# Statistilisi probleeme arvutibioloogia, genoomika ja bioinformaatika valdkonnas

**ESS konverents "Statistika ja eluteadused"**


## Maido Remm

**Tartu Ülikool**
**Eesti Biokeskus**
**Genoomika tippkeskus**


14.aprill 2010

# Mis on genoom?

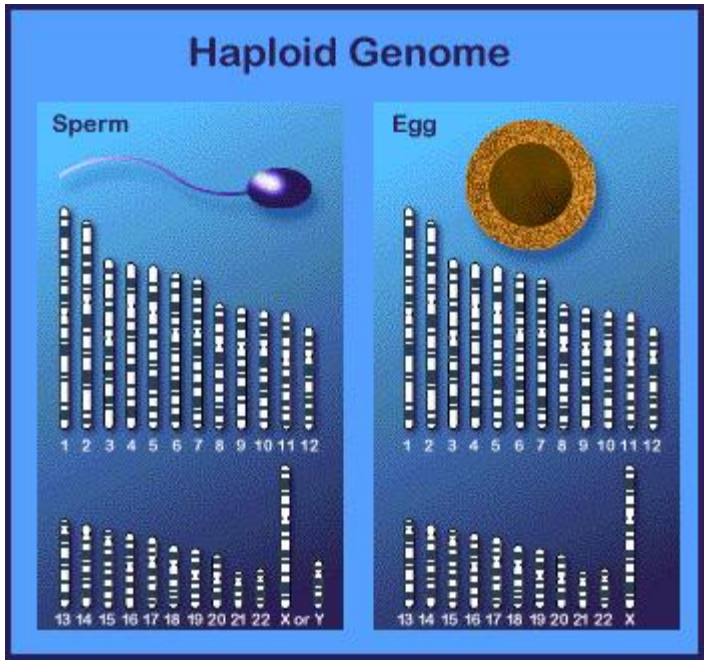
Genoom on **kogu** antud liigi või indiviidi **geneetiline materjal**.
Genoomi **järjestus** on kogu antud liigi või antud indiviidi geneetilise materjali järjestus.

Järjestus tähendab siin nukleotiidset järjestust, kus esineb 4 tähte: A, C, G, T

Genoomis on kirjas peaaegu kogu pärilik info.
Bakterigenoom koosneb paarist **miljonist** tähest, inimese genoom **3 miljardist** tähest.

Inimese rakkudes on enamuses rakkudes **2 genoomi koopiat** (isalt saadud genoom ja emalt saadud genoom).

Sugurakkudes on **1 genoomi koopia** (isa ja ema genoomide segu).



# Genoomi järjestuse näidis

**Esimene lehekülg *E.coli* genoomi järjestusest (jätkub ca 2000 leheküljel):**

AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTT
CTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATA
GGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACC
ACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACC
TGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGT
ACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCCAGGCAGGGGCAG
GTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGAAAAAACCATTAGC
GGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTTGACGGGACTCGCCGCC
GCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAAAACATGTC
CTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAA
ATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAA
AAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCA
AGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTG
GTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAG
ATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCG
ATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCC
ATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGT
GCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATGGCAATGTTCAGCGTT
TCTGGTCCGGGGATGAAAGGGATG

# Praeguseks järjestatud genoomide arv

(Oct 04, 2009)

| Organism | Complete | Draft assembly | In progress | total |
|---|---|---|---|---|
| Prokaryotes | 978 | 1044 | 1025 | 3047 |
| Archaea | 67 | 13 | 38 | 118 |
| Bacteria | 911 | 1031 | 987 | 2929 |
| Eukaryotes | 22 | 188 | 171 | 381 |
| Animals | 4 | 75 | 60 | 139 |
| Mammals | 2 | 28 | 19 | 49 |
| Birds | | 2 | 1 | 3 |
| Reptiles | | | 1 | 1 |
| Amphibians | | | 1 | 1 |
| Fishes | | 3 | 6 | 9 |
| Insects | 1 | 21 | 7 | 29 |
| Flatworms | | 2 | 2 | 4 |
| Roundworms | 1 | 9 | 11 | 21 |
| Other animals | | 12 | 14 | 26 |
| Plants | 2 | 11 | 46 | 59 |
| Land plants | 2 | 8 | 40 | 50 |
| Green Algae | | 3 | 6 | 9 |
| Fungi | 10 | 76 | 37 | 123 |
| Ascomycetes | 8 | 60 | 25 | 93 |
| Basidiomycetes | 1 | 10 | 8 | 19 |
| Other fungi | 1 | 6 | 4 | 11 |
| Protists | 6 | 24 | 24 | 54 |
| **Total:** | **1000** | **1232** | **1196** | **3428** |

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

# Genoomide varieeruvus

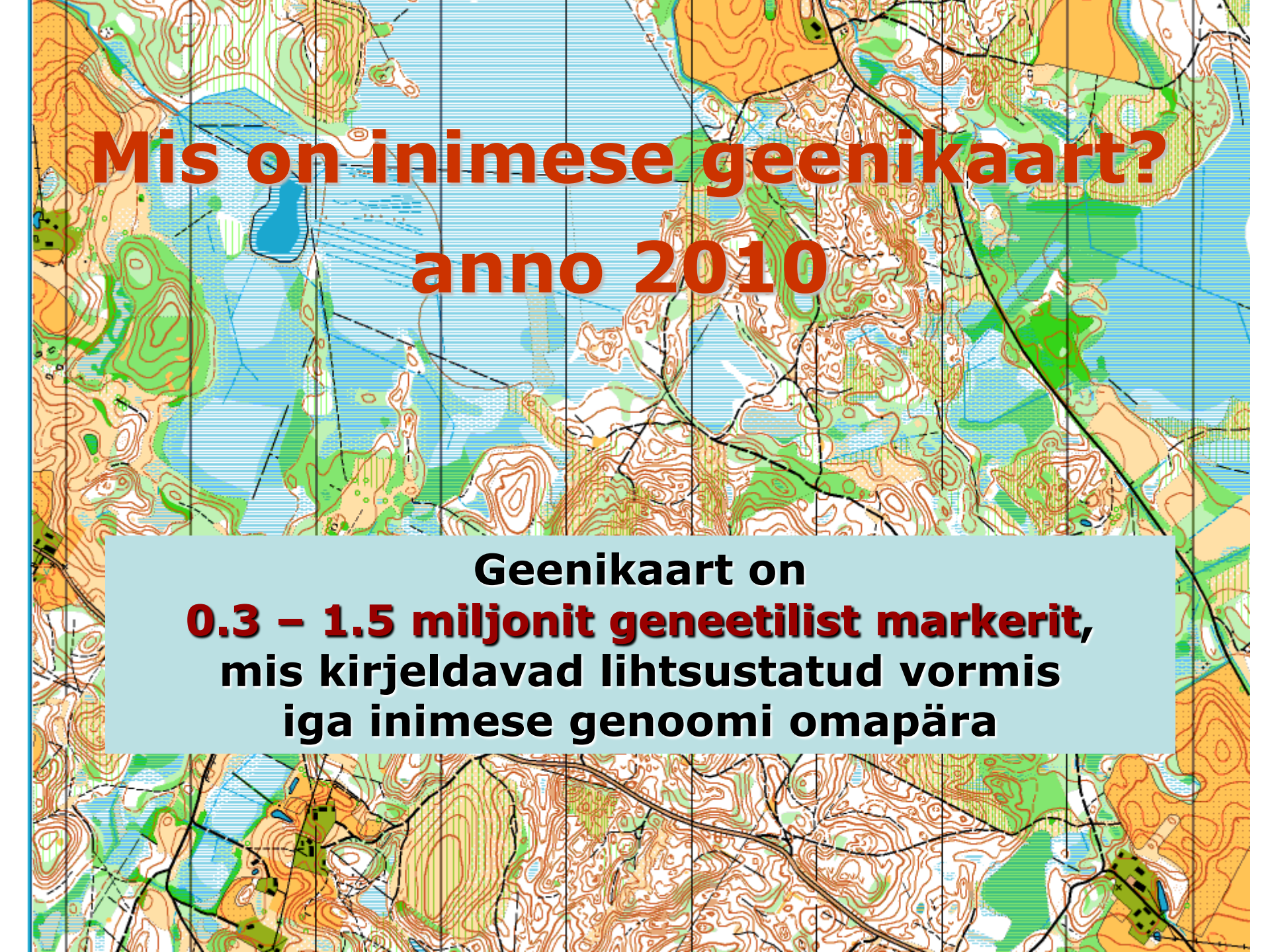
Mõned kohad genoomi järjestuses võivad olla varieeruvad – erineda liikide, populatsioonide ja üksikute inimeste vahel. Selle tõttu ongi iga liik, iga rahvas ja iga inimene omapärane ja teistest erinev.

