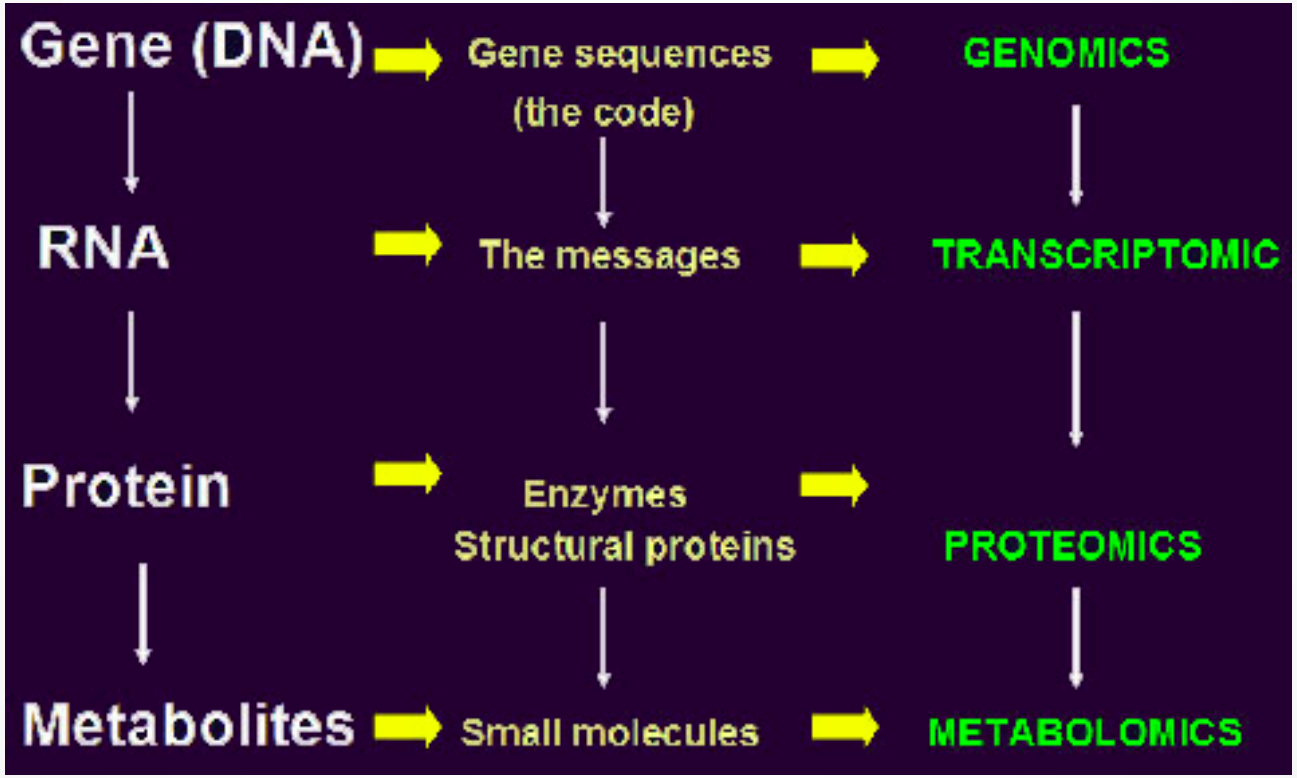
Transzkriptomika, RNA-seq

Ari Eszter, PhD
ELTE Genetikai Tanszék
arieszter@gmail.com

Omics...


Gene (DNA)
Gene sequences
(the code)
GENOMICS
RNA
The messages
TRANSCRIPTOMIC
Protein
Enzymes
Structural proteins
PROTEOMICS
Metabolites
Small molecules
METABOLOMICS

Miről lesz szó?

- Elmélet:
  - A transzkriptomikáról általában
  - Az RNA-seq módszer rövid ismertetése
- Gyakorlat:
  - Readek illesztése a genomra (read mapping)
  - Readek számlálása (counting)
  - Differenciáltan expresszálódó gének keresése

Transzkriptomika

- Transcriptome, Transzkriptom:
  - egy faj (sejt, szövet, szerv, egyed, populáció) összes transzkriptuma
  - meghatározott időben és környezetben a genomról átíródó RNS-ek összessége
- Transzkriptomika
  - A génexpresszió vizsgálata
  - A transzkripciós különbségek feltárása

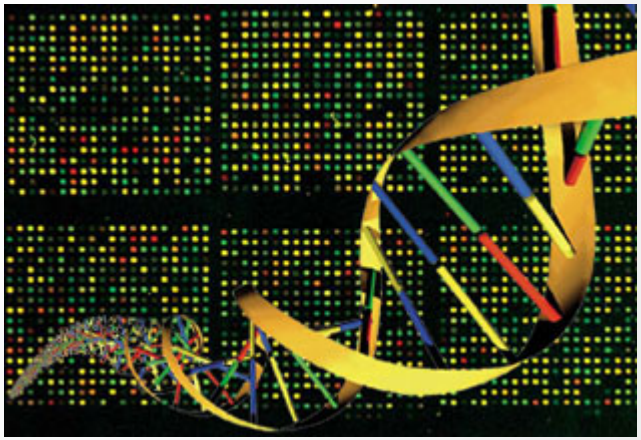
Transzkriptomika

- Alapfeltevés: az mRNS-szint jellemzi az adott gén expressziós szintjét és az adott **fehérje** mennyiségét is
  - Ez nem minden esetben igaz (poszttranszkripciós szabályozás)
  - Az RNS szinteket olcsóbb és egyszerűbb mérni, mint a proteomot
- Összehasonlíthatjuk a különböző, különböző állapotú vagy különböző kezelést kapott sejtek, szövetek, egyedek, populációk génjeinek expressziós szintjét.
- Ezekből azután következtethetünk a mögöttes biológiai folyamatokra

A transzkriptomika eredményeinek felhasználási területei

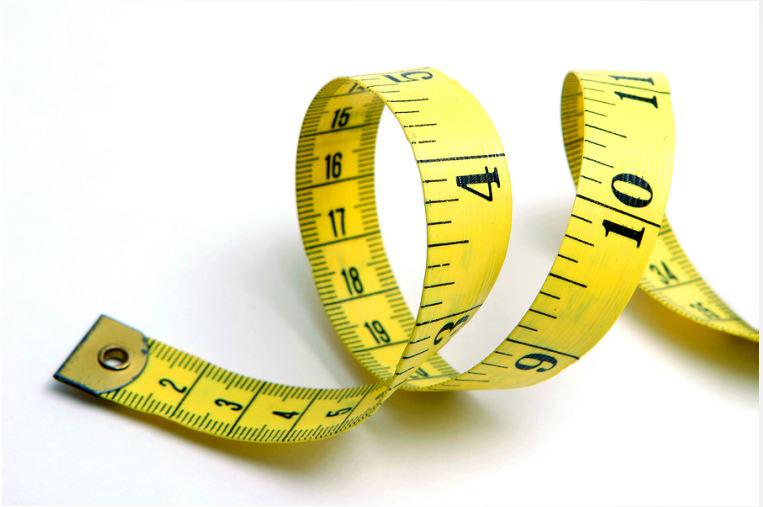
- Genetika
  - gének működése, annak szabályozása
- Genomika
  - gének helyének pontos meghatározása
- Rendszerbiológia
  - együtt kifejeződő fehérjék hálózatának feltárása
- Populációgenetika
  - különbözik-e két populáció génexpressziója?
- Orvostudomány
  - diagnosztika
  - gyógykezelési alapkutatás
- Gyógyszerfejlesztés

...



