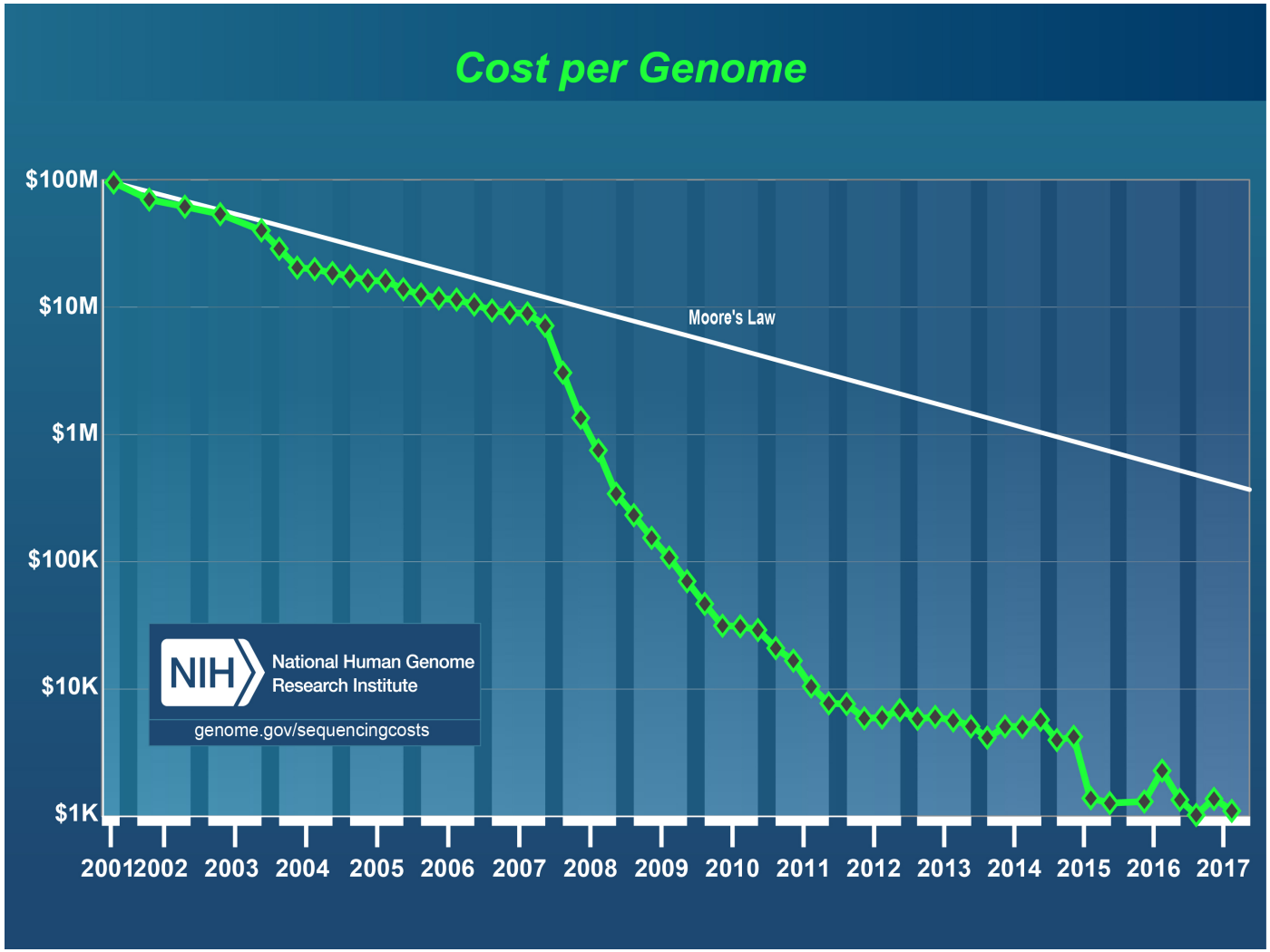
Predikcia vlastností polymorfných genómov zo sekvenačných dát

Werner Krampl
Školiteľ: doc. Mgr. Bronislava Brejová, PhD.