**Geneetilise varieeruvuse** uurimine ongi praeguse aja üks olulisemaid uurimisteemasid genoomikas.

Varieerumise uurimiseks on vaja võrrelda rohkem kui ühte genoomi.

Genoomi **sekveneerimine** = genoomi nukleotiidse järjestuse määramine.
Genoomi **genotüpiseerimine** = järjestuse määramine üksikutes genoomi kohtades.
Inimeste genotüpiseerimise käigus koostatakse inimese **geenikaart.**

Geenikaart kirjeldab neid kohti genoomis, mis on kõige varieeruvamad ja potentsiaalselt huvi pakkuvad verieeruvuse uurimiseks.



# Mis on inimese geenikaart?
# anno 2010

**Geenikaart on**
**0.3 – 1.5 miljonit geneetilist markerit,**
**mis kirjeldavad lihtsustatud vormis**
**iga inimese genoomi omapära**

# Kiibilt saadud genotüübi-info (geenikaart)

**Ühe indiviidi geenikaart**

40 km

AA TT AA CC GG AA GC GG AG GA TC -- -- -- TC CC CC -- GG GG CC TT GG TC TC TC GA AA GG AA TC GG -- CC

ca 500 000 markers

emalt saadud nukleotiid

isalt saadud nukleotiid

teadmata väärtusega nukleotiid

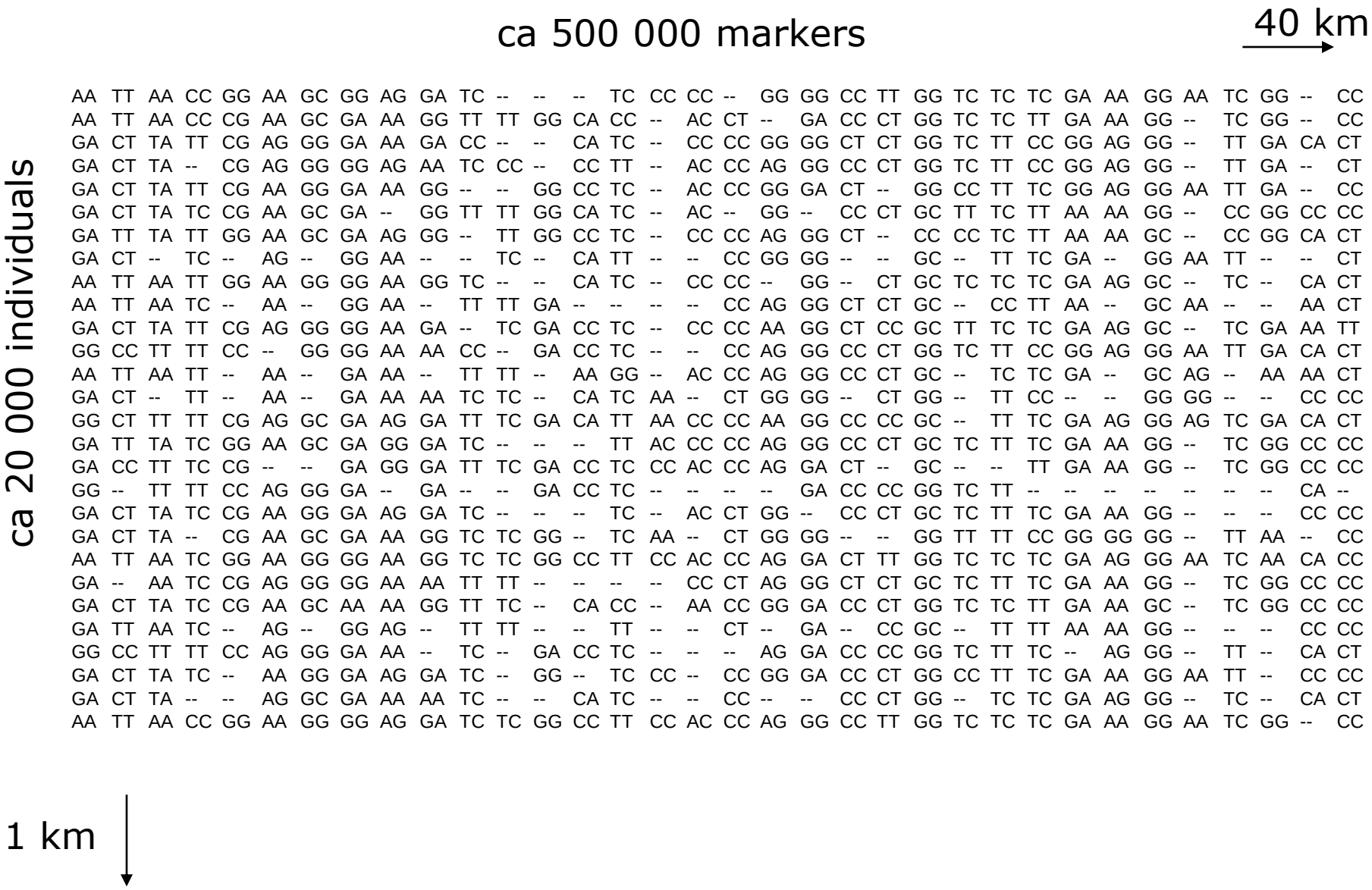
**Ühe indiviidi genoomi täisjärjestus**

300 000 km

AA TT AA CC CC GG TT GA TT GG AA CC -- CC TT TT TT CC AA GG TT AA TT AA TT TT CC AA CC AA TT CC TT CC TT CC

ca 3 000 000 000 nucleotides

# Kiibilt saadud genotüübi-info (geenikaart)



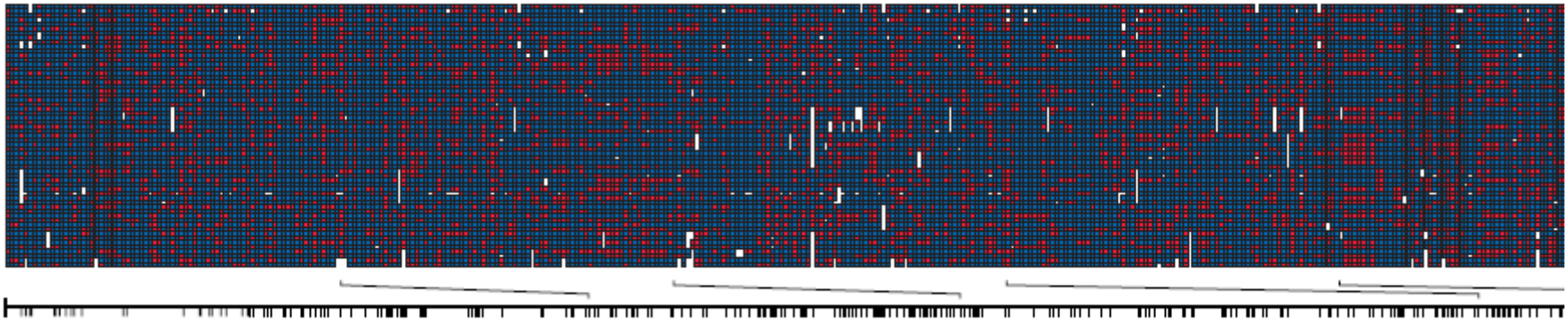

AA TT AA CC GG AA GC GG AG GA TC -- -- -- TC CC CC -- GG GG CC TT GG TC TC TC GA AA GG AA TC GG -- CC
AA TT AA CC CG AA GC GA AA GG TT TT GG CA CC -- AC CT -- GA CC CT GG TC TC TT GA AA GG -- TC GG -- CC
GA CT TA TT CG AG GG GA AA GA CC -- -- CA TC -- CC CC GG GG CT CT GG TC TT CC GG AG GG -- TT GA CA CT
GA CT TA -- CG AG GG GG AG AA TC CC -- CC TT -- AC CC AG GG CC CT GG TC TT CC GG AG GG -- TT GA -- CT
GA CT TA TT CG AA GG GA AA GG -- -- GG CC TC -- AC CC GG GA CT -- GG CC TT TC GG AG GG AA TT GA -- CC
GA CT TA TC CG AA GC GA -- GG TT TT GG CA TC -- AC -- GG -- CC CT GC TT TC TT AA AA GG -- CC GG CC CC
GA TT TA TT GG AA GC GA AG GG -- TT GG CC TC -- CC CC AG GG CT -- CC CC TC TT AA AA GC -- CC GG CA CT
GA CT -- TC -- AG -- GG AA -- -- TC -- CA TT -- -- CC GG GG -- -- GC -- TT TC GA -- GG AA TT -- -- CT
AA TT AA TT GG AA GG GG AA GG TC -- -- CA TC -- CC CC -- GG -- CT GC TC TC TC GA AG GC -- TC -- CA CT
AA TT AA TC -- AA -- GG AA -- TT TT GA -- -- -- -- CC AG GG CT CT GC -- CC TT AA -- GC AA -- -- AA CT
GA CT TA TT CG AG GG GG AA GA -- TC GA CC TC -- CC CC AA GG CT CC GC TT TC TC GA AG GC -- TC GA AA TT
GG CC TT TT CC -- GG GG AA AA CC -- GA CC TC -- -- CC AG GG CC CT GG TC TT CC GG AG GG AA TT GA CA CT
AA TT AA TT -- AA -- GA AA -- TT TT -- AA GG -- AC CC AG GG CC CT GC -- TC TC GA -- GC AG -- AA AA CT
