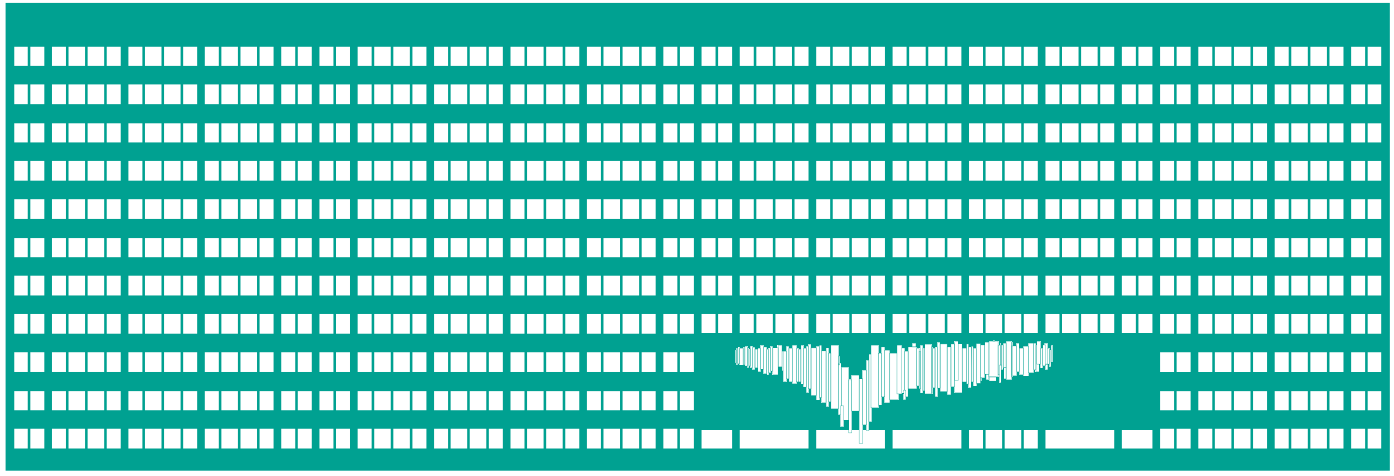
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Bioinformatika - algoritmy a analýza dat

Detekce variant

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA443

michal.vasinek@vsb.cz

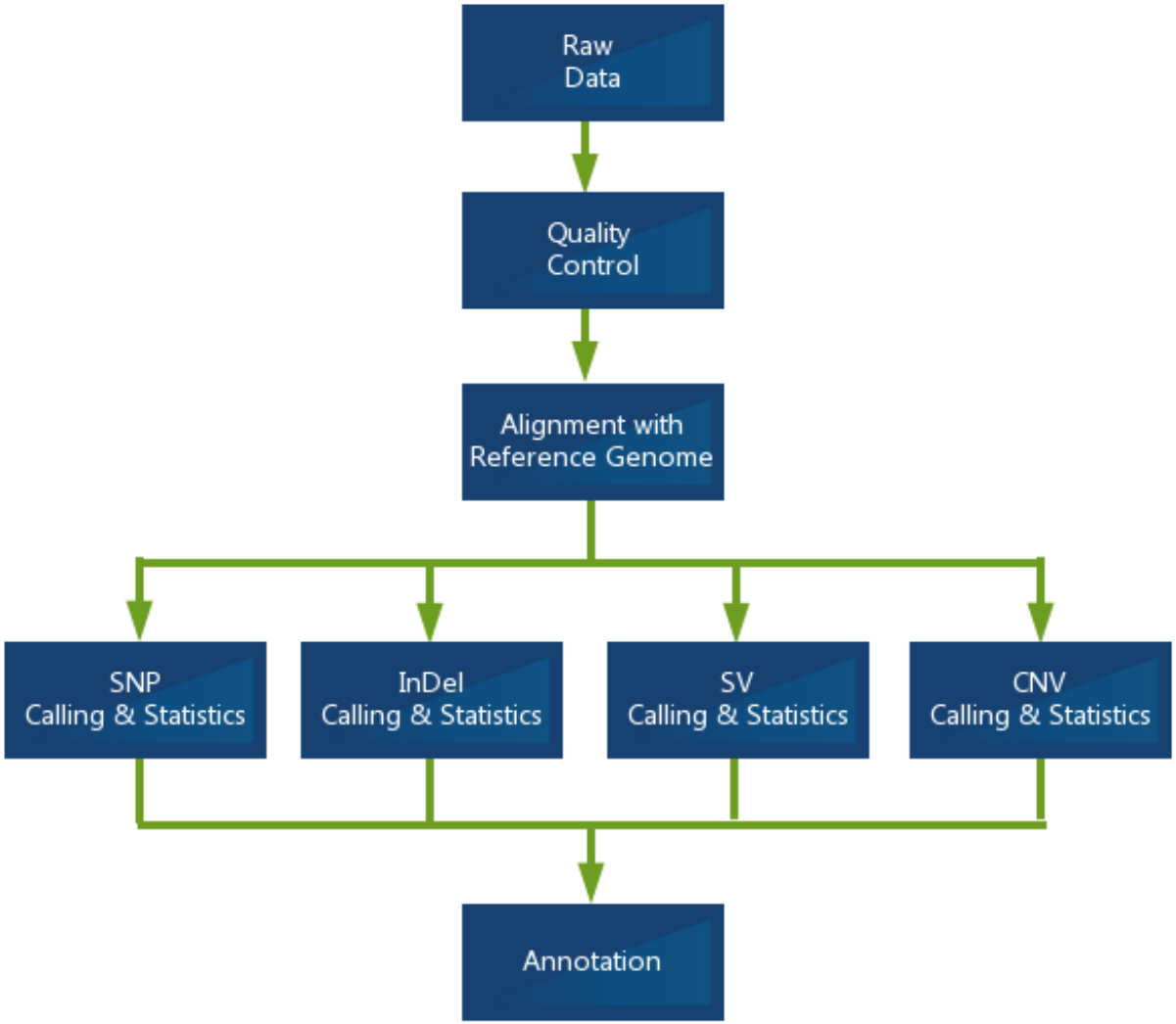
2. listopadu 2022

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Přehled přednášky

- Podstata detekce variant
  - Dědičnost
- Analýza
  - Model založený na binomickém rozdělení
  - Model založený na Bayesovské pravděpodobnosti

DNA pipeline



Dědičnost v kostce

Otázka

Kolik má jedna buňka zdravého člověka chromozómů?

- 23 párů, tudíž dohromady 46.
- Každý pár je tvořen dvěma chromozómy, jeden od každého rodiče.

Dědičnost v kostce

Otázka

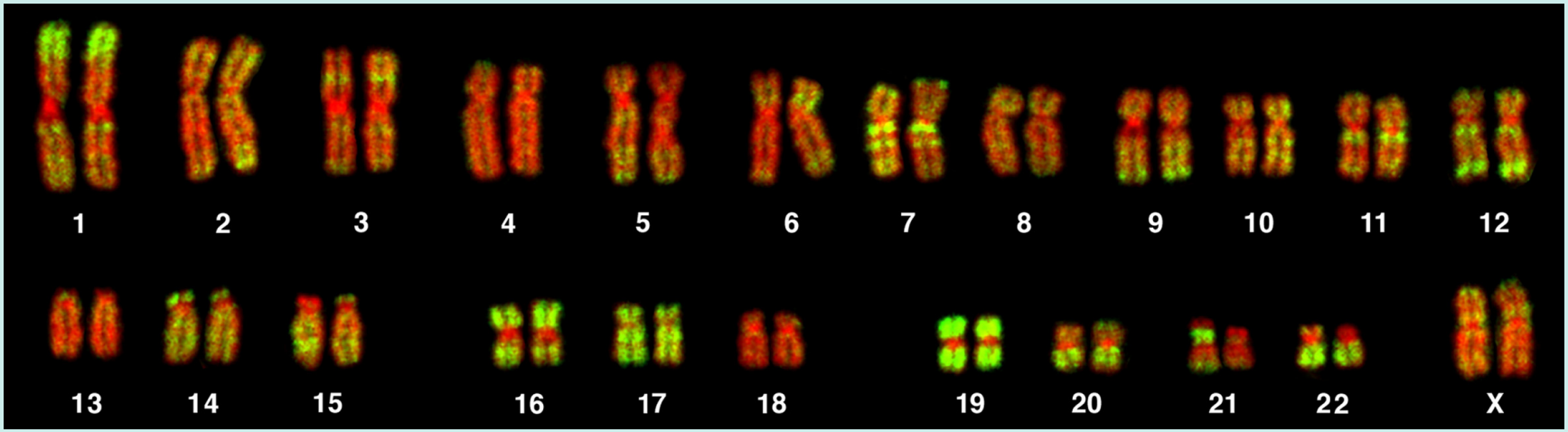
Kolik má jedna buňka zdravého člověka chromozómů?

- 23 párů, tudíž dohromady 46.
- Každý pár je tvořen dvěma chromozómy, jeden od každého rodiče.

