Introduction to linear models and multiple testing

Ghislain Durif, Laurent Modolo, Franck Picard

Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

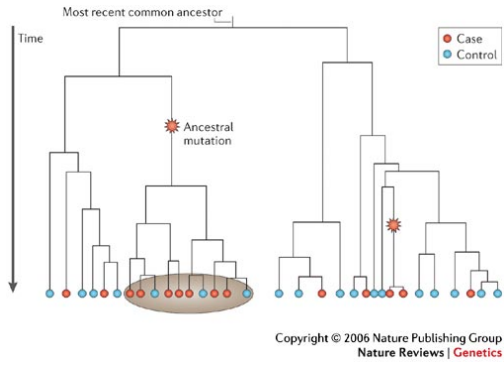
franck.picard@ens-lyon.fr

Outline

1. **Genetic Association studies**
2. ANOVA to test for association
3. Differential Expression Analysis for sequencing data
4. Multiple Testing

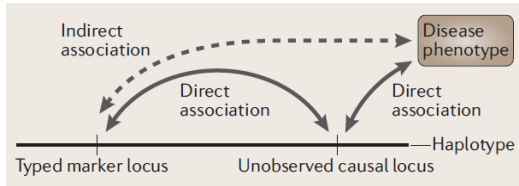
Basic Principles

- Identification of polymorphisms that impact mesurable phenotypes
- On non-related individuals (or distant kinship)
- Identify polymorphisms that systematically vary between individuals in different states
- The basic idea is to detect variants that are more present in cases wrt controls



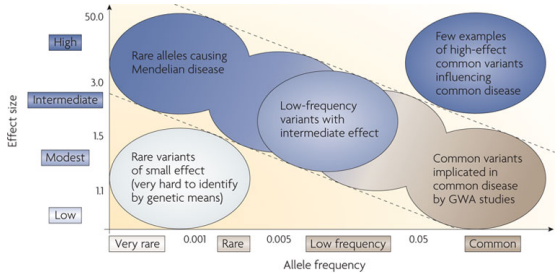
Different types of association studies

- Targetted polymorphism: one targetted locus (known)
- Candidate Gene Approach: known gene (associated with cases), study 5-10 SNPs in the gene
- Fine Mapping: a given region with many genes (1-10Mb), $\sim$100 SNP
- Genome-wide: need a catalog of SNPs ($\sim$300,000) for a search of case-associated variants



Challenges in association studies

- Easy cases: causal SNP with direct genotype/phenotype relation
- Complex cases: phenotype is difficult to define/measure. Complex and partially known genotype/phenotype relation
- Environmental impact ?
- Frequency and size of effects are two main components of association studies



QTL/Association

- Association case/control: association between a locus with a discrete response (contingency table, Fisher test)
- The testing strategy does not allow the inclusion of other factors (weight, age, clinic)
- QTL (quantitative trait loci): association of a locus with a quantitative trait
- In this case we use a regression model with the trait as response variable, and the number of alleles as covariates (explanatory variables)

Data

	M_1	M_2	...	M_p	status	Age	Sex	Glycemia
$i = 1$	0	1		0	0	38	F	0.8
$i = 2$	1	0		2	1	15	M	0.2
$\vdots$								
$i = N$	0	2		1	0	90	F	1.5

- For each individual we know the genotype on many markers (for instance p SNPs).
- We can have clinial data (non genomic)
- How to explain the variations of a response given genotyping and clinical data ?

Basic Framework in Quantitative Genetics

- The general framework in quantitative genetics relates the observed phenotype to genetic and environmental components: $P = G + E + G \times E$
- Supposing independence between components (?), the variance can be written such that $V_P = V_G + V_E + V_{G \times E}$
- Heritability (broad sense)can be defined as: $H^2 = V_G / V_P$
- Heritability (strict sense) can also be written only with the additive part of the genetic variance $h^2 = V_A / V_P$

Missing Heritability ?

- Despite many association studies part of the variability of many traits remains unexplained
- Can we include other genetic markers (CNV, epistasy)
- Question the "Common variant / Common disease" hypothesis ?
- Impact of rare variants
- Individual variations should be modeled more carefully in statistical models

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Outline

Notations

- Denote by I the total number of genotypes, and $i \in \{1, \ldots, I\}$ the i^{th} genotype for instance
$$I = 3, \quad \text{and } i \in \{\texttt{AA}, \texttt{Aa}, \texttt{aa}\}$$
- Denote by j the j^{th} replicate of genotype i, with n_i the number of replicates in genotype I.
- Denote by y_{ij} the trait of the individual j of genotype i
- We observe $\mathbf{y}$ a vector of size $n = \sum_i n_i$ made of I vectors $\mathbf{y}_i$:
$$\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_I]$$
- Each $\mathbf{y}_i = [y_{i1}, \ldots, y_{i,n_i}]$, with $\mathbf{y}_i = [y_{ij}]$ $j = 1, \ldots, n_i$.

