PGC Worldwide Lab Call Details

DATE: Friday, March 15th, 2013
PRESENTER: Shaun Purcell, Mt Sinai School of Medicine, NYC
TITLE: "An overview of sequencing studies and their application to psychiatric disease"
START: We will begin promptly on the hour.

1000 EST - US East Coast
0700 PST - US West Coast
1400 GMT - UK
1500 CET - Central Europe
0100 EDT – Australia (Saturday, March 16th, 2013)

DURATION: 1 hour

TELEPHONE:
- US Toll free: 1 866 515.2912
- International direct: +1 617 399.5126
- Toll-free number? See http://www.btconferencing.com/globalaccess/?bid=75_public
- Operators will be on standby to assist with technical issues. "*0" will get you assistance.
- This conference line can handle up to 300 participants.

PASSCODE: 275 694 38

Lines are Muted **NOW**

Lines have been automatically muted by operators as it is possible for just one person to ruin the call for everyone due to background noise, electronic feedback, crying children, wind, typing, etc.

Operators announce calls one at a time during question and answer sessions.

Dial *1 if you would like to ask a question of the presenter. Presenter will respond to calls as time allows.

Dial *0 if you need operator assistance at any time during the duration of the call.

UPCOMING PGC Worldwide Lab

DATE: Friday, April 12th, 2013
PRESENTER: To Be Announced
TITLE: To Be Announced
START: We will begin promptly on the hour.
1000 EDT - US East Coast
0700 PDT - US West Coast
1500 BST - UK
1600 CEST - Central Europe
2400 EST – Australia
DURATION: 1 hour

TELEPHONE:
- US Toll free: 1 866 515.2912
- International direct: +1 617 399.5126
- Toll-free number? See http://www.btconferencing.com/globalaccess/?bid=75_public
- Operators will be on standby to assist with technical issues. "*0" will get you assistance.
- This conference line can handle up to 300 participants.

PASSCODE: 275 694 38

An overview of sequencing studies and their application to psychiatric disease

Shaun M. Purcell

shaun.purcell@mssm.edu

Overview

Models	*Rationale for studying rare variants* *Population genetics & selection*
Data	*Application of NGS technologies* *What do the data look like?* *Common error modes*
Designs	*Mendelian designs* *Multiplex families* *De novo studies* *Population-based approaches*
Analysis	*Power and rare variants* *Strategies to improve power* *Other potential problems*
Prospects	*Current literature* *Emerging studies* *PGC and sequencing studies*

MODELS

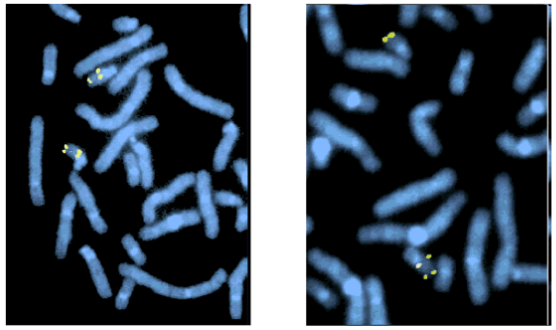
Enthusiasm for studying rare variation in common disease

- Precedent from Mendelian disease genetics
  - rare disease alleles strongly increase risk for rare disease
- Genome-wide association studies
  - "missing heritability" beyond specific, detected common variants
  - rare variation effectively not captured by common SNP platforms
- Population genetic theory (i.e. natural selection works)
  - most new mutations expected to be mildly deleterious
  - highly penetrant disease alleles will be selected against
  - (accepting viability) at the extreme, *de novo* mutation is uncensored w.r.t natural selection
- Single, highly-penetrant alleles may be easier to characterize functionally
  - particularly if the variant induces loss-of-function for a single gene
- Next generation sequencing
  - because now we can…
  - although note that decades of *linkage analysis* also constituted a window into rare variation and common disease

Rare and *de novo* mutations already documented in schizophrenia (and other psychiatric disease)

Deletion on 22q11.2, 1 in 4000 live births (Velo-Cardio-Facial Syndrome, VCFS)

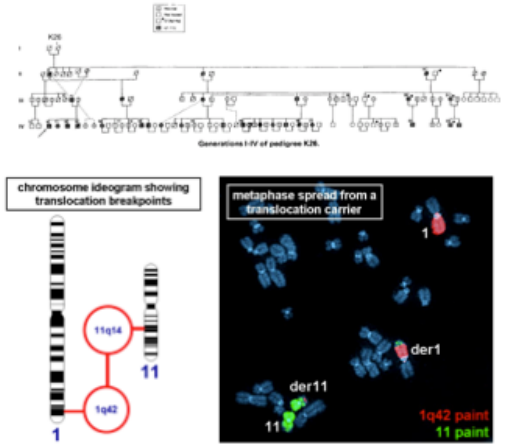
Fluorescence in situ hybridization (FISH)



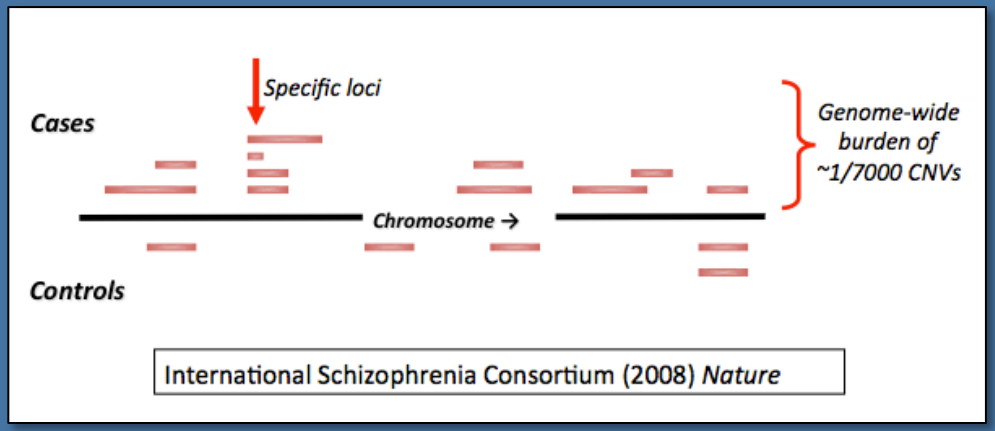
Unaffacted individual *VCFS patient*

~1% of all SCZ patients

Linkage analysis in a large pedigree detects a segregating translocation with breakpoint in *DISC1* gene



St Clair et al. Lancet 1990; 336:13-6;
Blackwood et al. Am J Hum Genet. 2001; 69:428-33



International Schizophrenia Consortium (2008) *Nature*

From GWAS/CNV studies, cases have a greater burden of "singleton" (ultra rare/*de novo*) micro-deletions and duplications

Enthusiasm versus realism

PERSPECTIVE

nature genetics

Exome sequencing and the genetic basis of complex traits

Adam Kiezun[1,2,16], Kiran Garimella[2,16], Ron Do[2,3,16], Nathan O Stitziel[2,4,16], Benjamin M Neale[2,3,5], Paul J McLaren[1,2], Namrata Gupta[2], Pamela Sklar[6,7], Patrick F Sullivan[8], Jennifer L Moran[2], Christina M Hultman[9], Paul Lichtenstein[9], Patrik Magnusson[9], Thomas Lehner[10], Yin Yao Shugart[11], Alkes L Price[2,12,13,17], Paul I W de Bakker[1,2,14,15,17], Shaun M Purcell[5,17] & Shamil R Sunyaev[1,2,17]

Table 2 Summary of gene burden test results for rare variant studies

Trait	Gene	Test	AC[a] low	AC[a] high	*n*	*P*	Ref.
Triglycerides	*ANGPTL4*	Fisher's exact	13	2	1,775	0.016[b]	26
Triglycerides	*ANGPTL5*	Fisher's exact	9	1	1,775	0.022[b]	
HDL	*ABCA1*	RVE	28	4	519	<0.0001[b]	21
	APOA1		1	0	519		
	LCAT		6	1	519		
Blood pressure	*SLC12A1*, *SLC12A3*, *KCNJ1*	Fisher's exact	9	1	626	0.02	22
Obesity	Obesity[c]	Fisher's exact	73	97	757	0.061	25
Type 1 diabetes	*IFIH1*	Fisher's exact	21	39	960	0.025	24
Triglycerides	*APOA5*	Fisher's exact	1	5	765	0.25	23
	GCKR	Fisher's exact	5	20	765	0.024	
	LPL	Fisher's exact	8	44	765	2.47×10^{-5}	
	APOB	Fisher's exact	39	85	765	0.008	

Of genes with rare variants previously detected in candidate studies of common disease, none would surpass exome-wide statistical thresholds (despite the moderately large samples N)

Recent insights into rare variation from 1000 Genomes and other large-scale exome sequencing projects: bottom line, there is a lot of it…

Sciencexpress Research Articles

Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes

Jacob A. Tennessen,[1*] Abigail W. Bigham,[2*†] Timothy D. O'Connor,[1*] Wenqing Fu,[1] Eimear E. Kenny,[3] Simon Gravel,[3] Sean McGee,[1] Ron Do,[4,5] Xiaoming Liu,[6] Goo Jun,[7] Hyun Min Kang,[7] Daniel Jordan,[8] Suzanne M. Leal,[9] Stacey Gabriel,[4] Mark J. Rieder,[1] Goncalo Abecasis,[7] David Altshuler,[4] Deborah A. Nickerson,[1] Eric Boerwinkle,[6,10] Shamil Sunyaev,[4,8] Carlos D. Bustamante,[3] Michael J. Bamshad,[1,2‡] Joshua M. Akey,[1‡] Broad GO, Seattle GO, on behalf of the NHLBI Exome Sequencing Project

Deep-sequencing >2000 individuals, conclude the majority of protein-coding variation is :

rare (86% of sites have minor allele frequency < 0.5%)

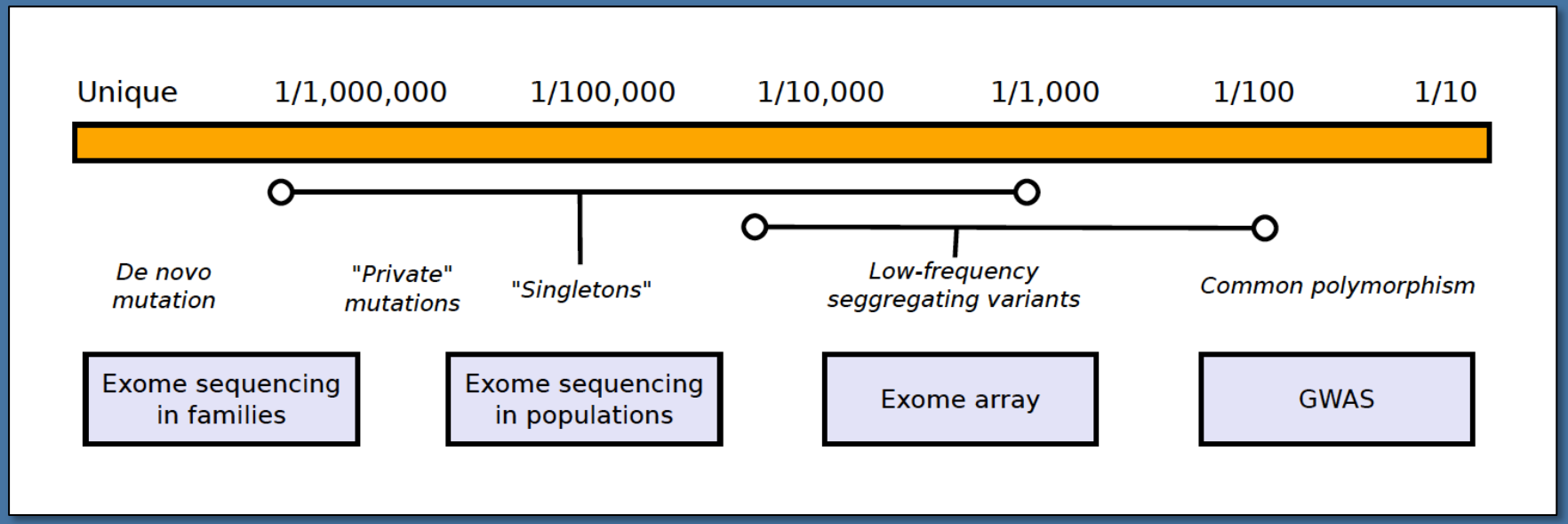
novel (82% never observed before)

population-specific (82% of sites)

and under **weak purifying selection**

Most people have **~300 genes whose function is deleteriously impacted** by a rare variant

