Reproducibility in an evolutionary context:

Its Ontology and Epistemology

Ken Weiss
Dept of Anthropology
Penn State

July, 2017

Three kinds of reproducibility in genetics

1. Causal (gene action)
2. Sampling (significance tests, &c.)
3. Phylogenetic (animal models)

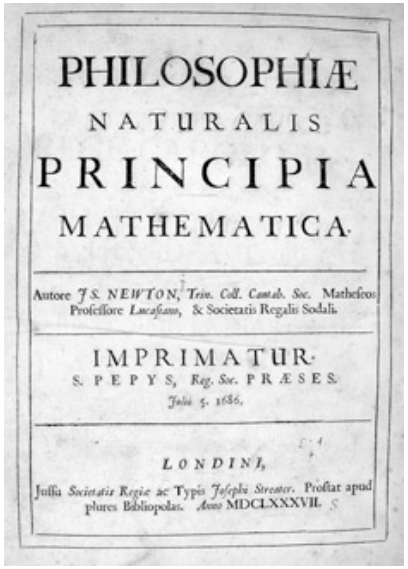
What should we expect of each?
How often is that actually tested?
Can it be tested?
If not, how confident can we be in our inferences?

What are scientific 'laws'? What is 'theory'?

Science since its foundation in the Enlightenment period 400 years ago has assumed Nature is law-like, the laws are universal, and our objective is to understand those laws, expressed in terms of 'theory'

As stated by Isaac Newton in his *Principia* (1687), if Nature is law-like, one can extend what one observes in a particular sample, to *all* of Nature

Life is a physical phenomenon. But is it law-like in this way?

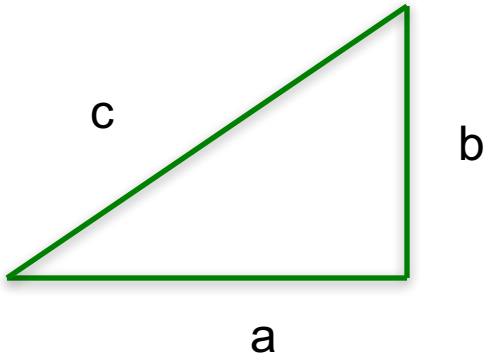

PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

Autore *IS. NEWTON*, Trin. Coll. Cantab. Soc. Matheseos Professore Lucasiano, & Societatis Regalis Sodali.

IMPRIMATUR.
S. PEPYS, Reg. Soc. PRÆSES.
Julii 5. 1686.

LONDINI,
Justu Societatis Regiæ ac Typis Josephi Streater. Prostat apud plures Bibliopolas. Anno MDCLXXXVII.

Euclid, Plato and others identified an essential principle: science seeks *abstract general* representations of Nature

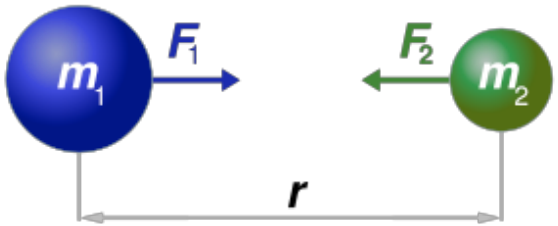


A*ll* right triangles contain 180^{o,}, and $a^2 + b^2 = c^2$ *always!*
But there is no such thing as a right triangle! There are only approximations to this 'true' relationship.
Note that this kind of theory provides external control and prediction.

We often think in this Platonic way. We refer to 'the' human genome, but it doesn't exist! And we do not have 'copies' of it! And it's not 'normal'! It is a totally arbitrary abstract *reference*.

**The universe actually has curvature so these principles don't hold exactly. But corrections for curvature exist

Newton, Galileo et al. introduced the modern concept of 'laws of Nature'



$$F_1 = F_2 = G\frac{m_1 \times m_2}{r^2}$$

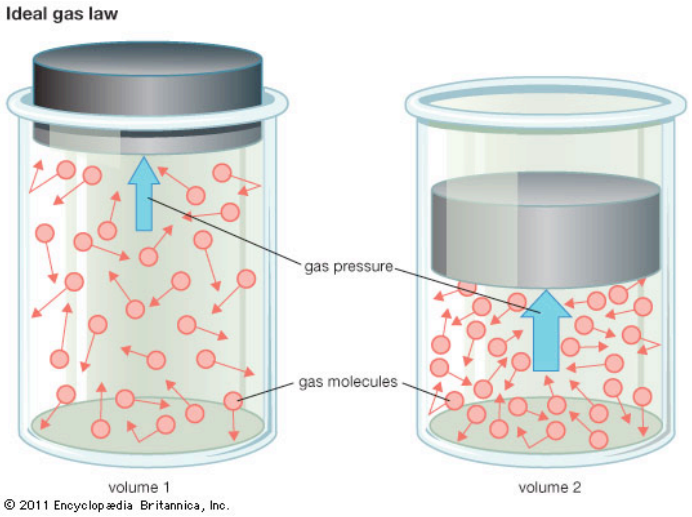
This led to *deterministic*, *universal* equations in mathematical form
This kind of theory provides external control and prediction

Charles Darwin had this kind of view of life, with natural selection the universal 'law'

There are some who think that fundamental physical parameters, like ***G*, the gravitational constant, may not actually be constant even within our universe. This may account for the apparent widespread 'dark matter'. At the very least, the values are close to constant *or* will obey some other universal 'law'.

Then came statistics: 'Laws of Nature' can be probabilistic.

But the interacting elements are identical and in large numbers



Pressure=Moles x Volume x Temperature

Each identical molecule acts like a deterministic Newtonian 'billiard ball', but there are so many that all we can measure is the net collective result, treated in probabilistic terms, and view such laws as *probabilistic determinism.* But this theory also provides external control and prediction.

Mendel's view was that transmission is probabilistic, but causation is deterministic. But what determines the probabilities, if they have specified values? Mendelian transmission has probability p=0.5….except when it doesn't!

Statistical criteria have become essential for biological inference:

Measurement is incomplete and imperfect
Inheritance is probabilistic (Mendelian transmission)
Reproductive success is probabilistic (fitness)
Mutation and recombination are probabilistic
Some functional effects may be truly probabilistic (antibody variation)

Epidemiological and evolutionary inference are often based on *sampling*
But most of these probabilisms are *not* simply based on sampling, yet we have no theory for their specifics.

