# Measuring significant relationships between sets of genomic features

Aaron Quinlan

quinlanlab.org

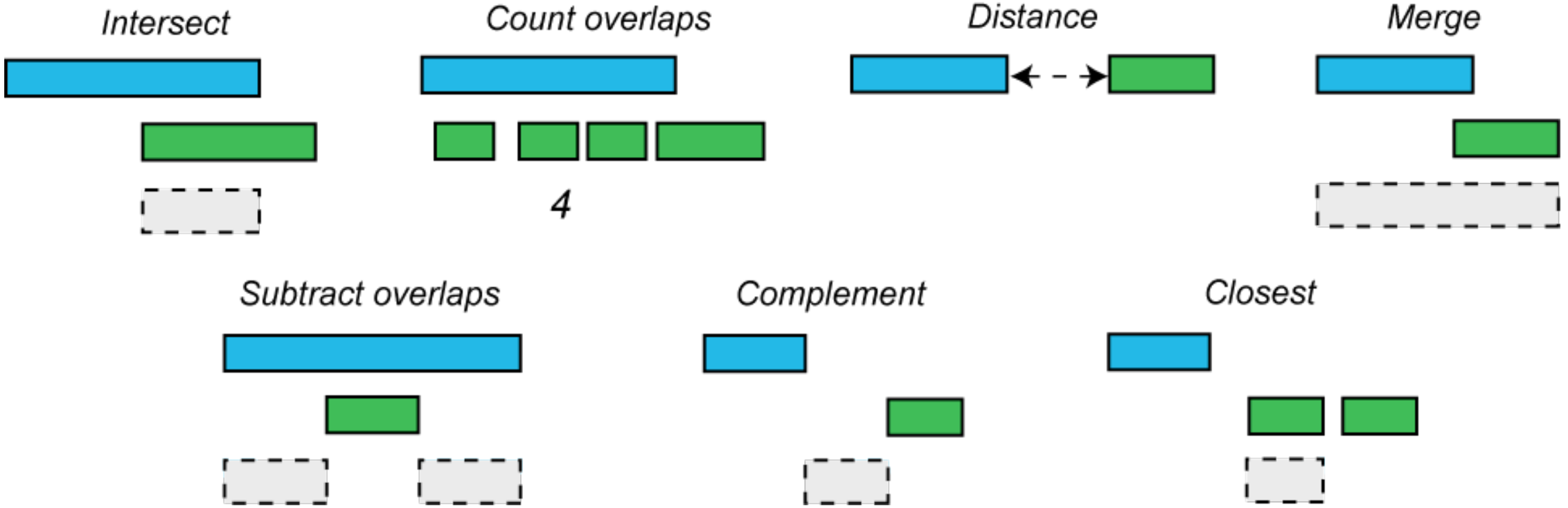
April 25, 2013

Public Health Sciences

Center for Public Health Genomics

Biochemistry and Molecular Genetics

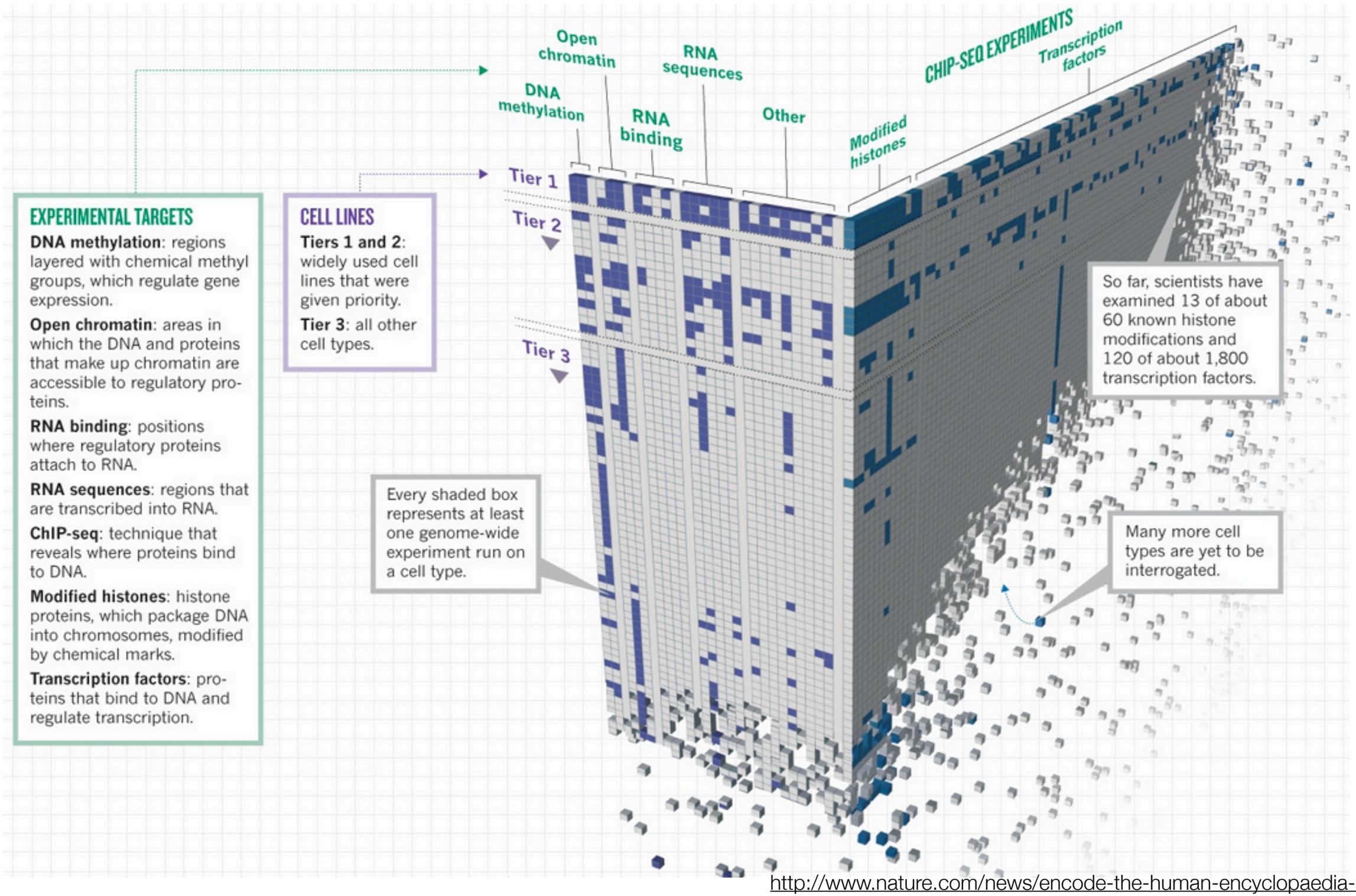

# genome arithmetic



# analytic limitations

- Existing methods such as bedtools work primarily on pairs of datasets
- These methods exhaustively list the **individual details** of relationships between genome annotations and experimental datasets.
- How do we reduce the huge dimensionality to something we can understand and compare?

# How do we make sense of **complex, multi-dimensional** datasets to gain insights into genome biology?




http://www.nature.com/news/encode-the-human-encyclopaedia-1.11312

“Low dimensional representations and human understanding are synonymous.”

Ewan Birney

Scale
chr21:
200 kb
hg19
32,700,000
32,800,000
32,900,000
33,000,000
33,100,000
33,200,000