GA CT -- TT -- AA -- GA AA AA TC TC -- CA TC AA -- CT GG GG -- CT GG -- TT CC -- -- GG GG -- -- CC CC
GG CT TT TT CG AG GC GA AG GA TT TC GA CA TT AA CC CC AA GG CC CC GC -- TT TC GA AG GG AG TC GA CA CT
GA TT TA TC GG AA GC GA GG GA TC -- -- -- TT AC CC CC AG GG CC CT GC TC TT TC GA AA GG -- TC GG CC CC
GA CC TT TC CG -- -- GA GG GA TT TC GA CC TC CC AC CC AG GA CT -- GC -- -- TT GA AA GG -- TC GG CC CC
GG -- TT TT CC AG GG GA -- GA -- -- GA CC TC -- -- -- -- GA CC CC GG TC TT -- -- -- -- -- -- -- CA --
GA CT TA TC CG AA GG GA AG GA TC -- -- -- TC -- AC CT GG -- CC CT GC TC TT TC GA AA GG -- -- -- CC CC
GA CT TA -- CG AA GC GA AA GG TC TC GG -- TC AA -- CT GG GG -- -- GG TT TT CC GG GG GG -- TT AA -- CC
AA TT AA TC GG AA GG GG AA GG TC TC GG CC TT CC AC CC AG GA CT TT GG TC TC TC GA AG GG AA TC AA CA CC
GA -- AA TC CG AG GG GG AA AA TT TT -- -- -- -- CC CT AG GG CT CT GC TC TT TC GA AA GG -- TC GG CC CC
GA CT TA TC CG AA GC AA AA GG TT TC -- CA CC -- AA CC GG GA CC CT GG TC TC TT GA AA GC -- TC GG CC CC
GA TT AA TC -- AG -- GG AG -- TT TT -- -- TT -- -- CT -- GA -- CC GC -- TT TT AA AA GG -- -- -- CC CC
GG CC TT TT CC AG GG GA AA -- TC -- GA CC TC -- -- -- AG GA CC CC GG TC TT TC -- AG GG -- TT -- CA CT
GA CT TA TC -- AA GG GA AG GA TC -- GG -- TC CC -- CC GG GA CC CT GG CC TT TC GA AA GG AA TT -- CC CC
GA CT TA -- -- AG GC GA AA AA TC -- -- CA TC -- -- CC -- -- CC CT GG -- TC TC GA AG GG -- TC -- CA CT
AA TT AA CC GG AA GG GG AG GA TC TC GG CC TT CC AC CC AG GG CC TT GG TC TC TC GA AA GG AA TC GG -- CC

# Kiibilt saadud genotüübi-info (geenikaart)

1500 markers

CEPH founders

60 individual chromosomes



## A first-generation linkage disequilibrium map of human chromosome 22

Elisabeth Dawson*†, Gonçalo R. Abecasis†‡§, Suzannah Bumpstead*, Yuan Chen*, Sarah Hunt*, David M. Beare*, Jagjit Pabial*, Thomas Dibling*, Emma Tinsley*, Susan Kirby*, David Carter*, Marianna Papaspyridonos*, Simon Livingstone*, Rocky Ganske‖, Elin Lõhmussaar¶#, Jana Zernant#, Neeme Tõnisson#, Maido Remm¶☆, Reedik Mägi#, Tarmo Puurand¶#, Jaak Vilo**, Ants Kurg¶, Kate Rice*, Panos Deloukas*, Richard Mott‡, Andres Metspalu¶☆, David R. Bentley*, Lon R. Cardon‡ & Ian Dunham*

Dawson et al. (2002) Nature 418: 544

# Genoomikas on peamiseks uurimisobjektiks DNA või valgu järjestus

**Tüüpilised andmestikud**

Kogu genoomi järjestus 1000 bakteriliigil

Kõikide geenide järjestus ühel bakteriliigil

Ühe inimese genoomi järjestus

Ühe inimese järjestus üksikutes kohtades (geenikaart)

**Tüüpilised küsimused**

Millised järjestused on seotud mingi organismi, rakkude või valkude bioloogilise omadusega (sõltuva tunnusega)?

Milline molekulaarne mehhanism seda omadust vahendab?

Raku ainevahetuse modelleerimine.

