Habilitation à Diriger des Recherches

Université Pierre et Marie Curie, Paris 6
Systématique, Adaptation, Evolution (UMR 7138)
Atelier de Bioinformatique

Guillaume Achaz
Maître de conférences

Evolution moléculaire,

des données aux modèles et *vice-versa*

Soutenue le 3 septembre 2009

Jury

Pr Martine Boccara	présidente
Pr John Brookfield	rapporteur
Dr Hugues Roest Crollius	rapporteur
Pr Xavier Vekemans	rapporteur
Pr Dominique Higuet	examinateur

Remerciements

Il arriva devant la lourde porte de l'Université. Fermée. Il fit résonner le titanesque heurtoir au prix d'un effort démesuré. On vint. Un judas dérobé s'ouvrit et il vit le visage patibulaire d'un homme qui lui décrocha un sourire forcé. "Qu'est-ce que vous voulez ?" demanda l'huissier d'une voie nasillarde. "Je viens pour la soutenance. J'ai... euh... J'ai une lettre de recommandation de ma grand-mère qui est malade et m'envoie lui chercher une galette et un pot de beurre". "Hum.... je vois. C'est bon ; vous pouvez entrer.", maugréa l'huissier à contre-coeur.

Il passa la porte et s'engouffra sous le sombre porche. Derrière, il découvrit émerveillé un superbe parvis, pavé de dalles en marbre et ombragé par quelques arbres exotiques d'où s'échappait le chant mélodieux de quelques oiseaux rares. Il croisa des étudiants en toge, qui déambulaient en ricanant. Il remarqua qu'ils étaient coiffés d'étranges petits chapeaux. Son regard se porta ensuite vers le centre du parvis, vers la massive tour sombre. Il s'en échappait des grognements sinistres. Il resta comme pétrifié d'effroi. "Le bureau du recteur !", lui annonça un étudiant, goguenard devant à sa mine déconfite. "N'y rentre pas. Ou bien, fais en sorte de ne jamais en ressortir... Tu as l'air perdu, que cherches-tu ici ?" Il lui demanda le chemin des laboratoires de ***. L'étudiant leva mollement la main pour lui indiquer un large bâtiment qui se dressait un peu à l'écart.

Il pénétra dans le bâtiment. Une forte odeur de formol lui chatouilla les narines. Dans ce grand hall froid, il n'y avait aucun bruit. Soudain, un bruit de pas vit vibrer la pierre. Le tac-tac cadencé des pas résonnait dans les os de son crâne. Il se retourna brusquement. Personne. Puis, d'un coup, Dominique H était là ; souriant. Il se détendit. "C'est par là. Cheminons ensemble", lui dit Dominique H. Il était content. Il finirent par atteindre une salle nichée au fond de ce dédale de couloirs, si déroutant pour les nouveaux arrivants.

A peine entré dans la salle, il perçut un bourdonnement presque imperceptible. Peut-être était-ce dans son crâne ? Il n'en su rien. Dominique H pris sa place auprès des quatre autres sages du département de ***. Le clan des sages était orchestré par Martine B, qui

lui adressa un sourire engageant. Les trois autres hommes, Hughes R, John B, et Xavier V attendaient, l'air quiet, absorbés dans leurs pensées. Au moment où il monta sur l'estrade, il fut saisi par une foule de souvenirs. Il sentit dans sa bouche le goût amer du café. Il sourit. Il repensa à ceux avec qui il partageait son bureau ; ceux avec qui les discussions sur tout ou rien pouvaient surgir à chaque instant, Sophie B, Joël P, Emmanuelle O et Eric D. Puis il repensa à tous les membres de l'Atelier. L'Atelier était un lieu collégial et mythique, îlot de liberté et refuge à l'abri des fauves de l'Université. Il eut une pensée particulière pour ceux avec qui il avait travaillé, Todd T et Eduardo R. Il eut une pensée particulièrement émue pour Etienne L, avec qui il avait partagé de nombreuses discussions scientifiques, mais aussi d'autres aventures dans le monde de l'imaginaire. Enfin, il se remémora le groupe de lecture qui s'efforçait d'extraire les connaissances, enfouies dans les arcanes du livre de Rice : Etienne L, encore, Mathilde C, Pierre N et Sarah S.

Le son marqué et incisif des coups de bâtons le sortit de sa rêverie. Il se retrouva bouche bée, face à ces gens qui le regardaient en silence. C'était à lui. Après un instant d'hésitation, il entama son exposé sur les conséquences à long terme de la duplication des êtres vivants. Il crut entrevoir, parmi les curieux qui étaient venu écouter ses élucubrations scientifiques, des collègues qui lui étaient chers, Pierre N, un autre, John W et tant d'autres encore. Il vit avec beaucoup de bonheur le regard complice de Stéphanie B. Il pensa aux enfants. Il se concentra sur son discours et fit place à la rhétorique scientifique. Lorsqu'il eut fini, il y eut un moment de pause. Puis, les cinq sages engagèrent une discussion à l'aide d'une série de questions. A la fin, il sourit. Il était content. Il allait enfin pouvoir célébrer l'événement. Non seulement avec ceux qui partageaient sa vie ou son travail, mais aussi avec ses amis, ceux qui ajoutent le sel de la vie.

Contents

Préambule

A propos de science

Démarche scientifique

La démarche scientifique à laquelle j'aspire adresse des questions en biologie évolutive par l'intégration de plusieurs niveaux d'organisation du vivant. Elle se fonde sur l'usage d'outils informatiques et mathématiques. Je m'attache plus particulièrement à étudier l'évolution moléculaire, champs disciplinaire visant à comprendre les bases moléculaires de la théorie de l'évolution. Je cherche à comprendre, modéliser et prédire les *pattern* évolutifs que l'on peut trouver au niveau des macromolécules du vivant et à proposer des *processus* qui les explique.

Dès lors que l'on s'intéresse à l'histoire de la biodiversité, on est assailli par une myriade d'observations intéressantes. Chacune d'entre elles soulève des problèmes et des questions spécifiques. Une grande partie de ces questions a cependant trait aux thèmes dérivant d'un but unique : celui de proposer une explication à une observation donnée. Par exemple, on cherche à comprendre pourquoi tel gène est très conservé au cours de l'évolution, ou encore pourquoi telle séquence est présente en milliers d'exemplaires dans un génome d'intérêt. On cherche une théorie explicative permettant de rendre compte du pattern mis en exergue. Dans cet exercice, plusieurs types d'outils peuvent être utilisés. On cherche, cependant, *in fine*, quel modèle permet d'expliquer au mieux notre observation.

Une autre classe de questions est celle regroupant les variations sur le thème de la prédictions des différents patterns possibles pour un processus donné. Par exemple, quelle est la nature des polymorphismes que l'on attend dans un modèle d'équilibre dérive-mutation. Ou bien, quels sont les types d'arbre phylogénétiques que l'on attend si la spéciation suit un modèle de Yule. On ne cherche plus à proposer une explication plausible permettant de rendre compte d'une observation, mais on cherche à comprendre les

conséquences d'un processus donné. En somme, on veut énumérer et caractériser les différentes histoires possibles qui découlent de certaines contraintes du modèle étudié. Si l'on impose peu de contraintes, de nombreuses histoires (et donc de nombreux *patterns*) sont possibles et, au contraire, si on impose toutes les contraintes, seules l'histoire "vraie" reste possible. Le nombre de contraintes représente le nombre de paramètre du modèle, qu'ils soient explicites ou non.

Les deux types de questions sont complémentaires. L'analyse de données permet de suggérer de nouveaux modèles et la modélisation permet de caractériser les modèles théoriques. En utilisant les prédiction du modèle, on pourra tester sa vraisemblance sur des données. Je suis séduit par l'idée de mener de front ces deux d'approches. J'aspire à m'embarquer sur le voyage fascinant qui relie données aux modèles, et *vice-versa*. Il m'est donc indispensable d'acquérir des compétences aussi bien en analyse de données qu'en modélisation, puis il me faut mettre en place des ponts nécessaires aux aller-retour entre l'un et l'autre.

L'heuristique génétique

Gageons que la vie soit une auto-catalyse. Dans cette idée, la vie serait un équilibre bio-chimico-physique capable de s'auto-entretenir. Une stratégie permettant de maintenir cet équilibre est sans doute la faculté de se dupliquer ; celle de générer de nouvelles copies de soi-même. Quiconque ne se duplique pas a potentiellement moins de chance de dresser des remparts solides contre l'érosion du temps et de l'environnement. On admettra bien volontier qu'une structure se dupliquant perdure plus longtemps qu'une structure ne se dupliquant pas. L'apparition de la vie est peut-être la première duplication de cette structure d'équilibre d'auto-catalytique. Dès lors que la structure se duplique, on comprendra que des variations peuvent être introduites au cours de cette duplication. Générations et variations sont suffisantes pour générer une évolution par sélection naturelle.

L'objet de la sélection naturelle est, dans cette idée, la structure auto-catalytique, c'est à dire, dans sa version moderne, l'organisme[1]. L'unité sélectionnée est celle qui se duplique. Si une structure présente un taux de duplication relatif plus important, elle est sélectionnée positivement. Cette vision n'exclue néanmoins pas que d'autres unités peuvent être sélectionnées. Dès lors qu'une structure se duplique, elle peut générer des variations et évoluer par sélection naturelle. La pertinence biologique d'unités non-organismiques soumises à la sélection est ici laissée à l'appréciation de chacun.

[1] Cette position n'est pas très novatrice et se conforme totalement à celle qui est communément admise.

Si l'unité est l'organisme, on est en droit de se demander quelle est la pertinence de l'approche réductionniste génétique ? Attention, je ne cherche pas ici à alimenter le débat sur la place du matériel génétique dans le vivant (Dawkins, 1976; Morange, 1998; Moss, 2002). Quelle que soit sa place, nous pouvons considérer le matériel génétique comme un marqueur de l'évolution des organismes qui le porte. Que ce matériel génétique soit cause ou conséquence de l'évolution moléculaire est hors de propos. Le matériel génétique est un marqueur. Il permet de suivre l'évolution des organismes. Bien évidement, dans l'approche génétique, on ignore superbement toute l'hérédité épigénétique (*i.e.* définie comme "non génétique"). Le matériel génétique (et les mutations qu'il porte) est transmis sur de très nombreuses générations. Par ailleurs, il est aujourd'hui relativement bien caractérisé dans de nombreux organismes ; il est, en tous cas, mieux caractérisée que les supports de l'hérédité non génétique.

Je perçois l'approche génétique comme une heuristique. Elle ne garantie pas d'être la manière optimale de suivre l'évolution, mais elle semble être un bon compromis entre fiabilité et utilité. La valeur exacte de cette heuristique est difficile à apprécier pleinement. Néanmoins, comme l'héritabilité du matériel génétique est très grande, je soupçonne que c'est une bonne heuristique.

J'ai découvert récemment, à travers la lecture du livre de Rice (2004), le théorème de Price (1970). Celui-ci se propose de modéliser l'évolution d'un ou plusieurs traits sans aucune référence à la génétique. C'est une généralisation des modèles d'évolution. Il considère l'organisme comme une agrégation de traits et cherche à prédire les variations de ces traits. Si ces traits sont des caractères discrets à héritabilité mendélienne, alors le modèle est génétique. Il ne l'est pas dans sa forme générale. Cette généralisation de la théorie de l'évolution m'apparaît comme très séduisante car elle permet de s'affranchir de l'heuristique génétique. Je n'ai malheureusement pas assez de recul pour juger de sa réelle utilité dans notre compréhension de la théorie de l'évolution, mais projette de m'y pencher plus sérieusement dans ma recherche future.

Science et société

Nous sommes probablement en train de vivre une époque charnière marquée par un changement de politique scientifique (attribution des crédits sur projets, renforcement des évaluations, etc.). C'est certainement vrai au niveau national et probablement extensible au niveau international. Ce changement de politique scientifique est sans doute

notre lorgnette par laquelle nous sommes témoins d'un changement général de la mentalité de notre société. Je me garderais bien de porter un jugement personnel, subjectif, à la qualité de cette évolution. Que nul n'interprète la suite de cet argumentaire comme un jugement de valeur. Néanmoins, la perspective d'un tel changement soulève la question de la pertinence et de l'utilité de notre métier dans une société différente. Après une expérience de projection de notre société dans un futur proche, je discuterais de la place d'un chercheur en science "fondamentale", d'un enseignement universitaire publique et enfin d'un enseignant-chercheur dans cette hypothétique future société.

Voyage dans le temps

Bien qu'il soit impossible de prédire ce que sera le monde de demain, jouons à en deviner les tendances. S'il me faut donner un trait qui marque le changement de mentalité auquel nous assistons, je me range à l'avis général. Individualisme. L'augmentation de l'envie d'un bonheur individuel est probablement l'un des traits les plus marquant. Ce trait n'est, bien entendu, pas unique, mais il est fort. Il serait malencontreux d'en tirer la conclusion hâtive que l'individualisme n'ait jamais existé par le passé. Bien au contraire.

L'individualisme précède certainement l'avènement des sociétés élaborées que nous connaissons. La sélection naturelle n'est-elle pas un exemple biologique d'un individualisme matériel physico-chimique poussé à son extrême ? Si oui, l'émergence de la pensée n'a pu se faire que dans un contexte où la maximisation de sa réalisation propre était de rigueur. Dans le cas d'organismes non-pensant, seule la dimension biologique peut être prise en compte ; la dimension psychologique est absente. Lorsque l'on parle d'organisme non-pensant, la réalisation d'un individu se mesure par sa *fitness* ; l'individu présentant un trait héritable qui augmente sa *fitness*, devient petit à petit majoritaire au fil des générations. Aujourd'hui, seuls les organismes ayant su maintenir une forte *fitness* ont laissé des descendants. Les autres lignées ne sont plus. De fait, la notion d'individualisme est inscrite dans notre être biologique.

C'est d'ailleurs l'un des arguments repris par les créationnistes qui militent contre la théorie de l'évolution (Yahya, 2002). L'avènement de la théorie de l'évolution a conduit certaines personnes à légitimer des actes amoraux au nom d'une science sociale pervertie. Par exemple, les partisans des théories racistes se sont trouvés une certaine accointance avec la théorie de l'évolution, via le darwinisme social. Ces partisans ont vu dans la théorie de l'évolution une explication "scientifique" à leur croyances. Il n'y a dans ces théories, bien entendu, aucune logique scientifique solide. En invoquant les dérives syllogiques historiques

de la théorie de l'évolution, les créationnistes proposent, sur des principes moraux, de rejeter la valeur scientifique de cette théorie. Néanmoins, construite sur une solide logique scientifique, la théorie de l'évolution est heureusement adoptée par l'ensemble de la communauté scientifique, exception faite de quelques excentriques obtus.

Revenons-en à l'individualisme. Bien que notre être biologique soit individualiste, notre être psychologique pourrait en être autrement. Si l'on se laisse tenter par l'idée que notre mental a émergé sous l'action de la sélection naturelle, il faut envisager que notre mental premier était purement individualiste. D'où le point avancé ci-dessus, l'individualisme n'est pas un nouveau trait de caractère, c'est notre état ancestral.

Il est, par ailleurs, peu réaliste de penser que l'individualisme psychologique va devenir, dans la société de demain, total, totalitaire ou omniprésent. On peut plutôt proposer que le nouvel équilibre sociétal vers lequel nous nous dirigeons aura une composante individualiste plus marquée.

"On en veut pour notre argent !". " Il faut plus de rentabilité". Voilà un discours récurrent que l'on peut entendre ci et là. Toute action entreprise doit être rentable. Ne rien entreprendre qui ne puisse potentiellement présenter un "retour sur investissement". Chaque individu veut maximiser son propre bonheur et, par conséquent, ne pas perdre de temps, d'argent ou d'énergie. Obsédé par l'efficacité et la rentabilité, il ne peut plus faire confiance à autrui. Il lui faut donc le contrôler de près et s'assurer du bon déroulement des opérations. Brandi au nom du bonheur de tous, ce contrôle extrême n'est peut-être que la simple conséquence d'un individualisme prononcé.

Science "fondamentale"

Il me semble difficile de définir clairement la science dite "fondamentale". En effet, dans l'usage courant, on définit une science fondamentale en la comparant à une autre plus appliquée. La biologie humaine est plus fondamentale que la médecine, mais plus appliquée que la biologie évolutive. On est, comme souvent, tenté d'emboîter linéairement les sciences les unes dans les autres sur un axe unique. Je crois cependant que notre vision unidimensionnelle de l'organisation du monde est une facilité qui tend à pervertir la réalité, plus complexe. Je défendrais plutôt une organisation des sciences en réseau, où chaque discipline, voire sous-discipline, est en interaction avec une autre.

Quelle serait, dans cette vision réticulée des sciences et du savoir, la place de sciences plus fondamentales ? Elles formeraient sans doute une partie de ce réseau de savoir. Une partie où les sciences plus fondamentales seraient d'avantage en interaction entre elles

qu'elles ne le sont avec des sciences très appliquées. L'absence de sciences étudiant le savoir fondamental laisserait en l'état cette partie du savoir. La non-remise en question d'une partie de ce réseau de savoir entraîne irrémédiablement une perte de la dynamique d'interaction entre les parties gelées et les autres parties du réseau. L'absence de changement de paradigme dans la partie fondamentale du savoir est sans doute la concrétisation d'une arrogance intellectuelle qui se targue d'en connaître suffisamment. Mais quiconque se penche sur les paradigmes fondamentaux est pris d'un vertige dû à notre étonnante absence de connaissance dans les domaines les plus enfouis. Nos savoirs empiriques masquent notre ignorance des mécaniques profondes qui régissent le monde et notamment le monde vivant.

Néanmoins, dans un contexte de maximisation de bonheur personnel, est-ce que cette perte de savoir et de dynamique est importante ? Est-ce qu'assumer notre ignorance n'est pas une solution pour palier à ces vertiges ? Je crois que, pour quelqu'un dont le bonheur personnel est maximal, ces questions sont sans importance. Cependant si ce bonheur individuel n'est pas atteint, où s'il est instable, alors la perte de compréhension du monde marque le début d'un déclin inéluctable. Dans notre hypothétique future société individualiste, rien ne sera acquis définitivement. Le bonheur personnel va et vient. Dans ces conditions, je pense que l'abandon complet des sciences fondamentales ne peut mener que, par effet de cascade, à la perte complète du bonheur individuel de chacun.

Enseignement universitaire publique

Aujourd'hui, l'enseignement supérieur est en partie assuré par l'état. Certaines disciplines, néanmoins, sont majoritairement assurées par des structures non universitaires. On peut, par exemple, prendre le cas des écoles de commerce ou d'ingénierie. Ma connaissance de l'éducation non-scientifique (privée ou publique) est si pauvre que je ne peux en discuter avec contenance. Cependant, j'entrevois que quelque soit la discipline, l'université propose un savoir plus fondamental que les écoles de métier. Aujourd'hui l'université n'enseigne aucun métier, mis à part, peut-être, celui de chercheur. C'est précisément ce qui lui est reproché par les corps de métier externes.

Faut-t-il que l'université deviennent une école de métier ? Je ne sais pas. Les choix politiques français ont fait de l'université une structure d'éducation de masse. Etait-ce le bon choix ? Si oui, il me parait raisonnable de se mettre à l'écoute du monde extérieur. En tout cas, je comprends la détresse des étudiants qui sont venus chercher un passeport de travail et qui réalise que l'université n'en délivre pas. Que leur a-t-on dit à propos de

l'université ? Si on leur fait miroiter l'apprentissage d'un métier, on les a égaré. Tant que les cours seront assurés par un personnel qui s'épanouit dans le savoir, leurs contenus seront orientés vers la soif de la connaissance et non vers des "connaissances utiles" pour lesquelles ils bénéficieront d'un "retour sur investissement" ; enfin je crois.

La place de l'université dans notre monde hypothétique futur est assez difficile à saisir. Faudra-t-il changer le personnel enseignant, changer la mission de l'université ou conserver une éducation généraliste ?

Comme nous l'avons vu plus haut, il m'apparaît peu opportun de supprimer la recherche dite fondamentale. Il serait, par conséquent maladroit de supprimer les filières telles qu'on les connaît : c'est-à-dire, orientées vers l'acquisition de connaissances généralistes. Cependant, je suspecte que la majorité des politiques et des étudiants souhaitent transformer l'université en école de métier. Il est donc possible que l'université assure la double compétence de former des étudiants à la recherche et à des métiers hors du monde de la recherche. Cela rentabiliserait le système éducatif, ce qui, comme vu précédemment, est une des priorités de notre hypothétique société du futur.

Enseignants-chercheurs

Alors, dans tout cela, quelle serait la place des enseignants-chercheurs ? On a du mal à imaginer les enseignants-chercheurs, recrutés puis évalués sur leur goût pour la recherche, dispenser des enseignements non tournés vers la quête de savoir. Les universités sont aujourd'hui dans le tourbillon des réformes. Quel sera *in fine* la qualité des réformes proposées ? Je n'en sais rien. On peut penser qu'elles vont chambouler une certaine vision universitaire, celle partagée par les mécontents de la réforme. Certains membres des universités voient cependant un appel d'air dans la mise en place de ces réformes.

Au delà des modalités de la réforme, l'aspect que je souhaite discuter est le but poursuivi par les promoteurs de ces réformes. On pourrait conjecturer sur les avantages et les inconvénients apportés à chacun, mais je soupçonne qu'il existe une direction plus ample et encore peu visible à ces réformes.

Laissons notre esprit imaginer ce que serait l'université de demain. Vraisemblablement, les universités sont toujours pressenties pour servir de structure à l'éducation supérieure de masse. Imaginer que cette éducation de masse sera, dans une société dont la rentabilité est une priorité, complètement assurée pour et par la recherche me semble illusoire. Il est plus probable d'imaginer que certaines filières universitaires seront plus orientées vers l'apprentissage d'un métier et que d'autres seront plus orientées vers la recherche. Cette

distinction se mettra en place soit entre les différentes universités soit au sein de chaque université où plusieurs filières co-existeront.

Dans ces futures hypothétiques universités, quelle serait la place des enseignants-chercheurs ? Je ne crois pas que les chercheurs seront massivement volontaires pour l'enseignement d'un autre métier que le sien. On doit donc imaginer que les enseignants-chercheurs tels que nous les connaissons seront moins nombreux et cantonnés aux filières "recherche". On peut d'ores et déjà imaginer que les formations de métier seront assurées par des enseignants "purs", voire par du personnel des corps de métier concernés.

Evidement, tous ces propos ne sont que de la pure science-fiction. Il convient de rappeler également que si l'on peut envisager l'hypothétique société de demain, on ne saurait envisager celle d'après-demain. Les mentalités changent comme le balancier d'une pendule. Pour preuve, j'invoquerais l'antériorité de l'individualisme sur l'émergence de toute forme société. On peut donc envisager que le mouvement qui s'amorce aujourd'hui sera vraisemblablement inversé dans plusieurs décennies, voire plusieurs siècles. Tant mieux pour certains, tant pis pour les autres.

Hier

Jusqu'ici, j'ai développé, en parallèle, plusieurs sujets ayant trait à différents aspects de l'évolution dite moléculaire. J'ai divisé ces sujets entre *génomique évolutive* et *génétique des populations*. Dans le cadre de ces études, j'ai utilisé, voire développé, des outils bioinformatiques et statistiques nécessaires à mon travail. La plupart de ces sujets ont été menés en collaboration avec d'autres scientifiques que j'ai eu la chance de rencontrer lors de mon parcours. C'est grâce à ce réseau collaboratif que j'ai pu travailler sur des sujets variés tant au niveau des organismes qu'au niveau des problématiques. Cela m'a fait réaliser le lien fort qui existe entre organismes et questions posées. Jusqu'à récemment, j'ai mené les deux types d'approches en parallèle. Ce n'est que plus récemment que j'ai commencé à emprunter les ponts qui relient *génomique évolutive* et *génétique des populations*.

Génomique évolutive

La génomique évolutive cherche à comprendre les mécanismes sous-jacents à la fluidité et à la dynamique des génomes. La modélisation analytique est en général peu représentée dans ce champs disciplinaire, au profit de l'algorithmique bioinformatique. J'ai choisi de présenter trois aspects de génomique évolutive de mon travail de recherche ayant tous trois une relation avec les séquences répétées.

Dynamique des répétitions

Depuis le début de mon doctorat, j'étudie les séquences génétiques qui co-existent en plusieurs exemplaires au sein d'un même génome. Quelque soit sa taille, une *répétition* est constituée d'au moins deux occurrences. Ce sont les *copies* de cette répétition. Si une nouvelle copie se crée, elle est soit due à une accumulation de substitutions soit la conséquence d'un événement de *duplication*.

On peut classer les différentes répétitions selon plusieurs critères : taille (de quelques nucléotides à des génomes entiers), position relative de leurs différentes copies (en tandem, proches, dispersées, inversées, intra- ou inter-chromosomiques), mécanisme de duplication (microsatellites, transposons, polyploïdies, etc.) ou encore impact fonctionnel (gènes, signaux de transcription, etc.). Ce catalogue instructif nous renseigne sur la diversité de structures et de fonctions de ces répétitions. Il nous permet également d'entrevoir leurs intrications complexes. Ayant largement développé ce catalogue dans mon manuscrit de thèse, je le passerais ici sous silence. Je reprendrai néanmoins un schéma récapitulatif qui intégre la dynamique de différents types d'éléments répétés.

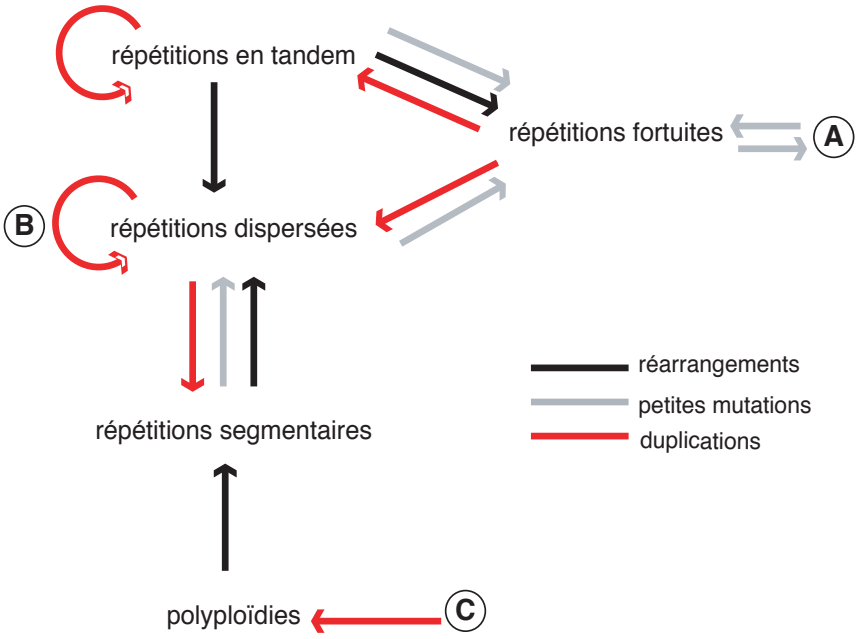


Figure 1: Dynamique "intégrée" de différents types de répétitions.

Ce schéma, mis à jour, est reporté sur la figure 1. Dans ce schéma, les répétitions fortuites désignent toutes les séquences répétées qui sont susceptibles d'apparaître par accumulation de substitutions (celles qui ne sont pas issues de duplications). Ces répétitions sont les amorces nécessaires à la genèse de répétitions plus larges. En effet, les mécanismes de duplication, qui créent des répétitions en tandem (voire en satellites) ou dispersées, nécessitent la présence de micro-homologie. C'est cette micro-homologie que représente les répétitions fortuites. Ces répétitions fortuites peuvent également être les reliquats de plus larges répétitions dont les copies sont si divergentes qu'il est impossible de les distinguer

d'une simple répétition créée par l'accumulation de substitutions indépendantes.

Dans ce schéma simpliste, il n'existe que trois points d'entrée pour les séquences répétées: (A) les répétitions fortuites (qui génèrent en cascade des répétitions plus grandes), (B) les répétitions dispersées qui peuvent être créées par un assemblage fonctionnel opportun (*i.e.* les transposons) et (C) les polyploïdies, issues d'accidents de division cellulaire.

Malgré l'abondance extraordinaire des éléments transposables dans certains génomes (e.g. 60% du génome du maïs), leur apparition *de novo* est vraisemblablement un événement très rare. Il faut donc supposer que leur propagation est principalement le fait de transfert horizontaux. La polyploïdisation stable d'une lignée est probablement également un évènement rare. Elle a été cependant clairement mise en évidence dans plusieurs lignées eucaryotes indépendantes : par exemple les vertébrés (Jaillon *et al.*, 2004), les téléostes (Amores *et al.*, 1998), les angiospermes (Blanc *et al.*, 2000) ou les levures (Wong *et al.*, 2002). Au même titre que la polyploïdie, l'aneuploïdie est un accident chromosomique méiotique fréquent (voir par exemple Griffiths *et al.* (1993), chapitre 9). Néanmoins, au contraire de la polyploïdie, il ne semble pas que ce soit un acteur majeur de l'évolution des génomes. Une multiplication partielle du matériel génétique est vraissemblablement délétère pour l'organisme porteur, argument soutenu par l'organisation en réseau des composants d'un génome dont la duplication partielle pourrait entrainer de graves dysfonctionnements (Shimeld, 1999). Peut-être aussi, plus simplement, le chromosome surnuméraire ne peut pas se maintenir en l'état au cours des divisions méiotiques successives qui ponctuent les générations. Je soupçonne que cette instabilité inhérente est suffisante pour rendre compte de l'absence d'aneuploïdies héritables.

Les événements de duplication les plus courants sont ceux impliquant une amorce de similarité apparaissant et disparaissant par des mutations ponctuelles (substitutions ou insertions/delététions de petite taille). Les mécanismes exacts à l'origine de ces duplications, observées largement dans les génomes, restent encore teintés de flou. Vraisemblablement, pour la création de répétitions en tandem, il s'agirait de recombinaison ectopique ou de dérapage de la polymérase au cours de la réplication. Pour les répétitions dispersées créées *de novo*, la conversion génique est un bon candidat de première intention mais la piste est confuse. La comparaison des séquences de l'homme et du chimpanzé a permis de mettre en évidence l'apparition de répétitions strictes (sans interruption) proches mais non organisées en tandem (Thomas *et al.*, 2004). Aucun mécanisme plausible n'a pu encore être proposé. Recombinaison ectopique, conversion, glissement de la polymérase reposent tous trois sur la présence de similarités pré-existantes permettant d'amorcer le mécanisme. Ils

sont donc tributaires des répétitions fortuites.

Si l'on admet que les répétitions sont souvent créées à partir de répétitions fortuites, on peut prédire que les génomes ayant un grand nombre de ces répétitions fortuites devraient avoir une plus grande quantité de répétitions de tout type. En conséquence, on prédit que la composition nucléotidique d'un génome est fortement corrélée à la densité de répétitions (Achaz *et al.*, 2002). En effet, plus la composition d'un génome $\{p_A, p_C, p_G, p_T\}$ s'écarte de $\{0.25, 0.25, 0.25, 0.25\}$ plus la densité de répétitions fortuites d'une taille donnée augmente, et donc potentiellement, plus le nombre de micro-homologie est grand.

On peut montrer aisément que la probabilité P_I de trouver deux nucléotides identiques à deux sites choisis au hasard est donnée par $P_I = \sum p_i^2$. En général, à l'échelle d'une grande séquence (comme un génome), on observe que $p_A = p_T$ et que $p_C = p_G$; la fréquence d'un nucléotide i est identique dans les deux brins d'ADN (règle de Chargaff, voir par exemple Griffiths *et al.* (1993), chapitre 11). Dans ce cas, la composition d'un génome n'est décrite que par une seule variable p : $p_A = p_T = p/2$ et $p_G = p_C = (1-p)/2$ et la probabilité de trouver deux nucléotides identiques se calcule comme suit : $P_I = \frac{1}{2}(p^2 + (1-p)^2)$. Si chaque site mute indépendamment de son contexte, la séquence est une suite aléatoire de mono-nucléotides. Dans ce cas, la distribution du nombre de couple de copies de taille m^+ (m ou plus), dans une séquence de longueur L et de fréquence p en A+T, peut être assimilée à une loi de Poisson composée à une loi géométrique (voir par exemple Robin *et al.* (2003)). La probabilité de trouver k couples de répétitions peut s'écrire :

$$P(k \text{ couples}) = \frac{N^k}{k!} e^{-N} \tag{1}$$

avec

$$\begin{aligned} N &= L \times (1 - P_I)(P_I)^m \\ &= L \times \left(1 - \frac{p^2 + (1-p)^2}{2}\right) \times \left(\frac{p^2 + (1-p)^2}{2}\right)^m \end{aligned} \tag{2}$$

où, N est le nombre moyen de couple de taille m^+.

Est représentée sur la figure 2, la densité par Mb du nombre de couples de copies de taille m^+ (pour $m = 10, 15$ ou 20) pour des compositions génomiques variables ($p \in [0.2, 0.8]$). Cette figure illustre l'effet majeur de la composition du génome sur la densité d'amorces d'une taille donnée. Plus la composition s'écarte de A+T=50%, plus on attend de répétitions fortuites. L'effet est d'autant plus frappant pour les répétitions fortuites de "grande" taille. Si la taille minimale, à partir de laquelle une répétition fortuite est perçue par un génome comme une amorce de similarité, est identique dans tous les génomes, ceux

qui présentent des compositions en A+T très biaisées vont contenir un très grand nombre d'amorces potentielles ; ce sont des génomes répétés et plus instables (comme celui de *Plasmodium falciparum*, voir à ce sujet DePristo *et al.* (2006)). Ainsi, dans cette idée, la composition d'un génome est une des composantes majeures de prédiction de la densité de répétitions fortuites et, par effet de cascade, des autres séquences répétées au sein un génome.

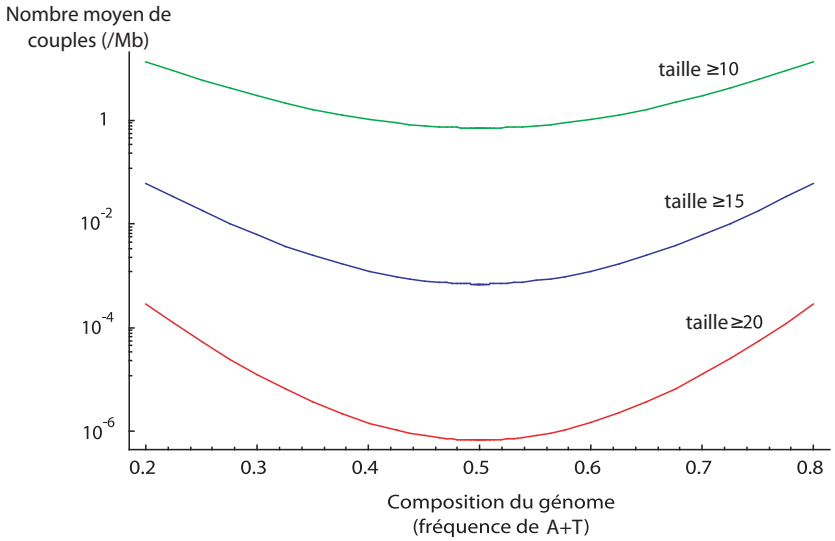


Figure 2: Effet de la composition sur la densité de répétitions fortuites.

Les prédictions faite à partir de ce modèle simple peuvent être testées. Est-ce que la composition A+T d'un génome est un prédicteur fort de la densité de répétitions et de la densité de gènes dupliqués ? Quels liens existent réellement entre composition, répétition génétique et redondance fonctionnelle (e.g. gènes dupliqués).

Divergence de protéine et divergence d'expression

L'étude des duplications de gènes et de leur implication dans l'évolution des répertoires géniques a donné lieu à une littérature extrêmement vaste. Elle s'est démultipliée lors des dernières années et la revue que j'avais constituée, en 2002, dans mon mémoire de thèse est presque obsolète. Historiquement, on peut faire remonter la mise en avant du mécanisme de duplication de gène comme un facteur important de l'évolution aux travaux de Haldane (1932). Bien que plusieurs auteurs aient également participé à l'essor des études de duplications de gènes (pour revue, voir Taylor and Raes (2004)), la communauté scientifique

cite abondamment, à juste titre, le livre de Ohno (1970) comme étape clef du domaine. Je suis tenté d'ajouter aux idées phares plus récentes concernant les duplications géniques, l'idée d'évolution par sous-fonctionnalisation, c'est à dire d'évolution par accumulation de mutations délétères complémentaires entre les copies (Hughes, 1994; Force *et al.*, 1999; Lynch and Force, 2000; Lynch *et al.*, 2001). La venue des données de biologie des systèmes a également permis de replacer les gènes dupliqués au coeur de l'évolution des réseaux génétiques puisqu'ils pourraient permettre d'expliquer certaines caractéristiques clefs des réseaux d'interactions observés aujourd'hui, notamment cette fameuse propriété d'absence d'échelle (*scale-free networks*).

Je me suis intéressé, en collaboration avec Cristian Castillo-Davis, lors de mon stage post-doctoral, à la relation existant entre l'évolution des séquences codantes et celle de leurs séquences régulatrices. Pour cela, nous avons développé une méthode permettant d'estimer la divergence qui existe entre deux promoteurs d'un gène homologue. Si les deux séquences non codantes qui sont comparées sont très peu divergentes et non réarrangées, un simple alignement des deux séquences suffit à calculer la similarité entre les deux séquences. C'est ce qui est classiquement fait pour les séquence codantes elles-même. On peut alors transformer l'identité de séquence en distance en utilisant les méthodes corrigeant pour les substitutions multiples (voir par exemple Nei (1987)). Cependant, les séquences régulatrices ne sont pas soumises aux même contraintes que les séquences codantes. Les éléments de régulations peuvent présenter un ordre différent, voire une orientation différente, tout en conservant leurs propriétés régulatrices. C'est pour cela que nous avons recherché toutes les zones de similarité entre les deux promoteurs, indépendamment de leur orientation ou de leur arrangement. Les résultats d'une telle méthode sont illustrés dans la figure 3.

La mesure de distance que nous avons adoptée est la proportion de la séquence non-couverte par ces zones de similarité partagée. Nous l'avons nommée d_{SM}, pour *distance of Shared Motifs*. Lorsque $d_{SM} = 0$, cela signifie que l'intégralité du promoteur est conservé. A l'inverse, lorsque $d_{SM} = 1$, aucune zone similaire ne peut être mise en évidence.

Nous avons pu montrer que les zones de similarité ainsi mises en évidence contenaient très souvent des motifs connus de transcription (voir pour illustration la figure 3). De plus, nous avons montré que notre métrique de divergence des promoteurs, d_{SM}, est corrélée avec une différence d'expression pour des paralogues du génome de *Caernorhabditis elegans*. Ceci suggère que mesurer d_{SM} entre deux promoteurs homologues est une mesure indirecte de la divergence du patron d'expression entre ces gènes.

La comparaison de la divergence des séquences protéiques (*i.e.* d_N) et de celle des promo-

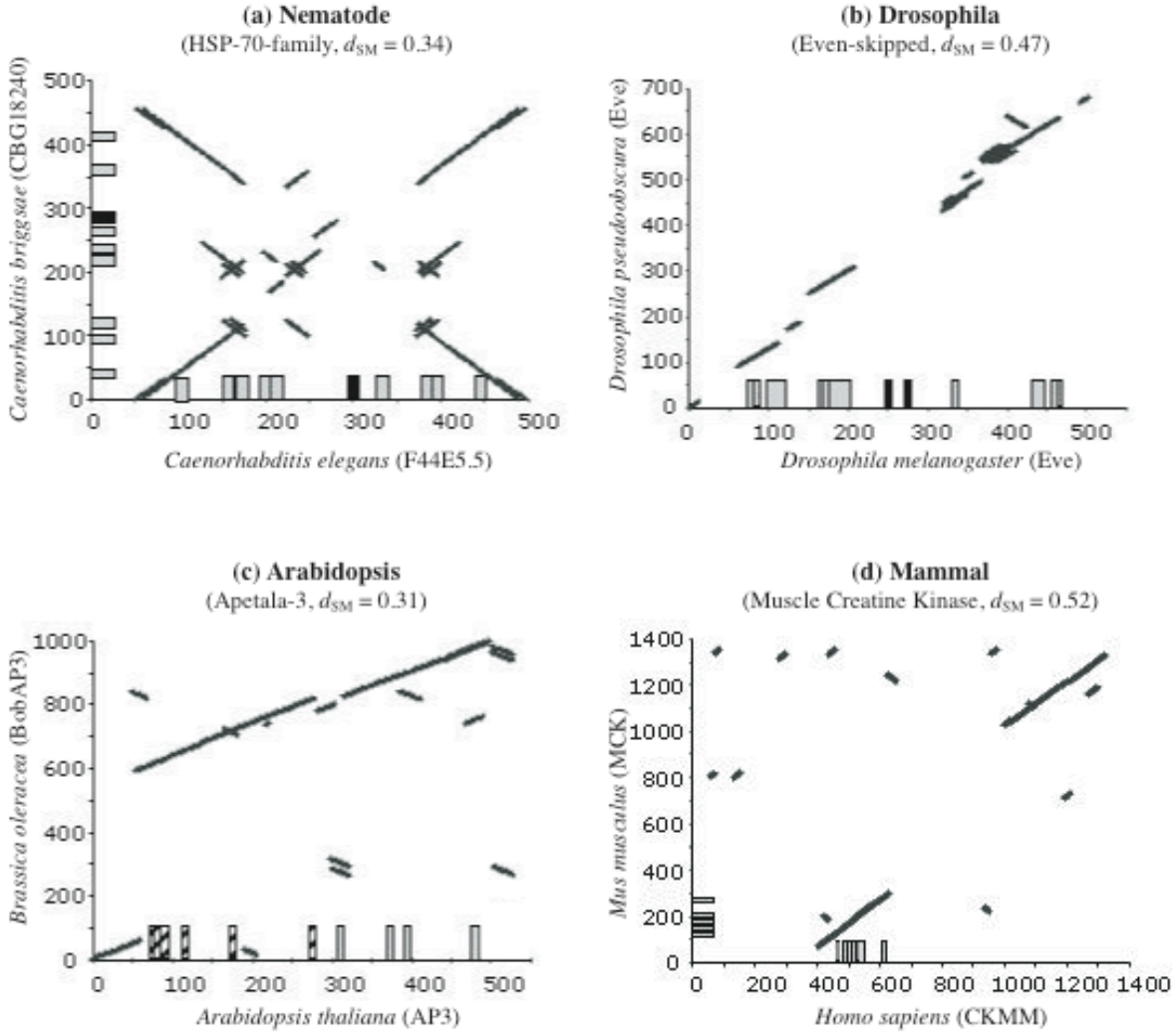


Figure 3: Dot plot des zones de similarité entre des promoteurs de gènes homologues. Les rectangles gris indiquent la position de facteurs de transcription lorsqu'ils sont connus.

teurs (d_{SM}) a révélé une faible corrélation positive. Ainsi, il existerait un certain couplage entre l'évolution d'une protéine et celle de son patron d'expression. Nous avons observé ce résultat aussi bien pour les orthologues des génomes de *C. elegans* et *Caernorhabditis briggsae* (ayant divergé il y a 50 millions d'années environ) que pour les paralogues au sein de chacune des deux espèces.

Il est cependant naturel d'obtenir un tel résultat. En effet, la divergence est une fonction croissante monotone du temps. La séquence codante et son promoteur divergent depuis un même temps identique et ont donc tout deux, en moyenne, subit autant de mutations. Leur divergence est donc naturellement corrélée. Il faut donc corriger pour cette corrélation naturelle liée au temps de divergence. Un estimateur de ce temps est donné par d_S, la divergence des sites synonymes de la séquence codante. En calculant une corrélation partielle entre d_N et d_{SM}, corrigée pour l'effet du temps (d_S), le résultat reste robuste pour les orthologues mais disparaît pour les paralogues.

Le couplage entre les divergences des séquences protéiques et promotrices pour les gènes orthologues suggère une certaine unité fonctionnelle, au regard de l'évolution, entre un gène et son patron d'expression. Si l'un tend à diverger rapidement, l'autre présente une tendance similaire. Pour des paralogues, ceci n'est pas vrai. Une première explication serait d'ordre méthodologique : tous les orthologues entre deux espèces ont divergé depuis un temps similaire alors que des duplicats présentent des temps de divergence très variables. La correction seraient donc plus forte (meilleure ?) pour les paralogues. Néanmoins, la divergence des gènes après un événement de duplication n'obéit pas aux même contraintes que celle de gènes après un événement de spéciation. Dans le cas de gènes dupliqués, les deux copies du gène présentent une certaine redondance fonctionnelle qui peut être perdue (symétriquement ou non) ou mener à la genèse d'une fonction (légèrement) différente de l'originale (à court terme, les chiens ne font pas des chats !). Pour le cas des orthologues, le couplage entre promoteur et séquence protéique est cependant faible. Pourquoi ?

Il faut, en première intention, préciser que la métrique que nous avons utilisé pour mesurer la divergence entre les promoteurs est très rudimentaire. On peut, d'ores et déjà, envisager des améliorations qui devrait la rendre plus proche des méthodes utilisées pour calculer des divergences entre séquences codantes. Premièrement, le score minimum d'un segment de similarité était fixé arbitrairement à combien sur des critères empiriques. Fondé sur des approches probabilistes, proches de celles mentionnées plus haut pour les répétitions, on peut calculer un score seuil en deçà duquel la similarité est attendue par hasard (de la similarité fortuite). Ce seuil serait choisi sur la base d'une probabilité et

remplacerait avantageusement notre choix empirique. En second lieu, à l'instar d'une identité de séquence, l'estimateur de distance que nous avions proposé est borné entre 0 et 1. Nous pourrions donc transformer cette métrique en une distance évolutive bornée entre 0 et $+\infty$, comme classiquement fait dans le domaine d'évolution moléculaire.

Quand bien même la métrique serait amendée par les deux points mentionnés ci-dessus, il est possible que la corrélation entre évolution du promoteurs et évolution de la séquence protéique reste faible. Cette corrélation est d'ailleurs nulle lorsque l'on compare des paralogues. Alors, quelle est la pertinence biologique d'un tel couplage ? Quel modèle théorique serait compatible avec ces résultats ? Si l'étude du couplage entre divergence d'expression et divergence protéique a été l'objet de plusieurs analyses de données (voir, par exemple, Wagner (2000); Gu *et al.* (2002); Makova and Li (2003)), peu de modélisation n'a encore été proposée. Des observations sans modèle forment une excellente première étape mais ne nous renseigne que peu sur les causes de nos observations.

Il serait intéressant de poursuivre ce projet qui s'était mis en place à un moment où les études sur les gènes dupliqués et sur les divergences d'expression ont envahi la littérature. Je ne souhaitais pas me lancer participer à une course scientifique. J'ai le sentiment, m'étant éloigné un peu de ce domaine que l'engouement scientifique s'est peut-être un peu écarté de ces sujets. J'envisage de reprendre ce type d'études dans le futur.

Des répétitions dans les gènes

Depuis 2006, je co-encadre avec Pierre Netter, la thèse d'Etienne Loire sur l'étude des répétitions intragéniques. Etienne prévoit de soutenir sa thèse fin octobre 2009. Son travail porte sur les séquences microsatellites présentes au sein des séquences codantes et notamment sur leur impact sur la robustesse et la mutabilité des gènes. Définir la robustesse n'est pas un exercice facile. Cependant, dans notre cadre génétique, elle doit être comprise comme la robustesse d'un gène vis à vis des mutations qu'il subit. Pour un métazoaire, ces mutations peuvent subvenir aussi bien dans la lignée germinale (coût sélectif différé) que dans la lignée somatique (coût sélectif immédiat).

A chaque génération, un gène subit des mutations diverses (substitution, insertion ou délétion, réarrangement, etc.). En moyenne, si l'on fait l'approximation que le taux de substitution pour un mammifère est d'environ 10^{-8} /base / génération (Drake *et al.*, 1998), une séquence codante humaine, d'environ 1 kb, est mutée par substitution toutes les 10^5 générations. On peut mesurer la mutabilité d'un gène par le nombre moyen de mutations subit par ce gène à chaque génération. On peut d'emblée noter qu'au moins

deux facteurs vont influer grandement la mutabilité des gènes : (1) leur taille et (2) leur composition nucléotidique, chaque motif nucléotidique ayant une mutabilité propre (par exemple PLATT (2004)).

Chaque substitution, si elle n'est pas silencieuse, peut modifier la séquence protéique (mutation faux-sens), voire l'interrompre (mutation non-sens). Considérons que la robustesse d'un gène est sa capacité à conserver un phénotype moléculaire identique pour une mutabilité donné. En ce sens, la robustesse d'un gène peut être rapprochée de la canalisation telle qu'elle est proposée par WADDINGTON (1942)[2].

Nous nous sommes intéressés aux gènes qui présentent des séquences microsatellites dans leurs séquences codantes. Ces séquences microsatellites ont une probabilité de muter par insertion ou délétion beaucoup plus importante que le taux de substitution. Ainsi, comme $\approx 1/20$ des codons sont des codons STOP, un gène de 1 kb présente un taux de substitution non-sens d'environ $\tau_{\text{sub}} \approx 5.10^{-7}$ par génération. Les estimations du taux d'insertion ou délétion d'un microsatellite varient entre $\tau_{\text{indel}} \in [10^{-3} - 10^{-6}]$ par génération, c'est à dire entre 10 et 10 000 fois plus que le taux de substitution. Lorsque l'unité répétée n'est pas un multiple de 3, l'insertion ou la délétion d'une unité entraîne un décalage de la phase de lecture et donc l'apparition d'un STOP prématuré (sauf cas exceptionnel).

