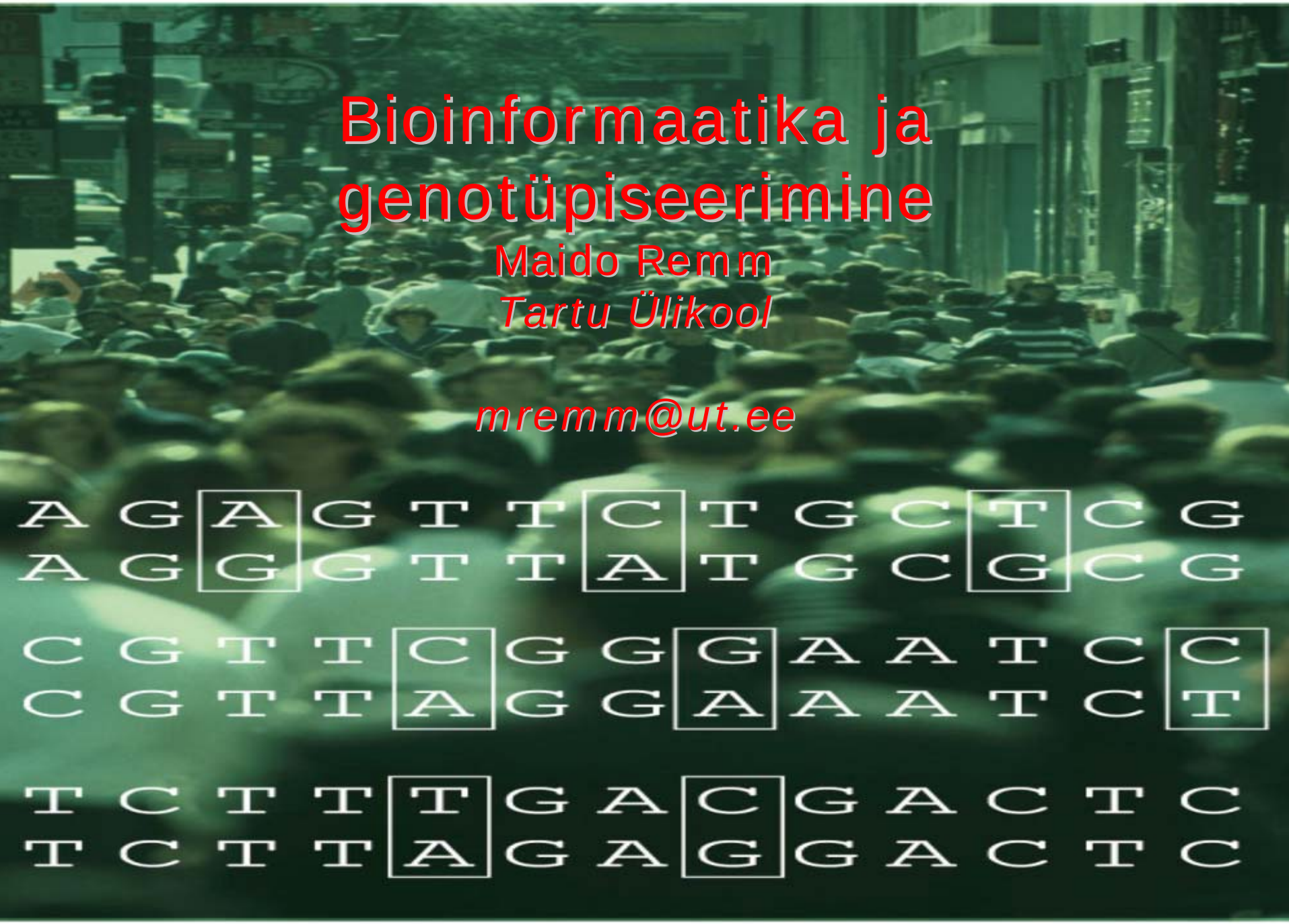
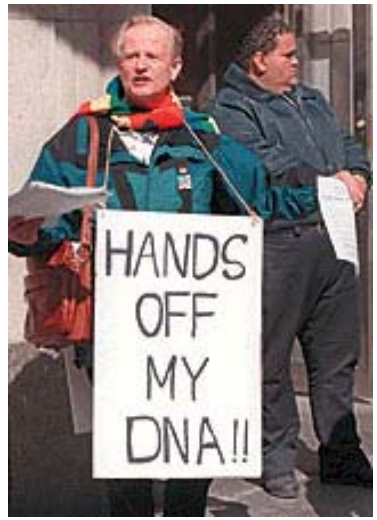
# Bioinformaatika ja genotüpiseerimine

Maido Remm
*Tartu Ülikool*

*mremm@ut.ee*



# Mis on genotüpiseerimine?




Indiviidi genoomi analüüs indiviidide (inimeste) omavahelise erinevuse leidmiseks

Inimese genoom 3 000 000 000 tähte (nukleotiidi)

Seni on uuritud peamiselt polümorfisme, mis on sagedasemad kui 10%. Neid on ca 3 000 000, seega iga 1000 nukleotiidi tagant võib esineda üks erinevus.

Need muutused ei teki igal inimesel ega ka paari viimase põlvkonna jooksul, vaid pigem peegeldavad populatsioonide ajalugu

Arvatakse, et liithaiguste põhjuste leidmisel oluline just sagedaste polümorfismide analüüs

# Genotüpiseerimine

**Terminid**

- **SNP**: Ühenukleotiidiline marker inimese genoomis. On toimunud 1 nukleotiidi asendus. Enamusel SNPdel on ainult 2 alleeli (bialleelne marker). Esineb ka teistsuguseid markereid (mikrosatelliidid, insertsioonid/deletsioonid), kuid massgenotupiseerimiseks need ei sobi.

- **Alleel**: Üks antud SNP variantidest. Kõikide alleelide sagedused kokku on 100%. SNPde puhul on tavaliselt ainult 2 alleeli.

- **Genotüüp**: Segu mõlema kromosoomi alleelist. SNPde puhul on tavaliselt 3 võimalikku genotüüpi AA, AB, BB ehk siis 11,12,22

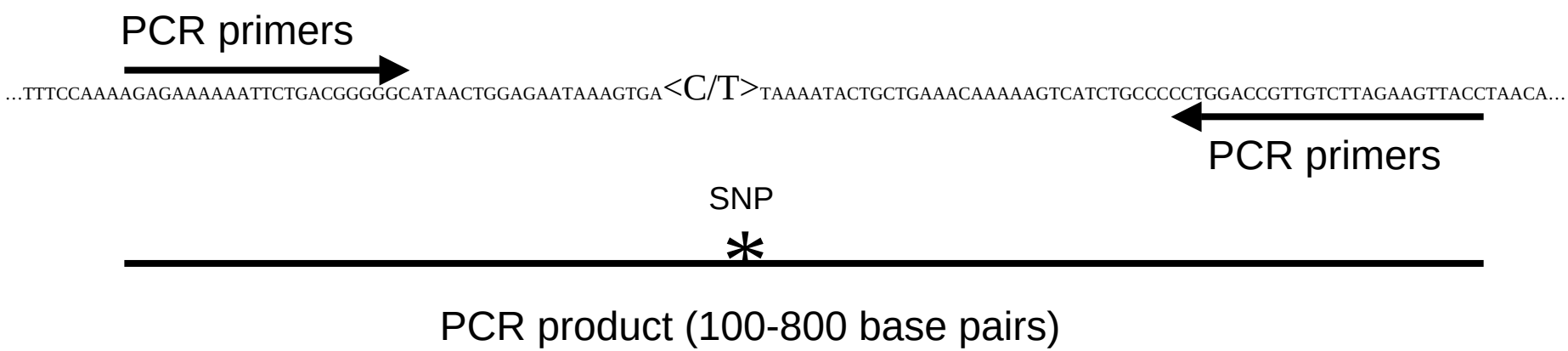
- **Haplotüüp**: Ühel kromosoomil järjestikku esinevad alleelid. Eksperimentaalselt on lihtsam määrata genotüüpe kui haplotüüpe. Haplotüüpide määramiseks on vajalik kromosoomid füüsiliselt eraldada või analüüsida perekondade genotüüpe.

- **Haplotüübi blokk**: Järjestikku esinev konserveerunud alleelide blokk. Tüüpiline haplotüübi blokk on vahemikus 1-100 kb, keskmiselt 10 kb pikkune.

# Kuidas genotüpiseerida?

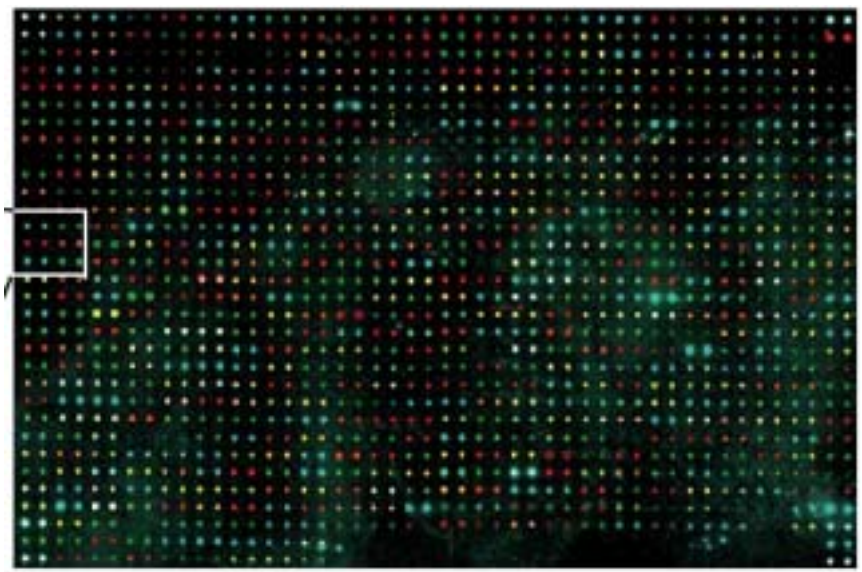
1. PCR (polymerase chain reaction)