# Peamised uurimisteemad meie töögrupis

- **Inimese evolutsioon ja geneetika**

**- Suuremahuliste geen-omadus uuringute planeerimine ja analüüs**
**- Inimese ja ahvide genoomide võrdlus ning inimesele kui liigile omaste geenijärjestuste otsimine ja visualiseerimine**
**- Mendeli seadusi rikkuvate genoomi piirkondade uurimine**
**- Viiruste DNA järjestused inimese genoomis**

- **Võrdlev genoomika**

**- Tigude genoomide järjestamine ja analüüs**
**- Valgu sünteesi mõjutavate motiivide otsimine geenide järjestustes**
**- Bakterigenoomides olevate kordusjärjestuste leidmine ja kirjeldamine**
**- Liikide automaatne määramine metagenoomika uuringutes**

- **DNA-DNA seondumiste modelleerimine**

**- Polümeraasi ahelreaktsiooni (PCR) statistiline modelleerimine**
**- PCR praimerite disaini tarkvara täiendamine ja parandamine**
**- PCR praimerite valimine diagnostilisteks testideks**

# Statistilise modelleerimise kasutamine molekulaarbioloogias

**Millised järjestuse omadused on olulised uuritava tunnuse väljendumiseks (määravad uuritava tunnuse avaldumise)?**

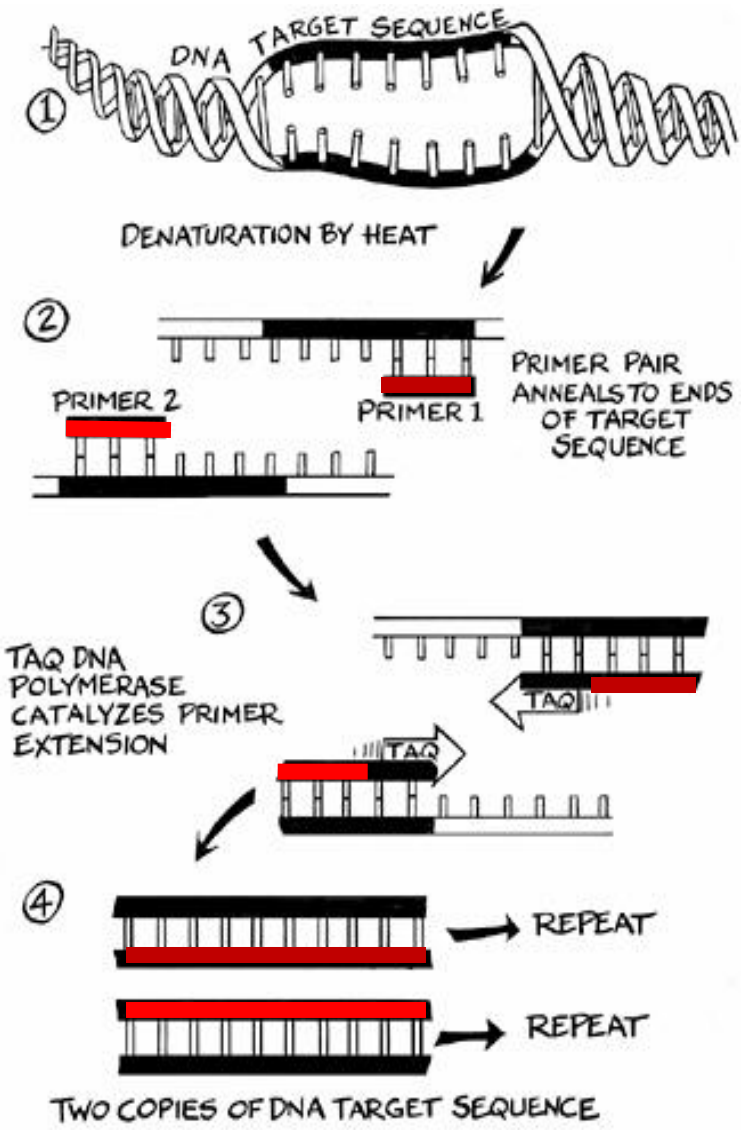
Peamine eesmärk on mõista protsessi olemust, mitte ennustada uuritavat tunnuse väärtust.

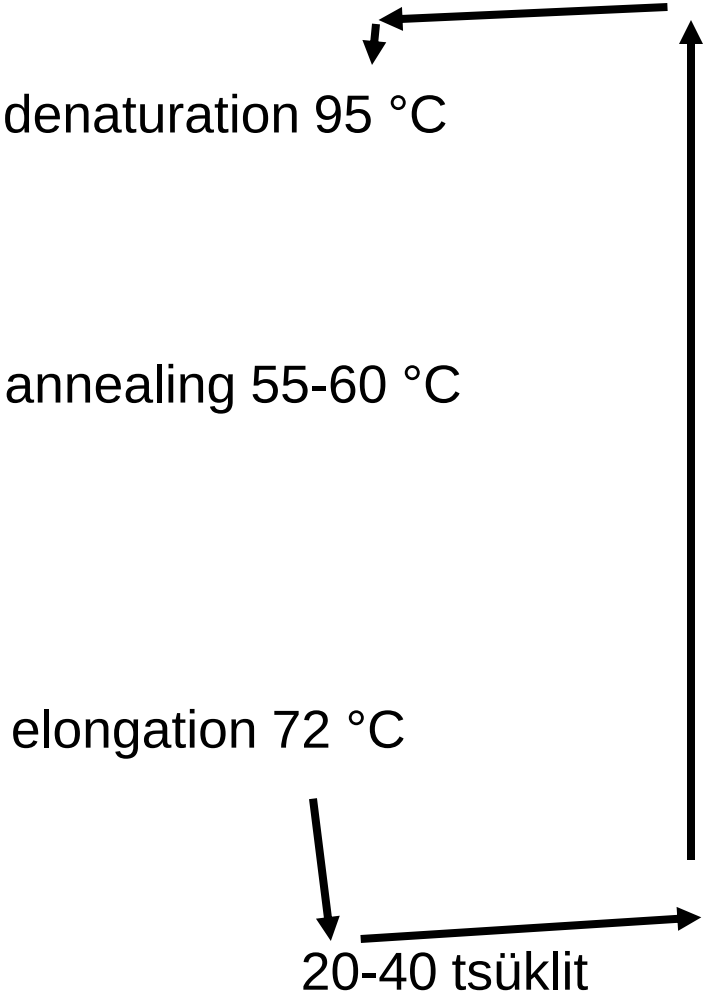
Selleks püüame aru saada milline faktor mõjutab uuritavat tunnust kõige rohkem. See võimaldab mõista milline molekulaarne mehhanism on seotud tunnuste avaldumisega;

Sõltumatuteks tunnusteks on näiteks järjestuse motiiv/muster, kindla aminohappe/nukleotiidi esinemine kindlas positsioonis, aminohapete või nukleotiidide biokeemilised või biofüüsilised omadused

Uuritavateks tunnusteks võib olla valgu tootmise efektiivsus antud geenilt, geeni tundlikkus antibiootikumiga mõjutamisele jms.