If Nature is law-like, and we had a valid theory of life, then our data should approach these values *asymptotically.* That, for example, is the implication of the phrase 'precision genomic medicine'

But if life evolved as we think it did, asymptotic approaches are perhaps not to be expected! That is because each population, time, species, etc. is different.

'Sampling' in biology is permanent: e.g., who does not reproduce cannot be resampled

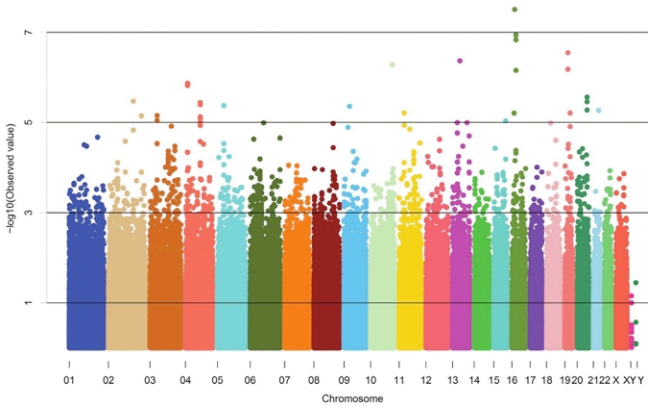
Internal comparative testing is not the same as — and is far less rigorous than inference from an externally derived *causal* theory

Causal?

Yes!

Hope so

Nah!



GWAS results plot

This is basically *retro-fitting* to existing data, and in the absence of adequate theory that has no formal way to ensure that its results are *predictive* other than assuming things won't change.

Life is an *ad hoc* evolutionary history of relationships among unique, non-homogeneous *polymeric* elements (modular sequences, genes, regulatory regions, interaction networks). Each gene, trait, and population is *inherently* different. That difference is the essence of evolution.

Could life's causal architecture be *not* precisely predictable—***in principle****?*

Epistemology: How do we know Nature?

1. *Induction: repeated observations replicate and hence support a hypothesis*
2. *Verifiability: you should be able to confirm your hypothesis*
3. *Falsifiability: a mistaken idea can be falsified*
4. *Deduction from theory*: first principles may enable valid prediction
5. *Prediction: if it's true you should be able to predict future instances*
6. *Parsimony: the simplest explanation is best (what justifies this criterion?)*
7. *Probabilistic causation: 'significance', power, importance*
8. *Probabilistic hypotheses: what guesses about Nature seem most plausible?*
9. *Who cares??: Use whatever you want!*
10. *Gang warfare: Use whatever will get your desired result published!*

Note that all of these criteria assume some form of replicability

A famous example of the problem with inference by induction:

River Ouse, Cambridgeshire, 2005 (photo by A Buchanan)





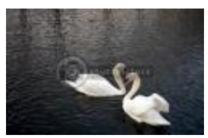




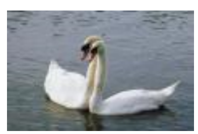













River Ouse, Cambridgeshire, 2005 (photo by A Buchanan)

Common responses to the black swan 'exception':

Define effects in terms of observations. A 'swan' is a white water bird. A black water bird is some other sort of bird.

In genetics, disease can be *defined*, or studied, in terms of a specific cause, such as variation in a specified gene.

Other outcomes of similar symptoms can be given a different diagnostic name, or called idiopathic phenocopies, but viewed as if they shared the same proximate cause.

Genotyping can be useful for treatment, but can also be used even to raise doubt about whether the person 'really' has the disease.

If risks associated with the same genotype vary over time or among samples or populations, how do we explain that?

Time for a tea break! After all, tea was the context of statistical significance!

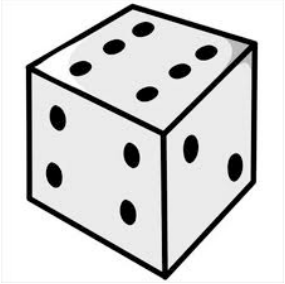
Significance tests are *arbitrary* and *subjective* and so are confidence intervals.
But if for evolutionary reasons, populations and samples are *not* replicable, what use should we make of them?

Milk or tea first?

In the 1920s, British biologist Dr Muriel Bristol tastes 8 cups, half milk first, half tea first. She gets them all right. The chance of that was about 1 in 70 (p~0.014). Could she really tell, or was she just lucky?

Probabilistic cause, and statistical thinking

Dice and coins: are they fair? Are they identical? Are flips and rolls identical?
Are they really probabilistic?
What do these questions *mean*? How can you *tell*?



Probabilistic Causation: Rolling Dice

WHAT CAN WE SAY ABOUT A DIE (singular of 'dice')?
SUPPOSE, FOR EXAMPLE, THAT YOU ROLL A DIE AND GET

4

* THIS DIE WHEN ROLLED COMES UP *4*
* THIS IS A FAIR DIE
* THIS DIE COMES UP 4 SOME SPECIFIC FRACTION OF THE TIME. WHY?
* THIS DIE HAS SOME TRUE PROBABILITY OF COMING UP 4 (WHAT ABOUT 1, 2, 3, 5, 6?)
* WE CAN'T SAY ANYTHING ABOUT THE DIE
* WE CAN ONLY SAY SOMETHING ABOUT THIS DIE BY LOOKING AT OTHER DICE

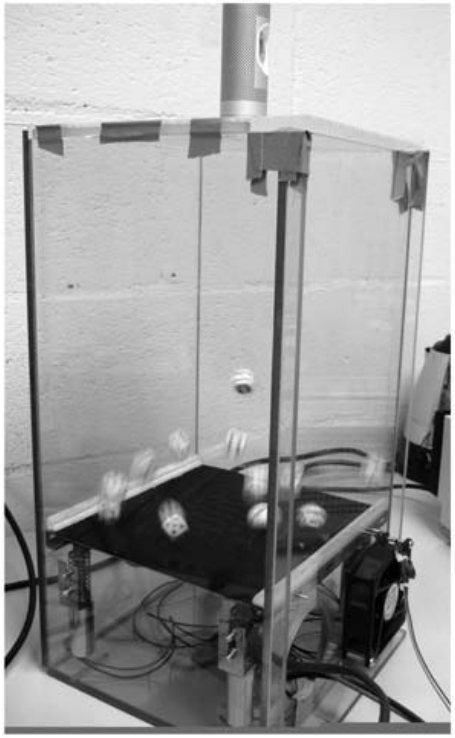
How would you know?

