Post-processing of results

Øystein A. Haaland

Post-processing of results

- Genome-wide and candidate gene analyses often result in very much output
- Not obvious how this should be handled
- This session covers the following strategies
  - p-value adjusting
    - Per comparison error rate
    - Family-wise error rate
    - False discovery rate
  - Ploting
    - QQ-plot
    - Volcano plot
    - Manhattan plot
    - Regional plot

Multiple testing

Normally we reject H0 if $p_i < \alpha = 0.05$

What if the number of tests is very large?

Test 1: $H_0^1 \; vs. H_1^1 \Rightarrow p_1$

Test 2: $H_0^2 \; vs. H_1^2 \Rightarrow p_2$

$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$

Test N: $H_0^N \; vs. H_1^N \Rightarrow p_N$

Multiple testing

		Result of test		
		Keep H_0	Reject H_0	Sum
The truth	H_0 true	N_{00}	N_{01}	N_0
	H_0 false	N_{10}	N_{11}	N_1
	Sum	n_0	n_1	N

Number of correct results: $N_{00}+N_{11}$

Number of incorrect results: $N_{10}+N_{01}$

Multiple testing

		Result of test		
		Keep H_0	Reject H_0	Sum
The truth	H_0 true	N_{00}	N_{01}	N_0
	H_0 false	N_{10}	N_{11}	N_1
	Sum	n_0	n_1	N

Number of correct results: $N_{00}+N_{11}$

Number of incorrect results: $N_{10}+N_{01}$

We know: N, n_0, n_1

Unknown: $N_0, N_1, N_{00}, N_{01}, N_{10}, N_{11}$

Multiple testing

Rejecting H0

1. Per comparison error rate (PCER)
   - Control type I error rate (false positive rate) for a single test.
     - Type I error rate: N_{01}/N_0
   - Marginal test: Reject H_0^i if $p_i < \alpha$
   - Ignoring multiple testing
   - Too liberal when N is large (rejects H0 far too often: $N_0 \times \alpha$)
     - $N_0 = 100000 \Rightarrow N_{01} \approx N_0 \times \alpha = 100000 \times 0.05 = 5000$ false positives

		Result of test		
		Keep H_0	Reject H_0	Sum
The truth	H_0 true	N_{00}	N_{01}	N_0
	H_0 false	N_{10}	N_{11}	N_1
	Sum	n_0	n_1	N

Multiple testing

Rejecting H0

2. Familywise error rate (FWER)
   - Control overall probability of type I errors (false positives)
   - Probability of «at least one type I error» < α
     - $P(N_{01} > 0) < \alpha$

		Result of test		
		Keep H_0	Reject H_0	Sum
The truth	H_0 true	N_{00}	N_{01}	N_0
	H_0 false	N_{10}	N_{11}	N_1
	Sum	n_0	n_1	N

Multiple testing

Rejecting H0

2. Familywise error rate (FWER)
    - Control overall probability of type I errors
    - Probability of «at least one type I error» < α
        - $P(N_{01} > 0) < \alpha$
    - Bonferroni: Reject H_0^i if $p_i < \frac{\alpha}{N}$
    - Sidak: Reject H_0^i if $p_i < 1 - (1 - \alpha)^{1/N}$
    - Too conservative when N is large (keeps H0 far too often)
        - $N = 100000 \Rightarrow$ reject H0 only if $p < \frac{0.05}{100000} = 0.0000005$

		Result of test		
		Keep H_0	Reject H_0	Sum
The truth	H_0 true	N_{00}	N_{01}	N_0
	H_0 false	N_{10}	N_{11}	N_1
	Sum	n_0	n_1	N

Multiple testing

Rejecting H0

3. False discovery rate (FDR)
   - Focuses only on tests where H0 was rejected (N_{01} and N_{11})
   - Expected proportion of rejections that are false rejections
     - $E\left[\frac{N_{01}}{N_{01}+N_{11}}\right] < q$

