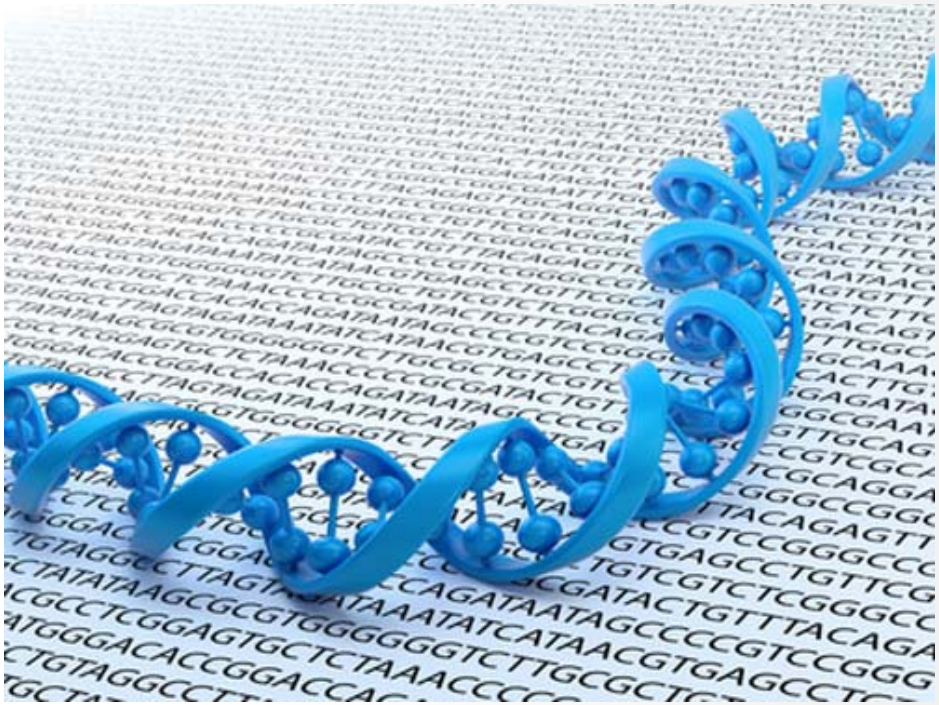
Genomika és transzkriptomika

Nagy áteresztőképességű (high-throughput) módszerek

Ari Eszter
ELTE, Genetikai Tsz.
arieszter@gmail.com
genetics.elte.hu
username: genetika2017
password: genetika2017

Miről lesz szó?

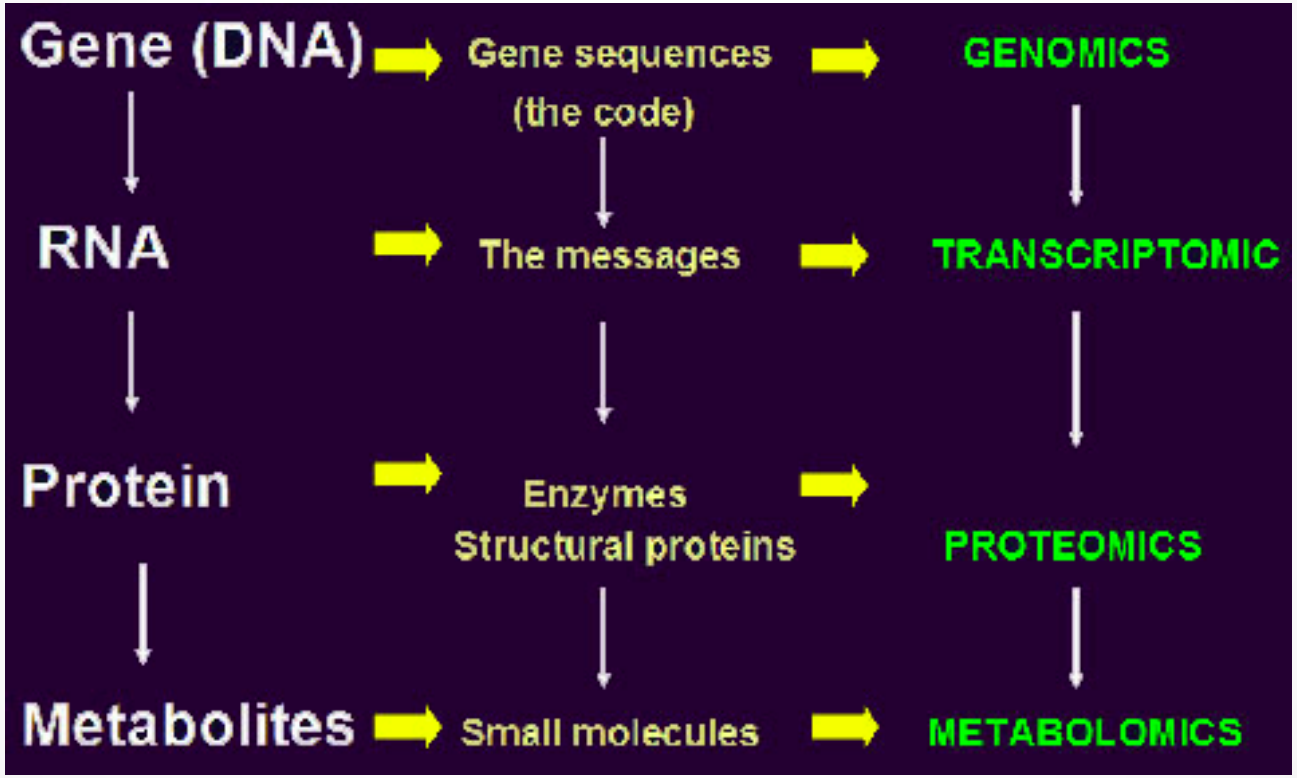
- Genomika
  - Genomok, projektek
  - Felhasználási területek
  - Genom szekvenálás
    - de novo szekvenálás
    - újraszekvenálás
- Transzkriptomika
  - Felhasználási területek
  - A microarray technológia
  - RNA-Seq adatok előállítása és elemzése
  - Expressziós különbségek vizsgálata



Genomika

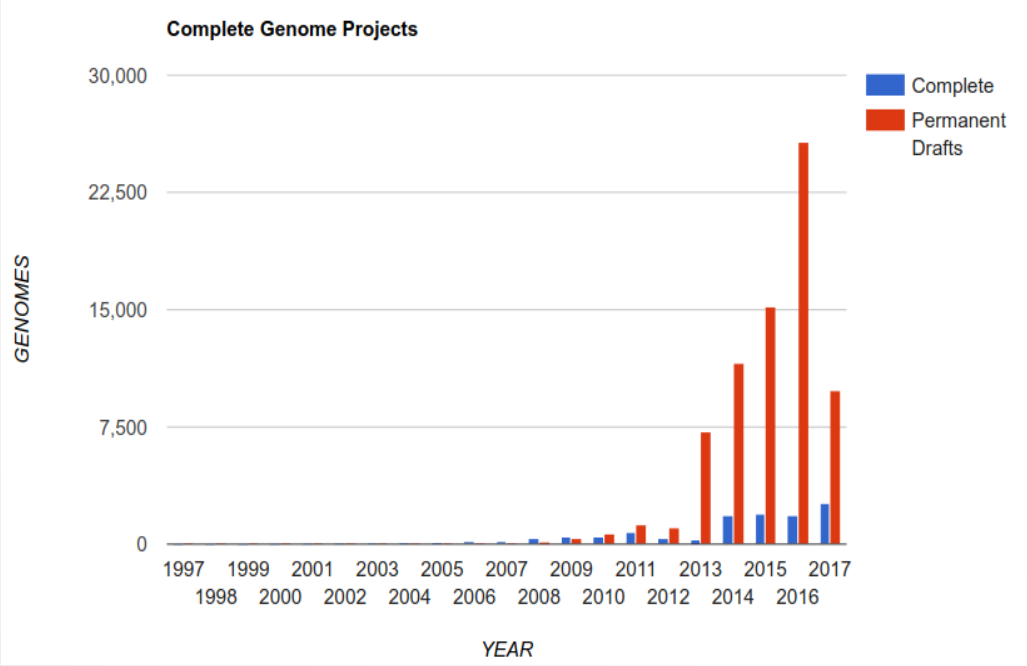
- Genom: Egy szervezet teljes örökítő információja
  - DNS kódolja (egyes vírusokban RNS)
  - gének és a nem kódoló szekvenciák
- A genommal foglalkozó tudományág a genomika,
  - eltér a genetikától, hiszen az utóbbi általában egy adott gén funkcióit vizsgálja.
- A genomika
  - a genomot,
  - a gének és a nem kódoló régiók kölcsönhatásait,
  - a genomok szerkezetét,
  - a gének elhelyezkedését és
  - az egyes élőlények genomja közötti különbségeket vizsgáló multidiszciplináris tudomány.
- Az élőlények genomjában rejlő információkat a bioinformatika segítségével dolgozza fel.

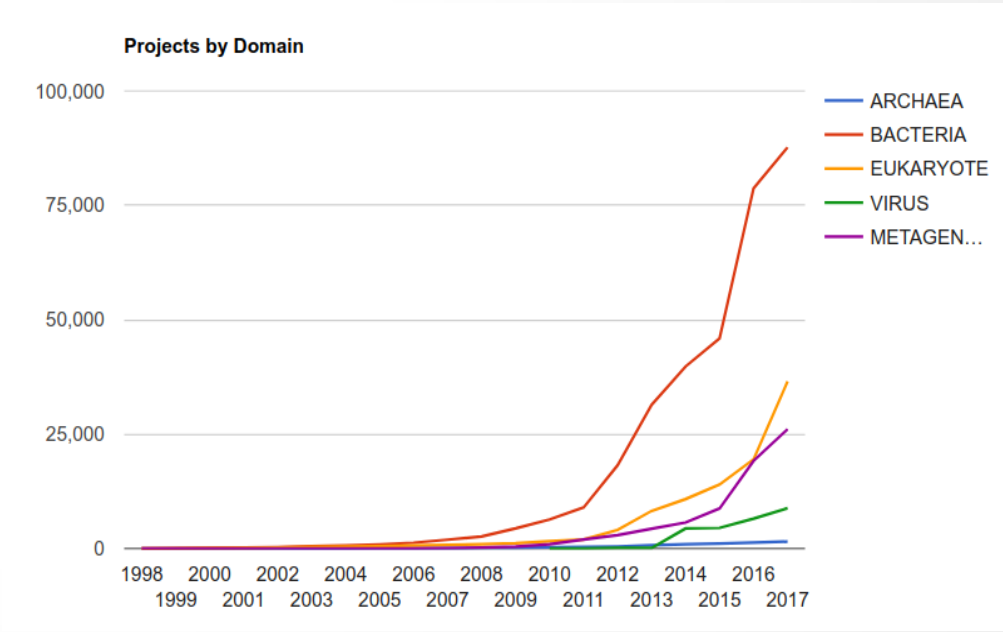
Omics


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

Genom programok

- GOLD, Genomes online database: https://gold.jgi.doe.gov/





A genomika eredményeinek felhasználási területei

- Genetika
  - pl: gének helye, környezete, szabályozása; rekombinációs hot-spotok feltárása
- Populációgenetika
  - pl: a populáció múltjának feltátsa SNP-k segítségével
- Evolúciógenetika
  - pl: szelekció alatt álló genomrészek feltérképezése; filogenetika
- Paleantológia
- Orvostudmány
  - diagnosztika
  - egyénre szabott terápia, pl: génterápia
- Gyógyászat:
  - pl. rákkutatás
- Gyógyszerfejlesztés
- Mezőgazdaság (GMO)
- Környezetvédelem
- Élelmiszeripar
- Kriminalisztika, igazságügy

Hogy kapjuk az adatokat?