Módszerek

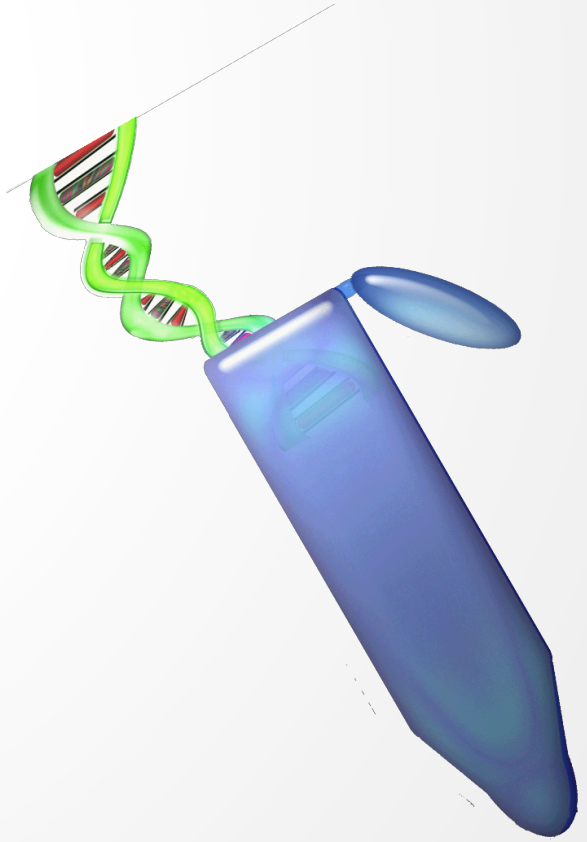
- Mit tudunk mérni?
  - RNS-ek szintje
  - Fehérjék szintje
- Hogyan?
  - Northen blot (1977)
  - Reverse-transcription RT-PCR (1992)
  - Reverse-transcription quantitative qRT-PCR
  - High-throughput módszerek
    - RNS Microarray vagy CHIP (1999)
    - Szekvenálás - RNA-Seq (2008)
    - Protein-array (Fehérje-csippek)



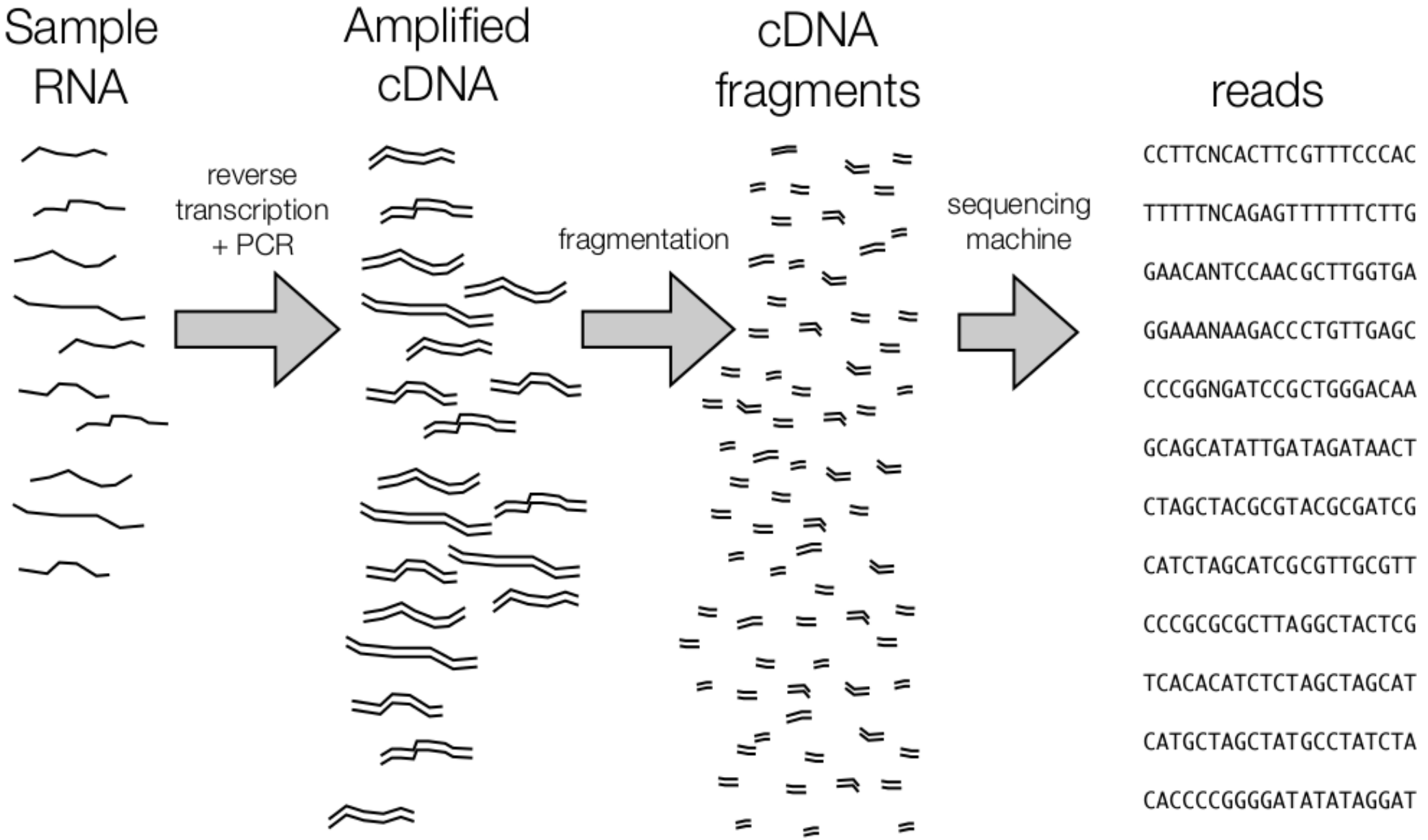
Általános RNA-seq minta preparálás

1. RNS izolálás, rRNS depláció vagy mRNS-ek izolálása (oligo-dT)
2. Hőkezelés → fragmentáció (94°C)
3. cDNS szintézis, RT-PCR amplifikáció random primerekkel
4. Fragmentáció
5. A 3' végek adenilálása
6. Adapter ligálás (+ barcoding)
   → szálspecifikus szekvenálás
7. PCR amplifikáció
8. Tisztítás
9. **Újgenerációs szekvenálás**

