European Journal of Human Genetics (2014) 22, 1040–1045

www.nature.com/ejhg

EJHG Open

ARTICLE

Variation and association to diabetes in 2000 full mtDNA sequences mined from an exome study in a Danish population

Shengting Li[1,2,8], Soren Besenbacher[1,8], Yingrui Li[2], Karsten Kristiansen[3], Niels Grarup[4], Anders Albrechtsen[3], Thomas Sparsø[4], Thorfinn Korneliussen[3], Torben Hansen[4], Jun Wang[2], Rasmus Nielsen[4,5], Oluf Pedersen[4], Lars Bolund[2,6] and Mikkel H Schierup*[,1,7]

In this paper, we mine full mtDNA sequences from an exome capture data set of 2000 Danes, showing that it is possible to get high-quality full-genome sequences of the mitochondrion from this resource. The sample includes 1000 individuals with type 2 diabetes and 1000 controls. We characterise the variation found in the mtDNA sequence in Danes and relate the variation to diabetes risk as well as to several blood phenotypes of the controls but find no significant associations. We report 2025 polymorphisms, of which 393 have not been reported previously. These 393 mutations are both very rare and estimated to be caused by very recent mutations but individuals with type 2 diabetes do not possess more of these variants. Population genetics analysis using Bayesian skyline plot shows a recent history of rapid population growth in the Danish population in accordance with the fact that >40% of variable sites are observed as singletons.
European Journal of Human Genetics (2014) **22,** 1040–1045; doi:10.1038/ejhg.2013.282; published online 22 January 2014

Keywords: diabetes; mtDNA; population history

INTRODUCTION

The mitochondrion is the energy engine of the cell and possible associations of mitochondrial function and metabolic disorders have therefore been sought.[1,2] However, whether such association are causative or an effect of the cell environment is difficult to disentangle. The mitochondrion is genetically highly variable and is inherited maternally as a single, non-recombining unit. Classification of variation is therefore done as haplogroups that fit into a mitochondrial tree with a root estimated around 200 000 years back. Several attempts of association studies of mtDNA variation and diabetes have been performed with different results in different human populations. All of these are based on typing of a subset of mitochondrial SNPs, defining major haplogroups. An early study reported evidence for association of a common variant and type 2 diabetes[3] in a British population but later, and larger studies, have failed to replicate this finding.[4,5] However, associations of mtDNA variants in other populations are still reported,[6,7] and it is speculated whether there is an indirect effect of mtDNA variation.[8] Therefore, the common mtDNA variation is also included in the new metabochip, which will be used in large-scale association studies of metabolic disorders.[9]

Next generation sequencing now offers cheap sequencing of the complete mitochondrion. Complete sequencing allows higher resolution inference of the demographic history of the population because sequencing also identifies the rare (and recurrent), and more recent mutations. This has been exploited to investigate demographics from samples of 100–200 mitochondria.[10,11] However, for use in association mapping, thousands of mitochondria need to be sequenced. Recently, Picardi and Pesole[12,11] demonstrated how complete mtDNA can be gleaned from exome sequencing studies, not because the mitochondrion is captured on exome chips but because there are 100–1000 times as many copies of the mtDNA as of any given nuclear sequence, thus there will be many mtDNA sequences among the off-target sequences. However, they also warn that nuclear copies derived from the mitochondrion (NUMTs) should be filtered before any inference is made.

Here, we take advantage of a large effort to sequence 2000 exomes (1000 cases and 1000 controls, for details, see Albrechtsen *et al.*[13]) in order to find new variants associated with type 2 diabetes in a Danish population sample. We mine the complete mtDNA sequwnces from the off-target sequences of this effort. We show that by mapping the exome sequences to the mitochondrial reference sequence, subtracting sequences also mapping to NUMTs, we can get an average coverage of $\sim 25\times$ with a Q20 quality threshold. In our final data set of 2000 mitochondrial sequences, we have $<1\%$ missing data. We use this data to infer 393 novel mutations, infer widespread heteroplasmy, make a detailed inference of the recent population history of the Danish population and we associate all variation with case/control

