

Associative Transcriptomics workshop

Part Two

Andrea Harper, Zhesi He

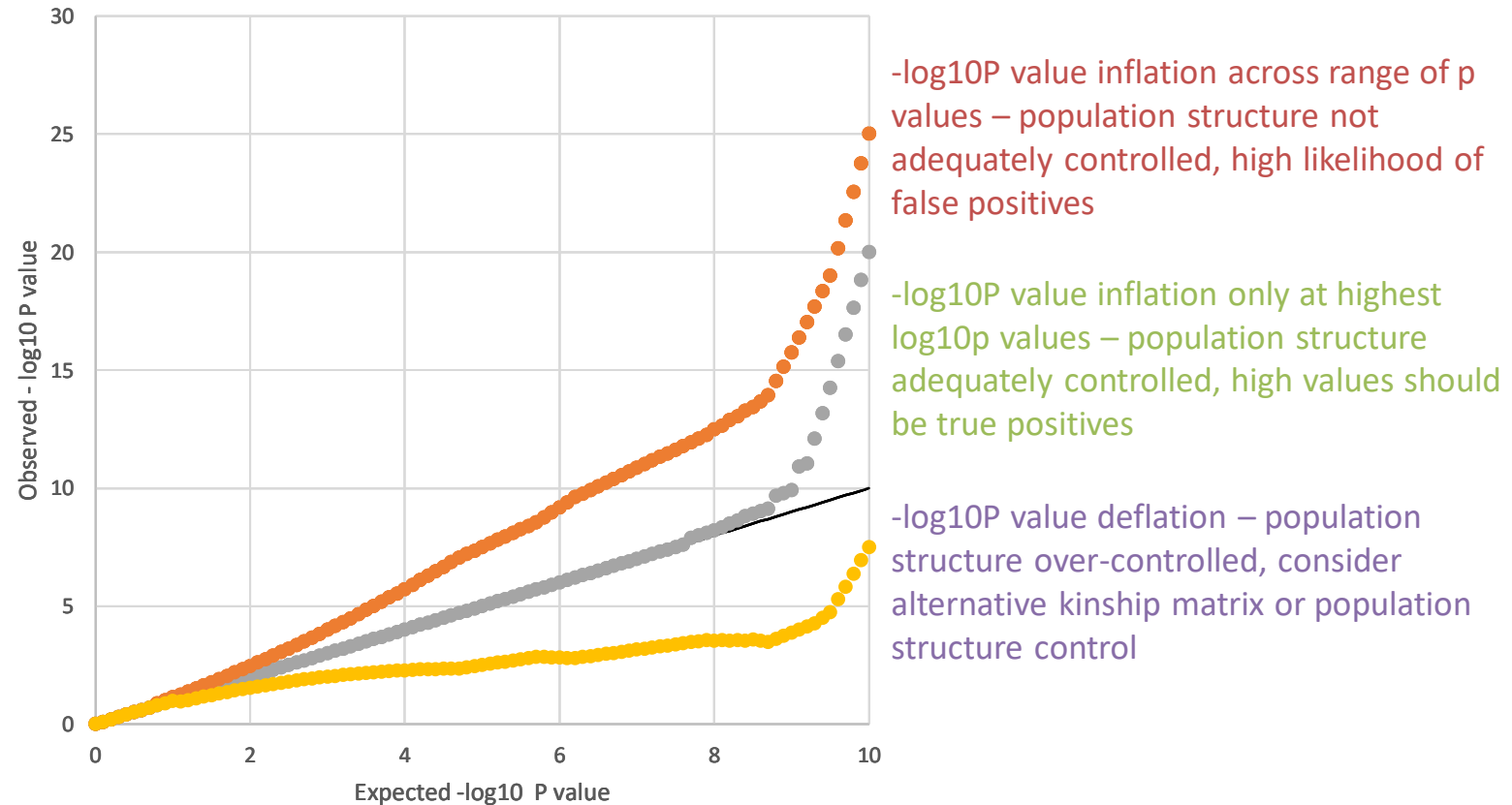
Interpreting Results

SNP and GEM results look similar

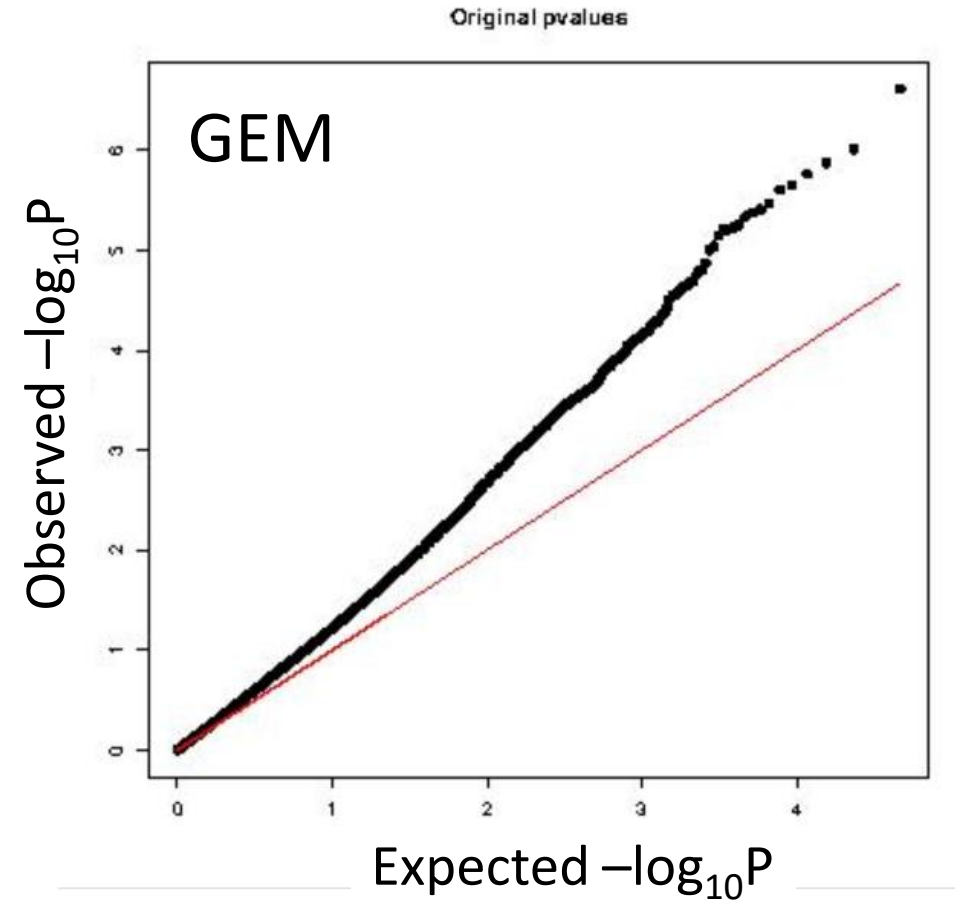
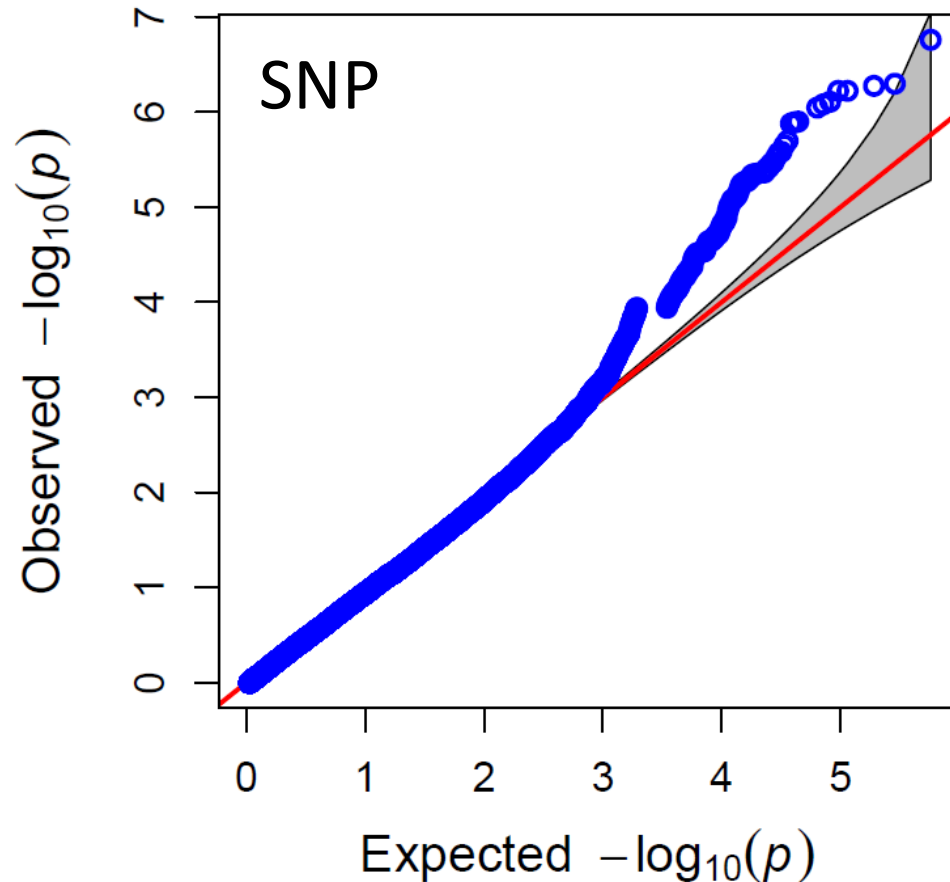
- You will find that the results outputted by GAPIT (SNPs) and Regress (GEMs) look quite similar
- Both will include results tables, QQ Plots and Manhattan Plots
- GAPIT also produces some additional files
- We will first look at the raw results tables...

