bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Haplotypes associated to gene expression in breast cancer: can they lead us to the susceptibility markers?

Hege Edvardsen[1,2], Bettina Kulle[3], Anya Tsalenko[4], Grethe Irene Grenaker Alnæs[1], Fredrik Ekeberg Johansen[1], Espen Enerly[1,2], Aditya Vailaya[4], Per-Eystein Lønning[5], Åslaug Helland[6], Ann-Christine Syvänen[7], Zohar Yakhini[4], Anne-Lise Børresen-Dale[1,2], Arnoldo Frigessi[3] and Vessela N. Kristensen[1]

1 Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Center, Oslo, Norway, 2 Medical Faculty, University of Oslo, Oslo, Norway, 3 Department of Biostatistics, University of Oslo, Norway, 4 Agilent Technologies, Santa Clara, CA, USA, 5 Department of Oncology, Haukeland Hospital, Bergen, Norway, 6 Department Medical Oncology and Radiotherapy, Rikshospitalet-Radiumhospitalet Medical Center, Oslo, Norway, 7 Department of Medical Sciences, University of Uppsala, Uppsala, Sweden

Correspondence should be addressed to:
Vessela N. Kristensen
Department of Cancer Genetics
Institute for Cancer Research
Oslo University Hospital, Radiumhospitalet
Montebello
0310 Oslo
Norway
e-mail: v.n.kristensen@medisin.uio.no

Running title: Screening for Intergenic Haplotype Structures within Given Functional Pathways

Keywords: haplotypes, breast cancer, Gene ontology analysis, Phase, R

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Abstract

We have undertaken a systematic haplotype analysis of the positional type of biclusters analysing samples collected from 164 breast cancer patients and 86 women with no known history of breast cancer. We present here the haplotypes and LD patterns in more than 80 genes distributed across all chromosomes and how they differ between cases and controls. We aim by this to 1) identify genes with different haplotype distribution or LD patterns between breast cancer patients and controls and 2) to evaluate the intratumoral mRNA expression patterns in breast cancer associated particularly to the cancer susceptibility haplotypes. A significant difference in haplotype distribution between cases and controls was observed for a total of 35 genes including *ABCC1*, *AKT2*, *NFKB1*, *TGFBR2* and *XRCC4*. In addition we see a negative correlation between LD patterns in cases and controls for neighboring markers in 8 genes such as *CDKN1A*, *EPHX1* and *XRCC1*.

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Introduction

The common disease common variant hypothesis is the foundation for large scale whole genome analyses of extensive population cohorts aiming at identifying low penetrant markers that in concert result in an increased risk of cancer. Nearly 1300 published GWAS studies have so far identified 6551 markers associated with various diseases and traits such as asthma, multiple sclerosis and various cancer types (*1*). For breast cancer specifically SNPs in 41 genes including *FGFR2*, *TOX3*, *TERT* and *ERBB4* have been associated to the disease (*2-5*). These studies view the risk of common genetic variation only and the number of markers is restricted to the number of SNPs on the studied arrays without focus on particular genes or functionality. Here we have taken an alternative route based on a candidate gene approach without restriction to frequency. Moreover, we did not study single disease associated SNPs but looked for differences in haplotype distribution and LD patterns between cases and controls. Linkage disequilibrium (LD) is the association between two markers (SNPs) resulting from common inheritance of two typically nearby loci. LD is eroded by mutations, gene conversions and recombination events, and is influenced by the age of the mutations as well as the history and size of the populations in which they are studied. Several measurements are used to estimate LD such as D' (*6*) and $r^2$ (*7*). D' shows larger variability within and between populations and is more influenced by sample size (8,9). D' takes into account the history of the markers and is more robust with regards to

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

frequency while $r^2$ is less affected by problems related to sampling (*8*). It is also possible to use statistical estimates of population recombination rates (ρ) instead of pairwise measures of LD (*9*). This measure correlates well across populations and relates the LD pattern directly to the underlying recombination process (*7*). Haplotypes are strings or combinations of co-inherited SNPs residing at regions of high LD and separated by areas with high recombination and low LD (*8*). They are inherited from parents as a single unit and tend to break at recombination hotspots (3). In population studies in contrast to linkage analysis in families, an absolute determination of haplotypes is not possible, but studies of phased estimations have proven these to be a very good approximation. The results from these studies indicate an error in assigning phase to genotypes of approximately 5 % in unrelated individuals (*10*). This uncertainty can be adjusted for as we have previously described (*11*).

The choice of LD block was motivated by our studies in eQTLs. Our findings indicate that the breast cancer risk variants found by the GWASs may exert their effect through the regulation of expression, and that the genes harboring these risk variants are significantly differentially expressed between the well established breast cancer subtypes (12). Given the significant role mRNA expression patterns play in the development of breast cancer, we hypothesize that SNPs associated to clusters of deregulated co-expressed mRNA transcripts may lead us to novel susceptibility markers. We have previously described that among 583 candidate SNPs in 203 genes of the reactive oxygen species metabolism/signaling, there are SNPs significantly

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

associated over the random to the expression of subsets of unselected transcripts in the tumor of breast cancer (13). Furthermore, these subsets of transcripts are enriched for given functional pathways also over the random. Multiple SNPs (biclustes) that together share significantly many common associations to a set of transcripts were identified. These biclusters were either located in different genes on different chromosomes, suggesting a multi-locus regulatory effect on a pathway (functional biclusters) or clustered in the same gene or chromosomal region (positional biclusters). With the present study we have undertaken a systematic haplotype analysis of these positional biclusters extending the analysis to samples collected from 164 breast cancer patients and 86 women with no known history of breast cancer. We aim by this study to 1) formally assess the degree of LD between the SNPs in the positional biclusters associated to expression 2) use these eQTL hits to identify cancer susceptibility haplotypes by comparing the distribution or LD patterns between 1592 breast cancer patients and 1892 controls.

## Material and Methods

### *Genotyping*

We have genotyped 164 breast cancer patients and 86 healthy women with no known history of cancer (two negative mammography screenings). The panel of SNPs genotyped are thoroughly described in (*14*) but in short, SNPs in candidate genes involved in the metabolism of reactive oxygen species and

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

xenobiotics, DNA repair, cell cycle and apoptosis were genotyped using a mini-sequencing (SNP-IT) method multiplexing up to twelve SNPs in one tube. The polymerase chain reaction, clean up with ExoI and SAP and SNP-IT reaction are performed in one tube and the reaction mix hybridized to an array. Each of the twelve SNP-IT primers contain a tag that utilizes sorting of the multiplex reaction on the array. The mini-sequencing reaction is a two colour reaction and signal is detected after laser excitation of the fluorophores on the SNP-stream UHT system.

### *Validation analysis of selected SNPs*

Validation of selected SNPs were done using the Sequenom MassARRY platform and iPLEX genotyping assays (www.sequenom.com/home/) (15).

### *Microarray expression analysis.*

For 50 of the breast cancer patients, expression data were also available. Tumour tissue (20-50mg) was dissected and powdered in liquid nitrogen and total RNA was prepared by standard procedures. Whole genome microarray expression analysis has been performed using cDNA microarrays as described in (16-18).

### *SNP-expression association analysis*

Unselected subset of 3351 mRNA transcripts was obtained by filtering for signal quality (ratio of spot intensity over background exceeding 1.5 in at least 80% of the experiments in each dye channel). The analysis of the SNP-expression associations are published earlier in (*13*). For these patients

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

additional 28 SNPs were available for the haplotype analysis. In short, the correlation between genotypes and expression level of the different mRNA transcripts were assessed using three, different statistical approaches; ANOVA, QMIS and LOOCV. For each SNP locus and each transcript, the one-way ANOVA p-value was computed for the expression vector and grouping of the samples based on SNP locus genotypes (19) assuming the null hypothesis that the expression level distributions are the same, regardless of the genotype class. QMIS (Quantitative Mutual Information Score). For a SNP locus s and an expression vector q of transcript t, let G be a partition of samples induced by the genotype values at locus s. For an expression level threshold p, let Cp be a partition of samples defined by the $q<p$ and $q\geq p$. The mutual information score (MIS) is the difference between the entropy of the partition Cp and the conditional entropy of Cp given G: MIS(Cp, G) = H(Cp) – H(Cp |G), where H is the entropy function. The quantitative mutual information score is defined to be the maximum possible MIS, i.e., QMIS(C,G) = maxmin(q) ≤p≤max(q)MIS(Cp,,G). An exact p-value for the mutual information score can be computed exactly by an efficient exhaustive approach (20). In this case, the null hypothesis is that genotype values have the same distribution, regardless of expression levels. For QMIS, 769 SNP-transcript association pairs with p-values ≤ 1.0E -04 were observed, representing an FDR of 0.2. LOOCV (Leave Out Cross Validation) for a given SNP in the data set, its genotypes were utilized to group samples. For each grouping, leave-one-out-cross-validation analysis was performed, trying to

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

predict from the expression data which genotype group each sample belongs to (similar to the methods described in (21)).

