Rare variants and de novo variants association studies

2019 Dragon Star Bioinformatics Course (Day 4)

Why Study Rare Variants

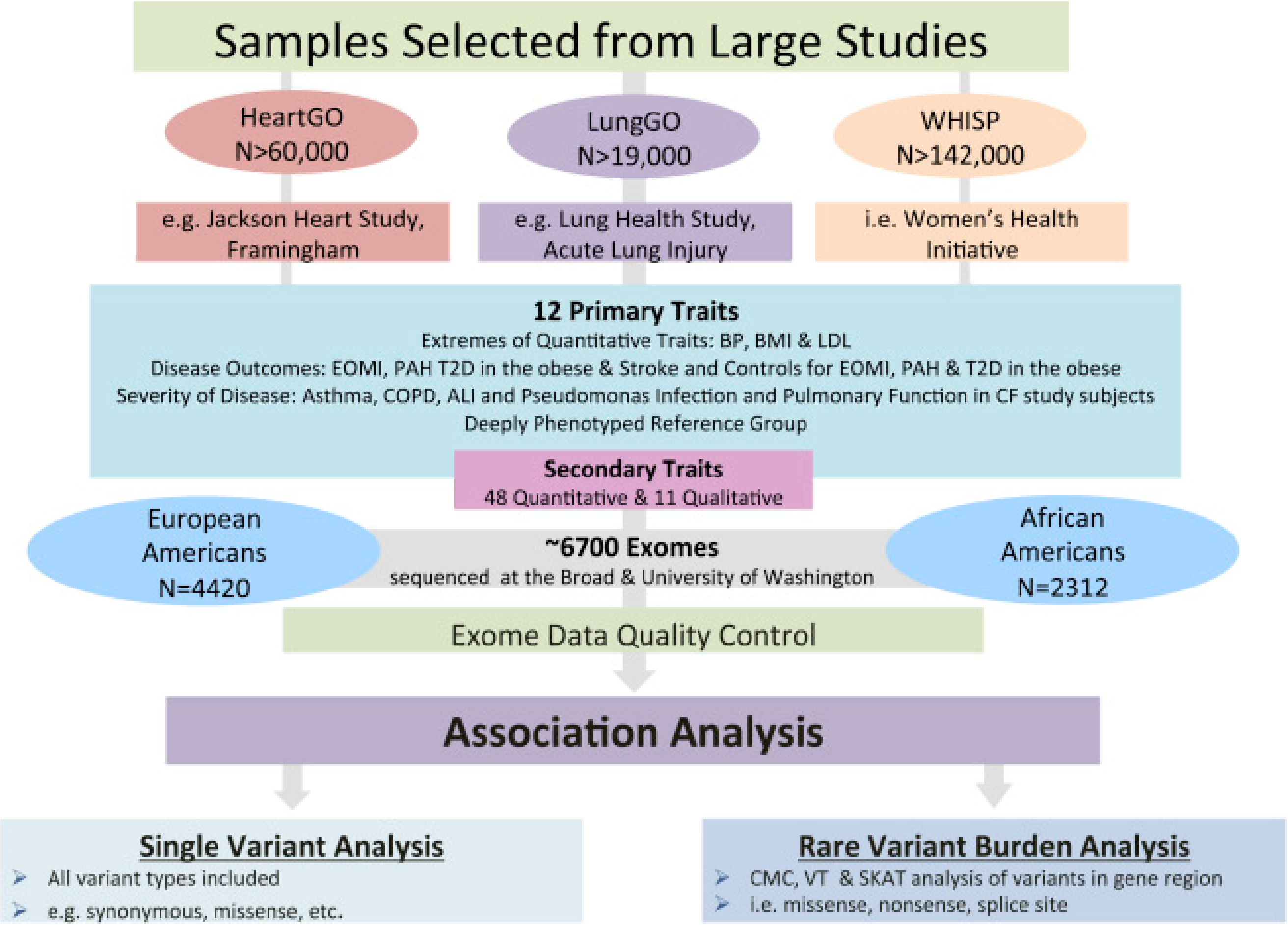
- GWAS have been extensively used to dissect the genetic architecture of complex diseases and quantitative traits.
- GWAS rely on SNP genotyping arrays or low-coverage genome sequencing, focus on common variants, typically with MAF>5% or >1%
- However, the heritability that can be explained by these GWAS findings is generally low
    - Type 2 diabetes: >70 loci identified from GWAS >150,000 individuals only explain ~11% of T2D heritability
    - Crohn’s disease: >70 loci identified from GWAS in >210,000 individuals only explain ~23% of heritability

Why Study Rare Variants

- In general, GWAS loci have modest effects on disease risk or quantitative trait variation

- Possible explanations for “missing heritability”
  - De novo, rare and low frequency (MAF<1%) variants may explain additional disease risk or trait variability
  - Structural variants
  - Epigenetic changes

NHLBI exome sequencing project (ESP)



Samples Selected from Large Studies

HeartGO
N>60,000

LungGO
N>19,000

WHISP
N>142,000

e.g. Jackson Heart Study, Framingham

e.g. Lung Health Study, Acute Lung Injury

i.e. Women's Health Initiative

12 Primary Traits

Extremes of Quantitative Traits: BP, BMI & LDL

Disease Outcomes: EOMI, PAH T2D in the obese & Stroke and Controls for EOMI, PAH & T2D in the obese

Severity of Disease: Asthma, COPD, ALI and Pseudomonas Infection and Pulmonary Function in CF study subjects

Deeply Phenotyped Reference Group

Secondary Traits

48 Quantitative & 11 Qualitative

European Americans
N=4420

~6700 Exomes

sequenced at the Broad & University of Washington

African Americans
N=2312

Exome Data Quality Control

Association Analysis

Single Variant Analysis

- All variant types included
- e.g. synonymous, missense, etc.

Rare Variant Burden Analysis

- CMC, VT & SKAT analysis of variants in gene region
- i.e. missense, nonsense, splice site

Auer et al, Am J Hum Genet, 2016

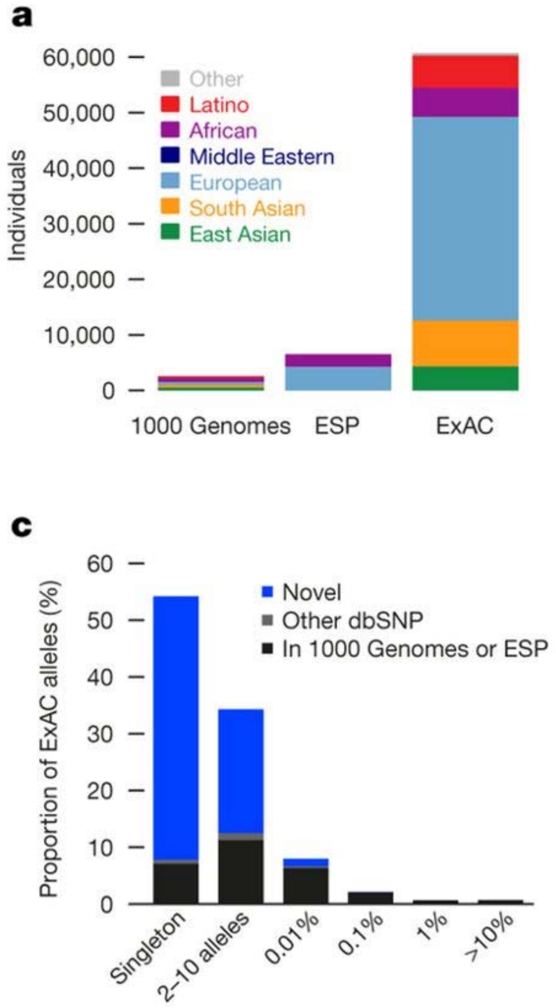
NHLBI exome sequencing project (ESP)

SET	# SNPs	Singletons	Doubletons	Tripletons	>3 Occurrences
Synonymous	270,263	128,319 (47%)	29,340 (11%)	13,129 (5%)	99,475 (37%)
Nonsynonymous	410,956	234,633 (57%)	46,740 (11%)	19,274 (5%)	110,309 (27%)
Nonsense	8,913	6,196 (70%)	926 (10%)	326 (4%)	1,465 (16%)
Non-Syn / Syn Ratio		1.8 to 1	1.6 to 1	1.4 to 1	1.1 to 1

There is a very large reservoir of extremely rare, likely functional, coding variants.

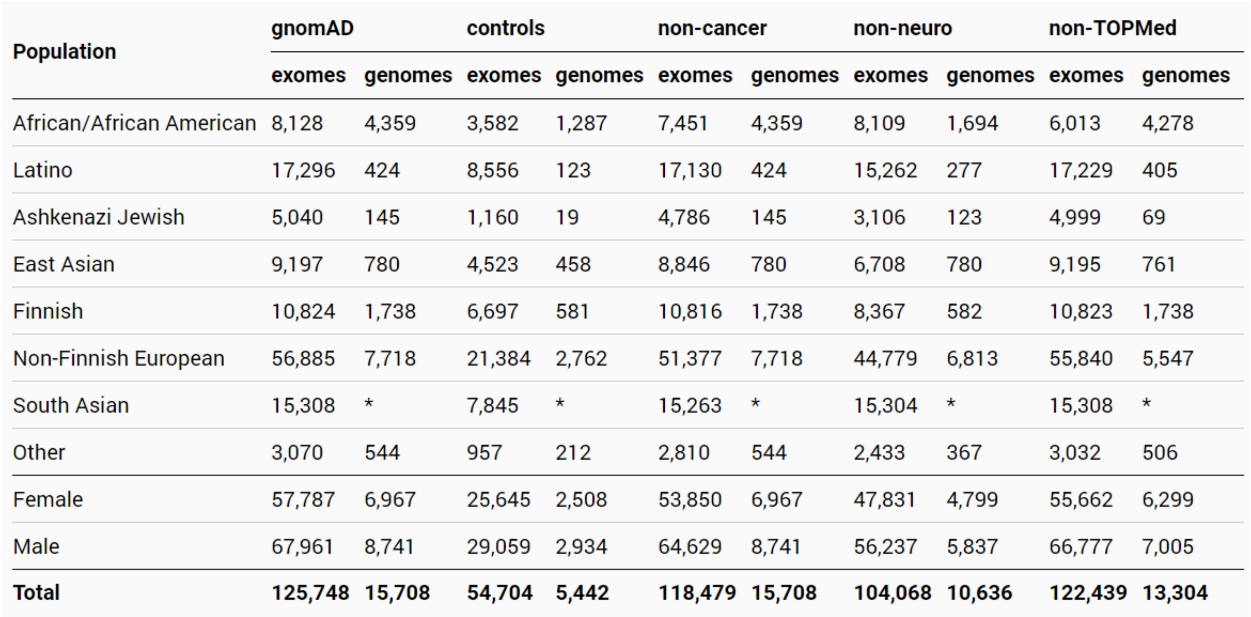
Exome Aggregation Consortium (ExAC)

- ExAC includes DNA sequence data for 60,706 individuals of diverse ancestries
- These data can be used for the efficient filtering of candidate disease-causing variants, and for the discovery of human 'knockout' variants in protein-coding genes.

