Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity

Christina L Usher1, Robert E Handsaker1–3, Tönü Esko1,2,4–6, Marcus A Tuke7, Michael N Weedon7, Alex R Hastie8, Han Cao8, Jennifer E Moon1,2,4,5, Seva Kashin2,3, Christian Fuchsberger9,10, Andres Metspalu6,11, Carlos N Pato12, Michele T Pato12, Mark I McCarthy13–15, Michael Boehnke9,10, David M Altshuler1,2,16, Timothy M Frayling7, Joel N Hirschhorn1,2,4,5 & Steven A McCarroll1–3

Hundreds of genes reside in structurally complex, poorly understood regions of the human genome1–3. One such region contains the three amylase genes (AMY2B, AMY2A and AMY1) responsible for digesting starch into sugar. Copy number of AMY1 is reported to be the largest genomic influence on obesity4, although genome-wide association studies for obesity have found this locus unremarkable. Using whole-genome sequence analysis5,6, droplet digital PCR6 and genome mapping7, we identified eight common structural haplotypes of the amylase locus that suggest its mutational history. We found that the AMY1 copy number in an individual’s genome is generally even (rather than odd) and partially correlates with nearby SNPs, which do not associate with body mass index (BMI). We measured amylase gene copy number in 1,000 obese or lean Estonians and in 2 other cohorts totaling ~3,500 individuals. We had 99% power to detect the lower bound of the reported effects on BMI4, yet found no association.

Like hundreds of human genes, the amylase genes reside in a structurally complex locus, one with inversions, deletions and duplications8. Each of the three amylase genes, which encode enzymes that digest starch into sugar, varies widely in copy number, with AMY1 varying from 2–17 copies9,10, AMY2A varying from 0–8 copies10 and AMY2B varying from 2–6 copies. Given the role of these genes in starch metabolism and the greater average copy number of AMY1 in three populations with high-starch diets9, it has been hypothesized that AMY1 copy number shapes the metabolic response to diet. A recent study reported that each copy of AMY1 decreases the risk of obesity by 1.2-fold4, potentially a profound effect as AMY1 copy number varies so widely (2–17 copies; s.d. of 2.4 copies). The effect of AMY1 copy number, reported to explain 11% of the genetic contribution to obesity (far greater than the effect of SNPs at FTO, the largest effect detected in genome-wide association studies (GWAS)4), was not detected in a GWAS of 339,224 people11. The discordance between these results raises questions about the completeness of GWAS and other genome-scale approaches in human genetics4. However, complex copy number variations (CNVs) are notoriously difficult to measure12, and CNV association studies often involve rough copy number estimates, which can be confounded by technical factors that would be readily detected in molecularly precise data13–15. We sought to understand these issues at the amylase locus.

We first measured the copy number of the three amylase genes in two large, partially overlapping cohorts. We analyzed whole-genome sequence data from 569 individuals from Phase 1 of the 1000 Genomes Project16 using our Genome STrip5,2 algorithm (Supplementary Tables 1 and 2). We also measured copy number in 114 parent-offspring trios from HapMap17 using droplet digital PCR (ddPCR) (Supplementary Figs. 1 and 2). These data, which were concordant for overlapping samples (Supplementary Fig. 3), identified two relationships: (i) individuals were four times more likely to have an even (2, 4, 6, etc.) than an odd (1, 3, 5, etc.) number of AMY1 copies (Fig. 1a) and (ii) AMY1 and AMY2A showed parity—the copy numbers of AMY1 and AMY2A were almost always both odd or both even (Fig. 1b). These features have not been observed in studies that used lower-precision molecular methods, such as RT-PCR and array comparative genomic hybridization (CGH), or lower-precision analyses of whole-genome sequencing data to measure copy number2,4,9,18.

If these observations are correct, then they would by necessity arise from an underlying set of structural alleles, only some of which have previously been identified8,9,19. To ascertain the gene content of these
amylase structural alleles, we extended an approach we developed for the 17q21.31 locus, one of the first structurally complex loci to be resolved into structural alleles20,21. We precisely measured and followed the segregation of copy number for each amylase gene in 114 father-mother-offspring trios (from HapMap cohorts of European and West African ancestry), allowing us to assign copy numbers to transmitted and untransmitted chromosomes and thereby to assemble models of the gene contents of each structural allele (Fig. 1c). We further evaluated these inferences by (i) quantifying how many individuals had genotypes that could be explained by a modest number of common haplotypes and (ii) comparing our inferred structural haplotypes to the haplotypes previously identified by fiber FISH and restriction mapping of clones8,9,19.

We found that eight common haplotypes could explain 98% of the combinations of AMY1, AMY2A and AMY2B copy numbers we observed in 480 European-ancestry individuals from the 1000 Genomes Project16. We identified common haplotypes consistent with five of the six previously identified haplotypes8,9,19, along with three new haplotypes, in the European-ancestry (CEU) trios and evidence for additional, rarer haplotypes in the West African (YRI) trios (Fig. 1c, Supplementary Fig. 4 and Supplementary Table 3). Because these analyses do not specify the order of the genes on the structural haplotypes, we used earlier data from cosmid mapping8,19 and fiber FISH9 and performed NanoChannel-based genome mapping analysis7 to predict the order of structural features on these alleles (Fig. 1d and Supplementary Fig. 5).