Il découle de ce rapide calcul que la présence d'un unique microsatellite au sein d'une séquence codante *a minima* décuple sa chance d'être interrompue par une mutation non-sens. En conséquence, les gènes contenant un microsatellite non multiple de 3 peuvent être qualifiés d'hypermutables. Ils ont une chance très accrue d'être muté dans la lignée somatique et donc de conduire à un phénotype déficient (ceci est également vraie pour la lignée germinale). C'est sans doute pour cette raison que les séquences microsatellites sont généralement sélectionnées négativement dans les séquences codantes : elles tendent à être plus petites et moins nombreuses qu'attendu sous un modèle aléatoire neutre (*i.e.* substitutions indépendantes). Néanmoins, on peut dénombrer 1,935 gènes humains (8.7% du génome) contenant un microsatellite non multiple de 3 et suffisamment long pour être instable (e.g. 8 unités pour un satellite de mono-nucléotides). Pour ces gènes, la probabilité d'être interrompu par une mutation non-sens est *a priori* bien plus grande pour les autres gènes.

Nous avons recherché parmi ces 1,935 gènes "hypermutables", des fonctions biologiques surreprésentées. Pour cela, nous avons développé une méthode de recherche de surreprésentation de fonctions. Utilisant les annotations standardisées de *Gene Ontology*,

[2]Bien que sa définition s'adresse plus à des patterns de développement, elle englobe explicitement la canalisation génétique.

nous avons montré que ces gènes hypermutables sont préférentiellement annotés comme appartenant au maintien de l'intégrité du génome et au cycle cellulaire. Pourquoi existe-t-il des classes de gènes plus mutables que d'autres ?

Une explication attrayante serait que les microsatellites codants seraient sélectionnés positivement dans certains gènes. L'observation que beaucoup des gènes impliqués dans le *mismatch repair* ont de tels microsatellites dans leurs séquences codantes ont amené certains auteurs à proposer l'existence de mutateurs chez les eucaryotes (Chang *et al.*, 2001). Ces gènes du *mismatch repair* sont impliqués dans la correction des substitutions et leur déficience entraîne une forte hausse du taux global de mutation du génome. Ainsi, on pourrait faire l'hypothèse que l'augmentation du taux global de mutation en condition de stress serait associée à un avantage sélectif. Hypothèse osée néanmoins car la présence de recombinaison dans les génomes eucaryotes abolit *a priori* totalement l'avantage de ces gènes modulateurs du taux de mutation global (*i.e.* les mutateurs) (Tenaillon *et al.*, 2000).

Une explication alternative serait que c'est un simple fait du hasard. Par exemple, la nature de ces gènes (longueur et composition) pourrait être la cause de cette observation. En effet des gènes longs et ayant une composition en nucléotide biaisée (différente de $\{0.25, 0.25, 0.25, 0.25\}$) présentent une probabilité accrue de contenir des microsatellites. Il est possible de calculer la probabilité qu'un microsatellite de taille m^+ soit présent dans une séquence de longueur L et de composition $\{p_A, p_C, p_G, p_T\}$. Pour un groupe de gène, on peut calculer une fraction attendue de gènes hypermutables. Le modèle utilisé pour ce calcul est très proche de celui développé ci-dessus pour les répétitions fortuites (équation 2) ; cependant, ici, les fréquences des quatre bases sont différentes.

Nous avons, pour chaque classe fonctionnelle de gènes, comparé la fraction observée du nombre de gènes présentant un microsatellite codant à la fraction attendue. Les résultats (figure 4) révèlent que dans tous les regroupements fonctionnels (ou presque), les microsatellites sont moins abondants qu'attendu par un modèle neutre. Les fonctions que nous avons trouvé surreprésentées ont toutes, compte tenu de leurs effectifs, des fractions attendues élevées. Les gènes de ces fonctions sont typiquement plus grand et/ou plus biaisés que les autres gènes.

Aucune fonction annotée dans le génome humain ne possède significativement plus de gènes hypermutables qu'attendu sous un modèle neutre. Au vu de ces résultats, l'hypothèse d'une sélection positive dans certains type de fonction me paraît improbable. La figure 4 montre que les différentes fonctions révèlent néanmoins une grande variabilité dans le ratio

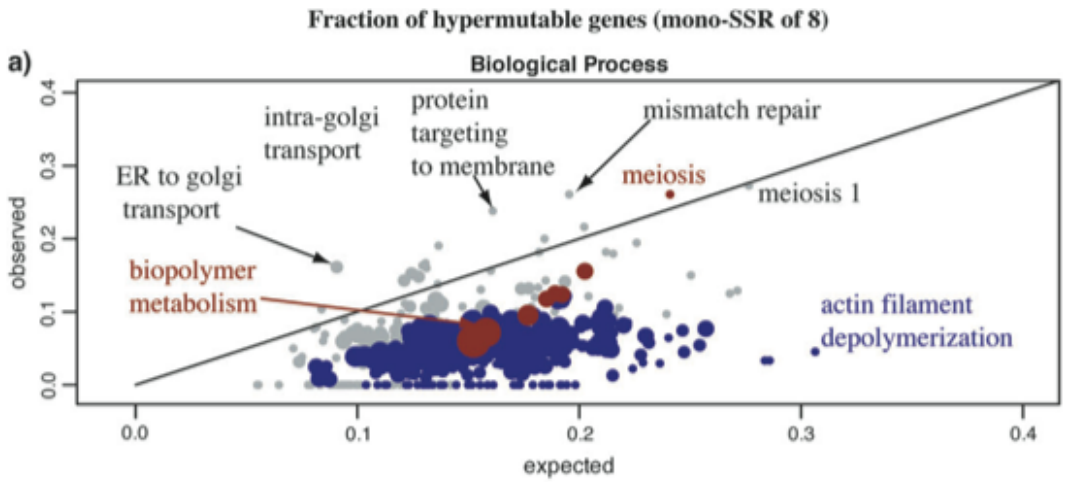


Figure 4: Fraction observée et attendue de gènes contenant des microsatellites *a priori* instable pour toutes les fonctions biologiques annotées dans le génome humains. La taille des points reflète le nombre de gènes annotés pour cette fonction. Les fonctions ayant moins de 20 gènes ont été exclues du dessin. En bleu, les fonctions pour lesquelles l'observé est significativement différent de l'attendu. En rouge, les fonctions détectées comme surreprésentées en gènes hypermutables.

observé/attendu. Certaines fonctions, comme *actin filament depolymerization*, montrent un ratio très bas ; d'autres, comme le *mismatch repair*, ont un ratio plus grand. Nous avons pu montré que cette hétérogénéité n'est pas le simple fait du hasard. En conclusion de ce travail, on peut penser que la pression de sélection qui agit pour purger ces microsatellites des séquences codantes varie d'une fonction à une autre. Les gènes de certaines fonctions toléreraient mieux ces séquences hypermutables que d'autres.

Génétique des populations

Depuis mon stage post-doctoral, je me suis intéressé à l'évolution des séquences génétiques au sein des populations. Comme expliqué ci-dessus le traitement massif des données de génomique nécessite la mise en place d'outils bioinformatique. Cette mise en oeuvre se fait, bien souvent, au détriment, d'une approche plus mathématique. Pour la génétique des populations, la situation est plutôt inversée. Il y a, bien sûr, de la modélisation fondée sur des simulations, qui requiert un certain bagage informatique, mais la composante algorithmique en génétique des populations reste souvent mineure. La génétique des populations

est une discipline séculaire qui prend ces racines dans une tradition mathématique et non informatique. Les problématiques abordées sont, en général, tournées vers la modélisation analytique. J'ai abordé la génétique des populations à travers l'analyse de données ; en particulier, j'ai analysé des séquences issues de populations de HIV-1 prélevées à différents temps. J'ai, par la suite, participé à l'analyse d'autres données populationnelles. Mû par l'envie de plonger plus avant dans la modélisation, j'ai abordé des problèmes liés à la théorie de la coalescence, notamment en développant des tests de neutralité.

J'ai choisi de présenter deux aspects de mes travaux ayant trait à la génétique des populations. Je présenterai d'abord mes travaux concernant les tests de neutralité, puis poursuivrai par ceux concernant les échantillons hétérochroniques des populations de HIV-1.

Tests de neutralité

La théorie neutraliste, depuis son avènement, a pris une importance capitale en génétique des populations et, plus généralement, en génétique évolutive. La forte popularité de cette théorie a imposé le modèle neutre comme modèle de référence. Plus précisément, le modèle de référence est celui d'un équilibre mutation-dérive. Lorsqu'aucune information n'est disponible pour une population, le modèle de référence est adopté par défaut. C'est l'hypothèse nulle que l'on cherchera à rejeter à l'aide des tests dits de neutralité. La mise en évidence d'événements de sélection naturelle s'accompagne aujourd'hui nécessairement du rejet du modèle neutre.

Dans ce modèle de réference, les polymorphismes génétiques apparaissent par mutation et leurs fréquences varient, voire se réduisent à zéro, uniquement sous l'action de la dérive génétique. Deux approches ont été développées pour étudier le modèle neutre. Dans la première approche, le temps se déroule dans le sens prospectif ; c'est l'approche "classique". Dans la seconde approche, on remonte le temps dans le sens rétrospectif ; c'est l'approche "coalescence". La simplicité qui accompagne la modélisation rétrospective a, sans aucun doute, été à l'origine de l'essor extraordinaire qu'a connu la théorie de la coalescence dans les dernières décennies (voir les livres de Hein *et al.* (2005); Wakeley (2009)).

Dans le domaine de la coalescence, on s'intéresse à la généalogie de loci homologues échantillonnés dans une population. La théorie de la coalescence est une approche probabiliste qui vise à caractériser la distribution des généalogies sous un modèle particulier. Dans le modèle de référence, on montre que, dans une population de N chromosomes, la hauteur moyenne de l'arbre généalogique est d'environ $2N$. Ainsi, il faut remonter env-

iron $2N$ générations en arrière pour trouver l'ancêtre commun à un échantillon, résultat extensible à l'ensemble de la population. Plus la taille de la population est grande, plus la profondeur de l'arbre est importante. Cet arbre représente l'apparentement entre les individus, soit leur consanguinité. Ainsi, comme le suggère notre bon sens, la consanguinité est une simple conséquence de la taille de la population. Coalescence et consanguinité sont deux facettes similaires de la dérive génétique, vue rétrospectivement.

La diversité génétique mesurable, comme par exemple l'hétérozygotie, le nombre de sites polymorphes ou encore la différence moyenne en deux séquences, est la conséquence des mutations qui se produisent le long des lignées évolutives de l'arbre de coalescence. Lorsque les individus sont très consanguins, ils sont génétiquement proches ; les temps de la généalogie qui les lie sont courts. Sous le modèle neutre, les indices de diversité attendus sont tous des fonctions du paramètre θ, caractéristique de l'équilibre mutation-dérive. Ce paramètre est défini comme $2N\mu$, où N est le nombre de chromosomes et μ le taux de mutation par génération. Il faut souligner que ce paramètre n'a réellement de sens que dans le cadre du modèle neutre de référence.

A partir d'un alignement de séquences, on peut aisément mesurer le nombre de sites polymorphes (S), la différence moyenne entre deux séquences (π) ou toute autre indice de la diversité génétique. Si l'on admet que les séquences échantillonnées évoluent sous le modèle de référence, chacun de ces indices peut servir à estimer le paramètre θ. On dispose donc d'autant d'estimateurs qu'il y a d'indices de diversité. Par exemple, $\hat{\theta}_S$ est l'estimateur calculé à partir du nombre de sites polymorphes (Watterson, 1975) et $\hat{\theta}_\pi$ est l'estimateur calculé à partir de la différence moyenne entre deux séquences (Tajima, 1983). Si le modèle de référence est pertinent tous ces estimateurs sont égaux en moyenne.

Dans le cadre de tests de neutralité, pour un alignement de séquences donné, on teste la vraisemblance du modèle de référence à partir de la différence entre deux estimateurs : $t = \hat{\theta}_1 - \hat{\theta}_2$. Sous le modèle de référence, $E[t] = 0$. En simulant le modèle neutre, on calcule la distribution attendue de t. Si le t_{obs} observé est en dehors de l'intervalle de confiance associé au risque choisi, on rejette, avec ce risque, le modèle neutre. En, cela on teste "la neutralité" de l'échantillon considéré.

J'ai, en premier lieu, étudié l'impact des erreurs de séquences sur les différents estimateurs de θ et sur les tests de neutralité qui en découlent. Ce travail a été motivé par le grand nombre d'erreurs de séquences que j'avais rencontré lors de mon travail sur les séquences de HIV-1. En effet, ces séquences sont obtenues à partir de clones issus de produits de RT-PCR, conditions particulièrement favorables à l'apparition d'erreurs de séquences. Si

l'on admet que les erreurs de séquences se répartissent uniformément le long des séquences et si le taux d'erreur n'est pas trop élevé, les erreurs de séquences sont typiquement des singletons (mutations présentes sur une seule séquence de l'échantillon).

En assimilant les erreurs de séquences à des singletons artéfactuels, j'ai montré que les différents estimateurs de θ n'étaient pas affectés de la même ampleur par ces erreurs. En conséquence, les différences entre deux estimateurs, t, n'ont plus une moyenne nulle sous le modèle de référence. En effet, à cause de ces erreurs expérimentales, on augmente fortement le risque de rejeter le modèle neutre. Pour exemple, considérons le premier test de ce type, D, proposé par Tajima (1989). D est construit sur la différence $\hat{\theta}_\pi - \hat{\theta}_S$. Les erreurs de séquences augmente plus $\hat{\theta}_S$ que $\hat{\theta}_\pi$. En conséquence, $E[D] < 0$. Ceci est d'autant plus prononcé que le taux d'erreur de séquence est fort. Comme illustré par la figure 5, lorsque le taux d'erreur de séquences atteint un certain seuil, le test basé sur D rejette le modèle neutre bien plus souvent qu'attendu. Si l'on ignorait la présence de ces erreurs, on pourrait conclure trop hâtivement, par exemple, que l'échantillon provient d'une population en croissance.

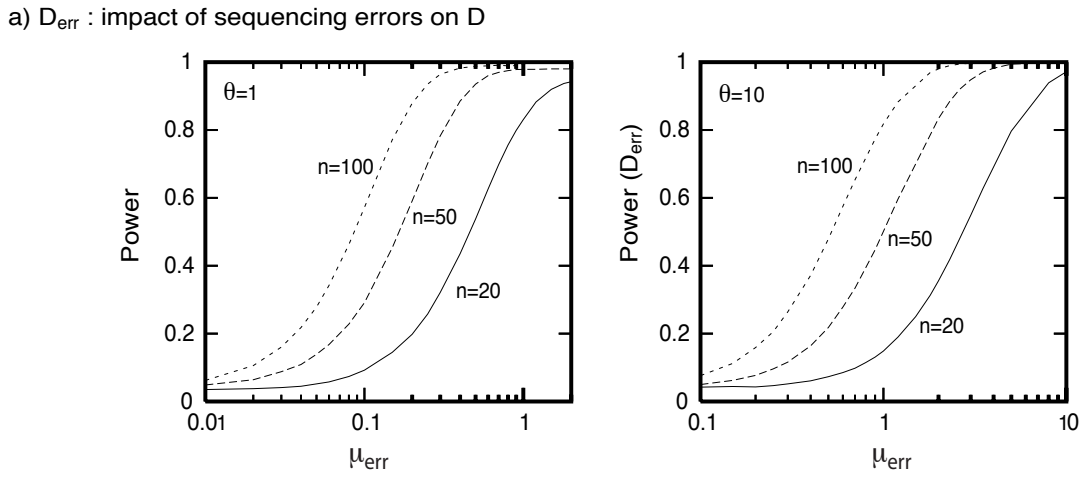


Figure 5: **Effet des erreurs de séquences sur D.** Puissance du test de neutralité basé sur D à rejeter le modèle neutre. Pourtant, ici le modèle est celui de référence auquel sont ajoutées des erreurs de séquences (des singletons artéfactuels). Le taux d'erreur est donné par séquence. Augmenter le nombre de séquences augmente le nombre d'erreurs.

Si, entre deux séquences, 1 différence sur 10 est une erreur de séquence, les chances de rejet du modèle de référence sont de 40-50%, pour un échantillon de 50 séquences. Cela correspond à un taux d'erreur de l'ordre de $[10^{-5}, 10^{-4}]$ erreurs /bp pour un échantillon

provenant d'une population humaine et de $[10^{-4}, 10^{-3}]$ erreurs /bp pour un échantillon provenant d'une population de HIV-1. Pour palier à ces problèmes d'erreurs de séquences, j'ai développé de nouveaux estimateurs : $\hat{\theta}_{S_{-\eta_1}}$ et $\hat{\theta}_{\pi_{-\eta_1}}$ dérivés du nombre de sites polymorphes S et du nombre de différences entre deux séquences π, mais pour lesquels les singletons sont ignorés. Si l'on dispose d'un groupe externe, on peut faire la différence entre les singletons ancestraux et dérivés ; dans ce cas, seuls les singletons dérivés doivent être ignorés, ce qui permet de construire de nouveaux estimateurs : $\hat{\theta}_{S_{-\xi_1}}$ et $\hat{\theta}_{\pi_{-\xi_1}}$. En utilisant ces nouveaux estimateurs, j'ai construit de nouveaux tests de neutralité (Y et Y^*) qui sont *a priori* insensibles aux erreurs de séquences. Ces tests ignorent une partie des données. Il présentent donc une certaine diminution de la puissance à rejeter le modèle de référence, notamment dans le cas de scénarios évolutifs qui augmente le nombre de singletons (e.g. population en croissance). Cependant, ils présentent l'avantage non-négligeable de ne pas être affectés par les erreurs de séquence si souvent présentes dans les alignements de séquences.

Plus récemment, j'ai travaillé sur une forme générique des estimateurs de θ dérivés du spectre de fréquence. Celui-ci est la distribution du nombre de mutations présentes sur 1, 2, 3, ... ou n séquences de l'échantillon. Le calcul du spectre de fréquence complet nécessite un groupe externe, car il faut faire la différence entre les mutations dérivées de fréquence i/n (avec $i < n/2$) et les mutations ancestrales de fréquence $1 - i/n$. Ce spectre de fréquence complet est noté $\boldsymbol{\xi}$, ξ_1 étant les singletons dérivés (mutations dérivées présentes sur une seule séquence). Lorsque l'on ne dispose pas de groupe externe, le spectre de fréquence est plié et est noté $\boldsymbol{\eta}$. η_1 représente l'ensemble des singletons, qu'ils soient dérivés ou ancestraux.

On peut montrer que tous les estimateurs de θ dérivés du spectre de fréquence sont une combinaison linéaire du "spectre de θ". Ce spectre de θ est construit, à partir du spectre de fréquence comme : $\hat{\theta}_i = i\xi_i$. Tous les estimateurs de θ dérivés du spectre de fréquence peuvent s'écrire comme $\sum_i \frac{\omega_i}{\sum \omega_i} \hat{\theta}_i$. Ainsi, un estimateur de θ est simplement un vecteur de poids normalisé associé au spectre de θ. Comme les tests de neutralité ne sont que de simples différences entre deux estimateurs, ils sont eux même caractérisés par des vecteurs de poids associés au spectre de θ. Ces vecteurs de tests, $\boldsymbol{\Omega}$, sont construit comme la différence entre les deux vecteurs normalisés. La somme de leurs éléments est 0 et aucune des valeurs ne peut être plus grande que 1 ou plus petite que -1. Tout vecteur répondant à ces critères est un test de neutralité. Des équivalents existent pour les spectres de fréquence pliés. Sur la figure 6 sont donnés quelques vecteurs $\boldsymbol{\Omega}$ associés à des tests de

neutralité classiques.

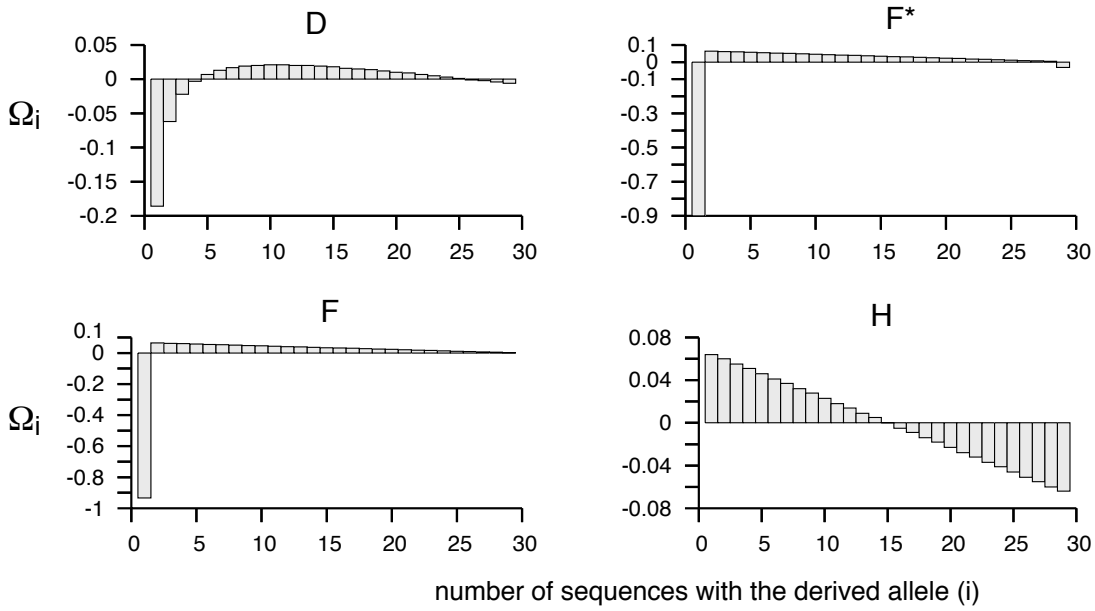


Figure 6: **Vecteurs de poids caractéristiques de tests de neutralité classiques.**

En utilisant cette forme générique, j'ai développé de nouveaux tests plus puissants que ceux qui existent déjà ou répondant à des contraintes particulières (*e.g.* corrigeant pour le biais de découverte des SNP humains). Le développement de nouveaux estimateurs et de nouveaux tests à l'aide de la forme générique est encore au stade expérimental. Je me suis, jusqu'ici, efforcé de montrer l'unicité des tests de neutralité et comment le développement de nouveaux tests est potentiellement très facile.

Ce travail devrait permettre la construction de tests puissants, prenant en compte des contraintes sur les données, comme par exemple la présence d'erreurs de séquence. Pour cela, il faudra mettre en oeuvre une procédure d'optimisation rationnelle qui permettra de créer de nouveaux estimateurs et tests répondant à certaines exigences. Par exemple, un des problèmes des tests Y et Y^* est qu'ils ont peu de puissance pour détecter les populations en croissance. Il faudrait donc trouver de nouveaux tests sensibles à l'excès de mutations à basse fréquence mais qui ignorent les singletons suspects. Que faut-il optimiser exactement ? Comment procéder ? Cela est encore peu clair.

Echantillons hétérochroniques, le cas du HIV-1

Comme expliqué ci-dessus, à partir de séquences échantillonnées à un temps donné, il est aisé de calculer des indices de diversité. Toutes ces mesures sont toujours le reflet de deux forces évolutives qu'il est impossible de séparer. Par exemple, sous le modèle neutre, on estime le paramètre composite θ, produit du taux mutation et du temps de coalescence. Comme illustré sur la figure 7, on montre que le nombre moyen de différence entre deux séquences a pour valeur attendue θ.

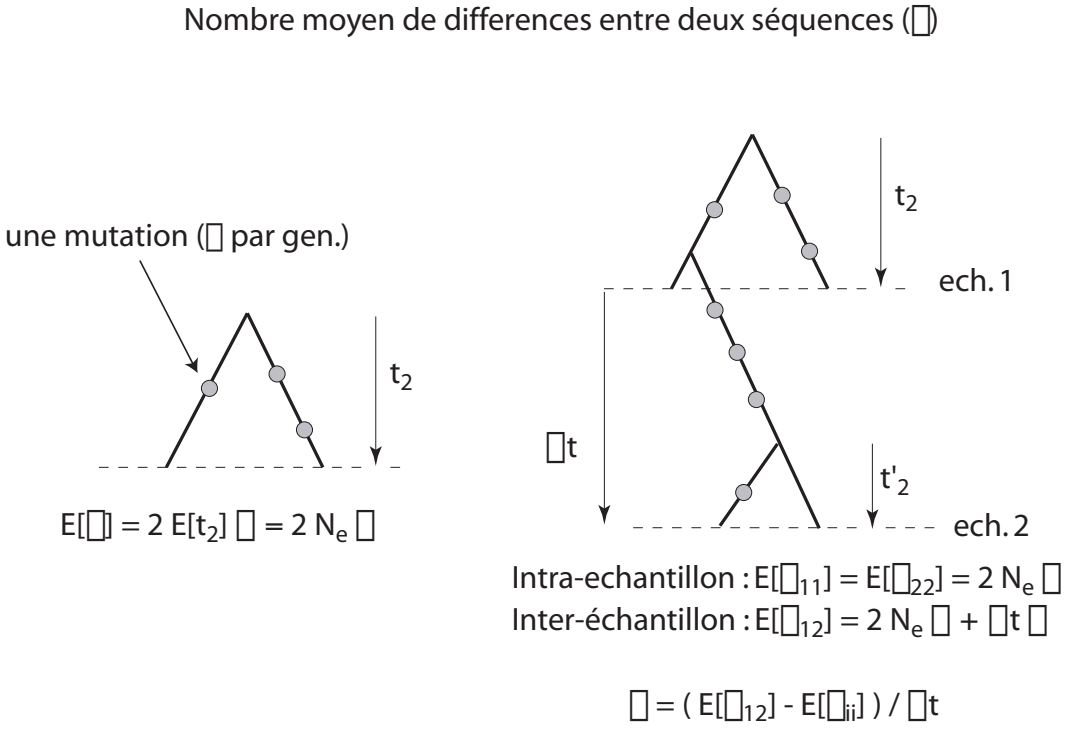


Figure 7: **Equilibre Mutation-Dérive en coalescence.** La diversité attendue dans un échantillon, comme, par exemple, le nombre moyen de différences entre les séquences prises deux à deux (π), dépend du temps de divergence entre les séquences (N_e dans le modèle de référence) et du taux de mutation par génération pour le locus (μ). Lorsque l'on dispose d'échantillons hétérochronique, il est possible de séparer taux de mutation et temps de divergence. Ainsi, on peut estimer μ et N_e indépendamment.

Si l'on dispose d'échantillons hétérochroniques (la même population prélevée à temps différents), il devient possible de découpler temps et taux de mutation (voir Fu (2001) et figure 7). Comme le nombre de générations entre les échantillons est connu, la divergence entre les échantillons permet d'estimer le taux de mutation et la taille efficace

indépendamment. Ainsi, les échantillons hétérochroniques présentent cette extraordinaire propriété de permettre la séparation du temps (ici lié à la dérive) et du processus considéré (ici la mutation).

L'équipe animée par J Coffin au *Drug Resistance Program* (NIH, USA) a mis au point une technique permettant de séquencer un fragment de génome des virus HIV-1 à partir d'un extrait de plasma sanguin. Plusieurs patients ont été prélevés à différentes dates. Je me suis, en premier lieu, focalisée sur les échantillons d'individus infectés depuis de nombreuses années qui maintiennent une virémie basse sans aucun traitement. Ces individus sont porteurs "sains" du virus : ils n'ont déclaré aucun symptôme de la maladie mais présentent une virémie d'environ 10^4 virus/ml ; ce qui représente une population totale d'au moins 10^7 virus circulant dans la plasma.

Nous avons analysé les échantillons de deux patients à l'aide d'un test de structure (Hudson *et al.*, 1992). Les tests de structures sont généralement utilisés pour mettre en évidence une structuration spatiale. Si la population est constituée de sous-population définie *a priori*, on peut tester l'hypothèse que ces sous-populations proviennent d'une unique population panmictique. Dans notre étude, nous avons utilisé le test pour déterminer si deux échantillons prélevés sur la même population, mais à des temps différents, pouvaient être considéré comme homogène. Nous avons utilisé la probabilité de panmixie (d'homogénéité) comme une mesure du *turn-over* de la population virale. Après deux années ($\approx$ 1000 générations), la fréquence des polymorphismes de la population virale a suffisamment changé pour que le test utilisé rejette systématiquement l'hypothèse d'homogénéité.

Nous avons ensuite comparé ce résultat aux attentes théoriques obtenues par simulation du modèle de référence. La comparaison entre observé et attendu nous a permis d'estimer que la taille efficace intra-individu N_e du virus était d'environ $[10^3, 10^4]$. La taille efficace est une mesure de la force de la dérive génétique. Ici, la force de la dérive observée est 1000 à 10 000 fois moindre que celle attendue par la taille réelle de la population ($N_e \approx 10^3 - 10^4$ versus 10^7 virus). La méthode, illustrée sur la figure 7, basée sur le nombre moyen de différences entre les séquences donne des résultats tout à fait comparables.

Ainsi, le *turn-over* de la population virale est beaucoup plus rapide qu'attendu par un modèle de dérive simple où la taille de la population et la taille efficace seraient proches. De nombreuses explications peuvent être avancées pour expliquer cette différence entre taille réelle et taille efficace : structuration de la population, nombre de virus donnant réellement naissance à une nouvelle génération, démographie complexe, etc. Cependant, celle qui a

retenu le plus mon attention, est celle qui a été suggérée par les travaux de Gillespie (1991, 2000a,b, 2001).

Ces travaux montrent que des balayages sélectifs récurrents dans des populations infinies aboutissent également à ce type de résultat. Au delà du problème de la taille efficace de HIV-1, ces travaux m'ont fait prendre conscience d'un problème majeur d'évolution moléculaire : pourquoi le modèle de référence est-il celui d'un équilibre mutation-dérive ? Les fluctuations des fréquences alléliques au cours du temps sont, par défaut, interprétées comme l'effet de la dérive génétique. Non comme l'action de la sélection naturelle. Or, il a été très clairement montré qu'un modèle, où le temps est ponctué d'événements sélectifs, mime quasi-parfaitement les effets de la dérive génétique. Alors, je me range à l'avis circonspect de Gillespie. Pourquoi avoir adopté le modèle neutre comme référence ?

Méthodologie

Faut-il utiliser des méthodes existantes ou mettre en place de nouvelles méthodes ? Voilà un dilemme que l'on se pose très fréquemment. Chaque question scientifique nécessite l'emploi d'une méthode précise et chaque méthode suggère une classe de questions. A l'instar de la relation entre espèce étudiée et question posée, le duo méthode-problème engendre une spirale synergique. Je ne peux nier avoir développé un certain goût à la mise en place de méthodes. J'utilise, importe et développe des méthodes informatiques et mathématiques pour répondre à des questions biologiques.

Algorithme

Dans le cadre de mon travail, j'ai développé plusieurs outils informatiques pour les appliquer tantôt à la génétique des populations (outils d'analyse mais également outil de simulation), tantôt à la génomique (outils d'analyse de données). Je ne voudrais pas ici me perdre dans les algorithmes investis pour mettre en place ces outils, cependant je pense que c'est un aspect important de mon travail. J'ai choisi de présenter brièvement l'algorithme de détection de répétitions développé tout au long de ma thèse et sur lequel je continue à travailler.

Au début de mon doctorat, il n'existait pas de méthode satisfaisante permettant de détecter efficacement les répétitions significatives dans un génome complet. Aussi, nous avons développé une méthode efficace de détection de répétitions. L'algorithme, implémenté dans le logiciel *repseek*, est basé sur les heuristiques classiques de recherches de similarité,

notamment sur celle utilisée dans l'algorithme de BLAST 2 (Altschul *et al.*, 1997). Si la structure générale de notre algorithme est très proche, nous avons incorporé les nombreuses spécificités liées au problème de la détection de répétitions.

La méthode se décompose en trois étapes clefs : (1) détection de similarités fortes (pas de substitution, ni d'insertion ou de délétion) (2) exploration de la similarité aux bornes de ces graines à similarité très forte et (3) filtrage des répétitions sur des critères statistiques. Pour la première étape, nous avons utilisé une version de l'algorithme de Karp, Miller et Rosenberg (aka KMR) (Karp *et al.*, 1972; Landraud *et al.*, 1989; Soldano *et al.*, 1995) dédiée aux répétitions et optimisée pour la consommation de mémoire. Pour la seconde étape, nous avons opté pour l'algorithme d'extension implémenté dans BLAST 2, à quelques modifications liées à la structure des répétitions. Pour le filtrage des répétitions non strictes, nous avons utilisé les statistiques développées par Karlin and Altschul (1993) sur les alignements locaux.

En pratique, le logiciel est suffisamment rapide pour détecter toutes les répétitions dans un génome bactérien en un temps très court (moins de 10 min) et en un temps raisonnable pour un chromosome eucaryote (quelques heures). Néanmoins, il faut mentionner que l'espace mémoire nécessaire à la première étape de la méthode est le principal facteur limitant. Cette limitation empêche le logiciel d'analyser des séquences de plus de quelques dizaines de mégabases (sur des ordinateurs de bureau). L'un des développements futurs est justement de modifier l'algorithme de cette étape pour se tourner vers des méthodes consommants *a priori* moins de mémoire (e.g. tableau de suffixes).

Le principal défaut de cette méthode est qu'elle ne traite que les couples de copies. Pour les répétitions ayant plus deux copies, tous les couples possibles sont traités indépendamment. Comme le nombre de couples croit quadratiquement avec le nombre de copies, les familles nombreuses représentent un grand nombre de couples, qui occupent une large partie des résultats. Cette limitation aux couples soulève au moins deux problèmes.

Le premier problème est d'ordre biologique. La similarité de la famille dans son entier n'est pas traitée, seules les paires de séquences similaires sont considérées. Nous ne produisons jamais d'alignements multiples de ces séquences, pourtant essentiels pour (1) estimer la taille de la famille et (2) mener des analyses évolutives phylogénétiques. Le logiciel *repseek* estime néanmoins le nombre de copies que contient la répétition à laquelle le couple considéré appartient. Ce n'est qu'un cache misère, mais cela permet de classer les couples sur la base du nombre de copies de la répétition et, le cas échéant, d'exclure les grandes familles. En ne considérant que les répétitions à deux copies, cette stratégie

m'a permis d'étudier des jeux de données homogènes et représentatifs de l'ensemble des répétitions (et non le fait de quelques très grandes familles).

Le second problème est d'ordre méthodologique. Les familles nombreuses génèrent un grand nombre de couples de répétitions et sont donc très consommatrices de temps de calcul. La complexité quadratique de l'algorithme en fonction du nombre de copies est une réelle limitation, tant au niveau de la mémoire que du temps de calcul. Ainsi, dès que le génome analysé contient des familles immenses, comme par exemple des milliers de copies d'éléments transposables ou des milliers de satellites (répétés sur eux-même et entre eux), l'algorithme n'est pas performant. Pour cette raison, le programme propose aux utilisateurs d'ignorer les régions trop répétées pour éviter des temps de calculs trop longs. Mais encore une fois, cette solution, bien que pratique, n'est pas entièrement satisfaisante.

Je participe, en collaboration avec Todd Treangen, à la mise en place d'un logiciel qui traitera l'ensemble des copies d'une répétition au lieu de les traiter par couples. L'algorithme suivrait les étapes exposées ci-dessus, mais serait étendu à des familles multi-copies. Todd Treangen a, d'ores et déjà, proposé une première solution (Treangen *et al.*, 2009), mais nous pensons que des améliorations notables sont à apporter. Traiter des copies multiples pose deux problèmes principaux. Tout d'abord, il n'existe pas de méthode complètement satisfaisante pour faire de l'alignement local multiple, pourtant nécessaire à la seconde étape de l'algorithme (exploration de la similarité aux bornes des graines). Notre meilleur candidat est, pour le moment, une heuristique basée sur une séquence consensus (Waterman, 1995; Price *et al.*, 2005). Par ailleurs, il n'existe pas de statistiques permettant d'associer une probabilité d'apparition par hasard aux répétitions de plusieurs copies. Comme expliqué ci-dessous, j'ai peut-être une solution à ce second problème.

Statistiques

Inspiré des statistiques développées pour la similarité entre deux séquences, j'ai développé un modèle permettant d'assigner des probabilités à une famille de répétitions. Un alignement est caractérisé par un score d'alignement. Le score d'alignement est la valeur qui est maximisée par les algorithmes d'alignements. Ce score, noté σ, est la somme des similarités (score individuel positif) et dissimilarités (score individuel négatif) de l'intégralité de l'alignement. Ce score augmente donc avec la longueur de l'alignement et avec la densité de similarités au sein de l'alignement. Les scores individuels de chaque paire de symboles sont donnés par une matrice de score (identité, HOXD, PAM, BLOSUM, ...). Le score prend également en compte les trous dans l'alignement, typiquement sous forme d'une

fonction affine : $\sigma_{\text{trou}} = \sigma_{\text{ouv}} + L \times \sigma_{\text{ext}}$, où L est la longueur de la l'alignement et où les scores associés (σ_{ouv} et σ_{ext}) sont négatifs. Ainsi, le score baisse avec l'augmentation du nombre de trous (via σ_{ouv}) et avec l'augmentation de leur longueur (via σ_{ext}).

Lors d'un alignement dit "local", on recherche la meilleure sous-séquence commune aux deux séquences alignées. On cherche à trouver le segment qui présente le meilleur score possible entre les deux séquences données en entrée; on ignore ainsi les parties des séquences sans similarité qui feront baisser le score. Pour le score d'un alignement global, il n'existe pas, à ma connaissance, de cadre probabiliste décrivant la distribution des scores dans un modèle de référence. *A contrario*, il existe une assez grande littérature concernant les probabilités associées aux scores d'alignement locaux (voir, par exemple, Karlin and Altschul (1990); Waterman (1995); Ewens and Grant (2001)).

La probabilité associée à un score pour une position donnée de l'alignement est donnée par une loi pseudo-géométrique :

$$P(\sigma \geq x) = Cp^x \tag{3}$$

où p est la probabilité associée à un score de 1 et C est une constante[3]. Cette équation est également notée sous la forme équivalente : $P(\sigma \geq x) = Ce^{-\lambda x}$, où p est exprimé comme $e^{-\lambda}$.

Une des façons de comprendre cette probabilité est la suivante : lorsque l'on choisit une position au hasard dans chacune des deux séquences entrées, la probabilité d'y trouver un score égal à 1 ou plus est donné par p. Il faut bien comprendre qu'il n'existe pas souvent un score individuel de 1 (aucun couple de symboles n'a un score de 1 dans la matrice de score), cette probabilité p ne correspond donc pas à l'appariement réel de deux symboles donnés. La probabilité de trouver une similarité de score individuel y à des positions choisies est donnée par p^y. Le score x d'un segment résulte de l'agencement bout à bout de symboles ayant des scores individuels (y_i) dont la somme fait x. Ici, on considère que chaque position est indépendante ; on multiplie donc les probabilités et l'on obtient $p^x = \prod_i p^{y_i}$. C'est la probabilité d'avoir un segment de score x à une position donnée sans tenir compte des symboles aux bornes du segment. Dans un tel système, les segments de score x peuvent se chevaucher. Pour éviter tout chevauchement, il faut considérer un segment de score x borné sur l'une des bornes par une dissimilarité. La probabilité qu'un segment de score x débute à une position donnée (ne pouvant pas s'étendre en amont, par exemple) est donné par Cp^x, où C est la probabilité que le segment ne s'étende pas au delà de la borne.

[3]Pour une loi géométrique, on a $C = (1 - p)$.

Eviter les possibles chevauchements rend indépendant chaque couple de positions. Ainsi, pour tous les couples de positions entre deux séquences de longueur m et n, le nombre moyen de segments de score x^+ est donné par:

$$E[N_{\sigma \geq x}] = Cmnp^x \tag{4}$$

C'est la E-value donnée par BLAST. Il s'en suit que la probabilité de trouver au moins un couple est donné par la loi de Poisson (mn tirages de probabilité Cp^x, avec $mn \gg 1$ et $Cp^x \ll 1$):

$$\begin{aligned} P(N_{\sigma \geq x} > 1) &= 1 - e^{-E[N_{x^+}]} \\ &= 1 - e^{-Cmnp^x} \end{aligned} \tag{5}$$

C'est la P-value donnée par BLAST. Dans l'équation 3 et celles qui en découlent (eq. 4 et 5), les valeurs C et p sont toutes deux dépendantes du système de score utilisé. Si les trous sont interdits (*i.e.* $\sigma_{\text{ouv}} = \sigma_{\text{ext}} = -\infty$), alors les valeurs de C et p peuvent être calculées analytiquement (Karlin and Altschul, 1990). Lorsque les trous sont permis, il n'existe pas de formule analytique. Ces valeurs sont donc estimées par simulation (Waterman and Vingron, 1994). Des corrections peuvent être apportées pour tenir compte des effets de bord lorsque les séquences sont "petites" (Altschul and Gish, 1996). On peut également calculer, à partir des ces valeurs p et C, la probabilité associée au n$^{\text{ieme}}$ meilleur score (Karlin and Altschul, 1993; Waterman and Vingron, 1994).

Pour une 2-répétition (une répétition à deux copies) dans le même brin, la même approche peut être appliquée. Cependant, le nombre de position dans une séquence de taille n est donné par $\binom{n}{2} = n(n-1)/2$ (et non plus mn). Lorsque l'on cherche la probabilité d'avoir r copies directes (dans le même brin) ayant un score de x, on a :

$$E[N_{\sigma \geq x}] = C\binom{n}{r}p^{(r-1)x} \tag{6}$$

Le problème se complique un peu lorsqu'il faut considérer que les r copies peuvent être réparties dans les deux brins. Ainsi, il faut envisager tous les cas : de r à 0 copies dans le brin direct. Notons, tout de suite, que par effet de complémentarité entre les brins, lorsque l'on a $r - k$ copies dans le brin direct et k dans le brin inversé, il existe une répétition miroir de séquences complémentaires inversées avec $r - k$ copies dans le brin inversé et k

dans le brin direct (voir figure 8).

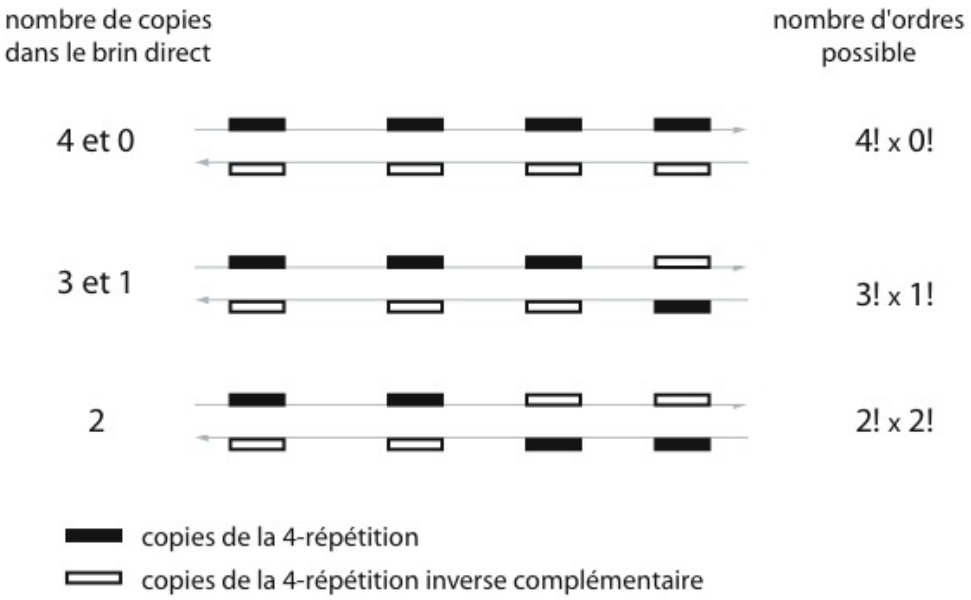


Figure 8: **Différents agencements pour une 4-répétition.** Les copies noires et blanches sont deux facettes de la même répétition : chaque famille de séquences est associée par effet de miroir à une famille de séquences complémentaire inversées. Les répétitions peuvent donc être comptées plusieurs fois. Pour une 4-répétition (une répétition de 4 copies), les copies peuvent être dans le même brin (schéma du haut), réparties en 3 dans un brin et 1 dans l'autre (schéma médian) ou bien 2 copies dans chaque brin (schéma du bas). Compter les 4-répétitions ayant les 4 copies dans un brin revient à compter celles qui ont les 4 copies dans l'autre brin (schéma du haut). Lorsque l'on compte toutes les répétitions qui ont 2 copies dans le brin direct, on compte deux fois toutes les répétitions.

Si l'on admet qu'aucune copie ne peut commencer à la même position (quel que soit l'orientation), le nombre de positionnements possibles pour r copies est $n \times (n-1) \times (n-2) \times \cdots \times (n-r+1)$, soit $n!/(n-r)!$, ce qui s'approxime pour $n \ll r$ à n^r. Pour chaque configuration, on peut échanger les copies sur un même brin sans changer la répétition ; ainsi, pour k copies sur le brin direct et $r-k$ copies sur le brin inversé, il existe $k!(r-k)!$ agencements équivalents. Enfin, il faut tenir compte de l'effet de miroir présenté ci-dessus et ne compter qu'une seule des deux répétitions qui existent en miroir l'une de l'autre. On a donc:

$$E[N_{\sigma \geq x}] \approx C \times \Big(\sum_{k=0}^{r} \frac{1}{2} \frac{n^r}{k!(r-k)!} \Big) \times p^{(r-1)x} \tag{7}$$

Dans l'équation présentée ci-dessus, seules les valeurs de p et C sont inconnues. J'ai confronté les prédictions de l'équation 7 à des répétitions obtenues dans des séquences aléatoires. Il n'existe aucun logiciel permettant de détecter toutes les répétitions multi-copies avec trou. Les meilleurs candidats (Price *et al.*, 2005; Treangen *et al.*, 2009) ne garantissent pas que la répétition trouvée soit maximale (pour un r donné). Par contre, on peut rechercher toutes les r-répétitions strictes (pas de différence ni de trou). Pour ces répétitions, j'ai considéré le système de score donné par la matrice HOXD (Chiaromonte *et al.*, 2002) pour les identités ($\sigma_{ii} = \text{hoxd}(i, i)$). Interdire les différences et les trous revient à considérer le cas où les score négatifs valent $\sigma_{ij} = \sigma_{\text{ouv}} = \sigma_{\text{ext}} = -\infty$. En estimant les paramètres C et p à partir des répétitions à deux copies ($r = 2$), je peux prédire la fonction de répartition, $P(\sigma < x)$, pour les répétitions d'autres multiplicités ($r > 2$). Comme illustré sur la figure 9, l'adéquation entre prédiction et observation est très bonne.

Logiciels et librairies

J'ai participé au développement de plusieurs logiciels et plusieurs librairies de fonctions dédiées aux problématiques biologiques. Le seul logiciel publié à ce jour auquel j'ai activement participé est *repseek*. Ecrit en C, ce logiciel est une implémentation de l'algorithme de détection des couples de répétitions présenté ci-dessus. Faisant suite à ce logiciel, je participe actuellement à la mise en place d'une nouvelle version du logiciel *repeatoire* developpé par Todd Treangen qui détectera toutes les familles de répétitions non-strictes en s'appuyant sur l'algorithme utilisé dans *repseek* mais généralisé aux cas des répétitions à plus de deux copies. Je projette de terminer ce projet en automne avant mon départ pour le Japon.

Je développe, en collaboration avec Sophie Brouillet, une librairie écrite en C d'analyse de données de séquences issues d'alignements multiples, notamment de données de génétique des populations. Deux versions en ligne d'exécutables sont disponibles à partir de pages web. Le premier est une implémentation d'un test de panmixie[4] et le second est une implémentation de plusieurs tests de neutralité[5].