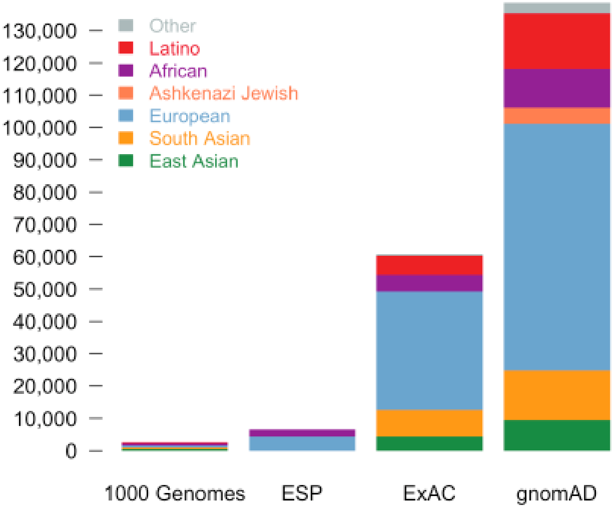


Lek et al, Nature, 2016

gnomAD is a public database of rare variants from genome/exome



Population	gnomAD		controls		non-cancer		non-neuro		non-TOPMed	
	exomes	genomes	exomes	genomes	exomes	genomes	exomes	genomes	exomes	genomes
African/African American	8,128	4,359	3,582	1,287	7,451	4,359	8,109	1,694	6,013	4,278
Latino	17,296	424	8,556	123	17,130	424	15,262	277	17,229	405
Ashkenazi Jewish	5,040	145	1,160	19	4,786	145	3,106	123	4,999	69
East Asian	9,197	780	4,523	458	8,846	780	6,708	780	9,195	761
Finnish	10,824	1,738	6,697	581	10,816	1,738	8,367	582	10,823	1,738
Non-Finnish European	56,885	7,718	21,384	2,762	51,377	7,718	44,779	6,813	55,840	5,547
South Asian	15,308	*	7,845	*	15,263	*	15,304	*	15,308	*
Other	3,070	544	957	212	2,810	544	2,433	367	3,032	506
Female	57,787	6,967	25,645	2,508	53,850	6,967	47,831	4,799	55,662	6,299
Male	67,961	8,741	29,059	2,934	64,629	8,741	56,237	5,837	66,777	7,005
Total	**125,748**	**15,708**	**54,704**	**5,442**	**118,479**	**15,708**	**104,068**	**10,636**	**122,439**	**13,304**



Karczewski, BioRxiv, 2019

Three Levels of Rare Variant Data

Level 1: Individual-level

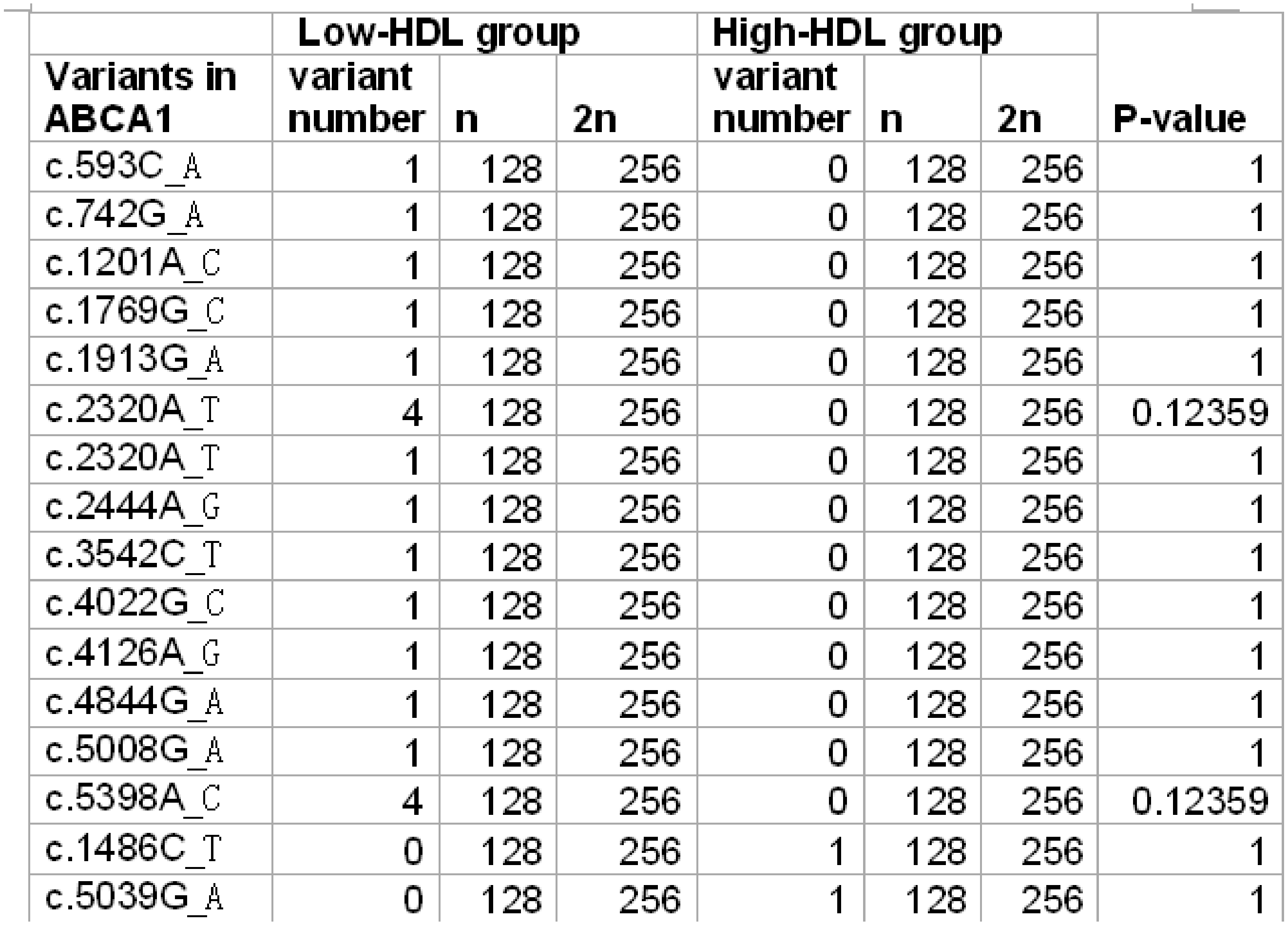
Level 2: Summarized over subjects

Level 3: Summarized over both subjects and variants

Level 1: Individual-level

Subject	V1	V2	V3	V4	Trait-1	Trait-2
1	1	0	0	0	90.1	1
2	0	1	0	.	99.2	1
3	0	0	0	0	105.9	0
4	0	0	0	0	89.5	0
5	0	.	0	0	97.6	0
6	0	0	0	0	110.5	0
7	0	0	1	0	88.8	0
8	0	0	0	1	95.4	1