QQ Plots

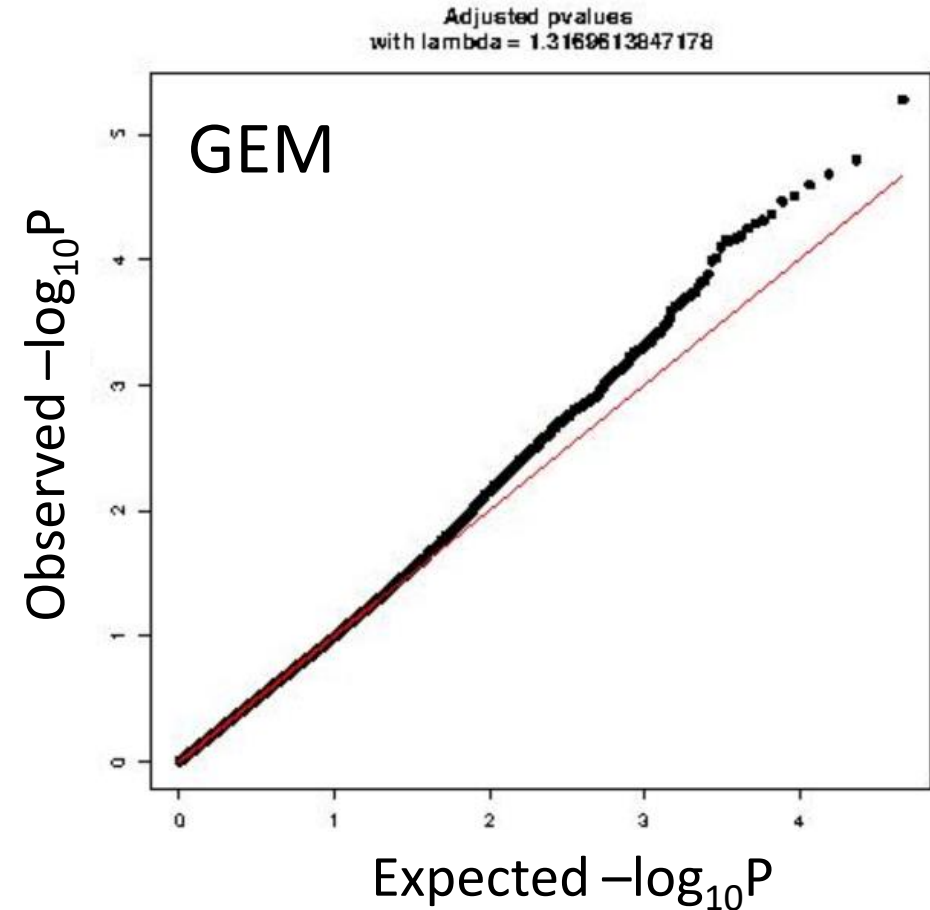
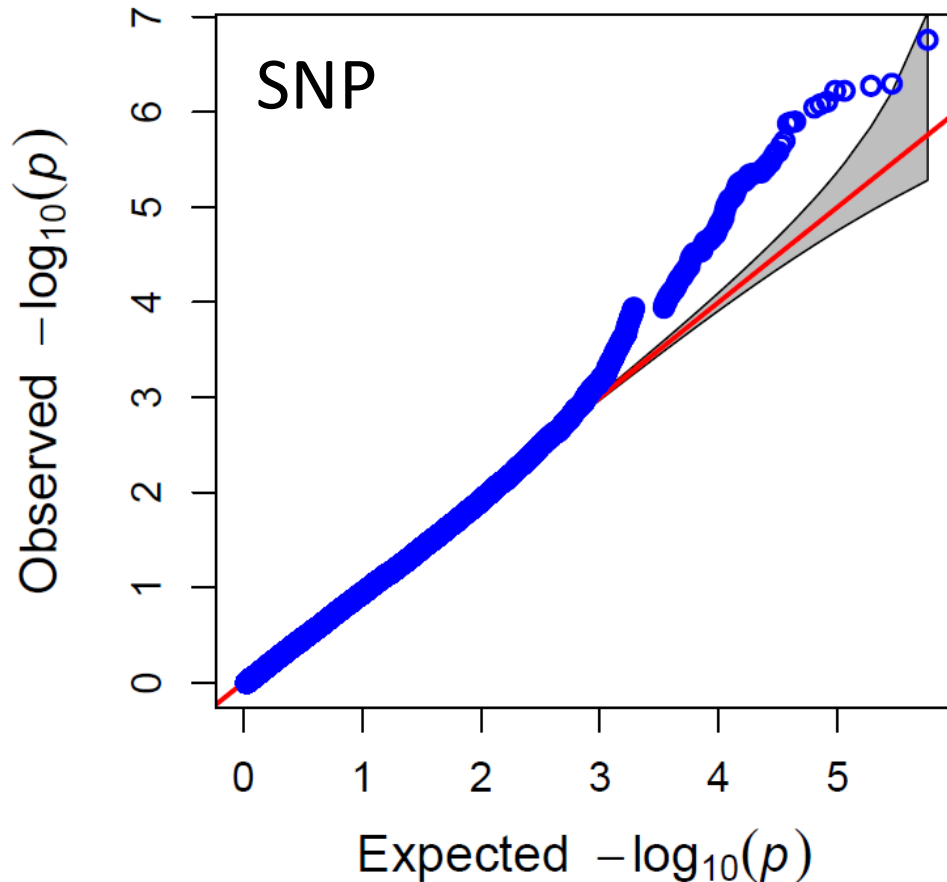
QQ-Plots are used to evaluate fit to model



More complex SNP model controls false positives more effectively than simpler GEM model