Enfin, j'ai implémenté une librairie pour générer des simulateurs d'arbres généalogiques en C++. Cette libraire permet de générer des arbres pour différents scénarios évolutifs (modèle de référence, croissance exponentielle, balayage sélectif, isolation avec migration, etc.). La librairie est suffisamment générique pour m'avoir servi dans mes travaux sur

[4]http://wwwabi.snv.jussieu.fr/achaz/hudsontest.html
[5]http://wwwabi.snv.jussieu.fr/achaz/neutralitytest.html

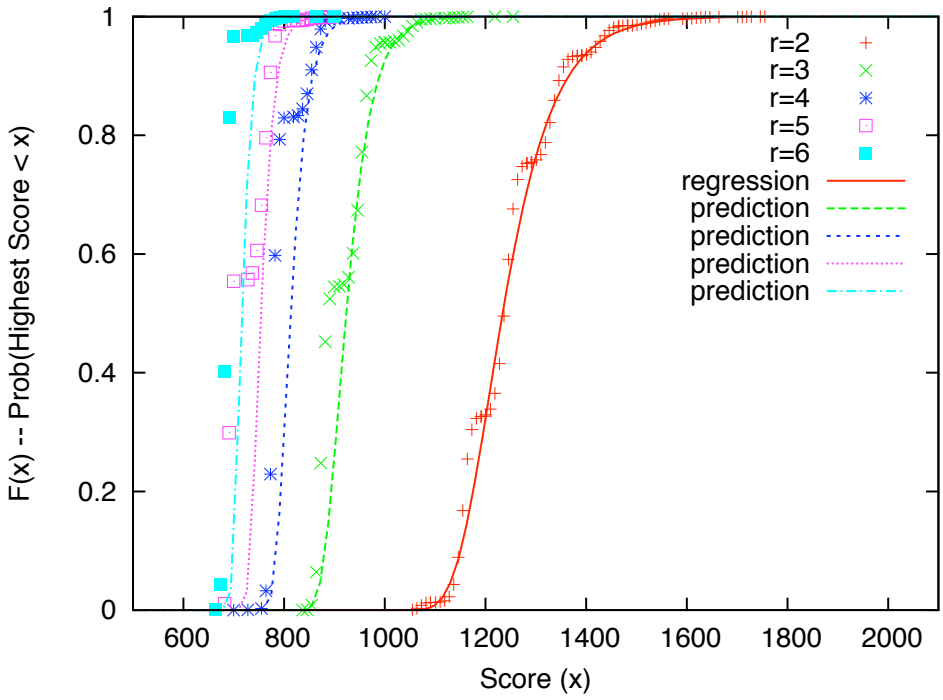


Figure 9: **Fonction de répartition pour les répétitions multi-copies.** Fonctions de répartitions des r-répétitions strictes (sans différence ni trou) ayant le plus haut score dans une séquence de longueur 10 kb et de composition $\{0.25, 0.25, 0.25, 0.25\}$. Le score est calculé selon le système suivant: $\sigma_{ii} = \text{hoxd}(i, i)$, $\sigma_{ij} = \sigma_{\text{ouv}} = \sigma_{\text{ext}} = -\infty$. La fonction de répartition empirique, calculée à partir de 10^4 séquences aléatoires, est donnée par les points pour $r \in [2, 6]$. Les paramètres C et p sont estimés à partir des données aléatoires pour $r = 2$: $C \approx 0.350$ et $p \approx 0.986$ (une identité ayant un score moyen de 96, on a $p^{96} = 0.25$). Les courbes pour $r \in [3, 6]$ sont prédites à partir de ces valeurs C et p.

les tests de neutralité mais également dans ceux sur les arbres de gènes dans des arbres d'espèces (*vide infra*).

Je ne pense pas distribuer ces librairies à court terme. Elles forment le coeur de mes outils d'analyse de données ou de simulation qui me permettent de travailler quotidiennement. Implémenter des méthodes m'a souvent permis de comprendre leurs détails techniques sous-jacents, bien souvent indispensables à la bonne maîtrise de ces méthodes. Il est malheureusement impossible, j'en conviens, de ré-implémenter tous les outils dont nous avons la nécessité par manque de temps. Néanmoins, la décomposition de la logique fine des méthodes est une des étapes clefs de la compréhension de l'ensemble d'une problématique scientifique.

Aujourd'hui

Ayant eut la chance de pouvoir explorer différents champs de biologie évolutive, je cherche dans mes projets de recherche actuels à établir des ponts entre les différentes approches. J'ai choisi de détailler trois projets qui s'inscrivent dans la continuité de mes travaux passés. Ces projets sont tous trois le fruit d'une collaboration scientifique. Le premier projet constitue le second volet de la thèse d'Etienne Loire dont je suis le principal encadrant. Le second concerne une seconde analyse des données temporelles de HIV-1 ; il est réalisé en collaboration avec Sophie Brouillet (Atelier de Bioinformatique, UPMC) et avec l'équipe du Pr John Coffin (Drug Resistance Program, USA). Enfin le troisième projet s'attache à l'étude des généalogies de gènes au sein de généalogies d'espèces et s'effectue en collaboration avec Nicolas Puillandre (Université d'Utah) et avec Amaury Lambert (Laboratoire des Probabilités et Modèles Aléatoires, UPMC).

De la génomique à la génétique des populations

Dans la première partie de sa thèse, Etienne Loire a réalisé un catalogue des microsatellites codants humains. Il a montré que ces séquences hypermutables étaient toutes soumises à une sélection négative dont la force est variable d'un groupe de gène à un autre. Afin d'appréhender mieux l'histoire évolutive de ces séquences hypermutables, nous avons étudié la dynamique d'apparition et de disparition des microsatellites dans les gènes orthologues de l'homme, du chimpanzé, de l'orang-outang et du macaque.

En premier lieu, nous avons montré que les substitutions s'accumulent deux à trois fois plus vite dans les loci microsatellites codants que dans le reste des gènes. Ceci semble vrai dans toutes les lignées considérées. Cette observation est réminiscente d'observations faites sur des régions régulatrices de gènes. En effet, il a été montré que, pour les régions de 2 kb en amont des gènes humains, les régions répétées évoluaient plus vite que les régions non répétées (Shankar *et al.*, 2007). Ceci suggère, qu'il est possible que cette accélération

de divergence soit lié à la nature des microsatellites plutôt qu'un effet lié à la sélection s'exerçant sur eux. Ceci demande cependant à être confirmé.

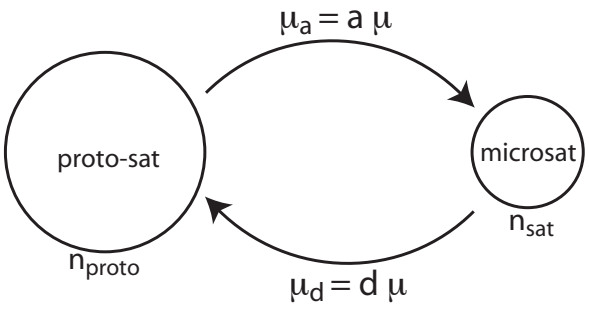
Nous avons également pu montrer que le nombre de microsatellites codants est constant le long des lignées évolutives, ce qui suggère qu'ils ont atteint un certain équilibre. Cet équilibre est pourtant dynamique puisque de nombreux loci apparaissent et disparaissent le long des lignées évolutives. Mathématiquement, cet équilibre se traduit par :

$$\mu_a \times n_{\text{proto-sat}} = \mu_d \times n_{\text{sat}}$$

avec μ_a le taux d'apparition de microsatellites codants, μ_d le taux de disparition, $n_{\text{proto-sat}}$ le nombre de loci qui sont susceptibles de se transformer en satellites et n_{sat} le nombre de microsatellites. Cet équilibre est illustré sur la figure 10.

Afin de caractériser finement la dynamique d'apparition et de disparition de ces microsatellites codants, nous nous sommes concentrés sur les microsatellites ayant une longueur de 8 exactement (e.g. $TAAAAAAAAG$). Nous avons également compté le nombre de proto-satellites, séquences non-satellites pouvant se transformer en microsatellites de taille 8 par une unique substitution (e.g. $TAAAACAAAG$). Pour transformer un proto-satellite en microsatellite, il n'existe qu'une une seule substitution possible. A l'inverse, il existe 24 substitutions transformant un microsatellite en proto-satellite. Ainsi, à l'équilibre, sans autre force évolutive, on attend 24 fois plus de proto-satellites que de satellites (si toutes les mutations se produisent avec une probabilité égale). Or, on dénombre 35 fois plus de proto-satellites que de microsatellites. Il ne semble pas que cette différence soit explicable par des substitutions non symétriques (les corrections prenant en compte les différences de composition ne changent pas l'observation). Il faut donc envisager que l'équilibre est déplacé légèrement par la sélection naturelle qui s'excerce à l'encontre des microsatellites codants.

Nous avons donc entrepris d'estimer le coût sélectif moyen associé à un microsatellite codant de taille 8. Pour cela, nous proposons, en première approche, le modèle décrit sur la figure 10 qui décrit les fréquences attendues de proto-satellites et de microsatellites dans un modèle d'équilibre mutation-sélection dans une population diploïde. Dans ce modèle, les microsatellites de taille 8 apparaissent avec un taux μ_a et disparaissent avec un taux μ_d. Ces deux taux sont rapportés au taux de substitution moyen du génome humain estimé à environ 10^{-8} par génération (Drake *et al.*, 1998), ici noté μ. Ainsi, nous avons $\mu_a = a\mu$ et $\mu_d = d\mu$. Si μ représente la probabilité d'être muté pour un site donné à une génération, alors $a = 1/3$ (une seule mutation parmi les 3 possibles) et $d = 8$ (8 sites peuvent être



phénotype	sat/sat	proto/sat	proto/proto
fréquences	p^2	$2p(1-p)$	$(1-p)^2$
fitness	1-s	1-hs	1

A l'équilibre:

$$s \sim \mu \{ p(a+d) - a \} / \{ p(p-1)[p - h(2p-1)] \}$$

Figure 10: **Equilibre mutation-sélection entre un microsatellite et les séquences proto-satellites**. Ces dernières sont celles qui peuvent se muer en microsatellites d'une taille donnée en une seule substitution. A l'équilibre, les changement de fréquences sont nuls et s peut être calculée à partir de p, μ, a, d et h. Pour $p = 1/36$, $\mu = 10^{-8}$, $a = 1/3$, $d = 8$ et $h = 0$, on a $s \approx 10^{-6}$ (lorsque $h = 1$, on obtient $s \approx 4.10^{-8}$).

affectés). s est le coût associé à un homozygote pour l'allèle satellite. h est indicatif de la dominance de l'allèle satellite: $h = 0$ signifie que le microsatellite codant est récessif, $h = 1$ qu'il est dominant.

Nous n'avons aucune donnée sur les fréquences des allèles satellites et proto-satellite pour un locus donné dans la population. Cependant, nous avons une très grande quantité de loci satellites et proto-satellites dans les génomes de plusieurs primates. Si nous cherchons l'effet moyen d'un satellite de taille 8, on peut imaginer, en première approximation, que pour un génome donné, chaque locus est un réplicat indépendant d'un tirage de Bernouilli ou p est la fréquence de l'allèle satellite dans la population. Ainsi, on pourra estimer la fréquence moyenne de l'allèle satellite dans la population en utilisant la fréquence des satellites au sein d'un seul génome : $\hat{p} = n_{\text{sat}}/(n_{\text{sat}} + n_{\text{proto}})$, soit $\hat{p} = 1/36$.

En utilisant les valeurs données ci-dessus, nous obtenons pour $h = 0$: $s \approx 10^{-6}$ et pour $h = 1$: $s \approx 4.10^{-8}$. Rapporté à la taille efficace de la population humaine (environ 2.10^4),

le coefficient de sélection devient $N_e s \approx 2.10^{-2}$ pour le cas récessif. En d'autres termes, un microsatellite codant de taille 8 pris séparément est assimilable à une séquence neutre ($N_e s \ll 1$).

Il faut donc envisager que c'est le nombre important de ces satellites (plus d'un millier dans le génome humain) qui accroît leur effet délétère. C'est leur présence massive et non individuelle qui est sélectionnée négativement.

Dans ce projet, nous sommes partis de données génomiques et phylogénétiques et nous avons tenté un passage vers la génétique des populations. Ce passage a pu être fait grâce à la présence répétée d'un présumé couple d'allèles. En effet, c'est en considérant que chaque copie de la répétition (ici le couple satellite/proto-satellite) est un réplicat indépendant du même phénomène (l'équilibre mutation-sélection) que nous avons estimé la fréquence des deux allèles au sein de chaque locus à partir d'un seul génome. Cette stratégie peut potentiellement s'appliquer dès que l'on a plusieurs réplicats indépendants d'un scénario évolutif. Toute forme de répétitions est, par exemple, candidate à ce genre d'approche.

De la population à la séquence

Le projet exposé ci-dessous est la suite directe de mon travail sur le *turn-over* des populations de HIV-1. Comme nous l'avons vu précédemment, le *turn-over* génétique des populations de HIV-1 au sein d'un patient correspond à celui d'une population assimilable au modèle de référence pourvu que l'on assimile la population virale à quelques milliers de génomes haploïdes (c'est la taille efficace). Ce résultat a été établi pour des populations infectant chroniquement des patients. Pour ces patients, la virémie est stable et n'augmente pas au delà d'une certaine limite. Ces individus infectés contiennent naturellement le virus et ne développent pas les symptômes d'immuno-déficience.

Cependant, cette situation reste exceptionnelle. Pour la plupart des patients infectés, la population virale croit rapidement, déborde le système immunitaire et entraîne la mort du patient. En conséquence, lorsque la virémie dépasse un seuil critique (environ 10^6 virions/ml), le patient se voit administrer un traitement anti-viral. En un mois, la virémie chute à moins de 50 virions/ml, preuve de l'efficacité du traitement. L'action des drogues est d'empêcher le virus de commencer un nouveau cycle réplicatif. On considérera ici, en première approche, que lorsque le traitement débute, le virus ne fait plus de nouvelles générations. Or, la durée de vie du virus sous forme libre est très courte. La virémie mesure donc le nombre de virus sortant juste des cellules infectées. La chute de la virémie

correspond à la diminution du nombre de virus relargués par les cellules infectées. *Grosso-modo*, toutes les cellules ont été infectées avant le début du traitement.

Grâce à des échantillons de la population virale prélevés avant le traitement et lors de la chute de la virémie, nous avons pu étudier les effets à court terme du traitement sur la population virale. Une première analyse montre que, malgré la baisse importante de la virémie, la diversité génétique de la population, reste constante. Ceci semble en accord avec notre intuition sur les effets du traitement. En effet, comme la diversité observée correspond aux virus ayant infecté les cellules avant le début du traitement, elle est celle de la population originale ; elle ne doit donc *a priori* pas diminuer. Notons que la diversité observée (environ 1% de différence entre deux séquences) est comparable aux observations faites sur la diversité génétique de HIV-1 au sein de patients infectés chroniquement (Leigh Brown, 1997; Seo *et al.*, 2002; Shriner *et al.*, 2004; Achaz *et al.*, 2004; Shriner *et al.*, 2006).

Nous avons également mesuré le *turn-over* de la population virale à partir du même test de structure utilisé précédemment. Rappelons que ce test (Hudson *et al.*, 1992) mesure la probabilité que deux échantillons soient homogènes. Les résultats (figure 11) montrent que lorsque des échantillons prélevés avant le traitement sont comparés à des échantillons prélevés après le traitement, ils apparaissent très souvent comme différents. Ceci signifie que la composition de la population virale a changé à la suite du traitement.

Nous avons cherché à savoir quels sites polymorphes étaient à l'origine de ce changement de composition. Dans, ce but, nous avons caractérisé l'effet individuel de chacun des 283 sites polymorphes en effectuant le test de structure pour chaque site pris individuellement[6]. En étudiant l'effet de chacun des sites pris indépendamment, nous avons pu montrer que seule une poignée de sites est responsable du changement de composition de la population. En effet, en ne considérant que les 5 sites ayant le plus fort effet, nous reproduisons un résultat tout à fait comparable à celui obtenu pour l'ensemble des 283 sites polymorphes. La différence de composition est la conséquence de changement de fréquence (en général une disparition d'un des deux variants) pour seulement 5 sites.

Nous avons également voulu savoir si certains types de séquences apparaissaient ou disparaissaient après le traitement. Pour cela, nous avons construit un arbre phénétique sur la base des distances génétiques entre individus par neighbor-joining. L'arbre obtenu (figure 12) montre qu'un groupe de virus (entouré sur la figure) bien représenté dans les échantillons pré-traitement (séquences notées a et b) est quasi-absent des échantillons post-

[6]Il est possible de calculer analytiquement les probabilités du test lorsqu'un seul site est considéré. Ceci rend le calcul pour chaque site instantané et s'affranchit des problèmes lié à l'échantillonnage.

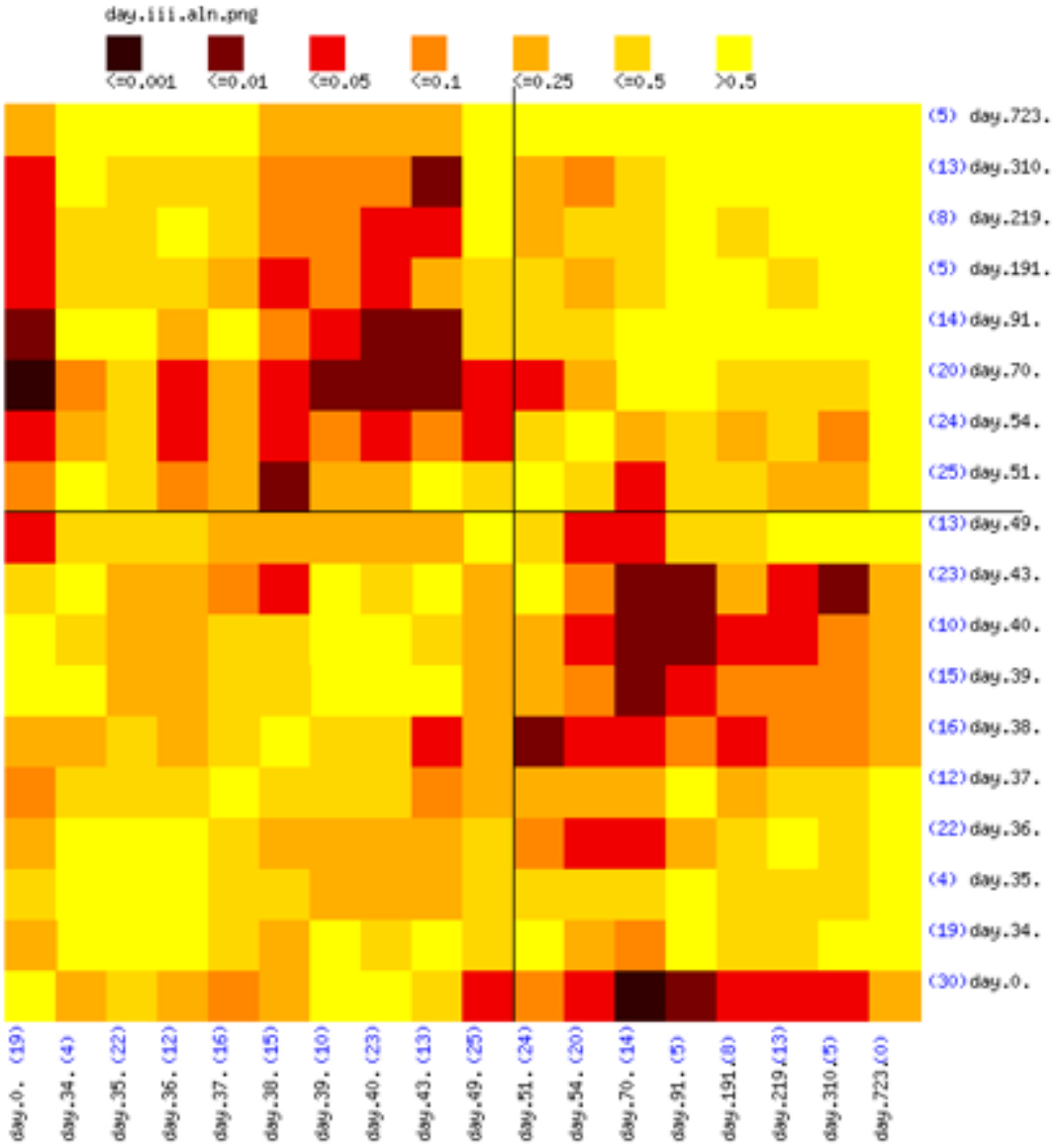


Figure 11: **Différence de composition de la population virale avant et après traitement.** Chaque case représente le résultat d'un test de structure appliqué à deux échantillons. Est reporté la probabilité que les deux échantillons proviennent d'une même population. Le traitement a débuté au jour 49, la ligne noire indique donc la frontière entre les échantillons avant et après traitement. En bleu et entre parenthèse sont donnés les effectifs de chaque échantillon.

traitement (séquences notées c et d).

Rappelons que les séquences des virus circulant dans le plasma sont un mélange des

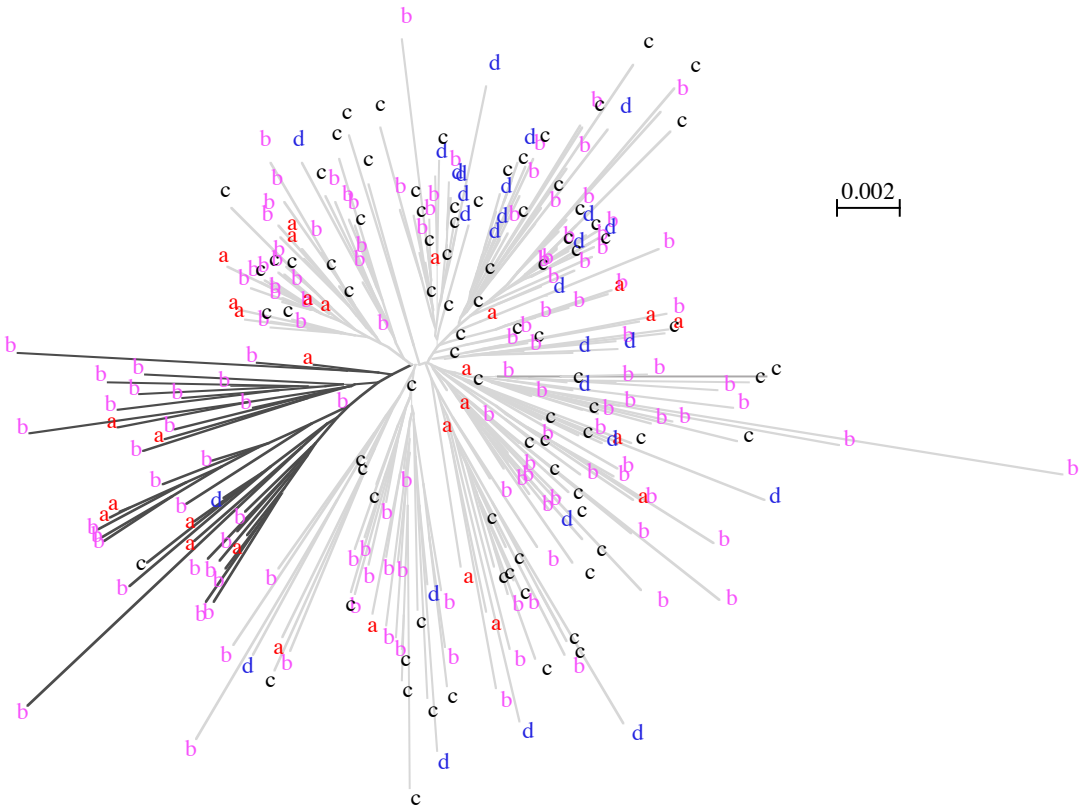


Figure 12: **Arbre phénétique de toutes les séquences.** Arbre reconstruit par la méthode *neighbor-joining* en utilisant les distances évolutives calulées sur un modèle F81, qui prend en compte la composition des séquences. Les séquences marquées *a* (rouge) et *b* (rose) sont issues d'échantillons avant le traitement et celles marquées *c* (noir) et *d* (bleu) proviennent d'échantillons post-traitement.

différents types de virus : ceux infectant les différents types cellulaires et les différents types tissulaires. Il a pu être clairement montré que les virus infectant les différents types tissulaires (Frost *et al.*, 2001) ou les différents types cellulaires (Potter *et al.*, 2004) présentent une structuration génétique. Notre interprétation de ces résultats est qu'un de ces types de virus serait absent des échantillons post-traitement.

Au moins deux interprétations sont possibles. En premier lieu, l'absence de ce type de virus pourrait s'expliquer par un effet différentiel de la drogue sur les différents types de virus (voire différents types de cellules infectées). Cependant, une autre explication, plus simple, est que la demi-vie des différents types de cellules infectées n'est pas identique. Certains types de cellules infectées meurent vite alors que d'autres vivent plus longtemps.

Dès le début du traitement, il n'y a plus de nouvelle infection. Toutes les cellules infectées dont le temps de demi-vie est court ne vont plus relarguer de virus après le début du traitement. Ce que nous voyons ici pourrait être la simple combinaison des différents temps de demi-vie des cellules infectées et de leur structuration génétique sous-jacente.

Un aspect intéressant de ce projet est qu'à partir d'une observation de génétique des populations (la population semble avoir changé rapidement après le traitement), nous sommes allé chercher les causes "génomiques" de notre observation (quelles séquences, quels sites). Ces observations nous ont permis d'entrevoir une explication populationnelle qui rendrait compte de l'observation.

Des arbres dans des arbres

Les arbres phylogénétiques représentent l'histoire évolutive des entités biologiques situées aux feuille de l'arbre (taxons, gènes, individus, etc). Ces arbres sont tous une vision rétrospective d'un processus de branchement qui se déroule dans le sens prospectif. Si plusieurs processus de branchement sont emboîtés les uns dans les autres, les arbres correspondant à ces différents processus seront également emboîtés les uns dans les autres. C'est le cas des populations de génomes qui évoluent au sein d'espèces. Rétrospectivement, les arbres de gènes sont emboîtés dans des arbres d'espèces.

L'approche *barcoding* (voir, par exemple, Frezal and Leblois (2008)) vise à systématiser les données de séquences disponibles pour toutes les espèces de la biosphère. Grâce à ce projet, nous disposons de séquences homologues (en général une partie du gène mitochondrial cytochrome c oxydase) pour un grand nombre d'individus, certains appartenant à la même espèce, d'autres appartenant à des espèces différentes. Un des intérêts de ce jeu de données est qu'il permet potentiellement de délimiter les espèces dans un clade donné sur la seule base des séquences génétiques. La distribution des distances entre deux séquences est typiquement bi-modale : les séquences appartenant à la même espèce présentent une faible distance et les séquences appartenant à des espèces différentes une plus grande distance. Ce trou dans la distribution des distances génétiques est connu sous le nom de *barcod gap*. J'ai mis au point une méthode permettant de détecter rapidement et automatiquement ce trou à partir de l'ensemble des distances deux à deux. En bref, les valeurs sont rangées par ordre croissant. Lorsque l'on est au voisinage du *barcod gap* les distances augmentent brusquement (voir figure 13). Cette augmentation se traduit par un pic dans la dérivée de la distance par le rang. La détection automatique du sommet de ce pic permet de trouver

la distance qui délimite les espèces entre elles.

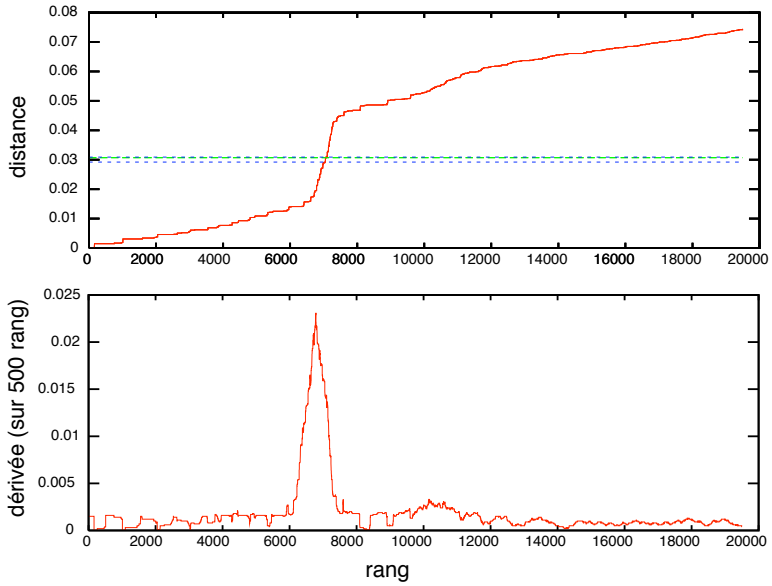


Figure 13: **Exemple de *barcod gap***. Cet exemple contient un large échantillon (un millier de séquences) d'organismes appartenant à la famille des *Turinae*. Les distances génétiques entre toutes les paires de séquences de *Turinae* ont été calculées. Dans la figure du haut, les distances sont rangées par ordre croissant. Dans la figure du bas est donnée la dérivée de la distance par le rang. La dérivée reportée sur cette figure est calculée par $\frac{d_{i+500}-d_i}{500}$.

Afin d'éprouver les performances et les limites de la méthode, nous avons décidé de simuler des données de type barcoding. Ces simulations nous permettront également de tester d'autres méthodes de délimitations d'espèces comme la méthode basée sur un modèle mélangeant explicitement spéciation et coalescent (*General Mixed Yule-Coalescent* (Pons *et al.*, 2006)). Le principe des simulations est illustré sur la figure 14 : (1) Simulation d'une phylogénie d'espèces, (2) reconstruction d'une généalogie de gène contrainte par l'arbre des espèces et (3) ajout de mutations dans l'arbre.

Pour la troisième étape, nous avons choisi un processus de Poisson dont la moyenne est donné par μG, où μ est le taux de mutation du locus considéré et G, la somme de toutes les branches de la généalogie. Les mutations sont distribuées uniformément dans la généalogie. Pour la seconde étape, nous avons opté, pour le moment, pour un processus de coalescence standard mais contraint par les espèces, assimilées à des populations totalement séparées (Simonsen *et al.*, 1995; Wakeley, 2009). Le choix du modèle de spéciation n'est pas une chose aisée. Il m'est apparu par le biais de discussions informelles qu'il n'existe pas de

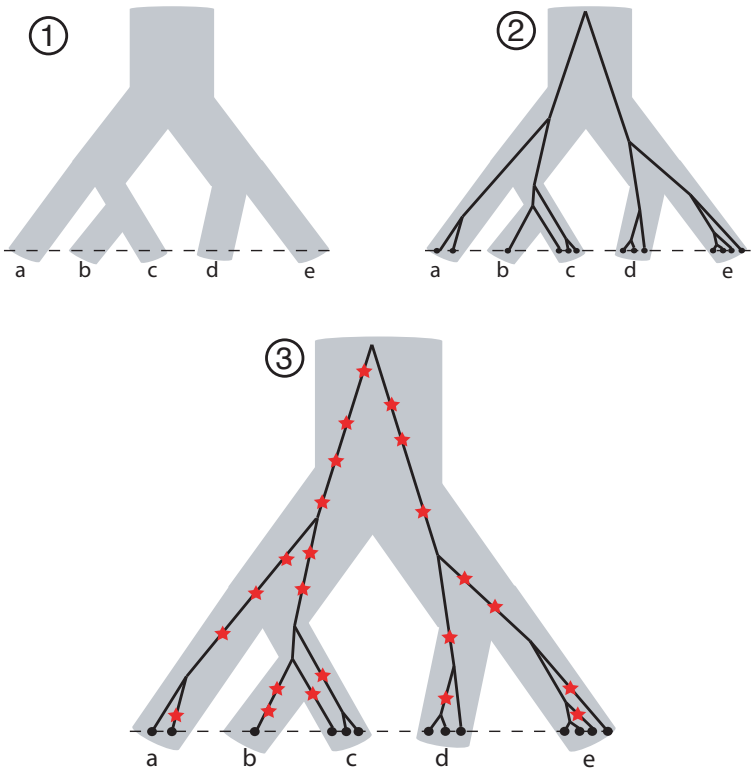


Figure 14: **Génalogie de gènes dans une phylogénie d'espèces**. En gris, la phylogénie des 5 espèces considérées. L'échantillon contient 13 gènes homologues dont la généalogie est contrainte par la phylogénie des espèces. Pour les espèces b et c, la généalogie des gènes ne suit pas celle des espèces, ce qui est possible si les temps de coalescence sont plus grands que les temps de spéciations. Ces phénomènes sont connus sous le nom de tri incomplet des lignées. Une fois la généalogie établie, des mutations (représentées par des étoiles rouges) sont ajoutées selon un processus de Poisson.

modèle de référence clairement identifié pour la spéciation. En génétique des populations, le modèle de référence est celui d'un équilibre mutation-dérive dans des populations de taille constante. Dans le domaine de la spéciation, il n'existe pas de consensus (sauf, peut-être une tendance à utiliser un modèle de Yule). Nous avons donc choisi de modéliser la spéciation par un processus de branchement généraliste. Ces processus sont défini par un taux de mort, τ_{d}, et un taux de naissance, τ_{b}. Nous avons, en premier lieu, exploré deux cas limites: le processus de Yule ($\tau_{\mathrm{d}} = 0$) et le processus critique ($\tau_{\mathrm{b}} = \tau_{\mathrm{d}}$). Les premiers résultats montrent que le choix exact du modèle ne modifie pas les conclusions de l'étude sur la fiabilité de la délimitation d'espèces. Grâce à des études antérieures (Lambert, 2007, 2008), on connaît certaines facettes des lois de distribution des temps de coalescence.

On peut donc très facilement simuler de tels processus.

Pour chaque simulation, la méthode de délimitation d'espèces basée sur la distribution des différences est appliquée. La performance de la méthode est évaluée par le nombre d'espèces parfaitement identifiées. Une espèce est parfaitement identifiée lorsque: (a) toutes les distances intra-espèces sont inférieures à la limite que propose la méthode et (b) toutes les distances des séquences de cette espèce avec toutes les séquences des autres espèces est supérieure à la limite.

Les premiers résultats des simulations montrent que la méthode de délimitation d'espèces fonctionne bien lorsque le taux de spéciation (*i.e.* τ_b) est 10 à 100 fois plus petit que les taux de coalescence intra-espèce. Dans ce cas, les noeuds de la phylogénie des espèces sont très anciens et la généalogie des gènes inter-espèces suit celle des espèces. A contrario, lorsque les taux des deux processus (spéciation et coalescence) sont proches, la généalogie des gènes est peu contrainte par celle des espèces. Des événements de tri incomplet des lignées se produisent (cas des espèces b et c dans la figure 14) et empêchent de faire la différence entre espèces sur la base de leur ressemblance génétique.

Demain

Pour l'année scolaire 2009-2010, je bénéficie d'une délégation au CNRS. Je vais mettre à profit cette disponibilité de temps pour partir, en famille, neuf mois à Kyoto. Ce séjour va me permettre de démarrer pleinement le sujet de recherche décrit ci-dessous. Je travaillerais dans le laboratoire d'écologie théorique d'Atsushi Yamauchi (Université de Kyoto) en étroite collaboration avec Fumio Tajima (Université de Tokyo), éminent spécialiste de la génétique des populations. Il est très probable que je bénéficie d'un financement du gouvernement japonais (JSPS) pour soutenir financièrement ce séjour scientifique. Ce voyage sera une aventure scientifique de qualité, mais se révélera, avant tout, je l'espère, une expérience humaine.

La longue marche adaptative d'un génome

La théorie de l'adaptation moléculaire prend ses racines aux fondements de la génétique des populations. Wright (1932) et Fisher (1930) proposent dès les années 30, de représenter l'espace d'évolution comme un paysage adaptatif dans lequel chaque combinaison d'allèle est associée à une valeur de fitness. Les séquences génétiques explorent ce paysage et la sélection naturelle tend à conserver celles qui sont associée à une meilleure fitness. La modélisation informatique des paysages adaptatifs a connu un enthousiasme certain (voir, par exemple, Maynard Smith (1970); Kaufman (1993)). Des approches de modélisation analytique de l'adaptation recentrées sur les contraintes liées aux systèmes vivants ont montré que, typiquement, les variations d'environnement n'induisent qu'un nombre limité de changements. En effet, les séquences génétiques ne sont pas aléatoires et donc souvent proches d'un optimum de fitness, même dans un nouvel environnement (Orr, 2002). En raison de la liaison génétique, l'adaptation d'un locus entraîne, par un effet dit "d'auto-stop", les locus neutres voisins (Maynard Smith and Haigh, 1974). Comme un locus en adaptation acquiert successivement plusieurs mutations bénéfiques, les auto-stops successifs

peuvent mimer de la dérive génétique aux loci avoisinants (Gillespie, 2000a,b). Cette observation essentielle soulève la pertinence de mesurer la force de la dérive génétique à travers les changements de fréquence des allèles.

Classiquement, il est implicitement fait l'hypothèse que ce sont les changements d'environement qui induisent l'adaptation d'un ou de quelques loci. Dans mon projet de recherche, je me propose d'étudier l'adaptation, non pas à l'échelle d'un locus ou de quelques loci, mais à celle d'un génome entier, c'est-à-dire d'un nombre de loci approximativement infini. J'étudierai l'influence des relations, dites épistatiques, que partagent ces loci sur le processus d'adaptation et chercherai à montrer que pour un environnement externe constant, les relations épistatiques induisent un nombre potentiellement infini de changements adaptatifs successifs.

Dès lors que l'on cherche à comprendre l'évolution de plusieurs loci, il est indispensable de prendre en compte les relations qui existent entre eux. L'importance de l'épistasie dans tous les phénomènes biologiques est reconnue depuis fort longtemps (voir la revue de Phillips (2008)). Cependant, l'épistasie a longtemps été mise à l'écart, sans doute par la "faute" du théorème fondamental de Fisher (Orr, 2005). Celui-ci stipule que seule la fitness additive (*i.e.* non épistatique) est perçue par la sélection naturelle (Fisher, 1930). Néanmoins, l'importance des relations épistatiques dans les processus adaptatifs a depuis été soulignée et est aujourd'hui explicitement prise en compte dans les modèles évolutifs (voir, par exemple, le livre de Rice (2004)).

Les relations épistatiques contraignent l'évolution des séquences génétiques. En ce sens, Maynard Smith (1970) propose que l'évolution des séquences génétiques peut être illustrée par un jeu dans lequel on doit changer un mot en un autre en ne changeant qu'une lettre à la fois ; chaque mot intermédiaire doit, bien entendu, avoir un sens. Récemment, Weinreich *et al.* (2006) ont montré comment l'épistasie contraint la succession de 5 mutations adaptatives dans le gène de la β-lactamase impliqué dans la résistance bactérienne à un antibiotique. Parmi tous les ordres possibles dans lesquels peuvent s'effectuer la succession de ces 5 mutations, seul un très petit nombre est compatible avec une augmentation croissante de fitness. De nombreuses études de modélisation par simulation ont étudié l'influence de l'épistasie sur l'évolution d'un nombre fini de loci en interaction (Kaufman, 1993; Yukilevich *et al.*, 2008).

Dans cette étude, je ne chercherai pas à comprendre comment l'épistasie contraint le paysage adaptatif (en ne laissant que quelques voies d'évolution ouvertes) mais au contraire comment elle peut devenir le moteur d'une adaptation constante lorsque le nombre de loci

tend vers l'infini. J'étudierai notamment, l'impact de la nature du réseau de relations entre les différents loci en interaction sur le temps d'adaptation d'un génome entier, dans un environnement externe constant.

Cette partie théorique se fera lors de mon séjour au Japon, en collaboration avec Atsushi Yamauchi et Fumio Tajima. La collaboration avec un laboratoire d'écologie me semble être une bonne occasion de m'initier aux modèles d'interactions si couramment employés en écologie. Ici, l'unité en interaction n'est pas un organisme mais bien un site dans un génome. Par ailleurs, la grande culture de Fumio Tajima en évolution moléculaire me semble un très bon atout dans ce projet. Ce dernier s'est d'ailleurs penché sur la modélisation de l'épistasie mais jusqu'ici confinée à deux loci (Takahasi and Tajima, 2005). Dans une seconde partie, je vais confronter les prédictions théoriques élaborées lors de mon séjour au Japon aux données biologiques réelles. Pour cela, je vais, dès mon retour, travailler avec Eduardo Rocha, collaborateur de longue date, pour mettre en place l'analyse des divergences observées dans trois groupes fonctionnels bactériens pour lesquels de nombreuses données de séquence sont disponibles. Julien Dutheil, auteur d'une méthode de détection de co-évolution (Dutheil *et al.*, 2005; Dutheil and Galtier, 2007), participera également à cette seconde partie.

J'étudierai les divergences qui s'accumulent sur les lignées évolutives des trois groupes fonctionnels: le ribosome, l'opéron lactose et les plasmides conjugatifs. Chacun de ces systèmes est composé d'un ensemble de gènes partageant entre eux un important nombre de relations épistatiques liées à leurs structures, à leurs interactions et à leurs fonctions. De plus, comme pour la plupart des gènes, des sites appartenant au même gène sont vraisemblablement soumis à une forte dépendance épistatique. Ces trois systèmes modèles sont donc emboîtés dans plusieurs niveaux de dépendance épistatiques : relations au sein du même gène, relations au sein du même groupe fonctionnel et relations avec le reste du génome. Je m'appuierai sur les données massives issues des projets de séquençage de génomes bactériens pour reconstruire l'enchaînement des mutations qui se sont succédées.

L'utilisation de génomes bactériens me permettra de choisir l'échelle de divergence souhaitée en fonction des divergences observées dans ces trois systèmes. En effet, sont disponibles aussi bien des collections de génomes appartenant à la même espèce que des génomes très divergents. Ainsi, j'aurai accès à différentes échelles de temps évolutifs et sélectionnerai la plus appropriée. Dans un premier temps, je débuterai sur des échelles de temps courtes pendant lesquelles la structure entière ne devrait que très peu être modifiée. En partant d'un état ancestral inféré, je pourrai reconstruire l'enchaînement des mutations

qui ont fait évoluer ces groupes fonctionnels dans les différentes lignées. Je tenterai de mettre en perspective cet enchaînement avec les informations ayant trait à la structure, la fonction et les interactions connues au sein de ces groupes.

Le premier système est le complexe ribosomique. Chez *Escherichia coli*, il est composé de 55 protéines et 3 sous-unités ARN. La structure tridimensionnelle de la totalité de ce complexe est connue (Schuwirth *et al.*, 2005; Wimberly *et al.*, 2000; Yusupov *et al.*, 2001) et de nombreuses données fonctionnelles sont disponibles (voir, par exemple, Carter *et al.* (2000)). Ceci ouvre la possibilité de cartographier tous les changements d'acides aminés ordonnés temporellement sur la structure du complexe ribosomique, afin d'en extraire une certaine cohérence. Il est cependant possible que le ribosome n'évolue que très lentement et donc qu'il n'y ait pas assez de divergence dans les souches de *E. coli* pour reconstruire une histoire pertinente sur son adaptation récente. Dans ce cas, nous pourrons utiliser les informations de séquences des très nombreuses entérobactéries.

Le second système est l'opéron lactose (Jacob and Monod, 1961), pour lequel les données moléculaires fonctionnelles de la littérature sont très abondantes (voir, par exemple, Abramson *et al.* (2003); Jacobson *et al.* (1994); Lewis *et al.* (1996)). Il faut souligner que l'opéron lactose est absent des espèces voisines de *E. coli*, ce qui suggère qu'il a été acquis par transfert horizontal, relativement récemment dans le génome de cette espèce (Ochman *et al.*, 2000). Ces gènes sont donc potentiellement encore en cours d'adaptation et constituent donc d'excellents candidats pour étudier la dynamique d'adaptation.

Enfin, j'analyserai les mutations observées dans les gènes des plasmides conjugatifs (Fernandez-Lopez *et al.*, 2006). Ces derniers sont des entités génétiques quasi-autonomes, capables de se répliquer dans un large spectre d'hôte (Amabile-Cuevas and Chicurel, 1992). A l'instar d'un parasite, ils assurent leur propre dispersion par un mécanisme invasif impliquant un système de transport de type IV constitué par un complexe de plus d'une dizaine de protéines. Ainsi, les plasmides contiennent des gènes en forte interaction fonctionnelle entre eux mais dans un génome hôte qui n'aurait que peu d'influence sur la réplication du plasmide. Un projet de séquençage de 125 plasmides de *E. coli* portant des gènes de résistance aux antibiotiques est en train d'être mis en place. Dans ce cadre, une demande de financement sera déposé auprès de l'ANR par Eduardo Rocha (entre autres), et devrait permettre d'obtenir des données intéressantes sur l'évolution et l'adaptation de ces entités parasites.

Je propose donc de m'intéresser à trois systèmes qui possèdent des caractéristiques très différentes. Le ribosome est une structure essentielle, ubiquitaire, très conservée et

impliquant un assemblage complexe de plusieurs dizaines de gènes. L'opéron lactose est constitué de peu de protéines, est non essentiel et a récemment été acquis ; il est donc a priori en train de s'adapter au génome de *E. coli*. Finalement, les plasmides conjugatifs ont des gènes codant pour des protéines en interaction étroite, mais qui sont remarquablement indépendantes du génome de l'hôte.

Conclusion

Après la fin de ma délégation, je reprendrai mes enseignements, indispensables à mon équilibre scientifique. J'ai perçu l'enseignement comme une excellente occasion de m'ouvrir à d'autres horizons, d'autres collègues que je n'aurais pas eu la chance de rencontrer via mon activité de recherche. L'enseignement représente également, pour moi, la phase ultime de compréhension d'une problématique scientifique. En extraire la quintessence et la rétablir sous forme d'un enseignement est l'un des exercices les plus difficiles qu'il m'ait été donné de pratiquer. Enseigner permet donc, non seulement d'accroître sa culture et sa curiosité scientifique, mais également met à l'épreuve des connaissances que l'on tenait pour acquises.

A l'aide d'outils d'analyse de données et de modélisation, j'ai développé des thèmes de recherche en évolution moléculaire. La plupart des projets scientifiques auxquels j'ai participé sont le fruit de collaboration, tantôt avec des étudiants, tantôt avec des chercheurs confirmés. Développer un large réseau de collaboration me parait l'une des étapes essentielles à la mise en place d'une recherche dynamique et alimentée par les flux d'idées. La diffusion des idées scientifiques s'opère par une opération impalpable, proche de la diffusion osmotique. Déterminer comment naissent les idées est une tache décourageante. Dans une vision poétique, les idées naissent dans l'éther ; elles mûrissent dans l'esprit de tous et finalement se concrétisent par l'intermédiaire d'un ou de quelques chercheurs providentiels.

La diffusion des idées scientifiques se fait via les publications et les conférences, mais également au détour de conversations *de visu*. La discussion est la seule forme de diffusion dynamique, qui autorise un aller-retour entre les participants. En ce sens, un réseau de collaboration permet de s'ouvrir aux flux d'idées qui circulent entre tous. Ce réseau est également un garde-fou contre l'hermétisme, si pénalisant pour le chercheur isolé. Une découverte scientifique totalement cachée ou hermétique est, à mes yeux, peine perdue.

Je compte poursuivre mon travail sur l'impact de la composition nucléotidique sur la densité de séquences répétées dans les génomes. Je chercherai en particulier à mettre en relation le taux de duplication prédit (par la composition et par la densité de répétitions

observée) et le nombre de séquences fonctionnelles répétées (e.g. les gènes dupliqués). Cette comparaison pourrait nous éclairer sur les relations qui existent entre répétitions génétiques et redondances fonctionnelles. En particulier, elle pourrait nous renseigner sur le rôle de la sélection naturelle ?

Pour ce qui est des tests de neutralité, le travail entrepris ouvre de larges perspectives de développement. D'une part, il serait intéressant d'élargir la forme générique afin qu'elle englobe d'autres tests de neutralité, comme ceux basés sur les haplotypes (voir, par exemple, Fu (1996); Depaulis and Veuille (1998)). D'autre part, ce travail appelle le développement d'une méthode d'optimisation qui permettra d'utiliser au mieux la forme générique pour répondre à certaines exigences.

J'ai participé à plusieurs études non détaillées dans ce manuscrit. Certaines sont encore en développement aujourd'hui. Soucieux de ne pas élargir inutilement un catalogue rébarbatif de projets, je les passerai sous silence. Je souhaite cependant mentionner les analyses de données populationnelles et phylogénétiques sur le genre *Flavobacterium*. Depuis 2008, je participe au projet financé par l'ANR et porté par Eric Duchaud "Flavophylogenomics" qui finance une partie de ma recherche. Ce projet a des visées aussi bien épidémiologiques (e.g. typage de souches de *Flavobacterium*) qu'évolutives (e.g. exploration de la diversité intra et inter-spécifique au sein du genre). L'analyse de ces données très riches devrait prendre pleinement son essor dans l'année à venir.

Si je devais dégager une grande ligne directrice pour mes travaux à long terme, je pencherais pour l'examen des modèles de références d'évolution moléculaire. A ce jour, le modèle de référence adopté par la communauté est celui imposé par la théorie neutraliste, celui d'un équilibre mutation-dérive dans une population de taille constante. Mon travail sur le HIV-1 m'a ouvert les yeux sur la fragilité des raisons qui justifie le modèle de référence. Je n'ai pas connaissance d'argument biologique fort qui défendrait le modèle neutre comme modèle de référence. Bien sûr, il présente cet avantage d'être simple et d'avoir été très étudié. Mais est-ce un argument recevable ? En l'état des connaissances actuelles peut-être. Cependant, ne serait-il pas plus confortable d'asseoir la justesse de ce choix sur une raison biologique solide ? Or, je n'en connais pas. Tout au contraire. L'un de mes objectifs de travail à long terme est de recenser les différences les plus fortes qui existent entre différents modèles candidats à être des références et d'en éprouver leur pertinence biologique.

Au moins deux modèles de références alternatifs sont à examiner. Le premier, proche du modèle actuel de référence, est celui d'un processus de branchement assez généraliste, dans

lequel la taille de la population n'est pas contrainte à être fixée. J'ai fait la connaissance de ces modèles grâce à ma collaboration avec Amaury Lambert. Dans ces modèles, la neutralité est respectée, mais la taille est assimilée à une variable aléatoire. Nous sommes actuellement en train d'examiner les différences quantitatives entre le modèle de référence actuel et celui d'un processus de branchement critique. Ce dernier est, parmi ces processus de branchement, celui qui se rapproche le plus du modèle de référence puisque la population a une taille constante en moyenne.

