Experimental design: variability, replication and power

Jeff Leek

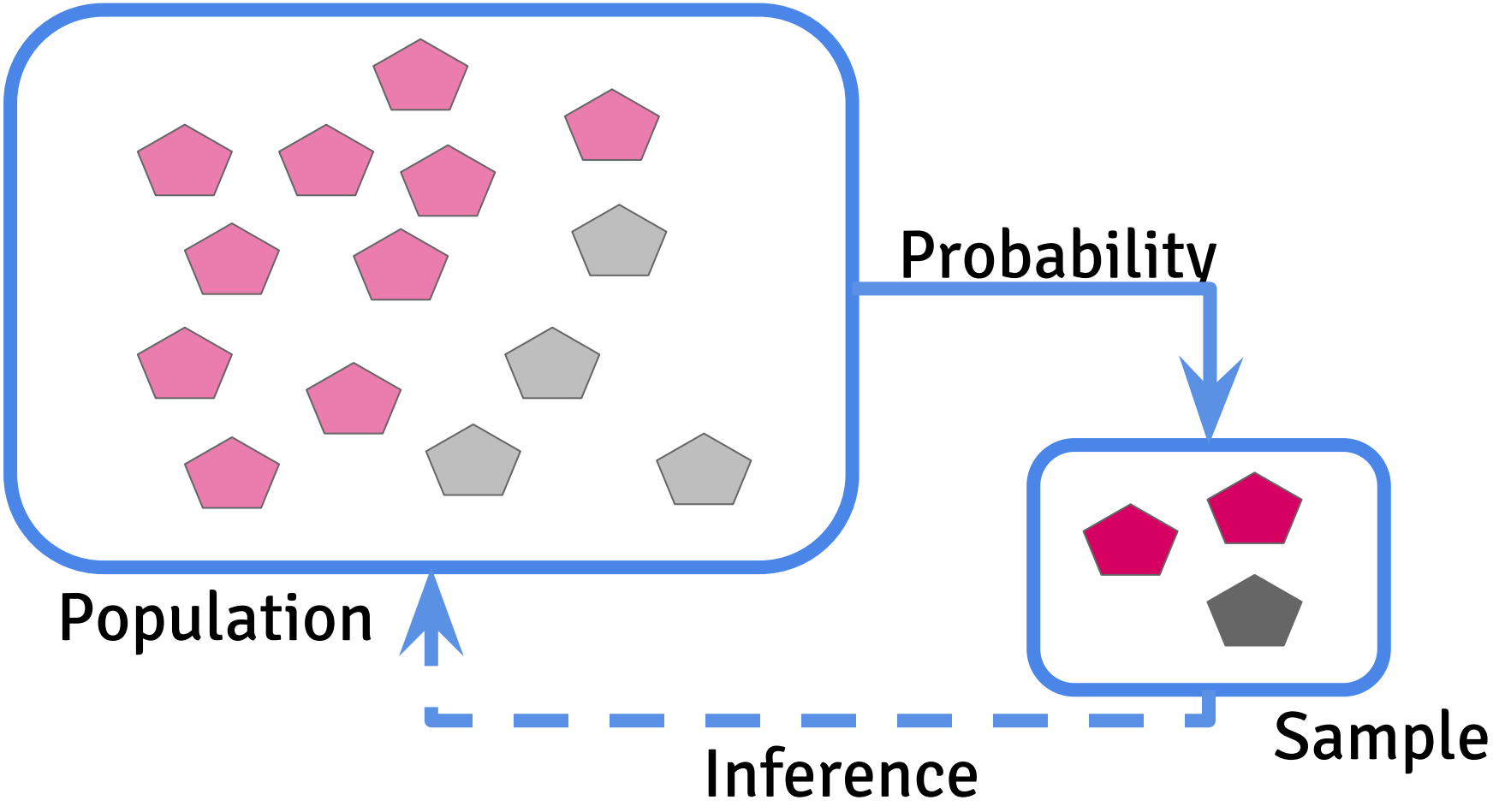
@jtleek

www.jtleek.com

Key ideas

- Sources of variation
- Power
- Replication

Central dogma of statistics


Probability
Population
Sample
Inference

Three types of variablity

Var(Genomic Measurement)

