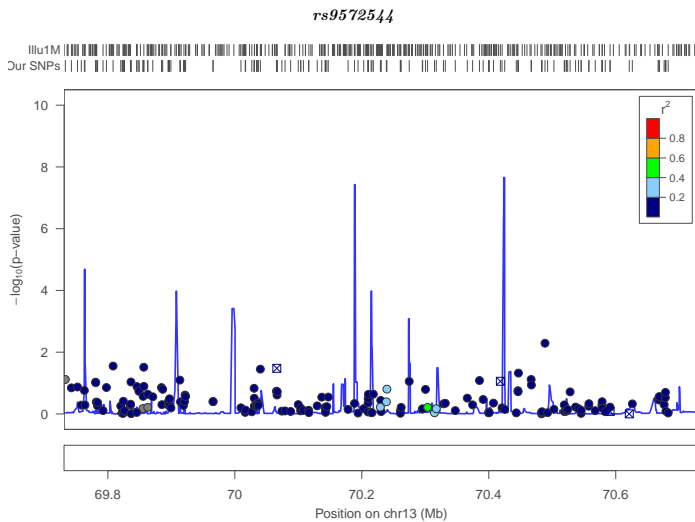


Figure A.274: NHS\_CNT(combined)

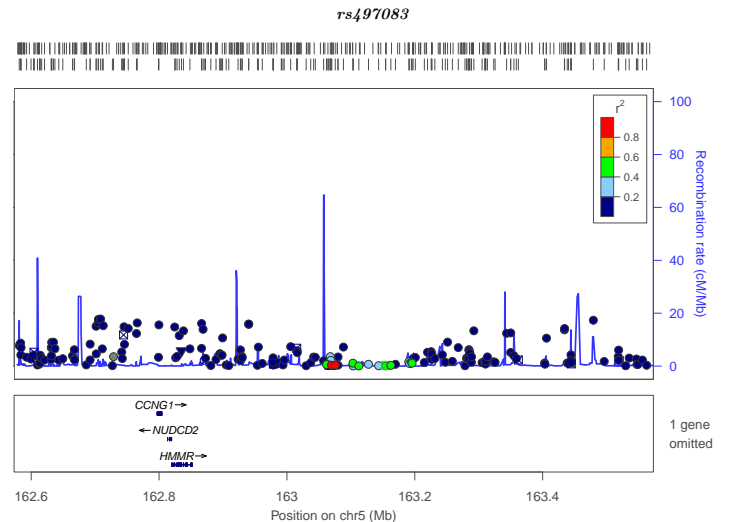


Figure A.275: NHS\_CNT(combined)

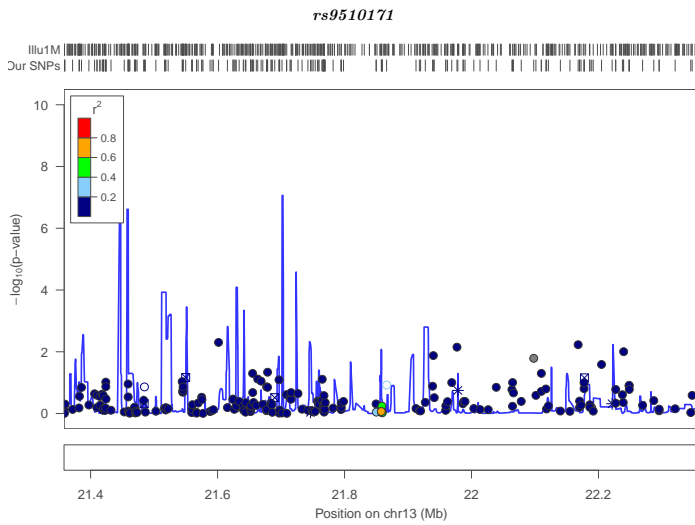


Figure A.276: NHS\_CNT(combined)

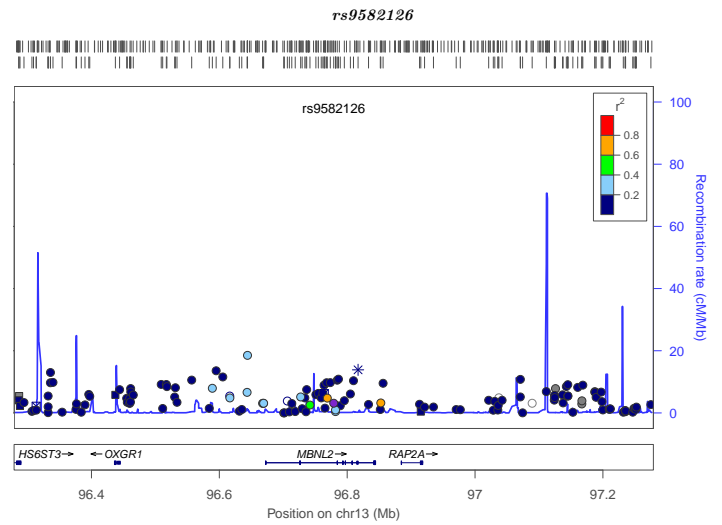


Figure A.277: NHS\_CNT(combined)

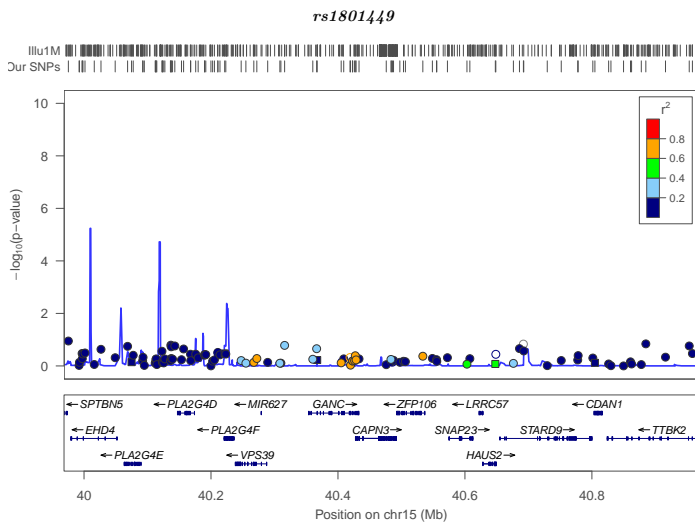


Figure A.278: NHS\_CNT(combined)

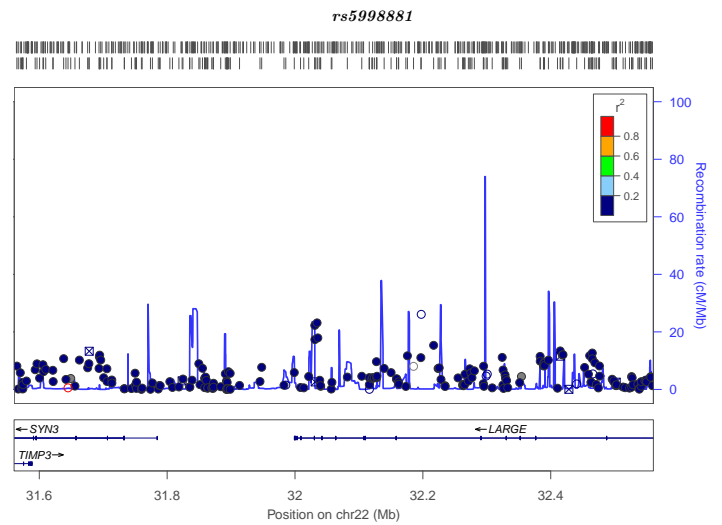


Figure A.279: NHS\_CNT(combined)

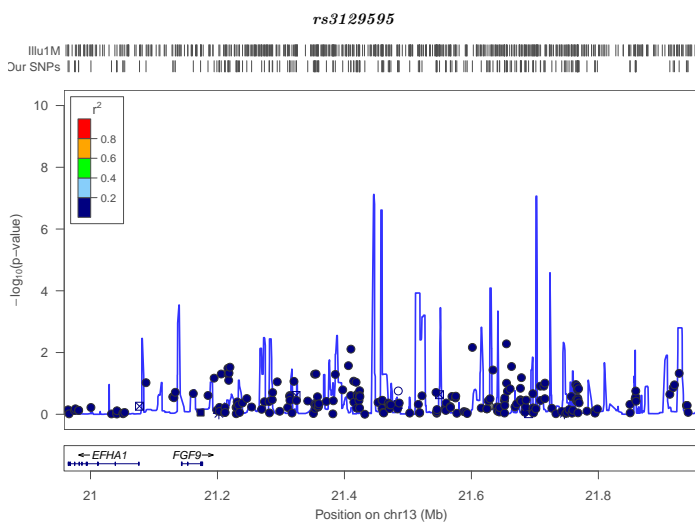


Figure A.280: NHS\_CNT(female)

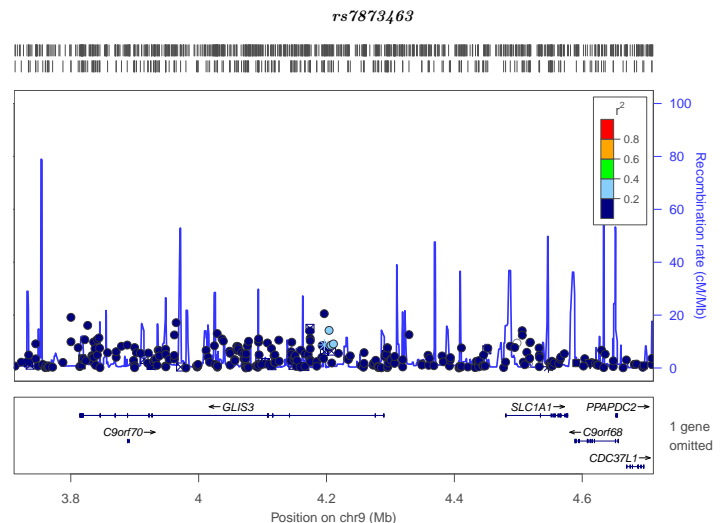


Figure A.281: NHS\_CNT(female)



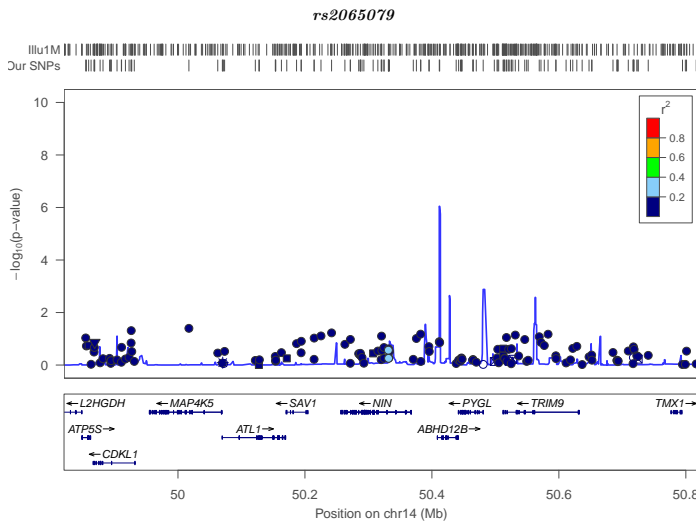


Figure A.282: NHS\_CNT(female)

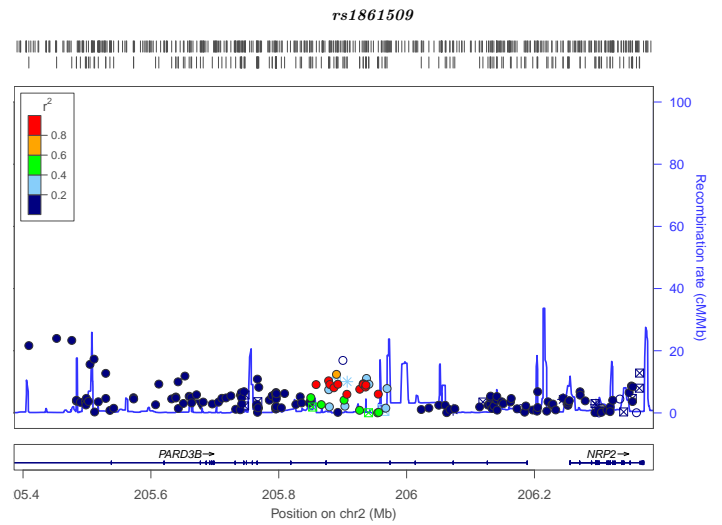


Figure A.283: NHS\_CNT(female)

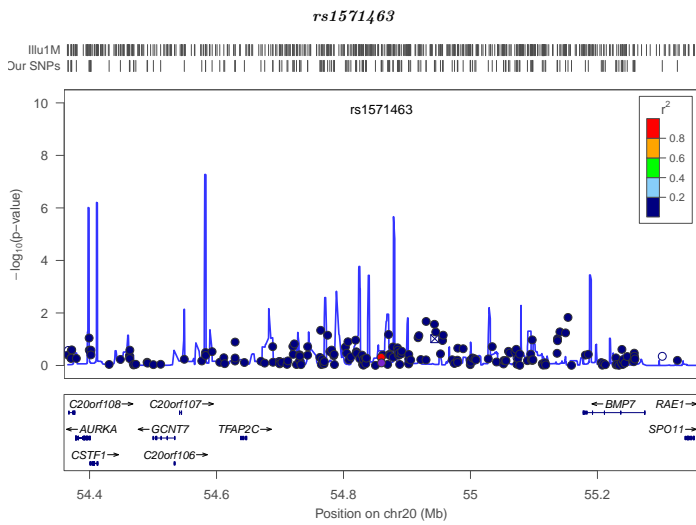


Figure A.284: NHS\_CNT(female)

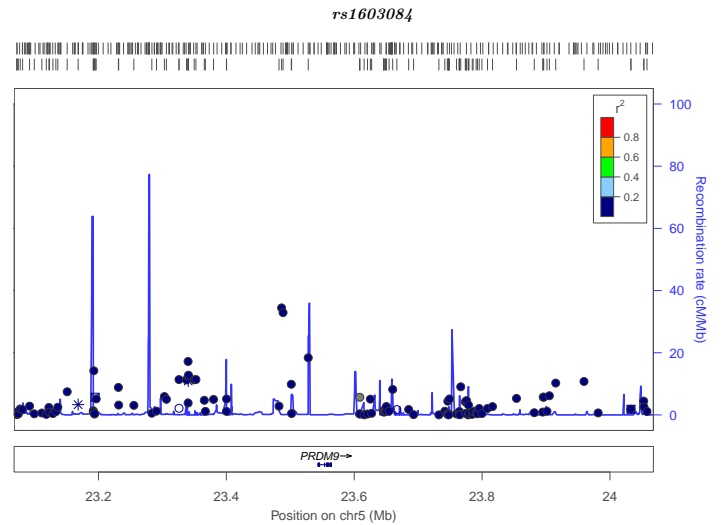


Figure A.285: NHS\_CNT(female)

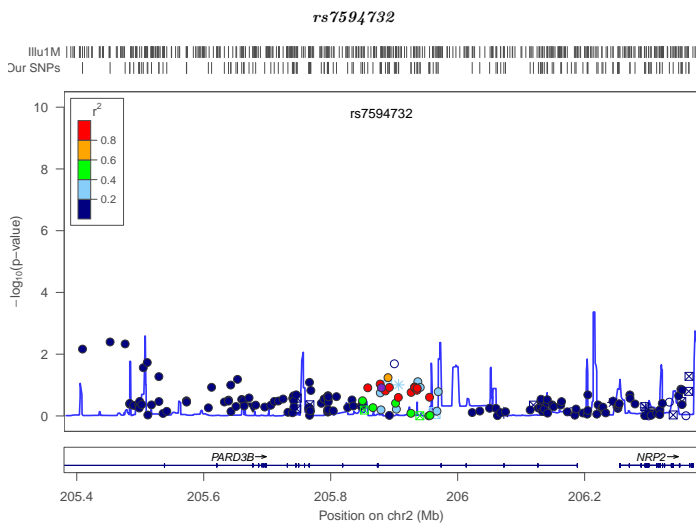


Figure A.286: NHS\_CNT(female)