Genom szekvenálás: múlt és jelen

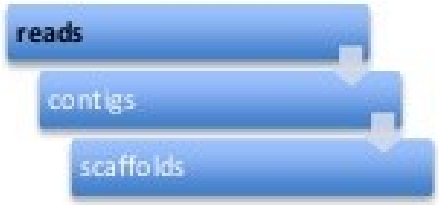
- A genomok szekvenálásához többféle stratégia használható:
  - a múlt:
    - Klón alapú, hierarchikus szekvenálás (BAC könyvtárakkal)
    - Whole genom shutgun (“genom-robbantás”) alapú szekvenálás
  - a jelen:
    - Massively paralell Next Generation Sequencing (NGS)

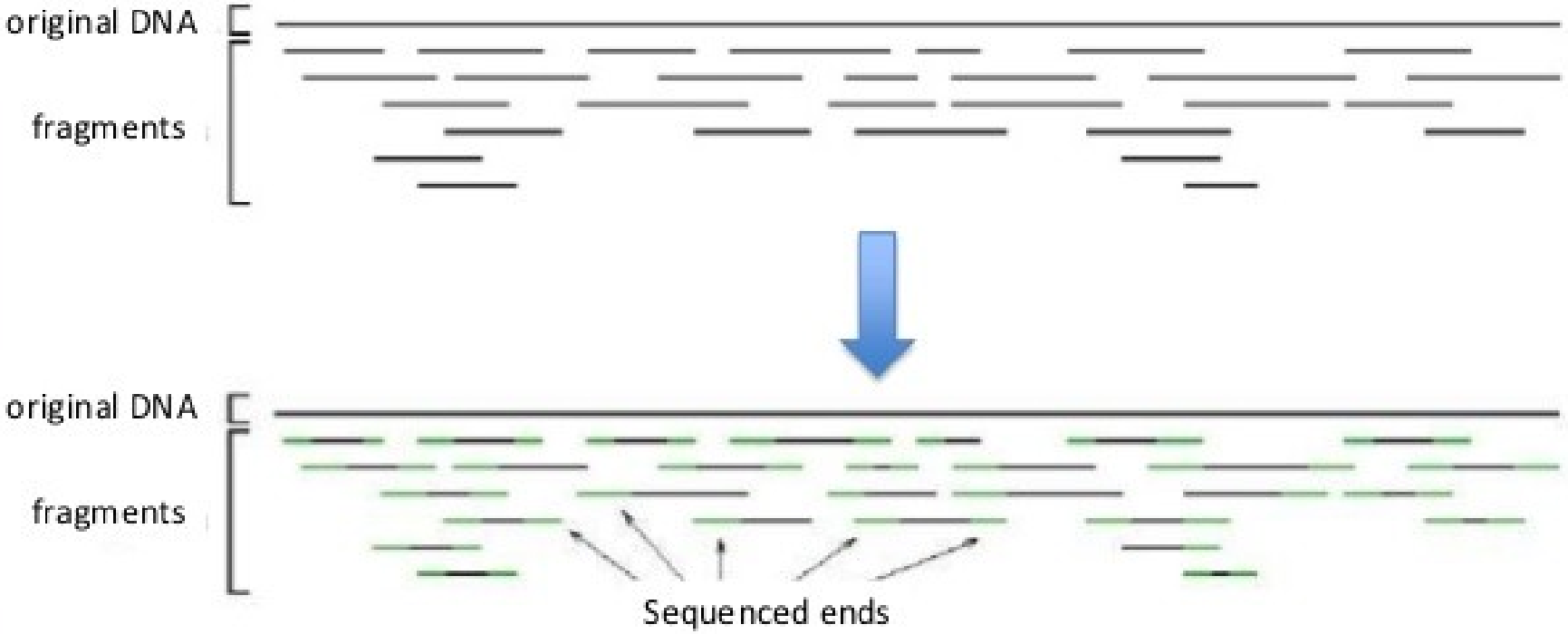
Az NGS lépései - általánosságban

1) DNS kivonás
2) Könyvtár készítés:
    1) Fragmentálás
    2) Amplifikálás
3) Fragmentek egyik vagy mindkét végének szekvenálása, párhuzamosan → readek
4) Bioinfo...

Sequence data

Reads

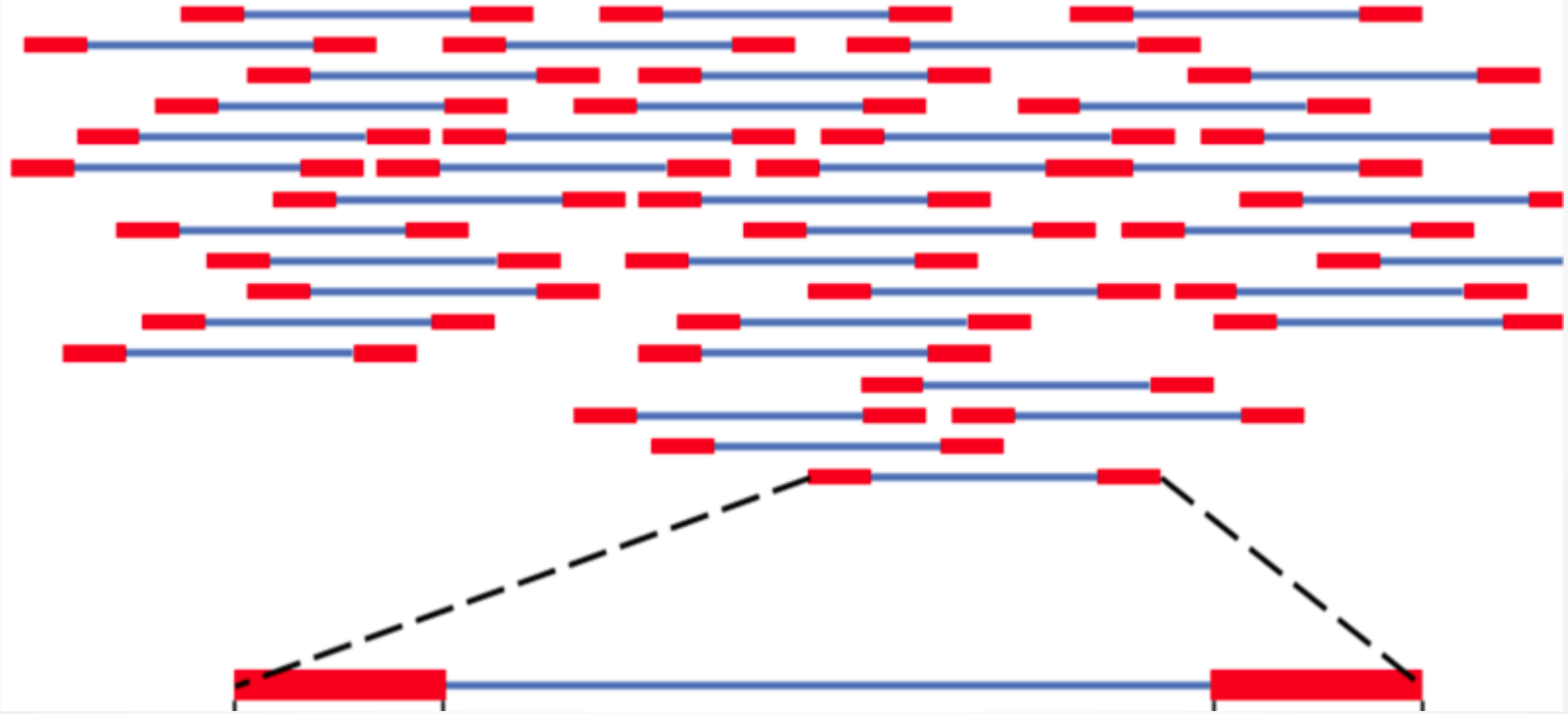




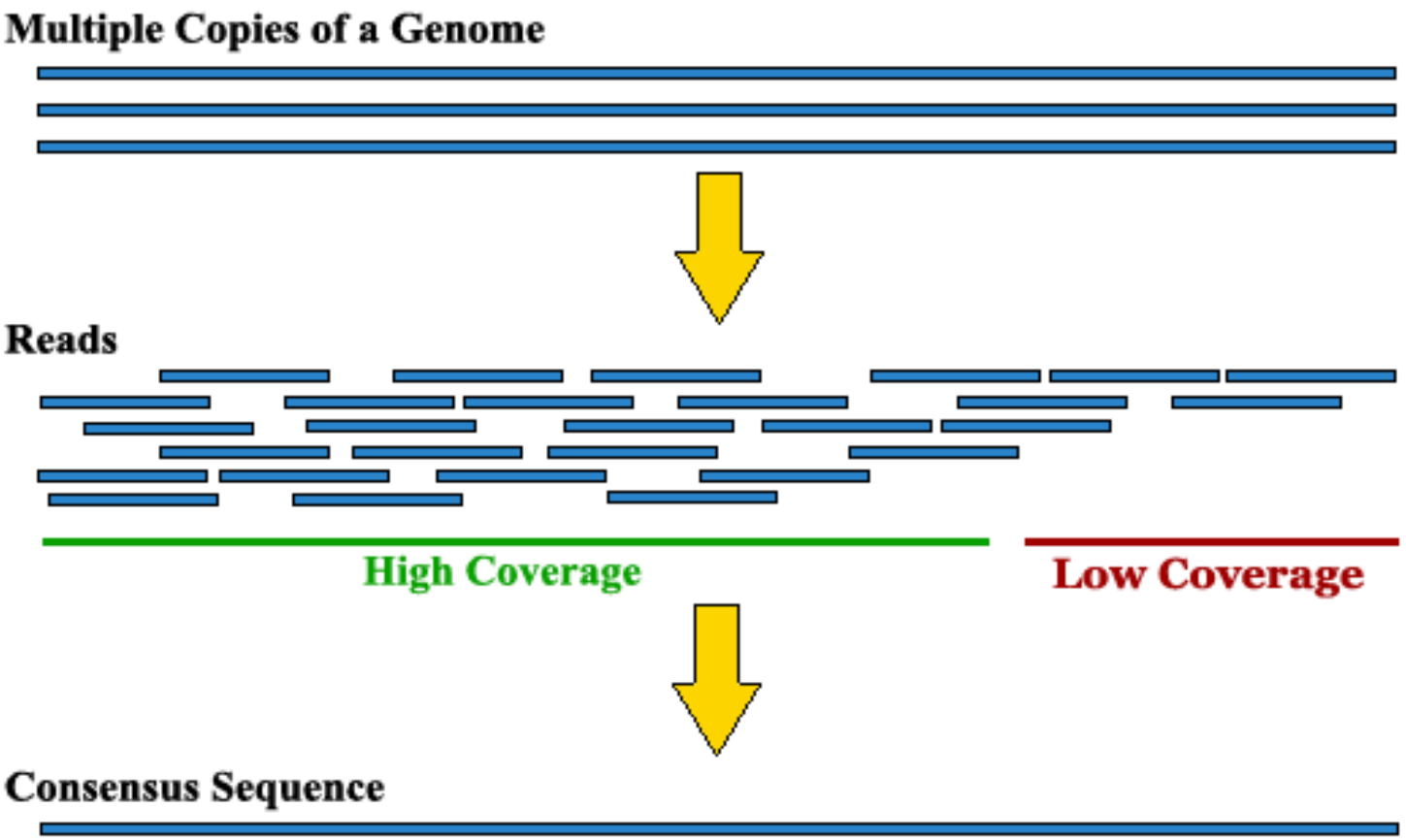
http://www.cbcb.umd.edu/research/assembly_primer.shtml