First model

- We suppose that the observed trait y_{ij} is the realization of a Gaussian variable Y_{ij} such that:

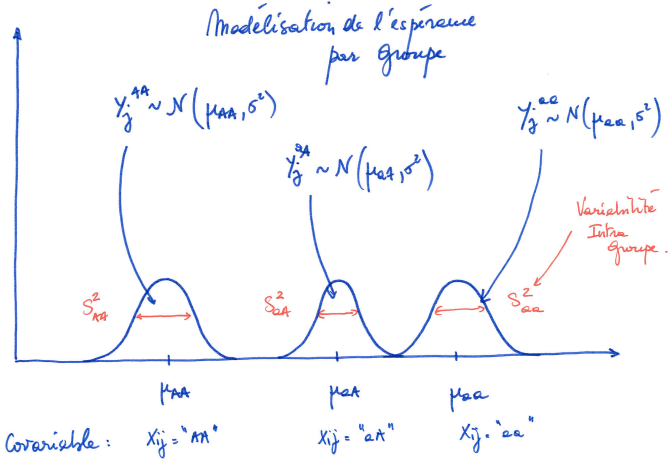
$$Y_{ij} \sim \mathcal{N}\Big(\mathbb{E}(Y_{ij}), \mathbb{V}(Y_{ij})\Big)$$

- The model concerns the expectation of Y_{ij}

$$\mathbb{E}(Y_{ij}) = \mu_i, \quad \mathbb{V}(Y_{ij}) = \sigma^2$$

- The model supposes that the expectation of the trait only depends on the genotype, and that the variance of the trait is constant
- We suppose that Y_{ij} are independents (intra and inter-genotype)

Illustration



New formulation with the conditional expectation

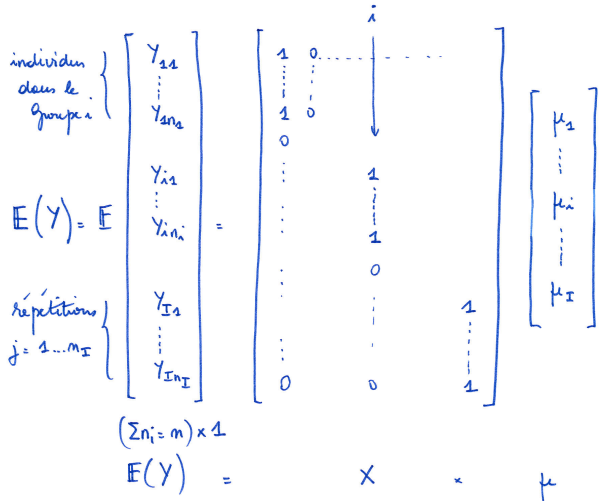
- Let's introduce the covariate X_{ij} such that $X_{ij} = 1$ if individual j has genotype i and $X_{ij} = 0$ otherwise
- If we had to model the distribution of X_{ij} we could consider a multinomial distribution
- But we are not interested in the variations of $\mathbf{X}$, but of those of $\mathbf{Y}$ with a given genotype ($\mathbf{X}$ is fixed)
- The model consists in writing the conditional expectation of $\mathbf{Y}$ once $\mathbf{X}$ has been observed:

$$\mathbb{E}(Y_{ij}|X_{ij} = 1) = \mu_i, \ \mathbb{V}(Y_{ij}|X_{ij} = 1) = \sigma^2$$

- This can also be written such that:

$$\mathbb{E}(Y_{ij}|X_{ij} = x_{ij}) = \sum_{i=1}^{I} \mu_i x_{ij}$$

Notation matricielle



Defining Residuals

- We want to decompose the expectation of the signal wrt covariates **X**
- Residuals are defined as what is left once this contribution has been removed
- Introduce a new variable called residual such that

$$E_{ij} = Y_{ij} - \mu_i, \ E_{ij} \sim \mathcal{N}(0, \sigma^2) \ (iid)$$

- It's a random error term, that is defined as the difference between the observations and what is expected by the model
- Using residuals the model becomes

$$Y_{ij} = \mu_i + E_{ij}, \ E_{ij} \sim \mathcal{N}(0, \sigma^2) \ (iid)$$

$$\text{Noisy Observations} = \text{signal} + \text{random errors}$$

- We suppose that the signal is in the expectation of the model

Parameters and estimators

- Parameters $(\mu_i)_i$,I parameters for the mean + 1 parameter for the variance
- The mean square criterion is

$$d^2(\mathbf{Y}, \boldsymbol{\mu}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} E_{ij}^2$$

- The MS estimator of the mean is the empirical mean

$$\widehat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i} Y_{i+} = Y_{i\bullet}$$

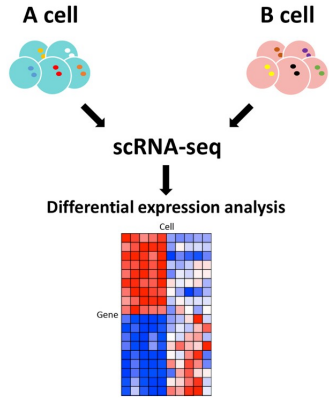
- One estimator of the variance is the so-called residual variance

$$\widehat{\sigma}^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \widehat{\mu}_i)^2$$

Outline

Basic Principles

- Many studies in Genomics can be restated as a comparison problem
- For each gene, is the expression different from one condition to another ?
- Compare average expression wrt biological variance
- How to compute biological variability ?
- Account for confounding effects (technical variability)