If your answer involves the *assumption* of replicability, what justifies that?

(When and why can that assumption be made about the outcome of a roll of the genetic dice?)

Are dice fair……or not??

This 2009 paper reflects an automated dice-throwing machine, and its results from 26,306 tosses. These results are statistically unclear (as to whether they represent deviation from 'fairness'), and they do *not* replicate the heavy-more/light-less face difference! What are we to make of this? How can we know?



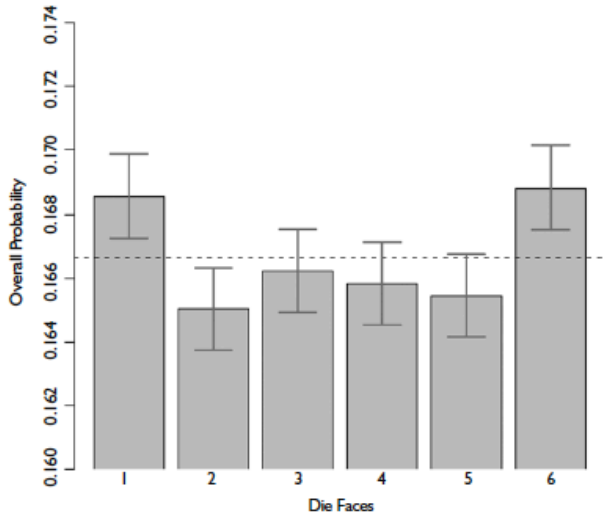
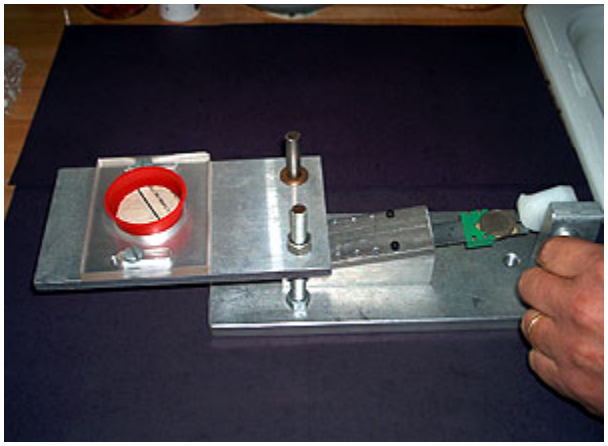


Figure 5. Probability of observing each number of pips out of 12 times 26,306 total rolls. The error bars are 95% confidence intervals according to binomial sampling, where $\sigma^2 = p(1-p)/315672$ and the dashed line shows the fair probability of 1/6 for each face.

"Oh!," you say, "these results aren't very different from…. 1/6?"
How many kinds of 'unfairness' would you have to test and how?
What do you mean not 'very' different? What are the consequences (if any) of that difference?
What if the results were *exactly* 1/6 for each outcome? Would that prove fairness?

Labby, Z . *Chance*, 2009 22:11-12

"God does not play dice with the universe" (so, approximately, said Albert Einstein)



A perfect coin-flipping machine would produce the same result every time. Persi Diaconis made one in 2004.

So why does flipping *appear* to be 'random'?
And then what does it mean to be 'fair' or 'unfair'?
Since it rotates during its flight, is it only hitting the table that seems 'random'? Does this mean the *table* is unfair??

Significance testing is subjective. Can you tell *objectively* if a coin is 'fair'? What would that require?

In terms of genetics, how can you tell if a SNP does something or not—for example, if you have it, you get a disease, or you get the disease with a different probability than if you did not carry that SNP? Or at a different age? "The" probability varies among populations, and presumably among individuals….but how could you tell? What would it mean for the prediction to be 'precise'?

Can an allele be selectively neutral? How could you possibly tell?

It is not entirely obvious what we mean by ‘replicate’

1. Exactly the same result every time we look?
2. The same result some percent of the time (e.g., 1 - *p*-value fraction?)
3. Almost the same result every time (then what does ‘precision’ mean here)?
4. Results of replications approach some fixed value asymptotically?
5. The gang warfare criterion: what the dominant group decides is ‘true’?

Genetic 'causation' is context-dependent

DNA sequence contains 'information' for the bearer's traits

DNA *variation* is responsible for corresponding variation in those traits.

Genetic function requires *interaction*, and is *context-dependent*.

Each nucleotide's context, or '*environment*', includes the rest of the genome in that cell, but also many factors beyond that.

The central property of evolution is *variation*: *non*-replicability

Functional and regulatory detail can arise by point mutation, that come and go quickly.

New genes and hence new function largely evolve by duplication.

Evolution generates *redundancy* in genetic causation.

Because of causal redundancy, complex traits exhibit *phenogenetic drift* and *phenogenetic equivalence* among samples and populations. Only the trait 'replicable'. Many loci contribute, most effects too small or too rare to detect.

One contributing gene

Small	Medium	Large
aa	Aa	AA
	aA	

Two identical contributing genes

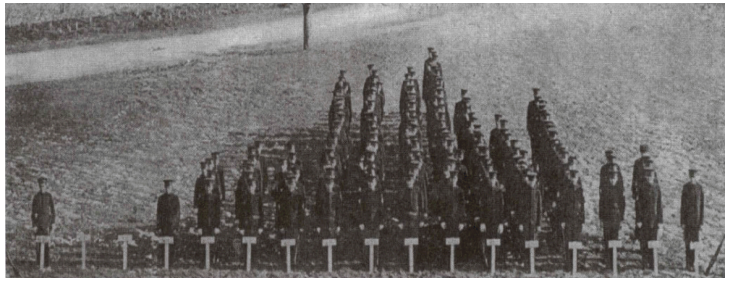
Small	Medium			Large
		AAbb		
		AaBb		
	Aabb	AabB	AABb	
	aAbb	aABb	AAbB	
	aaBb	aAbB	AaBB	
aabb	aabB	aaBB	aABB	AABB

