Data analysis in genetic epidemiology: an overview

(slides available at http://mga.bionet.nsc.ru/~yurii/, go to 'Courses' → 'SnpCourse_2010')

Yurii Aulchenko

Erasmus MC Rotterdam

What does statistical genomics give us?

- Loci and alleles, associated with the trait
- Knowledge of a **locus** and alleles allows individual risk prediction
- Knowledge of a **gene** provides information on biological networks involved in trait development
- This knowledge may allow development of new biomarkers, prevention and treatments strategies

Loci identified for complex traits

		# Loci		
	<2005	2008	2010	2012
Lipids	few	~30	95	±200
Height	0	~50	100+	±300

		%Var		
	<2005	2008	2010	2012
Lipids	~2%	5%	**10%**	±15%
Height	0	4%	**8%**	±10%

- %Variance attributable to genes:
  - height ~ 90%, lipids ~30%

Data analysis steps

Stage	Software
Genotype calling	BRLMM, GenomeStudio, Chiamo
Imputations	MACH, IMPUTE, BimBam, Beagle
Association analysis	PLINK, SNPTEST, *ABEL, SNPTEST, QTL2MACH, SNPAssoc, SNPMatrix, …
Meta-analysis	METAL, *ABEL, Mantel, MetaMapper

What we can do easy (and what we can not do easily)

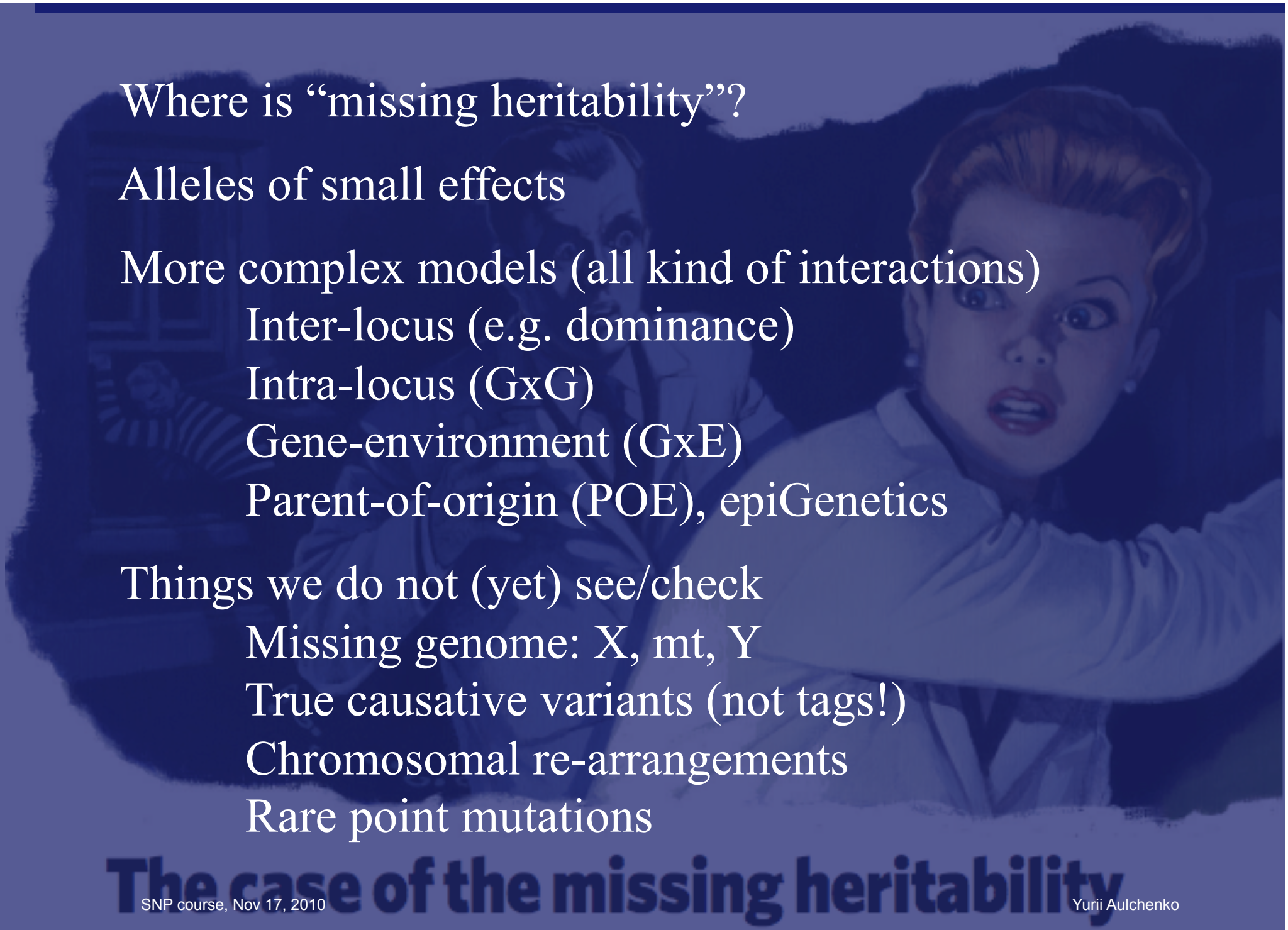
- Trait:
  - Quantitative, normally distributed [+]
  - Binary [+]
  - Categorical [-] (multinomial regression – feasibility)
  - Markedly non-normal QT [-] (methodological problem)
- Design:
  - Cross-sectional [+]
  - Follow-up [-] (GEE, LMM – feasibility)
- Ascertainment:
  - Random [+]
  - Case-control [+]

