



Modeling of linkage disequilibrium in whole genome genetic association studies

Randall Johnson

► To cite this version:

Randall Johnson. Modeling of linkage disequilibrium in whole genome genetic association studies. Genomics [q-bio.GN]. Conservatoire national des arts et metiers - CNAM, 2014. English. ⟨NNT : 2014CNAM0963⟩. ⟨tel-01208030⟩

HAL Id: tel-01208030

<https://theses.hal.science/tel-01208030v1>

Submitted on 1 Oct 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

ÉCOLE DOCTORALE TECHNOLOGIQUE ET PROFESSIONNELLE

Laboratoire Genomique, Bioinformatique et Application

THÈSE présentée par :

Randall Johnson

soutenue le : 19 Décembre 2013

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline/Spécialité : **Bioinformatique**

Modeling of linkage disequilibrium in
whole genome genetic association studies

THÈSE dirigée par :
Jean-François Zagury

Chaire de Bioinformatique, Conservatoire National
des Arts et Metiers, Paris, France

RAPPORTEURS :
Olivier Delaneau
Meredith Yeager

Department of Statistics,
University of Oxford, Oxford, UK
Cancer Genomics Research Laboratory
Frederick National Laboratory, Gaithersburg, MD, USA

JURY :
Cheryl Winkler
George Nelson

Basic Research Laboratory
Frederick National Laboratory, Frederick, MD, USA
BSP CCR Genetics Core
Frederick National Laboratory, Frederick, MD, USA

For Julie - my greatest cheer leader and support

Remerciements

First I would like to thank my family for all of their patience and support while I've worked on this. I couldn't have done it without them! Special thanks is also due to Cherie, George and Bailey who have supported me in this project all along. I would also like to thank Jean-Francois for inviting me to CNAM in the beginning, and who has been a wonderful facilitator to keep things on track for me. My reviewers, Meredith and Olivier, have also been great to work with, especially during the final push to get everything finalized. Thanks!

Résumé

Modélisation du déséquilibre de liaison dans les études d'association genome entier

1. Introduction

La découverte de la liaison génétique par Thomas Morgan il y a un siècle a eu d'immenses conséquences théoriques et pratiques. Gregor Mendel avait observé que différents traits étaient hérités séparément ; Morgan observa que certains traits avaient tendance à être hérités ensemble. L'observation de Morgan conduisit au concept de gènes agencés sur un génome linéaire, qui subit des recombinaisons au cours de la méiose lors de générations successives. En pratique, la liaison génétique permet l'identification de variants génétiques sous-tendants des maladies humaines : une forte association entre une maladie et un variant génétique suggère que la variation causant la maladie est proche du locus observé. Ce concept général conduit à de nombreuses stratégies spécifiques pour l'identification de facteurs génétiques causant la maladie, les questions statistiques étant cruciales pour toutes. D'une façon générale, les stratégies peuvent être divisées en études familiales et études d'association genome entier (GWAS). La cartographie par étude du déséquilibre de liaison dans des populations mixtes (MALD) représente une approche hybride entre les deux, où les populations ancestrales d'une population mixte jouent le rôle des parents dans une étude familiale. Cette thèse aborde les questions statistiques des GWAS et des MALD, démontrant spécifiquement la capacité de l'analyse en composantes principales (ACP) à gérer les questions découlant de la complexité de la liaison génétique.

Le déséquilibre de liaison (DL) correspond à l'association non aléatoire d'allèles de loci proches. L'apparition de nouvelles mutations, la dérive génétique et la sélection sont les processus à l'origine de cette association entre allèles, tandis que les événements de recombinaison rompent l'association. En raison de l'échelle temps/longueur de

recombinaison (~1 recombinaison pour 100 générations par mégabase), les loci distants sont non corrélés dans une population stable, mais les corrélations statistiquement utiles entre les facteurs pathologiques et les polymorphismes sur des longueurs de 1-100 kb abondent (International HapMap Consortium et al., 2007).

Les études familiales furent la première application du DL pour l'identification de gènes pathologiques. Les premiers gènes pathologiques cartographiés furent pour les troubles fonctionnels liés à l'X qui affectaient les porteurs masculins mais pas les porteurs féminins (par ex. facteur de déficience VIII, la cause de l'hémophilie (Toole et al., 1984)). La cartographie du gène autosomique responsable de la mucoviscidose a été complétée en 1989 (Riordan et al., 1989; Rommens et al., 1989). Depuis lors, la majorité des maladies mendéliennes avec une pénétrance moyenne ou élevée ont été cartographiées par le biais d'études familiales multiplexes ou de grands arbres généalogiques. Ces études emploient des polymorphismes mononucléotidiques (SNPs) ou des marqueurs microsatellites pour tester la liaison statistique entre les phénotypes et les marqueurs. La supposition sous-jacente est que le variant causal se situera sur le segment chromosomique hérité par tous les descendants affectés. On utilise alors le clonage positionnel pour identifier le gène responsable. Les études familiales ont connu un large succès pour les phénotypes hautement pénétrants avec héritage monogénique, mais elles ne peuvent pas être appliquées aux les maladies complexes où tous les membres de la famille ne reçoivent pas nécessairement la même exposition environnementale ou génétique, ou lorsque la pénétrance est faible.

Les études d'association pangénomique (GWAS) exploitent aussi le DL: le génotypage d'un ensemble suffisamment dense de variants génétiques distribués sur l'ensemble du génome rend vraisemblable le fait qu'un variant pathologique soit en DL suffisant avec un ou plusieurs marqueurs typés pour créer une association statistiquement significative entre le marqueur et la maladie. L'approche GWAS permet l'utilisation de données cas-contrôles pour les maladies communes avec des causes complexes, par opposition aux maladies mendéliennes causées par la variation d'un gène unique.

Bien que plus générales en applicabilité, les études d'association génétique cas-contrôles sont moins puissantes que les études familiales, nécessitant des échantillons beaucoup plus grands pour de nombreuses raisons : pénétrance, tests multiples, faible effet des variants les plus communs, mais également parce qu'il leur manque les grands blocs identiques par ascendance des membres d'une famille (Lander & Schork, 1994). Un autre inconvénient des études d'association génétique cas-contrôles est la fiabilité de la population contrôle (dans les études familiales, les membres non affectés de la famille offrent un contrôle interne) ; le groupe de contrôle approprié peut ne pas être évident ou difficile à recruter (Blackwell & Hodges, 1957). Cependant, les études cas-contrôles sont capables de modéliser une association entre une maladie non mendélienne complexe et des variants communs en DL avec des variants causaux. La liaison incomplète, le faible effet des variants à basse fréquence et les tests multiples limitent la puissance à détecter l'association génotype-phénotype (Lander & Schork, 1994). La plus problématique de ces questions reste la comparaison multiple : la comparaison pour association à $\sim 10^6$ marqueurs nécessite la correction d'un nombre semblable de comparaisons ; La correction précise à apporter est encore débattue et est l'un des thèmes abordés dans cette thèse, mais dans cette problématique créée de sérieuses questions de puissance pour les GWAS (Hoggart, Clark, De Iorio, Whittaker, & Balding, 2008).

Avec le développement de la liaison, l'effet de récents brassages sur les schémas de liaison a été reconnu comme une source potentielle d'informations pour cartographier les gènes associés à des phénotypes dont la prévalence diffère entre deux populations ancestrales. Les populations brassées conservent des haplotypes étendus intacts des populations fondatrices pendant de nombreuses générations, créant un déséquilibre de liaison dans les populations mixtes (ALD) qui peut être utilisé pour déduire l'ascendance génétique d'un locus (Gabriel et al., 2002; McKeigue, 1997). Dans les études de maladies avec une prévalence substantiellement différente entre les deux populations parentales, on peut déduire que les loci avec une association entre la maladie et l'ascendance locale abritent les gènes pathologiques. Comme ces haplotypes intacts sont beaucoup plus longs que l'étendue du DL au sein de la population, la cartographie ALD réduit substantiellement le nombre de comparaisons par

rapport aux GWAS, et peut donc substantiellement accroître la capacité à détecter les associations avec une maladie lorsqu'elle est applicable (McKeigue, 1997). Un problème technique cependant est que le DL de la population qui constitue la base des GWAS est un facteur perturbant dans l'identification de blocs chromosomiques hérités de populations ancestrales.

Cette thèse aborde le problème de la puissance dans les études d'association pangénomique (GWAS) au travers de l'étude du déséquilibre de liaison existant entre les marqueurs. La complexité du DL perturbe en effet les estimations simples du nombre de comparaisons dans une GWAS, ainsi que les tentatives simples de correction pour le DL d'une population dans une MALD. Dans cette thèse, je prends en considération ces questions et estime que l'ACP offre peut-être une solution optimale. Cela forme la base de mes deux objectifs, le premier étant de soigneusement passer en revue et tester les méthodes de seuil de significativité pour les GWAS pour tenter de trouver le seuil de significativité le plus précis pour un taux d'erreur de type I souhaité dans une étude. Sur les sept méthodes testées, la correction de Bonferroni pour les comparaisons multiples est à la fois la méthode la plus conservatrice et la plus communément utilisée. D'autres méthodes testées utilisent soit diverses mesures de DL entre les marqueurs pour estimer le nombre réel de comparaisons indépendantes faites, ou elles utilisent un système de test par permutation optimisé pour estimer le seuil de significativité pour le taux d'erreur de type I désiré à l'échelle de l'étude (Browning, 2008; Duggal, Gillanders, Holmes, & Bailey-Wilson, 2008; Gao, Starmer, & Martin, 2008; Han, Kang, & Eskin, 2009). L'assouplissement de ce seuil de significativité, même de façon modérée, peut avoir un impact conséquent sur la puissance statistique, mais peut aussi augmenter le taux d'erreur de type I de l'étude au-delà du niveau souhaité.

Le second objectif est de tirer parti des informations glanées à partir du déséquilibre de liaison dans des populations mixtes (ALD) dans les études de populations brassées. Les premiers algorithmes à effectuer une telle analyse ont été développés avant l'avènement des puces SNP couvrant des centaines de milliers ou des millions de marqueurs, et par conséquent ne permettent pas l'analyse de telles données (Falush, Stephens, & Pritchard, 2003; Hoggart,

Shriver, Kittles, Clayton, & McKeigue, 2004; Patterson et al., 2004). Les solutions actuelles d'analyse des puces à marqueurs denses ont tendance à être soit inefficace d'un point de vue informatique, soit à simplifier excessivement la modélisation du DL (Seldin, Pasaniuc, & Price, 2011). Un raffinement efficace sur le plan informatique des algorithmes de marqueurs peu abondants modélisant le DL local via l'utilisation des composantes principales est proposé ici. La mise en œuvre logicielle pour utiliser l'ALD pour cartographier les gènes pathologiques est également décrite.

1.1. Liaison génétique ou déséquilibre de liaison

La liaison génétique est fondamentale pour permettre aux études d'association génétique de réduire la recherche de variants causaux en identifiant une région chromosomique associée à la maladie. Dans de rares cas tels que celui de la découverte de l'association d'ACE avec l'infarctus du myocarde, ou de la découverte de l'homozygote CCR5 $\Delta 32$ bloquant l'infection du VIH, une connaissance claire de l'association biologique du gène avec la maladie a conduit directement à la découverte de la variation génétique responsable. Lorsque deux marqueurs sont physiquement proches l'un de l'autre sur un chromosome, il est plus vraisemblable qu'ils soient hérités ensemble car il y a une chance plus faible de recombinaison entre les deux chromosomes au cours de la méiose. La source initiale de DL est la mutation, en ce qu'un nouvel allèle provenant d'un locus par mutation se produit nécessairement sur un chromosome unique et est ainsi associé à tous les allèles portés par ce chromosome spécifique (Bateson & Kilby, 1905; Morgan, 1910; 1911). Au cours des générations successives, la recombinaison casse ce chromosome d'origine, mais même après 5000 générations – environ l'âge des humains modernes –, les segments de chromosome d'une longueur moyenne de 20 kb seront hérités ensemble (Matise et al., 2007). La dérive aléatoire de fréquences des allèles et la sélection d'allèles avantageux par rapport aux allèles délétères contribuent aussi à ce procédé, dans un schéma complexe et largement étudié (Keightley & Otto, 2006; Ohta, 1982; Palaisa, Morgante, Tingey, & Rafalski, 2004; Sober, 1993). Au fur et à mesure du vieillissement, les populations acquièrent plus de variants et ont plus d'occasions de se recombiner, ce qui résulte en blocs (ou haplotypes) plus courts de DL, tandis que les populations plus jeunes ont tendance à avoir des blocs de DL plus longs et

moins de variants communs (International HapMap Consortium et al., 2007). Les populations qui sont géographiquement séparées ou reproductivement isolées pour d'autres raisons acquerront des schémas de DL différents au fil du temps, formant des sous-populations génétiquement distinctes avec des caractéristiques phénotypiques distinctes. Cette sous-structure de population est évidente dans les génomes des Européens par exemple, avec des individus issus de régions géographiques différentes présentant des schémas de DL légèrement différents (P. Price, James, Fernandez, & Frencha, 2004).

La structure du DL entre les variants pathogènes et les loci proches permet aux chercheurs de déduire la présence du variant causal en vertu de sa liaison avec un marqueur proche. Ce postulat clé de la génétique est au cœur de mon travail dans diverses études d'association génétique (Freedman et al., 2011; Hendrickson et al., 2008; Kopp et al., 2011; Nelson et al., 2010). Cela forme aussi la base de la GWAS, ce qui permet l'interrogation du génome humain entier en utilisant seulement un sous-ensemble de marqueurs. Lorsque l'on trouve une association significative entre un marqueur génétique et le phénotype étudié, une interférence peut venir du fait que l'un ou plusieurs variants causaux en DL avec le marqueur génétique aient un effet fonctionnel sur le phénotype.

1.2. Liaison génétique dans des populations mixtes

Des populations récemment brassées ont des blocs de DL étendus beaucoup plus longs qui forment une mosaïque des populations ancestrales (par ex. les Afro-américains descendent principalement des Africains occidentaux avec un mélange européen de l'ordre de 20 %) (Parra et al., 1998). Les haplotypes étendus descendent de chaque population ancestrale et sont lentement cassés avec la recombinaison au cours de la méiose de chaque génération suivante. Les blocs haplotypiques consistent en un chromosome entier hérité de chaque population ancestrale dans la génération brassée initiale. La longueur de ces blocs, qui décline au fur et à mesure à chaque génération, est fonction de la distance génétique, mesurée en centimorgans (cM). Après une génération par exemple, chaque bloc haplotypique étendu variera, mais la taille de bloc moyenne sera d'environ 100 cM ou à peu près cent millions de bases. On parle de déséquilibre de liaison dans les populations mixtes (ALD) pour faire

référence à la corrélation de marqueurs à l'intérieur de ces blocs haplotypiques étendus. À l'intérieur de ces blocs haplotypiques étendus cependant, l'ALD est perturbé par le DL local subsistant des populations ancestrales car chaque haplotype étendu descend d'une population ancestrale (Sundquist, Fratkin, Do, & Batzoglou, 2008; Tang, Coram, Wang, Zhu, & Risch, 2006). Une structure de corrélation complexe émerge de l'ALD et du DL local, où la connaissance de l'allèle sur un locus peut fournir des informations significatives sur les loci voisins, et la connaissance des allèles dans un bloc haplotypique peut fournir des informations significatives sur l'ascendance de ce locus.

L'ALD peut aussi ajouter des informations intéressantes à l'analyse des études d'association en utilisant des populations brassées. Au lieu de faire une étude d'association cherchant les associations entre les marqueurs et un phénotype, on peut chercher la liaison entre un bloc haplotypique étendu dérivé d'une ou de l'autre population ancestrale avec un trait ou phénotype qui diffère en fréquence entre les deux populations parentales ancestrales. En d'autres termes, nous anticipons que le variant causal est plus susceptible d'être porté sur le même chromosome ancestral que la population ancestrale montrant une prévalence plus élevée de la maladie (J. C. Stephens, Briscoe, & O'Brien, 1994). Ceci est analogue aux études familiales qui utilisent de longs blocs de liaison pour identifier des mutations pathologiques dans les familles affectées ségrégant un gène pathologique (McKeigue, 1997).

1.3. Modélisation du DL par composantes principales

La sous-structure de la population due à des différences d'ascendance parmi les cas et les contrôles est une source courante de biais pouvant résulter en des associations de faux positifs si on ne la corrige pas (Campbell et al., 2005). Cette source de biais peut aisément être modélisée avec une analyse en composantes principales (ACP) des données génétiques (A. L. Price et al., 2006). Dans une GWAS, l'ACP démarre avec la caractérisation complète de la covariance des profils génétiques entre chaque paire d'individus de l'étude, suivie par le calcul de valeurs propres et de vecteurs propres de la matrice de covariance. Le premier vecteur propre définit une transformation des génotypes d'un individu qui explique autant que possible la variation génétique de l'échantillon. Chaque vecteur propre successif explique, de

même, autant que possible la variation restante. Chaque transformation du vecteur propre résume la relation dans la donnée et peut être utilisée pour expliquer les tendances générales de l'échantillon : on peut penser aux valeurs propres comme à la quantité de variation représentée par chaque vecteur propre. Les premières composantes principales d'une ACP de GWAS sont une bonne approximation de la sous-structure de la population, et son utilisation en tant que covariable fait partie intégrante de l'analyse classique des GWAS (A. L. Price et al., 2006).

Dans la GWAS, la convention consiste à établir le seuil de significativité pangénomique acceptable entre $p < 1 \times 10^{-8}$ et $p < 5 \times 10^{-8}$ (Hoggart et al., 2008; International HapMap Consortium, 2005; McCarthy et al., 2008; Risch & Merikangas, 1996). En raison du taux accru d'erreur de type II associé à la réduction de la probabilité d'erreurs de type I utilisant ce seuil, beaucoup estiment qu'il est trop conservateur (Duggal et al., 2008; Gao et al., 2008; Gu, 2007; Nicodemus, Liu, Chase, Tsai, & Fallin, 2005). L'ACP peut être appliquée à une GWAS en utilisant un algorithme simpleM pour estimer le nombre de comparaisons multiples indépendantes effectuées (Gao, Becker, Becker, Starmer, & Province, 2010). Ceci est l'une des méthodes utilisées dans mon premier objectif. Une autre application que j'utilise dans cette thèse est le résumé des covariables corrélées dans un modèle statistique (Hawkins, 1973). La stabilité des coefficients de régression linéaire souffre lorsqu'une combinaison de variables explicatives approche un état de collinéarité. Cette corrélation entre variables explicatives peut être éliminée en analysant à la place les composantes principales (CP) obtenues dans une ACP.

2. Résultats

2.1. Considération des comparaisons multiples dans une étude d'association pangénomique

Randall C. Johnson^{1,2}, George W. Nelson¹, Jennefer L. Troyer¹, James A Lautenberger³, Bailey D. Kessing¹, Cheryl A. Winkler¹, Stephen J. O'Brien³

1. Basic Research Program, SAIC-Frederick, Inc, NCI-Frederick, Frederick, MD
2. Chaire de Bioinformatique, Conservatoire National des Arts et Metiers, 75003, Paris, France
3. Laboratory of Genomic Diversity, NCI Frederick, Frederick, MD

Publié dans *BMC Genomics*, 2010, 11:724.

Résumé

Alors que nous entrons dans une ère où tester des millions de SNPs en une seule étude d'association génétique va devenir la norme, l'examen de comparaisons multiples est une partie essentielle de la détermination de l'importance statistique (Risch & Merikangas, 1996). Les ajustements Bonferroni peuvent être effectués, mais ils sont conservateurs en raison de la prépondérance du déséquilibre de liaison (DL) entre les marqueurs génétiques (Duggal et al., 2008), et le test par permutation n'est pas toujours une option viable. Trois grandes classes de corrections ont été proposées pour corriger la nature dépendante des données génétiques dans les ajustements Bonferroni : les tests par permutation et les alternatives apparentées (Browning, 2008), l'analyse en composantes principales (ACP) (Gao et al., 2008), et l'analyse de blocs de DL sur l'ensemble du génome (Duggal et al., 2008). Nous considérons sept applications de ces méthodes communément utilisées à l'aide de données provenant de 1 514 participants américains européens génotypés pour 700 078 SNPs dans une GWAS pour le SIDA.

Une correction de Bonferroni utilisant le nombre de blocs de DL par les trois algorithmes implémentés par Haploview a résulté en un seuil insuffisamment conservateur, correspondant à un niveau de significativité pangénomique de $\alpha = 0.15 - 0.20$. Nous avons observé une augmentation modérée en puissance avec l'utilisation de PRESTO, SLIDE et simpleM par rapport aux méthodes de Bonferroni traditionnelles pour des données de population génotypées sur la plate-forme Affymetrix 6.0 des Américains européens ($\alpha = 0.05$ seuils entre 1×10^{-7} et 7×10^{-8}).

La correction du nombre de blocs de DL a résulté en un ajustement Bonferroni anti-conservateur. SLIDE et simpleM sont particulièrement utiles lors de tests statistiques non

disponibles en paquet de test par permutation optimisée, et les p valeurs pangénomiques corrigées utilisant SLIDE sont beaucoup plus faciles à interpréter pour les consommateurs d'études GWAS.

2.2. ALDsuite: Marqueur dense MALD utilisant composantes principales de déséquilibre de liaison ancestrale

Randall C. Johnson^{1,2}, George W. Nelson¹, Jean-Francois Zagury², Cheryl A. Winkler¹

1. Basic Research Program, Leidos Biomedical Research, Inc, Frederick National Laboratory, Frederick, MD

2. Chaire de Bioinformatique, Conservatoire National des Arts et Metiers, 75003, Paris, France

3. Discussion

Comme le coût du génotypage et de séquençage n'a cessé de chuter, le flux d'informations génétiques a augmenté de façon substantielle et continue à augmenter à un rythme toujours plus soutenu. Cette abondance de données accroît grandement notre capacité à comprendre les facteurs de risque de maladies, mais les généticiens font également face à de nouveaux problèmes émanant du nombre de comparaisons statistiques nécessaires pour totalement explorer leurs données. Ces comparaisons multiples résultent en beaucoup plus d'erreurs de type I (tests d'hypothèse avec faux positifs) lorsque l'on utilise un seuil de significativité traditionnel, $\alpha = 0.05$. La solution évidente à ce problème est de baisser le seuil de significativité pour arriver à un taux acceptable d'erreur de type I sur l'ensemble de l'étude, où la probabilité de tout faux positif sur l'ensemble de l'étude est suffisamment bas. Ceci augmente cependant le taux d'erreur de type II (une mesure de tests d'hypothèse avec faux négatif) et avec lui, la perspective d'exclure des découvertes potentiellement importantes, car elles ne peuvent être différenciées de résultats faux positifs dans l'étude (Neyman & Pearson, 1967).

La façon la plus fondamentale de combattre cette baisse de puissance statistique est d'échantillonner plus d'individus, tirant par-là parti du théorème central limite de statistique –

l'agrégation d'un plus large échantillon diminuera le bruit statistique et résultera en des statistiques plus précises (Pólya, 1920). En moyenne, cette précision accrue de statistiques résultera en une significativité statistique accrue pour les « vraies » hypothèses alternatives, permettant ainsi au chercheur de différencier entre les associations statistiques significatives et le bruit statistique. Ceci peut être facilement accompli lorsque le financement et les cas sont disponibles, certaines études comprenant des dizaines de milliers de cas et jusqu'à 100 000 contrôles (Monda et al., 2013; Morris et al., 2012). Dans des conditions de contraintes budgétaires et pour l'étude de maladies rares, ceci peut être une tâche impossible. En pratique, une maladie n'a pas besoin d'être rare pour présenter des difficultés de recrutement de cas. Par exemple, notre étude pangénomique du VIH comprenait presque tous les individus infectés par le VIH enrôlés dans les cohortes d'histoire naturelle aux Etats-unis et ne comptait que 755 patients avec des dates connues de séroconversion – les cas les plus riches en informations. Bien que la prévalence du VIH soit extrêmement élevée dans certaines parties du monde, elle est beaucoup plus basse dans les sociétés occidentales où une majorité du financement et des infrastructures sont disponibles.

3.1. Modification des taux d'erreur de type I et de type II

Avec les récentes avancées technologiques et les efforts internationaux pour identifier le spectre de variation génétique dans de multiples populations humaines, les puces de SNPs sont devenues à la fois plus denses et contiennent une gamme plus large de fréquence d'allèles. Les puces les plus denses permettent une résolution plus fine de la structure haplotypique et le potentiel pour la cartographie fine des variants causaux. En raison des nombreuses p valeurs générées dans les GWAS, les puces denses ont exacerbé le désavantage de puissance des études cas-contrôles. Dans une tentative d'augmentation de la puissance des corrections de Bonferroni requises dans un scénario générateur d'hypothèses, de nombreux seuils de significativité et stratégies alternatives ont été recommandés (Benjamini & Hochberg, 1995; Misawa et al., 2008; Skol, Scott, Abecasis, & Boehnke, 2006; Storey & Tibshirani, 2003). Certains ont cherché à tirer particulièrement avantage des SNPs localisés dans les blocs haplotypiques avec un niveau élevé de liaison (Duggal et al., 2008). Cependant, à moins d'être parfaitement liés, ces SNPs conservent un certain niveau d'indépendance

statistique et doivent être comptés comme tests indépendants, et comme démontré en Section 2.1, les alternatives au seuil de significativité standard de 5×10^{-8} offrent seulement une modeste amélioration en puissance (Johnson et al., 2010). En dépit d'un consensus général établissant le seuil de significativité standard à 5×10^{-8} pour les GWAS, beaucoup demeurent insatisfaits des taux d'erreur de type II et cherchent une justification pour accepter davantage de résultats (Zaykin, Kuo, & Vsevolozhskaya, 2013). Bien que ce ne soit pas une mauvaise approche si l'on obtient la réplication indépendante des résultats, la puissance statistique est plus grande lorsque tous les échantillons sont compris ensemble, plutôt que d'organiser une validation indépendante (Skol et al., 2006).

Un passage en revue de tous les résultats GWAS publiés sur <http://www.genome.gov> a révélé que 75% des résultats qui ne passent pas le seuil significatif $p < 5 \times 10^{-8}$ n'ont pas été validés dans les études de réplication suivantes. Par comparaison, seulement 28% des résultats GWAS dépassant ce seuil conservateur n'ont pas été répliqués (Hindorff et al., n.d.). Le rapport de résultats significatifs limités est clairement l'une des causes principales du manque de reproductibilité dans les résultats GWAS à ce jour. Étant donné le coût des erreurs de type I en matière de temps et d'argent dépensés par les autres laboratoires pour répliquer les résultats, beaucoup questionnent la valeur de ces études, surtout lorsque l'importance des conséquences est prise en considération (Crow, 2011; Goldstein, 2009; McClellan & King, 2010). D'autres, cependant, ont noté que la découverte de gènes de GWAS a conduit à de nouvelles cibles médicamenteuses, a identifié des chemins pathologiques insoupçonnés auparavant et a apporté de nouveaux traitements.

La régulation des seuils de significativité dans les études de cas-contrôles demeurera importante, spécialement alors que nous entrons dans l'ère de l'analyse de séquence du génome entier, où le nombre de tests indépendants, comprenant des variants rares, peut être beaucoup plus élevé que dans une GWAS traditionnelle. Il est largement accepté que le moyen le plus efficace de réduire la publication d'associations de faux positifs est l'identification de variants causaux fonctionnels apparentés ; ou à défaut, la réplication de résultats dans une atmosphère collaborative avant la publication (GAIN Collaborative

Research Group et al., 2007). Reconnaissant le problème à la fois des faux positifs et des faux négatifs, de nombreux groupes sont en train de former de larges consortiums internationaux pour augmenter le nombre d'échantillons, accroissant ainsi la puissance pour détecter les variants à basse fréquence et/ou de faible effet qui n'atteindraient pas la significativité statistique dans de plus petites études (McLaren et al., 2013; Schork, Greenwood, & Braff, 2007; Wellcome Trust Case Control Consortium, 2007). Cependant, il est fondamental que la confirmation de la significativité statistique par réplication nécessite des méthodes exactes pour estimer la significativité dans les études individuelles et combinées. Dans les GWAS et les études similaires, cela nécessite d'estimer avec précision le nombre véritable de comparaisons indépendantes. Les simulations présentées dans cette thèse montrent que la méthode simple employant une analyse en composantes principales (CP) des données génétiques échantillonnées pour évaluer le nombre de tests indépendants offre une estimation du nombre de comparaisons qui n'est ni trop peu ou trop conservatrice, alors que d'autres méthodes publiées peuvent s'avérer dangereusement trop peu conservatrices.

3.2. Cartographie par déséquilibre de liaison dans des populations mixtes

Une option sous-exploitée pour augmenter la puissance est la cartographie par déséquilibre de liaison dans des populations mixtes (MALD), qui combine la puissance tirée du déséquilibre de liaison de l'ascendance à long terme (ALD) semblable à celle trouvée dans les études familiales avec la capacité d'études cas-contrôles pour modéliser la maladie complexe (McKeigue, 1997). Il est important de noter que la MALD présume l'existence d'un facteur génétique causal contribuant au brassage plus fréquent dans une population ancestrale que l'autre. Ceci est un scénario beaucoup plus vraisemblable lorsque le phénotype s'isole significativement dans les populations ancestrales. Cette présomption ajoute à la puissance de la MALD parce que l'association recherchée n'est pas entre le phénotype et un marqueur spécifique dans le déséquilibre de liaison (DL) avec le variant génétique fonctionnel, mais plutôt une association entre le phénotype et l'ascendance locale au niveau du locus abritant le variant génétique fonctionnel (Chakraborty & Weiss, 1986). Notamment, les variants *APOL1* affectant profondément les maladies rénales sont uniquement trouvés chez les Africains et des

populations mixtes avec une ascendance africaine récente (Freedman et al., 2011; 2010; Genovese et al., 2010; Kopp et al., 2011). On a trouvé depuis que ce gène est associé à d'autres maladies rénales y compris des maladies rénales de phase terminale non diabétiques, mais cette association n'a pas été trouvée par de précédentes études GWAS enrôlant beaucoup plus de monde parce que les variants causaux n'étaient pas couverts par les marqueurs de ces GWAS (Bostrom et al., 2010; Genovese et al., 2010; McDonough et al., 2011). Nous avons pu identifier le locus parce qu'il montrait un excès d'ascendance africaine comparé au reste du génome ou comparé au groupe contrôle de ce locus. L'absence de couverture complète de la diversité génétique dans les populations non européennes a commencé à être traitées par le projet 1000 Génomes, mais le séquençage d'individus dans chaque étude est le meilleur moyen de capturer la véritable portée de la diversité génétique humaine (Manry & Quintana-Murci, 2013).

Plusieurs très bonnes méthodes existent pour l'analyse de marqueurs d'information ancestrale (AIM) peu abondants et sont toujours très utilisées pour la découverte de gènes (McKeigue, Carpenter, Parra, & Shriver, 2000; Patterson et al., 2004; Pritchard, Stephens, & Donnelly, 2000). D'autres algorithmes sont capables d'analyser des ensembles de marqueurs denses avec des DL locaux, mais soit partagent les données de marqueurs en blocs haplotypiques discrets pour éviter la complexité informatique de modélisation des DL d'ordre supérieur, soit ils sont informatiquement inefficaces lorsqu'ils sont augmentés proportionnellement pour analyser de grands ensembles de données (Seldin et al., 2011).

ALDsuite, décrit dans cette thèse, tente d'éviter ces problèmes en modélisant un DL d'ordre supérieur en utilisant les composantes principales. Ceci permet non seulement un gain d'efficacité informatique grâce au fenêtrage, mais modélise aussi plus correctement les DL d'ordre supérieur qui peuvent s'étendre aux fenêtres voisines lorsqu'aucun bloc haplotypique clair n'existe. Sont également inclus dans ALDsuite la documentation complète d'entrées et de sorties de chaque fonction, les outils pour le contrôle qualité des données de génotypage, les outils pour le formatage de données et la préparation d'a priori requis pour l'analyse, et une

vignette couvrant deux analyses de données échantillons, depuis le formatage de données et la vérification jusqu'à l'analyse statistique et la préparation de tableaux et de chiffres.

4. Conclusion

La GWAS est un outil essentiel pour explorer la contribution génétique aux maladies humaines, mais convoit de graves problèmes de puissance statistique concernant les maladies pour lesquelles il n'est pas pratique d'échantillonner génétiquement des dizaines de milliers de sujets. Pour les maladies affectant les populations mixtes, où la maladie est substantiellement héritée d'une population ancestrale, l'analyse du déséquilibre de liaison dans les populations mixtes (ALD) est une stratégie pour parvenir à la puissance avec des nombres relativement faibles. Comme la plupart des populations d'étude seront typées de façon routinière pour un ensemble dense de marqueurs, l'utilisation d'ensembles complets de marqueurs est à la fois pratique et source de précision supplémentaire. Cependant, les algorithmes d'origine et les plus établis nécessitent un ensemble de marqueurs peu abondants pour éviter la confusion des mesures de déséquilibre des populations mixtes par le DL local. L'utilisation de marqueurs denses nécessite de prendre en compte le DL local mais ceci est difficile en raison de la haute dimensionnalité du DL. ALDsuite, présenté dans cette thèse, aborde ce problème avec l'ACP, corrigeant l'estimation d'ALD par les composantes principales majeures du DL ancestral. Les approches existantes alternatives à l'utilisation de marqueurs denses ont leurs limites, y compris les exigences informatiques qui ne s'adaptent pas facilement aux données très denses et aux plans de fenêtrages qui ne modélisent pas précisément le DL local dans des régions où les blocs haplotypiques discrets n'existent pas. ALDsuite ajoute en plus une nouvelle fonctionnalité en considérant des données d'associations pathologiques plus générales, en particulier les données de survie.

Pour des GWAS complètes de taille modeste, on ne peut échapper à la question de la puissance limitée face à $\sim 10^6$ comparaisons. L'absence de correction conduit à des résultats erronés, comme le montre l'incapacité à répliquer 75% des résultats issus des GWAS qui ne

respectent pas le seuil de significativité standard de $p < 5 \times 10^{-8}$. Pour minimiser les erreurs de type I et II, il est crucial de connaître la correction adéquate. Il a été suggéré de façon optimiste qu'en raison du DL, le nombre de comparaisons vraiment indépendantes est bien inférieur à $\sim 10^6$ SNPs testés, mais mes résultats n'appuient pas cette suggestion : les corrections utilisant le nombre de blocs haplotypique, par exemple, sont montrées par le test par permutation comme étant substantiellement trop peu conservatrices, même en considérant des critères grandement différents pour définir les blocs haplotypiques. D'un autre côté, la méthode simpleM, utilisant l'analyse en composantes principales (ACP), apporte une correction que le test par permutation montre comme étant exacte ; cette correction est modeste mais néanmoins utile.

Les résultats présentés ici - une correction efficace et nouvelle pour le DL de population ancestral local permettant l'utilisation de marqueurs denses avec le MALD et la démonstration que la méthode simpleM est effectivement optimale pour la correction de comparaisons multiple GWAS, réitère la valeur d'ACP pour capturer la part essentielle de la complexité des systèmes à nombreuses dimensions. L'ACP est déjà standard pour corriger la stratification des populations dans les GWAS, et mes résultats indiquent sa plus grande applicabilité comme stratégie générale pour traiter la haute dimensionnalité des données d'association génomiques.

La liaison est une part vitale des études d'association génétique et la modélisation appropriée du DL est nécessaire pour éviter les erreurs excessives de type I et le biais statistique. Une grande quantité de DL dans le génome humain n'exclut pas la nécessité d'une norme commune stricte pour la significativité pangénomique. Pour une GWAS traditionnelle, la réplication ou l'incapacité à répliquer 75% des résultats de la GWAS qui ne respectent pas le seuil de significativité standard de $p < 5 \times 10^{-8}$ appuie l'utilisation de ce seuil, mais d'autres méthodes, particulièrement la méthode simpleM utilisant une analyse de composantes principales (ACP) restera vraisemblablement pertinente pendant une bonne partie de l'ère de séquençage. Il y a de bons arguments envers d'autres seuils de significativité qui peuvent être appropriés dans certains cas, mais l'augmentation en puissance est généralement très modeste.

