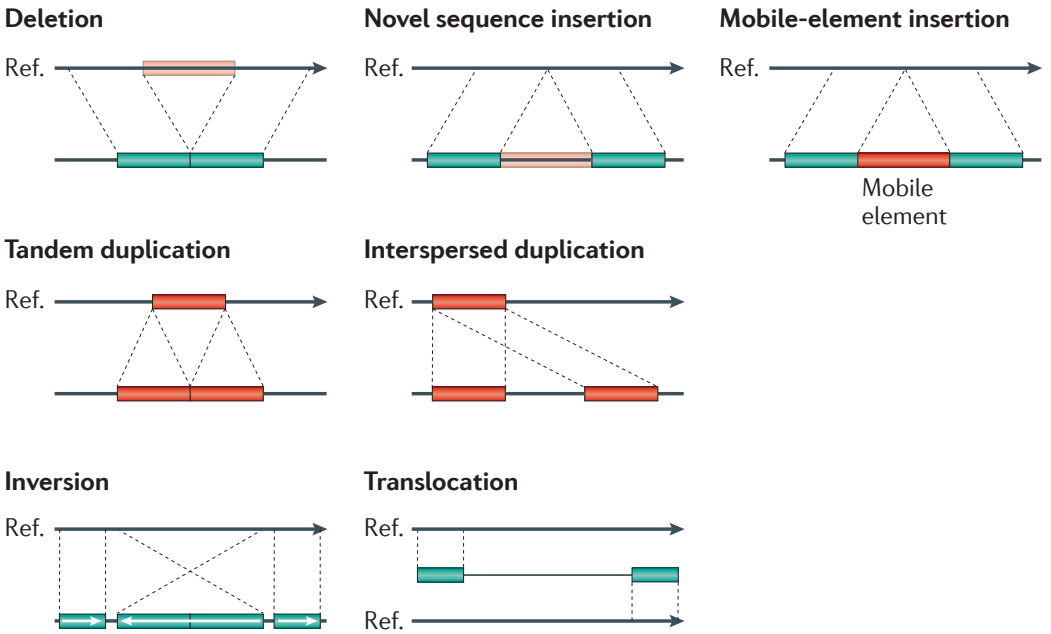
Variantes génomiques II

Variantes génomiques

- ⋆ polymorphisme d'un seul nucléotide (SNP=*Single Nucleotide Polymorphisms*), petits indels
- ⋆ variation structurale



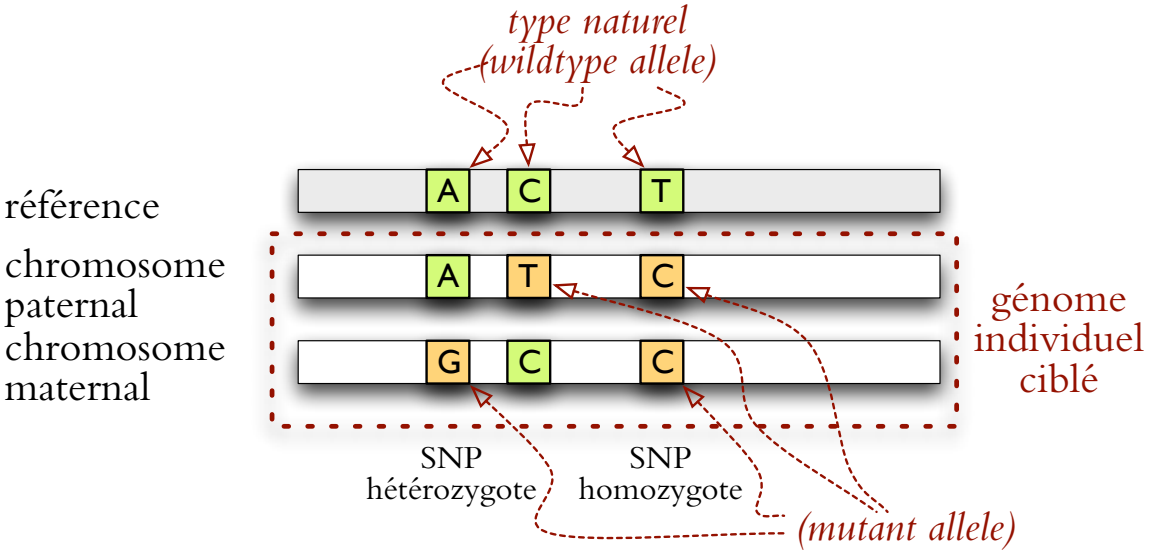
- ⋆ *copy number variation*
  (=duplication de $>$ 50pb, incluant chromosomes complets)

Alkan, Coe & Eichler *Nature Reviews Genetics* 12 :363 (2011)

Méthodes

⋆ CGH (comparative genomic hybridization) sur puce ⇒ copy number variation
⋆ puces SNP
⋆ séquençage

SNPs



Unphased

```
##fileformat=VCFv4.2
...
... REF ALT ... FORMAT smp
...   A   G ... GT     0/1
...   C   T ... GT     0/1
...   T   C ... GT     1/1
```

Phased

```
##fileformat=VCFv4.2
...
... REF ALT ... FORMAT smp
...   A   G ... GT     0|1
...   C   T ... GT     1|0
...   T   C ... GT     1|1
```

Variant calling

bases appellées dans une position : $\mathcal{Z} = \{(z_1, q_1), (z_2, q_2), \ldots, (z_n, q_n)\}$

génotypes possibles (non-ordonnés) : 4 homozygotes, 6 hétérozygotes

génotype inconnu : Y

$$\mathbb{P}\left\{Y = y_1 y_2 \,\middle|\, \mathcal{Z}\right\} \propto \underbrace{\mathbb{P}\left\{\mathcal{Z} \,\middle|\, Y = y_1 y_2\right\}}_{\text{vraisemblance de } y_1 y_2} \times \underbrace{\mathbb{P}\left\{y_1 y_2\right\}}_{\text{prob. de génotype } y_1 y_2}$$

calculer $\mathbb{P}\left\{\mathcal{Z} \,\middle|\, Y = y_1 y_2\right\}$ pour homozygotes ($y_1 = y_2$) ou hétérozygotes ($y_1 \neq y_2$) $\ldots$

Génotypes : Hardy-Weinberg

équilibre Hardy-Weinberg :

⋆ population infinie

⋆ générations discrètes

⋆ panmixie

⋆ $\emptyset$ mutations, immigration, sélection

fréquence d'allèles : type naturel (A) avec p, mutant (a) avec $q = 1 - p$

fréquence de génotypes diploïdes : $\mathsf{AA} \sim p^2$, $\mathsf{Aa} \sim 2pq$, $\mathsf{aa} \sim q^2$

reste constante . . .

Fréquences de SNPs

minor allele frequency (MAF) : fréquence d'allèle mutant dans la population (q)

SNPs fréquents ($\mathsf{MAF} > 10\%$) et rares ($\mathsf{MAF} < 5\%$)

HAPMAP : puces SNP (phase I 2005, phase II 2009)

1000 Genomes

populations : CEU (Amérique du Nord, ancêtres européens), YRI (Yoruba, Nigeria), JPT (Japon), CHB (Han) ; ASW (afro-américains), GIH (Gujarati de Houston), MEX (mexicains de Los Angeles), LWK (Luhya, Kenya), ...

$\Rightarrow$ MAF spécifique aux populations

dbSNP : base de données sur fréquence de SNPs

Haplotypage — phasing

⋆ séquençage directe de haplotypes (très cher)
⋆ à partir des lectures (distribuer les lectures à deux haplotypes)
⋆ par héritage (séquencer les parents)
⋆ par haplotypes de référence

Haplotypage — lectures

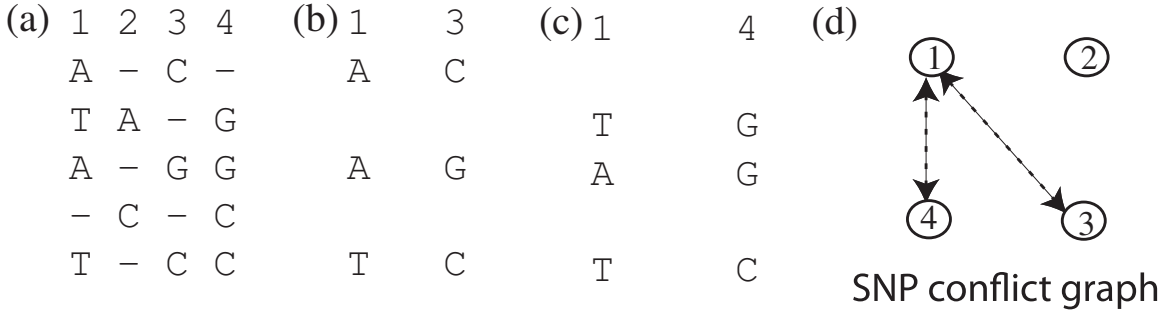


FIG. 3. *The SNP conflict graph. (a) A hypothetical data set consisting of five fragments sequenced at four sites. (b) Highlight of a SNP conflict between columns 1 and 3 of (a), using rows 1, 3, and 5. (c) Highlight of a SNP conflict between columns 1 and 4 of (a) using rows 2, 3, and 5. (d) The SNP conflict graph for (a), with edges corresponding to the conflicts shown in (b) and (c).*

on veut un graphe biparti — formulations différentes

Schwartz 2010

Haplotype assembly

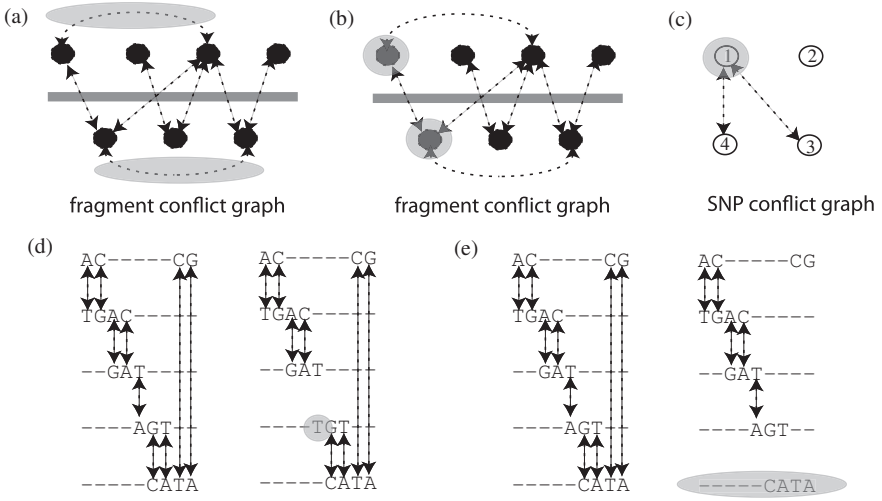


FIG. 4. *Illustration of haplotype assembly problem variants. (a) Minimum edge removal (MER). (b) Minimum fragment removal (MFR). (c) Minimum SNP removal (MSR). (d) Minimum error correction (MEC). (e) Longest fragment reconstruction (LFR).*

problèmes NP-difficiles . . .

Haplotypes-héritage

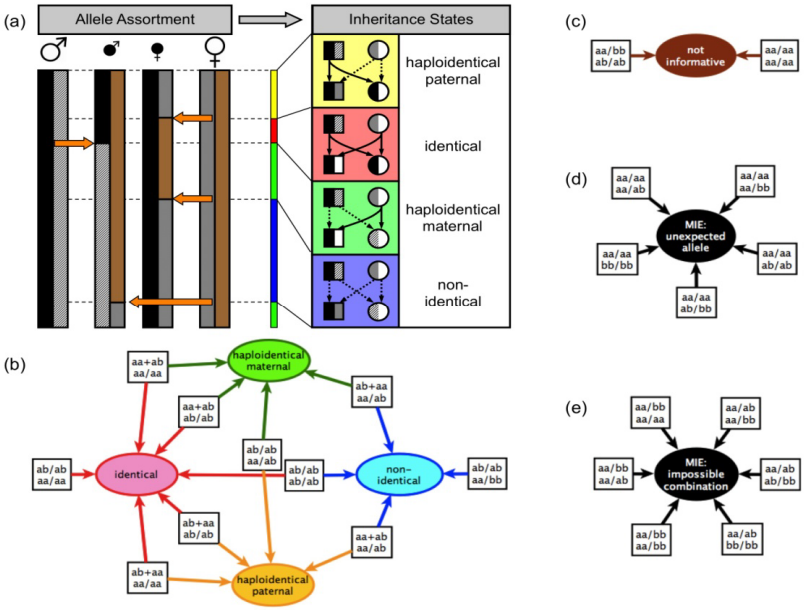
recombinaisons en meïose chez père ou mère — exemple : quartet (2 parents, 2 enfants)

enfants : identiques, haploidentiques (mère ou père seulement), ou différents ;

génotypage joint avec HMM : 4 états de héritage + 2 états d'erreur (compression/CNV et erreurs) ;

émissions : erreur avec 0.5% (ou 30% dans l'état d'erreur), sinon respecter règles ; compression : hétérozygotes fréquentes

Figure S2A



Roach et al. *Science* 328 :636 (2010)

Family quartet

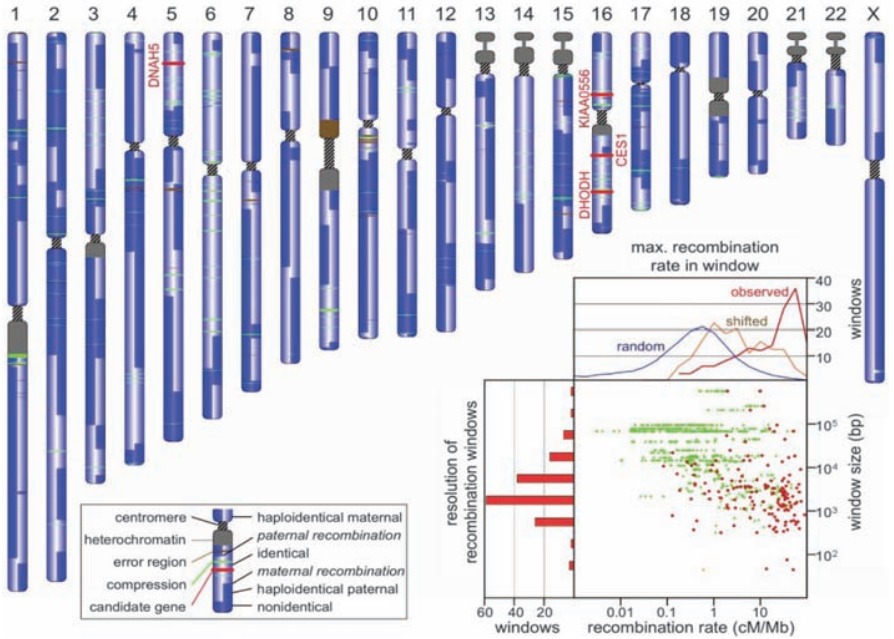
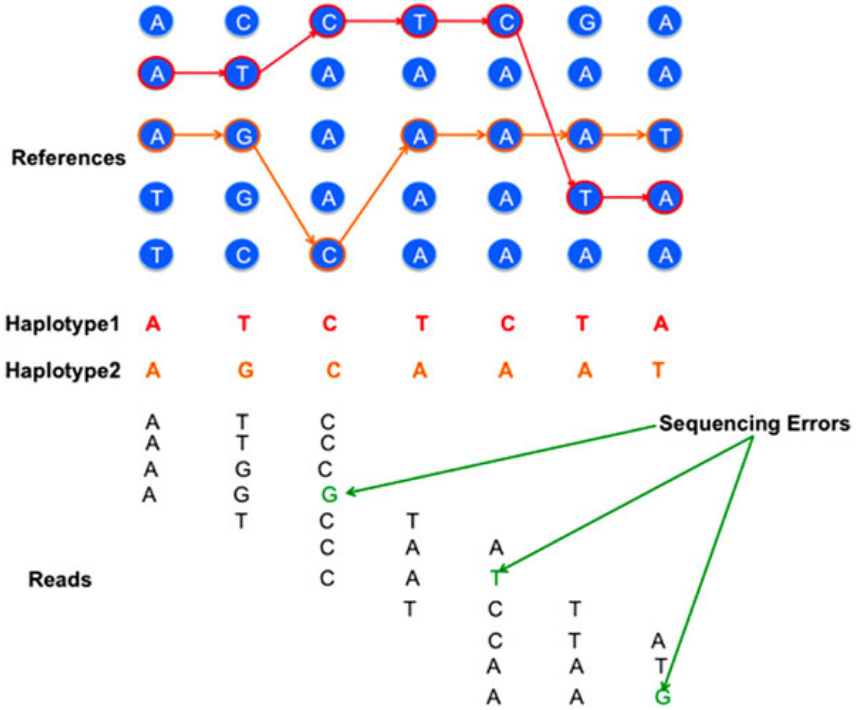


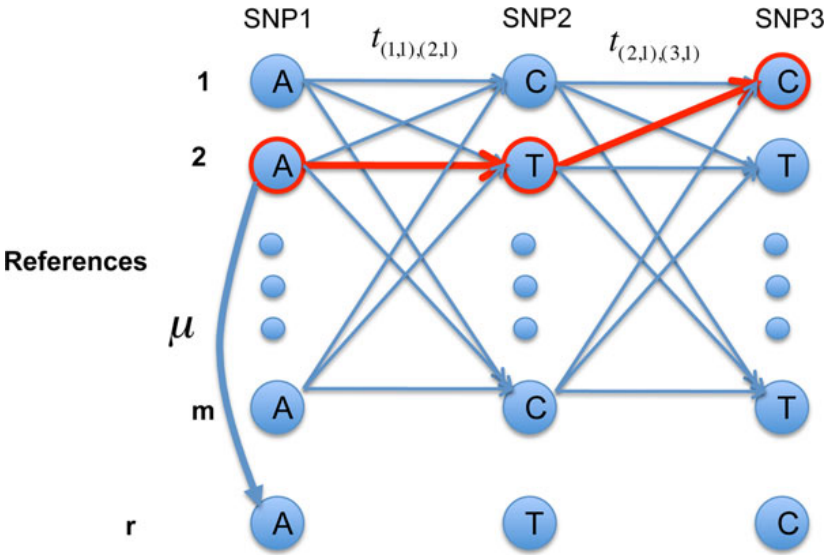
Fig. 1. The landscape of recombination. Each chromosome in this schematic karyotype is used to represent information abstracted from the four corresponding chromosomes of the two children in the pedigree. It is vertically split to indicate the inheritance state from the father (left half) and mother (right half), as shown in the key. The three compound heterozygous (*DHODH*, *DNAH5*, and *KIAA0556*) and one recessive (*CES1*) candidate gene, depicted by red bands, lie in "identical" blocks. (Inset) Scatterplot of HapMap recombination rates (in centimorgans per megabase) within the predicted crossover regions. The maximum value of centimorgans per megabase found in each window is shown in red. The left histogram shows the size distribution of recombination windows ($\log_{10}$ value of -0.58 ± 0.92). The top graph shows the centimorgans per megabase distribution for the observed maximal values (red), for similarly sized windows shifted by 6 kb (orange), and for similarly sized windows randomly chosen from the entire genome (blue). A shift of 6 kb from the observed locations eliminates the correlation with hotspots. Of 155 recombination windows, 92 contained a HapMap site with >10 cM/Mb. Only five randomly picked windows are expected to contain such high recombination rates.

Roach et al. *Science* 328 :636 (2010)

Imputation

HMM à partir des haplotypes connus dans la population :





He, Han & Eskin *J Comput Biol* 20 :80 (2013)

Imputation 2

HMM : n individus diploïdes, $2n$ haplotypes, états $(x, y) =$ indice des haplotypes identiq
transitions : constante ou incorporer fréquence de recombinaisons (LD = linkage disqeuilibrium)

émissions : même génotype que (x, y) ou permettre les mutations ?
aligner les lectures aux haplotypes (Hap-seq) ou génotypes indépendants (puces / MaCH) ?

calcul : MCMC ou forward-backward ou Viterbi