Suuremahulised andmed DNA sekveneerimisest

ESS 24. konverents "Uued suundumused statistikas"

Maido Remm

Tartu Ülikool
Eesti Biokeskus
Genoomika tippkeskus

28. september 2012

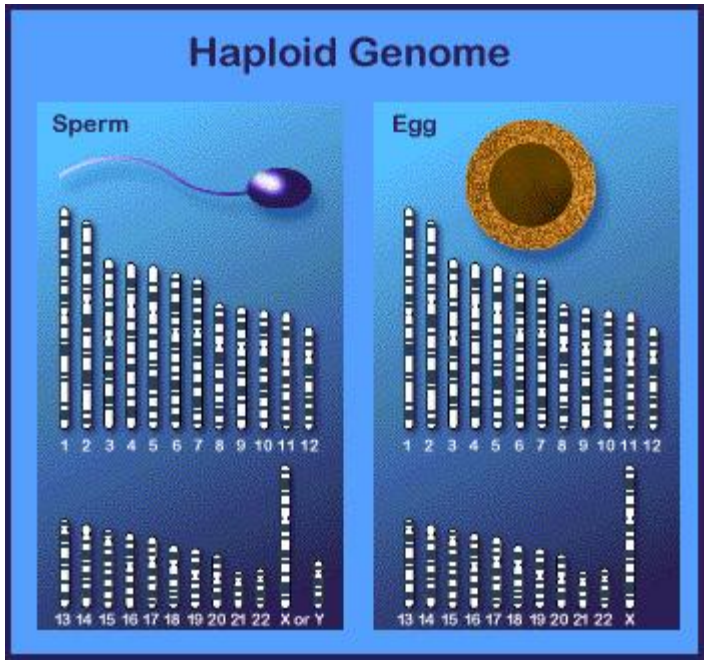
Mis on genoom?

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

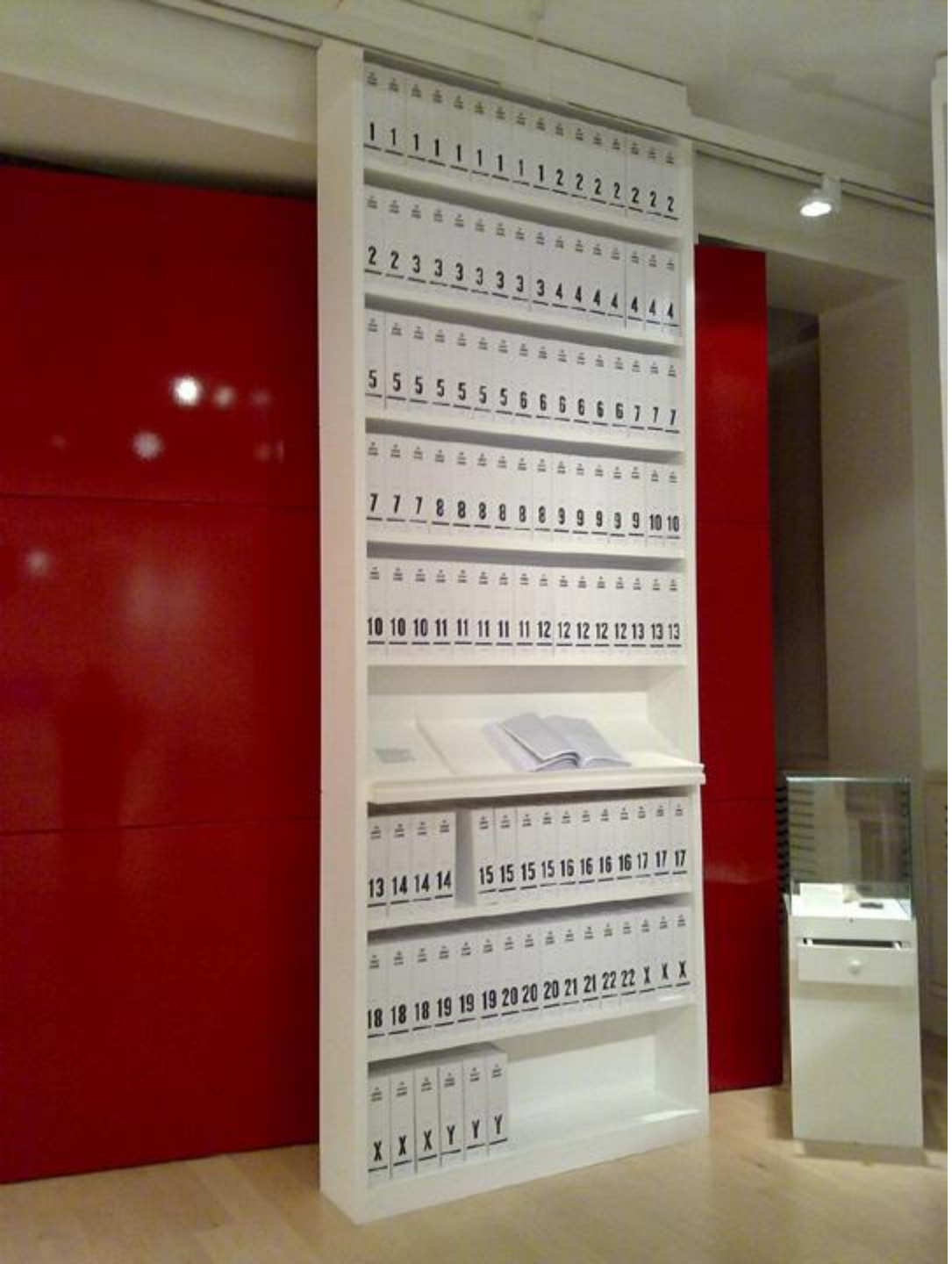
Järjestus tähendab siin nukleotiidset järjestust, kus esineb 4 tähte: A, C, G, T