Un autre modèle alternatif est celui défendu depuis longtemps par Gillespie (Gillespie, 1991). Dans ce modèle, les mutations neutres ont un rôle secondaire. Ce sont les mutations ayant un impact sur la sélection qui ont la première place. Le projet que je vais débuter au Japon sur l'adaptation des génomes s'inscrit pleinement dans l'étude de la théorie de l'adaptation. Depuis 1859, la principale cause de l'évolution était la sélection naturelle. Depuis l'avènement de la théorie neutraliste dans les années 1970, la place de la sélection naturelle en évolution moléculaire a régressé jusqu'à devenir une exception. Mais est-ce réellement justifié ? C'est précisément sur cette question que je souhaiterais me pencher dans le futur.

Bibliography

Abramson, J., I. Smirnova, V. Kasho, G. Verner, H. R. Kaback, *et al.*, 2003 Structure and mechanism of the lactose permease of escherichia coli. Science **301**: 610–615.

Achaz, G., S. Palmer, M. Kearney, F. Maldarelli, J. W. Mellors, *et al.*, 2004 A robust measure of hiv-1 population turnover within chronically infected individuals. Mol Biol Evol **21**: 1902–1912.

Achaz, G., E. P. C. Rocha, P. Netter, and E. Coissac, 2002 Origin and fate of repeats in bacteria. Nucleic Acids Res **30**: 2987–2994.

Altschul, S. F., and W. Gish, 1996 Local alignment statistics. Methods Enzymol **266**: 460–480.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, *et al.*, 1997 Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res **25**: 3389–3402.

Amabile-Cuevas, C. F., and M. E. Chicurel, 1992 Bacterial plasmids and gene flux. Cell **70**: 189–199.

Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya, *et al.*, 1998 Zebrafish hox clusters and vertebrate genome evolution. Science **282**: 1711–1714.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny, 2000 Extensive duplication and reshuffling in the arabidopsis genome. Plant Cell **12**: 1093–1101.

Carter, A. P., W. M. Clemons, D. E. Brodersen, R. J. Morgan-Warren, B. T. Wimberly, *et al.*, 2000 Functional insights from the structure of the 30s ribosomal subunit and its interactions with antibiotics. Nature **407**: 340–348.

Chang, D. K., D. Metzgar, C. Wills, and C. R. Boland, 2001 Microsatellites in the eukaryotic dna mismatch repair genes as modulators of evolutionary mutation rate. Genome Res **11**: 1145–1146.

Chiaromonte, F., V. B. Yap, and W. Miller, 2002 Scoring pairwise genomic sequence alignments. Proceedings of the Pacific Symposium on Biocomputing : 115–126.

Dawkins, R., 1976 *The Selfish Gene*. Oxford University Press.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol **15**: 1788–1790.

DePristo, M. A., M. M. Zilversmit, and D. L. Hartl, 2006 On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene **378**: 19–30.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148**: 1667–1686.

Dutheil, J., and N. Galtier, 2007 Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol **7**: 242.

Dutheil, J., T. Pupko, A. Jean-Marie, and N. Galtier, 2005 A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol **22**: 1919–1928.

Ewens, W. J., and G. R. Grant, 2001 *Statistical methods in Bioinformatics*. Springer-Verlag.

Fernandez-Lopez, R., M. P. Garcillan-Barcia, C. Revilla, M. Lazaro, L. Vielva, *et al.*, 2006 Dynamics of the incw genetic backbone imply general trends in conjugative plasmid evolution. FEMS Microbiol Rev **30**: 942–966.

Fisher, R. S., 1930 *The genetical theory of natural selection (2007 eds)*. Oxford University Press.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**: 1531–1545.

Frezal, L., and R. Leblois, 2008 Four years of dna barcoding: current advances and prospects. Infect Genet Evol **8**: 727–736.

Frost, S. D., M. J. Dumaurier, S. Wain-Hobson, and A. J. Brown, 2001 Genetic drift and within-host metapopulation dynamics of hiv-1 infection. Proc Natl Acad Sci U S A **98**: 6975–6980.

Fu, Y. X., 1996 New statistical tests of neutrality for dna samples from a population. Genetics **143**: 557–570.

Fu, Y. X., 2001 Estimating mutation rate and generation time from longitudinal samples of dna sequences. Mol Biol Evol **18**: 620–626.

Gillespie, J. H., 1991 *The causes of molecular evolution*. Oxford University Press.

Gillespie, J. H., 2000a Genetic drift in an infinite population. the pseudohitchhiking model. Genetics **155**: 909–919.

Gillespie, J. H., 2000b The neutral theory in an infinite population. Gene **261**: 11–18.

Gillespie, J. H., 2001 Is the population size of a species relevant to its evolution? Evolution **55**: 2161–2169.

Griffiths, A. J. F., J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart, 1993 *An introduction to genetic analysis (fifth edition)*. Freeman and Company.

Gu, Z., D. Nicolae, H. H.-S. Lu, and W. H. Li, 2002 Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet **18**: 609–613.

Haldane, J. B. S., 1932 *The causes of evolution*. Cornell Univ. Press.

Hein, J., M. H. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution*. Oxford University Press.

Hudson, R. R., D. D. Boos, and N. L. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol Biol Evol **9**: 138–151.

Hughes, A. L., 1994 The evolution of functionally novel proteins after gene duplication. Proc Biol Sci **256**: 119–124.

Jacob, F., and J. Monod, 1961 Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol **3**: 318–356.

Jacobson, R. H., X. J. Zhang, R. F. DuBose, and B. W. Matthews, 1994 Three-dimensional structure of beta-galactosidase from e. coli. Nature **369**: 761–766.

Jaillon, O., J.-M. Aury, F. Brunet, J.-L. Petit, N. Stange-Thomann, *et al.*, 2004 Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature **431**: 946–957.

Karlin, S., and S. F. Altschul, 1990 Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A **87**: 2264–2268.

Karlin, S., and S. F. Altschul, 1993 Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci U S A **90**: 5873–5877.

Karp, R. M., R. E. Miller, and A. L. Rosenberg, 1972 Rapid identification of repeated patterns in strings, trees and array. In *4th annual ACM symposium theory of computing*. ACM, 125–136.

Kaufman, S., 1993 *Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

Lambert, A., 2007 The contour of splitting trees is a levy process.

Lambert, A., 2008 The allelic partition for coalescent point processes.

Landraud, A. M., J. F. Avril, and P. Chretienne, 1989 An algorithm for finding a common structure shared by a family of strings. In *IEEE transactions on pattern analysis and machine intelligence*, volume 11. 890–895.

Leigh Brown, A. J., 1997 Analysis of hiv-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci U S A **94**: 1862–1865.

Lewis, M., G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, *et al.*, 1996 Crystal structure of the lactose operon repressor and its complexes with dna and inducer. Science **271**: 1247–1254.

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154**: 459–473.

Lynch, M., M. O'Hely, B. Walsh, and A. Force, 2001 The probability of preservation of a newly arisen gene duplicate. Genetics **159**: 1789–1804.

Makova, K. D., and W.-H. Li, 2003 Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res **13**: 1638–1645.

Maynard Smith, J., 1970 Natural selection and the concept of a protein space. Nature **225**: 563–564.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genetical research **23**: 23–35.

Morange, M., 1998 *La part des gènes*. Odile Jacob.

Moss, L., 2002 *What genes can't do*. The MIT Press.

Nei, M., 1987 *Molecular evolutionary genetics*. Columbia university press.

Ochman, H., J. G. Lawrence, and E. A. Groisman, 2000 Lateral gene transfer and the nature of bacterial innovation. Nature **405**: 299–304.

Ohno, S., 1970 *Evolution by gene duplication*. Springer-Verlag.

Orr, H. A., 2002 The population genetics of adaptation: the adaptation of dna sequences. Evolution **56**: 1317–1330.

Orr, H. A., 2005 The genetic theory of adaptation: a brief history. Nat Rev Genet **6**: 119–127.

Phillips, P. C., 2008 Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet **9**: 855–867.

Platt, A., 2004 *Properties and Implications of Context-sensitive Mutational Processes*. Ph.D. thesis, Harvard University.

Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, *et al.*, 2006 Sequence-based species delimitation for the dna taxonomy of undescribed insects. Syst Biol **55**: 595–609.

Potter, S. J., P. Lemey, G. Achaz, C. B. Chew, A.-M. Vandamme, *et al.*, 2004 Hiv-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy. J Leukoc Biol **76**: 562–570.

Price, A. L., N. C. Jones, and P. A. Pevzner, 2005 De novo identification of repeat families in large genomes. Bioinformatics **21 Suppl 1**: i351–8.

Price, G. R., 1970 Selection and covariance. Nature **227**: 520–521.

Rice, S., 2004 *Evolutionary theory: mathematical and conceptual foundations*. Sinauser Associates.

Robin, S., F. Rodolphe, and S. Schbath, 2003 *ADN, mots et modèles*. Belin.

Schuwirth, B. S., M. A. Borovinskaya, C. W. Hau, W. Zhang, A. Vila-Sanjurjo, *et al.*, 2005 Structures of the bacterial ribosome at 3.5 a resolution. Science **310**: 827–834.

Seo, T.-K., J. L. Thorne, M. Hasegawa, and H. Kishino, 2002 Estimation of effective population size of hiv-1 within a host: a pseudomaximum-likelihood approach. Genetics **160**: 1283–1293.

Shankar, R., A. Chaurasia, B. Ghosh, D. Chekmenev, E. Cheremushkin, *et al.*, 2007 Non-random genomic divergence in repetitive sequences of human and chimpanzee in genes of different functional categories. Mol Genet Genomics **277**: 441–455.

Shimeld, S. M., 1999 Gene function, gene networks and the fate of duplicated genes. Semin Cell Dev Biol **10**: 549–553.

Shriner, D., Y. Liu, D. C. Nickle, and J. I. Mullins, 2006 Evolution of intrahost hiv-1 genetic diversity during chronic infection. Evolution **60**: 1165–1176.

Shriner, D., R. Shankarappa, M. A. Jensen, D. C. Nickle, J. E. Mittler, *et al.*, 2004 Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. Genetics **166**: 1155–1164.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for dna polymorphism data. Genetics **141**: 413–429.

Soldano, H., A. Viari, and M. Champesme, 1995 Searching flexible repeated patterns using a non transitive relation. Pattern recognition letters **16**: 233–246.

Tajima, F., 1983 Evolutionary relationship of dna sequences in finite populations. Genetics **105**: 437–460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics **123**: 585–595.

Takahasi, K. R., and F. Tajima, 2005 Evolution of coadaptation in a two-locus epistatic system. Evolution **59**: 2324–2332.

Taylor, J. S., and J. Raes, 2004 Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet **38**: 615–643.

Tenaillon, O., H. Le Nagard, B. Godelle, and F. Taddei, 2000 Mutators and sex in bacteria: conflict between adaptive strategies. Proc Natl Acad Sci U S A **97**: 10465–10470.

Thomas, E. E., N. Srebro, J. Sebat, N. Navin, J. Healy, *et al.*, 2004 Distribution of short paired duplications in mammalian genomes. Proc Natl Acad Sci U S A **101**: 10349–10354.

Treangen, T. J., A. E. Darling, G. Achaz, M. A. Ragan, X. Messeguer, *et al.*, 2009 A novel heuristic for local multiple alignment of interspersed dna repeats. IEEE/ACM Trans Comput Biol Bioinform **6**: 180–189.

Waddington, C. H., 1942 Canalization of development and the inheritance of acquired characters. Nature **150**: 563–565.

Wagner, A., 2000 Decoupled evolution of coding region and mrna expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc Natl Acad Sci U S A **97**: 6579–6584.

Wakeley, J., 2009 *Coalescent theory, an Introduction*. Roberts and Company.

Waterman, M. S., 1995 *Introduction to computational biology. Maps, sequences and genomes*. Chapman and Hall.

Waterman, M. S., and M. Vingron, 1994 Rapid and accurate estimates of statistical significance for sequence data base searches. Proc Natl Acad Sci U S A **91**: 4625–4628.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7**: 256–276.

Weinreich, D. M., N. F. Delaney, M. A. Depristo, and D. L. Hartl, 2006 Darwinian evolution can follow only very few mutational paths to fitter proteins. Science **312**: 111–114.

Wimberly, B. T., D. E. Brodersen, W. M. J. Clemons, R. J. Morgan-Warren, A. P. Carter, *et al.*, 2000 Structure of the 30s ribosomal subunit. Nature **407**: 327–339.

Wong, S., G. Butler, and K. H. Wolfe, 2002 Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc Natl Acad Sci U S A **99**: 9272–9277.

Wright, S., 1932 The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress on genetics : 355–366.

Yahya, H., 2002 *Le mensonge de l'évolution*. Essalam.

Yukilevich, R., J. Lachance, F. Aoki, and J. R. True, 2008 Long-term adaptation of epistatic genetic networks. Evolution **62**: 2215–2235.

Yusupov, M. M., G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, *et al.*, 2001 Crystal structure of the ribosome at 5.5 a resolution. Science **292**: 883–896.

Curriculum vitae

09 décembre 1974
Maître de conférences à l'UPMC (Université Pierre et Marie Curie)
UMR 7138 (Systématique, Adaptation, Evolution)
http://wwwabi.snv.jussieu.fr/achaz/

Principaux intérêts scientifiques

Biologie Théorique, Biologie de l'Evolution, Bioinformatique

Diplômes universitaires

2002	Doctorat en Génétique - UPMC
1998	DEA en Génétique - UPMC
1996,1997	Licence puis Maîtrise en Biochimie - UPMC

Recherche

Depuis 2005	Maître de Conférences - UPMC
2002-2005	Stage Postdoctoral en Génétique des Populations - Harvard University avec J Wakeley - Turn-over des population de HIV-1
1998-2002	Doctorat en Génétique Evolutive - Institut Jacques Monod (Paris) avec P Netter et E Coissac- Etude des répétitions intrachromosomiques.

Financement recherche

2008-2010	Partenaire responsable dans le projet ANR *Flavophylogenomics*.
2003	Bourse sur projet - Fondation Singer Polignac.
2002	Bourse de stage post-doctoral - Fondation pour la Recherche Médical.

Encadrements et collaborations

Stage de doctorat

2006-2009	80% Etienne Loire, doctorant à l'école doctorale Logique du Vivant. *Impact des microsatellites codants sur la mutabilité des gènes* (évolution moléculaire *in silico*). Soutenance prévue pour octobre 2009.
2004	20% Jean-François Rual, doctorant au Dana Farber Cancer Institute (Harvard). *Détection d'un seuil de similarité minimum pour observer de l'ARN interférence* (biologie moléculaire *in silico*). Thèse soutenue en 2004.
2003	20% Cristian Castillo-Davis, doctorant au Organismic and Evolutionnary Biology (Harvard). *Etude du couplage entre l'évolution des gènes et de leurs promoteurs* (évolution moléculaire *in silico.*) Thèse soutenue en 2004.

Stages de M2

2006	100% Etienne Loire, étudiant en M2 - UPMC. *Etude des microsatellites codants sur la mutabilité des gènes humains* (évolution moléculaire *in silico*).

Stages de M1

2001	80% Warren Albertin, étudiante en Maîtrise de Génétique - Université Denis Diderot. *Etude de la recombinaison ectopique dans un plasmide de levure* (biologie moléculaire).

Collaborations an cours

2005-	S Brouillet (UPMC), F Maldrarelli (Drug Resistance Program, USA) et J Coffin (Drug Resistance Program, USA). Impact à court terme du traitement antiviral sur la population de HIV-1 (génétique des populations).
2008-	N Puillandre (Utah University) et A Lambert (UPMC - UFR de Mathématiques). Délimitation d'espèces et étude de la variabilité intra et inter espèce (spéciation et génétique des populations).
2005-	F Maldarelli (Drug Resistance Program, USA) et J Kovacs (National Insitute of Health, USA). Etude de la recombinaison chez *Pneumocystis jirovecii* (génétique des populations).

2008- T Treangen (Institut Pasteur) et EPC Rocha (Institut Pasteur). Mise au point et implémentation d'un algorithme de construction *de novo* de familles de séquences répétées (bioinformatique).

2007- E Duchaud (INRA) et P Nicolas (INRA). Génomique des populations de la bactérie *Flavobacterium psychrophilum* (génétique des populations).

Enseignement

Depuis 2005 Maître de conférences - UFR des Sciences de la Vie (UFR 927) - UPMC.

- Licence : Informatique, Statistiques et Mathématiques.
- Master et Doctorat : Génétique des Populations, Evolution Moléculaire et Bioinformatique
- Responsabilité : Co-responsable du module de Licence *Outils Mathématiques pour les Scientifiques* (LM130). Responsable du module de Master *Génétique Multifactorielle* (MV422)

1999-2002 Moniteur puis Vacataire - Université Denis Diderot.

- Licence : Génétique
- DESS : Bioinformatique

Publications scientifiques

Les publications correspondant aux encadrements de doctorants (en gras) sont indiquées par une ∗.

1. G Achaz. *Frequency spectrum neutrality tests, one for all and all for one.* **Genetics (2009).** Sous presse.
2. T Treangen, A Darling, G Achaz, M Ragan, X Messeguer, EPC Rocha. *A novel heuristic for local multiple alignment of interspersed DNA repeats.* **IEEE/ACM Transactions on Computational Biology and Bioinformatics (2009)**, 6(2):180-9.
3. * **E Loire**, F Praz, D Higuet, P Netter, G Achaz. *Hypermutability of genes in Homo sapiens due to the hosting of long mono-SSR.* **Molecular Biology and Evolution (2009)**, 26(1):111-121

4. G Achaz. *Testing for neutrality in samples with sequencing errors.* **Genetics (2008)**, 179(3):1409-24
5. P Nicolas, S Mondot, G Achaz, C Bouchenot, JF Bernardet, E Duchaud. *Population structure of the fish-pathogenic bacterium Flavobacterium psychrophilum.* **Applied Environnemental Microbiology (2008)**, 74(12):3702-9.
6. G Kutty, F Maldarelli, G Achaz, JA Kovacs. *Variation in the major surface glycoprotein genes in Pneumocystis jirovecii.* **Journal of Infectious Disease (2008)**, 198(5):741-9.
7. * **J-F Rual**, N Klitgord and G Achaz. *Novel insights into RNAi off-target degradation using C. elegans paralogs.* **BMC Genomics (2007)**, 8(1):106.
8. G Achaz, F Boyer, EPC Rocha, A Viari and E Coissac. *Repseek, a tool to retrieve non-exact repeats from large DNA sequences.* **Bioinformatics (2007)**, 23(1):119-21
9. * **CI Castillo-Davis**, DL Hartl and G Achaz. *Cis-regulatory and protein evolution in orthologous and duplicate genes.* **Genome Research (2004)**, 14(8):1530-1536.
10. G Achaz, S Palmer, M Kearney, F Maldarelli, JW Mellors, JM Coffin and J Wakeley. *A robust measure of HIV-1 population turnover within chronically infected individuals.* **Molecular Biology and Evolution (2004)**, 21(10):1902-12.
11. SJ Potter, P Lemey, G Achaz, CB Chew, AM Vandamme, DE Dwyer and NK Saksena. *HIV-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy.* **Journal of Leukocytes Biology (2004)**, 76(3):562-70.
12. G Achaz, P Netter and E Coissac and EPC Rocha. *Associations between inverted repeats and the structural evolution of bacterial genomes.* **Genetics (2003)**, 164(4):1279-89.
13. G Achaz, EPC Rocha, P Netter and E Coissac. *Origin and fate of repeats in bacteria.* **Nucleic Acid Research (2002)**, 30(13):2987-94.
14. G Achaz, P Netter and E Coissac. *Study of intrachromosomal duplications among eukaryotes genomes.* **Molecular Biology and Evolution (2001)**, 18(12):2280-2288.
15. G Achaz, E Coissac, A Viari and P Netter. *Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: a possible model for their origin.* **Molecular Biology and Evolution (2000)**, 17(8):1268-1275.

Communications scientifiques

2009 Séminaire invité, Population Biology (Tokyo University) ; Séminaire invité, Center for Ecological Research (Kyoto University) ; Séminaire invité, Musée de l'Homme (Paris, France)

2008 Exposé oral, Alignement et Phylogénie (Lyon, France) ; Exposé oral, GDR Génomique des Populations (Sète, France) ; Séminaire invité, Mathématiques, Informatique et Génomes (Jouy-en-Josas, France)

2007 Exposé oral, Population Genetics Group 06 (Manchester, UK) ; Poster Conférences Jacques Monod Population Genomics (Roscoff, France) ; Séminaire invité, Muséum, renouvellement du quadriennal (Paris, France) ; Séminaire invité, séminaires Génomique Institut Henri Poincaré (Paris, France)

2006 Séminaire invité, Midi-pile (Orsay, France) ; Séminaire invité, Institut Jacques Monod, séminaire de département (Paris, France) ; Poster, Journées Ouvertes en Biologie, Informatique et Mathématiques 2006 (Bordeaux, France) ; Séminaire invité, Séminaires de l'UMR 8016 (Lille, France)

2005 Exposé oral, Population Genetics Group 05 (Edinburgh, UK) ; Poster, Journées Ouvertes en Biologie, Informatique et Mathématiques 2005 (Lyon, France)

2004 Exposé oral, Molecular Biology and Evolution 2004 (College State, USA) ; Exposé oral, Journées Ouvertes en Biologie, Informatique et Mathématiques 2004 (Montréal, Canada) ; Exposé oral, Fifth HIV DRP Symposium on Antiviral Drug Resistance (Chantilly, USA) ; Séminaire invité, Harvard Center for Cancer Systems Biology (Boston, USA) ; Séminaire invité, Harvard Center for Genomic Research (Cambridge, USA) ; Séminaire invité, Tuft New England Medical Center (Somerville, USA).

2003 Exposé oral, Evolution 2003 (Chico, USA) ; Poster, Fourth HIV DRP Symposium on Antiviral Drug Resistance (Chantilly, USA).

2002 Poster, Third HIV DRP Symposium on Antiviral Drug Resistance (Chantilly, USA).

2001 Poster, Réparation, Recombinaison et Réplication (Villejuif, France) ; Poster, Gene and genome duplication (Aussois, France).

2000 Exposé oral, Journées Ouvertes en Biologie, Informatique et Mathématiques (Montpellier, France).

1999 Exposé oral, Réparation, Recombinaison et Réplication (Villejuif, France) ; Poster, International Conference on Yeast Genetics and Molecular Biology (Rimini, Italie) ; Poster, Levure, Modèle et Outil IV (Arcachon, France).

1998 Exposé oral, Club des Levuristes d'Ile de France (Orsay, France).

cis-Regulatory and Protein Evolution in Orthologous and Duplicate Genes

Cristian I. Castillo-Davis, Daniel L. Hartl, and Guillaume Achaz[1]

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, 02138 USA

The relationship between protein and regulatory sequence evolution is a central question in molecular evolution. It is currently not known to what extent changes in gene expression are coupled with the evolution of protein coding sequences, or whether these changes differ among orthologs (species homologs) and paralogs (duplicate genes). Here, we develop a method to measure the extent of functionally relevant *cis*-regulatory sequence change in homologous genes, and validate it using microarray data and experimentally verified regulatory elements in different eukaryotic species. By comparing the genomes of *Caenorhabditis elegans* and *C. briggsae*, we found that protein and regulatory evolution is weakly coupled in orthologs but not paralogs, suggesting that selective pressure on gene expression and protein evolution is quite similar and persists for a significant amount of time following speciation but not gene duplication. Additionally, duplicates of both species exhibit a dramatic acceleration of both regulatory and protein evolution compared to orthologs, suggesting increased directional selection and/or relaxed selection on both gene expression patterns and protein function in duplicate genes.

[Supplemental material is available online at www.genome.org.]

The relative importance of coding sequence change versus regulatory sequence change in evolution has vexed evolutionary geneticists for over 50 years. Given recent genomic analyses showing the conservation of many proteins among distantly related taxa, it has been proposed that regulatory changes play a key role in generating the great morphological diversity present in multicellular species. However, little is known about the evolution of gene regulation or its relationship to protein evolution. Do highly conserved genes also show conserved expression patterns? Or can gene expression evolve independently from protein function? The former pattern is expected if strong stabilizing selection acts on genes as integrated units in which protein sequence and expression pattern are not dissociable. At the same time it has been argued that "developmental systems drift" may result in reorganization of regulatory systems as long as general developmental patterns are preserved (True and Haag 2001). If so, a high turnover of gene regulatory elements may uncouple *cis*-element-mediated gene expression and protein evolution.

Differences in gene expression between species (or between duplicate genes) may entail changes in gene expression levels under the same conditions at the same developmental times, as well as gene expression changes in spatial, temporal, and environmental dimensions. Hereafter, we refer to the former changes as changes in *expression magnitude* and the latter as changes in *relative expression*. First attempts to find a correlation between the evolution of relative expression and protein evolution in yeast (using only duplicates) yielded contradictory results; one study argued that expression differences are not correlated with protein evolution (Wagner 2000), whereas more recent work suggests a weak correlation (Gu et al. 2002). However, a recent review (Wolfe and Li 2003) concluded that a wider analysis of regulatory and protein evolution is necessary. In particular, the dynamics of protein versus *cis*-regulatory evolution in duplicate genes, which are thought to play a central role in the evolution of novel molecular functions and the generation of genetic diversity (Haldane 1932; Ohno 1970), are still poorly understood.

[1]Corresponding author.
E-MAIL gachaz@oeb.harvard.edu; FAX (617) 496-5854.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2662504. Article published online before print in July 2004.

Although there is some evidence that duplicate genes undergo an increased rate of protein evolution (Lynch and Conery 2000; Kondrashov et al. 2002; Nembaware et al. 2002), a systematic analysis of *cis*-regulatory versus coding sequence change in orthologous and duplicate genes has not been carried out. This deficiency is due in large part to the lack of a biologically relevant measure of *cis*-regulatory evolution that relates directly to gene expression. The identity of *cis*-acting regulatory motifs is generally unknown, and such motifs are sparsely scattered within noncoding DNA that is under little or no selective constraint. It has thus been almost impossible to discriminate between functionally relevant and stochastic changes in putative regulatory DNA without time-consuming gene-by-gene experiments.

Accordingly, we set out to develop a method to quantify functional regulatory changes in the regulatory regions of homologous genes that does not depend on knowledge of experimentally characterized or computationally predicted DNA binding sites. We began by identifying orthologous genes between the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Next we identified duplicate genes within each genome and calculated rates of protein evolution in both duplicates and orthologs by maximum likelihood (Yang 1997). Finally, we used duplicate gene pairs in conjunction with microarray expression data to develop a method to measure regulatory evolution called the shared motif method (SMM), and used it to measure *cis*-regulatory evolution in duplicate and orthologous genes. These data are summarized in Table 1, and the list of genes is provided as Supplemental material.

RESULTS AND DISCUSSION

Regulatory Sequence Evolution

Because small intrachromosomal rearrangements resulting in changes in *cis*-element order, orientation, and spacing can occur over moderate stretches of evolutionary time, while leaving gene expression patterns intact (Ludwig et al. 2000), we detected "shared motifs" (regions of high local similarity) between the upstream regions of homologous genes. We define these motifs as conserved segments between sequences without respect to their order, orientation, or spacing (Fig. 1). By examining the

14:1530–1536 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

Table 1. Comparison of Mean Rates of Protein Evolution (d_N, d_S) and Regulatory Evolution (d_{SM}) in Orthologous and Duplicate Genes

	Number of pairs	d_S	d_N	d_{SM}
Orthologs between species	2,150	1.11 (0.31)	0.07 (0.06)	0.59 (0.22)
Duplicates within *C. elegans*	869	0.57 (0.43)	0.17 (0.15)	0.61 (0.30)
Duplicates within *C. briggsae*	542	0.60 (0.41)	0.22 (0.20)	0.64 (0.31)

Standard errors are given in parentheses.

cumulative fraction of shared motifs between sequences we defined a measure of functional regulatory sequence evolution called shared motif divergence (d_{SM}). By definition, d_{SM} is the fraction of both sequences that *does not* contain a region of significant local similarity by these criteria. For example, a d_{SM} of 0 indicates a complete sharing of motifs between the sequences, whereas a d_{SM} of 1 indicates an absence of shared motifs. Note that this measure is similar to a distance metric but has a maximum value of 1. Values of $d_{SM} = 1$ are not necessarily equally divergent and should not be compared because they are "saturated" with sequence differences. In this study, the mean d_{SM} was 0.59 between species and 0.61 and 0.64 among duplicate genes in the *C. elegans* and *C. briggsae* genomes, respectively (Table 1).

Because genome-wide expression data for *C. briggsae* are not available, we validated our measure of regulatory sequence evolution using pairs of duplicate genes within the *C. elegans* genome. Because little is known about the average size of regulatory regions in *C. elegans*, we looked for shared motifs 100, 500, and 1000 bp upstream from annotated translation start sites. Among genes with annotated transcription start sites in *C. elegans*, we found no significant difference in d_{SM} when calculated from transcription start versus translation start (Supplemental material).

Divergence between upstream sequences of each duplicate pair was measured by the shared motif method and was compared with (1) differences in the *magnitude* of expression across the life cycle of *C. elegans* in absolute numbers of transcripts as assessed by Affymetrix microarrays (Hill et al. 2000), and with (2) differences in *relative expression*, first across eight developmental stages (Hill et al. 2000) and then across 553 cDNA microarray experiments that included different nutrient conditions, developmental stages, and mutants (Kim et al. 2001). Data from cDNA microarrays describe only relative changes in gene expression, and therefore differences in gene expression magnitude cannot be determined from these results. Note also that all estimates of expression level are for genes that were reliably detected (Methods), and these are likely to be moderately to highly expressed.

We found only a marginally significant correlation between d_{SM} and difference in relative expression using Affymetrix expression data ($r_s = 0.23$, $P < 0.07$, Spearman rank correlation), and no significant correlation using the cDNA microarray data (Supplemental material; Kim et al. 2001). In contrast, we observed a highly significant correlation between d_{SM} and difference in gene expression magnitude ($r_s = 0.47$, $P < 10^{-3}$) for upstream sequences of 500 bp (Fig. 2). Shorter and longer upstream sequences were less correlated with expression difference (data not shown). Importantly, no significant correlation was detected between gene similarity (estimated by d_N) and expression difference ($r_s = -0.10$, $P = 0.18$) which is expected if cross-hybridization of transcripts on Affymetrix arrays is a significant phenomenon. Although a weak correlation of synonymous substitution rate (d_S) and expression difference was detected ($r_s = 0.23$, $P = 0.02$), it disappeared after correcting for the relationship between d_{SM} and d_S ($r_s = 0.40$, $P < 0.001$) in a multiple regression analysis (expression difference vs. d_S, B = 52.78, $P = 0.34$, multiple regression formula: expression difference $\sim d_{SM} + d_N + d_S$). These results taken together imply that d_{SM} correlates with a functional difference in the gene expression magnitude between genes which is not a simple consequence of gene similarity (time since divergence or duplication event). Considering that (1) the measure d_{SM} does not take into account differences in *trans*-acting factors or other mechanisms that mediate gene expression, and (2) that its performance is based on whole-animal assays, where changes in spatial expression will act to decrease the correlation between d_{SM} and expression difference, the correlation between d_{SM} and expression difference is remarkably high. Moreover, it is the first predictor available for estimating the expression difference between genes based on comparative sequence data alone.

As a further negative control, we examined the d_{SM} of orthologous sequences in regions located further upstream of the translation start (1–1.5 kb) where the density of functional motifs is presumably lower. Orthology of these noncoding regions was inferred if both species exhibited syntenic conservation of the adjacent gene. The 1–1.5-kb upstream region was significantly more diverged on average than the 0–500-bp upstream test region used in the analysis (mean $d_{SM} = 0.78$ versus 0.60, $n = 362$, $P \ll 10^{-4}$, Wilcoxon *U*-test). Further, the control region showed a distribution of d_{SM} similar to that obtained between random sequences (mean $d_{SM} = 0.89$) compared with the 0–500-bp test region (Supplemental Fig. 1). This result suggests that the conservation detected by the SMM is not a simple consequence of historical identity.

Although the SMM is not a motif discovery algorithm, the conserved blocks of sequences discovered by the method should include a high fraction of *cis*-acting elements that are experimentally known to be involved in gene regulation. We analyzed in detail the upstream sequence of a gene known to be up-regulated

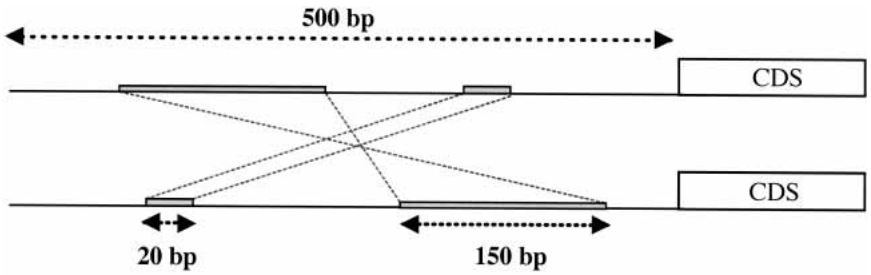


Figure 1 Illustration of the shared motif method (SMM). The SMM discovers regions of local similarity between DNA sequences without respect to their order, orientation, or spacing. In this example, two 500-bp noncoding sequences, upstream from homologous coding sequences (CDS), are compared. After iterative local alignment in both their native and inverted sequence orientations (Methods), two regions of significant local similarity between the sequences were discovered. One region is 150 bp long but has been inverted in one of the sequences. The other is 20 bp long but has been translocated. The fraction of "shared motifs" between these sequences is simply (20 + 150) / 500, or 0.34. We define shared motif *divergence* (d_{SM}) as one minus this fraction, or 1 − 0.34 = 0.66. Shared motif divergence is thus the fraction of the two sequences that does not contain a region of significant local alignment without respect to order, orientation, or spacing. Note, this example is a simplified caricature. Real sequence comparisons often exhibit more complex patterns of shared motif conservation (Supplemental Fig. 1).

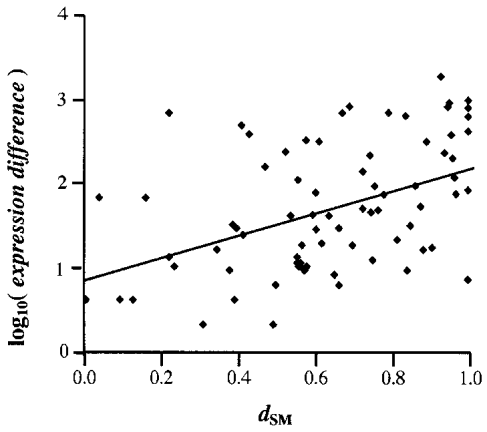


Figure 2 Correlation between d_{SM} and difference in magnitude of gene expression. We found a significant positive correlation between expression difference and shared motif divergence (d_{SM}) in sequences 0–500 bp upstream of translation start between duplicate genes in duplicate families with two to five members ($n = 76$, $r_s = 0.47$, $P < 10^{-3}$; Spearman rank correlation). The log(expression difference) is linearly correlated with d_{SM} ($R = 0.46$; $P \ll 10^{-3}$; Pearson linear correlation); a linear fit of the data is also plotted ($y = 0.85 + 1.37x$). Similar results were obtained with strict duplicate pairs or duplicate gene families of up to 10 members (data not shown).

in response to heat shock in *C. elegans* (*F44E5.5*), which contains experimentally characterized *cis*-regulatory elements conserved in *C. briggsae* (GuhaThakurta et al. 2002). The conserved sequences identified by the SMM contained eight of nine experimentally verified *cis*-elements shared between the species. We found similar results (Supplemental Fig. 2) for experimentally verified motifs at the *even-skipped* locus in *Drosophila melanogaster* / *D. pseudoobscura* (Ludwig et al. 2000), for *Apetala-3* in *Arabidopsis thaliana* / *Brassica oleracea* (Koch et al. 2001), and for *CKM* in *Homo sapiens* / *Mus musculus* (Wasserman et al. 2000). As a further positive control, we analyzed the upstream regions of a large set of human genes for which there was one or more experimentally characterized binding sites and a known orthologous gene in mouse (Supplemental material). Of 79 experimentally verified motifs among 20 different orthologous genes, 62 of 79 motifs (78%) were contained within conserved blocks discovered by the SMM (Supplemental material; Supplemental Table 2). The mean number of verified regulatory motifs in these 1-kb upstream regions was 3.95 and, on average, 3.10 motifs were found using the SMM.

cis-Regulatory and Protein Evolution are Weakly Coupled in Orthologs

We observed a positive correlation between functional regulatory evolution (d_{SM}) and protein evolution (d_N) in orthologs (Table 2, Fig. 3). As similarities in local mutation rate, or similar divergence times, may lead to the observed correlation between protein coding and noncoding change, we carried out multiple regressions involving d_N, d_{SM}, and d_S, using d_S as a simple measure of age/mutation rate.

Interestingly, there is a weak but significant correlation between protein and *cis*-regulatory evolution after controlling for the possibility that this correlation is a consequence of a similarity in local mutation rates (d_{SM} vs. d_N, B = 0.42, $P < 10^{-6}$, multiple regression formula: $d_{SM} \sim d_N + d_S$), a consequence of poor gene prediction (Supplemental material), or due to inclusion of genes in operons (data not shown).

This implies that, for a given gene, there exists a significant coupling between rates of coding sequence and *cis*-regulatory sequence change—and by inference, a potential coupling of protein function and gene expression change. Such a correlation, which has been shown for very young duplicate genes in yeast (Gu et al. 2002), seems to hold for orthologous genes shared between the more distantly related *C. elegans* and *C. briggsae*. Thus many genes that are conserved at the protein level also show conserved *cis*-regulation as predicted by the hypothesis that strong stabilizing selection acts on genes as integrated units of evolution. If divergence reflects the action of purifying selection, a correlation in *cis*-regulatory and protein divergence implies that the selective consequences of a deleterious mutation in either the *cis*-regulatory or protein coding sequence of a given gene are similar. If so, many genes are "selectively important" for an organism in a manner that is not dissociable into protein product and expression pattern components, even over long stretches of evolutionary time.

On the other hand, the observation that coupling between regulatory and protein evolution is generally weak argues that some amount of "network drift" in *cis*-element-mediated gene expression may indeed occur. The maintenance of stable gene expression patterns in the face of coevolution of transcription factors and their *cis*-acting DNA binding sites (Shaw et al. 2002) as well as wholesale rearrangement of promoter architecture (Ludwig et al. 2000) has been experimentally demonstrated for a least two loci in *Drosophila*, *bicoid* and *even-skipped*, respectively. Such co-evolutionary drift may act to weaken the correlation between *cis*-regulatory and protein evolution across the genome, as expected under a general model of stabilizing selection. A definitive test of this hypothesis will require an independent assessment of gene expression patterns in *C. briggsae* in conjunction with the analysis of *cis*-regulatory evolution carried out here.

Note that the distinction between protein and *cis*-regulatory evolution is not entirely clear-cut. For example, coding sequences can contain motifs that act to enhance and silence mRNA splicing in constitutively and alternatively spliced exons (Blencowe 2000), which is an arguably regulatory function. Thus, the dynamic and interrelated nature of coding and noncoding sequence change must be kept in mind when interpreting these results.

cis-Regulatory and Protein Evolution Are Not Coupled in Paralogs

Although we also observed a correlation between functional *cis*-regulatory evolution (d_{SM}) and protein evolution (d_N) in paralogs (Table 2, Fig. 3), in contrast to orthologs, we found that the correlation between protein (d_N) and regulatory evolution (d_{SM}) is a result of their correlation with d_S alone (d_{SM} vs. d_N, B = 0.046, $P = 0.565$, multiple regression formula = $d_{SM} \sim d_N + d_S$). In other

Table 2. Correlation Between Protein and Regulatory Evolution: Acceleration of Protein and Regulatory Evolution in Duplicate Genes

	Correlation between d_N and d_{SM}	d_N/d_S[a]	d_{SM}/d_S[a]	d_N/d_{SM}[a]
Orthologs between species	$r_s = 0.16$[b,c]	0.05	0.53	0.09
Duplicates within *C. elegans*	$r_s = 0.24$[b]	0.31	1.00	0.27
Duplicates within *C. briggsae*	$r_s = 0.21$[b]	0.37	1.06	0.33

[a]Median values of the distribution. Pairwise comparisons between orthologs and each set of duplicates were significant ($P \ll 10^{-4}$, Wilcoxon *U*-test).
[b]Spearman rank correlation coefficient. All correlations were significant ($P < 10^{-4}$).
[c]Correlation is still highly significant in a multivariate analysis including d_N, d_S and d_{SM} ($P \ll 10^{-4}$).

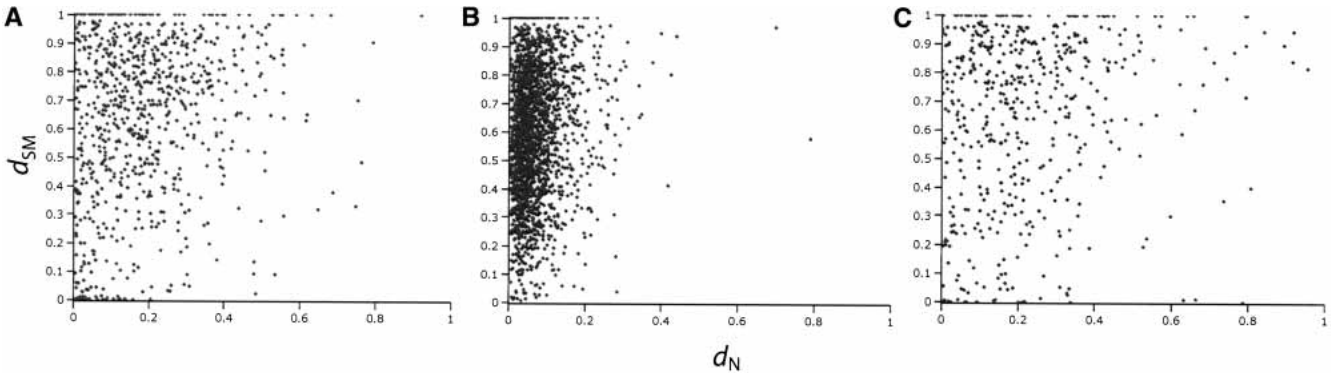


Figure 3 Correlation between protein evolution (d_N) and regulatory evolution (d_{SM}) in (*A*) paralogous genes in *C. elegans* ($r_s = 0.24$), (*B*) orthologous genes ($r_s = 0.16$), and (*C*) paralogous genes in *C. briggsae* ($r_s = 0.21$); $P \ll 10^{-4}$ for all tests. Multiple regressions that included d_S revealed that the correlation between d_N and d_{SM} in paralogs is primarily a function of d_S (i.e., duplicate age). No such effect was found in orthologs. Note that for some orthologs and duplicates, regulatory and protein evolution appear to be completely uncoupled.

words, d_N and d_{SM} increase together over time but are not themselves related. The observation that protein and regulatory evolution in paralogs is not coupled implies that these aspects of gene structure may evolve independently. It is interesting to note that this uncoupling can result from duplication events that do not encompass the entire regulatory region. Thus, shortly after duplication, d_{SM} may immediately be very high whereas d_N and d_S are close to zero. This pattern can be observed for many genes (Figs. 3, 4) and explains the seemingly paradoxical result that regulatory and protein evolution are not coupled in duplicate genes, despite the fact that both are higher in duplicates versus orthologs.

Apparent independence between *cis*-regulatory and protein sequence change is not entirely unexpected, as it has long been suggested that duplicate genes may evolve new functions (Ohno 1970; Ohta 1987; Walsh 1995) or lose them in complementary ways (Hughes 1994; Force et al. 1999) through changes in their *cis*-regulatory sequence, protein sequence, or both. Accordingly,

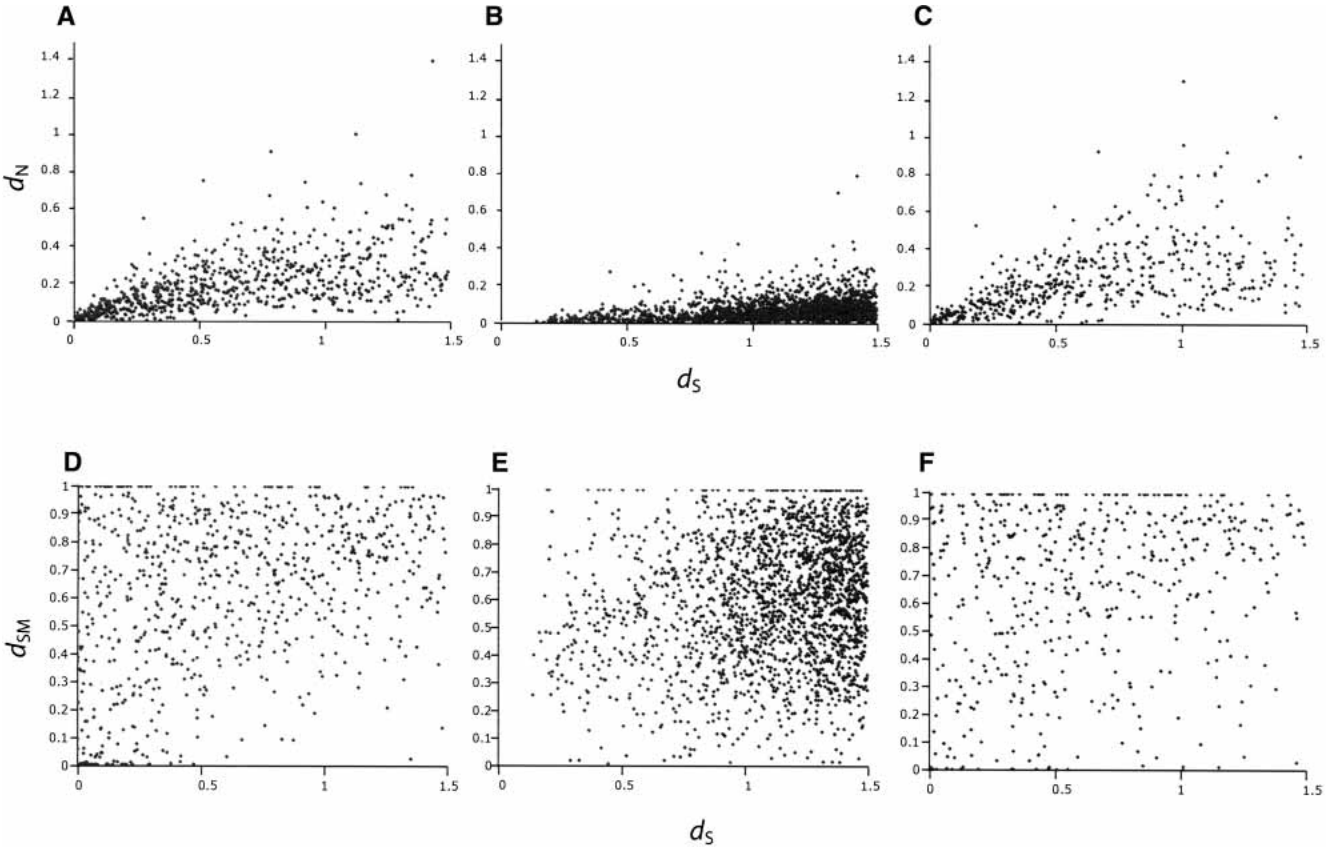


Figure 4 Rates of protein evolution (d_N/d_S) in (*A*) paralogous genes in *C. elegans*, (*B*) orthologous genes between *C. elegans* and *C. briggsae*, and (*C*) paralogous genes in *C. briggsae*. Rates of regulatory evolution (d_{SM}) in (*D*) paralogous genes in *C. elegans*, (*E*) orthologous genes between *C. elegans* and *C. briggsae*, and (*F*) paralogous genes in *C. briggsae*. In comparison with orthologs, duplicate genes in both the *C. briggsae* and *C. elegans* genomes exhibit a higher rate of amino-acid substitution and proximal *cis*-regulatory sequence evolution for the same amount of synonymous divergence.

it has been predicted that accelerated protein and/or regulatory sequence evolution will occur in duplicated genes.

Accelerated Protein and Regulatory Evolution in Duplicated Genes

Our results also indicate that, in comparison with orthologs, duplicate genes in both the *C. briggsae* and *C. elegans* genomes exhibit a significantly accelerated rate of amino-acid replacement and *cis*-regulatory evolution for the same amount of synonymous mutation (Figs. 4, 5). Overall, rates of protein evolution (d_N/d_S) are substantially accelerated in duplicate genes in both the *C. elegans* and *C. briggsae* genomes, compared to orthologs between the species ($P \ll 10^{-4}$ for each test, Wilcoxon *U*-test, Table 2). Likewise, the mean amount of regulatory evolution (d_{SM}/d_S) in duplicate genes in the genomes of both species is dramatically accelerated compared to orthologs, even though many are much younger ($P \ll 10^{-4}$ for each test, Wilcoxon *U*-test, Table 2). This pattern of accelerated evolution is similar for tandem and nontandem duplicates (Supplemental material).

At least three nonmutually exclusive scenarios could be envisaged to explain the accelerated evolution of duplicate genes. First, duplicate genes could experience weaker purifying selection than orthologs following a speciation event, that is, relaxed selection. Second, duplicate genes could experience greater positive selection than orthologs. Third, duplicates may simply be older than orthologs, that is, they predate speciation and therefore have had more time to evolve amino-acid substitutions (d_N). Under the last scenario, synonymous substitutions (d_S) would be expected to increase over time at a similar rate as d_N; however, if paralogs are more highly expressed than orthologs, they may be subject to high codon bias (Duret and Mouchiroud 1999; Castillo-Davis and Hartl 2002), which may result in underestimates of the synonymous substitution rate (d_S)—even when likelihood methods are used (Dunn et al. 2001). We find, however, that orthologs are more highly expressed (Hill et al. 2000) than paralogs ($P \ll 10^{-4}$, Wilcoxon *U*-test; data not shown), and we calculate that at least 93% (502/542) of the duplicates examined in *C. briggsae* either post-date speciation or have undergone post-speciation gene conversion (gene conversion is expected to reduce synonymous and nonsynonymous divergence equally; Methods). Thus the accelerated rates of protein and regulatory evolution in duplicate genes must be due to either relaxed selection or the action of positive selection, or both.

