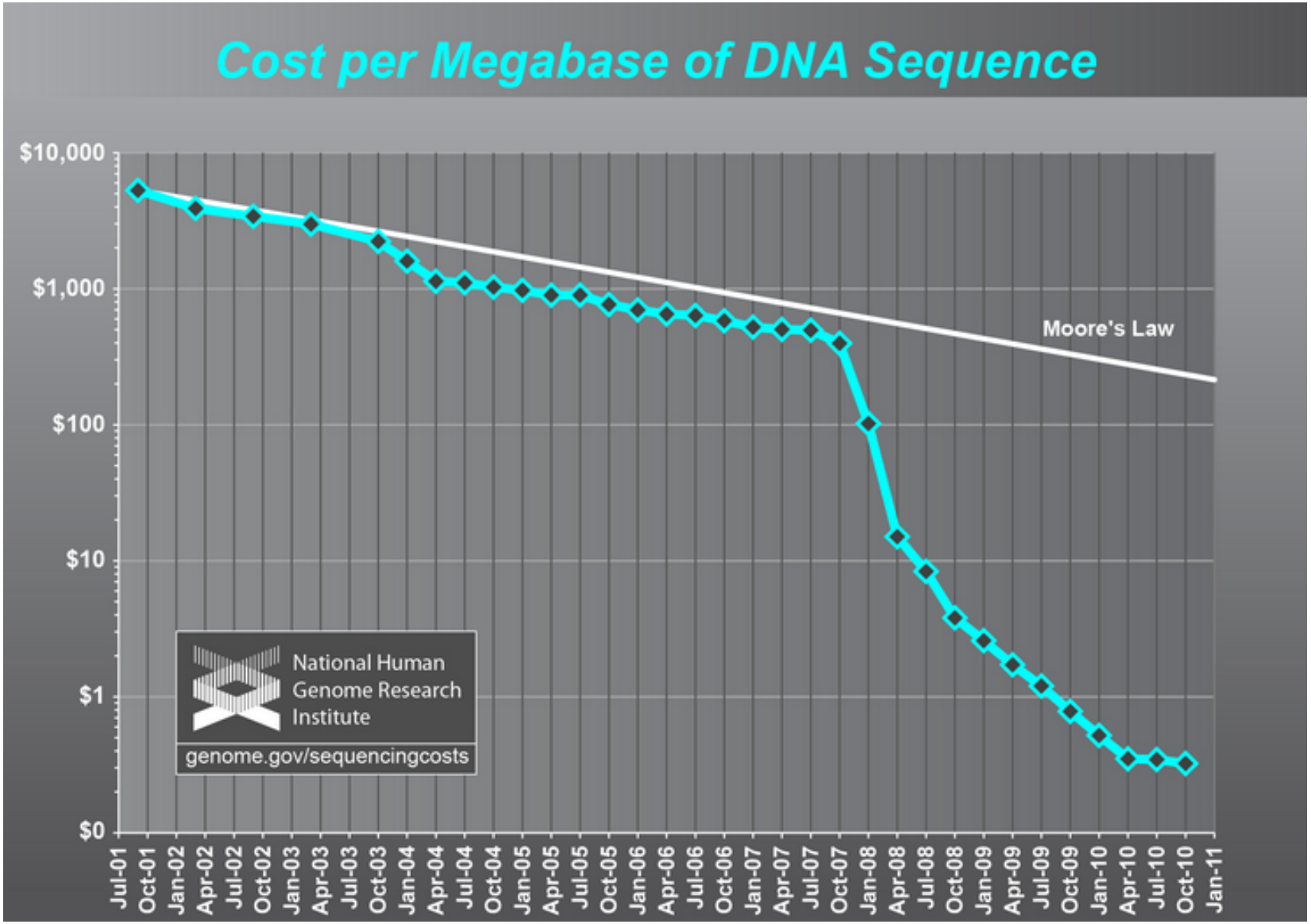
Investigating the genetic basis for intelligence and other quantitative traits

Steve Hsu

University of Oregon and BGI
www.cog-genomics.org

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

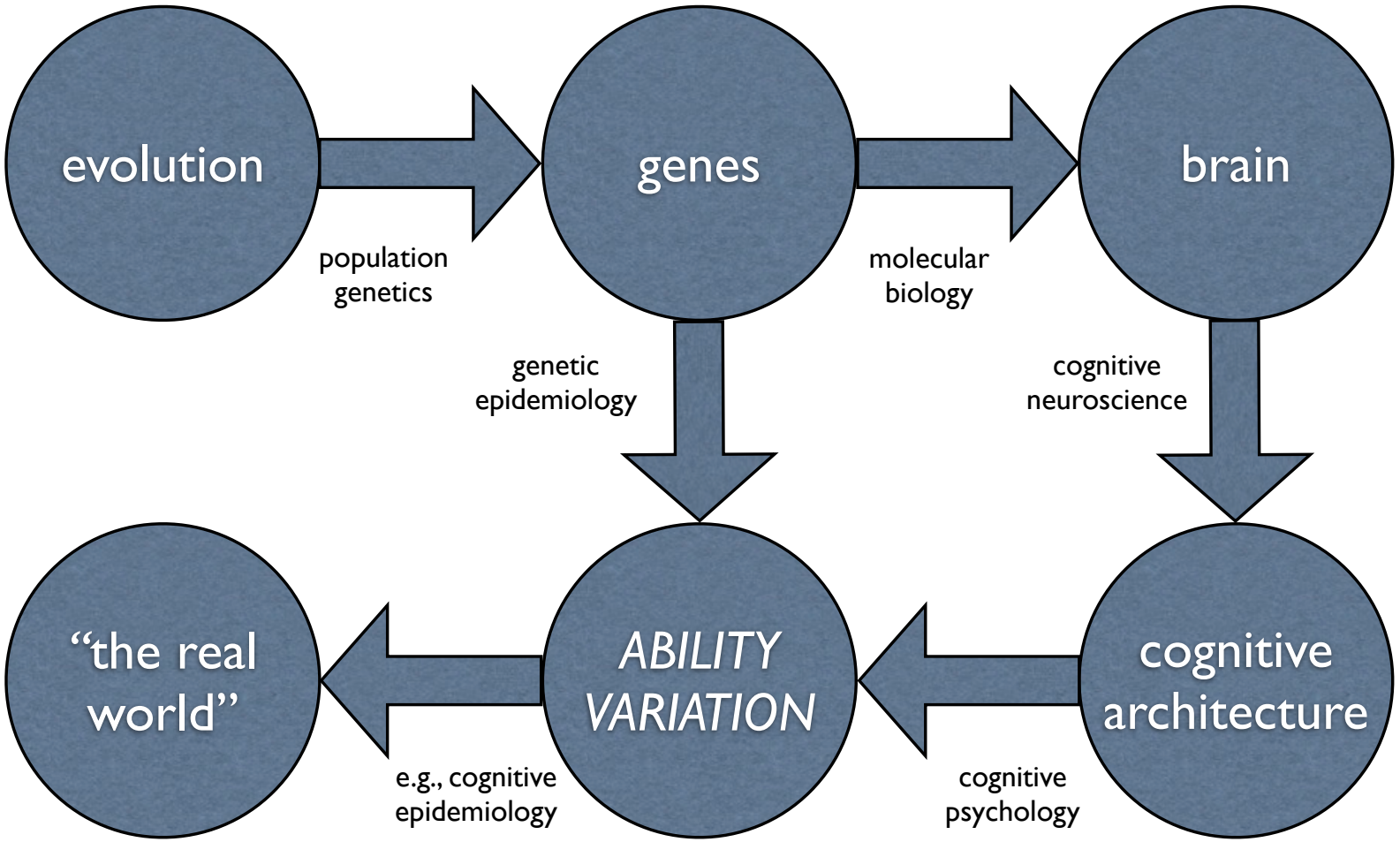
1. What is intelligence? Psychometrics

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

3. The future

www.cog-genomics.org

Outline: a multidisciplinary subject



What is Intelligence: IQ / SAT / GRE ?

