Big data in bioinformatics

Recap

Experimental design

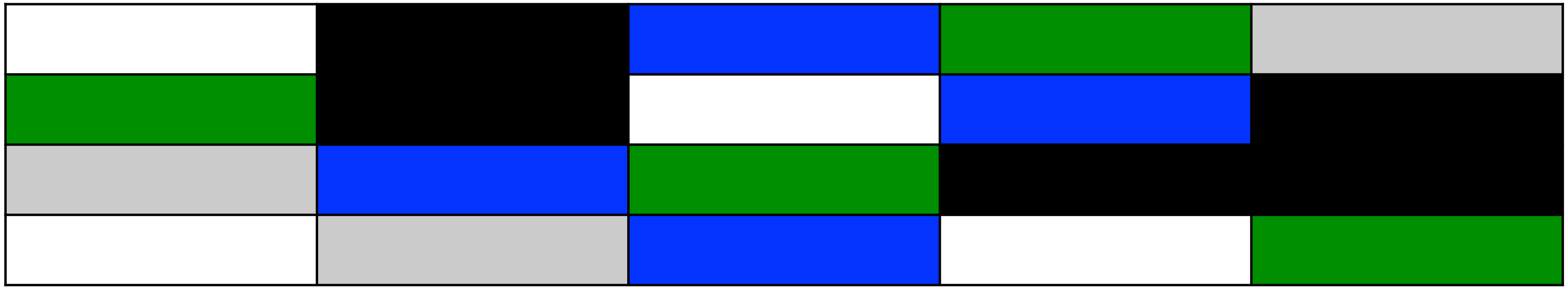
- Experimental design affects which tools we can use:
  - Categories? Continuous?
  - Nested? Orthogonal?
  - Blocks?
  - Time-series?
  - Fixed or random factors?
  - Replicates?

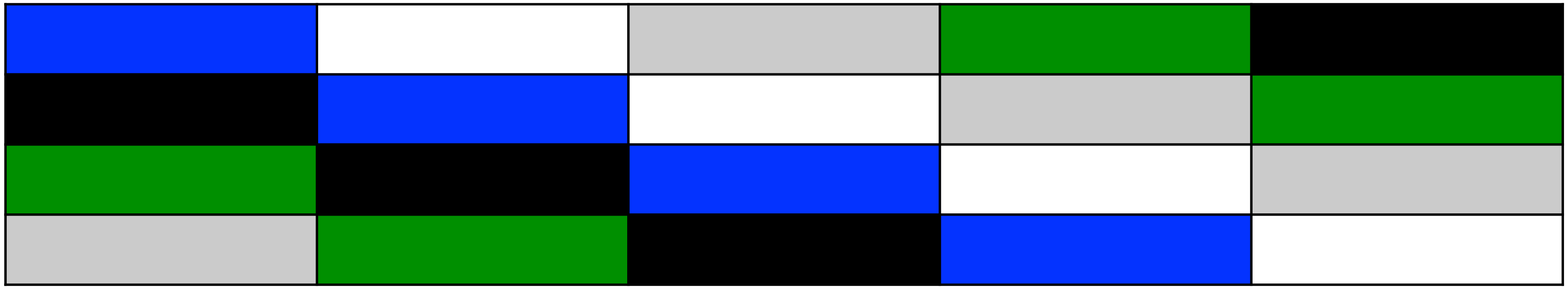
Blocks

- Experiments are typically divided into blocks
- Blocks may be used to collect replicates accounting for unknown systemic errors, or random ones

