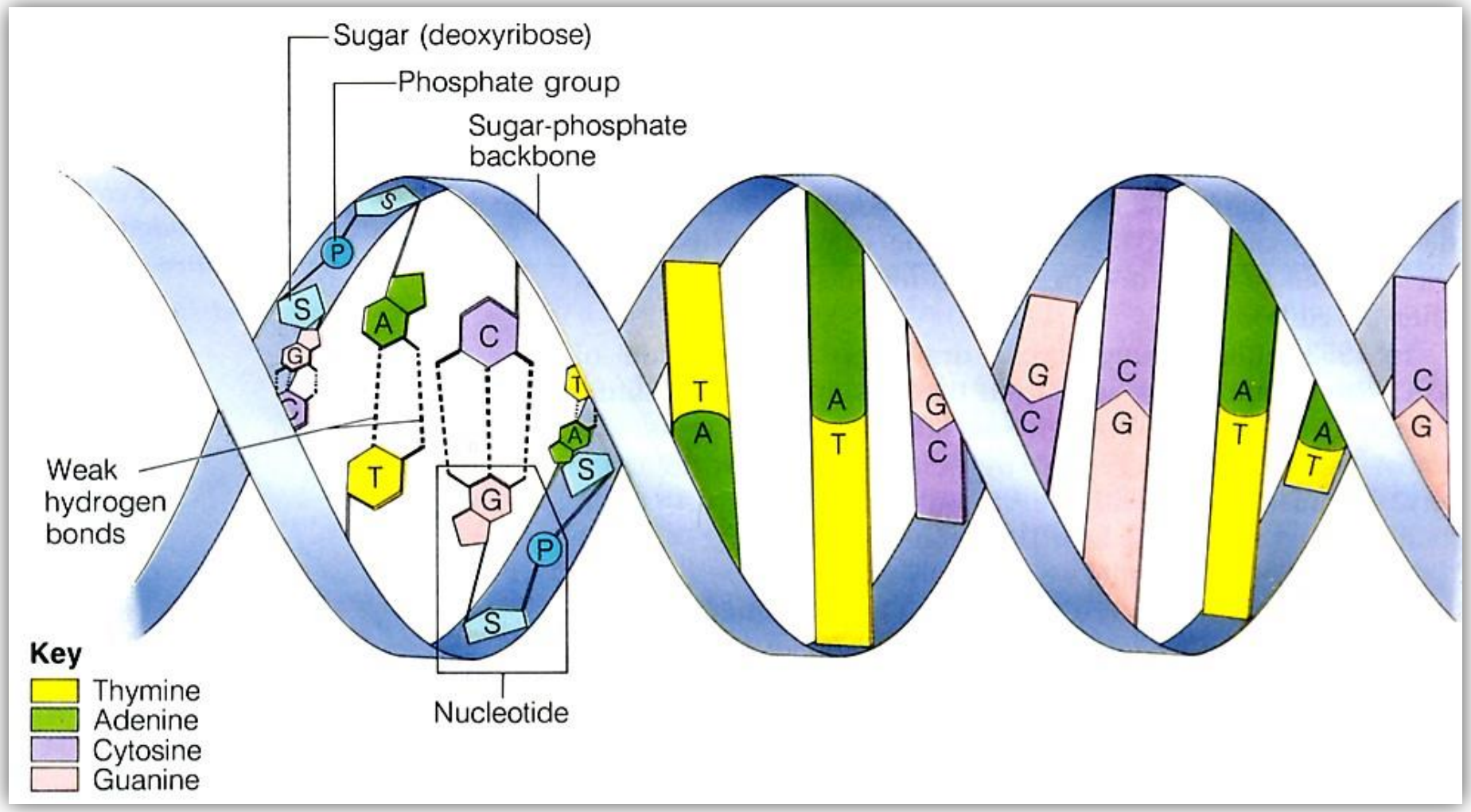
ALGORITMISKĀS METODES GENOMA VIRKŅU ANALĪZEI

Andreass Šulcs, 12.03.2014.

Kas, kāpēc, cik daudz?

- Genoms – gara, 4 dažādu nukleotīdu (A-T, C-G) virkne
- Satur **visu** informāciju par organisma uzbūvi
- Sadalīts vairākās hromosomās (cilvēkam – 2 x 23)
- Cilvēka DNS garums ~3 000 000 000 bp
- Nekompresētā veidā ~3GB, kompresētā ~800MB

Kas, kāpēc, cik daudz?



