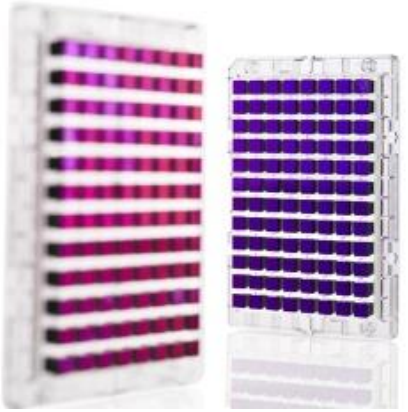
Genome-wide association studies (GWAS)



Outline for GWAS

- Review / Overview
- Design
- Analysis
  - QC
  - Prostate cancer example
  - Imputation
  - Replication & Meta-analysis

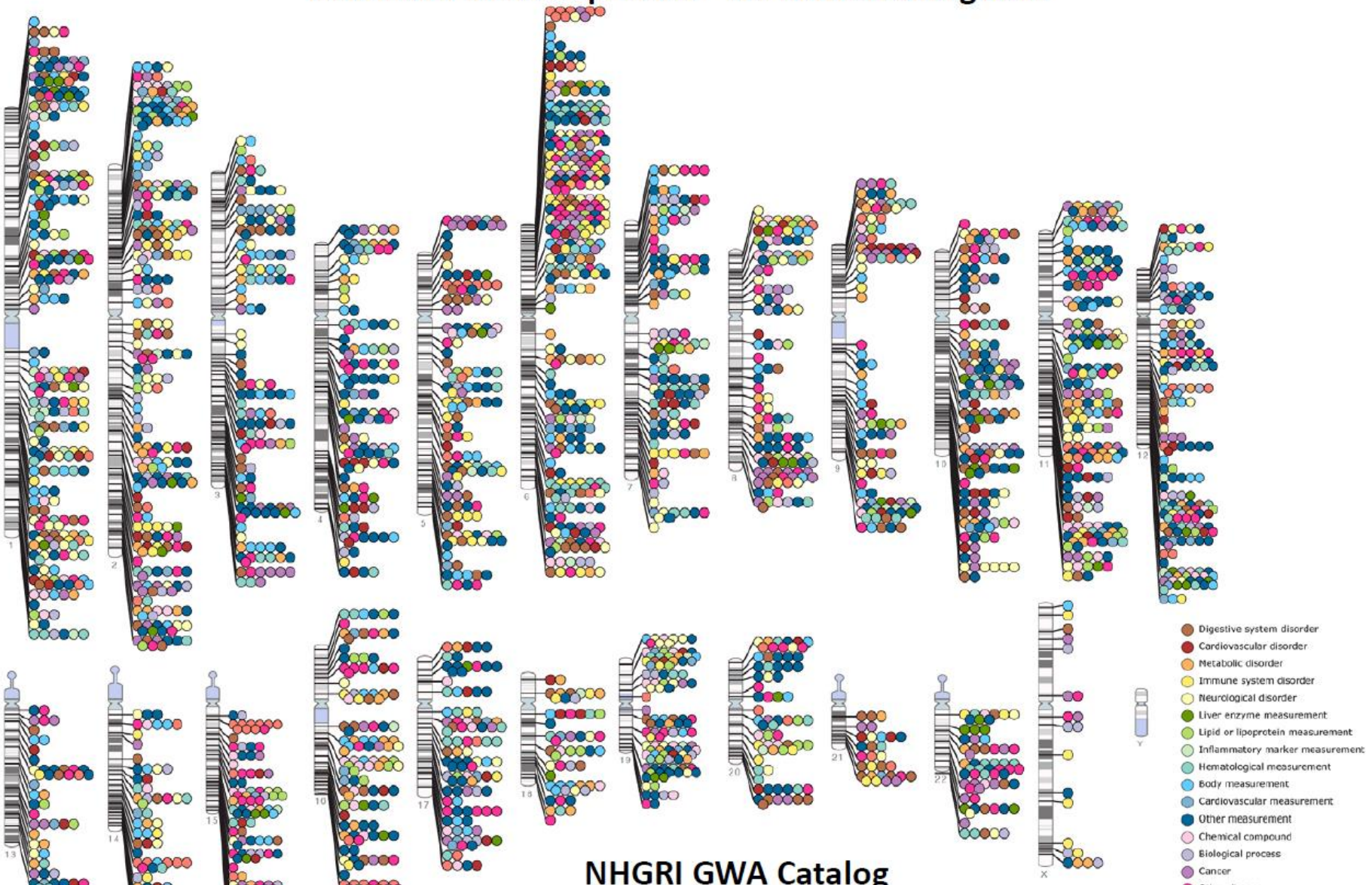
(Likely next time)

Outline for GWAS

- Review / Overview
- Design
- Analysis
  - QC
  - Prostate cancer example
  - Imputation
  - Replication & Meta-analysis

Published Genome-Wide Associations through 07/2012

Published GWA at $p \leq 5X10^{-8}$ for 18 trait categories



NHGRI GWA Catalog

www.genome.gov/GWAStudies

www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

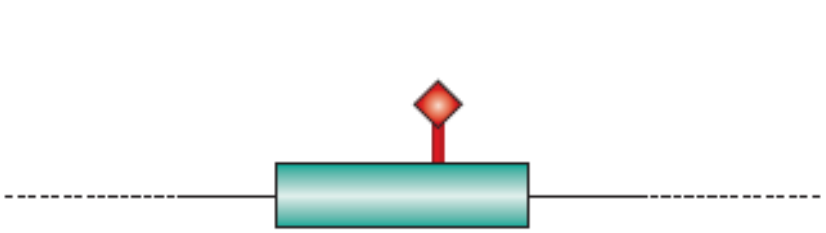
EMBL-EBI

01/25/2021

Outline for GWAS

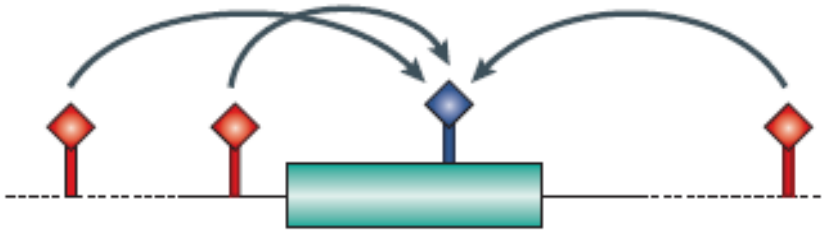
- Review / Overview
- Design
- Analysis
  - QC
  - Prostate cancer example
  - Imputation
  - Replication & Meta-analysis

Design 1: Microarray – Genotype a subset of markers available



Direct association

Disease locus directly typed



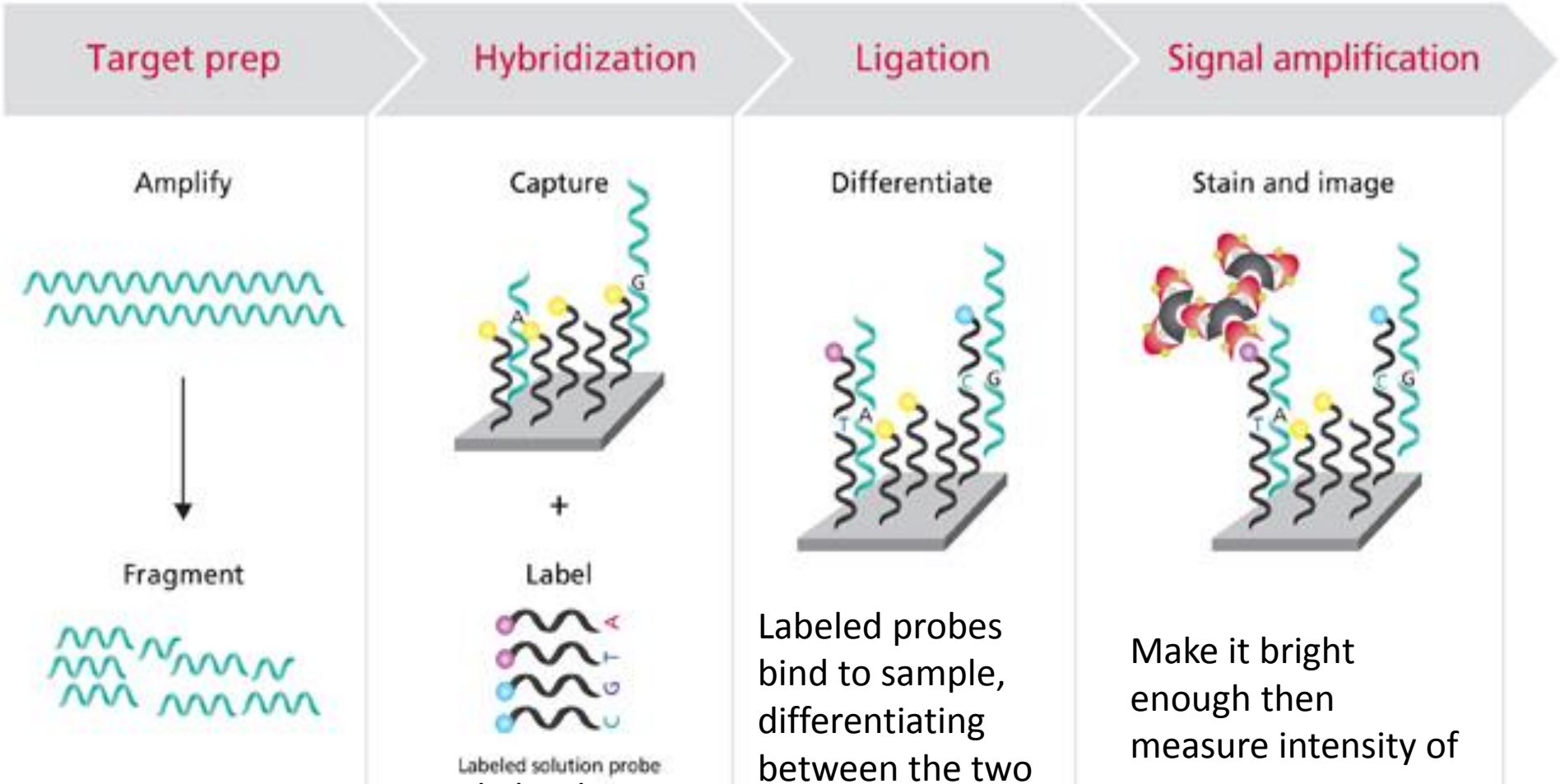
Indirect association (guilt by association)

Marker correlated with typed disease locus

- Previously we talked about using this to "tag" candidate genes (only Genotype a subset of the SNPs).
- Same principal applies to tagging the whole genome.

Hirschhorn & Daly, Nat Rev Genet 2005

Technology behind a GWAS Microarray



Sample binds to array

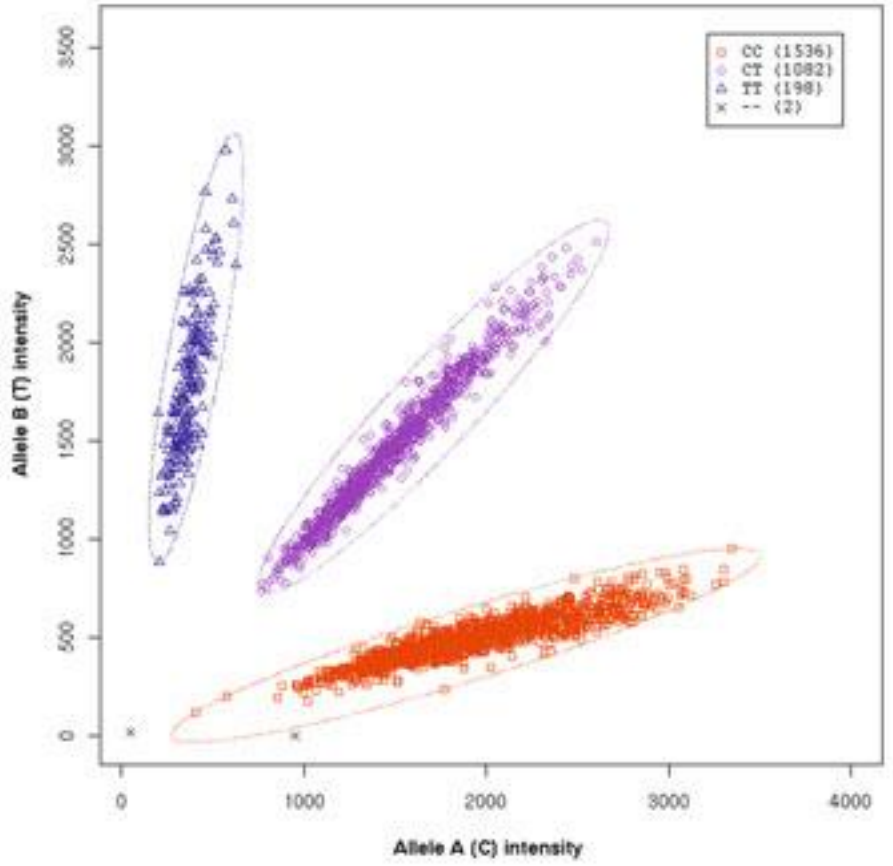
Labeled probes bind to sample, differentiating between the two alleles