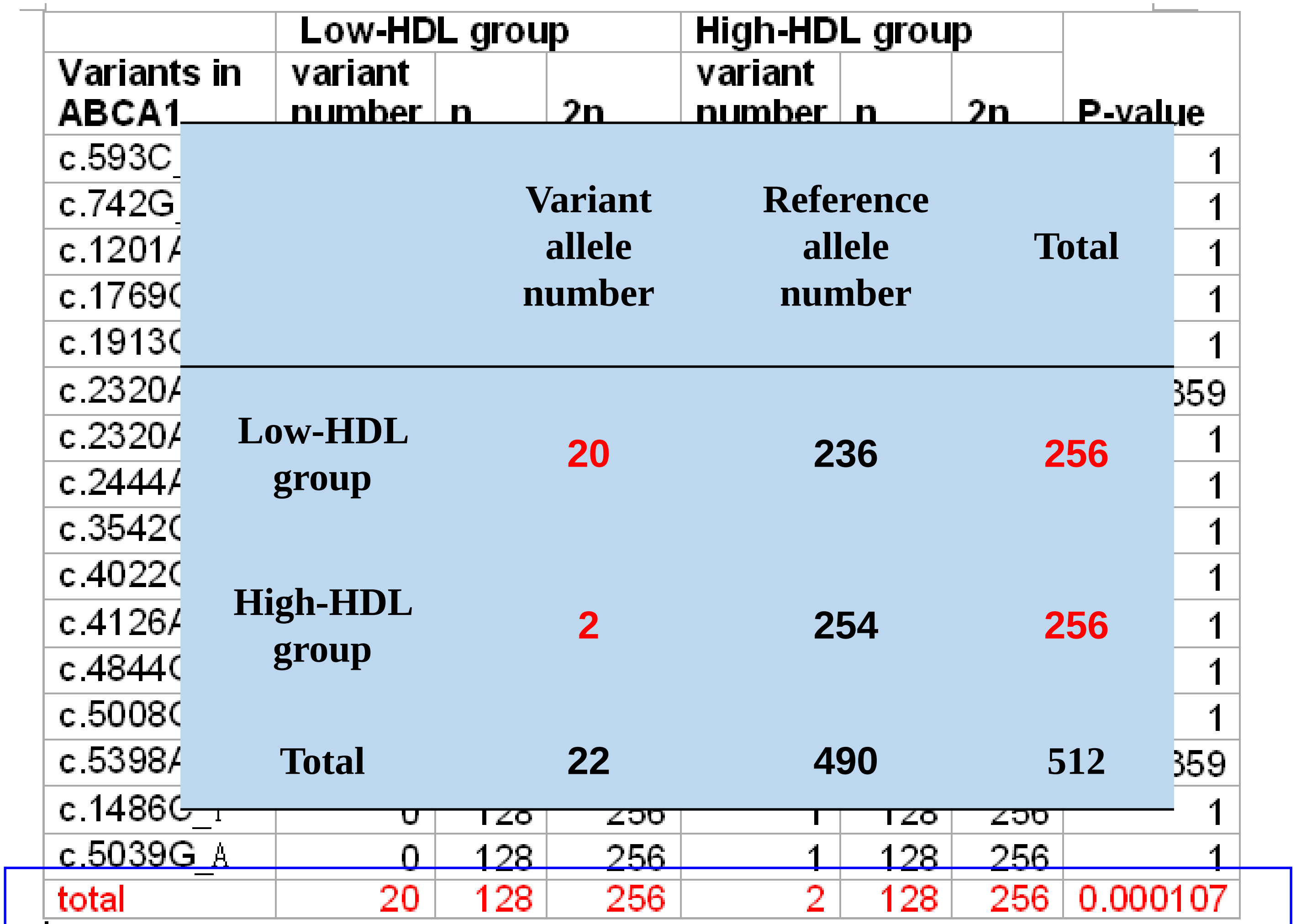
Level 2: Summarized by Subjects



	Low-HDL group			High-HDL group			
Variants in ABCA1	variant number	n	2n	variant number	n	2n	P-value
c.593C_A	1	128	256	0	128	256	1
c.742G_A	1	128	256	0	128	256	1
c.1201A_C	1	128	256	0	128	256	1
c.1769G_C	1	128	256	0	128	256	1
c.1913G_A	1	128	256	0	128	256	1
c.2320A_T	4	128	256	0	128	256	0.12359
c.2320A_T	1	128	256	0	128	256	1
c.2444A_G	1	128	256	0	128	256	1
c.3542C_T	1	128	256	0	128	256	1
c.4022G_C	1	128	256	0	128	256	1
c.4126A_G	1	128	256	0	128	256	1
c.4844G_A	1	128	256	0	128	256	1
c.5008G_A	1	128	256	0	128	256	1
c.5398A_C	4	128	256	0	128	256	0.12359
c.1486C_T	0	128	256	1	128	256	1
c.5039G_A	0	128	256	1	128	256	1

Cohen et al, Science, 2004

Level 3: Summarized by subjects and gene

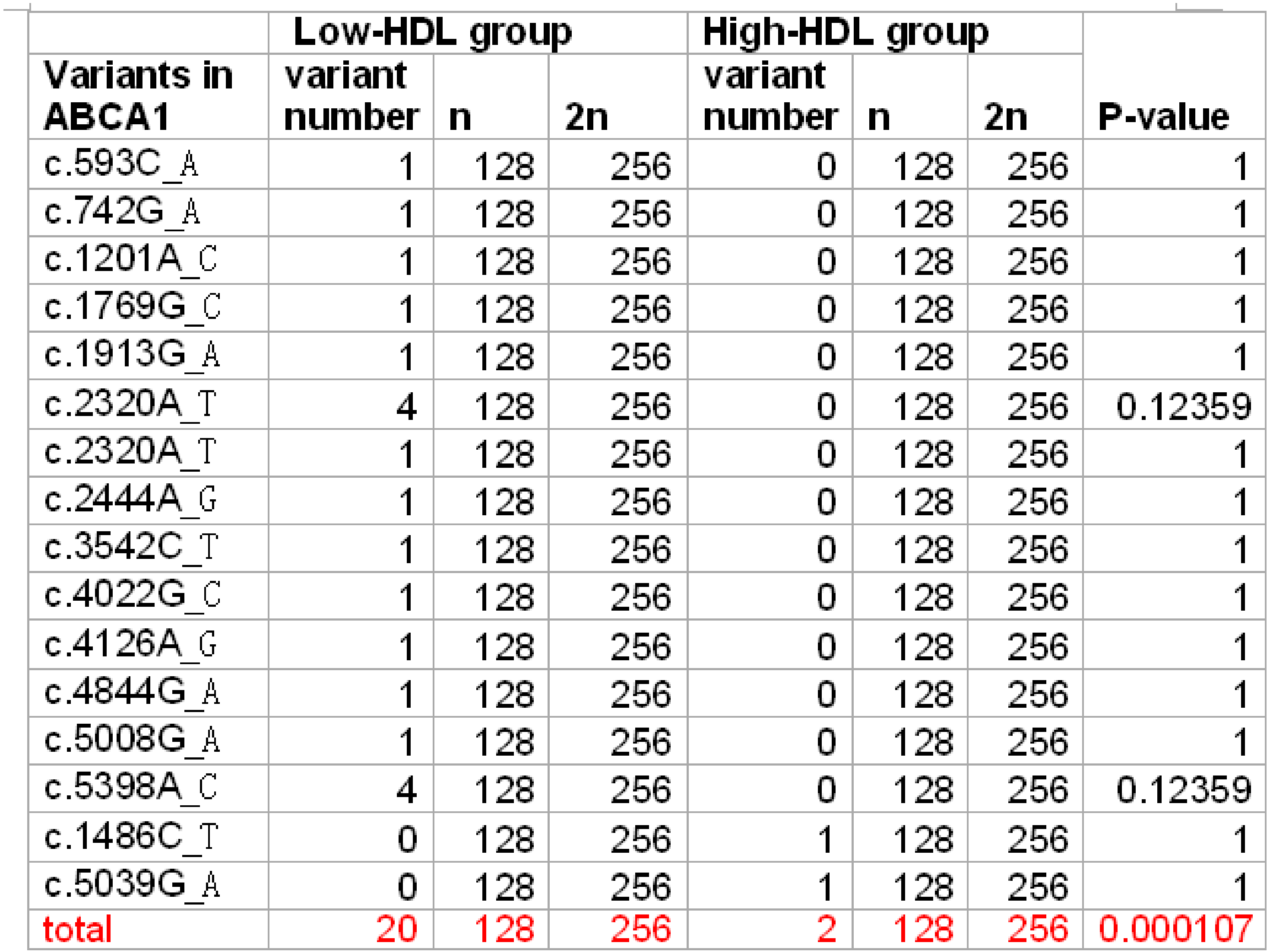


	Low-HDL group			High-HDL group			
Variants in ABCA1	variant number	n	2n	variant number	n	2n	P-value
c.593C_							1
c.742G_							1
c.1201A							1
c.1769C							1
c.1913C							1
c.2320A							359
c.2320A							1
c.2444A							1
c.3542C							1
c.4022C							1
c.4126A							1
c.4844C							1
c.5008C							1
c.5398A							359
c.1486G_T	0	128	256	1	128	256	1
c.5039G_A	0	128	256	1	128	256	1
total	20	128	256	2	128	256	0.000107

	Variant allele number	Reference allele number	Total
Low-HDL group	20	236	256
High-HDL group	2	254	256
Total	22	490	512

Fisher's exact test

Single-variant Test vs Total Freq Test



	Low-HDL group			High-HDL group			
Variants in ABCA1	variant number	n	2n	variant number	n	2n	P-value
c.593C_A	1	128	256	0	128	256	1
c.742G_A	1	128	256	0	128	256	1
c.1201A_C	1	128	256	0	128	256	1
c.1769G_C	1	128	256	0	128	256	1
c.1913G_A	1	128	256	0	128	256	1
c.2320A_T	4	128	256	0	128	256	0.12359
c.2320A_T	1	128	256	0	128	256	1
c.2444A_G	1	128	256	0	128	256	1
c.3542C_T	1	128	256	0	128	256	1
c.4022G_C	1	128	256	0	128	256	1
c.4126A_G	1	128	256	0	128	256	1
c.4844G_A	1	128	256	0	128	256	1
c.5008G_A	1	128	256	0	128	256	1
c.5398A_C	4	128	256	0	128	256	0.12359
c.1486C_T	0	128	256	1	128	256	1
c.5039G_A	0	128	256	1	128	256	1
total	20	128	256	2	128	256	0.000107

Cohen et al, Science, 2004

Burden Tests

- Tests
  - Binary collapsing: CAST
  - CMC
  - Count collapsing
  - Weighted sum test

- Power of burden tests depends on
  - Number of associated variants
  - Number of non-associated variants
  - Direction of the effects

- Powerful if most variants are causal and have effects in the same direction

Cohort Allelic Sums Test (CAST)

- A group of *n* variants (e.g., SNPs) in a unit (e.g. one gene)
- Collapse the genotypes across the variants
- Coding for individual *i*
  - x_i = 1, if rare alleles present at any of the *n* variants;
  - x_i = 0, otherwise
- Test if the proportions of individuals with rare variants in cases and controls differ
- Higher power than method testing single variant each time