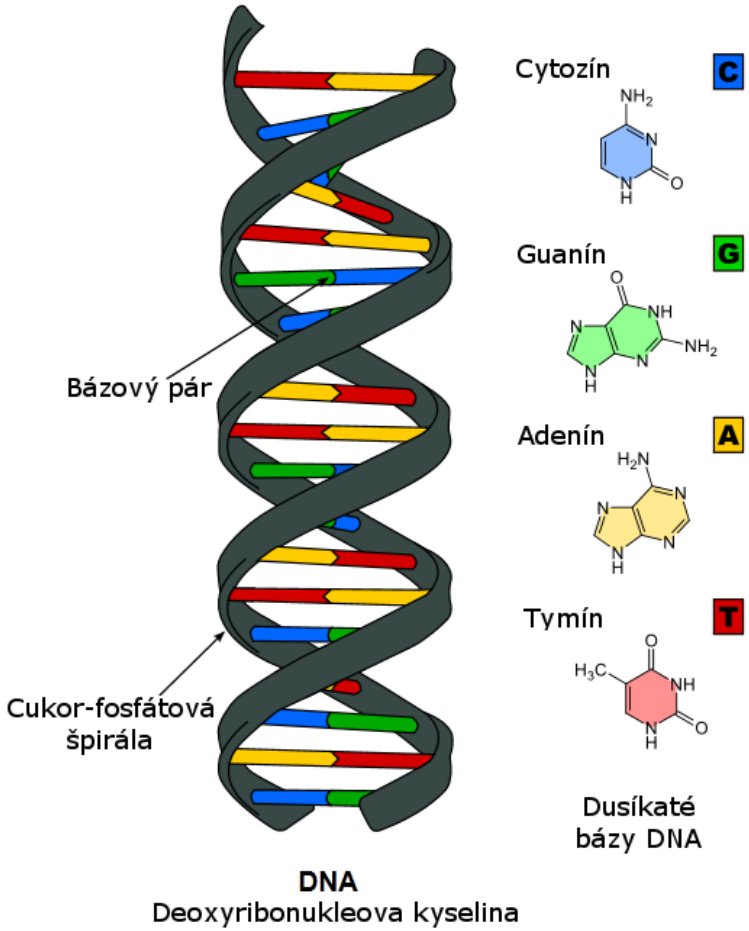
FMFI UK, 2018

Úvod

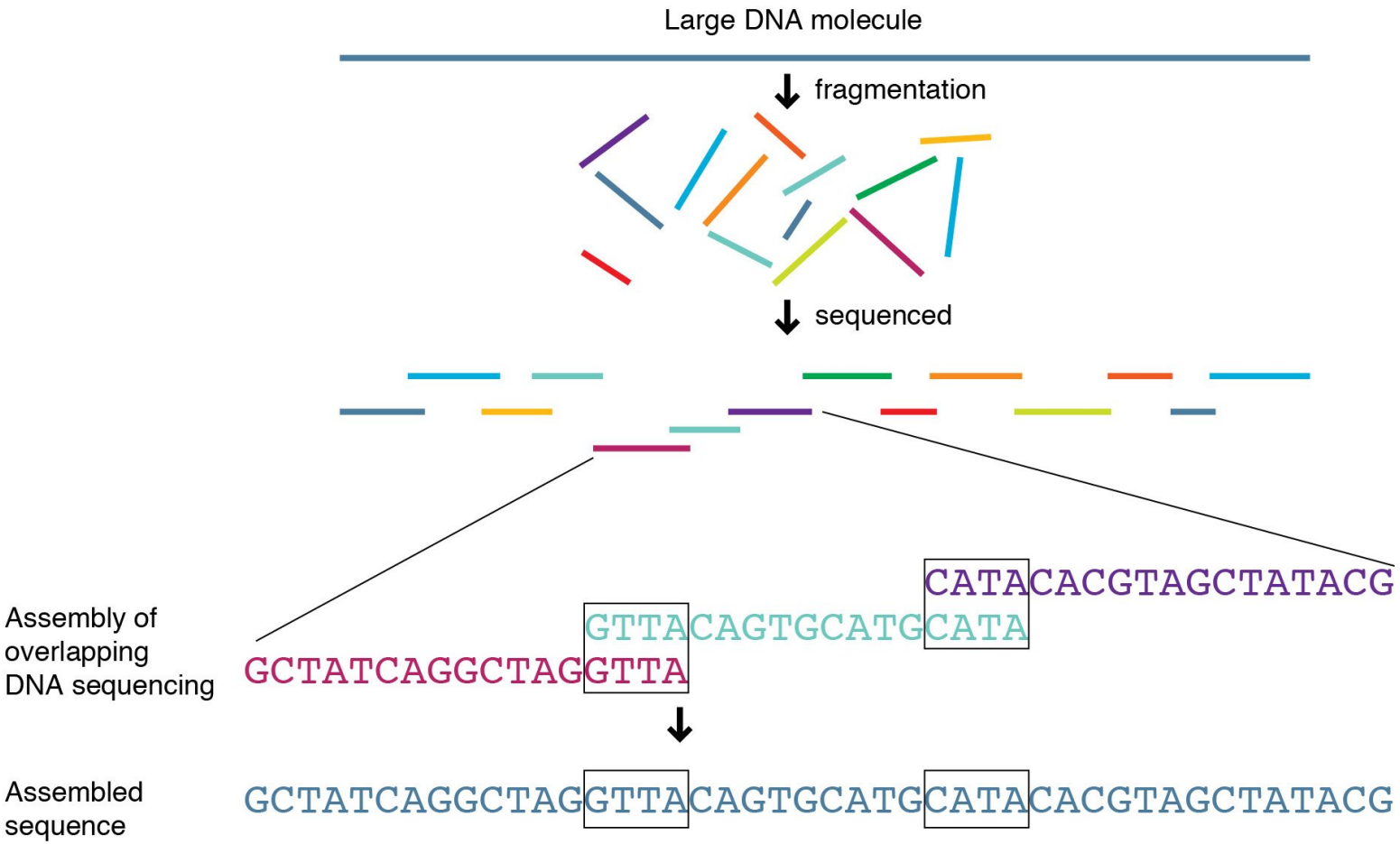


Obr. 1: Vývoj ceny za osekvenovaný genóm.