Operational perspective: who cares, as long as they have predictive power!

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

We'll see that intelligence 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)

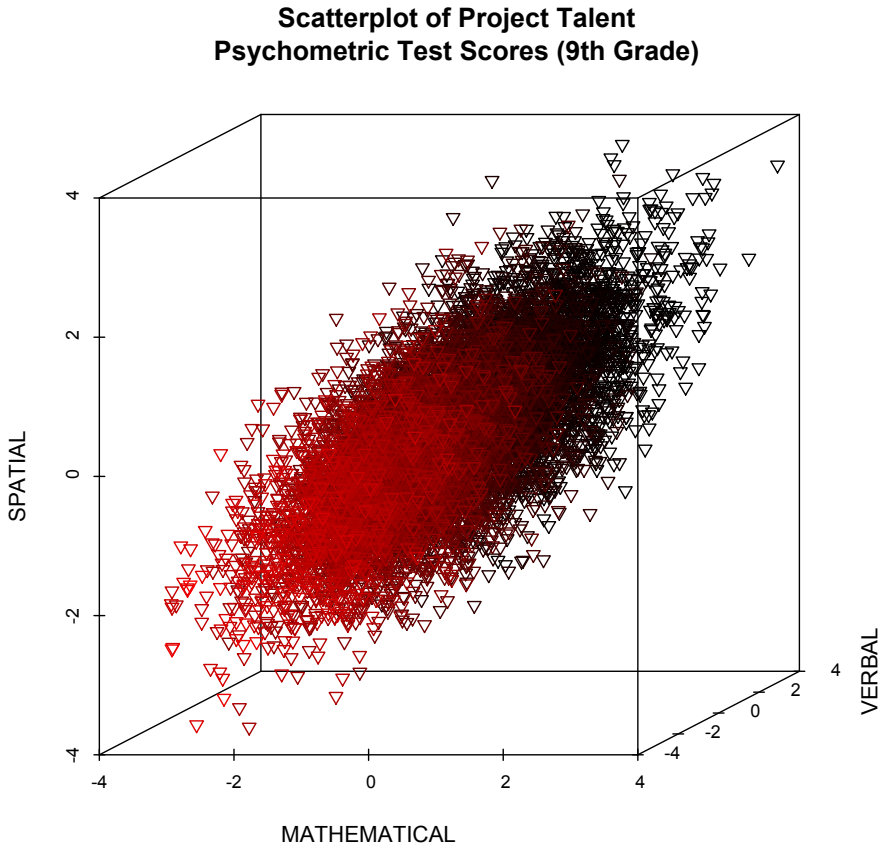
II. Test a lot of people.

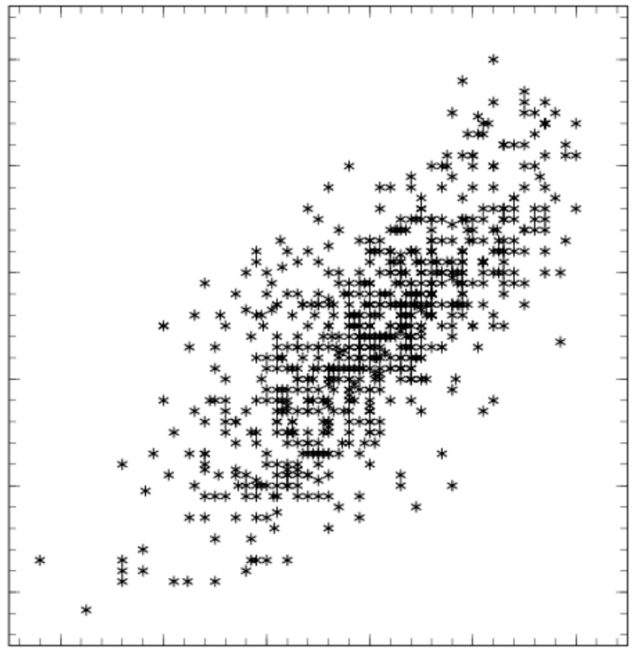
$$\{\text{individual}\} \rightarrow \text{n vector} \rightarrow \text{scalar (single number)}$$

(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 ≈ Dimension identified by factor analysis =

g factor = general factor of intelligence ≈ IQ ≈ SAT ≈ GRE

≈ overall goodness of cognitive functioning?

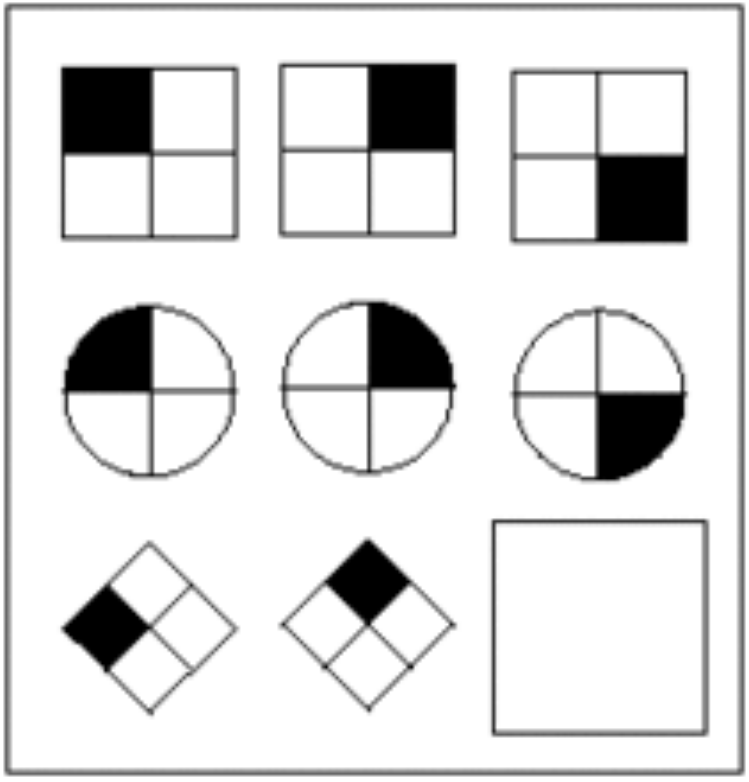
• Note these are *population level* correlations – compression may not work for a particular individual: value of g may not predict individual components of n-vector very well. But works for ”typical” individuals.