The ANOVA framework

- Y_{ijr} : expression (continuous) for gene i in condition j at replicate r
- Perform DE between conditions using model

$$Y_{ijr} \sim \mathcal{N}\left(\mathbb{E}(Y_{ijr}), \sigma^2\right)$$

$$\mathbb{E}(Y_{ikr}) = \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

- The parameters of the model are interpreted as :
  - α_i : mean expression of gene i (across conditions),
  - β_j : mean expression in condition j (across genes),
  - $(\alpha\beta)_{ij}$: interaction effect gene x condition
- Allows to integrate normalization while testing

Testing framework

- **Hypothesis** : no expression difference between conditions

$$\mathcal{H}_0^i : \{(\alpha\beta)_{i1} = (\alpha\beta)_{i2}\}$$

- The classical statistic for gene i is the Student statistic

$$T_i = \frac{|\widehat{\alpha\beta}_{i1} - \widehat{\alpha\beta}_{i2}|}{\widehat{\sigma}} \times \sqrt{2R-2} \underset{\mathcal{H}_0}{\sim} \mathcal{T}(2R-2)$$

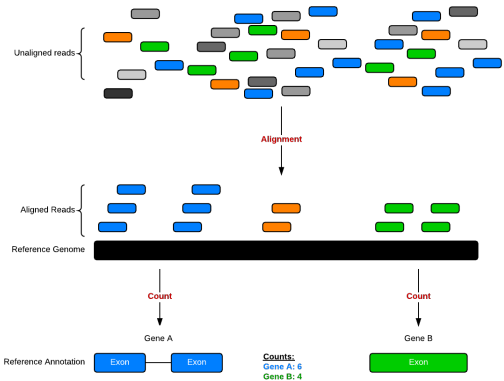
- Estimation of mean fixed effects is done by **Maximum Likelihood**

What about the estimation of the dispersion parameter ?

- Refinements / difficulties concern the **estimation of** σ, the dispersion parameter
- A **common variance** to all genes σ^2 : robust but lacks of power
- A **specific variance** to every gene σ_i^2 : powerful but sensitive to outliers,
  - Large sampling variance
  - To be stabilized empirically
- **Groups of variances** (combination of both)

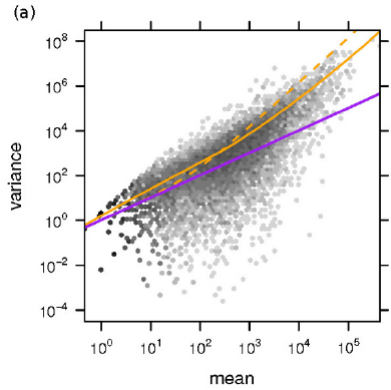
Sequencing data are count data

- Sequencing technologies provide read counts
- The underlying statistical distribution is Poisson
- Poisson variables show specific patterns of variability
- Heteroskedasticity : variance function of the mean



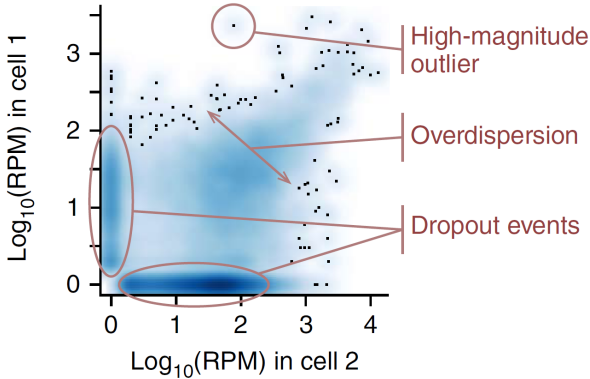
Overdispersed count distribution

- For Poisson variables $\mathbb{E}(X) = \mathbb{V}(X)$
- Sequencing data are overdispersed and the best model is the negative Binomial distribution:
  $V(\mu) = \mu + \kappa\mu^2$
- Very challenging to properly estimate the biological variability



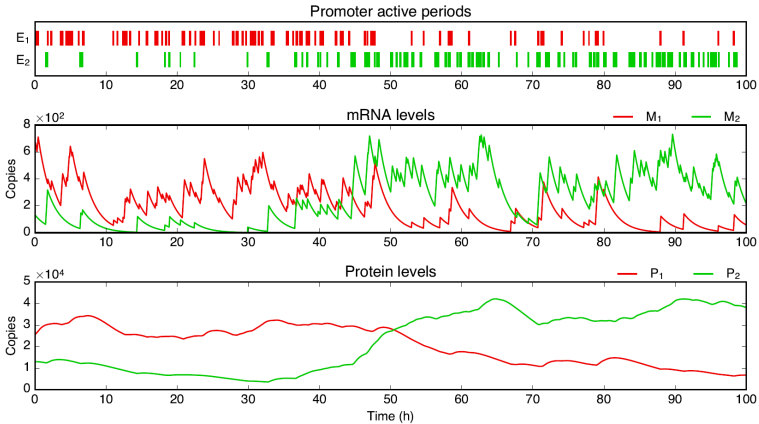
from [?]

How bad is the situation in single cell data ?