Dědičnost v kostce

Otázka

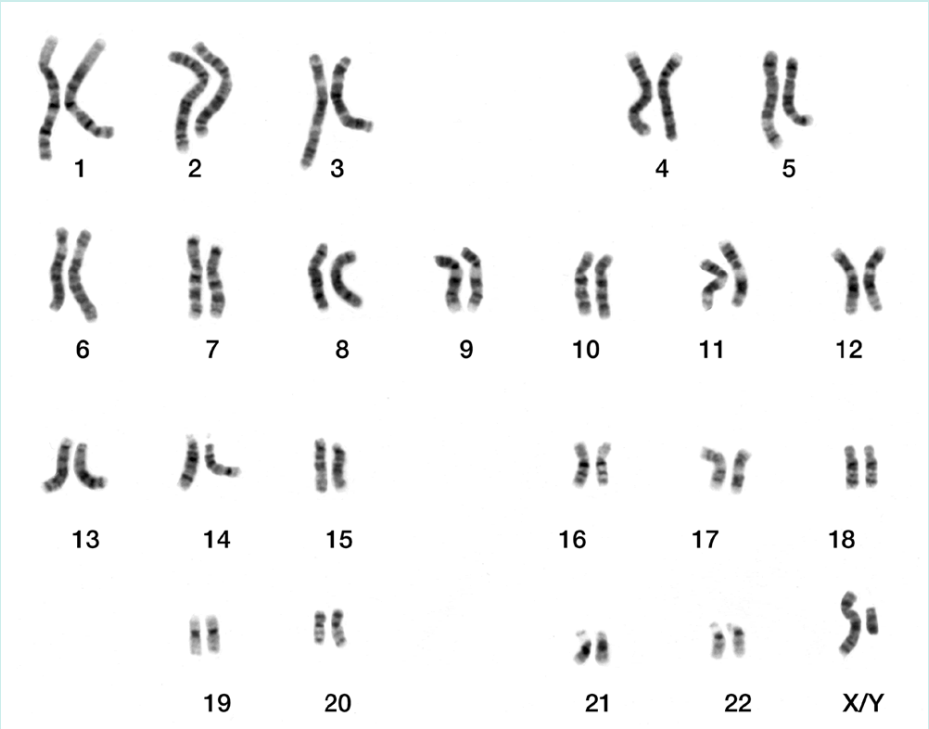
Díváme se na karyotyp(soubor všech chromozomů) muže či ženy?



Dědičnost v kostce

Otázka

Díváme se na karyotyp(soubor všech chromozomů) muže či ženy?



Dědičnost v kostce

K zamyšlení

Předpokládejte, že jste obdrželi sekvence (FASTQ) celého genomu (všech chromozómů) nějakého člověka a dále předpokládejte, že znáte referenční sekvence chromozomu X a Y. Jak byste zjistili, zda se jedná o muže či ženu?

- Zkusili byste sekvence namapovat na X a na Y, pokud na Y namapujete pouze několik náhodných sekvencí, pak se jedná o ženu. Pokud byste ovšem viděli rovnoměrné pokrytí obou chromozómů, pak by se jednalo o muže.

Dědičnost v kostce

K zamyšlení

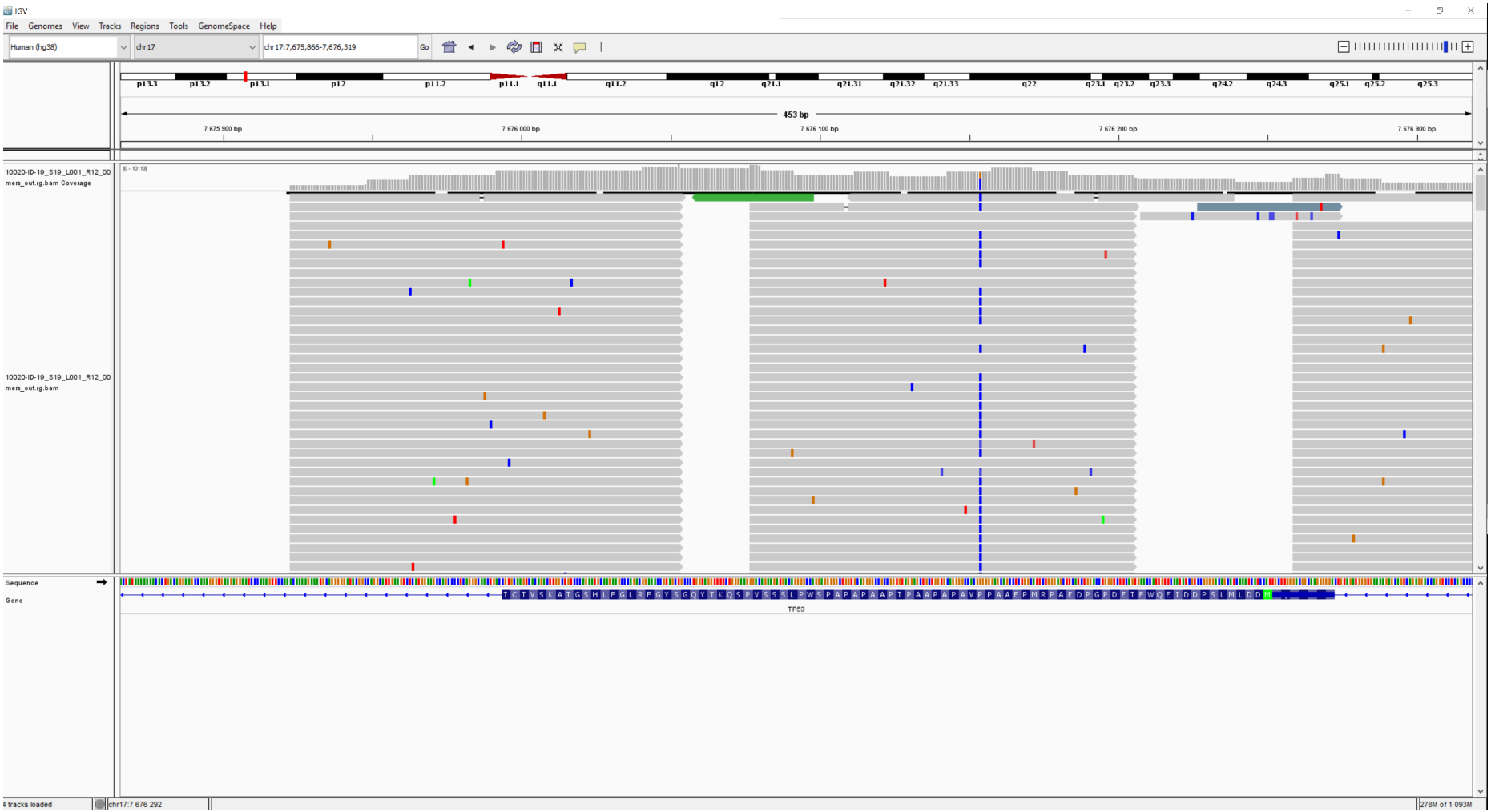
Předpokládejte, že jste obdrželi sekvence (FASTQ) celého genomu (všech chromozómů) nějakého člověka a dále předpokládejte, že znáte referenční sekvence chromozomu X a Y. Jak byste zjistili, zda se jedná o muže či ženu?

- Zkusili byste sekvence namapovat na X a na Y, pokud na Y namapujete pouze několik náhodných sekvencí, pak se jedná o ženu. Pokud byste ovšem viděli rovnoměrné pokrytí obou chromozómů, pak by se jednalo o muže.

