Analyzing GWAS Data

- Each SNP is an independent test
- Associations are tested by comparing the frequency of each allele in cases and controls
- The frequency of each of 3 possible genotypes can also be compared

Table 3. Association of Alleles and Genotypes of rs6983267 on Chromosome 8q24 With Colorectal Cancer[a]

	Number and Frequency of rs6983267 Alleles in Colorectal Cancer					Number and Frequency of rs6983267 Genotypes in Colorectal Cancer						
	C	T	χ^2 (1*df*)	*P* Value	OR	CC	CT	TT	χ^2 (2*df*)	*P* Value	OR	OR
Cases	875 (56.5)	675 (43.5)	24.8	6.3×10^{-7}	1.35[b]	250 (32.3)	375 (48.4)	150 (19.4)	24.5	4.7×10^{-6}	1.33[c]	1.81[d]
Controls	1860 (48.9)	1940 (51.1)				460 (24.2)	940 (49.4)	500 (26.3)				

Abbreviation: OR, odds ratio.
[a] Data are hypothetical; adapted from Tomlinson et al.[56]
[b] Denotes allelic odds ratio.
[c] Denotes heterozygote odds ratio.
[d] Denotes homozygote odds ratio.

Pearson *et al. JAMA*, 2008

Odds ratios

- measure of effect size, or strength of association
- odds = P / (1-P)
- Probability of winning is 50%:
  - odds is 0.5 / (1-0.5) = 1 (1 to 1, 50:50, "even money")
- If probably of winning is 75%
  - odds is 0.75 / (1-0.75) = 3
- Odds ratio = $\frac{\text{odds(event | exposure)}}{\text{odds(event | lack of exposure)}}$

Odds ratios

- P (D | genotype "AT") = 0.8
- P (D | genotype "TT") = 0.2
- OR for getting the disease with genotype AT compared to TT?
  - OR = (0.8 / 0.2) / (0.2 / 0.8) = 16
- What's the OR for AT individuals relative to an average population risk of 25%?
  - OR = (0.8 / 0.2) / (0.25 / 0.75) = 12

Analyzing a SNP for association
Allele Counting Method

Association of rs6983267 on 8q24 with colorectal cancer

	C	T
Cases	875 (56.5)	675 (43.5)
Controls	1860 (48.9)	1940 (51.1)

C is the risk allele

OR_C = odds(disease | C) / odds(disease | T)
= 875*1940 / 1860*675 = 1.35

Cases

C alleles = 2 * 250 "CC" + 375 "CT" = 875
T alleles = 2 * 150 "TT" + 375 "CT" = 675

Controls

C alleles = 2 * 460 "CC" + 940 "CT" = 1860
T alleles = 2 * 500 "TT" + 940 "CT" = 1940

Analyzing a SNP for association
Genotype Counting Method

Association of rs6983267 on 8q24 with colorectal cancer

	CC	CT	TT
Cases	250	375	150
Controls	460	940	500

C is the risk allele

These ORs are all relative to TT (lowest-risk genotype) - not meaningful to apply to an individual

OR_{TT} = odds(disease | TT) / odds(disease | TT) = 1

OR_{CT} = odds(disease | CT) / odds(disease | TT) = 375*500 / 150*940 = 1.33

OR_{CC} = odds(disease | CC) / odds(disease | TT) = 250*500 / 460*150 = 1.81

Analyzing a SNP for association
Likelihood Ratios

Association of rs6983267 on 8q24 with colorectal cancer

	CC	CT	TT	
Cases	250	375	150	775
Controls	460	940	500	1900

These LRs account for genotype frequencies in the population - can be directly applied to an individual

