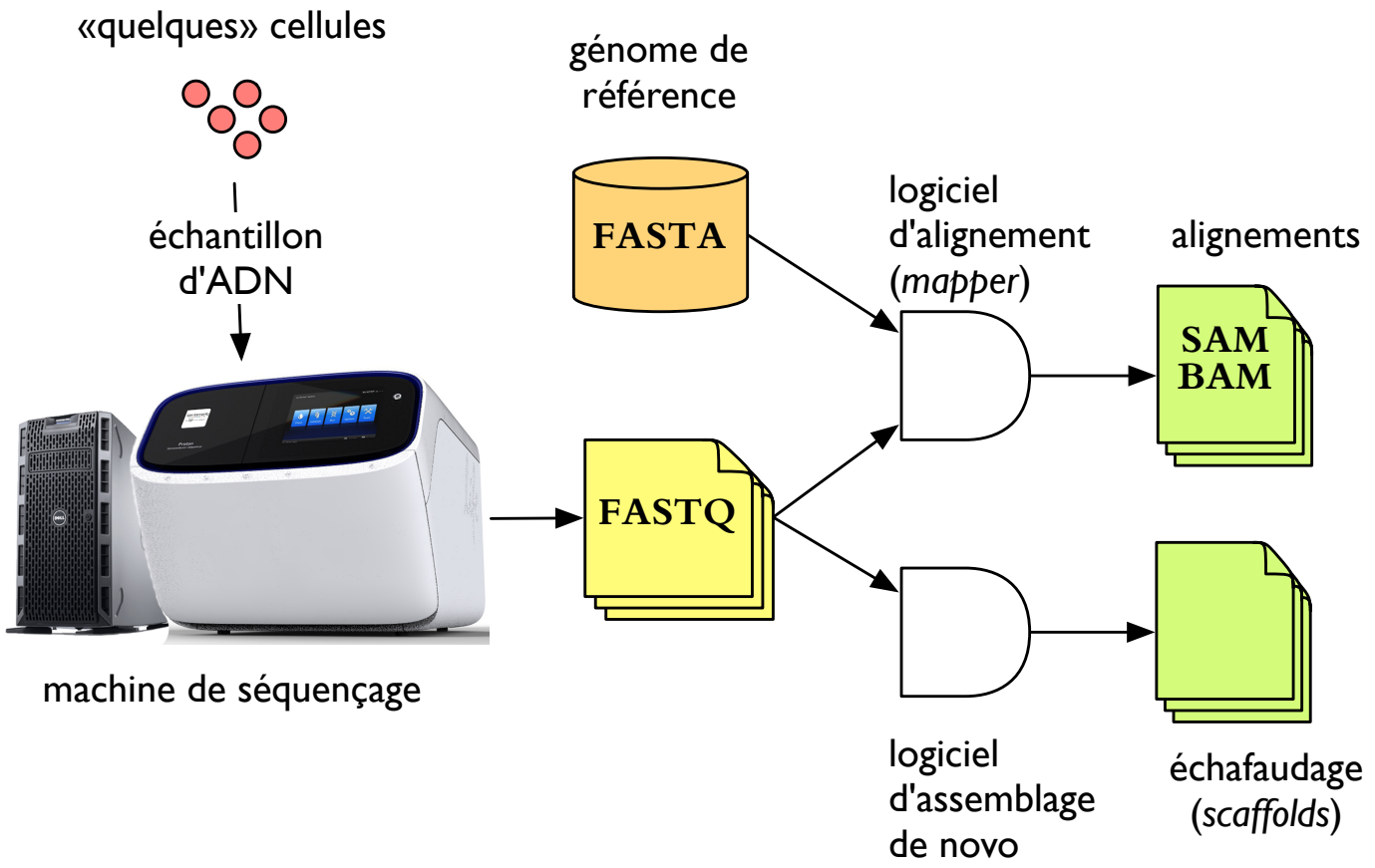
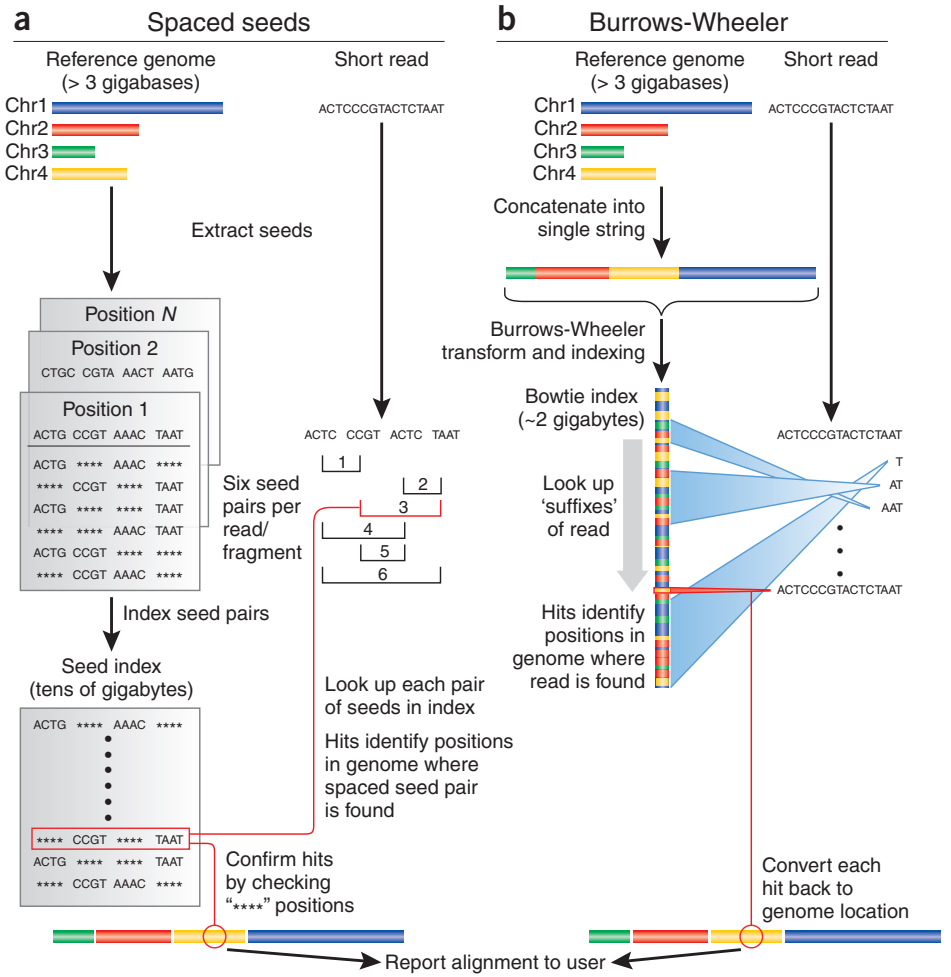
Variantes génomiques