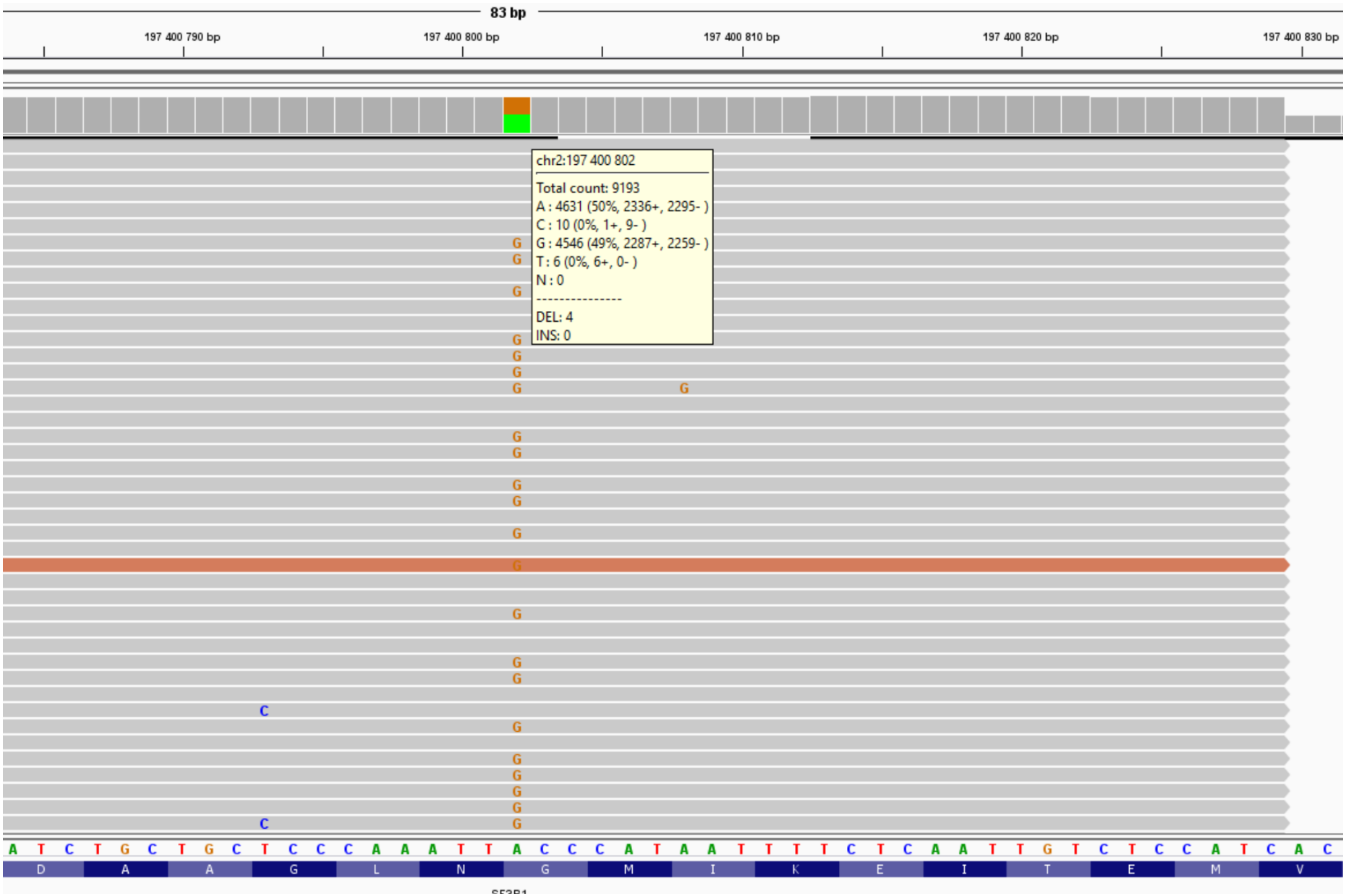
Typy variant

- Germline - zárodečné - vzniklé v důsledku rozmnožování meiozou.
  - Heterozygotní varianta - vidíme variantu cca 50:50, od obou rodičů jiná báze.
  - Homozygotní varianta - vidíme variantu cca 100%, od obou rodičů stejná báze ale odlišná od referenční.
- Somatické - tělní - vznikají v průběhu života organismu.

IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human (hg38)
chr17
chr17:7,675,866-7,676,319
Go
453 bp
7 675 900 bp
7 676 000 bp
7 676 100 bp
7 676 200 bp
7 676 300 bp
Sequence
Gene
TP53
4 tracks loaded
chr17:7 676 292
278M of 1 093M

Heterozygotní varianta



Model založený na binomickém rozdělení

- Předpokládejme, že sekvenační error je nezávislá náhodná veličina: záměna v jedné pozice nemá vliv na záměnu na dalších pozicích.
- Nechť P_e značí pravděpodobnost, že je jeden konkrétní nukleotid nesprávně přečten.

Problém

Mějme n sekvencí, které pokrývají pozici p_i v referenční genomu. Kolik z těchto sekvencí obsahuje na dané pozici nesprávně přečtený nukleotid, pokud je pravděpodobnost nesprávného čtení p_e?

Model založený na binomickém rozdělení

Problém

Mějme n sekvencí, které pokrývají pozici p_i v referenční genomu. Kolik z těchto sekvencí obsahuje na dané pozici nesprávně přečtený nukleotid, pokud je pravděpodobnost nesprávného čtení p_e?

- Lze modelovat binomickým rozdělením $B(n, p)$.

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

- $P[X = x]$ je pravděpodobnost, že chyba nastane právě x-krát v n sekvencích.

Sekvenační chyba pro různé platformy

Platform	P_e [%]
MiSeq	0.473
MiniSeq	0.613
NextSeq 500	0.429
HiSeq 2500	0.112
NovaSeq 6000	0.109
PacBio SMRT	15 [1]
Oxford Nanopore	7[2]

Tabulka: Sequencing error rates as number of mismatches.[3]

[1] https://www.genengnews.com/insights/dna-sequencing-accuracy-comes-a-long-way/

[2] Sahlin, Medvedev, Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis, Nature communications, 2021

[3] Stoler, Nekrutenko, Sequencing error profiles of Illumina sequencing instruments, NAR Genomics and Bioinformatics, 2021

Více readů => eliminace chyb

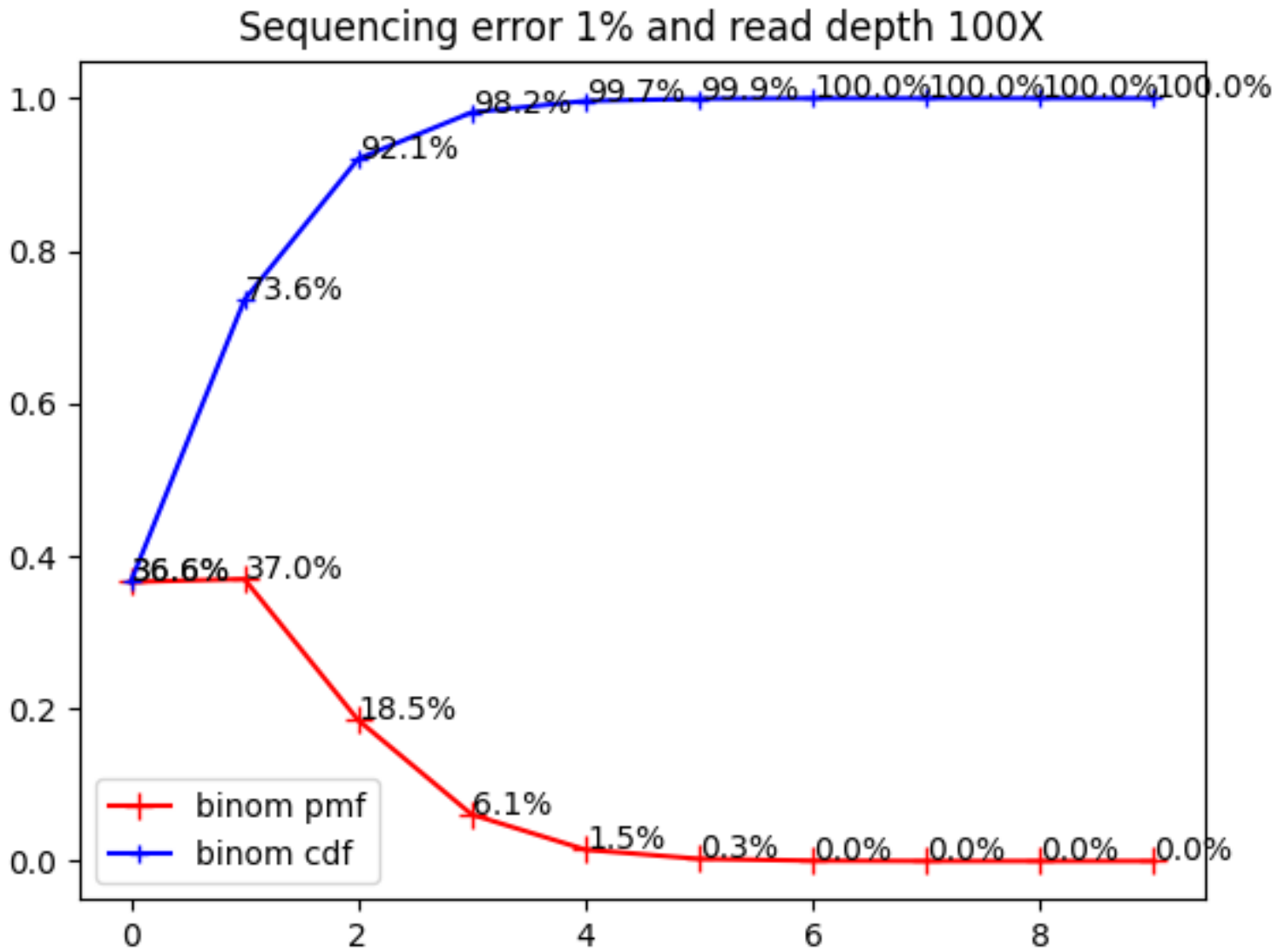
[4]PacBio reads typically have a really high error rate (15% compared with 0.1% for Illumina.) However, their errors tend to be random, so if the same region is sequenced several times, the errors average out resulting in a "consensus" sequence.