Több „read” / transzkript



Általános RNA-seq minta preparálás

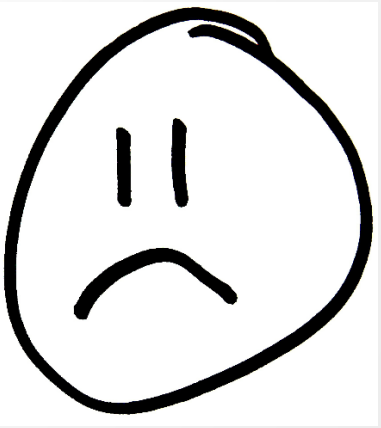


Az RNA-seq előnyei

- Jól reprodukálható, robusztus
- Érzékeny
- Az (m)RNS szintek „direkt” érzékelése
  - Nincs „telítődés”, mint a microarrynél
- Az átíródó RNS-ek szekvenciája is rekonstruálható
- Minden transzkript, még a tudományra újak is mérhetők
- Izoformák és splicing helyek feltárása
  - genom annotáció javítása
- Polimorfizmusok (SNP-k) feltárása
  - allél-specifikus expresszió vizsgálata
- Olyan fajoknál is alkalmazható, melyeknek nincs megszekvenálva a genomja

Az RNA-seq hátrányai

- Drágább ($300 - $1000 / minta), mint a microarray ($100 - $200 / minta)
  - Nagyobb számítási kapacitású számítógépeket igényel az adatok feldolgozása
- Nem detektálja a poszt-transzkripciós módosulásokat
- sem a poszt-transzkripciós szabályozásokat:
  - az mRNS mennyisége nem egyenesen arányos a fehérje mennyiségével
  - szabályozás: miRNA …
- Torzítások: a read-könyvtár mérete, fragment hossz, GC arány...
  - → normalizáció



Az RNA-seq adatok feldolgozása

0. RNS izolálás, szekvenálás → fastq file
1. Minőségellenőrzés, trimming (a rossz minőségű readek, vagy read részletek kiszűrése)
2. Read mapping a referencia genomhoz vagy de novo assembly

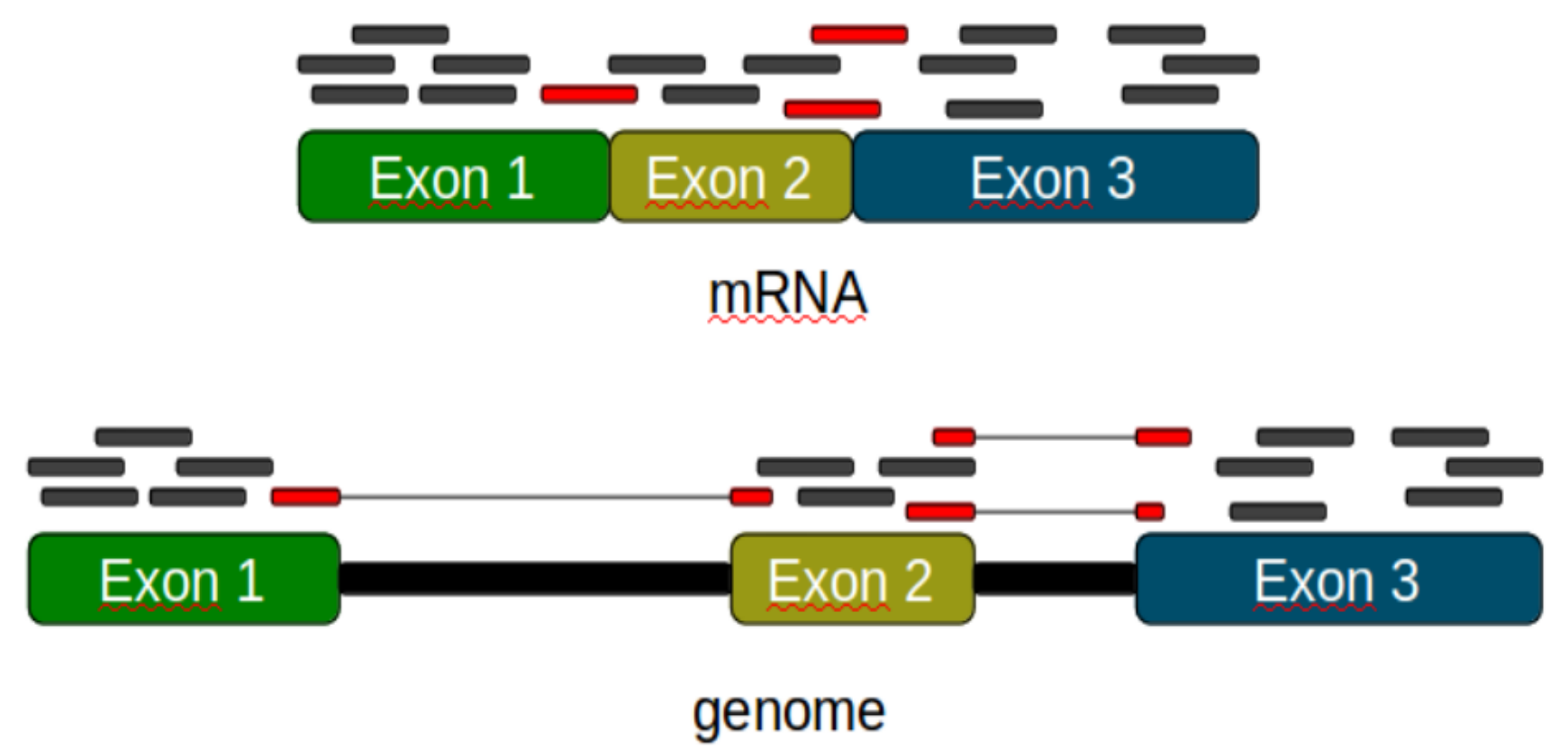
   Mappelés utáni minőségellenőrzés
3. Read counting
4. Kvantitatív analízis: az expressziós szintek összehasonlítása → Differenciál expresszió számítás
5. Funkcionális feldúsulás (Functional enrichment) vizsgálat: GO, útvonalak, TF-ok...