Základné pojmy



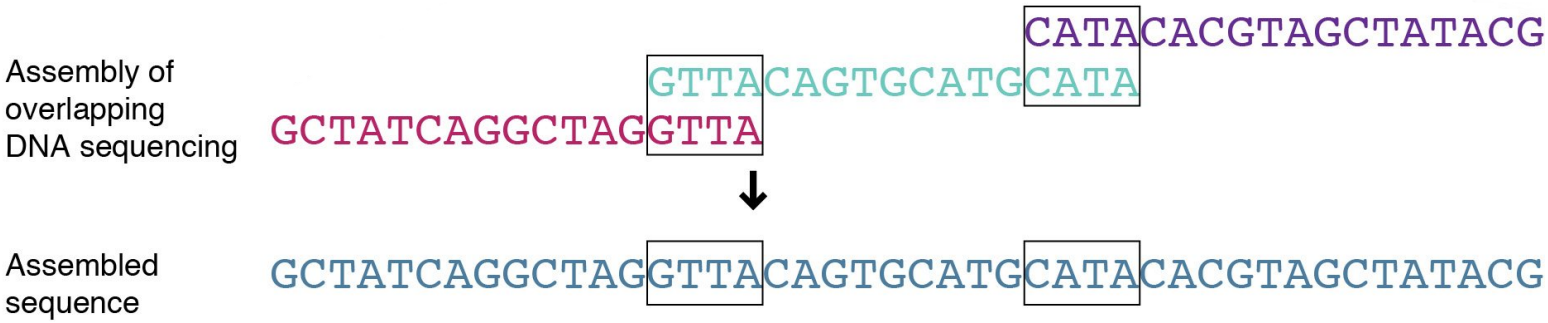
Obr. 2: Dvojzávitnica DNA.



Obr. 3: Štandardný postup sekvenácie.

- **Bázová abeceda:** abeceda $\Sigma_D = \{A, C, T, G\}$. Znaky abecedy Σ_D nazývame bázy.
- **Chromozóm:** konečné neprázdne slovo nad Σ_D.
- **Genóm:** konečný, neprázdny jazyk chromozómov.
- **Čítanie:** konečné, neprázdne podslovo chromozómu, môže obsahovať chyby.

Obr. 4: Príklad čítania (modrá) a všetkých jeho k-merov(červená).



Obr. 5: Skladanie troch čítaní do dlhšieho celku.

Skladanie:

- Veľká výpočtová zložitosť.
- Veľa informácií.

Optimalizácia:

- Vyššia rýchlosť.
- Niektoré približné hodnoty.

Články

- Williams et al.; BMC Genomics, 2013
- **Brejová, Hozza, Vinař; SPIRE, 2015**
- Vurture et al.; Bioinformatics, 2017

Existujúce modely

- **Error model (E):** Chyby v čítaniach
- **Repeats and errors model (RE):** Chyby v čítaniach, viac výskytov identických k-merov

Cieľ práce

- Preskúmanie vplyvu polymorfizmov na existujúce predikčné modely
- Rozšírenie predikčných modelov o polymorfizmus