Morgenthaler et al, Mutation Research, 2007

Combined Multivariate and Collapsing (CMC) Method

- Division and Collapsing
  - Divide SNPs into several sub-groups based on MAF
    - Ex. Subgroups : (0, 0.001), [0.001, 0.005), [0.005, 0.01)
  - SNPs are collapsed in each sub-group
    - x_{ij} = 1, if individual *j* has rare alleles present in the *i*-th subgroup;
    - x_{ij} = 0, otherwise

Li and Leal, American Journal of Human Genetics, 2008

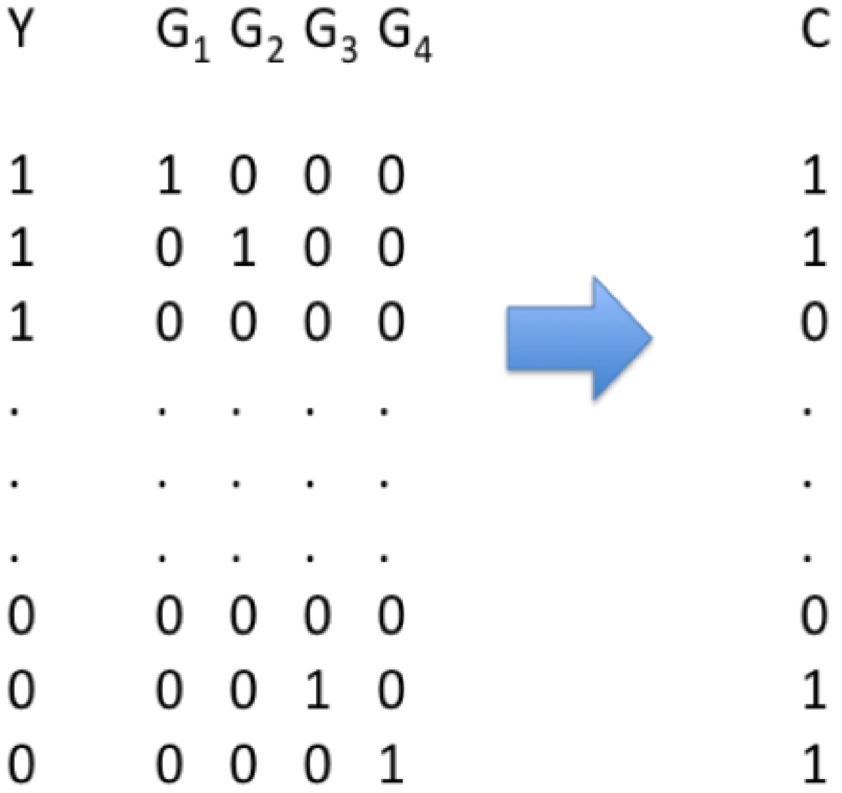
Combined Multivariate and Collapsing (CMC) Method

- Multivariate test of collapsed sub-groups
  - Hotelling T^2 test, MANOVA, Fisher's product method

- Comparison of power: often higher than CAST

- Different thresholds may have different power

Burden Tests: Mixed Effect Directions



	Y	G_1	G_2	G_3	G_4	C
	1	1	0	0	0	1
diseased	1	0	1	0	0	1
	1	0	0	0	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	0	0	0	0	0	0
normal	0	0	0	1	0	1
	0	0	0	0	1	1

Burden tests will lose power if variants have positive and negative effects.

Adaptive Burden Tests

- Several methods have been developed to estimate association directions and incorporate them in the burden test
  - Adaptive sum test
  - Estimated regression coefficient test

Adaptive Sum Test

- Model: weighted genotype score for individual i

$$C_i = \sum_{j=1}^{p} w_j g_{ij}$$

$$logit(Pr(Y=1)) = \alpha_0 + C_i\beta$$

- Fit individual SNP models

$$logit(Pr(Y=1)) = \alpha_0 + g_j\beta_j$$

Assign $w_j = -1$ if $\widehat{\beta}_j < 0$ and the p-value is small
$w_j = 1$ otherwise.

- Compute p-values by permutation

Han et al, Hum Hered. 2010

Variance-Component Tests

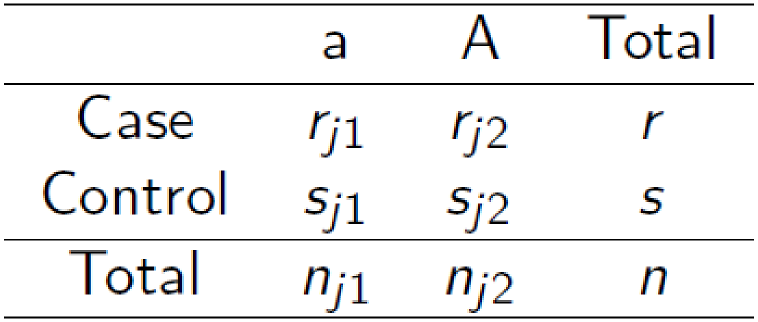
- Burden tests are not powerful, if there exist variants with different association directions or many non-causal variants.

- Adaptive burden tests are often computationally intensive due to permutation.

- Variance-component tests have been proposed to address these issues.

C-alpha Test

- Case-control studies without covariates
- For SNP *j* , the data can be summarized as



	a	A	Total
Case	r_{j1}	r_{j2}	r
Control	s_{j1}	s_{j2}	s
Total	n_{j1}	n_{j2}	n

- Under H0:

$$r_{j1} \sim Binomial(n_{j1}, q) \quad (q = r/n)$$

Neale BM, et al, PLoS Genet, 2011

C-alpha Test

- Risk increasing variant:

$$r_{j1} - qn_{j1} > 0$$

- Risk decreasing variant:

$$r_{j1} - qn_{j1} < 0$$

- Test statistic:

$$T_\alpha = \sum_{j=1}^{p} (r_{j1} - qn_{j1})^2 - \sum_{j=1}^{p} n_{j1} q(1-q)$$

- This test is robust in the presence of opposite association directions

C-alpha Test

- Weighting scheme

$$T_\alpha = \sum_{j=1}^{p} w_j (r_{j1} - q n_{j1})^2 - \sum_{j=1}^{p} w_j n_{j1} q(1-q)$$

- Test for the **over-dispersion** due to genetic effects

- Advantage: robust in the presence of different directions

- Disadvantage: cannot adjust for covariates

Sequence Kernel Sequential Test (SKAT)

- Standard regression model for individual *i*:

$$logit(\mu_i) = \alpha_0 + \mathbf{X}_i^T \boldsymbol{\alpha} + \mathbf{G}_i^T \beta$$

G*i*: genotype vector

X*i*: covariates

- Variance component test:

$$\text{Assume } \beta_j \sim dist.(0, w_j^2 \tau).$$

$$H_0 : \beta_1 = \cdots = \beta_p = 0 <=> \boxed{H_0 : \tau = 0.}$$

Wu et al, AJHG, 2011

SKAT vs Collapsing Tests

- Collapsing tests are more powerful when a large % of variants are causal and effects in the same direction
- SKAT is more powerful when a small % of variants are causal or the effects have mixed directions
- Both scenarios can happen when scanning the genome
- Best test to use depends on the underlying biology

There is a need to develop a unified test that works well in both situations → Omnibus tests.

Combine P-values of Burden & SKAT

Fisher method:

$$Q_{Fisher} = -2\log(P_{Burden}) - 2\log(P_{SKAT})$$

Q_{Fisher} follows χ^2 with 4 d.f when these two p-values are independent

Since they are not independent, p-values are calculated using resampling

Mist (Sun et al. 2013) modified the SKAT test statistics to make them independent

Derkach et al, Genet Epi, 2013

Unified Test Statistic — SKAT-O

- Combined test of Burden tests and SKAT

$$Q_\rho = (1-\rho)Q_{SKAT} + \rho Q_{Burden}, \quad 0 \leq \rho \leq 1.$$

- Use the smallest p-value from different *ρ*s:

$$T = minP_{\rho_b}, \quad 0 = \rho_1 < \ldots < \rho_B = 1.$$

where P_ρ is the p-value of Q_ρ for given ρ.

- The asymptotic p-value of SKAT-O can be calculated with computationally efficient 1-dim numerical integration.