2. Mapping

- read-ek: single- vagy paired-end, szál specifikus vagy nem
- a read-ek illesztése a genom szekvenciához

```
ATTCGATGCTAGGTCGGATGCGTAGATGCATAGGCATGATGCATGGCAT
ATTCGATGCT                                ATAGGCATGA
 TTCGATCCTA                               ATAGGCATGA
   CGATGCTAGG                              TAGGCATGAT
   CGATCCTAGG                                 GCATGATGCT
    GATGCTAGGT
```

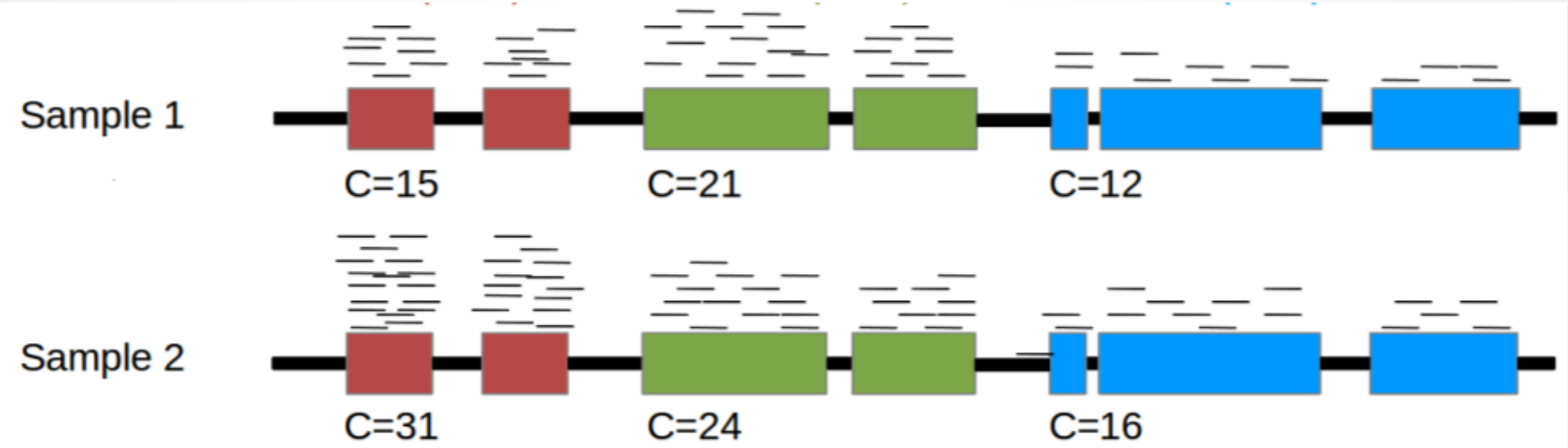
2. Mapping



2. Mapping

- Referencia genom: FASTA
  Szoftverek: GSNAP, STAR, TopHat2, Rsubread, etc.
  - genom indexelés
  - mapping → számításigényes lehet
- Readek szűrése
  - Illeszkedett, nem illeszkedett
  - Egy helyre vagy több helyre illeszkedett
  - A párja is illeszkedett-e, és ugyanarra a kromoszómára
- Readek sorba rendezése (sorting)
  - genomi koordináta vagy read név alapján

3. Read counting



- Hány read (pár) illeszkedett egy-egy génhez?
  - → count táblázat
- Genom annotáció a read számláláshoz: GFF, GTF
  - az exonok, gének és egyéb transzkriptok helyét tartalmazza

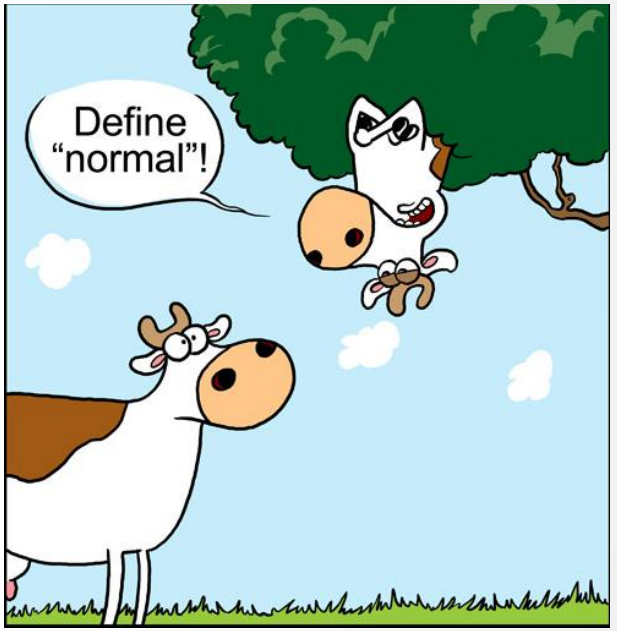
Count táblázat

	F1	F2	F3	F4	M1	M2	M3	M4
ENSG00000127720	14	14	23	16	32	35	10	19
ENSG00000242018	24	16	11	19	21	22	13	6
ENSG00000224440	0	0	0	0	0	0	0	0
ENSG00000214453	0	0	0	0	0	0	0	0
ENSG00000237787	1	0	0	0	0	0	1	0
ENSG00000051596	220	325	450	585	475	294	224	711

...

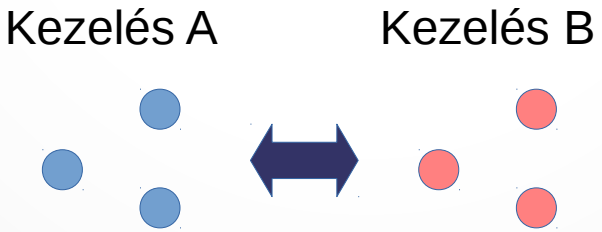
4. DE számítás: Normalizálás

- Normalizálás: az esetlegesen különböző lefedettséggel megszekvenált transzkriptom minták “összehasonlítható szintre hozása”
  - Ha az egyik csoport expresszióját 2x olyan lefedettséggel (alapossággal) mérték, mint a másik csoportét → ez normalizálás nélkül befolyásolná a végeredményt



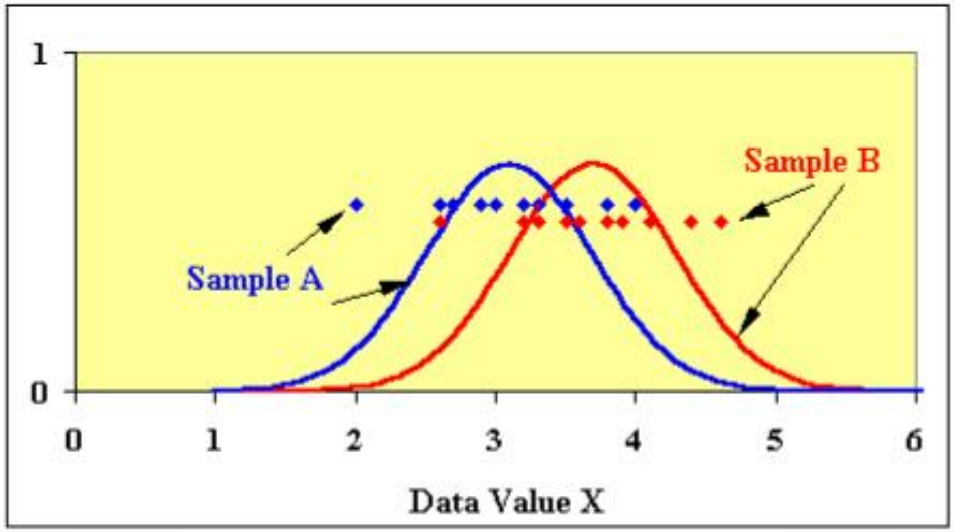
4. DE számítás: Kísérleti elrendezés, kérdésfelvetés

