Alzheimer's & Dementia 10 (2014) 592-601

Featured Articles

African-American TOMM40'523-*APOE* haplotypes are admixture of West African and Caucasian alleles

Allen D. Roses[a,b,c,*], Michael W. Lutz[a], Ann M. Saunders[a], Dmitry Goldgaber[d], Robert Saul[e], Scott S. Sundseth[c], P. Anthony Akkari[c], Stephanie M. Roses[a], W. Kirby Gottschalk[a], Keith E. Whitfield[f], Alexander A. Vostrov[d], Michael A. Hauser[g], R. Rand Allingham[h], Daniel K. Burns[b], Ornit Chiba-Falek[a], Kathleen A. Welsh-Bohmer[a]

[a]*Duke University Bryan Alzheimer's Disease Research Center, Duke University, Durham, NC, USA*
[b]*Zinfandel Pharmaceuticals Inc, Chapel Hill, NC, USA*
[c]*Cabernet Pharmaceuticals, Inc., Chapel Hill, NC, USA*
[d]*Department of Psychiatry, State University of New York, Stony Brook, NY, USA*
[e]*Polymorphic DNA Technologies, Alameda, CA, USA*
[f]*Department of Psychology, Duke University, Durham, NC, USA*
[g]*Center for Genetic Diseases, Duke University, Durham, NC, USA*
[h]*Department of Ophthalmology, Duke University, Durham, NC, USA*

Abstract

Background: Several studies have demonstrated a lower apolipoprotein E4 (*APOE* ε4) allele frequency in African-Americans, but yet an increased age-related prevalence of AD. An algorithm for prevention clinical trials incorporating TOMM40'523 (Translocase of Outer Mitochondria Membrane) and *APOE* depends on accurate TOMM40'523-*APOE* haplotypes.
Methods: We have compared the *APOE* and TOMM40'523 phased haplotype frequencies of a 9.5 kb TOMM40/*APOE* genomic region in West African, Caucasian, and African-American cohorts.
Results: African-American haplotype frequency scans of poly-T lengths connected in phase with either *APOE* ε4 or *APOE* ε3 differ from both West Africans and Caucasians and represent admixture of several distinct West African and Caucasian haplotypes. A new West African TOMM40'523 haplotype, with *APOE* ε4 connected to a short TOMM40'523 allele, is observed in African-Americans but not Caucasians.
Conclusion: These data have therapeutic implications for the age of onset risk algorithm estimates and the design of a prevention trial for African-Americans or other mixed ethnic populations.

Keywords: Phased Sanger sequencing; Phylogenetic analysis; Population admixture; *APOE*; *TOMM40*; haplotypes; African-American; Alzheimer's disease; Yoruban; Caucasian; Complex disease genetics; age of onset

1. Background

The apolipoprotein E4 (*APOE* ε4) allele frequency among Africans and African-Americans with late-onset Alzheimer's disease (LOAD) was lower than in Caucasians in several early *APOE* ε4 studies [1–3]. The relationship between *APOE* ε4 and the risk of LOAD in the Yoruban population, one of the Nigerian ethnic groups, showed the deleterious effect of *APOE* ε4, but the allele frequency of *APOE* ε4 in the Yorubans was lower than in Caucasians [1]. The lower *APOE* ε4 allele frequency in African-Americans with LOAD, coupled with a similar or higher incidence of dementia among African-Americans compared with Caucasians, has remained an area of great interest and discussion over the past two decades. These early studies

*Corresponding author. Tel.: 919-271-4076.
E-mail address: allen.roses@duke.edu

http://dx.doi.org/10.1016/j.jalz.2014.06.009

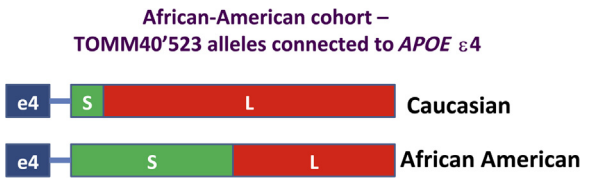

Fig. 1. Proportion of TOMM40'523 (Translocase of Outer Mitochondria Membrane) short and long alleles connected to apolipoprotein E4 (*APOE* ε4) chromosomes in a cohort of Ghanaian subjects compared with Caucasian subjects. The proportion of short alleles connected to *APOE* ε4 as a *cis*-haplotype in one of the two Caucasian discovery cohorts ("Canadian") was less than 2% (two alleles of >100 chromosomes) [8]. In the other cohort ("Arizona") there were no S alleles connected to *APOE* ε4, nor in other Caucasian cohorts subsequently tested. A large proportion of S alleles was attached to *APOE* ε4 chromosomes in a Ghanaian cohort tested for diversity in a series of ethnic cohorts [12]. These data formed the rationale for testing African and African-American cohorts once the poly-T sequence methodology allowed in-phase haplotyping and an accuracy of ± one base.

contributed to the proposal that another LOAD susceptibility gene may be operating in African-Americans [4,5]. This led to several association studies of promoter polymorphisms, which were noninformative [6,7].

The haplotype relationship between *APOE* ε4 and TOMM40'523 (Translocase of Outer Mitochondria Membrane) in an African-American cohort identified a distinct difference in the frequency of *APOE* ε4 alleles in-phase with a Short TOMM40'523 allele, a combination rarely observed in Caucasian cohorts (Fig. 1). An early study of a Ghanaian cohort found the allele frequency of *APOE* ε4 to be less than Caucasians or African-Americans. However, a significant proportion of *APOE* ε4 alleles was in phase with Short TOMM40'523 alleles. This suggested the presence of a possible cross-over event occurring in the Ghanaians, a distinct observation in West Africans that has been introduced in African-Americans and rarely observed in Caucasian cohorts [1] (Fig. 1). Several nonphased association studies of *APOE* regulatory region polymorphisms were performed in Caucasians and were not informative to the question of phased haplotypes in other ethnicities [6,7]. This study emphasizes the role of individual risk prediction for Mild Cognitive Impairment due to Alzheimer's Disease (MCI-AD) and/or LOAD based on primary TOMM40'523 length differences and the presence or absence of *APOE* ε2. Based on the comparison of allele frequencies in Caucasians, West Africans, and African-Americans reported in this communication, the conclusion is that individual risk prediction based on a new in-phase assay that directly determines haplotypes of TOMM40'523 and *APOE*, rather than imputing data based categorical poly-T lengths (short-S, long-L, and very long-VL) to determine which TOMM40'523 poly-T lengths are linked to *APOE* genotypes in other ethnicities which different variations of lengths were observed. Studies to precisely define a risk prediction algorithm for ethnicities other than Caucasians based on accurate age of onset data and TOMM40'523-*APOE* haplotypes are ongoing.

