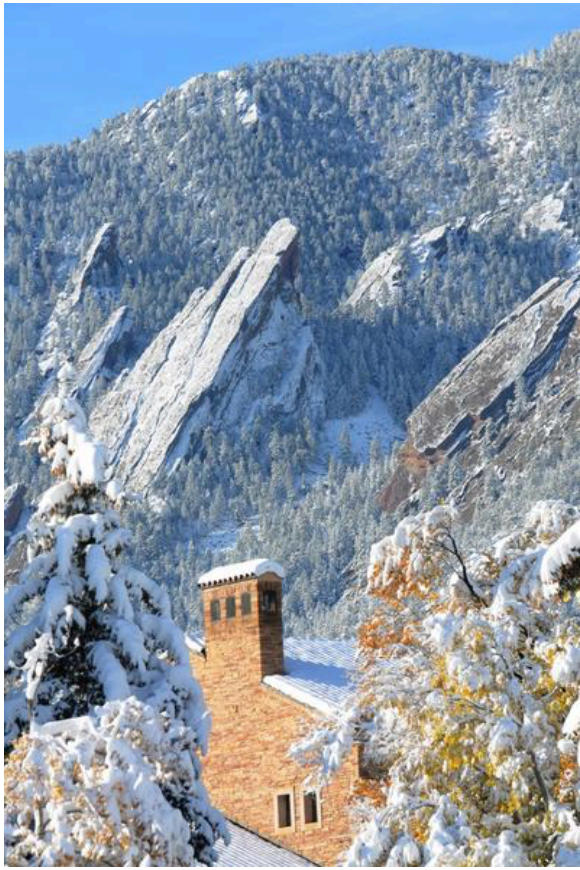
University of Colorado Boulder

Lecture 6: Various!

Matt McQueen | Associate Professor

Department of Integrative Physiology
Institute for Behavioral Genetics
Institute of Behavioral Science
University of Colorado Boulder

Department of Epidemiology (secondary)
Colorado School of Public Health
University of Colorado

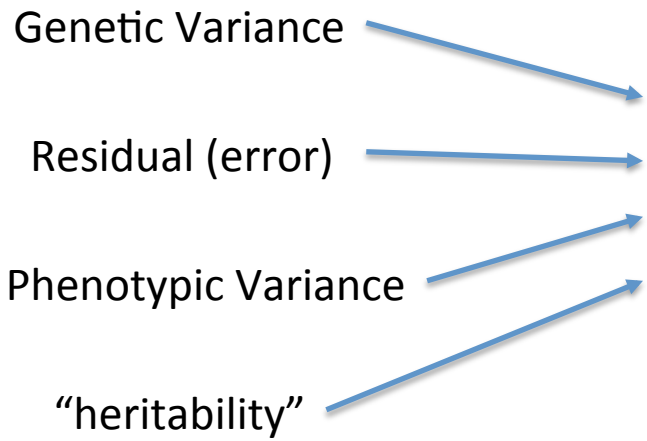


Wrap up topics

- Interpreting GCTA output
- Next Generation Sequencing
- Copy Number Variants
- Meta-Analysis

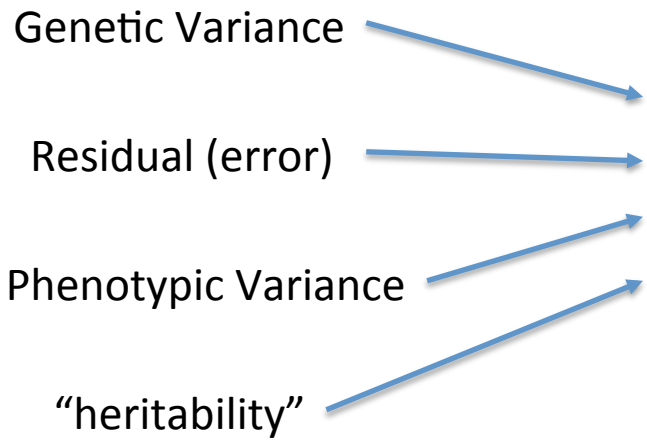
Interpreting GCTA Output

GCTA Output



Source	Variance	SE
V(1)	8.460930	5.852812
V(e)	9.985167	5.369622
Vp	18.446097	0.989077
V(1)/Vp	0.458684	0.304386
logL	−1791.054	
n	923	

GCTA Output

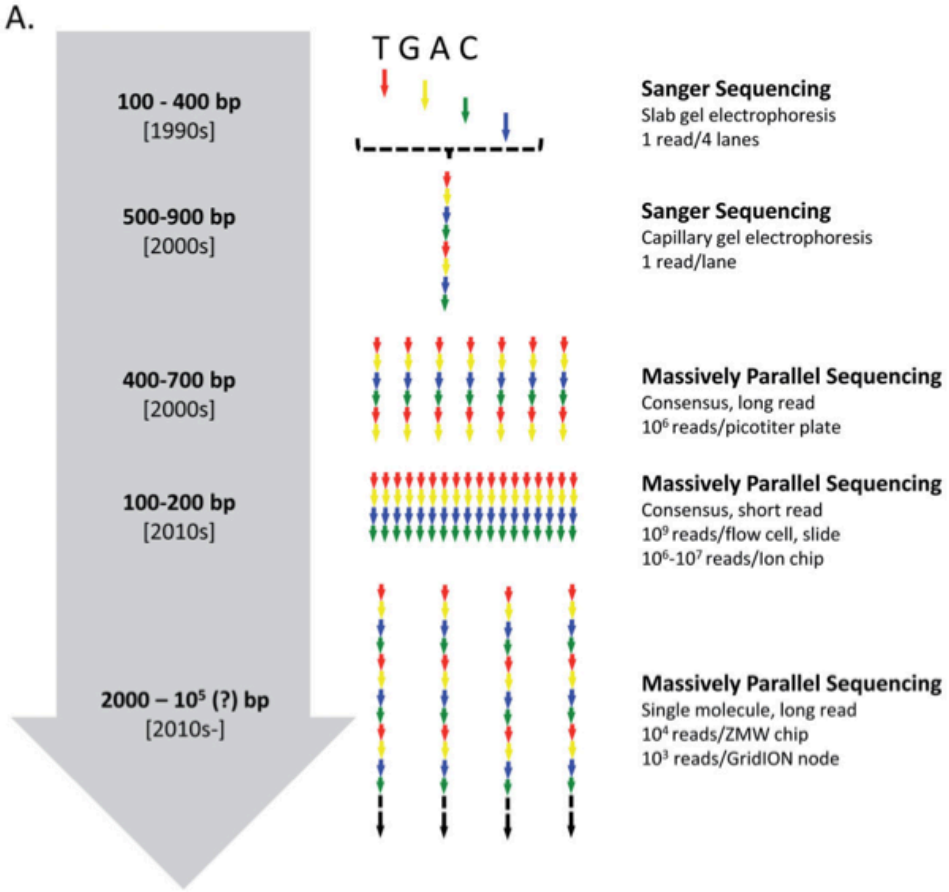


Source	Variance	SE
V(1)	8.460930	5.852812
V(e)	9.985167	5.369622
Vp	18.446097	0.989077
V(1)/Vp	**0.458684**	0.304386
logL	−1791.054	
n	923	