### *Gene Ontology analysis (GO)*

The group of transcripts associated to the same SNP or group of SNPs was analysed with regards to enrichment of GO terms based on GO terms downloaded from Source (http://source.stanford.edu/cgi-bin/source/sourceSearch), for this analysis the p-value cut-off for the SNP-expression association was set at 0.05 and 0.01. The significant overrepresentation for a GO term was calculated taking into account the total number of; 1) genes on the expression array, 2) genes associated with the GO term, 3) genes associated to the SNP and 4) the number of genes associated with the SNP or group of SNPs, that belong to the GO term. The z score was calculated according (14) by subtracting the expected number of genes in a GO term from the observed and diving this by the standard deviation of the observed.

$$z = \frac{(observed - \exp ected)}{std(observed)}$$

### *Calculation of LD and Spearman's correlation coefficient*

SNPs that had discovery rate lower than 75% were excluded. Initially, the panels of SNPs were screened for clusters containing a minimum of 3 SNPs with no more than 100 kb between neighboring SNPs. For the genes represented in

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

these clusters – all genotyped SNPs were included. LD estimations were done in two steps. First, we estimated the haplotypes for cases and controls separately from our population genotype data using the recombination model implemented in the program PHASE (Stephens, M. et al.) with 5 different seeds and 100. The significance of the difference in haplotype distribution between cases and controls was calculated in Phase. The second step was the evaluation of the LD for all included genes. For this purpose, we calculated pairwise D' for cases and controls separately for all possible SNP combinations within a gene and under consideration of the uncertainty in phase estimation (*11*). PHASE also provides the recombination rate as a measure of dependency between the SNPs for all adjacent SNPs within a gene. To evaluate the difference for each gene between the LD-patterns of cases and controls, we calculated Spearman's correlation coefficient $\rho$ as done in (*9*). The correlation is given as a value between -1 ≤ 0 ≤-1, where 0 indicates no correlation, whereas -1 and 1 indicates high negative and positive correlation respectively. We calculated this nonparametric correlation coefficient 1) using all markers for D' and 2) using only adjacent markers for $\rho$.

### *Analysing the relationship between haplotypes and expression levels of transcripts associated to multiple SNPs within a gene*

The non-parametric Mann Whitney or Kruskal Wallis test was used to analyse the possible connection between the haplotypes estimated for a gene and the expression levels of transcripts associated to all or a subset of the SNPs within the given haplotypes. The analysis were performed using SPSS v15.0, the

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

p-values are exact (50 iterations), two-tailed and not corrected for multiple testing.

***Estimating population subdivision – calculating the fixation index***

Population subdivision was estimated using the Arlequin Software to calculate the Fixation index (Fst). This index measures the population differentiation between two groups and its values range from 0 to 1 (with 0 meaning that the populations are completely similar with regard to allele frequencies and 1 being that the populations are completely differentiated (22).

## Results and discussion

The overall study design is given in **Supplementary Figure 1**. A total of 687 SNPs in 203 genes selected from pathways related to the ROS metabolism and signaling were genotyped in 169 breast cancer patients and 86 controls (*14*). Haplotypes were inferred and of the 687 SNPs, a subset of 457 SNPs were available at HapMap with associated frequency information. The full list of SNPs used in the analysis can be found in **Supplementary Table 1** together with information on gene affiliation, chromosomal position, allelic variation and strand genotyped.

### *Impact of multiple SNPs (biclusters) on the expression profile;*

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For 50 of the patients genotyped here expression data were available and we have previously reported the association of 538 SNPs to the intratumoral mRNA expression in these patients (*13*). Many of the studied genes, e.g. *ABCC1, ALOX12, DPYD, GSTM3, NOX3, IL10* and *IL8* were shown to harbor multiple SNPs significantly associated to the level of transcripts *in cis* and *trans* (for full list see **Supplementary Table 2**). We have formally assessed the degree of LD between the multiple SNPs regulating the same group of transcripts and observe that many of these are in strong linkage disequilibrium such as in the genes of *DPYD*, *TXNIP*, *GSTA4*, *PPP1R9A*, *NFKBIA*, *IGF1R*, *ABCC1* and as shown for *XDH* and *IL1R1* on chromosome 2 **Figure 1** (figures for all other chromosomes are given in **Supplementary Figure 2a-u**). Further analyzing the characteristics of these subsets of coexpressed transcripts by gene ontology analysis (p-value cut-off for the SNP-transcript association: 0.01), we find for SNPs in more than 25 genes a significant overrepresentation of GO terms in the list of regulated transcripts at p-value< 0.001 (**Table 1**). Compelling examples are: 1) 18 SNPs in DPYD (involved in pyrimidine base degradation) which together with a SNP in *GSTM4* all are associated to the expression of a group of 10 transcripts among which there is an overrepresentation of the GO term *regulation of cell growth* and 2) 6 SNPs in *GSTA4* associated to a group of 20 transcripts with an overrepresentation of the GO term *transcriptional activator activity.*

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In addition, we also found transcripts such as *ANKS1, CREG, NFKB1, TYMS* and *USP1* that were associated each to multiple SNPs (**Supplementary Table 3**).

### *Analysis of the haplotype distribution and chromosome wise LD pattern in the case vs. the control population*

Haplotypes were estimated for all genes harboring more than 3 SNPs with a maximum distance between neighboring SNPs of 100 kb (n=83). Haplotypes were inferred for the case and control groups separately and the significance of the difference in their distribution was evaluated. A significant difference (p<0.05) in haplotype distribution between cases and controls was observed for 35 genes such as *ABCC1*, *AKT2*, *NFKB1*, *ALOX15B*, *GSR* and *PIK3CA* (**Table 2**).

The pairwise LD was estimated for: 1) all markers and 2) only between neighboring markers by the standard measurements D' and $r^2$ under consideration of the uncertainty in phase estimation as described in (*11*). In addition for the neighboring markers, $\rho$ (estimating the population recombination rate across multiple populations) was calculated as described by Evans and Cardon (*9*). Looking at neighboring markers there is a negative correlation ($\rho < -0.700$) between the LD patterns in cases and controls in 8 genes such as *CDKN1A*, *EPHX1* and *XRCC1* **(Table 3, panel A)**. When including all possible pairwise comparisons for the D' measure, the Spearman's correlation analysis revealed a negative correlation for *PQLC2*, *SOD2* and *PIK3CA* **(Table 3, panel B)**. Comparing the pairwise correlation analysis between cancers and controls

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with the haplotype distribution analysis we see that for the genes where we find a significant different haplotype distribution between cases and controls the correlation is either very low or positive. These results indicate that the difference between cases and controls may be identified by studying together the degree of correlation of LD patterns and the haplotype frequency distribution.

Additionally, we investigated neighboring clusters of genes for differences in LD structure and found a negative correlation between the LD values for cases and controls in neighboring regions for gene-pairs such as *IL1A*+*IL1B*, *RAF1*+*XPC* and *NFKBIA*+*FOS* (**Supplementary Table 4**). These results suggest that the impact of a SNP on susceptibility may be fortified by its organization into haplotype structure including more than one gene, which together may confer higher risk.

### *Impact of the identified putative susceptibility haplotypes on the expression profile;*

The haplotypes that were found significantly differently distributed between cases and controls in the genes *ABCC1, BCL2*, *IGF1R*, *LIG4*, *PPP1R9A* and *TXNIP*. were then tested for association to intratumoral expression. The increased complexity with increasing number of estimated haplotypes made it difficult to detect any significant trends for *ABCC1*, *BCL2*, *IGF1R* and partly *PPP1R9A* but for both *LIG4* and *TXNIP* a significant association between the expression level of several transcripts and the estimated haplotypes was identified. For TXNIP, the second most frequent haplotype (AAAGGAG, **Table 1**) was found associated

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to the expression level of *MADH4*, *NFE2L1* and *TRAP240* (exact p-value <0.001 and 0.001 respectively, **Figure 2 a** and **b**). For *LIG4*, three transcript probes linked to the overrepresented GO term “ubiquitin cycle” were available representing the expression levels of *FBXO11*, *TSG101* and *CDC34*. Combinations of the second most frequent haplotype (CACCT, **Table 1**) show a significantly different expression level for *FBXO11* (exact p-value 0.009, **Figure 2c**).

### *Frequency distribution of the htSNPs derived from the putative susceptibility haplotypes associated to expression in cases and controls.*

A total of 42 htSNPs in 9 genes (*ABCC1, IL1R1*, *PPP3CA*, *NFKB1*, *BCL2*, *IGF1R*, *LIG4*, *PPP1R9A* and *TXNIP*) with both significant difference in haplotype distribution between cases and controls and an association between multiple SNPs in the gene an intratumoral expression, either *in cis or trans,* were selected for case control analysis. All in all we genotyped 3484 samples divided in 1592 samples from BC patients/survivors and 1892 controls. 16 of the 42 investigated SNPs were found associated or borderline associated with case-control status (**Table 4**). Three SNPs, rs 215094 in *ABCC1*,(p<2.25E-04) rs878335 in *IGF1R* (p< 5.58E-09) and rs1805388 in *Lig4 (p*< 7.73E-6) were significant after BonFerroni correction with the SNP in *IGF1R* reaching genome wide significance level.

