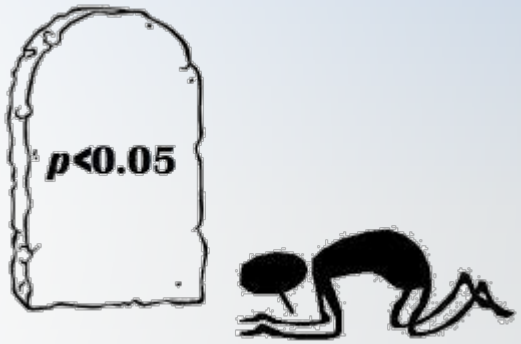
Evidential Metrics and Second-Generation p-values

Jeffrey D. Blume, PhD

Department of Biostatistics

VANDERBILT UNIVERSITY

Outline



- → Evidential Metrics
- → Second-generation p-value
- → High-dimensional example, 7128 Genes
  - → α=0.05 vs α=0.05/7128 vs SG p-value
- → Outrageous claims

Evidential metrics

Example:
Diagnostic Test

1. **Measure of the strength evidence**
   - → Axiomatic and intuitive justification
   - → Summary statistic, yardstick

Positive Test
Negative Test

2. **Propensity to collect data that will yield a misleading #1**
   - → Error rates
   - → Properties of the study design (!)

Sensitivity
Specificity

3. **Probability that an observed #1 is misleading**
   - → False Discovery rate, False Confirmation rate
   - → Chance that an observed result is mistaken
   - → Properties of the observed data (!)

PPV
NPV

Likelihood

Law of Likelihood: The hypothesis that does a better job at predicting the observed events is better supported by the data.

Evidential Metric	What it measures	Likelihood
1	strength of the evidence	Likelihood Ratio (LR)
2	propensity for study to yield misleading evidence	$P(LR > k \mid H_0)$ $P(LR < 1/k \mid H_1)$
3	propensity for observed results to be misleading	$P(H_0 \mid LR = k)$ $P(H_1 \mid LR = 1/k)$

The p-value (what it is)

- Number between 0 and 1
- Smaller ⇒ support for an alternative hypothesis
- Larger ⇒ data are inconclusive
- Clinical significance is ignored
- Sample size confounds comparisons
- Interpretation
  - awkward
  - assumes null hypothesis true
  - rooted in inductive reasoning
- Not clear if/when ‘adjustments’ are necessary

The ^*p*-value (what ~~it is~~)

2nd-generation

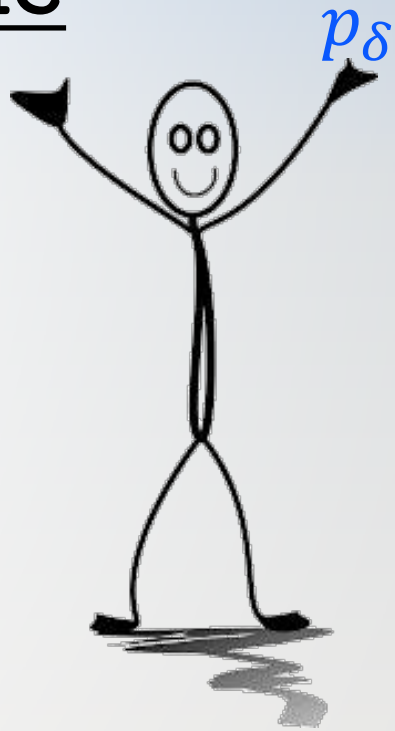
we want

Version 2.0

- ✔ Number between 0 and 1 → near 0 supports alt; near 1 supports null; near ½ inconclusive
- ✔ Smaller ⇒ support for an alternative hypothesis
- Larger ⇒ data ~~are inconclusive~~ support null
- Clinical significance is ~~ignored~~ incorporated
- ✗ ~~Sample size confounds comparisons~~
- Interpretation → Fraction of data-supported hypotheses that are null
  - ~~awkward~~ straightforward
  - assumes ~~null hypothesis true~~ conditions on observed data
  - ~~rooted in inductive reasoning~~ descriptive, summarizes
- ~~Not~~ clear if/when 'adjustments' are necessary

Ideally, never

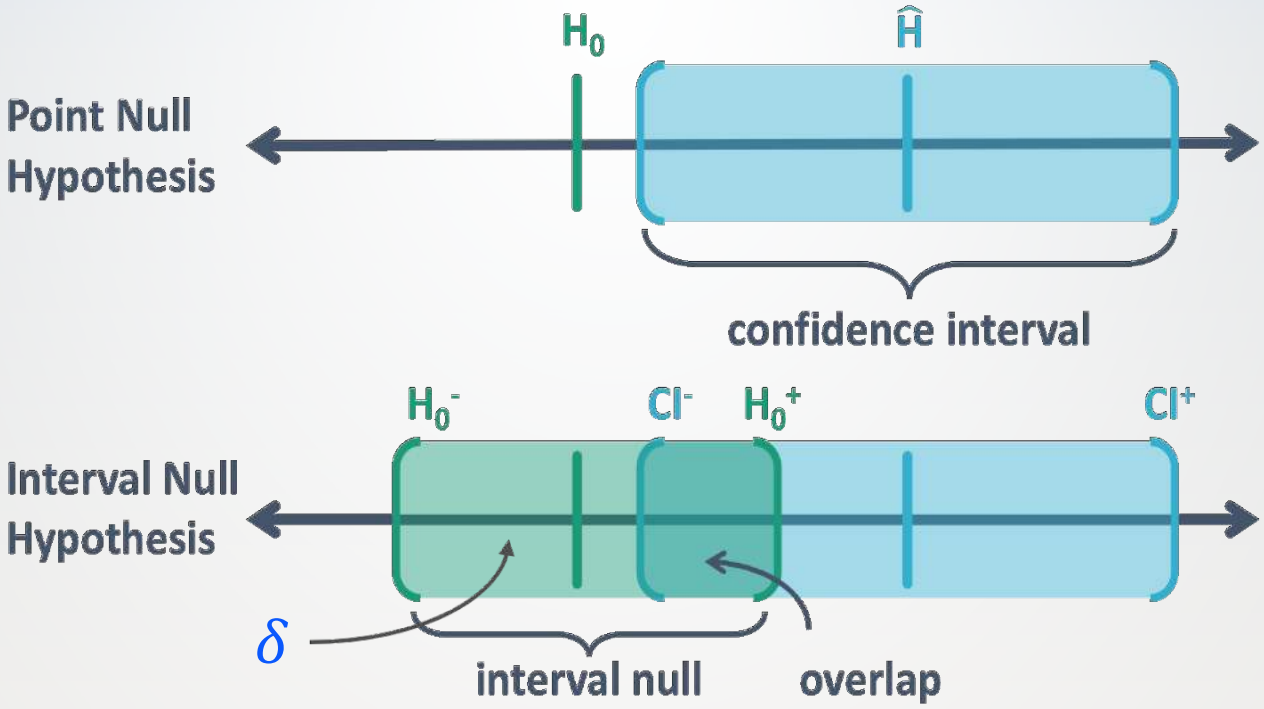
Second-generation p-value



- SGPV is in [0,1] and denoted by p_δ
- δ for scientific significance
  1. $p_\delta = 0 \Rightarrow$ null incompatible with data
  2. $p_\delta = 1 \Rightarrow$ null compatible with data
  3. $0 < p_\delta < 1 \Rightarrow$ data are inconclusive

- Fraction of data-supported hypotheses that are null
- Retains strict error control, all rates → 0