Traitement de lectures



(machine Ion Proton)

Principes algorithmiques



Trapnell & Salzberg *Nature Biotechnology* 27 :455 (2009)

Format SAM

```
Coor     12345678901234  5678901234567890123456789012345
ref1     AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r004                    ATAGCT..............TCAGC
-r001/2                                      CAGCGGCAT
```

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref1 LN:45
r001  163 ref1  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref1  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r004    0 ref1 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r001   83 ref1 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

En-tête : @HD, @SQ (références avec nom SN et longueur LN), . . .

Colonnes : (1) QNAME [lecture], (2) FLAG, (3) RNAME [référence], (4) POS, (5) MAPQ [*mapping quality*], (6) CIGAR [alignement], (7–9) info sur paires, (10) SEQ (séquence de l'amorce), (11) QUAL (qualité de SEQ)

BAM : même information, avec compression

Chaîne CIGAR

Encodage : $l_1\,\mathsf{op}_1\,l_2\,\mathsf{op}_2\,l_3\,\mathsf{op}_3\cdots$ avec l_k la longueur d'opération op_k. Opérations :

Op	Description
`M`	alignment match (can be a sequence match or mismatch)
`I`	insertion to the reference
`D`	deletion from the reference
`N`	skipped region from the reference
`S`	soft clipping (clipped sequences present in SEQ)
`H`	hard clipping (clipped sequences NOT present in SEQ)
`=`	sequence match
`X`	sequence mismatch