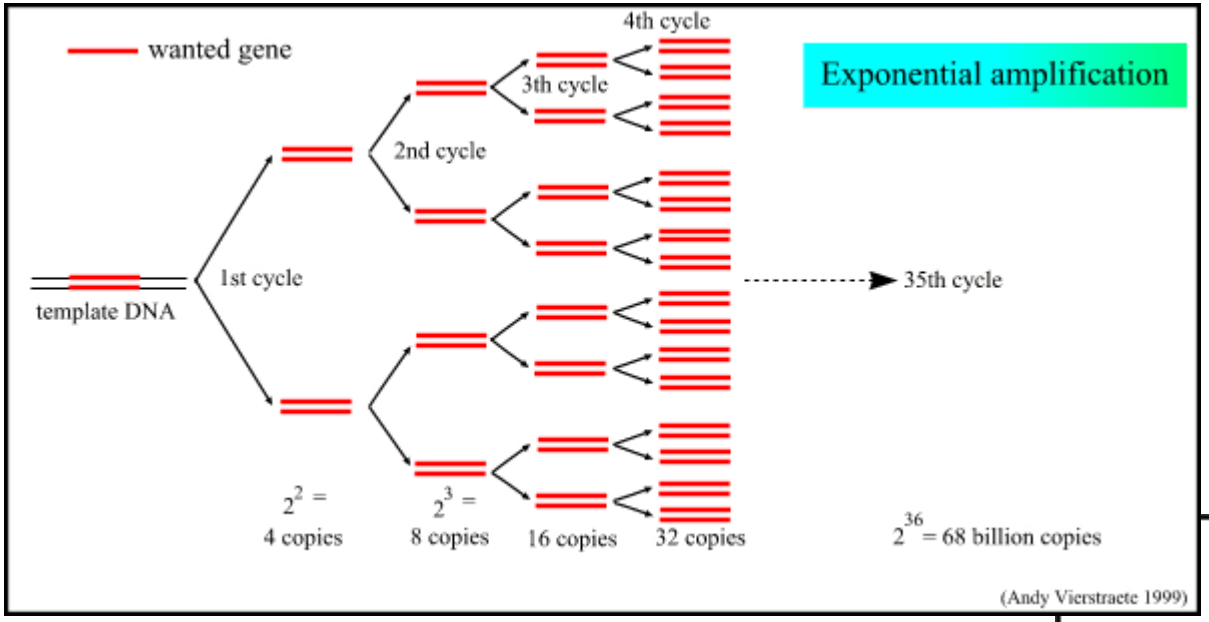
# Näide: Polümeraasi ahelreaktsiooni (PCR) modelleerimine

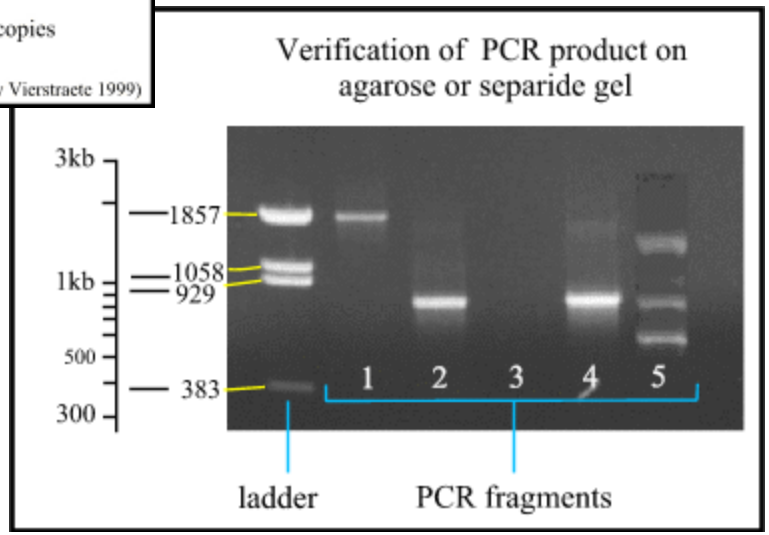



http://www.nap.edu/openbook/0309092051/html/121.html

# Näide: Polümeraasi ahelreaktsiooni (PCR) modelleerimine




http://users.ugent.be/~avierstr/principles/pcrgel.gif

# Näide: PCRi modelleerimine

**Andmestik:** DNA järjestuste (15-30 nukleotiidi pikad) ehk praimerite paarid, mida on kasutatud genoomse DNA paljundamiseks polümeraasi ahelreaktsioonis (PCR). Uuriti 1300 sõltumatut praimeripaari. Kokku 80 000 katset.

**Sõltuv tunnus:** Genoomse DNA paljundamise katseline edukus (näiteks - antud praimeripaar oli edukas 19 juhul 20-st).

**Sõltumatu tunnus:** ca 236 numbrit, mis praimerite järjestust teades on võimalik arvutada (ja mis peegeldavad PCRis toimuvaid füüsikalisi ja keemilisi protsesse).

**Otsesed eesmärgid:**

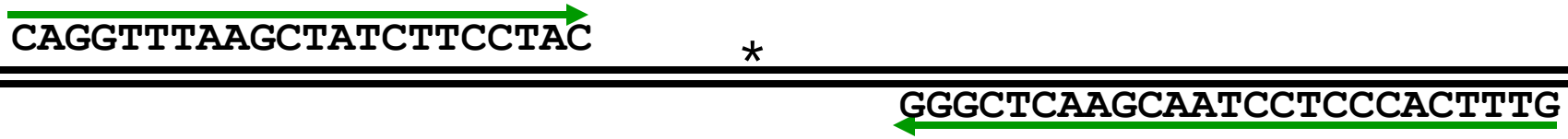
Milline on minimaalne arv tunnuseid, mis ennustab PCRi edukust?
Millised on kõige paremat ennustusjõudu omavad tunnused?

**Kaugem eesmärk:** Kirjutada arvutiprogramm, mis kiiresti valiks parima praimerite paari kasutaja poolt antud genoomi piirkonna paljundamiseks.

# PCRi bioinformaatiline probleem

Vajalik disainida 2 *praimerit*, mis seostuvad lähestikku uuritavale DNA-le, et nende abil uuritavat DNAd paljundada.