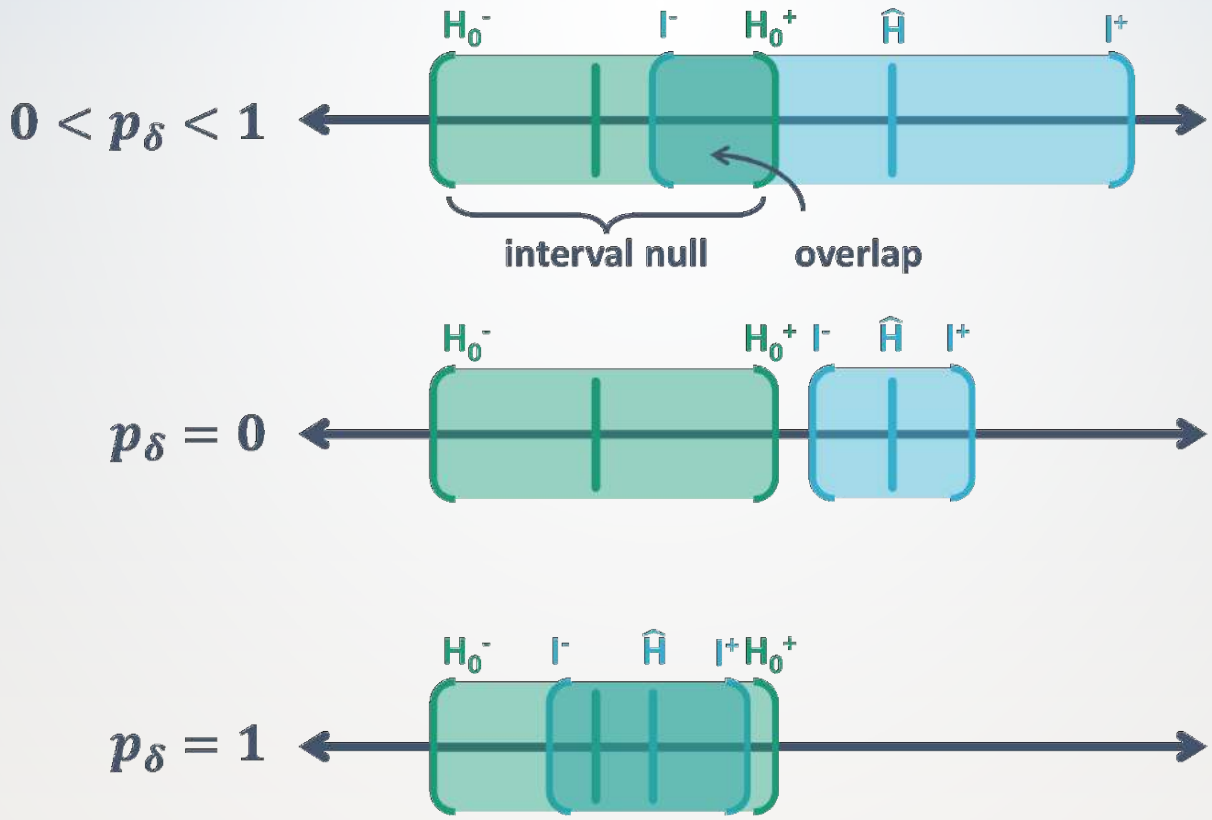
Illustration



Point null hypothesis H_0 and interval null hypothesis $[H_0^-, H_0^+]$

Data-supported hypothesis $\hat{H}$ and confidence interval $[CI^-, CI^+]$

Illustration



Works with confidence, credible, and support intervals

Definition

Second-generation *p*-value (SGPV)

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2|H_0|}, 1\right\}$$

Proportion of data-supported hypotheses that are also null hypotheses

Small-sample correction factor

shrinks proportion to ½ when $|I|$ wide

when $|I| > 2|H_0|$

Statistical Properties

Suppose interval I has coverage probability 1-α, then

Three 'Error' Rates

1. $P(p_\delta = 0 | H_0) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
2. $P(p_\delta = 1 | H_1) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
3. $P(0 < p_\delta < 1 | H)$ controlled through sample size

Two False Discovery Rates

1. $P(H_0 \mid p_\delta = 0)$
2. $P(H_1 \mid p_\delta = 1)$

Leukemia gene expression

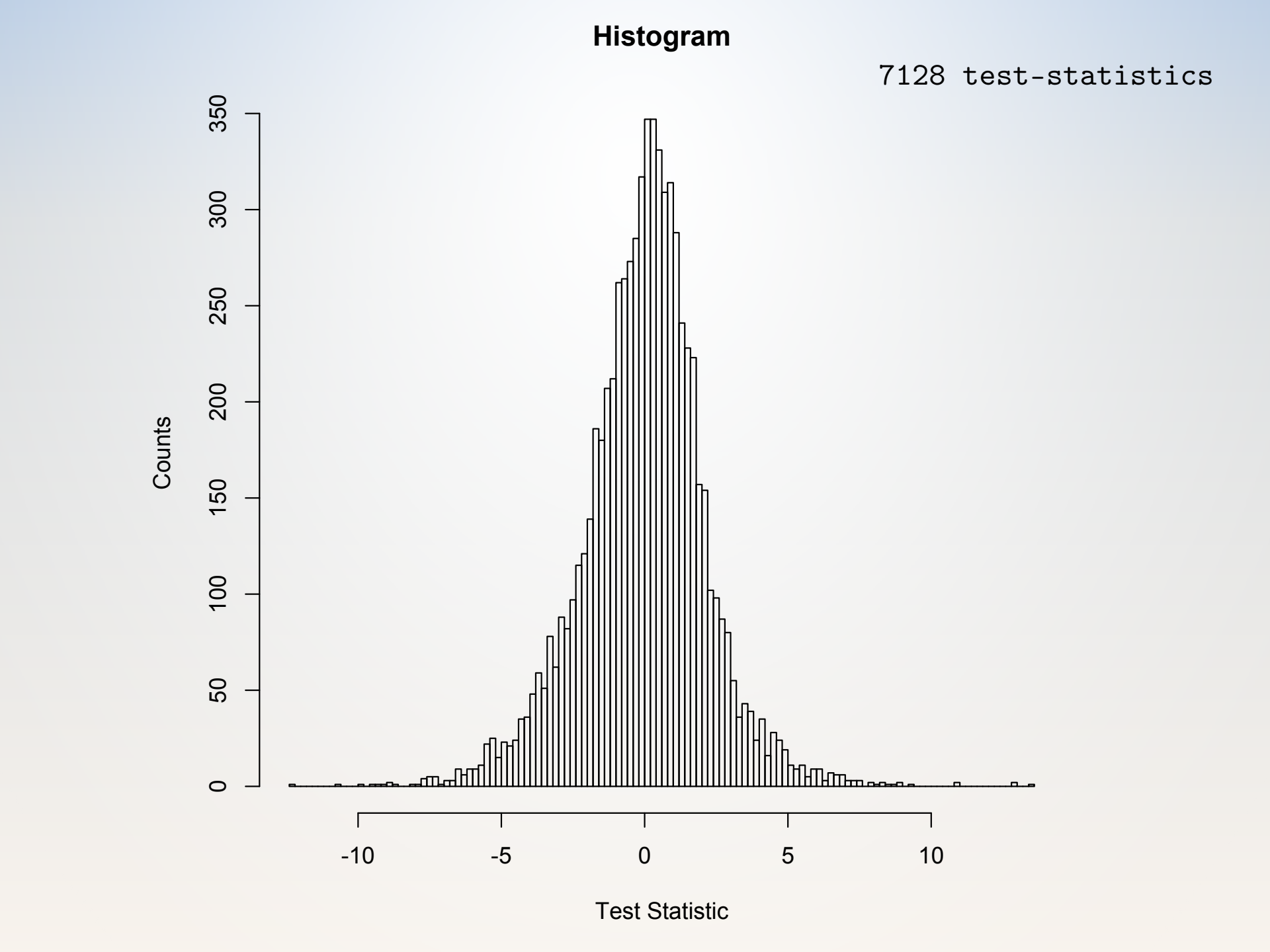
→ Classifying acute leukemia by precursors
(Golub 1999, *Science*)