PCR on vajalik genoomi osade võimendamiseks, amplifitseerimiseks.
Võimendav efekt ca $2^n$ , kus n on PCRi tsüklite arv

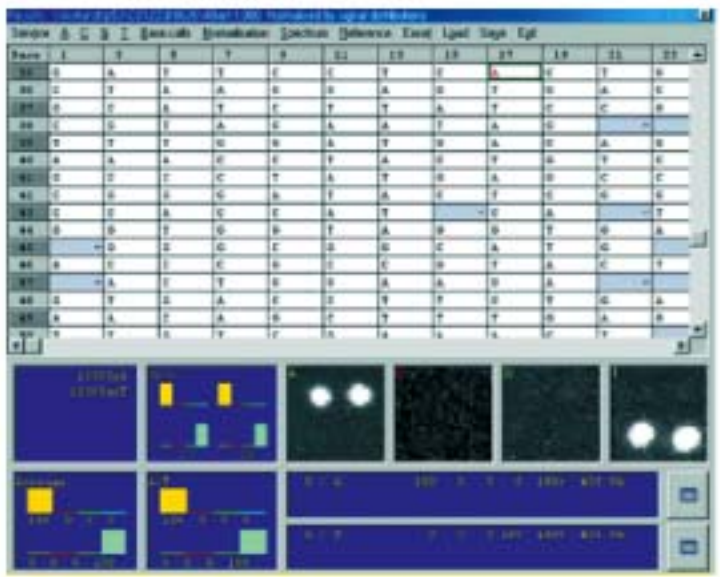


# Kuidas genotüpiseerida?

1. PCR (polymerase chain reaction)
2. PCR produktide genotüpiseerimine (näiteks mikrokiibil)
3. Pildianalüüs (base calling, signal detection)
4. Genotüüpide analüüs
   (evolutsiooni uurimine või haigusgeenide leidmine)



ABCR gene mutation detection assay for simultaneous screening of 350 mutations was designed for Columbia University. The image above is showing pseudocolor signals (A-yellow, C-red, G-green, T-cyan) and the image on the left is showing its zoomed fragment of the separate A, T, C, G channels.



Signal intensities of four images are compared and automatically translated into the presence of a nucleotide(s) on the image and analyzed sample.

# Bioinformaatika rakendusi genotüpiseerimisel

- ✓ Uuritavate markerite valik:
  - millised piirkonnad genoomis? (sidumine genoomi andmebaasiga)
  - kuidas valida ühtlase tihedusega markereid?
  - milliseid markereid valida? (maksimaalne infosisaldus)
- ✓ Tehniline teostus:
  - PCR praimerite disain
  - PCR praimerite unikaalsuse kontroll genoomis
  - PCR multiplex
  - APEX praimerite sobivus genotüpiseerimiseks
- ✓ Geneetiline epidemioloogia:
  - fenotüüp-genotüüp seoste leidmine

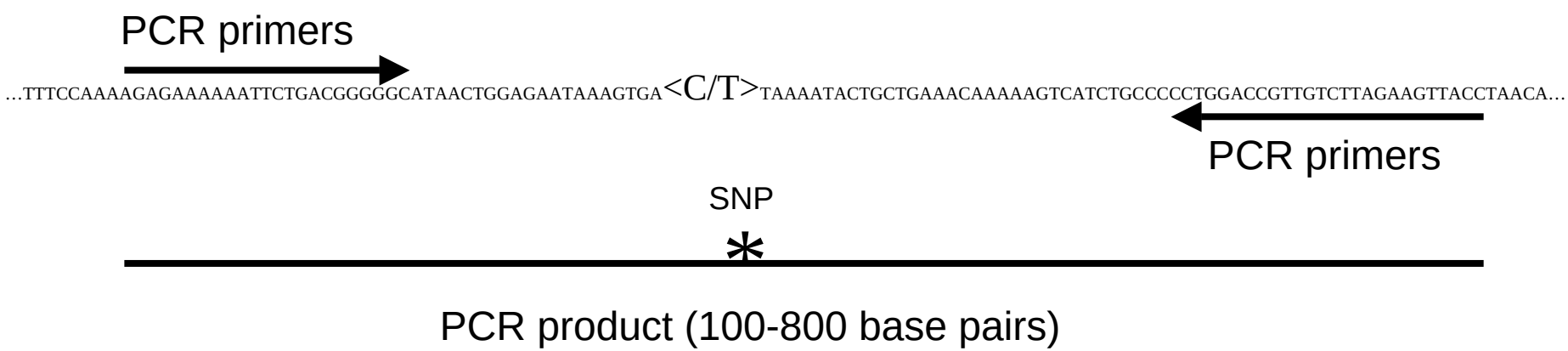
# Bioinformaatika rakendusi genotüpiseerimisel

- ✓ Uuritavate markerite valik:
  - millised piirkonnad genoomis? (sidumine genoomi andmebaasiga)
  - kuidas valida ühtlase tihedusega markereid?
  - milliseid markereid valida? (maksimaalne infosisaldus)
- ✓ Tehniline teostus:
  - PCR praimerite disain
  - PCR praimerite unikaalsuse kontroll genoomis
  - PCR multiplex
  - APEX praimerite sobivus genotüpiseerimiseks
- ✓ Geneetiline epidemioloogia:
  - fenotüüp-genotüüp seoste leidmine

# PCR

1. PCR (polymerase chain reaction)

PCR on vajalik genoomi osade võimendamiseks, amplifitseerimiseks.
Võimendav efekt ca $2^n$ , kus n on PCRi tsüklite arv



# PCR praimerite disain

Disainimise käigus leitakse 2 prraimerit (praimerite paar), mis seostuvad sobivasse kohta sobiva tugevusega andes sobiva pikkusega produkti.

Praimer on varieeruva pikkusega 16-26 nt.

Seostumist (seostumise tugevust) võib modelleerida mitmel erineval viisil:

1. evolutsioonilise kauguse kaudu (terve praimer või fikseeritud pikkusega substring)

(edit distance: +1 match, -1 mismatch, -2 gap).

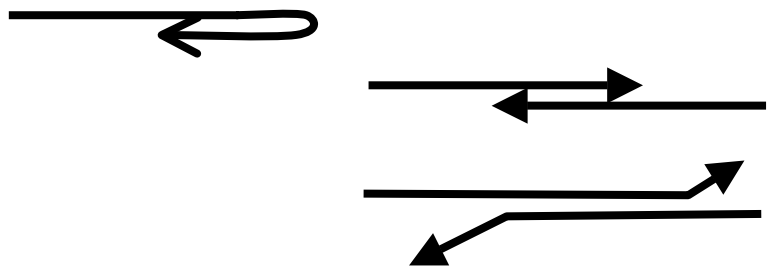
2. arvestada füüsilist seondumistugevust (varieeruva pikkusega substring).

J SantaLucia, JR. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. PNAS 95:1460.

# PCR praimerite kontroll genoomi vastu

Praimerite disaini juures on oluline jälgida et praimer ei seostuks valesse kohta:

- Iseenda külge
- Teise praimeri külge

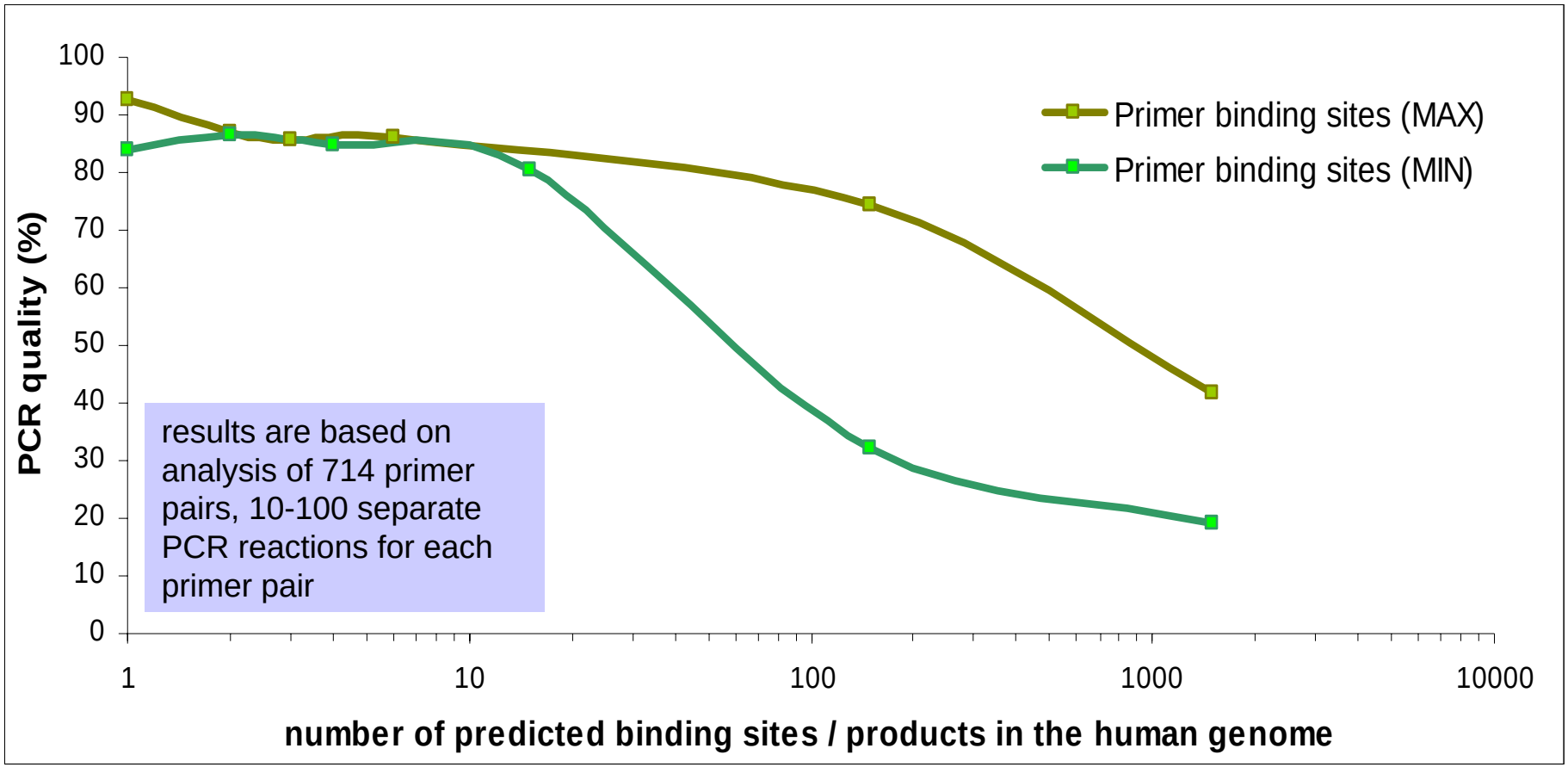


- Valesse kohta genoomis

Seega on oluline seostumise spetsiifilisus:

Praimer on ca 20 nukleotiidi pikk
Genoom 3 * $10^9$ nukleotiidi pikk

# PCR praimerite genoomitest




On oluline kontrollida genoomse PCRi praimerid genoomi vastu, et vältida kvaliteedi langust ja lisaprodukide teket!