LR_{CC} = P(CC | disease) / P(CC | control) = (250/775) / (460/1900) = 1.33

LR_{CT} = P(CT | disease) / P(CT | control) = (375/775) / (940/1900) = 0.978

LR_{TT} = P(TT | disease) / P(TT | control) = (150/775) / (500/1900) = 0.735

Analyzing a SNP for association

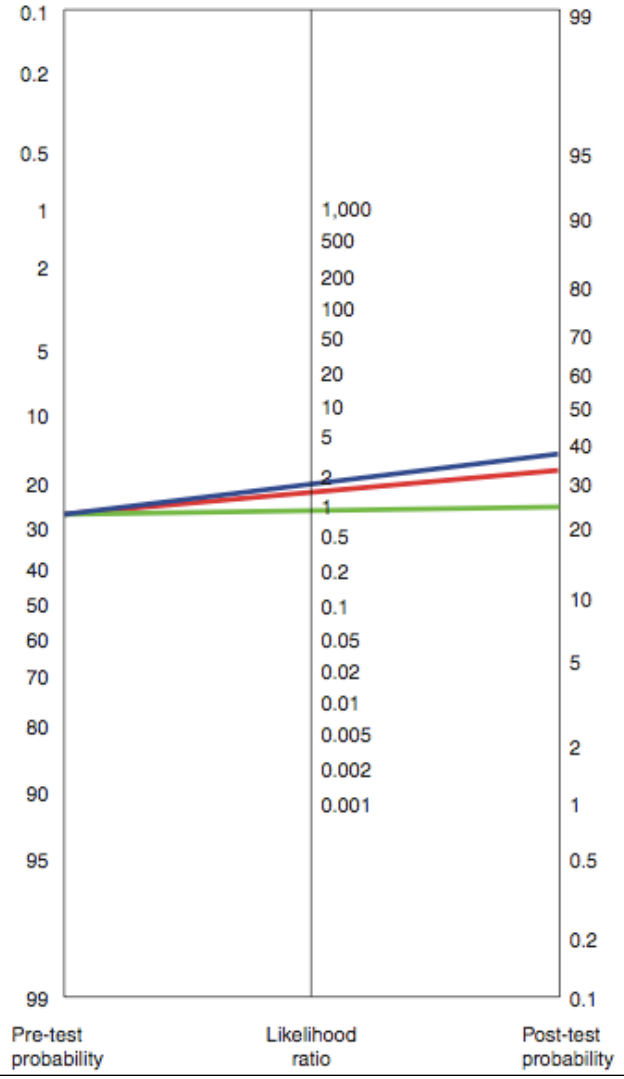
Likelihood Ratios

Likelihood ratio = Probability of genotype in diseased person/ Probability of genotype in non-diseased person = LR_i

Pre-test odds = Probability of disease/1 - Probability of disease

Pre-test odds × LR_1 × LR_2 ×...× LR_n = Post-test odds

Post-test probability = Post-test odds/Post-test odds + 1

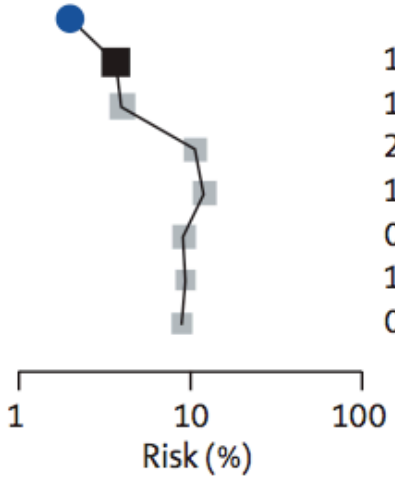


Analyzing a SNP for association

Likelihood Ratios

A Myocardial infarction

Gene*	SNP location	Patient genotype	LR	Studies†	Samples‡	Post-test probability (%)
						2·0%
LPA	rs3798220	CT	1·86	2	17031	3·7%
THBS2	rs8089	AC	1·09	1	4868	4·0%
LDLR	rs14158	GG	2·88	1	3542	10·6%
LIPC	rs11630220	AG	1·15	1	3542	12·0%
ESR2	rs1271572	CC	0·73	1	3089	9·1%
	rs35410698	GG	1·03	1	1094	9·4%
FXN	rs3793456	AA	0·94	1	1094	8·9%



Adjusting odds ratios

- GWAS most often report ORs relative to the low-risk allele or lowest-risk genotype
- To turn this into a meaning risk estimate, the prevalence of the disease and the genotype frequencies must be taken into account
- P(D) = prevalence

  $= P(D|AA)P(AA) + P(D|Aa)P(Aa) + P(D|aa)P(aa)$

  (incorporate genotype frequencies)