Overdispersion is mainly biological because diversity is high between cells [**?**]

Expression is a stochastic bursty process



The curse of Dropouts

- Low starting amount of RNAs: transcripts will be missed during RT
- Amplification is needed ($\times 10^6$), which creates distortions
- Stochasticity of gene expression (bursty process) sparsity of the data, high proportion of zeros
- Dropout depends on cells (different in different wells),
- Lowly expressed genes : sampling / amplification issues
- Highly expressed genes: is more likely to indicate a burst

The Generalized Linear Model framework

- Y_{ijr} : the **read count** (positive integer), for gene i in condition j
- Define the **Generalized Linear Model** (GLM) by setting

$$\begin{aligned} Y_{ijr} &\sim \mathcal{P}(\mu_{ij}) \\ \log \mathbb{E}(Y_{ijr}) = \log(\mu_{ij}) &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \end{aligned}$$

- Parameters have the same interpretation
- Testing hypotheses are similar : $\mathcal{H}_0^i : \{(\alpha\beta)_{i1} = (\alpha\beta)_{i2}\}$
- Dispersion parameter ? Test statistics ?

Testing Strategies based on LRT

- Compare different models, for instance

$$\begin{aligned} \log(\mu_{ij}) &= \mu + \alpha_i + \beta_j \\ \log(\mu_{ij}) &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \end{aligned}$$

- Use the Ratio of log likelihoods as a Statistics, which incorporates all infos:

$$LRT = -2 \log \left(\frac{\mathcal{L}(\widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{\alpha\beta})}{\mathcal{L}(\widehat{\mu}, \widehat{\alpha}, \widehat{\beta})} \right) \underset{\mathcal{H}_0}{\sim} \chi^2(\Delta df)$$

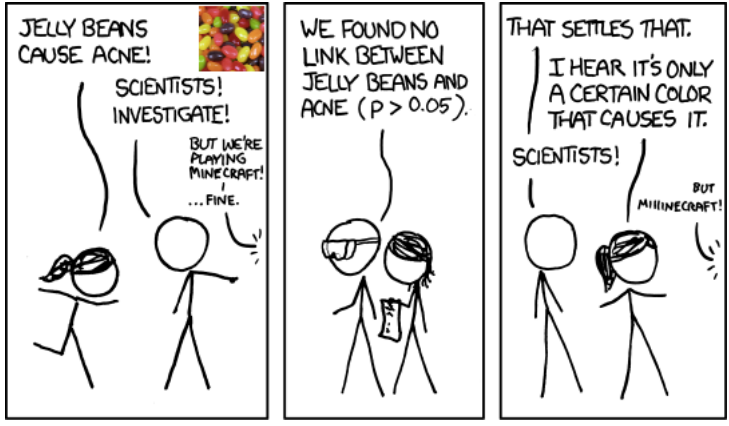
- This has been shown to be the best strategy on Sequencing data

Conclusion: don't think Normal !

- Use **Generalized Linear Models** to perform Count regression, and not Gaussian regression on the log-counts
- Incorporate effects in the model to perform a global analysis that **accounts for distributional characteristics**
- Do not perform tests that imply Poisson distribution when data are over-dispersed
- Use **Likelihood Ratio Tests** to compare models
- Overdispersion leads to estimation issues due to **numerical problems**

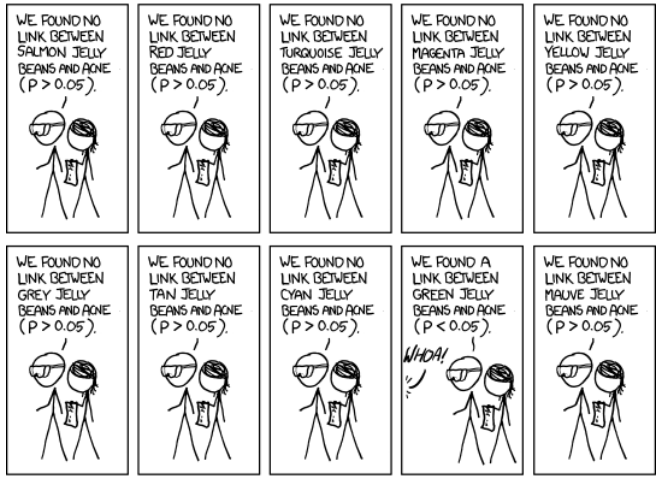
Outline

Example



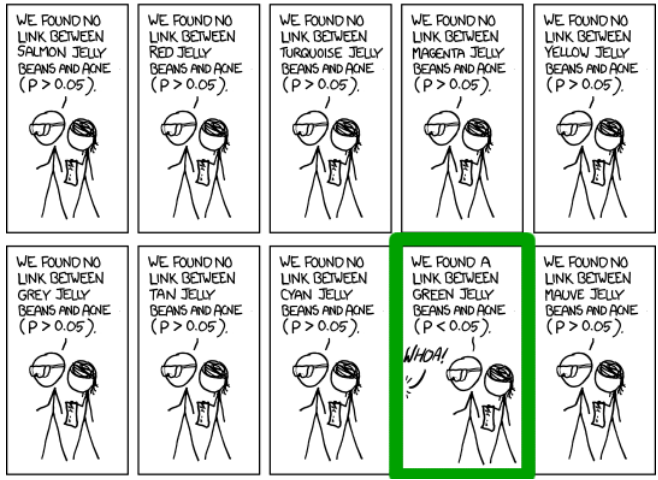
from `http://imgs.xkcd.com/comics/significant.png`