- <u>Két csoport:</u>
  - Kérdés: Mely gének (izoformák, exonok...) expressziója különbözik szignifikánsan a *két* csoport között? → p-érték
  - Azok expressziója melyik irányban változott? → Fold change
  - *páros (pairwise) DE analysis*

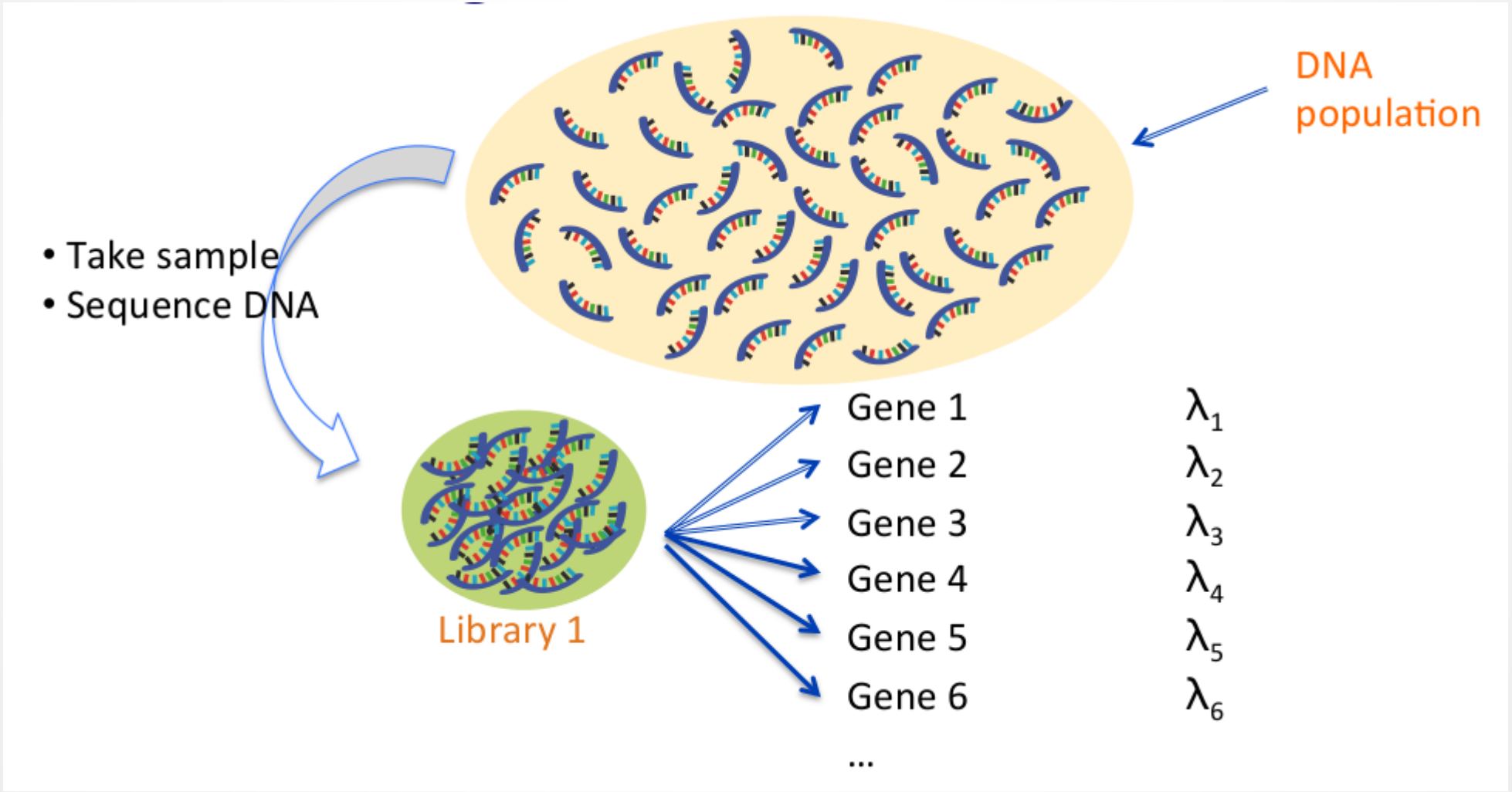


2 mintás t-próba (Welch teszt)

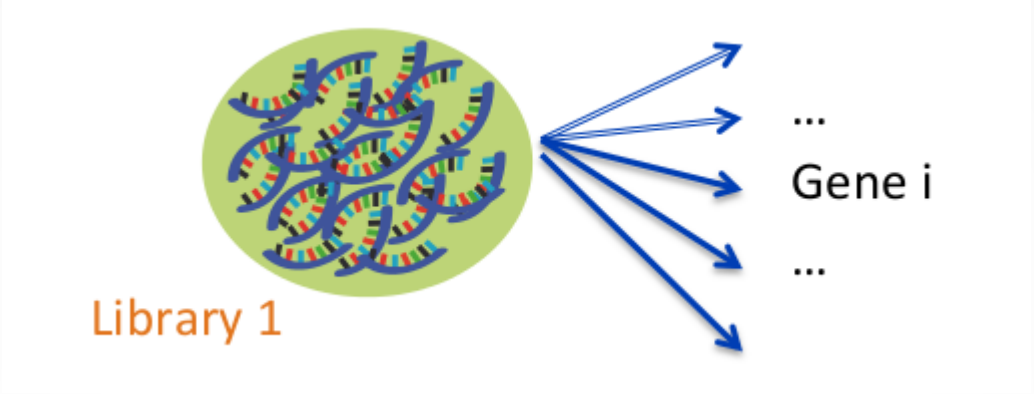
- Génenként megvizsgáljuk, hogy a *replikátok* által kirajzolt 2 eloszlás várható értéke azonos-e (H0)
- p-érték: Mekkora az esélye, hogy H0 helyes?
  - a p-érték egy valószínűség: annak a valószínűsége, hogy a H0 fennállása esetén a puszta véletlen folytán a H0-nak legalább annyira ellentmondó mintát kapunk, mint a ténylegesen megfigyelt minta.
- Feltételek:
  - normális eloszlású változók
  - a szórások ismeretlenek, és nem feltétlenül egyenlők