Bits de FLAG :

Bit	Description
0x1	template having multiple fragments in sequencing
0x2	each fragment properly aligned according to the aligner
0x4	fragment unmapped
0x8	next fragment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next fragment in the template being reversed
0x40	the first fragment in the template
0x80	the last fragment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

Alignement de lectures : pondération

Dépendances : $X \to Y \to Z$

X : référence, Y : variante séquencée, Z : séquence lue

échelle Phred — base z lue avec qualité $q = -10 \log_{10} \mathbb{P}\{Y \neq Z\}$

$$p(z|y) = \mathbb{P}\Big\{ Z = z \;\Big|\; Y = y \Big\} = \begin{cases} 1 - 10^{-q/10} & \text{si } z = y \\ \frac{1}{3} 10^{-q/10} & \text{si } z \neq y \end{cases}$$

Pour un alignement à position t, on assume indépendance :

$$\mathbb{P}\Big\{ Z_{1..\ell} = z_{1..\ell} \;\Big|\; Y_{1..\ell} = X_{t..t+\ell-1} \Big\} = \prod_{i=1}^{\ell} p\Big(z_i | x_{t+i-1}\Big)$$

Pondération LODS pour aligner z_i et $y_i = x_{t+i-1}$ ...

Erreur de placement (MAPQ)

[approche de MAQ]

On veut trouver T, la vraie position de $z_{1..\ell}$. L'aligneur choisit la position t qui maximise la probabilité $\mathbb{P}\left\{ Z_{1..\ell} = z_{1..\ell} \;\middle|\; Y_{1..\ell} = X_{t..t+\ell-1} \right\}$. Supposons que Ω dénote les positions que le mappeur a examiné pour l'alignement (*hits*) :

$$\begin{aligned}
\epsilon = \mathbb{P}\{T \neq t\} = & \underbrace{\mathbb{P}\{T = 0\}}_{\epsilon_1} && \text{mauvaise référence} \\
& + \underbrace{\mathbb{P}\{T > 0, T \neq t, T \notin \Omega\}}_{\epsilon_2(1-\epsilon_1)} && \text{position manquée} \\
& + \underbrace{\mathbb{P}\{T > 0, T \neq t, T \in \Omega\}}_{\epsilon_3(1-\epsilon_2)(1-\epsilon_1)} && \text{mauvais } \textit{hit}
\end{aligned}$$

Comme $\epsilon_i \approx 0$, on a $\epsilon \approx \max\{\epsilon_1, \epsilon_2, \epsilon_3\}$. En plus, on peut ignorer $\epsilon_1 << \epsilon_2, \epsilon_3$ (contamination ne s'aligne pas bien — filtrer par alignement à modèle iid)

Erreur de placement II

Position manquée ? On veut $\epsilon_2 = \mathbb{P}\left\{T \notin \Omega \;\middle|\; T > 0\right\}$. On ne trouve pas la vraie position :

- (1) l'alignement à la vraie position (T) est assez similaire au meilleur alignement trouvé (t),
- (2) mais on ne l'examine pas ($T \notin \Omega$) : à T il y a trop d'erreurs

(1) régions similaires dans le génome + erreurs : on a une position T choisie au hasard qui est lue avec autant d'erreurs que la lecture devient plus similaire à une autre région (t)

$\Rightarrow \epsilon_2$ est dominant quand on cherche au seuil de sensitivité du hachage (v. graines espacées), avec des lectures corrompues dans des régions répétitives

(on peut l'estimer . . .)

Erreur de placement III

Autres candidats pour le placement ? On veut $\epsilon_3 = \mathbb{P}\Big\{T \neq t \;\Big|\; T \in \Omega\Big\}$. On considère donc la probabilité postérieure du placement dans Ω :

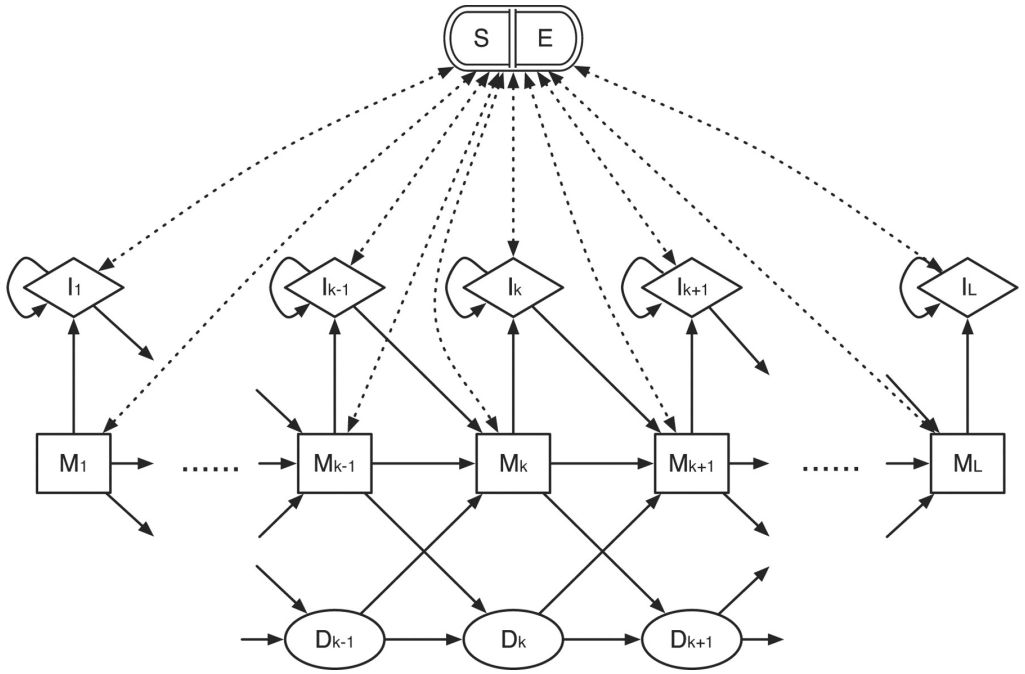
$$p_3(t) = \mathbb{P}\Big\{T = t \;\Big|\; T \in \Omega\Big\} = \frac{p\Big(z_{1..\ell} \;\Big|\; x_{t..t+\ell-1}\Big)}{\sum_{u\in\Omega} p\Big(z_{1..\ell} \;\Big|\; x_{u..u+\ell-1}\Big)}$$

(à priori $\mathbb{P}\{T = u\} = 1/|\Omega|$ pour tout $u \in \Omega$).

Il suffit de stocker les meilleurs alignements (MAQ : 1 meilleur, n 2e meilleur)

Qualité de bases alignées (QUAL)

SAMtools Base Alignment Quality (BAQ)



Génome de longueur L, lecture de longueur ℓ États : M, I, D, S, E

$$(a_{ij})_{5\times 5} = \begin{pmatrix} (1-2\alpha)(1-\gamma) & \alpha(1-\gamma) & \alpha(1-\gamma) & 0 & \gamma \\ (1-\beta)(1-\gamma) & \beta(1-\gamma) & 0 & 0 & \gamma \\ 1-\beta & 0 & \beta & 0 & 0 \\ (1-\alpha)/L & \alpha/L & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Émissions : $1/4$ dans I, selon qual en M, aucune en D,S, E

BAQ : probabilité postérieure de manque de homologie $1 - p(z_i \diamond x_t)$

Li *Bioinformatics* 27 :1157 (2011)

Variantes

Alignement de lectures dans la même position

1. réalignement (autour de trous)

2. inférence de génotypes (hétéro- et homozygotes ?) / haplotypes

Méthodes : (a) compter les votes, (b) inclure modèle d'erreurs et de mutation

Variant Call Format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

En-tête : information sur provenance, méthodes, explication de codes : filtres (FILTER), INFO et FORMAT

INFO : information sur le site et l'ensemble d'échantillons ; FORMAT montre les champs de génotypage dans les colonnes suivantes ; allèles encodés par 0,1,. . . (0=réf)