33,300,000
33,400,000
33,500,000
Uniqueness of 35bp Windows from ENCODE/OpenChrom(Duke)
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
TIAM1
TIAM1
TIAM1
TIAM1
BC014150
BC041449
SOD1
HUNK
LINC00159
SCAF4
SCAF4
SCAF4
SCAF4
SCAF4
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
GM12878 Syn Pk
H1-hESC Syn Pk
K562 Syn Pk
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
GM12878 Pk
GM12878 DS
GM12878 OS
H1hESC Pk
H1hESC DS
H1hESC OS
K562 Pk
K562 DS
K562 OS
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
GM12878 FAIRE Pk
GM12878 FAIRE DS
GM12878 FAIRE OS
H1-hESC FAIRE Pk
H1-hESC FAIRE DS
H1-hESC FAIRE OS
K562 FAIRE Pk
K562 FAIRE DS
K562 FAIRE OS
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
GM12878 Ht 1
GM12878 Pk 1
GM12878 Sg 1
GM12878 Ht 2
GM12878 Pk 2
GM12878 Sg 2
H1hESC Ht 1
H1hESC Pk 1
H1hESC Sg 1
K562 Ht 1
K562 Pk 1
K562 Sg 1
K562 Ht 2
K562 Pk 2
K562 Sg 2
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
K562 Sig
HepG2 Sig
HUVEC Sig