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

### *Controls vs hapmap Caucasians*

Population subdivision between our sample material and the HapMap samples was estimated by the Fixation index ($F_{st}$) which measures the population differentiation between two or more (22). The $F_{st}$ was calculated separately for the nine genes with ≤7 loci available for analysis (*BCL2*, *IGF1R*, *IL10*, *NFKB1*, *NOX3*, *TANK*, *TGFBR2*, *TXNIP* and *XRCC4*, **Table 1**) and then averaged over all genes. The average $F_{st}$ was 0.0065, indicating a negligible difference between the two populations.

## Conclusion

Several studies have looked into the relationship between single SNPs and risk of sporadic breast cancer both at the single SNP level and the GWAS level. The success of the former in identifying low penetrance alleles have been limited while the latter has identified regions of 10q26 (*FGFR2)*, 16q12.1 (*TNRC9*), 5q11.2 (*MAP3KI*), 8q24, 11p15.5, 5q12 and recently 1p11.2, 14q24.1 (RAD51L1), 3p24 and 17q23.2 to be linked to risk of sporadic breast cancer (3,23-27). In this study we have chosen to look at the association between haplotypes and LD patterns in more than 80 genes distributed across all chromosomes and how they differ between cases and controls and identify differences in both, interestingly not at the same time, in important cancer related genes such as *NFKB1*, *PIK3CA* and *CDKN1A*. We also link the results of our haplotype analysis to our previously published results revealing an association

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

between the germline variation and the expression level in the tumor itself (*13*). Our SNPs are not representative for the whole genome – they are selected from a candidate gene approach but they anyway make grounds for comparing haplotype patterns between cases and controls and to estimate to what extent these results can be extrapolated to other populations through the genetic similarity with data extracted for the Caucasian samples included in the HapMap project. If we manage to find SNPs in the classical and novel regulatory areas of the genes that correlate to the expression of genes in breast cancer, we will be able to predict the risk of developing certain molecular portraits of breast cancer before the cancer has at all occurred.

**Acknowledgement**

This project was supported by The Norwegian Cancer Association (project number D-03067 and D-99061) and The National Programme for Research in Functional Genomics in Norway (FUGE, project numbers 151924/150 and 15204/150), funded by The Research Council of Norway. Genotyping was partly performed at the Uppsala SNP technology platform, which is funded by the Swedish Wallenberg Consortium North (WCN).

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# References

References

(1) Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A *et al.* A Catalog of Published Genome-Wide Association Studies. 7-4-2012.

Ref Type: Internet Communication

(2) Haiman CA, Chen GK, Vachon CM *et al.* A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet* 2011; 43(12):1210-4.

(3) Thomas G, Jacobs KB, Kraft P *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* 2009.

(4) Kim HC, Lee JY, Sung H *et al.* A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. *Breast Cancer Res* 2012; 14(2):R56.

(5) Fletcher O, Johnson N, Orr N *et al.* Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J Natl Cancer Inst* 2011; 103(5):425-35.

(6) Lewontin RC. On measures of gametic disequilibrium. *Genetics* 1988; 120(3):849-52.

(7) Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; 69(1):1-14.

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(8) Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; 18(1):19-24.

(9) Evans DM, Cardon LR. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet* 2005; 76(4):681-7.

(10) Marchini J, Cutler D, Patterson N *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 2006; 78(3):437-50.

(11) Kulle B, Frigessi A, Edvardsen H, Kristensen V, Wojnowski L. Accounting for haplotype phase uncertainty in linkage disequilibrium estimation. *Genet Epidemiol* 2008; 32(2):168-78.

(12) Nordgard SH, Johansen FE, Alnaes GI *et al.* Genome-wide analysis identifies 16q deletion associated with survival, molecular subtypes, mRNA expression, and germline haplotypes in breast cancer patients. *Genes Chromosomes Cancer* 2008; 47(8):680-96.

(13) Kristensen VN, Edvardsen H, Tsalenko A *et al.* Genetic variation in putative regulatory loci controlling gene expression in breast cancer. *Proc Natl Acad Sci U S A* 2006; 103(20):7735-40.

(14) Edvardsen H, Irene Grenaker AG, Tsalenko A *et al.* Experimental validation of data mined single nucleotide polymorphisms from several databases and consecutive dbSNP builds. *Pharmacogenet Genomics* 2006; 16(3):207-17.

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(15) Craig DW, Millis MP, DiStefano JK. Genome-wide SNP genotyping study using pooled DNA to identify candidate markers mediating susceptibility to end-stage renal disease attributed to Type 1 diabetes. *Diabet Med* 2009; 26(11):1090-8.

(16) Perou CM, Sorlie T, Eisen MB *et al.* Molecular portraits of human breast tumours. *Nature* 2000; 406(6797):747-52.

(17) Sorlie T, Perou CM, Tibshirani R *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001; 98(19):10869-74.

(18) Sorlie T, Tibshirani R, Parker J *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003; 100(14):8418-23.

(19) Rice JA. Mathematical Statistics and Data Analysis. 1995.

(20) Tsalenko A, Ben-Dor A, Cox N, Yakhini Z. Methods for analysis and visualization of SNP genotype data for complex diseases. *Pac Symp Biocomput* 2003;548-61.

(21) Hedenfalk I, Ringner M, Ben-Dor A *et al.* Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc Natl Acad Sci U S A* 2003; 100(5):2532-7.

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(22) Schneider S, Roessli D, Excoffier L. Arlequin: A software for population genetics data analysis. [2.00]. 2002. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Ref Type: Computer Program

(23) Easton DF, Eeles RA. Genome-wide association studies in cancer. *Hum Mol Genet* 2008; 17(R2):R109-R115.

(24) Hunter DJ, Kraft P, Jacobs KB *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 2007; 39(7):870-4.

(25) Stacey SN, Manolescu A, Sulem P *et al.* Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2008; 40(6):703-6.

(26) Ahmed S, Thomas G, Ghoussaini M *et al.* Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* 2009.

(27) Vega A, Salas A, Milne RL *et al.* Evaluating new candidate SNPs as low penetrance risk factors in sporadic breast cancer: a two-stage Spanish case-control study. *Gynecol Oncol* 2009; 112(1):210-4.

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Table legends

**Table 1.** GO analysis of the set of transcripts associated to groups of SNPs in single genes reveals an overrepresentation of GO terms among these transcripts (p-value<0.001, Supplementary Table 5: p-value<0.05). Genes with a significant difference in haplotype distribution between cases and controls are given in bold (Table 3).

**Table 2**. Genes with significantly different haplotype distribution between cases and controls. *P value of 0.01 indicates 0.01 or less.*

**Table 3.** Spearmann's correlation between LD of cases and controls for neighbouring SNPs (panel A) and all SNPs (panel B) within a gene based on r and D' values respectively. Listed here are only genes with an absolute correlation between 0.7 and 1

**Supplementary Table 1.** SNPs included in analysis with information on gene affiliation, chromosomal position, allelic variants and strand genotyped.

**Supplementary Table 2.** Multiple SNPs located in the same gene were found associated to the expression level of a number of transcripts by both ANOVA and QMIS analysis in [1]. Listed here are the gene info, rs-numbers, probe id of associated transcripts, most significant p-value from association analysis as well as whether the association is *in cis* or *in trans.*

**Supplementary Table 3** Transcripts associated to genetic variation of multiple SNPs located within the same gene by both ANOVA and QMIS analysis in [1]. Listed here are gene info for identified transcripts,rs-numbers and gene info of associated SNPs, most significant p-value from association analysis as well as whether the association is *in cis* or *in trans.*

**Supplementary Table 2.** Spearmann's correlation based on D' values between LD of cases and controls calculated in the intergenic areas. Listed here are only intergenic regions with an absolute correlation between 0.4 and 1

**Supplementary Table 4.** List of transcripts regulated by several SNPs

**Supplementary Table 5.** GO analysis of the set of transcripts associated to groups of SNPs in single genes reveals an overrepresentation of GO terms among these transcripts (p-value < 0.05). "Top" indicates the number of the regulated transcripts associated with the given GO term.