Protein Evolution Appears to Outpace *cis*-Regulatory Evolution in Duplicate Genes

Finally, the distribution of d_N/d_{SM} (Fig. 5) among duplicates and orthologs indicates that, for a similar amount of regulatory divergence (d_{SM}), the mean rate of amino-acid substitution (d_N) is substantially higher in duplicate genes in both *C. elegans* and *C. briggsae* compared to orthologs between the two species ($P \ll 10^{-4}$, Wilcoxon *U*-test, Table 2). One possible explanation is that this pattern is due to an overall faster rate of saturation of d_{SM} compared to d_N; however, it should be noted that values of d_{SM} are similar among both orthologs and duplicates, whereas values of d_N differ greatly (Table 1). If not due to the saturation of d_{SM}, the accelerated divergence observed in the protein-coding regions of duplicate genes may simply reflect the less deleterious consequences of changes in amino-acid sequence versus changes in gene expression in "redundant" genes (i.e., stronger purifying selection on gene regulation). Lastly, it is possible that the accelerated rate of protein evolution in duplicate genes is due to positive selection on coding regions either in both copies or only in one copy, as it has been recently posited (Conant and Wagner 2003; Kellis et al. 2004).

One indicator of positive selection is a d_N/d_S ratio significantly greater than one. We found no evidence for increased rates of positive selection in paralogs versus orthologs by this criterion (data not shown). Because positive selection across limited regions of a protein may occur without a global excess of amino-acid replacements (Nielsen and Yang 1998), we performed more sensitive likelihood-ratio tests for positive selection (Yang 2000) for all duplicate gene families of three to five members in both species (Methods). Although such tests have low power with small gene family sizes, we again found no evidence of positive selection.

The alternative explanation of the observed pattern of faster protein change in duplicates—namely that expression of a gene in the wrong tissue, cell, or at the wrong developmental timepoint may incur a high fitness cost compared to loss of protein function in one duplicate copy—therefore cannot be ruled out. Such a scenario is plausible, especially if gain-of-function mutations are more common in *cis*-regulatory sequences compared to protein coding sequences.

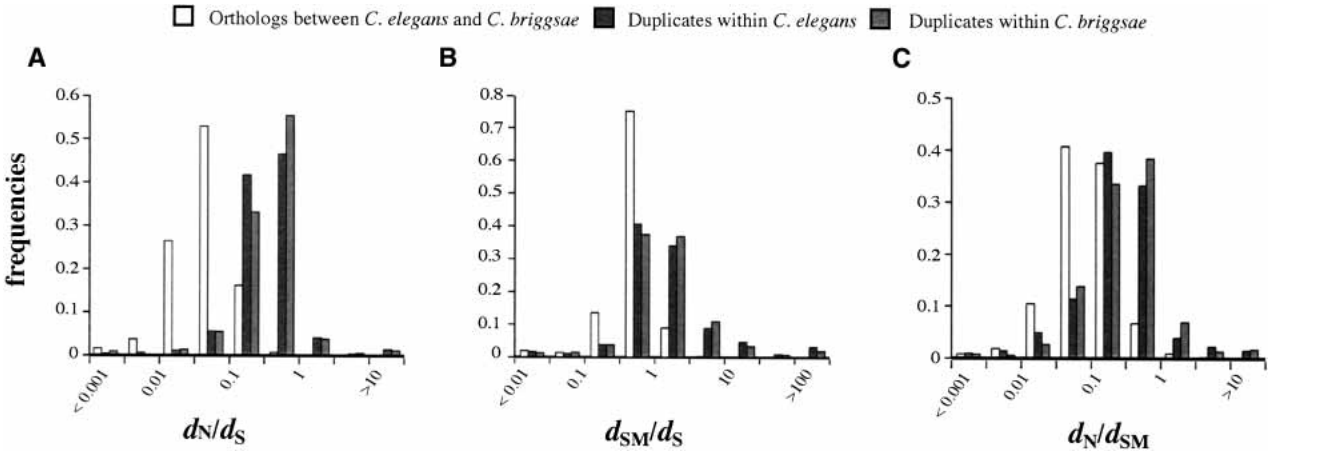


Figure 5 Histogram of the rate of (*A*) protein evolution and (*B*) regulatory evolution in orthologs between *C. elegans* and *C. briggsae* vs. paralogs within *C. elegans* and *C. briggsae*. Both protein (d_N) and regulatory evolution (d_{SM}) are accelerated in paralogs compared to orthologs for the same amount of synonymous divergence. (*C*) Histogram of protein vs. regulatory evolution in orthologs and paralogs. For the same amount of regulatory divergence, paralogs have an accelerated rate of protein evolution compared to orthologs.

Conclusions

Taken together, these observations suggest that, until genes duplicate, selection on proximal *cis*-regulation is weakly coupled to selection on protein sequences; when genes duplicate, however, it appears that selection can act independently on gene regulation and protein sequences. Selective pressure on gene expression and protein function is therefore inferred to be quite similar and persists over long stretches of evolutionary time following divergence due to speciation but not necessarily gene duplication. Additionally, compared to orthologs, duplicate genes are unique in that they exhibit dramatically accelerated rates of both *cis*-regulatory and protein evolution, suggesting increased positive and/or relaxed selection on both gene expression patterns and protein sequence in duplicate genes. Although we found no evidence for the action of positive selection in duplicate genes, the observation of accelerated rates of protein evolution over *cis*-regulatory evolution in duplicate genes is noteworthy. Further analyses should help reveal whether this pattern is due to positive selection or merely reflects the greater selective consequences of gene mis-regulation, versus the abrogation of protein function, in redundant genes.

METHODS

Protein Sequence Analysis

Coding sequences (CDSs) of the genomes of both *C. elegans* (The *C. elegans* Sequencing Consortium 1998) and *C. briggsae* (The Sanger Institute and The Genome Sequencing Center, Washington University, St. Louis, unpubl.) were obtained from WormBase (http://wormbase.org). All CDS were mapped onto genomic locations using BLASTN (Altschul et al. 1997) and when available, annotations. Only one occurrence of overlapping CDS was retained. The method of reciprocal best hits (Tatusov et al. 1997) using BLASTN was used to establish a set of orthologs between the two species ($E < 10^{-10}$ were considered significant matches). Orthologs obtained using only those genes not duplicated in either genome (1,765 / 2,150 = 82%) gave very similar results (data not shown). Duplicated genes within the *C. elegans* and *C. briggsae* genomes were obtained as follows. First a set of putative duplicate genes was obtained by significant BLASTN matches within each genome alone. Next each translated sequence of putative paralogs was globally aligned (Needleman and Wunsch 1970) against every other using the PAM250 matrix (Dayhoff et al. 1972), Gap(open) = −16 and Gap(ext) = −6. All scores were normalized using the length of the smallest of both sequences. Alignment scores >200 were considered significant. As a control, all translated *C. elegans* sequences were shuffled (Markov chains of order 0) and aligned in the same way. At this stringency, less than 0.001% of the random sequence alignments exhibit a significant score (data not shown). We considered families of five or fewer duplicate genes to avoid biases due to overrepresentation of very large gene families.

Next, all coding sequences were globally aligned by CLUSTALW (Thompson et al. 1994; default parameters) using the amino-acid translation of each sequence followed by back-translation into nucleotides. Maximum likelihood estimates of nonsynonymous substitution (d_N) and synonymous substitution (d_S) between pairwise alignments were obtained with PAML (Yang 1997) using a codon-based model of sequence evolution with d_N, d_S, and transition/transversion bias (κ) as free parameters and codon frequencies estimated from the data at each codon position (F3 × 4 model; Goldman and Yang 1994; Yang 1997). Based on simulations using random sequence pairs, pairs of sequences with $\kappa > 8$ or $d_S > 3$ were excluded from analysis. Finally, because values of $d_S > 1.5$ are prone to estimation error, we further restricted our dataset to orthologs and paralogs that exhibited a $d_S < 1.5$.

To determine the minimum proportion of duplicate genes that post-date speciation, we determined the ancestry of all *C. briggsae* duplicate genes on the basis of significant BLASTN matches to genomes of both species ($E < 10^{-10}$). Gene pairs that showed both of their closest matches within the *C. briggsae* genome were assumed to post-date speciation or have undergone gene-conversion post-speciation.

Likelihood ratio tests for positive selection within duplicate gene families were performed by comparing twice the log-likelihood difference of models M7 and M8 in PAML v3.13 (Yang 2000). This test compares the likelihood of the data under model M7 in which d_N/d_S among sites is constrained to be between 0 and 1, against model M8 where an additional category of sites with $d_N/d_S > 1$ is allowed. If the log-likelihood of the model allowing $d_N/d_S > 1$ (positive selection) is significantly greater, adaptive evolution may be inferred. Positive selection was inferred if $2(\ln L1 - \ln L2) \geq 9.21$, corresponding to $P < 0.01$ ($\sim\chi^2$, df = 2) and if d_N/d_S was greater than one among at least one of the site classes. For these tests, duplicate gene family trees were constructed using PHYLIP (Felsenstein 1993) using translated sequences with default parameters under a maximum parsimony criterion.

Regulatory Sequence Analysis

Regulatory sequences are often located 5′ to proteins, comprising 5′ UTRs, promoter regions, and other regulatory elements such as enhancers. We define a shared motif as a region of high local similarity between two DNA sequences regardless of their order, orientation, or spacing (Fig. 1).

Our method is a derived implementation of the recursive local alignment algorithm described by Waterman and Eggert (1987). This method is based on local alignment by dynamic programming (Smith and Waterman 1981) and is guaranteed to find all optimal and suboptimal alignments between two sequences. Briefly, we find the best local alignment between two sequences, then mask off this particular alignment. Next, we search for next best subalignment between the sequences and continue this process iteratively until the next-best alignment score falls below a specified threshold (Supplemental material).

Alignments were performed between sequences in their native orientation and also by inverting one of the sequences. The scoring matrix used for alignment was an identity matrix: match = +4, mismatch = −4, match(N) = +1, Gap(open) = −4, Gap(extension) = −4. The symbol X is assigned a very negative score (-10^5) so that it can be used to mask sequences (Xs will not be aligned). Any non-A,C,G,T,X symbols were treated as N.

We define d_{SM} as the fraction of both sequences that does not contain a region of significant local alignment, without respect to order or orientation (Fig. 1). In general, d_{SM} can be thought of as the fraction of the aligned sequences that cannot be posited as homologous according to the above method. We found that upstream sequences of 500 bp and a minimum score of 48 (a combination of matches, mismatches, and gaps that sum to 48) was most predictive of expression difference between paralogs (Results), and we used these parameters for all subsequent analyses. Further details of the algorithm and its implementation can be found in the Supplemental material.

Computation of the first sequence alignment is $O(n1 \times n2)$ in memory and CPU time, where n = sequence length. Subsequent iterations remain memory-intensive but are much less CPU-intensive, as only part of the matrix needs to be recomputed. The main limitation of the SMM is its memory usage, which limits analysis to pairs of sequences typically <5kb. In comparison, other recent approaches such as the DBA method (Jareborg et al. 1999) are able to handle alignment of much larger sequences. However, the latter method and others do not detect contiguous shared regions if their order or orientation is not conserved.

C source code of the SMM software (*sharmot*) is freely available for download at: http://www.oeb.harvard.edu/faculty/wakeley/.

Expression Data

Expression data were obtained from Hill et al. (2000) in which Affymetrix oligonucleotide microarrays were used to examine

mRNA expression at eight different stages of *C. elegans* development. We calculated the absolute value of the maximum difference in absolute number of transcripts for duplicate pairs through development where valid data for at least two timepoints for both genes was available [genes called "present" at least once by Hill et al. (2000)]. Similar results were found using mean expression difference through development (data not shown). Additionally, we measured changes in relative expression by computing Pearson's *r* correlation coefficient through development for duplicate pairs with valid data for ≥4 timepoints.

Data from 553 separate cDNA microarray experiments that included different nutrient conditions, developmental stages, and mutants (Kim et al. 2001) were also used to estimate differences in relative expression between pairs of duplicate genes. For each duplicate pair, we computed Pearson's *r* correlation coefficient across experiments on normalized $\log_2$ [Cy3/Cy5] ratios for genes with valid data (>2-fold change) across >30 experiments.

ACKNOWLEDGMENTS

We thank the anonymous reviewers and all members of the Wakeley and Hartl labs, as well as Eric Coissac, Eduardo P.C. Rocha, and Isabelle Gonçalves for their suggestions; Laura Garwin for her comments on an earlier version of the manuscript, and The Sanger Institute and the Genome Sequencing Center at Washington University for providing unfinished *C. briggsae* sequence. Special thanks to the Bauer Center for Genomics Research and Gordon L. Kindlmann at the University of Utah Scientific Computing and Imaging Institute for computational resources. G.A. was funded by La Fondation pour la Recherche Médicale.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Blencowe, B.J. 2000. Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25:** 106–110.

Castillo-Davis, C.I. and Hartl, D.L. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol. Biol. Evol.* **19:** 728–735.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Conant, G.C. and Wagner, A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13:** 2052–2058.

Dayhoff, M.O., Eck, R.V., and Park, C.M. 1972. A model of evolutionary change in protein sequences. In *Atlas of protein sequence and structure*, pp. 89–99, National Biomedical Research Foundation, Washington, D.C.

Dunn, K.A., Bielawski, J.P., and Yang, Z. 2001. Substitution rates in *Drosophila* nuclear genes: Implications for translational selection. *Genetics* **157:** 295–305.

Duret, L. and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci.* **96:** 4482–4487.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725–736.

Gu, Z., Nicolae, D., Lu, H.H., and Li, W.H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* **18:** 609–613.

GuhaThakurta, D., Palomar, L., Stormo, G.D., Tedesco, P., Johnson, T.E., Walker, D.W., Lithgow, G., Kim, S., and Link, C.D. 2002. Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Res.* **12:** 701–712.

Haldane, J.B.S. 1932. *The causes of evolution*. Longmans and Green, London.

Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G., and Brown, E.L. 2000. Genomic analysis of gene expression in *C. elegans*. *Science* **290:** 809–812.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. R Soc. Lond. B Biol. Sci.* **256:** 119–124.

Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.

Kellis, M., Birren, B.W., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617–624.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087–2092.

Koch, M.A., Weisshaar, B., Kroymann, J., Haubold, B., and Mitchell-Olds, T. 2001. Comparative genomics and regulatory evolution: Conservation and function of the Chs and Apetala3 promoters. *Mol. Biol. Evol.* **18:** 1882–1891.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biol.* **3:** research0008.0001–0008.0009.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564–567.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Nembaware, V., Crum, K., Kelso, J., and Seoighe, C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse–human orthologs. *Genome Res.* **12:** 1370–1376.

Nielsen, R. and Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Heidelberg.

Ohta, T. 1987. Simulating evolution by gene duplication. *Genetics* **115:** 207–213.

Shaw, P.J., Wratten, N.S., McGregor, A.P., and Dover, G.A. 2002. Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. *Evol. Dev.* **4:** 265–277.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

True, J.R. and Haag, E.S. 2001. Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.* **3:** 109–119.

Wagner, A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci.* **97:** 6579–6584.

Walsh, J.B. 1995. How often do duplicated genes evolve new functions? *Genetics* **139:** 421–428.

Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225–228.

Waterman, M.S. and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* **197:** 723–728.

Wolfe, K.H. and Li, W.H. 2003. Molecular evolution meets the genomics revolution. *Nat. Genet.* **33:** 255–265.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

———. 2000. Phylogenetic Analysis by Maximum Likelihood (PAML), version 3.0, http://abacus.gene.ucl.ac.uk/software/paml.html

WEB SITE REFERENCES

http://wormbase.org; Wormbase.

http://www.oeb.harvard.edu/faculty/wakeley/; C source code of the SMM software (*sharmot*).

Received April 7, 2004; accepted in revised form June 2, 2004.

A Robust Measure of HIV-1 Population Turnover Within Chronically Infected Individuals

G. Achaz, S. Palmer,† M. Kearney,† F. Maldarelli,† J. W. Mellors,‡ J. M. Coffin,† and J. Wakeley**

*Department of Organismic and Evolutionary Biology, Harvard University; †HIV Drug Resistance Program, NCI, NIH, Frederick, Maryland; and ‡Department of Infectious Diseases, University of Pittsburgh

A simple nonparameteric test for population structure was applied to temporally spaced samples of HIV-1 sequences from the *gag-pol* region within two chronically infected individuals. The results show that temporal structure can be detected for samples separated by about 22 months or more. The performance of the method, which was originally proposed to detect geographic structure, was tested for temporally spaced samples using neutral coalescent simulations. Simulations showed that the method is robust to variation in samples sizes and mutation rates, to the presence/absence of recombination, and that the power to detect temporal structure is high. By comparing levels of temporal structure in simulations to the levels observed in real data, we estimate the effective intra-individual population size of HIV-1 to be between 10^3 and 10^4 viruses, which is in agreement with some previous estimates. Using this estimate and a simple measure of sequence diversity, we estimate an effective neutral mutation rate of about 5×10^{-6} per site per generation in the *gag-pol* region. The definition and interpretation of estimates of such "effective" population parameters are discussed.

Introduction

There are at least two levels at which studies of HIV population genetics can be undertaken. The first is at a global level and considers dynamics of the virus in the whole population of infected individuals (Grassly, Harvey, and Holmes 1999). Even more broadly, this might include the whole immunodeficiency virus family (Mindell 1996). The second level, at a smaller scale, focuses on viral populations within an infected individual. The latter represents the intra-host, or intra-individual, population level and is the focus of the present study.

During HIV infection, changes in viral population size are typically characterized by three phases (Coffin 1999). For several weeks after the infection, the first phase is marked by an extensive increase in viral load, associated with a decrease in the CD4+ cells. That phase ends, in chronically infected individuals, when the immune system reacts, leading to a decrease in the viral load of up to two orders of magnitude (Daar et al. 1991). The third and last phase is characterized by a slow increase of the viral load. This phase usually ends when the virus infection overwhelms the individual's immune system, causing immunodeficiency, illness, and death. In some individuals, namely the long-term nonprogressors, this third phase is extended and the viral load can remain relatively low for decades.

In the present study, we analyzed ~1100 base pairs (bp) of the *gag-pol* region of HIV-1 sampled at different time points in two chronically infected patients. We propose a standard measure for analyzing the temporal structure in HIV-1 populations, which is based on a test for geographic population structure that was originally proposed by Hudson, Boos, and Kaplan (1992). The test compares the mean number of pairwise differences between sequences within each population to a theoretical distribution obtained by randomly shuffling the sequence labels. We adapt this straightforward test to the case of temporal structure between two samples of viral sequences taken from the same individual at two different times. We also use the test statistic as a measure of the amount of population turnover.

The evolution of a virus within a host has been shown to be strongly influenced by its environment. Some individuals are overwhelmed by the infection within a few years and others are able to resist disease progression for long periods of time. Two well-known examples of such intra-host environmental constraints are the genotype of the host, e.g., at histocompatibility and coreceptor loci, which can induce selective changes in viral genotype (Moore et al. 2002), and the application of therapeutic drugs, which leads to the emergence of predictable well-characterized drug resistant strains (Shankarappa 1999). Such observations underscore the importance of selection on intra-host evolutionary processes. Because the total population size of HIV in the third phase has been estimated to be very large, on the order of 10^7 to 10^8 infected cells and about 10^{10} individual viruses (Piatak et al. 1993; Haase et al. 1996), it may be appropriate to consider the dynamics of intra-host HIV genetic variation to be deterministic (Coffin 1995), as if the population size were infinite.

It has also been reported that, even if drugs do induce predictable mutations conferring a resistant phenotype, the frequency and timing of the fixation of these mutations is highly variable from one individual to another (Leigh Brown and Richman 1997). This observation has been interpreted as a consequence of random variation of the frequencies of resistant strains preexisting before the start of drug therapy, together with deterministic selection. Other possible explanations for differences among individuals include host-virus genotype interactions that affect either the fitness of resistance mutations or the mutation rate of the virus, and variation in the time for the mutation that confers resistance to appear. In any case, variation in

Key words: intra-host HIV evolution, effective population size, chronically infected individual.

E-mail: gachaz@oeb.harvard.edu.

Mol. Biol. Evol. 21(10):1902–1912. 2004
doi:10.1093/molbev/msh196
Advance Access publication June 23, 2004

the timing of events among individuals would seem to imply an important role for stochasticity, or random genetic drift, during infection.

To reconcile a very large population size with a nonnegligible effect of random genetic drift, it has been proposed that the effective population size (N_e) of viruses inside each individual is several orders of magnitude smaller than the real population size. The effective size is defined as the corresponding size of a hypothetical neutral idealized population (i.e., described by the standard Wright-Fisher model) that gives the same amount of genetic drift observed in the real population. Many factors are known to make the effective size very different from the real size, including selection, population structure, and fluctuating population size. All methods of estimating N_e assume that observed genetic variation is neutral. Indeed, the very concept of an effective population size is based on this notion. In this respect, it is interesting to note that, in a population of infinite size, a locus evolving under directional selection can drive the turnover at a partially linked neutral locus, mimicking genetic drift (Gillespie 2000). The neutral locus changes by a process similar to hitchhiking (Maynard Smith and Haigh 1974; Kaplan, Hudson, and Langley 1989). This "pseudohitchhiking" model (Gillespie 2000) shows that a reduced effective size at a genetic locus can be caused by selection even in an infinite population.

In studies of HIV-1, several estimates of the intra-host N_e have been obtained by analyzing polymorphisms in the *env* region (Leigh Brown 1997; Rodrigo et al. 1999; Shriner et al. 2004). These studies support a relatively small effective population size, on the order of 10^3 to 10^4. An alternative method to estimate N_e, using linkage disequilibrium between polymorphic sites, suggested that N_e is about 10^6 (Rouzine and Coffin 1999). However, a recent reanalysis of the same data suggests that this higher value was due to a bias in the analyzed polymorphisms (Shriner et al. 2004) and that, after correction, this value is on the order of 10^3. All of these estimates of N_e are much smaller than the actual population size of the virus within an infected person, which again may be 10^{10}, and imply an important role for genetic drift in the dynamics of genetic variation. A recent study that modeled resistance to Lamivudine (3TC) argued that even with N_e on the order of 10^6, random drift may still play an important role in *env* region (Frost et al. 2000). Gillespie's (2000) pseudohitchhiking model may help to reconcile these results, since strong selective effects have been observed in the *env* region (Nielsen and Yang 1998; Richman et al. 2003).

We show that temporally spaced samples (often referred to as serial samples) within two chronically infected individuals can be distinguished using the test mentioned above. In addition, we determine the power and size of the test using standard neutral coalescent simulations. A number of previous methods, reviewed in Drummond et al. (2003), have been developed to estimate the mutation rate and the population size from serial samples (Drummond and Rodrigo 2000; Rambaut 2000; Drummond, Forsberg, and Rodrigo 2001; Drummond et al. 2002). All these methods assume that there is no recombination, and they rely on the existence of a single simple coalescent history or genealogy for all sites in the locus. It is not known how such methods will perform when there is recombination. In contrast, the method we propose here does not rely on a common genealogy for all sites, and simulations show that it performs similarly well whether recombination is rampant or completely absent.

Using the neutral coalescent simulations of serial samples, we describe the rate of change of the test statistic in an evolving population both with and without recombination. This allows the estimation of the effective population size in one of the two patients by a comparison between expected and observed rates of change in the test statistic. Using a parametric bootstrap, i.e., by repeatedly simulating samples and applying the method to them, we can give confidence intervals on these estimates. We show that the intervals for no recombination and for free recombination are closely overlapping. We also calculate an effective mutation rate, which reflects the neutral mutation rate of these sequences.

Material and Methods

Origin and Analysis of the Sequences

Plasma samples were obtained at different times from two untreated individuals with well-established HIV-1 infection. DNA sequences from about 20–50 individual viral genomes were obtained for each sample using single genome RT-PCR sequence (SGS) analysis of approximately 1,098 bp, including the p6 region of gag, protease, and the first 900 nucleotides of RT (see more details on the method in S. Palmer et al. [in preparation]). Summary statistics of the sequences we used are described in table 1.

For each sample, sequences were aligned together by using ClustalW (Thompson, Higgins, and Gibson 1994) with default parameters. All alignments were visually inspected and frameshifts were removed using the sequence editor SEAVIEW (Galtier, Gouy, and Gautier 1996). The gap character was considered a fifth symbol in calculating pairwise differences between the sequences.

Testing for Population Subdivision

We implemented the series of tests for population subdivision described by Hudson, Boos, and Kaplan (1992). The tests were originally proposed to detect associations between genetic structure and geographic structure. However, the design of the tests, in which a matrix of pairwise sequence differences is calculated from the data then randomly permuted to assess the significance of structure, is quite general and nonparametric, so it is easily extended to other situations. Hudson, Boos, and Kaplan (1992) investigated two measures of subdivision, called K_s and K_s^*, defined below, and showed in simulations that the test using K_s^* had more power to detect geographic structure. Let n_1 be the number of sequences in the first sample and n_2 be the number of sequences in the second sample.

K_i is the mean number of differences between pairs of sequences in sample i. K_s is defined as $K_s = w_1K_1 + w_2K_2$, where $w_1 = n_1/(n_1 + n_2)$ and $w_2 = 1 - w_1$.

Table 1
Characteristics of Sequences from Patients A and B

		Date	Days	Number of Sequences	Viral Load (RNA/ml)	K_i (estimated Θ)
Patient A (1,098 sites)	First positive test	1991	—	—	—	—
	Sample 1	11/19/1998	0	16	27,252	8.88
	Sample 2	12/15/1998	26	22	18,604	5.78
	Sample 3	04/20/1999	142	7	24,648	9.52
	Sample 4	08/26/1999	280	17	22,164	8.78
	Sample 5	03/01/2000	468	13	136,760	10.95
	Sample 6	05/26/2000	554	42	10,520	11.31
	Sample 7	06/22/2000	561	15	11,934	9.31
	Sample 8	07/11/2000	600	10	19,285	8.19
	Sample 9	07/12/2000	601	7	15,362	9.95
	Sample 10	07/13/2000	602	7	16,018	12.95
	Sample 11	07/14/2000	603	16	16,446	11.09
	Sample 12	07/15/2000	604	18	16,419	9.88
	Sample 13	07/16/2000	605	16	16,904	9.89
	Sample 14	07/17/2000	606	13	20,392	12.09
	Sample 15	07/18/2000	607	9	18,918	12.00
	Sample 16	07/19/2000	608	14	24,855	11.09
	Sample 17	07/20/2000	609	17	21,600	11.16
	Sample 18	12/27/2001	1,134	21	21,760	11.83
	Sample 19	11/19/2002	1,461	53	30,111	11.00
Patient B (1,313 sites)	First positive test	July 2000		—	—	—
	Sample 1	07/23/2001	0	53	19,783	17.38
	Sample 2	01/07/2002	168	30	3,996	17.26
	Sample 3	01/07/2002	357	5	3,263	16.60

K_i^* is defined as $K_i^* = \sum_{a=1}^{n_i-1} \sum_{b=a+1}^{n_i} \log(1 + D_{ab})/\binom{n_i}{2}$, where D_{ab} is the number of differences between sequence a and sequence b of sample i, and $K_s^* = w_1K_1^* + w_2K_2^*$. Hudson, Boos, and Kaplan (1992) suggest that optimal weights for K_s^* are $w_1 = (n_1 - 2)/(n_1 + n_2 - 4)$ and $w_2 = 1 - w_1$.

This test generates a P value for the probability that the level of structure between two samples of sequences is due simply to chance. To do this, the sequences are randomly relabeled ("population 1" and "population 2") a large number of times, holding n_1 and n_2 constant, and the statistics are computed for each such permutation. Except where specified below, we used 10,000 relabelings/permutations to obtain P values. The P value of the observed statistic is equal to the fraction of times the value for the permuted data is less than or equal to the observed value. This procedure detects patterns of genetic structure in which pairwise differences within samples tend to be smaller than pairwise differences between samples. If the P value is less than the nominal level of significance, which we denote α, the null hypothesis of no structure is rejected.

Coalescent Simulations

We simulated samples of sequences to estimate the size of the test (i.e., the validity of the significance level α) and the power of the test to detect temporal structure as a function of the time between samples. We also used simulations to investigate how the average P value changes with the time between samples. Simulations followed the standard coalescent methods (see, e.g., Hudson 1990), in which a genealogy is constructed and then a Poisson-distributed number of neutral mutations is randomly placed on the genealogy. We assumed that each mutation gave rise to a unique polymorphic site (the infinite-sites mutation model). For each mutation, a branch is chosen randomly in proportion to its length and every descendent of that branch inherits the mutation. We allowed two different possibilities for recombination in this infinite-sites mutation model, either (1) no recombination occurred (Waterson 1975), or (2) recombination occurred freely between all pairs of sites (Kimura 1969). The results below are based on 10,000 simulation replicates for each set of parameters.

To build a genealogy, we first chose a sample size for the first and the second time points, respectively. We then used a standard neutral coalescence process (Kingman 1982*a*, 1982*b*; Tajima 1983), which depends on the neutral mutation parameter $\Theta = 2N_e\mu$ (where μ is the neutral mutation rate per sequence per generation) and on both sample sizes. We simulate the history of the second sample back to the time when the first sample was taken. The number of coalescent events during this part of the history depends on the time t_{2-1} (in number of generations) between the two samples. As usual in the coalescent, this time is rescaled so that the unit of measurement is N generations: $T_{2-1} = (t_{2-1})/N$ (where N is the population size). The expected number of mutations along a single lineage over this time period is equal to $\mu t_{2-1} = T_{2-1} \times \Theta/2$. When the simulation reaches time point 1, the sequences from sample 1 are added to the ancestral lineage(s) remaining from sample 2. The coalescent process continues until the most recent common ancestor of both samples is reached. A realization of such genealogy is shown in figure 1. This approach is identical to the way in which the coalescent process has previously been applied to HIV evolution (Rodrigo and Felsenstein 1999; Rodrigo et al. 1999). In the same manner, it is straightforward to include samples from more than two time points.

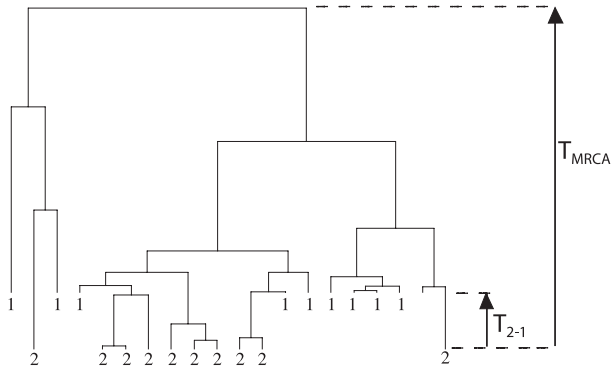


FIG. 1.—Coalescence of two time series samples. An example of a simulated neutral standard coalescent genealogy of two samples from the same population but separated by a defined time interval. In this case, we used $n_1 = 10$ sequences in the first sample (1) and $n_2 = 10$ in the second sample (2). We also used a value of $\Theta(2N\mu)$ of 10 and choose a time interval (T_{2-1}) of 0.4 (in number of N generations); one should note that the expected time to the most recent common ancestor is less than 2 and the expected time for two lineages (like the last two ones) to coalesce is 1.

Estimation of the Effective Population Size with Confidence Intervals

By matching the results of simulations with the results for the data, it is possible to estimate the effective population size N_e of HIV within a patient. There are a variety of ways this might be done. Here, we first estimate of the number of HIV generations between a pair of data samples such that there is a 50% chance of rejecting the null hypothesis at the $\alpha = 0.05$ level. Then, we equate this to the scaled time of separation in simulations that gives 50% power to reject the null hypothesis at the same $\alpha = 0.05$ significance level, and we solve for N_e. The value of 50% power was chosen based on preliminary simulations so we could use linear interpolation between different times of separation (on a log scale) without serious error. In contrast to the simulations in which the time of separation can be controlled, the sampling times for the data samples are fixed (table 1). Therefore, to estimate the 50%-power separation time for the data, we ordered the 171 sample pairs by time of separation then used a sliding window of 20 paired sample points to search (with the aid of interpolation) for the separation time that gave 10/20 rejections of the null hypothesis. We used the mean time of separation among the 20 points in the window as the estimate of the separation time.

We used a parametric bootstrap procedure to obtain confidence intervals for our estimate of N_e. Specifically, we assumed that the true values of N_e and Θ were those we estimated from the data and then simulated 10^4 genealogies of the 19 samples with separation times (table 1) rescaled by N_e, assuming either no recombination or free recombination. For each set of simulated sequences, we performed the test on all pairwise comparisons between samples and estimated N_e exactly as we did for the actual data. The upper and lower 2.5% cutoffs for these simulated distributions of estimates of N_e are taken as the 95% confidence interval.

Results

Because we were interested in finding a useful standard measure to compare population change through time within an individual, we investigated different sample sizes and times of separation of the serial samples in simulations. The coalescent process we used to create pseudosamples of sequences depended on four parameters: the sample sizes n_1 and n_2, the population mutation rate Θ, and the scaled time interval T_{2-1} between two samples. In most of what follows, we discuss results where Θ is equal to 10. This value is very close to the one we estimated for the data from patient A (see table 1) using average pairwise differences. Smaller and larger values for Θ, specifically $\Theta = 1$ and $\Theta = 100$, gave essentially the same results for all analyses and will be discussed later. We applied the test of Hudson, Boos, and Kaplan (1992) to each set of pseudosequences from the simulations to assess whether temporal structure could be detected. We examined the performance of both K_s and K_s^*.

Size of the Test

By setting T_{2-1} to 0, we create two sets of sequences sampled from a single time point. This case represents the null model of no temporal structure and can be used to measure the size of the test (i.e., frequency of false positive outcomes). To do this, we counted the number of times we would reject the null model, at the 5% significance level, for each set of parameter values. This addresses the concern that arises from the fact that an unknown genealogical structure exists and shapes genetic variation in the sample and that small or unbalanced samples ($n_1 < n_2$ or $n_1 > n_2$) might lead spurious rejections of the null hypothesis. Thus, we investigated different values of n_1 and n_2.

Results show that K_s^* is largely insensitive to sample size and to asymmetry of sample sizes from the two time points (fig. 2). The performance of K_s does depend on sample size when n_1 is small and n_2 is large. However, the direction of deviation is conservative (lower than expected chance of rejecting the null hypothesis when it is true). This result shows that these two measures exhibit the expected fraction (or fewer) of false positives even with small samples sizes. The test based on K_s^* was recommended by Hudson, Boos, and Kaplan (1992) because it was more sensitive for detecting geographical structure. In our simulations we found the same tendency and thus present only results from tests using K_s^*. However, using K_s results in only a subtle decrease of the power of the test.

Power to Detect Temporal Structure in a Neutrally Evolving Population

If we set the time interval between the two samples to a nonzero value, we simulate a sampling process at two different time points. To assess the power of the test, we chose a range of times. Again, as in a standard coalescent process, the time of separation T_{2-1} is scaled by the population size $T_{2-1} = (t_{2-1})/N$, where t_{2-1} is the number of generations between the second and the first sample. To assess the effect of recombination we created two series of artificial sequences. In the first series (no recombination; fig. 3*a*), all sites are so tightly linked that they always share the same genealogy. In the second one (free recombination; fig. 3*b*), all sites are segregating independently so that

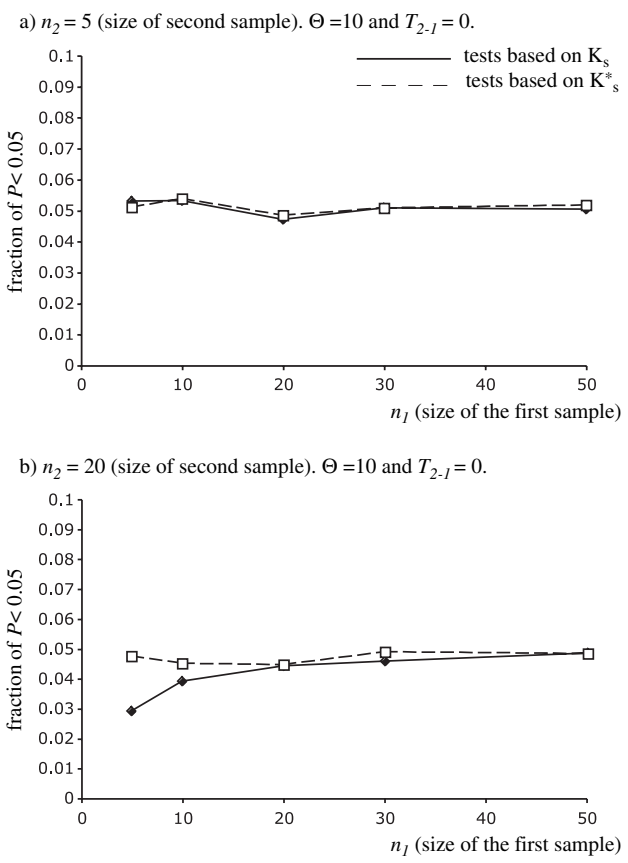


FIG. 2.—Size of the test using four different measures. The frequencies at which the null hypothesis is rejected (at an alpha risk of 5%) are plotted as a function of the first sample size. Since the time between the early and the late sample is $T_{2-1} = 0$, one would expect not to observe more than 5% of false positives. The coalescent parameters used here are $\Theta = 10$. In (*a*), the size of the late sample n_2 is set to 5 and the size of the early sample n_1 varies from 5 to 50. In (*b*), n_2 is set to 20 and n_1 varies from 5 to 50.

each site has its own genealogy. We fixed the sample size from each time point to be equal to 20.

The results show that, for $\Theta = 10$, the null hypothesis that the two samples come from the same time point is rejected at greater than 5% frequency if the scaled time of separation between the two samples is larger than ~0.01. They also show that after one scaled time unit (i.e., at $T_{2-1} = 1$ or greater), the samples are essentially always distinguishable from each other. This reflects the fact that many coalescent events will have occurred between the members of the later sample over this amount of time. In fact, from equations 6.1 and 6.2 in Tavaré (1984), the probability that there are more than three ancestral lineages of the later sample remaining at the time ($T_{2-1} = 1$) of the earlier sample is less than 0.05. Comparison of figure 3*a* to figure 3*b* shows that recombination increases the power of the test slightly, although mostly just in the vicinity of $T_{2-1} = 0.1$.

Analyses of other values of Θ show that for smaller values (i.e., $\Theta = 1$) the power of the test decreases and the effect of recombination almost disappears (fig. 3). In contrast, for higher values (i.e., $\Theta = 100$), the power of the test without recombination does not change but the effect of recombination is stronger (it increases the power of the test by an order of magnitude in T_{2-1}).

Application to HIV-1 *gag-pol* Sequences

These results show that the Hudson, Boos, and Kaplan (1992) test of subdivision provides a standard measure of the population structure through time, which can be used to tackle biological questions concerning the timing of population turnover. This can be done either with or without recombination in the sequences. We can now use this test to analyze the extent of intra-host HIV-1 population evolution. To do so, we used sequences sampled from two chronically infected individuals, here called A and B for reference, picked at different time points long after the primary infection (see table 1) and spaced by different time intervals.

As a visual test for structure, we reconstructed trees relating the samples, which are genealogies under the assumption that no recombination had occurred. This was done using the neighbor-joining method (Saitou and Nei 1987) with a Kimura two-parameter distance correction (Kimura 1980). An example tree of samples from two relatively distant time points from individual A is shown in figure 4. Although the tree does appear to show some structure, the significance of this structure is difficult to assess, both because it is just a visual comparison and because the assumption of no recombination is likely to be wrong. Recombination invalidates the usual interpretation that the branches of the tree represent ancestral lineages. Using the test based on K^*_s (or the one based on K_s) leads to rejection of the null hypothesis of no structure for these samples ($P < 0.003$). It is possible that the very shape of the tree implies recombination, because under complete linkage the expected tree should have long internal branches and short external ones (i.e., see fig. 1). As figure 3 shows, the presence of recombination only increases the power of the test, so the null hypothesis is rejected in either case for these samples.

We analyzed samples from 19 time points in individual A and the three time points in individual B. We made all 171 possible pairwise comparisons between samples from different time points in A and between the single pair of time points in B. The results, which are shown in figure 5, indicate that the test systematically rejects the null hypothesis (no population structure over time) after about 666 days, or 22 months. Compared to the very rapid adaptation that can sometimes be observed (e.g., in response to drug therapy; Shankarappa 1999), turnover of the HIV-1 population in these chronically infected individuals appears to be relatively slow, taking more than one year to be detectable at the 5% level using this test.

Estimation of the Effective Population Size

It is difficult to draw conclusions from the few comparisons we have for individual B. However, assuming neutrality of the observed mutations, it is possible to roughly estimate the effective population size of HIV-1 for individual A. The effective size N_e is defined as the

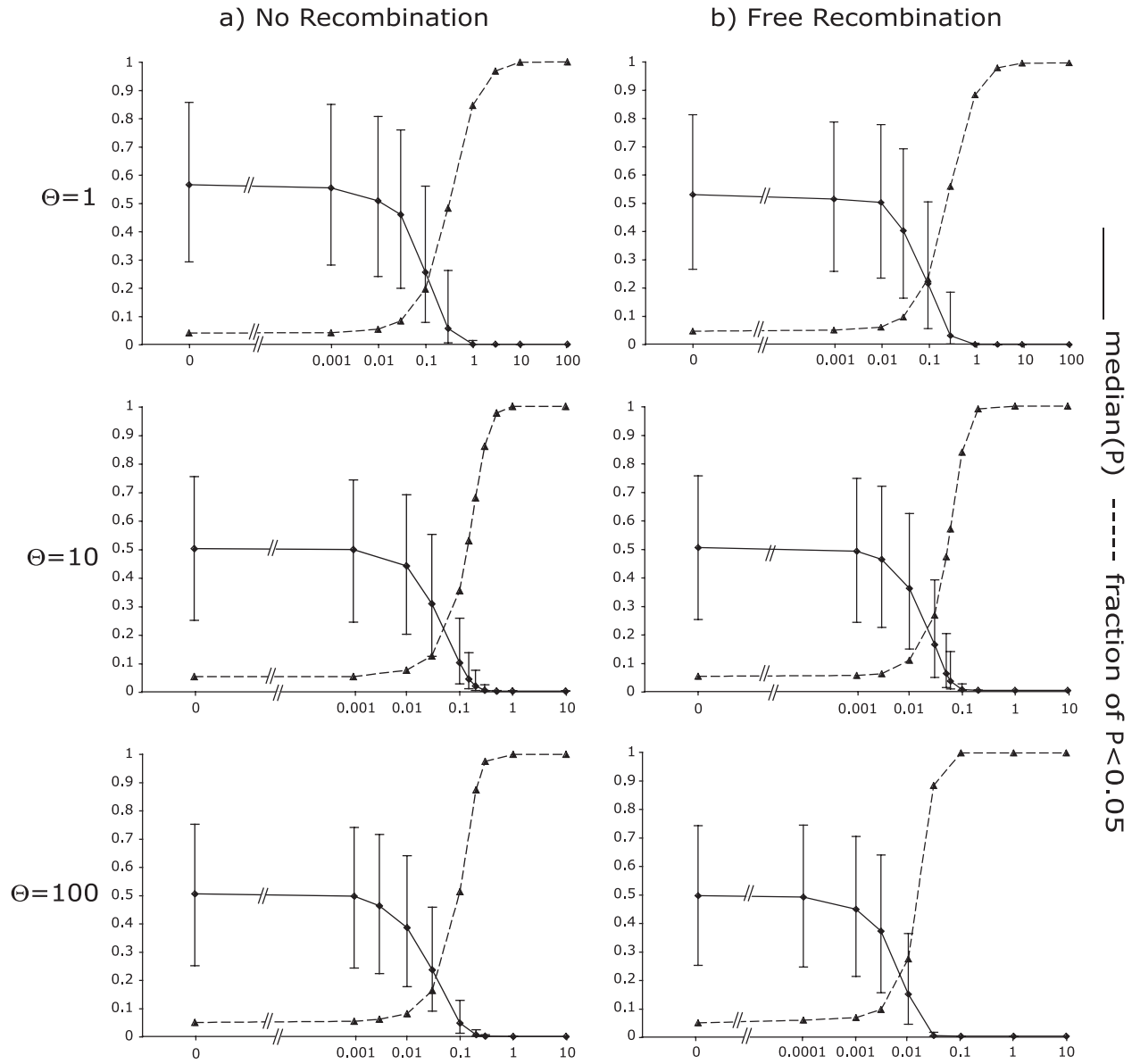


FIG. 3.—Temporal structure in neutrally evolving populations. Diamonds with continuous lines represent the median probability P (estimated by using K_s^*) and its first and third quartiles as a function of the rescaled number of generations (T_{2-1}) that separates the late from the early sample. Triangles and dashed lines represent the frequencies where the null hypothesis is rejected (at an alpha risk of 0.05) by the test as a function of T_{2-1}. We used the following set of parameters: $\Theta = 1$, 10, or 100 and $n_1 = n_2 = 20$. In (*a*) all sites share a single genealogy: there is no recombination and all sites are in complete linkage. In (*b*) each site has its own genealogy: all sites segregate independently.

population size in our simulations that gives the same amount of population change over time (by neutral drift alone) as the one observed in the data. To do so, we compared the expected rate of change for a neutrally evolving population shown in figure 3 to the corresponding data observed in figure 5.

We estimate the scaled time interval for which the test null hypothesis is rejected for half of the paired samples (see *Materials and Methods*). Assuming no recombination (fig. 3*a* with $\Theta = 10$), this time is estimated to be $T_{2-1} = 0.142$. In figure 5*b*, this time corresponds to 223 days. The generation time of HIV-1 in vivo has been estimated to be about 1.5 days (Rodrigo et al. 1999; Fu 2001; Seo et al. 2002; Markowitz et al. 2003). Thus, with the definition of the scaled time $T_{2-1} = (d_{2-1})/(1.5 \times N_e)$, where d_{2-1} is the number of days between the samples, we have $N_e = (d_{2-1})/(1.5 \times T_{2-1})$. This leads to an estimate of N_e equal to 1,047. Finally, by simulating 10^4 genealogies of the 19 samples and by using all pairwise comparisons (see *Materials and Methods*), we estimated the 95% confidence interval to be 445–2,655 under the assumption of no recombination. Assuming free recombination, but otherwise identical methods, we estimated N_e to be equal to 3,026 with a 95% confidence interval of 864–4,955.

We compared these estimates of N_e to that obtained from another method on the data from patient A. We used a recently proposed method that employs Monte Carlo simulations to estimate the likelihood of different

population sizes in a Wright-Fisher model (Anderson, Williamson, and Thompson 2000). The method uses changes in allele frequencies between two time points and tries to fit the real data to the expectation of a simulated time series sampling process of a neutral population. Since the method assumes that no mutations occur between the time points, we used frequency information from sites that are polymorphic in all samples of A. Note that this might lead to an upward bias because we have excluded some of the more extreme changes in allele frequency. We used the 11 samples with 15 sequences or more (see table 1) to increase the number of shared polymorphic sites to seven. This calculation gave an estimated N_e of about 2,800 haploid genomes, which is within the range we estimated using K_s^*.

Estimation of an Effective Mutation Rate

Based on our estimate of N_e, we can estimate an effective mutation rate per generation for the whole sequences corresponding to these data (1,098 sites). We have estimated $\Theta = 2N_e\mu$ to be about 10 using a method (average pairwise differences) that is unbiased under the assumptions of infinite-sites mutation and selective neutrality (Tajima 1983). Thus, μ is interpreted as the effective neutral mutation rate per sequence per generation. Then, $\Theta = 10$ translates into an estimate of $\mu = 10/(2 \times 1{,}047) = 4.8 \times 10^{-3}$, using the estimate of N_e obtained assuming no recombination. This gives a mutation rate per site per generation of 4.35×10^{-6} ($= 4.8 \times 10^{-3}$ /1,098). Using the 95 % confidence interval, we can compute a confidence interval for our estimation of the effective per site mutation rate that ranges from 1.7×10^{-7} to 1.0×10^{-5}. If we assume free recombination, the estimation of the effective mutation rate is 1.5×10^{-7} and the associated confidence interval ranges from 9.2×10^{-7} to 5.2×10^{-6}.

To validate our rough estimation, we used another approach to estimate this effective mutation rate. Fu (2001) developed a framework to estimate the mutation rate per day using multiple samples spaced by time interval. As with ours, this method uses the mean pairwise differences within and between the time points. The estimated effective mutation rate per day found by this method is 0.0058. Assuming a generation time of 1.5 days, we calculate a μ of 0.0024 for the whole sequence (about 1,100 sites) and, thus, a mutation rate per site per generation of 2.18×10^{-6}. This is in good agreement with our estimate based on K_s^*.