Next Generation Sequencing - NGS

- **Nagymértékben párhuzamosított, sok mintát szekvenálnak egyszerre, nagy áteresztőképességű módszerek, „high throughput”** → gyors, olcsó
- Whole genome shoutgun, a darabok elejének (single end sequencing), vagy az elejének és végének (paired end seq.) nt sorrendjét olvassák le.
- Párhuzamosan akár 1 milliót DNS darab szekvenálása!



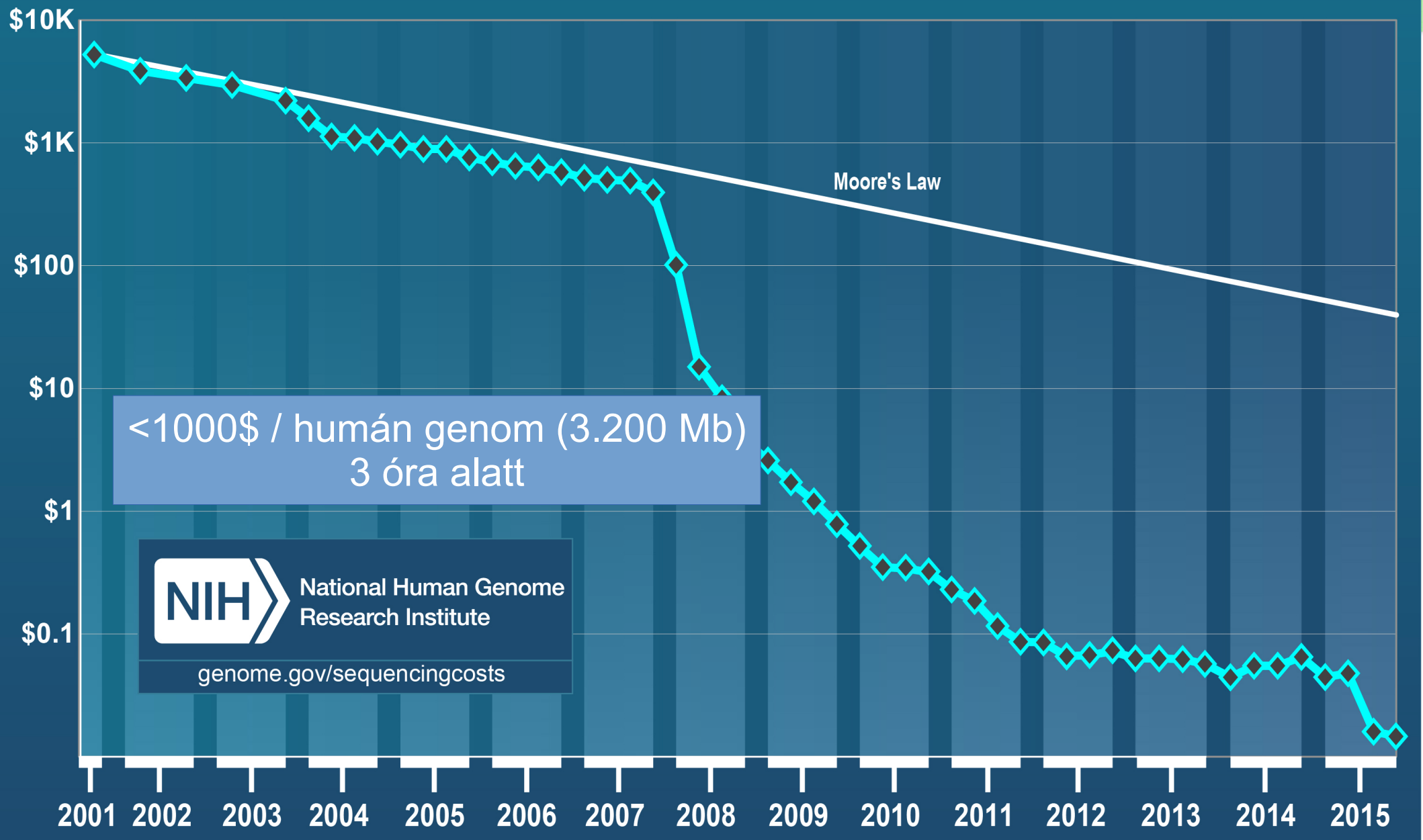
Lefedettség



Next Generation Sequencing - NGS

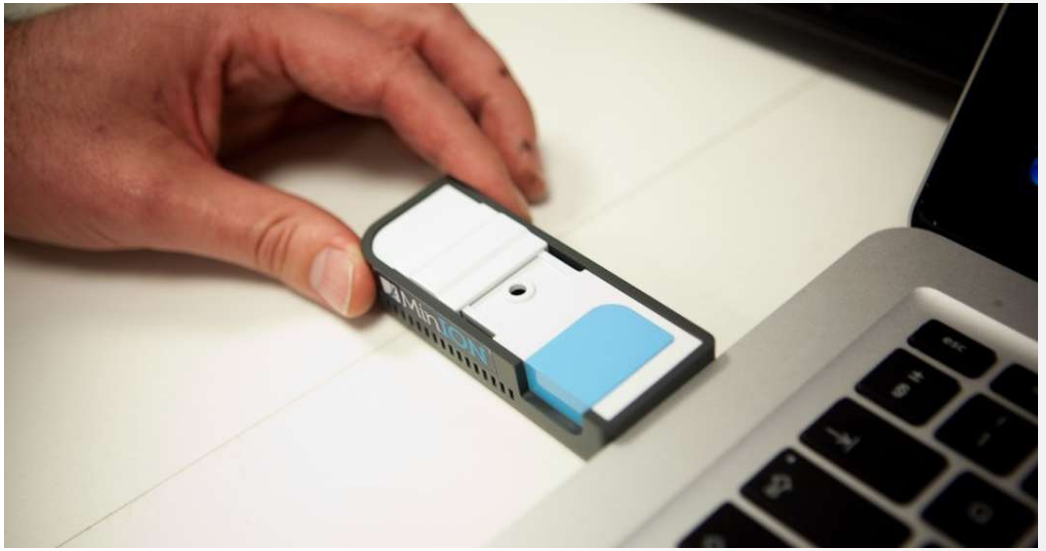
- Lehet szál-specifikus (forward, reverse)
- Eljárások (nem a Sanger-féle szekvenáláson alapulnak):
  - Illumina (Solexa) sequencing
  - SOLiD sequencing
  - Ion Torrent sequencing
  - Pyrosequencing (454)
  - PacBio
  - Oxford nanopore ...
- a readek hossza: 50-700 nt (egyféle eljárásban egyféle hossz) - rövidebb, mint a Sanger
- egy nap alatt milliós nagyságrendű read olvasható le
  - költség: ~ 1 $ - 5 cent / 1.000.000 nt
- A szekvenálás gyors (humán genom egy nap alatt), de az assembly bonyolult és nagyon számításigényes

Cost per Raw Megabase of DNA Sequence



NGS készülékek

Illumina HiSeq



MinION

Illumina sequencing

Video
https://www.youtube.com/watch?v=HMyCqWhwB8E

A genom analízis lépései

1. Minőség ellenőrzés
2. Trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3.a) Új genom: de novo assembly

3.b) Ismert genom: mapping

4. A genetikai változatosság feltárása: statisztikai elemzés

A genom analízis lépései

1. Minőség ellenőrzés
2. Trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3.a) Új genom: de novo assembly

3.b) Ismert genom: mapping

4. A genetikai változatosság feltárása: statisztikai elemzés

A readek

- a szekvenálás eredménye: pl. fastQ file
  - minőség ellenőrzés - quality checking (pl: FastQC)
  - trimming: a rossz minőségű readek, vagy read részletek kiszűrése