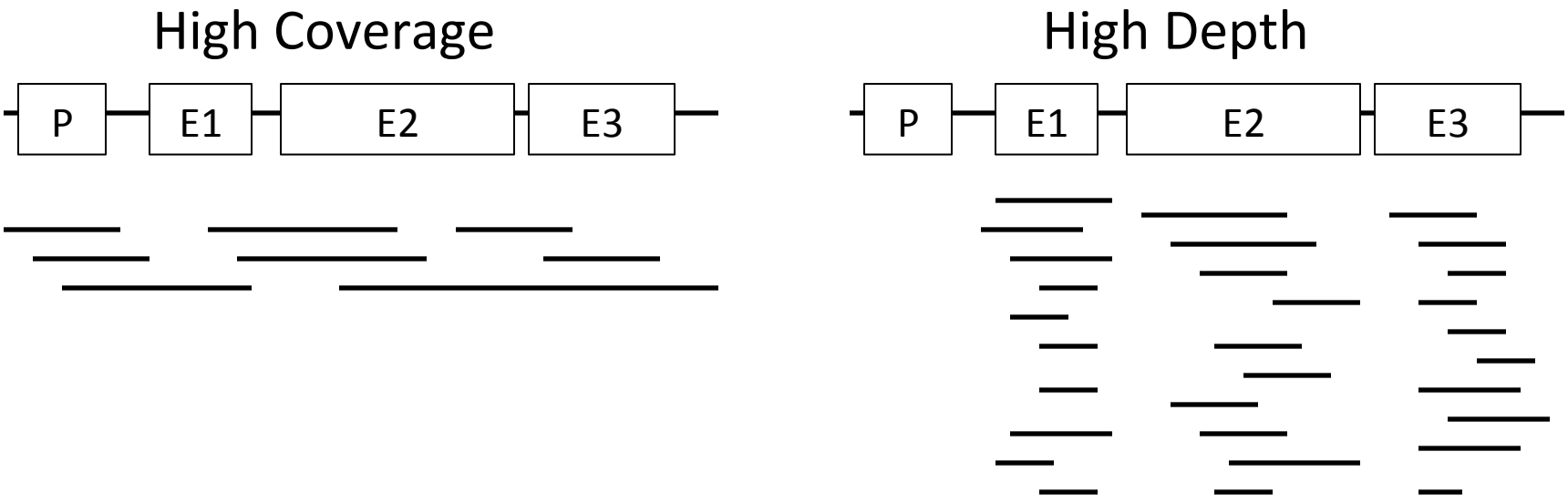
NOTE: This is the narrow sense heritability (additive effects)

Next Generation Sequencing

Sequencing


A.
100 - 400 bp
[1990s]
500-900 bp
[2000s]
400-700 bp
[2000s]
100-200 bp
[2010s]
2000 – 10⁵ (?) bp
[2010s-]
T G A C
Sanger Sequencing
Slab gel electrophoresis
1 read/4 lanes
Sanger Sequencing
Capillary gel electrophoresis
1 read/lane
Massively Parallel Sequencing
Consensus, long read
10⁶ reads/picotiter plate
Massively Parallel Sequencing
Consensus, short read
10⁹ reads/flow cell, slide
10⁶-10⁷ reads/Ion chip
Massively Parallel Sequencing
Single molecule, long read
10⁴ reads/ZMW chip
10³ reads/GridION node

Sequencing coverage vs depth



Next Generation Sequencing

- Moving fast
  - High depth, high coverage now possible
  - Prices falling

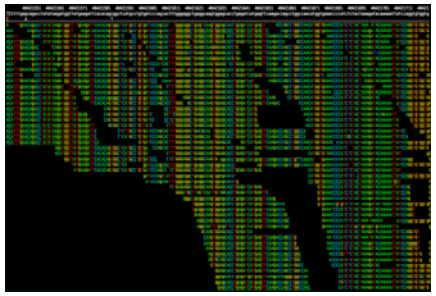
What are we expecting to find?

- Is this a looking under the lamp post issue?
  - More and more precise measurement
- Is there something new that we haven't seen?