Att. iegūts: http://karimedalla.wordpress.com/2012/11/01/3-3-7-1-dna-structure/

Genoma datu iegūšana

- Pirmo reizi pilns genoms iegūts: Human Genome Project (1990. – 2003.), izmaksas – $ 2 700 000 000
- Šobrīd – iegūšanas laiks dažas dienas, izmaksas zem $ 1 000 (pie lieliem apjomiem)
- Pēc dažiem gadiem, iespējams, $ 100

- Plašākai auditorijai: https://www.23andme.com/, tikai vispārīgi dati, izmaksas $ 99

Tehnoloģijas

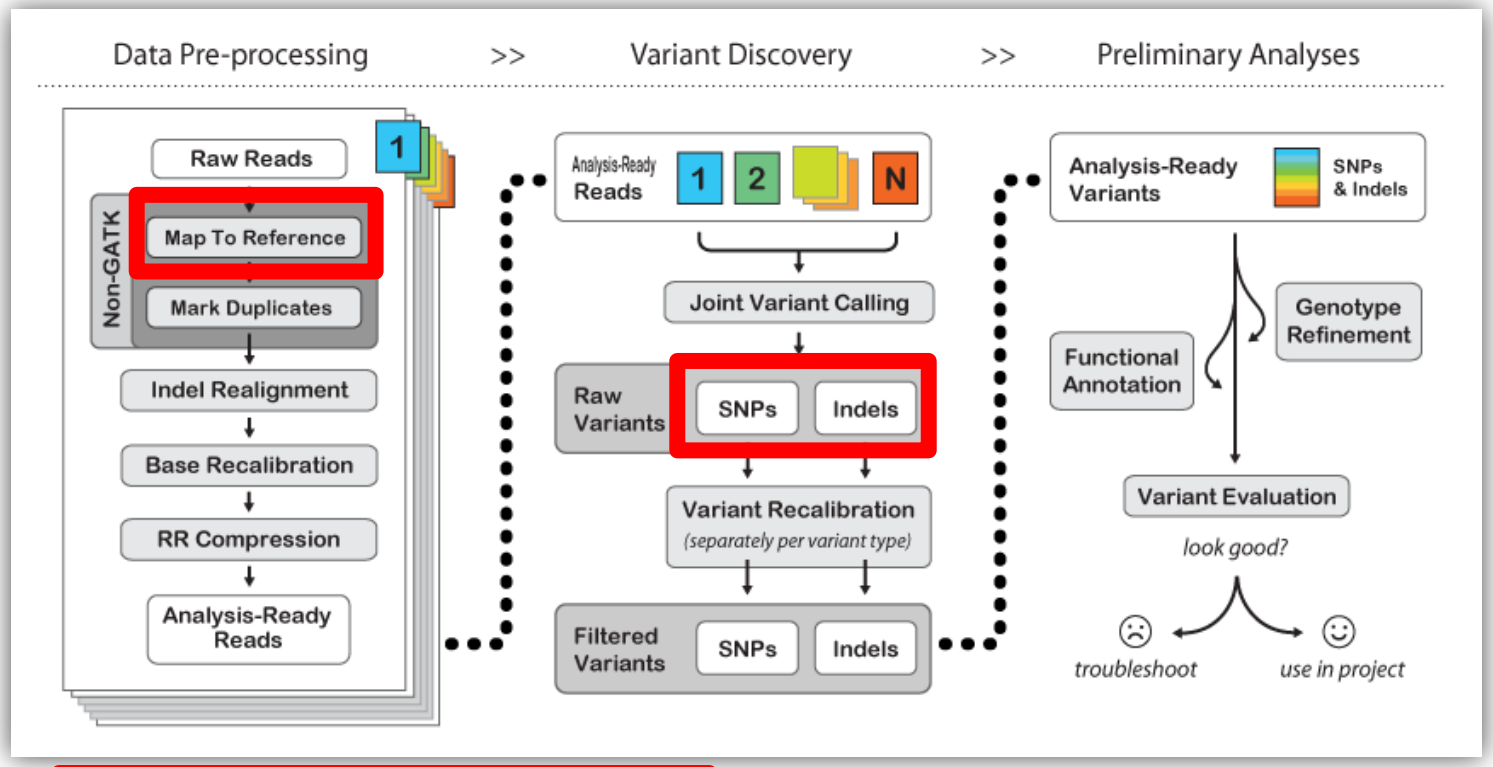
Method	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)
Read length	50 to 300 bp	50+35 or 50+50 bp
Accuracy	98%	99.9%
Reads per run	up to 3 billion	1.2 to 1.4 billion
Time per run	1 to 10 days, depending upon sequencer and specified read length[46]	1 to 2 weeks