A genom analízis lépései

1. Minőség ellenőrzés
2. Trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3.a) Új genom: de novo assembly

3.b) Ismert genom: mapping

4. A genetikai változatosság feltárása: statisztikai elemzés

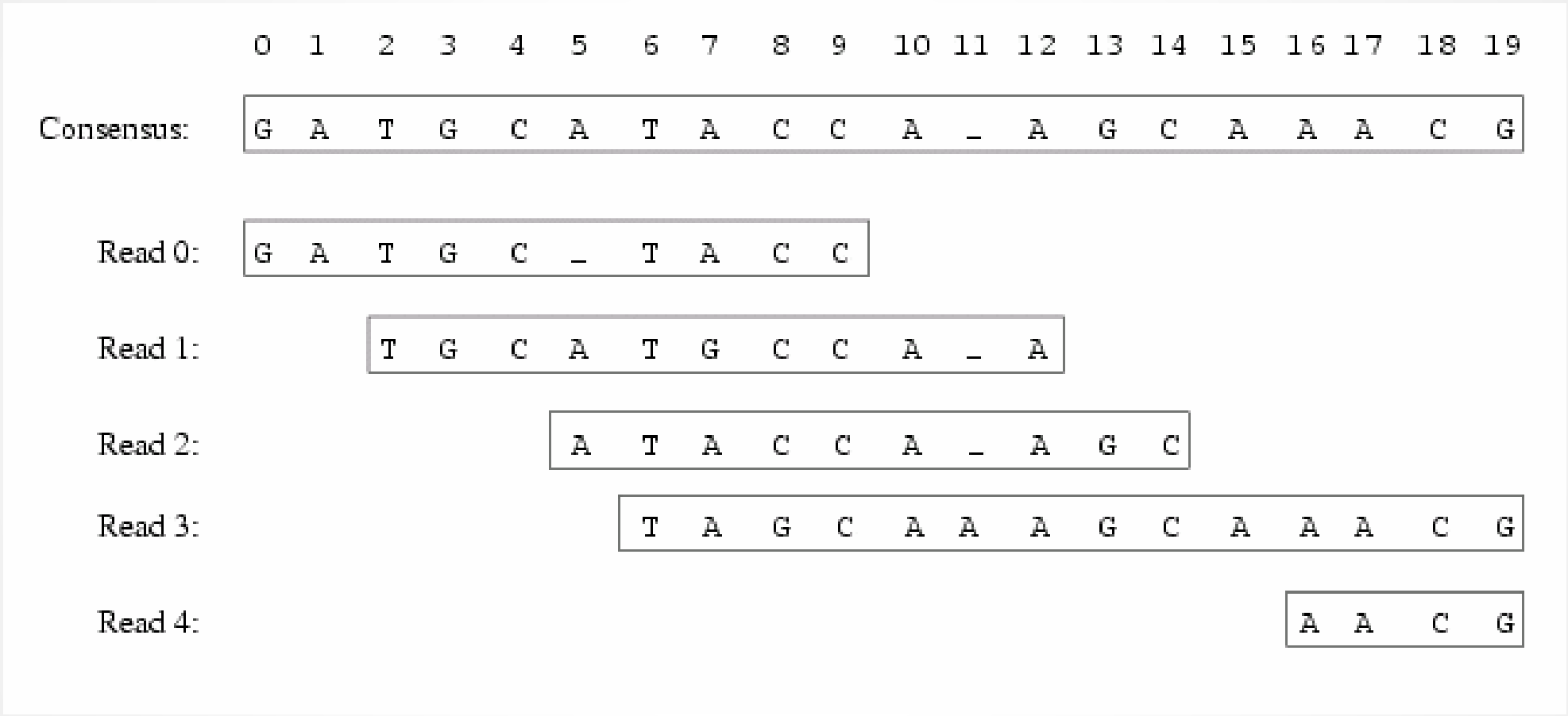
De-novo genomszekvenciák összeszerelése (assembly)

- A read-ek alapján az egész genom összeállítása
  - az eukarióták közül először a muslica genomját fejtették meg kizárólag ezzel a módszerrel
  - Humán genom: 2-3 milliárd(!) read (100X-os lefedettség)
- A mohó (gready) algoritmus:
  1. Minden lehetséges read páronkénti illesztése - szekvencia egyezés alapján
  2. A 2 legjobban átfedő readek kiválasztása és összevonása (merge)
  3. A 2. lépés ismétlése, amíg a readek el nem fogynak
- Assembler szoftverek: ABySS, Celera WGA, Edna, Euler, MIRA, Newbler, SOAPdenovo, …
- Probléma: egy újonnan megszekvenált genomnál nem tudjuk ellenőrizni, hogy helyes-e az assemly
  - okok lehetnek:
    - ismétlődő szakaszokat tartalmazó régiók - az innen származó readek kizárása
    - rossz helyre és/vagy orinentációban beillesztett readek

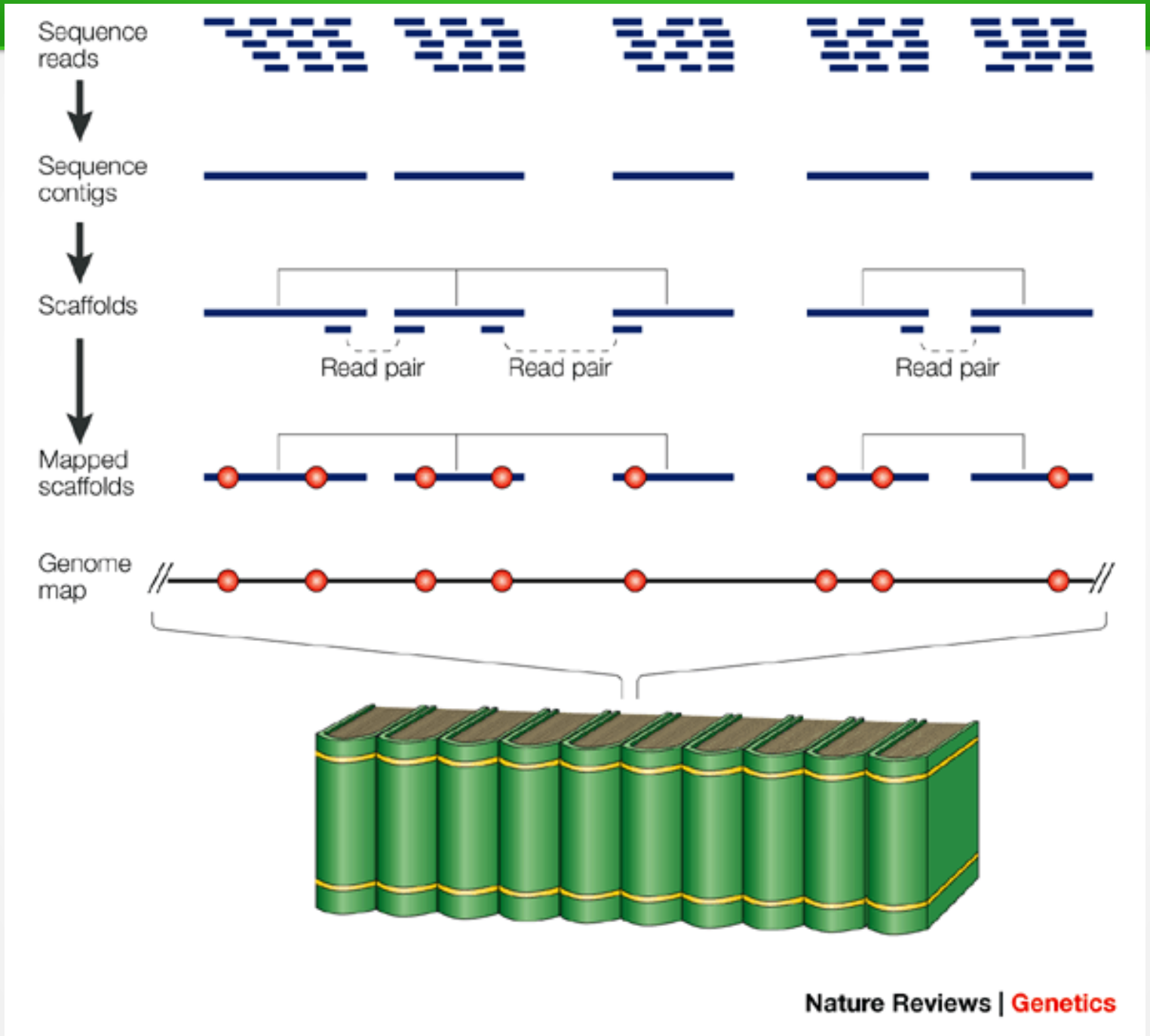
Illeszkedő és nem illeszkedő readek

```
ATTGTGCTAGTCGTAGCTAGCT
          |||||||||||||||||||
          CTAGTCGTAGCTAGCTGTCAA

TGATGATGCTCTAAGATCTCAT
```


0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Consensus: G A T G C A T A C C A _ A G C A A A C G
Read 0: G A T G C _ T A C C
Read 1: T G C A T G C C A _ A
Read 2: A T A C C A _ A G C
Read 3: T A G C A A A G C A A A C G
Read 4: A A C G

Genom összeszerelés



Genom annotáció

- Biológiai információ rendelése a genom szekvenciához
  - Gének, egyéb funkcióval bíró helyek megtalálása a genomban
- Struktúrális annotáció:
  - ORF-ek beazonosítása
  - gén struktúra (UTF, exon, intron...)
  - promóter szakaszok megtalálása: közös motívumok alapján
- Funkcionális annotáció:
  - biológiai funkció, pl: BLAST keresés
  - expressziós adatok
  - regulációs hálózatok …
- Annotációs projektek:
  - ENCyclopedia Of DNA Elements (ENCODE), Entrez Gene, Ensembl, GENCODE, Gene Ontology Consortium, GeneRIF, Uniprot, Vertebrate and Genome Annotation Project (Vega)

A genom analízis lépései

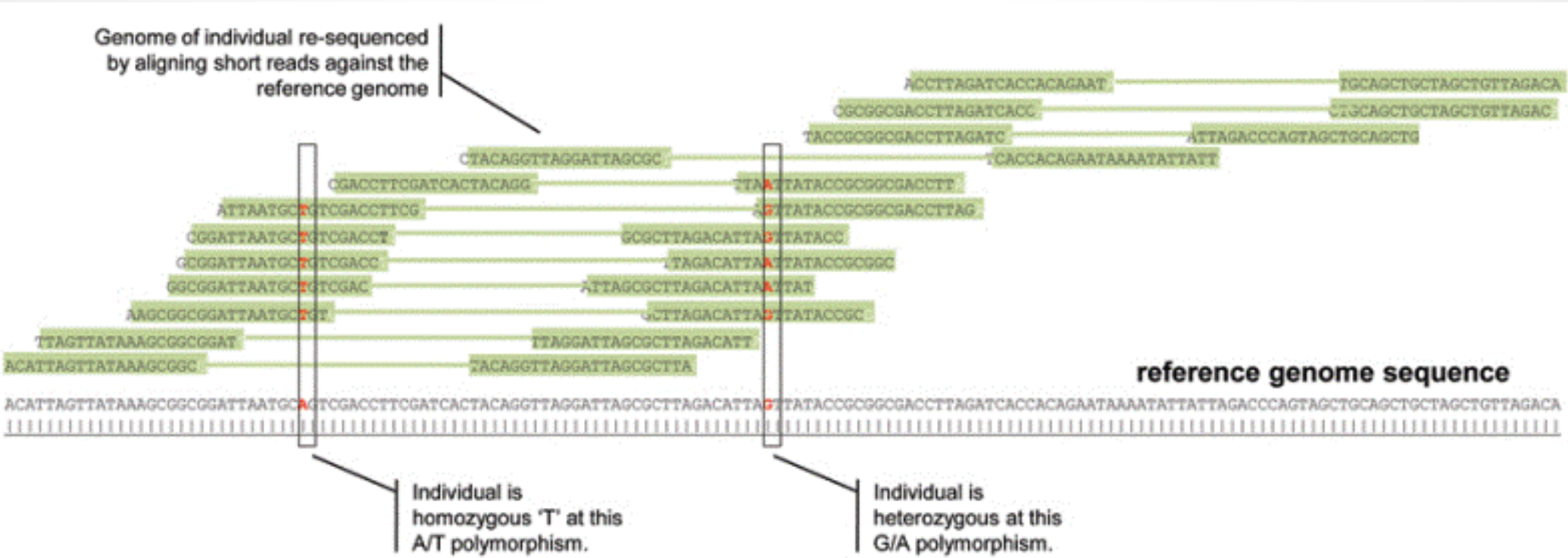
1. Minőség ellenőrzés
2. Trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3.a) Új genom: de novo assembly

