Association of polygenic risk scores with breast cancer

Sara Urru

University of Padua
PhD in Traslational Specialistic Medicine
Biostatistics and Clinical Epidemiology

YSM 2021

Index

*This thesis was realized during an internship at the *Genomic Lab - Edo &d Elvo Tempia Foundation*.

Introduction: ANDROMEDA

ANDROMEDA [1] is a multicentre prospective cohort study on women from Northern Italy, aged $46 - 67$ y.o, attending breast cancer screening. They were asked to provide the following:

SRQ - Short risk questionnaire	reproductive, hormonal, personal and familiar history
LRQ - Long risk questionnaire	diet, physical activity, smoking abits, psychological distress
Anthropometric measurements	height, weight, body composition, waist circumference
Blood sample	micro-RNA, SNPs

[1] Giordano, Livia. et al. "The ANDROMEDA prospective cohort study: predictive value of combined criteria to tailor breast cancer screening and new opportunities from circulating markers: study protocol." BMC cancer vol. 17,1 785. 22 Nov. 2017, doi:10.1186/s12885-017-3784-5

Aim

To define an appropriate women risk-based stratification for personalized screening considering different criteria such as:

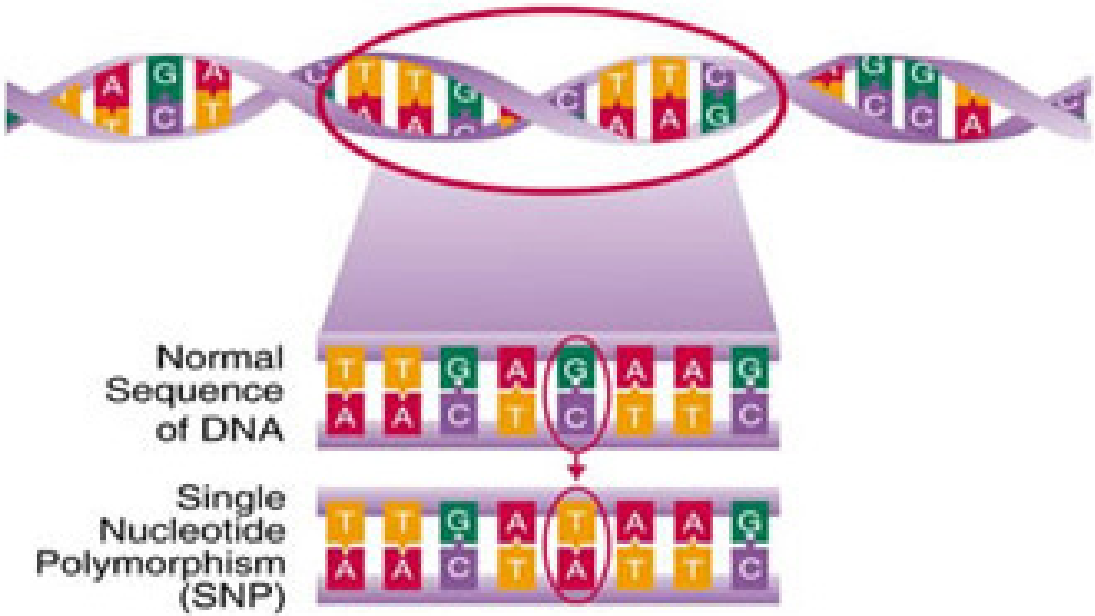
- genetics;
- anthropometric measurements;
- hormonal and reproductive history;
- personal and familiar history;
- lifestyle habits.

Study design

- A case-control study was nested in the cohort.
- Association between genetics and BC was analysed.
- Data from DNA sequencing and from the SRQ were considered.