Three identical contributing genes

A key fact is that individuals with the same trait have *different* genotypes

Small			Medium				Large
			AABbcc				
			AAbBcc				
			AAbbCc				
			AAbbcC				
			AaBBcc				
		AAbbcc	AaBbCc	aaBBCC			
		AaBbcc	AaBbcC	aAbBCC			
		AabBcc	AabBCc	aABbCC			
		AabbCc	AabBcC	aABBcC			
		AabbcC	AabbCC	aABBCc			
		aABbcc	aABBcc	AabBCC			
		aAbBcc	aABbCc	AaBbCC			
		aAbbCc	aABbcC	AaBBcC			
		aAbbcC	aABbCC	AaBBCc			
	Aabbcc	aaBBcc	aAbBcC	AAbbCC	AABBCc		
	aAbbcc	aaBbCc	aAbbCC	AAbBcC	AABBcC		
	aaBbcc	aaBbcC	aaBBCc	AAbBCc	AABbCC		
	aabBcc	aabBCc	aaBBcC	AABbcC	AAbBCC		
	aabbCc	aabBcC	aaBbCC	AABbCc	AaBBCC		
aabbcc	aabbcC	aabbCC	aabBCC	AABBcc	aABBCC	AABBCC	

Many contributing genes



Small — Medium — Large

(This presentation assumes equal deterministic effects, 2 alleles per gene, equal frequency, & no error or environment; but the overall conclusion is not affected by these simplifications)

Spooky action at a (very short) distance: complex *trans* causation

Complex *trans* 'communication' is pervasive within (and among) cells

It involves simultaneity and serially dynamic elements

How does it work?

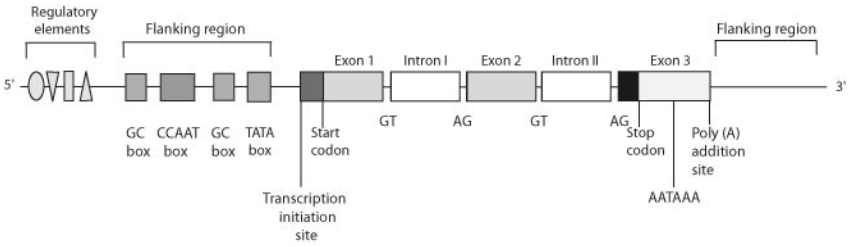
Are we missing something fundamental?

Can statistical thinking disentangle this or does that confuse epistemology with ontology?

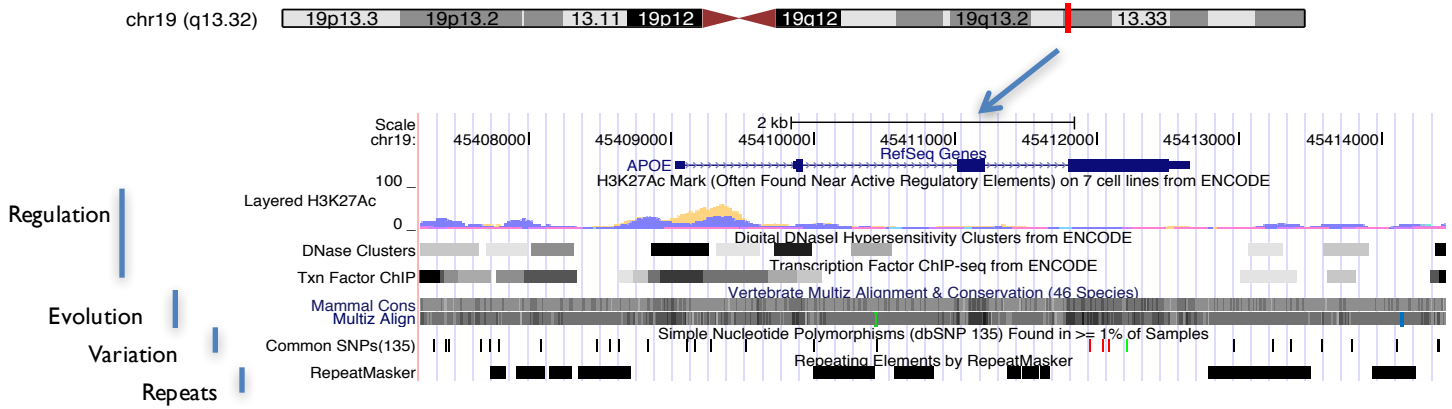
Life is about *context:* 'Boolean' relationships and arrangement:

Genomes are heterogeneous and modular—that's part of our *theory* of life.
That means 'replicability' is unpredictable in detail

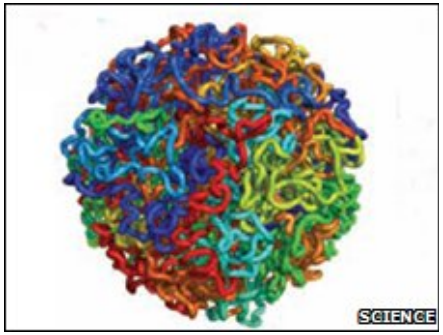
Cis: 'typical' gene (eukaryote)