# PCR praimerite genoomitest

Jõudsime järeldusele, et praimerite disaini juures on oluline jälgida et praimer(id) ei seostuks valesse kohta genoomis:

**Kuidas vältida valesid seostumisi?**

- Nimekiri kordusjärjestustest (PRIMER3) või
- Nimekiri sagedasematest 6-meersetest oligotest (OLIGO6)

**Selline lähenemine suurendab praimerite kvaliteeti, kuid ei garanteeri, et saadakse unikaalne PCR produkt**

**Tuleks leida kõik praimerite seondumiskohad genoomis!**
**Kuidas?**

1. BLAST (väga aeglane, probleem suuremate kromosoomidega)
2. MEGABLAST (aeglane)
3. SSAHA (kiire, kuid nõuab vähemalt 1 GB mälu)
4. GenomeTester (väga kiire nii väikese kui suure praimerite koguse korral)

# Genome test with 1280 PCR primer pairs

**BLAST (-F F):**
1 week on single processor, 1.5 GHz Pentium4

**MEGABLAST:**
> 2 hours, 1GB RAM

**SSAHA:**
45 minutes
12 GB disk space
1 GB of RAM

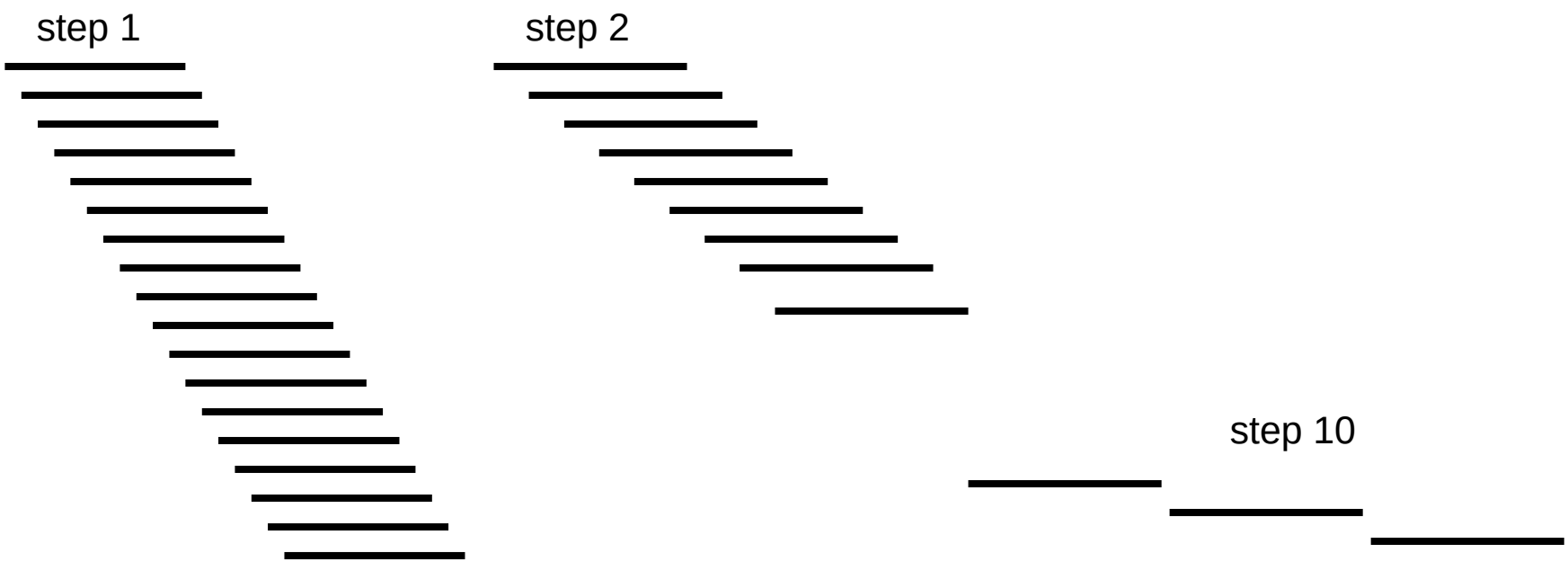
**GenomeTester:**
15 minutes
500 MB RAM
20 GB disk space
16 nt. from the primer's 3'end
all 16 nucleotide words from the human genome

# SSAHA

Sequence Search and Alignment by Hashing Algorithm

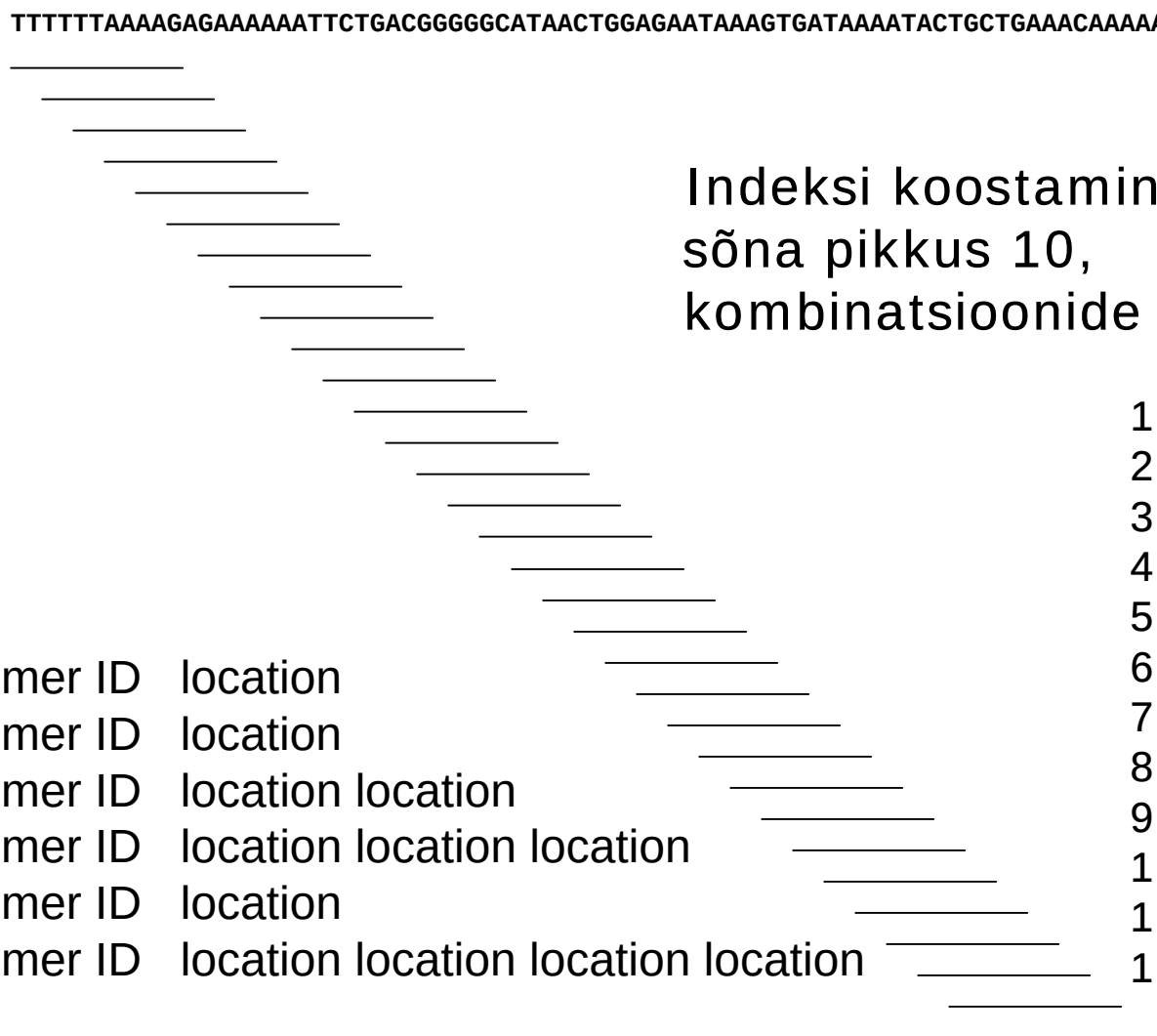
Koostab tabeli (indeksi) kõigist genoomis olevatest “sõnadest” ja jätab meelde nende asukoha genoomis

Tüüpiline sõna pikkus 10 nt.