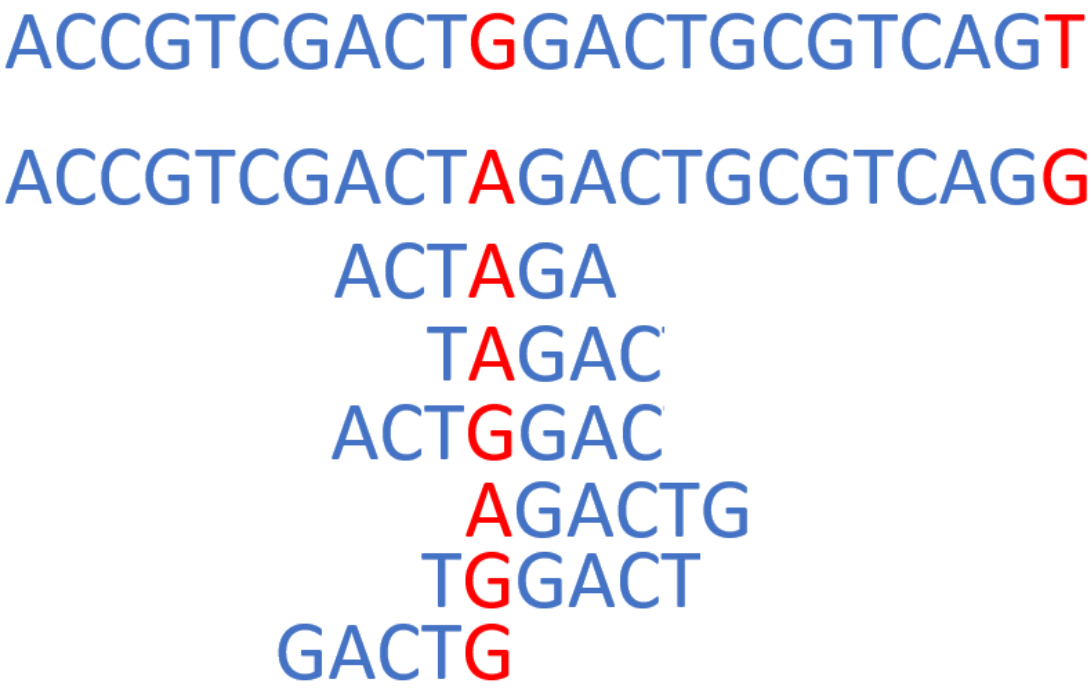
Polymorfné pozície v genóme

ACCGTCGACTGGACTGCGTCAGT

ACCGTCGACTAGACTGCGTCAGG

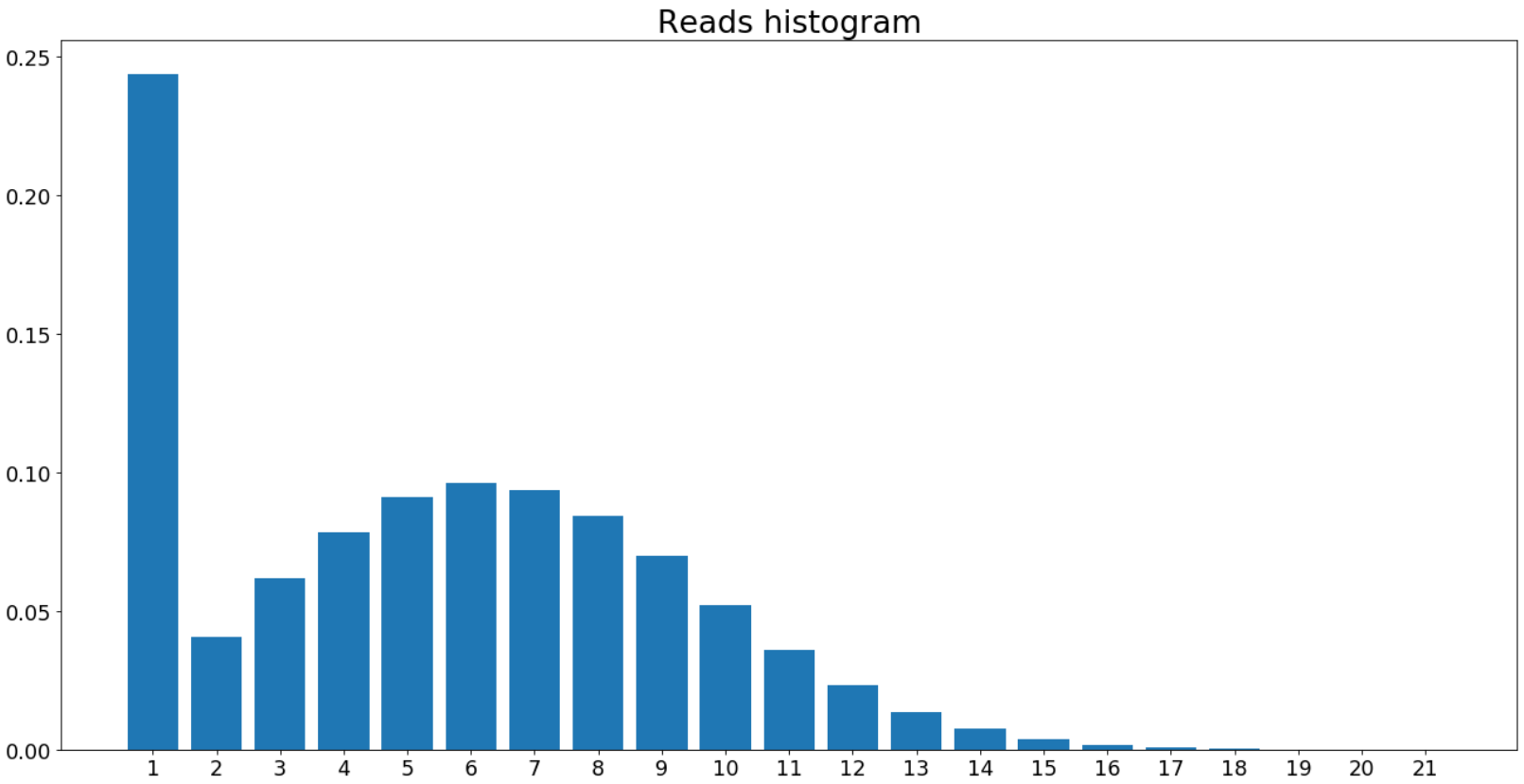
Obr. 6: Príklad polymorfizmu na dvoch pozíciách v časti genómu.

Polymorfné čítania



Obr. 7: Príklad polymorfných čítaní.

