

An Introduction to Second-Generation *p*-values & Their Application in Large- Scale Genetic Analyses

Jeffrey D. Blume, PhD

Departments of Biostatistics,
Biomedical Informatics, Biochemistry
Data Science Institute

Future work from
Valerie Welty,
PhD candidate in
Biostatistics

VANDERBILT  UNIVERSITY

Outline

- *p*-value overview
- Second-generation *p*-value
- Introductory examples (2)
- Leukemia gene expression, 7128 Genes
 - Affymetrix chip, 72 patients (47 ALL and 25 AML)
- High-dimensional example
 - Prostate Cancer SNP data (~247,000)
 - 3,894 participants: 2,511 cases & 1,383 controls
- Evidential metrics and remarks



Synopsis

- Classical p -values are
 - Ubiquitous, Sacrosanct, Imperfect, Misused
 - Misunderstood (Significance vs. Hypothesis testing roles)
 - Openly debated in practice and theory
- Trend toward estimation in reporting of results
 - Report a estimation interval (e.g. confidence interval)
 - Does interval contain only clinically significant values?
- Second-generation p -values (SGPVs)
 - Embody and formalize this trend
 - Maintain and improve error rate control
 - Define clinically significant before looking at the data

Second-generation p-values

Blume, June 2020

The p -value (what it is)

- Number between 0 and 1
- Smaller \Rightarrow support for an alternative hypothesis
- Larger \Rightarrow 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

Second-generation p-values

Blume, June 2020

2nd-generation
we want

The \hat{p} -value (what it is)

Version 2.0

- ✓ Number between 0 and 1 →
 - near 0 supports alt
 - near 1 supports null
 - near $\frac{1}{2}$ 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-values Blume, June 2020

Illustration

The diagram illustrates the relationship between different types of null and supported hypotheses:

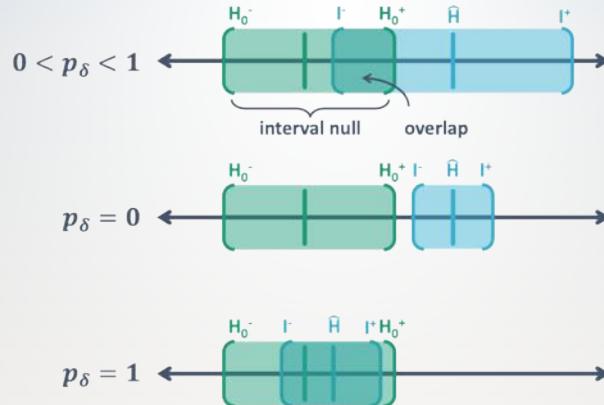
- Point Null Hypothesis:** A single vertical line labeled H_0 on a horizontal axis.
- Interval Null Hypothesis:** A double-headed arrow spanning from H_0^- to H_0^+ , labeled "interval null".
- Data-supported Hypothesis:** A blue shaded interval labeled \hat{H} .
- Confidence Interval:** A blue shaded interval labeled $[CI^-, CI^+]$.
- Overlap:** The region where the interval null and the confidence interval overlap.

Annotations below the diagram:

Point null hypothesis H_0 and interval null hypothesis $[H_0^-, H_0^+]$
 Data-supported hypothesis \hat{H} and confidence interval $[CI^-, CI^+]$

Second-generation p-values Blume, June 2020

Illustration



Works with confidence, credible, and support intervals

Second-generation p-values

Blume, June 2020

Definition

Second-generation
p-value (SGPV)

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

Proportion of data-supported hypotheses that are also null hypotheses

Small-sample correction factor

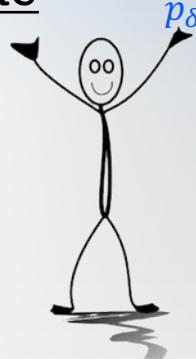
shrinks proportion to $\frac{1}{2}$ when $|I|$ wide

when $|I| > 2|H_0|$

Second-generation p-values

Blume, June 2020

Second-generation p -value



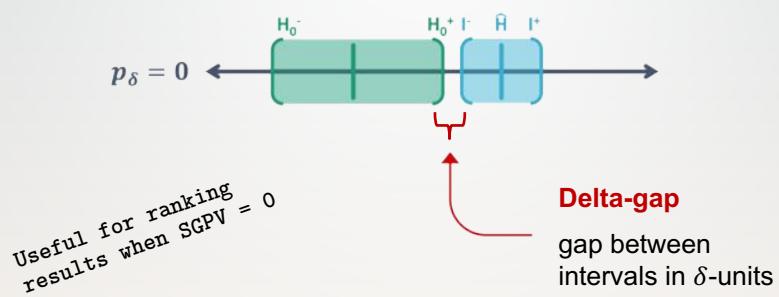
- SGpv is in $[0,1]$ and denoted by p_δ
- δ for scientific significance
 1. $p_\delta = 0 \Rightarrow$ null **incompatible** with data
 2. $p_\delta = 1 \Rightarrow$ null **compatible** with data
 3. $0 < p_\delta < 1 \Rightarrow$ data are **inconclusive**
- Fraction of data-supported hypotheses that are null
- Retains strict error control, all rates $\rightarrow 0$

Second-generation p -values

Blume, June 2020

The Delta-gap

When SGpv=0, there is a gap between the intervals. The length of that gap, in δ -units is the **delta-gap**.

