Genetic architecture and predictive modeling of quantitative traits

Steve Hsu

MSU and BGI
www.cog-genomics.org

References

Collaborators: Chris Chang (BGI/CG), Carson Chow (NIH), Chiu-Man Ho (MSU), James Lee (Minnesota), Laurent Tellier (BGI), Shashaank Vattikuti (NIH)

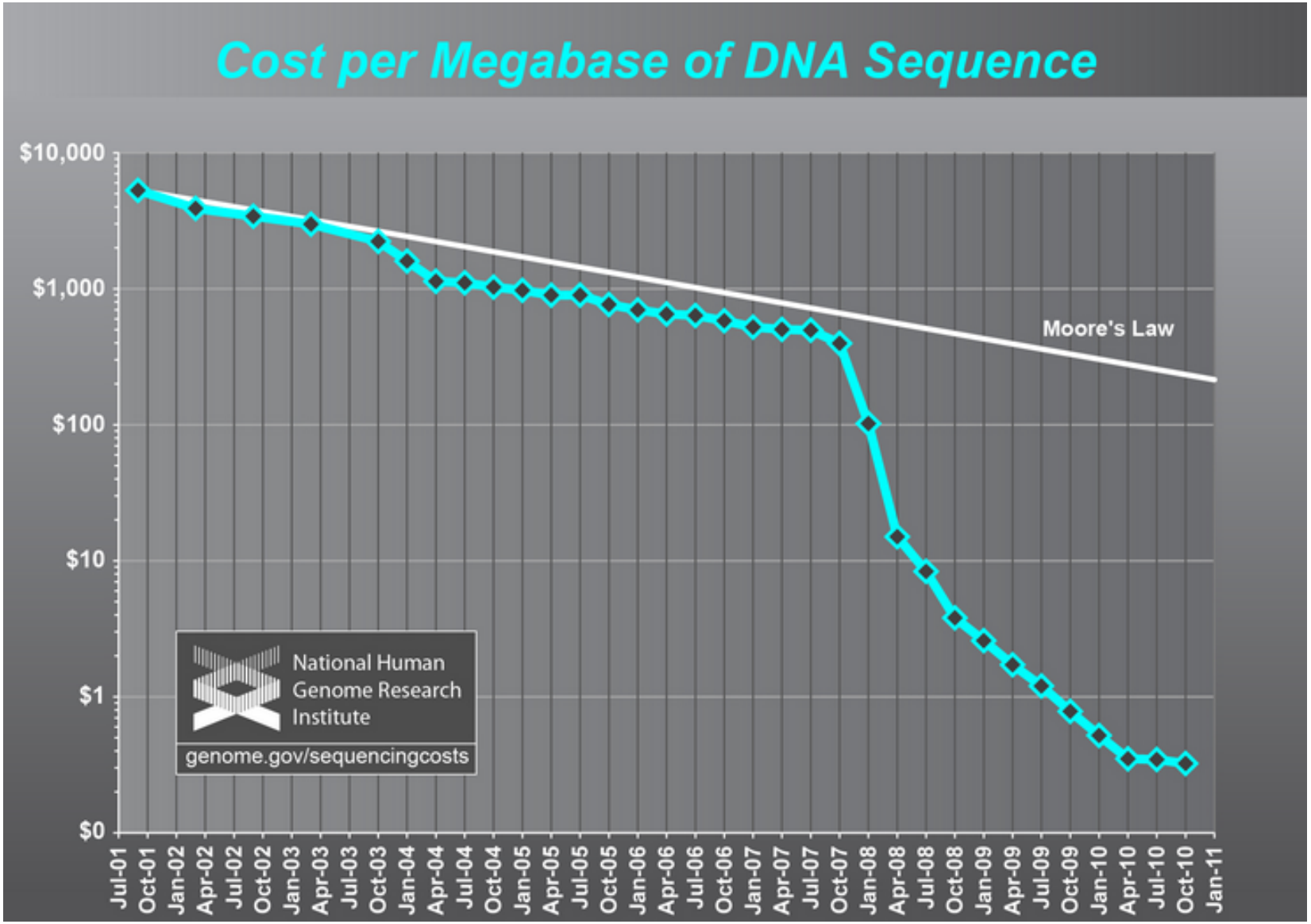
http://www.gigasciencejournal.com/content/3/1/10
http://arxiv.org/abs/1408.6583
http://arxiv.org/abs/1408.3421

Quick Summary

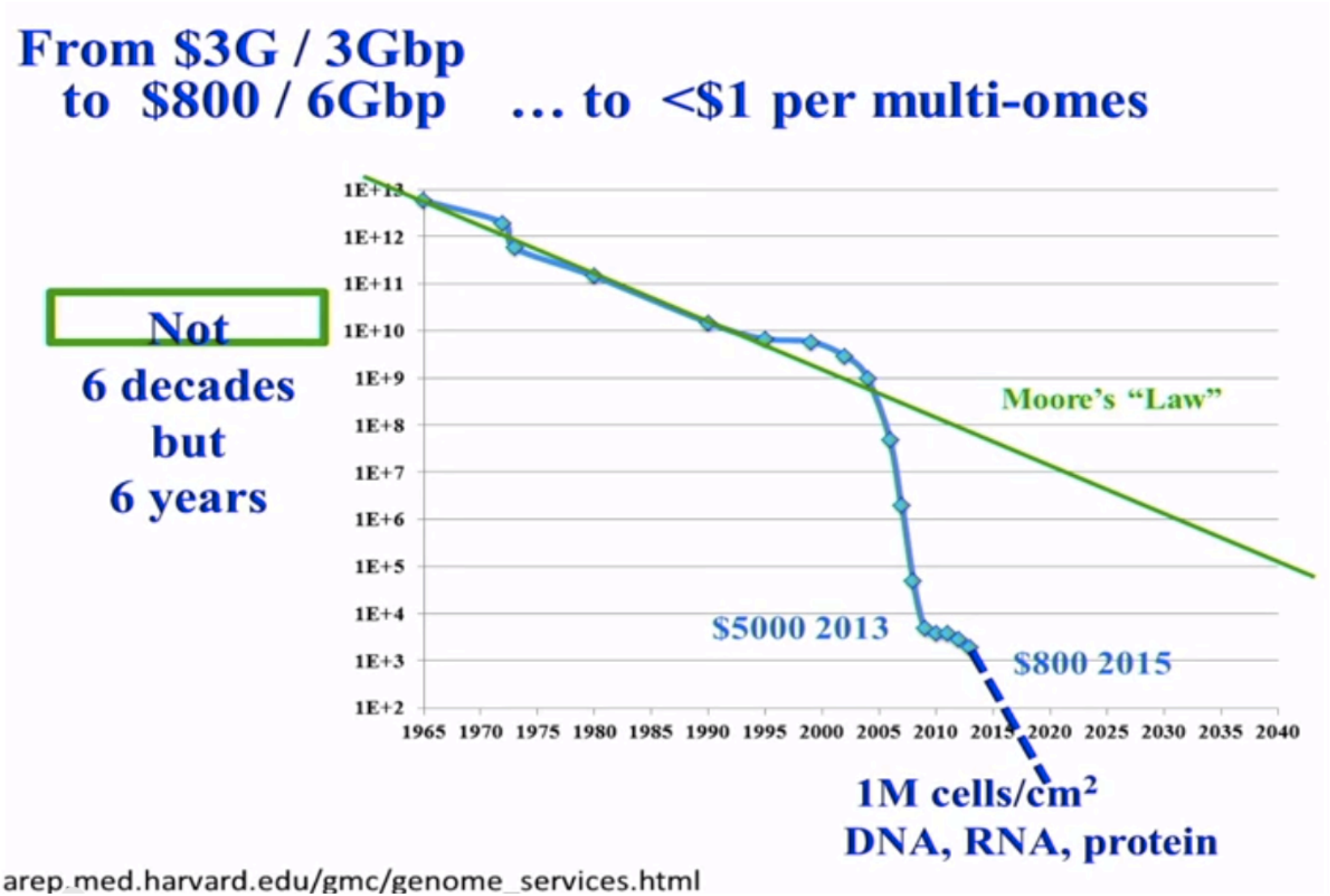
This talk will explain the background for the following prediction:

Using (phenotype | genotype) data sets with ~ million individuals, we can build predictive models for many complex traits, including a variety of human disease susceptibilities (e.g., with additive heritability h2 ~ 0.5).