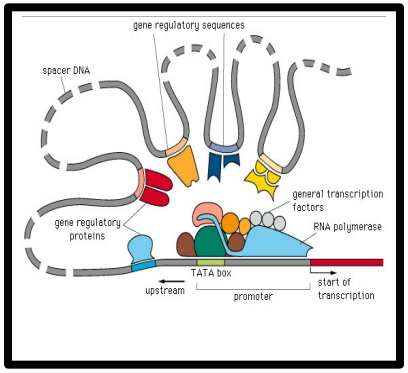


Cis: Genome contents around a 'typical' gene

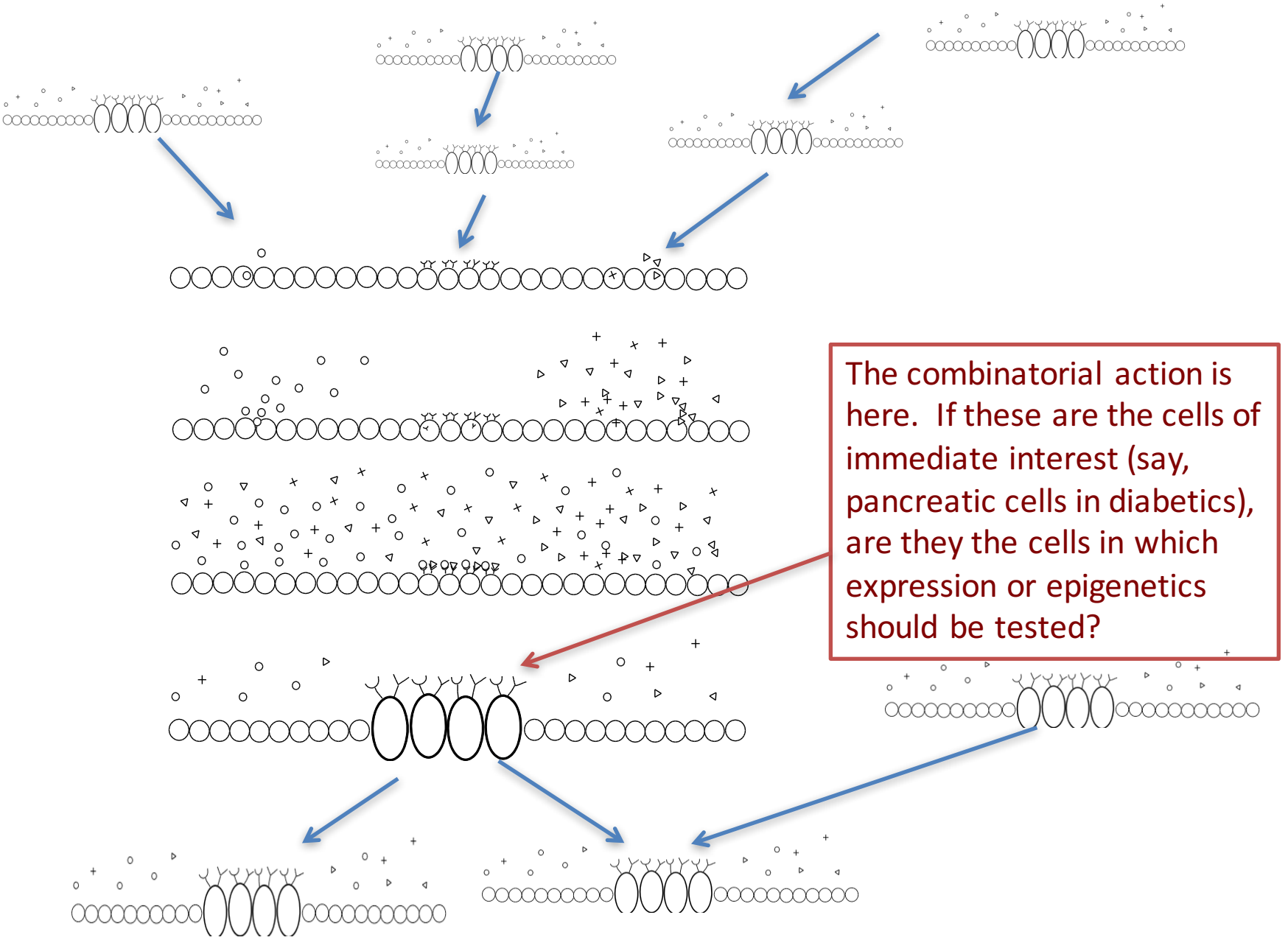


Trans: Non-linear 'typical' chromosome arrangement, and Boolean gene regulation





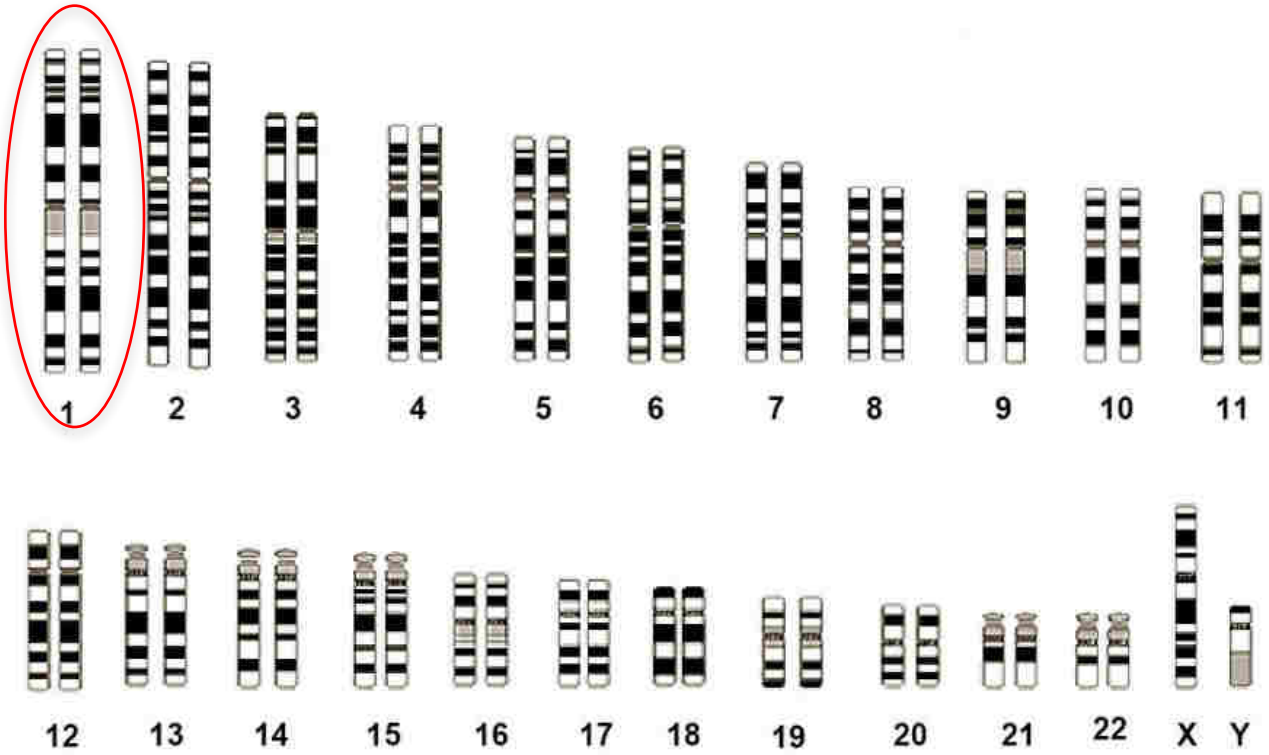
Signaling is a multidimensional Boolean *trans* phenomenon



The combinatorial action is here. If these are the cells of immediate interest (say, pancreatic cells in diabetics), are they the cells in which expression or epigenetics should be tested?

Context: Not all causation in a cell works in *cis*. Fundamental *trans* mysteries!