Odds ratios to Probabilities

rs2383207 is associated with risk of myocardial infarction (from pre-course survey)

P(D) = Prevalence = 0.212 = P(D|AA)* P(AA) + P(D | AG)* P(AG) + P(D | GG) * P(GG)

From CEU HapMap population
P(AA) = frequency in population = 0.183
P(AG) = frequency in population = 0.583
P(GG) = frequency in population = 0.233

P(D | "AA") = 17.5% (low risk homozygote)

P(D | "AG") = ? (can be determined from equation above)

P(D | "GG") = 24.8%

OR_{GG} = odds(disease | GG) / odds(disease | AA)
= (0.248/ (1-0.248)) / (0.175/(1-0.175)) = 1.55

OR_{GG} NOT (0.248/0.175)!!

OR^*_{GG} = odds(disease | GG) / odds(disease in avg pop)
= (0.248/ (1-0.248)) / (0.212/(1-0.212)) = 1.22

These OR*s are relative to the average population - can be directly applied to an individual

Odds ratios to Probabilities

OR* (i.e., odds ratios relative to the average population) for each SNP can be multiplied together to combine effects from multiple SNPs (OR^*_{SNP1} x OR^*_{SNP2} x OR^*_{SNP3} ... = Total OR*)

Total OR* = odds(disease | all SNP genotypes) / odds(disease in avg pop)

MI/rs2383207 example:
if OR* = 1.22 = odds(disease | genotype data) / odds(disease in avg pop)
= [P(D | G) / (1 - P(D | G)] / [0.212 / (1-0.212)]

let x = P(D|G))
0.328 = x / (1-x)
x = 0.328 / (1+0.328) = 0.248
i.e.your probability of disease given the genotype data is 24.8%

Odds ratios to Probabilities

To integrate other non-genetic risk factors, just multiply by the corresponding odds ratios:

$OR^*_{SNP1} \times OR^*_{SNP2} \times OR^*_{SNP3} \times OR^*_{cholesterol} \times OR^*_{exercise} = \text{Total gene + env } OR^*$

genetic risk | environmental risk

Overall probability can be derived now as before:

Consider someone with low-risk SNPs but high-risk cholesterol or lifestyle factor
Total OR* = 0.80 (genetic) x 4.0 (environment) = 3.2
OR* = 3.2 = odds(disease | gene+env data) / odds(disease in avg pop)
= [P(D | G+E data) / (1 - P(D | G+E data)] / [0.212 / (1-0.212)]

let x = P(D| G+E data))
0.861 = x / (1-x)
x = 0.861 / (1+0.861) = 0.46

i.e. overall probability of disease given the genetic and environmental data is 46%!
(much higher than average population risk of 21.2% and that predicted by SNPs alone)

Diabetes Risk Calculation Exercise

Calculate the diabetes risk for 4 individuals using their genotype data using odds ratios vs. likelihood ratios

Note whether the two methods of risk estimation provide concordant or discordant results

What is the effect size of the 2-6 SNPs examined in this analysis? Relatively small or large? Would effect sizes of environmental risk factors be larger or smaller?

Patient 1

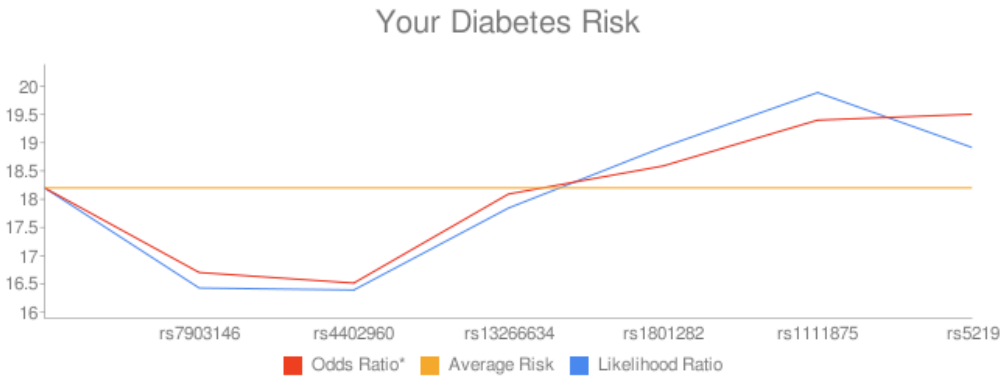
European Female

Diabetes Risk for Personal Genotype

Population given was: CEU
Gender given was: female

Pre-test probability for Caucasian female is: 18.200% (Odds = 0.222)

RSID	Genotype	Study Size	Odds Ratio*	Post-test odds (OR*)	Post-test probability (OR*)	Likelihood Ratio	Post-test odds (LR)	Post-test probability (LR)
Pre-test:				**0.222**	**18.2%**		**0.222**	**18.2%**
rs7903146	CC	69451	0.901	0.200	16.7%	0.883	0.197	16.4%
rs4402960	GT	11765	0.987	0.198	16.5%	0.997	0.196	16.4%
rs13266634	CC	6674	1.117	0.221	18.1%	1.108	0.217	17.8%
rs1801282	CC	2335	1.034	0.228	18.6%	1.075	0.233	18.9%
rs1111875	CT	1421	1.054	0.241	19.4%	1.064	0.248	19.9%
rs5219	CC	1034	1.007	0.242	19.5%	0.940	0.233	18.9%
Total:				**0.242**	**19.5%**		**0.233**	**18.9%**



Patient 3

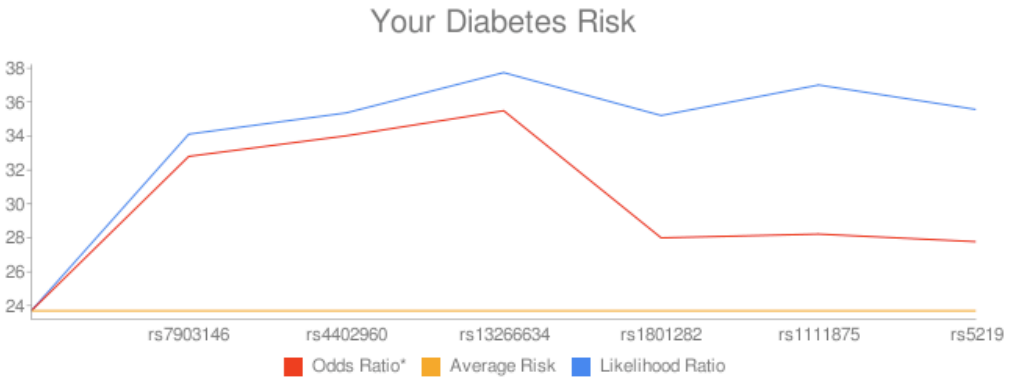
European Male

Diabetes Risk for Personal Genotype

Population given was: CEU
Gender given was: Male

Pre-test probability for Caucasian Male is: 23.700% (Odds = 0.311)

RSID	Genotype	Study Size	Odds Ratio*	Post-test odds (OR*)	Post-test probability (OR*)	Likelihood Ratio	Post-test odds (LR)	Post-test probability (LR)
Pre-test:				**0.311**	**23.7%**		**0.311**	**23.7%**
rs7903146	TT	69451	1.571	0.488	32.8%	1.667	0.518	34.1%
rs4402960	TT	11765	1.056	0.515	34.0%	1.057	0.547	35.4%
rs13266634	CC	6674	1.067	0.550	35.5%	1.108	0.606	37.7%
rs1801282	CG	2335	0.707	0.389	28.0%	0.897	0.543	35.2%
rs1111875	CC	1421	1.011	0.393	28.2%	1.081	0.587	37.0%
rs5219	CC	1034	0.978	0.384	27.8%	0.940	0.552	35.6%
Total:				**0.384**	**27.8%**		**0.552**	**35.6%**



Patient 4

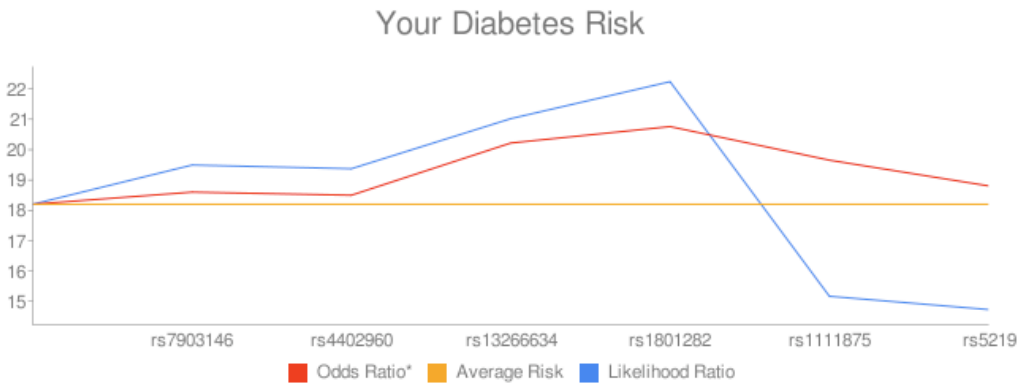
European Female

Diabetes Risk for Personal Genotype

Population given was: CEU
Gender given was: female

Pre-test probability for Caucasian female is: 18.200% (Odds = 0.222)

RSID	Genotype	Study Size	Odds Ratio*	Post-test odds (OR*)	Post-test probability (OR*)	Likelihood Ratio	Post-test odds (LR)	Post-test probability (LR)
Pre-test:				**0.222**	**18.2%**		**0.222**	**18.2%**
rs7903146	CT	69451	1.027	0.229	18.6%	1.088	0.242	19.5%
rs4402960	GG	11765	0.993	0.227	18.5%	0.992	0.240	19.4%
rs13266634	CC	6674	1.117	0.253	20.2%	1.108	0.266	21.0%
rs1801282	CC	2335	1.034	0.262	20.8%	1.075	0.286	22.2%
rs1111875	TT	1421	0.934	0.245	19.7%	0.625	0.179	15.2%
rs5219	CT	1034	0.947	0.232	18.8%	0.966	0.173	14.7%
Total:				**0.232**	**18.8%**		**0.173**	**14.7%**



Patient 6

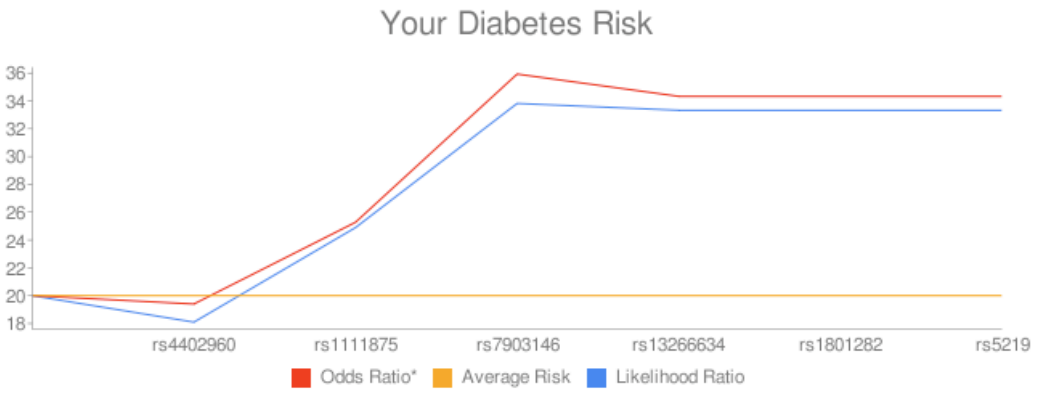
Japanese Male

Diabetes Risk for Personal Genotype

Population given was: JPT
Gender given was: Male

Pre-test probability for Japanese Male is: 20.000% (Odds = 0.250)

RSID	Genotype	Study Size	Odds Ratio*	Post-test odds (OR*)	Post-test probability (OR*)	Likelihood Ratio	Post-test odds (LR)	Post-test probability (LR)
Pre-test:				**0.250**	**20.0%**		**0.250**	**20.0%**
rs4402960	GG	3543	0.963	0.241	19.4%	0.884	0.221	18.1%
rs1111875	CC	3543	1.405	0.338	25.3%	1.500	0.331	24.9%
rs7903146	CT	2877	1.658	0.561	35.9%	1.542	0.511	33.8%
rs13266634	CT	1728	0.932	0.523	34.3%	0.978	0.500	33.3%
rs1801282	CC	None	Not found	-	-	Not found	-	-
rs5219	CT	None	Not found	-	-	Not found	-	-
Total:				**0.523**	**34.3%**		**0.500**	**33.3%**



Patient 7

Chinese Male

Diabetes Risk for Personal Genotype

Population given was: CHB
Gender given was: Male

Pre-test probability for Chinese Male is: 20.000% (Odds = 0.250)

RSID	Genotype	Study Size	Odds Ratio*	Post-test odds (OR*)	Post-test probability (OR*)	Likelihood Ratio	Post-test odds (LR)	Post-test probability (LR)
Pre-test:				**0.250**	**20.0%**		**0.250**	**20.0%**
rs13266634	TT	2443	0.975	0.244	19.6%	0.864	0.216	17.8%
rs1111875	CT	1410	1.044	0.255	20.3%	1.087	0.235	19.0%
rs7903146	CC	None	Not found	-	-	Not found	-	-
rs1801282	CC	None	Not found	-	-	Not found	-	-
rs5219	TT	None	Not found	-	-	Not found	-	-
rs4402960	GT	None	Not found	-	-	Not found	-	-
Total:				**0.255**	**20.3%**		**0.235**	**19.0%**

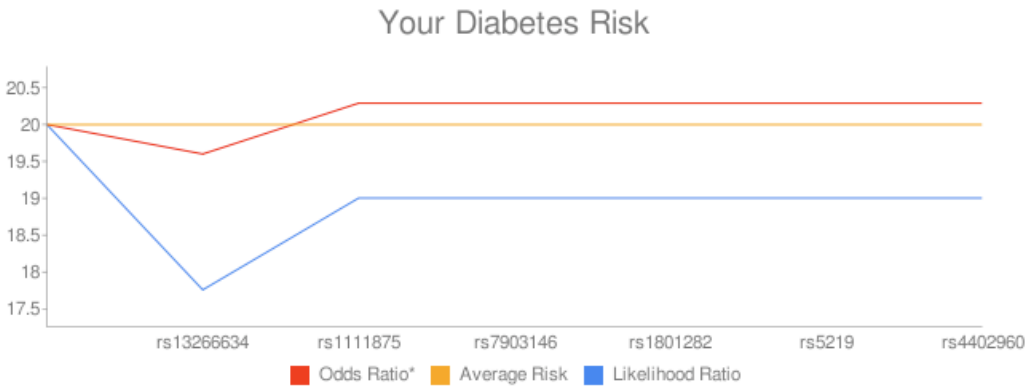


TABLE 1: PREDICTIONS FOR DISEASE RELATIVE RISKS FOR FIVE INDIVIDUALS

Disease	Female A	Female B	Female C	Male D	Male E
Breast cancer	↑↑	↑↑	↓↓		
Coeliac disease	↓↓	↓↓	↓↓	↓↓	↓↓
Colon cancer	==	==	=↓	↑↑	=↓
Crohn's disease	↓↑	↓↑	↓↓	↓↓	↓=
Heart attack	↓↓	=↓	=↓	=↓	↑↑
Lupus	↑↓	↓↓	↓↓	↑=	↑=
Macular degeneration	↓↓	↓↓	↑=	↓↓	↓↓
Multiple sclerosis	↑↑		↓↓	↓↓	↓↓
Prostate cancer				↑↑	↓↑
Psoriasis	↓↑		↑↓	↑↑	↓↓
Restless legs syndrome	=↓	↑↑	↓=	↓↑	↑↑
Rheumatoid arthritis	↑↑	↑↑	↓↓	↓↓	↑↑
Type 2 diabetes	↓↓	=↓	↓↓	↑↓	=↓

↑ increased risk (RR > 1.05), ↓ decreased risk (relative risk (RR) < 0.95), = average risk (0.95 ≤ RR ≤ 1.05). First prediction is from 23andMe; second prediction is from Navigenics. Different predictions are highlighted in beige.

Ng et al. *Nature* 2009