[1]Bioinformatics Research Centre, Aarhus University, C.F. Mollers Alle, 8000 Aarhus C, Denmark; [2]BGI-Shenzhen, Shenzhen, China; [3]Department of Biology, University of Copenhagen, Copenhagen, Denmark; [4]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [5]Department of Integrative Biology, University of California, Berkeley, USA; [6]Department of Bioscience, Aarhus University, Aarhus, Denmark; [7]Department of Bioscience, Aarhus University, Ny Munkegade, 8000 Aarhus C., Denmark
[8]These authors contributed equally to this work.
*Correspondence: Dr MH Schierup, Bioinformatics Research Centre, Aarhus University, CF Mollers Alle 8, DK-8000 Aarhus C., Denmark. Tel: +45 8715 6535; Fax: +45 8715 4102; E-mail:mheide@birc.au.dk
Received 31 January 2013; revised 22 August 2013; accepted 10 October 2013; published online 22 January 2014

status as well as with quantitative traits associated to metabolic disorders, but measured in the controls.

MATERIALS AND METHODS

Study populations

The 2000 study individuals are all from a Danish cohort described in detail by Albrechtsen *et al.*[13] Of these, 1000 were cases recruited based on presence of type 2 diabetes, BMI $>27.5\,kg/m^2$ and hypertension (blood pressure (BP) above 140/90 mmHg or use of anti-hypertensive medication) and 1000 were control individuals recruited from two Danish population-based cohorts and all had fasting plasma glucose <5.6 mmol/l, 2 h post-OGTT plasma glucose <7.8 mmol/l, BMI $<27.5\,kg/m^2$ and BP $<140/90$ mmHg (see also Albrechtsen *et al.*[13]).

Mining of mtDNA sequences

MtDNA sequences were mined from exome capture and sequencing data of the 2000 individuals. Exome capturing and sequencing were done twice on these individuals, initially at relatively low coverage of $\sim 8\times$ as reported in Albrechtsen *et al.*[13] Subsequently, to attain higher exon coverage, exome capture was performed on the same 2000 individual using the Agilent SureSelect Human All Exon Kit. A total 46 Mb of the genome consisting of 26 696 CCDS was targeted. The DNA was sequenced using Illumina pair-end sequencing resulting in an average sequencing depth of $56\times$ in the target regions (unpublished data).

Since most cell types contain many copies of the mitochondrion, mtDNA sequences are represented many more times than nuclear DNA. As a result a substantial amount of mtDNA is sequenced even though none of the mitochondrial genes were targeted by the exome capturing. Since the second exome capture experiment was more specific, a smaller proportion of sequences were of mitochondrial origin, so even though the exomes were covered seven times greater in the second experiment, it only yielded an average of $18.2\times$ of mtDNA sequences where the first experiment yielded $7.5\times$. In the present analysis, we have pooled mtDNA sequences gleaned from these two exome sequencing experiments after assuring that they do not contain conflicting evidence (see below).

A possible source of error when retrieving sequences from the mitochondrion is the presence of NUMTs (nuclear mitochondrial DNA sequences), so we have paid special attention to these elements to avoid errors. First we retrieved a list of known NUMTs from http://www.ianlogan.co.uk/numts/numt_chrs.html. By aligning (by blast) the assembled results of the exomes to these NUMT sequences, we then found that there is no overlap between these two data sets. So we consider that there are no NUMT sequences that are amplified on the exome chip. Then, the copies of NUMT sequences are supposed to be far less than the copies of mtDNA sequences ($<1\%$), which means that the NUMTs can be treated as a special kind of error (see below).

All sequences from the exome capture (an average of 30 million 101 base pair sequences per sample) were used.

When calling variants the following strategy was employed. First, we prepared a small database with the circular chrM (by appending the first 120 bp to the end) + NUMTs sequences. Then for each mitochondrion, the sequence was called assuming haploidy using the following steps:

(1) Extract the reads that align to the revised Cambridge reference sequence (http://www.ncbi.nlm.nih.gov/nuccore/NC_012920).
(2) Realign the reads set to the MT + NUMTs database as single-end reads by BWA[14] with the option '-n 10000' (to make sure the result shows all possible hits).
(3) Scan all aligned reads and store only the bases with quality score >20.
(4) Mark the bases with 'in_NUMTs' if one of the reads that covers it also maps to a NUMT sequence.
(5) Call the sequence position by position. For each position, the allele base with highest coverage is denoted as bp1, its coverage is cov1. If there exist alternative alleles, the base with second highest coverage is denoted as bp2, its coverage is cov2. We set the called base to 'N' if it satisfies one of the following conditions:

(a) cov1 = = cov2
(b) in_NUMTs and cov1 = = cov2 + 1
(c) cov1 < 3
(d) in_NUMTs and cov1 < 4

Otherwise, we call this base as bp1.