# This is a hard yet important problem.

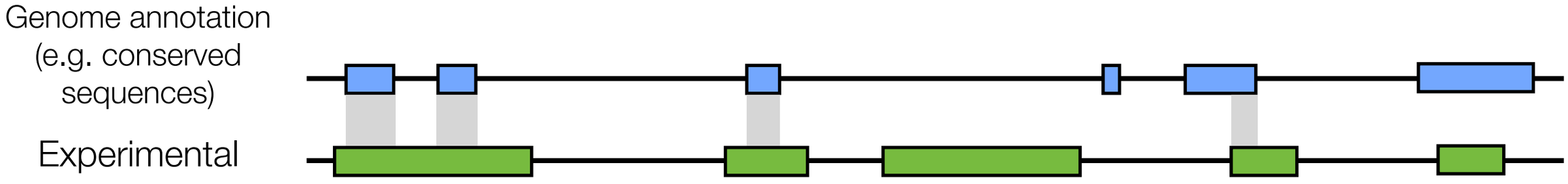
- Understand every base pair's function (or lack thereof) in different cell types and contexts.
- Challenges (among many):
  - Basic exploratory data analysis: slicing and dicing very large, heterogenous datasets.
  - Visualization: unbiased exploration; let the data tell its story.
  - **Testing for significant spatial relationships**

# Motivation

- How do we decide whether two observations (e.g., ChIP-seq peaks from two different assays) are correlated with one another?
- They overlap, but is it more than one would expect by chance?

# How do we detect genomic co-association?

*That is, do two sets of genomic features co-occur (overlap or have spatial consistency) more than expected?*

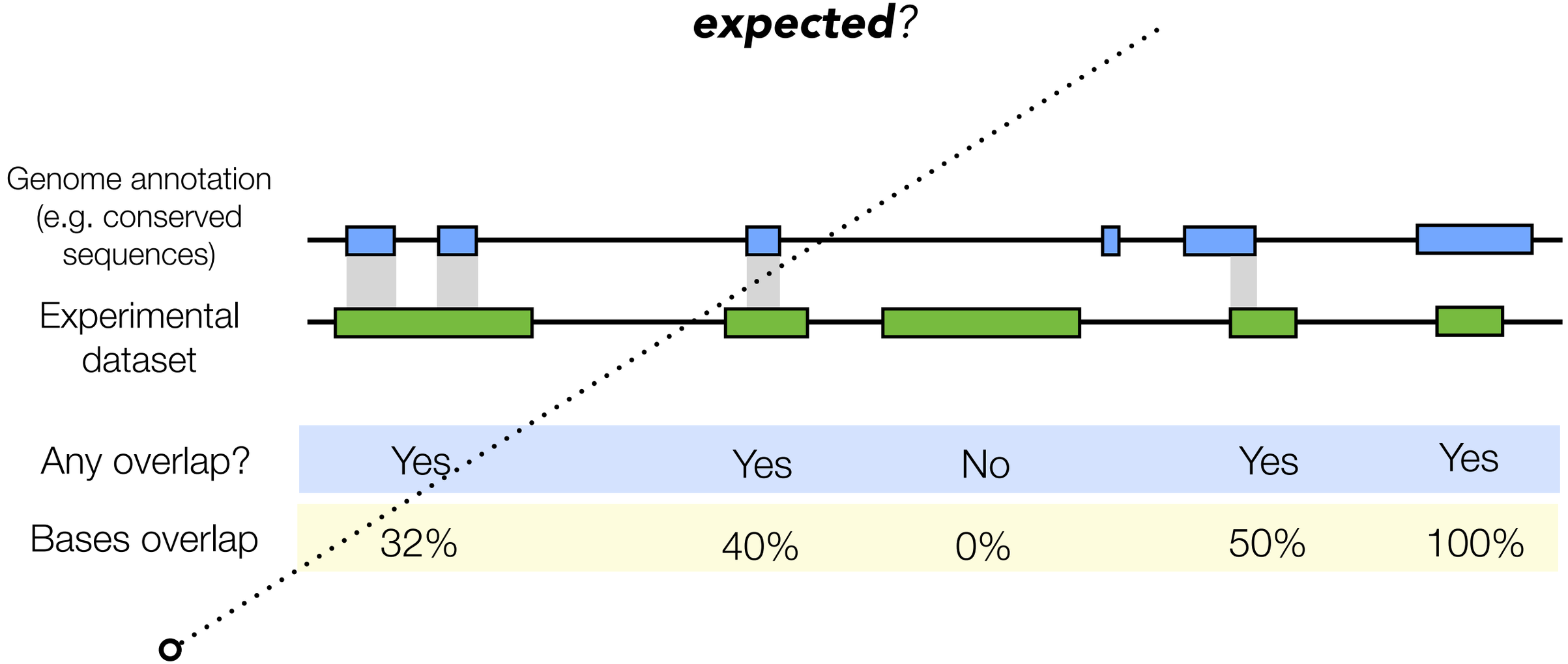



| | | | | | |
|---|---|---|---|---|---|
| Any overlap? | Yes | Yes | No | Yes | Yes |
| Bases overlap | 32% | 40% | 0% | 50% | 100% |

Adapted from slide by Ben Brown

# How do we detect genomic co-association?

*That is, do two sets of genomic features co-occur (overlap or have spatial consistency) more than* ***expected****?*