Vstupné dáta modelu



Obr. 8: Príklad vstupného histogramu množstva rôznych k-merov.

Optimalizované parametre

- Pokrytie C
- Chyby ϵ
- Polymorfné k-mery γ

EP model

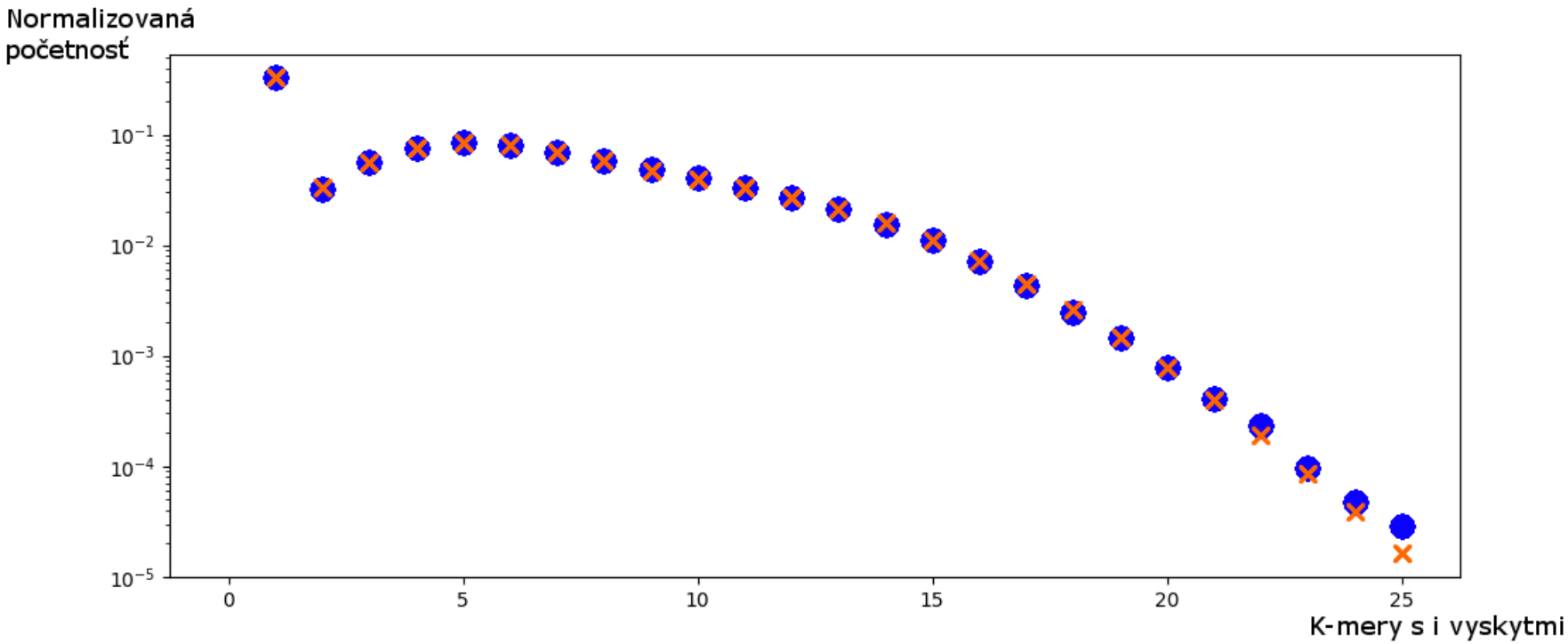
$$p_j = \sum_{s=0}^{k} \alpha_{s_n}((1-\gamma)f(j;\lambda_s)) + \alpha_{s_p}\gamma f(j;\frac{\lambda_s}{2}) \quad (1)$$

γ: pomer polymorfných k-merov ku všetkým k-merom v čítaniach

α_{s_n}, α_{s_p}: pravdepodobnosť výskytu s chýb v nepolymorfných a polymorfných k-meroch.

$\lambda_s = \epsilon^s(1-\epsilon)^{k-s}3^{-s}c_k$

Histogram z optimalizovaných parametrov



Obr. 9: Histogram z optimalizovaných parametrov. Modrá vstupné hodnoty, oranžová optimalizovaný odhad.

polymorphism rate	coverage accuracy in %								
0.064	59.92	60.02	60.17	60.50	61.06	62.26	64.96	70.69	
0.032	71.31	71.42	71.59	72.07	73.01	75.03	78.63	84.97	
0.016	81.86	82.04	82.25	82.61	83.52	85.25	88.34	93.08	
0.008	89.59	89.68	89.85	90.17	90.79	92.02	93.97	96.45	
0.004	94.37	94.46	94.54	94.71	95.13	95.86	96.98	97.80	
0.002	97.06	97.07	97.18	97.27	97.46	97.89	98.49	99.56	
0.001	98.48	98.52	98.58	98.66	98.72	98.98	99.14	99.65	
0	99.92	99.98	99.99	99.99	99.98	99.98	99.87	99.86	
	0	0.001	0.002	0.004	0.008	0.016	0.032	0.064	error rate