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).

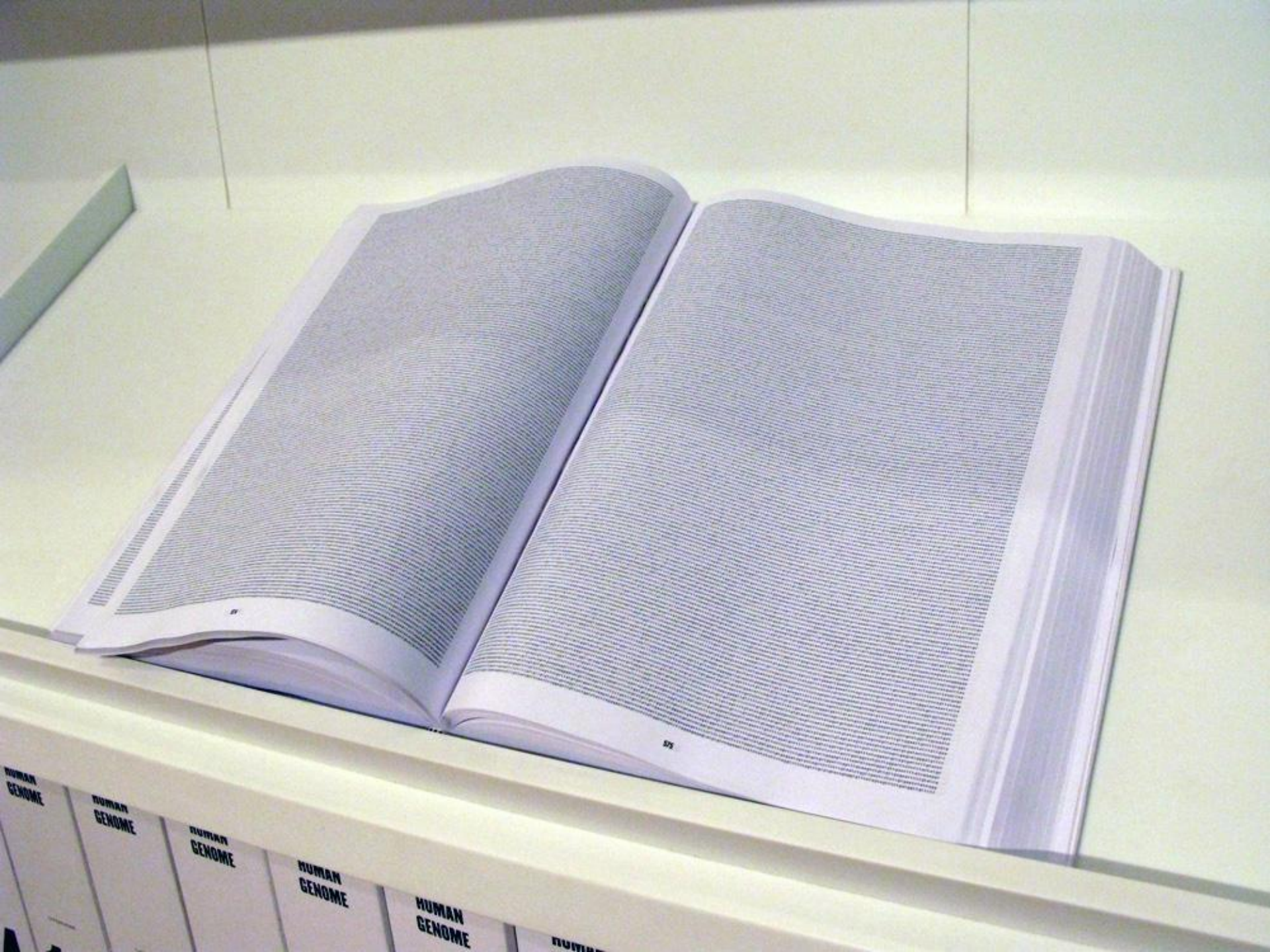



Genoomi järjestuse näidis
Inimese genoom, jätkub 1 miljonil sarnasel leheküljel ...



Genoomi järjestuse näidis

paberformaadis


HUMAN GENOME
HUMAN GENOME
HUMAN GENOME
HUMAN GENOME
HUMAN GENOME

Genoomide uurimiseks tuleb genoomide järjestusi määrata (genoome *sekveneerida*)

Sekveneerimise tehnoloogiad:

I generatsiooni tehnoloogiad (Sangeri tehnoloogia)
Selle tehnoloogiaga on määratud inimese, hiire, roti ja veel ca neljakümne imetaja genoomide järjestus.
Hind ca 1000 nukleotiidi / $

II põlvkonna tehnoloogiad (Illumina/Solexa, Roche/454, IonTorrent)
Hind ca 50 000 000 nukleotiidi / $

Illumina HiSeq 2000

Järjestuse andmed Illumina HiSeq2000 masinast:

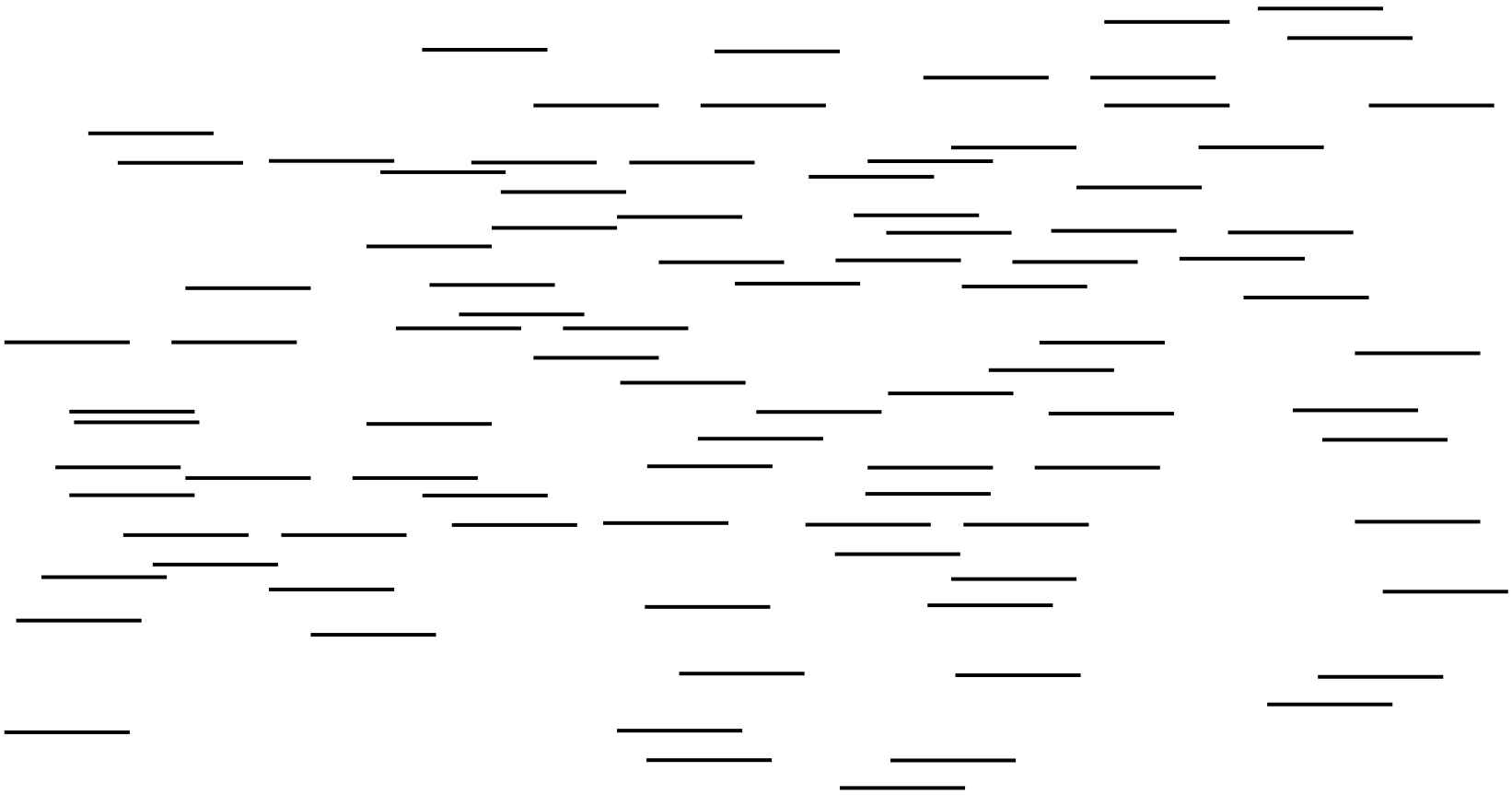
- Üks analüüs kestab 2 nädalat
  (2 päeva laboris ettevalmistust + 12 päeva masina töö)

- Masinast tuleb välja kuni 6 miljardit
  100 nukleotiidi pikkust ***lugemit*** (ik. *raw read*).