Princip

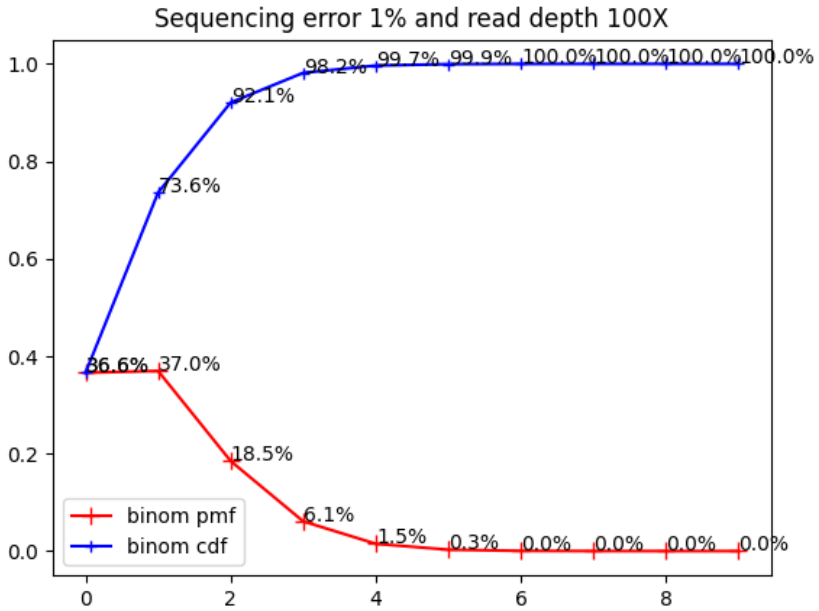
Shawn Baker, PhD, consultant at SanDiegOmics.com explains it like this: "If you have a 1% error rate and you sequence to 100X depth, at each base you'll get roughly 99 reads that look the same, say an 'A', and one base that is different, say a 'G'. The consensus call would be to assume that the base is truly an 'A' and ignore the 'G' call."

[4]https://www.genengnews.com/insights/dna-sequencing-accuracy-comes-a-long-way/

Pravděpodobnost, že N je readů je chybně přečteno

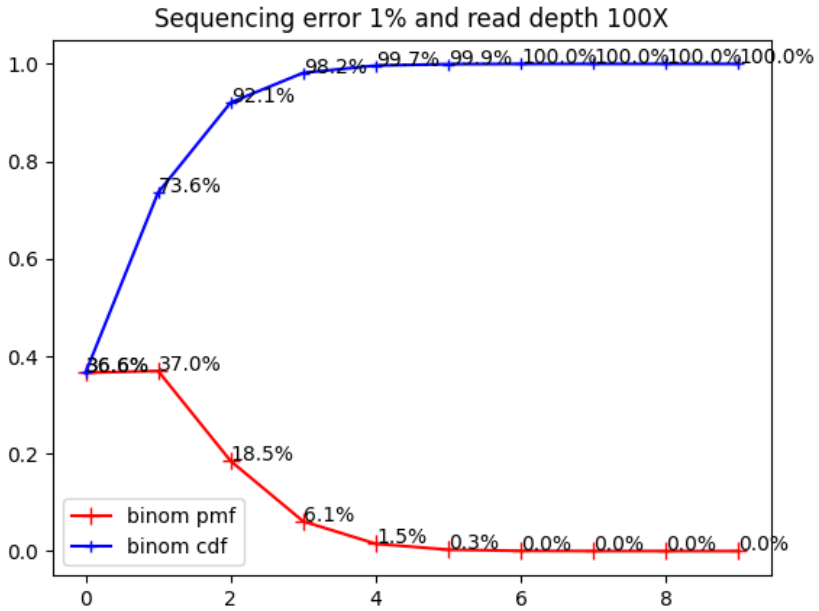


Pravděpodobnost, že N je readů je chybně přečteno



- Uvažujme, že sekvenujeme oblast o délce $d = 100bp$. Kolik bychom očekávali pozic, které budou mít nesprávně přečteny 4 báze?
- $\mathbb{E}(4 \text{ báze špatně}) = p[X = 4]d \approx 1.5$

Pravděpodobnost, že N je readů je chybně přečteno



- Uvažujme, že sekvenujeme oblast o délce $d = 100bp$. Kolik bychom očekávali pozic, které budou mít nesprávně přečteny 4 báze?
- $\mathbb{E}(\textit{4 báze špatně}) = p[X = 4]d \approx 1.5$

Pravděpodobnost, že N je readů je chybně přečteno

- Uvažujme, že sekvenujeme celý lidský genom $d = 3.2Gbp$ s pokrytím 100. Jak se změní čekávaný počet pozic s N chybně přečtenými bázemi v readech?

N	$\mathbb{E}(N$ *špatně*$)$
1	1183134840
2	591567420
3	195197330
4	47813487
5	9272918
6	1483042
7	201163
8	23621
9	2439
10	224
11	18
12	1

Opakování experimentu

Otázka

Pokud experiment zopakujeme, jaká je pravděpodobnost, že se na stejné pozici objeví stejný počet chybných čtení?

$$p[X = N, Y = N] = p[X = N]p[Y = N]$$

- X a Y jsou stejné genomy tudíž:

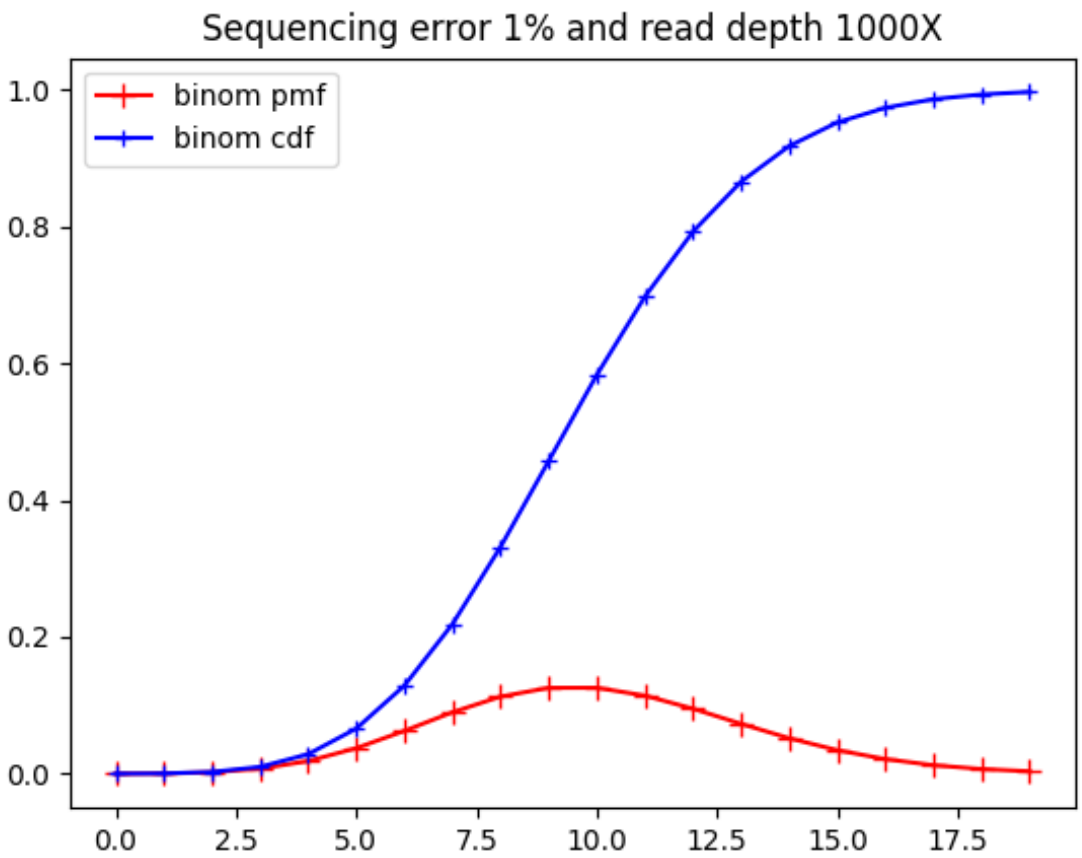
$$p[X = N, Y = N] = (p[X = N])^2$$

- Opakování vede k významné redukci chybné interpretace sekvence chybných čtení jako strukturální varianty!

Pravděpodobnost, že N je readů je chybně přečteno

Otázka

Co se stane, pokud zvýšíme pokrytí Cov?



