WHICH ARE THE VARIANTS BEHIND COLORECTAL CANCER?

Karmele Alapont

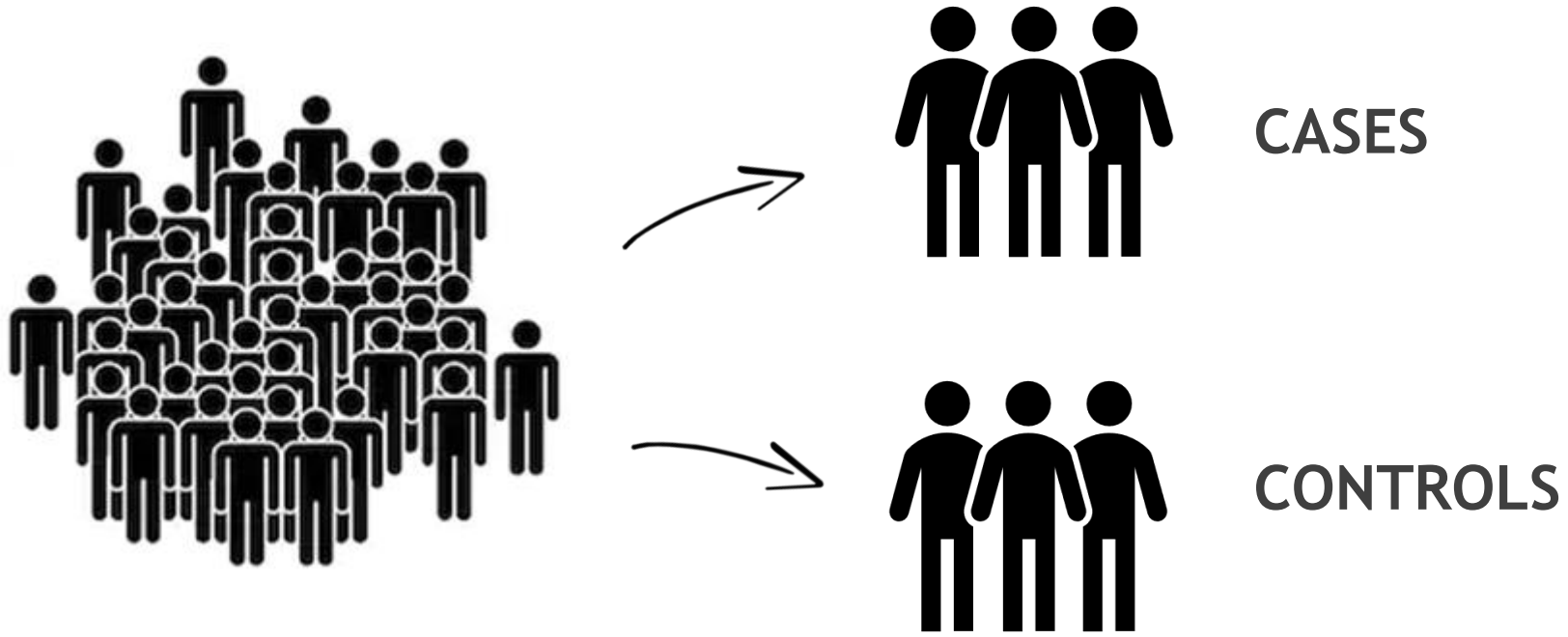
Marina Bataller

Nerea Carrón

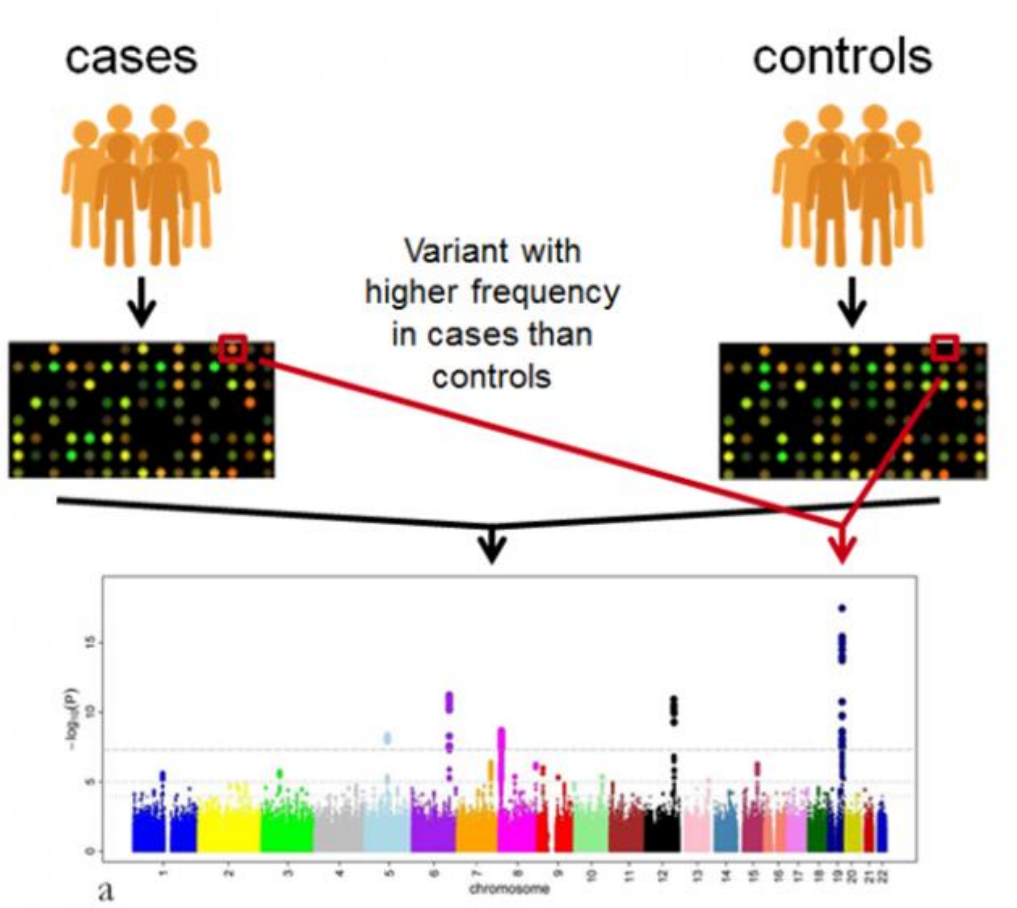
Judit García

INTRODUCTION

GWAS = Genome-Wide Association Studies



INTRODUCTION



EMBL-EBI Train Online, 2020, https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations-2019/what-gwas-catalog/what-are-genome-wide

Requires:

- Statistical tests
- Large number of subjects

In our analysis:

COLORECTAL CANCER

METHODS

PACKAGES AND TOOLS:

- ggplot2
- dplyr
- ggrepel
- devtools
- isglobal-brge/SNPassoc
- BiocManager
- snpStats
- SNPRelate

METHODS

DATA DESCRIPTION:

PLINK colorectal cancer dataset

- Colorectal.bed → Genomic SNP data
- Colorectal.bim → SNP annotation
- Colorectal.fam → Individual's family information

METHODS

DATA DESCRIPTION:

PLINK colorectal cancer dataset

- Colorectal.bed → Genomic SNP data
- Colorectal.bim → SNP annotation
- Colorectal.fam → Individual's family information

Upload to R

Colorectal.plink

- Colorectal.genotype
- Annotation
- Individuals

METHODS

DATA DESCRIPTION:

PLINK colorectal cancer dataset

- Colorectal.bed → Genomic SNP data
- Colorectal.bim → SNP annotation
- Colorectal.fam → Individual's family information

Upload to R

Colorectal.plink

- Colorectal.genotype
- Annotation
- Individuals

Full phenotype information → Additional txt files

Upload to R

- Colorectal.phenotype

METHODS

DATA DESCRIPTION:

```
```{r read-plink-data}
# Read PLINK data of the obestiy dataset
colorectal.plink <-  read.plink(bed = "Project3_cancer/data/colorectal_cancer/colorectal.bed",
                                bim = "Project3_cancer/data/colorectal_cancer/colorectal.bim",
                                fam = "Project3_cancer/data/colorectal_cancer/colorectal.fam")
```
```

METHODS

DATA DESCRIPTION:

```
```{r read-plink-data}
# Read PLINK data of the obestiy dataset
colorectal.plink <-  read.plink(bed = "Project3_cancer/data/colorectal_cancer/colorectal.bed",
                                bim = "Project3_cancer/data/colorectal_cancer/colorectal.bim",
                                fam = "Project3_cancer/data/colorectal_cancer/colorectal.fam")
```
```

```
```{r genotypes}
# Get genotypes information
colorectal.genotype <- colorectal.plink$genotypes
colorectal.genotype
```
```

METHODS

DATA DESCRIPTION:

```
```{r read-plink-data}
# Read PLINK data of the obestiy dataset
colorectal.plink <-  read.plink(bed = "Project3_cancer/data/colorectal_cancer/colorectal.bed",
                                bim = "Project3_cancer/data/colorectal_cancer/colorectal.bim",
                                fam = "Project3_cancer/data/colorectal_cancer/colorectal.fam")
```
```

```
```{r genotypes}
# Get genotypes information
colorectal.genotype <- colorectal.plink$genotypes
colorectal.genotype
```
```

```
```{r individuals}
# Get individuals information
individuals <- colorectal.plink$fam
head(individuals)
```
```

METHODS

DATA DESCRIPTION:

```
```{r read-plink-data}
# Read PLINK data of the obestiy dataset
colorectal.plink <-  read.plink(bed = "Project3_cancer/data/colorectal_cancer/colorectal.bed",
                                bim = "Project3_cancer/data/colorectal_cancer/colorectal.bim",
                                fam = "Project3_cancer/data/colorectal_cancer/colorectal.fam")
```
```

```
```{r genotypes}
# Get genotypes information
colorectal.genotype <- colorectal.plink$genotypes
colorectal.genotype
```
```

```
```{r annotation}
# Get annotation information
annotation <- colorectal.plink$map
head(annotation)
```
```

```
```{r individuals}
# Get individuals information
individuals <- colorectal.plink$fam
head(individuals)
```
```

METHODS

```{r obesity}
colorectal.phenotype <- read.delim("Project3_cancer/data/colorectal_cancer/colorectal.txt")
head(colorectal.phenotype)
```

DATA DESCRIPTION:

```{r read-plink-data}
# Read PLINK data of the obestiy dataset
colorectal.plink <-  read.plink(bed = "Project3_cancer/data/colorectal_cancer/colorectal.bed",
                                bim = "Project3_cancer/data/colorectal_cancer/colorectal.bim",
                                fam = "Project3_cancer/data/colorectal_cancer/colorectal.fam")
```

```{r genotypes}
# Get genotypes information
colorectal.genotype <- colorectal.plink$genotypes
colorectal.genotype
```

```{r annotation}
# Get annotation information
annotation <- colorectal.plink$map
head(annotation)
```

```{r individuals}
# Get individuals information
individuals <- colorectal.plink$fam
head(individuals)
```

METHODS

DATA DESCRIPTION:

```
```{r rename-rownames}
# Rename the rownames with the id
rownames(colorectal.phenotype) <- colorectal.phenotype$id
head(colorectal.phenotype)
```
```

```
```{r check-order}
# We check if the rownames of the two objects are identical
identical(rownames(colorectal.phenotype), rownames(colorectal.genotype))
```
```

```
[1] TRUE
```

METHODS

DATA DESCRIPTION:

```
```{r fix-individuals}

ids <- intersect(rownames(colorectal.phenotype), rownames(colorectal.genotype))
genotype <- colorectal.genotype[ids, ]
phenotype <- colorectal.phenotype[ids, ]
identical(rownames(phenotype), rownames(genotype))
individuals <- individuals[ids, ]
```
```

METHODS

Control individuals → subjects that do not have colorectal cancer

Cascon == 0

```
```{r controls}
# Controls are not subjects with colorectal cancer
controls <- phenotype$cascon == 0 & !is.na(phenotype$cascon)
genotype.controls <- genotype[controls, ]
info.controls <- col.summary(genotype.controls)
nrow(genotype.controls)
```
```

METHODS

QUALITY CONTROL:

Before GWAS analysis to make sure that the data is good enough for the analysis.