  Kokku 600 000 000 000 nukleotiidi koos iga nukleotiidi kvaliteedi (usaldusväärsuse) hinnanguga.

- Sellest mahust piisab, et analüüsida 2 inimese täisgenoomi või 192 erinevat bakteritüve.

Mis nende andmetega teha saab?

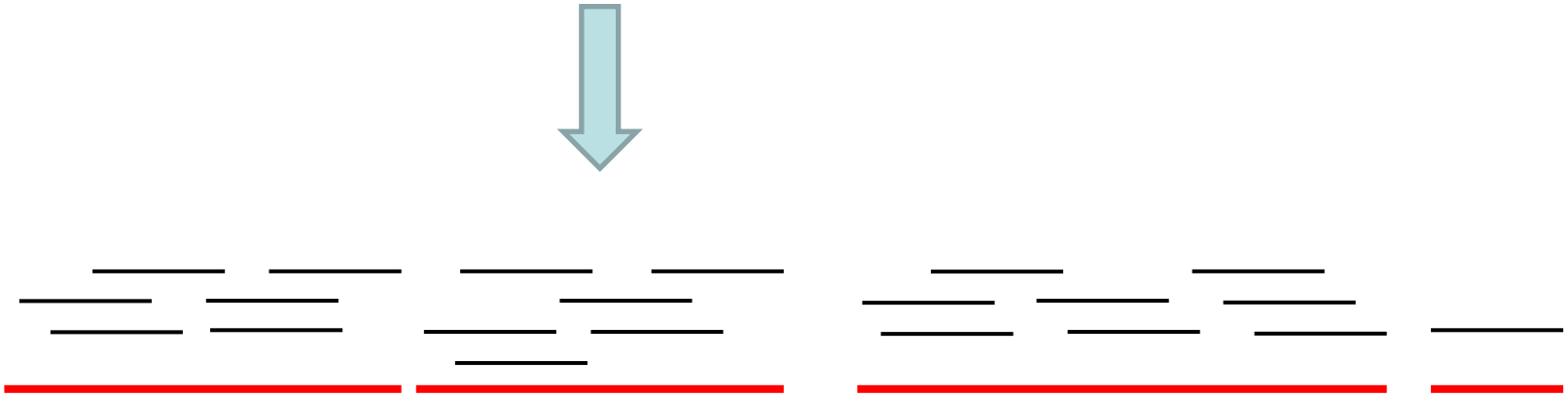
6 miljardit järjestuse juppi:



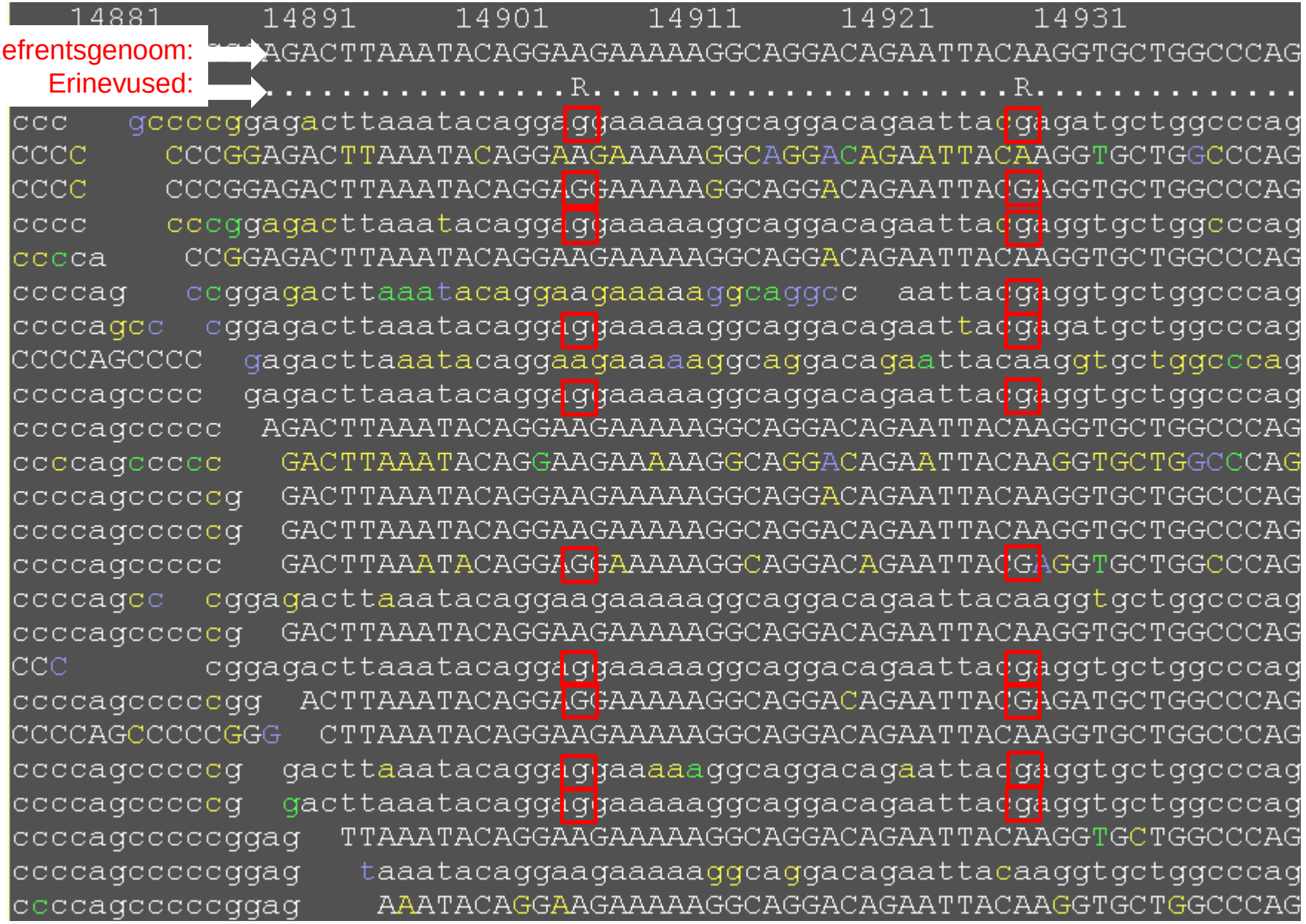
Mis nende andmetega teha saab?