Finally, we used an alternate method that estimates both the effective population size and the effective mutation rate. This method reconstructs the likely genealogies (under the assumption of no recombination) using Bayesian statistical inference and Markov Chain Monte Carlo integration (Drummond et al. 2002). A recent version of this strategy was implemented by Drummond and Rambaut (2003) in the program Beast (available at http://evolve.zoo.ox.ac.uk/beast/). Using an HKY model for mutation (with default parameters), we ran Beast on all our sequences. The chain appeared converged after 1.65×10^7 replicates and exhibited ESS (effective sample size) values above 100 (minimum ESS values recommended by the authors). As the estimated values given by Beast were per day, we rescaled them to compare them to our estimations per generation. This gives an effective population size of 9.1×10^3 and a 95% confidence interval ranging from 7.4×10^3 to 1.1×10^4. For the mutation rate, it gives an estimate of 1.9×10^{-5}, with a 95% confidence interval ranging from 1.5×10^{-5} to 2.3×10^{-5}. These estimates are larger than ours, but the estimate of N_e is still much smaller than the actual population size. Our methods and those of Drummond et al. (2002) differ in two significant ways: (1) we assume infinite-sites mutation in estimating Θ whereas Drummond et al. (2002) allow for multiple mutations, and (2) we examine both no recombination and free recombination (and show that our estimates are fairly robust) whereas Drummond et al. (2002) assume no recombination (and account for deviations from this in the data with multiple mutations). Presumably, the differences between the methods explain the differences in parameter estimates, including the fact that Beast gives an inferred value of Θ equal to $2 \times 1{,}098 \times 1.9 \times 10^{-5} \times 9.1 \times 10^3 = 380$, compared to our $\Theta = 10$.

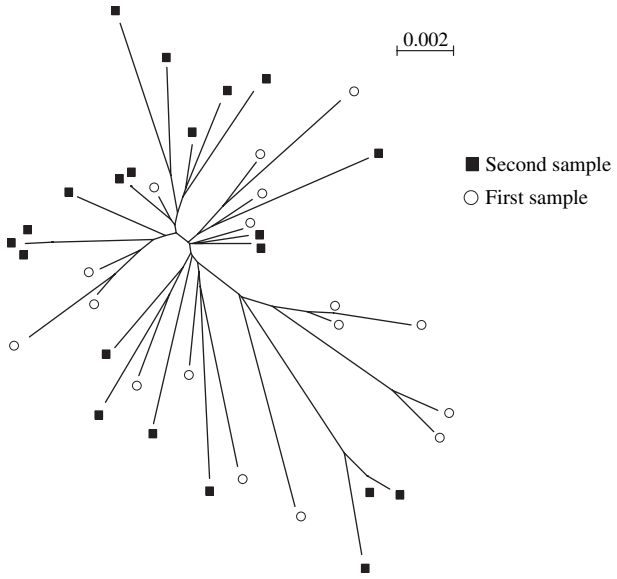


FIG. 4.—Phylogenetic trees of two samples of A. A simple neighbor-joining phylogenetic reconstruction of two samples from individual A. In table 1, the first sample is the "sample 17" and the second one is the "sample 18." These two samples are spaced by 525 days. Using the subdivision test (with 10^5 random labelings for the test), we obtain a probability of $P < 0.003$ that the two samples are picked from the same time point.

Discussion

A Standard Measure for the Rate of Population Change

Although temporal structure in HIV-1 sequences from the region analyzed (comprising ~1,100 bases from the P6 region of *gag* through *pro* and most of the RT region) from a chronically infected individual is not apparent (visually) in genealogical trees reconstructed from the data, we have found that such a structure can be detected if the interval between two samples is about 22 months or more. For this purpose, we used a test that was originally proposed to detect geographic structure

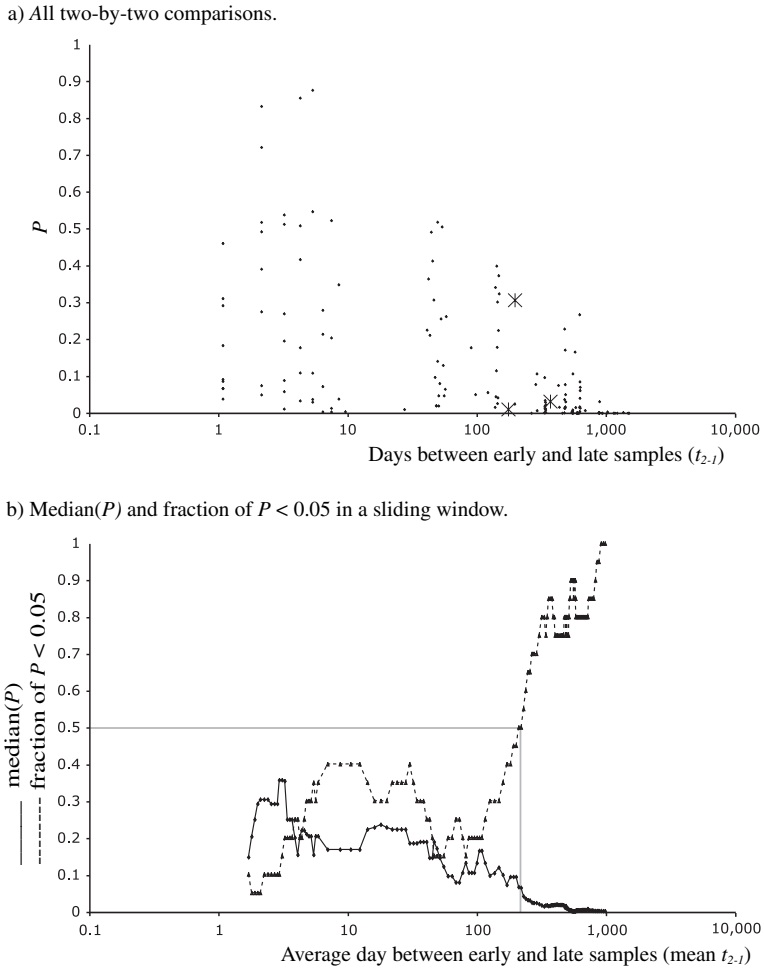


FIG. 5.—Temporal structure in real data. The probability P (estimated by using K_s^*) that two samples could be sampled from the same population is depicted as a function of the number of days between the samples. We used here all 171 possible two-by-two comparisons between the 19 samples of patient A and three comparisons between the three samples of patient B (see table 1). All values of P are plotted as dots for patient A and as stars for patient B. All tests were done using with 10^5 random labelings. In (*a*) all points are shown, whereas in (b) the median P value (diamonds with continuous lines) as well as the frequency where the null hypothesis is rejected, at an alpha risk of 5% (triangles with dashed lines), are given by the average day in a sliding window of 20 data points.

(Hudson, Boos, and Kaplan 1992) as a standard measure of temporal structure. This standard measure could be used to compare the rate of evolution of other populations under various conditions. For example, it might be interesting to measure the rate of evolution of HIV-1 population in a host treated with antiretroviral drugs.

An important assumption of the simulations above is that $\Theta = 2N_e\mu$ is constant over time. Note that this is not an assumption of the nonparametric test we have applied, but it is an important part of using the same statistics to estimate effective size. Changes in the diversity over time would tend to increase the power of the test because: (1) sequences sampled from a population with reduced diversity would tend to cluster together, and (2) bottlenecks between time points would increase the rate of population turnover. Although the mutation rate is contained in Θ, it is doubtful that the mutation rate would change over the times separating these samples. Interestingly, measures of sample sequence diversity and the independent viral load counts (table 1) do not show dramatic changes between time points, at least not in individual A. In individual B, there was a change in viral load but the diversity seems to be almost unaffected by it.

Another phenomenon that would influence the power of the test is recurrent mutations. It has been reported that the G→A and T→C mutations occur at higher frequencies than others (Mansky and Temin 1995). In the sequences of patient A, G/A and T/C pairs represents ~55% and ~30% of all polymorphisms, respectively. It is possible that multiple transition mutations have occurred at these sites, rendering some mutations unobservable. Indeed, we observed more sites that are polymorphic in both of the two patients than expected by chance (data not shown). This complication probably erodes the power of the tests since multiple, unobserved mutations would more likely affect pairwise comparisons between time points (K_{12}) than pairwise comparisons within time points (K_1 and K_2). It could lead to relatively larger K_s or K_s^* than if the infinite-sites mutation model was correct. This would then decrease the power of the test.

A Small Effective Population Size

The definition of effective population size is the size of an idealized population, exactly the population model of our simulations, which would give an equivalent rate of genetic drift as the one observed in the population under study. The "rate of genetic drift" can be defined in a variety of ways (Ewens 1979), leading to slightly different estimates of N_e. We estimated the expected rate of change of a neutrally evolving population by measuring the temporal structure of simulated samples. The comparison of this expected rate of change with the one we observed for real HIV-1 populations within a chronically infected individual leads to an estimate of its effective size of roughly 10^3 to 10^4. Like all previous estimates, this result is several orders of magnitude smaller than the actual count of replicating virus, which may be as high as 10^{10}. Most deviations from the idealized model—in fact all of them except some kinds of population structure—give values of N_e that are smaller than the actual population size. Thus, our observations are consistent with many possible causes.

The most obvious deviation from the null model, with regard to HIV-1 intra-host evolution, concerns the neutrality of the observed variation. Clearly HIV-1 is under tremendous selective pressure during an infection. In these data, evidence that selection is operating can be seen in the ratio of the rate nonsynonymous (d_N) to synonymous (d_S) mutations, which is estimated using Nei and Gojobori's method (Nei and Gojobori 1986) with Jukes and Cantor distance (Jukes and Cantor 1969) between 0.05 and 0.01 for the samples listed in table 1. A ratio of d_N/d_S of 1 is expected under neutrality and a ratio smaller than 1 under a purifying selection regime. These results then suggest that the sequences are under a regime of strong purifying selection for protein structure and function. In addition, there may be selection on synonymous sites for translation efficiency, as has been observed in other organisms (Duret and Mouchiroud 1999). There might also be selection for RNA structures in both nonsynonymous and synonymous sites. Selection on synonymous sites would tend to increase the ratio of d_N/d_S. This would reduce d_S and then imply that purifying selection for amino acid replacement is stronger than if synonymous sites where merely neutral. Interestingly, our estimation of the effective size, computed using the *gag-pol* region, is very similar to the estimates computed with the *env* gene. This suggests that even though there is evidence that the selection regime is very likely to be different in those two regions (*gag-pol* being mostly under purifying selection [see above] where *env* is subject to positive selection [Nielsen and Yang 1998; Richman et al. 2003]), other force(s), yet uncharacterized, constrain HIV populations all along their genomes.

A second possible explanation could be that HIV-1 populations do not evolve under panmixia, but rather with some population structure that causes a reduction in effective size. There is evidence that HIV populations are spatially structured because resistant strains can be in different frequencies in different organs (Epstein et al. 1991) or even in different cell types (Potter, Dwyer, and Saksena 2003). It has been suggested recently that HIV populations within patients might exhibit metapopulation structure (Frost et al. 2001), in which local populations of the virus become extinct and are recolonized by propagules from other local populations. It is well known that such patterns of colonization and replacement can reduce N_e dramatically (Slatkin 1977; Whitlock and Barton 1997; Rousset 2003; Wakeley 2004). Note that changes in population size over time are another possible cause of small N_e, but the metapopulation model we apply includes such changes, so we do not consider them as a separate force.

For illustration, we consider both metapopulation dynamics and natural selection as possible causes of the reduced intra-individual effective size of HIV. Prior estimates of intra-host N_e for HIV range from about 10^3 to 10^4 (Leigh Brown 1997; Rodrigo et al. 1999; Rouzine and Coffin 1999; Shriner et al. 2004). The estimates we made here are also in this range, although at the lower end of it. These estimates of N_e cover a broad range, but they are all much smaller than the actual intra-host population size of infected cells, which can be up to 10^{10} (Piatak et al. 1993; Haase et al. 1996). Thus, very roughly, there is between a 10^6-fold and 10^7-fold reduction in HIV effective population size that needs to be explained. Using either the standard metapopulation model (Slatkin 1977) or Gillespie's (2000) "pseudohitchhiking" model, it is possible to make a theoretical prediction for this ratio.

A general model of a metapopulation includes two kinds of dispersal: (1) regular migration and (2) recolonization after extinction (Slatkin 1977). For simplicity, we will assume that there is no regular migration among subpopulations (here cells) and, further, that extinct subpopulations are recolonized by single virus particles. In this case, the ratio of the effective size (N_e) of the population to the total size (N_T) of the population is given by $N_e/N_T = (1 - e_0)/\{N_L[1 - (1 - e_0)^2]\}$, in which e_0 is the proportion of local populations that go extinct and are recolonized every generation and N_L is the size of each local population (see Rousset [2003] or Wakeley [2004]). If we adopt a metapopulation model in which each infected cell is a local population, and we assume N_L to be about 100 viruses (Haase et al. 1996), we infer a high rate of extinction, or turnover, of local populations. In particular, the fraction of cells $(1 - e_0)$ that do contribute to the future intra-host population of HIV is between 10^{-4} to 10^{-5}.

In the pseudohitchhiking model, when no crossover is assumed, the ratio of the effective intra-host population size of HIV to the total intra-host population size is given by $N_e/N_T = 1/(1 + 2N_T\rho)$, where N_T is the total size of the population, which is assumed to be panmictic, and ρ is the per-generation probability of a selective sweep (Gillespie 2000). In this case, using $N_T = 10^{10}$, the rate of sweeps ρ ranges from 5×10^{-4} to 5×10^{-5}. The violation of the no-recombination assumption would lead to higher ρ values, depending on the frequency of crossover events. The high per-site, per-replication mutation rate of HIV, about 3.4×10^{-5} (Mansky and Temin 1995), might appear to violate the Poisson-process assumption of the pseudohitchhiking model. However, if two or more particular mutations are required for selective benefits or if only a small minority of

sites have the potential to drive a selective sweep, then this assumption might be reasonable. In contrast to the case of a metapopulation, under the pseudohitchhiking model even a small per-generation probability of a selective sweep can explain a large reduction in N_e, because N_T is so large and this appears in the denominator of the ratio. Clearly, selection and metapopulation dynamics are just two possibilities to consider, and even these are not mutually exclusive. It seems likely that a combination of factors act together to reduce the intra-host effective population size of HIV-1.

Acknowledgments

We thank all members of the Wakeley lab for their useful advice and their friendly support. We thank Cristian Castillo-Davis, Rob Kulathinal, Richard Watson, Igor Rouzine, and Daniel Shriner as well as the two anonymous reviewers for their constructive comments on the manuscript. We also thank Andrew Rambaut for his generous advice about keeping the Beast running. G.A. was funded by "La Fondation pour le Recherche Médicale." This work was supported by a Presidential Early Career Award for Scientists and Engineers from the NSF (DEB-013760) to J.W. and by a grant from the NIH (R01-CA089441) to J.M.C. J.M.C. was a Research Professor of the American Cancer Society.

Literature Cited

Anderson, E. C., E. G. Williamson, and E. A. Thompson. 2000. Monte Carlo evaluation of the likelihood for $N(e)$ from temporally spaced samples. Genetics **156**:2109–2118.

Coffin, J. M. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science **267**:483–489.

———. 1999. Molecular Biology of HIV. Pp. 3–40 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore, M.D.

Daar, E. S., T. Moudgil, R. D. Meyer, and D. D. Ho. 1991. Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection. New England J. Med. **324**:961–964.

Drummond, A., R. Forsberg, and A. G. Rodrigo. 2001. The inference of stepwise changes in substitution rates using serial sequence samples. Mol Biol. Evol. **18**:1365–1371.

Drummond, A., O. G. Pybus, A. Rambaut, R. Forsberg, and A. G. Rodrigo. 2003. Measurably evolving populations. Trends Ecol. Evol. **18**:481–487.

Drummond, A., and A. G. Rodrigo. 2000. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol. Biol. Evol. **17**:1807–1815.

Drummond, A. J., and A. Rambaut. 2003. BEAST v.1.0. available from http://evolve.zoo.ox.ac.uk/beast/.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**:1307–1320.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

Epstein, L. G., C. Kuiken, B. M. Blumberg, S. Hartman, L. R. Sharer, M. Clement, and J. Goudsmit. 1991. HIV-1 V3 domain variation in brain and spleen of children with AIDS: tissue-specific evolution within host-determined quasispecies. Virology **180**:583–590.

Ewens, W. J. 1979. Discrete stochastic models. Effective population size. Pp. 104–112 *in* K. Krickeberg and S. A. Levin, eds. Mathematical population genetics. Springer-Verlag, Berlin.

Frost, S. D., M. J. Dumaurier, S. Wain-Hobson, and A. J. Leigh Brown. 2001. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. Proc. Natl. Acad. Sci. USA **98**:6975–6980.

Frost, S. D., M. Nijhuis, R. Schuurman, C. A. Boucher, and A. J. Leigh Brown. 2000. Evolution of lamivudine resistance in human immunodeficiency virus type 1-infected individuals: the relative roles of drift and selection. J. Virol. **74**: 6262–6268.

Fu, Y. X. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol. Biol. Evol. **18**:620–626.

Galtier, N., M. Gouy, and C. Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**: 543–548.

Gillespie, J. H. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. Genetics **155**:909–919.

Grassly, N. C., P. H. Harvey, and E. C. Holmes. 1999. Population dynamics of HIV-1 inferred from gene sequences. Genetics **151**:427–438.

Haase, A. T., K. Henry, M. Zupancic et al. (14 co-authors). 1996. Quantitative image analysis of HIV-1 infection in lymphoid tissue. Science **274**:985–989.

Hudson, R. R. 1990. Gene genealogy and the coalescent process. Oxford Surv. Evol. Biol. **7**:1–44.

Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9**:138–151.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. Munro, ed. Mammalian protein metabolism III. Academic Press, New York.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisited. Genetics **123**:887–899.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**:893–903.

———. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kingman, J. F. C. 1982*a*. The coalescent. Stochastic processes and their applications. **13**:235–248.

———. 1982*b*. On the genealogy of large population. J. Appl. Prob. **19A**:27–43.

Leigh Brown, A. J. 1997. Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA **94**:1862–1865.

Leigh Brown, A. J., and D. D. Richman. 1997. HIV-1: gambling on the evolution of drug resistance? Nat. Med. **3**:268–271.

Mansky, L. M., and H. M. Temin. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. **69**:5087–5094.

Markowitz, M., M. Louie, A. Hurley, E. Sun, M. Di Mascio, A. S. Perelson, and D. D. Ho. 2003. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. J. Virol. **77**:5037–5038.

Maynard Smith, J., and J. Haigh. 1974. The hitchhiking effect of a favorable gene. Genet. Res. **23**:23–35.

Mindell, D. P. 1996. Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. Proc. Natl. Acad. Sci. USA **93**:3284–3288.

Moore, C. B., M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A. Mallal. 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science **296**:1439–1443.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

Piatak, M., Jr., M. S. Saag, L. C. Yang, S. J. Clark, J. C. Kappes, K. C. Luk, B. H. Hahn, G. M. Shaw, and J. D. Lifson. 1993. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. Science **259**:1749–1754.

Potter, S. J., D. E. Dwyer, and N. K. Saksena. 2003. Differential cellular distribution of HIV-1 drug resistance in vivo: evidence for infection of CD8+ T cells during HAART. Virology **305**:339–352.

Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics **16**:395–399.

Richman, D. D., T. Wrin, S. J. Little, and C. J. Petropoulos. 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc. Natl. Acad. Sci. USA **100**:4144–4149.

Rodrigo, A. G., and J. Felsenstein. 1999. Coalescent approaches to HIV population genetics. Pp. 233–272 *in* K. A. Crandall, ed. The evolution of HIV. Johns Hopkins University Press, Baltimore, M.D.

Rodrigo, A. G., E. G. Shpaer, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins. 1999. Coalescent estimates of HIV-1 generation time in vivo. Proc. Natl. Acad. Sci. USA **96**:2187–2191.

Rousset, F. 2003. Effective size in simple metapopulation models. Heredity **91**:107–111.

Rouzine, I. M., and J. M. Coffin. 1999. Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proc. Natl. Acad. Sci. USA **96**:10758–10763.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Seo, T. K., J. L. Thorne, M. Hasegawa, and H. Kishino. 2002. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics **160**:1283–1293.

Shankarappa, R. 1999. Evolution of HIV-1 resistance to antiviral agents. Pp. 3–40 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore, M.D.

Shriner, D., R. Shankarappa, M. A. Jensen, D. C. Nickle, J. E. Mittler, J. B. Margolick, and J. I. Mullins. 2004. Influence of random genetic drift on human immunodeficiency virus type 1 *env* evolution during chronic infection. Genetics **166**:1155–1164.

Slatkin, M. 1977. Gene flow and genetic drift in a species subject to frequent local extinctions. Theor. Popul. Biol. **12**:253–262.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26**:119–164.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Wakeley, J. 2004. Metapopulation models for historical inference. Mol Ecol **13**:865–875.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:256–276.

Whitlock, M. C., and N. H. Barton. 1997. The effective size of a subdivided population. Genetics **146**:427–441.

Edward Holmes, Associate Editor

Accepted June 15, 2004

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 1 2007, pages 119–121
doi:10.1093/bioinformatics/btl519

Genome analysis

Repseek, a tool to retrieve approximate repeats from large DNA sequences

Guillaume Achaz[1,2,*], Frédéric Boyer[3], Eduardo P. C. Rocha[1,4], Alain Viari[3] and Eric Coissac[3,5]

[1]Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris 6, 12, rue Cuvier, 75005 Paris, France, [2]UMR 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie-Paris 6, Bâtiment A, 7, quai St Bernard, 75252 Paris Cedex 05, France, [3]INRIA-Rhône Alpes projet HELIX, 655, avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France, [4]Unité Génétique des Génomes Bactériens, Institut Pasteur, 28, rue du Dr Roux, 75724 Paris Cedex 15, France and [5]UMR 5163 LAPM, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France

Received on July 20, 2006; revised on September 12, 2006; accepted on October 6, 2006
Advance Access publication October 11, 2006
Associate Editor: John Quackenbush

ABSTRACT

Summary: Chromosomes or other long DNA sequences contain many highly similar repeated sub-sequences. While there are efficient methods for detecting strict repeats or detecting already characterized repeats, there is no software available for detecting approximate repeats in large DNA sequences allowing for weighted substitutions and indels in a coherent statistical framework. Here, we present an implementation of a two-steps method (seed detection followed by their extension) that detects those approximate repeats. Our method is computationally efficient enough to handle large sequences and is flexible enough to account for influencing factors, such as sequence-composition biases both at the seed detection and alignment levels.

Availability: http://wwwabi.snv.jussieu.fr/public/RepSeek/

Contact: achaz@abi.snv.jussieu.fr, http://www.repetmasker.org

INTRODUCTION

The importance of genome redundancy has been strongly emphasized in the field of genome dynamics and evolution as well as in medical biology. A repeat is a sequence present twice or more with a high degree of similarity within a larger sequence (e.g. a chromosome) or set of sequences (e.g. a genome with several chromosomes). Each instance of the repeated sub-sequence is called a 'copy' of the repeat. Repseek aims at detecting as many as possible pairs of copies within or between large DNA sequences. Unlike RepeatMasker (Smit *et al*., 2004), we do not search for already well characterized repeated elements but instead we retrieve all repeated sequences without any a priori on the nature of the repeats. Furthermore, we do not construct families of repeats, which is the objective of multiple seeds extension (Price *et al*., 2005) or of clustering algorithms (Bao and Eddy, 2002; Pevzner *et al*., 2004), though our program can be used to feed the clustering algorithms. The detection of repeats is not a trivial problem and there is no satisfactory methodology available apart from recursive local alignment (using dynamic programming) of sequences with themselves (Waterman and Eggert, 1987). Such algorithms, however, are quadratic in computation time and memory and cannot be used for large sequences. Our approach, like most current methods to detect similarity in large sequences (Altschul *et al*., 1997; Vincens *et al*., 1998), works around the problem through a two-step strategy (Fig. 1). First, it detects seeds (strict repeats, i.e. repeats with neither indels nor substitutions) and, then it extends them into larger approximate repeats. The statistical evaluation of the repeats can be undertaken on seeds length or on repeats score (setting $L_{\min}$ and/or $S_{\min}$ parameters). Starting with longer seeds is faster but increases the chance to miss degenerate repeats. Both statistics can be used for the detection of repeats within a single sequence or between two sequences.

ALGORITHM

Several efficient algorithms are already available for computing the seeds (Abouelhoda *et al*., 2002; Kurtz and Schleiermacher, 1999) (see user's guide for a complete comparison). Repseek can accept as input a list of seeds; however, for simplicty, it also provides an exact builtin seeds detection algorithm, based on the KMR algorithm (Karp *et al*., 1972) that proves to be very efficient in practice. One of the main advantage of KMR in this context is that it can be implemented in a memory efficient way. Our current implementation requires $9n$ bytes direct repeats (where n is the sequence length) and $17n$ bytes for inverted repeats. All pairs of seeds are then extended on both sides, accepting substitutions or indels, by using a dynamic programming approach (Smith and Waterman, 1981). The edit matrix is filled as in the classical local alignment procedure, but the optimal path is anchored at the seeds extremities and ends up at a maximum of the matrix. To reduce the time and memory requirements, we use a heuristic similar to the one introduced in BLAST2 (Altschul *et al*., 1997). At the end, if more than one repeat share the same localization, only the one with the highest score is kept (users can tune how much overlap is required to do so). The use of a simple

*To whom correspondence should be addressed.

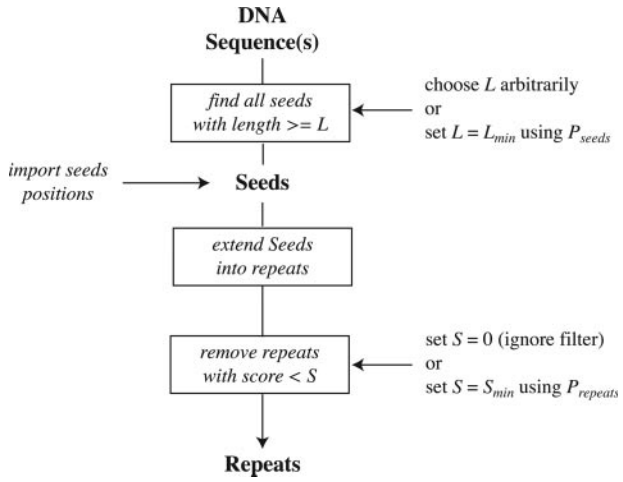


Fig. 1. Schematic workflow of repseek.

identity substitution matrix can create biases for sequences where the relative frequency of each nucleotide is not 1/4 (Achaz *et al*., 2003). In highly biased sequences, this results in longer alignments composed of the most abundant nucleotides. To fix this potential problem, repseek uses a matrix based on nucleotide frequencies that can correct for biases in sequence composition. The score for a match or a mismatch is scaled by the log of the product of the corresponding nucleotide frequencies. Other programs, such as reputer (Kurtz and Schleiermacher 1999) or RepeatScout (Price *et al*., 2005) also handle mismatches; though, to the best of our knowledge, no published program accepts indels or corrects for composition bias by using an adapted substitution matrix.

STATISTICS

Repseek proposes two statistics (P_{seeds} or P_{repeats}) to evaluate analytically the significance of a repeat. P_{seeds} is usually expressed as the probability $P(L_{\text{longest}-\text{seed}} \geq L)$ that the longest seed observed in a random sequence of same size and nucleotide composition is longer than L (Karlin and Ost 1985). Reciprocally, by imposing a statistical threshold, one can calculate the smallest length $L_{\min}$ above which no such seed is expected to occur by chance in a random sequence. An equivalent statistics is available for the analysis of seeds between two sequences. P_{repeats} is the probability $P(S_{\text{best}-\text{repeat}} \geq S)$ that the score of the best local alignment observed between two random sequences of size n and m is larger than a given score S. This probability can be well approximated by $P = e^{-\gamma mne^{tS}}$ (Karlin and Altschul, 1993). We evaluated the unknown parameters γ and t using the method proposed by (Waterman and Vingron 1994) for a range of sequence lengths (1 kb, 10 kb, 100 kb and 1 Mb) and compositions ($d_{GC} = |GC\% - 50|$ ranging from 0 to 35 by step of 5%). This was done by randomizing 10 000 random sequences for each combination of length and composition and using a least-square regression to estimate both parameters. Hence, we can associate a chosen probability with a minimum score $S_{\min}$ above which no repeats are expected to be found in a random sequence of same size and same composition.

PERFORMANCE

Repseek's memory and time consumption are typically small enough to handle large DNA sequences. On a G4 MacOSX, it takes 1 min with $L = 24$, $S = 0$ (i.e. $P_{\text{seed}} = 10^{-3}$) and around 3 min with $L = 16$, $S = 31.01$ (i.e. $P_{\text{repeats}} = 10^{-3}$) to retrieve all repeats from the genome of *Escherichia coli* (4.6 Mb). It takes 49 min or 5 h (depending on the chosen statistics) to detect all repeats on the chromosome V of *Caenorhabditis elegans* (20 Mb). Memory consumption is maximum at the seed detection step and is 80 Mb for the genome of *E.coli* and 359 Mb for the chromosome V of *C.elegans*. This shows that repseek can be potentially used to detect repeats on very large sequences on modern computers.

We performed a comparison of the repeated elements annotated by RepeatMasker in each *C.elegans* chromosome with the ones detected by repseek (using $P_{\text{repeats}} = 10^{-3}$, i.e. $L_{\min} = 17$ or 18 and $34.34 \leq S_{\min} \leq 34.94$ depending on the chromosome). Results shows that, on average, the sequence is composed at 12% of repeats detected by both Repeat-Masker and repseek, at 15% of repeats detected by repseek only and at 1% of repeats detected by Repeat-Masker only. This shows that, not only repseek retrieves almost all characterized repeats annotated by Repeat-Masker, but it also unravels a lot of yet uncharacterized repeated sequences. Interestingly, these later repeats are not only located in exons (i.e. gene duplicates), but span mostly intronic and intergenic regions.

Repseek is a fast and handy software that can detect approximate repeats in large chromosomes. The statistical pertinence of the detected repeats is evaluated considering the length and composition of the analyzed sequence. The C sources as well as a more-detailled user's guide can be found at the URL given above. Sources are publicly available and users are more than welcome to make improvements that will be incorporated in forthcoming releases.

ACKNOWLEDGEMENTS

The authors thank A. Platt and J. Pothier. G.A. was funded by La Fondation Singer-Polignac. The authors acknowledge the support of IMPBIO grant to EVOLREP. Funding to pay the Open Access publication charges for this article was provided by INRIA.

Conflict of Interest: none declared.

REFERENCES

Abouelhoda,M.I., Kurtz,S. and Ohlebusch,E. (2002) The enhanced suffix array and its applications to genome analysis. In *Proceedings of the Second Workshop on Algorithms in Bioinformatics*. Lecture Notes in Computer Science 2452, Springer-Verlag, pp. 449–463.

Achaz,G. *et al*. (2003) Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics*, **164**, 1279–1289.

Altschul,S.F. *et al*. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bao,Z. and Eddy,S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*., **12**, 1269–1276.

Karlin,S. and Altschul,S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA*, **90**, 5873–5877.

Karlin,S. and Ost,F. (1985) Maximal segmental match length among random sequences from a finite alphabet. In Cam,L.M.L. and Olshen,R.A. (eds), *Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*. Association for Computing Machinery, NY, Vol. **1**, pp. 225–243.

Karp,R.M., Miller,R.E. and Rosenberg,A.L. (1972) Rapid identification of repeated patterns in strings, trees and array. In *4th annual ACM symposium theory of computing,* ACM, pp. 125–136.

Kurtz,S. and Schleiermacher,C. (1999) Reputer: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.

Pevzner,P.A. *et al.* (2004) De novo repeat classification and fragment assembly. *Genome Res.*, **14**, 1786–1796.

Price,A.L., Jones,N.C. and Pevzner,P.A. (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics*, **21**, i351–i358.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Smit,A.F.A, Hubley,R. and Green,P. (1996–2004) , RepeatMasker Open-3.0.

Vincens,P. *et al.* (1998) A strategy for finding regions of similarity in complete genome sequences. *Bioinformatics*, **14**, 715–725.

Waterman,M.S. and Eggert,M. (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.*, **197**, 723–728.

Waterman,M.S. and Vingron,M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl Acad. Sci. USA*, **91**, 4625–4628.

Hypermutability of Genes in *Homo sapiens* Due to the Hosting of Long Mono-SSR

Etienne Loire,†‡§||¶#** *Françoise Praz,†††‡‡ Dominique Higuet,§||¶# Pierre Netter,*†‡ and Guillaume Achaz§||¶#***

*Université Pierre et Marie Curie-Paris 6, Unité Mixte de recherche (UMR) 7592, Institut Jacques Monod, Paris, France; †Centre National de la Recherche Scientifique (CNRS), UMR 7592, Institut Jacques Monod, Paris, France; ‡Université Denis Diderot-Paris7, UMR 7592, Institut Jacques Monod, Paris, France; §Université Pierre et Marie Curie Paris 6, UMR 7138, Systématique, Adaptation, Evolution, Paris, France; ||CNRS, UMR 7138, Systématique, Adaptation Evolution, Paris, France; ¶Museum National d'Histoire Naturelle, UMR 7138, Systématique Adaptation Evolution, Paris, France; #Institut National de al Sauté et de la Recherche Médicale, UMR 7138, Systématique, Adaptation Evolution, Paris, France; **Université Pierre et Marie Curie-Paris 6, Atelier de Bioinformatique, Paris, France; ††Université Pierre et Marie Curie-Paris 6, UMR_S 893, CdR Saint-Antoine, Paris, France; and ‡‡INSERM, UMR_S 893, CdR Saint-Antoine, Paris, France

Simple sequence repeats (SSRs) are very common short repeats in eukaryotic genomes. "Long" SSRs are considered "hypermutable" sequences because they exhibit a high rate of expansion and contraction. Because they are potentially deleterious, long SSRs tend to be uncommon in coding sequences. However, several genes contain long SSRs in their exonic sequences. Here, we identify 1,291 human genes that host a mononucleotide SSR long enough to be prone to expansion or contraction, being called hypermutable hereafter. On the basis of Gene Ontology annotations, we show that only a restricted number of functions are overrepresented among those hypermutable genes including cell cycle and maintenance of DNA integrity. Using a probabilistic model, we show that genes involved in these functions are expected to host long SSRs because they tend to be long and/or are biased in nucleotide composition. Finally, we show that for almost all functions we observe fewer hypermutable sequences than expected under a neutral model. There are however interesting exceptions, for example, genes involved in protein and RNA transport, as well as meiosis and mismatch repair functions that have as many hypermutable genes as expected under neutrality. Conversely, there are functions (e.g., collagen-related genes) where hypermutable genes are more often avoided than in other functions. Our results show that, even though several functions harbor unusually long SSR in their exons, long SSRs are deleterious sequences in almost all functions and are removed by purifying selection. The strength of this purifying selection however greatly varies from function to function. We discuss possible explanations for this intriguing result.

Introduction

Microsatellites or simple sequence repeats (SSRs) are arrays of DNA with short motifs—1–6 nt—repeated in tandem (Tautz 1994). SSRs are ubiquitous in all genomes explored so far and are especially abundant in eukaryote genomes (Toth et al. 2000). Strikingly, the number and the sizes of SSRs in genomes are typically much larger than expected from simple substitution models (Pupko and Graur 1999). This overabundance of SSRs is, most likely, a consequence of their specific mutational properties; these repeats are prone to expansion and contraction through polymerase strand slippage (Levinson and Gutman 1987) and, to a lesser extent, to recombination (Li et al. 2002). For strand slippage, after the replication fork has run, template and neosynthesized strands can be reannealed with the slippage of one (or more) motifs. If the "mismatch repair" (MMR) complex does not correct the resulting loop, a subsequent round of replication changes the number of repeated units by a specific amount. This translates into an insertion or a deletion of one (or more) motifs in the SSR.

Various factors have been shown to modulate the rate of SSR expansion/contraction, although their relative strength varies from species to species (Toth et al. 2000). It appears that, for "long" SSRs, contraction prevails over expansion (Xu et al. 2000). This bias in favor of contraction, along with a higher chance of being interrupted by a substitution for even longer SSRs, prevents their infinite growth (Kruglyak et al. 1998; Ellegren 2000; Dieringer and Schlotterer 2003). The nature of the motif itself also greatly modulates the mutation rate. For example, GC-rich SSRs are more unstable than others (Sagher et al. 1999; Gragg et al. 2002), and long motifs are more stable than short ones (Rose and Falush 1998; Legendre et al. 2007). Overall, the relative role of all these factors makes difficult to predict a mutation rate for a given SSR. However, it remains true that the SSR mutation rate is typically several orders of magnitude higher than the average substitution rate (Drake et al. 1998).

A number of studies in a wide range of organisms have attempted to delimit characteristics of SSRs that are predictive of the variability of the repeats. They concluded that the number of repeats was among the strongest predictors of the slippage probability during replication (Rose and Falush 1998; Lai and Sun 2003b; Legendre et al. 2007; Kelkar et al. 2008). Repeat variability is not an all-or-nothing phenomenon but rather increases exponentially with increasing number of repeat units, as initially established in yeast (Sia et al. 1997).

At what length should SSRs be deemed "hypermutable"? Using a simple probabilistic model, Rose and Falush (1998) proposed a threshold size for slippage mutations around 8 bp for mono-, di-, and tetranucleotide SSRs. Based on a different model, similar thresholds were proposed: 9 units for mononucleotide SSRs (mono-SSRs) and 4 units for dinucleotide (8 bp) and tetranucleotide (16 bp) SSRs (Lai and Sun 2003a). Using a human/chimpanzee complete genome comparison, it appears that, in this lineage, a mononucleotide of 9 units exhibit a similar mutability than a dinucleotide of 6 units or a tetranucleotide SSR of 5 units (Kelkar et al. 2008). Alternatively, we can also infer that a mononucleotide of 8 units exhibit

Key words: microsatellites, SSR, evolution, mutability, *Homo sapiens*.

E-mail: loire@abi.snv.jussieu.fr.

Mol. Biol. Evol. 26(1):111–121. 2009
doi:10.1093/molbev/msn230
Advance Access publication October 8, 2008

a similar mutability than a dinucleotide of 5 units or a tetranucleotide SSR of 4 units.

Interestingly, similar threshold sizes for mononucleotide and dinucleotide SSRs instability were observed in vitro during polymerase chain reaction (Lai and Sun 2003a; Shinde et al. 2003). For mono-SSRs, the observation of human oncogenesis associated with microsatellite instability (MSI) also highlights 8 units as an instability threshold. MSI has been shown to underlie hereditary nonpolyposis colorectal cancer (HNPCC) (Aaltonen et al. 1993). HNPCC patients carry a germ-line mutation in one of the postreplicative MMR genes, mainly MLH1 or MSH2 (Jacob and Praz 2002; Woerner et al. 2006). Once the corresponding normal allele is lost through somatic inactivation, cells become totally devoid of MMR activity and are left with unrepaired polymerase errors that arise during replication. Rates of mutation arising in microsatellite repeats are drastically enhanced by mutations affecting postreplicative DNA MMR (Strand et al. 1993). In this context, only genes with an SSR of at least 8 nt have been reported to exhibit a significant instability (Duval and Hamelin 2003; Woerner et al. 2006; Miquel et al. 2007). Altogether, these results suggest common features of microsatellite mutation mechanisms both in vivo and in vitro with evidence of a slippage mutation threshold at around 8 or 9 units for mono-SSRs.

SSRs tend to be less common in coding sequences (Metzgar et al. 2000; Ackermann and Chao 2006) as a change in nucleotide number often has disastrous functional consequences. If the unit length is a multiple of three, there will be an expansion or a contraction of the particular amino acids encoded by the 3-mer (codon). It is well established that long expansions of such coding microsatellites are responsible for many neurodegenerative disorders (Everett and Wood 2004). When the unit length is not a multiple of three, a change in unit number produces a frameshift (Strauss 1999). If the slippage occurs during the replication process, it may create an allele that contains a premature stop codon either in somatic cells or in the germ line. Slippage can also occur during transcription (Fabre et al. 2002), leading to abnormal messenger RNA that is usually degraded by the nonsense-mediated mRNA decay system (Conti and Izaurralde 2005). Because SSRs in coding sequences are typically associated with deleterious effects, they tend to be subject to purifying selection. We want to emphasize that SSRs that have unit lengths that are not a multiple of three have a direct, harmful potential in coding sequences because no slippage can be tolerated; therefore, they should be even less common within exons. Intriguingly, it has been observed that many genes involved in DNA repair, including MMR, carry a long mono-SSR in their coding sequences (Mori et al. 2001; Miquel et al. 2007). If these particular SSR experience an expansion or a contraction, the MMR system will become deficient and will lead to a higher mutation rate (as observed in some HNPCC-associated tumors). It has been postulated that a deficient MMR system could be advantageous when the environment is stressful. In this case, organisms with a higher mutation rate could adapt more easily to environmental challenges. Consequently, mono-SSRs in these genes could have been positively selected for their mutational potential (Moxon and Wills 1999; Chang et al. 2001; Kashi and King 2006).

In the present study, we have detected all strict SSRs—that is perfect repeats without any "interruption" in the pattern—in all human genes. We used the presence of a long SSR (with at least 8 [or 9] units for mono-SSR, 5 [or 6] units for di-SSR, and 4 [or 5] units for tetra- and penta-SSRs) as a proxy for the hypermutability of genes (Rose and Falush 1998; Lai and Sun 2003b). Even though many other factors can influence the mutability of genes, the presence of a long SSR greatly increases the chances for a gene to be inactivated. Indeed, the probability of a nonsense substitution is several orders of magnitude lower than the rate of slippage of a long enough SSR. We found mono-SSRs to be the most abundant unstable SSR as well as the most biased in term of hosting genes' function. Consequently, we focused our study on mono-SSR and used the term "hypermutable genes" to refer to genes that carry a long (and therefore potentially unstable) mono-SSR in their coding sequence hereafter.

Using annotations from the Gene Ontology (GO) database (Ashburner et al. 2000), we performed an in silico functional analysis of all genes that are a priori hypermutable. We found a cohesive restricted subset of functions that are overrepresented among hypermutable genes. To take into account differences due to the length and the composition of genes, we computed for each gene the probability to host a long mono-SSR. In this statistical framework, we observe less hypermutable genes than expected in almost all functions, including the ones we found overrepresented. This shows that, typically, hypermutable genes are removed by purifying from the human genome because of their deleterious potential. Interestingly, we observe that the strength of the purifying selection, that removes long mono-SSR, varies from function to function.

Materials and Methods

Microsatellites in Human-Coding Sequences

We extracted all exons and introns from all transcripts of the 22,218 genes from the human genome of the database Ensembl v37. Each gene sequence was then reduced to its exonic sequences only. When exons of different transcripts were overlapping, we merged them into an artificial exonic-like sequence. For each gene, we then concatenated all its nonredundant exonic-like sequences into a single sequence and inserted an "*X*" at each junction. The *X* tag ensures that no microsatellite can be detected astride two different exons. The same procedure was applied to introns to build up a unique intronic sequence for each gene. We built two sets, each composed of 22,218 artificial exonic sequences and 18,384 artificial intronic sequences derived from all transcripts.

We detected all strict SSRs (no interruption in the pattern) of a motif whose length ranges from 1 to 5.

Statistics on Mono-SSR

The following model is very similar to previous models that were used to describe the probability of observing a given SSR in sequences (de Wachter 1981).

Interestingly, the functional bias of hypermutability is only driven by mono-SSR. Therefore, the statistical

framework focused on mono-SSR exclusively. Extensions for longer motifs are given in Robin et al. (2005).

Probability of a Given Mono-SSR

We will give here an approximation of the probability to observe at least one occurrence of an X-SSR of size m^+ (m or more) in a random sequence of L independent letters of the {A, T, G, and C} alphabet. P_x will denote the probability to generate a nucleotide X in such random sequence.

Let us first note that the number of occurrences of an X-SSR of size m^+, denoted by N_x, is exactly the number of clumps of the m-mer $(X)_m$. A clump of a motif is defined here as the maximal set of overlapping occurrences of this motif in the sequence (Robin et al. 2005). The expectation of N_x is thus given by

$$E(N_X) = (1 - P_X) \times (P_X)^m \times (L - m + 1),$$

and N_x can be approximate by a Poisson random variable (Robin et al. 2005). Therefore, we have

$$P(N_X \geq 1) = 1 - P(N_X = 0),$$

where

$$P(N_X = 0) = e^{-E(N_X)}.$$

Expected Size of a Mono-SSR

We first computed, for each sequence and for each type of nucleotide, the m value that corresponds to $P(N_x \geq 1) \geq 0.5$. This value will be named $m_{1/2}$. If the model fits the data, a given gene has 50% chance of having its longest SSR larger than $m_{1/2}$. We can then affect all genes to either a "larger" or a "smaller" category depending whether its longest mono-SSR is larger or smaller than its $m_{1/2}$. Because these are independent Bernoulli trials, we expect for a set of genes that half of it should be in the larger category. We can then test if the genes tend to have a smaller/larger mono-SSR than expected using a χ^2 test.

Expected Fraction of Hypermutable Genes

We also computed the expected fraction of genes carrying a long mono-SSR (m fixed) in their coding sequences for a set of genes. To do so, we calculated, for each gene of this set, the probability of observing at least one mono-SSR of length m^+ of any type of nucleotide (with $m = 8$ or 9). We assume that the probability for each type of SSR is independent. Because m is not very small, this approximation is reasonable. In a given gene, the probability to find at least one mono-SSR of length m^+ is

$$P\left(N_{\mathrm{A,C,G,T}} \geq 1\right) = 1 - P(N_{\mathrm{A}} = 0) \times P(N_{\mathrm{C}} = 0) \times P(N_{\mathrm{G}} = 0) \times P(N_{\mathrm{T}} = 0).$$

The average of all these probabilities for a given function is an unbiased estimator of the expected fraction of hypermutable genes in this function.

Finally, using this model, we can compute the confidence interval (CI) associated with its expected fraction of hypermutable gene. To do so, one needs to compute the probability that, among N genes, each having a probability $P(N_{\mathrm{A,\ C,\ G,\ T}} \geq 1)$ to host a mono-SSR at size m^+, n genes have such an SSR. These are N independent Bernoulli trials with different probabilities of success. We estimate the probability to obtain at least n hypermutable genes for a given term by simulations. For each term, we randomly run N Bernoulli trials with respect to the individual probability of each gene. This procedure is repeated 10^5 times for a given term. The empirical distribution is then used to compute a 95% CI for a given set of genes.

Functional Group of Human Genes

We used GO (Ashburner et al. 2000) as well as Panther Ontology (Mi et al. 2005) to assign human genes to functional groups. Both databases are based on organized ontologies, a controlled vocabulary for the description of gene products. More precisely, there are constituted of terms (i.e., GO term or PantherID) that describe a "biological process" (BP), a "molecular function" (MF), or a "cellular component" (CC) (although this latter category does not exist in Panther Ontology). For all genes, we considered all available annotations. We retrieved GO terms from Ensembl (http://www.ensembl.org/biomart/martview/) and Panther IDs from the Panther database Web page (http://www.pantherdb.org/).

Here, we defined the level of a term as the number of nodes that exists between this term and the root of the graph (level 0). In the cases of multiple paths, we keep the shortest one. We decided to compare only terms lying at the same level. We used the annotated term of a gene to browse the ontologies and collect all its parental terms. For each level, we considered only genes that have at least one defined term.

Representation of Gene Functions among the Data Set

We wanted to test if any function were overrepresented among genes carrying a long SSR. For that purpose, we used a cumulative hypergeometric law (see e.g., Castillo-Davis and Hartl 2003 as suggested Rivals et al. 2007).

We perform our tests level by level to compare comparable terms. For each level of the ontologies, we performed one test per term. To correct for multiple tests, we considered that terms lying at the same level of the ontology were independent and therefore can be corrected using the Bonferroni correction. On the contrary, we considered that tests between levels were fully dependent because they use the same annotations but with different accuracies.

Results

In this study, we restricted ourselves to strict SSRs that contain no nucleotide interruption, which tend to stabilize microsatellites (Ellegren 2004) and thus lower their intrinsic mutability. For each of the 22,218 annotated genes in the

human genome, we detected all strict SSRs in concatenated exonic and intronic sequences. Because we were interested in studying genes that are susceptible to direct inactivation by SSR contraction or expansion, we excluded SSR that had a unit length that was a multiple of three.

Long SSRs in Coding Sequences Are Mononucleotide and Dinucleotide SSRs

For each gene, we identified the largest mono-, di-, tetra-, and pentanucleotide SSR (frameshifting SSR) in exonic and, when available, in intronic sequence. Because the rate of insertion/deletion grows exponentially with the number of repeat units (Tran et al. 1997; Legendre et al. 2007), the longest SSR in the exons of a given gene provides a good approximation for gene hypermutability.

Results (fig. 1) show that all types of SSRs are smaller in exons than in introns. Indeed, intronic SSRs are four times longer than exonic SSR for mono-SSR (5.8 vs. 21.9) and 2.5 times longer for pentanucleotide SSR (1.3 vs. 3.4). One could relate this observation to the purifying selection that acts against the expansion of SSR in coding sequence; however, intronic sequences are much longer than exonic sequences (i.e., on average 30 times). Both factors contribute to this difference, as it will be shown below.

As mentioned above, mono-SSRs are estimated to be unstable when they reach a length of 8 units (Rose and Falush 1998) or 9 units (Lai and Sun 2003b). If we consider a threshold of 8 units, the corresponding mutabilities are reached for di-, tetra-, and penta-SSRs for 5, 4 and 4 units, respectively. In this case, the numbers of genes, in the human genome, having an SSR longer or equal than the threshold, are 1,291 for mono-SSR (5.8% of all genes), 678 for di-SSR (3.1%), 39 for tetra-SSRs (0.2%), and 11 for penta-SSRs (<0.1%) and a total of 1,935 (8.7%) genes. Using thresholds of 9, 6, 5, and 5 units for mono-, di-, tetra-, and penta-SSRs yields to 417 for mono-SSR (1.9%), 116 for di-SSR (0.52%), 8 for tetra-SSRs (<0.1%), and 1 for penta-SSRs (<<0.1%) and a total of 475 (2.1%) genes.