Two levels:

1. Quality control of SNPs
2. Quality control of individuals

METHODS

QUALITY CONTROL: 1. QUALITY CONTROL OF SNPs

Measures:

- SNPs with high rate of missing → SNPs with a call rate less than 95% are removed
- Rare SNPs (MAF) → SNPs with less than 5% minor allele frequency are deleted
- SNPs that do not pass the HWE test → controls with a Z-value bigger than 3.3 are removed

METHODS

QUALITY CONTROL: 1. QUALITY CONTROL OF SNPs

```
```{r quality2}
# Filter QC
use <- info.snps$call.rate > 0.95 &
       info.snps$MAF > 0.05 &
       abs(info.controls$z.HWE < 3.3)
mask.snps <- use & !is.na(use)

# We keep those SNPs that pass the QC
genotype.qc.snps <- genotype[, mask.snps]
genotype.qc.snps
annotation <- annotation[mask.snps, ]

# Original SNPs
genotype
# Filtered SNPs
genotype.qc.snps
```
```

METHODS

QUALITY CONTROL: 1. QUALITY CONTROL OF SNPs

Number of deleted individuals:

```
```{r snp-quality-report}
# Number of SNPs removed for a bad call rate
sum(info.snps$call.rate < 0.95, na.rm = TRUE)

# Number of SNPs removed for low MAF
sum(info.snps$MAF < 0.05, na.rm = TRUE)

# Number of SNPs removed that do not pass HWE
sum(abs(info.controls$z.HWE > 3.3), na.rm = TRUE)

# The total number of SNPs removed for any reason
sum(!mask.snps)
```
```

```
[1] 875

[1] 10669

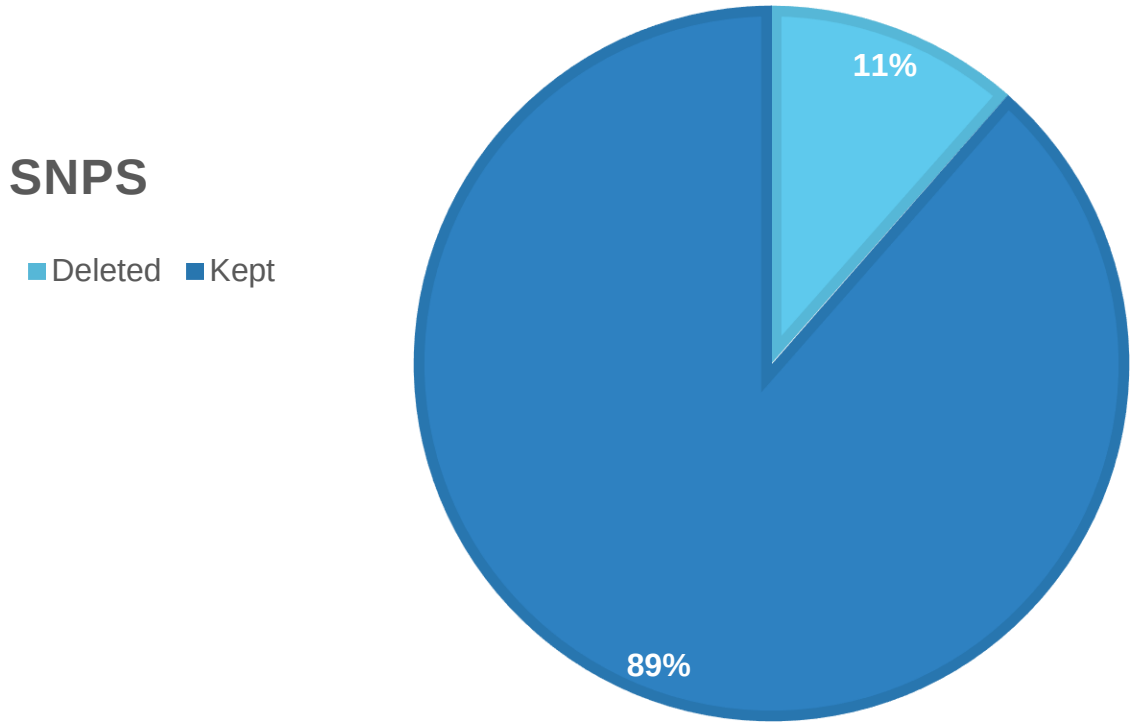
[1] 72

[1] 11479
```

METHODS

QUALITY CONTROL: 1. QUALITY CONTROL OF SNPs

From 100,000 SNPs, we keep 88,521



METHODS

QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS

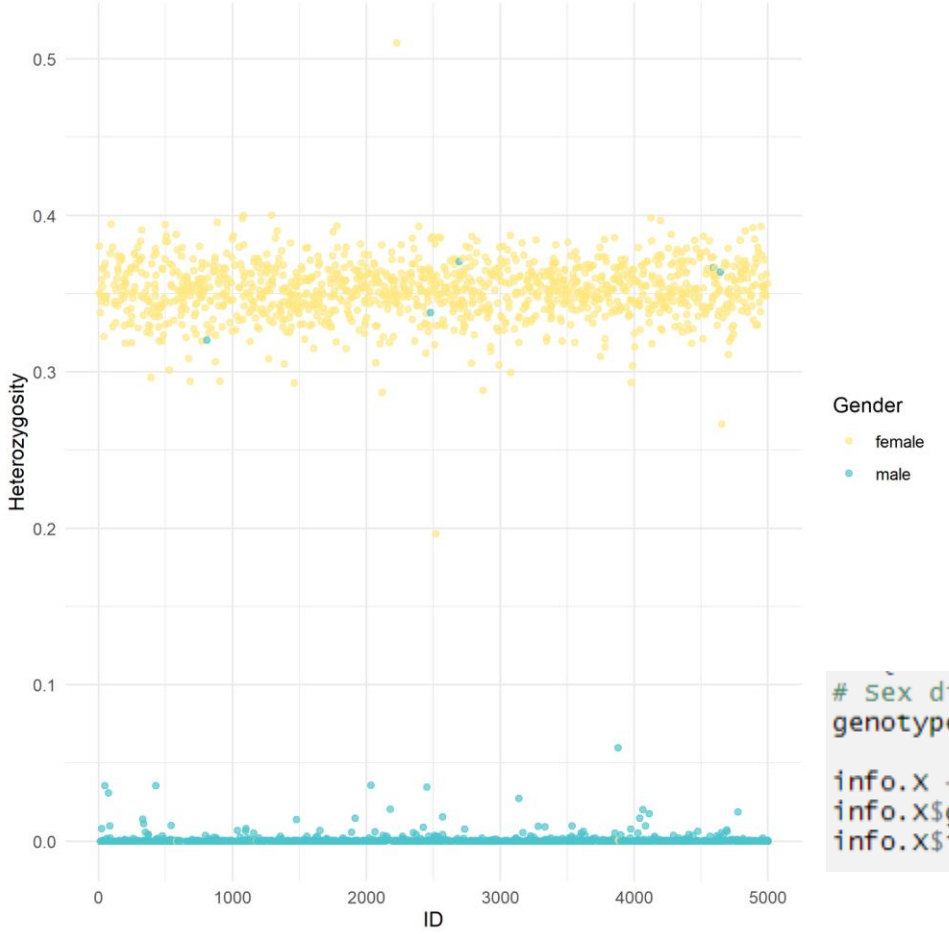
Information at individuals' level

Steps

- Sex discrepancies → Heterozygosity of chromosome X
- Individuals with outlying heterozygosity from the overall genomic heterozygosity
- Delete close familiar relatedness between individuals
- Remove individuals with more than 5% missing genotypes

METHODS

QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS



SEX DISCREPANCIES:

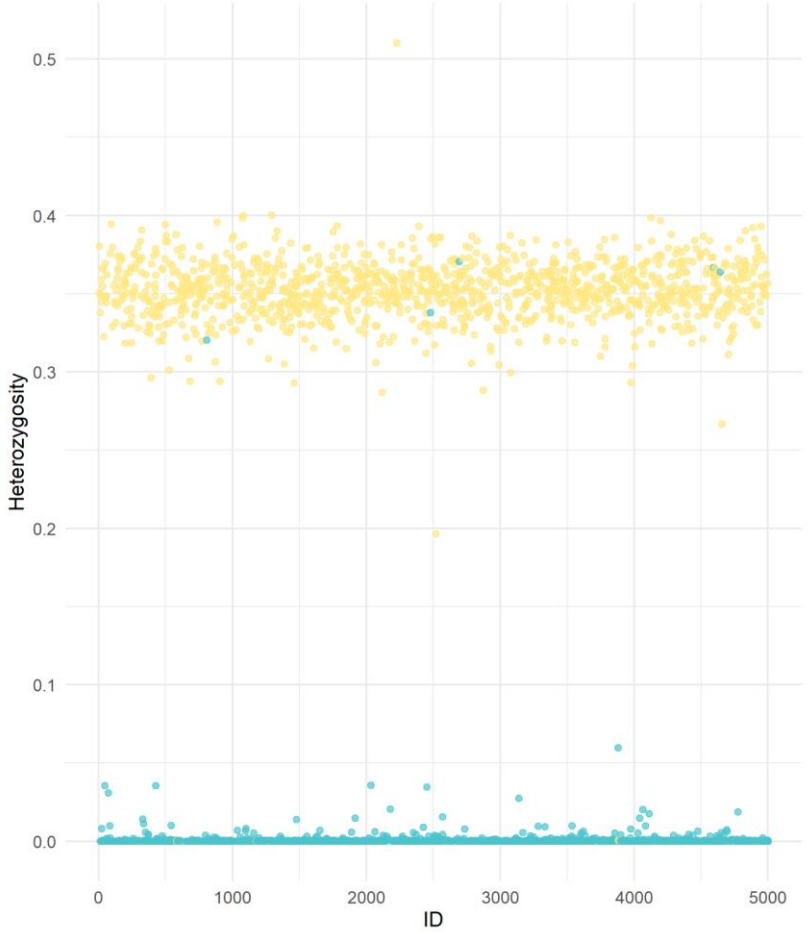
- Males → expected heterozygosity = 0
- Females → expected heterozygosity = 0.30

```
# Sex discrepancies
genotype.X <- genotype.qc.snps[,annotation$chromosome=="23" & !is.na(annotation$chromosome)]

info.X <- row.summary(genotype.X)
info.X$gender <- phenotype$sex
info.X$id <- phenotype$id
```

METHODS

QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS



SEX DISCREPANCIES:

- Males → expected heterozygosity = 0
- Females → expected heterozygosity = 0.30

```
# Plot with ggplot2
ggplot(info.X, aes(y = Heterozygosity, x = id)) +
  geom_point(aes(color=gender), alpha = 0.7) +
  labs(y = "Heterozygosity", x = "ID", color = "Gender") +
  theme_minimal() + scale_color_manual(values = c("#FFE882", "#4DC4CC"))
```

```
# Sex discrepancies
genotype.X <- genotype.qc.snps[,annotation$chromosome=="23" & !is.na(annotation$chromosome)]

info.X <- row.summary(genotype.X)
info.X$gender <- phenotype$sex
info.X$id <- phenotype$id
```

METHODS

QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS

SEX DISCREPANCIES:

- Males → expected heterozygosity = 0
- Females → expected heterozygosity = 0.30

```
```{r sex-discr2}
sex.discrep <- (info.x$gender == "Male" &
                 info.x$Heterozygosity > 0.2) |
                 (info.x$gender=="Female" &
                  info.x$Heterozygosity < 0.2)
```
```

METHODS

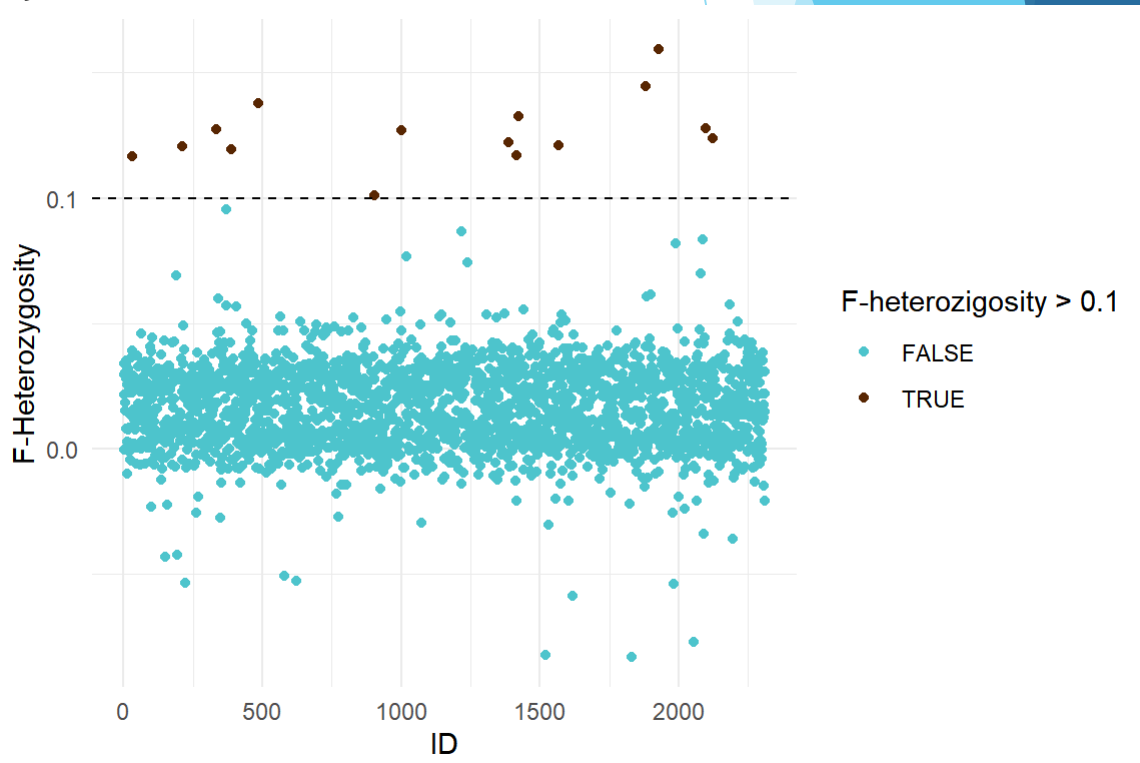
QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS

OUTLYING HETEROZYGOSITY FROM THE OVERALL GENOMIC HETEROZYGOSITY

- Heterozygosity rate lower than 0.32 or F-heterozigosity > 0.1 → outliers that need to be removed

```
MAF <- col.summary(genotype.qc.snps)$MAF
callmatrix <- !is.na(genotype.qc.snps)
hetExp <- callmatrix %*% (2*MAF*(1-MAF))
hetObs <- with(info.indv,
               Heterozygosity*(ncol(genotype.qc.snps))*Call.rate)
info.indv$hetF <- 1 - (hetObs/hetExp)
head(info.indv)
```

```
ggplot(info.indv, aes(x = 1:nrow(info.indv), y = hetF))
  geom_point(aes(color = hetF > 0.1)) +
  geom_hline(yintercept = 0.1, linetype = "dashed") +
  labs(y = "F-Heterozygosity", x = "ID", color = "F-heterozigosity > 0.1") +
  theme_minimal() + scale_color_manual(values = c("#4DC4CC", "#582602"))
```



METHODS

QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS

RELATEDNESS

```
# Open the file
genofile <- snpgdsOpen("colorectalGDS")
# Using a seed allows to reproduce the analysis
set.seed(12345)
snps.qc <- colnames(genotype.qc.snps)
snp.prune <- snpgdsLDpruning(genofile,ld.threshold = 0.2, snp.id=snps.qc)
```

```
snps.ibd <- unlist(snp.prune, use.names=FALSE)
ibd <- snpgdsIBDMoM(genofile, kinship = TRUE,
                    snp.id = snps.ibd,
                    num.thread = 1)
ibd.kin <- snpgdsIBDSelection(ibd)
head(ibd.kin)
```

Individuals with higher kindship than 0.1 are removed

```
ibd.kin.thres <- subset(ibd.kin, kinship > 0.1)
head(ibd.kin.thres)
```

```
##          ID1  ID2        k0        k1   kinship
## 46484   1049  188 0.2933008 0.5060649 0.2268334
## 232848  1202 1330 0.0000000 0.0000000 0.5000000
## 281069  1237  872 0.2903871 0.4285650 0.2476652
## 640474   155 1682 0.2608786 0.4202985 0.2644860
## 806337   170 2015 0.2593817 0.5409794 0.2350643
## 1158509 2055  825 0.0000000 0.0000000 0.5000000
```

We removed them by using their Id

```
ids.rel <- related(ibd.kin.thres)
ids.rel
```

METHODS

QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS

SUMMARY

```
 use <- info.indv$Call.rate > 0.95 &
  abs(info.indv$hetF) < 0.1 &     # or info.inv$Heterozygosity < 0.32
  !sex.discrep &
  !rownames(info.indv)%in%ids.rel
mask.indiv <- use & !is.na(use)
genotype.qc <- genotype.qc.snps[mask.indiv, ]

phenotype.qc <- colorectal.phenotype[mask.indiv, ]
identical(rownames(phenotype.qc), rownames(genotype.qc))

dim(phenotype)
dim(phenotype.qc)
```

METHODS

QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS

Number of deleted individuals:

```
# Number of individuals removed to bad call rate
sum(info.indv$Call.rate < 0.95)
## [1] 32

# Number of individuals removed for heterozygosity problems
sum(abs(info.indv$hetF)>0.1)
## [1] 15

# Number of individuals removed for sex discrepancies
sum(sex.discrep)
## [1] 9

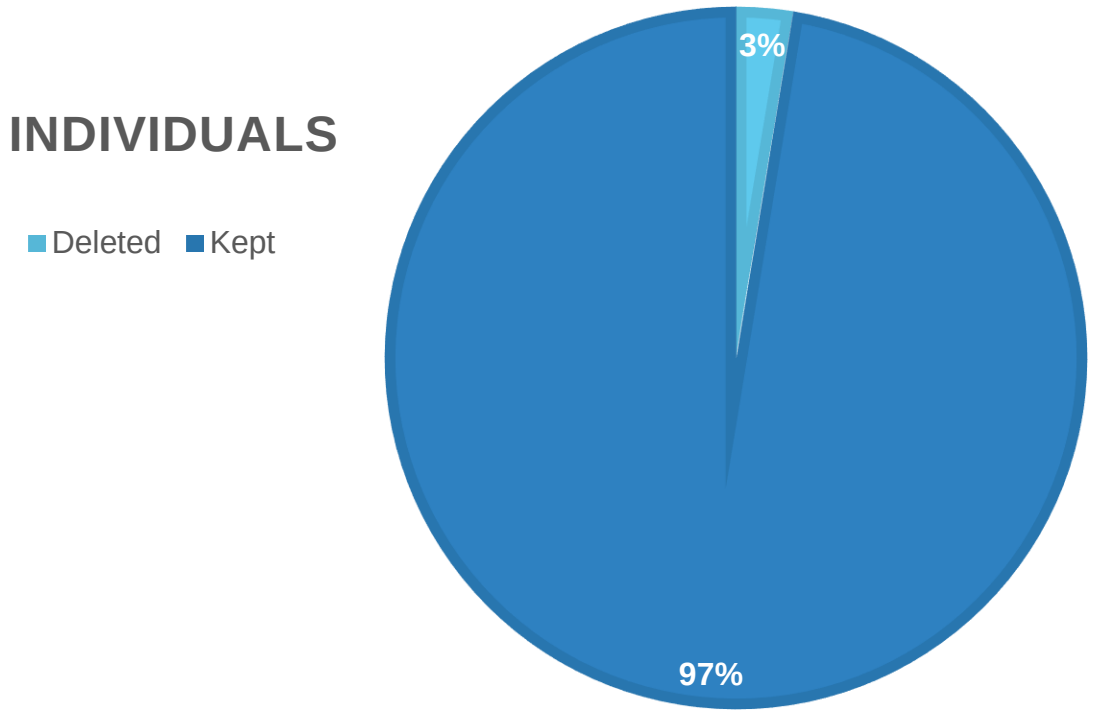
# Number of individuals removed to be related with others
length(ids.rel)
## [1] 15

# The total number of individuals that do not pass QC
sum(!mask.indiv)
## [1] 69
```

METHODS

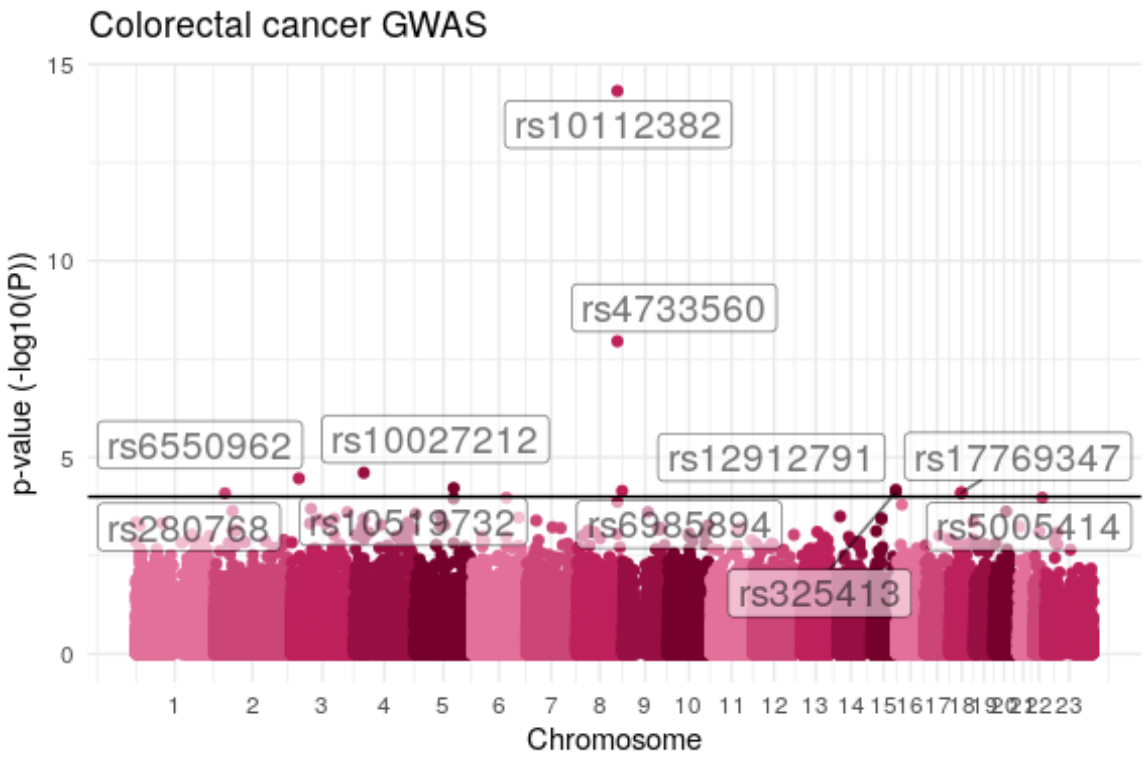
QUALITY CONTROL: 2. QUALITY CONTROL OF INDIVIDUALS

From 2312 individuals, we kept 2243



RESULTS

GWAS ⟶ Manhattan plot



One point for every SNP.

Using the Bonferroni-corrected threshold:

$$\frac{0{,}05}{\text{Number of SNPs}}$$

DISCUSSION

SNPs associated to colorectal cancer

SNP	CHROMOSOME
rs280768	2
rs6550962	3
rs10027212	4
rs10519732	5
rs4733560	8
rs10112382	8
rs6985894	8
rs12912791	15
rs325413	15
rs5005414	18
rs17769347	18

DISCUSSION

SNPs associated to colorectal cancer

SNP	CHROMOSOME
rs280768	2
rs6550962	3
rs10027212	4
rs10519732	5
rs4733560	8
rs10112382	8
rs6985894	8
rs12912791	15
rs325413	15
rs5005414	18
rs17769347	18

DISCUSSION

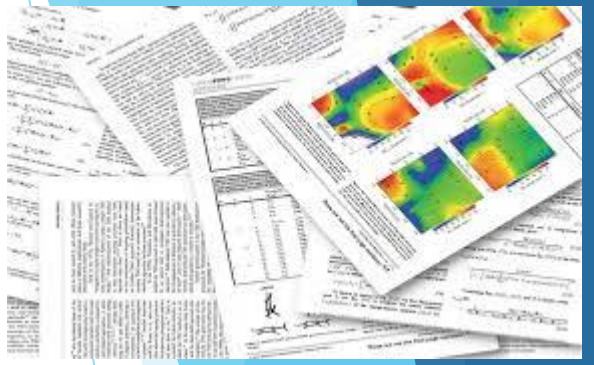
SNPs associated to colorectal cancer

SNP	CHROMOSOME
rs280768	2
rs6550962	3
rs10027212	4
rs10519732	5
rs4733560	8
rs10112382	8
rs6985894	8
rs12912791	15
rs325413	15
rs5005414	18
rs17769347	18