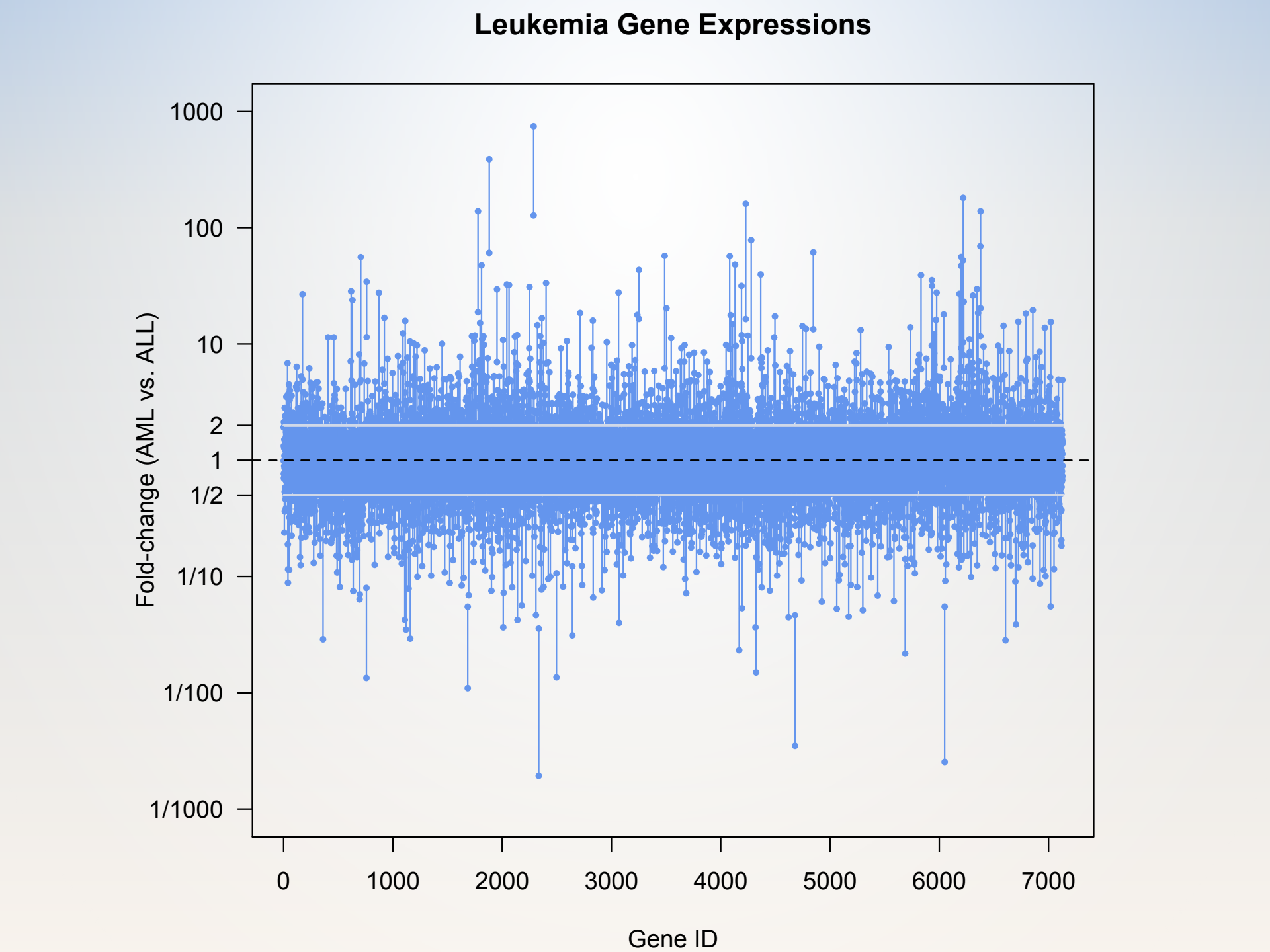
→ 7128 genes ; 72 patients (47 ALL and 25 AML)

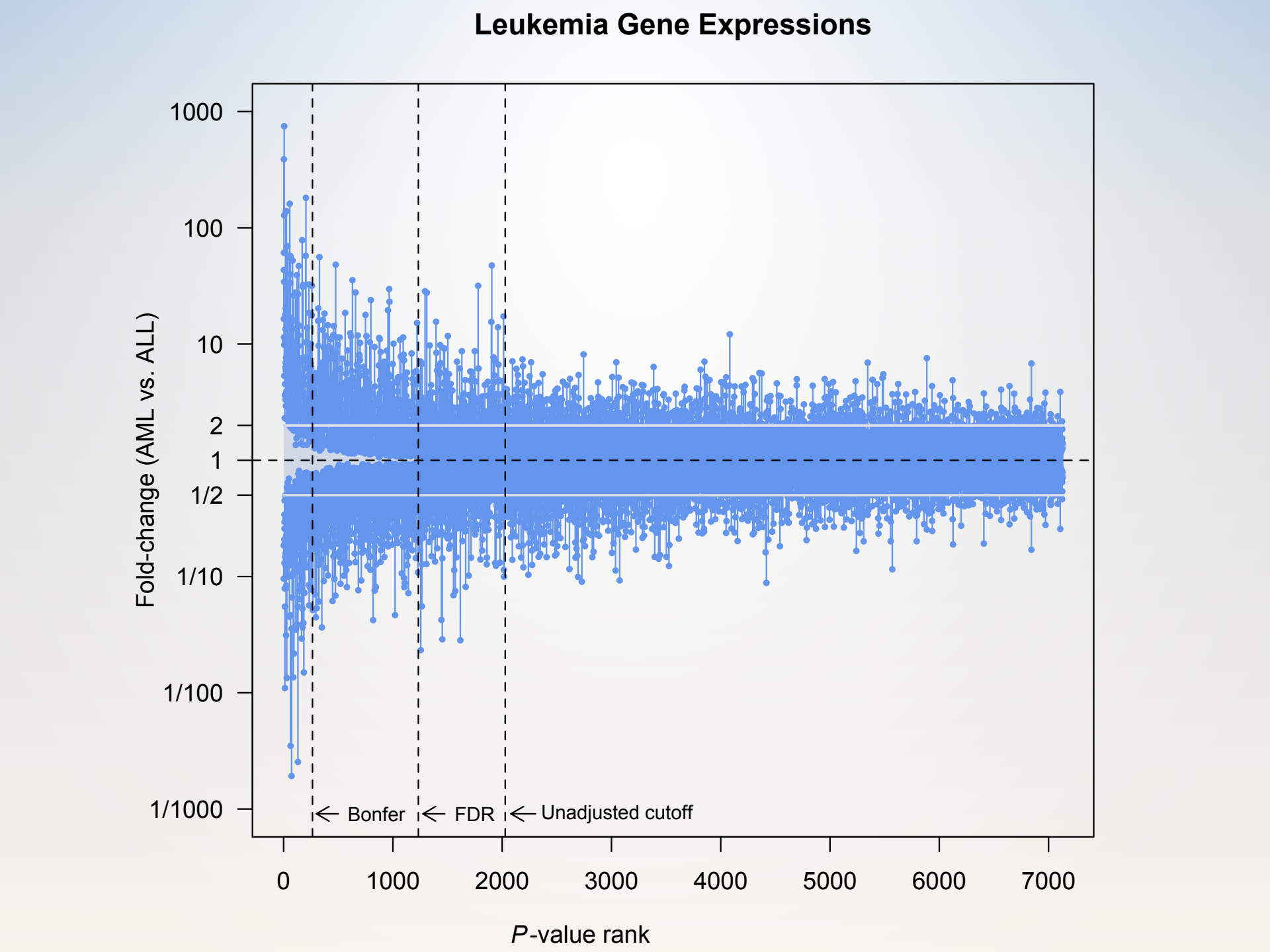
→ Affymetrix chip collected expression levels

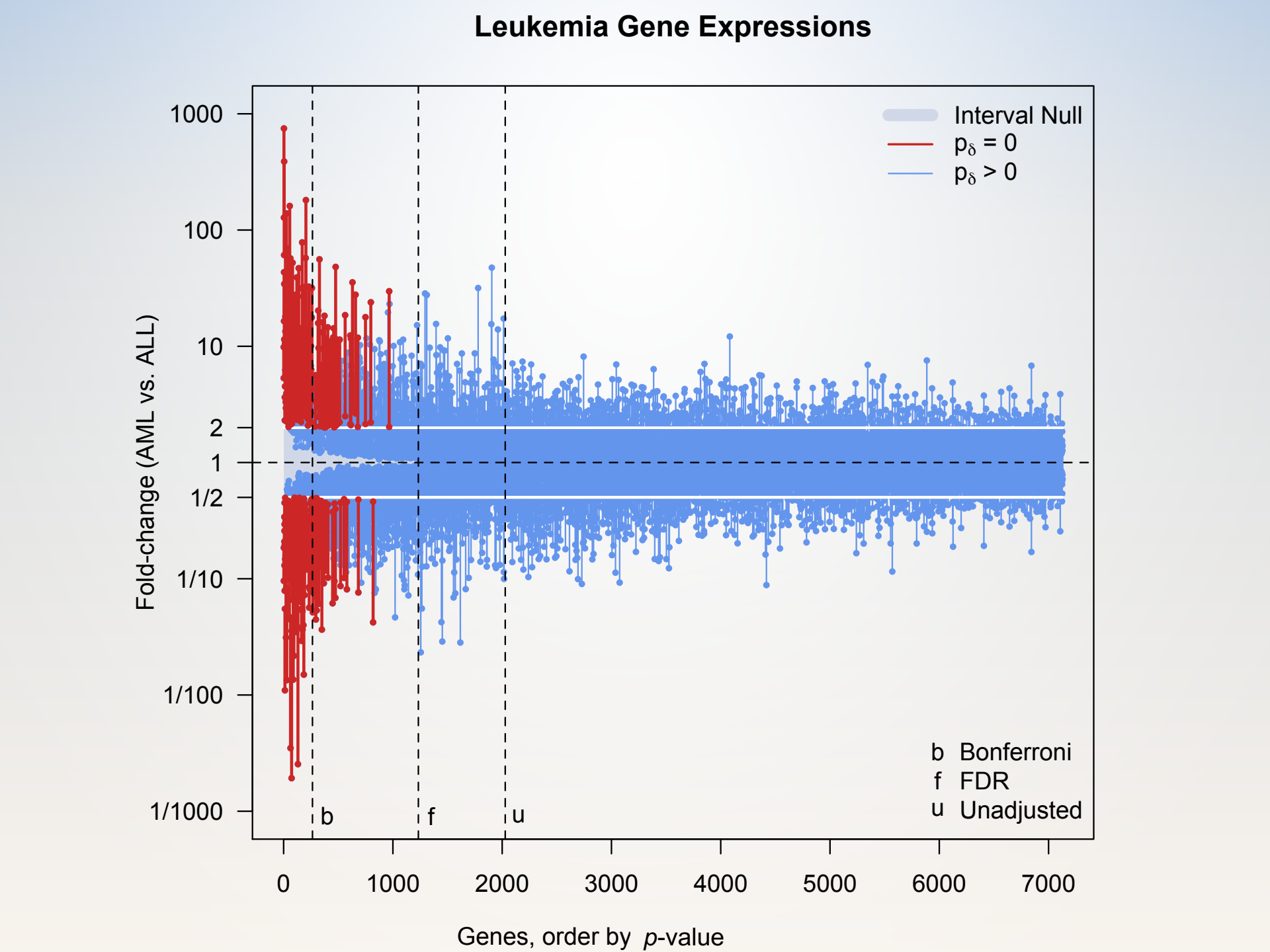
→ Goal: Identify `interesting' genes whose
expression levels differ between
All and AML subjects.

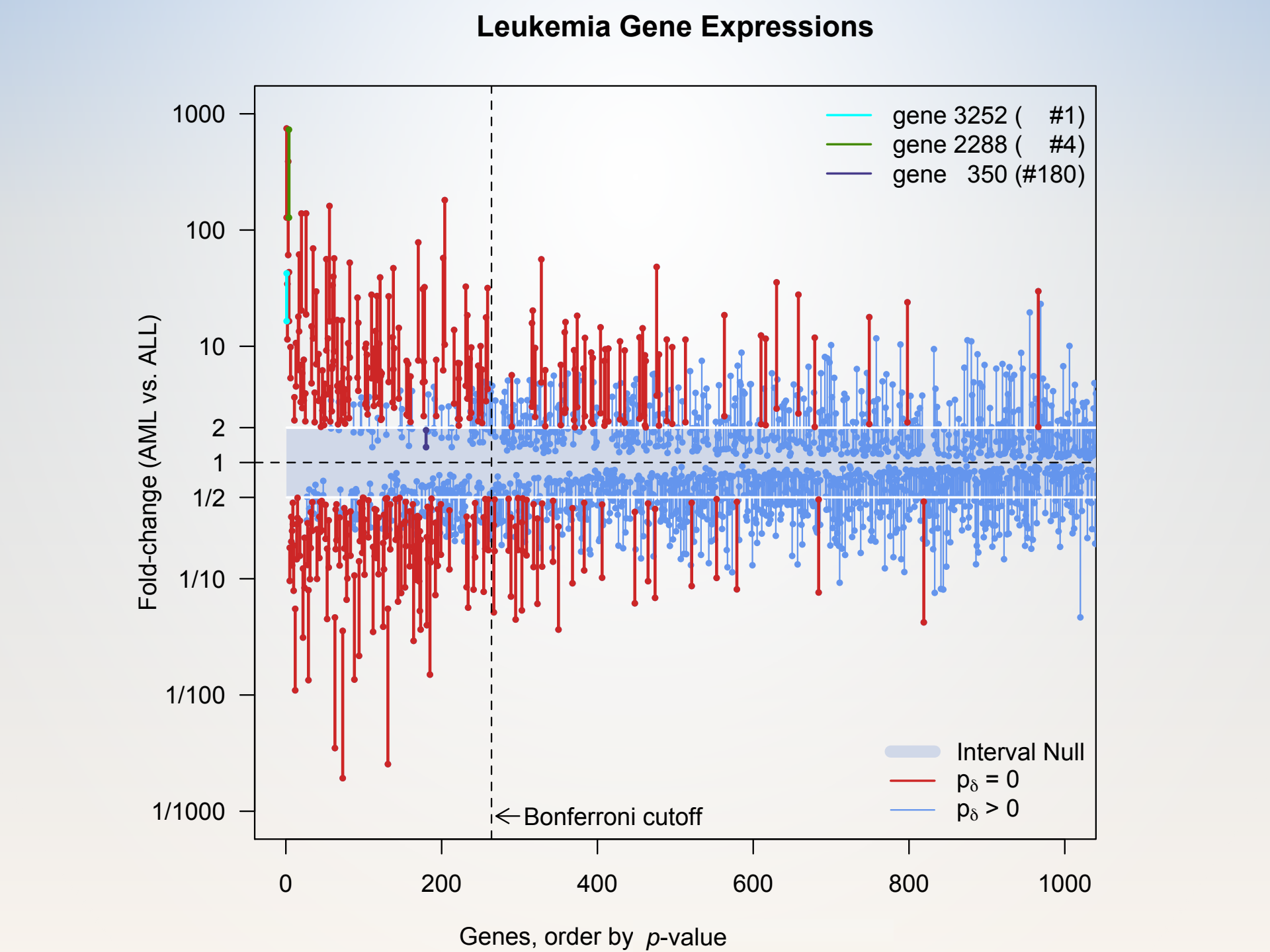
→ Looking for fold changes of 2 or more


Histogram
7128 test-statistics
Counts
0
50
100
150
200
250
300
350
-10
-5
0
5
10
Test Statistic


Leukemia Gene Expressions
1000
100
10
2
1
1/2
1/10
1/100
1/1000
Fold-change (AML vs. ALL)
0
1000
2000
3000
4000
5000
6000
7000
Gene ID


Leukemia Gene Expressions
1000
100
10
2
1
1/2
1/10
1/100
1/1000
Fold-change (AML vs. ALL)
← Bonfer
← FDR
← Unadjusted cutoff
0
1000
2000
3000
4000
5000
6000
7000
P-value rank


Leukemia Gene Expressions
Interval Null
$p_\delta = 0$
$p_\delta > 0$
1000
100
10
2
1
1/2
1/10
1/100
1/1000
Fold-change (AML vs. ALL)
b
f
u
b Bonferroni
f FDR
u Unadjusted
0
1000
2000
3000
4000
5000
6000
7000
Genes, order by *p*-value


Leukemia Gene Expressions
gene 3252 (#1)
gene 2288 (#4)
gene 350 (#180)
1000
100
10
2
1
1/2
1/10
1/100
1/1000
Fold-change (AML vs. ALL)
Interval Null
$p_\delta = 0$
$p_\delta > 0$
←Bonferroni cutoff
0
200
400
600
800
1000
Genes, order by p-value

Leukemia study findings

- Findings: Bonferroni 264, SVPG 229
  - Agree on 164 findings
  - Bonferroni +100, SGPV +65
- Effective Type I error rate: 0.037 vs. 0.032
- FDR of 2.45% captures all $p_\delta = 0$, 737 others
- Moving cutoff trades Type I for Type II errors
- SGPV changes the *ranking* of findings
  - Three categories now: null, alt, inconclusive
  - Null findings not illustrated here

Some SGPV findings have a priori published validation

False discovery rates

- Impact of α=0.05 vs α=0.05/7128
- False Discovery Rate (FDR)

$\mathrm{LR}(H_1, H_0 \mid p<\alpha)$

$$P(H_0|p<\alpha) = \left[1 + \frac{(1-\beta)}{\alpha} r\right]^{-1}$$

- False Confirmation Rate (FCR)

$$P(H_1|p>\alpha) = \left[1 + \frac{(1-\alpha)}{\beta} \frac{1}{r}\right]^{-1}$$

$\mathrm{LR}(H_0, H_1 \mid p>\alpha)$

$r = P(H_1)/P(H_0)$

False discovery rates

- Second-generation p-values
- False Discovery Rate (FDR)

$$P(H_0|p_\delta = 0) = \left[1 + \frac{P(p_\delta = 0|H_1)}{P(p_\delta = 0|H_0)} r\right]^{-1}$$

$\mathrm{LR}(H_1, H_0 | p_\delta = 0)$

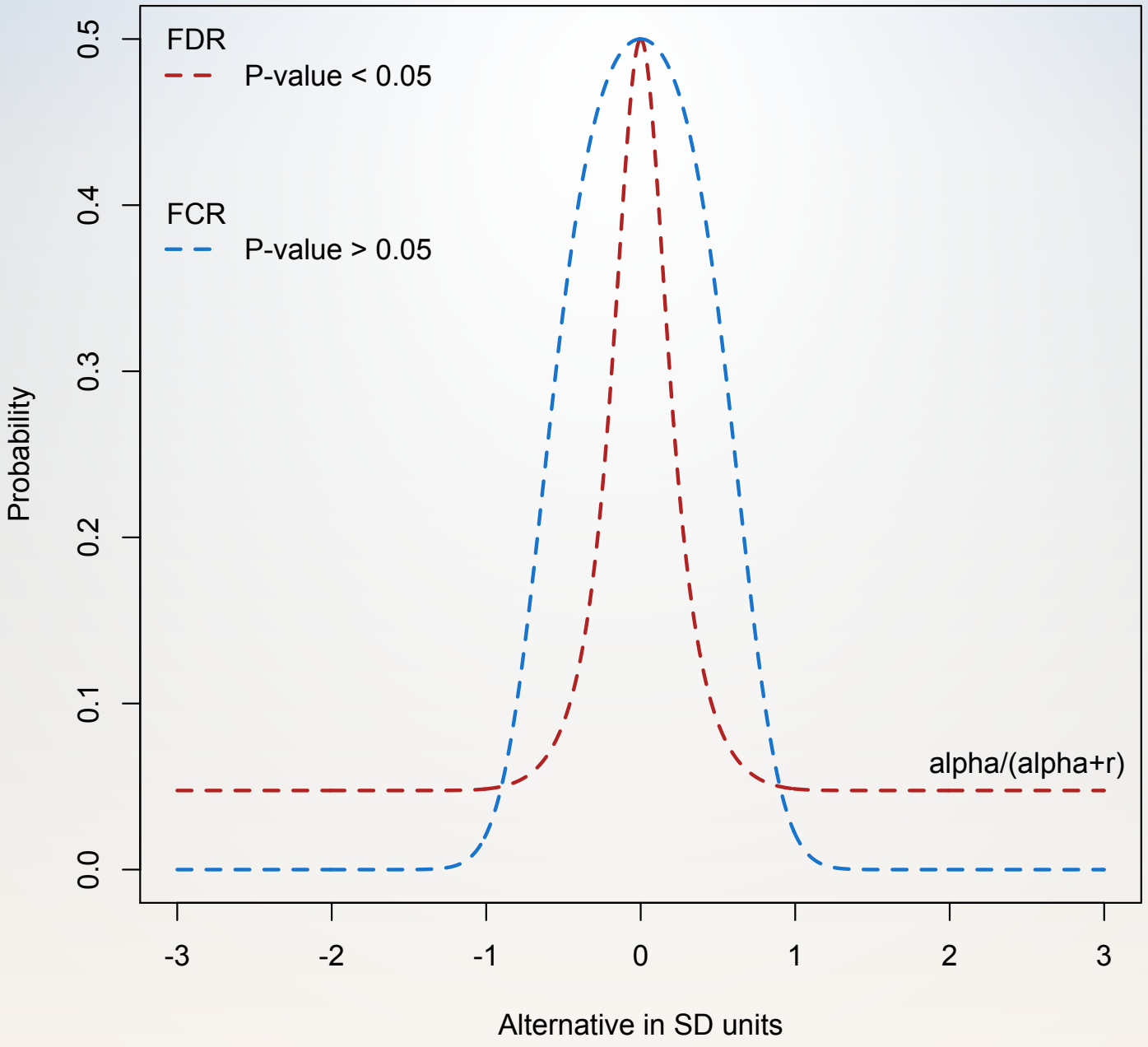
- False Confirmation Rate (FCR)

$$P(H_1|p_\delta = 1) = \left[1 + \frac{P(p_\delta = 1|H_0)}{P(p_\delta = 1|H_1)} \frac{1}{r}\right]^{-1}$$

$\mathrm{LR}(H_0, H_1 | p_\delta = 1)$

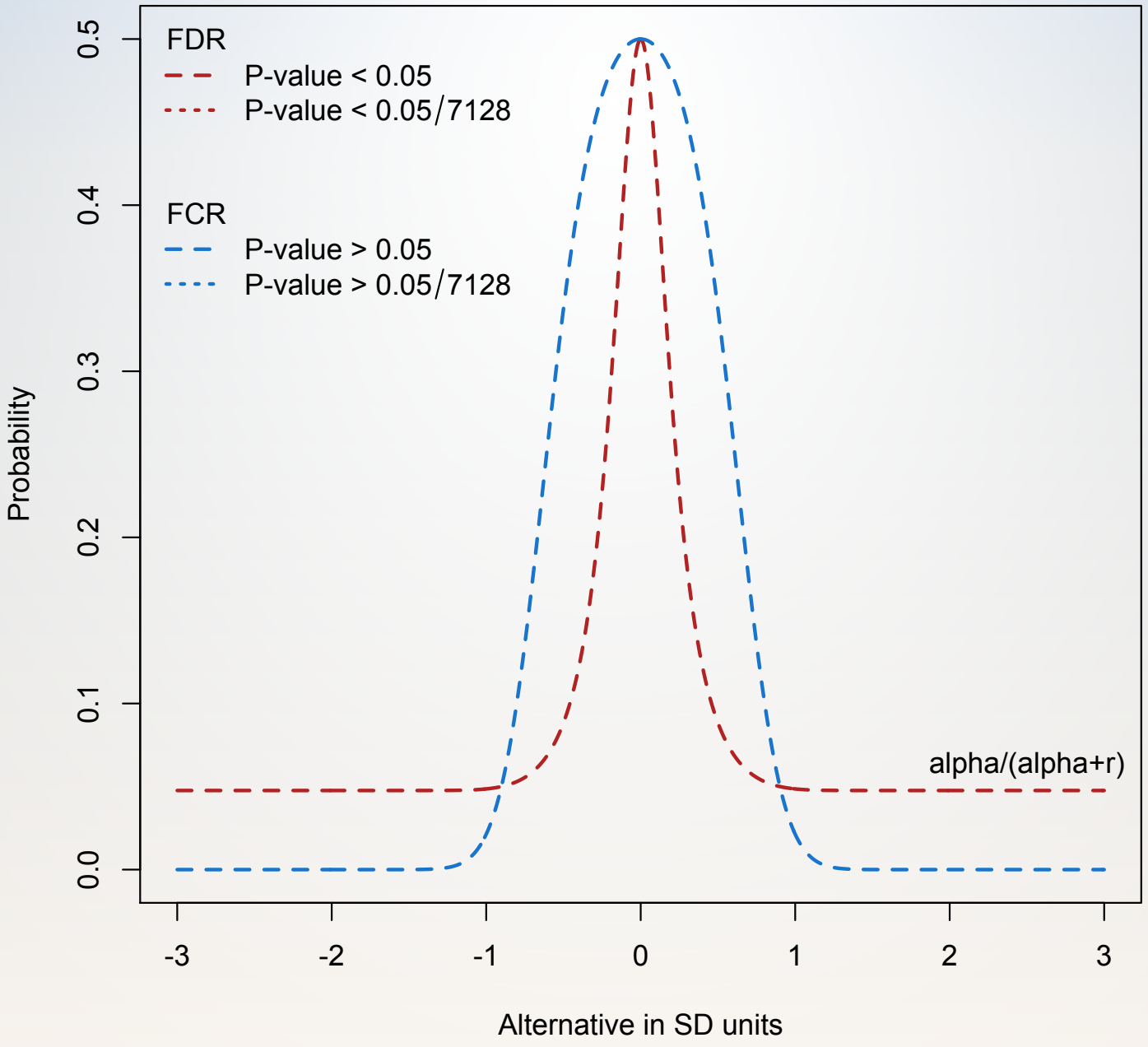
$r = P(H_1)/P(H_0)$

False discovery and confirmation rates



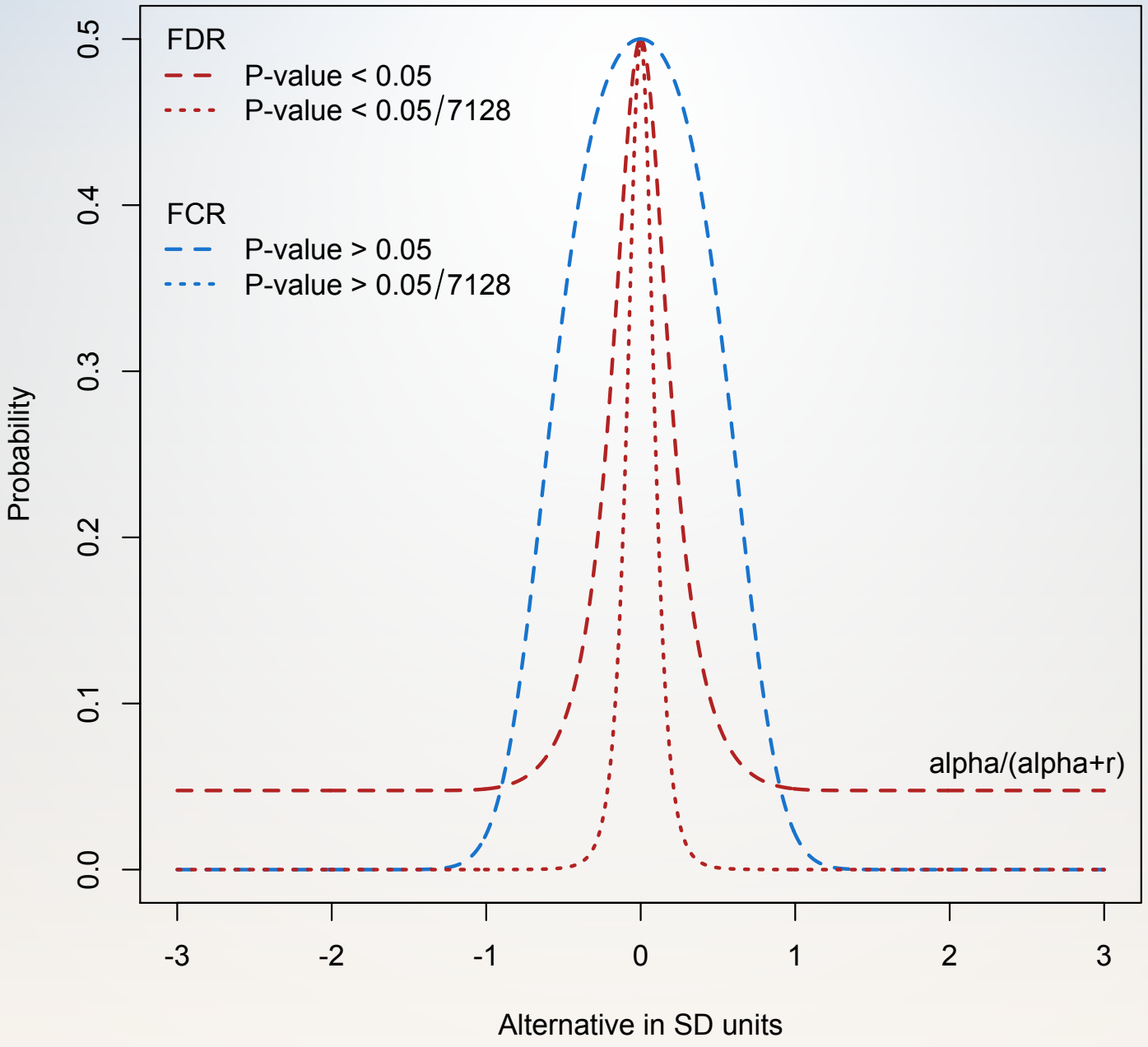
r=1

False discovery and confirmation rates



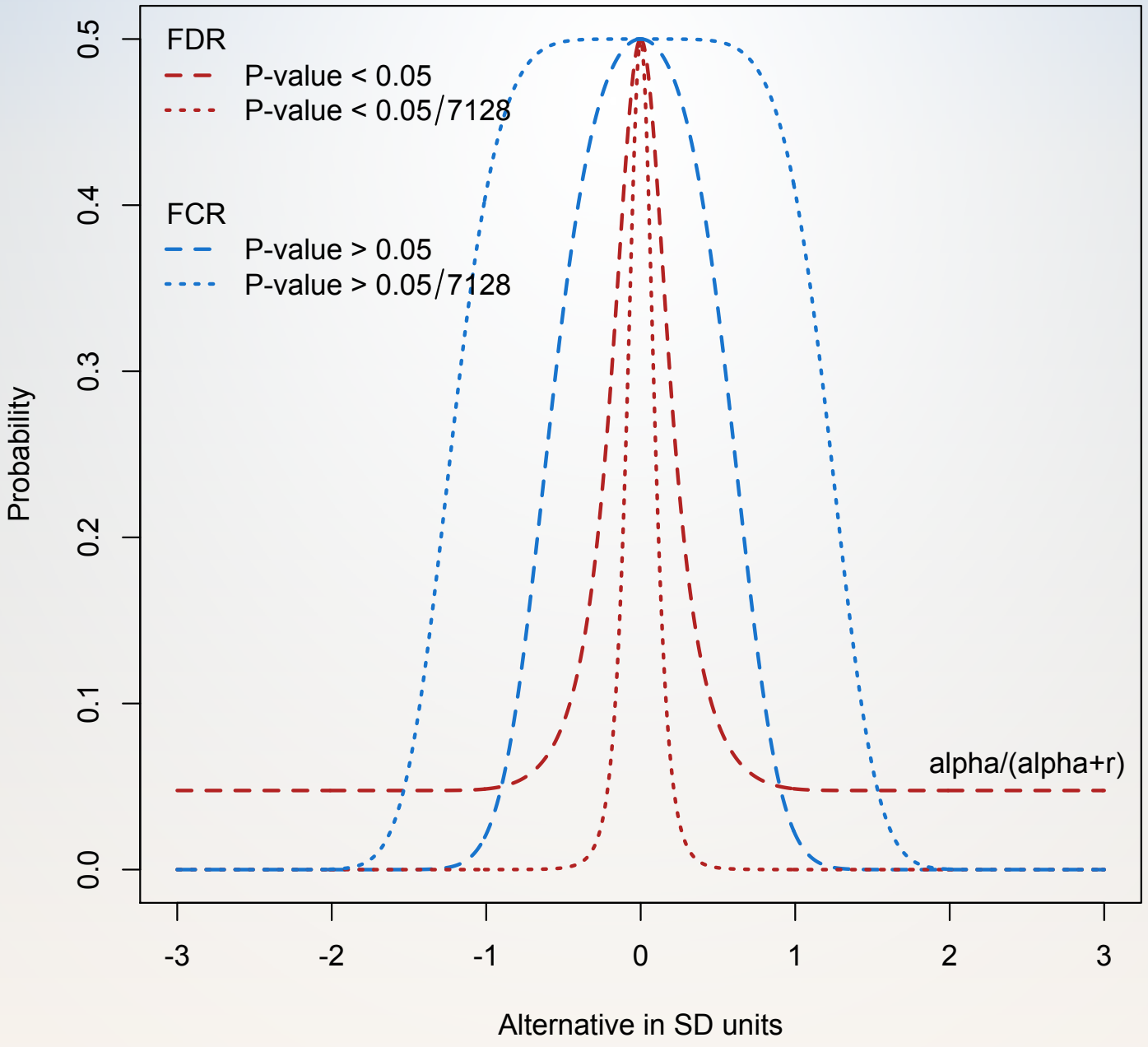
r=1

False discovery and confirmation rates



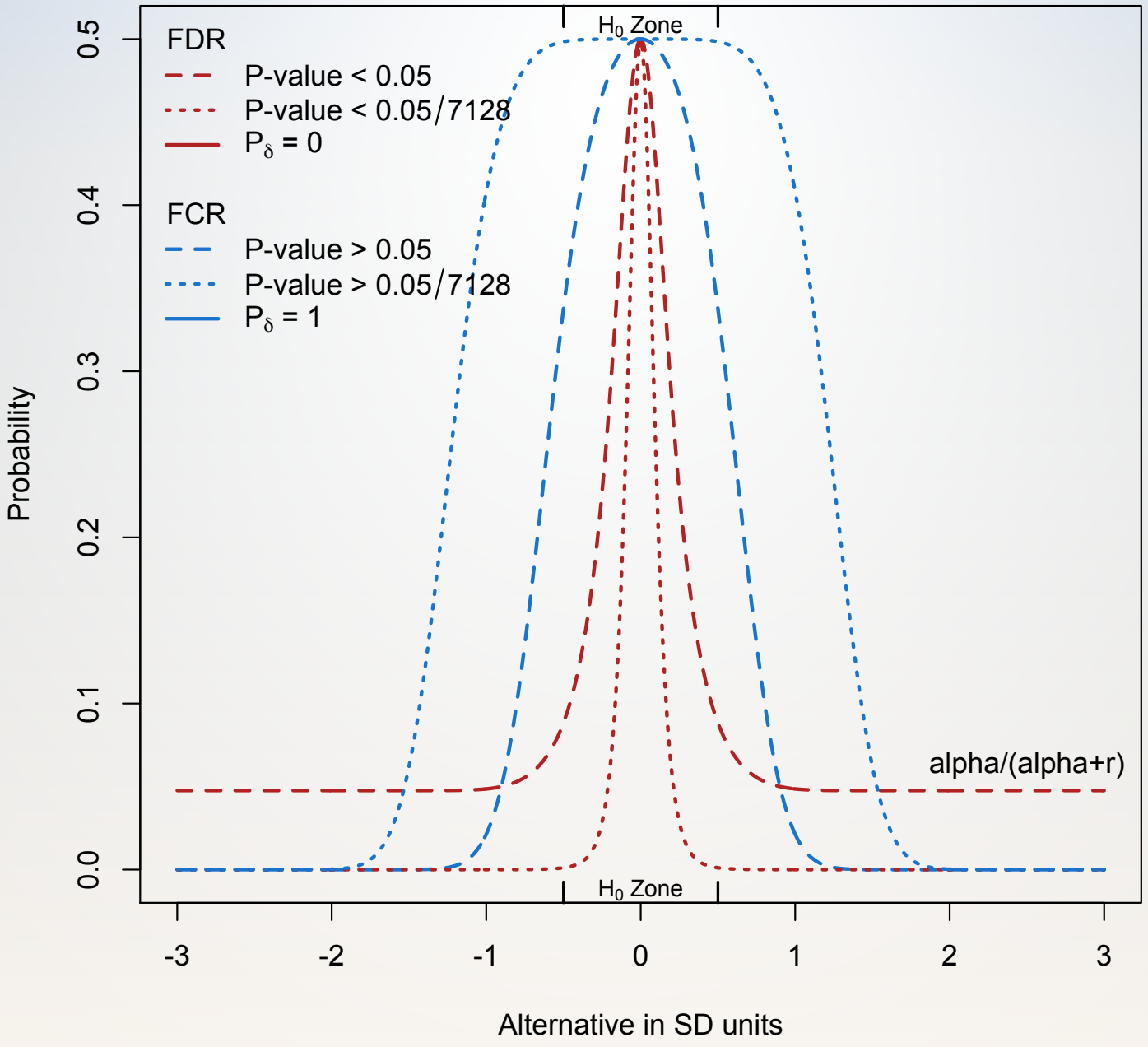
r=1

False discovery and confirmation rates



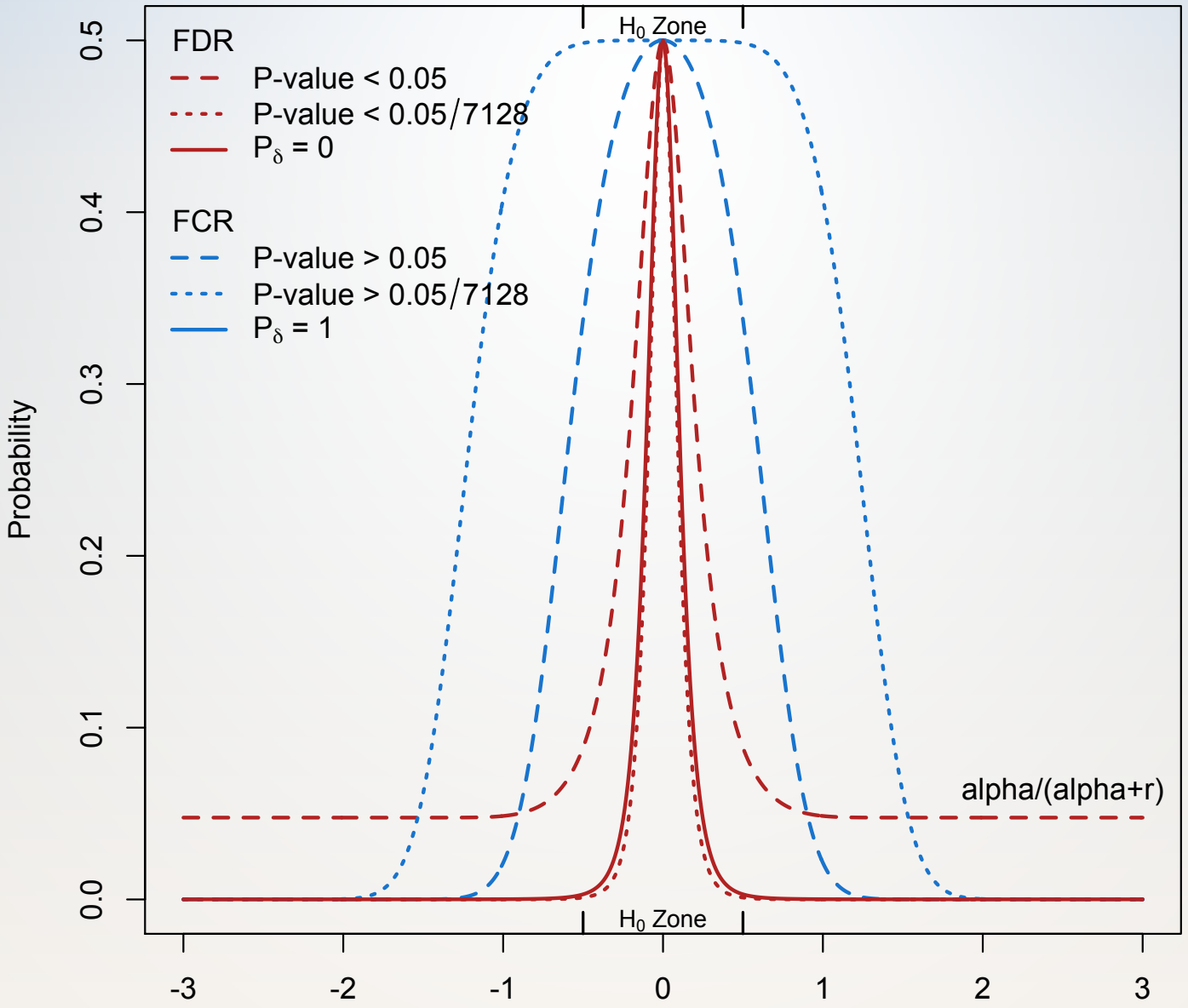
r=1

False discovery and confirmation rates



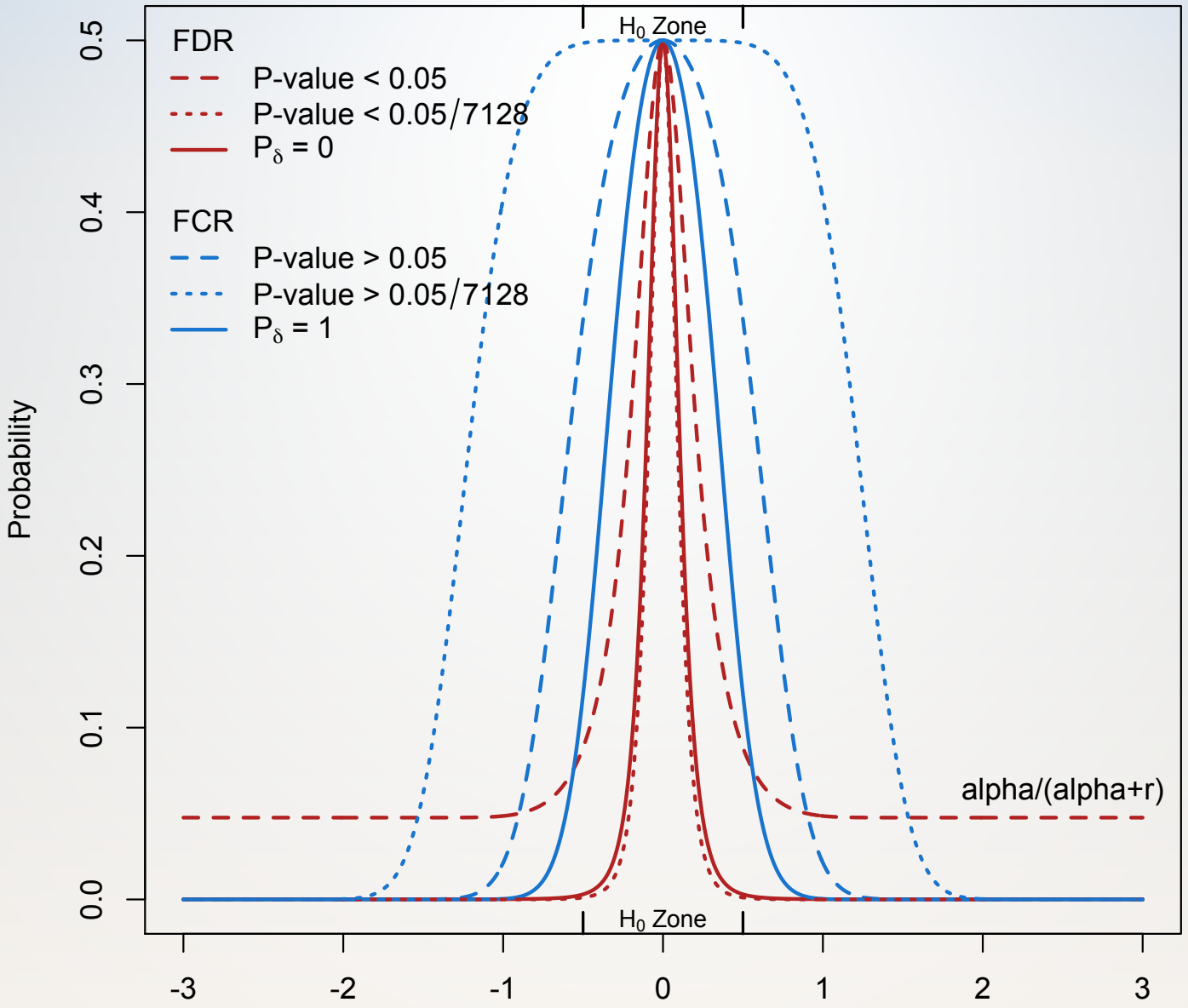
r=1

False discovery and confirmation rates