What we can do easy (and what we can not do easily)

- Genetic structure:
  - Unrelated [+]
  - Related [+] (but feasibility problem with large sample sizes, not standard designs and models)
- Analysis model
  - Standard single-marker [+] (does not work for rare variation)
  - Multiple-marker models, especially for rare variation [+/-] (methodological and feasibility)
  - Interaction models [+/-] (methodological and feasibility)

- In essence, we can test well single-marker models for binary and quantitative traits in cross-sectional design

The case of the missing heritability


Where is “missing heritability”?
Alleles of small effects
More complex models (all kind of interactions)
Inter-locus (e.g. dominance)
Intra-locus (GxG)
Gene-environment (GxE)
Parent-of-origin (POE), epiGenetics
Things we do not (yet) see/check
Missing genome: X, mt, Y
True causative variants (not tags!)
Chromosomal re-arrangements
Rare point mutations
The case of the missing heritability

Progressively bigger studies?

Sample size	R2 with 50% power
30,000	0.1%
50,000	0.06%
100,000	0.03%
200,000	0.015%
400,000	0.0075%
800,000	0.00375%

P. Holman's 'detectability limit'

- 'Detectability limit' hypothesis: in a study showing residual inflation of test statistics, there is a limit on detectable effect size, whatever the sample size is
- In study with residual inflation, significance threshold grows proportional to N c1 ('noise' constant), while power grows as N c2 ('power' constant). If c1 > c2 'noise' grows faster then power
- Read it other way: we hoped that brute force approach will always work (by big N's we can compensate for imperfect methodology). Apparently this is not true.