Example



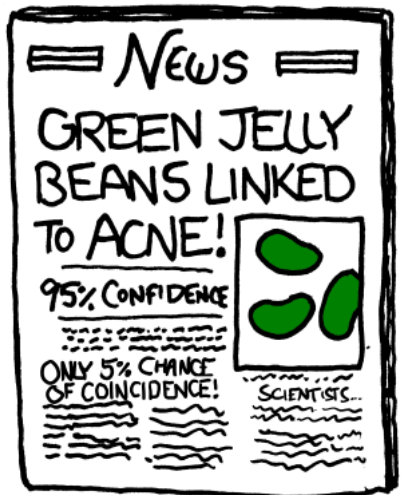
from `http://imgs.xkcd.com/comics/significant.png`

Example



from `http://imgs.xkcd.com/comics/significant.png`

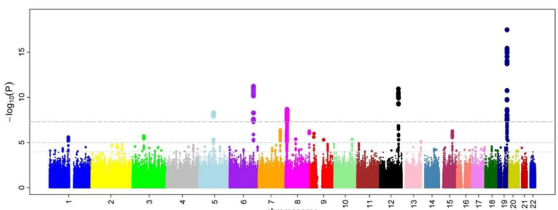
Example



from http://imgs.xkcd.com/comics/significant.png

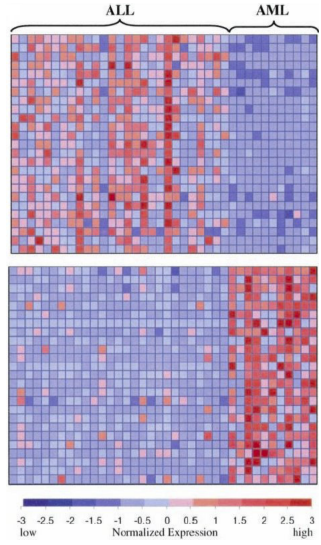
GWAS

- $m \sim 10^6$ tests (genomic markers)
- $n \sim 10^3 - 10^4$ observations (individuals)
- Which markers are significantly associated with a phenotype of interest?



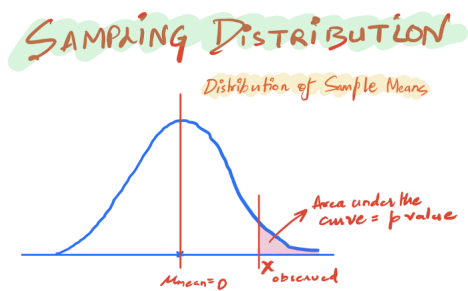
Gene expression analysis

- One test for each gene!
- $m \sim 10^4$ tests (genes/transcripts/splicing variants)
- $n \sim 10^1 - 10^3$ observations (individuals)
- Which genes are differentially expressed?



Why is the p-value so important ?

- A null hypothesis supposes the absence of effect $\mathcal{H}_0$
- We can determine what could be the expected behavior of the data if the null hypothesis were true
- Risk of taking the wrong decision under $\mathcal{H}_0$
- p-value quantifies the risk of a procedure



https://towardsdatascience.com/

Definition of the risk for one hypothesis

- The test procedure provides a test statistics used to build a decision rule
- $p_v(\mathbf{x}) = \mathbb{P}(T(\mathbf{X}) > t_{obs}(\mathbf{x}) \mid \mathcal{H}_0)$
- Reject if $p_v(\mathbf{x}) < \alpha$
- α is an admissible risk
- Control α while maximizing power $1 - \beta$

	$\mathcal{H}_0$ true	$\mathcal{H}_0$ false
Accept	$1 - \alpha$	β
Reject	α	$1 - \beta$

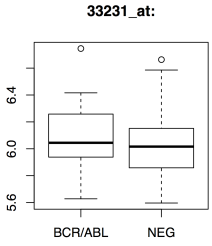
Example with the Leukemia dataset

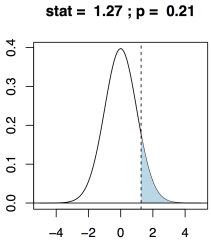
- Chiaretti et. al., Clinical cancer research, 11(20):7209–7219, 2005
- Data and code available from `https://pneuvial.github.io/sanssouci/`
- Expression measurements (mRNA) $m = 9838$ genes
- Marginal testing (gene by gene)

BRC/ABL	NEG	n
37	42	79

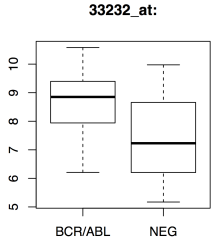
Which genes differ between BRC/ABL and NEG ?

