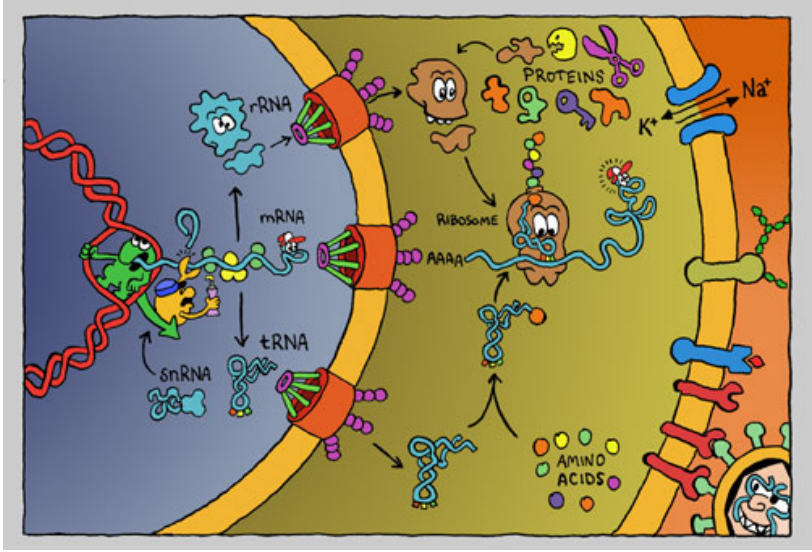
Statistical Genomics

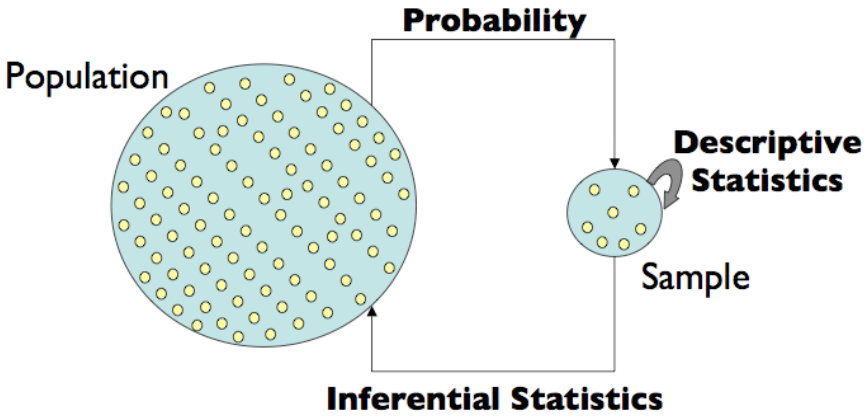
An introduction and some basic considerations

The central dogma of biology

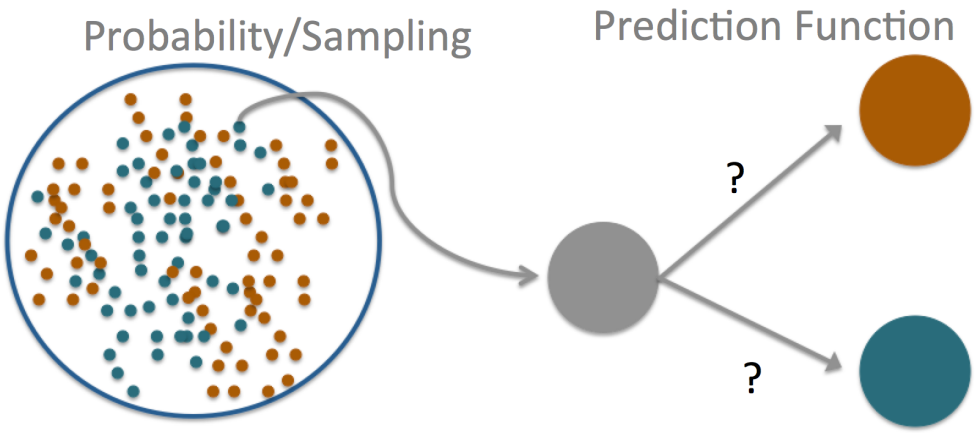


Ebbe Andersen (http://mb.au.dk)

The central dogma of statistics



The central dogma of prediction



Typical steps in a genomic study

- Determine the scientific question (biological).
- Select the study design (biological and statistical).
- Conduct the experiment (biological).
- Pre-process / normalize the data (statistical).
- Find differentially expressed genes, associations of genomic variants with a phenotype, … (statistical).

[LM]

Why do we look at genomic data?