		Result of test		
		Keep H_0	Reject H_0	Sum
The truth	H_0 true	N_{00}	N_{01}	N_0
	H_0 false	N_{10}	N_{11}	N_1
	Sum	n_0	n_1	N

Multiple testing

		Result of test		
		Keep H_0	Reject H_0	Sum
The truth	H_0 true	N_{00}	N_{01}	N_0
	H_0 false	N_{10}	N_{11}	N_1
	Sum	n_0	n_1	N

Rejecting H0

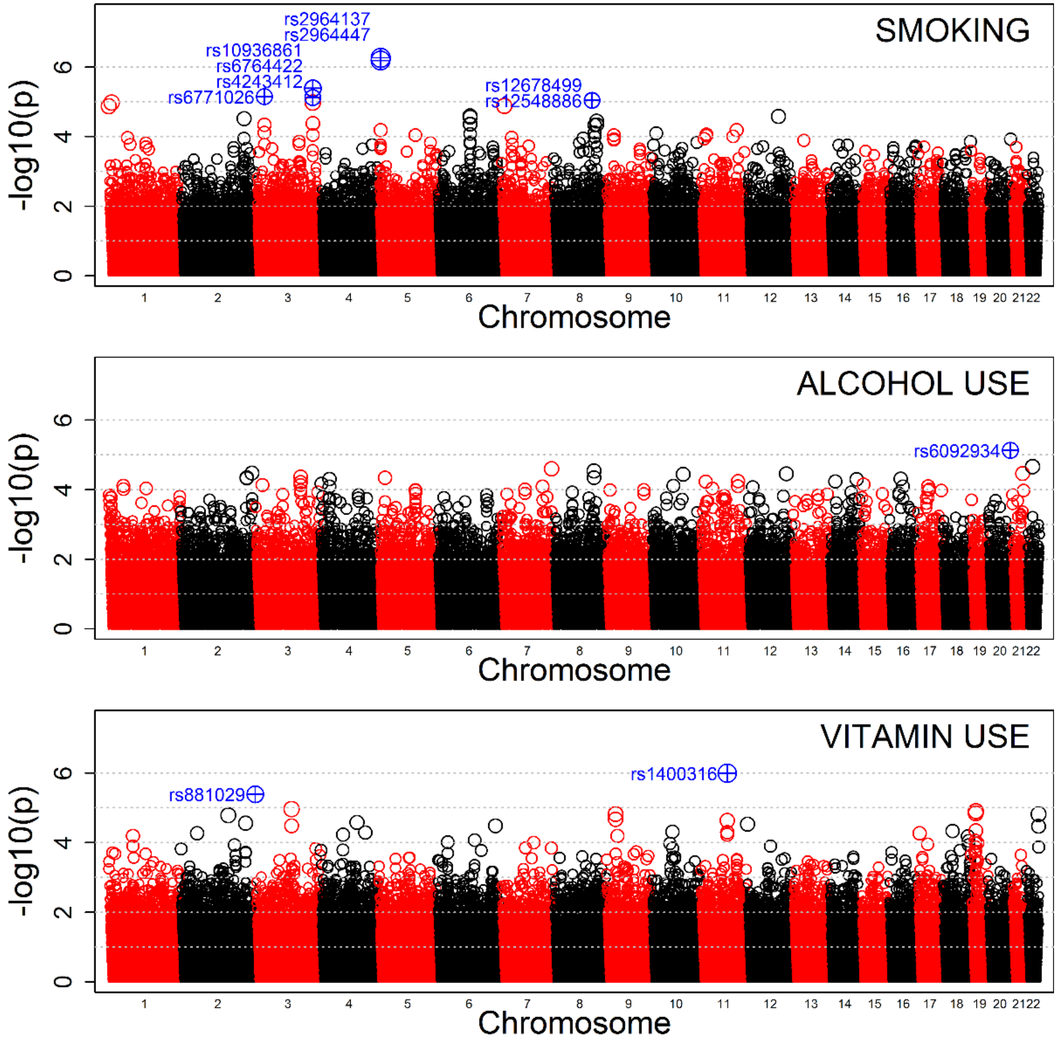
3. False discovery rate (FDR)
   - Focuses only on tests where H0 was rejected (N_{01} and N_{11})
   - Expected proportion of rejections that are false rejections
     - $E\left[\frac{N_{01}}{N_{01}+N_{11}}\right] < q$
   - q-values:
     - Transform p-values to q-values
     - Example:
       Among the tests where $q < 0.1$, we expect a proportion of 90% to be true positives.
       Among the tests where $q < 0.2$, we expect a proportion of 80% to be true positives.
     - Storey & Tibshirani (2003) Statistical significance for genomewide studies. PNAS