Next Generation Sequencing

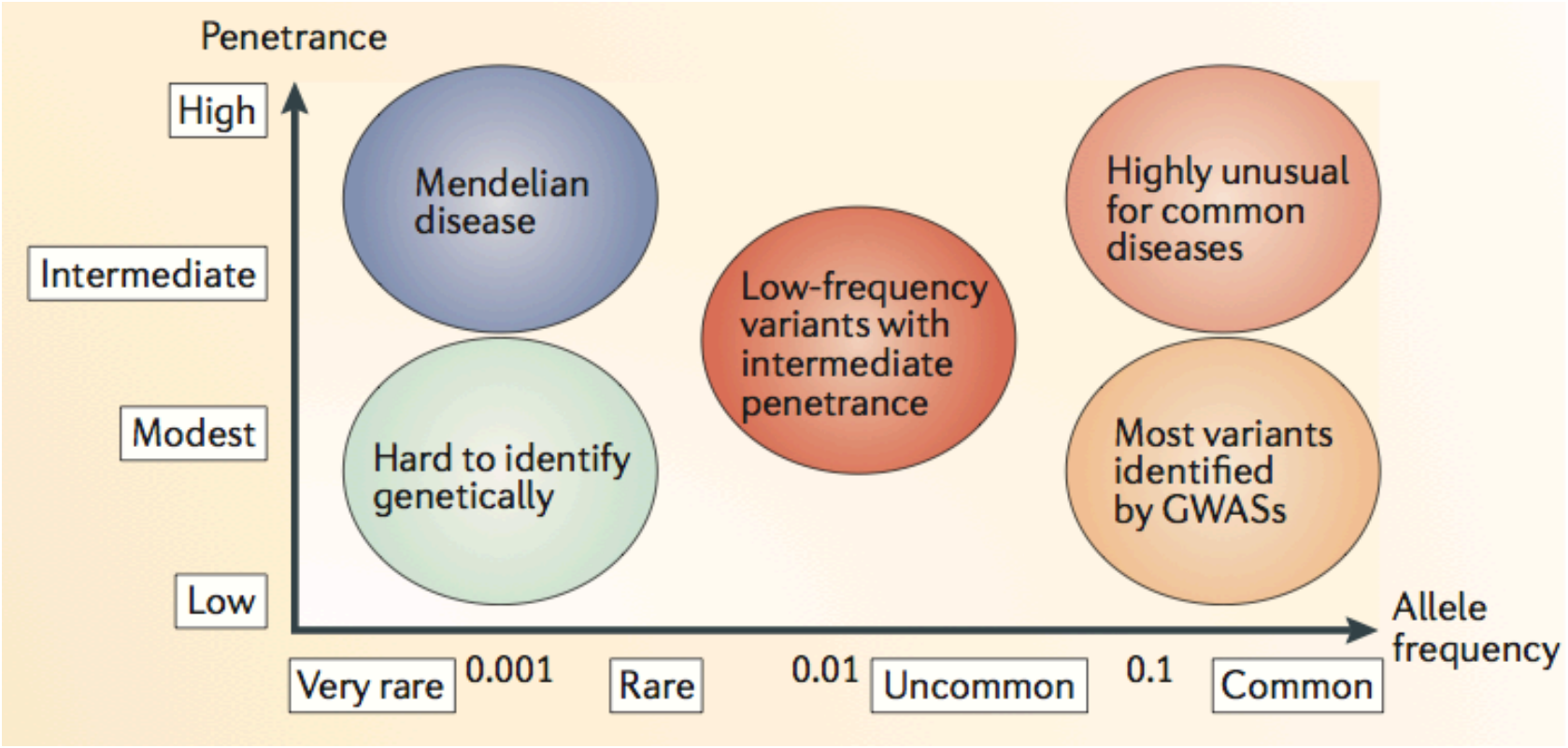
- Will this provide more answers than GWAS?

Sequencing



- Objective
  - Find **rare/common variants** associated with disease
- Design
  - Cohort, case-control, family-based
- Molecular information
  - 3B base-pair
- Desired outcome
  - Find genetic variation underlying disease

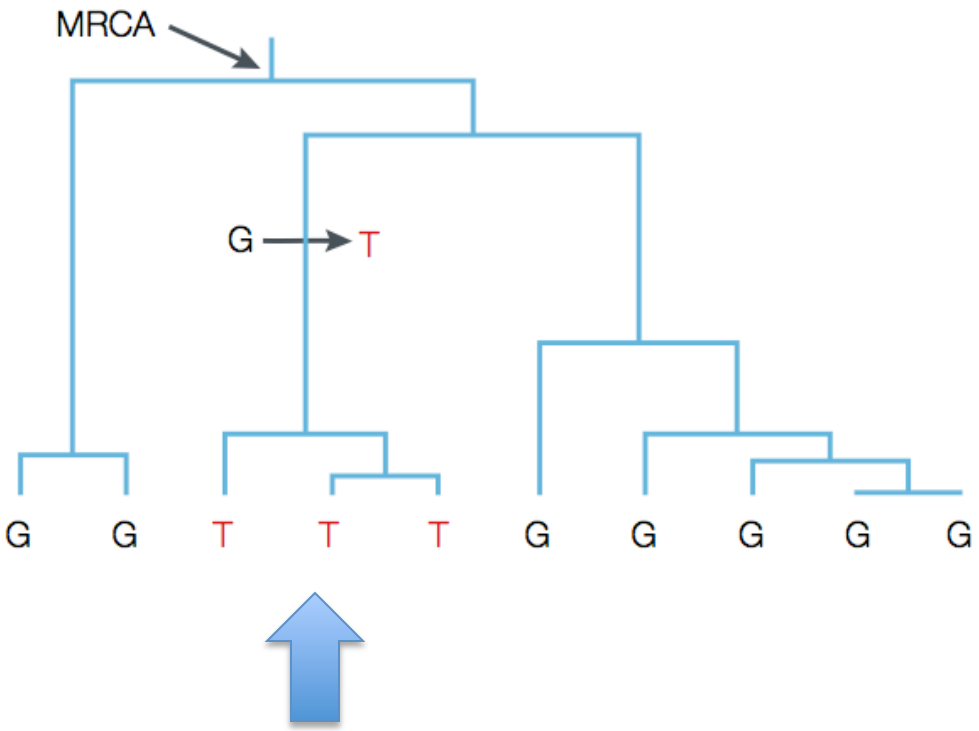
Disease and DNA Variation



Penetrance: P(D | G)