A read sampling eredénye multinomiális eloszlású

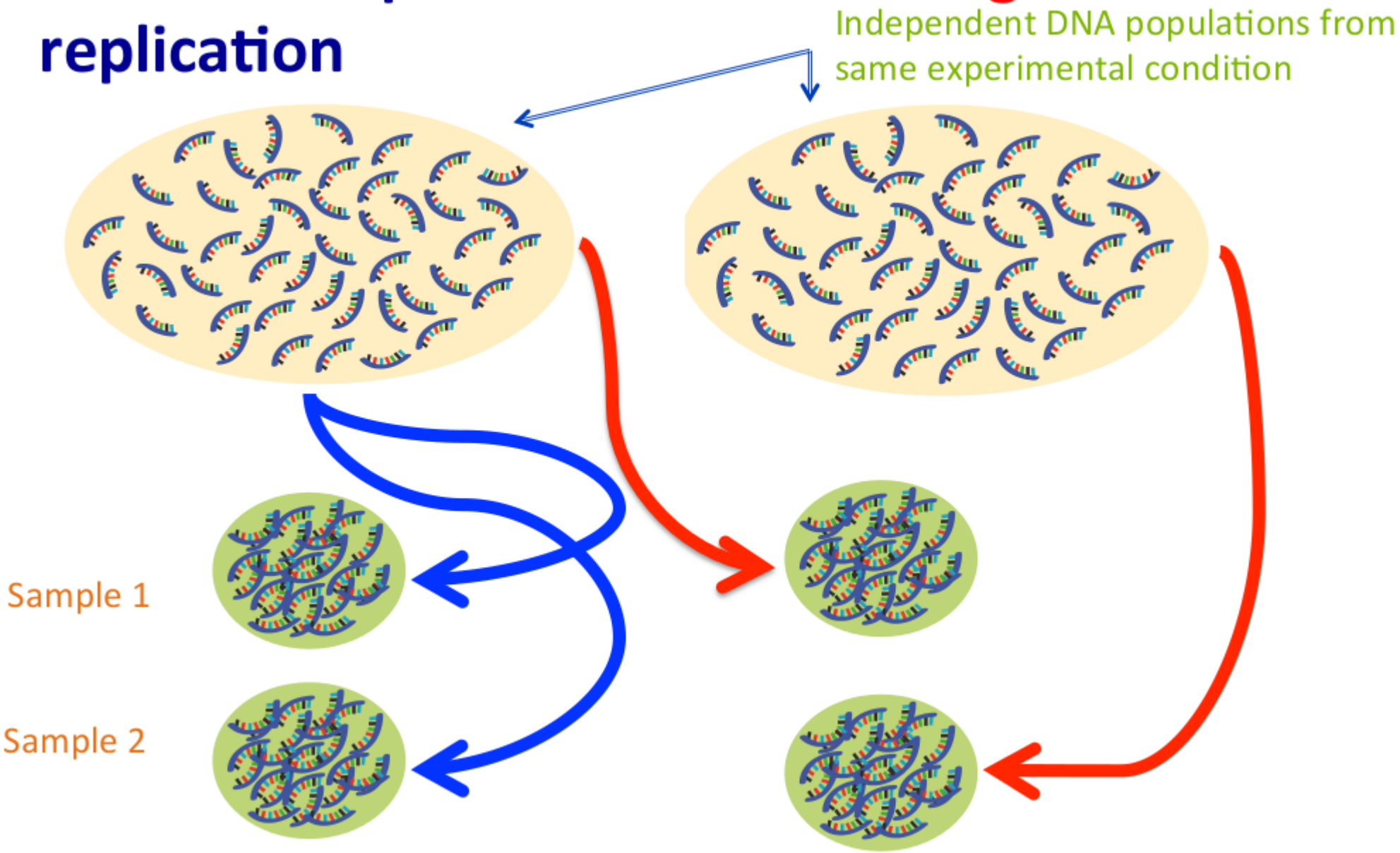


Egy gént vizsgálva binomiális eloszlású (pénz feldobás)

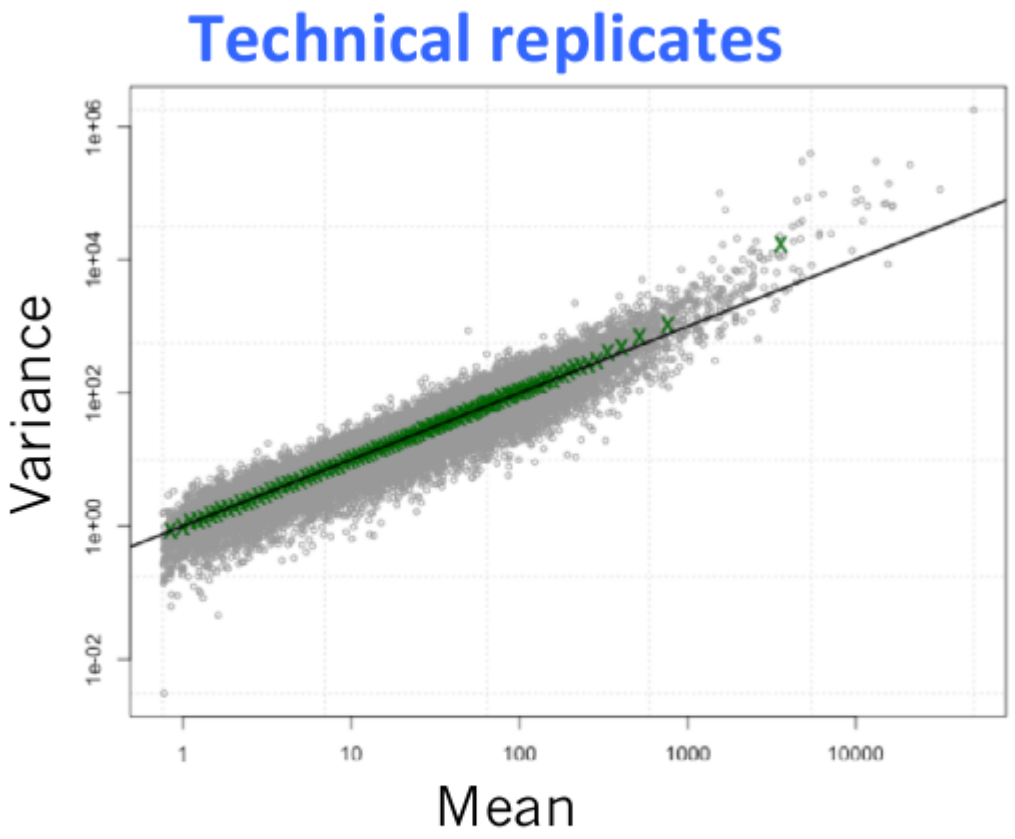


- Y_i ~ Bionomiális(M, λ_i)
  - Y_i a readek megfigyelt száma i gén esetén
  - M a readek összege a mintában (könyvtár méret)
  - λ_i az i gén readjeinek relatív aránya