Frequency spectrum of disease alleles

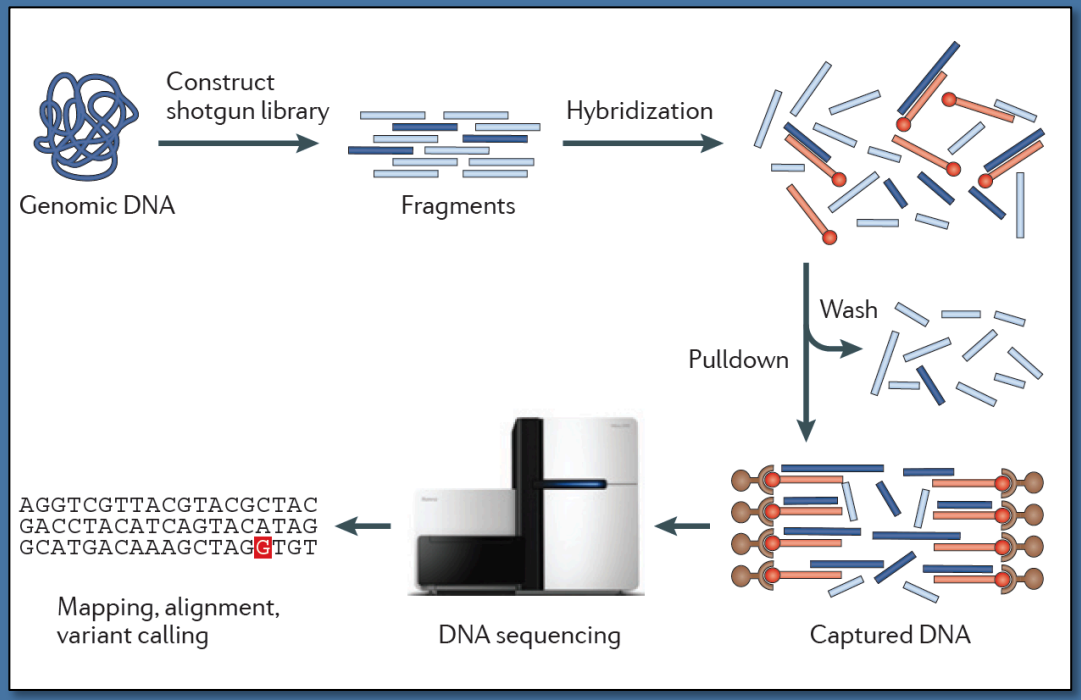


For complex polygenic disease, working assumption that the *pathways* hit by different types of variant will be similar

Motivates strategies that look for convergence across this spectrum

DATA

Exome sequencing



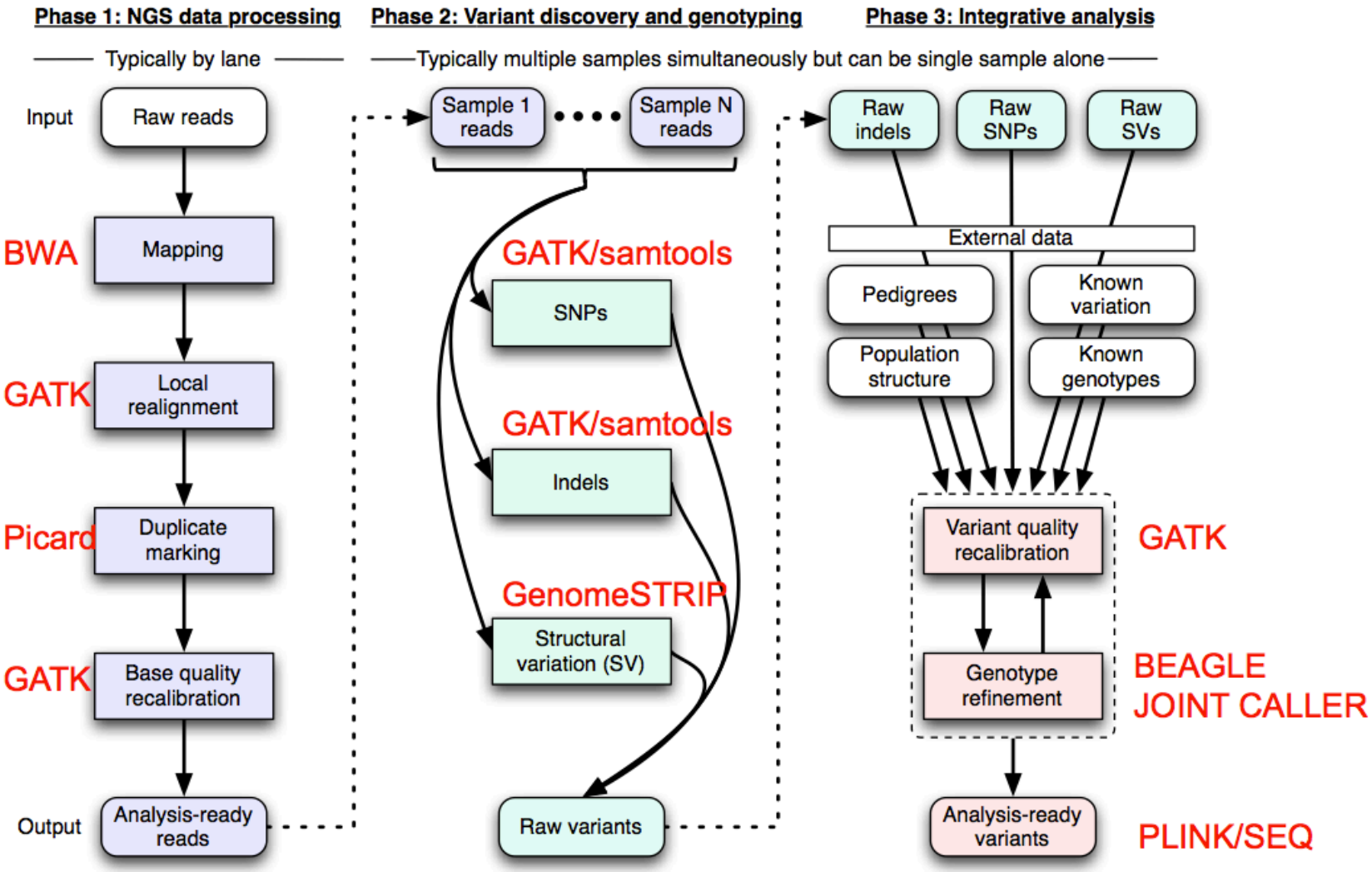
~200,000 "targets" (~exons)
- each ~150bp in length
- from ~20,000 genes
- ~30-50Mb of genomic sequence

Each targeted site, on average, covered by a hundred or more short reads, each ~70-100 bases

~20,000 variants per individual
- ~4M in whole-genome

Bamshad et al, *Nature Reviews Genetics*, 2011

The Picard/GATK NGS analysis pipeline



Exomes versus genomes

- Target ~1% of the genome (primarily coding exons in CCDS/RefSeq)
- ~10% of the cost of sequencing the whole genome
- Typically “deep coverage”, meaning high probability of detecting even variants observed only once
- Pros/cons (versus whole-genome sequencing)
  - \+ Enriches for the regions of the genome most strongly associated with disease. Even for common disease, where many GWAS hits do not map to genes, the relative rate of hits in genes is still much greater.
  - \+ Allows function to be ascribed to variants (filtering for deleterious variants, etc)
  - \+ Any positive result is more likely to be readily interpretable
  - \+ Currently more affordable to apply to large samples
  - – Targeting procedure introduces extra costs, steps in the sequencing pipeline, and biases in coverage
  - – Expanding definition of “the exome” (regulatory regions, rare transcripts, ncRNAs, etc)

Deep versus low-pass sequencing

- For a fixed $ amount of sequencing, how should I distribute it among samples?

10 individuals Mean 100x coverage Most sites >30x coverage	100 individuals Mean 10x coverage Many sites 0-2x coverage
Much better detection of singletons (inc. *de novo* mutations) and very rare variants, but in a smaller pool of variation	For low to moderate frequency variants, will be more powerful to *detect* variants: a less accurate sampling of a larger pool of variation
Greater depth allows other analysis, e.g. a) read depth analysis to detect CNVs b) better ability to QC/filter out bad variants	But can take advantage of the fact that reads at nearby sites are often informative due to local LD. **Imputation** will often be able to infer individuals' genotypes even at sites with, e.g. <<10x coverage

Exome sequencing → Exome chip

http://genome.sph.umich.edu/wiki/Exome_Chip_Design

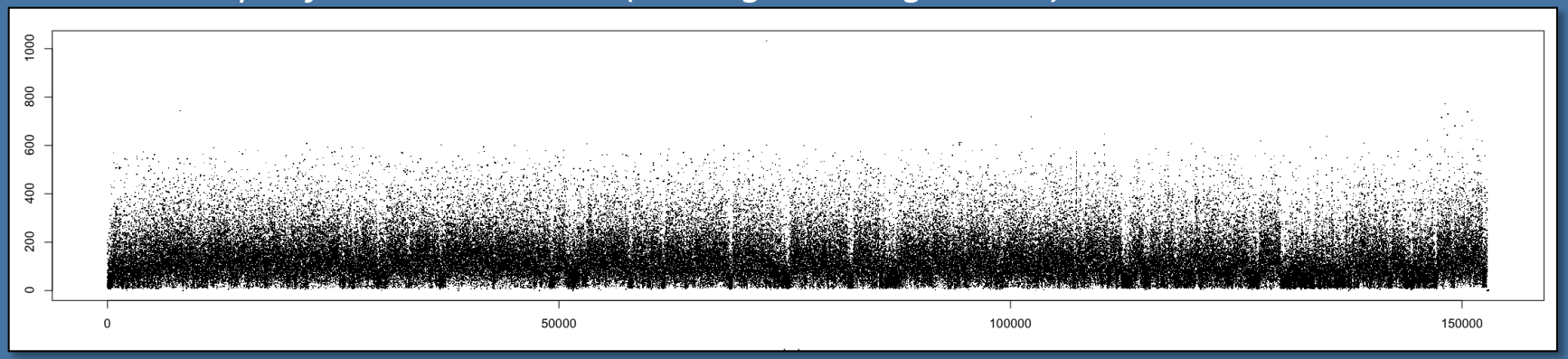
- Genotyping using microarrays is still cheaper and more accurate than exome-sequencing
- A very large proportion of all low frequency (e.g. >0.5%) coding variation will already have been observed in the 10,000+ exomes collectively sequenced at various centers
- Consortium to select a panel of these SNPs and manufacture an Illumina array at a reasonable price point, to enable testing in very large cohorts

What to expect from exome sequencing

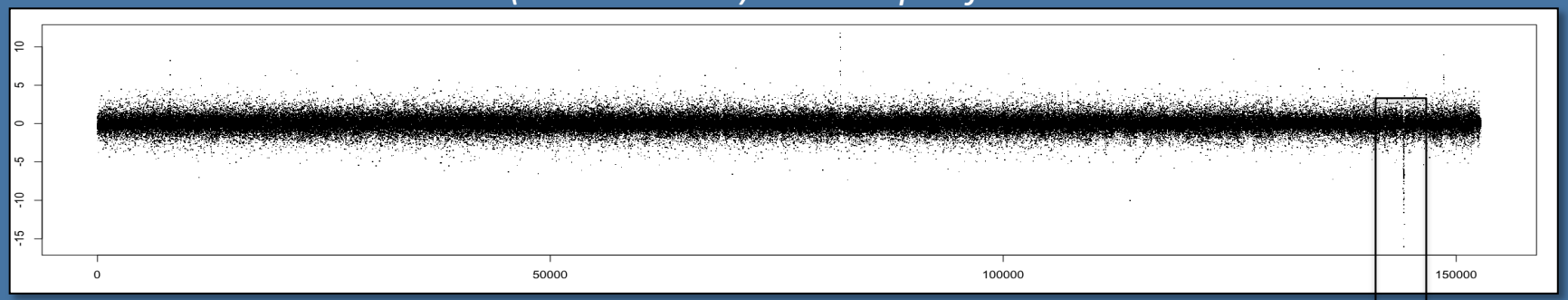
- From 1 individual ("case" or "control")
  - ~15,000 – 20,000 variant sites
  - ~10,000 of these nonsynonymous (of which 200-300 will be novel)
  - ~100 nonsense mutations (of which ~10 will be novel)
  - Vast majority of sites are **common and known** (in dbSNP) : over 95%

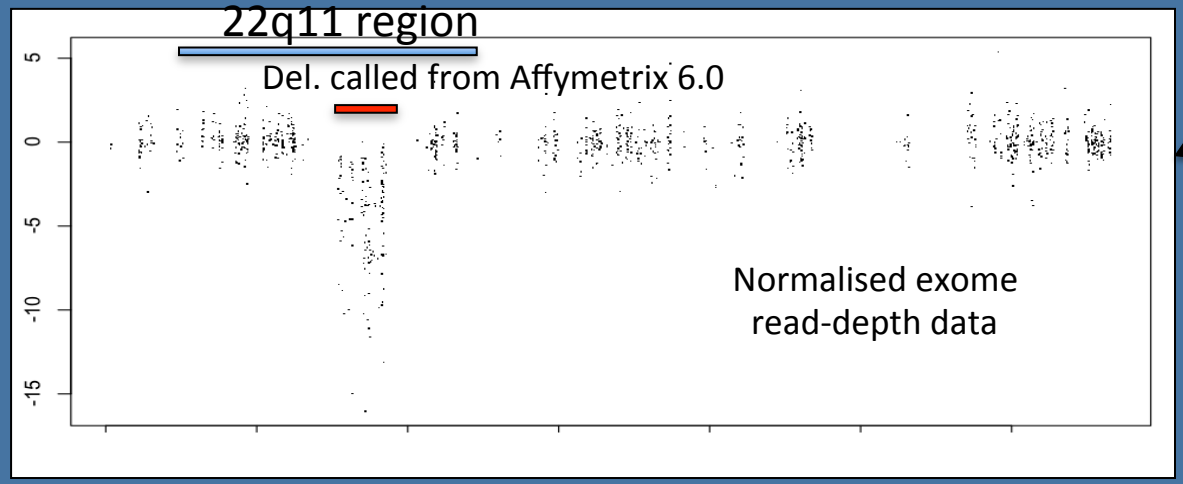
- From 5000 individuals
  - ~15,000+ gene-disruptive mutations (nonsense, splice, frameshift), of which most are novel
  - ~300,000 missense mutations (~100,000 – 150,000 of which are "damaging")
  - ~200,000 silent mutations
  - ~50% of all sites observed only once in the sample ("singletons")
  - Majority of variants **very rare and novel** (not in dbSNP)