Dati no: http://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods

Motivācija

- Absolūti primārais - cilvēka DNS
- Atšķirība cilvēku DNS ir ar kārtu 0.1%
- Pielietojums genoma analīzei – iedzimtu slimību izpausmes prognozēšana (gan individuāli, gan cilveķu kopai) kā arī gēnu terapijā.
- Ņemot vērā sekvencēšanas izmaksu samazināšanos, jāspēj pārskatāmā laikā apstrādāt datus un sniegt pacientam ieteikumus, etc.
- Šobrīd visaktīvāk tiek pētīti vēža pacientu DNS, kodējošos reģionos (1.5% no kopējā cilv. DNS)

Tipiskā darba plūsma DNS datu apstrādē

- Genoma de-novo salikšana +
- 

- + haplotipu atšifrēšana

Att. iegūts: http://www.broadinstitute.org/gatk/guide/best-practices

De-novo salikšana

- Daudz miljonu nolasījumu (atkarībā no tehnoloģijas, 30-50 bp), kuri jāsavieno vienā garā virknē
- Salikšana bez ‘references’ genoma
- Metodes – garākiem lasījumiem OLC (overlap/layout/consensus) algoritmi (parasti ‘greedy’), īsākiem uz de Bruijn grafu balstītas metodes
- Nav vairs tik aktuāli, jo cilvēka genoms sekvencēts un salikts ļoti daudz reižu.

Meklēšana pēc ‘references’

- Meklē jau saliktā genomā
-
```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGAC
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCC
GTTGCAAAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGC.......
```

- Eksperiments izdod daudz milj. nolasījumus garumā ~50-150 bp
-
```
@r1
CCAGCCGGACTTCAGGCCTGCCATCCAGTTCCCGCGAAGCTGGTCTTCAGCCGCCCAGGTCA
+
7%&%E,3@->*/4>0):4)17/4A=H<3!72"$97HDB(*66B(,&$+41&A,8+=,7/216
@r2
AAGCAGTAAGGGGCATACCCCGCGCGAAGCGAAGGACAACCTGAAGTCCACGCAGTTGCTGA
+
60./".G76C,0220H8@4B)"+&F2,0>$+%%%.&@><*%.=!!%G&$GB"+!*G)5<!;&
@r3
```

