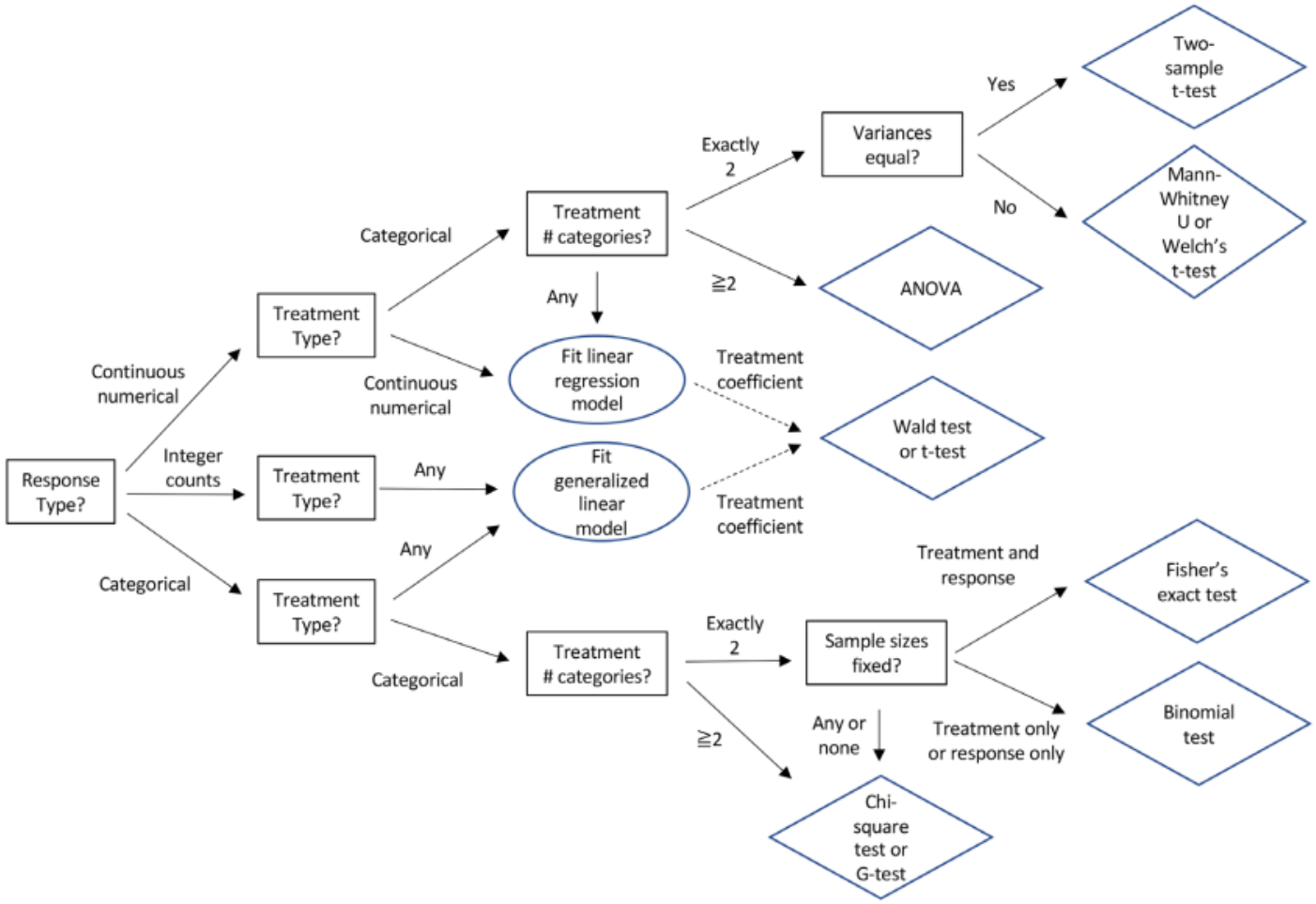
Statističko zaključivanje – praktični primjeri

26.01.2024.

Rosa Karlić

Kako izabrati statistički test



Povezani i nepovezani podaci

- Predstavljaju li observacije u sljedećim studijama povezane (paired) ili nepovezane (independent) podatke?
- a) Procjenjujemo razliku u ocjenama učenika u istom razredu na početku i na kraju školske godine.

- b) Želimo procijeniti postoji li statistički značajna razlika u ocjenama iz statistike učenika koji pohađaju školu X i školu Y. Za potrebe analize uzmemo nasumičan uzorak od po 100 studenata iz svake škole.

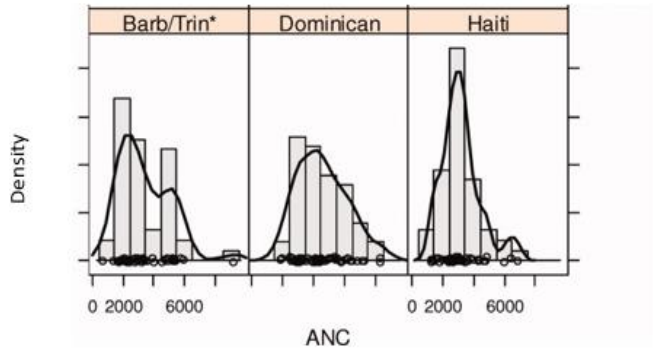
- **Proveli smo eksperiment kako bismo izmjerili i usporedili učinkovitost različitih dodataka prehrani na brzinu rasta pilića. Novoizlegle piliće nasumično smo podijelili u dvije skupine i svaka skupina je primala drugačiji dodatak prehrani tijekom 6 tjedana. Nakon toga smo izvagali piliće u svakoj skupini i izračunali sljedeće numeričke sažetke**:

skupina	mean	sd	n
skupina A	321.48	70.54	40
skupina B	157.33	34.94	40

- Koji bismo test koristili kako bismo utvrdili postoji li statistički značajna u težini pilića ovisno o dodatku prehrani koji su primali (pretpostavite da su težine pilića normalno distribuirane).
  - A. t-test za povezane uzorke
  - B. Welch test
  - C. Wilcoxon-Mann-Whitney Rank Sum test
  - D. t-test za dva neovisna uzorka – jednake varijance

- Provedeno je istraživanje čiji je cilj usporedba razine krvnog tlaka kod strogih vegetarijanaca (SV), laktovegetarijanaca (LV), koji jedu mliječne proizvode, ali ne i drugu hranu životinjskog porijekla, te ljudi koji jedu standardnu europsku prehranu (STD). Svim ispitanicima izmjerena je razina sistoličkog krvnog tlaka i uspoređene su vrijednosti za sve tri skupine. Prikladan statistički test za ovu usporedbu je:
-
- A. koeficijent korelacije
- B. hi-kvadrat test
- C. t-test
- D. analiza varijance