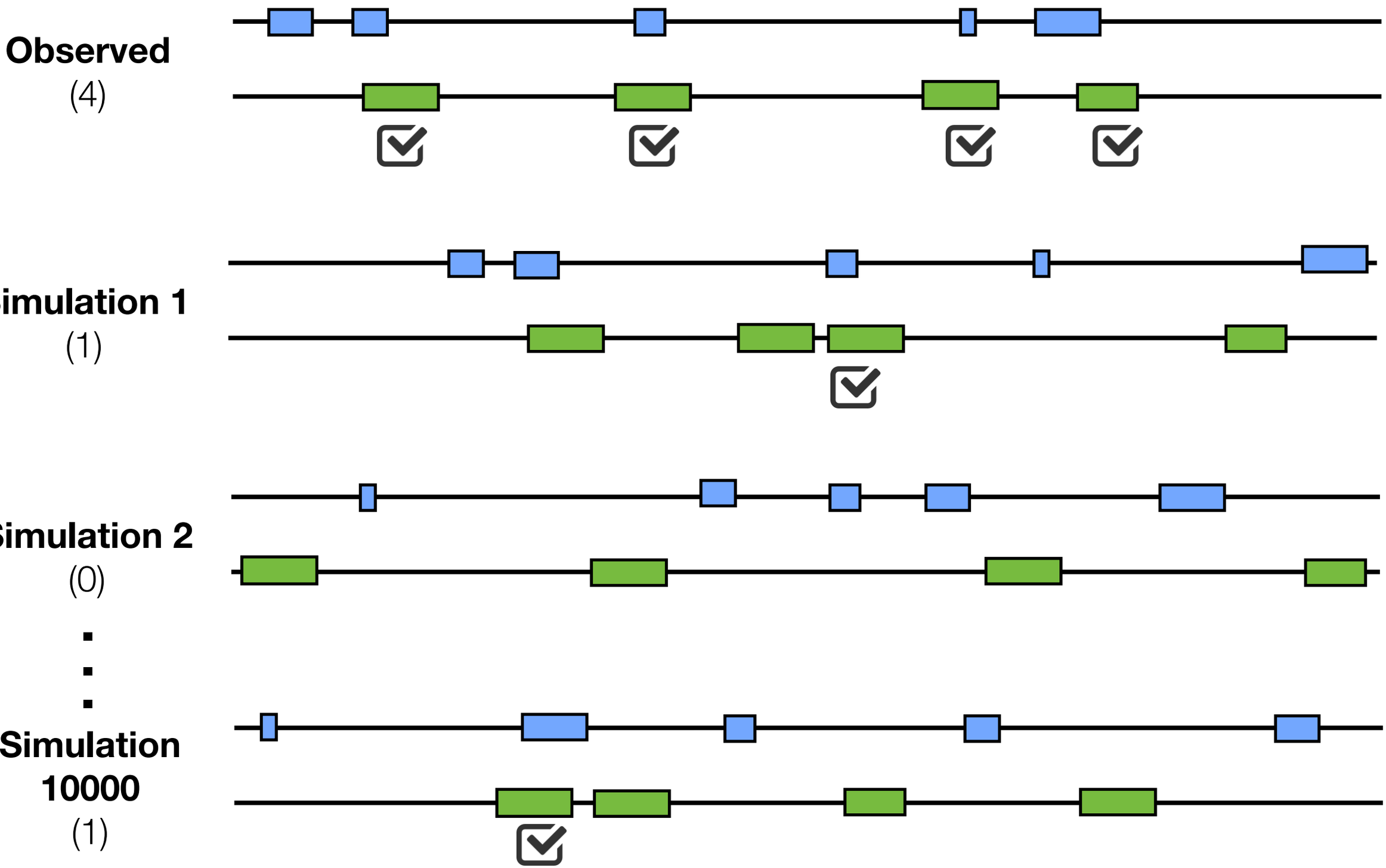



How do we develop a proper null expectation in order to reduce the dimensionality of the data to an informative statistic?