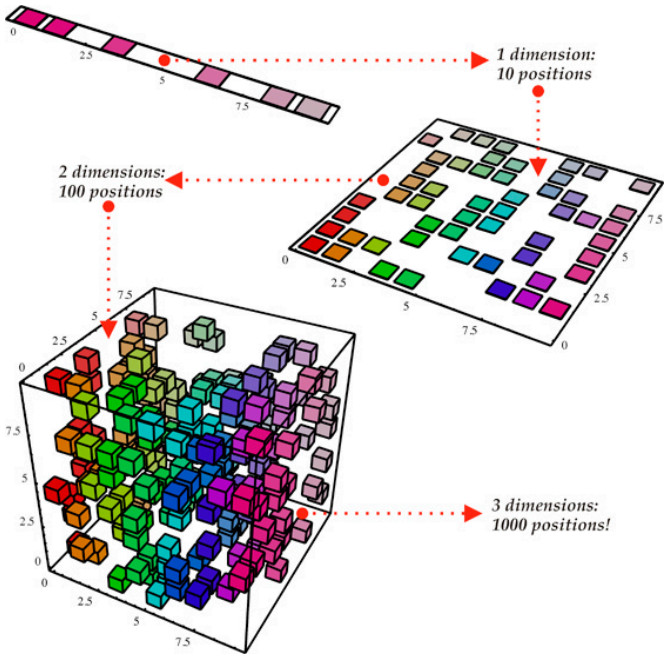
- Learn about basic biology.
- Identify drug targets.
- Find biomarkers.
  - Disease risk prediction.
  - Early detection of disease onset.
  - Prediction of response to treatment.
  - Diagnosis and disease monitoring.

Bioinformatics and Computational Biology are super-exciting fields to be in! With tons of genomic data being generated, this is a great time to use those skills for clinical and translational research (headed towards personalized medicine).

However, there are some aspects to all of this that are less than super-exciting. Sometimes they get lost in all the hype.

- Even the best device can have poor predictive performance.
- Poor experimental design is common, and can easily do people in. The Hall of Shame is well populated.
- Mistakes are easy to make with these high dimensional data, even with the best of intentions.
- It is not unusual that the technical artifacts in the genomic data are much larger than any biological signal.
- Quite frequently, you do have a "needle in the haystack" problem. The haystack will be hard to move, and your barn might not be large enough for the hay.
- Meet the curse of dimensionality!

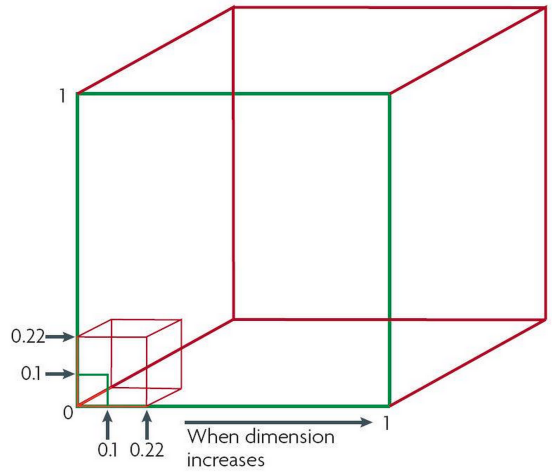
The curse of dimensionality



www.iro.umontreal.ca/~bengioy

The curse of dimensionality

```
> (0.01)^(1/(1:20))
 [1] 0.0100000 0.1000000 0.2154435 0.3162278 0.3981072 0.4641589 0.5179475
 [8] 0.5623413 0.5994843 0.6309573 0.6579332 0.6812921 0.7017038 0.7196857
[15] 0.7356423 0.7498942 0.7626986 0.7742637 0.7847600 0.7943282
> (0.01)^(1e-6)
[1] 0.9999954
```

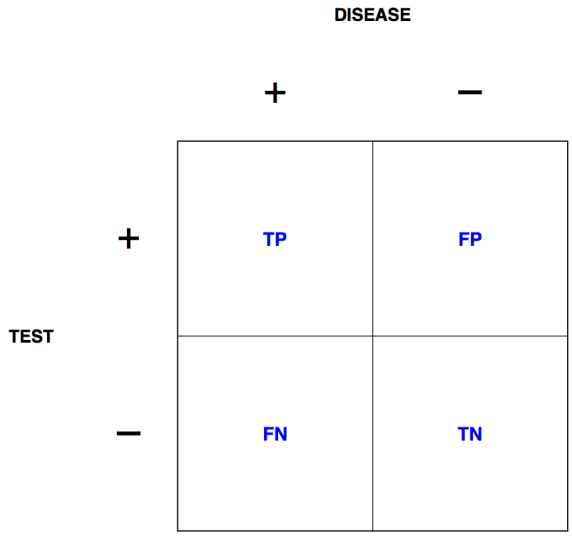


[PMID 18097463]

Personalized medicine

Assume you identified a gene signature that predicts the early onset of disease with 99% sensitivity and 99% specificity. What is the probability of a person having the disease given the result is positive, if we randomly select a subject from

- the general population with 0.1% disease prevalence?
- a high risk sub-population with 10% disease prevalence?



		DISEASE +	DISEASE −
TEST	+	TP	FP
	−	FN	TN

Sensitivity	→ Pr (positive test | disease)
Specificity	→ Pr (negative test | no disease)
Positive Predictive Value	→ Pr (disease | positive test)
Negative Predictive Value	→ Pr (no disease | negative test)
Accuracy	→ Pr (correct outcome)

		DISEASE +	DISEASE −
TEST	+	99	999
	−	1	98901

		DISEASE +	DISEASE −
TEST	+	99	999
	−	1	98901

Sensitivity	$\rightarrow$ 99 / (99+1) = 99%
Specificity	$\rightarrow$ 98901 / (999+98901) = 99%
Positive Predictive Value	$\rightarrow$ 99 / (99+999) $\approx$ 9%
Negative Predictive Value	$\rightarrow$ 98901 / (1+98901) $>$ 99.9%
Accuracy	$\rightarrow$ (99+98901) / 100000 = 99%

