THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

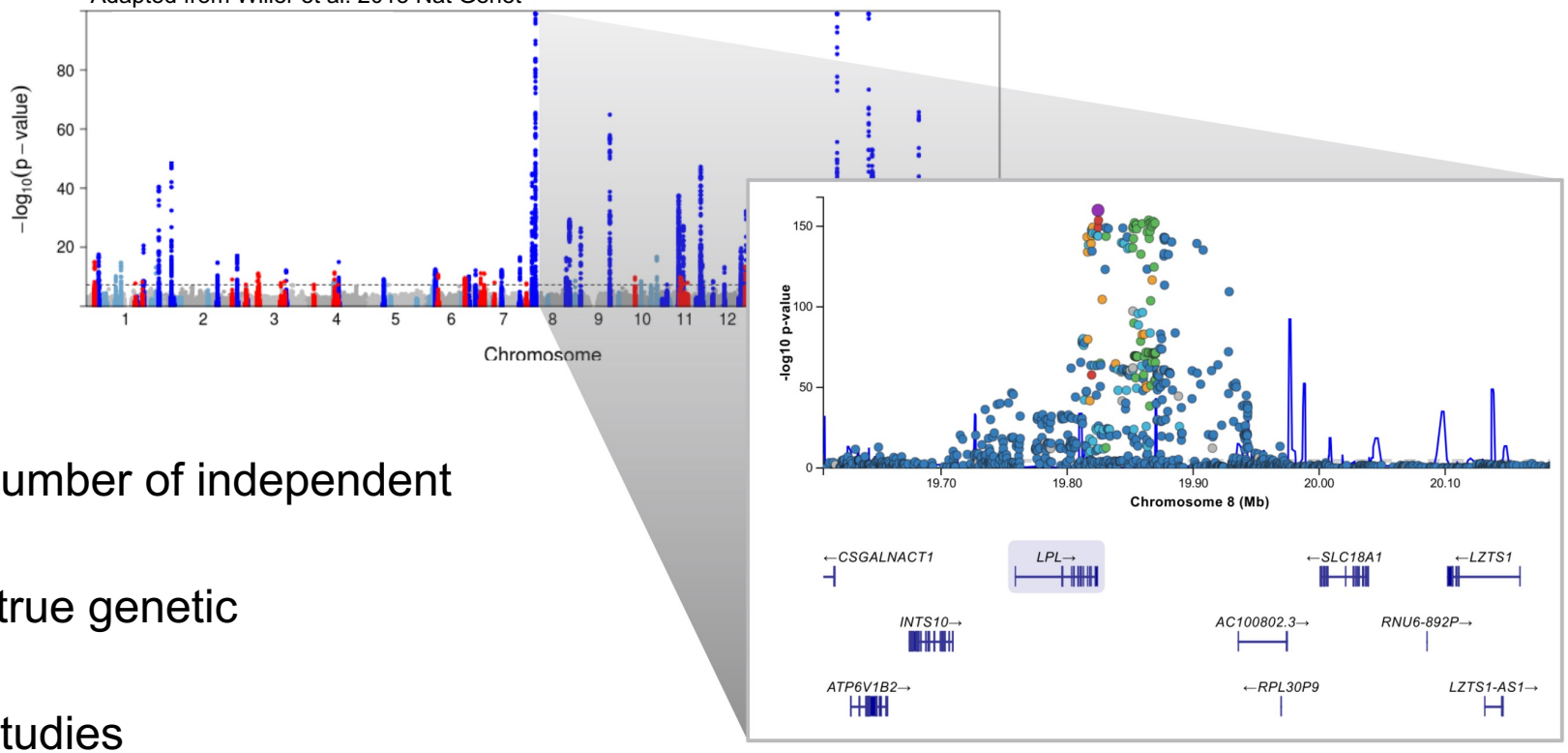
MODULE 1 | GENETIC MAPPING

Session 8. Independent GWAS associations

25 June 2024

- GWAS associations are typically clustered at specific loci
- Often, these clusters include only one causal variant, and the other variants are in LD with the causal variant

Adapted from Willer et al. 2013 Nat Genet



- Why is important to assess the number of independent associations?
  - Better understanding of the true genetic architecture of the trait
  - Inform functional follow-up studies
  - Out-of-sample prediction (module 5)
- How can we identify independent associations?
  - LD-based clumping
  - Conditional analyses