Make it bright enough then measure intensity of array

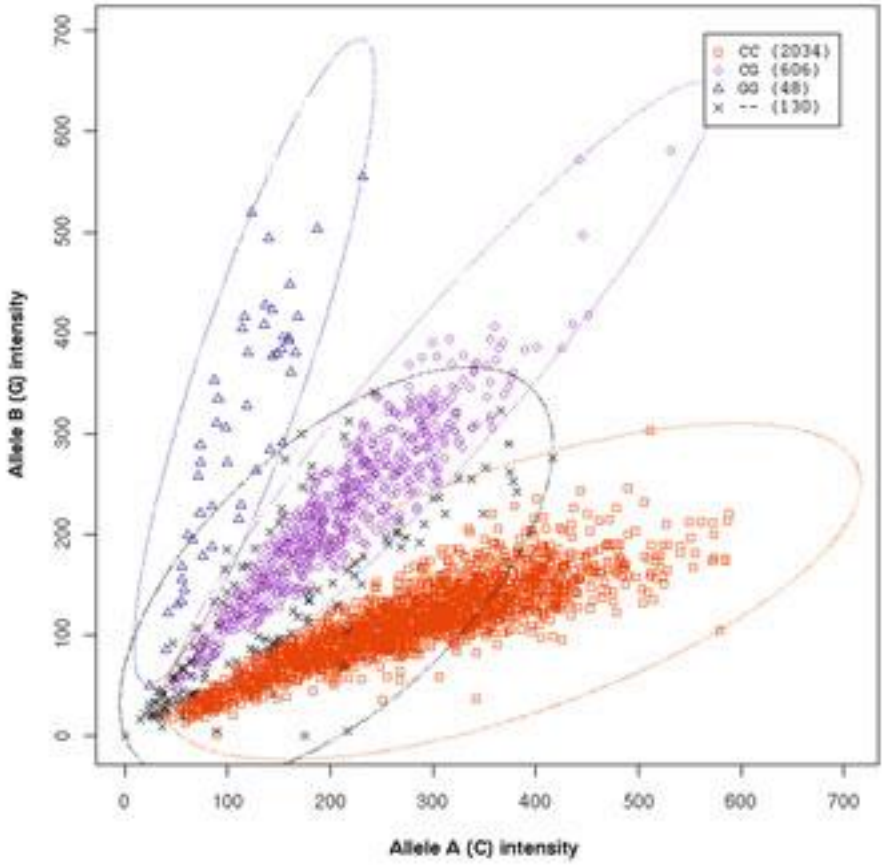
Assay ~ 0.7 - 5M SNPs (keeps increasing)

Affymetrix, http://www.affymetrix.com

Raw microarray data: Genotype calls



Good calls!



Bad calls!

Class Exercise (groups of 3-4)!: How would you design a *general* GWAS microarray?

- It will be linked to an EHR, all diseases of interest
- Do you have to genotype *everything*?
- Regions/SNPs to prioritize?

Hoffmann et al., Next generation genome-wide association tool. Design and coverage of a high-throughput European-optimized SNP array. Genomics, 2012.
Hoffmann et al., Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics, 2012.

Class Exercise (groups of 3 ideal)!: How would you design a *general* GWAS microarray?

- It will be linked to an EHR, all diseases of interest
- Do you have to genotype *everything*/*everyone*? [tag SNPs (population?), minor allele frequency limit, old multistage design]
- Regions/SNPs to prioritize? [genes?, prev hits, functional significance?; some SNPs cost more; some SNPs work better]
- [Similar to candidate genes before, but grander scale]

Hoffmann et al., Next generation genome-wide association tool. Design and coverage of a high-throughput European-optimized SNP array. Genomics, 2012.
Hoffmann et al., Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics, 2012.

Class Exercise!: How would you design a *general* GWAS microarray?

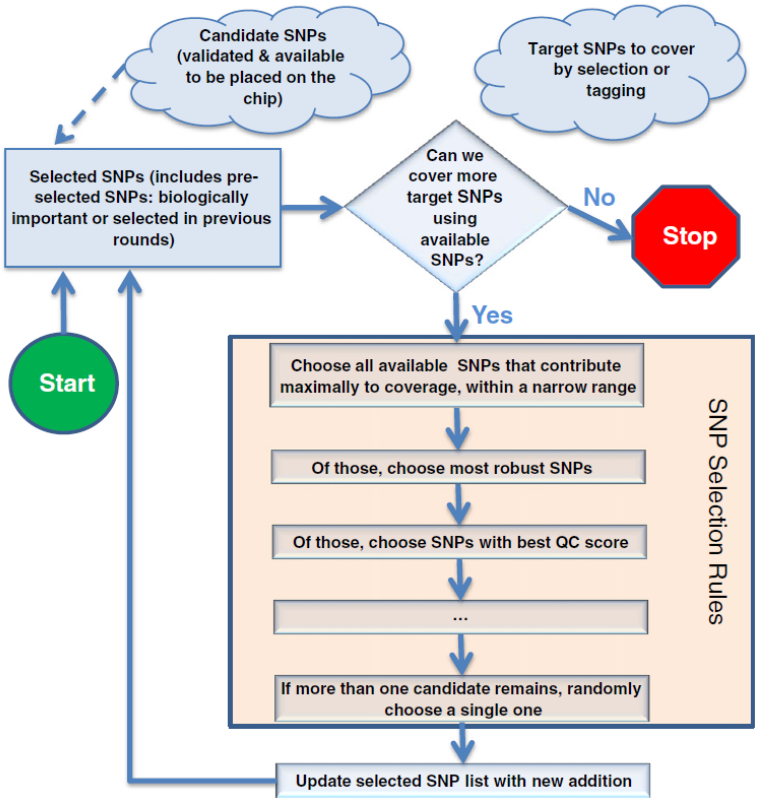
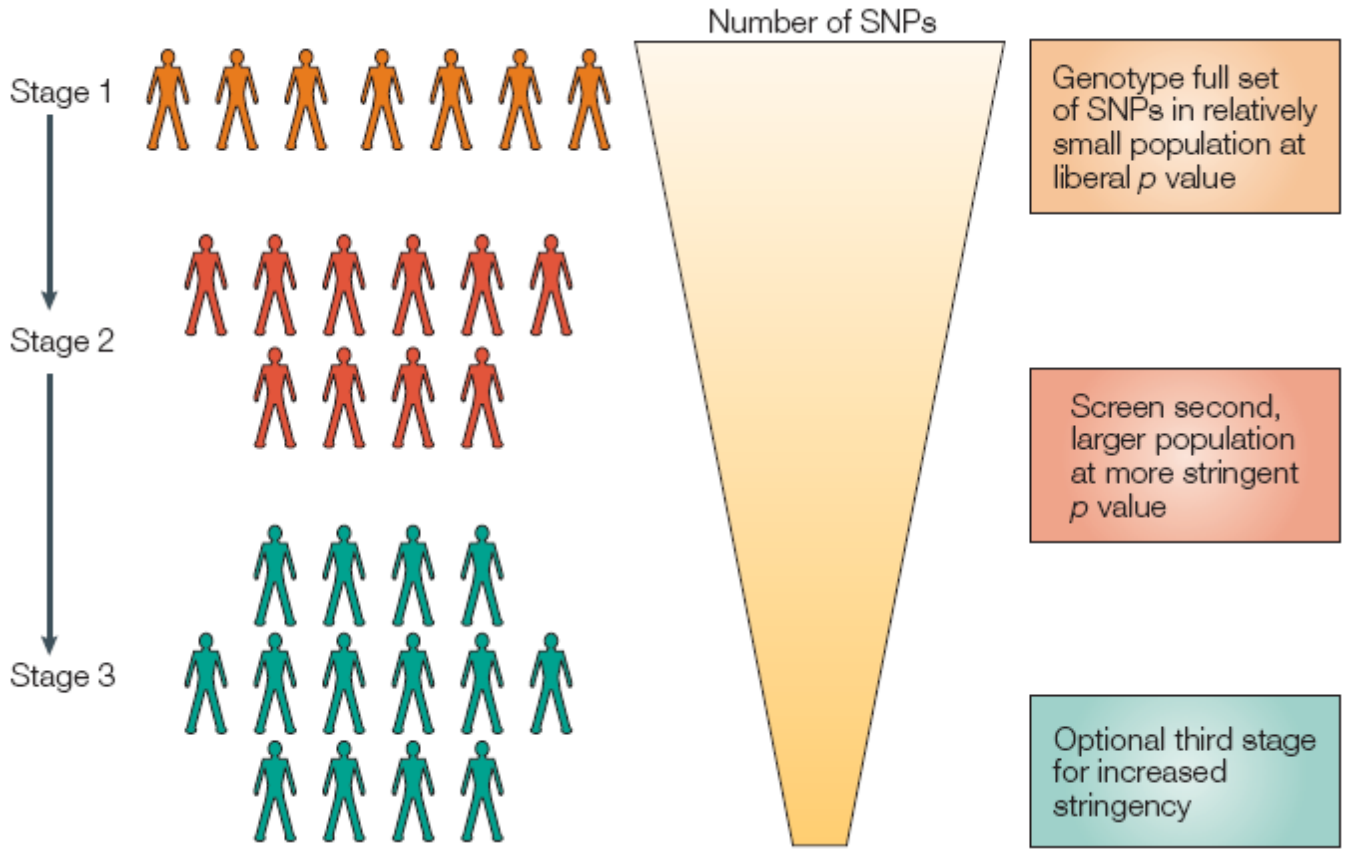


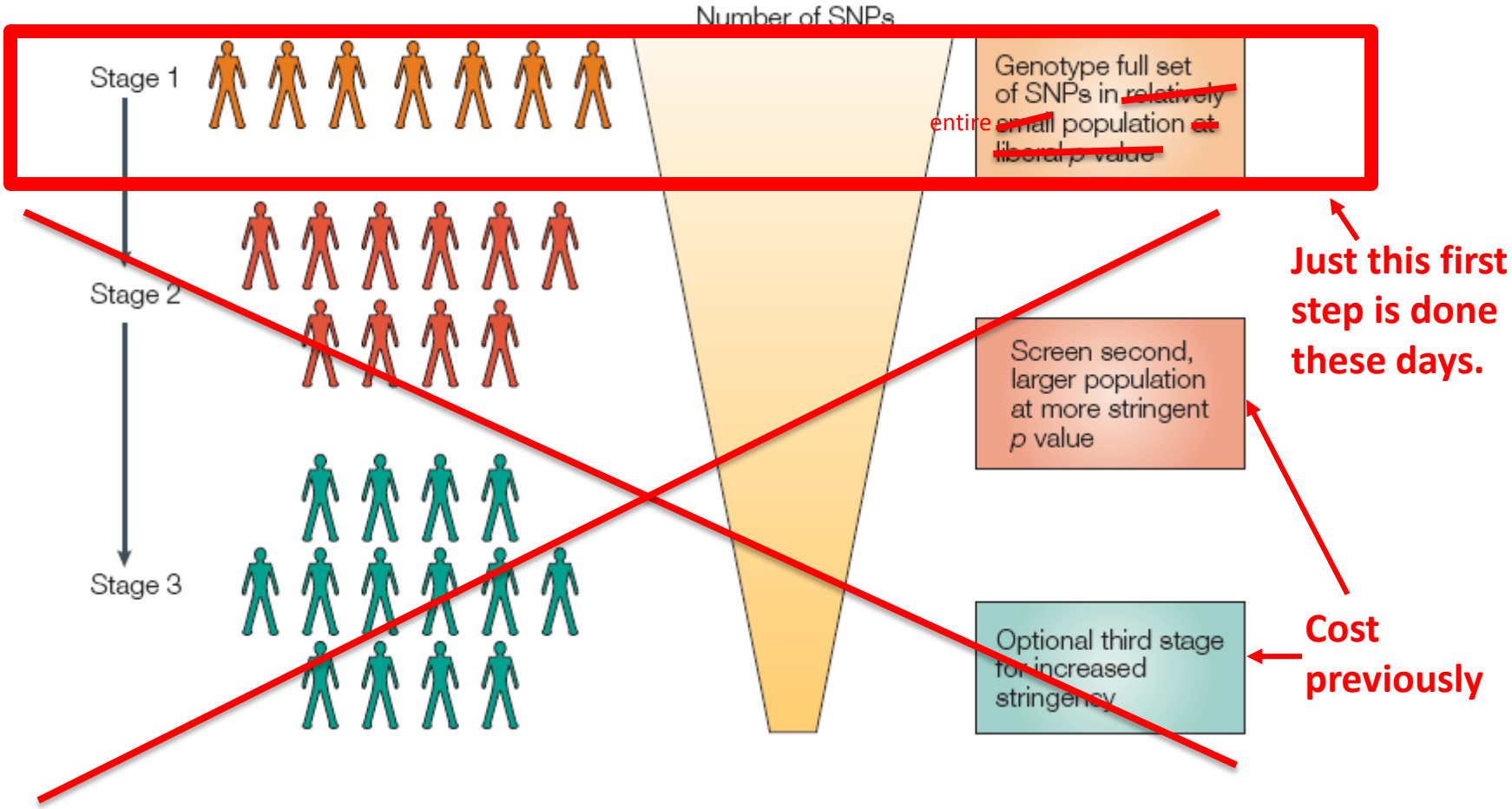
Fig. 7. Greedy SNP selection algorithm. A set of SNPs are chosen for reasons of biological importance, significance in published GWAS, etc., or as the result of previous rounds of greedy SNP selection. The "target" set of SNPs to be covered by tagging is established to fit the purpose of the current round of SNP selection, e.g., maximize coverage of SNPs in coding regions, or maximize general coverage of the genome. Then, SNPs which are available to be placed on the microarray are assessed for their ability to increase coverage of the target set. If coverage can be increased, a set of decision rules is applied to select the best single SNP to add to the selected list, as described in the text. This process continues until maximum coverage of the target set is achieved, or no space for additional SNPs on the microarray remains.

