UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE BIOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

TESE DE DOUTORADO

EVOLUÇÃO EM POPULAÇÕES AMERÍNDIAS: ASPECTOS ESTOCÁSTICOS, ADAPTATIVOS E CULTURAIS

CARLOS EDUARDO GUERRA AMORIM

ORIENTADOR: PROF. FRANCISCO MAURO SALZANO

CO-ORIENTADOR: PROF. SANDRO LUÍS BONATTO

COLABORADOR: PROF. LAURENT EXCOFFIER

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE BIOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

EVOLUÇÃO EM POPULAÇÕES AMERÍNDIAS: ASPECTOS ESTOCÁSTICOS, ADAPTATIVOS E CULTURAIS

TESE SUBMETIDA AO PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR DA UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL COMO REQUISITO PARCIAL PARA A OBTENÇÃO DO TÍTULO DE DOUTOR EM CIÊNCIAS

CARLOS EDUARDO GUERRA AMORIM

ORIENTADOR: PROF. FRANCISCO MAURO SALZANO

CO-ORIENTADOR: PROF. SANDRO LUÍS BONATTO

COLABORADOR: PROF. LAURENT EXCOFFIER

Porto Alegre, Julho de 2013

Este trabalho foi desenvolvido no Laboratório de Evolução Humana e Molecular da Universidade Federal do Rio Grande do Sul, no Laboratório de Biologia Genômica e Molecular da Pontifícia Universidade Católica do Rio Grande do Sul, no *Computational and Molecular Population Genetics Laboratory* da *Universität Bern* e no *Computational Population Genetics Group* do *Swiss Institute of Bioinformatics* entre julho de 2009 e junho de 2013, com o financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Programa de Amparo a Núcleos de Excelência (PRONEX).

À minha família

I believe in intuition and inspiration. … At times I feel certain I am right while not knowing the reason. When the eclipse of 1919 confirmed my intuition, I was not in the least surprised. In fact, I would have been astonished had it turned out otherwise. Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution. It is, strictly speaking, a real factor in scientific research.

– Albert Einstein *in* "Cosmic Religion and Other Opinions and Aphorisms", p. 97

AGRADECIMENTOS

Comecei este trabalho em julho de 2009. Desde então algumas pessoas foram essenciais para que eu chegasse ao ponto em que me encontro agora. Algumas dessas pessoas me ensinaram a fazer ciência, outras me deram atenção e suporte emocional e também serviram como modelos, como um exemplo a ser seguido. A todas essas pessoas eu tenho muita gratidão.

A mais importante delas foi o meu orientador, o Professor Francisco M. Salzano, quem desde o início foi muito aberto para as minhas idéias e investiu seu tempo (e muito mais do que isso!) para que eu conseguisse crescer e explorar territórios desconhecidos. Foi ele que direta ou indiretamente me guiou durante todo esse caminho, me ensinando desde o básico da Genética de Populações até os menores detalhes acerca das populações nativas americanas, sempre sendo muito rigoroso, detalhista e motivador. Seu trabalho árduo e diligência, sua seriedade, coragem e genialidade, bem como sua amabilidade são certamente características que vou lembrar para sempre. Espero que pelo menos parte delas permaneça em mim além da minha memória e que possam ser tomadas como um exemplo para minha carreira acadêmica.

Não menos importante é a minha família. Em todas pequenas vitórias, desde a aprovação na prova de seleção do doutorado até os últimos meses deste trabalho, meus pais me acompanharam com muita alegria e satisfação. A eles agradeço poder ter sido capaz de investir todo o tempo que pude neste trabalho e por ter feito isso com a máxima tranquilidade e segurança. Junto a eles, agradeço à minha amada irmã pela motivação e companheirismo e o restante dos meus familiares pelo suporte (pontual, mas essencial), em especial ao meu tio-avô Félix, quem há muito tempo com palavras simples me trouxe grande motivação para seguir meu caminho.

Gostaria de agradecer ao meu co-orientador, o Professor Sandro L. Bonatto, por ter me introduzido na Biologia Computacional e por ter me proporcionado as melhores condições para o meu amadurecimento na área. Seu nível intelectual excelente e maleabilidade certamente foram um estímulo para o meu crescimento. Na PUCRS, fui recebido da melhor maneira possível e por isso também devo agradecer à equipe do Laboratório de Biologia Genômica dessa instituição que me recebeu de portas abertas.

Obrigado ao Professor Laurent Excoffier, meu orientador no estágio sanduíche. Sua ajuda inicial com as várias questões burocráticas relativas à minha mudança para Berna e o estímulo intelectual constante foram essenciais na minha formação. Obrigado por ter me recebido da melhor maneira possível em Berna, por ter sido extremamente paciente ao me ensinar, e por ter me colocado em contato com metodologias de ponta na área da genômica populacional. Grande parte do meu bem-estar em Berna deve-se ao cuidado que este professor teve comigo e, indiretamente, com o nosso grupo de pesquisa aqui no Brasil. Ao grupo coordenado por ele na Universidade de Berna, o *Computational and Molecular Population Genetics Laboratory*, agradeço por todos os ensinamentos, em especial à Heidi Lischer e à Isabel Dupanloup pelo suporte constante, à Josephine Daub pela colaboração e ensinamentos na área de matemática, e muito especialmente à Isabel Alves por toda nossa amizade. Obrigado também ao Gerald, Vanessa, Matthieu, Anna, Robert, Matthias, Miriam, Stephan, Tom e Susan pela excelente atmosfera de trabalho.

À professora Maria Cátira Bortolini, agradeço os estímulos intelectuais e motivação constantes. Sua criatividade e inteligência são exemplares e, no meio das adversidades relativas ao meu doutorado e vida pessoal, sua personalidade amável me salvou de diversos problemas. Certamente esta Tese não teria chegado a este ponto sem a sua ajuda e dedicação. A esta professora, juntamente ao professor Aldo M. Araújo, agradeço pelos valiosos comentários no meu seminário de qualificação.

Muito obrigado também à Dra. Tábita Hünemeier, uma grande amiga, colaboradora e professora, alguém que me deu as boas-vindas em Porto Alegre e que constantemente me estimula a fazer novas pesquisas, sem contar a ajuda técnica, metodológica e teórica que ela sempre me ofereceu durante todos os momentos da realização deste trabalho.

Como será demonstrado nesta tese, o processo de povoamento de um novo território é um processo complexo e árduo. Aos meus amigos Pietra, Reinaldo, Nina, Mariana Botton, Evandro, Clarissa, Rodrigo, Bernardo, Graziela, Enzo, Hudson, Renata Renz, Renata Maieron, Luciana Marques, Susana, Luciano, Adriano e Ana Paula, tenho muita gratidão por terem me dado todo o suporte necessário para que o meu processo de "povoamento" de Porto Alegre tenha sido muito bem sucedido. Nesse âmbito, tenho enorme gratidão pelo amigo Rogério por ter me acompanhado por muito tempo nessa jornada, ter feito a maior parte das ilustrações desta tese e ter me ajudado incondicionalmente durante todos os momentos; à "irmã" Luciana Tissot, por ter me oferecido todo o conforto e segurança necessários na reta final, e à colega Camila Zanella, no início deste percurso. Também gostaria de agradecer aos amigos e professores de Três Coroas, em especial à Lama Sherab Drolma e Michael.

Obrigado ao pessoal do Laboratório de Evolução Humana e Molecular, Vanessa, Clênio, Rafael, Eli, Carla, Caio, Carlos, Virgínia, Pamela, Lucas, Agatha e Pedro pelo convívio tranquilo e suporte; ao Elmo e à Laci pela prontidão e seriedade; e aos demais colegas e professores do PPGBM, em especial à professora Mara H. Hutz, por permitir o acesso ao seu laboratório e uso de seus equipamentos num estágio inicial deste trabalho, à professora Eliane Bandinelli pela disponibilidade em ajudar nos vários momentos da realização desta tese, bem como à professora Sídia Maria Callegari-Jacques pelo suporte estatístico.

Todo esse processo não seria possível sem os auxílios e bolsas que as seguintes instituições de fomento nos concederam: Conselho Nacional de Desenvolvimento Científico e Tecnológico, Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, Programa de Apoio a Núcleos de Excelência, e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

E, por fim, muito obrigado a todos os meus amigos de Brasília, que me ajudaram a percorrer este longo caminho com simples gestos e companheirismo: Ivan, Thiago, Patrícia, Lula, André, Fabíola, Mary Jane, Manuela, Alice, Priscila, Luana, Domitila, Juliana Freitas, Juliana Gama, Chipe, Juca, Victor Hugo, Marília, June, Erika, Isolda, Denise, Symone, Renata, Lucas, Xan, Alisson, Édi, Carol e Silviene.

Sem a participação de cada uma dessas pessoas e instituições certamente essa jornada teria sido muito mais complicada. A todos eles gostaria de demonstrar a minha sincera e imensa gratidão.

Mestre não é quem sempre ensina, mas quem de repente aprende.

– Riobaldo Tatarana *in* "Grande Sertão: Veredas"

SUMÁRIO

RESUMO

Para o melhor entendimento da biologia de uma espécie ou população, faz-se necessária a análise detalhada das forças evolutivas que moldaram sua diversidade genética. Mecanismos evolutivos randômicos e direcionais devem ser entendidos separadamente e em sua interação para que uma interpretação mais acurada possa ser realizada. No que concerne a humanos, uma camada adicional de complexidade deve ser considerada: a cultura. No presente trabalho, aspectos da história evolutiva de populações nativas americanas são analisados levando em conta os efeitos da deriva genética, história demográfica e seleção natural sobre sua diversidade genética e estrutura de recombinação (desequilíbrio de ligação, DL). Adicionalmente, é realizada uma comparação direta entre a evolução cultural e a evolução biológica. O corpo principal da tese é composto por quatro artigos que visam de maneira geral ao entendimento desses fatores em populações humanas atuais, mas que, no contexto da presente tese, serão discutidos com foco nos ameríndios. Os resultados desses artigos podem ser resumidos como seguem:

1) Amorim *et al.* (2011) *X-chromosomal genetic diversity and linkage disequilibrium patterns in Amerindians and non-Amerindian populations* (*Am J Hum Biol* 23: 299-304). Os resultados deste trabalho revelam baixa diversidade genética no cromossomo X aliada à alta proporção de *loci* em DL em populações ameríndias, quando comparadas a grupos com ancestralidade genética distinta. A deriva genética seria o principal candidato a agente causal para o padrão observado.

2) Amorim *et al. Detecting Genome-wide Signals of Human Adaptation to Tropical Forests in a Convergent Evolution Framework* (manuscrito em preparação). Aqui analizaram-se SNPs distribuídos nos autossomos e no cromossomo X para a detecção de sinais de seleção positiva. Populações amazônicas e africanas vivendo na floresta tropical foram comparadas com outras que viviam em ambiente não-florestal. Os resultados apontam para a existência de seleção positiva especialmente sobre genes aparentemente relacionados à imunidade, fluxo de colesterol e altura corporal, entre outros.

3) Amorim *et al*., (2013) *A Bayesian Approach to Genome/Linguistic Relationships in Native South Americans* (*PLoS ONE* 8: e64099). Neste trabalho avaliou-se o ajuste de modelos demográficos, construídos a partir de classificações linguísticas, à diversidade genômica de populações da América do Sul. As análises revelam uma maior adequação da classificação proposta por Joseph Greenberg em 1987. De acordo com esse cenário, o ancestral comum dos principais grupos linguísticos da América do Sul teria uma idade de cerca de 3,1 mil anos e a separação mais recente entre esses grupos teria ocorrido há 2,8 mil anos, entre os Tupi e os Aruaque. Os resultados sugerem ainda que, neste contexto, línguas e genes apresentam taxa de evolução semelhante.

4) Amorim *et al. Differing evolutionary histories of the ACTN3*R577X polymorphism among the major human geographic groups* (manuscrito em preparação). Neste artigo a diversidade genética do gene *ACTN3* e mais especificamente de sua variante funcional rs1815739 em populações ameríndias foi comparada às de outros continentes, revelando um cenário evolutivo que sugere que este gene foi alvo de seleção positiva no passado,

mas atualmente o sinal desse processo foi apagado ou suavizado nas Américas.

Esses resultados em conjunto sugerem que fatores biológicos seletivos e neutros foram ambos importantes durante o povoamento do Novo Mundo e destacam a importância de analisá-los em conjunto com características culturais, reconhecendo as peculiaridades de cada um e a interação entre eles.

ABSTRACT

For a better understanding of the biology of a species or population, it is necessary to analyze the forces that shaped its genetic diversity in detail. Neutral and selective evolutionary mechanisms should be interpreted independently and in their interaction for a better interpretation of facts. Regarding human evolution, an additional level of complexity should be considered: culture. In the present work, some aspects of the evolutionary history of Native American populations is considered in relation to the effects of genetic drift, demography, and natural selection upon their genetic diversity and recombination structure (linkage disequilibrium, LD). Additionally, a direct comparison between cultural and biological evolution is made. The nucleus of this Thesis is composed by four articles that aim at the understanding of the role of these factors in the evolution of extant human populations, but for the purposes of this Thesis they will be discussed with a main focus in Amerindians. The results of these articles can be briefly summarized as follows:

1) Amorim *et al.* (2011*) X-chromosomal Genetic Diversity and Linkage Disequilibrium Patterns in Amerindians and non-Amerindian Populations* (*Am J Hum Biol* 23: 299-304). The results of this work reveal low X-chromosomal genetic diversity and high proportion of loci in linkage disequilibrium when Amerindian populations were compared to other groups with a distinct genetic background. Genetic drift is the best candidate for generating the observed pattern.

2) Amorim *et al. Detecting Genome-wide Signals of Human Adaptation to Tropical Forests in a Convergent Evolution Framework* (manuscript in preparation). Here a positive selection analysis of autosomal and X-chromosomal SNPs was

conducted. Amazonian and African tropical forest populations were compared to those living outside forests. The results suggest the action of positive selection especially upon genes related to immunity, cholesterol cellular flux and body height, among others.

3) Amorim *et al*., (2013) *A Bayesian Approach to Genome/Linguistic Relationships in Native South Americans* (*PLoS ONE* 8: e64099). In this work, we tested the fit of current genomic South Amerindian diversity to demographic models based on linguistic classifications. The analyses revealed a better fit of the classification proposed by Joseph Greenberg in 1987. According to this scenario, the common ancestor of the main South American linguistic groups would have an age of circa 3.1 thousand years before present (BP) and the most recent fission between these groups would be that of the Tupí and Arawakan at 2.8 thousand years BP. The results also suggest that, in this context, language and genes evolve at a similar pace.
4) Amorim *et al. Differing evolutionary histories of the ACTN3*R577X polymorphism among the major human geographic groups* (manuscript in preparation). The genetic diversity of the *ACTN3* gene, and more specifically of its functional variant rs1815739 in Amerindian populations was compared to these of other continents, revealing an evolutionary scenario that suggests that this gene might have been a target of positive selection in the past; currently however, the signal of this process was erased or smoothed in the Americas.

These results suggest that selective and neutral biological evolutionary factors were important during the settlement of the New World and highlight the importance of

analyzing them together with cultural characteristics, considering their peculiarities and the interaction between them.

PARTE I

I.I) APRESENTAÇÃO

O entendimento da origem e da manutenção da diversidade genética em humanos e outros organismos em escala local e genômica tem aplicações em vários domínios da ciência e tecnologia, como, por exemplo, na genética médica, no melhoramento vegetal e na biologia evolutiva. Ignorar a ação da seleção natural sobre essa diversidade pode levar a um desentendimento da biologia dos organismos estudados em sua totalidade e de suas relações com o meio-ambiente e outros organismos. Da mesma forma, o desconhecimento das peculiaridades da história demográfica de uma população e da ação dos mecanismos randômicos de manutenção da diversidade genética pode levar a conclusões precipitadas acerca dos processos adaptativos dos organismos em questão. Por exemplo, estudos de associação e varreduras totais do genoma podem falhar na identificação de *loci* candidatos, relatando associações espúrias como verdadeiras, pois grande parte da variação genética humana está sujeita a ampla variação inter-populacional em decorrência dos efeitos de mecanismos apenas neutros (Hoffer *et al*., 2009) e boa parte dos estudos de seleção natural e associação são baseados na identificação de *loci* com distribuição aberrante (Beaumont, 2005); além do que, em alguns casos, mecanismos puramente neutros podem mimetizar os efeitos da seleção positiva (Excoffier e Ray, 2008). Apesar disso, um número crescente de estudos tem demonstrado que parte da genética de nossa espécie foi moldada por mecanismos adaptativos (Sabeti *et al.,* 2007; López Herráez *et al*., 2009; Williamson *et al.,* 2007; Coop *et al*., 2010; Daub *et al.*, 2013; Grossman *et al.,* 2013). Assim sendo, compreender as peculiaridades da história evolutiva de uma espécie ou população, separando os efeitos de mecanismos evolutivos direcionados daqueles de

ação randômica é um primeiro passo para o enriquecimento do conhecimento acerca dessa população ou espécie.

No que concerne aos estudos de evolução humana, esse problema adquire uma nova dimensão: a cultura apresenta uma dinâmica evolutiva peculiar e interdependente da evolução biológica. Atualmente tem-se disponível um número considerável de trabalhos que relatam casos em que a cultura e a biologia interagem, estudos que demonstram que ambas evoluem e exemplos de eventos em que uma se adapta à outra e modifica suas pressões seletivas (Laland e Brown, 2002; Richerson e Boyd, 2005; Reali e Griffiths, 2010; Tovo-Rodrigues *et al*., 2010; Hünemeier *et al.,* 2012a, b).

No presente trabalho, aspectos da história evolutiva de populações nativas americanas são analisados levando em conta os efeitos da deriva genética, demografia e seleção natural sobre sua diversidade genética e estrutura de recombinação (desequilíbrio de ligação, DL). Adicionalmente, uma comparação direta entre a evolução cultural e a evolução biológica é estabelecida de forma a contribuir para a compreensão da história de populações autóctones da América do Sul. Em última instância, o trabalho visa à compreensão detalhada dos mecanismos evolutivos aos quais os nativos americanos estiveram submetidos ao longo de sua história, inclusive no momento posterior à sua entrada nas Américas.

O corpo principal desta tese está estruturado em três partes: I) esta apresentação, junto a uma breve revisão bibliográfica e o relato dos objetivos do trabalho; 2) uma compilação de quatro artigos – publicados ou em preparação –, que versam em geral sobre evolução humana, mas que no contexto desta tese serão usados para a discussão da história evolutiva dos ameríndios; e 3) conclusões gerais de acordo com essa

perspectiva. Em anexo está reproduzido um trabalho realizado com a minha colaboração, que foi utilizado para a discussão dos artigos e redação da tese.

I.II) UMA BREVE INTRODUÇÃO À PRÉ-HISTÓRIA AMERICANA

Um dos poucos pontos consensuais no que concerne à pré-história das Américas é a origem asiática de seus primeiros habitantes humanos (Salzano, 2007). O processo de povoamento desse continente é entendido atualmente como tendo ocorrido de acordo com três estágios (Kitchen *et al.*, 2008). O primeiro envolve um gargalo populacional em decorrência da saída de uma parcela de uma população do leste asiático – provavelmente da região da Sibéria (Bortolini *et al.,* 2003) – em direção à Beríngia (correspondente hoje à região onde se localiza o estreito de Bering). A este primeiro estágio, está associada a redução da diversidade genética da população migrante, devido ao fato de que esta representa uma amostragem possivelmente não-aleatória da população-fonte.

Na Beríngia, a população migrante permanece isolada por cerca de 5.000 anos (Fagundes *et al.*, 2008), o que consiste no segundo estágio do processo de povoamento do continente (Kitchen *et al.*, 2008). Neste estágio, a população passa por um novo gargalo populacional por conta da redução do tamanho populacional (possivelmente devido à degradação do ambiente decorrente da ocupação prolongada do território e limitação natural de recursos) e do isolamento (uma vez que massas de gelo impediam o a entrada na América do Norte por uma rota interna). Estudos sugerem que este gargalo pode ter sido bastante severo, com o tamanho efetivo da população tendo sido estimado em cerca de no mínimo mil mulheres (Fagundes *et al*., 2008).

Com o aumento da temperatura do planeta após o último glacial máximo há cerca de 20.000 anos (Yokoyama *et al*., 2000), as massas de gelo que bloqueavam o acesso ao

continente americano se derreteram permitindo o acesso de humanos ao seu interior, o que consiste no terceiro estágio do povoamento do continente americano (Kitchen *et al*., 2008). É possível que rotas costeiras também tenham sido usadas em maior ou menor grau – como aparentemente ocorreu na América do Sul (Dillehay *et al.,* 2008) – dependendo ou não da tecnologia de navegação marítima incipiente da época. Essa entrada teria ocorrido via América do Norte há cerca de 18 mil anos (Fagundes *et al*., 2008).

Membros da nossa espécie já estariam presentes no extremo sul do continente há mais de 14.000 anos, como indicado pelos artefatos humanos encontrados no sítio arqueológico de Monte Verde, localizado no sul do Chile (Dillehay *et al.*, 2008). Isso sugere que os humanos devem ter se deslocado ao longo de toda a extensão do continente americano – com mais de 14 mil km – em pouquíssimo tempo e que o povoamento de todo o continente deve ter ocorrido de maneira muito veloz (Dillehay, 2009). Neste curto período de tempo, uma população extremamente móvel teria se deslocado ocupando diferentes ambientes, ao longo de uma extensa variação de formações florestais, recursos alimentícios, geografia, diversidade patogênica e, em suma, de pressões seletivas variadas. É, portanto, bastante plausível pensar que essas populações tenham desenvolvido algumas adaptações biológicas a essa grande variedade de ambientes e que pelo menos algumas devam ter deixado sinais genéticos, como, por exemplo, a variante ABCA1*R230C, de acordo com o cenário evolutivo proposto por Hünemeier *et al*. (2012a).

Adicionalmente, algumas dessas populações passaram por gargalos de garrafa posteriores devido ao processo de fissão-fusão ao qual particularmente as populações das terras baixas sul-amerincas estariam submetidas (Neel e Salzano, 1967). Em

consequência desses fenômenos de redução da diversidade genética causado pela estruturação e reduzido tamanho efetivo dessas populações, observa-se por um lado uma alta homogeneidade intra-populacional e por outro uma alta divergência inter-populacional, em especial com relação aos indígenas amazônicos (Wang *et al.*, 2007). Tal fenômeno é menos pronunciado em indígenas andinos, o que se deve ao fluxo gênico recorrente entre populações dos Andes e maior tamanho efetivo dessas populações (Scliar *et al*., 2012).

Como populações de diversidade genética tão reduzida em decorrência dos sucessivos gargalos populacionais e da ação da deriva genética foram capazes de lidar com a adaptação a esta ampla gama de pressões seletivas? Como os processos adaptativos biológicos podem responder tão rapidamente a essa variação extrema? Processos exclusivamente biológicos devem ter ocorrido ao longo da história dos povos nativos americanos, porém os efeitos da cultura e suas vantagens adaptativas são fatos que não podem ser negligenciados.

I.III) GENES E CULTURA(S)

A cultura humana se diferencia da dos outros animais por apresentar uma alta complexidade e capacidade de acumulação sucessiva de novidades (Richerson e Boyd, 2005). Desde os primórdios, essa cultura tem sido associada com um aumento na vantagem adaptativa de nossa espécie, propiciando a saída da África e o povoamento da Eurásia (Mellars, 2006) e, mais recentemente, acionando mecanismos seletivos que alteram o valor adaptativo de indivíduos que apresentam um fenótipo específico (Hünemeier *et al*., 2012b) ou de certos genes a partir da construção de nichos em que estes podem ser vantajosos (Hünemeier *et al*., 2012a). Nesse sentido, a co-evolução

gene-cultura[1] tem sido estudada em populações ameríndias utilizando-se a variação genética do gene que codifica o receptor de dopamina-D4 (*DRD4*) em populações agricultoras e caçadoras-coletoras (Tovo-Rodrigues *et al*., 2010), a diversidade crânio-facial dos Xavantes associada à evolução de marcadores putativamente neutros (Hünemeier *et al*., 2012b) e a relação de um gene associado ao influxo de lipídios para a célula com o surgimento da agricultura nas Américas (Hünemeier *et al*., 2012a).

Uma forma de abordagem alternativa a essa é considerar aspectos culturais que apresentam evolução independente da genética. Exemplos desse tipo de abordagem são os estudos que visam a entender a variação linguística a partir da variação biológica e ainda aqueles que avaliam a relação entre a evolução dessas duas entidades: as línguas (ou culturas) e os genes. Esses estudos, apesar de serem baseados em modelos demográficos simplificados e pouco realistas, têm contribuído enormemente para o entendimento de aspectos da história demográfica e da variação linguística das populações nativas das Américas, como, por exemplo, ao esclarecer a dinâmica de surgimento e dispersão de família linguísticas (Callegari-Jacques *et al*., 2011), ao analisar a estruturação dentro de certos grupos linguísticos (Fagundes *et al*., 2002) e ao validar determinadas classificações linguísticas a partir da variação genética (Hunley e Long, 2005).

Esses estudos comparativos entre genes e línguas são possíveis devido ao fato de que ambos evoluem, isto é, apresentam variação herdável, que está sujeita a mecanismos de manutenção randômicos e direcionais (Richerson e Boyd, 2005; Reali e Griffiths, 2010). Apesar disso, línguas, como outros elementos culturais, também são

[1] "Co-evolução gene-cultura" é um termo utilizado por Peter Richerson e Robert Boyd (2005:191-193) para descrever o sistema em que os componentes biológico e cultural humanos interagem de forma a modificarem o valor adaptativo um do outro respectivamente.

transmitidas horizontalmente e possuem menos restrições evolutivas. Por conta disso, a cultura, em princípio, pode ser substituída mais rapidamente que genes e apresentar taxas de evolução mais velozes (Perreault, 2012).

I.IV) A IMPORTÂNCIA DOS MECANISMOS ADAPTATIVOS NO POVOAMENTO DAS AMÉRICAS

A interação de espécies com o ambiente – quer seja este ambiente manipulado pela cultura ou não – é um fator normalmente associado à ocorrência de adaptações biológicas que muitas vezes geram impacto na diversidade genética da população submetida a novas pressões ambientais (Coop *et al*., 2010). Apesar da evolução molecular ser influenciada principalmente por mecanismos evolutivos neutros (Kimura e Ohta, 1974; Hughes, 2012), a genômica populacional tem apontado uma série de casos em que genes, redes de genes e regiões regulatórias foram aparentemente alvos de seleção positiva. Essas análises revelam por um lado que episódios de varreduras adaptativas clássicas[2] devem ter sido raros durante a evolução humana (Hernandez *et al*., 2011) e por outro que algumas classes de genes como aqueles relacionados à reprodução, percepção sensorial e principalmente à resposta imune foram os alvos mais importantes da seleção positiva (Sabeti *et al.*, 2006, 2007; Williamson *et al.,* 2007; López Herráez *et al*., 2009; Daub *et al*., 2013), além de regiões regulatórias (Grossman *et al.*, 2013). Tais estudos, de escala genômica, representam a mudança de uma perspectiva em que os estudos se baseavam em hipóteses *a priori* a partir da investigação de genes candidatos, para os estudos que são geradores de hipótese, a partir dos dados, trazendo

[2] As varreduras adaptativas clássicas ou *classic selective sweeps* são aqueles casos em que uma mutação nova com alto valor adaptativo é rapidamente fixada numa população (Sabeti *et al.*, 2006), em contraposição com a seleção positiva sobre variação pré-existente ou seleção sobre múltiplos alelos (Hernandez *et al.*, 2011), o que levaria à fixação de alelos de frequência intermediária.

novas perspectivas para o entendimento da evolução humana e dos mecanismos adaptativos.

As metodologias para inferência de seleção positiva são as mais variadas e partem de diferentes pressupostos. Duas das mais utilizadas em nível intraespecífico são baseadas nas ideias de que alelos que apresentam alto valor adaptativo em determinado ambiente devem exibir maior diversidade interpopulacional do que alelos em *loci* neutros (Beaumont, 2005) e os haplótipos em que se situam devem ser encontrados em alta frequência populacional atrelada à ampla extensão de desequilíbrio de ligação (Sabeti *et al.*, 2002). Apesar das críticas constantes (Hughes, 2012), ambos os métodos – principalmente em suas versões mais recentes – são robustos o suficiente para não serem afetados por efeitos demográficos ou por particularidades dos sistemas e das regiões utilizadas (Beaumont, 2005; Sabeti *et al.*, 2006, 2007; Foll e Gaggiotti, 2008). Dessa forma, é possível ter, hoje em dia, uma certeza maior de que parte significativa do genoma humano evoluiu sob o efeito da seleção positiva, colocando em cheque ideias neutralistas clássicas mais radicais ou, pelo menos, levantando questionamentos sobre as situações especiais em que mecanismos neutros devem ter sido mais ou menos importantes que mecanismos adaptativos.

Nesse âmbito, é plausível imaginar que mecanismos evolutivos randômicos devem ter sido de grande importância na origem e manutenção da diversidade genética de populações ameríndias, uma vez que a história demográfica dessas populações incluiu uma série de gargalos populacionais e fenômenos de redução de tamanho efetivo (Neel e Salzano, 1967; Fagundes *et al.,* 2008; Kitchen *et al.*, 2008). Apesar disso, tal ideia é meramente especulativa, já que poucos estudos em escala genômica incluíram populações ameríndias em suas amostras. São alguns exemplos desse tipo os estudos de

López Herráez *et al*. (2009), Coop *et al*. (2010) e Daub *et al*. (2013). Não obstante, em nenhum desses casos essas populações foram o principal foco. Dentro desse cenário, é difícil imaginar quais genes e funções biológicas devem ter exercido alguma importância na pré-história americana, se de fato a ampla variação ambiental representou uma barreira para o povoamento de determinadas partes do continente, como foram os modos de adaptação no Novo Mundo e quais redes genéticas podem ser ativadas em um curto período de tempo para efetivarem a adaptação de seres humanos às pressões seletivas de novos territórios.

Apesar da escassez de varreduras totais do genoma para esse fim, esforços têm sido direcionados para a busca de sinais de seleção natural em genes específicos. Podem-se citar como exemplos a busca de *loci* candidatos para adaptação à alta altitude em populações andinas (Bigham *et al.*, 2009); a inferência de ação de seleção natural em genes relacionados à imunidade (Tarazona-Santos *et al.*, 2011); o exemplo do *ABCA1* na Mesoamérica, que sugere que a interação dos indivíduos com o ambiente modificado pela agricultura teria ocasionado o aumento do valor adaptativo do um alelo que favorece o influxo de colesterol para a célula (Hünemeier *et al*., 2012a); e o caso em que a deriva genética obliterou o sinal da seleção natural existente no gene *KIR* devido às peculiaridades da história demográfica ameríndia (Augusto *et al.*, 2013), o mesmo não tendo ocorrido com o *CCR5*, um gene com indícios persistentes de seleção balanceadora (Ramalho *et al*., 2010). Com o aumento de exemplos como esses e com a realização de estudos de varredura genômica especialmente desenhados para a análise de povos nativos americanos, o cenário complexo de povoamento do Novo Mundo poderá ser esclarecido de uma maneira mais eficaz. Espera-se que parte dos sinais genéticos de adaptação deva ter sido apagada por conta da deriva genética, dos sucessivos gargalos

populacionais e do reduzido tamanho efetivo de grande parte das populações nativas das Américas. No entanto, uma vez que estudos com genes e regiões candidatas já revelaram alguns casos em que a seleção natural moldou a diversidade genética ameríndia, é provável que estudos de varredura total do genoma apontem para novas direções, sugerindo novos genes candidatos, e que possam explicitar de que maneira povos nativos americanos foram capazes de lidar com a diversidade de ambientes do continente em um curto período de tempo.

I.V) OBJETIVOS

O presente trabalho tem como objetivo geral analisar os processos evolutivos culturais e biológicos (adaptativos ou neutros) em populações nativas americanas, procurando entender as particularidades de cada um e avaliando seus efeitos sobre a diversidade genética. Os resultados, apresentados na Parte II a seguir, são divididos em quatro artigos que, no geral, visam:

1) à análise dos padrões de desequilíbrio de ligação e diversidade genética do cromossomo X e sua relação com a história demográfica de populações ameríndias, comparando-as com grupos que também apresentam certo grau de isolamento, mas que possuem contribuição genética distinta dessas;

2) à identificação e interpretação de sinais genéticos de adaptação à Floresta Amazônica numa abordagem que considera as convergências evolutivas como evidência para a adaptação, utilizando-se para esse fim uma comparação com populações africanas que vivem em ambiente semelhante;

3) ao entendimento da relação entre a evolução genômica e linguística em populações nativas sul-americanas, relacionando-a à história demográfica de

populações pertencentes aos cinco principais grupos linguísticos dessa região e utilizando modelos demográficos realistas; e

4) à descrição dos mecanismos evolutivos que levaram ao aumento da frequência de uma mutação não-sinônima no gene *ACTN3* em populações ameríndias, em contraposição com o corrido em outras populações do mundo.

PARTE II

II.I) ARTIGO 1

Amorim CEG, Wang S, Marrero AR, Salzano FM, Ruiz-Linares A, Bortolini MC (2011) *X-chromosomal genetic diversity and linkage disequilibrium patterns in Amerindians and non-Amerindian populations*. *Am J Hum Biol* 23:299-304. doi: 10.1002/ajhb.21110.

AMERICAN JOURNAL OF HUMAN BIOLOGY 23:299–304 (2011)

Original Research Article

X-Chromosomal Genetic Diversity and Linkage Disequilibrium Patterns in Amerindians and Non-Amerindian Populations

CARLOS EDUARDO G. AMORIM,[1] SIJIA WANG,[2] ANDREA R. MARRERO,[1] FRANCISCO M. SALZANO,[1] ANDRÉS RUIZ-LINARES,[2] AND MARIA CÁTIRA BORTOLINI[1]*
[1]*Programa de Pós-Graduação em Genética e Biologia Molecular and Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, 91501-970, Porto Alegre, RS, Brazil*
[2]*The Galton Laboratory, Department of Biology, University College London, London, United Kingdom*

Objectives: We report X-chromosomal linkage disequilibrium (LD) patterns in Amerindian (Kogi, Wayuu, and Zenu) and admixed Latin American (Central Valley of Costa Rica and Southern Brazilian Gaucho) populations.

Methods: Short tandem repeats (STRs) widespread along the X-chromosome were investigated in 132 and 124 chromosomes sampled from the Amerindian tribes and the admixed Latin American populations, respectively. Diversity indexes (gene diversity and average numbers of alleles per locus) were estimated for each population and the level of LD was inferred with an exact test.

Results: The Amerindian populations presented lower genetic diversity and a higher proportion of loci in LD than the admixed ones. Two haplotype blocks were identified in the X-chromosome, both restricted to the Amerindians. The first involved DXS8051 and DXS7108 in Xp22.22 and Xp22.3, while the second found only among the Kogi, included eight loci in a region between Xp11.4 and Xq21.1.

Conclusions: In accordance to previous work done with other populations, human isolates, such as Amerindian tribes, seem to be an optimal choice for the implementation of association studies due to the wide extent of LD which can be found in their gene pool. On the other hand, the low proportion of loci in LD found in both admixed populations studied here could be explained by events related to their history and similarities between the allele frequencies in the parental stocks. Am. J. Hum. Biol. 23:299–304, 2011.

Understanding patterns of linkage disequilibrium (LD), i.e., the nonrandom association of alleles at two or more loci, is the basis for human gene mapping and for the design of association studies (reviewed by Slatkin, 2008). It also provides information on many aspects of population history and evolution, such as the occurrence of natural selection, past demographic events, gene flow, population structure, and breeding system (Ardlie et al., 2002; Pfaff et al., 2001; Reich et al., 2001; Slatkin, 2008). LD serves as the theoretical foundation for association mapping—a marker and a functional locus need to be in LD for the association to be detected. Consequently, populations in high LD are the best choices for designing gene mapping strategies, as the number of markers employed for the identification of an associated allele can be reduced (Ardlie et al., 2002; Slatkin, 2008). In particular, small isolates present the desired profile for gene mapping studies because they commonly have high levels of LD caused by drift effects (Katoh et al., 2002; Varilo and Peltonen, 2004; Service et al., 2006). Admixed populations are also of interest, because their gene pool was formed by relatively recent events involving distinct parental stocks, which may result in long-range LD (Pfaff et al., 2001).

Many Amerindian and Latin American populations present these two characteristics and therefore seem to be optimal choices for gene mapping and for the implementation of association studies. Notwithstanding linkage disequilibrium investigations among them are scarce. We therefore developed an investigation design, which is an extension of previous and recent investigations on X-STR LD in distinct pools of Amerindian and non-Amerindian samples (Leite et al., 2009; Wang et al., 2010). In our study three relatively isolated Native American populations and two non-native communities were tested for a fast-evolving genetic system, with the following questions in mind: (a) since genetic variability clearly influences the power to detect LD, do the populations investigated here show differences in this characteristic that could influence the patterns obtained? (b) In the populations tested what factors could be more adequate to explain the LD levels? and (c) Could haplotype blocks be distinguished that would provide new tools for the investigation of the phylogeography of these populations? The results indicated that the Amerindian tribes presented less genetic diversity and wider LD extent than the remaining populations, where LD was virtually absent. This pattern is most likely associated to the evolutionary history of these populations and also to the genetic systems chosen for the analyses.

SUBJECTS AND METHODS

Population description and data source

New data for 47 X-chromosomal short tandem repeats (STRs) were generated for three South Amerindian populations from Colombia. Seventeen of these markers were

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsors: Institutos do Milênio and Apoio a Núcleos de Excelência Programs; Conselho Nacional de Desenvolvimento Científico e Tecnológico; Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul.

Carlos Eduardo G. Amorim and Sijia Wang contributed equally to this work.

*Correspondence to: Maria Cátira Bortolini, Programa de Pós-Graduação em Genética e Biologia Molecular and Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, 91501-970, Porto Alegre, RS, Brazil.
E-mail: maria.bortolini@ufrgs.br

Received 5 April 2010; Revision received 8 September 2010; Accepted 14 September 2010

DOI 10.1002/ajhb.21110

Published online 3 February 2011 in Wiley Online Library (wileyonlinelibrary.com).

TABLE 1. Geographical location of the five populations analyzed

Population	Ethnic composition	Country	Geographical coordinates	
Central Valley of Costa Rica	Admixed	Costa Rica	9°56′N	84°05′W
Southern Brazilian Gaucho	Admixed	Brazil	31°00′S	54°00′W
Kogi	Amerindian	Colombia	11°00′N	74°00′W
Wayuu	Amerindian	Colombia	9°00′N	75°00′W
Zenu	Amerindian	Colombia	11°00′N	73°00′W

also investigated in a sample of inhabitants of the Pampa region, which corresponds to parts of Argentina, Uruguay, and southern Brazil. These people are known as "Gaucho" (Marrero et al., 2007).

Our Amerindian sample comprises 132 chromosomes sampled from Kogi (Chibchan-Paezan linguistic stock), Wayuu (Equatorial-Tucanoan linguistic stock), and Zenu (currently speaking Spanish; Mesa et al.,2000). For each population, 13 women and 18 men have been genotyped. The Gaucho genotyped here ($N = 70$) are a subset of the male sample collected in the Brazilian Pampa region in the cities of Bagé and Alegrete in the Brazilian state of Rio Grande do Sul previously described by Marrero et al. (2007).

A fifth population, located in the Central Valley of Costa Rica (CVCR), was included in the analysis. This admixed population was previously described by Service et al. (2001; data kindly provided) and, in our work we analyzed the 54 male individuals. Briefly, CVCR was founded by Spaniards and Amerindians in the 16th to 18th centuries, and the current population number is two to three million; they are relatively isolated from communities of the Pacific and Atlantic coastal regions.

Additional information about the populations studied is as follows: (a) the Kogi live in Sierra Nevada de Santa Marta, and have been relatively isolated from non-Amerindians (Zarante et al., 2000). A summary of their environment and culture, basically derived from the work of Gerardo Reichel-Dolmatoff performed in the 1940s, can be found in Wilson (1999); (b) Wayuu and Zenu represent the other Amerindian populations which, differently from Kogi, show signs of admixture with non-native groups during America's colonization (Mesa et al., 2000; Wang et al., 2007); and (c) the Gaucho population was basically formed at 18th century through intermarriages between the European colonizers (mainly Portuguese and Spaniard males) and Amerindian females. The African component was introduced later with the first slaves in the region (Marrero et al., 2007; Wang et al., 2008). Geographical location of each population can be found in Table 1.

Ethical approval for the present study was provided by the Brazilian National Ethics Commission (CONEP; resolution number: 1333/2002) and by ethics committees in the countries where the non-Brazilian samples were collected.

Genotyping procedure

Forty-seven STRs were genotyped for the Amerindian populations. This set of loci comprises DXS1039, DXS1216, plus the markers included in ABI Prism linkage Panels 28, 83 (except for DXS8088), 84, 85, and 86. The same information was compiled for the 54 CVCR male individuals studied earlier (Service et al., 2001). A subset of these loci, encompassing 17 markers of Panel 28 (except DXS8043), was typed for the Gaucho. Genotyping was performed according to the user's manual provided by the manufacturer (ABI Prism) using an ABI 3730xl sequencer and with GENEMAPPER v3.5. All the newly obtained genotypes (classified according to allele length) are included in the Supporting Information Dataset S1.

Statistical analysis

Phase v.2.1 (Stephens and Scheet, 2005; Stephens et al., 2001) was employed to resolve the haplotype phase of the 39 female Amerindians using default settings. The stepwise mutation model was chosen for the analysis. Males from these populations were randomly paired to form pseudo-diploid individuals and were used as known-phase individuals, as suggested by Phase's authors (Mathew Stephens, personal communication). To control for inference errors, three runs were performed for each sample and the three outputs were then compared. No differences were observed.

Allele frequencies and mean number of alleles per locus were estimated by direct count. The Arlequin package v.3.11 (Excoffier et al., 2005) was employed to calculate the average gene diversity ($\hat{H}$; i.e., the probability that two randomly chosen haplotypes are different in a sample) and to perform the LD analysis for all populations. For the LD estimates, Arlequin employs an exact test to check for non-random association of alleles at different loci. The output *P*-values for the association tests were then corrected following the Benjamini and Hochberg false discovery rate procedure (Benjamini and Hochberg, 1995). $\hat{H}$ and mean number of alleles per locus were compared across populations with Dunn's test ($\alpha = 0.05$) using the BioEstat 5.0 software (available at www.mamiraua.com.br) after significant differences in these statistics had been found by a Kruskal-Wallis test ($\alpha = 0.05$).

We then looked for haplotype blocks by examining sequences of two or more markers in a row with significant linkage disequilibrium. Further recombination analyses were conducted for each population with Phase v.2.1 (Li and Stephens, 2003; Crawford et al., 2004) with the presumed haplotype blocks to estimate their background recombination rate (ρ) in "per base pair" units. This software gives a ρ distribution of values generated by "*n*" successive runs, which we set to 100. Haplotype networks for both blocks were generated with the Network 4.5.1.0 package (Bandelt et al., 1999) using the Median Joining method, while the mean number of pairwise differences between haplotypes were estimated with the Arlequin package v.3.11 (Excoffier et al., 2005).

RESULTS

For the 47 X-chromosomal STRs, average number of alleles per locus and $\hat{H}$ were both significantly lower for the three Amerindian tribes as compared to CVCR (Table 2). When the Gaucho sample was considered, decreasing the number of markers to 17, a similar pattern was observed: the mean number of alleles per locus was lower for the Amerindians as compared to CVCR and Gaucho, and $\hat{H}$ was significantly different between Kogi and Gaucho (Table 3).

In the 17-loci dataset the Gaucho presented the lowest degree of LD, followed by CVCR and Zenu. The latter showed the lowest values in the 47-loci results. Kogi and Wayuu presented the highest proportion of loci in LD, but

TABLE 2. Genetic information for Amerindians and the Central Valley of Costa Rica (CVCR) based on the allelic distribution of 47 STRs[a]

Characteristic	Kogi	Wayuu	Zenu	CVCR
Average ± SD gene diversity ($\hat{H}$)	$0.51^{a} \pm 0.25$	$0.57^{a} \pm 0.28$	$0.59^{a} \pm 0.29$	$0.67^{b} \pm 0.33$
Average no. of alleles per locus	3.70^{a}	4.64^{ab}	5.15^{b}	6.87^{c}
Proportion of loci in LD before correction (%)	31.69	22.11	15.82	6.75
Proportion of loci in LD after correction (%)	15.65	3.89	2.59	1.76

[a]Values followed by the same letter present no statistically significant differences according to Dunn's test ($\alpha = 0.05$).

TABLE 3. Genetic information for Amerindian (Kogi, Wayuu, and Zenu) and non-Amerindian (CVCR and Gaucho) populations based on the allelic distribution of 17 STRs[a]

	Kogi	Wayuu	Zenu	CVCR	Gaucho
Average ± SD gene diversity ($\hat{H}$)	$0.52^{a} \pm 0.27$	$0.61^{ab} \pm 0.32$	$0.59^{abc} \pm 0.31$	$0.72^{bc} \pm 0.36$	$0.72^{c} \pm 0.37$
Average no. of alleles per locus	3.23^{a}	4.47^{a}	4.82^{a}	7.06^{b}	8.18^{b}
Proportion of loci in LD before correction (%)	28.33	35.29	23.53	6.62	4.41
Proportion of loci in LD after correction (%)	14.17	22.79	8.09	0	0

[a]Values followed by the same letter present no statistically significant differences according to Dunn's test ($\alpha = 0.05$).

TABLE 4. DXS8051 and DXS7108 haplotype distribution in three Colombian Amerindian populations[a]

			Frequencies			
Haplotype	DXS8051	DXS7108	Kogi	Wayuu	Zenu	Total
h1	112	252	–	–	1	1
h2	114	260	1	–	1	2
h3	116	252	–	3	1	4
h4	116	262	–	2	–	2
h5	118	252	5	1	3	9
h6	118	254	3	–	–	3
h7	118	256	–	–	3	3
h8	118	260	5	2	4	11
h9	118	262	4	1	9	14
h10	120	252	–	–	2	2
h11	120	260	–	–	1	1
h12	122	252	1	–	–	1
h13	122	260	8	6	–	14
h14	122	262	–	7	–	7
h15	124	252	5	5	2	12
h16	124	260	–	6	–	6
h17	124	262	2	–	–	2
h18	124	262	–	2	–	2
h19	126	252	4	–	–	4
h20	126	256	1	–	–	1
h21	126	260	–	–	7	7
h22	126	262	1	–	–	1
h23	128	252	–	4	–	4
h24	128	258	–	–	1	1
h25	128	260	2	–	2	4
h26	128	262	1	–	–	1
h27	128	264	–	–	2	2
h28	130	260	1	–	–	1
h29	132	252	–	–	2	2
h30	132	262	–	5	–	5
h31	134	252	–	–	2	2
h32	134	262	–	–	1	1
Alelles	12	7				132

[a]Alleles are classified according to allele length.

showed distinct ranks when the two different sets of markers were employed (Tables 2 and 3).

Two haplotype blocks of interest were identified, both in Amerindians. The first is common to all three Amerindian populations. It is defined by two loci (DXS8051 and DXS7108) and is located from Xp22.22 to Xp22.3, involving a region of 1,080 kb, with a genetic distance estimated as 2.2 cM (http://www.appliedbiosystems.com). Average $\hat{H}$ for DXS8051 and DXS7108 in Amerindians was estimated as 0.81 and 0.71 respectively, while the recombination rates between these loci were 0.0005, 0.0013, and 0.0005 for the Kogi, Wayuu, and Zenu populations, respectively. The haplotype network presented several reticulations and therefore is not shown. Possible causes for this level of reticulation include back mutation, homoplasy, and statistical limitations of the methodological procedures used to generate the networks. As given in Table 4, the most common haplotypes in the whole Amerindian sample were h9 (118-262) and h13 (122-260); one of them was also the most common haplotype in the Zenu and Kogi, but the haplotype with the highest frequency among the Wayuu was h14 (122-262). Kogi, Wayuu, and Zenu presented eight, six, and ten private haplotypes respectively.

The Kogi presented many haplotype blocks defined by few markers. However, one of these blocks deserves mention due to its size. It includes eight loci (DXS993, DXS8080, DXS8083, DXS1055, DXS1039, DXS991, DXS1216, and DXS986) with a background recombination rate (ρ) estimated at 6.3×10^{-5}. The first and last markers in this block are 11 kb apart from each other, and delimit a region between Xp11.4 and Xq21.1, therefore including the centromere. All these loci except DXS1039, presented higher $\hat{H}$ than the overall mean. DXS1039 presents the lowest proportion of significant LD *P*-values and very low $\hat{H}$ value. The network with this eight-loci haplotype block is shown in Figure 1, while the haplotype frequencies are given in Table 5. Haplotype h8 (275-82-183-152-189-331-250-175) was the most frequent (22%), followed by h7 (275-82-183-150-189-333-250-175; 14%). Mean number of pairwise differences for this haplotype block was estimated as 4.73 ± 2.36 and $\hat{H}$ as 0.59 ± 0.22.

DISCUSSION

Many aspects of population evolutionary and demographic trajectories can affect LD patterns. It is known, for instance, that natural selection, drift, and gene flow can raise the number of associated alleles in a sample, but the exact way these mechanisms affect linkage is not fully understood. In our study, we examined five distinct populations with different evolutionary or demographic histories to better understand the forces that can lead to distinct patterns of LD in such groups. The Amerindian populations presented lower intra-population genetic diversity

and higher extent of LD than the other two (CVCR and Gaucho) groups. While the pattern observed in the three Amerindian populations can be explained according to the evolutionary and demographic dynamics of such groups in the Americas, the low level of LD observed in CVCR and Gaucho can be associated with their admixture dynamics and the set of markers employed.

The first populations that arrived in the Americas were initially subjected to a moderate population bottleneck during the entry into the continent (Fagundes et al., 2008). Additionally, they experienced successive posterior population bottlenecks due to the dramatic effects of the fission–fusion events of village propagation and tribalization (Bortolini et al., 2003; Neel and Salzano, 1967). As a consequence of these demographic and evolutionary phenomena, genetic drift became important. Presently, Amerindian populations show distinct genetic characteristics when compared to those of other continents, such as the highest interpopulation divergence and the lowest intrapopulation diversity when a large number of autosomal fast-evolving markers are considered (Wang et al., 2007). Additionally, private Native American alleles have emerged (Schroeder et al., 2007, 2009; Acuña-Alonzo et al., 2010).

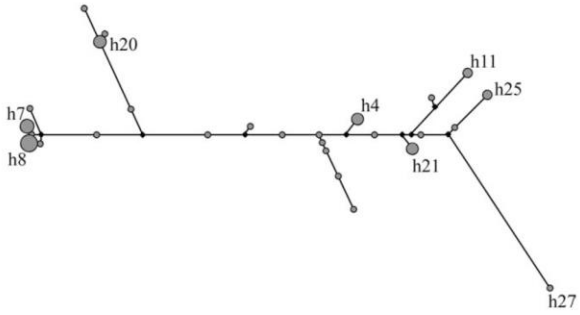

Fig. 1. Kogi's haplotype network considering eight X-chromosome loci. Grey circles represent the haplotypes and their sizes are proportional to haplotype frequencies. Black diamonds represent median vectors; h28 was not included in this analysis because its inclusion would lead to reticulation.

It is known that population bottlenecks influence LD, since genetic drift may easily cause haplotypes loss, generally resulting in increased LD (Slatkin, 2008). Moreover, small effective sizes, which are a characteristic of many Native American populations, may decrease the occurrence of recombination and consequently may also raise LD levels. Thus, the lower genetic diversity and the higher level of LD observed in the Native populations, as compared to the CVCR and Gaucho samples, might be associated with these phenomena. Similar results were obtained in previous independent studies involving South Amerindian (Leite et al., 2009; Wang et al., 2010) and other small and isolated populations such as the Khoton from Mongolia (Katoh et al., 2002), the Saami (Laan and Pääbo, 1997), and Kuusamo (Varilo et al., 2000) from Finland.

The lower genetic diversity of the Kogi in comparison to Wayuu and Zenu might be associated with the fact that they present a higher degree of isolation, with less influence of admixture with nonautochthonous groups (Mesa et al., 2000; Zarante et al., 2000; Wang et al., 2007). This fact, in addition to their reduced population size, may also be responsible for the Kogi's increased LD and for the occurrence of the private long-range haplotype block spanning Xp11.4 and Xq21.1.

TABLE 5. Haplotype distribution considering eight loci in the Kogi, a Colombian Amerindian population[a]

Haplotype	DXS993	DXS8080	DXS8083	DXS1055	DXS1039	DXS991	DXS1216	DXS986	Frequency
h1	273	96	181	150	189	325	254	175	1
h2	273	96	181	150	189	327	254	167	1
h3	273	96	181	150	189	327	254	175	1
h4	273	96	185	154	189	327	254	177	3
h5	273	98	181	150	189	325	254	175	1
h6	275	82	181	156	189	335	250	175	1
h7	275	82	183	150	189	333	250	175	4
h8	275	82	183	152	189	331	250	175	6
h9	275	82	183	152	189	333	250	175	1
h10	275	82	183	154	189	333	254	167	1
h11	275	96	181	150	189	333	260	175	2
h12	275	96	181	154	189	331	254	177	1
h13	275	96	181	154	189	331	260	167	1
h14	277	82	183	154	189	333	250	175	1
h15	277	96	181	150	189	327	260	167	1
h16	277	100	181	150	189	325	254	167	1
h17	277	100	181	150	189	325	254	175	1
h18	281	82	181	154	189	331	260	167	1
h19	281	82	185	152	189	327	260	175	1
h20	281	82	185	154	189	327	260	175	3
h21	281	96	181	154	189	327	254	177	3
h22	281	96	181	154	189	333	254	167	1
h23	281	96	181	154	189	333	254	175	1
h24	281	96	183	152	189	331	260	175	1
h25	281	96	185	154	189	327	254	167	2
h26	289	82	185	154	189	327	260	175	1
h27	295	90	175	148	187	331	250	169	1
h28	275	96	181	154	189	331	254	175	
Alleles	6	5	4	5	2	5	3	4	43

[a]Alleles are classified according to allele length.

On the other hand, the low proportion of loci in LD in the CVCR and Gaucho samples deserves additional attention. These populations present a complex pattern of admixture for the X-chromosome, involving mainly Amerindians and Europeans (the African contribution is also detected, but at lower proportions). The relative ancestral contribution to their X-chromosomes is not very different [in percentages, CVCR, European (E): 40; Amerindian (Am): 42; African (Af): 18; Gaucho, E: 47; Am: 31; Af: 22; (Wang et al., 2008)].

Different admixture models can create distinct LD patterns. Long (1991) proposed two extreme models of admixture: the hybrid-isolation (HI) and the continuous-gene-flow (CGF) models. In the HI model, admixture occurs in a single generation producing long-range LD, which progressively decays in each successive generation as a result of independent assortment and recombination between loci. In the CGF model, admixture occurs at a steady rate in every generation, and the amount of LD increases during the first few generations as continual admixture generates more disequilibrium than is broken down by independent assortment and recombination. After a few generations, the amount of LD begins to decay, although at a much slower rate than that observed in the HI model (Pfaff et al., 2001). The last authors, using a simulation approach, observed that under the HI model and considering unlinked loci (genetic distances between loci greater than 5 cM), only five generations are necessary after admixture for LD to decay to values close to zero (Pfaff et al., 2001). Applying this theory to our work, based on age of admixture estimates generated elsewhere (6.57 generations for the Gaucho and 14.42 for CVCR; Wang et al., 2008), the observed low proportion of loci in LD (Tables 2 and 3) could be explained assuming the hybrid-isolation model for unlinked markers for both admixed populations studied here. Furthermore, similarities between allele frequencies in the parental stocks could also be related to the pattern observed to Gaucho and CVCR populations.

The comparison between X-chromosomal diversity and LD between Amerindian and admixed groups was already examined in a previous investigation (Leite et al., 2009). This work shows some similarities to ours, with lower locus diversity, lower number of alleles per site, and higher LD among Amerindians as compared to the admixed populations. The diversity indices, however, are slightly lower in our work, indicating that the set of markers used by Leite et al. (2009) is more variable than ours, which may also be linked to the wider extent of LD detected in our study. Mention should also be made of the differences between the correction procedures and methods for LD detection used in both studies, which may have influenced the LD comparison.

Finally, the two haplotype blocks observed in the Amerindian X-chromosomes could be of potential interest, since LD analysis can also bring new insights into human evolutionary studies. It is reasonable to suggest that the most frequent haplotype observed in a specific population is the founder haplotype. However, the nature of the Amerindian evolutionary history, as mentioned before, strongly influenced by genetic drift due to isolation and fragmentation, prevents the identification of the Native American founder haplotype considering the *DXS8051* and *DXS7108* loci (Table 4). On the other hand, the network of Figure 1, based on eight loci (DXS993, DXS8080, DXS8083, DXS1055, DXS1039, DXS991, DXS1216, and DXS986), shows no evident signal of diversification from a founder haplotype, and the variable haplotype frequencies can also be associated with a scenario where genetic drift has a strong influence.

Despite these uncertainties, the two blocks reflect low-recombination regions, and it is possible to suggest that both, or at least parts of them, could be found in other Native American populations. In this context it is worth mentioning that a haplotype block between Xq13.3 and Xq21.3 was found in an investigation of 51 different populations, including South Amerindians (Santos-Lopes et al., 2007). In our analysis, only one marker (DXS1196) was included in this region, and therefore no appropriate comparison can be made between their findings and our data.

CONCLUSIONS

Now it is possible to answer the questions raised in the introduction. The Amerindian populations clearly show less intrapopulation variability than the two others and only among them LD blocks were detected. In this case, random drift should be considered as the most important factor for this detection. On the other hand, events as described in the hybrid-isolation model for unlinked markers and similarities between the allele frequencies in the parental stocks could be responsible for the absence of LD in both admixed populations. The two observed haplotype blocks can provide valuable new tools for Amerindian X-chromosome phylogeopraphic investigations, and this possibility will be explored by us in future studies.

ACKNOWLEDGMENTS

The authors thank all the blood donors for their participation, as well as Rogério Schmidt Campos for the designing support, and Sidia Maria Callegari-Jacques, Carla Martins Lopes, and Nelson Jurandi Rosa Fagundes for statistical and methodological help.

LITERATURE CITED

Acuña-Alonzo V, Flores-Dorantes T, Kruit JK, Villarreal-Molina T, Arellano-Campos O, Hünemeier T, Moreno-Estrada A, Ortiz-López MG, Villamil-Ramírez H, León-Mimila P, Villalobos-Comparan M, Jacobo-Albavera L, Ramírez-Jiménez S, Sikora M, Zhang LH, Pape TD, Granados-Silvestre Mde A, Montufar-Robles I, Tito-Alvarez AM, Zurita-Salinas C, Bustos-Arriaga J, Cedillo-Barrón L, Gómez-Trejo C, Barquera-Lozano R, Vieira-Filho JP, Granados J, Romero-Hidalgo S, Huertas-Vázquez A, González-Martín A, Gorostiza A, Bonatto SL, Rodríguez-Cruz M, Wang L, Tusié-Luna T, Aguilar-Salinas CA, Lisker R, Moises RS, Menjivar M, Salzano FM, Knowler WC, Bortolini MC, Hayden MR, Baier LJ, Canizales-Quinteros S. 2010. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. Hum Mol Genet 19:2877–2885.

Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299–309.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300.

Bortolini MC, Salzano FM, Thomas MG, Stuart S, Nasanen SP, Bau CH, Hutz MH, Layrisse Z, Petzl-Erler ML, Tsuneto LT, Hill K, Hurtado AM, Castro-de-Guerra D, Torres MM, Groot H, Michalski R, Nymadawa P, Bedoya G, Bradman N, Labuda D, Ruiz-Linares A. 2003. Y-chromosome evidence for differing ancient demographic histories in the Americas. Am J Hum Genet 73:524–539.

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700–706.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50.

Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA Jr, Zago MA, Ribeiro-dos-Santos AK, Santos SE, Petzl-Erler ML, Bonatto SL. 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet 82:583–592.

Katoh T, Mano S, Ikuta T, Munkhbat B, Tounai K, Ando H, Munkhtuvshin N, Imanishi T, Inoko H, Tamiya G. 2002. Genetic isolates in East Asia: a study of linkage disequilibrium in the X chromosome. Am J Hum Genet 71:395–400.

Laan M, Pääbo S. 1997. Demographic history and linkage disequilibrium in human populations. Nat Genet 17:435–438.

Leite FP, Santos SE, Rodrigues EM, Callegari-Jacques SM, Demarchi DA, Tsuneto LT, Petzl-Erler ML, Salzano FM, Hutz MH. 2009. Linkage disequilibrium patterns and genetic structure of Amerindian and non-Amerindian Brazilian populations revealed by long-range X-STR markers. Am J Phys Anthropol 139:404–412.

Li N, Stephens M. 2003. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. Genetics 165:2213–2233.

Long JC. 1991. The genetic structure of admixed populations. Genetics 127:417–428.

Marrero AR, Bravi C, Stuart S, Long JC, Pereira-das-Neves-Leite F, Kommers T, Carvalho CM, Pena SD, Ruiz-Linares A, Salzano FM, Bortolini MC. 2007. Pre- and post-Columbian gene and cultural continuity: the case of the Gaucho from southern Brazil. Hum Hered 64:160–171.

Mesa NR, Mondragón MC, Soto ID, Parra MV, Duque C, Ortíz-Barrientos D, García LF, Velez ID, Bravo ML, Múnera JG, Bedoya G, Bortolini MC, Ruiz-Linares A. 2000. Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre- and post-Columbian patterns of gene flow in South America. Am J Hum Genet 67:1277–1286.

Neel JV, Salzano FM. 1967. Further studies on the Xavante Indians. X. Some hypotheses-generalizations resulting from these studies. Am J Hum Genet 19:554–574.

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD. 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68:198–207.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Santos-Lopes SS, Pereira RW, Wilson IJ, Pena SD. 2007. A worldwide phylogeography for the human X chromosome. PLoS One 2:e557.

Schroeder KB, Jakobsson M, Crawford MH, Schurr TG, Boca SM, Conrad DF, Tito RY, Osipova LP, Tarskaia LA, Zhadanov SI, Wall JD, Pritchard JK, Malhi RS, Smith DG, Rosenberg NA. 2009. Haplotypic background of a private allele at high frequency in the Americas. Mol Biol Evol 26:995–1016.

Schroeder KB, Schurr TG, Long JC, Rosenberg NA, Crawford MH, Tarskaia LA, Osipova LP, Zhadanov SI, Smith DG. 2007. A private allele ubiquitous in the Americas. Biol Lett 3:218–223.

Service SK, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, Ospina J, Ruiz-Linares A, Macedo A, Palha JA, Heutink P, Aulchenko Y, Oostra B, van Duijn C, Jarvelin MR, Varilo T, Peddle L, Rahman P, Piras G, Monne M, Murray S, Galver L, Peltonen L, Sabatti C, Collins A, Freimer N. 2006. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38:556–560.

Service SK, Ophoff RA, Freimer NB. 2001. The genome-wide distribution of background linkage disequilibrium in a population isolate. Hum Mol Genet 10:545–551.

Slatkin M. 2008. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462.

Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L. 2000. Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. Eur J Hum Genet 8:604–612.

Varilo T, Peltonen L. 2004. Isolates and their potential use in complex gene mapping efforts. Curr Opin Genet Dev 14:316–323.

Wang S, Bedoya G, Labuda D, Ruiz-Linares A. 2010. Brief communication: patterns of linkage disequilibrium and haplotype diversity at Xq13 in six Native American populations. Am J Phys Anthropol 142:476–480.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado AM, Labuda D, Klitz W, Barrantes R, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A. 2007. Genetic variation and population structure in Native Americans. PLoS Genet 3:e185.

Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM, Camrena B, Nicolini H, Klitz W, Barrantes R, Molina JA, Freimer NB, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Dipierri JE, Alfaro EL, Bailliet G, Bianchi NO, Llop E, Rothhammer F, Excoffier L, Ruiz-Linares A. 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet 4: e1000037.

Wilson DJ. 1999. Indigenous South Americans of the past and the present. An ecological perspective. Boulder: Westview Press.

Zarante I, Ossa D, Mendosa R, Valvuena G. 2000. Descripción etnográfica, demográfica y características en salud de las comunidades indígenas visitadas por La gran expedición humana. In: Vásquez AO, organizer. Geografía humana de Colombia: variación biológica y cultural en Colombia—Tomo 1. Santafé de Bogotá, DC: Editora Guadalupe. p 157–227.

II.II) ARTIGO 2

Amorim CEG, Daub J, Foll M, Bonatto SL, Salzano FM, Excoffier L (2013) *Detecting Genome-wide Signals of Human Adaptation to Tropical Forests in a Convergent Evolution Framework* (manuscrito em preparação a ser submetido a *PLoS One*).

Detecting Genome-wide Signals of Human Adaptation to Tropical Forests in a Convergent Evolution Framework

Carlos Eduardo G. Amorim[1,2,3], Josephine Daub[1,2], Matthieu Foll[2,4], Sandro L. Bonatto[5], Francisco M. Salzano[3]*, Laurent Excoffier[1,2]

1 Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, 3012 Berne, Switzerland, **2** Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, **3** Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, 91501-970 Rio Grande do Sul, Brazil, **4** École Polytechnique Fédérale de Lausanne, Switzerland, **5** Faculdade de Biociências, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil.

Running headline: Human genomic adaptations to tropical forests

Keywords: Cholesterol metabolism; HGDP; Immune system; Positive selection; Pygmy phenotype

Funding: CNPq (Brazil); CAPES (Brazil); FAPERGS, PRONEX (Brazil); SNSF (Swizerland)

Competing Interests: The authors have no conflicts of interest to declare.

***E-mail:** francisco.salzano@ufrgs.br

Abstract Tropical forests are believed to be very harsh environments for human life. It is unclear if human beings would have ever subsisted in those environments without external resources. However, it is possible that humans have developed recent biological adaptations in response to specific selective pressures to this challenge. To understand such biological adaptations we analyzed genome-wide SNP data under a Bayesian statistics framework, looking for outlier markers with overly large

differentiation between populations living in a tropical forest, as compared to genetically related populations living outside forest. The most significant positive selection signals were found in genes related to lipid metabolism, the immune system, body development, and axon guidance. The results are discussed in the light of putative tropical forest selective pressures, namely food scarcity, high prevalence of pathogens, difficulty to move, and inefficient thermoregulation. Agreement between our results and previous studies on pygmy phenotype, a putative prototype of forest adaptation, were found, suggesting that a few genetic regions previously described as associated with short stature may be evolving under positive selection in Africa and the Americas.

INTRODUCTION

Tropical forests are characterized by a high diversity of plants, with tall trees, dense canopies and low light penetration (Ratnam et al., 2011). Their climate is generally warm with minimum temperatures well above the freezing point, and mean annual rainfall above 1,000 mm (Bailey et al., 1989). Despite being one of the most productive environments of the world, tropical forests provide only few resources for humans (Hart and Hart, 1986). Indeed, in these environments plants invest most of their energy in structure maintenance and not into the reproductive parts that are the most edible parts for humans and their preys (Bailey et al., 1989). In addition, the instability of food resources in response to the high seasonality of rain falls raises the costs of foraging, further reducing its capacity to support human life (Bailey et al., 1989; Hart and Hart, 1986).

Besides food limitation, other characteristics of tropical forests may also contribute to the hostility of these environments. For instance, tropical areas harbor on average 70% higher pathogen diversity as compared to more temperate areas (Guernier

et al., 2004). As a consequence, infant and child mortality rates among tropical forest dwellers should be high (Ohenjo et al., 2006). Moreover, the small differences between air and skin relative humidities and high temperature, coupled with little air movement, make sweat production and evaporation difficult in tropical forests, potentially compromising thermoregulation (Perry and Dominy, 2009).

Due to the hostility of this environment, it is unclear if humans would have ever subsisted in tropical forests without depending on external resources, such as agriculture or possible exchanges with neighboring populations. Evidence of societies living in such harsh conditions is scarce for contemporary and extinct modern humans (Bailey et al., 1989), as well as for early *Homo* (Mercader, 2002). Nonetheless, it is possible that humans have developed recent biological adaptations to tropical forests. A few examples of such adaptations have indeed been documented, the most well-known being the pygmy phenotype. Short-statured individuals may have, for instance, advantages to cope with food limitation, thermoregulation, and mobility hardship in a dense forest (Perry and Dominy, 2009). However, it has also been suggested that this phenotype could be a by-product of selection for early onset of reproduction (Migliano et al., 2007), which could enable populations to overcome problems related to their life history and increased mortality (Walker and Hamilton, 2008).

To investigate whether tropical forest dwellers have developed specific biological adaptations to this harsh environment, we searched for genome-wide signals of positive selection in populations from the Americas and Africa. Populations living in this environment and others, genetically related, living outside forests were investigated with the specific aim of identifying signals in these two continents.

SUBJECTS AND METHODS

Populations and samples

Genome-wide single nucleotide polymorphism (SNP) data were downloaded for seven populations included in the Human Genetic Diversity Panel database (HGDP; Cann et al., 2002; Rosenberg et al., 2002). Two African (Biaka and Mbuti pygmies) and two American (Surui and Karitiana) tropical forest populations were selected, as well as three other populations from these continents to serve as non-tropical forest comparisons (Mandenka and Yoruba in Africa; Pima in America). Considering the genetic similarity between Mandenka and Yoruba (Li et al., 2008), these populations were grouped into a single set, hereafter called "West Africa", to increase sample size and the statistical power of the analyses. More information on the chosen populations can be found in Hart and Hart (1986); Cann et al. (2002); Rosenberg et al. (2002); Lee and Daly (2004); and ISA (2013). We excluded atypical and duplicated samples, keeping only those present in the H1048 subset of Rosenberg (2006).

The six populations were combined into four distinct population sets (PS), each including one tropical forest and one control population per continent as follows:

PS1: West Africa, Biaka; Pima, Surui.

PS2: West Africa, Mbuti; Pima, Surui.

PS3: West Africa, Biaka; Pima, Karitiana.

PS4: West Africa, Mbuti; Pima, Karitiana.

Each PS was analyzed separately to find loci and genomic regions that would putatively be under natural selection. The rationale behind the use of different population sets is to look for signals of adaptation across data sets, and thus to eliminate signals potentially due to particular tropical forest populations, which have been shown

to present high rates of genetic drift due to their small effective population sizes (Wang et al., 2007; Tishkoff et al. 2009).

Genetic data

Data on 660,918 SNPs were downloaded from the Stanford University HGDP-CEPH SNP genotyping data supplement 1 (Li et al., 2008; <ftp://ftp.cephb.fr/hgdp_supp1/>). We initially discarded 250 markers that were monomorphic in all populations, those that presented only missing data, and those located on either the Y-chromosome, the pseudoautosomal region of the sex chromosomes, or the mtDNA, leaving us with 660,668 SNPs. After selecting the different PSs as described above, those markers with minor allele frequency less than 5%, when all populations were pooled, were discarded, yielding 582,074, 581,855, 584,205, and 577,345 SNPs for population sets PS1-PS4, respectively.

Detecting outlier SNPs

A modified version of BayeScan (Foll and Gaggiotti, 2008) was used to identify candidate targets for natural selection. The methodology is based on the multinomial-Dirichlet likelihood-based approach (Balding, 2003) implemented via a Markov chain Monte Carlo (MCMC) algorithm (Beaumont and Balding, 2004) and uses a hierarchical island model (Excoffier et al., 2009) – in which the subpopulations' allele frequencies are correlated through a common migrant gene pool from which they differ by varying degrees – to calculate a population specific F_{ST} coefficient. Logistically transformed F_{ST} coefficients are then decomposed into a population-specific component (β), shared by all loci, and a locus-specific component (α), shared by all the populations (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008). Selection is inferred when α is significantly different from zero. For each locus, two alternative evolutionary models including α

(selection) or neutrality can thus be explored. The posterior probability of each model (selection vs. neutrality) is estimated with a reversible-jump MCMC algorithm (Foll and Gaggiotti, 2008) and indicates how likely the model with selection is in comparison to the neutral one. Significantly positive values are indicators of an overly large level of differentiation of a given SNP and thus positive selection; whereas significantly negative values of α are indicative of balancing selection. Further information on this methodology can be found in the Bayescan manual or other methodological papers on the F-model (Balding 2003; Beaumont and Balding, 2004; Foll and Gaggiotti, 2008).

The newest version of BayeScan includes a hierarchical island model accounting for the relative closer similarity of certain populations in comparison to others as should be the case of populations that are distributed in different continents. In this regard, when considering two pairs of populations in two different continents one ends up with four alternative selection models for each locus: (1) neutral variability; (2) selection in one continent; (3) selection in the other continent; and (4) selection in both continents. We shall assume hereafter that convergent selection occurred when selection presents a higher probability than neutrality and model 4 has a higher posterior probability than models 2 and 3.

The modified version of BayeScan estimates for each marker the posterior probability of each one of the four tested models. These posterior probabilities are then transformed into *q*-values for each marker in order to control for the False Discovery Rate (FDR; Benjamini and Hochberg, 1995). FDR is defined as the expected proportion of false positives among outlier markers. In this paper, we considered all SNPs with *q*-values lower than 0.1 as significant. This procedure yielded a list of outlier SNPs for each PS, which were then considered as candidate loci for natural selection targets.

Gene and regulatory elements annotation

All SNPs were assigned to genes using PLINK v1.07 (Purcell et al., 2007; available at http://pngu.mgh.harvard.edu/purcell/plink). SNPs not present in coding regions were assigned to a given gene if located less than 50 kb away from it. When more than one gene was within this range, the closest gene was chosen for the subsequent analyses. The amount of outlier SNPs falling in genic regions (or <50kb way from them) was compared to the amount of SNPs in non-genic regions. This proportion was then compared to the distribution of the 660,668 HGDP SNPs in genic or non-genic regions with a chi-square test using R (R Development Core Team, 2011) to check if these outlier SNPs were enriched for genic SNPs in comparison to all available markers.

The coordinates of the 19,668 protein-coding genes located on the human autosomal and X chromosomes were obtained from NCBI Entrez Gene website (Maglott et al., 2011; <http://www.ncbi.nlm.nih.gov/gene>, accessed on January 4, 2012) for the hg19 assembly (NCBI Build 37.3). Twenty-six genes presented multiple locations; in these cases we took the outermost start and end positions. The original positions of the SNPs on the hg18 reference genome (NCBI Built 36.3) obtained from the original dataset (Li et al., 2008; <ftp://ftp.cephb.fr/hgdp_supp1/>) were remapped on hg19 with the NCBI Genome Remapping Service (<http://www.ncbi.nlm.nih.gov/genome/tools/remap>). In doing so, we were not able to remap 74 SNPs, which were then excluded from the following analyses.

Information on the DNaseI Hypersensivity clusters were obtained online (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/>, volume 2, accessed on April 4, 2013) for the hg19 assembly and added to the annotation file. These clusters show DNaseI hypersensitive areas assayed in 125 human

cell types by the ENCODE Project (ENCODE Project Consortium, 2012) and may be indicative of regulatory regions. This information was used to calculate the distance of an outlier SNP to a putative functional region.

Detecting clusters of outlier SNPs

A sliding-window approach was implemented with R to identify significant clusters of candidate SNPs and remove isolated loci. We considered consecutive windows of size 500 kb-wide with 25 kb overlap. The *q*-value associated to each window was assumed as the 95% quantile of all included SNPs. Low density windows, i.e. those with less than one fifth of the average SNP density per chromosome, were set as non-significant. A graphical representation of this procedure was plotted with R taking into account the distribution of the SNPs, their particular *q*-value, the window *q*-value, and the best supported model of selection for each outlier SNPs. For the outlier SNPs, we also included information on their nearest gene if it was at most 50 kb apart.

The sliding-window approach yielded a second set of outlier markers. This SNP set is more refined than the one with the first candidates, since it ignores the low SNP density regions of the genome and those candidates that are isolated, highlighting genomic regions with a higher density of outlier SNPs. For each of these outliers, we also flagged cases in which the difference between the probabilities of the best and the second-best model of selection was less than 0.15, since the exact selection model (selection in one or both continents) is difficult to identify with high confidence.

Genetic bases of adaptation

To infer the biological processes and pathways associated with genes containing one or more outlier SNPs, we used the STRING 9.0 software (Franceschini et al., 2013) to identify Gene Ontology (GO) terms (Ashburner et al., 2000) and biological pathways

described in the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000; Kanehisa et al., 2012) that were enriched in the outlier gene lists considering a FDR of 0.1. For each PS only one list with all selection models (selection in one continent or both) was used. In doing so, we allow that two or more genes belonging to the same pathway, but that were inferred to be evolving under different selection models, to be considered as evolving under convergent evolution at the pathway-level and not only at the SNP- or gene-levels as the previous analysis.

When the STRING database used a gene alias different from that used for the annotation file, we resolved the ambiguity by considered the GeneCards Encyclopedia (<www.genecards.org>; Safran et al., 2010), except for five genes for which no correspondence was found in this online encyclopedia, which were then ignored in subsequent analyses.

To account for the fact that a gene with many SNPs is more likely to contain high scoring SNP than a gene with fewer SNPs just by chance, we checked if the outlier gene lists were biased for high SNP-count genes. For this, a Kolmogorov-Smirnov (KS) test implemented via R was employed to check if the distribution of SNP-counts from all annotated genes differed significantly from those of the outlier gene lists.

RESULTS

According to the F_{ST}-based Bayesian approach implemented via BayeScan (Foll and Gaggiotti, 2008), 1,482, 1,222, 1,579, and 1,365 outlier SNPs were identified in populations sets PS1 to PS4, respectively (the list, very extensive, will be available on request). From these, 75 are significant regardless of which PS is analyzed. Using a threshold of 50 kb to associate a SNP with a candidate gene, these sets yielded 568, 474, 620, and 517 genes respectively, of which 57 were found to be co-occurring in all four

PSs (*ABLIM3, ACSS2, AKAP6, ANKRD26, ARHGEF10, ATIC, BCAT1,* C20orf111, C2orf73, *CBLN1, CNTN4, CNTNAP5, COL22A1, CPA5, CRTC3, CWH43, DCUN1D4, DHCR7, EPHB4, FAM188B, FKBP6, GALNT16, GLIS3, GLRB, GPC6, GRIK2, HLA-DPA1, HMG20B, HSF2, IQGAP1, KIAA1598, KLHL29, LRRC66, MASTL, MPST, NAALADL2, NRG1, NRP2, NSUN5, PARK2, PKIB, PPP2R2C, RABGGTB, RAD51B, RBFOX1, RBM9, RBMS3, RFX3, ROBO2, SCP2, SGCB, SHISA6, SPAG16, SPATA13, SPATA18, ST6GAL1,* and *TRG@).*

The sliding-window procedure yielded 440, 399, 505, and 471 SNPs with significant signal of positive selection and included in significant sliding-windows in each case. There were no cases in which balancing selection could be inferred, i.e. α never assumed a significant negative value, which may be due to the characteristics of the genetic system employed (e.g. ascertainment bias) and not necessarily to the absence of this phenomena in the evolutionary history of these populations. Supplementary Information on chromosomal location of these SNPs, nearest gene (if located in a genic region), as well as the *q*-value and best supported model of selection, highlighting those cases in which the two best supported models present similar (<0.15) probabilities furnishes a very extensive list of 1,815 lines, which will be available on request.

Outlier SNPs are significantly enriched (p-values < 0.00017) for genic SNPs in comparison to the whole HGDP SNP set. In average 72.0% of the outlier SNPs after the sliding-window procedure are located in genic regions considering all four PSs, while in the HGDP database this number is only 63.6%. The outlier SNPs were located at maximum 45.8kb apart from a gene or a putative regulatory element indicated by a hypersensibility to DNaseI as described by the ENCODE project.

Figure 1 shows the Manhattan plots of the distribution of the SNPs in the genome for the different PSs and their estimated *q*-value, as well as the best model for selection (color-coded) into the sliding-window approach. It was possible to identify seven clusters of outlier SNPs co-occurring in all four PSs (indicated by gray-shaded boxes at Figure 1).

For the first two clusters – located at 1p32.3 and 2q32.1 (Figures 2 and 3) – the best supported model is positive selection in the Americas, although convergent evolution cannot be ruled out in a few cases. Cluster 1 is delimited by two SNPs (rs7550236 and rs6679819) that are 43.1 kb apart from each other. The latter and a few other SNPs from this cluster occur inside *SCP2* (Figure 2). Cluster 2 contains four SNPs (rs17715017, rs1733497, rs1439771, and rs2119047; the outermost SNPs are separated by 25.2 kb) in an inter-genic region (Figure 3). This signal occurs in a wider region for PS3 and PS4, reaching SNPs rs12617731 and rs2165172, that are 206.6kb apart from each other.

The next two clusters of significant SNPs occur on the same chromosome (Cluster 3 at 4p11 and Cluster 4 at 4q12; Figure 1) and selection in Africa is in general the best supported model. Cluster 3 comprises rs2605267 and rs2572363 (the latter could also be evolving under convergent adaptation in Africa and the Americas) and is associated to *CWH43* (Figure 4). Cluster 4 is delimited by rs4865414 and rs1460554, which are 177.4 kb apart from each other, and comprises SNPs that fall into genes such as *DCUN1D4, LRRC66,* and *SGCB* (Figure 4).

Convergent evolution is the best supported model of selection for the SNPs falling into the other three clusters, which are located at 5p12, 6p22.31, and 7q11.23 (Figure 1). One of the extremities of Cluster 5 is defined by rs6875400, which is found to be

significant in all PSs and is located in an inter-genic region. The other extremity of this cluster is defined by two different SNPs depending on the analyzed PS. For PS1 and PS3, this SNP is rs6895327, located circa 365 kb away from rs6875400, and occurring inside C5orf34, delimiting a genomic region in which *NNT* is included and presents significant outlier SNPs associated to it (Figures 5A and 5C,). For PS2 and PS4, the outermost significant SNP is rs4264950, located more than 445 kb from rs6875400 and around 27 kb from *CCL28* (Figures 5B and 5D). Although the whole region delimited by significant SNPs in the four different PSs include C5orf34 and *NNT*, no outlier SNP was found to be associated to these genes in PS2 and PS4. Cluster 6 comprises *HSF2* and *PKIB* and is delimited by rs3778348 and rs9320878 (Figure 6). The analyses of PS1 and PS3 revealed an additional significant SNP at this cluster (rs487098), which is located at *SERINC1.* The last cluster (no. 7) includes two SNPs that are 27.3 kb apart: rs1178970, located in a *FKBP6* intron*;* and rs1880948, located 0.12 kb apart from *NSUN5* and 3.55 kb from *TRIM50*.

The abovementioned 14 genes that include or are close to SNPs that present signals of positive selection after the sliding-windows approach for all four PSs are described in Table 1.

Genes with at least one significant outlier SNP presented strong SNP-count bias (KS test p-value < 1e-15). Hence, these gene lists were not analyzed for GO terms enrichment or KEGG pathways. Nonetheless the gene lists after the sliding-windows procedure were unbiased when all PSs were pooled and repeated genes were eliminated (KS test p-value = 0.07) and also for PS2-4 independently (p-values for the KS test equals to 0.07, 0.26, and 0.79 respectively); but not for PS1 (KS test p-value = 0.03). Although several GO terms and KEGG pathways were enriched at the analyses, no GO term

remained significant after correction for multiple testing (FDR=0.1; results not shown), even considering the 57 ones found to be co-occurring in all four PSs, and only two KEGG pathways remained significant, namely "Apoptosis" (with a *q*-value of 0.07 for PS2) and "Axon guidance" (for PS2-4, with *q*-values of 0.09, 0.02, and 0.01 respectively). The first pathway includes the following genes with outlier SNPs for PS2: *PPP3CA, PPP3CB, PRKAR2B,* and *XIAP.* Genes occurring in the "Axon guidance" pathway with a significant signal of positive selection are *PPP3CA, PPP3CB, ROBO2*, and *UNC5C* for PS2; *ABLIM3, EPHB4, LIMK2*, *NTN4, ROBO2*, and *UNC5D* for PS3; and *EPHB4, LIMK2*, *NTN4, PPP3CB, ROBO2*, and *UNC5D* for PS4.

DISCUSSION

Tropical forests are believed to be very harsh habitats for human beings (Hart and Hart, 1986). In addition to being almost deprived from energy-rich food and edible plants (Bailey et al., 1989), these environments are very propitious for the development of diseases (Guernier et al., 2004) and might also compromise thermoregulation (Perry and Dominy, 2009). In this work we sought to identify human adaptations to these environments by employing a Bayesian method (Foll and Gaggiotti, 2008) to identify SNPs in a genome-wide dataset showing overly large or low extent of differentiation between tropical and non-tropical forest populations. We sought to infer signals of convergent evolution by comparing native populations from tropical forests (Biaka, Mbuti, Karitiana, and Surui) with genetically related populations living elsewhere (Mandenka, Yoruba, and Pima) combined in four different population sets (PSs). The analysis suggested some SNPs, genes, and biological pathways in which convergence and positive selection could be inferred, highlighting biological functions associated to immunology, nervous system development, and lipid metabolism, among others, that

may have presented adaptive advantage in tropical forests. The few cases in which the same signal could be identified by employing different combinations of populations (clusters 1-7 at Figure 1), indicated that the inferred signal is likely due to environmental selective pressures and adaptation rather than any particularity of the demographic history of the analyzed populations. The outlier regions are enriched for genic loci. Those few cases that have fallen inside an intergenic region, such as Cluster 2 (Figure 2), could always be associated (distance < 50 kb) to a putative regulatory element described by the ENCODE project. The results show that at least some recent human adaptations are indeed based in regulatory regions (Grossman et al., 2013) and that selection had played an important role in shaping human genetic diversity.

In a previous study involving Native Americans, Hünemeier et al. (2012) suggested that *ABCA1*, a gene encoding a cholesterol efflux regulatory protein, was evolving under positive selection due to limited food resources that Native American encountered during their history. We also found significant signals of positive selection in genes that are related to lipid circulation and metabolism, such as *SCP2* and *CWH43* (Figures 1, 2 and 4; Table 1). While the first gene is found to be probably evolving under positive selection in the Americas, the latter presents this pattern in African populations. Both might have developed cellular mechanisms for energy maintenance in response to food scarcity pressures. However other factors could be involved, since cholesterol plays an important role in various infectious processes such as virus invasion and replication (Lee at al., 2008) and resistance against malaria (Combes et al., 2005).

In this regard, Sabeti et al. (2006) in their review noticed the preponderance of genes related to the immune system in the available genome-wide scans for positive selection. This preponderance was further confirmed by Williamson et al. (2007) and

López Herráez et al. (2009), and ultimately by our analyses. Besides the two abovementioned genes, *SCP2* and *CWH43*, which have a possible effect on immunology, the protein encoded by *CCL28*, which was found in the vicinities of Cluster 5 and present SNPs with signal of convergent evolution in PS2 and PS4 (Figure 5, right column), modulates immunity to HIV infection and skin-related inflammatory diseases (Table 1).

Another category of genes frequently presenting signals of positive selection is fertility, more specifically, spermatozoid development (Sabeti et al., 2006). In this regard, *FKBP6,* a male fertility factor (Table 1), was also suggested to be evolving under positive selection with the analyses of the tropical forest populations of Africa and the Americas (Figure 7).

Heat-shock transcription factors, such as that encoded by *HSF2* (Cluster 6; Figures 1 and 6), are activated by stress and respond to elevated temperatures. One of the consequences of inefficient thermoregulation is the increase of body temperature. The observed positive selection signals at SNPs found in this gene could be due to an adaptation to the tropics, initiating gene(s) transcription in response to high body temperatures. Another study with African-, European-American, and Chinese populations also found a significant signal of positive selection in heat shock genes (Williamson et al., 2007), suggesting that this category might have some importance in human adaptation to different environments.

In the search for enriched KEGG pathways, the "Axon guidance" category is significant in more than one PS (PS2-4), even after correction for multiple testing. The exact nature of this signal is unknown and thus deserves further investigation.

It is generally accepted that the pygmy phenotype might have evolved as an adaptation to life in dense tropical forests, to thermoregulation, and to food scarcity

(Diamond, 1991; Perry and Dominy, 2009); or as a by-product of selection for early onset of reproduction (Migliano et al., 2007). Our research design enables the comparison of two African pygmy populations with two other non-pygmy populations from the same continent besides this tropical/non-tropical dichotomy in order to identify possible targets for positive selection that could be related to the pygmy phenotype. The two best candidate regions for this comparison are those of Clusters 3 and 4 (Figures 1 and 4). The first presents a gene involved in lipid metabolism (*CWH43*, discussed above). The second includes *DCUN1D4*, *LRRC66*, and *SGCB*, and is found in a locus that, when subjected to a homozygous microdeletion, leads to severe limb-girdle muscular Duchenne-like dystrophy, combined with hyperlaxity and join contractures, chest pain, palpitations, and dyspnea (Kaindl et al., 2005). Both could be involved in body growth and development, and these regions deserve further attention in future investigations on human height and pygmy evolution.

Fifteen genomic regions have been found to be associated with the pygmy phenotype by means of covariation between allele frequencies and body height in Africa (Mendizabal et al., 2012). In four of these regions we also found signals of positive selection in the African continent. By comparing Mbuti to West Africa two regions were highlighted (regions 12 and 14 as named by Mendizabal et al., 2012). In our study, the first region is among those with the highest *q*-values, and is located in the long arm of chromosome 10 (Figure 1) spanning over 644 kb. It presents 14 outlier SNPs in the following 11 genes: *P4HA1, NUDT13, ECD, FAM149B1, DNAJC9, MRPS16, ANXA7, ZMYND17, USP54, PPP3CB,* and *TTC18*. One of them, *PPP3CB,* encodes a subunit of calcineurin, a protein that regulates bone formation by osteoblast differentiation (Sun et al., 2005). Additionally, two polymorphisms in this region – rs2271904 and rs4294502 –

present signals of positive selection and are both non-synonymous mutations in *ECD* and *TTC18* respectively.

The second region is defined by two SNPs (rs7174731 and rs7181518) in *TRIP4*. This gene encodes a thyroid hormone receptor interactor and was also suggested to be related to the pygmy phenotype by López Herráez et al. (2009). Since African pygmies live in inland tropical forests, regions with limited access to iodine-rich foods, these authors suggested that genetic changes in this gene could be related to an adaptation to their iodine-deficient environment and responsible for their short stature, since changes in the thyroid hormone pathway can cause growth deficiency (Sultan et al., 2008).

Finally, considering Biaka instead of Mbuti, we found two genes with positive selection signals: *USP46* and *MLL3*. The latter is involved in histone modification, a category associated with height in a genome-wide association study (Lango Allen et al., 2010). *USP46* has been associated with cognition and behavior of humans and mice (Fukuo et al., 2011; Zhang et al., 2011).

An additional region suggested by Mendizabal et al. (2012) to be associated with the pygmy phenotype presents signals of convergent adaptation using all four different combinations of populations in our study. They found significant SNPs in *NNT*, while we observed a more diffuse significant signal also including *CCL28* in some cases (Figure 5).

Other genes listed in Table 1 indicating a possible role in human adaptation to the tropics and not yet discussed, whose function is known, include: (a) *PKIB* related to cell division; and (b) *NSUN5* and *TRIM50*, involved with the development of the vascular system and with calcium metabolism. Four genes (*PPP3CA, PPP3CB, PRKAR2B,* and *XIAP*), involved with apoptosis, present significant values in PS2 and could also play a role in this ontogenetic process.

CONCLUSIONS

The F_{ST}-based Bayesian method employed in our study was able to detect some regions with positive selection signal suggesting that the following biological functions and pathways may play a role in human adaptations to tropical forest: lipid metabolism, immunology, body development, axon guidance, and heat stress response. The same signal in different population sets may assure that they are due to environment adaptation and not to any demographic peculiarity of the chosen populations. Further refinement of these analyses with the input of full genome or exome sequence information for these populations could reveal which particular mutations are responsible for the fitness variation among individuals and the specific biological function affected by them.

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The GeneOntology Consortium. Nat Genet 25: 25-9

Bailey RC, Head G, Jenike M, Owen B Rechtman R, Zechnter E (1989) Hunting and gathering in tropical rain forest: is it possible? Am Anthropologist 91: 59-82.

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol 63: 221-30.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13: 969-80.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57: 289–300.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. Science 296: 261-2.

Castelletti E, Lo Caputo S, Kuhn L, Borelli M, Gajardo J, Sinkala M, Trabattoni D, Kankasa C, Lauri E, Clivio A, Piacentini L, Bray DH, Aldrovandi GM, Thea DM, Veas F, Nebuloni M, Mazzotta F, Clerici M (2007) The mucosae-associated epithelial chemokine (MEC/CCL28) modulates immunity in HIV infection. PLoS One 2: e969.

Chung S, Tamura K, Furihata M, Uemura M, Daigo Y, Nasu Y, Miki T, Shuin T, Fujioka T, Nakamura Y, Nakagawa H (2009) Overexpression of the potential kinase serine/ threonine/tyrosine kinase 1 (STYK 1) in castration-resistant prostate cancer. Cancer Sci 100: 2109-14.

Combes V, Coltel N, Alibert M, van Eck M, Raymond C, Juhan-Vague I, Grau GE, Chimini G (2005) ABCA1 gene deletion protects against cerebral malaria: potential pathogenic role of microparticles in neuropathology. Am J Pathol 166: 295-302.

Diamond JM (1991) Anthropology. Why are pygmies small? Nature 354: 111-2.

Doll A, Grzeschik KH (2001) Characterization of two novel genes, WBSCR20 and WBSCR22, deleted in Williams-Beuren syndrome. Cytogenet Cell Genet 95: 20-7.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103: 285-98.

Ezzat MH, Sallam MA, Shaheen KY (2009) Serum mucosa-associated epithelial chemokine (MEC/CCL28) in atopic dermatitis: a specific marker for severity. Int J Dermatol 48: 822-9.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977-93.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808-15.

Fukuo Y, Kishi T, Kushima I, Yoshimura R, Okochi T, Kitajima T, Matsunaga S, Kawashima K, Umene-Nakano W, Naitoh H, Inada T, Nakamura J, Ozaki N, Iwata N (2011) Possible association between ubiquitin-specific peptidase 46 gene and major depressive disorders in the Japanese population. J Affect Disord 133: 150-7.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, Cabili M, Adegbola RA, Bamezai RN, Hill AV, Vannberg FO, Rinn JL; 1000 Genomes Project, Lander ES, Schaffner SF, Sabeti PC. (2013) Identifying recent adaptations in large-scale genomic data. Cell 152: 703-13.

Guernier V, Hochberg ME, Guégan JF (2004) Ecology drives the worldwide distribution of human diseases. PLoS Biol 2: e141.

Hart TB, Hart JA (1986) The ecological basis of hunter-gatherer subsistence in African rain forests: the Mbuti of Eastern Zaire. Hum Ecol 14: 29-55.

Hünemeier T, Amorim CEG, Azevedo S, Contini V, Acuña-Alonzo V, Rothhammer F, Dugoujon JM, Mazières S, Barrantes R, Villarreal-Molina MT, Paixão-Côrtes VR, Salzano FM, Canizales-Quinteros S, Ruiz-Linares A, Bortolini MC. (2012) Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. PLoS One 7: e38862.

ISA – Instituto Socioambiental (2013) Povos Indígenas do Brasil <http://pib.socioambiental.org> Website accessed on April 26, 2013.

Kaindl AM, Jakubiczka S, Lücke T, Bartsch O, Weis J, Stoltenburg-Didinger G, Aksu F, Oexle K, Koehler K, Huebner A (2005) Homozygous microdeletion of chromosome 4q11-q12 causes severe limb-girdle muscular dystrophy type 2E with joint hyperlaxity and contractures. Hum Mutat 26: 279-80.

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109-14

Kriska T, Pilat A, Schmitt JC, Girotti AW (2010) Sterol carrier protein-2 (SCP-2) involvement in cholesterol hydroperoxide cytotoxicity as revealed by SCP-2 inhibitor effects. J Lipid Res 51: 3174-84.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Mägi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S,

Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, König IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Müller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpeläinen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Paré G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietiläinen KH, Pouta A, Ridderstråle M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kähönen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimäki T, Melander O, Mosley TH Jr, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tönjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC,

Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Grönberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Völzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-8.

Lee RB, Daly R (2004) The Cambridge Encyclopedia of Hunters and Gatherers. Cambridge University Press. 534 p.

Lee CJ, Lin HR, Liao CL, Lin YL (2008) Cholesterol effectively blocks entry of flavivirus. J Virol 82: 6470-80.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-4.

López Herráez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M (2009) Genetic variation and recent positive selection in

worldwide human populations: evidence from nearly 1 million SNPs. PLoS One 4: e7888.

Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39 (Database issue): D52-7.

Meimaridou E, Kowalczyk J, Guasti L, Hughes CR, Wagner F, Frommolt P, Nürnberg P, Mann NP, Banerjee R, Saka HN, Chapple JP, King PJ, Clark AJ, Metherell LA (2012) Mutations in NNT encoding nicotinamide nucleotide transhydrogenase cause familial glucocorticoid deficiency. Nat Genet 44: 740-2.

Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African Pygmy phenotype. Hum Genet 131: 1305-17.

Mercader J (2002) Forest people: the role of African rainforests. Evol Anthropol 11: 117-24.

Micale L, Fusco C, Augello B, Napolitano LM, Dermitzakis ET, Meroni G, Merla G, Reymond A (2008) Williams-Beuren syndrome TRIM50 encodes an E3 ubiquitin ligase. Eur J Hum Genet 16: 1038-49.

Migliano AB, Vinicius L, Lahr MM (2007) Life history trade-offs explain the evolution of human pygmies. Proc Natl Acad Sci U S A 104: 20216-9.

Ohenjo N, Willis R, Jackson D, Nettleton C, Good K, Mugarura B (2006) Health of indigenous people in Africa. Lancet 367: 1937-46.

Perry GH, Dominy NJ (2009) Evolution of the human pygmy phenotype. Trends Ecol Evol 24: 218-25.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-75.

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ratnam J, Bond WJ, Fensham RJ, Hoffmann WA, Archibald S, Lehmann CER, Anderson MT, Higgins SI, Sankaran M (2011) When is a 'forest' a savanna, and why does it matter? Glob Ecol Biogeogr 20: 653-60.

Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70: 841-7.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298: 2381-5.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. Science 312: 1614-20.

Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A and Lancet D (2010) GeneCards Version 3: the human gene integrator database. doi: 10.1093/database/baq020

Sandqvist A, Björk JK, Akerfelt M, Chitikova Z, Grichine A, Vourc'h C, Jolly C, Salminen TA, Nymalm Y, Sistonen L (2009) Heterotrimerization of heat-shock factors 1 and 2 provides a transcriptional switch in response to distinct stimuli. Mol Biol Cell 20: 1340-7.

Sultan M, Afzal M, Qureshi SM, Aziz S, Lutfullah M, Khan SA, Iqbal M, Maqsood SU, Sadiq N, Farid N (2008) Etiology of short stature in children. J Coll Physicians Surg Pak 18: 493-7.

Sun L, Blair HC, Peng Y, Zaidi N, Adebanjo OA, Wu XB, Wu XY, Iqbal J, Epstein S, Abe E, Moonga BS, Zaidi M (2005) Calcineurin regulates bone formation by the osteoblast. Proc Natl Acad Sci USA 102: 17130-5.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035-44.

Umemura M, Fujita M, Yoko-O T, Fukamizu A, Jigami Y (2007) *Saccharomyces cerevisiae* CWH43 is involved in the remodeling of the lipid moiety of GPI anchors to ceramides. Mol Biol Cell 18: 4304-16.

Walker RS, Hamilton MJ (2008) Life-history consequences of density dependence and the evolution of human body size. Curr Anthr 49: 115-22.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado AM, Labuda D, Klitz W, Barrantes R, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A. (2007) Genetic variation and population structure in Native Americans. PLoS Genet 3: e185.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90.

Zhang W, Zhang S, Xiao C, Yang Y, Zhoucun A (2007) Mutation screening of the FKBP6 gene and its association study with spermatogenic impairment in idiopathic infertile men. Reproduction 133: 511-6.

Zhang W, Tian QB, Li QK, Wang JM, Wang CN, Liu T, Liu DW, Wang MW (2011) Lysine 92 amino acid residue of USP46, a gene associated with 'behavioral despair' in mice, influences the deubiquitinating enzyme activity. PLoS One 6: e26297.

Table 1. Genes with signals of positive selection in the four sets of comparisons made (PS1-4) suggesting human adaptations to tropical forests in Africa and the Americas.

Gene	Alias	Cluster	Biological function in mammalians or associated human diseases
Sterol carrier protein-2	*SCP2*	1	Involved in cholesterol trafficking and metabolism[1].
Cell wall biogenesis 43 C-terminal homolog	*CWH43*	3	Enhance lipid remodeling to ceramides[2].
Defective in cullin neddylation 1 domain containing 4	*DCUN1D4*	4	Unknown.
Leucine rich repeat containing 66	*LRRC66*	4	Unknown.
Sarcoglycan beta	*SGCB*	4	Is located in a genomic region where a microdeletion causes limb-girdle muscular dystrophy type 2E with joint hyperlaxity and contractures[3].
Chromosome 5 open reading frame 34	C5orf34	5	Unknown.
Chemokine (C-C motif) ligand 28	*CCL28*	5	Modulate immunity to viral infection[4] and skin-related inflammatory diseases[5].
Nicotinamide nucleotide transhydrogenase	*NNT*	5	Produces high concentrations of NADPH at mitochondria and the resulting energy is used for biosynthesis and in free-radical detoxification[6].
Heat shock transcription factor 2	*HSF2*	6	Involved in the activation of heat-shock response genes under conditions of heat[7].
cAMP-dependent protein kinase catalytic inhibitor beta	*PKIB*	6	Associated to the aggressive phenotype of prostate cancer[8].
Serine incorporator 1	*SERINC1*	6	Unknown.
FK506 binding protein 6	*FKBP6*	7	May play a role in modifying the susceptibility to idiopathic spermatogenic impairment[9].
NOP2/Sun domain family member 5	*NSUN5*	7	Deleted in Williams-Beuren syndrome (vascular system and calcium metabolism problems)[10].
Tripartite motif containing 50	*TRIM50*	7	May be involved in the Williams-Beuren syndrome[11].

References: [1]Kriska et al. (2010); [2]Umemura et al. (2007); [3]Kaindl et al. (2005); [4]Castelletti et al. (2007); [5]Ezzat et al. (2009); [6]Meimaridou et al. (2012); [7]Sandqvist et al. (2009); [8]Chung et al. (2009); [9]Zhang et al. (2007); [10]Doll and Grzeschik (2001); [11]Micale et al. (2008).

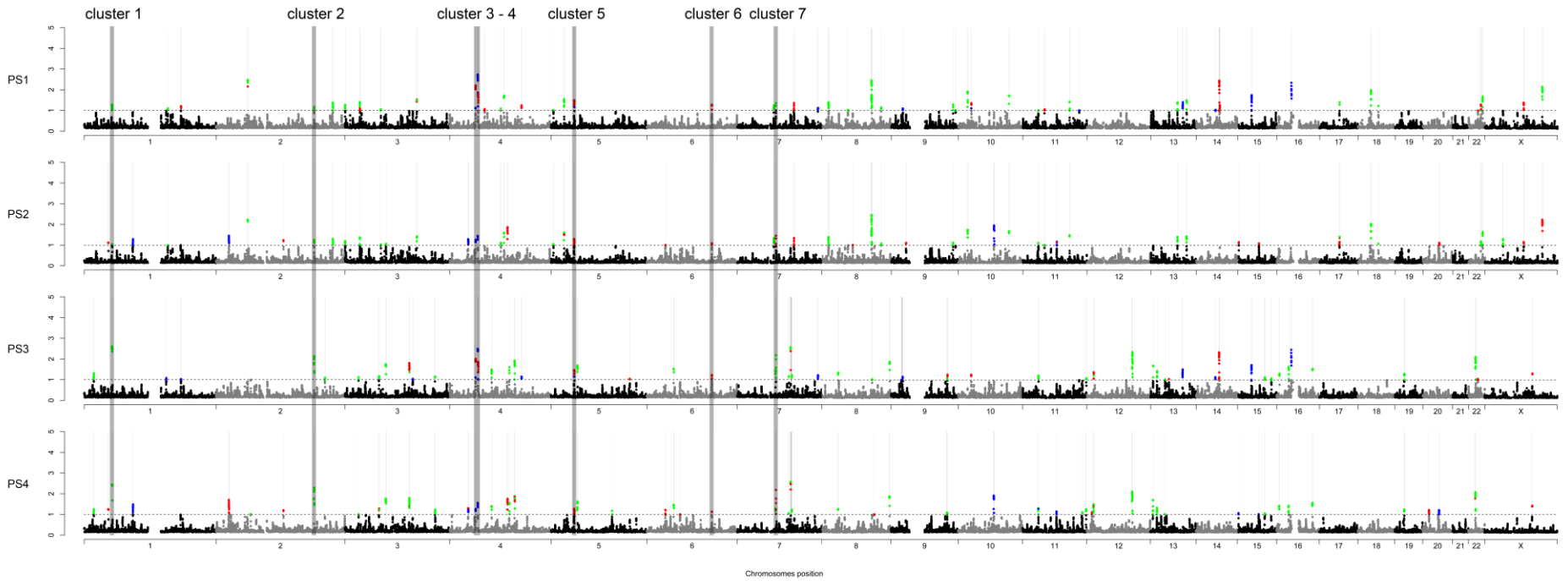

Figure 1: Manhattan plots of the genome-wide distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black or grey), in Africa (blue), in the Americas (green), and in both continents (convergent evolution, red). Different sets of populations were used in the analysis yielding four different population sets (PS1-4). Congruent clusters of outlier SNPs considering all four PSs are highlighted with a grey box.

Figure 2A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 1. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 2A; PS2: 2B; PS3: 2C; PS4: 2D).

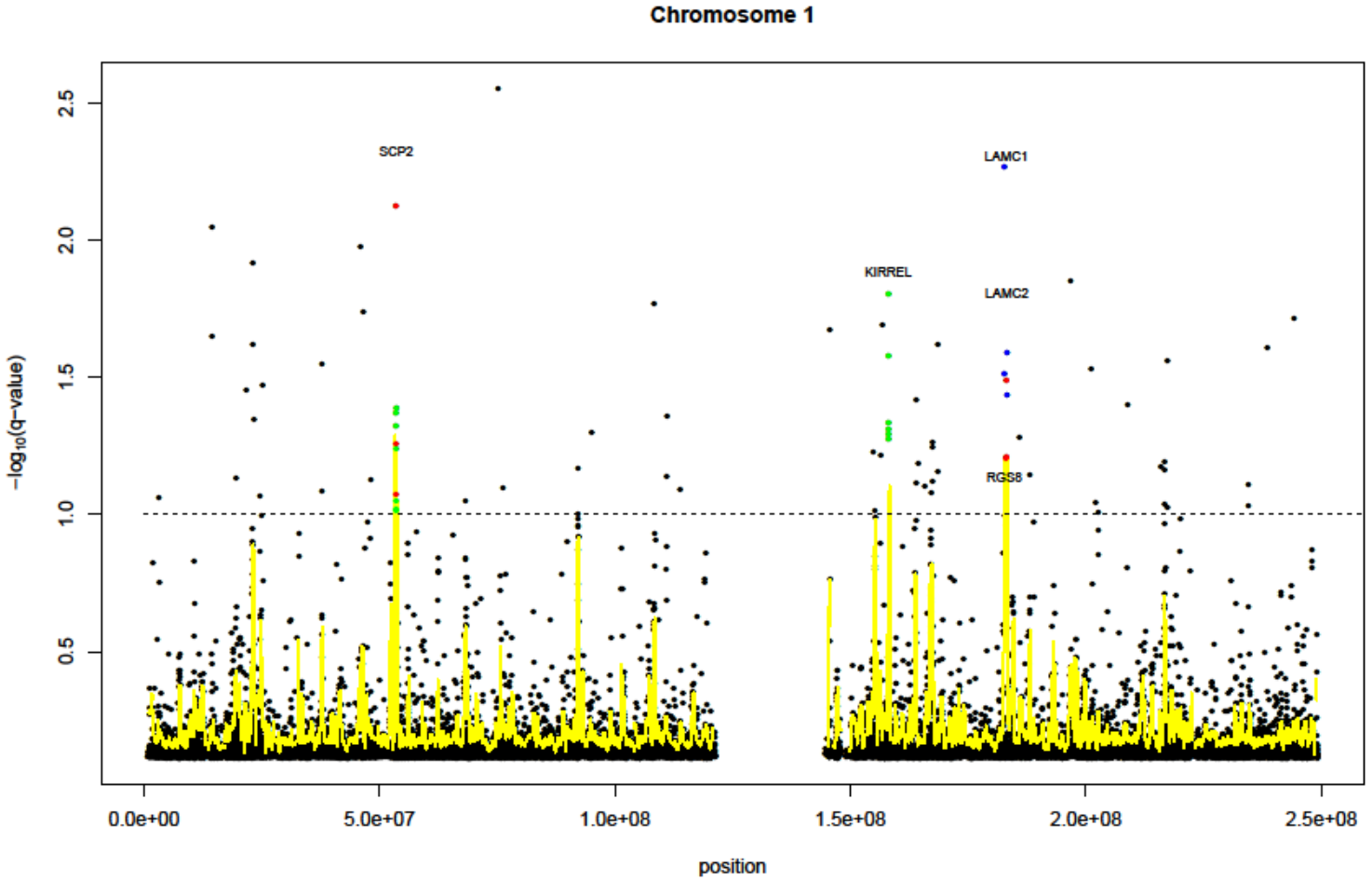
Chromosome 1
SCP2
LAMC1
KIRREL
LAMC2
RGS8
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 2A

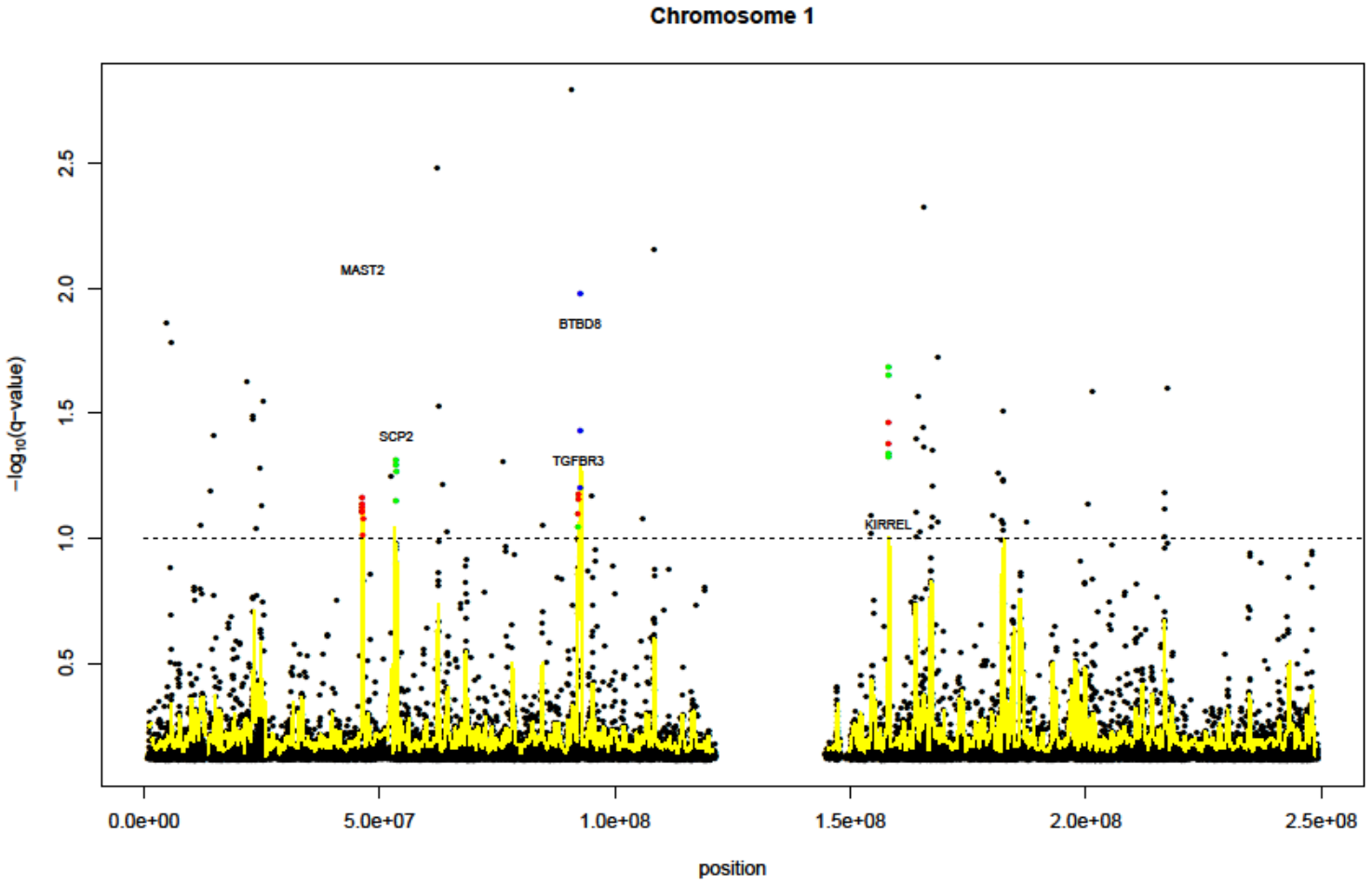
Chromosome 1
−log10(q−value)
0.5
1.0
1.5
2.0
2.5
MAST2
BTBD8
SCP2
TGFBR3
KIRREL
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 2B

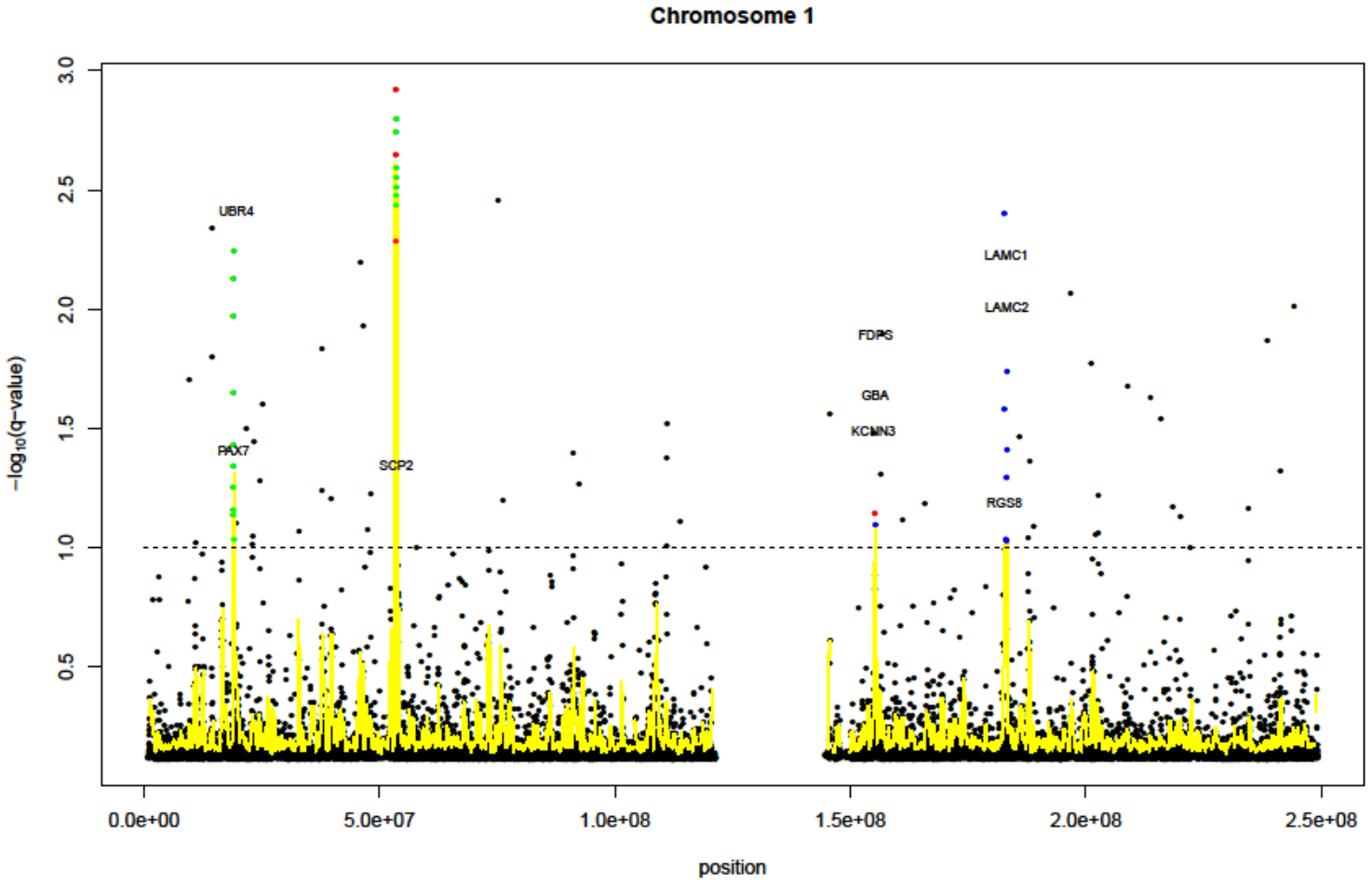
Chromosome 1
−log10(q−value)
3.0
2.5
2.0
1.5
1.0
0.5
UBR4
PAX7
SCP2
FDPS
GBA
KCNN3
LAMC1
LAMC2
RGS8
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 2C

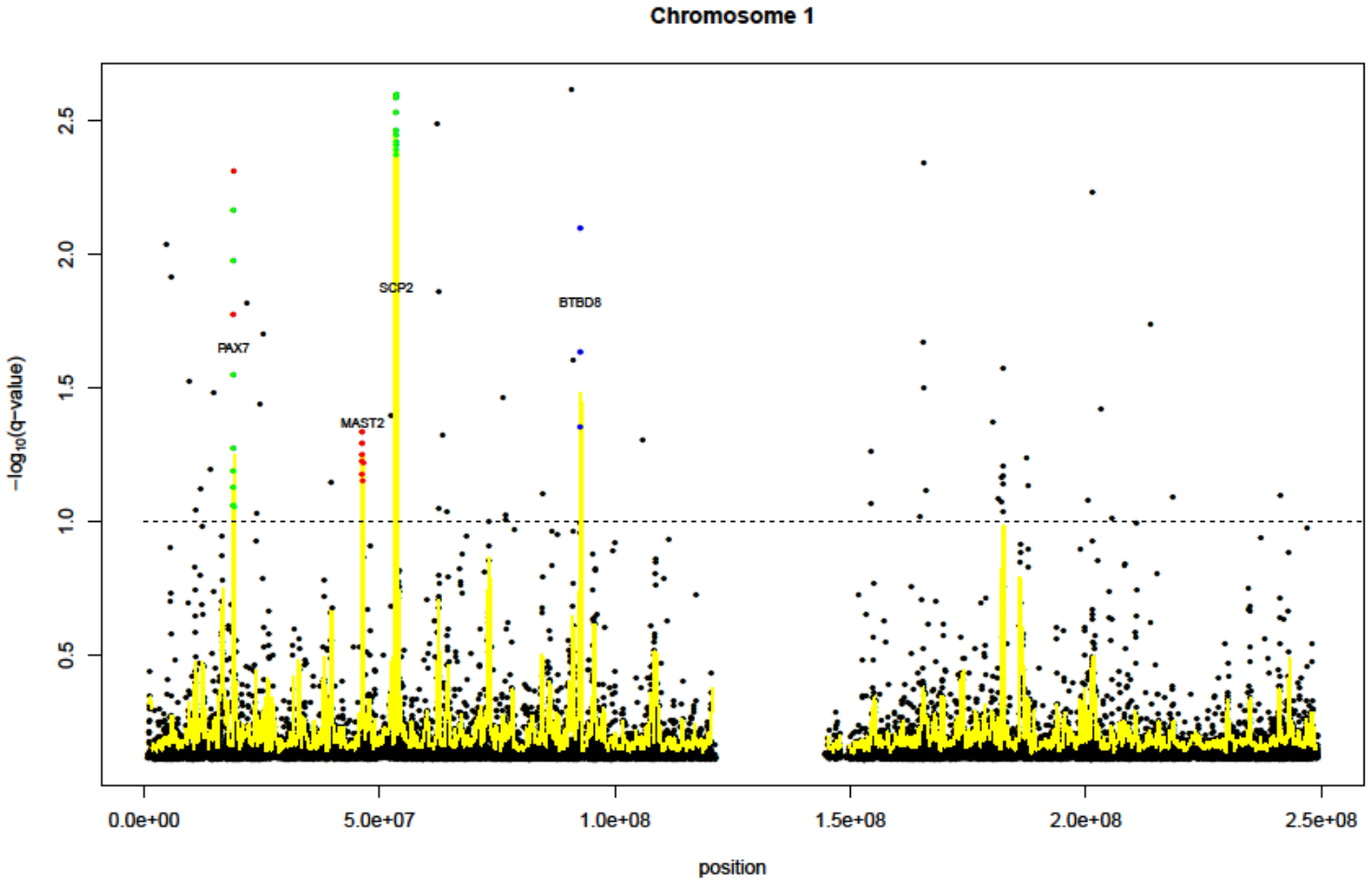
Chromosome 1
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
PAX7
SCP2
BTBD8
MAST2
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 2D

Figure 3A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 2. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 3A; PS2: 3B; PS3: 3C; PS4: 3D).

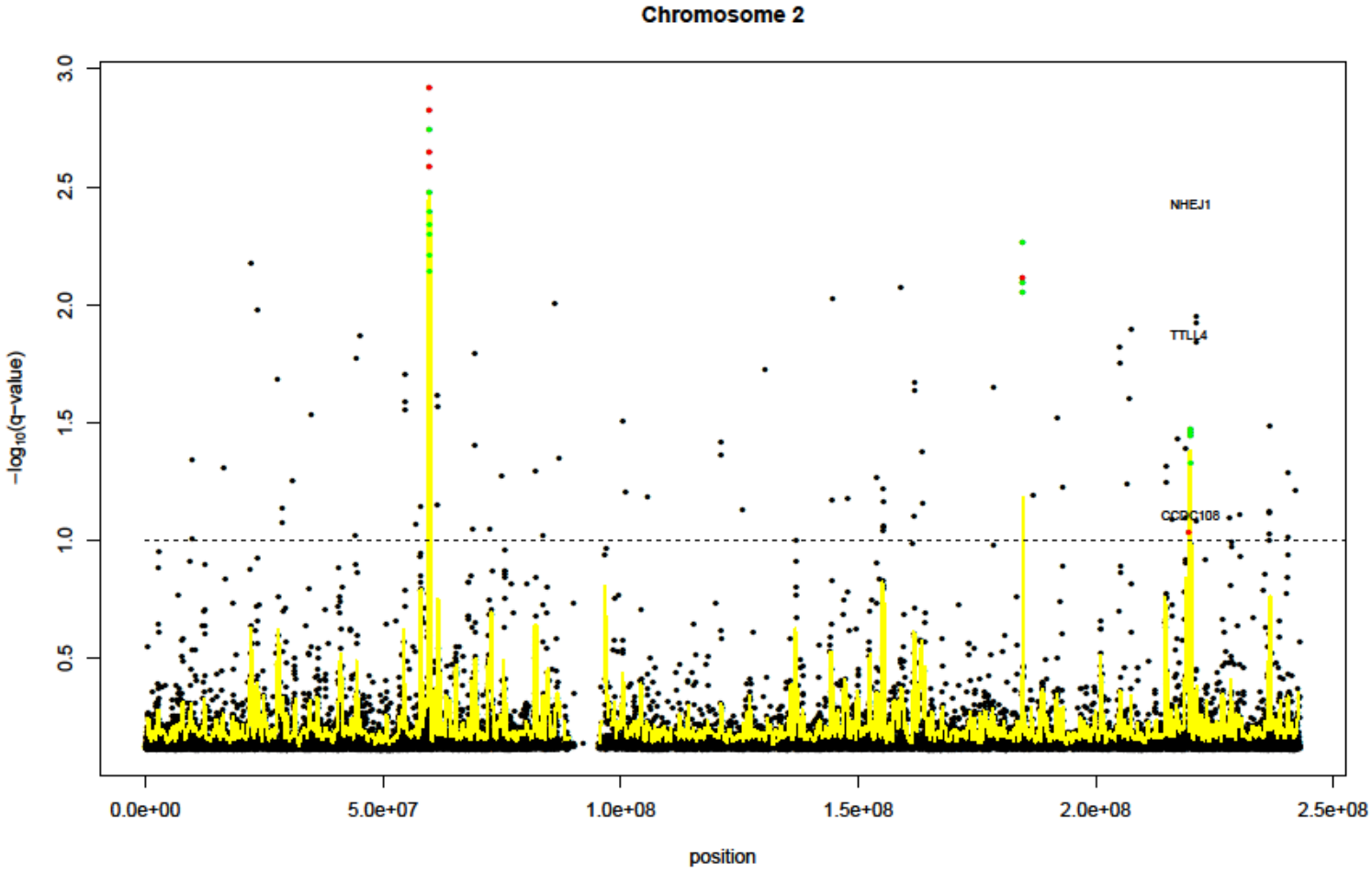
Chromosome 2
−log10(q−value)
3.0
2.5
2.0
1.5
1.0
0.5
NHEJ1
TTLL4
CCDC108
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 3A

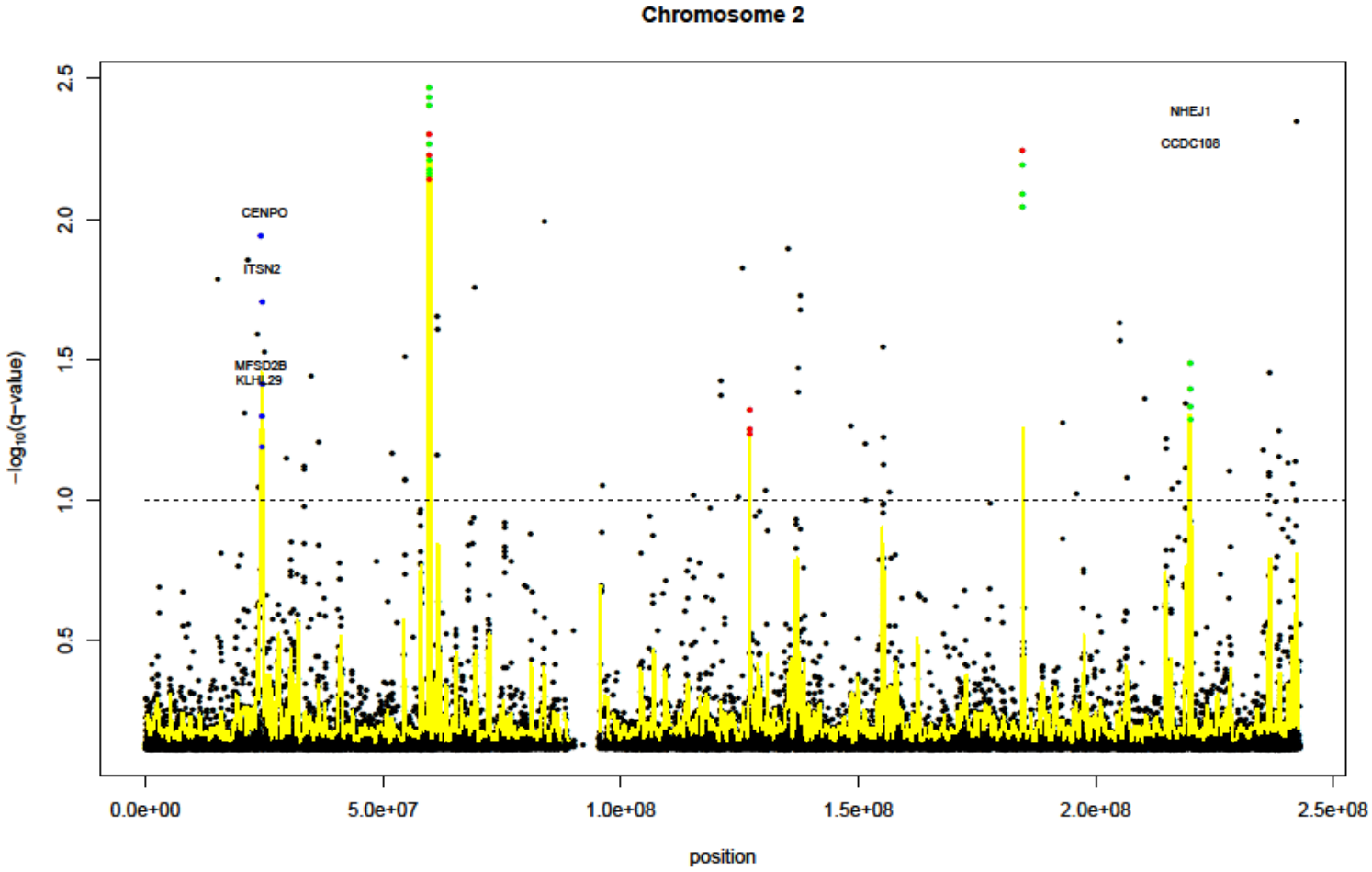
Chromosome 2
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
CENPO
ITSN2
MFSD2B
KLHL29
NHEJ1
CCDC108
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 3B

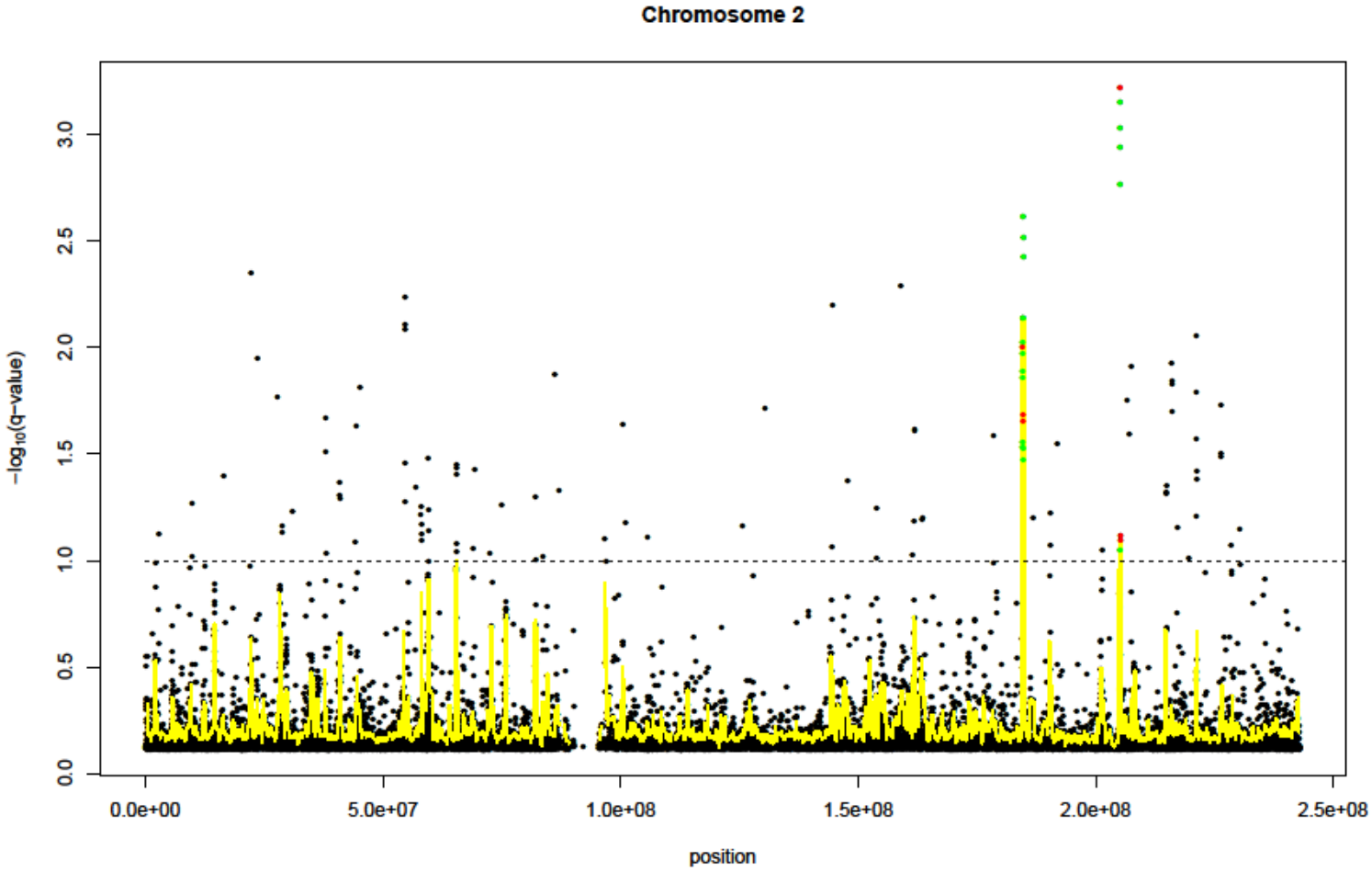
Chromosome 2
−log10(q−value)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 3C

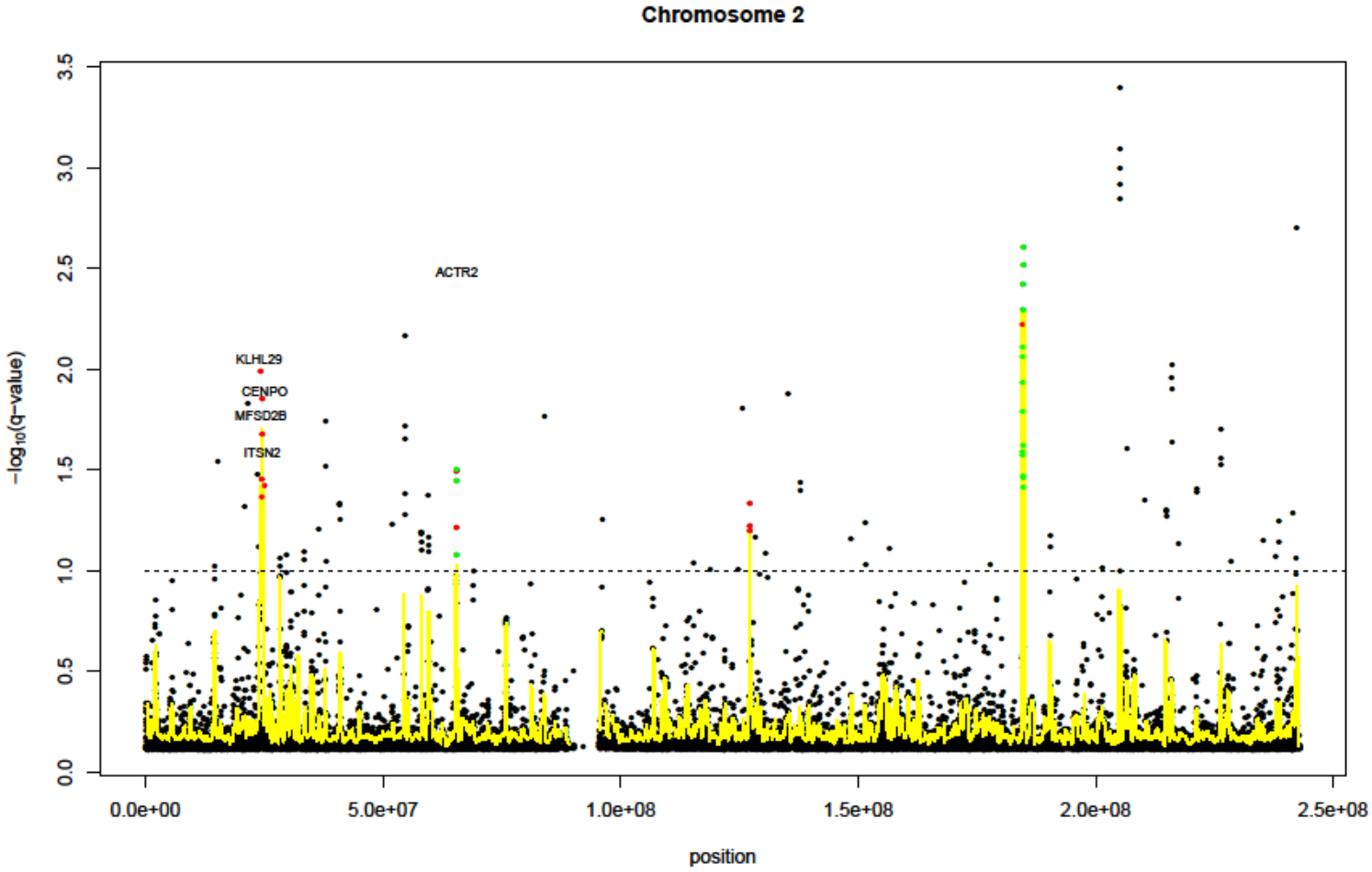
Chromosome 2
−log10(q−value)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
ACTR2
KLHL29
CENPO
MFSD2B
ITSN2
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
position

Figure 3D

Figure 4A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 4. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 4A; PS2: 4B; PS3: 4C; PS4: 4D).

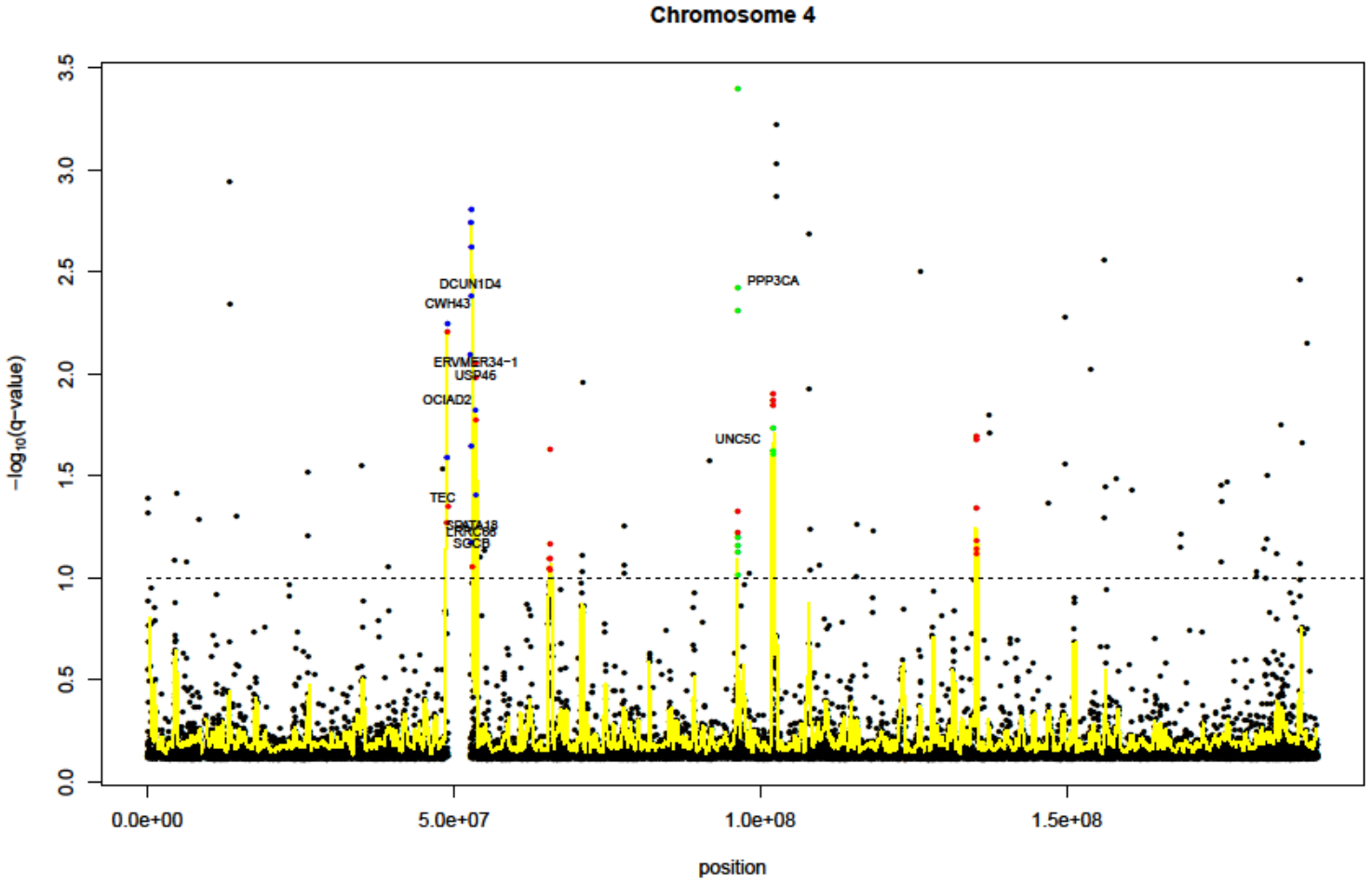
Chromosome 4
−log10(q−value)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
DCUN1D4
CWH43
ERVMER34−1
USP45
OCIAD2
TEC
SGCB
PPP3CA
UNC5C
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 4A

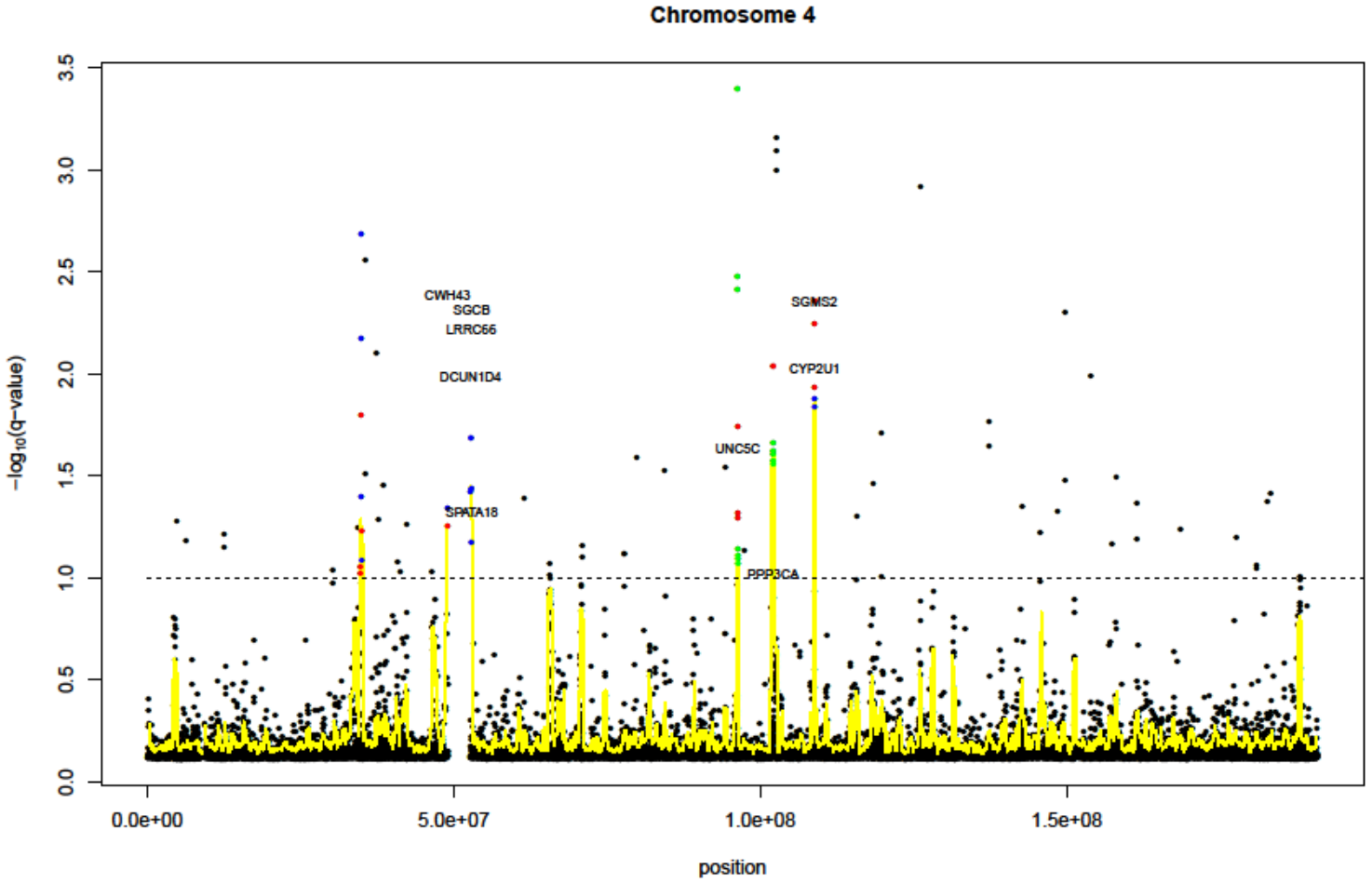
Chromosome 4
−log10(q−value)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
CWH43
SGCB
LRRC66
DCUN1D4
SPATA18
UNC5C
PPP3CA
SGMS2
CYP2U1
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 4B

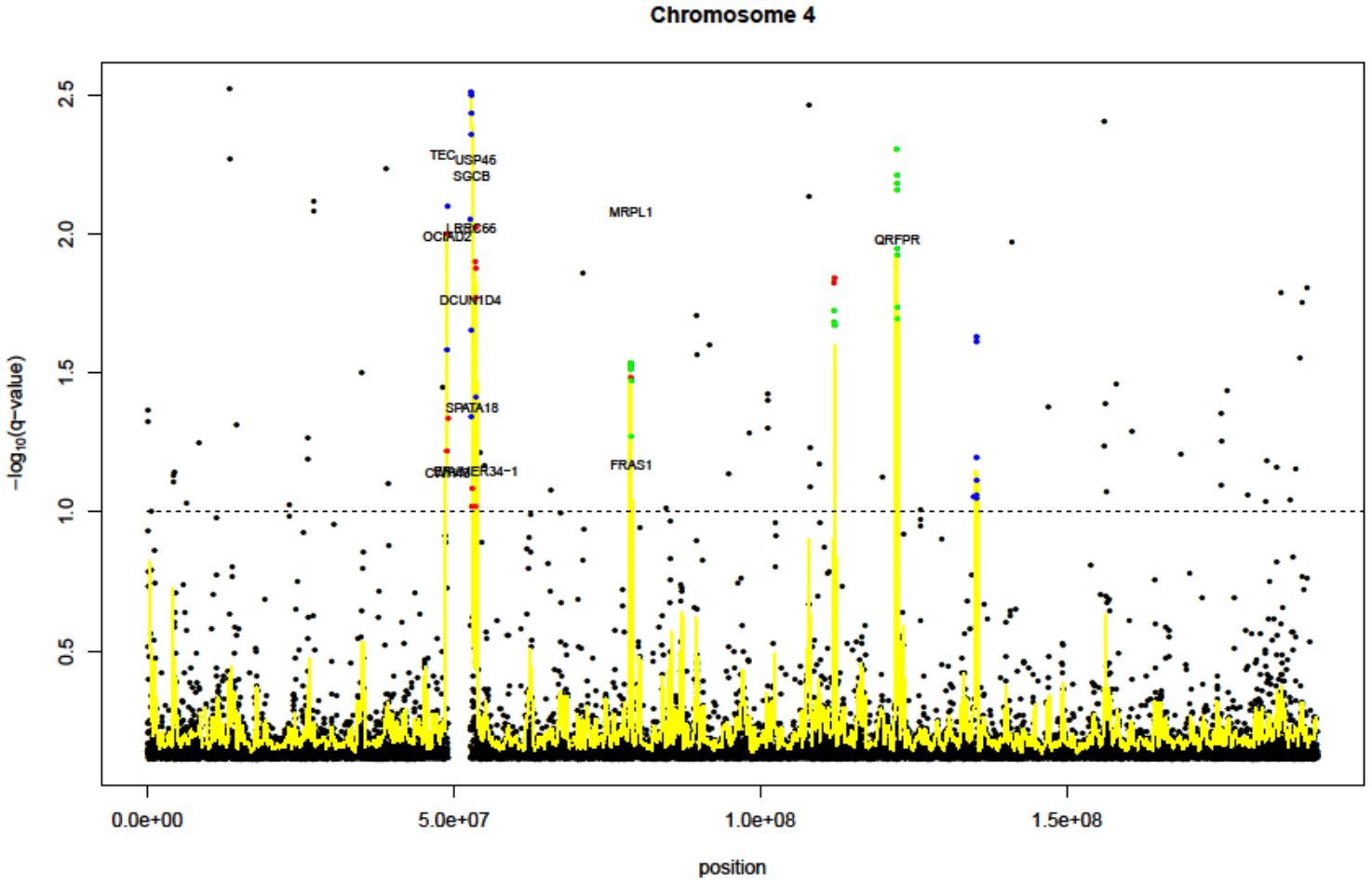
Chromosome 4
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
TEC
USP45
SGCB
MRPL1
QRFPR
DCUN1D4
SPATA18
FRAS1
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 4C

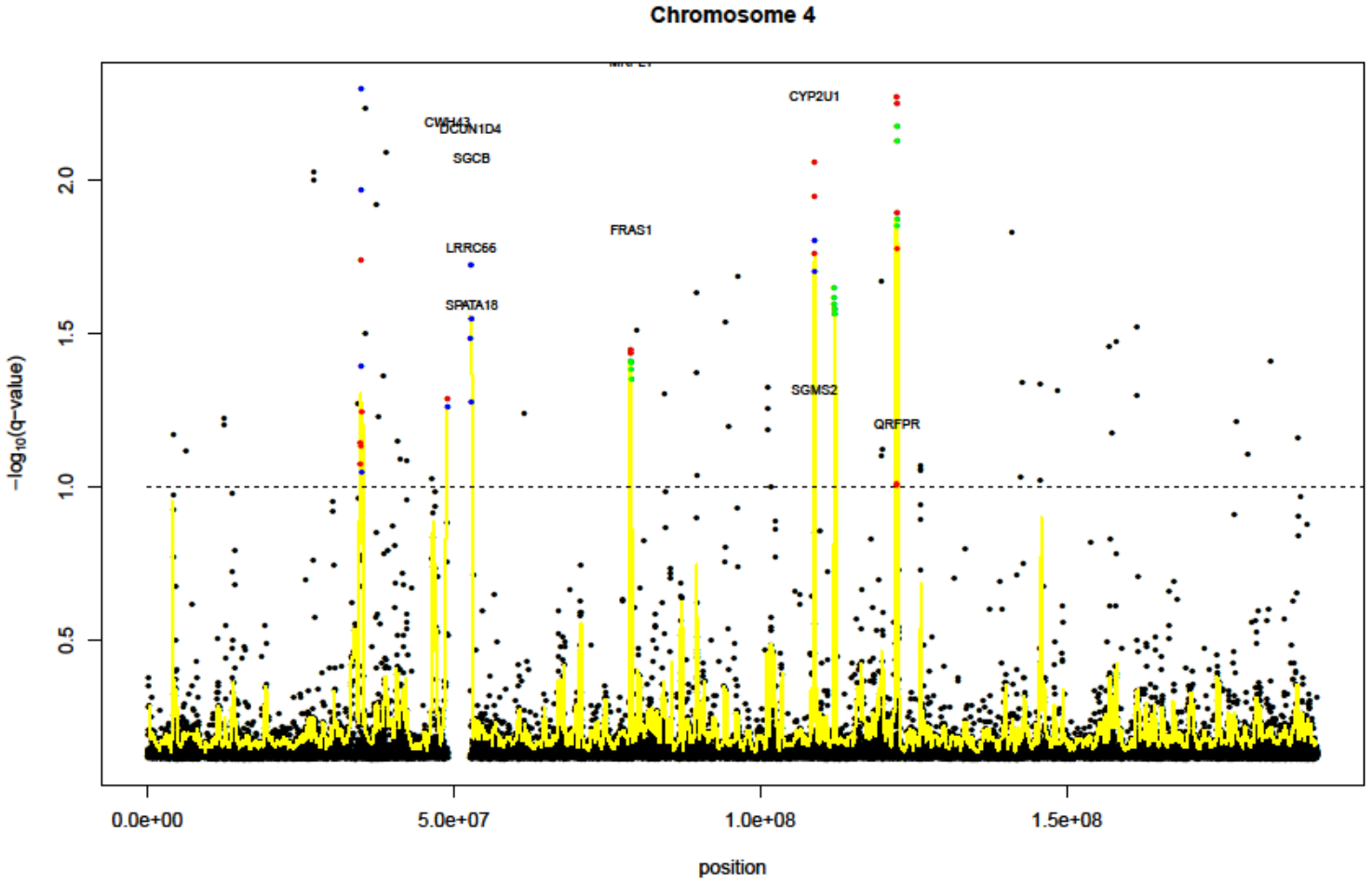
Chromosome 4
CWH43
DCUN1D4
SGCB
LRRC56
SPATA18
FRAS1
CYP2U1
SGMS2
QRFPR
−log10(q−value)
0.5
1.0
1.5
2.0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 4D

Figure 5A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 5. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 5A; PS2: 5B; PS3: 5C; PS4: 5D).

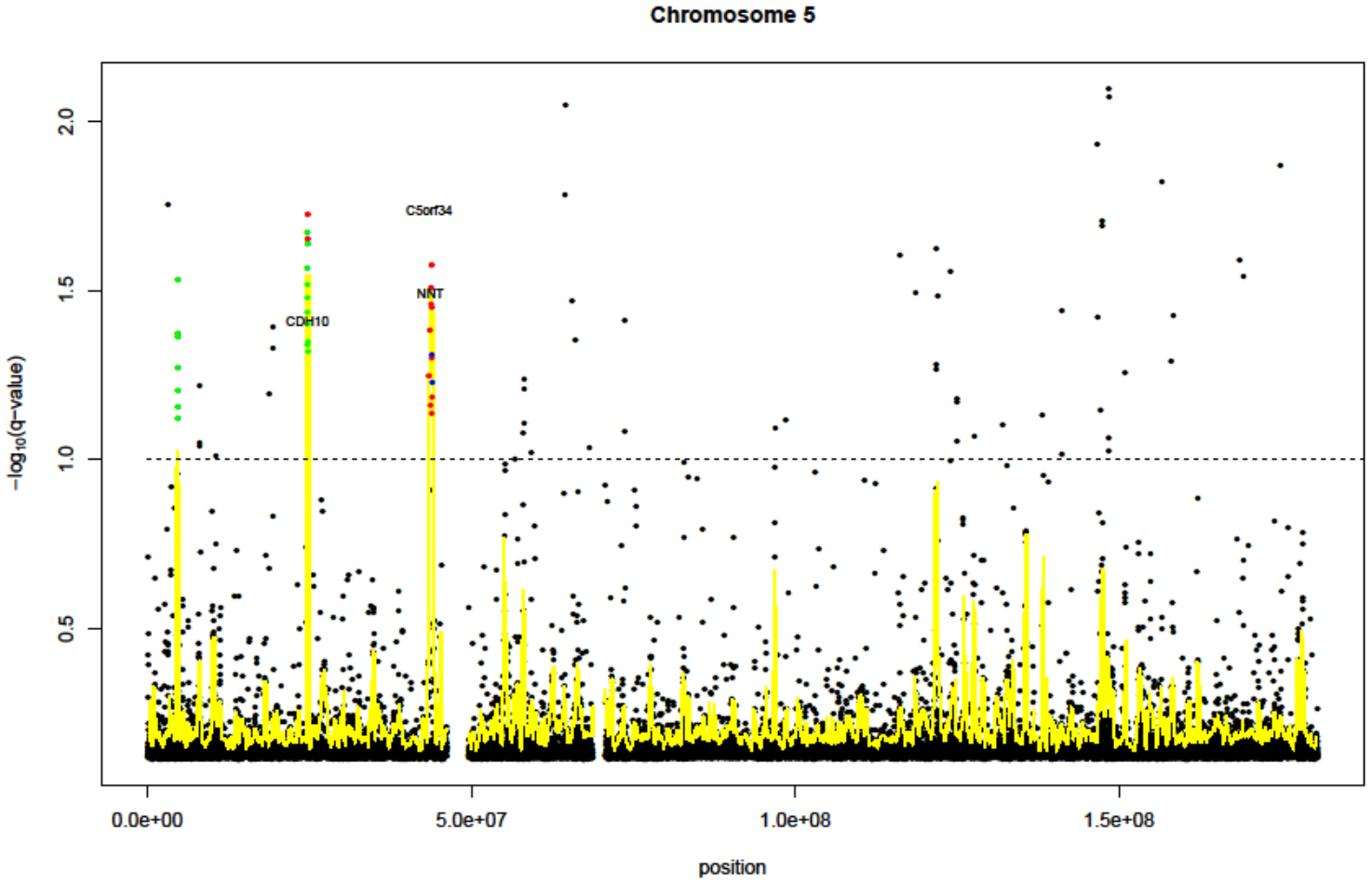
Chromosome 5
C5orf34
NNT
CDH10
−log10(q−value)
2.0
1.5
1.0
0.5
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 5A

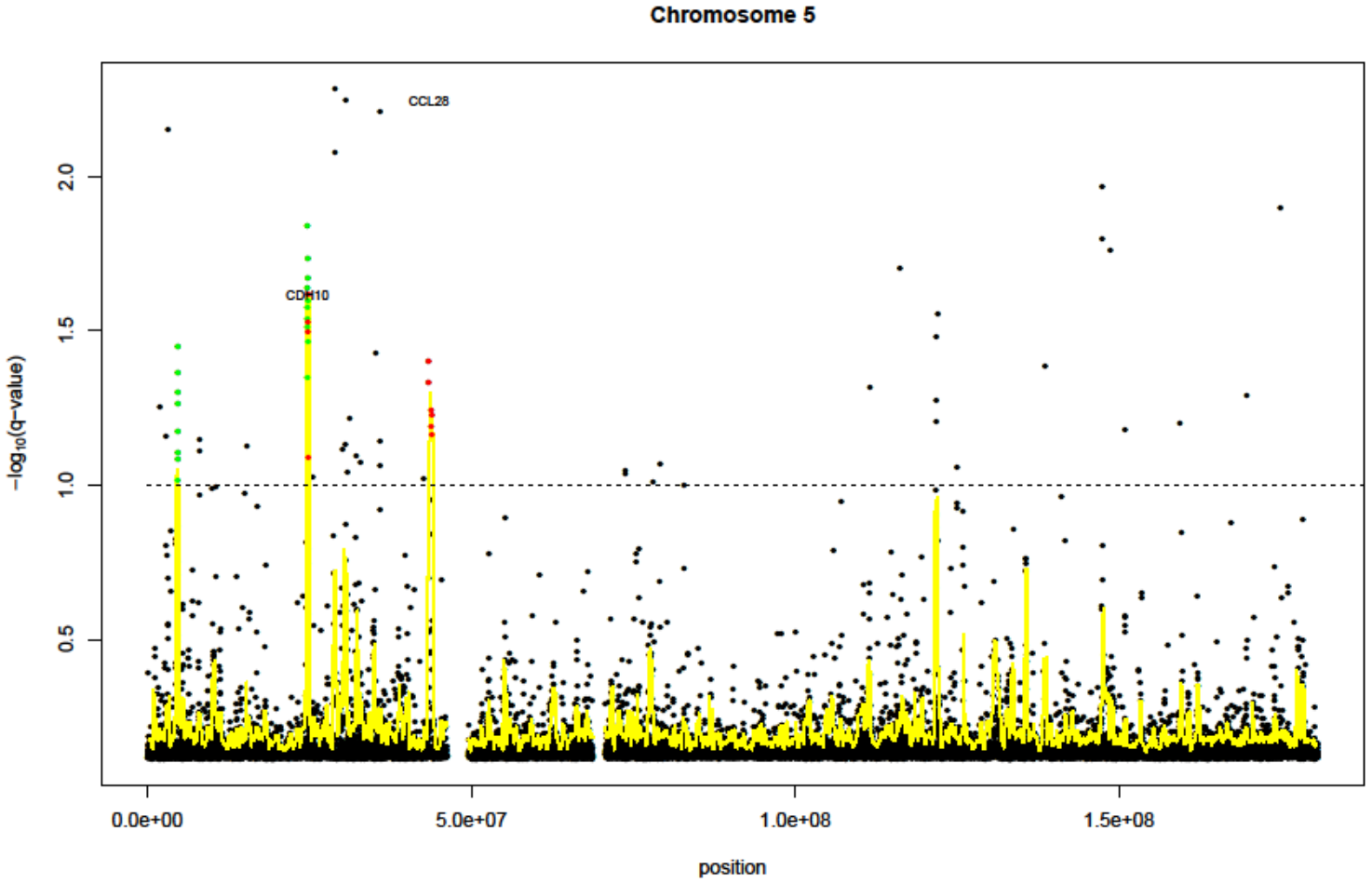
Chromosome 5
CCL28
CDH10
−log10(q−value)
2.0
1.5
1.0
0.5
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 5B

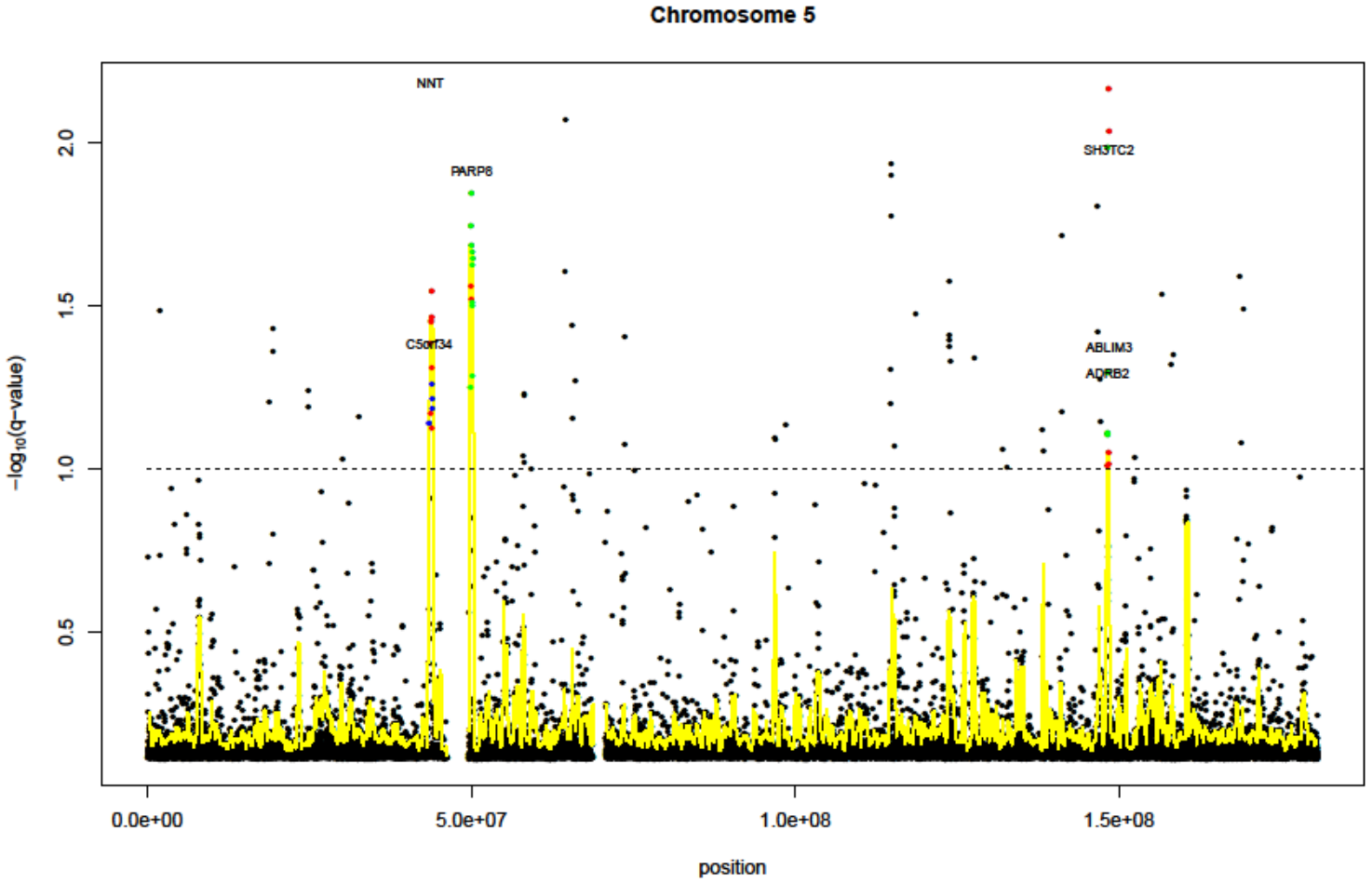
Chromosome 5
NNT
PARP6
C5orf34
SH3TC2
ABLIM3
ADRB2
−log10(q−value)
2.0
1.5
1.0
0.5
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 5C

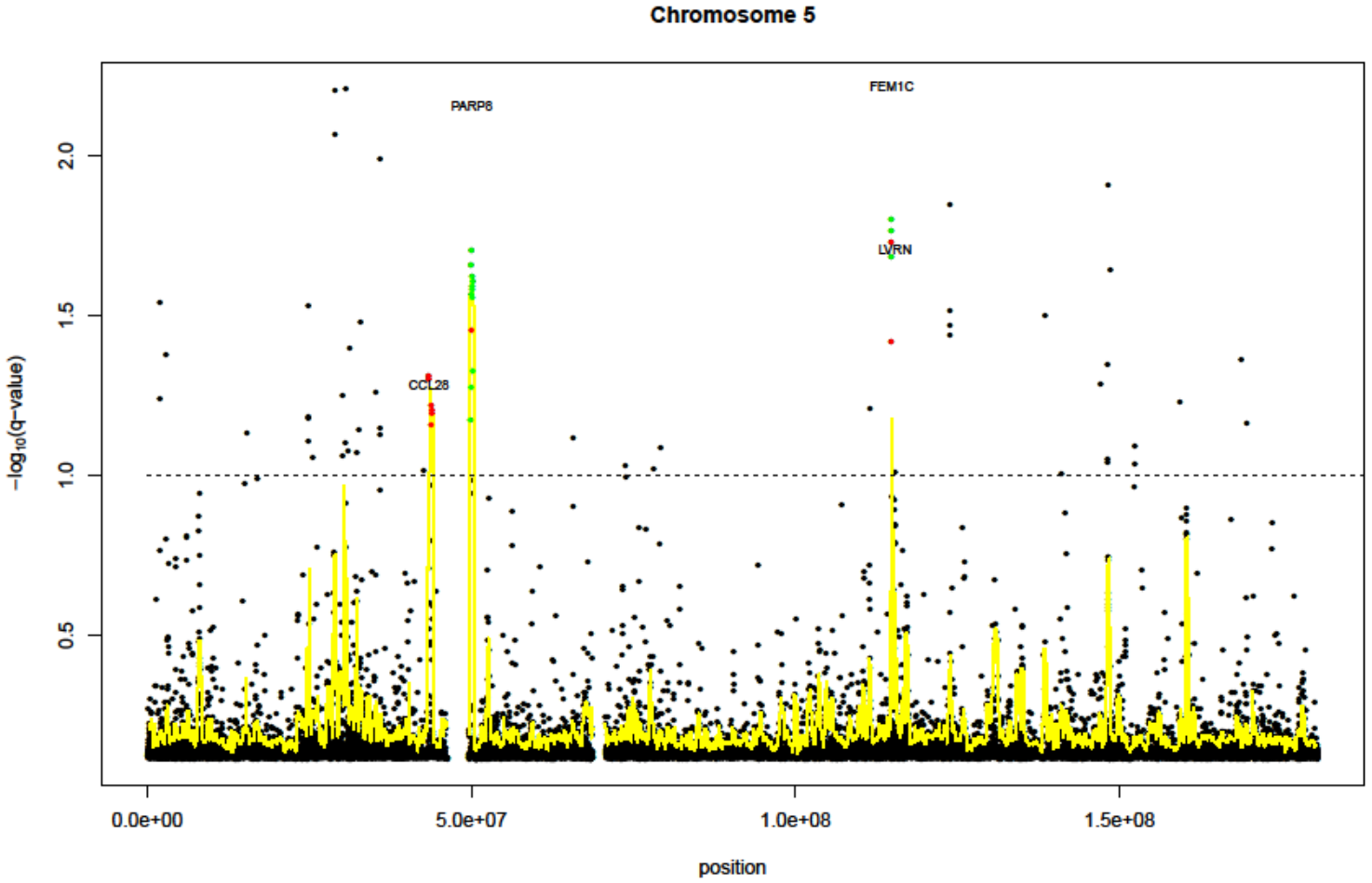
Chromosome 5
PARP6
FEM1C
LVRN
CCL28
2.0
1.5
1.0
0.5
−log10(q−value)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 5D

Figure 6A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 6. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 6A; PS2: 6B; PS3: 6C; PS4: 6D).

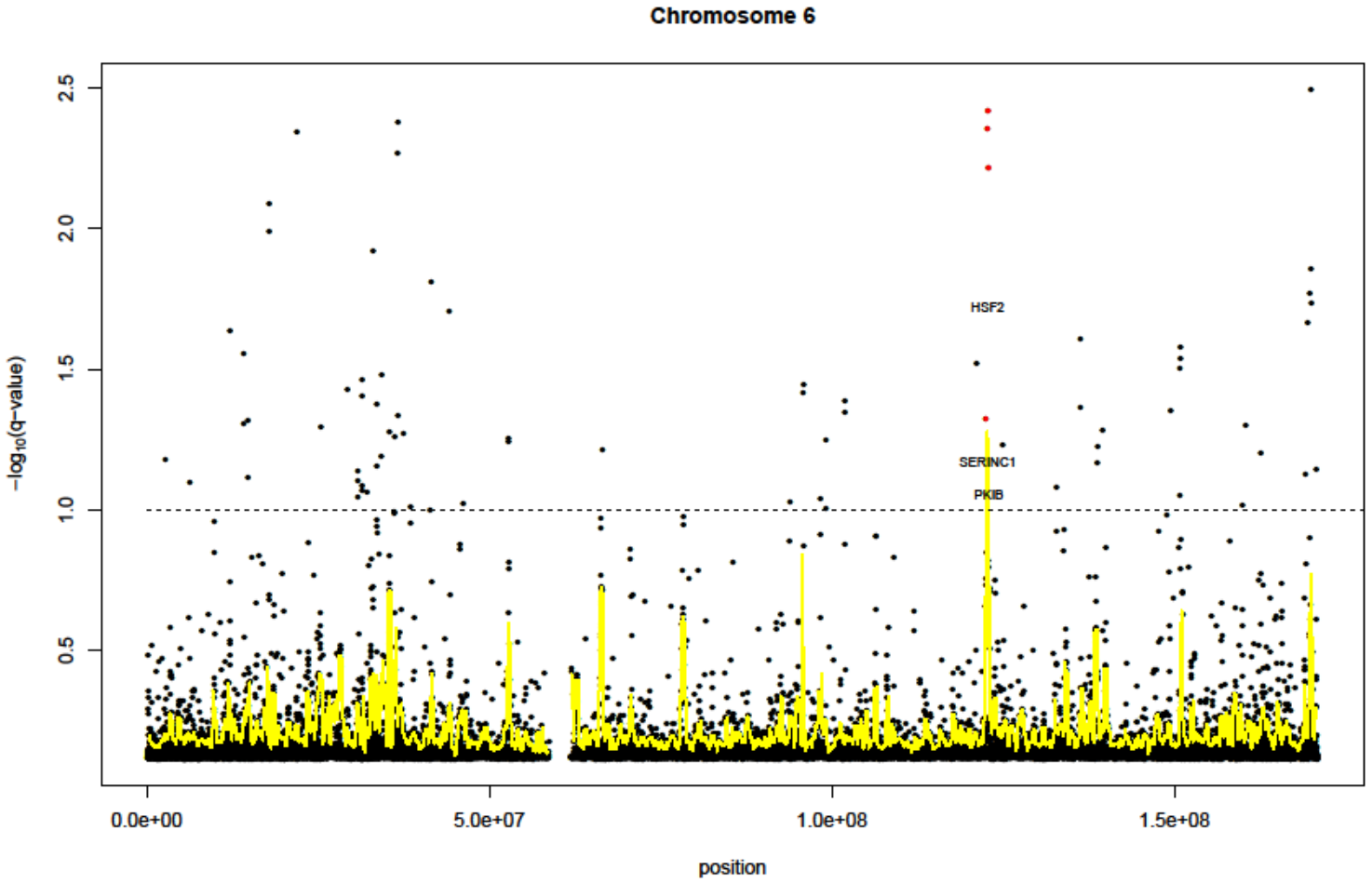
Chromosome 6
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
HSF2
SERINC1
PKIB
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 6A

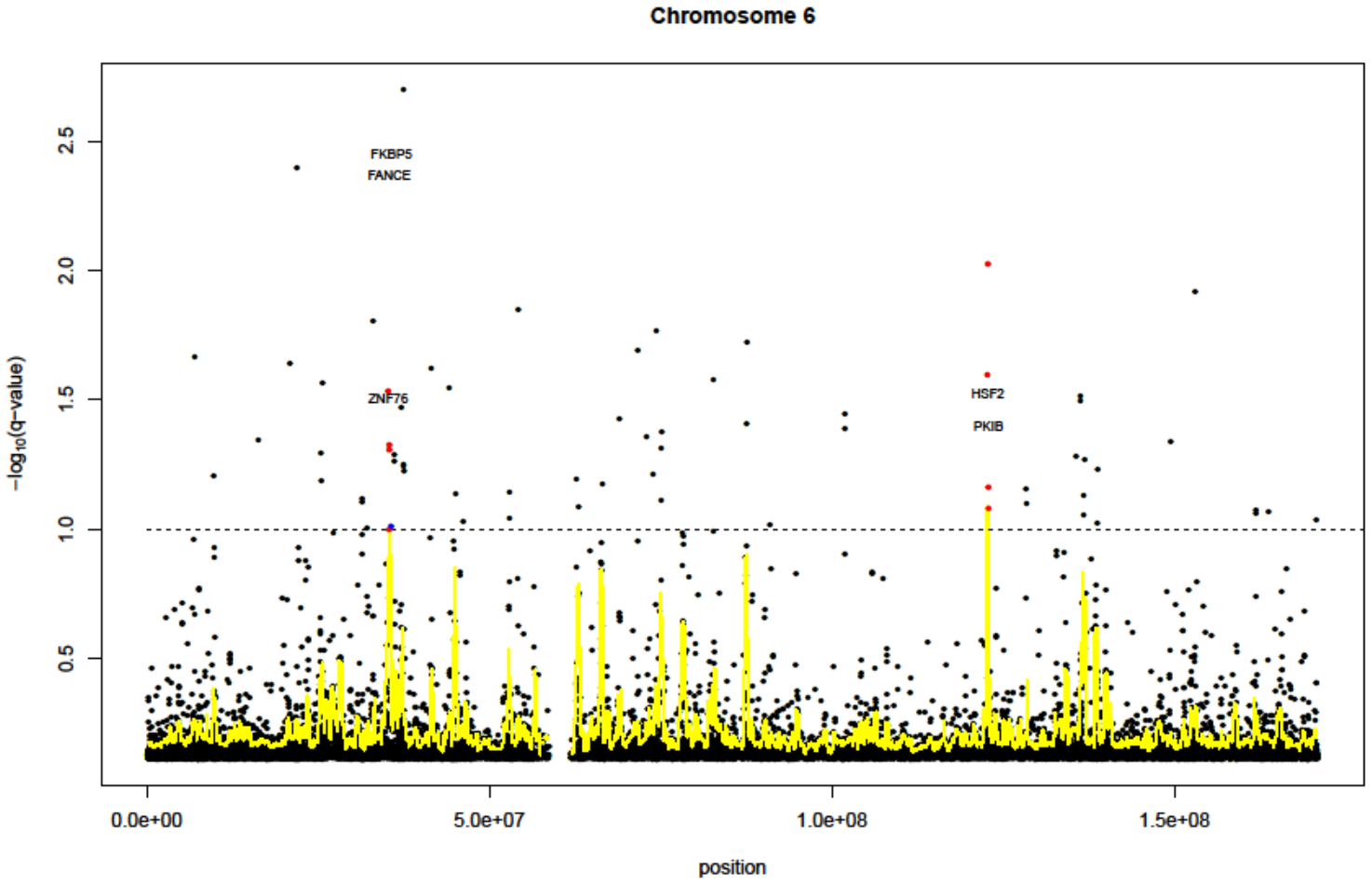
Chromosome 6
FKBP5
FANCE
ZNF76
HSF2
PKIB
2.5
2.0
1.5
1.0
0.5
−log10(q−value)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 6B

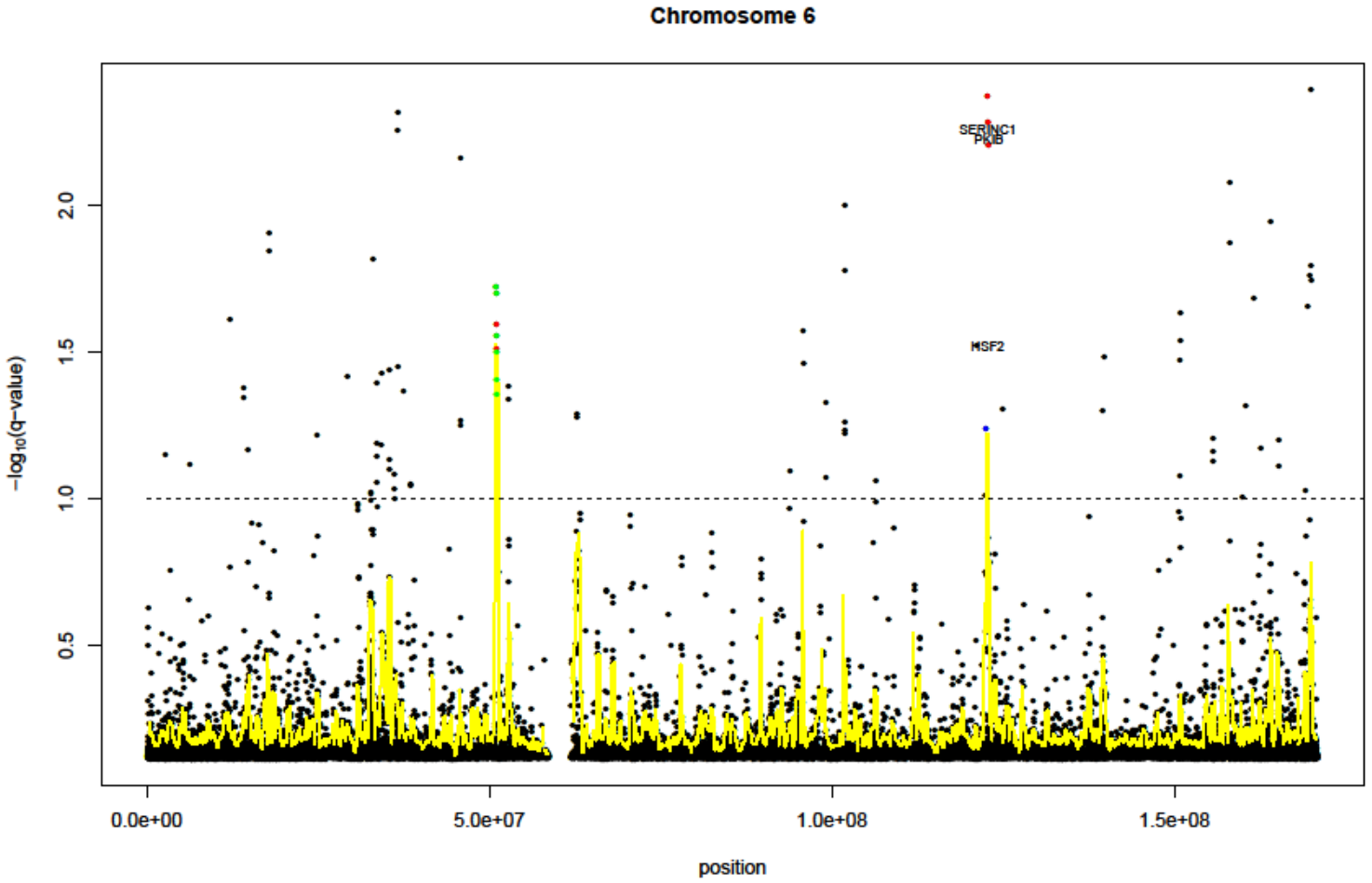
Chromosome 6
SERINC1
PKIB
HSF2
2.0
1.5
1.0
0.5
−log10(q−value)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 6C

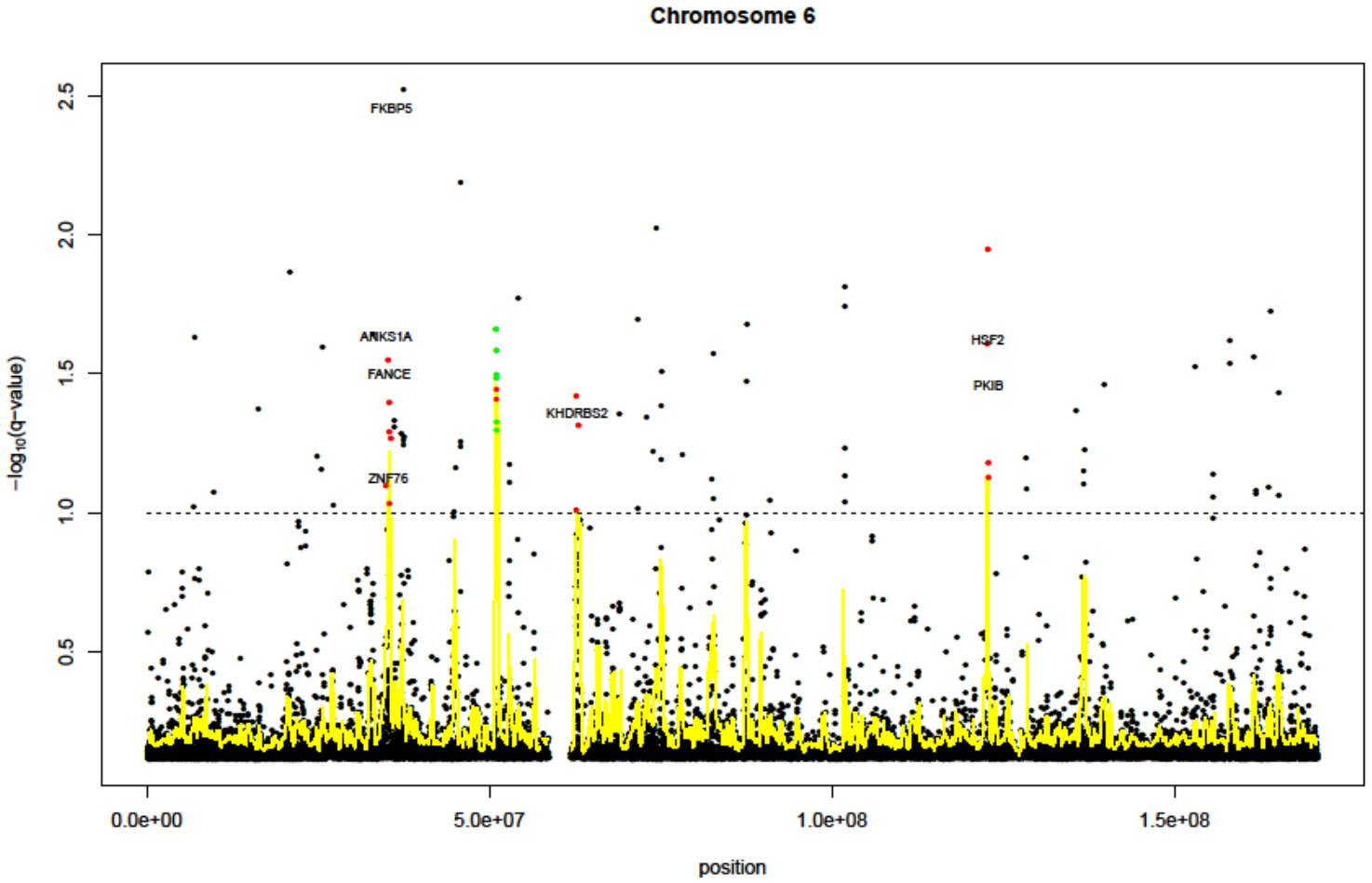
Chromosome 6
FKBP5
ANKS1A
FANCE
ZNF76
KHDRBS2
HSF2
PKIB
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 6D

Figure 7A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 7. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 7A; PS2: 7B; PS3: 7C; PS4: 7D).

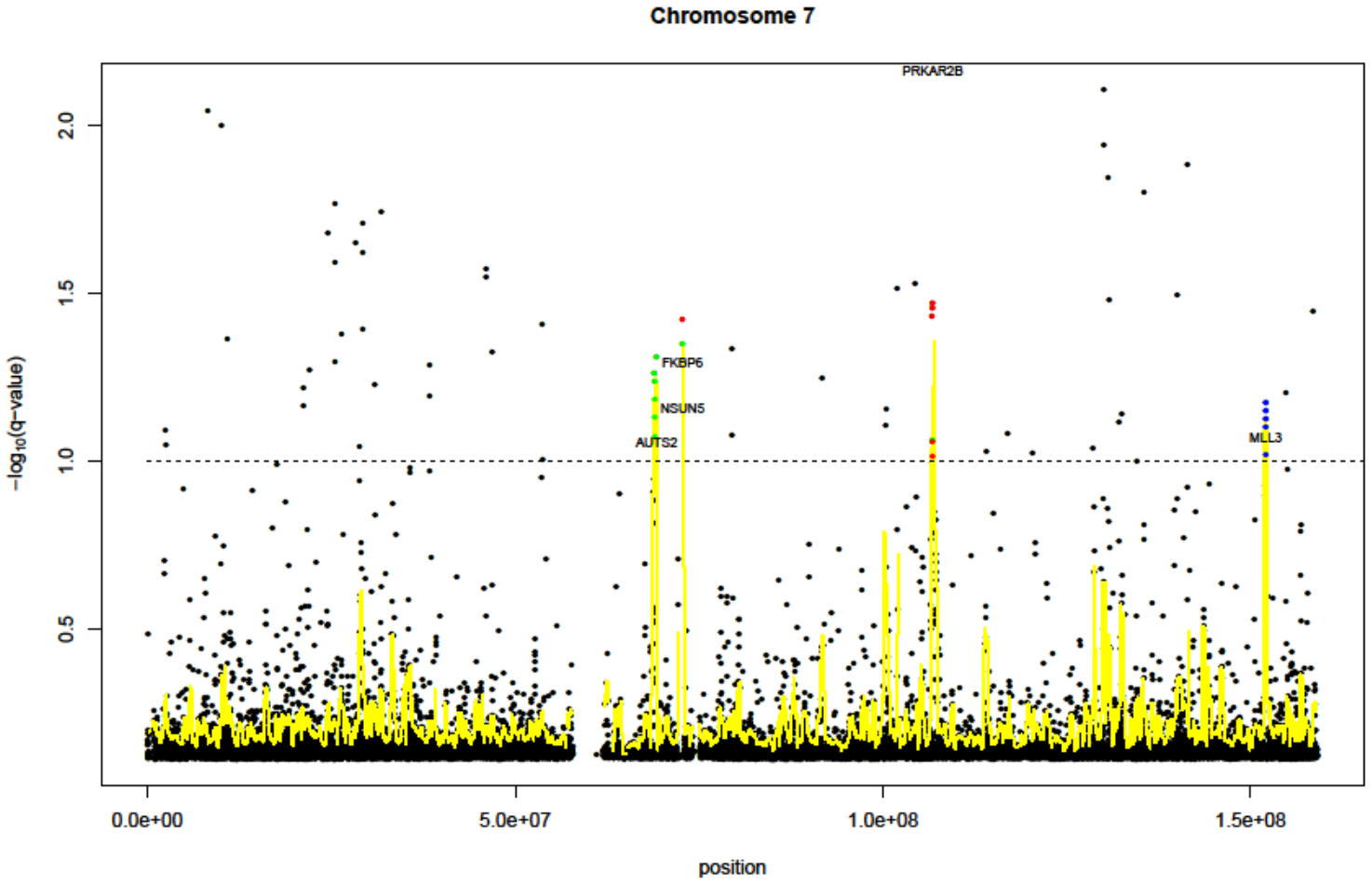
Chromosome 7
PRKAR2B
FKBP6
NSUN5
AUTS2
MLL3
2.0
1.5
1.0
0.5
−log10(q−value)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 7A

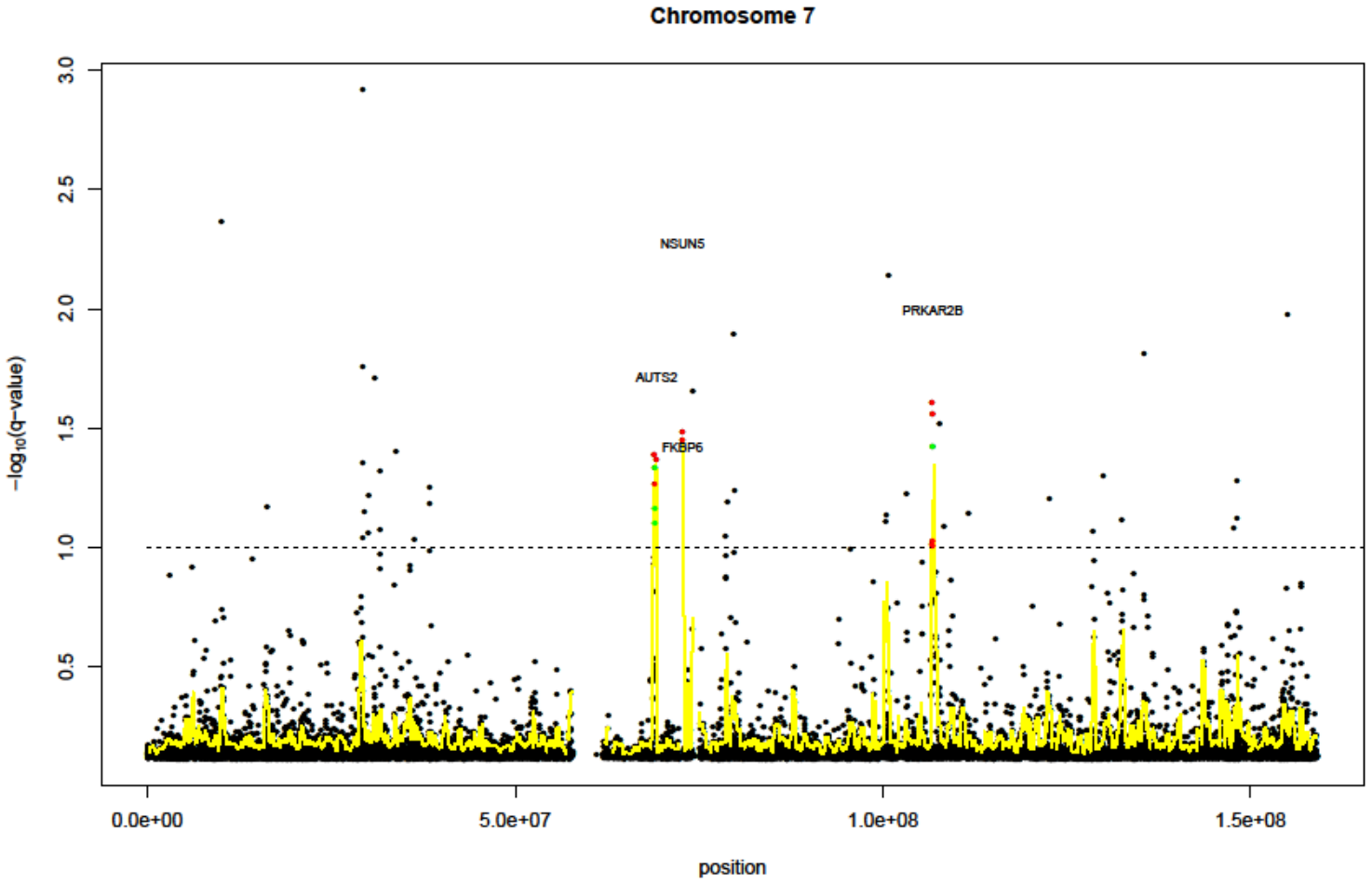
Chromosome 7
NSUN5
PRKAR2B
AUTS2
FKBP6
−log10(q−value)
3.0
2.5
2.0
1.5
1.0
0.5
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 7B

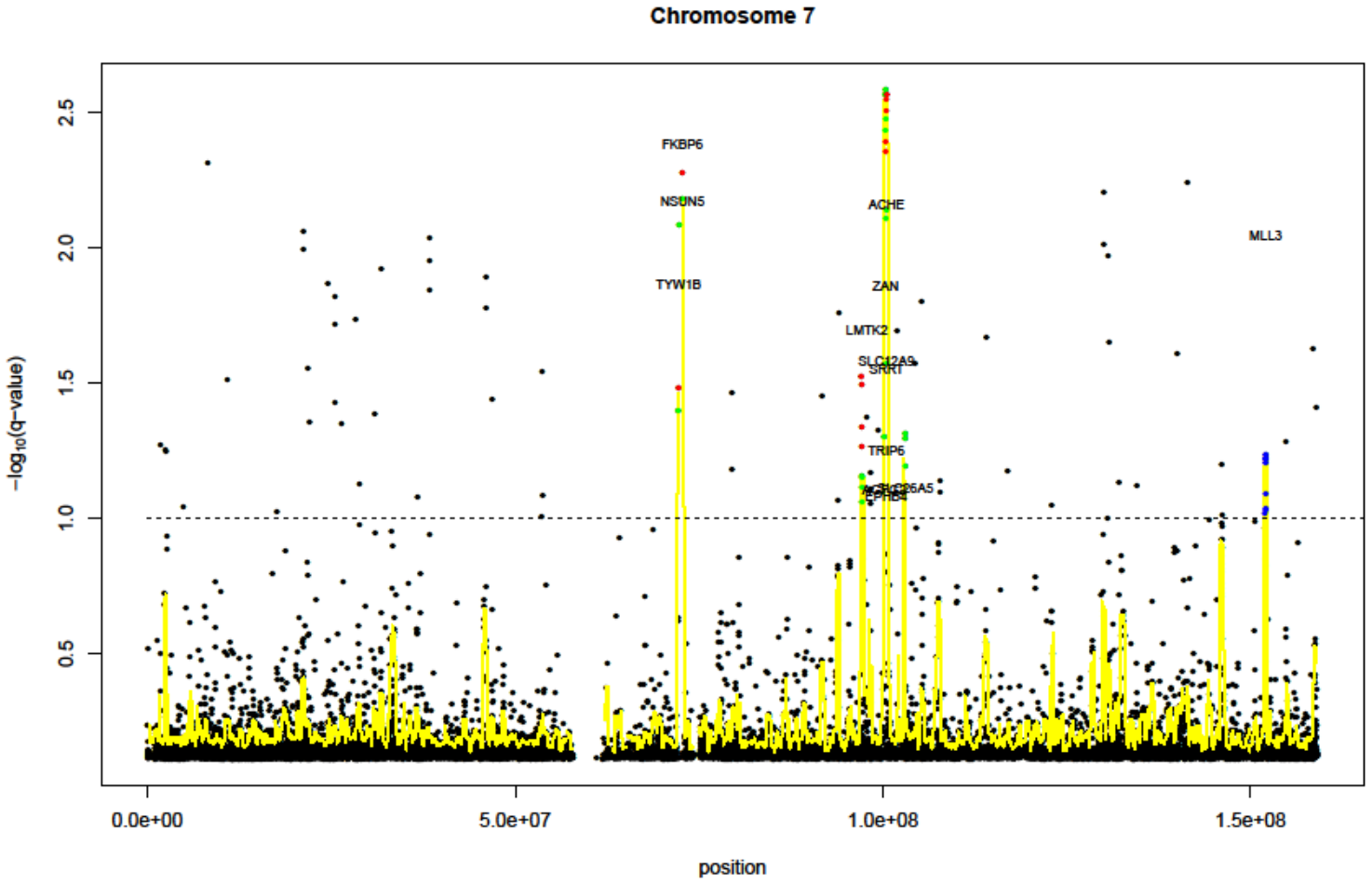
Chromosome 7
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
FKBP6
NSUN5
TYW1B
ACHE
ZAN
LMTK2
SLC12A9
SRRT
TRIP6
MLL3
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 7C

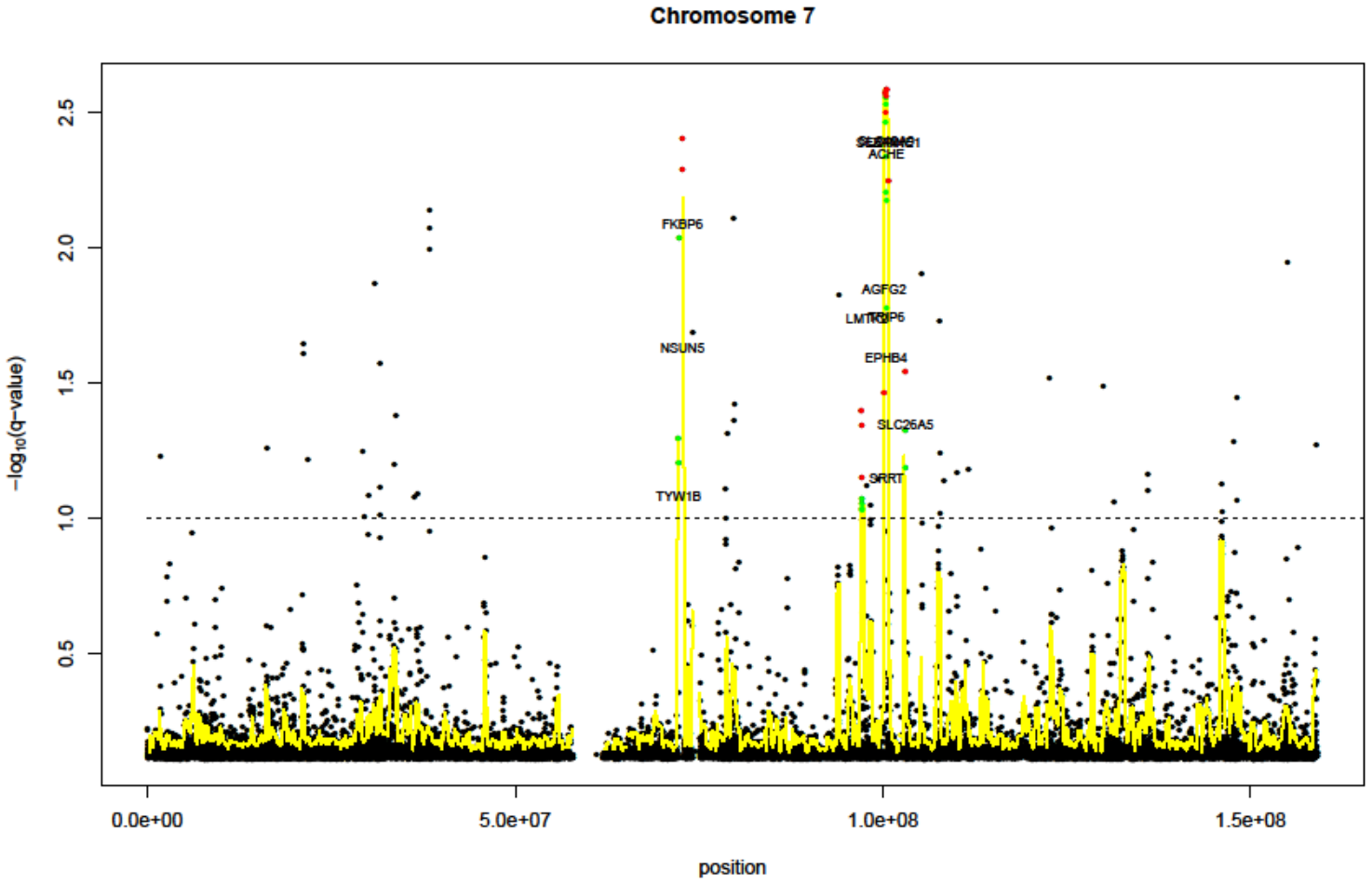
Chromosome 7
−log10(q−value)
2.5
2.0
1.5
1.0
0.5
FKBP6
NSUN5
TYW1B
AGHE
AGFG2
EPHB4
SLC26A5
SRRT
0.0e+00
5.0e+07
1.0e+08
1.5e+08
position

Figure 7D

II.III) ARTIGO 3

Amorim CEG, Bisso-Machado R, Ramallo V, Bortolini MC, Bonatto SL, Salzano FM, Hünemeier T (2013) *A Bayesian Approach to Genome/Linguistic Relationships in Native South Americans*. *PLoS ONE* 8: e64099.

OPEN ACCESS Freely available online

PLOS ONE

A Bayesian Approach to Genome/Linguistic Relationships in Native South Americans

Carlos Eduardo Guerra Amorim[1], Rafael Bisso-Machado[1], Virginia Ramallo[1], Maria Cátira Bortolini[1], Sandro Luis Bonatto[2], Francisco Mauro Salzano[1]*, Tábita Hünemeier[1]

1 Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil, **2** Faculdade de Biociências, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Abstract

The relationship between the evolution of genes and languages has been studied for over three decades. These studies rely on the assumption that languages, as many other cultural traits, evolve in a gene-like manner, accumulating heritable diversity through time and being subjected to evolutionary mechanisms of change. In the present work we used genetic data to evaluate South American linguistic classifications. We compared discordant models of language classifications to the current Native American genome-wide variation using realistic demographic models analyzed under an Approximate Bayesian Computation (ABC) framework. Data on 381 STRs spread along the autosomes were gathered from the literature for populations representing the five main South Amerindian linguistic groups: Andean, Arawakan, Chibchan-Paezan, Macro-Jê, and Tupí. The results indicated a higher posterior probability for the classification proposed by J.H. Greenberg in 1987, although L. Campbell's 1997 classification cannot be ruled out. Based on Greenberg's classification, it was possible to date the time of Tupí-Arawakan divergence (2.8 kya), and the time of emergence of the structure between present day major language groups in South America (3.1 kya).

Citation: Amorim CEG, Bisso-Machado R, Ramallo V, Bortolini MC, Bonatto SL, et al. (2013) A Bayesian Approach to Genome/Linguistic Relationships in Native South Americans. PLoS ONE 8(5): e64099. doi:10.1371/journal.pone.0064099

Editor: Keith A. Crandall, George Washington University, United States of America

Received December 21, 2012; **Accepted** April 9, 2013; **Published** May 16, 2013

Funding: This study was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS, PRONEX), Brazil. These funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: francisco.salzano@ufrgs.br

Introduction

The patterns of genetic and linguistic variation have been compared for over three decades. These studies rely on the hypothesis that languages, as many other cultural traits, evolve in a gene-like manner, accumulating diversity through time and being subjected to evolutionary mechanisms of change [1,2]. However, it should be mentioned that language, as a culturally mediated trait, is also transmitted horizontally (between unrelated individuals) in a Lamarckian way. This fact may lead to its undergoing a faster mutation rate and being subject to additional evolutionary forces [1,3–5]. Thus, linguistic and genetic evolution may or may not agree [1,6–13].

Studies involving Native American language and gene parallel evolutions are scarce ([3,8,9,12,14,15] and references therein), but have brought relevant contributions to our understanding of the peopling of the Americas. However, some important parameters, such as population size differences, demographic fluctuations, or gene flow among demes, were not considered [8,12,15,16].

In the present work, we revisited the problem considered by Salzano et al. [3] –*i.e.* use of genetic data to evaluate different native language classifications in South America – comparing discordant models with the current patterns of genetic variation. We propose realistic evolutionary models based on the Coalescent [17] and developed under a robust statistical framework, the Approximate Bayesian Computation (ABC; [18,19]). Differently from earlier studies, this approach considers variances in population effective size through time, among demes, and gene flow; dates fission events, and can handle a large set of genetic markers (in the present case, 381 microsatellite loci).

In this analysis, we addressed three main questions: (a) Which language classification better fits the current South American genome-wide diversity? (b) How old are the interpopulation branch connections? and (c) Do the divergence dates between language groups, as estimated by genetic and linguistic data, agree?

Subjects and Methods

Linguistic classifications

From the six classifications that cover South Native American languages: Loukotka [20], Rodrigues [21], Greenberg [22], Campbell [23], Urban [24], and Lewis [25]; only three could be used here, since Rodrigues' and Urban's classification are restricted to certain groups and Lewis' to recent branches (which are identical among these classifications). Five major South American linguistic groups were considered: Andean, Arawakan, Chibchan-Paezan, Macro-Jê, and Tupí.

Loukotka [20], Greenberg [22], and Campbell [23] recognize roughly the same large language groups:

1) Andean: distributed along the Andean Cordillera (mainly Chile, Peru, and Bolivia). Examples: Aymara and Quechua;

2) Arawakan: distributed along most of the equatorial latitude. Includes the Piapoco and Wayuu;
3) Chibchan-Paezan: occupying the extreme northwestern territories of the subcontinent. Examples: Arhuaco, Kogi, and Waunana;
4) Macro-Jê: found in Central and Eastern Brazil (example, Kaingang); and
5) Tupí: distributed from the Amazon Forest southwards. Guarani is its most southern group.

Despite this agreement, each of these linguists employed different methods to classify the relationships between these groups. Greenberg [22] used multilateral comparisons, examining many languages simultaneously to detect similarities in a small number of basic words and grammatical elements. Campbell [23] used a more orthodox analysis: the comparative method, considering that proposals of remote linguistic relationships are only plausible when a series of other possible explanations have been eliminated. And finally, Loukotka [20] made use of two different methods in his classification: the lexicostatistical in some and the comparative in other cases.

May be due to these different methodologies, there are differences between the three language classifications. Campbell [23], recognizes similarities between the Andean and Maipurean (Arawakan in the above-mentioned classification), grouping them in a stock named Quechumaran. He also noticed resemblances between the Tupí and Macro-Jê languages, while also proposing a third group, which would be that composed by the Chibchan-Paezan languages. The deeper relationship between these three groups is not resolved.

Greenberg [22] clustered the Tupí together with the Arawakan in a group called Equatorial-Tucanoan. He did not clarify the relationship between this group and the remaining three, but assembled those in a large group called Amerindian, including all the native languages spoken in South and Central America, and a few from North America.

Loukotka's [20] classification agrees with Greenberg's [22] in relation to the close relationship between the Tupí and Arawakan. However, Loukotka groups the Chibchan-Paezan with the Andean languages. The relationship of these two groups and their connections with the Macro-Jê are not detailed. Table S1 (Supporting Information) provides a more detailed classification of the languages belonging to each of these groups according to these and additional authors.

In 2007 a close collaborator of Greenberg, Merritt Ruhlen, published a posthumous revision of his Amerindian linguistic family classification [26]. This work considered all the previous criticisms from other scholars and also new studies, making this new classification somewhat closer to Loukotka's proposition. Given this proximity, the present work will not make use of this more recent study, although it can be seen in comparison to the others in Table S1.

Genetic markers

Starting from the 678 autosomal microsatellite loci (STRs) reported in [10], 297 were removed from the analyses due to a high (>5%) percentage of missing data for at least one of the populations studied here. The remaining 381 STRs were formatted for the genetic analyses software employed here by using the PGDSpider [27] and in-house written scripts (STR IDs are listed in Table S2).

Populations and samples

From an initial set of 30 populations studied in [10], five were selected to represent the above-mentioned major linguistic groups as follows: Aymara (2n = 18; Andean), Piapoco (2n = 13; Arawakan), Kogi (2n = 17; Chibchan-Paezan), Kaingang (2n = 7; Jean), Guarani (2n = 10; Tupí). See Table S1 for a detailed classification of these languages and [10] for alternative language names and geographic coordinates of each population.

The selection of a single population to represent a whole linguistic group was based on two assumptions. First, the discrepancies between the three linguistic classifications were observed only at deep branches (involving the final relationship among the five language groups); and second, this procedure reduces the number of parameters of the complex demographic models used here, what is important for both statistical and computational reasons [19].

Ethical approval for the original study from which the STR information was obtained was given in Brazil (Kaingang, Guarani) by the Brazilian National Ethics Commission (CONEP Resolution no. 123/98); in Colombia (Piapoco, Kogi) by the Ethics Commission of Universidad de Antioquia, Medellin, Colombia; and in Chile (Aymara) by the Ethic Commission of Universidad de Chile, Santiago, Chile. Individual and tribal informed oral consent was obtained from all participants, since they were illiterate, and they were obtained according to the Helsinki Declaration. The ethics committees approved the oral consent procedure, as well as the use of these samples in population and evolutionary studies.

Overview of demographic and genetic modeling

Three demographic scenarios (Figure 1) were modeled with Fastsimcoal 1.1.2 [28], which is a simulator of genetic diversity based on the Coalescent [17]. All scenarios presented the same configuration between times T_0 and T_1: a small ancestral population of effective size N_0 (at T_0) undergoes exponential growth until it reaches effective size N_1 (at T_1), time in which the ancestral population undergoes subdivision for the first time as depicted in Figure 1. Further structure arises at T_2 separating populations that diverged more recently. For each pair of populations in such fission events, an independent T_2 value was sampled from the prior distribution in each simulation, with a restriction, no sampled value for the date of a more recent fission event (T_2) could represent older dates than T_1. Symmetric gene flow was allowed to happen among any pair of populations at a rate of m, that is the probability of a gene in the source population to be sent to the sink population. As for T_2, m may also assume different values for each pair of populations. Current average deme size was represented by N_P, which was assumed to be Gamma (10, $10/N_P$) distributed. The populations were thus allowed to have different sizes and different susceptibility to genetic drift. Time was measured in years, with a generation time of 25 years. Effective population sizes are given in number of diploid individuals. Prior distributions (based on results from recent Native American evolutionary studies) for the main model parameters are given in Table 1.

Under a strict stepwise mutation model (SMM), the average STR mutation rate (υ) was set to 6.4×10^{-4} per generation [32]. Since the observed variance between different loci may affect population genetic statistics, and to take this point into consideration, mutation rates were allowed to vary according to the Gamma distribution ($\alpha,\alpha/\upsilon$; where α is a hyperparameter drawn from an uniform 1–20 distribution). Thus υ was allowed to vary in each simulation and among loci by several orders of magnitude, depending on sampled α values.

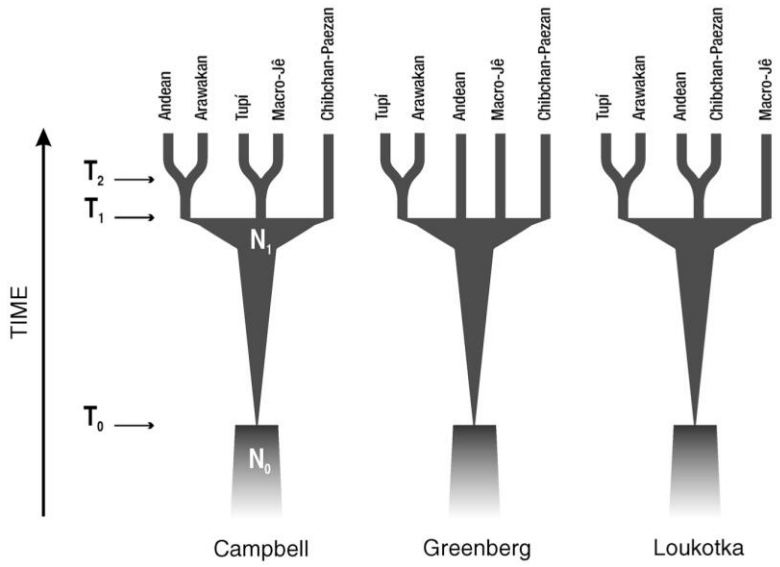

Figure 1. Alternative demographic models tested against the genetic variation in 381 autosomal STRs. Parameters are explained in Table 1. Current average deme size (N_P) and gene flow (m) between populations are not shown.
doi:10.1371/journal.pone.0064099.g001

Model choice

The first approach to compare the scenarios was to see if they could generate simulated populations that closely matched the observed data in relation to the distribution of the genetic diversity observed in the 381 loci sampled. The posterior probability of each modeled scenario was then calculated under the ABC framework [18,19] using the ABCtoolbox [33]. Briefly, for each scenario, 100,000 simulations were generated with Fastsimcoal using the empirical sampling configuration and the previously described models. For each simulation a certain value for each model parameter was sampled from the prior distribution (Table 1) using Fastsimcoal for simulating genetic diversity. Pairwise and global R_{ST}, a F_{ST} analogue for STR data which takes into account the difference between STR allelic sizes, were then calculated for each simulated sample and for the empirical dataset with the Arlequin 3.5.1.2 command line version [34] yielding a total 11 summary-statistics. This procedure was conducted with the ABCsampler software implemented with the ABCtoolbox.

The reference tables containing the model parameters used to generate the 100,000 simulations under each scenario and corresponding summary-statistics were then compared to the empirical dataset with the ABCestimator software, also implemented with the ABCtoolbox. This software compares the vectors defined by the summary-statistics estimated for each simulated data set (S) with that estimated for the empirical data (S*) by calculating Euclidian distances $\delta = ||S\text{-}S^*||$ between them. Half a percent (0.5%) of the simulations matching closest the empirical data were retained for the estimation of the marginal densities of each model. These are then used for the assessment of the posterior odds (Bayes factors; [35]) for each model given the observed data.

To check for potential biases in model choice, 100 additional simulations were generated under each scenario and used as pseudo-empirical data. The same procedure was performed for the empirical data for each of these 300 simulations and the rate of false model inference could then be calculated.

Table 1. Prior distributions of selected model parameters.

Parameter[1]	Distribution	Range	References
T_0 – Time for the onset of expansion	Uniform	10,000–19,000	[29,30]
T_1 –Time for the first emergence of structure	Uniform	800–6,400	[23,31]
T_2 –Time for the second emergence of structure	Uniform	800–6,400	[23,31]
N_0 – Ancestral effective population size	Uniform	2–1,000	[29]
N_1 – Effective population (continental) size	Uniform	1,000–100,000	[29]
N_P – Current effective deme size	Gamma (10, 10/N_P)	50–1,000	[29]
m- symmetric migration rate	Uniform	0.00001–0.001.	[29]

[1]Time is given in years before present and effective population size in number of diploid individuals (2n). T_1 and T_2 prior distributions may present deviations from uniformity, since $T_1 > T_2$.
doi:10.1371/journal.pone.0064099.t001

An additional methodology for inferring model posterior probabilities is that proposed by Pritchard et al. [36], which could be described as follows: From the initial 100,000 simulations conducted according to each model, the 100 with smallest associated Euclidian distances to the empirical dataset were retained. This set of 300 simulations was then ranked by ascending Euclidian distances and the posterior probability of a given model was then computed as the proportion of simulations performed under this model included among the 100 first simulations.

Model parameter estimates

The posterior distributions of the selected parameters (T_0, T_1, T_2, N_0, N_1, and N_P) of the model with higher posterior odds were inferred according to the same framework used for model choice, but with a new reference table with 500,000 simulations. The ABCestimator [33] computes point estimates (mode and median) and confidence intervals (highest posterior density interval) for these distributions. It also checks for potential bias using, in our case, 1,000 pseudo-empirical data, generating a quantiles distribution of the known parameter values in relation to the inferred posterior confidence interval [33], which is then examined statistically for its uniformity according to a Kolmogorov-Smirnov test with $\alpha = 0.05$ using R [37]. Visual histogram examination was also performed. R was also used to calculate the parameter regression against the summary-statistics, which indicates the proportion of the parameter variance explained by it [38].

Results

The empirical distribution for the 11 summary-statistics – namely pairwise and global R_{ST}s – estimated using the genetic variation of the 381 STRs in the above-mentioned Amerindian populations could be reproduced in the bulk of simulations generated, with no particular better performance for any model. The inference is that all modeled scenarios were able to capture the reality of the STR genome-wide diversity.

Table 2 describes the posterior odds of each scenario according to the two adopted methods to infer posterior probabilities [35,36]. Both indicate a higher posterior probability for Greenberg's model, followed by Campbell's. Loukotka's model presented virtually no correspondence with the tested genome-wide diversity.

To control for the quality of the model inference, we used the reference table containing the 300 simulations, each 100 generated under a specific model. The known correct model was properly inferred 86% of the times among all inferences performed with the pseudo-empirical data, a rate much higher than that expected by chance (~33%); the conclusion is that the model fitting procedure was strongly reliable.

Table 2. Posterior probability of three linguistic classifications for South American languages given the genetic diversity of 381 autosomal STRs.

Linguistic classification	Posterior probability (%)	
	Method I [35]	Method II [36]
Campbell [23]	40.3	43.0
Greenberg [22]	59.1	51.0
Loukotka [20]	00.6	6.0

doi:10.1371/journal.pone.0064099.t002

Figure 2 presents the prior (with all 500,000 runs), retained (0.5%) best simulations and posterior distributions for the selected parameters (T_0, T_1, T_2, N_0, N_1, and N_P) of the demographic model based on Greenberg's language classification. Their characteristics (point estimates and confidence intervals) are given in Table 3 together with the indicators of estimation accuracy. Root mean squared errors (Table 3) indicate that the median was more accurate than the mode in all measures.

Figure 3 shows the histograms of the posterior quantiles of the model parameters. T_1, T_2, and N_P present sharp distributions (Figure 2), ideal for ABC estimation. Most of the parameters also present uniform posterior quantiles distribution in the pseudo-empirical dataset (Figure 3) and corresponding Kolmogorov-Smirnov non-significant *p*-values (Table 3). T_2 and N_P also show high R^2 values (Table 3) suggesting their estimate may be very reliable. In spite of that R^2 for T_1 was low. To further test the reliability of the T_1 estimate, we evaluated the effect of including four additional summary-statistics in its estimation, namely mean and standard deviation of both heterozygosity (H) and number of alleles per locus (K). After this procedure, R^2 presented a higher value (0.16) and its posterior distribution gave a narrower high posterior density interval (HPDI = 2,835–5,571 years before present-YBP) mostly overlapping with the previous estimate (Table 3). To standardize the analyses performed for parameters' estimation, we will consider only the first estimate for T_1 and will use the second one just in this step for assuring quality.

The remaining parameter posterior point estimates (T_0, N_0, and N_1) are likely not reliable, since these parameters are poorly explained by the summary statistics ($R^2 < 10\%$) (see [38]). The posterior distributions of these parameters did not present clear peaks (Figure 2) and almost no difference from the prior distributions (Tables 1 and 3). However, they present no bias according to the posterior quantiles distribution (Figure 3), except for T_0, which showed a significant *p*-value for the Kolmogorov-Smirnov test (Table 3).

Discussion

Campbell's [23], Greenberg's [22], and Loukotka's [20] classifications present marked differences on the relationships of the five South American major linguistic groups. Studies have been conducted to assess which of these propositions presented better correlation with the population relationships suggested by the genetic data. Campbell's and Greenberg's had received genetic support previously ([39] and [40]; [7] and [12], respectively), while Loukotka's classification has not received any. Our results agree with these previous results, since Loukotka's is significantly rejected by the genetic variation observed in a large dataset of fast-evolving autosomal markers widespread along the human genome, while Greenberg's classification receives the greatest support although it is just slightly more adequate than Campbell's (Table 2). The difference between the Loukotkas and the Greenberg's models that may explain why the former is significantly worst fitted to the data is probably the grouping of Andean with Chibchan-Paezan languages.

Comparisons between linguistic and genetic models are very informative for the understanding of human evolution, and may contribute to the knowledge of language evolutionary dynamics; but it should be remembered that they start from quite different methodological assumptions [2]. The main Native American linguistic varieties are classified in well-established language families, but the connection among them to establish major lineages remain controversial. Greenberg's linguistic classification [22] and its multilateral or mass comparison approach have been

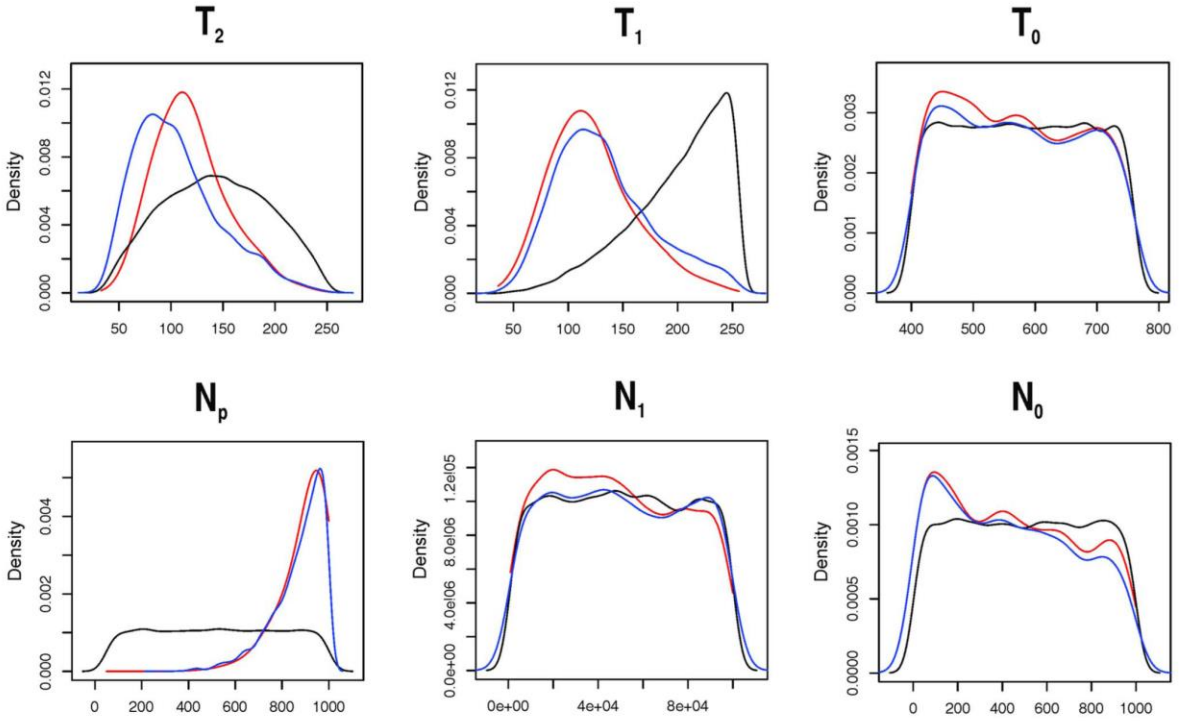

Figure 2. Prior (black), posterior (red) and retained (blue) simulations distributions of time (in generations) and size (2n) of parameters of the demographic model based on Greenberg's [22] language classification.
doi:10.1371/journal.pone.0064099.g002

harshly criticized from a methodological point of view [41–43]. According to Greenberg [22], with the exception of the Na-Dene and Eskimo-Aleut language groups, all other Native American languages belong to the single macro-family, named Amerind. This classification was regarded as reductionist by some scholars [44]. In this context, an important issue to consider is the pace of change; language, like other cultural traits, can change in a single generation [5]. The reconstruction of remote language families could be very different if the time period considered is 10,000 or 200,000 YBP [45]. Apart from these caveats and criticisms, it is noteworthy that Reich et al. [46] using information from~365,000 SNPs genotyped in individuals from 69 Siberian and Native American populations, suggested that the latter descend from at least three streams of Asian gene flow, a compatible scenario with the three major linguistic divisions originally proposed by Greenberg (Amerind, Eskimo–Aleut and Na-Dene).

Table 3. Posterior characteristics of the parameters of the model designed based on Greenberg's [22] classification given the genetic diversity of 381 autosomal STRs.

Parameter	Posterior distribution			Estimation accuracy			
				R^2 [2]	RMSE [3]		P-value [4]
	Mode	Median	HPDI[1] (95%)		Mode	Median	
T_0	10,905	14,040	10,136–18,683	0.00	3,625	2,675	0.00
T_1	2,779	3,094	1,480–5,294	0.03	1,300	1,000	0.05
T_2	2,666	2,812	800–4,382	0.40	925	850	0. 71
N_0	52	419	2–985	0.00	423	292	0.47
N_1	19,905	45,852	2,492–96,020	0.00	40,407	28,474	0.92
N_P	967	912	709–1,000	0.74	117	106	0.57

[1]Highest posterior density interval, which is the continuous interval of parameter values with highest posterior density.
[2]Coefficient of determination (R^2) obtained when regressing the parameter against the summary-statistics.
[3]Root mean squared error.
[4]P-value considering Kolmogorov-Smirnoff's test for uniformity of posterior quantiles.
doi:10.1371/journal.pone.0064099.t003

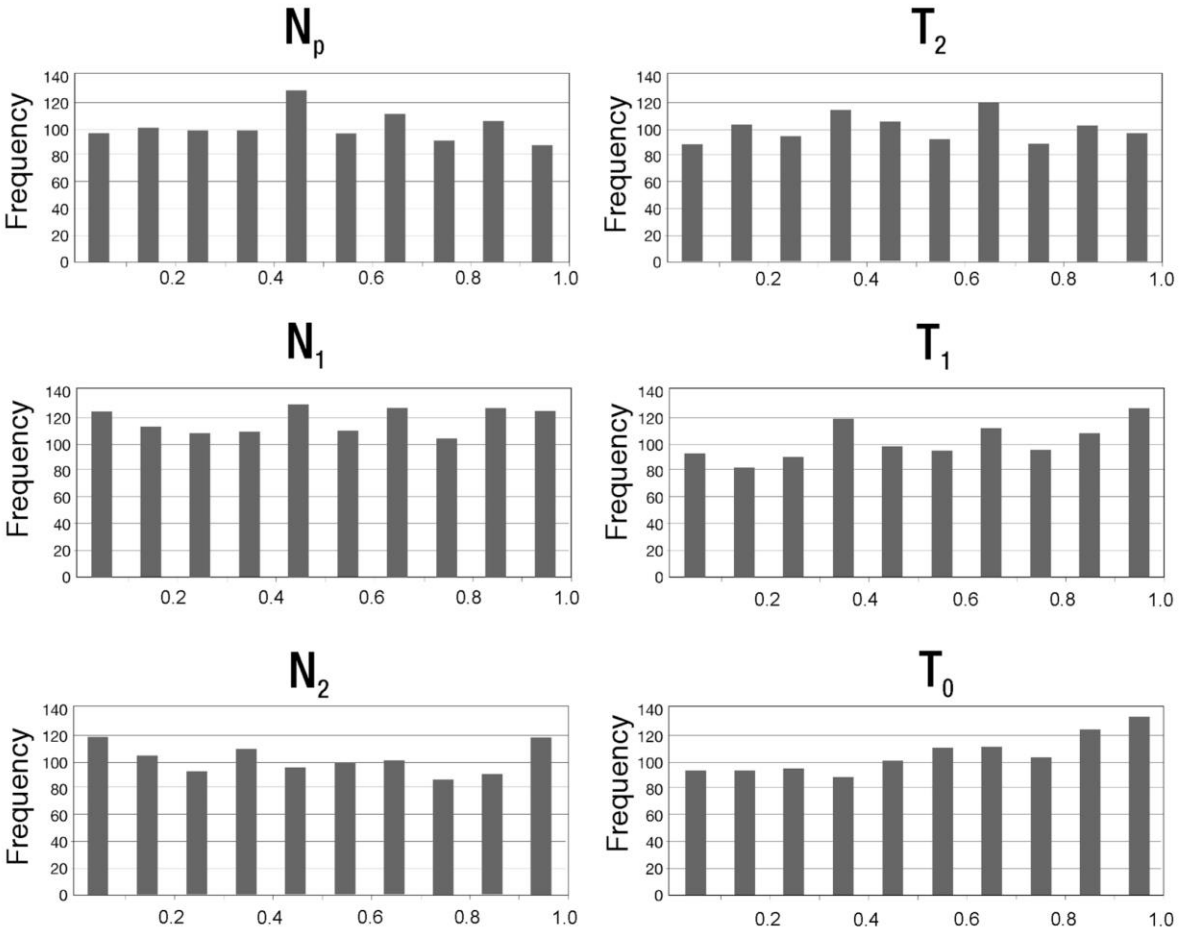

Figure 3. Quantile distributions (x-axis) of the known parameter values as inferred from the posterior distributions for 1,000 pseudo-observed data sets generated under Greenberg's [22] model.
doi:10.1371/journal.pone.0064099.g003

Greenberg's classification links the Tupí and Arawakan in the Equatorial-Tucanoan group and denies any closer relationship between the Tupí and the Macro-Jê or Arawakan and Andean, as proposed Campbell [23]; or between the Chibchan-Paezan and the Andean, as suggested Loukotka [20].

Notice that for the first time a study relating genetics and language in South America employed the ABC, a statistical framework that allows the use of realistic models which include gene flow and variances in effective population sizes along time and among populations, as well as the use of methods for controlling the quality of the estimates. Therefore, the relationship between any pair of population groups more likely reflects common origin rather than recent gene flow.

As explained in the results, the posterior estimates of T_0, N_0, and N_1 in the model based on Greenberg's classification were not very informative given their confidence intervals being very similar to the prior distributions (Fig. 2 and Table 3) and also not very reliable given their very low coefficients of determination (R^2). However, since the focus of this investigation was to unravel between-population relationships, these parameters are not of interest and could be considered 'nuisance parameters' (see [19]), i.e. they are not of immediate interest but must be accounted for in the analysis of the other parameters.

On the other hand, T_2, N_P and possibly T_1 estimates from Greenberg's scenario seem to be reliable based on the R^2 values (Table 3). The current effective deme size (N_P, 709 to 1,000 diploid individuals) matches Ray's et al. [29] estimates, which range from 751 to 904. T_1 and T_2 are exclusive to our models, and it is not possible to compare them with other genetic estimates. The Tupí and Arawakan divergence (T_2) was estimated to have happened from 800 to 4,382 years ago, with a higher probability of having occurred 2,812 years before present, while the time for the first emergence of structure in South Amerindian groups (T1), indicative of a most recent common ancestor, was dated from 1,480 to 5,294 YBP, with a higher probability at 3,094 years ago (Table 3).

How do these values compare with those obtained from linguistic information? Quechua, an Andean language, emerged 1,150 years before present according to Campbell [23]. The Arawakan group appears to have been formed at 3,000 [24] to 4,000 [47] years ago. The origin of the Chibchan-Paezan languages is dated at sometime between 3,000 and 5,600 before present [23]. Swadesh [48] and Brown [31] estimates for the

Chibchan languages emergence are included in this range (5,000 and 4,484 respectively). Jê languages origin is dated between 3,000 to 6,856 years before present according to different authors [24,48,49]; more specifically the Kaingang might have emerged 3,000 years ago [24]. The origin of the Tupi-Guarani is dated at some point between 2,000 and 5,000 YBP [24], while Guarani, according to Noelli [50] is 2,000 years old.

Confidence intervals in our genomic approach are large, and those calculated using linguistic data have not been obtained through rigorous statistical criteria. All in all, however, the numbers are not very different, pointing to a relative concordance between the interpopulation genomic and linguistic splits.

Conclusion

The questions raised in the introduction can now be answered. (a) Greenberg's language classification [22] presents a better fit to the current genome-wide diversity in South America when compared to those of the other linguists, although Campbell's is also compatible with the genomic data; (b) We estimated the time for the emergence of the structure between present day major language groups in South America around 3,100 ago, while the Tupí and Arawakan languages fission seem to have been more recent, around 2,800 years ago; and (c) Although confidence intervals are large, there is general agreement between split times estimated through genomic and linguistic data.

Supporting Information

Table S1 Classification of the five languages considered in this study. When available, the date of origin of the language is given in parenthesis.
(DOCX)

Table S2 Identification numbers of the 381 STR used in our analyses.
(DOCX)

Acknowledgments

We thank the members of the investigation team that published the STR data set. Nelson J. Fagundes for providing information about software use, and members of the Laboratório de Alto Desempenho/PUCRS for providing access to the computer clusters used for the analyses.

Author Contributions

Conceived and designed the experiments: CEGA RBM MCB SLB TH. Performed the experiments: CEGA. Analyzed the data: CEGA. Contributed reagents/materials/analysis tools: LSB. Wrote the paper: CEGA RBM VR MCB SLB FMS TH.

References

1. Reali F, Griffiths TL (2010) Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. Proc R Soc B 277: 429–436.
2. Hunley K, Bowern C, Healy M (2012) Rejection of a serial founder effects model of genetic and linguistic coevolution. Proc R Soc B 279: 2281–2288.
3. Salzano FM, Hutz MH, Salamoni SP, Rohr P, Callegari-Jacques SM (2005) Genetic support for proposed patterns of relationship among Lowland South American languages. Curr Anthropol 46: S121–S129.
4. Richerson PJ, Boyd R, Henrich J (2010) Gene-culture coevolution in the age of genomics. Proc Natl Acad Sci USA 107: 8985–8992.
5. Perreault C (2012) The pace of cultural evolution. PLoS One 7: e45150.
6. O'Rourke DH, Suarez BK (1986) Patterns and correlates of genetic variation in South Amerindians. Ann Hum Biol.13(1): 13–31.
7. Cavalli-Sforza LL (1997) Genes, peoples, and languages. Proc Natl Acad Sci USA 94: 7719–7724.
8. Fagundes NJ, Bonatto SL, Callegari-Jacques SM, Salzano FM (2002) Genetic, geographic, and linguistic variation among South American Indians: Possible sex influence. Am J Phys Anthropol.117:68–78.
9. Hunley KL, Cabana GS, Merriwether DA, Long JC (2007) A formal test of linguistic and genetic coevolution in native Central and South America. Am J Phys Anthropol 132: 622–631.
10. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, et al. (2007) Genetic variation and population structure in Native Americans. PLoS Genet 3: e185.
11. Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, et al. (2011) Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol 28: 2905–2920.
12. Jay F, François O, Blum MG (2011) Predictions of Native American population structure using linguistic covariates in a hidden regression framework. PLoS One 6: e16227.
13. Sharma G, Tamang R, Chaudhary R, Singh VK, Shah AM, et al. (2012) Genetic affinities of the central Indian tribal populations. PLoS One 7: e32546.
14. Hunley K, Long JC (2005) Gene flow across linguistic boundaries in Native North American populations. Proc Natl Acad Sci USA 102: 1312–1317.
15. Callegari-Jacques SM, Tarazona-Santos EM, Gilman RH, Herrera P, Cabrera L, et al. (2011) Autosome STRs in native South America - Testing models of association with geography and language. Am J Phys Anthropol 145: 371–381.
16. Long JC, Kittles RA (2003) Human genetic diversity and the nonexistence of biological races. Hum Biol 75: 449–471.
17. Kingman JFC (1982) The coalescent. Stochastic Process Appl 13: 235–248.
18. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in population genetics. Genetics 162: 2025–2035.
19. Csilléry K, Blum MG, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol 25: 410–418.
20. Loukotka Č (1968) Classification of South American Indian languages. Los Angeles: Latin American Studies Center, University of California.
21. Rodrigues AD (1986) Línguas brasileiras: Para o conhecimento das línguas indígenas. São Paulo: Edições Loyola.
22. Greenberg JH (1987) Languages in the Americas. Stanford: Stanford University Press.
23. Campbell L (1997) American Indian languages: The historical linguistics of native America. New York: Oxford University Press.
24. Urban G (1998) A História da cultura brasileira segundo as línguas nativas. In: História dos índios no Brasil (ed. MC . Cunha), 87–102. São Paulo: Companhia da Letras.
25. Lewis MP (2009) Ethnologue: languages of the world, 16th edition. Dallas, Tex.: SIL International. Available: http://www.ethnologue.com/. Accessed 10 november 2012.
26. Greenberg JH, Ruhlen M (2007) An Amerind etymological dictionary. Stanford: Stanford University Press.
27. Lischer HE, Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28: 298–299.
28. Excoffier L, Foll M (2011) Fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27: 1332–1334.
29. Ray N, Wegmann D, Fagundes NJ, Wang S, Ruiz-Linares A, et al. (2010) A statistical evaluation of models for the initial settlement of the American continent emphasizes the importance of gene flow with Asia. Mol Biol Evol 27: 337–345.
30. González-José R, Bortolini MC, Santos FR, Bonatto SL (2008) The peopling of America: Craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. Am. J. Phys. Anthropol 137: 175–187.
31. Brown CH (2010) Lack of linguistic support for Proto-Uto-Aztecan at 8900 BP. USA Proc Nat Acad Sci 107: E34.
32. Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72: 1171–1186.
33. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: A versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11:116.
34. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10: 564–567.
35. Kass RE, Raftery AE (1995) Bayes Factor. J Am Statist Assoc 90: 773–795.
36. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Mol Biol Evol 16: 1791–1798.
37. R Development Core Team. (2011) R: a language and environment for statistical computing.Vienna, Austria, R Foundation for Statistical Computing. Available: http://www.R-project.org/. Accessed 29 October 2012.
38. Neuenschwander S, Largiadèr CR, Ray N, Currat M, Vonlanthen P, et al. (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. Mol Ecol 17: 757–772.

39. Bolnick DA, Shook BA, Campbell L, Goddard I (2004) Problematic use of Greenberg's linguistic classification of the Americas in studies of Native American genetic variation. Am J Hum Genet 75: 519–523.
40. Cavalli-Sforza LL, Minch E, Mountain JL (1992) Coevolution of genes and languages revisited. Proc Natl Acad Sci USA 89: 5620–5624.
41. Dürr M, Whittaker G (1995) The methodological background to the Na-Dene controversy. In: Language and culture in native North America – Studies in honor of Heinz-Jürgen Pinnow. (eds. M. Dürr, E. Renner, W. Oleschinski), 102–122. München and Newcastle: LINCOM.
42. Matisoff JA (1990) On megalocomparison. Language 66: 106–120.
43. Campbell L (2008) How to show languages are related: Methods for distant genetic relationship. In: The handbook of historical linguistics (eds. BD. Joseph, RD. Janda), 262–282. Oxford: Blackwell.
44. Adelaar WFH (1989) Review of Language in the Americas by Joseph H. Greenberg. Lingua 78: 249–255.
45. Trask RL (1999) Why should a language have any relatives? In: Nostratic: Examining a linguistic macrofamily (eds. C. Renfrew, D. Nettle), 157–176. Cambridge: McDonald Institute for Archaeological Research.
46. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, et al. (2012) Reconstructing Native American population history. Nature 488: 370–374.
47. Hornborg A (2005) Ethnogenesis, regional integration, and ecology in prehistoric Amazonia: Toward a system perspective. Curr Anthropol 46: 589–620.
48. Swadesh M (1959) Mapas de clasificación linguística de México y las Américas México, DF: Universidad Nacional Autónoma de México.
49. ASJP – The Automated Similarity Judgement Program (2012) Available: http://email.eva.mpg.de/awichmann/ASPJHomePage.htm. Accessed 10 October 2012.
50. Noelli FS (1998) The Tupi: Explaining origin and expansions in terms of archeology and of historical linguistics. Antiquity 72: 648–663.

Supplementary Table S1. Classification of the five languages considered in this study. When available, the date of origin of the language is given in parenthesis.

Population	Hierarchic level	Campbell [23]	Urban [24]	Greenberg [22]	Greenberg and Ruhlen [26]	Lewis [25]	Loukotka [20]	Rodrigues [21]
Kogi	0	Chibchan-Paezan (Chibchan + Chocoan)						
	1	Chibchan (56 centuries ago or sometime after 3000 B.C.)		Amerind	Amerind	Chibchan	Languages of Andean tribes	
	2	Chibchan B		Chibchan-Paezan	Southern	Aruak	Northern division	
	3	Eastern Chibchan		Chibchan	Andean-Chibchan-Paezan	Kogi	Chibcha, stock	
	4	Colombian subgroup		Nuclear Chibchan	“Chibchan-Paezan”		Arhuaco group	
	5	Northern Colombian group		Aruak	Chibchan		Koghi (Kogi)	
	6	Arhuacan		Kagaba (Kogi)	Nuclear Chibchan			
	7	Cágaba (Kogi)			Aruak			

	8				Kagaba (Kogi)			
Aymara	0	Quechumaran stock (Quechumaran + Aymaran)						
	1	Aymaran		Amerind	Amerind	Aymaran	Languages of Andean tribes	
	2	Aymara		Andean	Southern	Aymara	South Central division	
	3			Aymara	Andean-Chibchan-Paezan		Aymara, stock	
	4			Aymara	Andean		Aymara	
	5				Aymara			
	6				Aymara			
Piapoco	0	Quechumaran stock (Quechumaran + Aymaran						
	1	Maipurean	Arawak	Amerind	Amerind	Arawakan	Languages of tropical forest tribes	
	2	Northern division	Maipure (3000 years ago)	Equatorial-Tucanoan	Southern	Maipuran	North Central division	

	3	Upper Amazon branch	Setentrional	Equatorial	"Equatorial-Tucanoan-Ge-Pano-Carib"	Northern Maipuran	Arawak, stock	
	4	Western Nawiki subbranch	Piapoco	Macro-Arawakan	Equatorial-Tucanoan	Inland	Caquetio group	
	5	Piapoko group		Arawakan	Equatorial	Piapoco	Piapoco	
	6	Piapoco		Maipuran	Macro-Arawakan			
	7			Piapoco	Arawakan			
	8				Maipuran			
	9				Piapoco			
Guarani	0	(Similarities between Tupian and Jean languages)[1]						
	1	Tupían stock	Tupi-Karib-Macro-Jê (Before 6000 years ago)	Amerind	Amerind	Tupi	Languages of tropical forest tribes	Tronco Tupí
	2	Tupí-Guaraní family	Macro-Tupi (Between 3000 and 5000 years ago)	Equatorial-Tucanoan	Southern	Tupi-Guarani	North Central division	Família Tupí-Guaraní

	3	Guaraní group	Tupi-Guarani (2000 or 3000 years ago)	Equatorial	"Equatorial-Tucanoan-Ge-Pano-Carib"	Subgroup I	Tupi, stock	Guaraní
	4	Guaraní language (área)	Guarani	Kariri-Tupi	Equatorial-Tucanoan	Guaraní (Guarani)	Guarani group	"vários" Guaraní (Guarani)
	5	"several" Guarani		Tupi	Equatorial		Guaraní (Guarani)	
	6			Guaraní	Kariri-Tupi			
	7				Tupi			
	8				Guaraní			
Kaingang	0	(Similarities between Tupian and Jean languages) [1]						
	1	Jean	Tupi-Karib-Macro-Jê (Before 6000 years ago)	Amerind	Amerind	Macro-Ge	Languages of Paleo-American tribes	Tronco Macro-Jê
	2	Southern branch	Macro-Jê (At least 5000 or 6000 years ago)	Ge-Pano-Carib	Southern	Ge-Kaingang	Division of Central Brazil	Família Jê

	3	Kaingang	Jê (3000 years ago or more)	Macro-Ge	"Equatorial-Tucanoan-Ge-Pano-Carib"	Kaingang	Kaingán, stock	Kaingáng (Kaingang)
	4		Kaingang	Ge-Kaingan	Ge-Pano-Carib		Kaingán (Kaingang)	
	5			Kaingan	"Ge-Pano"			
	6			Kaingan (Kaingang)	Macro-Ge			
	7				Ge-Kaingang			
	8				Kaingang			
	9				Kaingang			

[1] Campbell [23] sees similarities between Tupian and Jean languages. In our models, those languages were grouped when Campbell's classification was considered.

Supplementary Table S2. List of the microsatellite (STR) loci used in the analysis.

Loci	Loci	Loci	Loci
AAAAC001_9	AGAT119M_1	ATAA009_8	D11S1981
AAAT105ZP_2	AGAT126_5	ATAA018P_8	D11S1998
AAAT111_5	AGAT128_3	ATAC026P_14	D11S4459
AAAT121P_8	AGAT130_5	ATAC037P_7	D11S4463
AAAT134_13	AGAT132_17	ATAG042_8	D11S4464
AAC023_3	AGAT133_7	ATAG053P_10	D12S1064
AAC030_3	AGAT136M_20	ATAG078P_5	D12S1300
AACAT001_13	AGAT140P_9	ATAG089P_18	D12S2070
AAT071_3	ATA009_1	ATC033_6	D12S297
AAT107_16	ATA063_16	ATC3D09_3	D12S372
AAT226_16	ATA069P_14	ATC4D07_3	D12S373
AAT238_1	ATA103C03P_16	ATCT018_4	D12S395
AAT245_17	ATA16D09_2	ATCT035_20	D13S1807
AAT246_4	ATA18C09P_9	ATCT050_18	D13S317
AAT253P_12	ATA20B07_10	ATGA020_6	D13S793
AAT256P_6	ATA21F01_4	ATGT006Z_10	D13S796
AAT261_9	ATA25D12_11	ATT023_8	D13S800
AAT268_11	ATA27C11_11	ATTT019M_22	D13S894
AATA019_8	ATA29E07M_2	ATTT030_6	D13S895
AATA053_15	ATA2E04_1	CATA002Z_16	D14S1426
ACT3F12_13	ATA31F09M_7	CTAT016_9	D14S608
AGAT021_2	ATA38A05_1	D10S1208	D14S617
AGAT030P_5	ATA42G04P_9	D10S1222	D14S742
AGAT049P_7	ATA44F05P_4	D10S1230	D15S642
AGAT060_18	ATA57D10M_3	D10S1239	D15S643
AGAT084_12	ATA58E08ZP_17	D10S1426	D15S659
AGAT099P_5	ATA65H08P_9	D10S1432	D15S816
AGAT110P_13	ATA73A08M_1	D10S2327	D16S2616
AGAT115_8	ATA73C05P_12	D10S2470	D16S2621

AGAT116P_14	ATA80B10Z_10	D10S677	D16S2624
AGAT118_1	ATA85B10P_3	D11S1392	D16S3253
D16S539	D20S480	D3S3039	D6S305
D16S769	D20S482	D3S3045	D6S474
D17S1294	D21S1432	D3S4523	D6S942
D17S2180	D21S1437	D4S1629	D7S1799
D17S2196	D21S1440	D4S1644	D7S1802
D18S1370	D21S2052	D4S1647	D7S1808
D18S1371	D22S683	D4S2368	D7S1818
D18S1376	D22S686	D4S2397	D7S1824
D18S535	D2S1328	D4S2431	D7S2204
D18S542	D2S1352	D4S2623	D7S3051
D18S851	D2S1360	D4S2632	D7S3056
D18S858	D2S1394	D4S3243	D7S3061
D18S877	D2S1399	D4S3248	D7S3070
D19S246	D2S1400	D4S3360	D7S821
D19S254	D2S1776	D5S1456	D8S1048
D19S559	D2S1780	D5S1457	D8S1108
D19S589	D2S2944	D5S1462	D8S1110
D19S714	D2S2952	D5S1501	D8S1113
D1S1589	D2S2968	D5S1505	D8S1136
D1S1594	D2S2972	D5S2488	D8S1477
D1S1596	D2S410	D5S2845	D8S261
D1S1597	D2S427	D5S2849	D8S592
D1S1653	D2S434	D5S817	D9S1120
D1S1677	D3S1763	D5S820	D9S1121
D1S3462	D3S1764	D6S1006	D9S1122
D1S3669	D3S2387	D6S1009	D9S1838
D1S3721	D3S2398	D6S1017	D9S2157
D1S518	D3S2409	D6S1277	D9S2169
D1S551	D3S2427	D6S2410	D9S922

D20S1143	D3S2432	D6S2436	D9S930
D20S164	D3S2460	D6S2439	D9S934
D9S938	GATA156H01M_8	GATA51F04P_14	MFD433-AGAT010_3
F13A1-D6S	GATA157H01_18	GATA5E06P_9	MFD442-GTTT002_7
GAAT1A5_2	GATA165A11M_9	GATA61F04_9	MFD455-AAT052_9
GATA036_18	GATA167C12_12	GATA63B12P_15	MFD466-TTA001_16
GATA045_14	GATA169F02_17	GATA63F01_2	NA-D10S-2
GATA060_8	GATA173A03_18	GATA6F05P_22	NA-D12S-1
GATA10H07P_17	GATA174G01_2	GATA70F12M_2	NA-D12S-2
GATA126A06M_2	GATA175H06M_9	GATA71E06_11	NA-D13S-1
GATA129D03M_4	GATA178C11M_3	GATA72A06_3	NA-D14S-1
GATA129G03P_6	GATA193D02_1	GATA73B08M_11	NA-D17S-1
GATA12A08P_5	GATA194A05M_2	GATA73D05_18	NA-D18S-1
GATA131D09_3	GATA194B06P_2	GATA73D11P_5	NA-D1S-2
GATA134F03P_10	GATA194H05Z_1	GATA7F09_12	NA-D1S-3
GATA135F02P_1	GATA196C10P_10	GATA81E09_20	NA-D6S-1
GATA136A04_14	GATA22F01_15	GATA81F06_10	NA-D8S-2
GATA137A12M_7	GATA22H04M_9	GATA85D10_18	NA-D9S-1
GATA137B09_13	GATA23A02_2	GATA87D11_7	SCA10_22
GATA138B05_5	GATA23G09_1	GATA8H05_2	TAAA014P_6
GATA139B09P_5	GATA27Z_9	GATA90G05P_10	TAAAA006_4
GATA140E03_16	GATA29A06M_2	GATA90G11M_14	TACA003_10
GATA141B10M_5	GATA29B11_11	GATA91D12M_2	TAGA002M_2
GATA143C02_15	GATA29C09P_6	GATA91G06_14	TAT028P_7
GATA145G10M_7	GATA2B02Z_1	GATA91H01_12	TAT032Z_15
GATA146B10_3	GATA30A08M_6	GGAA19H02_12	TATC012_8
GATA146D07_3	GATA30B11_4	GGAA20F08_1	TATC028_1
GATA148F04P_21	GATA31B11_17	GGAA22C05_12	TATC057_21
GATA148G10P_2	GATA31H11P_5	GGAA23C07_1	TATG002P_7
GATA149B10M_2	GATA3H11_8	GGAA30H04_14	TCAT006ZP_22
GATA152F04M_3	GATA4E04_7	GGAT2G03_3	TCTA015M_22

GATA152F05L_1	GATA51A07P_5	GGAT2G06M_12	TCTA020_9
GATA153F11_15	GATA51D11P_5	GGAT3G09M_9	TCTA023P_14
TCTA025_11	TTA032Z_6	TTAT027P_15	TTTA001M_7
TPO-D2S	TTAT023Z_16	TTCA004P_8	TTTA040_3
TTA008P_11			

II.IV) Artigo 4

Amorim CEG, Salzano FM, Bortolini MC, Hünemeier T (2013) *Differing evolutionary histories of the ACTN3*R577X polymorphism among the major human geographic groups* (Manuscrito em preparação a ser submetido a *European Journal of Human Genetics*).

Differing evolutionary histories of the ACTN3*R577X polymorphism among the major human geographic groups

Carlos Eduardo G. Amorim[1], Francisco M. Salzano[1], Maria C. Bortolini[1] and Tábita Hünemeier[1]

[1]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil.

Running headline: Differing *ACTN3* evolutionary histories

A proposal was made that the functional *ACTN3R577X polymorphism might have evolved due to selection in Eurasian human populations. To test this possibility we surveyed all available population-based data for this polymorphism and made a comprehensive evolutionary analysis of its genetic diversity, trying to assess the action of adaptive and random mechanisms on its variation in the major human groups throughout the world. The derived 577X allele increases in frequency with distance from Africa, reaching the highest frequencies in the Americas. Positive selection, detected by an extended haplotype homozygosisty test, was consistent with the Eurasian data only, but simulations with neutral models could not fully explain the Amerindian results. It is possible that the peculiar Native American population structure could be responsible for the observed allele frequencies, which would have resulted from a complex interaction between selective and random factors.**

Keywords: alfa-actinin; human dispersal; positive selection; rs1815739.

INTRODUCTION

The human *ACTN3* (α-actinin skeletal muscle isoform 3) gene is located in the long arm of chromosome 11 and encodes an α-actinin binding protein solely expressed in fast twitching skeletal muscle fibers (Mills et al., 2001). A non-deleterious C→T transversion (rs1815739) converts arginine at residue 557 of the ACTN3 protein to a premature stop-codon (North et al., 1999).

Evidence shows that this polymorphism may affect muscle performance (Alfred et al., 2011) in ways that the derived X allele is under-represented in sprint/power athletes (Yang et al., 2003). Hence it is believed that ACTN3 is required for optimal muscle performance at high velocity, but it is still doubtful if the derived allele could improve endurance performance (Alfred et al., 2011). Experimental approaches with knock-out mice demonstrated that the lack of ACTN3 in muscle cells results in a significant shift in the metabolic pathways of the lactate-based fast fibers, leading to a slower but more efficient aerobic pathway normally associated with slow muscle fibers (MacArthur et al., 2007). The lack of this protein is then compensated by an ACTN2 up-regulation (Mills et al., 2001).

These data suggest that the high frequency of the 557X allele in some human populations could have had resulted from selective pressures for increased metabolic efficiency, enhancing the capability for endurance running. The consequence could be the emergence of novel kinds of hunting – such as persistence hunting – that could also have played a role in range expansions (Bramble and Lieberman, 2004). It was then hypothesized that this gene could have had evolved under balancing selection earlier in *Homo* history (MacArthur and North, 2004) and that more recently it could have been

subjected to positive selection in European and Asian populations (MacArthur et al., 2007), but not in Africa (Schlebusch et al., 2012).

Despite the intricate scenario that was proposed, only few attempts to analyze it into an evolutionary perspective context was made (for instance, Friedlander et al., 2013), and the focus was on association of muscle metabolism and athletes´ performance (Alfred et al., 2011 and references therein).

We gathered all the available population-based rs1815739 polymorphism data creating a databank of over 5600 genotyped chromosomes distributed in 121 autochthonous populations worldwide, the most complete databank so far on the variability of the "athlete gene". The question was how the genetic variation observed at the human *ACTN3* has been impacted by human evolutionary history? To answer this question we used the Extended Haplotype Homozygosity test, estimated the age of emergence of the derived allele, and devised five alternative models for the peopling of the Americas, to verify how the empirical data would fit them.

SUBJECTS AND METHODS

Data source and populations

Data were first obtained from three public databanks: 1000 Genomes Project (1000 Genomes Project Consortium et al., 2012), HapMap (International HapMap Consortium, 2003), and Human Genetic Diversity Panel (HGDP; Li et al., 2008). Then an extensive literature review was conducted and three articles with rs1815739 genotypes from healthy human individuals using random population samples were retrieved (Fattahi and Najmabadi, 2012; Reich et al., 2012; Schlebusch et al., 2012). Association studies were not included in our analyses, since samples were not taken at random. A total of 3989 genotypes distributed in 150 population samples were obtained from these

sources. We first eliminated the populations with known substantial recent admixture. When individual information on recent admixture was available, we also removed these recent migrants. Since some of the remaining samples overlaped in the different reports, only the largest sample was included, and we merged populations with the same name that were separated in the original study. This procedure left us with 2806 genotypes for 121 populations worldwide. Readers are referred to the primary publications for further information on samples and populations.

Data processing and handling were performed with PGDSpider (Lischer and Excoffier, 2012) and with in-house built bash scripts.

ACTN3*R577X Allele Age

To estimate the age of emergence of the derived allele, we used equation $E(t1) = [-2p/(1-p)]\ln(p)$, where $E(t1)$ is its expected age in units of 2N generations and p is the derived allele frequency (Slatkin and Rannalla, 2000). We assumed a generation time of 25 years and $N = 6000$, which is often regarded as a minimum estimate for the effective population size of modern humans during the period before recent growth. Since this formula assumes neutral evolution for the locus under investigation, we used the average derived allele frequency in all pooled selected African populations in the formula, since there is no signal of selection in this continent (Schlebusch et al., 2012).

Positive selection inference

Additional information on the flanking region of rs1815739 was available for three major human populations: 384 individuals from Central and East Asia, 162 from Europe, and 260 from the Americas (Reich et al., 2012). This information includes the genotypes for 30 SNPs (rs4930359, rs905770, rs2282529, rs7947391, rs7925108, rs2511224, rs2305535, rs11227501, rs7951189, rs2298806, rs11227516, rs3816492,

rs3867132, rs490998, rs2275998, rs540874, rs560556, rs498045, rs556759, rs519380, rs624561, rs10791889, rs3782079, rs664297, rs4930390, rs569818, rs11601241, rs7948839, rs2167457, and rs7119426) located ~187 kb upstream and ~322 kb downstream of rs1815739, therefore spanning circa 510 kb. This dataset was used for the identification of positive selection in this genetic region and the same sampling procedure employed by the source publication (Reich et al., 2012) was followed.

To infer positive selection we used the Extended Haplotype Homozygosity (EHH) test, defined as the probability that any two randomly chosen chromosomes carrying the haplotype of interest are identical by descent from the core region to a distance *x;* and the Relative EHH (REHH), the factor by which EHH decays on the tested haplotype compared to the decay on all other haplotypes combined. We selected one SNP at a time for the core marker and the relative extended haplotype homozygosity (REHH) was computed for each haplotype and compared at increasing distances from these markers. These analyses were performed with Sweep (Sabeti et al., 2007) and haplotype phases were estimated with BEAGLE 3.3.2 (Browning and Browning, 2007).

Neutral Demographic Simulations

To test how the observed patterns of *ACTN3* genetic diversity are correlated to neutral evolution, we simulated genetic data for 1000 SNPs with ms (Hudson, 2002) under a wide range of demographic scenarios mimicking the prehistoric settlement of the Americas, since this is the continent where the highest frequencies of the 577X allele was observed (Schlebusch et al., 2012). In a similar approach to that developed by Schroeder et al. (2009), we analyzed this dataset – simulated under strict neutrality – to determine how often we could observe an allele with a similar distribution to that described for the polymorphism under investigation, which, in our case, is rs1815739,

that presents an allele with high frequency in the Americas and an average allele frequency differential of at least 0.29 in comparison to any other region of the globe (Li et al., 2008; Reich et al., 2012).

Five scenarios (Figure 1) were modeled for the split of Native Americans from Asians, as follows:

- Model A considers two derived populations with the same size;
- Model B adds a bottleneck to the previous model, in ways that the ratio between the effective population sizes (N_{EF}) of the derived populations is 0.15;
- Model C, in which this ratio is 0.06 and both populations undergo exponential growth, so that the ratio between N_{EF}s of the derived populations is 0.15 in the present (T_0);
- Model D, the same as B, but with population substructure in both derived populations;
- Model E, the same as C, but with population substructure in both derived populations.

Population samples and deme sizes were defined based on the empirical data available for the rs1815739 polymorphism, filtering out those populations with less than 4 available genotypes and considering only East Asians as the derived population living in Asia, since this is the region most genetically related to the present day Amerindians. That yielded 25 demes for the simulated population that mimicked Asia and 18 for the simulated populations that mimicked the Americas.

Different splitting times (T_2) between Asia and the Americas were adopted for each model in different runs as follows: 1,001, 740, and 500 generations before the present. Current Asian N_{EF} was assumed to be n=9,000 and the ancestral N_{EF} was equal

to that estimated for Asia during the split, which may vary according to the model as explained before.

For models D and E structure emerged at T_1, which assumed different values in different runs (295, 195, or 75 generations ago) and migration was allowed to happen according to the stepping-stone and island models separately. Under the island model, 5% of each subpopulation consists of migrants from all other subpopulations at random in each generation; while for the stepping-stone model, the same percentage was made of migrants from two other subpopulations. We also considered circumarctic migration for these models; gene flow would also occur between one Asian and one American subpopulation at the rate of 0.0022, and they would not exchange migrants with the other subpopulations from their respective continents.

This procedure replicated Schroeder et al.'s (2009) strategy. The ms command-lines employed by us were all based on those kindly provided by them and will be available on request. Modifications were done to accommodate differences between the sampling schemes and genetic systems.

RESULTS

Table 1 shows the rs1815739 derived allele frequencies in the analyzed populations pooled according to the six major geographical regions. Mozabite was excluded from Africa due its geographical location and possible influence of Middle Eastern gene flow. Supplementary Table S1 presents the 577X allele distribution for each population before pooling. Considering data from all Africans the allele age estimated for ACTN3*577X using population frequency information was 61,373 YBP (~2,455 generations).

The data from Asian and European populations (Figure 2 – A and B respectively) showed that the REHH of the 577X allele core haplotypes is significantly higher than the REHH of the ancestral allele core haplotypes on both sides of rs1815739 (REHH = 5.5 and 4.4 at 187 kb of distance on the 5' side; and REHH = 5.3 and 5.6 at 322 kb on the 3' region, for Asian and Europeans, respectively). Native Americans presented lower extension values (REHH = 1.7 at 187 kb on the 5' side, and REHH = 0.7 at 322 kb on the 3' region; Figure 2 – C).

In average 6.1% of the SNPs reproduced the pattern observed for the rs1815739 polymorphism. Models A, B, C, D, and E present respectively 0.8, 1.8, 10.5, 6.9, 6.4% of SNPs with at least 0.29 of a frequence differential between the simulated populations. In general, this percentage was higher with increasingly more complex and realistic models. The inclusion of a bottleneck for the population simulating Amerindians was the factor that most contributed for the emergence of alleles with a high frequency in this continent. Different migration models (i.e. island and stepping-stone) did not affect the allele's emergence with such intercontinental differentiation; and circumarctic migration did it just slightly, lowering the proportion of simulations presenting the rs1815739 polymorphism pattern.

DISCUSSION

Human populations are known to have been subjected to different population bottlenecks as they expanded their range along the planet (Fagundes et al., 2007; 2008). These bottlenecks affect genetic diversity in ways that one allele may become fixed in some groups whereas it may reach intermediate or very low frequencies in others. In fact, high interpopulation diversity due solely to neutrality is expected to be seen in a large proportion of loci in the human genome (Hofer et al., 2009); on the other hand the

different selective pressures that humans encountered during the colonization of the world may also cause allele frequency heterogeneity between populations (Hancock et al., 2010), since some alleles may be more advantageous than others in specific environments. Dissecting the effects of drift from those of natural selection may be a hard task in some cases. In this work, we considered the *ACTN3* evolutionary history in the major human continental groups and showed the importance of both processes on its evolution.

Some studies have suggested that *ACTN3* may be evolving under natural selection (Bramble and Lieberman, 2004; MacArthur and North, 2004; Friedlander et al., 2013) and another has demonstrated it empirically (MacArthur et al., 2007). Our study confirms the hypothesis of an adaptive evolutionary history for rs1815739 at *ACTN3*, although random drift may also be of importance in some regions of the world.

The rs1815739 derived allele frequency shows a general trend of increase with distance from Africa, reaching its highest frequencies in the Americas. Despite the high frequency of this allele in this continent, the corresponding adaptive sweep could not be confirmed for these populations, since they presented REHH values about five times lower than those found in Asia and Europe for the haplotypes carrying the derived 577X allele. For these latter populations, there is a consistent signal of strong positive selection acting upon this locus. A few hypotheses for the high fitness of carriers of the 577X allele in Eurasia can be suggested; among them, the most obvious would be that of advantages in muscle activity in endurance running in scenarios such as those proposed by Bramble and Lieberman (2004): ease of access to carcasses and the emergence of new strategies for acquiring food, such as persistence hunting. Additionally, both alleles could have been evolving under selection for a certain time.

The present estimate of the allele´s emergence at 84.2 thousand years ago is in the range calculated for the time humans left Africa (45.0 to 87.5 years ago; Laval et al., 2010), and gives further support for the hypothesis that the derived allele may have presented selective advantages during human dispersal and the out-of-Africa migration.

In accordance to the lack of evidence for *ACTN3* positive selection in the Americas as indicated by the EHH tests, the neutral demographic simulations also suggest that, given the peculiar demographic scenario for the peopling of the Americas, it is plausible to observe alleles that have raised in frequency in this continent due to drift effects, without any need for evoking adaptive processes. However, the proportion of loci with similar intercontinental differentiation to that of the rs1815739 polymorphism was low, indicating that the devised models cannot completely reproduce ACTN3*R577X evolutionary and demographic history.

The simulated scenarios considered several different aspects of the theories on the demographic dynamics of the peopling of the New World, such as the variance in the strength of the bottleneck during the entrance in the continent (Fagundes et al., 2008; Kitchen et al., 2008); the presence of circumarctic migration (González-José *et al*., 2008); different patterns of migration among demes; population growth (Fagundes et al., 2008) and stasis; and population structure or "tribalization" (Neel and Salzano, 1967). There might be other factors influencing the rs1815739 distribution that could not be reproduced in these scenarios or detected by the EHH tests. Hence, neither a situation in which this allele presents selective advantage or that it evolves solely due to random drift solely can be indicated. A combination of both, with possible changes in the selective pressure over time and space and the loss or fixation of the beneficial allele in certain groups due to drift could be suggested.

South Amerindian structure can distort positive selection signals and make them hard to be identified. Similar results as those found here were obtained for *PAX9* (Paixão-Côrtes et al., 2011), *ABCA1* (Hünemeier et al., 2012), and *KIR* (Augusto et al., 2013), for which no signal of positive selection was found in the Americas (or specifically in South America for *ABCA1*), despite the evidence for non-neutrality in other populations. As more realistic simulations and robust selection inference methods are developed, a clearer evolutionary scenario for such genes in the Americas could be proposed.

The hypothesis of a recent high increase in frequency of the rs1815739 derived allele due to selection in the Americas is much appealing. An allele that could enhance human dispersal would be likely advantageous during the settlement of the New World, especially since it has happened in relatively short time (Dillehay, 2009). Moreover, the involved populations certainly have used persistence hunting and this way of living could act as a selective pressure favoring the derived allele. For instance, the Tarahumara Amerindians, besides practicing persistence hunting, also have customs related to endurance running (Balke and Snow, 1965) and the derived X allele is nearly fixed among them (Victor Acuña-Alonzo, personal communication). Thus in spite of the doubts about the exact role of this allele in enhancing human fitness (Alfred et al., 2011), our study does not exclude the view that the X allele may enhance human capability for enduring exercising and related practices.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGEMENTS

We thank the authors that generated the original genetic datasets and Dr. Kari Schroeder for kindly providing the ms command-lines. This work was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, Programa de Apoio a Núcleos de Excelência, Brazil.

1000 Genomes Project Consortium, Abecasis GR, Auton A *et al*: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56-65.

Alfred T, Ben-Shlomo Y, Cooper R *et al*: ACTN3 genotype, athletic status, and life course physical capability: meta-analysis of the published literature and findings from nine studies. *Hum Mutat* 2011; **9**: 1008-18.

Augusto DG, Piovezan BZ, Tsuneto LT, Callegari-Jacques SM, Petzl-Erler ML KIR gene content in amerindians indicates influence of demographic factors. *PLoS One* 2013; **8**: e56755.

Balke B, Snow C: Anthropological and physiological observations on Tarahumara endurance runners. *Am J Phys Anthropol* 1965; **23**: 293-301.

Bramble DM, Lieberman DE: Endurance running and the evolution of *Homo*. *Nature* 2004; **432**: 345-52.

Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084-97.

Dillehay TD: Probing deeper into first American studies. *Proc Natl Acad Sci U S A* 2009; **106**: 971-8.

Fagundes NJ, Kanitz R, Eckert R *et al*: Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet* 2008; **82**: 583-92.

Fagundes NJ, Ray N, Beaumont M *et al*: Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* 2007; **104**: 17614-9.

Fattahi Z, Najmabadi H: Prevalence of ACTN3 (the athlete gene) R577X polymorphism in Iranian population. *Iran Red Crescent Med J* 2012; **14**: 617-22.

Friedlander SM, Herrmann AL, Lowry DP *et al:* ACTN3 allele frequency in humans covaries with global latitudinal gradient. *PLoS One* 2013; **8**: e52282.

González-José R, Bortolini MC, Santos FR, Bonatto SL: The peopling of America: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. *Am J Phys Anthropol* 2008; **137**: 175-87.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A: Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philos Trans R Soc Lond B Biol Sci* 2010; **365**: 2459-68.

Hofer T, Ray N, Wegmann D, Excoffier L: Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet* 2009; **73**: 95-108.

Hudson, RR: Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 2002; **18**: 337-8.

Hünemeier T, Amorim CEG, Azevedo S *et al*: Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. *PLoS One* 2012; **7**: e38862.

International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789-96.

Kitchen A, Miyamoto MM, Mulligan CJ: A three-stage colonization model for the peopling of the Americas. *PLoS One* 2008; **3**: e1596.

Laval G, Patin E, Barreiro LB, Quintana-Murci L: Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 2010; **5**: e10284.

Li JZ, Absher DM, Tang H *et al:* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **319**: 1100-4.

Lischer HE, Excoffier L: PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 2012; **28**: 298-9.

MacArthur DG, North KN: A gene for speed? The evolution and function of alpha-actinin-3. *Bioessays* 2004; **26**: 786-95.

MacArthur DG, Seto JT, Raftery JM *et al*: Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nat Genet* 2007; **39**: 1261-5.

Mills M, Yang N, Weinberger R *et al*: Differential expression of the actin-binding proteins, alpha-actinin-2 and -3, in different species: implications for the evolution of functional redundancy. *Hum Mol Genet* 2001; **10**: 1335-46.

Neel JV, Salzano FM: Further studies on the Xavante Indians. X. Some hypotheses-generalizations resulting from these studies. *Am J Hum Genet* 1967; **19**: 554-74.

North KN, Yang N, Wattanasirichaigoon D, Mills M, Easteal S, Beggs AH: A common nonsense mutation results in alpha-actinin-3 deficiency in the general population. *Nat Genet* 1999; **21**: 353-4.

Paixão-Côrtes VR, Meyer D, Pereira TV *et al:* Genetic variation among major human geographic groups supports a peculiar evolutionary trend in PAX9. *PLoS One* 2011; **6**: e15656.

Reich D, Patterson N, Campbell D *et al:* Reconstructing Native American population history. *Nature* 2012; **488**: 370-4.

Sabeti PC, Varilly P, Fry B *et al:* Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007; **449**: 913-8.

Slatkin M, Rannala B: Estimating allele age. *Annu Rev Genomics Hum Genet* 2000; **01**: 225–49.

Schlebusch CM, Skoglund P, Sjödin P *et al:* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 2012; **338**: 374-9.

Schroeder KB, Jakobsson M, Crawford MH *et al*: Haplotypic background of a private allele at high frequency in the Americas. *Mol Biol Evol* 2009; **26**: 995-1016.

Yang N, MacArthur DG, Gulbin JP *et al*: ACTN3 genotype is associated with human elite athletic performance*. Am J Hum Genet* 2003; **73**: 627-31.

Table 1 Frequencies of the rs1815739 derived allele in autochthonous populations pooled into six geographical groups

Geographical group	Sample sizes (2n)	Allele frequencies
Africa	794	0.07
Middle East	146	0.38
Europe	445	0.44
Asia	992	0.49
Oceania	35	0.50
Americas	394	0.76

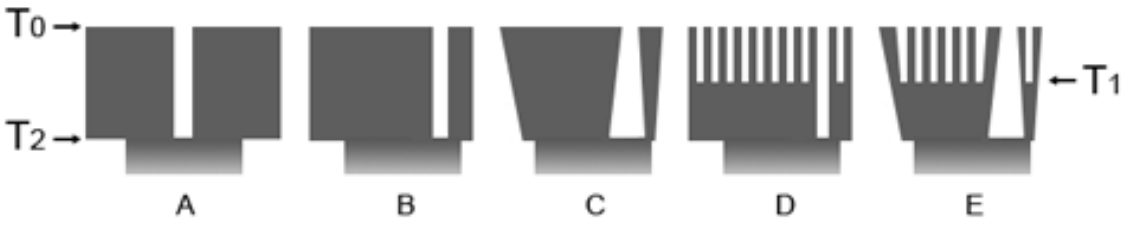

Figure 1 Demographic models used for the coalescent simulations: **A** – population split at T_2 with two derived populations with the same size; **B** – population split at T_2 with two derived populations with ratio between their effective population sizes (N_{EF}) equal to 0.15; **C** – population split at T_2 with two derived populations with ratio between their N_{EF}s equal to 0.06 at T_2 and 0.15 presently (T_0); **D** – same as model B, but with population substructure arising at T_1 in both derived populations; **E –** same as model C, but with population substructure arising at T_1 in both derived populations. The number of demes depicted does not correspond to the actual simulations. For further details see methods section.

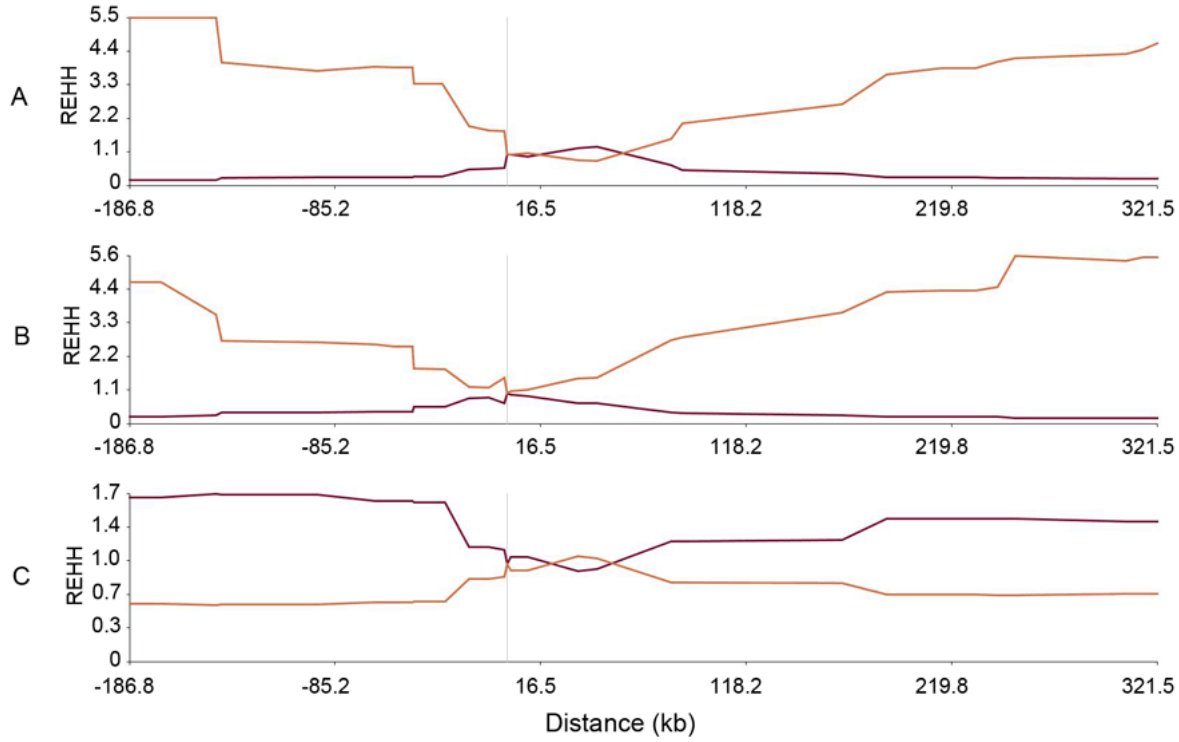

Figure 2 Relative extended haplotype homozygosity (REHH) for 30 SNPs located at a distance described in the X-axis from *ACTN3* rs1815739 spanning circa 510 kb. The orange line represents the homozygosis of the flanking SNPs relative to rs1815739 for haplotypes carrying the derived 577X mutation, while the red line represents the same for haplotypes carrying the ancestral allele. A: Asian; B: European; and C: American data.

Supplementary Table S1. Frequency of the rs1815739 derived allele in autochthonous populations worldwide

	Population	Sample size (2n)	577X allele frequency	Reference
Africa	Bantu in Kenya	12	0.08	[2]
	Bantu in South Africa	38	0.11	[6]
	Biaka Pygmies	30	0.13	[2]
	Gui Ghana Kgal	14	0	[6]
	Herero	16	0.12	[6]
	Juhoansi	34	0	[6]
	Karretjie	24	0	[6]
	Khomani	34	0.06	[6]
	Khwe	34	0.03	[6]
	Luhya	97	0.08	[3]
	Luhya	90	0.07	[1]
	Maasai	143	0.19	[1]
	Mandenka	24	0.17	[2]
	Mbuti Pygmies	15	0.03	[2]
	Mozabite	30	0.48	[2]
	Nama	14	0.07	[6]
	San	6	0	[2]
	Xun	26	0.08	[6]
	Yoruba	113	0.08	[1]
Middle East	Bedouin	48	0.42	[2]
	Druze	47	0.29	[2]
	Palestinian	51	0.42	[2]
Europe	Adygei	17	0.56	[2]
	British	89	0.49	[3]
	Finnish	93	0.35	[3]
	French	29	0.41	[2]
	French Basque	24	0.46	[2]
	Spanish	14	0.39	[3]
	North Italian	12	0.42	[2]
	Orcadian	16	0.59	[2]
	Russian	25	0.3	[2]
	Sardinian	28	0.48	[2]
	Toscan	98	0.42	[3]
Asia	Algonquin	2	1	[5]
	Altaian	3	0.67	[5]
	Balochi	25	0.46	[2]
	Brahui	25	0.32	[2]

	Population	Sample size (2n)	577X allele frequency	Reference
	Burusho	25	0.52	2
	Buryat	9	0.61	5
	Cambodians	11	0.45	2
	Chukchi	30	0.3	5
	Dai	10	0.4	2
	Daur	9	0.61	2
	Dolgan	4	0.62	5
	Evenki	10	0.4	5
	Han Chinese	97	0.41	3
	Hazara	17	0.38	2
	Hezhen	8	0.44	2
	Iran	210	0.44	4
	Japanese	89	0.49	3
	Kalash	25	0.86	2
	Ket	1	1	5
	Khanty	5	0.6	5
	Koryak	10	0.3	5
	Lahu	10	0.35	2
	Makrani	25	0.46	2
	Miaozu	10	0.4	2
	Mongolia	10	0.5	2
	Naukan	16	0.41	5
	Naxi	9	0.56	2
	Nganasan	20	0.43	5
	Oroqen	10	0.55	2
	Pathan	22	0.48	2
	Selkup	4	0.5	5
	She	10	0.35	2
	Sindhi	25	0.54	2
	Southern Han Chinese	100	0.38	3
	Tu	10	0.5	2
	Tujia	10	0.35	2
	Tuvinians	9	0.39	5
	Uygur	10	0.5	2
	Xibo	9	0.33	2
	Yakut	25	0.48	2
	Yizu	10	0.5	2
	Yukaghir	13	0.5	5
Oceania	Melanesian	18	0.64	2
	Papuan	17	0.35	2

	Population	Sample size (2n)	577X allele frequency	Reference
Americas	Arara	1	0	[5]
	Arhuaco	1	1	[5]
	Aymara	19	0.89	[5]
	Bribri	4	0.87	[5]
	Cabecar	30	0.6	[5]
	Chané	2	1	[5]
	Chilote	4	0.75	[5]
	Chipewyan	7	1	[5]
	Chono	1	1	[5]
	Colombian Amerindians	13	0.81	[2]
	Cree	2	1	[5]
	Diaguita	3	1	[5]
	Inuit	4	0.25	[5]
	Embera	5	0.8	[5]
	Guahibo	6	0.67	[5]
	Guarani	5	0.7	[5]
	Guaymi	5	0.8	[5]
	Huilliche	3	0.83	[5]
	Inga	6	1	[5]
	Jamamadi	1	0	[5]
	Kaingang	1	0	[5]
	Kaqchikel	10	0.7	[5]
	Karitiana	24	0.9	[2]
	Kogi	3	0.5	[5]
	Maleku	2	0.75	[5]
	Maya	25	0.88	[2]
	Mixe	17	0.91	[5]
	Mixtec	5	0.75	[5]
	Ojibwa	2	1	[5]
	Palikur	3	1	[5]
	Parakanã	1	0.5	[5]
	Pima	25	0.88	[2]
	Purepecha	1	1	[5]
	Quechua	38	0.89	[5]
	Surui	21	0.48	[2]
	Tepehuano	23	0.76	[5]
	Teribe	3	1	[5]
	Ticuna	6	0.5	[5]
	Toba	4	0.62	[5]
	Waunana	3	1	[5]

Population	Sample size (2n)	577X allele frequency	Reference
Wayuu	9	0.72	[5]
Wichi	5	0.8	[5]
Yaghan	2	1	[5]
Zapotec	39	0.86	[5]

[1]International HapMap Consortium (2003); [2]Li et al. (2008); [3]1000 Genomes Project Consortium et al. (2012); [4]Fattahi and Najmabadi (2012); [5]Reich et al. (2012); [6]Schlebusch et al. (2012).

PARTE III

III.I) CONCLUSÕES GERAIS

Para melhor entender a biologia de uma população, é necessário o detalhamento das forças evolutivas que moldaram a sua diversidade genética. Forças seletivas direcionais e mecanismos evolutivos randômicos, como a deriva genética, devem ser analisados de maneira separada e quanto à sua interação para que uma interpretação mais acurada possa ser realizada. Adicionalmente, no que concerne a humanos, o fator cultura não pode ser negligenciado.

O papel da deriva genética na evolução de nativos americanos já foi bastante explorado na literatura científica com respeito à diversidade genética. Os sucessivos gargalos de garrafa a que essas populações estiveram submetidas ao longo de sua história (Neel e Salzano, 1967; Fagundes *et al.,* 2008) ocasionaram uma grande divergência entre elas e, concomitantemente, uma baixa diversidade intrapopulacional (Wang *et al.,* 2007). O papel da deriva genética sobre o desequilíbrio de ligação (DL), por sua vez, foi relativamente pouco explorado. Como demonstrado na seção II.I desta tese (Amorim *et al.*, 2011), os padrões de DL também são afetados pela deriva genética, de modo que a proporção de *loci* associados pode aumentar concomitantemente à diminuição da diversidade genética. Tal fenômeno de associação entre a redução da diversidade genética e o aumento do DL deve-se ao fato de que, com a perda de alguns haplótipos, novas associações entre os alelos restantes podem surgir (Slatkin, 2008). Dessa forma, é de se esperar que em populações autóctones das Américas os níveis de DL sejam altos comparativamente a outras populações do mesmo continente, como as populações miscigenadas analisadas no mesmo estudo, e mesmo a populações autóctones de outros continentes. Assim sendo, a alta proporção de *loci* em DL, bem como sua extensão, devem ser levadas em conta para a formulação de estudos de

seleção natural (uma vez que parte da metodologia disponível para inferência de seleção positiva baseia-se na extensão do DL; ver revisão por Sabeti *et al*., 2006) e de associação (já que, com o maior tamanho dos blocos de marcadores em DL, um menor número de marcadores seria suficiente para a detecção de regiões de interesse).

Ademais, as consequências da deriva genética na distribuição alélica em populações ameríndias podem ser bastante extremas, de forma a apagar ou suavizar os sinais de seleção positiva que podem ser detectados em populações ancestrais à dos Ameríndios, como sugerimos na discussão dos resultados da análise da variante rs1815739 no gene da *ACTN3* descritos na seção II.IV (Artigo 4). Os casos dos genes *PAX9* (Paixão-Côrtes *et al*., 2011), *ABCA1* (Hünemeier *et al*., 2012a) e *KIR* (Augusto *et al*., 2013) também podem ser exemplos desse fenômeno, o qual aparentemente está emergindo como um padrão para populações ameríndias. Tal fato deve-se em parte à possibilidade de haver variações na pressão seletiva ao longo do tempo e em diferentes ambientes; aos efeitos da deriva genética, que pode fixar os diferentes alelos, quer seja vantajoso ou não; e às limitações inerentes às metodologias disponíveis.

Apesar da capacidade da deriva genética de mimetizar ou suavizar os efeitos da seleção positiva, é possível encontrar uma grande quantidade de *loci* para os quais um modelo de evolução não-neutra é o mais provável. O método estatístico utilizado na seção II.II (Artigo 2) para a detecção de sinais de seleção natural baseia-se em um método robusto que leva em consideração os efeitos da demografia sobre a diversidade genética das populações sob análise (Foll e Gaggiotti, 2008). Adicionalmente, uma abordagem específica do trabalho permitiu que se identificassem regiões que apresentam o mesmo indício de seleção positiva independentemente da população utilizada, revelando *loci* e funções biológicas que foram importantes para a adaptação

dos povos americanos à Floresta Amazônica e filtrando rigorosamente aqueles *loci* que apresentam distribuição aberrante devido às peculiaridades de uma população específica.

Naquela seção, foi possível identificar *loci* que desempenharam função adaptativa no ambiente tropical da Floresta Amazônica. Para tanto, utilizou-se a ideia de que a evolução convergente[3] é indício de adaptação a um conjunto de pressões seletivas associado a uma característica ambiental compartilhada (Losos, 2011) e foram incluídas na análise populações que vivem em um ambiente semelhante ao amazônico, *i.e.* a Floresta Tropical do Congo, de forma a identificar alterações na distribuição alélica que sugerissem adaptação a este nicho. Uma série de genes e funções biológicas associados principalmente à imunidade e metabolismo de lipídios foram destacadas. Ao povoarem as florestas do Novo Mundo, os primeiros americanos estiveram submetidos a um novo ambiente com uma diversidade patogênica elevadíssima (Guernier *et al*., 2004). Alguns genes, como *CCL28,* devem ter respondido a esta pressão seletiva permitindo com que o povoamento desse ambiente, extremamente inóspito segundo Bailey *et al.* (1989), tenha ocorrido de forma eficaz. Por outro lado, o gene *SCP2,* possivelmente relacionado à obteção de energia nutritiva – muitas vezes dificultada pelo fato de que florestas tropicais têm disponível poucos alimentos ricos em energia comestível para humanos e suas presas (Bailey *et al.,* 1989; Hart e Hart, 1986) – também apresentou sinais sugestivos de seleção positiva. Hünemeier *et al.* (2012a) descreveram um mecanismo adaptativo semelhante em populações mesoamericanas. Tanto nesse estudo quanto na presente tese, genes relacionados ao fluxo celular de colesterol apresentam sinais

[3] Evolução convergente, como define Losos (2011), é a similaridade fenotípica derivada independentemente em duas ou mais linhagens em contraposição à similaridade resultante da herança de um ancestral comum.

condizentes com uma história evolutiva não-neutra. É possível que esses genes tenham desempenhado funções que contornaram a deficiência de energia disponível e aumentado a capacidade adaptativa das populações autóctones das Américas, permitindo com que essas atingissem a distribuição atual e a ocupação dos mais variados ambientes, mesmo aqueles aparentemente mais inóspitos como a Floresta Amazônica.

O valor adaptativo do fenótipo pigmeu nas Américas foi um aspecto levantado na seção II.II que merece atenção especial. Nesta seção, encontrou-se congruência entre os resultados das análises de seleção positiva nos genomas de habitantes da floresta (Karitiana e Surui) em comparação com habitantes de uma região mais desértica no norte do México (Pima) e resultados de um estudo de associação com a altura de africanos conduzidos por Mendizabal *et al.* (2012). Uma das regiões descritas por esses autores inclui o *NNT*, um gene alvo de seleção positiva nas Américas de acordo com nossas análises. Essa congruência de resultados sugere que o fenótipo pigmeu ou alguma característica associada a este pode apresentar alto valor adaptativo em outras regiões do mundo que não a África. Indivíduos de baixa estatura podem, por exemplo, apresentar vantagens para lidar com a limitação de recursos alimentícios, com a termorregulação e com dificuldade de movimentação (Perry e Dominy, 2009). A importância dessa característica no povoamento das Américas e mais especificamente da Floresta Amazônica merece mais atenção em investigações subsequentes.

É provável que novos estudos como esse, utilizando diferentes populações expostas a diversas pressões seletivas das Américas, revelem novos *loci* de interesse e tornem mais claro o cenário evolutivo do povoamento do continente, esclarecendo de

que forma os primeiros americanos lidaram com a imensa variedade de ambientes a que foram expostos durante o povoamento pré-histórico do Novo Mundo.

Um caso curioso é o da ACNT3*R577X, uma variante protéica associada ao desempenho muscular de atletas de elite. Yang *et al.* (2003) demonstraram que o alelo derivado ocorre significativamente menos em atletas de velocidade. Posteriormente, foi sugerido que este alelo poderia desempenhar vantagem para o desempenho de atividades de resistência e que isso seria a causa dos sinais de seleção positiva sobre esse gene em populações europeias e asiáticas (MacArthur *et al.,* 2007; Friedlander *et al.,* 2013). Entre outras características, uma das possíveis vantagens da capacidade para desempenho de esforço muscular durante longos períodos é a possibilidade de emergência de modalidades de caça alternativas como o acesso facilitado a carcaças e a caça por exaustão da presa (Bramble e Lieberman, 2004), esta última é uma forma de caça típica dos Tarahumara (Balke e Snow, 1965), uma etnia habitante do noroeste mexicano. É possível que, além dessa, outras características tenham sido afetadas pela diversidade do *ACTN3*, entre elas algo poderia estar relacionado com a rápida dispersão dos primeiros americanos no Novo Mundo. Apesar de não ter sido identificado sinal de seleção positiva neste gene nas Américas, é possível que o polimorfismo no *ACNT3* tenha, de alguma forma, exercido alguma influência sobre esse aspecto da história americana.

De fato, o que torna o povoamento das Américas um episódio peculiar na história da humanidade é a grande rapidez com que uma vasta diversidade de ambientes ainda não povoados foi ocupada. Durante esse longo percurso, os primeiros americanos se depararam com diversas pressões seletivas e apesar de uma diversidade genética reduzida em decorrência dos sucessivos gargalos populacionais a que foram

submetidos, foram capazes de se estabelecer por quase toda a extensão do continente. A cultura, nesse caso, deve ter desempenhado um papel importante, quer seja na modificação do valor adaptativo de determinados alelos pela construção de nichos (Hünemeier *et al*., 2012a), quer seja pela aceleração da evolução populacional (Hünemeier *et al*., 2012b). Compreender com que velocidade a cultura de nativos americanos evolui em comparação com os genes pode trazer informações mais detalhadas sobre a história dos ameríndios. Um estudo comparando as taxas de evolução genética e cultural evidenciou empiricamente que as taxas evolutivas são maiores para a cultura (Perreault, 2012), o que pode estar relacionado, por exemplo, com a sua forma de transmissão horizontal, isto é, entre indivíduos não aparentados (Reali e Griffiths, 2010). Entretanto, a inclusão de datas para a fissão de populações sul-americanas dentro do intervalo estimado para eventos de fissão linguística correspondentes indica que, ao menos no que concerne aos grupos linguísticos analisados, língua e genes têm taxas evolutivas semelhantes (Seção II.III – Amorim *et al.*, 2013). A existência de limitações na evolução de determinados traços culturais, mas não em outros, é uma possibilidade a ser investigada. É possível ainda que em alguns casos as línguas apresentem maiores limitações para dispersão que genes (Hunley e Long, 2005), o que ocasionaria, por exemplo, uma maior taxa evolutiva para os genes em comparação com as línguas.

Estudos como os explorados na presente tese contribuirão para o entendimento da biologia de populações nativas americanas em especial no que concerne à dinâmica populacional e evolução pós-povoamento. Novas perspectivas devem surgir na medida em que novos genomas e exomas de ameríndios forem sequenciados. Para tanto, será necessária a utilização de uma perspectiva integradora, que leve em conta as

peculiaridades da história demográfica dessas populações, bem como a existência de respostas às pressões seletivas, e que não descarte a cultura como um elemento evolutivo essencial. Na medida em que novos estudos de seleção natural forem realizados, outros *loci* devem se destacar como alvos de seleção positiva, como foi o caso do *ABCA1* (Hünemeier *et al.*,, 2012a)*, SCP2* e *CCL28* entre outros. Apesar disso, exemplos como o dos genes *PAX9* (Paixão-Côrtes et al., 2011), *KIR* (Augusto *et al*., 2013) e *ACTN3* devem ser os mais comuns, já que a história evolutiva dessas populações deve ter sido em grande parte influenciada pela deriva genética**.**

III.II) REFERÊNCIAS BIBLIOGRÁFICAS

Amorim CEG, Wang S, Marrero AR, Salzano FM, Ruiz-Linares A, Bortolini MC (2011) X-chromosomal genetic diversity and linkage disequilibrium patterns in Amerindians and non-Amerindian populations. *Am J Hum Biol* 23: 299-304.

Amorim CEG, Bisso-Machado R, Ramallo V *et al.* (2013) A Bayesian Approach to Genome/Linguistic Relationships in Native South Americans. *PLoS One* 8: e64099.

Augusto DG, Piovezan BZ, Tsuneto LT, Callegari-Jacques SM, Petzl-Erler ML (2013) KIR gene content in amerindians indicates influence of demographic factors. *PLoS One* 8: e56755.

Bailey RC, Head G, Jenike M, Owen B Rechtman R, Zechnter E (1989) Hunting and gathering in tropical rain forest: is it possible? *Am Anthropologist* 91: 59-82.

Balke B, Snow C (1965) Anthropological and physiological observations on Tarahumara endurance runners. *Am J Phys Anthropol* 23: 293-301.

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol Evol* 20: 435-40.

Bigham AW, Mao X, Mei R *et al.* (2009) Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Hum Genomics* 4: 79-90.

Bortolini MC, Salzano FM, Thomas MG et al. (2003) Y-chromosome evidence for differing ancient demographic histories in the Americas. *Am J Hum Genet* 73: 524-39

Bramble DM, Lieberman DE (2004) Endurance running and the evolution of *Homo*. *Nature* 432: 345-52.

Callegari-Jacques SM, Tarazona-Santos EM, Gilman RH *et al.* (2011) Autosome STRs in native South America - testing models of association with geography and language. *Am J Phys Anthropol* 145: 371-81.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185:1411-23.

Daub JT, Hofer T, Cutivet E *et al.* (2013) Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Mol Biol Evol* [Artigo no prelo com publicação electrônica antecipada].

Dillehay TD, Ramírez C, Pino M, Collins MB, Rossen J, Pino-Navarro JD (2008) Monte Verde: seaweed, food, medicine, and the peopling of South America. *Science* 320: 784-6.

Dillehay TD (2009) Probing deeper into First American studies. *Proc Natl Acad Sci U S A* 106: 971-8.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol Evol* 23: 347-51.

Fagundes NJ, Bonatto SL, Callegari-Jacques SM, Salzano FM (2002) Genetic, geographic, and linguistic variation among South American Indians: possible sex influence. *Am J Phys Anthropol* 117: 68-78.

Fagundes NJ, Kanitz R, Eckert R *et al.* (2008) Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet* 82: 583-92.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180: 977-93.

Friedlander SM, Herrmann AL, Lowry DP *et al.* (2013) ACTN3 allele frequency in humans covaries with global latitudinal gradient. *PLoS One* 8: e52282.

Grossman SR, Andersen KG, Shlyakhter I *et al*. (2013) Identifying recent adaptations in large-scale genomic data. *Cell* 152: 703-13.

Guernier V, Hochberg ME, Guégan JF (2004) Ecology drives the worldwide distribution of human diseases. *PLoS Biol* 2: e141.

Hart TB, Hart JA (1986) The ecological basis of hunter-gatherer subsistence in African rain forests: the Mbuti of Eastern Zaire. *Hum Ecol* 14. 29-55.

Hernandez RD, Kelley JL, Elyashiv E *et al.* (2011) Classic selective sweeps were rare in recent human evolution. *Science* 331: 920-4.

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection*. Ann Hum Genet* 73: 95-108.

Hughes AL (2012) Evolution of adaptive phenotypic traits without positive Darwinian selection. *Heredity* 108: 347-53.

Hünemeier T, Amorim CEG, Azevedo S *et al.* (2012a) Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. *PLoS One 7*: e38862.

Hünemeier T, Gómez-Valdés J, Ballesteros-Romero M *et al.* (2012b) Cultural diversification promotes rapid phenotypic evolution in Xavánte Indians. *Proc Natl Acad Sci U S A* 109: 73-7.

Hunley K, Long JC (2005) Gene flow across linguistic boundaries in Native North American populations. *Proc Natl Acad Sci USA* 102: 1312-7.

Kimura M, Ota T (1974) On Some Principles Governing Molecular Evolution. *Proc Natl Acad Sci USA* 71: 2848-52.

Kitchen A, Miyamoto MM, Mulligan CJ (2008) A three-stage colonization model for the peopling of the Americas. *PLoS One* 3: e1596.

Laland KN, Brown G (2002) *Sense and nonsense: evolutionary perspectives on human behaviour*. Oxford: Oxford University Press, pp. 264.

López Herráez D, Bauchet M, Tang K *et al.* (2009) Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* 4: e7888.

Losos JB (2011) Convergence, adaptation, and constraint. *Evolution* 65:1827-40.

MacArthur DG, Seto JT, Raftery JM *et al.* (2007) Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nat Genet* 39: 1261-5.

Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African Pygmy phenotype. *Hum Genet* 131: 1305-17.

Mellars P (2006) Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc Natl Acad Sci U S A* 103:9381-6.

Neel JV, Salzano FM (1967) Further studies on the Xavante Indians. X. Some hypotheses-generalizations resulting from these studies. *Am J Hum Genet* 19: 554-74.

Paixão-Côrtes VR, Meyer D, Pereira TV *et al* (2011) Genetic variation among major human geographic groups supports a peculiar evolutionary trend in PAX9. *PLoS One* 6: e15656.

Perreault C (2012) The pace of cultural evolution. *PLoS One* 7: e45150.

Perry GH, Dominy NJ (2009) Evolution of the human pygmy phenotype. *Trends Ecol Evol* 24: 218-25.

Ramalho RF, Santos EJ, Guerreiro JF, Meyer D (2010) Balanced polymorphism in bottlenecked populations: the case of the CCR5 5' cis-regulatory region in Amazonian Amerindians. *Hum Immunol* 71: 922-8.

Reali F, Griffiths TL (2010) Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. *Proc R Soc B* 277: 429-36.

Richerson PJ, Boyd R (2005) *Not by genes alone: how culture transformed human evolution.* Chicago: University of Chicago Press, pp. 342.

Sabeti PC, Reich DE, Higgins JM *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832-7.

Sabeti PC, Schaffner SF, Fry B *et al.* (2006) Positive natural selection in the human lineage. *Science* 312: 1614-20.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913-8.

Salzano FM (2007) The prehistoric colonization of the Americas. *In*: Crawford MH (Ed.) *Anthropological genetics. theory, methods and applications*. Cambridge: Cambridge University Press, p. 433–55.

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9: 477–85.

Scliar MO, Soares-Souza GB, Chevitarese J *et al.* (2012) The population genetics of Quechuas, the largest native South American group: autosomal sequences, SNPs, and microsatellites evidence high level of diversity. *Am J Phys Anthropol* 147: 443-51.

Tarazona-Santos E, Castilho L, Amaral DR *et al.* (2011) Population genetics of GYPB and association study between GYPB*S/s polymorphism and susceptibility to P. falciparum infection in the Brazilian Amazon. *PLoS One* 6: e16123.

Tovo-Rodrigues L, Callegari-Jacques SM, Petzl-Erler ML, Tsuneto L, Salzano FM, Hutz MH (2010) Dopamine receptor D4 allele distribution in Amerindians: a reflection of past behavior differences? *Am J Phys Anthropol* 143: 458-64.

Wang S, Lewis CM, Jakobsson M *et al.* (2007) Genetic variation and population structure in Native Americans. *PLoS Genet* 3: e185.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet* 3: e90.

Yang N, MacArthur DG, Gulbin JP *et al* (2003) ACTN3 genotype is associated with human elite athletic performance. *Am J Hum Genet* 73: 627-31.

Yokoyama Y, Lambeck K, Deckker PD, Johnston P, Fifield LK (2000) Timing of the last glacial maximum from observed sea-level minima. *Nature* 406: 713–6.

APÊNDICE

Hünemeier T, Amorim CEG, Azevedo S, Contini V, Acuña-Alonzo V, Rothhammer F, Dugoujon JM, Mazières S, Barrantes R, Villarreal-Molina MT, Paixão-Côrtes VR, Salzano FM, Canizales-Quinteros S, Ruiz-Linares A, Bortolini MC (2012) *Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. PLoS One* 7: e38862.

OPEN ACCESS Freely available online

PLoS one

Evolutionary Responses to a Constructed Niche: Ancient Mesoamericans as a Model of Gene-Culture Coevolution

Tábita Hünemeier[1], Carlos Eduardo Guerra Amorim[1], Soledad Azevedo[2], Veronica Contini[1], Víctor Acuña-Alonzo[3], Francisco Rothhammer[4,5], Jean-Michel Dugoujon[6], Stephane Mazières[7], Ramiro Barrantes[8], María Teresa Villarreal-Molina[9], Vanessa Rodrigues Paixão-Côrtes[1], Francisco M. Salzano[1], Samuel Canizales-Quinteros[10,11], Andres Ruiz-Linares[12], Maria Cátira Bortolini[1]*

1 Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil, **2** Centro Nacional Patagónico, CONICET, U9120ACV, Puerto Madryn, Argentina, **3** Molecular Genetics Laboratory, Escuela Nacional de Antropología e Historia, Mexico City, Mexico, **4** Programa de Genética Humana, Instituto de Ciencias Biomédicas, Facultad de Medicina, Universidad de Chile, Santiago, Chile, **5** Instituto de Alta Investigación, Universidad de Tarapacá, Arica, Chile, **6** Laboratoire d'Anthropologie Moléculaire et d'Imagerie de Synthèse, UMR 5288 CNRS, Université Paul Sabatier (Toulouse3), Toulouse, France, **7** Anthropologie Bio-culturelle, Droit, Ethique et Santé (ADES), UMR 7268, Aix-Marseille-Université/CNRS/EFS, Marseille, France, **8** Escuela de Biología, Universidad de Costa Rica, San José, Costa Rica, **9** Laboratorio de Genómica de Enfermedades Cardiovasculares, Instituto Nacional de Medicina Genómica, Mexico City, Mexico, **10** Unit of Molecular Biology and Genomic Medicine, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico, **11** Departamento de Biología, Facultad de Química, Universidad Nacional Autónoma de México, Mexico City, Mexico, **12** The Galton Laboratory, Department of Biology, University College London, London, United Kingdom

Abstract

Culture and genetics rely on two distinct but not isolated transmission systems. Cultural processes may change the human selective environment and thereby affect which individuals survive and reproduce. Here, we evaluated whether the modes of subsistence in Native American populations and the frequencies of the *ABCA1*Arg230Cys* polymorphism were correlated. Further, we examined whether the evolutionary consequences of the agriculturally constructed niche in Mesoamerica could be considered as a gene-culture coevolution model. For this purpose, we genotyped 229 individuals affiliated with 19 Native American populations and added data for 41 other Native American groups (n = 1905) to the analysis. In combination with the SNP cluster of a neutral region, this dataset was then used to unravel the scenario involved in *230Cys* evolutionary history. The estimated age of *230Cys* is compatible with its origin occurring in the American continent. The correlation of its frequencies with the archeological data on *Zea* pollen in Mesoamerica/Central America, the neutral coalescent simulations, and the F_{ST}-based natural selection analysis suggest that maize domestication was the driving force in the increase in the frequencies of *230Cys* in this region. These results may represent the first example of a gene-culture coevolution involving an autochthonous American allele.

Citation: Hünemeier T, Amorim CEG, Azevedo S, Contini V, Acuña-Alonzo V, et al. (2012) Evolutionary Responses to a Constructed Niche: Ancient Mesoamericans as a Model of Gene-Culture Coevolution. PLoS ONE 7(6): e38862. doi:10.1371/journal.pone.0038862

Editor: Toomas Kivisild, University of Cambridge, United Kingdom

Received February 8, 2012; **Accepted** May 12, 2012; **Published** June 21, 2012

Funding: This study was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), PRONEX (Brazil); Programme Interdisciplinaire CNRS: Amazonie-Analyse, Modélisation et Ingénierie des Systèmes Amazoniens (France); and Convenio de Desempeño Mecesup 2/UTA (Chile). These funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: maria.bortolini@ufrgs.br

Introduction

Human cultural practices have drastically modified environmental conditions and behaviors, promoting rapid and substantial genomic changes often associated with positive selection and adaptation (gene-culture dynamics [1,2]). In the history of *Homo sapiens sapiens*, a particularly important event that triggered a new and striking gene-culture-coevolution cycle was the development of agriculture and animal domestication during the Neolithic period (~10,000 years ago). Further, the human gene-culture coevolution mediated by the domestication of plants and animals has been argued to provide some of the clearest and most spectacular examples of niche construction. The Niche Construction Theory can be defined as a branch of evolutionary biology that emphasizes on the ability of organisms to modify the pressure of natural selection in their environment and thereby act as co-directors of their own evolution, as well as that of other directly associated species [3–6].

Although more than 100 regions/genes had been identified as the likely targets of recent positive selection resulting from cultural pressures in newly constructed niches [1], well-documented examples are scarce. One of the best-known cases of gene-culture coevolution is lactase persistence (LP; the ability of adult humans to digest the lactose found in fresh milk) and dairying. High frequencies of LP are generally observed in traditional pastoralist populations. For example, LP reaches ~64% in Beni Amir pastoralists from Sudan, whereas its frequency in a neighboring non-pastoralist community is only ~20%. In Europe, LP varies from 15–54% in eastern and southern regions, 62–86% in central and western regions, and 89–96% in northern regions [5,7–13].

Multiple independent mutations have been associated with this characteristic, some of which are located in an intron of the *MCM6* gene, a region fundamental to lactase expression [5,14]. The alleles that led to lactose persistence in Europe, such as *MCM6 13,910*T*, first underwent selection among dairying farmers around 7,500 years ago, possibly in association with the dissemination of the Neolithic Linearbandkeramik culture over Central Europe [13]. The high copy number variation of the amylase gene and the spread of the corresponding alleles in agricultural societies are another well-studied example [7,8,11,15,16]. Additionally, the West African Kwa-speaking agriculturalists cut and clear the forest to grow yams, increasing the amount of standing water after rain, therefore providing better breeding grounds for malaria-carrying mosquitoes [1] favoring the *HbS* allele, which confers protection against malaria in heterozygous individuals [17].

America was the last continent colonized by modern humans in prehistoric times. In less than 15,000 years before present (YBP), these first migrants had to adapt to an immensely wide variety of environments. In some regions during this evolutionary trajectory, as in Mesoamerica and the Andes, hunter-gatherer/forager societies gave rise to agriculturalist and urban communities, while others remained with a hunter-gatherer/forager subsistence system until the time of contact with Europeans or even until the present day. Thus, studies with Native American populations can provide useful information for better understanding gene-culture coevolution and the niche construction processes.

Based on studies with blood groups and other classical genetic polymorphisms, J. V. Neel and F. M. Salzano were pioneers in identifying complex population processes highly dependent on cultural factors in Native Americans (e.g., fission-fusion dynamic; [18]). Other examples are related to the coevolution of genes and languages [19,20], but only two more recently reported examples might be associated with positive selection: (1) Tovo-Rodrigues *et al.* [21] investigated the distribution of D4 dopamine receptor (*DRD4*) alleles in several South Amerindian populations and found a significant difference in the allelic distributions between hunter-gatherers and agriculturalists, with an increase of the *7R* allele among the former; and (2) Acuña-Alonzo *et al.* [22] showed that the *230Cys* allele (*Arg230Cys*, rs9282541) of the ATP-binding cassette transporter A1 (*ABCA1*) gene, which was previously associated with low HDL-cholesterol levels and obesity-related comorbidities, was exclusively present in Native American and mestizo individuals. These authors verified that cells expressing the *ABCA1*230Cys* allele showed a 27% cholesterol efflux reduction, confirming that this Native American autochthonous variant has a functional effect *in vitro*. Other investigations have shown that the presence of *ABCA1*230Cys* explains almost 4% of the variation in plasma HDL-C concentrations in Mexican admixed populations [23]. This variation in HDL-C concentration was the highest one associated with a single nucleotide polymorphism (SNP) among different continental populations in these genome-wide association studies, corroborating its functionality [23].

Acuña-Alonzo *et al.* [22] demonstrated that *230Cys* resides in a haplotype that was the target of an ongoing directional selective sweep, suggesting that *230Cys* conferred an advantage during periods of food deprivation in the past. On the other hand, under the current modern lifestyle, *230Cys* may have become a major susceptibility allele for low HDL levels and has been correlated with metabolic diseases [22]. This study provides an example of the "thrifty" genotype hypothesis, which postulates that variants that increase the efficiency of energy use and storage during periods of famine would have been positively selected in prehistoric times but can be associated with diseases of affluence in contemporary societies, where food is usually abundant [24].

Here, we expand the investigations of the *Arg230Cys* polymorphism in Native Americans and integrate the thrifty genotype concept with the gene-culture coevolution process, considering the human ability to create new ecological niches that may lead to the selection of genetic variants.

Materials and Methods

(a) Populations

New data for the *Arg230Cys* polymorphism were generated for 19 Amerindian populations (n = 229) from Meso/Central America and South America (Table 1). Additional information about these tribes can be found in Bortolini *et al.* [25,26], Wang *et al.* [27], and Mazières *et al.* [28,29]. These new *Arg230Cys* data were then analyzed together with those of an earlier published report [22], providing a total of 1905 investigated individuals. One hundred and twenty-six individuals of our sample were also previously genotyped for ~680,000 SNPs using Illumina Human 610-Quad BeadChips(Ruiz-Linares *et al.*, unpublished data), and a part of this information was used in some of our analyses. The populations were clustered according to their geographical location and ancient mode of subsistence as follows: (1) Mesoamerican agriculturalists, (2) Andean agriculturalists, and (3) South American hunter-gatherers/foragers. Of course, caution is needed regarding this classification, since subsistence modes are not stable over time and may not be unique. However, the two categories adopted here (agriculturalists and hunter-gatherers/foragers) represent general pre-Colombian subsistence conditions, providing a starting point for research related to gene-culture dynamics in Native Americans.

Ethical approval for the present study was provided by the Brazilian National Ethics Commission (CONEP Resolution no. 123/98) for the Brazilian samples, as well as by ethics committee of: (a) Universidad de Antioquia, Medellin, Colombia (Colombian samples); (b) Universidad Nacional Autónoma de México, Ciudad de México, México (Mexican samples); (c) Universidad de Costa Rica, San José, Costa Rica (Costa Rican and Panamanian samples); (d) Universidad of Chile, Santiago, Chile (Chilean samples); (e) Université Paul Sabatier Toulouse 3, Toulouse, France (Bolivian and French Guianian samples). Individual and tribal informed oral consent was obtained from all participants, since they were illiterate, and they were obtained according to the Helsinki Declaration. The ethics committees approved the oral consent procedure as well as the use of these samples in population and evolutionary studies.

(b) SNP Genotyping and Intra- and Inter-subdivision Structures

The *Arg230Cys* polymorphism was genotyped using TaqMan assays (ABI Prism 7900HT Sequence Detection System; Applied Biosystems). Allele frequencies were obtained by direct counting. The level of the population structure observed within and between Mesoamerican agriculturalist, Andean agriculturalist, and South American hunter-gatherer/forager groups was estimated using F statistics and the Arlequin 3.5.1 software [30]. Allele frequencies were compared between the 3 population subdivisions with the student's t-test (alfa = 0.05) using the R Stats package (R Development Core Team; http://www.R-project.org/).

(c) Allele Age and Neutrality/selection Tests

A large Asian and Native American sample, including the 126 individuals investigated here, were genotyped for a major panel of

Table 1. Genotypes, allele frequencies, and geographic locations of the Native American populations investigated.

Population	N^1	Genotype frequency			Allele frequency		Country	Geographical coordinates	References
		Arg 230 Arg	*Arg 230 Cys*	*Cys 230 Cys*	*Arg230*	*230Cys*			
Mesoamerican agriculturalist[2] (1218)									
Yaqui	45	30	11	4	0.79	0.21	Mexico	27° 29′ N 110° 40′ W	Acuña-Alonzo et al. (2010)
Tarahumara	109	81	23	5	0.85	0.15	Mexico	26° 49′ N 107° 04′ W	Acuña-Alonzo et al. (2010)
Teenek	67	45	20	2	0.82	0.18	Mexico	21° 36′ N 98° 58′ W	Acuña-Alonzo et al. (2010)
Cora	123	62	51	10	0.71	0.29	Mexico	22° 3′ N 104° 55′ W	Acuña-Alonzo et al. (2010)
Purepecha	35	22	11	2	0.79	0.21	Mexico	19° 36′ N 102° 14′ W	Acuña-Alonzo et al. (2010)
Mazahua	83	68	15	0	0.91	0.09	Mexico	19° 26′ N 100° 00′ W	Acuña-Alonzo et al. (2010)
Mixe	19	15	4	0	0.89	0.11	Mexico	17° N 96°W	Present study
Mixtec	4	4	0	0	1.00	0.00	Mexico	17° N 97°W	Present study
Nahuatl	267	185	73	9	0.83	0.17	Mexico	19° 58′ N 97° 37′ W	Acuña-Alonzo et al. (2010)
Totonaco	113	86	24	3	0.87	0.13	Mexico	19° 57′ N 97° 44′ W	Acuña-Alonzo et al. (2010)
Otomíes	42	35	7	0	0.92	0.08	Mexico	20° 28′ N 99° 13′ W	Acuña-Alonzo et al. (2010)
Zapotec	125	71	50	4	0.76	0.24	Mexico	17°14′ N 96°14′ W	Present study; Acuña-Alonzo et al. (2010)
Mayan	110	68	39	3	0.80	0.20	Mexico	20°13′ N 90°28′ W	Acuña-Alonzo et al. (2010)
Kaqchikel-Quiche	17	13	3	1	0.85	0.15	Guatemala	15° N 91°W	Present study
Cabecar	24	19	5	0	0.90	0.10	Costa Rica	9° 30′ N 84°W	Present study
Guaymí	35	26	8	1	0.85	0.15	Costa Rica/ Panamá	8°30′ N 82° W	Present study
South American hunter-gatherer/forager[2] (572)									
Parkatejê (Gavião)	78	65	12	1	0.91	0.09	Brazil	05° 03′ S 48° 36′ W	Acuña-Alonzo et al. (2010)
Jamamadi	26	26	0	0	1.00	0.00	Brazil	07° 15′ S 66° 41′ W	Acuña-Alonzo et al. (2010)
Mekranoti (Kayapó)	25	24	1	0	0.98	0.02	Brazil	08° 40′ S 54° W	Acuña-Alonzo et al. (2010)
Mura (Pirahã)	18	11	6	1	0.78	0.22	Brazil	03°34′ S 59° 12′ W	Acuña-Alonzo et al. (2010)
Pacaás-Novos (Wari)	25	23	2	0	0.96	0.04	Brazil	11° 08′ S 65° 05′ W	Acuña-Alonzo et al. (2010)
Sateré-Mawé	25	20	4	1	0.88	0.12	Brazil	03° S 57° W	Acuña-Alonzo et al. (2010)
Apalaí	22	15	7	0	0.84	0.16	Brazil	01°20′ N 54°40′ W	Acuña-Alonzo et al. (2010)
Arara	24	15	9	0	0.81	0.19	Brazil	03° 30′ S 54°10′ W	Acuña-Alonzo et al. (2010)
Guarani	31	30	1	0	0.98	0.02	Brazil	25° 20′ S 52° 30′ W	Present study; Acuña-Alonzo et al. (2010)
Gorotire (Kayapo)	7	6	0	1	0.86	0.14	Brazil	07° 44′ S 51° 10′ W	Acuña-Alonzo et al. (2010)
Karitiana	20	20	0	0	1.00	0.00	Brazil	08° 45′ S 63° 51′ W	Acuña-Alonzo et al. (2010)
Xavante	21	10	9	2	0.69	0.31	Brazil	13° 20′ S 51° 40′ W	Acuña-Alonzo et al. (2010)
Xikrin (Kayapo)	17	16	1	0	0.97	0.03	Brazil	05°55′ S 51°11′ W	Acuña-Alonzo et al. (2010)
Yanomama	25	20	4	1	0.88	0.12	Brazil	02°30 ′–04°30′ N 64° W	Acuña-Alonzo et al. (2010)
Txukahamae (Kayapo)	30	26	4	0	0.93	0.07	Brazil	10° 20′ S 53° 5′ W	Acuña-Alonzo et al. (2010)
Tiriyó (Trio)	25	21	4	0	0.92	0.08	Brazil	01° 57′ N 55°49′ W	Acuña-Alonzo et al. (2010)
Içana River (Baniwa)	19	13	3	3	0.76	0.24	Brazil	01° N 67° 50′ W	Acuña-Alonzo et al. (2010)
Kuben Kran Keng (Kayapo)	17	13	4	0	0.88	0.12	Brazil	08°10′ S 58°8′ W	Acuña-Alonzo et al. (2010)
Lengua	29	29	0	0	1.00	0.00	Paraguay	23° S 56° W	Acuña-Alonzo et al. (2010)
Ache (Guayaki)	23	23	0	0	1.00	0.00	Paraguay	23° S 58° W	Acuña-Alonzo et al. (2010)
Ayoreo	30	30	0	0	1.00	0.00	Paraguay	16–22° S 58–63° W	Acuña-Alonzo et al. (2010)
Zenu	4	4	0	0	1.00	0.00	Colombia	9° N 75° W	Present study
Kogi	7	7	0	0	1.00	0.00	Colombia	11° N 74° W	Present study
Ticuna	1	1	0	0	1.00	0.00	Colombia	3° 53′ S 70°W	Present study
Embera	3	3	0	0	1.00	0.00	Colombia	7° N 76° W	Present study
Wayuu	17	15	2	0	0.94	0.06	Colombia	11° N 73° W	Present study
Palikur	3	1	2	0	0.67	0.33	French Guiana	4° N 51° 45′ W	Present study

Table 1. Cont.

Population	N[1]	Genotype frequency			Allele frequency		Country	Geographical coordinates	References
		Arg 230 Arg	*Arg 230 Cys*	*Cys 230 Cys*	*Arg230*	*230Cys*			
Andean agriculturalist[2] (115)									
Mapuche	40	40	0	0	1.00	0.00	Chile	40° 30′ S 69° 20′ W	Acuña-Alonzo et al. (2010)
Aymara	16	16	0	0	1.00	0.00	Bolivia	16°30′ S 68°9′ W	Present study
Quechua	16	15	1	0	0.97	0.03	Bolivia	14°30′ S 69° W	Present study
Aymara	22	20	2	0	0.95	0.05	Chile	22° S 70° W	Present study
Chilote	2	2	0	0	1.00	0.00	Chile	42°30′ S 73°55′ W	Present study
Hulliche	13	10	3	0	0.89	0.11	Chile	41° S 73° W	Present study
Ingano	6	5	1	0	0.92	0.08	Colombia	1° N 77° W	Present study

[1]Samples genotyped in present study = 229;
[2]Caution is needed regarding the classification of these modes of subsistence, since they are not stable over time and may not be unique. However, the two categories adopted here (agriculturalist and hunter-gatherer/forager) represent general pre-Columbian subsistence conditions of the investigated populations in accordance with what is known about them. AMOVA results: (a) Among the subdivisions (F_{CT}): 3.6% ($p = 0.000$); (b) Among populations within the Mesoamerican Agriculturalist subdivision (F_{ST}): 1.8% ($p = 0.008$); (c) Among populations within the South American hunter-gatherer/forager subdivision: 5.3% ($p = 0.005$); Among populations within the Andean Agriculturarist group: 0% ($p = 0.36$).
doi:10.1371/journal.pone.0038862.t001

~680,000 SNPs (Ruiz-Linares *et al.*, unpublished data). Based on this additional information (Tables S1, S2 and S3), the following analyses were performed:

(c.1) *ABCA1*230Cys* allele age. Since estimates of allele age depend on assumptions about demographic history and natural selection, we have performed two approaches to estimate the age of the variant allele:

(1) Kimura and Ohta [31] were the first to consider the relations between allele age and its frequency. With this purpose they developed the equation $E(t1) = [-2p/(1-p)]\ln(p)$, where $E(t1)$ = expected age, time is measured in units of $2N$ generations, and p = population frequency [31]. For *ABCA1*230Cys*, we considered the average of frequencies of all populations and only those from Mesoamerica/Central America ($p = 9.6$ and 15.4, respectively; Table 1). A generation time of 25 years and $N = 720$ (number of generation considering the upper limit for the peopling of America, 18,000 YBP [32]) were assumed.

(2) Slatkin and Rannala [33] began to exploit Linkage Disequilibrium (LD) to estimate allele ages, based on variation among different copies of the same allele, where the age of an allele is estimated by the intra-allelic variation following the LD exponential decay due to recombination and mutation rates. Rannala and Reeve [34], on the other hand, explored the use of LD to map genes, as well as to obtain the allele age using a Markov Chain Monte Carlo framework. We applied this method to obtain a second *ABCA1*230Cys* age estimative using the DMLE+ v2.2 software (http://www.dmle.org). This program allows a Bayesian inference of the mutation age using an intra-allelic coalescent model to assess LD across the nineteen SNPs that occur around *ABCA1*230* (rs2065412, rs2515601, rs2472386, rs2274873, rs2487054, rs4149290, rs2487039, rs2472384, rs2253174, rs2230806, rs2230805, rs2249891, rs4149281, rs4743764, rs1929841, rs2000069, rs2275542, rs3904998, and rs4149268). Taken into account the historical information about our sampled populations, three parameters were introduced: (a) Generation time of 25 years; (b) Proportion of population growth of 0.005; and (c) Proportion of population sampled of 0.0002. The program was run in the haplotype mode using two million of iterations.

(c.2) Test to detect deviations from neutrality. Based on the long-range haplotype test [35] and integrated haplotype scores [9,36] Acuña-Alonzo et al. [22] suggested that the autochthonous Native American ABCA1*230Cys allele could have been positively selected. However, demographic events, population structure, and other stochastic processes can create complex patterns in the genome, obscuring signals of natural selection or mimicking adaptive processes [37]. Additionally, positive and balancing selections show different effects on the genetic diversity patterns within and between populations [38]. Therefore, we performed additional analyses to explore these issues and elucidate the factors responsible for the eventual effect of natural selection on the ABCA1*230 locus.

To detect loci under selection, we used a method that contrasted the observed population differentiation (F_{ST}) with that generated for a null simulated distribution under a hierarchical island model using a coalescent approach. In this model, demes exchange more migrants within groups than between groups to generate the joint distribution of genetic diversity within and between populations [38]. Thus, a p value can be estimated from the joint distribution for the population heterozygosity (H_e) and F_{ST} using a kernel density estimation procedure [30]. The analysis was performed using Arlequin 3.5.1 in consideration of 126 Native Americans whose results for the *ABCA1*230* locus were known. Data from 20 other autosomal SNPs (rs6559725, rs11140096, rs4877767, rs4014024, rs11140109, rs7872891, rs7850633, rs17086298, rs10746709, rs5014093, rs10868019, rs11140116, rs3860938, rs3860941, rs4097644, rs9942844, rs12551103, rs7863524, rs4877785, and rs70439590) from these same individuals were compiled from a major SNP panel (Table S2). These 20 additional SNPs were selected based on their location (chromosome 9: from position 85252250 to 85317359) inside the putative neutral region, defined by Schroeder and colleagues [39], which comprises ~76,000 bp around the *D9S1120* locus. Using this database (Table S2 and Figure S1), we were able to evaluate whether the joint distribution of the observed H_e and F_{ST} for the *ABCA1*230*

polymorphism data departed from the expected outcome at neutrality. Fifty thousand coalescent simulations were performed with a 100-demes island model. Four comparative analyses were conducted: (1) Mesoamerican agriculturalists (n = 68) *vs.* Andean agriculturalists (n = 35), (2) Mesoamerican agriculturalists *vs.* South American hunter-gatherers/foragers (n = 23), (3) Mesoamerican agriculturalists *vs.* South Americans (agriculturalists + hunter-gatherers/foragers), and (4) Andean agriculturalists *vs.* South American hunter-gatherer/foragers.

In addition, we simulated 100,000 neutral genealogies for a region containing two distinct sets of 20 biallelic markers under demographic scenarios mimicking the settlement of the Americas [32,40]. To accomplish these simulations, we used the msABC software [41], a modification of ms software [42] that uses the coalescent to generate samples under a neutral Wright-Fisher model. The demographic parameters included: (1) a current effective population size of 830 individuals [40], a number not very different from that used in the simulations of Schroeder *et al.* [39]; (2) three demes that corresponded to the sampled subdivisions (Mesoamerican agriculturalists, Andean agriculturalists, South American hunter-gatherers/foragers); (3) a single ancestral population that existed from 6,350 to 18,000 YBP [32,40]; and (4) different exponential growth rates to include the possibility of an ancestral population of 70 to 830 individuals (i.e., constant population size [40]).

The genetic diversity obtained in the simulations–summarized by intra- (heterozygosity) and interpopulation (global and pairwise F_{ST}) statistics–was compared to the observed genetic diversity of two genetic datasets: (1) that of the *ABCA1*230* locus plus the same 19 flanking SNPs listed in item c.1; (Table S3), and (2) the same 20 SNPs mentioned in section c.2 (Table S2; Figure S1). The summary statistics calculated for each simulation (S vectors) were then compared to the summary statistics of the observed data (S* vectors) using an Euclidean distance measure $\partial = ||S\text{-}S^*||$ with the ABCestimator software, implemented with the ABCtoolbox [43]. The rationale of the analysis was to check which observed dataset could be reproduced with higher fidelity among the range of neutral simulations.

(d) Allele Frequencies vs. Maize Domestication

Genetic, archeological, botanical, and paleoecological data furnished evidence that maize (*Zea mays ssp. mays*) had a single domestication origin from the wild grass teosinte (*Zea mays ssp. parviglumis*) in the Río Balsas region, southwestern Mexico, approximately 6,300–10,000 calendar years before present [44–53]. Pollen samples taken from sediments in lakes, swamps, and archeological deposits have provided evidence for the presence or absence of *Zea* (maize and/or teosinte) in the Americas and have been used to estimate the age of maize domestication and dispersion [44]. Blake [44] summarized the *Zea* pollen dates from several American archeological sites, and we selected this data set to perform our analysis. To test the connection between maize culture and the *ABCA1*230Cys* variant, we used allele frequencies from Mesoamerica/Central America populations (Zapotec, Maya, Nahuatl,Kaqchikel-Quiche, Totonac, Cabecar, and Guaymí) as well as *Zea* pollen dates obtained in archeological sites located geographically near these populations (Table 2). Spearman rank order correlations between the two data sets (*Zea* pollen archeological records and *ABCA1*230Cys* allele frequencies) were obtained using the Statistica 7.0 software (StatSoft, Inc©).

Results

SNP Genotyping and Intra- and Inter-group Structures

Table 1 presents the genotype and allele frequencies for the 1905 individuals analyzed, including the new samples genotyped in the present study. A molecular analysis of variance (AMOVA) test was performed to quantify the level of population structure observed within and between the 3 subdivisions adopted here (Mesoamerican agriculturalists, Andean agriculturalists, and South American hunter-gatherer/forager; Table 1). A significant difference was observed between the subdivisions (F_{CT} = 0.036; $p = 0.000$). On the other hand, the highest F_{ST} among populations within subdivisions was observed in the South American hunter-gatherers/foragers (0.053; $p = 0.005$), and a value 5 times lower was found among Mesoamerican agriculturalists (0.013; $p = 0.008$); no sign of structuration was found in the Andes area ($p = 0.36$).

No significant differences in allelic frequencies were found when subsistence modes (hunter-gatherer/foragers *vs.* agriculturalists) were compared using the student's *t*-test ($p = 0.1316$; Table 1). However, significant differences were observed between Mesoamerican agriculturalists and Andean agriculturalists or South American hunter-gatherers/foragers ($p = 0.0022$ and $p = 0.0174$, respectively). A comparison of South American hunter-gatherers/foragers and Andean agriculturalists revealed no significant differences ($p = 0.351$).

(c.1) *ABCA1*230Cys* allele age. The *ABCA1*230Cys* allele age estimates, using population frequency information, were 12,097 YBP and 19,409 YBP, considering data from all populations and only those from Mesoamerica/Central America,

Table 2. Zea pollen relics ages and 230Cys*ABCA1 populations frequencies used for the regression analysis.

Native Americans		Zea Pollen relics		Site Ages (BP)[2]		Allele frequency
Population[1]	**Geographic region**	**Archeological site**	**Geographic region**	**Radiocarbon years**	**Calendar years**	***230Cys*ABCA1***
Zapotec	Oaxaca	Guilá Naquitz	Oaxaca	8240	9212	0.24
Maya	Tabasco	San Andrés	Tabasco	6208	7122	0.20
Nahuatl	Mexico state	Zoalpilco	Mexico state	5090	5835	0.17
Kaqchikel-Quiche	Guatemala	Zipacate	Guatemala	4600	5318	0.15
Totonac	Veracruz	Laguna Pompal	Veracruz	4250	4818	0.13
Cabecar	Costa Rica	Lago Cote	Costa Rica	2940	3096	0.10
Guaymí	Panama/Costa Rica	Gatun Lake	Panamá	4000	4468	0.15

[1]Located near the archeological sites of *Zea* pollen relics; 2 Conversion according to www.radiocarbon.ldeo.columbia.edu/radcarbcal.htm.
[2]Data relative to archeological information were obtained from Blake (2006).
doi:10.1371/journal.pone.0038862.t002

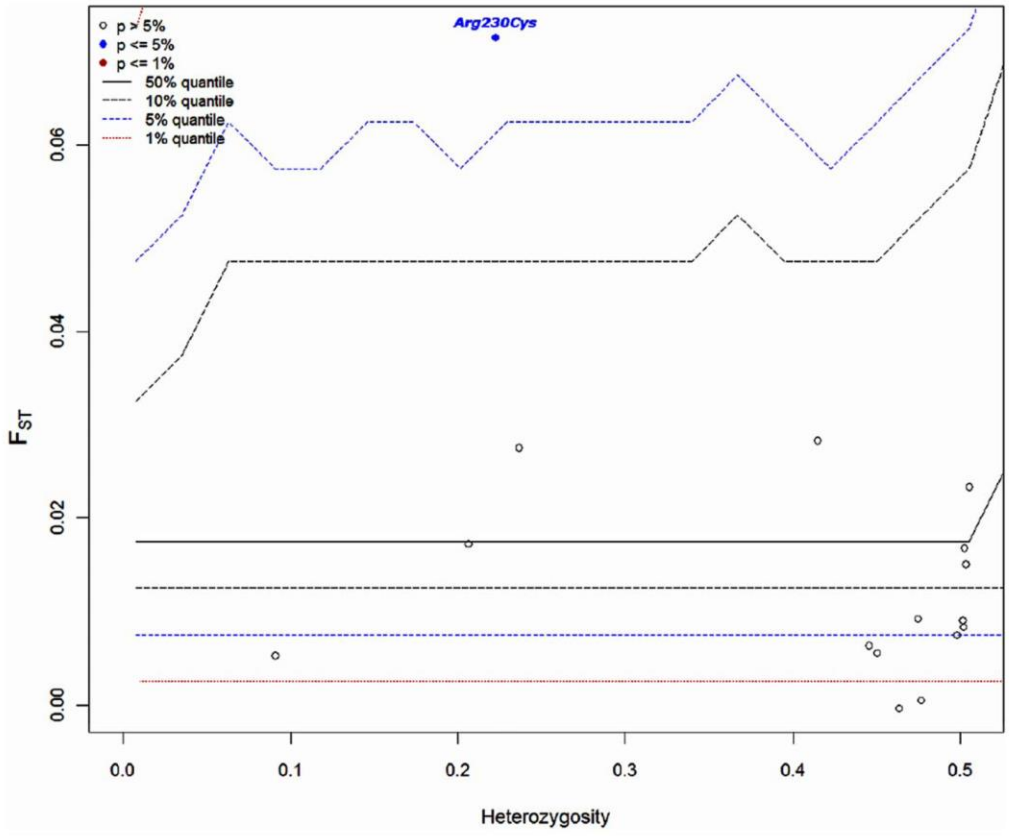

Figure 1. Plot of the joint F_{ST} and *He* distributions for the Mesoamerican agriculturalist *versus* South American (agriculturalists + hunter-gatherer/foraging) groups. Each dot indicates a SNP (listed in item c.2 in the Materials and Methods section). The lines represent confidence intervals. Only the *ABCA1* locus showed significance at the 5% level (filled blue circle). Five selected SNPs were not plotted in the figure because of monomorphic sites in all subdivisions, missing data, and/or dot superposition.
doi:10.1371/journal.pone.0038862.g001

respectively. But the allele age obtained using the observed LD and a Bayesian approach, is relatively younger, 7,540 YBP, with a posterior probability of 99%. Although the numbers generated with these two distinct methods are compatible with an American origin of the *ABCA1*230Cys* allele [32], the last seems more realistic since methods based on LD, rather than frequencies, have the property of reflecting what happened to an allele more accurately [33]. Discrepancies between estimates obtained from these two approaches are usually taken as evidence that selection has increased the frequency of the allele to higher levels than expected by random genetic drift [33,54].

(c.2) Detecting candidate loci for selection. Patterns of genetic diversity between populations can be used to detect loci under selection [30]. The joint distributions of *He* and F_{ST} of the *ABCA1*230* locus and 20 other SNPs (listed in item c.2 in the Materials and Methods section) were examined to test whether the *ABCA1*230* locus and these SNPs departed from neutral expectation. The values obtained indicated that only the *ABCA1*230* polymorphism departed significantly from neutral expectation ($p = 0.02$; Figure 1). However, when the comparisons excluded the Mesoamericans (e.g., Andean agriculturalists *vs.* South American hunter-gatherer/forager subdivisions), no significant departure from the expected under neutrality was found (data not shown). These results suggest that the *ABCA1*230* allele frequencies in Mesoamerica are incompatible with a simple neutral model.

Our neutral demographic simulation analysis showed results in the same direction. The region containing the *ABCA1*230* polymorphism and 19 flanking SNPs presented a slightly lower average heterozygosity than the putative neutral region dataset (0.32 vs. 0.34 respectively); but global and pairwise F_{ST} were higher for the *ABCA1* region (global F_{ST} 0.03 vs. 0.01; average pairwise F_{ST} 0.05 vs. 0.01). Considering that both genomic regions were studied using the same quantity of markers and the same sampling strategy, in populations that were subjected to the same demographic history, the observed differences may occur due to diverse factors, one of them being natural selection. To test this hypothesis each dataset, summarized by the above-mentioned statistics, was also compared to each of 100,000 neutral simulations by means of Euclidian distances. The empirical dataset containing the *ABCA1*230* polymorphism presented a poorer fit to neutrality than the putative neutral region dataset, showing Euclidian distances that were twofold higher than those of the neutral simulations (70.64 *vs.* 35.12). Interestingly, when the Mesoamerican agriculturalist subdivision was excluded from the analysis, this difference dramatically decreased (45.80 *vs.* 35.12). Thus, the poor fit to neutrality observed at the *ABCA1*230* site

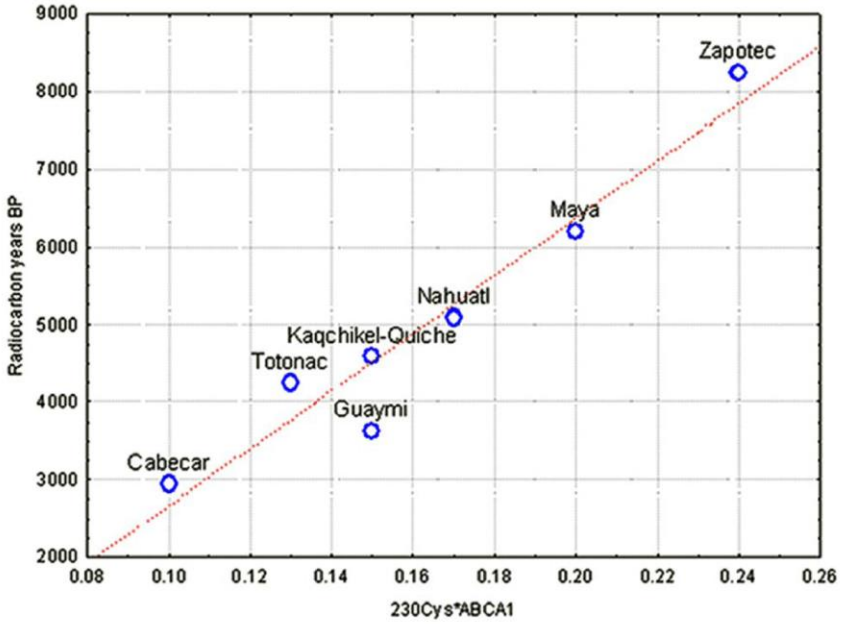

Figure 2. *ABCA1230Cys frequencies versus radiocarbon ages of maize domestication (*Zea* pollen relics; Blake [42]).** Spearman's rho value = 0.936975 and $p = 0.0019$.
doi:10.1371/journal.pone.0038862.g002

and its flanking regions may be associated with the genetic pattern found in Mesoamerica.

(d) Allele frequencies vs. maize domestication. A significant correlation was observed between the *ABCA1*230Cys* allele frequencies and the distribution of the *Zea* pollen relics in Mesoamerica (Figure 2; r = 0.9, $p = 0.002$). It is important to note that the populations used in the correlation analysis performed here (Zapotec, Maya, Nahuatl, Kaqchikel-Quiche, Totonac, Guaymí, and Cabecar) were investigated for microsatellites and other genetic markers in previous studies conducted by our and other groups [27,55,56]. These studies indicated that these populations have a substantial Amerindian substrate, a generally small European contribution, and almost no African influence. For instance, Wang et al. [27] showed that the Maya and Guaymí showed the highest and the lowest numbers of individuals with some level of recent European and African admixture, respectively. This indicates an opposite trend from what would be expected if the level of admixture with non-Indians was influencing our findings, since the variant allele is absent in Europeans and Africans.

Discussion

We can now examine some hypotheses in an attempt to explain the results and to draw the evolutionary scenario associated with the pattern of diversity of the *ABCA1* Arg230Cys* polymorphism.

Maize is considered the most important native crop of the Americas [44–53]. Several lines of evidence indicate that the Mesoamerican village lifestyle began with maize domestication [44,45,49–51,57,58]. Originating in the Mexican southwestern lowlands, maize journeyed southwards, traveling hand-in-hand with pottery and bringing sedentary life to the Andes, although the date of its entry, as well as the dispersion pattern of this crop into and throughout South America, remain controversial [52].

Other crops were also present in the pre-Columbian Mesoamerican civilizations (squash and beans; [50,59]), but maize was the dietary base for most of these civilizations. For example, Benedict and Steggerda [60] showed that 75% of the calories consumed by the Mayas were derived from maize. In addition, Mesoamerica was the only region in the world where an ancient civilization lacked a domesticated herbivore. Therefore, protein from domesticated animal sources would have been scarce in Pre-Hispanic Mesoamerica in comparison to other parts of the ancient urbanized world, including the Andes [61]. As a whole, these studies demonstrated that the diet of the first Mesoamerican sedentary communities was extremely dependent on maize. These early farmers, however, suffered periods of plantation loss, questioning the common assumption that farming and sedentary lifestyle brought increased dietary stability and health homeostasis [62]. Several studies have revealed that homeostasis should have declined with sedentary farming, and bioarcheologists and paleopathologists have also detected a deterioration in Mesoamerican health indices from ~8,000 to ~500 years before present-YBP ([63] and references therein). Domestic crops are more vulnerable than wild ones, crowding promotes crop diseases, and storage systems often fail (estimates suggest that as much as 30% of stored food is lost even in a modern sophisticated system [62]). In other study, based on the molecular analysis of dietary diversity for three archaic Native Americans, Poinar et al. [64] found evidence that, as compared to individuals dependent on agriculture, the diet of hunter-gatherers seems to have been more varied and nutritionally sound. Clearly, a diet based on one or only a few crops should have been deleterious to health in the pre-Columbian era [65]. These different lines of evidence illustrate that the incipient farming niches of Mesoamerica, when communities of hunters/gatherers/foragers started to cultivate and domesticate wild plants, could have been remarkably unstable like those of other pre-industrial societies [6].

Based on what was discussed above, as well as in our results (allele age and neutrality/selection tests), it is reasonable to suppose that *ABCA1*230Cys* has an American origin and it could have had a selective advantage during the periods of food scarcity experienced by Mesoamericans during the implementation of the sedentary life style based on maize. The strong correlation between maize culture propagation and *230Cys* frequencies in this region reinforces this suggestion, even when considering that the advantage of the allele may have been lost after technological innovations had been implemented and agricultural production stabilized. Peng *et al.* [66] presented evidence for a similar case of gene-culture coevolution, suggesting that positive selection for the *ADH1B*47His* allele was caused by the emergence and expansion of rice domestication in East Asia.

Noteworthy is that other environmental factors may also have been involved in the distribution of the *ABCA1* alleles, since cholesterol plays an important role in various infectious processes, such as the entry and replication of Dengue virus type 2 and flaviviral infection [67]. Additionally, the ABCA1 transporter participates in infectious and/or thrombotic disorders involving vesiculation, since homozygous *ABCA1* gene deletions confer complete resistance against cerebral malaria in mice [68,69]. These findings can be considered as additional causal factors to the *ABCA1*230Cys* selective sweep associated with agricultural development. A sedentary village lifestyle with a corresponding growth in the density of the local population can promote an increase in the mortality rate, particularly in children under 5 years of age [70]. For example, archeological and paleoecological evidence in Europe showed that during the Neolithic demographic transition, the causes of increased infant mortality would have included a lack of drinking water supplies, contamination by feces, emergence of highly virulent zoonoses, as well as an increase in the prevalences of other germs such as *Rotavirus* and *Coronavirus* (causing diarrhea, one of the main killers of children under 5 years of age), *Streptococcus*, *Staphylococcus*, *Plasmodium* (*P. falciparum* and *P. vivax*, which are believed to have emerged more recently), and *Herpesvirus* [70]. However, the real impact of the *ABCA1*230Cys* variant in these infectious processes will require additional functional studies.

In agreement with this historical scenario, the genetic variation in *Arg230Cys* presented a worse fit to neutrality than loci known to be neutral, indicating that selective mechanisms are necessary to explain the genetic diversity of *Arg230Cys*, especially when the Mesoamerican agriculturalist subdivision is considered in the analysis. This result also supports our hypothesis that maize domestication in Mesoamerica lead to changes in the gene pool of the natives from that region.

South America presents much more diversity in relation to habitats, people, and culture than Mesoamerica. For instance, maize arrived in South America, but apparently the level of consumption seen in Mesoamerica/Central America was rarely found there. Archaeological data indicate that only during the implantation and expansion of the Inca Empire (800–500 YBP) was the level of maize consumption important, but the level of consumption was not comparable to that of Mesoamerica/Central America [71–73]. Additionally, South Amerindian hunter-gatherers/foragers present lower intrapopulation genetic variation and higher levels of population structure when compared to those seen in Andean populations [27,74]. This same tendency was also observed in the present study. These results indicate low levels of gene flow between villages/populations and low effective population sizes, favoring the role of genetic drift. Conversely, the Andean groups show opposing characteristics. These findings correlate well with distinguishing patterns of gene flow and historical effective sizes in these indigenous populations, with cultural differences, as well as with paleoclimatic and environment changes in their habitats [74]. Therefore, the significant role of random processes and/or more heterogeneous cultural and ecological scenarios makes it difficult to define a particular pattern associated with the *Arg230Cys* polymorphism in South American groups, a situation different from that in Mesoamerica.

In conclusion, our analyses demonstrate for the first time a robust correlation between a constructed niche and a selected Native American autochthonous allele. The *230Cys* allele, with a probable origin in America continent, seems have been the target for an ongoing directional selective sweep as a result of the origin and spread of the maize culture in ancient Mesoamerica.

Supporting Information

Figure S1 Twenty SNPs selected based on their location (chromosome 9: from position 85252250 to 85317359) inside the putative neutral region, defined by Schroeder and colleagues [39], which comprises ~76,000 bp around the D9S1120 locus.
(JPG)

Table S1 Populations included in the selection analyses.
(DOC)

Table S2 Allele frequency by region of 20 SNPs located around the D9S1120 locus.
(DOCX)

Table S3 Allele frequency by region of 19 SNPs located around the ABCA1*230 locus.
(DOCX)

Acknowledgments

We are very grateful to the volunteers who contributed in the sampling. We thank M. Villena and R. Vasquez for their assistance in the sample collection in Bolivia, Claiton H. D. Bau for laboratory support, and Rolando González-José, Lavínia Schüler-Faccini, Matthew Nunes, Sidia M. Callegari-Jacques, and Jorge Gomez-Valdez for their comments on the manuscript. We also thank R. Bisso-Machado for his help with editing the tables.

Author Contributions

Conceived and designed the experiments: TH MCB. Performed the experiments: TH VC. Analyzed the data: TH CEGA SA VRPC. Contributed reagents/materials/analysis tools: VAA FR JMD SM RB MTVM FMS SCQ ARL. Wrote the paper: TH FMS MCB.

References

1. Laland KN, Odling-Smee J, Myles S (2010) How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet 11: 137–148.
2. Richerson PJ, Boyd R, Henrich J (2010) Gene-culture coevolution in the age of genomics. Proc Natl Acad Sci USA 107: 8985–8992.
3. Odling-Smee FJ (1988) Niche construction phenotypes. In: The role of behavior in evolution (ed. H. C. Plotkin), 73–132. Cambridge: MIT Press.
4. Odling-Smee FJ, Laland KN, Feldman MW (2003) Niche construction: the neglected process in evolution. Monographs in Population Biology 37. Princeton: Princeton Univ. Press.
5. Gerbault P, Liebert A, Itan Y, Powell A, Currat M, et al. (2011) Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond B Biol Sci 366: 863–877.

6. Rowley-Conwy P, Layton R (2011) Foraging and farming as niche construction: stable and unstable adaptations. Philos Trans R Soc Lond B Biol Sci: 366: 849–862.
7. Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. Hum Biol 69: 605–628.
8. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner S F, et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111–1120.
9. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
10. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31–40.
11. Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. Hum Biol 81: 597–619.
12. Ingram CJ, Mulcare CA, Itan Y, Thomas MG, Swallow DM (2009) Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet 124: 579–591.
13. Itan Y, Jones BL, Ingram CJ, Swallow DM, Thomas MG (2010) A worldwide correlation of lactase persistence phenotype and genotypes. BMC Evol Biol 10: 36.
14. Lewinsky RH, Jensen T G, Møller J, Stensballe A, Olsen J, et al. (2005) T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Hum Mol Genet 14: 3945–3953.
15. Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, et al. (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nat Genet 35: 311–313. Erratum in: Nat Genet (2004) 36: 106.
16. Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG (2007) Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc Natl Acad Sci U S A 104: 3736–3741.
17. Livingstone FB (1958) Anthropological implications of sickle-cell distribution in West Africa. Am Anthropol 60: 533–562.
18. Neel JV, Salzano FM (1967) Further studies on the Xavante Indians. X. Some hypotheses-generalizations resulting from these studies. Am J Hum Genet 19: 554–574.
19. Hunley KL, Cabana GS, Merriwether DA, Long JC (2007) A formal test of linguistic and genetic coevolution in native Central and South America. Am J Phys Anthropol 132: 622–631.
20. Kemp BM, González-Oliver A, Malhi RS, Monroe C, Schroeder KB, et al. (2010) Evaluating the Farming/Language Dispersal Hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica. Proc Natl Acad Sci USA 107: 6759–6764.
21. Tovo-Rodrigues L, Callegari-Jacques SM, Petzl-Erler ML, Tsuneto L, Salzano FM, et al. (2010) Dopamine receptor D4 allele distribution in Amerindians: a reflection of past behavior differences? Am J Phys Anthropol 143: 458–464.
22. Acuña-Alonzo V, Flores-Dorantes T, Kruit JK, Villarreal-Molina T, Arellano-Campos O, et al. (2010) A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. Hum Mol Genet 19: 2877–2885.
23. Romero-Hidalgo S, Villarreal-Molina T, González-Barrios JA, Canizales-Quinteros S, Rodríguez-Arellano ME, et al. (2012) Carbohydrate intake modulates the effect of the *ABCA1-R230C* variant on HDL cholesterol concentrations in premenopausal women. J Nutr 142: 278–283.
24. Neel JV (1962) Diabetes mellitus: a 'thrifty' genotype rendered detrimental by 'progress'? Am J Hum Genet 14: 353–362.
25. Bortolini MC, Salzano FM, Bau CH, Layrisse Z, Petzl-Erler ML, et al. (2002) Y-chromosome biallelic polymorphisms and Native American population structure. Ann Hum Genet 66: 255–259.
26. Bortolini MC, Salzano FM, Thomas MG, Stuart S, Nasanen SP, et al. (2003) Y-chromosome evidence for differing ancient demographic histories in the Americas. Am J Hum Genet 73: 524–539.
27. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, et al. (2007) Genetic variation and population structure in Native Americans. PLoS Genet 3: e185.
28. Mazières S, Guitard E, Crubézy E, Dugoujon JM, Bortolini MC, et al. (2008) Uniparental (mtDNA, Y-chromosome) polymorphisms in French Guiana and two related populations–implications for the region's colonization. Ann Hum Genet 72: 145–156.
29. Mazières S, Sevin A, Callegari-Jacques SM, Crubézy E, Larrouy G, et al. (2009) Population genetic dynamics in the French Guiana region. Am J Hum Biol 21: 113–117.
30. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10: 564–567.
31. Kimura M, Ohta T (1973) The age of a neutral mutation persisting in a finite population. Genetics 75: 199–212.
32. González-José R, Bortolini MC, Santos FR, Bonatto SL (2008) The peopling of America: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. Am J Phys Anthropol 137: 175–187.
33. Slatkin M, Rannala B (2000) Estimating allele age. Annu Rev Genom Hum Genet 1: 225–249.
34. Rannala B, Reeve JP (2008) Joint Bayesian estimation of mutation location and age using linkage disequilibrium. Pacific Symp Biocomp 8: 526–534.
35. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
36. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Re 19: 826–837.
37. Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol Evol 23: 347–351.
38. Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103: 285–298.
39. Schroeder KB, Jakobsson M, Crawford MH, Schurr TG, Boca SM, et al. (2009) Haplotypic background of a private allele at high frequency in the Americas. Mol Biol Evol 26: 995–1016.
40. Hey J (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol 3: e193.
41. Pavlidis P, Laurent S, Stephan W (2010) msABC: a modification of Hudson's ms to facilitate ABC analysis. Mol Ecol Res 10: 723–727.
42. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18: 337–338.
43. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11: 116.
44. Blake M (2006) Dating the initial spread of *Zea mays*. In Histories of maize (eds. J. E Staller, R H Tykot, BF Benz), 55–68. San Diego: Academic Press.
45. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez JG, Buckler E, et al. (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci USA 99: 6080–6084.
46. Dull R (2006) The maize revolution: a view from El Salvador. In: Histories of maize (eds. J. E Staller, R H Tykot, B F Benz), 357–363. San Diego: Academic Press.
47. Jaenicke-Després VR, Smith BD (2006) Ancient DNA and the integration of archaeological and genetic approaches to the study of maize domestication. In: Histories of maize (eds. J E Staller, RH Tykot, BF Benz), 83–92. San Diego: Academic Press.
48. Lesure RG (2008) The Neolithic demographic transition in Mesoamerica? Larger implications of the strategy of relative chronology. In The Neolithic demographic transition and its consequence (eds. JP Bocquet-Appel, O Bar-Yosef), 107–138. New York: Springer.
49. Ranere AJ, Piperno DR, Holst I, Dickau R, Iriarte J (2009) The cultural and chronological context of early Holocene maize and squash domestication in the Central Balsas River Valley, Mexico. Proc Natl Acad Sci USA 106: 5014–5018.
50. Piperno DR (2007) Late Pleistocene and Holocene environmental history of the Iguala Valley, Central Balsas Watershed of Mexico. Proc Natl Acad Sci USA 104: 11874–11881.
51. van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, et al. (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc Natl Acad Sci USA 108: 1088–1092.
52. Lia VV, Confalonieri VA, Ratto N, Hernández J A, Alzogaray AM, et al. (2007) Microsatellite typing of ancient maize: insights into the history of agriculture in southern South America. Proc Biol Sci 274: 545–554.
53. Tian F, Stevens NM, Buckler ES 4th. (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc Natl Acad Sci USA 106: 9979–9986.
54. Ding YC, Chi HC, Grady DL, Morishima A, Kidd JR, et al. (2002) Evidence of positive selection acting at the human dopamine receptor D4 gene locus. Proc Natl Acad USA 99: 309–314.
55. Vargas-Alarcon G, Moscoso J, Martinez-Laso J, Rodriguez-Perez JM, Flores-Dominguez C, et al. (2007) Origin of Mexican Nahuas (Aztecs) according to HLA genes and their relationships with worldwide populations. Mol Immunol 44: 747–755.
56. González-Martín A, Gorostiza A, Rangel-Villalobos H, Acunha V, Barrot C, et al. (2008) Analyzing the genetic structure of the Tepehua in relation to other neighbouring Mesoamerican populations. A study based on allele frequencies of STR markers. Am J Hum Biol 20: 605–613.
57. Raymond JS, Deboer WR (2006) Maize on the move. In Histories of maize (eds. JE Staller, RH Tykot, BF Benz), 337–341. San Diego: Academic Press.
58. Chisholm B, Blake M (2006) Diet in prehistoric Soconusco. In Histories of maize (eds. JE Staller, RH Tykot, BF Benz), 161–167. San Diego: Academic Press.
59. Brown CH (2010) The pastoral niche in pre-Hispanic Mesoamerica. In Pre-Columbian foodways: interdisciplinary approaches to food, culture, and markets in ancient Mesoamerica (eds. J Staller, M Carrasco), 273–291. New York: Springer.
60. Benedict FG, Steggerda M (1936) The food of present-day Maya Indians of Yucatan. Washington: Carnegie Institute of Washington.
61. Parsons JR (2010) The pastoral niche in pre-Hispanic Mesoamerica. In Pre-Columbian foodways: interdisciplinary approaches to food, culture, and markets in ancient Mesoamerica (eds. J Staller, M Carrasco), 109–136. New York: Springer.
62. Cohen NM (2008) Implications of the NDT for worldwide health and mortality in prehistory. In The Neolithic demographic transition and its consequence (eds. JP Bocquet-Appel, O Bar-Yosef), 481–500. New York: Springer.
63. Kennett DJ, Voorhies B, Martorana D (2006) An ecological model for the origins of maize-based food production on the Pacific Coast of Southern Mexico.

In Behavioral ecology and the transition to agriculture (eds. DJ Kennett, B Winterhalder), 103–136. Berkeley: University of California Press.
64. Poinar HN, Kuch M, Sobolik KD, Barnes I, Stankiewicz AB, et al. (2001) A molecular analysis of dietary for three archaic Native Americans. Proc Natl Acad USA 98: 4317–4322.
65. Steckel RH, Rose JC, Larsen CS, Walker PL (2002) Skeletal health in the western hemisphere from 4000 B.C. to the present. Evol Anthropol 11: 142–155.
66. Peng Y, Shi H, Qi XB, Xiao CJ, Zhong H, et al. (2010) The *ADH1B* Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. BMC Evol Biol 20: 10–15.
67. Lee CJ, Lin HR, Liao CL, Lin YL (2008) Cholesterol effectively blocks entry of flavivirus. J Virol 82: 6470–6480.
68. Simons K, Toomre D (2000) Lipid rafts and signal transduction. Nat Rev Mol Cell Biol 1: 31–39.
69. Combes V, Coltel N, Alibert M, van Eck M, Raymond C, et al. (2005) *ABCA1* gene deletion protects against cerebral malaria: potential pathogenic role of microparticles in neuropathology. Am J Pathol 166: 295–302.
70. Bocquet-Appel JP (2008) Explaining the Neolithic demographic transition. In The Neolithic demographic transition and its consequence (eds. JP Bocquet-Appel, O Bar-Yosef), 35–56. New York: Springer.
71. Zarrillo S, Pearsall DM, Raymond JS, Tisdale MA, Quon DJ (2009) Directly dated starch residues document early formative maize (*Zea mays L.*) in tropical Ecuador. Proc Natl Acad Sci USA 106: 9979–9986.
72. Schwarcz HP (2006) Stable carbon isotope analysis and human diet: a synthesis. In Histories of maize (eds. JE Staller, RH Tykot, BF Benz), 315–324. San Diego: Academic Press.
73. Tykot RH, Burger RL, Van der Merwe N (2006) The importance of maize in initial period and early horizon Peru. In Histories of maize (eds. JE Staller, RH Tykot, BF Benz), 187–233. San Diego: Academic Press.
74. Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, et al. (2001) Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am J Hum Genet 68: 1485–1496.