The average coverage per sample and per base pair across samples was calculated from the high quality reads.

Heteroplasmy

We scored a given position in an individual as heteroplasmic if at least three high-quality reads ($>$Q20) supported each of two different base pairs.

Quantitative measurements on controls

In the analysis of quantitative metabolic traits, only the 1000 control individuals were used. The following quantitative measurements were used:

- pglu0, fasting plasma glucose
- pglu30, 30 min glucose post oral glucose load
- pglu120, 120 min glucose post oral glucose load
- insu0, fasting serum insulin
- insu30, 30 min insulin post oral glucose load
- insu120, 120 min insulin post oral glucose load
- Xinsulin, the insulinogenic index (a beta cell measure)
- homair, insulin resistance measure
- BIG_AIR, measure of insulin release from betacells
- BIG_SI, measure of insulin resistance
- trig, fasting serum triglyceride
- chol, fasting serum cholesterole
- hdlc, high density lipoprotein cholesterole
- bmi, body mass index.

Assignment of haplogroups to sequences

Haplogroups were assigned to each sequence using the following approach.

For each full mtDNA sequence, we calculated the weighted edit distance to each node in the Build 13 from http://phylotree.org/, weighting variants according to the number of times they have occurred in the tree (ie a variant which has occurred n times is given a weight of $1/n$). The node with the minimal distance is then the haplogroup of the sequence. If two groups have the exact same distance, then the most recent common ancestor of the two groups is used. Thus, some sequences are assigned to major haplogroups only, whereas some sequences can be assigned to sub haplogroups. For the results of the full assignment, see Supplementary Table S1.

Statistical analyses of association

Association was tested both at the level of haplogroups and at the level of individual variants. For testing of cases versus controls we used Fisher's exact test of independence in a 2×2 Table for each variant. For testing association of haplogroups we used a Fisher's exact test on all internal nodes in the mitochondrial tree based on the haplogroup assignment explained above. Sequences that are not assigned to a haplogroup sub-group, are only included in tests that are at the level they belong to or above.

For testing of association with quantitative traits, we used a rank test of traits values that first order them according to their value irrespective of the genetic variant at the given position (or the haplogroup assignment). The values are then transformed into a normal distribution and a linear model is used to test for the effect of the variant (or haplogroup) on the trait value, using sex as a covariate. In cases where sex is not significant, we report results from the model without this covariate.

Detection of new variants

We use the variation inventory of mitomap.org to determine which of our observed variants have not been reported previously. All these variants were annotated using SNPeff and are shown in Supplementary Table S2.

For investigation of enrichment of novel variants in individual genes in cases we focused on variants predicted to either affect RNA structure or being non-synonymous. For each gene, we counted the number of cases and controls with such variants and tested differences using Fisher's exact test.

Skyline plot of demography

BEAST (v1.7.2) with BEAGLE (v1.0) was used to infer a Bayesian Skyline Plot of demographic inference.[15,16] For this analysis, we removed the control region (leaving the positions 577–16 023 bp).

RESULTS

Mining mtDNA from exome data

From an average $>50\times$ exome sequencing, we managed to get an average of $25.7\times$ coverage of the complete mitochondrion. The coverage and detected variants for all 2000 full mtDNA sequences are shown in Supplementary Table S1 and the full sequences are also deposited in GenBank, accession numbers KF161060–KF163059. The coverage along the mtDNA sequence when pooling the coverage among all samples is not even (Figure 1a) but most regions are covered by at least an average of $10\times$. With the criteria outlined for calling sequence we get ~99% complete mtDNA sequence and the fall out is very limited along the sequence except in a couple of positions (Figure 1b). We used the two independent capture experiments to evaluate accuracy by comparing concordance and found that we had an average discordance of 0.62 bps/sample, with the first experiment most often suggesting the reference base pair and the second capture experiment the alternative base pair. We believe that this difference is due to the higher coverage in the second experiment, which allows us to call the alternative variant in more of the heteroplasmic sites, and therefore that most of the inconsistencies are due to heteroplasmy. We conclude that complete mtDNA sequences can be gleaned from Agilent captured exomes as also found for a limited number of mitochondria.[12]