Raw read-depth for one individual (all targets along exome)



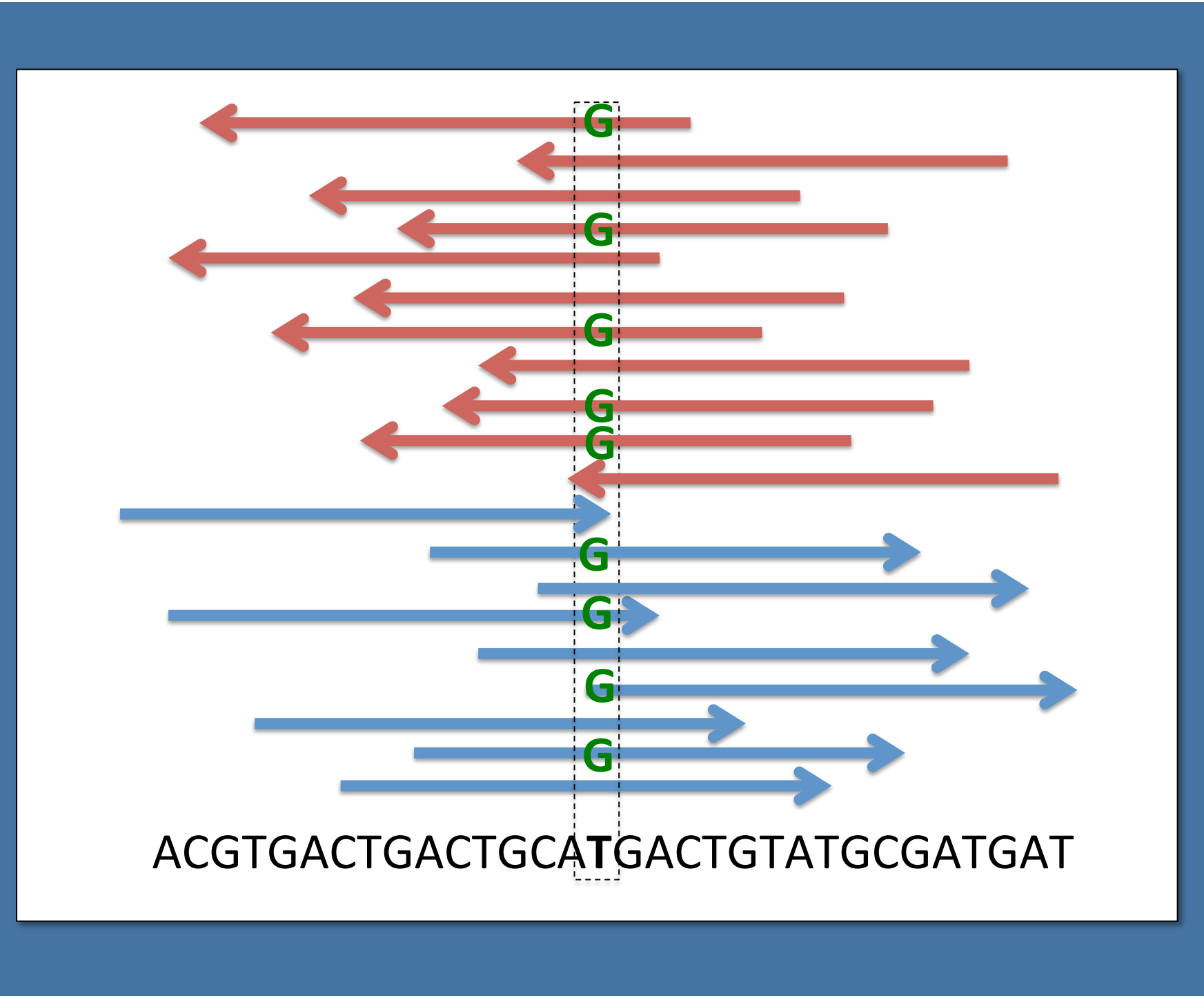
Normalised and de-trended (SVD method) read-depth for same individual



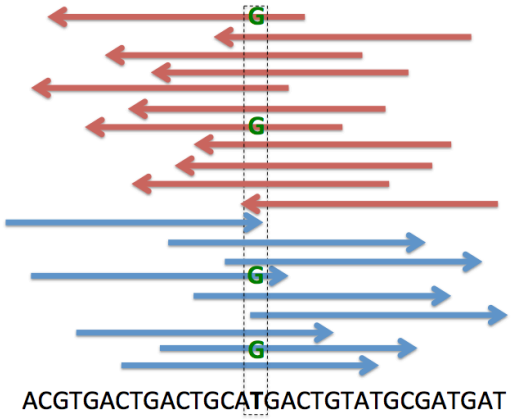


Calling CNVs from exome data

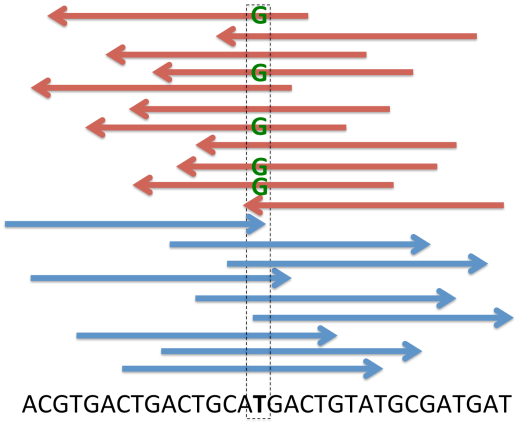
XHMM: Fromer et al (AJHG. 2012)


G
G
G
G
G
G
G
G
G
ACGTGACTGACTGCATGACTGTATGCGATGAT

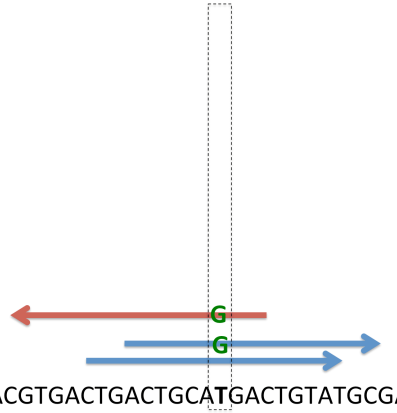
Common diagnostics of noise and bias in variant calling



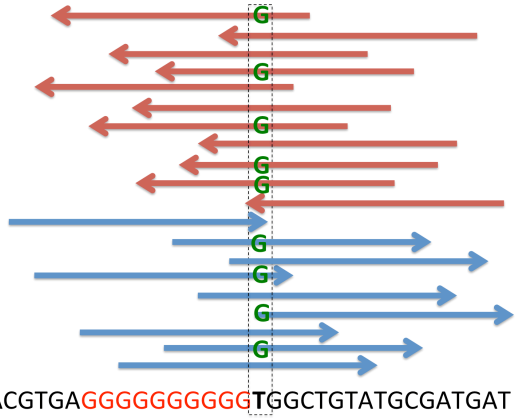
Allele balance, the ratio of reference : non-reference reads



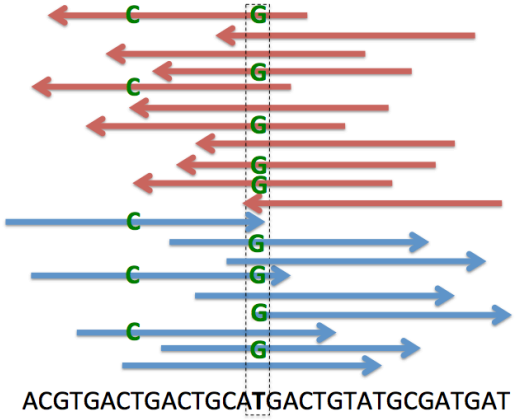
Strand bias, non-reference reads preferentially +ve or –ve



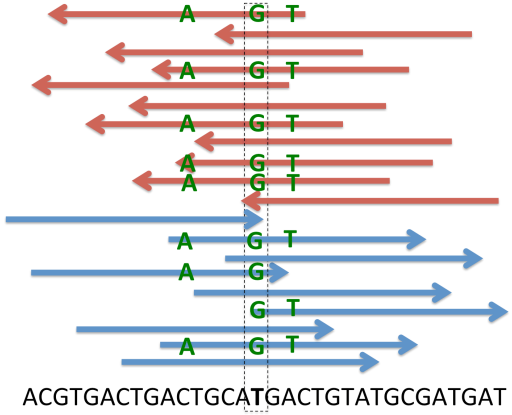
Low coverage / read-depth



Homopolymer runs, difficult sequence context



Reads are inconsistent with two distinct haplotypes

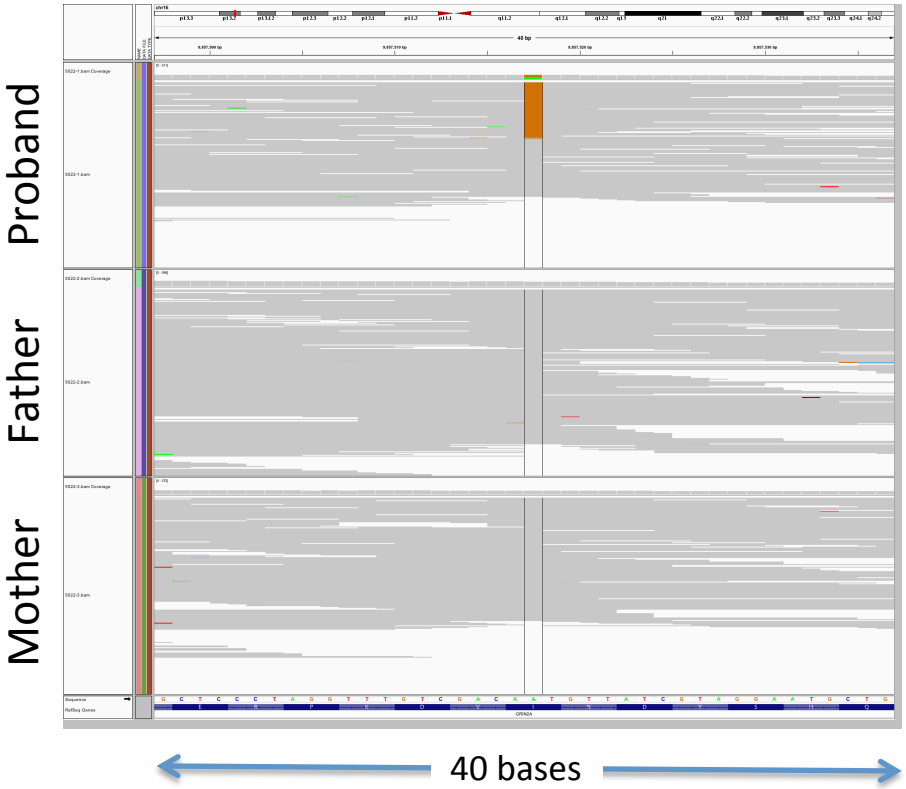


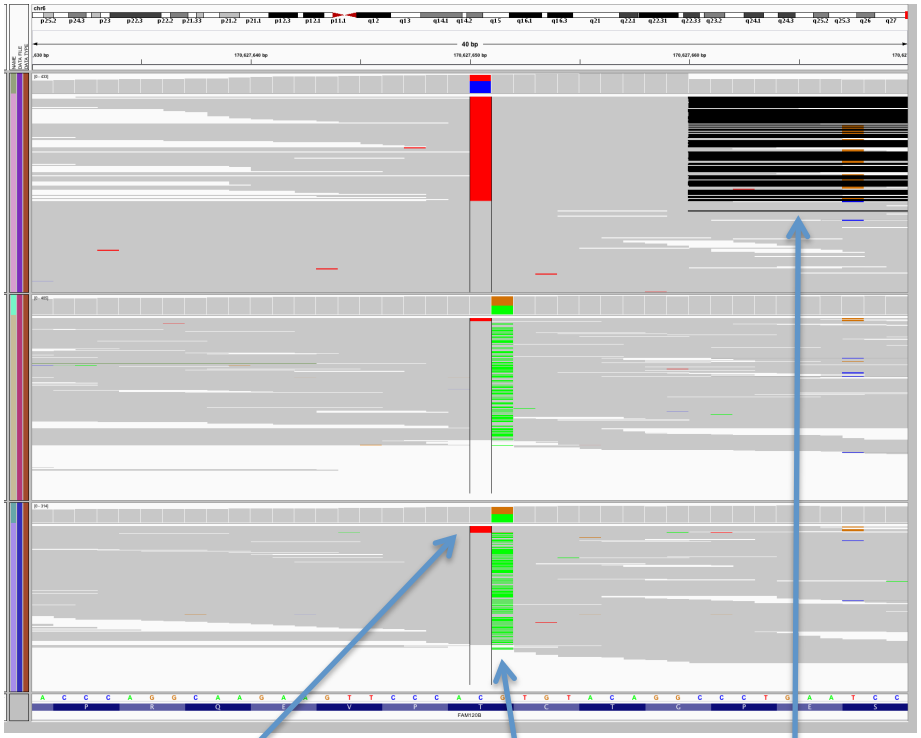
Clusters of SNPs, indicative of mis-mapping reads (e.g. due to cryptic indels, or other region of high homology)

Reflect often unknown, complex confounding influence of factors in DNA preparation, exome capture technology, alignment, SNP calling algorithm or QC filters:
means that directly combining data across different studies is difficult.

A validated *de novo* mutation…

… and one that did not





Parents also have some (red) non-reference reads

Both parents heterozygous for adjacent base

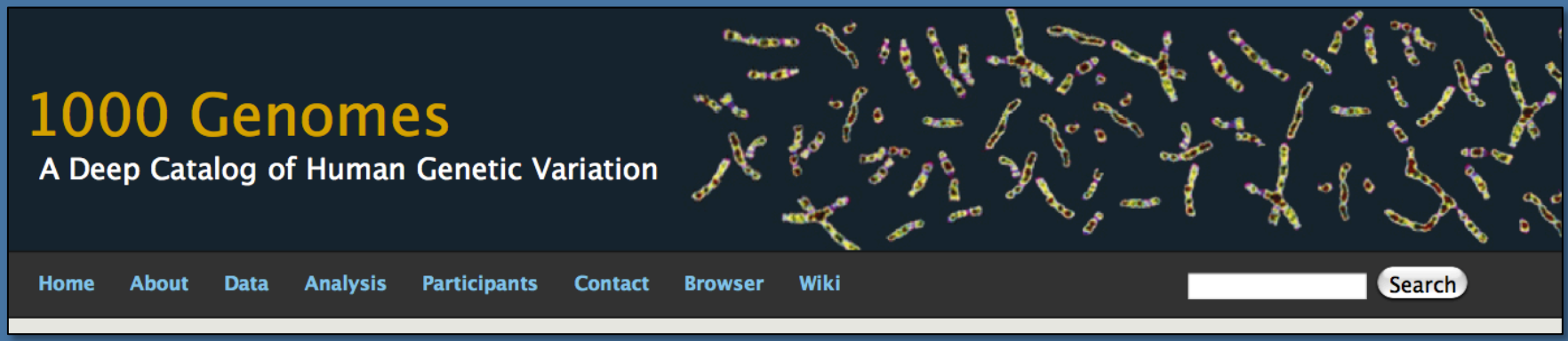
Proband also has apparent de novo deletion nearby (black)

Gray lines = 10s-100s of reads with reference allele piled up across region

Colored : reads containing a non-reference site

IGV : Integrative Genome Viewer

Data to play with: 1000 Genomes



A great resource for publicly available NGS data.

Both read-level data (BAM) and called variant/genotype datasets (VCF) available.

Whole genome and whole-exome.

VCF: An extensible text-based format for representing variant and genotype information and meta-information.

Tools to work with VCFs: PLINK/Seq, vcftools, vtools, others

- What does this genotype mean?

GT : AD :DP :GQ: PL

0/0:366,11:200:99:0,600,5980

PL={0/0 , 0/1 , 1/1 }
={REF , HET , HOM }

- GT hard genotype call
- AD and DP read-depth information
- GQ quality score
- PL is (phred-scaled) *genotype-likelihoods* (soft-calls)
  - Above, the heterozygote is 10^{-60} as likely as the reference homozygote

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

$$Q = -10 \log_{10} P \qquad P = 10^{\frac{-Q}{10}}$$

- A less compelling genotype call

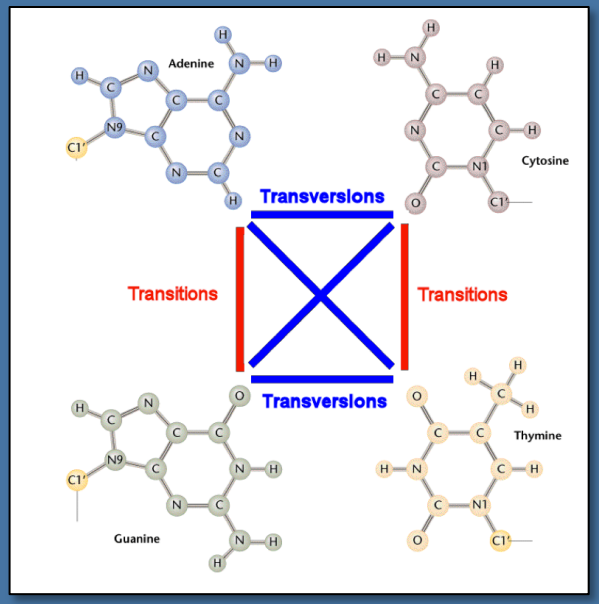
```
GT  :  AD  :DP  :GQ:  PL
0/1:6,1  :7   :9  :9,0,187
```

```
PL={0/0  ,  0/1  ,  1/1 }
  ={REF  ,  HET  ,  HOM }
```

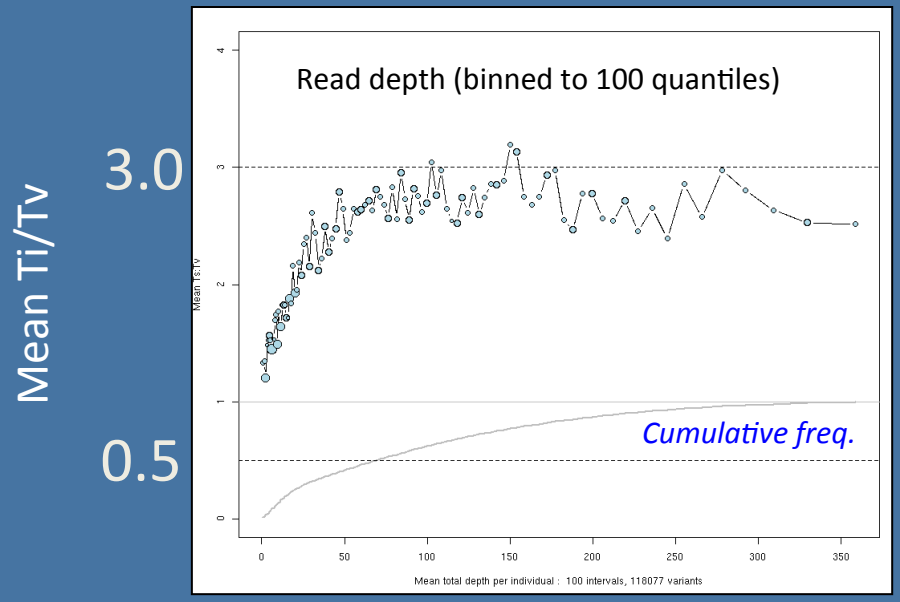
- Heterozygote is most likely call
  - the reference homozygote has likelihood of $10^{-0.9}$ = 0.12 compared to heterozygote
- But, not a high confidence call
  - Based on a relatively low number of reads (7)
  - Ratio of reference to alternate reads skewed from 50:50 (6:1)
  - Rule of thumb in deep exome data: PL > 20 or 30 and DP > 10 defines “high confidence”

Transition/transversion ratio as a figure of merit

two-ring **purines** one-ring **pyrimidines**



http://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html



Twice as many possible transversions as transitions: in practice transitions (A/G, C/T) more common

Expected Ti/Tv (or Ts/Tv): **Errors:** ~0.5 **Real exome variants:** ~3.0

Study how Ti/Tv (or other metrics such as dbSNP%) vary with technical attributes (e.g. read depth)

This type of process formalized in GATK's variant quality score recalibration (VQSR) procedure

Functional annotation of variants

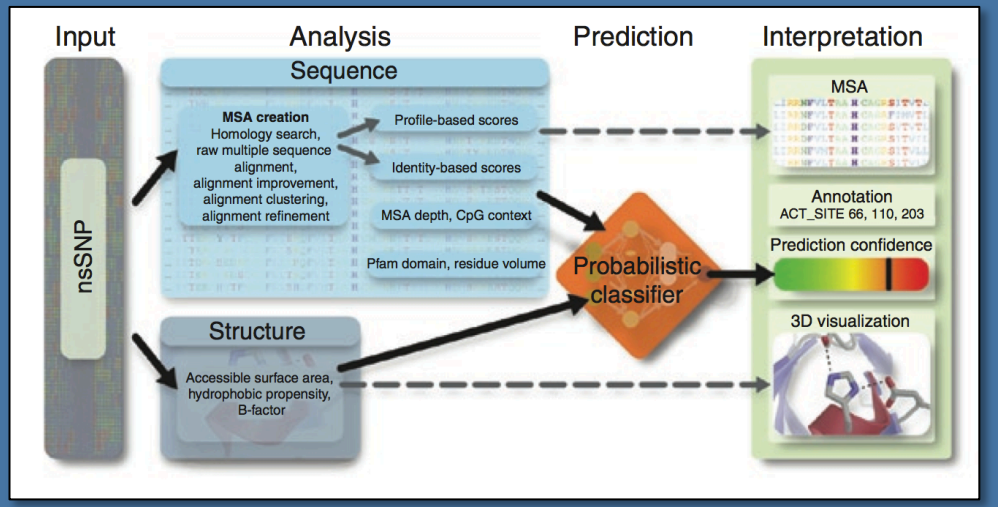
Type of variation in amino-acid sequence

Missenses not all equally likely to have an impactful change on protein

AMINO ACID TABLE

First Position	Second Position: U		Second Position: C		Second Position: A		Second Position: G		Third Position
	code	Amio Acid	code	Amio Acid	code	Amio Acid	code	Amio Acid	
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

PolyPhen2: predicting the damaging effects of missense mutations



Adzhubei *et al*. Nature Methods (2010).

1) Truncates protein (nonsense)
2) Changes protein (missense)
3) Doesn't change protein (silent)

In silico prediction of "damaging" or "deleterious" mutation

Functional annotation of variants

In large samples, allele frequency can be used to evaluate prediction methods:
natural selection implies that more damaging mutations should on average be rarer

Type	MAF	% singletons
Intronic (off-target)	0.066	25
Silent	0.047	40
Missense	0.025	51
Benign	0.038	45
Possibly damaging	0.020	52
Probably damaging	0.010	59
Essential splice site	0.016	54
Nonsense	0.010	66

(based on RefSeq transcripts and hg19; missense ranking w/ PolyPhen2)

Different prediction methods diverge…

Mean correlation between raw scores only 0.35 (median 0.11)

	Gerp(NS)	Gerp(RS)	PhyloP	PPH2(HumDiv)	PPH2(HumVar)	SIFT	LRT
Gerp(RS)	0.351						
PhyloP	0.346	0.903					
PPH2(HumDiv)	0.076	0.418	0.465				
PPH2(HumVar)	0.094	0.405	0.455	0.911			
SIFT	0.025	0.244	0.263	0.456	0.441		
LRT	0.152	0.386	0.376	0.276	0.287	0.150	
MutationTaster	0.275	0.425	0.448	0.392	0.423	0.244	0.350

Scores calculated for set of variants from exome sequencing in ~5000 Swedish individuals

…but appear to have (independent) information

Class	Propotrion of missenses that are singleton
Mean missense	**0.480**
PPH2 HumDiv	0.622
PPH2 HumVar	0.634
LRT	0.624
SIFT	0.613
MT	0.626
Any	0.604
All	**0.664**

Singleton status calculated in same Swedish sample

Annotation issues

- Which transcripts to use?
  - Trade-offs using more or less restrictive definitions of the CDS (CCDS, RefSeq, ENSEMBL, etc)
  - Multiple transcripts v.s. aggregated “gene" v.s. a single canonical transcript per gene
  - Prioritizing transcripts based on expression in tissue of interest and/or RNA-seq data, etc
- Catching likely errors
  - “Rogue transcripts”, e.g. many stop codons in reference, CDS not mod 3, invalid start codon
- Complex variants
  - Multi-nucleotide mutations often misannotated
  - Edge cases: a single base insertion at the intron/exon boundary: splice or frameshift?
- Weighting within existing functional classes
  - Optimal use of *in silico* prediction tools for deleteriousness of missense variants
- Noncoding variants
  - Variants in ncRNAs and other functional elements in exome-seq: miRNAs, UTRs, etc

DESIGNS

Study designs

- Mendelian disease and "filtering" approaches
  - Assumes very rare disease, very highly penetrant mutation and low locus heterogeneity
- Multiplex families to ascertain "familial"(*) cases
  - Assumes a private mutation of large effect largely accounts for disease in the family
  - Assumes that co-segregation will be informative
- Trio studies of "sporadic" cases(*)
  - Focus on *de novo* rather than inherited mutation
  - Particularly suitable for early-onset diseases that reduce reproductive success
- Population-based case/control studies
  - Less efficient to the extent that private/*de novo* mutations account for most disease risk
  - But a more general, potentially more scalable design (e.g. if families hard to collect)
  - Likely(?) better suited to tackle more heterogeneous & complex architectures

*Yang et al (2010) *Sporadic cases are the norm for complex disease*. *EJHG*.

Rare variant burden analysis

$P(D)$ — Prevalence of disease

$P(G_D)$ — Prior probability of carrying a disease *allele*

$P(D \mid G_D)$ — Penetrance of (an average) disease allele

Rare variant burden analysis

$P(D)$ — Prevalence of disease

$P(G_D)$ — Prior probability of carrying a disease *allele*

$P(D \mid G_D)$ — Penetrance of (an average) disease allele

$$P(G_D \mid D) = \frac{P(D \mid G_D)\, P(G_D)}{P(D)}$$

Allele frequency in cases

$P(G_D \mid \bar{D})$ — Allele frequency in controls

Mendelian disease

$P(D)$ **V. LOW** Prevalence of disease

$P(G_D)$ **V. LOW** Prior probability of carrying a disease *allele*

$P(D|G_D)$ **V. HIGH** Penetrance of (an average) disease allele

$$P(G_D|D) = \frac{P(D|G_D)\,P(G_D)}{P(D)}$$

V. HIGH · V. LOW · V. LOW

HIGH Allele frequency in cases

why simple filtering works

$P(G_D|\bar{D})$ **V. LOW** Allele frequency in controls

Filtering approaches and Mendelian disease

Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome

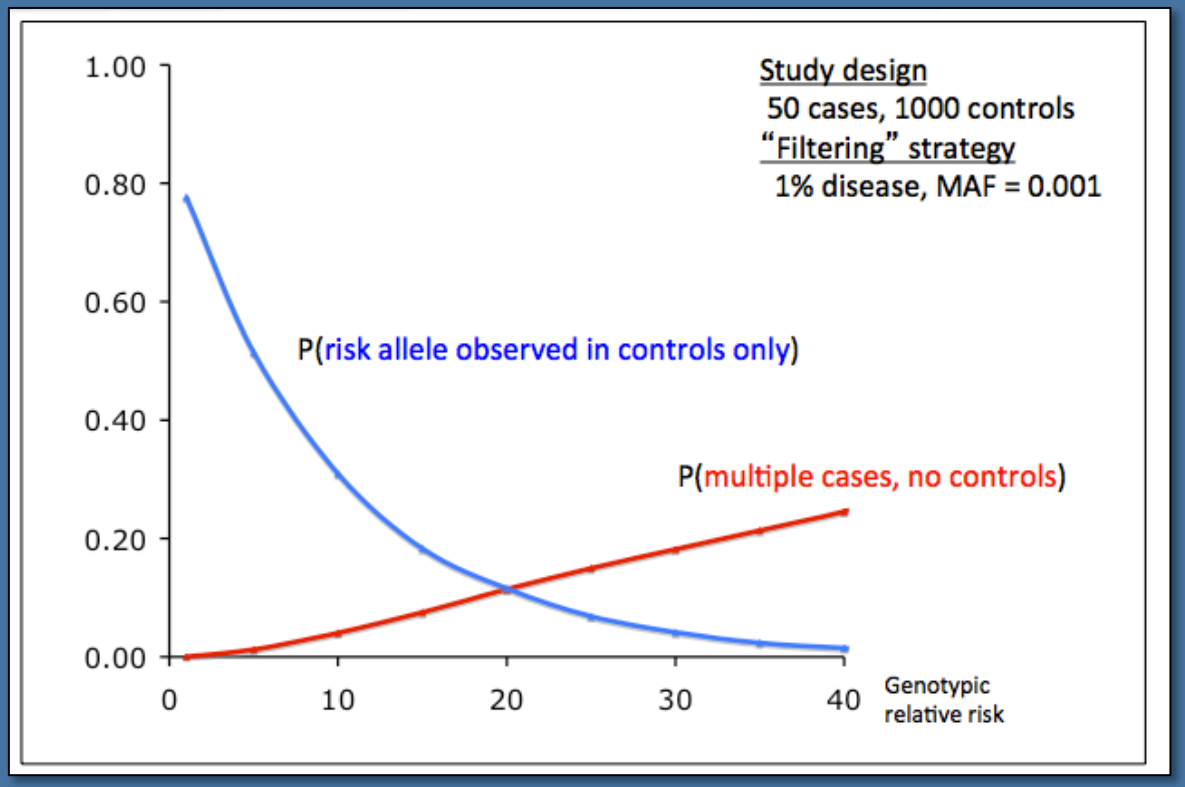
Nat Genet (2010)

Sarah B Ng[1,7], Abigail W Bigham[2,7], Kati J Buckingham[2], Mark C Hannibal[2,3], Margaret J McMillin[2], Heidi I Gildersleeve[2], Anita E Beck[2,3], Holly K Tabor[2,3], Gregory M Cooper[1], Heather C Mefford[2], Choli Lee[1], Emily H Turner[1], Joshua D Smith[1], Mark J Rieder[1], Koh-ichiro Yoshiura[4], Naomichi Matsumoto[5], Tohru Ohta[6], Norio Niikawa[6], Deborah A Nickerson[1], Michael J Bamshad[1–3] & Jay Shendure[1]

Table 1 Number of genes common to any subset of *x* affected individuals.

Subset analysis (any *x* of 10)	1	2	3	4	5	6	7	8	9	10
NS/SS/I	12,042	8,722	7,084	6,049	5,289	4,581	3,940	3,244	2,486	1,459
Not in dbSNP129 or 1000 Genomes	7,419	2,697	1,057	488	288	192	128	88	60	34
Not in control exomes	7,827	2,865	1,025	399	184	90	50	22	7	2
Not in either	6,935	2,227	701	242	104	44	16	6	3	1
Is loss-of-function (non-sense or frameshift indel)	753	49	7	3	2	2	1	0	0	0

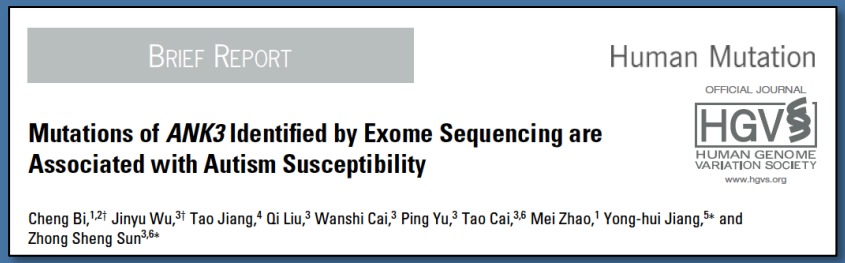
Much harder for complex, common disease: basic filtering approaches not a good strategy



One is still just as likely to observe this rare disease allele of 20-fold increase in risk in a large sample of (screened) controls, compared to observing it recurrently in 50 cases

Application of "filtering" to common disease

- 20 autism probands
- Detect "novel" variants
- Prioritize based on function/ gene
- Whatever is left is the "finding" (*ANK3*)
- Find other *ANK3* mutations in other individuals: "additional support"



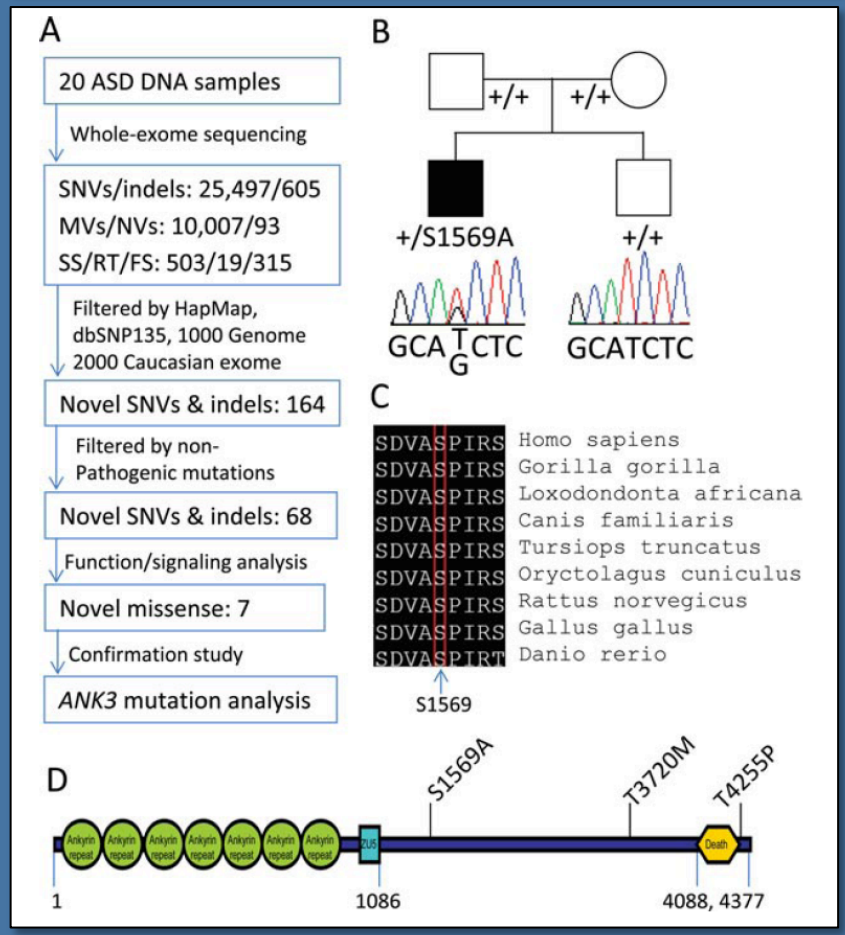
BRIEF REPORT

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Mutations of *ANK3* Identified by Exome Sequencing are Associated with Autism Susceptibility

Cheng Bi,[1,2†] Jinyu Wu,[3†] Tao Jiang,[4] Qi Liu,[3] Wanshi Cai,[3] Ping Yu,[3] Tao Cai,[3,6] Mei Zhao,[1] Yong-hui Jiang,[5*] and Zhong Sheng Sun[3,6*]



Novel *ANK3* mutations in healthy individuals

- Exome sequence data from ~2500 healthy Caucasian individuals
  - Part of the Swedish Schizophrenia Seqeuncing Study
- Screen for novelty against dbSNP (>50,000,000 variants)
- 95 novel mutations detected in controls
  - 1 nonsense
  - 32 missense
  - 24 rated as “damaging” by PolyPhen2
- In other words, if you look at enough samples and/or genes, it isn’t hard to pull out “interesting mutations”

Familial co-transmission to filter variant lists

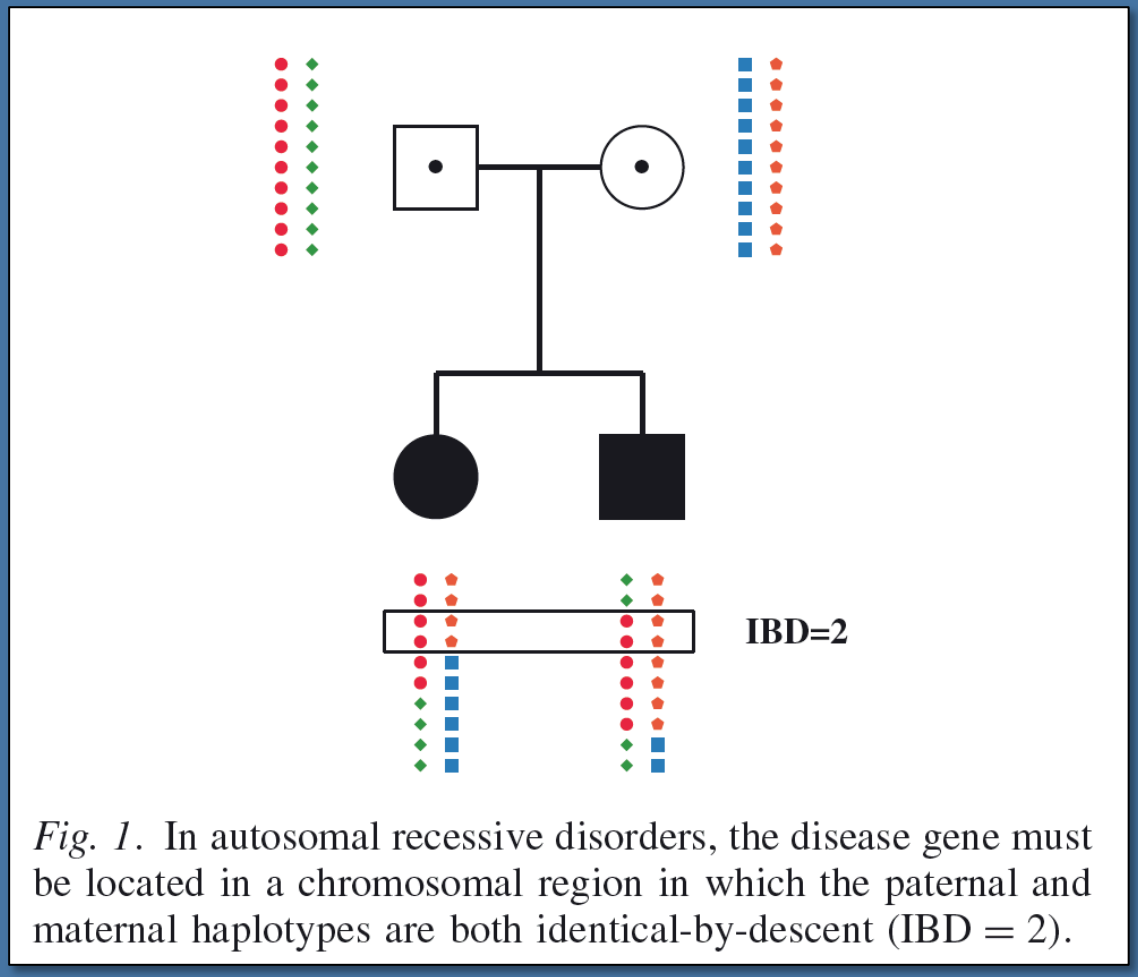
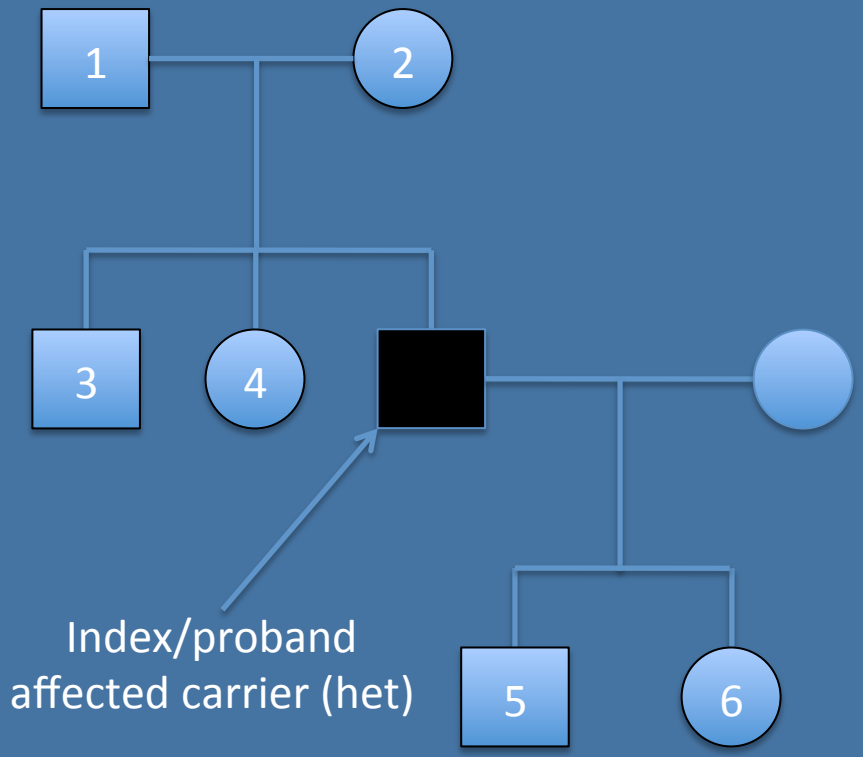


Fig. 1. In autosomal recessive disorders, the disease gene must be located in a chromosomal region in which the paternal and maternal haplotypes are both identical-by-descent (IBD = 2).

Robinson et al, 2011

Simulation: using co-segregation in common disease



Family: proband and 6 1st-degree relatives

Ascertainment: require at least *X* of 6 to be affected, where *X* = 1,2 or 3.

Disease: 1% prevalence, h^2=.6, c^2=.05

Genetic model:
MAF: 1/100, 1/1,000 and 1/10,000
GRRs: range for each MAF (next table)
Dominant gene-action

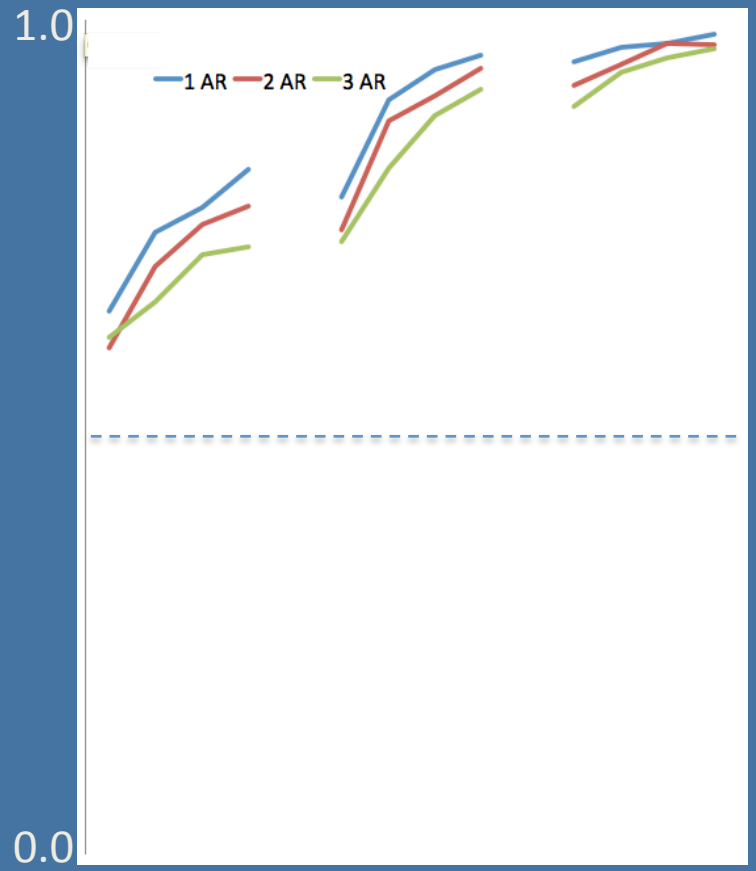
Questions:

1) How likely an *affected* relative shares the index's rare allele?
2) How likely an *unaffected* relative shares it?
3) Expectation that a 1st-degree relative is "consistent"?
4) Probability of complete co-segregation of allele and disease in family?

Power/sample size for a standard, population-based case/control study

MAF	GRR	*N* for 80% power at alpha = <0.01	*N* for 80% power at alpha = <5e-8	Penetrance
1/100	2	1802	6110	0.020
	3	610	2070	0.029
	4	344	1168	0.038
	5	236	800	0.046
1/1,000	5	2153	7303	0.050
	10	784	2661	0.098
	20	340	1156	0.193
	30	219	743	0.284
1/10,000	20	3309	11223	0.199
	40	1537	5213	0.397
	60	1002	2643	0.593
	80	744	2524	0.788

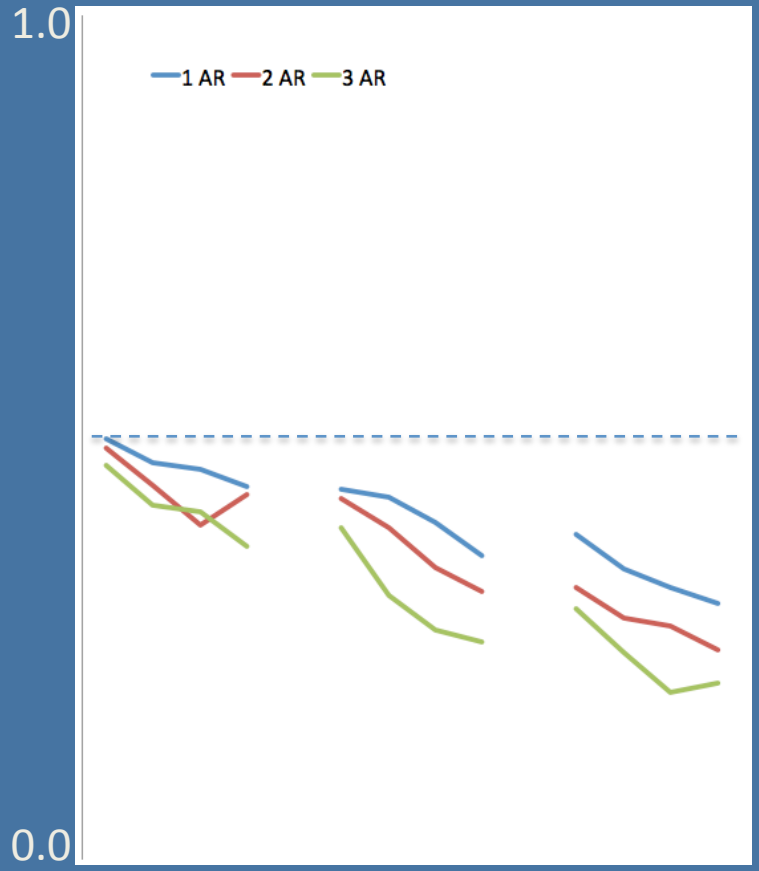
P(affected relative shares allele)



MAF: 1/100 1/1,000 1/10,000

OR: 2 – 3 5 – 30 20 – 80

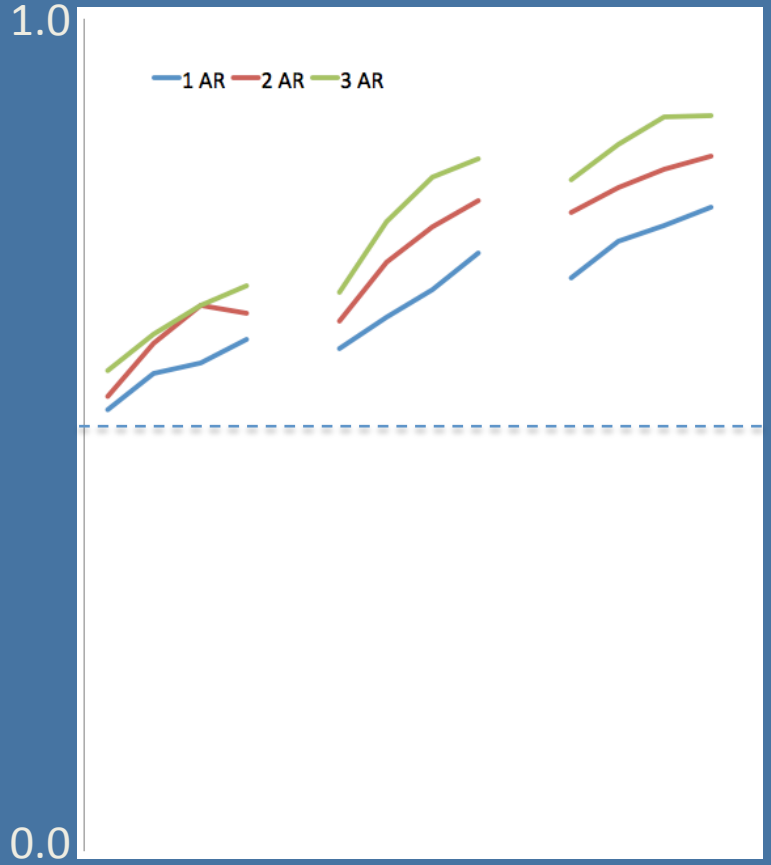
P(unaffected relative shares allele)



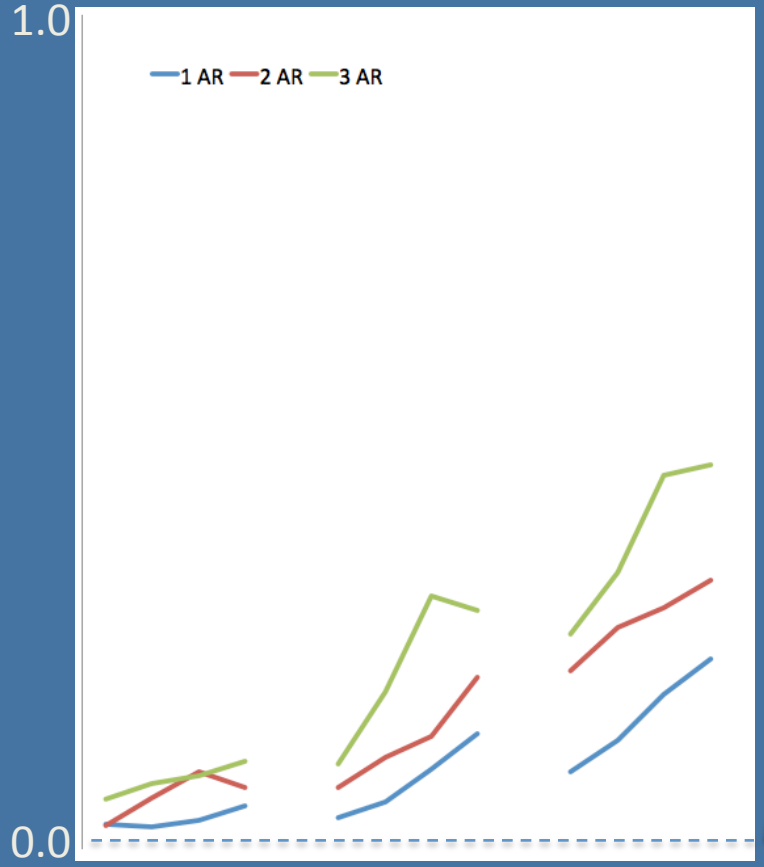
1/100 1/1,000 1/10,000

2 – 3 5 – 30 20 – 80

P(any one 1st degree relative co-segregates)



P(whole family co-segregates)



~1% under the null

MAF:	1/100	1/1,000	1/10,000	1/100	1/1,000	1/10,000
OR:	2 – 3	5 – 30	20 – 80	2 – 3	5 – 30	20 – 80

1) Under all models here, it is more likely than not that the true disease allele will ***not*** segregate with disease
2) In contrast, 1% of *all* of the proband's rare null alleles would be expected to perfectly co-segregate by chance
3) Naturally, larger families and/or more "Mendelian" alleles would change the balance
4) (But unlikely that different families will segregate at the same locus given polygenicity, linkage findings…)

1/1000 20-fold variant example, for a disease with 1/100 prevalence

	Penetrance, P(disease | genotype)
Reference	0.0096
Heterozygote	0.193
Homozygote	0.193

But only 4% of cases would be expected to carry such an allele

	P(genotype | affected)
Reference	0.962
Heterozygote	**0.038**
Homozygote	0.000

Even under the most optimistic circumstances, for common diseases:

1) the vast majority of patients will not carry the risk allele
2) the majority of carriers will not be affected (penetrance < 50%)

Rare variants & complex disease

$P(D)$ **MED/LOW** Prevalence of disease

$P(G_D)$ **LOW** Prior probability of carrying a disease *allele*

$P(D|G_D)$ **MED/LOW** Penetrance of (an average) disease allele

$$P(G_D|D) = \frac{\overset{\text{MED/LOW}}{P(D|G_D)}\ \overset{\text{LOW}}{P(G_D)}}{\underset{\text{MED/LOW}}{P(D)}}$$

LOW Allele frequency in cases

under-powered tests of association

$P(G_D|\bar{D})$ **LOW** Allele frequency in controls

Ways to improve power beyond ⇧sample *N*

$P(D)$ ce of disease

$P(G_D)$ disease *allele*

$P(D|G_$ disease allele

$P(G_D|D$ ency in cases

sociation

$P(G_D|\overline{D})$ LOW Allele frequency in controls

(In loose terms,) ideas to drive up $P(G_D|D) - P(G_D|\overline{D})$

Ascertain on family history: increase $P(G_D)$

Aggregate tests ("super-alleles") : increase $P(G_D)$

Subsets of genes/variants : decrease $P(G_D)$ but increase $P(D|G_D)$

Extreme/subtypes of disease: decrease $P(D)$, increase $P(D|G_D)$

"Ultra-healthy" controls : reduce $P(G_D|\overline{D})$

Etc.

(Of course, not all equally feasible or effective...)

ANALYSIS

The challenge of interpreting rare-variant studies: a lot of the data will look either like…

	Alternate allele	Reference allele
Case	1	999
Control	0	1000

… or …

	Alternate allele	Reference allele
Case	0	1000
Control	1	999

Gene based tests

Group variants within a region and test for aggregate distributional differences between cases and controls (or with a quantitative trait).

For exome studies, “the gene” is the natural unit of grouping.

<u>A large number of tests developed in the last couple of years. Main differences:</u>

1) Are all variants assumed to have a similar magnitude of effect, or does the test allow for differential weighting, e.g. rarer variants can have larger effects?

2) Are all the variants assumed to have similar direction of effect, or does the test allow for a mixture of risk and protective variants in the same region?

3) Practically, can covariates be included? Application to quantitative traits? Reliance on permutation versus *accurate* asymptotic statistics.

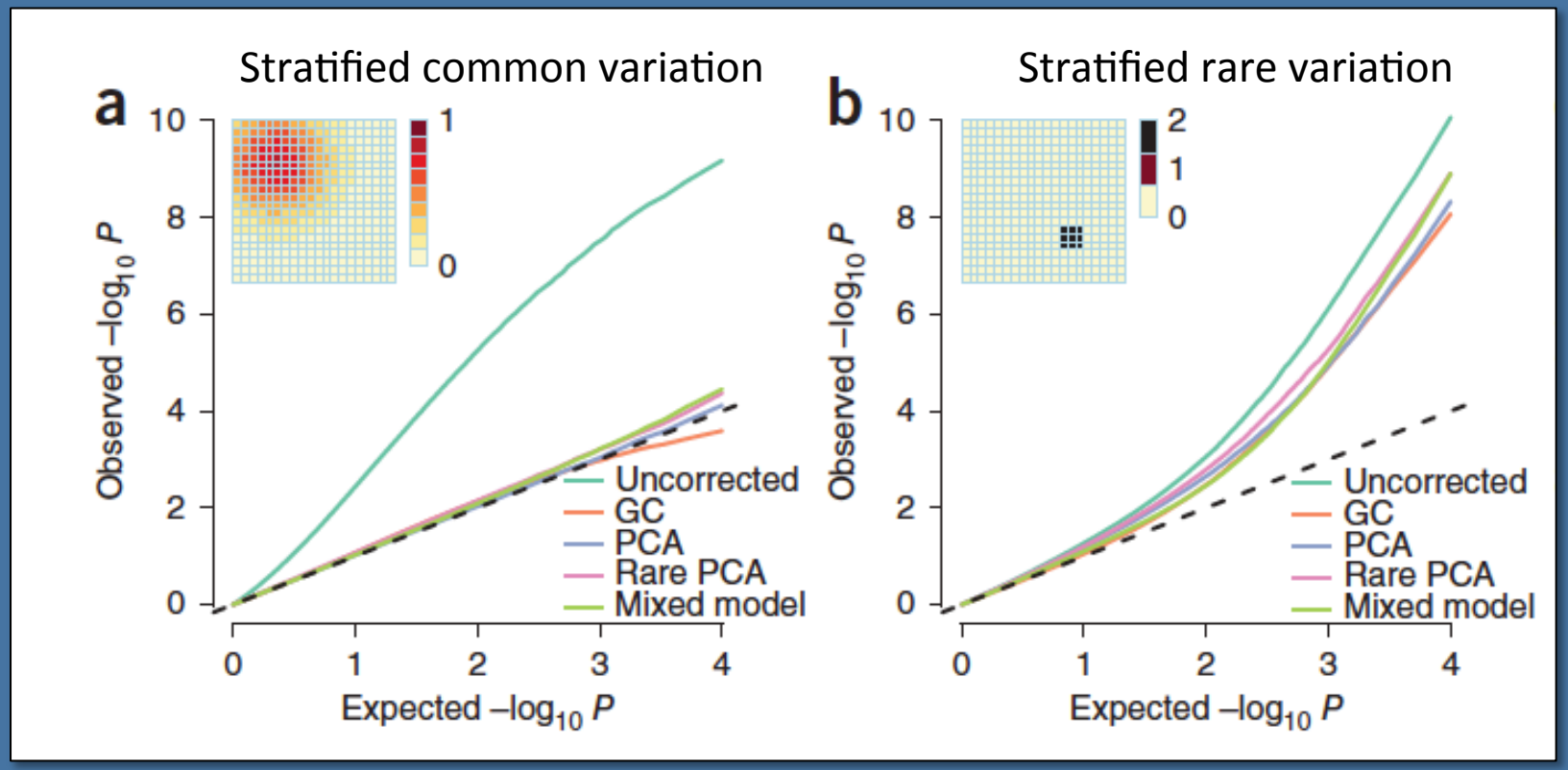
Test	Description
Cohort allelic sums test (CAST), Morgenthaler & Thilly (2007)	Carrier rate of 1+ rare allele compared between cases and controls
Burden	Count of rare alleles compared between cases and controls
Burden of case-specific variants	Burden of cases-only variants, assessed by permutation
CMC (Li & Leal, 2008)	Combines collapsed rare variant counts with more common alleles
Madsen & Browning (2009)	Up-weight rarer variants in a burden test analytically
Variable-threshold (Price et al, 2010)	Optimal definition of “rare” found empirically from the data, adjusted by permutation
C-Alpha (Neale et al, 2010)	Frames a 2-sided test, allowing a mixture of risk and protective variants
SKAT (Wu et al, 2011)	Generalization of C-Alpha based on kernel machine regression
(many other variations upon these themes)	….

Burden analyses in population-based exome studies

- To a large extent, both GWAS and CNV studies of psychiatric disease initially relied on demonstrating genome-wide burden effects
  - Genome-wide burden of rare microdeletions and duplications in AUT and SCZ
  - Polygenic analysis of GWAS data in SCZ and BIP
- Following this, seems likely we'll end up following a similar path for sequencing studies. However, the high baseline levels of rare SNVs means that (unlike for rare, large CNVs) we would not expect a simple "exome-wide increased burden of deleterious SNVs" analysis to yield much.
- Need to focus either on specific **classes of variant**, or **classes of gene** in which to frame burden questions, i.e. that we believe are *a priori* more likely.
- Demonstrated increased exome-wide burden of particular classes of SNV
  - Gene-disruptive *de novo* mutation in autism
  - Rare-recessive loss-of-function mutations also in autism
- Similar logic to stratify by *class of gene* (as well as SNV type), based on candidates and/or prior genetic literature
  - Do genes flagged by *de novo* or CNV or GWAS studies show an increased burden in cases?
  - Do genes involved in candidate "pathways" show an increased burden?

Population stratification

Under geographically realistic models of gene-flow, methods that successfully correct population of common variants do not necessarily work well for rare variants



Mathieson & McVean (2012) Nat Genet.

Population homogeneity and rare variants

- In gene-based tests, one often uses the *sample frequency* to define which variants are "rare"
  - sample frequency as an estimate of *population frequency* as a proxy for *causal potential*
- Potential problems when dealing with heterogeneous samples, even if false positives are controlled
  - "Singleton" allele in a large, homogeneous sample likely has low population frequency
  - "Singleton" carried by the one Asian in an otherwise Caucasian sample likely won't
- For rare variant tests, a single individual can often have a strong leverage on the sample test statistic
  - An 8/0 becoming an 8/1 can make a result much less impressive, even in large sample
  - For common variants, or rare variant tests across many genes, harder for a small proportion of individuals to have a great influence

Heterogeneity and power: simulation

- Three gene models with equivalent power in a standard population-based case/control in population "A"
  - 1000 case/control pairs; each has ~77% power for type I error 0.001 in simple burden test

Population "A" (N=1,000 & 1,000)

Genic/aggregate allele frequency	Mean genotypic relative risk
1 / 10,000	70
1 / 1,000	10
1 / 5	1.5

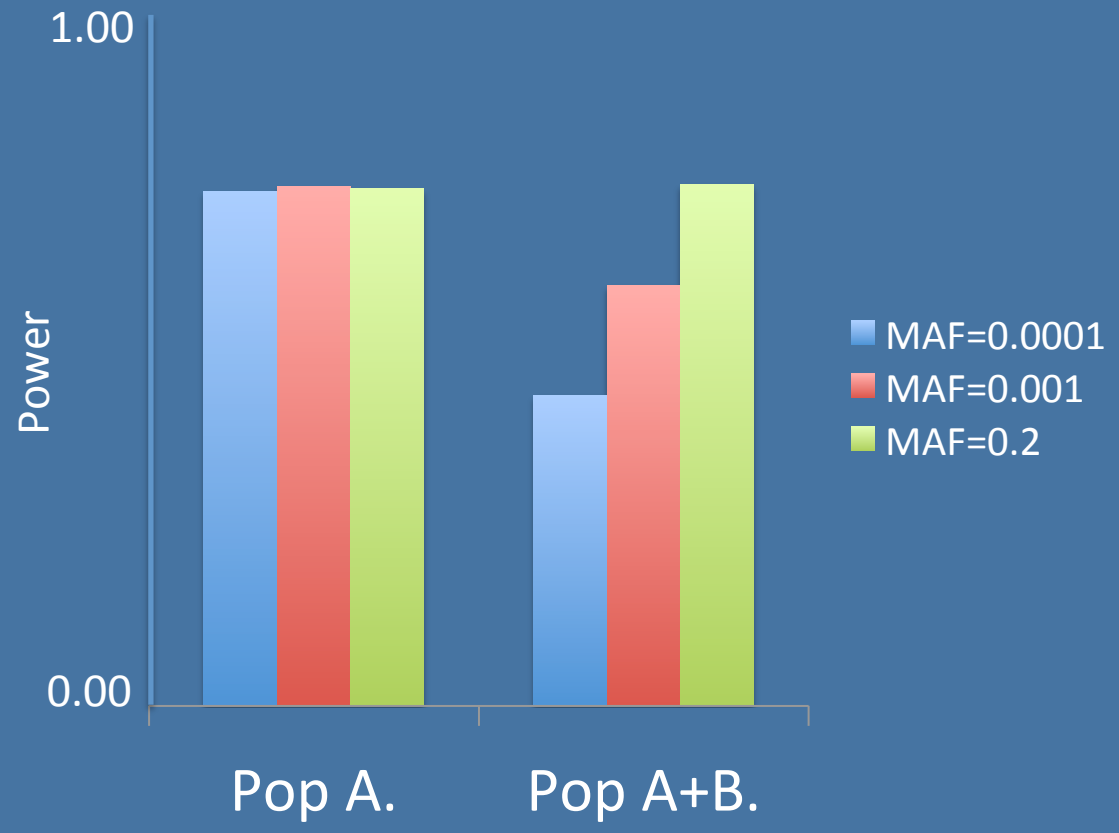
- Second population, different properties
  - i.e. null hypothesis is true

Population "B" (N=100 & 100)

Genic/aggregate allele frequency	Mean genotypic relative risk
1 / 100	1.0

What happens in a combined analysis?