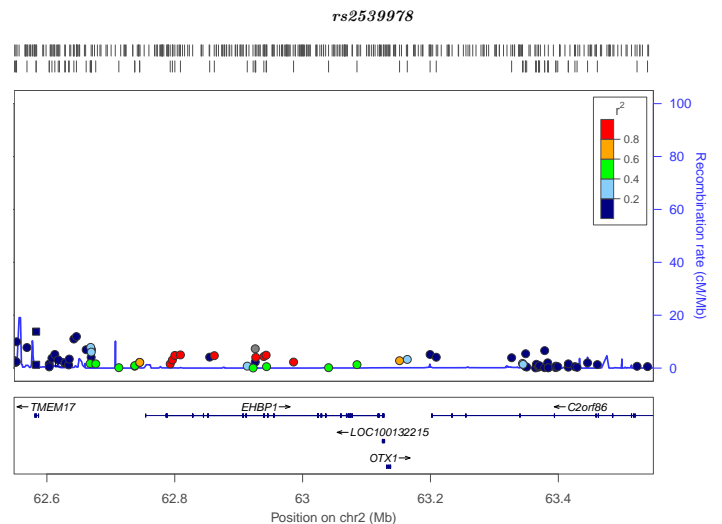


Figure A.287: NHS\_CNT(female)

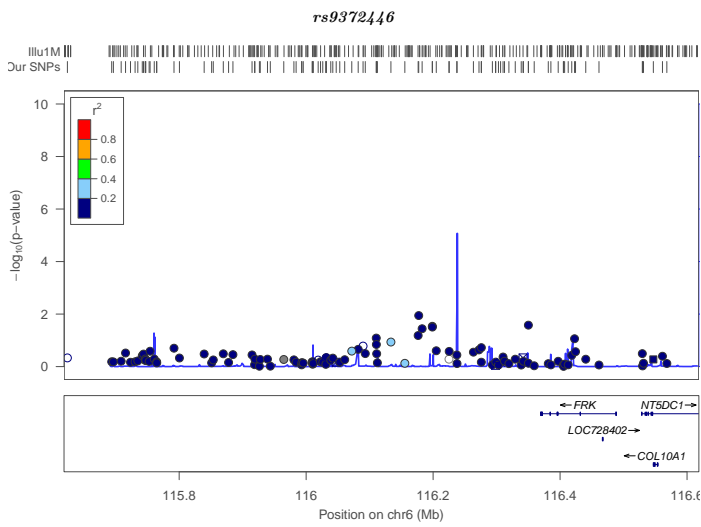


Figure A.288: NHS\_CNT(female)

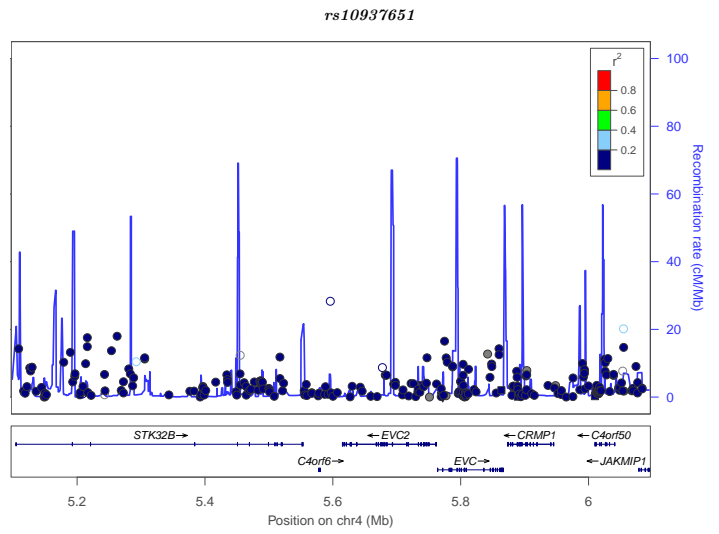


Figure A.289: NHS\_CNT(male)

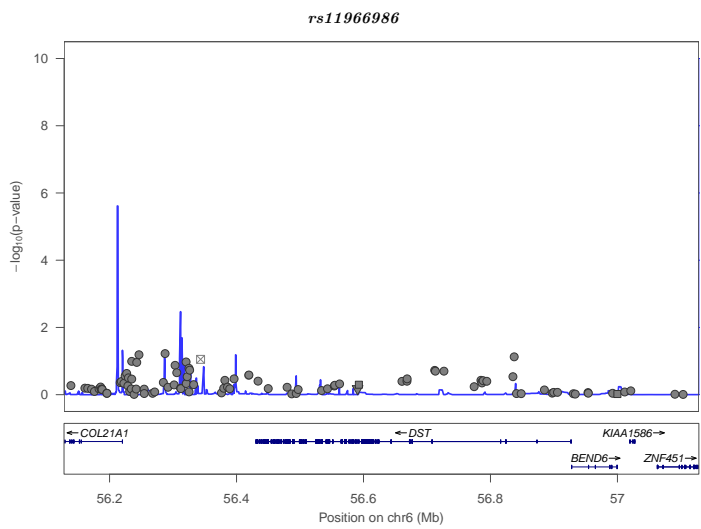


Figure A.290: NHS\_CNT(male)

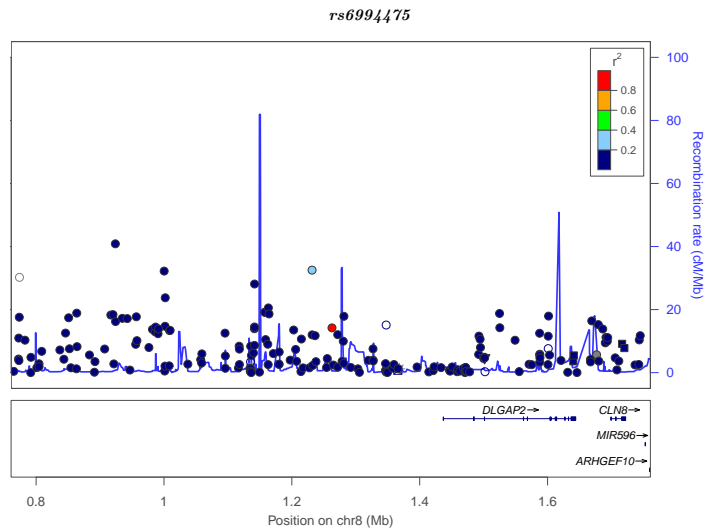


Figure A.291: NHS\_CNT(male)

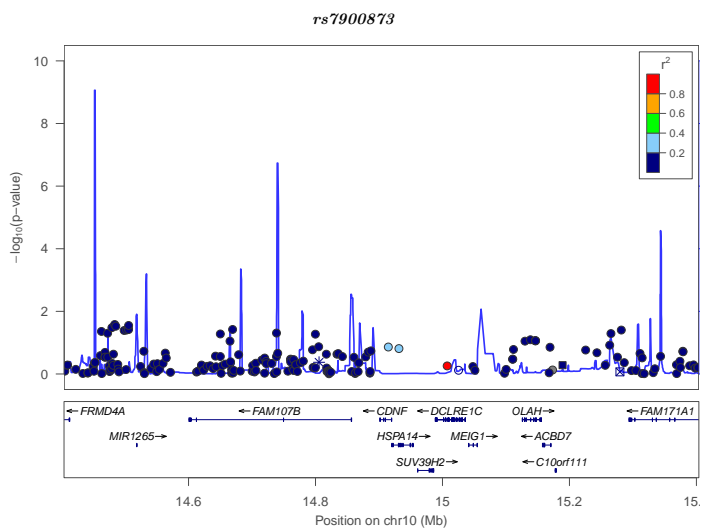


Figure A.292: NHS\_CNT(male)

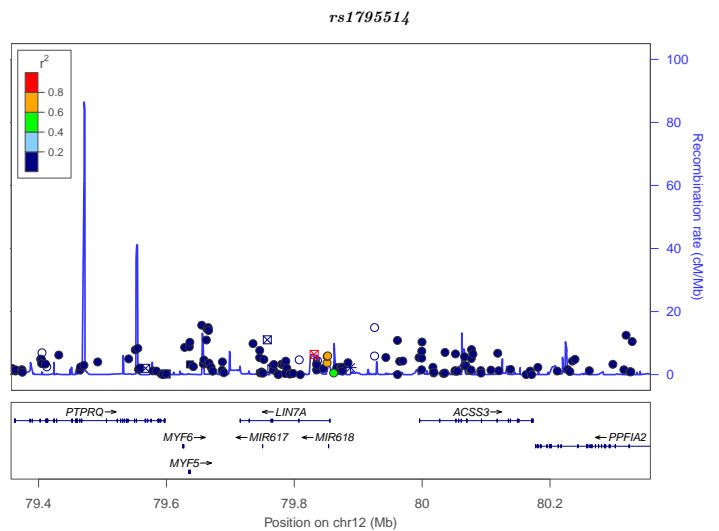


Figure A.293: NHS\_CNT(male)

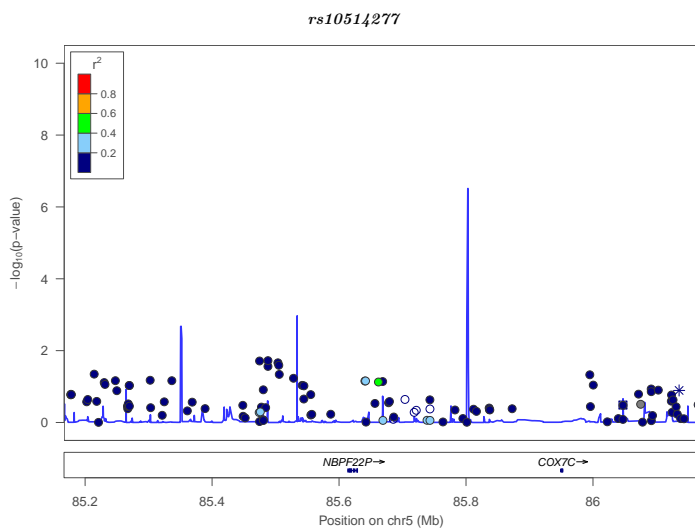


Figure A.294: NHS\_CNT(male)

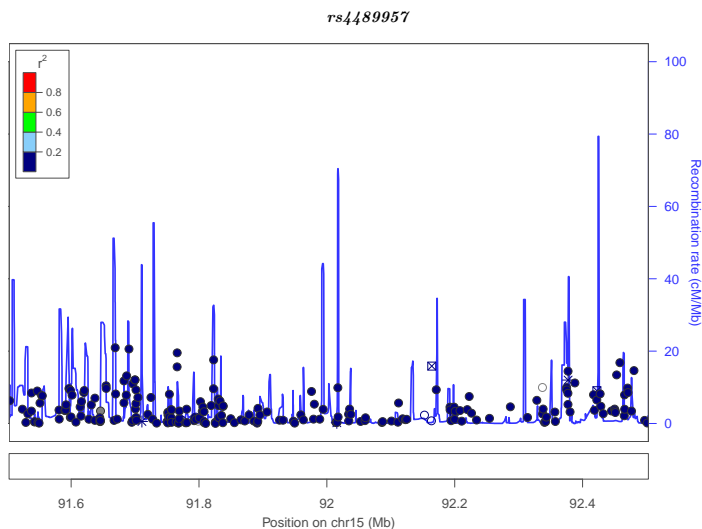


Figure A.295: NHS\_CNT(male)

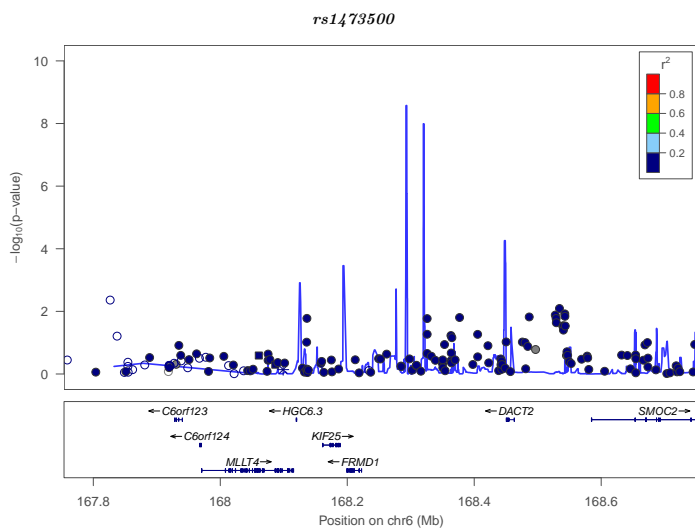


Figure A.296: NHS\_CNT(male)

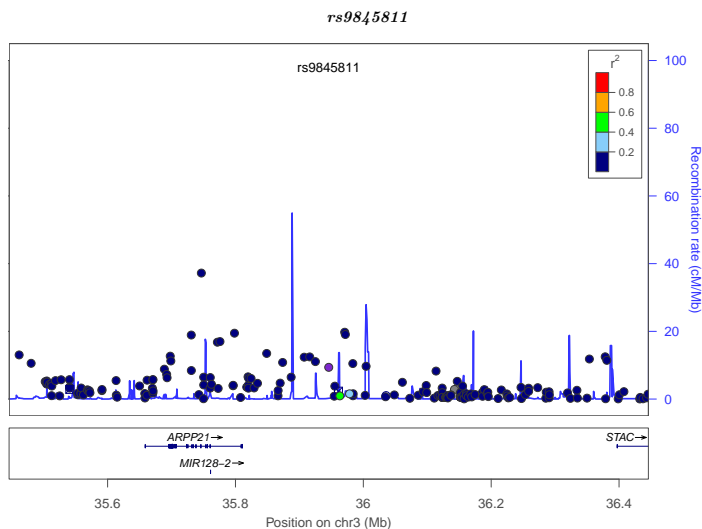


Figure A.297: NHS\_CNT(male)

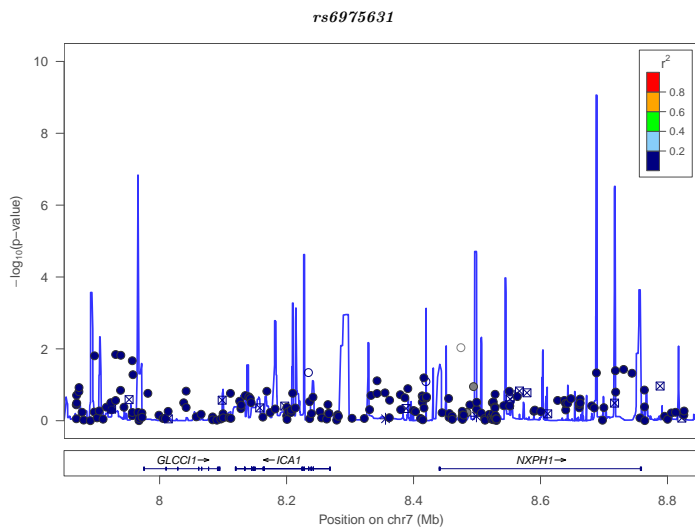


Figure A.298: NHS\_CNT(male)