Figure 1 Copy number distributions of the amylase genes and structural haplotypes of the amylase locus. Whole-genome sequence data from European-ancestry (CEU), Yoruba (YRI), Chinese (CHB) and Japanese (JPT) population samples from the 1000 Genomes Project were analyzed for AMY1, AMY2A and AMY2B copy number using Genome STRIP. (a) The populations have similar distributions of AMY2A and AMY2B copy numbers but a different distribution for AMY1 copy number (KS test, P = 3 × 10−6). For AMY1, even copy numbers greatly outnumber odd copy numbers. (b) In the YRI and CEU population samples, AMY2A and AMY1 have parity in 98% of individuals: when AMY2A copy number is odd, so is AMY1 copy number. JPT and CHB samples had very few odd copy numbers and are not shown. Numbers on the plot represent AMY2B copy number. (c,d) A preliminary map of eight common structural forms of the amylase locus in European-ancestry individuals, derived from the above analysis and with structural features ordered using earlier data from cosmid mapping (c) and genome mapping (d) experiments to analyze the haplotypes (AH2, AH3, AH4, AH2B2 and AH4B2). An asterisk indicates previously unidentified haplotypes (predicted structure), a superscript “a” indicates the chimpanzee haplotype (presumed ancestral) and a superscript “r” indicates the human reference sequence haplotype. Additional structural forms found only in African genomes are described in Supplementary Figure 4.
This set of common haplotypes and their frequencies (Fig. 1c) explained both the predominance of even AMY1 copy numbers in diploid genomes and the sharing of odd or even parity for AMY1 and AMY2A copy numbers. Most European-ancestry chromosomes (89%) contained an odd number of AMY1 copies, which naturally sum to an even number in diploid genomes. In addition, the AMY1-odd haplotypes (those that had an odd number of copies of AMY1) each had one copy of AMY2A, whereas the AMY1-even haplotypes had either zero or two copies of AMY2A (Fig. 1c), resulting in odd AMY2A and AMY1 copy numbers segregating together and explaining the odd/even parity of these genes.

The structural haplotypes (Fig. 1c) also suggest the mutational history of the locus. The more common AMY1-odd haplotypes differ in the copy number of a cassette duplicated in tandem that contains two head-to-head AMY1 genes. We found that these AMY1-odd haplotypes (haplotypes AH1, AH3, AH5 and AH7 in Fig. 1c) segregated on many of the same SNP haplotypes (Fig. 2a and Supplementary Fig. 6), and we identified different historical recombination sites within their intergenic regions (Supplementary Figs. 7 and 8). Frequent non-allelic homologous recombination (NAHR) involving the tandem array could have generated these many structural forms from one another (haplotypes AH1, AH3, AH5 and AH7 in Fig. 1c). In contrast, the haplotypes containing even numbers of AMY1 copies appeared to segregate on distinct SNP haplotype backgrounds, consistent with their having arisen from unique mutational events that involved complex rearrangements by a rarer mutation mechanism (Fig. 1c).

On the basis of these AMY1 structures and their relationships to surrounding SNP haplotypes (Fig. 2a), we hypothesized that individual SNPs near the amylase genes might at least partially correlate with AMY1 copy number within populations. We compared the AMY1 copy number of European-ancestry individuals from the 1000 Genomes Project to their SNP genotypes and found SNPs that had an average difference of 0.6 to 2.0 AMY1 copies per minor allele of the SNP (Table 1). Permutation tests established that these correlations were statistically significant ($\alpha = 6.5 \times 10^{-8}$). The partial correlations between AMY1 copy number and these SNPs replicated in 2 independent cohorts of 768 and 2,807 European-ancestry individuals sampled in the United States and Europe, respectively (Fig. 2b, Supplementary Fig. 9 and Supplementary Tables 4 and 5).

Although each of these SNPs explains only a small fraction of variation in AMY1 copy number, power in GWAS arises from the product of linkage disequilibrium (LD; $r^2$) and sample size; in the GIANT Consortium's meta-analysis of SNP data from 339,224 individuals, a contribution of AMY1 copy number to BMI as strong as that reported would be 99.9% likely to bring about a nominal ($P = 0.05$) association with the more correlated SNPs. However, none of the 17 SNPs in the GIANT meta-analysis reached even nominal ($P = 0.05$) significance, and the SNPs as a group showed no trend toward low association statistics (Table 1 and Supplementary Fig. 10).

Because this lack of evidence for association of AMY1 with BMI is indirect, we conducted our own association analyses by directly measuring the copy numbers of the amylase genes using our high-resolution methods in three European-ancestry cohorts.

We began by analyzing DNA from 1,000 Estonians selected from a broader Estonian Biobank cohort (51,535 individuals) for being in the tails of the BMI distribution—500 individuals with BMI < 22 and 500 individuals with BMI > 33. Among these 1,000 individuals, we observed associations with SNPs that have been associated with BMI in earlier studies, including SNPs in the FTO ($P = 3.5 \times 10^{-7}$), SEC16B ($P = 5.3 \times 10^{-4}$) and MTCHE ($P = 9.6 \times 10^{-3}$) loci, and

### Table 1 Association of SNPs with amylase copy number and BMI in large cohorts

| Gene | SNP | Minor allele freq. | Change in copy number/minor allele | $r^2$ | $P$ value | BMI association $P$ value |
|------|-----|-------------------|-----------------------------------|------|----------|--------------------------|
| AMY1 | rs2443472 | 0.33 | -1.23 | 0.111 | $<10^{-6}$ | 0.09 |
|      | rs11573790 | 0.07 | 2.08 | 0.104 | $<10^{-6}$ | 0.13 |
|      | rs1566154 | 0.19 | 0.90 | 0.044 | $<10^{-6}$ | 0.11 |
|      | rs1930212 | 0.18 | -0.89 | 0.041 | $<10^{-6}$ | 0.74 |
|      | rs108811197 | 0.35 | -0.66 | 0.037 | $<10^{-6}$ | 0.75 |
|      | rs1329597 | 0.03 | -1.39 | 0.036 | $<10^{-6}$ | 0.73 |
|      | rs11185098 | 0.26 | 0.70 | 0.032 | $<10^{-6}$ | 0.80 |
|      | rs1999478 | 0.18 | -0.76 | 0.030 | $<10^{-6}$ | 0.53 |
|      | rs1330403 | 0.14 | 0.82 | 0.029 | $<10^{-6}$ | 0.42 |
|      | rs6696797 | 0.35 | -0.60 | 0.028 | $<10^{-6}$ | 0.63 |
| AMY2B | rs12076610 | 0.11 | 0.80 | 0.582 | $<10^{-6}$ | ND |
|      | rs11185098 | 0.26 | 0.91 | 0.270 | $<10^{-6}$ | ND |
| AMY2A | rs28558115 | 0.11 | 0.90 | 0.398 | $<10^{-6}$ | ND |
|      | rs11185098 | 0.26 | 0.42 | 0.154 | $<10^{-6}$ | 0.80 |