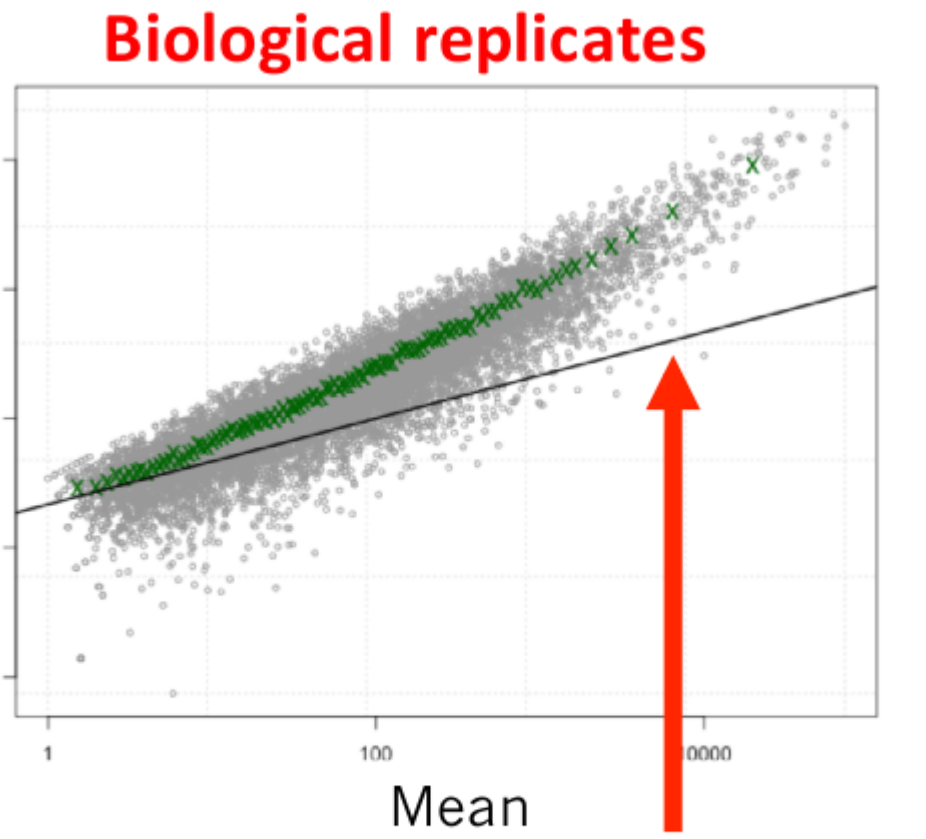
Technical replication versus biological replication



Átlag - variancia plot



Data from Marioni et al. Genome Research 2008



Data from Parikh et al.
Genome Biology 2010

Az RNA-seq-hez nem jó a t-teszt

- A read countok eloszlása nem normál
- Nagyobb a szórásuk, mint az átlaguk
- A Poisson eloszlásnál az átlag = variancia
- A negatív binomiális eloszlásnál a variancia lehet nagyobb, mint az átlag
  - A Poisson- és a negatív binomiális eloszlást akkor használjuk, ha a kérdés „hány. . . ?”, „hányszor. . . ?”, de nem előre rögzített számú *n* megfigyelésből.

Fold change (FC)

- Nagyságrendi változás (nem statisztika):

$$\log 2 \left(\frac{\textit{Egyik csoport countjainak átlaga}}{\textit{Másik csoport countjainak átlaga}} \right)$$

  - +2 v -2 (4*-es expressziós különbség) már nagy logFC-nek minősül

p-érték korrigálás

- A p-érték korrekciója a többszörös tesztelés miatt (multiple testing correction):
  - ugyanazt a mérést végezzük szimultán minden génen - melyek egymástól nem függetlenek
  - Bonferroni, Benjamini-Hochberg
- Ha a korrigált p-érték alacsonyabb, mint 0.05 (5% valószínűséggel fogadjuk el a helytelen hipotézist) → az adott gén expressziója szignifikánsan különbözik a kezelések között

- Ingyenes, nyílt forráskódú és nyíltan fejleszthető projekt
- http://bioconductor.org/
- Több, mint 1000 R csomag:
  - OMICS és mol. biol kutatási adatok feldolgozásához. pl:
    - genom annotációk
    - microarray, RNA-seq
    - enrichment analízis
    - ...