Pravděpodobnost, že N je readů je chybně přečteno

Otázka

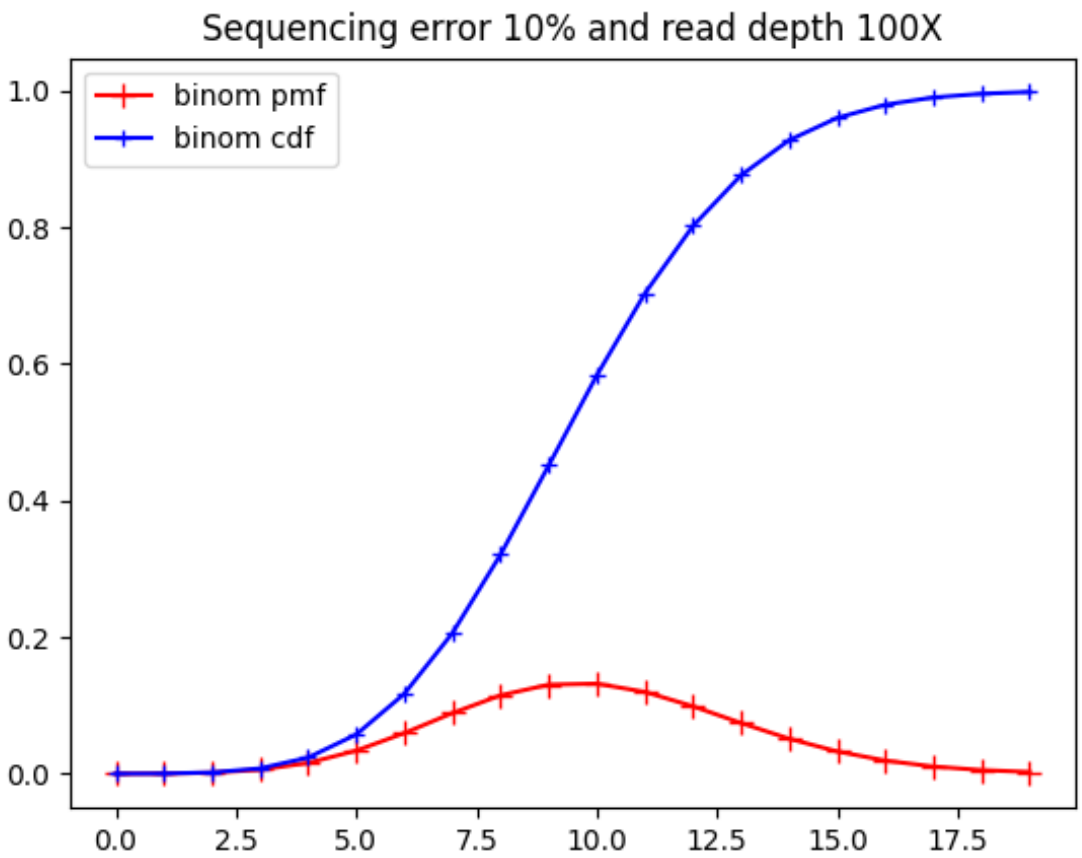
Co se stane, pokud zvýšíme pokrytí Cov?

- Při pokrytí 100X, jsme měli 99% pravděpodobnost 4 a méně chyb. Nicméně 4 chyby mohou potenciálně znamenat 4 procentní variantu zámeny v jednom nukleotidu (tzv. SNP-single nucleotide polymorphism).
- Při pokrytí 1000X, jsme měli 99% pravděpodobnost 20 a méně chyb. Nicméně v tomto případě se jedná pouze o $20/1000 = 2\%$ variantu.
- Pozor: Vyšší pokrytí znamená větší finanční náklady.

Pravděpodobnost, že N je readů je chybně přečteno

Otázka

Co se stane, pokud přejdeme na technologii s vyšší sekvenační chybou?

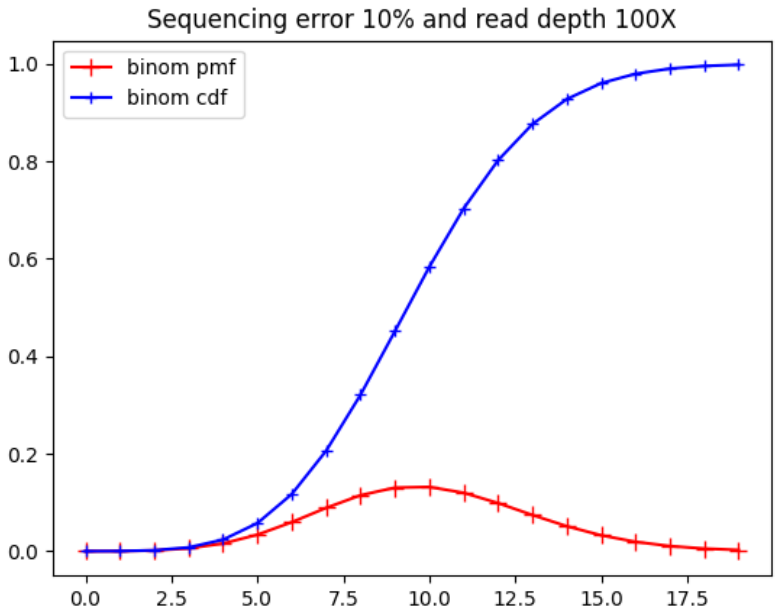


Pravděpodobnost, že N je readů je chybně přečteno

Otázka

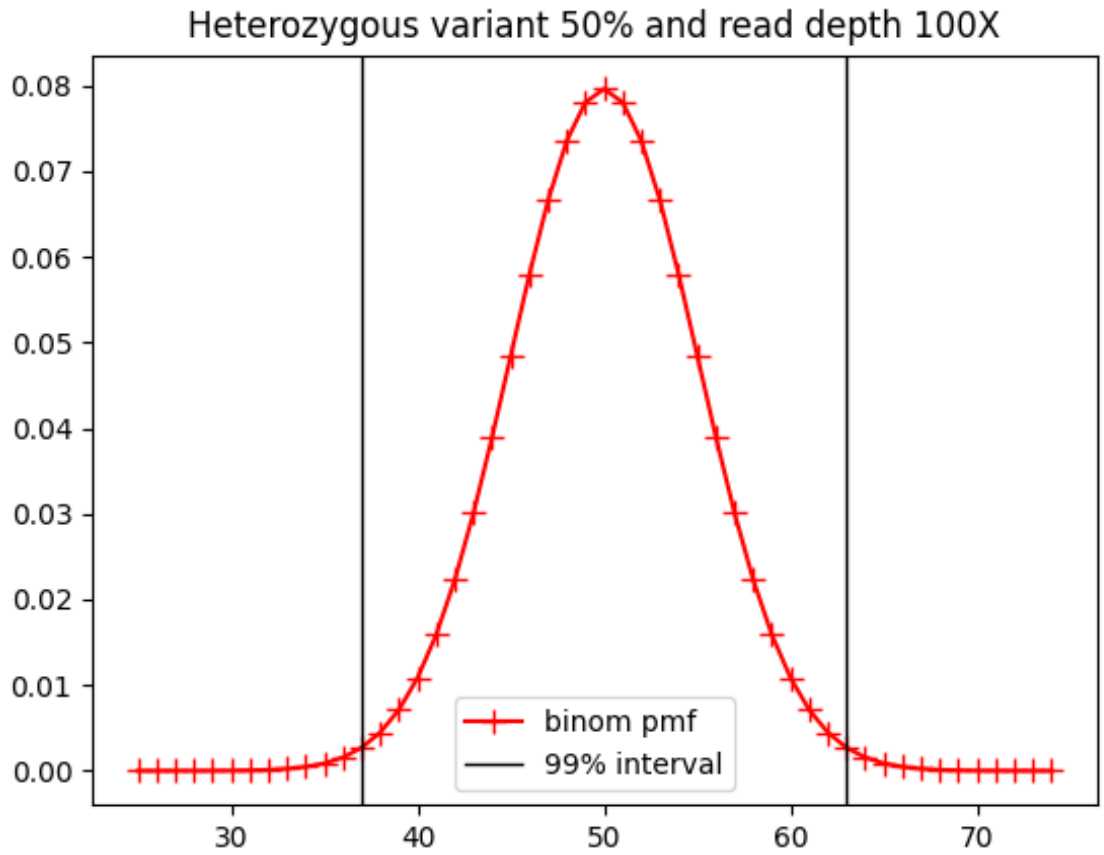
Co se stane, pokud přejdeme na technologii s vyšší sekvenační chybou?

- Sníží se naše citlivost detekce x-procentní varianty.
- Více, než 50% pozic bude pokryto více, než 10% chybných čtení.



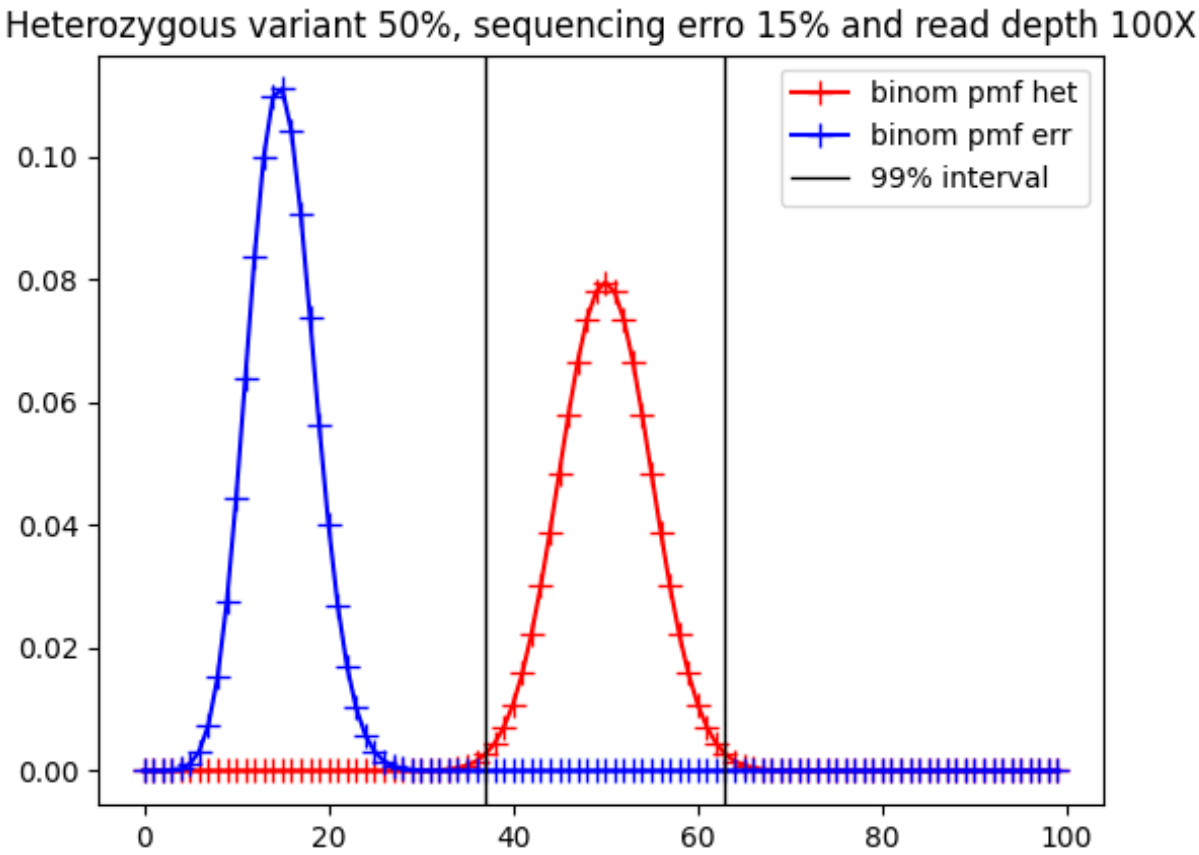
Heterozygotní varianta

- Cokoli mezi 35% a 65% lze interpretovat jako heterozygotní variantu.