Blocks

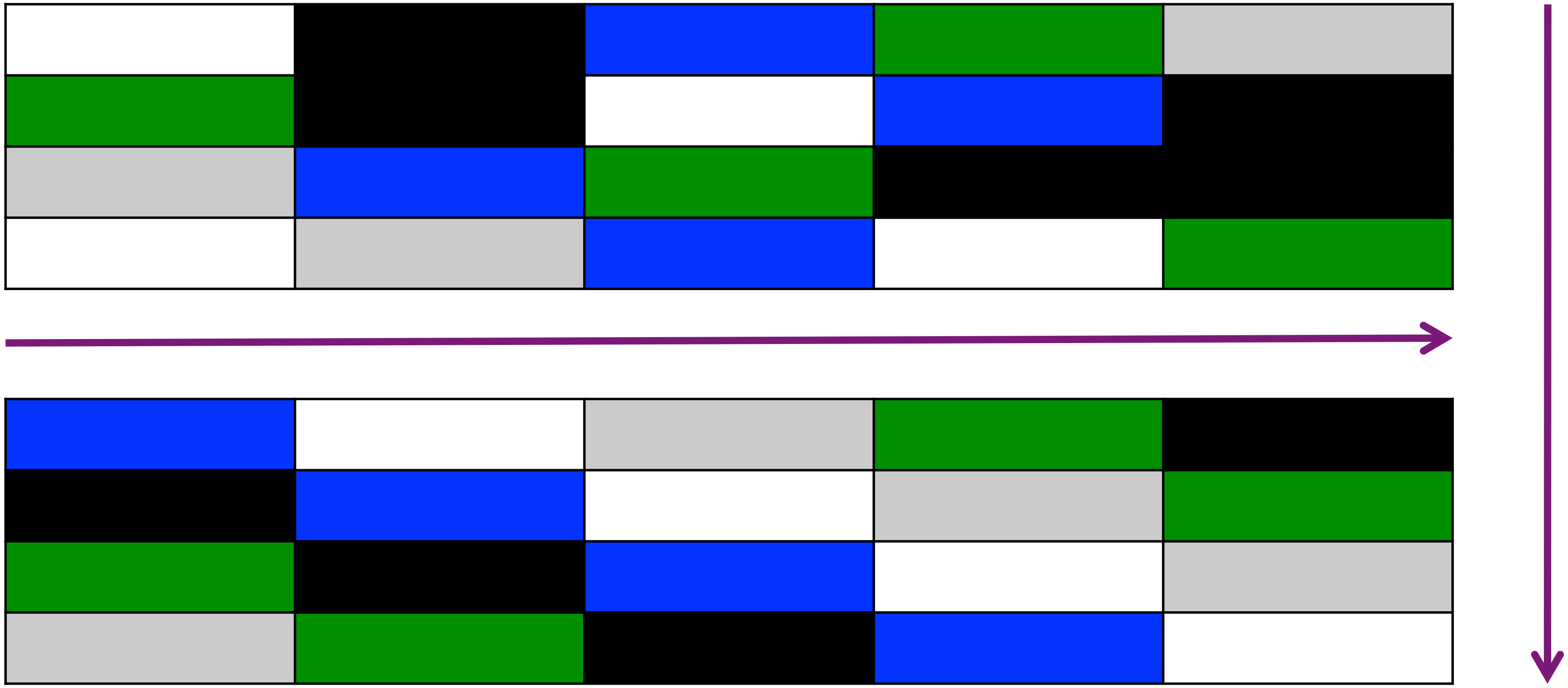
Where treatment levels are used, they can be arranged onto blocks randomly or regularly.





