On the genetic architecture of intelligence and other quantitative traits

Steve Hsu

MSU and BGI
www.cog-genomics.org

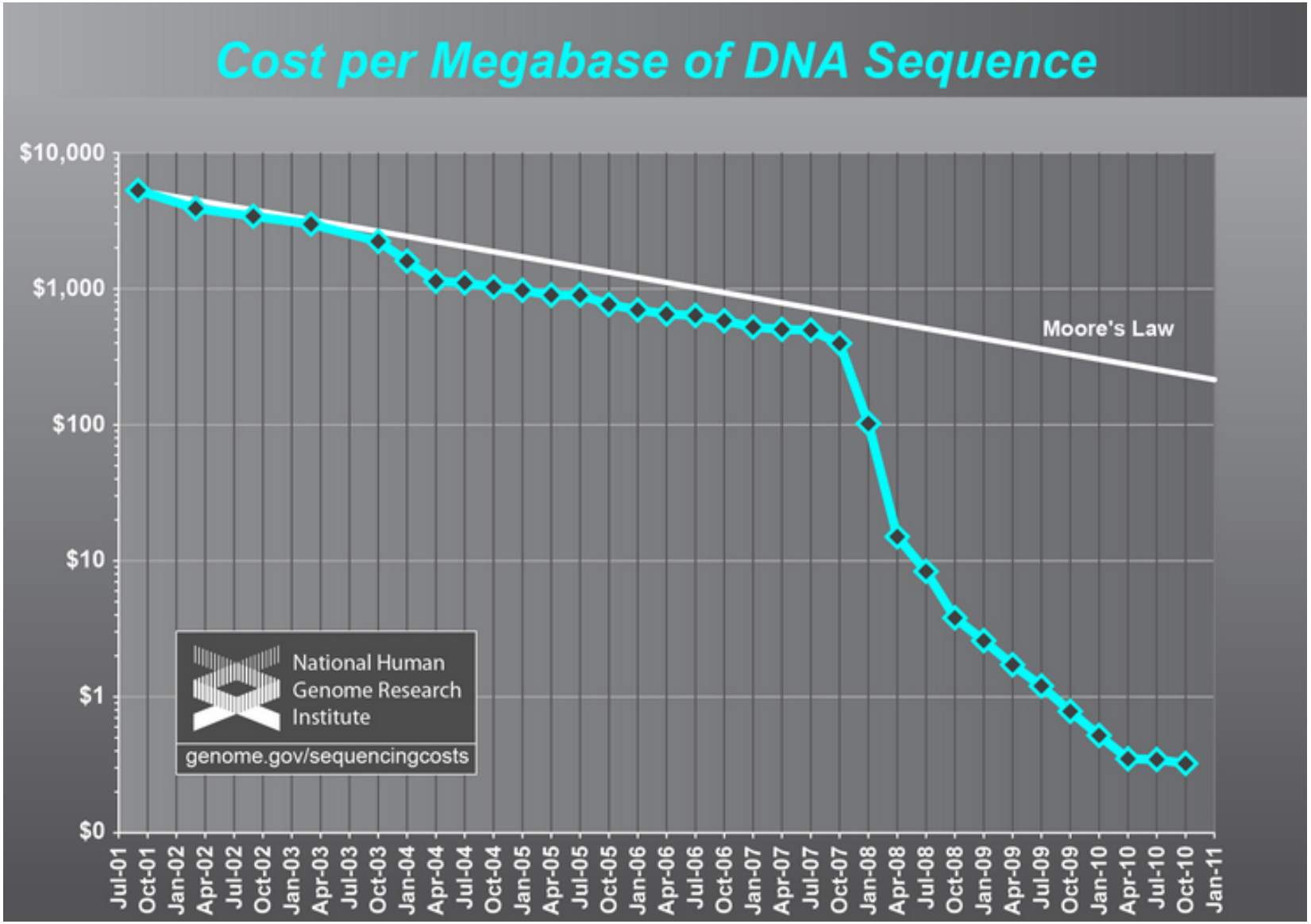
Not this talk... perhaps another time

Macroscopic superpositions and black hole unitarity

We discuss the black hole information problem, including the recent claim that unitarity requires a horizon firewall, emphasizing the role of decoherence and macroscopic superpositions. We consider the formation and evaporation of a large black hole as a quantum amplitude, and note that during intermediate stages (e.g., after the Page time), the amplitude is a superposition of macroscopically distinct (and decohered) spacetimes, with the black hole itself in different positions on different branches. ...

arXiv:1302.0451

Technology and economics drive science



Human genetics primer for physicists

1 genome $\approx 10^9$ base pairs, variation at rate 10^{-3}; compressible to few MB.

SNPs = Single Nucleotide Polymorphisms $\approx 10^6$ sites where variation is common. Informative sampling of whole genome.

2012: SNP genotyping cost = \$100 ; Whole Genome Sequencing = \$1000.

A few years from now SNPs will be a historical oddity of old technology.

Outline: a multidisciplinary subject

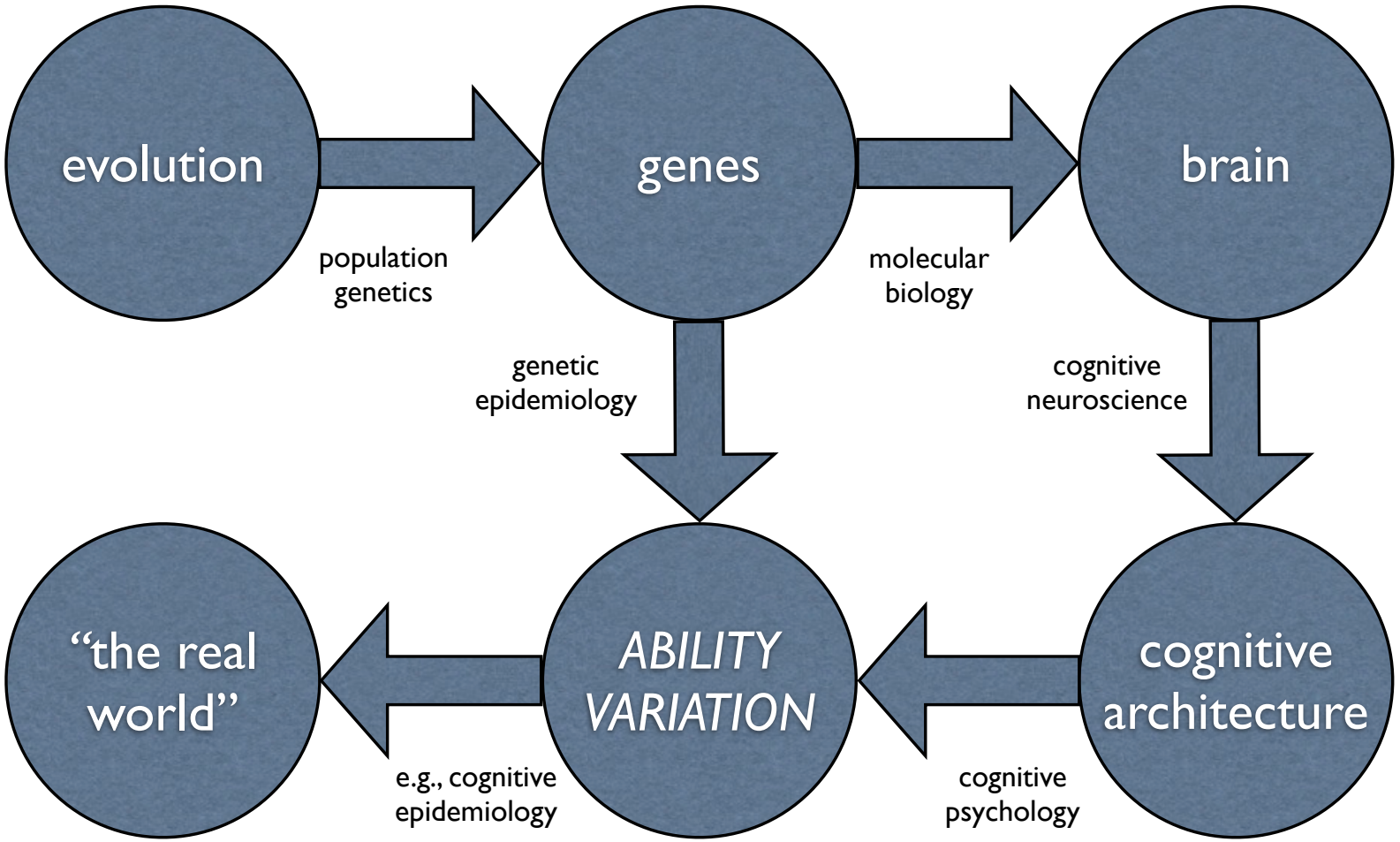
1. What is intelligence? Psychometrics

2. g and GWAS: a project with BGI (formerly Beijing Genomics Institute)

3. Genetic architecture from SNP distance analysis

www.cog-genomics.org

Outline: a multidisciplinary subject



Quantitative phenotypes

1. Stability / Reliability (measured value doesn't change)
2. Validity (predictive power; measures something *real*?)
3. Heritability (genetic causes)

Cognitive ability (properly defined) is comparable to height on each of these criteria.

What are IQ / SAT / GRE ?