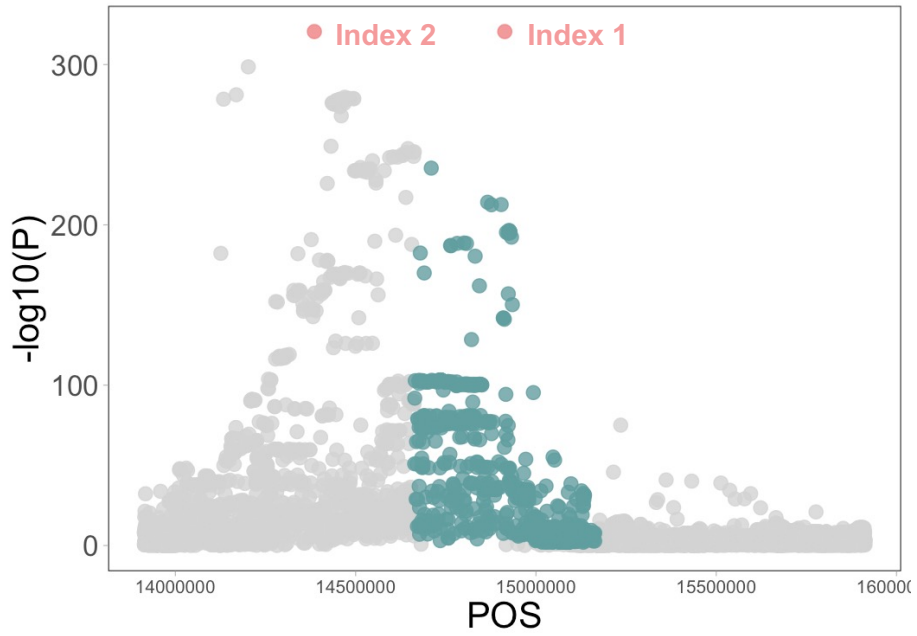
What is LD clumping?

- A method to restrict GWAS summary statistics to a subset of variants that are uncorrelated with each other.
- This method is implemented in PLINK

How does it work?

1. The SNP with lowest *P*-value is defined as an index SNP
2. All variants that are:
   - within a certain distance from the index SNP
   - in LD with the index SNP, based on a given LD r^2
   - associated at a given *P*-value threshold

   are assigned to the same cluster as the index SNP
3. Start again with another index SNP that was not in the clump above



What is LD clumping?

- A method to restrict GWAS summary statistics to a subset of variants that are uncorrelated with each other.
- This method is implemented in PLINK

How does it work?

1. The SNP with lowest *P*-value is defined as an index SNP
2. All variants that are:
   - within a certain distance from the index SNP
   - in LD with the index SNP, based on a given LD r^2
   - associated at a given *P*-value threshold

   are assigned to the same cluster as the index SNP
3. Start again with another index SNP that was not in the clump above

PLINK example:

```
plink --bfile $ld_ref \
      --clump $sumstats_file \
      --clump-p1 5e-8 \
      --clump-p2 0.01 \
      --clump-r2 0.5 \
      --clump-kb 250 \
      --out $output_file
```

Needs individual-level data from the GWAS or a reference dataset to estimate LD

What is LD clumping?

- A method to restrict GWAS summary statistics to a subset of variants that are uncorrelated with each other.
- This method is implemented in PLINK

How does it work?

1. The SNP with lowest *P*-value is defined as an index SNP
2. All variants that are:
   - within a certain distance from the index SNP
   - in LD with the index SNP, based on a given LD r^2
   - associated at a given *P*-value threshold

   are assigned to the same cluster as the index SNP
3. Start again with another index SNP that was not in the clump above

PLINK example:

```
plink --bfile $ld_ref \
      --clump $sumstats_file \
      --clump-p1 5e-8 \
      --clump-p2 0.01 \
      --clump-r2 0.5 \
      --clump-kb 250 \
      --out $output_file
```

Lets user specify:

- Maximum *P*-value of index (p1)
- Maximum P-value for clumping (p2)
- LD r^2 threshold for clumping
- Distance from index for clumping

Example output file

Each line represents one clump.

Clump index SNP

Number of SNPs clumped with index SNP

Number of clumped SNPs with *P* < 0.0001

CHR	F	SNP	BP	P	TOTAL	NSIG	S05	S01	S001	S0001	SP2
22	1	rs4821083	33056341	2.81e-36	16	0	0	0	0	16	rs762899(1),rs437300
22	1	rs6007043	45838646	6.75e-32	14	0	0	0	0	14	rs9614466(1),rs21426
22	1	rs7291090	33108150	4.31e-31	9	0	0	0	0	9	rs1475979(1),rs16991
22	1	rs7286215	28397709	3.01e-25	0	0	0	0	0	0	NONE

Chromosome, base pair and *P*-value of index SNP

List of SNPs clumped with index SNP

In some cases, we may have variants that are in close proximity and in LD with each other, but that represent independent associations. Here is an example:

							Conditional on rs12794714		Conditional on rs116970203	
Chr	SNP	Position	EA	EAF	beta	*P*-value	bC	pC	bC	pC
11	rs61883501	13882754	A	0.97	0.002	0.7492	-0.012	0.0373	0.006	0.2683
11	rs116970203	14876718	G	0.97	0.377	0.0E+00	0.415	0.0E+00	NA	NA
11	rs117576073	14912573	G	0.99	0.147	5.9E-62	0.182	1.1E-93	0.159	2.8E-72
11	rs12794714	14913575	G	0.58	0.088	0.0E+00	NA	NA	0.105	0.0E+00
11	rs61891388	66079818	T	0.54	-0.013	4.1E-10	-0.013	4.1E-10	-0.013	4.1E-10
11	rs78168201	70971149	C	0.99	-0.089	5.2E-25	-0.089	5.3E-25	-0.089	5.3E-25
11	rs1660839	71094232	G	0.75	-0.029	6.4E-37	-0.029	6.6E-37	-0.029	6.6E-37
11	rs12803256	71132868	A	0.22	-0.104	0.0E+00	-0.104	0.0E+00	-0.104	0.0E+00
11	rs12798050	71223256	C	0.17	-0.110	0.0E+00	-0.110	0.0E+00	-0.110	0.0E+00
11	rs72997623	75488054	C	0.92	-0.028	1.3E-14	-0.028	1.3E-14	-0.028	1.3E-14
11	rs1149605	76485216	T	0.83	-0.022	1.3E-16	-0.022	1.3E-16	-0.022	1.3E-16

- The highlighted SNPs would likely be clumped together because they are:
  - within 36,857 bp of each other and in high LD ($r^2 = 1$)
- But, conditional analysis shows that the effects are independent

Revez et al. 2020 Nat Commun

Stepwise conditional analysis

An alternative to clumping to identify independent associations is through stepwise conditional analysis. This is typically done with individual-level data, but GCTA implements a method (COJO) that can do these analyses based on summary-level data and an LD reference (Yang et al. 2012 Nat Genet).

How does COJO work?

1. Identify the most significant GWS SNP in the GWAS
2. Calculate the *P*-values of all other SNPs conditional on the SNP above
3. Identify the next most significant GWS SNP based on the conditional *P*-values
4. Fit all the selected SNPs jointly in a model and drop the SNP with the largest *P*-value that is greater than the cutoff *P*-value.
5. Repeat processes (2), (3) and (4) until no SNPs can be added or removed from the model

GCTA example:

```
gcta --bfile $ld_ref \
     --cojo-file $sumstats \
     --cojo-p 5e-8 \
     --cojo-slct \
     --out Height_meta
```

User can change specify maximum *P*-value of SNPs selected to condition on

Example output file

Each line is an independent locus identified with COJO

Info about genetic variant tested

Marginal effects
Each variant fitted individually in the model

Joint effects
All variants fitted jointly in the same model

Chr	SNP	bp	refA	freq	b	se	p	n	freq_geno	bJ	bJ_se	pJ	LD_r
22	rs2540653	18973549	G	0.9089	-0.0087	0.0029	0.0029	707907	0.912525	-0.0398	0.003	6.62E-41	0.0034
22	rs1155419	20787850	T	0.5695	-0.0133	0.0017	6.50E-15	700238	0.536779	-0.0265	0.0017	2.35E-53	0.0487
22	rs5754102	21916272	C	0.8226	0.0121	0.0022	4.77E-08	697485	0.824056	0.0317	0.0022	3.36E-45	-0.0342

Effect allele
Estimates are relative to this allele

GWAS summary statistics

Some examples of information gained through COJO

SNPs in modest LD

SNP	Chr	bp	Gene	refA	freq	b	p	freq_geno	bJ	pJ	LD_r
rs17720281	4	145763226	HHIP	T	0.406	0.047	1.70E-30	0.437	0.031	9.20E-13	–0.389
rs7689420	4	145787802	HHIP	T	0.163	–0.069	1.10E-41	0.164	–0.054	2.90E-23	
rs1367226	2	55943044	EFEMP1	A	0.434	–0.005	2.00E-01	0.428	–0.027	5.00E-11	–0.421
rs3791675	2	55964813	EFEMP1	T	0.234	–0.050	1.10E-28	0.249	–0.063	3.00E-37	

- When increasing alleles are positively correlated, marginal SNP effects are overestimated
- When increasing alleles are negatively correlated, marginal SNP effects are underestimated
- When increasing alleles are negatively correlated, joint analysis is more powerful

Multiple associations in a single locus

SNP	Chr	bp	Gene	refA	freq	b	p	freq_geno	bJ	pJ	LD_r
s4932429	15	87164536	*ACAN*	C	0.51	–0.025	2.70E-10	0.52	–0.022	2.30E-08	0.047
rs16942341	15	87189909	*ACAN*	T	0.031	–0.134	2.20E-24	0.027	–0.116	1.80E-18	–0.122
rs2280470	15	87196630	*ACAN*	A	0.334	0.039	1.50E-22	0.33	0.036	6.30E-19	

Yang et al. 2012 Nat Genet

Links for further information

- PLINK clump
- Yang et al. 2012 Nat Genet
- GCTA COJO