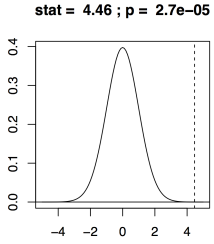
Illustration





No evidence of difference between groups





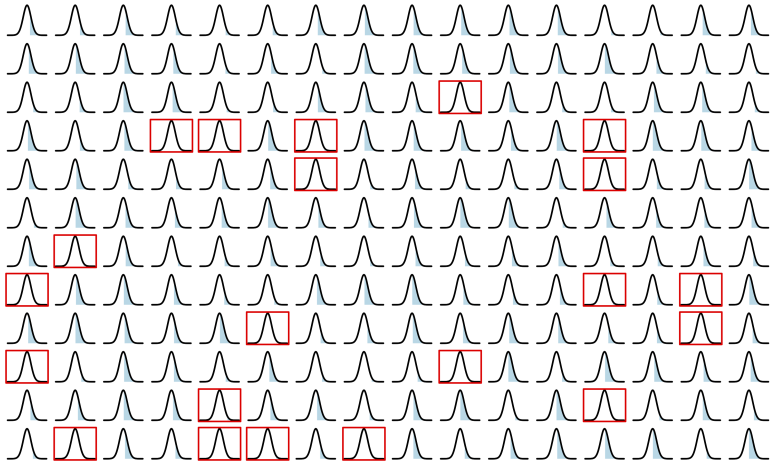
Some evidence of difference between groups.
"Significant"?

Definition of the risk for many hypotheses

- Consider m hypotheses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$
- Perform m tests based on $p_1, \ldots, p_m$
- m_0 is the total number of true null hypotheses
- Consider a threshold t_j such that $\mathcal{H}_0^j$ is rejected if $p_j < t_j$
- For many hypothesis, control the expected number of false positives $\mathbb{E}(V)$

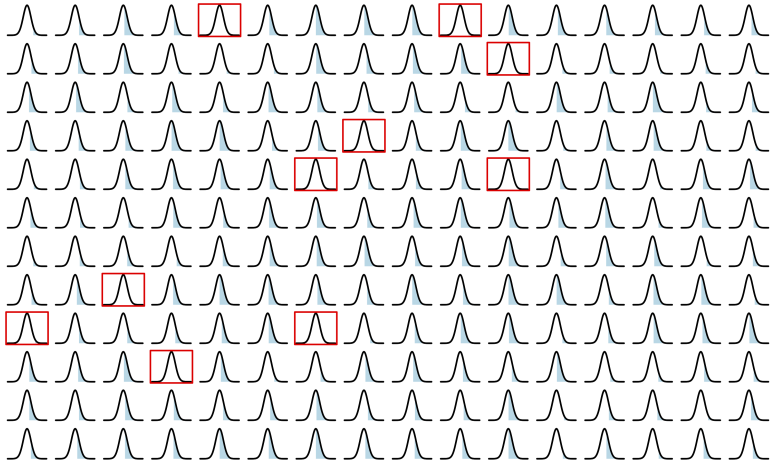
	$\mathcal{H}_0$ true	$\mathcal{H}_0$ false	Total
Accept	U	T	$m - R$
Reject	V	S	R
Total	m_0	$m - m_0$	m

First 192 genes of the Leukemia data set:



Genes with p-value < 0.05 highlighted in red

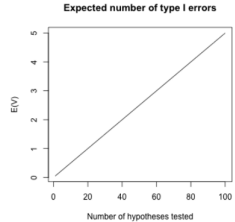
When considering 192 simulated with no effect



Genes with p-value < 0.05 highlighted in red

Considering a fixed threshold does not control the risk

- Data: $p_1, \dots p_m$: p-values for m tests
- Strategy: reject $\mathcal{H}_0$ for all i such that $p_i \leq \alpha$
- Expected # of type I errors scales linearly with m
- Probability of a type I error quickly grows to 1



Notations

- $\mathcal{H} = \{1, \dots m\}$ m null hypotheses to be tested
- $\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses,
- $\mathcal{H}_1 = \mathcal{H} \setminus \mathcal{H}_0$
- $m_0 = |\mathcal{H}_0|$, $\pi_0 = m_0/m$
- $(p_i)_{1 \leq i \leq m}$: p-values
- R: a set of rejected hypotheses
- $V = |R \cap \mathcal{H}_0|$: number of "false positives" within R.

Multiple testing risks and their control

- Family-Wise Error Rate:
$$FWER = \mathbb{P}(V > 0)$$
- False Discovery Rate:
$$FDR = \mathbb{E}\left(\frac{V}{|R| \vee 1}\right)$$
- Aim : from the data determine the set of rejected hypotheses R, by choosing a threshold $\hat{t}$ such that:
$$R = \left\{i \in \mathcal{H} \mid p_i < \hat{t}\right\}$$
- How can we control these risks ? (dependency assumptions, power/conservativeness, algorithms and their implementations)

Expected # of type I errors scales linearly with m

- $p_1, \ldots p_m$: p-values for m tests
- Strategy: reject $\mathcal{H}_0$ for all i such that $p_i \leq \alpha$
- Recall: $V = \sum_{i \in \mathcal{H}_0} 1_{p_i \leq \alpha}$

$$E(V) = \sum_{i \in \mathcal{H}_0} E_{\mathcal{H}_0}(1_{p_i \leq \alpha})$$

$$E(V) = \sum_{i \in \mathcal{H}_0} P_{\mathcal{H}_0}(p_i \leq \alpha) = \sum_{i \in \mathcal{H}_0} \alpha = |\mathcal{H}_0|\alpha = \pi_0 m \alpha$$