It is not even known how homologous 'sister' chromatids find each other to align before cell division. Whatever the mechanism, it must over-ride the similar negative charges of the homologs, and be robust to synteny rearrangements



Monoallelic expression: Olfactory Receptor clusters in the human genome

~1000 genes in tandem arrays with some singletons, many (60%) often-polymorphic pseudogenes, and located disproportionately near telomeres in humans and probably in the original vertebrate cluster (color coding indicates gene subfamilies)

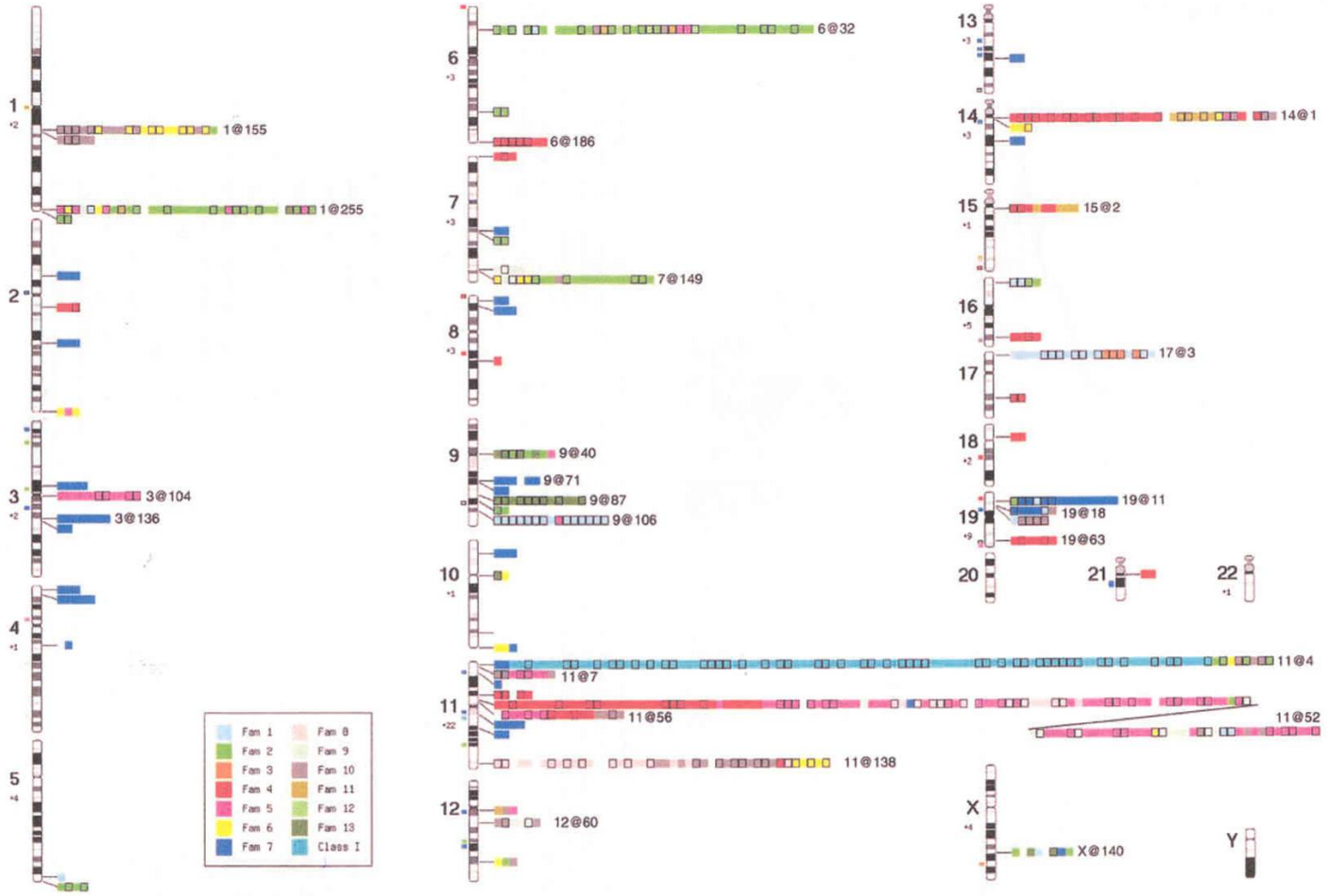
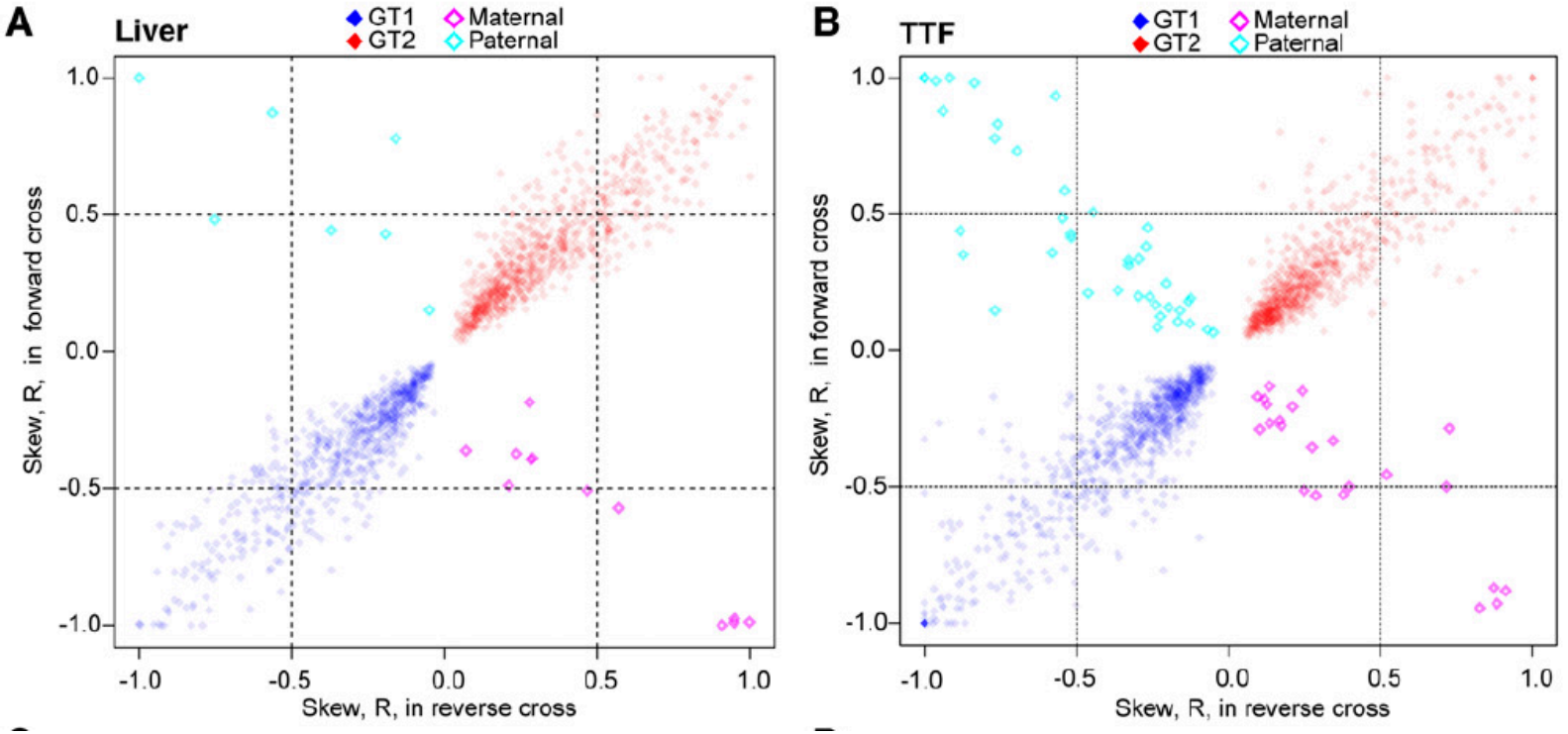