		DISEASE +	DISEASE −
TEST	+	9900	900
	−	100	89100

		DISEASE +	DISEASE −
TEST	+	9900	900
	−	100	89100

Sensitivity → 9900 / (9900+100) = 99%

Specificity → 89100 / (900+89100) = 99%

Positive Predictive Value → 9900 / (9900+900) ≈ 92%

Negative Predictive Value → 89100 / (100+89100) ≈ 99.9%

Accuracy → (9900+89100) / 100000 = 99%

Bayes rule

$$\Pr(A \mid B) =$$

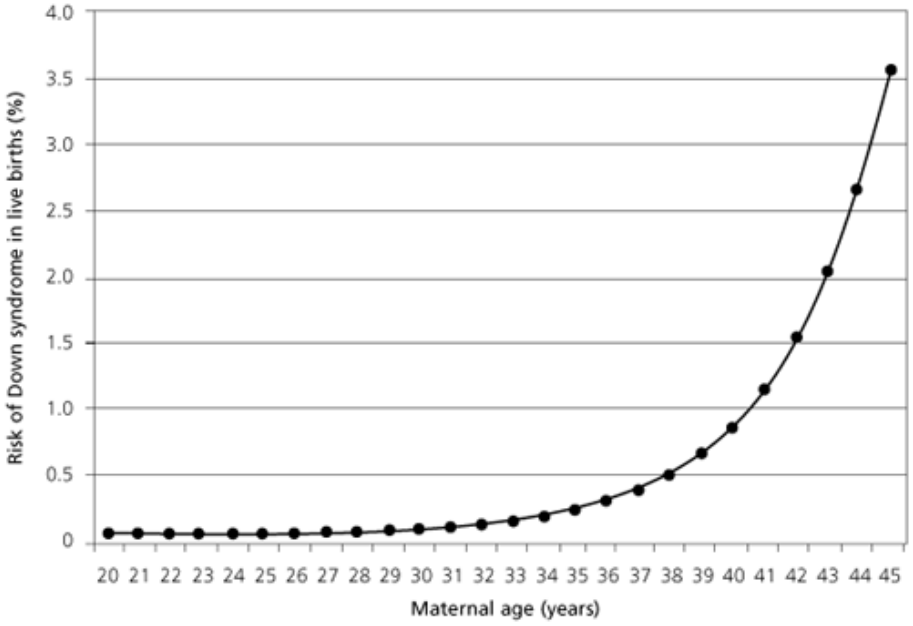
$$\Pr(A) \times \Pr(B \mid A) \;/\; \Pr(B) =$$

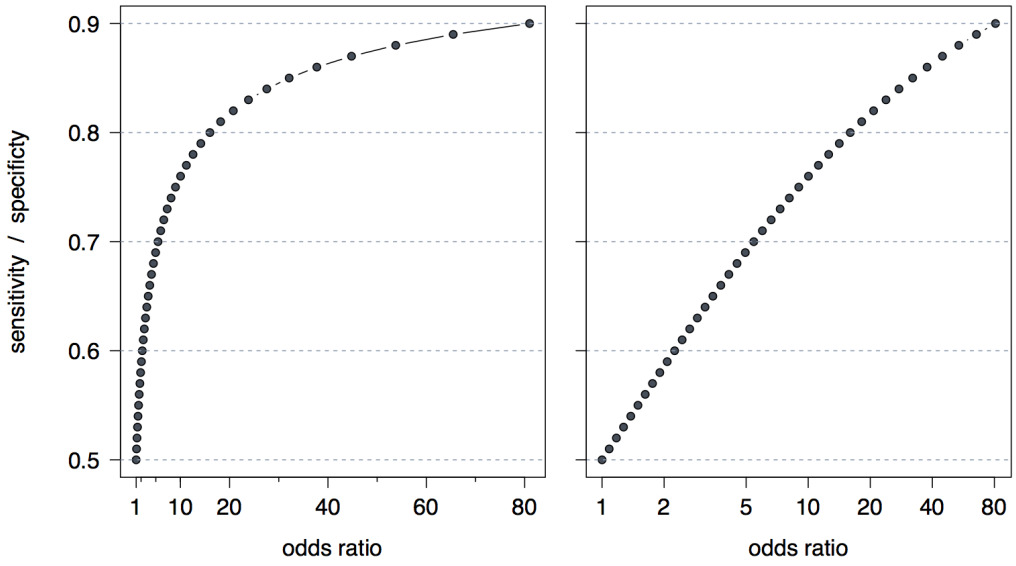
$$\Pr(A) \times \Pr(B \mid A) \,/\, \{\, \Pr(A) \times \Pr(B \mid A) + \Pr(\text{not } A) \times \Pr(B \mid \text{not } A) \,\}$$

Let A denote disease, and B a positive test result!

- ⟶ Pr(A | B) is the probability of disease given a positive test result.
- ⟶ Pr(A) is the prevalence of the disease.
- ⟶ Pr(not A) is 1 minus the prevalence of the disease.
- ⟶ Pr(B | A) is the sensitivity of the test.
- ⟶ Pr(not B | not A) is the specificity of the test.
- ⟶ Pr(B | not A) is 1 minus the specificity of the test.