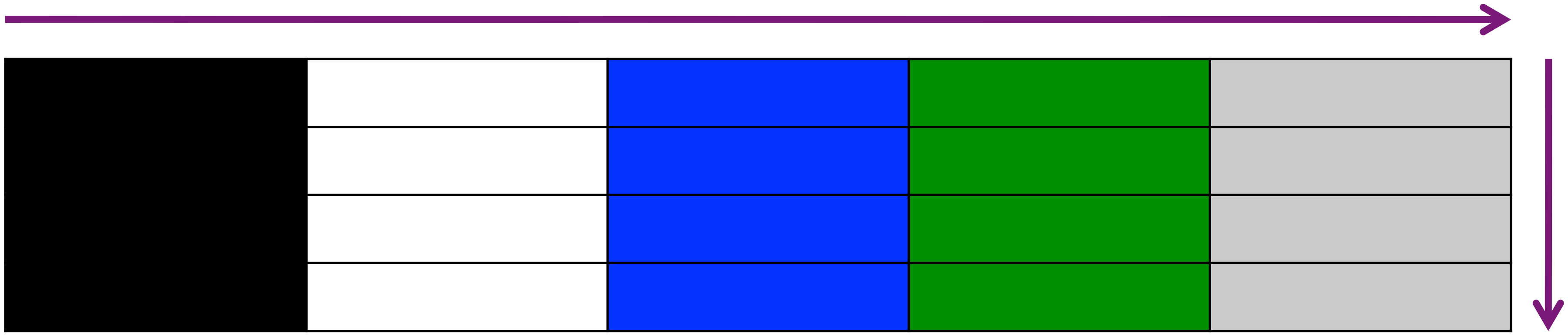
Blocks

If systemic confounding variables are known to exist but aren't of interest, we can use block effect to account for them – if **appropriately laid out…**

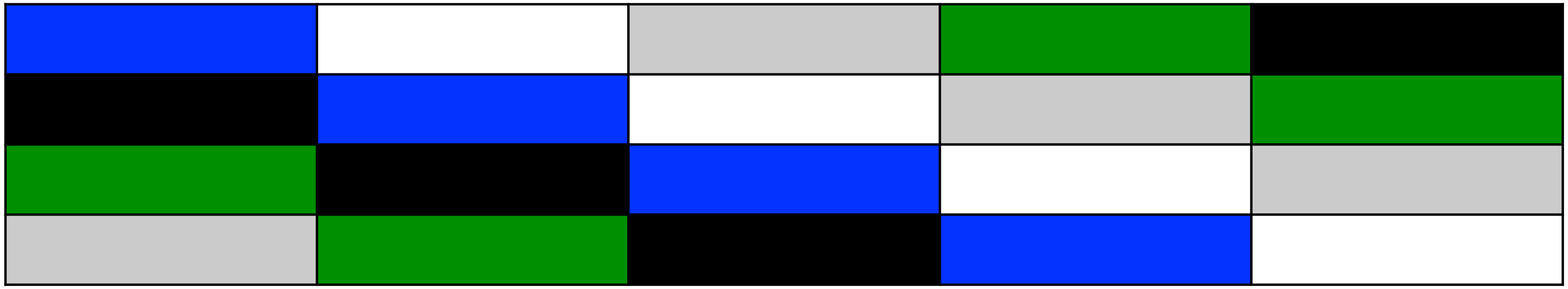


Blocks

… or not



If we want to account for **unknown** random effects we can use a random effects model



Orthogonality

- We get our information by comparing levels of our different factors, e.g. drug Hi/Low vs age Old/Young
- This is why we may refer to these as *contrasts*
- Ideally we want every possible combination to be represented, and with equal samples
- This is the ideal of **orthogonality** and simplifies analysis

Factorial designs

- Most of our designs aim to be factorial, e.g. we have multiple levels of one or more treatments, and we collect/test them simultaneously
- This (hopefully) controls for some of the variation that could arise if we collected and analysed data sequentially
- Big genomic data can violate this fairly frequently, e.g. several transcriptomes collected and sequenced over a year

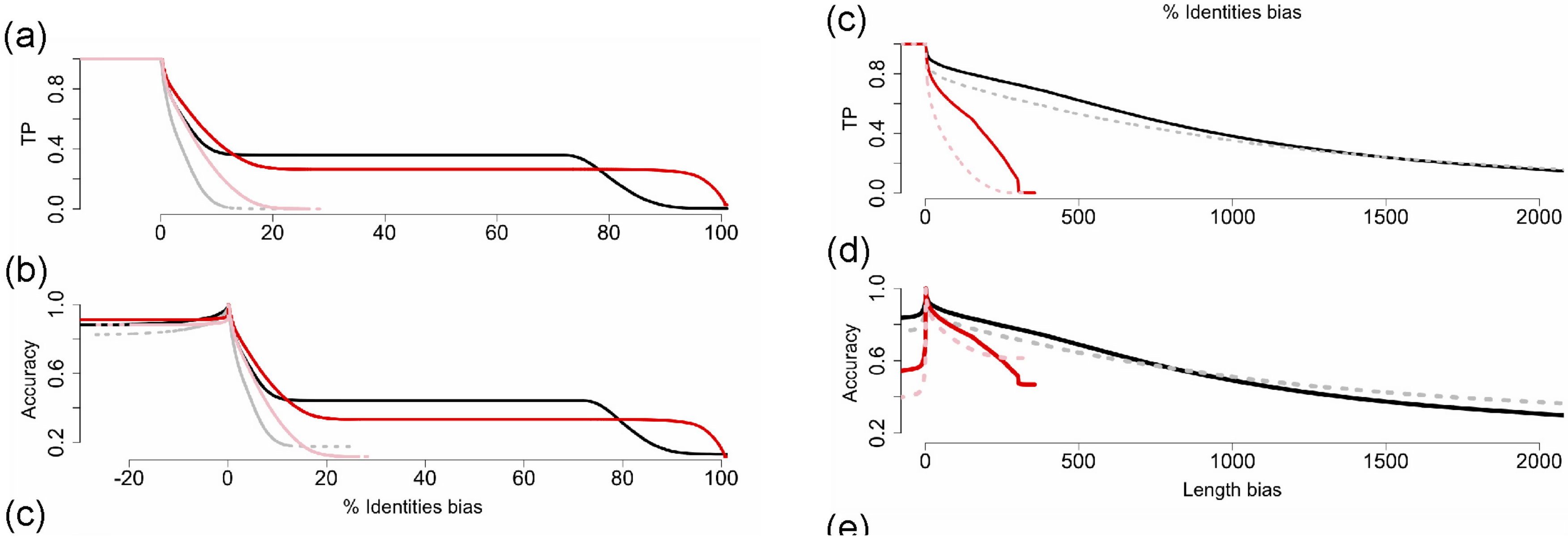
Nested designs

- We may want/need to perform a nested experiment
- In a nested experiment, replicates at different levels

Types of error and statistic choice

	There is an effect	There isn't an effect
We detect an effect	☺ Power; 1-β	☹ Type I rate (α)
We don't detect an effect	☹ Type II rate (β)	☺

Figure 2



Types of error and statistic choice

	There is an effect	There isn’t an effect
We detect an effect	☺ Power; 1-β	☹ Type I rate (α)
We don’t detect an effect	☹ Type II rate (β)	☺

- Not all statistics are created equal: error rates may vary
- Neither are all effects: a statistic which efficiently detects with large effects may perform poorly with weak ones, and vice versa
- To compare statistics we may use a ROC plot, or power curve

Power and effect size

- Based on the assumed power and the expected effect size, we can calculate the sample size needed to detect an effect (if one is there)
- Equivalently, if we have a finite sampling resource, and know which approach we will use, we can determine what magnitude effects we will realistically be able to detect.

Null model choice

- The comparison between a/the null mode is our primary means for assigning significance to findings
- Only works if null is valid
- Valid nulls should be as simple as possible (but no simpler)
- We *must* state the null model before we get to work collecting data/designing work

False discoveries and multiple tests

- Multiple tests or linked *p*-values carry an inherent risk that we wrongly reject the null hypothesis.
- Recall that *e.g.* ‘$p \leq 0.05$’ is equivalent to $P(D|H_0) = 0.05$ = 5%
- 20 x 5% = 100%(!)
- Corrections:
  - Raise ‘significance’ threshold ($p \leq 0.001$)
  - Adjusted / synthetic *p*-values (K-S; Benjamini-Hochberg
  - Explicitly combine models to eliminate repeated tests in the first place

Uses of simulation

We love to simulate. We may use simulation to:

- Evaluate significance by estimating the null distribution, where we cannot compute it directly
- Save time and/or €€€€
- Discover where boundary conditions are, and what goes on there
- Explore the consequences of the fitted model