# SSAHA

## Sequence Search and Alignment by Hashing Algorithm




Indeksi koostamine:
sõna pikkus 10,
kombinatsioonide arv $4^{10} = 10^6$

1: 16345, 23881,5543906
2:
3: 554262,
4: 777624, 2228511, ...
5:
6:
7:
8:
9:
10:
11:
12:

10-mer ID location
10-mer ID location
10-mer ID location location
10-mer ID location location location
10-mer ID location
10-mer ID location location location location

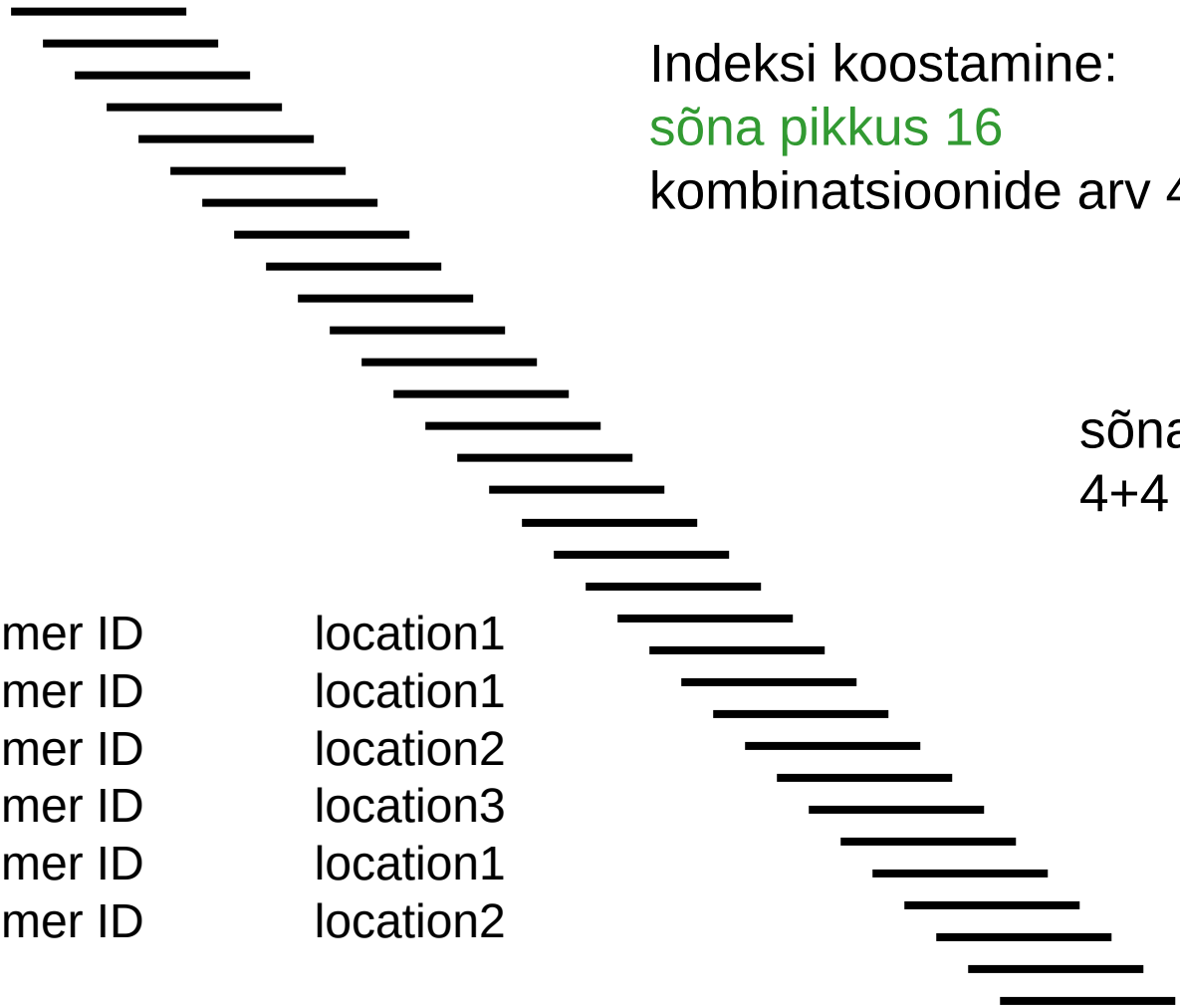
# SSAHA

## Sequence Search and Alignment by Hashing Algorithm


sõnade koguarv $3*10^9$ ,
4 bytes each = 12 GB

# Genome Tester:

TTTTTTAAAAGAGAAAAAATTCTGACGGGGGCATAACTGGAGAATAAAGTGATAAAATACTGCTGAAACAAAAAGTCATCTG

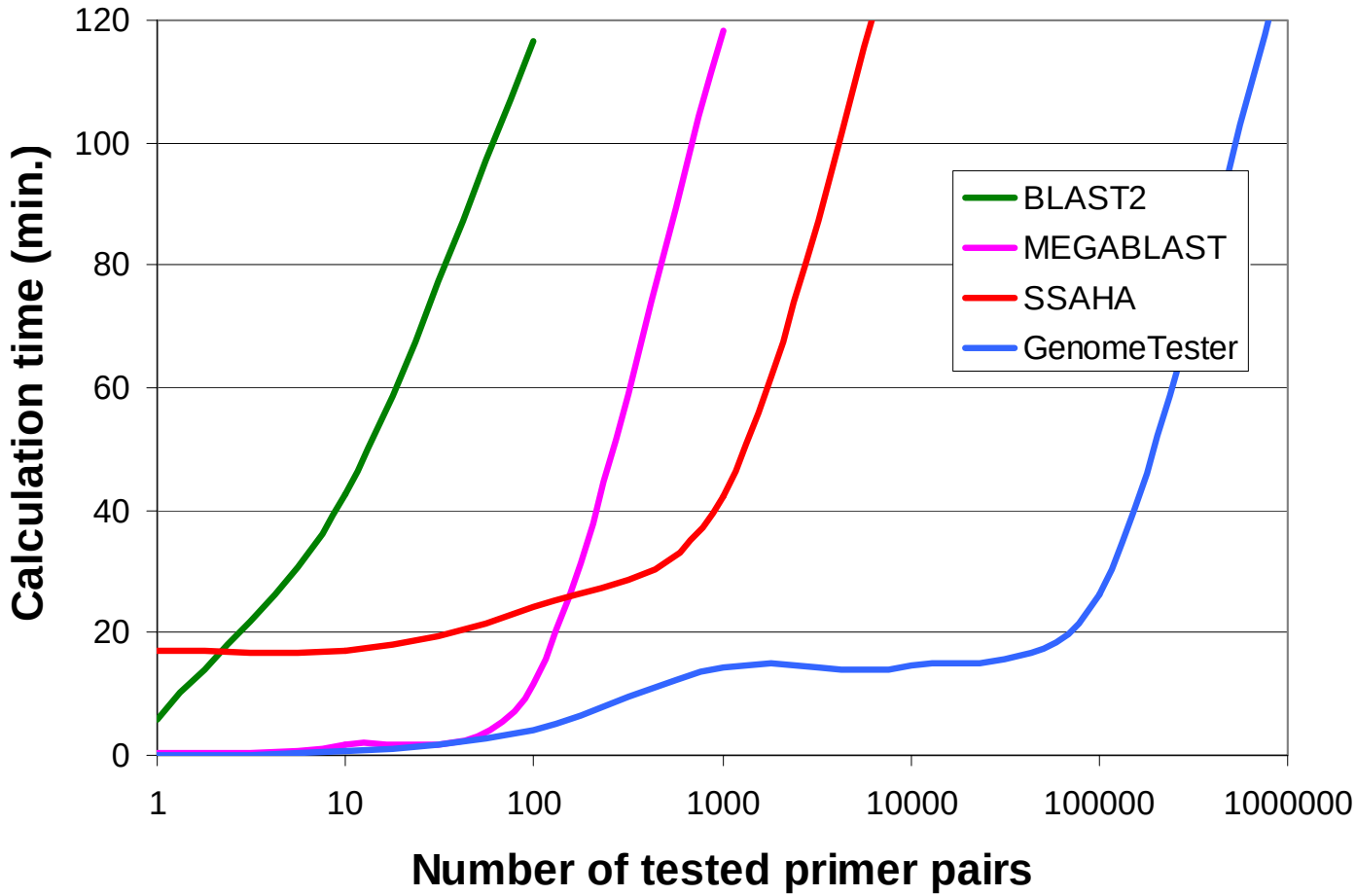


Indeksi koostamine:
sõna pikkus 16
kombinatsioonide arv $4^{16} = 4*10^9$

sõnade koguarv $3*10^9$ ,
4+4 bytes each = 24 GB

| | |
|---|---|
| 16-mer ID | location1 |
| 16-mer ID | location1 |
| 16-mer ID | location2 |
| 16-mer ID | location3 |
| 16-mer ID | location1 |
| 16-mer ID | location2 |

# GenomeTester is signficantly faster for the ‘genome test’ than any other program.




The ‘genome test’ here means finding locations of all primers (16 nt. from the 3’ end) in the human genome and calculation of possible PCR products. Tests were performed on PC-Linux based server, Pentium III, 2 GB RAM, SCSI-RAID0 hard drives. BLAST and MEGABLAST were used without dust filter, word length was 12 for MEGABLAST and 10 for SSAHA.

# PROBLEEM 1
# PCR praimerite genoomitest varieeruva pikkusega sõnade leidmine

Kui modelleerida seostumist füüsilise seondumistugevusega, on vaja teada sõnade asukohta genoomis pikkuste vahemikus 12-28

Lahendus peaks sobima nii 1 praimeripaari kui miljoni praimeripaari analüüsi jaoks.

# PCR praimer multiplex

Selle asemel et teha 10 000 PCRi eraldi võiksime teha 200 PCRi 50-kaupa kokku segatuna ?
Tohutu kokkuhoid DNA ja materjalide osas.

Erinevaid PCRi reaktsioone saab teha samas tuubis kui praimerid ja produktid üksteist ei sega.