Lee et al, Biostatistics, 2012

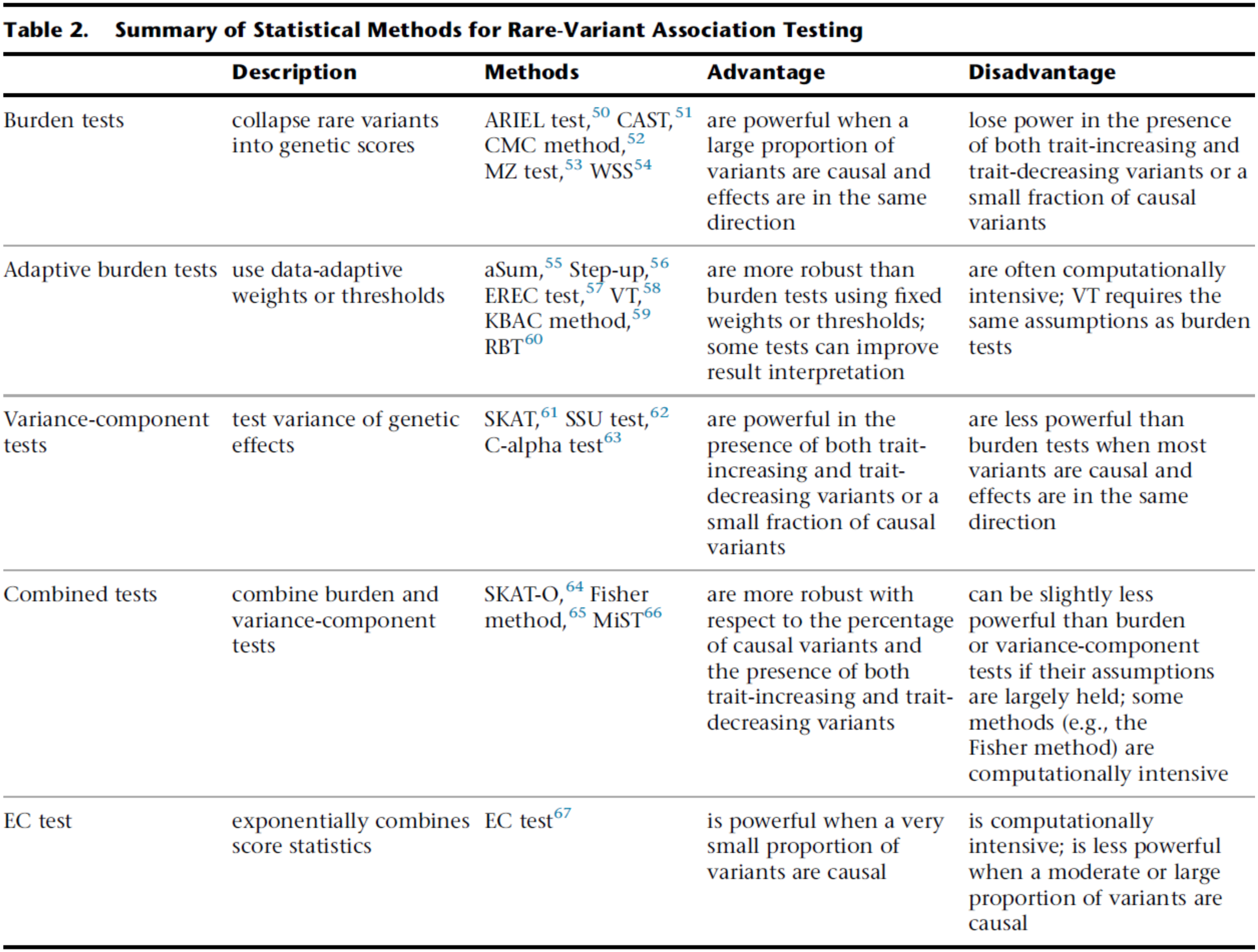


Table 2. Summary of Statistical Methods for Rare-Variant Association Testing

	Description	Methods	Advantage	Disadvantage
Burden tests	collapse rare variants into genetic scores	ARIEL test,[50] CAST,[51] CMC method,[52] MZ test,[53] WSS[54]	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants
Adaptive burden tests	use data-adaptive weights or thresholds	aSum,[55] Step-up,[56] EREC test,[57] VT,[58] KBAC method,[59] RBT[60]	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests
Variance-component tests	test variance of genetic effects	SKAT,[61] SSU test,[62] C-alpha test[63]	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction
Combined tests	combine burden and variance-component tests	SKAT-O,[64] Fisher method,[65] MiST[66]	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive
EC test	exponentially combines score statistics	EC test[67]	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal

Lee et al, Am J Hum Genet, 2014

Maximizing the Power

- Power of rare variant analysis depends on summed frequency -- threshold for defining rare is critical, but difficult to specify.

- Rare causal variants can be enriched in extreme phenotypic samples.

- Given the fixed budget, increase power by sequencing extreme phenotypic samples.

- For binary traits, focus on individuals with family history of disease, or select super controls.

Strategies to find high impact novel risk genes

- Select cases that have strong genetic contribution
  - Familial (breast cancer)
  - Early onset (developmental disorders)
  - Extreme forms (diabetes and obesity studies)

- Focus on most rare and deleterious variants

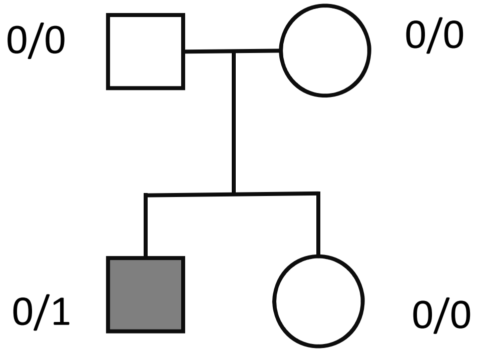
Developmental disorders

- Developmental delay and intellectual disability
- Autism spectrum disorders
- Epilepsy
- Structural birth defects: congenital heart disease, congenital diaphragmatic hernia, etc

Severely selected: either lethal without surgery, or difficult to establish stable families and produce offsprings.

De novo mutations

Major contributor to developmental disorders



Absent in parents

Present in proband(s)

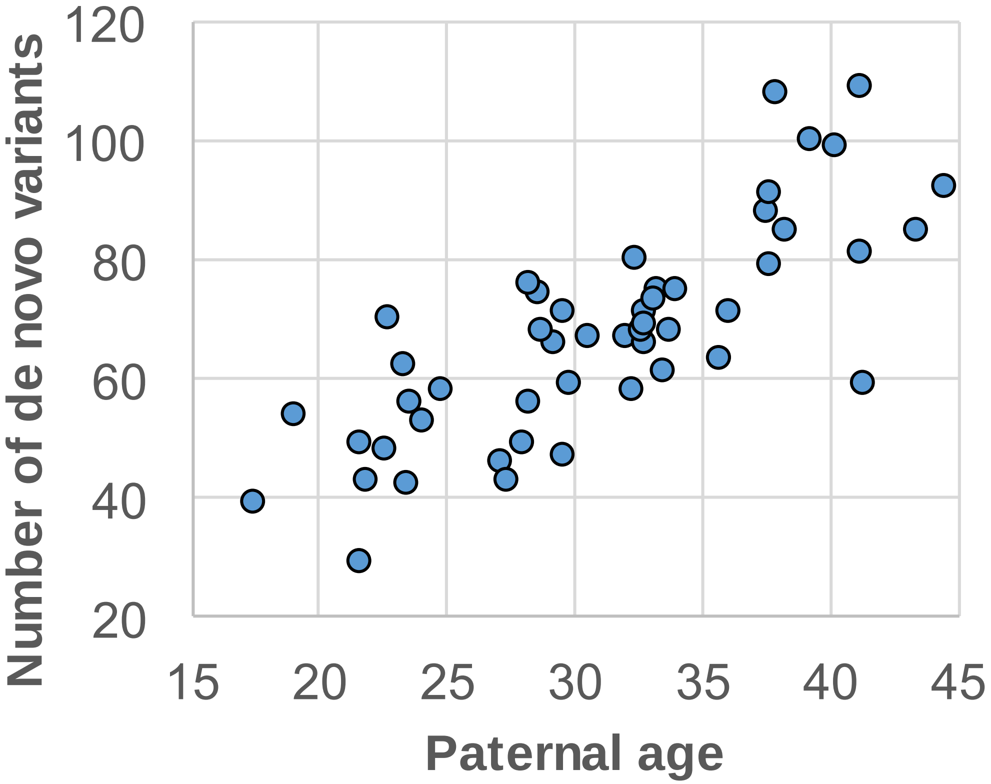
De novo mutations

- Expectation: background rate

- Background rate (Samocha et al 2014):
    - Most important thing: transition and transversion, 10x difference in rate

    - Local context
    - Replication timing
    - Paternal age (germline cell biology)