2. Results

2.1. Background of prior results

We reported in 2010 that base length variation of a poly-T tract (short sequence repeat (SSR), rs10524523; TOMM40'523) in the translocase of outer mitochondrial membrane 40 homolog (*TOMM40*) gene has been associated with LOAD age of onset [8]. *TOMM40* is adjacent to, and in LD with, the *APOE* gene. With the use of phylogenetic mapping within the LD region containing the *TOMM40* and *APOE* genes, considerable progress has been made in analyzing the role of genetics in LOAD. Rather than simply observing *APOE* allele frequencies for this complex neurological disease, it became critically important to review the data for haplotypes of TOMM40 and *APOE* to support Alzheimer's disease clinical trials using pharmacogenetics, and especially prevention studies using enrichment algorithms in their design. TOMM40'523 length-sequence groups were categorized as short (S: T ≤ 19), long (L: ≤ 20T ≤ 29), or very long (VL: T ≥ 30). The TOMM40'523 genotypes are informative for the LOAD age of onset for individuals who carry all *APOE* genotypes, not just *APOE* ε4-bearing genotypes: 97% of Caucasian individuals, unlike *APOE* ε4-bearing genotypes which are informative in only 29% of the population. If TOMM40'523 had been discovered before the association of *APOE* ε4 and LOAD in 1993, then TOMM40'523 would have been considered far more informative for age of onset distributions due to the richness of poly-T length variations at a single genetic locus [rs10524523] [9]. Two frequent *APOE* genotypes, *APOE* ε3/3, and *APOE* ε2/3, representing almost two-thirds of the Caucasian population were uninformative for LOAD risk using *APOE* ε4 carriage, but as TOMM40'523 genotyping includes all the *APOE* ε4 genotypes, the *APOE* ε3, and *APOE* ε2 data are now also informative for risk [10] (Fig. 2).

Using phylogenetic analysis, we now recognize that each copy of *APOE* ε3 (and *APOE* ε2) is attached on the same chromosome (in phase) to either of two TOMM40'523 allele sizes, S or VL, in Caucasians. The distinction was clear and S and VL forms were each found grouped separately within two major divergent phylogenetic clades. These data were used to describe the categories for simplicity [8,11]. This allows *APOE* ε3 to be resolved into two *APOE* ε3 haplotypes: *APOE* ε3-S and *APOE* ε3-VL. When observed as a genotype using these haplotypes there are three distinct *APOE* ε3/3–TOMM40'523 haplotypes (S/S, VL/VL, and S/VL), each with a distinct age of onset distribution. *APOE* ε2 also is also attached to S or VL poly-T, lengths but provide a later age of onset for *APOE* ε2/3 patients as previously observed [10] (Fig. 1). *APOE* ε4 haplotypes account for almost all Caucasian *APOE* ε4 carriage. There are six different TOMM40'523 genotypes: *APOE* ε4/4 each have as L/L genotype; for *APOE* ε3/4 there are two TOMM40'523 genotypes with one allele of L and the

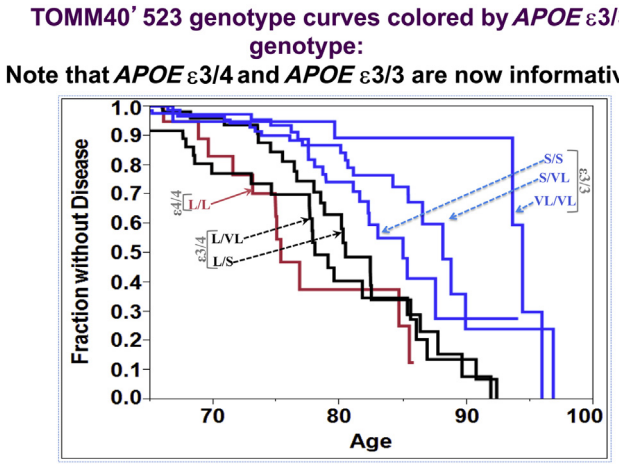

Fig. 2. Age at the onset of cognitive impairment as a function of TOMM40'523 (Translocase of Outer Mitochondria Membrane) genotypes. The Bryan Alzheimer's Disease Research Center (ADRC) Memory, Health, and Aging cohort (n = 508, 106 conversion events) was followed prospectively at the Bryan ADRC at Duke University. Cognitive Status was determined using standard neuropsychological tests. The age at which cognitive impairment occurred was retrospectively stratified by TOMM40'523 genotype, and Kaplan-Meier curves were constructed. TOMM40'523 genotypes and the corresponding apolipoprotein E (*APOE*) genotypes are indicated on the figure. The red line corresponds to *APOE* ε4/4; the two green lines correspond to *APOE* ε3/4,and the three blue lines correspond to *APOE* ε3/3. Note that within this cohort there were 78 individuals who carried an *APOE* ε2 allele, but only five developed cognitive impairment during the study (VL/L (*APOE* ε2/4), n = 1; S/L (*APOE* ε2/4), n = 2; VL/VL (*APOE* ε3/3), n = 2). The data for these individuals are indicated as points on the appropriate *TOMM40* genotype curve; open circles and filled diamonds indicate the age at onset of symptoms in the *APOE* ε2/4 and *APOE* ε2/3 groups, respectively. The numbers of subjects per TOMM40'523 genotype (number case to converted status) were L/L, (*APOE* ε4/4) n = 23 (11); VL/L (*APOE* ε3/4) n = 54 (24); S/L(*APOE* ε3/4) n = 72(23); S/S(*APOE* ε3/3 or *APOE* ε3/2), n = 100 (20); S/VL(*APOE* ε3/3 or *APOE* ε3/2) n = 138(22); VL/VL(*APOE* ε3/3 or *APOE* ε3/2), n = 51 (9). Abbreviations: S, short; L, long; VL, very long.

APOE ε3 allele being either S or VL resulting in two *APOE* ε3/4 genotypes: S/L or VL/L (Table 1 and Fig. 2) [9]. Most Caucasians are *APOE* ε3/3 and there are three possible TOMM40'523 genotypes: S/S, S/VL, or VL/VL, each with a different age of onset distribution. There are still too few *APOE* ε2/2 and *APOE* ε2/4 LOAD subjects observed in prospective age of onset cohorts to determine the accurate LOAD age of onset distributions for these two uncommon genotypes, but *APOE* ε2/3 distributions are displaced to the right [older onset] with later ages of onset due to the presence of an *APOE* ε2 allele [10]. One important practical consequence of these findings is that in planning LOAD prevention trials, it is more informative to review the genotypic age of onset data for the TOMM40'523-*APOE* haplotypes instead of the current practice of simply observing *APOE* allele frequencies for this complex neurological disease.

It became evident very early in the TOMM40'523 research that different ethnic groups have different haplotype frequency profiles of TOMM'523 poly-T lengths

Table 1
Caucasian TOMM40'523-*APOE* haplotypes and age at high risk used in TOMMORROW trial

523 or *APOE* genotype	Haplotype frequency	Age at high risk
523 L/L	2%	All high risk
523 L/VL	13%	All high risk
APOE ε2/4	3%	All high risk
523 S/L	14%	74
523 S/S	17%	77
523 S/VL	36%	76
523 VL/VL	18%	All low risk
APOE ε2/2	1%	All low risk
APOE ε2/3	13%	All low risk

Abbreviations: *APOE*, apolipoprotein E; S, short; L, long; VL, very long; *TOMM40*, Translocase of Outer Mitochondria Membrane.