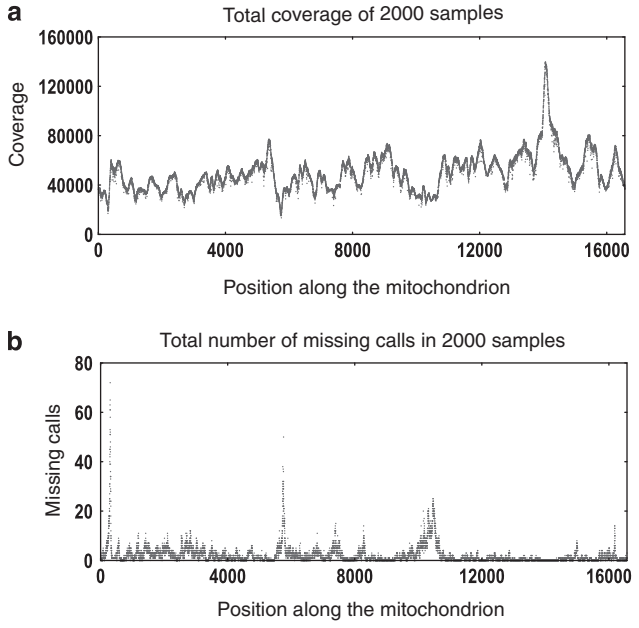

Figure 1 (**a**) The total coverage of mitochondrial sequence from 2000 samples over the length of the mitochondrion. (**b**) The number of mitochondrial sequences with missing data (out of 2000) along the length of the mitochondrion.

The amount of genetic variation

Table 1 shows the number of polymorphisms in the full set of mitochondria for each gene and the control region separately. A total of 2025 polymorphisms are found with 1920 transitions and 166 transversions, corresponding to a ts/tv = 23.1, which is in line with other studies.[10,11] This indicates that sequencing errors are not widespread since random sequencing errors should lower the ration of transitions to transversions. Furthermore, positions with variation have about the same coverage (23.2) as positions without variation (25.0), suggesting that low coverage regions are not causing false SNP calls. The selective constraint of the different genes is highly variable as indicated by different dn/ds values, in line with previous observations. The proportion of polymorphic sites (Table 1, last column) is similar among protein coding genes and lower for RNA coding genes, which are under more selective constraint. The site frequency spectrum (SFS, Table 2 and Figure 2) shows a large proportion of singleton (~40%) variation in line with recent population growth (see below). There is only a slight enrichment of singleton non-synonymous variation over synonymous variation, suggesting that purifying selection on the mitochondrial lineages recently hit by non-synonymous mutations is weak. This supports that singleton variation identified is not associated with a greater proportion of false positives since this should increase the fraction of rare non-synonymous variants. The SFS for RNA mutations appear more skewed suggesting stronger selection against these and the three frameshift mutations observed are all singletons suggesting strong selection. Intergenic variants have an SFS with a much smaller fraction of rare variants, likely caused by recurrent mutations on these, generally, hypervariable sites.

The amount of detected heteroplasmy is high with >4000 sites showing heteroplasmy in at least one individual and 65 sites showing heteroplasmy in >10 individuals (Supplementary Table S3). A detailed investigation of heteroplasmy will be reported elsewhere.

Among the variation observed, a total of 393 variants, of which 68 are found in RNA genes, 125 are non-synonymous variants and 177 are synonymous variants have not previously been recorded in mitomap. All of this newly identified variation is rare, with 72.0% being singletons, and 17.1% doubletons (Table 2, last column) and most of these mutations are likely to have occurred recently in the Danish population. The new variants have a threefold lower ts/tv than the