• 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



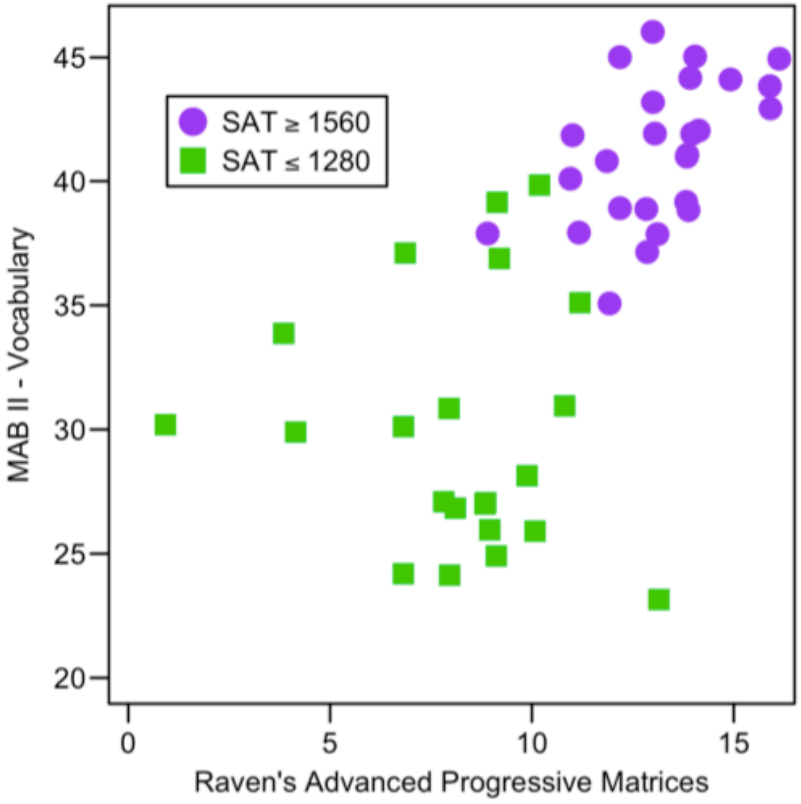
Highly g loaded but relatively culture neutral and abstract.

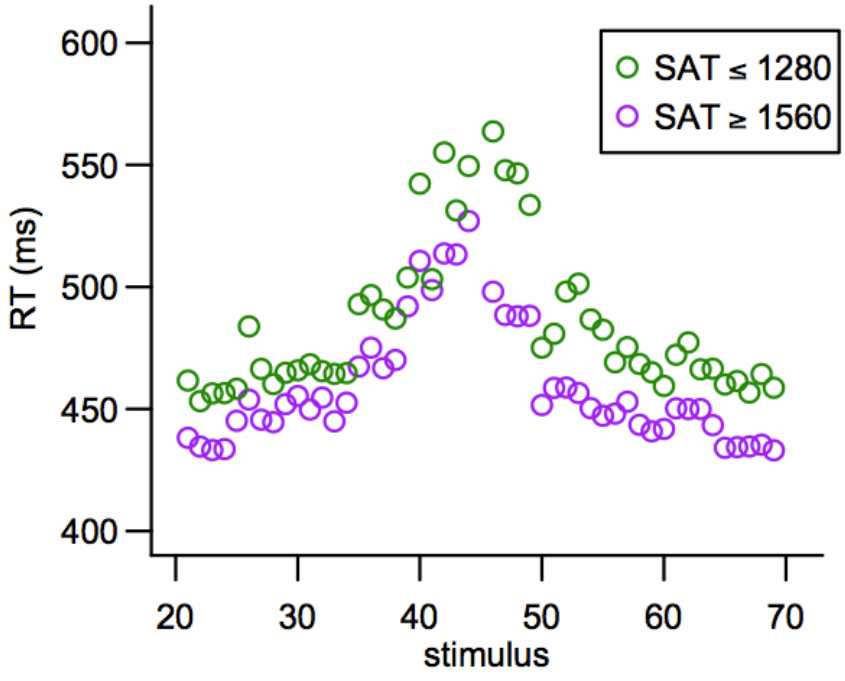
Pattern recognition and algorithmic reasoning.

Results

left fig. Vocabulary, SAT and RAPM intercorrelation.

right fig. Reaction time differences for two groups.





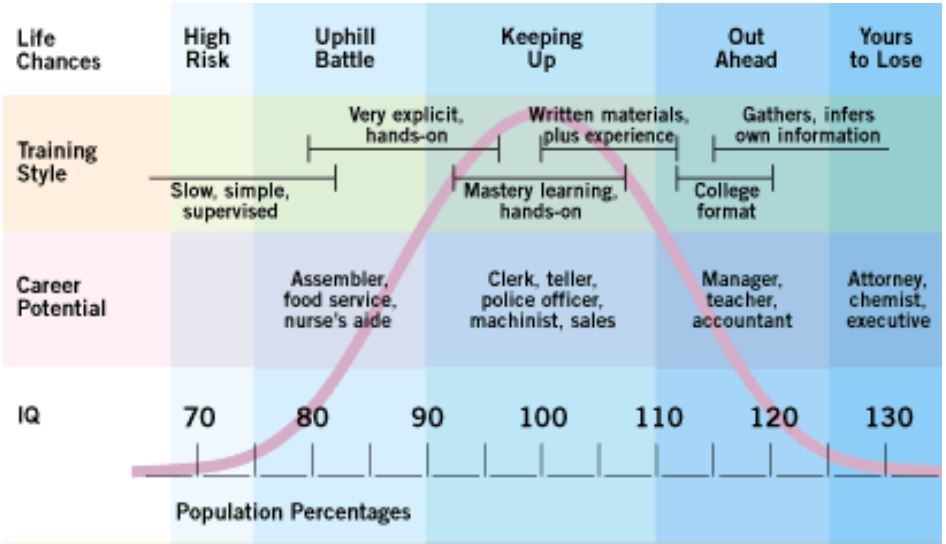
What good is it?

PRO: 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 studies). Gulp!

CON: Only explains small fraction of variance in life outcomes.

Life outcomes



	Population Percentages				
Total population distribution	5	20	50	20	5
Out of labor force more than 1 month out of year (men)	22	19	15	14	10
Unemployed more than 1 month out of year (men)	12	10	7	7	2
Divorced in 5 years	21	22	23	15	9
Had illegitimate children (women)	32	17	8	4	2
Lives in poverty	30	16	6	3	2
Ever incarcerated (men)	7	7	3	1	0
Chronic welfare					

Can further control for SES (socio-economic status) by considering sibling pairs with different IQs.

College outcomes: two factor working model

Factor 1: SAT (cognitive ability)
Factor 2: Conscientiousness, work ethic, motivation ...

These factors are only weakly correlated with each other.