NOTE. For the TOMMORROW Study, all *APOE* ε2 carriers are assigned to the high-risk or low-risk categories according to *APOE* genotype. Risk is assessed for all other subjects according to the *TOMM40* genotype. Subjects with *APOE* ε2/2, *APOE* ε2/3, and VL/VL are considered to be at low risk. Subjects with *APOE* ε2/2 and *APOE* ε2/4 do not have sufficient AD onset data, but represent 3% of the population., *APOE* ε2/4 were classified as high risk so that they would all be entered and randomized in the "high risk" treatment or placebo groups. Certain genotypes are considered to be high risk in the context of the targeted age range for the delay-of-onset clinical trial. Subjects with any of three 523 genotypes, S/L, S/S, and S/VL, will be placed in the high-risk category if the age at entry to the trial is greater than the age indicated. The TOMM40'523 genotype frequencies are from the Cache County Study of Memory cohort (n = 2042). The *APOE* genotype frequencies are those for Caucasian controls from the Farrer et al. meta-analysis study (n = 6262) [46].

than those found in several geographical diverse Caucasian cohorts [12] (Fig. 1). This manifests itself in several different ways. One relates to the differences between the lengths of poly-Ts attached to *APOE* ε4 or *APOE* ε3. The second ethnic variation can be the presence of a different haplotype not observed in Caucasians. Both of these differences will be presented in this communication regarding highly variable inheritance from two different ethnicities that are combined in African-Americans. The first type of poly-T length heterogeneity is due to the fact the TOMM40'523 poly-T occurs at a single locus but range in length from 10 to 50 bases. A distinct and broader sets of poly-T lengths characterizes many ethnic groups. The second involves a unique TOMM40'523 combination not observed in Caucasians, and one of these distinct haplotypes is present in West Africans and transmitted to African-Americans. The number of TOMM40'523 haplotypes is far richer for generating genotype combination than the two SNP loci defining *APOE* coding sequences (at amino acid positions 112 and 152 of 299).

2.2. New data

In this communication, we present new data for African-American and West African populations. We examined TOMM40'523 length distributions in African-Americans and Yorubans (Figs. 1, 3, 4). We will demonstrate that using a new phased sequencing method for *APOE* and

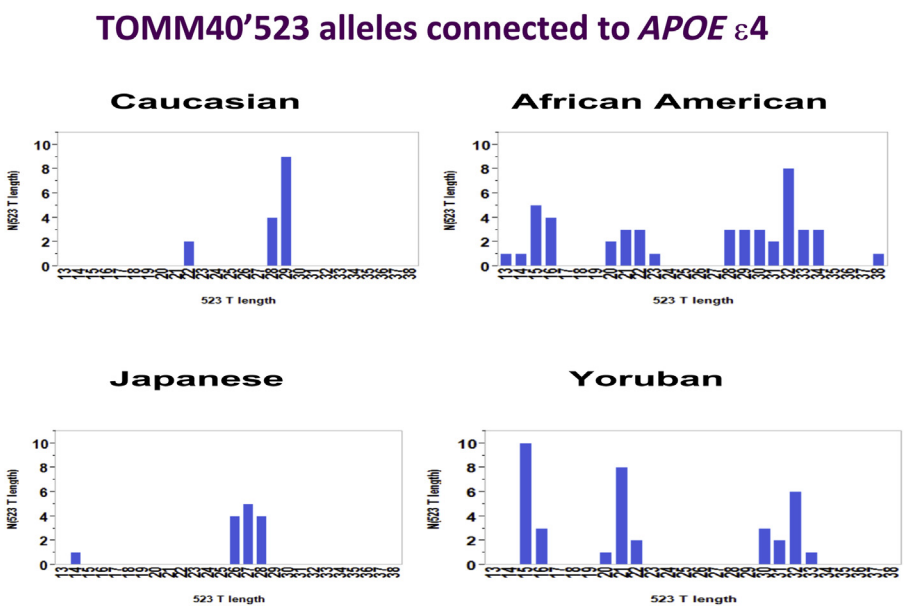

Fig. 3. Haplotypes of TOMM40'523 (Translocase of Outer Mitochondria Membrane) alleles connected to apolipoprotein E4 (*APOE* ε4). The length sizes of the poly-T alleles are indicated in each of the panels: Panel A upper left: Caucasians; Panel B upper right, African-Americans; Panel C lower right, Yorubans (West Africans), and Panel D lower left, Japanese. Note that the Caucasians in Panel A have two L poly-T groupings at T22 and T28/29. Yorubans in Panel C have three groupings: two L groupings at T20-22 and T30-33, the second peak being slightly longer than in Caucasians. In addition there is a distinct peak at S-T15-16 not observed in the Caucasians, In Panel B African-Americans carry the S poly-Ts at T13-16, like the Yorubans but rarely observed in Caucasians, In African-Americans L poly-T lengths including T28-T34 which includes the Caucasian contribution of T28-30 and the Yoruban contribution T30-33. Japanese are shown to demonstrate that other ethnicities have characteristic patterns of poly-Ts connected to *APOE* ε4 and, in this group, only T26-28—a bit shorter than the Caucasian long peak but with no T20-23. Abbreviations: S, short; L, long; VL, very long.

TOMM40 on the same [cis] haplotype fragment for the Caucasian, Yoruban, and African-American populations, it is possible to directly determine in-phase TOMM40'523-*APOE* haplotypes that were previously not possible to impute based on historical Caucasian data, particularly where the TOMM40'523 lengths overlap in the L–VL size transition. We have also demonstrated in both West African cohorts that that there are a large proportion of *APOE* ε4 alleles physically connected in-phase to S alleles, which is extremely rare or non-existent for Caucasians examined. There is now the opportunity to assess whether the TOMM40'523 risk algorithm developed for the TOMMORROW AD prevention trial can be generalized for other ethnic groups using quantitative lengths and distinct haplotypes [9].

In a previously reported study, we examined control sets of Caucasians and Ghanaians with respect to *APOE* and *TOMM40* allele frequencies [12] (Fig. 1). At that time the technique to perform phased haplotyping in this region was not available, thus we had not been able to accurately determine the proportion of TOMM40'523 alleles connected to *APOE* alleles [12]. The data were generated using the categorical sizes for TOMM40'523 based on Caucasian classification. However, in Caucasians 98–100% of the TOMM40'523 L variants were *APOE* ε4-connected and <2% were connected to a S-TOMM40'523 poly-T. Nonphased sequencing of Ghanaians suggested that the S-TOMM40'523 alleles were connected to both *APOE* ε4 and *APOE* ε3 alleles (Fig. 1). There was a striking proportion of *APOE* ε4 alleles linked to S TOMM40'523 alleles observed in African-Americans but not observed in Caucasians (Fig. 3). Confirmation experiments were planned using publically available population control samples from the Coriell Cell Repository to determine if the original Caucasian and Ghanaian data could be replicated and to provide a standardized data set for other investigators to confirm [12]. Fig. 3 compares the TOMM40'523 genotypes from four defined ethnic groups: Caucasian, Yoruban, and African-American; a fourth ethnicity from Japan is also illustrated to further demonstrate the differences in the relationship of *APOE* ε4 and the two L-TOMM40'523 allele peaks in another ethnic group [12]. In the Caucasian population (Fig. 3, Panel A) the length of the TOMM40'523 alleles connected to *APOE* ε4 is displayed, and do not differ from the other normal Caucasian series previously published [8,13]. There are two distinct and separated TOMM40'523 L peaks at L-T = 22–23 and L-T28-29) in Caucasians (only one in Japanese; Fig. 3, Panel A and D) S poly-T lengths do not link with *APOE* ε4 alleles in Caucasians, but there is a previously undetected overlap in length with the smallest of the VL lengths observed in on *APOE* ε3 and *APOE* ε2 strands that would be labeled L in the convention used since 2010 (Fig. 3, Panel A). In

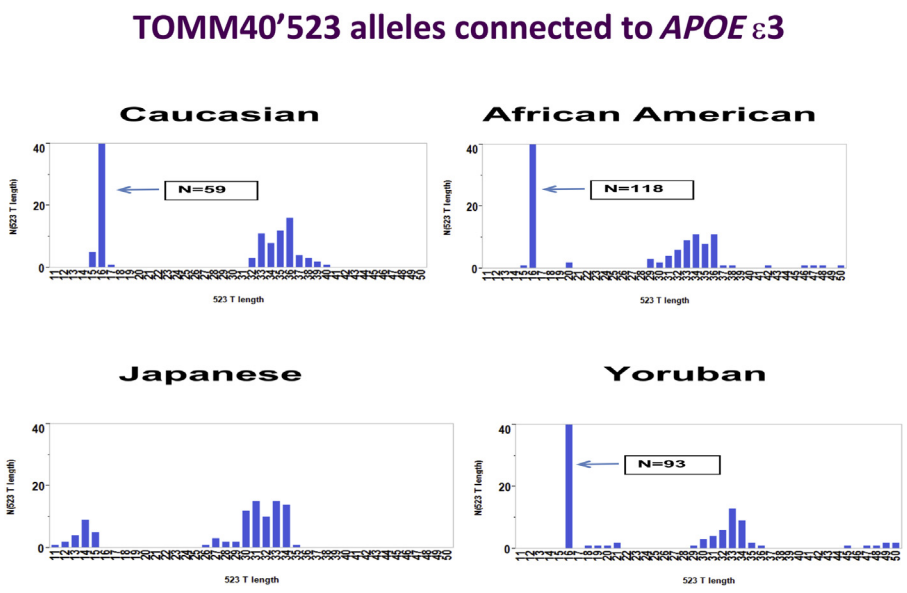

Fig. 4. Haplotypes of TOMM40'523 (Translocase of Outer Mitochondria Membrane) alleles connected to apolipoprotein E3 (*APOE* ε3). The panels are arranged as in Fig. 3. Caucasians in several series have two well-characterized peaks in the S-T15-16 range and the VL-T32-40 range. The Yorubans have a larger proportion of S to VL poly-Ts connected to *APOE* ε3 and a greater proportion of S-T15 alleles than Caucasians (assay = /−1 base accuracy, but the proportions are different in these analyses). In the Yorubans, the range of VL alleles extends to the 45-50T range, extremely rare in Caucasians. In the Yorubans S range there are also T19-21 alleles connected to *APOE* ε3, also not observed in Caucasians. The African-Americans demonstrate all the alleles carried in Caucasians and Yorubans, particularly the T19-20 and the T45-50 characteristic of the Yorubans. Again the Japanese demonstrate a "shorter" VL range of 26 to 35 compared with the Caucasians, Yorubans, and the African-Americans. There are a greater proportion of VL to S alleles in *APOE* ε3-connected haplotypes in the Japanese, consistent with a later age of onset of AD in Japanese population who have a higher proportion of the *APOE* ε3–VL/VL genotypes. Abbreviations: S, short; L, long; VL, very long.

the Yoruban population (Fig. 3, Panel C), there are three peaks of poly-Ts linked to the *APOE* ε4 alleles: a S-T15 allele which is uncommon in Caucasians, the L-T20-23 peak is similar to the shorter of the two L peaks in Caucasians, and only a few longer longs (T30-33) from the previously defined (in Caucasians) overlap region between L and VL. The predominant L (T 28–29) peak observed in Caucasians is conspicuously absent in Yorubans. The African-American samples (Fig. 3, Panel B) illustrate the overlapping range of L sizes attached to *APOE* ε4 strands from both the Yoruban (T30-33) and Caucasians (T28-29). Note that the Japanese population only has the predominant L (T24-28) peak (Fig. 3, Panel C) and, in this relatively small illustrated cohort, excludes the shorter L (T20-23) of the two L Caucasian peaks previously published [1,8,9]. This further illustrates the heterogeneity between ethnicities with recognizable genetic variations [5,8].

Fig. 4 represents the histograms for *APOE* ε3-connected TOMM40'523 poly-T lengths. Connected to *APOE* ε3 in the African-American samples are S lengths, including short (T14-16) peaks observed frequently in Caucasians, and an approximately equal number of longer VL lengths of TOMM40'523. The Yorubans have a predominant S-T16 peak attached to *APOE* ε3 (Fig. 4, Panel D) compared with a predominant S-T15 attached to *APOE* ε4 (Fig. 3, Panel D and Fig. 1). African-Americans have both sets of data (Fig. 3 and 4, Panel B).

3. Discussion

APOE ε3 is the most common *APOE* allele in all populations. TOMM40'523-*APOE* in-phase haplotypes are more accurate than simply *APOE* ε4 alone in Caucasians. We hypothesize that these informative genotypes (both inherited haplotypes) are generalizable and predictive in other ethnic populations. Because the *APOE* ε4-TOMM40'523 haplotypes contain all the *APOE* ε4-containing genotypes, the addition of *APOE* ε3-TOMM40'523 and *APOE* ε2-TOMM40'523 haplotypes is especially important in ethnicities with low allele frequencies of *APOE* ε4. In Chinese, Korean, and Japanese populations, the *APOE* ε4 allele frequency is approximately one third of Caucasians yet the prevalence of LOAD in the populations also grows with age with the main genotype being *APOE* ε3/3. It is now clear that *APOE* ε3-*TOMM40* haplotypes do not provide a neutral effect on onset of LOAD, but map with later age of onset distributions. The literature frequently considered *APOE* ε3 as a normal comparator to *APOE* ε4 carriage, but only the *APOE* ε2/2 genotype appears almost totally preventive.

The joint interaction of the expressed and metabolized protein products of the *APOE* and *TOMM40* genes, and possibly other promoters or modifiers within the same LD region, has supported the hypothesis that both sets of polymorphisms contribute to the early, preclinical pathogenesis of AD, whereas still other interactions may accelerate the

pathogenic processes observed later in the clinical and pathological course of the disease [14].

The "missing heritability" of AD can be defined with fully informed TOMM40'523-*APOE* haplotypes, derived from a phylogenetic map and the ability to assess grouping of TOMM40'523 lengths grouped on the phylogenetic map [8,9]. Searching the genome with more than two dozen genome-wide association studies (GWAS) has not defined "missing heritability". There have been many gene hypotheses from GWAS, and imputing the role of SNPs in noncoding regions near genes, but little successful translation to drug discovery or development for LOAD. TOMM40'523-*APOE* haplotypes account for >98% of the age-dependent risk of onset. Four of the five SNPs represented on the commercially available chips as in the "*APOE*" LD region are Single nucleotide polymorphisms SNPs from the *TOMM40* gene, yet these data are frequently interpreted as support only for *APOE* solely because they are in LD. Our data would suggest a single structural variation located in a *TOMM40* intron may affect the expression of *TOMM40*, *APOE*, and possibly other nearby genes or structural variants to influence the metabolism leading to age-dependent expression of early metabolic imaging studies and expression of clinical disease [13-15]. It is interesting to observe the recent interest in "*APOE*" in the LOAD scientific field, especially when the data suggests that another critical gene adjacent to *APOE* is involved in the pathogenesis. In fact, the two specific *APOE* SNPs (at positions 112 and 152 of 299 amino acids) were not included on either of the two popular GWAS testing platforms nor could they alone generate the extremely high statistical association of the "*APOE* region." GWAS actually tested *TOMM40* SNPs and a SNP located in the 3′ non-coding region of *APOE* to generate all the GWAS data, but haplotypes from this region of LD involving both the *APOE* and *TOMM40* genes predict the age of onset distribution better than *APOE* alone. The poly-T variations at the TOMM40'523 locus are used to identify haplotypes associated with specific clades by phylogenetic mapping. The locations of similar poly-T length variants supports the hypothesis that much of the hereditability is located at this highly variant locus. Human triplet repeat diseases with intra- and intergenerational variability in the length at a single locus is usual, but this is not the same as the stable clad-specific inheritance pattern as TOMM40'523. However, triplet repeat disease variable lengths specifically affect the expression of the disease phenotype [16–18].

We hypothesize the likelihood of developing LOAD among African-Americans is more accurately determined by consideration of the TOMM40'523-*APOE* haplotype than the *APOE* isoform alone, which is based on Caucasian data; however we cannot presently test this hypothesis. Several African-American MCI-AD/LOAD age-of-onset studies have been initiated, but it will be years before prospective data will be available. Although the current TOMMORROW clinical trial permits enrollment of normal African-American subjects, registration will be based on Caucasian data. Those data are also years away.

Our data supports the hypothesis, derived from the data of several African-American cohorts in the United States and two from West Africa, appear to represent a genetic cross between West Africans and Caucasians. A large proportion of the African-American population carries the L-TOMM40'523 allele connected to an *APOE* ε3 allele, which until now has only been observed in West Africans. In Caucasians, >98% of the long TOMM40'523 alleles are connected to *APOE* ε4. A new S(T = 15) TOMM40'523 length allele connected to *APOE* ε4 is also characteristic of West African populations and is observed in the African-Americans in far greater frequency than in Caucasians. The fact that African-Americans share both sets of L-connected alleles might explain the higher frequency of AD in this population than would be expected by *APOE* ε4 alone [1–7]. Confirmation of the role of these specific inherited poly-T lengths relative to age of onset will require in-phase assays that are more accurate for more exact haplotype calls in individual subjects.

The analysis of individuals in the African-American and West African (Yoruban and Ghanaian) population showed that both of the short TOMM40'523 allele groups (T = 15 and T = 16) are connected to *APOE* ε3 and *APOE* ε4 in approximately equal proportions. Similarly, in the Yorubans and the African-Americans, there is connection of *APOE* ε4 to a S (T15) allele, which is observed only rarely in Caucasians. Up to now, this large proportion of *APOE* ε4 attached to the S (T15) allele has not been observed in any other population [19]. The presence of S TOMM40'523 alleles connected to *APOE* ε4 in approximately equal proportions to those connected to *APOE* ε3 in African-Americans can be explained by admixture in recent history, where Caucasians contributed mostly L-T28-29) *APOE* ε4-TOMM40'523 chromosomes and West Africans contributed the L-T30-32) chromosomes. The connection of either short T15 or T16 to *APOE* ε4 in Caucasians is extremely uncommon in all cohorts studied to date. This highlights the importance of looking at an individual's ancestry. As was demonstrated using two new independently performed poly-T assay methods, the S-*APOE* ε4-haplotype is inherited from the Yorubans has a specific and slightly different number of S-T15 poly-Ts in this genomic region. This strongly supports the proposition that a cross-over mutation occurred in the West African population, that is not observed in Caucasians. We have no data to implicate whether the S-T15 allele affects the age of onset distribution of *APOE* ε4 differently, but his can now be tested in prospective age of onset study. However, this haplotype was transferred to the African-Americans from West African ancestors.

At the time of the earlier Ghanaian experiments, the accuracy of the original assay was within ±2 Ts. It was

not until the phased sequencing of a 9.5 kb *TOMM40/APOE* genomic region (Polymorphic DNA Technologies, Inc; see Supplementary Methods) was developed that it became possible to differentiate the S-T15 peak from the S-T16 peak, and linkage of S-T15 to *APOE* ε4 in the West Africans and, subsequently, in the African-Americans. Simply put, the African-American samples represent a mixture of the characteristic TOMM40-*APOE* haplotypes in both West African Yorubans (*APOE* ε4 and *APOE* ε3 short) and Caucasians (only *APOE* ε4). Before assays for the continuity of TOMM40'523 and *APOE* were possible, *APOE* ε4 measurements in African-Americans underestimated the L-TOMM40'523 allele because the haplotype S15-L had not been encountered previously attached to *APOE* ε4. Because a significant proportion of *APOE* ε3/4 patients could be homozygous for *APOE* ε4, underestimating the S-*APOE* ε4 homozygous effect simply using the S classification in the context of the *APOE* ε3/3 and *APOE* ε3/4 subject groups. Similarly it would skew interpretations of S/S homozygotes and each of the other categories. The situation is straightforward: the TOMM40'523-*APOE* haplotypes used for age of onset stratification for African-Americans would be more accurate if all of the TOMM40'523-*APOE* haplotypes were incorporated. An algorithm for predicting high risk African-American individuals to enter a prevention trial would be based on all of the African-American haplotypes instead of using the original Caucasian poly-T length-based claasification of the risk algorithm.

One must be careful defining "African-American" populations based on a social nomenclature. In fact, Africans who may have emigrated from Africa recently, with no Caucasian genetic interactions, can be characterized socially as African-Americans when they simply reside in the United States. In other countries where immigration is common, African-Americans are referred to with different social classifications, and are not defined genetically. African-Caribbean is a commonly used classification in the UK. Due to diffuse migration patterns, scientists should be very careful to avoid expectations that social classifications reflect the specific genetic heritage in all cases. Geographic relationships among these groups may be more powerful than any social classification. For example, in a previously reported Caucasian series from Canada, two individuals were observed to carry a S poly-T length linked to *APOE* ε4 [8]. In that study, L was reported as "98%" attached to *APOE* ε4 strands, but was later observed to be 100% in several other dispersed Caucasian analyses, including the Arizona series in the discovery study and Caucasian cohorts from Tomsk, Siberia, and Russian immigrants to Israel. Thus, it is possible that those two Canadian individuals will have had African ancestors but still self-reported as Causasian [19].

The presence of S TOMM40'523 alleles connected to *APOE* ε4 in approximately equal proportions to those connected to *APOE* ε3 in African-Americans can be explained by admixture in recent history, where Caucasians contributed mostly long (T 28–29) *APOE* ε4-TOMM40'523 chromosome and West Africans contributed the long (T 30–32) chromosomes that could have been characterized as VL using the Caucasian-based convention. The connection of either S-T15 or S-T16 to *APOE* ε4 in Caucasians is extremely uncommon in all cohorts studied to date. This highlights the importance of looking at an individual's ancestry in studied populations. As was demonstrated using two new performed poly-T assay methods (see Supplementary Methods), the S-TOMM40'523-*APOE* ε4 haplotype inherited from the Yorubans has a specific and slightly different number of S15 poly-Ts in this genomic region. This strongly supports the proposition that a cross-over event occurred in the West African population, that was not observed in Caucasians.

3.1. *Possible role of* TOMM40 *in mitochondria and metabolism*

TOMM40 encodes the principle mitochondrial protein import pore, and is therefore critically important for mitochondrial biogenesis and function [20]. Research is underway in African-American AD subjects and normal controls to determine how accurately the *APOE* genotype or the specific *TOMM40* genotype describes the major susceptibility influence for LOAD, or whether both interact and may contribute to the pathogenesis in mitochondria leading to characteristic neuronal death and neuropathology [21,22]. *APOE* fragments generated within neurons may affect mitochondrial dynamic functions [23,24]. If this role is verified, this could also explain both the early and long-term effects of *APOE* genotypes on glucose utilization observed using PET glucose and/or BOLD oxygen consumption models [25–31]. It is critically important to note that the conclusions of such age-related metabolic imaging studies, especially positron emission tomography (PET) imaging, demonstrate life-long neuroanatomically consistent regional decreased glucose utilization related to *APOE* genotypes [32]. All the DNA samples from these previous studies could be analyzed, if they were available and had been consented properly in the past, so that the imaging studies could be related to specific TOMM40'523 genotypes. Our hypothesis would be that the long allele lengths depict the genetic risk of the neuronal degeneration leading to the later manifestations of the disease pathogenesis, including the amyloid cascade, tau deposition, atrophy, and the tragic clinical disease.

The mechanism by which TOMM40'523-*APOE* haplotype variation affects mitochondrial structure and function will require extensive research. It is generally proposed that structural DNA mutations can exert their effects through transcriptional and post-translational maturation and processing, which can result in differences in alternative splicing or protein expression. Recently, we demonstrated that TOMM40'523 affects the expression of *APOE* and TOMM40 mRNAs in brain tissues (temporal and occipital

cortexes) from LOAD patients and normal subjects. The affect of this poly-T variation on transcription regulation was supported in a cell-based luciferase reporting system [12]. The number of protein import channels per mitochondrion may be regulated by TOMM40'523 variations. The downstream effects may appear over a comparatively long time span (lifespan) and be associated with other environmental factors. With neuronal diseases affecting mitochondria, it should be recalled that neurons receive their initial complement of mitochondria maternally, and the life spans of these mitochondria are limited to the reproductive efficiency of these mitochondria. Unlike most other tissues in the body, the brain has little cellular capacity to replace mitochondria damaged by ROS generated by the high rates of glucose combustion, compared with dividing cells in other tissues, so the reproductive capacity and dynamics of neuronal mitochondria is a critical factor in determining neuronal health.

Consistent with this hypothesis, the TOMMORROW clinical trial now in-progress will test whether a drug that increases glucose and oxygen metabolism by increasing mitochondrial function and number can delay the onset of MCI-AD in those subjects of highest risk for onset within the five years of the study [9]. From an exploratory viewpoint, the data derived from this large population-based study may provide extensive evidence for the role of TOMM40'523 variants in the presentation and progression of MCI to LOAD clinical signs and symptoms [11,33,34].

It is also reasonable to expect that TOMM40'523 variants may involve mitochondria in the pathogenesis of other neurodegenerative and neuropsychiatric diseases. The relationship of other aggregating peptides, such as α-synuclein, may affect mitochondrial dynamics and select other specific neuronal cell types characteristic of other neurodegenerative diseases [35–40]. Clearly there is value in investigating the participation of other molecules, like synuclein and TDP-43 (transactive response DNA-binding protein 43, for mitochondrial interactions in neuronal cells specifically affected in other neuropathologically defined diseases [39,40]). Similar neuronal mitochondrial mechanisms processes may also underlie other neurodegenerative and metabolic diseases. The coexistence of multiple disease-associated pathologies in AD, Parkinson's disease, fronto-temporal dementia, amyotrophic lateral sclerosis and other diseases may stimulate new discoveries and insights into early pathogenesis for focused targets for discovery and drug development.

The genetic association of a GGGGCC (G4C2) hexanucleotide repeat expansion in the chromosome 9 open reading frame 72 (C9ORF72) gene is a common cause of amyotrophic lateral sclerosis and frontotemporal dementia should be also be investigated genetically using phylogenetic analysis to identify specific haplotype sharing and translational downstream mechanisms [41–44].

RESEARCH IN CONTEXT

1. Systematic Review: Allele frequency data for *TOMM40'523* poly-T lengths have been replicated using three new assays. The current assays used for this research are each accurate to ± 1 T. One of the new assays is an in-phase assay that allows chromosome strand-specific haplotypes to be determined accurately, without an imputation step based on Caucasian poly-T lengths, especially clarifying the haplotypes in the region of long (L) and very long (VL) overlap.
2. Interpretation: In the Yorubans there has been a cross-over mutation leading to a new haplotype. This haplotype combines apolipoprotein E4 (*APOE* ε4) with a Short-T15 allele, and represents many of the *APOE* ε4 containing haplotypes in Ghanaian and Nigerian Yoruban inheritance. Thus the *APOE* ε4 allele frequency is attached to a new S-T15 TOMM40'523 allele but there are no age of onset data for the new haplotype.
3. Future directions: African-Americans carry both Caucasian-specific sizes linked to *APOE* ε4, and S-T15- *APOE* ε4, as well as slightly larger sized *APOE* ε4 alleles in *TOMM40'523*-haplotypes. This is evidence of genetic admixture between West Africans and Caucasians.

References

[1] Hendrie HC, Ogunniyi A, Hall KS, Baiyewu O, Unverzagt FW, Gureje O, et al. Incidence of dementia and Alzheimer disease in 2 communities: Yoruba residing in Ibadan, Nigeria, and African Americans residing in Indianapolis, Indiana. JAMA 2001;285:739–47.
[2] Hendrie HC, Osuntokun BO, Hall KS, Ogunniyi AO, Hui SL, Unverzagt FW, et al. Prevalence of Alzheimer's disease and dementia in two communities: Nigerian Africans and African Americans. Am J Psychiatry 1995;152:1485–92.
[3] Hall KS, Gao S, Baiyewu O, Lane KA, Gureje O, Shen J, et al. Prevalence rates for dementia and Alzheimer's disease in African Americans: 1992 versus 2001. Alzheimers Dement 2009;5:227–33. PMCID: 2718566.
[4] Tang MX, Maestre G, Tsai WY, Liu XH, Feng L, Chung WY, et al. Relative risk of Alzheimer disease and age-at-onset distributions, based on APOE genotypes among elderly African Americans, Caucasians, and Hispanics in New York City. Am J Hum Genet 1996; 58:574–84. PMCID: 1914582.
[5] Tang MX, Cross P, Andrews H, Jacobs DM, Small S, Bell K, et al. Incidence of AD in African-Americans, Caribbean Hispanics, and Caucasians in northern Manhattan. Neurology 2001;56:49–56.
[6] Tycko B, Lee JH, Ciappa A, Saxena A, Li CM, Feng L, et al. APOE and APOC1 promoter polymorphisms and the risk of Alzheimer disease in African American and Caribbean Hispanic individuals. Arch Neurol 2004;61:1434–9.

[7] Tang MX, Stern Y, Marder K, Bell K, Gurland B, Lantigua R, et al. The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. JAMA 1998;279:751–5.
[8] Roses AD, Lutz MW, Amrine-Madsen H, Saunders AM, Crenshaw DG, Sundseth SS, et al. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. Pharmacogenomics J 2010;10:375–84.
[9] Crenshaw DG, Gottschalk WK, Lutz MW, Grossman I, Saunders AM, Burke JR, et al. Using genetics to enable studies on the prevention of Alzheimer's disease. Clin Pharmacol Ther 2013;93:177–85.
[10] Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 1993;261:921–3.
[11] Roses AD, Lutz MW, Crenshaw DG, Grossman I, Saunders AM, Gottschalk WK. TOMM40 and APOE: requirements for replication studies of association with age of disease onset and enrichment of a clinical trial. Alzheimers Dement 2013;9:132–6.
[12] Linnertz C, Saunders AM, Lutz MW, Crenshaw DM, Grossman I, Burns DK, et al. Characterization of the poly-T variant in the TOMM40 gene in diverse populations. PloS one 2012;7:e30994. PMCID: 3281049.
[13] Caselli RJ, Dueck AC, Huentelman MJ, Lutz MW, Saunders AM, Reiman EM, et al. Longitudinal modeling of cognitive aging and the TOMM40 effect. Alzheimers Dement 2012;8:490–5. PMCID: 3483561.
[14] Roses AD, Saunders AM, Lutz MW, Zhang N, Hariri AR, Asin KE, et al. New applications of disease genetics and pharmacogenetics to drug development. Curr Opin Pharmacol 2014;14:81–9.
[15] Reiman EM, Caselli RJ, Yun LS, Chen K, Bandy D, Minoshima S, et al. Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E. N Engl J Med 1996; 334:752–8.
[16] Morales F, Couto JM, Higham CF, Hogg G, Cuenca P, Braida C, et al. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. Hum Mol Genet 2012;21:3558–67.
[17] Budworth H, McMurray CT. A brief history of triplet repeat diseases. Methods Mole Biol 2013;1010:3–17. PMCID: 3913379.
[18] Krzyzosiak WJ, Sobczak K, Wojciechowska M, Fiszer A, Mykowska A, Kozlowski P. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. Nucleic Acids Res 2012; 40:11–26. PMCID: 3245940.
[19] Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science 2014; 343:747–51.
[20] Chang S, ran Ma T, Miranda RD, Balestra ME, Mahley RW, Huang Y. Lipid- and receptor-binding regions of apolipoprotein E4 fragments act in concert to cause mitochondrial dysfunction and neurotoxicity. Proceedings of the National Academy of Sciences of the United States of America 2005;102:18694–9. PMCID: 1311737.
[21] Mahley RW, Huang Y, Weisgraber KH. Detrimental effects of apolipoprotein E4: potential therapeutic targets in Alzheimer's disease. Curr Alzheimer Res 2007;4:537–40.
[22] Roses AD, Einstein G, Gilbert J, Goedert M, Han SH, Huang D, et al. Morphological, biochemical, and genetic support for an apolipoprotein E effect on microtubular metabolism. Ann N Y Acad Sci 1996; 777:146–57.
[23] Roses AD, Saunders AM, Huang Y, Strum J, Weisgraber KH, Mahley RW. Complex disease-associated pharmacogenetics: drug efficacy, drug safety, and confirmation of a pathogenetic hypothesis (Alzheimer's disease). Pharmacogenomics J 2007;7:10–28.
[24] Chen K, Ayutyanont N, Langbaum JB, Fleisher AS, Reschke C, Lee W, et al. Correlations between FDG PET glucose uptake-MRI gray matter volume scores and apolipoprotein E epsilon4 gene dose in cognitively normal adults: a cross-validation study using voxel-based multi-modal partial least squares. NeuroImage 2012; 60:2316–22. PMCID: 3325642.
[25] Chen K, Bandy D, Reiman E, Huang SC, Lawson M, Feng D, et al. Noninvasive quantification of the cerebral metabolic rate for glucose using positron emission tomography, 18F-fluoro-2-deoxyglucose, the Patlak method, and an image-derived input function. J Cereb Blood Flow Metab 1998;18:716–23.
[26] Langbaum JB, Chen K, Launer LJ, Fleisher AS, Lee W, Liu X, et al. Blood pressure is associated with higher brain amyloid burden and lower glucose metabolism in healthy late middle-age persons. Neurobiology Aging 2012;33:827.e11–9. PMCID: 3236809.
[27] Nicholson RM, Kusne Y, Nowak LA, LaFerla FM, Reiman EM, Valla J. Regional cerebral glucose uptake in the 3xTG model of Alzheimer's disease highlights common regional vulnerability across AD mouse models. Brain Res 2010;1347:179–85. PMCID: 2974951.
[28] Nicolson TJ, Bellomo EA, Wijesekara N, Loder MK, Baldwin JM, Gyulkhandanyan AV, et al. Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. Diabetes 2009;58:2070–83. PMCID: 2731533.
[29] Protas HD, Chen K, Langbaum JB, Fleisher AS, Alexander GE, Lee W, et al. Posterior cingulate glucose metabolism, hippocampal glucose metabolism, and hippocampal volume in cognitively normal, late-middle-aged persons at 3 levels of genetic risk for Alzheimer disease. JAMA Neurol 2013;70:320–5. PMCID: 3745014.
[30] Xu G, McLaren DG, Ries ML, Fitzgerald ME, Bendlin BB, Rowley HA, et al. The influence of parental history of Alzheimer's disease and apolipoprotein E epsilon4 on the BOLD signal during recognition memory. Brain 2009;132(Pt 2):383–91. PMCID: 2724919.
[31] Roses AD, Saunders AM, Burns DK, Saul RL, Lutz MW. From Bench to bedside: translation into drug development space. Expert Opin Drug Disov 2013. *in press*.
[32] Reiman EM. Alzheimer's disease and other dementias: advances in 2013. Lancet Neurol 2014;13:3–5.
[33] Lyall DM, Harris SE, Bastin ME, Munoz Maniega S, Murray C, Lutz MW, et al. Alzheimer's disease susceptibility genes APOE and TOMM40, and brain white matter integrity in the Lothian Birth Cohort 1936. Neurobiol Aging 2014;35:1513.e25–33.
[34] Lyall DM, Royle NA, Harris SE, Bastin ME, Maniega SM, Murray C, et al. Alzheimer's disease susceptibility genes APOE and TOMM40, and hippocampal volumes in the Lothian birth cohort 1936. PloS one 2013;8:e80513. PMCID: 3829876.
[35] Irwin DJ, Lee VM, Trojanowski JQ. Parkinson's disease dementia: convergence of alpha-synuclein, tau and amyloid-beta pathologies. Nat Rev Neurosci 2013;14:626–36.
[36] Linnertz C, Anderson L, Gottschalk W, Crenshaw D, Lutz MW, Allen J, et al. The cis-regulatory effect of an Alzheimer's disease-associated poly-T locus on expression of TOMM40 and apolipoprotein E genes. Alzheimers Dementia 2014.
[37] Mastaglia FL, Rojana-udomsart A, James I, Needham M, Day TJ, Kiers L, et al. Polymorphism in the TOMM40 gene modifies the risk of developing sporadic inclusion body myositis and the age of onset of symptoms. Neuromuscul Disord 2013;23:969–74.
[38] Wilson RS, Yu L, Trojanowski JQ, Chen EY, Boyle PA, Bennett DA, et al. TDP-43 pathology, cognitive decline, and dementia in old age. JAMA Neurol 2013;70:1418–24. PMCID: 3830649.
[39] Xia Y, Rohan de Silva HA, Rosi BL, Yamaoka LH, Rimmler JB, Pericak-Vance MA, et al. Genetic studies in Alzheimer's disease with an NACP/alpha-synuclein polymorphism. Ann Neurol 1996;40:207–15.
[40] Consortium A, Ahmeti KB, Ajroud-Driss S, Al-Chalabi A, Andersen PM, Armstrong J, et al. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. Neurobiol Aging 2013; 34:357.e7–35719. PMCID: 3839234.
[41] Fecto F, Siddique T. UBQLN2/P62 cellular recycling pathways in amyotrophic lateral sclerosis and frontotemporal dementia. Muscle Nerve 2012;45:157–62.
[42] van Blitterswijk M, DeJesus-Hernandez M, Niemantsverdriet E, Murray ME, Heckman MG, Diehl NN, et al. Association between

repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. Lancet Neurol 2013;12:978–88. PMCID: 3879782.