Hoffmann et al., Next generation genome-wide association tool. Design and coverage of a high-throughput European-optimized SNP array. Genomics, 2012.
Hoffmann et al., Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics, 2012.

Genome-wide Assocation Studies (GWAS)



Genome-wide Assocation Studies (GWAS)



Class Exercise!: How would you sequence a genome?

- Capture *every* base (or at least try to)…

Class Exercise!: How would you sequence a genome?

- Capture *every* base (or at least try to)…
- Can only sequence fragments of a certain size… (you can chop the genome into pieces…)

Class Exercise!: How would you sequence a genome?

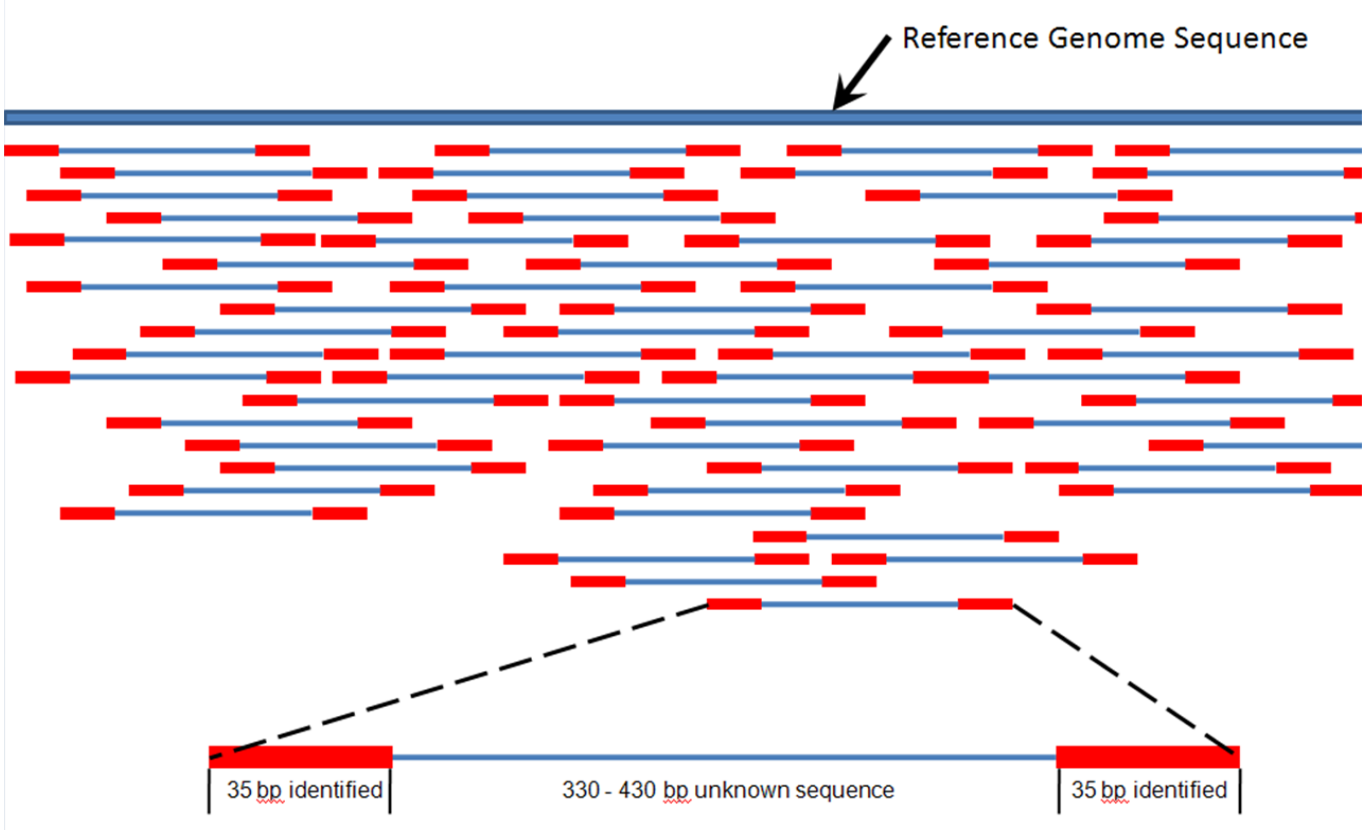
- Capture *every* base (or at least try to)…
- Can only sequence fragments of a certain size… (you can chop the genome into pieces…)
  - How many times should you sequence it?
  - Are some regions harder than others?

Class Exercise!: How would you sequence a genome?

- Capture *every* base (or at least try to)…
- Can only sequence fragments of a certain size… (you can chop the genome into pieces…)
  - How many times should you sequence it?
  - Are some regions harder than others?
- De novo, or from a reference?

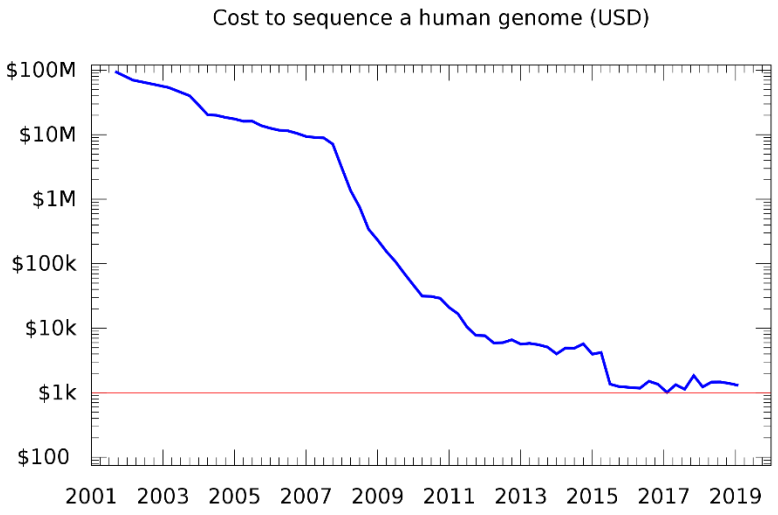
Design 2: Next generation sequencing

- Genotype essentially entire genome
- A bunch of ~random reads all over the genome are assembled (not trivial) to produce the genome



Design 2: Next generation sequencing

- Genotype essentially entire genome
- A bunch of ~random reads all over the genome are assembled (***not trivial***) to produce the genome
- Some regions of the genome have more reads than others; can choose how much you want to sequence (5x coverage 1000 Genomes, e.g.)
- More expensive
  - Old thought: Extreme phenotypes,
    - since can't do everyone?
  - New thought: Just sequence everyone.



Design choices

- **GWAS Microarray**
  - Only assay SNPs designed into array (0.7-5 million)
  - Cheaper (so more subjects tradeoff)

GWAS Sequencing

"De novo" discovery (particularly good for rare variants)

More expensive (but costs are falling) (many less subjects)

Need much more expansive IT support

Lots of interesting interpretation problems (field rapidly evolving)

Can also do a hybrid design (use sequencing on some to inform array design on rest; complicates analysis)

Design choices

- **Exome Microarray**
  - Only assay SNPs designed into array (~300K+custom); in **exons only** and that could affect protein coding function
  - Cheapest (so many more subjects)

Exome Sequencing

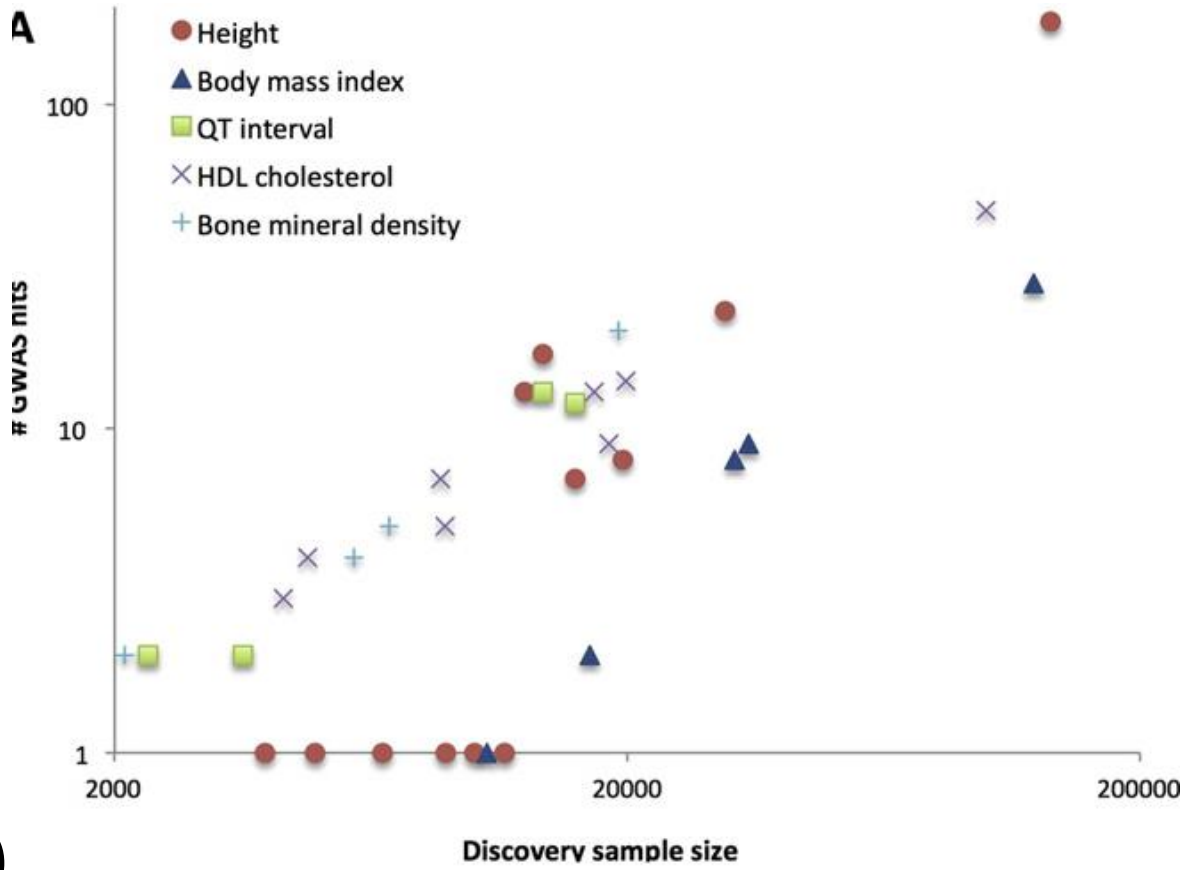
"De novo" discovery (particularly good for rare variants); **%age of exons only** (first step is a pull down that does not capture all exons)

More expensive than microarrays, less expensive than gwas sequencing

Need more expansive IT support, but not as much as whole genome

Size of study matters

- Large # SNPs tested – multiple comparisons
- Very small effect sizes
- Tag SNP, rather than actual SNP
- Consortiums
- (This is an old figure, but has certainly held up)



Visscher, AJHG 2012

Biggest studies...

- GWAS Microarray:
  - US:
    - Million Veteran's Program (1M in process, >300K now)
    - 100K people in the Kaiser RPGEH (Hoffmann et al., Genomics, 2011a&b)
    - 23&me, crude self report, but sample size can win for discovery!
  - UK Biobank: nearly 500K people
  - Other big groups:
- Whole Genome Sequencing: 1000 Genomes Project (2.5K), UK10K (~4K), TopMed (150K)
- Exome Sequencing: GO ESP (12,031 subjects, for exome microarray design), Geisinger, all of UKB planned (200K so far)

Outline for GWAS

- Review / Overview
- Design
- Analysis
  - QC
  - Prostate cancer example
  - Imputation
  - Replication & Meta-analysis

Why QC data?

nature medicine

Retraction | Published: 08 October 2019

Retraction Note: CCR5-Δ32 is deleterious in the homozygous state in humans

Xinzhu Wei & Rasmus Nielsen

Nature Medicine **25**, 1796(2019) | Cite this article

7795 Accesses | **2** Citations | **211** Altmetric | Metrics

The original article was published on 03 June 2019

Retraction to: *Nature Medicine* https://doi.org/10.1038/s41591-019-0459-6, Published online 3 June 2019.

The authors are retracting this Brief Communication to correct the scientific literature. Arising from exchanges with David Reich and colleagues, they have been made aware of a genotyping calling bias in the underlying UK Biobank data from which the main results of the study were drawn. Further analyses confirmed that the central finding of the study – that homozygous CCR5-Δ32 mutation is associated with increased mortality in the UK Biobank – is a result of this technical artifact. Because the main conclusion of

HIV resistance

QC Steps: Class exercise!

- Can you think about potential QC steps you might take at an individual and SNP level basis? Which one first?
- What tests have we talked about in this course?
- Special chromosomes?
- (Could families be problematic? useful?)

QC Steps

- "garbage rises to the top"
- Filter SNPs and Individuals
  - MAF (e.g., 1%, but very sample size dependent!)
  - Low call rates (means genotype probably didn't work correctly)

•Test for HWE among controls & within ethnic groups. Use conservative alpha-level (10^{-6}, e.g., but opinions vary) (again, to remove artifacts)

•Check for relatedness.