Table 1 Overview of variation in the sample of Danish mtDNA

Region/gene	*Var no.*	*Transitions*	*Transversions*	*Non-syn*	*Syn*	*pN/pS*	*% of polymorphism*
Control region	304	289	45				27.1
Other non-coding	27	21	7				
12S rRNA	68	62	9				6.7
16S rRNA	94	90	5				6.0
tRNAs	117	116	2				7.8
MT-ATP6	128	121	8	70	58	0.52	18.8
MT-ATP8	27	26	1	14	13	0.46	13.0
MT-CO1	164	154	15	40	124	0.14	10.4
MT-CO2	80	74	8	23	57	0.17	11.7
MT-CO3	97	93	5	34	63	0.23	12.4
MT-CYB	155	144	14	59	96	0.26	13.6
MT-ND1	121	114	9	37	84	0.19	12.7
MT-ND2	131	126	7	36	95	0.16	12.5
MT-ND3	39	37	2	13	26	0.21	11.3
MT-ND4	147	140	8	30	117	0.11	10.7
MT-ND4L	30	30	2	6	24	0.11	9.8
MT-ND5	221	209	15	59	162	0.16	12.1
MT-ND6	75	74	4	30	46	0.28	14.3
Total	2025	1920	166	451	965	0.20	

Table 2 The site frequency of variants divided into coding variants (synonymous and non-synonymous), variants in RNA genes, intergenic variants and new variants

Site frequency	Non-synonymous		Synonymous		RNA		D-loop		New variants	
	No.	%	No.	%	No.	%	No.	%	No.	%
1	193	42.8	399	41.3	128	45.9	73	24.0	286	72.0
2	71	15.7	163	16.9	47	16.8	27	8.9	68	17.1
3	41	9.1	73	7.6	20	7.2	28	9.2	18	4.5
4	26	5.8	55	5.7	15	5.4	15	4.9	9	2.3
5	21	4.7	40	4.1	12	4.3	10	3.3	6	1.5
6–10	37	8.2	111	11.5	20	7.2	40	13.2	5	1.3
11–20	28	6.2	38	3.9	7	2.5	28	9.2	1	0.3
21–100	22	4.9	66	6.8	17	6.1	55	18.1	0	0.0
100–2000	12	2.7	20	2.1	13	4.7	28	9.2	0	0.0

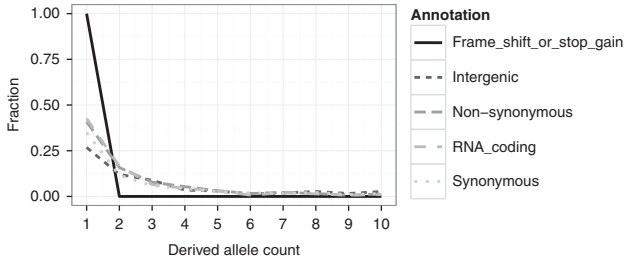

Figure 2 The low range of the folded SFS for the variation observed, divided into synonymous, non-synonymous, RNA coding, intergenic and non-sense. Variants observed up to 10 times are shown.

variants recorded previously but this is in accordance with the fact that >60% of all possible synonymous transitions have already been seen previously. Table 3 shows the number of potentially functional new variation (ie all new RNA variants and non-synonymous coding variants) found in each gene separately. The number of cases and controls harbouring these new mutations is not different suggesting that new mutations are not enriched in cases. The only weak indication is for the MT-TW tRNA (tryptophane) where only cases harbour new variants.

Finally, among the variants recorded we found a few cases of mutations known to cause syndromes (as reported by mitomap) (Table 4). All of these are supported by at least 10 reads and none of them shows any sign of heteroplasmy. There is a statistically non-significant indication of enrichment of these in the case group. The total frequency of pathogenic mutations is in accordance with the expected population frequency in a previous report by Elliot *et al.*[17]

Patterns of mtDNA variation in Denmark

We assigned 2000 mtDNA sequences to a haplogroup using the phylotree.org[18] classification and the assignment algorithm outlined in the Methods section. The phylogenetic relationships between the main observed haplogroups is shown in Figure 3. Figure 4 shows the counts of the main haplogroups divided into cases and controls. No large differences are observed between cases and controls.

Population history (Bayesian skyline)

The Bayesian skyline inference of population size history is shown in Figure 5, assuming a mutation rate of 1.7×10^{-8} per year. An increase in the effective population size by two orders of magnitude since the last ice age is evident. The large sample size of the present study allows more accurate inference of the very recent past (last 1000 years) and this is likely the cause that a greater increase in population size is estimated than in other European studies with smaller sample sizes. Other recent studies using sequencing of nuclear genes in thousands of individuals also report a plethora of variation and an extreme recent rise in effective population sizes.[19]

Table 3 The number of potentially functional variants (non-synonymous or changing an RNA gene) not previously recorded in mitomap, divided into genes and with the number of cases and controls having such variants

Gene	Odds ratio	No. of new variants	In cases	In controls	P-value
MT-TW	NA	3	6	0	0.031
MT-CO1	0.35	11	5	14	0.062
MT-ATP6	3.02	12	12	4	0.076
MT-RNR2	1.43	35	34	24	0.230
MT-ND5	0.66	28	16	24	0.263
MT-ND1	0.58	12	7	12	0.357
MT-TH	4.01	4	4	1	0.374
MT-TT	0.50	6	4	8	0.386
MT-CO2	1.67	14	10	6	0.453
MT-TF	0.00	2	0	2	0.500
MT-TL2	0.00	2	0	2	0.500
MT-RNR1	1.45	18	13	9	0.521
MT-CO3	0.83	14	15	18	0.726
MT-TS1	0.78	4	7	9	0.803
MT-ND6	1.29	11	9	7	0.803

P-value is Fisher's exact test of independence of case/control status.

Table 4 Variants found in the present data set that have been confirmed (according to mitomap.org) to be involved in diseases, with number of cases and controls harbouring them.

Variant	No. of Cases	No. of Controls	Phenotype	Number of reads supporting variant/total reads in individuals
A1555G	3	4	DEAF	102/107
G11778A	4	0	Progressive dystonia, LHON	22/22
T14484C	1	0	LHON	49/49
T14674C	2	0	Reversible COX deficiency myopathy	47/47

No association with case/control status or metabolic traits

There are no clear cases of association of single SNPs or haplogroups with either case/control status or with any of the quantitative measures. The strongest associations for the different tests are shown in Supplementary Tables 4–7, but none of the variants approach significance when controlling for multiple testing using a permutation test.

DISCUSSION

In the present study, we show how a very large set of complete mtDNA sequences can be gleaned from a high coverage exome study. This should be feasible to repeat for many other exome studies in other populations. The resulting mtDNA sequences increases the number of sequenced mitochondria in the Danish population more

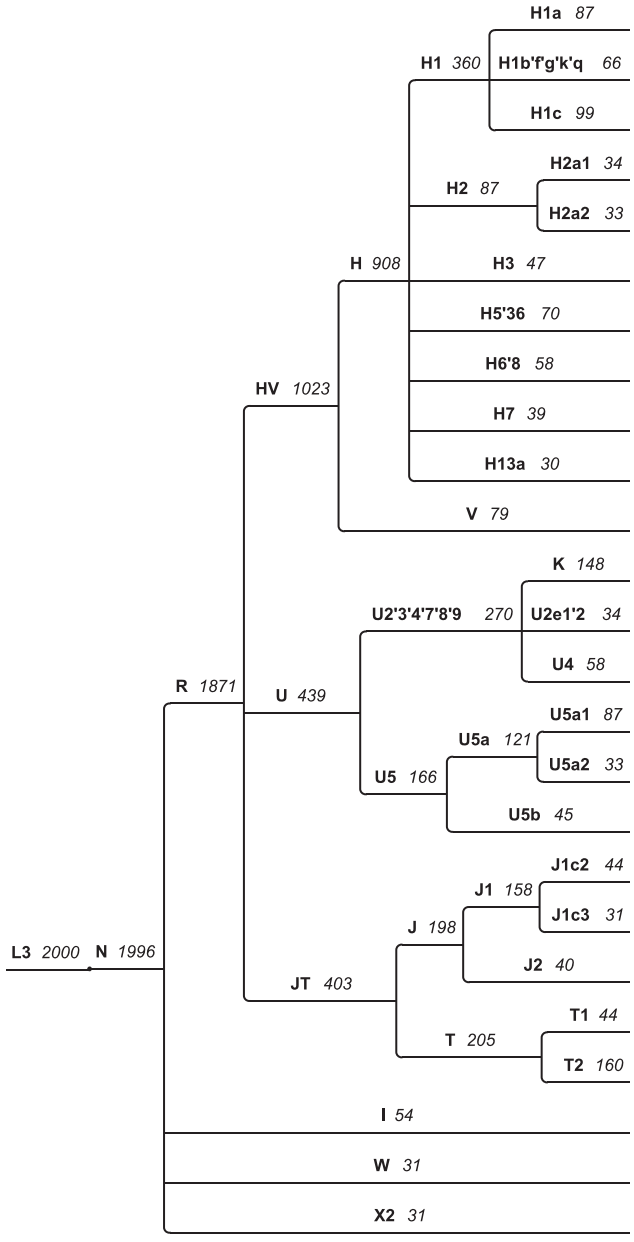

Figure 3 The phylogenetic tree relating the haplogroups with >30 representatives among the 2000 mtDNA sequences sampled in the current study. Haplogroup designation and number of samples are shown, for internal branches and the tips.

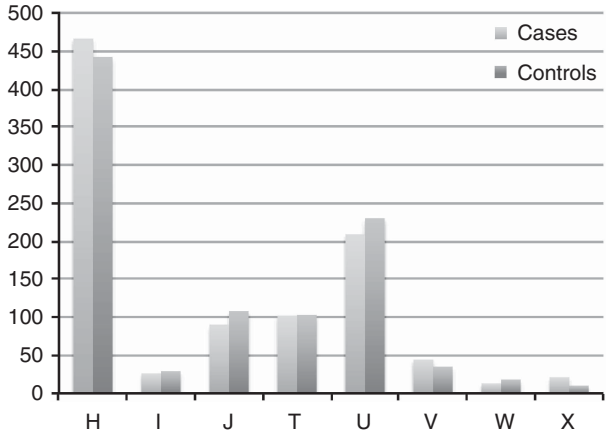

Figure 4 The haplogroup distribution in cases and controls. None of the differences are significant with Bonferroni correction for multiple testing.

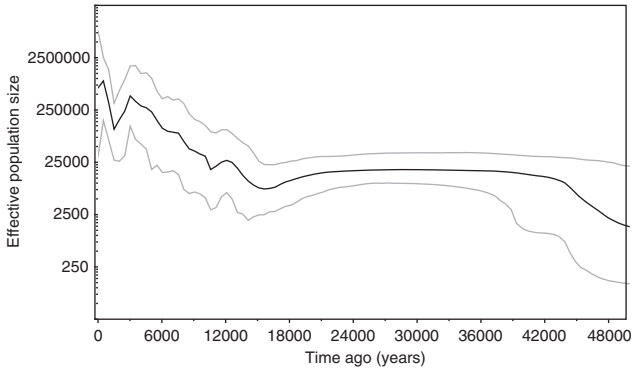

Figure 5 Bayesian skyline plot of the 2000 sequences. A mutation rate of 1.7×10^{-8} per year was used to convert substitution rates into years (*x*-axis) and coalescent intensities into effective population sizes (*y*-axis).

than 10-fold and give a precise picture of the genetic variation segregating, both in terms of haplogroup frequencies and in terms of new and rare variation. This in turn allowed a very detailed estimation of rapid recent population growth in the Danish population. Even though many of the haplogroup defining mutations are found in high frequency, almost half of the variants are singletons and 393 (19%) of the variants have not been reported in mitomaps catalogue of variation based on >8000 fully sequenced mitochondria. This shows that as for nuclear genes the recent European population growth has resulted in a large set of new variants within the last tens of generations[19,20] and many of these variants will be specific to the Danish population. In contrast to Nelson *et al.*[19] there is only a slight enrichment of non-synonymous variants among the very rare variants (here the 40% singletons). This suggests that in the recent history of the Danish population there has been a limited number of slightly deleterious non-synonymous mutations in the mitochondrion and a relative large set of strongly deleterious non-synonymous variants that never reach a frequency so they can be found in a set of 2000 mitochondria. This appears in conflict with the report by Subramanian[21] of a high fraction of non-synonymous variation being slightly deleterious based on a decreasing dn/ds ratio over time. It is conceivable that the recent explosion in population size makes natural selection so inefficient that we observe only a minor skew of the site frequency distribution, ie, no lineages are presently dying out due to selection.