Back to R!

Manhattan plot

Zoom out to get an overview of where on the genome low p-values are prevalent

- X-axis: Chromosome and position on chromosome
- Y-axis: -log10(p-value)



Right:
Example from analyses looking for gene-environment effects on the risk of facial clefts.
SNPs with p-values less than 0.00001 are colored blue.

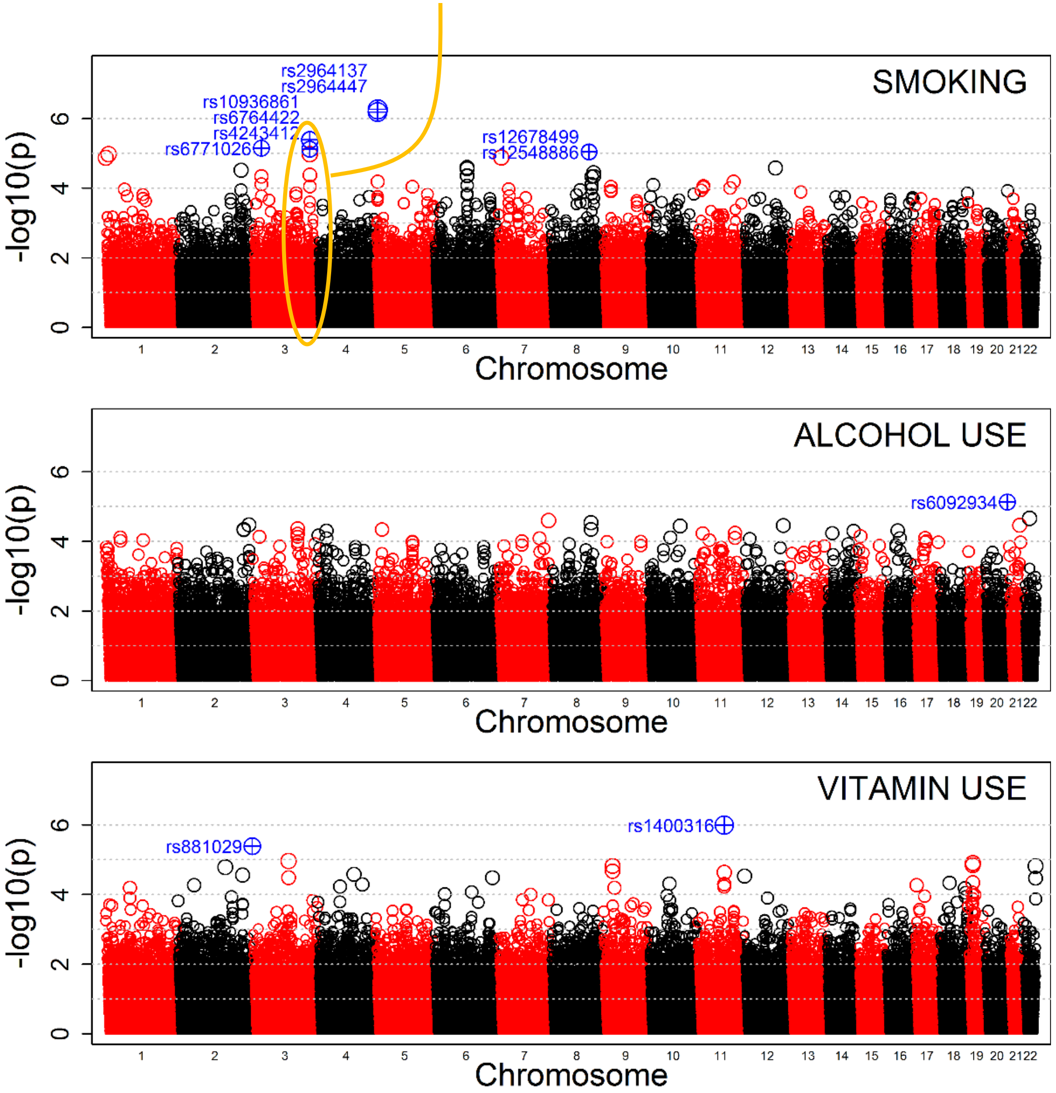
Manhattan plot

Zoom out to get an overview of where on the genome low p-values are prevalent

- X-axis: Chromosome and position on chromosome
- Y-axis: -log10(p-value)

Right:
Example from analyses looking for gene-environment effects on the risk of facial clefts.
SNPs with p-values less than 0.00001 are colored blue.

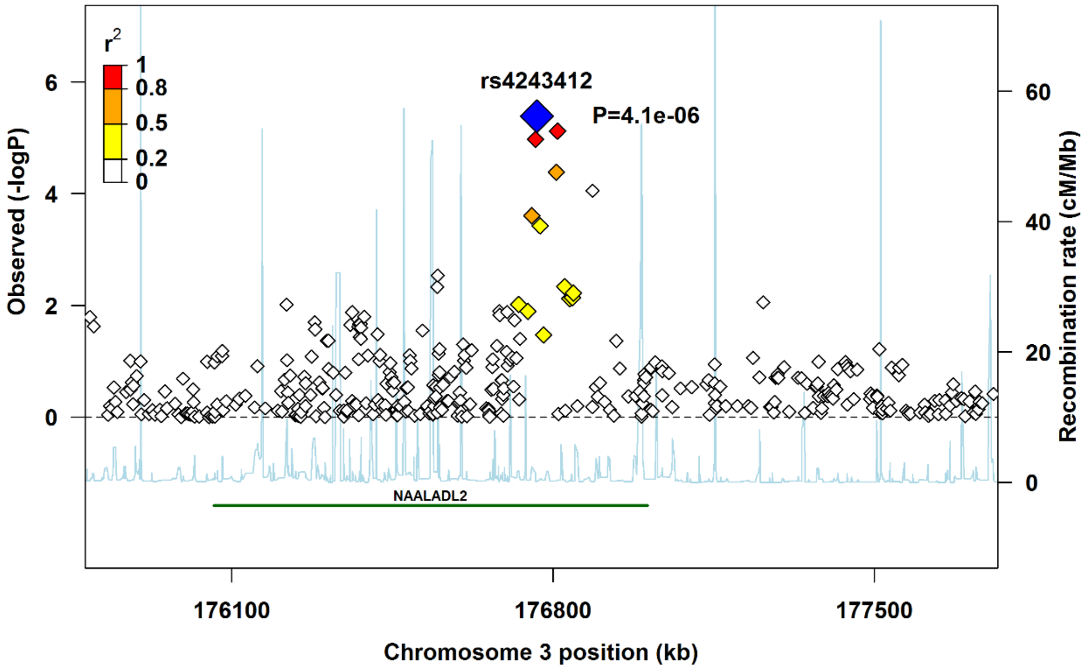
What is going on here?



Regional plot

Zoom back in to get more detail on the areas of interest

- X-axis:
  Position on chromosome
  Genes
  Recombination rate at position
- Y-axis:
  -log10(p-value)
  Recombination rate



Right:
Regional plot for rs4243412 (blue). Linkage disequilibrium with rs4243412 is indicated by colors (red, orange, yellow, white). Light blue lines indicate recombination rate.

Thank you for your attention!