Figure 1 The human OR subgenome at a glance. ORs are depicted as squares, colored by family (see key). All Class I families are colored equally. Unclassified ORs are indicated in light gray. Framed squares denote intact genes. ORs to the left of each chromosome indicate singletons, and those to the right are in clusters of two or more. Gene order within each cluster is only approximate. The largest cluster on chromosome 11 is shown split for convenience. Small plus numbers under each chromosome name indicate the number of additional ORs for which a coordinate was not determined. Megabase coordinates (distance from the p telomere, excluding heterochromatic regions) were translated linearly into chromosomal localization. Therefore, the correspondence between cytogenetic bands and the indicated OR locations may be somewhat shifted.

Monoallelic expression seems common. Non-Mendelian inheritance of traits? Replicable widespread results in a GT1 – GT2 F1 mouse cross:



Dots are the expression-level skew (boxes=imprinted sites. Liver and Tail Tip Fibroblasts (TTF) were tested. Gene-specific results were replicable in crosses in both directions.

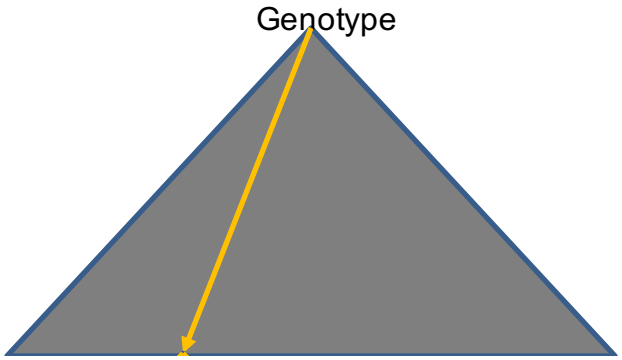
Is Mendelian inheritance of traits the diploid 'law'. . . . except when it isn't?

Pinter et al. Genetics, June 2015

Profound limits to replicability, and hence to predictability: A *fundamental* epistemological trap for personalized genomic medicine and evolution