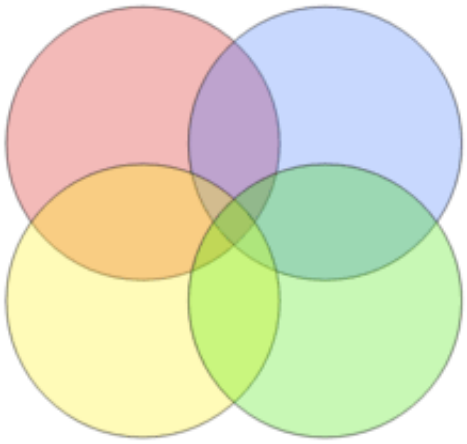
Why is a theoretical physicist doing genomics?



Plateau to end soon?



Human Capital: progress will require multiple domains of expertise



Genomics

Population Genetics

Behavior Genetics / Psychometrics

Algorithms and High Dimensional Statistics

Complex phenotypes

Many traits with $h^2 \sim 0.5$ and numerous causal variants.
Height, Cognitive Ability: $h^2 \sim 0.8$

Medical Condition / Topic	Heritability Est.	References
Alcoholism	50 - 60%	[PMID 19785977]
Alzheimer's disease	58 - 79%	[PMID 16461860]
Anorexia nervosa	57 - 79%	[PMID 19828139]
Asthma	30%	[PMID 16117840]
Attention deficit hyperactivity disorder	70%	[PMID 22833045]
Autism	30 - 90%	[PMID 17033636]
Bipolar disorder	70%	[PMID 14601036]
Bladder cancer	7 - 31%	[PMID 21927616]
Blood pressure, diastolic	49%	[PMID 19858476]
Blood pressure, systolic	30%	[PMID 22479213]
Body mass index	23 - 51%	[PMID 25383972, PMID 18271028]
Bone mineral density	44 - 87%	[PMID 15750698, PMID 16025191]
Breast cancer	25 - 56%	[PMID 11979442, PMID 2491011]
Cervical cancer	22%	[PMID 11979442]
Colon cancer	13%	[PMID 11979442]
Coronary artery disease	49%	[PMID 15710764]

General model for quantitative phenotype

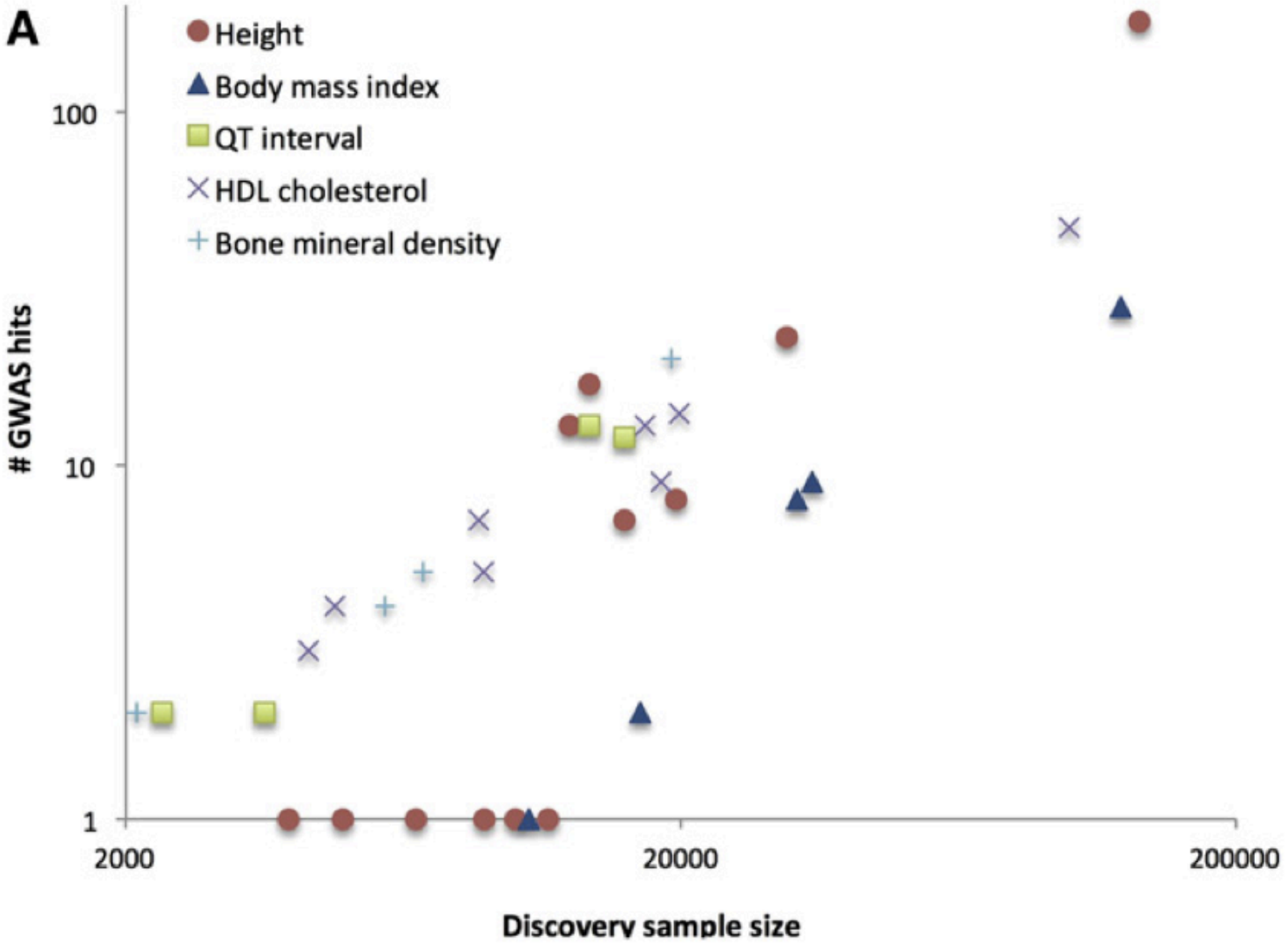
y = individual phenotype
g_i = individual genotype (e.g., list of 1M SNPs or 3B loci)
x_i = linear effect sizes
z_{ij} = tensors of nonlinear effect sizes