Sobivuse selgitamiseks peab kontrollima, et:

- ei teki praimer-praimer seostumisi

- ei teki praimer-produkt seostumisi

- ei teki lisaprodukte genoomist (genoomitest)

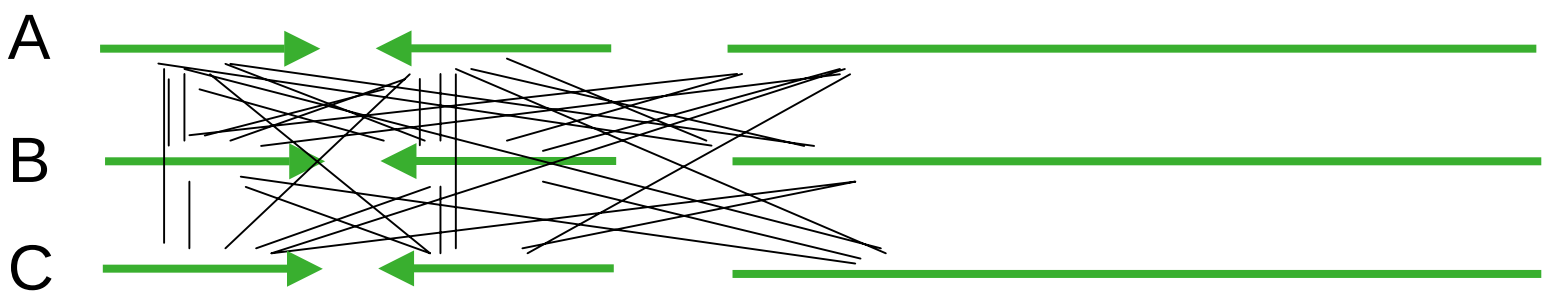
- produktid on sarnase pikkusega

# MultiPLX: tööpõhimõte

1. Testida kõik võimalikud praimeripaarid:

- **PRIMER_SELF_ANY**
- **PRIMER_SELF_END**
- **PRIMER_PRODUCT_END**
- **GENOME_TEST**
- **PRODUCT_LENGTH_DIFFERENCE**

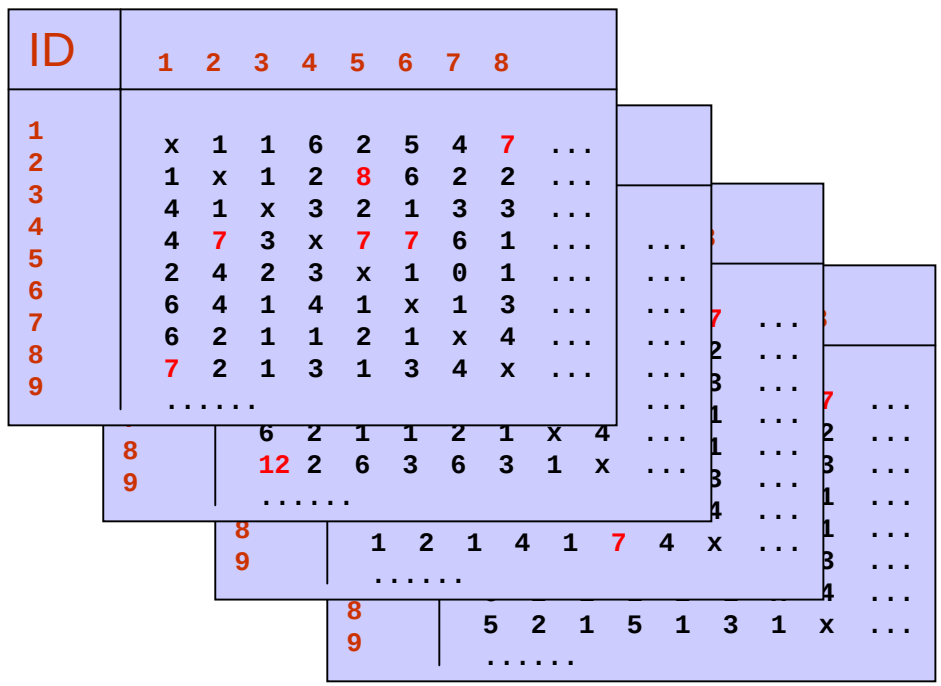
iga parameetri kohta 2*n*(n-1) interaktsiooni



# MultiPLX: tööpõhimõte

2. Koosta iga parameetri jaoks sobivustabelid ja leia antud parameetrile bioloogiliselt relevantne cutoff-väärtus

1. **PRIMER_SELF_ANY**
2. **PRIMER_SELF_END**
3. **PRIMER_PRODUCT_END**
4. **GENOME TEST**
5. **...**



# MultiPLX: tööpõhimõte

## 3. Koosta summaarne sobivustabel:

| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | x | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 4 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 0 |
| 2 | 1 | x | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | x | 3 | 0 | 1 | 0 | 0 | 1 | 1 | 2 | 2 | 4 | 1 | 2 | 1 | 0 |
| 4 | 0 | 0 | 3 | x | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 5 | 2 | 0 | 0 | 0 | x | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | 1 | x | 1 | 3 | 3 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 2 | 0 | 1 | 0 | 1 | x | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 2 | 0 | 0 | 1 | 3 | 0 | x | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 9 | 4 | 0 | 1 | 1 | 0 | 3 | 0 | 0 | x | 0 | 1 | 0 | 0 | 1 | 2 | 0 | 2 |
| 10 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | x | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 11 | 1 | 3 | 2 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | x | 0 | 0 | 1 | 1 | 0 | 0 |
| 12 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | x | 0 | 0 | 0 | 1 | 3 |
| 13 | 0 | 0 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | x | 0 | 0 | 0 | 0 |
| 14 | 4 | 0 | 1 | 1 | 4 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | x | 1 | 0 | 0 |
| 15 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 1 | x | 0 | 0 |
| 16 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | x | 1 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | x |

# MultiPLX: tööpõhimõte

## 4. Leia omavahel sobivate praimerite grupid, nii et saavutatakse minimaalne gruppide arv :

| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | x | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 4 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 0 |
| 2 | 1 | x | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | x | 3 | 0 | 1 | 0 | 0 | 1 | 1 | 2 | 2 | 4 | 1 | 2 | 1 | 0 |
| 4 | 0 | 0 | 3 | x | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 5 | 2 | 0 | 0 | 0 | x | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | 1 | x | 1 | 3 | 3 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 2 | 0 | 1 | 0 | 1 | x | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 2 | 0 | 0 | 1 | 3 | 0 | x | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 9 | 4 | 0 | 1 | 1 | 0 | 3 | 0 | 0 | x | 0 | 1 | 0 | 0 | 1 | 2 | 0 | 2 |
| 10 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | x | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 11 | 1 | 3 | 2 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | x | 0 | 0 | 1 | 1 | 0 | 0 |
| 12 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | x | 0 | 0 | 0 | 1 | 3 |
| 13 | 0 | 0 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | x | 0 | 0 | 0 | 0 |
| 14 | 4 | 0 | 1 | 1 | 4 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | x | 1 | 0 | 0 |
| 15 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 1 | x | 0 | 0 |
| 16 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | x | 1 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | x |

# MultiPLX: tööpõhimõte

## 5. Prindi tulemused:

**Group 1:** 1 4 6 17 24 26 31 79 96 103 168 170 177 270 277 287 304
**Group 2:** 2 5 9 12 13 32 36 64 128 341
**Group 3:** 3 7 22 28 29 35 44 47 49 87 102 106 157 201 327
**Group 4:** 8 10 16 19 23 33 71 110 123 165 262
**Group 5:** 11 21 34 37 62 82 86 108 122 147 317 362
**Group 6:** 14 18 27 40 63 81 118 135 136 145 179 197 219 260 321
**Group 7:** 15 20 41 51 57 68 78 95 107 117 190 318
**Group 8:** 25 30 38 45 46 76 127 142 166 195
**Group 9:** 39 48 58 65 85 92 100 140 164 185 236
**Group 10:** 42 50 52 60 67 75 114 115 126 132 137 193 213
**Group 11:** 43 54 59 66 72 88 94 151 162 212 249 283
**Group 12:** 53 55 73 77 80 83 91 237 252 357
**Group 13:** 56 61 84 90 104 146 180 231 309 326
**Group 14:** 69 70 93 98 101 111 112 149 191 215 272 298 307
**Group 15:** 74 97 105 120 125 138 143 202 259 308 383
**Group 16:** 89 99 113 121 160 183 200 206 247 273
**Group 17:** 109 116 119 144 150 187 188 189 221 319
**Group 18:** 124 129 131 133 139 192 255 256 279 334 363 392
**Group 19:** 130 141 148 153 154 163 174 330 353 384 397
**Group 20:** 134 152 159 173 205 210 243 305 329 345
**Group 21:** 155 156 169 208 211 225 322 325 388
**Group 22:** 158 161 167 175 204 242 246 291 371
**Group 23:** 171 172 196 223 244 289 306 378 379
**Group 24:** 176 182 194 199 203 257 281 311 372
**Group 25:** 178 184 207 214 234 258 350 389
...

# PROBLEEM 2
# PCRi praimerite multiplex

Kas on v6imalik leida optimaalne lahendus ?

Milline on lahenduse ajaline ja ruumiline kompleksus ?