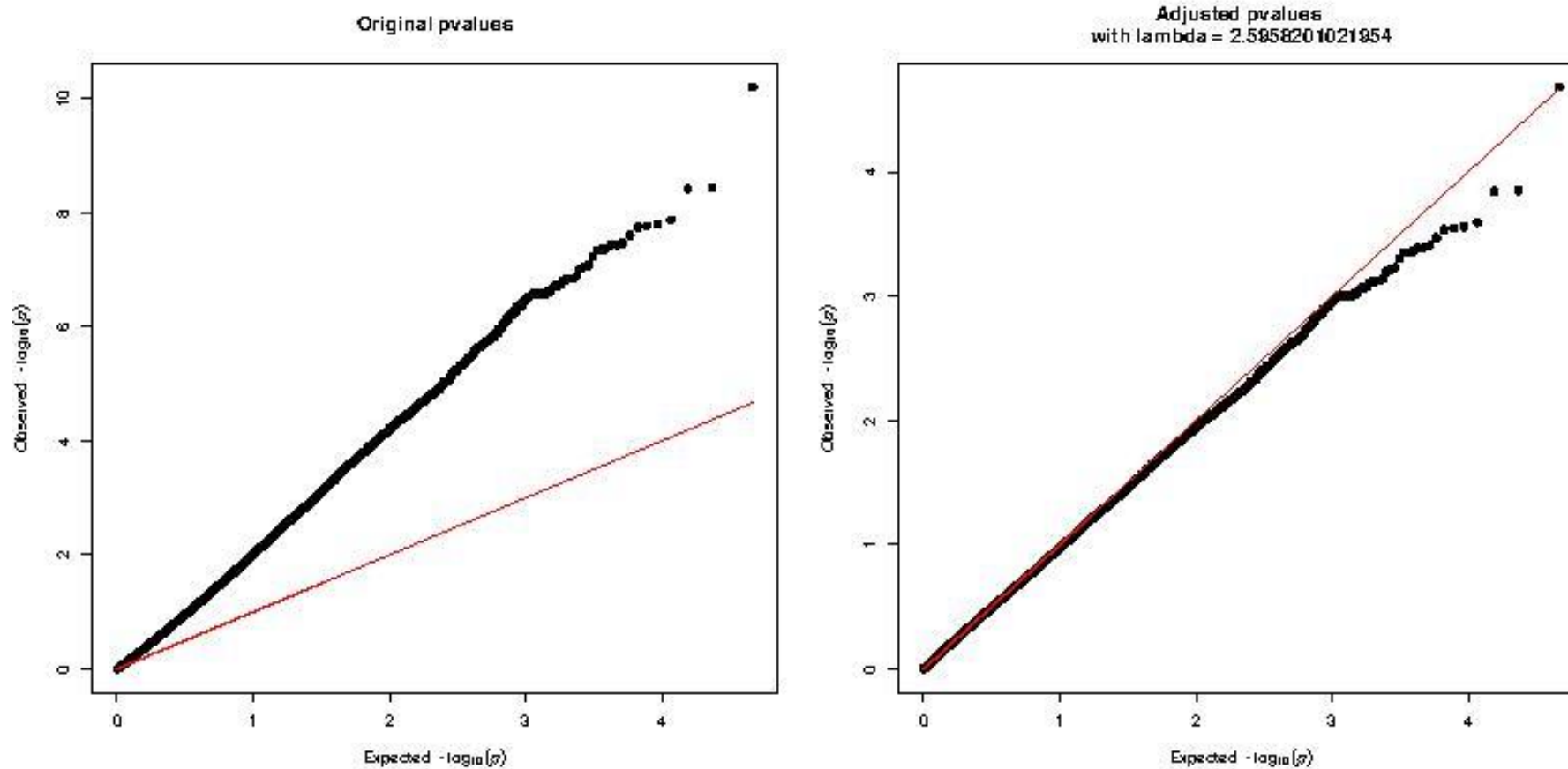


More complex SNP model controls false positives more effectively than simpler GEM model

