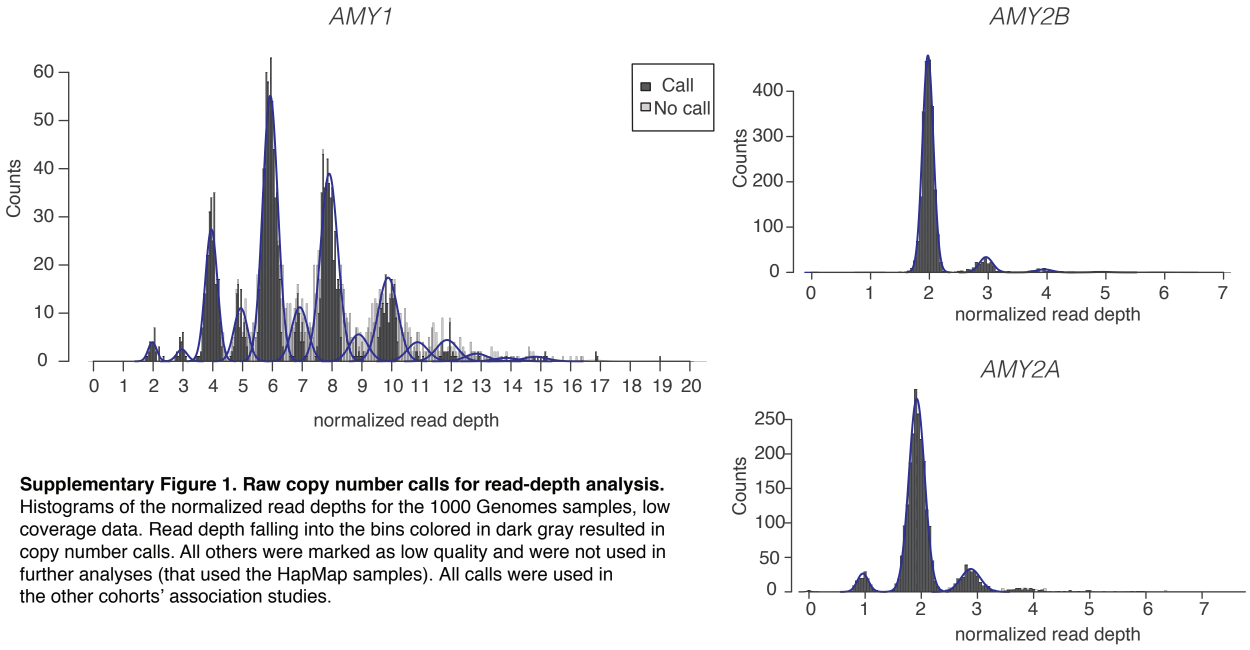
Supplementary Material for:

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

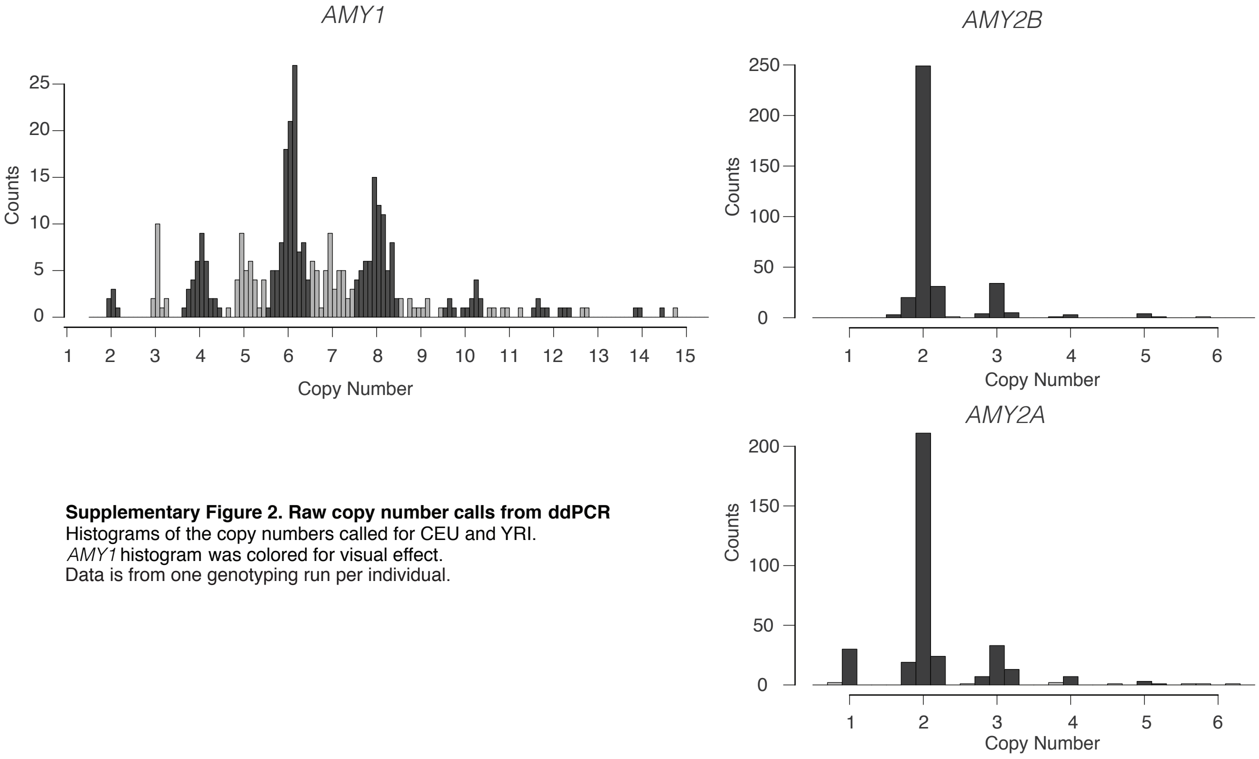
Christina L Usher, Robert E Handsaker, Tõnu Esko, Marcus A Tuke, Michael N Weedon, Alex R Hastie, Han Cao, Jennifer E Moon, Seva Kashin, Christian Fuchsberger, Andres Metspalu, Carlos N Pato, Michele T Pato, Mark I McCarthy, Michael Boehnke, David M Altshuler, Timothy M Frayling, Joel N Hirschhorn, Steven A McCarroll