By construction:

I. Choose a battery of n "cognitive" tests, e.g.,

(1) digit recall (short term memory)
(2) vocabulary
(3) math puzzles
(4) spatial rotations
...
(n-1) reaction time
(n) pitch recognition (music)

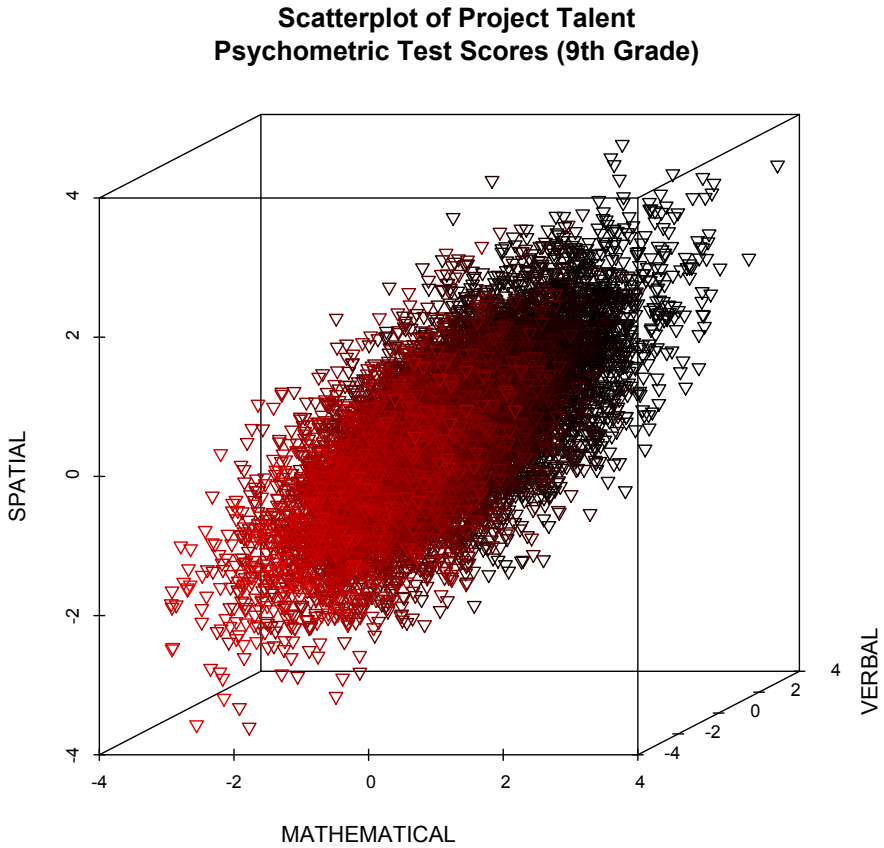
II. Test a lot of people.

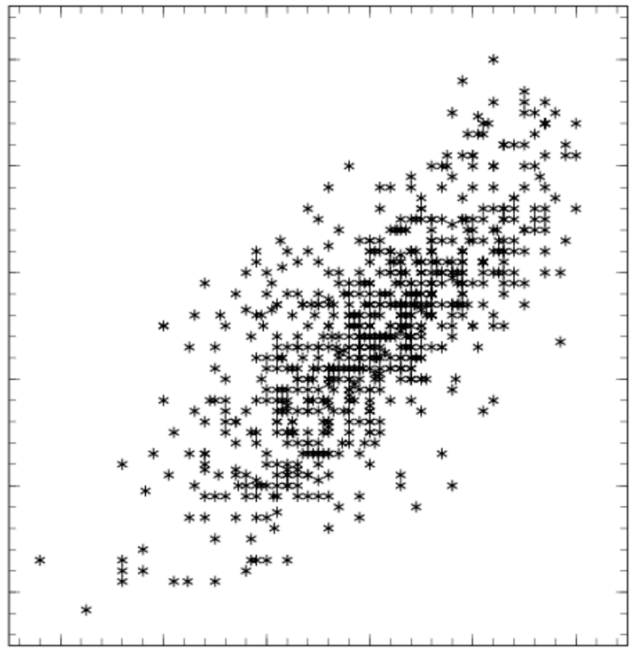
$$\{\text{individual}\} \rightarrow \text{n vector} \rightarrow \text{scalar (single number)}$$

(LOSSY) COMPRESSION!

Results

- All "cognitive" observables seem to be *positively* correlated
- Use factor analysis or principal components to isolate direction of largest variation in the n-dimensional space





General factor of intelligence

Largest principal component of variation ≈ g factor = general factor of intelligence ≈ IQ ≈ SAT ≈ GRE

≈ overall goodness of cognitive functioning?

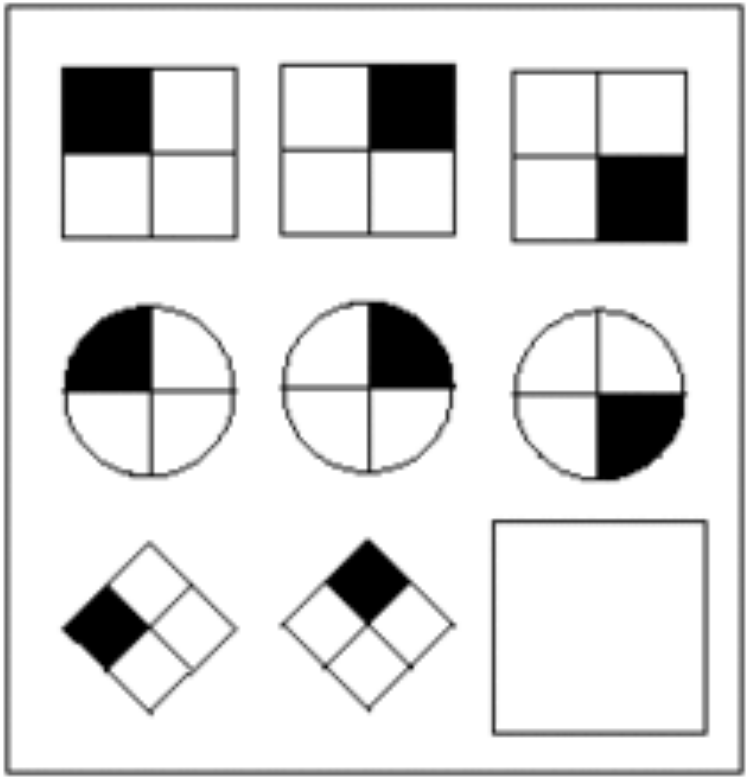
• Note these are *population level* correlations – compression may not work for a particular individual: value of g may not predict individual components of n-vector very well. But works for ”typical” individuals.

• SAT, GRE heavily g-loaded: high correlation with g or IQ; ”SAT is an IQ test”

IQ: mean 100, SD 15 (normally distributed)

SAT (M+V): mean 1000, SD ~ 200 (1995 ”recentering”)

Progressive Matrices



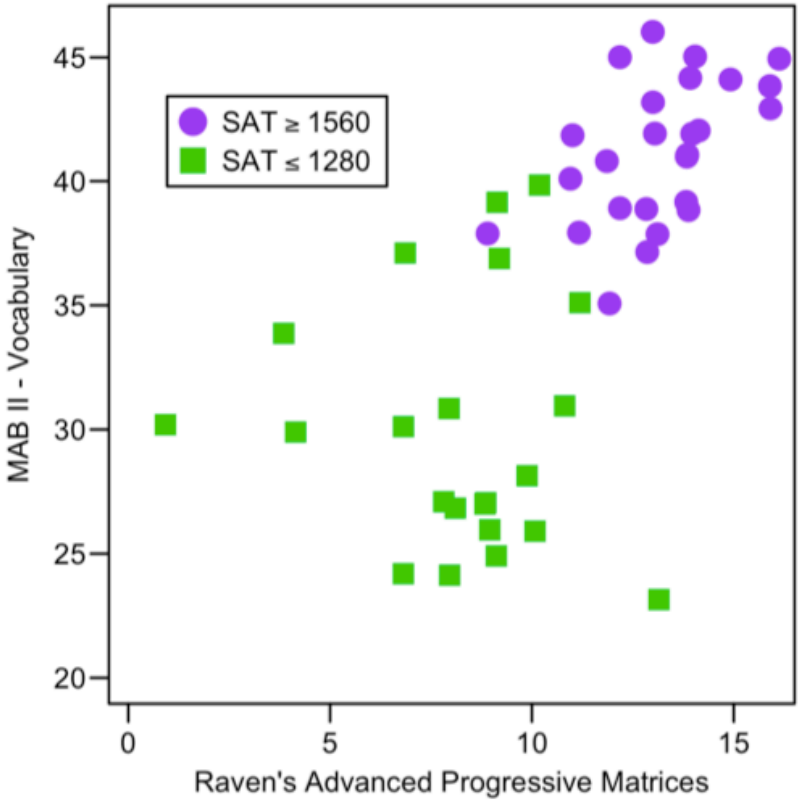
Highly g loaded but relatively culture neutral and abstract.