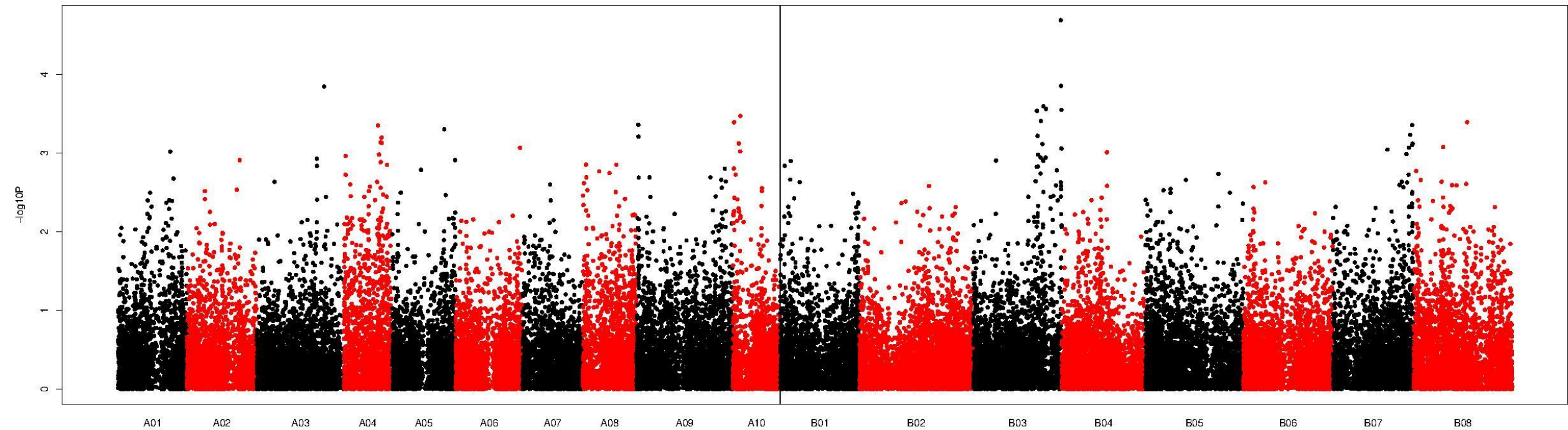


We adjust p-values for Genomic Inflation
in GEM analysis

Correction for genomic inflation is a rough tool – use alongside Manhattan plots



Correction for genomic inflation is a rough tool – use alongside Manhattan plots

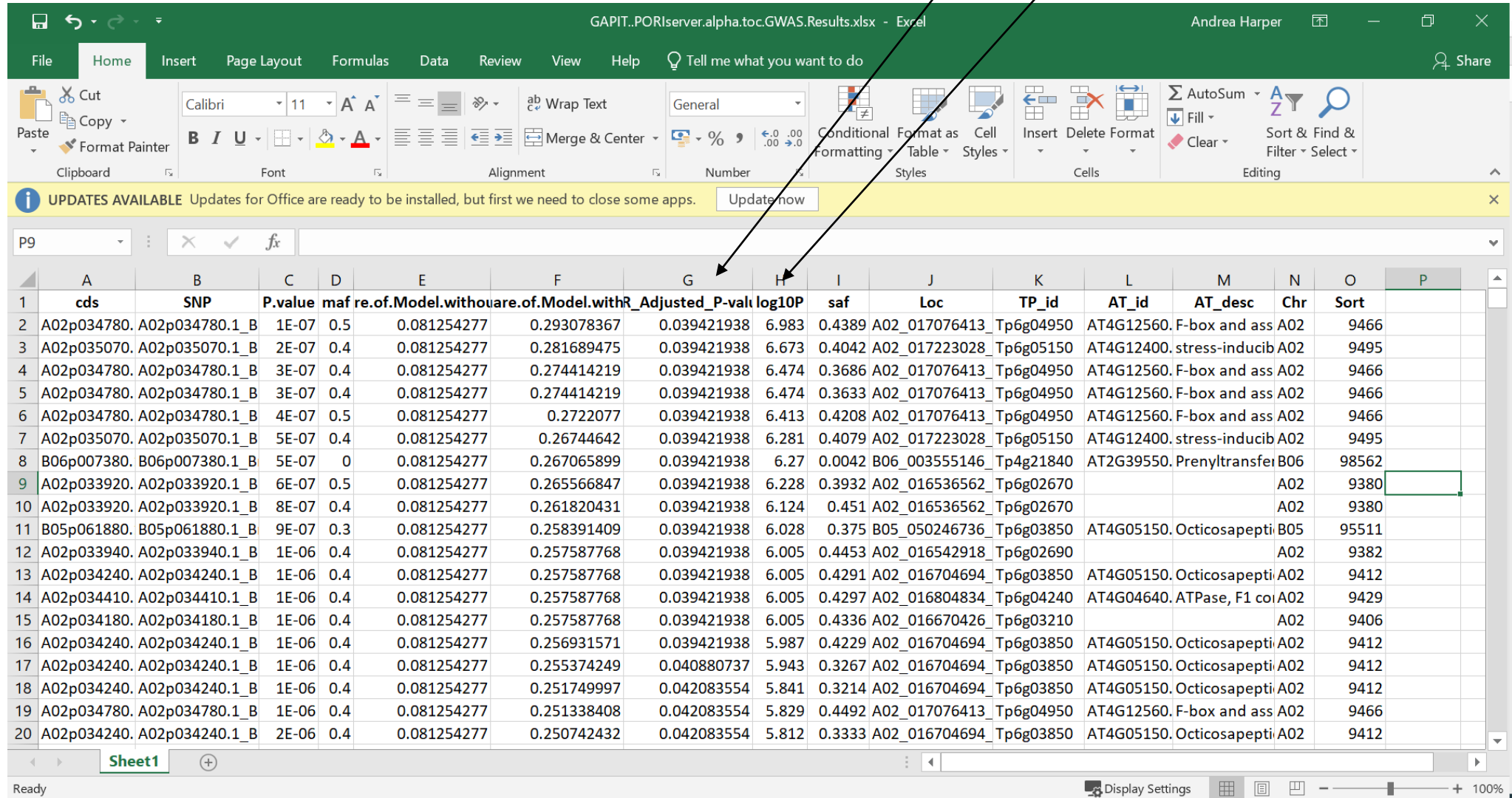


Association test results files

SNP results file

Adjusted p-values

$-\log_{10}$ adjusted p-values



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	cds	SNP	P.value	maf	re.of.Model.withouare.of.Model.withR	Adjusted_P-val	log10P	saf	Loc	TP_id	AT_id	AT_desc	Chr	Sort		
2	A02p034780.	A02p034780.1_B	1E-07	0.5	0.081254277	0.293078367	0.039421938	6.983	0.4389	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
3	A02p035070.	A02p035070.1_B	2E-07	0.4	0.081254277	0.281689475	0.039421938	6.673	0.4042	A02_017223028	Tp6g05150	AT4G12400.	stress-inducib	A02	9495	
4	A02p034780.	A02p034780.1_B	3E-07	0.4	0.081254277	0.274414219	0.039421938	6.474	0.3686	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
5	A02p034780.	A02p034780.1_B	3E-07	0.4	0.081254277	0.274414219	0.039421938	6.474	0.3633	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
6	A02p034780.	A02p034780.1_B	4E-07	0.5	0.081254277	0.2722077	0.039421938	6.413	0.4208	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
7	A02p035070.	A02p035070.1_B	5E-07	0.4	0.081254277	0.26744642	0.039421938	6.281	0.4079	A02_017223028	Tp6g05150	AT4G12400.	stress-inducib	A02	9495	
8	B06p007380.	B06p007380.1_B	5E-07	0	0.081254277	0.267065899	0.039421938	6.27	0.0042	B06_003555146	Tp4g21840	AT2G39550.	Prenyltransfer	B06	98562	
9	A02p033920.	A02p033920.1_B	6E-07	0.5	0.081254277	0.265566847	0.039421938	6.228	0.3932	A02_016536562	Tp6g02670			A02	9380	
10	A02p033920.	A02p033920.1_B	8E-07	0.4	0.081254277	0.261820431	0.039421938	6.124	0.451	A02_016536562	Tp6g02670			A02	9380	
11	B05p061880.	B05p061880.1_B	9E-07	0.3	0.081254277	0.258391409	0.039421938	6.028	0.375	B05_050246736	Tp6g03850	AT4G05150.	Octicosapepti	B05	95511	
12	A02p033940.	A02p033940.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4453	A02_016542918	Tp6g02690			A02	9382	
13	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4291	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
14	A02p034410.	A02p034410.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4297	A02_016804834	Tp6g04240	AT4G04640.	ATPase, F1 coi	A02	9429	
15	A02p034180.	A02p034180.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4336	A02_016670426	Tp6g03210			A02	9406	
16	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.256931571	0.039421938	5.987	0.4229	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
17	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.255374249	0.040880737	5.943	0.3267	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
18	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.251749997	0.042083554	5.841	0.3214	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
19	A02p034780.	A02p034780.1_B	1E-06	0.4	0.081254277	0.251338408	0.042083554	5.829	0.4492	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
20	A02p034240.	A02p034240.1_B	2E-06	0.4	0.081254277	0.250742432	0.042083554	5.812	0.3333	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	

GEM results file

Adjusted p-values

$-\log_{10}$ adjusted p-values

File

Home

Insert

Page Layout

Formulas

Data

Review

View

Help

Tell me what you want to do

</

Multiple Test Correction

- Both SNP and GEM associations include many individual statistical tests
- In general, if we perform x tests, what is the chance of seeing at least 1 false positive?

$$P(\text{making an error}) = \alpha$$

$$P(\text{not making an error}) = 1 - \alpha$$

$$P(\text{not making an error in } x \text{ tests}) = (1 - \alpha)^x$$

$$P(\text{making at least one error in } x \text{ tests}) = 1 - (1 - \alpha)^x$$

- So, if we have a significance threshold of 0.05, and we do 20 tests...

$$P = 1 - (1 - 0.05)^{20} = 0.64$$

This number rises with the number of tests, 100 tests...

$$P = 1 - (1 - 0.05)^{100} = 0.994$$

Multiple Test Correction

- So, you will have false positives, and lots of them!
- Multiple test correction must be used to reduce the chance of picking up false positives
- There are two main ways to do this:
 1. Bonferroni correction – This changes the p-value significance threshold to make it more stringent, based on the number of tests you have done
 2. False Discovery Rate Adjustment (FDR)– This adjusts each p-value resulting from your statistical tests to correct for the expected rate of false positives

1. Bonferroni correction

- This is the simplest and quickest approach, but also the most stringent
- Because of this, it can mean that all your results become non-significant!
- However, as a result, if you do have p-values that pass the Bonferroni correction, they are extremely robust results!

1. Bonferroni correction

- To calculate a Bonferroni-corrected significance threshold, simply divide your usual threshold (usually 0.05) by the number of tests you have done (ie. total number of SNPs/GEMs or rows in the results file)
- $\alpha_{\text{adj}} = 0.05/5000 = 0.00001$
- So now only p-values below this threshold will be deemed significant
- As p-values are transformed using $-\log_{10}$ for plotting, we can also transform this threshold (ie. $-\log_{10}P(\alpha_{\text{adj}}) = 5$)

2. False Discovery Rate Adjustment (FDR)

- FDR treats every statistical result individually, and is less stringent than Bonferroni
- This makes it useful when no p-values are significant under more stringent corrections ie. (Bonferroni)
- It takes into account the the number of tests, the p-value of each individual test, and their overall ranking in the total set of tests

2. False Discovery Rate Adjustment (FDR)

- FDR adjusted p-values:
 1. P-value of each gene ranked in order from the smallest to the largest.
 2. Largest p-value multiplied by the number of genes in test.
 3. The remaining p-values are multiplied by the total number of markers divided by their rank positions

Rank	Gene	P value	Correction
1	A	0.0005 ***	$0.0005 \times (6/1) = 0.003$ **
2	B	0.004 **	$0.004 \times (6/2) = 0.012$ *
3	C	0.01 **	$0.01 \times (6/3) = 0.02$ *
4	D	0.02 *	$0.02 \times (6/4) = 0.03$ *
5	E	0.045 *	$0.04 \times (6/5) = 0.054$ (NS)
6	F	0.08 (NS)	$0.08 \times 6 = 0.48$ (NS)

GEMs

GEM results file

R2 test statistic – How well the model fits the data (proportion variance)

Intercept
Gradient

Enables you to predict trait values based on RPKM

Regress_PORIsrver.alpha.toc.Results.xlsx - Excel

Andrea Harper

FileHomeInsertPage LayoutFormulasDataReviewViewHelpTell me what you want to doShare