= Phenotypic variability

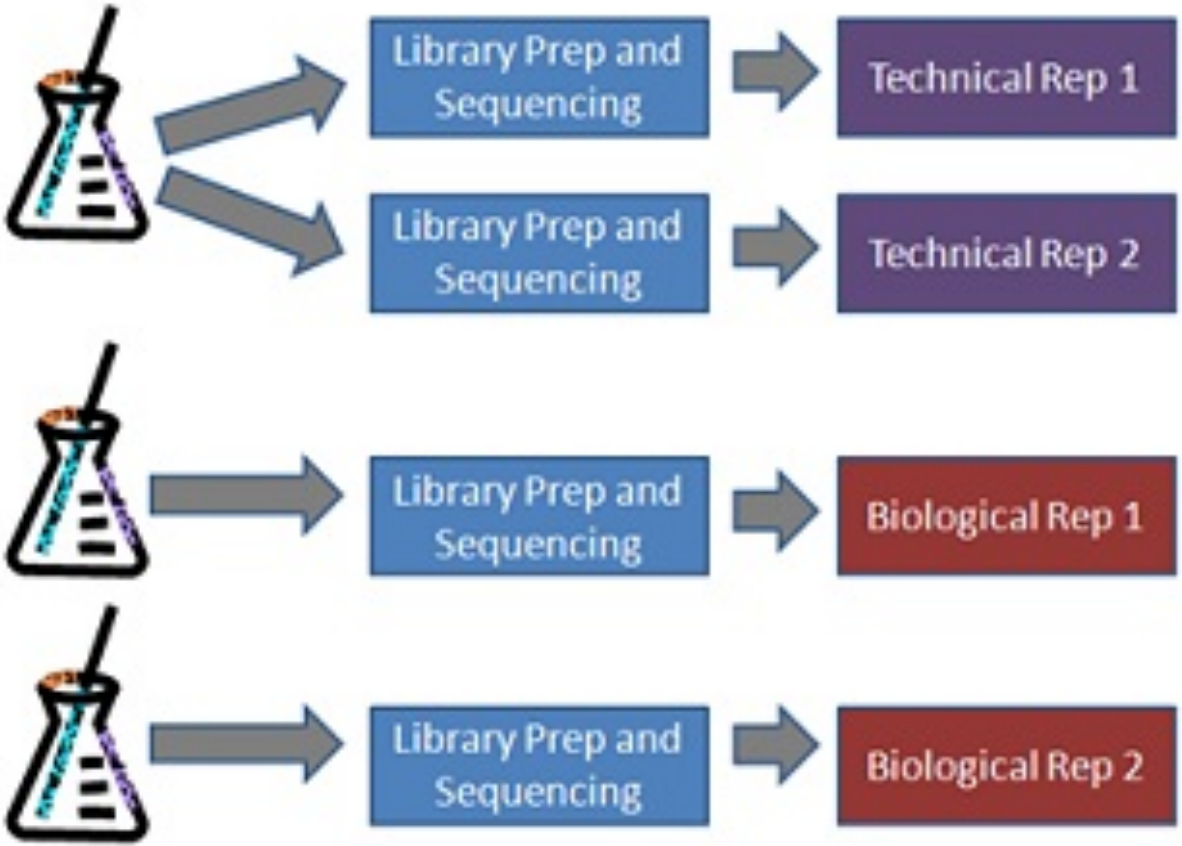
+ Measurement error

+ Natural biological variation

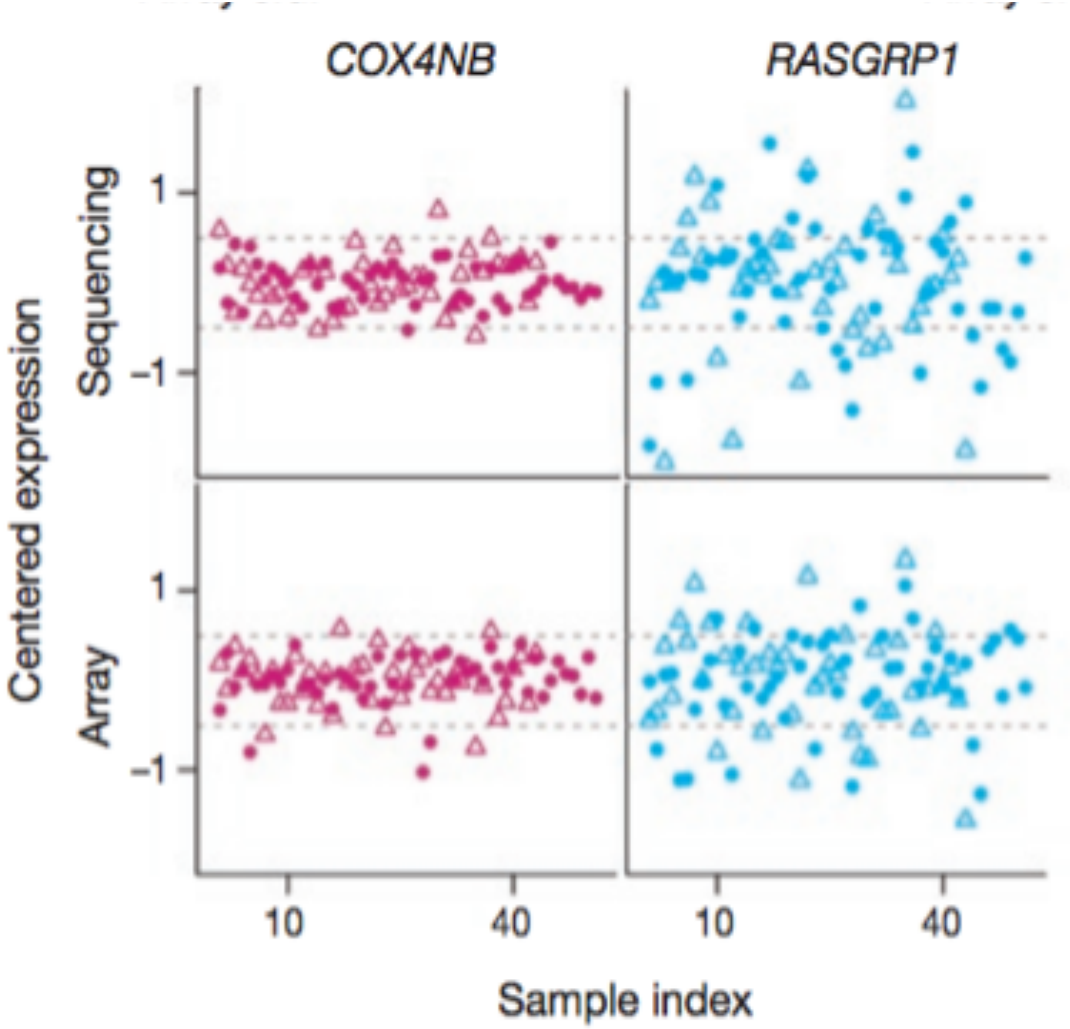
Replicates

- Biological
- Technological

http://bioinformatics.bc.edu/marthlab/scotty/help.html



New technology doesn’t eliminate variability


COX4NB
RASGRP1
Centered expression
Sequencing
Array
1
−1
1
−1
10
40
10
40
Sample index

Sample size

N = Number of Measurements

$$N = \frac{(\$ \text{ you have})}{(\$ \text{ / measurement})}$$

Sample size depends on measurement

Rare mendelian disease

N ≈ 3-5

RNA-Sequencing study

N ≈ 10 - 1,000

DNA methylation study

N ≈ 10 - 1,000

Common disease genome-wide association

N ≈ 10,000 - 1,000,000+

New technology doesn’t eliminate variability

http://www.nature.com/nbt/journal/v29/n7/full/nbt.1910.html

NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

Sequencing technology does not eliminate biological variability

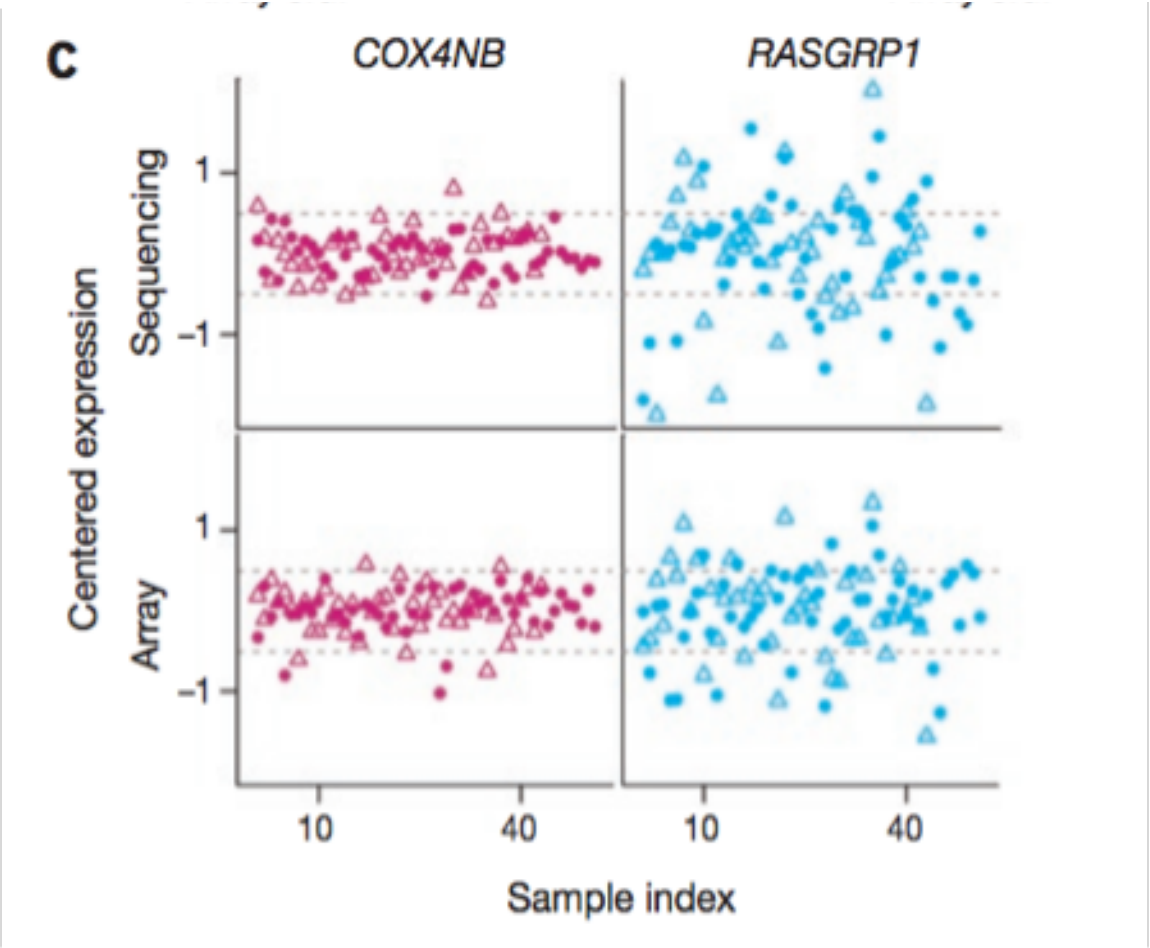
Kasper D Hansen, Zhijin Wu, Rafael A Irizarry & Jeffrey T Leek

Affiliations | **Corresponding authors**

Nature Biotechnology **29**, 572–573 (2011) | doi:10.1038/nbt.1910

Published online 11 July 2011

http://www.nature.com/nbt/journal/v29/n7/full/nbt.1910.html



This is often ignored

http://www.nature.com/nbt/journal/v29/n7/full/nbt.1910.html

Pubmed ID	Journal (Year)	# of Biological Groups	# of Technical Replicates	# of Biological Replicates
18451266 [10]	Science (2008)	1	2	2
19056941 [11]	Science (2008)	1	1	2
18516045 [6]	Nature Methods (2008)	3	2	1
18599741 [12]	Science (2008)	2	2	1
18516046 [13]	Nature Methods (2008)	2	3	1
18978772 [14]	Nature (2008)	15	1	1 (6 in 1 group)
19234113 [15]	PNAS (2009)	2	1	2
19581875 [16]	Nature Biotechnology (2009)	1	3	1
19349980 [17]	Nature Methods (2009)	4	1	2
20436464 [18]	Nature Biotechnology (2010)	4	1	1
20810668 [19]	Genome Research (2010)	9	1	1
20980679 [20]	Blood (2010)	2	1	2
21057496 [21]	Nature Methods (2010)	2	1	1
20452967 [22]	Genome Research (2010)	4	1	2
20363980 [23]	Genome Research (2010)	1	1	1
20220758 [1]	Nature (2010)	1	2	69
20220756 [2]	Nature (2010)	1	1	60

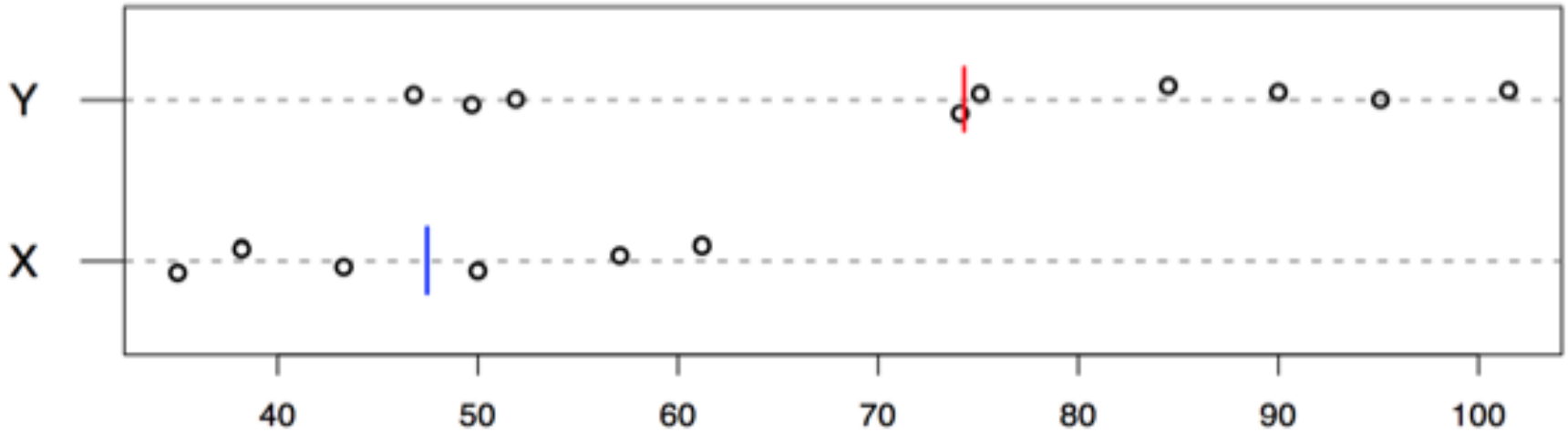
Variability and power

Power =

Probability of discovering a real signal if it is there

- Power is typically set at 80%
- Calculations are based on made up assumptions
- Higher power is better
- Low powered studies don’t replicate

Slide courtesy: Ingo Ruczinski



Slide courtesy: Ingo Ruczinski

- n = 10 for each group; effect = $\Delta = 5$; pop'n SD = $\sigma = 10$

```
power.t.test(n=10, delta=5, sd=10)
```

$\longrightarrow$ 18%

- power = 80%; effect = $\Delta = 5$; pop'n SD = $\sigma = 10$

```
power.t.test(delta=5, sd=10, power=0.8)
```

$\longrightarrow$ n = 63.8 $\longrightarrow$ 64 for each group

- power = 80%; effect = $\Delta = 5$; pop'n SD = $\sigma = 10$; one-sided

```
power.t.test(delta=5, sd=10, power=0.8,
             alternative="one.sided")
```

$\longrightarrow$ n = 50.2 $\longrightarrow$ 51 for each group

Slide courtesy: Ingo Ruczinski