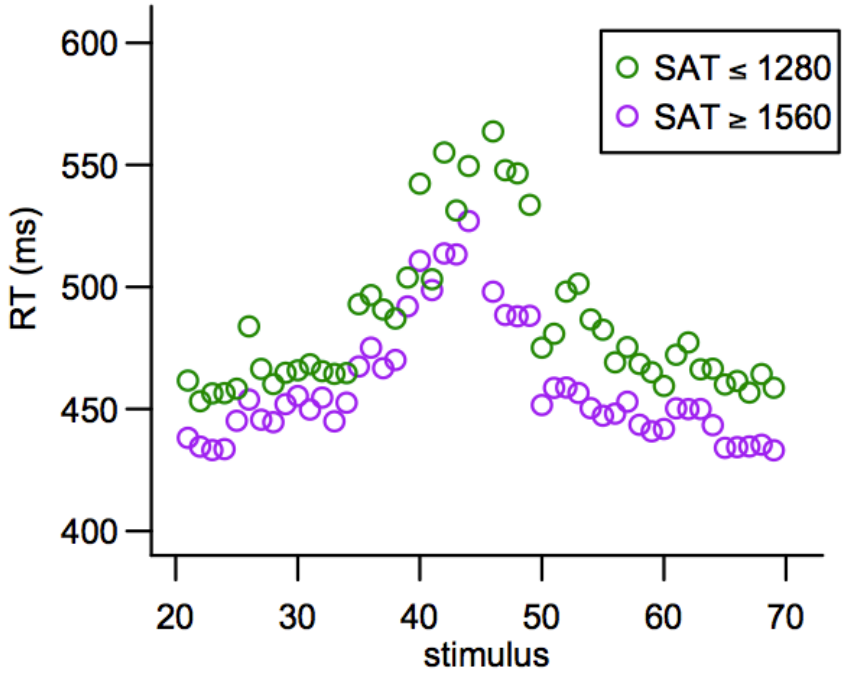
Pattern recognition and algorithmic reasoning.

Results

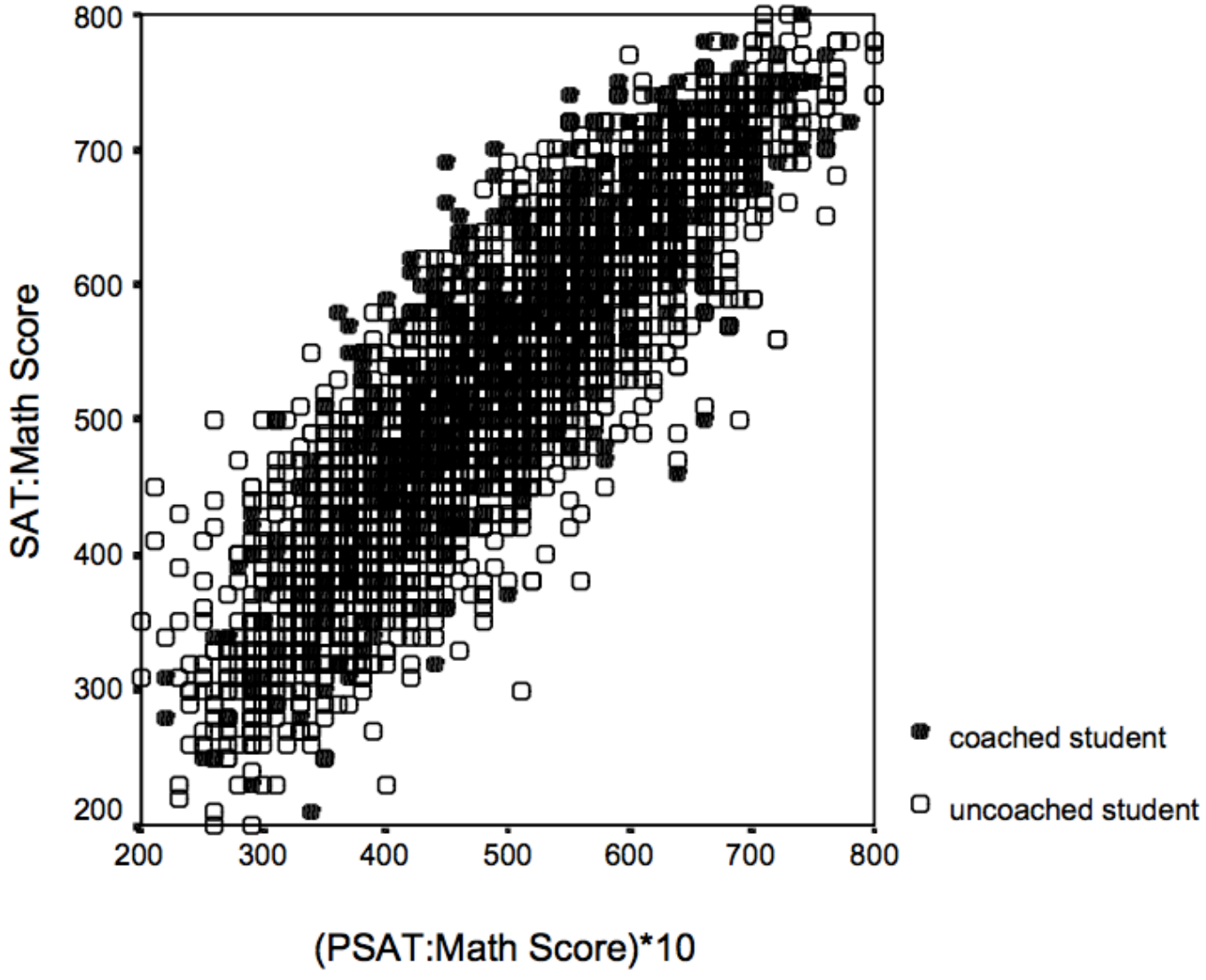
left fig. Vocabulary, SAT and RAPM intercorrelation.

right fig. Reaction time differences for two groups.





Scores are difficult to change



g: what good is it?

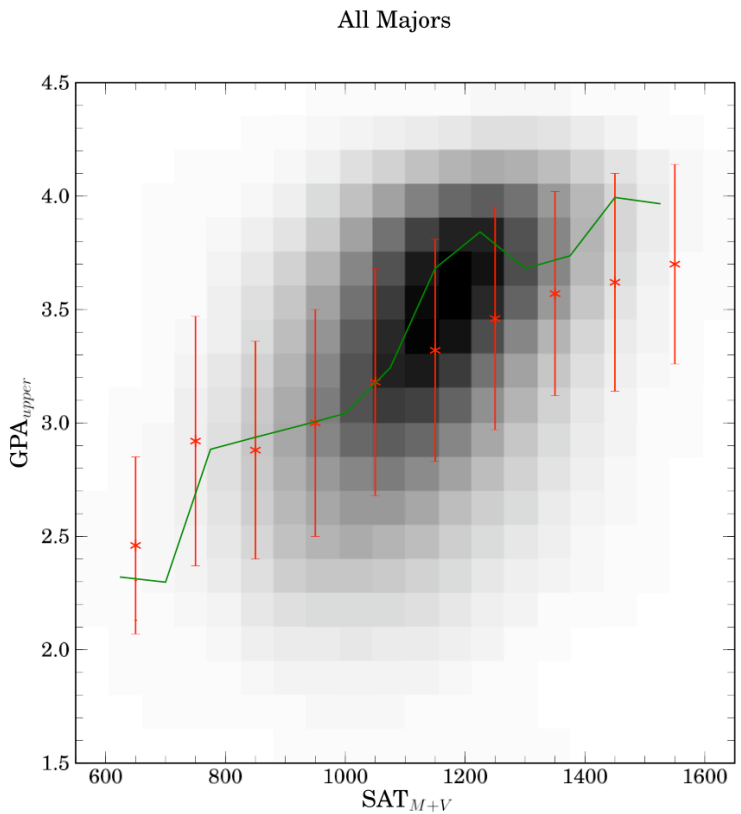
Among the most impressive quantitative results in all of psychology.

1. Results are stable after late adolescence (reliability). One year retest correlation .9 or higher.
2. Results are predictive (validity).
3. It's heritable (twin and adoption studies).

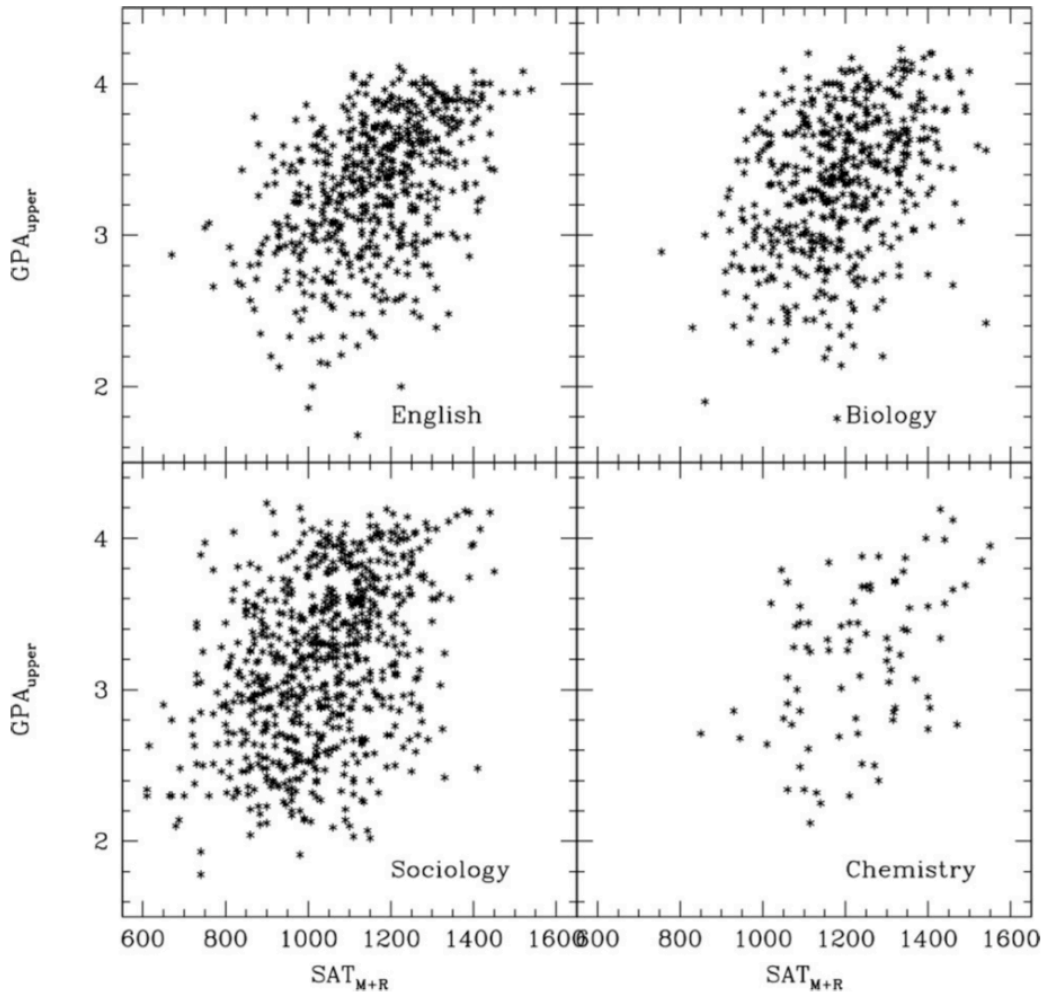
College outcomes

Data Mining the University, Hsu and Schombert, arXiv:1004.2731

Analysis of 5 years of student records at the University of Oregon.