Here, rare variants are “more sensitive” to inclusion of unassociated alleles from additional samples and different populations



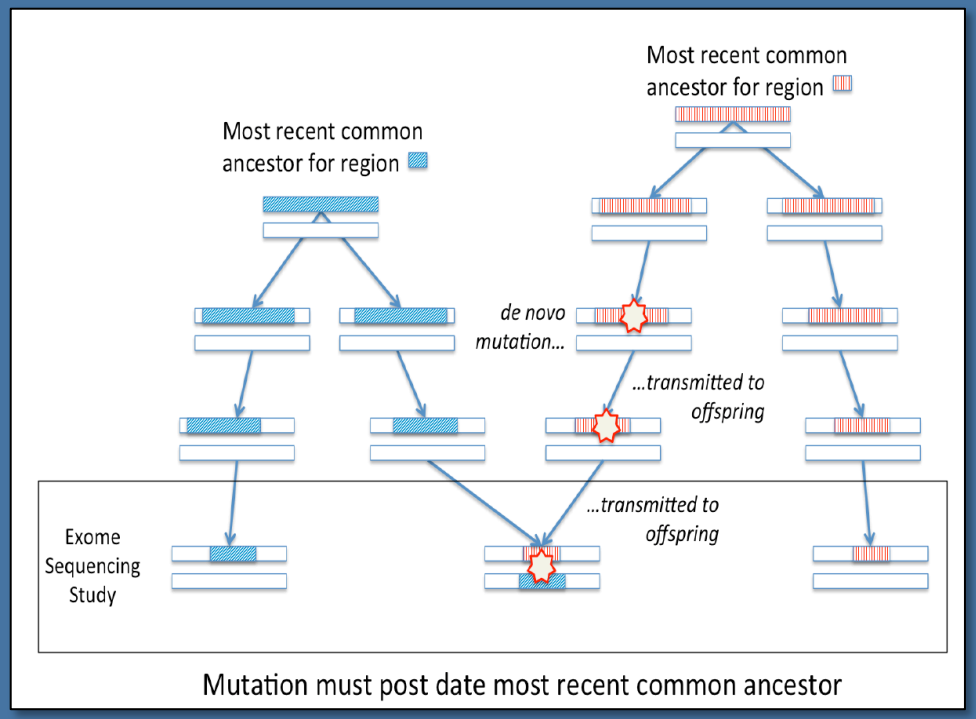
Note: A+B test conditions on population – i.e., type I error is still controlled, but power is reduced

Analysis questions for *de novo* studies

- Is the **rate** of mutations higher than { in controls | expected }?
- Is the **ratio** of { nonsense } to { missense } higher?
- Is any particular gene **recurrently** hit, more than expected by chance (given total # of mutations, coverage of exome, gene size, mutation rate)?
- Are genes with *de novos* more **closely related** (in terms of functional class, or position within PPI or co-expression network, etc)?

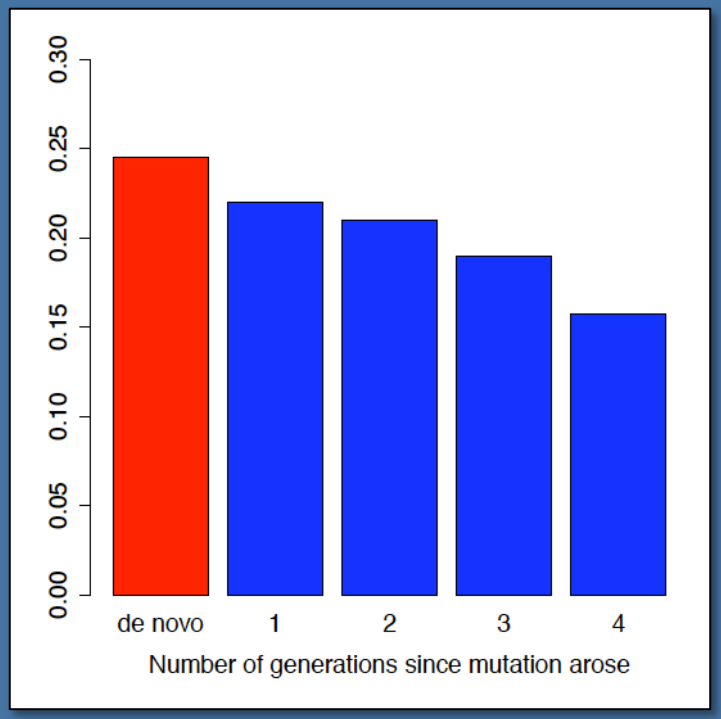
Finding “recent” mutations in populations

Using patterns of shared ancestry between seemingly unrelated individuals to flag “recent” mutations



Compared to all singleton mutations, those flagged as “recent” are more likely to be novel and to be nonsense mutations (as are *de novo* mutations).
[data from Swedish Sequencing Study]

“Recently *de novo*” mutations transmitted to cases: unless complete selection and penetrance, expect enrichment



Age of disease mutation | observed in a case?
Results from simulation, modeled on mutation rate, penetrance and estimated selection coefficient of the 15q13.3 deletion.

Nadia Solovieff

APPLICATIONS

(***not*** any kind of comprehensive review: rather, just highlighting a couple of approaches that clearly didn't work and a couple that clearly did)

“Classical (aka 2005)” candidate gene sequencing

Deep resequencing and association analysis of schizophrenia candidate genes

Molecular Psychiatry (2013) **18,** 138–140; doi:10.1038/mp.2012.28; published online 3 April 2012

In 2005, we selected 10 genes for which there was reasonable evidence for involvement in the etiology of schizophrenia (*COMT*, *DAOA*, *DISC1*, *DRD2*, *DRD3*, *DTNBP1*, *HTR2A*, *NRG1*, *SLC6A3* and *SLC6A4*, Supplementary Table S1).[1] Although these genes have not received support from far larger and comprehensive subsequent studies, and may not contain common etiological variations,[2] it is possible that they contain uncommon variations of etiological importance. To test this hypothesis, we conducted a multistage resequencing study.

Crowley *et al* (2012) Mol Psych.

~700 cases, ~700 controls

No support that rare variants in these genes play a significant role in schizophrenia risk.

Early exome studies in SCZ delimit the genetic architecture

Exome Sequencing Followed by Large-Scale Genotyping Suggests a Limited Role for Moderately Rare Risk Factors of Strong Effect in Schizophrenia

Anna C. Need,[1,2,3,*] Joseph P. McEvoy,[3] Massimo Gennarelli,[4,5] Erin L. Heinzen,[1,2] Dongliang Ge,[1] Jessica M. Maia,[1] Kevin V. Shianna,[1,2] Min He,[1] Elizabeth T. Cirulli,[1] Curtis E. Gumbs,[1] Qian Zhao,[1] C. Ryan Campbell,[1] Linda Hong,[1] Peter Rosenquist,[6] Anu Putkonen,[7] Tero Hallikainen,[7] Eila Repo-Tiihonen,[7] Jari Tiihonen,[7,8] Deborah L. Levy,[9] Herbert Y. Meltzer,[10] and David B. Goldstein[1,11,*]

Exome sequenced 166 cases

Selected 5,155 variants (e.g. novel, seen in multiple cases)

Genotyped in further 2,617 cases, 1,800 controls

No single variant study-wide significant.

"Rather, multiple rarer genetic variants must contribute substantially to the predisposition to schizophrenia"

Focus on unusual genomic events

Neuron
Report

Cell PRESS

Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders

Elaine T. Lim,[1,4,5,6,7] Soumya Raychaudhuri,[4,6,9] Stephan J. Sanders,[10] Christine Stevens,[4] Aniko Sabo,[11] Daniel G. MacArthur,[1,4,6] Benjamin M. Neale,[1,4,5,6] Andrew Kirby,[1,4,6] Douglas M. Ruderfer,[1,3,4,5,6,8,12,14,15] Menachem Fromer,[1,3,4,5,6,8,12,14,15] Monkol Lek,[1,4,6] Li Liu,[18] Jason Flannick,[1,2,4,6] Stephan Ripke,[1,4,5] Uma Nagaswamy,[11] Donna Muzny,[11] Jeffrey G. Reid,[11] Alicia Hawes,[11] Irene Newsham,[11] Yuanqing Wu,[11] Lora Lewis,[11] Huyen Dinh,[11] Shannon Gross,[11] Li-San Wang,[19] Chiao-Feng Lin,[19] Otto Valladares,[19] Stacey B. Gabriel,[4] Mark dePristo,[4] David M. Altshuler,[1,2,4,6] Shaun M. Purcell,[1,3,4,5,6,8,12,14,15] NHLBI Exome Sequencing Project, Matthew W. State,[10] Eric Boerwinkle,[11,21] Joseph D. Buxbaum,[13,14,15,16,17] Edwin H. Cook,[22] Richard A. Gibbs,[11] Gerard D. Schellenberg,[20] James S. Sutcliffe,[23] Bernie Devlin,[24] Kathryn Roeder,[18] and Mark J. Daly[1,4,5,6,*]

Table 1. Population Distribution of Rare and Common LoFs

	Average Number of Homozygous Variants	Number of Unique Genes with a Homozygous Variant	Average Number of Heterozygous Variants	Number of Unique Genes with a Heterozygous Variant
Rare (≤5%) LoFs	0.05 variants per individual	33 genes	13 variants per individual	3,409 genes
Common (>5%) LoFs	5 variants per individual	96 genes	36 variants per individual	99 genes

The average number of rare (≤5%) and common (>5%) homozygous LoF variants, as well as the average number of such variants calculated from the BI case-control data set.

Excess of rare complete knockouts provides support for inherited component in ASD

3% contribution to ASD risk for rare autosomal complete knockouts

2% contribution to ASD risk in males from X-linked complete knockouts

One of four recent autism trio studies

Neuron
Article

Cell PRESS

De Novo Gene Disruptions in Children on the Autistic Spectrum

Ivan Iossifov,[1,6] Michael Ronemus,[1,6] Dan Levy,[1] Zihua Wang,[1] Inessa Hakker,[1] Julie Rosenbaum,[1] Boris Yamrom,[1] Yoon-ha Lee,[1] Giuseppe Narzisi,[1] Anthony Leotta,[1] Jude Kendall,[1] Ewa Grabowska,[1] Beicong Ma,[1] Steven Marks,[1] Linda Rodgers,[1] Asya Stepansky,[1] Jennifer Troge,[1] Peter Andrews,[1] Mitchell Bekritsky,[1] Kith Pradhan,[1] Elena Ghiban,[1] Melissa Kramer,[1] Jennifer Parla,[1] Ryan Demeter,[2] Lucinda L. Fulton,[2] Robert S. Fulton,[2] Vincent J. Magrini,[2] Kenny Ye,[3] Jennifer C. Darnell,[4] Robert B. Darnell,[4,5] Elaine R. Mardis,[2] Richard K. Wilson,[2] Michael C. Schatz,[1] W. Richard McCombie,[1] and Michael Wigler[1,*]

343 families, each with a single child on the autism spectrum and 1+ unaffected sibling

No significant difference in *de novo* missense rate in affected vs. unaffected children

Gene-disrupting mutations (nonsense, splice site, frameshifts) twice as frequent

Estimate 350 – 400 autism susceptibility genes

Many of the disrupted genes associated with the fragile X protein, FMRP, reinforcing links between autism and synaptic plasticity

Broadly similar picture in other studies (Neale et al.; O'Roak et al.; Sanders et al., *Nature* 2012)

Combining data, individual genes recurrently hit by disruptive mutations can be identified.

In the pipeline…

- As well as the published large-scale sequencing in autism in 2012 and smaller published schizophrenia *de novo* studies, other emerging large-scale projects (e.g. as presented at WCPG, Hamburg): e.g.
  - Exome sequencing in schizophrenia in >5000 individuals (Swedish, population-based)
  - UK10K sequencing study including schizophrenia
  - Exome sequencing in >600 trios (Bulgarian)
  - Multiple moderately-sized exome studies in bipolar disorder (population and family)
- On balance, seems clear (for schizophrenia) both that a) promising and convergent results are emerging, but b) not "game-changing" at this point and harder to extricate signal than, e.g., for autism.
- Rather (as PGC'ers might agree), 2012 was definitely the year of GWAS delivering, for schizophrenia at least…

PGC2 and sequence data

- Several emerging disease-focused sequencing consortia
  - Autism Sequencing Consortium
  - Bipolar disorder Sequencing Consortium

- Different models:
  - **"Share BAMs"**
  - Pool raw data, establish joint-calling pipeline(s), centralized data repository / analysis hub
  - cf. PGC CNV model

  - **"Share VCFs"**
  - Variants called at individual sites, attempt to reconcile/merge downstream for central analysis
  - cf. PGC GWAS model

  - **"Share results"**
  - All analyses performed centrally, share all case/control counts and meta-information
  - cf. standard "meta-analysis" model from GWAS

  - **"Lookup/replication"**
  - Loose consortium of groups that agree to look up specific sites of interest in a directed manner
  - cf. replication samples included in PGC GWAS efforts

  - **"Targeted genotyping"**
  - As above, but with facility to perform large-scale genotyping of select rare variants across many cohorts
  - cf. aspects of PsychChip development

- In all cases, possible to share best practice from an analytic perspective, e.g. via the central analysis group.

ACKNOWLEDGEMENTS

MSSM
Menchem Fromer
Douglas Ruderfer
Eli Stahl
Pamela Sklar

Stanley Center | Broad | MGH
Nadia Solovieff
Stephan Ripke
Jennifer Moran
Ed Scolnick
Steve Hyman
Steve Haggarty
Ben Neale
Colm O'Dushlaine
Steven McCaroll
Mark Daly

Bulgarian Trio Study
Mick O'Donovan (Cardiff)
Mike Owen (Cardiff)
Peter Holmans (Cardiff)
George Kirov (Cardiff)
Andrew Pocklington (Cardiff)
Davy Kavanagh (Cardiff)
Aarno Palotie (Sanger)
Stanley Center
MSSM

Swedish Study
Patrick Sullivan (UNC)
Christina Hultman (Karolinska)
MSSM
Stanley Center