**Kuidas võimalike kandidaatide hulgast valida kõige paremad praimerid?**



# Andmestiku näidis:

| ID | PRAIMER_1 | PRAIMER_2 | PALJUNDATUD DNA LÕIK |
|---|---|---|---|
| 1573737 | GCATTTGATGACAGTTCATTTCAGCC | ACATGAATCTTGCGGACATTTTGTCA | GCATTTGATGACAGTTCATTTCAGCCTCTTTCCTGAAACTCCTTTTCTAGGCTTC... |
| 963515 | TGGACAGAAAGAACAATGAGTGTCCA | AACATTCTGGCAAACCCGAATTTCA | TGGACAGAAAGAACAATGAGTGTCCATGCCCCCCCACTGTCTTGGGCTGGGATCT... |
| 2482878 | TTCGGAATGCAATCTGATGCTGTTTT | TTAGATGAATGAGGCACCAAACACAA | TTCGGAATGCAATCTGATGCTGTTTTAGTGTTGGTGCAGAATGTGGTTTTAAAGG... |
| 1323851 | TCAAAAGAGTTCTTTGCAGAGTGCCA | TTGGAAAATTCACTTGGGTGGAACA | TCAAAAGAGTTCTTTGCAGAGTGCCATTTAAGGGCTCGTGCAGCCTTTAGAAGTT... |
| 1984178 | TTTTCATGTTGCCCTGCACCCTT | TATTTTCACTTTTCCCCACAGGCCAA | TTTTCATGTTGCCCTGCACCCTTCTTTCTACTCGATTTGTAGGAATGGGCAGCCA... |
| 3190 | TGGCAAAAGAAGCTTGGGTGATATTT | TTGTCAGCAATTAAGCTGCCATCAT | TGGCAAAAGAAGCTTGGGTGATATTTATGGGCAATAAATTTAAAAAAAACAGTTT... |
| 2383270 | TTATTCCATGTTGCCACAAATTGACA | CCGGGAGCTCAATCACTCATCATTT | TTATTCCATGTTGCCACAAATTGACAGCATAGTTGAAAGATGAATATTCCATGAT... |
| 2618657 | TGCTTCGTCACAGGGCCTATAATTT | CCTGCCACTTTAAATGGGAAGACACA | TGCTTCGTCACAGGGCCTATAATTTTTGATGGAAAGTTTTAAAAATGTATTCCAA... |
| 3003553 | TTCAACAAATCATTTTGGCCATCCA | TTCTGATTGCTCATCCATTCTTGCAT | TTCAACAAATCATTTTGGCCATCCAATAAATTTATTGAGAGGGTATATCCTCAAG... |
| 169757 | TTCAAAATGATGCCTGATGTTCCTGC | TTTATCCATCAGATTTTGCCGTGGA | TTCAAAATGATGCCTGATGTTCCTGCATACTGTGTTCCAAATTTAGGTAAATACA... |
| 2338975 | ATTTGGAATGAAGTCTGCTGGTCAAG | TTGTCCAGAATCTCCTTTGCTCATGC | ATTTGGAATGAAGTCTGCTGGTCAAGCTACTCATGGAAAAAAGCTCAGGTATTAC... |
| 303299 | AAAGAATGGATGAAATGTGACTGGCG | CAATTCAACCCTTGCAGAACATCACA | AAAGAATGGATGAAATGTGACTGGCGCCAGCTGAAATACTGTTGCAGCAATGTGA... |
| 303308 | AAAATGCCCTCTCTTCACTGCTTCCA | GCCCAGAATTTGGTCGATATTGTCA | AAAATGCCCTCTCTTCACTGCTTCCATTTAACAGTATACTACAGAGCCTAACCAG... |
| 212134 | CATTTTCACCCCACCAATGTCAACTT | TGGATTCTTGGCACCTTTGTTGAAT | CATTTTCACCCCACCAATGTCAACTTCTTCGTATCATTACACTATAAAAATTCCA... |
| 192766 | TTGGCCATCTTGCATCACTCTAAACA | AAGATGGGGATGCCATAACCCAAAAT | TTGGCCATCTTGCATCACTCTAAACATTTCTTTTTTAGCCCTAGTTCTTAATTCC... |
| 1818297 | TGAAGCACAGAGGGGAATTGTTTTCA | TGATTCTTCTGCTGGTTGATGTCGTT | TGAAGCACAGAGGGGAATTGTTTTCAATATAAATTTTTTTGATATAGCAGATATT... |
| 2338980 | TCTGGGAATCAGGGCCTAGATCAAAA | CGGGAGCATTCACTACAAACTGACTT | TCTGGGAATCAGGGCCTAGATCAAAAACTGTAAAAGTCTACCTAGTGTTTTATTG... |
| 212107 | TTGCAGTGAGCTGAGTGGAGCC | TTTTCGTGCTTTGCATCATGGC | TTGCAGTGAGCTGAGTGGAGCCACTGCACTCCAGCCTGGTGACAGAGCGAGACTC... |
| 212120 | TCCAGATGTGCTTTGCAGGACAGTTT | TGGCAAAGATACCTTCAGTCCAGCTT | TCCAGATGTGCTTTGCAGGACAGTTTCTCAGGCCCAGGATGCAGGCACACAGCTG... |
| 240447 | AGGCGGAGTCTCACTCTCACCGA | CCCAAACCCAAACCCACAATCAT | AGGCGGAGTCTCACTCTCACCGAGGCTGGAGTGCAGTGGAGCCATCTTGGCTCAC... |
| 469956 | TGGGAACATGTCAAGTCAAAAGGACA | GGTATTGCCAGATTTATTCGCTGACA | TGGGAACATGTCAAGTCAAAAGGACACAGGGGCTAGCTTTTAGGGAGTCCCACTG... |
| 468929 | AATGAGAACACGTGGACCCAGGAA | ACCCACAAAGGAAGCTGAACACTTGA | AATGAGAACACGTGGACCCAGGAAGGGGAACATCACACTCTGGGGACTGTTGTGG... |
| 371899 | TTAGCTAGGCATGGTGGTTGTGTGC | TGGTCATCTCCAAAGCCTTTCTTTGA | TTAGCTAGGCATGGTGGTTGTGTGCCTGTGGTCTCCAGCTACATCAGAGGCTGAG... |
| 469841 | TTCACCAAATGGTCAGCACAAATCA | CAAGGGTGTTGCAGATCTTCTGGCT | TTCACCAAATGGTCAGCACAAATCACCCCACTAAACTTACAAATCGACAAGCACC... |
| 2094140 | CTCAAGTTGAAAACTGCACAGCTGAA | CCAACAGAACTCAACAAAATCCCA | CTCAAGTTGAAAACTGCACAGCTGAAGATCACTTATAGTCAATAACACTTTTCAA... |
| 2879320 | TTCAATCAAACACACATCATCCTGGG | TCGTGATCTGCCTGCCACCTC | TTCAATCAAACACACATCATCCTGGGTATAAGTCTTGGGTGGGTCTAAAATTACA... |
| 1831133 | TTCCAGTGGCCATGACACTTTATTCA | GCATTCTCTCTCCTCTGTGCTCACTC | TTCCAGTGGCCATGACACTTTATTCAAATATGTAAGTTTTATTAAGACTGAGTTC... |
| 2838161 | TCATTTGAGATGAGTGTGCTGAGGAA | ACAAAGCCTTTTGTCAGCTTTCCTGT | TCATTTGAGATGAGTGTGCTGAGGAATAAAATGGTTAAGTAACTTGTCCAAGGTC... |

# Näide: PCRi modelleerimine

**Table 1.** The complete list of factors used in study for building models

| Factor description | Factor name | GM1 | GM1MM | GM2 | GM2MM | PCR | Number of factors |
|---|---|---|---|---|---|---|---|
| The number of binding sites of PCR primers (exact, with one and two mismatches allowed) with different word sizes from the 3′-end and from random positions in the primer sequence. | MAX/MIN[8,10,12,14,15,16]; | + | + | | | + | 12 |
| | MAX/MIN _FULL; | | | | | + | 2 |
| | MAX/MIN[8,10,12,14,15,16]_RAND; | | | | | + | 12 |
| | MAX/MIN[12,14,15,16]_1MM; | | + | | | + | 8 |
| | MAX/MIN_FULL_1MM; | | | | | + | 2 |
| | MAX/MIN[12,14,15,16] _1MM _RAND; | | | | | + | 8 |
| | MAX/MIN[12,14,15,16]_2MM; | | + | | | + | 8 |
| | MAX/MIN_FULL_2MM; | | | | | + | 2 |
| | MAX/MIN[12,14,15,16] _2MM _RAND | | | | | + | 8 |
| The number of binding sites of PCR primers (exact, with one and two mismatches allowed) with variable word sizes from the 3′-end and from random positions in the primer sequence. The word size for each primer is extended until three different free energy levels are achieved: $\Delta G < -10$, $-15$, $-20$ kcal/mol. | MAX/MIN_DG[10,15,20]; | | | + | + | + | 6 |
| | MAX/MIN_DG[10,15,20]_RAND; | | | | | + | 6 |
| | MAX/MIN_DG[10,15,20]_1MM; | | | | + | + | 6 |
| | MAX/MIN_DG[10,15,20]_1MM RAND; | | | | | + | 6 |
| | MAX/MIN_DG[10,15,20]_2MM; | | | | + | + | 6 |
| | MAX/MIN_DG[10,15,20]_2MM _RAND | | | | | + | 6 |
| The number of all binding sites of PCR primers (exact, with one and two mismatches allowed) counted with NCBI BLASTN (-F F). | [MAX,MIN]_BLASTALL | | | | | + | 2 |
| PCR primer length | PRIM_LENGTH_[MAX,MIN] | | | | | + | 2 |
| GC content of PCR primer with different word sizes from the 3′-end and full primer | PRIM_GC_PRC_[8,12,16]_[MAX,MIN]; | + | + | + | + | + | 6 |
| | PRIM_GC_PRC_[MAX,MIN] | | | | | + | 2 |
| The free energies of different sub-sequences from the primer 3′-end | PRIM_ DG[3,4,5,6,7,8,9]_[MAX,MIN] | + | + | + | + | + | 14 |