Obr. 10: Presnosť odhadu C E modelom pri rôznych nastaveniach parametrov a pokrytí 8.

polymorphism rate	coverage accuracy in %								
0.064	99.82	99.77	99.99	99.92	99.97	99.92	99.18	82.78	
0.032	99.65	99.83	99.79	99.40	99.58	99.91	97.24	97.83	
0.016	99.71	99.92	99.79	99.66	99.93	99.85	99.35	93.70	
0.008	99.94	99.93	99.90	99.78	99.88	99.77	99.63	95.27	
0.004	99.84	99.93	99.97	99.88	99.39	99.91	99.84	96.31	
0.002	98.60	99.89	99.90	99.99	98.45	99.21	99.91	96.88	
0.001	99.37	99.92	99.94	99.82	99.62	96.37	99.65	97.82	
0	97.46	99.82	99.92	99.86	98.23	97.87	98.89	96.10	
	0	0.001	0.002	0.004	0.008	0.016	0.032	0.064	error rate

Obr. 11: Presnosť odhadu C EP modelom pri rôznych nastaveniach parametrov a pokrytí 8.

polymorphism rate	coverage accuracy in %								
0.064	57.42	57.48	57.52	57.65	58.04	58.83	60.82	65.91	
0.032	67.36	67.46	67.58	67.94	68.79	70.59	73.94	80.06	
0.016	81.88	82.05	82.23	82.69	83.58	84.10	85.80	89.94	
0.008	89.56	89.70	89.85	90.19	90.85	90.65	92.72	95.05	
0.004	94.38	94.44	94.55	94.75	95.14	95.55	96.38	97.76	
0.002	97.05	97.11	97.18	97.29	97.51	97.89	98.18	98.82	
0.001	98.51	98.52	98.56	98.64	98.72	98.93	99.12	99.42	
0	99.99	100.00	100.00	99.99	100.00	99.96	99.98	99.95	
	0	0.001	0.002	0.004	0.008	0.016	0.032	0.064	error rate

Obr. 12: Presnosť odhadu C E modelom pri rôznych nastaveniach parametrov a pokrytí 16.

polymorphism rate	coverage accuracy in %								
0.064	99.99	99.90	99.97	99.93	99.90	99.93	99.90	99.80	
0.032	99.95	99.87	99.98	99.93	99.93	99.95	98.79	98.25	
0.016	99.92	99.89	99.97	99.96	99.91	99.85	99.48	99.83	
0.008	99.99	99.89	99.95	100.00	99.94	99.80	99.63	99.41	
0.004	99.93	99.96	99.97	99.96	99.91	99.94	99.90	99.28	
0.002	98.96	100.00	99.88	100.00	99.18	99.99	99.97	99.91	
0.001	98.28	99.98	99.96	99.87	99.05	97.47	99.99	98.46	
0	99.94	99.97	97.75	99.97	99.96	99.91	99.98	99.89	
	0	0.001	0.002	0.004	0.008	0.016	0.032	0.064	error rate

Obr. 13: Presnosť odhadu C EP modelom pri rôznych nastaveniach parametrov a pokrytí 16.

Extended RE model

RE model

- opakovania k-merov v genóme - copy number
- tri copy number parametre: β_0, β_1, β_2
- vyššie copy number hodnoty z geometrického rozdelenia

REP model

- d nových copy number parametrov

Extended RE model - výsledky

- zanedbatelné zlepšenie odhadu pri ERE modeli
- odhalená chybná interpretácia copy number parametrov
  - $\beta_{R,o}$: copy number v čítaniach
- navrhnutý model pre skutočné copy number v genóme
  - $\beta_{G,o}$: copy number v genóme

repeats	**k-mers**	**real** $\beta_{G,o}$	**estimated** $\beta_{G,o}$	**real** $\beta_{R,o}$	**estimated** $\beta_{R,o}$
1	4999970	0.8333	0.8327	0.7872	0.8174
2	999985	0.1667	0.1673	0.2128	0.1826

Tabuľka 1: Predikované $\beta_{G,o}$, dĺžka genómu 7 miliónov báz, $C = 32$, $\epsilon = 0.1\%$.