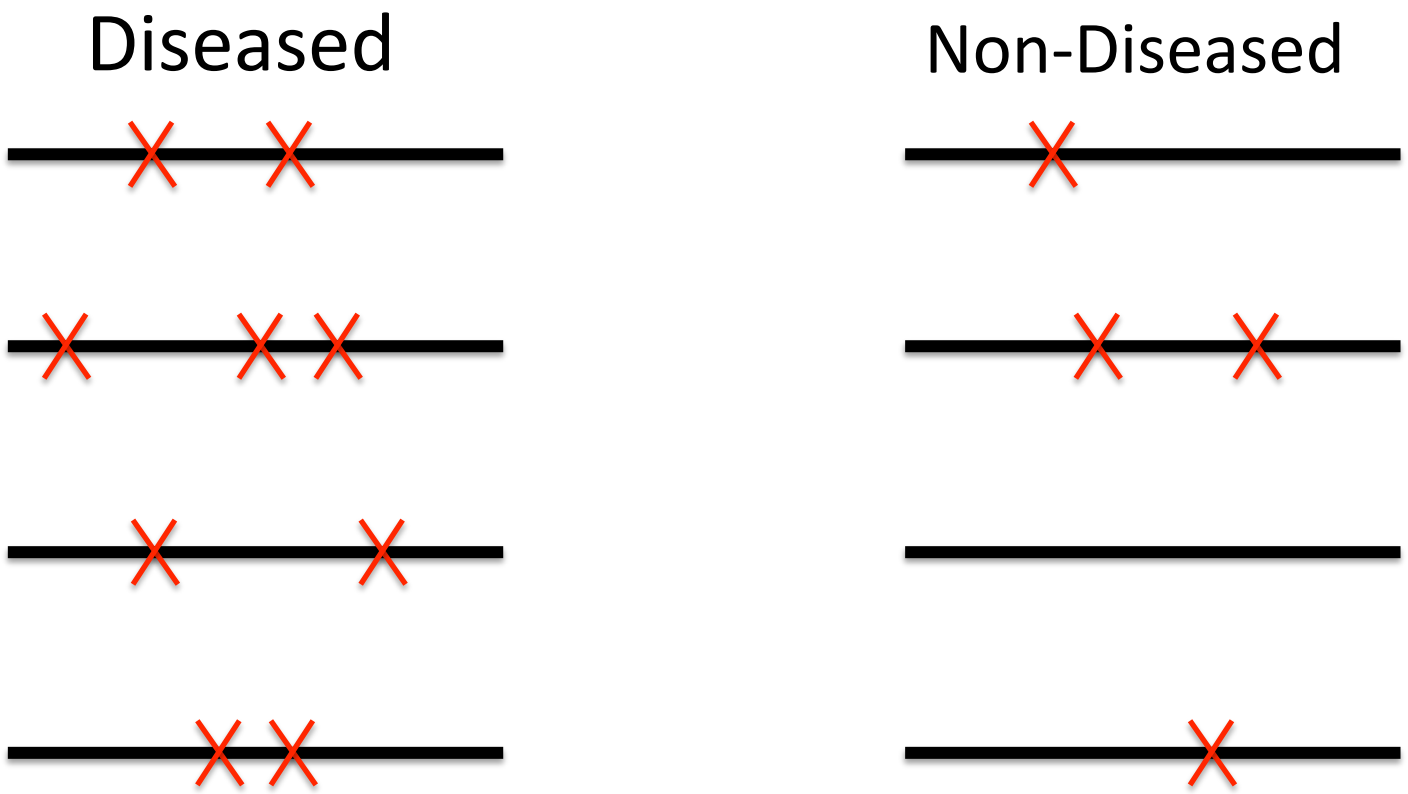
2012 Nature Reviews | Genetics

GWAS: Common Disease / Common Variant



Higher disease prevalence associated with T allele

Sequencing: Rare Variant Hypothesis

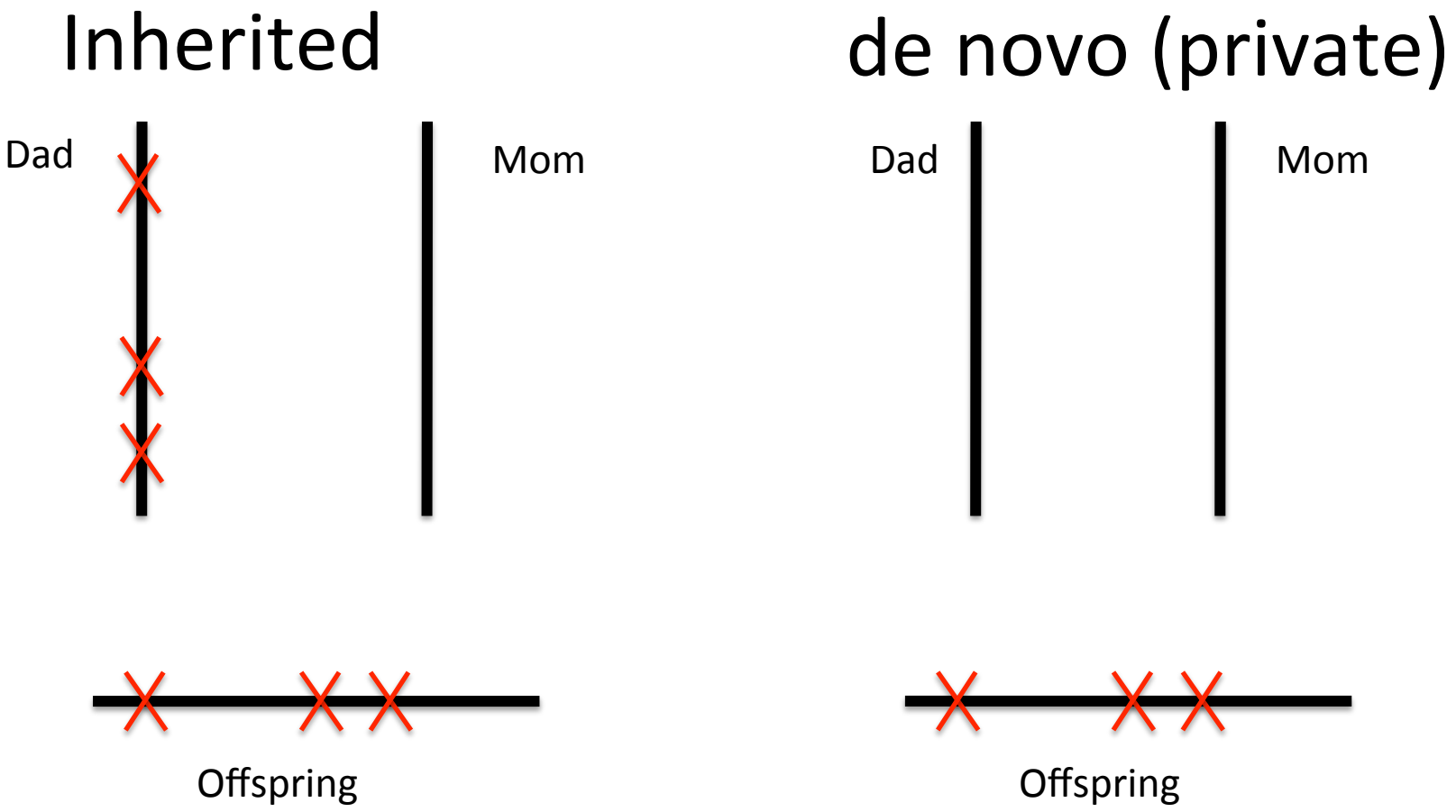


Inherited vs de novo mutation

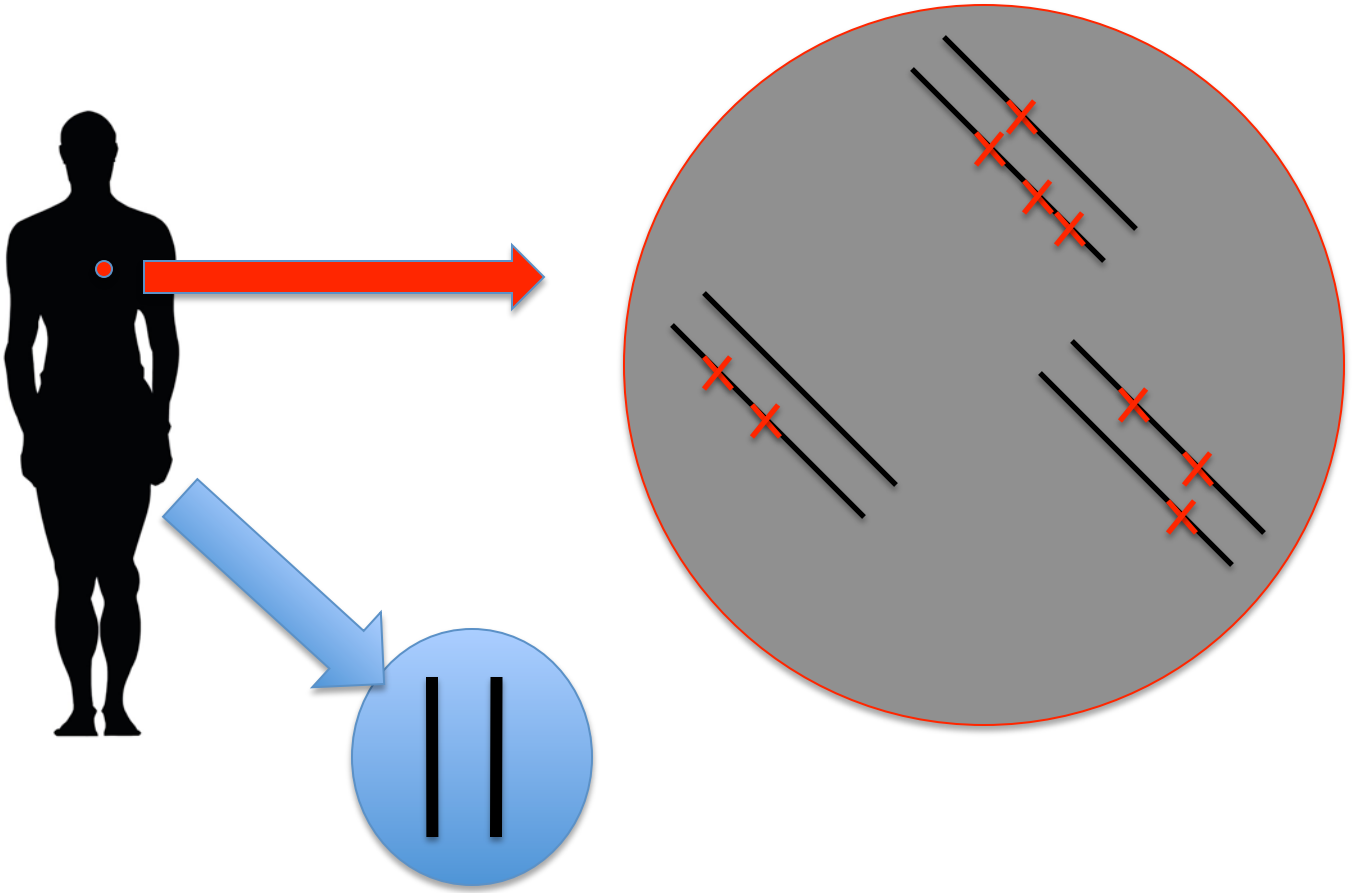
Offspring

Offspring

Inherited vs de novo mutation