Correlations among amylase copy number and SNP minor alleles are calculated from two cohorts analyzed by whole-genome sequencing, the Genomic Psychiatry Cohort (768 European-ancestry individuals sampled in the United States and GoT2D (2,863 individuals sampled in Europe). BMI association $P$ values are from the GIANT Consortium meta-analysis of 339,224 individuals. Freq., frequency; ND, not determined.

### Table 2 Association analyses of AMY1 copy number and previously BMI-associated SNPs in multiple obesity and BMI cohorts

| Cohort | Sample size | Gene | Variant genotyped | Power* | Odds ratio, obesity* | $P$ value |
|--------|-------------|------|------------------|--------|---------------------|---------|
| Estonian | 1,000 | FTO | rs1558902 | Polygenic | $>0.99$ | 1.64 (1.41, 1.84) | $3.7 \times 10^{-12}$ |
|         |           | AMY1 | Copy number | $>0.99$ | 1.01 (0.96, 1.06) | 0.7     |
| InCHIANTI | 657 | FTO | rs1558902 | Polygenic | $>0.99$ | 1.16 (1.34, 1.93) | $3.5 \times 10^{-7}$ |
|         |           | AMY1 | Copy number | $>0.99$ | 1.16 (1.41, 1.84) | $3.7 \times 10^{-12}$ |
| GoT2D controls | 1,370 | FTO | rs1558902 | Polygenic | $>0.99$ | 1.01 (0.96, 1.06) | 0.7     |
|         |           | AMY1 | Copy number | $>0.99$ | 1.01 (0.96, 1.06) | 0.7     |
| GoT2D cases | 1,437 | FTO | rs1558902 | Polygenic | $>0.99$ | 1.01 (0.96, 1.06) | 0.7     |
|         |           | AMY1 | Copy number | $>0.99$ | 1.01 (0.96, 1.06) | 0.7     |
| meta-analysis | 2,807 | FTO | rs1558902 | Polygenic | $>0.99$ | 1.01 (0.96, 1.06) | 0.7     |
|         |           | AMY1 | Copy number | $>0.99$ | 1.01 (0.96, 1.06) | 0.7     |

AMY1 copy number was inferred by ddPCR in the Estonian cohort and by whole-genome sequencing in the other cohorts. For the 11-SNP polygenic score, odds ratio or $\beta$ is reported per standard deviation of score. *Power is for a significance level of 0.05. 95% confidence intervals are shown in parentheses.
association with a polygenic score calculated from 11 SNPs ($P = 3.7 \times 10^{-15}$) (Table 2, Supplementary Fig. 11 and Supplementary Table 6). With these positive controls validating the study design and demonstrating power to detect the much larger reported effect of $AMY1$ (ref. 4), we used ddPCR to obtain the integer copy numbers of all three amylase genes, again observing the preponderance of even $AMY1$ copy numbers (Fig. 3a). We had >99% power to detect (at nominal significance) effects as strong as those reported4. However, we did not observe even a nominal association between obesity and the copy number of any amylase gene ($P = 0.70$ for $AMY1$) (Fig. 3b and Supplementary Table 7).

We then analyzed two other cohorts of European-ancestry individuals—one consisting of 2,807 individuals (1,437 type 2 diabetes cases and 1,370 controls) sequenced to >5x average coverage (Genetics of Type 2 Diabetes (GoT2D) cohort) and the other of 657 European-ancestry individuals sequenced to 7x median coverage (InCHIANTI25). Analysis of amylase gene copy number (using Genome STRiP3,5) again showed the enrichment of even, relative to odd, copy numbers (Fig. 3a), validating the precision of the analysis. The GoT2D control and set controls each had 95% power to detect the reported effect4 of $AMY1$ at nominal significance, whereas InCHIANTI samples had 77% power (Table 2). Yet, $AMY1$ copy number did not associate with BMI in any group ($P = 0.31$ for GoT2D controls, $P = 0.24$ for GoT2D cases and $P = 0.53$ for InCHIANTI samples) (Fig. 3) or in a meta-analysis of all 3,464 replication samples ($P = 0.38$). By contrast, SNPs at FTO and other loci implicated in GWAS had the associations expected given sample size and statistical power (Table 2 and Supplementary Table 8).

These results contrast with a recent report finding that $AMY1$ copy number exerts a stronger effect on BMI and obesity than do SNPs at FTO and other loci4. We believe that the difference from the reported observation likely comes from our use of higher-resolution approaches for both molecular and computational analysis (Supplementary Fig. 12). Many studies have found that low-resolution, poorly clustering molecular data conceal technical effects that can create the false impression of strong association13-15,26,27. We also considered the possibility that our study could have failed to detect a real genetic effect. Our study used an Estonian study cohort, in addition to two other European-ancestry cohorts with elevated body weight. The Estonian diet is slightly different from that of other European countries28, although it seems to be similarly rich in starch29. We also considered the possibility that amylase acts in ways that are specific to lean individuals, but we saw no evidence for this hypothesis in our BMI cohorts (Supplementary Table 9) and we note that other BMI-associated variants have tended to associate in ways that are consistent across the BMI spectrum30. We also note that a subsequent study of a different obesity cohort31 did not observe the previously reported shifting of the distribution of $AMY1$ copy number between obese and lean individuals but instead described an outlier set of control samples with unusually high $AMY1$ copy number measurements31. We believe that these results constitute a different hypothesis rather than a replication of the earlier finding at $AMY1$.