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Figure legends

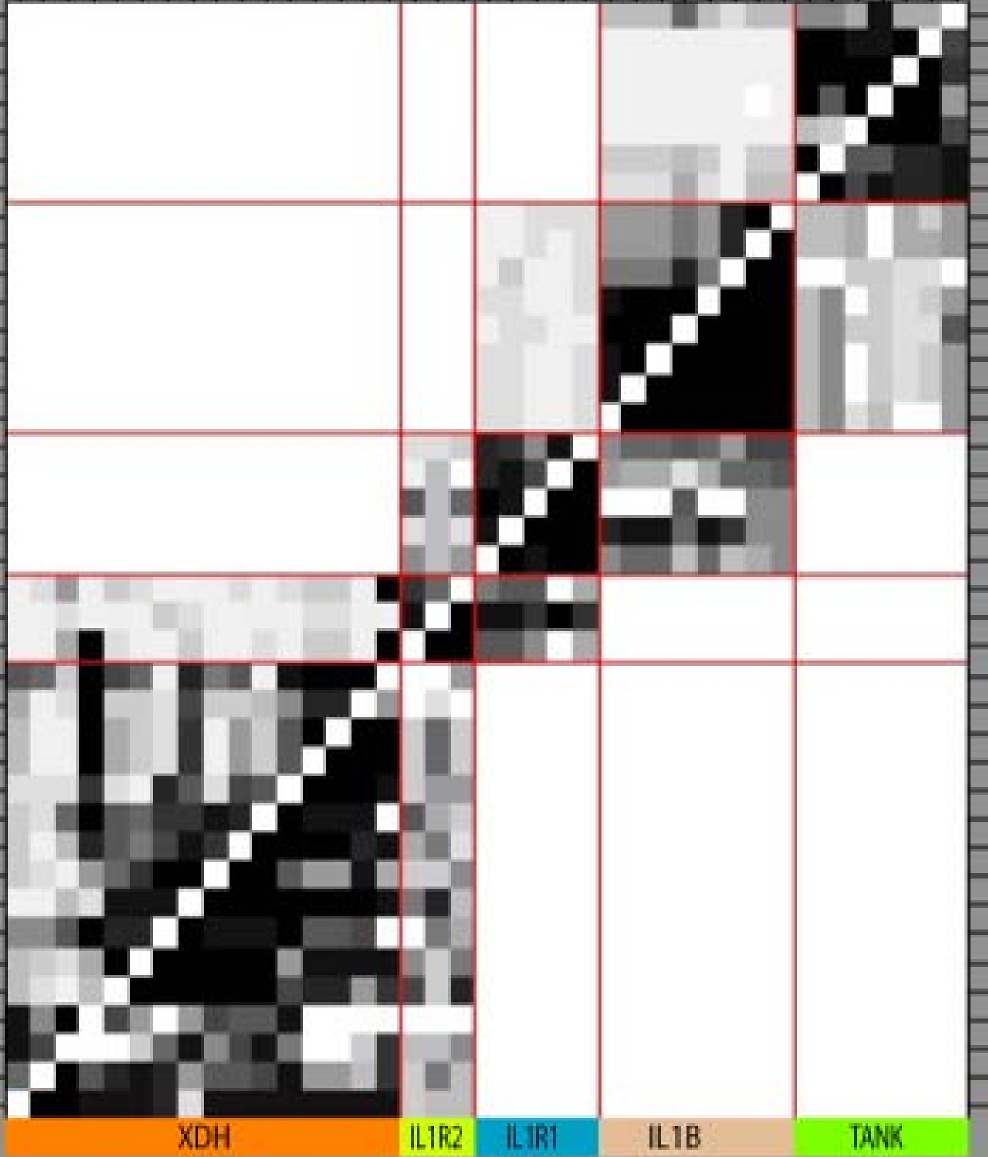
Figure 1 LD pattern in chromosome 2 for the cases together with information on overrepresented GO terms among associated transcripts. X-axis indicates the significance level of the LD while the |D'| values are plotted on the Y-axis, values along the diagonal are intragenic, adjacent panels give information on intergenic regions.

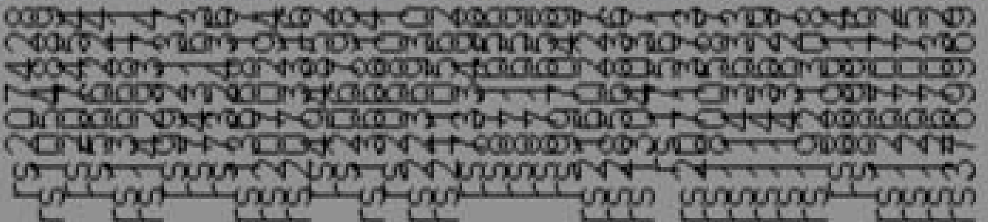
Figure 2 Boxplots showing the spread in the expression levels of the transcripts probes for *MADH4* (A) and *NFE2L1* (B) for the different haplotype combinations of *TXNIP* as well as the spread in the expression levels of the transcripts *FBXO11* for the different haplotype combinations of LIG4

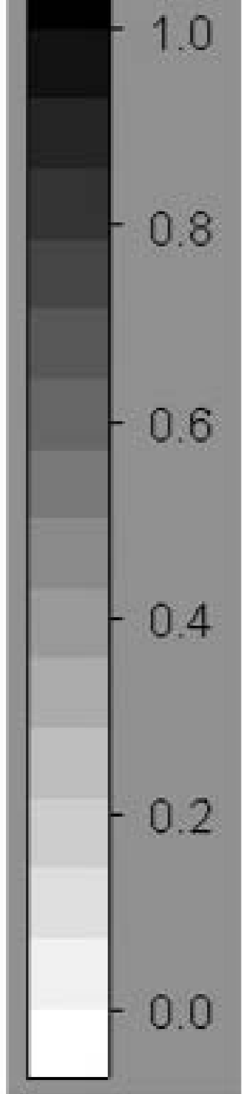
Supplementary Figure 1 Flow chart of the sample material and analysis

Supplementary Figure 2a-u.Chromosome wise LD for the cases together with information on overrepresented GO terms among associated transcripts. X-axis indicates the significance level of the LD while the |D'| values are plotted on the Y-axis, values along the diagonal are intragenic, adjacent panels give information on intergenic regions.

Chromosome 2




| GO term | Associated transcripts |
| --- | --- |
| blood coagulation | TFPI, WAS, THBD |
| transmembrane receptor activity | IL1R1, FCER1G, [illegible] |

| GO term | Associated transcripts |
| --- | --- |
| rho small monomeric gtpase activity | ARHG, ARHE, CDC42 |

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2 Boxplots showing the spread in the expression levels of the transcripts probes associated to haplotypes of 1) *TXNIP*: *MADH4* (A) and *NFE2L1* (B) and 2) *LIG4*: *FBXO11* (C). The haplotypes presented in the figure is as follows (1=CAAGGAG, 3=CAAACTG, 4=CGGGGAG and 5=AAAGGAG) for *TXNIP* and (1=TACCT, 2=TATCT, 3=TATTT and 4= CACCT) for *LIG4*, (for full list of the haplotypes with a frequency of more than 1% in the studied sample set and identified frequency in the controls and cases separately see Supplementary Table 2).

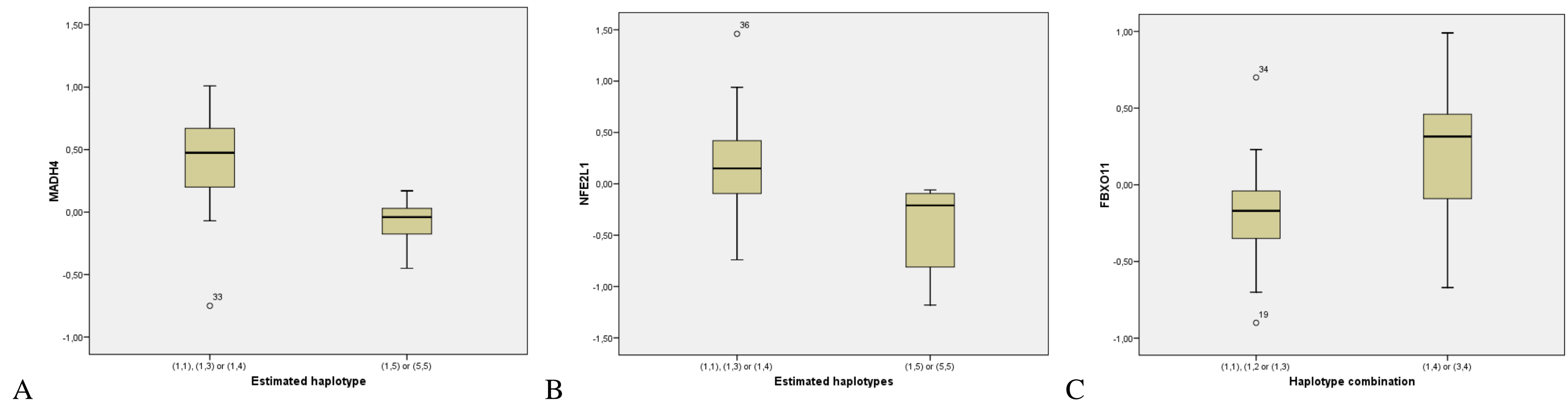


**Table 1. GO analysis of the set of transcripts associated to groups of SNPs in single genes reveals an overrepresentation of GO terms among these transcripts ( p-value<0.001). Genes with a significant difference in haplotype distribution between cases and controls are given in bold.**