- Nizak apsolutni broj neutrofila (absolute neutrophil count) može odgoditi ili spriječiti završetak odgovarajuće kemoterapije i utjecati na preživljanje raka. Budući da je etnička pripadnost također povezana s preživljavanjem, autori studije su uspoređivali ANC u zdravih žena s Barbadosa / Trinidad-Tobaga, Dominikanske Republike i Haitija. Koji test biste koristili kako biste utvrdili postoji li statistički značajna razlika između razine ANC kod zdravih žena iz Dominikanske Republike i s Haitija.



a) Wilcoxon signed-rank test

b) t-test za neovisne uzorke

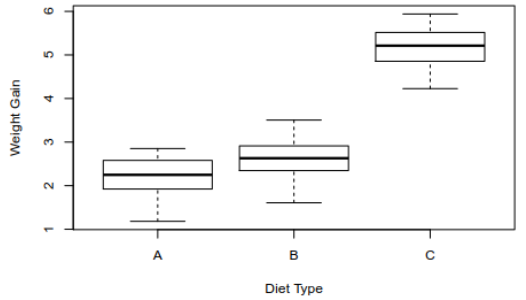
c) t-test za povezane uzorke

d) Wilcoxon-Mann-Whitney Rank Sum test (Mann-Whitney U test)

Proveli smo eksperiment u kojem su ljudi iz tri različite zemlje (Velika Britanija, SAD ili Njemačka) stavljeni na jednu od tri dijete (dijeta tipa A, B ili C) kako bi se potaknulo debljanje. Nakon mjesec dana ponovno ćemo izvagati sve sudionike i zabilježiti koliko su dobili na težini (Weight.gain, vidi boxplot).

Zatim provodimo analizu varijance i Tukey's HSD post hoc test i dobijemo sljedeće rezultate za glavni učinak tipa dijete (Diet.type):