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

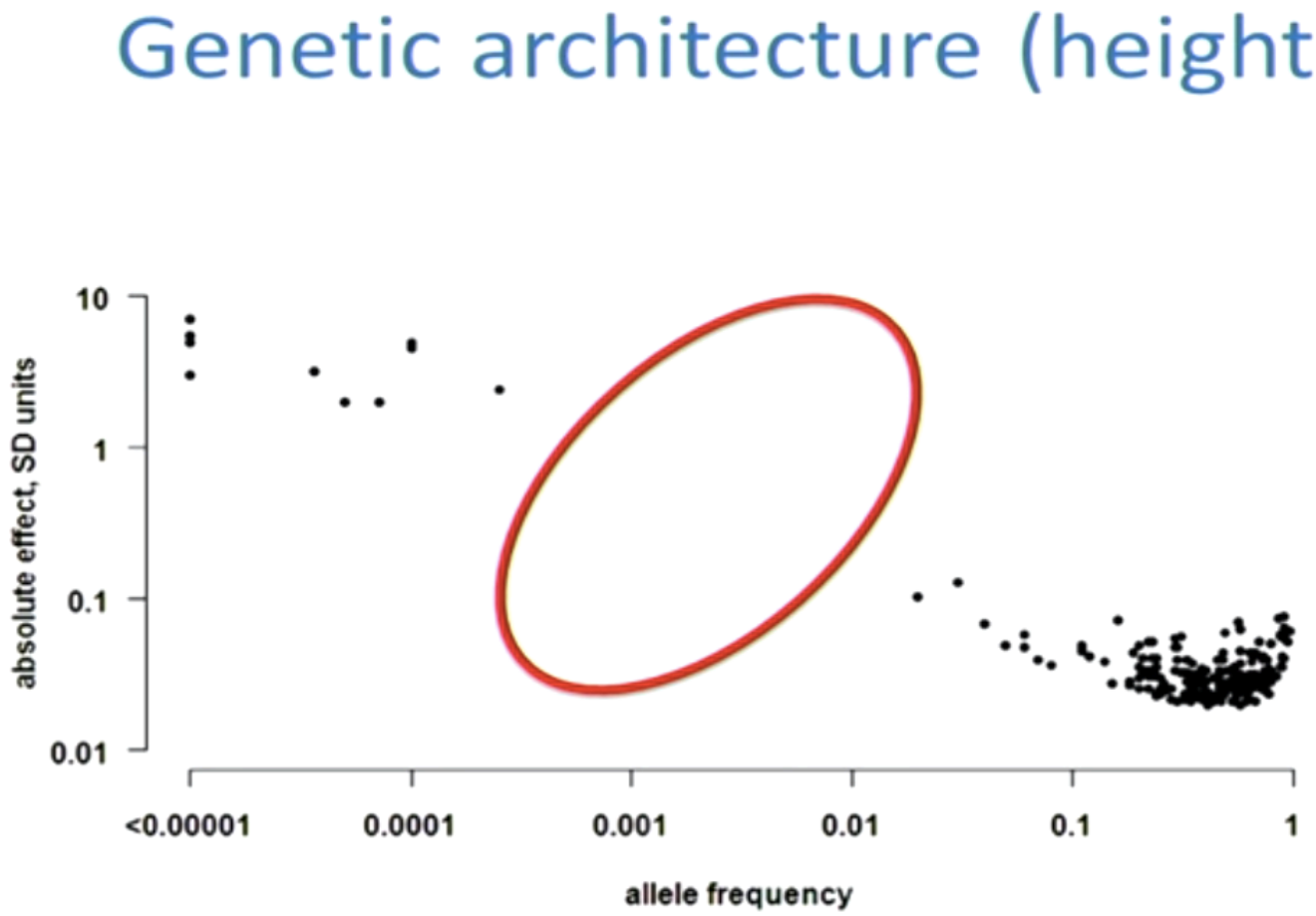
Plausible that linear term dominates (largest component of variance), even if nonlinear terms are important in certain circumstances.