Fully understanding human genetic variation and its relationship to phenotypes will require characterizing hundreds of complex loci such as the amylase locus, which mutate at high frequencies in ways that cause large-scale changes in the dosage and expression of genes. Some of these loci could, as has been proposed32-34, represent loci capable of rapid evolutionary adaptation. The amylase locus offers several insights to guide studies of structurally complex loci. First, the high apparent complexity observed in measurements from diploid genomes may arise combinatorially from a modest number of common structural forms that appear in different combinations in different diploid genomes. Second, structurally complex loci reflect both ancient and recent mutations and may be best understood through combinations of analysis methods developed for common and rare variants, including tagging, imputation and direct molecular analysis. Third, although GWAS may miss or underestimate the relationships of structurally complex loci to phenotypes, accurately typed SNP markers can help constrain plausible expectations about the strength of a CNV’s potential effect on a phenotype. Whole-genome sequencing of large cohorts will ultimately determine the extent to which this locus and many other structurally complex loci contribute to human phenotypes.

URLs. Estonian Genome Center of the University of Tartu (EGCUT), http://www.geenivaramu.ee/en; SNP genotypes from the 1000 Genomes Project, Omni chip data, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120313_omni_genotypes_and_intensities/; GIANT Consortium, http://www.broadinstitute.org/collaboration/giant/; Genetic Power Calculator, http://pngu.mgh.harvard.edu/~purcell/gpc/.

METHODS

Methods and any associated references are available in the online version of the paper.
ACKNOWLEDGMENTS

This work was supported by a grant from the National Human Genome Research Institute (R01 HG006855) to S.A.M. to support C.L.U., R.E.H. and S.A.M. Work by T.E. and A.M. was supported through the Estonian Genome Center of the University of Tartu (EGCUT) by Targeted Financing from the Estonian Ministry of Science and Education (SF0180142608), the Development Fund of the University of Tartu (SPIP-GVARENG) and the European Regional Development Fund to the Centre of Excellence in Genomics (3.2.0304.11-0312) and through Framework Programme 7 grant 313010. T.E., A.M. and J.N.H. were further supported by the US National Institutes of Health (R01 DK075787). T.M.F. is supported by European Research Council funding (Framework Programme 7, SZ-50371-GLUCOSEGENES), M.A.T. and M.N.W. are supported by the Wellcome Trust Institutional Strategic Support Award (WT097835MF), and M.B. is supported by US National Institutes of Health grant DK062370.

AUTHOR CONTRIBUTIONS

C.L.U., J.N.H. and S.A.M. conceived the project. C.L.U. pursued molecular (ddPCR) and statistical analyses of amylase locus structural variation. R.E.H. contributed analyses of whole-genome sequence data. T.E., A.M., C.L.U., J.E.M. and J.N.H. analyzed the Estonian cohort. M.A.T., M.N.W. T.M.F., R.E.H. and S.K. analyzed the InCHIANTI cohort. M.I.M., M.B., D.M.A., R.E.H., C.L.U. and C.E. analyzed the GoT2D cohort. C.N.P., M.T.P., C.L.U. and R.E.H. analyzed the GPC cohort. A.R.H. and H.C. performed the NanoChannel-based genome mapping. C.L.U., J.N.H. and S.A.M. wrote the manuscript, with contributions from D.M.A., T.M.F., M.B., M.I.M. and T.E.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712 (2010).
2. Sudmant, P.H. et al. Diversity of human copy number variation and multicolony data. Science 330, 641–646 (2010).
3. Handsaker, R.E. et al. Large multiallelic copy number variations in humans. Nat. Genet. 47, 296–303 (2015).
4. Falchi, M. et al. Low copy number of the salivary amylase gene predisposes to obesity. Nat. Genet. 46, 492–497 (2014).
5. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. 43, 269–276 (2011).
6. Hindson, B.J. et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal. Chem. 83, 8604–8610 (2011).
7. Hastie, A.R. et al. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. PLoS ONE 8, e55864 (2013).
8. Groot, P.C. et al. The human α-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics 5, 29–42 (1989).
9. Perry, G.H. et al. Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39, 1256–1260 (2007).
10. Groot, P.C., Mager, W.H. & Frants, R.R. Interpretation of polymorphic DNA patterns in the human α-amylase multigene family. Genomics 10, 779–785 (1991).
11. Locke, A.E. et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197–206 (2015).
12. Cantaliieris, S. & White, S.J. Correlating multiallelic copy number polymorphisms with disease susceptibility. Trends Genet. 28, 245–257 (2012).
13. Barnes, C. et al. A robust statistical method for case-control association testing with copy number variation. Nat. Genet. 40, 1245–1252 (2008).
14. Clayton, D.G. et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat. Genet. 37, 1243–1246 (2005).
15. Zanda, M. et al. A genome-wide assessment of the role of untagged copy number variants in type 1 diabetes. PLoS Genet. 10, e1004367 (2014).
16. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).
17. International HapMap Consortium. The International HapMap Project. Nature 426, 789–796 (2003).
18. Carpenter, D. et al. Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. Hum. Mol. Genet. 24, 3472–3480 (2015).
19. Groot, P.C. et al. Evolution of the human α-amylase multigene family through unequal, homologous, and inter- and intrachromosomal crossovers. Genomics 8, 97–105 (1990).
20. Boettiger, L.M., Handsaker, R.E., Zody, M.C. & McCarroll, S.A. Structural haplotypes and recent evolution of the human 17q21.31 region. Nat. Genet. 44, 881–885 (2012).
21. Steinberg, K.M. et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat. Genet. 44, 872–880 (2012).
22. Lupski, J.R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. 1, e49 (2005).
23. Leitsalu, L. et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. doi:10.1093/ije/dys286 (2014).
24. Speliotes, E.K. et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. 42, 937–948 (2010).
25. Ferrucci, L. et al. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J. Am. Geriatr. Soc. 48, 1618–1625 (2000).
26. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).
27. Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713–720 (2010).
28. Tognon, G. et al. Mediterranean diet, overweight and body composition in children from eight European countries: cross-sectional and prospective results from the IDEFICS study. Nutr. Metab. Cardiovasc. Dis. 24, 205–213 (2014).
29. Mottus, R. et al. Personality traits and eating habits in a large sample of Estonians. Health Psychol. 31, 806–814 (2012).
30. Berndt, S.I. et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. 45, 501–512 (2013).
31. Mejia-Benitez, M.A. et al. Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican children. Diabetologia 58, 290–294 (2015).
32. Iskow, R.C., Gokcumen, O. & Lee, C. Exploring the role of copy number variants in human adaptation. Trends Genet. 28, 245–257 (2012).
33. Stankiewicz, P. & Lupski, J.R. Structural variation in the human genome and its role in disease. Annu. Rev. Med. 61, 437–455 (2010).
34. Zhang, F., Gu, W., Hurles, M.E. & Lupski, J.R. Copy number variation in human health, disease, and evolution. Annu. Rev. Genomics Hum. Genet. 10, 451–481 (2009).
ONLINE METHODS