3.b) Ismert genom: mapping

4. A genetikai változatosság feltárása: statisztikai elemzés

Re-szekvenálás



Re-szekvenálás

- Létező referencia-genomhoz illesztjük a readeket
  - kisebb - de még mindig jelentős - számításigény
    - genom variabilitás
  - viszont néhány variáció rejtve maradhat (pl. kromoszóma átrendeződések)
  - az illesztéshez használt genom szekvenciában is lehetnek hibák, hiányosságok
- Cél lehet pl. a genetikai változatosság feltárása
- Illesztő (mapping) szoftverek: BWA (Burrow's Wheeler Transform Algorithm), Bowtie, GSNAP, SOAP2, …

A genom analízis lépései

1. Minőség ellenőrzés
2. Trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3.a) Új genom: de novo assembly

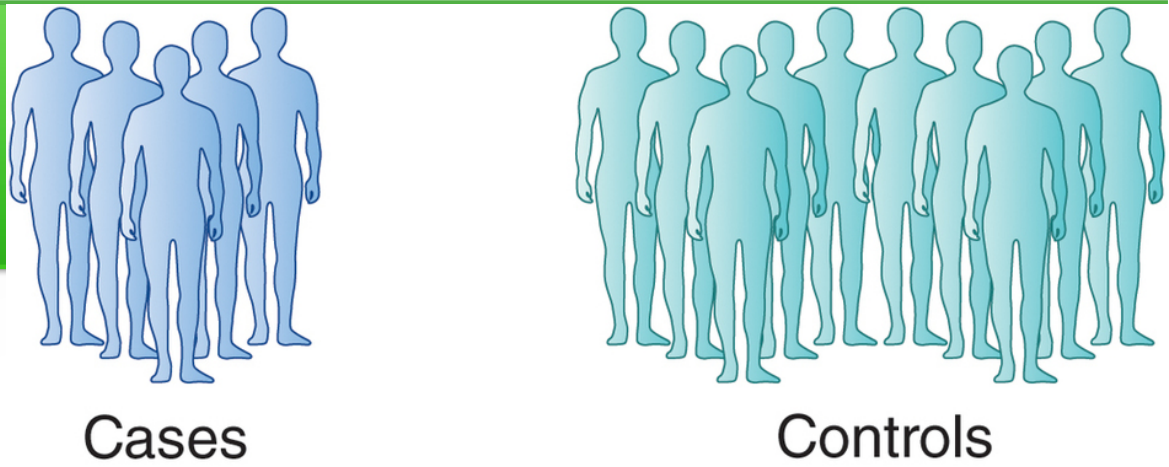
3.b) Ismert genom: mapping

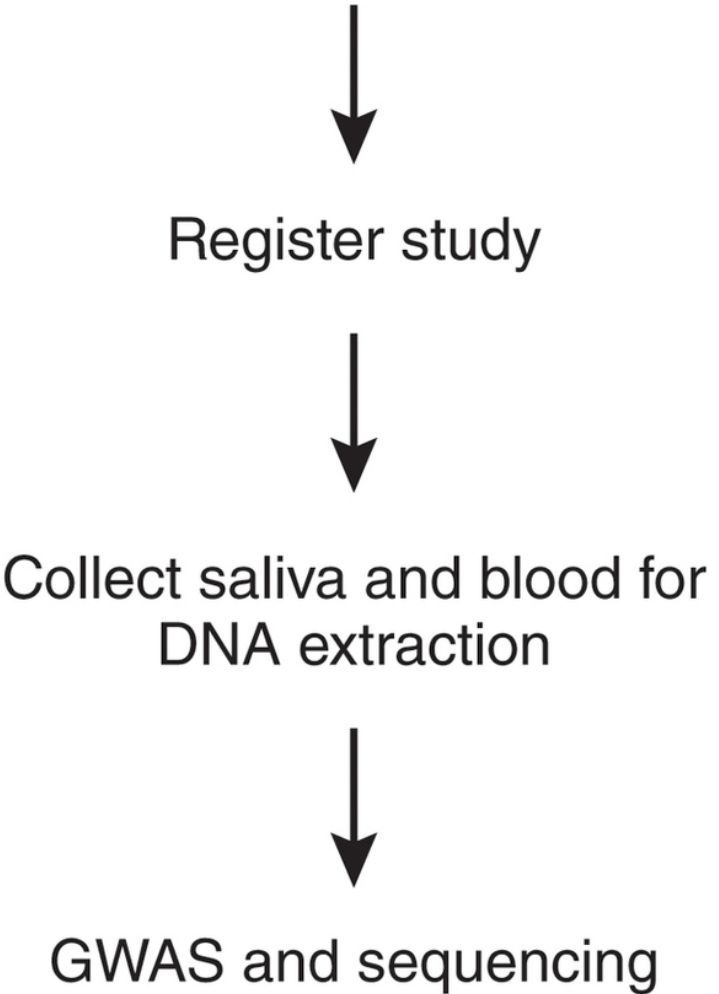
4. A genetikai változatosság feltárása: statisztikai elemzés

A genetikai változatosság feltárása

- 2 egyed genomja sokféleképpen különbözhet: SNPk, in/del-ek, kópia szám különbségek, kromoszóma átrendeződések
  - ezek közül bármelyek felelősek lehetnek egy adott fenotípus vagy betegség kialakulásáért
- SNP analízis / GWAS: genome-wide association study / Genetikus asszociáció kutatás
  - kapcsolat keresése egy tulajdonság (trait) és az egyedek genom-variabilitása között
  - többnyire SNP különbségeken alapszik → allélfrekvenciák feltárása
  - a tulajdonság lehet egy betegség vagy pl. az egyedek mérete, hőtűrése, szemszíne, etc.

- A genetikai változatosság feltárása
- SNP analízis
- GWAS: genome-wide association study, Genetikus asszociáció kutatás





Replikátok

- Ahhoz, hogy statisztikai tesztelést végezzünk megfelelő mennyiségű mintára van szükségünk
  - replikátok: azonos “kezelést” kapott minták
  - vizsgálattól függően 2-3-100 replikát / kezelés szükséges

Transzkriptomika

- Transzkriptom:
  - a genomról átíródó RNS-ek összessége
  - (Egyik) kulcs a DNS - fenotípus kapcsolatának megfejtésében
- RNS típusok:
  - mRNS, rRNS, tRNS
  - Poszt traszkripciós módosítás: small nuclear snRNS, small nucleolar snoRNS, ...
  - RNS regulálás: micro miRNS, piwi-interacting piRNS, small interfering siRNS
- Transzkriptomika
  - A génexpresszió vizsgálata
  - A transzkripciós különbségek feltárása

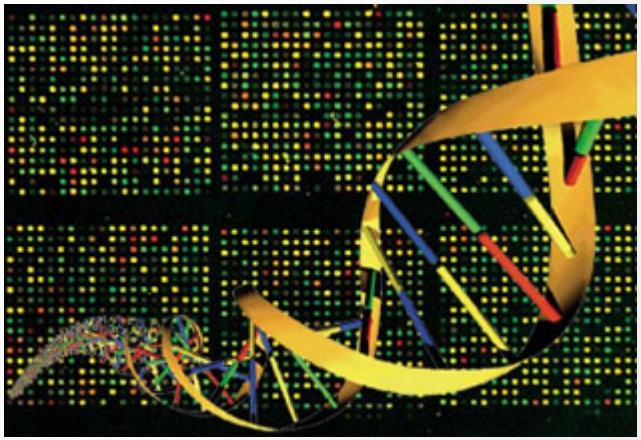
Transzkriptomika

- Alapfeltevés: az mRNS-szint jellemzi az adott gén expressziós szintjét és az adott fehérje mennyiségét is
- A sejt, szövet, egyed vagy populáció különböző állapotaiban (pl. sejtciklus különböző fázisai, ill. más-más kezelés) vehetünk mRNS-mintát és megmérhetjük az egyes gének expressziós szintjét.
- Összehasonlíthatjuk a különböző sejtek, szövetek, egyedek, populációk génexpresszióját.
- Ezekből 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

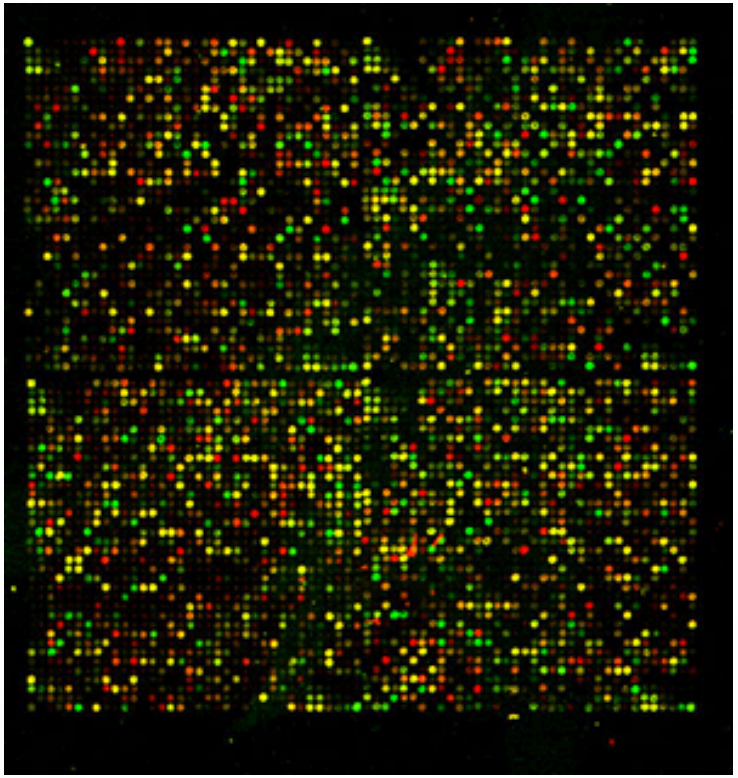
...