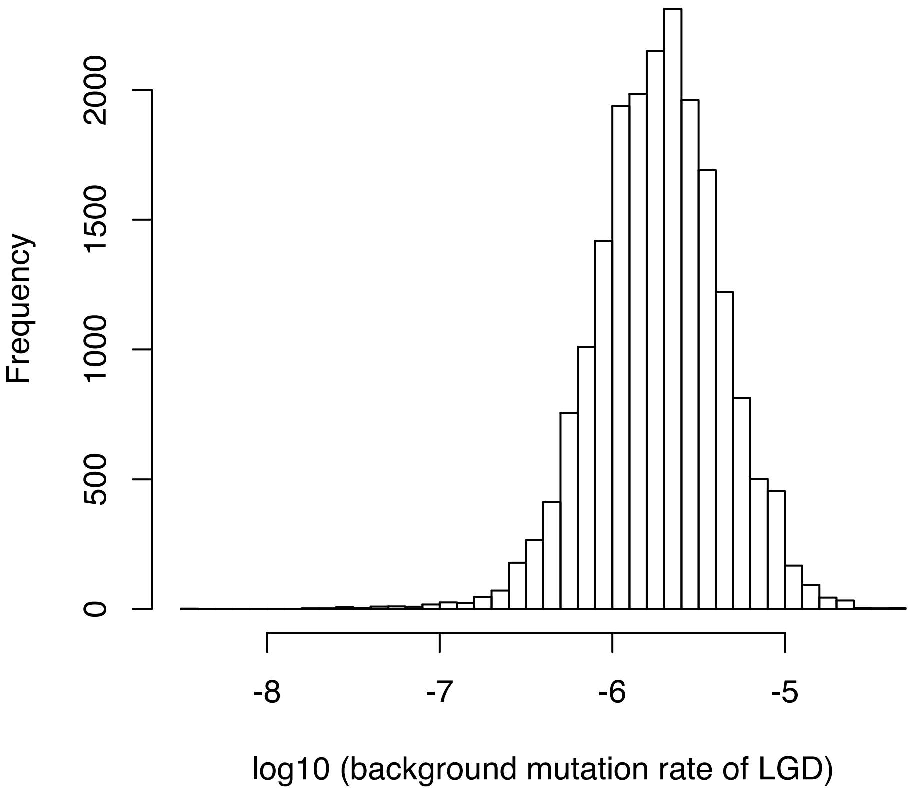
Paternal age and *de novo* mutations

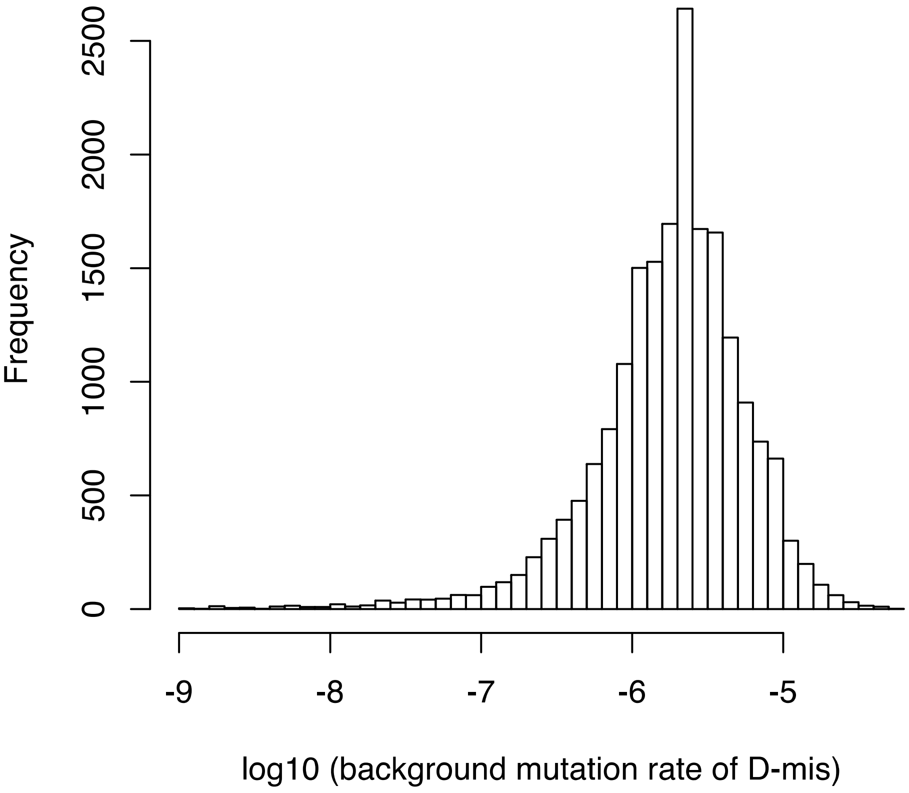


Data from WGS
of 50 CHD trios

2 de novo mutations / year;
R^2 ~ 0.6

Background mutation rate



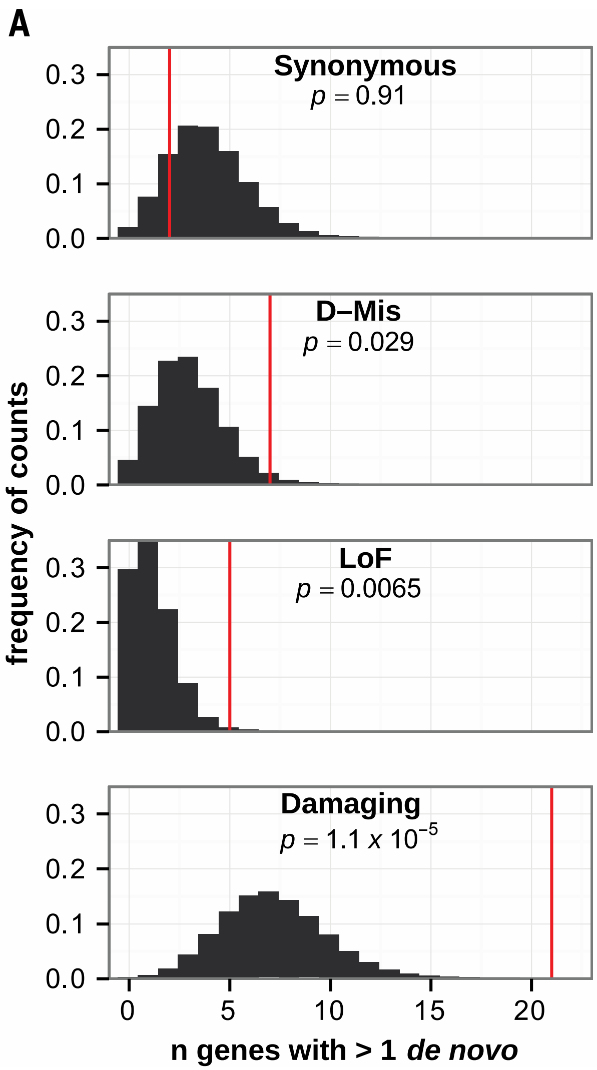


Median: ~ $2x10^{-6}$

Strategies of finding novel risk genes from de novo mutations

- Statistical evidence alone:
    - Poisson test: M0: N ~ Poisson(λ). Bonferroni threshold: $2.5\text{x}10^{-6}$
- Data-driven FDR:
    - Group genes based on haploinsufficiency, mutation intolerance, or gene expression (aka relevancy to a disease)
    - Can prioritize even singletons
- Network enrichment
    - There are limited number of risk pathways (or functional modules) → True risk genes (among putative risk genes) are more likely to be functionally related
- Rank by functional similarity to known risk genes
    - ToppGene
    - Phenolyzer etc

Genes with >=2 damaging de novo mutations



B

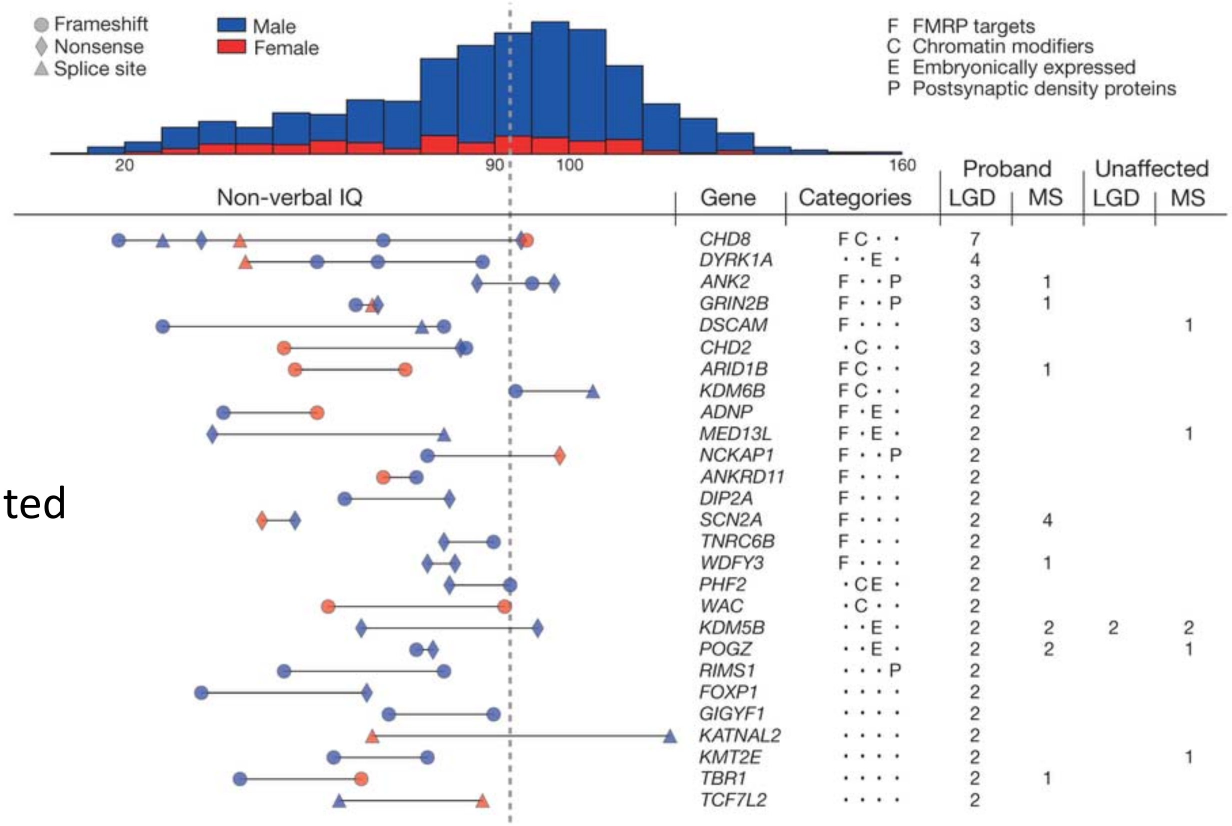
Gene	LoF	D-Mis	p
PTPN11	0	4	2.90 *x* 10^{-11}
KMT2D	4	2	4.10 *x* 10^{-09}
RBFOX2	3	0	3.40 *x* 10^{-08}
KDM5B	3	0	2.88 *x* 10^{-06}
KRT13	0	2	1.01 *x* 10^{-05}
MYH6	0	3	2.41 *x* 10^{-05}
CAD	0	3	3.73 *x* 10^{-05}
NAA15	2	0	4.67 *x* 10^{-05}
SMAD2	1	1	1.08 *x* 10^{-04}
RABGAP1L	1	1	4.01 *x* 10^{-04}
POGZ	1	1	4.33 *x* 10^{-04}
JAG1	1	1	4.47 *x* 10^{-04}
GANAB	1	1	4.52 *x* 10^{-04}
DTNA	1	1	4.68 *x* 10^{-04}
PPL	1	1	5.98 *x* 10^{-04}
CHD7	2	0	6.16 *x* 10^{-04}
ZEB2	1	1	6.18 *x* 10^{-04}
FBN1	0	2	6.79 *x* 10^{-04}
CHD4	0	2	1.16 *x* 10^{-03}
AHNAK	1	1	2.88 *x* 10^{-03}
NOTCH1	1	1	4.35 *x* 10^{-03}