Cohort collection. Estonian sample collection. The Estonian Biobank is the population-based biobank of the Estonian Genome Center of the University of Tartu (EGCUT). The EGCUT is conducted according to the Estonian Gene Research Act, and all participants are volunteers who have signed broad informed consent. The cohort size is currently 51,535 individuals of 18 years of age and older. All subjects are recruited randomly by general practitioners and physicians in hospitals. A computer-assisted personal interview is conducted at the doctor’s office to record personal data, genealogical data, lifestyle data, and the subjects’ educational and occupational history.

Height and weight are measured at recruitment by a medical professional, and all diagnosed diseases are recorded according to the ICD10 system. Anthropometric measurements are taken, along with blood pressure (sitting position at the end of the interview) and resting heart rate. Thirty to 50 milliliters of venous blood is collected into EDTA Vacutainers. These are transported to the central laboratory of EGCUT at 4–6 °C within 6 to 36 h of collection. Upon arrival, DNA, plasma and white blood cells are immediately isolated and kept in aliquots in MAP straws in liquid nitrogen.

A Hamilton Robotics Automated Sample Management system with a capacity of 100,000 tubes is used for intermediate storage of normalized DNA samples (30–100 ng/µl) in tubes with two-dimensional barcodes. This enables quick and highly accurate delivery of the samples by cherry-picking according to the selected barcodes.

The 1,000 Estonian samples used in the current study were selected from the BMI extremes of the EGCUT population cohort of 51,535 samples. The lean (BMI < 22) and obese (BMI > 33) groups were matched on the basis of age at recruitment and year of birth to account for cohort and life-course effects on overall body composition (demographic details given in Supplementary Table 10), resulting in the obese being in the 98th percentile (females, 99.5th percentile) of the cohort's BMI distribution and the lean being in the 87th (females, 82nd) percentile. The lower threshold for the lean control samples is due to age and sex matching to the obese samples. The extreme sets included 250 samples from both sexes and were limited to only individuals of Estonian nationality whose self-reported mother tongue was Estonian. Further measures are taken to account for potential population structure are described below.

GoT2D sample collection. The GoT2D study aims to characterize the genetic architecture of type 2 diabetes and related quantitative traits through low-coverage whole-genome sequencing, deep (~100x) exome sequencing and SNP genotyping by Illumina HumanOmniv2.5 array of cases and controls from four large European-ancestry cohorts: the Diabetes Genetics Initiative (DGI), the Finland–United States Investigation of NIDDM Genetics (FUSION), GoT2D–UK and Kooperative Gesundheitsforschung in der Region Augsburg (KORA). They are a mixture of smaller ones obtained in Finland, Sweden, the UK and Germany. Owing to the confounding effect of diabetes on BMI, cases and controls were analyzed separately while controlling for the cohort of origin. Only the low-coverage data were used in the current study (Supplementary Table 11).

InChIANTI cohort sample collection. We selected 680 individuals from the InChIANTI study—a study of aging from the Chianti region in Tuscany, Italy—for low-pass whole-genome sequencing. Individuals were selected for sequencing on the basis of the availability of gene expression and circulating biomarker data (Supplementary Table 12).

Whole-genome sequencing was performed at the Beijing Genomics Institute (BGI; Shenzhen, China) using the Illumina HiSeq 2000 platform to obtain a minimum read depth of 6x. An average of 240 million paired-end 90-bp reads per sample were aligned to the 1000 Genomes Project implementation of the Genome Reference Consortium’s Build 37 of the human reference genome, using the Burrows-Wheeler aligner (BWA), version 1.5.9 (ref. 45).

GPC cohort sample collection. Headed by the Center for Genomic Psychiatry at the University of Southern California, the GPC cohort is a collection of individuals with schizophrenia and bipolar disorder, along with controls. Participants are recruited in this study consisted of 768 self-reported (and genetically confirmed) European-ancestry patients from the United States who underwent whole-genome sequencing from blood to a depth of 30x. Data on BMI were not available.