- Varbūtības kodē ar 96 ASCII simboliem

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

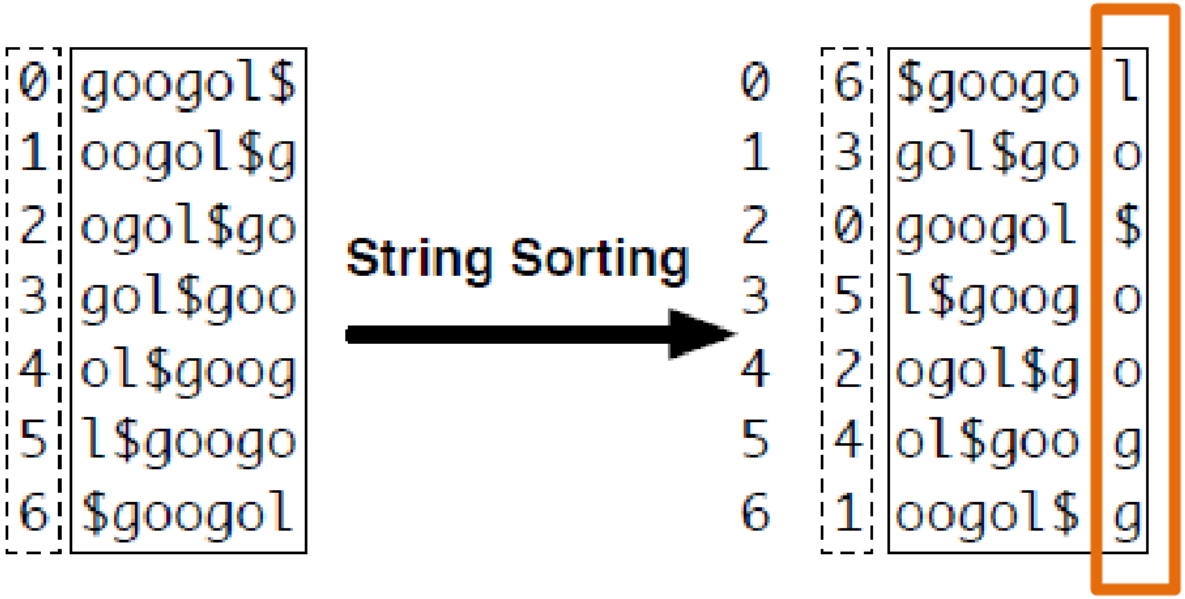
- Tipiski 1-2% kļūdaini nolasījumi

Meklēšana pēc ‘references’

- Nepieciešams konkrēta gadījuma analīzei (rase, tautība, ķermeņa uzbūve, slimības uc.)
- Nepieciešamība pēc algoritmiem/programmatūras parastai darba stacijai. Šobrīd tāda ir – Bowtie2, BWA, GMAP
- Visi izmanto Burrows-Wheeler transformāciju un tās pielāgotu indeksāciju, kas nodrošina mazu atmiņas nepieciešamību meklējot, un iespēju saspiest uzglabājot

Meklēšana pēc ‘references’ - BWT

1. Original text = “googol”
2. Append ‘$’ to mark the end = X = “googol$”
3. Sort all rotations of the text in lexicographic order
4. Take the last column.



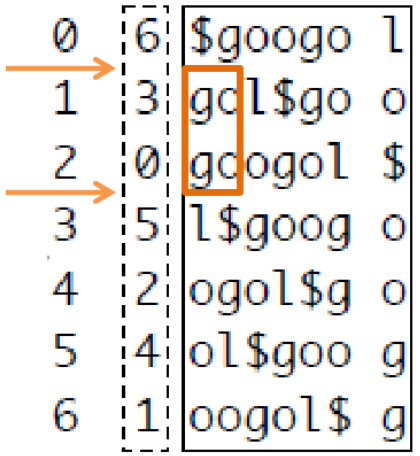
Att. iegūts: http://bozeman.genome.washington.edu/compbio/mbt599_2011/Genome540_BWA_15Feb2011.pdf

Meklēšana pēc ‘references’ - BWT

- No transformācijas viennozīmīgi var atgūt orģinālteksta
- Meklēšana notiek kāpjoties atpakaļ
- Precīzai meklēšanai sarežģ. O(|W|)
- Neprecīzai meklēšanai izmanto ‘backtracking’, sarežģītība parametriska O(|W|)*f(|hg|, d)

- All occurrences of substrings with a common suffix, W appear next to each other

W = “go”



X = googol$

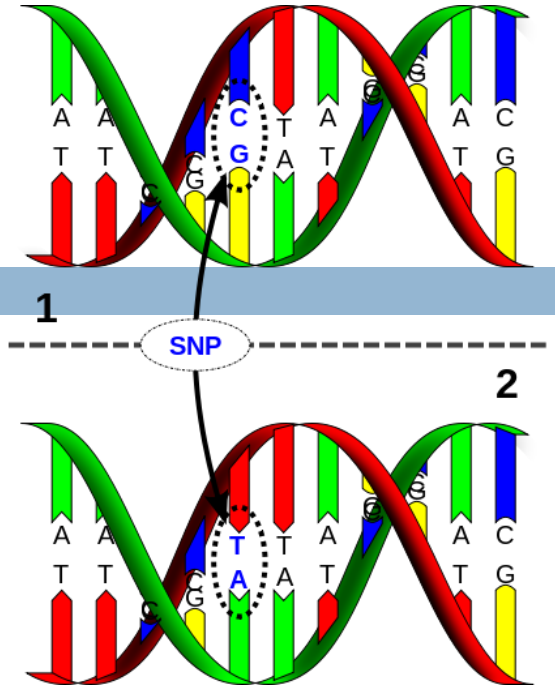
0 3

Att. iegūts: http://bozeman.genome.washington.edu/compbio/mbt599_2011/Genome540_BWA_15Feb2011.pdf