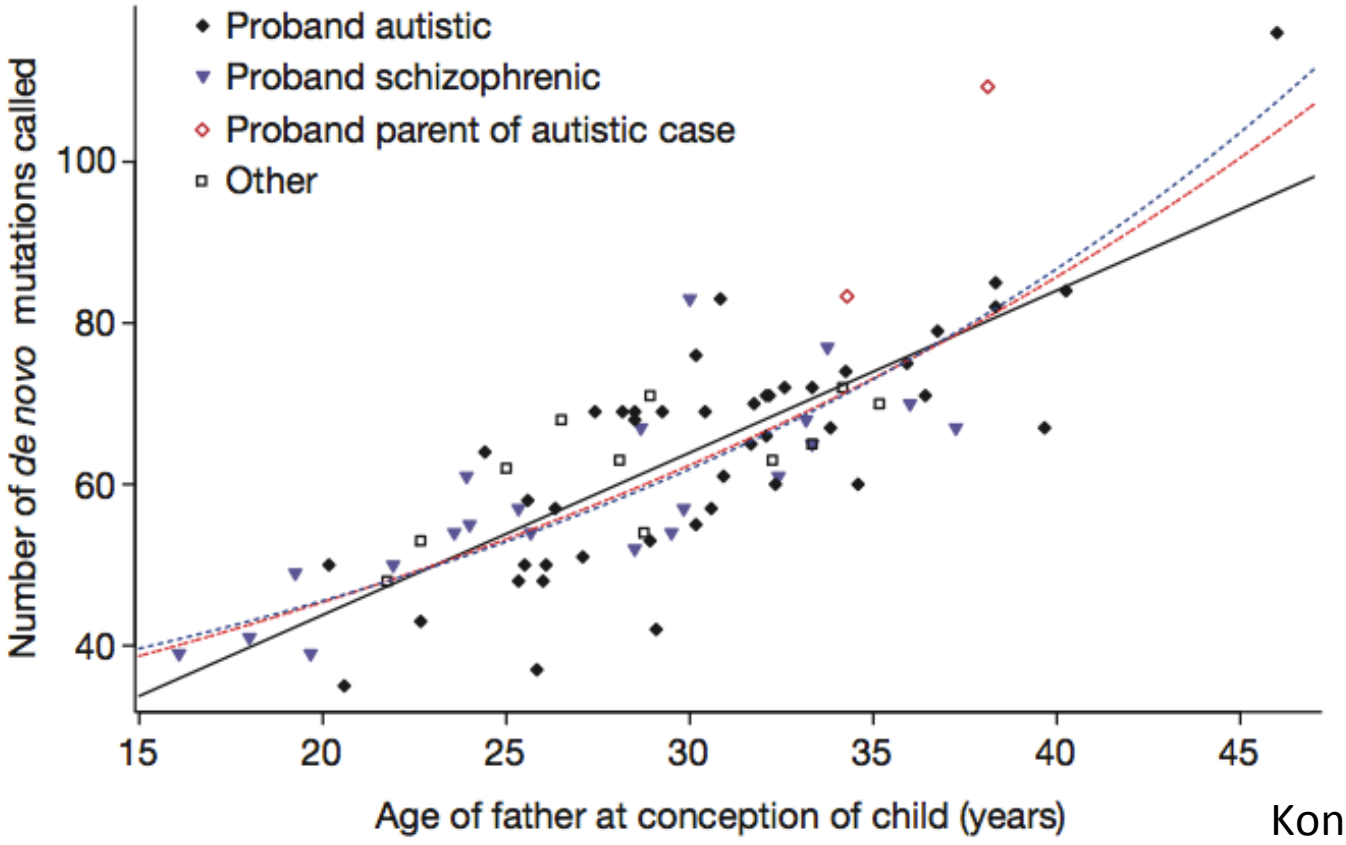


Tumor genomes



Gerlinger et al (2012) | NEJM

Paternal Age, Autism and Mutations



Kong et al., 2012

Disease characteristic vs prediction

- Mutations and genetic variation may be part of the disease process
- However, can we use our DNA to predict future disease?
  - Using “clones” (monozygotic twins) might help us answer the question…