Módszerek

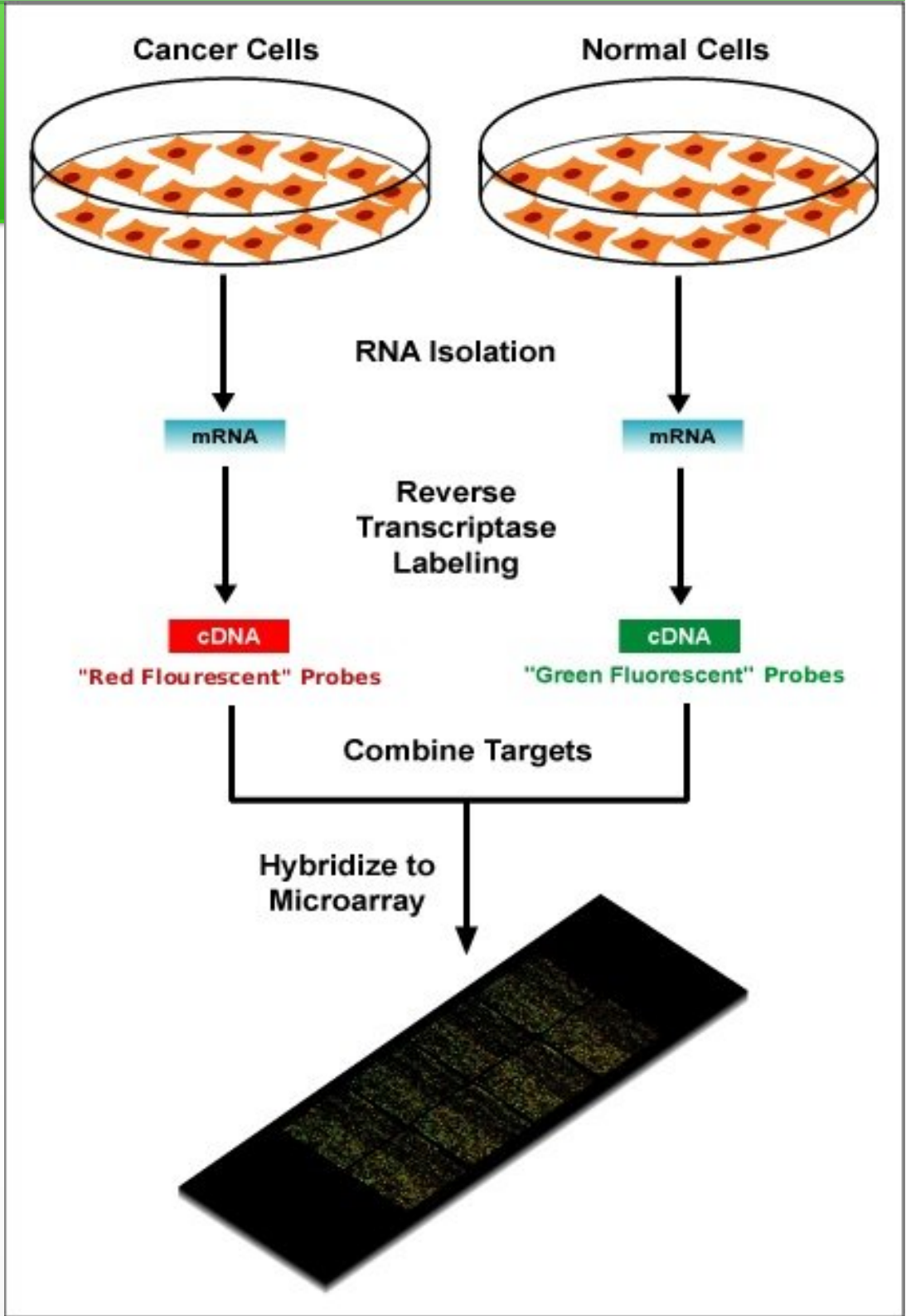
- Mit tudunk mérni?
  - RNS-ek szintje
  - Fehérjék szintje
- Hogyan?
  - Northen blot (1977)
  - reverse-transcription RT-PCR (1992)
  - Real-Time quantitative qRT-PCR
  - high-throughput módszerek
    - RNS Microarray vagy CHIP (1999)
    - Szekvenálás - RNA-Seq (2008)
    - Protein-array (Fehérje-csippek)

Microarray



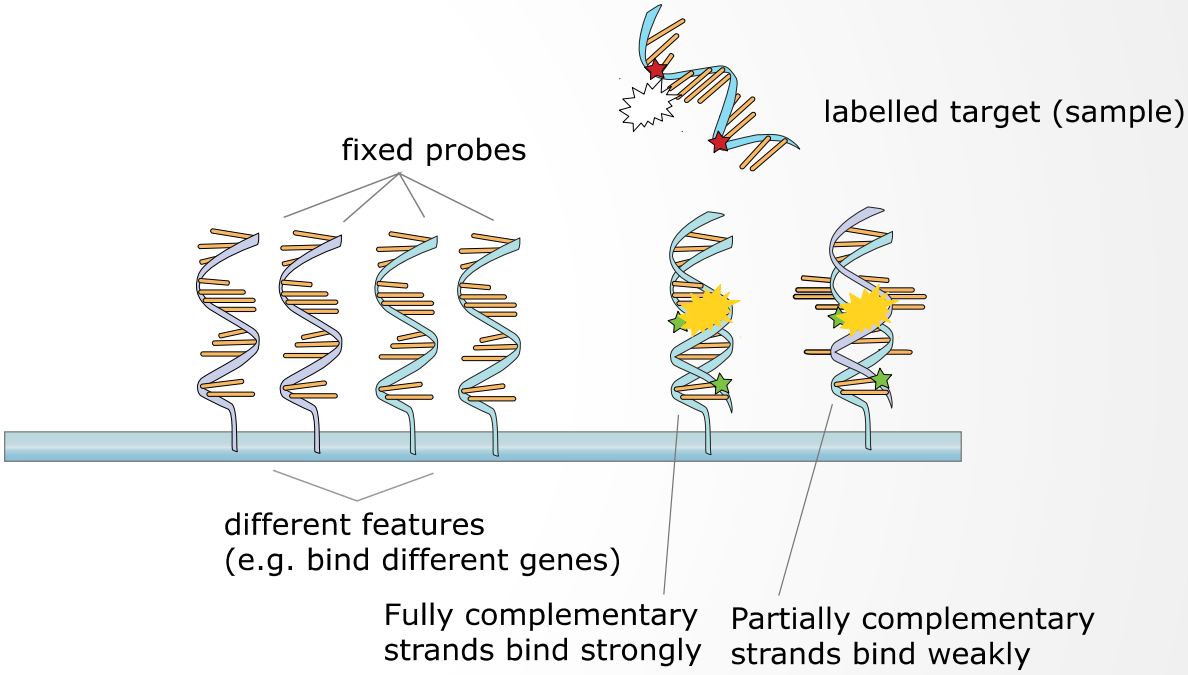
Mire jó?

- Mennyiségi információt ad a „teljes” transzkriptomról egy 1×1 cm-s lemezen
  - Egészséges vs. beteg
  - Kezelt vs. kezeletlen
- Mely gének expressziós szintje különbözik szignifikánsan?

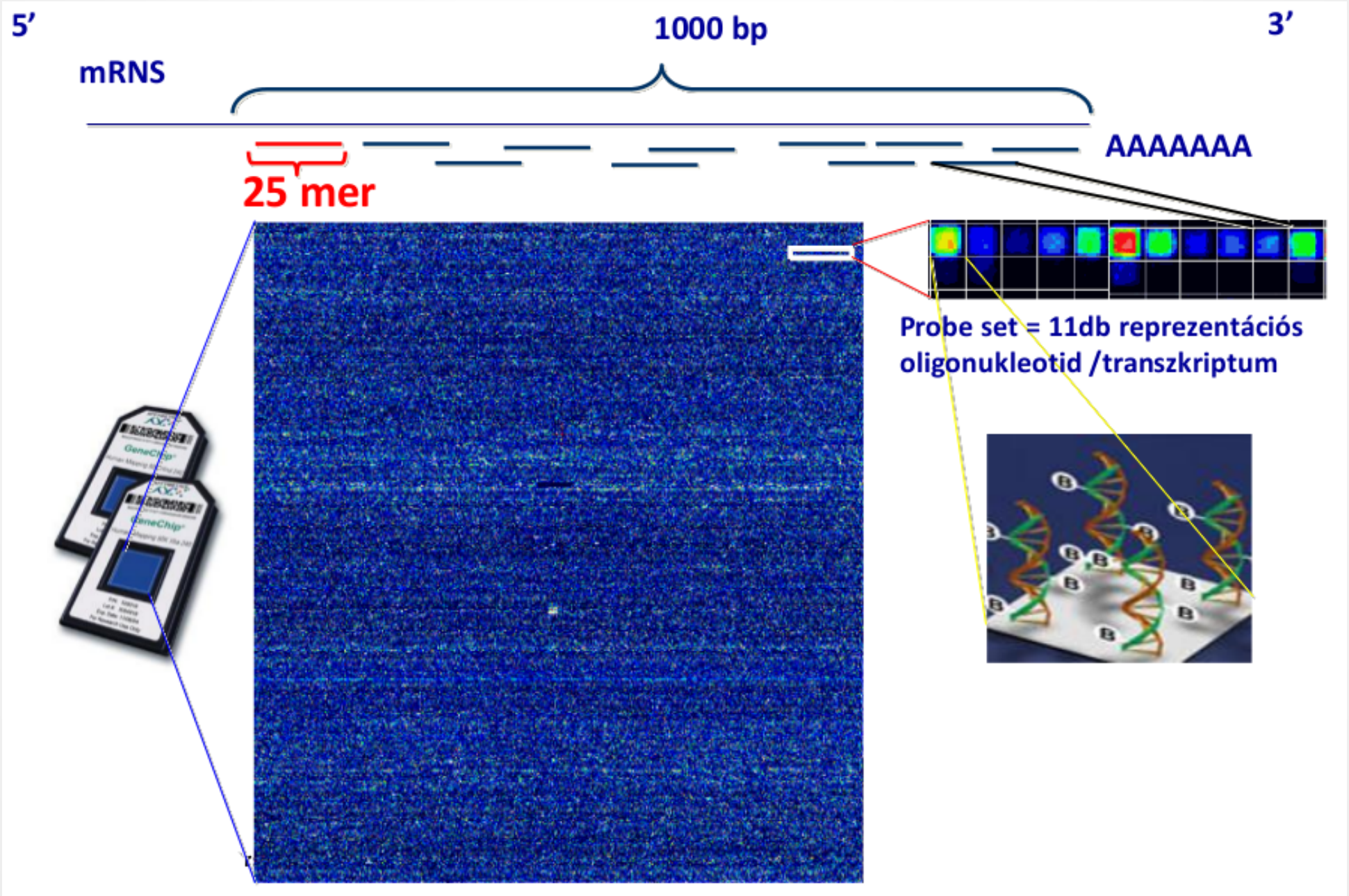


Tulajdonságok

- Oligo típusú:
  - 20-60 bp hosszú DNS darabka a chipen
  - *Affimetrix* - 1 minta, 1 szín
  - *Agilent* - 2 minta, 2 szín
- több probe / DNS szakasz → probe set
- több probeset / mRNS
- mismatch control