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

GWAS history

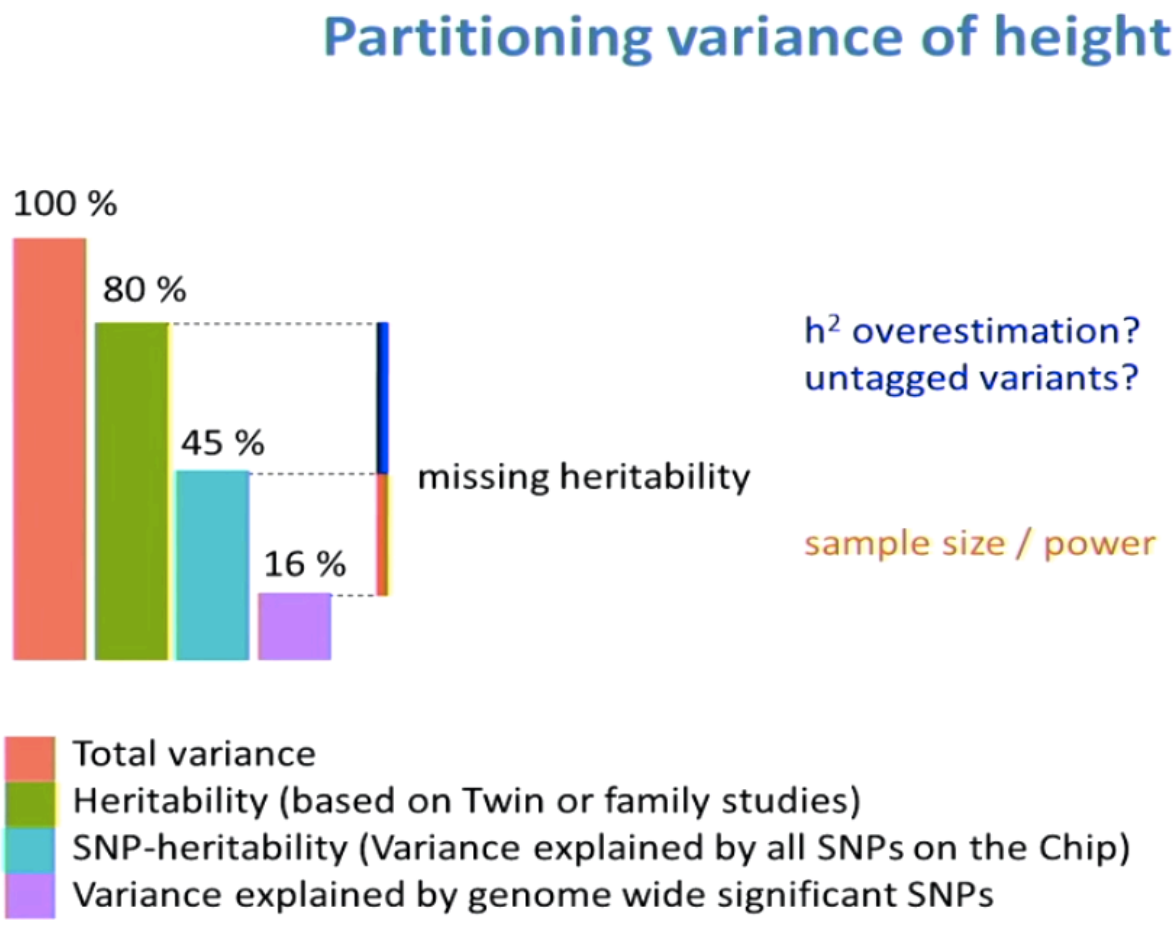


Common vs Rare Variants



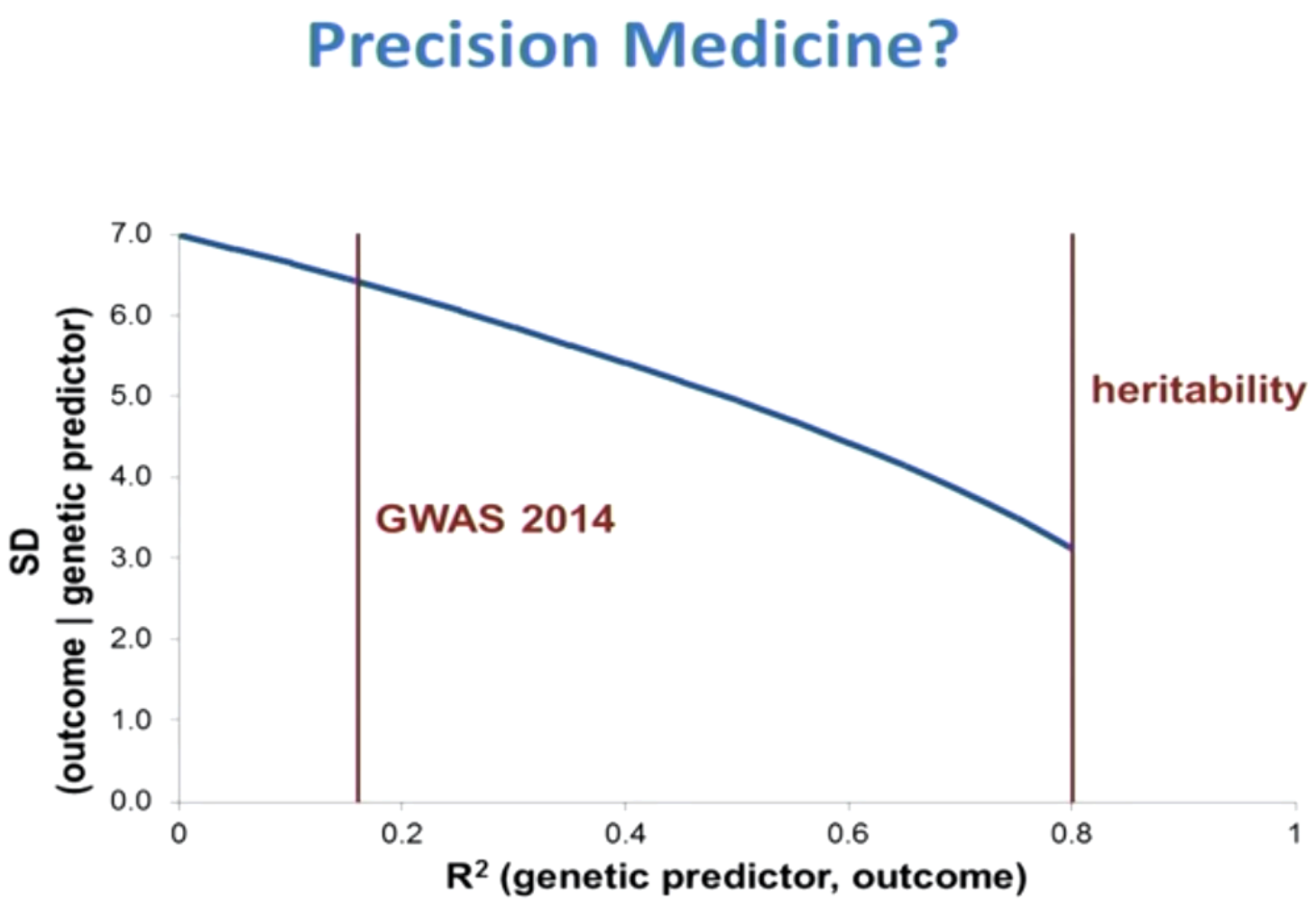
Kemper et al. (2012) Genome Biology

No dark matter



Slide by Robert Maier

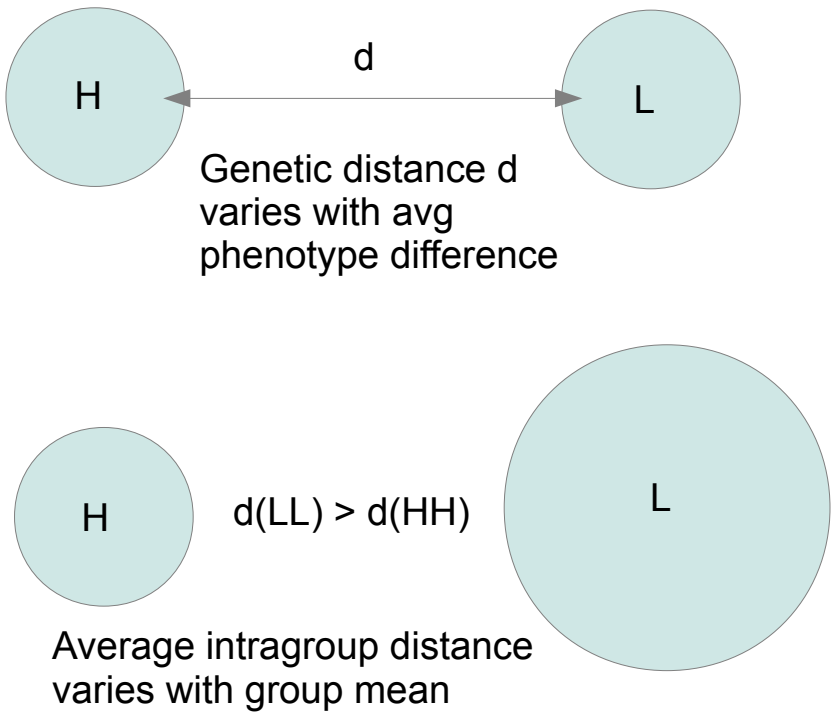
Predictive accuracy



How many causal loci? Height / Cognitive Ability (g)

Rough estimate: $s \sim 10k$

Based on analysis of pairwise genetic distance versus phenotype difference: ~ 100 or fewer SNPs per SD of height or g.



Compressed Sensing

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

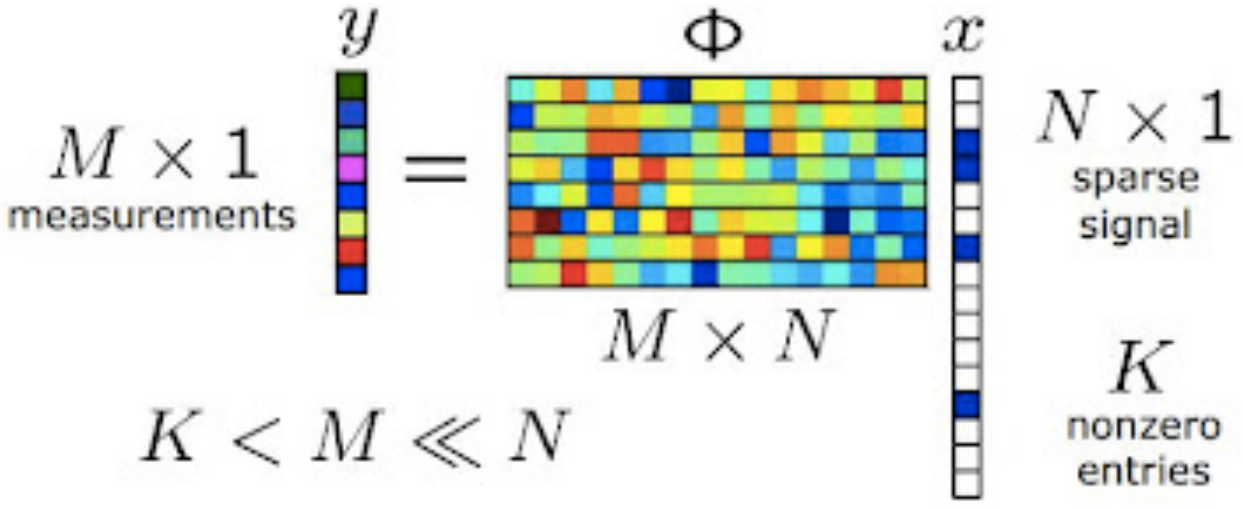
$$y_i = \sum_j G_{ij}\, x_j + \epsilon_i$$

1. x is sparse (e.g., $s \sim$ 10k causal variants among $p =$ 1M SNPs)

2. at least for next few years, an *underdetermined* problem: $p \gg n$

Surprising, and nearly optimal results, from Compressed Sensing (L1 norm penalty enforcing sparseness; LASSO). Required data scales as $s \log p$.