Table of Contents

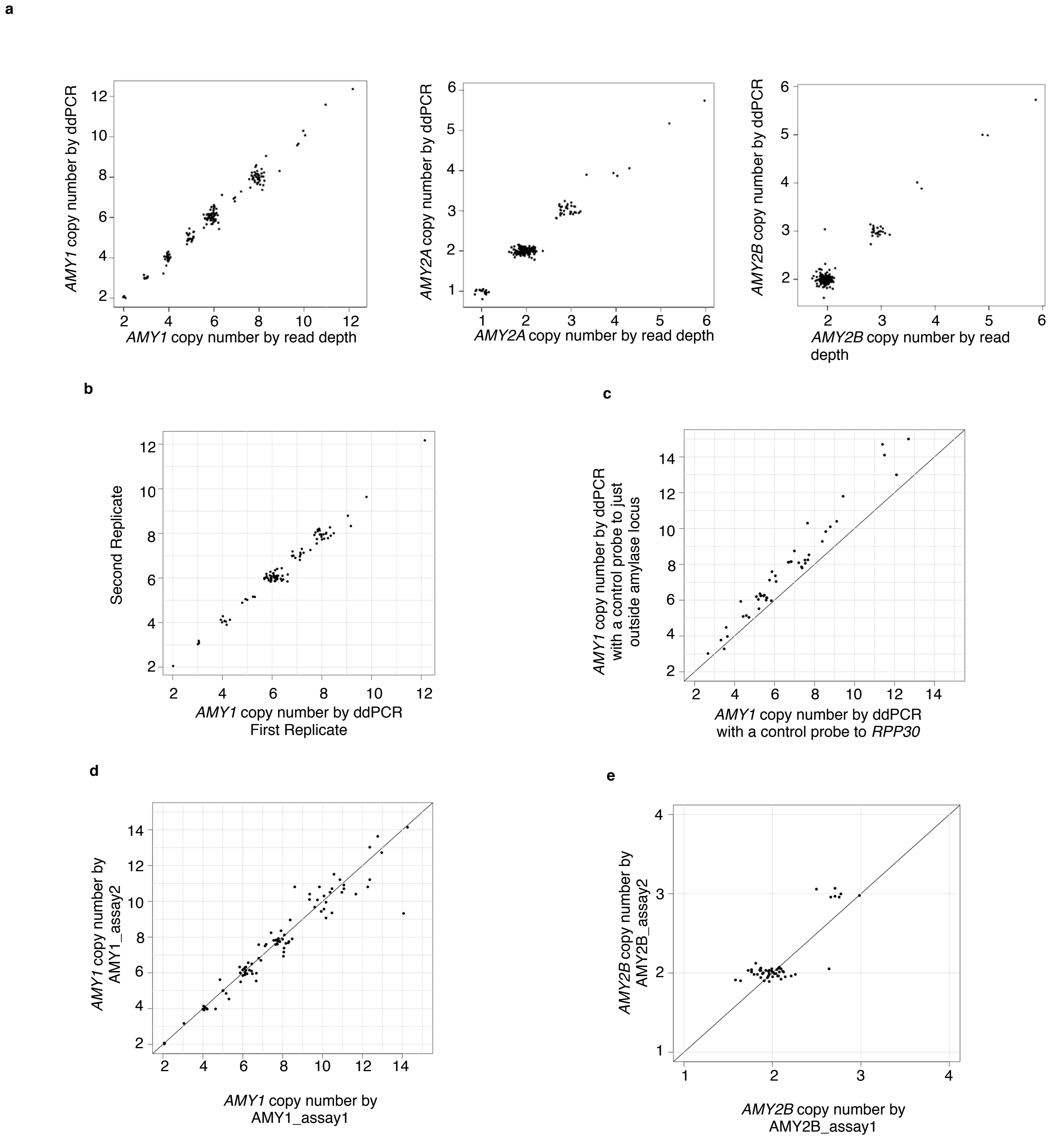
SUPPLEMENTARY FIGURES.. 1
Supplementary Figure 1. Raw copy number calls for read-depth analysis... 1
Supplementary Figure 2. Raw copy number calls from ddPCR... 1
Supplementary Figure 3. Agreement between genotyping methods on the copy number of the amylase genes.. 2
Supplementary Figure 4. Population evidence for novel haplotypes.. 3
Supplementary Figure 5. BioNano optical mapping assemblies 4
Supplementary Figure 6. SNPs tagging haplotypes.. 5
Supplementary Figure 7. Detailed structure of the amylase locus.. 5
Supplementary Figure 8. Detailed list of amylase haplotypes.. 6
Supplementary Figure 9. Cross-cohort concordance of SNPs' association to *AMY1* copy number................ 7
Supplementary Figure 10. SNPs' association to *AMY1* copy number and obesity.. 7
Supplementary Figure 11: Power calculations for the detection of obesity-related SNPs in the Estonian cohort 8
Supplementary Figure 12. Comparison of CNV analysis programs using the same data set.......................... 9
Supplementary Figure 13. Output of ddPCR 10
Supplementary Figure 14. Estonian *AMY1* copy number calls... 11
Supplementary Figure 15: Concordance between ddPCR and GenomeSTRIP in InCHIANTI samples...... 12
Supplementary Figure 16: Alternative read-depth bins for measuring *AMY1* copy number......................... 12
Supplementary Figure 17: Principle component analysis of the ancestry of the Estonians........................... 13



Supplementary Figure 1. Raw copy number calls for read-depth analysis. Histograms of the normalized read depths for the 1000 Genomes samples, low coverage data. Read depth falling into the bins colored in dark gray resulted in copy number calls. All others were marked as low quality and were not used in further analyses (that used the HapMap samples). All calls were used in the other cohorts' association studies.



Supplementary Figure 2. Raw copy number calls from ddPCR
Histograms of the copy numbers called for CEU and YRI.
AMY1 histogram was colored for visual effect.
Data is from one genotyping run per individual.



Supplementary Figure 3. Agreement between genotyping methods on the copy number of the amylase genes
Europeans (CEU) and Yorubans (YRI) from the 1000Genomes project were genotyped for amylase copy number using read-depth analysis (GenomeSTRiP) and ddPCR.