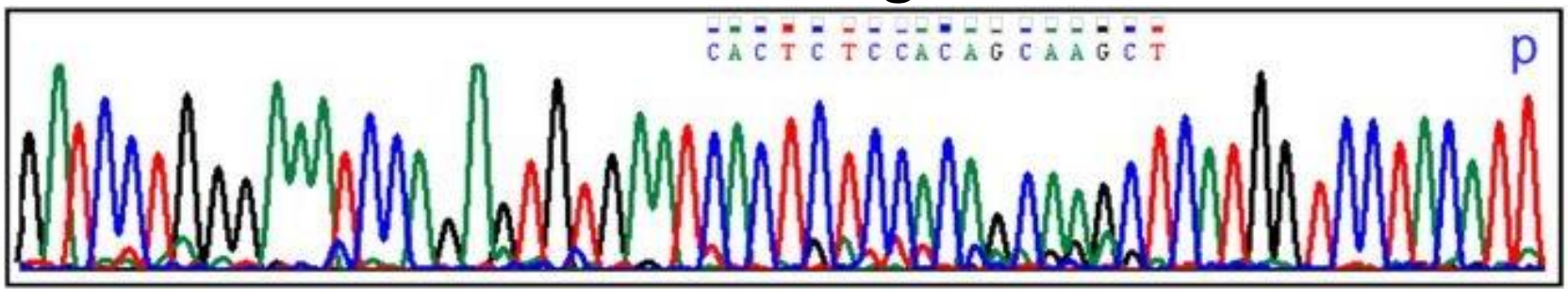
Meklēšana pēc ‘references’

- Nozīmīgākās problēmas, ar ko jātiek galā: SNP/indels, atkārtojošies reģioni.
  - SNP/indels – dažādi sodi par dažādām neatbilstībām + max. robeža vai *x* pieļaujamas neatbilstības, resp. rēķina labošanas distanci.
  - Atkārtojošies reģ. - vairāku atbilstību gadījumā nolasījumu piekārto varbūtiski vienai no atbilstošajām pozīcijām
- Visi ātrākie un populārākie ir ‘short read aligners’ un radīti Illumina datiem, tātad problēmas palielinoties raksturīgajiem eksperimentālo datu garumiem (>500bp).
- Atrākai un precīzākai meklēšanai var izmantot SNP datu bāzes. hg ~3Gbp, dbSNP ~30Mbp. Vai var iekļaut BWT indeksa veidošanas laikā?

SNP



- Single nucleotide polymorphism (parasti 2 varianti, nav vienmērīgi)
- Nozīmīgi meklējot noteiktas cilvēku grupas un slimības korelāciju
- Meklēšanā plaši lieto bezmaksas programmatūru: GATK, SAMtools.
- Abas darbojas līdzīgi - izmantojot nolasījumu varbūtības un ‘references’ genomu



Haplotipu atšifrēšana

- Diploīdos organismos (katrā šūnā 2x no katras hromosomas) – katrs SNP dod 3 dažādus variantus. Piem SNP = A/T ➔ 3 haplotipi (AA, AT, TT)
- Vienīgā viennozīmīgā metode - sekvencēšana
- Izvēlētai cilvēku grupai var noteikt varbūtību.
- Izmantotās metodes šobrīd – novērojumi par noteiktu haplotipu biežumu noteiktās kopās
- Ja pieder dažādām kopām – pēc iespējas mazāks skaits haplotipu, kas izmantots kombinācijā

Apzinātie resursi

- http://www.ncbi.nlm.nih.gov/
- http://www.ebi.ac.uk/
- EBI (A.Brāzma, N. Kurbatova)
- LV Biomedicīnas centrs (D. Frīdmanis)
- Bioloģijas Mg/PhD studenti (A. Kalviša)

Tuvākie mērķi

- Esošās programmatūras/metožu salīdzinājums uz reāliem (ne simulētiem) datiem
- Algoritms precīzākai/lielāka garuma (200 bp < l <1 kbp) eksperimentālo datu nolasījumu meklēšanai pēc ‘references’ (izmantojot SNP datu bāzes)
- Metodes implementācija (iespējams, balstīta uz esošo Bowtie2, BWA indeksēšanu, jo šobrīd BW transformācija piedāvā labāko atmiņas optimizāciju)

- NB! Ļoti nepieciešama cieša sadarbība ar attiecīgās nozares (DNS sekvencēšana) biologiem (reāli dati, konsultācijas, nozares standarti)

Paldies par uzmanību!