Basic concepts: SNP

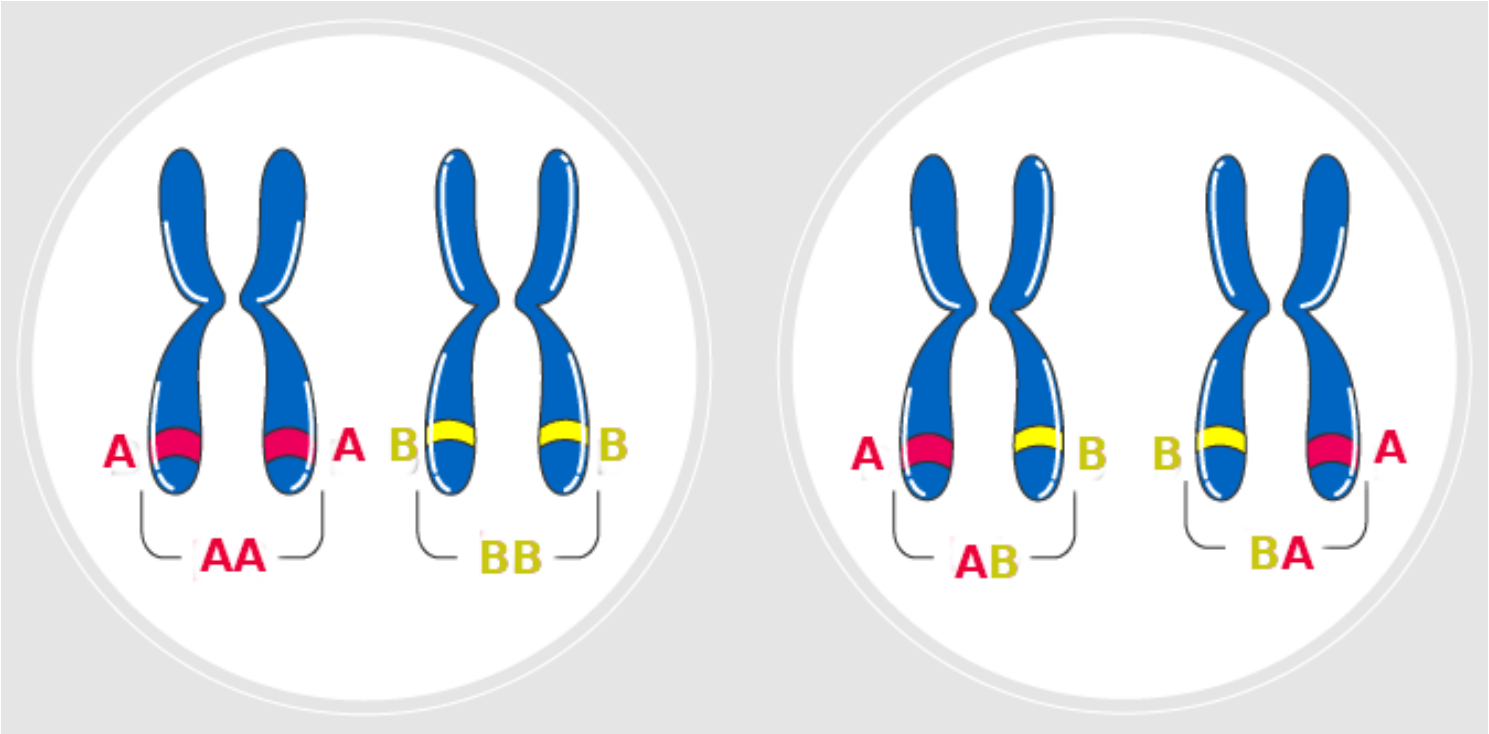
A single nucleotide polymorphism (SNP) is a variation of one nucleotide in the DNA. The expected bases in a specific locus is defined as the reference base, the variant, instead, is defined as the alternate base. SNPs occur at least in 1% of the population.



Basic concepts: genotype

A SNP can be present in one or in both alleles of a chromosome; this is indicated with the term *genotype* which is called

- *wild type* (0) if the variant base is absent;
- *heterozygous* (1) if the variant base is only on one of the two alleles;
- *homozygous* (2) if the variant base is on both alleles.



Basic concepts: PRS

- Polygenic risk score (PRS) summarises the combined effect of many genetic variants.
- *Mavaddat et al. (2015)* [2] developed a PRS to study the association between breast cancer risk and the joined effect of 77 SNPs on a cohort of $\sim$ 67000 European women.

[2] Mavaddat, Nasim et al. "Prediction of breast cancer risk based on profiling with common genetic variants." Journal of the National Cancer Institute vol. 107,5 djv036. 8 Apr. 2015, doi:10.1093/jnci/djv036

Basic concepts: PRS

PRS is calculated for every individual as:

$$PRS = \beta_1 x_1 + \cdots + \beta_n x_n \tag{1}$$

- β_i is the log-odds ratio for the SNP i;
- $x_i = \{0, 1, 2\}$ is the genotype of the SNP i;
- $n = 77$ is the total number of SNPs.