Compressed Sensing



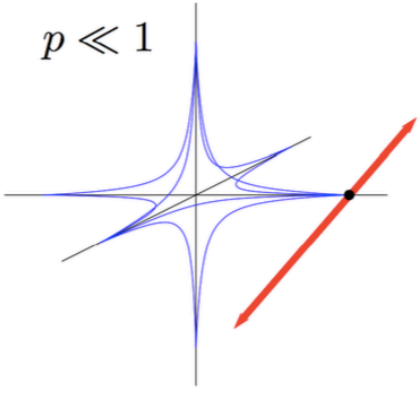
The most valuable list of ~ 10k numbers in the world? :-)

L1 penalization enforces sparsity

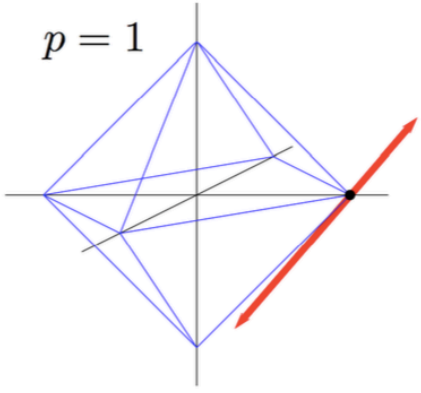
Now consider, for some fixed $p > 0$, the optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \underbrace{\|\boldsymbol{x}\|_p}_{\sqrt[p]{\sum_n |x_n|^p}} \quad \text{s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

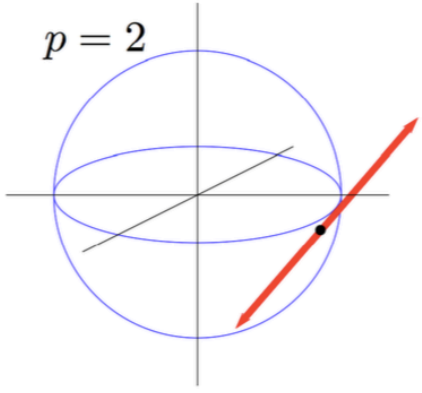
The solution can be found by growing the ℓ_p-ball until it touches the ϵ-rod:



Solution definitely sparse but problem is **NP hard**.



Solution usually sparse and problem is **convex**!



Solution is **not sparse**; $\Leftrightarrow$ LS when $\epsilon = 0$.

This suggests to use the ℓ_1 norm as a surrogate for the ℓ_0 norm!

Compressed Sensing

Objective function:

$$O = ||y - G\hat{x}||_{L2} + \lambda||\hat{x}||_{L1}$$

O is convex, so optimization is fast. Many recently proved theorems: performance guarantees given sufficient conditions on G, largely independent of properties of x other than sparsity.

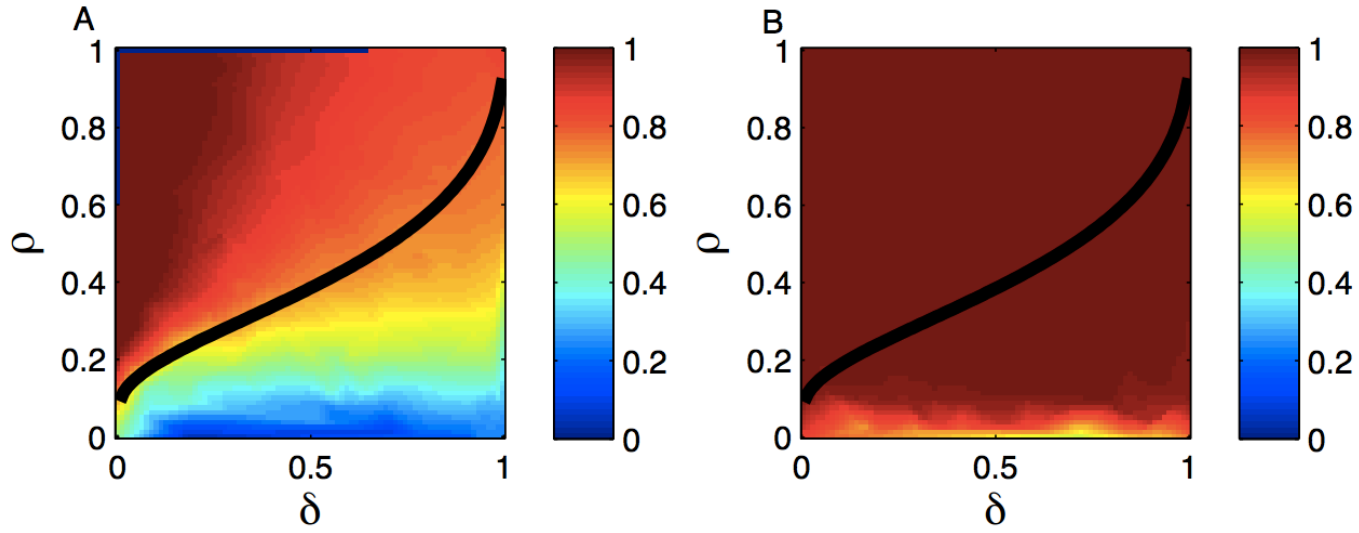
Phase transition in performance of algorithm as a function of s/n: for g and height ($h^2 \sim 0.5$), expect at $n \sim 30s$.

In good phase, can select ALL s causal variants at once:

support of $\hat{x}$ = support of x

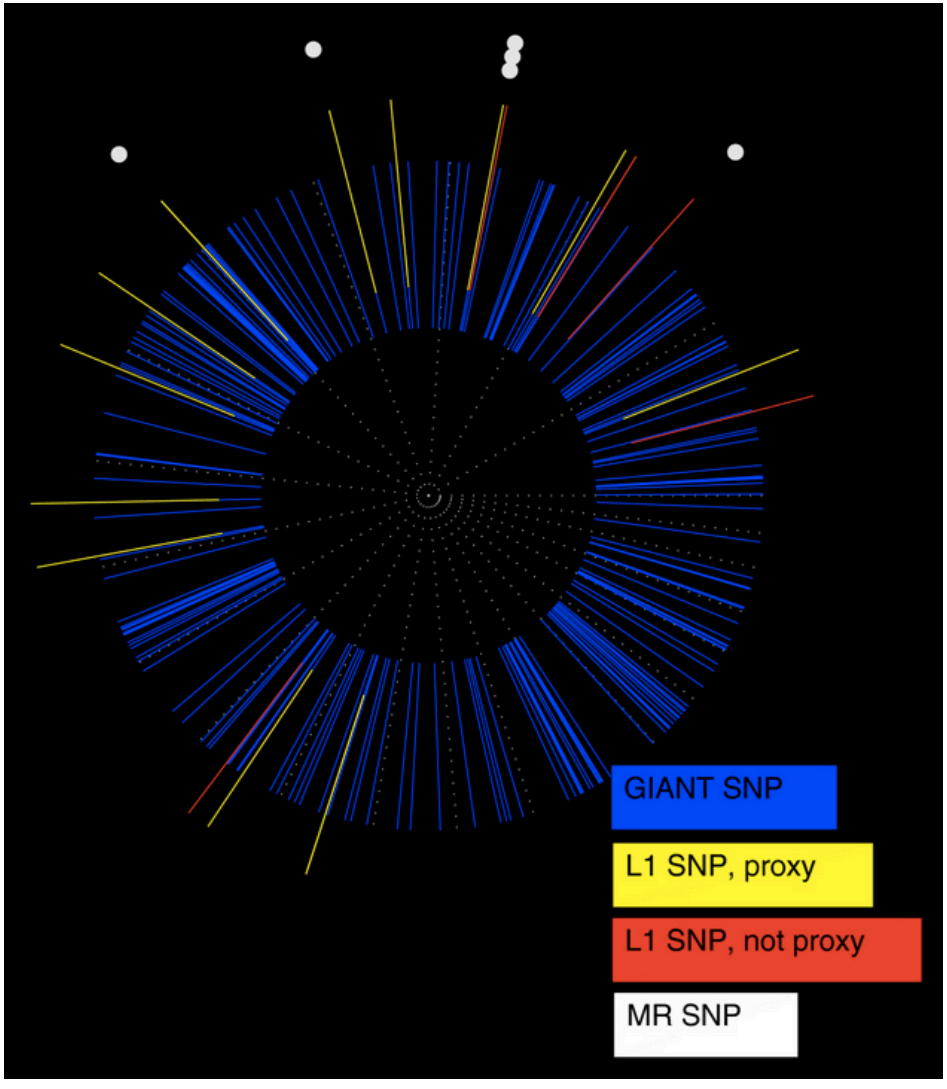
Compressed Sensing

Phase transition in performance of algorithm as a function of $\rho = s/n$: for g and height ($h^2 \sim 0.5$), expect at $n \sim 30s \sim 300$k individuals.



Need ~ 1 million phenotype-genotype pairs to extract causal loci and build predictive genomic model. Sometime in the next decade? Sooner for height.

Wrong phase: increase penalization



If sample size insufficient for full recovery (wrong phase), can increase penalization and scan for phase transition. This corresponds to an effective model with smaller s and heritability. Recovery is much better than for simple regression.

(red + yellow): $n \sim 7$k individuals.
GIANT (blue): $n \sim 130$k.

Nonlinear case

Gene-gene interactions (indices: $1 \leq a \leq n$ labels individuals and $1 \leq i, j \leq p$ label genomic loci)

$$y^a = \sum_i g_i^a z_i + \sum_{ij} g_i^a Z_{ij} g_j^a + \epsilon^a \quad ,$$

g is an $n \times p$ dimensional matrix of genomes, z is a vector of linear effects, Z is a matrix of nonlinear interactions, and ϵ is a random error term. Could include higher order (i.e., gene-gene-gene) interactions if desired. Rewrite as

$$y^a = G^a(g) \cdot X + \epsilon^a$$

where G is a nonlinear function of g but X is *linear* in z, Z.

Nonlinear case

Step 1. Run CS on (y, g) data, using linear model $y = gx$. Determine support of x: subset defined by s loci of nonzero effect.

Step 2. Compute $G(g)$ over this subspace. Run CS on $y = G(g) \cdot X$ model to extract nonzero components of X. These can be translated back into the linear and nonlinear effects of the original model (i.e., nonzero components of z and Z).

Simulation results show that with enough data to do linear analysis (Step 1), can recover half or more of nonlinear variance in Step 2.

Cognitive ability: the most interesting phenotype of all

Human brain arguably most complex known object in the universe. Constructed from a small (~ gigabit) program.

Humans and Chimps: differ at ~ tens per thousand bp. Huge gap in cognition.

Humans (modern) and Neanderthals: differ at ~ few per thousand bp (vs one per thousand among moderns). 300ky of technological stasis vs 50ky of acceleration and global impact.

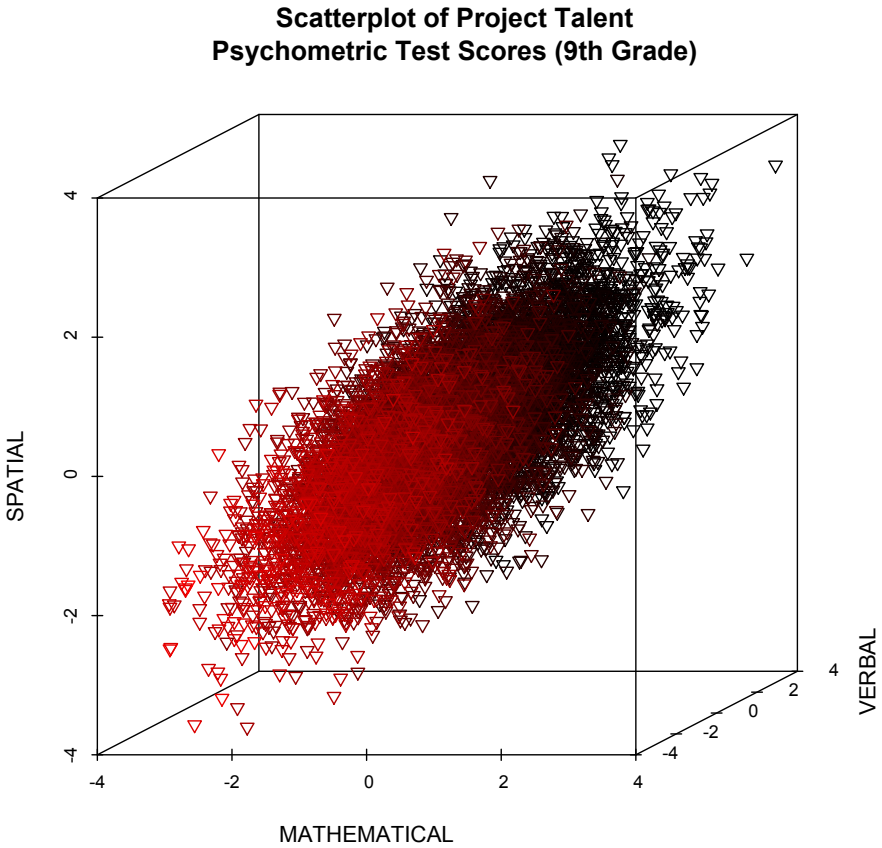
Majority of scientific / technological progress attributable to far outliers in cognitive ability (g).

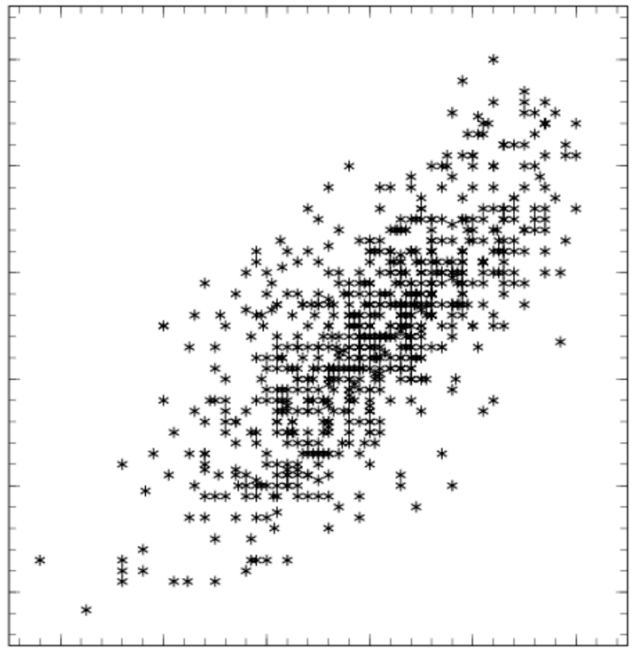
Conjecture:

Small genotype tweaks can elevate human cognition substantially.

We can (crudely) measure cognitive ability = ”g”

- 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





Cognitive ability and longevity / health outcomes

Cognitive ability measured at age 11 is a better predictor of longevity than adult body mass index, total cholesterol, blood pressure or blood glucose.

Results survive correction for family social status and economic class.

Scottish Mental Survey of 1932 and longitudinal follow up. Nature 456, 175-176 (13 November 2008) — doi:10.1038/456175a

Smarter people take better care of themselves? Or high g correlated to overall high-functioning of bodily systems?

Final thoughts

Key threshold at $n \sim$ million sample size. But must be able to implement LASSO or similar algorithm. Pooling summary statistics is not enough.

1. Need to improve data sharing incentives and infrastructure.

2. Good phenotyping is key. Sequencing is still expensive – careful planning of large studies will pay off.

3. Generalization from SNPs to whole genomes still needs work. What are key features / degrees of freedom? For example, how to treat structural variants and CNVs?

Final thoughts: NIH policy recommendation

NIH funded studies which gather (phenotype | genotype) data should be required to make the data available for pooled analysis in the cloud, under a reasonable privacy / security model.

Storage and analysis capabilities could be provided by NIH, but if not, general guidelines should be formulated for universities, institutes and private entities that are willing to become providers.

Basic cloud model: researchers submit code, which is run in the cloud, and only results are returned to researcher. Researcher agrees not to attempt identification of study participants, etc.

WARNING!

Recall first Human Genome Project: Celera vs. government.

EXTRA SLIDES

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:

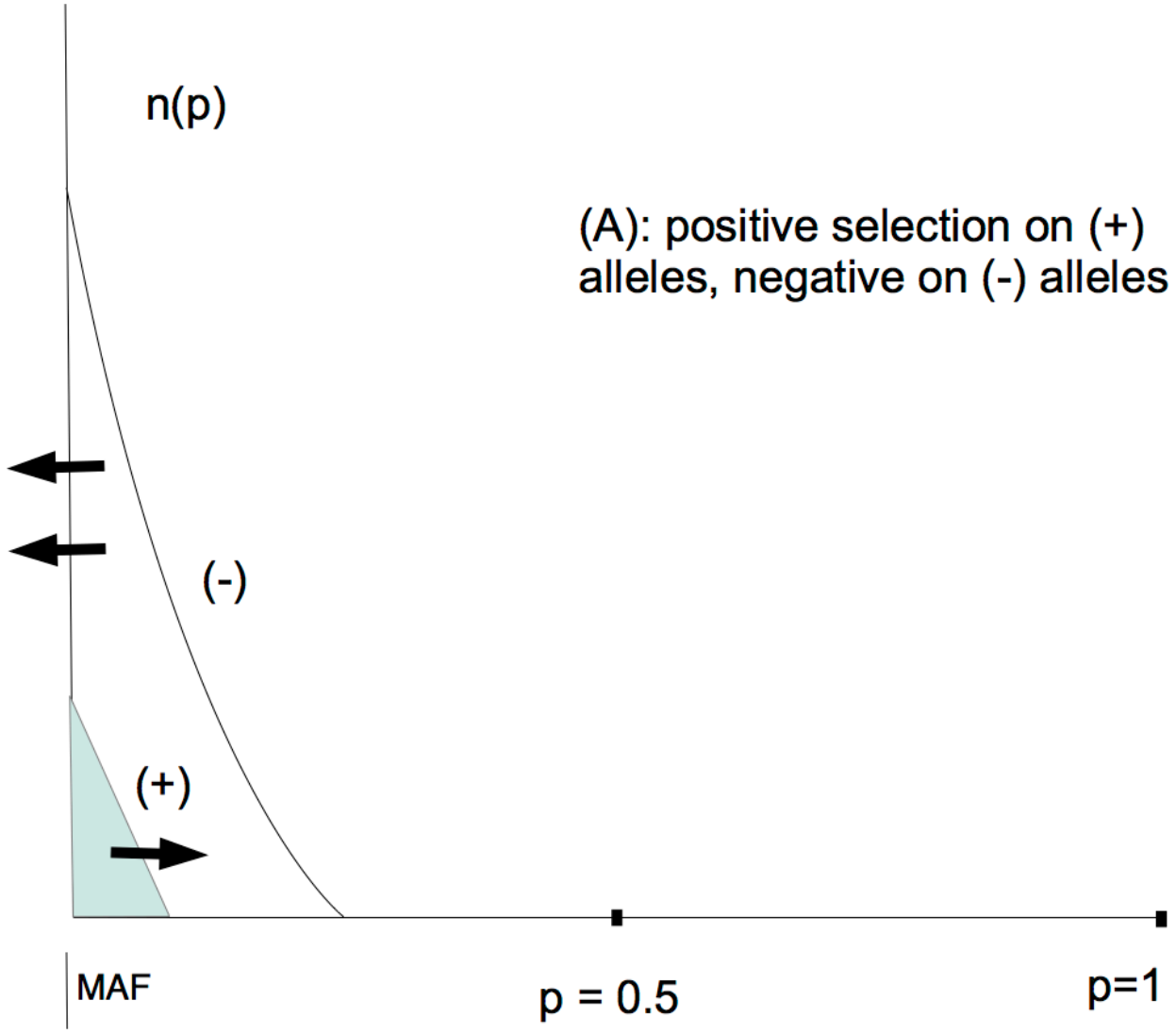
$$\frac{d\langle F\rangle}{dt} \approx \sigma_A^2$$

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

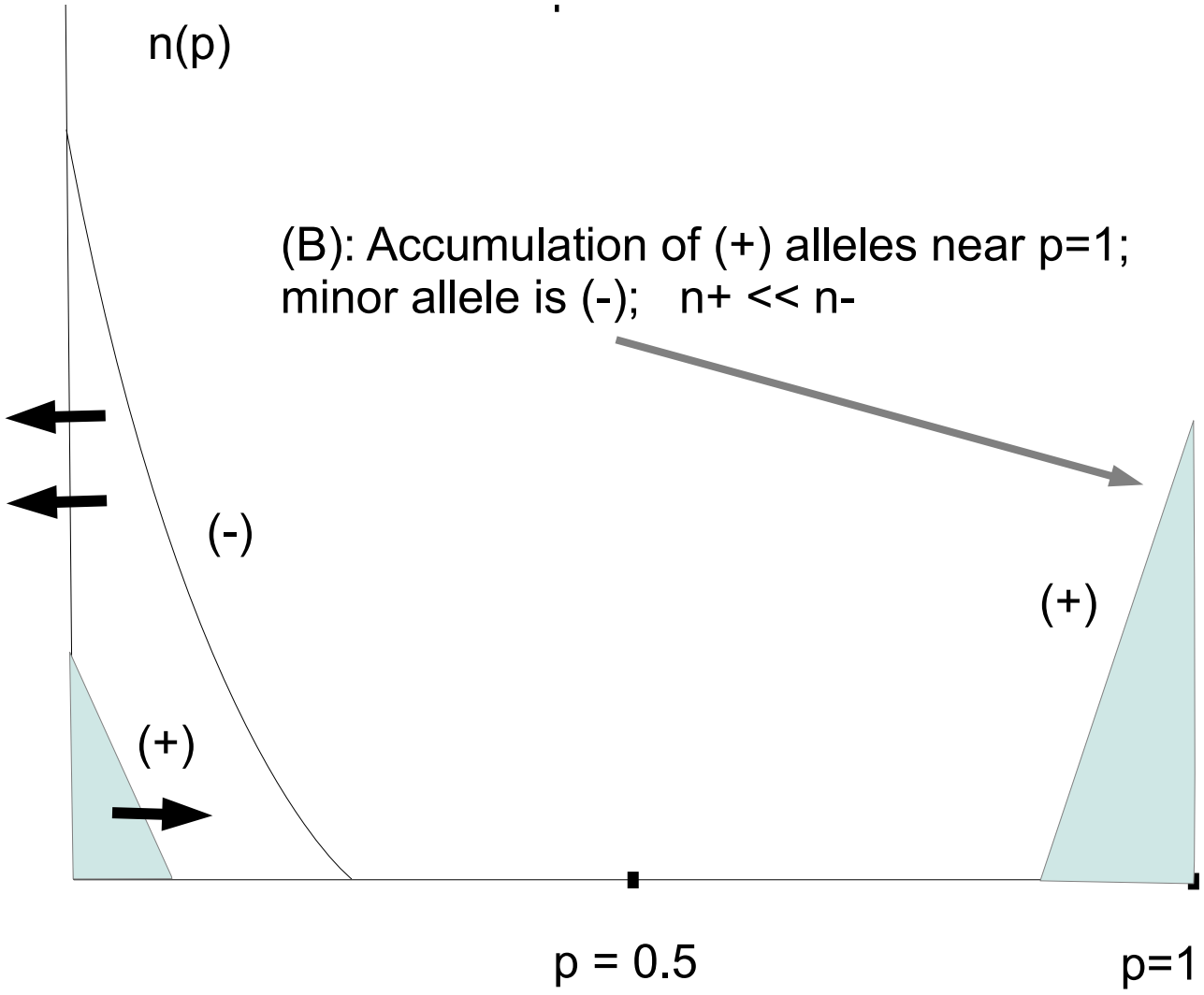
Animal and plant breeders have been using additive variance for millennia.

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!

Selection and MAF distribution



Selection and MAF distribution



Simplified additive model: spherical cow

(1) N causal variants, minor alleles have (−) effect on IQ

(2) Typical $MAF \sim 0.1$

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

For $N \sim 10$k, get 1 SD change in intelligence per ~ 40 extra (−) variants.

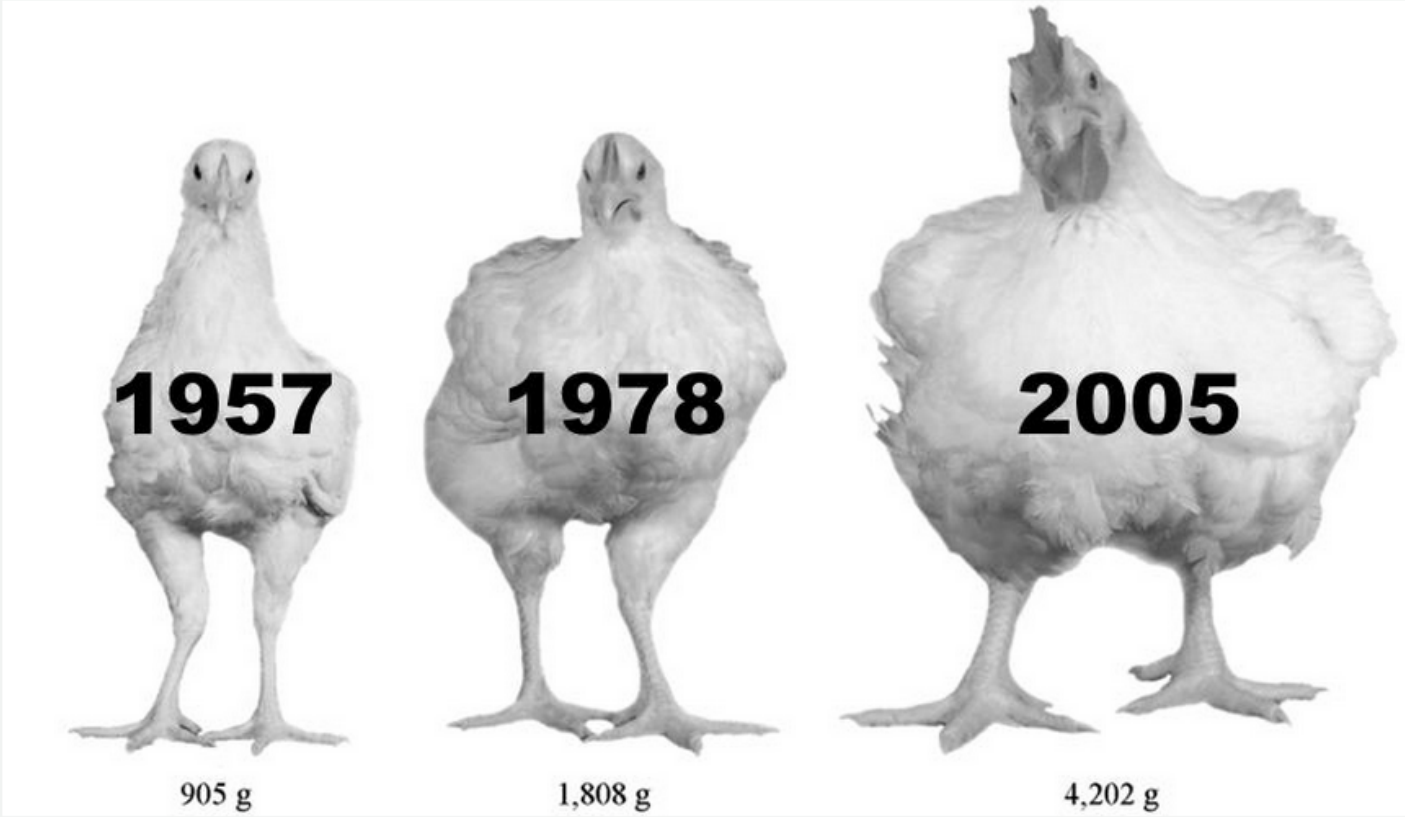
$$(+\,-\,+\,+\,-\,+\,+\,+\;\cdots\;+\,+\,-\,+\,+\,+\,+\;\cdots\;+\,+\,+\,+\,+\,+\,-\,+\,+)$$

Typical individual has 1000 ± 40 (-) alleles. A highly intelligent individual might only have 900 (-) alleles, etc.

No negative variants: +25 SD ?!?

Reservoir of variation is LARGE: $N \gg \sqrt{N}$.

Big Chickens



The left-hand chicken is a breed from 1957. The middle chicken is a breed from 1978. The right-hand one is a breed from 2005. They were all raised in the same manner for this paper and were photographed at the same age. Vox added the dates to this image. **(Zuidhof, MJ, et al. 2014 Poultry Science 93 :1–13)**