Disease/Condition	Sex	Number of MZ Twin Pairs	Number MZ Disease Concordant Pairs	Number MZ Disease Discordant Pairs	Disease Prevalence in Cohort (*CR*)
Bladder Cancer	Male & Female	15668	5	189	0.6%
Breast Cancer	Female	8437	42	505	3.5%
Colorectal Cancer	Male & Female	15668	30	416	1.5%
Leukemia	Male & Female	15668	2	103	0.3%
Lung Cancer	Male & Female	15668	18	296	1.1%
Ovarian Cancer	Female	8437	3	125	0.8%
Pancreatic Cancer	Male & Female	15668	3	123	0.4%
Prostate Cancer	Male	7231	40	299	2.6%
Stomach Cancer	Male & Female	15668	11	223	0.8%
Thyroid Autoimmunity	Male & Female	284	7	17	5.5%
Type 1 Diabetes	Male & Female	4307	3	20	0.3%
Gallstone Disease	Male & Female	11073	112	956	5.3%
Type 2 Diabetes	Male & Female	4307	29	113	2.0%
Alzheimer's Disease	Male & Female	398	2	8	1.5%
Dementia	Male & Female	398	3	16	2.8%
Parkinson Disease	Male & Female	3477	7	60	1.1%
Chronic Fatigue	Female	1803	133	526	22.0%
Chronic Fatigue	Male	1426	48	266	12.7%
Gastro Esophageal Reflux Disorder (GERD)	Female	1260	63	284	16.3%
Gastro Esophageal Reflux Disorder (GERD)	Male	918	32	185	13.6%
Irritable Bowel Syndrome	Male & Female	1252	14	97	5.0%
Coronary heart disease (CHD) Death	Female	2004	97	424	15.4%
Coronary heart disease (CHD) Death	Male	1640	153	451	23.1%
Stroke-related Death	Male & Female	3852	35	316	5.0%
General Dystocia	Female	928	40	173	13.6%
Pelvic Organ Prolapse	Female	3376	34	157	3.3%
Stress Urinary Incontinence	Female	3376	13	87	1.7%

MZ: Monozygotic. Disease prevalence in cohort (cohort risk, *CR*) was determined as described in the Materials and Methods.

Roberts et al., 2012

NGS Analytic Considerations

- Common variation
  - GWAS pipeline applies
- Rare variation
  - Might require new methods/thinking

Analysis of rare variants

- Effectively count data
  - Number of mutations/variants
- Accumulation of rare variants
  - Genome-wide
  - Genic region
  - Pathway/system

Analysis of rare variants

- Counts follow a Poisson distribution
  - "rate" of mutational load
- Weight variants
  - Prior biological information
  - Up-weight specific variants

Hecht *et al. BMC Genomics* 2015, **16**(Suppl 8):S1
http://www.biomedcentral.com/1471-2164/16/S8/S1

BMC Genomics

RESEARCH Open Access

Better prediction of functional effects for sequence variants

Maximilian Hecht[1*], Yana Bromberg[2,3,4], Burkhard Rost[1,4]

From VarI-SIG 2014: Identification and annotation of genetic variants in and disease
Boston, MA, USA. 12 July 2014

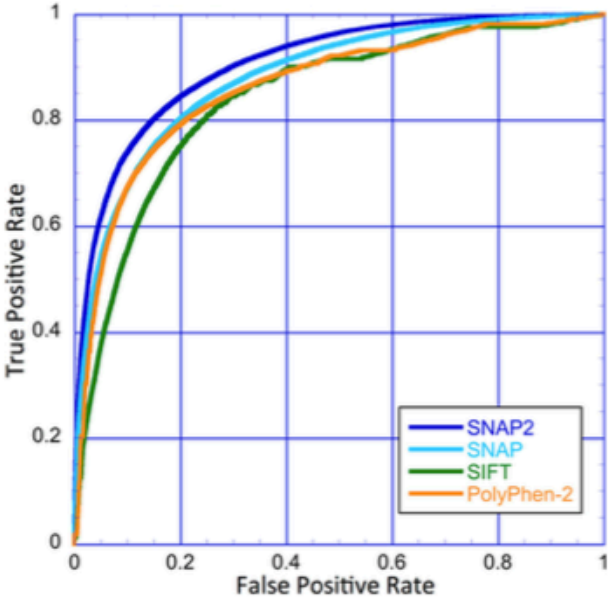


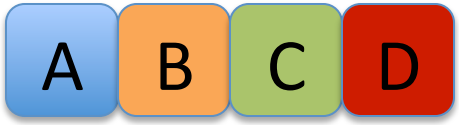
Figure 1 SNAP2 performs best for the ALL data set. This figure shows performance estimates for the *ALL* data set. Our new method SNAP2 (dark blue, AUC = 0.905) outperforms its predecessor SNAP (light blue, AUC = 0.880), PolyPhen-2 (orange, AUC = 0.853) and SIFT (green, AUC = 0.838) over the entire spectrum of the Receiver Operating Characteristic (ROC) curve. Curves are significantly different from each other at a significance level of P < 10-4 as measured by the DeLong method [59]. All SNAP2 results were computed on the test sets not used in training after a rigorous split into training, cross-training and testing. Results for PolyPhen-2 and our original SNAP included some of those proteins in their training, suggesting over-estimated performance.