If we assume that those thresholds represent the minimum numbers of units to observe instability, the SSRs that mostly participate to gene hypermutability are clearly mono-SSR and di-SSR.

Hypermutable Genes Are Overrepresented in a Restricted Subset of Functions

Using either the lower (8 units for mono-SSRs) or the higher (9 units for mono-SSRs) threshold, we define a set of genes that have, a priori, a high probability to be disrupted by a nonsense mutation due to the expansion/contraction of the SSR they host. We then searched for overrepresented terms of GO (Ashburner et al. 2000) among the set of genes.

We worked on the subset of 15,385 genes (69% of total) that had at least one term in one of the three graphs. Note that 57% of all genes have one term in BP, 63% in MF, 54% in CC, and 48% in the three. The fraction of genes with a long SSR within each subset is identical (data not shown). It is however important to note that more specific levels are made up of fewer annotated genes.

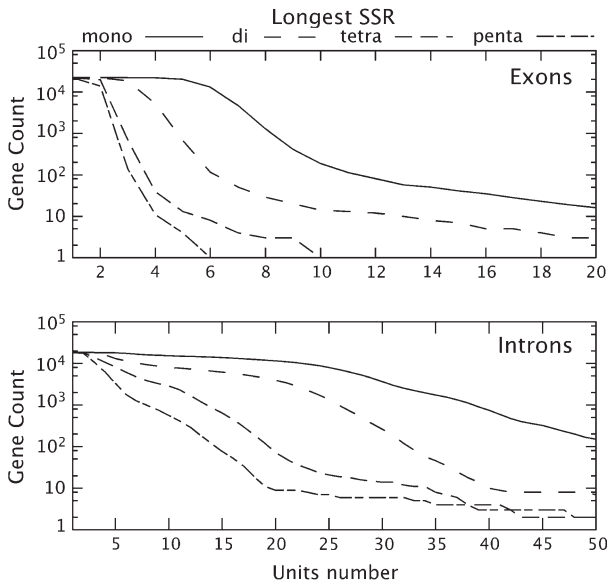


FIG. 1.—Distribution of SSR length in human genes. Counts of human genes that contain an SSR which size is equal or larger than the value given in *x* axis. The size is expressed in number of units. We only report the results for SSRs whose motif length is not a multiple of three. In the top panel, we report results for exonic sequences, whereas results for intronic sequences are displayed on the bottom panel. This figure illustrates that introns carry larger SSRs than exons do and that long SSRs in exons are mostly mono- or di-SSRs.

From all terms that were annotated at least once in the human genome (supplementary table S1, Supplementary Material online), only a few were found overrepresented. No function was overrepresented if only genes hosting a long tetra- or a penta-SSR were considered, and their removal has no impact on the results. More surprisingly, there is no function overrepresented among genes with long di-SSR, and their removal leaves the results almost unchanged (supplementary table S2, Supplementary Material online). Therefore, the only SSRs that are not uniformly distributed among functions are the mono-SSRs.

Figure 2 shows all terms we found overrepresented in hypermutable genes when mono-SSRs of 8 bp or more are considered. Results with mono-SSR of 9 bp are consistent with the former and are presented in supplementary table S2 (Supplementary Material online). Among the 3,122 BP terms, only 10 were statistically overrepresented (fig. 2*a*). Interestingly, genes with mono-SSRs are enriched for functions involved in either "cell cycle" or "response to DNA damage stimulus." Many of these hypermutable genes carry both types of annotations or related ones. The overrepresented terms are more or less precise descriptions of the same subset of functions. Following Alexa et al. (2006), if we remove the 12 genes that are annotated as functioning in meiosis (the most specific overrepresented term), no BP terms are found to be overrepresented. Therefore, genes with this function are responsible for the more general terms found to be overrepresented. Because there is no reason to believe that the most precise terms are most informative, we present results for all levels.

The same trend is observed for MF (fig. 2*b*) and CC (fig. 2*c*). In MF, out of the 2,600 terms, only 15 highly

connected terms are found overrepresented. These terms all relate to "hydrolase" (especially "ATPase"), "helicase," "GTPase regulator," and "ATP binding." Removing ATPase and GTPase regulator genes from the data set suppresses other overrepresentations in MF. As for CC, only five terms (out of 583) are overrepresented and all are related to "nucleus." The "intracellular nonmembrane-bound organelle" term encompasses intracellular molecular components such as the kinetochores, the chromosomes, and the nucleosome. Ignoring genes from nucleus does not alter the overrepresentation in intracellular nonmembrane-bound organelle and vice versa. Obviously, removing genes annotated by the latter suppresses the overrepresentations of shallower related terms.

Among Overrepresented Functions, Genes Are Longer and/or More Biased in Composition

Only a restricted number of functions are overrepresented in hypermutable genes. In the three graphs, these functions all relate to cell cycle and "DNA maintenance." We wanted to test whether genes involved in these functions have a higher chance of hosting a long mono-SSR. In this respect, we computed, for each gene, the probability of finding a long mono-SSR (8 bp or more) given its length and composition. The probability model we used here assumes that mono-SSRs are only generated by several independent substitutions that keep the average nucleotide content of the gene unchanged. It is therefore used to check whether the presence of a given mono-SSR in a given gene can be explained by random point mutations only. This model does not include the possibility of slippage for modifying the size of coding mono-SSR. Indeed, insertion or deletion of 1 or 2 units in a coding SSR whose motif length is not a multiple of three leads to a frameshift mutation. Thus, fixation of such events must be extremely rare.

The average probability of having a mono-SSR of 8 units or more in genes involved in the function we find overrepresented is 0.184, that is higher than 0.142, the average for the other annotated genes ($P << 10^{-16}$, Wilcoxon U test). This shows that, on average, genes involved in the function we found overrepresented have a higher probability to host a long mono-SSR.

Mono-SSRs Are Typically Shorter than Expected in Exons

Because this model assumes that all substitutions can occur freely with respect to the gene nucleotide composition, this model can be used as a neutral model. Indeed, this model corrects for local composition and therefore for potential local mutation biases. Furthermore, it assumes that all substitutions occur freely within the sequence, which implies the neutrality of substitutions. From the comparison of what is expected under the model to what we observe, we are able to test for the neutrality of mono-SSR.

We first tested whether the length of mono-SSR, we observe in genes, is expected under the neutral model. To do so, we computed for each gene, $m_{1/2}$, the size of the SSR that corresponds to a probability $P = 0.5$. If an SSR originates from several independent selectively neutral substitutions, half of the genes will have a mono-SSR larger than $m_{1/2}$, the other half will have a mono-SSR smaller than $m_{1/2}$. We counted all genes that were hosting a smaller or a larger SSR than $m_{1/2}$. Results are given in table 1.

In exons, we find that all types of SSRs are smaller than expected (χ^2 test; $P < 10^{-16}$), which agrees with previous studies (Metzgar et al. 2000; Ackermann and Chao 2006). For introns, we find that G-SSR and C-SSR are smaller than expected (χ^2 test; $P < 10^{-16}$), whereas A-SSR and T-SSR are longer than expected (χ^2 test; $P < 10^{-16}$). Interestingly, introns where "Alu" were removed by RepeatMasker (Smit 1999) show the same pattern. Actually, masking Alu reduces the length of intronic sequence and increases the number of sequences that host larger than expected G-SSR or C-SSRs.

There Are Less Hypermutable Genes than Expected in Almost All Functions

If we find as many hypermutable genes (i.e., genes with a mono-SSR of 8 bp or more) as the neutral model predicts, we will have to acknowledge that long mono-SSRs are virtually neutral for these genes. If we find more hypermutable genes than expected, it suggests that mono-SSRs were positively selected in these genes. Indeed, in exons, mono-SSRs are created by the accumulation of substitutions and if they improve the fitness of their host genome, they will be selected for. If we find less long mono-SSR than expected, it suggests that mono-SSRs are removed by purifying selection from the coding sequences.

In all, 1,291 genes (5.8% of the total) contain a mono-SSR of 8 units or more. Using the model, we expect 14.2% of such genes (with a 95% conservative CI of [13.8%, 14.7%]). This again highlights that, on average, there are less long mono-SSR in genes than expected by chance, most likely due to their removal by purifying selection.

We further wanted to test if this trend is shared by all functions taken individually. Therefore, we compared for each term of GO, the expected fraction of genes with long mono-SSR to the expected one. Results are shown in figure 3. Overall, among the functions that have at least 20 genes, 734/1,238 functions (59.3%) exhibit a fraction of hypermutable genes outside the 95% CI that was computed under the neutral model—406/679 (59.8%) in BP, 233/404 (57.7%) in MF, and 95/155 (61.3%) in CC. These functions are colored in blue in figure 3. For all, except one, there are less hypermutable genes than predicted by the neutral model. Taking into account also the terms with less than 20 genes, we observe a lower, though significant, number of terms outside the 95% CI: 788/6,305 terms (12.5%). This demonstrates that for almost all functions, hypermutable genes are removed by purifying selection. Considering mono-SSR of 9 bp or more (instead of 8 bp or more) leads to identical results (supplementary fig. S1, Supplementary Material online).

The functions that we found overrepresented among hypermutable genes (colored in red)—the functions given

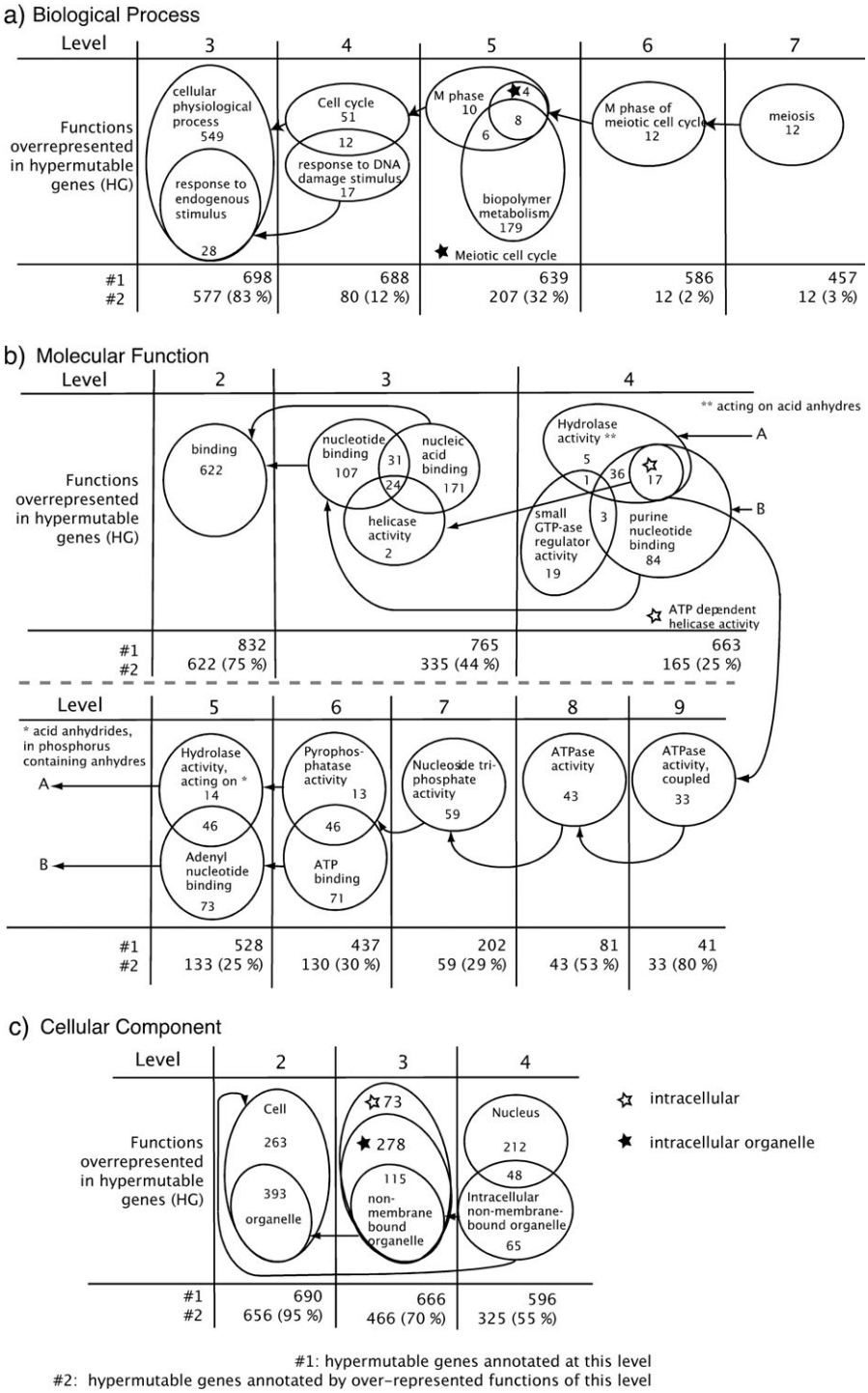


FIG. 2.—GO terms overrepresented among hypermutable genes. Here, we report for all three branches of the ontology, the functions we found overrepresented among hypermutable genes in the human genome. Results are given for (*a*) BP, (*b*) MF, and (*c*) CC. Each column is a level in the ontology (the higher the level, the more precise the annotation). It contains ellipses representing overrepresented functions lying at this level. The encapsulated numbers are the numbers of hypermutable genes in these functions. Genes shared by several functions are given in the intersection of ellipses. Arrows indicate a complete inclusion into another term at a shallower adjacent level. We also give, under the picture, the total number of hypermutable genes that is annotated at this level as well as the number among them that is embedded in the functions we found overrepresented.

in figure 2—have a larger observed fraction than the average. They, however, exhibit usually less hypermutable genes than expected from neutrality. This shows that even though we find them overrepresented, hypermutable genes are also avoided in these functions. It is noteworthy to mention that the hypergeometric statistics we used to estimate the overrepresentation among hypermutable genes depends on the number of genes within a term. Therefore, we

Table 1
Mono-SSR Probability in Human Exons and Introns

	Exons		Introns		Alu-masked Introns	
Mono-SSR	Smaller	Larger	Smaller	Larger	Smaller	Larger
A	20,271	1,947	3,441	1,4943	5,045	13,339
T	20,279	1,939	3,254	1,5130	4,458	13,926
G	21,660	558	16,059	2,325	13,651	4,733
C	21,342	976	15,139	3,245	12,484	5,900
Number expected	11,109	11,109	9,192	9,192	9,192	9,192

NOTE.—For each type of mono-SSRs (A, C, G, and T), we compute for each human gene an expected length value ($m_{1/2}$) beyond which there is a 50% chance of finding an SSR of size $m_{1/2}$ or longer. Each gene was then assigned to the larger or the smaller category depending on the comparison of the length of its longest mono-SSR to $m_{1/2}$. If the neutral model were fitting, we would expect half of the genes to host a mono-SSR larger than $m_{1/2}$. This table shows the results for exonic and intronic sequences and for each type of repeat nucleotide. We also examined intronic sequences masked for Alu sequences because their presence in an intron adds A/T repeats to these sequences. Deviation from the expectation (0.5 vs. 0.5) is significant for all types of sequences and mono-SSRs (χ^2 test, $P < 10^{-16}$ for all tests).

observed terms with a high fraction of hypermutable genes that are not significantly overrepresented (e.g., MMR with an observed fraction of 26.1%) and, conversely, terms we found significantly overrepresented even though they exhibit a moderate fraction of hypermutable genes (e.g., "biopolymer metabolism," which has an observed fraction of 7.3%). This latter case happens when the number of genes is very large for a given function, which improves the power of the statistical test we used.

Generally, the comparison between the observed and the expected fraction of hypermutable genes for all terms (fig. 3) reveals a weak though positive correlation between the observed and the expected values ($r = 0.35$ for BP, $r = 0.56$ for MF, and $r = 0.43$ for CC, $P << 10^{-4}$ for all regressions). This implies that typically the presence of long mono-SSR in genes can be partially explained by their length and their nucleotide composition.

The Strength of Purifying Selection Varies from Function to Function

Results highlight interesting functions that appear different from the others. First, we observed some functions with a particularly small observed/expected ratio. The most striking example is the "collagen" term (fig. 3*c*) for which the ratio is 0.075. Even though one would expect a large proportion (40.0%) of hypermutable genes within this term, we found only very few (3.1%). Conversely, there are 36 terms with a ratio observed/expected larger than 1 (e.g., "endoplasmic reticulum to Golgi transport" as well as meiosis and MMR in fig. 3*a*).

This variation could be solely due to the random sampling of genes within functions. Modeling the probability of having long mono-SSR under purifying selection may allow the test for this hypothesis. As a first approximation, we used the observed density of long mono-SSR in coding sequences to compute an average rate of SSR per base. If all genes were under the same selective constraints, the number of SSR per gene should be Poisson distributed with this average rate multiplied by their length. Accordingly, we computed the probability to host at least one long mono-SSR (i.e., to be an hypermutable gene) for all genes. We then computed, for each function, a 95% CI for the expected number of hypermutable genes. Among terms with more than 20 genes, we found 171/1,238 (13.8%) terms outside the CI; this is larger than the 5% we expected if coding mono-SSRs were under the same selective pressure in all functions.

Discussion

In this study, we assumed that all genes hosting a long enough mono-SSR can be considered as hypermutable genes. Whatever the chosen threshold for hypermutability, we show that only a cohesive restricted set of functions are overrepresented among hypermutable genes. Interestingly, we show that this is only due to the mono-SSR within genes, the other type of SSRs being uniformly distributed among functions. Using a probabilistic model, we were able to show that mono-SSRs are shorter than expected by a model of neutral substitution (which is coherent with previous studies, e.g., Metzgar et al. [2000]; Ackermann and Chao [2006]) and that hypermutable genes are avoided in almost all functions. Finally, our study shows that the strength of purifying selection, that removes hypermutable genes from the human genomes, varies greatly from function to function.

SSRs Are Kept Small by Purifying Selection in Exons

The comparison between introns and exons suggests that frameshifting SSRs are subject to a strong purifying selection in coding sequences. Indeed, if one considers that intron evolution is almost neutral, then the length of intronic SSRs must be solely the consequence of their mutation process. The differences observed between length of exonic and intronic SSRs reflect the existence of selection that acts against free expansion of those SSRs in coding sequence.

Indeed, using a model that predicts the size of the longest mono-SSR expected in a coding sequence of a given length and composition, we showed that, in exons, mono-SSR length is globally smaller than expected. In introns, G/C-SSRs are also shorter than expected but A/T-SSRs are usually longer than expected. This is consistent with the observation that G/C-SSRs are generally smaller than A/T-SSRs (Li et al. 2002). This suggests that A/T- and G/C-SSRs should be considered separately. Insertion of Alu sequences in introns contributes to the abundance of long A/T-SSRs but is not sufficient to explain their

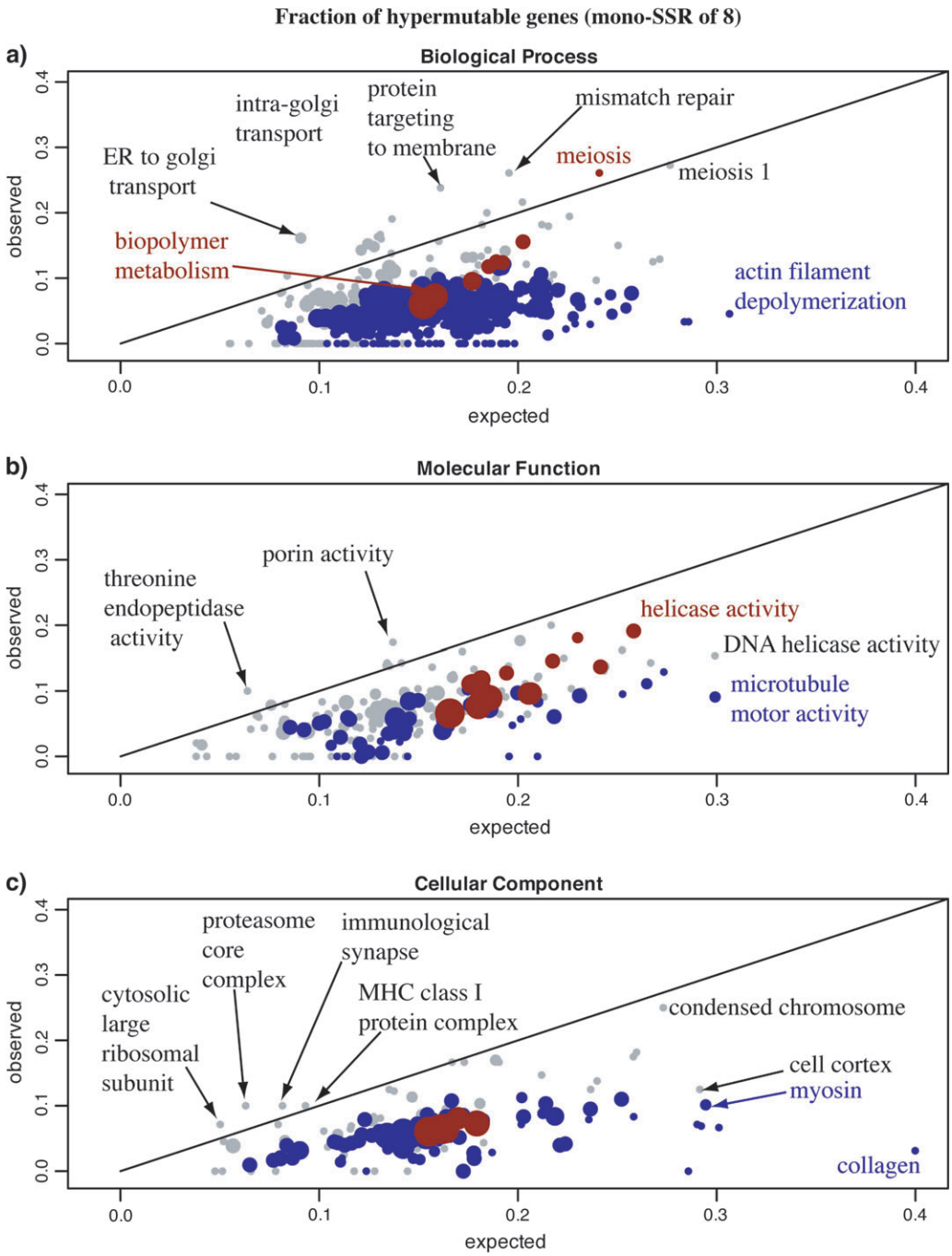


FIG. 3.—Expected and observed fractions of hypermutable genes for all GO terms. Here we represent for each term, the observed proportion of genes that contain a mono-SSR larger than eight as a function of its expected fraction. The terms are extracted from (*a*) BP, (*b*) MF, and (c) CC ontologies. The size of each dot is proportional to the total number of genes that term encompasses (taken as discrete intervals: [20, 50], [50, 100], [100, 500], [500, 10^3], and [10^3, infinity]). Terms with less than 20 genes were not represented. Terms we found statistically overrepresented among hypermutable genes (terms from fig. 2) are colored in red. The line represents the ratio observed/expected = 1. Terms that are significantly outside the 95% CI predicted under neutrality are colored in blue. This figure shows that almost all functions contain less genes carrying a long mono-SSR than expected. This again illustrates that most, if not all, long mono-SSR tends to be removed by purifying selection. Although, it also shows that some functions (e.g., meiosis, MMR, and "condensed chromosome") encompass many genes with long mono-SSR along with an observed/expected ratio close to 1. This suggests that genes involved in those functions are under relaxed purifying selection.

abundance. Because there is no reading frame, one could imagine that A/T-SSRs can undergo free expansion. The model we used as a reference assumes that all SSRs are created by an accumulation of substitutions. Beyond a threshold size, SSRs experience expansions through replication slippage (or recombination) and then become longer than expected. Obviously, there are additional factors that prevent G/C-SSRs to expand. As for coding sequences, we

suspect that G/C-SSRs are kept short by purifying selection in introns. Two molecular evidences are compatible with this hypothesis. First, G-rich tracts are known to adopt unusual DNA structure (parallel quadruplex) involved in different biological functions (Sen and Gilbert 1988). Second, G-rich tracts are also prone to electron transfer that causes oxidative damage (Hall et al. 1996). For one or the other (or both) reasons, there is a good chance that G/C-SSRs have an impact on fitness even in introns. This effect should equally apply in exons. Those deleterious effects certainly add up with those previously highlighted (selection against frameshifts).

Functions of Hypermutable Genes

Despite this global underrepresentation of SSRs in exonic sequences, several genes still host a long SSR. Defining a threshold for long SSRs is not trivial. Thus, we used two sets of values that are relevant for the minimum size beyond which SSRs are subject to expansion and contraction. It is important to mention that both sets of thresholds lead to extremely similar results. This highlights the robustness of our results to the choice of a threshold for hypermutability. Among the very large number of terms that were annotated in the human genes, only a restricted number exhibits an overrepresentation of hypermutable genes.

Legendre et al. (2007) conducted a similar analysis on a data set that includes all genes that contain any type of SSR. BP overrepresented among this data set is different from the ones we report here. An analysis of the 1,266 genes hosting a long tri-SSR reveals a similar set of functions (data not shown), with the exception of neurogenesis and related terms. We suspect that the difference in metric for hypermutability explains this difference. Because many neurological disorders are caused by the presence of a coding tri-SSR, we conclude that the overrepresentation of the functions described by Legendre et al. is mainly driven by genes hosting a tri-SSR that we ignored in our study.

Importantly, one could argue that this is a consequence of large duplicate families that share often the same annotations. However, using Ensembl definition of gene family (Enright et al. 2002), we computed for each function the fraction of genes that contain a duplicate within the function. No differences were observed between the overrepresented functions and the others (0.35 vs. 0.41, $P = 0.18$ when considering all genes, 0.25 vs. 0.31, $P = 0.32$ when considering genes with mono-SSR, Mann–Whitney U test). Therefore, this overrepresentation is not an artifact of large duplicate families. Our analysis shows that those functions are generally devoted to cell cycle and maintenance of genome integrity (DNA repair, meiosis, cell cycle, helicase domain–containing genes, nuclear localized genes, etc.). It should be mentioned that a similar set of functions is overrepresented among genes that host at least two long mono-SSRs (data not shown). Furthermore, the same analysis with annotations from PantherDB (Mi et al. 2005) also leads to a similar set of functions (data not shown). Overall, we think that our results are robust to the most obvious artifacts and that the restricted cohesive set of functions we find overrepresented in hypermutable genes are meaningful.

The Strength of Purifying Selection against Hypermutable Genes Varies from Function to Function

We computed an expected fraction of hypermutable genes in all functional groups of genes and compared it with the observed fraction. We show that almost all functions clearly harbor less hypermutable genes than expected under neutrality. This strongly suggests that the vast majority of long mono-SSRs are kept out of coding sequences by purifying selection.

Functions overrepresented among the hypermutable genes (i.e., those dedicated to genomic stability and cell cycle) are expected to contain a large fraction of hypermutable genes. They are longer and/or more biased in composition than the average genes. Therefore, the overrepresentation of hypermutable genes in those functions can be explained by the length and the nucleotide composition of genes among those functions. This points out the importance of using a statistical framework that tests for the effect of length and composition of the genes.

An overestimation of the expected number of long mono-SSRs would diminish the strength of the purifying selection we observe. At least three properties of DNA-coding sequences were neglected in our model. First, slippage process was ignored, although almost none is expected in coding sequence. Slippage, however, leads to larger mono-SSR than what is observed in coding sequence. Therefore, ignoring slippage lowers the expected number and size of mono-SSR. Second, we also ignored the dependency of nucleotide context in coding sequences. We estimated the probabilities of mono-SSR in coding sequences using a simulated data set of random sequences modeled by a Markov model of size 2 (using the frequency of the 3-mers). Using these probabilities instead of the one given by the Poisson model does not qualitatively changes our results. Finally, we ignored the amino acid sequences of the genes. Ackermann and Chao (2006) fixed the amino acid sequences of genes and showed that mono-SSRs are underrepresented.

There are few functions for which we observed as many hypermutable genes as expected under a neutral model. For these functions, long mono-SSRs are virtually neutral. On another extreme, we shall consider functions that are expected to contain long mono-SSR but do not (e.g., cytoskeleton- and collagen-related genes). Overall, we have to acknowledge that the strength of the purifying selection that acts against long mono-SSR varies from function to function, from very strong (e.g., for collagen) up to its complete absence (e.g., for ER to Golgi transport). We can propose several hypotheses to explain this observation.

First, the rate of instability for SSR within the same genome may greatly vary from one locus to another. Therefore, we can imagine that the hypermutable genes are located in peculiar loci in the genome where SSRs are stabilized.

Alternatively, it is possible that the functions where mono-SSRs are apparently neutral could be composed of genes that are more “dispensable” than others. Here we used dispensable to refer to a low cost in fitness when the gene is not properly expressed. For the human genome, we however do not have a list of phenotype associated with the absence of all genes. The use of Online Mendelian

Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim/) seems inappropriate because although half of the genes carry an entry, the entries clearly do not have the same meaning in terms of individual fitness and the genes with no annotation cannot be considered as dispensable.

Finally, it is possible that the apparent neutrality of SSR could be the result of a balance between positive and negative selection. If the expression of a gene is associated to sometimes positive, sometimes negative fitness, one could imagine that the evolution of such a gene would look neutral even though it is always under selection. Here, we find that genes that host a long SSR are devoted to the maintenance of DNA integrity. Why did such genes retain hypermutable motifs in their coding sequences? Previous studies (Moxon and Wills 1999; Chang et al. 2001; Rocha et al. 2002; Kashi and King 2006) reported the presence of long mono-SSR in MMR genes and proposed that these genes tune the global mutation rate of the organism by switching on and off after a loss-of-frame mutation caused by replication slippage. Mutator phenotypes, generally caused by a mutated MMR gene (Rosenberg et al. 1998), have been shown to be evolutionary advantageous in bacteria facing an environmental challenge (Taddei et al. 1997). Among a population under stress, individuals with a new advantageous mutation (most likely individuals bearing the mutator allele) will improve in fitness. Thus, this advantageous mutation will increase in frequency along with the mutator (by hitchhiking). If genetic linkage is likely to be strong in bacteria, it is not in eukaryotes. Therefore, the possibility of mutators in the human lineage seems difficult. We can nonetheless intuitively suspect that selection could favor a premutator state (i.e., unstable mono-SSR hosted in coding sequence) in some function (e.g., genes devoted to genomic stability), although it would require more theoretical investigations that will not be conducted here.

It seems difficult at this stage to definitely support or reject one of the hypotheses. However, we would like to mention that the last hypothesis (hidden positive selection) should be regarded with caution. If long mono-SSR looks neutral in these genes, the most parsimonious explanation is that they are neutral.

As a consequence, we do not favor this "oscillating mode of selection" hypothesis and challenge the existence of mutator genes in human and more generally in eukaryotes.

Conclusion

The hypermutability of the human genes (when considering only potentially unstable SSR) is typically a consequence of their length and/or nucleotide composition. Most long SSRs are removed from coding sequence by purifying selection. However, a restricted set of functions seems to be insensitive to the presence of a priori deleterious long SSR. The mystery of this apparent relaxed purifying selection needs more thought and data. In that respect, we think that there is a need for more theory along with a phylogenetic perspective on the evolution of coding SSR to gather further insight in this unclosed debate.

Supplementary Material

Supplementary figure S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Authors would like to thank S. Schbath, E. Rocha, I. Gonçalves, S. Baulac, E. Leguern, and C. Castillo-Davis for comments on previous versions of the manuscript.

Literature Cited

Aaltonen LA, Peltomaki P, Leach FS, et al. (15 co-authors). 1993. Clues to the pathogenesis of familial colorectal cancer. Science. 260:812–816.

Ackermann M, Chao L. 2006. DNA sequences shaped by selection for stability. PLoS Genet. 2:e22.

Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 22:1600–1607.

Ashburner M, Ball CA, Blake JA, et al. (17 co-authors). 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 25:25–29.

Castillo-Davis CI, Hartl DL. 2003. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 19:891–892.

Chang DK, Metzgar D, Wills C, Boland CR. 2001. Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. Genome Res. 11: 1145–1146.

Conti E, Izaurralde E. 2005. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. Curr Opin Cell Biol. 17:316–325.

de Wachter R. 1981. The number of repeats expected in random nucleic acid sequences and found in genes. J Theor Biol. 91: 71–98.

Dieringer D, Schlotterer C. 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 13: 2242–2251.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics. 148:1667–1686.

Duval A, Hamelin R. 2003. Replication error repair, microsatellites, and cancer. Med Sci (Paris). 19:55–62.

Ellegren H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet. 24:400–402.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 5:435–445.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Everett CM, Wood NW. 2004. Trinucleotide repeats and neurodegenerative disease. Brain. 127:2385–2405.

Fabre E, Dujon B, Richard GF. 2002. Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. Nucleic Acids Res. 30:3540–3547.

Gragg H, Harfe BD, Jinks-Robertson S. 2002. Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in *Saccharomyces cerevisiae*. Mol Cell Biol. 22:8756–8762.

Hall DB, Holmlin RE, Barton JK. 1996. Oxidative DNA damage through long-range electron transfer. Nature. 382:731–735.

Jacob S, Praz F. 2002. DNA mismatch repair defects: role in colorectal carcinogenesis. Biochimie. 84:27–47.
Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22:253–259.
Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18:30–38.
Kruglyak S, Durrett RT, Schug MD, Aquadro CF. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA. 95:10774–10778.
Lai Y, Sun F. 2003a. Microsatellite mutations during the polymerase chain reaction: mean field approximations and their applications. J Theor Biol. 224:127–137.
Lai Y, Sun F. 2003b. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol. 20:2123–2131.
Legendre M, Pochet N, Pak T, Verstrepen KJ. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res. 17:1787–1796.
Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 4:203–221.
Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 11:2453–2465.
Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 10:72–80.
Mi H, Lazareva-Ulitsky B, Loo R, et al. (12 co-authors). 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 33:284–288.
Miquel C, Jacob S, Grandjouan S, Aime A, Viguier J, Sabourin JC, Sarasin A, Duval A, Praz F. 2007. Frequent alteration of DNA damage signalling and repair pathways in human colorectal cancers with microsatellite instability. Oncogene. 26:5919–5926.
Mori Y, Yin J, Rashid A, Leggett BA, Young J, Simms L, Kuehl PM, Langenberg P, Meltzer SJ, Stine OC. 2001. Instabilotyping: comprehensive identification of frameshift mutations caused by coding region microsatellite instability. Cancer Res. 61:6046–6049.
Moxon ER, Wills C. 1999. DNA microsatellites: agents of evolution? Sci Am. 280:94–99.
Pupko T, Graur D. 1999. Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. J Mol Evol. 48:313–316.
Rivals I, Personnaz L, Taing L, Potier MC. 2007. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics. 23:401–407.
Robin S, Rodolphe F, Schbath S. 2005. DNA words and models. Cambridge: Cambridge University Press.
Rocha EPC, Matic I, Taddei F. 2002. Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? Nucleic Acids Res. 30: 1886–1894.
Rose O, Falush D. 1998. A threshold size for microsatellite expansion. Mol Biol Evol. 15:613–615.
Rosenberg SM, Thulin C, Harris RS. 1998. Transient and heritable mutators in adaptive evolution in the lab and in nature. Genetics. 148:1559–1566.
Sagher D, Hsu A, Strauss B. 1999. Stabilization of the intermediate in frameshift mutation. Mutat Res. 423:73–77.
Sen D, Gilbert W. 1988. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature. 334:364–366.
Shinde D, Lai Y, Sun F, Arnheim N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res. 31:974–980.
Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD. 1997. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol. 17:2851–2858.
Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 9:657–663.
Strand M, Prolla TA, Liskay RM, Petes TD. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature. 365:274–276.
Strauss BS. 1999. Frameshift mutation, microsatellites and mismatch repair. Mutat Res. 437:195–203.
Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. 1997. Role of mutator alleles in adaptive evolution. Nature. 387:700–702.
Tautz D. 1994. Simple sequences. Curr Opin Genet Dev. 4:832–837.
Toth G, Gaspari Z, Jurka J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10: 967–981.
Tran HT, Keen JD, Kricker M, Resnick MA, Gordenin DA. 1997. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. Mol Cell Biol. 17:2859–2865.
Woerner SM, Kloor M, von Knebel Doeberitz M, Gebert JF. 2006. Microsatellite instability in the development of DNA mismatch repair deficient tumors. Cancer Biomark. 2:69–86.
Xu X, Peng M, Fang Z. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet. 24: 396–399.

Naruya Saitou, Associate Editor

Accepted October 5, 2008

Frequency spectrum neutrality tests, one for all and all for one.

Guillaume Achaz[1,2]

July 21, 2009

[1]: Systématique, Adaptation et Evolution (UMR 7138), Université Pierre et Marie Curie-Paris VI, CNRS, MNHN, IRD, Paris, France

[2]: Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris VI, Paris, France

Running Head: Frequency spectrum neutrality tests

Keywords: coalescent theory, neutrality tests

Corresponding Author:
Guillaume ACHAZ
Atelier de Bioinformatique
Université Pierre et Marie Curie
4, place Jussieu
Boîte courrier 1202
75005 PARIS
tel: +33-1-44-27-65-82
fax: +33-1-44-27-63-12
email: achaz@abi.snv.jussieu.fr

Abstract

Neutrality tests based on the frequency spectrum (e.g. Tajima's D or Fu and Li's F) are commonly used by population geneticists as routine tests to assess the likeliness of the standard neutral model on their dataset. Here, I show that these neutrality tests are specific instances of a general model that encompasses them all. I illustrate how this general framework can be taken advantage of to devise new more powerful tests that better detect deviations from the standard model. Finally, I exemplify the usefullness of the framework on SNP data by showing how it supports the selection hypothesis in the lactase human gene by overcoming the ascertainment bias. The framework presented here paves the way for constructing novel tests optimized for specific violations of the standard model and ultimately help to unravel scenarios of evolution.

Introduction

The standard models of population genetics (*i.e.* Wright-Fisher model and related ones) constitute null models for which an amazing amount of theory has been developed. Population geneticists have used some aspect of the theory (e.g. summary statistics) to test the goodness of fit of the standard model on a given dataset. Rejection of the standard model typically suggests that alternative hypotheses, such as selection or demographic history, have to be accounted for. Although they test for more than neutrality, tests that compute the goodness of fit of the standard model have been referred to as "neutrality tests". Since different neutrality tests have varying sensitivity to different violations of the standard model, one typically uses a plethora of tests on the dataset of interest. One then hopes that the evolutionary processes that generated the dataset will be, at least partially, uncovered by the tests. Although neutrality tests based on population samples exhibit important diversity, they can be assigned to families such as: "haplotypes tests" (e.g. Fu (1997); Depaulis and Veuille (1998)) that use the distribution of haplotypes, "tree shape tests" that try to capture specific tree deformations (e.g. Ramos-Onsins and

Rozas (2002)) and "frequency spectrum tests" that are based on the frequency spectrum (e.g. Tajima (1989); Fu and Li (1993b); Fay and Wu (2000); Achaz (2008)).

In the present study, I will investigate neutrality tests based on the frequency spectrum (hereafter refered simply as neutrality tests) and show that they are all specific instances of a general framework. Neutrality tests compare two estimators of the population mutation parameter θ that characterizes the mutation-drift equilibrium. It is defined as $\theta = 2pN_e\mu$, where p is the ploidy (1 for haploids and 2 for diploids), N_e the effective population size and μ the locus neutral mutation rate. When the standard model is true, the expectations of the several unbiased estimators of θ are equal.

Typical estimators of θ, in a sample of n sequences, are: $\hat{\theta}_S = S/a_n$, where S is the number of polymorphic sites and $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ (Watterson, 1975), $\hat{\theta}_\pi = \pi$, where π is the average pairwise differences between all sequences in the sample (Tajima, 1983). If an outgroup is available, mutations at frequency i/n can be distinguished from mutations at frequency $1 - i/n$. Following Fu (1995)'s notations, $\boldsymbol{\xi}$ is a vector that represents the unfolded frequency spectrum composed of ξ_i, the number of polymorphic sites at frequency i/n in the sample ($i \in [1, n-1]$). When no outgroup is available, the frequency spectrum is folded and is given by a vector $\boldsymbol{\eta}$, composed of η_i, the number of polymorphic sites at both frequencies i/n and $1 - i/n$. Accordingly, it has been shown that θ can be estimated from $\hat{\theta}_{\xi_1} = \xi_1$, with ξ_1 the number of derived singletons (Fu and Li, 1993b), from $\hat{\theta}_{\eta_1} = \frac{n-1}{n}\eta_1$, with η_1 the total number of singletons (derived and ancestral) (Fu and Li, 1993b) and from $\hat{\theta}_H = \sum_{i=1}^{n-1} \frac{2i^2}{n(n-1)}\xi_i$ (Fay and Wu, 2000). Recently, it has been suggested that singletons should be ignored when θ is estimated in samples with sequencing errors ; this leads to estimators such as $\hat{\theta}_{\pi-\xi_1}$, $\hat{\theta}_{S-\xi_1}$, $\hat{\theta}_{\pi-\eta_1}$ and $\hat{\theta}_{S-\eta_1}$ (Achaz, 2008). Other estimators of θ, such as $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$, were designed to minimize their variance (Fu, 1994b), although they can be only computed using recursions for a given value of θ.

Neutrality tests compute the goodness of fit of a statistic T, which is the difference between two estimators of θ, normalized by its standard deviation:

$$T = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{Var[\hat{\theta}_1 - \hat{\theta}_2]}} = \frac{t}{\sqrt{\alpha_n \theta + \beta_n \theta^2}} \quad (1)$$

For a given θ, under the standard model, T has a mean of $E[T] = 0$ and a variance of $Var[T] = 1$. Lower case letters (e.g. t) will denote the absolute difference (*i.e.* the numerator only) and upper case (e.g. T) the normalized difference (equation 1) throughout this work. Interestingly, the variance in the denominator is a function of both θ and θ^2. Because θ is unknown, the denominator cannot be computed as such. In practice, unbiased estimators of θ and θ^2 must be used instead. Because the variance of $\hat{\theta}_S$ vanishes asymptotically in a very large sample ($\lim_{n\to\infty} Var[\hat{\theta}_S] = 0$), θ and θ^2 are, in practice, substituted by estimators based on S (Tajima, 1989), which changes the mean and the variance of T to: $E[T] \approx 0$ and $Var[T] \approx 1$.

Tajima's D (Tajima, 1989) is defined by $d = \hat{\theta}_\pi - \hat{\theta}_S$; the statistics proposed by Fu and Li (1993b) are $f = \hat{\theta}_\pi - \hat{\theta}_{\xi_1}$, $f^* = \hat{\theta}_\pi - \hat{\theta}_{\eta_1}$, $d_2 = \hat{\theta}_S - \hat{\theta}_{\xi_1}$ and $d_2^* = \hat{\theta}_S - \hat{\theta}_{\eta_1}$. Another classical statistics is $h = \hat{\theta}_\pi - \hat{\theta}_H$ (Fay and Wu, 2000), even though its variance was not given by the authors. Finally, two other related neutrality tests that are, *a priori*, immune to sequencing errors were proposed: $y = \hat{\theta}_{\pi_{-\xi_1}} - \hat{\theta}_{S_{-\xi_1}}$ and $y^* = \hat{\theta}_{\pi_{-\eta_1}} - \hat{\theta}_{S_{-\eta_1}}$ (Achaz, 2008). Other tests based on θ_ξ and θ_η (which are optimized for a given θ value) as well as the difference between the observed and the expected values of the frequency spectrum were also proposed (Fu, 1996).

Here, I show that when using a general weighted linear combination of $\hat{\theta}_i = i\xi_i$ (or $\hat{\theta}_i^*$ when no outgroup is available) any estimators of θ (*i.e.* $\hat{\theta}_\omega = \frac{1}{\sum \omega_i} \sum_i \omega_i i\xi_i$) and consequently any neutrality tests can be derived. Nawa and Tajima (2008) recently advocated the use of the $\hat{\theta}_i^*$ spectrum, which is expected to be uniform under the standard model, as a visual test for neutrality instead of the classical frequency spectrum. This last proposal is in complete agreement with the current work. Importantly, it has been previously reported that some θ estimators and neutrality tests could be expressed as specific linear combinations of ξ_i or η_i (Tajima, 1997; Wakeley, 2009). Furthermore, Fu (1997) show that several θ estimators can be expressed as specific linear combinations of

$\hat{\theta}_i$ ($\hat{\theta}_{L(x)} = \frac{1}{\sum i^{-x}} \sum_i i^{-x} i\xi_i$) or in a related framework that uses $\hat{\theta}_i^*$ instead of $\hat{\theta}_i$. $\hat{\theta}_H$ was subsequently designed as $\hat{\theta}_{L(-1)}$ (Fay and Wu, 2000). However, some estimators (like $\hat{\theta}_\pi$, $\hat{\theta}_{\pi_{-\xi_1}}$ or $\hat{\theta}_{S_{-\xi_1}}$) cannot be expressed using the Fu (1997) framework. To the best of my knowledge, no previous study has explicitly derived the framework presented here. No work has yet highlighted the striking simplicity of θ estimators and related tests, when expressed in this framework. I further show how the use of such a simple framework greatly facilitates the study of previous θ estimators and their related neutrality tests and how it opens the door for constructing yet undiscovered interesting θ estimators and neutrality tests with enhanced power.

Model

With an outgroup

According to Fu (1995), we know that:

$$E[\xi_i] = \theta/i \tag{2}$$

$$Var[\xi_i] = \theta/i + \sigma_{ii}\theta^2 \tag{3}$$

$$Cov[\xi_i, \xi_j] = \sigma_{ij}\theta^2 \tag{4}$$

where σ_{ii} and σ_{ij} only depend on n and are given in Equation 2 of Fu (1995). This shows that $E[i\xi_i] = \theta$ and therefore that any ξ_i can be used to construct an unbiased estimator of θ:

$$\hat{\theta}_i = i\xi_i \tag{5}$$

Consequently, a linear combination $\hat{\theta}_\omega$ of the $\hat{\theta}_i$s (in which the weights sum to 1) is also an unbiased estimator of θ. Mathematically, it is expressed as:

$$\hat{\theta}_\omega = \frac{1}{\sum_i \omega_i} \sum_{i=1}^{n-1} \omega_i \; i\xi_i \tag{6}$$

where ω_i is the weight of each $\hat{\theta}_i$ in the combined estimator. Therefore, any estimator based on the frequency spectrum can be solely described by an $\boldsymbol{\omega}$ vector. Importantly, it should be mentioned that Fu (1997) also proposed a linear combination of $i\xi_i$, but in which only a subset of the weight vectors were used. Namely, the proposed weight vectors were restricted to $\omega_i = i^{-x}$.

Using Equations 3 and 4 the variance of $\hat{\theta}_\omega$ can be shown to be:

$$\begin{aligned} \mathrm{Var}[\hat{\theta}_\omega] &= (\sum_i \omega_i)^{-2}\Big(\sum_{i=1}^{n-1} \omega_i^2 i^2 Var[\xi_i] + 2\sum_i \sum_{j>i} ij\omega_i\omega_j Cov[\xi_i, \xi_j]\Big) \\ &= (\sum_i \omega_i)^{-2}\Big(\theta\big(\sum_i \omega_i^2 i\big) + \theta^2\big(\sum_i \omega_i^2 i^2\sigma_{ii} + 2\sum_i \sum_{j>i} ij\omega_i\omega_j\sigma_{ij}\big)\Big) \end{aligned} \tag{7}$$

Following Tajima (1989), using Equation 1, one can compute a normalized statistic that is, in the general framework:

$$T_\Omega = \frac{\hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2}}{\mathrm{Var}^{1/2}[\hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2}]} \tag{8}$$

which can be expressed as a function of an Ω vector:

$$T_\Omega = \frac{\sum_i \Omega_i\ i\xi_i}{\sqrt{\alpha_n\theta + \beta_n\theta^2}} \tag{9}$$

with:

$$\begin{aligned} \Omega_i &= \frac{\omega_{1i}}{\sum_j \omega_{1j}} - \frac{\omega_{2i}}{\sum_j \omega_{2j}} \\ \alpha_n &= \sum_i i{\Omega_i}^2 \\ \beta_n &= \sum_i i^2{\Omega_i}^2\sigma_{ii} + 2\sum_i \sum_{j>i} ij\Omega_i\Omega_j\sigma_{ij} \end{aligned}$$

The Ω vector results from the difference between two weight vectors normalized to 1. As a consequence, (1) all elements of the Ω vector sum to 0 and (2) the sum of all positive values cannot be greater than 1 and the sum of all negative values cannot be smaller than

-1. Any vector that fits these two constraints can be considered, along with Equation 9, as a neutrality test.