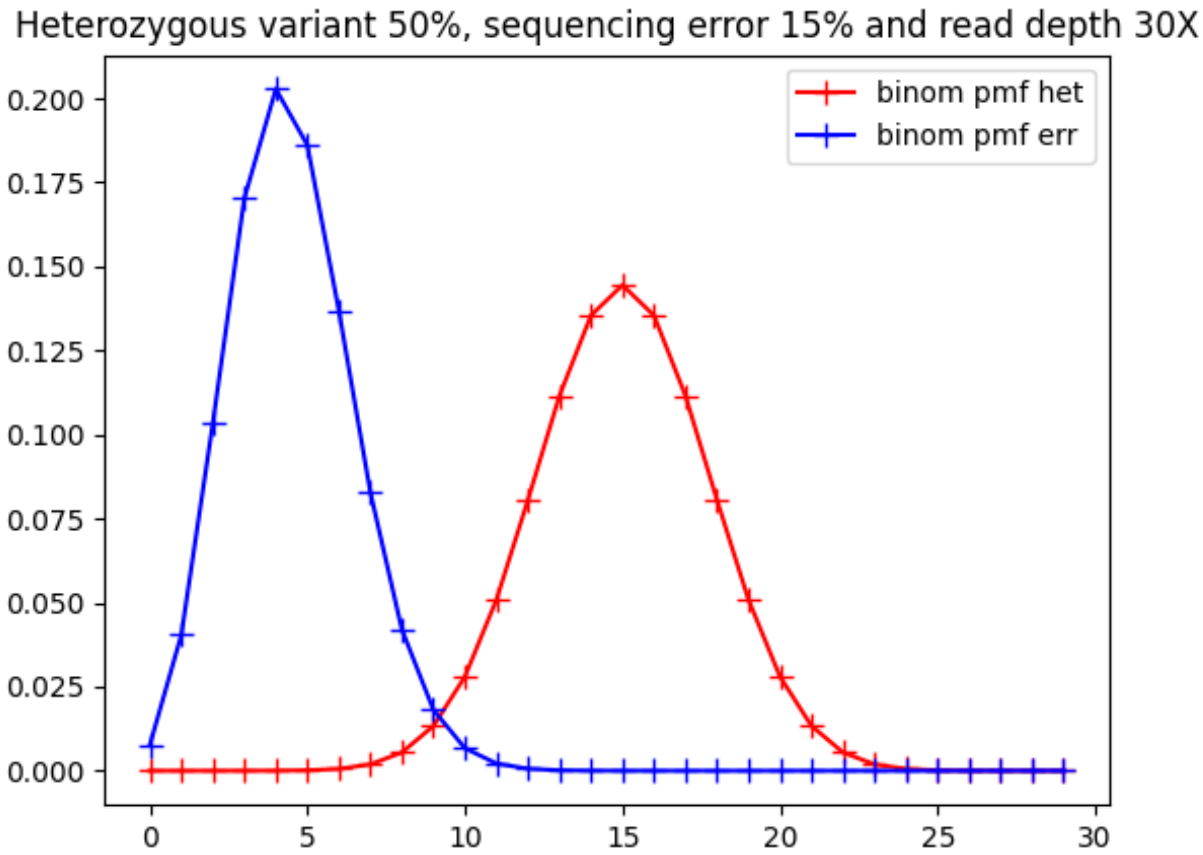
Heterozygotní varianta

- Cokoli mezi 35% a 65% lze interpretovat jako heterozygotní variantu.



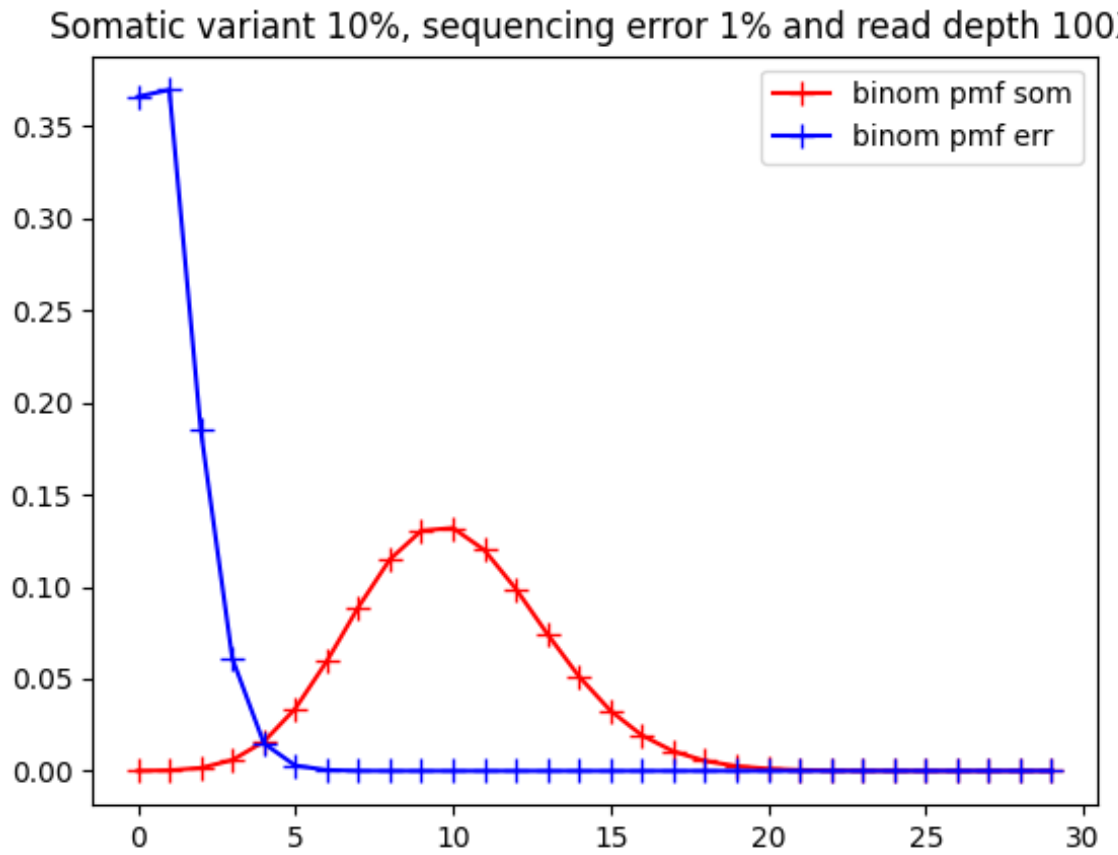
Heterozygotní varianta

- Cokoli mezi 35% a 65% lze interpretovat jako heterozygotní variantu.