RNS Microarray



Microarray adatfeldolgozás

- Bioinformatika
  - Háttérkorrekció
    - Kivonni a zajt az adatokból
    - Segítenek a negatív kontrol probe-ok (nincsenek a genomban)
  - Normálás
    - Probe-okat egymáshoz viszonyítani
  - Aggregálás
    - A probe-okon mért szinteket összegezni probe setekre, majd mRNS szintekre

Differenciál expresszió analízis

- Beteg vs. egészséges állapot összehaonlítása
- Hipotézismentes kutatás
- *Fold change*
- *t*-teszt: *P*-érték

Fold change

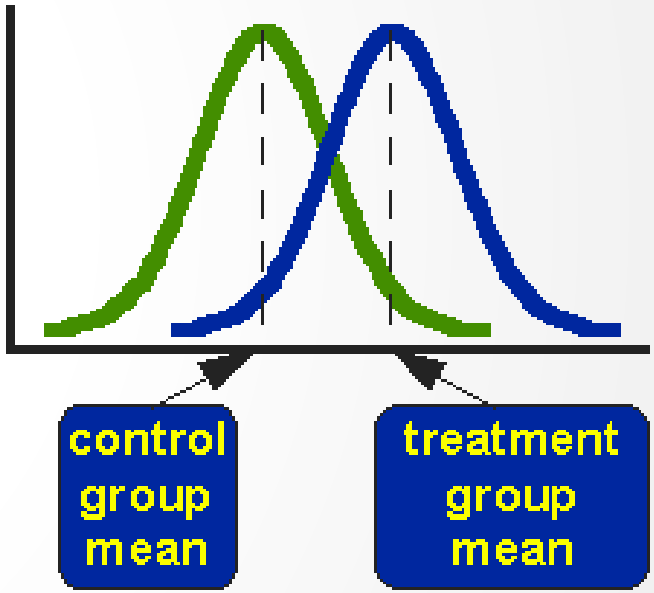
- Nagyságrendi expressziós szint változást lehet mérni vele

$$\log_2 FC = \log_2\left(\frac{egyik\ állapot\ expr.\ szint\ átlaga}{másik\ állapot\ expr.\ szint\ átlaga}\right)$$

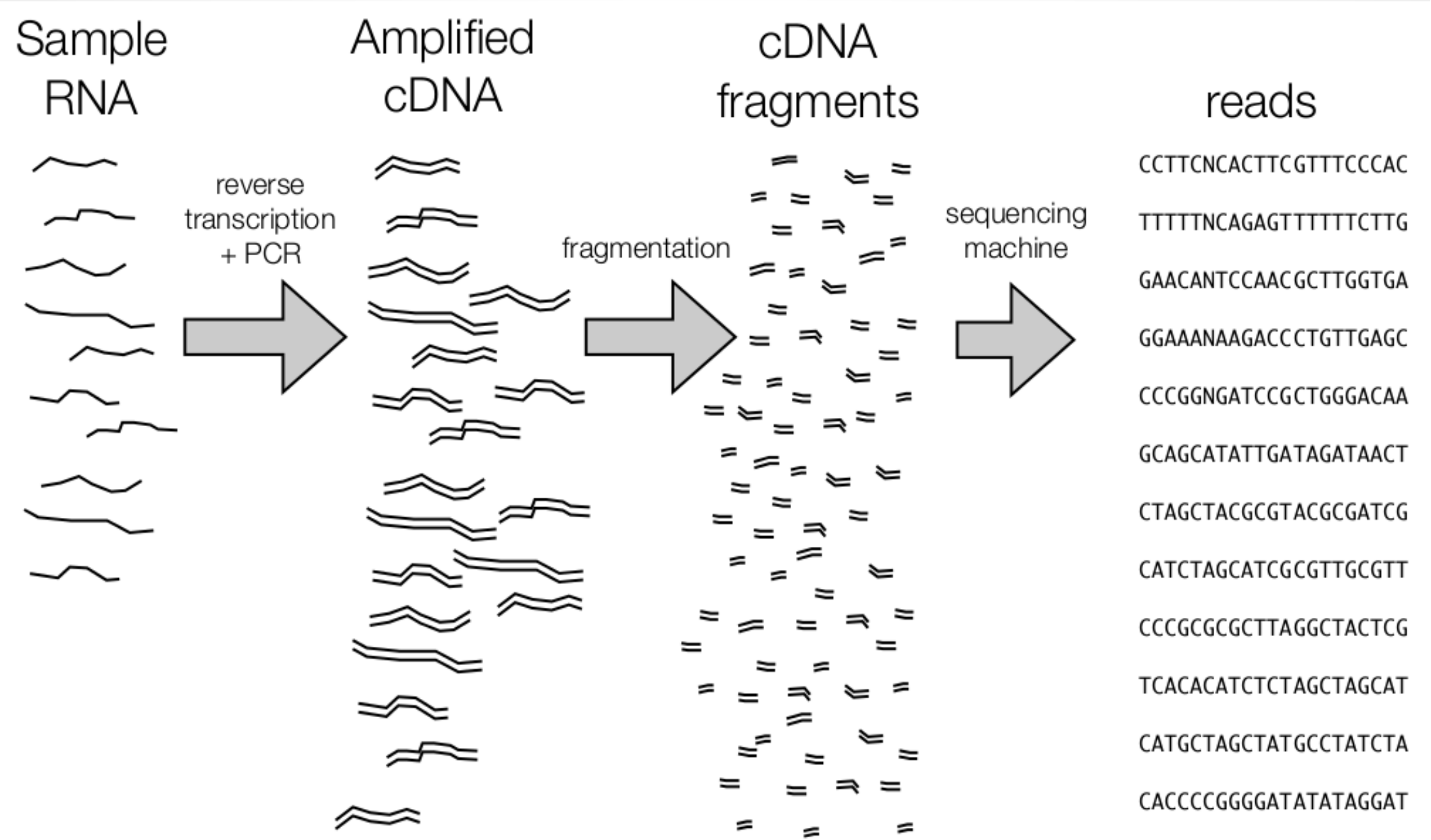
- a + - 2 (4×-es változás) már nagynak számít
- De itt csak az átlagokat hasonlítjuk össze! Nem statisztikai teszt.

Hipotézis vizsgálat

- 2 mintás *t*-próba
  - H0: a két eloszlás várható értéke azonos
    - → honnan vannak eloszlásaink?
    - → replikátok (több mintánk van mindkét állapotból)
  - *P*-érték: annak a valószínűsége, hogy a H0 fennállása esetén a véletlen játéka a H0-nak legalább annyira ellentmondó mintát produkál, mint a ténylegesen megfigyelt minta.



RNA-Seq


Sample RNA
reverse transcription + PCR
Amplified cDNA
fragmentation
cDNA fragments
sequencing machine
reads
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
CCCGCGCGCTTAGGCTACTCG
TCACACATCTCTAGCTAGCAT
CATGCTAGCTATGCCTATCTA
CACCCCGGGGATATATAGGAT

Az RNA-seq előnyei

- robusztusság, jól reprodukálható
- érzékenység
- az mRNS-ek szintjének „direkt” mérése (?)
- az átíródó RNS szekvenciákat is rekonstruálhatjuk
- minden transzkript mérése - még az addig ismeretleneké is
- transzkripciós izoformák, transzkript variánsok, szplicing helyek megismerése
- SNP-k feltárása
- olyan fajoknál is alkalmazható, ahol még nem áll rendelkezésre genom szekvencia

Problémák az mRNS expresszió vizsgálatokkal

- Az RNA-seq drágább, mint a microarray
  - Több bioinformatika és erősebb számítógépek kellenek az adatfeldolgozáshoz
- Ahhoz, hogy az „összes” gén működését lemérjük nagy lefedettség kell → drágább vizsgálat
- Nem mutathatók ki a poszt-transzkripciós módosulások
- Sem a poszt-transzkripciós reguláció hatásai:
  - Az mRNS expresszió még nem jelenti azt, hogy lesz belőle fehérje
  - A transzláció szabályozása az mRNS átíródása után (pl: miRNS gátlás)
    - bizonyos RNA-Seq technikákkal a miRNS-ek is mérhetők

Az RNA-Seq adatok feldolgozása

1. Szekvenálás → fastq file
2. Minőség ellenőrzés, trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3a. De-novo transzkriptom assembly