Then

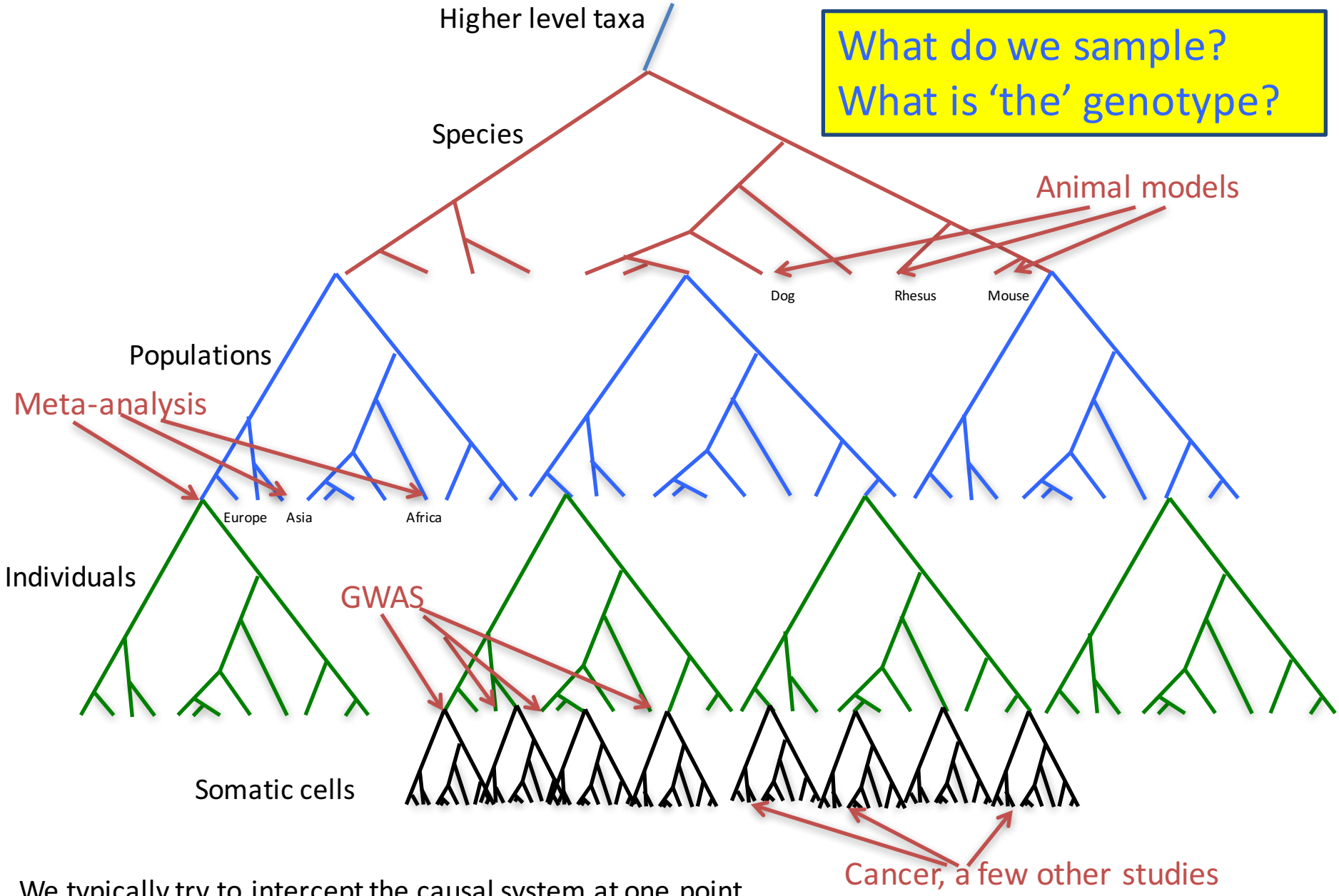
Now



Environmental exposures known or knowable

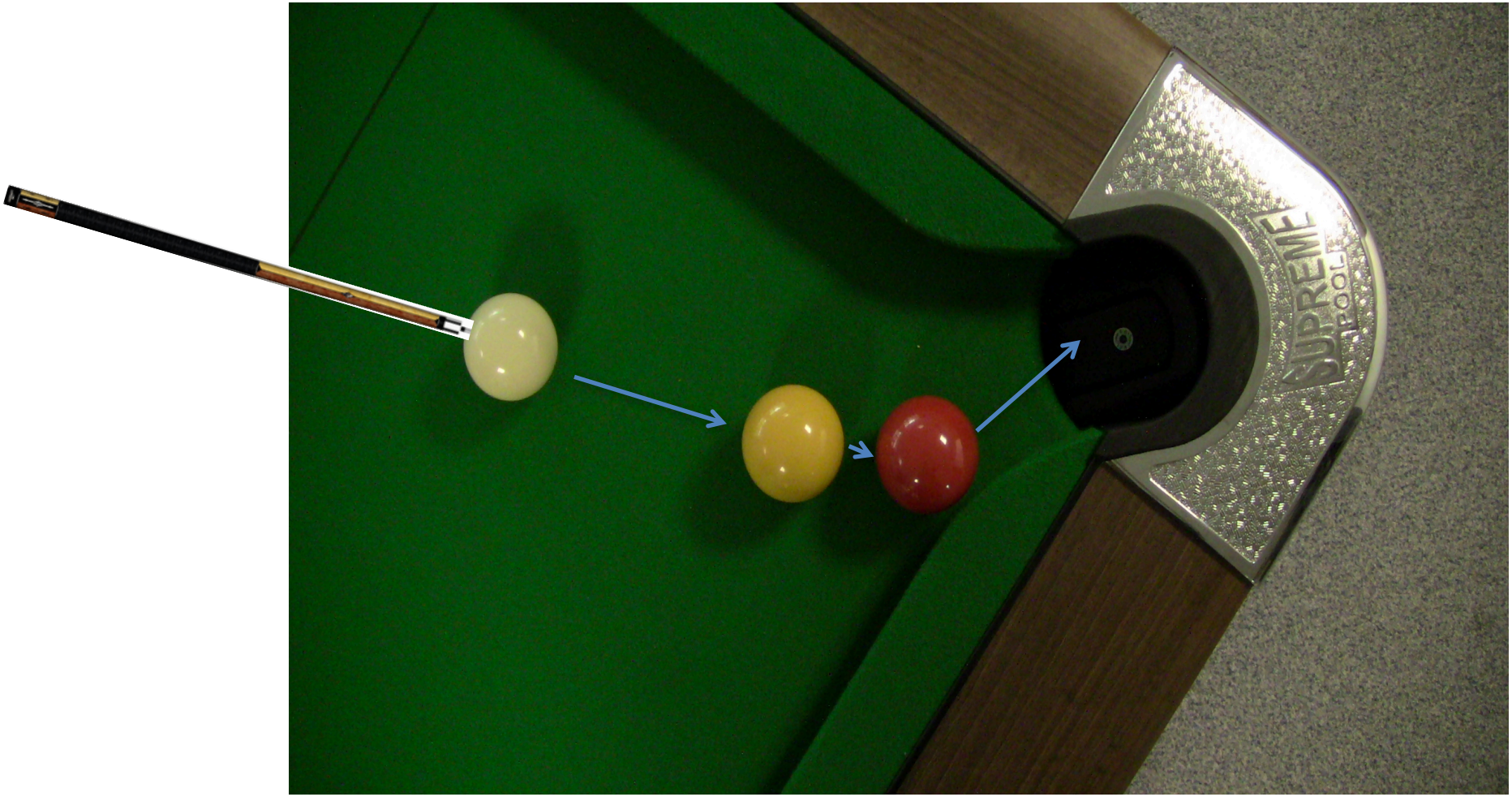
Possible outcomes

Estimates of risk are *retrospective*. But the risk we want to know is *prospective.* Culture-based definitions of traits are also part of the causal landscape.



We typically try to intercept the causal system at one point, rather than deal with its nested generating process as a whole. Can we find better strategies?

Red ball in pocket! What is the ‘cause’? Who decides what counts?
Do we always know what we mean by ‘cause’? Where does causation ‘start’?



Source: Wikipedia ‘Billiard table’

Some core reproducibility problems

Evolution is about variation, *non*-replicability

Evolution generates redundancy, but only probabilistically

Most of our inferential tools are based on replicability with proper distributions

Each trait has its own history and characteristics

Analysis of one local subset does not extrapolate automatically to others

Finally . . . .
There are obvious bioethical problems with promises like 'precision' genomic medicine—and with *making* such promises. We might even mislead ourselves, because precision assumes an asymptote towards a true value.

But because we got here by evolution, such values may not exist

Formulating a well-posed question or hypothesis:

1. Has an answer
2. The answer is unique to the question
3. The answer changes in an orderly way with initial conditions

If we can't form well-posed questions, we can't expect clear answers.

The usefulness of the answers, or their 'precision', may depend on unmeasured or unknown factors, or unstated or untestable assumptions, or misapplied mathematical or statistical methods

THE END.