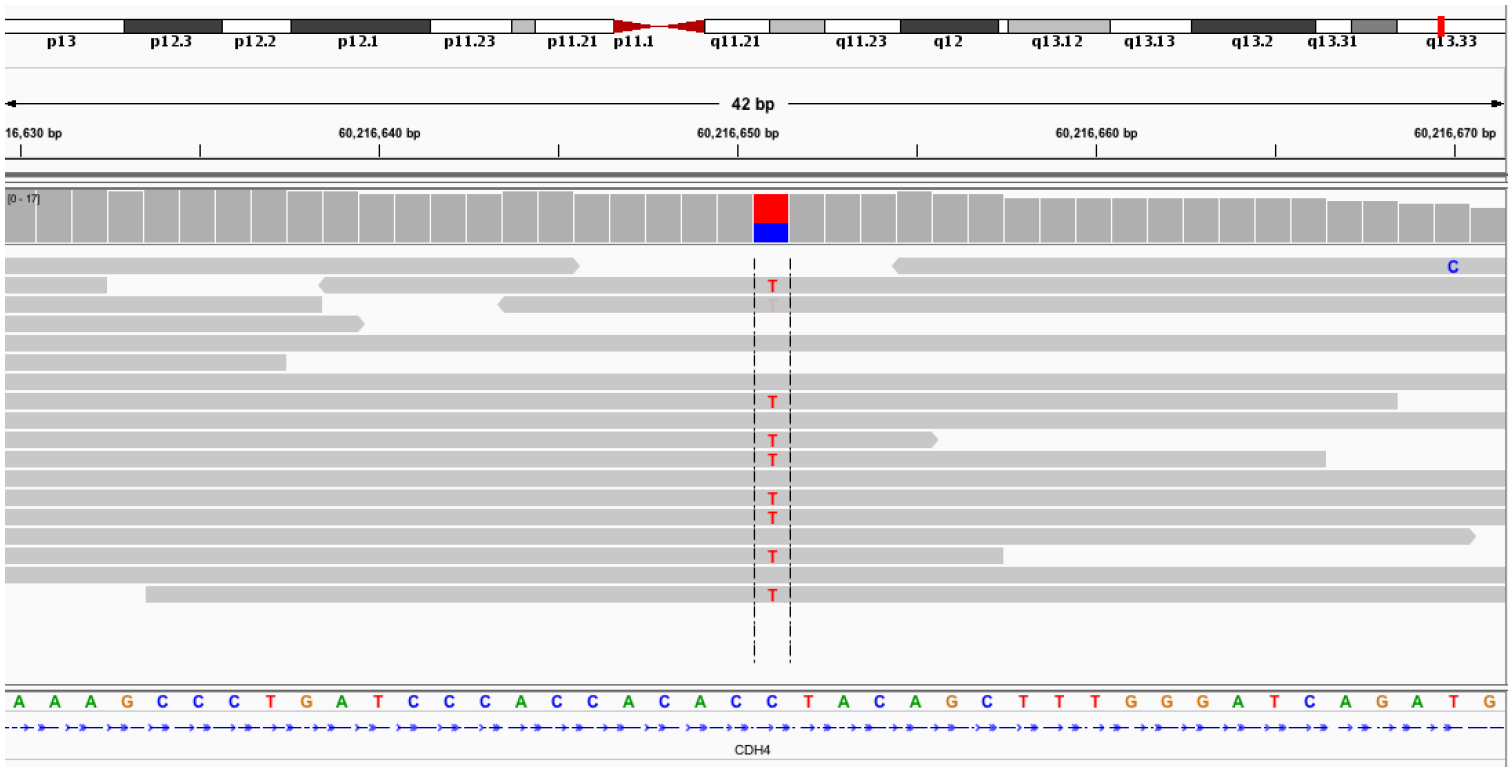
PRS has a normal distribution in the population with mean and variance

$$\mu = 2\sum_{i=1}^{n} p_i \beta_i \qquad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2 = 2\sum_{i=1}^{n} p_i q_i \beta_i^2 \tag{2}$$

where p_i is the population minor allele frequency (MAF) of SNP i and $q_i = 1 - p_i$.

Data collection: genetic data

- DNA was extracted from buffy-coat of 384 women.
- DNA was sequenced using the next generation sequencing method.
- 80 SNPs were evaluated.
- BAM files were obtained for every sample.



Data collection: variant call process

- BAM files were processed using the *variant call* algorithm to obtain the genotype of the SNPs.
- The result of this process is a VCF file containing SNPs information for each sample.
- The final dataset includes information about the quality of 80 for the 384 women.
- The genotype is identified if some filter parameters are satisfied, otherwise the genotype is missing.

Statistical methods: imputation

Missing genotype data are known as *No call* and due to poor quality sample and sequencing issues.