r=1

False discovery and confirmation rates



r=1

Acknowledgements

- Collaborators
  - William D. Dupont
  - Robert A. Greevy
  - Lucy D'Agostino McGowan

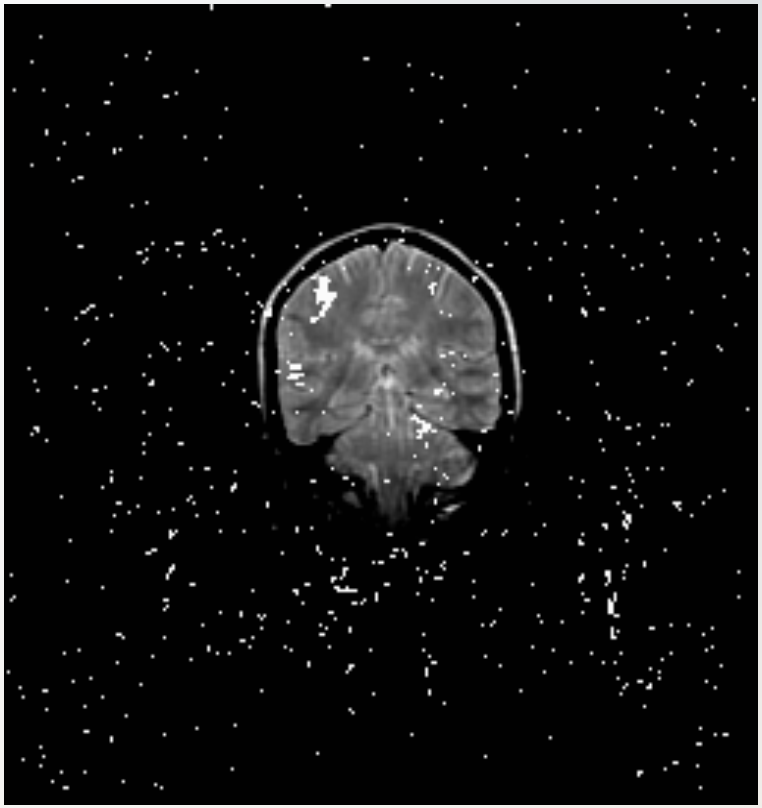
- Website / Paper
  - statisticalevidence.com
  - ArXiv.org ID 1709:09333
  - Google "Second-Generation p-value"

Outrageous Claim (!?)

The SGPV achieves the inferential properties that many scientists hope, or believe, are attributes of the classic p-value.

Setting interval null

- Before analyzing data (!)
- Measurement error
- Subject matter knowledge
- Impact of findings
- Community standard
- Get creative (fMR example)
- Width not critical, buffer



Testing

Evidential Metric	What it measures	Hypothesis Testing	Significance Testing
1	strength of the evidence	Absent	**Tail-area probability** (*p*-value)
2	propensity for study to yield misleading evidence	**Tail-area probability** (error rates)	Absent
3	propensity for observed results to be misleading	misinterpret #2	misinterpret #1

- Using one mathematical concept (tail-area probability) used to measure three distinct metrics creates paradoxes