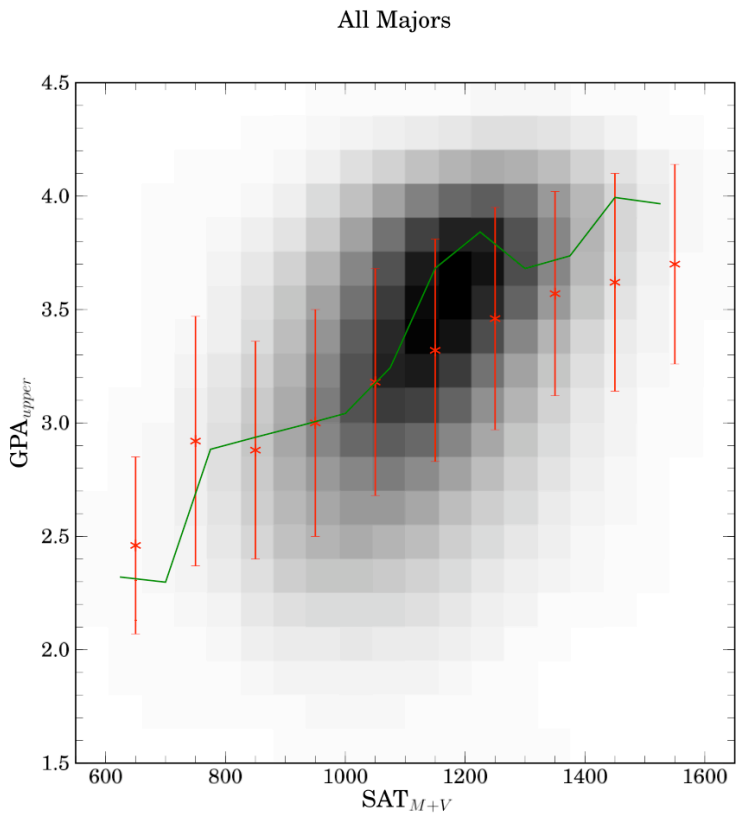
To what extent can Factor 2 compensate for Factor 1? For fixed values of SAT, what is the range of outcomes in college performance?

Are there *cognitive thresholds* for certain subjects, such that mastery is very unlikely below a certain SAT threshold (i.e., no matter how dedicated or hard working the student)?