Genotype data can be missing at random because of the dependence on other variables such as:

- quality (Phred quality score $Q = -10log_{10}(P)$);
- coverage, total number of reads aligned;
- allele coverage, total number of reads aligned containing the variant;
- strand bias, bias due to the alignment of positive and negative strands;
- signal shift, shift between predicted and observed allele.

Statistical methods: imputation

- Missing genotypes were imputed using *multinomial logistic regression*.
- For every SNP data were divided in complete and missing set.
- Complete data were splitted in training set (65%) and testing set (35%).
- Cases and controls were balanced.
- The multinomial regression model was applied on testing sets to get the imputation errors and on missing sets to impute the missing values.

Summary of imputation error

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.000	0.000	0.000	0.005	0.008	0.043

Statistical methods: multinomial logistic regression

Multinomial logistic regression is a logistic model where the outcome variable has more than 2 levels. Let Y be the outcome variable which assumes three possible values coded $0, 1$ and 2 and $X = (x_1, \ldots, x_p)$ a vector of p independent variables. The model needs two logit functions:

$$\begin{aligned} g_1(x) &= \log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p \\ g_2(x) &= \log \frac{\mathbb{P}(Y = 2|X = x)}{\mathbb{P}(Y = 0|X = x)} = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p \end{aligned} \tag{3}$$

The conditional probabilities of each outcome class are:

$$\begin{aligned} \mathbb{P}(Y = 0|X = x) &= \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}} \\ \mathbb{P}(Y = 1|X = x) &= \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}} \\ \mathbb{P}(Y = 2|X = x) &= \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}} \end{aligned} \tag{4}$$

Statistical methods: cross-validation

- Cross-validation was used to calculate the coefficients of the PRS and to evaluate the performance average of the models.
- Cross-validation is a procedure which allows to derive training and testing sets from the same data set.
- *k-fold cross-validation* with $k = 10$ was used.
- In 10-fold cross-validation, the starting data set D is partitioned in 10 subsets $S_1, \ldots, S_{10}$ and for $i = 1, \ldots, 10$:
  - S_i = testing set;
  - $D \setminus S_i$ = training set.

Statistical methods: model selection and performance

- Stepwise logistic regression was applied to select the variables to include in the model.
  - Bidirectional elimination was used.
  - The best model was chosen according to the AIC criterion.
- Receiver operating curve (ROC) and area under the curve (AUC) were calculated:
  - To evaluate the performance of the models.
  - To compare the models.

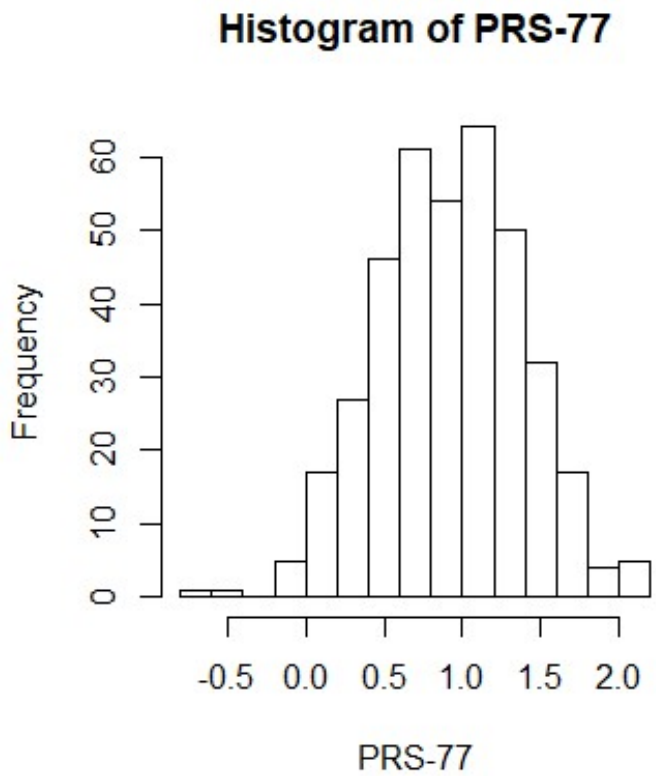
Results: PRS

Genotypes of 80 SNPs for 384 women (115 cases and 269 controls) were obtained and two PRS were computed:

- PRS-77 using the log-odds found in literature;
- PRS-80 using the log-odds derived from our data and including 3 more SNPs associated with breast cancer prognosis.