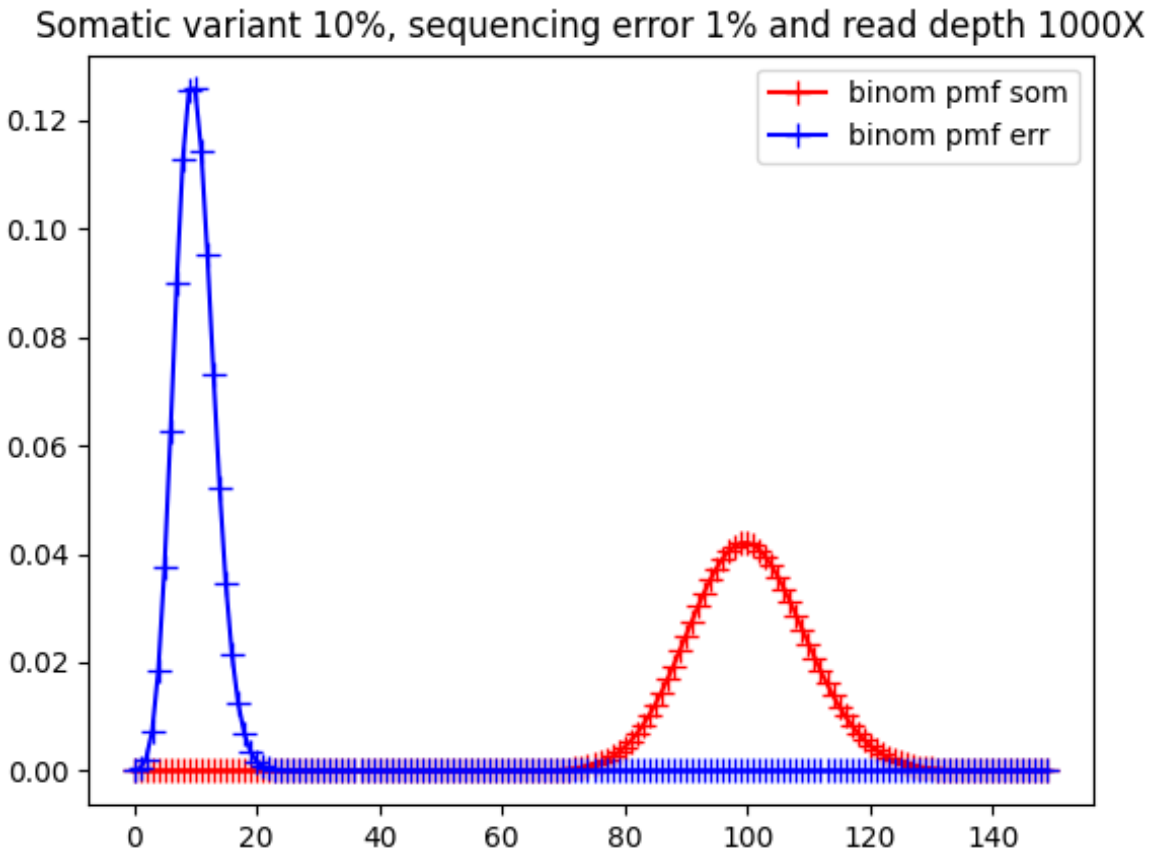
Somatická varianta

- Somatická varianta se obvykle vyskytuje v řádu jednotek a nižších desítek procent.
- Vyšší zastoupení nerozlišitelné od heterozygotní varianty.



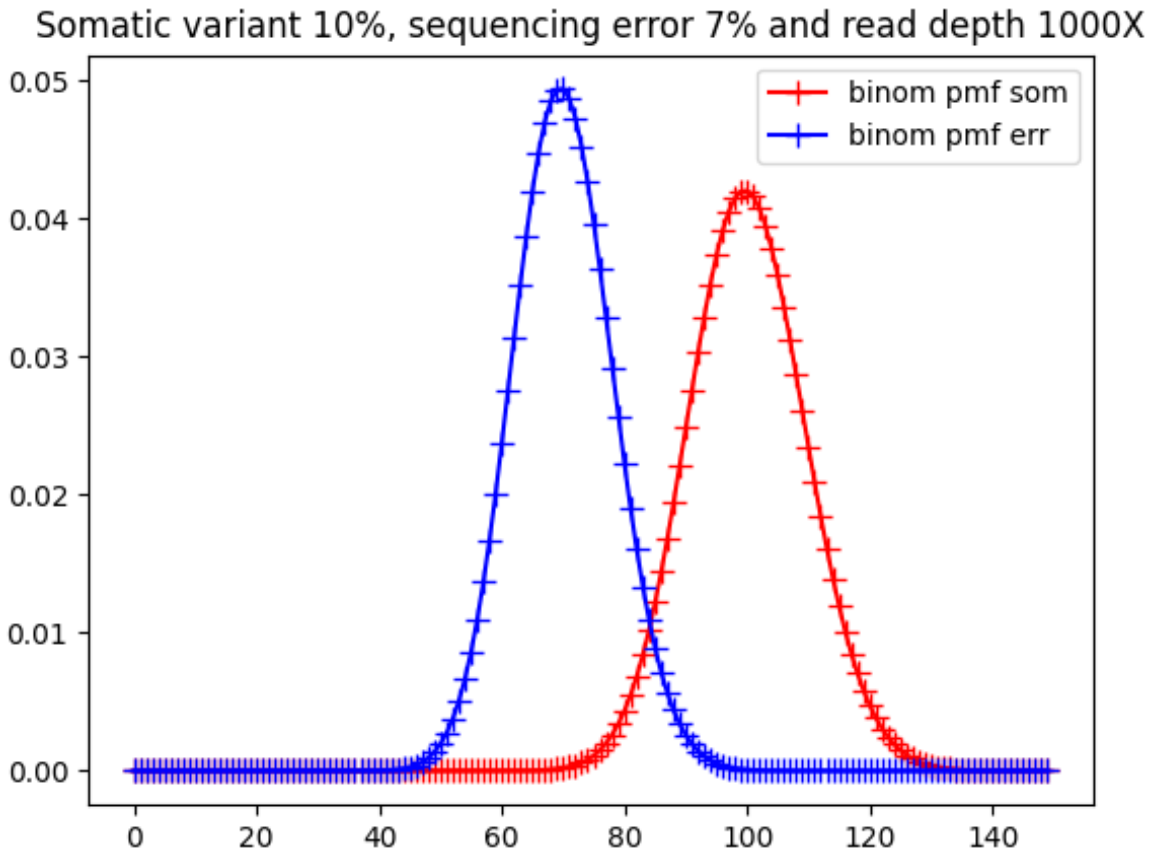
Somatická varianta

- Opět lze chybu oddělit zvýšením pokrytí.



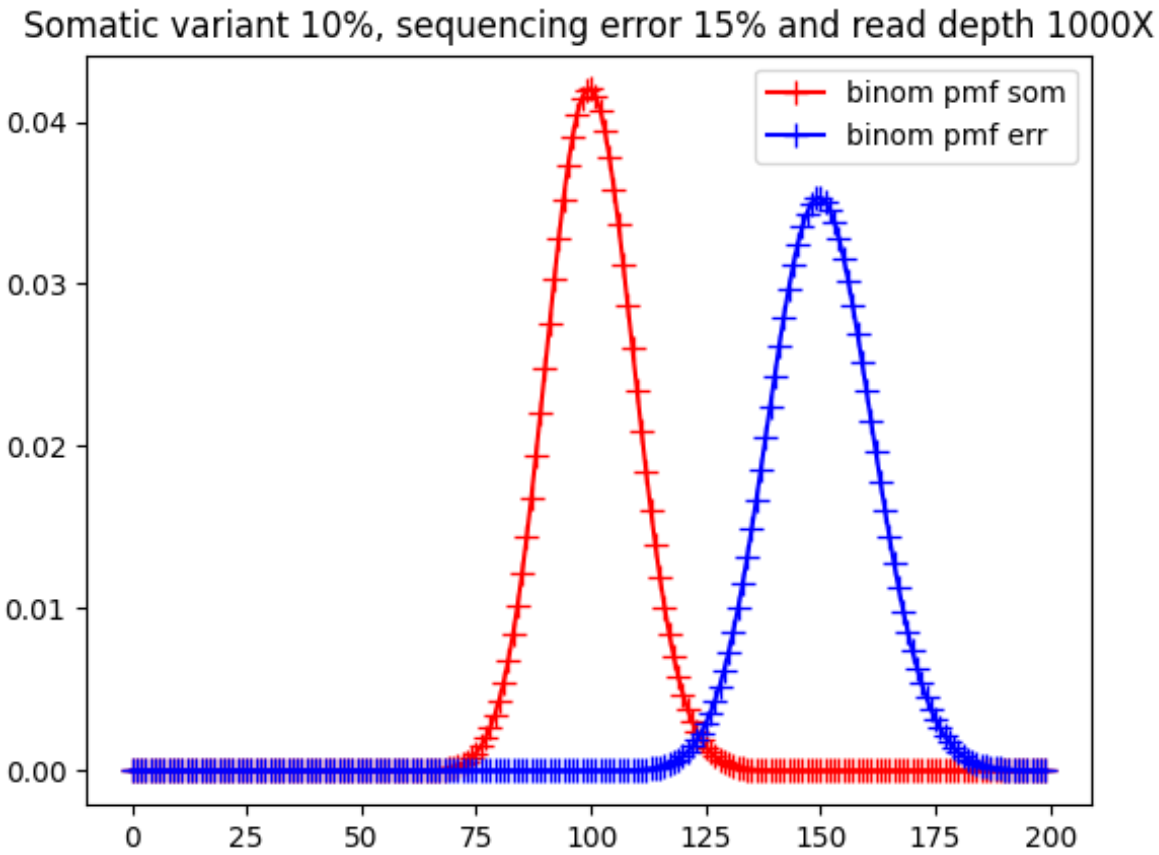
Somatická varianta

- Pokrytí 1000X u Oxford Nanopore.



Somatická varianta

- Pokrytí 1000X u PacBio.



Model je jenom model

- Předložená analýza je pouze základní model.
- To, že nám vyjde určitý počet chyb neznámá, že všechny mají stejný nukleotid.
- Sekvenační chyby nejsou nutně nezávislé a záleží na použité sekvenační platformě.
- Sekvence mají obvykle nižší kvalitu čtení blízko konci sekvence.
- Obecně, kvůli vysokým nákladům na opakovaná měření preferujeme nižší falešnou positivitu strukturálních variant.