Risk of Down syndrome





Sensitivity / Specificity / Prevalence
Positive predictive value

- 60% / 60% / 0.1% 0.15%
- 80% / 80% / 0.1% 0.4%
- 80% / 80% / 1.0% 3.9%
- 80% / 80% / 10% 30.8%

CD Genomics offers genetic testing panel which is based on a technology that assesses a complex but specific set of sites on the human genome -- Single Nucleotide Polymorphisms (SNPs) – which determines an individual's likelihood of disease.

GenSeq™ Disease Susceptibility Panel

Cancers (15)	Breast/Ovarian Cancer, Colorectal Cancer, Pancreatic Cancer, Endometrial Cancer, Esophageal Cancer, Renal Cancer, Bladder Cancer, Prostate Cancer, Hodgkin's Lymphoma, Follicular Lymphoma., Chronic Lymphocytic Leukemia, Meningioma, Abdominal Aortic Aneurysm, Melanoma
Cardiovascular Diseases (3)	Hypertension, Coronary Heart Disease, Venous Thromboembolism.
Neurological Diseases (3)	Parkinson's disease, Multiple Sclerosis, Alzheimer's Disease
Metabolic Disease (4)	Obesity, Gout, Kidney Stones, Gallstones
Immune System Diseases (3)	Type I Diabetes, Asthma, Rheumatoid Arthritis
Endocrine Diseases (3)	Type II Diabetes, Endometriosis, Hypothyroidism.
Inflammation (3)	Chronic Kidney Disease, Ankylosing Spondylitis, Chronic Obstructive Pulmonary Disease

By identifying your carrier status for mutations linked to 34 common diseases' susceptibility, we provide you and your family with the knowledge to help you prepare for the future.

By knowing more about your underlying health risks, you and your doctor can make more informed decisions about your healthcare.

CD Genomics
The Genomics Services Company

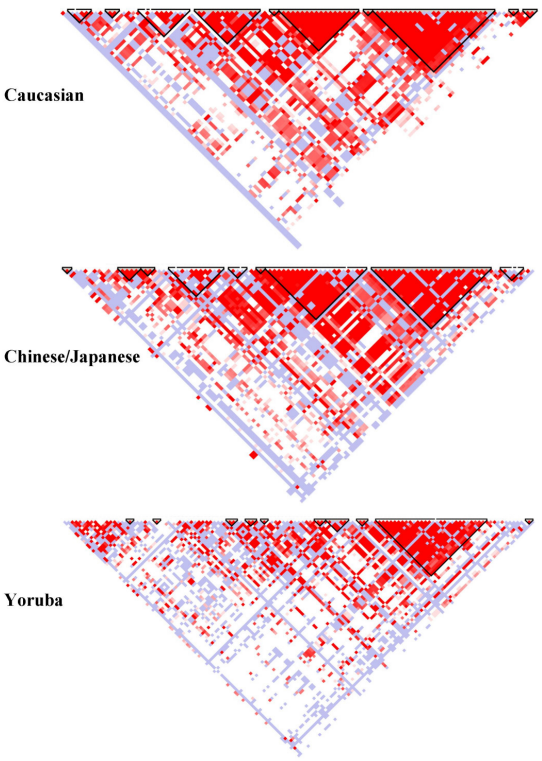
As of 06/19/14, the catalog includes 1922 publications and 13395 SNPs.

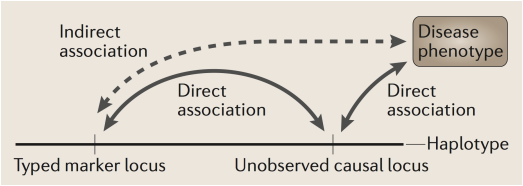
38 studies were returned in the search below

Download Spreadsheet of Search Results (For a description of the file spreadsheet column headings, go to: Tab Delimited File Heading Descriptions.)

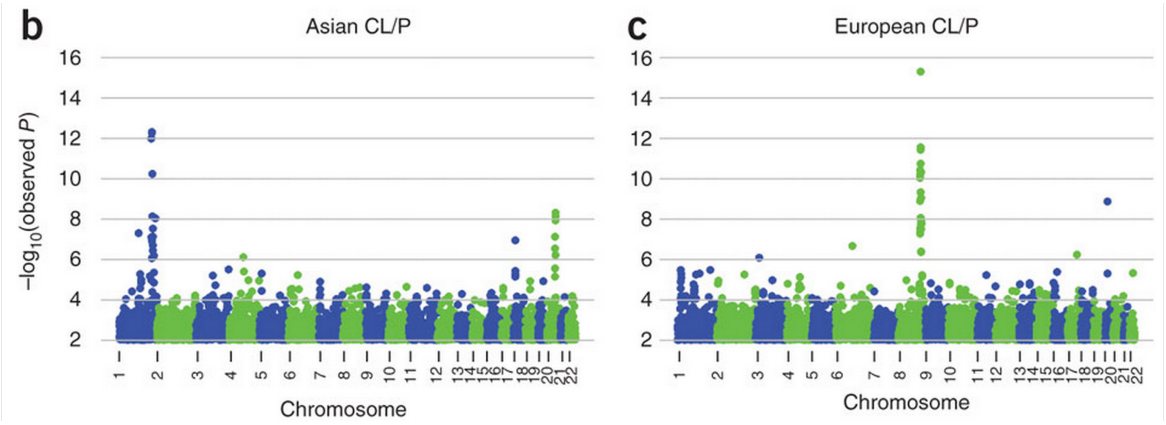
Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Mapped Gene(s)	Strongest SNP-Risk Allele	Context	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]
05/01/14	Hanson RL October 07, 2013 *Diabetes* A genome-wide association study in American Indians implicates DNER as a susceptibility locus for type 2 diabetes.	Type 2 diabetes	278 American Indian young-onset cases, 129 American Indian cases, and 424 American Indian controls from 514 sibships	1,273 American Indian cases, 1,635 American Indian controls, 793 cases, 3,133 controls	11p15.4	*KCNQ1*	KCNQ1	rs8181588-A	intron	0.48	5 x 10-9	1.3 [NR]	Affymetrix [453,654]
					2q36.3	*DNER*	DNER	rs1861612-T	intron	0.64	7 x 10-8	1.29 [NR]	
03/11/14	Hara K August 14, 2013 *Hum Mol Genet* Genome-wide association study identifies three novel loci for type 2 diabetes.	Type 2 diabetes	5,976 Japanese ancestry cases, 20,829 Japanese ancestry controls	18,207 Japanese ancestry cases, 6,780 Japanese ancestry controls, 6,209 Chinese ancestry cases, 7,205 Chinese ancestry controls	View full set of 16 SNPs								Illumina [6,209,637] (Imputed)
					11p15.4	*KCNQ1*	KCNQ1	rs2237892-C	intron	0.61	4 x 10-29	1.3 [1.24-1.36]	
					9p21.3	*CDKN2A, CDKN2B*	UBA52P6 - DMRTA1	rs10811661-T		0.55	1 x 10-18	1.23 [1.18-1.29]	
					10q25.2	*TCF7L2*	TCF7L2	rs7903146-T	intron	0.04	2 x 10-15	1.48 [1.34-1.63]	
					3q27.2	*IGF2BP2*	IGF2BP2	rs1470579-C	intron	0.34	5 x 10-14	1.19 [1.14-1.24]	
					6p22.3	*CDKAL1*	CDKAL1	rs7754840-C	intron	0.42	2 x 10-13	1.18 [1.13-1.23]	
					7q32.1	*MIR129, LEP*	LOC101928423	rs791595-A	intron	0.08	3 x 10-13	1.17 [1.12-1.22]	
					17p13.1	*SLC16A13*	SLC16A13	rs312457-G	intron	0.078	8 x 10-13	1.2 [1.14-1.26]	
					Xq28	*DUSP9*	KRT18P48 - DUSP9	rs5945326-A		0.68	2 x 10-12	1.14 [1.10-1.18]	
					9q34.3	*GPSM1*	GPSM1	rs11787792-A	intron	0.874	2 x 10-10	1.15 [1.10-1.20]	
					10q23.33	*HHEX*	HHEX - EXOC6	rs1111875-C		0.29	2 x 10-8	1.14 [1.09-1.20]	

www.genome.gov/gwastudies





PMIDs 16983374, 17214878



PMID 20436469

ARTICLE

Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data

Sebastian Zöllner and Jonathan K. Pritchard

Genomewide association studies are now a widely used approach in the search for loci that affect complex traits. After detection of significant association, estimates of penetrance and allele-frequency parameters for the associated variant indicate the importance of that variant and facilitate the planning of replication studies. However, when these estimates are based on the original data used to detect the variant, the results are affected by an ascertainment bias known as the "winner's curse." The actual genetic effect is typically smaller than its estimate. This overestimation of the genetic effect may cause replication studies to fail because the necessary sample size is underestimated. Here, we present an approach that corrects for the ascertainment bias and generates an estimate of the frequency of a variant and its penetrance parameters. The method produces a point estimate and confidence region for the parameter estimates. We study the performance of this method using simulated data sets and show that it is possible to greatly reduce the bias in the parameter estimates, even when the original association study had low power. The uncertainty of the estimate decreases with increasing sample size, independent of the power of the original test for association. Finally, we show that application of the method to case-control data can improve the design of replication studies considerably.

PMID 17357068

Table: Selected Personalized Medicine Drugs, Treatments and Diagnostics as of September 2011*

Indications in quotes and otherwise unattributed, are cited from the therapeutic or diagnostic product label.

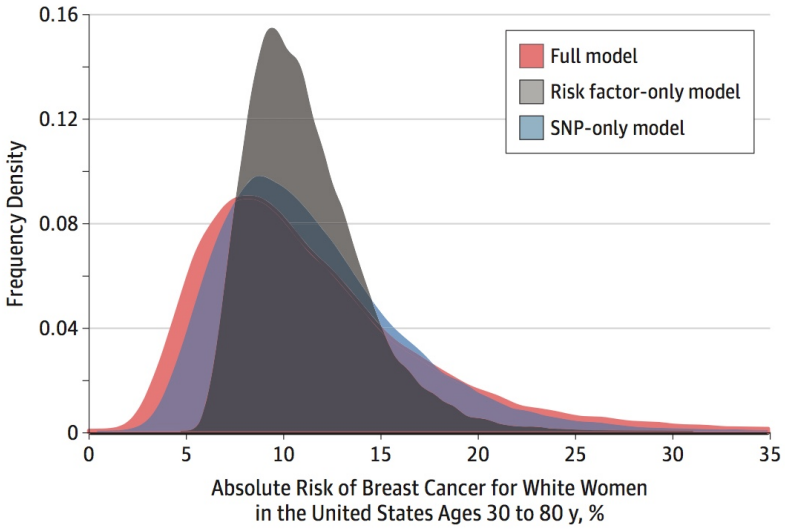
Therapeutic product labels contain pharmacogenomic information as:

- *Information only*
- *Recommended*
- *Required*
- *Unhighlighted products have no pharmacogenomic information, recommendations or requirements in the label.*

THERAPY	BIOMARKER/TEST	INDICATION
Mivacron® (mivacurium)	Cholinesterase gene	**Anesthesia adjunct**: "Mivacron is metabolized by plasma cholinesterase and should be used with great caution, if at all, in patients known to be or suspected of being homozygous for the atypical plasma cholinesterase gene."
Ansaid® (flurbiprofen)	*CYP2C9*	**Arthritis**: "*In vitro* studies have demonstrated that cytochrome P450 2C9 plays an important role in the metabolism of flurbiprofen to its major metabolite, 4'-hydroxy-flurbiprofen."
Depakote® (divalproex)	UCD *(NAGS; CPS; ASS; OTC; ASL; ARG)*	**Bipolar disorder**: "Hyperammonemic encephalopathy, sometimes fatal, has been reported following initiation of valproate therapy in patients with urea cycle disorders [UCDs]...particularly ornithine transcarbamylase deficiency [OTC]."
Aromasin® (exemestane) **Arimidex® (anastrozole)** **Nolvaldex® (tamoxifen)**	Estrogen Receptor (ER)	**Breast cancer**: Exemestane is indicated for adjuvant treatment of post-menopausal women with ER-positive early breast cancer. Anastrozole is for treatment of breast cancer after surgery and for metastases in post-menopausal women. Tamoxifen is the standard therapy for estrogen receptor-positive early breast cancer in pre-menopausal women.
Chemotherapy	Mammostrat®	**Breast cancer**: Prognostic immunohistochemistry (IHC) test used for postmenopausal, node negative, estrogen receptor expressing breast cancer patients who will receive hormonal therapy and are considering adjuvant chemotherapy.
Chemotherapy	MammaPrint®	**Breast cancer**: Assesses risk of distant metastasis in a 70-gene expression profile.
Chemotherapy	Onco*type* DX® 16-gene signature	**Breast cancer**: A 16-gene signature (plus five reference genes) indicates whether a patient has a low, intermediate, or high risk of having a tumor return within 10 years. Low-risk patients may be treated successfully with hormone therapy alone. High-risk patients may require more aggressive treatment with chemotherapy.

www.personalizedmedicinebulletin.com

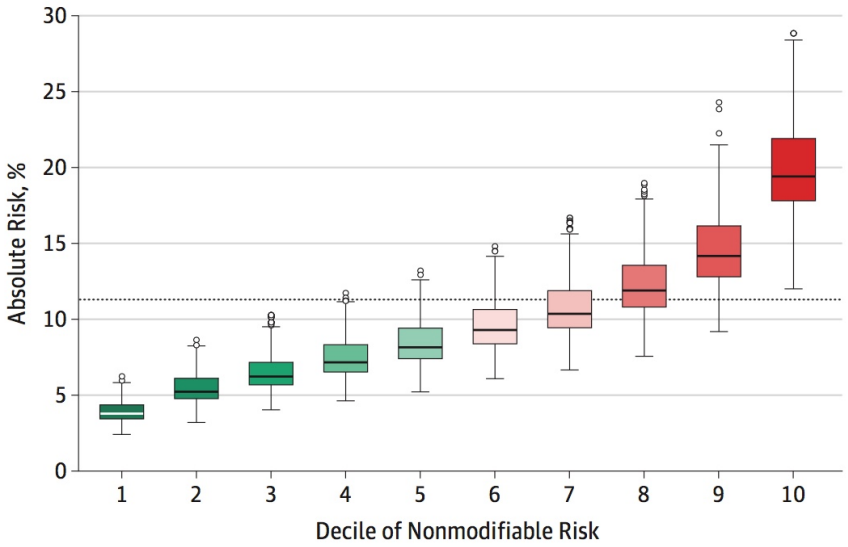
Figure 1. Projected Distribution of Absolute Lifetime Risk of Breast Cancer for White Women in the United States Ages 30 to 80 Years



SNP indicates single nucleotide polymorphisms.

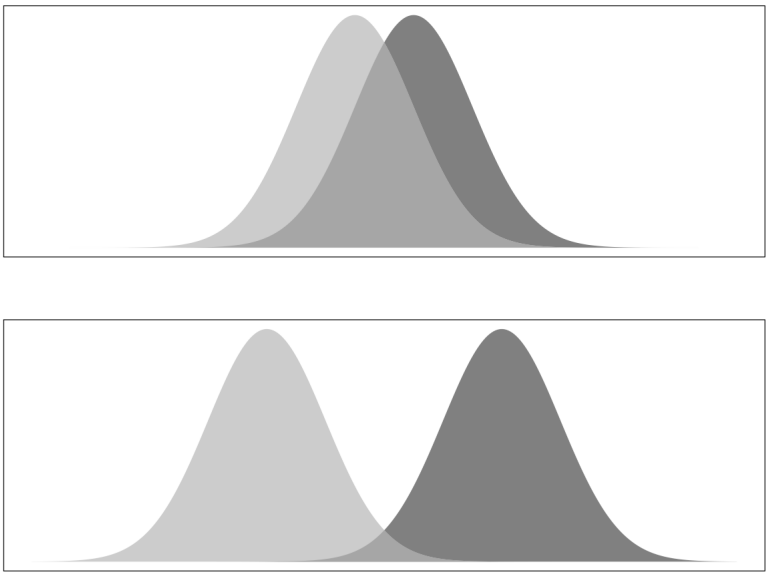
PMID 27228256

Figure 3. Distribution of Absolute Lifetime Risk Associated With Modifiable Risk Factors Stratified by Deciles of Nonmodifiable Risk for White Women in the United States