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Weight.gain ~ Diet.type, data = diet.data)
##
## $Diet.type
##          diff         lwr       upr     p adj
## B-A 0.4511019 0.003617744 0.8985861 0.0478053
## C-A 2.9709286 2.523444425 3.4184128 0.0000000
## C-B 2.5198267 2.072342500 2.9673109 0.0000000
```



Na temelju ovih rezultata možemo zaključiti da je na razini značajnosti α = 0.01:

A. Prosječno povećanje tjelesne težine sudionika na tipu dijete A statistički se značajno razlikuje od prosječnog povećanja tjelesne težine sudionika na tipu dijete B i prosječnog povećanja težine sudionika na tipu dijete C.

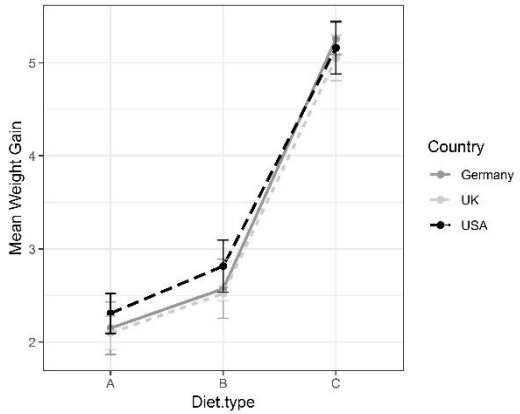
B. Prosječno povećanje tjelesne težine sudionika na tipu dijete C statistički se značajno razlikuje od prosječnog povećanja tjelesne težine sudionika na tipu dijete A i prosječnog povećanja težine sudionika na tipu dijete B.

C. Prosječno povećanje tjelesne težine sudionika na tipu dijete B statistički se značajno razlikuje od prosječnog povećanja tjelesne težine sudionika na tipu dijete A i prosječnog povećanja težine sudionika na tipu dijete C.

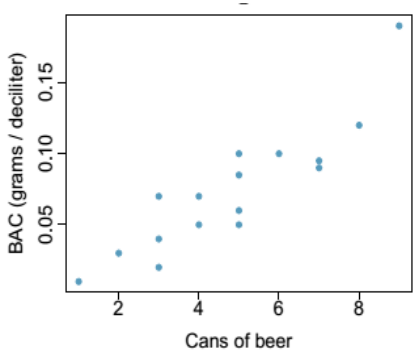
D. Postoji statistički značajna razlika kod svih usporedbi svake sa svakom skupinom (all pairwise comparisons).

Također želimo ispitati potencijalne interakcijske učinke tipa dijete i zemlje u kojoj ispitanik živi (Diet.type i Country). S obzirom na sljedeći grafički prikaz, pretpostavili bismo da će interakcijski učinak:

- Biti začajan
- Neće biti značajan



- Pivo i sadržaj alkohola u krvi. Mnogi ljudi vjeruju da su spol, težina, navike pijenja i mnogi drugi čimbenici puno važniji u predviđanju sadržaja alkohola u krvi (BAC) od pukog razmatranja broja pića koje je osoba popila. Ovdje ispitujemo podatke od šesnaest studenata volontera na Državnom sveučilištu Ohio od kojih je svaki popio nasumično dodijeljen broj limenki piva. Ti su studenti bili ravnomjerno podijeljeni na muškarce i žene, a razlikovali su se po težini i navikama u piću. Trideset minuta kasnije policajac im je izmjerio sadržaj alkohola u krvi (BAC) u gramima alkohola po decilitru krvi. Dijagram raspršenosti i regresijska tablica sažimaju nalaze.