College outcomes



The far tail

What about the far tail?

+2 SD	130	top few percent
+3 SD	145	1 in 1000
+4 SD	160	1 in 30,000

Diminishing returns above some threshold (e.g., 120)?

OR

It's good to have a big brain ... BIGGER IS BETTER :-)

The far tail

Roe study (1950's): 64 randomly selected eminent scientists had IQs much higher than the general population of science PhDs. Almost all of the eminent scientists in the sample scored above +(3-4) SD in at least one of M / V categories.

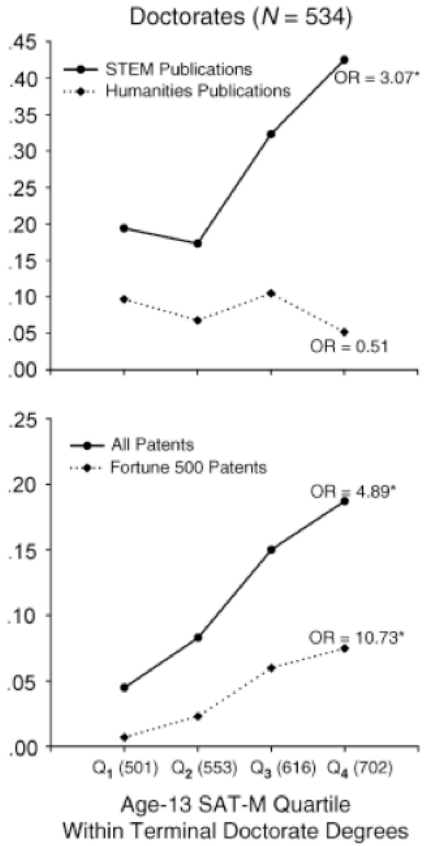
Mean score in both categories was roughly +4 SD.

Average for science PhDs around +2 SD, so eminent group highly atypical among scientists.

Positive returns to IQ > +2 SD in scientific research?

The far tail: SMPY longitudinal study.

Tested at age 13 or younger. First quartile Q1 roughly top percentile, top quartile Q4 roughly 1 in 10,000. Q4 > NSF Fellows at top 5 departments in later careers.



Heritability

Heritability is defined relative to a specific range of environments.

High heritability estimates are obtained in cases where subjects have generally experienced good environments. In the absence of deprivation, genes have a big effect: probably determine upper limit to height, cognitive ability, etc.

However, in studies where subjects have experienced a wider range of environmental conditions, such as poverty, malnutrition or lack of education, heritability estimates are much smaller (Turkheimer). When environmental conditions are unfavorable, individuals do not achieve their full potential.

Heritability and Linearity

g is highly heritable and effect of individual genes is mostly linear: many genes, each of small effect. (Additive heritability about .6; broad sense heritability about .8; similar to height!)

kinship	IQ correlation	number of pairs
monozygotic twins reared apart	0.77	87
monozygotic twins reared together	0.82	1684
full siblings reared apart	0.38	144
full siblings reared together	0.47	100,000+
biologically unrelated adoptive siblings	0.05	471

Heritability and Linearity

Table 1 Intraclass twin correlations and 95% confidence intervals for general cognitive ability (*g*) at each site by zygosity

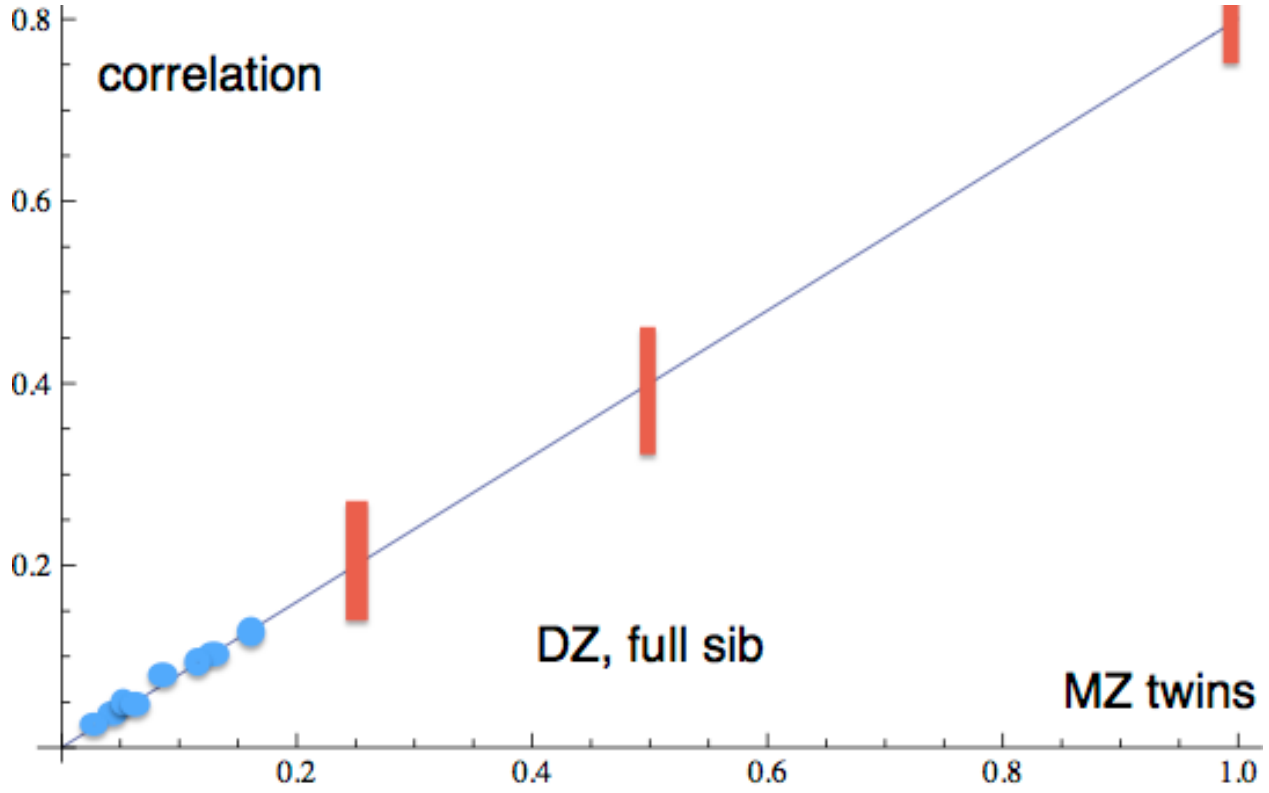
GHCA site	*MZ*	*DZ*
US Ohio	0.76 (0.68–0.83) ($n = 121$)	0.55 (0.44–0.65) ($n = 171$)
United Kingdom	0.67 (0.64–0.69) ($n = 1518$)	0.43 (0.40–0.46) ($n = 2500$)
US Minnesota	0.76 (0.73–0.78) ($n = 1177$)	0.50 (0.44–0.55) ($n = 679$)
US Colorado	0.82 (0.80–0.84) ($n = 1288$)	0.53 (0.49–0.56) ($n = 1559$)
Australia	0.83 (0.79–0.86) ($n = 338$)	0.48 (0.41–0.54) ($n = 513$)
Netherlands	0.83 (0.80–0.86) ($n = 434$)	0.58 (0.52–0.63) ($n = 517$)

Heritability and Linearity

Nature Molecular Psychiatry 16, 996-1005 (October 2011)

... We conducted a genome-wide analysis of 3511 unrelated adults with data on 549,692 single nucleotide polymorphisms (SNPs) and detailed phenotypes on cognitive traits. **We estimate that 40 percent of the variation in crystallized-type intelligence and 51 percent of the variation in fluid-type intelligence between individuals is accounted for by linkage disequilibrium between genotyped common SNP markers and unknown causal variants. These estimates provide lower bounds for the narrow-sense heritability of the traits.** ...

Heritability and Linearity



Modern genomic method: we can effectively vary genetic relatedness (horizontal axis) *continuously* over many pairs, and measure changes in phenotype correlation or difference (vertical axis).

Linearity: many genes of small effect

1. phenotype is normally distributed
2. genetic component is approximately linear in effect (e.g., for g, additive heritability .6 out of .8 total)

Can think in terms of $+$ and $-$ effects from alleles. (I suppress effect sizes ϵ_i.)