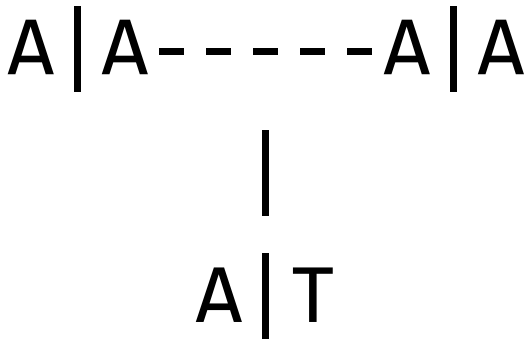
  - MZ twins, or accidentally genotyped same sample twice?
  - Remove first degree, or account for.

•Check genotype gender (mislabeled samples)

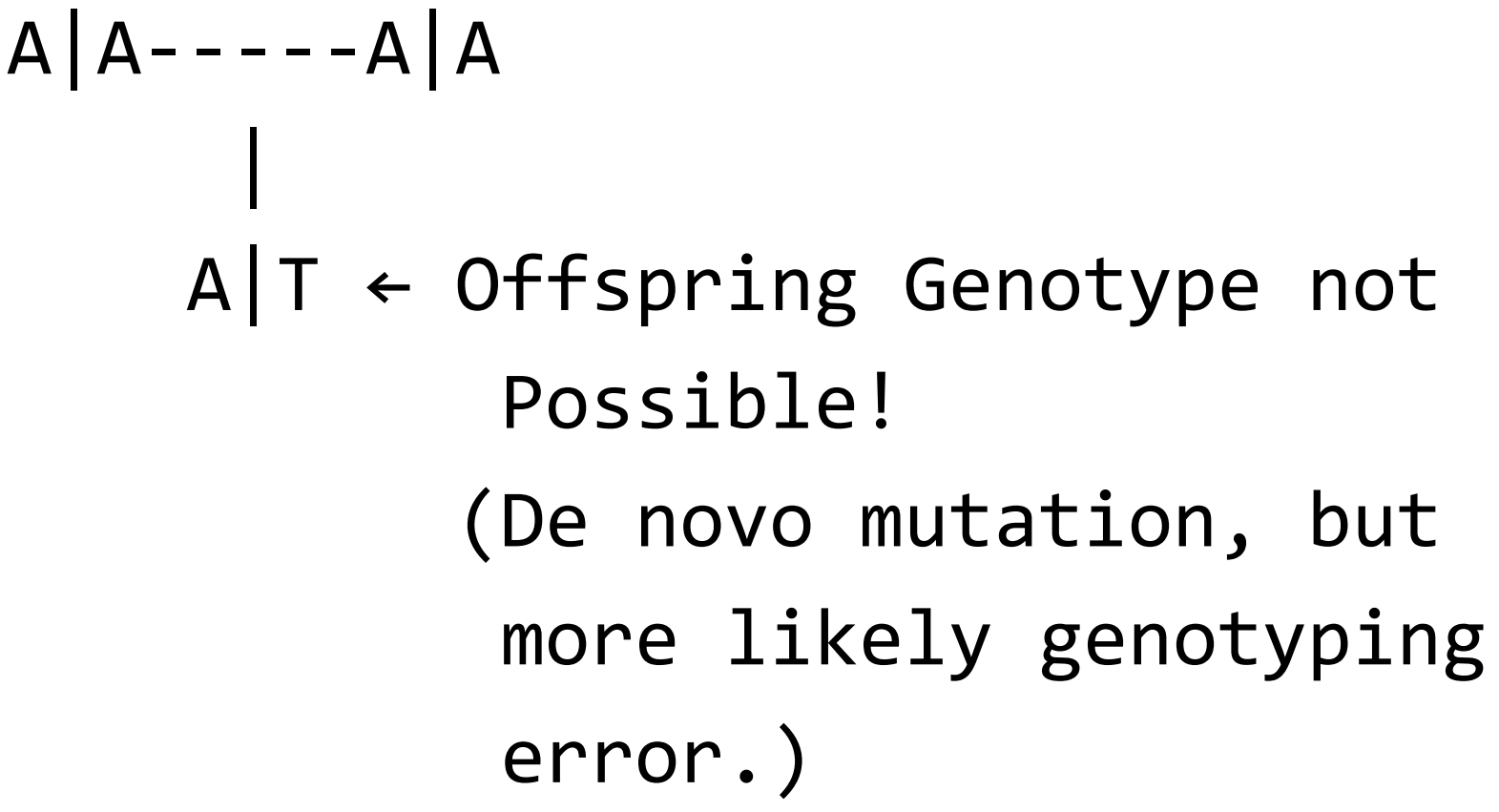
•Filter Mendelian inhertance (family-based, or potentially cryptics, if large enough sample)

•plink sofware: pngu.mgh.harvard.edu/~purcell/plink/reference.shtml

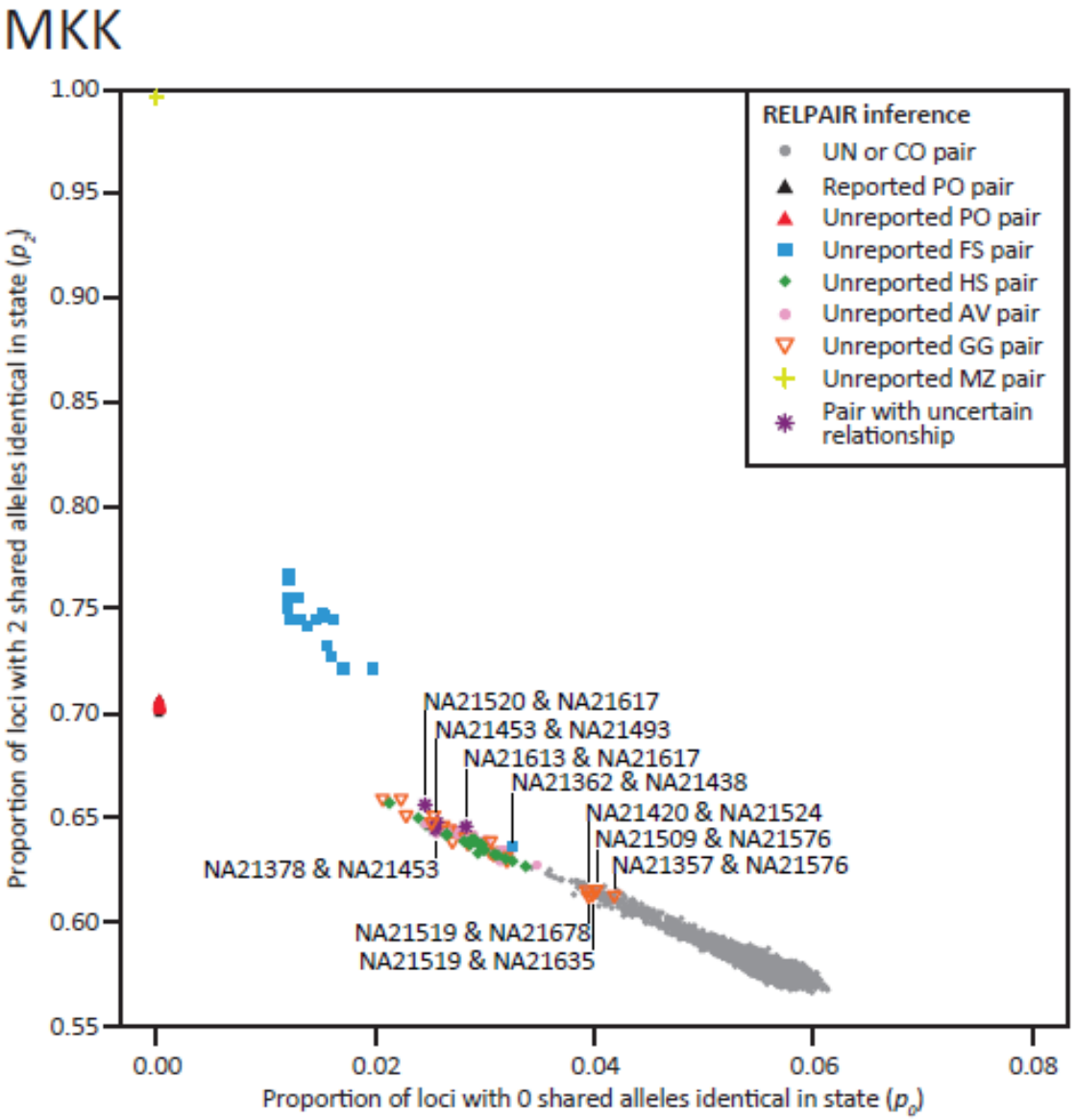
Filter for Mendelian inheritance



Filter for Mendelian inheritance



Check for relatedness, e.g., HapMap



- Take overall average of all SNPs of how many alleles are shared.
- E.g., parent-offspring never shares zero alleles.
- HapMap was supposed to be unrelateds (this was a bit of a surprise!)

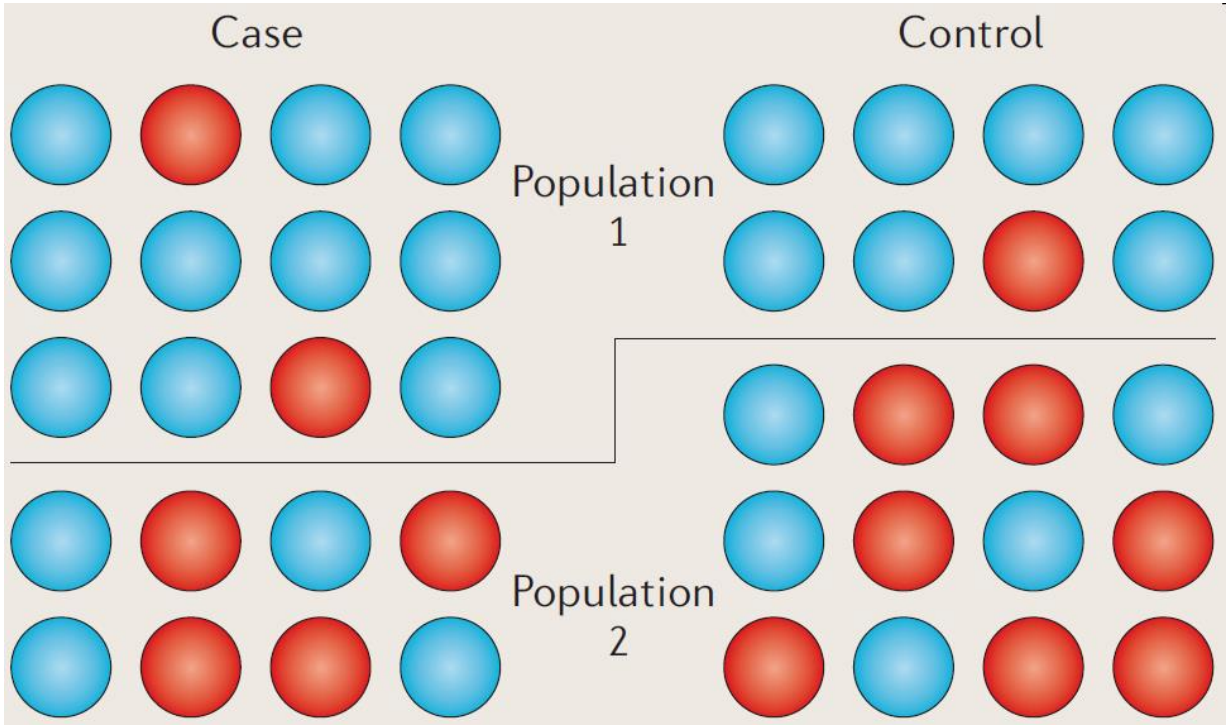
Pemberton et al., AJHG 2010

GWAS analysis

- Most common approach: Model each SNP separately
- Additive coding of SNP most common, just a covariate in a regression framework
- Dichotomous phenotype: logistic regression
- Continuous phenotype: linear regression
- Correct for multiple comparisons
  - e.g., Bonferroni, 1 million gives $\alpha=5x10^{-8}$
  - [Often use $5x10^{-8}$ for larger GWAS arrays (or imputed SNPs, more later) also, SNPs are correlated, so Bonferroni is a little too conservative…]
- Adjust for potential population stratification (details next slides…)
  - principal components (PC's), on best performing SNPs
  - software usually does LD filter (e.g., Eigensoft)