1. Varieeruvuse uurimine teadaoleva genoomiga liikides
(mapping of reads and variant calling)

2. Genoomide järjestuse määramine seni järjestamata genoomiga liikidele
(*de novo* genome assembly)



Lugemite paigutamine (mapping)



Lugemite paigutamine (mapping)

Tulemus ei ole alati nii ilus kui loodetud:

- Sekveneeritud meeste Y-kromosoom on paljudes kohtades heterosügootne (näiteks: peaks olema A aga arvuti pakub AG)

- Perekondade sekveneerimisel nähakse palju mittemendeliaalselt päranduvaid positsioone:

ema **AA** isa **TT**

laps **AA**

Leidsime 5000 sellist positsiooni valke kodeerivates piirkondades) !

Lugemite paigutamine genoomile

Lugemite genoomile paigutamist segavad asjaolud:

- Masinast tulevates 100 bp lugemite järjestustes esineb vigu (1-2%)

- Loomulik varieeruvus inimeste vahel
  Võimalikud on nii asendused kui deletsioonid/insertsioonid; seetõttu ei saa paigutamisel nõuda 100%-list kokkulangemist teadaoleva referentsgenoomiga.

- Genoomis on palju korduva järjestusega piirkondi, ca 45% kogupikkusest

Uurisime probleeme kontrollitud keskkonnas:

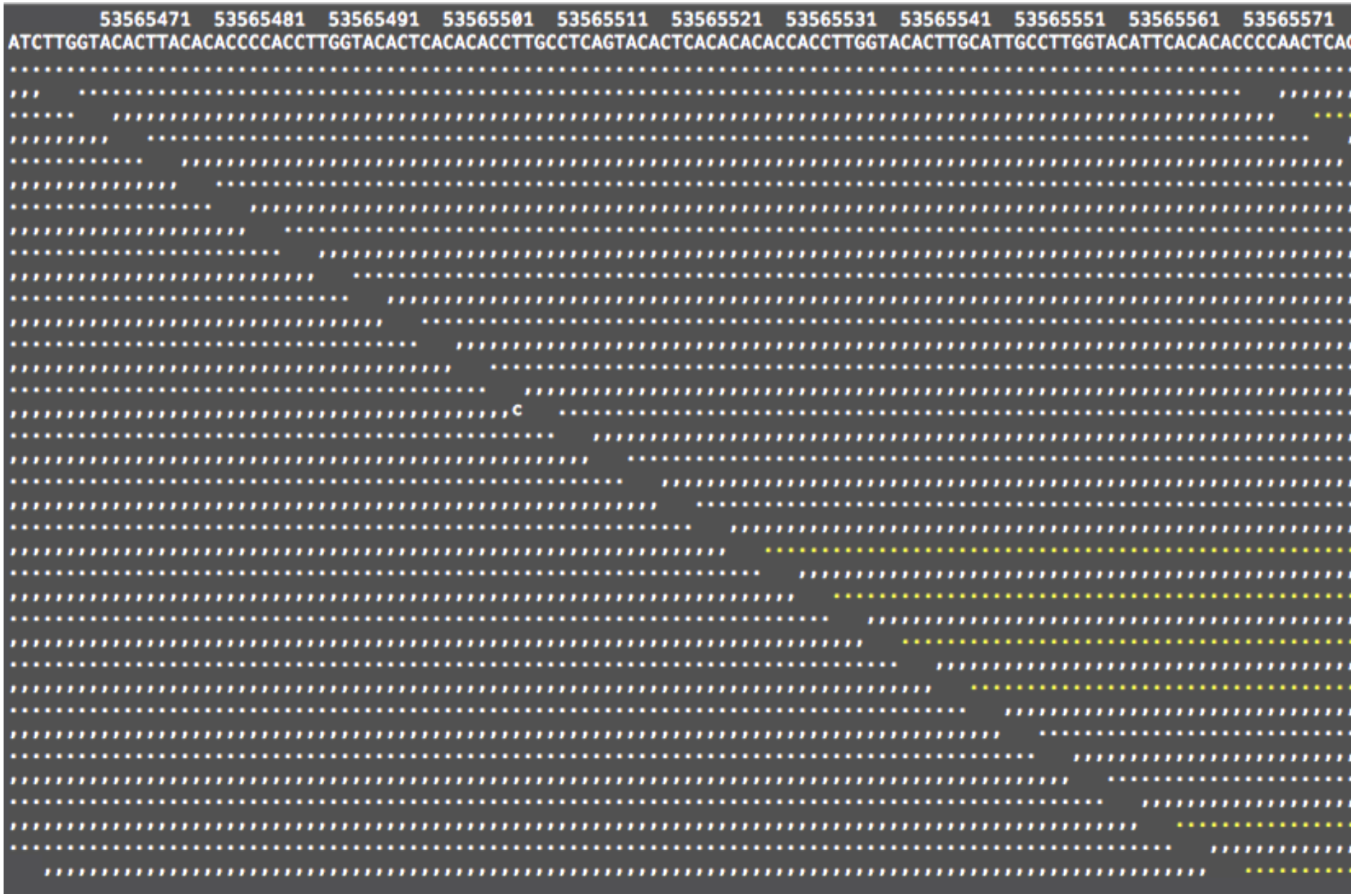
- Tegime genoomist arvutiga kunstlikke 100 bp pikkusi lugemeid, viisime sisse tegelikult esinevaid mutatsioone ja üritasime need lugemid genoomile tagasi paigutada.