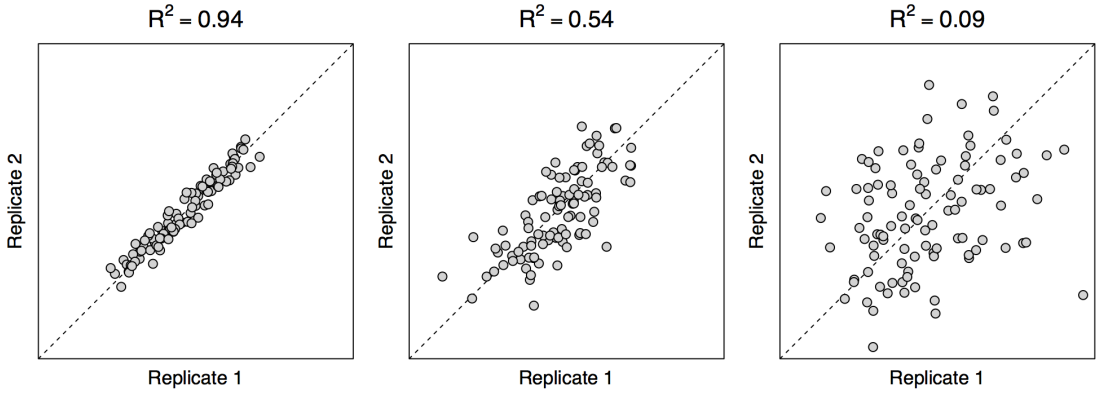


PMID 27228256

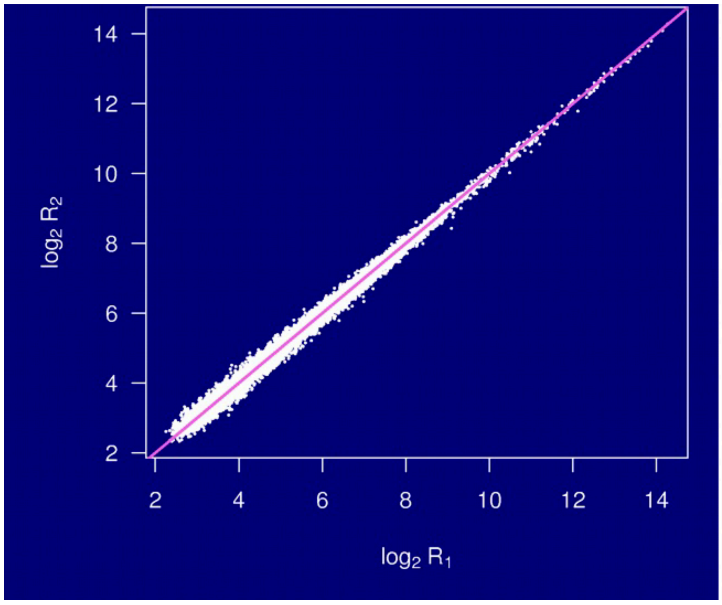
→ Association versus Prediction



Technical replicates



Technical replicates



McIntyre *et al. BMC Genomics* 2011, **12**:293
http://www.biomedcentral.com/1471-2164/12/293

BMC Genomics

RESEARCH ARTICLE Open Access

RNA-seq: technical variability and sampling

Lauren M McIntyre[1*], Kenneth K Lopiano[2], Alison M Morse[1], Victor Amin[1], Ann L Oberg[3], Linda J Young[2] and Sergey V Nuzhdin[4]

Abstract

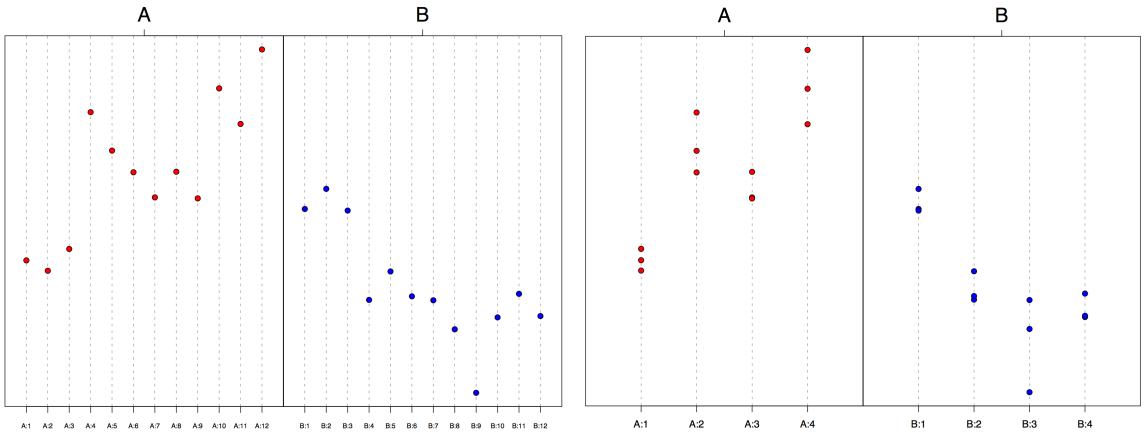
Background: RNA-seq is revolutionizing the way we study transcriptomes. mRNA can be surveyed without prior knowledge of gene transcripts. Alternative splicing of transcript isoforms and the identification of previously unknown exons are being reported. Initial reports of differences in exon usage, and splicing between samples as well as quantitative differences among samples are beginning to surface. Biological variation has been reported to be larger than technical variation. In addition, technical variation has been reported to be in line with expectations due to random sampling. However, strategies for dealing with technical variation will differ depending on the magnitude. The size of technical variance, and the role of sampling are examined in this manuscript.