Adapted from slide by Ben Brown

# Monte-Carlo simulation (slow)

*Are the observed feature overlaps more than expected by chance?*



# How do we speed this up?

*We invent a scalable new algorithm, of course!*

---

**Binary Interval Search (BITS):**
**A Scalable Algorithm for Counting Interval Intersections**

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3]*

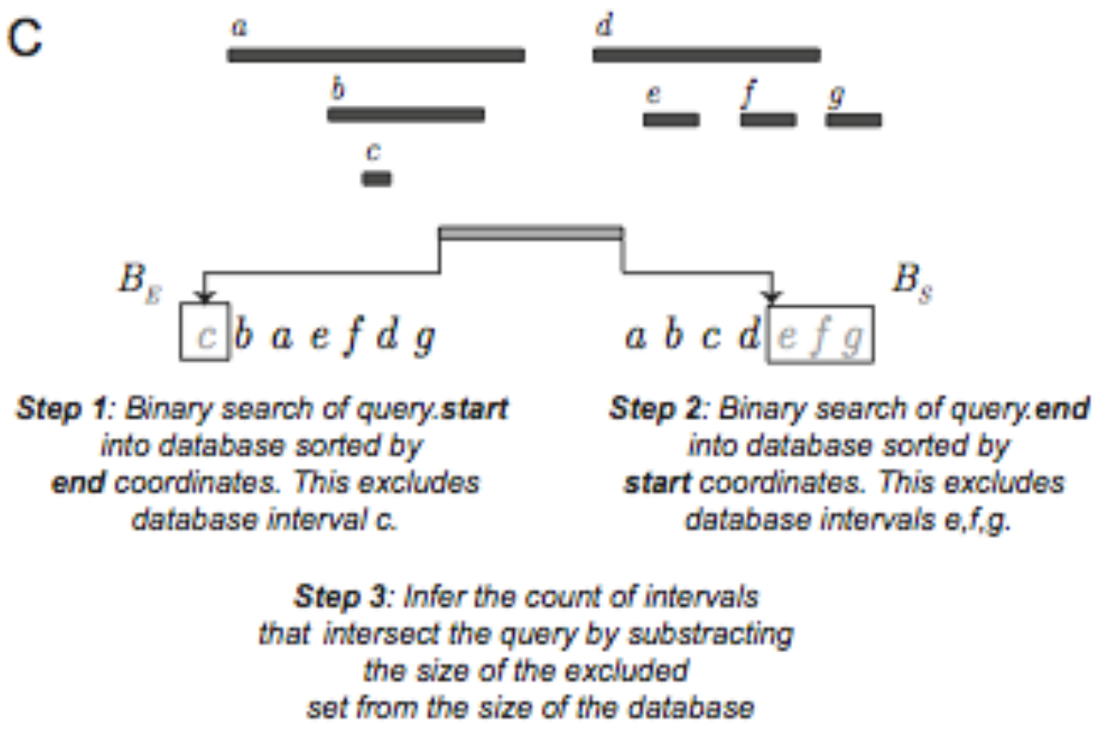
[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA

Ryan Layer

https://github.com/arq5x/bits

Advance access at Bioinformatics

# Binary InTerval Search BITS

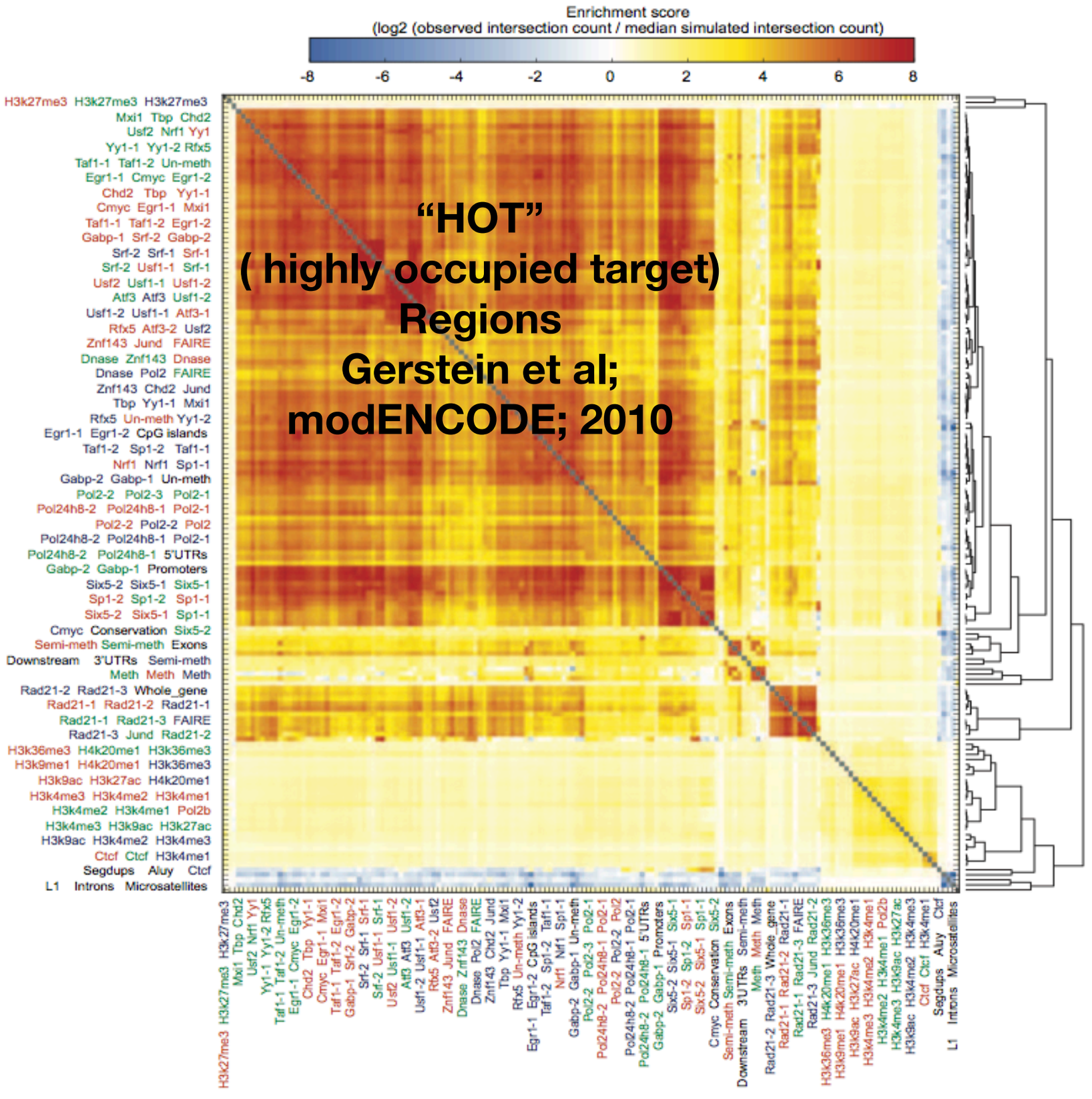



**Algorithm 1:** Single interval intersection counter

**Input**: Sorted interval starts and ends $B_S$ and $B_E$, query interval $a$
**Output**: Number of intervals $c$ intersecting $a$

**Function** ICOUNT($B_S, B_E, a$) **begin**
    $first \leftarrow$ BINARYSEARCH($B_S, a.end$)
    $last \leftarrow$ BINARYSEARCH($B_E, a.start$)
    $c \leftarrow first - last$  /* $= |B| - (last + (|B| - first))$ */
    **return** $c$

- Entirely novel algorithm for detecting genomic interval intersections.
- Clever aspect: unlike any other algorithm, it can deduce the **count** of overlaps without having to **enumerate** each individual intersection.
- Uses two binary searches. Very fast. ***Spaceballs*** **fast on a GPU**.
- If you haven't heard, faster is better.