Cut

Copy

Format Painter

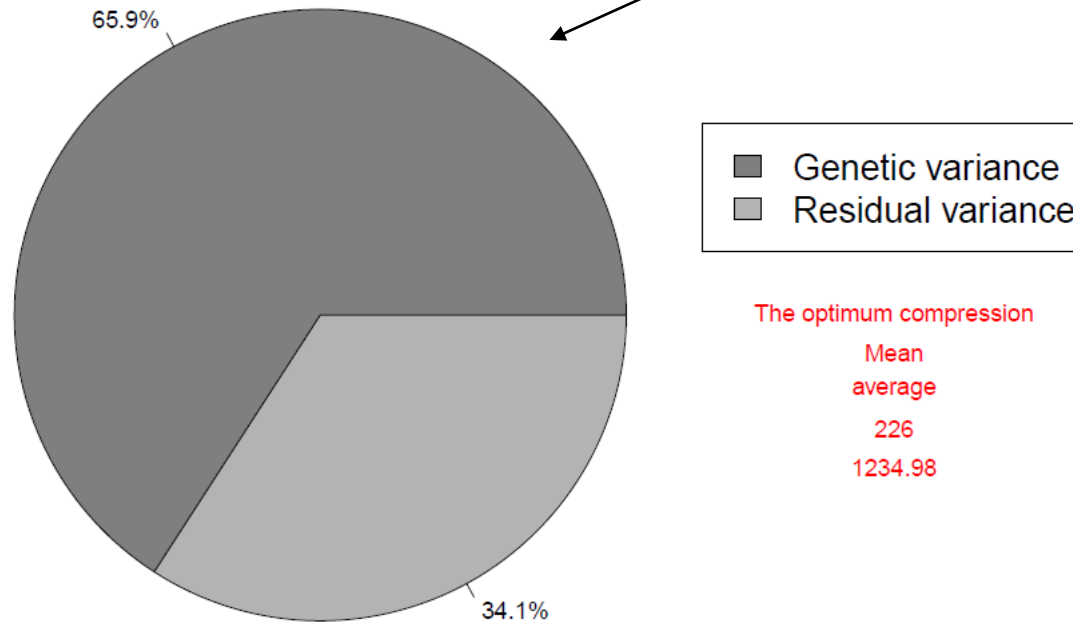
Clipboard

Calibri11

SNPs

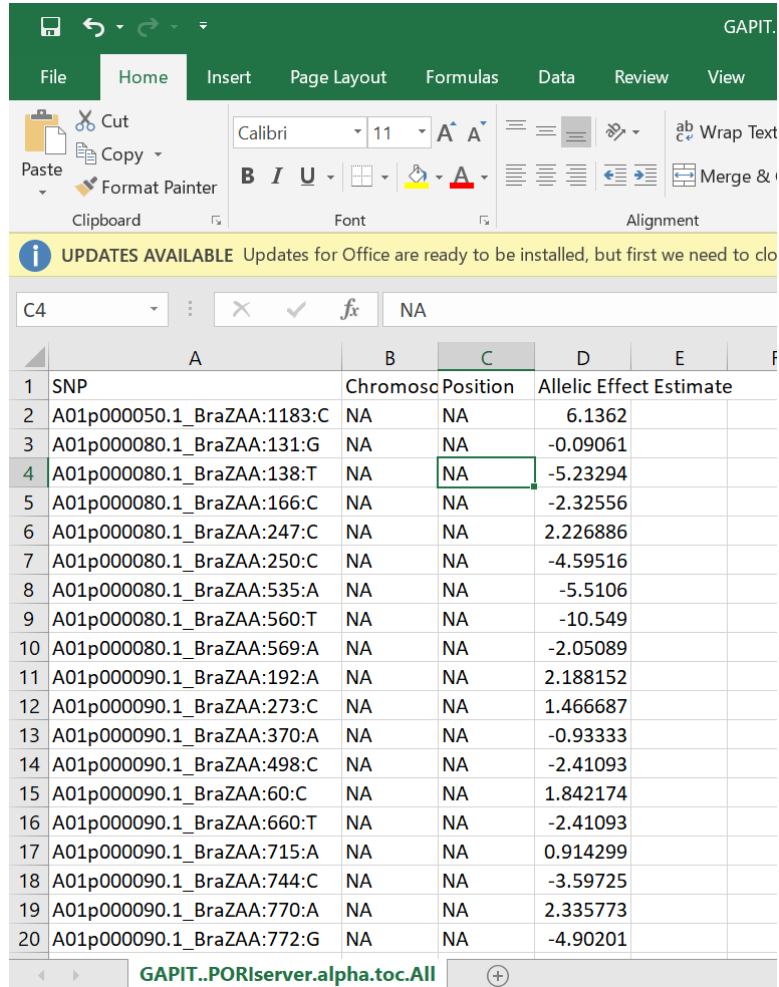
Broad sense heritability H^2

“Optimum” file



The higher this value, the more variation can be explained by SNP genotypes

SNP allele effects



1	SNP	Chromosome	Position	Allelic Effect Estimate
2	A01p000050.1_BraZAA:1183:C	NA	NA	6.1362
3	A01p000080.1_BraZAA:131:G	NA	NA	-0.09061
4	A01p000080.1_BraZAA:138:T	NA	NA	-5.23294
5	A01p000080.1_BraZAA:166:C	NA	NA	-2.32556
6	A01p000080.1_BraZAA:247:C	NA	NA	2.226886
7	A01p000080.1_BraZAA:250:C	NA	NA	-4.59516
8	A01p000080.1_BraZAA:535:A	NA	NA	-5.5106
9	A01p000080.1_BraZAA:560:T	NA	NA	-10.549
10	A01p000080.1_BraZAA:569:A	NA	NA	-2.05089
11	A01p000090.1_BraZAA:192:A	NA	NA	2.188152
12	A01p000090.1_BraZAA:273:C	NA	NA	1.466687
13	A01p000090.1_BraZAA:370:A	NA	NA	-0.93333
14	A01p000090.1_BraZAA:498:C	NA	NA	-2.41093
15	A01p000090.1_BraZAA:60:C	NA	NA	1.842174
16	A01p000090.1_BraZAA:660:T	NA	NA	-2.41093
17	A01p000090.1_BraZAA:715:A	NA	NA	0.914299
18	A01p000090.1_BraZAA:744:C	NA	NA	-3.59725
19	A01p000090.1_BraZAA:770:A	NA	NA	2.335773
20	A01p000090.1_BraZAA:772:G	NA	NA	-4.90201

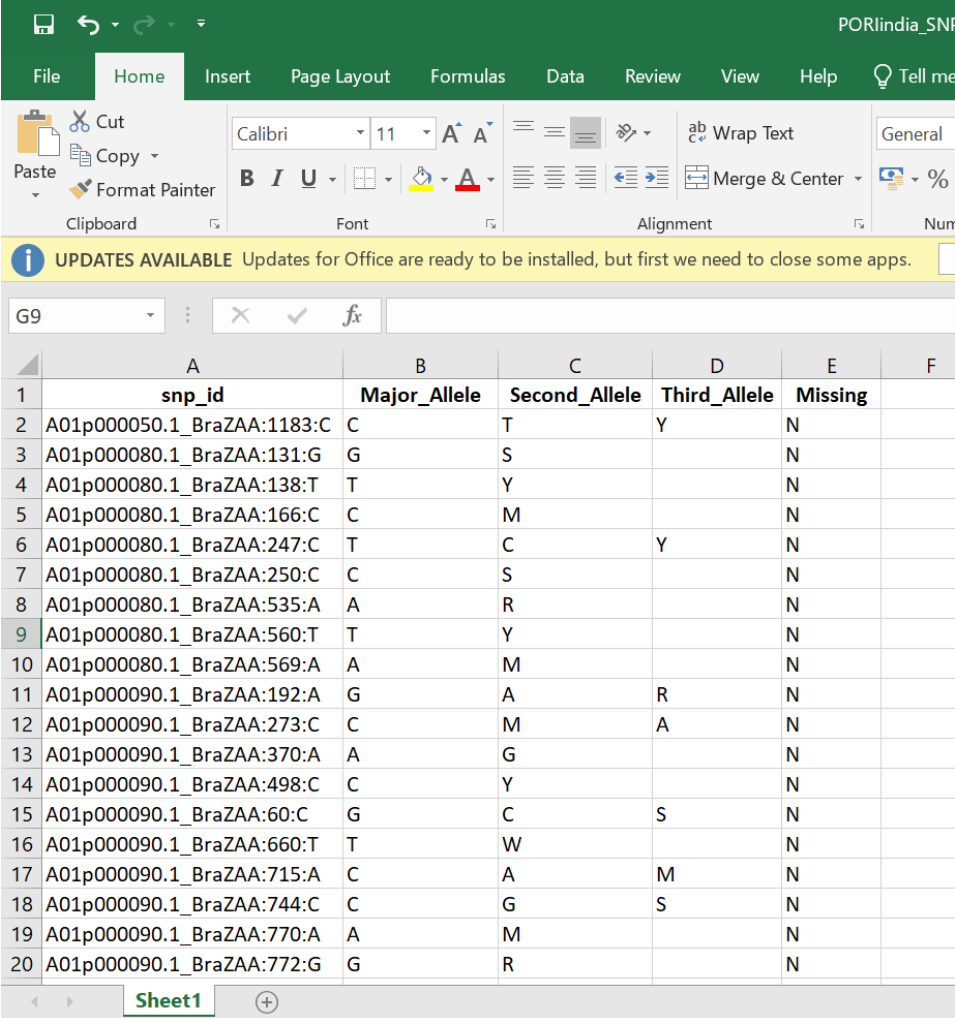
- Having identified significant SNPs, you can see how much of an effect they have on the trait in the “Allele Effect Estimates” file
- Allele effects are presented in the same units as the trait file
- Each SNP is assigned an Allele Effect with respect to the nucleotide that is second in alphabetical order. For example, if the nucleotides at a SNP are “A” and “T”, then a positive allelic effect indicates that “T” is favourable

SNP alleles

- To make finding the SNP alleles easy, we have provided a file for download called “PORlindia_SNP_alleleTable.xlsx”
- Taking the top SNP in this alpha tocopherols example:

	A	B	C	D	E	F
1	SNP	Chromosome	Position	Allelic Effect Estimate		
2	A01p000050.1_BraZAA:1183:C	NA	NA	6.1362		
3	A01p000080.1_BraZAA:131:G	NA	NA	-0.09061		

- The top SNP has C and T alleles
- The T allele is estimated to have a positive effect on the trait of 6.1 mg kg⁻¹



	A	B	C	D	E	F
1	snp_id	Major_Allele	Second_Allele	Third_Allele	Missing	
2	A01p000050.1_BraZAA:1183:C	C	T	Y	N	
3	A01p000080.1_BraZAA:131:G	G	S		N	
4	A01p000080.1_BraZAA:138:T	T	Y		N	
5	A01p000080.1_BraZAA:166:C	C	M		N	
6	A01p000080.1_BraZAA:247:C	T	C	Y	N	
7	A01p000080.1_BraZAA:250:C	C	S		N	
8	A01p000080.1_BraZAA:535:A	A	R		N	
9	A01p000080.1_BraZAA:560:T	T	Y		N	
10	A01p000080.1_BraZAA:569:A	A	M		N	
11	A01p000090.1_BraZAA:192:A	G	A	R	N	
12	A01p000090.1_BraZAA:273:C	C	M	A	N	
13	A01p000090.1_BraZAA:370:A	A	G		N	
14	A01p000090.1_BraZAA:498:C	C	Y		N	
15	A01p000090.1_BraZAA:60:C	G	C	S	N	
16	A01p000090.1_BraZAA:660:T	T	W		N	
17	A01p000090.1_BraZAA:715:A	C	A	M	N	
18	A01p000090.1_BraZAA:744:C	C	G	S	N	
19	A01p000090.1_BraZAA:770:A	A	M		N	
20	A01p000090.1_BraZAA:772:G	G	R		N	