Second-generation p -values

Blume, June 2020

Steps

- SGPV ~ the fraction of data-supported hypotheses that are null or practically null
1. Specify an the interval null hypothesis or a point null with indifference zone
 2. Find confidence, support or credible interval
 3. Measure the fraction of interval (#2) that is in the null zone
 4. Apply small-sample correction factor, as necessary

Second-generation p-values

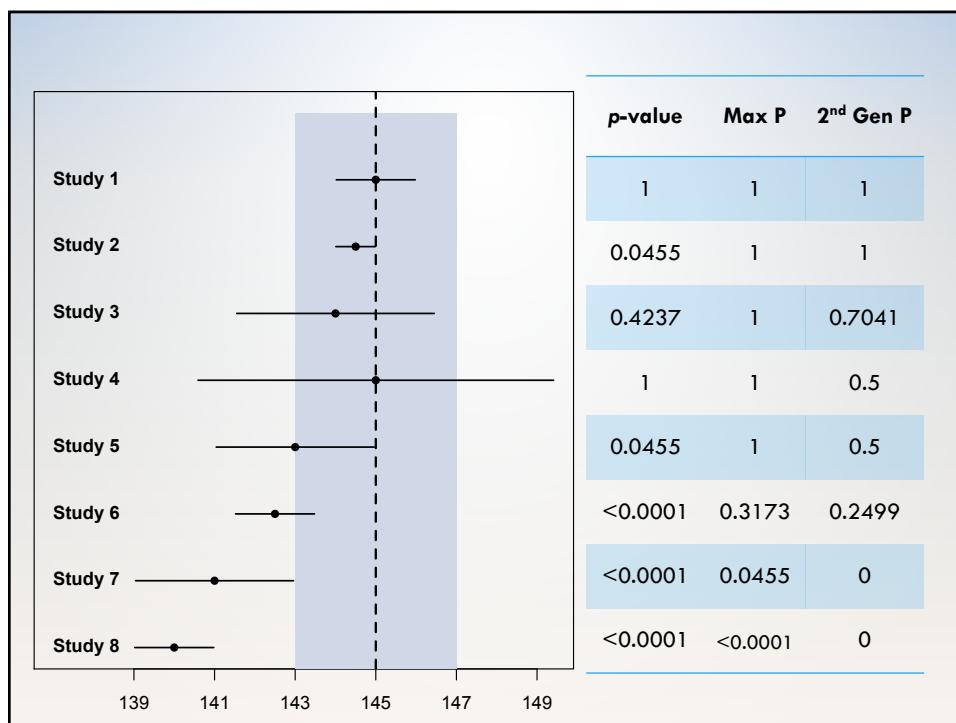
Blume, June 2020

Systolic Blood Pressure

- SBP is reported to the nearest 2 mmHg
- Null Hypothesis: mean SPB is 145 mmHg
- Interval Null hypothesis: mean is 143 to 147 mmHg
- Results from 8 mock studies

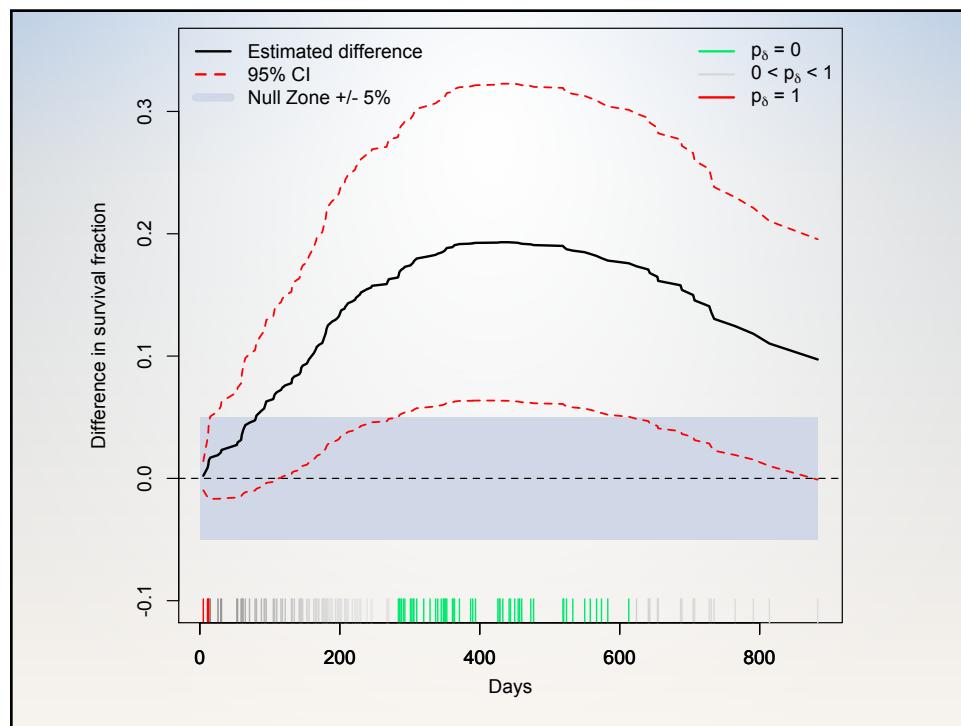
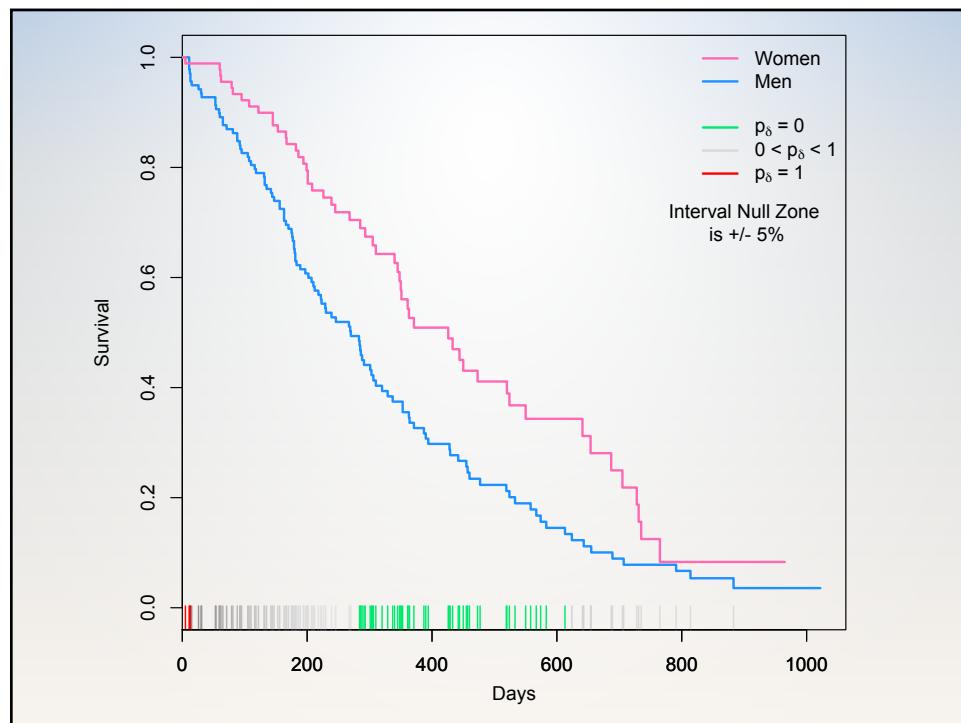
Second-generation p-values

Blume, June 2020



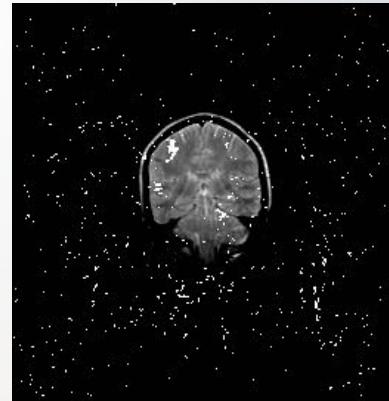
Lung Cancer Survival

- Survival time in patients with advanced lung cancer (days)
- Potential for gender dissimilarities
- Trial by North Central Cancer Treatment Group (1994)



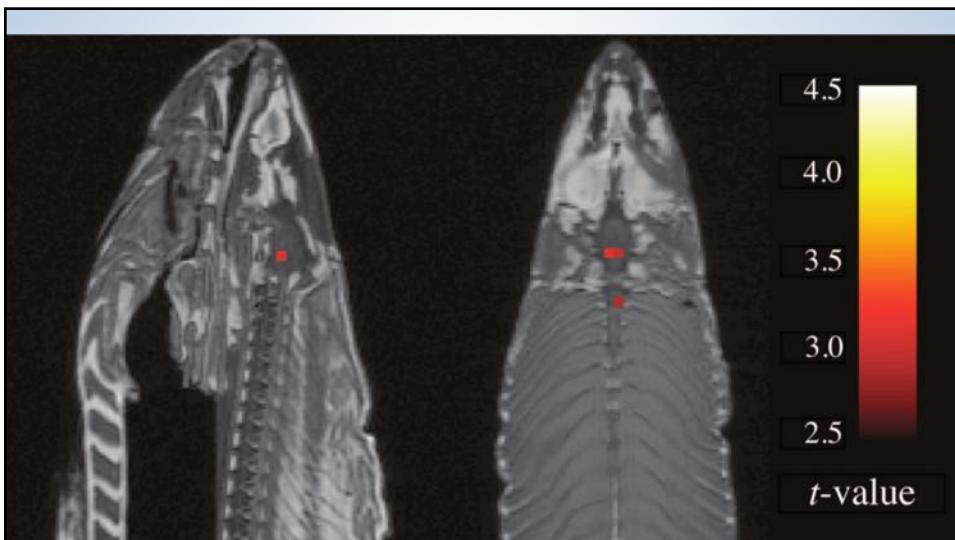
Setting interval null

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



Second-generation p-values

Blume, June 2020



"Sagittal and axial images; $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster...in the medial brain cavity and another...in the upper spinal column."

From Bennett et. al., 2010, JSUR 1:1 1-5. **8064 total voxels; 16 identified.**

Leukemia gene expression

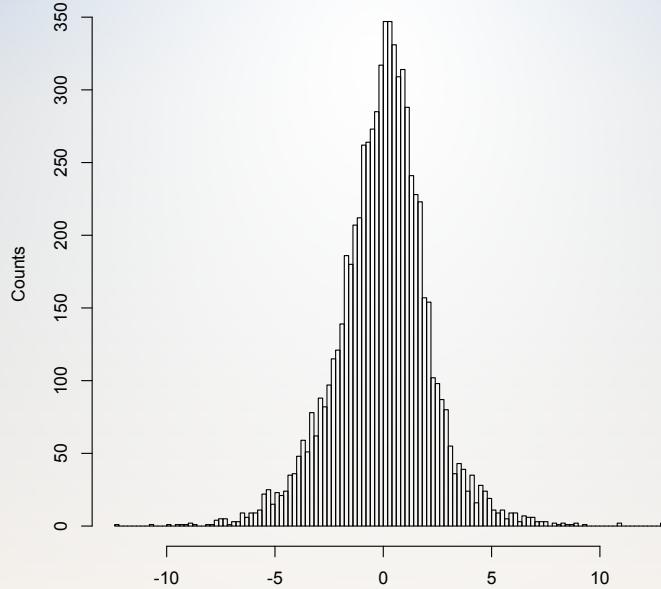
- Classifying acute leukemia by precursors
(Golub 1999, *Science*)
- 7128 genes ; 72 patients (47 ALL and 25 AML)
- Affymetrix chip collected expression levels
- Goal: Identify 'interesting' genes whose expression levels differ between All and AML subjects.
- Looking for fold changes of 2 or more

Second-generation p-values

Blume, June 2020

Histogram

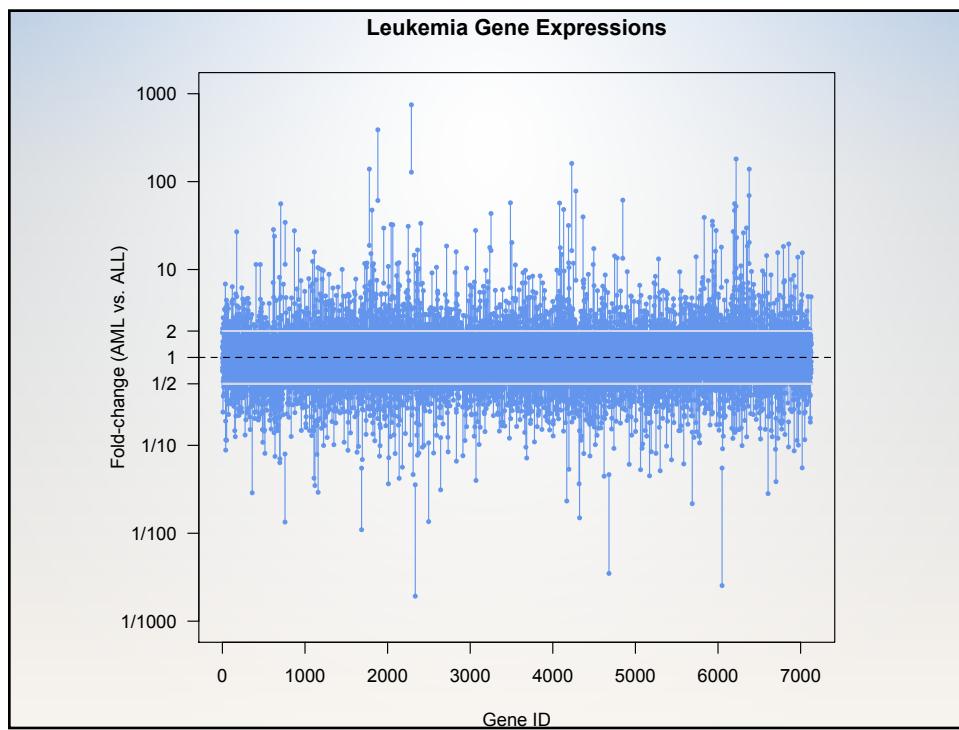
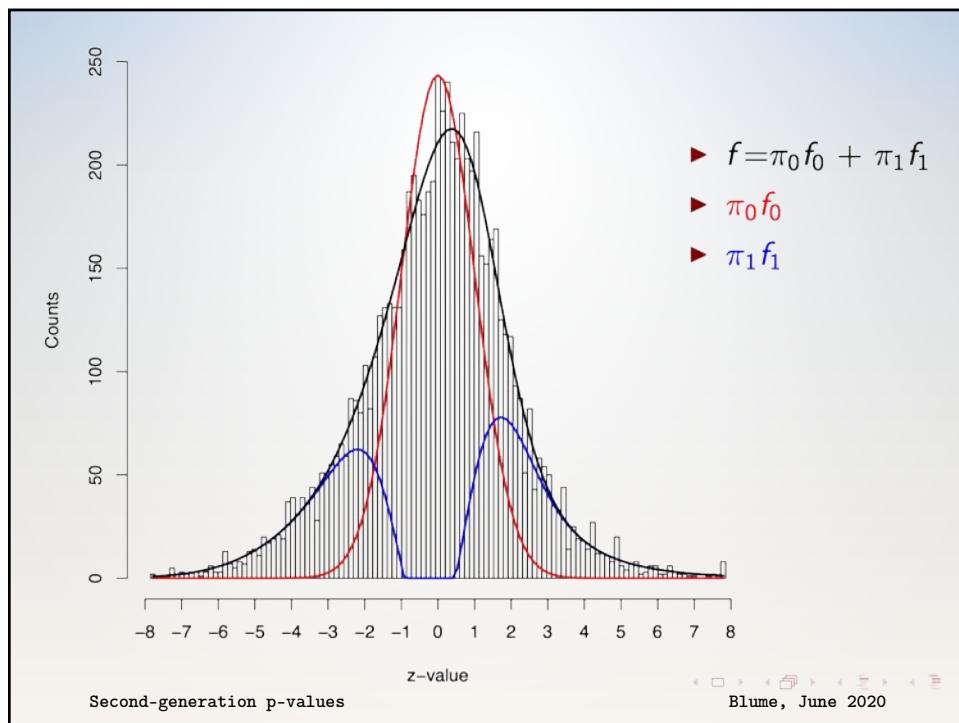
7128 test-statistics

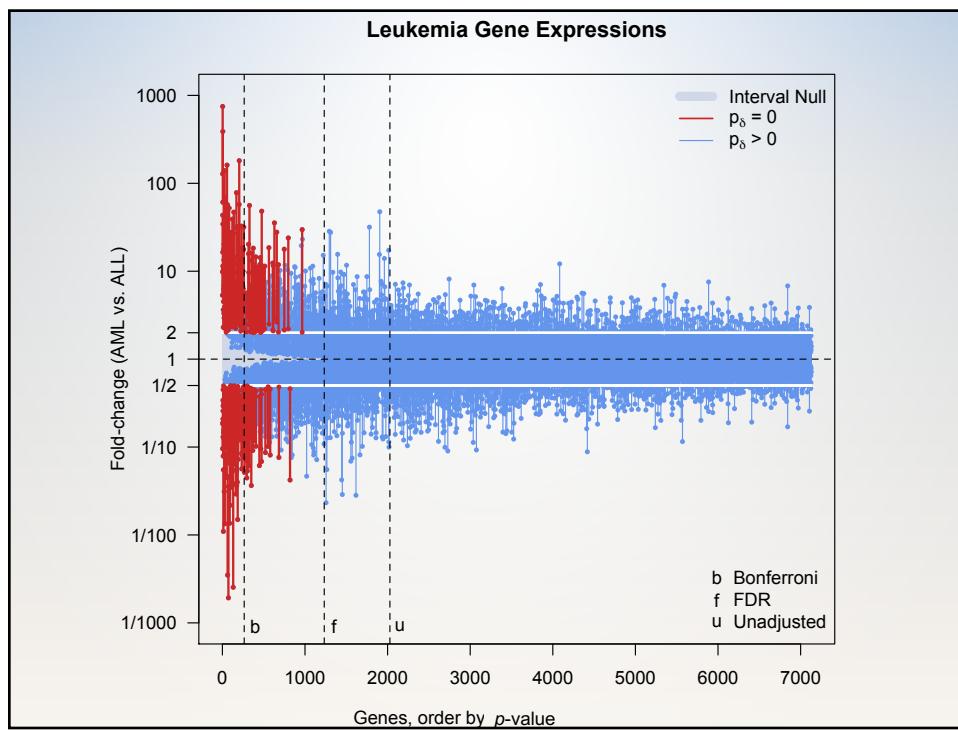
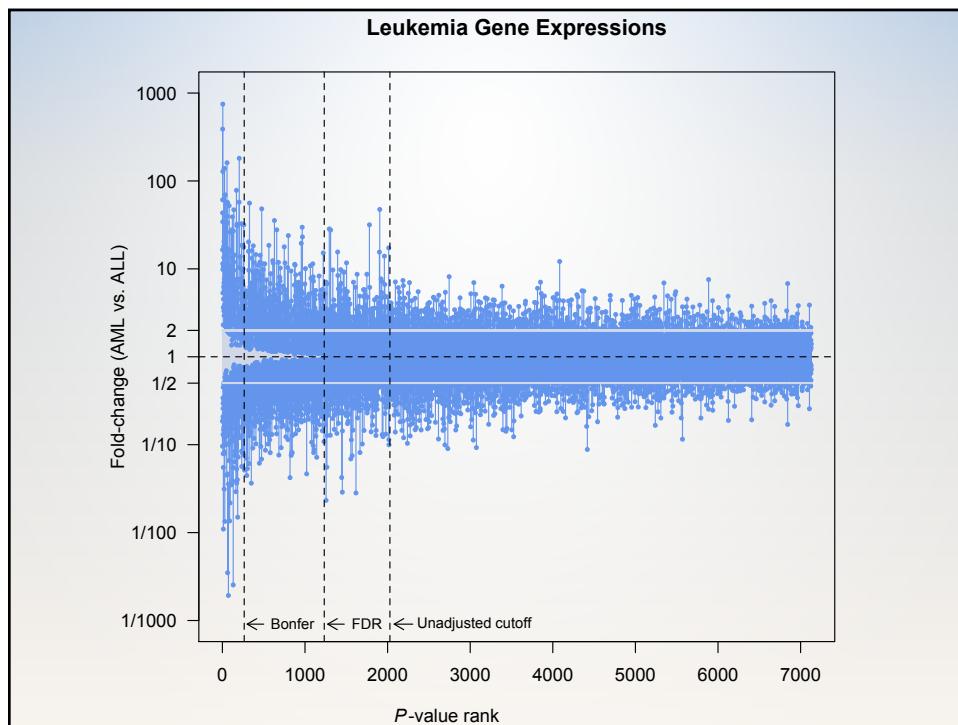


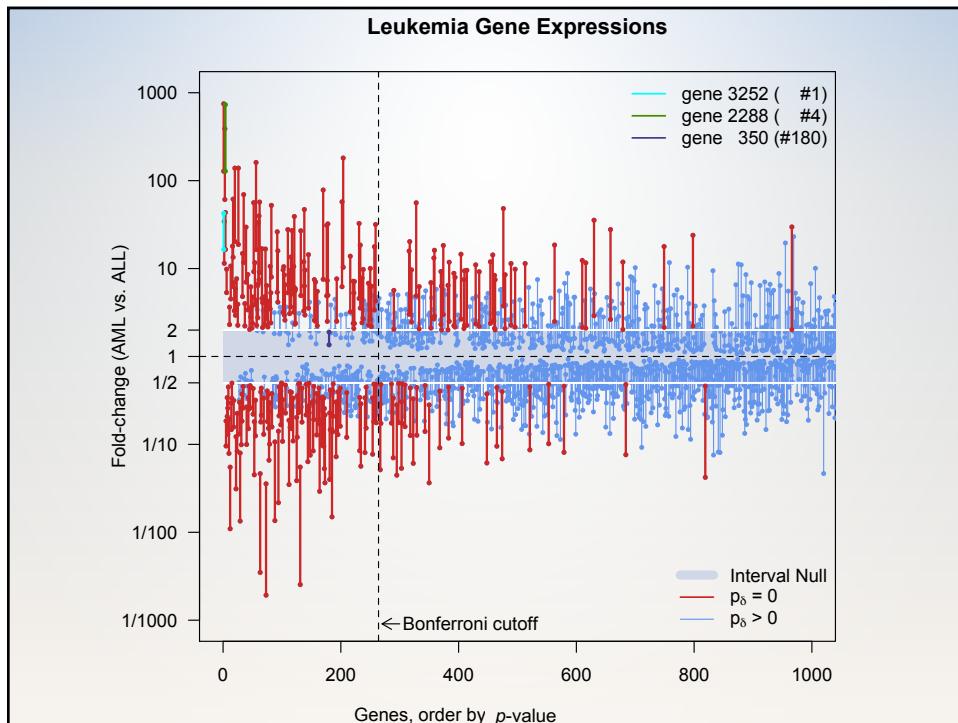
Second-generation p-values

Test Statistic

Blume, June 2020







Leukemia study findings

- Findings: Bonferroni 264, SGPV 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

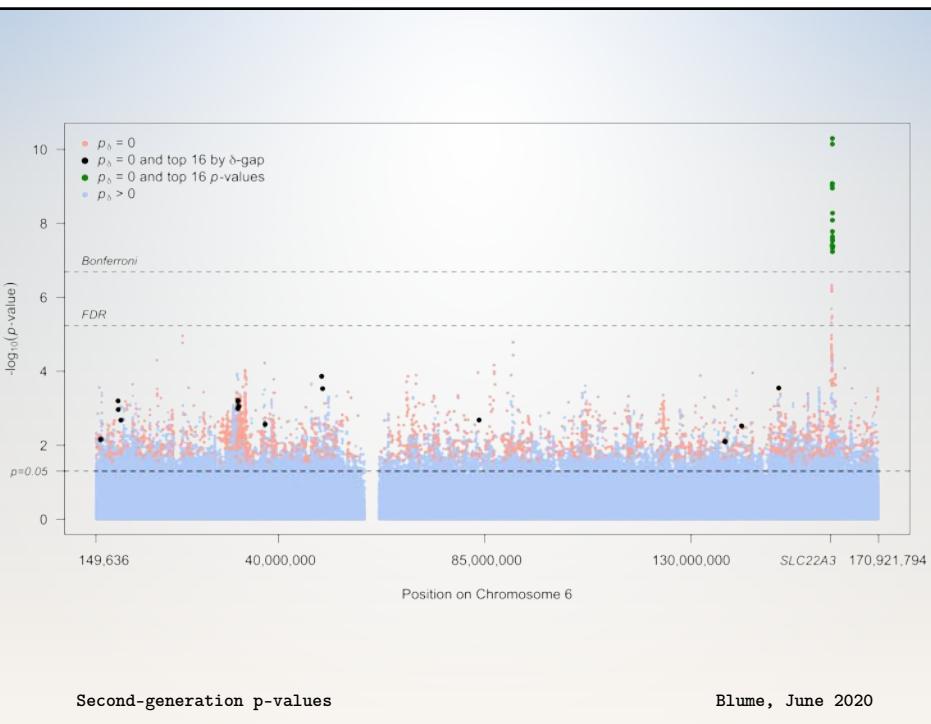
Some SGPV findings
have a priori
published validation

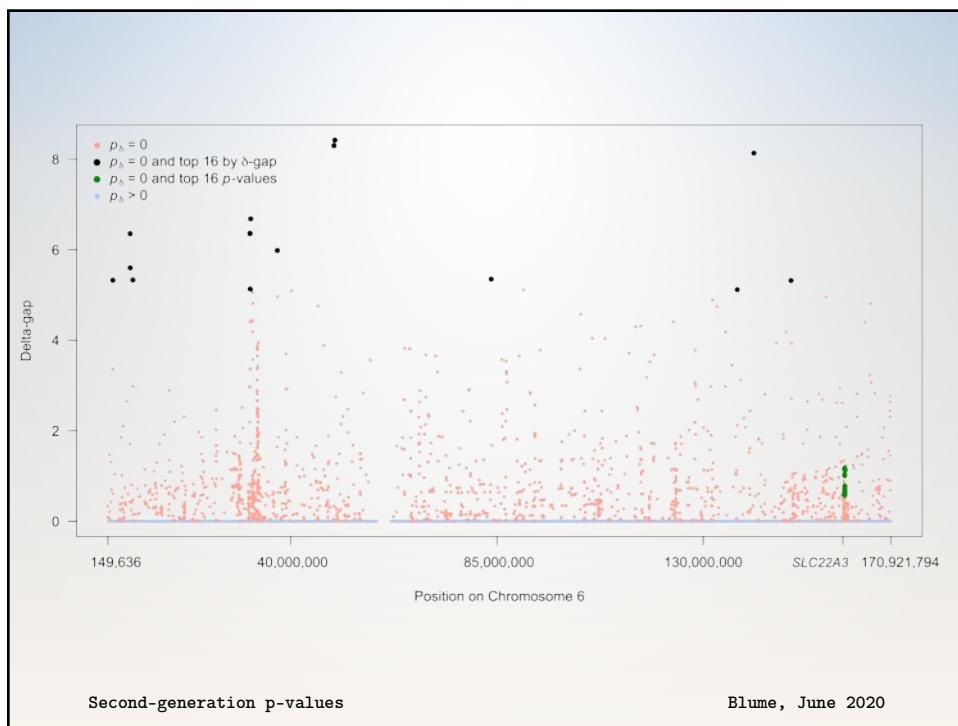
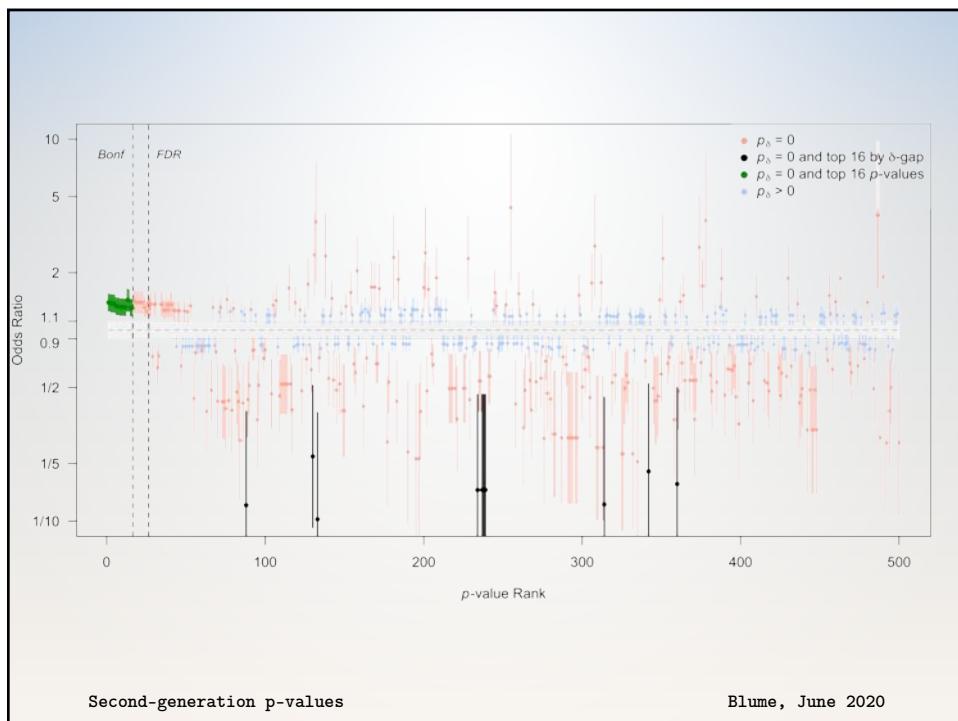
Prostate Cancer SNPs

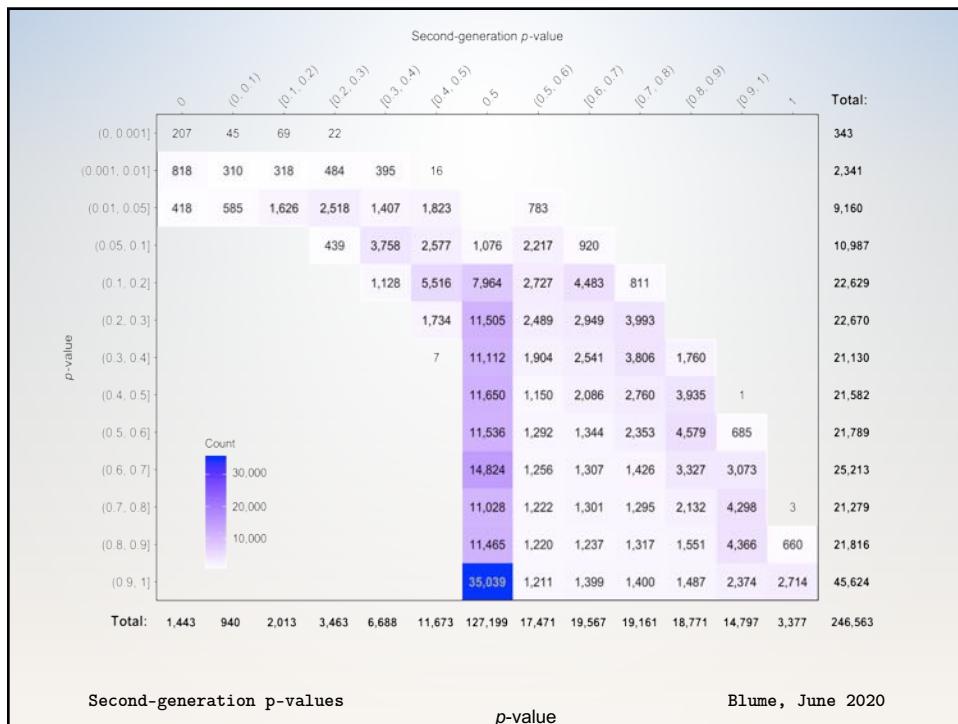
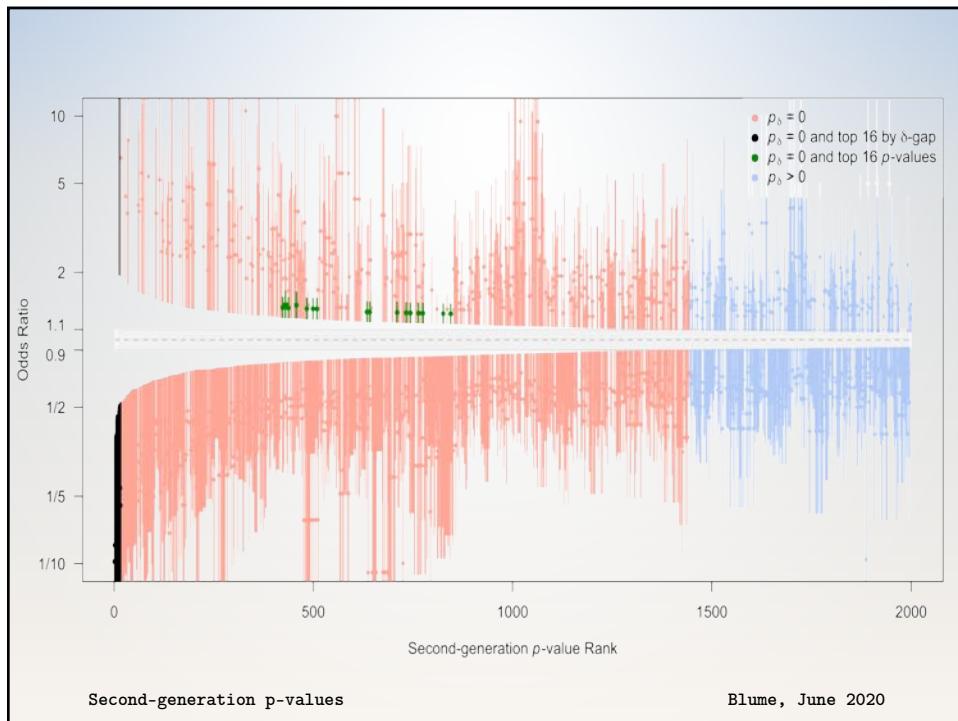
- International Consortium for Prostate Cancer Genetics (Schaid and Chang 2055; ICPCG 2018)
- 3,894 subjects: 2,511 cases & 1,383 controls
- 247,000 single-nucleotide polymorphisms (SNPs) from Chromosome 6
- Goal: Identify 'interesting' SNPs potentially associated with prostate cancer
- Looking for odds ratios of <0.9 or >1.11

Second-generation p-values

Blume, June 2020







False Discovery Rates

- FDR for 5 SGpv=0 findings; computed under various null and alternative configurations (w/ flat prior).

| SNP ID | SGPV rank | p-value rank | OR | 1/8 SI lower limit | 1/8 SI upper limit | FDR ₁ | FDR ₂ | FDR ₃ |
|--------------|-----------|--------------|-------|--------------------|--------------------|------------------|------------------|------------------|
| kgp4568244_C | 1 | 133 | 0.10 | 0.03 | 0.37 | 2.9% | 17.1% | 3.3% |
| kgp8051290_G | 13 | 2002 | 15.58 | 1.95 | 124.68 | 4.3% | 30.3% | 4.9% |
| kgp4497498_A | 28 | 255 | 4.37 | 1.80 | 10.64 | 2.5% | 8.6% | 3.1% |
| rs3123636_G | 423 | 1 | 1.39 | 1.26 | 1.55 | 0.01% | 0.1% | 0.4% |
| kgp7460928_G | 1443 | 3310 | 1.78 | 1.11 | 2.87 | 2.4% | 2.0% | 3.0% |

Second-generation p-values

Blume, June 2020

Statistical Properties

Suppose interval I has coverage probability $1-\alpha$, then

Three ‘Error’ Rates

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

Two False Discovery Rates

- $P(H_0 | p_\delta = 0)$
- $P(H_1 | p_\delta = 1)$

Second-generation p-values

Blume, June 2020

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

Second-generation p-values

Blume, June 2020

Testing

| Evidential Metric | What it measures | Hypothesis Testing | Significance Testing |
|-------------------|---|-------------------------------------|--|
| 1 | strength of the evidence | Absent | Tail-area probability (<i>p</i> -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 |

- The *tail-area probability* is used to measure *three* distinct metrics

Second-generation p-values

Blume, June 2020

Second-generation p -value

- StatisticalEvidence.com
- Examine statistical properties in module 2
- Retains strict error control

| Evidential Metric | What it measures | SGPV |
|-------------------|---------------------------|---|
| 1 | Summary measure | p_δ |
| 2 | Operating characteristics | $P(p_\delta = 0 H_0)$ $P(p_\delta = 1 H_1)$ $P(0 < p_\delta < 1 H)$ |
| 3 | False discovery rates | $P(H_0 p_\delta = 0)$ $P(H_1 p_\delta = 1)$ |

Second-generation p -values

Blume, June 2020

Remarks

- Second-generation p -values...
 - Indicate compatibility with null or alternative
 - Indicate when the data are inconclusive (!)
 - Straightforward to compute and interpret
 - Maintains controls error rates
 - Reduces the false discovery rate
- Anchoring the scale of the effect size...
 - Eliminates most Type I Errors
 - Improves scientific translation of statistical model

Second-generation p -values

Blume, June 2020

Acknowledgements

- Collaborators

- William D. Dupont
- Robert A. Greevy
- Valerie Welty
- Jeffrey R. Smith

- Website / Papers / Code

- statisticalevidence.com
- PLOS One (2018) ; The American Stat. (2019)
- Google “Second-Generation *p*-value”
- `devtools::install_github("weltybiostat/sgpv")`

Outrageous Claim (!?)

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

Second-generation *p*-values

Blume, June 2020