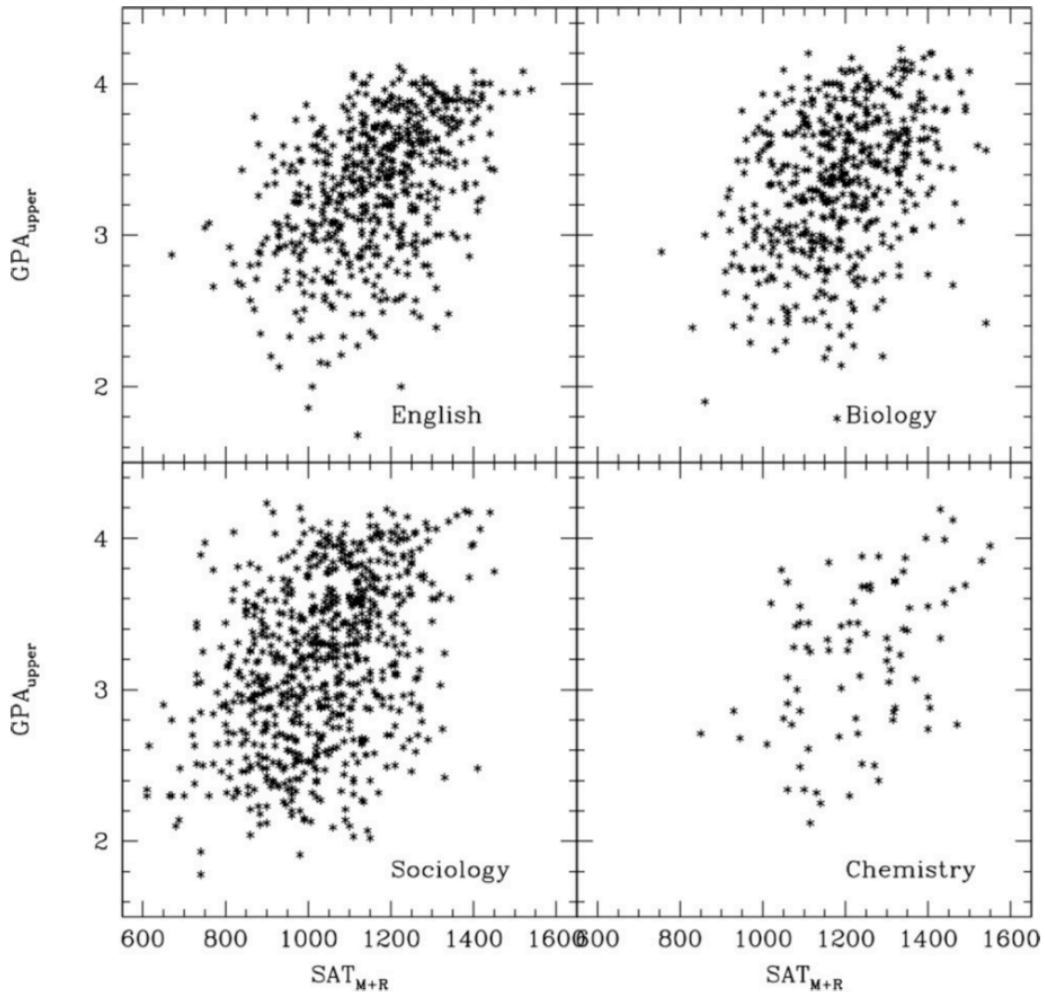
College outcomes

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

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

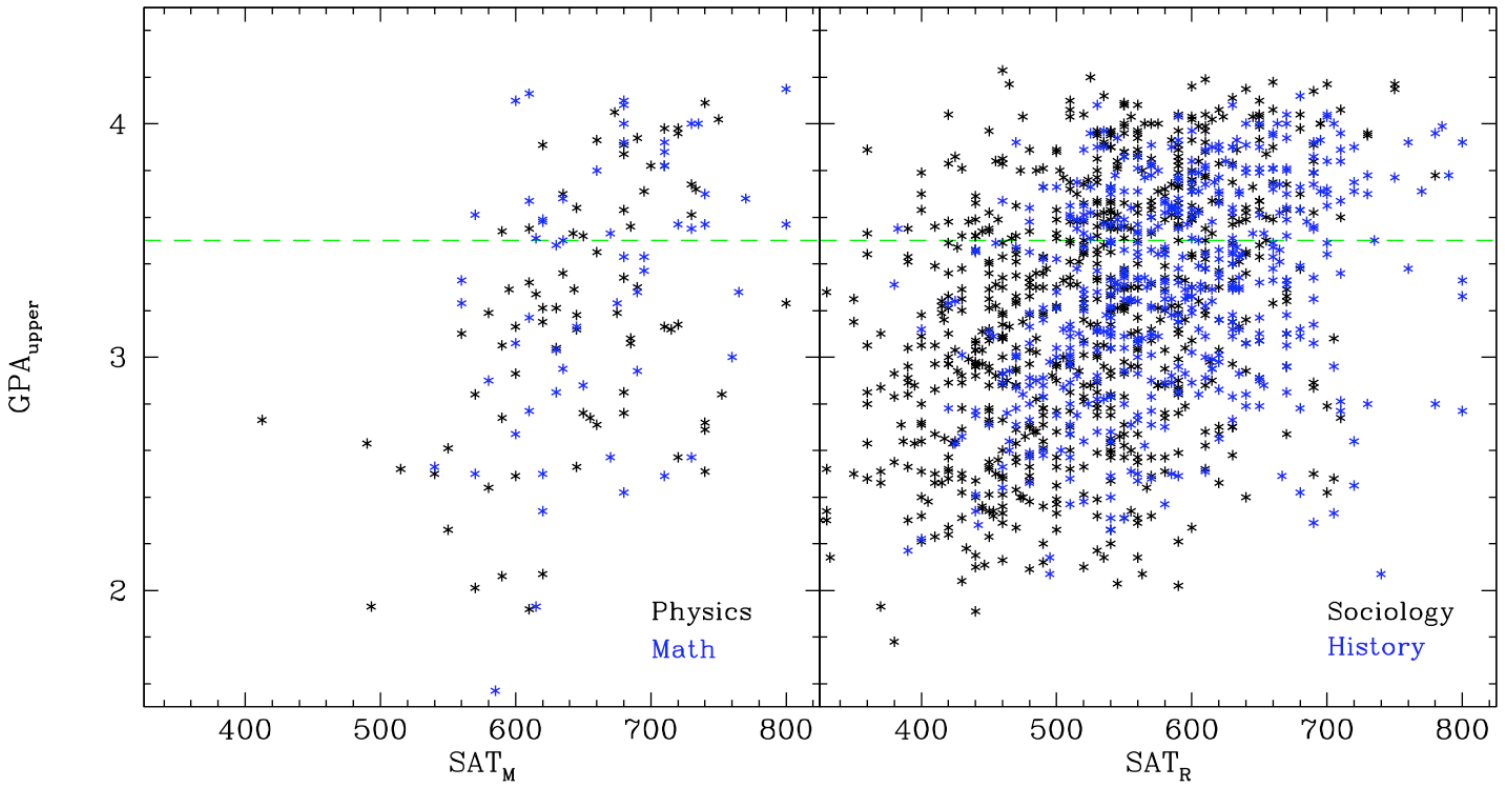


College outcomes

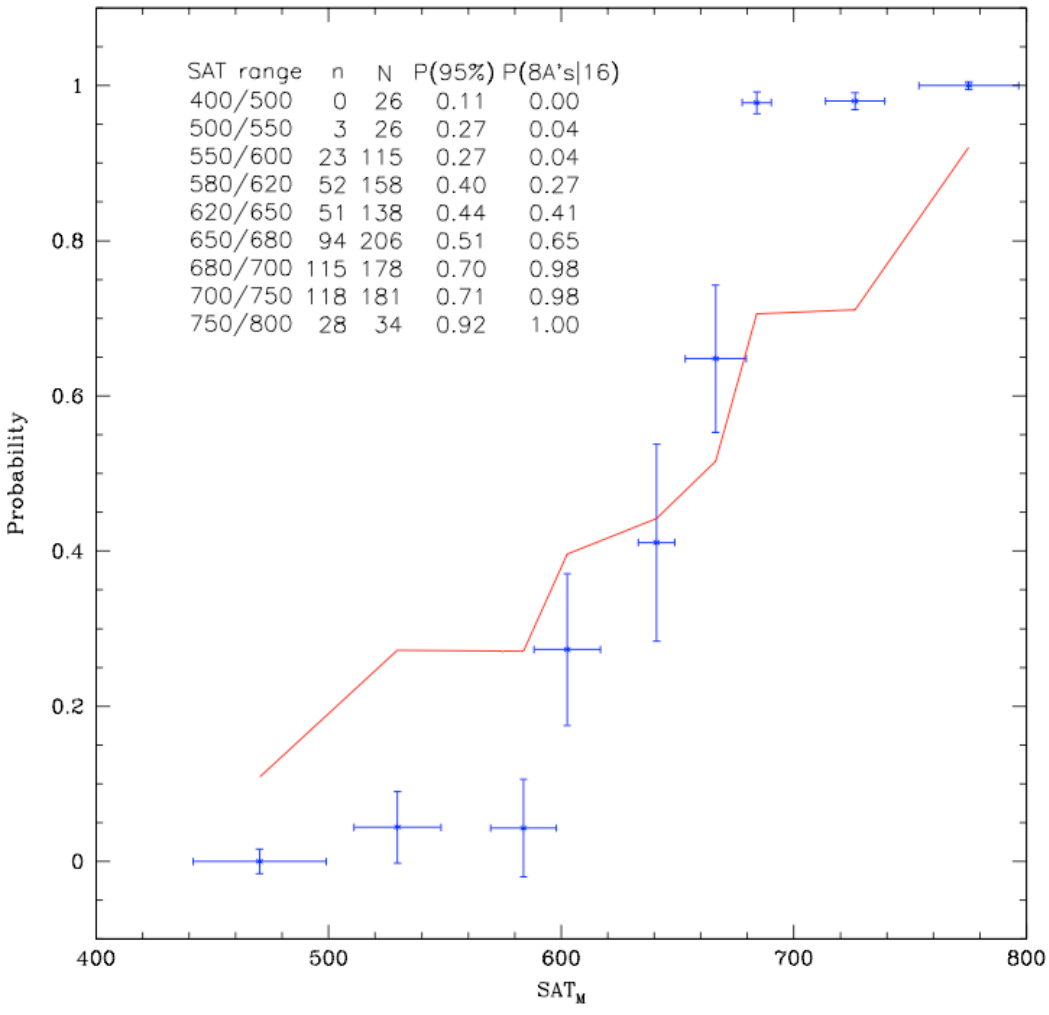


College outcomes: thresholds?

Nonlinear Psychometric Thresholds for Physics and Mathematics, Hsu and Schombert, arXiv:1011.0663



College outcomes: thresholds?



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 :-)

Answer: IT DEPENDS ...

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.

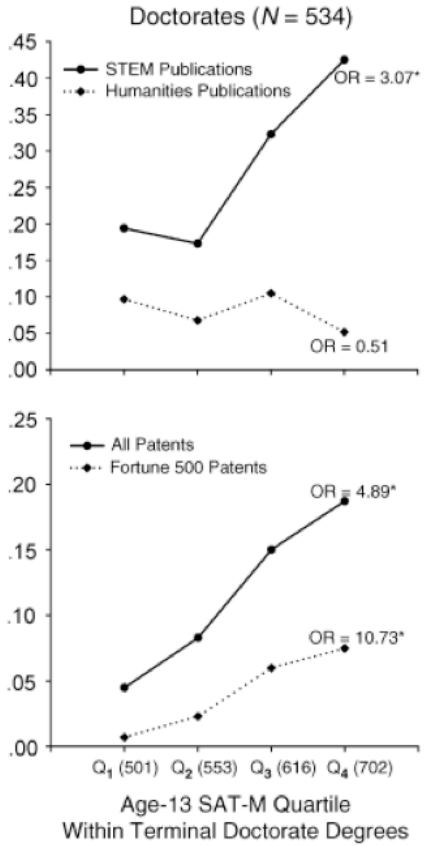
Mean score in both categories was roughly +4 SD.

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.



Your kids and regression

Assuming parental midpoint of n SD above the population average, the kids' IQ will be normally distributed about a mean which is around +.6n with residual SD of about 13 points. (The .6 could actually be anywhere in the range (.5, .7), but the SD doesn't vary much from choice of empirical inputs.)

So, e.g., for n = 4 (parental midpoint of 160 – very smart parents!), the mean for the kids would be 136 with only a few percent chance of any kid to surpass 160 (requires ~ 2 SD fluctuation). For n = 3 (parental midpoint of 145) the mean for the kids would be 127 and the probability of exceeding 145 less than 10 percent.