159 ENCODE ChIPseq, RNA-seq,
from 3 cell types (GM12878, hESC, k562). UCSC tracks as well

25,281 (159^2) pairwise dataset comparisons

10,000 simulations per 25,281 comparisons

**44 trillion** intersection measurements

**6 days is better than 138!**

159 ENCODE ChIPseq, RNA-seq,
from 3 cell types (GM12878, hESC, k562). UCSC tracks as well

# How do we develop a null expectation?

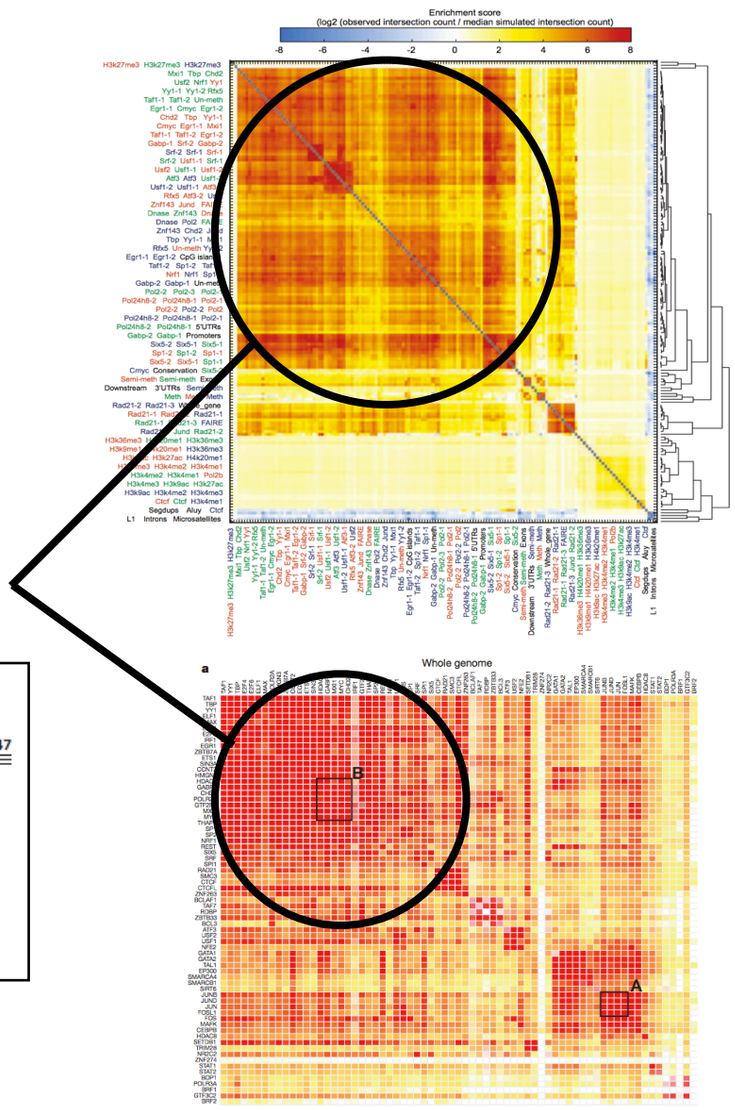
**1. Monte-Carlo simulation**

**Binary Interval Search (BITS):**
**A Scalable Algorithm for Counting Interval Intersections**

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3]*

[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA

"HOT regions"




**2. Block bootstrap sampling (Peter Bickel)**

ARTICLE
doi:10.1038/nature11247

**An integrated encyclopedia of DNA elements in the human genome**

The ENCODE Project Consortium*

NON PARAMETRIC METHODS FOR GENOMIC INFERENCE

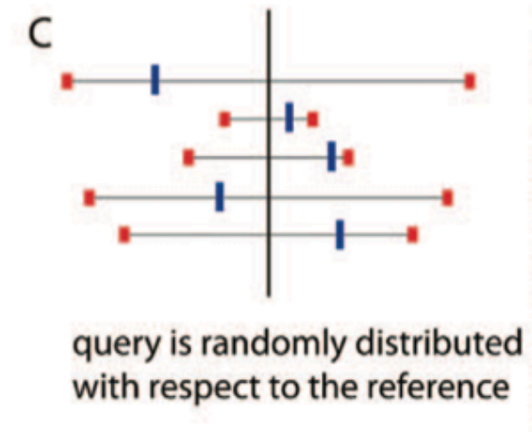
BY PETER J. BICKEL*,†, NATHAN BOLEY*,†, JAMES B. BROWN*,†, HAIYAN HUANG*,†, NANCY R. ZHANG*,‡

**3. Spatial methods (Sarah Wheelan)**

**Exploring Massive, Genome Scale Datasets with the GenometriCorr Package**

**Alexander Favorov[1,2,3]*, Loris Mularoni[1]¤a, Leslie M. Cope[1], Yulia Medvedeva[2,3]¤b, Andrey A. Mironov[4,5], Vsevolod J. Makeev[2,3], Sarah J. Wheelan[1]***

**1** Department of Oncology, Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **2** Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia, **3** Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia, **4** Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, **5** Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia




*Integrating these and other, novel association statistics into* ***bedtools2***

# “Block bootstrap” method

- Sampling a la Monte Carlo
- However, don’t shuffle randomly.
- Instead, preserve spacial relationships.
- http://encodestatistics.org/
- *Example on whiteboard.*

# Alternative (better?) methods?

PLOS | COMPUTATIONAL BIOLOGY



OPEN ACCESS PEER-REVIEWED

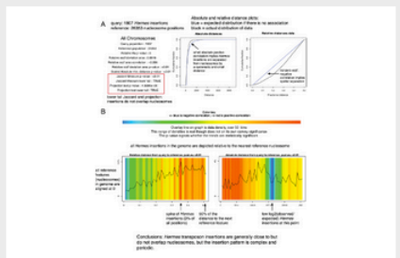
RESEARCH ARTICLE | FEATURED IN PLOS COLLECTIONS

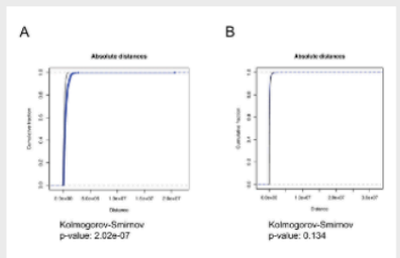


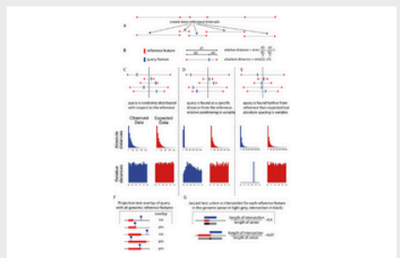
## Exploring Massive, Genome Scale Datasets with the GenometriCorr Package