# Näide: PCRi modelleerimine

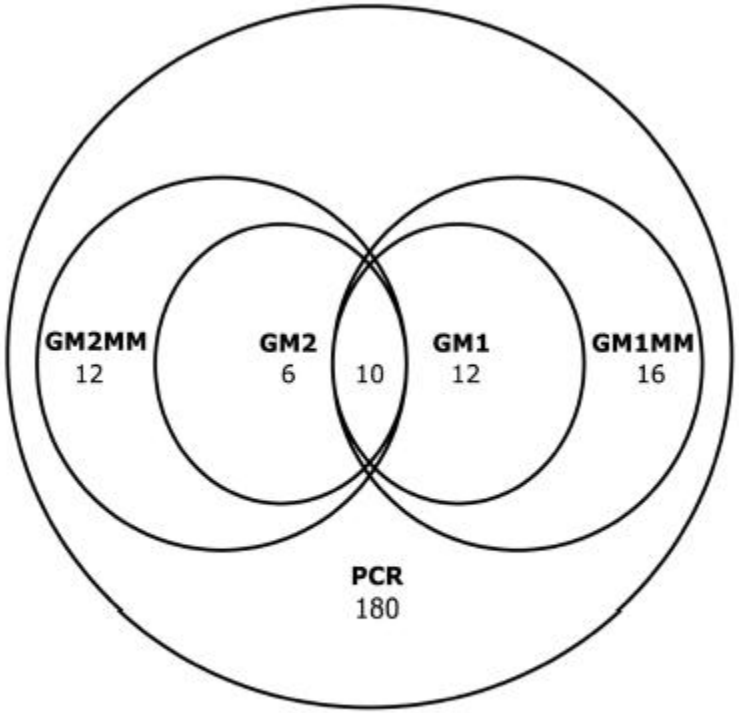
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| DUST score of PCR primer | PRIM_DUS_[MAX,MIN] | | | | | + | 2 |
| The strongest free energies of the dimers of primers alone and in pairs using local and global alignment approaches | MAX/MIN_PRIM_END1; | | | | | + | 2 |
| | PRIM_PAIR_END1; | | | | | + | 1 |
| | MAX/MIN_PRIM_END2; | | | | | + | 2 |
| | PRIM_PAIR_END2; | | | | | + | 1 |
| | MAX/MIN_PRIM_ANY; | | | | | + | 2 |
| | PRIM_PAIR_ANY | | | | | + | 1 |
| The strongest secondary structure of the PCR primers in a given pair predicted with MFOLD at 55°C | [MAX,MIN]_PRIM_MFOLD | | | | | + | 2 |
| The $T_m$ of the primer, difference of melting temperatures between the two primers in a given pair and the difference between annealing (used in PCR experiments) and melting temperature | TM_[MAX,MIN]; | | | | | + | 2 |
| | TM_DIFF; | | | | | + | 1 |
| | TM_TA_[MAX,MIN]_DIFF | | | | | + | 2 |
| Total number of SNPs in both primers and the position of the SNP closest to the 3′-end | NO_OF_SNPS; | + | + | + | + | + | 1 |
| | ALL_POS_FROM_3_END; | + | + | + | + | + | 1 |
| | NO_OF_VALID_SNPS; | + | + | + | + | + | 1 |
| | VALID_POS_FROM_3_END | + | + | + | + | + | 1 |
| The terminal and last two nucleotides of primer sequence, also the first nucleotide of amplicon following the primer sequence. These are categorical values (0 – given nuc. is not present in both primers, 1 – is present at least in one primer, 2 – is present in both primers). | PRIM_LAST_ONE_NUC_[A,C,G,T]; | | | | | + | 4 |
| | PRIM_LAST_TWO_NUC_[AA,AC, AG,AT,CC,CG,CT,GG,GT,TT]; | | | | | + | 10 |
| | PROD_FIRST_ONE_NUC_[A,C,G,T] | | | | | + | 4 |
| The number of predicted products with maximum length of 1000, 3000 and 10 000 nt for exact | PROD[8,10,12,14,15,16]_1000; | | | | | + | 6 |
| | PROD_FULL_1000; | | | | | + | 1 |
| | PROD[8,10,12,14,15,16]_1000_RAND; | | | | | + | 6 |

(continued)

# Näide: PCRi modelleerimine

| | | | |
|---|---|---|---|
| binding sites with different word sizes from the 3′-end and from random positions in the primer sequence. | PROD[8,10,12,14,15,16]_3000; | + | 6 |
| | PROD_FULL_3000; | + | 1 |
| | PROD[8,10,12,14,15,16]_3000_RAND; | + | 6 |
| | PROD[8,10,12,14,15,16]_10000; | + | 6 |
| | PROD_FULL_10000; | + | 1 |
| | PROD[8,10,12,14,15,16]_10000_RAND | + | 6 |
| The number of predicted products with maximum length of 1000, 3000 and 10 000 nt for exact binding sites with variable word sizes from the 3′-end and from random positions in the primer sequence. The word size for each primer is extended until three different free energy levels are achieved: $\Delta G < -10$, $-15$, $-20$ kcal/mol. | PROD_DG[10,15,20]_1000; | + | 3 |
| | PROD_DG[10,15,20]_1000_RAND; | + | 3 |
| | PROD_DG[10,15,20]_3000; | + | 3 |
| | PROD_DG[10,15,20]_3000_RAND; | + | 3 |
| | PROD_DG[10,15,20]_10000; | + | 3 |
| | PROD_DG[10,15,20]_10000_RAND | + | 3 |
| PCR product length | PROD_LENGTH | + | 1 |
| GC content of PCR product | PROD_GC_PRC | + | 1 |
| Area under the GC curve and above 65% of the PCR product (7) | PROD_AUCGC | + | 1 |
| Number of GC windows with values above 65% divided by the length of the PCR product ($\times 100$) (7) | PROD_RATIOGC_100 | + | 1 |
| PROD_AUCGC × PROD_RATIOGC (7) | PROD_AUCGC2 | + | 1 |
| The strongest secondary structure of PCR product predicted with MFOLD at 55°C | PROD_MFOLD_55 | + | 1 |
| Percentage of masked nucleotides of PCR product using DUST | PROD_DUST_PRC | + | 1 |
| Percentage of masked nucleotides of PCR product using Repeat Masker with different sensitivity parameters (-s, -q, -qq) | PROD_RMs_PRC; | + | 1 |
| | PROD_RMq_PRC; | + | 1 |
| | PROD_RMqq_PRC | + | 1 |
| Percentage of masked nucleotides of PCR product using GenomeMasker with different word sizes (exact matches) | PROD_GM[8,10,12,14,16]_PRC | + | 5 |
| **Total number of factors** | | | **236** |

# Näide: PCRi modelleerimine




**Total: PCR**=236 **GM2MM**=28 **GM2**=16 **GM1MM**=38 **GM1**=22

Figure 1. The distribution of factors between different model types.

Programmi kirjutamiseks oli võimalik valida erineva keerukusega algoritme. Seetõttu võrdlesime 4 erinevat tunnuste komplekti ja lisaks täiskomplekti.