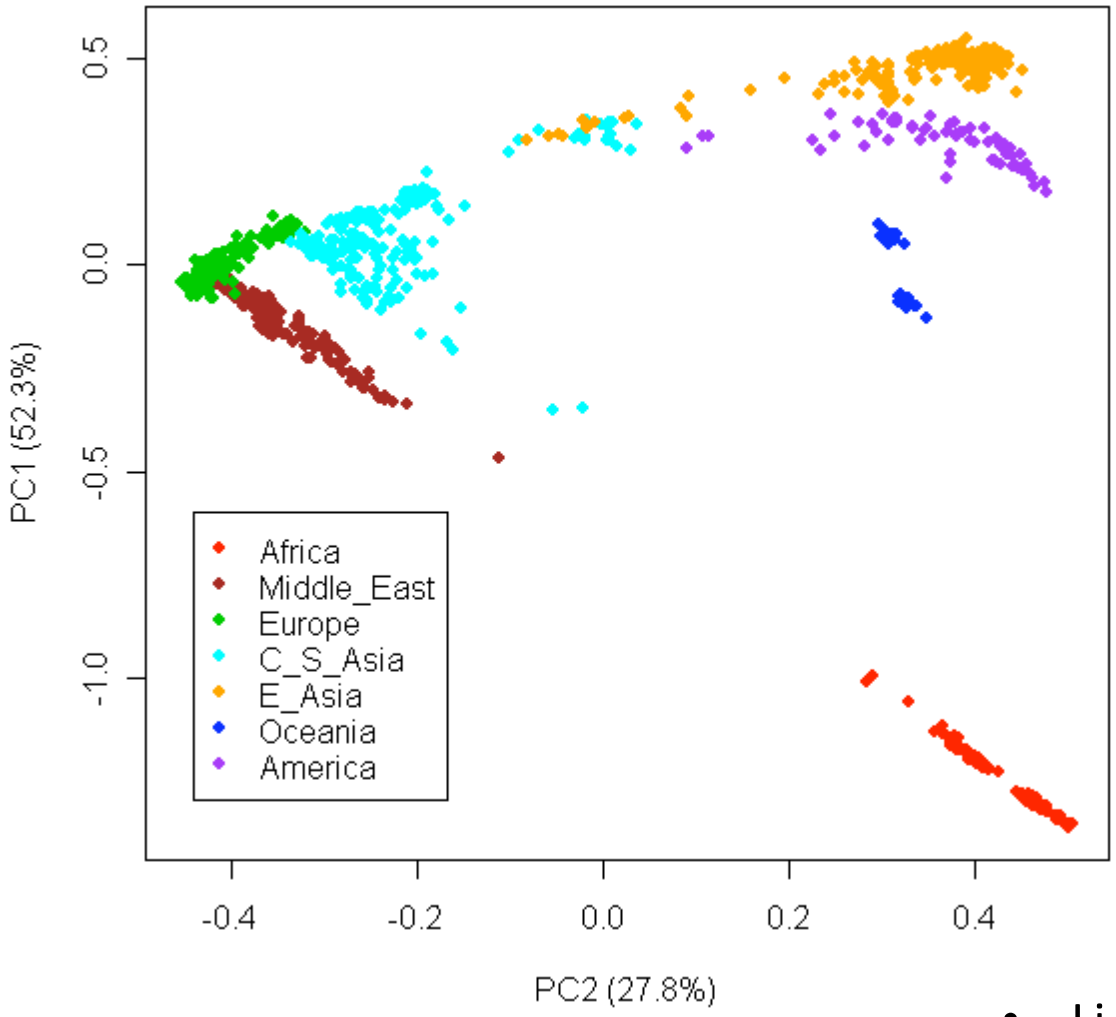
Recap of population substructure

- Two populations have different disease frequency, and different allele frequency (Recall Pima & Papago example).
- Association picks up they are different populations!



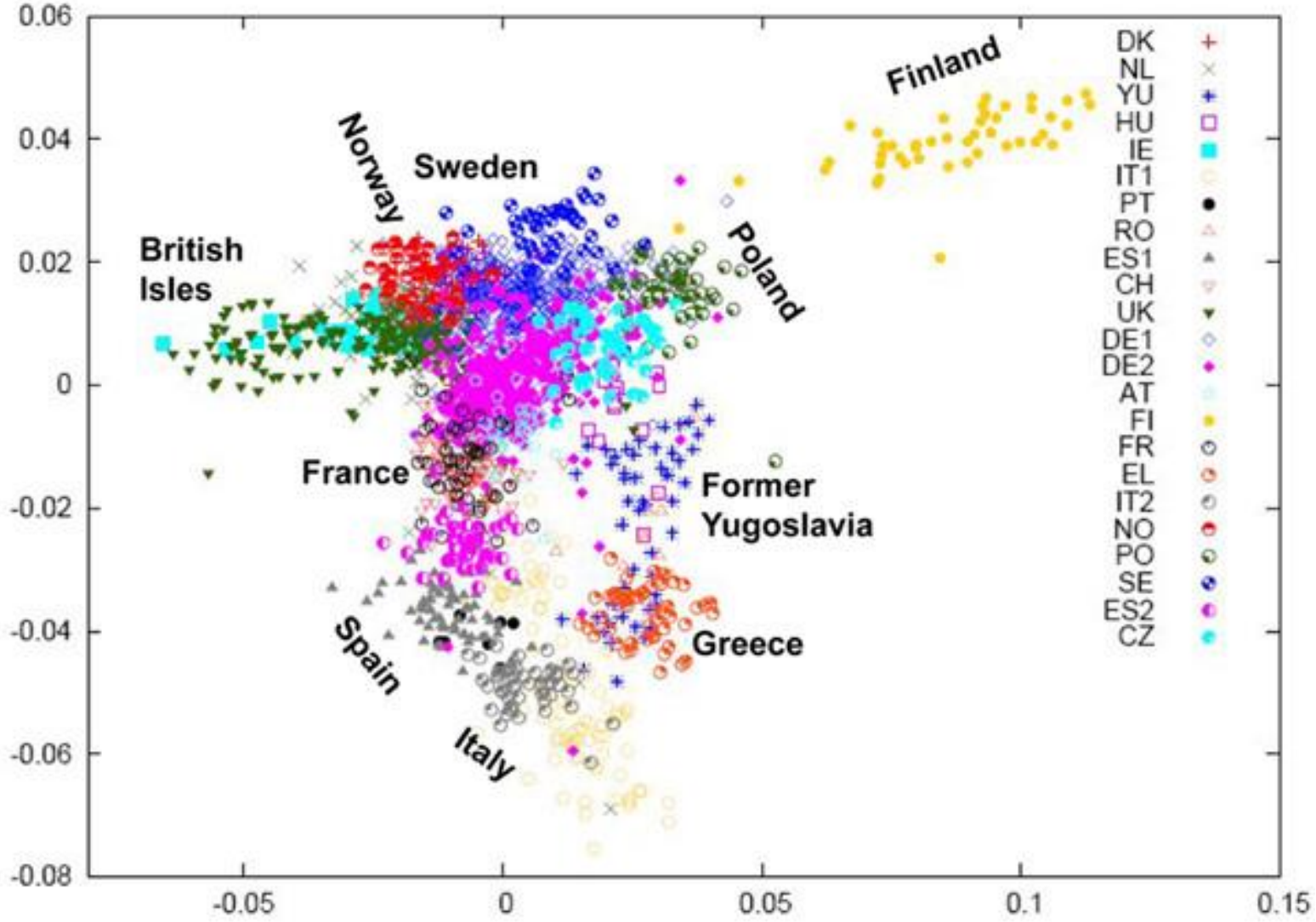
Balding, Nature Reviews Genetics 2010

Control for race/ethnicity/ancestry: Principal Components (PCs) as covariates in regression model



- Li et al., Science 2008

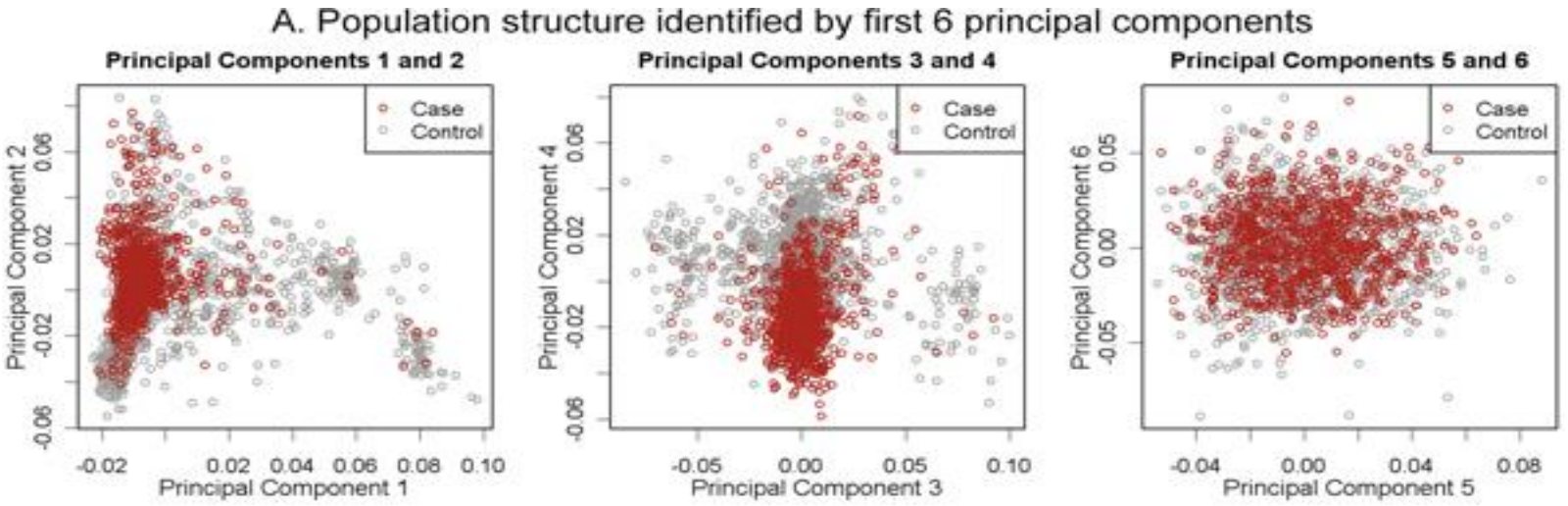
PCs pick up fine population structure

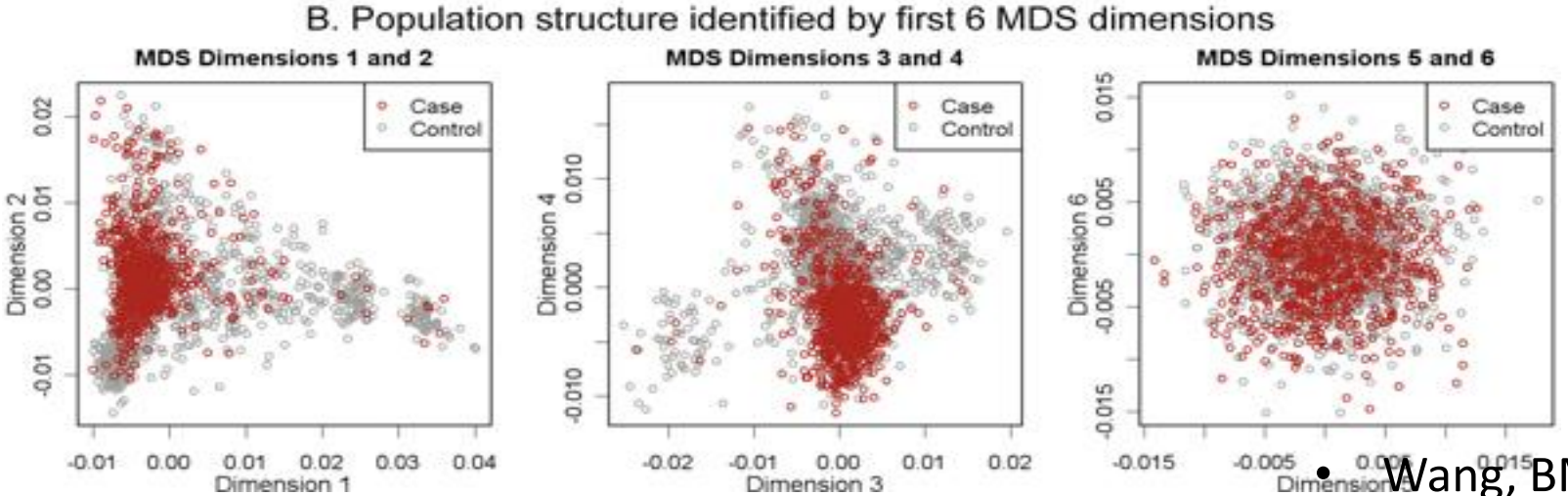


- Razib, Current Biology 2008

Adjusting for PC's

Do not want separate cluster for cases as controls! – Why want individuals from similar population for controls (and randomization). [These look reasonable, but if see a separation…]





- Wang, BMC Proc 2009

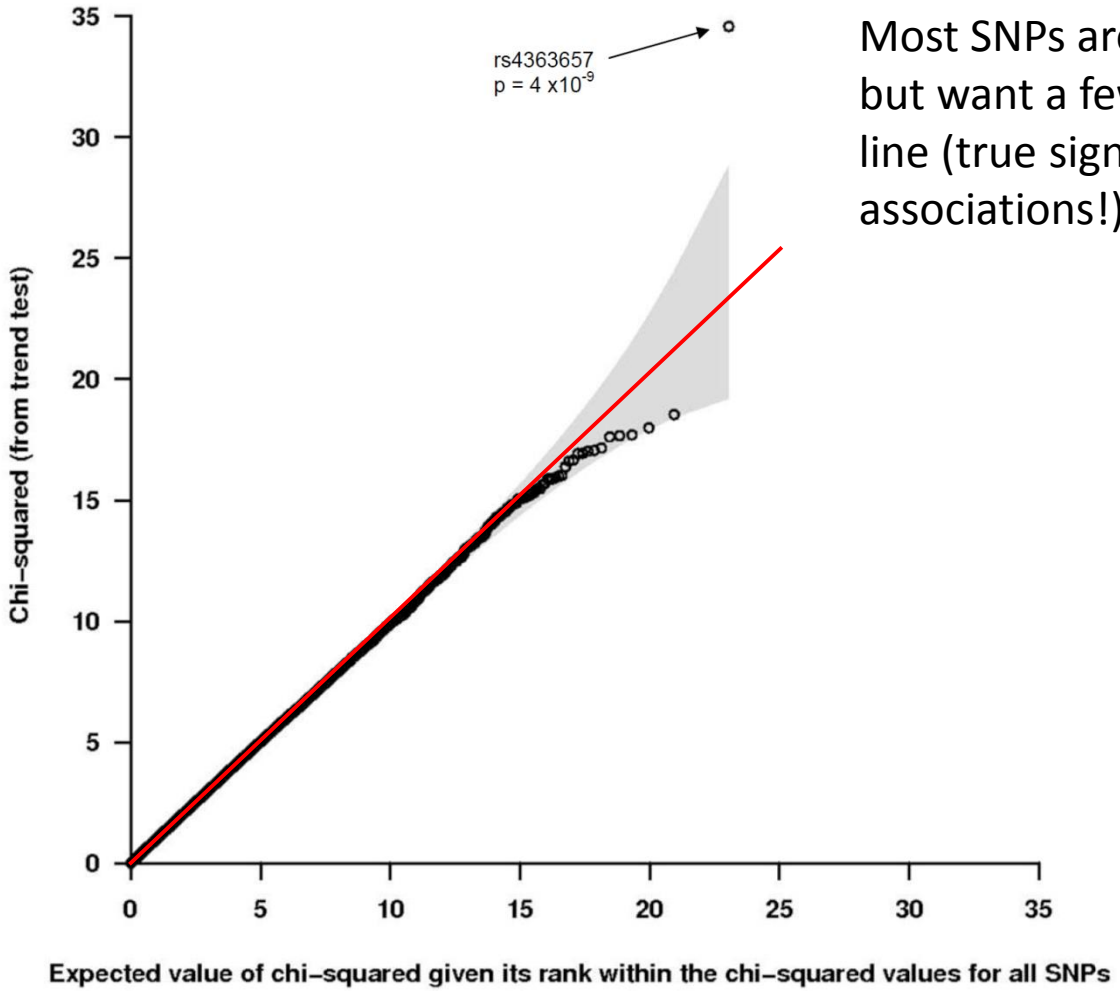
Instead of adjusting for PCs, adjust for the rest of the genome (mixed models)