Results: In this study three independent Solexa/Illumina experiments containing technical replicates are analyzed. When coverage is low, large disagreements between technical replicates are apparent. Exon detection between technical replicates is highly variable when the coverage is less than 5 reads per nucleotide and estimates of gene expression are more likely to disagree when coverage is low. Although large disagreements in the estimates of expression are observed at all levels of coverage.

Conclusions: Technical variability is too high to ignore. Technical variability results in inconsistent detection of exons at low levels of coverage. Further, the estimate of the relative abundance of a transcript can substantially disagree, even when coverage levels are high. This may be due to the low sampling fraction and if so, it will persist as an issue needing to be addressed in experimental design even as the next wave of technology produces larger numbers of reads. We provide practical recommendations for dealing with the technical variability, without dramatic cost increases.

PMID 21645359

Technical replicates

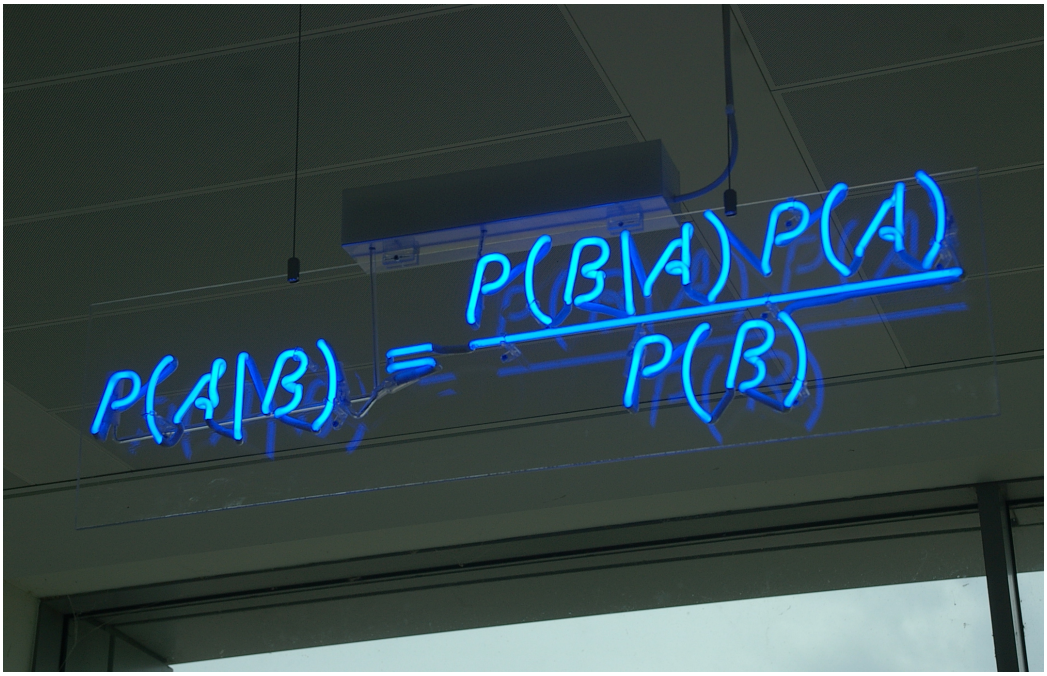


Distributions you should know

- Normal (Gaussian) distribution
- t distribution
- Chi-square distribution
- F distribution
- Binomial distribution
- Poisson distribution
- Gamma distribution
- Negative Binomial distribution
- Beta distribution
- Beta Binomial distribution
- Multinomial distribution

1 parameter
2 parameters
3+ parameters

And how can we estimate some of those parameters…?



MODELLING

Bayesian statistical methods for genetic association studies

Matthew Stephens and David J. Balding*[‡,§]

Abstract | Bayesian statistical methods have recently made great inroads into many areas of science, and this advance is now extending to the assessment of association between genetic variants and disease or other phenotypes. We review these methods, focusing on single-SNP tests in genome-wide association studies. We discuss the advantages of the Bayesian approach over classical (frequentist) approaches in this setting and provide a tutorial on basic analysis steps, including practical guidelines for appropriate prior specification. We demonstrate the use of Bayesian methods for fine mapping in candidate regions, discuss meta-analyses and provide guidance for refereeing manuscripts that contain Bayesian analyses.

PMID 19763151