Like previous studies looking for associations between common mtDNA sequence variation and diabetes risk,[4] we see no significant association results for the common SNPs and the major mitochondrial haplogroups. Since we have the full mitochondrial genome sequence and not just SNP data we have also been able to test for an effect of rare mtDNA sequence variants on diabetes. This is interesting since several studies have reported rare mutations in mitochondrial tRNA genes that cause maternally inherited diabetes with deafness (MIDD)(OMIM #520000). Furthermore, a recent study used transmitochondrial mice to show that a rare mutation in the *MT-ND6* gene regulates diabetes development in mice.[22] None of the previously reported MIDD mutations were present in our data and since we have a sample of unrelated cases it is unlikely that one

variant that is very rare in controls should be present in many cases and thus significant. So in addition to the normal single marker tests, we have also implemented a gene-level test that groups all carriers of previously unreported and likely functional variants in a given gene. The most significant gene in this analysis (MT-TW) had a *P*-value of 0.03, which is not significant when the number of tests is taken into account.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We acknowledge the Lundbeck Foundation for funding through the LUCAMP centre.

1 Lowell BB, Shulman GI: Mitochondrial dysfunction and type 2 diabetes. *Science* 2005; **307**: 384–387.
2 Patti ME, Corvera S: The role of mitochondria in the pathogenesis of type 2 diabetes. *Endocr Rev* 2010; **31**: 364–395.
3 Poulton J, Luan J, Macaulay V, Hennings S, Mitchell J, Wareham NJ: Type 2 diabetes is associated with a common mitochondrial variant: evidence from a population-based case–control study. *Hum Mol Genet* 2002; **11**: 1581–1583.
4 Saxena R, de Bakker PI, Singer K *et al*: Comprehensive association testing of common mitochondrial DNA variation in metabolic disease. *Am J Hum Genet* 2006; **79**: 54–61.
5 Chinnery PF, Mowbray C, Patel SK *et al*: Mitochondrial DNA haplogroups and type 2 diabetes: a study of 897 cases and 1010 controls. *J Med Genet* 2007; **44**: e80.
6 Yang TL, Guo Y, Shen H *et al*: Genetic Association Study of Common Mitochondrial Variants on Body Fat Mass. *PLoS One* 2011; **6**: e21595.
7 Liou CW, Chen JB, Tiao MM *et al*: Mitochondrial DNA coding and control region variants as genetic risk factors for type 2 diabetes mellitus. *Diabetes* 2012; **61**: 2642–2651.
8 Achilli A, Olivieri A, Pala M *et al*: Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole. *PLoS One* 2011; **6**: e21029.
9 Voight BF, Kang HM, Ding J *et al*: The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 2012; **8**: e1002793.
10 Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M: High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res* 2011; **21**: 1–11.
11 Schonberg A, Theunert C, Li M, Stoneking M, Nasidze I: High-throughput sequencing of complete human mtDNA genomes from the Caucasus and West Asia: high diversity and demographic inferences. *Eur J Hum Genet* 2011; **19**: 988–994.
12 Picardi E, Pesole G: Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods* 2012; **9**: 523–525.
13 Albrechtsen A, Grarup N, Li Y *et al*: Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* 2013; **56**: 298–310.
14 Li H, Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010; **26**: 589–595.
15 Drummond AJ, Rambaut A, Shapiro B, Pybus OG: Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 2005; **22**: 1185–1192.
16 Minin VN, Bloomquist EW, Suchard MA: Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* 2008; **25**: 1459–1471.
17 Elliott H, Samuels D, Eden J, Relton C, Chinnery P: Pathogenic mitochondrial DNA mutations are common in the general population. *Am J Hum Genet* 2008; **83**: 254–260.
18 van Oven M, Kayser M: Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 2009; **30**: E386–E394.
19 Nelson MR, Wegmann D, Ehm MG *et al*: An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 2012; **337**: 100–104.
20 Fu W, O'Connor TD, Jun G *et al*: Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2013; **493**: 216–220.
21 Subramanian S: Temporal trails of natural selection in human mitogenomes. *Mol Biol Evol* 2009; **26**: 715–717.
22 Hashizume O, Shimizu A, Yokota M *et al*: Specific mitochondrial DNA mutation in mice regulates diabetes and lymphoma development. *Proc Natl Acad Sci USA* 2012; **109**: 10528–10533.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)