Probability of a type I error quickly grows to 1

- $p_1, \dots p_m$: p-values for m tests
- Strategy: reject $\mathcal{H}_0$ for all i such that $p_i \leq \alpha$
- Recall: $V = \sum_{i \in \mathcal{H}_0} 1_{p_i \leq \alpha}$

$$P(V = 0) = \mathbb{P}(\forall i \in \mathcal{H}_0, p_i > \alpha)$$

Assuming independent tests:

$$\begin{aligned} P(V = 0) &= \prod_{i \in \mathcal{H}_0} \mathbb{P}(p_i > \alpha) \\ &= \prod_{i \in \mathcal{H}_0} (1 - \alpha) \\ &= (1 - \alpha)^{m_0} \end{aligned}$$

Hence $P(V > 0) = 1 - (1 - \alpha)^{m_0}$

FWER control with the Bonferroni procedure

- Definition: Reject all i such that $p_i \leq \alpha/m$
- Properties: FWER control at level $\pi_0\alpha(\leq \alpha)$ under arbitrary dependence
- Limitation: Conservativeness: α/m can be small!
- Directions for increased power :

  - other dependency assumptions: independence, positive dependence;

  - estimation of π_0

Proof

- Let $V(t)$ be the number of false positives obtained by rejecting all p-values less than t:

$$V(t) = \sum_{i \in \mathcal{H}_0} 1_{p_i \leq t}$$

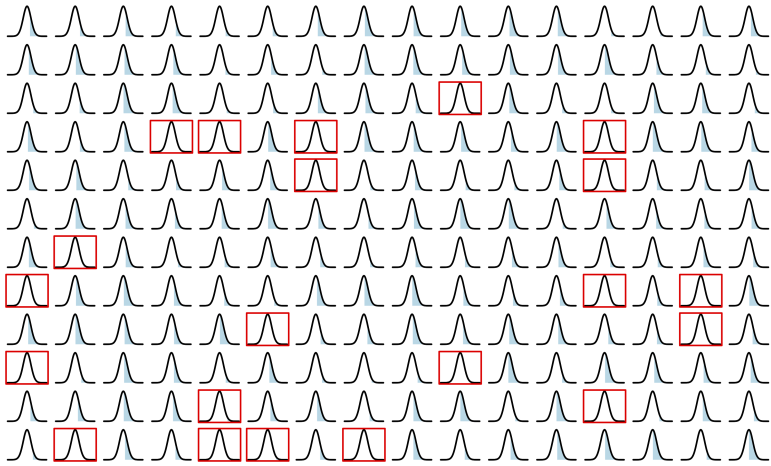
- We have:

$$P(V(t) > 0) \leq \sum_{i \in \mathcal{H}_0} \mathbb{P}_{\mathcal{H}_0}(p_i \leq t) = \sum_{i \in \mathcal{H}_0} t = m_0 t$$

- The Bonferroni procedure at level α rejects all p-values less than $t = \alpha/m$
- Its FWER is

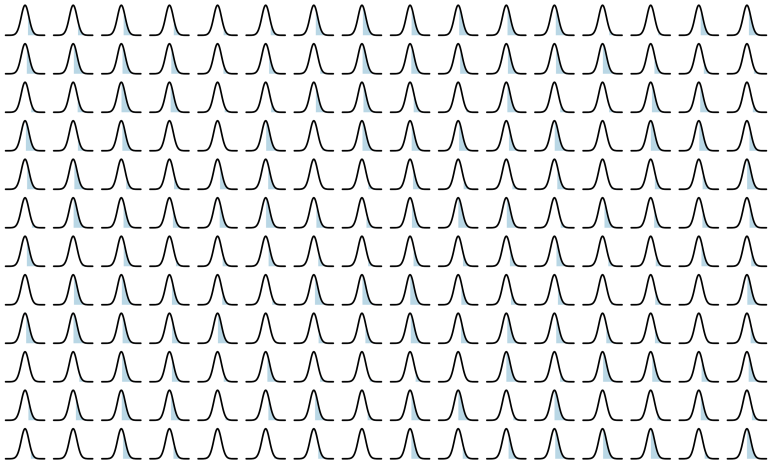
$$\mathbb{P}(V(\alpha/m) > 0) \leq \pi_0 \alpha.$$

Leukemia data set: no multiple testing correction



971 genes called significant genes at (uncorrected) level $\alpha = 0.05$

Leukemia data set: FWER thresholding by Bonferroni



20 genes called significant at FWER level $\alpha = 0.05$

FWER control with the Sidak procedure

- Definition: Reject all i such that $p_i \leq 1 - (1 - \alpha)^{1/m}$
- Properties: FWER control at level $1 - (1 - \alpha)^{\pi_0} \leq \alpha$ under independence
- Sidak is slightly more powerful than Bonferroni, but at the price of a much narrower applicability
- In genomic applications, Bonferroni should be preferred to Sidak