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

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.** ...

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.

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 $n \to \infty$.

Tests of additive variance (linearity) using height

Regress genetic distance (measured in SNP differences) on difference in phenotype height.

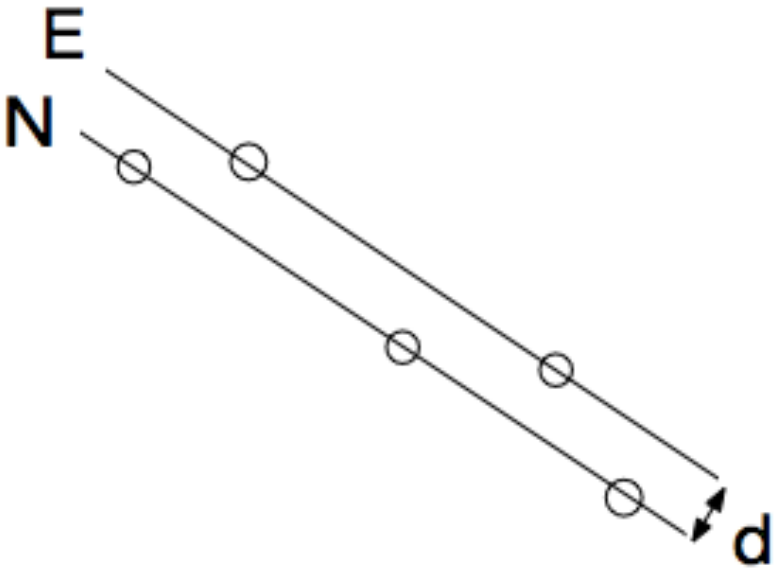
Each SD of height corresponds, on average, to about 30 SNP differences.

For example, a 2 meter tall male has about 100 more (+) alleles than an average male.

Can also estimate average effect size and number of causal loci: ~ 5000.

Note there are many different ways (at the genetic level) to be tall or smart!

Genetic distance between surfaces of constant phenotype



Typical distance between individuals: $3 \cdot 10^5 \pm 10^4$ SNPs.

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

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* genetic variance:

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

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!

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 300 genes found so far correlated to height, with > 100K pheno-genotype pairs analyzed. Only 10 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.

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. Eventually, 1000 genomes sequenced per day at less than $ 1000 per genome.

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 perform whole genome sequencing later this year.

Search for additional volunteers among this population under way – please volunteer!

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

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!

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

What can we do with this information?

Please help! Free genotyping!

If you are cognitively gifted, please participate in our study.
www.cog-genomics.org

Free genotyping: initially SNPs, later possibly exome or whole genome sequencing.

Learn about your ancestry and health.

Please help! Free genotyping!

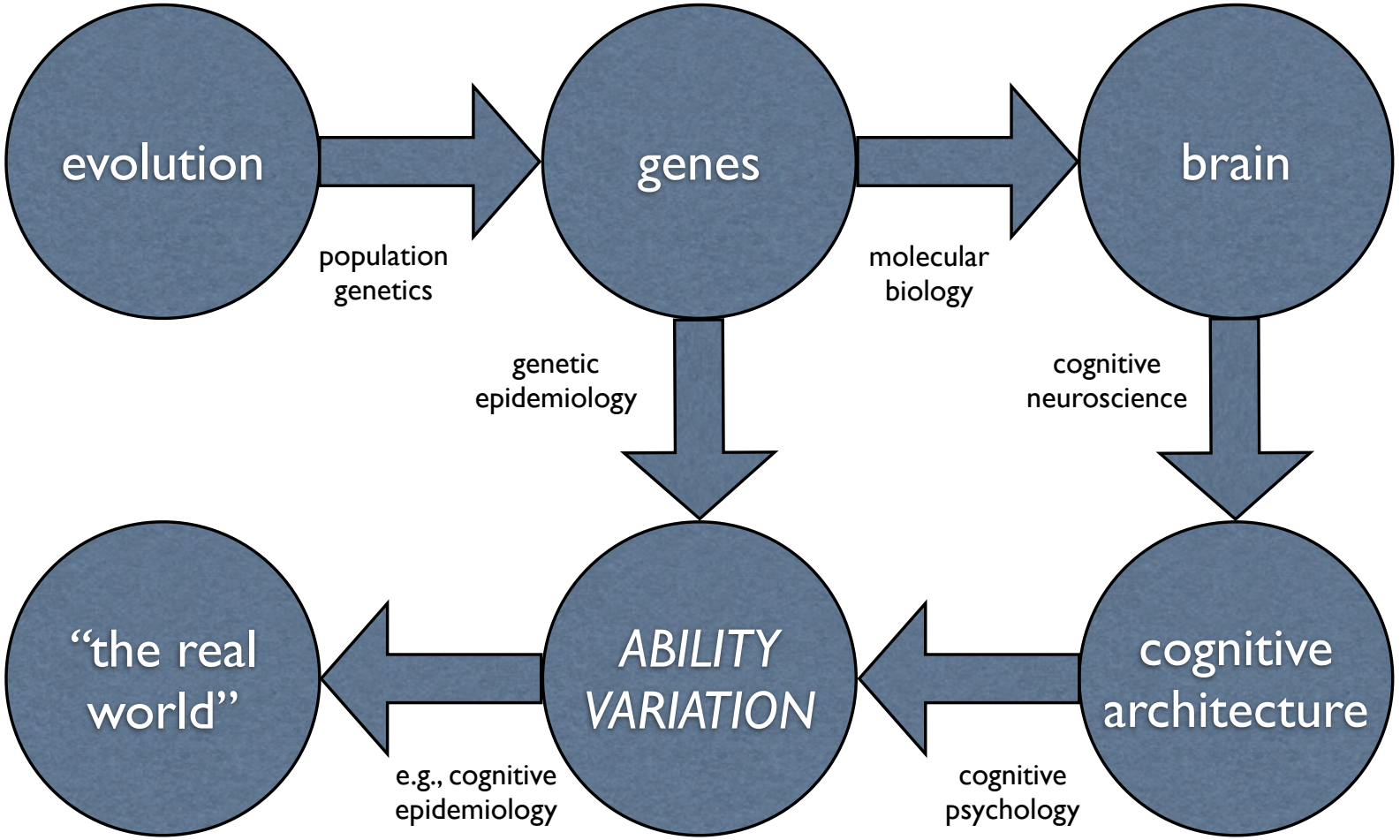
Some automatic qualifying criteria:

SAT (post 1995) M800 V760
SAT (pre 1995) M780 V700
GRE Q800 V700
ACT 35

PhD from top 5 ranked US program in physics, math, EE, theoretical computer science.

Honorable mention or better (top 50 or so in US) in Putnam competition.