- Sünteetiliselt tekitatud lugemite genoomile tagasipaigutamisel ei leia 2-8% lugemitest enam oma õiget kohta.

- Valede asukohtade tõttu on oht, et leitakse olematuid variatsioone.

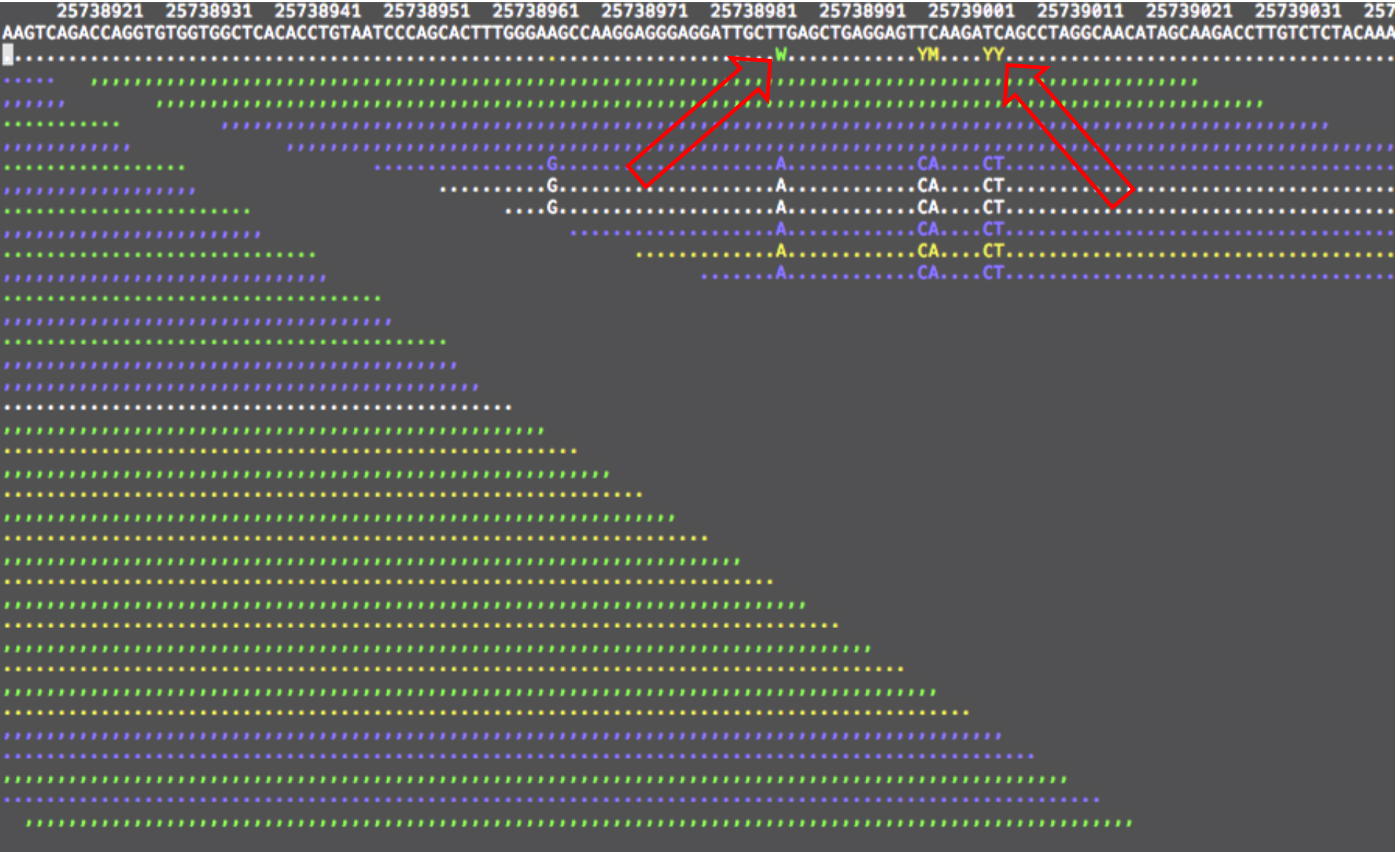
Ohustatud ca 150 000 kohta genoomis ja potentsiaalselt on sellest probleemist puudutatud kuni 7500 geeni (kolmandik kõigist geenidest). Ka mõned heades ajakirjades raporteeritud haigusriski suurendavad mutatsioonid on sellistes piirkondades.

Selline peaks olema tulemus:



Paljud lugemid on paigutatud valesse kohta:

Paljud lugemid on paigutatud valesse kohta ja valedeks genotüüpideks määratud:



Kuidas vale-paigutuse probleeme lahendada?

A) Tõenäoslik lugemite paigutamine

Lugemite paigutamisel leida igale lugemile TÕENÄOSUS tema paiknemiseks just selles kohas.

Seejuures on mitmeid parameetreid mida vaja eelnevalt hinnata:

- alternatiivsete paiknemiste arvu
- mutatsiooni tekke tõenäosust igas positsioonis
- sekveneerimisvea tõenäosust
- jpms.