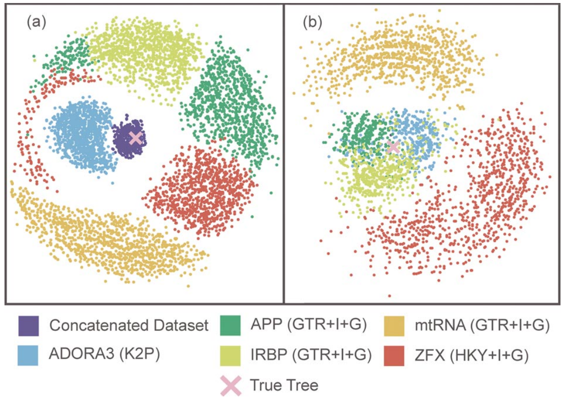
Some studies *may* even be wholly simulational

GWAs

- Genome-wide association studies(GWAS) are *extremely* common
- Compare 100s, or even 1000s of loci for SNP/haplotypes etc, millions of dimensions
- **Very** large numbers of *p*-values effectively, so controlling for multiple tests essential.

PCA and multidimensional reduction

- Many datasets are *extremely* high-dimensional
- Visualising and model selection are extremely hard
- Often most variation contained in a handful of parameters / dimensions
- Techniques to reduce dimensionality *e.g.*:
  - PCA (principal component analysis)
  - MDS (multidimensional scaling)
  - AI-*type* approaches



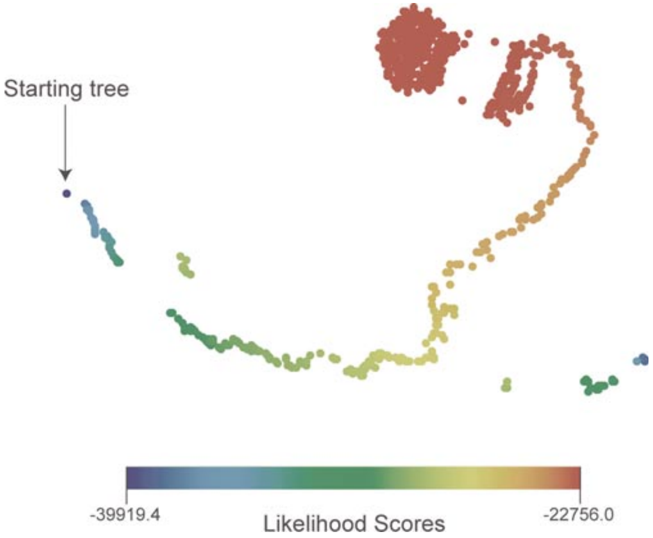


FIGURE 8. Progress in a Bayesian MCMC analysis. The progress in the search can be visualized in the Tree Set Visualization program, as a demonstration of how an MCMC analysis functions. In the visualization, the progress of the chain through tree-space moves from regions of low optimality scores (blue) to regions of high optimality scores (red).

Validation and ‘ground truthing’

- Frequently we may be developing a new model to fit unusual data
- Great. But remember to keep checking against intuition / previous / partial results, especially if high-dimensional
- Previous results, predictions, slices of the data and boundary cases can all help reassure us we’re not *bonkers*

Reproducibility

- Everyone wants to live longer. Datasets aren't any different
- Reproducing results is ***central*** to the scientific methods
- Be extremely suspicious of apparently 'landmark' studies which are hard to reproduce
- Applies to software, environments, etc
- Also applies to model selection if done using *in silico* criteria/algorithms ('best' model selection should be stable/robust)
- As you prepare to publish your Big Finding, make sure
  - All code is accessible and documented;
  - Data **and metadata** available;
  - Methods are clearly described, including software versionsing and dependencies

Summary

- Experimental design is the biggest factor in what we can infer
- Power, sensitivity, and effect size can all help us calculate samples needed
- We must have a valid null
- We need to select models, checking assumptions
- Beware of multiple tests, whatever the context
- If working with highly multidimensional data, keep constantly validating results
- Do reproducible science