Watch this space

- Methods are changing fast

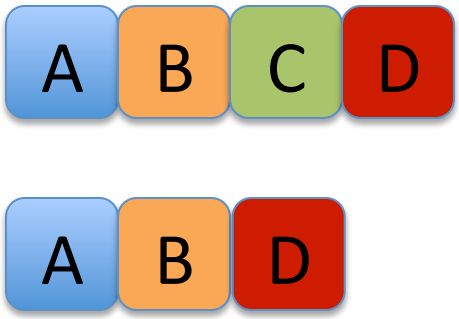
Copy Number Variation

CNV



Reference

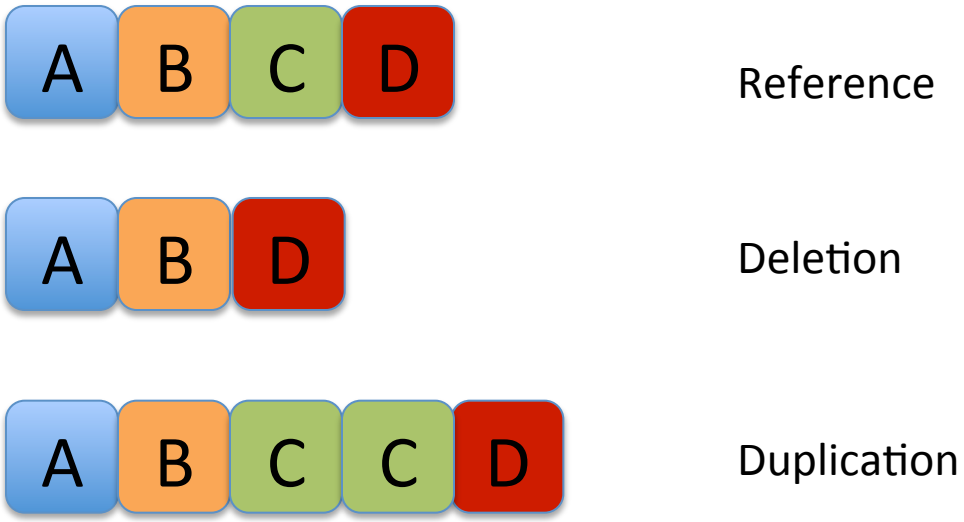
CNV



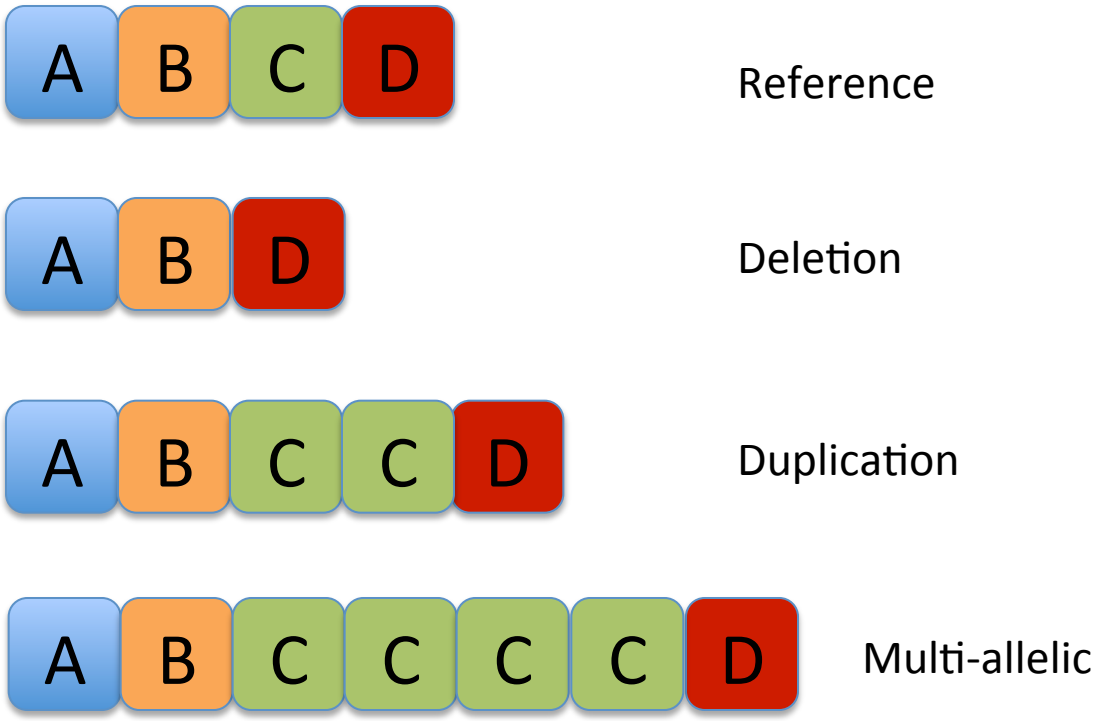
Reference

Deletion

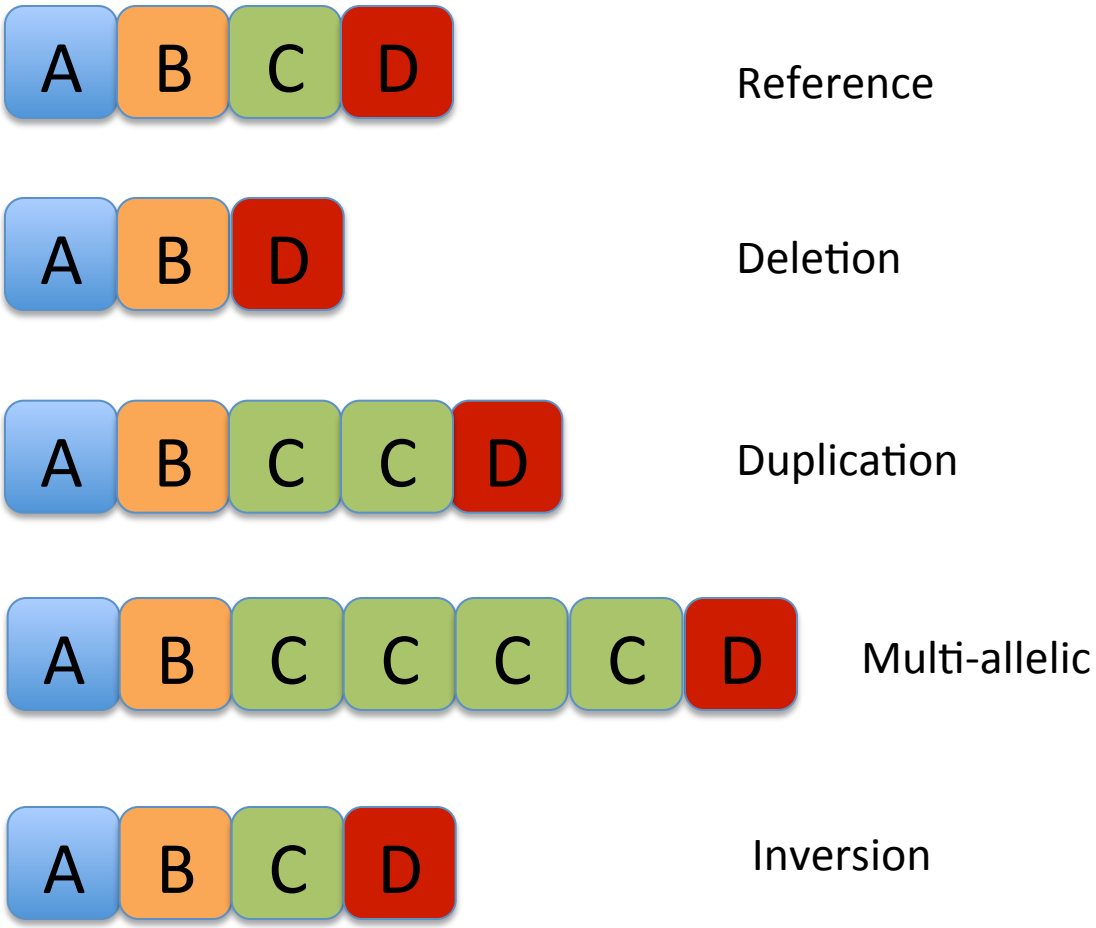
CNV



CNV


A B C D
Reference
A B D
Deletion
A B C C D
Duplication
A B C C C C D
Multi-allelic

CNV


A B C D
Reference
A B D
Deletion
A B C C D
Duplication
A B C C C C D
Multi-allelic
A B C D
Inversion

How do we measure CNVs?

- GWAS platforms
- RT PCR and dPCR methods
- Next Gen Sequencing

GWAS Platform

- PennCNV is a common tool designed to harness Illumina and Affy data
  - Reliable and well-documented

CNV Analysis

Analysis of copy number variations at 15 schizophrenia-associated loci

Elliott Rees, James T. R. Walters, Lyudmila Georgieva, Anthony R. Isles, Kimberly D. Chambert, Alexander L. Richards, Gerwyn Mahoney-Davies, Sophie E. Legge, Jennifer L. Moran, Steven A. McCarroll, Michael C. O'Donovan, Michael J. Owen and George Kirov

CNV Analysis

Table 1 Findings from our data-set for previously implicated copy number variation (CNV) loci in schizophrenia[a]

Locus	Position in Mb	Case group ($n = 6882$)		Control group ($n = 6316$)		OR (95% CI)	P
		CNVs, n	Frequency, %	CNVs, n	Frequency, %		
1q21.1 del	chr1:146,57-147,39	12	0.17	1	0.016	11.03 (1.43–84.86)	**0.0027**
1q21.1 dup	chr1:146,57-147,39	8	0.12	5	0.079	1.47 (0.48–4.49)	0.35
NRXN1 del	chr2:50,15-51,26	11	0.16	0	0.00	NA (1.25–∞)	**7.7×10^{-4}**
3q29 del	chr3:195,73-197,34	4	0.058	0	0.00	NA (0.44–∞)	0.074
WBS dup	chr7:72,74-74,14	3	0.044	1	0.016	2.75 (0.29–26.48)	0.35
VIPR2 dup	chr7:158,82-158,94	1	0.015	6	0.095	0.15 (0.02–1.27)	0.99
15q11.2 del	chr15:22,80-23,09	44	0.64	26	0.41	1.56 (0.96–2.53)	**0.046**
AS/PWS dup	chr15:24,82-28,43	8	0.12	0	0.00	NA (0.90–∞)	**0.0055**
15q13.3 del	chr15:31,13-32,48	4	0.058	2	0.032	1.84 (0.34–10.03)	0.38
16p13.11 dup	chr16:15,51-16,30	24	0.35	12	0.19	1.84 (0.92–3.68)	0.056
16p11.2 distal del	chr16:28,82-29,05	0	0.00	2	0.032	NA (0–3.82)	1
16p11.2 dup	chr16:29,64-30,20	27	0.39	0	0.00	NA (3.09–∞)	**2.3×10^{-8}**