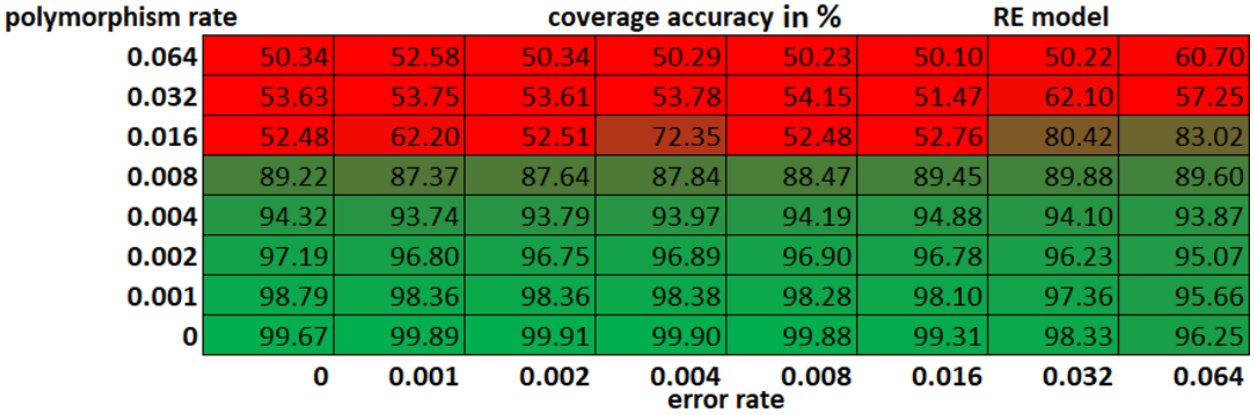
Equal Repeats for Non-polymorphism and Polymorphism and Error

- pokrytie C
- chybovosť ϵ
- pomer polymorfných k-merov γ
- tri copy number parametre: $\beta_{R,0}$, $\beta_{R,1}$, $\beta_{R,2}$

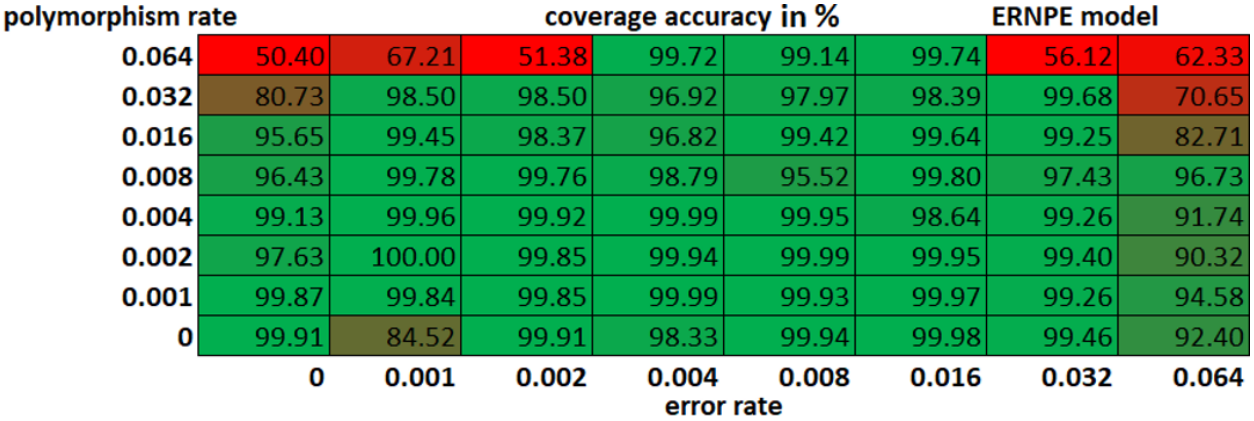
- Ak je k-mer polymorfný, všetky opakovania sú polymorfné.
  - Silný predpoklad.
  - Zjednodušuje predikcie.

$$p_j = (1-\gamma)\sum_{o=0}^{\infty}\beta_o p_{n,o,j} + \gamma\sum_{o=0}^{\infty}\beta_o p_{p,o,j} \quad (2)$$

kde $p_{n,o,j} = E(o \cdot C, \epsilon)$ a $p_{p,o,j} = E(o \cdot C/2, \epsilon)$



Obr. 14: Presnosť odhadu C RE modelom pri rôznych nastaveniach parametrov a pokrytí 16.

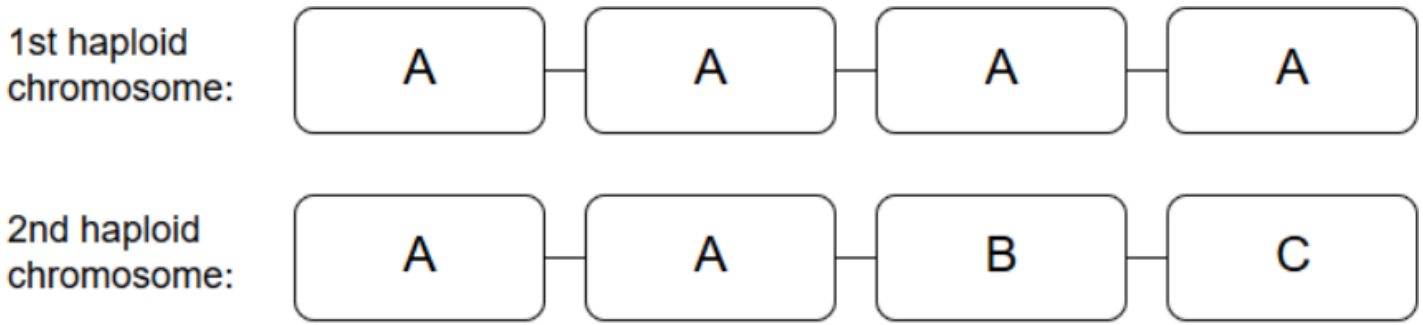


Obr. 15: Presnosť odhadu C ERNPE modelom pri rôznych nastaveniach parametrov a pokrytí 16.