Une façon alternative d'augmenter la puissance, même face à des ensembles de données toujours plus importants, est une analyse du déséquilibre de liaison dans des populations mixtes (ALD). Cette forme de sous-structure de population peut être une source de biais significatif, résultant en une augmentation du taux d'erreur de type I, mais peut être efficacement contrôlée en utilisant une ACP des données génétiques. En fait, plutôt que d'être un inconvénient, l'ALD peut être une puissante source d'information pour identifier les régions du génome abritant les gènes pathogènes. Le milieu doit encore décider d'un bon algorithme pour déduire l'ascendance locale et ALDsuite répond à ce besoin. Il modélise les DL locaux sur des fenêtres flexibles en utilisant l'ACP, plutôt que des blocs haplotypiques discrets, permettant l'analyse de données de marqueurs à la fois denses ou peu abondants. Le logiciel est suffisamment efficace sur le plan informatique, bien documenté, et convivial, et il fournit les outils nécessaires pour préparer, analyser et présenter ces données complexes.

mots-clés : GWAS, association génétique, génome-entier, statistiques, correction, Analyse par composantes principales

Bibliographie

- Bateson, W., & Kilby, H. (1905). Experimental studies in the physiology of heredity. *Royal Soc Reports to the Evolution Committee*, 2, 1–55.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1), 189–300.
- Blackwell, D., & Hodges, J. L. (1957). Design for the Control of Selection Bias. *The Annals of Mathematical Statistics*, 28(2), 449–460. doi:10.1214/aoms/1177706973
- Bostrom, M. A., Lu, L., Chou, J., Hicks, P. J., Xu, J., Langefeld, C. D., et al. (2010). Candidate genes for non-diabetic ESRD in African Americans: a genome-wide association study using pooled DNA. *Human Genetics*, 128(2), 195–204. doi:10.1007/s00439-010-0842-3
- Browning, B. L. (2008). PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics*, 9, 309. doi:10.1186/1471-2105-9-309
- Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., et al. (2005). Demonstrating stratification in a European American population. *Nature Genetics*, 37(8), 868–872. doi:10.1038/ng1607
- Chakraborty, R., & Weiss, K. M. (1986). Frequencies of complex diseases in hybrid populations. *American Journal of Physical Anthropology*, 70(4), 489–503. doi:10.1002/ajpa.1330700408
- Crow, T. J. (2011). 'The missing genes: what happened to the heritability of psychiatric disorders?'. *Molecular Psychiatry*, 16(4), 362–364. doi:10.1038/mp.2010.92
- Duggal, P., Gillanders, E. M., Holmes, T. N., & Bailey-Wilson, J. E. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, 9, 516. doi:10.1186/1471-2164-9-516
- Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), 1567–1587.
- Freedman, B. I., Kopp, J. B., Langefeld, C. D., Genovese, G., Friedman, D. J., Nelson, G. W., et al. (2010). The apolipoprotein L1 (APOL1) gene and nondiabetic nephropathy in African Americans. *Journal of the American Society of Nephrology*, 21(9), 1422–1426. doi:10.1681/ASN.2010070730
- Freedman, B. I., Langefeld, C. D., Lu, L., Divers, J., Comeau, M. E., Kopp, J. B., et al. (2011). Differential effects of MYH9 and APOL1 risk variants on FRMD3 Association with Diabetic ESRD in African Americans. *PLoS Genetics*, 7(6), e1002150. doi:10.1371/journal.pgen.1002150
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576), 2225–2229. doi:10.1126/science.1069424
- GAIN Collaborative Research Group, Manolio, T. A., Rodriguez, L. L., Brooks, L., Abecasis, G., Collaborative Association Study of Psoriasis, et al. (2007). New models of collaboration in

- genome-wide association studies: the Genetic Association Information Network. *Nature Genetics*, 39(9), 1045–1051. doi:10.1038/ng2127
- Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., & Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, 34(1), 100–105. doi:10.1002/gepi.20430
- Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32(4), 361–369.
- Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, 329(5993), 841–845. doi:10.1126/science.1193032
- Goldstein, D. B. (2009). Common genetic variation and human traits. *The New England Journal of Medicine*, 360(17), 1696–1698. doi:10.1056/NEJMp0806284
- Gu, X. (2007, January 1). *Stepwise forward multiple regression for complex traits in high density genome-wide association studies*. Texas Medical Center. Retrieved from <http://digitalcommons.library.tmc.edu/dissertations/AAI3283834>
- Han, B., Kang, H. M., & Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, 5(4), e1000456. doi:10.1371/journal.pgen.1000456
- Hawkins, D. M. (1973). On the investigation of alternative regressions by principal component analysis. *Journal of the Royal Statistical Society, Series C*, 22(3), 275–286.
- Hendrickson, S. L., Jacobson, L. P., Nelson, G. W., Phair, J. P., Lautenberger, J. A., Johnson, R. C., et al. (2008). Host Genetic Influences on HAART Efficacy and AIDS-Free Survival. *Journal of Acquired Immune Deficiency Syndromes*, 48(3), 263–271.
- Hindorff, L. A., MacArthur, J., Morales, J., Junkins, H. A., Hall, P. N., Klemm, A. K., & Manolio, T. A. (n.d.). A Catalog of Published Genome-Wide Association Studies. Retrieved from <http://www.genome.gov/gwastudies>
- Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., & Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, 32(2), 179–185. doi:10.1002/gepi.20292
- Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2004). Design and analysis of admixture mapping studies. *The American Journal of Human Genetics*, 74(5), 965–978. doi:10.1086/420855
- International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, 437(7063), 1299–1320. doi:10.1038/nature04226
- International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851–861. doi:10.1038/nature06258
- Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., & O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics*, 11, 724. doi:10.1186/1471-2164-11-724

- Keightley, P. D., & Otto, S. P. (2006). Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107), 89–92. doi:10.1038/nature05049
- Kopp, J. B., Nelson, G. W., Sampath, K., Johnson, R. C., Genovese, G., An, P., et al. (2011). APOL1 Genetic Variants in Focal Segmental Glomerulosclerosis and HIV-Associated Nephropathy. *Journal of the American Society of Nephrology*, 22(11), 2129–2137. doi:10.1681/ASN.2011040388
- Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265(5181), 2037–2048.
- Manry, J., & Quintana-Murci, L. (2013). A genome-wide perspective of human diversity and its implications in infectious disease. *Cold Spring Harbor Perspectives in Medicine*, 3(1), a012450. doi:10.1101/cshperspect.a012450
- Matisse, T. C., Chen, F., Chen, W., La Vega, De, F. M., Hansen, M., He, C., et al. (2007). A second-generation combined linkage physical map of the human genome. *Genome Research*, 17(12), 1783–1786. doi:10.1101/gr.7156307
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5), 356–369. doi:10.1038/nrg2344
- McClellan, J., & King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, 141(2), 210–217. doi:10.1016/j.cell.2010.03.032
- McDonough, C. W., Palmer, N. D., Hicks, P. J., Roh, B. H., An, S. S., Cooke, J. N., et al. (2011). A genome-wide association study for diabetic nephropathy genes in African Americans. *Kidney International*, 79(5), 563–572. doi:10.1038/ki.2010.467
- McKeigue, P. M. (1997). Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *The American Journal of Human Genetics*, 60(1), 188.
- McKeigue, P. M., Carpenter, J. R., Parra, E. J., & Shriver, M. D. (2000). Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annals of Human Genetics*, 64(Pt 2), 171–186. doi:doi:10.1017/S0003480000008022
- McLaren, P. J., Coulonges, C., Ripke, S., van den Berg, L., Buchbinder, S., Carrington, M., et al. (2013). Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathogens*, 9(7), e1003515. doi:10.1371/journal.ppat.1003515
- Misawa, K., Fujii, S., Yamazaki, T., Takahashi, A., Takasaki, J., Yanagisawa, M., et al. (2008). New correction algorithms for multiple comparisons in case-control multilocus association studies based on haplotypes and diplotype configurations. *Journal of human genetics*, 53(9), 789–801. doi:10.1007/s10038-008-0312-0
- Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., et al. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nature Genetics*, 45(6), 690–696. doi:10.1038/ng.2608
- Morgan, T. H. (1910). Sex limited inheritance in drosophila. *Science*, 32(812), 120–123.

- Morgan, T. H. (1911). Random segregation versus coupling in medellian inheritance. *Science*, 34(873), 384.
- Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9), 981–990. doi:10.1038/ng.2383
- Nelson, G. W., Freedman, B. I., Bowden, D. W., Langefeld, C. D., An, P., Hicks, P. J., et al. (2010). Dense mapping of MYH9 localizes the strongest kidney disease associations to the region of introns 13 to 15. *Human Molecular Genetics*, 19(9), 1805–1815. doi:10.1093/hmg/ddq039
- Neyman, J., & Pearson, E. S. (1967). The testing of statistical hypotheses in relation to probabilities a priori. *Joint Statistical Papers*.
- Nicodemus, K., Liu, W., Chase, G., Tsai, Y., & Fallin, M. D. (2005). Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. *BMC Genetics*, 6 Suppl 1, S78. doi:10.1186/1471-2156-6-S1-S78
- Ohta, T. (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences*, 79(6), 1940–1944.
- Palaisa, K., Morgante, M., Tingey, S., & Rafalski, A. (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proceedings of the National Academy of Sciences*, 101(26), 9885–9890. doi:10.1073/pnas.0307839101
- Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., et al. (1998). Estimating African American admixture proportions by use of population-specific alleles. *The American Journal of Human Genetics*, 63(6), 1839–1851. doi:10.1086/302148
- Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., et al. (2004). Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5), 979–1000.
- Pólya, G. (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift*, 8(3-4), 171–181. doi:10.1007/BF01206525
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics*, 38(8), 904–909.
- Price, P., James, I., Fernandez, S., & Frencha, M. A. (2004). Alleles of the gene encoding IL-1 α may predict control of plasma viraemia in HIV-1 patients on highly active antiretroviral therapy. *AIDS*, 18(11), 1495–1501.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Riordan, J. R., Rommens, J. M., Kerem, B.-S., Alon, N., Rozmahel, R., Grzelczak, Z., et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922), 1066–1073.
- Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281), 1516–1517.

- Rommens, J. M., Iannuzzi, M. C., Kerem, B.-S., Drumm, M. L., Melmer, G., Dean, M., et al. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245(4922), 1059–1065.
- Schork, N. J., Greenwood, T. A., & Braff, D. L. (2007). Statistical genetics concepts and approaches in schizophrenia and related neuropsychiatric research. *Schizophrenia Bulletin*, 33(1), 95–104. doi:10.1093/schbul/sbl045
- Seldin, M. F., Pasaniuc, B., & Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 12(8), 523–528. doi:10.1038/nrg3002
- Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics*, 38(2), 209–213. doi:10.1038/ng1706
- Sober, E. (1993). *The nature of selection: evolutionary theory in philosophical focus*. Chicago: University of Chicago Press.
- Stephens, J. C., Briscoe, D., & O'Brien, S. J. (1994). Mapping by Admixture Linkage Disequilibrium in Human Populations: Limits and Guidelines. *The American Journal of Human Genetics*, 55, 809–824.
- Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.
- Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4), 676–682. doi:10.1101/gr.072850.107
- Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1), 1–12. doi: 10.1086/504302
- Toole, J. J., Knopf, J. L., Wozney, J. M., Sultzman, L. A., Buecker, J. L., Pittman, D. D., et al. (1984). Molecular cloning of a cDNA encoding human antihemophilic factor. *Nature*, 312(5992), 342–347.
- Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. doi:10.1038/nature05911
- Zaykin, D. V., Kuo, C. L., & Vsevolozhskaya, O. A. (2013, October 25). The more you test, the more you find: massive multiple testing does not promote spurious findings among top hits of association studies. *American Society of Human Genetics*. Boston, MA. Retrieved from <http://abstracts.ashg.org/cgi-bin/2013/ashg13s.pl?pgmnr=1789&sort=ptimes&sbutton=Detail&absno=130121288&sid=850896>

Table des matières

Remerciements	3
Résumé	4
1. Introduction	5
1.1. Liaison génétique ou déséquilibre de liaison.....	9
1.2. Liaison génétique dans des populations mixtes.....	10
1.3. Modélisation du DL par composantes principales	11
2. Résultats	12
2.1. Considération des comparaisons multiples dans une étude d'association pangénomique.....	12
2.2. ALDsuite: Marqueur dense MALD utilisant composantes principales de déséquilibre de liaison ancestrale	14
3. Discussion	14
3.1. Modification des taux d'erreur de type I et de type II	15
3.2. Cartographie par déséquilibre de liaison dans des populations mixtes	17
4. Conclusion.....	19
Bibliographie	22
Table des matières	27
Liste des tableaux	29
Liste des figures	30
Introduction	31
1. Genetic Linkage	36
1.1. From the birth of genetics to personalized genomics	37
1.2. Family Linkage Studies	40
1.3. Case / Control studies	43
1.4. Genotyping Technologies	57
2. Admixture Linkage.....	67
2.1. Admixture Linkage Disequilibrium.....	68
2.2. Ancestry Inference.....	69

2.3. Mapping by Admixture Linkage Disequilibrium	76
3. PCA Modeling of LD	82
3.1. Principal Components Analysis	83
3.2. Measurement of Population Substructure.....	87
4. Objectives	91
4.1. Objective 1	91
4.2. Objective 2.....	91
Results	93
5. Accounting for multiple comparisons in a genome-wide association study	94
6. ALDsuite: Dense marker MALD using principal components of ancestral linkage disequilibrium.....	102
7. Additional publications	110
Discussion	112
8. Modification of Type I and Type II error rates	115
9. Mapping by Admixture Linkage Disequilibrium	119
Conclusions	125
Bibliography	128
Appendices	148
A. Review of admixture publications from 2013	149
B. Currently available admixture software	152

Liste des tableaux

1 . Transmission Disequilibrium Test	42
2 . Example haplotype block definitions	44
3 . Null and alternative hypothesis probabilities	54
4 . Individual parameters in the HMM	69
5 . Population parameters in the HMM	70
6 . Individual and population parameter updates	72
7 . Hyperparameter updates	73
(1) . Summary of analysis results	98
(2) . Difference in significance threshold in a subset of the data	98
(1) . Currently available admixture inference software	105
8 . GWAS power calculations	112
9 . Review of admixture publications in 2013	146
10 . Currently available admixture software	154

Liste des figures

1 . Significant GWAS results	46
2 . False association from batch effect	47
3 . B allele frequency and log R ratio plots	48
4 . Example Manhattan plots	53
5 . Example QQ plots	53
6 . Automated Sanger sequencing read	57
7 . The PCR cycle	58
8 . Agarose gel	59
9 . Allelic discrimination plot	60
10 . IBD chromosomal segments following admixture	68
11 . Hidden Markov Model for ancestry inference	69
12 . Illustration of problems with relying solely on first order LD	74
13 . Disease gene mapping by admixture linkage disequilibrium	76
14 . Cancer incidence rate ratios	79
15 . PCA example in two dimensions	83
16 . Illustration of the first five PCs in a preliminary analysis of ESRD	84
17 . False associations from population substructure	87
18 . First two eigenvectors of a PCA illustrating population structure	88
19 . Parallel coordinates plot of 10 eigenvectors illustrating population structure..	89
(1) . Change in computation time and significance threshold for varying region size	99
(1) . Hidden Markov Model for ancestry inference	104
(2) . Representative chromosomes from one individual in the ASW population ...	105
20 . Replication rates of all GWAS results	114

Introduction

The discovery of genetic linkage by Thomas Morgan a century ago had immense theoretical and practical consequences. Gregor Mendel had observed that different traits are inherited separately; Morgan observed that certain traits tend to be inherited together. Morgan's observation led to the concept of genes arranged along a linear genome, which undergoes recombination during meiosis in successive generations. Practically, genetic linkage allows identification of genetic variation underpinning human diseases: strong association of disease with variation at any genomic locus suggests that the disease-causing variation is close to the observed locus. This general concept leads to many specific strategies for identifying genetic factors causing disease, and in all of them statistical issues are critical. Broadly, the strategies may be divided into family studies and GWAS. Mapping by admixture linkage disequilibrium (MALD) represents a hybrid approach between the two, with the ancestral populations of an admixed population playing the role of parents in a family study. This dissertation addresses statistical issues of GWAS and MALD, specifically demonstrating the ability of principal components analysis (PCA) to deal with issues arising from the complexity of genetic linkage.

Family studies were the first and most basic application of LD to disease gene identification. Family studies have been remarkably successful in assigning genes to chromosomal regions. The first disease genes mapped were for X-linked loss of function disorders that affected males but not female carriers (e.g. factor VIII deficiency, the cause of hemophilia (Toole et al., 1984)). The mapping of the autosomal gene responsible for cystic fibrosis was completed in 1989 (Riordan et al., 1989; Rommens et al., 1989). Since then the majority of Mendelian disorders with moderate to high penetrance have been mapped using large pedigree or multiplex linkage studies. These studies employ single nucleotide polymorphisms or microsatellite markers to test for statistical linkage between the segregating phenotype and markers. The underlying assumption is that the causal variant will be located on the chromosomal segment inherited by all the affected offspring. Positional cloning is then used to identify the responsible gene. Family studies have been extremely successful at mapping genes for highly penetrant phenotypes with single gene inheritance, but they cannot be used for complex and common diseases where all family members may not have the same genetic or environmental exposures or when penetrance is low.

Other methods of disease gene mapping depend on linkage disequilibrium (LD), a nonrandom association between alleles at nearby loci, determined by the balance of the evolutionary processes of mutation, genetic drift, and selection, which create association between alleles, and recombination, which breaks down the association. Because of the time/length scale of recombination (~ 1 recombination per 100 generations per megabase) distant loci are uncorrelated in a stable population, but statistically useful correlations between disease factors and marker polymorphisms over lengths of 1-100 kb abound (International HapMap Consortium et al., 2007). Genome-wide association studies (GWAS) exploit this correlation: genotyping of a sufficiently dense set of genetic variants distributed across the genome, makes it likely that a disease variant is in sufficient LD with one or more typed markers to create a statistically significant association between the marker and disease. The GWAS approach allows use of case-control data for common diseases, with complex causes, as opposed to Mendelian disorders caused by variation of a single gene.

While more general in applicability, case-control genetic association studies are less powerful than family studies, requiring much larger sample sizes for numerous reasons—penetrance, multiple tests, small effect size of most common variants, but also because they lack the large, identical-by-descent blocks shared by family members (Lander & Schork, 1994). Another drawback of case-control genetic association studies is their reliance on a control population sample (in family studies unaffected family members provide an internal control); the appropriate control group may not be obvious or difficult to recruit (Blackwell & Hodges, 1957). Case-control studies, on the other hand, are able to model complex, non-Mendelian disease association with common variants in LD with causal variants. Incomplete linkage, small effect sizes, low frequency variants, and multiple testing erode power to detect genotype-phenotype association (Lander & Schork, 1994). The most problematic of these issues is multiple testing: testing for association at $\sim 10^6$ markers requires correcting for a similar number of comparisons; the precise correction is debated, and is a subject of this dissertation, but in any case multiple comparisons create severe power issues for GWAS (Hoggart, Clark, De Iorio, Whittaker, & Balding, 2008).

As understanding of linkage developed, the effect of recent admixture on linkage patterns was recognized as a potential source of information to map genes associated with phenotypes that differ in prevalence between two ancestral populations. Admixed populations retain intact extended haplotypes from founding populations for many generations, creating an admixture linkage disequilibrium (ALD) which can be used to infer the genetic ancestry of a locus (Gabriel et al., 2002; McKeigue, 1997). In studies of disease with substantially different prevalence between the two parental populations, loci with an association between disease and local ancestry can be inferred to harbor disease genes. Since these intact haplotypes are much longer than the extent of population LD, ALD mapping substantially reduces the number of comparisons, as compared to GWAS, and where applicable can substantially increase power to detect disease associations (McKeigue, 1997). A technical problem, however, is that the population LD that provides the basis for GWAS is a confounding factor in identifying chromosomal blocks inherited from the ancestral populations.

This dissertation approaches the problem of power in genome-wide association studies (GWAS) through a study of genetic linkage between markers. A common thread of the difficulties encountered is the complexity of LD. This confounds simple estimates of the number of comparisons in a GWAS, as well as simple attempts to correct for population LD in MALD. In this dissertation I consider these issues and find that PCA offers a possibly optimum solution. This forms the basis of my two objectives, the first of which is to carefully review and test significance threshold methods for GWAS in an attempt to find the most accurate significance threshold for a desired study wide Type I error rate. Of the seven proposed methods tested, the Bonferroni correction for multiple comparisons is both the most conservative and the most commonly used method. Other methods tested either use various measures of linkage disequilibrium (LD) between markers to estimate the true number of independent comparisons being made, or they employ an optimized permutation testing scheme to estimate the significance threshold for the desired study-wide Type I error rate (B. L. Browning, 2008; Duggal, Gillanders, Holmes, & Bailey-Wilson, 2008; Gao, Starmer, & Martin, 2008; Han, Kang, & Eskin, 2009). Relaxing this significance threshold even moderately can have a significant effect on statistical power but can also inflate the study wide Type I error rate beyond the desired level.

The second objective is to more fully take advantage of information gleaned from admixture linkage disequilibrium (ALD) in studies of admixed populations. The first algorithms to perform such an analysis were developed before the advent of SNP chips with hundreds of thousands to millions of features and consequently do not allow the analysis of dense marker sets (Falush, Stephens, & Pritchard, 2003; Hoggart, Shriver, Kittles, Clayton, & McKeigue, 2004; Patterson et al., 2004). Current dense marker solutions tend to either be computationally inefficient when analyzing today's whole genome data sets, or they oversimplify the modeling of higher order LD (Seldin, Pasaniuc, & Price, 2011). A computationally efficient refinement of sparse marker algorithms, modeling higher order local LD indirectly using principal components, is proposed here, and a software implementation for using ALD to map disease genes is described.

1. Genetic Linkage

Genetic linkage is fundamental to genetic association studies as a means to narrowing the search for causal variants, by identifying a chromosomal region associated with disease. In rare cases, such as the discovery of the association of ACE with myocardial infarction, and the discovery of CCR5 $\Delta 32$ homozygosity blocking HIV infection, clear knowledge of the biological association of the gene with disease led directly to the discovery of the responsible genetic variation. When two markers are physically near each other on a chromosome, they are more likely to be inherited together, because there is a smaller chance of a crossover between the two during meiosis. The initial source of LD is mutation, in that a new allele arising at a locus by mutation necessarily occurs on a single chromosome, and is thus associated with all alleles carried on that specific chromosome (Bateson & Kilby, 1905; Morgan, 1910; 1911). In successive generations recombination breaks up this original chromosome, but even after 5000 generations—roughly the age of fully modern humans—chromosome segments of an average length of 20 kb will be inherited unbroken with probability less than 0.0001 (Matise et al., 2007). Random drift of allele frequencies, and selection for advantageous alleles against deleterious alleles contribute to this process, in a complex and extensively studied pattern (Keightley & Otto, 2006; Ohta, 1982; Palaisa, Morgante, Tingey, & Rafalski, 2004; Sober, 1993). As populations age, they acquire more variants and have more opportunity for recombination, which results in shorter blocks (or haplotypes) of LD, while younger populations tend to have longer LD blocks and fewer common variants (International HapMap Consortium et al., 2007). Populations that are geographically separated, or reproductively isolated for other reasons, will acquire different LD patterns over time, forming genetically distinct subpopulations with distinctive phenotypic characteristics. This population substructure is evident in the genetics of Europeans, for example, with individuals from different geographic regions exhibiting slightly different patterns of LD (P. Price, James, Fernandez, & Frencha, 2004).

LD structure between disease causing variants and nearby loci allows the researcher to infer the presence of the causal variant by virtue of its linkage to a nearby marker. This key

postulate of genetics is at the core of my work in various genetic association studies (B. I. Freedman et al., 2011; Hendrickson et al., 2008; Jeffrey B Kopp et al., 2011; Nelson et al., 2010). This also forms the basis for GWAS, which permits interrogation of the entire human genome using only a subset of markers. When a significant association is found between a genetic marker and the phenotype under study, an inference can be made that one or more causal variants in LD with the genetic marker have a functional effect on phenotype.

1.1. From the birth of genetics to personalized genomics

1.1.1. Birth of Genetics

In 1903, Walter Sutton postulated that chromosomes, known to segregate in a Mendelian fashion, contain hereditary units or genes (Sutton, 1903). Thomas Morgan greatly expanded upon this work using the fruit fly, *Drosophila melanogaster*, as a model and discovered the white-eyed mutant, a sex-linked, heritable trait. Because the trait was sex-linked, he hypothesized that the *white* gene was located on the sex chromosome and inferred that other genes are probably located on chromosomes as well (Morgan, 1910).

As Morgan's team began finding more mutants and comparing them together, they noticed that the miniature wing mutation was also sex-linked, but was not always inherited with the white mutation. This not only led to the idea of genetic linkage and crossing over of chromosomes, but he recognized that by measuring the amount of crossing over between the two genes, a measure of the distance between the two could be obtained (Morgan, 1911). This was closely followed by the first genetic map in 1913 (Sturtevant, 1913).

DNA was first discovered by Johann Friedrich Miescher in 1869, just a few years after Mendel's experiments concluded. He was able to identify the elemental makeup of DNA and recognized that DNA was a prime candidate for furthering understanding about heredity (Dahm, 2008). The significance of Miescher's discovery of DNA went unrecognized for some 75 years until, in 1933, Jean Brachet was able to show that DNA was present in chromosomes (Sapp, 1997), and nine years later Edward Tatum and George Beadle showed that proteins

were coded by genes. It wasn't until Oswald T. Avery, Colin MacLeod and Maclyn McCarty proved that DNA carries genetic information, however, that its full importance began to be realized (Avery, Macleod, & McCarty, 1944). Over the following decade, transposons were discovered, and the relationship between adenine/thymine and cytosine/guanine pairs was discovered (Chargaff, Zamenhof, & Green, 1950; McClintock, 1950). A few years later, Alfred Hershey and Martha Chase confirmed that DNA was the molecule responsible for heritability (Hershey & Chase, 1952). Many others were working on the problem at this time, including Rosalind Franklin and Maurice Wilkins, who had gathered X-ray data on the structure of DNA. Building on this work, James Watson and Francis Crick published their famous paper on the structure of DNA in 1953 (J. Watson & Crick, 1953).

1.1.2. Modern genetics

Continued work in the field of genetics lead to the discovery of restriction enzymes by Danna and Nathans over the decades following the publication on the structure of DNA. This new tool enabled scientists to cut and paste DNA for the first time, facilitating both the rapid growth of the field of genetics as well as the rise of the biotech industry (Roberts, 2005).

A year later the sequence of the first gene, the bacteriophage MS2 coat protein, was fully characterized (Min Jou, Haegeman, Ysebaert, & Fiers, 1972), and five years later the sequence of the first genome was published (Sanger et al., 1977). Sequencing and genotyping was extremely slow and labor intensive at the time, but in the mid 1980's the polymerase chain reaction (PCR) was developed, making the process much easier and paving the way for the current explosion of data generating technologies (Mullis et al., 1986; Saiki et al., 1985).

This helped facilitate the discovery of locus causing cystic fibrosis (CF), the first disease causing gene to be identified using genetics (Riordan et al., 1989; Rommens et al., 1989). CF has been shown to be caused by defects in the CFTR gene on chromosome 7, and over 1500 CF causing mutations have been identified since then (Bobadilla, Macek, Fine, & Farrell, 2002). The penetrance is nearly perfect, that is, inheritance of a mutation from each parent

absolutely causes disease, because it is a recessive Mendelian disease. This made the CFTR gene a relatively easy gene to find.

The first locus associated with cancer, BRCA1, was described a year later (Hall et al., 1990). It took a little more time to identify BRCA1 because of its dominant effect and intermediate penetrance. BRCA1 acts dominantly on breast cancer incidence, but has a more intermediate penetrance than the CFTR gene. There have been over 50 disease causing BRCA1 mutations discovered since the first causal mutations were reported in 1994 (Miki et al., 1994).

1.1.3. Personal genomics

The decade since the completion of the human genome project has seen the development of microarray technologies for gene expression, single nucleotide polymorphism (SNP) and copy number polymorphism (CNV) genotyping, and genome sequencing. The amount of available information continues to mount as genotyping technologies and statistical methodologies continue to progress, and the ability to locate disease genes has increased dramatically. These technologies are covered in more detail in Section 1.4.

The identification of disease causing variants has progressed from the relative ease of high penetrance Mendelian diseases, to the arduous identification of low penetrance genetic affects identified using GWAS, and now the field of personalized medicine is rapidly expanding. One obvious application is in personalized oncology. Many current treatments are available for specific mutations, both in the germ cell and somatic cell lines involved. Currently, an individual may be tested for the most common known mutations, but these next generation sequencing technologies will provide an opportunity to investigate the entire genome for a much broader range of mutations for less cost (Cronin & Ross, 2011).

Gene therapy is another field of application in personal genomics. While early attempts have resulted in some successes, there have also been some adverse outcomes including cancers induced by the virus used to deliver the modified genes and the death of one individual

(Sibbald, 2001; R. Weiss, 2005). This remains an active area of research with active clinical trials in the treatment of cystic fibrosis and HIV infection (Alton et al., 2013; Didigu et al., 2013). There has also been significant progress in treating chromosome 21 trisomy in vitro using the XIST gene to silence the third chromosome (Jiang et al., 2013). It is conceivable that scientific advances in the next few decades will make identification of very rare or personal variants with clinically actionable effects on disease for each individual an economically reasonable prospect.

1.2. Family Linkage Studies

The locus containing the CFTR gene discussed in Section 1.1.2 was discovered using family linkage studies (Knowlton et al., 1985; Wainwright et al., 1985; White et al., 1985). The most appealing aspects of family linkage studies have to do with the sample size required to achieve good statistical power. This statistical power comes from the large regions of identical by descent (IBD) haplotypes shared by closely related individuals, by virtue of their shared inheritance. Loci in a region containing a disease gene like CFTR or BRCA1 with a strong effect on disease outcome will also be associated to disease outcome because of their IBD linkage with the disease gene.

The most common way to test for an association between a locus and disease in family linkage studies, is to calculate the odds of recombination between sampled genetic markers and the disease gene being searched for. Given a genetic marker, A , and a disease gene, D , the log odds (LOD) of the two being linked is

$$(1) \quad \begin{aligned} \text{LOD} &= \log_{10} \left(\frac{P(\text{pedigree} | \text{linkage})}{P(\text{pedigree} | \text{no linkage})} \right) \\ &= \log_{10} \left(\frac{(1-\theta)^N \theta^R}{0.5^{N+R}} \right) \end{aligned}$$

where N is the number of offspring with both A and D , R is the number of offspring with A or D but not both, and θ is the recombination fraction, which can be estimated as $R / (N + R)$. The

marker with the highest LOD score is taken to be the marker most closely linked to *D*, and the magnitude of the LOD score gives a measure of statistical significance (Morton, 1955).

One assumption this statistic makes is that presence of the disease gene is perfectly associated with manifestation of disease. As is often the case, however, the disease gene does not cause disease in every individual. This is referred to as incomplete penetrance of the disease phenotype. Other relevant factors, including disease prevalence, mode of inheritance (i.e. dominant, recessive, etc...) and nearby genetic markers are also important to consider when modeling family linkage. A number of software packages exist to calculate LOD scores across the genome, which take these factors into account (Abecasis, Cherny, Cookson, & Cardon, 2002; Cottingham, Idury, & Schäffer, 1993; Fishelson & Geiger, 2002; Gudbjartsson, Thorvaldsson, Kong, Gunnarsson, & Ingolfsson, 2005; Kruglyak, Daly, Reeve-Daly, & Lander, 1996; Lange, Weeks, & Boehnke, 1988; Lathrop, Lalouel, Julier, & Ott, 1985; O'Connell, 2001; Ott, 1974), most using the Lander-Green algorithm, the Elston-Stewart algorithm or some combination of the two (Elston & Stewart, 1971; Lander & Green, 1987). There has also been renewed interest in this problem in recent years as the amount of information available in whole-genome sequencing studies is not computationally practical to analyze with any of the earlier software solutions (Silberstein et al., 2013).

Another computationally friendly algorithm, the transmission disequilibrium test (TDT), is also prevalent in family studies. The TDT relies on Mendel's law of segregation, stating that each individual possesses two alleles of each chromosome and that one will be passed on to offspring with equal probability. The TDT was first formally described by Spielman et. al. when a problem arose with some observed associations between Insulin-dependent Diabetes Mellitus (IDDM) and a 5' flanking polymorphism adjacent to an insulin gene on chromosome 11 (Spielman, McGinnis, & Ewens, 1993).

	non-Transmitted Allele
Transmitted	

Allele	A	A	Total
A	a	b	$a + b$
A	c	d	$c + d$
Total	$a + b$	$b + d$	n

Table 1. Transmission Disequilibrium Test. Table shows how individuals are accounted for in a TDT.

Given n children there are $2n$ parents. Each is assigned according to which allele is transmitted and which allele is not.

The simplest way to carry out a TDT is with family trios, consisting of an affected child and their unaffected parents. For linkage of a disease, D, at a given locus with alleles A_1 (risk allele) and A_2 (normal allele), a 2×2 table is set up as shown in Table 2, and the χ^2 statistic for association is

$$(2) \quad T = \frac{(b - c)^2}{b + c} \sim \chi_1^2.$$

McNemar's test incorporates a continuity correction, but is otherwise similar:

$$(3) \quad T_M = \frac{(|b - c| - 0.5)^2}{b + c} \sim \chi_1^2.$$

There are a few important assumptions and implications for this test. Firstly, because a homozygous parent will both transmit and not transmit the same allele, that parent will be uninformative in the test. Secondly, it is assumed that the parents will pass each allele randomly to their child, and the only significant associations observed will be due to the risk allele or something in linkage disequilibrium with the risk allele. This observation also allows us to devise an exact test for the TDT. The two sided exact TDT p-value will follow the binomial distribution, $B(b + c, 0.5)$, and will be equal to:

$$(4) \quad p = \begin{cases} 2 \sum_{i=b}^{b+c} \binom{b+c}{i} 0.5^{b+c}, & b > c \\ 2 \sum_{i=0}^b \binom{b+c}{i} 0.5^{b+c}, & b < c \\ 1, & b = c \end{cases}$$

There has also been some work to extend the TDT to families with multiple infected children. A parametric test can rapidly become complicated, as demonstrated by Martin et. al., but such tests are possible (Martin, Kaplan, & Weir, 1997). Another option is to use an MCMC approach (Cleves, Olson, & Jacobs, 1997).

1.3. Case / Control studies

While the extended linkage blocks add power to family linkage studies, they are limited to the study of Mendelian traits. Case - control studies, on the other hand, can be used to search for genetic variants involved in complex traits but rely only on local LD for linkage. This reliance on local LD comes with the requirement that many more markers must be tested in order to fully interrogate the entire human genome. Because of this and the relative high cost of genotyping, genome-wide association studies (GWAS) were not practical until SNP genotyping chips became prevalent. Prior to this, candidate genes were identified and a minimal set of SNPs in and around the gene of interest were carefully chosen. Post GWAS analysis has moved from testing candidate gene hypotheses to hypothesis generation, with followup studies of regions containing significant GWAS results to verify findings and identify the cause of the observed association.

1.3.1. Candidate gene association studies

Identification of linkage between genes and disease is much less straightforward in diseases with lower penetrance, usually because of the complex nature of these diseases. Most have environmental components that affect disease outcome and are impacted by multiple biological systems and pathways. When a candidate gene is identified, a careful analysis of

Definition	Reference
A contiguous set of markers with a minimum D' between each marker pair.	(Reich et al., 2001)
A block of markers where a small subset of haplotypes account for nearly all the observed variation.	(Patil et al., 2001; Kui Zhang, Deng, Chen, Waterman, & Sun, 2002)
A region with low levels of haplotype diversity.	(Zhao et al., 2003)
A region with both limited diversity and strong LD, with the possible exception of a few rare variants.	(Dawson et al., 2002)
A region with no historical evidence of recombination between any markers.	(N. Wang, Akey, Zhang, Chakraborty, & Jin, 2002)

Table 2. Example haplotype block definitions.

the genetic variants in the region is undertaken in cases and controls to identify possible modifiers of disease outcome.

Prior to the current sequencing technologies, great care was taken to minimize the number of markers to be genotyped. Haplotypes, defined as a haploid, multi-locus genotype, are often used to help optimize this minimal marker set. Haplotype blocks can be defined in a number of ways, but the unifying principle that they have a low inter-block recombination rate (see Table 2). As such, they have strong LD across the block and can be represented by very few genetic markers, referred to as haplotype tagging SNPs. Haplotype tagging is used both to increase power and decrease genotyping costs in candidate gene studies (Zhao, Pfeiffer, & Gail, 2003). These haplotype blocks can also be very useful in inference of population history (Tishkoff et al., 1996).

Data from the International HapMap Project can be useful when identifying haplotype tagging SNPs (International HapMap Consortium, 2005), but may not be sufficient in populations where good population data is lacking. Many populations in Africa, for example, retain a good deal of variability that has been lost by populations migrating out to Europe, requiring information not found in HapMap. In these populations, investigators can infer haplotype

blocks from sequence of a subset of the sample. Once haplotype blocks are identified, rather than genotyping the entire cohort using the full set of all identified SNPs, the smallest subset of SNPs uniquely identifying all haplotypes is chosen to represent the region (Martin et al., 2000). This allows the researcher to sample a larger group of individuals, reduce the number of multiple comparisons, and save time in narrowing down the search for disease causing variants. The HapMap project increased the efficiency of this method significantly by providing the LD profiles of various populations across the genome, and current bioinformatics tools exploiting these data can deliver good haplotype tagging SNPs for most regions with relatively minimal effort.

These studies are still performed when trying to identify the disease causing variant associated with a GWAS finding, but haplotype tagging as a means of minimizing the number of markers genotyped is becoming less common with the increased economy of targeted sequencing and increased availability of whole-genome sequence. The dense mapping of *MYH9* and *APOL1* (Jeffrey B Kopp et al., 2011; Nelson et al., 2010), for example, were follow up studies we performed after our observed association of that region with focal segmental glomerular sclerosis (Jeffery B Kopp et al., 2008).

1.3.2. Genome wide association studies (GWAS)

As previously discussed, genome wide studies have been carried out for decades in family studies of Mendelian diseases with high penetrance. As genotyping technology continued to advance, genotyping chips, discussed briefly in Section 1.4.6, became available with increasing numbers of SNPs. The first SNP chip with genome wide coverage was made available by the mid 2000's, with the first studies to use this technology published in 2005 . (Klein et al., 2005). As of November, 2013 there had been 1726 peer reviewed GWAS studies, with 14,767 findings reported (See Figure 1 for all SNPs reaching genome-wide significance) (Hindorff et al., n.d.).

Mapping disease genes by admixture linkage disequilibrium is another GWAS method which uses long-range linkage extant in recently admixed populations. The linkage blocks are

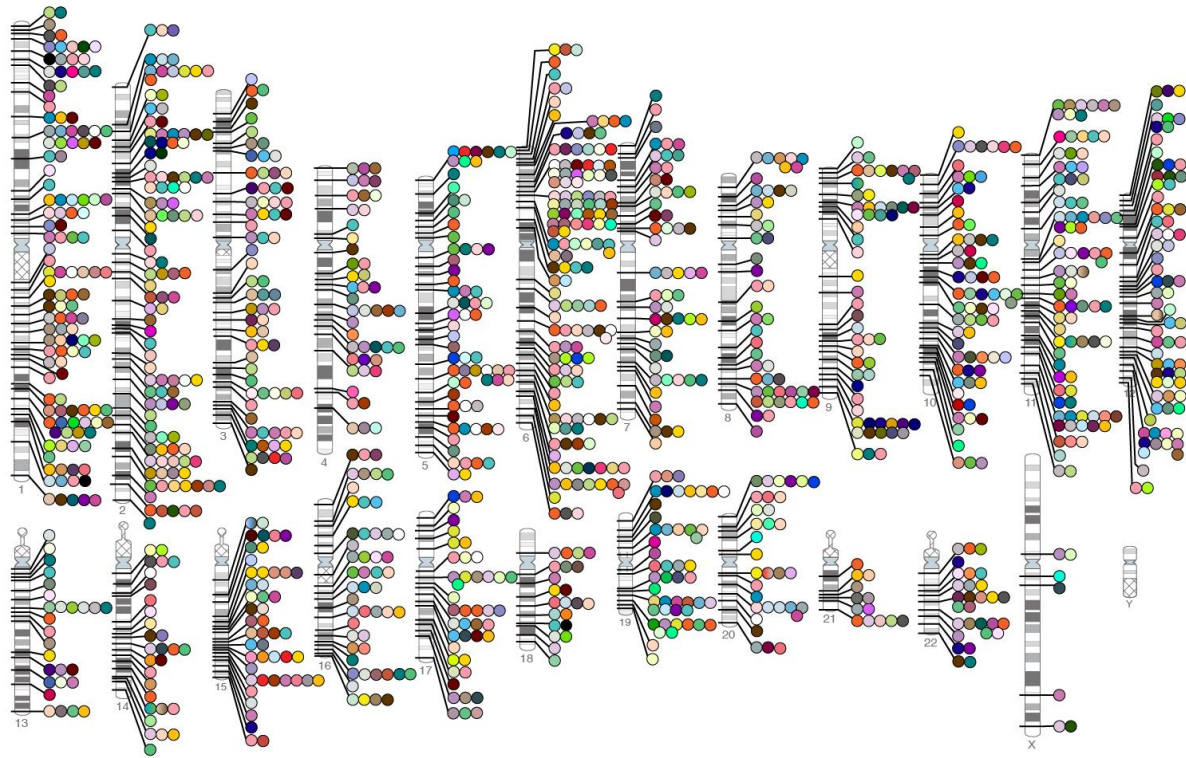


Figure 1. Significant GWAS results. Graphical representation of published SNPs reaching genome-wide significance (Hindorff et al., n.d.).

smaller than in family studies, but significantly larger than the local LD blocks found in non-admixed populations typically studied in case-control studies, resulting in a significant power advantage (Chakraborty & Weiss, 1988). In this section we will focus on traditional GWAS and treat this specialized study design in Section 2 after a more rigorous definition of admixture disequilibrium.

1.3.3. Quality control of GWAS

Batch effects

A batch of individuals in a GWAS consists of a group of samples which share some attribute of the genotyping process (e.g. day of the week they were processed, person who did the

sample processing, machine they were run on, month they were genotyped, etc...). A batch effect arises when samples in one group end up with a systematic difference in their genotyping data. Because no two sets of samples can possibly be run under exactly the same

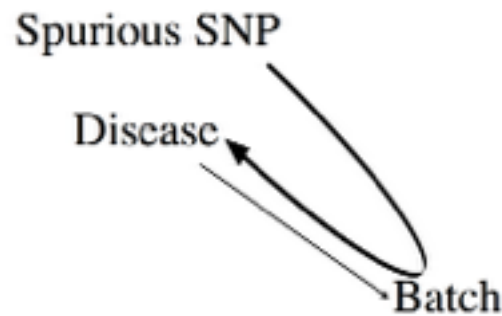


Figure 2. False association by batch effect. False positive associations induced when batch is associated with disease outcome.

circumstances, batch effects are common and can be associated with something as simple as the scanner or lot of reagents used. While batch effects are not desirable, a truly detrimental situation arises when the batch is associated with the disease outcome (Laurie et al., 2010). This can occur when the order samples are genotyped is dependent in some way on disease status. In this case, true associations may be minimized and any SNP that is randomly associated with the batch, will also be associated with disease outcome (see Figure 2).

The best way to avoid this scenario is to randomize samples as much as possible in the genotyping phase of the study. Randomization effectively breaks the association between disease and batch, also breaking induced associations. In the worst case, statistical models should be conditioned on known batch effects.

Checking genotyping quality

Genotyping quality scores are automatically calculated for each called genotype in most genotype calling software. Any SNP with a low call rate should be excluded from the analysis, as should any individual with low call rates, because they will likely introduce additional noise and bias to the data. Typical thresholds for call rates are between 3% and 5%

(Laurie et al., 2010). Low call rates among SNPs is indicative of a problem with the assay and low call rates among individuals is indicative of poor sample quality.

Checking Hardy-Weinberg Equilibrium (HWE) assumptions is another test which can identify genotyping errors. Some SNPs will not fall within HWE expectations simply as a natural result of frequency deviations, but research indicates that most of these deviations are due to genotyping error (Laurie et al., 2010).

Chromosomal abnormalities

Chromosomal abnormalities result in genotype data that does not adhere to the assumed distribution, which should be dropped from the data set. It is also possible for chromosome abnormalities to cause all data for an individual to be flagged for omission during the quality control of genotyping results, when a closer look would identify a single chromosome omission to be sufficient.

An analysis of the intensity scores for each SNP has been shown to provide both an excellent measure to automatically identify potential anomalies and graphical tools to identify the nature of the anomaly. Two measures can be obtained from the intensity scores after calling the genotypes: the BAAllele Frequency (BAF) and the log R Ratio (LRR). These two statistics

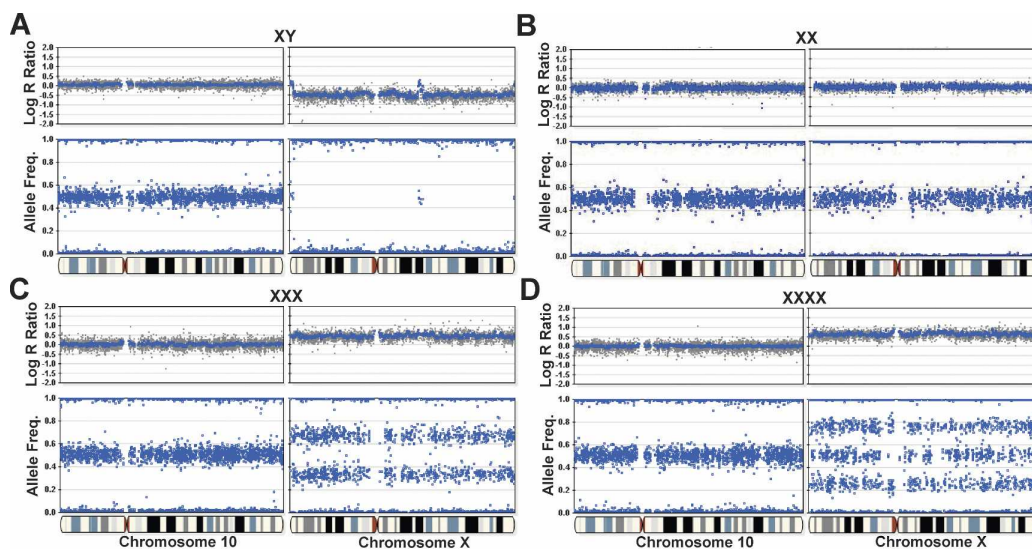


Figure 3. B-allele-frequency and log R ratio plots. Plots for a normal male and a normal female are shown in panels A and B, respectively, and plots for a female with three and four X chromosomes are shown in panels C and D. Chromosome 10 is placed beside the X chromosome for comparison purposes (Peiffer et al., 2006).

calculated by transforming the probe intensities, x and y , into polar coordinates and normalizing as described below (Peiffer et al., 2006). A change in the local median or the local variance of the BAF or LRR, defined as the median or variance of all statistics within 500 kb of a marker, can signal a chromosome aberration (see Figure 3).

Sample contamination

The B allele frequency (BAF), a measure of allelic imbalances, and the log R ratio (LRR) statistics described by Peiffer et. al. can also be used to identify sample contamination (Peiffer et al., 2006). The BAF is a measure of the mean allele frequency of a population of cells from a single sample, and has an easily recognizable distribution under normal circumstances with bands at 0, 0.5 and 1. The LRR is a measure of the relative allele intensity from the raw genotype calling data, and should be consistent across the genome of each sample. Figure 3 illustrates how the amount of contamination can be identified using BAF and LRR statistics. Normal individuals and individuals with an extra X chromosomes are shown here, but patterns similar to those observed will emerge when a contaminations exist (Peiffer et al., 2006). This also highlights the potential use of analysis of the BAF and LRR to identify large chromosomal abnormalities (Laurie et al., 2010).

Population substructure

Population substructure is a common source of bias. This source of bias can be easily controlled for with a principal components analysis (PCA) of the genetic data. A PCA starts with the full characterization of the covariance of the genetic profiles between each pair of individuals in the study, followed by the calculation of eigenvalues and eigenvectors of the covariance matrix (A. L. Price et al., 2006). Eigenvalues can be thought of as a measure of the amount of variation accounted for in the data by the associated eigenvector, and each

eigenvector can be used to summarize relationships in the data (see Section 3.2 for more details).

Race and gender check

Race and gender are probably the two most common covariates in the statistical analysis of biological data due to major genetic and epidemiological differences between populations. Clinical data are not always reliable and can introduce unnecessary bias into an analysis.

One of the easiest ways to use the genetic data available in a GWAS to check an individual's race, is through a PCA of the genetic data. Misidentified individuals will be readily identifiable when graphically examining population substructure. Gender misclassification is generally less common, but clerical mistakes and some types of chromosomal abnormalities can lead to the introduction of bias. Checking for excess X heterozygosity in males and loss of X heterozygosity in females, in addition to verifying the presence or absence of Y markers, is a simple way to check these assumptions (Laurie et al., 2010).

Checking for duplicate / related individuals

Duplicate individuals add nothing to a study, and related individuals can add significant bias unless analyzed with appropriate methods. Identification of duplicates is a simple process of comparing the genotypes from each pair of individuals, looking for any pairs that are near perfect matches to each other. Identification of parent-offspring relationships, full siblings, half siblings, and cousins can be accomplished through an analysis of identity by descent (Laurie et al., 2010).

1.3.4. GWAS Analysis

Statistical modeling in a GWAS is very similar to case-control studies of smaller scale. The main differences that arise are a direct result of testing millions of hypotheses. The three most important considerations to make are how to appropriately power the study, identify

deviations from modeling assumptions and ascertain statistical significance. Also discussed here are considerations for optimizing power and the cost of genotyping when under budgetary constraints.

Power considerations

As with the analysis of GWAS data, power calculations are similar to the power calculations of any other study, with a few important caveats. The first and foremost thing to consider is how significance will be assessed (see *Assessing Significance* below). Other important considerations are the sample size, the number of statistical tests, the statistic to be tested, the expected effect size, the frequency of the associated allele, and the distribution of the outcome data. Since most of these variables are not modifiable in the design of a study, the main emphasis is usually placed on gathering an adequate sample.

SNP imputation can be used to increase the power of a GWAS. Several options exist for imputing genotypes, including IMPUTE, fastPHASE and BEAGLE (S. R. Browning, 2006; Howie, Donnelly, & Marchini, 2009; Scheet & Stephens, 2006). Analysis of imputed genotypes needs to include the uncertainty of each genotype call, which can readily be accomplished in a frequentist setting by weighting each genotype by the corresponding posterior probability. Bayes factors can also readily accommodate this uncertainty (Marchini & Howie, 2010).

In some cases, particularly when the budget is limited, it may be useful to take a tiered approach to a GWAS. In a tiered approach, a subset of the cohort is fully genotyped and any significant SNPs are subsequently genotyped in the remainder of the cohort. This approach can provide significant cost savings, but will cause a decrease in statistical power. Skol et. al. developed a software package, CaTS, to minimize this loss of power while maximizing the benefit of limited budget funds (Skol, Scott, Abecasis, & Boehnke, 2006).

Visualization - Manhattan plots and QQ-plots

Careful thought should always be used in developing a statistical model for a particular analysis. Potential confounding effects in the population should be appropriately accounted for, and care should be taken to test the hypothesis best able to answer the question of interest. In a GWAS setting, individual attention cannot be paid to each test, so visual tools have been developed to help identify discrepancies.

Manhattan plots are a common way to visualize all test results across the genome, making significant results easily to identify. Significance is plotted in terms of $-\log_{10}$ p-value, as a function of genomic position. Because variants associated with a true cause of morbidity will be linked with other variants, causing a peak to rise from the background, single variants with high significance and no other linked significant variants can usually be flagged as false positives. Other anomalies may also appear that can be recognized using a Manhattan plot (see Figure 4).

QQ plots are another common way to visualize test results from the entire genome. SNPs are rank ordered and plotted as a function of the expected significance, based on an assumed random uniform distribution under the null hypothesis of no true associations. The corresponding inflation factor, λ , is the slope of the line fit to all results below the 90th percentile. These results can safely be assumed to be non-significant and give a good idea of how the statistical model fits the expected distribution under the null hypothesis. Values above 1 indicate anti-conservative results and potential problems with the statistical model or the data. Values below 1 indicate overly conservative results or inadequate statistical power. It is common to observe a dramatic departure from the expected distribution in the tail of the distribution. This is an indication that the results are more significant than expected under the null hypothesis, and without this it is difficult to assert relevance of the results (See Figure 5).

Assessing significance - Significance thresholds, False Discovery Rates, and Bayes Factors

The most common method of significance assessment in a GWAS is to use a Bonferroni threshold. Many have advocated for a uniform standard threshold of approximately 5×10^{-8} for

a study wide error rate of $\alpha = 0.05$. This comes from a number of careful analyses of commonvariation that have concluded there on the order of one million tests performed in a GWAS, which has gradually lead to the acceptance of this significance threshold for GWAS studies (Hoggart et al., 2008; International HapMap Consortium, 2005; Risch & Merikangas, 1996).

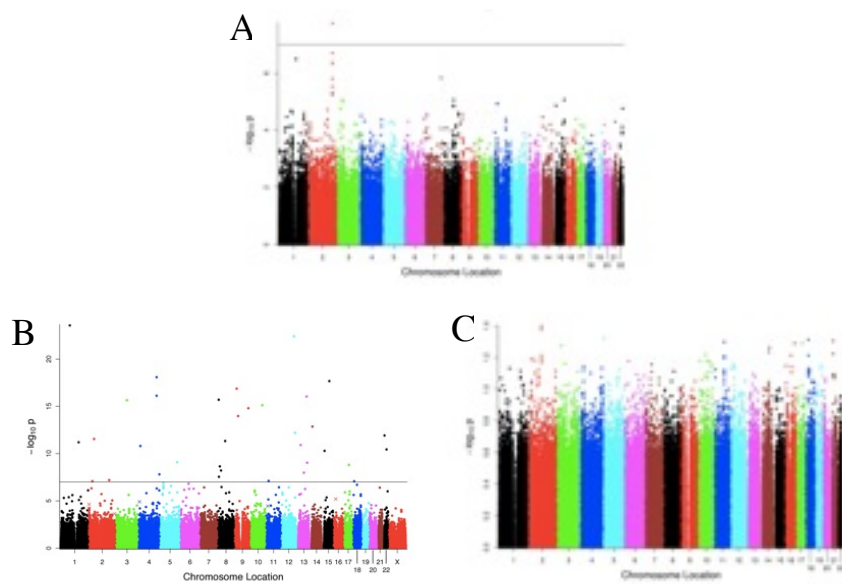


Figure 4. Example Manhattan plots. A) Typical Manhattan plot showing a peak on chromosome 2. B) Manhattan plot showing excess of significance. No clear peaks and many singletons, indicate problems with either the data or the model used. C) Manhattan plot showing lack of significance. Possibly due to low power, problems with the data or an inappropriate statistical model (data for these plots was simulated to represent the three scenarios represented).

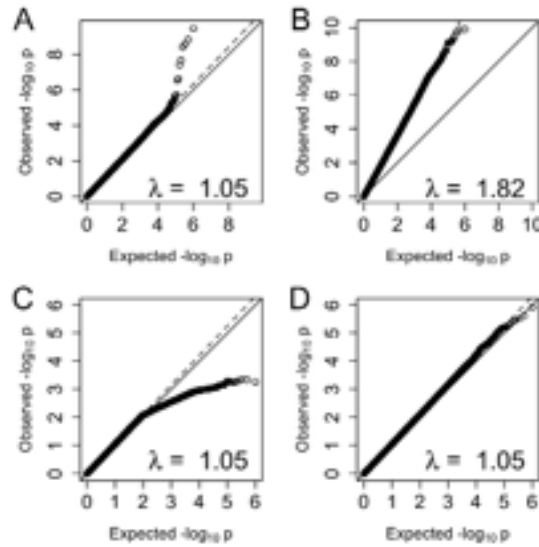


Figure 5. Example QQ-plots. A) Typical QQ-plot with λ inflation factor. The tail departs from the expected, indicating significance of test results more than expected by random chance. B) QQ-plot showing excess significance. Inflation factor is high, indicating a problem with the data or statistical model. C) QQ-plot showing lack of significance. Inflation factor is normal, but tail drops off, indicating inadequate power. D) QQ-plot adhering to the expected distribution. Inflation factor is very near 1 and no departure from the expected distribution is evident near the tail, indicating results are likely significant merely by chance.

A Bonferroni correction is easy to apply, but there are some noted problems with its application to a GWAS. One philosophical point made by Stern et. al. is that in the effort to increase objectivity, a significance threshold can stifle critical thought and interpretation of test results. Highly statistically significant results clearly indicate more believable results than those just passing the predetermined threshold, which are not substantively better than results narrowly missing this threshold. Inclusion of 95% confidence intervals has helped put some perspective on the interpretation of a test result, but these confidence intervals still rely on an arbitrary $\alpha = 0.05$ rule (Sterne & Davey Smith, 2001).

Another serious problem is the focus of these significance thresholds on Type I errors, falsely rejecting the null hypothesis, while ignoring Type II errors, falsely accepting the null hypothesis. This has led to the use of false discovery rates (FDR) by many and refined to include the Type II error rate in the false positive report probability (FPRP) (Benjamini & Hochberg, 1995; Wacholder, Chanock, Garcia-Closas, Ghormli, & Rothman, 2004). Given the

Type I and Type II error rates, α and β , and the proportion of expected true associations, π , an intuitive estimation of the FDR and FPRP can be calculated using Table 3 as a reference (see Equations 5 and 6). A statistic could be derived from the FPRP to include the respective cost of false positive and false negative results by weighting the Type I and Type II errors.

Truth of H	Outcome of the statistical test		Totals
	Reject H	Accept H	
H	$(1 - \beta) \pi$	$\beta \pi$	π
H	$\alpha (1 - \pi)$	$(1 - \alpha) (1 - \pi)$	$1 - \pi$
Totals	$(1 - \beta) \pi + \alpha (1 - \pi)$	$\beta \pi + (1 - \alpha) (1 - \pi)$	1

Table 3. Null and alternative hypothesis probabilities. Joint probabilities of the truth of the null hypothesis and outcome of the statistical test are given. α = Type I error rate, β = Type II error rate, π = proportion of true associations.

$$(5) \quad \text{FDR} = \alpha(1 - \pi)$$

$$(6) \quad \text{FPRP} = \alpha(1 - \pi) + \beta\pi$$

The q-value is another refinement to the FDR and is similar to a p-value, except instead of measuring significance in terms of the Type I error rate, it measures significance in terms of the FDR. Thus, a frequentist interpretation of the q-value could be the probability of a statistical test, as or more significant than the current test, being a false discovery. The distinction between the two is that the FDR relies on the practitioner supplied value, π , while the algorithm for generation of q-values estimates this proportion based on the distribution of test results above the $\alpha = 0.5$ level, assuming that all such tests are truly null (Storey & Tibshirani, 2003).

Another approach to this problem, the use of Bayes Factors (BF), takes both error rates into account in a Bayesian decision theory approach, which can also include the cost of Type I and Type II errors. A Bayes Factor is the ratio of the probability of the data under the null and alternative hypotheses:

$$(7) \quad \text{BF} = \frac{P(\mathbf{y} | H_0)}{P(\mathbf{y} | H_1)}.$$

The calculation of a Bayes factor requires the specification of a prior distribution, which can be difficult in some instances. An asymptotic Bayes factor (ABF), requiring only the specification of the prior distribution of the test statistic, has been proposed as well (Wakefield, 2007). Given a logistic regression maximum likelihood estimator (MLE), $\hat{\theta}$, with a prior distribution $N(0, W)$ and variance, V ,

$$(8) \quad \text{ABF} = \sqrt{\frac{V+W}{V}} \exp\left(\frac{-z^2}{2} \frac{W}{(V+W)}\right),$$

where $z^2 = \hat{\theta}^2 / V$ is the usual Wald statistic. The main difference between the ABF and a Wald statistic is that the ABF relies on both the test statistic, z , and on the power of the test, via V .

The reliance of Bayes factors on the power of the test has a few consequences that need to be considered. The major advantage is that V not only relies on the sample size, as the FPRP does, but also implicitly relies on the minor allele frequency for each individual test. Bayes factors also do not have a one to one relationship with p-values, unless the appropriate prior is chosen (Marchini & Howie, 2010).

Perhaps the best method of assessing significance is to use both frequentist and Bayesian measures. This approach includes the consideration of p-values in accordance with standard practice, but also reduces the FPRP through the use of q-values and Bayes factors (Wakefield, 2008).

Replication and validation

While initial results from GWAS were exciting, it quickly became evident that with the massive number of tests being performed, more needed to be done to control the number of false positive findings. Stringent significance levels are demanded by many journals, and some sort of replication or biological evidence is required before publishing results. In deed, a common practice is to include both independent replication results from similar cohort(s) and some function biological evidence validating the inference (Limou et al., 2009; Troyer et al., 2011).

1.4. Genotyping Technologies

This section contains a brief overview of the development of genotyping and sequencing technology, beginning with Sanger sequencing in 1975 and ending with a short discussion of some technologies currently under development.

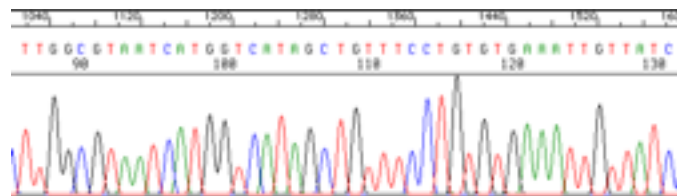


Figure 6. Automated Sanger sequencing read. Peaks show the relative fluorescence observed as each band of DNA passes the laser.

1.4.1. Sanger Sequencing

Sanger sequencing was first introduced in the 1970's and with the exception of the automation of the process, has changed little (Sanger & Coulson, 1975). The first step of Sanger sequencing is amplification of the DNA. This is done using DNA primers to ensure the correct portion of the genome is amplified. Trace amounts of fluorescently labeled dideoxynucleotides are added to the four basic nucleotides, each base fluorescing with distinct color. These dideoxynucleotides, when added to the elongating strand of replicating DNA, cause polymerase to disassociate and stop the reaction. These copies of the DNA are terminated at every possible base pair location along the length of the region to be sequenced.

The second step requires passing the DNA fragments through a gel or capillary array. This sorts the DNA fragments by length, and the fluorescence can be read as the DNA passes through a computer operated laser (see Figure 6). This process is limited to about 1 kb of DNA sequence per run.

In order to sequence long stretches of DNA, libraries of bacterial artificial chromosome (BAC) clones containing 100 - 300 kb of contiguous human DNA would be generated at great cost. DNA from these BAC clones would then be sequenced via the primer walking method (Stone et al., 1996). Primer walking is used to sequence a contiguous stretch of DNA, which is longer than the 1 kb of DNA that can be sequenced at one time using Sanger sequencing. When primer walking, the initial kb of DNA is sequenced in the normal fashion, and a new primer is chosen from near the end of the known DNA sequence. This primer is used to sequence the next kb of the region and the process is repeated until the end of the desired region is reached. This does not require the individual sequences to be reassembled after sequencing, but is significantly more labor intensive. This method is now abandoned in favor of newer technologies discussed in Section 1.4.7 which use modern computer algorithms and hardware to reassemble an entire genome's worth of reads at one time, decreasing the time and cost (Voelkerding, Dames, & Durtschi, 2009).

1.4.2. PCR

The polymerase chain reaction (PCR) is a process used to quickly make multiple copies of the selected region of DNA, that became available in the mid 1980s (Mullis et al., 1986; Saiki et al., 1985). The PCR reaction is a three step cycle using a thermo-stable DNA polymerase, typically Taq polymerase from the bacterium *Thermus aquaticus* (see Figure 7). The cycle begins with heating the desired DNA sequence to 94°-96° C to denature it. Cooling to 50°-65° C allows the primers to anneal to the DNA. Heating the reaction to an optimal temperature, typically 72° for Taq polymerase, provides the conditions needed for DNA polymerase to elongate each DNA strand.

The amount of PCR product is doubled by the end of each cycle when it is carried out under the proper circumstances, and therefore, increases exponentially with each cycle. When the amount of reagents or enzyme is insufficient for the desired number of cycles, however, the reaction will slow to a linear or log-linear pattern of growth as the reaction consumes reagents, and will eventually plateau when the reagents or enzyme are completely exhausted.

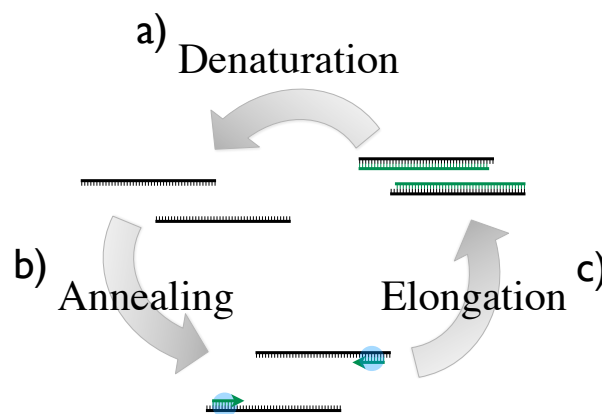


Figure 7. The PCR cycle: a) Denaturation, b) Annealing of the primers, c) Elongation of the desired fragment.

PCR is a powerful laboratory technique, but it is also highly sensitive to contamination. If contaminating sequence is introduced during the preparatory phase, it can bias the results under the best circumstances and can lead to the amplification of the wrong sequence under the worst circumstances.

1.4.3. Microsatellites and RFLP

A microsatellite, or a short tandem repeat (STR), consists of a 2-6 base pair DNA sequence that is repeated as many as 100 times or more. They occur frequently in most genomes and are highly polymorphic because they are prone to higher than average rates of mutation, usually because of slipped strand mis-pairing during DNA replication. They can be very good for identifying relationships between individuals and alleles of genetic markers. The process now used to identify microsatellites involves excision using restriction enzymes, amplification by PCR, and running through a gel electrophoresis. Microsatellites of varying sizes will result in differentiable bands on the gel.

Restriction fragment length polymorphism (RFLP) is another technique used to identify genetic variance between individuals. The general technique involves the use of restriction enzymes to fragment the DNA at loci where a specific sequence occurs. Because DNA is negatively charged, the resulting DNA fragments can be passed through an agarose gel using an electrical current in a process called gel electrophoresis (see Figure 8).

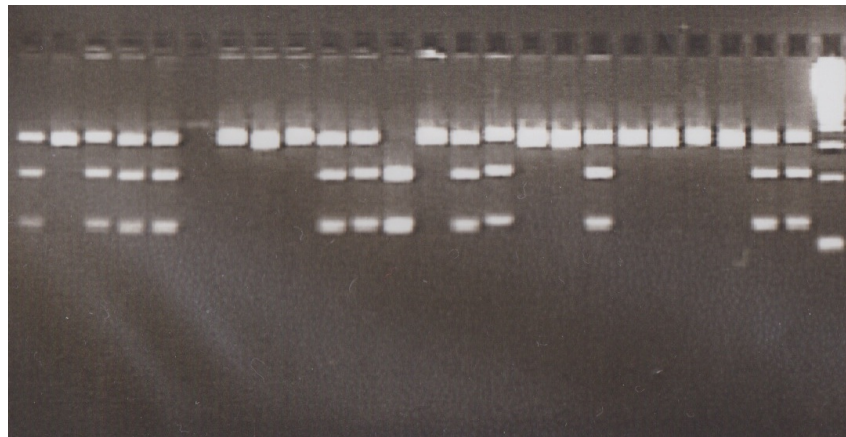


Figure 8. Agarose gel. An agarose gel showing RFLP bands after gel electrophoresis.

1.4.4. SNP genotyping and Real-time PCR

The genotyping of single nucleotide polymorphisms (SNPs) results in quicker, more accurate results than RFLP. SNP genotyping uses two paired primers, each matching only one of the two alleles of the SNP. Primers are chosen such that the length of the restriction fragments are distinguishable from one another, and after the fragments are amplified using PCR, they are run through a gel electrophoresis. Individuals with one band are homozygous for the given allele, and individuals with two bands are heterozygous.

In 1991, researchers at Cetus Corporation reported a novel method of PCR which allowed genotyping of SNPs in real time (Holland, Abramson, Watson, & Gelfand, 1991). Two TaqMan probes are added to the reaction, one straddling each variant. The TaqMan probes have a fluorophore attached to the 5' end of the probe and a quencher attached to the 3' end of the probe, preventing the fluorophore from fluorescing. During the elongation phase of PCR, the DNA polymerase interrupts the bond attaching the fluorophore, releasing it and allowing it

to fluoresce. Each probe fluoresces at a different wavelength, and groups are inferred for each genotype given the probe intensity data. An individual's genotype can be determined by the light intensity for each probe during the reaction, provided it falls within one of the clusters identified in the allelic discrimination plot (see Figure 9). Data points falling outside of the clusters are unable to be called with certainty and are removed from further analyses. Assays resulting in allelic discrimination plots with no clear clusters or with many uncalled individuals are flagged as bad and are either excluded from further analyses or re-genotyped.

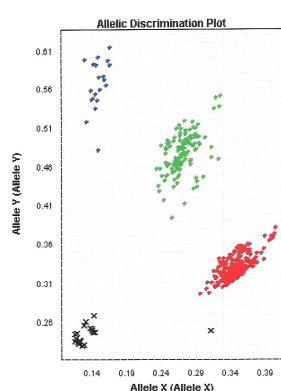


Figure9. Allelic discrimination plot. Individuals homozygous for the reference allele are shown in red, heterozygous individuals in green, and individuals homozygous for the alternate allele in blue. Data points marked with a black 'x' indicate individuals for which the assay results were inconsistent.

1.4.5. Gene Expression Arrays

A gene expression array is a powerful tool used to capture mRNA sequences and assess their relative abundance in the sampled tissue. Before sequencing, mRNA sequences are isolated and fragmented. The mRNA fragments are used to make cDNA sequence, which is then sequenced using microarray technology. These sequenced reads are then mapped back to the reference genome and relative abundance is calculated.

Because gene expression is tissue specific, multiple tissues and/or tumors from each individual can be profiled and compared. In one comparison of adult mouse brain, liver and muscle tissues, the abundance of *Myf6*, known to be highly specific to muscle cell function, is clearly differentially expressed in muscle tissue, compared to the other two cell types (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008).

1.4.6. SNP arrays

Several chips exist for this technology, each with a different number of markers sampled. The Affymetrix 6.0 array, for example, consists of 906,600 SNP probes and 946,826 non-polymorphic probes for detection of potential CNVs. The SNP probes are spaced an average of 3.2 kb apart along the genome (median spacing of 1.3 kb), and the non-polymorphic probes are spaced an average of 3.2 kb apart along the genome (median spacing of 2.2 kb). Each feature on the array is 5 μm^2 and each probe is replicated 3 - 4 times per feature. A number of markers are repeated and mismatches are included in different areas of the chip for quality control purposes. The Affymetrix technology requires 500 ng of genomic DNA and uses two restriction enzymes, Nsp I and Sty I, to digest the DNA into fragments. Each fragment is ligated to a four base pair (bp) overhang, regardless of the size of the fragment. At this point, a generic primer is used to amplify the adaptor-ligated fragments using a PCR protocol that has been optimized to preferentially amplify fragments ranging between 200 and 1,100 bp in size. The PCR product is then purified using activated beads, after which the amplified DNA is labeled and hybridized to the gene chip. After hybridization, the gene chip is run through the GeneChip Scanner to optically read probe intensities in preparation for genotype calling.

Illumina also has a number of SNP arrays, some of which are extensively customizable. Each HumanOmni2.5S bead chip, for example, genotypes 2,015,318 standard markers for each individual and up to 500,000 custom markers. The markers are spaced an average of 1.45 kb apart along the genome (median spacing of 0.79). Each marker is associated with a 3-micron silica bead which has been manufactured with an oligo unique to a specific marker. Beads are uniformly, randomly spaced about 5.7 microns apart on the chip, and each bead type is present an average of about 30 times on the chip, providing ample replication of each marker. After random self assembly, the beads are quality checked and the locations are decoded. Highly specific, dye-labeled oligonucleotides, complementary to the bead sequences, are added in a series of hybridizations. Each bead type has a unique sequence of dyes and is able to be uniquely identified at the end of the last stage of decoding hybridizations (Gunderson et al., 2004).

1.4.7. Next-generation sequencing

The Human Genome Project began an unparalleled journey to fully characterize the human genome in the early 1990's. In a report filed in 1989, it was estimated that it would cost \$200 million annually and was expected to take 15 years (Barnhart, 1989). The project officially started the following year, and after a few years was already behind schedule and over budget. In 1998, after some heated disagreements between several leading genome researchers over sequencing techniques, Celera, lead by Craig Venter, began a private genome project to compete with the public Human Genome Project.

The major differences of Venter's private venture was the use of shotgun sequencing in place of BAC library development (Venter, Smith, & Hood, 1996). The shotgun sequencing process begins with randomly breaking the genome into small fragments, which are then amplified and individually sequenced. At the completion of sequencing, however, the fragments need to be pieced together using a computer. Many leading researchers in the public effort were skeptical of the feasibility of shotgun sequencing, citing primarily the lack of computing power to reassemble all of the sequence reads once the sequencing was complete (Weber & Myers, 1997). As proof of principle, Celera started by sequencing the first multicellular eukaryote, *Caenorhabditis elegans* (C elegans Sequencing Consortium, 1998). A few years later the rough draft of the human genome was released in 2001, followed by its completion in 2003 (Lander et al., 2001; Venter et al., 2001). Shotgun sequencing has replaced primer walking and the use of BACs as the preferred method for genome wide sequencing, due to its significant time and cost savings.

As sequences are aligned, large overlapping regions, called contigs, begin to become clear, but the association of these contigs with each other can be difficult to ascertain. One refinement that was made early in the development of shotgun sequencing, is the sequencing of larger fragments, termed the scaffold. The ends of these DNA fragments are sequenced and included in the alignment. Because the distances between the paired ends can be measured, this provides a framework that can be used to paste otherwise disparate contigs together.

The current generation of sequencing methodologies has focused mainly on automation and parallelization of sequencing reactions. The first commercially viable massively parallel sequencing method was described by scientists at 454 Life Sciences Corp. in 2005 incorporating approximately 1.6 million simultaneous reactions in picoliter wells of a slide (Margulies et al., 2005). Since then, several other companies have marketed massively parallel sequencing technologies, including ABI's SOLiD platform, Illumina sequencing and Ion Torrent sequencing.

Sequencing an entire genome, while much more cost effective than just a few years ago, is still quite expensive to perform on each individual. By using exome capture arrays, significant savings can be realized. DNA samples are randomly fragmented into small fragments and hybridized to a microarray with as many as 200,000 exome probes. In this manner, DNA fragments containing known exome regions are captured and the rest of the DNA can be discarded (Choi et al., 2009; Ng et al., 2010; 2009). The captured exome is amplified, and the rest of the sequencing process proceeds using the same protocol as for a whole genome. Because of the significantly smaller size of the exome, a much greater read depth can be accomplished with a similar amount of starting material.

Alternatively, by sequencing RNA where the introns have been excised, the results give gene expression data as well as the expression of non-translated RNA sequences. The technology known as RNA-Seq gives an over-arching picture of what is going on within the cell by considering the different RNA sequences that are present and that are both time and tissue sensitive (Chu & Corey, 2012). This can be especially useful to see how the cellular pathways are affected by the progression of a particular infection like West Nile Virus or diseases such as lung cancer (Beane et al., 2011; Qian et al., 2013).

ChIP-Seq is another new generation sequencing approach that identifies protein-DNA interactions to understand the cellular interactions that are regulating changes in both normal or disease states (D. S. Johnson, Mortazavi, Myers, & Wold, 2007). A panel of known DNA

probes are used to capture the desired DNA-interacting fragments for amplification and sequencing. A whole map of these binding sites can be produced identifying a set of proteins that interact directly with the DNA for regulatory purposes and possible gene-therapy target sites (Marban et al., 2011).

A contemporary approach using sequencing to determine DNA methylation sites is receiving more attention as the importance of methylation in disease progression is becoming more fully recognized. Gene regulation is changed as a methyl group is added to the carbon-5 position of cytosine. As methylation patterns change over time, corresponding changes have been observed in gene expression of the genes in affected regions. In colon cancer, 10 genes were found to be hypermethylated in 100 percent of cases and many more genes were frequently changed (Schnekenburger & Diederich, 2012). By treating DNA with bisulfide, cytosine is converted to uracil when a methyl group is not present. This allows the identification or a map of epigenetic markers for the DNA region of choice. Further studies looking at methyl changes over nearly the entire human genome have used a methylation assay with a chip containing possible methylation sites spread over 14,495 genes to identify changes in methylation patterns (Weisenberger, VanDenBerg, Pan, Berman, & Laird, 2008). These technologies and new variations are allowing new disease models to be developed, and some longitudinal trait changes, some of which appear to be heritable, can now be better explained. For example, Soubry et al. found that obesity was associated with lower methylation levels at some genes while the newborns of obese parents had altered methylation patterns at multiple imprint regulatory regions (Soubry et al., 2013).

The next generation of sequencing technologies currently being developed focus mainly on the sequencing of single molecules and increasing the length of accurate reads (M. Xu, Fujita, & Hanagata, 2009). Very long sequencing reads from PacBio, while error prone, have already been effectively employed to improve genome assemblies, and increases in length and accuracy are regularly being made (Xiaojing Zhang et al., 2012). Oxford Nanopore has also reported some success with a micro sequencing platform that could fit on a USB stick and plugged directly into a laptop for sequencing (Eisenstein, 2012). These technologies will

likely result in cheaper, longer read lengths, providing solutions to current problems like the sequencing of highly structured regions. The Y chromosome, for example, has been very difficult to sequence in its entirety because it contains many palindromic regions and repeats (Francalacci et al., 2013).

2. Admixture Linkage

Recently admixed populations have much longer, extended LD blocks that form a mosaic of the ancestral populations (e.g. African Americans are primarily descended from West Africans with about 20% European admixture) (Parra et al., 1998). Extended haplotypes descend from each ancestral population and are slowly broken up through recombination during meiosis in each subsequent generation. Haplotype blocks consist of an entire chromosome inherited from each ancestral population in the initial admixed generation. The length of these blocks as they decay with each generation is a function of genetic distance, measured in centimorgans (cM). After one generation, for example, each extended haplotype block will vary, but the average block size will be approximately 100 cM or roughly one hundred million bases. The correlation of markers within these extended haplotype blocks is referred to as admixture linkage disequilibrium (ALD). Within these extended haplotype blocks, however, ALD is confounded by extant local LD retained from ancestral populations, because each extended haplotype is descended from one ancestral population (Sundquist, Fratkin, Do, & Batzoglou, 2008; Tang, Coram, Wang, Zhu, & Risch, 2006). A complex correlation structure arises from ALD and local LD, in which knowing the allele at one locus can give significant information about alleles of neighboring loci, and knowing the alleles in one haplotype block can give significant information about the ancestry of that locus.

ALD can also add meaningful information to the analysis of association studies using admixed populations. Instead of doing an association study looking for association between markers and a phenotype, one can look for linkage between an extended haplotype block derived from one or the other ancestral populations with a trait or phenotype that differs in frequency between the two ancestral parental populations. In other words, we anticipate that the causal variant is more likely to be harbored on the same ancestral chromosome as the ancestral population showing the higher disease prevalence (J. C. Stephens, Briscoe, & O'Brien, 1994). This is analogous to family studies that use long linkage blocks to identify disease mutations in affected families segregating a disease gene (McKeigue, 1997).

2.1. Admixture Linkage Disequilibrium

Linkage disequilibrium (LD) is the non-random association of the alleles from a group of loci, due to shared inheritance of the alleles over many generations, while admixture linkage disequilibrium (ALD) discussed above is the non-random association of alleles from each ancestral population over a few generations following admixture. Regions of high LD, often referred to as LD blocks, are not generally recombined during meiosis and are inherited together. This is most often due to the loci in the LD block being in close physical proximity. These regions vary somewhat in size between populations, mainly due to the size of the founding population and the number of generations since the population's founding. Africans, for example, have an overall lower level of LD than Europeans.

When there is admixture between two populations, individuals of the initially admixed generation (Generation 0) inherit one entire chromosome from each parent and thus inherit one entire chromosome from each ancestral population. The LD structure across each of these chromosomes can be quite different, making them easily distinguishable from each other. In the first generation after admixture (Generation 1) the contiguous chromosomal segments from each ancestral population are 100 cM long on average, which is quite large in comparison to the background identical by descent (IBD) segments originating from each of the ancestral populations. These chromosomal segments form extended haplotypes that are unique to each ancestral population, which can be described by local LD patterns specific to each ancestral population, resulting in considerable ALD in addition to the local LD present in the ancestral populations. Through succeeding generations, the IBD chromosomal segments resulting from admixture will become smaller by the process of genetic recombination (see Figure 10), but ALD continues to be significant and measurable for many generations (Lautenberger, Stephens, O'Brien, & Smith, 2000; Patterson et al., 2012).

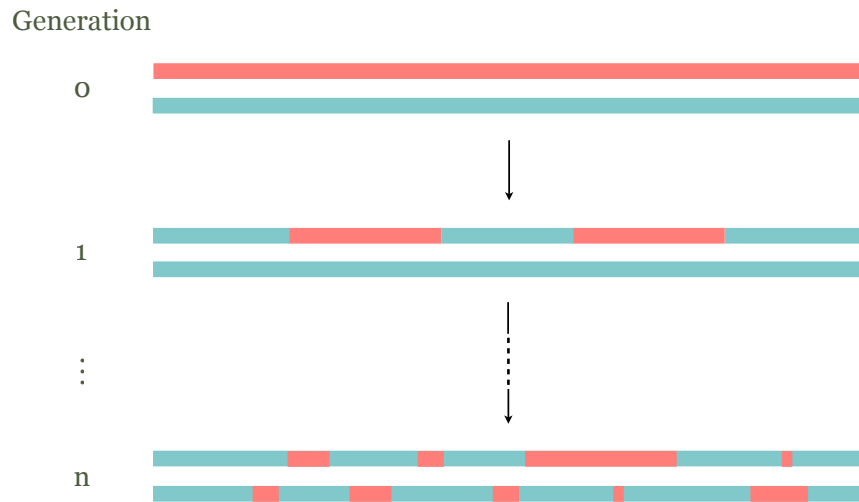


Figure 10. IBD chromosomal segments following admixture. Progression of identical by descent (IBD) chromosomal segments resulting from admixture are shown over n generations. Admixed chromosomes typical of individuals in initially admixed (Generation 0) and following generations (1 - n) are depicted. The ancestral population of each IBD segment is identified by color. Average IBD segment size decreases by way of genetic recombination each generation.

2.2. Ancestry Inference

Most established ancestry inference algorithms use a Hidden Markov Model (HMM) to model ancestry from genotypic data (see Figure 11). The three most established software packages, STRUCTURE / MALDsoft, ANCESTRYMAP and ADMIXMAP, use a similar HMM to model the likelihood of each ancestral state at each locus, conditional on unobserved parameters. This is used iteratively in conjunction with an MCMC algorithm to estimate the likelihood distribution of each individual's true ancestry profile (Falush et al., 2003; Hoggart et al., 2003; Patterson et al., 2004). A general outline of these algorithms follows in Sections 2.2.1 - 4. These algorithms do not model local LD from ancestral populations, but focus instead only on modeling of admixture LD. Because of this, they are only appropriate for analysis of sparse marker sets that are unlinked in ancestral populations. The general outline of these sparse marker algorithms is followed by a discussion of dense marker strategies in Section 2.2.5.

Parameter	Definition
$A \sim \text{Dirichlet}(\omega_1, \dots, \omega_k)$	• Global ancestry
$A^X \sim \text{Dirichlet}(\omega^X A_1, \dots, \omega^X A_k)$	• Global X chromosome ancestry
$\lambda \sim \text{Gamma}(\alpha_1, \alpha_2)$	• Mean number of generations since admixture

Table 4. Individual parameters in the HMM. Parameters for individual global ancestry, on both the autosomes and X chromosome, and mean number of generations since admixture are given. Distributions for these parameters are also given as a function of study-wide hyperparameters.

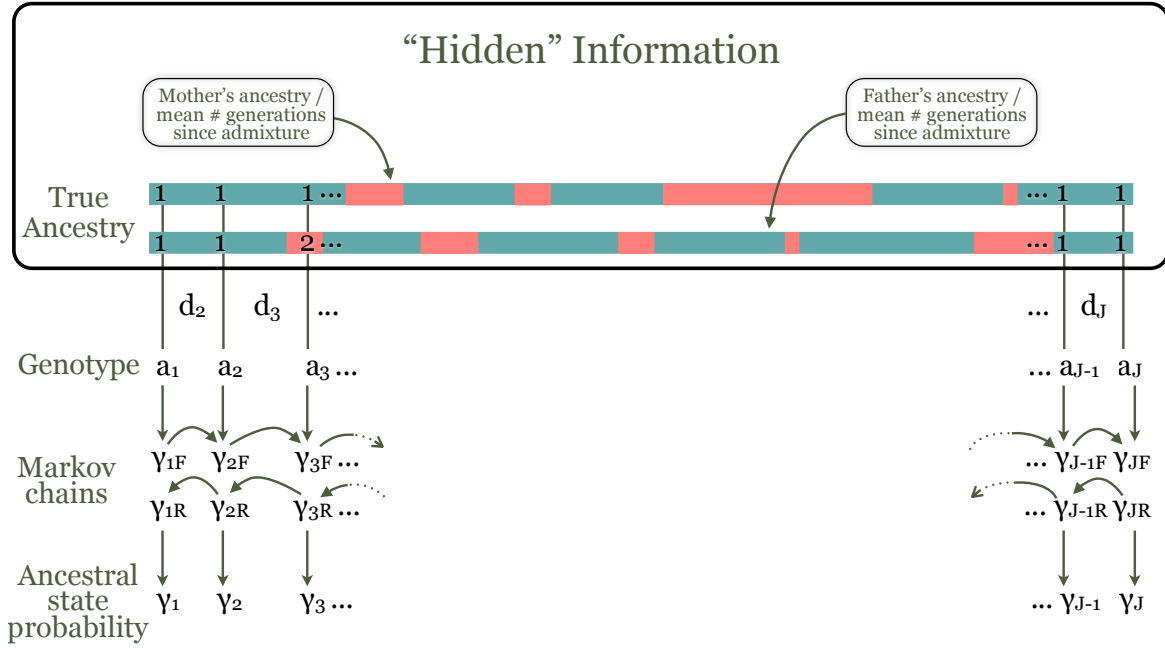


Figure 11. Hidden Markov Model for ancestry inference. An individual's ancestral chromosomal composition is a random process stemming from each parent's ancestral history since admixture. True ancestry of each chromosome segment is designated by color. Markers are chosen with d_j cM between the j^{th} and $j-1^{\text{th}}$ markers, and individuals are genotyped. Ancestral state probabilities are calculated conditional on the observed genotype, the ancestral state probability of the previous marker in the chain and other relevant parameters. Conditional probabilities are generated, one starting at each end of the chromosome, and are combined to obtain the overall ancestral state probabilities at each locus, conditional on the ancestral state probabilities of neighboring loci.

2.2.1. Setup

In setting up the algorithm, individual parameters for global ancestry, A , and mean number of generations since admixture, λ , are either estimated using maximum likelihood or randomly sampled from a non-informative or minimally-informative prior distribution describing the admixed population. Global admixture proportions are modeled using a beta distribution in the case of two-way admixture or dirichlet distribution, and the mean number of generations since admixture is modeled using a gamma distribution (see Table 4).

2.2.2. Step 1

The workhorse of these algorithms is the HMM described in Figure 11. The hidden ancestral state of each parent, represented by γ , is inferred using the observed genotypes at each sampled locus, conditional on genetic distance between markers (d), individual parameters defined in Table 4 and population parameters defined in Table 5. Markers with large allele frequency differences in ancestral populations are typically chosen to maximize the power to differentiate between the two (A. L. Price et al., 2007; M. W. Smith et al., 2004). Two Markov chains are propagated from each end of the chromosome, in which the probability of each allele having originated from a given ancestral population is calculated is conditional on previous markers in the chain. The two chains, one in the forward and one in the reverse direction, are then multiplied and normalized to obtain the final ancestral state probabilities (Patterson et al., 2004; Pritchard, Stephens, & Donnelly, 2000).

Parameter	Definition
$p_k \sim \text{Beta}(\tau_k P_k, \tau_k (1 - P_k))$	• Allele frequency in chromosomes descended from population k

Table 5. Population parameters in the HMM. Population parameters are defined for the allele frequencies in chromosomes descended from each population. Distributions for each population parameter are also defined in terms of study-wide hyperparameters.

The iterative portion of the algorithm starts with the evaluation of the HMM to obtain ancestral state probabilities, conditional on the unobserved parameters. A theoretical distribution of the ancestral state probabilities, γ , that is the probability of each chromosomal segment originating in a specific ancestral population, k , can be generated, conditional on the previously sampled parameters, A and λ , as well as the observed genotype, a , and ancestral states, γ_{prev} , of the surrounding markers. This follows from an application of Bayes Rule:

$$(9) \quad \gamma = P(g = k | a = x) = \frac{P(a = x | g = k)P(g = k)}{P(a = x)},$$

$$(10) \quad P(a = x | g = k) = \begin{cases} (1 - p_k)^2 & , x = 0 \\ 2(1 - p_k)p_k & , x = 1 \\ p_k^2 & , x = 2 \end{cases}$$

$$(11) \quad P(g = k) = A_k P(\text{recombination} | \lambda) + \gamma_{prev} (1 - P(\text{recombination} | \lambda))$$

$$(12) \quad P(a = x) = \sum_{k \in K} P(a = x | g = k)P(g = k).$$

Equation 10 is a function of p_k , the allele frequency in chromosomes descended from population k . The ancestral state probability, unconditional on genotype (Equation 11), is a function of the distance to the previous marker in cM, d , and the mean number of generations since admixture in that individual, λ . If there is a recombination with the previous locus, the probability of observing a chromosome from population k is simply the global proportion of chromosomes from population k in that individual's ancestors, A_k . If there is no recombination, then the probability is the same as that of the previous locus.

To start with, the ancestral state probability of the first marker on the chromosome, γ_{1F} , is calculated, conditional only on the individual's global ancestry, A , and the observed genotype at that locus, essentially setting the probability of recombination to 1 in Equation 11. The ancestral state probability, γ_{jF} , of subsequent markers on the chromosome can then be

calculated, conditional on the ancestral state probability at the previous locus, γ_{j-1F} , as well as the individual's observed genotype and global parameters A and λ . The same process is repeated starting with the marker at the opposite end of the chromosome and moving in the reverse direction to calculate γ_R . The final ancestral state probability at each locus is proportional to the product of γ_F and γ_R , which can be normalized to sum to 1:

$$(13) \quad \gamma = \frac{\gamma_F * \gamma_R}{\|\gamma_F * \gamma_R\|}.$$

2.2.3. Step 2

In the second step of this iterative process, each individual's ancestral state at each locus is sampled conditional on γ , and each individual's parameters are updated by a Metropolis Hastings update, conditional on the sampled ancestral state and the hyperparameters describing their distributions in the admixed population (see Table 6). Each parameter is sampled from the posterior distribution, defined by each individual's sampled ancestral state at each locus and is kept with probability

$$(14) \quad \min\left(1, \frac{L(\text{data}|\text{new parameter})}{L(\text{data}|\text{old parameter})}\right).$$

Parameter Update	Definition
$A_i \sim \text{Dirichlet}\left(\omega_1 + \sum_i \gamma_i, \dots, \omega_k + \sum_i \gamma_k\right)$	• Global ancestry
$A_i^X \sim \text{Dirichlet}\left(\omega^X A_{i1} + \sum_i \gamma_i^X, \dots, \omega^X A_{ik} + \sum_i \gamma_k^X\right)$	• Global X chromosome ancestry
$\lambda_i \sim \text{Gamma}\left(\alpha_1 + \# \text{crossovers}, \alpha_2 + \sum_j d\right)$	• Mean number of generations since admixture
$p_k \sim \text{Beta}\left(\begin{array}{c} \tau_k P_k + \# \text{reference alleles}, \\ \tau_k (1 - P_k) + \# \text{variant alleles} \end{array}\right)$	• Allele frequency in chromosomes descended from population k

Table 6. Individual and population parameter updates. Posterior distributions are given for individual parameter updates, conditional on the sampled ancestral states and number of sampled crossovers. Updates of allele frequencies are also conditional on the sampled ancestral states and observed genotypes.

Hyperparameter Updates	Definition
$\log_{10}(\omega) \sim N(1, 0.5)$	• Ancestral population contribution hyperparameter
$\log_{10}(\omega^X) \sim N(1, 0.5)$	• Prior similarity between A and A^X
$\log_{10}\left(\frac{\alpha_1}{\alpha_2}\right) \sim N(1, 0.5)$	• Hyperparameters for λ
$\log_{10}\left(\frac{\alpha_1}{\alpha_2^2}\right) \sim N(1, 0.5)$	
$P_k \sim \text{Beta}(\Omega_k)$	• Allele frequency in ancestral population k
$\Omega_k \sim \text{Unif}(1, 10)$	• Allele counts in modern-day surrogate population (sampled if unknown)
$\log_{10}(\tau) \sim N(2, 0.5)$	• Weight for each modern-day surrogate population

Table 7. Hyperparameter updates. Distributions are given for hyperparameter updates. Updated values in the Gibbs sampler are kept conditional upon the updated parameters from Table 6.

This is followed by a Metropolis Hastings update of the population parameters and a Gibbs sample of the hyperparameters that describe the prior distributions used above (see Table 7). In this sample, each hyperparameter is sampled from the non-informative prior distribution and kept as a function of the prior likelihoods of the new parameters, similar to Equation 14. This is repeated enough times to reasonably explore the hyperparameter space, and a representative value is chosen for the new hyperparameters.

2.2.4. Inference of ancestry

Steps 1 and 2 are repeated hundreds or thousands of times in an MCMC chain until the likelihood distribution of individual ancestry is sufficiently characterized and each individual's ancestry is summarized by the mean of the sampled distribution. Other measures can also be summarized, including disease association statistics as discussed in Section 2.3.1.

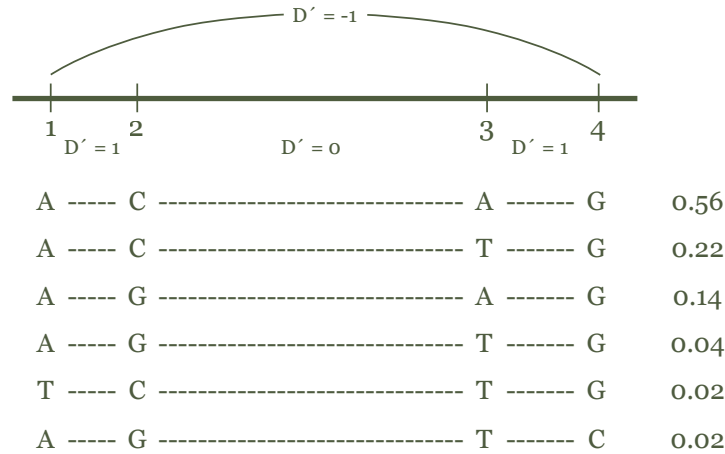


Figure 12. Illustration of problems with relying solely on first order LD. Markers 2 and 3 are very old and unlinked. Markers 1 and 4 are relatively recent, and having occurred on different haplotypes are in nearly complete negative LD. Conditioning only on LD with the previous marker in this case will break the linkage between markers 1 and 4, adding bias to the model. This bias can lead to wildly incorrect inferences (A. L. Price et al., 2008).

2.2.5. Dense marker strategies

The algorithm above depends on a marker set where individual markers are unlinked in the ancestral population. Linked markers will introduce bias unless local LD is correctly modeled accounted for by a sufficiently accurate model (Wall & Pritchard, 2003). SABER implements the first algorithm allowing dense marker sets by incorporating first order local LD estimates into the model (Tang et al., 2006). There have been problems reported using this method, however, indicating higher order LD can play an important roll in in some cases (A. L. Price et al., 2008; Tang et al., 2007). An example of where this could get one into trouble is when two very old, and thus unlinked, markers are sampled between two fairly recent mutations (see Figure 12).

LAMP infers ancestry over windows of unlinked SNPs and uses a ranking system to merge inferences over multiple windows, and a more recent update, LAMP-LD, includes unlinked haplotype blocks (Baran et al., 2012; Sankararaman, Sridhar, & Kimmel, 2008). A recently published method, RFMix, takes a similar approach, but models the unlinked haplotypes and infers ancestry using a random forest approach (Maples, Gravel, Kenny, & Bustamante, 2013). These both allow local modeling of LD on a manageable scale, while increasing power

with a greater number of genotype information. Another recently published fully frequentist approach, EILA, models the odds of a set of SNPs coming from a given parental population as a function of the observed genotypes in the admixed individual (James J Yang, Li, Buu, & Williams, 2013). While computationally friendly, this approach is likely to suffer when the proxy ancestral populations are at all divergent from the ancestral populations they are chosen to represent.

HAPAA and HAPMIX use a HMM similar to that used in imputation of genotypes to model local LD (A. L. Price et al., 2009; Sundquist et al., 2008). HAPMIX differs from HAPAA mainly in that it allows for rare miscopy events resulting in differences between admixed and ancestral haplotypes. These fully Bayesian algorithms, while accurate, are computationally intensive and do not scale well as the marker set increases in size (Baran et al., 2012; Seldin et al., 2011).

The extent of European admixture in African Americans has been studied extensively, and admixture inference algorithms are quite accurate in assigning ancestry proportions for two-way admixed populations (Seldin et al., 2011). However, many populations have multiple populations contributing recent admixture. In Latinos, for example, there is 3-way admixture between Europe, Africa and indigenous populations. Accurate and efficient admixture in these populations is more complex, and evidence exists that many algorithms for this application are not sufficiently developed (Pasaniuc, Sankararaman, Kimmel, & Halperin, 2009; Tian et al., 2007). Europeans and many African populations have been well characterized, but getting a good prior for Native Americans is more problematic (Bryc et al., 2010; Tian et al., 2007). To add to this problem, there is evidence that Native American chromosomes are more often falsely inferred to be European than African. This added bias can affect disease association statistics in unexpected ways (Adler et al., 2010; Jun J Yang et al., 2011).

2.3. Mapping by Admixture Linkage Disequilibrium

Ancestry plays a critical role in our genetic makeup and affects distinct phenotypic traits in various racial groups. As discussed in Section 1.3.3, this can negatively impact genetic studies through unrecognized substructure, but it can also be exploited to map genes under the right circumstances. For several decades admixed populations have been recognized for their potential to elucidate genetic causes of observed phenotypic differences or disease disparities between racial groups (MacLean & Workman, 1973; Thoday, 1969), but it was not until relatively recently that genotyping and computing technologies have evolved sufficiently to truly harness the capabilities of this method.

The basic premise of Mapping by Admixture Linkage Disequilibrium (MALD) is to infer the location of disease genes by way of association between local ancestry and phenotype (McKeigue, 1998). Power to observe statistically significant associations between ancestry and phenotype increases as a function of the magnitude of the incidence rate ratio of disease between the ancestral populations, but even in populations with a large difference in incidence rates a genetic cause is not guaranteed. In many instances, an environmental factor can be responsible for some or all of the observed difference in incidence rates (Risch, Burchard, Ziv, & Tang, 2002).

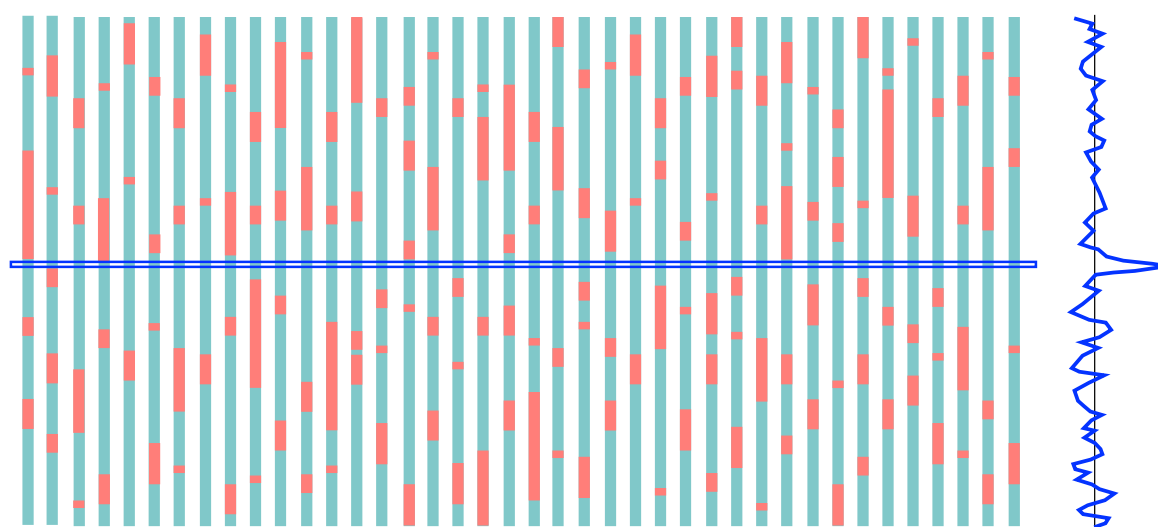


Figure 13. Disease gene mapping by admixture linkage disequilibrium. Human chromosomes are symbolized above. Chromosomal segments with different ancestry are denoted by change in color. After inferring local ancestry, average ancestry from the blue ancestral population (heavy blue line) is summarized for the

entire sample and compared with the genome-wide ancestry proportion (black vertical line). Loci with an unusually high average risk population ancestry are inferred to harbor disease genes (highlighted with the blue box).

Ancestry at each sampled locus in the genome is typically inferred using one of the methods discussed in Section 2.2. When the local ancestry at a locus under study is significantly different than in controls, or when it is significantly different from global ancestry in a case only study design, it can be inferred that a variant specific to that ancestral population is nearby, and that this variant is responsible for a change in prevalence or character of the phenotype of interest (see Figure 13). Statistical power is greatly increased when dense marker sets are used to infer local ancestry, primarily due to the increased accuracy of inferred local ancestry, but care must be taken to correctly account for local LD. The two most common consequence of failing to appropriately handle local LD is that the model will tend to overestimate the divergence of the admixing populations and it will tend to infer significant admixture in unadmixed individuals (Falush et al., 2003). An ongoing area of research is how best to efficiently and accurately model local LD to make use of this increase in power (Seldin et al., 2011).

2.3.1. Disease association statistics

The first test proposed for MALD was a Transmission Disequilibrium Test (TDT, see Section 3.1) utilizing parent/child trios (McKeigue, 1997). The TDT is a powerful method, but requires additional resources to sample both parents of each case. Testing unrelated individuals avoids this problem and is also a valid test for linkage when done properly (McKeigue, 1998).

One option in admixture mapping is a case-only test. This is not normally an option in traditional genetic association studies, but in MALD studies the genome of each case can be effectively used as a substitute for a more traditional control population. In these case-only tests, local ancestry is compared to genome-wide ancestry, with large deviations between the two resulting in statistically significant test results. Two examples of this class of test statistics

are a Bayes factor proposed by Patterson et al. and a non-parametric statistic proposed by Montana et al. (Montana & Pritchard, 2004; Patterson et al., 2004).

Given the i^{th} individual's global ancestry proportion of the risk population: A_i , the estimated probability of the number of chromosomes from the risk population at the locus being tested, $\gamma_{i0} = P(0 \text{ chromosomes})$, $\gamma_{i1} = P(1 \text{ chromosome})$ and $\gamma_{i2} = P(2 \text{ chromosomes})$; and the prior risk increase associated with one or two risk chromosomes, ψ_1 and ψ_2 , the Bayes factor testing association of the locus and disease is defined in Equation 10 (Patterson et al., 2004).

$$(15) \quad Bf = \prod_i \frac{\gamma_{i0} + \psi_1 \gamma_{i1} + \psi_2 \gamma_{i2}}{(1 - A_i)^2 + \psi_1 2A_i(1 - A_i) + \psi_2 A_i^2}$$

Alternately, the non-parametric test statistic defined in Equation 11 is defined in terms of the sample averages for local ancestry at the locus of interest, $\bar{\gamma}$, and global ancestry, (Montana & Pritchard, 2004) $\bar{A}$.

$$(16) \quad T = \frac{\bar{\gamma} - \bar{A}}{SD(\bar{\gamma})}$$

Case / control statistics are typically explored independently of case-only statistics. Pasaniuc et al, however, have proposed a χ^2 test, merging both case / control GWAS results and case only admixture mapping results (Pasaniuc et al., 2011).

Accounting for confounding variables and testing for associations with continuous traits has also been modeled using generalized linear regression techniques (Hoggart et al., 2003). Linear models typically take the form of

$$(17) \quad f(Y_i) = \beta_0 + \bar{A}_i \beta_1 + g(\gamma_i) \beta_3 + X_i \beta_4 + \dots + \epsilon_i,$$

where $f()$ is the link function, $\bar{A}_i$ is the global ancestry of the risk population, $g()$ is a function of the local ancestry probabilities (e.g. under an additive model $g(\gamma) = \gamma_1 + 2\gamma_2$) and X is a

vector of environmental and/or confounding variables. There have also been concerns that differences in admixture between chromosomes inherited from each parent can also introduce bias in disease association. It has been shown that the product of each parent's global ancestry proportion, $\bar{A}_{1i}\bar{A}_{2i}$, is sufficient to account for this potential bias (Redden et al., 2006):

$$(18) \quad f(Y_i) = \beta_0 + \bar{A}_i\beta_1 + \bar{A}_{1i}\bar{A}_{2i}\beta_3 + g(\gamma_i)\beta_4 + X_i\beta_5 + \dots + \varepsilon_i.$$

2.3.2. MALD applications

Phenotypes with large differences in incidence between ancestral populations that make up an admixed population provide the greatest power in MALD studies. Given the genetic diversity that exists between continental populations, it is not surprising that a number of diseases are more common in one or another population. Figure 14 illustrates an analysis of the SEER database tracking cancer in the United States over a five year period, for example, identified 13 cancers with higher incidence in those with African ancestry and 14 cancers with higher incidence in those with European ancestry (Winkler, Nelson, & Smith, 2010).

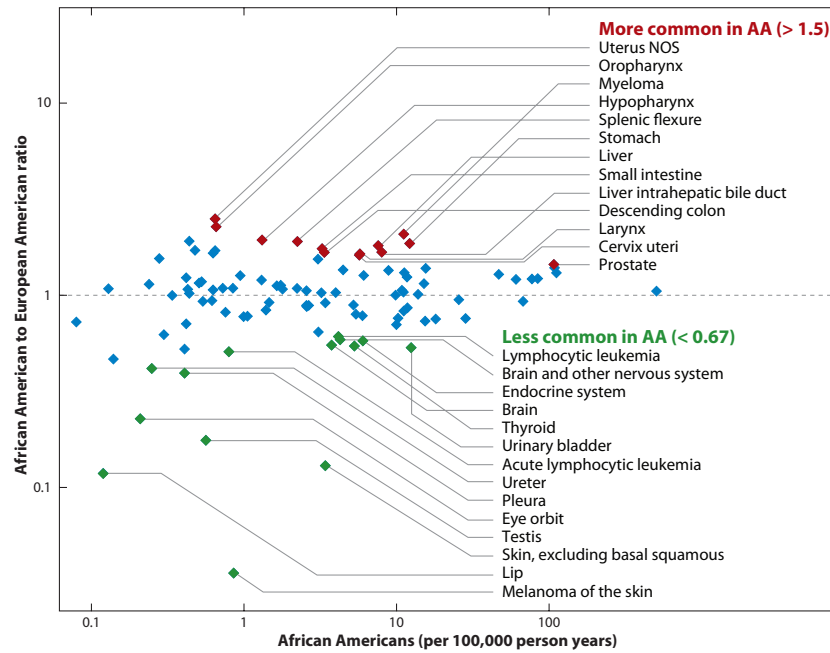


Figure 14. Cancer Incidence Rate Ratios. Cancer incidence among African Americans (x-axis) is plotted against incidence rate ratios comparing African American and European American rates (y-axis) for the incidence of various cancers between 2000 and 2005 as reported in the SEER database. Red points indicate cancers with

significantly elevated incidence in African Americans, and green points indicate cancers with significantly elevated incidence in European Americans. Blue points do not reach statistical significance (Winkler et al., 2010).

Prostate cancer, the most prevalent of the cancers identified in Figure 14, was the first MALD study published (M. L. Freedman et al., 2006). Many reports of disease causing variants have been made over the ensuing seven years in diseases as diverse as renal disease, asthma and neutrophil count (Jeffery B Kopp et al., 2008; Nalls et al., 2008; Torgerson et al., 2012). MALD has not only been applied to the study of human populations, but also to organisms of agricultural importance and in ecological modeling (Mezaka, Legzdina, Waugh, Close, & Rostoks, 2012; Rogers, 2012).

Only recently admixed populations have reliably distinguishable contiguous chromosome segments from identifiable ancestral populations (Hoggart et al., 2004). With newer methods, however, this condition is being relaxed and characterizations of anciently admixed populations are also shedding light on the history of human migration (Patterson et al., 2012; A. L. Price et al., 2009). Admixture mapping is also being used to refine the human reference sequence with information from admixed populations (Genovese, Handsaker, Li, Kenny, & McCarroll, 2013). Even with recent advances, more work is needed, particularly in populations with more complicated admixture patterns (e.g. Latinos and Southern Africa Cape Coloreds (James J Yang et al., 2013)).

3. PCA Modeling of LD

As noted in Section 1.3.3, Population substructure due to differences in ancestry among cases and controls is a common source of bias and may result in inflation of false positive associations if left uncorrected (Campbell et al., 2005). This source of bias can be easily modeled with a principal components analysis (PCA) of the genetic data (A. L. Price et al., 2006). A PCA in a GWAS starts with the full characterization of the covariance of the genetic profiles between each pair of individuals in the study, followed by the calculation of eigenvalues and eigenvectors of the covariance matrix. The first eigenvector defines a transformation of an individual's genotypes that explains as much of the genetic variation in the sample as possible. Each successive eigenvector likewise explains as much of the remaining variation as possible. Each eigenvector transformation summarizes relationships in the data and can be used to explain general trends in the sample, and eigenvalues can be thought of as a measure of the amount of variation accounted for by each eigenvector. The first few principal components of a GWAS PCA are a good proxy for population substructure and their use as confounding covariates is a standard part GWAS analysis (A. L. Price et al., 2006).

In GWAS, the convention is to set the acceptable threshold for genome wide significance between $p < 1 \times 10^{-8}$ and $p < 5 \times 10^{-8}$ (Hoggart et al., 2008; International HapMap Consortium, 2005; McCarthy et al., 2008; Risch & Merikangas, 1996). Because of the increased Type II error rate associated with reducing the probability of Type I errors using this threshold, many believe it is too conservative (Duggal et al., 2008; Gao et al., 2008; Gu, 2007; Nicodemus, Liu, Chase, Tsai, & Fallin, 2005). PCA can be applied to a GWAS using the simpleM algorithm to estimate the number of independent multiple comparisons performed (Gao, Becker, Becker, Starmer, & Province, 2010). This is one of the methods used in my first objective. Another application I use in this dissertation is the summarization of correlated covariates in a statistical model (Hawkins, 1973). Stability of linear regression coefficients suffers when a combination of predictor variables approaches a state of collinearity. This

correlation between predictor variables can be eliminated by instead analyzing the principal components (PCs) obtained in a PCA.

3.1. Principal Components Analysis

PCA is a linear transformation of observations of a series of variables, $X_1, \dots, X_n$, into a series of principal components (PCs), $PC_1, \dots, PC_m$, where $m \leq n$ (Hotelling, 1933; K. Pearson, 1901). The choice of these PCs results two unique properties that can be very useful in practice: 1. Each PC is orthogonal to all other PCs, and 2. Each PC accounts for as much variance in $X_1, \dots, X_n$ that is unexplained by previous PCs, conditional on its orthogonality to previous PCs. As mentioned above, each PC is associated with an eigenvector of the variance covariance matrix of $X_1, \dots, X_n$. The variance of a PC is dependent on how the original variables are scaled, so they are rescaled to have unit sample variance, and the eigenvalue associated with that PC is directly proportional to the amount of the total sample variance it accounts for. This process can also be conceptualized as a rotation in $\mathbb{R}^n$ resulting in n uncorrelated variables when $n = m$ or the projection of $\mathbb{R}^n \rightarrow \mathbb{R}^m$ most closely fitting the full data set, resulting in m uncorrelated variables when $n > m$.

For example, if we have two correlated variables, X_1 and X_2 shown in Figure 15, PC_1 will be associated with the green-dashed least squares line and PC_2 will be associated with the blue-dashed line, perpendicular to it. As such, PC_1 will contain information with regard to the linear relationship between X_1 and X_2 , while PC_2 will contain information about the spread of X_1 and X_2 about the least squares line. An analysis of PC_1 alone would have a slight modeling advantage, since we only have one variable to consider. Two variables are easily visualized, and interactions between the two are easily modeled. In higher dimensions, however (e.g. with hundreds of thousands of genetic markers), visualizing and modeling all variables can be problematic, especially when many highly correlated variables are present. Since the amount of variation captured by each successive PC usually rapidly diminishes, the first few PCs can be sufficient to statistically model extremely complex systems.

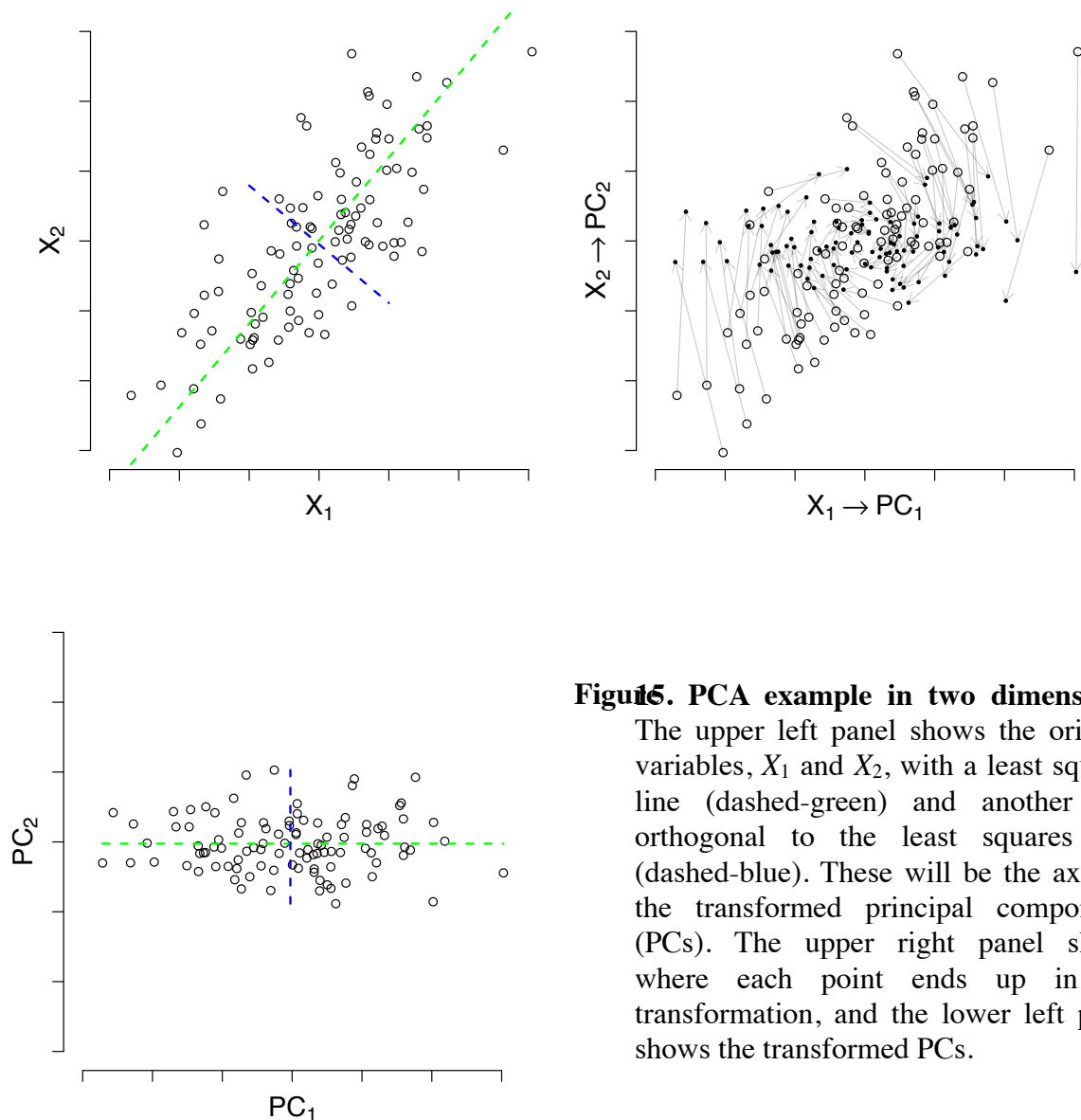


Figure 15. PCA example in two dimensions.

The upper left panel shows the original variables, X_1 and X_2 , with a least squares line (dashed-green) and another line orthogonal to the least squares line (dashed-blue). These will be the axes of the transformed principal components (PCs). The upper right panel shows where each point ends up in the transformation, and the lower left panel shows the transformed PCs.

3.1.1. Use of PCA in Bioinformatics

PCA is a classical method with many uses in bioinformatics (Ma & Dai, 2011; Ramsay & Silverman, 2005). There are at least three primary applications of PCA in bioinformatics: 1) exploratory analysis and data visualization, 2) cluster analysis and 3) regression analysis. The primary benefit in all of these groups of applications is in the consideration of only a few summarizing PCs, greatly reduction of the number of variables in the model.

In many settings it is impossible to visualize all possible variables in a data set, for example when looking for gene / gene interactions in a gene expression analysis. PCA methods can be used in these situations to summarize the many variables into a small enough number of PCs to visualize. This can be useful, for example, when exploring how different classes of genes cluster in cases and controls (Hibbs, Dirksen, Li, & Troyanskaya, 2005). In other exploratory analyses, a graphical representation of PCs can provide insight into how variable are related to one another. An example is given in my preliminary analysis of clinical factors important in the development and progression of end stage renal disease (ESRD). In Figure 16 we note that

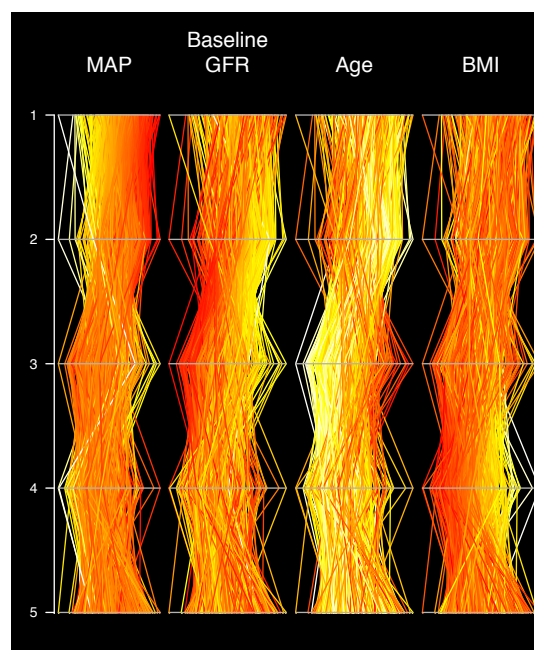


Figure 16. Illustration of the first five principal components (PCs) in a preliminary analysis of variables confounding end stage renal disease (ESRD) outcomes. Each PC is represented in one dimension, with the first at the top through the fifth at the bottom (labels for each PC are located on the left of the figure). Each confounding variable is listed across the top of the figure, including mean arterial blood pressure (MAP), baseline glomerular filtration rate (GFR), age and body mass index (BMI). All PCs are repeated for each confounding variable, but are colored differently. Each line represents a different observation, and individual observations are colored according to the level of the confounding variable. Low MAP levels, for example are colored yellow with high values colored red. PCs with a smooth gradient from yellow to red indicate an association between the PC and the confounding variable.

the first PC associates most strongly with mean arterial blood pressure (MAP) and can conclude that MAP may be the most important factor of those sampled, followed by baseline glomerular filtration rate (GFR), age and body mass index (BMI). It is also interesting to note that both the second and third principal components associate with baseline GFR and age, probably due to the natural decline in GFR with age, but also indicating some other interaction between the two with regard to ESRD.

When the first few PCs are used to summarize a large number of variables, the remaining, unused PCs are generally considered only to capture residual noise in the data. Clustering algorithms can use PCs to combine important variables and minimize residual noise, increasing both the power and computational efficiency of these algorithms (Yeung & Ruzzo, 2001). PCs can be used in a similar way in regression analyses, including many variables in one or two covariates (Hawkins, 1973). Stability of linear regression coefficients suffers when a combination of predictor variables approaches a state of collinearity. This correlation between predictor variables can be eliminated by instead analyzing the PCs obtained in a PCA. An example of this is the analysis of gene expression and gene-gene interactions (Sharov, Dudekula, & Ko, 2005). Some gene combinations may have a different effect on disease outcome than any individual genes would have by themselves. By combining genes with a PCA, combined changes in gene expression data can be more concisely modeled and compared, and by analyzing how the PCs are clustered, important interactions may be recognized that otherwise would have been missed in the crowd of multiple comparisons.

Another application in GWAS is to identify the number of independent genetic markers sampled. The convention in setting a significance threshold is to divide the desired study-wide significance threshold by the number of independent tests performed and use this threshold, called a Bonferroni threshold, as the standard for each individual test. The easiest way to do this is to simply count up the total number of markers sampled, but this will be a conservative estimate, since the markers are not independent due to the presence of local LD. Because of the increased Type II error rate associated with reducing the probability of Type I errors using this threshold, many believe it is too conservative. PCA can be applied to a GWAS using the

simpleM algorithm to estimate the number of independent multiple comparisons performed (Gao et al., 2010). This is one of the methods used in my first objective. When applying the simpleM algorithm, the same basic steps of a PCA are followed, but the covariance matrix is calculated comparing pairs of markers, rather than pairs of individuals. This could result in a very large covariance matrix, so chromosomal segments or even entire chromosomes are analyzed to ease the computational burden. Breaking chromosomes up into smaller segments will have the effect of inflating the final estimate, since this essentially makes the assumption that markers in different segments are unlinked, but this effect is very slight in practice. Because each eigenvalue is directly proportional to the amount of variation it represents in the data, the number of summed, normalized eigenvalues it takes to reach 0.995, or 99.5% of the variation accounted for, is used for the estimate. All other PCs are inferred to provide no further independent information.

3.2. Measurement of Population Substructure

The presence of population substructure in the sample is particularly problematic in GWAS. Population substructure results when an apparently homogeneous population is genetically heterogeneous. This can lead to multiple false positive findings. The association normally sought after is that of genotype causing disease, but false associations can be induced by population substructure in several ways (see Figure 17).

Figure 17a illustrates a real association between the subpopulation and disease. In such a case, the subpopulation confounds the relationship between any structured SNPs and the disease, inducing false associations with structured SNPs. In Figure 17b the chance of a false association is not necessarily increased, but any random association between subpopulations and disease status will result in an increase in false positive results. In both cases, false associations will rise with significantly increased probability in regions that are structured.

The classical illustration of a population association inducing a false association with a structured SNP is a European study which associated height with the lactase gene, LCT. As height varies widely by geographic location, it is associated with ancestry and therefore prone

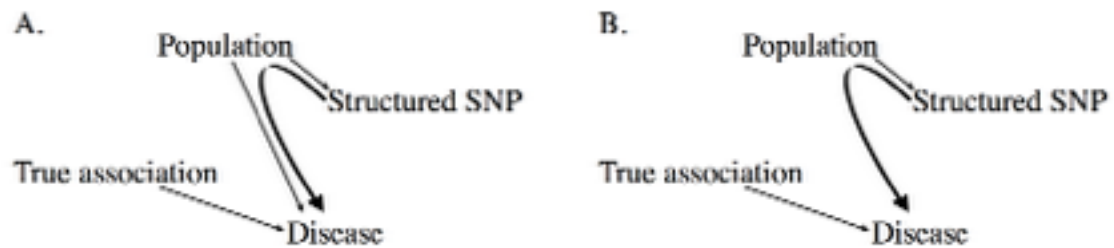


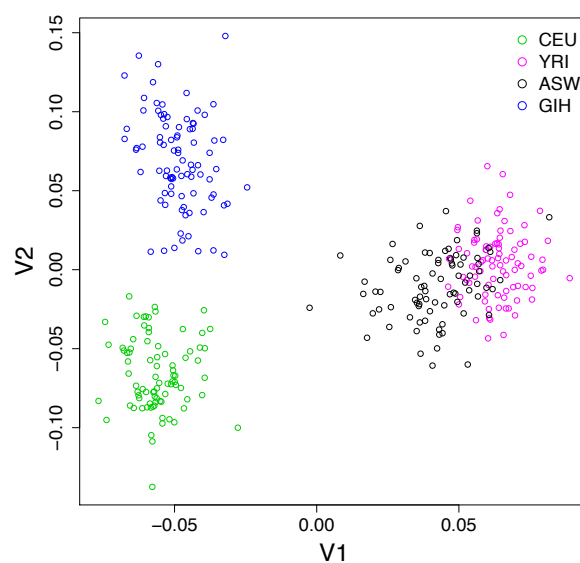
Figure 17. False associations from population substructure. Two ways a false association could arise in the presence of population substructure. Light arrows represent true associations, and heavy, curved arrows represent false associations. A false association can arise via an association with a subset of the population: A) when the disease status is truly associated with that subpopulation, or B) when disease status is merely associated with the study sample of the subpopulation because of poor control selection or some other unrecognized factor.

to false associations due to population substructure. LCT has been more under the influence of selection in some regions than others and is also associated with ancestry. Even though no statistically significant population substructure was detected in the sample under study, when ancestry was taken into account, the observed association between LCT and height disappeared (Campbell et al., 2005).

An example of a false association being induced even when the population under study is likely not associated with the disease outcome is the finding discussed at the beginning of the Introduction between DARC and HIV acquisition in African Americans (He et al., 2008). The comparison of a sample of cases with low amounts of European ancestry relative to the control group led to an overestimation of the association between HIV and alleles more common in Africans. Thus the association statistics for DARC, which nearly perfectly differentiates between African and European ancestry, were more inflated. Several

independent studies with appropriate samples and population substructure statistics were unable to replicate this finding (Walley et al., 2009; Winkler, An, Johnson, Nelson, & Kirk, 2009).

Population substructure of can be identified using a PCA, comparing the genetic profiles of each pair of individuals with each other. The PCs in this application summarize genetic information across all markers, and the first few PCs have been shown to correlate well with population substructure (A. L. Price et al., 2006). Figure 18 illustrates how each subpopulation can be clustered using this method. Included in the analysis are 482 markers in Europeans, Indians, Africans, and African Americans from the HapMap project. Given the recent admixture of African Americans, it is not surprising to see some African Americans clustering closer to the Europeans than the Africans. The amount of population information is shown for the first 10 principal components in Figure 19, further illustrating the power of this summary method. The first few eigenvectors are typically sufficient to characterize population substructure, as they are in this example. These first few eigenvectors may be easily added to a linear model to condition on population substructure and remove this source of bias.



Figure

18. First two eigenvectors of a PCA illustrating population structure. The first two eigenvectors of a principal components analysis of the genetic data include Europeans (CEU, green), Yorubans (YRI, magenta), African Americans (ASW, black), and Gujarati Indians living in Texas (GIH, blue).

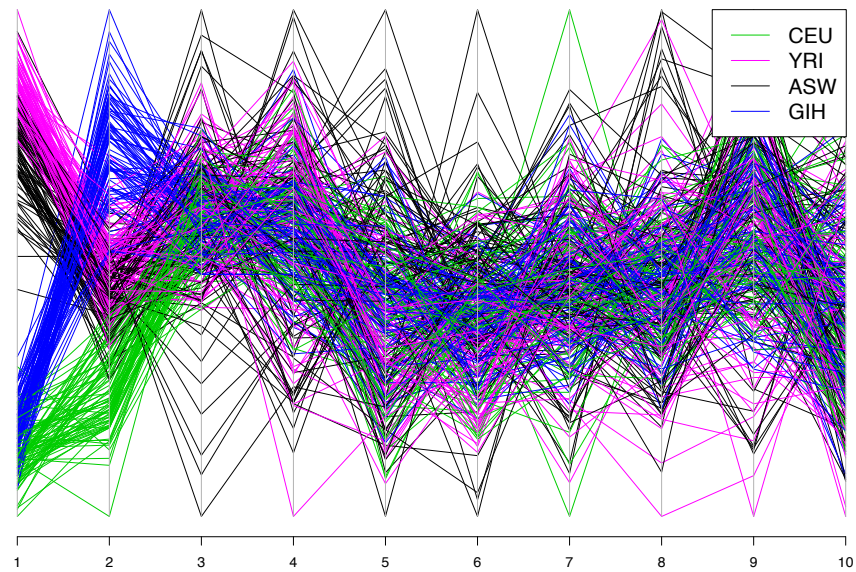


Figure 19. Parallel coordinates plot of 10 eigenvectors illustrating population structure. The first 10 eigenvectors are plotted for a principal components analysis are plotted for Europeans (CEU, green), Yorubans (YRI, magenta), African Americans (ASW, black), and Gujarati Indians living in Texas (GIH, blue). Note that most of the clustering occurs in the first two eigenvectors.

4. Objectives

The overarching goal of this dissertation is to explore ways to increase power in genome-wide association studies (GWAS) of underrepresented populations through more accurate modeling of genetic linkage.

4.1. Objective 1

Putting restrictions on the Type I error rate (false positive rate) directly influences the Type II error rate (false negative rate). Because Type I errors lead to wasted time and money, Type II errors are generally preferred to Type I errors. Increases in Type II error rates associated with large studies, however, have their own costs stemming from missed discoveries and opportunities. In order to reevaluate the appropriateness of using strict Bonferroni corrections, which can lead to important results being overlooked, various methods to determine appropriate genome-wide significance levels are explored.

4.2. Objective 2

An indirect modification of Type II error rates can be achieved most simply by decreasing the number of statistical comparisons made in a GWAS. One powerful method for reducing the number of comparisons is to focus on extended haplotype blocks prevalent in recently admixed populations. A typical GWAS study, with a resolution of less than 0.1 cM, while providing a rich marker set, also suffers from a lack of power, especially when studying rare diseases or underrepresented populations. This requires a stringent genome-wide significance threshold to avoid too many Type I errors at the cost of a corresponding increase in Type II errors. Mapping by Admixture Linkage Disequilibrium (MALD), on the other hand, has a resolution of 5 - 10 cM, and can be efficiently carried out with fewer genetic markers, a potential difference of several orders of magnitude (M. W. Smith et al., 2004).

Current algorithms to infer local ancestry, needed for MALD, generally have one of three shortcomings: they rely on sparse marker sets with no local LD between them, they partition dense marker sets into discrete haplotype blocks to avoid having to model higher order LD, or they are computationally inefficient. Additionally, some of these software packages were designed with other purposes in mind, and obtaining the required estimates can be difficult. A new software package including inference, hypothesis testing and graphics was developed to streamline the analysis and publication of MALD studies. The algorithm, described in this dissertation, models local LD using principal components regression, allowing both the computational efficiency of windowing approaches and appropriate modeling of higher order LD in regions without well defined haplotype blocks.

Results

5. Accounting for multiple comparisons in a genome-wide association study

**Randall C. Johnson^{1,2}, George W. Nelson¹, Jennefer L. Troyer¹, James A Lautenberger³,
Bailey D. Kessing¹, Cheryl A. Winkler¹, Stephen J. O'Brien³**

1. Basic Research Program, SAIC-Frederick, Inc, NCI-Frederick, Frederick, MD
2. Chaire de Bioinformatique, Conservatoire National des Arts et Metiers, 75003, Paris, France
3. Laboratory of Genomic Diversity, NCI Frederick, Frederick, MD

Published in *BMC Genomics*, 2010, 11:724.

Summary

The assessment of statistical significance is one of the key stages in a GWAS, due to the large amount of data and the number of tests performed. Three approaches to assessing statistical significance are controlling the significance threshold by accounting for multiple comparisons, controlling the significance threshold through maintenance of a desired false discovery rate, and by using Bayes Factors (BF). This publication focuses on Bonferroni corrections for multiple comparisons and other alternatives developed to improve upon violated assumptions of Bonferroni corrections. While other methods may be superior in some respects, this focus is primarily due to the preference of this method by so many researchers.

Bonferroni corrections are ideal for their intended purpose in a GWAS setting because there are many comparisons made and there are no well defined hypotheses. One of the key assumptions of a Bonferroni correction, however, is that all tests are independent of one another. This assumption is violated due to the preponderance of LD throughout the genome, and substantial effort has gone into the development of alternate strategies that take this LD into account.

Four methods are explored and contrasted to a simple Bonferroni correction: 1) Permutation test using PRESTO (B. L. Browning, 2008); 2) PCA of the genetic data to identify the number of independent SNPs using the simpleM method (Gao et al., 2010); 3) Multivariate normal p-value correction using SLIDE (Han et al., 2009); 4) Counting the number of haplotype blocks using the three haplotype block calling algorithms implemented in Haploveiw (International HapMap Consortium et al., 2007).

Permutation testing is the ideal way to correct p-values, but the compute time is unreasonable for statistical tests that lack optimized permutation algorithms. We therefore used a simple case-control outcome for this publication. 1,255 HIV cases and 259 exposed uninfected controls, were genotyped on the Affy 6.0 platform, and it is assumed the results extend to other statistical tests, populations, and genotyping platforms.

All methods explored had reasonable computational requirements and execution times. The thresholds identified by each method were standardized, using the permutation results from PRESTO, and compared with a simple Bonferroni correction using the total number of tests as a correction factor. The $\alpha = 0.05$ significance thresholds were similar for the simpleM and SLIDE methods ($\alpha_{\text{standardized}} = 0.055$ and 0.068 respectively), but the significance thresholds obtained by using the number of haplotype blocks were much less conservative ($\alpha_{\text{standardized}}$ between 0.15 and 0.20). These results are tempered by the fact that the simple Bonferroni correction was only very slightly more conservative ($\alpha_{\text{standardized}} = 0.045$) and that the absolute significance threshold was 8×10^{-8} - not far from the uniform standard of 5×10^{-8} for this study wide significance level.

The PRESTO, simpleM, and SLIDE methods appear to do equally well in these data, and we recommend using one of them instead of relying on a one-size-fits all uniform standard. Not much difference was noted in this example, but scenarios are discussed where relying on a uniform standard would be unwise.

RESEARCH ARTICLE

Open Access

Accounting for multiple comparisons in a genome-wide association study (GWAS)

Randall C Johnson^{1,2}, George W Nelson¹, Jennifer L Troyer¹, James A Lautenberger³, Bailey D Kessing¹, Cheryl A Winkler¹, Stephen J O'Brien^{3*}

Abstract

Background: As we enter an era when testing millions of SNPs in a single gene association study will become the standard, consideration of multiple comparisons is an essential part of determining statistical significance. Bonferroni adjustments can be made but are conservative due to the preponderance of linkage disequilibrium (LD) between genetic markers, and permutation testing is not always a viable option. Three major classes of corrections have been proposed to correct the dependent nature of genetic data in Bonferroni adjustments: permutation testing and related alternatives, principal components analysis (PCA), and analysis of blocks of LD across the genome. We consider seven implementations of these commonly used methods using data from 1514 European American participants genotyped for 700,078 SNPs in a GWAS for AIDS.

Results: A Bonferroni correction using the number of LD blocks found by the three algorithms implemented by Haploview resulted in an insufficiently conservative threshold, corresponding to a genome-wide significance level of $\alpha = 0.15 - 0.20$. We observed a moderate increase in power when using PRESTO, SLIDE, and simpleM when compared with traditional Bonferroni methods for population data genotyped on the Affymetrix 6.0 platform in European Americans ($\alpha = 0.05$ thresholds between 1×10^{-7} and 7×10^{-8}).

Conclusions: Correcting for the number of LD blocks resulted in an anti-conservative Bonferroni adjustment. SLIDE and simpleM are particularly useful when using a statistical test not handled in optimized permutation testing packages, and genome-wide corrected p-values using SLIDE, are much easier to interpret for consumers of GWAS studies.

Background

Since the first successful genome-wide association studies (GWAS) in 2005, over 600 GWAS have been reported [1]. Due in large part to rapid advances in genotyping technology and standardized guidelines for reporting statistical evidence, the multitude of comparisons made in a GWAS will result in both false positive (Type 1 errors) and, if the correction for multiple comparisons is overly conservative or power is inadequate, false negative (Type 2 errors) results.

The probability of a Type I error (incorrectly ascribing scientific significance to a statistical test) is generally controlled by setting the significance level, α , for a test, but the probability of making at least one Type I error in a study,

$$P(\text{Study-wide Type I error}) = 1 - (1 - \alpha)^n,$$

is a function of n , the number of independent comparisons made, as well as α . The direct application to a GWAS is that, with a significance level typical to small studies and candidate gene studies (e.g. $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$), the probability of not committing a GWAS-wide Type I error is very small.

The standard for evidence of significance in GWAS to securely identify a genotype-phenotype association in European Americans is generally considered to be $p < 5 \times 10^{-8}$ or $p < 1 \times 10^{-8}$, for $\alpha = 0.05$ and 0.01 , respectively [2-5]. This standard is based on a Bonferroni correction for an assumed million independent variants in the human genome. As a consequence, the avoidance of Type 1 errors may inflate Type 2 errors. This is especially true for analyses with low power, such as rare diseases where patient numbers are limited, low

* Correspondence: obrien@ncicrf.gov

³Laboratory of Genomic Diversity, NCI-Frederick, Frederick, MD, USA
Full list of author information is available at the end of the article

frequency alleles, or genetic factors with small effect sizes. This conundrum can be resolved with extremely large study sizes, but in practice this is not always cost efficient or practical. These issues should be major considerations both for designing GWAS and interpreting GWAS results.

Several methods are commonly used to control the GWAS-wide Type I error rate: p-value adjustments for multiple comparisons have long been used when making multiple comparisons [6]; the use of q-values, a measure of the false discovery rate, has been proposed as a way to indirectly measure and control the Type I error rate [7]; a two-stage analysis of the data can be used not only to decrease the Type I error rate [8], but also to decrease the genotyping costs incurred [9]; genotype imputation can result in a net increase in statistical power [10,11].

A Bonferroni adjustment fits our problem particularly well because many comparisons are made and a GWAS is considered agnostic, with no prior hypotheses [12]. Several studies have estimated the number of statistical comparisons made in a GWAS [2-5], but the universal application of a one-size-fits-all significance level to GWAS studies is inappropriate. Power to detect associations is determined, in large part, by allele frequencies and their effect sizes; since these variables are constants, only sample size can be adjusted. As the sample size increases, the power to detect low frequency and/or small effect size genetic variants also increases. Newer SNP arrays, designed to more fully capture the range of SNPs in diverse human populations and to include rare SNPs hypothesized to be more likely to have larger effect sizes, will increase the number of independent statistical comparisons [4]. Additionally, the dependent nature of genetic data, where SNPs in linkage disequilibrium (LD) are correlated to some degree, may lead to over-correction when using Bonferroni adjustments. One of the key assumptions of a Bonferroni adjustment is that all comparisons are independent. Neighboring SNPs on a chromosome tend to be inherited together in blocks and are not independent [3], making a strict Bonferroni adjustment overly conservative.

One relevant question is then not how many SNPs are being tested, but how many independent statistical comparisons are being made. In the context of a principal components analysis (PCA) of the genotype data, the number of independent comparisons can be defined as the number of principal components accounting for a large portion (99.5% has been suggested) of the variance in the data [13]. The set of informative SNPs represented by these components could be used to infer the

remainder of the data set with a high degree of fidelity, and can be used to make a Bonferroni adjustment with the desired GWAS-wide significance level:

$$\alpha_{\text{GWAS}} = \frac{\alpha}{n_{\text{Informative}}}.$$

What is not clear, however is which SNPs fall into the informative set, so all SNPs are tested. The assumption is then made that the test statistics are distributed similarly to the test statistics from an analysis including only the informative SNPs. Based on the simulations done by Gao et. al. this seems to be a reasonable assumption [13].

Another relevant question is how to adjust the p-values directly, rather than relying on a significance threshold [14]. These corrected p-values, measuring significance on the genome-wide scale, have the added benefit of easier interpretation. For example, comparing two uncorrected p-values, 6.8×10^{-8} and 4.1×10^{-10} , becomes much more tractable after a genome-wide correction, resulting in corrected p-values of 0.0291 and 0.0004, respectively.

There have been a number of studies attempting to provide an accurate picture of how SNPs, and/or statistical tests of SNPs, are correlated in genome-wide studies. These fall into three general categories: variations and alternatives to permutation testing [14,15], principal components analysis [13,16-18], and analysis of the underlying LD structure in the genome [19-21].

We have recently genotyped 1514 European Americans for 700,078 SNPs using the Affymetrix 6.0 platform in a GWAS to search for AIDS restriction genes. Here we compare traditional Bonferroni significance thresholds with methods from each of these statistical correction strategies to identify an appropriate measure of significance in our GWAS: 1) PRESTO, an optimized permutation algorithm [15] verified by PERMORY [22]; 2) the Sliding-window method for Locally Inter-correlated markers with asymptotic Distribution Errors corrected (SLIDE) program, an alternative to permutation testing, developed to correct p-values in a GWAS using a multivariate normal distribution-based correction [14,23]; 3) the simple M method, specifically developed to calculate the number of informative SNPs being tested in a GWAS using a principal components analysis [13]; 4) the number of LD blocks found by the Gabriel, Solid Spine of LD, and 4-Gamete algorithms, as implemented in Haploview [24].

Our aim is to identify the most appropriate method for obtaining accurate GWAS-wide significance thresholds

and/or corrected p-values among 700,000 linked SNPs, the best method being one that results in an accurate estimate of the number of comparisons and has reasonable computational requirements.

Methods

GWAS Data

After filtering for a 90% sample call rate, 1,514 European Americans were successfully genotyped on the Affymetrix 6.0 platform. These subjects consisted of 1,255 HIV- infected and 259 HIV-negative individuals at risk of HIV infection; clinical categories were distributed randomly across plates and batch effects were monitored. We chose 700,078 SNPs, after filtering each SNP for >95% call rate, Hardy-Weinberg equilibrium, Mendel errors, and a minor allele frequency below 1%. After re-clustering and filtering bad SNPs, all sample call rates were >95% with an average call rate of 98.9%. Individuals were unrelated, with the exception of 8 CEPH trios used to check for Mendel errors in the genetic data. A principal components analysis of the genetic data using Eigensoft was used to identify population structure. No significant outliers were identified, however, since there is some stratification in European American populations, SNPs that contributed significantly to population structure were tagged in subsequent analyses [25]. Association statistics were not used for the purposes of this paper, except where indicated in the multiple comparisons methods below.

To address the concern that an excess number of cases to controls would lead to less generalizable results, we analyzed a random sample of 259 cases with all 259 controls. Other than the changes in case/control ratio and sample size, all other variables were left unchanged.

Variations and Alternatives to Permutation Testing

PRESTO: The software package, PRESTO, was used to permute case/control status 10,000 times, and the minimum Mantel trend test p-value for all SNPs in the genome, comparing cases with controls, was recorded for each permuted data set [15]. These minimum p-values were then used to estimate the uncorrected distribution of p-values under the null hypothesis of no true associations in the study. Each p-value was then corrected by finding the corresponding percentile of the distribution of uncorrected p-values, and a significance threshold for a study-wide significance level of α was obtained by finding the α^{th} percentile of the uncorrected distribution. This distribution was used as the standard by which each method's accuracy is gauged, and corresponding significance levels for all other methods were estimated using this distribution. Results from PRESTO were compared with the results from PERMORY,

another optimized permutation testing software package that was recently released [22].

SLIDE: The SLIDE software package was used to implement a multivariate normal distribution-based approximation to a permutation test, using the quantitative trait option, with 10,000 iterations [14,23]. For comparisons with the other methods considered, SLIDE corrected p-values were used to estimate the GWAS-wide significance threshold by finding a corrected p-value equal to the desired study-wide significance level, α .

Principal Components Analysis

SimpleM: The simpleM method [13], based on a principal components analysis of the data, was implemented in R, version 2.9.0 [26], following the example code provided by Gao et al. https://dsgweb.wustl.edu/rgao/simpleM_Ex.zip. This measure of the number of informative SNPs was then used in a Bonferroni adjustment to estimate the GWAS-wide significance threshold. Each chromosome was broken into regions of approximately 5,000 SNPs due to computational constraints. To choose appropriate regions, with as little LD between adjacent regions as possible, we chose cut points between LD blocks identified by Haploview. A second analysis using the largest regions possible, given the memory available, was also explored to see if results were dependent on the region size.

Analysis of Underlying LD

LD blocks were inferred in our GWAS data using the three methods available in Haploview [24]. The number of LD blocks across the human genome, including inter-block SNPs (i.e. singleton SNPs), was used in a Bonferroni adjustment to estimate GWAS-wide significance thresholds [27]. Entire chromosomes could not be analyzed, due to memory constraints, so smaller regions were analyzed. All SNPs from the last full LD block of the previous region were included in the analysis of the next region to ensure complete LD blocks.

The Gabriel protocol, the default method for Haploview, was used with an upper D' confidence interval bound of 0.98, a lower D' confidence interval bound of 0.70, and with 5% of informative markers required to be in strong LD [28]. The Solid Spine of LD algorithm [29] was used with minimum D' value of 0.8, as suggested by Duggal et al. [21]. The 4-Gamete test was run setting the cutoff for frequency of the fourth pairwise haplotype at 1% [30,31].

Results and Discussion

Variations and Alternatives to Permutation Testing

PRESTO: The permutation based significance threshold from PRESTO was 7.6×10^{-8} (see Table 1). By

Table 1 Summary of Analysis Results

Method	Significance Threshold	Corresponding α level
Bonferroni	0.71×10^{-7}	0.046
PRESTO	0.76×10^{-7}	0.05
simple.M	0.82×10^{-7}	0.053
SLIDE	1.09×10^{-7}	0.068
Gabriel	2.72×10^{-7}	0.151
4-Gamete	3.06×10^{-7}	0.166
Solid spine	3.71×10^{-7}	0.195

The significance threshold for each method is shown ($\alpha_{\text{GWAS}} = 0.05$), as well as the corresponding genome-wide α level when compared with the PRESTO method. A strict Bonferroni significance threshold is also given.

comparison, the PRESTO analysis of the smaller sample had a significance threshold of 1.4×10^{-7} (see Table 2); this corresponds to an α level of 0.09 when compared to the analysis of the full data set. These results were consistent with an analysis using PERMORY on the same subset and probably reflect the decrease in statistical power associated with the smaller sample size. Permutation tests are the gold standard for identifying appropriate significance thresholds, and are computationally efficient when optimized solutions exist for a particular statistical test. As we see in Table 2, these results are very specific to each study. One drawback of permutation testing is the computational burden that arises when no optimized solutions exist (e.g. when modeling survival or longitudinal data). In such a case, permutation testing can be impractical and one of the other methods considered here would be more appropriate.

Bonferroni: The standard Bonferroni correction, simply using the total number of SNPs tested in the genome-wide significance level calculation, was 7.1×10^{-8} , which corresponded to a genome-wide significance level of $\alpha \approx 0.05$ when compared with PRESTO (see Table 1). While a permutation test may not result in a large improvement in the corresponding genome-wide significance level when compared with a standard Bonferroni correction in this SNP set, other, denser SNP sets will result in greater disparities in significance levels.

Table 2 Difference in Significance Threshold in a Subset of the Data

Method	Δ Significance Threshold
simple.M	-8×10^{-11}
4-Gamete	-8×10^{-10}
SLIDE	-5×10^{-9}
Gabriel	-6×10^{-8}
PRESTO	7×10^{-7}
Solid spine	-8×10^{-7}

The difference in significance threshold is given, comparing an analysis of the full data set to a subset of the data with an equal number of cases and controls (1,514 and 518 individuals, respectively).

SLIDE: The significance threshold identified by SLIDE was 1.1×10^{-7} , which corresponded to a genome-wide significance level of $\alpha = 0.07$ when compared with PRESTO (see Table 1). The significance threshold found in the analysis of the smaller sample was remarkably similar, differing only by 5×10^{-9} (see Table 2). Over all, these results indicate that SLIDE is an excellent alternative to permutation testing. Additionally, the corrected p-values provide increased ease in interpretation of GWAS results.

Principal Components Analysis

simple.M: The significance threshold based on the number of effective SNPs identified by the simple.M algorithm was 8.2×10^{-7} , corresponding to a genome-wide significance level of $\alpha \approx 0.05$ when compared with the PRESTO results. As with SLIDE, the analysis of the smaller sample was remarkably similar, differing only by 8×10^{-10} . These results indicate that simple.M is also an excellent alternative to a full permutation test. However, because of the concern of how variations in region size would affect the accuracy of the simple.M analysis, regions with as many SNPs as we had computational resources to analyze (some regions included nearly 30,000 SNPs, others consisted of entire chromosomes) were compared to the results in Table 1. The corresponding thresholds differed by less than 6×10^{-9} . It is important to note, however, that since this is an $O(n^2)$ problem, the memory and serial time required to analyze these larger regions increases rapidly with the size of the regions analyzed. Regions containing more than a few thousand SNPs, however, seem to result in very similar significance thresholds in this data set, and the computational resources required are reasonable for regions of a few thousand SNPs (see Figure 1).

The simple.M method is currently the fastest way to calculate the effective number of independent tests in a GWAS [32], but due to the $O(n^2)$ nature of this algorithm the genome needs to be broken up into small regions to maintain this computational speed. This adds complexity to the analysis and requires a significant amount of pre-analysis. Considering the many examples of long range LD across the genome, simple.M could also lead to a slightly more conservative estimate in some studies [14].

Analysis of Underlying LD

The three LD-based methods using Haploview are the least conservative, with significance thresholds between 2.72×10^{-7} and 3.71×10^{-7} , corresponding to α levels between 0.15 and 0.20 as compared to permutation testing using PRESTO (see Table 1). Thus, it appears that the use of LD blocks to construct Bonferroni significance thresholds is anti-conservative in this data set. We also explored alternate parameters but did not observe a

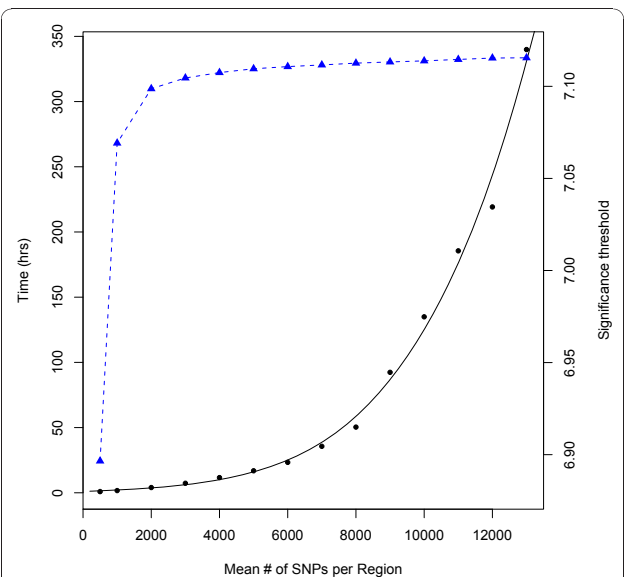


Figure 1 Change in computation time and significance threshold for varying region sizes. The change in serial computation time (solid black line) and significance threshold (dotted blue line) are plotted as a function of the mean number of SNPs in each region in a GWAS-wide analysis using the simple. *M* method.

sufficient improvement in the corresponding significance level when severely restricting the definition of haplotypes (see Table 3).

Nicodemus et al. [27] noted that estimates may be more or less conservative under varying levels of LD. An alternate LD algorithm or parameter constraints could be found that would result in a more accurate estimate [33], but this would vary significantly depending on the sample size, the set of SNPs, and the underlying level of LD structure in the population. This is further illustrated in the large differences found using the Gabriel and Solid Spine of LD

Table 3 Comparison of α levels when restricting the definition of a haplotype

Method	Parameters	Significance Threshold	Corresponding α level
Gabriel	D'U > 0.98	2.72×10^{-7}	0.151
	D'L > 0.70		
	D'U > 0.98	2.11×10^{-7}	0.12
	D'L > 0.85		
4-Gamete	Cutoff = 1%	3.06×10^{-7}	0.166
	Cutoff = 0.5%	2.50×10^{-7}	0.139
Solid spine	D' = 0.80	3.71×10^{-7}	0.195
	D' = 0.95	2.79×10^{-7}	0.155

Significance thresholds with corresponding α levels are given for each haplotype calling method with the standard parameters and a set of more restricted parameters. D'U and D'L represent the upper and lower confidence limits of D', respectively.

algorithms on a subset of the individuals in this study (see Table 2). While LD blocks do provide key information on patterns of LD and how SNPs are correlated, providing invaluable information for interpreting GWAS results and for the planning of follow-up studies, we find the use of significance thresholds derived from LD blocks to be too variable for general application to GWAS data.

Conclusions

A one-size-fits-all Bonferroni correction, although conservative, may not result in a large Type II error rate with a sample size in the tens of thousands, but as the sample size drops, so does statistical power. In studies where gathering large numbers of cases is prohibitive (e.g. when disease prevalence is low), a Bonferroni correction becomes overly conservative by detrimentally inflating the Type II error rate. The methods considered here can ameliorate this loss of power and make interpretation of study results less enigmatic.

The results from the PRESTO, SLIDE and simple. *M* methods appear to be equally good in population data genotyped on the Affymetrix 6.0 platform in European Americans ($\alpha = 0.05$ thresholds between 1×10^{-7} and 8×10^{-8}), and each presents a modest gain in power over the strict Bonferroni thresholds advocated by some [2-5]. The SLIDE and simple. *M* methods may be less dependent on the number of individuals in the study, and will be particularly useful when using a statistical test that is not supported by optimized permutation packages (e.g. when modeling survival or longitudinal data) and when the SNPs being tested are sufficiently dense. SLIDE not only has much nicer computational properties when compared to simple. *M*, but the corrected p-values measuring significance on the genome-wide scale are easier to interpret. While the idea of an even standard across studies is appealing, the traditional standard of presenting p-values in the context of the study more accurately represents the data.

Acknowledgements

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD <http://biowulf.nih.gov>. We thank the individuals who participated in the HGDS, MACS, MHCS, and SFCC cohort studies, as well as the physicians and researchers responsible for recruitment and sample collection. We also thank Michelle Hall, Michael Malasky, Lisa Maslan, Mary McNally, and Jami Troxler who performed the genotyping, Leslie Chinn, Sher Hendrickson, Carl McIntosh, and Joan Pontius who helped with quality control and annotation of the GWAS data, and Julie Johnson for her comments and discussion. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This research was supported [in part] by the Intramural Research Program of NIH, National Cancer Institute, Center for Cancer Research.

Author details

¹Basic Research Program, SAIC-Frederick, Inc. NCI-Frederick, Frederick, MD, USA. ²Chaire de Bioinformatique, Conservatoire National des Arts et Metiers, 75003, Paris, France. ³Laboratory of Genomic Diversity, NCI-Frederick, Frederick, MD, USA.

Authors' contributions

RCJ conceived and carried out the analysis. GWN, CAW, and SJO contributed to the study design. JLT, JAL, BDK, RCJ, CAW, GWN, and SJO contributed to the GWAS data. RCJ wrote the manuscript with contributions from GWN, CAW, JAL, and SJO. All authors read and approved the final manuscript.

Received: 19 August 2010 Accepted: 22 December 2010

Published: 22 December 2010

References

1. A Catalog of Published Genome-Wide Association Studies. [http://www.genome.gov/26525384].
2. Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996, **273**(5281):1516-1517.
3. International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005, **437**(7063):1299-1320.
4. Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ: Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology* 2008, **32**(2):179-185.
5. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis PA, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008, **9**(5):356-369.
6. Miller RG: Simultaneous Statistical Inference. New York, NY: Springer-Verlag Inc.; 2 1981.
7. Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proceedings of the National Academy of Science* 2003, **100**(16):9440-9445.
8. Zheng G, Song K, Elston RC: Adaptive Two-Stage Analysis of Genetic Association in Case-Control Designs. *Human Heredity* 2007, **63**:175-186.
9. Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* 2006, **38**(2):209-213.
10. Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* 2007, **39**(7):906-913.
11. Spencer CCA, Su Z, Donnelly P, Marchini J: Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* 2009, **5**(5):e1000477.
12. Perneger TV: What's wrong with Bonferroni adjustments. *British Medical Journal* 1998, **316**(7139):1236-1238.
13. Gao X, Starmer J, Martin ER: A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology* 2008, **32**(4):361-369.
14. Han B, Kang HM, Eskin E: Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics* 2009, **5**(4):e1000456.
15. Browning BL: PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics* 2008, **9**:309.
16. Cheverud JM: A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 2001, **87**(1):52-58.
17. Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics* 2004, **74**(4):765-769.
18. Galwey NW: A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology* 2009, **33**(7):559-568.
19. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly M, Reich D: Methods for High-Density Admixture Mapping of Disease Genes. *American Journal of Human Genetics* 2004, **74**(5):979-1000.
20. Smith MW, O'Brien SJ: Mapping by Admixture Linkage Disequilibrium: Advances, Limitations and Guidelines. *Nature Genetics* 2005, **6**:623-632.
21. Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE: Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* 2008, **9**:516.

22. Pahl R, Schäfer H: PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics* 2010, **26**(17):2093-2100.
23. Han B, Eskin E: Sliding-window method for Locally Inter-correlated markers with asymptotic Distribution Errors corrected (SLIDE). 2009 [http://slide.cs.ucla.edu].
24. International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007, **449**(7164):851-861.
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics* 2006, **38**(8):904-909.
26. R Development Core Team: R: A language and environment for statistical computing. (Vienna Austria) 2009 [http://www.R-project.org].
27. Nicodemus KK, Liu W, Chase GA, Tsai YY, Fallin MD: Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. *BMC Genet* 2005, **6**(Suppl 1):S78.
28. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, Defelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: The structure of haplotype blocks in the human genome. *Science* 2002, **296**(5576):2225-2229.
29. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005, **21**(2):263-265.
30. Hudson RR, Kaplan NL: Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 1985, **111**(1):147-164.
31. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L: Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics* 2002, **71**(5):1227-1234.
32. Gao X, Becker LC, Becker DM, Starmer JD, Province MA: Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology* 2010, **34**(1):100-105.
33. Schwartz R, Halldórsson BV, Bafna V, Clark AG, Istrail S: Robustness of inference of haplotype block structure. *J Comput Biol* 2003, **10**(1):13-19.

doi:10.1186/1471-2164-11-724

Cite this article as: Johnson et al.: Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 2010 **11**:724.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



6. ALDsuite: Dense marker MALD using principal components of ancestral linkage disequilibrium

Randall C. Johnson^{1,2}, George W. Nelson¹, Jean-Francois Zagury² and Cheryl A. Winkler³

1. BSP CCR Genetics Core, Leidos Biomedical Research, Inc, Frederick National Laboratory, Frederick, MD
2. Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, 75003, Paris, France
3. Basic Research Laboratory, Leidos Biomedical Research, Inc, Frederick National Laboratory, Frederick, MD

In preparation for submission to *Bioinformatics*

Summary

Mapping by admixture linkage disequilibrium (MALD) is a powerful whole genome gene mapping method that uses LD from extended blocks of ancestry inherited from parental populations among admixed individuals to map associations for diseases varying in prevalence among human populations. The extended LD queried for marker association with ancestry results in a greatly reduced number of comparisons compared to standard genome wide association studies (GWAS). As ancestral population LD tends to confound the analysis of admixture LD, the earliest algorithms for MALD required marker sets sufficiently sparse to lack significant ancestral LD between markers. However current genotyping technologies routinely provide data more dense than this, which convey more information than sparse sets, if they can be efficiently used. Several algorithms for MALD with dense data have been presented; we here present an R package, ALDsuite, which accounts for local LD using principal components of haplotypes from surrogate ancestral population data, and includes tools for quality control of data, MALD and downstream analysis of results.

There are currently no software solutions which both offer analysis of dense marker data from more than two admixing populations and disease association statistics. ALDsuite offers a fast,

accurate estimation of global and local ancestry and comes bundled with the tools needed for MALD, from data quality control through mapping of and visualization of disease genes.

ALDsuite: Dense marker MALD using principal components of ancestral linkage disequilibrium

Randall C. Johnson^{1,2}, George W. Nelson¹, Jean-Francois Zagury² and Cheryl A. Winkler^{3*}

¹BSP CCR Genetics Core, Leidos Biomedical Research, Inc, Frederick National Laboratory, Frederick, MD

²Chaire de Bioinformatique, Conservatoire National des Arts et Metiers, 75003, Paris France

³Basic Research Laboratory, Leidos Biomedical Research, Inc, Frederick National Laboratory, Frederick, MD

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Motivation: Mapping by admixture linkage disequilibrium (MALD) is a powerful whole genome gene mapping method that uses LD from extended blocks of ancestry inherited from parental populations among admixed individuals to map associations for diseases varying in prevalence among human populations. The extended LD queried for marker association with ancestry results in a greatly reduced number of comparisons compared to standard genome wide association studies (GWAS). As ancestral population LD tends to confound the analysis of admixture LD, the earliest algorithms for MALD required marker sets sufficiently sparse to lack significant ancestral LD between markers. However current genotyping technologies routinely provide data more dense than this, which convey more information than sparse sets, if they can be efficiently used. Several algorithms for MALD with dense data have been presented; we here present an R package, ALDsuite, which accounts for local LD using principal components of haplotypes from surrogate ancestral population data, and includes tools for quality control of data, MALD and downstream analysis of results.

Results: There are currently no software solutions which both offer analysis of dense marker data from more than two admixing populations and disease association statistics. ALDsuite offers a fast, accurate estimation of global and local ancestry and comes bundled with the tools needed for MALD, from data quality control through mapping of and visualization of disease genes.

Availability: ALDsuite and ALDdata are freely available at <https://ccrod.cancer.gov/confluence/display/BCGC/BCGC+Software> and will soon be available on the Comprehensive R Archive Network.

Contact: johnsonra@mail.nih.gov

1 INTRODUCTION

It is well established that phenotypes differ among human populations. Observed differences between ancestral groups can be attributed to two general causes: a differences in environmental exposures or factors or a difference in underlying genetic variation.

It has been recognized for some time that individuals with mixed ancestry could provide a way to map phenotype / genotype associations to specific loci (MacLean and Workman, 1973; Thoday, 1969). When two populations combine to form a new admixed population, large chromosomal segments from each of the ancestral populations remain in circulation for many generations. The difference in allele and haplotype frequencies between the populations induces an admixture linkage disequilibrium (ALD) that extends over much greater distances than the local LD inherited from ancestral populations. With each new generation chromosomes recombine and the extent of ALD becomes smaller, but with the sequencing of the human genome and the advances in genotyping technology of the last decade, the ancestral origin of chromosomal segments can be inferred with high accuracy for many generations post-admixture (Seldin *et al.*, 2011).

The application of ALD information to association studies, also referred to as Mapping by Admixture Linkage Disequilibrium (MALD), is a statistically powerful method to identify genetic associations with disease in admixed populations when there is a difference in disease risk among ancestral groups (McKeigue, 1997). The key advantage of this approach over the standard genome wide association study (GWAS) approach is that the effective number of statistical comparisons, for associations between markers and disease, is inversely related to the length of LD between markers and the causal disease locus. In African Americans, for example, ALD between loci as distant as 20 cM has been identified, while LD in non-admixed populations rarely extends longer than 0.1 cM (Parra *et al.*, 1998; Smith and O'Brien, 2005). This increases the power over classical GWAS by drawing focus to a specific region of interest with 200-500 fold fewer comparisons that must be corrected using multiple comparisons techniques (Smith and O'Brien, 2005).

As computational power has increased and the cost of genotyping and sequencing has decreased, MALD studies have become more common and successfully applied to identify a number of genetic variants associated with common diseases (Kopp *et al.*, 2008). Several software packages, ADMIXMAP, ANCESTRYMAP and STRUCTURE, provided good estimates of global ancestry as well as statistics for association between phenotype and local ancestry (Hoggart *et al.*, 2004; Patterson *et al.*, 2004; Falush *et al.*, 2003).

*to whom correspondence should be addressed

These early software packages, however, were limited in their ability to analyze dense marker sets, due to their reliance on the lack of local LD among sampled markers. This reliance on sparse marker sets results from the additional complexity involved with the modeling of local LD. An attempt was made in one software package, SABER, to model 2-way LD of a marker with its immediate neighbors, but this was later shown to allow bias into the model from higher order local LD with more distantly linked markers (Tang *et al.*, 2006; Price *et al.*, 2008). The consequences of this bias include a tendency to overestimate the divergence of admixing populations and possible inference of significant admixture in unadmixed individuals (Seldin *et al.*, 2011).

Two recent software packages, HAPAA and HAPMIX, have modeled local LD in a Bayesian framework similar to that of genotype imputation with very good results (Sundquist *et al.*, 2008; Price *et al.*, 2009). These methods, however, are computationally intensive and do not scale well with increasingly dense marker sets (Seldin *et al.*, 2011). Other recent algorithms, including LAMP-LD, MULTIMIX and RFMix, have mainly focused on local ancestry inference using disjoint haplotype blocks (Baran *et al.*, 2012; Churchhouse and Marchini, 2013; Maples *et al.*, 2013). While this approach is much more computationally efficient, many regions do not segregate well into haplotype blocks.

With the R package described here we provide local ancestry estimates using a hidden Markov model (HMM) algorithm similar to that used by existing software (Hoggart *et al.*, 2004; Patterson *et al.*, 2004; Falush *et al.*, 2003), with higher order local LD modeled indirectly using principal components of neighboring markers. Additional features not provided in most admixture software packages include MALD association statistics, quality control measures and data formatting tools. Analysis of phased haplotypes or unphased genotype data is supported and further statistical and graphical analysis using the powerful tool set available in R is readily available.

2 METHODS

2.1 HMM Algorithm

We use an HMM to model the ancestral states of each individual (see Figure 1). In the first step, ancestral states, γ , are sampled for each individual at each locus in the sample, followed in the second step by an update of the parameters on which γ is conditioned. This is repeated iteratively, typically with 100 burn in iterations and 200 follow on iterations when good priors are used. A basic overview is given here; complete details are given in the appendix.

Higher order LD information is captured using principal components (PCs) of the surrounding, linked markers, which is included in the Hidden Markov Model (HMM). PCAdmix also uses PCs of nearby markers, but does not include closely linked markers and does not capture information from higher order LD (Brisbin *et al.*, 2012). Additionally, our software includes a number of other distinguishing features including the calculation of MALD statistics for a wide range of disease association models, quality control measures of the supplied data and analysis of both phased and unphased data. In preparing the sample priors for our model, consisting of modern-day surrogates for ancestral populations, a principal component analysis (PCA) of linked markers of each analysis marker is performed using the modern-day surrogates for ancestral populations. This is best done on a phased data set such as that provided by the International HapMap Project (International HapMap Consortium, 2007), which can be found in the

accompanying companion package, ALDdata. The markers included in the PCA for a locus are all markers that are significantly linked in the modern-day surrogate populations or are within w centimorgans (cM) of the analysis marker (default is 0.1 cM).

All surrogate populations are included in the PCA, and PCs accounting for 80% of the observed variation are chosen to model the likelihood of each ancestral at the current locus. The transformation of the genotype data is illustrated in Equation 1 where the principal component matrix for a locus is the matrix multiplication of the genotype or the phased haplotype matrix, A (one row per individual/chromosome, one column per marker in the haplotype) with the eigenvector, v , associated with the PC:

$$PC_l(A) = Av_l. \quad (1)$$

Prior estimates of the regression coefficients for the first $K-1$ populations, β_k , are calculated on the modern-day surrogate populations, which are used as starting points for local ancestral probability calculation in the HMM. In the case of two ancestral populations, this simplifies to logistic regression, and a multinomial logistic regression is used to model admixture between more than two populations (see Equation 2). In sparsely sampled regions, where only one marker was sampled, observed alleles are used in the model instead of PCs. Given the ancestral state of the haplotype, k :

$$\begin{aligned} \log \left(\frac{P(g=1|A)}{P(g=K|A)} \right) &= \beta_{1,0} + \beta_{1,1}PC_1(A) + \dots \\ &\vdots \\ \log \left(\frac{P(g=K-1|A)}{P(g=K|A)} \right) &= \beta_{(K-1),0} + \beta_{(K-1),1}PC_1(A) + \dots \\ P(g=K|A) &= \frac{1}{1 + \frac{P(g=1|A)}{P(g=K|A)} + \dots + \frac{P(g=K-1|A)}{P(g=K|A)}}. \end{aligned} \quad (2)$$

This has been further extended to the analysis of unphased genotypes, where combinations of local ancestry from two possible populations are modeled. The PCA for unphased data is performed on random pairings of 100 surrogate ancestral haplotypes for each possible local ancestral state (e.g. 'YRI/YRI', 'YRI/CEU', 'CEU/CEU'). An example for two admixing populations is given here:

$$\begin{aligned} \log \left(\frac{P(g_1=1 \cap g_2=1|A)}{P(g_1=2 \cap g_2=2|A)} \right) &= \beta_{11,0} + \beta_{11,1}PC_1(A) + \dots \\ &= LR_{1,1} \\ \log \left(\frac{P(g_1=1 \cap g_2=2|A)}{P(g_1=2 \cap g_2=2|A)} \right) &= \beta_{12,0} + \beta_{12,1}PC_1(A) + \dots \\ &= LR_{1,2} \\ P(g_1=2 \cap g_2=2|A) &= \frac{1}{1 + e^{LR_{1,1}} + e^{LR_{1,2}}}. \end{aligned} \quad (3)$$

Before sampling ancestral states, ancestral state probabilities are first calculated at each locus for each parent. We calculate these ancestral state probabilities using a forward-backward algorithm similar to other admixture HMMs (Falush *et al.*, 2003; Patterson *et al.*, 2004; Hoggart *et al.*, 2004). The ancestral state probabilities in each Markov chain (i.e. from the forward or backward pass) consist of the ancestral state probabilities defined in Equations 2 and 3, conditioned on the ancestral state probability of the previous marker in the chain and the likelihood of recombination between the two:

$$\begin{aligned} \gamma_1 &= P(g_1|A) \\ \gamma_j &= P(g_j|A) * (P(r_j)G + (1 - P(r_j))\gamma_{j-1}), \end{aligned} \quad (4)$$

where A is the observed genotype or haplotype data and g is the ancestral state, G is the global ancestry and $P(r)$ is the probability of recombination between the current and previous loci. These probabilities further dependent

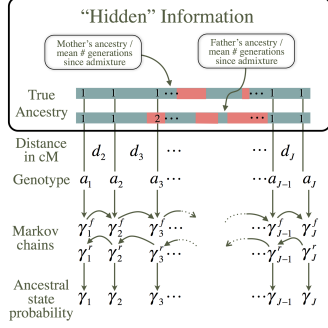


Fig. 1. Hidden Markov Model for ancestry inference. Each individuals local ancestral state probability, γ , is modeled as a function of preceding ancestral state probabilities in each Markov chain, genetic distance to neighboring markers, d , individual global ancestry parameters and observed haplotype or genotypes, a , in a region.

on the number of generations since admixture, λ , and the genetic distance between loci, d . The product of these two Markov chains, γ^f and γ^r , is normalized (so that they sum to one) to obtain the final ancestral state probabilities for each locus, conditional on admixture linkage disequilibrium with nearby markers,

$$\gamma = \left\| \gamma^f * \gamma^r \right\|, \quad (5)$$

and these ancestral state probabilities are used to sample the local ancestral state for each chromosome.

Parameters informing the HMM, particularly those on which γ is conditioned, are updated at the conclusion of each iteration, using the sampled ancestral states discussed in the preceding paragraphs (see the appendix for more details).

Support for the parallelization of the computation is also included in ALDsuite using a distributed MCMC approach in which a separate analysis, or chain, is run for each parallel process (Murray, 2010; Wu *et al.*, 2012). In order to avoid unnecessary duplication of effort during the burn in phase, each chain reports back to the main process after each iteration, where a remote proposal of each parameter is calculated based on the average of all parallel chains. Each chain then updates its own parameter using a weighted sum of the local and remote proposals:

$$\frac{iter}{n \text{ burn}} * local \text{ proposal} + \left(1 - \frac{iter}{n \text{ burn}}\right) * remote \text{ proposal}, \quad (6)$$

where *iter* is the current iteration and *n burn* is the total number of burn in iterations. This results in a quicker convergence to the equilibrium distribution while allowing each chain to start sampling at an independent state.

2.2 Error Checking

Marker checks Several quality control checks are done on each marker to identify potential genotyping errors, mapping errors, flipped markers and irregular variations in allele frequency:

1. Hardy-Weinberg Equilibrium is tested using the `hwexact()` function in the `hwde` package (Mandonald, 2013).
2. Markers with a missing data rate exceeding a user-defined threshold are screened (default threshold is 5%).
3. Allele frequencies from genotypic data coded as A/C/T/G are compared among populations to identify potential A-T / G-C flips that may have

occurred in data originating from different sources. The default is to drop these markers from the analysis set.

4. Allele frequencies in the admixed population are compared with modern-day, ancestral surrogate population allele frequencies to identify potentially irregular loci.

Individual checks Several quality control checks can be performed for individuals to identify potentially bad samples:

1. Individuals with a missing data rate exceeding a user-defined threshold are screened (default threshold is 5%).
2. When sex chromosome data are available, simple gender checks are performed and possible issues are flagged.
3. The sample is screened for potentially related individuals, and matches are flagged.

2.3 Statistical Association

Locus-genome statistics The case-only locus-genome statistic from ANCESTRYMAP is also implemented in our software (Patterson *et al.*, 2004). The locus-genome statistic compares the ancestry at each locus, conditional on a prior risk model, to the genome-wide ancestry estimates under the same model. This is calculated as the product of likelihood ratios for each individual across the entire sample at the j th locus:

$$L(k)_j = \frac{\prod_i P(g_1 \neq k \cap g_2 \neq k) + P(g_1 = k \oplus g_2 = k) \psi_1 + P(g_1 = k \cap g_2 = k) \psi_2}{\prod_i (1-G)^2 + 2G(1-G) \psi_1 + A^2 \psi_2} \quad (7)$$

where the ancestral probabilities are given in Equation (7) of the appendix. The risk model is defined by ψ_1 , the increased risk associated with one allele from population k , with $\psi_2 = \psi_1^2$ equal to the increased risk associated with two alleles from population k . We also implement the genome-wide association statistic proposed by Patterson *et al.* (2004) for this model. This can also be implemented to calculate a chromosome-wide association statistic. Typically LJ is considered moderately significant at $L_J > 2$ and significant at $L_J > 3$.

$$L_J(k) = \frac{1}{|J|} \sum_{j \in J} L(k)_j \quad (8)$$

In addition to the model implemented in ANCESTRYMAP, we also include a locus-genome statistic for association with alternate risk models. The first alternate risk model allows the user to specify other relationships between ψ_1 and ψ_2 . The two logical choices would be to specify a dominant ($\psi_1 = \psi_2$) or a recessive ($\psi_1 = 1, \psi_2 \neq 1$) model.

Another built in alternate risk model allows for the association of disease with increased ancestry of one population to vary with a continuous variable, c . Similar to Equation 7, this is calculated as the product of likelihood ratios for each individual across the entire population at locus j :

$$L(k)_j = \frac{\prod_i P(g_1 \neq k \cap g_2 \neq k) + P(g_1 = k \oplus g_2 = k) c \psi_1 + P(g_1 = k \cap g_2 = k) c \psi_2}{\prod_i (1-G)^2 + 2G(1-G) c \psi_1 + G^2 c \psi_2}, \quad (9)$$

where the default weight of c comparing ψ_1 and ψ_2 is $2\psi_1 = \psi_2$. This locus-genome statistic is an undocumented feature of ANCESTRYMAP (Cheng *et al.*, 2009).

Our software also allows the user to specify their own alternate locus-genome statistic models in the form of an R function. States of the MCMC chain can be saved to allow this process to be handled independently from the MCMC calculations. Locus-genome statistics are calculated for each iteration of the MCMC chain and averaged over all samples from the chain.

Regression modeling Global ancestry estimates are reported for each individual and each individuals parents as well as ancestral state probabilities for each individual. With this information the user can also perform any generalized linear regression technique according to the framework proposed by Redden *et al.* (2006):

$$Y_i \sim \beta_0 + \beta_1 G_i + \beta_2 G_{1i} G_{2i} + \beta_3 P(g_{ij1} = k \oplus g_{ij2} = k) + \beta_4 P(g_{ij1} = k \cap g_{ij2} = k) + \dots \quad (10)$$

This modeling framework includes terms for global ancestry for the population of interest, G_i , the product of each parents global ancestry proportion for the population of interest, $G_{1i} * G_{2i}$, and the probability of one or two alleles from the population of interest at the specified locus, $P(g_{ij1} = k \oplus g_{ij2} = k)$ and $P(g_{ij1} = k \cap g_{ij2} = k)$. The model should also be weighted by a measure of the inverse variance of the estimates of, G_i , G_{1i} and G_{2i} (Redden *et al.*, 2006).

2.4 Simulations and Power

2.4.1 Control populations Random recombination rates, λ , and ancestral proportions, G , were sampled for parents of 200 individuals. Values for λ were Gamma distributed with a mean of 6 and standard deviation of 2, and values for G were Beta distributed with a mean of 0.82 and standard deviation of 0.1. These parameters were chosen to simulate a typical African American sample. A random chromosome from the YRI and CEU populations of the HapMap was randomly chosen as a function G . Each time a recombination is randomly sampled between two markers, based on individual values for λ , a new chromosome was chosen in a like manner. In this way, admixed chromosomes were simulated with appropriate admixture linkage patterns across the chromosome. The CEU and YRI populations were also used as surrogate populations, but with the prior estimates for p and β randomly modified to simulate imperfect surrogates. This was done by taking a sample from the prior distributions of P and B with $\tau = 300$.

A sample of 100 individuals from each simulation above was analyzed using ALDSuite, MULTIMIX and PCAdmix (Brisbin *et al.*, 2012; Churchhouse and Marchini, 2013), and the proportion of correct and incorrect inferences are reported.

2.5 Empirical Data

The ASW population from the International HapMap Project were analyzed using YRI and CEU populations as surrogate ancestral populations. These populations were analyzed using ALDSuite as well as MULTIMIX and PCAdmix (Brisbin *et al.*, 2012; Churchhouse and Marchini, 2013), and a representative sample of the results on chromosome 20 are shown.

2.6 Additional Tools

Several tools are included in the R package, additional to the local ancestry inference and disease association statistics described above. These include input and output data formatting aids, quality control and analysis of the data, and useful data sets. Formatting functions are available for generating prior parameter estimates for different populations using HapMap populations contained in the ALDdata package, and calculation of genetic distance in humans is performed using one of several maps, including the International HapMap Project and those generated by Matise *et al.* (International HapMap Consortium, 2007; Matise *et al.*, 2007; Nato *et al.*, 2013). Error checking functions for quality control measures discussed in the Error Checking section are included as well as some basic graphics. Additional downstream statistical analysis and custom generation of graphics using the diverse and powerful toolset provided by R is also directly available (R Development Core Team, 2013).

3 DISCUSSION

ALDSuite provides accurate inference of local ancestry, while indirectly modeling local, higher order LD remaining from ancestral populations. The analysis of our simulation resulted in 96.3% accuracy of local ancestry inference, compared to the 98.1% accuracy of PCAdmix and the 98.7% accuracy of MULTIMIX, which is on par with other leading analysis software (Price *et al.*,

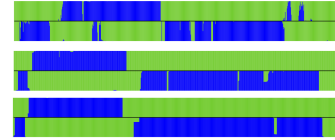


Fig. 2. Representative chromosomes from one individual in the ASW population. Chromosomal ancestry along chromosome 20 is shown for ALDSuite (top), PCAdmix (middle) and MULTIMIX (bottom). African ancestry confidence is represented by green bars and European ancestry confidence by blue bars with one phased haplotype stacked on top of the second phased haplotype.

Table 1. Currently available admixture inference software.

Software	Dense Markers	MALD	>2 Pops	Cited	Ref
STRUCTURE			✓	3278	
ADMIXMAP		✓		189	
ANCESTRYMAP		✓		338	
FRAPPE	✓		✓	227	
SABER+	✓		✓	141	
LAMP-LD	✓		✓	125	
HAPAA	✓		✓	41	
SWITCH-MHMM	✓		✓	29	
WINPOP	✓		✓	41	
HAPMIX	✓			174	
ADMIXTURE	✓		✓	189	
PCAdmix	✓		✓	4	
MULTIMIX	✓		✓	6	
SEQMIX	✓				
ALDER	✓		✓	6	
RFMix	✓		✓		
ALLOY	✓		✓	1	
ELIA	✓		✓		
DBM-Admix	✓		✓		
MaCH-Admix	✓		✓	8	

Analysis of dense marker data, inclusion of disease association statistics, number of supported populations and number of citations listed on GoogleScholar as of December 10, 2013 are listed.

2009; Yang *et al.*, 2013). Comparison of chromosomes from an analysis of the ASW population using ALDSuite, MULTIMIX and PCAdmix also shows a good degree of concordance between the methods used (see Figure 2).

Of the seven MALD studies published in 2013 that we were able to identify, four used sparse marker panels for disease gene mapping (Divers *et al.*, 2013; Bensen *et al.*, 2013; Kim-Howard *et al.*, 2013; Molineros *et al.*, 2013), at least one of which explicitly thinned their dense marker data to accommodate the software used (Kim-Howard *et al.*, 2013). An additional 15 GWAS studies we identified from 2013 used various software listed in Table 1 to control for population substructure resulting from admixture, mostly using dense marker strategies (citations not listed here). This trend highlights the need for a dense marker software package that, like most sparse marker software, includes disease association statistics for MALD.

While sparse marker panels are more cost effective and have proven powerful in the detection several important disease risk genes, dense data provide more accurate ancestry inference and a finer resolution of recombination points (Tang *et al.*, 2006). One strategy that has been used is to follow up a MALD study with fine typing around an associated locus (Nelson *et al.*, 2010). One use potential use of ALDSuite is to analyze both sparse and dense marker data in combination, resulting in better global ancestry estimates, while being able to infer local ancestry on a much finer scale in areas of particular interest.

4 CONCLUSION

Admixture inference software can be categorized using a few different metrics including the number of admixing populations it can simultaneously infer, the way it models local LD when analyzing dense marker data, the number of admixing populations it will simultaneously infer and support of disease gene mapping (see Table 1). There are currently no software solutions which both offer analysis of dense marker data from more than two admixing populations and disease association statistics, requiring the use of several software programs, often with very different input and output data formats. ALDSuite offers a fast, accurate estimation of global and local ancestry with the tools needed from data quality control through mapping of disease genes, along with the rich statistical and graphical utilities provided with R.

ACKNOWLEDGEMENT

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN26120080001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

REFERENCES

Baran, Y., Pasaniuc, B., Sankaranarayanan, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintrón, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., and Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, **28**(10), 1359–1367.

Bensen, J. T., Xu, Z., McKeigue, P. M., Smith, G. J., Fonham, E. T. H., Mohler, J. L., and Taylor, J. A. (2013). Admixture mapping of prostate cancer in African Americans participating in the North Carolina-Louisiana Prostate Cancer Project (PCaP). *The Prostate*.

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., and Bustamante, C. D. (2012). PCAmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human biology*, **84**(4), 343–364.

Cheng, C.-Y., Kao, W. H. L., Patterson, N., Tandon, A., Haiman, C. A., Harris, T. B., Xing, C., John, E. M., Ambrosone, C. B., Brancati, F. L., Coresh, J., Press, M. F., Parekh, R. S., Klag, M. J., Meoni, L. A., Hsueh, W.-C., Fejerman, L., Pawlikowska, L., Freedman, M. L., Jandorf, L. H., Bandera, E. V., Ciupak, G. L., Nalls, M. A., Akylbekova, E. L., Orwoll, E. S., Leak, T. S., Miljkovic, I., Li, R., Ursin, G., Bernstein, L., Ardlie, K., Taylor, H. A., Boerwinkle, E., Zmuda, J. M., Henderson, B. E., Wilson, J. G., and Reich, D. (2009). Admixture mapping of 15,280 African

Americans identifies obesity susceptibility loci on chromosomes 5 and X. *PLoS Genetics*, **5**(5), e1000490.

Churchhouse, C. and Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology*, **37**(1), 1–12.

Divers, J., Palmer, N. D., Lu, L., Register, T. C., Carr, J. J., Hicks, P. J., Hightower, R. C., Smith, S. C., Xu, J., Cox, A. J., Hruska, K. A., Bowden, D. W., Lewis, C. E., Heiss, G., Province, M. A., Borecki, I. B., Kerr, K. F., Chen, Y.-D. I., Palmas, W., Rotter, J. I., Wassel, C. L., Bertoni, A. G., Herrington, D. M., Wagenknecht, L. E., Langefeld, C. D., and Freedman, B. I. (2013). Admixture mapping of coronary artery calcified plaque in African Americans with type 2 diabetes mellitus. *Cardiovascular Genetics*, **6**(1), 97–105.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**(4), 1567–1587.

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., and McKeigue, P. M. (2004). Design and analysis of admixture mapping studies. *The American Journal of Human Genetics*, **74**(5), 965–978.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861.

Kim-Howard, X., Sun, C., Molineros, J. E., Maiti, A. K., Chandru, H., Adler, A., Wiley, G. B., Kaufman, K. M., Kottyan, L., Guthridge, J. M., Rasmussen, A., Kelly, J., Sánchez, E., Raj, P., Li, Q.-Z., Bang, S.-Y., Lee, H.-S., Kim, T.-H., Kang, Y. M., Suh, C.-H., Chung, W. T., Park, Y.-B., Choe, J.-Y., Shim, S. C., Lee, S.-S., Han, B.-G., Olsen, N. J., Karp, D. R., Moser, K., Pons-Estel, B. A., Wakeland, E. K., James, J. A., Harley, J. B., Bae, S.-C., Gaffney, P. M., Alarcón-Riquelme, M., on behalf of GENLES, Looger, L. L., and Nath, S. K. (2013). Allelic heterogeneity in NCF2 associated with systemic lupus erythematosus (SLE) susceptibility across four ethnic populations. *Human Molecular Genetics*.

Kopp, J. B., Smith, M. W., Nelson, G. W., Johnson, R. C., Freedman, B. I., Bowden, D. W., Oleksyk, T., McKenzie, L. M., Kajiyama, H., Ahuja, T. S., Berns, J. S., Briggs, W., Cho, M. E., Dart, R. A., Kimmel, P. L., Korbet, S. M., Michel, D. M., Mokrzycki, M. H., Schelling, J. R., Simon, E., Trachtman, H., Vlahov, D., and Winkler, C. A. (2008). MYH9 is a Major-Effect Risk Gene for Focal Segmental Glomerulosclerosis. *Nature Genetics*, **40**(10), 1175–1184.

MacLean, C. J. and Workman, P. L. (1973). Genetic studies on hybrid populations. I. Individual estimates of ancestry and their relation to quantitative traits. *Annals of Human Genetics*, **36**(3), 341–351.

Maindonald, J. H. (2013). The hwd Package. <http://cran.r-project.org/web/packages/hwd/>.

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*.

Matise, T. C., Chen, F., Chen, W., De La Vega, F. M., Hansen, M., He, C., Hyland, F. C. L., Kennedy, G. C., Kong, X., Murray, S. S., Ziegler, J. S., Stewart, W. C. L., and Buyske, S. (2007). A second-generation combined linkage physical map of the human genome. *Genome Research*, **17**(12), 1783–1786.

McKeigue, P. M. (1997). Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *The American Journal of Human Genetics*, **60**(1), 188.

Molineros, J. E., Maiti, A. K., Sun, C., Looger, L. L., Han, S., Kim-Howard, X., Glenn, S., Adler, A., Kelly, J. A., Niewold, T. B., Gilkeson, G. S., Brown, E. E., Alarcón, G. S., Edberg, J. C., Petri, M., Ramsey-Goldman, R., Reveille, J. D., Vilá, L. M., Freedman, B. I., Tsao, B. P., Criswell, L. A., Jacob, C. O., Moore, J. H., Vyse, T. J., Langefeld, C. L., Guthridge, J. M., Gaffney, P. M., Moser, K. L., Scofield, R. H., Alarcón-Riquelme, M. E., BIOLUPUS Network, Williams, S. M., Merrill, J. T., James, J. A., Kaufman, K. M., Kimberly, R. P., Harley, J. B., and Nath, S. K. (2013). Admixture mapping in lupus identifies multiple functional variants within IFIH1 associated with apoptosis, inflammation, and autoantibody production. *PLoS Genetics*, **9**(2), e1003222.

Murray, L. (2010). Distributed Markov chain Monte Carlo. In *LCCC: NIPS Workshop on Learning on Cores, Clusters and Clouds*.

Nato, A. J., Buyske, S., and Matise, T. C. (2013). The Rutgers Map: A third-generation combined linkage-physical map of the human genome. *In Preparation*.

Nelson, G. W., Freedman, B. I., Bowden, D. W., Langefeld, C. D., An, P., Hicks, P. J., Bostrom, M. A., Johnson, R. C., Kopp, J. B., and Winkler, C. A. (2010). Dense mapping of MYH9 localizes the strongest kidney disease associations to the region of introns 13 to 15. *Human Molecular Genetics*, **19**(9), 1805–1815.

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., Ferrell, R. E., and Shriver, M. D. (1998). Estimating African American admixture proportions by use of population-specific alleles. *The American Journal of Human Genetics*, **63**(6), 1839–1851.

- Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J., and Reich, D. (2004). Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, **74**(5), 979–1000.
- Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., Goldstein, D. B., and Reich, D. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *The American Journal of Human Genetics*, **83**(1), 132–135.
- Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, **5**(6), e1000519.
- R Development Core Team (2013). R: A language and environment for statistical computing. *Manual*. <http://cran.r-project.org>.
- Redden, D. T., Divers, J., Vaughan, L. K., Tiwari, H. K., Beasley, T. M., Fernández, J. R., Kimberly, R. P., Feng, R., Padilla, M. A., Liu, N., Miller, M. B., and Allison, D. B. (2006). Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. *PLoS Genetics*, **2**(8), e137.
- Seldin, M. F., Pasaniuc, B., and Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, **12**(8), 523–528.
- Smith, M. W. and O'Brien, S. J. (2005). Mapping by Admixture Linkage Disequilibrium: Advances, Limitations and Guidelines. *Nature Genetics*, **6**, 623–632.
- Sundquist, A., Fratkin, E., Do, C. B., and Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, **18**(4), 676–682.
- Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, **79**(1), 1–12.
- Thoday, J. M. (1969). Limitations to genetic comparison of populations. *Journal of biosocial science*, **Suppl 1**, 3–14.
- Wu, X.-L., Sun, C., Beissinger, T. M., Rosa, G. J., Weigel, K. A., Gatti, N. d. L., and Gianola, D. (2012). Parallel Markov chain Monte Carlo - bridging the gap to high-performance Bayesian computation in animal breeding and genetics. *Genetics, selection, evolution : GSE*, **44**, 29.
- Yang, J. J., Li, J., Buu, A., and Williams, L. K. (2013). Efficient Inference of Local Ancestry. *Bioinformatics*.

7. Additional publications

Harris M, Burns CM, Becker EA, Braasch AT, Gostick E, **Johnson RC**, Broman KW, Price DA, Friedrich TC, O'Connor SL. **Acute-phase CD8 T cell responses that select for escape variants are needed to control live attenuated simian immunodeficiency virus**, *Journal of Virology*, 2013, 87 (16), 9353–9364

Torres AR, Westover JB, Gibbons C, **Johnson RC**, Ward DC. **Activating killer-cell immunoglobulin-like receptors (KIR) and their cognate HLA ligands are significantly increased in autism**, *Brain, Behavior and Immunity*, 2012, 26 (7), 1122-7

Kopp JB, Nelson GW, Sampath K, **Johnson RC**, Genovese G, An P, Friedman D, Briggs W, Dart R, Korbet S, Mokrzycki M, Kimmel PL, Limou S, Ahuja TS, Berns JS, Simon E, Smith MC, Trachtman H, Michel DM, Schelling JR, Vlahov D, Pollak M, Winkler CA. **APOL1 trypanolytic sequence variants in focal segmental glomerulosclerosis and HIV-associated nephropathy: genetic risk and clinical phenotype**, *Journal of the American Society of Nephrology*, 2011, 22 (11), 2129-37

Freedman BI, Langefeld CD, Lu L, Divers J, Comeau ME, Kopp JB, Winkler CA, Nelson GW, **Johnson RC**, Palmer ND, Hicks PJ, Bostrom MA, Cooke JN, McDonough CW, Bowden DW. **Differential effects of MYH9 and APOL1 risk variants on FRMD3 association with diabetic ESRD in African Americans**, *PLoS Genetics*, 2011, 7(6), e1002150

Troyer JL, Nelson GW, Lautenberger JA, Chinn L, McIntosh C, **Johnson RC**, Sezgin E, Kessing BD, Malasky M, Hendrickson S, Guan L, Minzhong T, An P, Winkler CA, Limou S, Le Clerc S, Delaneau O, Zagury J-F, Schuitmaker H, van Manen D, Bream JH, Gompert. **PARD3B-based AIDS restriction discovered by genome-wide association study**, *Journal of Infectious Disease*, 2011, 203 (10), 1491-502

Nelson GW, Freedman BI, Bowden DW, Langefeld CD, An P, Hicks PJ, Bostrom MA, **Johnson RC**, Kopp JB, Winkler CA. **Dense mapping of MYH9 localizes the strongest kidney disease associations to the region of introns 13 to 15**, *Human Molecular Genetics*, 2010, 19 (9), 1805-15

O'Connor SL, Lhost JJ, Becker EA, Detmer AM, **Johnson RC**, Macnair CE, Wiseman RW, Karl JA, Greene JM, Burwitz BJ, Bimber BN, Lank SM, Tuscher JJ, Mee ET, Rose NJ, Desrosiers RC, Hughes AL, Friedrich TC, Carrington M, O'Connor DH. **MHC heterozygote advantage in simian immunodeficiency virus-infected Mauritian cynomolgus macaques**, *Science Translational Medicine*, 2010, 2 (22), 22ra18

An P, **Johnson RC**, Phair J, Kirk GD, Yu XF, Donfield S, Buchbinder S, Goedert JJ, Winkler CA. **Loss of APOBEC3B Gene Increases Risk of HIV-1 Acquisition**, *Journal of Infectious Diseases*, 2009, 200 (7), 1054-8

Roelke ME, Brown MA, Troyer JL, Winterbach H, Winterbach C, Hemson G, Smith D, **Johnson RC**, Pecon-Slattery J, Roca AL, Alexander K, Klein L, Martinelli P, Krishnasamu K, O'Brien SJ. **Pathological Manifestations of Feline Immunodeficiency Virus (FIV) Infection in Wild African Lions**, *Virology*, 2009, 390 (1), 1-12

Streeck H, Jolin JS, Qi Y, Yassine-Diab B, **Johnson RC**, Kwon DS, Addo MM, Brumme C, Routy JP, Little S, Jessen HK, Kelleher AD, Hecht FM, Sekaly RP, Rosenberg ES, Walker BD, Carrington M, Altfeld M. **HIV-1-Specific CD8+ T Cell Responses During Primary Infection are Major Determinants of the Viral Set Point and Loss of CD4+ T Cells**, *Journal of Virology*, 2009, 83 (15), 7641-8

Winkler CA, An P, **Johnson RC**, Nelson GW, Kirk G. **Expression of Duffy Antigen Receptor for Chemokines (DARC) has no Effect on HIV-1 Acquisition or Progression to AIDS in African Americans**, *Cell Host & Microbe*, 2009, 5 (5), 411-13

Guo X, **Johnson RC**, Deng H, Liao J, Guan L, Nelson GW, Tang M, Zheng Y, de The G, O'Brien SJ, Winkler CA, Zeng Y. **Evaluation of Nonviral Risk Factors for Nasopharyngeal Carcinoma in a High-Risk Population of Southern China**, *International Journal of Cancer*, 2009, 124 (12), 2942-7

Discussion

As previously stated, the overarching goal of this dissertation is to increase power in genome-wide association studies (GWAS) of underrepresented populations by more fully taking advantage of and more accurately modeling of genetic linkage. Section 5 explores different options for assessing statistical significance in a traditional GWAS, showing that many ideas for setting significance thresholds are anti-conservative, lending credence to a more standard Bonferroni correction between 5×10^{-8} to 1×10^{-8} . It is important to note, however, that some slight increase in power can be obtained by modeling linkage disequilibrium (LD) with a principal components analysis (PCA) to obtain a more accurate number of independent tests performed in the GWAS (R. C. Johnson et al., 2010). Section 6 describes a new admixture inference software package, ALDsuite, written as a package for use in R (R Development Core Team, n.d.). ALDsuite models local LD more accurately than previous admixture software, using a principal components (PC) regression. This also allows considerable computational savings, even as the density of the marker set increases dramatically (R. C. Johnson, Winkler, & Nelson, n.d.).

As the cost of genotyping and sequencing has increasingly fallen, the flow of genetic information has increased substantially and continues to do so at an ever-increasing pace. This wealth of data greatly increases our ability to understand determinants of disease, but geneticists are also faced with new problems stemming from the number of statistical comparisons needed to fully explore their data. These multiple comparisons result in many more Type I errors (false positive hypothesis tests) when using a traditional significance threshold, $\alpha = 0.05$. The obvious solution to this problem is to lower the significance threshold to achieve an acceptable study wide Type I error rate, where the likelihood of false positives in the entire study is sufficiently low. This, however, raises the Type II error rate (a measure of false negative hypothesis tests) and with it, the prospect of throwing out potentially important findings because they cannot be differentiated from false positive results in the study (Neyman & Pearson, 1967).

The most fundamental way to combat this decrease in statistical power is to sample more individuals, thereby taking advantage of the unique relationship between the central limit

theorem (CLT) and the law of large numbers (LLN). Simply put, the CLT states that the distribution of the sample average will be approximately normally distributed (Pólya, 1920), and the LLN implies that the variance of the sample average will decrease with the size of the sample (Bernoulli, 1713; Tchebichef, 1846). Thus, with an increased sample size, statistical noise of the association statistics will decrease and result in more accurate inferences. The speed of this convergence and the associated reduction in variance is order $n^{-1/2}$, so for larger samples a larger increase in sample size is required to achieve the same level of improvement (Berry, 1941). For example, with an equal number of cases and controls, a disease associated allele with frequency 0.1 and odds ratio 1.7, and a significance threshold of 5×10^{-8} , we would need to increase our sample size by 374 individuals to achieve an increase in statistical power from 50% to 60%. To achieve the same 10% power increase from 70% to 80%, however, we would need to increase the sample size by 514 (see Table 8).

On average, this increased accuracy of statistics will result in higher statistical significance for “true” alternate hypotheses, thus helping the researcher to differentiate between meaningful statistical associations and statistical noise. This can be easily achieved when funding and cases are available, with some studies enrolling tens of thousands of cases and as many as 100,000 controls (Monda et al., 2013; Morris et al., 2012). Due to the large sample sizes required to adequately power a GWAS of a rare condition or under restricted budgetary conditions, recruitment of a cohort to meet minimum standards can be an impossible task. In

Power	n
50%	3926
60%	4300
70%	4718
80%	5232

Table 8. GWAS power calculations. Sample size, n, required to achieve various levels of statistical power in a study design aiming to detect an odds ratio of 1.7 in disease associated alleles of 10% frequency that passes GWAS significance threshold of 5×10^{-8} . The calculation assumes a 1:1 case-control match.

fact, a disease need not be rare to have difficulties recruiting cases. Our genome-wide study of HIV which included nearly all HIV infected individuals enrolled in natural history cohorts in the US, as an example, only enrolled 755 patients with known seroconversion dates—the most informative cases for time to event (survival) analysis (Troyer et al., 2011). Even though HIV prevalence is extremely high in some parts of the world, it is much lower in western societies where a majority of funding and infrastructure are available to support large studies.

8. Modification of Type I and Type II error rates

With the recent technological advances and international efforts to identify the spectrum of genetic variation in multiple human populations, SNP arrays have become both denser and contain a wider range of allele frequencies. The denser arrays allow for finer resolution of haplotype structure and the potential for fine mapping for causal variants. Because of the many p values generated in GWAS, dense arrays have exacerbated case-control studies' power disadvantage. In an attempt to increase power over standard Bonferroni corrections required in a hypothesis generating scenario, many alternate significance thresholds and strategies have been recommended (Benjamini & Hochberg, 1995; Misawa et al., 2008; Skol et al., 2006; Storey & Tibshirani, 2003). Some have sought to take particular advantage of SNPs in haplotype blocks with high levels of linkage (Duggal et al., 2008). Unless perfectly linked, however, these SNPs retain some level of statistical independence and need to be counted as independent tests, and as demonstrated in Section 5, alternates to the standard significance threshold of 5×10^{-8} only offer a modest improvement in power (R. C. Johnson et al., 2010). There is now a general consensus setting the standard 5×10^{-8} significance threshold for GWAS, stemming from studies showing the number of independent tests performed in a GWAS of common genetic variants to be on the order of one million (Anttila et al., 2010; Hoggart et al., 2008; Kamatani et al., 2010; Kathiresan et al., 2008). Many still remain unsatisfied with Type II error rates in GWAS and are either seeking other methods to reduce this burden or are advocating a change in the current standard, including false discovery rates and false positive report probabilities, bayes factors and data reduction methods, which are

covered in more detail in Section 1.3.4. One group is currently advocating the elimination of the use of significance thresholds. Relying on the fact that genetics affects health outcomes, they presume that if enough tests are performed, those associations will be found, and that the most significant test results in a large study with multiple hypotheses, regardless of the observed significance, will likely be true associations (Zaykin, Kuo, & Vsevolozhskaya, 2013). While these approaches are not bad if independent replication of results is obtained and can actually be beneficial in resource limited studies, statistical power is greatest when including all samples together, combining both discovery and replication cohorts, rather than staging for independent validation (Skol et al., 2006).

A visual analysis of replication rates of all GWAS results published on <http://www.genome.gov> is shown as a function of replication in Figure 20. A review of these data revealed that 75% of results that do not pass the $p < 5 \times 10^{-8}$ significance threshold have not been validated in subsequent replication studies. By comparison, only 28% of GWAS results

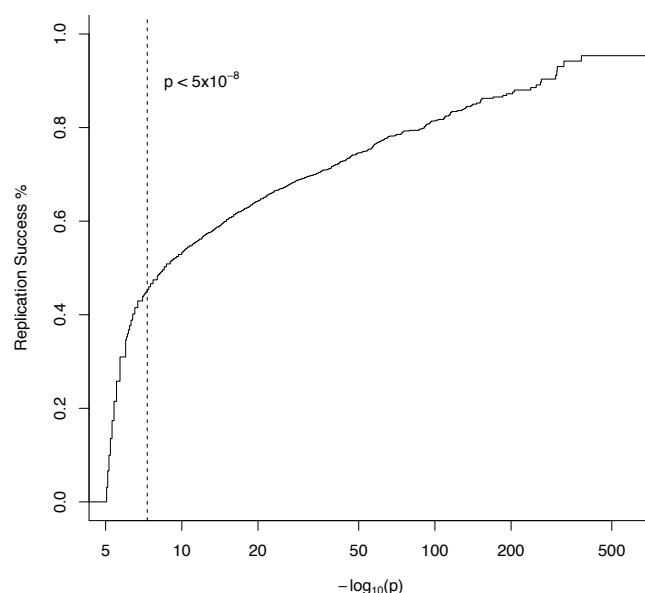


Figure 20. Replication rates of all GWAS results. An inverse survival curve represents the proportion of replicated GWAS results as a function of the significance of the original finding. The vertical dashed line represents the significance threshold of

5×10^{-8} . Data were accessed from <http://www.genome.gov> on October 29, 2013 (Hindorff et al., n.d.).

passing this conservative threshold have not been replicated (Hindorff et al., n.d.). Reporting of borderline significant results is clearly a major cause for the lack of reproducibility in GWAS findings to date, but there are still clinically relevant results in this range (Panagiotou, Ioannidis, Genome-Wide Significance Project, 2012). Given the cost of Type I errors in the form of time and money spent by other labs to replicate findings or to perform follow-up functional studies, this has led to some investigators questioning of the worth of these studies, especially when the size of the effects is taken into consideration. One GWAS, for example, identified 20 genetic markers associated with adult height, but estimated that they only accounted for 3% of the population variation (Weedon et al., 2008). A very crude extrapolation based on this estimate concluded that as many as 93,000 further markers would need to be identified in order to explain the genetic heritability of height (Goldstein, 2009). Other investigators have pointed to common diseases already shown to be caused by rare variants and suggested that association of disease with common variants is a less efficient use of resources (Goldstein, 2009; McClellan & King, 2010). An example of common disease caused by rare variants is breast cancer. Germline mutations in at least ten genes have been shown to increase risk of breast cancer, including thousands of mutations observed in BRCA1 alone (Walsh & King, 2007). Others have implicated epigenetic factors in an attempt to explain GWAS' perceived failures (Crow, 2011).

Other investigators, however, have noted that GWAS gene discovery have led to new drug targets, identified previously unsuspected pathological pathways, and resulted in novel treatments of disease (Visscher, Brown, McCarthy, & Yang, 2012). Additional examples from GWAS discoveries highlight the role of balancing selection between common alleles that confer both benefits and increased risks (Klein, Xu, Mukherjee, Willis, & Hayes, 2010). Examples of these include *TCF2* variants that both confer protection against Crohn's disease while increasing risk of type 1 diabetes (K. Wang et al., 2010), and the *APOL1* variants discussed throughout this dissertation, that both protect against trypanosome infection and

increases risk of kidney disease (although *APOLI* was mapped by admixture linkage rather than a traditional GWAS (Genovese et al., 2010; Jeffery B Kopp et al., 2008)).

Whether in a GWAS with a few million common variants or a whole-genome sequence data analysis with tens of millions of variants, ascertainment of statistical significance in case-control studies remains critically important. In fact, as we enter into the whole-genome sequence analysis era, where the number of independent tests, including rare variants, can be much larger than in a traditional GWAS, this will likely gain in importance. It is widely accepted that the most effective way to reduce publication of false positive associations is the identification of linked functional, causal variant(s); or failing that, replication of results in a collaborative atmosphere prior to publication (GAIN Collaborative Research Group et al., 2007). Recognizing the problem of both false positives and false negatives, many groups are forming large international consortia to increase sample number, thereby increasing power to detect low frequency and/or small effect variants that would not reach statistical significance in smaller studies (McLaren et al., 2013; Schork, Greenwood, & Braff, 2007; Wellcome Trust Case Control Consortium, 2007). However, it is fundamental that confirming statistical significance by replication requires accurate methods for estimating significance, in the individual and combined studies. In GWAS and similar studies this requires accurately estimating the true number of independent comparisons.

The simulations presented in this dissertation show that a simple method, employing a principal components analysis of sampled genetic data to assess the number of independent tests provides an estimate of the number of comparisons that is neither under- nor over-conservative, while some other published methods may be dangerously under-conservative. This study, published in BMC Genomics (R. C. Johnson et al., 2010), makes an important contribution to the field, because it provides empirical evidence through simulations that a principal components analysis of sampled data provides a basis for setting genome-wide significance levels for a given study. It also provides objective criteria for investigators to evaluate the noteworthiness of their discoveries and for reviewers and readers to interpret the level of statistical support for reported associations.

9. Mapping by Admixture Linkage Disequilibrium

An underused option for increasing power is mapping by admixture linkage disequilibrium (MALD), which combines the power gained from long range ancestry linkage disequilibrium (ALD) similar to that found in family studies with the ability of case-control studies to model complex disease (McKeigue, 1997). It is important to note that MALD assumes the existence of a causal genetic factor, more frequent in one ancestral population compared to the other ancestral population(s) contributing to the admixture. Importantly, computer simulations exploring the effect of the difference in disease allele frequency between ancestral populations indicates that statistical power remains stable with as little as a 20% difference in disease prevalence between populations (J. C. Stephens et al., 1994).

A further benefit of MALD is that the association sought after is not between phenotype and a specific marker in linkage disequilibrium (LD) with the functional genetic variant, but rather an association between phenotype and local ancestry at the locus harboring the functional genetic variant (Chakraborty & Weiss, 1986). Notably, the *APOL1* variants profoundly affecting kidney disease were discovered only after an admixture linkage study identifying the C22 q13.1 region with glomerular disease and non-diabetic end stage renal disease. The causal alleles are only found in Africans, where they are under selection by trypanosomes, and in admixed populations with recent African ancestry (B. I. Freedman et al., 2011; 2010; Genovese et al., 2010; Jeffrey B Kopp et al., 2011). This gene is strongly associated with end stage renal disease (OR=7), focal segmental glomerulosclerosis (OR=17), and HIV-associated nephropathy (OR=29), as well as increased rate of progression to end stage renal disease in persons with chronic kidney disease (Jeffrey B Kopp et al., 2011). This locus, however, was not found by prior GWAS studies for chronic and end stage kidney disease enrolling many more people because the causal variants were not tagged by markers included in the earlier GWAS (Bostrom et al., 2010; Genovese et al., 2010; McDonough et al., 2011). We were able

to identify the locus using MALD because it showed an excess of African ancestry compared to the rest of the genome in cases and when compared to the control group at that locus using under 200 cases (Jeffery B Kopp et al., 2008).

The lack of complete coverage of genetic variation in non-European populations is being remedied by the 1000 Genomes project; but sequencing of each individual in studies is the only way to capture the true spectrum of variants contributing to disease (Manry & Quintana-Murci, 2013). This is particularly important in light of the impact of rare variants on some common diseases, as discussed in Section 8. Although African American admixed populations have been used to map several disease phenotype, there are many potentially informative admixed populations that remain less well genetically characterized. Examples include Latino and Cape Coloured populations. Latino populations are three-way admixed between European and Native American populations with a small but significant African component, and are more heterogeneous than African Americans due to varied Native American input spanning a broad geographical range (A. L. Price et al., 2007). Several atopic hypersensitivity disorders have higher prevalence in Latino populations. Recently, admixture mapping was used to identify genetic variants associated with atopy in Latino children (Kumar et al., 2013). The Cape Coloured population, which stems from the interaction of many peoples in the heavily traveled route from Europe to Asia, is composed of admixture between Khoisan and Bantu Africans, Europeans, South Asians and East Asians (Tishkoff et al., 2009). While most readily available SNP arrays were developed to capture European genetic diversity, there is enough information in these arrays to easily distinguish ethnic groups (A. L. Price et al., 2006; 2007; Tishkoff et al., 2009).

Early admixture studies relied entirely on sparse panels of ancestry informative markers (AIMs), many derived from SNP arrays, to map disease genes (M. L. Freedman et al., 2006; Jeffery B Kopp et al., 2008; Reich et al., 2005). The three software packages for analysis of sparse AIMs panels discussed in Section 2.2 remain valuable and are still in use (McKeigue, Carpenter, Parra, & Shriver, 2000; Patterson et al., 2004; Pritchard et al., 2000). In fact, a review of publications in 2013 revealed some striking trends with respect to the use of

admixture software (see Appendix A). Of the seven publications that were distinctly identified as MALD studies, four used sparse marker panels, at least one of which had dense marker data available and pruned the set back to meet local LD requirements of the software (Kim-Howard et al., 2013). Of the remaining three MALD studies, all three used SABER+ and one verified the SABER+ results with a combination of several other dense marker softwares (see Appendix A for remaining references).

Other algorithms able to analyze dense marker sets with local LD, including SABER+, HAPMIX, LAMP-LD and HAPAA, seem to be more widely used for control of local ancestry in 14 GWAS, many of which also employed sparse marker strategies to estimate global ancestry as well (see Appendix A for complete list of references). Another common use for dense marker software is in characterizing the genetics in population history studies (see Appendix A). It is possible that these preferences have developed partly as a result of the output offered by these programs. The sparse marker software packages used for MALD in this survey both incorporate disease association statistics in their output, while the more commonly used dense marker software outputs only local ancestry estimates (see Appendix B). Other advantages of sparse marker software are mostly related to the cost of genotyping, which is becoming less of a factor with recent advances in genotyping and sequencing technologies. When dense marker data or sequence data are available, however, dense marker strategies provide more accurate inference of local ancestry and a finer resolution of each recombination point in each individual (Tang et al., 2006).

Accurate modeling of local LD has proven difficult, and research in this area remains active, as evidenced by the fact that nearly 40% of the 65 papers referenced in Appendix A describe new ancestry inference methods. Perhaps the most difficult problem with the modeling of local LD is finding a computationally efficient algorithm (Seldin et al., 2011). The most common solution to this problem is to discretize the sampled markers into windows of predetermined length (see Appendix B). This decreases the complexity of modeling local LD substantially, resulting in more efficient scalable software. A good example of this benefit is the recently described RFMix package. By considering windows of markers it is able to

achieve a ~30 fold decrease in run time as compared to HAPMIX, which models local LD in a more continuous fashion (Maples et al., 2013; A. L. Price et al., 2009).

ALDsuite, described in Section 6, also uses a type of windowing to achieve better computational efficiency, but the window sizes and the markers included in each window are less arbitrary than in other software. A preprocessing module is available that picks a set of representative markers that are both minimally spaced with respect to their distance (default is 0.1 cM) and maximally differentiates between the populations under study. In special cases this may be altered to fit specific study designs, such as a MALD study looking specifically for Native American loci associated with disease, in which differentiation between African and European ancestry is less important (A. L. Price et al., 2007). All markers within a short distance (default is 0.1 cM) and any additional markers that are significantly linked in user provided training data or surrogate ancestral populations in the International HapMap Project are used to infer the likelihood of markers in the analysis set, conditional on each possible local ancestral state, using principal components (PC) regression model for each ancestral population (International HapMap Consortium et al., 2007; R. C. Johnson et al., n.d.). Additional information from the 1000 Genomes Project would also be worth adding to the reference database that ALDsuite refers to (1000 Genomes Project Consortium et al., 2010). The use of PC regression modeling of the ancestry conditioned likelihood of each marker in the analysis set was chosen as an indirect model of local LD for two main reasons: 1) A conditional PC regression captures a measure of the underlying LD while maintaining good statistical properties, and 2) A conditional PC regression has excellent computational properties.

As noted in Section 2.2.5, incorrect modeling of local LD can lead to incorrect inference of the number of generations since admixture (Wall & Pritchard, 2003). This is the primary reason sparse marker software requires markers to be unlinked in ancestral populations (McKeigue et al., 2000). Including a two-way LD measure in the calculation was attempted in the first dense marker software package, SABER, but was shown to allow long range LD bias into the model (A. L. Price et al., 2008; Tang et al., 2007). Rather than directly modeling

higher order local LD, which becomes exponentially complex as marker density increases, we posited that modeling higher order LD with a PC regression would both control for extant local LD and provide a computational benefit. Directly modeling every possible haplotype in a window can also be computationally demanding (A. G. Clark et al., 1998; Excoffier & Slatkin, 1995), and modeling one representative marker as a function of the other markers in the haplotype can suffer from multicollinearity associated with tightly linked markers. As noted in Section 3.1 principal components analysis (PCA) is an excellent way to extract relevant information from highly correlated variables and avoids statistical instability that otherwise would enter into the model (Hotelling, 1933; K. Pearson, 1901).

The added benefit of using conditional PC regression to model the likelihood of each marker in the analysis set is that it has very good computational properties. As noted in our PCA of GWAS data in Section 5, adding many additional markers to a data set continues to add measurable information (R. C. Johnson et al., 2010). It has also been shown, however, that most of the useable information in a data set can be captured in the first few PCs (Hawkins, 1973). Thus, the number of PCs included in the regression model may be fairly constant, even as the marker density supporting each model grows.

Another possible reason for the observed uses of admixture software in Appendix A is usability. One distinguishing feature of some very successful software packages that have become industry standards and continue to be used years after introduction (e.g. R, STRUCTURE and PLINK), is that they are useable (Falush et al., 2003; Pritchard et al., 2000; Purcell et al., 2007; R Development Core Team, n.d.). In software engineering, usability heuristics play an important roll in the development lifecycle (Seffah & Metzker, 2004). Some heuristics that we have paid especial interest to in the development of ALDsuite include: help and documentation, consistency and standards, user control and freedom, flexibility and efficiency of use, and error prevention (Nielsen, 1994).

Many of these usability heuristics are enhanced by making ALDsuite into an R package. All functions are fully documented in R's documentation system, a vignette is available with worked out examples that users can replicate, graphics generation is consistent with and integrated with R's graphics system, scripting of analyses using ALDsuite will already be familiar to R users, and output can be analyzed using R's powerful statistical functionality. Functions are also available to read in and format data from various sources. Quality control of input data is performed and reported to the user as described in Section 6, and standard defaults are easily modifiable. Perhaps the most significant feature included in ALDsuite is the function to compare a list of provided rs numbers to a database of samples from the International HapMap Project (International HapMap Consortium et al., 2007), automatically pick a maximally informative set of analysis markers, performs a PCA on HapMap samples from representative modern day populations and returns a properly formatted data object for use in ancestry inference. This is provided by some software packages, where required, but takes considerable time for other packages (see Appendix B).

The ALDsuite package presented in this dissertation provides a local admixture inference option that is not only quick and accurate, but also allows efficient use of the powerful statistical and graphical utilities provided in R (R. C. Johnson et al., n.d.; R Development Core Team, n.d.). Furthermore, quality control measures and tools for the generation of a custom, maximally informative prior based on International HapMap Project data make this a complete package, rather than one step in a complicated pipeline. Additional features will continue to be added including tools to query databases from the 1000 genomes project in support of whole genome sequencing studies, and because of the data reduction qualities of PCA, it is anticipated that analysis of these sequence data will remain computationally reasonable while maintaining maximal information.

Conclusions

GWAS is an essential tool for genetic contribution to human disease, but has severe problems of statistical power for diseases for which it is impractical to genetically sample tens of thousands of subjects. For diseases in admixed populations, where the disease is substantially inherited from one ancestral population, the analysis of admixture linkage disequilibrium (ALD) is a strategy to achieve power with relatively small numbers. Since most study populations will be routinely typed for a dense marker set, it is both convenient and a source of additional precision to use the full marker set. However the original and most established algorithms require a sparse marker set to prevent confounding of the admixture disequilibrium measures by local LD. Using dense markers requires accounting for local LD, but this is difficult because of the LD's high dimensionality. ALDsuite, presented in this dissertation, addresses this problem with PCA, correcting the ALD estimate by the major principal components of the ancestral LD. Existing alternate approaches to using dense markers have limitations, including computational demands that do not scale well with super dense data and windowing schemes that do not accurately model local LD in regions where discrete haplotype blocks do not exist. ALDsuite further adds new functionality by considering more general disease association data, in particular survival data.

For full GWAS of modest size, the issue of limited power in the face of $\sim 10^6$ comparisons is inescapable. Lack of correction leads to spurious results, as shown by the failure to replicate 75% of GWAS findings that fail to meet the standard significance threshold of $p < 5 \times 10^{-8}$. To minimize both type I and II errors, it is critical to know the proper correction. It has been optimistically suggested that due to LD the number of truly independent comparisons is much less than $\sim 10^6$ SNPs tested, but my results do not support this: corrections using the number of haplotype blocks, for example are shown by permutation testing to be substantially under conservative, even considering widely differing criteria for defining haplotype blocks. On the other hand the simpleM method, using principal components analysis (PCA), supplies a correction that permutation testing shows to be accurate; this correction is modest, but still useful.

The results presented here—a novel, effective correction for local ancestral population LD allowing use of dense markers in MALD, and the demonstration that the simpleM method is effectively optimum for GWAS multiple comparisons correction, reiterate the value of PCA for capturing the essential part of the complexity of high-dimensional systems. PCA is already standard for correcting for population substructure in GWAS; my results point to its broader applicability as a general strategy for dealing with the high dimensionality of genomic association data.

Linkage is a vital part of genetic association studies, and appropriate modeling of LD is needed to avoid excessive Type I errors and statistical bias. High amounts of LD in the human genome do not preclude the need for a strict, common standard for genome-wide significance. For a traditional GWAS, the replication or failure to replicate 75% of GWAS findings that don't meet the standard significance threshold of $p < 5 \times 10^{-8}$ supports the use of this threshold, but other methods, particularly the simpleM method using principal components analysis (PCA), will likely remain relevant well into the sequencing era. There are sound arguments that other significance thresholds may be appropriate in some cases, but the increase in power is generally quite modest.

One alternate way to increase power, even in the face of increasingly large data sets, is an analysis of admixture linkage disequilibrium (ALD). This form of population substructure can be a source of significant bias, resulting in an increase in the Type I error rate, but can be effectively controlled using a PCA of the genetic data. In fact, rather than being a liability, ALD can be a powerful source of information to identify regions of the genome harboring disease genes. The field has yet to settle on a good algorithm for inferring local ancestry, and ALDsuite fills this need. It models local LD over flexible windows using PCA, rather than in discrete haplotype blocks, allowing the analysis of both sparse and dense marker data. The package is sufficiently computationally efficient, well documented and user friendly, and it provides the tools needed to prepare, analyze and present these complex data.

Bibliography

- 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–1073. doi:10.1038/nature09534
- Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1), 97–101. doi:10.1038/ng786
- Adler, S., Pahl, M., Abboud, H., Nicholas, S., Ipp, E., & Seldin, M. (2010). Mexican-American admixture mapping analyses for diabetic nephropathy in type 2 diabetes mellitus. *Seminars in Nephrology*, 30(2), 141–149. doi:10.1016/j.semnephrol.2010.01.005
- Alton, E. W. F. W., Boyd, A. C., Cheng, S. H., Cunningham, S., Davies, J. C., Gill, D. R., et al. (2013). A randomised, double-blind, placebo-controlled phase IIB clinical trial of repeated application of gene therapy in patients with cystic fibrosis. *Thorax*, 68(11), 1075–1077. doi:10.1136/thoraxjnl-2013-203309
- Anttila, V., Stefansson, H., Kallela, M., Todt, U., Terwindt, G. M., Calafato, M. S., et al. (2010). Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. *Nature Genetics*, 42(10), 869–873. doi:10.1038/ng.652
- Avery, O. T., Macleod, C. M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Introduction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, 79(2), 137–158.
- Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10), 1359–1367. doi:10.1093/bioinformatics/bts144
- Barnhart, B. J. (1989). DOE human genome program. *Human Genome Quarterly*, 1(1).
- Bateson, W., & Kilby, H. (1905). Experimental studies in the physiology of heredity. *Royal Soc Reports to the Evolution Committee*, 2, 1–55.
- Beane, J., Vick, J., Schembri, F., Anderlind, C., Gower, A., Campbell, J., et al. (2011). Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer Prevention Research*, 4(6), 803–817. doi:10.1158/1940-6207.CAPR-11-0212

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1), 189–300.
- Bernoulli, J. (1713). *Ars Conjectandi: Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis*.
- Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1), 122–136.
- Blackwell, D., & Hodges, J. L. (1957). Design for the Control of Selection Bias. *The Annals of Mathematical Statistics*, 28(2), 449–460. doi:10.1214/aoms/1177706973
- Bobadilla, J. L., Macek, M., Fine, J. P., & Farrell, P. M. (2002). Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening. *Human Mutation*, 19(6), 575–606. doi:10.1002/humu.10041
- Bostrom, M. A., Lu, L., Chou, J., Hicks, P. J., Xu, J., Langefeld, C. D., et al. (2010). Candidate genes for non-diabetic ESRD in African Americans: a genome-wide association study using pooled DNA. *Human Genetics*, 128(2), 195–204. doi:10.1007/s00439-010-0842-3
- Browning, B. L. (2008). PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics*, 9, 309. doi:10.1186/1471-2105-9-309
- Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *The American Journal of Human Genetics*, 78(6), 903–913. doi:10.1086/503876
- Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., et al. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences*, 107 Suppl 2, 8954–8961. doi:10.1073/pnas.0914618107
- C elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396), 2012–2018.
- Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., et al. (2005). Demonstrating stratification in a European American population. *Nature Genetics*, 37(8), 868–872. doi:10.1038/ng1607
- Chakraborty, R., & Weiss, K. M. (1986). Frequencies of complex diseases in hybrid populations. *American Journal of Physical Anthropology*, 70(4), 489–503. doi:10.1002/ajpa.1330700408
- Chargaff, E., Zamenhof, S., & Green, C. (1950). Composition of human desoxypentose nucleic acid. *Nature*, 165(4202), 756–757.

- Charkraborty, R., & Weiss, K. M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci (Vol. 85, pp. 9119–9123). Presented at the Proceedings of the National
- Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, 106(45), 19096–19101. doi:10.1073/pnas.0910672106
- Chu, Y., & Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, 22(4), 271–274. doi:10.1089/nat.2012.0367
- Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengård, J., et al. (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *The American Journal of Human Genetics*, 63(2), 595–612. doi:10.1086/301977
- Cleves, M. A., Olson, J. M., & Jacobs, K. B. (1997). Exact transmission-disequilibrium tests with multiallelic markers. *Genetic Epidemiology*, 14(4), 337–347. doi:10.1002/(SICI)1098-2272(1997)14:4<337::AID-GEPI1>3.0.CO;2-0
- Cottingham, R. W., Idury, R. M., & Schäffer, A. A. (1993). Faster sequential genetic linkage computations. *The American Journal of Human Genetics*, 53(1), 252–263.
- Cronin, M., & Ross, J. S. (2011). Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology. *Biomarkers in Medicine*, 5(3), 293–305. doi:10.2217/bmm.11.37
- Crow, T. J. (2011). 'The missing genes: what happened to the heritability of psychiatric disorders?'. *Molecular Psychiatry*, 16(4), 362–364. doi:10.1038/mp.2010.92
- Dahm, R. (2008). The first discovery of DNA. *American Scientist*, 96(4), 320.
- Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., et al. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418(6897), 544–548. doi:10.1038/nature00864
- Didigu, C. A., Wilen, C. B., Wang, J., Duong, J., Secreto, A. J., Danet-Desnoyers, G. A., et al. (2013). Simultaneous zinc-finger nuclease editing of the HIV coreceptors ccr5 and cxcr4 protects CD4+ T cells from HIV-1 infection. *Blood*. doi:10.1182/blood-2013-08-521229
- Duggal, P., Gillanders, E. M., Holmes, T. N., & Bailey-Wilson, J. E. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, 9, 516. doi:10.1186/1471-2164-9-516
- Eisenstein, M. (2012). Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*.

- Elston, R. C., & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, 21(6), 523–542.
- Excoffier, L., & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5), 921–927.
- Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), 1567–1587.
- Fishelson, M., & Geiger, D. (2002). Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18 Suppl 1, S189–98.
- Francalacci, P., Morelli, L., Angius, A., Berutti, R., Reinier, F., Atzeni, R., et al. (2013). Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science*, 341(6145), 565–569. doi:10.1126/science.1237947
- Freedman, B. I., Kopp, J. B., Langefeld, C. D., Genovese, G., Friedman, D. J., Nelson, G. W., et al. (2010). The apolipoprotein L1 (APOL1) gene and nondiabetic nephropathy in African Americans. *Journal of the American Society of Nephrology*, 21(9), 1422–1426. doi:10.1681/ASN.2010070730
- Freedman, B. I., Langefeld, C. D., Lu, L., Divers, J., Comeau, M. E., Kopp, J. B., et al. (2011). Differential effects of MYH9 and APOL1 risk variants on FRMD3 Association with Diabetic ESRD in African Americans. *PLoS Genetics*, 7(6), e1002150. doi:10.1371/journal.pgen.1002150
- Freedman, M. L., Haiman, C. A., Patterson, N., McDonald, G. J., Tandon, A., Waliszewska, A., et al. (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences*, 103(38), 14068–14073.
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576), 2225–2229. doi:10.1126/science.1069424
- GAIN Collaborative Research Group, Manolio, T. A., Rodriguez, L. L., Brooks, L., Abecasis, G., Collaborative Association Study of Psoriasis, et al. (2007). New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genetics*, 39(9), 1045–1051. doi:10.1038/ng2127
- Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., & Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, 34(1), 100–105. doi:10.1002/gepi.20430

- Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32(4), 361–369.
- Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, 329(5993), 841–845. doi:10.1126/science.1193032
- Genovese, G., Handsaker, R. E., Li, H., Kenny, E. E., & McCarroll, S. A. (2013). Mapping the Human Reference Genome's Missing Sequence by Three-Way Admixture in Latino Genomes. *The American Journal of Human Genetics*. doi:10.1016/j.ajhg.2013.07.002
- Goldstein, D. B. (2009). Common genetic variation and human traits. *The New England Journal of Medicine*, 360(17), 1696–1698. doi:10.1056/NEJMp0806284
- Gu, X. (2007, January 1). *Stepwise forward multiple regression for complex traits in high density genome-wide association studies*. Texas Medical Center. Retrieved from <http://digitalcommons.library.tmc.edu/dissertations/AAI3283834>
- Gudbjartsson, D. F., Thorvaldsson, T., Kong, A., Gunnarsson, G., & Ingolfsson, A. (2005). Allegro version 2. *Nature Genetics*, 37(10), 1015–1016. doi:10.1038/ng1005-1015
- Gunderson, K. L., Kruglyak, S., Graige, M. S., Garcia, F., Kermani, B. G., Zhao, C., et al. (2004). Decoding randomly ordered DNA arrays. *Genome Research*, 14(5), 870–877. doi:10.1101/gr.2255804
- Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988), 1684–1689.
- Han, B., Kang, H. M., & Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, 5(4), e1000456. doi:10.1371/journal.pgen.1000456
- Hawkins, D. M. (1973). On the investigation of alternative regressions by principal component analysis. *Journal of the Royal Statistical Society, Series C*, 22(3), 275–286.
- He, W., Neil, S., Kulkarni, H., Wright, E., Agan, B. K., Macroni, V. C., et al. (2008). Duffy Antigen Receptor for Chemokines Mediates trans-Infection of HIV-1 from Red Blood Cells to Target Cells and Affects HIV-AIDS Susceptibility. *Cell Host and Microbe*, 4, 52–62.
- Hendrickson, S. L., Jacobson, L. P., Nelson, G. W., Phair, J. P., Lautenberger, J. A., Johnson, R. C., et al. (2008). Host Genetic Influences on HAART Efficacy and AIDS-Free Survival. *Journal of Acquired Immune Deficiency Syndromes*, 48(3), 263–271.

- Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36(1), 39–56.
- Hibbs, M. A., Dirksen, N. C., Li, K., & Troyanskaya, O. G. (2005). Visualization methods for statistical analysis of microarray clusters. *BMC Bioinformatics*, 6, 115. doi:10.1186/1471-2105-6-115
- Hindorff, L. A., MacArthur, J., Morales, J., Junkins, H. A., Hall, P. N., Klemm, A. K., & Manolio, T. A. (n.d.). A Catalog of Published Genome-Wide Association Studies. Retrieved from <http://www.genome.gov/gwastudies>
- Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., & Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, 32(2), 179–185. doi:10.1002/gepi.20292
- Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2003). Control of confounding of genetic associations in stratified populations. *The American Journal of Human Genetics*, 72(6), 1492–1504. doi:10.1086/375613
- Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2004). Design and analysis of admixture mapping studies. *The American Journal of Human Genetics*, 74(5), 965–978. doi:10.1086/420855
- Holland, P. M., Abramson, R. D., Watson, R., & Gelfand, D. H. (1991). Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proceedings of the National Academy of Sciences*, 88(16), 7276–7280.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 2, 417–441.
- Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6), e1000529. doi:10.1371/journal.pgen.1000529
- International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, 437(7063), 1299–1320. doi:10.1038/nature04226
- International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851–861. doi:10.1038/nature06258
- Jiang, J., Jing, Y., Cost, G. J., Chiang, J.-C., Kolpa, H. J., Cotton, A. M., et al. (2013). Translating dosage compensation to trisomy 21. *Nature*, 500(7462), 296–300. doi:10.1038/nature12394

- Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830), 1497–1502. doi:10.1126/science.1141319
- Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., & O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics*, *11*, 724. doi:10.1186/1471-2164-11-724
- Johnson, R. C., Winkler, C. A., & Nelson, G. W. (n.d.). ALDsuite: Dense marker MALD using principal components of ancestral linkage disequilibrium. *Bioinformatics*.
- Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., et al. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature Genetics*, *42*(3), 210–215. doi:10.1038/ng.531
- Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burt, N. P., Rieder, M. J., et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics*, *40*(2), 189–197. doi:10.1038/ng.75
- Keightley, P. D., & Otto, S. P. (2006). Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, *443*(7107), 89–92. doi:10.1038/nature05049
- Kim-Howard, X., Sun, C., Molineros, J. E., Maiti, A. K., Chandru, H., Adler, A., et al. (2013). Allelic heterogeneity in NCF2 associated with systemic lupus erythematosus (SLE) susceptibility across four ethnic populations. *Human Molecular Genetics*. doi:10.1093/hmg/ddt532
- Klein, R. J., Xu, X., Mukherjee, S., Willis, J., & Hayes, J. (2010). Successes of genome-wide association studies. *Cell*, *142*(3), 350–1– author reply 353–5. doi:10.1016/j.cell.2010.07.026
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389. doi:10.1126/science.1109557
- Knowlton, R. G., Cohen-Haguenauer, O., Van Cong, N., Frézal, J., Brown, V. A., Barker, D., et al. (1985). A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. *Nature*, *318*(6044), 380–382.
- Kopp, Jeffery B, Smith, M. W., Nelson, G. W., Johnson, R. C., Freedman, B. I., Bowden, D. W., et al. (2008). MYH9 is a Major-Effect Risk Gene for Focal Segmental Glomerulosclerosis. *Nature Genetics*, *40*(10), 1175–1184.
- Kopp, Jeffrey B, Nelson, G. W., Sampath, K., Johnson, R. C., Genovese, G., An, P., et al. (2011). APOL1 Genetic Variants in Focal Segmental Glomerulosclerosis and HIV-

- Associated Nephropathy. *Journal of the American Society of Nephrology*, 22(11), 2129–2137. doi:10.1681/ASN.2011040388
- Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *The American Journal of Human Genetics*, 58(6), 1347–1363.
- Kumar, R., Nguyen, E. A., Roth, L. A., Oh, S. S., Gignoux, C. R., Huntsman, S., et al. (2013). Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: the Genes-environments and Admixture in Latino Asthmatics (GALA II) study. *The Journal of Allergy and Clinical Immunology*, 132(4), 896–905.e1. doi:10.1016/j.jaci.2013.02.046
- Lander, E. S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8), 2363–2367.
- Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265(5181), 2037–2048.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. doi:10.1038/35057062
- Lange, K., Weeks, D., & Boehnke, M. (1988). Programs for Pedigree Analysis: MENDEL, FISHER, and dGENE. *Genetic Epidemiology*, 5(6), 471–472. doi:10.1002/gepi.1370050611
- Lathrop, G. M., Lalouel, J. M., Julier, C., & Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *The American Journal of Human Genetics*, 37(3), 482–498.
- Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34(6), 591–602. doi:10.1002/gepi.20516
- Lautenberger, J. A., Stephens, J. C., O'Brien, S. J., & Smith, M. W. (2000). Significant Admixture Linkage Disequilibrium across 30 cM around the FY Locus in African Americans. *The American Journal of Human Genetics*, 66(3), 969–978. doi:10.1086/302820
- Limou, S., Le Clerc, S., Coulonges, C., Carpentier, W., Dina, C., Delaneau, O., et al. (2009). Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *The Journal of Infectious Diseases*, 199(3), 419–426. doi:10.1086/596067
- Ma, S., & Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, 12(6), 714–722. doi:10.1093/bib/bbq090

- MacLean, C. J., & Workman, P. L. (1973). Genetic studies on hybrid populations. I. Individual estimates of ancestry and their relation to quantitative traits. *Annals of Human Genetics*, 36(3), 341–351.
- Manry, J., & Quintana-Murci, L. (2013). A genome-wide perspective of human diversity and its implications in infectious disease. *Cold Spring Harbor Perspectives in Medicine*, 3(1), a012450. doi:10.1101/cshperspect.a012450
- Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*. doi:10.1016/j.ajhg.2013.06.020
- Marban, C., Su, T., Ferrari, R., Li, B., Vatakis, D., Pellegrini, M., et al. (2011). Genome-wide binding map of the HIV-1 Tat protein to the human genome. *PloS One*, 6(11), e26894. doi:10.1371/journal.pone.0026894
- Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7), 499–511. doi:10.1038/nrg2796
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376–380. doi:10.1038/nature03959
- Martin, E. R., Kaplan, N., & Weir, B. (1997). Tests for linkage and association in nuclear families. *The American Journal of Human Genetics*, 61, 439–448.
- Martin, E. R., Lai, E. H., Gilbert, J. R., Rogala, A. R., Afshari, A. J., Riley, J., et al. (2000). SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *The American Journal of Human Genetics*, 67(2), 383–394. doi:10.1086/303003
- Matise, T. C., Chen, F., Chen, W., La Vega, De, F. M., Hansen, M., He, C., et al. (2007). A second-generation combined linkage physical map of the human genome. *Genome Research*, 17(12), 1783–1786. doi:10.1101/gr.7156307
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5), 356–369. doi:10.1038/nrg2344
- McClellan, J., & King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, 141(2), 210–217. doi:10.1016/j.cell.2010.03.032
- McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6), 344–355.

- McDonough, C. W., Palmer, N. D., Hicks, P. J., Roh, B. H., An, S. S., Cooke, J. N., et al. (2011). A genome-wide association study for diabetic nephropathy genes in African Americans. *Kidney International*, 79(5), 563–572. doi:10.1038/ki.2010.467
- McKeigue, P. M. (1997). Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *The American Journal of Human Genetics*, 60(1), 188.
- McKeigue, P. M. (1998). Mapping Genes that Underlie Ethnic Differences in Disease Risk: Methods for Detecting Linkage in Admixed Populations, by Conditioning on Parental Admixture. *The American Journal of Human Genetics*, 63, 241–251.
- McKeigue, P. M., Carpenter, J. R., Parra, E. J., & Shriver, M. D. (2000). Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annals of Human Genetics*, 64(Pt 2), 171–186. doi:doi:10.1017/S0003480000008022
- McLaren, P. J., Coulonges, C., Ripke, S., van den Berg, L., Buchbinder, S., Carrington, M., et al. (2013). Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathogens*, 9(7), e1003515. doi:10.1371/journal.ppat.1003515
- Mezaka, I., Legzdina, L., Waugh, R., Close, T. J., & Rostoks, N. (2012). Genetic Diversity in Latvian Spring Barley Association Mapping Population (pp. 25–35). Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-4682-4_2
- Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182), 66–71.
- Min Jou, W., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237(5350), 82–88.
- Misawa, K., Fujii, S., Yamazaki, T., Takahashi, A., Takasaki, J., Yanagisawa, M., et al. (2008). New correction algorithms for multiple comparisons in case-control multilocus association studies based on haplotypes and diplotype configurations. *Journal of human genetics*, 53(9), 789–801. doi:10.1007/s10038-008-0312-0
- Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., et al. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nature Genetics*, 45(6), 690–696. doi:10.1038/ng.2608
- Montana, G., & Pritchard, J. K. (2004). Statistical tests for admixture mapping with case-control and cases-only data. *The American Journal of Human Genetics*, 75(5), 771–789. doi:10.1086/425281
- Morgan, T. H. (1910). Sex limited inheritance in drosophila. *Science*, 32(812), 120–123.

- Morgan, T. H. (1911). Random segregation versus coupling in medellian inheritance. *Science*, 34(873), 384.
- Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9), 981–990. doi:10.1038/ng.2383
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621–628. doi:10.1038/nmeth.1226
- Morton, N. E. (1955). Sequential tests for the detection of linkage. *The American Journal of Human Genetics*, 7(3), 277–318.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51 Pt 1, 263–273.
- Nalls, M. A., Wilson, J. G., Patterson, N. J., Tandon, A., Zmuda, J. M., Huntsman, S., et al. (2008). Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *The American Journal of Human Genetics*, 81(1), 81–87. doi:10.1016/j.ajhg.2007.09.003
- Nelson, G. W., Freedman, B. I., Bowden, D. W., Langefeld, C. D., An, P., Hicks, P. J., et al. (2010). Dense mapping of MYH9 localizes the strongest kidney disease associations to the region of introns 13 to 15. *Human Molecular Genetics*, 19(9), 1805–1815. doi:10.1093/hmg/ddq039
- Neyman, J., & Pearson, E. S. (1967). The testing of statistical hypotheses in relation to probabilities a priori. *Joint Statistical Papers*.
- Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1), 30–35. doi:10.1038/ng.499
- Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), 272–276. doi:10.1038/nature08250
- Nicodemus, K., Liu, W., Chase, G., Tsai, Y., & Fallin, M. D. (2005). Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. *BMC Genetics*, 6 Suppl 1, S78. doi:10.1186/1471-2156-6-S1-S78
- Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods*. New York, NY: John Wiley & Sons.

- O'Connell, J. R. (2001). Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. *Human Heredity*, 51(4), 226–240.
- Ohta, T. (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences*, 79(6), 1940–1944.
- Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *The American Journal of Human Genetics*, 26(5), 588–597.
- Palaisa, K., Morgante, M., Tingey, S., & Rafalski, A. (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proceedings of the National Academy of Sciences*, 101(26), 9885–9890. doi:10.1073/pnas.0307839101
- Panagiotou, O. A., Ioannidis, J. P. A., Genome-Wide Significance Project. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1), 273–286. doi:10.1093/ije/dyr178
- Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., et al. (1998). Estimating African American admixture proportions by use of population-specific alleles. *The American Journal of Human Genetics*, 63(6), 1839–1851. doi:10.1086/302148
- Pasaniuc, B., Sankararaman, S., Kimmel, G., & Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12), i213–21. doi:10.1093/bioinformatics/btp197
- Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. H. L., et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARE and a Breast Cancer Consortium. *PLoS Genetics*, 7(4), e1001371. doi:10.1371/journal.pgen.1001371
- Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., et al. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547), 1719–1723. doi:10.1126/science.1065573
- Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., et al. (2004). Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5), 979–1000.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics*, 192(3), 1065–1093. doi:10.1534/genetics.112.145037

- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2, 559–572.
- Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research*, 16(9), 1136–1148. doi:10.1101/gr.5402306
- Pólya, G. (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift*, 8(3-4), 171–181. doi:10.1007/BF01206525
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics*, 38(8), 904–909.
- Price, A. L., Patterson, N., Yu, F., Cox, D. R., Waliszewska, A., McDonald, G. J., et al. (2007). A genomewide admixture map for Latino populations. *The American Journal of Human Genetics*, 80(6), 1024–1036. doi:10.1086/518313
- Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., et al. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6), e1000519. doi:10.1371/journal.pgen.1000519
- Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., et al. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *The American Journal of Human Genetics*, 83(1), 132–135. doi:10.1016/j.ajhg.2008.06.005
- Price, P., James, I., Fernandez, S., & Frencha, M. A. (2004). Alleles of the gene encoding IL-1 α may predict control of plasma viraemia in HIV-1 patients on highly active antiretroviral therapy. *AIDS*, 18(11), 1495–1501.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: A Toolset for Whole-Genome Association and Population-Based Linkage Analysis. *The American Journal of Human Genetics*, 81(3), 559–575.
- Qian, F., Chung, L., Zheng, W., Bruno, V., Alexander, R. P., Wang, Z., et al. (2013). Identification of genes critical for resistance to infection by West Nile virus using RNA-Seq analysis. *Viruses*, 5(7), 1664–1681. doi:10.3390/v5071664
- R Development Core Team. (n.d.). R: A language and environment for statistical computing. *Manual*. Vienna, Austria.

- Ramsay, J. O., & Silverman, B. W. (2005). Principal components analysis for functional data. In P. Bickel, P. J. Diggle, S. Fienberg, U. Gather, I. Olkin, & S. L. Zeger (Eds.), *Functional Data Analysis* (2nd ed., pp. 147–172). New York: Springer.
- Redden, D. T., Divers, J., Vaughan, L. K., Tiwari, H. K., Beasley, T. M., Fernández, J. R., et al. (2006). Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. *PLoS Genetics*, 2(8), e137. doi:10.1371/journal.pgen.0020137
- Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834), 199–204. doi:10.1038/35075590
- Reich, D., Patterson, N., De Jager, P. L., McDonald, G. J., Waliszewska, A., Tandon, A., et al. (2005). A Whole-Genome Admixture Scan Finds a Candidate Locus for Multiple Sclerosis Susceptibility. *Nature Genetics*, 37(10), 1113–1118.
- Riordan, J. R., Rommens, J. M., Kerem, B.-S., Alon, N., Rozmahel, R., Grzelczak, Z., et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922), 1066–1073.
- Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281), 1516–1517.
- Risch, N., Burchard, E., Ziv, E., & Tang, H. (2002). Categorization of humans in biomedical research: genes, race and disease. *Genome Biology*, 3(7), comment2007.
- Roberts, R. J. (2005). How restriction enzymes became the workhorses of molecular biology. *Proceedings of the National Academy of Sciences*, 102(17), 5905–5908. doi:10.1073/pnas.0500923102
- Rogers, S. M. (2012). Mapping the genomic architecture of ecological speciation in the wild: does linkage disequilibrium hold the key? *Molecular Ecology*, 21(21), 5155–5158. doi:10.1111/mec.12019
- Rommens, J. M., Iannuzzi, M. C., Kerem, B.-S., Drumm, M. L., Melmer, G., Dean, M., et al. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245(4922), 1059–1065.
- Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732), 1350–1354.
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3), 441–448.

- Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596), 687–695.
- Sankararaman, S., Sridhar, S., & Kimmel, G. (2008). Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics*.
- Sapp, J. (1997). Jean Brachet, l'hérédité générale et les origines de l'embryologie moléculaire. *History and Philosophy of the Life Sciences*, 19(1), 69–87.
- Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4), 629–644. doi:10.1086/502802
- Schnekenburger, M., & Diederich, M. (2012). Epigenetics Offer New Horizons for Colorectal Cancer Prevention. *Current Colorectal Cancer Reports*, 8(1), 66–81. doi:10.1007/s11888-011-0116-z
- Schork, N. J., Greenwood, T. A., & Braff, D. L. (2007). Statistical genetics concepts and approaches in schizophrenia and related neuropsychiatric research. *Schizophrenia Bulletin*, 33(1), 95–104. doi:10.1093/schbul/sbl045
- Seffah, A., & Metzker, E. (2004). The obstacles and myths of usability and software engineering. *Communications of the ACM*, 47(12), 71–76. doi:10.1145/1035134.1035136
- Seldin, M. F., Pasaniuc, B., & Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 12(8), 523–528. doi:10.1038/nrg3002
- Sharov, A. A., Dudekula, D. B., & Ko, M. S. H. (2005). A web-based tool for principal component and significance analysis of microarray data. *Bioinformatics*, 21(10), 2548–2549. doi:10.1093/bioinformatics/bti343
- Sibbald, B. (2001). Death but one unintended consequence of gene-therapy trial. *Canadian Medical Association Journal - Journal de l'Association Médicale Canadienne*, 164(11), 1612.
- Silberstein, M., Weissbrod, O., Otten, L., Tzemach, A., Anisenia, A., Shtark, O., et al. (2013). A system for exact and approximate genetic linkage analysis of SNP data in large pedigrees. *Bioinformatics*, 29(2), 197–205. doi:10.1093/bioinformatics/bts658
- Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics*, 38(2), 209–213. doi:10.1038/ng1706
- Smith, M. W., Patterson, N., Lautenberger, J. A., Truelove, A. L., McDonald, G. J., Waliszewska, A., et al. (2004). A high-density admixture map for disease gene

- discovery in African Americans. *The American Journal of Human Genetics*, 74(5), 1001–1013.
- Sober, E. (1993). *The nature of selection: evolutionary theory in philosophical focus*. Chicago: University of Chicago Press.
- Soubry, A., Murphy, S. K., Wang, F., Huang, Z., Vidal, A. C., Fuemmeler, B. F., et al. (2013). Newborns of obese parents have altered DNA methylation patterns at imprinted genes. *International Journal of Obesity*. doi:10.1038/ijo.2013.193
- Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *The American Journal of Human Genetics*, 52(3), 506–516.
- Stephens, J. C., Briscoe, D., & O'Brien, S. J. (1994). Mapping by Admixture Linkage Disequilibrium in Human Populations: Limits and Guidelines. *The American Journal of Human Genetics*, 55, 809–824.
- Sterne, J. A., & Davey Smith, G. (2001). Sifting the evidence-what's wrong with significance tests? *BMJ*, 322(7280), 226–231.
- Stone, N. E., Fan, J. B., Willour, V., Pennacchio, L. A., Warrington, J. A., Hu, A., et al. (1996). Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene. *Genome Research*, 6(3), 218–225.
- Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.
- Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14, 43–59.
- Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4), 676–682. doi: 10.1101/gr.072850.107
- Sutton, W. S. (1903). The chromosomes in heredity. *Biological Bulletin*, 4, 231–251.
- Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G., & Risch, N. J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. *The American Journal of Human Genetics*, 81(3), 626–633. doi:10.1086/520769
- Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1), 1–12. doi:10.1086/504302
- Tchebichef, P. (1846). Démonstration élémentaire d'une proposition générale de la théorie des probabilités. *Journal für die reine und angewandte Mathematik*, (33), 259–267.

- Thoday, J. M. (1969). Limitations to genetic comparison of populations. *Journal of biosocial science, Suppl 1*, 3–14.
- Tian, C., Hinds, D. A., Shigeta, R., Adler, S. G., Lee, A., Pahl, M. V., et al. (2007). A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *The American Journal of Human Genetics*, 80(6), 1014–1023.
- Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., et al. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271(5254), 1380–1387.
- Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., et al. (2009). The genetic structure and history of Africans and African Americans. *Science*, 324(5930), 1035–1044. doi:10.1126/science.1172257
- Toole, J. J., Knopf, J. L., Wozney, J. M., Sultzman, L. A., Buecker, J. L., Pittman, D. D., et al. (1984). Molecular cloning of a cDNA encoding human antihemophilic factor. *Nature*, 312(5992), 342–347.
- Torgerson, D. G., Gignoux, C. R., Galanter, J. M., Drake, K. A., Roth, L. A., Eng, C., et al. (2012). Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *The Journal of Allergy and Clinical Immunology*, 130(1), 76–82.e12. doi:10.1016/j.jaci.2012.02.040
- Troyer, J. L., Nelson, G. W., Lautenberger, J. A., Chinn, L., McIntosh, C., Johnson, R. C., et al. (2011). Genome-wide association study implicates PARD3B-based AIDS restriction. *The Journal of Infectious Diseases*, 203(10), 1491–1502. doi:10.1093/infdis/jir046
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science*, 291(5507), 1304–1351. doi:10.1126/science.1058040
- Venter, J. C., Smith, H. O., & Hood, L. (1996). A new strategy for genome sequencing. *Nature*, 381(6581), 364–366. doi:10.1038/381364a0
- Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1), 7–24. doi:10.1016/j.ajhg.2011.11.029
- Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, 55(4), 641–658. doi:10.1373/clinchem.2008.112789
- Wacholder, S., Chanock, S. J., Garcia-Closas, M., Ghormli, El, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6), 434–442.

- Wainwright, B. J., Scambler, P. J., Schmidtke, J., Watson, E. A., Law, H. Y., Farrall, M., et al. (1985). Localization of cystic fibrosis locus to human chromosome 7cen-q22. *Nature*, 318(6044), 384–385.
- Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *The American Journal of Human Genetics*, 81(2), 208–227. doi:10.1086/519024
- Wakefield, J. (2008). Reporting and interpretation in genome-wide association studies. *International Journal of Epidemiology*, 37(3), 641–653. doi:10.1093/ije/dym257
- Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8), 587–597. doi:10.1038/nrg1123
- Walley, N. M., Julg, B., Dickson, S. P., Fellay, J., Ge, D., Walker, B. D., et al. (2009). The Duffy antigen receptor for chemokines null promoter variant does not influence HIV-1 acquisition or disease progression. *Cell Host and Microbe*, 5(5), 408–10– author reply 418–9. doi:10.1016/j.chom.2009.04.011
- Walsh, T., & King, M.-C. (2007, February). Ten genes for inherited breast cancer. *Cancer Cell*, pp. 103–105. doi:10.1016/j.ccr.2007.01.010
- Wang, K., Baldassano, R., Zhang, H., Qu, H.-Q., Imielinski, M., Kugathasan, S., et al. (2010). Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Human Molecular Genetics*, 19(10), 2059–2067. doi:10.1093/hmg/ddq078
- Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., & Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *The American Journal of Human Genetics*, 71(5), 1227–1234. doi:10.1086/344398
- Watson, J., & Crick, F. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–738.
- Weber, J. L., & Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Research*, 7(5), 401–409.
- Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., et al. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, 40(5), 575–583. doi:10.1038/ng.121
- Weisenberger, D. J., VanDenBerg, D., Pan, F., Berman, B., & Laird, P. W. (2008). *Comprehensive DNA methylation analysis on the Illumina Infinium assay platform*. Illumina Application Note.

- Weiss, R. (2005, March 4). Boy's cancer prompts FDA to halt gene therapy. *The Washington Post*, p. A02. Retrieved from <http://www.washingtonpost.com/wp-dyn/articles/A3926-2005Mar3.html>
- Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. doi:10.1038/nature05911
- White, R., Woodward, S., Leppert, M., O'Connell, P., Hoff, M., Herbst, J., et al. (1985). A closely linked genetic marker for cystic fibrosis. *Nature*, 318(6044), 382–384.
- Winkler, C. A., An, P., Johnson, R. C., Nelson, G. W., & Kirk, G. (2009). Expression of Duffy Antigen Receptor for Chemokines (DARC) Has No Effect on HIV-1 Acquisition or Progression to AIDS in African Americans. *Cell Host and Microbe*, 5(5), 411–413.
- Winkler, C. A., Nelson, G. W., & Smith, M. W. (2010). Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics*, 11, 65–89. doi:10.1146/annurev-genom-082509-141523
- Xu, M., Fujita, D., & Hanagata, N. (2009). Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small*, 5(23), 2638–2649. doi:10.1002/smll.200900976
- Yang, James J, Li, J., Buu, A., & Williams, L. K. (2013). Efficient Inference of Local Ancestry. *Bioinformatics*. doi:10.1093/bioinformatics/btt488
- Yang, Jun J, Cheng, C., Devidas, M., Cao, X., Fan, Y., Campana, D., et al. (2011). Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nature Genetics*, 43(3), 237–241. doi:10.1038/ng.763
- Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9), 763–774.
- Zaykin, D. V., Kuo, C. L., & Vsevolozhskaya, O. A. (2013, October 25). The more you test, the more you find: massive multiple testing does not promote spurious findings among top hits of association studies. *American Society of Human Genetics*. Boston, MA. Retrieved from <http://abstracts.ashg.org/cgi-bin/2013/ashg13s.pl?pgmnr=1789&sort=ptimes&sbutton=Detail&absno=130121288&sid=850896>
- Zhang, Kui, Deng, M., Chen, T., Waterman, M. S., & Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences*, 99(11), 7335–7339. doi:10.1073/pnas.102186799
- Zhang, Xiaojing, Davenport, K. W., Gu, W., Daligault, H. E., Munk, A. C., Tashima, H., et al. (2012). Improving genome assemblies by sequencing PCR products with PacBio. *BioTechniques*, 53(1), 61–62. doi:10.2144/0000113891

Zhao, H., Pfeiffer, R., & Gail, M. H. (2003). Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, 4(2), 171–178. doi:10.1517/phgs.4.2.171.22636

Appendices

A. Review of admixture publications from 2013

Type	Software Used	Map	Citation
Fine Mapping			(Foster et al. 2013)
Fine Mapping	HAPMIX		(Kantor et al. 2013)
Fine Mapping			(Ng et al. 2013)
GWAS	STRUCTURE	Sparse	(Chen et al. 2013)
GWAS	LAMP-LD	Dense	(Chimusa et al. 2013)
GWAS	LAMP-LD	Dense	(Drake et al. 2013)
GWAS	HAPMIX	Dense	(Fejerman et al. 2013)
GWAS	ADMIXTURE, EIGENSTRAT, HAPMIX	Dense	(Gamazon et al. 2013)
GWAS	LAMP-LD, MULTIMIX	Dense	(Lalli et al. 2013)
GWAS	HAPMIX	Dense	(J. Li et al. 2013)
GWAS	HAPMIX, EIGENSOFT	Dense	(Lopes et al. 2013)
GWAS	STRUCTURE, EIGENSOFT	Dense	(Maranville et al. 2013)
GWAS	HAPMIX	Dense	(Monda et al. 2013)
GWAS		Dense	(Payne 2013)
GWAS	HAPMIX	Dense	(Song et al. 2013)
GWAS	HAPMIX	Dense	(Wheeler et al. 2013)
GWAS	HAPMIX	Dense	(Witte et al. 2013)
GWAS	HAPMIX	Dense	(Xie et al. 2013)
MALD	ADMIXMAP	Sparse	(Bensen et al. 2013)
MALD	ADMIXMAP	Sparse	(Divers et al. 2013)
MALD	ANCESTRYMAP	Sparse	(Kim-Howard et al. 2013)
MALD	STRUCTURE, ANCESTRYMAP, ADMIXMAP	Sparse	(Molineros et al. 2013)
MALD	SABER+	Dense	(Beleza et al. 2013)
MALD	SABER+	Dense	(Coram et al. 2013)
MALD	HAPMIX, LAMP, SABER	Dense	(Ochs-Balcom et al. 2013)
Methods	ADMIXMAP	Sparse	(McKeigue et al. 2013)
Methods	STRUCTURE	Sparse	(Porras-Hurtado et al. 2013)
Methods	LOCO-LD	Dense	(Baran et al. 2013)
Methods	invertFREGENE	Dense	(Seich Al Basatena et al. 2013)
Methods	MULTIMIX	Dense	(Churchhouse & Marchini 2013)

Type	Software Used	Map	Citation
Methods		Dense	(Crawford & Nielsen 2013)
Methods	LFMM	Dense	(Frichot et al. 2013)
Methods	LAMP-LD, HAPMIX, MULTIMIX, RFMIX	Dense	(Genovese, Handsaker, H. Li, Kenny, et al. 2013b)
Methods	LAMP-LD, HAPMIX, MULTIMIX, RFMIX	Dense	(Genovese, Handsaker, H. Li, Altemose, et al. 2013a)
Methods	Hap-seq	Dense	(He et al. 2013)
Methods	SEQMIX	Dense	(Hu et al. 2013)
Methods		Dense	(Jin et al. 2013)
Methods	MaCH-Admix	Dense	(E. Y. Liu et al. 2013a)
Methods		Dense	(Y. Liu 2013)
Methods	ALDER	Dense	(Loh et al. 2013)
Methods	HAPMIX	Dense	(Mao et al. 2013)
Methods	RFMix	Dense	(Maples et al. 2013)
Methods	MaCH-Admix	Dense	(Martin et al. 2014)
Methods		Dense	(Morrison 2013)
Methods	LAMP-LD, ALLOY, WINPOP, PCAdmix	Dense	(Pasaniuc et al. 2013)
Methods		Dense	(Paul 2013)
Methods	ARGweaver	Dense	(Rasmussen & Siepel 2013)
Methods	ALLOY	Dense	(Rodriguez et al. 2013)
Methods	ELIA	Dense	(Yang et al. 2013)
Methods	DBM-Admix	Dense	(Zhang 2013)
Other	PCA	Sparse	(Isobe et al. 2013)
Population History		Dense	(Alcala et al. 2013)
Population History		Dense	(Gourjon 2013)
Population History		Dense	(Gravel et al. 2013)
Population History		Dense	(Hudjašov 2013)
Population History	HAPMIX	Dense	(Hufford et al. 2013)
Population History	STRUCTURE, EIGENSOFT	Dense	(McTavish et al. 2013)
Population History	ADMIXTURE, HAPMIX	Dense	(Moreau et al. 2013)
Population History		Dense	(Skoglund 2013)
Review			(Barbujani et al. 2013)
Review			(Gompert & Buerkle 2013)
Review			(Haralambieva et al. 2013)
Review			(Y. Liu et al. 2013b)

Type	Software Used	Map	Citation
Review			(Shriner 2013)
Review			(Thompson 2013)

Table 9. Review of publications using admixture inference in 2013.

B. Currently available admixture software

Software	Dense Markers	MALD	>2 Pops	Cited	References
STRUCTURE			✓	12427	Pritchard <i>et al.</i> (2000); Falush <i>et al.</i> (2003) McKeigue <i>et al.</i> (2000);
ADMIXMAP		✓		201	McKeigue <i>et al.</i> (2013)
ANCESTRYMAP		✓		361	Patterson <i>et al.</i> (2004)
FRAPPE	✓		✓	255	Tang <i>et al.</i> (2005)
SABER+	✓		✓	157	Tang <i>et al.</i> (2006)
LAMP-LD	✓		✓	131	Sankararaman <i>et al.</i> (2008a); Baran <i>et al.</i> (2012)
HAPAA	✓		✓	48	Sundquist <i>et al.</i> (2008)
SWITCH-MHMM	✓		✓	35	Sankararaman <i>et al.</i> (2008b)
WINPOP	✓		✓	54	Pasaniuc <i>et al.</i> (2009)
HAPMIX	✓			210	Price <i>et al.</i> (2009)
ADMIXTURE	✓		✓	293	Alexander <i>et al.</i> (2009)
PCAdmix	✓		✓	14	Brisbin <i>et al.</i> (2012)
MULTIMIX	✓		✓	9	Churchhouse and Marchini (2013)
SEQMIX	✓				Hu <i>et al.</i> (2013)
ALDER	✓		✓	20	Loh <i>et al.</i> (2013)
RFMix	✓		✓	7	Maples <i>et al.</i> (2013)
ALLOY	✓		✓	1	Rodriguez <i>et al.</i> (2013)
EILA	✓		✓	2	Yang <i>et al.</i> (2013)
DBM-Admix	✓		✓		Zhang (2013)
MaCH-Admix	✓		✓	14	Liu <i>et al.</i> (2013)
ELAI	✓		✓	2	Guan (2014)

Table 10. Currently available admixture inference software. Software title, ability to analyze dense marker data, support for mapping by admixture linkage disequilibrium, number of admixing populations supported, number of citations and references are included.

Bibliography

- Alcala, N., Streit, D., Goudet, J., & Vuilleumier, S. (2013). Peak and persistent excess of genetic diversity following an abrupt migration increase. *Genetics*, 193(3), 953–971. doi:10.1534/genetics.112.147785
- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 1655–1664. doi:10.1101/gr.094052.109
- Baran, Y., Quintela, I., Carracedo, A., Pasaniuc, B., & Halperin, E. (2013). Enhanced Localization of Genetic Samples through Linkage-Disequilibrium Correction. *The American Journal of Human Genetics*, 92(6), 882–894. doi:10.1016/j.ajhg.2013.04.023
- Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10), 1359–1367. doi:10.1093/bioinformatics/bts144
- Barbujani, G., Ghirotto, S., & Tassi, F. (2013). Nine things to remember about human genome diversity. *Tissue Antigens*, 82(3), 155–164. doi:10.1111/tan.12165
- Beleza, S., Johnson, N. A., Candille, S. I., Absher, D. M., Coram, M. A., Lopes, J., et al. (2013). Genetic architecture of skin and eye color in an African-European admixed population. *PLoS Genetics*, 9(3), e1003372. doi:10.1371/journal.pgen.1003372
- Bensen, J. T., Xu, Z., McKeigue, P. M., Smith, G. J., Fonthan, E. T. H., Mohler, J. L., & Taylor, J. A. (2013). Admixture mapping of prostate cancer in African Americans participating in the North Carolina-Louisiana Prostate Cancer Project (PCaP). *The Prostate*. doi:10.1002/pros.22722
- Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., et al. (2012). PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human Biology*, 84(4), 343–364. doi:10.3378/027.084.0401
- Chen, F., Chen, G. K., Stram, D. O., Millikan, R. C., Ambrosone, C. B., John, E. M., et al. (2013). A genome-wide association study of breast cancer in women of African ancestry. *Human Genetics*, 132(1), 39–48. doi:10.1007/s00439-012-1214-y
- Chimusa, E. R., Zaitlen, N., Daya, M., Möller, M., van Helden, P. D., Mulder, N. J., et al. (2013). Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Human Molecular Genetics*. doi:10.1093/hmg/ddt462

- Churchhouse, C., & Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology*, 37(1), 1–12. doi:10.1002/gepi.21692
- Coram, M. A., Duan, Q., Hoffmann, T. J., Thornton, T., Knowles, J. W., Johnson, N. A., et al. (2013). Genome-wide Characterization of Shared and Distinct Genetic Components that Influence Blood Lipid Levels in Ethnically Diverse Human Populations. *The American Journal of Human Genetics*, 92(6), 904–916. doi:10.1016/j.ajhg.2013.04.025
- Crawford, J. E., & Nielsen, R. (2013). Detecting adaptive trait loci in nonmodel systems: divergence or admixture mapping? *Molecular Ecology*, 22(24), 6131–6148. doi:10.1111/mec.12562
- Divers, J., Palmer, N. D., Lu, L., Register, T. C., Carr, J. J., Hicks, P. J., et al. (2013). Admixture mapping of coronary artery calcified plaque in African Americans with type 2 diabetes mellitus. *Cardiovascular Genetics*, 6(1), 97–105. doi:10.1161/CIRCGENETICS.112.964114
- Drake, K. A., Torgerson, D. G., Gignoux, C. R., Galanter, J. M., Roth, L. A., Huntsman, S., et al. (2013). A genome-wide association study of bronchodilator response in Latinos implicates rare variants. *The Journal of Allergy and Clinical Immunology*. doi:10.1016/j.jaci.2013.06.043
- Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies., 164(4), 1567–1587.
- Fejerman, L., Hu, D., Huntsman, S., John, E. M., & Stern, M. C. (2013). Genetic ancestry and risk of mortality among US Latinas with breast cancer. *Cancer Research*.
- Foster, M. C., Coresh, J., Fornage, M., Astor, B. C., Grams, M., Franceschini, N., et al. (2013). APOL1 variants associate with increased risk of CKD among African Americans. *Journal of the American Society of Nephrology*, 24(9), 1484–1491. doi:10.1681/ASN.2013010113
- Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, 30(7), 1687–1699. doi:10.1093/molbev/mst063
- Gamazon, E. R., Pinto, N., Konkashbaev, A., Im, H. K., Diskin, S. J., London, W. B., et al. (2013). Trans-population analysis of genetic mechanisms of ethnic disparities in neuroblastoma survival. *Journal of the National Cancer Institute*, 105(4), 302–309. doi:10.1093/jnci/djs503

- Genovese, G., Handsaker, R. E., Li, H., Altemose, N., Lindgren, A. M., Chambert, K., et al. (2013a). Using population admixture to help complete maps of the human genome. *Nature Genetics*, 45(4), 406–14– 414e1–2. doi:10.1038/ng.2565
- Genovese, G., Handsaker, R. E., Li, H., Kenny, E. E., & McCarroll, S. A. (2013b). Mapping the Human Reference Genome's Missing Sequence by Three-Way Admixture in Latino Genomes. *The American Journal of Human Genetics*. doi:10.1016/j.ajhg.2013.07.002
- Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics*, 196(3), 625–642. doi:10.1534/genetics.113.160697
- Gompert, Z., & Buerkle, C. A. (2013). Analyses of genetic ancestry enable key insights for molecular ecology. *Molecular Ecology*, 22(21), 5278–5294. doi:10.1111/mec.12488
- Gourjon, G. (2013). *Anthropological issues in genetic admixture* (Vol. 25, pp. 65–82). Paris: Bulletins et mémoires de la Société d'anthropologie de Paris. doi:10.1007/s13219-012-0071-5
- Gravel, S., Zakharia, F., Byrnes, J. K., & Muzzio, M. (2013). Reconstructing Native American Migrations from Whole-genome and Whole-exome Data. *arXiv.org*.
- Haralambieva, I. H., Ovsyannikova, I. G., Pankratz, V. S., Kennedy, R. B., Jacobson, R. M., & Poland, G. A. (2013). The genetic basis for interindividual immune response variation to measles vaccine: new understanding and new vaccine approaches. *Expert Review of Vaccines*, 12(1), 57–70. doi:10.1586/erv.12.134
- He, D., Han, B., & Eskin, E. (2013). Hap-seq: an optimal algorithm for haplotype phasing with imputation using sequencing data. *Journal of Computational Biology*, 20(2), 80–92. doi:10.1089/cmb.2012.0091
- Hu, Y., Willer, C., Zhan, X., Kang, H. M., & Abecasis, G. R. (2013). Accurate Local-Ancestry Inference in Exome-Sequenced Admixed Individuals via Off-Target Sequence Reads. *The American Journal of Human Genetics*, 93(5), 891–899. doi:10.1016/j.ajhg.2013.10.008
- Hudjašov, G. (2013). *Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations*. University of Tartu.
- Hufford, M. B., Lubinsky, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, 9(5), e1003477. doi:10.1371/journal.pgen.1003477
- Isobe, N., Gourraud, P.-A., Harbo, H. F., Caillier, S. J., Santaniello, A., Khankhanian, P., et al. (2013). Genetic risk variants in African Americans with multiple sclerosis. *Neurology*, 81(3), 219–227. doi:10.1212/WNL.0b013e31829bfe2f

- Jin, W., Li, R., Zhou, Y., & Xu, S. (2013). Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *European Journal of Human Genetics*. doi:10.1038/ejhg.2013.265
- Kantor, D. B., Palmer, C. D., Young, T. R., Meng, Y., Gajdos, Z. K., Lyon, H., et al. (2013). Replication and fine mapping of asthma-associated loci in individuals of African ancestry. *Human Genetics*, 132(9), 1039–1047. doi:10.1007/s00439-013-1310-7
- Kim-Howard, X., Sun, C., Molineros, J. E., Maiti, A. K., Chandru, H., Adler, A., et al. (2013). Allelic heterogeneity in NCF2 associated with systemic lupus erythematosus (SLE) susceptibility across four ethnic populations. *Human Molecular Genetics*. doi:10.1093/hmg/ddt532
- Lalli, M. A., Cox, H. C., Arcila, M. L., Cadavid, L., Moreno, S., Garcia, G., et al. (2013). Origin of the PSEN1 E280A mutation causing early-onset Alzheimer's disease. *Alzheimer's & Dementia*. doi:10.1016/j.jalz.2013.09.005
- Li, Jin, Glessner, J. T., Zhang, H., Hou, C., Wei, Z., Bradfield, J. P., et al. (2013). GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. *Human Molecular Genetics*, 22(7), 1457–1464. doi:10.1093/hmg/ddt534
- Liu, E. Y., Li, M., Wang, W., & Li, Y. (2013a). MaCH-admix: genotype imputation for admixed populations. *Genetic Epidemiology*, 37(1), 25–37. doi:10.1002/gepi.21690
- Liu, Y. (2013). *Fast and Accurate Haplotype Inference with Hidden Markov Model*. University of North Carolina.
- Liu, Y., Nyunoya, T., Leng, S., Belinsky, S. A., Tesfaigzi, Y., & Bruse, S. (2013b). Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics*, 7, 1. doi:10.1186/1479-7364-7-1
- Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4), 1233–1254. doi:10.1534/genetics.112.147330
- Lopes, A. M., Aston, K. I., Thompson, E., Carvalho, F., Gonçalves, J., Huang, N., et al. (2013). Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS Genetics*, 9(3), e1003349. doi:10.1371/journal.pgen.1003349
- Mao, X., Li, Y., Liu, Y., Lange, L., & Li, M. (2013). Testing genetic association with rare variants in admixed populations. *Genetic Epidemiology*, 37(1), 38–47. doi:10.1002/gepi.21687

- Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*. doi:10.1016/j.ajhg.2013.06.020
- Maranville, J. C., Baxter, S. S., Witonsky, D. B., Chase, M. A., & Di Rienzo, A. (2013). Genetic mapping with multiple levels of phenotypic information reveals determinants of lymphocyte glucocorticoid sensitivity. *The American Journal of Human Genetics*, 93(4), 735–743. doi:10.1016/j.ajhg.2013.08.005
- Martin, A. R., Tse, G., Bustamante, C. D., & Kenny, E. E. (2014). Imputation-based assessment of next generation rare exome variant arrays. *Pacific Symposium on Biocomputing*, 19, 241–252.
- McKeigue, P. M., Colombo, M., Agakov, F., Datta, I., Levin, A., Favro, D., et al. (2013). Extending admixture mapping to nuclear pedigrees: application to sarcoidosis. *Genetic Epidemiology*, 37(3), 256–266. doi:10.1002/gepi.21710
- McKeigue, P. M., Carpenter, J. R., Parra, E. J., & Shriver, M. D. (2000). Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annals of Human Genetics*, 64(Pt 2), 171–186. doi:10.1017/S0003480000008022
- McTavish, E. J., Decker, J. E., Schnabel, R. D., Taylor, J. F., & Hillis, D. M. (2013). New World cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences*, 110(15), E1398–406. doi:10.1073/pnas.1303367110
- Molineros, J. E., Maiti, A. K., Sun, C., Looger, L. L., Han, S., Kim-Howard, X., et al. (2013). Admixture mapping in lupus identifies multiple functional variants within IFIH1 associated with apoptosis, inflammation, and autoantibody production. *PLoS Genetics*, 9(2), e1003222. doi:10.1371/journal.pgen.1003222
- Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., et al. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nature Genetics*, 45(6), 690–696. doi:10.1038/ng.2608
- Moreau, C., Lefebvre, J.-F., Jomphe, M., Bhérer, C., Ruiz-Linares, A., Vézina, H., et al. (2013). Native American admixture in the Quebec founder population. *PloS One*, 8(6), e65507. doi:10.1371/journal.pone.0065507
- Morrison, J. (2013). Characterization and correction of error in genome-wide IBD estimation for samples with population structure. *Genetic Epidemiology*, 37(6), 635–641. doi:10.1002/gepi.21737

- Ng, M. C. Y., Saxena, R., Li, J., Palmer, N. D., Dimitrov, L., Xu, J., et al. (2013). Transferability and fine mapping of type 2 diabetes loci in African Americans: the Candidate Gene Association Resource Plus Study. *Diabetes*, 62(3), 965–976. doi: 10.2337/db12-0266
- Ochs-Balcom, H. M., Preus, L., Wactawski-Wende, J., Nie, J., Johnson, N. A., Zakharia, F., et al. (2013). Association of DXA-derived bone mineral density and fat mass with African ancestry. *The Journal of Clinical Endocrinology and Metabolism*, 98(4), E713–7. doi:10.1210/jc.2012-3921
- Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., et al. (2004). Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5), 979–1000.
- Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Zaitlen, N., Eng, C., et al. (2013). Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, 29(11), 1407–1415. doi: 10.1093/bioinformatics/btt166
- Pasaniuc, B., Sankararaman, S., Kimmel, G., & Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12), i213–21. doi: 10.1093/bioinformatics/btp197
- Paul, J. (2013). *Conditional Sampling Distributions for Coalescent Models Incorporating Recombination*. University of California at Berkeley.
- Payne, E. K. (2013). *Genetic studies of salivary cortisol profiles and their influence on chronic disease risk factors*. University of Michigan.
- Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., & Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in Genetics*, 4, 98. doi:10.3389/fgene.2013.00098
- Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., et al. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6), e1000519. doi:10.1371/journal.pgen.1000519
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Rasmussen, M. D., & Siepel, A. (2013). Genome-wide inference of ancestral recombination graphs. *arXiv.org*.
- Rodriguez, J. M., Bercovici, S., Elmore, M., & Batzoglou, S. (2013). Ancestry inference in complex admixtures via variable-length Markov chain linkage models. *Journal of Computational Biology*, 20(3), 199–211. doi:10.1089/cmb.2012.0088

- Sankararaman, S., Kimmel, G., Halperin, E., & Jordan, M. I. (2008a). On the inference of ancestries in admixed populations. *Genome Research*, 18(4), 668–675. doi:10.1101/gr.072751.107
- Sankararaman, S., Sridhar, S., & Kimmel, G. (2008b). Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics*, 82(2), 290–303.
- Seich A1 Basatena, N.-K., Hoggart, C. J., Coin, L. J., & O'Reilly, P. F. (2013). The effect of genomic inversions on estimation of population genetic parameters from SNP data. *Genetics*, 193(1), 243–253. doi:10.1534/genetics.112.145599
- Shriner, D. (2013). Overview of admixture mapping. *Current Protocols in Human Genetics*, Chapter 1, Unit 1.23. doi:10.1002/0471142905.hg0123s76
- Skoglund, P. (2013). *Reconstructing the Human Past using Ancient and Modern Genomes*. Uppsala University.
- Song, C., Chen, G. K., Millikan, R. C., Ambrosone, C. B., John, E. M., Bernstein, L., et al. (2013). A genome-wide scan for breast cancer risk haplotypes among African American women. *PloS One*, 8(2), e57298. doi:10.1371/journal.pone.0057298
- Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4), 676–682. doi:10.1101/gr.072850.107
- Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1), 1–12. doi:10.1086/504302
- Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, 28(4), 289–301. doi:10.1002/gepi.20064
- Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2), 301–326. doi:10.1534/genetics.112.148825
- Wheeler, H. E., Gamazon, E. R., Wing, C., Njiaju, U. O., Njoku, C., Baldwin, R. M., et al. (2013). Integration of cell line and clinical trial genome-wide analyses supports a polygenic architecture of Paclitaxel-induced sensory peripheral neuropathy. *Clinical Cancer Research*, 19(2), 491–499. doi:10.1158/1078-0432.CCR-12-2618
- Witte, J. S., Mefford, J., Plummer, S. J., Liu, J., Cheng, I., Klein, E. A., et al. (2013). HOXB13 mutation and prostate cancer: studies of siblings and aggressive disease. *Cancer Epidemiology, Biomarkers and Prevention*, 22(4), 675–680. doi:10.1158/1055-9965.EPI-12-1154

- Xie, P., Kranzler, H. R., Yang, C., Zhao, H., Farrer, L. A., & Gelernter, J. (2013). Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. *Biological Psychiatry*, 74(9), 656–663. doi:10.1016/j.biopsych.2013.04.013
- Yang, J. J., Li, J., Buu, A., & Williams, L. K. (2013). Efficient Inference of Local Ancestry. *Bioinformatics*. doi:10.1093/bioinformatics/btt488
- Zhang, Y. (2013). De novo inference of stratification and local admixture in sequencing studies. *BMC Bioinformatics*, 14 Suppl 5, S17. doi:10.1186/1471-2105-14-S5-S17

Randall Johnson

**Modeling of linkage disequilibrium in
whole genome genetic association studies**

Résumé

GWAS est un outil essentiel pour la découverte de gènes de la maladie, mais il a de graves problèmes de puissance statistique quand il est impossible d'échantillonner génétiquement des dizaines de milliers de sujets. Les résultats présentés ici—ALDsuite, un programme en utilisant une correction nouvelle et efficace pour le déséquilibre de liaison (DL) ancestrale de la population locale, en permettant l'utilisation de marqueurs denses dans le MALD, et la démonstration que la méthode simpleM fournit une correction optimale pour les comparaisons multiples dans le GWAS—réaffirment la valeur de l'analyse en composantes principales (APC) pour la capture essentiel de la complexité des systèmes de grande dimension. L'APC est déjà la norme pour corriger la structure de la population dans le GWAS; mes résultats indiquent qu'elle est une stratégie générale pour faire face à la forte dimensionnalité des données génomiques d'association.

mots-clés : GWAS, association génétique, génome-entier, statistiques, correction, Analyse par composantes principales

Résumé en anglais

GWAS is an essential tool for disease gene discovery, but has severe problems of statistical power when it is impractical to genetically sample tens of thousands of subjects. The results presented here—a novel, effective correction for local ancestral population LD allowing use of dense markers in MALD using the ALDsuite and the demonstration that the simpleM method provides an optimum Bonferroni correction for multiple comparisons in GWAS, reiterate the value of PCA for capturing the essential part of the complexity of high-dimensional systems. PCA is already standard for correcting for population substructure in GWAS; my results point to it's broader applicability as a general strategy for dealing with the high dimensionality of genomic association data.

keywords: GWAS, genetic association, genome-wide, statistics, correction, principal components analysis