17p12 del	chr17:14,16-15,43	4	0.058	3	0.047	1.22 (0.27–5.47)	0.55
17q12 del	chr17:34,81-36,20	1	0.015	0	0.00	NA (0.11–∞)	0.52
22q11.2 del	chr22:19,02-20,26	20	0.29	0	0.00	NA (2.28–∞)	**2.2×10^{-6}**
Totals		171	2.48	58	0.92		**1.4×10^{-12}**

del, deletion; dup, duplications, NA, not applicable; WBS, Williams–Beuren syndrome; AS/PWS, Angelman/Prader–Willi syndrome.
a. Copy number variation positions are in UCSC Build 37. Significant results are in bold (using Fisher exact test, 1-tailed).

Meta-Analysis

Aggregating the evidence

- Often, we are interested in combining evidence across independent studies
- There are a variety of ways to do this

Differing approaches…

- Mega-Analysis
- Combining Significance
- Meta-Analysis
- Weighted Hypothesis Testing

Mega-Analysis

- Combine two or more samples
- Requires access to raw data
- Many consortia utilize this approach

Mega-Analysis

- Strengths
  - Unprecented statistical power
- Weaknesses
  - Combining across heterogeneous samples
  - Ignore variation between studies

Combining significance

- Rather than combine raw data, you combine test statistics and/or p-values
- Simplest approach
  - Fisher's Method

$$X_{2k}^2 \sim -2\sum_{i=1}^{k} \ln(p_i)$$

Fisher's Method

- Strengths
  - Simple approach
  - Does not require raw data
- Weaknesses
  - Assumptions
    - Independent tests
    - Uniform distribution of p-values
  - Lack of effect size (only p-values)

Meta-Analysis

- Combining effect size estimates across studies
  - Odds ratios, risk ratios, etc.
- Important distinction
  - Random vs Fixed Effects

Fixed vs Random Effects

- Fixed Effects Meta-Analysis
  - Ignores between-study variance
- Random Effects Meta-Analysis
  - Incorporates between-study variance
  - More conservative (wider confidence intervals)

Conducting a meta-analysis

- Requirements
  - Proper extensive literature search
  - Parameter estimate (i.e. odds ratio)
  - Standard error
- Various tools to conduct a meta-analysis
  - R packages
    - Metafor is a good option
    - Provides graphics

Examples

- See alzgene.org