Congenital heart disease

Homsy et al , Science, 2015

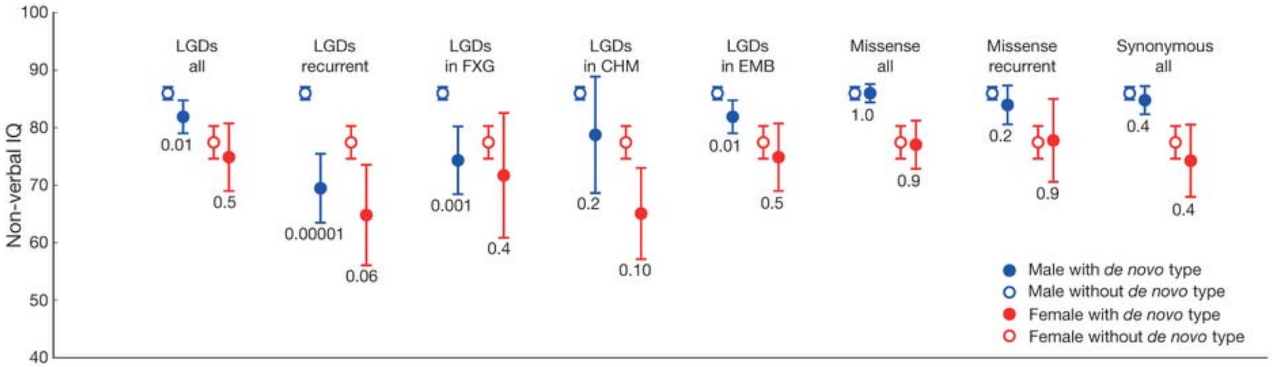
Autism:

Recurrently mutated genes and IQ



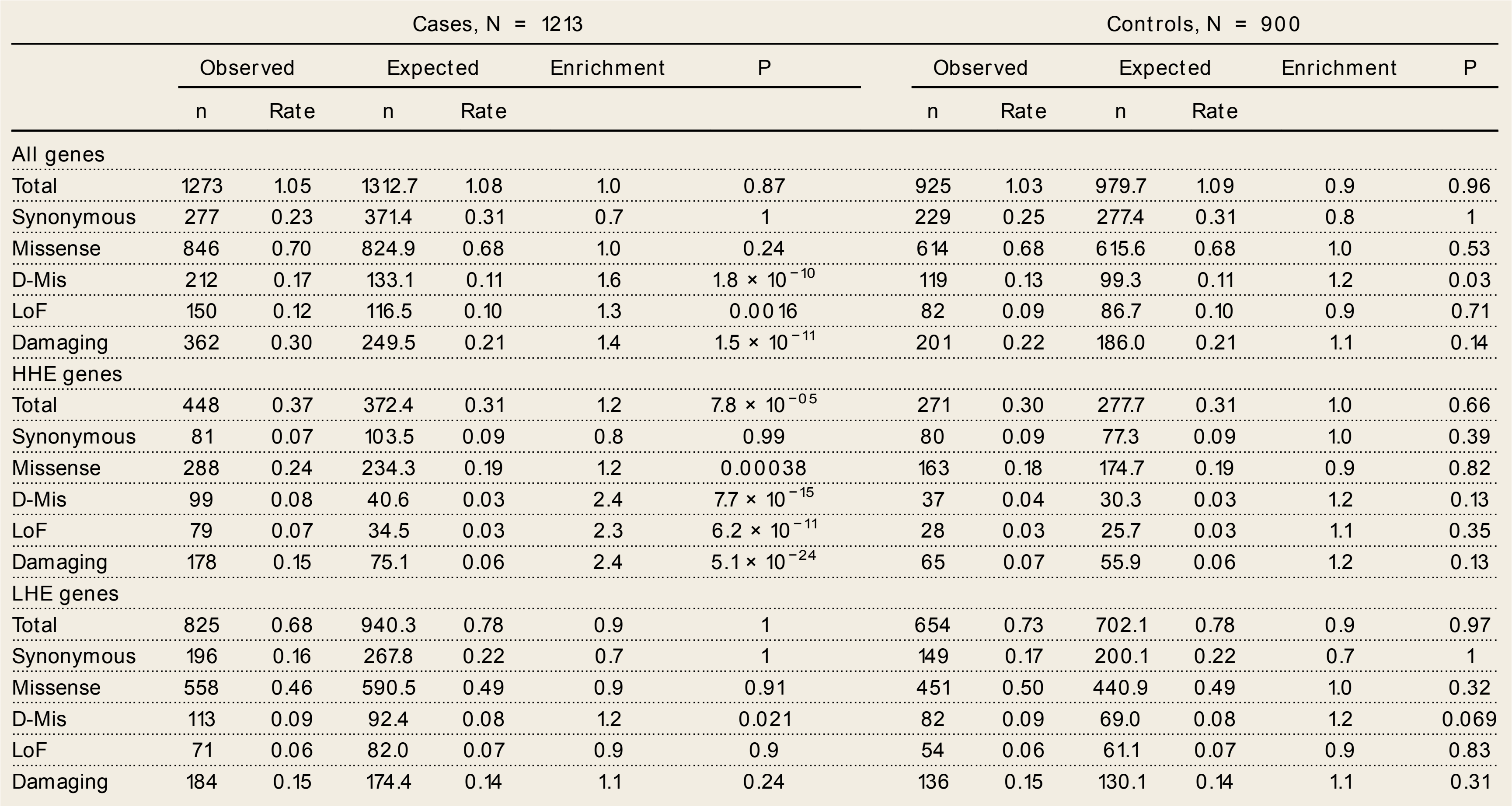
Gene	Categories	Proband LGD	Proband MS	Unaffected LGD	Unaffected MS
CHD8	F C · ·	7			
DYRK1A	· · E ·	4			
ANK2	F · · P	3	1		
GRIN2B	F · · P	3	1		
DSCAM	F · · ·	3			1
CHD2	· C · ·	3			
ARID1B	F C · ·	2	1		
KDM6B	F C · ·	2			
ADNP	F · E ·	2			
MED13L	F · E ·	2			1
NCKAP1	F · · P	2			
ANKRD11	F · · ·	2			
DIP2A	F · · ·	2			
SCN2A	F · · ·	2	4		
TNRC6B	F · · ·	2			
WDFY3	F · · ·	2	1		
PHF2	· C E ·	2			
WAC	· C · ·	2			
KDM5B	· · E ·	2	2	2	2
POGZ	· · E ·	2	2		1
RIMS1	· · · P	2			
FOXP1	· · · ·	2			
GIGYF1	· · · ·	2			
KATNAL2	· · · ·	2			
KMT2E	· · · ·	2			1
TBR1	· · · ·	2	1		
TCF7L2	· · · ·	2			

Mutation type and IQ



Iossifov et al, Nature, 2014

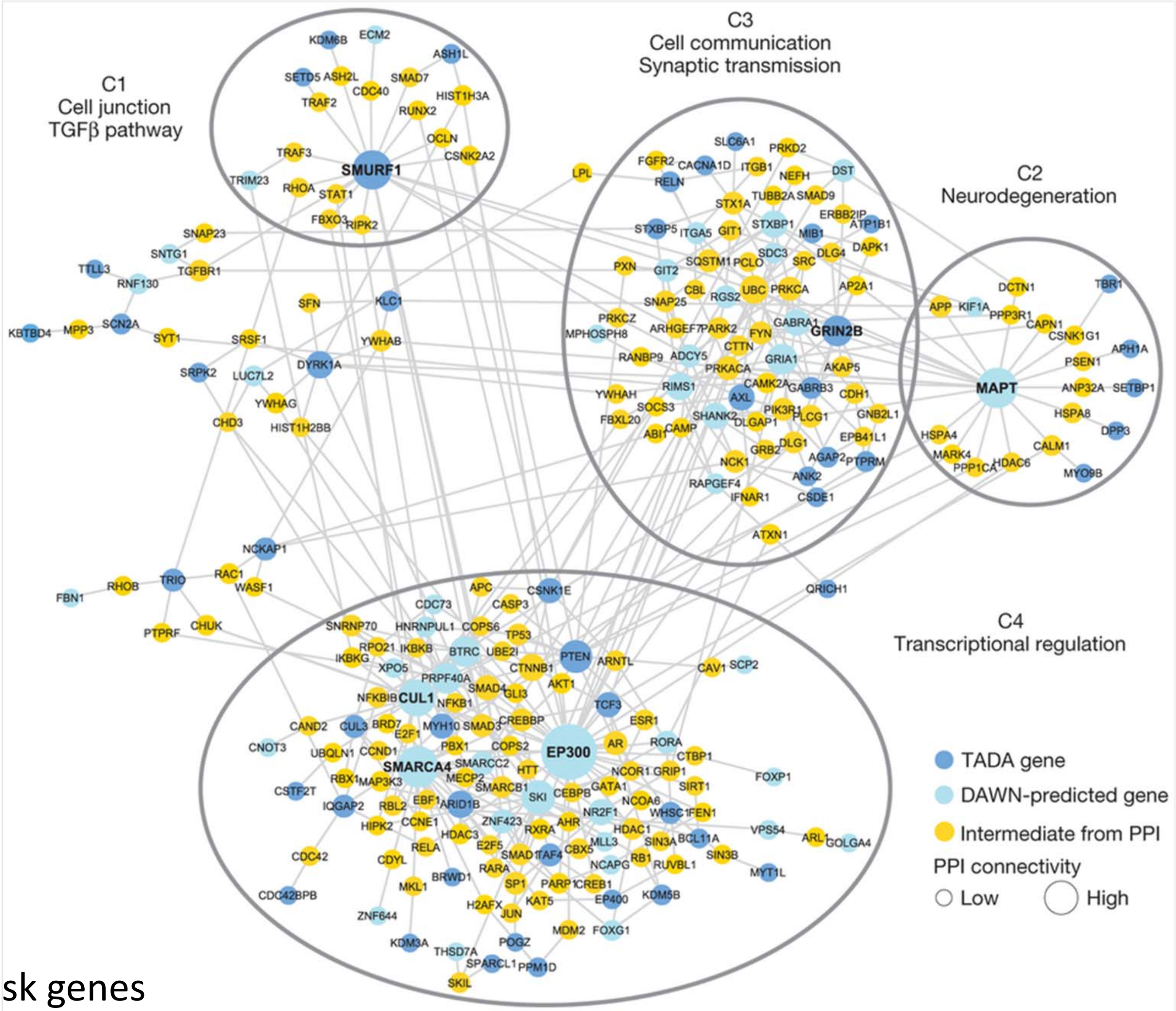
Most CHD risk genes are highly expressed in developing heart