- (1) Increase power by jointly modelling all variants, and (2) adjust for population structure
- Adjust for genetic relatedness between all pairs of individuals (kinship matrix)
  - This is huge, need approximations to handle this in modern cohorts.
  - In practice -- leave one chromosome out analysis
    - E.g., chromosome 1, adjust for kinship based on chromosome 2-22
    - Don’t really want to adjust for variants immediately nearby (LD, proximal contamination)
- Software: Emmax, GEMMA, BOLT-LMM (fast)

After run analysis

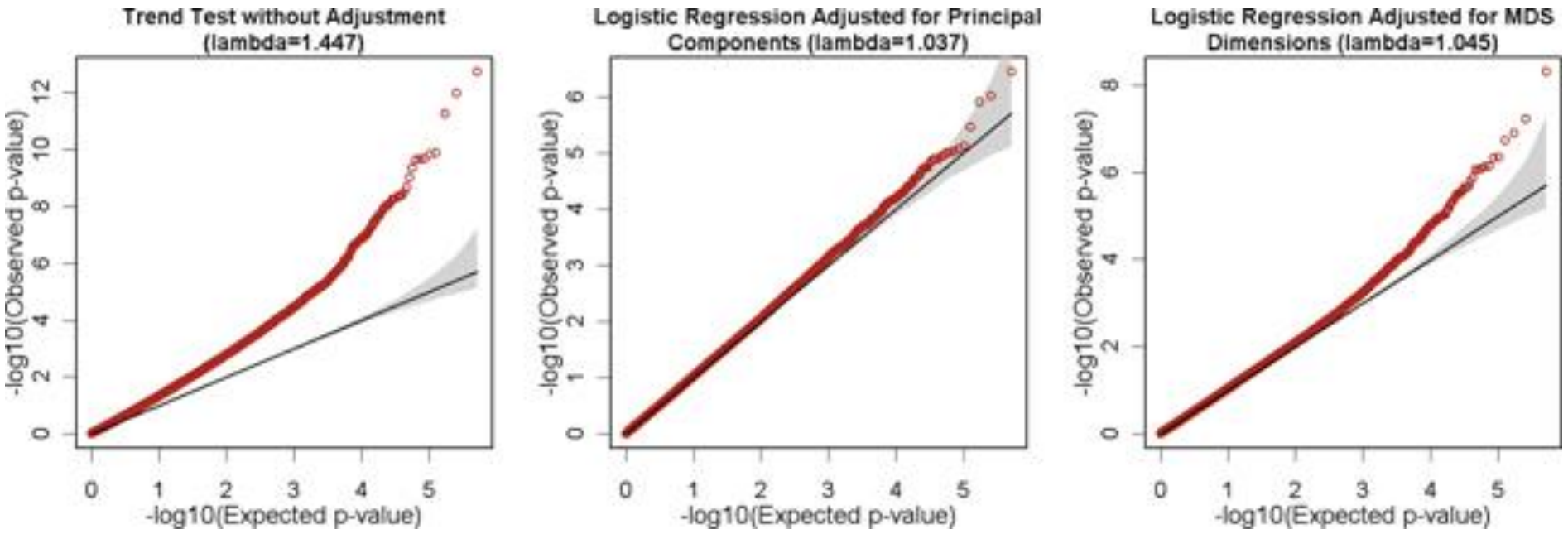
- Assess the fit with Q-Q plot (saw last lecture)
- Look at results in Manhattan plots
- **Replication** of hits in a separate cohort

Quantile-quantile (Q-Q) plot review



Most SNPs are on the line, but want a few hits off the line (true significant associations!)

Q-Q plots and PC adjustment recap



- Exactly on line if no signal at all.
- If don't adjust for PC's, then p-values are all inflated artificially.
- Adjusting for PC's fixes the problem.

Wang, BMC Proc 2009

Manhattan plot example: Kaiser GWAS of Prostate Cancer

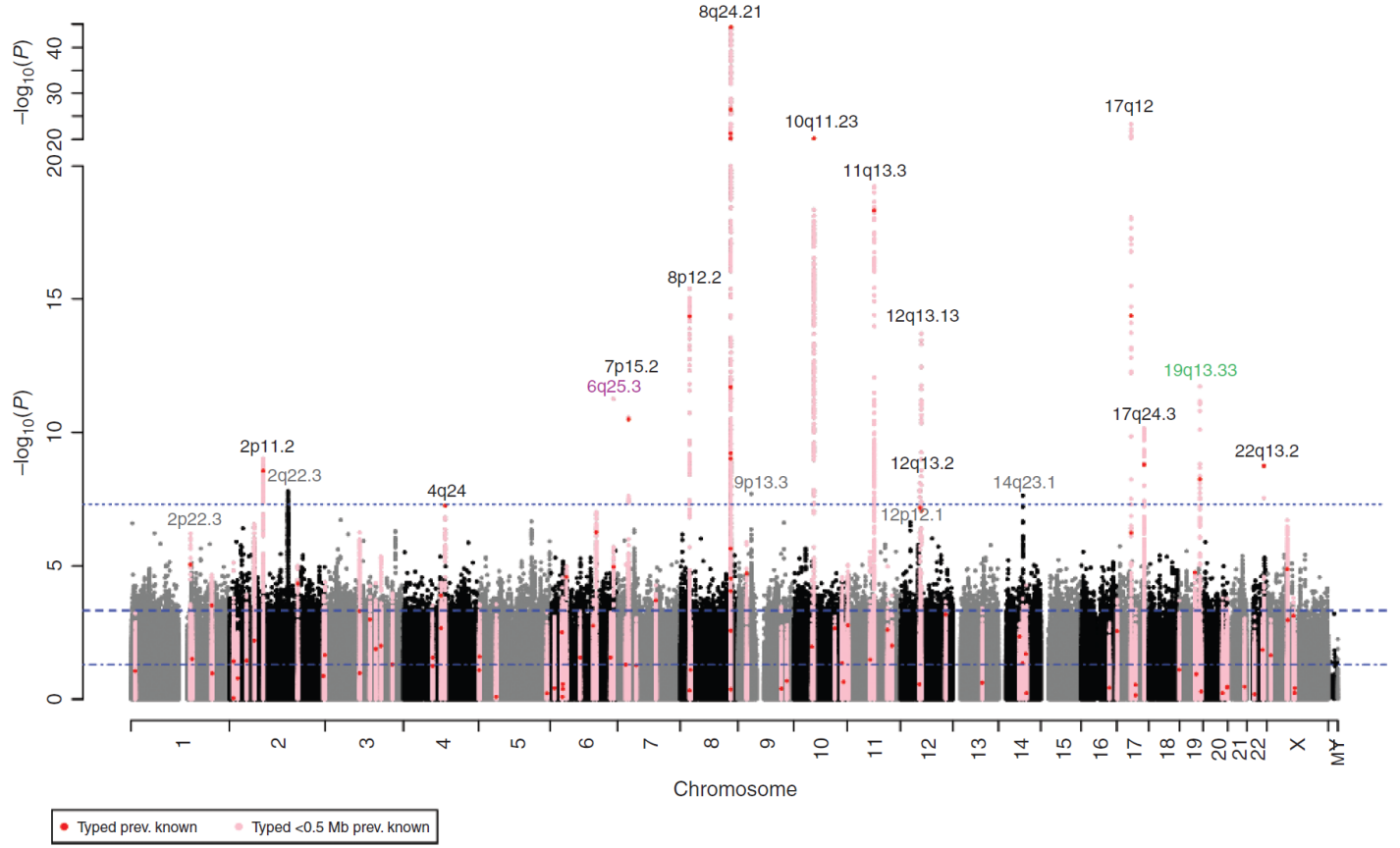
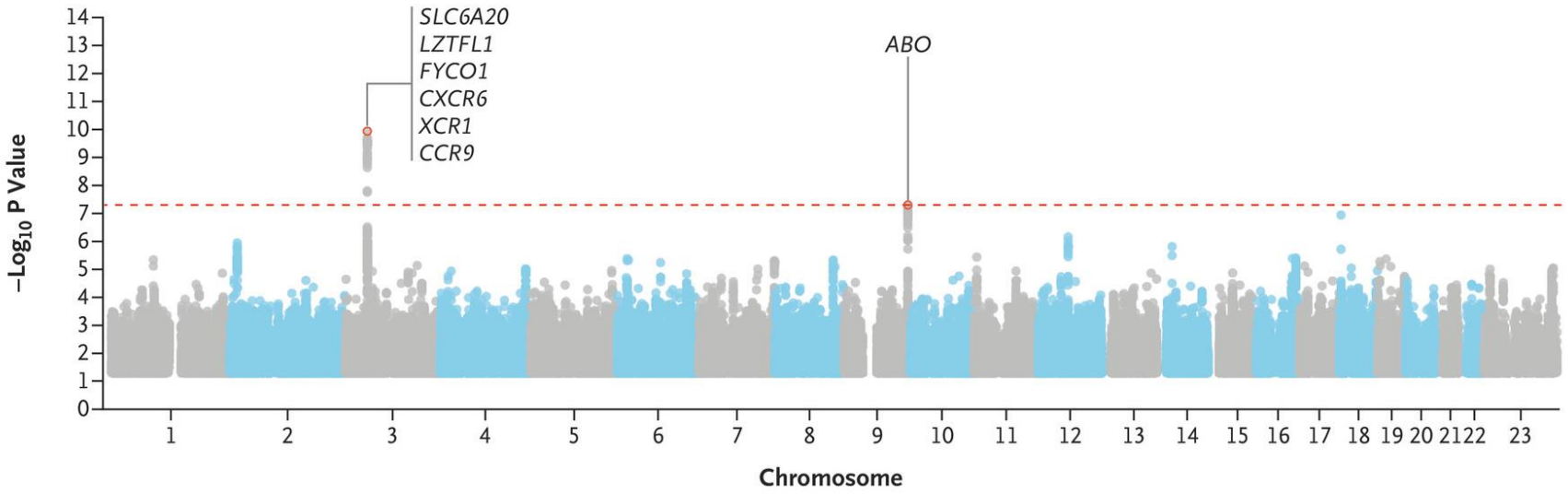


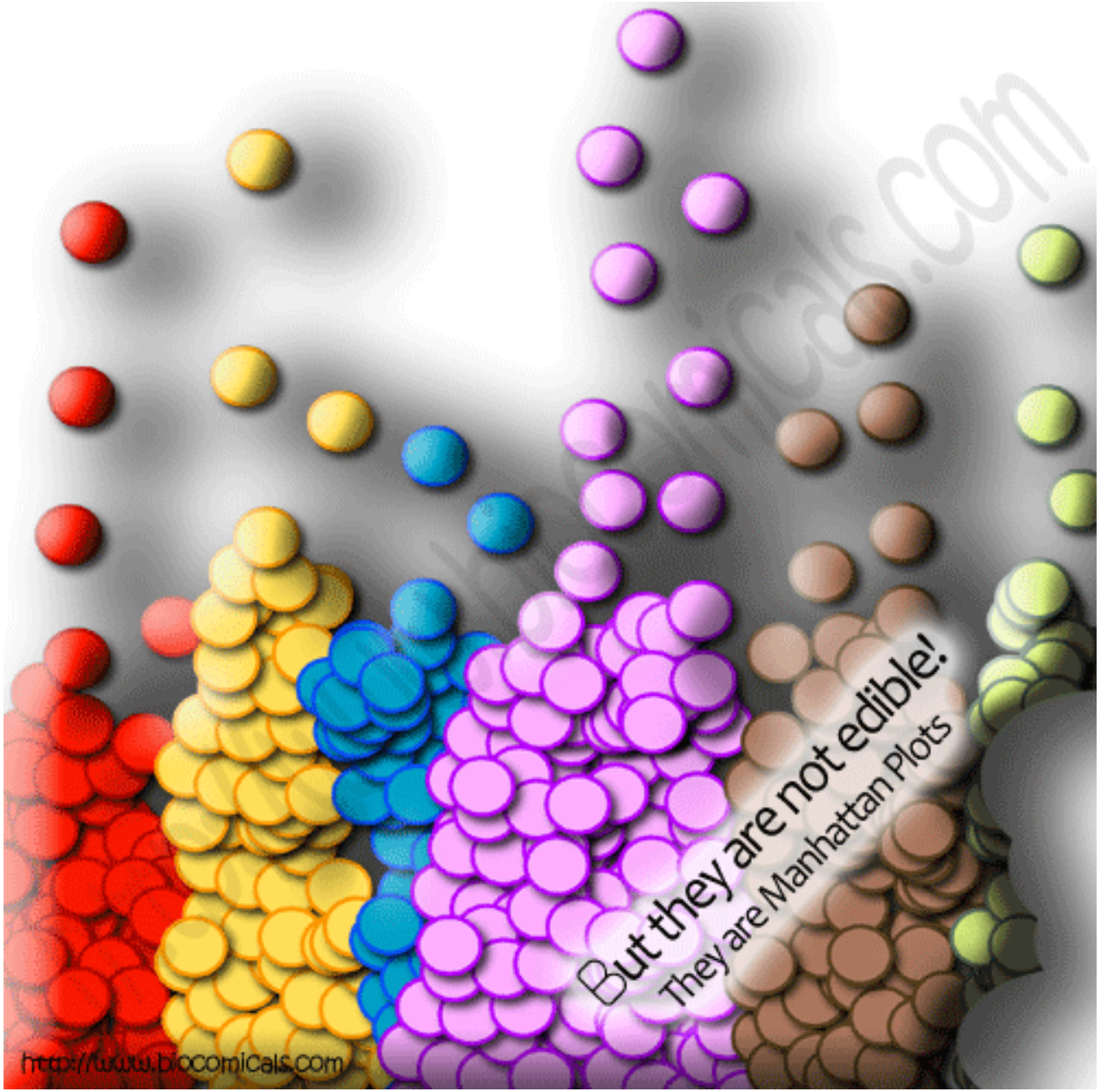
Figure 1. Results from a GWAS of prostate cancer in the KP population (8,399 cases and 38,745 controls), highlighting key chromosomal regions. *P* values are for variant associations with prostate cancer from transethnic fixed-effects meta-analysis of four race/ethnicity GWAS (non-Hispanic white, Latino, African-American, East Asian), each adjusted for age and ancestry principal components. Horizontal dashed lines indicate genome-wide statistical significance ($P < 5 \times 10^{-8}$), Bonferroni-corrected significance for 105 known prostate cancer risk variants ($P < 0.05/105$), and nominal significance ($P < 0.05$). Red points denote our findings for the 105 known risk SNPs, and pink indicates results for SNPs within a 0.5 Mb window around these SNPs. The new 6q25.3 indel rs4646284 is colored magenta, and the 19q13.33 prostate cancer or PSA SNP rs2659124 is noted in green. Those loci containing previously reported variants that we replicated at genome-wide significance are noted in black text. Loci with variants that were novel in the KP GWAS but failed to replicate are noted in gray text.

Hoffmann et. Al, Cancer Discovery 2015

Manhattan plot example: Covid-19



The severe covid-19 GWAS group. Genomewide association study of severe covid-19 with respiratory failure. NEJM, 2020; 383:1522-1534.


But they are not edible!
They are Manhattan Plots
http://www.biocomicals.com

Replication, Replication, Replication

- To replicate:
  - Association test for replication sample significant at 0.05/{Number of SNPs replicating} alpha level
  - Same genetic model (e.g. additive, dominant)
  - Same direction
  - Sufficient sample size for replication!
    - [e.g., do a power calculation]
- Non-replications not necessarily a false positive
  - LD structures, different populations (e.g., flip-flop)
  - covariates, phenotype definition, underpowered
  - Winner's curse

Meta-analysis

- Combine multiple studies to increase power
- Either combine p-values (Fisher's test),
- or coefficient estimates + standard error (better)

Replication & Meta-analysis

Table 2. Prostate cancer genome-wide association study results for new risk indel rs4646284 at 6q25.3 and for risk SNP rs2659124 at 19q13.33

A

Variant	Risk allele	Ref. allele	Group	Risk AF	No conditioning		Conditioning on other 6q25.3 SNPs		
					OR (95% CI)	*P*	OR (95% CI)	*P*	r^2_{info}
Indel: rs4646284	TG (indel)	T	KP NH white	0.298	1.17 (1.12–1.23)	6.7×10^{-11}	1.16 (1.10–1.23)	3.3×10^{-7}	0.91
			KP Latino	0.250	1.13 (0.94–1.36)	0.20	1.06 (0.85–1.31)	0.61	0.89
			KP Asian	0.324	1.03 (0.84–1.27)	0.76	0.91 (0.68–1.23)	0.56	0.89
			KP Af-Amer	0.365	1.18 (1.02–1.37)	0.029	1.21 (1.03–1.42)	0.020	0.85
			KP Meta FE	—	1.17 (1.12–1.22)	5.5×10^{-12}	1.15 (1.09–1.21)	8.5×10^{-8}	—
			KP Meta RE	—	1.17 (1.12–1.22)	5.5×10^{-12}	1.15 (1.07–1.22)	7.7×10^{-5}	—
			MEC	0.361	1.13 (1.03–1.25)	0.0094	1.13 (1.03–1.25)	0.014	0.81
			PEGASUS	0.278	1.27 (1.17–1.37)	1.4×10^{-8}	1.20 (1.09–1.32)	0.00024	0.80
			Overall Meta FE	—	1.18 (1.14–1.22)	1.0×10^{-19}	1.16 (1.11–1.21)	5.4×10^{-12}	—
			Overall Meta RE	—	1.18 (1.13–1.23)	9.8×10^{-17}	1.16 (1.11–1.21)	5.4×10^{-12}	—

Hoffmann et. Al, Cancer Discovery 2015`

Replication and Meta-analysis

Table 2. Susceptibility Loci Associated with Severe Covid-19 with Respiratory Failure.*

Chromosome and Analysis	Meta-analysis		Italian Panel				Spanish Panel			
	P Value	Odds Ratio (95% CI)	P Value	Odds Ratio (95% CI)	Allele Frequency		P Value	Odds Ratio (95% CI)	Allele Frequency	
					patient	control			patient	control
3p21.31†										
Main analysis	1.15×10^{-10}	1.77 (1.48–2.11)	1.98×10^{-7}	1.74 (1.27–2.38)	0.14	0.09	1.32×10^{-4}	1.85 (1.50–2.28)	0.09	0.05
Analysis corrected for age and sex	9.46×10^{-12}	2.11 (1.70–2.61)	7.02×10^{-8}	1.95 (1.53–2.48)	0.14	0.09	1.17×10^{-5}	2.79 (1.76–4.42)	0.09	0.05
9q34.2‡										
Main analysis	4.95×10^{-8}	1.32 (1.20–1.47)	2.90×10^{-6}	1.37 (1.20–1.57)	0.42	0.35	3.55×10^{-3}	1.26 (1.08–1.48)	0.42	0.35
Analysis corrected for age and sex	5.35×10^{-7}	1.39 (1.22–1.59)	5.31×10^{-5}	1.37 (1.17–1.60)	0.42	0.35	2.81×10^{-3}	1.45 (1.13–1.84)	0.42	0.35

* The meta-analysis included 1610 patients and 2205 control participants; the Italian analysis, 835 and 1255, respectively; and the Spanish analysis, 775 and 950, respectively. Allele frequencies of the minor or risk allele (see below) are shown among the patients and the control participants. All the association test statistics were adjusted for the top 10 principal components from the principal-component analysis. Two analyses were performed: a main analysis, which was corrected for 10 principal components, and an analysis that was corrected for age and sex in addition to 10 principal components. In the analyses that were corrected for age and sex, 25 control participants were excluded from the Spanish analysis and the meta-analysis because of missing covariate data. The P values and corresponding odds ratios and 95% confidence intervals (CIs) are shown with respect to the minor allele. Association results for the recessive and heterozygous models for both meta-analyses (main and corrected for age and sex) are shown in Supplementary Appendix 3. Covid-19 denotes coronavirus disease 2019.

† For chromosome 3p21.31, the association boundaries for each index single-nucleotide polymorphism (SNP; see the Supplementary Methods section), with the genomic positions retrieved from genome build hg38, were chr3:45800446 through 46135604. The Single Nucleotide Polymorphism database (dbSNP) identifier was rs11385942 (the rs identifier from the National Center for Biotechnology Information, rs11385942, is annotated as chr3:45834968 through 45834969:AAA:AA in dbSNP, version 153, and as chr3:45834967:GA:G in the TransOmics for Precision Medicine [TOPMed] imputation reference panel). The SNP rs11385942 was imputed according to TOPMed with high confidence (TOPMed estimated imputation accuracy, $R^2=0.94$ and $R^2=0.95$ for the Italian and Spanish panels, respectively) (Supplementary Appendix 2).The minor or risk allele was GA, and the major allele was G. The key genes (i.e., the candidate genes in the region) were *SLC6A20*, *LZTFL1*, *FYCO1*, *CXCR6*, *XCR1*, and *CCR9*.

‡ For chromosome 9q34.2, the association boundaries for each index SNP, with the genomic positions retrieved from genome build hg38, were chr9:133257521 through 133279871. The SNP rs657152 was genotyped according to the Global Screening Array (GSA) in the Italian and Spanish panels (Supplementary Appendix 2). The minor or risk allele was A, and the major allele was C. The key gene was *ABO*.

The severe covid-19 GWAS group. Genomewide association study of severe covid-19 with respiratory failure. NEJM, 2020; 383:1522-1534.

Beyond Genotyped SNPs: Imputation of SNP Genotypes

- Combine data from different platforms (e.g., Affy & Illumina) (for replication / meta-analysis).
- Estimate unmeasured or missing genotypes.
- Based on measured SNPs and external info (e.g., haplotype structure of HapMap).
- Increase GWAS power (impute and analyze all), e.g. Sick sinus syndrome, most significant was 1000 Genomes imputed SNP (Holm et al., Nature Genetics, 2011)
- HapMap as reference, now 1000 Genomes Project; UK10K, Haplotype reference consortium?
- Hybrid sequencing/genotyping approaches

Generalization of Tag SNPs: Imputed SNPs using a Reference Panel

- Generalization of pairwise r^2 of tagSNPs to multi-marker correlation

Study sample

....A.....A....A...
....G.....C....A...

Reference haplotypes

CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC

Class Exercise!

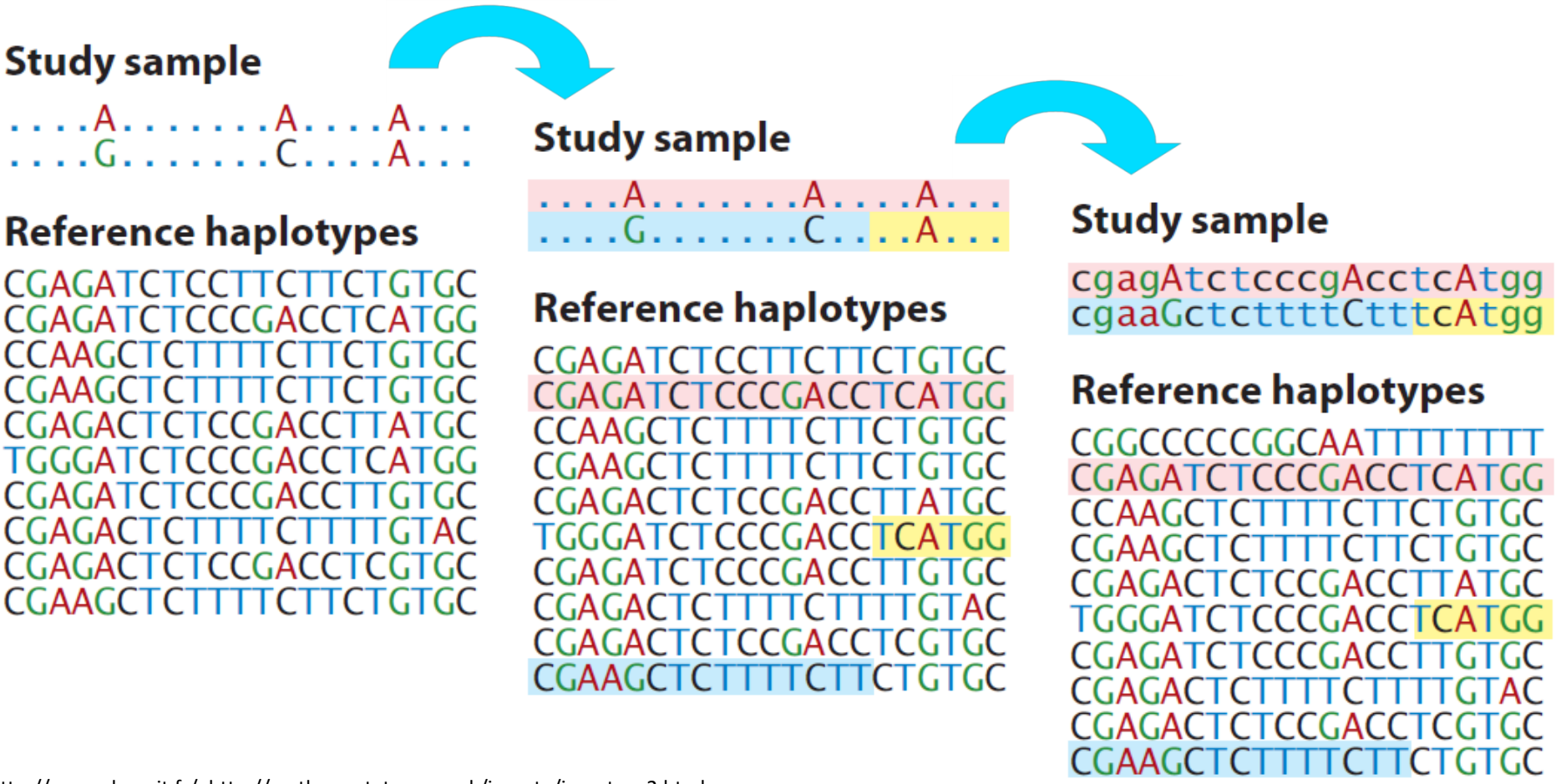
How would you fill in the remaining genotypes of this study sample individual, given the reference haplotypes?

- http://www.shapeit.fr/, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://faculty.washington.edu/browning/beagle/beagle.html
- http://www.sph.umich.edu/csg/abecasis/MACH/download/

Li et al., Ann Rev Human Genet, 2009

Generalization of Tag SNPs: Imputed SNPs using a Reference Panel

- Generalization of pairwise r^2 of tagSNPs to multi-marker correlation

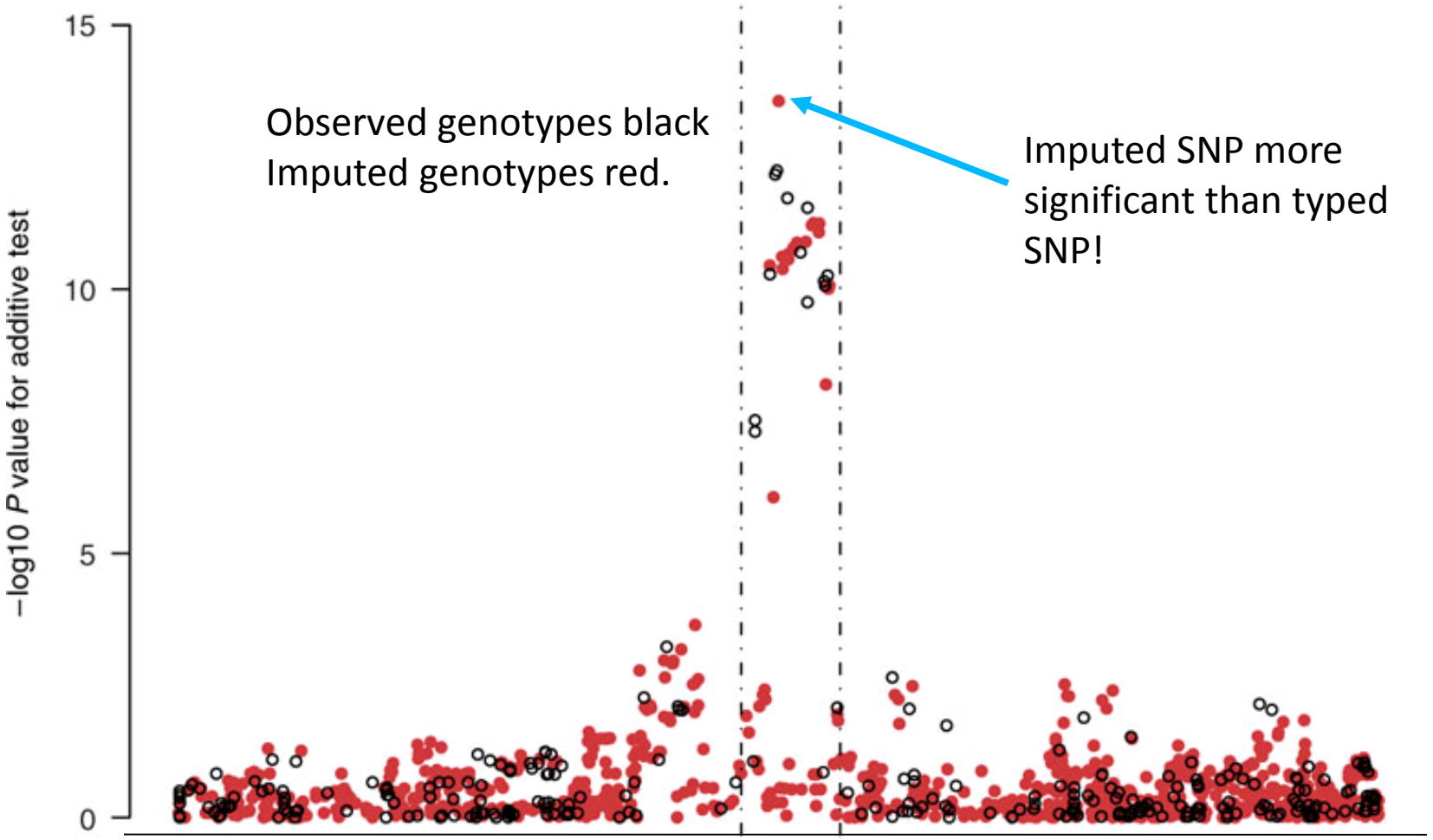


- http://www.shapeit.fr/, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://faculty.washington.edu/browning/beagle/beagle.html
- http://www.sph.umich.edu/csg/abecasis/MACH/download/

Li et al., Ann Rev Human Genet, 2009

Imputation Application

TCF7L2 gene region & T2D from the WTCCC data



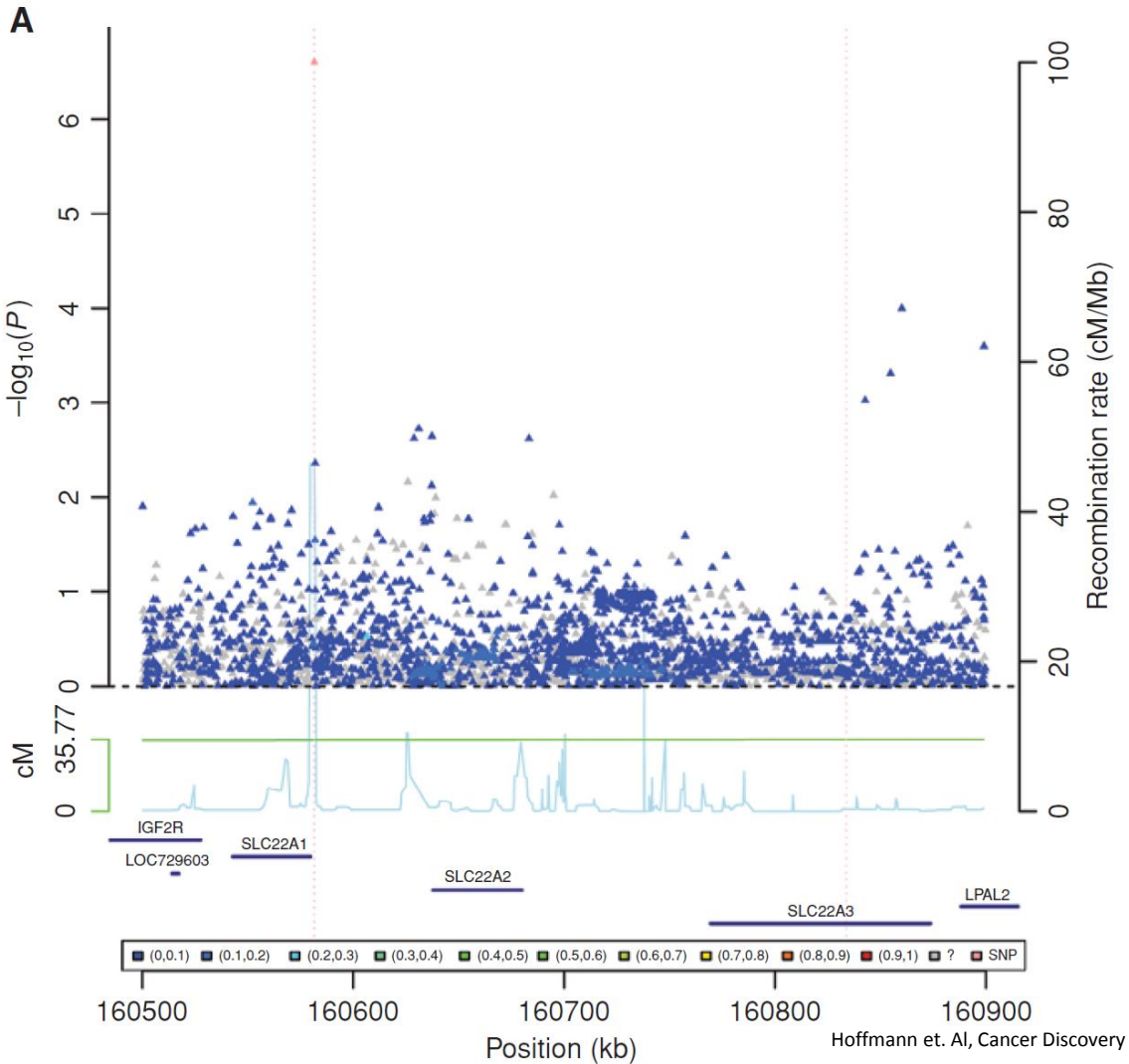
Marchini Nature Genetics2007

http://www.stats.ox.ac.uk/~marchini/#software

Imputation Application #2 – Do not always see a jet of points, even for common SNP…

Figure 2. Regional fixed-effects meta-analysis plots from GWAS in KP population of the two risk variants that replicated: (**A**) 6q25.3 (rs4646284), (**B**) 19q13.33 (rs2659124). The color code for the points represents the r^2 of each SNP with the risk variant (ranges defined in the legend). The dotted vertical lines pass through the risk variants and the other significant SNPs in the region, on which analyses were conditioned.

(Same SNP as 4 slides ago)



Hoffmann et. Al, Cancer Discovery 2015`

Imputation example: How rare can you go?

- G84E example talked about in Austin pg. 62
- MAF 0.0034 in European ancestry. r^2=0.57.
- In general, a moving target, still unknown how rare is possible

PLOS | GENETICS

RESEARCH ARTICLE

Imputation of the Rare *HOXB13* G84E Mutation and Cancer Risk in a Large Population-Based Cohort

Thomas J. Hoffmann[1,2], Lori C. Sakoda[3], Ling Shen[3], Eric Jorgenson[3], Laurel A. Habel[3], Jinghua Liu[2], Mark N. Kvale[2], Maryam M. Asgari[3], Yambazi Banda[2], Douglas Corley[3], Lawrence H. Kushi[3], Charles P. Quesenberry Jr.[3], Catherine Schaefer[3], Stephen K. Van Den Eeden[3,4], Neil Risch[1,2,3], John S. Witte[1,2,4]*

1 Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America, **2** Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America, **3** Division of Research, Kaiser Permanente, Northern California, Oakland, California, United States of America, **4** Department of Urology, University of California San Francisco, San Francisco, California, United States of America

HOXB13 G84E (continued)

Table 1. Pleiotropic effect of G84E mutation on risk of cancer.

Cancer[a]	# Case carriers[b]	# Cases[b]	Frequency[c]	Odds Ratio[d] (95% CI)	Corrected Odds Ratio[e] (95% CI)	P-value[d]
Kidney	5	303	0.94%	2.32 (0.94, 5.76)	3.32 (0.89, 9.99)	0.034
Bladder	5	335	0.85%	2.05 (0.81, 5.22)	2.84 (0.70, 8.94)	0.065
Non-Hodgkin's Lymphoma	10	846	0.67%	1.81 (0.98, 3.36)	2.42 (0.96, 5.52)	0.029
Melanoma	15	1301	0.66%	1.50 (0.89, 2.55)	1.88 (0.81, 3.95)	0.066
Pancreas	2	149	0.77%	1.49 (0.28, 7.83)	1.86 (0.18, 13.70)	0.32
Breast	38	3183	0.68%	1.48 (1.04, 2.10)	1.84 (1.07, 3.16)	0.014
Endometrium	6	578	0.59%	1.44 (0.64, 3.24)	1.77 (0.49, 5.22)	0.19
Colon	11	1119	0.56%	1.42 (0.77, 2.60)	1.74 (0.65, 4.06)	0.13
Thyroid	3	217	0.79%	1.35 (0.35, 5.22)	1.61 (0.23, 8.81)	0.33
Ovary	2	198	0.58%	1.32 (0.32, 5.49)	1.56 (0.20, 9.26)	0.35
Multiple Myeloma	1	135	0.42%	1.26 (0.20, 7.92)	1.46 (0.12, 13.75)	0.40
Lung	5	667	0.43%	1.10 (0.44, 2.72)	1.18 (0.30, 4.22)	0.42
Oral	2	283	0.40%	0.98 (0.23, 4.10)	0.97 (0.14, 6.78)	NA
Lymphocytic Leukemia	1	218	0.26%	0.53 (0.05, 5.24)	0.39 (0.03, 9.06)	NA
Any[f]	123	10493	0.67%	1.36 (1.13, 1.63)	1.63 (1.22, 2.29)	5.8×10^{-4}
Any single cancer[g]	106	9532	0.63%	1.41 (1.14, 1.75)	1.72 (1.23, 2.51)	8.9×10^{-4}
Two or more cancers[h]	17	961	1.01%	2.21 (1.35, 3.61)	3.12 (1.60, 6.14)	0.0011