Characterize an individual in terms of which variants they inherit at each of n sites:

$$(+ + + - + + - \quad \cdots \quad + + - - + ++)$$

Coin flips with probability p_i at each site yields normal distribution as $n \to \infty$.

Evolution and additive variance

Why are phenotype differences linear functions of genotype?

Consider diploid genotypes: *CC*, *cC*, *cc*

Non-linear interactions (*epistasis*): effect of cc may not be twice effect of cC. (Also multi-locus interactions.)

But if variants c are relatively rare (e.g., p = 0.1 - 0.2), the effect of non-linearity is suppressed and non-linear effects are small *as a fraction of total variation*.

A high degree of non-linearity at the genetic level can still correspond to almost linear aggregate variation between two individuals.

Biology ≈ linear combination of non-linear gadgets!

Evolution and additive variance

Additive variation is easier for evolution to act on, and polygenic traits do not easily exhaust their variation.

Fisher's Fundamental Theorem says rate of increase of fitness is approximately the *additive* (linear) genetic variance:

$$\frac{d\langle F\rangle}{dt} \approx \sigma_A^2$$

(for sexually reproducing species with recombination timescale smaller than evolutionary timescale).

Animal and plant breeders have been using additive variance for millennia.

Example: Maize experiments over 100 generations of selection have produced a difference in oil content between the high and low selected strains of 32 times the original standard deviation!

General model for quantitative phenotype

y = individual phenotype
g_i = individual genotype (e.g., list of 1M SNPs or 3B loci)
x_i = linear effect sizes
z_{ij} = tensors of nonlinear effect sizes

$$y = \sum_i g_i x_i + \sum_{ij} g_i g_j z_{ij} + \mathcal{O}(g^3)$$

Plausible that linear term dominates, even if nonlinear terms are important in certain circumstances.

We will extract the effect sizes x_i for a variety of human traits in the next decade, allowing for approximate genomic prediction.

g and GWAS

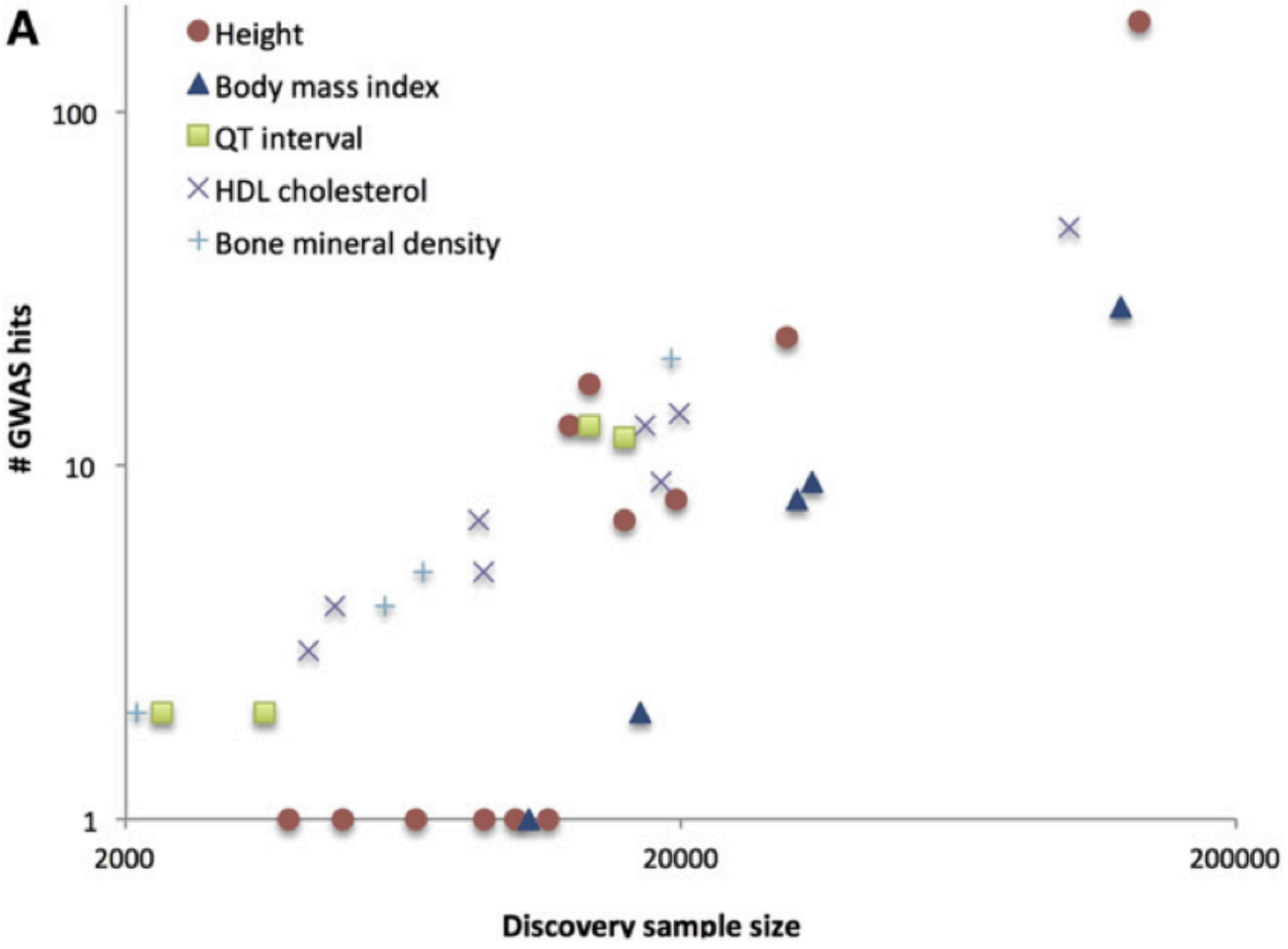
GWAS = Genome Wide Association Study. Thus far, little success in finding genes linked to intelligence (Plomin 2010).

Candidate hits have not been successfully replicated.

Compare to the situation with height: about 750 genes found so far correlated to height, with > 100k pheno-genotype pairs analyzed. Only 20 percent of total variance associated with specific loci ("missing heritability"), but over 50 percent or more of total variance from global fit.

There is a historic opportunity to conduct the first study that finds a significant number of IQ-associated genes.

GWAS history



Nature Molecular Psychiatry 29 January 2013

Here, we report the first genome-wide association study (GWAS) on childhood intelligence (age range 6–18 years) from 17,989 individuals in six discovery and three replication samples... aggregate effects of common SNPs explain 22–46% of phenotypic variation in childhood intelligence in the three largest cohorts... FNBP1L, previously reported to be the most significantly associated gene for adult intelligence, was also significantly associated with childhood intelligence (P=0.003). ... these genetic prediction results are consistent with expectations if the genetic architecture of childhood intelligence is like that of body mass index or height.

BGI: formerly Beijing Genomics Institute

Headquarters in Shenzhen, China. Raised funding of US $ 1.6 billion. Nearly 5000 employees (1000 in software development alone).

More sequencing power than any academic lab in US or Europe. Aims to become leading platform for sequencing and bioinformatics.

Previous successes: participant in original Human Genome Project (1 percent), rice genome, Panda genome, Tibetan altitude adaptation, early hominid sequence, over 1000 Han genomes sequenced.

BGI

BGI

Table 1 Top 10 Institutions in *Nature Publishing Index 2010 China*

2010				2009		
RANK	INSTITUTION	CORRECTED COUNT	ARTICLES	RANK	CORRECTED COUNT	ARTICLES
1	Chinese Academy of Sciences	13.354	40	1	12.008	31
2	Tsinghua University	6.155	16	2	2.696	8
3	University of Science and Technology of China	3.725	7	3	2.674	8
4	BGI Shenzhen	3.572	9	19	0.520	1
5	Peking University	3.440	17	4	2.615	8
6	Nanjing University	3.088	7	7	1.413	5
7	The University of Hong Kong	2.172	8	8	1.292	4
8	Southeast University	2.049	3	–	–	–
9	Xiamen University	1.829	3	10	1.000	1
10	Zhejiang University	1.650	12	14	0.664	4

(Table is quoted from the journal of *Nature Publishing Index 2010 China*)

BGI

High–normal (case:control) design

Seek thousands of subjects with IQ +3 SD or higher (roughly 1 in 1000).

US gifted education in last 20 years: SAT at age 12. Ceiling very high: above 1 in 10,000.

We have obtained 2000 DNA samples from this 1 in 10,000 population and will obtain whole genome sequences.

Rough estimates

Simple model: n genes of equal small effect. (i.e., $n \sim 10^3$). Let + allele have slightly positive effect on IQ, and − allele have slightly negative effect.

Assume high group average is $+k$ SD, so $k \sim (3-4)$. Then difference in frequencies between high and normal groups is

$$f_+^H - f_+^N \sim \frac{k}{2\sqrt{n}}$$

How well can we measure f_+ in the two populations? Statistical fluctuations: $\frac{1}{2\sqrt{M}}$, where M is population size.

Once $M > n$, have good power to detect + alleles. (False positives: 10^3 variants of interest, 10^6 SNPs on chip; need signal to noise ratio of $> 10^3$ or so.)

Power calculations

Expected hits assuming IQ allele frequencies and effect sizes similar to height.

2000 CASES, 4000 CONTROLS

case lower threshold = 3.5 SD **total expected hits: *3.51***

	average effect					
MAF	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**
0.1	—	—	0.02	0	0.26	0
0.2	—	0.07	0.28	0	0	0
0.3	—	0.18	0.36	0	0	0
0.4	—	0.20	0	1.24	0.90	0

Genetic architecture of intelligence

Preliminary results as presented to the 2012 Behavior Genetics Association meeting in Edinburgh.