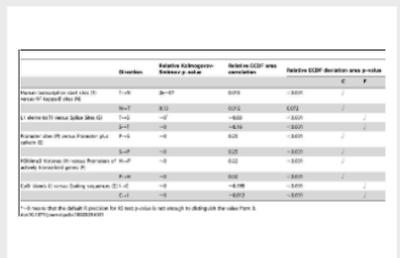
Alexander Favorov, Loris Mularoni, Leslie M. Cope, Yulia Medvedeva, Andrey A. Mironov, Vsevolod J. Makeev, Sarah J. Wheelan







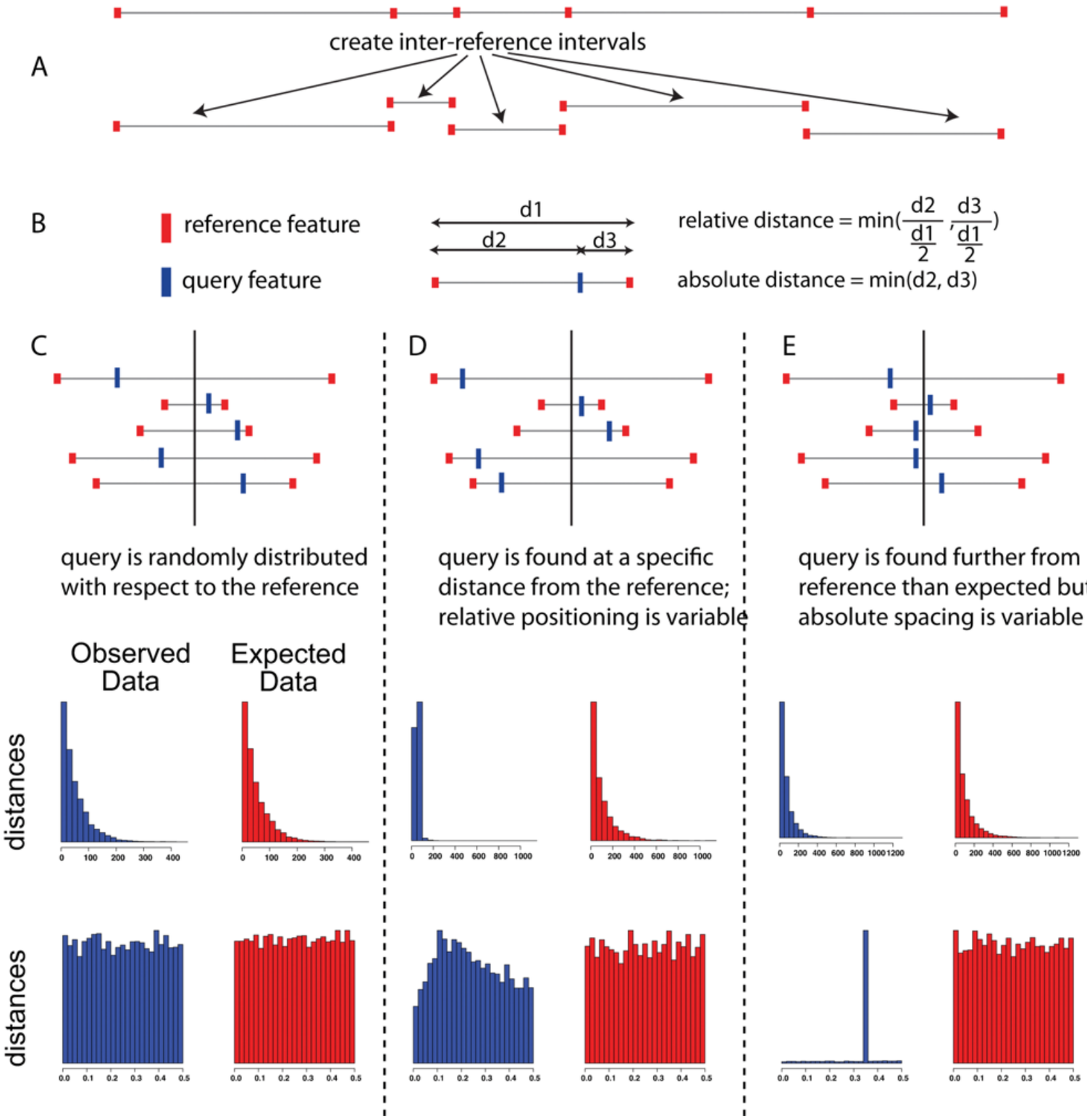


# “Distance measures”

What if features don’t *overlap*
yet are most always nearby?

For example: TF binding peaks and TSS

# “Distance measures”



# Other measures.