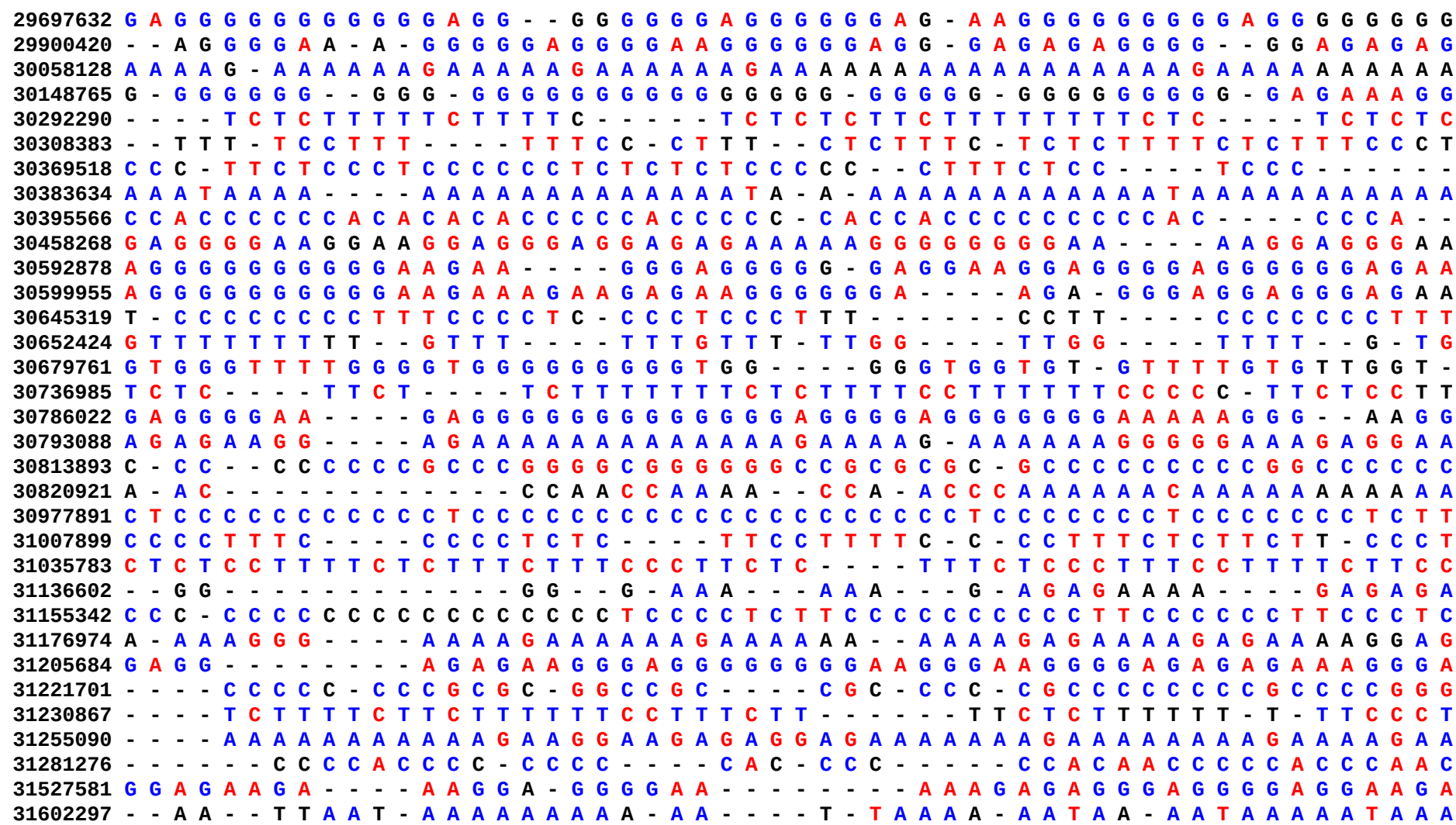
NB! Grupeerimisel v6ivad olla lisatingimused:

- Gruppide min. max. suurus (liikmete arv) piiratud
- Suurima ja väikseima grupi liikmete arvu vahe ei tohi olla suurem kui n

# 3. teema
# Genotüpiseerimine
# Efektiivne markerite valik

Marker

55 kromosoomi

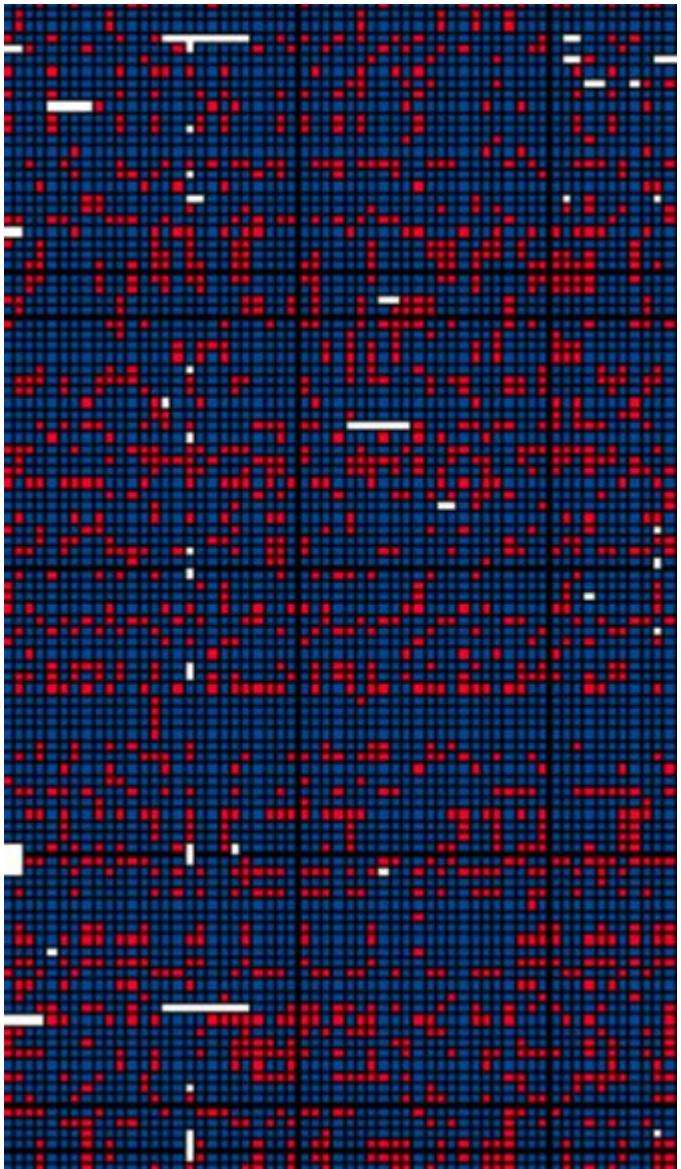


# Haplotüüpide blokid, kromosoom 22

59 haplotypes from CEPH founders

**Dawson et al. (2002) Nature 418: 544**

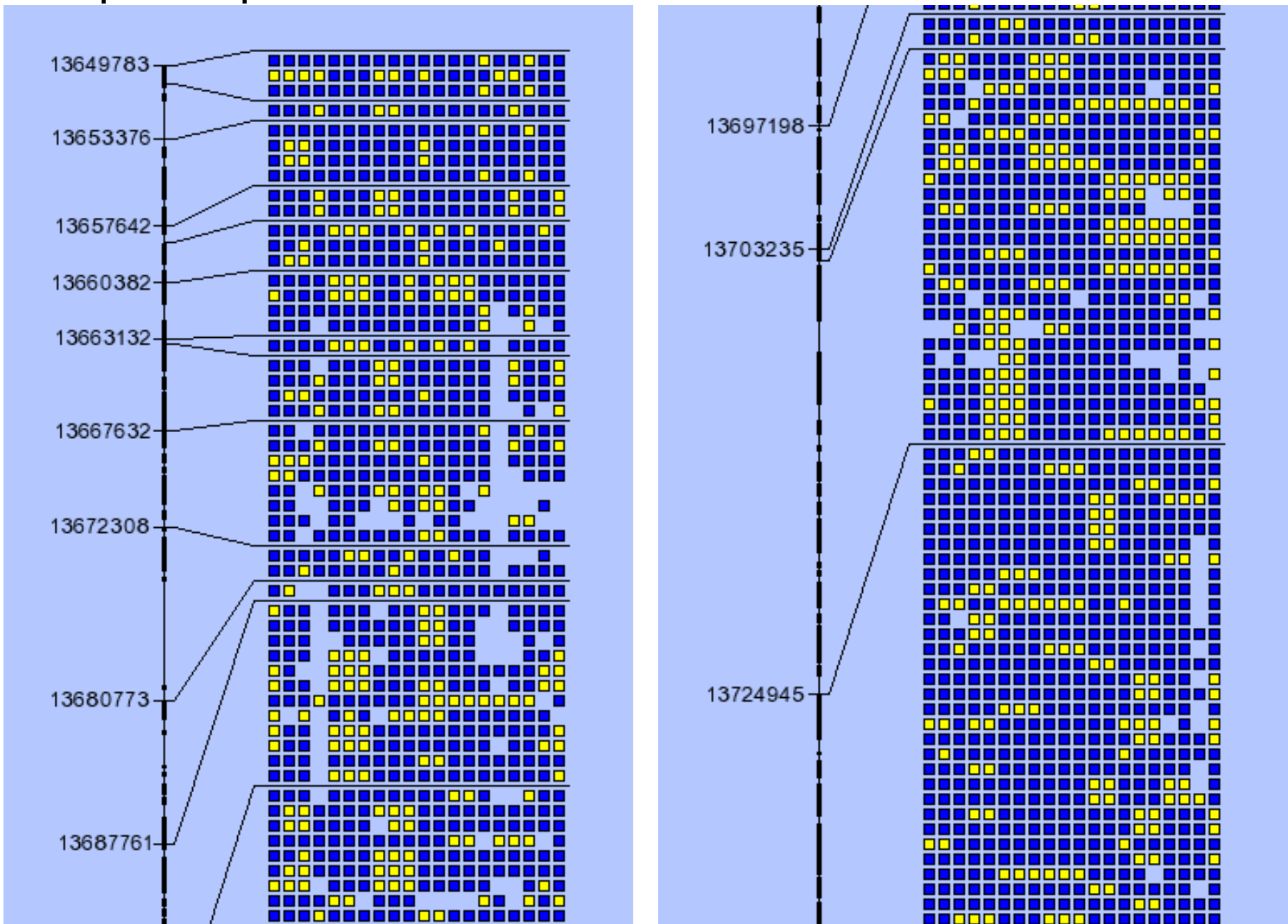
1504 SNPs



Eesmärk:
Vähendada markerite (SNPde) arvu, nii et säiliks X% informatsioonist.
Ridade järjekord peab säiluma.