3b. Illesztés a genom szekvenciához

4. Expressziós szint számolás (count): egy génhez hány read(-pár) illeszkedett?

5. DE analízis: normalizálás, statisztikai összehasonlítás, az eredmények feldolgozása

Az RNA-Seq adatok feldolgozása

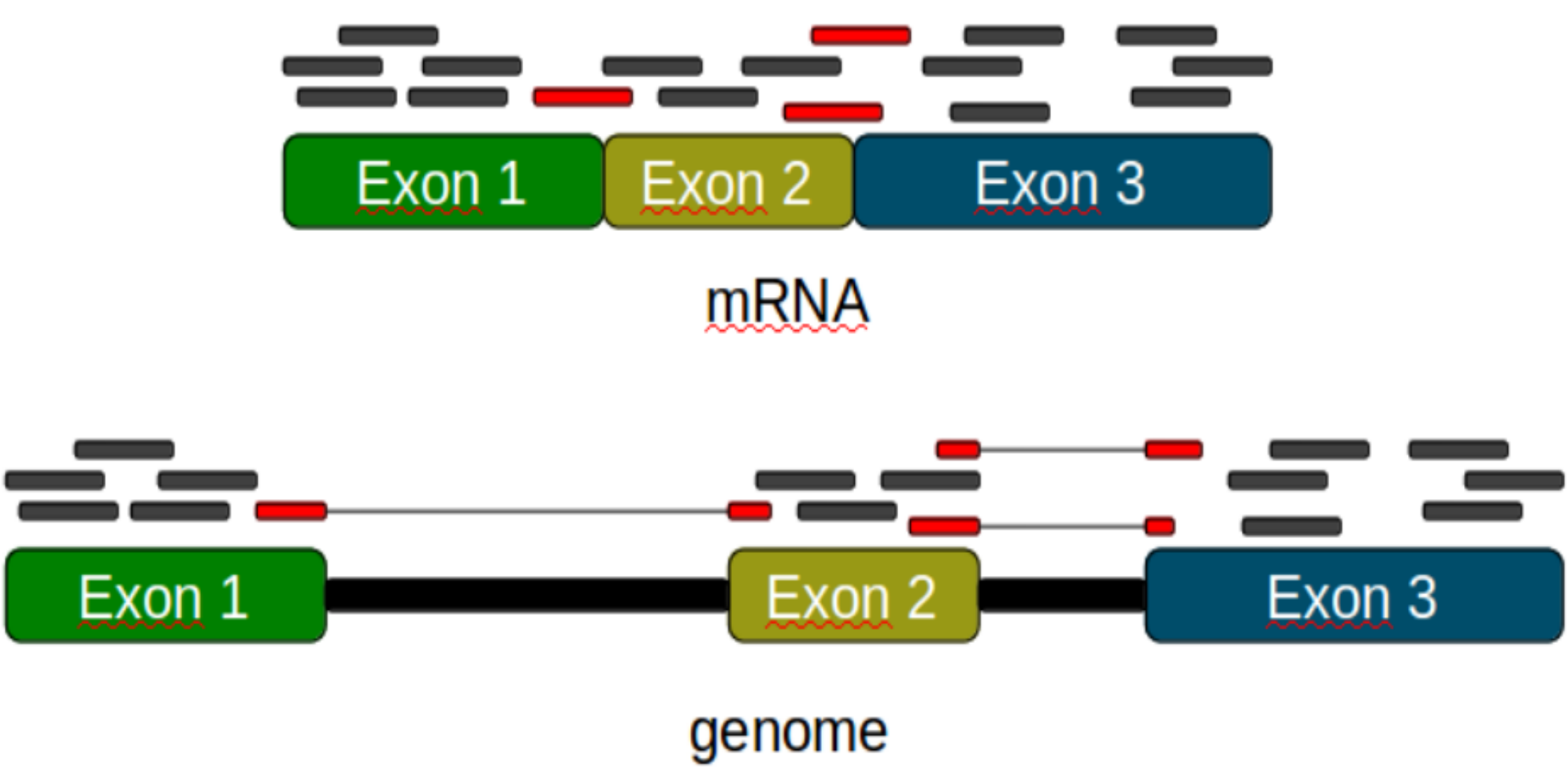
1. Szekvenálás → fastq file
2. Minőség ellenőrzés, trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3a. De-novo transzkriptom assembly

3b. Illesztés a genom szekvenciához

4. Expressziós szint számolás (count): egy génhez hány read(-pár) illeszkedett?
5. DE analízis: normalizálás, statisztikai összehasonlítás, az eredmények feldolgozása

Az RNA-Seq readek illesztése a genomhoz: mappelés



TopHat, GSNAP

Az RNA-Seq adatok feldolgozása

1. Szekvenálás → fastq file
2. Minőség ellenőrzés, trimming: a rossz minőségű readek, vagy read részletek kiszűrése

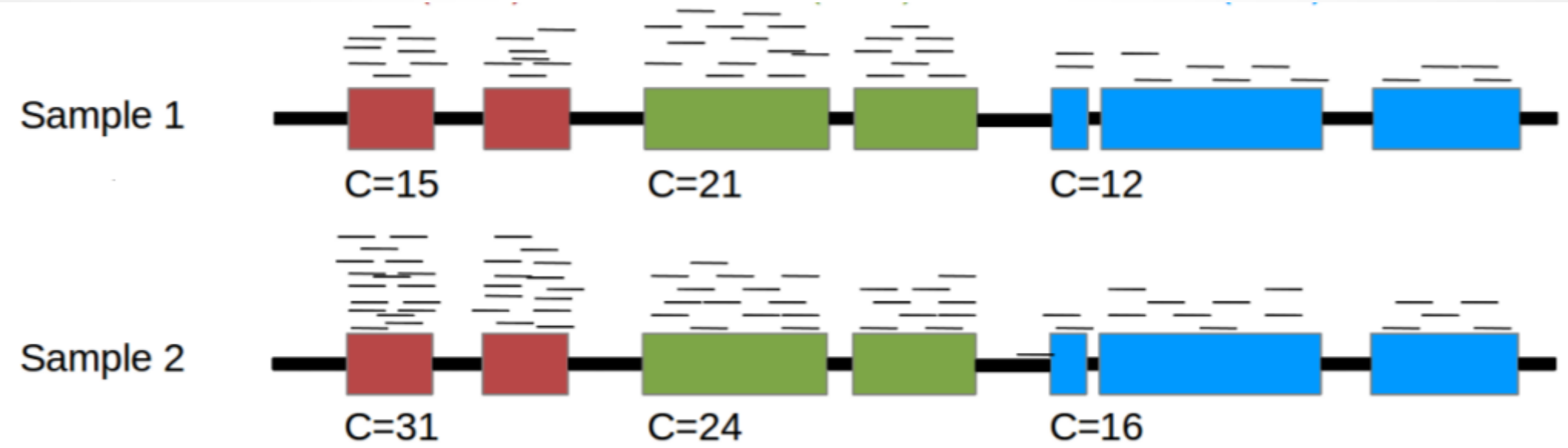
3a. De-novo transzkriptom assembly

3b. Illesztés a genom szekvenciához

4. Expressziós szint számolás (count): egy génhez hány read(-pár) illeszkedett?

5. DE analízis: normalizálás, statisztikai összehasonlítás, az eredmények feldolgozása

Expressziós szint mérés → countok



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

...

Az RNA-Seq adatok feldolgozása

1. Szekvenálás → fastq file
2. Minőség ellenőrzés, trimming: a rossz minőségű readek, vagy read részletek kiszűrése

3a. De-novo transzkriptom assembly

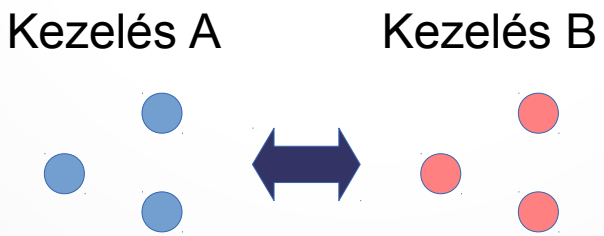
3b. Illesztés a genom szekvenciához

4. Expressziós szint számolás (count): egy génhez hány read(-pár) illeszkedett?

5. DE analízis: normalizálás, statisztikai összehasonlítás, az eredmények feldolgozása

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*
  - *pairwise DE analysis*

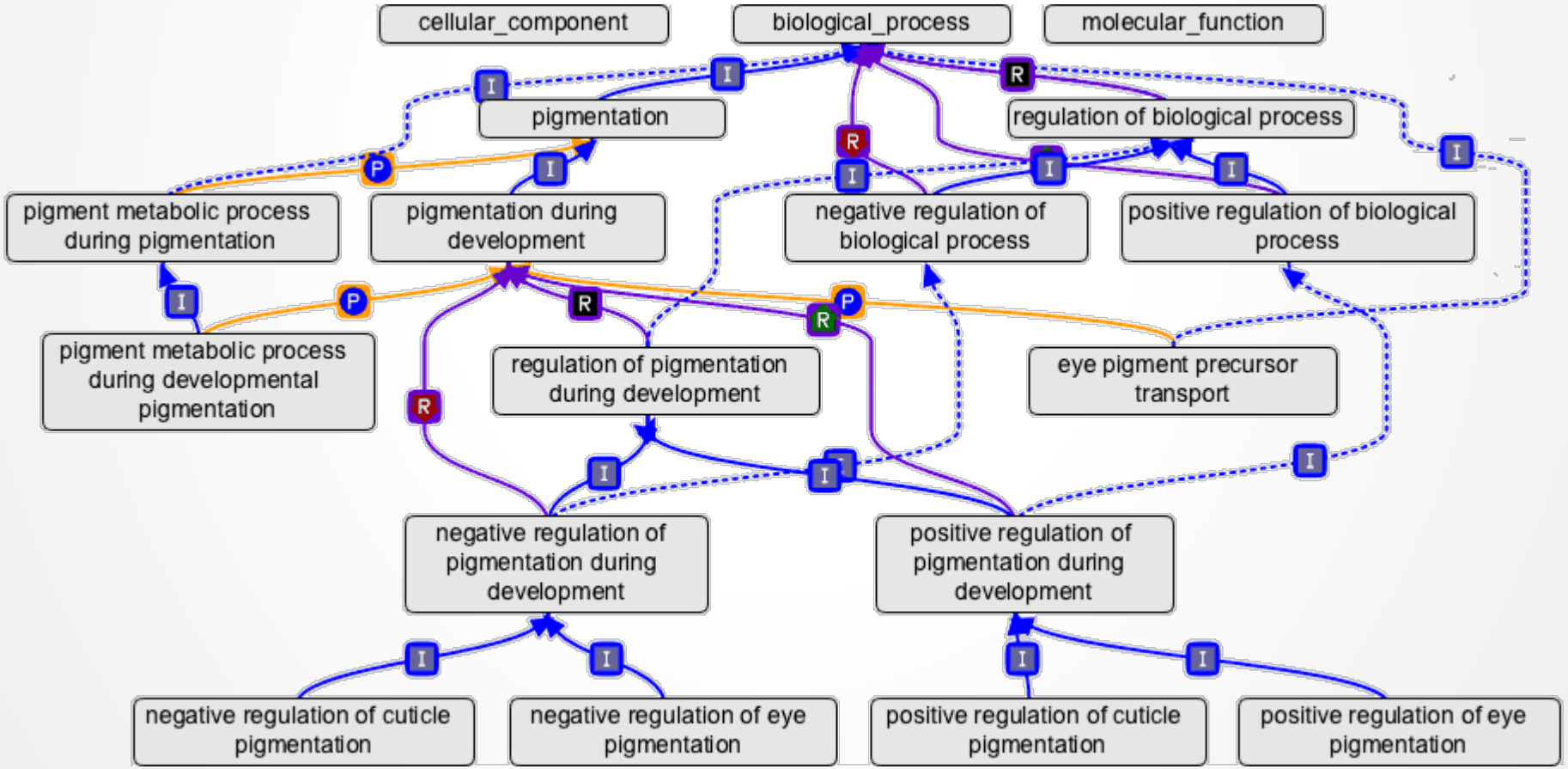


P-érték korrekció - Többszörös tesztelés

- rengetegszer végezzük el ugyanazt a tesztet → az összes génre vagy a chip minden probe setjére
- ha *p*=0.05-nél húzzuk meg a határt az azt jelenti, hogy 5% eséllyel elfogadunk szignifikánsan különbözőnek nem különböző expressziót is (a két kezelés között)
- → *False discovery rate* (FDR) korrekció az összes *p*-érték alapján. - pl: Bonferroni vagy Benjamini-Hochberg korrekció

Az RNA-seq analízis után: a DE gének funkciójának megismerése

- Gene Ontology - GO: http://geneontology.org



Köszönöm a figyelmet!