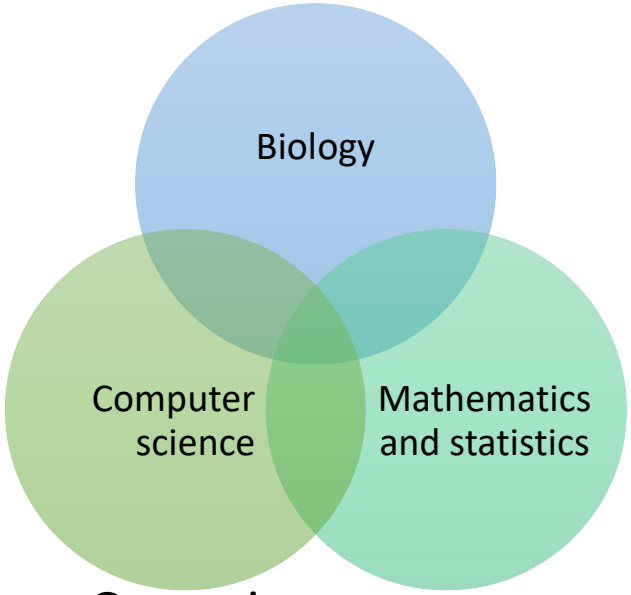
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

(a) Opišite odnos između broja limenki piva i BAC-a.
(b) Napišite jednadžbu regresijske linije. Protumačite nagib i intercept u kontekstu.
(c) Pružaju li podaci čvrste dokaze da je ispijanje više limenki piva povezano s povećanjem alkohola u krvi? Navedite nultu i alternativnu hipotezu, navedite p-vrijednost i iznesite svoj zaključak.
(d) Koeficijent korelacije za broj limenki piva i BAC je 0,89. Izračunajte R2 i protumačite ga u kontekstu.

Statistika u bioinformatici

Neće biti dio ispita

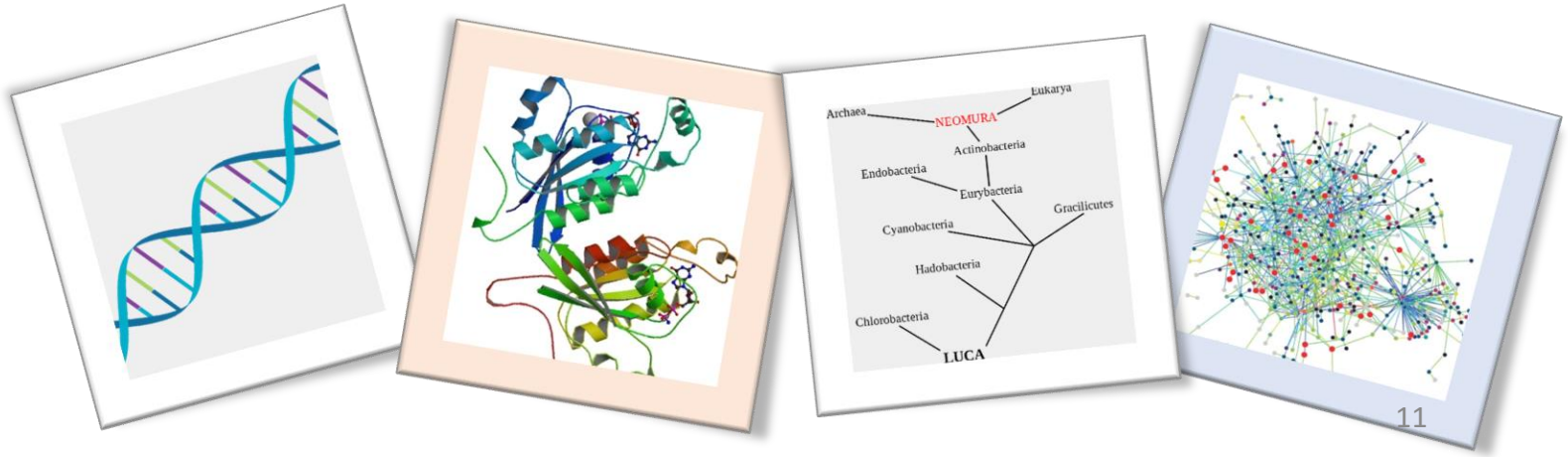
Bioinformatics



- Genomics
- Proteomics
- Evolutionary biology
- Systems biology
- ….

Margaret Oakley Dayhoff

- Dayhoff & Ladley, 1962, *COMPROTEIN: A computer program to aid primary protein structure determination*
- Using computers to determine evolutionary relationships from protein sequences



What is genomics

- Genome – complete set of an organism's genetic material (genes, regulatory sequences, noncoding regions, …)
- Genomics – the study of genomes
- Relatively young discipline which relies on DNA sequencing

  - Comparative genomics
  - Functional genomics
  - Translational genomics

Brief history of DNA sequencing

2001

- Human Genome Project - launched in 1989 –expected to take 15 years
- Competing Celera project launched in 1998
- 1st Draft released in 2000
- “Complete” genome released in 2003

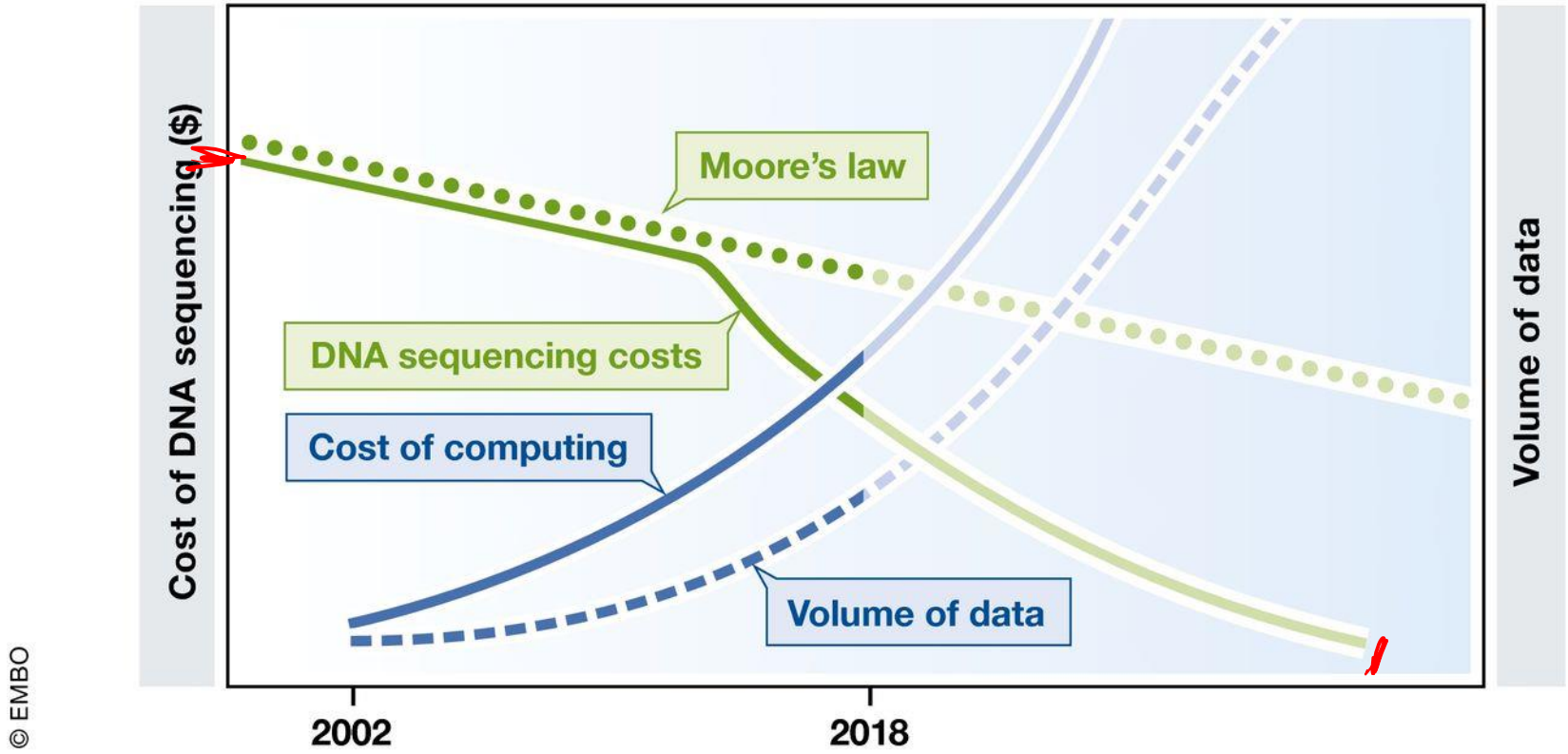
- Cost:
  - Human Genome Project ~$3 billion
  - Celera ~$300 million

Next generation sequencing

- Second generation sequencing
- Masivelly parallel sequencing
- PCR amplification
- Sequencing by synthesis / sequencing by ligation
- Large amounts of short reads

    - Roche/454 FLX: 2004
    - Illumina Solexa Genome Analyzer: 2006
    - Applied Biosystems SOLiD™ System: 2007

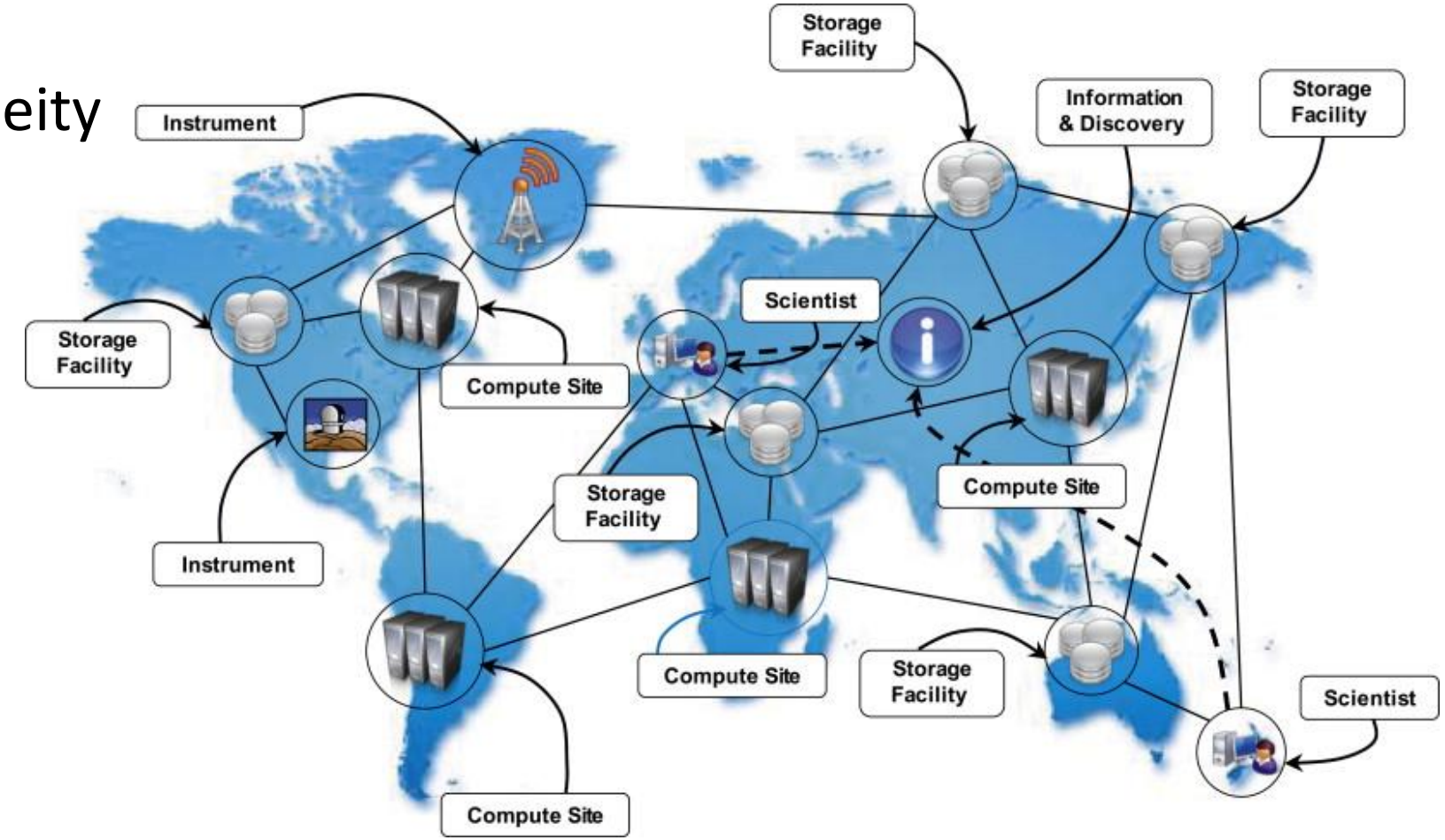
Cost and availability of sequencing



Todd J Treangen, and Mihai Pop EMBO Rep. 2018;19:e47036

Additional challenges

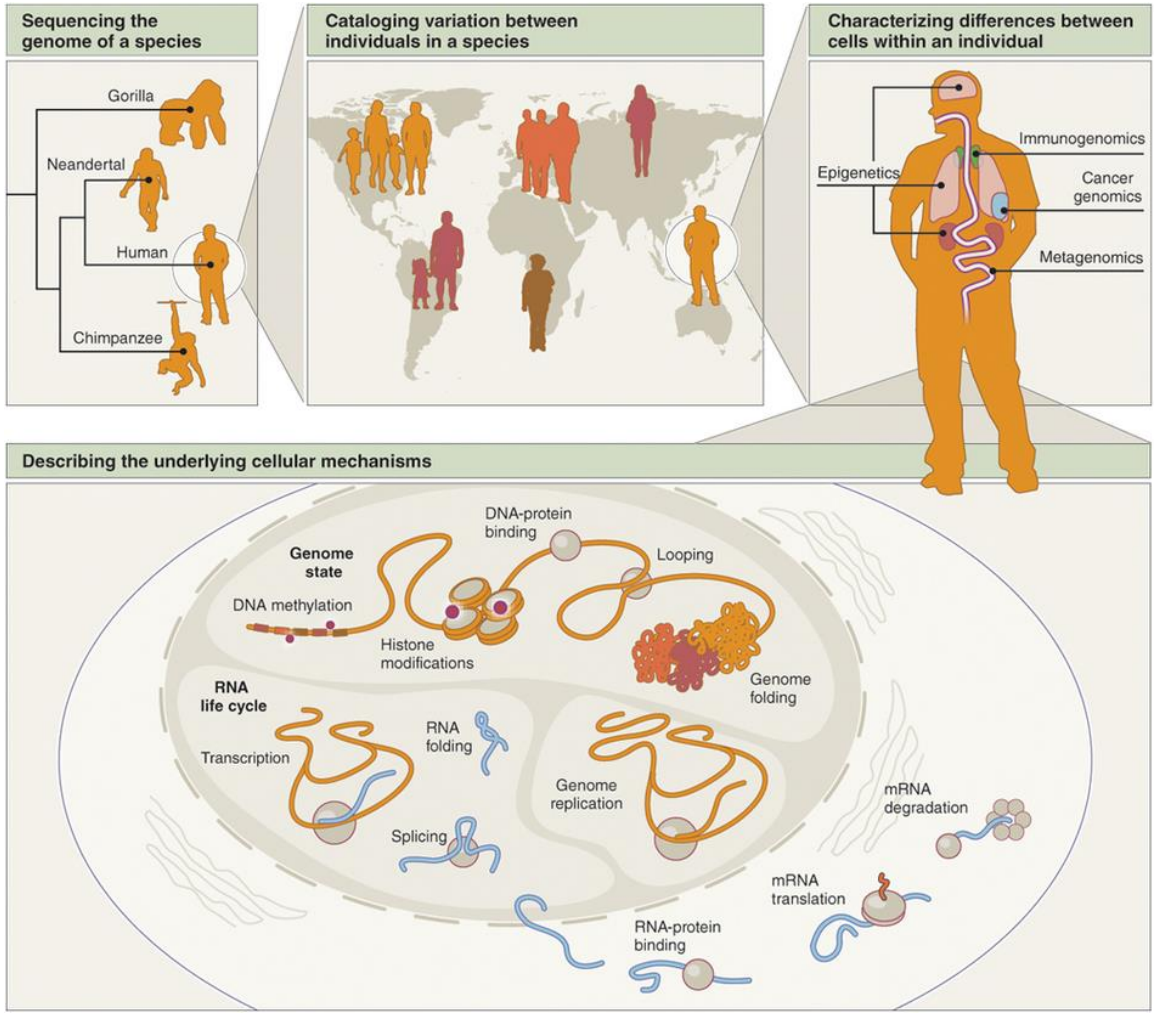
- Data heterogeneity
- Data security



Taken from https://www.sciencedirect.com/topics/computer-science/data-heterogeneity

Applications of NGS

- DNA-Seq
- RNA-Seq
- Methyl-Seq
- ChIP-Seq
- 3D Chromatin structure
- …….



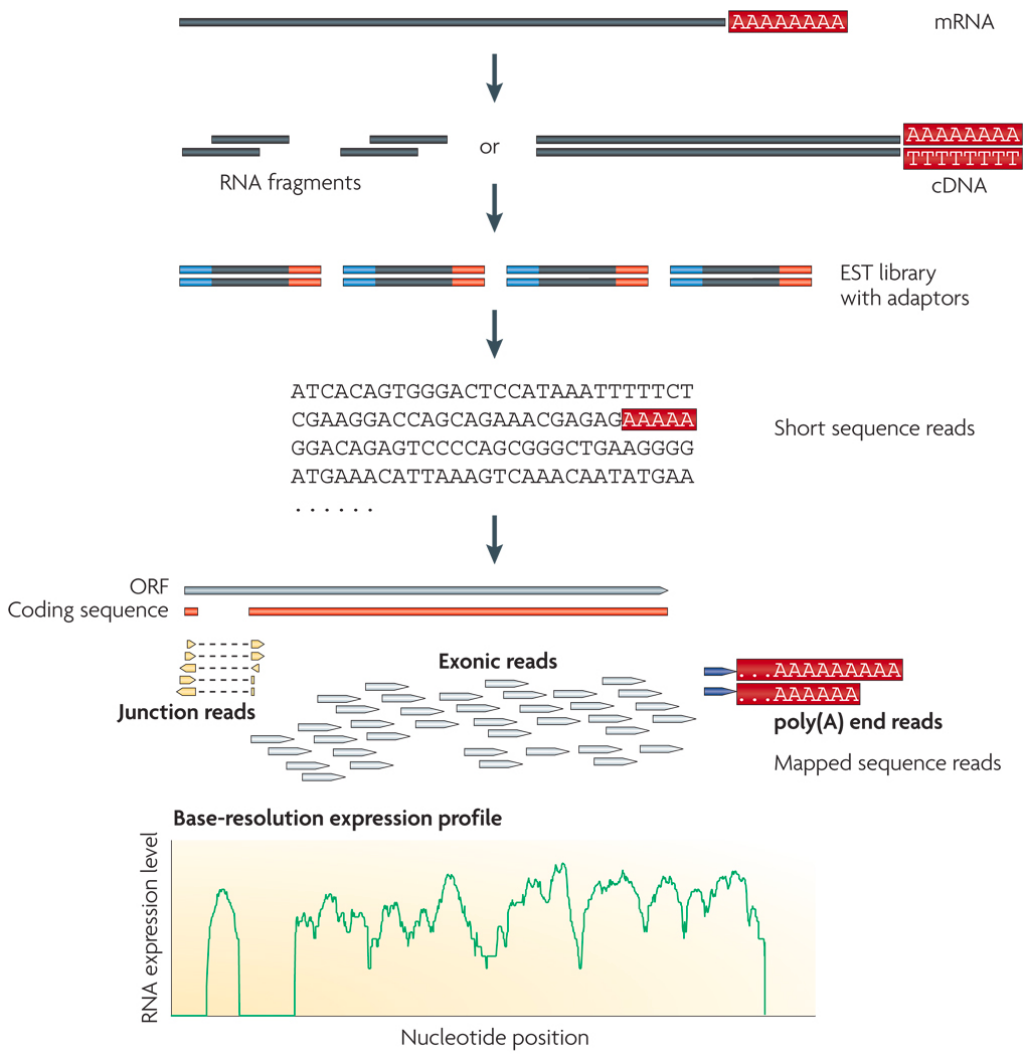
Shendure and Liberman Aiden, 2012, Nature Biotech

Specific types of experiments: RNA-Seq

- Global expression differences
- Annotating genes from a newly sequenced genome
- Discovery of novel genes or transcripts
- Discovery of antisense or other regulatory transcripts
- Variability of isoform expression across conditions

RNA-Seq

- Extract RNA from samples
- Enrich for mRNAs
- Make cDNA from RNA
- Fragment the cDNA
- Library construction
- Sequencing



RNA-Seq

- Number of samples needed (conditions and replicates)
- Number of reads needed
- Length of reads
- Single end or paired end sequencing
- Two methods of analysis:
  - Align then assemble
  - Assemble then align
- Measuring transcript levels by RNA-Seq

RNA-seq - How many reads do I need – sequencing depth

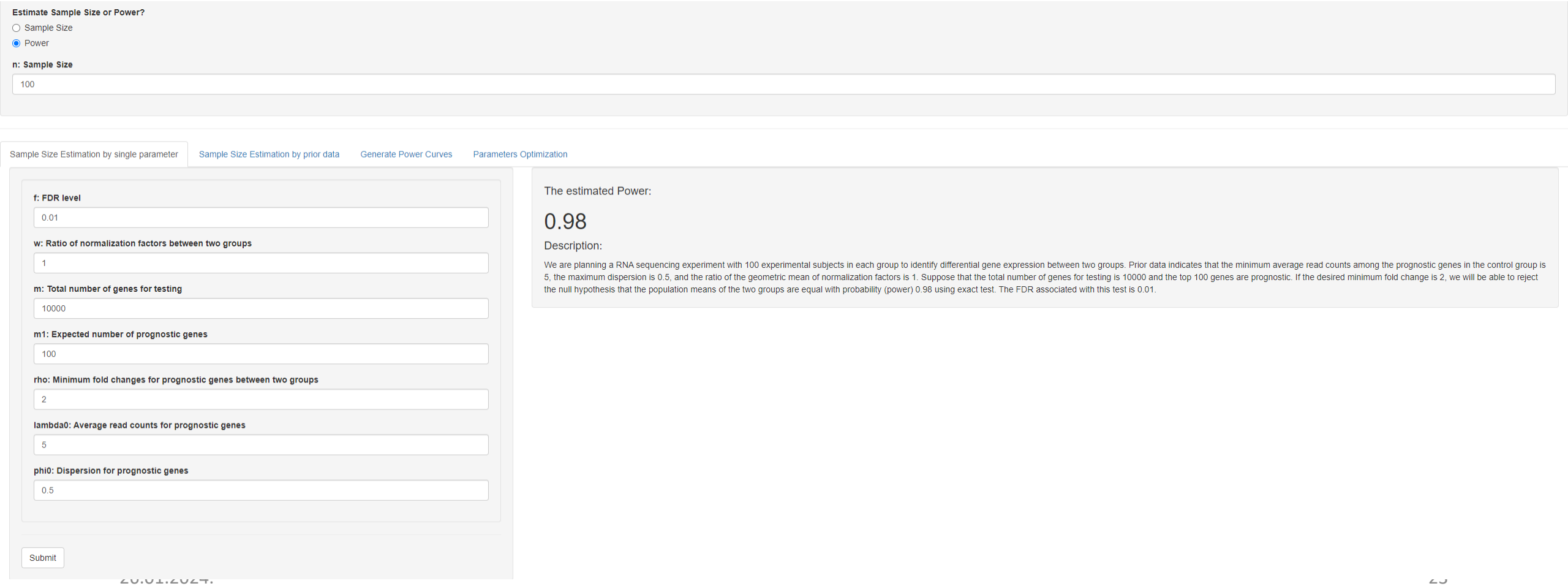
Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

RNA-Seq – how many samples?

- Number of conditions or tissues determined by experiment:
  - For differential expression, what are you comparing
  - For novel discovery, what are the relevant tissues, conditions, or time points?
- Number of replicates determined by biological variability among replicates
- Various tools to estimate optimal power

RNA-Seq: estimate power

https://cqs-vumc.shinyapps.io/rnaseqsamplesizeweb/



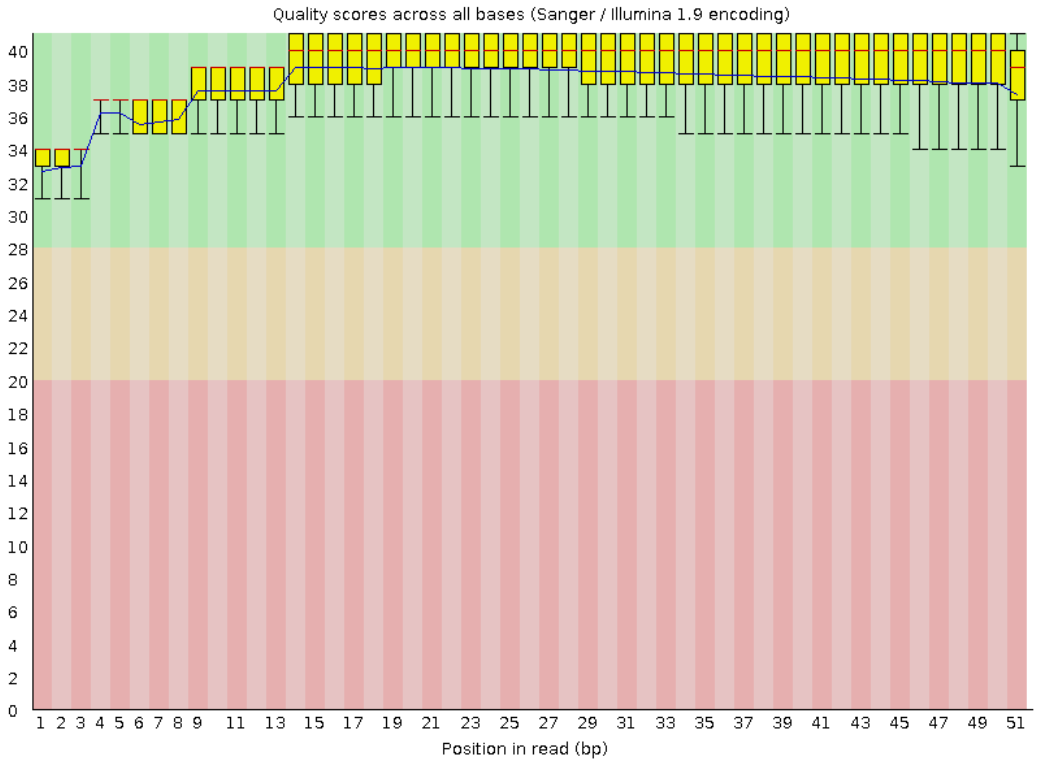
Read quality control

Phred score of a base:

$$Q = -10 * \log10(P)$$

P - estimated probability of a base being wrong

- For example: If a base is estimated to have a 0.1% chance of being wrong, it gets a Phred score of 30.



- Filtering
- Trimming

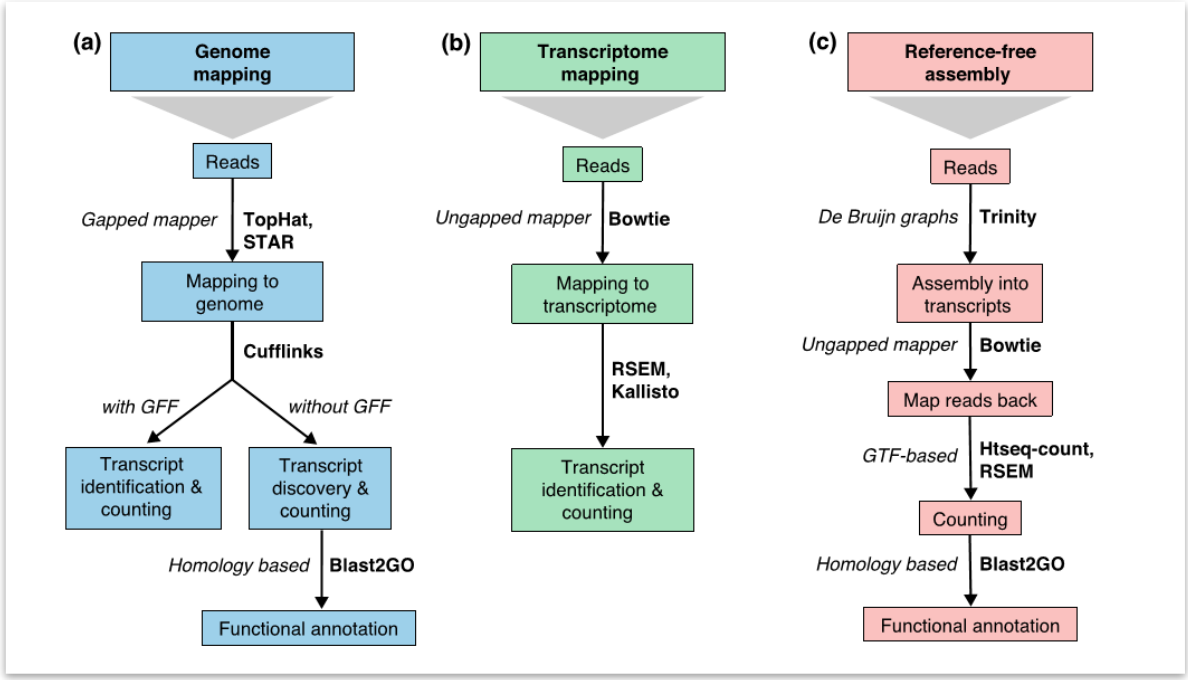
Quality control of aligned reads

Based on:

- mapping quality
- presence of duplicates (possible PCR artifacts)
- overlaps with blacklists
- library complexity
- reads in blacklists

Transcriptome reconstruction

- Genome-guided or genome-independent (*de novo*)
- Transcript identification or discovery



Conesa et al., 2016, Genome Biology

RNA-Seq: estimating transcript levels

- Read count from a transcript is proportional to transcript levels, with two considerations:
  - Transcripts differ in length

  Normalize: divide read count by length in kb
  - Experiments differ in total read count

  Normalize: divide read count by millions total reads
- Resulting value in **RPKM**
- For paired end sequencing, count each fragment once whether one or two read align = **FPKM**

RNA-Seq: estimate transcript levels

- PCR duplicates don’t represent actual counts of RNA fragments, so you need to remove them for quantitation
- Need to be careful about variance:
  - Biological Variance, e.g. Biological variability between replicates of the same conditions may be greater than what is needed to determine statistically significant gene expression changes between conditions
  - Statistical Variance, e.g. When you align reads, they may map to multiple isoforms or multiple paralogs, so you need to assign those reads fractionally to get total transcription levels

DEseq2 – Analysis example

- Test for the effect of dexamethasone (dex) controlling for the effect of different donors' cells (cell)

