IDENTITY BY DESCENT
Consanguinité

entre deux individus reliés, on copte sur l’identité de haplotypes par descente

Figure 1

Identity by descent (IBD) on chromosome 1 for half-siblings and fifth cousins. Chromosome 1 is approximately 250 Mb long and has a genetic length of approximately 280 cM. The common ancestors’ copies of chromosome 1 are shown in various colors, and tan represents all other haplotypes. Regions of IBD are shown with black bars.

Shared ancestry are approximately exponentially distributed with a mean of 100 $m^{-1}$ cM.

Thus, for example, fifth cousins (see Figure 1) are separated by 12 meioses. On average, 0.05% of their genome, or approximately 1.5 cM ($\sim 1.5$ Mb), is identical by descent through their great-great-great-great grandmother. If they are full fifth cousins, they may also have IBD sharing through their great-great-great-great grandfather, which doubles the expected IBD proportion to 0.1% of their genome, or approximately 3 cM ($\sim 3$ Mb). However, fifth cousins usually have no detectable IBD sharing, and when they do have IBD sharing it is usually composed of a single IBD segment with a mean length of 8.3 cM ($\sim 8$ Mb). Extrapolating further, individuals who have shared ancestry through a certain common ancestor 25 generations ago (with 50 meioses of separation) almost always share none of their genome identical by descent through that ancestor, but if they do have an IBD segment through that ancestor it will have a mean length of 2 cM ($\sim 2$ Mb).

Any given pair of individuals is related through many common ancestors. For a pair of individuals on different continents, the relationships may be too distant to result in detectable IBD sharing. However, pairs of individuals from the same geographic region may have many recent common ancestors. Such individuals may, however, have only one or two detectable IBD segments, as many of the relationships have not resulted in any IBD sharing.

In a data set with $N$ unrelated individuals, there are $N(N-1)/2$ pairs of individuals. Although any given pair has very little IBD sharing, the total amount of IBD sharing in the sample, and the total amount per individual, can be large.
Segments IBD

centiMorgan : distance sur laquelle 0.01 recombinainesons occurrent en une génération
genome humain : $10^6 \text{pb} \approx 1 \text{cM}$

distance de $m$ générations : proportion de genome partagée $1/2^{m-1}$
longueur de segment : $100/m \text{cM}$

$\Rightarrow$ cousins de 5e degré : partagé $\frac{3000 \text{ Mb}}{2^{11}} \approx 1.5 \text{ cM}$ en moyenne venant de chaque grand parent, de longueur $8.3 \text{ cM}$ en moyenne ; aucun segment en commun avec $\geq 65\%$ probabilité :

$$\frac{2 \cdot 1.5}{3.5} = 0.35$$
Reine Victoria et Prince Albert

Csuros (2014)
Mode d’identité (Jacquard)

<table>
<thead>
<tr>
<th>IBD mode</th>
<th>A’s genotype</th>
<th>B’s genotype</th>
<th>identity coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x</td>
<td>x</td>
<td>(\Delta_1)</td>
</tr>
<tr>
<td>B</td>
<td>x</td>
<td>x</td>
<td>(\Delta_1)</td>
</tr>
<tr>
<td>A</td>
<td>x</td>
<td>x</td>
<td>(\Delta_1)</td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>y</td>
<td>(\Delta_1)</td>
</tr>
<tr>
<td>A</td>
<td>x</td>
<td>y</td>
<td>(\Delta_1)</td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>z</td>
<td>(\Delta_1)</td>
</tr>
<tr>
<td>A</td>
<td>x</td>
<td>y</td>
<td>(\Delta_1)</td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>z</td>
<td>(\Delta_1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IBD mode</th>
<th>A’s genotype</th>
<th>B’s genotype</th>
<th>identity coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x</td>
<td>y</td>
<td>(\Delta_2)</td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>y</td>
<td>(\Delta_2)</td>
</tr>
<tr>
<td>A</td>
<td>x</td>
<td>z</td>
<td>(\Delta_2)</td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>y</td>
<td>(\Delta_2)</td>
</tr>
<tr>
<td>A</td>
<td>x</td>
<td>y</td>
<td>(\Delta_2)</td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>z</td>
<td>(\Delta_2)</td>
</tr>
<tr>
<td>A</td>
<td>x</td>
<td>y</td>
<td>(\Delta_2)</td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>w</td>
<td>(\Delta_2)</td>
</tr>
</tbody>
</table>

The matrix in Eq. (5) is:

The system of equations (4) can be written in matrix form as:

\[ \begin{pmatrix} a & b & c & d & e & f & g & h & i \\ j & k & l & m & n & o & p & q & r \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} s \\ t \\ u \\ v \end{pmatrix} \]

By Equation (6),

\( (\bar{z} - \bar{x}) \cdot (\bar{y} - \bar{w}) = \bar{u} + \bar{v} \)

where \( \bar{z} \) and \( \bar{w} \) are the genotypes of the individuals.

Type distribution: let

\( \text{let } \mathbf{A} \text{ shares } \mathbf{B} \)
IBD → IBS

(Identity-by-descent et identity-by-state)

Table 2 | Joint genotypic probabilities

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Genotypic state</th>
<th>Number of shared alleles</th>
<th>General</th>
<th>Non-inbred</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $A_iA_i, A_iA_i$</td>
<td>Hom/hom</td>
<td>2</td>
<td>$\Delta_i P_i + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7) P_i^2 + (\Delta_4 + \Delta_6 + \Delta_8) P_i^3 + \Delta_9 P_i^4$</td>
<td>$k_2 P_i^2 + k_1 P_i^3 + k_0 P_i^4$</td>
</tr>
<tr>
<td>2 $A_iA_i, A_jA_j$</td>
<td>Hom/hom</td>
<td>0</td>
<td>$\Delta_2 P_j P_i + \Delta_4 P_j P_i^2 + \Delta_6 P_j P_i^3 + \Delta_8 P_j P_i^4$</td>
<td>$k_0 P_j P_i^2$</td>
</tr>
<tr>
<td>3 $A_iA_i, A_jA_j$</td>
<td>Hom/het</td>
<td>1</td>
<td>$\Delta_3 P_j P_i + (2 \Delta_4 + \Delta_8) P_j P_i^2 + 2 \Delta_9 P_j P_i^3$</td>
<td>$k_1 P_j P_i^2 P_j + 2 k_0 P_j P_i^3 P_j$</td>
</tr>
<tr>
<td>4 $A_iA_i, A_jA_m$</td>
<td>Hom/het</td>
<td>0</td>
<td>$2 \Delta_4 P_j P_i P_m + 2 \Delta_9 P_j P_i^2 P_m$</td>
<td>$2 k_0 P_j P_i^2 P_m$</td>
</tr>
<tr>
<td>5 $A_iA_j, A_iA_j$</td>
<td>Het/het</td>
<td>2</td>
<td>$2 \Delta_7 P_j P_i P_j + \Delta_8 P_j P_i (P_i + P_j) + 4 \Delta_9 P_j P_i^2 P_j$</td>
<td>$2 k_2 P_j P_i + k_1 P_j P_i (P_i + P_j) + 4 k_0 P_j P_i^2 P_j$</td>
</tr>
<tr>
<td>6 $A_iA_j, A_iA_m$</td>
<td>Het/het</td>
<td>1</td>
<td>$\Delta_8 P_j P_i P_m + 4 \Delta_9 P_j^2 P_m P_i$</td>
<td>$k_1 P_j P_i P_m + 4 k_0 P_j P_i^2 P_m$</td>
</tr>
<tr>
<td>7 $A_iA_j, A_mA_i$</td>
<td>Het/het</td>
<td>0</td>
<td>$4 \Delta_9 P_j P_i P_m P_l$</td>
<td>$4 k_0 P_j P_i P_m P_l$</td>
</tr>
</tbody>
</table>

The table shows seven distinct patterns of genotypes that are possible for two unordered individuals, and the probabilities of these pairs of genotypes in general, or assuming no inbreeding. Two genotypes could be homozygous (hom) for the same or different alleles (rows 1 and 2), one could be homozygous and the other heterozygous (het) with one or zero shared alleles with the homozygote (rows 3 and 4), or both individuals could be heterozygous with two, one or zero shared alleles (rows 5–7). There are nine pairs of genotypes if the ordering of individuals is important (not shown), as the genotypes in rows 3 and 4 (one homozygote and one heterozygote) each have two orders. $k$, the probability of sharing $i$ number of alleles that are identical-by-descent (where $i = 0–2$; see also FIG. 1); $P$, allele frequency; $\Delta_i$, Jacquard coefficients, which are measures of identity-by-descent status (BOX 1; FIG. 1).

→ probabilité d’émission pour HMM avec états $\Delta_i$

Weir & al. (2012)
Détecter les segments IBD

Problèmes

I. On a 2 individus — on veut identifier le niveau exact de parenté : kcoeffs, CARROT

II. Identifier des liens de parenté dans une population : on a \( n \) génomes (diploïdes) — on veut identifier les paires d’individus avec segments IBD : fastIBD, GERMLINE, ...
Quantification

coefficient de consanguinité $\gamma$ (inbreeding coefficient) :
probabilité d’IBD entre les 2 allèles du même individu

modèle iid (allèle $A$ avec $p$, $a$ avec $q = 1 - p$) + consanguinité : proba de génotypes non-ordonnés

$$
\begin{align*}
\phi(Aa) &= 2(1 - \gamma)pq \\
\phi(AA) &= p^2 \left(1 + \frac{q}{p}\right) \\
\phi(aa) &= q^2 \left(1 + \frac{p}{q}\right)
\end{align*}
$$

avec $\gamma = 0$ : coefficients de Cotterman pour fréquences de IBD0, IBD1, IBD2
Méthode de Lee

homozygotes discordantes ($\mathcal{D}$) : (AA, aa) ou (aa, AA)
hétérozygotes concordantes ($\mathcal{C}$) : (Aa, Aa)

probabilités sans IBD : $d = \mathbb{P}(\mathcal{D}) = 2p^2q^2$, $c = \mathbb{P}(\mathcal{C}) = 4p^2q^2$

compter seulement les sites $\mathcal{C}$ ou $\mathcal{D}$ : $X_1, X_2, \ldots, X_n \in \{0, 1\}$ où $X_i = 1$ dénote des hétérozygotes concordantes

définir le compte $N_C = \sum_{i=1}^{n} X_i$ ; on a

$\mathbb{E} N_C = \frac{2n}{3}$ ; $\text{Var} N_C = \frac{2n}{9}$

$\Rightarrow$ test d’hypothèse : $\frac{N_C}{n} \sim \mathcal{N}\left(\frac{2}{3}, \frac{\sqrt{2}}{3n}\right)$ si indépendents

si IBD, alors $\mathbb{E} N_C > 2n/3$ ;
si populations différentes (aucun IBD, $p$ différentes), $\mathbb{E} N_C < 2n/3$

distribution jointe de $\text{IBS2}^*$ _ratio_ $\approx \frac{N_C}{N_C+N_D}$ et $\frac{N_C+N_D}{N_C+N_D+N_{\text{autres}}}$
Segments à fine échelle

HMM avec état [∅ consanguinité] comme vecteurs de 8 indicateurs :
\( m_C \in \{0, 1\} : E_1 \text{ et } F_1 \) héritent le même haplotype du parent \( C \) ; \( m_D \)
\( m_E \in \{0, 1\} : E_2 \) reçoit l’allèle de \( C \) ; \( m_F \) (reçoit de \( D \))
\( d_E = 0 \) si \( E_1 \rightarrow E_2 \cdots \rightarrow E_k \)
\( p_A \) : phase de \( A/E_{\text{gen}_A} \)

Transitions : par coordonnées, indépendamment!

Kyriazopoulou-Panagiotopoulou & al ISMB 2011
remaining subset. This process was repeated 10 times and the ability of CARROT to distinguish between relationship was known, so predictions were made only within each relationship. To additional relationships, we implemented a classifier that uses \textit{Max likelihood} relationship.

We refer to relationship types using the notation (mrcas allele frequencies to compute the emission probabilities (Epstein

\[
\begin{align*}
\text{Average accuracy over all the pairs of the corresponding relationships, the average accuracy was } 34.45\% \text{ (full results not shown).}
\end{align*}
\]

\[
\begin{align*}
\text{The value at row } i \text{ and column } j \text{ is the percentage of pairs of relationship } i \text{ that were predicted to be of relationship } j: (2, 0, 2): \text{ great aunt–niece}; (2, 1, 1): \text{ first cousins}; (2, 2, 0): \text{ great niece–aunt}; (1, -1, 2): \text{ great grandparent–grandchild}; (1, 0, 1): \text{ half aunt–niece}; \text{(1, 1, 0): \text{ great niece–aunt}; (1, 2, -1): \text{ great grandchild–grandparent.}
\end{align*}
\]

\textit{émissions selon qualité des bases de sequençage classification de segments (fenêtre) par probabilité postérieure d'}IBD

\[
\begin{array}{cccccccc}
\text{Degree} & \text{Relationship} & (\text{mrcas, gen}_A, \text{gen}_B) \\
1 & \text{Full siblings} & (2, 0, 0) \\
& \text{Parent–child} & (1, -1, 0) \\
& \text{Half siblings} & (1, 0, 0) \\
2 & \text{Aunt–niece} & (2, 0, 1) \\
& \text{Avuncular} & (2, 0, 1) \text{ or } (2, 1, 0) \\
& \text{Grandparent–grandchild} & (1, -1, 1) \\
3 & \text{First cousins} & (2, 1, 1) \\
& \text{Great grandparent–grandchild} & (1, -1, 2) \\
4 & \text{Great aunt–niece} & (2, 0, 2) \\
& \text{Half aunt–niece} & (1, 0, 1) \\
& \text{Half first cousins} & (1, 1, 1) \\
\end{array}
\]

\[
\begin{align*}
\text{Table 2. Transition type frequencies (estimated from a training set of genotyped pairs).}
\end{align*}
\]
Segments IBD dans une population

(méthode de GERMLINE : segments longues)

On a des haplotypes [phases !] : matrice binaire $H[1..2n][1..s]$ avec haplotypes sur $s$ sites dans $n$ génomes

segment IBD : $H[i][j..j'] = H[i'][j..j']$

But : identifier les segments IBD les plus longs

1. groupage de haplotypes identiques (tableau de hachage) dans bloc

2. fusion de paires dans blocs consécutifs ; maintenir début de segment pour la paire

paires de haplotypes $(i, i')$ avec bloc $(k - 1)$ IBD : soit extension à bloc $k$ (avec erreurs permises), soit tester longueur (dépasse $L_{\text{min}}$ ?)

Gusev & al Genome Research 19 :318 (2009)
Segments IBD dans une population 2

Beagle et fastIBD : modèle LD (linkage disequilibrium) pour haplotypes

HMM : états = allèle 0/1 dans le haplotype ; transitions par fréquence dans les haplotypes ; émissions incluent erreur de séquençage

Figure 1. Example of an LD Model on Four SNPs
SNP 1 is represented by edges \( e_A \) and \( e_B \), SNP 2 by edges \( e_D, e_E, e_F \); SNP 3 by edges \( e_D, e_G, e_H \); and SNP 4 by edges \( e_D, e_E, e_F, e_G, e_H \). For each SNP, allele 1 is represented by a solid line, whereas allele 2 is represented by a dashed line. Haplotype H1 (1 1 1 1) follows the orange path \((e_A, e_B, e_E, e_I)\), and haplotype H2 (2 1 1 1) follows the blue path \((e_B, e_E, e_D, e_H)\).
To process the data in an efficient manner, we use a slight variant of the algorithm of Browning [2006], which is adapted from that of Ron et al. [1998]. The algorithm is the same as that in Browning [2006] except that the merging threshold is changed as described below. This algorithm processes the markers in chromosomal order, starting with all haplotypes in a single node, and in turn splitting the nodes by considering the alleles at the next marker, and then merging nodes based on the Markov criterion, as illustrated in Figure 1. To implement the Markov criterion in the merging procedure, a score is calculated for each pair of nodes at the current level. The score is described in detail in Browning [2006]. A low score for a pair of nodes means that the two nodes (i.e. haplotype clusters) have similar probabilities for sequences of alleles at markers $t_1, t_2, y, t_k$, for all $k$. A pair of nodes may be merged if the score is less than a threshold $m(n/C_0 + n/C_0)y^{-1} = 2 + b$.

In this formula, $m$ and $b$ are scale and shift parameters, respectively, that we have introduced because simulation studies (unpublished data) demonstrated increased power using a more parsimonious model with fewer clusters at each position (this is achieved by increasing the threshold). In the results presented here we use $m = 4$ and $b = 0.2$, which gives good power over a range of conditions. These parameters were not used in Browning [2006] (effectively $m = 1$ and $b = 0$ were used).

The outcome of this procedure is a directed acyclic graph. Edges of the graph are labeled by alleles, and each haplotype in the sample traces a path through the graph from the root node to the terminal node, following the sequence of alleles. Edges of the graph correspond to localized haplotype clusters. The localized haplotype cluster given by an edge of the graph consists of all haplotypes in the sample tracing their path through that edge. Figure 2 illustrates these concepts.

**INTERPRETATION OF MODEL**

The fitted graph model describing the localized haplotype clusters does not correspond to any specific population genetic model, yet it flexibly models the empirical pattern of LD seen in the data. Roughly speaking, merges (where two edges direct into the same node) correspond to historical recombination. At recombination hot-spots one would expect to see many merges, and a low number of haplotype clusters. In fact, in regions of very low LD (due to recombination hot-spots or to widely spaced markers) the number of nodes at each position may be reduced to one and the haplotype clusters will simply correspond to alleles, with one edge (cluster) for each allele at each marker. In this case the haplotypic tests reduce to single-marker tests.

**Fig. 1.** Initial steps in procedure for creating localized haplotype clusters. (A) Starting with all haplotypes in a single node, the node is split by the alleles at the first marker (a diallelic marker is shown here, however the method can be used with multiallelic markers). The resulting nodes may be merged if the merging criterion is met (for this example the nodes are not merged). (B) The nodes are split by the alleles at the second marker. The four nodes represent the four possible haplotypes when considering the first two markers. (C) Pairs of nodes may be merged. In this example, the first and third nodes are merged. The resulting merged node corresponds to all haplotypes with allele 1 at the second marker. (D) The three nodes are split by the third marker. The process continues by considering merging the resulting nodes, then splitting by the fourth marker and so on.

(fusionner les haplotypes les plus proches selon quelques critères)
Chromosome 1 in the high variant coverage sequence data is longer than that in the low variant coverage data (Table 3), suggesting not just greater resolution but also greater detection of IBD using sequence data. This may be due to the fact that fastIBD is able to detect many more small segments of size 0.5 cM or smaller with the use of sequence data (high variant density) than that using the low density genotype dataset.

Discussion
In this study, we examined how the density of genetic variants in a dataset affects the power to detect IBD between individuals. We found that analysis of sequence data with high SNP density improves resolution and power for detecting IBD relative to microarray-based genotyping, particularly for small segments. In our simulation, there was good power (80%) to detect IBD segments of size 0.4 cM using high coverage sequence data with a low false positive rate, compared to a power of approximately (77%) for segments of size 1 cM using microarray genotype data (WTCCC).

It is possible that the methods we examined in this study may be further refined to improve the power to detect even smaller IBD segments. We found that Germline has slightly higher power to detect IBD using sequence data compared to fastIBD, but it has a much higher false positive rate. That is, for high variant density data, Germline detects many small segments, where around 25% of them are false positives. We set the detectable minimum length to 0.1 cM while running Germline, which allows Germline to detect small segments, but it increases the false positive rate. Germline also provided lower power for detecting IBD segments using the microarray dataset (from WTCCC). These results indicate that fastIBD provides more robust and reliable IBD detection than Germline for these types of datasets. Given these observations, the current implementation of fastIBD appears to be better than the current implementation of Germline for detecting IBD.

WTCCC : puce avec SNPs éparse ; Hapmap : puce dense ; 1000G : Illumina ; Complete Genomics

Fins points de simulation

«Vrai» haplotypes artificiels après mixage (pas de relations non-détectées)

IBD par copiage au hasard

relatively small population could be affected dramatically by some aspects of population history (e.g., growth type and internal subdivision) [16]. In fact, recent studies have shown that both the amount of the genome shared identical by descent and the proportion of the genome that is covered by long runs of homozygosity differs by population [5,17,18]. A more detailed assessment of IBD across populations could help determine to what extent whole genome sequence data can improve the power of these mapping approaches. Additionally, the continued improvement of IBD detection methods and the testing of those methods on dense genetic data can provide a foundation for future genetic studies.

Materials and methods

Data

To assess the statistical power to detect chromosomal segments that are shared identical by descent between two individuals, we conducted a simulation study. First, we collected genotype data from four sources that represent different levels of coverage (that is, the proportion of all variants in the genome that are assayed by a given platform), ranging from microarray genotype data to deep coverage whole genome sequence data.

These empirical genotype data include:
- Microarray genotype data (WTCCC) from the Wellcome Trust Case Control Consortium (WTCCC) study. We obtained genotype data on 1000 controls from UK National Blood Donors (NBS) cohort genotyped on the Illumina 1.2 M chip. We used the SNP set released from the WTCCC database, which represents a cleaned set of data from their default QC procedures [19]
- Denser genotype data (HapMap) from the HapMap phase II project. We obtained genotype data on 60 unrelated samples from the CEU population (Utah residents with ancestry from northern and western Europe).
- Low coverage sequence data (1000 g) from the 1000 Genomes Project. We obtained genotype data on 283 individuals that originate from Europe sequenced with 4X coverage (2010.08 release).
- Deep coverage sequence data (complete) from University of California at San Francisco Whole Genome Sequencing Consortium and Complete Genomics [20]. We obtained genotype data on 54 samples of European origin sequenced by Complete Genomics with an average of 50X coverage. We used the Complete Genomics default cut offs for full genotype calls (excluding partial and no calls), which pass a strict quality score metric.

Construction of artificial IBD for assessing power

Next, to investigate the power to detect IBD of various lengths, we constructed artificial IBD segments by

---

Figure 4 Empirical false positive rate of Germline. Empirical false positive rate of Germline to detect an IBD segment as a function of the number of SNPs within a segment in the simulation study. Each plot presents different lengths of IBD segments examined. The error rate of each dataset is represented by different colored circles and plotted against the number of SNPs contained within a given region.

Table 2 The average false positive rate of fastIBD and Germline

<table>
<thead>
<tr>
<th>Segment Size (cM)</th>
<th>WTCCC</th>
<th>1000 g</th>
<th>WTCCC</th>
<th>1000 g</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2 cM</td>
<td>0.009</td>
<td>0.009</td>
<td>0.015</td>
<td>0.269</td>
</tr>
<tr>
<td>0.4 cM</td>
<td>0.007</td>
<td>0.009</td>
<td>0.006</td>
<td>0.239</td>
</tr>
<tr>
<td>0.6 cM</td>
<td>0.008</td>
<td>0.007</td>
<td>0.009</td>
<td>0.225</td>
</tr>
<tr>
<td>1.0 cM</td>
<td>0.010</td>
<td>0.009</td>
<td>0.006</td>
<td>0.223</td>
</tr>
<tr>
<td>2.0 cM</td>
<td>0.011</td>
<td>0.009</td>
<td>0.003</td>
<td>0.218</td>
</tr>
</tbody>
</table>

---

Su et al. BMC Bioinformatics 2012, 13:121
http://www.biomedcentral.com/1471-2105/13/121
la distribution de segments IBD dépend de l’histoire de la population
Segments IBD et paramètres

(taille constante, croissance exponentielle)
Démographie des Ashkenazim

meilleur fit pour données d’Ashkenazim ($n = 500$ génotypés, $\ell = 750$ k sites) : EFE avec $N_c = 2300$ (-200 générations), $\Uparrow 45$ k (-34 générations) $\downarrow 270 / \Uparrow 4.3$ M

meilleur fit pour données de Maasaï ($n = 78$ génotypés, $\ell = 1.5 \text{M sites}$) : EFE avec $N_c = 23500$ (-23 générations) , $\geq 500$

explication : villages avec migration de taux bas ($\Rightarrow$ ancêtres communs sont plus anciens)
Population des Amériques

Identity by descent analysis

We used germline [55] and the trio-phased OMNI data above to identify segments identical-by-descent (IBD) within and across populations (see Text S1). Not surprisingly, we found more IBD

Figure 2. (a) Individual ancestry proportions in the 1000 Genomes CLM, MXL, and PUR populations according to ADMIXTURE, (b) Map showing the sampling locations for the populations most closely related to the Native components of the 1000 Genomes populations. (c) Principal component analysis restricted to genomic segments inferred to be of Native Ancestry in these populations, compared to a reference panel of Native American groups from [40], pooled according to country of origin as a proxy for geography. Populations sampled across many locations are labeled according to the country of the centroid of locations. (d) Zoomed version of the PCA plot, showing specific Native American population labels, colored according to country of origin.

doi:10.1371/journal.pgen.1004023.g002

The MXL population exhibits significantly less within-population IBD compared to the other two panels (Figure 4). The amount of IBD among unrelated individuals can be used to infer the underlying population size under panmictic assumption: the larger a population, the more distant the expected relationship between any two individuals [56]. Using IBD segments longer than 4 cM, we infer effective population sizes of 140,000 in MXL, 15,000 in CLM, and 10,000 in PUR. As we will show, these largely reflect post-ADMixTURE population sizes.

We expect long IBD segments to be inherited from a recent common ancestor, and therefore to have identical continental ancestry. Comparing the RFMIX ancestry assignments on chromosomes that have been identified as IBD by germline thus provides a measure of the consistency of the two methods (see [57] for a related metric). Rates of IBD-Ancestry mismatch ranged from $2\%$ in segments of 5 Mb to less than $0\%$ for segments longer than 40 Mb (Figure S4).

Patterns of ancestry in IBD segments within a population differ markedly from those across populations (Figure 5): IBD segments within populations contain many ancestry switches. This indicates that many common ancestors lived after contact, and that the effective population sizes estimated using IBD largely reflects post-contact demography. The IBD patterns in cross-population IBD segments exhibited fewer ancestry switches than a random control (Figure S5), as may be expected if common ancestors often predate the onset of ADMIXTURE. Cross-population IBD segments were also found to be overwhelmingly of European origin: among the 120 longest cross-population IBD segments, 117 are in European-inferred segments, two are among Native segments, and one is among African segments. This is not due to overall ancestry proportions, as can be observed by considering the alternate (non-IBD) haplotypes at the same positions (Figure S5). This is likely a result of the colonization history, in which European colonists rapidly spread from a relatively specific region over a large continent. This interpretation is supported by the ADMIXTURE analysis (Figure S6), showing a common cluster of ancestry for the European component dominant in PUR, CLM, MXL, and Andean populations, but not in CEU, Eskimo-Aleut, and Na-Dene. Finally, we were interested in testing whether the relationship between IBD and ancestry can be used to date...
By calibrating our results using $T \sim 16\, \text{kya}$, towards the most recent end of the range of plausible values for the peopling of the Americas (see e.g., [6] and references therein), we find a mutation rate of $1.44 \times 10^{-8}$ bp $^{-1}$ gen$^{-1}$ (bootstrap 95% CI: $1.32 \times 10^{-8}$ bp $^{-1}$ gen$^{-1}$, $1.53 \times 10^{-8}$ bp $^{-1}$ gen$^{-1}$), within the range of recently published human mutation rates [63]. The narrowest confidence interval reported in [63] was $1.05 \times 10^{-8}$ bp $^{-1}$ gen$^{-1}$, obtained from a de novo exome sequencing study [64]. Our sampling confidence interval is narrower than this value, but the main source of uncertainty here is the degree to which the bottleneck in our model reflects the bottleneck at the founding of the Americas, or the earlier split with the ancestors to the Chinese (CHB) and Japanese (JPT) sample, as well as uncertainty with respect to the timing of these two events (see Figure 7). The effect of changing the founding time or mutation rate assumptions would be to scale all parameters and confidence intervals according to $T \propto N \propto \frac{1}{m}$: Thus the absolute uncertainty on individual parameters is larger than the sampling uncertainty suggests.

Estimating Native American allele frequencies

There is scarce publicly available, genome-wide data about Native American genomic diversity. The 1000 Genomes dataset offers the opportunity to provide a diversity resource for Native American genomics by reconstructing the genetic makeup of Native American populations ancestral to the PUR, CLM, and MXL. This is particularly interesting in the case of the Puerto Rican population, where such reconstruction may be the only way to understand the genetic make-up of the pre-Columbian inhabitants of the Islands. Using the expectation maximization method presented in the Methods section, we estimated the allele frequencies in the Native-American-inferred part of the genomes of the sequenced individuals. These estimates are available at http://genomes.uprm.edu/cgi-bin/gb2/gbrowse/.

Figure 8 shows the distribution of the number of Native American haplotypes per site and the resulting confidence intervals for allele frequency in each population for exome capture target regions. Absolute confidence intervals are narrow for rare variants, and reach a maximum for SNPs at intermediate

arrivée : 16000 ans, MXL 12200 ans, PUR/CLM 11700 ans, migrations