### A.3.5 Phenotype: MOTIF

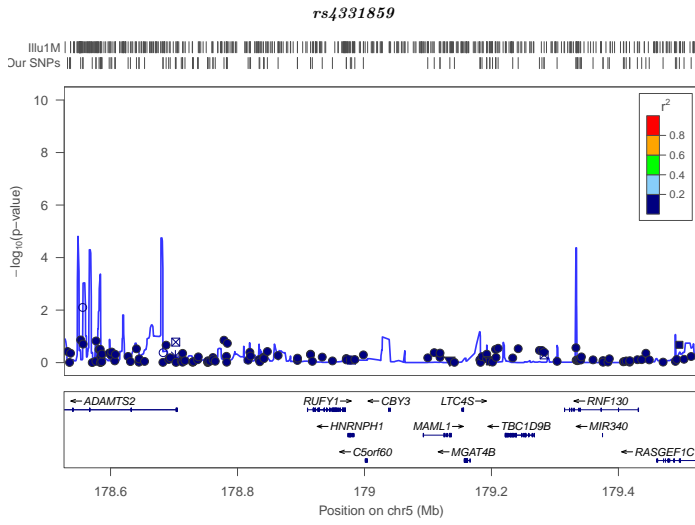


Figure A.299: MOTIF(combined)

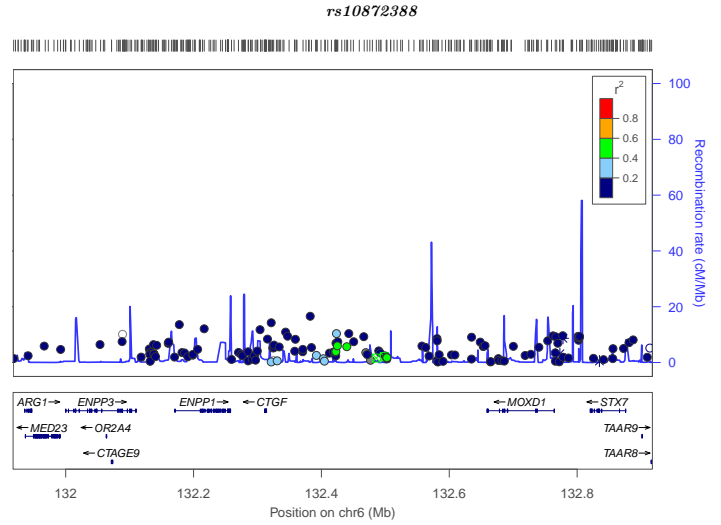


Figure A.300: MOTIF(combined)

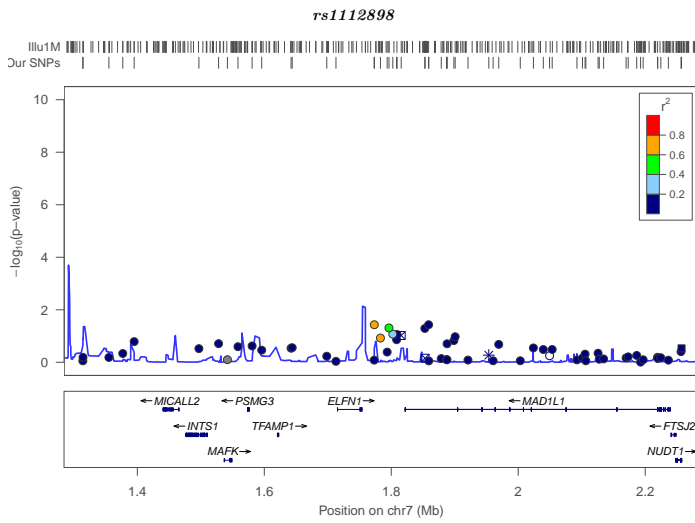


Figure A.301: MOTIF(combined)

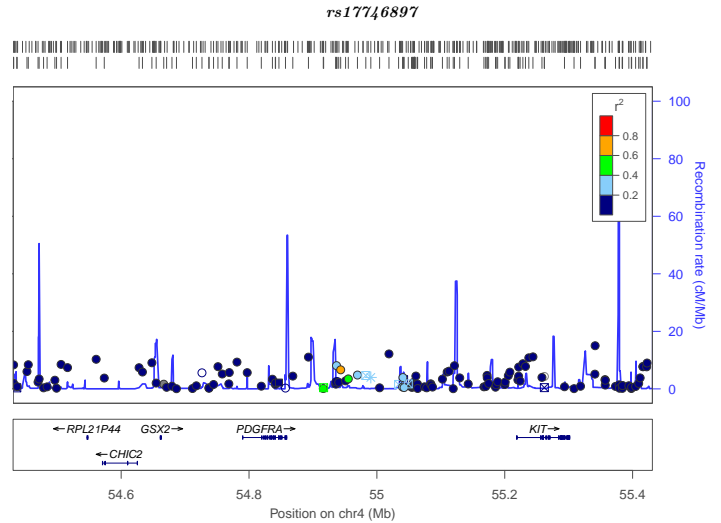


Figure A.302: MOTIF(combined)

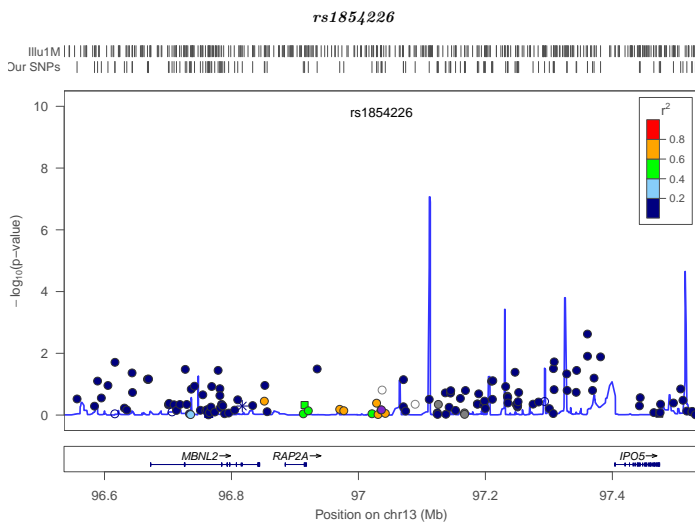


Figure A.303: MOTIF(combined)

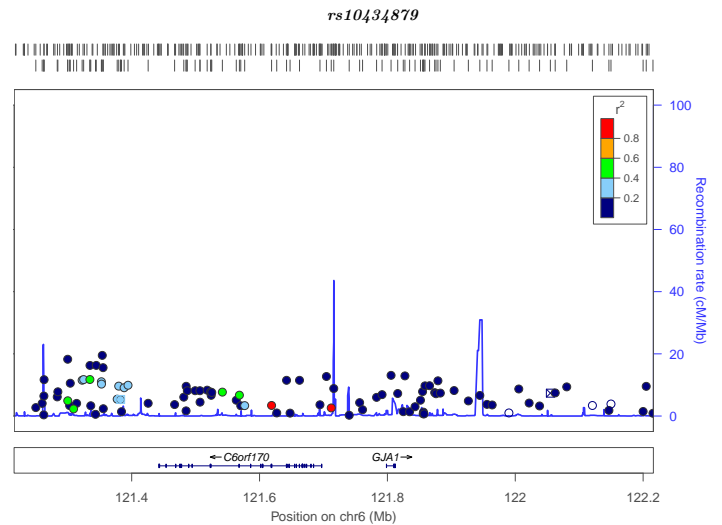


Figure A.304: MOTIF(combined)

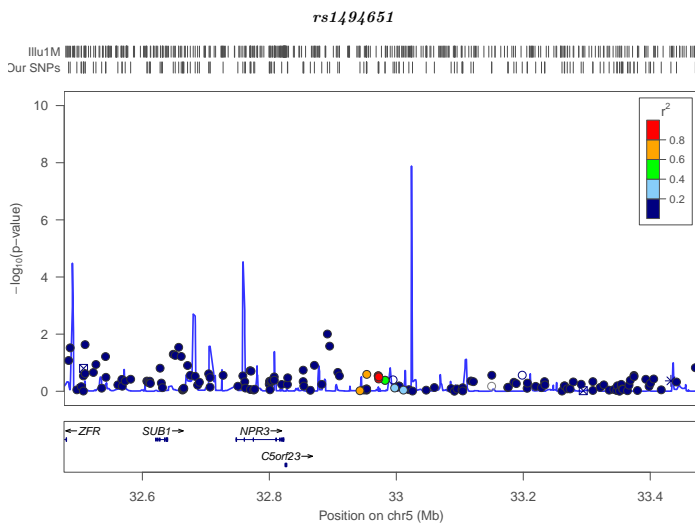


Figure A.305: MOTIF(combined)

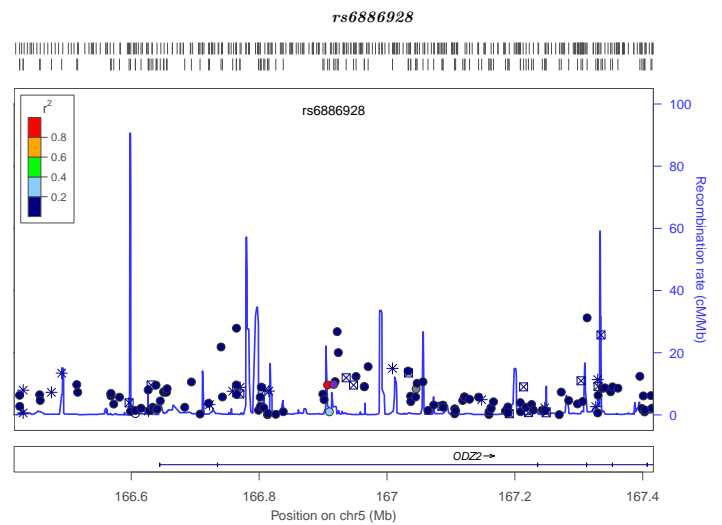


Figure A.306: MOTIF(female)

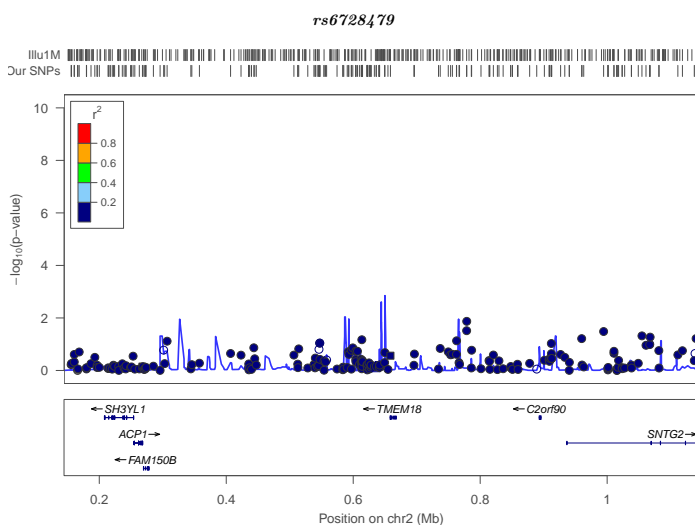


Figure A.307: MOTIF(female)

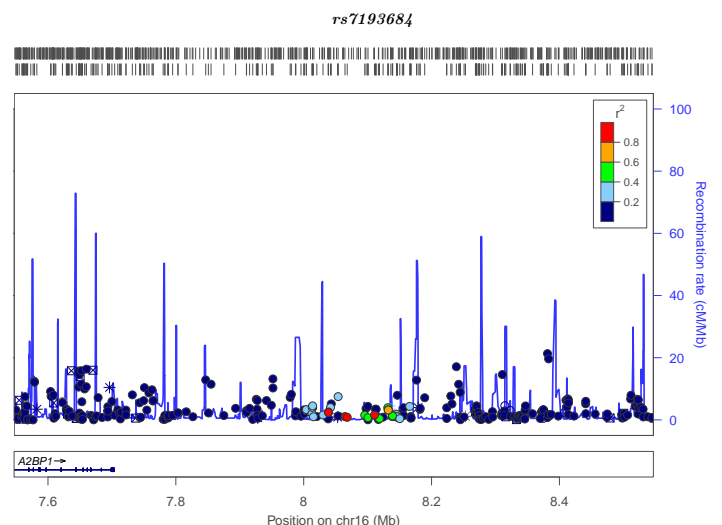


Figure A.308: MOTIF(female)

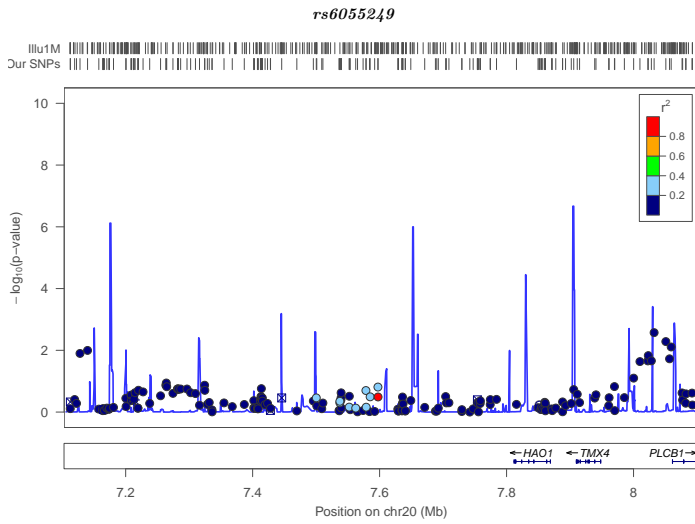


Figure A.309: MOTIF(female)

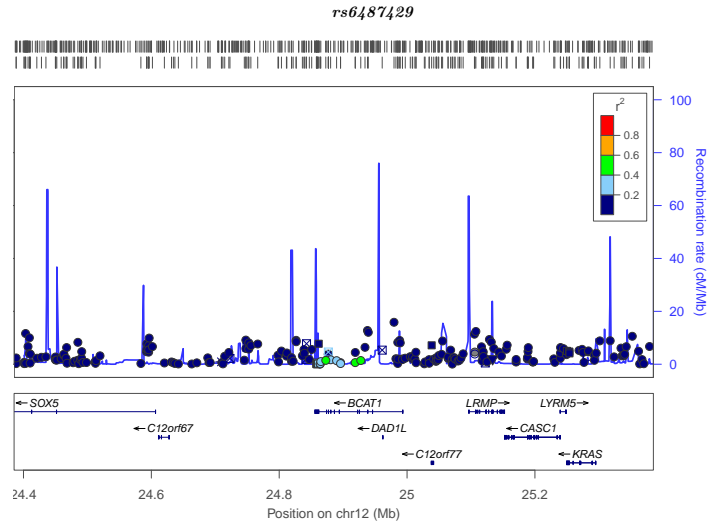


Figure A.310: MOTIF(female)

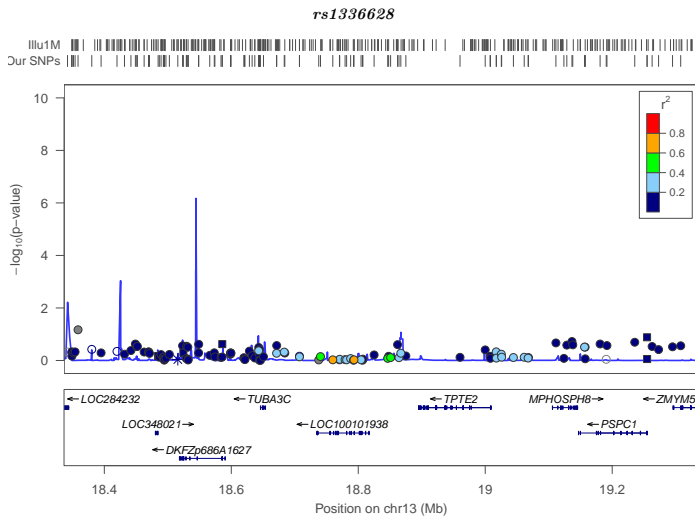


Figure A.311: MOTIF(male)

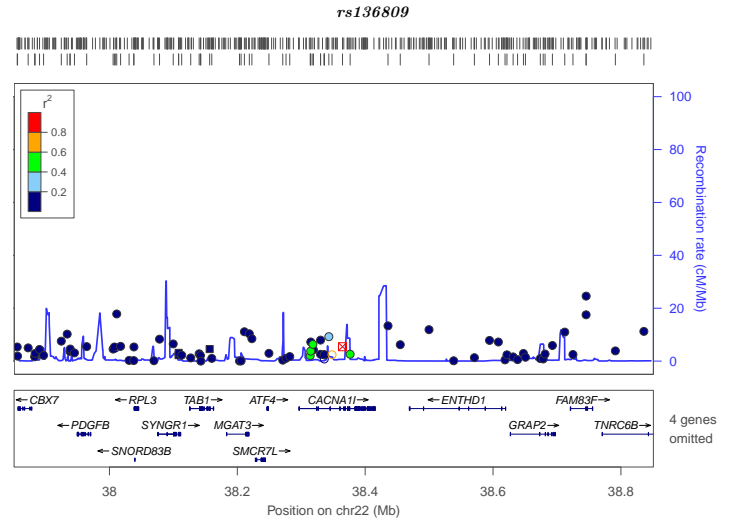


Figure A.312: MOTIF(male)

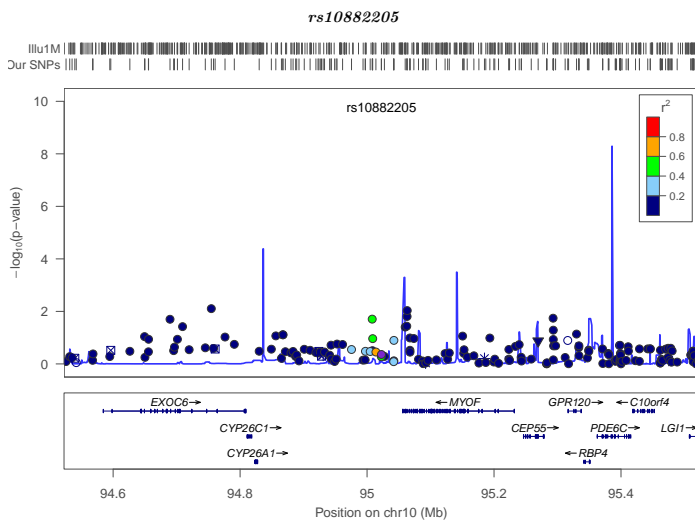


Figure A.313: MOTIF(male)

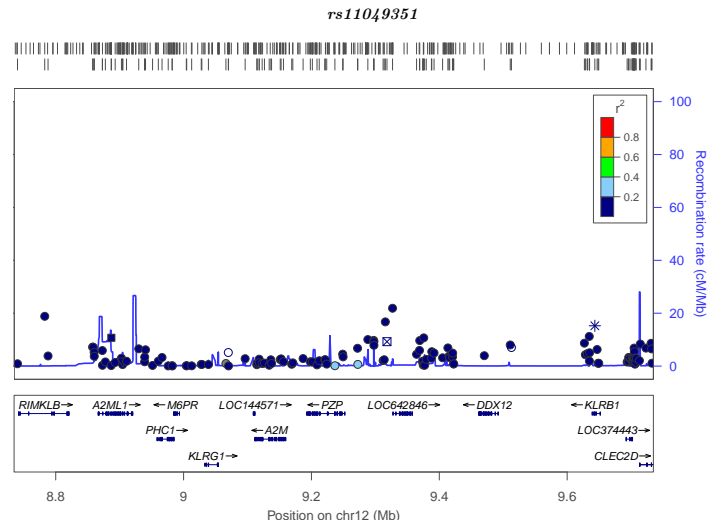


Figure A.314: MOTIF(male)

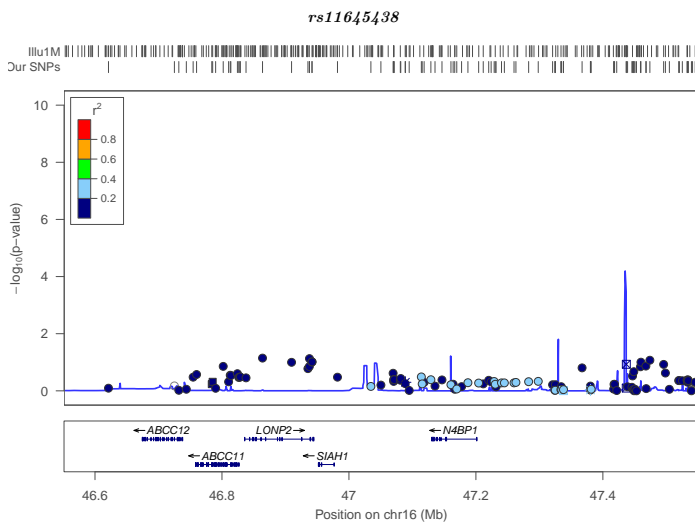


Figure A.315: MOTIF(male)

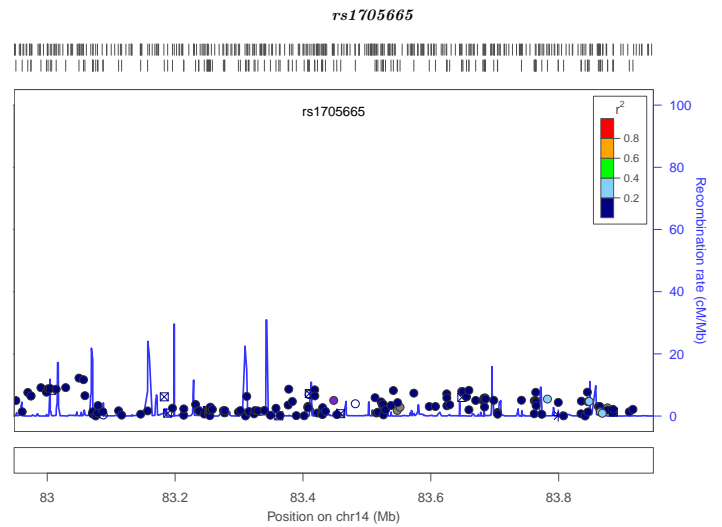


Figure A.316: MOTIF(male)

#### A.4 LOCUSZOOM PLOT OF *MAPT* GENE IN OUR META-ANALYSIS ACROSS PHENOTYPES

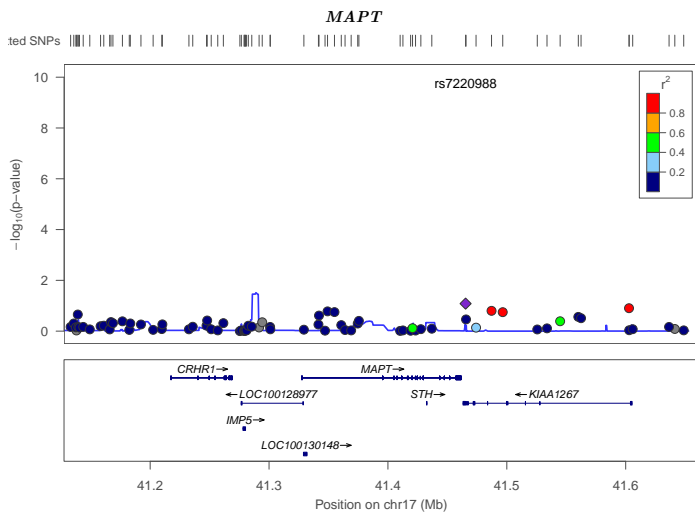


Figure A.317: ARC(combined)

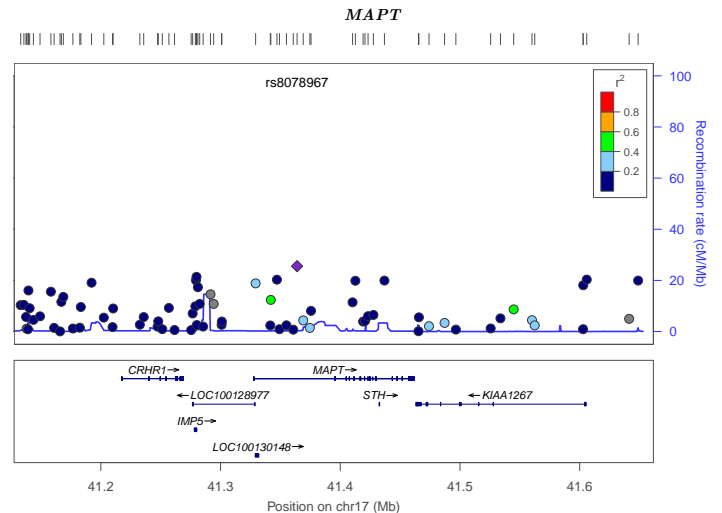


Figure A.318: ARC(female)

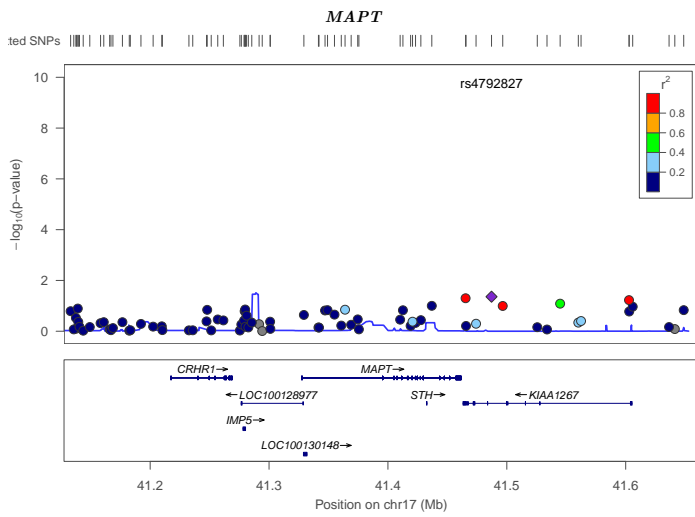


Figure A.319: ARC(male)

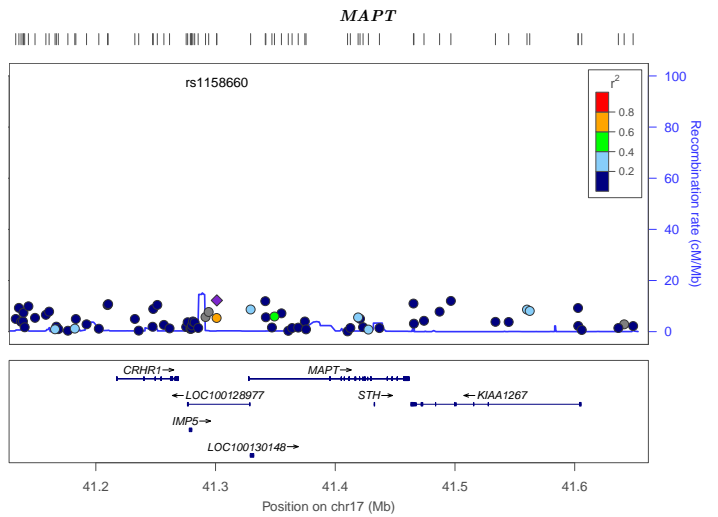


Figure A.320: HS\_PCT(combined)

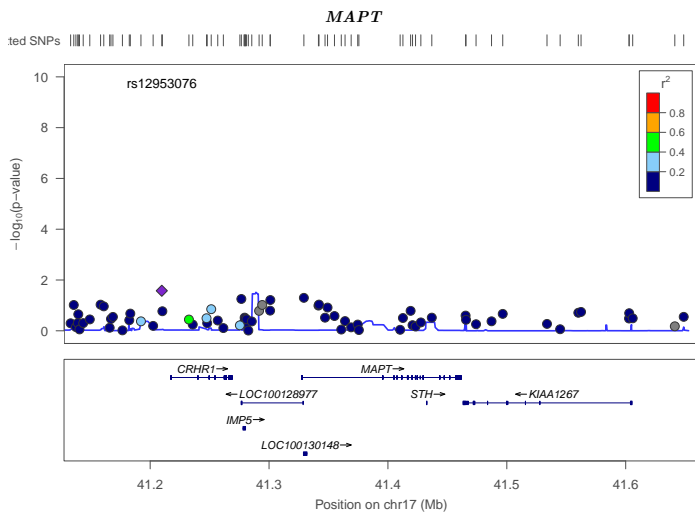


Figure A.321: HS\_PCT(female)

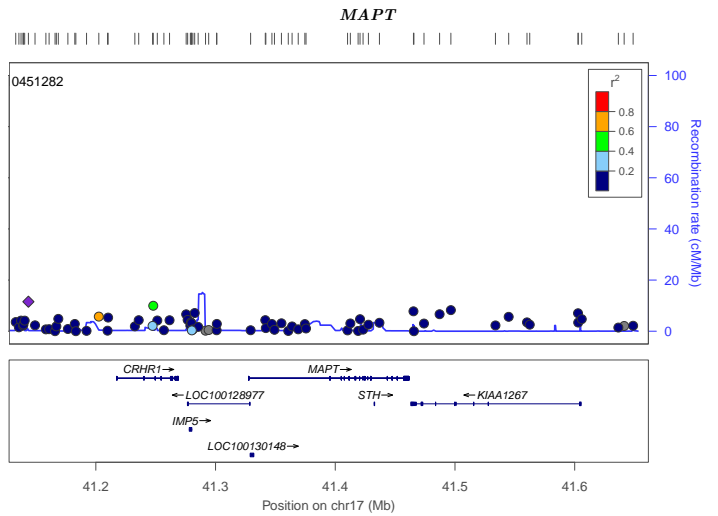


Figure A.322: HS\_PCT(male)

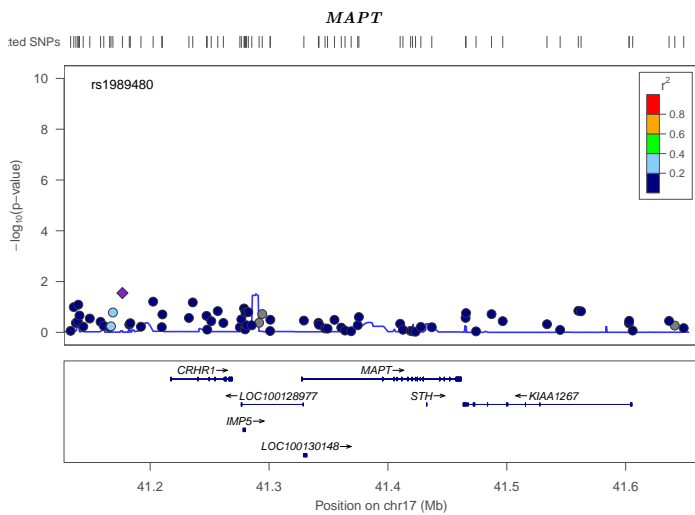


Figure A.323: HS\_CNT(combined)

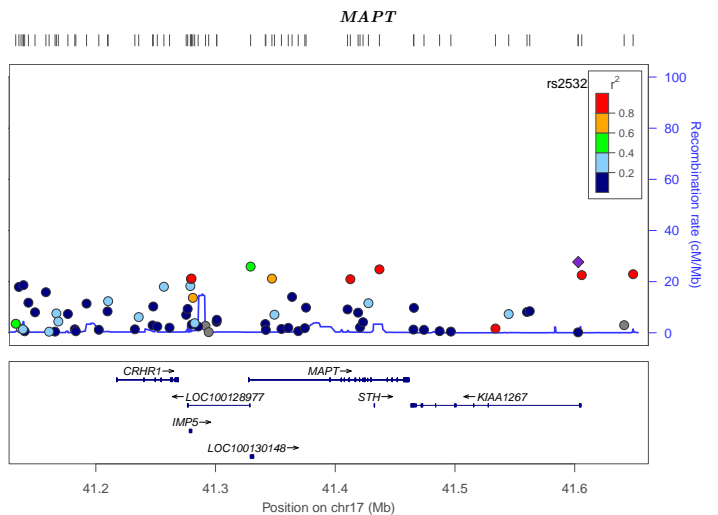


Figure A.324: HS\_CNT(female)



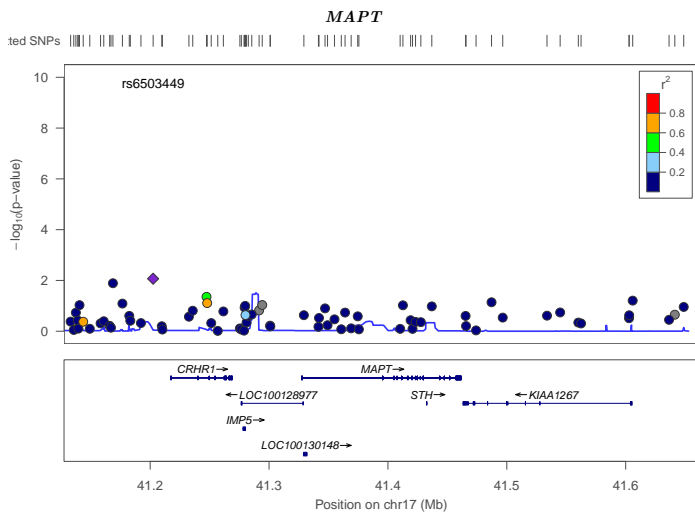


Figure A.325: HS\_CNT(male)

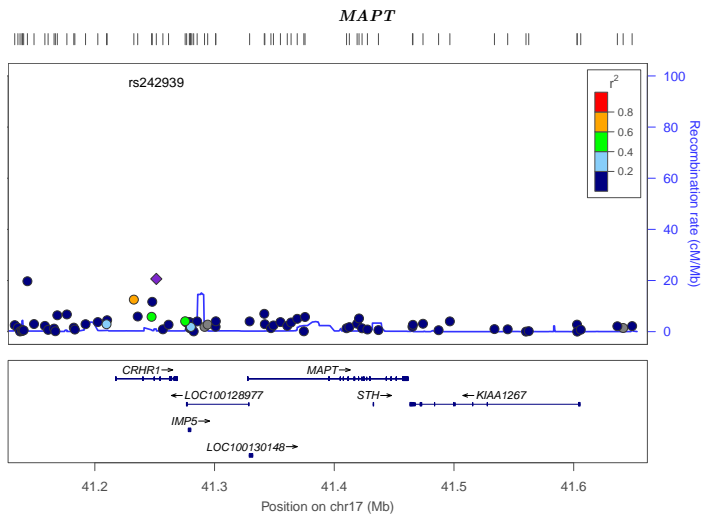


Figure A.326: NHS\_CNT(combined)

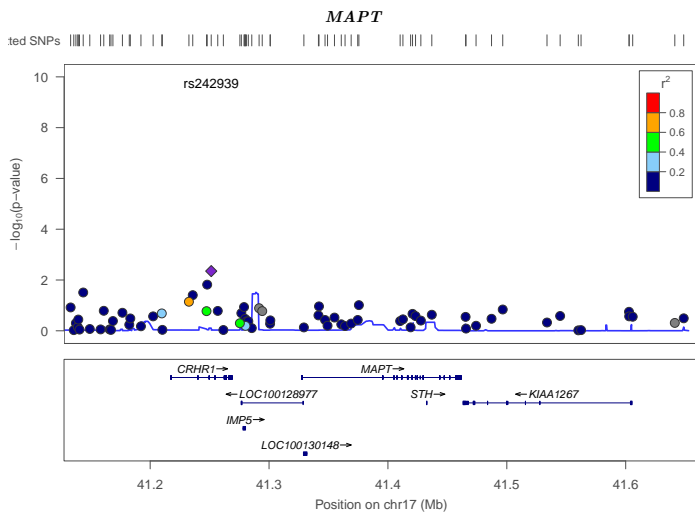


Figure A.327: NHS\_CNT(female)

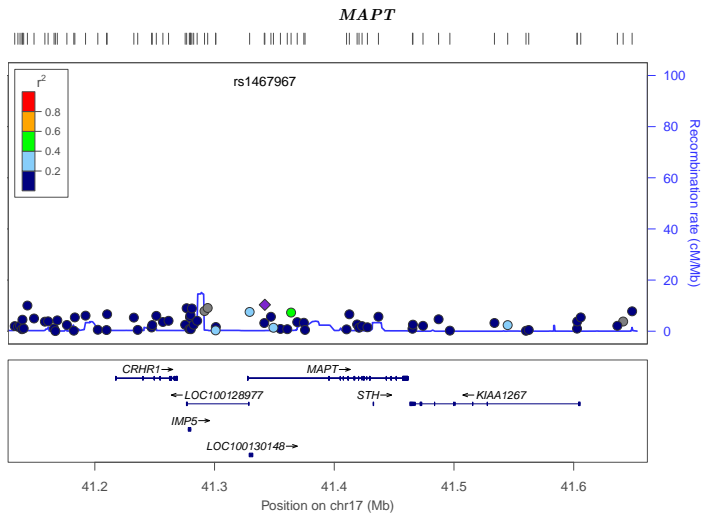


Figure A.328: NHS\_CNT(male)

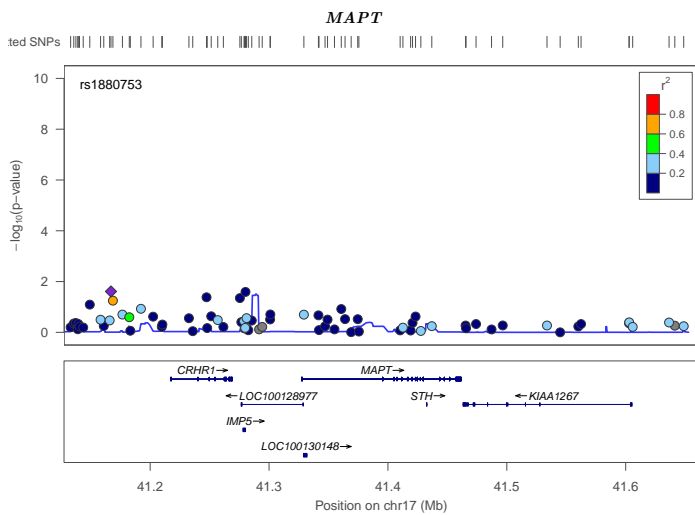


Figure A.329: MOTIF(combined)

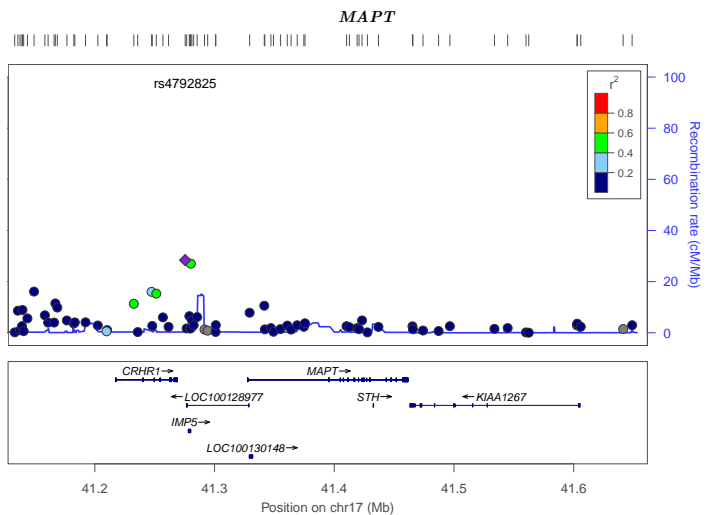


Figure A.330: MOTIF(female)

# MAPT\_M2

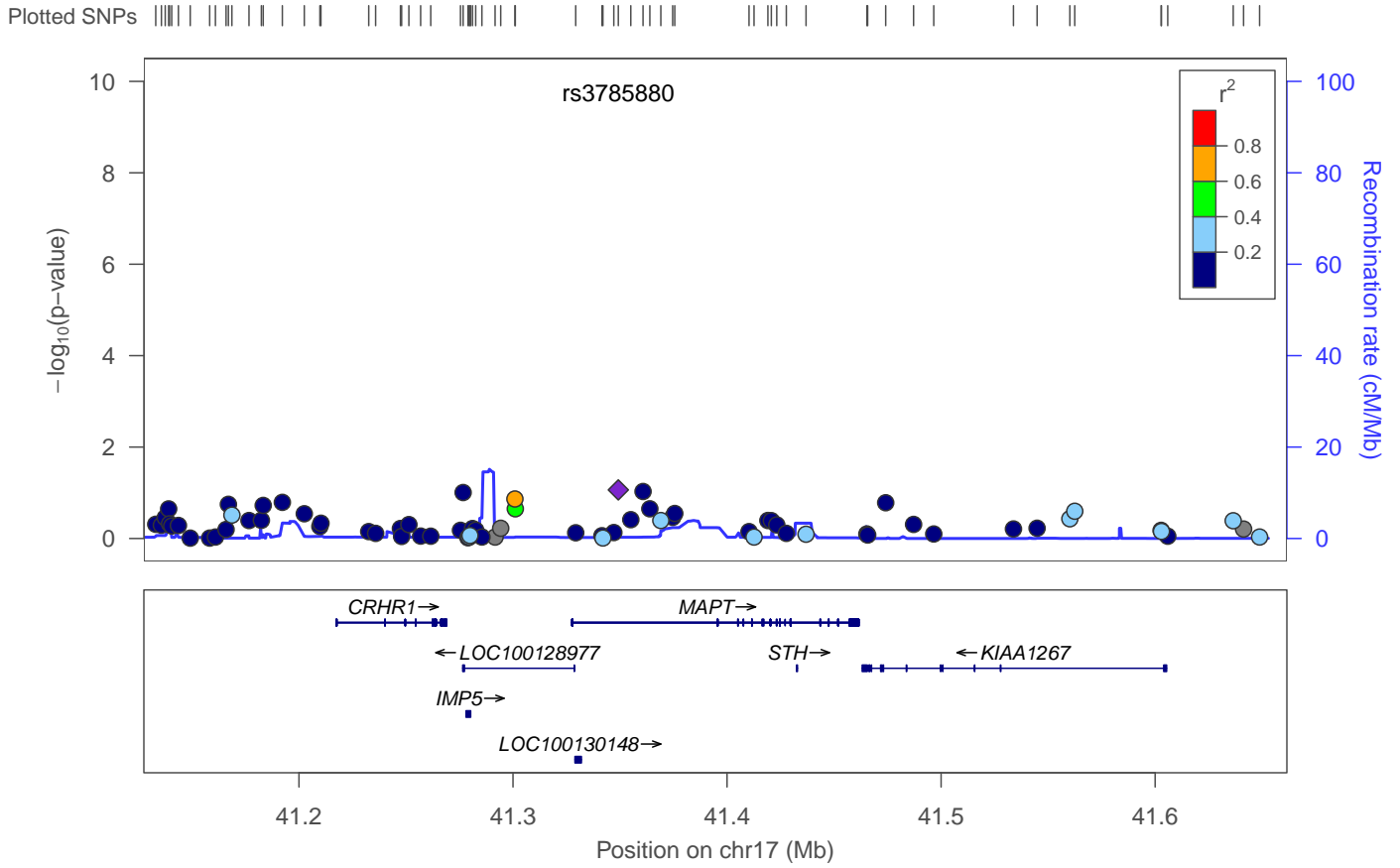


Figure A.331: MOTIF(male)

## APPENDIX B

### SCORING RECOMBINATION IN COMPLEX PEDIGREE STRUCTURES INCLUDING HALF-SIBLINGS

Scoring recombination in different new pedigree structures is presented below. In section one, we present the informative SNP tables for three-generation families allowing varying missing genotypes. In section two, we present the informative SNP tables for two and three-generation families including half-siblings. In section three, we present informative SNP tables for the mixed half and full-siblings in two and three generation families.

## Section One

### **Three-generation Pedigrees and tables:**

Source of Allele: Grandfather = 1 and Grandmother = 0

Solid color square or circle in the figure represents genotyped male or female and empty Square or circles represents missing genotyped male or female.

### **List of uninformative pedigrees:**

- a) Six pedigrees with none or one person genotyped is not informative for recombination.
- b) Among 10 combinations, eight pedigrees with two people genotyped are uninformative
- c) Among 10 combinations, five pedigrees with three people genotyped are uninformative
- d) Among 5 combinations, one pedigree with four people genotyped is uninformative.

### **List of informative pedigrees:**

Two people genotyped:

- a). Grandmother and grandchild
- b). Grandfather and grandchild

Three people genotyped:

- a). Grandfather, grandmother and grandchild
- b). Grandfather, mother and grandchild
- c). Grandmother, mother and grandchild
- d). Grandfather, father and grandchild
- e) Grandmother, father and grandchild

Four people genotyped

- a). Grandfather, grandmother, mother and grandchild
- b). Grandfather, mother, father and grandchild

c). Grandmother, mother, father and grandchild

d). Grandfather, grandmother, father and grandchild

Five persons genotyped:

a). Grandfather, grandmother, mother, father and grandchild

**Two persons genotyped:**

1. Two persons genotyped: Grandfather, grandchild

	Grandfather	Grandchild	Source of Allele
	AA	AA	?
		AB	?
		BB	0
	AB	AA	?
		AB	?
		BB	?
	BB	AA	0
		AB	?
BB		?	

2. Two persons genotyped: Grandmother, grandchild

	Grandmother	Grandchild	Source of Allele
	AA	AA	?
		AB	?
		BB	1
	AB	AA	?
		AB	?
		BB	?
	BB	AA	1
		AB	?
BB		?	

**Three people Genotyped:**

1. Grandfather, grandmother and grandchild genotyped

Grandfather	Grandmother	Mother (Poss genotype)	Grandchild	Source of Allele	
AA	AA	AA	AA	?	
			AB	?	
			BB	NP	
	AB	AA	AA or AB	AA	?
				AB	?
				BB	0
	BB	AA	AB	AA	1
				AB	?
				BB	0
AB	AA	AA or AB	AA	?	
			AB	?	
			BB	1	
	AB	AB	AA or AB or BB	AA	?
				AB	?
				BB	?
	BB	BB	AB or BB	AA	1
				AB	?
				BB	?
BB	AA	AB	AA	0	
			AB	?	
			BB	1	
	AB	AB	AB or BB	AA	0
				AB	?
				BB	?
	BB	BB	BB	AA	NP
				AB	?
				BB	?

2. Grandfather, Mother, and grandchild are genotyped:

Grandfather	Mother	Grandmother (Poss genotype)	Grandchild	Source of Allele
AA	AA	AA or AB	AA AB BB	? ? NP
	AB	AB or BB	AA AB BB	1 ? 0
	BB NP			
AB	AA	AA or AB or BB	AA AB BB	? ? NP
	AB	AA or AB or BB	AA AB BB	? ? ?
	BB	AB or BB	AA AB BB	NP ? ?
BB	AA NP			
	AB	AA or AB	AA AB BB	0 ? 1
	BB	AB or BB	AA AB BB	NP ? ?

3. Grandmother, Mother, and grandchild are genotyped:

Grand mother	Mother	Grandfather (Poss genotype)	Grandchild	Source of Allele
AA AB BB	AA	AA or AB	AA AB BB	? ? NP
	AB	AB or BB	AA AB BB	0 ? 1
	BB NP			
AB	AA	AA or AB	AA AB BB	? ? NP
	AB	AA or AB or BB	AA AB BB	? ? ?
	BB	AB or BB	AA AB BB	NP ? ?
BB	AA NP			
	AB	AA or AB	AA AB BB	1 ? 0
	BB	AB or BB	AA AB BB	NP ? ?



4. Maternal Grandfather, Father, and grandchild are genotyped:

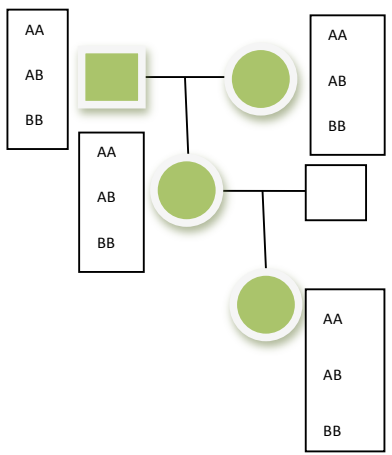
Grandfather	Father	Mother (Poss genotype)	Grandchild	Source of Allele
AA	AA	AA or AB	AA AB BB	? 0 NP
	AB	AA or AB	AA AB BB	? ? 0
	BB	AA or AB	AA AB BB	NP ? 0
AB	AA	AA or AB or BB	AA AB BB	? ? NP
	AB	AA or AB or BB	AA AB BB	? ? ?
	BB	AA or AB or BB	AA AB BB	NP ? ?
BB	AA	AB or BB	AA AB BB	0 ? NP
	AB	AB or BB	AA AB BB	0 ? ?
	BB	AB or BB	AA AB BB	NP 0 ?

5. Maternal Grandmother, Father, and grandchild are genotyped:

Grandmother	Father	Mother (Poss genotype)	Grandchild	Source of Allele
AA	AA	AA or AB	AA AB BB	? 1 NP
	AB	AA or AB	AA AB BB	? ? 1
	BB	AA or AB	AA AB BB	NP ? 1
AB	AA	AA or AB or BB	AA AB BB	? ? NP
	AB	AA or AB or BB	AA AB BB	? ? ?
	BB	AA or AB or BB	AA AB BB	NP ? ?
BB	AA	AB or BB	AA AB BB	1 ? NP
	AB	AB or BB	AA AB BB	1 ? ?
	BB	AB or BB	AA AB BB	NP 1 ?

**Four people Genotyped:**

**1. Grandfather, grandmother, mother and grandchild**



Grandfather	Grandmother	Mother	Father (Poss genotype)	Grandchild	Source of Allele
AA	AA	AA	AA or AB or BB	AA AB BB	? ? NP
	AB	AA	AA or AB or BB	AA AB BB	? ? NP
		AB	AA or AB or BB	AA AB BB	1 ? 0
	BB	AB	AA or AB or BB	AA AB BB	1 ? 0
AB	AA	AA	AA or AB or BB	AA AB BB	? ? NP
		AB	AA or AB or BB	AA AB BB	0 ? 1
	AB	AA AB BB (Nine poss cases, all uninformativ)	AA AB BB	? ? ?	
	BB	AB	AA or AB or BB	AA AB BB	1 ? 0
		BB	AA or AB or BB	AA AB BB	NP ? ?
BB	AA	AB	AA or AB or BB	AA AB BB	0 ? 1
	AB	AB	AA or AB or BB	AA AB BB	0 ? 1

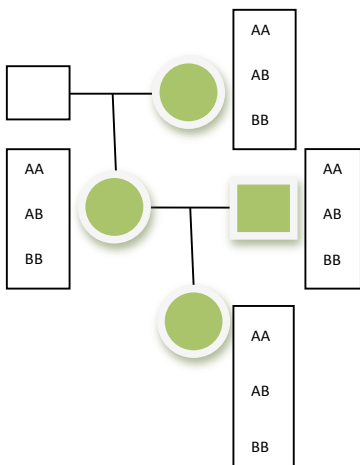
			BB		AA AB BB	NP ? ?
		BB	BB	AA or AB or BB	AA AB BB	NP ? ?

## 2. Grandfather, mother, father and grandchild

	Grandfather	Grandmother (poss genotype)	Mother	Father	Grandchild	Source of Allele
		AA	AA of AB	AA	AA	AA AB BB
AB					AA AB BB	? ? NP
BB					AA AB BB	NP ? NP
AB or BB			AB		AA	1 0 NP
					AB	1 ? 0
					BB	NP 1 0
AB		AA or AB	AA	AA	AA AB BB	? NP NP
				AB	AA AB BB	? ? NP
				BB	AA AB BB	NP ? NP
		AA or AB or BB		AB	AA	? ? NP

		AB or BB	BB	AB	AA AB BB	? ? ?
				BB	AA AB BB	NP ? ?
				AA	AA AB BB	NP ? NP
				AB	AA AB BB	NP ? ?
				BB	AA AB BB	NP NP ?
				BB		
	BB	AA or AB	AB	AA	AA AB BB	0 1 NP
				AB	AA AB BB	0 ? 1
				BB	AA AB BB	NP 0 1
				BB		
		AB or BB	BB	AA	AA AB BB	NP ? NP
				AB	AA AB BB	NP ? ?
				BB	AA AB BB	NP NP ?
				BB		

### 3. Grandmother, mother, father and grandchild



Grandmother	Grandfather (Poss genotype)	Mother	Father	Grandchild	Source of Allele	
AA	AA of AB	AA	AA	AA AB BB	? NP NP	
			AB	AA AB BB	? ? NP	
			BB	AA AB BB	NP ? NP	
	AB or BB	AB	AA	AA AB BB	0 1 NP	
			AB	AA AB BB	0 ? 1	
			BB	AA AB BB	NP 0 1	
			BB NP			
	AB	AA or AB	AA	AA	AA AB BB	? NP NP
				AB	AA AB BB	? ? NP
				BB	AA AB BB	NP ? NP
AA or AB or BB		AB	AA	AA AB BB	? ? NP	
			AB	AA AB BB	? ? ?	
			BB	AA AB BB	NP ? ?	
AB or BB		BB	AA	AA AB BB	NP ? NP	
			AB	AA AB BB	NP ? ?	

				BB	AA AB BB	NP NP ?
	BB		AA NP			
		AA or AB	AB	AA	AA AB BB	1 0 NP
				AB	AA AB BB	1 ? 0
				BB	AA AB BB	NP 1 0
		AB or BB	BB	AA	AA AB BB	NP ? NP
				AB	AA AB BB	NP ? ?
				BB	AA AB BB	NP NP ?

4. Four people genotyped: Grandfather, grandmother, father and grandchild

Grandfather	Grandmother	Mother (Poss genotype)	Father	Grandchild	Source of Allele		
AA	AA	AA	AA	AA	? NP NP		
			AB	AA	? ? NP		
			BB	AA	NP ? NP		
		AB	AA or AB	AA	AA	? 0 NP	
				AB	AA	? ? 0	
				BB	AA	NP ? 0	
	BB	AB	AB	AA	AA	1 0 NP	
				AB	AA	1 ? 0	
				BB	AA	NP 1 0	
		AB	AA	AA or AB	AA	AA	? 1 NP
					AB	AA	? ? 1
					BB	AA	NP ? 1
AB	AA or AB or BB		AA or AB or BB	AA	AA	? ? NP	
				AB	AA	? ? ?	
				BB	AA	NP ? ?	
BB	AB or BB	AB or BB	AA	AA	1 AB ?		



					BB	NP
				AB	AA AB BB	? ? ?
				BB	AA AB BB	NP 1 ?
	BB	AA	AB	AA	AA AB BB	0 1 NP
				AB	AA AB BB	0 ? 1
				BB	AA AB BB	NP 0 1
		AB	AB or BB	AA	AA AB BB	0 ? NP
				AB	AA AB BB	0 ? ?
				BB	AA AB BB	NP 0 ?
		BB	BB	AA	AA AB BB	NP ? NP
				AB	AA AB BB	NP ? ?
				BB	AA AB BB	NP NP ?

**Five people genotyped**

	Grandfather	Grandmother	Mother	Father	Grandchild	Source of Allele			
	AA	AA	AA	AA	AA	?			
					AB	NP			
						BB	NP		
				AB	AA	?			
					AB	?			
					BB	NP			
				BB	AA	NP			
					AB	?			
					BB	NP			
				AB	AB	AA	AA	AA	?
								AB	NP
								BB	NP
							AB	AA	?
								AB	?
								BB	NP
							AB	AA	NP
								AB	?
								BB	NP
AB	BB	AB	AA	1					
			AB	0					
			BB	NP					
AB	BB	AB	AA	1					
			AB	?					
			BB	0					
AB	BB	AB	AA	NP					
			AB	1					
			BB	0					
AB	AA	AA	AA	AA	?				
				AB	NP				
				BB	NP				
			AB	AA	?				
				AB	?				
				BB	NP				
AB	AA	AB	AA	O					
			AB	1					
AB	AA	BB	AA	NP					
			AB	NP					

					AB BB	? 1			
				BB	AA AB BB	NP O 1			
				AB	AA	AA	AA AB BB	? ? NP	
						AB	AA AB BB	? ? NP	
						BB	AA AB BB	NP ? NP	
						AB	AA	AA AB BB	? ? NP
							AB	AA AB BB	? ? ?
							BB	AA AB BB	NP ? ?
					BB	AA	AA AB BB	NP ? NP	
						AB	AA AB BB	NP ? ?	
						BB	AA AB BB	NP NP ?	
					BB	AB	AA	AA AB BB	1 O NP
							AB	AA AB BB	1 ? O
							BB	AA AB BB	NP 1 O
				BB		AA	AA AB BB	NP ? NP	
						AB	AA AB BB	NP ? ?	
						BB	AA AB BB	NP NP ?	
				BB		AA	AA AB BB	0 1 NP	

		AB		AB	AA AB BB	0 ? 1	
				BB	AA AB BB	NP 0 1	
			AB	AB	AA	AA AB BB	0 1 NP
					AB	AA AB BB	0 ? 1
					BB	AA AB BB	NP 0 1
			BB	BB	AA	AA AB BB	NP ? NP
		AB			AA AB BB	NP ? ?	
		BB			AA AB BB	NP NP ?	
		BB	BB	AA	AA AB BB	NP ? NP	
				AB	AA AB BB	NP ? ?	
				BB	AA AB BB	NP NP ?	

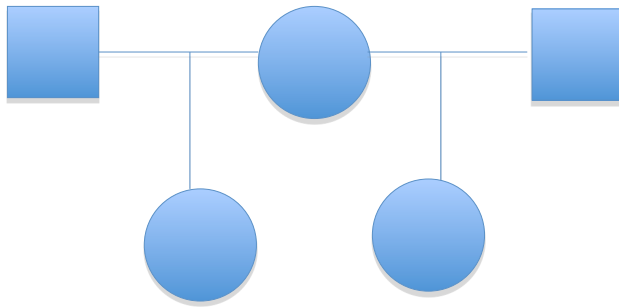
## Section 2

### Pedigree tables for half-siblings:

S: same, D: Different

Solid color square or circle in the figure represents genotyped male or female and empty Square or circles represents missing genotyped male or female.

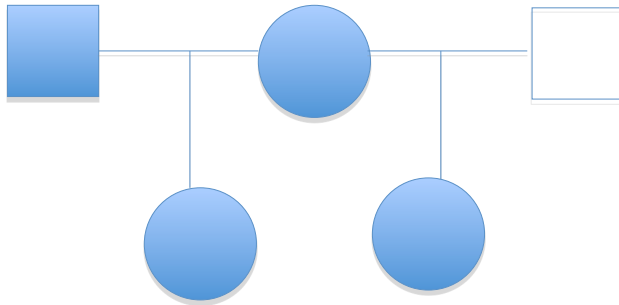
Five people genotyped:



Mom	DAD1	Child1	Dad2	Child2	Grandparental allele
AB	AA	AA	AA	AA AB	S D
			AB	AA AB BB	S ? D
			BB	AB BB	S D
		AB	AA	AA AB	D S
		AB	AA AB BB	D ? S	
		BB	AB BB	D S	
	AB	AA	AA	AA AB	S D
			AB	AA AB BB	S ? D
			BB	AB BB	S D
		AB	AA	AA AB	? ?

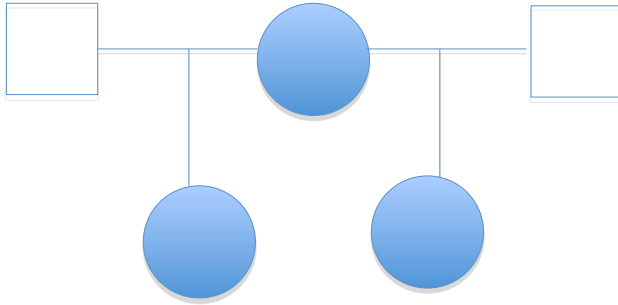
			AB	AA AB BB	? ? ?
			BB	AB BB	? ?
		BB	AA	AA AB	D S
			AB	AA AB BB	D ? S
			BB	AB BB	D S
		BB	AB	AA	AA AB
	AB			AA AB BB	S ? D
	BB			AB BB	S D
	BB		AA	AA AB	D S
			AB	AA AB BB	D ? S
			BB	AB BB	D S
	AA or BB				?

**Four people genotyped:**



Mom	DAD1	Child1	Dad2 Prob genotype	Child2	Grandparental allele
AB	AA	AA	AA or AB or BB	AA AB BB	S ? D
		AB	AA or AB or BB	AA AB BB	D ? S
	AB	AA	AA or AB or BB	AA AB BB	S ? D
		AB	AA or AB or BB	AA AB BB	? ? ?
		BB	AA or AB or BB	AA AB BB	D ? S
	BB	AB	AA or AB or BB	AA AB BB	S ? D
		BB	AA or AB or BB	AA AB BB	D ? S
	AA or BB				?

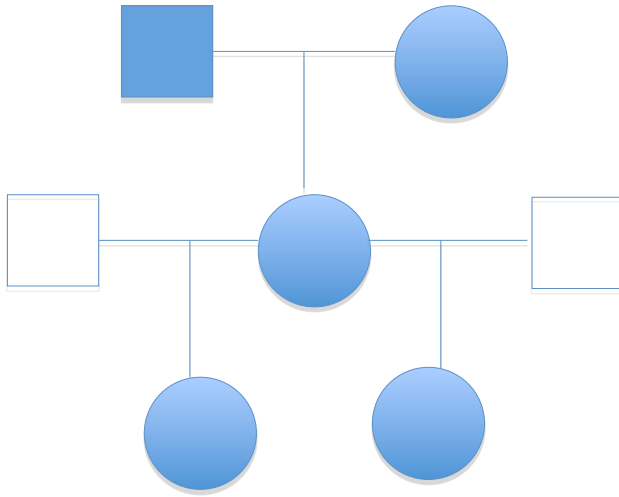
**Three people genotyped:**



Mom	DAD1 Prob genotype	Child1	Dad2 Prob genotype	Child2	Grandparental allele
AB	AA or AB or BB	AA	AA or AB or BB	AA AB BB	S ? D
		AB	AA or AB or BB	AA AB BB	? ? ?
		BB	AA or AB or BB	AA AB BB	D ? S
AA or BB					?

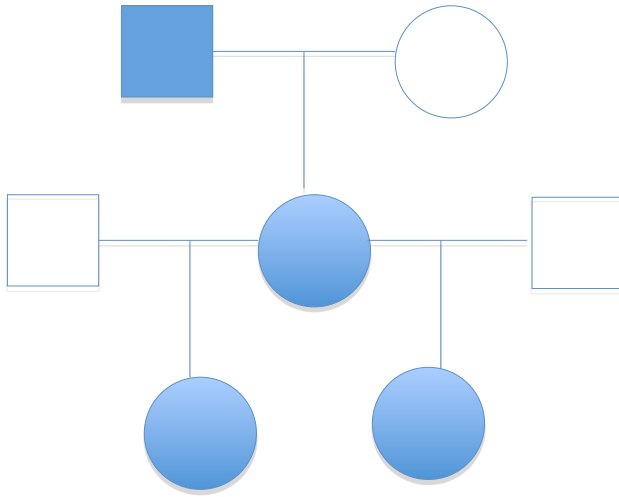


**Five people genotyped:**



Grand Dad	Grand Mom	Mom	DAD1	Child1	DAD2	Child2	Allele status
AA	BB	AB		AA		AA AB BB	S ? D
				AB		AA AB BB	? ? ?
				BB		AA AB BB	D ? S
	AB	AB		AA		AA AB BB	S ? D
				AB		AA AB BB	? ? ?
				BB		AA AB BB	D ? S
AB	AA	AB					
	BB	AB					
BB	AA	AB					
	AB	AB					

**Four people genotyped:**



Grand_Dad	Mom	DAD1	Child1	DAD2	Child2	Grandparental Allele status
AA	AB		AA		AA	S
			AB		AB	?
			BB		BB	D
BB	AB		AA		AA	S
			AB		AB	?
			BB		BB	D

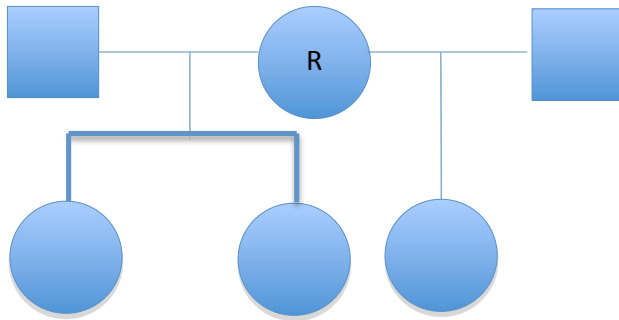
### Section Three

**Pedigree tables for mixed two-generation families with half-siblings:**

S: same, D: Different

Solid colored squares or circles in the figure represent genotyped male or female and empty squares or circles represent missing genotyped male or female.

**Six people genotyped:**



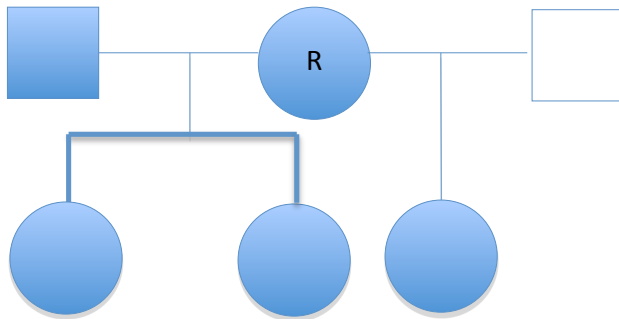
Mom	D1	C1	C2	D2	C3	C1&C2	HS_C1&C3	HS_C2&C3
AB	AA	AA	AA	AA	AA	S	S	S
				AB	AB	S	D	D
				BB	BB	S	D	D
			AB	AA	AA	D	D	S
				AB	AB	D	?	?
				BB	BB	D	D	S
		AB	AA	AA	AA	D	D	S
				AB	AB	D	?	?
				BB	BB	D	S	D

			AB	AA	AA	S	D	D
					AB	S	S	S
				AB	AA	S	D	D
					AB	S	?	?
					BB	S	S	S
			BB	AB	BB	S	D	D
					BB	S	S	S
	AB	AA	AA	AA	AA	S	S	S
					AB	S	D	D
				AB	AA	S	S	S
					AB	S	?	?
					BB	S	D	D
			BB	AB	BB	S	S	S
					BB	S	D	D
			AB	AA	AA	?	S	?
					AB	?	D	?
				AB	AA	?	S	?
					AB	?	?	?
					BB	?	D	?
			BB	AB	BB	?	S	?
					BB	?	D	?
		AB	AA	AA	AA	D	S	D
					AB	D	D	S
				AB	AA	D	S	D
					AB	D	?	?
					BB	D	D	S
			BB	AB	BB	D	S	D
					BB	D	D	S
		AB	AA	AA	AA	?	?	S
					AB	?	?	D
				AB	AA	?	?	S
					AB	?	?	?
					BB	?	?	D
			AB	AA	AA	?	?	S
					AB	?	?	?
					BB	?	?	D
			BB	AA	AA	?	?	D
					AB	?	?	S
				AB	AA	?	?	D
					AB	?	?	?

				BB	?	?	S
			BB	AB BB	? ?	? ?	D S
	BB	AA	AA	AA AB	D D	D S	S D
			AB	AA AB BB	D D D	D ? S	S ? D
			BB	AB BB	D D	D S	S D
		AB	AA	AA AB	? ?	D S	? ?
			AB	AA AB BB	? ? ?	D ? S	? ? ?
			BB	AB BB	? ?	D S	? ?
		BB	AA	AA AB	S S	D S	D S
			AB	AA AB BB	S S S	D ? S	D ? S
			BB	AB BB	S S	D S	D S
BB	AB	AB	AA	AA AB	S S	S D	S D
			AB	AA AB BB	S S S	S ? D	S ? D
			BB	AB BB	S S	S D	S D
		BB	AA	AA AB	D D	S D	D S
			AB	AA AB BB	D D D	S ? D	D ? S
			BB	AB BB	D D	S D	D S
	BB	AB	AA	AA AB	D D	D S	S D
			AB	AA AB BB	D D D	D ? S	S ? D
			BB	AB BB	D D	D S	S D

			BB	AA	AA	S	D	D
				AB	AA	S	D	D
					AB	S	?	?
					BB	S	S	S
				BB	AB	S	D	D
					BB	S	S	S

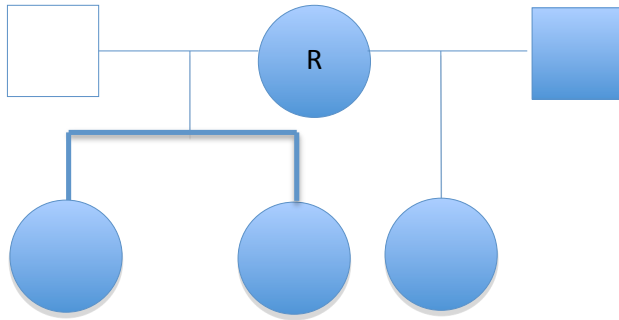
**FIVE people genotyped:**



Mom	D1	C1	C2	HS_C3	C1&C2	HS_C1&C3	HS_C2&C3	
AB	AA	AA	AA	AA	S	S	S	
				AB	S	?	?	
				BB	S	D	D	
		AB	AA	AA	AB	D	S	D
					AB	D	?	?
					BB	D	D	S
	AB	AA	AA	AA	D	D	S	
				AB	D	?	?	
				BB	D	S	D	
	AB	AA	AA	AA	S	D	D	
				AB	S	?	?	
				BB	S	S	S	
AB	AA	AA	AA	S	S	S		
			AB	S	?	?		
			BB	S	D	D		
AB	AA	AA	AB	?	S	?		
			AB	?	?	?		
			BB	?	D	?		
AB	AA	AA	BB	D	S	D		

			AB BB	D D	? D	? S
	AB	AA	AA AB BB	? ? ?	? ? ?	S ? D
		AB	AA AB BB	? ? ?	? ? ?	? ? ?
		BB	AA AB BB	? ? ?	? ? ?	D ? S
	BB	AA	AA AB BB	D D D	D ? S	S ? D
		AB	AA AB BB	? ? ?	D ? S	? ? ?
		BB	AA AB BB	S S S	D ? S	D ? S
BB	AB	AB	AA AB BB	S S S	S ? D	S ? D
		BB	AA AB BB	D D D	S ? D	D ? S
	BB	AB	AA AB BB	D D D	D ? S	S ? D
		BB	AA AB BB	S S S	D ? S	D ? S

**FIVE people genotyped:**

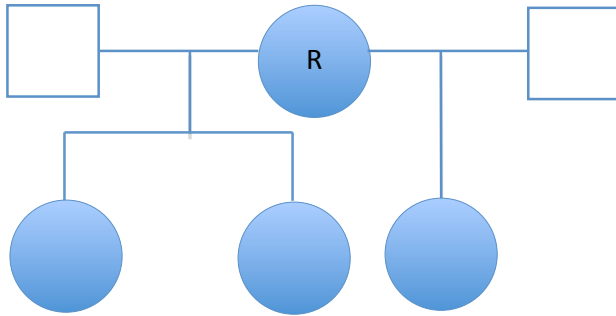


Mom	C1	C2	D2	C3	C1&C2	HS_C1&C3	HS_C2&C3
AB	AA	AA	AA	AA	S	S	S
				AB	S	D	D
			AB	AA	S	S	S
				AB	S	?	?
				BB	S	D	D
			BB	AB	S	S	S
			BB	S	D	D	
		AB	AA	?	S	?	
			AB	?	D	?	
			AB	AA	?	S	?
				AB	?	?	?
				BB	?	D	?
		BB	AA	D	S	D	
			AB	D	S	D	
			AB	D	?	?	
			BB	D	D	S	
			BB	AB	D	S	D
			BB	BB	D	D	S
		AB	AA	AA	?	?	S
				AB	?	?	D
		AB	AA	?	?	S	
			AB	?	?	?	
			BB	?	?	D	
		AB	AA	?	?	?	
			AB	?	?	?	
			AB	AA	?	?	



			AB BB	? ?	? ?	? ?
		BB	AB BB	? ?	? ?	? ?
	BB	AA	AA AB	? ?	? ?	D S
		AB	AA AB BB	? ? ?	? ? ?	D ? S
		BB	AB BB	? ?	? ?	D S
	BB	AA	AA AB	D D	D S	S D
		AB	AA AB BB	D D D	D ? S	S ? D
		BB	AB BB	D D	D S	S D
		AB	AA AB	? ?	D S	? ?
		AB	AA AB BB	? ? ?	D ? S	? ? ?
		BB	AB BB	? ?	D S	? ?
	BB	AA	AA AB	S S	D S	D S
		AB	AA AB BB	S S S	D ? S	D ? S
		BB	AB BB	S S	D S	D S

**FOUR people genotyped:**



Mom	C1	C2	HS C3	C1&C2	HS C1&C3	HS C2&C3
AB	AA	AA	AA AB BB	S S S	S ? D	S ? D
		AB	AA AB BB	? ? ?	S ? D	? ? ?
		BB	AA AB BB	D D D	S ? D	D ? S
	AB	AA	AA AB BB	? ? ?	S ? D	? ? ?
		AB	AA AB BB	? ? ?	? ? ?	? ? ?
		BB	AA AB BB	? ? ?	? ? ?	D ? S
	BB	AA	AA AB BB	D D D	D ? S	S ? D
		AB	AA AB BB	? ? ?	D ? S	? ? ?
		BB	AA AB BB	S S S	D ? S	D ? S

## BIBLIOGRAPHY

- [1] “Review Manager (RevMan) [Computer program]. Version 5.1. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2011.” In: ().
- [2] S. E. Antonarakis et al. “Chromosome 21 and down syndrome: from genomics to pathophysiology”. In: *Nature Reviews. Genetics* 5.10 (2004), pp. 725–38.
- [3] J. Baudat F Fau Buard et al. “PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice”. In: 1095-9203 (Electronic) (2009).
- [4] F. Begum et al. “Comprehensive literature review and statistical considerations for GWAS meta-analysis”. In: *Nucleic acids research* (2012).
- [5] I. L. Berg et al. “PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans”. In: *Nature genetics* 42.10 (2010), pp. 859–63.
- [6] I. L. Berg et al. “Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.30 (2011), pp. 12378–83.
- [7] J. C. Biancotti et al. “Human embryonic stem cells as models for aneuploid chromosomal syndromes”. In: *Stem cells* 28.9 (2010), pp. 1530–40.
- [8] H. G. Blitzblau and A. Hochwagen. “Genome-wide detection of meiotic DNA double-strand break hotspots using single-stranded DNA”. In: *Methods in molecular biology* 745 (2011), pp. 47–63.

- [9] K. W. Broman et al. “Comprehensive human genetic maps: individual and sex-specific variation in recombination”. In: *American journal of human genetics* 63.3 (1998), pp. 861–9.
- [10] M. Bugge et al. “Non-disjunction of chromosome 18”. In: *Human molecular genetics* 7.4 (1998), pp. 661–9.
- [11] W. S. Bush and J. H. Moore. “Chapter 11: Genome-wide association studies”. In: *PLoS Computational Biology* 8.12 (2012), e1002822.
- [12] R. M. Cantor, K. Lange, and J. S. Sinsheimer. “Prioritizing GWAS results: A review of statistical methods and recommendations for their application”. In: *American journal of human genetics* 86.1 (2010), pp. 6–22.
- [13] V. G. Cheung et al. “Polymorphic variation in human meiotic recombination”. In: *American journal of human genetics* 80.3 (2007), pp. 526–30.
- [14] V. G. Cheung, S. L. Sherman, and E. Feingold. “Genetics. Genetic control of hotspots”. In: *Science* 327.5967 (2010), pp. 791–2.
- [15] T. Chiang, R. M. Schultz, and M. A. Lampson. “Meiotic origins of maternal age-related aneuploidy”. In: *Biology of reproduction* 86.1 (2012), pp. 1–7.
- [16] R. Chowdhury et al. “Genetic analysis of variation in human meiotic recombination”. In: *PLoS genetics* 5.9 (2009), e1000648.
- [17] W. G. Cochran. “The combination of estimates from different experiments.” In: *Biometrics* 10 (1954), pp. 101–129.
- [18] G. Coop and M. Przeworski. “An evolutionary view of human recombination”. In: *Nature reviews. Genetics* 8.1 (2007), pp. 23–34.
- [19] G. Coop et al. “High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans”. In: *Science* 319.5868 (2008), pp. 1395–8.
- [20] D. Curtis, A. E. Vine, and J. Knight. “A simple method for assessing the strength of evidence for association at the level of the whole gene”. In: *Advances and applications in bioinformatics and chemistry : AABC* 1 (2008), pp. 115–20.

- [21] T. R. Dawber, G. F. Meadors, and Jr. Moore F. E. “Epidemiological approaches to heart disease: the Framingham Study”. In: *American journal of public health and the nation’s health* 41.3 (1951), pp. 279–81.
- [22] P. I. de Bakker et al. “Practical aspects of imputation-driven meta-analysis of genome-wide association studies”. In: *Human molecular genetics* 17.R2 (2008), R122–8.
- [23] P. I. de Bakker, B. M. Neale, and M. J. Daly. “Meta-analysis of genome-wide association studies”. In: *Cold Spring Harbor protocols* 2010.6 (2010), pdb top81.
- [24] A. F. Dernburg et al. “Meiotic recombination in *C. elegans* initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis”. In: *Cell* 94.3 (1998), pp. 387–98.
- [25] C. Ellermeier et al. “RNAi and heterochromatin repress centromeric meiotic recombination”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.19 (2010), pp. 8701–5.
- [26] A. Fledel-Alon et al. “Variation in human recombination rates and its genetic determinants”. In: *PLoS ONE* 6.6 (2011), e20321.
- [27] M. H. Gail et al. “Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies”. In: *Biostatistics* 9.2 (2008), pp. 201–15.
- [28] M. H. Gail et al. “Probability that a two-stage genome-wide association study will detect a disease-associated snp and implications for multistage designs”. In: *Annals of human genetics* 72.Pt 6 (2008), pp. 812–20.
- [29] I. P. Gorlov et al. “GWAS meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example”. In: *PLoS ONE* 4.8 (2009), e6511.
- [30] R. Guerra and D. R. Goldstein. *Meta-analysis and Combining Information in Genetics and Genomics*. Mathematical and Computational Biology Series. CRC press, Taylor et al., 2010.
- [31] T. Hassold and P. Hunt. “To err (meiotically) is human: the genesis of human aneuploidy”. In: *Nature reviews. Genetics* 2.4 (2001), pp. 280–91.

- [32] T. Hassold et al. “Cytogenetic and molecular studies of trisomy 13”. In: *Journal of medical genetics* 24.12 (1987), pp. 725–32.
- [33] T. J. Hassold et al. “Molecular studies of non-disjunction in trisomy 16”. In: *Journal of medical genetics* 28.3 (1991), pp. 159–62.
- [34] D. P. Hibar et al. “Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects”. In: *NeuroImage* 56.4 (2011), pp. 1875–91.
- [35] J. P. Higgins and S. G. Thompson. “Quantifying heterogeneity in a meta-analysis”. In: *Statistics in medicine* 21.11 (2002), pp. 1539–58.
- [36] J. P. Higgins et al. “Measuring inconsistency in meta-analyses”. In: *BMJ* 327.7414 (2003), pp. 557–60.
- [37] A. G. Hinch et al. “The landscape of recombination in African Americans”. In: *Nature* 476.7359 (2011), pp. 170–5.
- [38] M. Hirakawa et al. “JSNP: a database of common gene variations in the Japanese population”. In: *Nucleic Acids Res* 30.1 (2002), pp. 158–162.
- [39] J. Hodgkin, H. R. Horvitz, and S. Brenner. “Nondisjunction Mutants of the Nematode *Caenorhabditis Elegans*”. In: *Genetics* 91.1 (1979), pp. 67–94.
- [40] A. J. Iafrate et al. “Detection of large-scale variation in the human genome”. In: *Nature genetics* 36.9 (2004), pp. 949–51.
- [41] A. Iliadis, D. Anastassiou, and X. Wang. “Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled DNA data”. In: *BMC genetics* 13 (2012), p. 94.
- [42] J. P. Ioannidis. “Non-replication and inconsistency in the genome-wide association setting”. In: *Human heredity* 64.4 (2007), pp. 203–13.
- [43] J.P. A. Ioannidis, N. A. Patsopoulos, and E. Evangelou. “Heterogeneity in Meta-analysis of Genome-wide Association Investigations.” In: *PLOS ONE* e841.9 (2007).
- [44] M. Kanehisa and S. Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

- [45] L. Kauppi, A. J. Jeffreys, and S. Keeney. “Where the crossovers are: recombination distributions in mammals”. In: *Nature reviews. Genetics* 5.6 (2004), pp. 413–24.
- [46] S. T. Kim et al. “Prostate cancer risk-associated variants reported from genome-wide association studies: meta-analysis and their contribution to genetic Variation”. In: *The Prostate* 70.16 (2010), pp. 1729–38.
- [47] K. Kimura et al. “Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes”. In: *Genome research* 16.1 (2006), pp. 55–65.
- [48] K. E. Koehler et al. “Spontaneous X chromosome MI and MII nondisjunction events in *Drosophila melanogaster* oocytes have different recombinational histories”. In: *Nature genetics* 14.4 (1996), pp. 406–14.
- [49] A. Kong et al. “Fine-scale recombination rate differences between sexes, populations and individuals”. In: *Nature* 467.7319 (2010), pp. 1099–103.
- [50] A. Kong et al. “Sequence variants in the RNF212 gene associate with genome-wide recombination rate”. In: *Science* 319.5868 (2008), pp. 1398–401.
- [51] M. D. Krawchuk and W. P. Wahls. “Centromere mapping functions for aneuploid meiotic products: Analysis of *rec8*, *rec10* and *rec11* mutants of the fission yeast *Schizosaccharomyces pombe*”. In: *Genetics* 153.1 (1999), pp. 49–55.
- [52] A. W. Kung et al. “Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies”. In: *American journal of human genetics* 86.2 (2010), pp. 229–39.
- [53] C. L. Kuo and E. Feingold. “What’s the best statistic for a simple test of genetic association in a case-control study?” In: *Genetic epidemiology* 34.3 (2010), pp. 246–53.
- [54] N. E. Lamb et al. “Characterization of susceptible chiasma configurations that increase the risk for maternal nondisjunction of chromosome 21”. In: *Human molecular genetics* 6.9 (1997), pp. 1391–9.
- [55] N. E. Lamb et al. “Association between maternal age and meiotic recombination for trisomy 21”. In: *American journal of human genetics* 76.1 (2005), pp. 91–9.

- [56] E. S. Lander and P. Green. “Construction of multilocus genetic linkage maps in humans”. In: *Proceedings of the National Academy of Sciences of the United States of America* 84.8 (1987), pp. 2363–7.
- [57] Green Lander. *CRIMAP*. 1987.
- [58] J. Lau, J.P. Ioannidis, and C.H. Schmid. “Quantitative synthesis in systematic reviews.” In: *Ann Intern Med* 126 (1997), pp. 820–826.
- [59] J. Lau, J. P. Ioannidis, and C. H. Schmid. “Summing up evidence: one answer is not always enough”. In: *Lancet* 351.9096 (1998), pp. 123–7.
- [60] C. C. Laurie et al. “Quality control and quality assurance in genotypic data for genome-wide association studies”. In: *Genetic epidemiology* 34.6 (2010), pp. 591–602.
- [61] J. Li and G. C. Tseng. “An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies.” In: *The Annals of Applied Statistics* 5.2A (2011), pp. 994–1019.
- [62] M. Li et al. “ATOM: a powerful gene-based association test by combining optimally weighted markers”. In: *Bioinformatics* 25.4 (2009), pp. 497–503.
- [63] M. X. Li et al. “IGG3: a tool to rapidly integrate large genotype datasets for whole-genome imputation and individual-level meta-analysis”. In: *Bioinformatics* 25.11 (2009), pp. 1449–50.
- [64] M. X. Li, J. S. Kwan, and P. C. Sham. “HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis”. In: *American journal of human genetics* 91.3 (2012), pp. 478–88.
- [65] D. Y. Lin and P. F. Sullivan. “Meta-analysis of genome-wide association studies with overlapping subjects”. In: *American journal of human genetics* 85.6 (2009), pp. 862–72.
- [66] D. Y. Lin and D. Zeng. “Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data”. In: *Genetic epidemiology* 34.1 (2010), pp. 60–6.
- [67] Y. C. Lin et al. “Using maximal segmental score in genome-wide association studies”. In: *Genetic epidemiology* 36.6 (2012), pp. 594–601.



- [68] A. J. Lorenz, M. T. Hamblin, and J. L. Jannink. “Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley”. In: *PLoS ONE* 5.11 (2010), e14079.
- [69] L. Ma, A. G. Clark, and A. Keinan. “Gene-based testing of interactions in association studies of quantitative traits”. In: *PLoS genetics* 9.2 (2013), e1003321.
- [70] I. Maayan. “”Meiosis in Humans””. In: *Embryo Project Encyclopedia* (2011).
- [71] R. Magi and A. P. Morris. “GWAMA: software for genome-wide association meta-analysis”. In: *BMC bioinformatics* 11 (2010), p. 288.
- [72] J. Marchini and B. Howie. “Genotype imputation for genome-wide association studies”. In: *Nature reviews. Genetics* 11.7 (2010), pp. 499–511.
- [73] R. H. Martin and A. W. Rademaker. “The effect of age on the frequency of sperm chromosomal abnormalities in normal men”. In: *American journal of human genetics* 41.3 (1987), pp. 484–92.
- [74] K. M. May et al. “The parental origin of the extra X chromosome in 47,XXX females”. In: *American journal of human genetics* 46.4 (1990), pp. 754–61.
- [75] C. Minelli et al. “The choice of a genetic model in the meta-analysis of molecular association studies”. In: *International journal of epidemiology* 34.6 (2005), pp. 1319–28.
- [76] C. Minelli et al. “How should we use information about HWE in the meta-analyses of genetic association studies?” In: *International journal of epidemiology* 37.1 (2008), pp. 136–46.
- [77] J. H. Moore and M. D. Ritchie. “STUDENTJAMA. The challenges of whole-genome approaches to common diseases”. In: *JAMA : the journal of the American Medical Association* 291.13 (2004), pp. 1642–3.
- [78] M. R. Munafo and J. Flint. “Meta-analysis of genetic association studies”. In: *Trends in genetics : TIG* 20.9 (2004), pp. 439–44.
- [79] S. Myers et al. “Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination”. In: *Science* 327.5967 (2010), pp. 876–9.

- [80] H. Nakaoka and I. Inoue. “Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner’s curse”. In: *Journal of human genetics* 54.11 (2009), pp. 615–23.
- [81] M. J. Neale. “PRDM9 points the zinc finger at meiotic recombination hotspots”. In: *Genome biology* 11.2 (2010), p. 104.
- [82] C. O’Connor. “Meiosis, Genetic Recombination, and Sexual Reproduction”. In: *Nature Education* 1.1 (2008).
- [83] N. A. Patsopoulos and J. P. Ioannidis. “Susceptibility variants for rheumatoid arthritis in the TRAF1-C5 and 6q23 loci: a meta-analysis”. In: *Annals of the rheumatic diseases* 69.3 (2010), pp. 561–6.
- [84] S. A. Pendergrass et al. “Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis”. In: *BioData mining* 3 (2010), p. 10.
- [85] T. V. Pereira et al. “Discovery properties of genome-wide association signals from cumulatively combined data sets”. In: *American journal of epidemiology* 170.10 (2009), pp. 1197–206.
- [86] M. Petronczki, M. F. Siomos, and K. Nasmyth. “Un menage a quatre: the molecular biology of chromosome segregation in meiosis”. In: *Cell* 112.4 (2003), pp. 423–40.
- [87] S. Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. In: *American journal of human genetics* 81.3 (2007), pp. 559–75.
- [88] H. Q. Qu et al. “In silico replication of the genome-wide association results of the Type 1 Diabetes Genetics Consortium”. In: *Human molecular genetics* 19.12 (2010), pp. 2534–8.
- [89] W. P. Robinson et al. “Maternal meiosis I non-disjunction of chromosome 15: dependence of the maternal age effect on level of recombination”. In: *Human molecular genetics* 7.6 (1998), pp. 1011–9.
- [90] L. O. Ross, R. Maxfield, and D. Dawson. “Exchanges are not equally able to enhance meiotic chromosome segregation in yeast”. In: *Proceedings of the National Academy of Sciences of the United States of America* 93.10 (1996), pp. 4979–83.

- [91] G. Salanti, S. Sanderson, and J. P. Higgins. “Obstacles and opportunities in meta-analysis of genetic association studies”. In: *Genetics in medicine : official journal of the American College of Medical Genetics* 7.1 (2005), pp. 13–20.
- [92] Carmen Sandovici I Fau Sapienza and C. Sapienza. “PRDM9 sticks its zinc fingers into recombination hotspots and between species. LID - 37 [pii]”. In: 1757-594X (Electronic) (2010).
- [93] S. Sarbajna et al. “A major recombination hotspot in the XqYq pseudoautosomal region gives new insight into processing of human gene conversion events”. In: *Human molecular genetics* (2012).
- [94] A. R. Savage et al. “Elucidating the mechanisms of paternal non-disjunction of chromosome 21 in humans”. In: *Human molecular genetics* 7.8 (1998), pp. 1221–7.
- [95] A. V. Segre et al. “Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glyceic traits”. In: *PLoS genetics* 6.8 (2010).
- [96] L. Segurel, E. M. Leffler, and M. Przeworski. “The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans”. In: *PLoS biology* 9.12 (2011), e1001211.
- [97] Q. Sha, R. Tang, and S. Zhang. “Detecting susceptibility genes for rheumatoid arthritis based on a novel sliding-window approach”. In: *BMC proceedings* 3 Suppl 7 (2009), S14.
- [98] J. R. Shaffer et al. “Genome-wide association scan for childhood caries implicates novel genes”. In: *Journal of dental research* 90.12 (2011), pp. 1457–62.
- [99] A. J. Sutton et al. *Methods for meta-analysis in medical research*. John Wiley and Sons, Chichester, 2000.
- [100] N. Takaesu et al. “Nondisjunction of chromosome 21”. In: *American journal of medical genetics. Supplement* 7 (1990), pp. 175–81.
- [101] R. Tang et al. “A variable-sized sliding-window approach for genetic association studies via principal component analysis”. In: *Annals of human genetics* 73.Pt 6 (2009), pp. 631–7.

- [102] A. Thakkinstian et al. “A method for meta-analysis of molecular association studies”. In: *Statistics in medicine* 24.9 (2005), pp. 1291–306.
- [103] N. S. Thomas et al. “Maternal sex chromosome non-disjunction: evidence for X chromosome-specific risk factors”. In: *Human molecular genetics* 10.3 (2001), pp. 243–50.
- [104] J. R. Thompson, J. Attia, and C. Minelli. “The meta-analysis of genome-wide association studies”. In: *Briefings in bioinformatics* 12.3 (2011), pp. 259–69.
- [105] S. G. Thompson. “Why sources of heterogeneity in meta-analysis should be investigated”. In: *BMJ* 309.6965 (1994), pp. 1351–5.
- [106] L.H.C. Tippett. *The Methods in Statistics*. 1st ed. Williams and Norgate, Ltd., 1931.
- [107] Wolfgang Viechtbauer. “Conducting meta-analyses in R with the metafor package.” In: *Journal of Statistical Software* 36.3 (2010), pp. 1–48.
- [108] A. M. Villeneuve. “A cis-acting locus that promotes crossing over between X chromosomes in *Caenorhabditis elegans*”. In: *Genetics* 136.3 (1994), pp. 887–902.
- [109] X. Wan et al. “HapBoost: A fast Approach to Boosting Haplotype Association Analyses in Genome-Wide Association Studies”. In: *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* (2013).
- [110] L.A. Weiss et al. “Association between microdeletion and microduplication at 16p11.2 and autism.” In: *N Engl J Med* 358.7 (2008), pp. 737–9.
- [111] M. C. Whitlock. “Combining probability from independent tests: the weighted Z-method is superior to Fishers approach”. In: *J. Evol. Biol* 18.2005 (2005), pp. 1368–1373.
- [112] C. J. Willer, Y. Li, and G. R. Abecasis. “METAL: fast and efficient meta-analysis of genomewide association scans”. In: *Bioinformatics* 26.17 (2010), pp. 2190–1.
- [113] Wu Y-w et al. “Analysis of Lingo1 variant in sporadic and familial essential tremor among Asians”. In: *Acta Neurologica Scandinavica* (2010).
- [114] K. Yu et al. “Pathway analysis by adaptive combination of P-values”. In: *Genetic epidemiology* 33.8 (2009), pp. 700–9.

- [115] E. Zeggini and J. P. Ioannidis. “Meta-analysis in genome-wide association studies”. In: *Pharmacogenomics* 10.2 (2009), pp. 191–201.
- [116] M. C. Zetka et al. “Synapsis and chiasma formation in *Caenorhabditis elegans* require HIM-3, a meiotic chromosome core component that functions in chromosome segregation”. In: *Genes development* 13.17 (1999), pp. 2258–70.
- [117] M. C. Zetka and A. M. Rose. “The meiotic behavior of an inversion in *Caenorhabditis elegans*”. In: *Genetics* 131.2 (1992), pp. 321–32.
- [118] F. Zhang and R. Drabier. “IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis”. In: *BMC bioinformatics* 13 Suppl 15 (2012), S7.
- [119] J. Zhang et al. “Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome”. In: *Cytogenetic and genome research* 115.3-4 (2006), pp. 205–14.
- [120] E. Zintzaras and J. Lau. “Synthesis of genetic association studies for pertinent gene-disease associations requires appropriate methodological and statistical approaches”. In: *Journal of clinical epidemiology* 61.7 (2008), pp. 634–45.
- [121] E. Zintzaras and J. Lau. “Trends in meta-analysis of genetic association studies”. In: *Journal of human genetics* 53.1 (2008), pp. 1–9.