```
se$cell

## [1] N61311  N61311  N052611 N052611 N080611 N080611 N061011 N061011
## Levels: N052611 N061011 N080611 N61311

se$dex

## [1] untrt trt   untrt trt   untrt trt   untrt trt
## Levels: trt untrt
```

Love et al., RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000Research 2016, 4:1070

DESeq2 – variance stabilization

- For RNA-seq raw counts the variance grows with the mean.
- Variance stabilization strategies:
  - logarithm of the normalized count values plus a small pseudocount
  - regularized-logarithm transformation (rlog2)
  - variance stabilizing transformation (vst package)

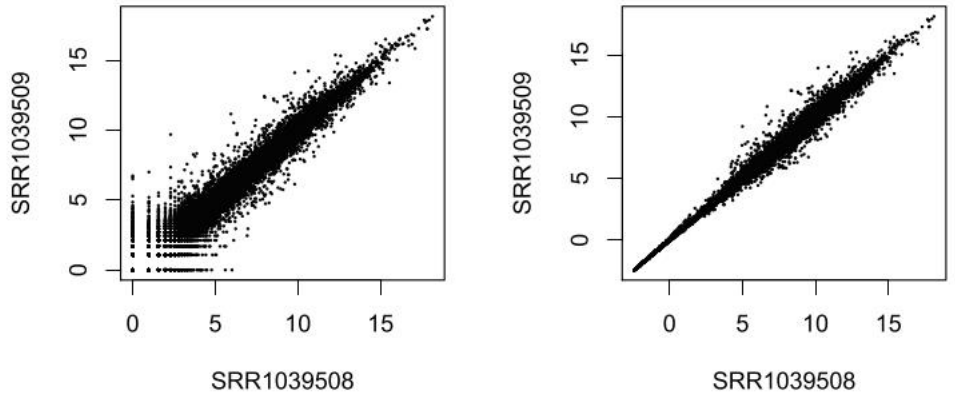


Figure 2. Scatterplot of transformed counts from two samples. Shown are scatterplots using the log2 transform of normalized counts (left side) and using the rlog (right side).

Love et al., RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000Research 2016, 4:1070

DESeq2 – clustering of samples

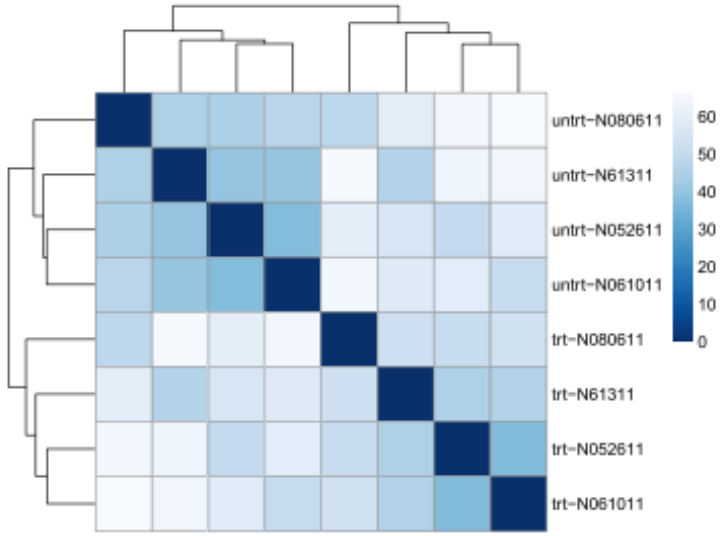


Figure 3. **Heatmap of sample-to-sample distances using the rlog-transformed values.**

Love et al., RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000Research 2016, 4:1070

DESeq2 – PCA plot of samples

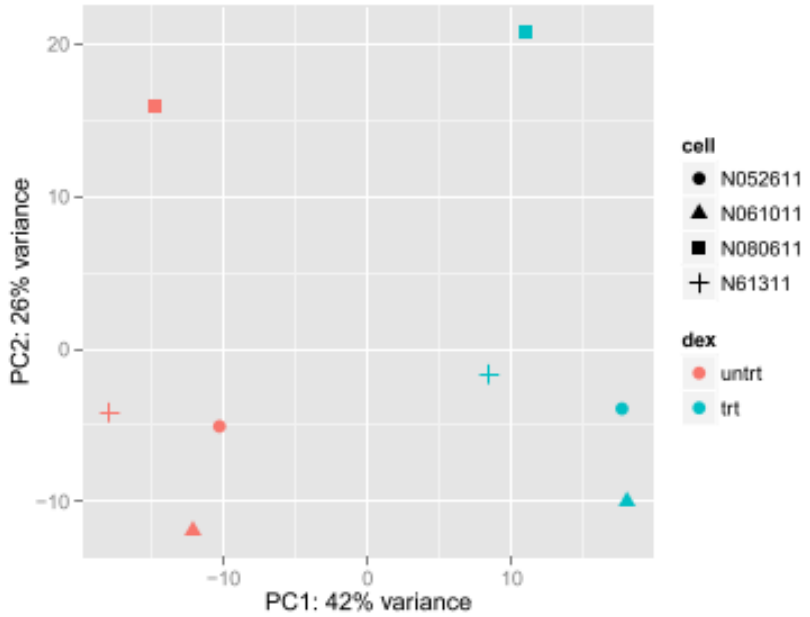


Figure 6. **PCA plot using the rlog-transformed values with custom *ggplot2* code.** Here we specify cell line (plotting symbol) and dexamethasone treatment (color).

Love et al., RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000Research 2016, 4:1070

DESeq2 results

Down-regulated genes

```
## log2 fold change (MAP): dex trt vs untrt
## Wald test p-value: dex trt vs untrt
## DataFrame with 6 rows and 6 columns
##                 baseMean log2FoldChange     lfcSE      stat       pvalue        padj
##                <numeric>      <numeric> <numeric> <numeric>    <numeric>   <numeric>
## ENSG00000162692 508.17023     -3.452454 0.1763751 -19.574503 2.551125e-85 3.460700e-82
## ENSG00000146006  46.80760     -2.856273 0.3366877  -8.483451 2.186122e-17 1.073879e-15
## ENSG00000105989 333.21469     -2.850960 0.1754638 -16.248133 2.302720e-59 1.194366e-56
## ENSG00000214814 243.27698     -2.759539 0.2224907 -12.402938 2.519140e-35 4.113429e-33
## ENSG00000267339  26.23357     -2.743928 0.3511985  -7.813041 5.582443e-15 2.182846e-13
## ENSG00000013293 244.49733     -2.646116 0.1981216 -13.356020 1.092517e-40 2.240295e-38
```

Up-regulated genes

```
## log2 fold change (MAP): dex trt vs untrt
## Wald test p-value: dex trt vs untrt
## DataFrame with 6 rows and 6 columns
##                 baseMean log2FoldChange     lfcSE      stat       pvalue        padj
##                <numeric>      <numeric> <numeric> <numeric>    <numeric>   <numeric>
## ENSG00000179593  67.24305      4.880507 0.3308119   14.75312  2.937594e-49 9.418996e-47
## ENSG00000109906 385.07103      4.860877 0.3321627   14.63403  1.704000e-48 5.181040e-46
## ENSG00000152583 997.43977      4.315374 0.1723805   25.03400 2.608143e-138 4.599460e-134
## ENSG00000250978  56.31819      4.090157 0.3288246   12.43872  1.610666e-35 2.679631e-33
## ENSG00000163884 561.10717      4.078073 0.2103212   19.38974  9.421379e-84 1.038413e-80
## ENSG00000168309 159.52692      3.991146 0.2547755   15.66534  2.610147e-55 1.180255e-52
```

We need to perform multiple hypothesis testing!

Love et al., RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000Research 2016, 4:1070

DESeq2 - Results

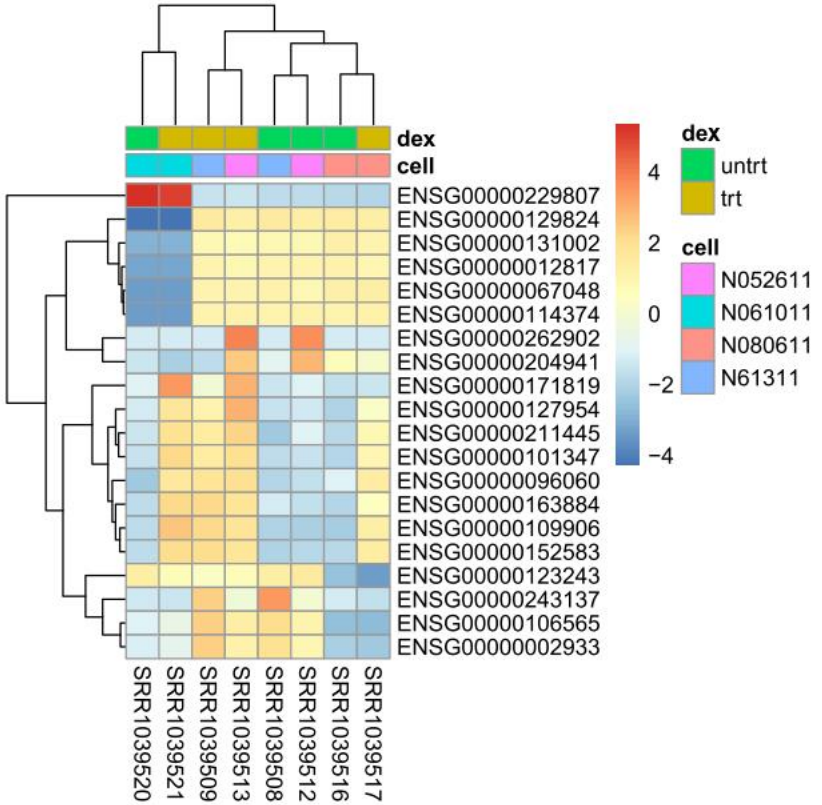


Figure 13. Heatmap of relative rlog-transformed values across samples. Treatment status and cell line information are shown with colored bars at the top of the heatmap. Note that a set of genes at the top of the heatmap are separating the N061011 cell line from the others. In the center of the heatmap, we see a set of genes for which the dexamethasone treated samples have higher gene expression.

Love et al., RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000Research 2016, 4:1070