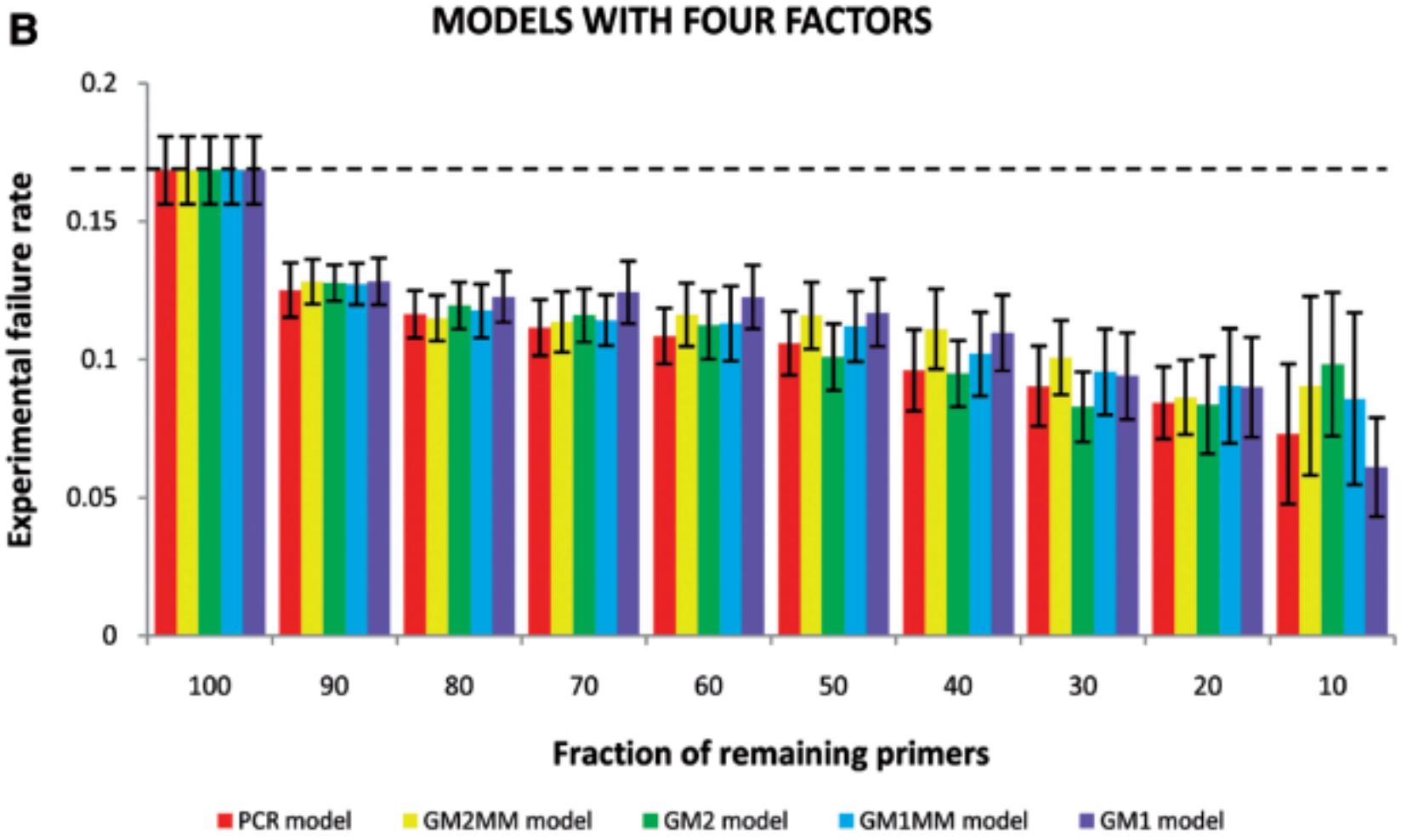
# Näide: PCRi modelleerimine

**Table 2.** List of the best factors (top 4) and the corresponding one-degrees-of-freedom chi-squares ($\chi^2(1)$) from the GENMOD Type I analysis using whole dataset

| Factor name | $\chi^2$ (1) | Model |
|---|---|---|
| MAX_DG15_2MM | 4862 | **PCR** |
| MAX_DG15_RAND*MAX_DG15_RAND | 1374 | |
| PROD_DG20_1000_RAND* PROD_DG20_1000_RAND | 378 | |
| PROD_LENGTH*PROD_LENGTH | 298 | |
| MAX_DG15_2MM | 4862 | **GM2MM** |
| MAX_DG15_1MM*MAX_DG15_1MM | 1091 | |
| MAX_DG20_2MM | 244 | |
| MAX_DG20_1MM*MAX_DG20_1MM | 262 | |
| MAX_DG20 | 4085 | **GM2** |
| MAX_DG15*MAX_DG15 | 1106 | |
| PRIM_GC_PRC_8_MIN | 386 | |
| MIN_DG20*MIN_DG20 | 277 | |
| MAX15_2MM | 2854 | **GM1MM** |
| MAX12_1MM*MAX12_1MM | 1681 | |
| MAX12 | 1291 | |
| PRIM_GC_PRC_16_MAX | 789 | |
| MAX16 | 2507 | **GM1** |
| MAX15*MAX15 | 2394 | |
| PRIM_GC_PRC_16_MAX | 1126 | |
| MAX14 | 272 | |

All factors are significant at $P < 0.0001$. Asterisks in factor names mark the polynomial regression of given independent variable. $\chi^2$-values illustrate the estimated simultaneous (Type I) effects of the best four factors on each model.

# Näide: PCRi modelleerimine



# Näide: PCRi modelleerimine

**Leidsime, et kõige efektiivsem on ennustada PCRi edukust nelja tunnuse abil, mis on kombineeritud mudeliks GM1.**

**Hinnanguliselt vähendab meie mudel PCRi ebaõnnestumise sagedust kuni 3x (17%-lt 6%-ni).**

*Published online 20 May 2008*

*Nucleic Acids Research, 2008, Vol. 36, No. 11* ***e66***
*doi:10.1093/nar/gkn290*

## Predicting failure rate of PCR in large genomes

**Reidar Andreson[1,2], Tõnu Möls[1] and Maido Remm[1,2,*]**

[1]Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu and
[2]Estonian Biocentre, Tartu, Estonia

Received March 12, 2008; Revised April 24, 2008; Accepted April 28, 2008

# Tulemuste tõlgendamine

**Saadud tulemustest saab ka hinnata, millised molekulaarsed mehhanismid mõjutavad tulemust (PCRi edukust).**

**Sarnase metoodikaga on analüüsitud ka muid olulisi bioloogilisi protsesse.**
**Näiteks leiti et peamiseks faktoriks, mis mõjutab valkude tootmise efektiivsust rakus on mRNA algusotsa sekundaarstruktuur:**

## Coding-Sequence Determinants of Gene Expression in *Escherichia coli*

Grzegorz Kudla,[1]* Andrew W. Murray,[2] David Tollervey,[3] Joshua B. Plotkin[1]†

www.sciencemag.org SCIENCE VOL 324 10 APRIL 2009 

# Probleemid statistika rakendamisel

**Mõned probleemid nede olulisuse kasvavas järjekorras:**

- **tunnuste jaotus pole ei normaalne ega pidev;**
- **sõltumatud tunnused ei ole tegelikult sõltumatud;**
- **andmestikus on segiläbi kategooriana esinevad tunnused ja pidevad tunnused;**
- **palju uuritavaid tunnuseid, vähe katseid;**
- **halvasti kirjeldatud tunnused, mürasisaldus >50%;**
- **statistilise hariduse puudumine katse planeerijatel**