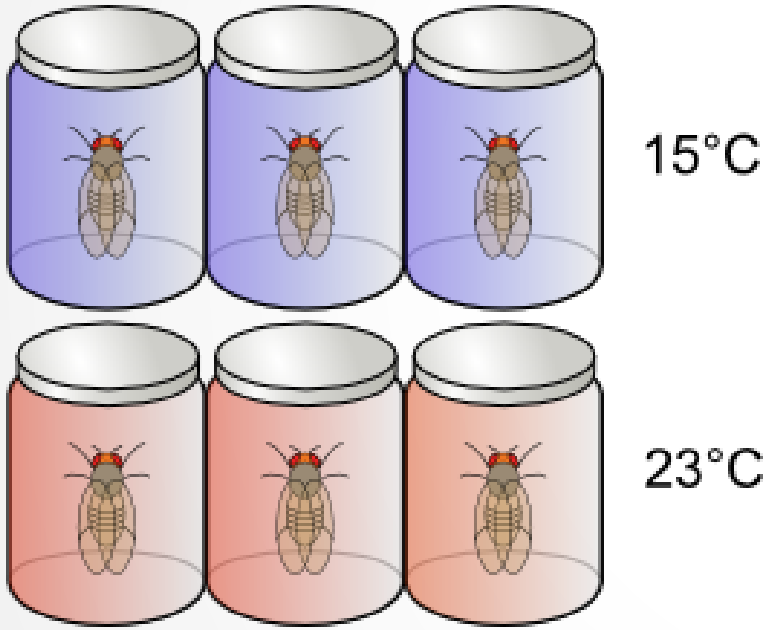
Ajánlott weboldalak, cikkek

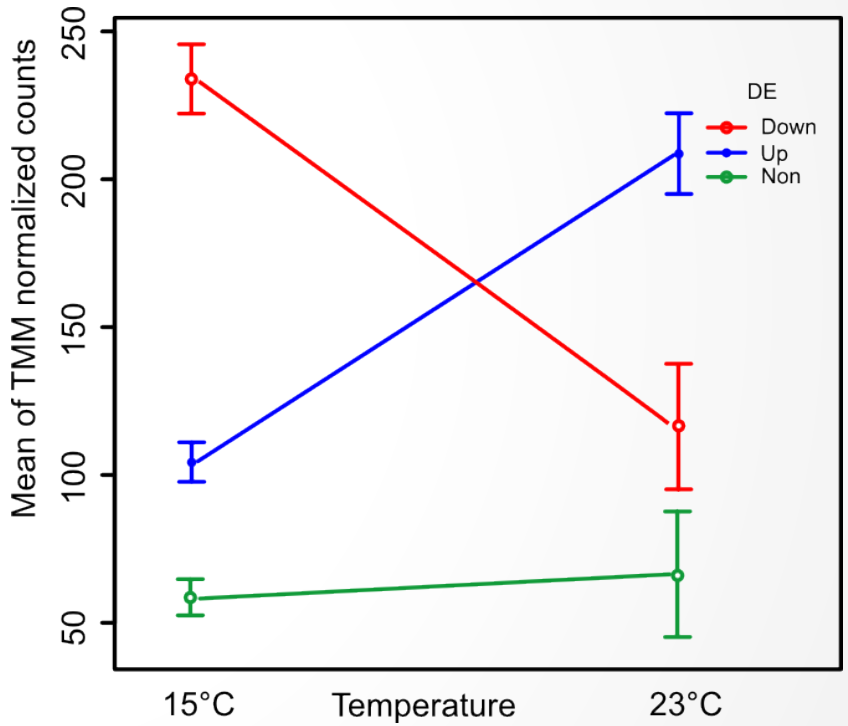
- http://www.rna-seqblog.com/
- http://rnaseq.uoregon.edu/
- Huber W et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Meth* **12**: 115
- De Wit P, et al. (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Mol Ecol Resour* **12**: 1058
- Malone JH & Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol* **9**: 34
- Oshlack A et al. (2010) From RNA-seq reads to differential expression results. *Genome Biol* **11**: 220
- Anders S et al. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**: 1765

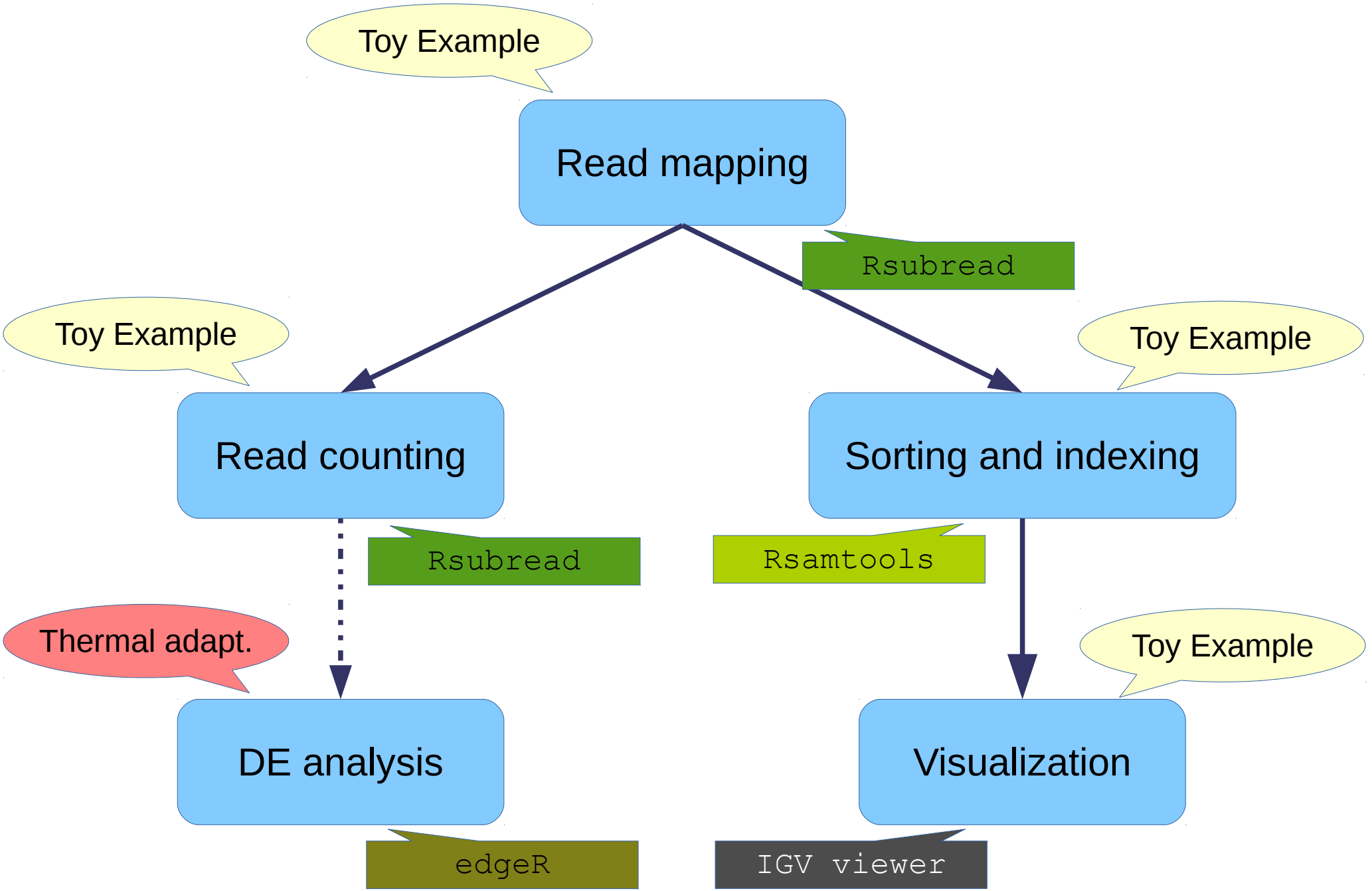
Az adatok

- Poikilotherm állatoknál erősen eltérhet a gének kifejeződése hideg és meleg környezetben
- Ezt génplaszticitásnak hívják.
- A jelenséget *Drosophila simulans* populációkon vizsgálták
- 3-3 biológiai replikát populáció (r1 r2 r3) génexpresszióját mérték le 15°C és 23°C-on
- A kérdés az volt, hogy mely gének expressziója változik a külső hőmérséklet hatására és hogyan?
  - szignifikáns különbségek
  - up és down regulált gének
- Szekvenálás: pooled, szál specifikus, paired-end, Illumina, a readek hossza: 100

Az adatok és a várható eredmények






Toy Example
Read mapping
Rsubread
Toy Example
Toy Example
Read counting
Sorting and indexing
Rsubread
Rsamtools
Thermal adapt.
Toy Example
DE analysis
Visualization
edgeR
IGV viewer