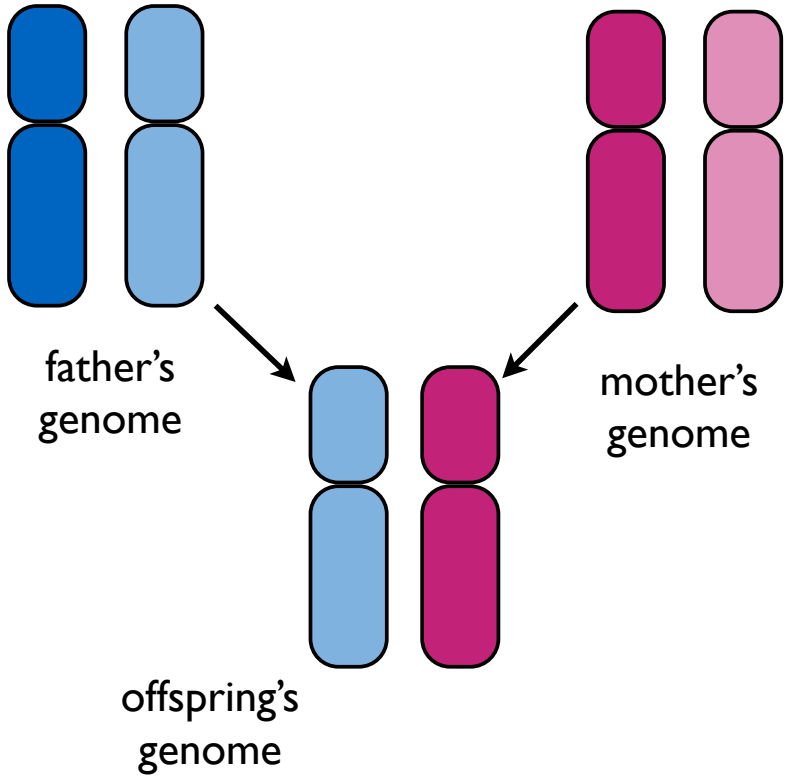
Consider making a donation!



The Future: genetic engineering

- Random segregation is a large part of the reason why children can differ so much from their parents and from each other.

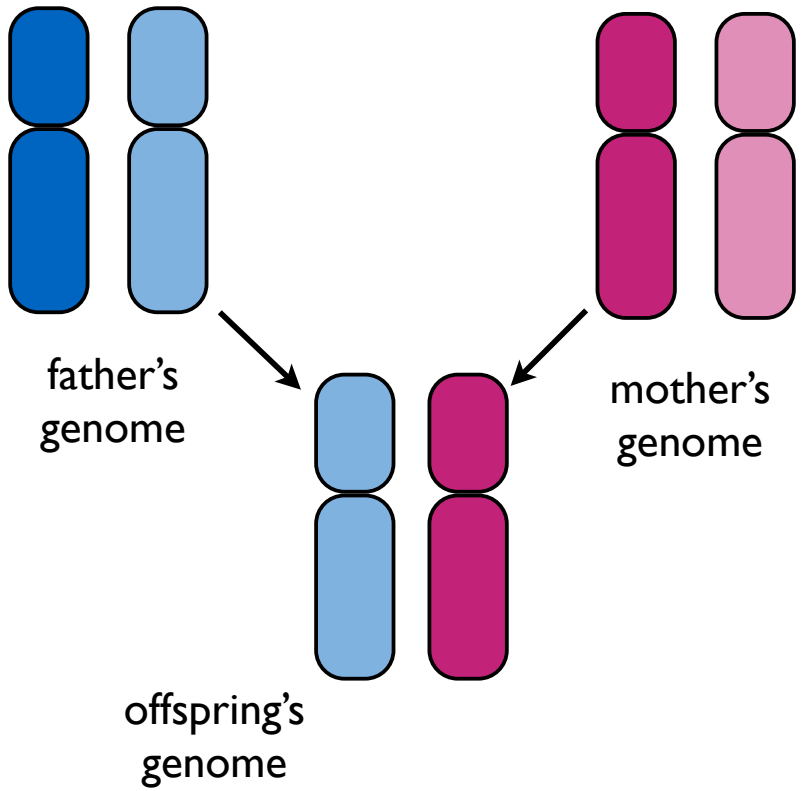
- "The stars above us govern our conditions; else one self mate and mate could not beget such different issues."—WILLIAM SHAKESPEARE, *KING LEAR*



The Future: genetic engineering

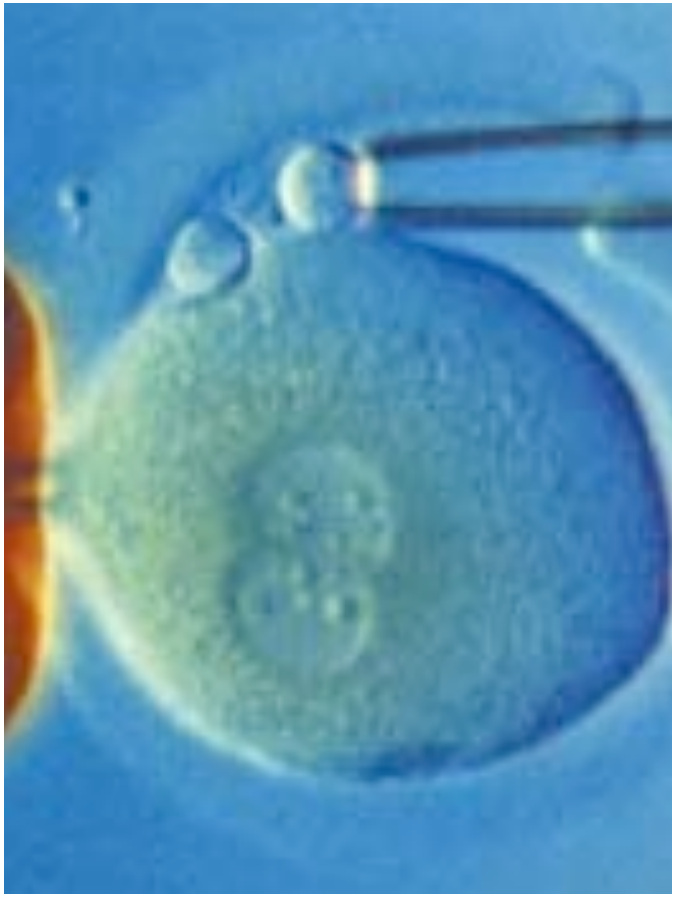
- Suppose that two parents are heterozygous at **all** trait-affecting loci. If a child is lucky and inherits many + alleles, he will be smart. If a child is unlucky and inherits many − alleles ...

- But suppose that the parents can **choose** which allele at a heterozygous locus to pass on. How might this be done?



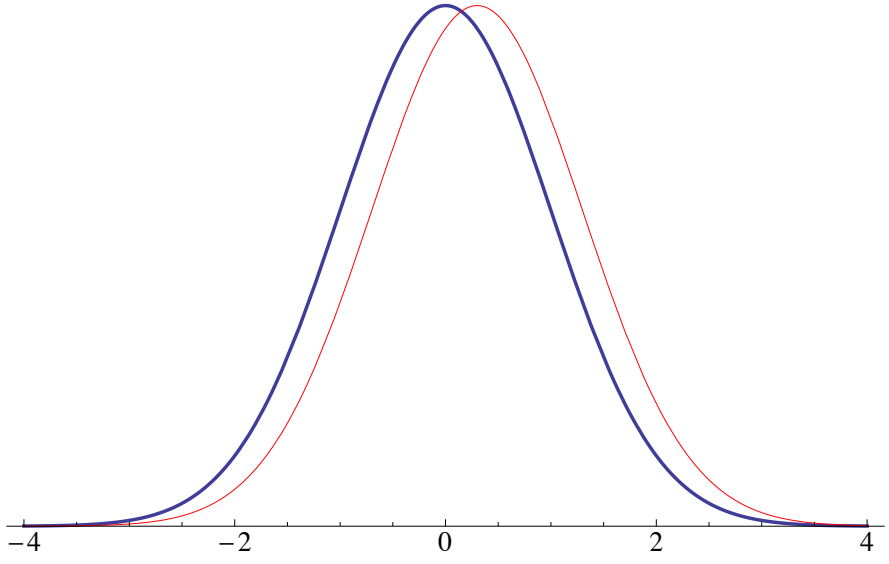
The Future: genetic engineering

- Suppose that we can non-destructively sequence gametes (sperm and egg cells).
- We can imagine parents **choosing** which gametes to unite in order to constitute their offspring.
- In particular, they might choose to unite gametes bearing many *g*-enhancing alleles.