Paigutada genoomile ainult sellised lugemid, millel on selgelt üks suure tõenäosusega asukoht genoomis, ülejäänud eemaldada andmestikust.

Tulemus: Seda lähenemist on erinevate töögruppide poolt proovitud, kuid praktikas pole laialt levinud.

Kuidas vale-paigutuse probleeme lahendada?

B) Katseline probleemsete piirkondade leidmine ja eemaldamine.

1. Leida regioonid genoomis, mis antud metoodika korral valesti paigutatakse
2. Koostada valesti paigutuvate piirkondade must nimekiri (*mask*)
3. Eemaldada need regioonid andmestikust

Tulemus: Tänu *mask*i kasutamisele ja mõnede muude lugemite paigutamist mõjutavate parameetrite muutmisele saame pärandumisvigade arvu vähendada kuni 10x (geenide sees 5000 pealt 500-ni).

Järeldus:

=> Pimesi ei saa ühtegi II põlvkonna sekveneerimise tulemust usaldada

=> Ülejäänud vigade tekkepõhjus vajab täiendavat uurimist

Tänuavaldused

Ulvi Gerst Talas

Mikk Eelmets

ning ...

Reidar Andreson
Tarmo Puurand
Andres Veidenberg
Reedik Mägi

jpt. praegused ja endised
TÜMRI bioinformaatika
töögrupi liikmed