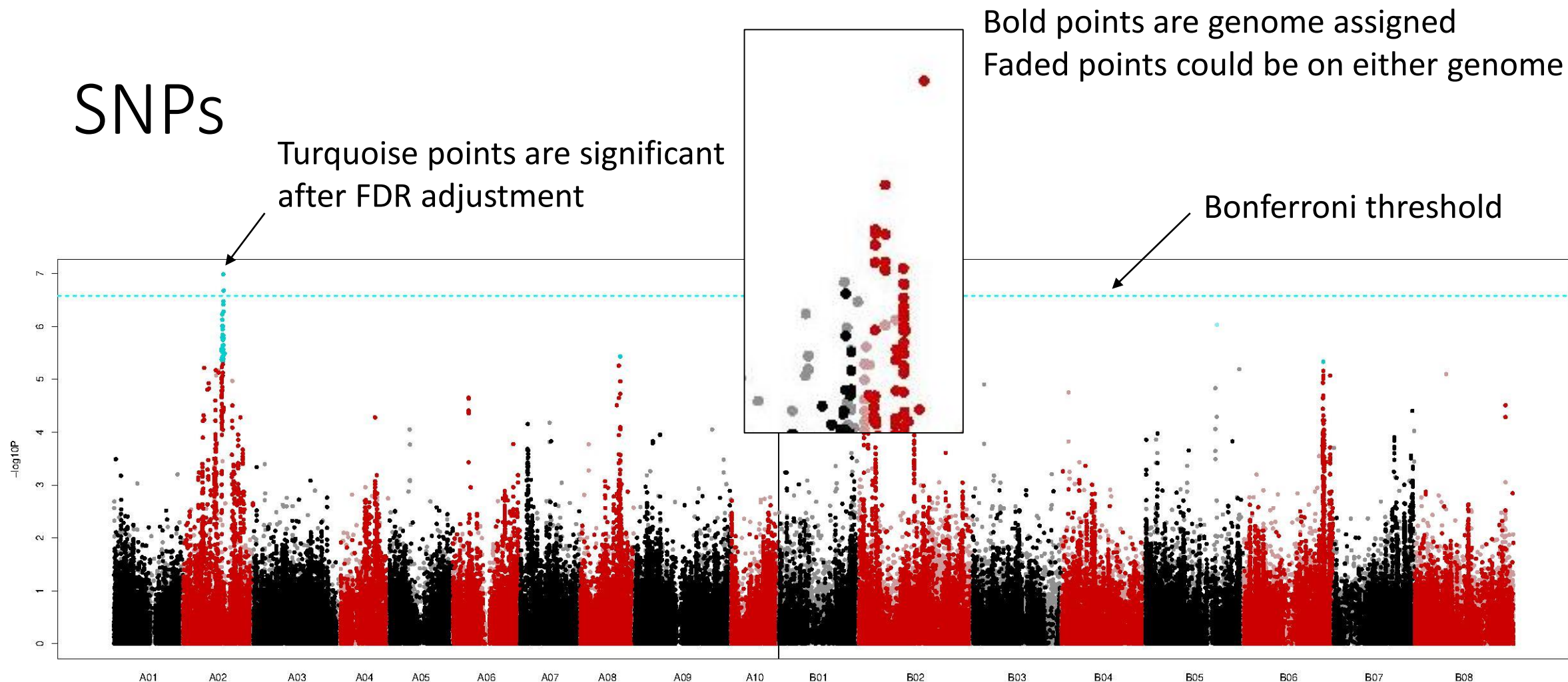
A note on allele effects

- Most complex traits are *additive*
- This means that many loci contribute to the phenotype
- Allele effects estimates are the total effect of all additive loci contributing to the trait
- So, if you select several SNPs as markers, don't expect their effect to be the sum of the estimated allele effects

Manhattan Plots

Interpreting Manhattan Plots

- Manhattan Plots may look similar for SNP and GEM results, but they should be interpreted in different ways

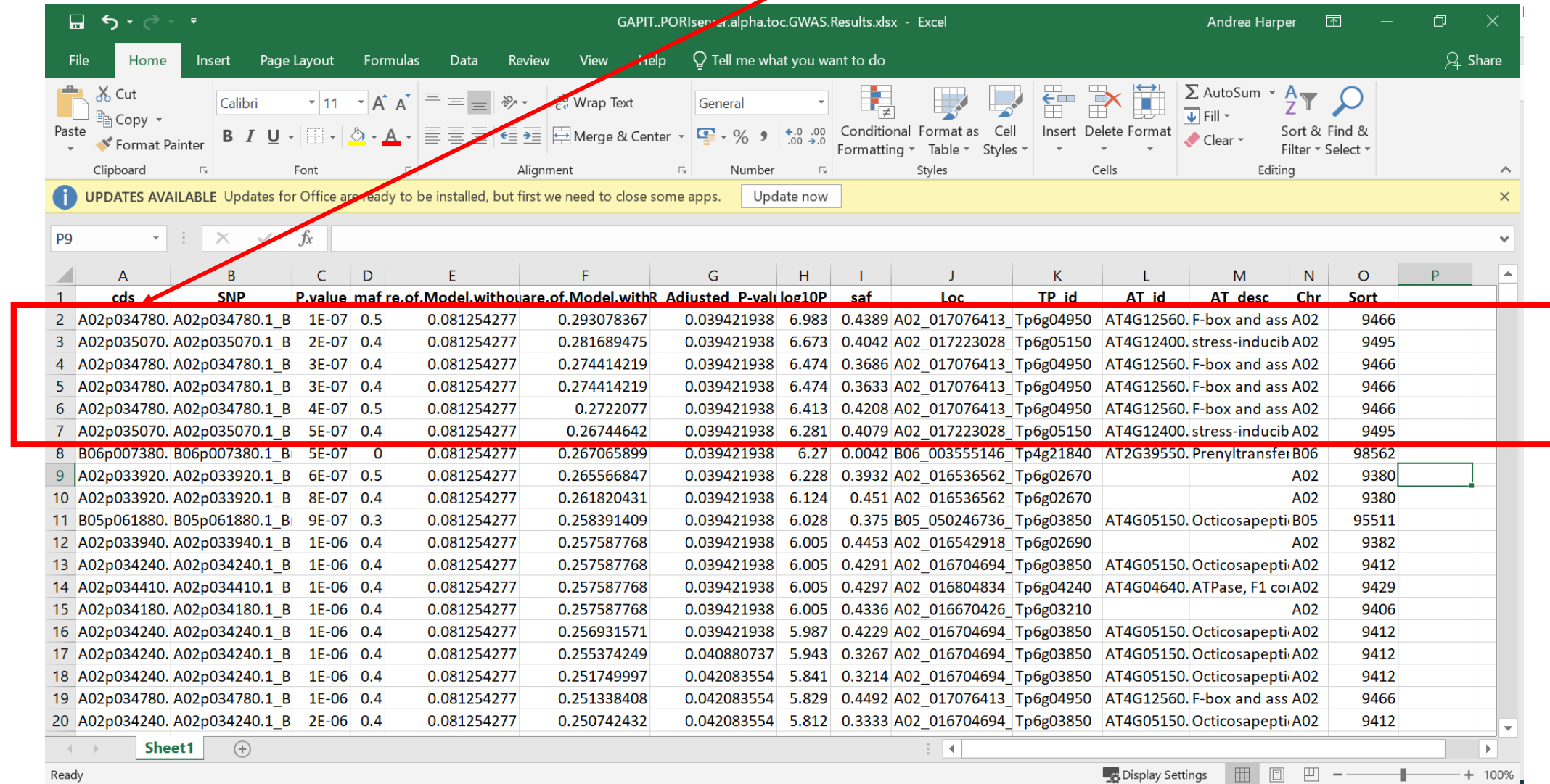


Theory suggests that we should see multiple SNPs within a block of linkage disequilibrium, so we should only consider **markers in peaks**, not single points on their own!

The candidate gene/s could be anywhere within the peak (but usually closest to the top markers)

SNPs

Peak on A2 should be closest to the top gene



Excel spreadsheet showing SNP data. The table has columns: A (cds), B (SNP), C (P.value), D (maf), E (re.of.Model.without), F (are.of.Model.with), G (R Adjusted), H (P-value), I (log10P), J (saf), K (Loc), L (TP id), M (AT id), N (AT desc), O (Chr), and P (Sort). The data is sorted by P-value (ascending). A red box highlights rows 2-7, and a red arrow points from the text 'Peak on A2 should be closest to the top gene' to the 'cds' column in row 2.

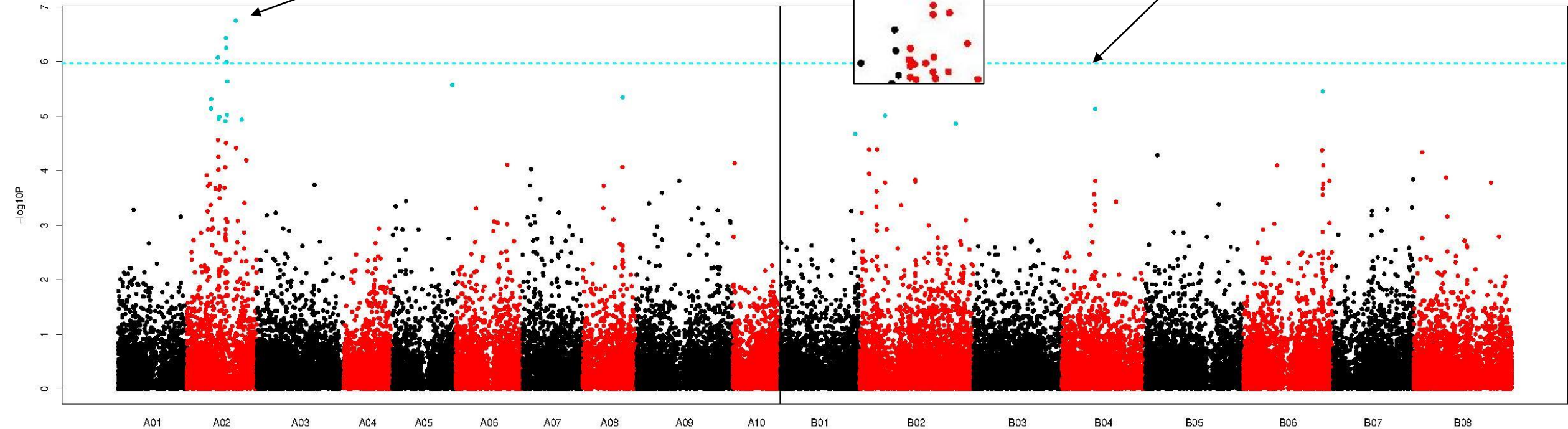
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	cds	SNP	P.value	maf	re.of.Model.without	are.of.Model.with	R Adjusted	P-value	log10P	saf	Loc	TP id	AT id	AT desc	Chr	Sort
2	A02p034780.	A02p034780.1_B	1E-07	0.5	0.081254277	0.293078367	0.039421938	6.983	0.4389	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
3	A02p035070.	A02p035070.1_B	2E-07	0.4	0.081254277	0.281689475	0.039421938	6.673	0.4042	A02_017223028	Tp6g05150	AT4G12400.	stress-inducib	A02	9495	
4	A02p034780.	A02p034780.1_B	3E-07	0.4	0.081254277	0.274414219	0.039421938	6.474	0.3686	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
5	A02p034780.	A02p034780.1_B	3E-07	0.4	0.081254277	0.274414219	0.039421938	6.474	0.3633	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
6	A02p034780.	A02p034780.1_B	4E-07	0.5	0.081254277	0.2722077	0.039421938	6.413	0.4208	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
7	A02p035070.	A02p035070.1_B	5E-07	0.4	0.081254277	0.26744642	0.039421938	6.281	0.4079	A02_017223028	Tp6g05150	AT4G12400.	stress-inducib	A02	9495	
8	B06p007380.	B06p007380.1_B	5E-07	0	0.081254277	0.267065899	0.039421938	6.27	0.0042	B06_003555146	Tp4g21840	AT2G39550.	Prenyltransfer	B06	98562	
9	A02p033920.	A02p033920.1_B	6E-07	0.5	0.081254277	0.265566847	0.039421938	6.228	0.3932	A02_016536562	Tp6g02670			A02	9380	
10	A02p033920.	A02p033920.1_B	8E-07	0.4	0.081254277	0.261820431	0.039421938	6.124	0.451	A02_016536562	Tp6g02670			A02	9380	
11	B05p061880.	B05p061880.1_B	9E-07	0.3	0.081254277	0.258391409	0.039421938	6.028	0.375	B05_050246736	Tp6g03850	AT4G05150.	Octicosapepti	B05	95511	
12	A02p033940.	A02p033940.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4453	A02_016542918	Tp6g02690			A02	9382	
13	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4291	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
14	A02p034410.	A02p034410.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4297	A02_016804834	Tp6g04240	AT4G04640.	ATPase, F1 coi	A02	9429	
15	A02p034180.	A02p034180.1_B	1E-06	0.4	0.081254277	0.257587768	0.039421938	6.005	0.4336	A02_016670426	Tp6g03210			A02	9406	
16	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.256931571	0.039421938	5.987	0.4229	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
17	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.255374249	0.040880737	5.943	0.3267	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
18	A02p034240.	A02p034240.1_B	1E-06	0.4	0.081254277	0.251749997	0.042083554	5.841	0.3214	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	
19	A02p034780.	A02p034780.1_B	1E-06	0.4	0.081254277	0.251338408	0.042083554	5.829	0.4492	A02_017076413	Tp6g04950	AT4G12560.	F-box and ass	A02	9466	
20	A02p034240.	A02p034240.1_B	2E-06	0.4	0.081254277	0.250742432	0.042083554	5.812	0.3333	A02_016704694	Tp6g03850	AT4G05150.	Octicosapepti	A02	9412	

GEMs

Turquoise points are significant after FDR adjustment

All GEMs are genome assigned

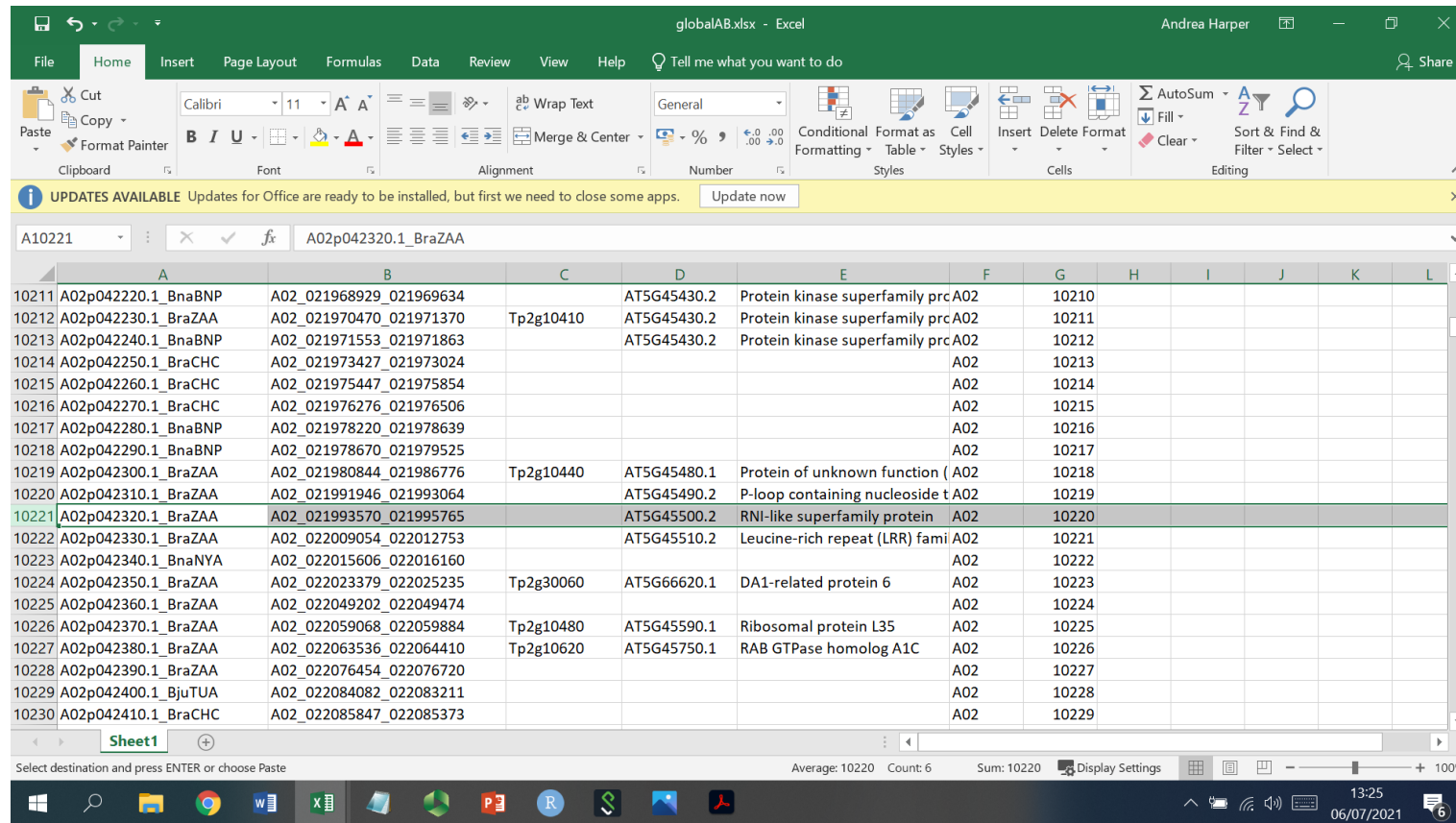
Bonferroni threshold



For GEMs, **peaks mean changes to expression of several genes**, often due to being in *cis* with a deletion, rearrangement etc. **Candidate gene could be anywhere within peak**

Individual points may also be candidate genes that are subject to *trans*-regulation

Identifying candidates in peak regions



globalAB.xlsx - Excel

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Clipboard Font Alignment Number Conditional Formatting Styles Cells Insert Delete Format AutoSum Fill Clear Sort & Find & Filter Select Editing

UPDATES AVAILABLE Updates for Office are ready to be installed, but first we need to close some apps. Update now

A10221 A02p042320.1_BraZAA

	A	B	C	D	E	F	G	H	I	J	K	L
10211	A02p042220.1_BnaBNP	A02_021968929_021969634		AT5G45430.2	Protein kinase superfamily prcA02		10210					
10212	A02p042230.1_BraZAA	A02_021970470_021971370	TP2g10410	AT5G45430.2	Protein kinase superfamily prcA02		10211					
10213	A02p042240.1_BnaBNP	A02_021971553_021971863		AT5G45430.2	Protein kinase superfamily prcA02		10212					
10214	A02p042250.1_BraCHC	A02_021973427_021973024				A02	10213					
10215	A02p042260.1_BraCHC	A02_021975447_021975854				A02	10214					
10216	A02p042270.1_BraCHC	A02_021976276_021976506				A02	10215					
10217	A02p042280.1_BnaBNP	A02_021978220_021978639				A02	10216					
10218	A02p042290.1_BnaBNP	A02_021978670_021979525				A02	10217					
10219	A02p042300.1_BraZAA	A02_021980844_021986776	TP2g10440	AT5G45480.1	Protein of unknown function (A02		10218					
10220	A02p042310.1_BraZAA	A02_021991946_021993064		AT5G45490.2	P-loop containing nucleoside tA02		10219					
10221	A02p042320.1_BraZAA	A02_021993570_021995765		AT5G45500.2	RNI-like superfamily protein A02		10220					
10222	A02p042330.1_BraZAA	A02_022009054_022012753		AT5G45510.2	Leucine-rich repeat (LRR) famiA02		10221					
10223	A02p042340.1_BnaNYA	A02_022015606_022016160				A02	10222					
10224	A02p042350.1_BraZAA	A02_022023379_022025235	TP2g30060	AT5G66620.1	DA1-related protein 6	A02	10223					
10225	A02p042360.1_BraZAA	A02_022049202_022049474				A02	10224					
10226	A02p042370.1_BraZAA	A02_022059068_022059884	TP2g10480	AT5G45590.1	Ribosomal protein L35	A02	10225					
10227	A02p042380.1_BraZAA	A02_022063536_022064410	TP2g10620	AT5G45750.1	RAB GTPase homolog A1C	A02	10226					
10228	A02p042390.1_BraZAA	A02_022076454_022076720				A02	10227					
10229	A02p042400.1_BjuTUA	A02_022084082_022083211				A02	10228					
10230	A02p042410.1_BraCHC	A02_022085847_022085373				A02	10229					

Sheet1

Average: 10220 Count: 6 Sum: 10220 Display Settings

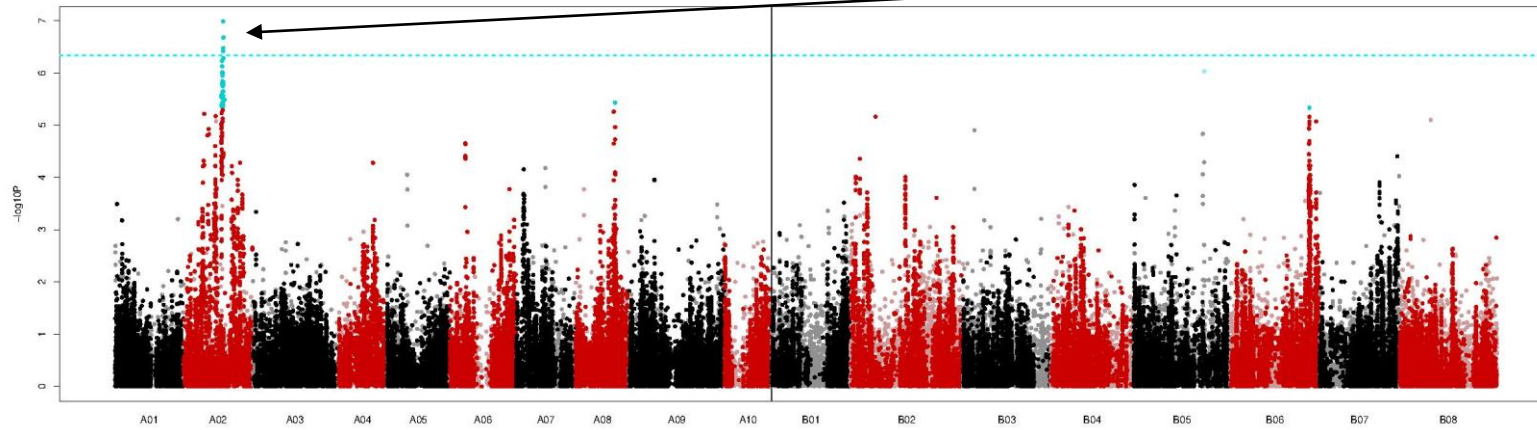
13:25 06/07/2021

GlobalAB.xlsx is available for download

Use it to look at regions surrounding significant markers

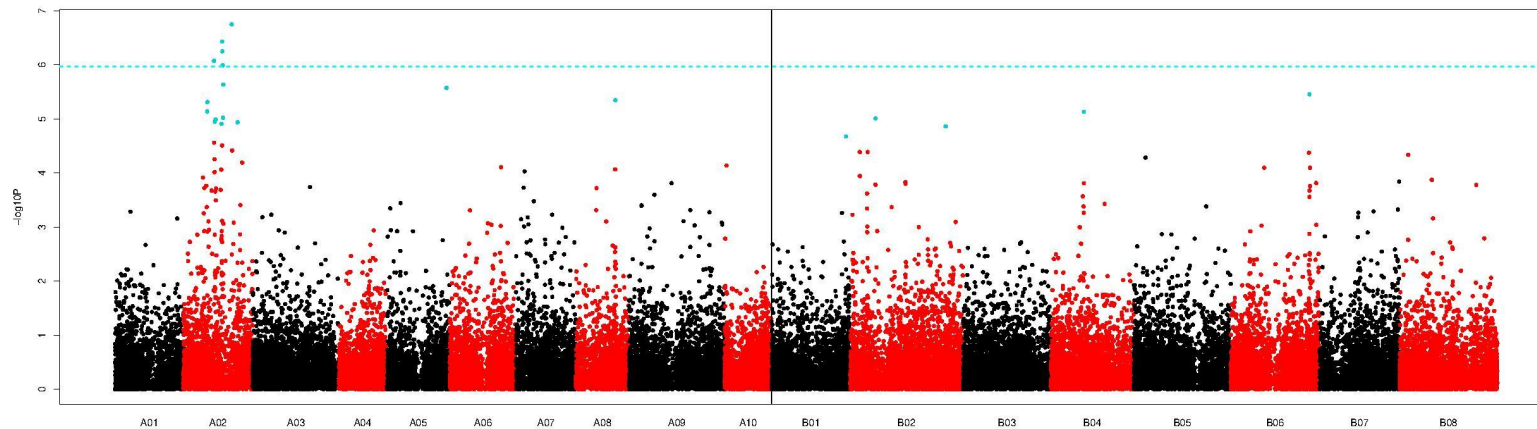
Co-localised marker associations

Looking at both SNP and GEM plots can be useful



Co-localised peaks suggest sequence variation is affecting expression of genes in *cis* with it

Width of expression peak suggests genomic extent of this effect



It is possible to have multiple genes potentially affecting the TOI in these regions