Without an outgroup

If no adequate outgroup is available, the unfolded frequency spectrum and consequently the $\hat{\theta}_i$ spectrum, cannot be computed. This implies that one has to use the $\boldsymbol{\eta}$ folded frequency spectrum. Following Fu (1995), we define $\eta_i = (\xi_i + \xi_{n-i})/(1 + \delta_{i,n-i})$ and therefore we have:

$$E[\eta_i] = \phi_i\theta \quad (10)$$

$$Var[\eta_i] = \phi_i\theta + \rho_{ii}\theta^2 \quad (11)$$

$$Cov[\eta_i, \eta_j] = \rho_{ij}\theta^2 \quad (12)$$

where $\delta_{i,n-i}$ is a Kroneker delta (1 if $i = j$, 0 otherwise) and where:

$$\begin{aligned}
\phi_i &= \frac{n}{(1+\delta_{i,n-i})i(n-i)} \\
\rho_{ii} &= \frac{\sigma_{ii} + \sigma_{(n-i)(n-i)} + 2\sigma_{i(n-i)}}{(1+\delta_{i,n-i})^2} \\
\rho_{ij} &= \frac{\sigma_{ij} + \sigma_{i(n-j)} + \sigma_{(n-i)j} + \sigma_{(n-i)(n-j)}}{(1+\delta_{i,n-i})(1+\delta_{j,n-j})}
\end{aligned}$$

Although, we cannot compute the $\hat{\theta}_i = i\xi_i$ spectrum (as defined above), we can compute a folded $\hat{\theta}_i^*$ spectrum defined as:

$$\hat{\theta}_i^* = \phi_i^{-1}\eta_i \quad (13)$$

This folded $\hat{\theta}_i^*$ spectrum is the visual neutrality test proposed by Nawa and Tajima (2008). Using a similar reasoning as above, a linear combination of $\hat{\theta}_i^*$ leads to a generic unbiased estimator of θ defined as:

$$\hat{\theta}_\omega^* = \frac{1}{\sum_i \omega_i^*} \sum_{i=1}^{n/2} \omega_i^* \ \phi_i^{-1}\eta_i \quad (14)$$

which variance is given by:

$$\mathrm{Var}[\hat{\theta}^*_\omega] = (\sum_i \omega^*_i)^{-2}\Big(\theta\big(\sum_i \omega^{*2}_i \phi_i^{-1}\big) + \theta^2\big(\sum_i \omega^{*2}_i \phi_i^{-2}\rho_{ii} + 2\sum_i \sum_{j>i} \phi_i^{-1}\phi_j^{-1}\omega^*_i\omega^*_j\rho_{ij}\big)\Big) \quad (15)$$

Consequently, the corresponding neutrality test T^*_Ω is:

$$T^*_\Omega = \frac{\sum_i \Omega^*_i \ \phi_i^{-1}\eta_i}{\sqrt{\alpha^*_n\theta + \beta^*_n\theta^2}} \quad (16)$$

with:

$$\begin{aligned}
\Omega^*_i &= \frac{\omega{1_i}^*}{\sum_j \omega{1_j}^*} - \frac{\omega{2_i}^*}{\sum_j \omega{2_j}^*} \\
\alpha^*_n &= \sum_i \phi_i^{-1}\Omega^{*2}_i \\
\beta^*_n &= \sum_i \phi_i^{-2}\Omega^{*2}_i\rho_{ii} + 2\sum_i\sum_{j>i}\phi_i^{-1}\phi_j^{-1}\Omega^*_i\Omega^*_j\rho_{ij}
\end{aligned}$$

It is important to mention that Tajima (1997) previously showed that D, F^* and D^*_2 could be expressed as a linear combination of η_i. More precisely, the vectors used then correspond in the present framework to $\frac{\sum_i \Omega^*_i \ \phi_i^{-1}}{\sqrt{\alpha^*_n\hat{\theta}_S + \beta^*_n\hat{\theta}^2_S}}$. This vector definition emphasizes the weight on each η_i rather than on each $\hat{\theta}^*_i$.

With or without an outgroup

Using both definitions of $\hat{\theta}_i$ (Equation 5) and $\hat{\theta}^*_i$ (Equation 13), it is easy to show that we have:

$$\hat{\theta}^*_i = \frac{1}{n}\big((n-i)\ \hat{\theta}_i + i\ \hat{\theta}_{n-i}\big) \quad (17)$$

As a consequence, the use of an $\boldsymbol{\omega^*}$ vector along with the $\boldsymbol{\eta}$ folded frequency spectrum is only equivalent to the use of an $\boldsymbol{\omega}$ vector with the $\boldsymbol{\xi}$ unfolded frequency spectrum when

we have:

$$\frac{1}{n}\big((n-i)\ \omega_i^*\ \hat{\theta}_i + i\ \omega_i^*\ \hat{\theta}_{n-i}\big) = \frac{1}{(1+\delta_{i,n-i})}\ \big(\omega_i\hat{\theta}_i + \omega_{n-i}\hat{\theta}_{n-i}\big)$$

$$\begin{aligned}\omega_i^* &= \frac{n}{(n-i)(1+\delta_{i,n-i})}\omega_i \\ &= \frac{n}{i(1+\delta_{i,n-i})}\omega_{n-i}\end{aligned} \tag{18}$$

This makes clear that there is an equivalent $\boldsymbol{\omega^*}$ vector for any $\boldsymbol{\omega}$ vector that adheres to the following constraint:

$$i\ \omega_i = (n-i)\ \omega_{n-i} \tag{19}$$

To fold the frequency spectrum, the weight $i\omega_i$ associated to ξ_i (and not to $\hat{\theta}_i$) has to be the same as the weight $(n-i)\omega_{n-i}$ associated to ξ_{n-i}. This translates into a $i\omega_i$ vector that is symmetric around $n/2$. Furthermore, when the constraint (expressed in Equation 19) is fulfilled we can write, for any $0 \leq f \leq 1$:

$$\omega_i^* = \frac{n}{(1+\delta_{i,n-i})}\Big(f\frac{\omega_i}{(n-i)} + (1-f)\frac{\omega_{n-i}}{i}\Big)$$

which leads interestingly for $f = (n-i)/n$ to:

$$\omega_i^* = (\omega_i + \omega_{n-i})\frac{1}{(1+\delta_{i,n-i})} \tag{20}$$

The weights on $\hat{\theta}_i^*$ simply result from the sums of the weights on $\hat{\theta}_i$ and on $\hat{\theta}_{n-i}$ that are pooled when the spectrum is folded. In that respect, any $\boldsymbol{\omega}$ vector complying to Equation 19 can be used without the help of an outgroup. The $\boldsymbol{\omega^*}$ vectors are then a subset of all possible values of the $\boldsymbol{\omega}$ vectors. The former can be computed from the latter by using Equations 18 or 20.

Because $\boldsymbol{\Omega}$ is the difference between two normalized $\boldsymbol{\omega}$ vectors, all relationships between $\boldsymbol{\omega}$ and $\boldsymbol{\omega^*}$ expressed above also hold for $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega^*}$.

Results

The model described above shows that all estimators of θ based on the frequency spectrum are linear combinations of $\hat{\theta}_i = i\xi_i$, weighted by a specific vector $\boldsymbol{\omega}$. When no outgroup is available, one can use a linear combination of $\hat{\theta}_i^* = \phi_1^{-1}\eta_i$, weighted by a vector $\boldsymbol{\omega}^*$. Consequently, neutrality tests can be expressed as a linear combination of $\hat{\theta}_i$ (or $\hat{\theta}_i^*$) weighted by a vector $\boldsymbol{\Omega}$ (or $\boldsymbol{\Omega}^*$), for which a variance can be computed easily. Three applications of the model will be developed below. First, I will first re-investigate the previous estimators of θ and their corresponding neutrality tests and frame their intrinsic properties in terms of $\hat{\theta}_i$ ($\hat{\theta}_i^*$) spectrum. Then, since previous tests are only specific instances of the framework, I will show how the model can be used to build new tests that are more powerful than previous ones. Finally, I will exemplify the benefit of the framework on real data that are known to be subject to an ascertainment bias.

Previous θ estimators and neutrality tests

Using Equation 6, all previously reported θ estimators are given by an $\boldsymbol{\omega}$ vector (Table 1). When defined, the corresponding $\boldsymbol{\omega}^*$ vectors are also provided (Table 1). A graphical representation of four estimators of θ is shown in Figure 1. This figure highlights that both $\hat{\theta}_S$ and $\hat{\theta}_\pi$ emphasize the low-frequency polymorphic sites in their estimation of θ (although not as much as $\hat{\theta}_{\xi_1}$, which is solely based on derived singletons) and that on the contrary, $\hat{\theta}_H$ gives more weight to ancestral polymorphisms. Framed in the folded spectrum, $\hat{\theta}_S$ still weights more low+high frequencies whereas $\hat{\theta}_\pi$ has a uniform weight. Potentially, using other weight vectors, one could express any undiscovered estimator of θ based on the frequency spectrum.

The numerical variance of the previous estimators of θ are reported in Table 1 (for $n = 30$ and $\theta = 1, 10, 100$). They can be computed either by their original derivations or by Equation 7. This clearly shows that, among previous estimators of θ, the variance of $\hat{\theta}_S$ is the smallest and the one from $\hat{\theta}_H$ is the largest. This can be explained by the fact that the variance of $\hat{\theta}_i$ increases with i. As a consequence $\hat{\theta}_H$, which puts more weight on

ancestral alleles, shows a larger variance. Interestingly, estimators without singletons have relatively small variances.

Previous neutrality tests are given in Table 2. A graphical representation of the $\mathbf{\Omega}$ vectors (and $\mathbf{\Omega^*}$ when defined) used in four previous tests is reported in Figure 2. This figure shows that the sensitivity of the different tests differ although they share some common features. For example, D and F^* both are negatively sensitive to both low and high frequencies (although more sensitivity to low frequencies). D shows opposite sensitivity between medium frequencies and low/high frequency, whereas F^* shows poor sensitivity to medium frequency polymorphisms. F and F^* have opposite effects on doubletons and singletons. Thus, deviations that enhance both will have opposite effects. Finally, H is oppositely skewed by low and high frequencies.

One crucial aspect of neutrality tests is their important variance under the neutral model. This variance induces a large confidence interval and therefore drecreases their power to detect a deviation. It has been argued that this variance is a consequence of the tree shape variance and that neutrality tests based on the frequency spectrum are doomed to exhibit low power (Felsenstein, 1992b).

As a consequence, an ideal neutrality test should minimize its variance under the standard model. The variances of the denominator of previous neutrality tests are given in Table 2 (for $n = 30$ and $\theta = 1, 10, 100$). It is also important to mention that previous derivations of f, f^*, y and y^* variances give different values. Simulations show that the new derivations are the correct ones (Supplementary Table 1). First, it should be noted that the original D test has a very low variance when compared to all other tests. This is connected to the low variance of both $\hat{\theta}_S$ and $\hat{\theta}_\pi$. Second, Y and Y^* tests have also a small variance, although they ignore an important fraction of the data (*i.e.* singletons). All other tests have a similar variance.

This predicts that D typically will be sensitive to low, medium and high frequencies and should be more powerful because it has a relatively low variance under neutrality. Therefore, it has the potential to be an excellent neutrality test and it appears that it is often one of the most powerfull tests (Simonsen *et al.*, 1995; Fu, 1997). H is sensitive

either low or high frequencies, however its larger variance predicts that it will be useful only when the distortion in θ spectrum is very strong. In practice, it is powerful only when there is a large excess of high frequency polymorphisms. The singleton tests appear to be good candidates to capture an excess of singletons, although they neglect other deviations in the spectrum. The Y and Y^* tests have low variance, although ignoring singletons can lead to low power especially when they are in excess (Achaz, 2008).

Building new tests

To design new neutrality tests using this framework I started by analyzing the deviation of the average $\hat{\theta}_i$ spectrum, which is expected to be uniform under the standard models. Furthermore, because Fu (1995) showed that the covariance between ξ_i is weak when compared to their variance, visual inspection of the variance of $\hat{\theta}_i$ provides a first approximation to the expected variance of the $\hat{\theta}_\omega$ and therefore of their related T_ω tests. I studied two deviations from the standard model: a severe bottleneck and isolated populations with migration.

The severe bottleneck was simulated as a sudden change of size from N chromosomes to $N/100$ that lasts for a time $T_l = 0.1$ (in N generations). Accordingly, the coalescent rates within the bottleneck are accelerated by one hundredth and the simulations were performed as in Simonsen *et al.* (1995). Sampling was performed after a time T_b has elapsed after the bottleneck. The mean and the standard deviation of the $\hat{\theta}_i$ are given in the top panels of Figure 3 for two times $T_b = 0.03$ and $T_b = 0.3$. Figure 3 shows that most of the deviation comes from the sites with low frequency. Therefore, I designed a new test that captures the deviations within low frequencies. In this test, I used a first vector of $\omega_{1i} = e^{-\alpha i}$, with $\alpha = 0.9$ and a second uniform vector $\omega_{2i} = 1$. This results in an exponentially decreasing weight for low frequency mutations (Figure 3) that is positive for frequency $i/n \leq 0.13$. The choice of $\alpha = 0.9$ was mostly empirical, although, using $\alpha = 0.8$ or $\alpha = 1$ lead to similar results (data not shown). As stressed in the discussion, the present study aims at illustrating how easy it is to create new tests with enhanced power ; power

optimization deserves an entire new work. A graphical view of the $\mathbf{\Omega}$ vector associated with this new T_Ω test is given in Figure 3 and its variance is reported in Table 2. Most of the weight of this test is given to low frequencies and its variance is comparable to those of other neutrality tests. The power of this new tests along with the ones of D, F and H are reported in Figure 3. Results show that the new test outperforms the previous tests by 20% and is able to detect the deviation for a longer time.

The 95% confidence intervals were built using coalescent simulations under the standard model using a fixed number of segregating sites (Hudson, 1993; Depaulis and Veuille, 1998). Although there has been much debate on how confidence intervals should be set (Markovtsova *et al.*, 2001; Depaulis *et al.*, 2001; Wall and Hudson, 2001), it has been clearly shown that the choice of a particular method does not alter the results in standard models (Ramos-Onsins *et al.*, 2007) and therefore will not be discussed here.

In the second scenario, I compared the power of neutrality tests in detecting a case of isolation with migration (e.g. (Nielsen and Wakeley, 2001)). In the simulations, the isolation event happened at time $T_i = 3$ and both populations were sampled equally ($n_a = n_b = 15$). The migration rate between the two populations is variable. Similar to the analysis of the bottleneck, I first report the mean and the standard deviation of the $\hat{\theta}_i$ spectrum. Figure 4 show that most of the deviation comes from the sites at frequency 15/30. Additionally, for small enough migration rate ($M = 0.1$), there are almost no polymorphisms with frequency larger than 0.5. Although the standard deviations are large, the coefficient of variations (variance/mean) are relatively small. To design a new test, I used for the first $\boldsymbol{\omega}$ vector the probabilities given by a Binomial law $\omega_{1i} = \binom{i}{n} p^i (1-p)^{n-i}$ with $p = 0.5$ and $n = 30$ and a uniform vector $\omega_{2i} = 1$ as a second vector. This was motivated by the idea of designing a test that specifically captures an excess of medium frequency polymorphisms. A graphical view of the resulting $\mathbf{\Omega}$ vector is given in Figure 4 and its variance given in Table 2. Almost all the weight of this test is given to the $13 < i < 17$ sites. The variance of this new test is large, and this is to be related to the large variance of $\hat{\theta}_{n/2}$ in sample with even n. Despite this large variance, the test clearly outperforms all previous tests (Figure 4).

Overcoming the ascertainment bias

As an example of the power of designing new neutrality tests, I analyzed SNP data (from HapMap) around the Lactase gene (LCT) which has been shown to exhibit footprint of recent strong selective sweep in European populations (Bersaglieri *et al.*, 2004) as well in east african populations (Tishkoff *et al.*, 2007). This pattern of recent selection is one of the strongest in the human genome (Nielsen *et al.*, 2005). Indeed, its has been advanced that the lactase-persistance phenotype (the ability to digest milk as an adult) has been advantageous in European populations of farmers (especially in Northern European ones). The SNPs that are tightly associated with the selective sweep in Europeans are located at 13-22 kb upstream the gene start (Bersaglieri *et al.*, 2004). From HapMap (release 27 - feb 09) I gathered all SNP in a window of 100kb centered at the start of the lactase gene. This includes 50 kb upstream and the entire gene. I only considered SNP which sample size was at least 85 chromosomes. Because the sample size of all SNP was not identical, I used the observed frequencies to generate a folded frequency spectrum of 85 chromosomes for populations CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan) and YRI (Yoruban in Ibadan, Nigeria).

According to literature, one expects to find a trace of an ongoing selective events in the CEU population only. Without the help of an outgroup, this would translate into an excess of low frequency polymorphism in the folded frequency spectrum (typically, a negative D, F^* and D_2^*). Computation of the standard neutrality tests all show a deficit of low-frequency polymorphism rather than an excess. This deficit is even often significant (Table 3). This is clearly caused by the ascertainement bias in the dataset. Because the polymorphisms were first screened in a small group and further genotyped in larger groups, rare variants are underrepresented (e.g. Kuhner *et al.* (2000); Clark *et al.* (2005)). This ascertainment bias has been subject to various corrections (e.g., Wakeley *et al.* (2001); Nielsen *et al.* (2004)). To avoid any correction, I computed a T_Ω^* test where the weights of both ω_1^* and ω_2^* vectors was set to 0 for $i < 8$. The remaining two vectors are

computed using $\hat{\theta}_\pi$ and $\hat{\theta}_S$. As a consequence, this test is D-like in that it only considers polymorphisms with frequencies in the range $[0.09, 0.91]$. This is reminiscent of ignoring the singletons dataset where sequencing errors are suspected (Achaz, 2008). Results (Table 3) show that this test significantly deviates from the standard model for the CEU population. Ignoring less polymorphisms (e.g. only the 5% of low frequency) or changing the minimum sample size leads to similar results (data not shown).

Discussion

Here I developed a unifying framework for θ estimators based on the frequency spectrum. Namely, all known estimators of θ are linear combinations of $\hat{\theta}_i = i\xi_i$ (or $\hat{\theta}_i^* = \phi_i^{-1}\eta_i$). Because neutrality tests based on the frequency spectrum are simple functions of these θ estimators, the framework can be used to derive them. All tests (of this family) proposed so far are embedded in the framework. Using the model, I have shown that estimators of θ based on a folded spectrum always have an unfolded equivalent. The reciprocal however is not true.

Besides its unifying appeal, the model developed here can be used in several ways. First, I showed how it can be used to compute the variance of all estimators of θ and consequently of statistics such as $t = \hat{\theta}_1 - \hat{\theta}_2$. All variances of all estimators can be computed either using this framework or from their previous derived analytical formula. The same should be true for all t. Importantly, the computation of f, f^*, y and y^* revealed differences between both methods. Simulations demonstrate that the previous formulae were not correct while the new ones are. Besides a minor error in the f and f^* variance (corrected in Simonsen *et al.* (1995)), it appears that the $Cov[\pi, \xi_1]$ that was derived by Fu and Li (1993b) is inexact. Therefore the variances of f and f^* (Fu and Li, 1993b) as well as the variances of y and y^* (Achaz, 2008) that were using this covariance, carried along the error. Framed within the model presented here, all variances are correct. Finally, it can be used to compute the variance of h that was not given by the authors (Fay and Wu, 2000).

One potentially interesting development is to find an ω vector that minimize variance

of the associate estimator of θ. This problem has been previously addressed throughly (Felsenstein, 1992b,a; Fu and Li, 1993a; Fu, 1994b,a). Indeed, it has been shown that phylogenetic estimates have lower variance than estimators based on summary statistics (Felsenstein, 1992b; Fu and Li, 1993a; Fu, 1994b). Moreover, Fu (1994b,a) proposed a general method to find weight vectors that minimize the variance of the estimators and show that the best vector actually depends on the value of θ itself. Nonetheless, it remains true that some estimators have less variance than other (*i.e.* $\hat{\theta}_S$ *vs* $\hat{\theta}_\pi$) whatever is the value of θ. This latter observation suggest that re-exploring this question of minimizing the variance may be of interest.

Nawa and Tajima (2008) recently proposed to use the $\hat{\theta}_i^*$ spectrum instead of the classical frequency spectrum as a visual test for neutrality. This can be extended to the unfolded $\hat{\theta}_i$ spectrum if an outgoup is available. The study presented here fully supports this idea. The visual inspection of the $\hat{\theta}_i$ spectrum indicates why some tests will reject neutrality. Contrarily to what intuition may suggest, when one is interested in θ estimation, the appropriate representation for weight vectors is the $\boldsymbol{\omega}$ vector as defined above rather than weights on the ξ_i themselves (or on the η_i as in (Tajima, 1997)).

When an outgroup is used to unfold the spectrum, the choice of the appropriate outgoup is of critical importance. If the outgroup is not adequate (too distant or too close), mis-oriented sites will have a disastrous effect on θ estimations and therefore on related neutrality tests (Baudry and Depaulis, 2003). This adds to the difficulty of using tests based on the full ξ spectrum. However, when low and high frequencies can be sorted apart, much power is gained in terms of choosing the adequate evolutionary scenario. For example, no high frequencies are over-represented under recent growth or severe bottlenecks.

Specific problems that concern only some area of the spectrum can be handled easily by setting to 0 all weights in the suspicious area. For example, the sequencing errors can be avoided when the singletons are ignored (Achaz, 2008). With the current frawmework, by ignoring the low frequency polymorphisms, the ascertainment bias can be overcome and the pattern expected from selection at the lactase gene appears. This strategy has endless extensions as long as we have some prior knowledge of the suspicious area.

Finally, I think that this framework opens the door for new estimations of θ and the related neutrality tests. Using simple examples, I show how the power of neutrality tests can easily improved to detect deviations from the standard model. To optimize the power of the future new tests, one could (1) minimize their variance under the standard model, (2) select their area of sensitivity based on prior knowledge of the impact of specific deviations and (3) use recombination estimate to compute smaller confidence intervals (Wall, 1999) (as recombination results in quasi-independent replicates which lower the variance of the θ estimators). By building specific tests that will be sensitive to specific deviations, one could envision how several selected tests will be able to help the population geneticist to choose between different possible scenarios for a given dataset. Another interesting alternative would be to use the different θ estimators as summary statistics to infer the best parameters for a given evolutionary scenario (e.g. using ABC analysis).

Acknowledgment

The source code was designed as a C++ library for the simulations and a C library for sequence analysis and is available upon request. A dedicated web version of the tests is available at: `http://wwwabi.snv.jussieu.fr/achaz/neutralitytest.html`. Furthermore, the tests will be incorporated in future release of DNAsp. I would like to also thank F Tajima, EPC Rocha, J Wakeley, P Nicolas and D Higuet for their interesting comments on the manuscript and T Treangen for english improvement. I would also like to thank two anonymous reviewers for their constructive comments. This work was supported by the 07-GMGE-004-04 grant from the ANR (Agence Nationale de la Recherche).

References

Achaz, G., 2008 Testing for neutrality in samples with sequencing errors. Genetics **179**: 1409–1424.

Baudry, E., and F. Depaulis, 2003 Effect of misoriented sites on neutrality tests with outgroup. Genetics **165**: 1619–1622.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, *et al.*, 2004 Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet **74**: 1111–1120.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen, 2005 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res **15**: 1496–1502.

Depaulis, F., S. Mousset, and M. Veuille, 2001 Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol Biol Evol **18**: 1136–1138.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol **15**: 1788–1790.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155**: 1405–1413.

Felsenstein, J., 1992a Estimating effective population size from samples of sequences: a bootstrap monte carlo integration method. Genet Res **60**: 209–220.

Felsenstein, J., 1992b Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet Res **59**: 139–147.

Fu, Y. X., 1994a Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of dna sequences. Genetics **138**: 1375–1386.

Fu, Y. X., 1994b A phylogenetic estimator of effective population size or mutation rate. Genetics **136**: 685–692.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor Popul Biol **48**: 172–197.

Fu, Y. X., 1996 New statistical tests of neutrality for dna samples from a population. Genetics **143**: 557–570.

Fu, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**: 915–925.

Fu, Y. X., and W. H. Li, 1993a Maximum likelihood estimation of population parameters. Genetics **134**: 1261–1270.

Fu, Y. X., and W. H. Li, 1993b Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Hudson, R. R., 1993 *Mechanism of molecular evolution*, chapter The how and why of generating gene genealogies. Sinauer, Sunderland, Mass, 23–36.

Kuhner, M. K., P. Beerli, J. Yamato, and J. Felsenstein, 2000 Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics **156**: 439–447.

Markovtsova, L., P. Marjoram, and S. Tavaré, 2001 On a test of depaulis and veuille. Mol Biol Evol **18**: 1132–1133.

Nawa, N., and F. Tajima, 2008 Simple method for analyzing the pattern of dna polymorphism and its application to snp data of human. Genes Genet Syst **83**: 353–360.

Nielsen, R., M. J. Hubisz, and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168**: 2373–2382.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a markov chain monte carlo approach. Genetics **158**: 885–896.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, *et al.*, 2005 Genomic scans for selective sweeps using snp data. Genome Res **15**: 1566–1575.

Ramos-Onsins, S. E., S. Mousset, T. Mitchell-Olds, and W. Stephan, 2007 Population genetic inference using a fixed number of segregating sites: a reassessment. Genet Res **89**: 231–244.

Ramos-Onsins, S. E., and J. Rozas, 2002 Statistical properties of new neutrality tests against population growth. Mol Biol Evol **19**: 2092–2100.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for dna polymorphism data. Genetics **141**: 413–429.

Tajima, F., 1983 Evolutionary relationship of dna sequences in finite populations. Genetics **105**: 437–460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics **123**: 585–595.

Tajima, F., 1997 *Progess in Population Genetics and Human Evolution*, chapter Estimation of the amount of DNA polymorphism and statistical tests of the neutral mutation hypothesis based on DNA polymorphism. Springer-Verlag, 149–164.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, *et al.*, 2007 Convergent adaptation of human lactase persistence in africa and europe. Nat Genet **39**: 31–40.

Wakeley, J., 2009 *Coalescent theory, an Introduction*. Roberts and Company.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero, and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms–and inferences about human demographic history. Am J Hum Genet **69**: 1332–1347.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genetical research **74**: 65–79.

Wall, J. D., and R. R. Hudson, 2001 Coalescent simulations and statistical tests of neutrality. Mol Biol Evol **18**: 1134–1135.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7**: 256–276.

Figure Legends

Figure 1 - Estimators of θ

A graphical view of the weight vectors of four typical estimators of θ (for $n = 30$). All values of the normalized vector sum to 1. In the top four panels, the $\boldsymbol{\omega}$ vectors that are defined for the unfolded frequency spectrum ($\boldsymbol{\xi}$) are given, whereas the two bottom ones are the $\boldsymbol{\omega^*}$ vectors that are defined for the folded frequency spectrum ($\boldsymbol{\eta}$). For estimators that can be defined in terms of both $\boldsymbol{\omega}$ and $\boldsymbol{\omega^*}$ (here $\hat{\theta}_\pi$ and $\hat{\theta}_S$), the latter can be computed from the former with $\omega_i^* = \omega_i + \omega_{n-i}$ (when $i \neq n - i$) or $\omega_i^* = \omega_i$ (when $i = n - i$).

Figure 2 - Neutrality tests

A graphical view of the weight vectors of four typical neutrality tests (for $n = 30$). Because the $\boldsymbol{\Omega}$ vectors used for neutrality tests are computed as a difference between two normalized vectors, all values of $\boldsymbol{\Omega}$ sum to 0. In the top four panels, the $\boldsymbol{\Omega}$ vectors that are defined for the unfolded frequency spectrum ($\boldsymbol{\xi}$) are given, whereas the two bottom ones are the $\boldsymbol{\Omega^*}$ vectors that are defined for the folded frequency spectrum ($\boldsymbol{\eta}$). For estimators that can be defined in terms of both $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega^*}$ (here D and F^*), the latter can be computed from the former in the way that ω_i^* can be deduced from ω_i.

Figure 3 - Example of a Severe Bottleneck

The top two panels reports the mean and the standard deviation of the $\hat{\theta}_i$ spectrum that is observed in simulations ($n = 30$, 10^4 replicates) of a standard model or of a recent severe bottleneck (reduction of $f = 1/100$ for a time $T_l = 0.1$). In both times after the bottleneck ($T_b = 0.03$ and $T_b = 0.3$), the observed trend is similar: excess in low frequency of $\hat{\theta}_i$, though stronger for $T_b = 0.03$. In the left bottom panels, the weight vector of a new neutrality test (here T_Ω) is reported. It focuses its sensitivity on low frequencies: $\Omega_i = \frac{e^{-0.9i}}{\sum_j e^{-0.9j}} - \frac{1}{30}$. In the right bottom panel the power of four neutrality tests is compared in detecting a severe bottleneck as a function of the time elapsed after the bottleneck. The new test

shows enhanced power to detect the bottleneck (more power for a longer time).

Figure 4 - Isolation with Migration

The top two panels report the mean and the standard deviation of the $\hat{\theta}_i$ spectrum that is observed in simulations ($n = 30$, 10^4 replicates) of a standard model or of a isolation with migration model (two populations equally sampled $n_a = n_b = 15$ that were a single ancestral panmictic population at time $T_i = 3$). In both sampling migration rates between the two populations ($M = 0.1$ and $M = 1$), the observed trend is similar: an excess of $\hat{\theta}_{15}$, though much stronger for $M = 0.1$. In the left bottom panel, the weight vector of a new neutrality tests (T_ω), that focuses its sensitivity on $i = 15$, is shown. The weight vector used here is: $\Omega_i = \frac{\binom{30}{i}0.5^{30}}{\sum_j \binom{30}{j}0.5^{30}} - \frac{1}{30}$, where $\binom{30}{i}0.5^{30}$ is obtained using a Binomial with $p = 0.5$ and $n = 30$. In the right bottom panel the power of four neutrality tests is compared when detecting the population structure as a function of the migration rate. The new test displays much more power to detect the population structure.

Table 1: Basic characteristics of previous estimators of θ

estimators	$\boldsymbol{\omega}$	$\boldsymbol{\omega^*}$ (when defined)	Variance $(n=30)$ $\theta=1$	$\theta=10$	$\theta=100$
$\hat{\theta}_S$	$\omega_i = i^{-1}$	$\omega_i^* = \frac{n}{i(n-i)(1+\delta_{i,n-i})}$	0.36	12.8	1,052
$\hat{\theta}_\pi$	$\omega_i = n-i$	$\omega_i^* = \frac{n}{(1+\delta_{i,n-i})}$	0.59	27.4	2,419
$\hat{\theta}_{\eta_1}$	$\omega_1 = (n-1)$, $\omega_{n-1} = 1$, $\omega_{1<i<n-1} = 0$	$\omega_1^* = n$, $\omega_{i>1}^* = 0$	1.22	35.1	2,839
$\hat{\theta}_{S-\eta_1}$	$\omega_1 = \omega_{n-1} = 0$, $\omega_{1<i<n-1} = i^{-1}$	$\omega_1^* = 0$, $\omega_{i>1}^* = \frac{n}{i(n-i)(1+\delta_{i,n-i})}$	0.52	21.4	1,833
$\hat{\theta}_{\pi-\eta_1}$	$\omega_1 = \omega_{n-1} = 0$, $\omega_{1<i<n-1} = (n-i)$	$\omega_1^* = 0$, $\omega_{i>1}^* = \frac{n}{(1+\delta_{i,n-i})}$	0.68	31.9	2,825
$\hat{\theta}_{\xi_1}$	$\omega_1 = 1$, $\omega_{i>1} = 0$	-	1.15	25.0	1,599
$\hat{\theta}_H$	$\omega_i = i$	-	1.55	65.0	5,597
$\hat{\theta}_{S-\xi_1}$	$\omega_1 = 0$, $\omega_{i>1} = i^{-1}$	-	0.51	20.3	1,730
$\hat{\theta}_{\pi-\xi_1}$	$\omega_1 = 0$, $\omega_{i>1} = (n-i)$	-	0.68	31.5	2,790

Table 2: Basic characteristics of neutrality tests

Test	$\hat{\theta}_1$	$\hat{\theta}_2$	Mandatory outgroup	Variance $\theta = 1$	Variance $\theta = 10$	Variance $\theta = 100$
d	$\hat{\theta}_\pi$	$\hat{\theta}_S$		0.18	8.2	728
f	$\hat{\theta}_\pi$	$\hat{\theta}_{\xi_1}$	no	1.62	51.9	4,084
d_2	$\hat{\theta}_S$	$\hat{\theta}_{\xi_1}$		0.93	25.8	1,910
y	$\hat{\theta}_{\pi_{-\xi_1}}$	$\hat{\theta}_{\pi_{-\xi_1}}$		0.12	6.2	558
h	$\hat{\theta}_\pi$	$\hat{\theta}_H$		0.98	40.0	3,417
f^*	$\hat{\theta}_\pi$	$\hat{\theta}_{\eta_1}$	yes	1.71	63.8	5,314
d_2^*	$\hat{\theta}_S$	$\hat{\theta}_{\eta_1}$		0.99	34.5	2,805
y^*	$\hat{\theta}_{\pi_{-\eta_1}}$	$\hat{\theta}_{S_{-\eta_1}}$		0.12	5.8	524
	$\omega_{1i} = e^{-0.9i}$	$\omega_{2i} = 1$		1.19	37.1	2,895
T_Ω	$\omega_{1i} = \binom{30}{i}0.5^{30}$	$\omega_{2i} = 1$	yes	2.48	151.4	14,167

Table 3: Neutrality tests in the lactase region (*: $P < 0.05$; **: $P < 0.01$)

Population	D	D_2^*	F^*	T_{Ω^*}
CEU	-0.16	0.81	0.48	-1.79*
CHB	2.04*	1.97**	2.29*	0.25
JPT	0.92	2.22**	1.93*	0.28
YRI	1.00	2.18**	1.94*	-0.04

Table 4: Supplementary Table 1 - Variance of neutrality tests ($\theta = 10$, $n = 30$, 10^5 replicates). Please note that for F^*, the typo Fu and Li (1993b) equation was corrected, as it is in Simonsen *et al.* (1995).

Test	Publication	Original Derivation	Current Framework	Simulation
d	Tajima (1989)	8.21	8.21	8.24
d_2	Fu and Li (1993b)	25.83	25.83	25.89
d_2^*	Fu and Li (1993b)	34.47	34.47	34.58
h	Fay and Wu (2000)	na	39.96	39.55
f^*	Fu and Li (1993b)	61.44	63.84	63.98
f	Fu and Li (1993b)	49.59	51.85	51.83
y^*	Achaz (2008)	7.72	5.81	5.80
y	Achaz (2008)	7.40	6.18	6.17

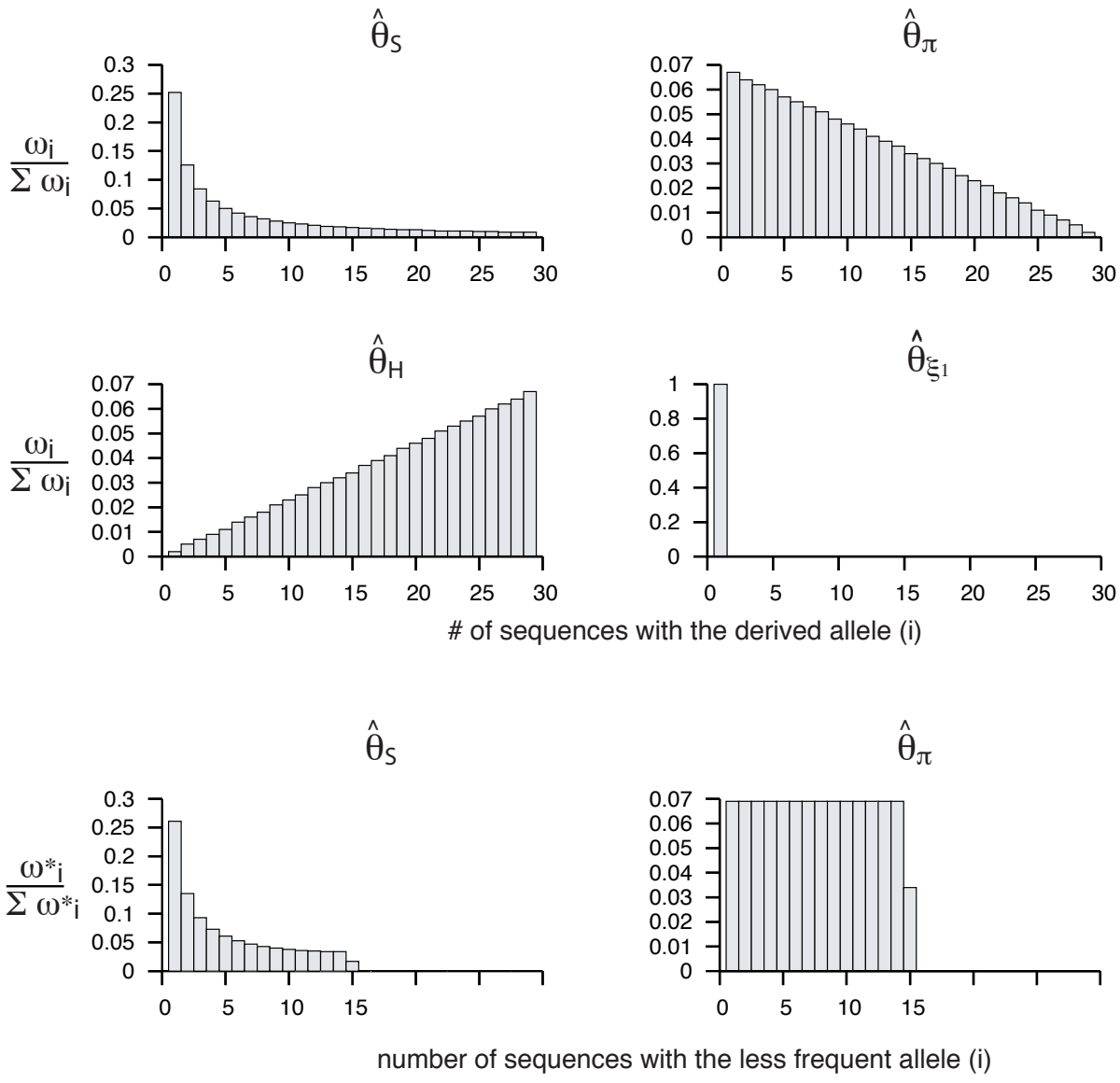


Figure 1: Normalized vectors of four typical estimators of θ

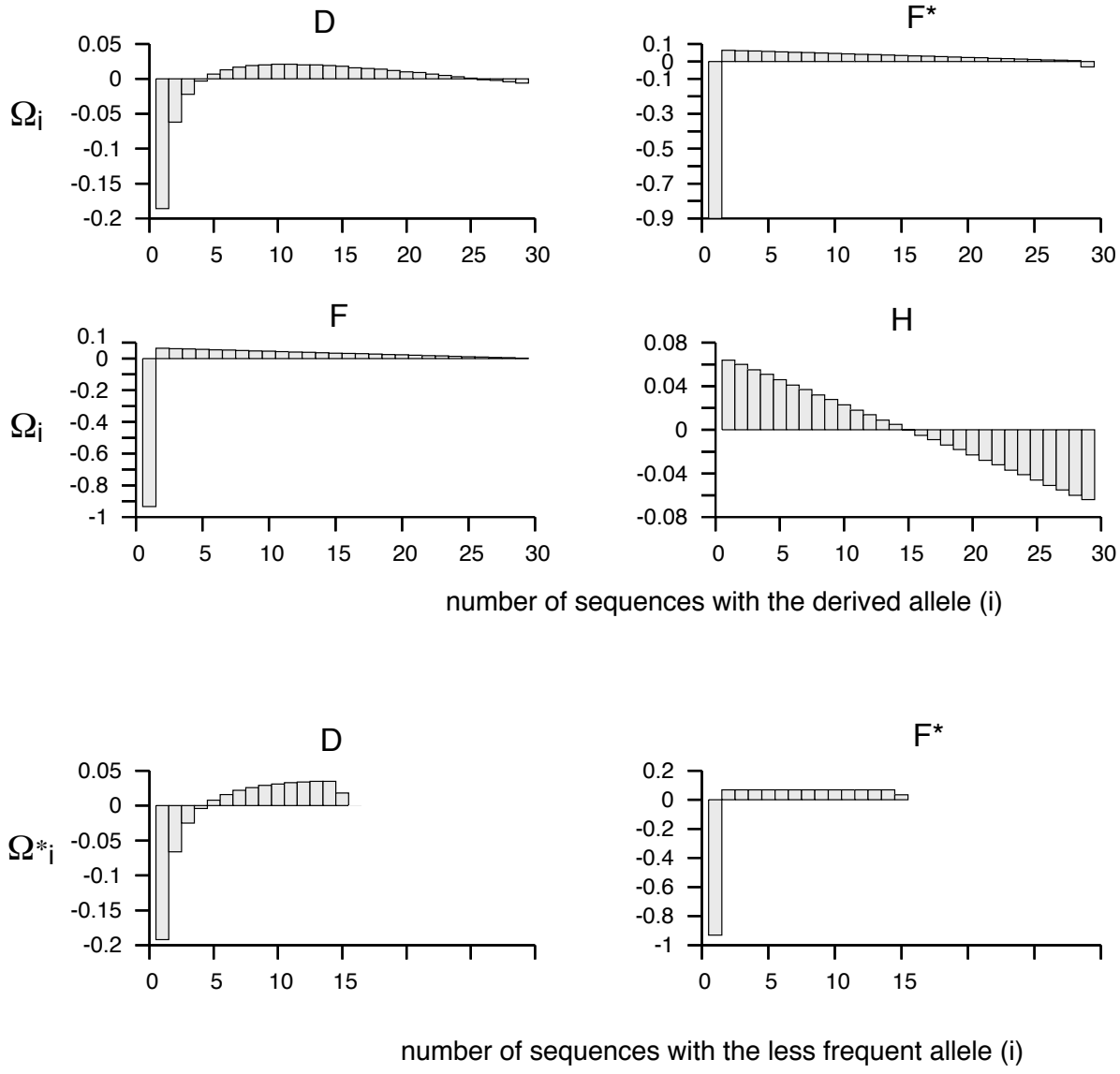


Figure 2: Normalized vectors of four typical neutrality tests

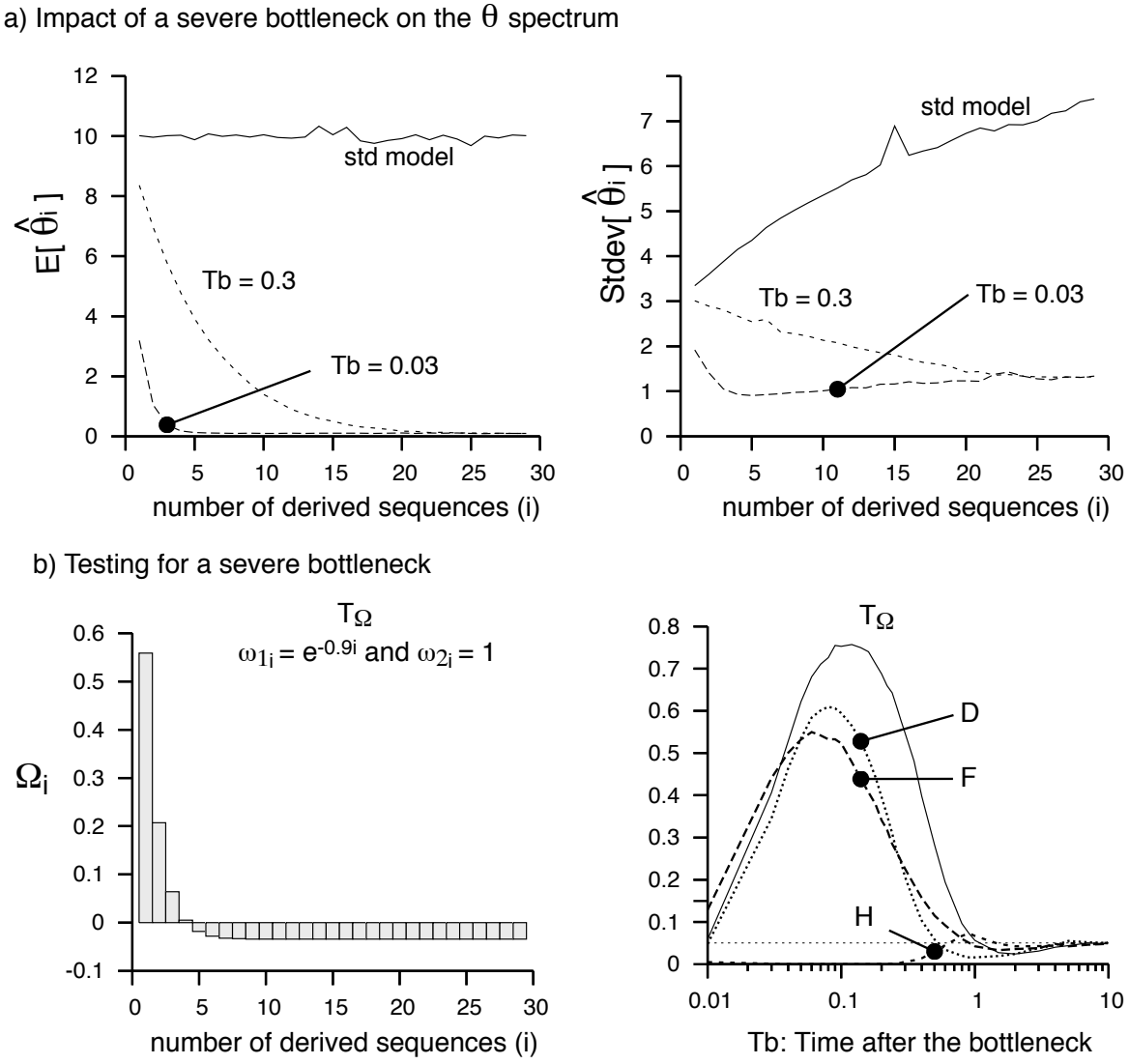


Figure 3: Severe bottleneck

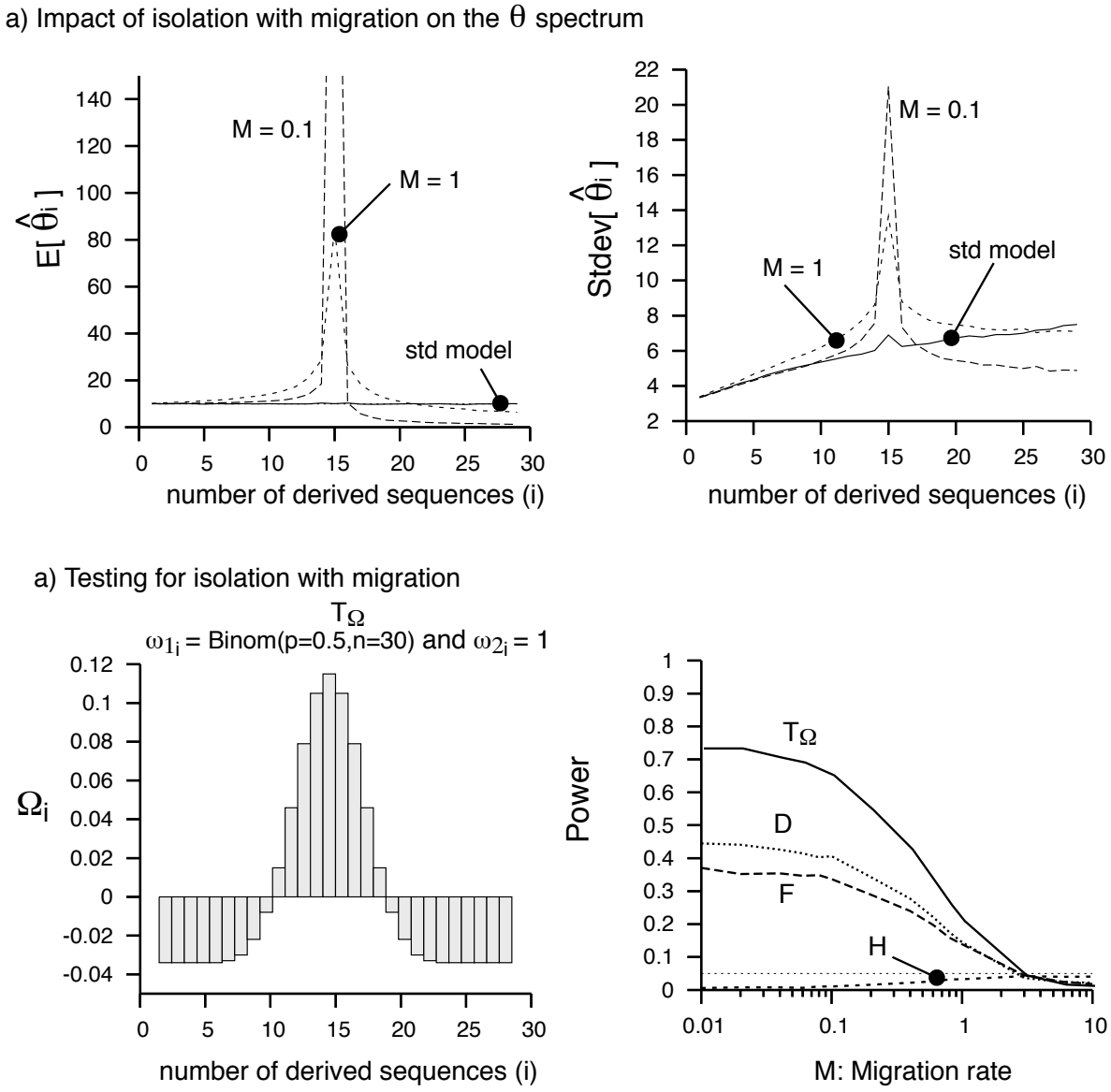


Figure 4: Isolation with migration