[43] Zu T, Liu Y, Banez-Coronel M, Reid T, Pletnikova O, Lewis J, et al. RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. Proceedings of the National Academy of Sciences of the United States of America 2013;110:E4968–77. PMCID: 3870665.

[44] Fazekas A, Steeves R, Newmaster S. Improving sequencing quality from PCR products containing long mononucleotide repeats. BioTechniques 2010;48:277–85.

[45] Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA 1997;278:1349–56.

Supplemental methods

Clinical data sets

The Ghanaian, Japanese, and original African-American samples were published previously [12]. The population controls of African-Americans and Yoruban samples were purchased from Coriell Cell Repositories (Camden, NJ), and are independent of the previously published collections referred to above and cohorts published in the original TOMM40'523 (Translocase of Outer Mitochondria Membrane) report. [8]

(A) APOE *Genotyping (Polymorphic DNA Technologies, Inc)*

Apolipoprotein E (*APOE*) genotyping was performed at polymorphic DNA Technologies, (Alameda, CA). Two Single nucleotide polymorphisms SNPs (dbSNP ID rs429358 and rs7412) within the gene *APOE* were genotyped by the method of amplicon sequencing using "nested" PCR followed by automated Sanger sequencing. For each SNP, human genomic DNA samples were first PCR-amplified with a pair of "outside" oligonucleotide primers, and those amplification products were then reamplified with a second pair of "inside" primers. Final PCR products were cleaned to remove small products and then sequenced using ABI Big Dye chemistry with detection and analysis on an ABI3730xl DNA sequencer. Forward and reverse electropherograms were processed by sequence analysis software that made alignments against the known reference sequences, and the genotype calls were taken from the base-calls at the known SNP sites.

(B) TOMM40'523 poly-T Genotyping, method 1 (Polymorphic DNA Technologies, Inc)

A variable length poly-T (dbSNP ID rs10524523) within the gene TOMM40 was genotyped [45]. The method exploits the fact that during cycle-sequencing reactions of short amplicons, Taq DNA polymerase adds fluorescent 3′-deoxyadenosine tags to full-length primer extension products that can be detected and used as a measure of total amplicon length. Two different assays (Assay 1 and Assay 2) were used, each using a different set of primers. A nested PCR was performed for each assay using the same conditions described above for *APOE* genotyping. After the second PCR step, an internal standard (containing an amplicon with the artificial allele poly$(T)_8$) was added to the product. A Sanger sequencing reaction was performed for each assay using only the forward primer as the sequencing primer, and fluorescent products were detected with an ABI 3730xl DNA sequencer. Because of PCR "stutter" of the poly-T sequence, a complex pattern of terminal A-peaks was seen. Proprietary software was used to identify and assign allele lengths to the A-peaks in the complex pattern, and to deduce the allele lengths present in the original genomic sample. Final genotypes were reported as the allele averages of the two assays.

(C) TOMM40'523 poly-T length analysis, Method 2 (Dr Dmitry Goldgaber)

To analyze the precise length of the poly-T stretch of the rs10524523 polymorphic site of the TOMM40 gene PCR was performed using 5′- fluorescently labeled forward primer ACCTCAAGCTGTCCTCTTGC and unlabeled reverse primer GAGGCTGAGAAGGGAGGATT each at 0.5 μM concentration. The coordinates of the amplified fragment were chr19:50094819 + 50094997 (NCBI36/hg18 human genome assembly). The PCR reaction was performed in 20μl volumes with FailSafe PCR kit (Epicentre Biotechnologies, Madison, WI) using buffer G. The PCR reaction was carried under the following conditions: 3 minutes at 95°C, 40 seconds at 98°C, then 10 cycles with 30 seconds at 95°C, 90 seconds at 63°C, then 18 cycles with 30 seconds at 95°C, 1 minute at 61°C, then final extension 7 minutes at 72°C. One microliter of each PCR product was diluted with 15 μl of HiDi Formamide (Applied Biosystems, Foster City, CA) and supplemented with 0.3 μl of 600LIZ size standard (Applied Biosystems, Foster City, CA). The size of PCR products was determined by electrophoresis using an ABI 3730 DNA analyzer (Applied Biosystems, Foster City, CA) at the Cornell University Genomics Core Laboratory (Ithaca, NY). The output data from the ABI 3730 were analyzed using GeneMarker software (SoftGenetics, State College, PA). The presence of the poly-T stretch caused "stuttering" of DNA polymerase during PCR resulting in generation of PCR products of various lengths that appeared as a staggered peak [45]. To interpret these complex peak patterns we generated DNA fragments with known number of T residues by subcloning individual rs10524523-containing DNA fragments that were amplified from several human genomic DNA samples. Each resulting plasmid DNA was Sanger sequenced and the length of poly-T was precisely determined. Thus a series of individual plasmid DNA samples with known number of T residues in the poly-T stretch was generated. Plasmid DNA was PCR amplified along with tested human DNA samples in each analysis. The resulting peaks from plasmid DNA was compared with peaks generated from human DNA samples.

*(D) Phased Sequencing of a 9.5 kb TOMM40/*APOE *genomic region (Polymorphic DNA Technologies, Inc), Method 3*

For selected genomic samples in this study, the haplotype phases between the *APOE* SNPs and the TOMM40'523 poly-T were determined by the method of long-range PCR, DNA cloning into plasmids, and Sanger sequencing. Long-range PCR was performed using Takara LA Taq Polymerase (Takara Mirus Bio, Mountain View, CA) and very long cycle times to create a 9.5 kb amplicon that contained both the TOMM40 poly-T region and the *APOE* SNPs rs429358 and rs7412. This amplicon was purified by agarose gel electrophoresis and cloned into a TA cloning vector (TOPO XL PCR, Life Technologies, Inc.). The vector was then used to

transform electro-competent cells using electroelution. Cells were plated with antibiotic selection, incubated, and 12 colonies were picked from each plate and cultured. Diluted cultures were transferred to a denaturing buffer and were prepped using a TempliPhi DNA Sequencing Template Amplification kit (GE HealthCare/Amersham Biosciences, Pittsburgh, PA). For each original sample, twelve plasmid templates were then used in Sanger sequencing reactions using sequencing primers having sites within the cloned region. Because only the phases of the poly-T with the *APOE* SNPs were needed, only the 800 bp at each end of the 9.5 kb insert were sequenced. A sequencing alignment report was made for each clone of each sample. For each sample, two sequencing patterns were seen, with some clones giving the sequence of one haplotype and the other clones giving the sequence of the other haplotype. The consensus sequences for each haplotype were used directly to report the phases of the poly-T alleles with the *APOE* alleles.