Selleks jagatakse read gruppideks (haplotyybi blokid)

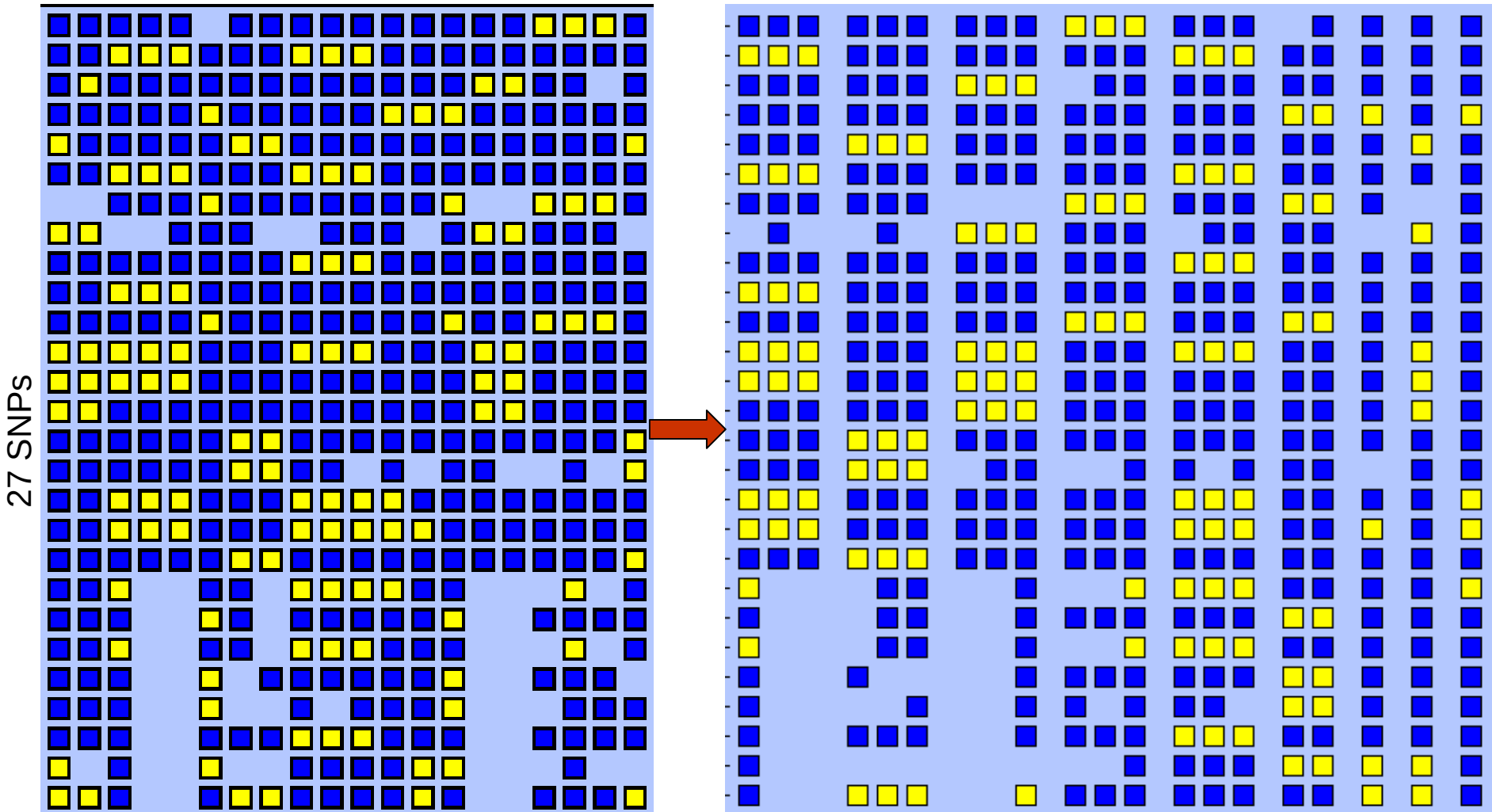
# Haplotüüpide blokid



# Haplotüüpide blokid

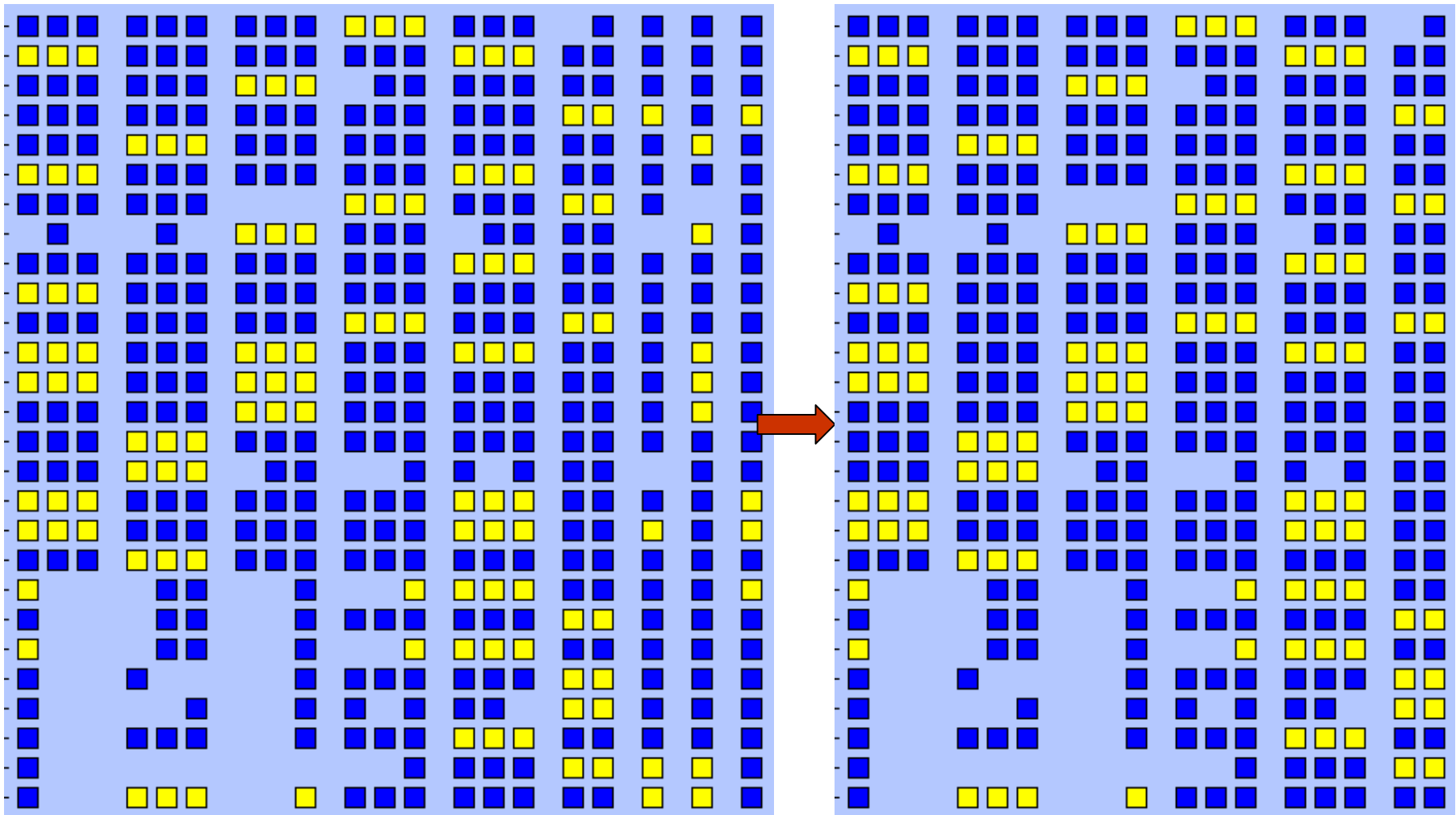
**Patil et al. 2001 Science 294:1719**

Haplotype data from human chromosome 21

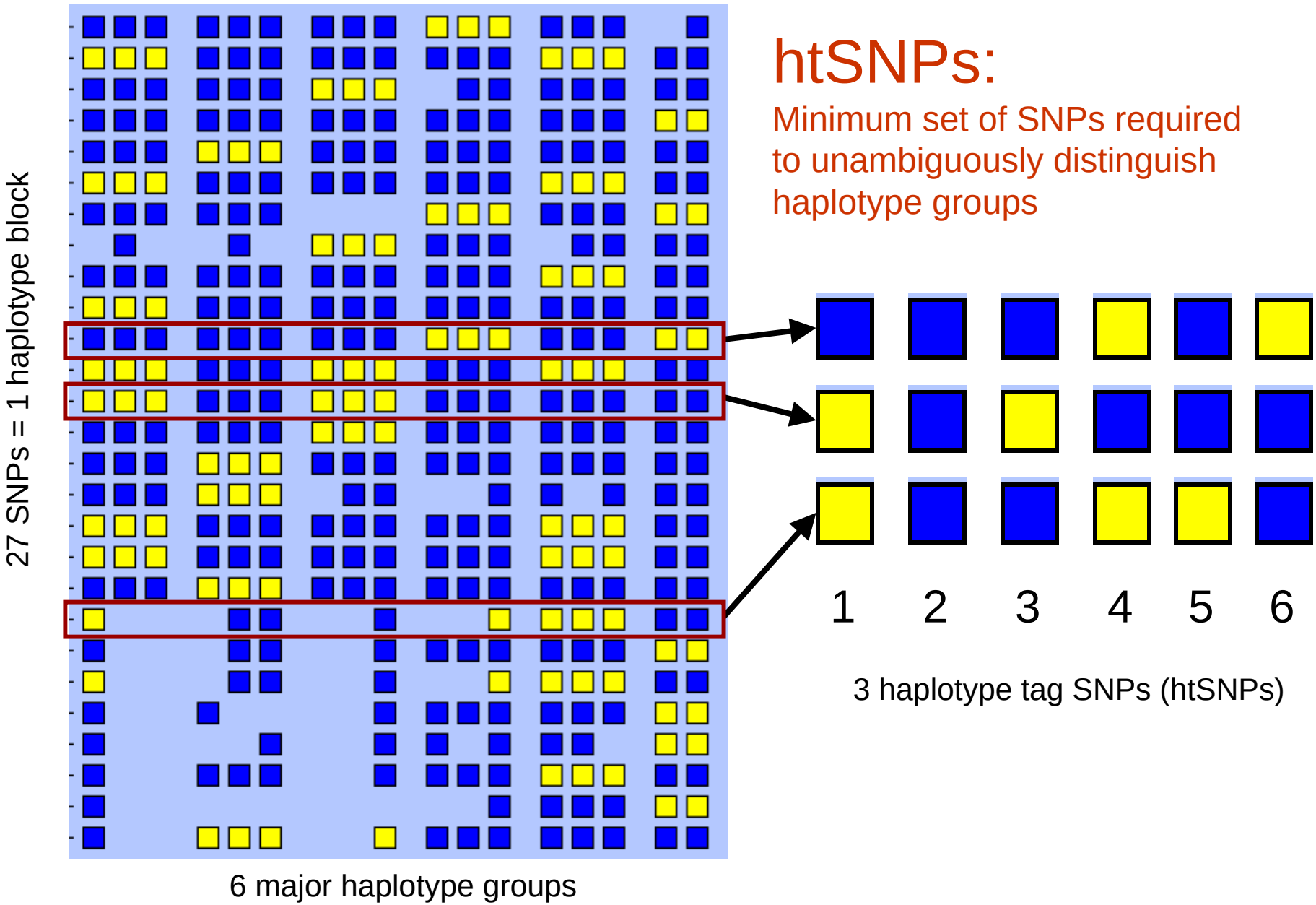


# Haplotüüpide blokid

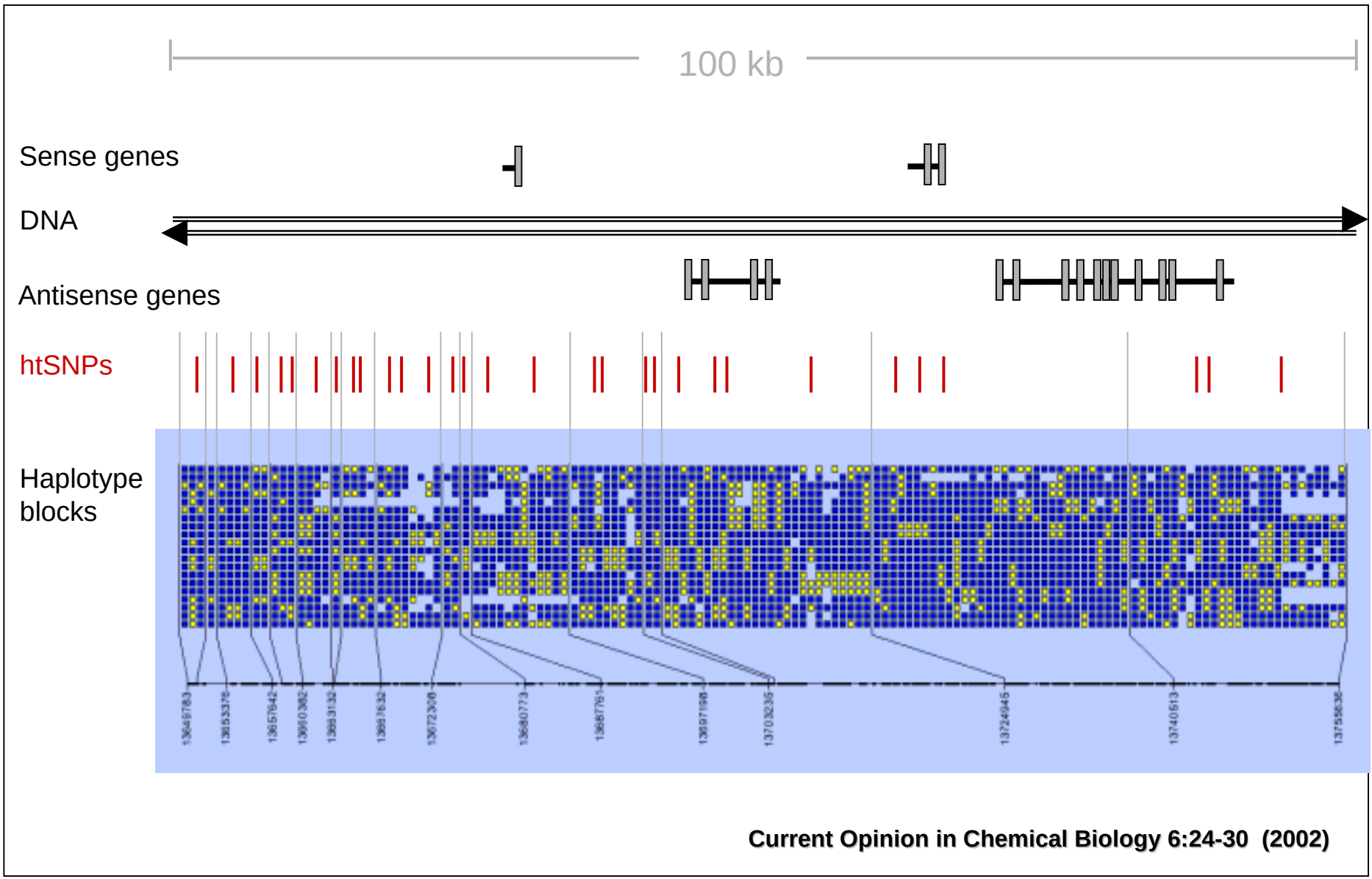
**Patil et al. 2001 Science 294:1719**



# Haplotüüpide blokid

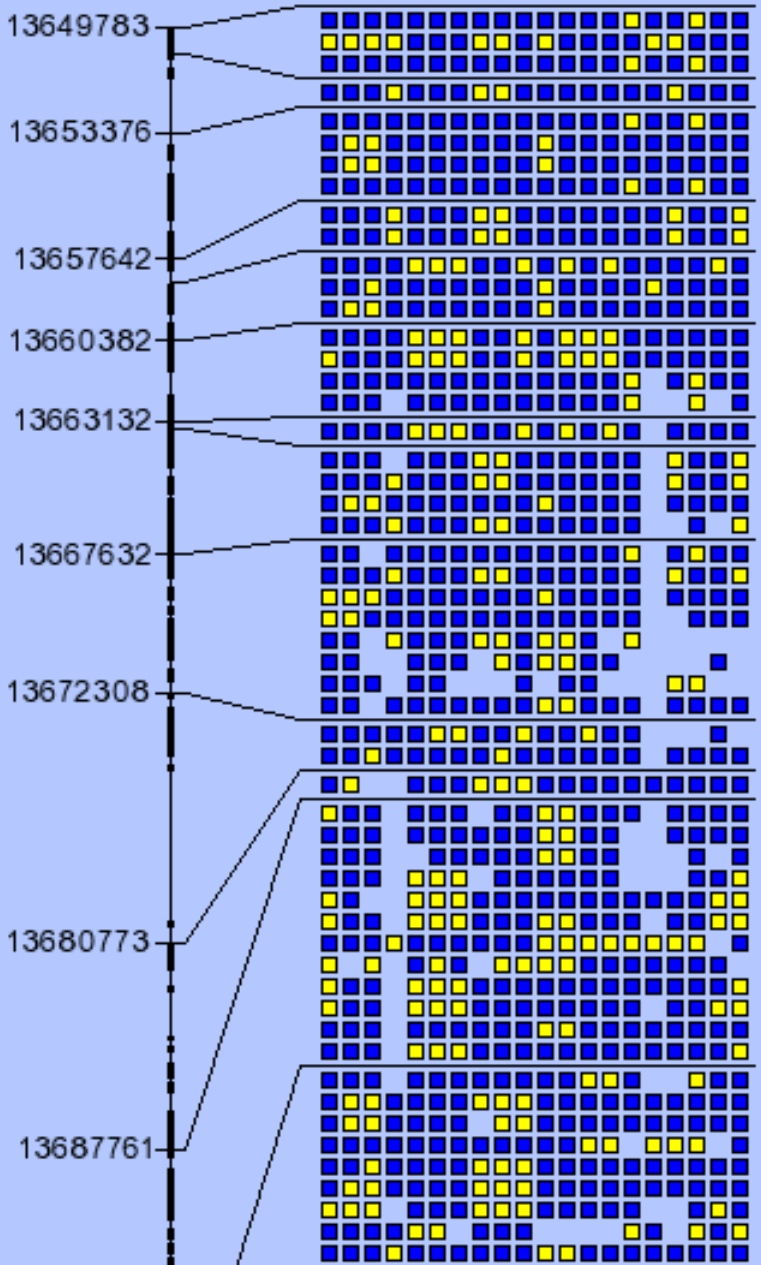


# Eksonitel baseeruvad assotsiatsiooniuuringud




**Current Opinion in Chemical Biology 6:24-30 (2002)**

# Haplotüüpide blokid




Mõned artiklid, mis kirjeldavad erinevaid blokkide leidmise algoritme:

Daly et al. 2001 Nat. Genet. 29: 229

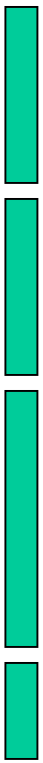
Patil et al. 2001 Science 294: 1719

Zhang et al. 2002 PNAS 99:7335

Gabriel et al. 2002 Science 296: 2225

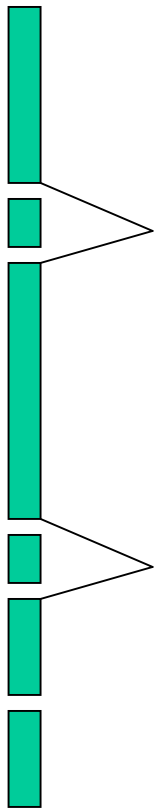
Dawson et al. 2002 Nature 418: 544

# PROBLEEM 3
## Kas mittepidevate blokkidega on võimalik sama infot efektiivsemalt kirjeldada?



Seni on kasutatud dünaamilist programmeerimist et leida optimaalseimad bloki piirid

3952 htSNPs esindavad 24046 SNPd



?