Collaborators

ALSPAC / University of Bristol: George McMahon, George Davey Smith

TEDS / King's College London: Ken Hanscombe, Robert Plomin

NIDDK / NIH: James Lee, Shashaank Vattikuti, Carson Chow (Height data from ARIC)

Cognitive Genomics Lab, BGI: Christopher Chang, Laurent Tellier, Rui Yang, Bowen Zhao

Genetic distance measures

Quantitative traits: many alleles, each of small effect. GWAS discovery of individual loci is hard.

But, phenotype differences must be associated with LARGE NUMBER of genetic differences.

Investigate pairwise genetic distance as g score (or height) are varied. Extract underlying genetic architecture:

1. Distribution of associated alleles dominated by small MAF (Minor Allele Frequency)
2. More (−) than (+) minor alleles (MAF < 0.5)
3. Rough estimate of 10k causal alleles in total

Data sources and Results

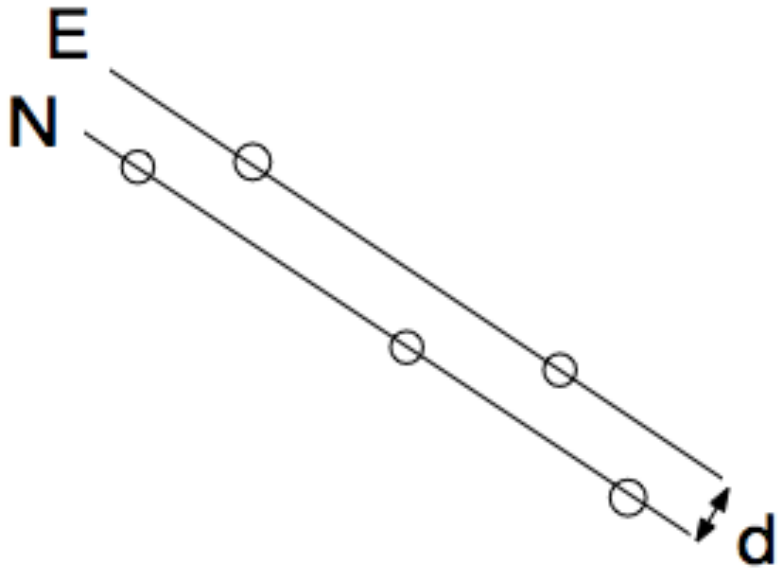
ALSPAC: 4000 individuals, age 15 IQ; 2000 individuals, age 8 IQ
TEDS: 2400 individuals, age 12 IQ
ARIC: 5700 adult heights

ALSPAC: 488k SNPs on chip. Average pairwise distance = 261k ± 1.5k SNPs.

Select outlier groups H and L. Averaging over pairs eliminates fluctuations in distance which are uncorrelated to phenotype.

Average pairwise genetic distance changes with mean IQ and IQ difference: ~ 40 SNPs per population SD

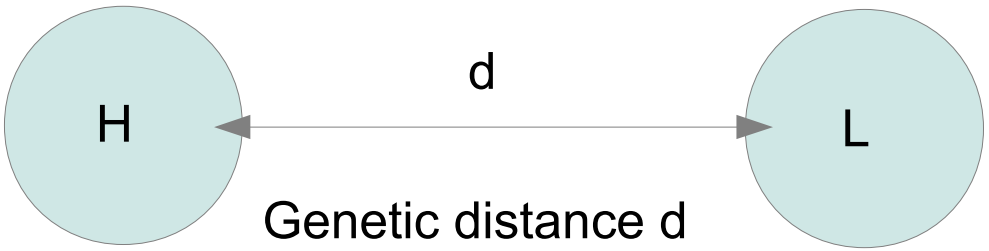
Genetic distance between surfaces of constant phenotype



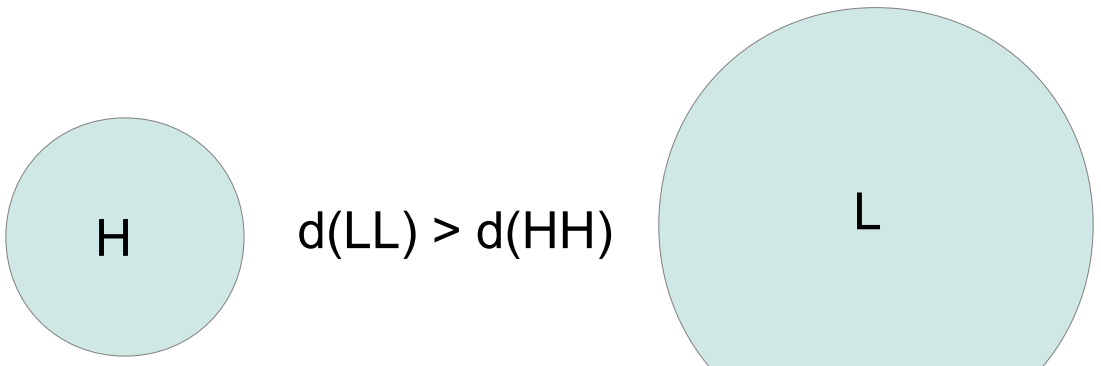
Typical distance between individuals: $3 \cdot 10^5 \pm 10^3$ SNPs.

Detect $d \sim 100$ using 10^6 pairs (fluctuations cancel).

Results



Genetic distance d
varies with avg
phenotype difference



Average intragroup distance
varies with group mean

Genetic distance: architecture from geometry

These two genotypes have a relative Hamming distance of 2:

$\{+ + + + \ominus + + + + + + + + +\}$ vs $\{+ + + + + + + + + + + \ominus + +\}$

These two genotypes have a relative Hamming distance of 6:

$\{+ + \ominus + + + + + + \ominus + + \ominus +\}$ vs $\{+ + + + + \ominus \ominus + + + \ominus + +\}$

More $\ominus$ alleles means greater Hamming distance.

Note we've made the assumption that (+) is common (MAF > 0.5) and $\ominus$ is uncommon (MAF < 0.5). Otherwise, more (+) alleles would mean greater Hamming distance.

Genetic distance: architecture from geometry

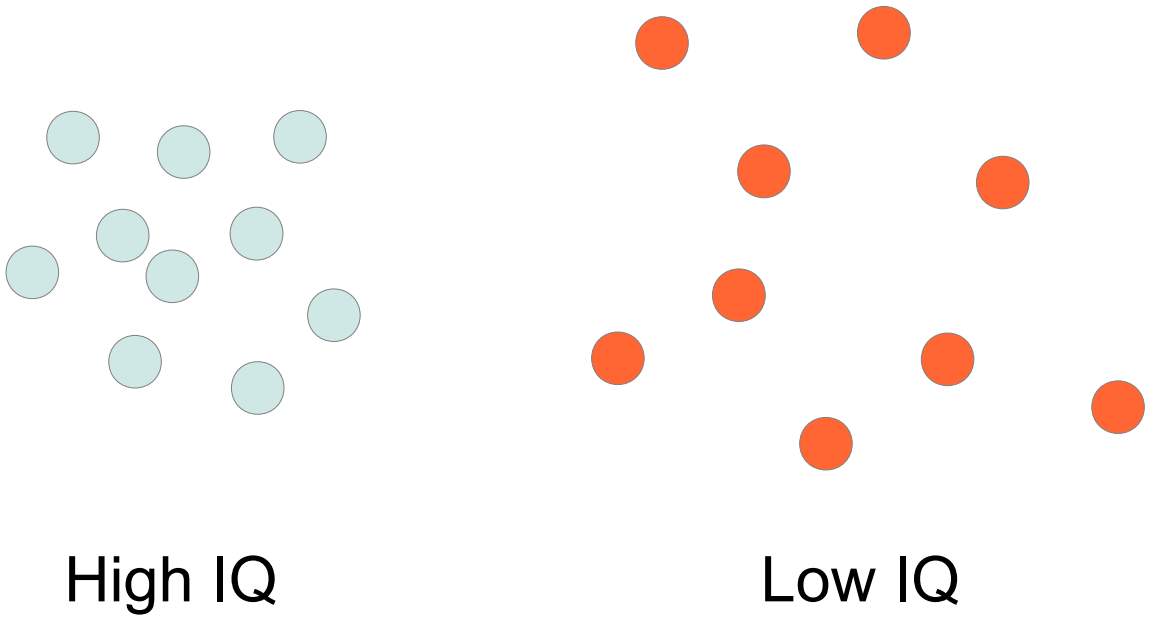
Real genomes are diploid.

Simplest distance measure, analogous to Hamming distance:

AA AA	0
AA Aa	1
AA aa	2

Can also weight by factors of MAF or standardize to obtain different distance measures (e.g., relatedness).

Genetic distance: architecture from geometry



Low IQ = more rare (−) variants. Larger genetic distances between individuals. Similar results for height.

Additive model

n_+ minor alleles with (+) effect on intelligence (MAF < 0.5).

$n_\ominus$ minor alleles with ($-$) effect on intelligence (MAF < 0.5).

Result $d(LL) > d(HH)$ implies that

$$n_\ominus > n_+$$

Plausible that

$$n_\ominus \gg n_+$$

Simplified additive model: spherical cow

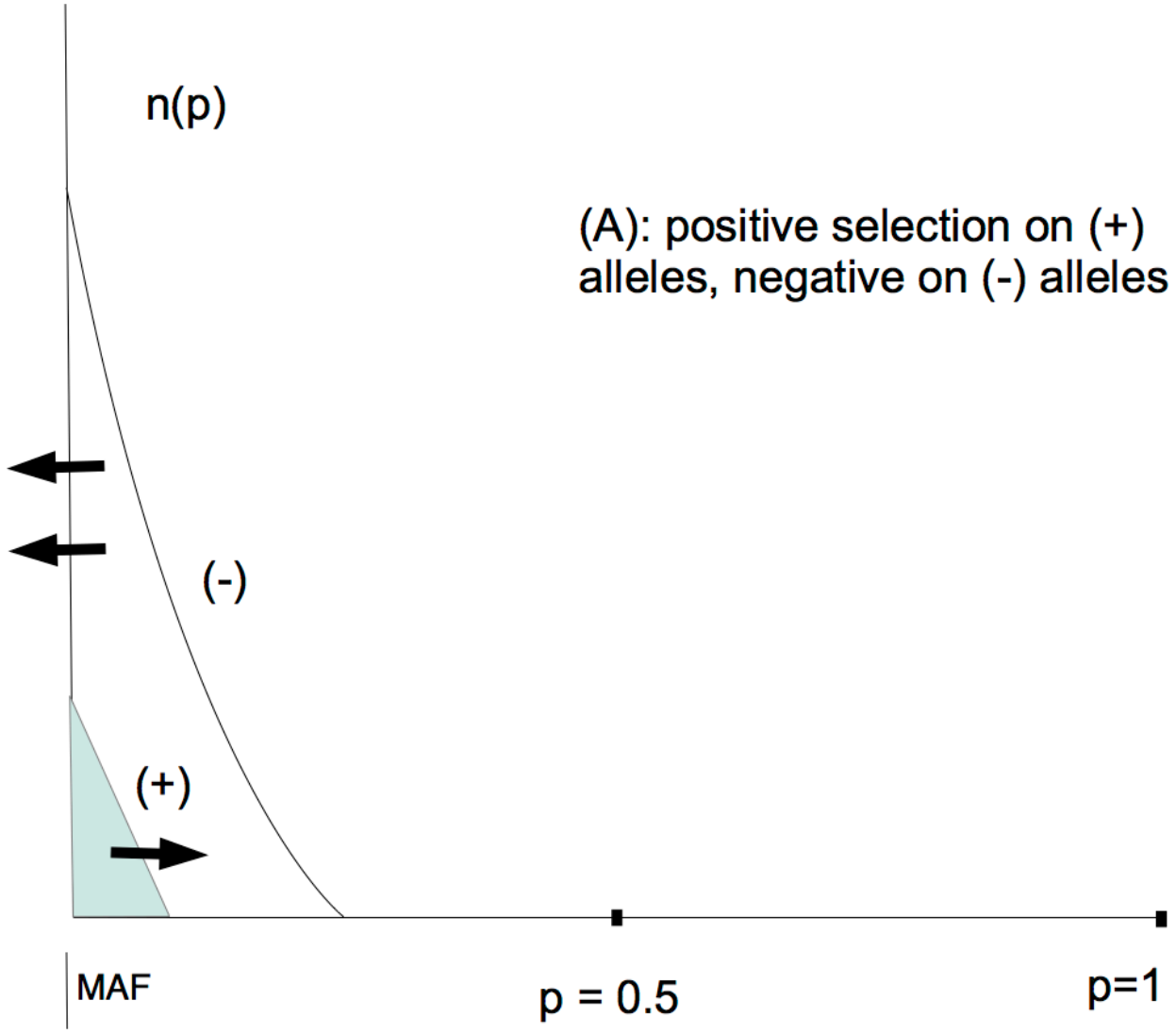
(1) N causal variants, ALL minor alleles have (−) effect on IQ ($n_+ = 0$; $n_\ominus = N$)

(2) Typical $MAF < 0.1$

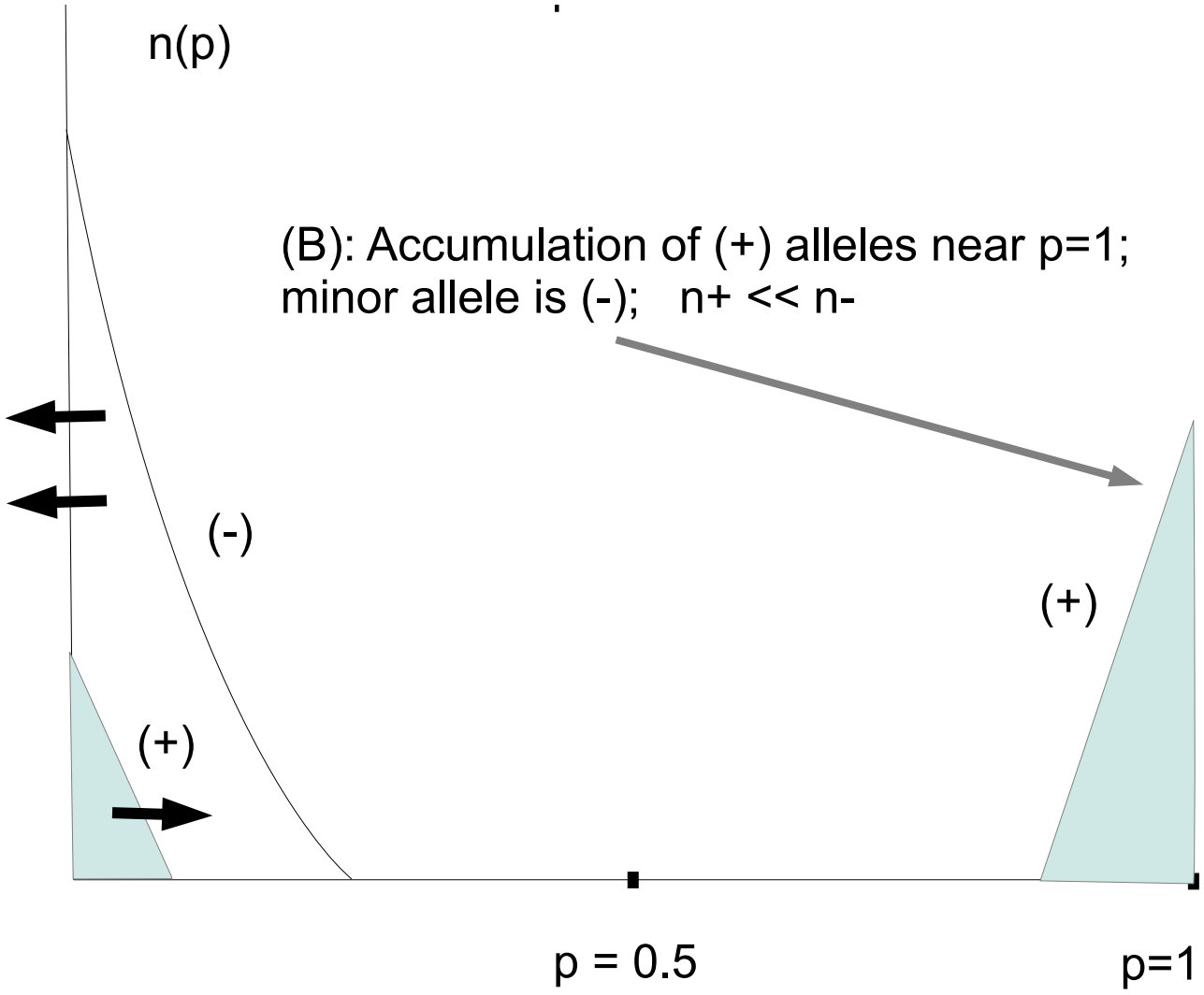
(3) Binomial distribution: 1 SD $\sim (0.1\,N)^{1/2}$

For $N \sim$ 10k, get 1 SD change in intelligence per 30 extra (−) variants.