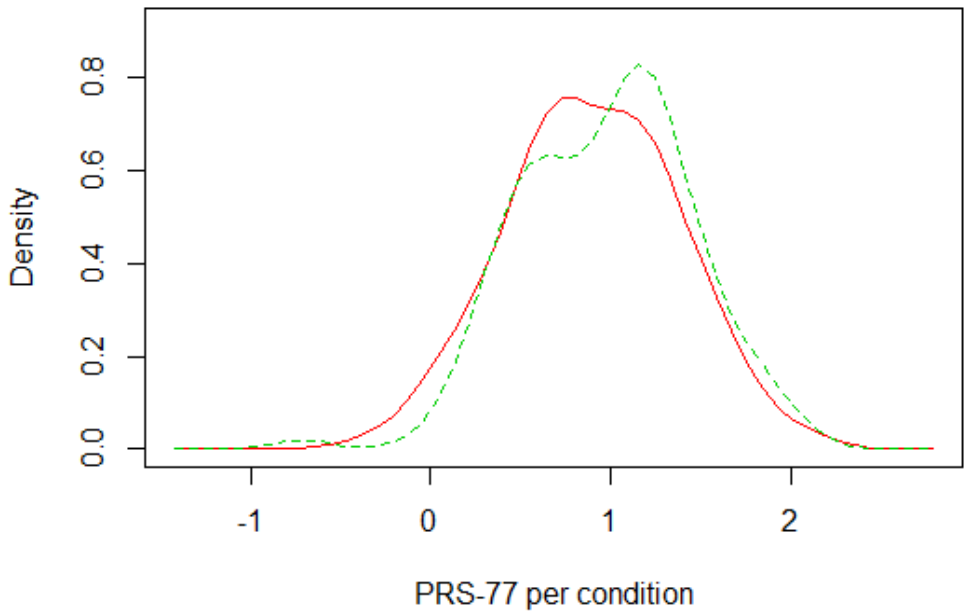
Results: PRS-77

	Mean	SD	Median	Min	Max
Overall	0.918	0.468	0.925	-0.715	2.082
Cases	0.972	0.465	1.023	-0.715	2.043
Controls	0.895	0.467	0.847	-0.426	2.082



Results: PRS-77 cases vs controls

The differences between cases (green) and controls (red) did not result statistically significant.



	OR	CI	p
(Intercept)	0.31	0.18 – 0.50	<0.001
PRS-77	1.43	0.89 – 2.31	0.138

Results: *SNP Class*

Since PRS-77 does not allow to detect differences between cases and controls, we decide to classify samples using the sign of variants. For every SNP i and sample j we defined

$$v_{ij} = \begin{cases} 0 & G = 0 \\ 1 & G \neq 0 \end{cases}, \qquad RP_j = \frac{1}{47}\sum_{i=1}^{47} v_{ij}, \qquad RN_j = \frac{1}{30}\sum_{i=1}^{30} v_{ij} \tag{5}$$

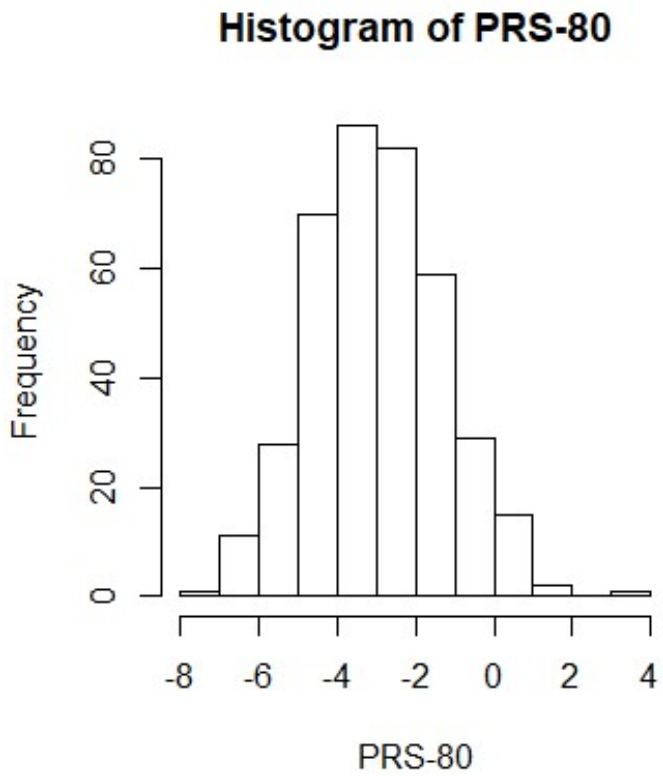
$$SC_j = \begin{cases} 0 & RP_j < RN_j \\ 1 & RP_j \geq RN_j \end{cases} \tag{6}$$