Repeats, Errors and Polymorphisms

- pokrytie C
- chybovosť ϵ
- tri copy number parametre: $\beta_{G,0}$, $\beta_{G,1}$, $\beta_{G,2}$

- Nie všetky opakovania k-meru v genóme sú polymorfné.
- Silný biologický základ.



Obr. 16: Príklad k-meru s copy number 5 v haploide. Dve kópie sú polymorfné.

Výsledky na reálnych dátach

- Diploidný červ Caenorhabditis elegans, kmeň JU258 z Madeiry
- Illumina sekvenátor

Model	Pokrytie	Dĺžka genómu	Miera chybovosti	Miera polymorfizmu
Reálne	33.3159	100 286 401	0.0011%	2.428%
RE	19.7410	179 025 988	0.0012%	N/A
ERNPE	37.03	95 440 588	0.0012%	N/A
GenomeScope	16.75	98 897 045	0.122%	0.178%

Tabuľka 2: Výsledky rôznych modelov na Caenorhabditis elegans.

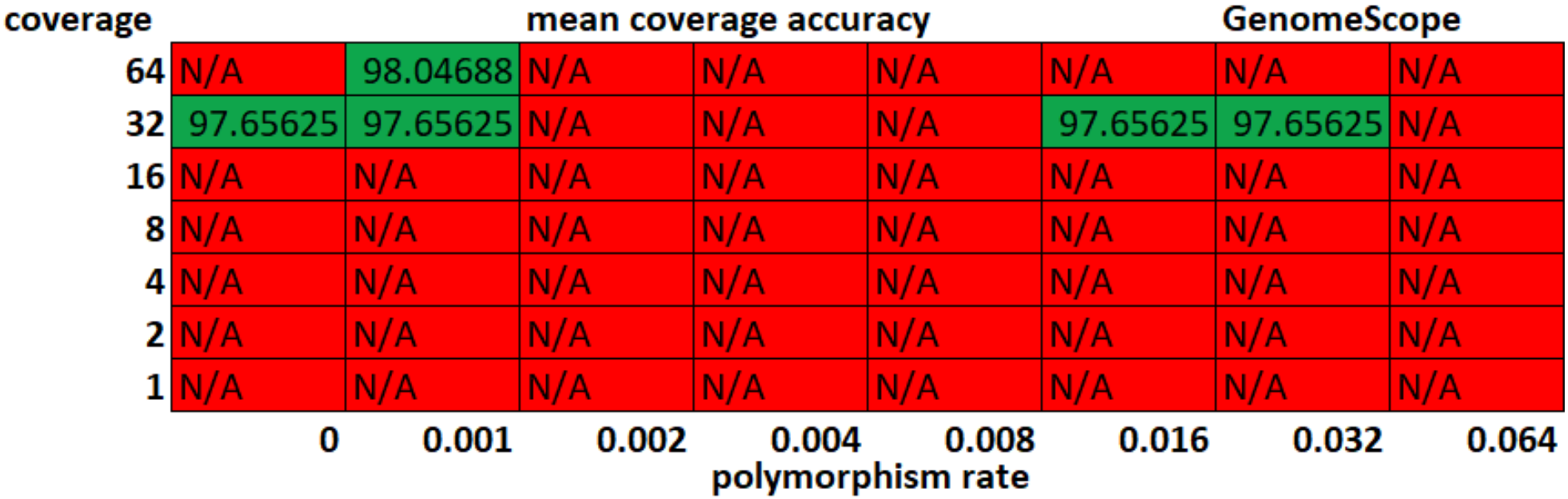
Záver

- Preskúmanie vplyvu polymorfných sekvenačných dát na existujúce predikčné modely.
- Vytvorenie EP modelu.
- Vytvorenie ERE modelu.
- Odhalenie chyby v interpretácií copy number parametrov.
  - Ovplyvňuje už publikované články.
  - Navrhnutý postup na získanie hodnôt pre pôvodnú interpretáciu.
- Vytvorenie ERNPE modelu.
- Navrhnutie REP modelu.

Možné rozšírenia

- Preskúmanie REP modelu.
- GC skreslenie.
- Metagenomické odhady.

Ďakujem za pozornosť



Obr. 17: Presnosť odhadu C GenomeScope modelom pri rôznych nastaveniach parametrov.

$$X_{p_a} = \gamma W_d \sum_{i=1}^{\infty} Z_i \cdot i \tag{3}$$

kde:

- X_{p_a}: počet všetkých polymorfných k-merov v čítaniach
- W_d: počet rôznych k-merov v čítaniach
- Z_i: pravdepodobnosť množstva k-merov s i opakovaniami z $E(C/2, \epsilon)$
- γ: pomer rôznych polymorfných k-merov ku všetkým rôznym k-merom v čítaniach