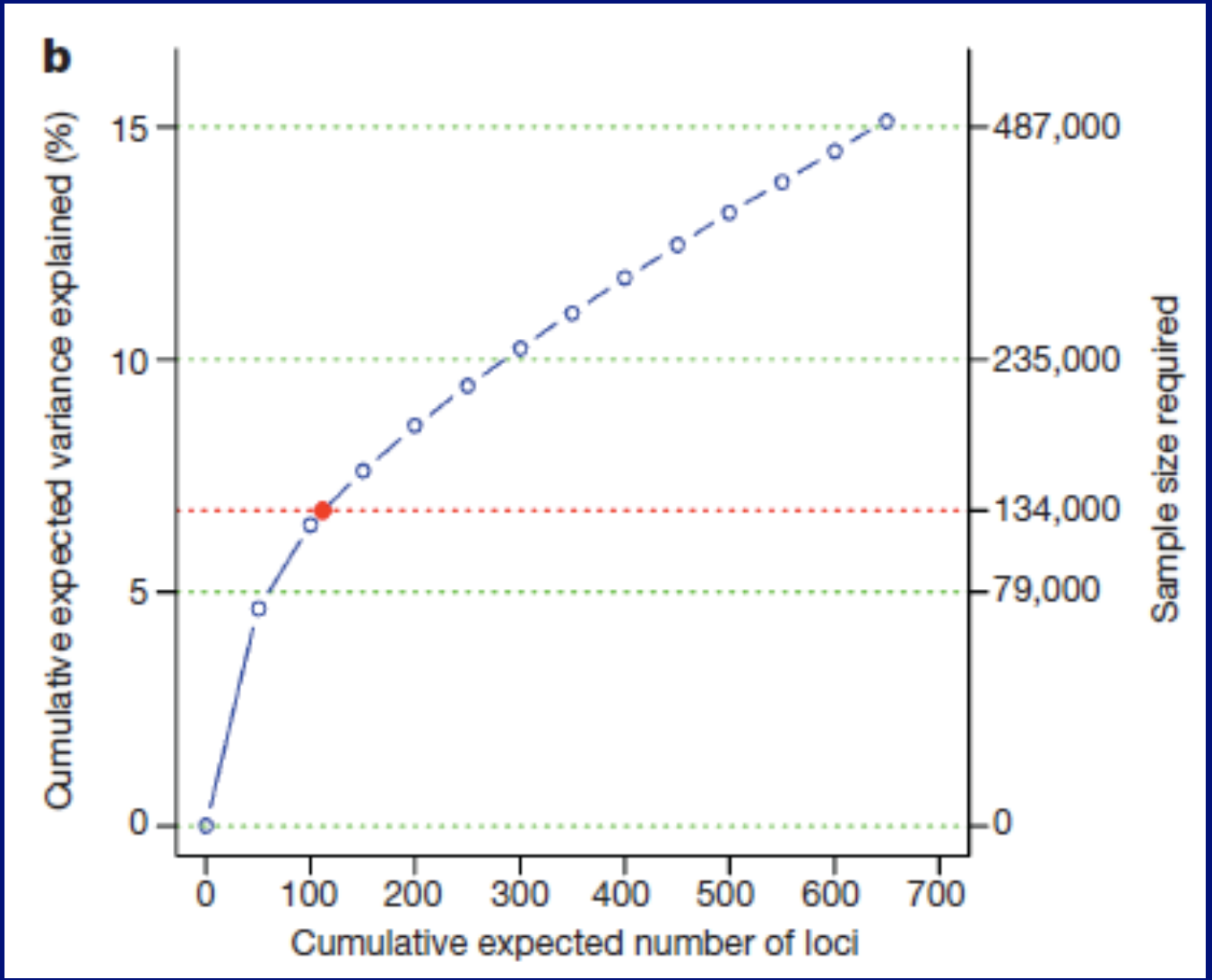
Progressively bigger studies should use progressively better methods!

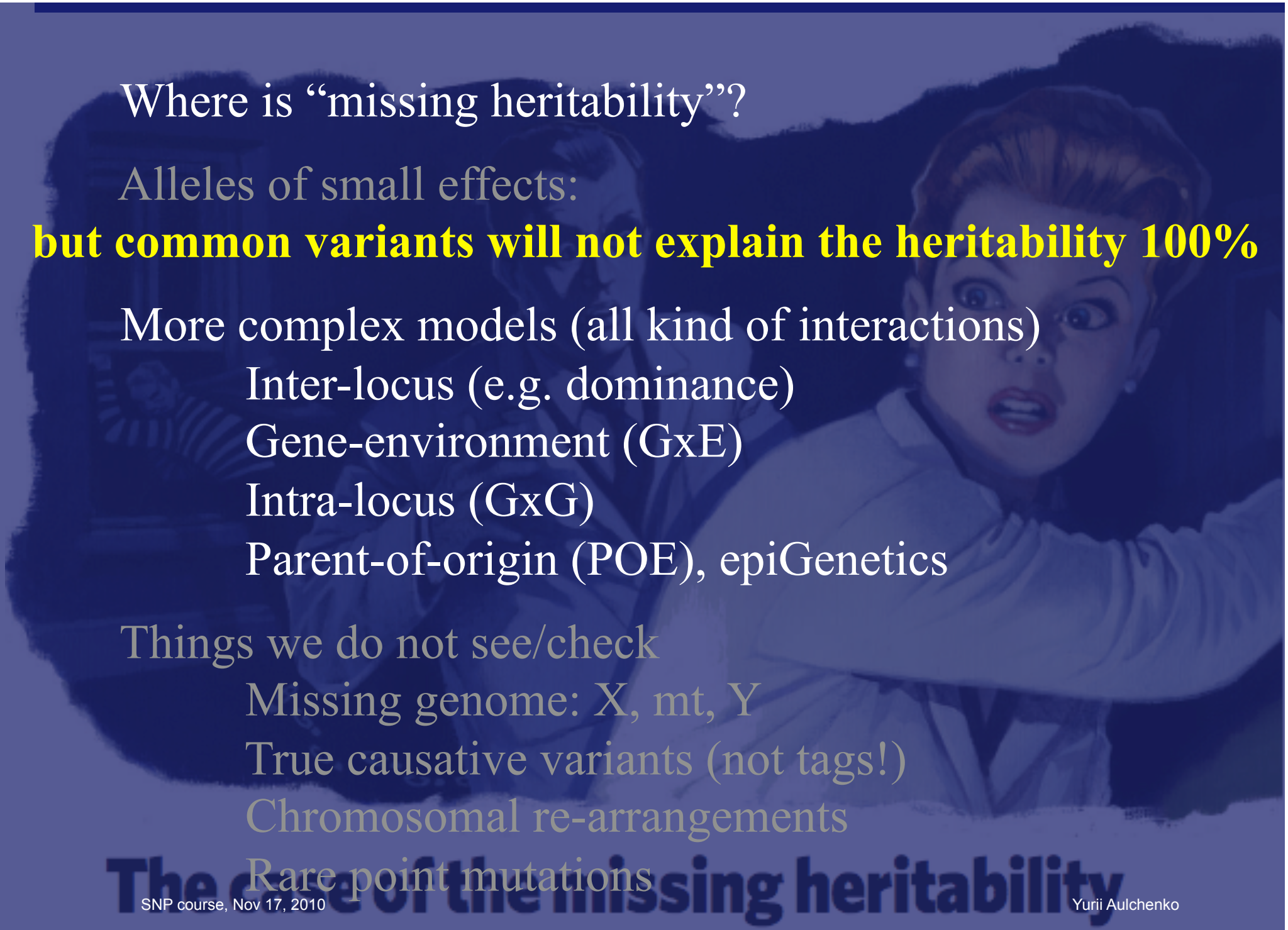
Sample size	R2 with 50% power	Limiting λ_{1000}
50,000	0.06%	1.0202
100,000	0.03%	1.0101
200,000	0.015%	1.0050
400,000	0.0075%	1.0025
800,000	0.00365%	1.0012

Problems with alleles of small effects

- Need very good methods to account for confounding (note: also for causality!)
- Mixed models?
  - Similarity matrix is estimated as correlation matrix between genomes; in many ways too simplistic approach **(open)**
  - Computational aspect: computational time ~ N^2
- Better meta-analysis methods: J. Lebrec presented several methods during EMGM-2009, yet not available as software **(open)**

Proportion of variance explained for height





Where is “missing heritability”?

Alleles of small effects:

but common variants will not explain the heritability 100%

More complex models (all kind of interactions)

- Inter-locus (e.g. dominance)
- Gene-environment (GxE)
- Intra-locus (GxG)
- Parent-of-origin (POE), epiGenetics

Things we do not see/check

- Missing genome: X, mt, Y
- True causative variants (not tags!)
- Chromosomal re-arrangements
- Rare point mutations

Problems with GxE when E is known: λ's going all the way around 1

- Rotterdam study: **population-based** cohort used for genetic research for over 15 years
- In GWAS performed over many traits, always λ < 1.05
- G x E results for some traits:

	Environmental factor			
	cov 1	cov 2	cov 3	cov 4
trait 1	1.13	1.13	1	1.14
trait 2	0.98	1.04	1.02	1.04
trait 3	1.12	**1.22**	1	1.09
trait 4	1.05	1.01	1.03	0.97
trait 5	1.1	1.09	1.07	1.01
trait 6	1.02	1.01	0.92	1.03
trait 7	0.94	0.95	**0.89**	1

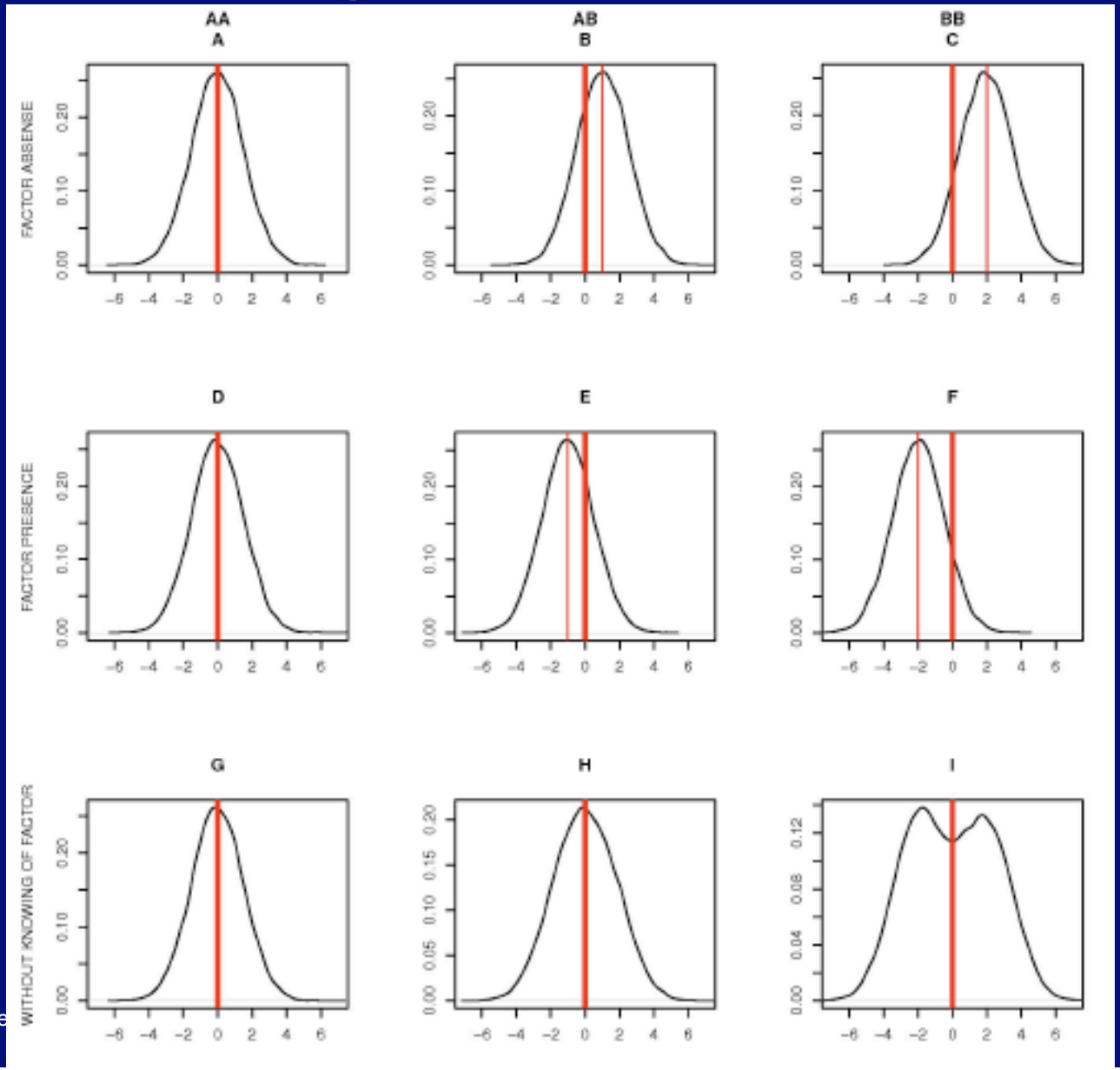
Solution: use robust (co)variances?

- Suggested by T. Lumley
- Implemented in ProbABEL v. ≥ 0.1-1

	Environmental factor			
	cov 1	cov 2	cov 3	cov 4
trait 1	1.03	1.04	1.03	1.02
trait 2	1.03	1.01	1.03	1.02
trait 3	1.02	1.04	1.03	1.02
trait 4	1.04	1.03	1.03	1.01
trait 5	1	1.02	1.03	1.01
trait 6	1.03	1.01	1.02	1.01
trait 7	1.02	1.03	1.03	1.01

- Still to high for small effects! **(open)**

Gx?: detecting interacting loci without knowing what it interacts with



Gx? method indeed works!

Trait	Interacting SNP	MAF	Chr	Position (Kb)	Nearest Gene	Type	Covariable	Variance of A1A1* (N)	Variance of A1A2* (N)	Variance of A2A2* (N)	Levene's P-value	Interaction P-value
CRP												
	rs12753193	0.38	1	65942.3	*LEPR*	-	BMI	1.27 (8491)	1.47 (10126)	1.68 (3167)	1.6E-29	7.2E-10
sICAM-1												
	rs17999[illegible]9	0.11	19	10255.8	*ICAM1*	Missense	Smoking	6621 (17063)	5316 (4421)	4104 (300)	2.1E-09	4.8E-09
	rs738409	0.22	22	42656.1	*PNPLA3*	Missense	BMI	6087 (13098)	6743 (6965)	9205 (1110)	1.9E-10	1.6E-07

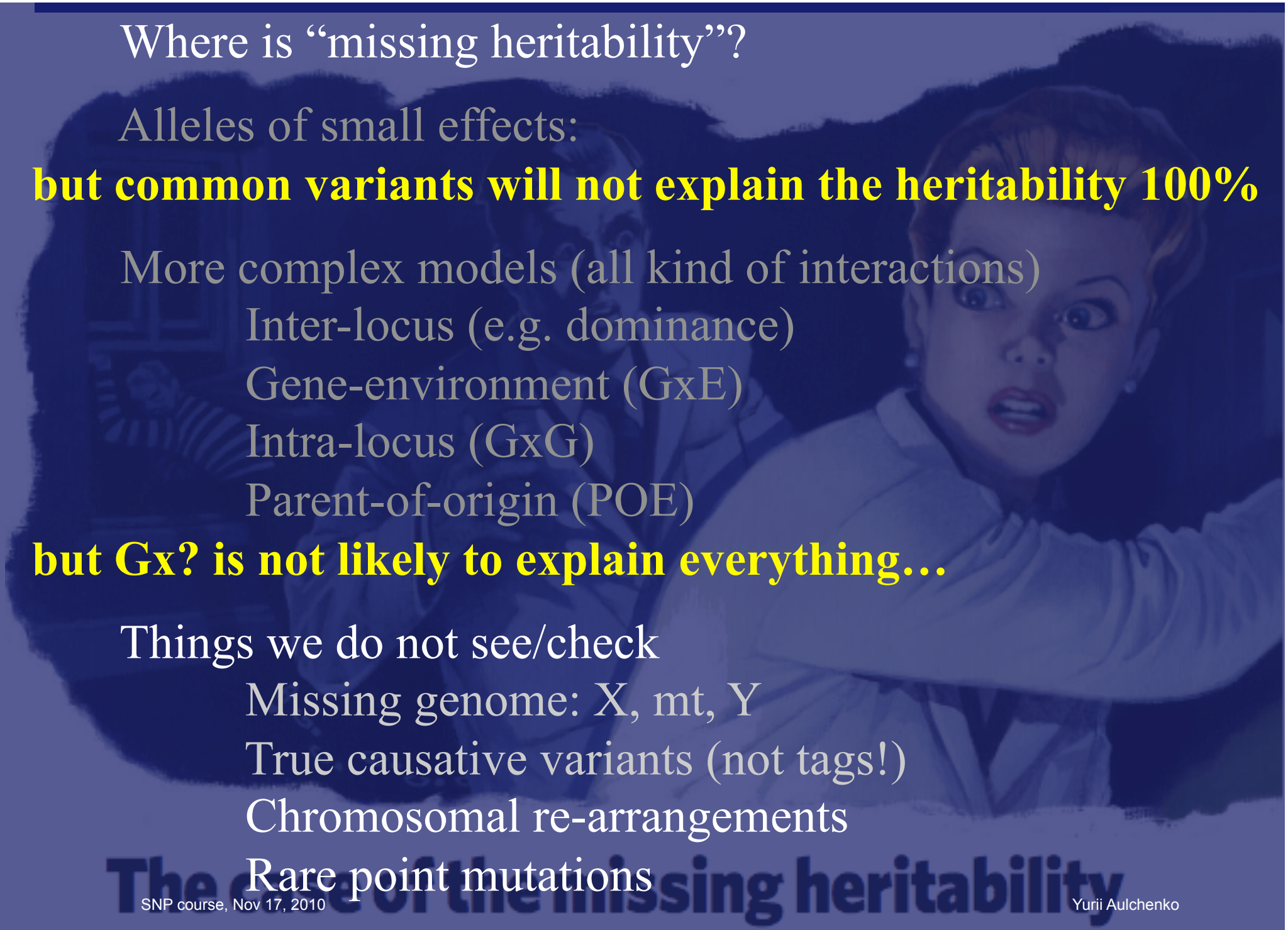
*A1A1: Homozygous Major Allele; A1A2: Heterozygous; A2A2: Homozygous Minor Allele.
doi:10.1371/journal.pgen.1000[illegible]81.t001

Pare *et al.*, PLoS Genet, 2010

Replicated by Struchalin *et al.*, BMC Genet, in press

An R package for GWV analysis:
VariABEL (under development)

Globally, few hits => not a lot of interactions, at least for common variants!



Where is “missing heritability”?

Alleles of small effects:

but common variants will not explain the heritability 100%

More complex models (all kind of interactions)

- Inter-locus (e.g. dominance)
- Gene-environment (GxE)
- Intra-locus (GxG)
- Parent-of-origin (POE)

but Gx? is not likely to explain everything…

Things we do not see/check

- Missing genome: X, mt, Y
- True causative variants (not tags!)
- Chromosomal re-arrangements
- Rare point mutations

Approaches to analysis of rare variants (rare = no power)

- Pooling approaches:
  - On frequency: large effect (on fitness) => low frequency
    - Large effect on trait not necessarily = large effect on fitness
    - Evolutionary modeling is tough: shifting & balancing selection, antagonistic pleiotropy…
  - On potential functionality: computational prediction
    - Works well for coding sequence
    - Rather miserable for non coding
  - On genomic context?
- Mixed models: assume certain distribution of effect (with some relation to anything above), integrate… very demanding!
- We need to see the data first? – coming!

Computational throughput

- Average throughput for fixed effects model
  - 30k tests per second (tps)
- Analysis of single trait with 10M markers
  - 1e7/3e4 = 5 minutes
- What about analyzing 30,000 expression traits?
  - 100 days = 2.5 months!
- Mixed models: time is quadratic on sample size
  - 3,000 people => 1 week
  - 10,000 people => 10 weeks = 2.5 months
- What about analyzing 30,000 expression traits using mixed models?!

Parallel computations

- Every test is independent
- If we had 1,000 computers, we could use the first to test traits 1 to 100, second to test 101 to 200, …
- Speed-up of ~1,000 times, thus 100 days → 2 ½hours

High-throughput analysis (of omics data)

- Smart factorization and approximation, and parallelization
- Parallelization for clusters, grid, cloud computing, GPU, (FPGA?)
  - Standard data format allowing (easy) data parallelization for different types of analysis
  - Computational ‘cache’ / ‘recyclable computations’? Unified cache ‘language’?

Storage and Input/Output (IO)

- 10,000 people imputed at 10M markers. Data size in plain text ~ 0.6 Tb
- Reading in RAM from plain text: 10Mb per second
- Time for reading = 0.6Tb/10Mbps = 16 hours (and analysis is done in 5 minutes!)

- Way out:
  - Binary format => 2 hours
  - Use advanced storage technology (RAID arrays) => 30 minutes

Conclusions

- Great progress achieved in the area of complex genetics during last 5 years, in part because of GWAS technology. Even continuing along the same lines will be very beneficial!
- Still, many methodological problems are to be solved, especially for small effects detection and interaction analyses. Mixed models is one way to go.
- General sequence variation analysis: a lot of work to be done – no conventional methodology yet, no software
- Storage and access to the data, computational throughput are important issues

Further courses

- Advanced statistical analysis of genomic variation, in particular mixed models:
  - GE03 (‘Advances …’)
  - GE05 (‘Family-based …’),
- Planned courses:
  - ‘High throughput computations for scientists’ (L. Karssen)
  - ‘Genomic variation analysis using R’ (Y. Aulchenko)