DISCUSSION

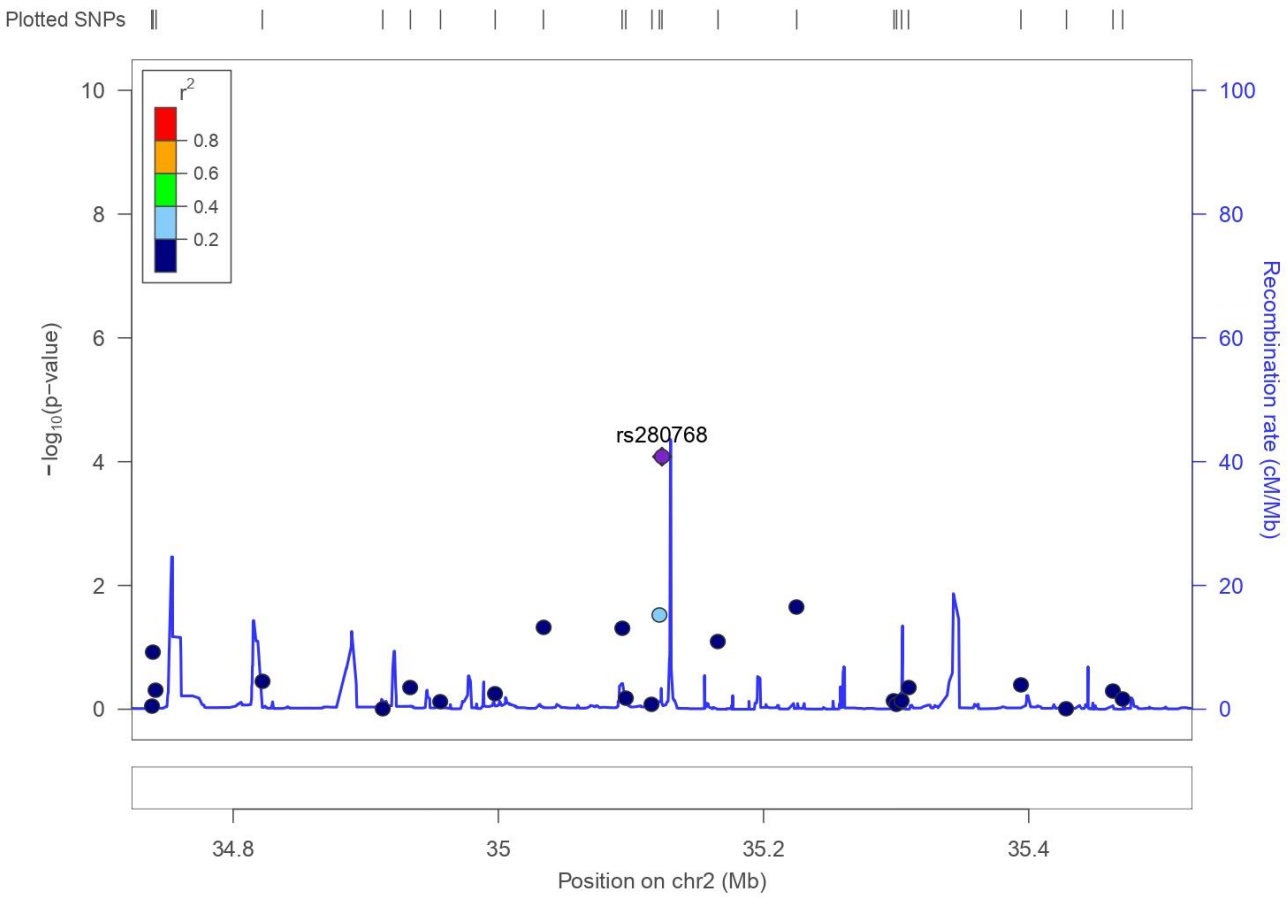
SNPs associated to colorectal cancer

SNP	CHROMOSOME
rs280768	2
rs6550962	3
rs10027212	4
rs10519732	5
rs4733560	8
rs10112382	8
rs6985894	8
rs12912791	15
rs325413	15
rs5005414	18
rs17769347	18



DISCUSSION

rs280768



C > T

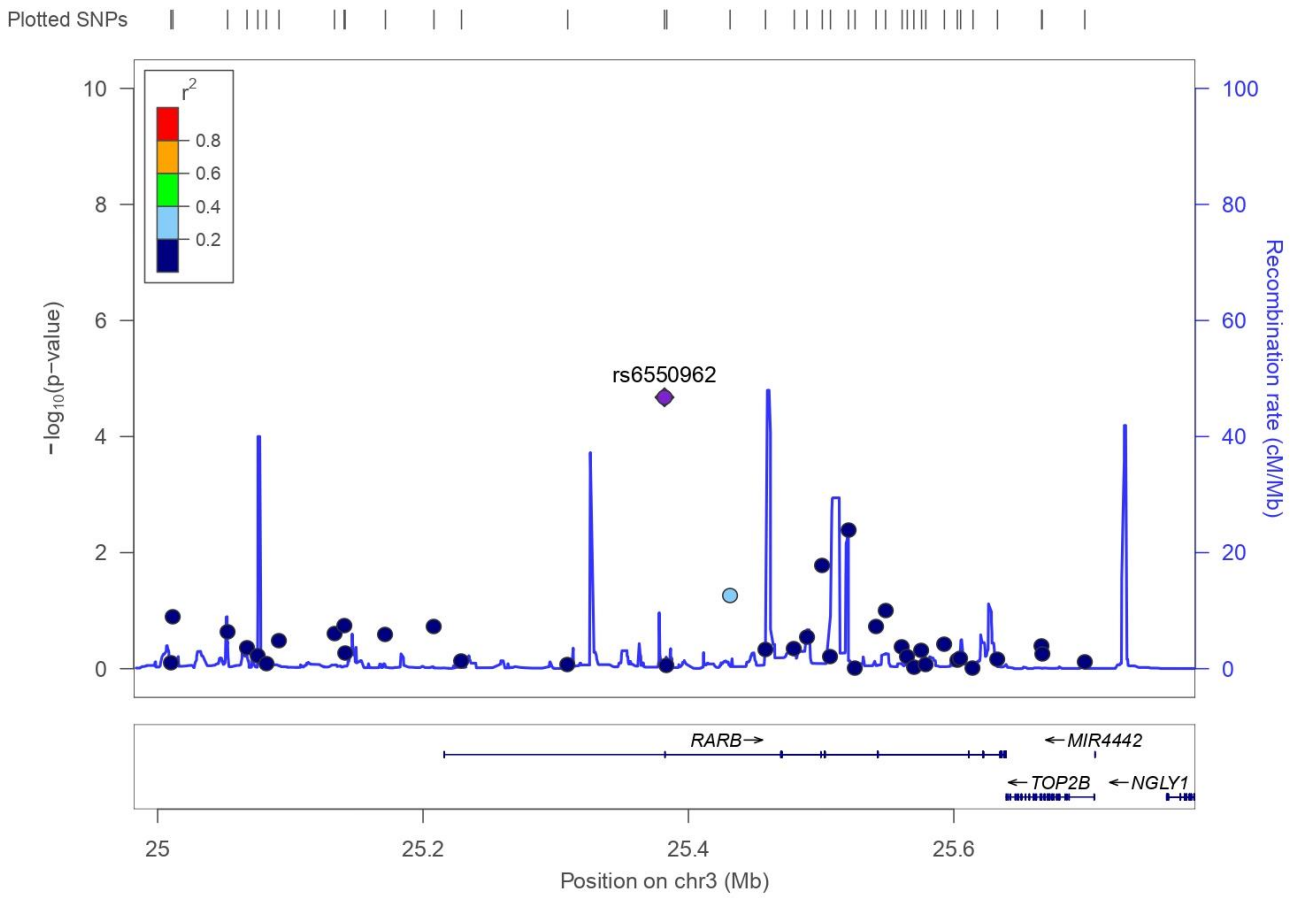
CHROMOSOME 2

FREQUENCY:

- C → 0,529
- T → 0,471

DISCUSSION

rs6550962



A > G

CHROMOSOME 3

- *RARB* gene

FREQUENCY:

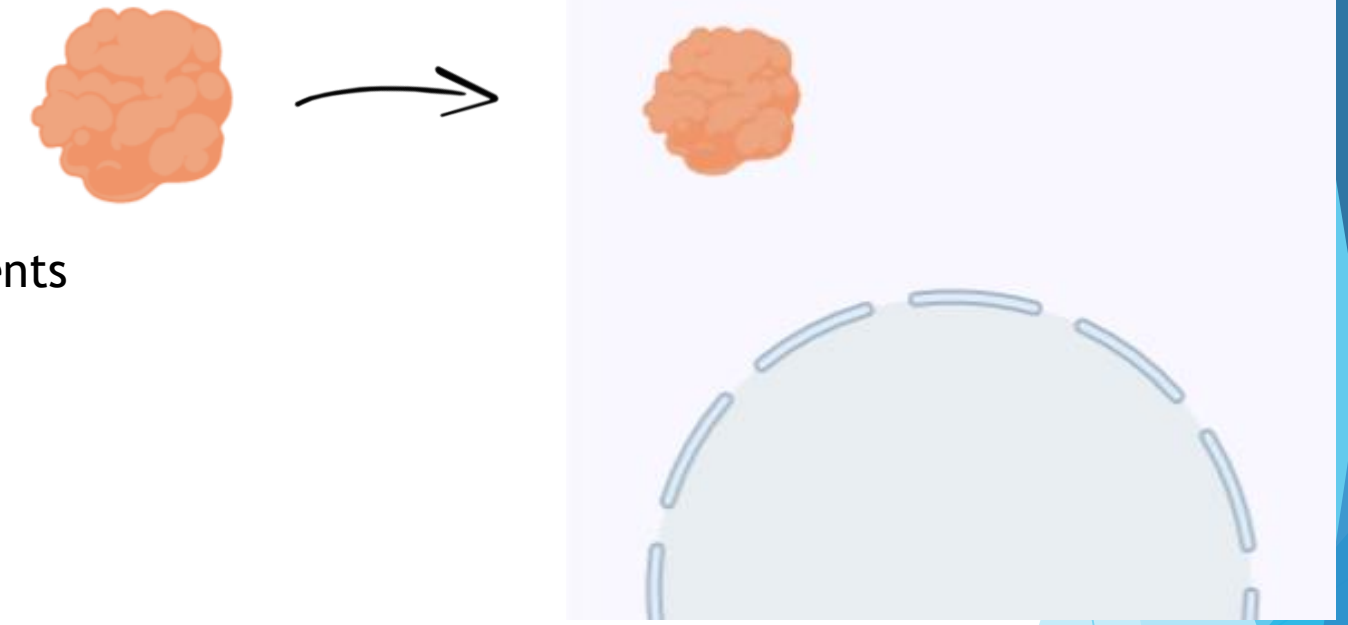
- A → 0,868
- G → 0,132

SO Term

- Genic Upstream Transcription Variant
- Intron variant

DISCUSSION

rs6550962 – *RARB* gene

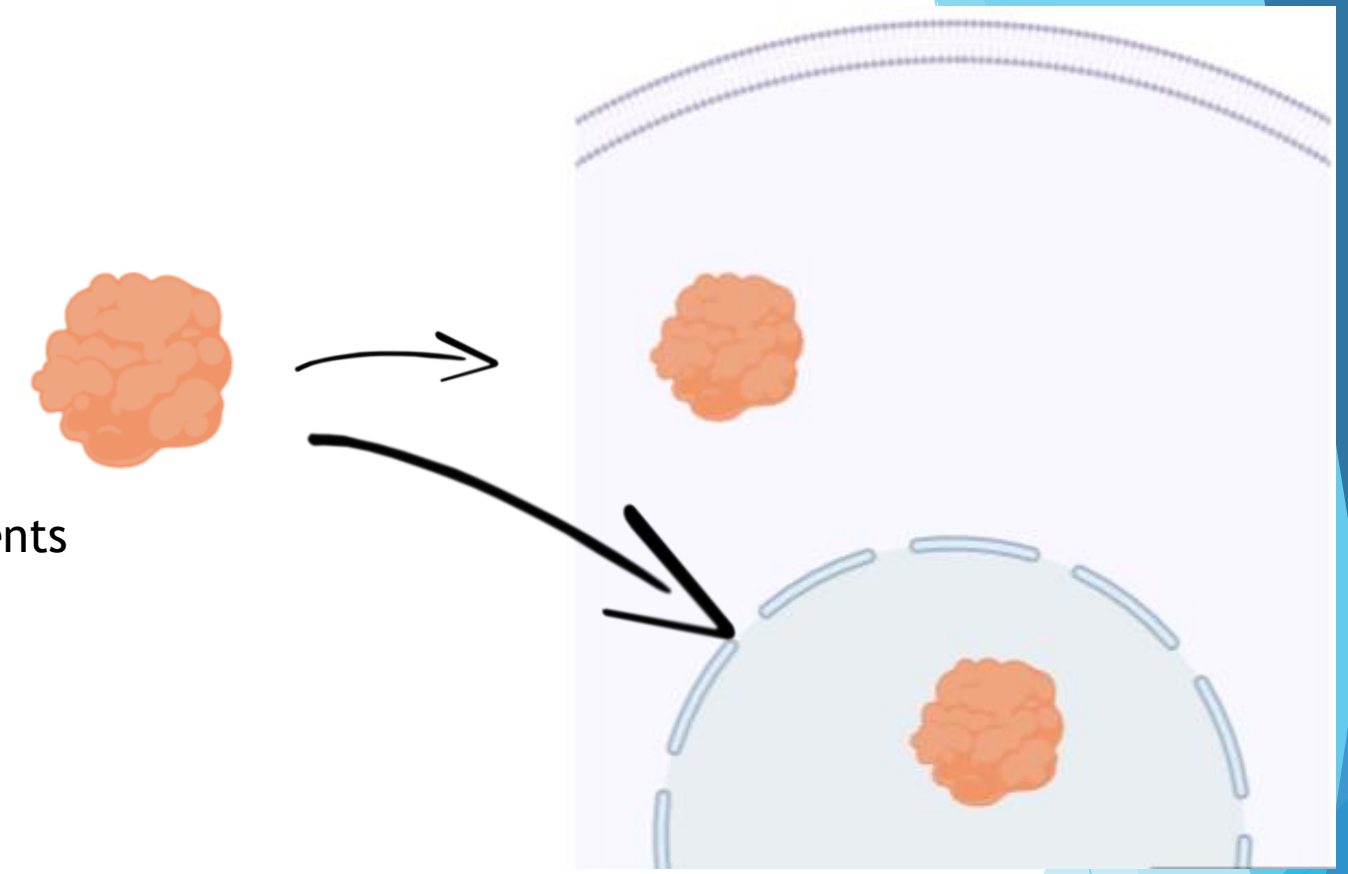


- Nuclear transcriptional regulator
- Cytoplasm and subnuclear compartments

DISCUSSION

rs6550962 – *RARB* gene

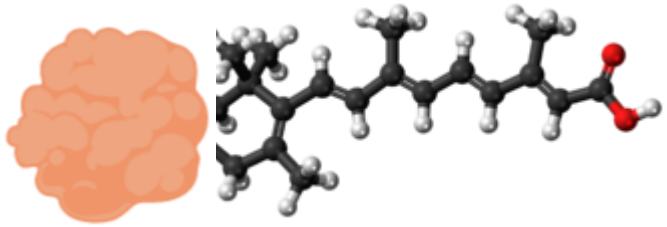
- Nuclear transcriptional regulator
- Cytoplasm and subnuclear compartments