To assess the validity of this new classification the logistic regression was performed:

	OR	**CI**	**p**
(Intercept)	0.34	0.25 – 0.44	<0.001
SNP Class	1.99	1.26 – 3.14	0.003

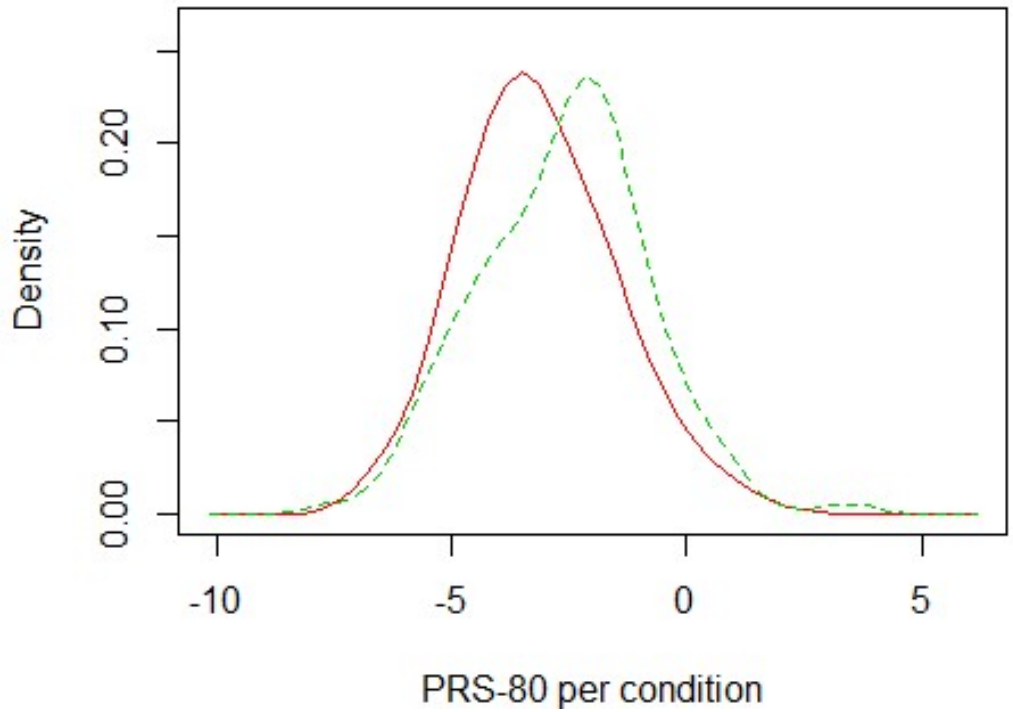
Results: PRS-80

	Mean	SD	Median	Min	Max
Overall	-2.974	1.674	-3.046	-7.429	3.451
Cases	-2.599	1.767	-2.432	-7.429	3.451
Controls	-3.135	1.609	-3.300	-6.813	1.922



Results: PRS-80 cases vs controls

The following Figure shows that PRS-80 density in cases (green) is slightly shifted to the right with respect to controls (red) as theory predicts.



	OR	CI	p
(Intercept)	0.74	0.48 – 1.14	0.177
PRS-80	1.21	1.06 – 1.39	0.004

Results: description of the sample

		Controls (n=269)	Cases (n=115)	p
Age (mean (SD))		58.07 (5.98)	59.66 (6.22)	0.019
BMI (%)	<25.00	173 (64.3)	59 (51.3)	0.044
	25.00-29.99	62 (23.0)	33 (28.7)	
	≥30.00	34 (12.6)	23 (20.0)	
Menopause (%)	No	51 (19.0)	22 (19.1)	1.000
	Yes	218 (81.0)	93 (80.9)	
MHT (%)	No	259 (96.3)	106 (92.2)	0.149
	Yes	10 (3.7)	9 (7.8)	
Children (%)	No	57 (21.2)	16 (13.9)	0.128
	Yes	212 (78.8)	99 (86.1)	
Family history (%)	No	241 (89.6)	100 (87.0)	0.566
	Yes	28 (10.4)	15 (13.0)	
Previous biopsies (%)	No	234 (87.0)	93 (80.9)	0.165
	Yes	35 (13.0)	22 (19.1)	
Age at menarche (%)	<12	73 (27.1)	34 (29.6)	0.853
	12-13	132 (49.1)	56 (48.7)	
	>13	64 (23.8)	25 (21.7)	
Education (%)	Primary school	88 (32.7)	45 (39.1)	0.211
	High school	134 (49.8)	46 (40.0)	
	University	47 (17.5)	24 (20.9)	
Physical activity at work (%)	Sitting	81 (30.1)	44 (38.3)	0.373
	Medium	79 (29.4)	34 (29.6)	
	Standing	71 (26.4)	24 (20.9)	
	Tiring	38 (14.1)	13 (11.3)	
Physical activity in the free time (%)	<2h per week	131 (48.7)	66 (57.4)	0.296
	2h-4h per week	99 (36.8)	35 (30.4)	
	>5h per week	39 (14.5)	14 (12.2)	
Alcohol consumption (%)	Never	81 (30.1)	34 (29.6)	0.986
	In the past	10 (3.7)	4 (3.5)	
	Occasionally	178 (66.2)	77 (67.0)	

Results: multivariable models

Stepwise logistic regression was performed on the standard risk factors to select the most explicative predictors using the AIC criterion.

Predictors	OR	CI
(Intercept)	**0.05**	**0.00 – 0.53**
Age	**1.04**	**1.00 – 1.08**
BMI: 25.00-29.99	1.48	0.86 – 2.52
BMI: $\geq$30	**1.98**	**1.04 – 3.72**
Education: High school	**0.58**	**0.33 – 1.00**
Education: University	1.11	0.57 – 2.18
MHT: Yes	**2.81**	**1.05 – 7.50**
Physical activity at work: Medium	0.68	0.38 – 1.21
Physical activity at work: Standing	**0.48**	**0.26 – 0.90**
Physical activity at work: Tiring	0.47	0.20 – 1.04
AIC		**465.22**

Results: multivariable models

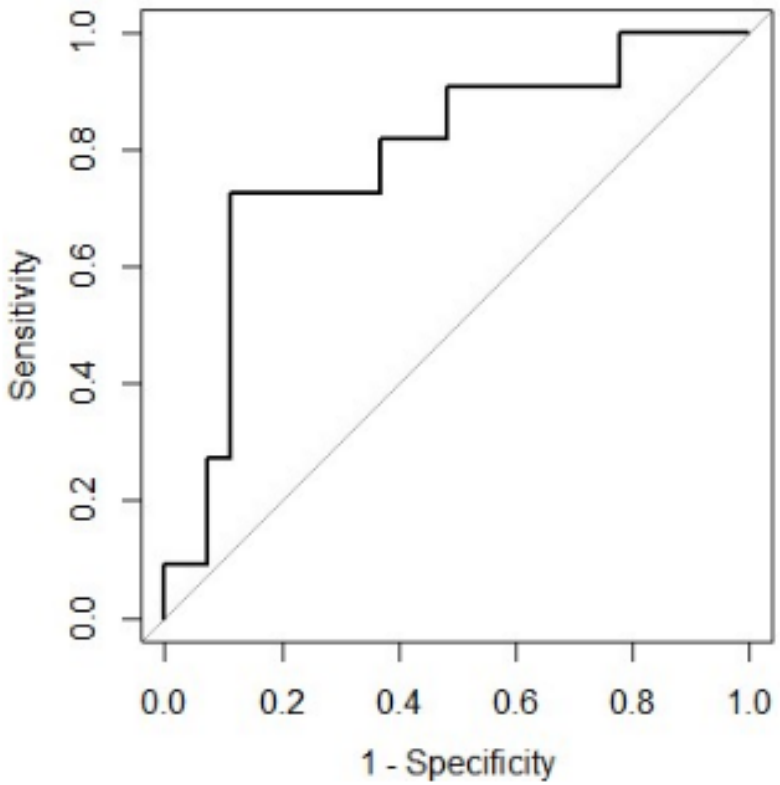
We compared the risk models with and without the genetic component: 10-fold cross-validation was applied and the average AUC computed.

Model	**AUC**
Step.model	0.6161
Step.model + PRS-77	0.6164
Step.model + PRS-80	0.6396
Step.model + SNP Class	0.6247

