Experimental Gerontology 46 (2011) 979–986

Contents lists available at SciVerse ScienceDirect

Experimental Gerontology

journal homepage: www.elsevier.com/locate/expgero

Complex phenotypes and phenomenon of genome-wide inter-chromosomal linkage disequilibrium in the human genome

Alexander M. Kulminski *

Center for Population Health and Aging, Duke University, Box 90408, Trent Hall, Room 002, Durham, NC 27708, USA

ARTICLE INFO

Article history:
Received 8 April 2011
Received in revised form 29 July 2011
Accepted 23 August 2011
Available online 31 August 2011

Section Editor: R. Westendorp

Keywords:
Linkage disequilibrium
Complex phenotypes
Epistasis
Gene networks
Epistatic evolution

ABSTRACT

Studies of non-human species show that loci on non-homologous chromosomes can be in linkage disequilibrium (LD). I focus on the Framingham Heart Study (FHS) participants to explore whether the phenomenon of inter-chromosomal LD can be caused by non-stochastic bio-genetic mechanisms in the human genome and be associated with complex, polygenic phenotypes. This paper documents remarkably strong and extensive LD among SNPs at loci on multiple non-homologous chromosomes genotyped using two independent (Affymetrix 50 K and 500 K) arrays. The analyses provided compelling evidences that the observed inter-chromosomal LD was unlikely generated by stochasticity, population or family structure, or mis-genotyping. The analyses show that this LD is associated with complex heritable phenotypes characterizing poor health. The inter-chromosomal LD was observed in parental and offspring generations of the FHS participants. These findings suggest that inter-chromosomal LD can be caused by bio-genetic mechanisms possibly associated with favorable or unfavorable epistatic evolution. This phenomenon can challenge our understanding of the role of genes and gene networks in regulating complex, polygenic phenotypes in humans.

1. Introduction

A fundamental problem for the aging populations worldwide is extending years of healthy life. Numerous studies show that complex (non-Mendelian) phenotypes characterizing health and longevity are heritable (Bergman et al., 2007; Christensen et al., 2006; Gogele et al., 2011; Levy et al., 2000; Levy et al., 2007; Martin et al., 2007). Heritability of such phenotypes motivates studies underpinning the role of genetic factors in their regulation that can lead to major breakthroughs in extending healthy lifespan. This is, however, an extremely complex problem because: (i) various processes accompanying aging might be associated with multiple genes and complex gene networks (Dato et al., 2010; Ungvari et al., 2010); (ii) genes can exhibit antagonistic pleiotropy (Alexander et al., 2007; Kulminski et al., 2010a; Martin, 2007; Summers and Crespi, 2010; Williams and Day, 2003); (iii) some heritable risk factors which are detrimental in midlife might become beneficial at older ages (Kulminski et al., 2008; Le Couteur and Simpson, 2011; McAuley et al., 2010); (iv) the same genes can predispose to some diseases but protect against the others (Charlesworth, 1996; Finch, 2010; Kulminski et al., 2011; Martin, 1999), (v) and even single nucleotide polymorphisms (SNPs) showing relatively strong linkage disequilibrium (LD) in the same gene can exhibit antagonistic effects on the same aging-related phenotypes due to gene–gene or gene–environment interactions (Kulminski et al., 2010b). Although many of these problems are well acknowledged, especially in candidate-gene studies, majority of them are not yet in the mainstream of current genome-wide association studies (GWAS).

Furthermore, GWAS often discovers allelic variants with unknown biological function. To more effectively map functional genes causing complex phenotypes in humans, the LD concept, referring to statistical connections of alleles at different loci in a population, is frequently exploited (Lander and Schork, 1994; Slatkin, 2008). Although the LD concept has been originally introduced assuming that loci can be on homologous as well as on non-homologous chromosomes, current common understanding is that LD characterizes statistical connections between nearby loci (Slatkin, 2008). Moreover, simulation studies suggested that notably useful levels of LD in humans should be within relatively narrow range of about 3 Kb on the same chromosome (Kruglyak, 1999). Recent studies have shown, however, that LD can span large areas on the same chromosome both in humans (Sabeti et al., 2007) and in non-human species (Dyer et al., 2007). Studies of LD also suggested that modest LD could extend across the entire genome in plants (Rostoks et al., 2006) and animals (Farnir et al., 2000). This genome-wide LD was attributed to specific of population structure of plants and animals arising due to largely high inbreeding rate (Farnir et al., 2000; Flint-Garcia et al., 2003; Rostoks et al., 2006).

A recent study also suggested that rare variants might create synthetic associations with common variants if they are associated with the same phenotype, i.e., rare variants can stochastically occur more often in association (i.e., LD) with one of the alleles of the common

Abbreviations: LD, linkage disequilibrium; GWAS, genome-wide association studies; FHS, Framingham Heart Study; MAF, minor allele frequency.

* Tel.: +1 919 684 4962.
E-mail address: Alexander.Kulminski@duke.edu.

0531-5565/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.exger.2011.08.010

variant linked to the same phenotype (Dickson et al., 2010). Such synthetic associations may span over a 2.5 Mb interval on homologous chromosomes (Dickson et al., 2010). Several studies reported also on weak LD observed between two alleles of genes located on non-homologous chromosomes in malaria parasites (Duraisingh et al., 2000; Happi et al., 2006). The observed LD was explained as a result of selective pressure through chloroquine.

Because LD plays an important role in GWAS, there is also a renewing interest in studies of epistatically-driven evolutionary selection and, consequently, in non-random transmission of genetically unlinked loci that can cause LD among loci on non-homologous chromosomes (Rohlfs et al., 2010). Furthermore, several recent studies (Graber et al., 2006; Petkov et al., 2005) have documented large-scale functional organization in the mammalian genome extending to different domains on non-homologous chromosomes that can provide a deterministic basis for inter-chromosomal LD. The history of these studies goes back to early 1920s, when numerous experiments documented apparently non-random transmission of parental genotypes to progeny (see, e.g., (Clegg et al., 1972; Korol et al., 1994; Malinowski, 1927; Sapre and Deshpande, 1987; Sivagnanasundaram et al., 2004) and references therein). These unusual results either resemble genetic linkage resulting in excess of parental genotypes in progeny (called quasi-linkage) or document excess of non-parental genotypes in progeny (called super-recombination) (Robinson, 1971). These phenomena were extensively studied in relation to population structure (Malinowski, 1927; Mike, 1977), and inheritance of phenotypes (Michie, 1953) and oncogenic viruses (Boyse, 1977).

This study addresses the question whether a phenomenon of the inter-chromosomal LD can be caused by fundamental bio-genetic mechanisms in the human genome and be associated with complex, polygenic traits. The paper focuses on 9274 genotyped participants of the Framingham Heart Study (FHS).

2. Methods

2.1. Data and quality control

The FHS data are available from the National Institutes of Health Genome-Wide Association Study Data Repository (SHARe) through the dbGaP. The FHS SHARe includes 14,428 participants comprising three cohorts of successive generations, i.e., the FHS (launched in 1948, 5209 respondents), the FHS Offspring (FHSO, launched in 1971–1975, 5124 offspring), and the 3rd Generation (launched in 2002, 4,095 grandchildren) cohorts. The FHS is a population-based longitudinal study following its participants for up to about 60 years. Selection criteria, study design and phenotypic data have been previously described (Dawber, 1980; Govindaraju et al., 2008; Splansky et al., 2007). Genotype data are available for 9274 participants for whom DNA samples have been drawn (Cupples et al., 2009). Genotyping was conducted using the Affymetrix 500 K (250 K Nsp and 250 K Sty) and supplemental 50 K Human Gene Focused arrays having no overlapping SNPs. After quality control (excluding if: missingness > 10%, Hardy–Weinberg equilibrium p-values $< 10^{-2}$, Mendel errors > 2%, and minor allele frequency (MAF) < 2%) and exclusion of non-autosomal chromosomes, about 368 K and 38 K SNPs were retained in each set.

2.2. Two-stage approach

The analyses have been performed using an original two-stage strategy. The first stage was to pre-select SNPs of interest from the 368 K + 38 K SNPs set on the basis of their associations with complex phenotypes. The second stage was the analysis of LD of the pre-selected SNPs. The same sample of genotyped individuals was used at each stage.

Prevalence of cardiovascular disease (CVD; diseases of heart and stroke occurred within the entire follow up period through 2007) and three quantitative traits, i.e., total cholesterol (TC), and systolic (SBP) and diastolic (DBP) blood pressures were representatively chosen as phenotypes for the pre-selection analyses at the first stage. The quantitative phenotypes were assessed at the baseline examination (mean age 38 years, standard deviation 9.3 years) in 8842 (N = 8311 for TC) genotyped FHS participants. 1711 cases of CVD were documented in this sample. The largest correlation was observed between the SBP and DBP, i.e., Pearson two-tiled correlation coefficient was $r = 0.82$. The other correlations were relatively weak with $r <= 0.29$.

To quantify the SNP-phenotype associations, I used a traditional univariate technique of GWAS. The associations were assessed using logistic/linear regression, as appropriate, and an additive genetic model (using *plink* v. 1.06 (Purcell et al., 2007)). All SNPs were ranked according to the p-values for associations with the phenotypes. A cut-off for significance of the SNP-phenotype associations was used for pre-selection of a reasonable number of SNPs for the second stage (i.e., analysis of LD); the choice of the cut-off neither affects validity nor conclusions of the LD analysis. This cut-off was representatively set at $p = 10^{-7}$.

The LD analysis at the second stage was performed by calculating pair-wise r^2 statistics among SNPs pre-selected at the first stage using *plink* v. 1.06 (Purcell et al., 2007). The LD statistics is based on allele frequencies estimated via the EM algorithm. Frequency of alleles for all analyses was evaluated using the exact test for founders only (Wigginton et al., 2005), which is more accurate for rare genotypes.

3. Results

3.1. Two stage strategy

A simple but powerful method to ascertain whether the inter-chromosomal LD could be considered as a deterministic phenomenon driven by non-trivial bio-genetic mechanisms in the human genome is to use a two stage approach (see Methods). The first stage in this approach is focused on pre-selection of SNPs with potentially useful levels of LD on the basis of a meaningful hypothesis. The second stage is to evaluate LD in the pre-selected set of SNPs. I hypothesize that a meaningful pre-selection of SNPs of interest can be accomplished on the basis of associations of SNPs with multiple complex phenotypes. The rationale for this hypothesis is that complex phenotypes are of polygenic nature (Gibson, 2009) and, thus, they are controlled by gene networks rather than by individual loci. Consequently, SNPs involved in a regulation of complex phenotypes might exhibit potentially useful levels of LD and, therefore, might be enriched in the pre-selected set. An important aspect of the pre-selection hypothesis is that specific of phenotypes to be used for SNP pre-selection is less important than their number or/and complexity. Thus, given the proposed approach, it should be expected that: (i) SNPs associated with complex phenotypes should be in LD, (ii) LD (if any) should be stronger among SNPs exhibiting more extensive pleiotropic effect, and (iii) SNPs with useful levels of LD should be over-represented in the pre-selected set compared to the set with no phenotype-based pre-selection.

3.2. SNP pre-selection

Candidate SNPs were pre-selected on the basis of their associations with prevalence of CVD (diseases of heart and stroke), total cholesterol (TC), and systolic (SBP) and diastolic (DBP) blood pressures (see Methods). This resulted in 5236 SNPs at loci on all autosomal chromosomes in genome. A majority of them, i.e., 3537 SNPs (67.6%), exhibits pleiotropic effect, i.e., they are associated with two to four phenotypes (Fig. 1A). Pleiotropy results in 12,782 associations for the 5236 SNPs at significance level $p \leq 10^{-7}$ with median for the most significant association (in the case of pleiotropy each SNP can be associated with up to four phenotypes) at $p = 5 \times 10^{-16}$. Given genome-wide significance level ($p < 5 \times 10^{-8}$), only 4.2% (536 associations) of these 12,782 associations have significance level above that, i.e., $5 \times 10^{-8} < p \leq 10^{-7}$.