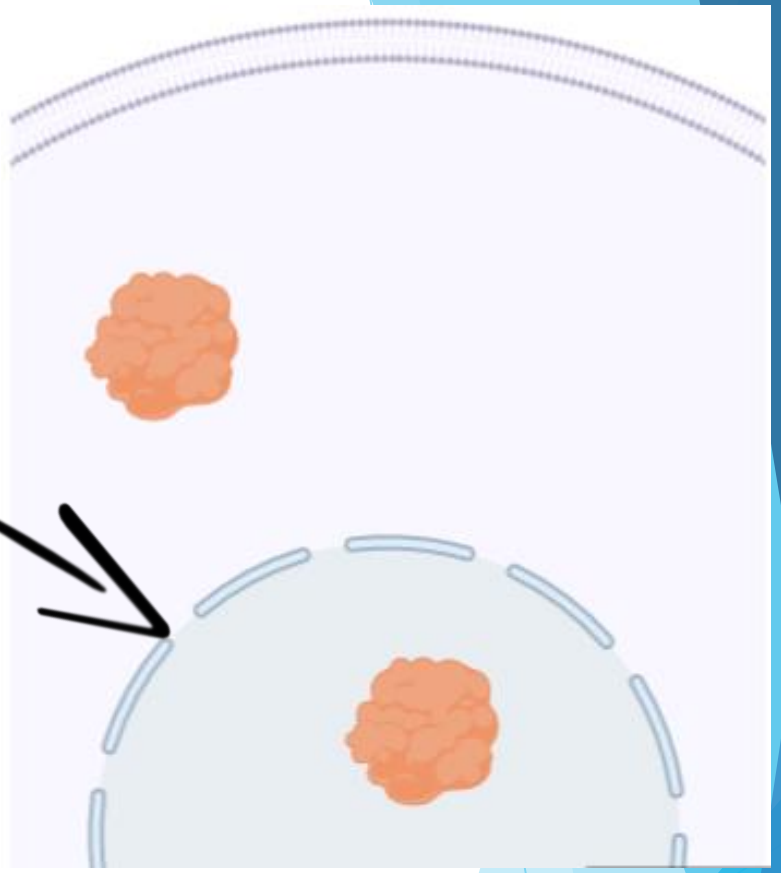


DISCUSSION

rs6550962 – *RARB* gene

- Nuclear transcriptional regulator
- Cytoplasm and subnuclear compartments





Cellular signalling in embryonic morphogenesis

Cell growth

Differentiation

¿ Cancer ?

DISCUSSION

rs6550962 – *RARB* gene

2019

Research Article

High Expression of RARβ Is a Favorable Factor in Colorectal Cancer

Wei Wang,[1] Shuang Liu,[2] Chunyi Jiang,[1] Yan Wang,[1] Huijun Zhu,[1] and XuDong Wang[3,4]

RARB expression was strongly correlated with several clinicopathological factors of colorectal cancer and may represent a favourable prognostic marker in patients with this cancer

DISCUSSION

rs6550962 – *RARB* gene

2019

Research Article

High Expression of RARβ Is a Favorable Factor in Colorectal Cancer

Wei Wang,[1] Shuang Liu,[2] Chunyi Jiang,[1] Yan Wang,[1] Huijun Zhu,[1] and XuDong Wang[3,4]

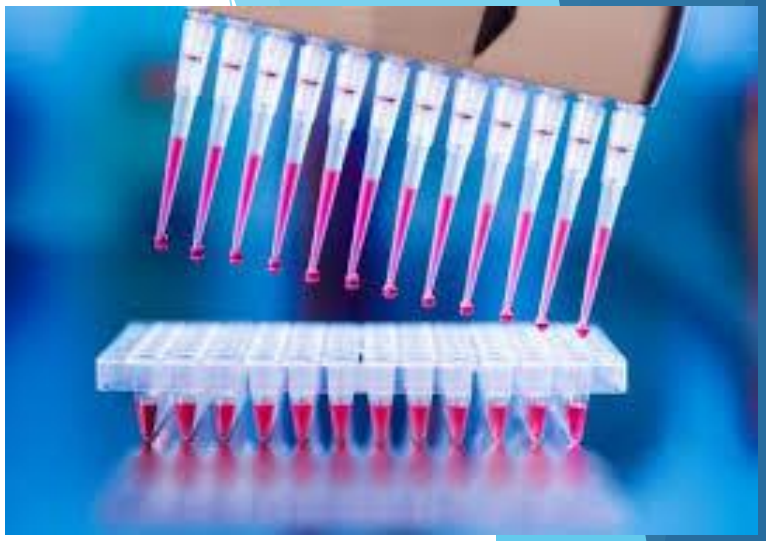
RARB expression was strongly correlated with several clinicopathological factors of colorectal cancer and may represent a favourable prognostic marker in patients with this cancer

COSMIC
Catalogue Of Somatic Mutations In Cancer

67 samples with mutation

2340 samples tested

DISCUSSION



rs6550962 – *RARB* gene

2019

Research Article

High Expression of RARβ Is a Favorable Factor in Colorectal Cancer

Wei Wang,[1] Shuang Liu,[2] Chunyi Jiang,[1] Yan Wang,[1] Huijun Zhu,[1] and XuDong Wang[3,4]

RARB expression was strongly correlated with several clinicopathological factors of colorectal cancer and may represent a favourable prognostic marker in patients with this cancer

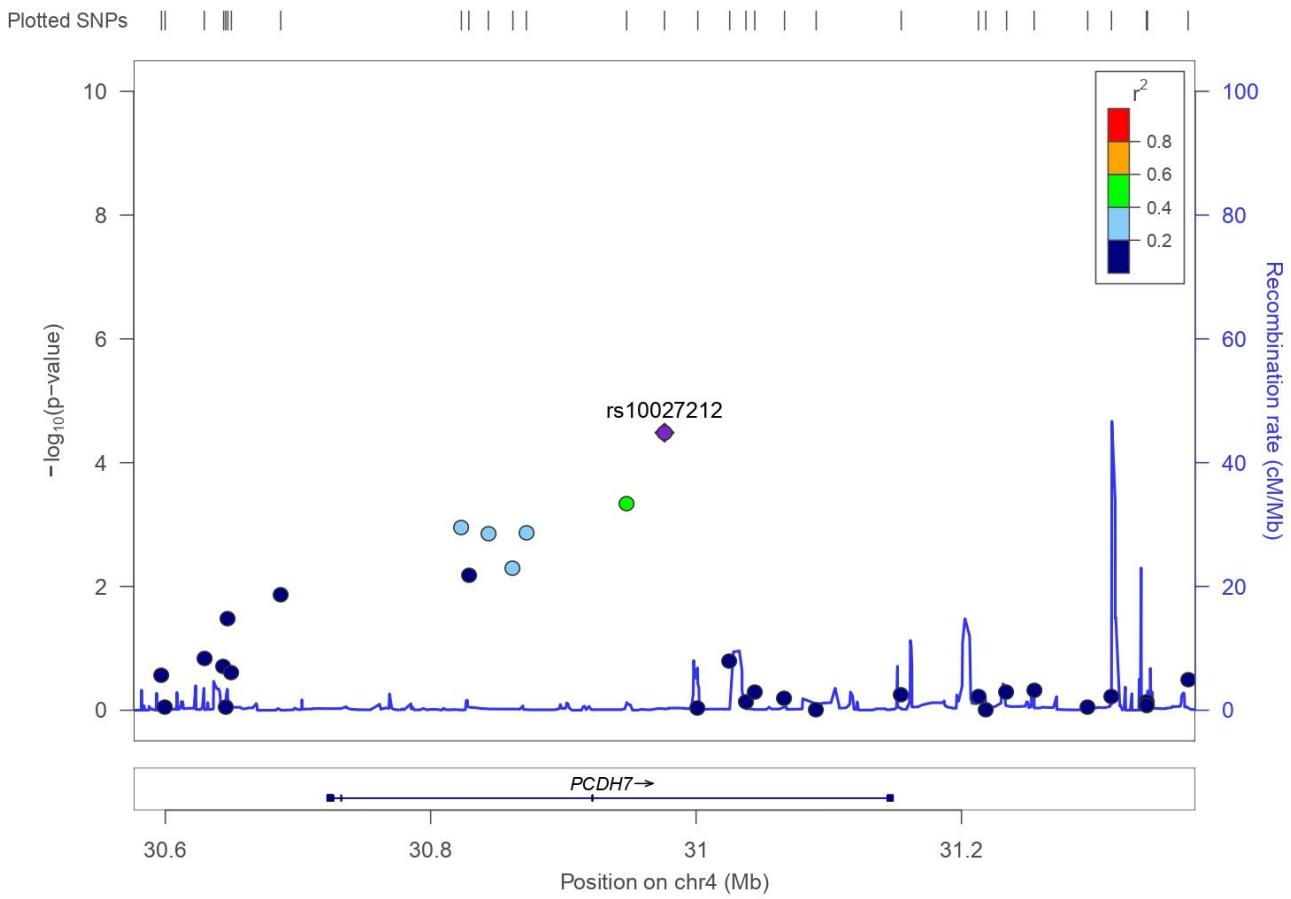
COSMIC
Catalogue Of Somatic Mutations In Cancer

67 samples with mutation

2340 samples tested

DISCUSSION

rs10027212



T > G

CHROMOSOME 4

- *PCDH7* gene

FREQUENCY:

- T → 0,544
- G → 0,456

SO Term

- Genic Downstream Transcription Variant
- Intron variant

DISCUSSION

rs10027212 – *PCDH7* gene

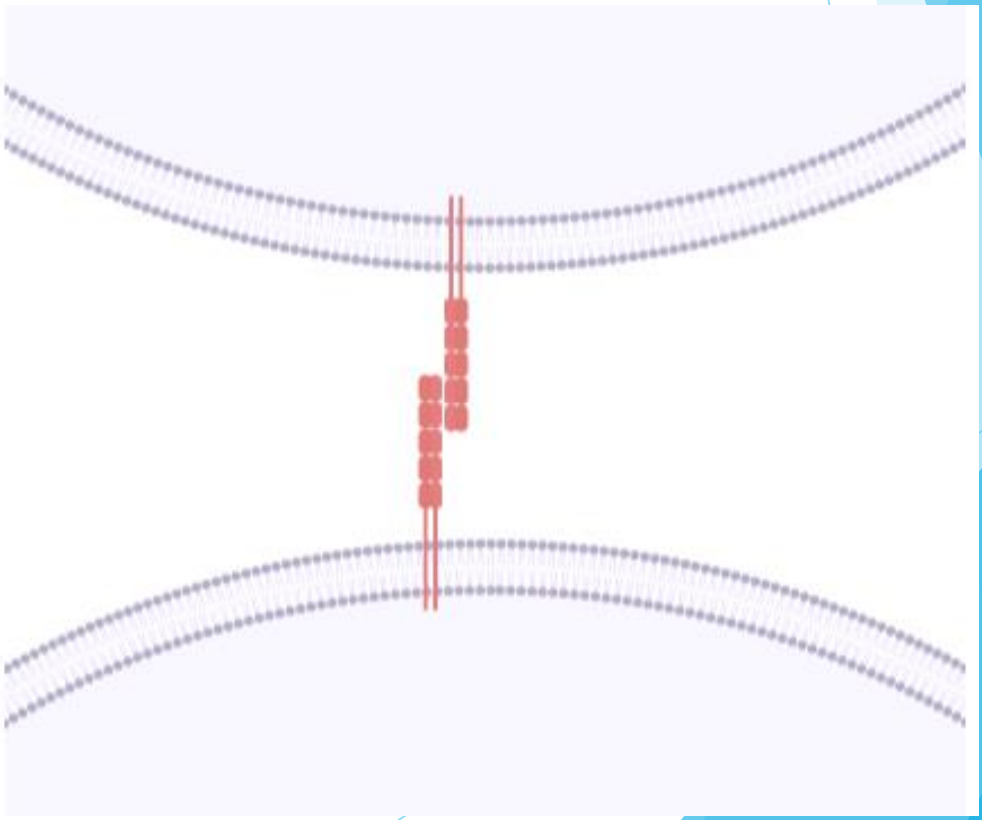
- Integral membrane protein

DISCUSSION

rs10027212 – *PCDH7* gene

- Integral membrane protein

- Cell-Cell recognition and adhesion



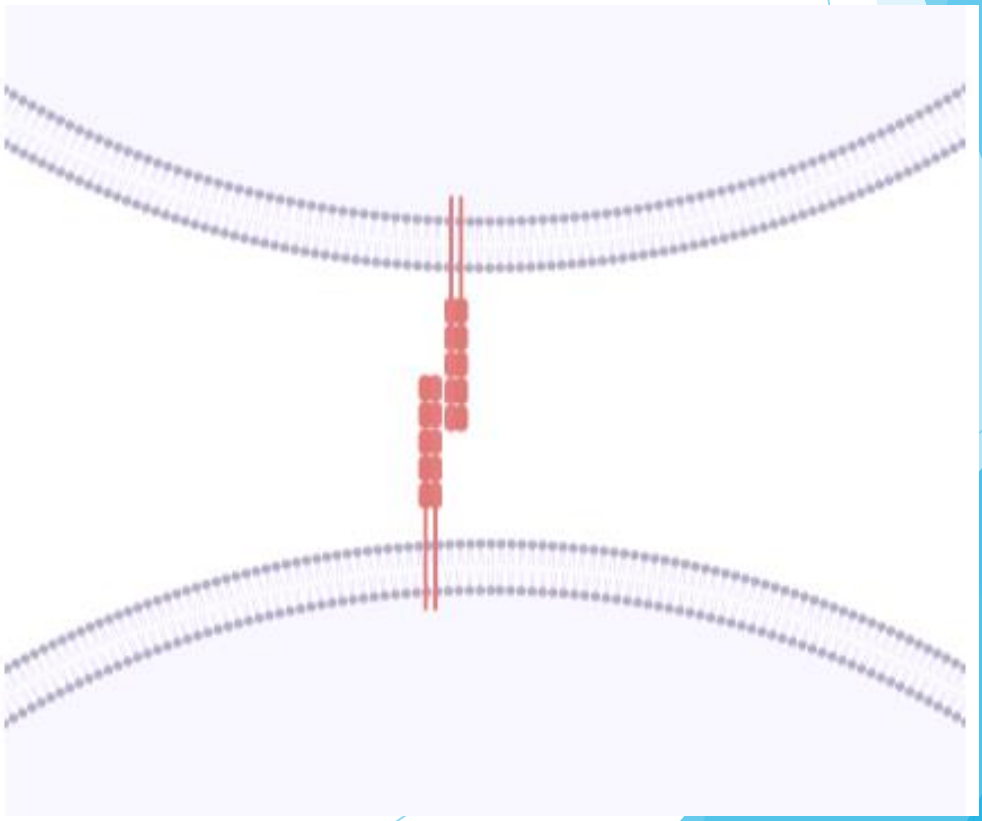
DISCUSSION

rs10027212 – *PCDH7* gene

- Integral membrane protein

¿ Cancer ?

- Cell-Cell recognition and adhesion



DISCUSSION

rs10027212 – *PCDH7* gene

2018

AQP8 inhibits colorectal cancer growth and metastasis by down-regulating PI3K/AKT signaling and PCDH7 expression

De Qing Wu[1,2], Zi Feng Yang[2], Ke Jian Wang[3], Xing Yu Feng[2], Ze Jian Lv[2], Yong Li[2], Zhi Xiang Jian[1,2]

They found molecular evidences which support the vital role of a novel AQP8-PCDH7 signalling axis in growth and metastasis of colorectal carcinoma.

It was already known that PCDH7 is overexpressed in several malignancies and are correlated with cancer cells metastasis.

DISCUSSION

rs10027212 - *PCDH7* gene

2018

AQP8 inhibits colorectal cancer growth and metastasis by down-regulating PI3K/AKT signaling and PCDH7 expression

De Qing Wu[1,2], Zi Feng Yang[2], Ke Jian Wang[3], Xing Yu Feng[2], Ze Jian Lv[2], Yong Li[2], Zhi Xiang Jian[1,2]

They found molecular evidences which support the vital role of a novel AQP8-PCDH7 signalling axis in growth and metastasis of colorectal carcinoma.

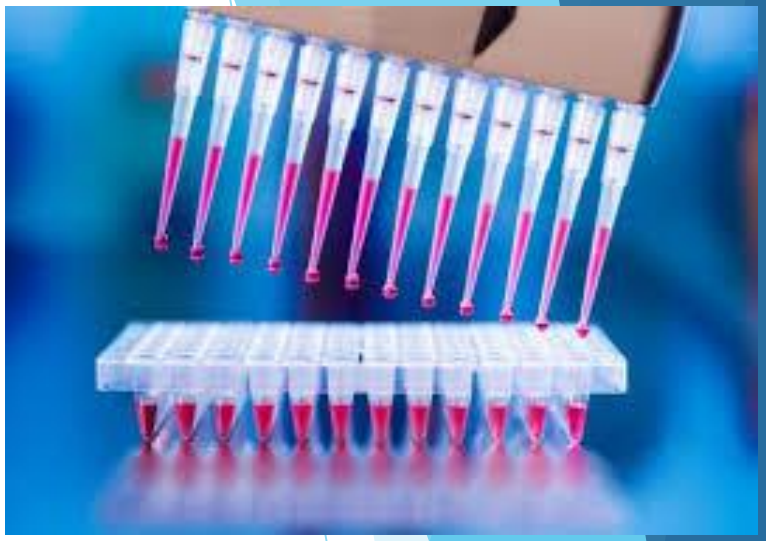
It was already known that PCDH7 is overexpressed in several malignancies and are correlated with cancer cells metastasis.

COSMIC
Catalogue Of Somatic Mutations In Cancer

125 samples with mutation

2314 samples tested

DISCUSSION



rs10027212 - *PCDH7* gene

2018

AQP8 inhibits colorectal cancer growth and metastasis by down-regulating PI3K/AKT signaling and PCDH7 expression

De Qing Wu[1,2], Zi Feng Yang[2], Ke Jian Wang[3], Xing Yu Feng[2], Ze Jian Lv[2], Yong Li[2], Zhi Xiang Jian[1,2]

They found molecular evidences which support the vital role of a novel AQP8-PCDH7 signalling axis in growth and metastasis of colorectal carcinoma.

It was already known that PCDH7 is overexpressed in several malignancies and are correlated with cancer cells metastasis.

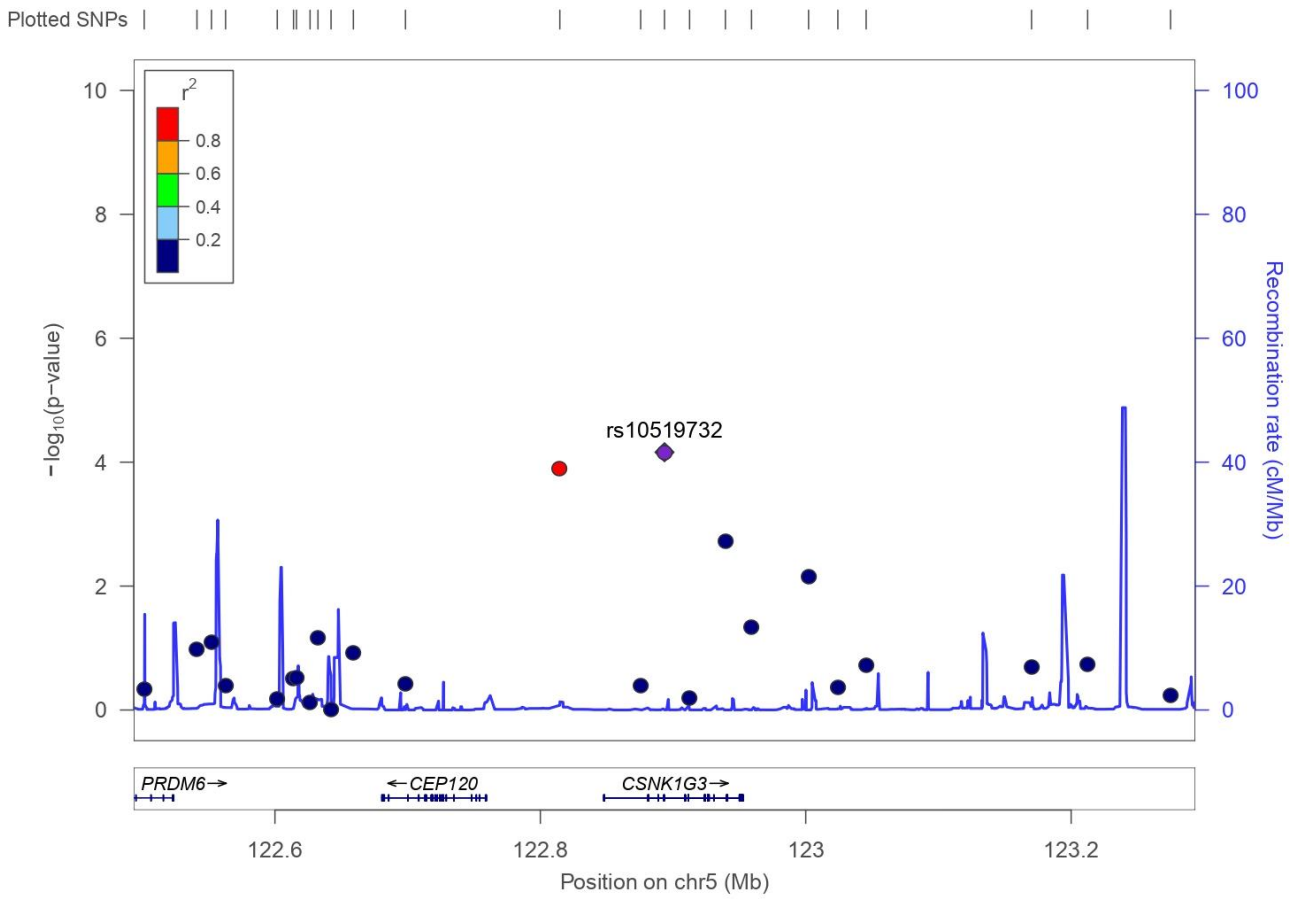
COSMIC
Catalogue Of Somatic Mutations In Cancer

125 samples with mutation

2314 samples tested

DISCUSSION

rs10519732



T > C

CHROMOSOME 5

- *CSNK163* gene

FREQUENCY:

- T → 0,893
- C → 0,107

SO Term

- Intron variant

DISCUSSION

rs10519732 – *CSNK163* gene

- Serine/Threonine protein kinases

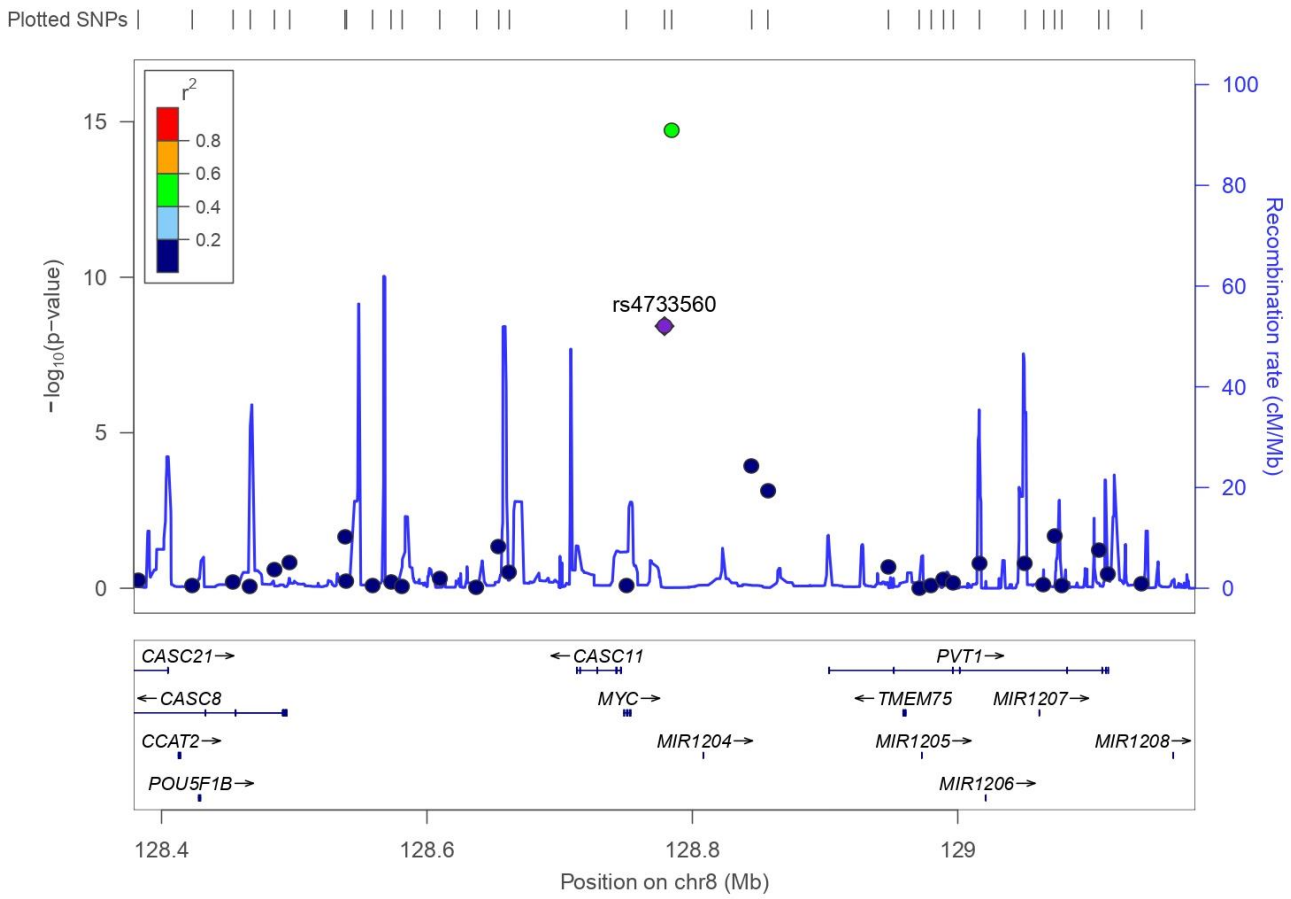
DISCUSSION

rs10519732 – *CSNK163* gene

- Serine/Threonine protein kinases

DISCUSSION

rs4733560



G > A

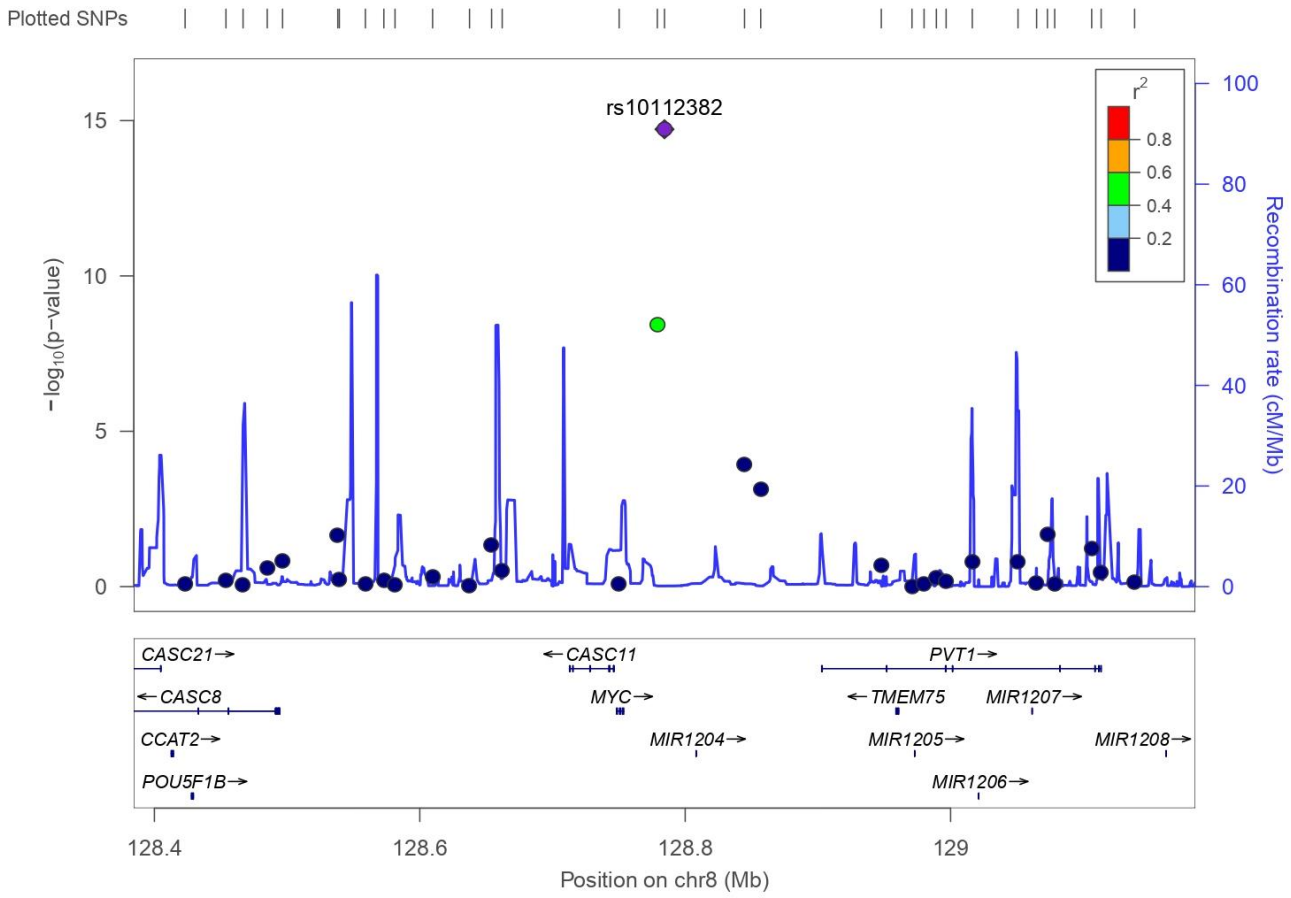
CHROMOSOME 8

FREQUENCY:

- G → 0,740
- A → 0,260

DISCUSSION

rs10112382



T > C

CHROMOSOME 8

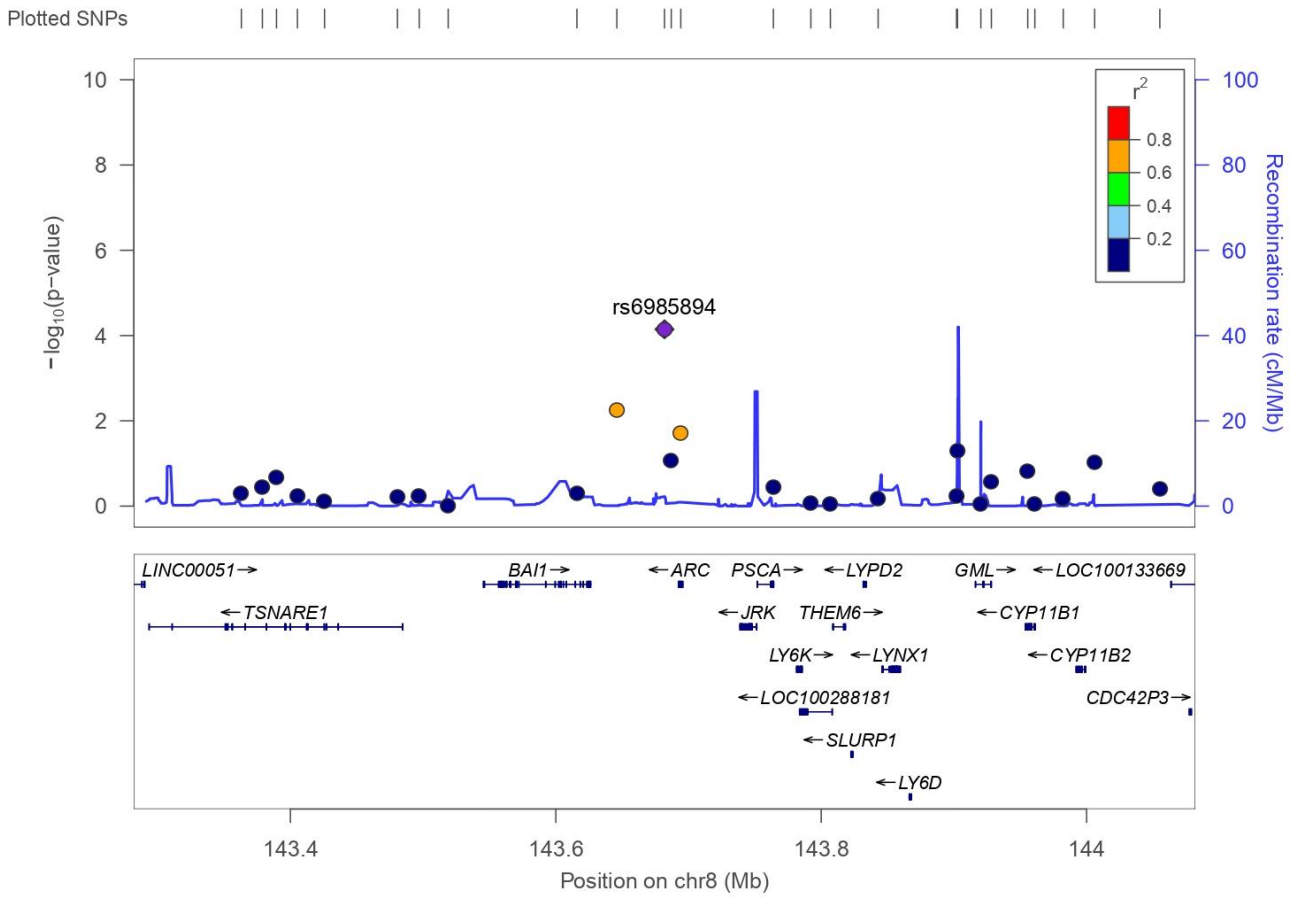
FREQUENCY:

- T → 0,339
- C → 0,661

DISCUSSION

A > G

rs6985894



CHROMOSOME 8

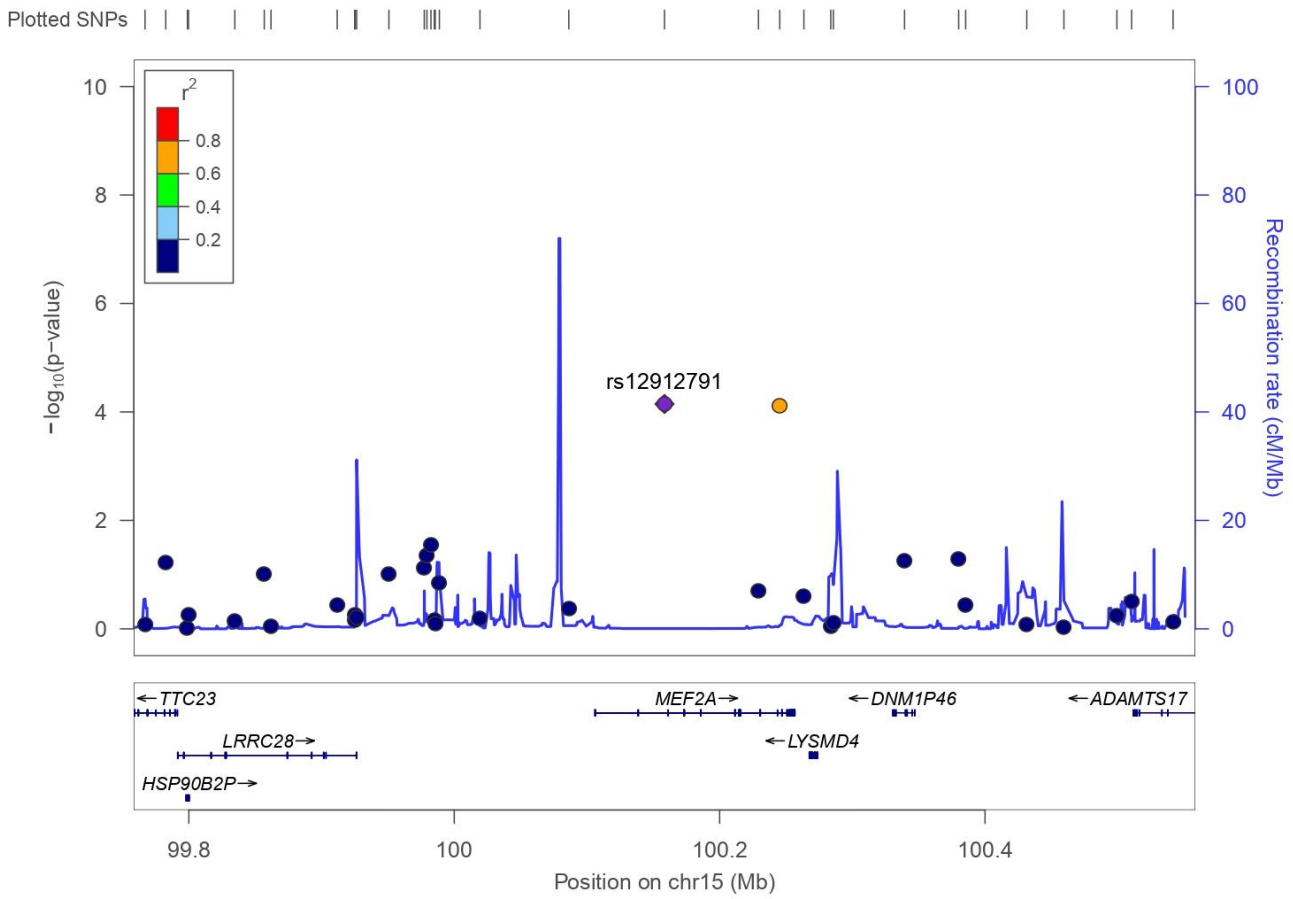
FREQUENCY:

- A → 0,218
- G → 0,782

DISCUSSION

T > C

rs12912791



CHROMOSOME 15

- *MEF2A* gene

FREQUENCY:

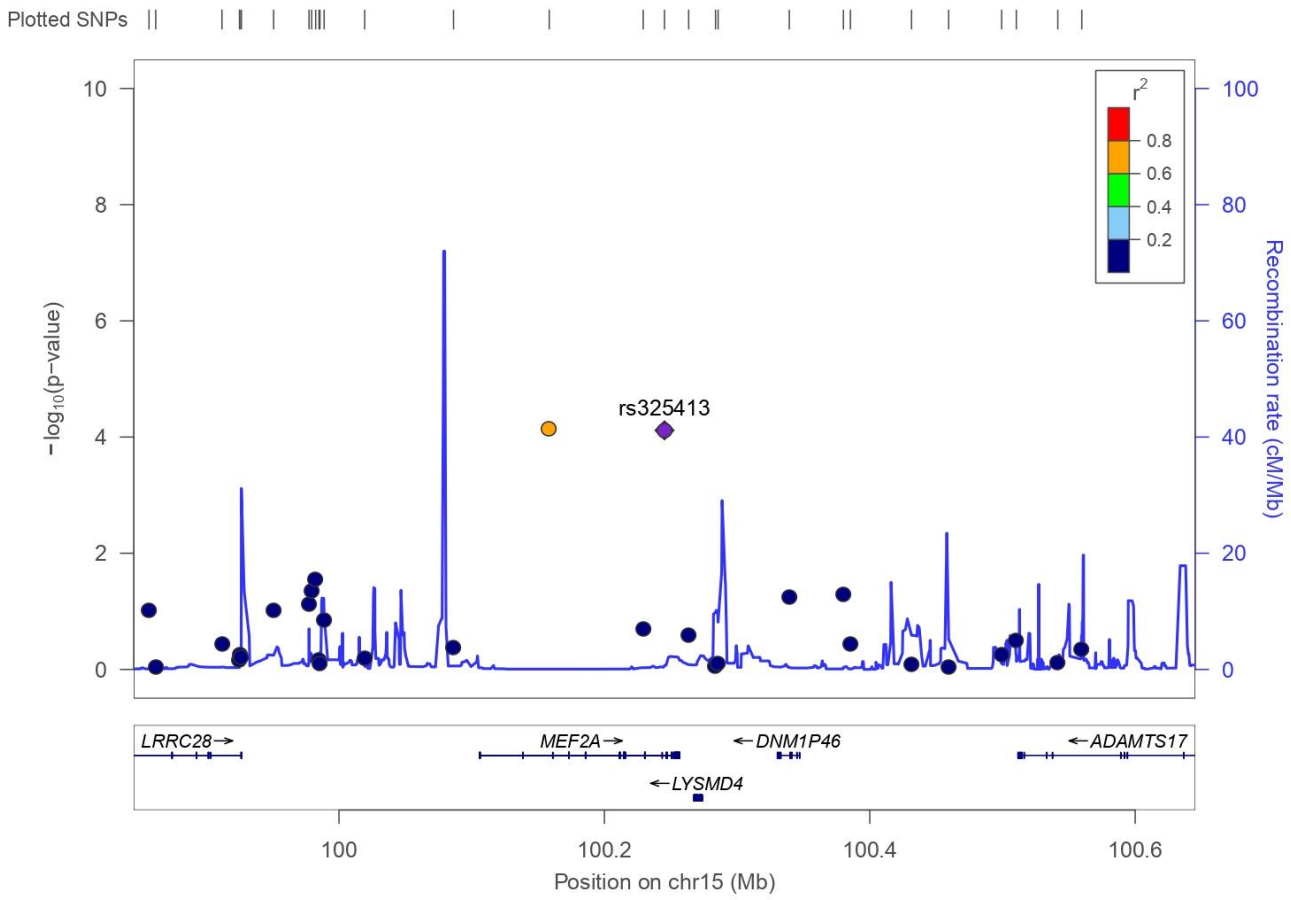
- T → 0,732
- C → 0,268

SO Term

- Intron variant
- Genic Upstream Transcript Variant

DISCUSSION

rs325413



G > A

CHROMOSOME 15

- *MEF2A* gene

FREQUENCY:

- G → 0,379
- A → 0,621

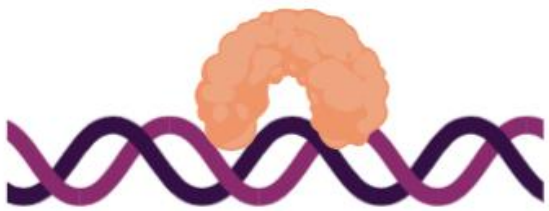
SO Term

- Intron variant
- Genic Downstream Transcript Variant

DISCUSSION

rs12912791 and rs325413 – *MEF2A* gene

- DNA-binding transcription factor



Muscle-specific genes

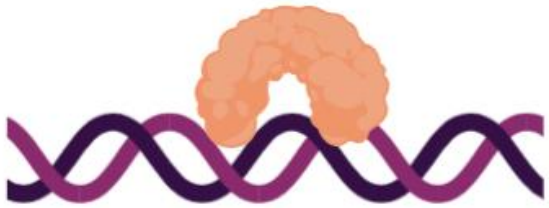
Growth factor-induced genes

Stress-induced genes

DISCUSSION

rs12912791 and rs325413 – *MEF2A* gene

- DNA-binding transcription factor



Muscle-specific genes

Growth factor-induced genes

Stress-induced genes

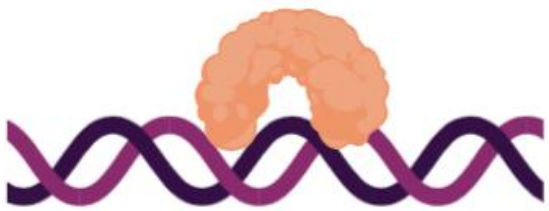
¿ Cancer ?

- Defects could be a cause of autosomal dominant coronary artery disease 1 with myocardial infarction (ADCAD1)

DISCUSSION

rs12912791 and rs325413 – *MEF2A* gene

- DNA-binding transcription factor



Muscle-specific genes

Growth factor-induced genes

Stress-induced genes

¿ Cancer ?

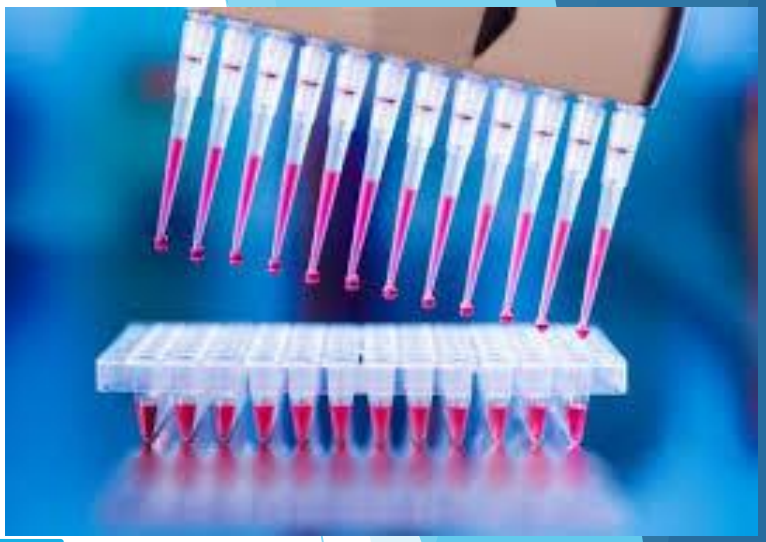
- Defects could be a cause of autosomal dominant coronary artery disease 1 with myocardial infarction (ADCAD1)

COSMIC
Catalogue Of Somatic Mutations In Cancer

54 samples with mutation

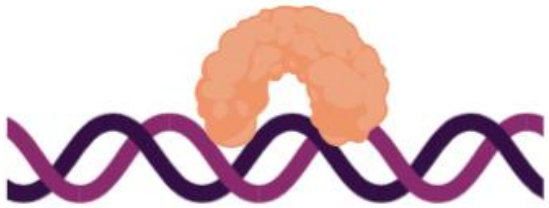
2303 samples tested

DISCUSSION



rs12912791 and rs325413 – *MEF2A* gene

- DNA-binding transcription factor



- Muscle-specific genes
- Growth factor-induced genes
- Stress-induced genes

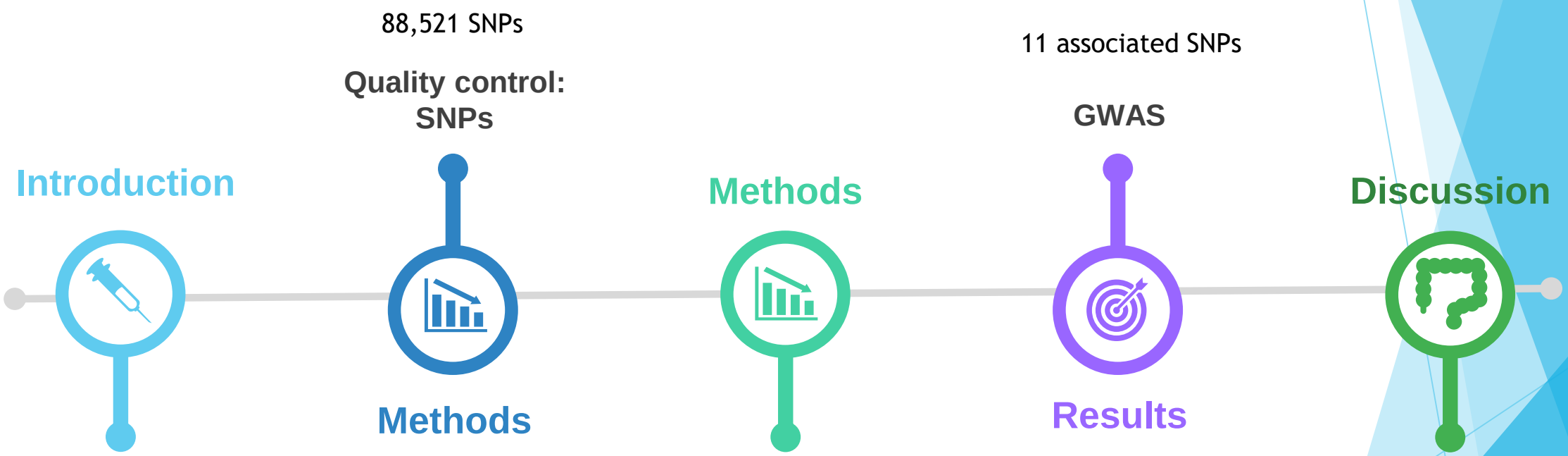
- Defects could be a cause of autosomal dominant coronary artery disease 1 with myocardial infarction (ADCAD1)

COSMIC
Catalogue Of Somatic Mutations In Cancer

54 samples with mutation

2303 samples tested

SUMMARY


Introduction
DNA samples collection
2,312 individuals
100,000 SNPs
88,521 SNPs
Quality control: SNPs
Methods
Methods
Quality control: individuals
2,243 individuals
11 associated SNPs
GWAS
Results
Discussion
SNPs analysis
4 SNPs considered for molecular studies