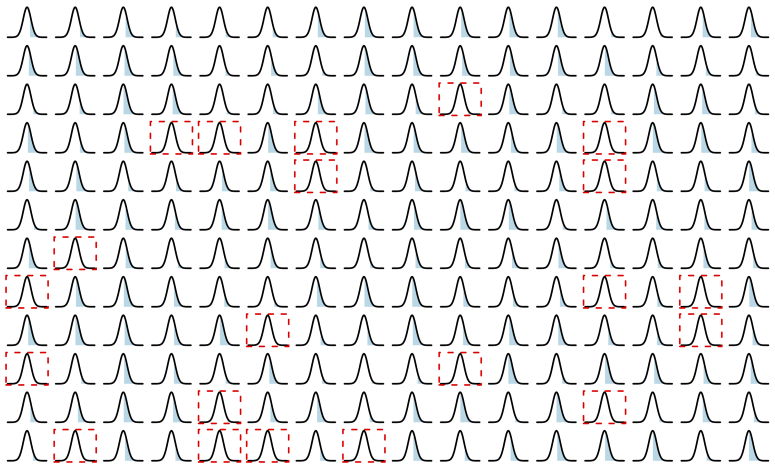
FWER control with the Holm procedure

- Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered p-values.
- Definition: Reject all i such that $\forall j \leq i, p_{(j)} \leq \alpha/(m-j+1)$
- Properties: FWER control at level α under arbitrary dependence
- same guarantees as Bonferroni, at least as powerful:

$$\alpha/(m-j+1) \geq \alpha/m$$

  !
- Holm should be preferred to Bonferroni

Leukemia data set: FWER thresholding by Bonferroni

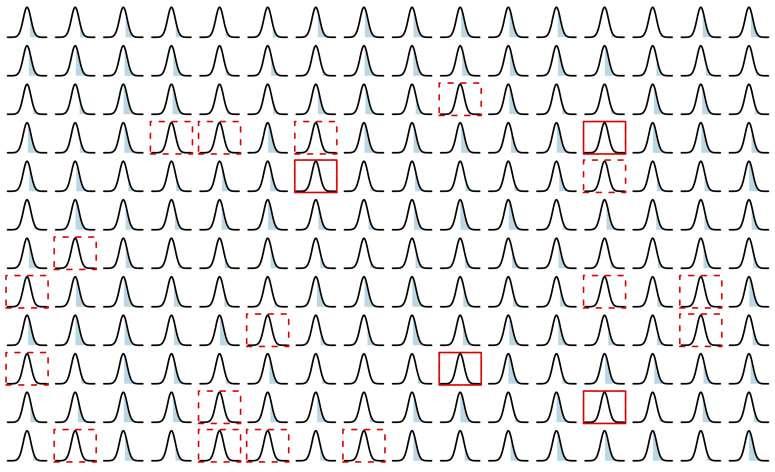


20 genes called significant at FWER level $\alpha = 0.05$

FDR control

- sort p-values: $p_{(1)} \leq \ldots \leq p_{(m)}$
- define $\widehat{l} = \max\left\{k | p_{(k)} \leq \alpha \frac{k}{m}\right\}$
- reject all i such that $p_i \leq p_{(\widehat{l})} (= \alpha \widehat{l}/m)$
- $\pi_0 = |\mathcal{H}_0|/m$: proportion of true null hypotheses
- FDR $= \mathbb{E}\left(\frac{V}{|R| \vee 1}\right)$: expected proportion of false positive among rejections
- BH (α) provides FDR control at level $\pi_0 \alpha$ if the p-values under $\mathcal{H}_0$ are either independent or positively associated
- Improvements in the statistical literature: general dependence: Benjamini and Yekutieli (2001), estimation of π_0, in the hope of a sharper FDR control

Leukemia data set: FDR control by BH



163 genes called significant at FDR level $\alpha = 0.05$

The BH procedure is widely used

- Controlling the false discovery rate: A practical and powerful approach to multiple testing, Y. Benjamini, Y. Hochberg, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol 57(1), pp. 289–300. 1995.
- 6,000 publications in the PubMed database with "False Discovery Rate" in their title or abstract
- 60,000 citations according to scholar.google.com.
- Kaplan, Meier. Nonparametric estimation from incomplete observations: 57,000
- Dempster, Laird, Rubin. Maximum likelihood from incomplete data via the EM algorithm (1977): 56,000
- Cox. Regression and life tables (1975): 50,000
- Bland, Altman. Statistical methods for assessing agreement between two methods of clinical measurement: 43,000
- Tibshirani. Regression shrinkage and selection via the lasso (1996): 30,000

Conclusion

- (large-scale) multiple testing is ubiquitous in biomedical data analysis
- multiple testing risks $\neq$ multiple testing procedures
- FWER and FDR control different risks:
  - FWER for confirmatory analyses
  - FDR for "exploratory" analyses
- Some caveats
  - interpretation of FDR control: FDR is an expectation!
  - applicability conditions (dependence assumptions)
- Related topics not explicitly discussed:
  - scientific reproducibility, hidden multiplicity and selective inference
  - online multiple testing

References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[2] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nat Methods*, 11(7):740–742, Jul 2014.