Droplet digital PCR. General. ddPCR® is similar in concept and preparation to a real-time quantitative PCR reaction but with a few important modifications. Before amplification, the DNA is first digested with a restriction enzyme to physically separate the copies of a CNV that are on the same DNA strand. The PCR reaction mixture is prepared similarly to that for quantitative PCR, with each primer at a concentration of 900 nM, the fluorescent probes at a concentration of 250 nM and the input DNA at a concentration of around 1 ng/µl. The reaction is then emulsified into approximately 20,000 water droplets surrounded by oil using a droplet generator (Bio-Rad). The droplets are thermocycled using a standard thermocycler with the Bio-Rad–supplied PCR protocol (with an additional 10 cycles), and the droplets containing probe targets then become fluorescent. Fluorescent droplets are counted by a droplet reader (Bio-Rad). At low DNA input concentrations, each fluorescent droplet contains only one PCR target, thus allowing the measurement of nearly exact numbers of targets within the reaction—as opposed to comparing amplification curves, as in quantitative PCR. At higher DNA concentrations, a Poisson correction factor is applied to count droplets possibly having more than one target.

Control probes. The standard control probe for ddPCR is targeted to RPP30. However, because amylase is in a late-replicating region, DNA isolated from replicating cells will naturally have less of it than other parts of the genome—the parts that have already replicated (Supplementary Fig. 3c). To counteract this bias, we used a probe assay targeted to just outside the amylase region, called Near_AMY (Supplementary Table 1).

Genotype calling. The output of the droplet reader is a scatterplot with FAM fluorescence on the y axis and HEX or VIC fluorescence on the x axis. Each dot represents a droplet (Supplementary Fig. 13). QuantaSoft software draws suggested thresholds for droplets positive for FAM and HEX. The experimenter checks these thresholds and redraws if needed (while still blinded to sample identity). A CNV copy number call is determined by dividing the number of droplets fluorescing FAM, corresponding to the CNV target, by the number of droplets fluorescing VIC, the control target (with both numbers Poisson corrected). Before final genotype calling, the raw CNV calls from each plate are corrected by a plate-wide correction factor, generally between 0.97 and 1.05 (Supplementary Fig. 13c).

HapMap samples. Plates containing the HapMap DNA samples from the CEU and YRI populations were subjected to ddPCR in three reactions for AMY1B, AMY2A and AMY1 using the “assay” assays listed in Supplementary Table 1. All except half the CEU individuals on the AMY2B run were analyzed using the control assay Near_AMY. DNA inputs varied, owing to variation in DNA concentration across the plate, but the ideal DNA concentrations for which we aimed were 1 ng/µl for AMY1 reactions and 0.5 ng/µl for AMY2B and AMY2A reactions. All copy numbers reported are from a single reaction for each gene. We did not average multiple replicates to obtain copy numbers. However, we do have multiple runs on file for these assays and others (Supplementary Fig. 3b.d.e).

Estonian samples. The Estonian DNA samples were aliquotted into 96-well plates, with a random distribution of under- and overweight samples (KS test, P = 0.51). The ddPCR runs were performed within a 3-week period in the same laboratory and using the same machines, with an experimenter blinded to the case-control status of the samples—thus reducing the risk of batch effects and biases. Each sample had one genotyping run of each of the following assays: AMY2B_assay1, AMY2A_assay1, AMY1_assay1 and AMY1_assay2 (with the exception of plate 1, which did not have an AMY1_assay2 run). AMY2B. Initially, 28 Estonians had copy number calls less than 2 for AMY2B (copy number of 0 or 1), a call that should be impossible given the copy number distribution for AMY2B. We hypothesized that an Estonian-specific SNP might be interfering with the assay and ran these samples again using AMY2B_assay2. All the samples, except for three, then had calls consistent with the known AMY2B distribution (copy number of two or three). In the association analysis, the AMY2B_assay2 genotype calls were used for these samples.

AMY1. Two different assays targeting AMY1 were used to reduce the noise that a single assay might have. Two different DNA input concentrations were used with the AMY1 assays to ensure that each sample had at least one genotype call acquired when it was within the optimal concentration range for ddPCR. In the concentrated reaction, each sample on the plate was...
precalculated to have an input DNA concentration of >0.2 ng/μl and was genotyped with AMY1_assay1. However, given the wide distribution of sample concentrations on each plate, the concentrated run resulted in many of the samples having oversaturated reactions. In the diluted run (AMY1_assay2), each sample input was precalculated to produce >10% probe-negative droplets, thus diluting the previously oversaturated samples.

AMY2A-adjusted averaging of the AMY1 copy number calls. To avoid biases that might arise from differences in sample DNA concentration between cases and controls, we did not filter or clean the data on the basis of concentration and used every genotype call the Bio-Rad QuantaSoft software provided. However, a straight average of the two AMY1 replicate genotype calls was not ideal, as many of the samples had one genotype call obtained when the sample was too dilute or overly concentrated, thus adding noise to the better genotype call (Supplementary Fig. 14).

Given that AMY1 and AMY2A have parity (that is, their copy numbers are either both odd or both even), we could check the correctness of the AMY1 copy number call using the AMY2A call. In practice, this meant checking each individual’s two replicate AMY1 calls for concordance with the AMY2A call. If both AMY1 calls were concordant, they were averaged (70% of samples). If only one call was concordant, only the concordant AMY1 genotype was used (24% of samples). If both calls were discordant, they were averaged (6% of samples). This resulted in better clustering of copy numbers at integers (average deviation from integers of 0.152 in comparison to 0.179 when using the AMY1-adjusted AMY1 call). The variability of all segments, except for the intergenic region, was confirmed by designing ddPCR assays targeting these segments and carrying out ddPCR experiments to confirm the sequencing-based results.

The bins used for the read depth analysis are listed in Supplementary Table 1. Even though the bins for AMY1 were substantially larger than the AMY1 repeated segment, most of the signal Genome STRIP used to call genotypes arose from the AMY1 repeated segment (Supplementary Fig. 16).