Shrnutí

- Technologie 3.generace (dlouhé ready Oxford Nanopore, PacBio) produkují 100kb sekvence, ale vyšší chybovost neumožňuje provádět detekci somatických variant.
- Technologie 2.generace má velmi nízkou chybovost, ale krátké ready nám neumožní detekovat strukturální varianty delší než několik desítek bazí.
- Jistotu o detekované variantě můžeme zvýšit opakováním experimentu případně zvýšením coverage => obě možnosti však vedou na násobně vyšší náklady.

Principy GATK

- GATK - Genome Analysis ToolKit[5]
- Snažíme se najít Bayesovu pravděpodobnost, že data podporují určitý genotyp.
- Předpokládejme, že v referenční sekvenci máme A, rozlišujeme následujících deset genotypů:
  - Homozygotní s referencí - AA
  - Homozygotní varianta - CC, GG, TT
  - Heterozygotní vůči referenci - AC, AG, AT
  - Heterozygotní mimo referenci - CG, CT, GT

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928508/pdf/1297.pdf

Bayesovo pravidlo

- $P(A|B)$ je podmíněná pravděpodobnost, že nastane A pokud nastalo B.
- $P(A)$ resp. $P(B)$ je pravděpodobnost A resp. B.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayesovo pravidlo - příklad Hello World

Example

nemoc/příznak	Ano	Ne	Celkem
Ano	10	0	10
Ne	90	900	990
Celkem	100	900	1000

$$P(Nemoc|Priznak) = \frac{P(Priznak|Nemoc)P(Nemoc)}{P(Priznak)}$$

$$P(Nemoc|Priznak) = \frac{1 \times 0.01}{100/1000} = 0.1 = \frac{10}{90}$$

Bayesovo pravidlo - detekce variant

- Jaká je pravděpodobnost, že jsme detekovali konkrétní genotyp G, když jsme naměřili data D?
- Např. podmíněná pravděpodobnost, že jsme naměřili genotyp AA, pokud jsme naměřili data D:

$$P(AA|D) = \frac{P(D|AA)P(AA)}{P(D)}$$

- Neznáme přesné hodnoty v tabulce, viz. předchozí slide, známe pouze data D.
- Potřebujeme znát $P(D)$, nicméně je konstantní napříč všemi genotypy => můžeme zanedbat.

Bayesovo pravidlo

Udává pravděpodobnost, že na základě pozorovaných dat D se jedná o daný genotyp G.

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}$$

- G - konkrétní genotyp
- D - data na určité pozici v genomu

$$p(D|G) = \prod_{b \in B} p(b|G)$$

- B je soubor znalostí z namapovaných readů popisující báze A,C,G,T => všechny báze z readů.

Bayesovo pravidlo pro detekci variant

$$p(b|G) = p(b|\{A_1, A_2\}) = \frac{1}{2}p(b|A_1) + \frac{1}{2}p(b|A_2)$$

- A_1 a A_2 jsou příslušné báze, pokud genotyp rozložíme na jednotlivé alely(báze). Např. pro genotyp CT znamená, že $A_1 = C$ a $A_2 = T$.

$$P(b|A) = \begin{cases} e/3 & \text{if } b \neq A \\ 1 - e & \text{if } b = A \end{cases}$$

- e je ohodnocení kvality dané báze. ASCII reprezentace převedená zpět na pravděpodobnost.

Příklad $P(D|G)$

Báze	PHRED	ASCII	P_{err}
A	B	66	0.0005
A	?	63	0.001
T	5	53	0.01

Tabulka: Příklad bází a jejich ohodnocení chyby

$$P(A|AA) = \frac{1}{2}p(A|A) + \frac{1}{2}p(A|A)$$
$$= 0.5(1 - 0.0005) + 0.5(1 - 0.0005) = 0.9995$$

Příklad $P(D|G)$

Báze	PHRED	ASCII	P_{err}
A	B	66	0.0005
A	?	63	0.001
T	5	53	0.01

Tabulka: Příklad bází a jejich ohodnocení chyby

$$P(A|AT) = \frac{1}{2}p(A|A) + \frac{1}{2}p(A|T)$$
$$= 0.5(1 - 0.0005) + 0.5(0.0005/3) = 0.499$$

Příklad $P(D|G)$

Báze	PHRED	ASCII	P_{err}
A	B	66	0.0005
A	?	63	0.001
T	5	53	0.01

Tabulka: Příklad bází a jejich ohodnocení chyby

$$P(A|CG) = \frac{1}{2}p(A|C) + \frac{1}{2}p(A|G)$$
$$= 0.5(0.0005/3) + 0.5(0.0005/3) = 0.00016$$

Příklad $P(D|G)$

Báze	PHRED	ASCII	P_{err}
A	B	66	0.0005
A	?	63	0.001
T	5	53	0.01

Tabulka: Příklad bází a jejich ohodnocení chyby

$$P(D|AT) = P(A|AT) \times P(A|AT) \times P(T|AT)$$

- Co se stane, pokud budete násobit 1000 malých čísel?
- Obvyklé řešení použít logaritmy: $\log abc = \log a + \log b + \log c$

Určení pravděpodobnosti genotypu $P(G)$

- Parametr heterozygosita $h = 0.01$. S jakou pravděpodobností se vyskytuje heterozygotní varianta. Populační charakteristika.
- Parametr chyby reference $e_{err} = 0.01$. Různé populace mají různé referenční sekvence.
- Pokud máme heterozygotní variantu mimo referenci A, tedy CG, CT, GT, pak předpokládáme, že se jedná o chybu reference:

$$p(hetvar_{non-ref}) = h \times e_{err}$$

Určení pravděpodobnosti genotypu $P(G)$

- Pokud máme homozygotní variantu - CC, GG, TT:

$$p(homvar) = h/2$$

- Pokud máme heterozygotní variantu s jednou referenční bází - AC, AG, AT:

$$p(hetvar) = h$$

- Shoda s referencí - AA:

$$p(match) = 1 - \frac{3h}{2}$$

GATK - dokončení

- Vybíráme dva nejpravděpodobnější genotypy.
- Spočítáme LOD (log odds) skóre a ověříme, že je větší než předdefinovaný limit, pokud ano, pak daný genotyp vybereme jako detekovanou variantu.
- $odds$ je poměr mezi pravděpodobností nejpravděpodobnější varianty p_1 vůči druhé nejpravděpodobnější variantě p_2:

$$LOD(p_1, p_2) = \log \frac{p_1}{p_2} = \log p_1 - \log p_2$$

GATK - příklad dokončení

Báze	PHRED	ASCII	P_{err}
A	B	66	0.0005
A	?	63	0.001
T	5	53	0.01

Tabulka: Příklad bází a jejich ohodnocení chyby

$$P(AA|D) = P(AA) \times P(D|AA) = 0.985 \times 0.9995 \times 0.9995 \times 0.0033 = 0.0032$$

$$P(AT|D) = P(AT) \times P(D|AT) = 0.01 \times 0.496 \times 0.499 \times 0.499 = 0.0012$$

$$LOD = \log 0.0032 - \log 0.0012 = 1.41$$

Výsledek VCF soubor

- VCF - variant calling format.
- Soubor pro zápis detekovaných variant.

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  4695-JK-16_S82_L001_R12_001.primers_out
chr17   7674326 rs17880604      C       G       2134.60 PASS    AC=1;AF=0.500;AN=2;BaseQRankSum=8.198;DB;DP=115;
chr17   7674797 rs1625895       T       C       5441.03 PASS    AC=2;AF=1.00;AN=2;DB;DP=127;ExcessHet=3.0103;FS=
chr17   7675327 rs2909430       C       T       11254.03        PASS    AC=2;AF=1.00;AN=2;BaseQRankSum=2.095;DB;
chr17   7676154 rs1042522       G       C       27150.03        PASS    AC=2;AF=1.00;AN=2;BaseQRankSum=-0.434;DB
chr17   7676483 rs1642785       G       C       24897.03        PASS    AC=2;AF=1.00;AN=2;BaseQRankSum=2.896;DB;
chr12   25225782        .       GA      G       643.64  REJECT  AC=1;AF=0.500;AN=2;BaseQRankSum=-0.830;DP=404;Ex
chr17   7668836 .       GAAAA   G,GAAA  673.12  REJECT  AC=1,1;AF=0.500,0.500;AN=2;BaseQRankSum=1.929;DP=411;Exc
chr17   7669644 .       GT      G       137.64  REJECT  AC=1;AF=0.500;AN=2;BaseQRankSum=-3.241;DP=463;ExcessHet=
chr17   7675393 rs5819162       CTTTT   C,CT    14597.10        PASS    AC=1,1;AF=0.500,0.500;AN=2;BaseQRankSum=
chr17   7676317 rs1332047620    CCCCCAGCCCCCCAGCCCTCCAGGT       C       100.60  REJECT  AC=1;AF=0.500;AN=2;BaseQ
chr17   7676325 .       CCCCCAGCCCTCCAGGTCCCCAGCCCTCCAGGT       *,CCCCCAGCCCTCCAGGT     2892.04 PASS    AC=1,1;A
```

DÄŻkuji za pozornost

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA443

michal.vasinek@vsb.cz

2. listopadu 2022

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY