Histoire épidémiologique du sous-type C du VIH-1 dans la pandémie mondiale

Le sous-type C du virus de l'immunodéficience humaine de type 1 (VIH-1) est responsable de près de 50% des infections mondiales au VIH-1, mais il est surtout prévalent en Afrique australe, en Afrique de l'est, en Inde et au sud du Brésil. Certaines études d'épidémiologie moléculaire montrent que ce variant génétique s'est propagé au Brésil et en Inde à partir du Burundi et de l'Afrique du Sud respectivement. Ces études se basent systématiquement sur un échantillon réduit des souches disponibles et les migrations du sous-type C au sein même de l'Afrique restent mal connues. Nous proposons une étude visant à connaître l'origine géographique de l'épidémie du sous-type C, ainsi que ses migrations dans le monde entier, incluant l'Afrique, en utilisant toutes les souches disponibles de ce variant plus 528 souches séquencées par l'équipe TransVIHMI. La phylogénie obtenue, comprenant plus de 3 600 souches, est difficilement interprétable « à la main ». Plusieurs indices basés sur les transitions entre pays (reconstruites par parcimonie) sont proposés afin de donner une vision synthétique des flux migratoires du sous-type C à l'échelle mondiale. Le logiciel PhyloType est ensuite utilisé pour mettre en valeur des liens entre les évènements fondateurs probables. La plupart des flux migratoires du sous-type C décrits dans la littérature sont observés, par exemple le lien entre le Brésil et le Burundi, et d'autres sont différents, par exemple, la Zambie est suggérée être à l'origine de l'épidémie en Inde. En Afrique, ce variant se propage indépendamment de la Zambie, épicentre de l'épidémie, vers l'Afrique australe et vers l'Afrique de l'est.

Sommaire

6.1 Introduction

Les erreurs lors de la rétrotranscription, les phénomènes de recombinaison, la pression de sélection immunitaire et médicamenteuse, un fort taux de réplication virale ont donné lieu à de nombreuses variantes génétiques du virus de l'immunodéficience humaine (VIH) que l'on nomme sous-types ou formes recombinantes circulantes (*circulating recombinant forms*, CRF) (Rambaut *et al*, 2004). Ceci est le résultat de l'adaptation du virus à son environnement (Brun-Vézinet *et al*, 1999).

De ce fait, les souches du groupe pandémique du VIH-1 (groupe M) présentent une diversité génétique importante. Elles sont répertoriées dans 9 sous-types (A à D, F à H, J et K), 6 sous-sous-types (A1 à A4, F1 et F2) et 51 CRF (CRF01_AE à CRF51_01B) ; sans compter les nombreuses formes recombinantes uniques découvertes (*unique recombinant forms*, URF). Le variant génétique du groupe M le plus répandu est, sans conteste, le sous-type C (VIH-1C). Ce sous-type est responsable de 48,23% des infections mondiales liées au VIH-1[8], mais sa distribution géographique est hétérogène et en constante évolution. Cette distribution géographique hétérogène est la résultante de nombreux facteurs, tant biologiques que sociologiques (Perrin *et al*, 2003). En effet, le VIH-1C a une virulence moindre par rapport à celle des autres sous-types (Abraha *et al*, 2009) et donc une phase asymptomatique plus longue, laissant plus d'opportunité de transmission (Ariën *et al*, 2007). Le sous-type C a aussi une prédisposition plus élevée à se localiser dans les muqueuses génitales des femmes (Walter *et al*, 2009) favorisant ainsi la transmission par contact hétérosexuel.

L'épidémie du VIH en Afrique australe, en Éthiopie et en Inde est presque exclusivement due au sous-type C. Pour ces pays, les études épidémiologiques estiment respectivement que 98,31%,

[8] D'après Hemelaar *et al*. (2011), sur la période 2004-2007.

97,44% et 97,77% des infections au VIH-1 sont dues au sous-type C[8]. Elles ont aussi montré que deux épidémies différentes du sous-type C (C et C') co-circulent en Éthiopie (Abebe *et al*, 2000), probablement d'origine différente. Dans le reste de l'Afrique, le sous-type C est aussi observé à l'est, où il est responsable de 22,97% des infections (Éthiopie exclue)[8], avec une forte prévalence au Burundi, où il est responsable de plus de 80% des infections (Vidal *et al*, 2007; Koch *et al*, 2001), et au centre où il est responsable de 5,75% des infections[8], notamment à Lubumbashi et à Mbuji-Mayi, deux villes situées au sud de la République Démocratique du Congo (RDC), où il est responsable de 51,3% et de 16,3% des infections respectivement (Vidal *et al*, 2005, 2000). À l'ouest et au nord de l'Afrique le sous-type C est assez rare (<1%[8]). Toutefois, il est retrouvé avec une prévalence de 40% chez les MSM sénégalais (Ndiaye *et al*, 2009) (cf. Chapitre 5). Sur le continent américain, le sous-type C est surtout prévalant au sud du Brésil (Soares *et al*, 2005) et il est aussi observé dans quelques autres pays voisins, comme l'Argentine ou l'Uruguay (Carrion *et al*, 2004). En Amérique du nord et centrale des cas sont observés mais restent rares (Sides *et al*, 2005; Cuevas *et al*, 2002). En Asie (sauf Inde), la prévalence du sous-type C est faible (2,93%[8,9]), mais elle est élevée en Océanie (particulièrement aux îles Fidji (Ryan *et al*, 2009) et en Papouasie-Nouvelle-Guinée (Ryan *et al*, 2007)) où le VIH-1C est responsable de 66,34% des infections[8]. En Europe, le sous-type C est très peu prévalant et est souvent observé chez des patients qui ont des liens avec l'Afrique (Paraschiv *et al*, 2011; Giuliani *et al*, 2009; Vercauteren *et al*, 2008; Tatt *et al*, 2004; Couturier *et al*, 2000; Alaeus *et al*, 1997).

L'émergence ou l'introduction d'un nouveau variant dans une population donnée, où un autre variant génétique est déjà prédominant, peut provoquer des phénomènes de recombinaison lors des co- ou surinfections. C'est le cas de l'introduction du sous-type C chez les utilisateurs de drogues intraveineuses (*intravenous drug users*, IDU) en Asie de l'est, où le sous-type B était prédominant, qui engendra l'épidémie des CRF08_BC et CRF07_BC (Takebe *et al*, 2010). Au sud du Brésil, l'introduction du sous-type C a produit l'épidémie du CRF31_BC (Passaes *et al*, 2009).

L'origine et la diffusion des virus restent d'un intérêt majeur pour les épidémiologistes, car la diversité génétique peut avoir des conséquences sur l'efficacité d'outils diagnostiques (sérologiques et/ou moléculaires), le développement de résistances aux antirétroviraux, la pathogénicité virale ou la possibilité de développement d'un vaccin. Depuis l'avènement de la phylogénie moléculaire, de nombreuses méthodes permettent aujourd'hui de répondre aux questions relatives à la dynamique des épidémies, sur la base des séquences nucléotidiques. La plupart de ces méthodes utilisent ou infèrent un arbre phylogénétique et déduisent, à partir de celui-ci, les régions géographiques correspondantes aux nœuds ancestraux de l'arbre connaissant celles associées aux souches contempo-

[9] Ce chiffre prend en considération quelques pays de l'Europe de l'est.

raines (représentées par les feuilles dans l'arbre). Par exemple, Véras *et al*. (2011a) utilisent le principe de parcimonie (décrit à la fin du Chapitre 1) et Faria *et al*. (2011) utilisent une méthode bayésienne implémentée dans la suite de logiciels BEAST. Cette branche de la phylogénie moléculaire porte le nom de phylogéographie dont Avise (2000) donne la définition suivante : « *field of study concerned with the principles and processes governing the geographical distributions of geographical lineages, especially those within and among closely related species* ».

Des études phylogénétiques de ce type décrivent déjà les migrations de l'épidémie du VIH-1C (de Oliveira *et al*, 2010; Bello *et al*, 2008; Fontella *et al*, 2008; Qiu *et al*, 2005; Gehring *et al*, 1997; Dietrich *et al*, 1995). L'épidémie du sous-type C en Afrique du Sud s'est propagée en Inde suite à un évènement fondateur d'origine inconnue (Shen *et al*, 2011; Dietrich *et al*, 1995, 1993), puis s'est introduite en Chine où quelques souches virales se sont recombinées avec du sous-type B (plus précisément avec un sous-cluster particulier du B, appelé B' ou Thai B) pour former les recombinants CRF08_BC et CRF07_BC qui circulent chez les IDU (Qiu *et al*, 2005). L'épidémie du sous-type C qui sévit en Éthiopie s'est répandue en Israël en 1991 à la suite de l'opération *Salomon* qui permit à plus de 14 000 juifs d'Éthiopie de rejoindre l'Israël (Gehring *et al*, 1997). Notons que ce variant est resté endémique à cette communauté sur le territoire israélien. En Amérique du sud, l'épidémie du sous-type C a pour centre de dispersion le sud du Brésil (Véras *et al*, 2011a; de Oliveira *et al*, 2010; Jones *et al*, 2009; Bello *et al*, 2008; Fontella *et al*, 2008), puis s'est répandue en Argentine et en Uruguay par le biais d'immigrants et de touristes (Carrion *et al*, 2004). Dans la littérature, l'introduction du sous-type C sur le territoire d'Amérique Latine a connu plusieurs origines géographiques différentes. Cependant l'hypothèse la plus citée reste celle de l'Afrique de l'est (le Burundi est souvent mentionné). Récemment, de Oliveira *et al*. (2010) suggèrent que l'épidémie s'est propagée du Burundi vers l'Angleterre puis de l'Angleterre vers le Brésil, mais Véras *et al*. (2011a) apportent une controverse à cette théorie. En revanche, toutes ces études s'accordent sur le fait que l'origine de l'épidémie du VIH-1C en Amérique du sud est monophylétique (ou avec un nombre faible d'introductions marginales).

Malgré le nombre grandissant d'études de ce genre, à notre connaissance, aucune ne montre ou ne discute de l'origine exacte de l'épidémie du VIH-1C, qui, au vu des informations relevées dans la littérature, semble être en Afrique. De plus, aucune donnée générale sur le mouvement de l'épidémie du sous-type C en Afrique n'est disponible. Seule observation à noter, les souches collectées en Afrique de l'est se regroupent dans un cluster (Thomson & Fernández-García, 2011).

La plupart de ces études moléculaires utilisent seulement une partie des souches disponibles. En effet, ces études utilisent généralement des méthodes probabilistes, lourdes en temps de calcul, qui

permettent rarement de dépasser quelques centaines de séquences. Cependant, la sélection de souches peut introduire un biais dans les résultats. L'exemple du rôle de l'Angleterre dans l'origine de l'épidémie du VIH-1C en Amérique du sud est caractéristique : de Oliveira *et al*. (2010) se sont restreint à certaines séquences et en ont déduit que l'Angleterre a eu un rôle dans la diffusion de l'épidémie au Brésil. L'étude de Véras *et al*. (2011a) utilise les mêmes souches d'Angleterre mais avec une couverture plus large de souches du VIH-1C collectées en Amérique du sud et trouve des conclusions différentes.

Nous présentons ici la première étude de phylogénie moléculaire visant à retracer l'histoire épidémiologique mondiale et l'origine du VIH-1C. Pour ce faire, nous inférons une phylogénie contenant toutes les souches du VIH-1C disponibles, sans aucune sélection aléatoire ou arbitraire. Elles proviennent de la base de données sur le VIH du laboratoire de Los Alamos, soit 3 081 séquences utilisées dans cette étude (cf. Chapitre 5), auxquelles nous ajoutons 528 séquences collectées en Afrique, continent suspecté d'être à l'origine de l'épidémie du VIH en général et donc aussi du sous-type C. À l'aide de cette phylogénie et de la connaissance de l'origine géographique des souches nos objectifs sont de : 1) déterminer l'origine géographique de l'épidémie du VIH-1C ; 2) connaître les voies ayant mené à la diffusion de ce variant à travers le monde. Toutefois, une telle phylogénie, associée aux origines géographiques des séquences, est très difficilement interprétable de façon manuelle et nécessite l'utilisation d'outils ou logiciels permettant les analyses phylogénétiques d'un grand nombre de souches dans un temps relativement court. Aussi, nous proposons deux manières différentes mais complémentaires pour analyser ces données, utilisant toutes deux le principe de parcimonie (détaillé à la fin du Chapitre 1). La première utilise des indices basés sur les transitions entre pays (reconstruites par parcimonie) pour donner une information synthétique retraçant les grandes tendances (pays donneurs ou receveurs, symétrie des échanges, etc.). Ces indices sont proches de ceux proposés par Slatkin et Maddison (1989) ou encore Salemi *et al*. (2005) aussi définis à partir du nombre de transitions entre annotations. Associés à des sorties graphiques appropriées, ces indices permettent très rapidement de se faire une idée globale sur l'épidémie. La seconde approche est basée sur l'outil PhyloType (Chevenet, Jung, de Oliveira et Gascuel, en cours de soumission) qui permet d'identifier des évènements fondateurs probables, comme par exemple l'introduction du VIH-1C au Brésil ou chez les MSM du Sénégal. La section suivante présente les méthodes et logiciels utilisés pour la préparation des données, la définition des indices utilisés permettant de synthétiser l'information de la phylogénie, une présentation sommaire du logiciel PhyloType et enfin son paramétrage. Les deux dernières sections présentent respectivement les résultats obtenus et leurs discutions.

6.2 Préparation des données

6.2.1 Conception de l'alignement

Pour cette étude, nous réutilisons l'alignement du Chapitre 5 contenant tous les sites, même ceux associés à des mutations de résistance (Hué *et al*, 2004). Pour mémoire, il contient 3 081 séquences du sous-type C du VIH-1, collectées à travers le monde et à différentes dates, d'une longueur de 1 011 sites et correspondant à la région génomique 2 253-3 263 (*pol*) d'HXB2 (codant l'intégralité de la protéase et le début de la transcriptase inverse).

À cet alignement, 528 nouvelles séquences, collectées et séquencées par l'équipe TransVIHMI, sont ajoutées. Parmi ces 528 séquences, 199 sont collectées au Burundi en 2008 et 20 en 2010, 1 en République Démocratique du Congo (RDC) en 2007 et 66 en 2008, 1 en République Centrafricaine en 2006, 2 en République du Congo en 2007, 1 en Éthiopie en 1999 et 238 au Swaziland en 2008. Ces séquences sont toutes confirmées comme appartenant au sous-type C soit par l'application web RE-GA HIV-1 & 2 *Automated Subtyping Tool* (de Oliveira *et al*, 2005), soit à l'aide d'analyses de similarité et *bootscan* réalisés avec le logiciel SimPlot (Lole *et al*, 1999). La séquence HXB2 (sous-type B ; numéro d'accession : K03455) sert d'*outgroup* pour enraciner l'arbre de maximum de vraisemblance construit dans cette étude.

Un alignement séquences contre profil est effectué afin d'ajouter les 528 nouvelles séquences à l'alignement initial de 3 081 séquences. La méthode d'alignement et le logiciel utilisés sont les mêmes qu'au Chapitre 5, à savoir MAFFT version 6 (Katoh *et al*, 2002) avec la méthode L-INS-i (Katoh *et al*, 2005). Quelques corrections manuelles sont apportées avec MEGA version 5 (Tamura *et al*, 2011) et les sites contenant un nombre excessif de gaps ($\geq 50\%$) sont supprimés.

Au final, nous obtenons un alignement de 1 011 sites contenant 3 609 séquences collectées dans 63 pays différents entre 1986 et 2010. Le Tableau 4 liste le nombre de souches pour chaque pays présents dans cette étude.

6.2.2 Inférence phylogénétique

Les mêmes paramètres et les mêmes options sont utilisés pour calculer l'arbre de maximum de vraisemblance, que pour celui contenant les 3 081 souches du Chapitre 5. Pour rappel, il est inféré sous le modèle *general time reversible* (GTR) avec des sites invariants et une loi gamma discrète avec 4 catégories de taux (GTR+I+Γ4), comme conseillé par Posada et Crandall (2001), avec le logiciel PhyML v3.0 (Guindon *et al*, 2010). L'option *subtree pruning and regrafting* (SPR) est choisie pour explorer l'espace des arbres. Tous les paramètres sont évalués et optimisés par PhyML. Les supports

de branche sont déterminés par la méthode *approximate likelihood ratio test* (aLRT) (Anisimova & Gascuel, 2006), option SH-like.

Tableau 4. Liste des pays utilisés dans cette étude, ainsi que le nombre de séquences associées en nombre et en pourcentage.

Afrique				**2 615**	**72,46%**
Afrique du Sud	689	19,09%	Mozambique	98	2,72%
Botswana	133	3,69%	Niger	4	0,11%
Burundi	310	8,59%	Ouganda	16	0,44%
Congo	2	0,06%	Rép. Centrafricaine	1	0,03%
Djibouti	1	0,03%	Rép. Démo. du Congo	86	2,38%
Érythrée	2	0,06%	Sénégal	56	1,55%
Éthiopie	100	2,77%	Somalie	1	0,03%
Gabon	1	0,03%	Soudan	10	0,28%
Guinée Équatoriale	1	0,03%	Swaziland	285	7,90%
Kenya	4	0,11%	Tanzanie	82	2,27%
Malawi	71	1,97%	Zambie	633	17,54%
Mali	1	0,03%	Zimbabwe	28	0,76%
Amérique				**299**	**8,28%**
Argentine	8	0,22%	Honduras	1	0,03%
Brésil	253	7,01%	Uruguay	2	0,06%
Cuba	25	0,69%	Venezuela	1	0,03%
États-Unis	9	0,80%			
Asie				**380**	**10,53%**
Birmanie	1	0,03%	Israël	5	0,14%
Chine	7	0,19%	Philippines	1	0,03%
Corée du Sud	2	0,06%	Taiwan	1	0,03%
Inde	355	9,84%	Yémen	7	0,19%
Europe				**315**	**8,73%**
Allemagne	7	0,19%	Luxembourg	3	0,08%
Autriche	3	0,08%	Norvège	16	0,44%
Belgique	35	0,97%	Pays-Bas	8	0,22%
Chypre	8	0,22%	Pologne	2	0,06%
Danemark	21	0,58%	Portugal	28	0,78%
Espagne	26	0,72%	Rép. Tchèque	11	0,30%
Finlande	6	0,17%	Roumanie	35	0,97%
France	7	0,19%	Russie	1	0,03%
Géorgie	1	0,03%	Slovaquie	1	0,03%
Grande-Bretagne	3	0,08%	Suède	64	1,77%
Grèce	3	0,08%	Suisse	2	0,06%
Italie	22	0,61%	Ukraine	3	0,08%

Compte tenu du nombre important de séquences, la méthode du *bootstrap* n'est pas utilisée. De même, aucune approche bayésienne n'est possible.

6.2.3 Reconstruction des états ancestraux

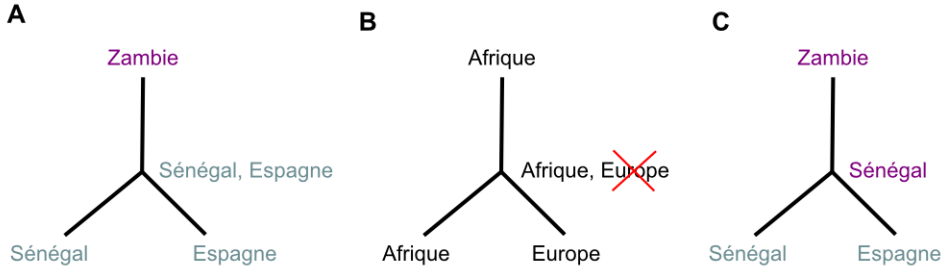
Afin de comprendre le mouvement de l'épidémie du VIH-1C dans son intégralité, les états géographiques ancestraux de chaque nœud de la phylogénie sont calculés à partir de l'information sur les

pays de collecte des souches contemporaines. La méthode choisie pour inférer les états ancestraux est la parcimonie (Hartigan, 1973; Fitch, 1971) parce qu'elle nécessite peu de temps de calcul comparé aux méthodes probabilistes où l'utilisation d'une telle quantité de données rend le temps de calcul prohibitif. Par ailleurs, il n'existe pas (comme pour les séquences) de modèle simple et consensuel pour les mouvements géographiques et le passage d'un pays à un autre.

Deux souches virales sont initialement écartées de cette analyse. L'*outgroup* (83FR-HXB2) et l'isolat 02ZA-1752 (numéro d'accession EF602195). Cet isolat a été collecté en 2002 au Cap en Afrique du Sud par Jacobs *et al*. (2008) et il est ancestral à toutes les souches de la phylogénie hormis 83FR-HXB2. Cependant, dans la phylogénie de l'étude sur le VIH-1C au Sénégal (cf. Chapitre 5) et dans les phylogénies de Jacobs *et al*. (2008), cet isolat se situe dans un clade contenant d'autres souches d'Afrique du Sud. Donc, cette souche semble être une *rogue taxon* (Trautwein *et al*, 2011), c'est-à-dire une souche « avec un placement phylogénétique incertain et variable qui a généralement un effet négatif sur la reconstruction topologique et les valeurs supports ». Pour cette raison, elle n'est pas considérée dans l'interprétation de la phylogénie.

La première phase du calcul des états ancestraux est effectuée avec l'algorithme UPPASS, décrit à la section 1.6. Les calculs par parcimonie nécessitent l'utilisation d'une seconde phase afin que chaque nœud interne (exception faite du nœud racine) exploite l'information de toute la phylogénie. Les algorithmes DOWNPASS (Maddison & Maddison, 2003), ACCTRAN (Fitch, 1971) et DELTRAN (Swofford & Maddison, 1987) sont choisis pour cette seconde phase. Ils sont aussi décrits à la section 1.6. Comme beaucoup de nœuds internes sont ambigus (plus d'un état est assigné) à la fin de la seconde phase, deux règles sont ajoutées pour résoudre ces ambiguïtés au cours de cette seconde phase. La première utilise la parcimonie pour résoudre l'ambiguïté d'un nœud, mais en utilisant non plus les pays, mais les continents auxquels ils appartiennent. Ainsi, les états attribués à ce nœud sont uniquement les pays appartenant au(x) continent(s) inféré(s) par parcimonie (Figure 42). La seconde règle choisit aléatoirement un état parmi ceux restants (s'il en reste plusieurs) et l'assigne comme état final. De ce fait, chaque nœud de la phylogénie contient exactement une seule annotation. Lorsque le choix aléatoire est utilisé pour résoudre les ambiguïtés, la procédure complète est répétée 1 000 fois et les nombres finals de transitions sont obtenus par des moyennes sur les 1 000 cas. Notons que les ambiguïtés au niveau du nœud racine peuvent uniquement être résolues avec le choix aléatoire (du moins si comme ici on décide de ne pas prendre en compte l'*outgroup*, très éloigné phylogénétiquement de l'*ingroup*). Les algorithmes de parcimonie utilisant ces deux règles seront précédés par le terme « Rand » afin de les distinguer des algorithmes originaux.

Figure 42. Illustration de la première règle pour la résolution de nœuds ambigus.
Les annotations correspondantes à la phase ascendante de la parcimonie sont surlignées en bleu, tandis que les annotations de la phase descendante en mauve. La figure A montre un nœud où, par exemple, la parcimonie ACCTRAN ne peut résoudre l'ambiguïté puisque l'annotation Zambie n'est pas associée au nœud interne. La figure B montre ce même nœud mais en regardant les continents associés aux pays. Maintenant la parcimonie ACCTRAN peut résoudre l'ambiguïté et calcule que l'annotation correspondante est le continent africain. La figure C montre les annotations associées au nœud après sa résolution à l'aide de la règle des continents.



6.2.4 Mesure des taux de migrations entre pays

Une fois la reconstruction des états ancestraux achevée, chaque nœud interne de la phylogénie contient un unique état correspondant à la localisation géographique (dans notre cas, le pays) la plus parcimonieuse de l'ancêtre commun représenté par ce nœud. Certaines branches de la phylogénie, dont les localisations aux deux extrémités diffèrent, symbolisent alors les migrations, aussi appelées transitions, du VIH-1C au cours de son histoire évolutive. Nous proposons ici des indices, basés sur ces transitions, qui visent à donner une vision synthétique des migrations du VIH-1C. Une méthode de ré-échantillonnage aléatoire (ou *shuffling*) permet de dégager la significativité de ces mesures, et de s'affranchir (au moins pour une part) des effets liés aux tailles variables d'échantillon par pays (Wallace *et al*, 2007). Autrement, il serait nécessaire d'utiliser des tailles similaires d'échantillon par pays, et par conséquent réduire considérablement le nombre de séquences étudiées, afin d'éviter tout biais potentiel sur les estimations de ces mesures (Véras *et al*, 2011a). Également, en procédant ainsi, on considère que le nombre de souches disponibles dans chaque pays représente peu ou prou la prévalence du sous-type C dans le pays. Avec des échantillons ramenés à la même taille cette information disparaît.

Notations

Soit $\mathcal{E}$ l'ensemble de toutes les annotations (pays) et soient a et b deux annotations de cet ensemble. Le nombre de transitions de l'annotation a vers l'annotation b se note $N_{a\to b}$, c'est-à-dire qu'il représente le nombre de branches dont le nœud adjacent le plus proche de la racine est annoté a et le plus éloigné b. Considérons, de plus, le nombre de transitons de l'annotation b vers l'annotation a, $N_{b\to a}$, le nombre de fixations (branches dont les deux nœuds aux extrémités ont la même annotation) de l'annotation a, $N_{a\to a}$, et le nombre de fixations de l'annotation b, $N_{b\to b}$. De manière générale, $N_{a\to X} = \sum_{u\in\mathcal{E},u\neq a} N_{a\to u}$ désigne le nombre de toutes les transitions sortantes de a et $N_{X\to a} = \sum_{u\in\mathcal{E},u\neq a} N_{u\to a}$ le nombre de toutes les transitions entrantes dans a. Si $N_{X\to a} > 0$, le pays

a est dit receveur (ou IN) et si $N_{a\to X} > 0$, le pays a est dit donneur (ou OUT). Paraskevis *et al*. (2009) utilisent une terminologie similaire. Les indices proposés dans la suite vont nous servir à quantifier les tendances principales (plutôt donneur ? vers quel pays ? plutôt receveur ? de quel pays ?). Notons aussi par $|a| = N_{a\to X} + N_{a\to a}$, le nombre de transitions ayant l'annotation a en entrée et supposons que la phylogénie contient n arêtes.

Indice de dispersion

Nous proposons un premier indice D_a qui indique le degré de dispersion d'une annotation $a \in \mathcal{E}$ au sein de la phylogénie, c'est-à-dire si les feuilles annotées a sont plus ou moins regroupées (dans le cas extrême, elles forment un clade) ou dispersées (dans le cas extrême, aucune feuille annotée a n'est à côté d'une autre feuille a). Cet indice est construit à partir de l'indice $D_{a\to X}$, qui normalise le nombre de transitions sortantes de a (OUT), et de l'indice $D_{X\to a}$, qui normalise le nombre de transitions entrantes dans a (IN). Posons

$$D_{a\to X} = \frac{N_{a\to X}}{N_{a\to X} + N_{X\to a} + N_{a\to a} - 1},$$

$$D_{X\to a} = \max\left\{0, \frac{N_{X\to a} - 1}{N_{a\to X} + N_{X\to a} + N_{a\to a} - 1}\right\}$$

et

$$D_a = D_{a\to X} + D_{X\to a}.$$

La fonction maximum de l'indice $D_{X\to a}$ permet d'éviter les valeurs négatives lorsque l'annotation a n'est pas receveuse (quand la racine est annotée a et que cette annotation est uniquement donneuse), c'est-à-dire lorsque $N_{X\to a} = 0$. Si l'annotation a est totalement dispersée au sein de la phylogénie et qu'aucun état ancestral a n'a pu être inféré, alors $N_{a\to a} = N_{a\to X} = 0$ et donc $D_{X\to a} = D_a = 1$. Au contraire, lorsque l'annotation a est totalement régionalisée dans l'arbre au sein d'un clade unique, la mesure $N_{a\to X} = 0$ et $N_{X\to a} = 1$ (sauf si a est l'unique annotation de la phylogénie), donc $D_{a\to X} = D_{X\to a} = D_a = 0$ (Figure 43). Le dénominateur est nul lorsque l'annotation a est uniquement représentée par une feuille. Dans ce cas (peu intéressant), cette mesure n'est pas utilisée.

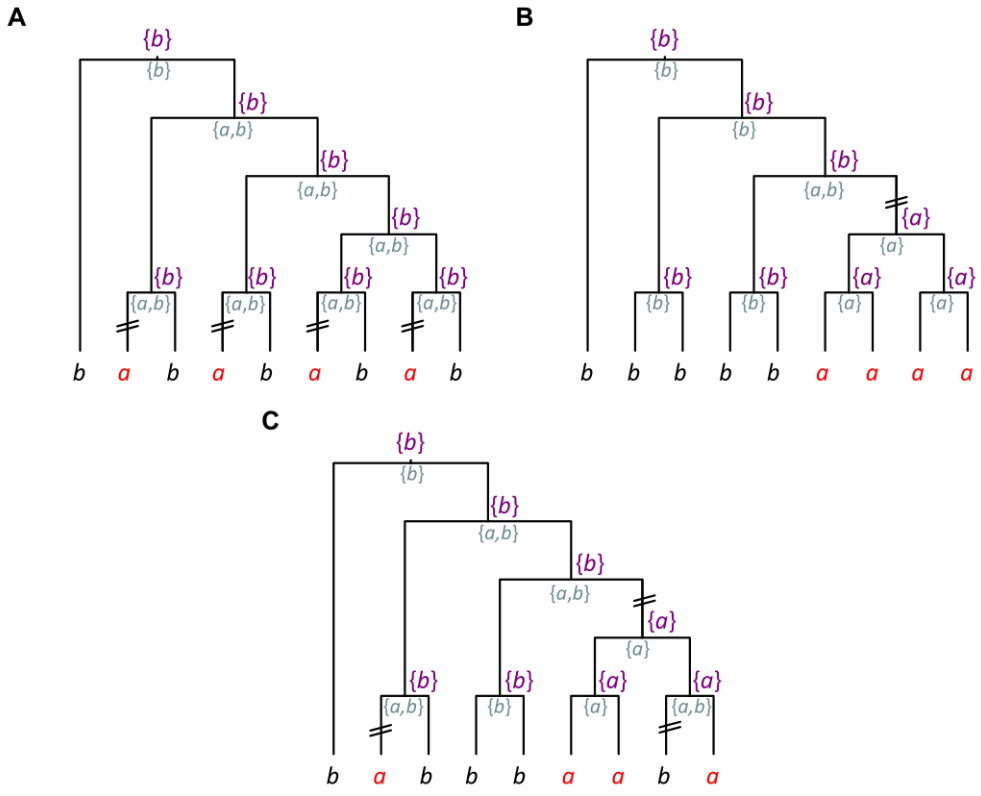
Indice de flux

Nous proposons un second indice $F_{a\to b}$ qui renseigne sur la fréquence des transitions de l'annotation a vers l'annotation b. Cet indice est largement utilisé par d'autres (Salemi *et al*, 2008; Wallace *et al*, 2007) et certains logiciels proposent de le calculer (Maddison & Maddison, 2003). Toutefois, nous lui ajoutons une normalisation afin d'augmenter la lisibilité des graphiques. L'indice de flux $F_{a\to b}$, avec $a, b \in \mathcal{E}$, $a \neq b$, est défini par

$$F_{a\to b} = \frac{N_{a\to b}}{\sum_{i\in\mathcal{E}, i\neq b} N_{i\to b}}.$$

Lorsqu'un pays b n'est receveur que d'un seul pays, disons a, la mesure $F_{a\to b}$ vaut 1 et $F_{i\to b}$ vaut 0 quel que soit le pays i. Pour tout b, la somme des mesures $F_{i\to b}$ vaut 1, $i \in \mathcal{E}$, $i \neq b$. Cet indice représente donc la proportion du nombre de transitions de a vers b parmi toutes les transitions vers b. Notons que si a est uniquement un pays receveur alors $F_{a\to i} = 0$, pour tout $i \in \mathcal{E}$, $i \neq a$. Il est donc inutile de reporter ces mesures. En exemple, la phylogénie A de la Figure 43 montre que $F_{a\to b} = 0$ et la mesure $F_{b\to a} = 1$.

Figure 43. Exemples d'application avec l'indice de dispersion.
Ces deux figures illustrent le comportement de l'indice de dispersion. Les annotations ancestrales calculées lors de la phase ascendante de la parcimonie sont indiquées en bleu, celles obtenues lors de la phase descendante (par ACCTRAN) en mauve. Les doubles barres obliques indiquent les branches où une transition se produit. La figure A montre un cas où l'annotation a est totalement dispersée dans la phylogénie, $D_a = D_{X\to a} = 1$ et $D_{a\to X} = 0$, la figure B un cas où les feuilles associées à l'annotation a forment un clade, $D_a = D_{a\to X} = D_{X\to a} = 0$, et la figure C un cas intermédiaire $D_{a\to X} = D_{X\to a} = 1/6$ et $D_a = 1/3$.



Indice de symétrie

Le dernier indice introduit indique la quantité de transitions échangées entre deux pays a et b de $\mathcal{E}$. Il permet de vérifier si l'échange est symétrique (autant de transitions de a vers b que de b vers a), unidirectionnel (que des transitions de a vers b ou que des transitions de b vers a) ou bidirectionnel (des transitions en quantité variable de a vers b et de b vers a). Afin d'intégrer les effectifs

correspondant à chaque annotation a et b, il s'inspire dans sa définition d'un processus de Markov stationnaire réversible dans le temps, c'est-à-dire que pour tout a et b de $\mathcal{E}$, $a \neq b$,

$$p_a p_{a\to b} = p_b p_{b\to a}$$

où p_a (respectivement p_b) est la probabilité d'observer un nœud interne annoté a (resp. b) et $p_{a\to b}$ (resp. $p_{b\to a}$) la probabilité d'observer une transition de a vers b (resp. b vers a). Or $p_i = |i|/n$, où n représente le nombre d'arêtes de la phylogénie, et $p_{i\to y} = N_{i\to y}/|i|$. Après simplification, la relation devient tout simplement

$$N_{a\to b} = N_{b\to a}.$$

L'indice de symétrie est donc défini pour tout a et b par (Véras *et al*, 2011a)

$$S_{a\leftrightarrow b} = N_{a\to b} - N_{b\to a}.$$

Remarquons que $S_{a\leftrightarrow b} = -S_{b\leftrightarrow a}$. Si l'échange est parfaitement symétrique alors $N_{a\to b} = N_{b\to a}$ et la mesure vaut 0. Si la mesure est positive (resp. négative), alors il y a plus de transitions de a vers b (resp. b vers a) que l'inverse. Si l'échange est unidirectionnel alors $|S_{a\leftrightarrow b}/(N_{a\to b} + N_{b\to a})| = 1$. Il est évident que cet indice est uniquement appliqué s'il existe des échanges entre a et b.

Test de ré-échantillonnage aléatoire

Afin de vérifier la significativité statistique des résultats observés, la procédure de ré-échantillonnage aléatoire (*shuffling*) est utilisée 1 000 fois pour comparer les valeurs observées à celles de l'hypothèse nulle ou panmixie (cf. Chapitre 1). Pour chaque paire de pays et chaque indice, la valeur observée est comparée à la distribution des valeurs aléatoires obtenues. Un indice observé est jugé statistiquement significatif avec une p-valeur de 5%, s'il est plus grand (ou plus petit, suivant que les valeurs remarquables de l'indice sont élevées ou au contraire faibles) que le quantile à 95% (ou 5%) de cette distribution. Ce test permet de comparer les valeurs de l'indice de flux F à celles de l'hypothèse nulle ; on s'attend à des valeurs plus grandes que celles obtenues aléatoirement, et on se positionne donc par rapport au quantile à 95%. Au contraire, pour juger de la significativité de la dispersion D, dont on s'attend à ce qu'elle soit plus faible en raison de réalité des frontières entre pays que dans l'hypothèse nulle, on se positionne donc au quantile à 5%. Enfin, les valeurs observées par l'indice S peuvent être grandes (proches de 1) ou petites (proches de -1) et on doit faire un test « *two-sided* » en se positionnant par rapport aux quantiles à 2,5% et 97,5%.

6.2.5 Recherche d'évènements fondateurs à l'aide de PhyloType

Les indices décrits ci-dessus permettent de décrire les grands flux géographiques. On décrit ici la méthode PhyloType[10] (Chevenet, Jung, de Oliveira et Gascuel, en cours de soumission), qui va nous permettre de rechercher les grands événements fondateurs expliquant l'essentiel de la pandémie.

6.2.5.1 Présentation de PhyloType

En épidémiologie, un évènement fondateur correspond à l'introduction d'un nouvel élément pathogène dans une population donnée et à sa diffusion au sein de celle-ci. Ce genre d'évènement peut être observé à l'aide d'une phylogénie. En effet, les séquences de l'agent pathogène collectées après sa diffusion sont toutes issues d'une même séquence ancestrale, celle à l'origine de l'évènement fondateur. Ainsi, les feuilles d'une phylogénie associées à ces séquences forment un clade. L'identification de clades dans une phylogénie reste, en pratique, assez aisée. Mais cela se complique lorsque le nombre de séquences dans la phylogénie est important, si les séquences étudiées proviennent de plusieurs évènements fondateurs différents, imbriqués ou non les uns dans les autres, ou si une quantité non négligeable de séquences parasites y sont mêlées, par exemple, celles provenant de plusieurs chaines de transmission marginales entre plusieurs populations, ou bien en raison d'erreurs de reconstruction.

La méthode PhyloType (Chevenet *et al*) aide à la localisation d'évènements fondateurs (clades parfaits ou imbriqués) sur une phylogénie, celle-ci doit être racinée afin de connaître l'orientation du temps, et on doit avoir la connaissance des pays de collecte associés à chaque séquence échantillonnée. Les groupes de séquences mis en valeur par ce logiciel sont appelées des *phylotypes*. Un *phylotype* est un sous-ensemble de souches dont chaque souche x et leur ancêtre commun ρ partagent la même annotation A, et telle que A est conservée le long du chemin de ρ à x (Figure 44). Par la suite nous utiliserons le terme « membre » pour désigner les souches appartenant à un phylotype. Pour faire cela, PhyloType doit connaître les annotations ancestrales associées à chaque nœud interne de la phylogénie. Il les calcule par parcimonie (ACCTRAN ou DELTRAN).

Des critères de sélection sont utilisés pour restreindre le nombre de *phylotypes* et leur garantir de fortes propriétés spécifiques. Par exemple, en limittant le nombre de séquences à l'intérieur du clade définissant le *phylotype* ayant une annotation différente de celle du *phylotype*. Le choix des critères à utiliser et leur seuil de validité sont choisis par l'utilisateur. Ils sont définis récursivement et leur complexité en temps de calcul est en $O(n)$, où n est le nombre de feuilles dans la phylogénie. Notons par P un phylotype potentiel d'annotation A, par L et R les nœuds racines de ses sous-arbres gauche

[10] Mise en œuvre dans un serveur web (http://lamarck.lirmm.fr/phylotype)

et droit respectivement et par F son nœud père. Neuf critères, plus trois qui sont des rapports entre deux autres critères, sont proposés par PhyloType :

- *size* (Sz) : ce critère correspond au nombre de membres du *phylotype* (et non au nombre de feuilles contenu dans le clade de même racine que le *phylotype*). Il est défini par

> *Size*(P, A)
> Si Annotation(P) est différent de A, alors 0
> Sinon si P est une feuille, alors 1
> Sinon *Size*(L, A) + *Size*(R, A) ;

- *different* (Df) : ce critère compte le nombre de sous-arbres et/ou de feuilles inclus dans le *phylotype* et ayant une annotation différente de ce dernier (dans le cas d'un sous-arbre, seule l'annotation à la racine est considérée). Il est défini par

Different(P, A)
Si Annotation(P) est différent de A, alors 1
Sinon si P est une feuille, alors 0
Sinon *Different*(L, A) + *Different*(R, A) ;

- *total* (Tt) : ce critère compte le nombre total de feuilles incluses dans le clade de même racine que celle du *phylotype*. Il est défini par

Total(P)
Si P est une feuille, alors 1
Sinon *Total*(L) + *Total*(R) ;

- *persistence* (Ps) : ce critère mesure le degré de conservation de l'annotation d'un *phylotype*, de la racine de celui-ci jusqu'à ses descendants. Il débute à la racine du *phylotype* et est égal au nombre minimum de générations où l'annotation est conservée dans chaque lignée. Il est défini par

Persistence(P, A)
Si (P est une feuille) ou (Annotation(L) est différent de A) ou (Annotation(R) est différent de A), alors 0
Sinon 1 + Min{*Persistence*(L, A), *Persistence*(R, A)} ;

- *local separation* (Sl) : ce critère correspond à la longueur de la branche parente à la racine du *phylotype*, c'est-à-dire la longueur de la branche qui sépare le *phylotype* du reste de la phylogénie ;
- *global separation* (Sg) : ce critère est utile lorsque la longueur de la branche parente de la racine du *phylotype* est courte, mais que les longueurs des autres branches qui séparent la racine du *phylotype* et la racine de la phylogénie sont grandes, indiquant ainsi une grande séparation du *phylotype* par rapport au reste de la phylogénie. Il est défini par

  Global_separation(P)
  Si P n'est pas la racine, alors
  Local_separation(P) + *Total*(P) × *Global_separation*(F)/*Total*(F)
  Sinon 0 ;

- *diversity* (Dv) : ce critère mesure la diversité génétique des membres du *phylotype*. Il est défini par

  Diversity(P)
  Sum(P, Annotation(P))/*Size*(P, Annotation(P)),

  avec (où l et r sont respectivement les longueurs des branches parentes à L et R)

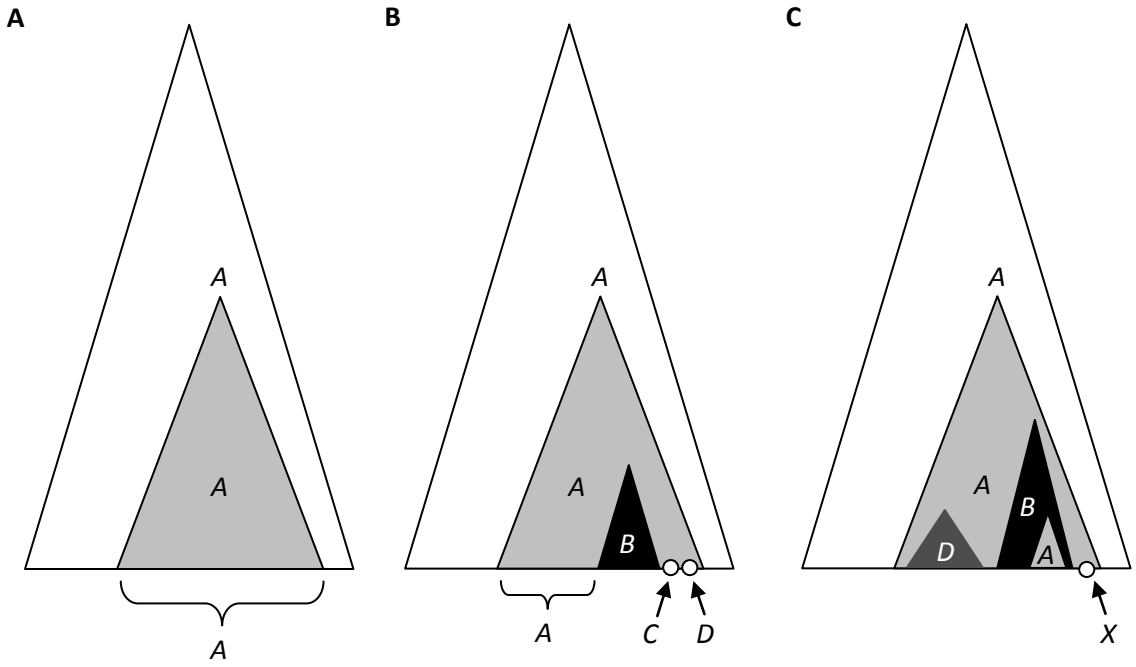
  Sum(P, A)
  Si Annotation(P) est A et si P n'est pas une feuille, alors
  Sum(L, A) + l × *Size*(L, A) + *Sum*(R, A) + r × *Size*(R, A)
  Sinon 0 ;

- *support* (Sp) : ce critère renvoie le support de branche (lorsqu'il est présent) associé à la racine du *phylotype* ;
- *global support* (SpG) : ce critère renvoie la plus forte valeur entre *support* et une pondération entre les supports des branches se trouvant sur le chemin reliant la racine du *phylotype* à celle de la phylogénie. Il est défini par

  Global_support(P)
  Si P n'est pas la racine, alors
  Max{*Support*(P), Total(P) × *Global_support*(F)/*Total*(F)}
  Sinon 0.

Les trois autres critères restant sont les rapports *size*/*different* (Sz/Df), *local separation*/*diversity* (Sl/Dv) et *global separation*/*diversity* (Sg/Dv). Par exemple, le critère *size*/*different* permet à un *phylotype* donné de contenir un nombre de sous-arbres et/ou de feuilles ayant une annotation différente et qui varie en fonction de la taille du *phylotype* donnée.

Figure 44. Exemples de *phylotypes*.
La figure A montre la définition la plus simple d'un *phylotype* (un clade) où la racine du *phylotype* et toutes ses feuilles ont la même annotation (A). La figure B montre un *phylotype* annoté A qui contient un *phylotype* annoté B et deux feuilles annotées C et D. La figure C montre un *phylotype* annoté A contenant deux *phylotypes* annotés D et B, ainsi qu'une feuille annotée X, et le *phylotype* annoté B contient un autre *phylotype* annoté A.
Extrait de Chevenet et al.



PhyloType possède aussi une procédure statistique qui permet de savoir si les résultats obtenus sont statistiquement significatifs ou non, grâce à une p-valeur associée à chaque critère (sélectionné par l'utilisateur) de chaque *phylotype*. Cette p-valeur est obtenue par ré-échantillonnage aléatoire (*shuffling*) (cf. Chapitre 1). Par exemple, pour un *phylotype* donné, si une valeur de 2 est obtenue pour le critère *size* avec une p-valeur de 3/1 000, cela signifie que 3 *shufflings* sur les 1 000 ont au moins un *phylotype* de même annotation avec une valeur supérieure ou égale à 2 pour le critère s*ize*. En pratique, si ce nombre est supérieur à 5% du nombre de *shufflings*, alors le *phylotype* résultant est généralement considéré comme non significatif.

En plus de cela, PhyloType propose une interface permettant d'enraciner la phylogénie de l'utilisateur (au cas où elle ne l'est pas) de trois manières différentes à l'aide d'un logiciel que j'ai développé pour cette occasion. La première manière positionne la racine sur le point qui minimise la variance de la distance séparant chaque feuille de la racine, de sorte à rendre la phylogénie la plus

ultramétrique possible (toutes les feuilles sont plus ou moins à égale distance de la racine). Si l'utilisateur dispose de dates de collecte associées à chaque feuille et que les souches étudiées proviennent d'une population à évolution mesurable (MEP) (Drummond *et al*, 2003b), une deuxième méthode est proposée. Il s'agit de la régression linéaire *Root-to-tip* pour laquelle la minimisation de la somme des résidus permet d'obtenir l'emplacement optimal de la racine (cf. Chapitre 2). La dernière méthode proposée localise la racine en fonction de séquences sélectionnées (supposées *outgroup*) par l'utilisateur.

6.2.5.2 Association de certains pays afin de favoriser l'apparition de *phylotypes*

PhyloType est une méthode qui met en exergue des *phylotypes* correspondant à diverses annotations et les inclusions de *phylotypes* au cœur de la phylogénie enracinée suggèrent les migrations successives du virus au cours du temps ou « chaines de transmission ». Les annotations dont les feuilles sont dispersées dans la phylogénie ont peu de chance d'être interprétées par PhyloType, puisque les *phylotypes* sont dérivés de clades dans lesquels une annotation donnée est très représentée. En revanche, PhyloType permet de grouper deux ou plusieurs annotations ensembles dans le but de favoriser l'apparition de *phylotypes* correspondant à l'union de ces annotations. Ces groupements doivent avoir un sens épidémiologique, et typiquement correspondre à des pays voisins et/ou ayant des échanges forts et identifiés.

Dans ce but, nous proposons un indice qui renseigne sur la régionalisation d'une annotation, que l'on pourra comparer avec celle de l'union de deux annotations ou plus. En accord avec PhyloType, le principe de parcimonie est utilisé (même procédé qu'à la section 6.2.3) mais en ne considérant que deux annotations dans la phylogénie (l'annotation ou les annotations étudiées a et la réunion des autres annotations $X = \neg a$). En réutilisant les mêmes notations qu'à la section 6.2.4, l'indice de régionalisation R_a, pour une annotation a, est définie par

$$R_a = \frac{N_{a\to X} + N_{X\to a} - 1}{|a| - 1},$$

où $|a|$ représente ici le nombre de feuilles annotées a. Cette mesure vaut 1 lorsque l'annotation a est totalement dispersée ($N_{a\to X} = 0$ et $N_{X\to a} = |a|$), et vaut 0 lorsqu'elle est totalement régionalisée ($N_{a\to X} = 0$ et $N_{X\to a} = 1$). Cet indice ne peut pas être utilisé avec des annotations représentées par une seule souche (le dénominateur est nul). Pour vérifier si l'union de deux annotations a et b est plus régionalisée que les deux annotations prises séparément, il suffit de comparer $R_{a\cup b}$ à $\min\{R_a, R_b\}$. Si la comparaison est favorable (c'est-à-dire $R_{a\cup b} < \min\{R_a, R_b\}$) alors le *phylotype* de l'union a plus de chance d'apparaître que ceux correspondant aux annotations a et b séparées.

En plus d'aboutir à une meilleure régionalisation, une association de pays est uniquement proposée s'ils partagent une frontière géographique (les migrations ou le commerce, et donc le transport de germe viral, en est facilité) et s'ils sont trop peu représentés pour qu'on puisse espérer obtenir des *phylotypes* pertinents en considérant chacun de ces pays pris séparément (par exemple, s'il y a moins de 20 souches par annotation et si le critère *size* est supérieur ou égal à 20).

6.2.5.3 Paramétrage de PhyloType

Trois analyses PhyloType successives sont faites pour chaque option de parcimonie disponible (ACCTRAN et DELTRAN) et en faisant varier la taille (*size*) minimale des *phylotypes* : 20, 10 et 5 ; ce qui correspond donc à des analyses plus ou moins détaillées, avec des niveaux d'exigence variables. Les trois autres critères choisis sont fixes et sont *persistence* ≥ 1, *size*/*different* ≥ 1 et *support* ≥ 70% (valeur aLRT minimum pour la branche aboutissant à la racine du *phylotype*). Mille s*hufflings* sont calculés pour chaque analyse et les *phylotypes* dont la p-valeur est supérieure à 10/1 000 (1%) pour le critère s*ize* ne sont pas considérés dans les résultats.

6.3 Résultats

6.3.1 Séquences *pol* du VIH-1C incluses dans l'étude

Les 3 081 séquences *pol*, couvrant plus de 1 000 paires de bases, de l'étude sur l'origine géographique et temporelle du VIH-1C au Sénégal (cf. Chapitre 5) sont incluses dans cette étude. À celles-ci, 528 nouvelles séquences sont ajoutées dont 219 sont collectées au Burundi, 67 en RDC, 1 en République Centrafricaine, 2 en République du Congo, 1 en Éthiopie et 238 au Swaziland.

L'ensemble des séquences provient de 63 pays différents listés dans le Tableau 4. Le continent africain reste le plus représenté ; il contient à lui seul 72% du nombre de séquences et 75% des souches collectées en Afrique proviennent de l'Afrique australe. L'Afrique du Sud et la Zambie sont toujours les deux pays les plus représentés (37% du nombre total de séquences). Aucune souche collectée en Amérique ou en Asie n'est ajoutée à cette étude. Ces continents sont donc toujours majoritairement représentés par le Brésil (253 séquences sur 299) et l'Inde (355 séquences sur 380) respectivement. Le continent européen représente seulement 9% du nombre total de séquences et aucun pays de collecte ne se démarque fortement en nombre de souches disponibles.

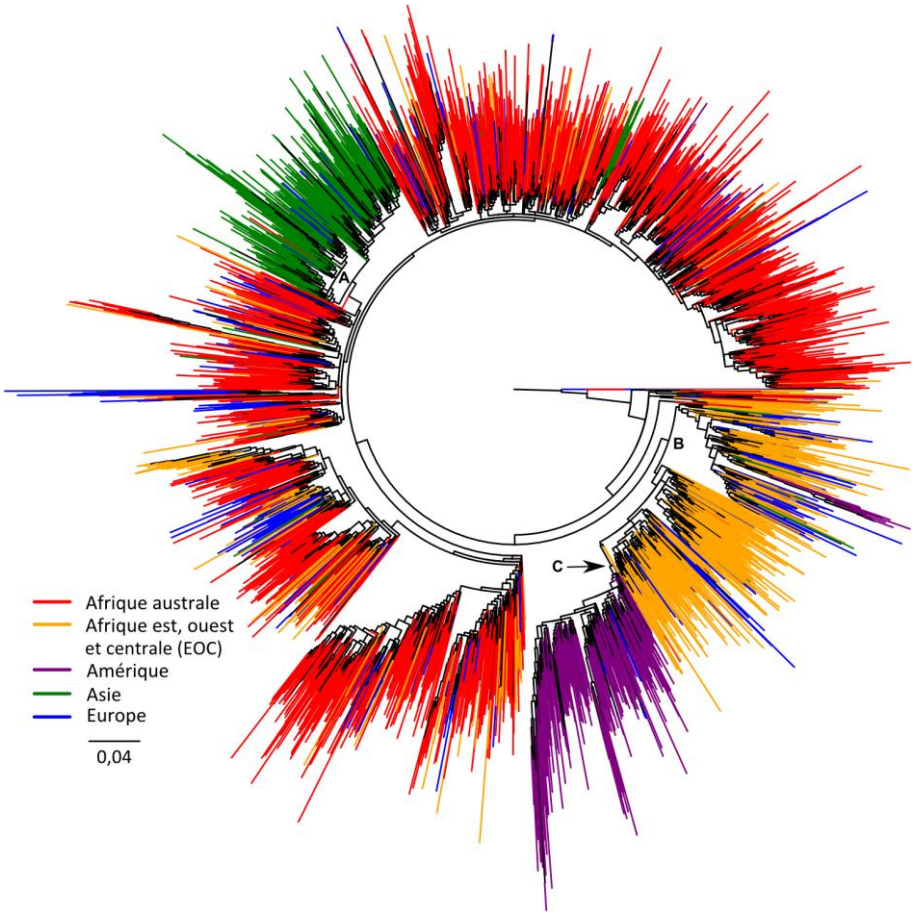
6.3.2 Phylogénie des séquences *pol* du VIH-1C

La phylogénie du maximum de vraisemblance (PhyML) des 3 609 souches du VIH-1C collectées à travers le monde est présentée à la Figure 45. Les souches sont coloriées en fonction de leur pays de collecte : l'Afrique australe (Afrique du Sud, Botswana, Malawi, Mozambique, Swaziland, Zambie et

Zimbabwe) en rouge, les pays africains de l'est, de l'ouest et centrale (EOC) en orange, l'Amérique en mauve, l'Asie en vert et l'Europe en bleu. À l'instar de la phylogénie du Chapitre 5, trois clusters importants contenant la majorité des souches de l'Amérique (essentiellement des souches du Brésil), de l'Asie (essentiellement des souches de l'Inde) et de l'Afrique EOC (majoritairement des souches du Burundi) sont observés. Ils sont respectivement supportés en valeur aLRT à 99,2% (nœud C), 74,0% (nœud A) et 86,1% (nœud B). Très peu de souches de l'Afrique australe sont visibles dans le cluster de l'Afrique EOC. Les souches collectées en Europe sont dispersées dans toute la phylogénie, sans formation de clusters importants. Aucune souche d'Afrique n'est visible à l'intérieur des clusters formés par les souches de l'Amérique et de l'Asie (hormis une souche collectée en Zambie qui se situe à l'intérieur du cluster asiatique). Ces observations suggèrent de multiples introductions du sous-type C en Europe, provenant principalement de pays africains, et une introduction majeure de ce variant, suivie d'une diffusion efficace, aussi bien en Afrique EOC (majoritairement représentée par le Burundi) que sur les continents américain (majoritairement représenté par le Brésil) et asiatique (majoritairement représenté par l'Inde).

Tout comme les souches d'Inde et du Brésil, certaines souches de pays africains forment des clusters d'intérêt au sein de la phylogénie. La Figure 46 montre la phylogénie où seulement les souches appartenant aux pays africains d'intérêt sont coloriées : l'Afrique du Sud en jaune, le Burundi en bleu, l'Éthiopie en vert, la Tanzanie en orange et la Zambie en rouge. Les souches collectées en Afrique du Sud sont disséminées dans la phylogénie mais un cluster remarquable, supporté en valeur aLRT à 88,4% (nœud A), apparaît. Cela suggère une introduction majeure du sous-type C dans ce pays, accompagnée de multiples introductions mineures. Aucune souche du Burundi n'est mélangée à l'Afrique australe et la majorité de celles-ci forme un cluster, supporté en valeur aLRT à 76,2% (nœud C), dans lequel un cluster de souches collectées en Tanzanie apparaît, supporté en valeur aLRT à 87,9% (nœud D). Ceci suggère une chaîne de transmission depuis l'origine épidémique, sans doute la Zambie (cf. ci-dessous) vers le Burundi, et ensuite la Tanzanie. Les souches de l'Éthiopie sont aussi bien retrouvées parmi des souches de l'Afrique EOC que de l'Afrique australe, mais elles y sont régionalisées (nœud B et E supportés respectivement en valeur aLRT à 88,2% et 90,6%), indiquant que l'épidémie du sous-type C en Éthiopie proviendrait de deux origines géographiques différentes. Malgré le nombre important de souches collectées en Zambie, aucun cluster important n'apparaît dans la phylogénie. Cependant, elles sont uniquement mélangées avec d'autres souches de l'Afrique australe (une souche est également vue en Asie). Ceci indique une origine possible de l'épidémie en Zambie, qui sera confirmée par les analyses suivantes basées sur nos indices et PhyloType.

Figure 45. Phylogénie basée sur le gène *pol* des 3 609 souches du VIH-1C.
Cette figure montre la phylogénie obtenue par maximum de vraisemblance (PhyML) des 3 609 souches *pol* d'une longueur de 1 011 paires de bases. Les souches appartenant au continent africain sont séparées en deux groupes. Un groupe contient toutes les souches de l'Afrique australe (en rouge), région géographique où l'épidémie du VIH-1C est très intense, et l'autre contient toutes celles de l'Afrique de l'est, de l'ouest et centrale (EOC, en orange) qui forment un clade. Le continent américain (en mauve) montre une introduction majeure du sous-type C sur ce continent, tout comme sur le continent asiatique (en vert). D'autres introductions sont visibles mais sont marginales. Les souches du continent européen (en bleu) sont mélangées dans la phylogénie, suggérant des introductions multiples de ce virus en Europe.



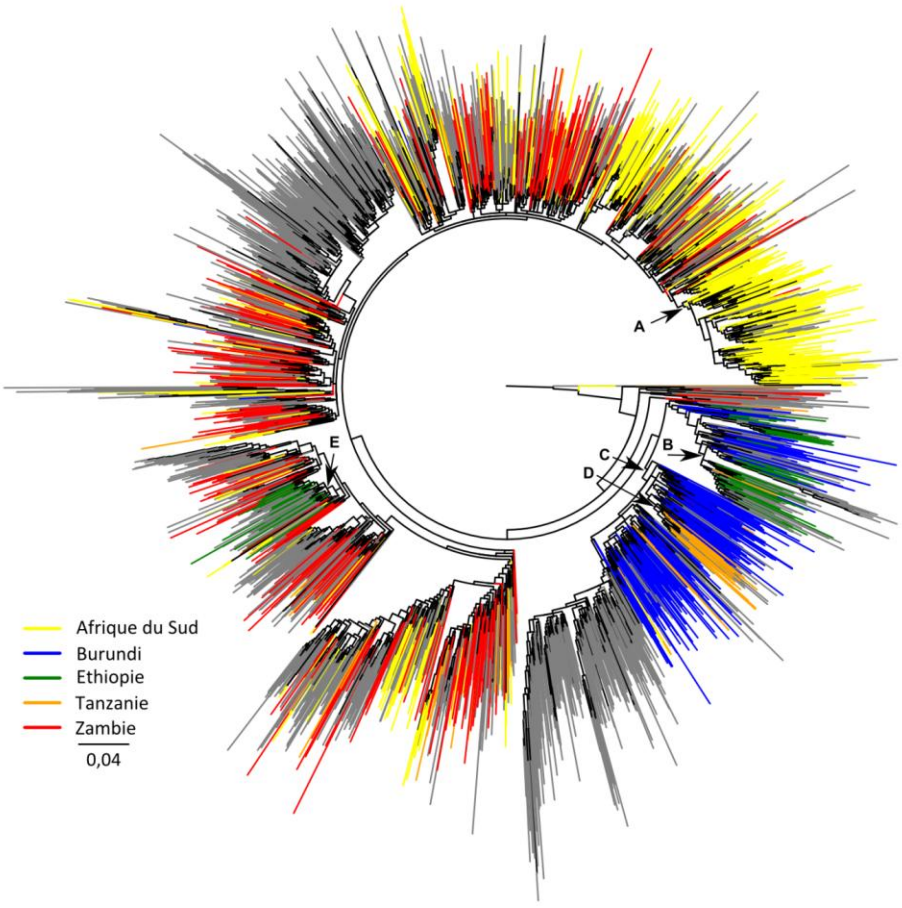
La Figure 47 montre plus en détail les souches situées à proximité de la racine de la phylogénie des 3 609 souches *pol* du VIH-1C. Les annotations ancestrales déduites des parcimonies ACCTRAN (en bleu), DELTRAN (en rouge) et DOWNPASS (en vert) sont reportées sur chaque nœud interne de la phylogénie ; les estimations communes sont en noir. Les souches 83FR-HXB2 (*outgroup*) et 02ZA-1752 (*rogue taxa*) sont écartées de l'étude (cf. section 6.2.3). L'annotation correspondante à l'ancêtre commun de toutes les souches du VIH-1C est donc RDC/Tanzanie/Zambie et cela quelle que soit la parcimonie utilisée. Les annotations RDC et Tanzanie sont dues au groupe de trois souches (deux collectées en RDC et une en Tanzanie) qui sont ancestrales aux autres souches du VIH-1C. Hormis ce groupe de trois séquences et pour les trois parcimonies, l'annotation résultante à l'ancêtre commun des autres souches du VIH-1C est la Zambie. Les différentes parcimonies renvoient des estimations assez similaires sur les autres nœuds proches de la racine, seulement six nœuds sont discordants. Enfin, un cluster remarquable de treize souches collectées en RDC, et supporté à 63,7% en aLRT, apparaît (nœud A). L'origine géographique de l'épidémie du VIH-1C est donc indéterminée, elle

est soit en RDC, soit en Zambie ou soit en Tanzanie. Notons que ces trois pays africains sont limitrophes.

6.3.3 Étude des flux migratoires du VIH-1C

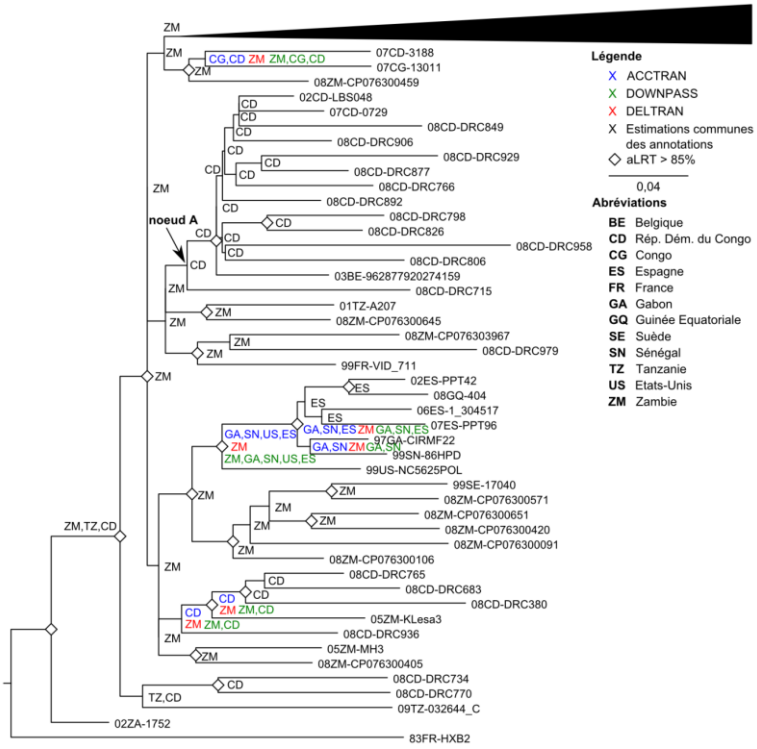
Afin de synthétiser les flux migratoires du VIH-1C de la phylogénie, nous avons développé trois indices basés sur les transitions entre pays. Ces transitions sont obtenues de trois manières différentes, en utilisant les algorithmes de parcimonie RandACCTRAN, RandDELTRAN et RandDOWNPASS (cf. section 6.2.3). Une méthode de ré-échantillonnage aléatoire, le *shuffling*, est utilisée pour mesurer la significativité statistique de ces mesures. Seuls les résultats correspondant à la parcimonie RandDOWNPASS sont présentés dans ce chapitre ; cette méthode étant la moins arbitraire des trois. Les résultats des autres parcimonies sont disponibles dans l'Annexe A.

Figure 46. Phylogénie basée sur le gène *pol* des 3 609 souches du VIH-1C (zoom sur les pays africains d'intérêt).
Cette figure montre la phylogénie des 3 609 souches où seulement les souches appartenant aux pays d'intérêt sont coloriées. Les souches appartenant à l'Afrique du Sud sont en jaune, celles du Burundi en bleu, celles de l'Éthiopie en vert, celles de Tanzanie en orange et celles de la Zambie en rouge. Les souches des autres pays sont grisées. La majorité des souches du Burundi et de l'Afrique du Sud forment deux clusters importants. Les souches de Tanzanie forment un cluster au sein des souches du Burundi et les souches de l'Éthiopie forment deux clusters. Le premier est à proximité des souches du Burundi et l'autre parmi les souches de l'Afrique australe. Enfin, les souches de Zambie sont ubiquitaires au sein des souches de l'Afrique australe.



Nous présentons d'abord les graphiques des mesures obtenues par les trois indices, puis nous discutons des résultats.

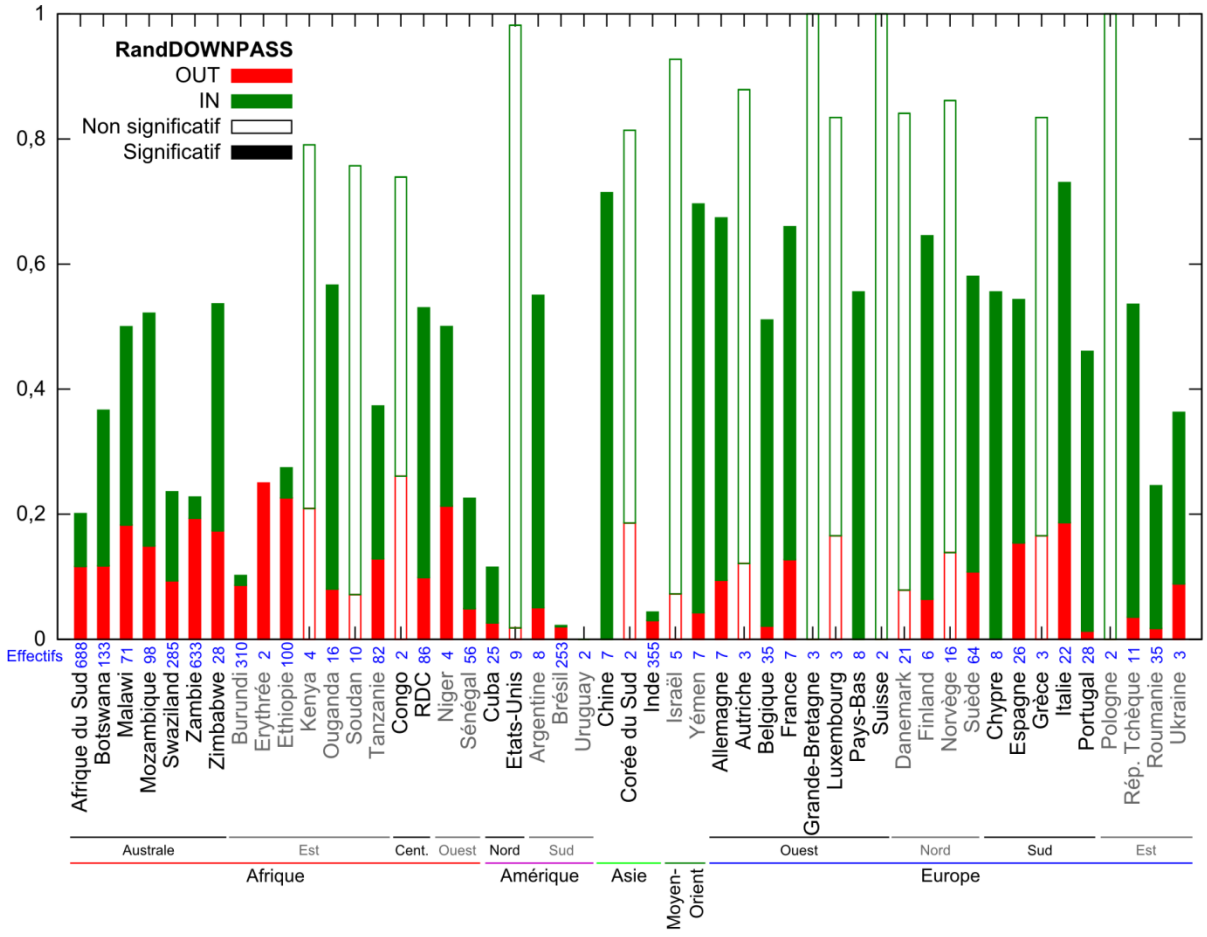
Figure 47. Souches à proximité de la racine.
Phylogénie du maximum de vraisemblance (PhyML) des 3 609 souches du VIH-1C où uniquement les souches à proximité de la racine sont représentées. Les annotations ancestrales inférées par les parcimonies ACCTRAN (en bleu), DELTRAN (en rouge) et DOWNPASS (en vert) sont indiquées sur chaque nœud. Les estimations communes sont en noir. Les nœuds avec un losange blanc indiquent des valeurs supports (aLRT) plus grandes que 85%. Chaque nom de souche est précédé de l'année de collecte (par exemple, 02 pour 2002 ou 99 pour 1999) puis d'une abréviation représentant le pays de collecte (SN pour Sénégal, etc.). La liste complète des abréviations utilisées est présentée dans la figure.



La Figure 48 montre les estimations de l'indice de dispersion D pour chaque pays étudié représenté par plus d'une souche et pour la parcimonie RandDOWNPASS. Ces pays sont organisés par région géographique, puis par ordre alphabétique. Les deux composantes IN et OUT de cet indice sont respectivement représentées en vert et rouge. Les mesures non significatives ont un intérieur vide (obtenus en se positionnant par rapport aux quantiles à 5%) et les mesures significatives sont représentées par une barre pleine. Le nombre de souches associées à chaque pays est donné en bleu en abscisse. Plus l'indice de dispersion est grand (proche de 1), plus les souches sont éparpillées dans la phylogénie, plus il est petit (proche de 0), plus elles sont régionalisées et forment un clade lorsque cet indice vaut 0. Par exemple, l'indice de dispersion de l'Uruguay vaut zéro, ce qui signifie que les souches forment une cerise (cette phylogénie ne contient que deux souches collectées en Uruguay), tandis que les souches collectées en Pologne sont isolées les une des autres (l'indice vaut 1). Par définition de l'indice de dispersion, un pays donneur ne peut être totalement dispersé dans la phylogénie. Les Figures 54 et 55 de l'Annexe A correspondent aux mesures de l'indice de dispersion pour les méthodes de parcimonie RandACCTRAN et RandDELTRAN respectivement. Outre le fait de rensei-

gner sur la régionalisation des souches dans la phylogénie (hauteur des barres), comme pour l'Inde et le Brésil, cet indice permet aussi d'identifier les pays fortement donneurs (donc probablement à l'origine de la diffusion de l'épidémie) et les pays fortement receveurs (donc au bout d'une chaîne de transmission). On voit ici nettement que les pays africains sont plutôt donneurs, donc à l'origine de la diffusion de l'épidémie du sous-type C, alors que, par exemple, les pays européens, sont globalement receveurs. Cependant, cet indice n'indique pas vers quels pays est transmise l'épidémie ou de quels pays elle provient ; c'est le rôle de l'indice de flux.

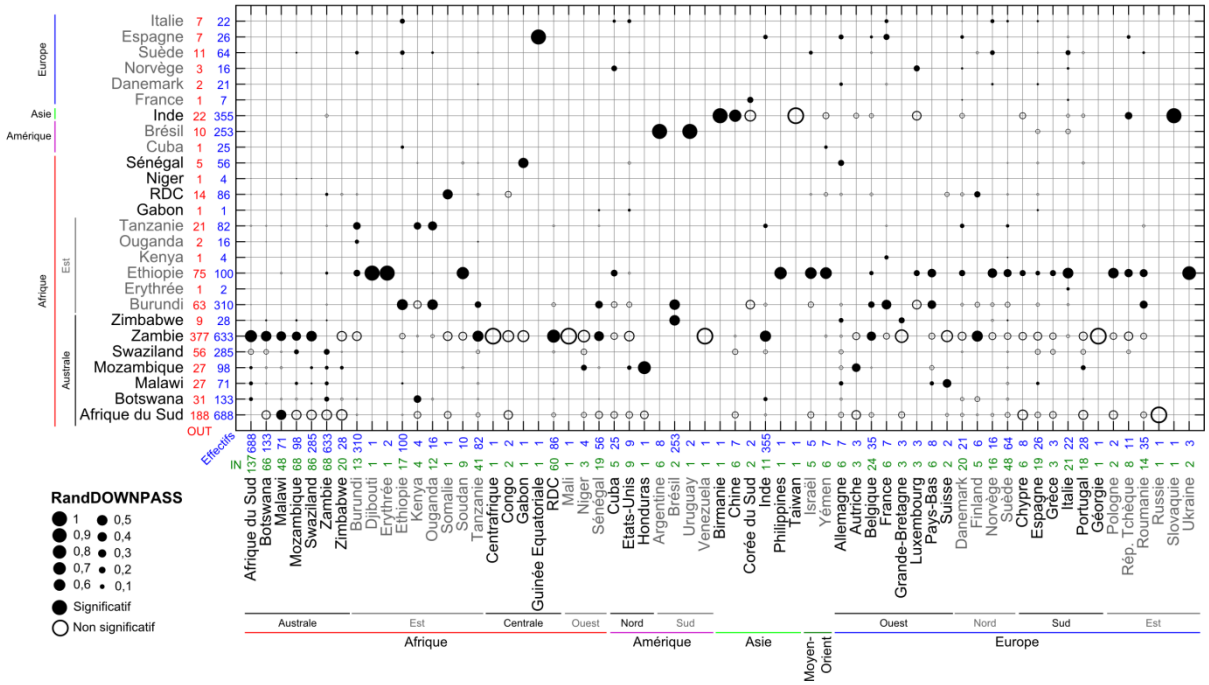
Figure 48. Estimations de l'indice de dispersion avec la parcimonie RandDOWNPASS.
Le graphique indique, pour chaque pays de collecte ayant au moins deux souches, les valeurs correspondantes à l'indice de dispersion IN (en vert) et OUT (en rouge) ; leur somme correspondant à l'indice de dispersion totale. Ces résultats sont obtenus avec la parcimonie RandDOWNPASS. Les mesures significatives de l'indice de dispersion sont représentées par des barres pleines tandis que les mesures non significatives par des barres vides. Les pays sont regroupés par zone géographique, puis par ordre alphabétique. Une mesure totale de 1 signifie que les souches sont totalement dispersées dans la phylogénie, sans possibilité de formation d'annotations ancestrales (par exemple la Grande-Bretagne). Une mesure de zéro signifie que toutes les souches d'un pays forment un clade monophylétique (seul exemple, l'Uruguay). Pour chaque pays, le nombre de souches présentes dans la phylogénie est rappelé en bleu en abscisse.



La Figure 49 indique les estimations de l'indice de flux F obtenues pour chaque couple de pays avec la parcimonie RandDOWNPASS. Chaque point reflète la proportion des transitions reçues par le pays en abscisse, du pays en ordonnée. Par exemple, le Brésil a deux introductions une provenant du Burundi et l'autre du Zimbabwe, tandis que la Roumanie ne donne qu'une seule fois vers la Grèce. Les

cercles vides représentent les mesures non significatives (obtenus en se positionnant par rapport aux quantiles à 95%) et les cercles pleins les mesures significatives. Par définition, la somme des mesures d'une colonne vaut 1. Le nombre de souches (en bleu), le nombre de transitions OUT (en rouge) et le nombre de transitions IN (en vert) sont indiqués pour chaque pays en dessous ou à côté des axes correspondants. Seuls les pays dont le nombre de transitions OUT est supérieur ou égal à 1 sont représentés en ordonnée. Les pays sont classés de la même façon qu'à la Figure 48. Les résultats des méthodes de parcimonie RandACCTRAN et RandDELTRAN sont disponibles aux Figures 56 et 57 de l'Annexe A. Cet indice est utile pour identifier une partie des chaînes de transmission, par exemple, de quel(s) pays est venue l'épidémie du VIH-1C en Éthiopie ? Et à quelle proportion ? On voit pour l'Éthiopie une source principale significative issue du Burundi, une deuxième source moindre non significative issue de la Zambie, et des sources accessoires venant de nombreux pays (dont l'Europe). Cet indice permet aussi d'identifier l'épicentre de l'épidémie du sous-type C : c'est le pays qui est le plus souvent donneur, donc probablement la Zambie.

Figure 49. Estimations de l'indice de flux avec la parcimonie RandDOWNPASS.
Chaque point sur le graphique reflète la proportion de transitions IN pour le pays en abscisse issues du pays en ordonnée. La somme des points d'une colonne vaut 1. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique. Le nombre de souches de chaque pays est indiqué sur les deux axes en face des pays concernés. Sur l'axe des abscisses, la mesure IN indique le nombre de transitions entrantes dans ce pays. Sur l'axe des ordonnées, la mesure OUT indique le nombre de transitions sortantes de ce pays. Les mesures significatives sont représentées par un cercle plein et les mesures non significatives par un cercle vide. Les pays avec un nombre de transitions OUT inférieur à 1 ne sont pas représentés en ordonnée. Le graphique présenté correspond à la parcimonie RandDOWNPASS.



La Figure 50 montre les mesures obtenues par l'indice de symétrie S avec la parcimonie RandDOWNPASS. Cet indice renseigne sur la symétrie des échanges entre pays donneurs. Par commodité, la mesure reportée sur le graphique entre un couple d'annotations a et b correspond à $S_{a\leftrightarrow b}/(N_{a\to b}+$

$N_{b\to a}$). Lorsque le point est rouge (respectivement bleu) il y a plus de mouvement du pays en ordonnée vers le pays en abscisse (resp. du pays en abscisse vers le pays en ordonnée) que l'inverse. Les cercles vides représentent des échanges non statistiquement supportés (obtenus en se positionnant par rapport aux quantiles à 2,5% et 97,5%) et les cercles pleins de mesures statistiquement significatives. Les croix indiquent qu'il n'y a aucune transition entre le couple de pays en question. Plus la taille des points est proche de 1, plus les échanges sont asymétriques, tandis que plus la taille des points est proche de zéro plus les échanges sont symétriques. Par exemple, l'échange entre la Zambie et le Danemark est parfaitement asymétrique (le flux migratoire va uniquement de la Zambie vers le Danemark), tandis que l'échange entre le Niger et l'Afrique du Sud est parfaitement symétrique (autant de flux migratoires du Niger vers l'Afrique du Sud que de l'Afrique du Sud vers le Niger). Les pays sont organisés de la même manière qu'aux Figures 47 et 48. Le nombre de souches correspondant à chaque pays est rappelé en bleu en abscisse et en ordonnée. Le graphique est évidemment symétrique (cf. section 6.2.4). Les Figures 58 et 59 de l'Annexe A correspondent aux estimations de l'indice de symétrie des méthodes de parcimonie RandACCTRAN et RandDELTRAN respectivement. Cet indice est utile pour identifier le pays ou les pays à l'épicentre de l'épidémie puisque le flux doit logiquement être plus important en sortie qu'en entrée, et, dans notre cas, deux pôles impotants sont clairement identifiés : la Zambie dont les flux sont tous majoritairement sortants, même si certains ne sont pas significatifs, et l'Afrique du Sud. Remarquons tout de même que la Zambie est bien l'épicentre de l'épidémie (en accord avec les observations précédentes) puisque le flux épidémique est plus intense de la Zambie vers l'Afrique du Sud.

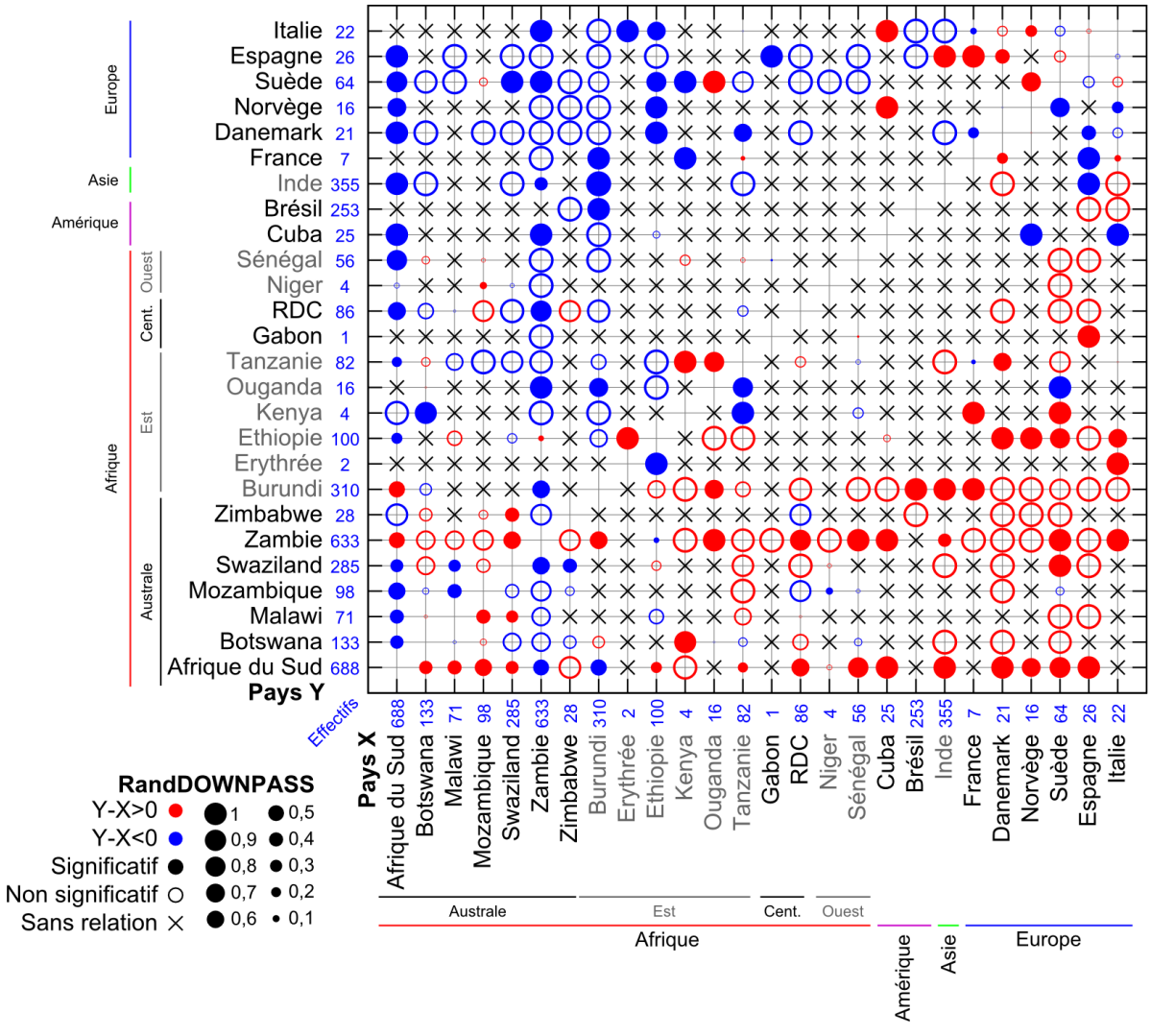
Origine de l'épidémie du sous-type C du VIH-1

L'origine géographique de l'épidémie du VIH-1C est déterminée par l'annotation associée au nœud racine de la phylogénie et les différentes méthodes de parcimonie s'accordent sur la même incertitude en hésitant entre la Zambie, la Tanzanie et la RDC (cf. section 6.3.2). Toutefois, il reste possible d'identifier l'épicentre de cette épidémie pouvant, à l'instar de l'épidémie du VIH-1 (cf. Chapitre 5), être différent du pays d'origine. L'épicentre de l'épidémie est délicat à observer puisque l'annotation correspondante se situe sur les nœuds internes de la phylogénie, à proximité de la racine et suffisamment éloigné des feuilles de celle-ci.

Les mesures de l'indice de dispersion obtenues par la méthode de parcimonie RandDELTRAN montrent que les pays donneurs sont majoritairement des pays africains localisés dans la région australe et est (12 sur 19 donneurs), suggérant que l'épicentre de l'épidémie se situe en Afrique (Figure 56 de l'Annexe A). Le nombre de transitions OUT et le nombre de pays différents vers lesquels la Zambie donne suggèrent que ce pays est l'épicentre de l'épidémie du sous-type C du VIH-1 (518 tran-

sitions OUT sur un total de 976, soit 53% du nombre total de transitions, vers 46 pays sur 62, soit 74% ; Figure 58). Notons toutefois que seule une portion de ces mesures est significative (12 flux sur 46, dont 5 vers d'autres pays de l'Afrique australe), sans doute en raison du nombre très important de souches de Zambie (633). Ce résultat est confirmé par l'indice de symétrie qui indique que la Zambie est le seul pays où le flux migratoire est le plus important en sortie qu'en entrée, quel que soit le deuxième pays auquel on se réfère (Figure 59). L'Afrique du Sud peut être considérée comme un deuxième pôle épidémique, puisque, pour ce pays aussi, le flux est plus important en sortie qu'en entrée. Mais le flux entre la Zambie et l'Afrique du Sud reste plus important du premier vers le second pays, confortant l'hypothèse d'un épicentre en Zambie.

Figure 50. Estimations de l'indice de symétrie avec la parcimonie RandDOWNPASS.
Ce graphique renseigne sur la symétrie des échanges entre les pays donneurs pour la parcimonie RandDOWNPASS. Pour en faciliter la lecture, la mesure $S_{a\leftrightarrow b}/(N_{a\to b}+N_{b\to a})$ est reportée sur le graphique pour tout a et b. Si la mesure est représentée par un point rouge (resp. bleu) cela signifie qu'il y a plus de mouvement du pays en ordonnée (resp. abscisse) vers le pays en abscisse (resp. ordonnée), que l'inverse. Lorsque la mesure vaut 1 et que le point est rouge (resp. bleu), seuls des mouvements du pays Y (resp. X) vers le pays X (resp. Y) sont observés. Lorsqu'elle vaut 0, l'échange est symétrique. Les cercles vides montrent les mesures non significatives et les cerlces pleins les mesures significatives. Les croix indiquent deux pays sans relations. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique, et leur nombre de souches est rappelé sur les axes. Seuls les pays dont le nombre de transitions OUT est supérieur à 1 sont représentés.



L'indice de dispersion de la méthode de parcimonie RandDOWNPASS montre aussi que, pour les pays africains, il y a un flux plus important en sortie qu'en entrée (100% des pays africains sont donneurs), suggérant donc également que l'épidémie est originaire du continent africain (Figure 48). L'indice de flux montre, tout comme celui de la méthode RandDELTRAN, que la Zambie est le pays donnant le plus souvent (377 transitions OUT sur un total de 965, soit 39% du nombre total de transitions, vers de nouveau 46 pays sur 62, soit 74%, Figure 49), mais cette fois-ci 11 flux (sur les 46) sont significatifs, au lieu de 12, et ils ne correspondent pas forcément au même pays receveur. Par exemple, avec RandDELTRAN (resp. RandDOWNPASS), le flux de la Zambie vers le Mozambique (resp. Belgique) est significatif alors qu'il ne l'est pas avec RandDOWNPASS (resp. RandDELTRAN). L'indice de symétrie (Figure 50), quant à lui, montre qu'à nouveau la Zambie est le seul pays dont les flux sont plus importants en sortie qu'en entrée (sauf avec l'Éthiopie ou les flux sont quasiment symétriques), mais la majorité des points (19 sur 24) est non significative. Une observation similaire à celle apportée précédemment à l'Afrique du Sud peut être faite.

Une conclusion identique (Zambie épicentre de l'épidémie du sous-type C) est obtenue avec la méthode de parcimonie RandACCTRAN, dont les résultats sont assez semblables à ceux de la méthode de parcimonie RandDOWNPASS. Cela provient du fait que les nombres de nœuds internes ambigus résultant de l'application des parcimonies RandDOWNPASS (239 nœuds ambigus) et RandACCTRAN (119 nœuds ambigus) sont tous deux élevés, alors que ce nombre est bien plus faible avec la parcimonie RandDELTRAN (11 nœuds ambigus). Les deux premières méthodes sont donc proches, car elles utilisent largement les choix aléatoires, alors que ceux-ci sont bien plus rares avec RandDELTRAN. Notons toutefois qu'avec RandACCTRAN, il y a 335 transitions OUT pour la Zambie sur 969, soit 35% du nombre total de transitions OUT (Figure 57), contre 39% avec RandDOWNPASS (et 53% avec RandDELTRAN). C'est un effet attendu, compte tenu des choix (plus ou moins arbitraires) effectués par ACCTRAN et DELTRAN de « pousser » les transitions vers les feuilles ou la racine, alors que DOWNPASS ne tranche pas. Malgré ces différences quantitatives, les trois méthodes s'accordent sur l'essentiel, à savoir que la Zambie est probablement l'épicentre de l'épidémie du sous-type C.

Flux migratoires du sous-type C du VIH-1

L'observation des flux migratoires se fait à l'aide de l'indice de flux qui indique précisément que tel pays donne à tel autre pays. Globalement, les mesures de cet indice montrent de fortes interactions entre les pays de l'Afrique australe ou pratiquement chaque pays donne et reçoit de tous les autres pays de l'Afrique australe. Une observation identique peut être faite avec les pays de l'Afrique de l'est, mais le signal est moins soutenu. Ces deux observations indiquent que les échanges semblent se faire principalement entre pays géographiquement proches. Les pays européens reçoivent

majoritairement des pays africains, mais des échanges marginaux entre pays européens sont aussi à noter. Précisément, les flux significatifs entre pays représentés par au moins 20 souches et pour lesquels le pays en entrée a au moins 10 transitions IN, sont présentés dans le Tableau 5. Ces flux indiquent donc des mouvements épidémiques importants (plusieurs introductions) entre deux pays, qui peuvent être à l'origine d'évènements fondateurs ou de cas isolés (les conséquences de ces introductions ne peuvent pas être connues avec cet indice). Les résultats montrent bien les pays ayant des liens épidémiques forts, notamment les pays de l'Afrique australe (11 flux sur 19 sont identifiés entre deux pays de l'Afrique australe).

Tableau 5. Flux significatifs déduits des mesures de l'indice de flux.
Les flux présentés dans ce tableau proviennent de pays représentés par au moins 20 souches, avec un minimum de 10 transitions IN. Le nombre de transitions IN est indiqué pour chaque algorithme de parcimonie, et surligné en gris lorsque le flux est significatif. Seuls les flux significatifs pour au moins un algorithme sont indiqués. Les mesures entre parenthèses ne respectent pas la condition du minium de 10 transitions IN, mais sont données à titre indicatif.

Pays		Nombre de transitions IN		
De	Vers	RandDELTRAN	RandACCTRAN	RandDOWNPASS
Zambie	Afrique du Sud	114	80	87
	Botswana	51	27	32
	Malawi	28	17	20
	Mozambique	41	21	25
	Swaziland	54	39	43
	Tanzanie	28	20	22
	RDC	52	43	45
	Inde	10	(4)	(6)
	Sénégal	12	(7)	(8)
Afrique du Sud	Malawi	22	20	22
Botswana	Zambie	(1)	10	(1)
	Afrique du Sud	10	(9)	(9)
Malawi	Afrique du Sud	(6)	11	(6)
Swaziland	Afrique du Sud	11	19	16
	Zambie	(6)	13	10
Éthiopie	Suède	15	14	14
	Italie	12	10	10
Burundi	Éthiopie	11	(8)	(9)
	Tanzanie	11	(6)	(7)

On voit à nouveau dans le Tableau 5 une origine ou épicentre probable en Zambie, avec cependant des flux significatifs de retour vers la Zambie de souches venant du Botswana et du Malawi. La présence de flux IN et OUT significatifs est également visible entre le Malawi et l'Afrique du Sud qui est la plaque tournante du transport (notamment aérien) en Afrique australe et donc logiquement aussi pour la diffusion de l'épidémie. On retrouve des transmissions connues (par exemple du Burundi vers l'Éthiopie) ou historiquement explicable (par exemple l'Éthiopie vers l'Italie).

Le Tableau 6 donne les mesures entre les pays a et b donneurs étant chacun représentés par plus de 20 séquences et pour lesquels le nombre de transitions $N_{a\to b} + N_{b\to a}$ est supérieur ou égal à 10, afin d'observer la tendance du mouvement de l'épidémie entre ces pays (plutôt de a vers b ou plutôt

de b vers a ?). Ces mesures montrent presque tout le temps des mouvements unidirectionnels, par exemple, le flux migratoire est plus intense en sortie de la Zambie qu'en entrée (>0,4), confirmant un mouvement épidémique de la Zambie vers l'Afrique australe (Afrique du Sud, Botswana et Swaziland), la RDC et l'Inde. Les mouvements épidémiques avec l'Afrique du Sud sont nettement moins unidirectionnels, en particulier avec RandACCTRAN (<0,4), suggérant des échanges épidémiques réguliers, dans les deux sens, avec les pays de l'Afrique australe (Botswana, Malawi et Swaziland), comme déjà discuté ci-dessus. Il y a cependant une exception avec le Mozambique pour lequel le flux est largement unidirectionnel (>0,5). Enfin, les mouvements de l'Éthiopie vers les pays européens (Italie et Suède) sont aussi unidirectionnels (>0,5), tout comme ceux du Burundi vers les pays de l'Afrique de l'est (Éthiopie et Tanzanie ; >0,6), suggérant une chaîne de transmission du Burundi vers l'Éthiopie, puis de l'Éthiopie vers la Suède et Italie (cf. ci-dessus).

Tableau 6. Mesures significatives et remarquables de l'indice de symétrie.
Ce tableau présente les mesures de l'indice de symétrie entre les pays a et b donneurs à fort effectif (>20 séquences) et pour lesquels le nombre de transitions $N_{a\to b} + N_{b\to a} \geq 10$. La mesure correspondante à l'indice de symétrie est présentée pour chaque algorithme de parcimonie, et surlignée en gris lorsqu'elle est significative. Seules les mesures significatives pour au moins un algorithme sont indiquées. Les mesures entre parenthèses ne respectent pas la condition $N_{a\to b} + N_{b\to a} \geq 10$ et sont données à titre indicatif.

Pays		Mesure de l'indice de symétrie		
De	Vers	RandDELTRAN	RandACCTRAN	RandDOWNPASS
Zambie	Afrique du Sud	0,69	0,43	0,51
	Botswana	0,96	0,44	0,96
	RDC	0,99	0,80	0,87
	Swaziland	0,80	0,50	0,61
	Inde	0,67	(0,19)	(0,67)
Afrique du Sud	Botswana	0,11	0,37	0,35
	Malawi	0,57	0,27	0,40
	Swaziland	0,48	0,26	0,33
	Mozambique	0,72	0,50	0,59
Éthiopie	Italie	-	0,59	0,69
	Suède	0,76	0,86	0,79
Burundi	Éthiopie	0,69	0,47	0,55
	Tanzanie	0,69	0,25	0,41

Les mesures de flux présentées ici sont globales. Elles peuvent résulter d'un évènement fondateur unique et bien visible (par exemple du Brésil vers l'Uruguay), ou bien de transmissions multiples, sans qu'on discerne clairement les effets fondateurs, s'il y en a (par exemple entre la Zambie et l'Afrique du Sud). L'utilisation du logiciel PhyloType va précisément relever les chaînes de transmission complètes issues d'évènements fondateurs.

6.3.4 Recherche des chaînes de transmission majeures du VIH-1C avec PhyloType

Les chaînes de transmission du VIH-1C sont déterminées à l'aide de la méthode PhyloType qui met en exergue des *phylotypes*, reflets d'évènements fondateurs probables, mais surtout les liens qui les unissent. Ces liens sont difficilement observables avec les indices présentés ci-avant, en particulier si les chaînes de transmission traversent plus de deux pays ou correspondent à des flux faibles mais essentiels (fondateurs). Cette analyse vient donc en complément de celles présentées précédemment. Mais avant cela, nous proposons une analyse visant à déterminer quels regroupements peuvent être faits afin d'intégrer le maximum d'information (de feuilles) dans l'analyse PhyloType, puisque nous nous limiterons à des *phylotypes* d'une certaine taille, excluant d'office certaines annotations peu représentées.

6.3.4.1 Associations d'annotations pour l'analyse avec PhyloType

Les chaînes de transmission de l'épidémie du VIH-1C sont déduites de la phylogénie des 3 609 séquences à l'aide du logiciel PhyloType. Auparavant, les souches des pays peu régionalisées (probabilité faible de formation de *phylotypes*) et en faible effectif (représentés par moins de 20 souches) sont étudiées afin de proposer des associations d'annotations permettant l'émergence de *phylotypes* portant sur ces combinaisons d'annotation. Sans ces regroupements, il y a peu de chance qu'un *phylotype* avec ces annotations apparaisse et, donc, l'information fournie par ces souches est perdue. La Figure 51 présente l'indice de régionalisation R (cf. section 6.2.5.2) appliquée pour chaque pays africain (en bleu) et pour chaque paire de pays africains (en vert et rouge). Des tableaux similaires sont donnés en annexe pour les autres continents. Lorsque le point est rouge la régionalisation est strictement inférieure au minimum des régionalisations des deux pays. Une croix noire symbolise deux pays ayant une frontière géographique commune, règle importante dans le choix des associations. On trouve ainsi 20 points remarquables (rouges), pouvant relever d'une association. Sur ceux-ci, 15 sont observés sur des couples de pays appartenant à la même région géographique (Afrique australe, est, ouest, etc.) et 13 correspondent à un couple de pays partageant une frontière géographique. Ce résultat est particulièrement frappant et montre que la géographie des pays africains et la phylogénie sont fortement corrélées, confirmant ainsi que les échanges se font avant tout entre pays proches et que nos indices sont à même de détecter ce signal, bien que simples et basés sur des calculs rapides de parcimonie. Aucune différence entre les résultats obtenus par ACCTRAN, DELTRAN ou DOWNPASS n'est à noter puisque, pour cet indice, seulement deux annotations sont considérées (a versus $\neg a$, cf. section 6.2.5.2). Cela a pour effet que le nombre total de transitions de chaque méthode de parcimonie est égal. Plusieurs associations sont suggérées par ce graphique, comme la République Démocratique du Congo (RDC) avec la République du Congo ou encore la Tanzanie avec

l'Ouganda et/ou le Kenya. Remarquons qu'avec ce dernier cas, si nous souhaitons associer ces trois pays ensembles, il est nécessaire de calculer l'indice de régionalisation correspondant à l'union des trois annotations afin de le comparer aux indices de régionalisation de la Tanzanie, de l'Ouganda et du Kenya de manière à garantir un gain global en régionalisation.

À cet effet, le Tableau 7 récapitule la liste de toutes les associations successives suggérées par l'indice de régionalisation R, mais uniquement pour les pays à faible effectif (<20) et pour ceux qui partagent une frontière géographique. Les lignes grisées indiquent les associations retenues pour l'analyse PhyloType où le critère *size* est supérieur ou égal à 20, tandis que les autres lignes indiquent des associations temporaires ayant permis de les obtenir. Ce tableau est construit itérativement de la manière suivante :

1. Chaque annotation ayant un faible effectif (<20) est associée avec celles partageant une frontière géographique ;
2. L'indice de régionalisation est calculé pour les deux annotations et pour leur union ;
3. Si l'indice de l'union indique une meilleure régionalisation alors l'union de ces deux annotations est considérée par la suite, sinon l'association n'est pas retenue ;
4. Une fois la procédure finie, la première étape est répétée avec les nouvelles associations et cela jusqu'à convergence.

Les associations retenues par cette procédure sont donc :

- le Congo et la RDC ;
- la Birmanie, la Chine et l'Inde ;
- l'Espagne et la France ;
- l'Argentine et le Brésil ;
- le Kenya, l'Ouganda et la Tanzanie ;
- la Norvège et la Suède ;
- Djibouti, l'Érythrée, l'Éthiopie et le Soudan.

Toutefois, certaines de ces associations perdent leur intérêt lorsque le critère *size* est supérieur ou égal à 10, puisque certaines annotations sont alors suffisamment représentées. Ainsi, les associations retenues lorsque le critère *size* est supérieur ou égal à 10 sont :

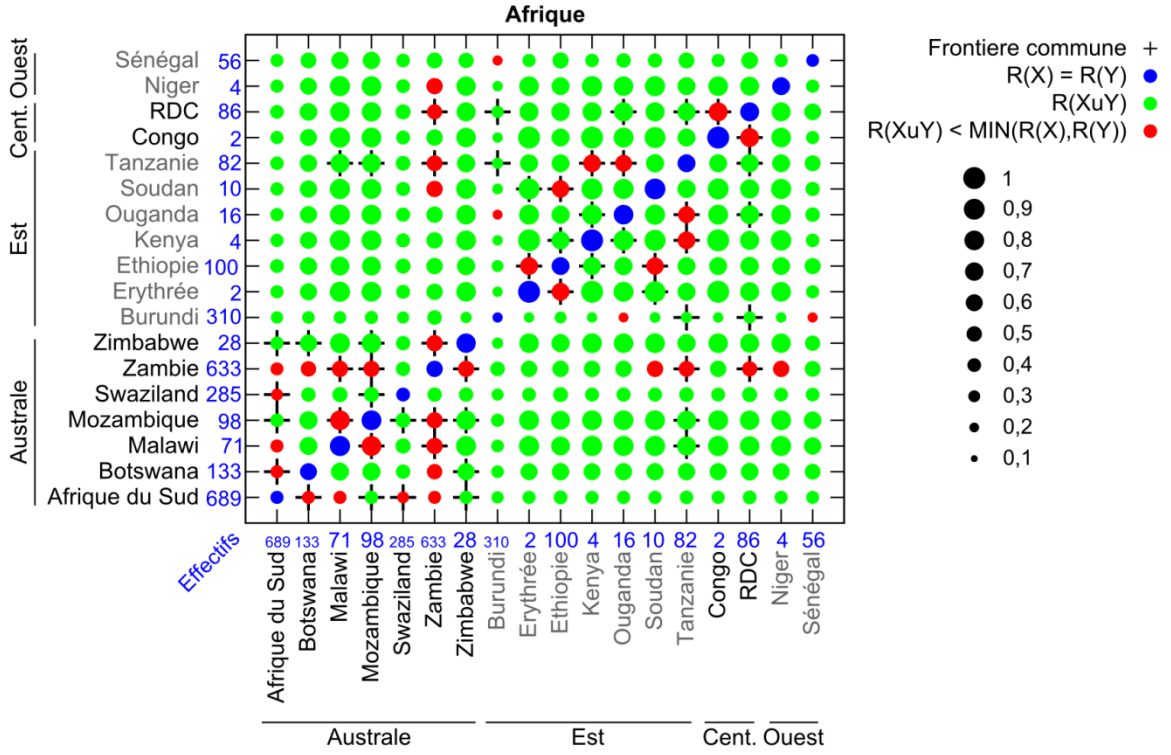
- le Congo et la RDC ;
- la Birmanie, la Chine et l'Inde ;
- l'Espagne et la France ;

- l'Argentine et le Brésil ;
- le Kenya et la Tanzanie ;
- l'Érythrée et l'Éthiopie.

Et lorsqu'il est supérieur ou égal à 5 :

- le Congo et la RDC ;
- la Birmanie et l'Inde ;
- le Kenya et la Tanzanie ;
- l'Érythrée et l'Éthiopie.

Figure 51. Estimations de l'indice de régionalisation entre les souches de pays africains.
Ce graphique renseigne sur la possibilité de grouper deux pays ensembles lors de l'analyse avec PhyloType. Les couples de pays ayant une croix sont ceux qui partagent une frontière géographique commune. Les points en bleu indiquent l'indice de régionalisation. Les points verts et rouges indiquent l'indice de régionalisation de l'union des deux pays situés sur l'axe des ordonnées et des abscisses. Lorsque ce point est en rouge la régionalisation de l'union est meilleure que la régionalisation des deux pays pris séparément. Par exemple, la mesure pour le couple RDC/Congo indique que l'union des deux pays est plus régionalisée (point rouge) que celle des pays pris séparément. De plus, ces deux pays partagent une frontière géographique commune (une croix), il est donc conseillé de les grouper ensembles lors de l'analyse avec PhyloType afin de maximiser les chances d'apparition de *phylotypes*. Les pays sont regroupés par zone géographique et seuls les pays ayant au moins deux souches sont représentés. Seuls des groupements entre pays d'un même continent sont envisagés. Il n'y a pas de différence entre les méthodes DELTRAN, ACCTRAN et DOWNPASS. Voir l'Annexe A pour les graphiques des autres continents.



6.3.4.2 Analyse des chaînes de transmission du VIH-1C avec PhyloType

Les chaînes de transmission du VIH-1C sont étudiées et analysées avec l'outil PhyloType (Chevenet *et al*) qui met en exergue des *phylotypes*, reflets d'évènements fondateurs, ainsi que les liens

épidémiologiques qui les unissent. Trois analyses sont faites (*size* ≥ 20, 10 et 5) avec les parcimonies ACCTRAN et DELTRAN respectivement. La parcimonie DOWNPASS n'est pas disponible dans PhyloType car générant trop d'ambigüités sur les annotations ancestrales. Quelles que soient les analyses, les autres critères utilisés sont *persistence*, *size*/*different* et *support* respectivement supérieurs ou égaux à 1, 1 et 70%. Enfin, pour mesurer la significativité statistique des résultats, 1 000 *shufflings* sont calculés et les *phylotypes* avec une p-valeur strictement supérieure à 10/1 000 (= 1%) pour le critère *size* ne sont pas conservés. À nouveau, les résultats des analyses PhyloType sont d'abord présentés, puis nous les discuterons.

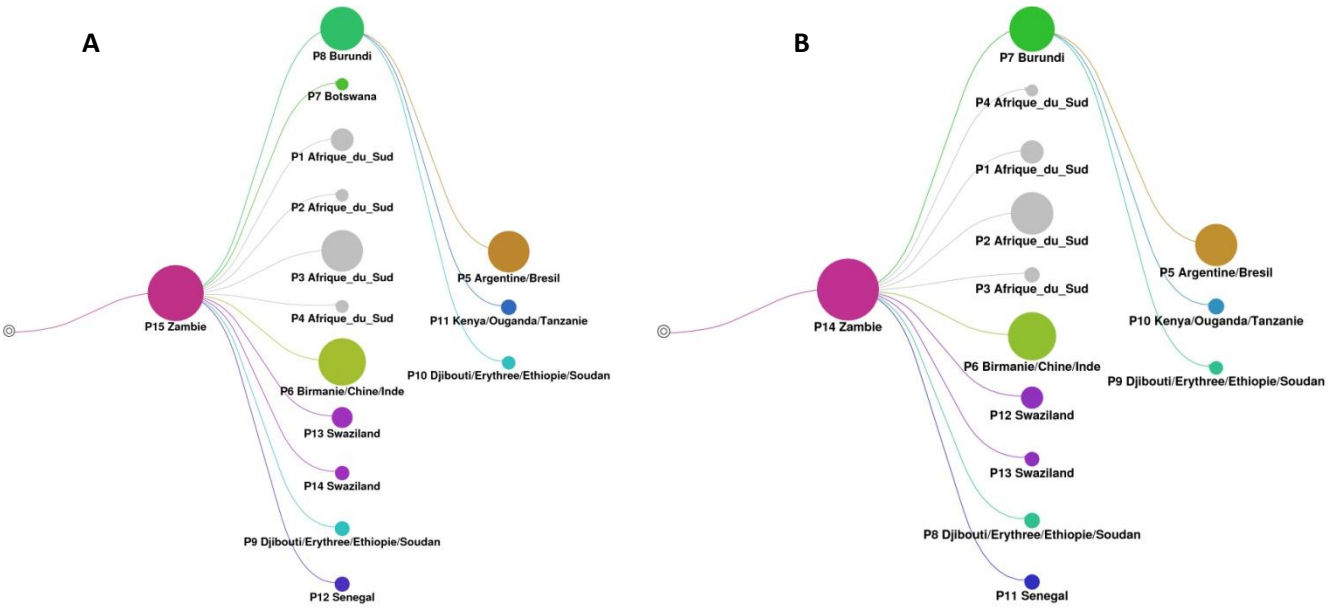
Tableau 7. Liste des associations suggérées par l'indice de régionalisation.
Ce tableau indique toutes les associations successives suggérées par l'indice de régionalisation. Les effectifs et l'indice de régionalisation correspondant à chaque annotation sont rappelés et l'estimation de leur union indiquée. Les lignes grisées montrent les associations retenues pour *size* ≥ 20, tandis que les lignes non grisées indiquent les associations temporaires. Les mesures sont identiques entre les parcimonies DELTRAN, ACCTRAN et DOWNPASS.

Association		**Effectifs**		**Régionalisation**		
$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{A} \cup \boldsymbol{B}$
Birmanie	Inde	1	355	-	0,090	0,087
Congo	RDC	2	86	1,000	0,741	0,736
Érythrée	Éthiopie	2	100	1,000	0,667	0,663
Kenya	Tanzanie	4	82	1,000	0,667	0,659
France	Espagne	7	26	0,833	0,800	0,781
Argentine	Brésil	8	253	0,714	0,044	0,023
Norvège	Suède	16	64	1,000	0,825	0,823
Chine	Birmanie/Inde	7	356	0,833	0,087	0,080
Soudan	Érythrée/Éthiopie	10	102	0,889	0,663	0,622
Ouganda	Kenya/Tanzanie	16	86	0,800	0,659	0,594
Djibouti	Érythrée/Éthiopie/Soudan	1	112	-	0,622	0,616

Les Figures 51 et 52 montrent un diagramme où les différents *phylotypes* significatifs observés, représentés par des cercles de surface proportionnelle à leur taille (*size*), sont disposés en fonction de leur apparition le long de la phylogénie. Les inclusions sont représentées par des arêtes reliant deux phylotypes. Ces figures correspondent aux résultats obtenus lorsque le critère *size* est respectivement supérieur ou égal à 20 et à 5. La Figure 52A montre les résultats de la parcimonie ACCTRAN, tandis que la Figure 52B et la Figure 53 donnent les résultats de la parcimonie DELTRAN. Les résultats de la parcimonie ACCTRAN, lorsque le critère *size* est supérieur ou égal à 10 et à 5, sont disponibles dans l'Annexe A et correspondent respectivement aux Figures 61 et 63. Les résultats de la parcimonie DELTRAN lorsque le critère *size* est supérieur ou égal à 10 sont aussi disponibles dans l'Annexe A (Figure 63). Tous les *phylotypes* représentés sont significatifs ($p \leq 1\%$ pour le critère *size*). Un numéro d'identification est attribué à chaque *phylotype*. Il est ainsi possible de connaître explicitement

les valeurs associées à chaque critère proposé par PhyloType (cf. section 6.2.5.1) en se reportant dans le tableau correspondant.

Figure 52. Cartes des liens entre les *phylotypes* des analyses avec *size* ≥ 20 pour ACCTRAN et DELTRAN.
Cartes des analyses PhyloType (ACCTRAN en figure A et DELTRAN en figure B) avec *size* ≥ 20, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70%. Tous les *phylotypes* sont statistiquement supportés (p-valeur inférieure ou égale à 1% pour le critère *size*). La taille des cercles est proportionnelle à la valeur du critère *size* pour le *phylotype* en question. Chaque couleur correspond à une annotation donnée, spécifiée au bas de chaque cercle. Un numéro d'identification est attribué à chaque *phylotype* et est indiqué avant l'annotation correspondante au *phylotype*.



Le Tableau 8 (resp. Tableau 9 de l'Annexe A) liste les *phylotypes* significatifs obtenus ($p \leq 1\%$ pour le critère *size*), les valeurs correspondantes à chaque critère et les p-valeurs (indiquées en rose) associées uniquement aux critères choisis (en gras) pour la parcimonie DELTRAN (resp. ACCTRAN) et lorsque le critère *size* est supérieur ou égal à 20. Les résultats de la parcimonie ACCTRAN (resp. DELTRAN) où le critère *size* est supérieur ou égal à 10 et à 5 correspondent respectivement aux Tableaux 10 et 12 (resp. Tableaux 11 et 13) de l'Annexe A. Le numéro d'identification de chaque phylotype est donné dans la colonne P. Lorsque le critère *different* (resp. *size/different*) est à 0 (resp. infini) alors le *phylotype* en question définit un clade. Par exemple, le *phylotype* n°11 du Tableau 8, représentant les souches des MSM de l'étude sur le Sénégal (cf. Chapitre 5), est un clade.

Origine de l'épidémie du sous-type C du VIH-1

Toutes les analyses PhyloType s'accordent à identifier un *phylotype* annoté Zambie (racine supportée à 88,8% en valeur aLRT (numéro 14 dans le Tableau 8 et numéro 15 dans le Tableau 5) à l'origine de l'épidémie du VIH-1C. Remarquons toutefois que le nombre total de souches incluses dans ce *phylotype* (critère *total*) n'englobe pas toutes les séquences de la phylogénie (3 605 souches sur 3 608 séquences étudiées), signifiant que la racine de ce *phylotype* ne correspond pas à la racine de la phylogénie comme déjà discuté plus haut. Les analyses PhyloType ne révèlent donc pas avec certitude l'origine de l'épidémie (indéterminée entre Zambie, RDC et Tanzanie, d'après les analyses

précédentes) mais plutôt l'épicentre de celle-ci, c'est-à-dire la région géographique à l'origine de la diffusion massive du VIH-1C. Ce résultat est en accord avec (et conforte) ceux observés précédemment.

Chaînes de transmission du sous-type C du VIH-1

Les chaînes de transmission majeures du VIH-1C sont facilement observables à l'aide des graphiques générés automatiquement par PhyloType qui synthétisent les liens entre les *phylotypes* significatifs observés. Les grands flux géographiques sont donnés par les analyses où le critère *size* est supérieur ou égal à 20. Ils sont aussi observés sur les autres analyses mais de nombreux *phylotypes* secondaires (de moindre importance en termes de nombre de membres) compliquent leur interprétation. Mais tous les *phylotypes* observés lorsque le critère *size* est supérieur ou égal à 20, le sont aussi lorsqu'il est supérieur ou égal à 10 et ceux observés lorsqu'il est supérieur ou égal à 10 sont, quant à eux, aussi observés lorsqu'il est supérieur ou égal à 5.

D'après l'analyse de la parcimonie DELTRAN présentant les grands flux migratoires (*size* ≥ 20), l'épidémie du VIH-1C se diffuse indépendamment de la Zambie vers l'Afrique australe (Swaziland [*phylotypes* n°13 et n°14] et Afrique du Sud [*phylotypes* n°1, n°2, n°3 et n°4]), vers l'Afrique de l'est (Burundi [*phylotype* n°7] et Djibouti/Érythrée/Éthiopie/Soudan [*phylotype* n°8]), vers le Sénégal (*phylotype* n°11 ; contenant les souches des MSM de l'étude sur le Sénégal) et vers le continent asiatique (Birmanie/Chine/Inde [*phylotype* n°6]) (Figure 52B). Du Burundi (*phylotype* contenant la presque totalité des souches collectées au Burundi, soit 300 sur 310, Tableau 8) l'épidémie se diffuse en Afrique de l'est (Kenya/Ouganda/Tanzanie [*phylotype* n°10]) et à nouveau vers les pays de la Corne de l'Afrique et le Soudan (Djibouti/Érythrée/Éthiopie/Soudan [*phylotype* n°9]). L'analyse avec ACCTRAN donne des résultats très similaires, mais ajoute un *phylotype* annoté Botswana (n°7), inclus dans le *phylotype* principal annoté Zambie (n°15) (Figure 52A). On retrouve notamment dans ces deux approches la double origine de l'épidémie en Éthiopie et dans les pays proches, issue directement de Zambie et du Burundi, et correspondant probablement aux variants C et C' référencés par Abebe *et al*. (2000).

Les analyses de la parcimonie DELTRAN où le critère *size* est moins restrictif (p. ex. Figure 53) montrent la formation de deux *phylotypes* annotés Roumanie (n°37 et n°36) (resp. Cuba [n°25 et n°26]) d'origine géographique différente (Burundi et Zambie). Hormis la Roumanie, les *phylotypes* représentant des pays européens (particulièrement la Belgique [*phylotype* n°14], le Portugal [*phylotype* n°35] et la Suède [*phylotype* n°39]) ont tous pour origine le *phylotype* principal Zambie. À nouveau les analyses avec la parcimonie ACCTRAN confirment ces observations, sauf en ce qui concerne un *phylotype* annoté Roumanie (n°40) qui n'a plus pour origine le *phylotype* principal Zambie (n°53)

mais un *phylotype* annoté Botswana (n°21) ayant pour origine le *phylotype* principal Zambie (Figure 64 de l'Annexe A).

Figure 53. Carte des liens entre *phylotypes* lorsque *size* ≥ 5 avec DELTRAN.
Carte de l'analyse PhyloType lorsque *size* ≥ 5, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie DELTRAN. Tous les *phylotypes* présentés sont statistiquement supportés (p-valeur inférieure ou égale à 1% pour le critère *size*). La taille des cercles est proportionnelle à la valeur du critère *size* pour le *phylotype* en question. Chaque couleur correspond à une annotation donnée, spécifiée au bas de chaque cercle. Un numéro d'identification est attribué à chaque *phylotype* et est indiqué avant l'annotation correspondante au *phylotype*.

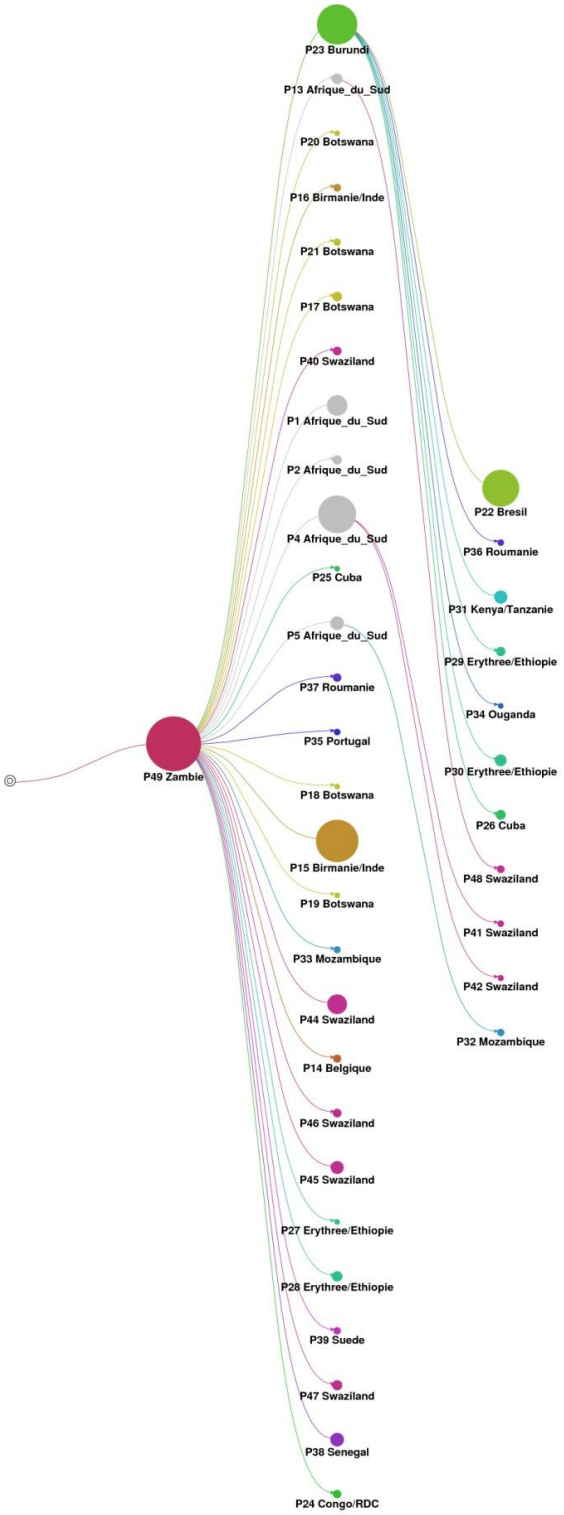


Tableau 8. Valeurs associées à chaque critère pour tous les *phylotypes* significatifs observés lorsque *size* ≥ 20 avec DELTRAN.

Ce tableau contient la liste de tous les *phylotypes* obtenus lorsque *size* ≥ 20, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie DELTRAN, ainsi que toutes les valeurs associées à chacun des critères que l'on peut choisir. Les valeurs en rose représentent les p-valeur obtenues par le *shuffling*. Seuls les *phylotypes* dont la p-valeur est strictement inférieure à 11/1 000 pour le critère *size* sont représentés. Les abréviations des pays sont les suivantes : AR, Argentine ; BI, Burundi ; BR, Brésil ; CN, Chine ; DJ, Djibouti ; ER, Érythrée ; ET, Éthiopie ; IN, Inde ; KE, Kenya ; MM, Birmanie ; SD, Soudan ; SN, Sénégal ; SZ, Swaziland ; TZ, Tanzanie ; UG, Ouganda ; ZA, Afrique du Sud ; ZM, Zambie. Les abréviations des titres sont les suivantes : P, identifiant du *phylotype* ; A, *annotation* ; Tt, *total* ; Sz, *size* ; Ps, *persistence* ; Df, *different* ; Sl, *local separation* ; Sg, *global separation* ; Dv, *diversity* ; Sp, *support* ; Spg, *global support*.

P	A	Tt	Sz	Ps	Df	Sz/Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
1	ZA	86	**76**	**2**	9	**8,444**	0,005	0,007	0,083	0,064	0,079	**0,855**	0,855
			0/1000	0/1000		0/1000						0/1000	
2	ZA	311	**265**	**3**	33	**8,030**	0,002	0,014	0,089	0,027	0,154	**0,831**	0,843
			0/1000	0/1000		0/1000						0/1000	
3	ZA	80	**33**	**2**	29	**1,138**	0,007	0,010	0,075	0,092	0,137	**0,876**	0,876
			0/1000	0/1000		0/1000						0/1000	
4	ZA	51	**20**	**2**	16	**1,250**	0,005	0,008	0,071	0,069	0,108	**0,874**	0,874
			0/1000	0/1000		0/1000						0/1000	
5	AR/BR	269	**260**	**2**	5	**52,000**	0,018	0,029	0,106	0,172	0,269	**0,992**	0,992
			0/1000	0/1000		0/1000						0/1000	
6	MM/CN/IN	356	**339**	**2**	14	**24,214**	0,004	0,022	0,081	0,044	0,265	**0,740**	0,882
			0/1000	0/1000		0/1000						0/1000	
7	BI	829	**300**	**3**	75	**4,000**	0,006	0,010	0,093	0,065	0,106	**0,861**	0,861
			0/1000	0/1000		0/1000						0/1000	
8	DJ/ER/ET/SD	71	**34**	**3**	34	**1,000**	0,005	0,014	0,051	0,095	0,264	**0,906**	0,906
			0/1000	0/1000		0/1000						0/1000	
9	DJ/ER/ET/SD	47	**26**	**2**	17	**1,529**	0,003	0,023	0,059	0,042	0,394	**0,773**	0,773
			0/1000	0/1000		0/1000						0/1000	
10	KE/UG/TZ	43	**36**	**3**	7	**5,143**	0,009	0,014	0,083	0,104	0,171	**0,926**	0,926
			0/1000	0/1000		0/1000						0/1000	
11	SN	33	**33**	**1**	0	∞	0,018	0,033	0,075	0,240	0,438	**0,980**	0,980
			0/1000	0/1000		0/1000						0/1000	
12	SZ	87	**70**	**3**	13	**5,385**	0,003	0,045	0,077	0,035	0,594	**0,749**	0,766
			0/1000	0/1000		0/1000						0/1000	
13	SZ	33	**30**	**2**	3	**10,000**	0,002	0,021	0,059	0,041	0,352	**0,781**	0,781
			0/1000	0/1000		0/1000						0/1000	
14	ZM	3605	**564**	**4**	490	**1,151**	0,014	0	0,114	0,120	0,000	**0,880**	0,880
			0/1000	0/1000		0/1000						0/1000	

6.4 Conclusion

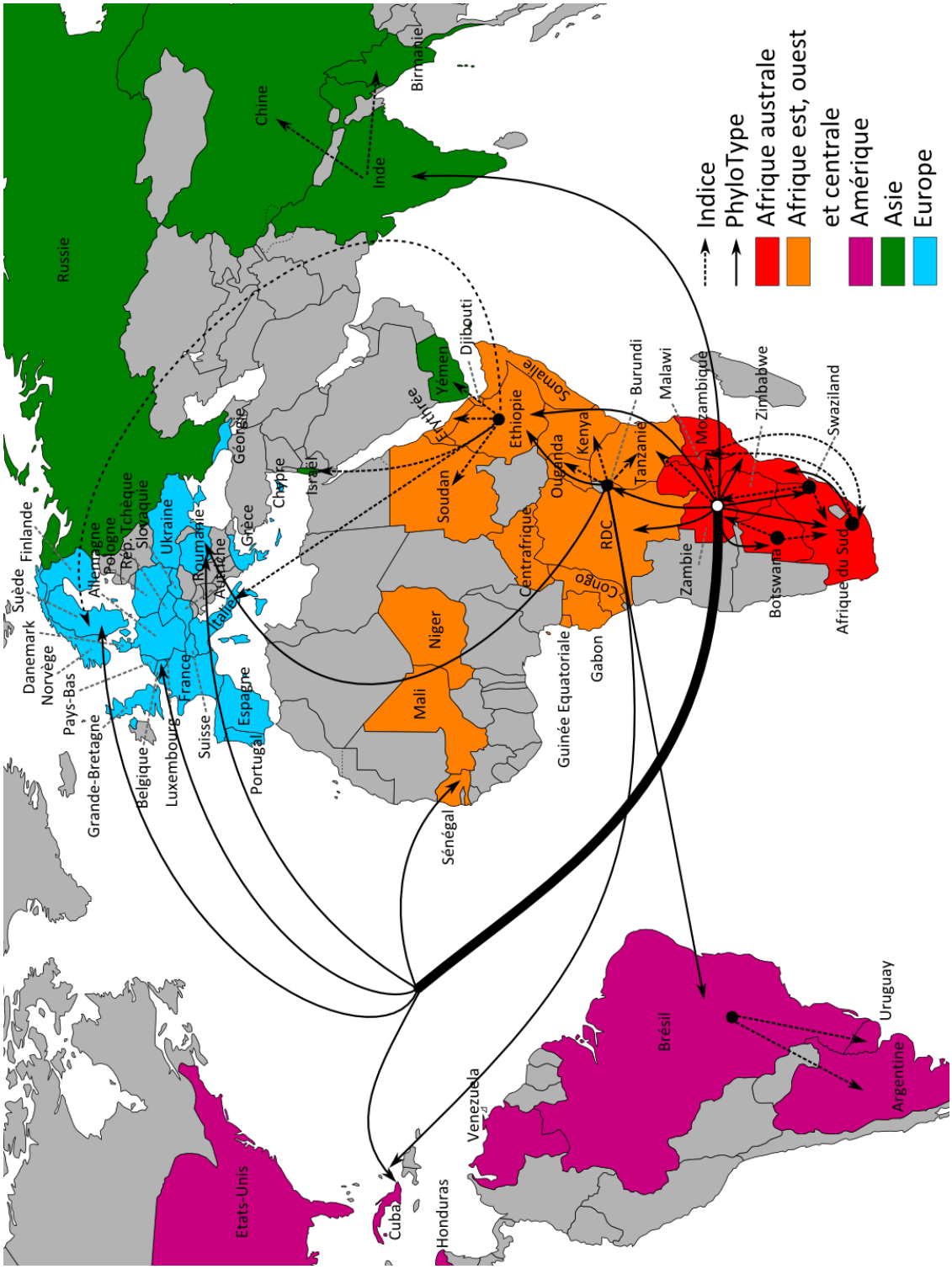
Nous présentons la première étude moléculaire visant à retracer l'histoire épidémiologique du sous-type C du VIH-1 à l'échelle mondiale en s'aidant du maximum de souches disponibles. Pour cela un arbre de maximum de vraisemblance est calculé et comprend 3 609 souches *pol* (dont 528 sont nouvelles) collectées à travers le monde (63 pays différents) et à différentes époques (entre 1986 et 2010). Cette phylogénie est exploitée de deux manières différentes mais complémentaire, toutes deux basées sur le principe de parcimonie. La première utilise des indices basés sur les transitions entre pays (reconstruites par parcimonie) pour synthétiser le mouvement de l'épidémie décrit par la phylogénie. La deuxième utilise le logiciel PhyloType qui met en exergue des *phylotypes*, groupes de séquences reflétant des évènements fondateurs probables, afin d'observer les chaînes de transmission majeures de l'épidémie du sous-type C du VIH-1.

En supposant que les séquences utilisées dans cette étude fournissent une représentation adéquate de la diversité globale du sous-type C du VIH-1 et que la phylogénie obtenue soit la plus juste possible, notre étude suggère que l'épicentre de l'épidémie du sous-type C du VIH-1 se situe en Zambie (Figure 54). Cette épidémie s'est ensuite diffusée indépendamment dans les pays de l'Afrique australe, au Burundi, en Éthiopie, au Sénégal et en Inde. Du Burundi, l'épidémie du sous-type C s'est diffusée dans les pays à l'est (Ouganda, Kenya et Tanzanie), à nouveau en Éthiopie et au Brésil (Véras *et al*, 2011a; de Oliveira *et al*, 2010; Jones *et al*, 2009; Bello *et al*, 2008; Fontella *et al*, 2008). De l'Éthiopie l'épidémie s'est répandue dans les autres pays de la Corne de l'Afrique (Érythrée, Djibouti), au Soudan et dans les pays du Moyen-Orient (Israël, Yémen) (Gehring *et al*, 1997). Du Brésil, l'épidémie s'est propagée dans le sud de l'Amérique Latine (Argentine et Uruguay) (Carrion *et al*, 2004) et de l'Inde dans les pays de l'Asie de l'est (Birmanie, Chine, Corée du Sud et Taiwan) (Lau *et al*, 2007). Enfin, notons les nombreuses introductions de ce variant en Europe provenant de pays africains et qui évoquent les liens sociologiques entre les pays européens et africains créés durant l'époque coloniale (Perrin *et al*, 2003).

Notre étude n'a pu déterminer avec exactitude l'origine géographique du sous-type C du VIH-1. Toutefois, les différentes méthodes de parcimonie utilisées (DELTRAN, ACCTRAN et DOWNPASS) s'accordent sur l'incertitude à donner au nœud racine (RDC, Zambie ou Tanzanie) (Figure 47). Ces trois pays se situent sur le continent africain et sont frontaliers. Au vu de l'étonnante diversité génétique présente en RDC (Vidal *et al*, 2000) et sachant que deux souches isolées à partir de matériels anciens (1958 et 1960) en RDC, présentaient déjà une diversité génétique étonnante (Worobey *et al*, 2008; Zhu *et al*, 1998), la RDC est généralement vue comme l'épicentre de l'épidémie du VIH-1 et il serait donc probable, dans cette hypothèse, que l'origine géographique du sous-type C se situe également en RDC. Ceci d'autant plus qu'un nombre important de séquences (21 sur 45, soit 47%) collectées en RDC se trouvent à proximité de la racine (Figure 47). Malgré cela, l'hypothèse d'une origine zambienne est aussi très probable, puisque, premièrement, ce pays est identifié comme l'épicentre de l'épidémie du sous-type C et, deuxièmement, la prévlence du sous-type C en RDC est surtout observée au sud du pays (Vidal *et al*, 2005), à proximité de la frontière avec la Zambie. D'ailleurs, sur les 21 souches de RDC situées à proximité de la racine, 20 sont collectées à Mbuji-Mayi ou Lumbumbashi, deux villes au sud de la RDC. Au vu de la situation géographique particulière entre la RDC et la Zambie (l'appendice au sud-est de la RDC traverse pratiquement la Zambie), facilitant certainement les migrations de populations entre ces deux pays (par exemple, pour traverser la Zambie d'est en ouest), l'argumentation en faveur de l'un ou l'autre pays devient difficile et il serait plus vraisemblable de supposer que l'origine de l'épidémie du sous-type C se situe au niveau de la région frontalière entre ces deux pays ; zone riche en industrie minière (p. ex. cuivre ou cobalt) et où

les mouvements de populations sont donc nombreuses. En revanche, l'hypothèse d'une origine tanzanienne reste assez peu probable, mais elle ne peut être complètement rejetée.

Figure 54. Planisphère résumant la diffusion de l'épidémie du sous-type C du VIH-1.
Les flèches représentent les mouvements de l'épidémie du sous-type C du VIH-1 dans le monde entier. Les flèches en pointillé indiquent les flux identifiés avec les indices, tandis que les flèches en trait continu indiquent ceux identifiés avec PhyloType et avec ou sous les indices. Les cercles indiquent que l'épidémie se diffuse de manière indépendante dans plusieurs pays différents. Le cercle blanc indique l'épicentre de l'épidémie. Seuls les pays représentés dans cette étude sont mentionnés et coloriés. Les pays de l'Afrique australe sont en rouge, ceux de l'Afrique de l'est, ouest et centrale en orange, ceux du continent asiatique en vert, ceux du continent américain en mauve et ceux du continent européen en bleu.



À la section 4.3.2, page 120, nous présentons une étude, basée sur cette phylogénie, visant à déterminer l'origine temporelle de l'épidémie du sous-type C du VIH-1. Deux méthodes de distances sont utilisées : la méthode ULS, présentée au Chapitre 4, et la régression linéaire *Root-to-tip* (cf. Chapitre 2). Pour mémoire, la méthode ULS date l'ancêtre commun aux souches du sous-type C à 1964, estimation qui semble cohérente avec celles retrouvées dans la littérature (Hemelaar, 2012), et *Root-to-tip* à 1782, estimation complètement différente de celles communément admises. Ici, et pour des raisons de temps de calculs, les intervalles de confiance ne sont pas calculés, mais rappelons que ceux associés aux estimations publiées dans la littérature sont larges (allant de 1933 à 1973), indiquant une grande incertitude, vraisemblablement liée à un faible signal global (cf. discussion à la section 4.3.2).

Nos analyses ont révélé deux origines géographiques différentes aux souches collectées en Éthiopie, probablement l'explication de l'observation de deux variants C et C' décrit précédemment (Abebe *et al*, 2000). La souche 86ET-ETH2220, présente dans cette étude, se situe dans le groupe qui a pour origine géographique le Burundi. Dans d'autres études, elle se place en-dehors du sous-cluster C' (Kassu *et al*, 2007; Abebe *et al*, 2000), ce qui suggère que les souches appartenant au sous-cluster C' ont pour origine épidémique la Zambie, tandis que le C (pour ce qui concerne l'Éthiopie) viendrait du Burundi. Toutefois, nous ne pouvons pas confirmer ces inférences sur la base d'informations publiées. Il faudrait aller plus loin dans les recoupements et disposer de plus de données. Notons toutefois que Kassu *et al*. (2007) n'observent pas la formation du sous-cluster C' sur la protéase et sur la transcriptase inverse, mais uniquement sur les gènes *gag* et *env*.

Plusieurs études indépendantes ont suggéré un lien épidémiologique entre l'Inde et l'Afrique du Sud (Shen *et al*, 2011; Dietrich *et al*, 1995, 1993). Toutefois, Dietrich *et al*. (1993) utilisent très peu de souches provenant de l'Afrique (peu disponibles à l'époque), tandis que le jeu de séquences de Shen *et al*. (2011) comprend en totalité 312 séquences réparties sur 27 pays différents (en-dehors de celles collectées en Inde). Ce chiffre représente moins de la moitié des souches collectées en Zambie ou en Afrique du Sud utilisées dans cette étude. Nos analyses qui intègrent une plus grande quantité de séquences collectées en Zambie et en Afrique du Sud (respectivement 633 et 689 souches), suggèrent un lien direct entre la Zambie et l'Inde et non avec l'Afrique du Sud. Notons que ces études utilisent le gène *env* et qu'une seule de leurs souches pertinentes (03ZA-PS057MB2) est dans notre analyse. Elle se place à l'intérieur d'un groupe contenant d'autres souches d'Afrique du Sud.

L'absence significative de séquences collectées au Royaume-Uni (due à la faible quantité de séquences disponibles publiquement) sur la région génomique considérée, ne permet pas de confirmer les liens épidémiologiques établies par de Oliveira *et al*. (2010), entre le Brésil, l'Afrique de l'est et le

Royaume-Uni. En effet, cette dernière étude suggère que l'épidémie du sous-type C s'est d'abord diffusée au Royaume-Uni avant d'être introduite au Brésil, suite à un évènement fondateur. Théorie en contradiction avec une étude récente (Véras *et al*, 2011a) et d'autres études anciennes (Bello *et al*, 2008; Fontella *et al*, 2008) qui suggèrent une introduction du sous-type C au Brésil directement par le Burundi ou un pays à proximité. Cette dernière version est corroborée par nos analyses. Notons que Bello *et al*. (2008) et Fontella *et al*. (2008) n'utilisent pas de souches d'Angleterre dans leurs études. Les quelques souches d'Angleterre considérées dans notre étude permettent uniquement de montrer le lien épidémiologique avec l'Afrique australe, laissant sous-entendre qu'elles proviennent probablement d'individus originaires ou ayant voyagé en Afrique (Dougan *et al*, 2005; Hughes *et al*, 2009). Rendre public toutes les séquences disponibles permettrait de considérer systématiquement un ensemble de séquences plus vaste et ainsi de mettre en évidence des liens épidémiologiques plus complexes, précis et exhaustifs.

Cette dernière remarque rappelle le problème du temps de calcul nécessaire pour inférer des phylogénies par des méthodes probabilistes (considérées comme les plus précises) et pose un double challenge aux chercheurs qui doivent développer des outils permettant l'inférence de phylogénies de plus en plus grandes, en des temps raisonnables, mais aussi de les visualiser et les interpréter simplement et rapidement.