Results: multivariable models

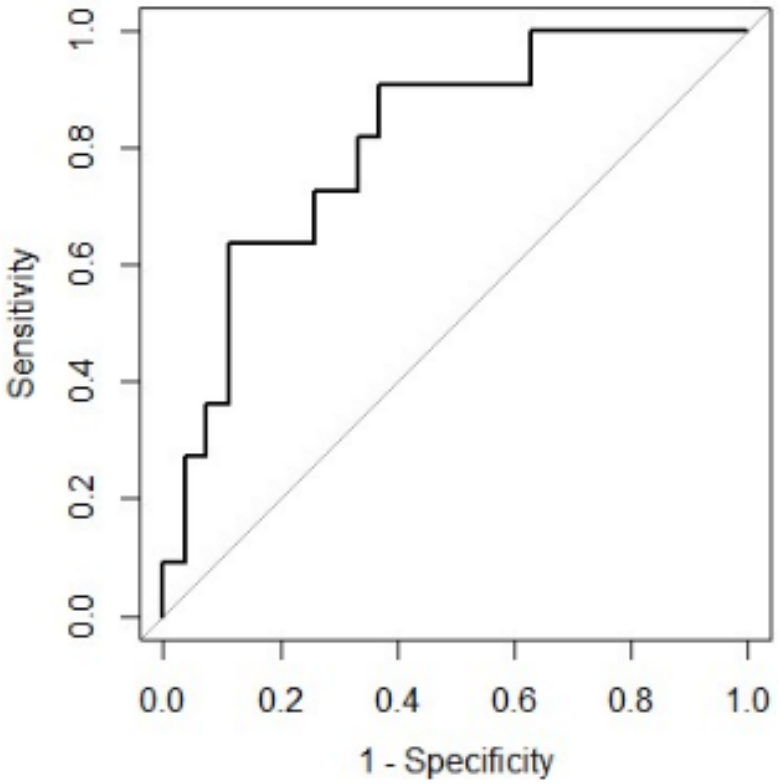
Step.model

AUC=0.788



Step.model+PRS-77

AUC=0.811



Results: model selection

We repeated the stepwise logistic regression including the three genetic components which were all selected and compared the models using AIC.

Selected variables by stepwise regression	**AIC**
Age, BMI, Education, MHT, Physical activity at work	465.22
Age, BMI, Education, MHT, Physical activity at work, PRS-77	464.78
Age, MHT, PRS-80	460.21
Age, BMI, Education, MHT, Physical activity at work, SNP Class	**459.10**

Results: interaction terms

We considered the interaction terms between *SNP Class* and the other variables: only the interaction with BMI resulted significant.

Predictors	OR	CI
(Intercept)	**0.03**	**0.00 – 0.33**
SNP Class: 1	**3.46**	**1.82 – 6.61**
BMI: 25.00-29.99	**2.17**	**1.07 – 4.37**
BMI: $\geq$30	**3.18**	**1.44 – 6.99**
Age	**1.04**	**1.00 – 1.09**
Education: High school	**0.57**	**0.32 – 1.00**
Education: University	1.13	0.56 – 2.24
MHT: Yes	**2.76**	**1.01 – 7.57**
Physical activity at work: Medium	0.72	0.40 – 1.30
Physical activity at work: Standing	**0.46**	**0.24 – 0.87**
Physical activity at work: Tiring	**0.49**	**0.21 – 1.11**
SNP Class: 1* BMI: 25.00-29.99	**0.32**	**0.11 – 0.96**
SNP Class: 1* BMI: $\geq$30	**0.25**	**0.06 – 0.99**

Discussion

Liu et al. (2018) [3] found that:

- $5 kg/m^2$ increase in BMI corresponds to a 2% increase in BC risk;
- higher BMI can be a protective factor in breast cancer risk for premenopausal women.

[3] Liu, Kang et al. "Association between body mass index and breast cancer risk: evidence based on a dose-response meta-analysis." Cancer management and research vol. 10 143-151. 18 Jan. 2018, doi:10.2147/CMAR.S144619

Conclusion

From these preliminarly analyses we can conclude:

- age still remain a good risk indicator;
- the genetic component always improves the risk model;
- BMI is a factor to keep under control.

Developments of this thesis could be

- the inclusion of other factors collected in *ANDROMEDA*;
- the enlargement of the sample size;
- the sequencing of additional SNPs.

Acknowledgments

- Maria Teresa Giraudo from *Department of Mathematics, University of Turin, Italy*
- Giovanna Chiorino, Paola Ostano, Maurizia Mello-Grand, Ilaria Gregnanin from the *Genomic Lab - Edo & Elvo Tempia Foundation, Biella, Italy*
- Livia Giordano, Nereo Segnan, Federica Gallo from the *Centre for Cancer Prevention (CPO Piemonte), Unit of Epidemiology and Screening, AOU Città della Salute e della Scienza of Turin, Italy*
- Elisabetta Petracci from *Unity of Biostatistics and Clinical Trials, Istituto Scientifico Romagnolo per lo Studio e Cura dei Tumori, IRCCS, Meldola, Italy*

Thank you for listening!