Selection and MAF distribution

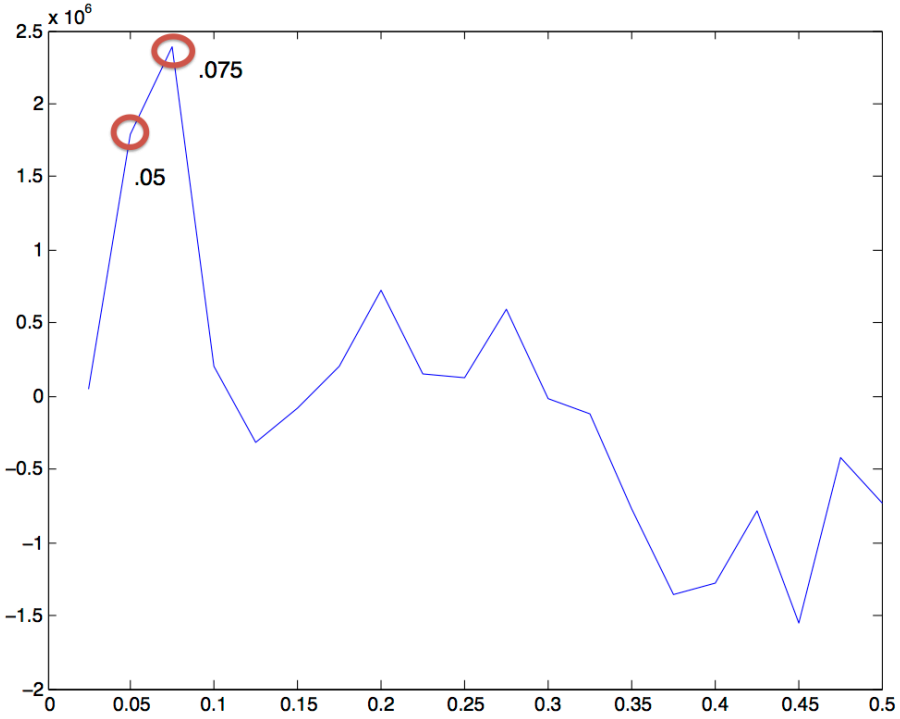


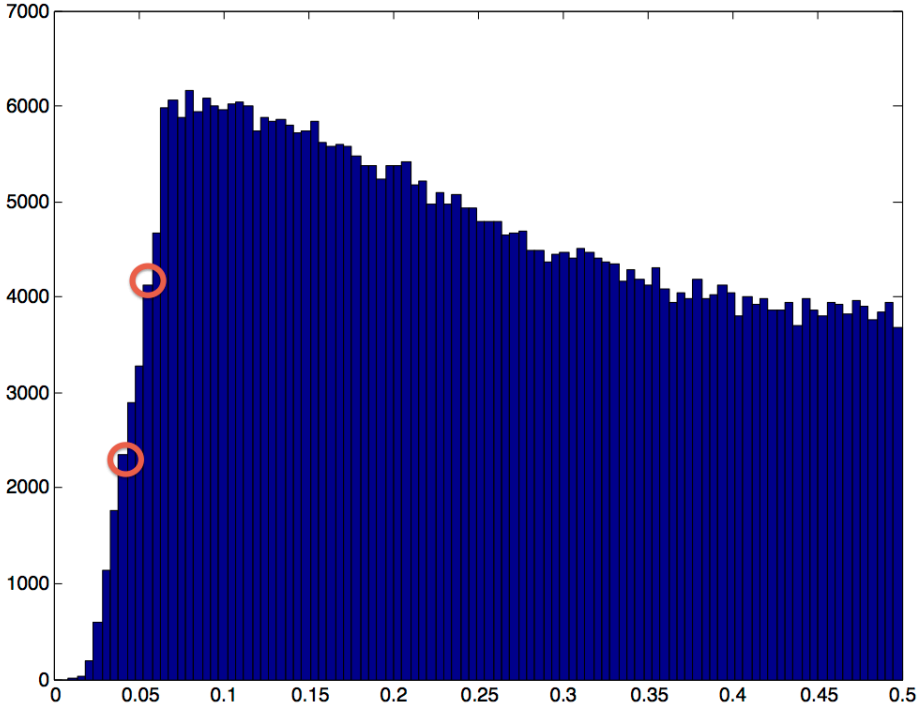
Selection and MAF distribution



MAF distributions

Distribution of associated alleles dominated by MAF < 0.1.





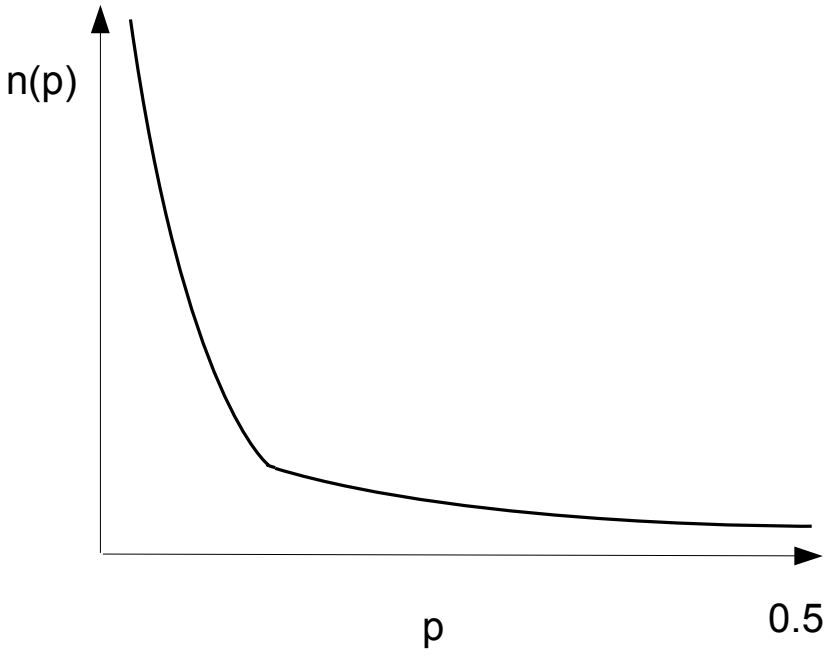
Left: contributions to H–L genetic distance by MAF. Right: density of SNPs on chip.

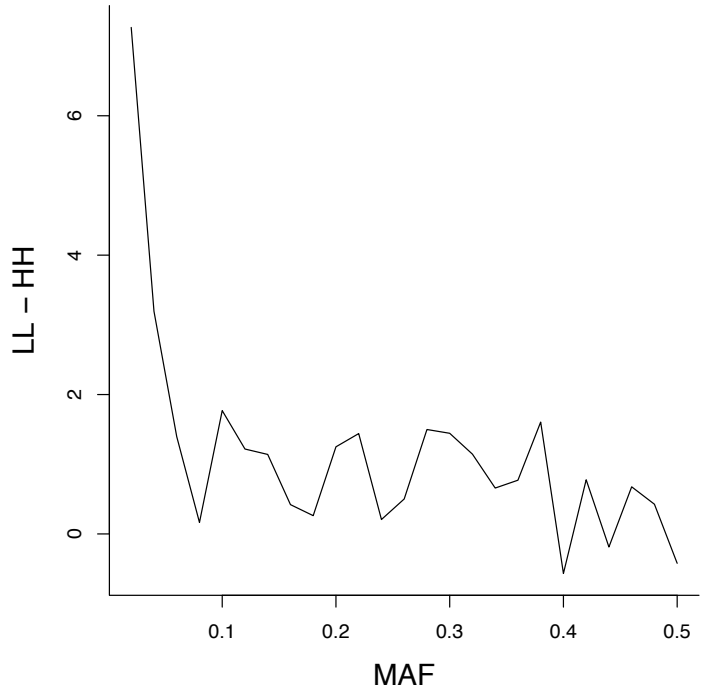
MAF distributions

Modulo statistical errors, can extract

$$n(p) = \text{density of associated SNPs}$$

Result consistent with "L shape" suggested by population genetics models.





Implications of low MAF: missing heritability

GCTA: heritability on chip is roughly $h^2 \sim 0.5$. (Specifically, 0.56 for ALSPAC.)

But, expect larger total additive heritability, perhaps even $h^2 \approx 0.8$!

Yang et al. 2010: causal variants at low MAF are poorly tagged by chip; if MAF of many causal variants < 0.1, can recover "missing heritability".

Implications of low MAF: epistasis, additivity and all that

Why is most of the variance additive? Where is the epistasis that our wet lab colleagues see every day?

If most causal variants are rare (e.g., MAF < 0.1), then when two individuals differ at a locus we likely find AA vs Aa. Very few individuals are aa.

Therefore, even if the effect of aa is not twice that of Aa (non-additivity or non-linearity), the relative size of population level non-additive effects is still small – suppressed relative to additive effects by of order MAF.

(Similar argument for gene-gene interactions, etc.)

Geniuses and Giants: Fewer deleterious alleles.

(A) $\sim$ 40 SNPs per SD of IQ suggests roughly 10k causal variants.

(B) Exceptional cognitive ability = of order 100's fewer rare (−) variants than an average person.

Many caveats to estimate (A); uncertainty in (B) is smaller due to SD $\sim \sqrt{N}$.

Toy model: 10k causal variants, typical MAF = 0.1 : average person has $\sim$ 1000 randomly distributed (−) variants; little overlap between individuals in locations of (−)'s. A genius or giant has $\sim$ 100 fewer (−) alleles: $\sim$ 900 (−) variants in total.

WDIST

WDIST genomic distance calculator

WDIST performs a variety of lengthy computations on PLINK-formatted genomic datasets. (Support for SNPTEST-formatted files is currently being added.)

Latest versions

v0.10.4L, 3 October 2012 (**Download: Linux x86_64 binary**, **OS X x86_64 binary**, **Amazon Linux/OS X source***.)
v0.10.2, 3 October 2012 (**Download: Linux/OS X source**.)

*: You can compile this on other Unix systems if you are familiar with linking with CBLAS/CLAPACK.

Recent version history

0.10.4: --indep bugfix.
(more...)

Source compilation instructions

Sample usage

```
wdist --bfile test --exponent 0.5 --distance gz
```

Compressed Sensing

Problem: Extract linear genetic model (effect sizes x) from statistical data (genomes G + phenotypes y).

$$y_i = \sum_j G_{ij} x_j + \epsilon_i$$

1. x is sparse (e.g., s = 10k causal variants among N = 1M SNPs)

2. at least for next few years, an *underdetermined* problem

Surprising, and nearly optimal results, from Compressed Sensing (L1 norm penalty enforcing sparseness). Required data scales as

$$s \log N$$

The Future

Expect full sequencing (not just SNP genotyping) of 10^6 individuals within next 5 years, paid for by science agencies of national governments. Total cost roughly US $1 billion or so ... comparable to first genome sequenced by Human Genome Project! Note: clinical applications of sequencing (e.g., whole genome cancer treatment) may produce even more data ...

IF sufficient phenotype data is collected about these individuals, will have very well-powered GWAS within next few years – enough statistical power to capture a good fraction of total additive variance (about .6 for intelligence).