BioNano Genomics, genome mapping. NanoChannel array–based genome mapping experiments were performed by BioNano Genomics. In brief, genome mapping can be thought of as next-generation restriction mapping. Long, whole strands of DNA (~300 kb) are labeled at sequence-specific nickase (Nt.BspQI) recognition sites, and the DNA backbone is stained with YOYO1. The DNA is electrophoresed through a NanoChannel array to straighten it for image capture. The nickase creates patterns that can be used to assemble a whole genome or pieces thereof15, in a manner similar to restriction mapping.

Each amylase gene has its own restriction pattern, and, because genome mapping uses whole strands of DNA, we can determine the order and orientation of the genes from these patterns.

We selected three individuals who together had three unreported haplotypes (AH2, AH4B2 and AH2B2), one partially assembled haplotype (AH4) and two known haplotypes to serve as positive controls (AH3 and AH1) (Supplementary Fig. 5). The haplotypes that had already been assembled (AH3 and AH2; ref. 44) were largely consistent. In contrast, the AH4 structure contradicts the structure in Perry et al.9, having one fewer inverted AMY1 copy. In addition, anonymous samples to which BioNano had access, as well as a European-American family, contained AH1, AH3 and AH5, which assembled into structures consistent with the known haplotypes. Of note, in several haplotypes, the AMY2A pseudogene was inverted. This feature appeared to be stably inherited but has not been confirmed using a second technology.

Genotyping the InCHIANTI cohort with the alternative read depth method. We analyzed 657 samples after quality control checks. Average depth was 7×. We aligned the reads to the repeat-masked GRC Build 37 reference genome using the msuFAST ultra version 3.3.1 algorithm, which aligns single reads to multiple positions in the genome and so is optimal for regions of variable copy number13. Repeats were detected and masked using both RepeatMasker Open-3.0 and Tandem Repeats Finder 4.07b46. Reads were mapped in single-read mapping mode with a Hamming distance threshold of ≤4 bp. We derived GC-corrected absolute copy number in 100-bp windows using mrCaNaVsR version 0.51 (ref. 47), a program that predicts from read depth and GC enrichment an absolute copy number. We calculated a mean copy number value for the three combined AMY1 regions, AMY2A and AMY2B. The distribution of AMY1 copy number is given in Supplementary Figure 12. The read depth bins are listed in Supplementary Table 1.
Phasing of HapMap samples. At first, trios were phased manually using only haplotypes described in previous literature8–10, resulting in successful phasing for only 7% of the trios and 15% of individuals. We noticed patterns in some of the unphased trios that could be explained by new haplotypes (Supplementary Table 3) and found population evidence to support these haplotypes (Fig. 1c and Supplementary Fig. 4). Adding the five new haplotypes resulted in successful phasing for 27% of trios and 39% of individuals. The remaining trios and individuals do not necessarily contain unknown haplotypes; rather, most of them just have genotypes that correspond with multiple combinations of known haplotypes. For instance, six was the most common copy number for AMY1 and can be accomplished with three different combinations of known haplotypes, resulting in phasing failure for nearly every individual with six copies.

Calculation of haplotype frequencies. We could not calculate haplotype frequencies on the basis of the individuals we could phase, as this would artificially enrich for haplotypes that can create unique genotypes that can be phased. Instead, we used haplotype AH2. Haplotype AH2 can be identified within individuals because it causes a characteristic decrease in AMY2A copy number, and its companion haplotype can be found by simple subtraction of copy numbers. We selected individuals carrying haplotype AH2 from the GPC cohort and the European–ancestry individuals of the 1000 Genomes Project (142 individuals in total) and identified their other haplotype. We calculated the frequency of each haplotype in this pool of other haplotypes and reported it in Figure 1c. The frequency of haplotype AH4 cannot be determined this way, as it causes an increase in AMY2A that balances out the decrease in haplotype AH2; thus, its frequency was determined by identifying individuals who carried haplotype AH4 (marked by an increase in AMY2A) and dividing it by the total.

Clustering of SNP haplotypes (spider plot). All unrelated individuals in the 1000 Genomes Project European–ancestry populations (CEU, TSI, GBR, FIN and IBS) who had amylase genotypes where the two structural haplotypes could be unambiguously determined were selected for SNP clustering in the spider plot shown in Figure 2a. These individuals’ amylase haplotypes and SNP genotypes (downloaded from the 1000 Genomes Project website, Omni chip data) were combined in a .bgl file and phased as a group using BEAGLE version 3.3.2 (ref. 48) under the default conditions with no reference panel. The spider plot created was from the 23 closest SNPs that had a minor allele frequency (MAF) greater than 1% and were outside of the variable region (resulting in 9 SNPs upstream and 14 SNPs downstream). The spider plot was created by traversing the set of SNP haplotypes in both directions from the target variant (amylase) and grouping the haplotypes according to their state at each successive SNP to form two tree structures representing the left and right flanking sequences. At each split, the branch corresponding to the minor allele was plotted above the branch corresponding to the major allele. The color of each horizontal segment indicates the allele frequency of the next SNP on the branch, and the thickness corresponds to the number of haplotypes sharing that segment.

Association of SNPs to amylase haplotypes, copy number and BMI. The individuals of the GPC cohort who had genotypes where the two amylase haplotypes could be unambiguously determined were used to search for tagging SNPs for each haplotype. SNPs with a MAF under 1% and those within the copy number–variable region were not used. Every remaining SNP was correlated with every haplotype in turn, using a Pearson test. During the test, all amylase haplotypes were recoded as 0 or 1, with the target haplotype being 1. P values were permuted by shuffling the amylase haplotypes 1 million times to create a distribution of possible $r^2$ values for each SNP.

The efficiency of imputation was calculated on the basis of leave-one-out trials. Briefly, each individual’s amylase haplotypes were masked in turn within the unphased data and phased using BEAGLE48 under default conditions with no reference panel. The amylase haplotypes that BEAGLE assigned were extracted and compared to the true values for the individuals with masked haplotypes. $r^2$ values were obtained by Pearson correlation, and $P$ values were calculated from 1 million permutations, creating an $r^2$ distribution.
Next, the residuals from the model are transformed using an inverse normal transformation and used in subsequent association analyses.

This protocol was slightly modified for normalizing the 1,000 Estonians to account for the extreme phenotype design and to more precisely capture the underlying trait distribution in the whole sampling cohort of ~51,000. First, the previously described trait normalization (only age and age² were used as covariates) was separately performed in females (n = 32,724) and males (n = 17,352), resulting in the normalized BMI statistics for the 1,000 extreme phenotype samples given in Supplementary Table 6. In subsequent analyses, both sexes were analyzed together, and sex and three MDS genetic vectors (estimated using ExomeChip data) were used as covariates to account for both sex differences and population stratification.

Power analysis in the Estonian cohort. To make sure that the Estonian extremes design had sufficient statistical power to find the associations reported in Falchi et al.⁴, we used the Genetic Power Calculator⁴⁹, as it has the option to account for threshold-selected quantitative trait design. Falchi et al. report that AMY1 copy number explains 0.66 to 4.40% (95% confidence interval) of the genetic variance in BMI. By using the same calculations that Falchi et al. used to estimate the total variance explained for obesity, we back-calculated the mean total BMI variance explained and estimated it as 1.11% (95% confidence interval = 0.461–1.79%). The phenotypic thresholds in standard deviation units for defining the case and control sets were obtained from the normalized BMI scores described in the previous paragraph and were as follows in standard deviation units: (i) case thresholds +2.0 and +4.0 and (ii) control thresholds −1.2 and −4.2.

The genetic effect sizes, trait-increasing allele frequencies and respective total trait variation explained for the SNPs were obtained from the latest GIANT Consortium full report¹⁶ and have been outlined in Supplementary Table 13. For the GIANT Consortium–based power analysis, the stage 1 sample size of 124,000 was used. For estimating our power with the polygenic extremes design had sufficient statistical power to find the associations (estimated using ExomeChip data) were used as covariates to account for both sex differences and population stratification.

Association analysis in the Estonian cohort. Both logistic and linear regression models were used to detect association between BMI and the 21 directly genotyped SNPs using PLINK⁵⁰. In the linear regression model, the normalized BMI scores were used, and sex and three genetic vectors were used in both models as covariates. We used the integer genotypes of the copy number of all 3 amylase genes obtained through ddPCR in the same 1,000 individuals. For the AMY1 gene, four different copy number estimates were used—the AMY2A−adjusted average, the straight average and both genotyping runs separately (as described in “AMY2A−adjusted averaging”). Again, both logistic and linear regression models were fitted using the same phenotype and covariates in R⁵¹. A similar analytical framework was used to detect the association between BMI and the constructed polygenic score. No association between amylase gene copy number and obesity was observed with either model (Supplementary Table 6).

Association analysis in the InCHANTI cohort. We regressed the copy number values against BMI corrected for age and sex. We used all copy number calls regardless of quality. The distribution of total (diploid) absolute copy number plotted against BMI is given in Figure 3e. Our analyses did not provide any evidence of an association between amylase copy number and BMI in AMY1 (P = 0.53), AMY2A (P = 0.37) and AMY2B (P = 0.49). Using the more refined method of classifying AMY1 copy number using the triplicated regions of AMY1 alone, we still did not see any association with BMI (P = 0.50).

Association analysis in the GoT2D cohort. The data set was first divided into T2D cases (n = 1,437) and controls (n = 1,374). The BMI phenotype was transformed using the GIANT protocol (described above; adjusting BMI scores in a sex-stratified way for age and age² by fitting a linear regression model and subsequently applying inverse normal transformation on resulting residuals from the model). Cases and controls were analyzed separately, and the copy number values were regressed against normalized BMI scores while adjusting for sex and source cohort. All copy number calls were used regardless of quality. The latter variable was included to correct for population stratification as the GoT2D sample consists of nine separate cohort collections (Botnia, Diabetes Registry, FUSION, Helsinki, KORA, Malmö, MPP, STT and WTCCC). Whereas both subcohorts (cases and controls) had >95% power to replicate the result from Falchi et al.⁴ (total variance explained of 1.11%), we did not observe nominally significant associations with any of the amylase locus copy numbers in either subcohort (Supplementary Table 8).

35. Zeggini, E. et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316, 1336–1341 (2007).
36. Scott, L.J. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316, 1341–1345 (2007).
37. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316, 1331–1336 (2007).
38. Heid, I.M. et al. Genetic architecture of the APM1 gene and its influence on adiponectin plasma levels and parameters of the metabolic syndrome in 1,727 healthy Caucasians. Diabetes 55, 375–384 (2006).
39. Meizer, D. et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 4, e1000072 (2008).
40. Church, D.M. et al. Modernizing reference genome assemblies. PLoS Biol. 9, e1001091 (2011).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
42. Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am. J. Hum. Genet. 91, 1033–1040 (2012).
43. Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience 3, 34 (2014).
44. Teague, B. et al. High-resolution human genome structure by single-molecule analysis. Proc. Natl. Acad. Sci. USA 107, 10848–10853 (2010).
45. Hach, F. et al. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat. Methods 7, 576–577 (2010).
46. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580 (1999).
47. Nukan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. 41, 1061–1067 (2009).
48. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097 (2007).
49. Purcell, S., C.C. & Sham, P.C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19, 149–150 (2003).
50. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
51. R Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2015).