	Cases, N = 1213						Controls, N = 900					
	Observed		Expected		Enrichment	P	Observed		Expected		Enrichment	P
	n	Rate	n	Rate			n	Rate	n	Rate		
All genes												
Total	1273	1.05	1312.7	1.08	1.0	0.87	925	1.03	979.7	1.09	0.9	0.96
Synonymous	277	0.23	371.4	0.31	0.7	1	229	0.25	277.4	0.31	0.8	1
Missense	846	0.70	824.9	0.68	1.0	0.24	614	0.68	615.6	0.68	1.0	0.53
D-Mis	212	0.17	133.1	0.11	1.6	1.8×10^{-10}	119	0.13	99.3	0.11	1.2	0.03
LoF	150	0.12	116.5	0.10	1.3	0.0016	82	0.09	86.7	0.10	0.9	0.71
Damaging	362	0.30	249.5	0.21	1.4	1.5×10^{-11}	201	0.22	186.0	0.21	1.1	0.14
HHE genes												
Total	448	0.37	372.4	0.31	1.2	7.8×10^{-05}	271	0.30	277.7	0.31	1.0	0.66
Synonymous	81	0.07	103.5	0.09	0.8	0.99	80	0.09	77.3	0.09	1.0	0.39
Missense	288	0.24	234.3	0.19	1.2	0.00038	163	0.18	174.7	0.19	0.9	0.82
D-Mis	99	0.08	40.6	0.03	2.4	7.7×10^{-15}	37	0.04	30.3	0.03	1.2	0.13
LoF	79	0.07	34.5	0.03	2.3	6.2×10^{-11}	28	0.03	25.7	0.03	1.1	0.35
Damaging	178	0.15	75.1	0.06	2.4	5.1×10^{-24}	65	0.07	55.9	0.06	1.2	0.13
LHE genes												
Total	825	0.68	940.3	0.78	0.9	1	654	0.73	702.1	0.78	0.9	0.97
Synonymous	196	0.16	267.8	0.22	0.7	1	149	0.17	200.1	0.22	0.7	1
Missense	558	0.46	590.5	0.49	0.9	0.91	451	0.50	440.9	0.49	1.0	0.32
D-Mis	113	0.09	92.4	0.08	1.2	0.021	82	0.09	69.0	0.08	1.2	0.069
LoF	71	0.06	82.0	0.07	0.9	0.9	54	0.06	61.1	0.07	0.9	0.83
Damaging	184	0.15	174.4	0.14	1.1	0.24	136	0.15	130.1	0.14	1.1	0.31

Homsy et al , Science, 2015

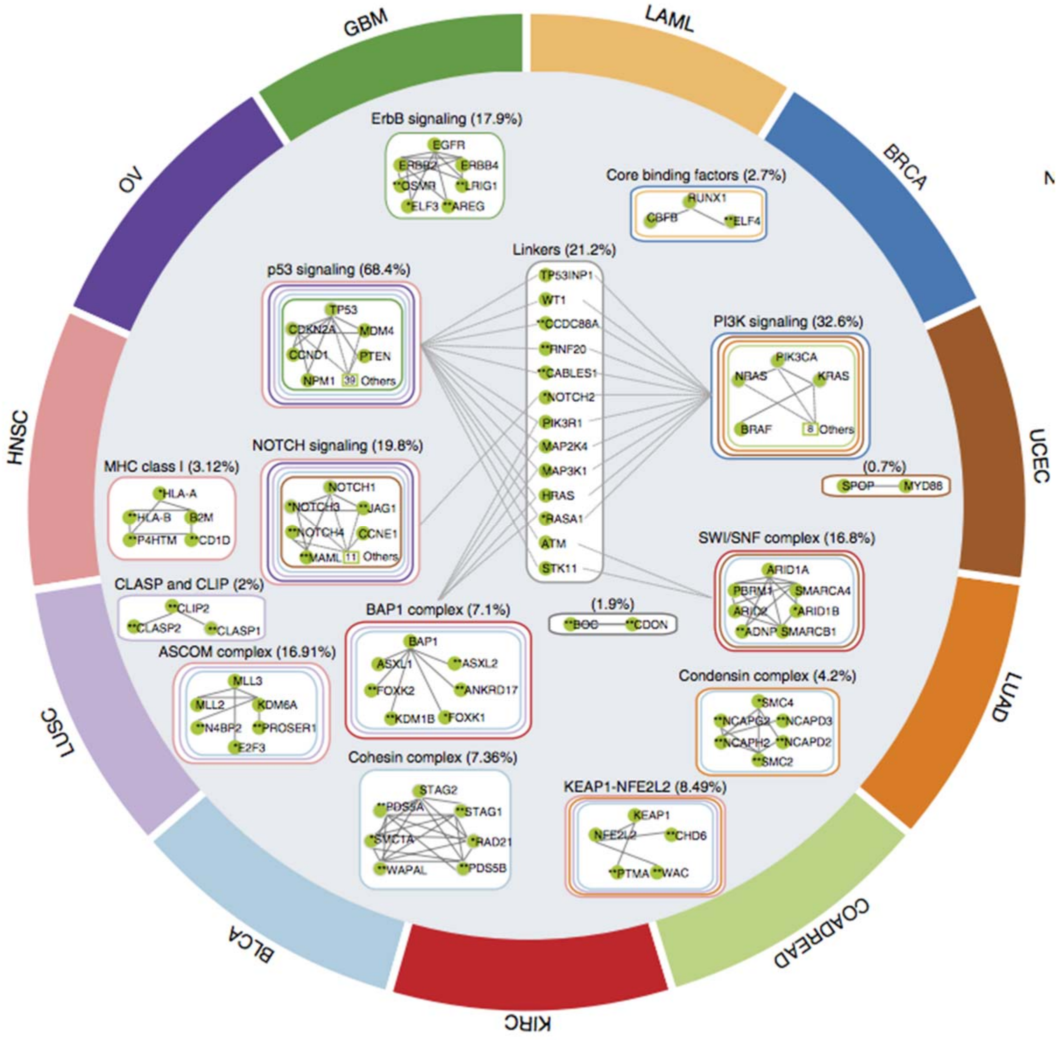
Network analysis



~100 candidate risk genes

De Rubeis et al, Nature, 2014

Network analysis



consensus subnetworks are arranged near the cancer types

Leiserson et al, Nature Genetics, 2015

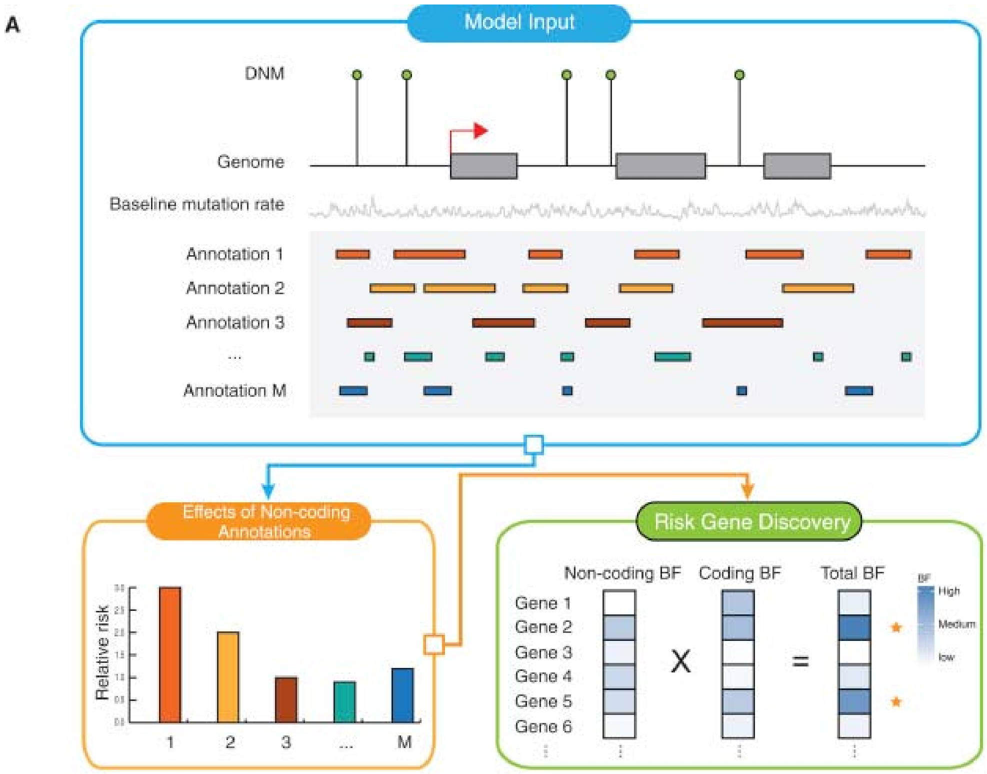
Putting de novo mutations and inherited mutations together

- TADA: Transmission And De novo Association
- Incorporate WES data regarding de novo mutations, inherited variants present, and variants identified within cases and controls.
- Integrates these data by a gene-based likelihood model involving parameters for allele frequencies and gene-specific penetrances.

He et al, PLoS Genetics, 2013

TADA-Annotations (TADA-A)

- It incorporates many functional annotations such as conservation and enhancer marks, to learn from data which annotations are informative of pathogenic mutations, and to combine both coding and non-coding mutations at the gene level to detect risk genes



Liu et al, Am J Hum Genet, 2018