The Future: genetic engineering

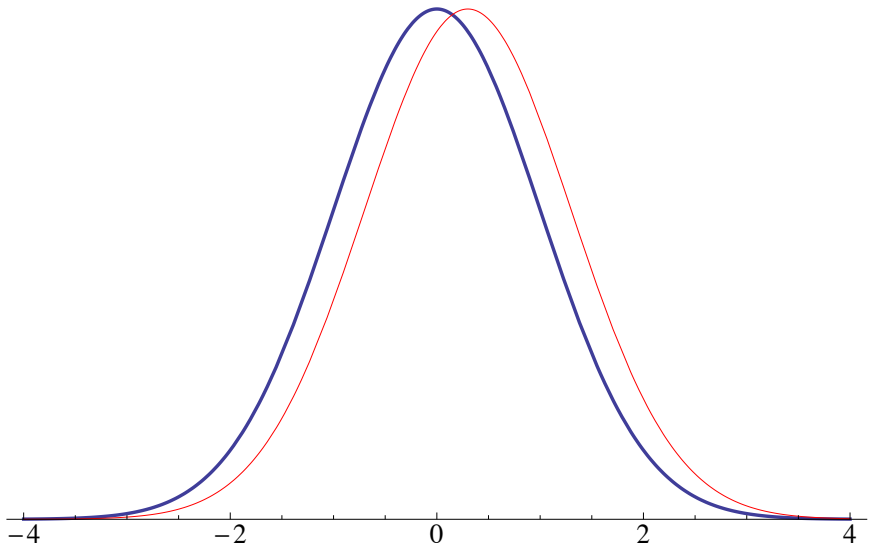
- Suppose that we have **100** *g*-affecting loci, each with minor allele frequency (MAF) of 0.1 and an average effect of 0.02.
- Instead of using a random sperm cell to conceive, a couple might go through 500 cells and choose the one with the **most** + alleles.
- If **all** couples in the population do this, what will happen to the level of *g*?



The Future: genetic engineering

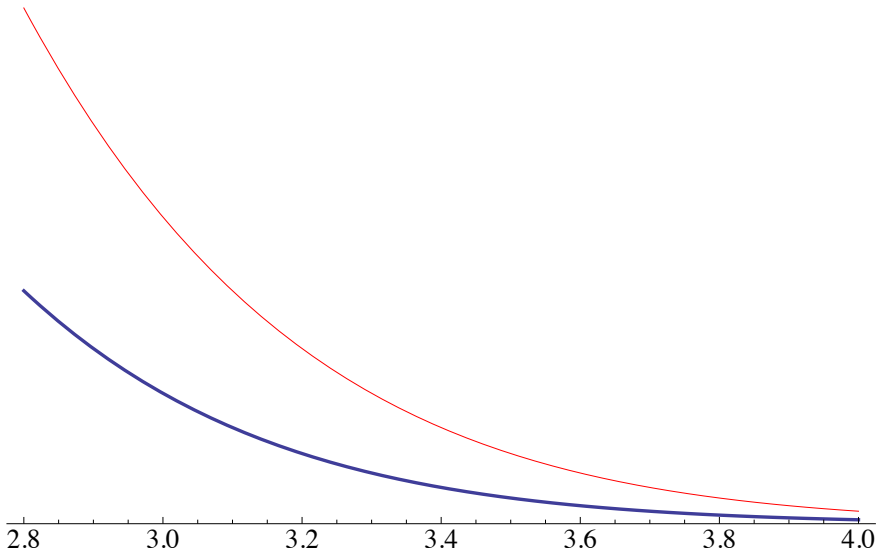
- In the offspring generation, the mean level of *g* will increase by **0.2 SDs**.

- Consequences will be especially prominent at the tails ...



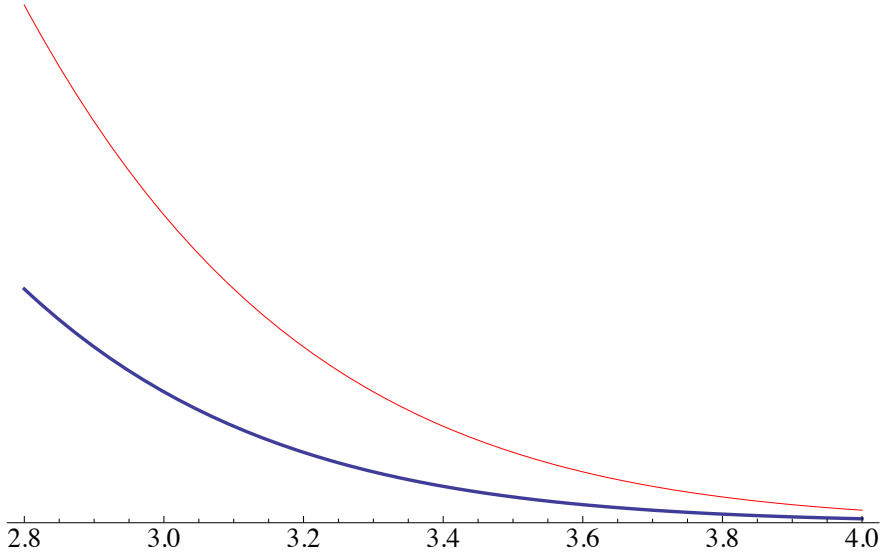
The Future: genetic engineering

- In the offspring generation, the mean level of *g* will increase by **0.2 SDs**.
- Consequences will be especially prominent at the tails ...



The Future: genetic engineering

- There will be more than **twice** as many individuals exceeding 4 SDs above the parental mean.
- Recall the accomplishments of the SMPY cohorts. An enrichment of individuals blessed with this potential!
- Since there are hundreds (if not thousands) of manipulable loci, this would be only the beginning ...



The Future of Human Intelligence

- "Suppose we knew, for instance, twenty [loci affecting] mental characters. These would combine in over a million [homozygous] mental types. In practice each of these would naturally occur rather less frequently than one in a billion, or in a country like England, about once in 20,000 generations.

The Future of Human Intelligence

- "It will give some idea as to the excellence of the best of these types when we consider that the Englishmen from Shakespeare to Darwin ... have occurred within ten generations; the thought of a race of men combining the illustrious qualities of these giants, and breeding true to them, is almost too overwhelming ...

The Future of Human Intelligence

- "... but such a race will inevitably arise in whatever country first sees the inheritance of mental characters elucidated."—RONALD A. FISHER, "MENDELISM AND BIOMETRY"