| Gene | # of associated SNPs | SNP(s) | # of transcripts regulated | GO term overrepresentated in group of transcripts | z-score | p-value | Top | Top transcript members of Go term |
|---|---|---|---|---|---|---|---|---|
| **ABCC1** | **5** | **212083, 212088, 215067, 215094, 2062541** | **51** | **intracellular signaling cascade** | **4.017746** | **2.93787E-05** | **6** | **SHC1, RAB2L, SYK, PARG1, HSPC163, AKAP13** |
| **BCL2** | **8** | **1381548, 2551402, 899966, 1481031, 720321, 1016860, 1982673, 2062011** | **16** | **heparin binding** | **3.4840574** | **0.000246937** | **3** | **AAMP, SERPINC1, SERPINE2** |
| DPYD | 19 | 1889229, 2151563, 2065943, 1023245, 2786507, 1337521, 1337522, 1801265, 290855, 866129, 1413229, 2039448, 828054, 1879371, 1415681, 827500, 1333727, 2811187, (569998, GSTM4) | 7 | ossification | 3.7774165 | 7.92318E-05 | 3 | SPARC, OSTF1, MGP |
| | | | 10 | inner membrane | 4.15958 | 1.59417E-05 | 4 | COX6A1, SURF1, UQCR, COX6C |
| | | | 10 | regulation of cell growth | 4.15958 | 1.59417E-05 | 4 | COVA1, TSG101, IGFBP5, CTGF |
| | | | 5 | mitochondrial electron transport chain | 4.7101364 | 1.23776E-06 | 3 | SURF1, UQCR, CYC1 |
| GSTA4 | 6 | 1032419, 316128, 316130, 316131, 316132, 367836 | 20 | transcriptional activator activity | 4.087544 | 2.17982E-05 | 3 | MYB, TP53BP1, FOXC1 |
| | | | 11 | protein kinase activity | 5.894784 | 1.87586E-09 | 3 | CCL2, CDK4, TRB2 |
| HIF1AN | 1 | 2295779 | 19 | extracellular matrix structural constituent | 14.076864 | <1E-14 | 3 | MFAP2, LUM, COL6A1 |
| **IER3** | **1** | **14350** | **28** | **structural constituent of ribosome** | **5.9378867** | **1.4436E-09** | **3** | **MRPS2, RPL31, RPS6** |
| **IGF1R** | **8** | **907799, 907807, 2137680, 1568502, 2229765, 871335, 1567811, 2715438** | **76** | **endoplasmic reticulum** | **4.275154** | **9.55026E-06** | **6** | **STS, SYNCRIP, ALG5, CYP1B1, VHL, GNAZ** |
| | | | **21** | **microsome** | **4.501634** | **3.37165E-06** | **3** | **STS, CYP1B1, STCH** |
| | | | **38** | **transcription coactivator activity** | **5.535717** | **1.54979E-08** | **5** | **ELF4, RNF4, NCOA2, DP1, TIF1** |
| **IL1R1** | **5** | **871656, 997049, 871658, 2160227, 871659** | **18** | **blood coagulation** | **4.4390535** | **4.51777E-06** | **3** | **TFPI, WAS, THBD** |
| | | | **17** | **transmembrane receptor activity** | **4.601597** | **2.09632E-06** | **3** | **IL1R1, FCER1G, THBD** |
| IL8 | 3 | 4073, 2227547, 2227306 | 19 | extracellular matrix structural constituent | 6.5541277 | 2.79841E-11 | 5 | FBN1, COL5A2, BGN, COL3A1, MFAP2 |
| LIG3 | 4 | 3136027, 2074516, 2074522, 1003918 | 73 | intracellular | 4.359937 | 6.50499E-06 | 3 | BAT4, RFP, ASB1 |
| **LIG4** | **4** | **868284, 1805388, 1805389, 1805386** | **28** | **ubiquitin cycle** | **4.4994664** | **3.40621E-06** | **3** | **FBXO11, TSG101, CDC34** |
| NDUFA8 | 2 | 6822, 1411445 | 19 | extracellular matrix structural constituent | 8.083895 | <1E-14 | 4 | COL4A2, COL4A1, COL6A2, COL6A1 |
| NFAT5 | 2 | 1437134, 920191 | 38 | transcription coactivator activity | 4.962491 | 3.47974E-07 | 3 | TAF7, TIF1, HTATIP2 |
| **NFKB1** | **10** | **230498, 230505, 230525, 230526, 230531, 1609798, 1585214, 1598857, 1020760, 1020759** | **10** | **epidermal differentiation** | **5.8582754** | **2.33849E-09** | **3** | **KRT5, PLOD, FLOT2** |
| | | | **8** | **central nervous system development** | **6.6552978** | **1.41364E-11** | **3** | **DRPLA, RPS6KA3, ADORA2A** |
| NFKBIA | 3 | 696, 2233415, 1022714 | 24 | response to stress | 7.2226477 | 2.54907E-13 | 3 | HIF1A, MAPK8, MKNK2 |
| NQO1 | 3 | 1800566, 1541979, 744972 | 20 | protein modification | 4.7224402 | 1.16516E-06 | 3 | AGPAT1, GPAA1, MMP15 |
| PDGFC | 2 | 1425492, 2113992 | 250 | integral to membrane | 3.3582497 | 0.000392189 | 3 | STX17, FLOT1, SLC39A1 |
| PPP1R15A | 3 | 638050, 557806, 626140 | 38 | transcription coactivator activity | 4.9209385 | 4.30651E-07 | 4 | TFDP1, ELF3, NFATC3, SF1 |
| **PPP1R9A** | **7** | **854549, 854518, 705377, 854537, 854524, 854523, 854539** | **29** | **inflammatory response** | **5.0928392** | **1.7637E-07** | **4** | **TLR5, NFATC3, RAC1, TNFRSF5** |
| **PPP3CA** | **3** | **1021965, 920559, 958379** | **11** | **antigen processing** | **6.4047303** | **7.53178E-11** | **3** | **HLA-DMA, HLA-DQB1, HLA-DPB1** |
| | | | **10** | **antigen presentation** | **6.7584443** | **6.97409E-12** | **3** | **HLA-DMA, HLA-DQB1, HLA-DPB1** |
| | | | **8** | **exogenous antigen** | **7.648024** | **1.02141E-14** | **3** | **HLA-DMA, HLA-DQB1, HLA-DPB1** |
| | | | **8** | **mhc class ii receptor activity** | **7.648024** | **1.02141E-14** | **3** | **HLA-DMA, HLA-DQB1, HLA-DPB1** |
| | | | **8** | **exogenous antigen via mhc class ii** | **7.648024** | **1.02141E-14** | **3** | **HLA-DMA, HLA-DQB1, HLA-DPB1** |
| TGFBR3 | 17 | 284170, 284176, 284190, 284873, 284874, 901917, 1192529, 2253316, 913059, 2038931, 2799547, 1805113, 2279455, 1192524, 2007686, 2634021, 717923 | 10 | core complex | 3.818641 | 6.70944E-05 | 3 | POLR2K, POLR2G, POLR2F |
| | | | 10 | dna-directed rna polymerase ii | 3.818641 | 6.70944E-05 | 3 | POLR2K, POLR2G, POLR2F |
| TNFAIP2 | 4 | 8126, 2234131, 2234143, 710100 | 45 | extracellular space | 5.1493545 | 1.30692E-07 | 5 | HSPG2, YARS, APOD, TNFAIP2, SERPING1 |
| TOP2B | 3 | 1881708, 1881709, 1001647 | 28 | structural constituent of ribosome | 4.6068473 | 2.0441E-06 | 3 | LAMR1, NHP2L1, MRPL15 |
| **TXNIP** | **4** | **4755, 7211, 7212, 9245** | **13** | **transcription cofactor activity** | **6.652163** | **1.44408E-11** | **3** | **MADH4, NFE2L1, TRAP240** |
| UGT2A1 | 3 | 1432314, 1432324, 1432336 | 56 | protein biosynthesis | 6.098483 | 5.35399E-10 | 5 | ETF1, MRPS21, KIAA0256, SCYE1, NACA |
| XDH | 15 | 2073316, 206798, 206801, 1042039, 1366814, 1366817, 494852, 992137, 732436, 1366811, 1429374, 1054889, 743163, 2070294, 207428 | 6 | rho small monomeric gtpase activity | 5.165976 | 1.19594E-07 | 3 | ARHG, ARHE, CDC42 |

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Table 2. Genes with significantly different haplotype distribution between cases and controls. P value of 0.01 indicates 0.01 or less.**

| Gene (s) | Nr. of SNPs | p-value* | Location | Tot. Nr of hap | No. Of hap > 1%" | Hap. freq.>1% " | Haplotype frequency (%) Controls | Cases |
|---|---|---|---|---|---|---|---|---|
| | | | | | | AATA | 0.871953 | 0.004498 |
| PTGS2 | 4 | 0.01 | 1q31.1 | 3 | 2 | AAGA | 0.122233 | 0.995413 |
| | | | | | | CAAGGAG | 0.912659 | 0.865665 |
| | | | | | | AAAGGAG | 0.052922 | 0.078901 |
| TXNIP | 7 | 0.02 | 1q21.1 | 13 | 3 | CAAACTG | 0.004535 | 0.035556 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | ACGCGCCG | 0.233689 | 0.168315 |
| | | | | | | ATCCGCCG | 0.177444 | 0.241212 |
| | | | | | | ATCTAATG | 0.170937 | 0.205880 |
| | | | | | | GCGTGCCG | 0.017846 | 0.003165 |
| | | | | | | ACGCGCCA | 0.029069 | 0.012455 |
| | | | | | | ATGCGCCG | 0.005971 | 0.017908 |
| | | | | | | ATGTAATG | 0.001858 | 0.016724 |
| | | | | | | ATCCGATG | 0.034942 | 0.031330 |
| IL10 | 8 | 0.03 | 1q32.1 | 30 | 9 | ATCCAATG | 0.017556 | 0.012814 |
| | | | | | | CTGCAAC | 0.514038 | 0.494155 |
| | | | | | | CTGTAAC | 0.293175 | 0.251378 |
| | | | | | | CTACGGC | 0.139476 | 0.224557 |
| TANK | 7 | 0.02 | 2q24.2 | 17 | 4 | CTGCGGC | 0.032224 | 0.000295 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | ACCAC | 0.221693 | 0.225895 |
| | | | | | | TTCAA | 0.173862 | 0.146918 |
| | | | | | | ATCAC | 0.078671 | 0.084747 |
| | | | | | | ATGAA | 0.091558 | 0.047522 |
| | | | | | | ATGAC | 0.013087 | 0.026784 |
| | | | | | | ATCAA | 0.016711 | 0.023701 |
| | | | | | | ACCAA | 0.028107 | 0.010536 |
| IL1R1 | 5 | 0.03 | 2q11.2 | 18 | 9 | ACCTC | 0.027812 | 0.006547 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | GAGCC | 0.223666 | 0.201174 |
| | | | | | | CAGCC | 0.174941 | 0.185088 |
| | | | | | | CGGAT | 0.131953 | 0.160631 |
| | | | | | | CAGAC | 0.130723 | 0.130002 |
| | | | | | | GAGAT | 0.052045 | 0.023673 |
| | | | | | | GAGAC | 0.033055 | 0.010527 |
| XPC, MGC3222 | 5 | 0.01 | 3p25.1 | 14 | 8 | GGGAT | 0.009673 | 0.014249 |
| | | | | | | ATC | 0.442989 | 0.503389 |
| | | | | | | CTC | 0.533736 | 0.456775 |
| PIK3CA | 3 | 0.01 | 3q26.32 | 5 | 3 | ATG | 0.010836 | 0.038121 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | GTGAGTGATCAGTG | 0.055400 | 0.081952 |
| | | | | | | GTAGACGGCTGCCC | 0.066236 | 0.060786 |
| | | | | | | GTGAACGGCTAGTG | 0.033521 | 0.060892 |
| | | | | | | GTAGATGGCTAGCC | 0.042002 | 0.050078 |
| | | | | | | GTAGATGGCTAGTG | 0.042402 | 0.049279 |
| | | | | | | GTAGATGGCCGCCC | 0.032147 | 0.034859 |
| | | | | | | TTAGACGGCTGGTG | 0.016937 | 0.033880 |
| | | | | | | GTGAACGGCTGCCC | 0.011456 | 0.035070 |
| | | | | | | GTAGATAATTAGTG | 0.021577 | 0.024749 |
| | | | | | | GTAGATGGCTGGCC | 0.036540 | 0.016798 |
| | | | | | | TTAGATAATCGGCC | 0.018181 | 0.025054 |
| | | | | | | GTGAATAATTAGTG | 0.028556 | 0.016599 |
| | | | | | | GTAGATGGCTGCTG | 0.012651 | 0.023713 |
| | | | | | | GTAGACGGCTGGCC | 0.017967 | 0.020195 |
| | | | | | | GTAGATGGCTGGTG | 0.023750 | 0.016411 |
| | | | | | | TTAGATAATCGCCC | 0.007683 | 0.024318 |
| | | | | | | TTAGACGGCTGCCC | 0.019633 | 0.017260 |
| | | | | | | TTAGATGGCTGCCC | 0.012387 | 0.015041 |
| | | | | | | GTGAACGGCTGGCC | 0.000746 | 0.020181 |
| | | | | | | TTGAACGGCTGCCC | 0.007766 | 0.014837 |
| | | | | | | GTAGATGGCCAGCC | 0.010506 | 0.012902 |
| TGFBR2 | 14 | 0.01 | 3p24.1 | 215 | 24 | GTAGATGGCCAGTG | 0.006027 | 0.013500 |
| | | | | | | AGCTCCTGCT | 0.034303 | 0.235505 |
| | | | | | | GATTTACGGC | 0.138890 | 0.130323 |
| | | | | | | AGCTCCTCCT | 0.249885 | 0.014257 |
| | | | | | | GGTCTACGGC | 0.066730 | 0.098851 |
| | | | | | | AGTTTCCGCC | 0.075465 | 0.064106 |
| | | | | | | GGTTTACGGC | 0.042786 | 0.056117 |
| | | | | | | AGCTCCTGCC | 0.009937 | 0.056837 |
| | | | | | | GGTCTACGGT | 0.021990 | 0.025563 |
| | | | | | | GGTTTCTGCC | 0.013938 | 0.011156 |
| NFKB1 | 10 | 0.01 | 4q24 | 51 | 11 | AGTTTCTGCC | 0.004538 | 0.015111 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | GATGC | 0.144510 | 0.177623 |
| | | | | | | GGTTC | 0.284364 | 0.005556 |
| | | | | | | AGTGC | 0.038653 | 0.074460 |
| | | | | | | GATTC | 0.068521 | 0.002134 |
| | | | | | | GATGA | 0.020773 | 0.017663 |
| | | | | | | AATGC | 0.008002 | 0.019636 |
| PDGFRA | 5 | 0.01 | 4q12 | 19 | 8 | AGTTC | 0.031469 | 0.000032 |
| | | | | | | AGC | 0.677841 | 0.805401 |
| | | | | | | AAC | 0.111075 | 0.068700 |
| | | | | | | AAG | 0.046110 | 0.061877 |
| | | | | | | GGC | 0.052271 | 0.012229 |
| PPP3CA | 3 | 0.01 | 4q24 | 8 | 5 | GAG | 0.091910 | 0.048143 |
| | | | | | | TCCGG | 0.391092 | 0.472078 |
| | | | | | | CTTTG | 0.439001 | 0.407780 |
| | | | | | | TCTGA | 0.073988 | 0.094056 |
| | | | | | | CCCGG | 0.005581 | 0.018047 |
| CCNB1 | 5 | 0.01 | 5q13.2 | 16 | 5 | TTCGG | 0.029482 | 0.000644 |

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| Gene (s) | Nr. of SNPs | p-value* | Location | Tot. Nr of hap | No. Of hap >1% | Hap. freq>1% | Haplotype frequency (%) Controls | Cases |
|---|---|---|---|---|---|---|---|---|
| | | | | | | CGGAACTCACGTG | 0.229599 | 0.353006 |
| | | | | | | CGGAACTAACGTG | 0.121548 | 0.051329 |
| | | | | | | CGGAGTCAGTGTG | 0.058801 | 0.081718 |
| | | | | | | TAAGACTCACACA | 0.026756 | 0.037067 |
| | | | | | | CGGAACTCACACA | 0.022659 | 0.038418 |
| | | | | | | TAAGACTAACGTG | 0.032373 | 0.002874 |
| | | | | | | CGGAGTCAGTACA | 0.011860 | 0.012605 |
| | | | | | | TAAGGTCAGTACA | 0.004883 | 0.014989 |
| | | | | | | TAAGGTCAGTGTG | 0.009525 | 0.012409 |
| XRCC4 | 13 | 0.02 | 5q14.2 | 60 | 11 | CGGAACTAACACA | 0.020216 | 0.006561 |
| | | | | | | CACC | 0.006744 | 0.823071 |
| | | | | | | CATC | 0.834971 | 0.171974 |
| | | | | | | CATT | 0.111069 | 0.003507 |
| IER3, FLOT1 | 4 | 0.01 | 6p21.33 | 6 | 4 | CTTC | 0.037016 | 0.001007 |
| | | | | | | GGGGTCA | [illegible] | [illegible] |
| | | | | | | GGGGTCC | 0.016381 | 0.332584 |
| | | | | | | GGGGCCA | 0.434521 | 0.011417 |
| | | | | | | GGGGCCC | 0.364135 | 0.009913 |
| | | | | | | GGAATCA | 0.037590 | 0.079271 |
| | | | | | | GGAATCC | 0.028165 | 0.069558 |
| | | | | | | GGGGTAC | 0.000116 | 0.021412 |
| NOX3 | 7 | 0.01 | 6q25.3 | 44 | 8 | GGGGCAC | 0.034837 | 0.000207 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | TTGGCT | 0.261208 | 0.306296 |
| | | | | | | TTGACT | 0.186323 | 0.131406 |
| | | | | | | ACAGCT | 0.050812 | 0.071011 |
| | | | | | | TCAACT | 0.075680 | 0.054396 |
| | | | | | | ACAACT | 0.033680 | 0.052567 |
| | | | | | | TCAGCT | 0.037238 | 0.028066 |
| | | | | | | TTGGTC | 0.045283 | 0.014458 |
| PPP1R9A | 6 | 0.04 | 7q21.3 | 25 | 9 | TCAGTC | 0.015332 | 0.018127 |
| | | | | | | CAG | 0.921753 | 0.960371 |
| GSR | 3 | 0.01 | 8p12 | 4 | 2 | CAA | 0.054588 | 0.036553 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | ACGATC | 0.151244 | 0.152154 |
| | | | | | | AGAATC | 0.142394 | 0.140139 |
| | | | | | | GGAGTC | 0.046576 | 0.091591 |
| | | | | | | GGAATC | 0.027317 | 0.057604 |
| | | | | | | ACAATC | 0.029419 | 0.037510 |
| | | | | | | AGGATC | 0.052660 | 0.018406 |
| | | | | | | ACAGTC | 0.015068 | 0.030906 |
| | | | | | | AGAGTG | 0.055505 | 0.000532 |
| PDGFRL | 6 | 0.01 | 8p22 | 31 | 10 | GGAGTG | 0.033848 | 0.000709 |
| | | | | | | ACC | 0.604080 | 0.638464 |
| | | | | | | GTT | 0.354126 | 0.339410 |
| GSTP1 | 3 | 0.04 | 11q13.2 | 7 | 3 | ATT | 0.017792 | 0.021478 |
| | | | | | | TAG | 0.357149 | 0.463277 |
| | | | | | | TGC | 0.406703 | 0.289934 |
| CCND1,FLJ42258 | 3 | 0.05 | 11q13.3 | 6 | 3 | TGG | 0.221147 | 0.237858 |
| | | | | | | CAGGGC | 0.832252 | 0.912600 |
| | | | | | | CAGGGG | 0.079164 | 0.075237 |
| | | | | | | CATGGC | 0.031895 | 0.003168 |
| CDK2,SILV, RAB5B | 6 | 0.01 | 12q13.2 | 12 | 4 | CAGGAC | 0.025254 | 0.002869 |
| | | | | | | TACCT | 0.664213 | 0.679079 |
| | | | | | | CACCT | 0.161037 | 0.131308 |
| | | | | | | TATCT | 0.132246 | 0.132383 |
| LIG4,C13orf6 | 5 | 0.04 | 13q33.3 | 10 | 4 | TATTT | 0.024655 | 0.056769 |
| | | | | | | GGAG | 0.911390 | 0.946262 |
| | | | | | | GGGG | 0.047386 | 0.046985 |
| NOX5 | 4 | 0.02 | 15q23 | 4 | 3 | TGAG | 0.041224 | 0.003765 |
| | | | | | | GCATAGA | [illegible] | [illegible] |
| | | | | | | GCATGGG | 0.089676 | 0.095857 |
| | | | | | | GCATGGA | 0.078899 | 0.094462 |
| | | | | | | GCATATA | 0.071077 | 0.061278 |
| | | | | | | GCACAGA | 0.065305 | 0.048441 |
| | | | | | | CTATGGA | 0.024052 | 0.051052 |
| | | | | | | CTATAGA | 0.025723 | 0.035314 |
| | | | | | | GCATGTA | 0.034747 | 0.029203 |
| | | | | | | CTATGGG | 0.022218 | 0.030210 |
| | | | | | | CTATATA | 0.014430 | 0.029300 |
| | | | | | | CCATGGA | 0.013170 | 0.024623 |
| | | | | | | CTACAGA | 0.019649 | 0.019941 |
| | | | | | | CCATAGA | 0.031387 | 0.012467 |
| | | | | | | GCATAGG | 0.024496 | 0.015017 |
| | | | | | | GCGTGGG | 0.021673 | 0.015105 |
| | | | | | | CTACGTA | 0.004887 | 0.021284 |
| | | | | | | GCACGTA | 0.008105 | 0.017565 |
| | | | | | | CTATGTA | 0.009085 | 0.015981 |
| | | | | | | CCATGGG | 0.016483 | 0.011514 |
| | | | | | | GCGTAGA | 0.011778 | 0.013432 |
| IGF1R | 7 | 0.04 | 15q26.3 | 115 | 22 | CTACGGG | 0.010612 | 0.012677 |
| | | | | | | TACACG | 0.060470 | 0.153020 |
| | | | | | | TATCTA | 0.068313 | 0.086060 |
| | | | | | | TATACG | 0.046879 | 0.078974 |
| | | | | | | TACCCG | 0.104373 | 0.034471 |
| | | | | | | CGCCCG | 0.044661 | 0.025139 |
| | | | | | | TACCTA | 0.049865 | 0.021073 |
| | | | | | | TACATA | 0.013110 | 0.043541 |
| | | | | | | CACCCG | 0.046067 | 0.021520 |
| | | | | | | CACACG | 0.020836 | 0.033933 |
| | | | | | | CGCACG | 0.022173 | 0.031074 |
| ABCC1 | 6 | 0.01 | 16p13.11 | 51 | 12 | TACCTG | 0.021234 | 0.012495 |
| | | | | | | GTCT | 0.077256 | 0.802388 |
| | | | | | | GCCT | 0.796801 | 0.035392 |
| | | | | | | ATCT | 0.010251 | 0.089543 |
| | | | | | | ACCC | 0.044981 | 0.021830 |
| | | | | | | GCCC | 0.013049 | 0.034648 |
| ALOX15B | 4 | 0.01 | 17p13.1 | 10 | 6 | ACCT | 0.056381 | 0.003313 |
| | | | | | | GGGC | 0.978139 | 0.819479 |
| MAPK7,MFAP4 | 4 | 0.01 | 17p11.2 | 5 | 2 | GGGT | 0.021688 | 0.176348 |

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| Gene (s) | Nr. of SNPs | p-value* | Location | Tot. Nr of hap | No. Of hap >1% | Hap. freq>1%" | Haplotype frequency (%) Controls | Cases |
|---|---|---|---|---|---|---|---|---|
| | | | | | | CACAA | 0.197403 | 0.160090 |
| | | | | | | CAAAG | 0.186905 | 0.118722 |
| | | | | | | CCCGA | 0.098155 | 0.118051 |
| | | | | | | CCCAA | 0.070620 | 0.098702 |
| | | | | | | TACAA | 0.097847 | 0.076954 |
| | | | | | | CACGA | 0.087234 | 0.039990 |
| | | | | | | TAAAG | 0.054397 | 0.039115 |
| | | | | | | CACAG | 0.019759 | 0.019017 |
| | | | | | | CCAAA | 0.005873 | 0.022615 |
| PRKCA | 5 | 0.01 | 17q24.2 | 26 | 12 | TCAAG | 0.000995 | 0.018576 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | CCTGT | 0.198348 | 0.172466 |
| | | | | | | CCCAC | 0.136531 | 0.114793 |
| | | | | | | ACTGT | 0.051837 | 0.055525 |
| | | | | | | CTCGT | 0.060769 | 0.049999 |
| | | | | | | CCCGT | 0.030241 | 0.063912 |
| | | | | | | ATTGT | 0.067519 | 0.034288 |
| COX10 | 5 | 0.03 | 17p12 | 24 | 8 | CTCAC | 0.028492 | 0.033244 |
| | | | | | | GTAGCTGG | 0.274158 | 0.267773 |
| | | | | | | GTAGCTGA | 0.174956 | 0.156950 |
| | | | | | | GTAATAGG | 0.068641 | 0.086111 |
| | | | | | | ATAGCTGG | 0.058198 | 0.034983 |
| | | | | | | GTAGCTAG | 0.041844 | 0.037817 |
| | | | | | | GGAGCTGG | 0.028740 | 0.043804 |
| | | | | | | GTAATAGA | 0.033515 | 0.035419 |
| | | | | | | GTAACTGG | 0.012179 | 0.042986 |
| | | | | | | GTAATTGG | 0.020741 | 0.033496 |
| | | | | | | GGAGCTGA | 0.028100 | 0.026647 |
| | | | | | | GTAATAAG | 0.034400 | 0.018203 |
| | | | | | | GGAATAGG | 0.01733 | 0.023541 |
| | | | | | | GTAATTGA | 0.019188 | 0.02208 |
| | | | | | | GTAACTGA | 0.009672 | 0.017863 |
| | | | | | | GTAGCAAG | 0.011460 | 0.016258 |
| | | | | | | ATAGCTGA | 0.025580 | 0.008679 |
| | | | | | | GTAGCTAA | 0.009443 | 0.015063 |
| BCL2 | 8 | 0.04 | 18q21.33 | 89 | 19 | GGAATAGA | 0.011744 | 0.012961 |
| | | | | | | [illegible] | [illegible] | [illegible] |
| | | | | | | TGGA | 0.149831 | 0.248644 |
| | | | | | | CGGA | 0.159086 | 0.177101 |
| | | | | | | TGAG | 0.209520 | 0.123112 |
| | | | | | | TGAA | 0.081176 | 0.156805 |
| | | | | | | TAGA | 0.001043 | 0.055969 |
| AKT2 | 4 | 0.01 | 19q13.2 | 11 | 7 | TAAA | 0.000000 | 0.028195 |
| POLD1 | 5 | 0.04 | 19q13.33 | | | | | |
| | | | | | | TGG | 0.982558 | 0.999882 |
| COX4I2 | 3 | 0.03 | 20q11.21 | 2 | 1 | AGG | 0.017442 | 0.000089 |
| | | | | | | TGC | 0.737205 | 0.825207 |
| | | | | | | TAA | 0.029907 | 0.163548 |
| TXN2 | 3 | 0.01 | 22q12.3 | 5 | 3 | CAA | 0.225907 | 0.010979 |
| | | | | | | GGC | 0.652538 | 0.566495 |
| | | | | | | GGG | 0.318276 | 0.415658 |
| GSTT2 | 3 | 0.04 | 22q11.23 | 6 | 3 | GAC | 0.020529 | 0.007874 |

* calculated based on the output from Phase which only gives two decimals. This means that the minimum p-value will be 0.01 for this calculations

"in cases and controls combined

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Table 3. Spearmann's correlation between LD of cases and controls for neighbouring SNPs (panel A) and all SNPs (panel B) within a gene based on ρ and D' values respectively. Listed here are only genes with an absolute correlation between 0.7 and 1. The tables are sorted by gene name.**

**A. Between all pairwise LD measurements of neighbouring SNPs**

| Gene (s) | Nr. of SNPs | Localisation | ρ |
|---|---|---|---|
| ABCB1 | 7 | 7q21.12 | 0.943 |
| ABCC1 | 6 | 16p13.11 | 0.800 |
| AKR7A2,PQLC2 | 3 | 1p36.13 | 1.000 |
| ALOX15B | 4 | 17p13.1 | 1.000 |
| BCL2 | 8 | 18q21.33 | 0.786 |
| CAT | 4 | 11p13 | 1.000 |
| CCND1,FLJ42258 | 3 | 11q13.3 | 1.000 |
| CDC42BPB | 3 | 14q32.32 | -1.000 |
| CDK2, SILV,RAB5B | 6 | 12q13.2 | 0.700 |
| CDKN1A | 3 | 6p21.31 | -1.000 |
| COX10 | 5 | 17p12 | 1.000 |
| COX4I2 | 3 | 20q11.21 | -1.000 |
| CYP2C8 | 3 | 10q23.33 | 1.000 |
| DPYD | 17 | 1p21.3 | 0.894 |
| EGF | 6 | 4q25 | 0.900 |
| EPHX1 | 6 | 1q42.12 | -0.700 |
| FGF2 | 4 | 4q27 | 1.000 |
| FOS | 3 | 14q24.3 | 1.000 |
| GADD45A | 3 | 1p31.2 | 1.000 |
| GCLC | 9 | 6p12.1 | 0.762 |
| GSR | 3 | 8p12 | 1.000 |
| GSTA4 | 6 | 6p12.1 | 0.900 |
| GSTM3 | 3 | 1p13.3 | 1.000 |
| GSTP1 | 3 | 11q13.2 | 1.000 |
| GSTT2 | 3 | 22q11.23 | 1.000 |
| IGF1 | 5 | 12q23.2 | 0.800 |
| IGF1R | 7 | 15q26.3 | 0.943 |
| IGF2R | 6 | 6q25.3 | 0.900 |
| IL10 | 8 | 1q32.1 | 0.750 |
| IL10RA | 3 | 11q23.3 | 1.000 |
| IL1A | 3 | 2q13 | 1.000 |
| IL1B | 4 | 2q13 | -1.000 |
| IL1R2 | 3 | 2q11.2 | 1.000 |
| KCNMB1 | 3 | 5q35.1 | 1.000 |
| LIG3 | 3 | 17q12 | -1.000 |
| LIG3,RFFL | 3 | 17q12 | -1.000 |
| LIG4,C13orf6 | 5 | 13q33.3 | 0.800 |
| MAPK9 | 3 | 5q35.3 | 1.000 |
| MGMT | 7 | 10q26.3 | 0.886 |
| NDUFA8 | 3 | 9q33.2 | 1.000 |
| NOX3 | 7 | 6q25.3 | 0.829 |
| NQO2 | 3 | 6p25.2 | 1.000 |

bioRxiv preprint doi: https://doi.org/10.1101/248047; this version posted February 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| | | | |
|---|---|---|---|
| PCNA, C20orf30,CDS2 | 3 | 20p12.3 | 1.000 |
| PDGFRB | 4 | 5q32 | 1.000 |
| PIK3CA | 3 | 3q26.32 | 1.000 |
| PLCG2 | 3 | 16q23.2 | -1.000 |
| PPP1R15A, PLEKHA4, TULP2 | 3 | 19q13.33 | -1.000 |
| PPP1R1A, PDE1B | 3 | 12.q13.2 | 1.000 |
| PPP3CA | 3 | 4q24 | 1.000 |
| PRKCA | 5 | 17q24.2 | 1.000 |
| RAF1 | 3 | 3p25.2 | 1.000 |
| SOD1, SFRS15 | 3 | 21q22.11 | 1.000 |
| SOD2 | 3 | 6q25.3 | 1.000 |
| TGFB2 | 4 | 1q41 | 1.000 |
| TNFRSF6 | 3 | 17q25.1 | 1.000 |
| TXN | 3 | 9q31.3 | 1.000 |
| TXN2 | 3 | 22q12.3 | 1.000 |
| TXNRD2 | 4 | 22q11.21 | 1.000 |
| XDH | 16 | 2p23.1 | 0.800 |
| XPC, MGC3222 | 5 | 3p25.1 | 0.800 |
| XRCC1 | 3 | 19q13.31 | -1.000 |
| XRCC4 | 13 | 5q14.2 | 0.797 |

**B. Between all pairwise LD measurements**

| Gene (s) | Nr. of SNPs | | D' |
|---|---|---|---|
| AKR7A2,PQLC2 | 3 | 1p36.13 | -1.000 |
| SOD2 | 3 | 6q25.3 | -1.000 |
| PIK3CA | 3 | 3q26.32 | -0.866 |
| AKT2 | 4 | 19q13.2 | 0.714 |
| IGF1 | 5 | 12q23.2 | 0.758 |
| IGF1R | 7 | 15q26.3 | 0.765 |
| NAT2 | 4 | 8p22 | 0.771 |
| TNFAIP2 | 4 | 14q32.32 | 0.771 |
| PDGFRL | 6 | 8p22 | 0.836 |
| GSTA4 | 6 | 6p12.1 | 0.846 |
| GSR | 3 | 8p12 | 0.866 |
| GSTP1 | 3 | 11q13.2 | 0.866 |
| TXN2 | 3 | 22q12.3 | 0.866 |
| CAT | 4 | 11p13 | 0.868 |
| IL1B | 4 | 2q13 | 0.886 |
| COX10 | 5 | 17p12 | 0.891 |
| ABCC1 | 6 | 16p13.11 | 0.925 |
| PDGFRB | 4 | 5q32 | 0.943 |
| EPHX1 | 6 | 1q42.12 | 0.943 |
| PPP1R3B | 5 | 8p23.1 | 0.988 |
| CDC42BPB | 3 | 14q32.32 | 1.000 |
| TXNRD2 | 4 | 22q11.21 | 1.000 |
| FOS | 3 | 14q24.3 | 1.000 |
| IL1R2 | 3 | 2q11.2 | 1.000 |
| MAPK9 | 3 | 5q35.3 | 1.000 |
| NDUFA8 | 3 | 9q33.2 | 1.000 |
| NQO2 | 3 | 6p25.2 | 1.000 |
| PPP1R1A, PDE1B | 3 | 12q13.2 | 1.000 |
| TNFRSF6 | 3 | 17q25.1 | 1.000 |
| TXN | 3 | 9q31.3 | 1.000 |