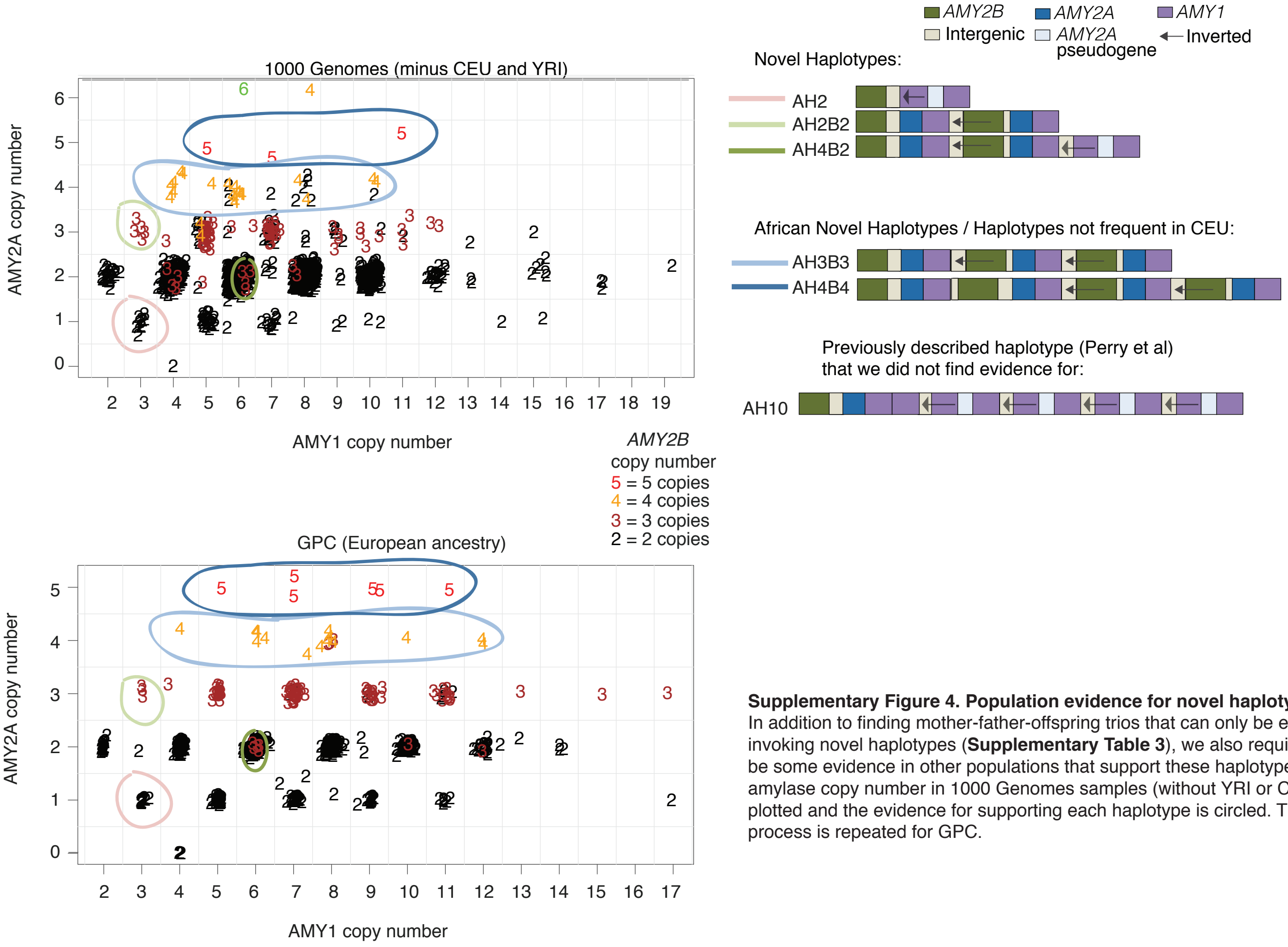
(A) Read-depth and ddPCR show high concordance for all three of the amylase genes.

(B) Two ddPCR reactions were run on the same plate to test the reproducibility of *AMY1* copy number calls. Concordance was 96%.

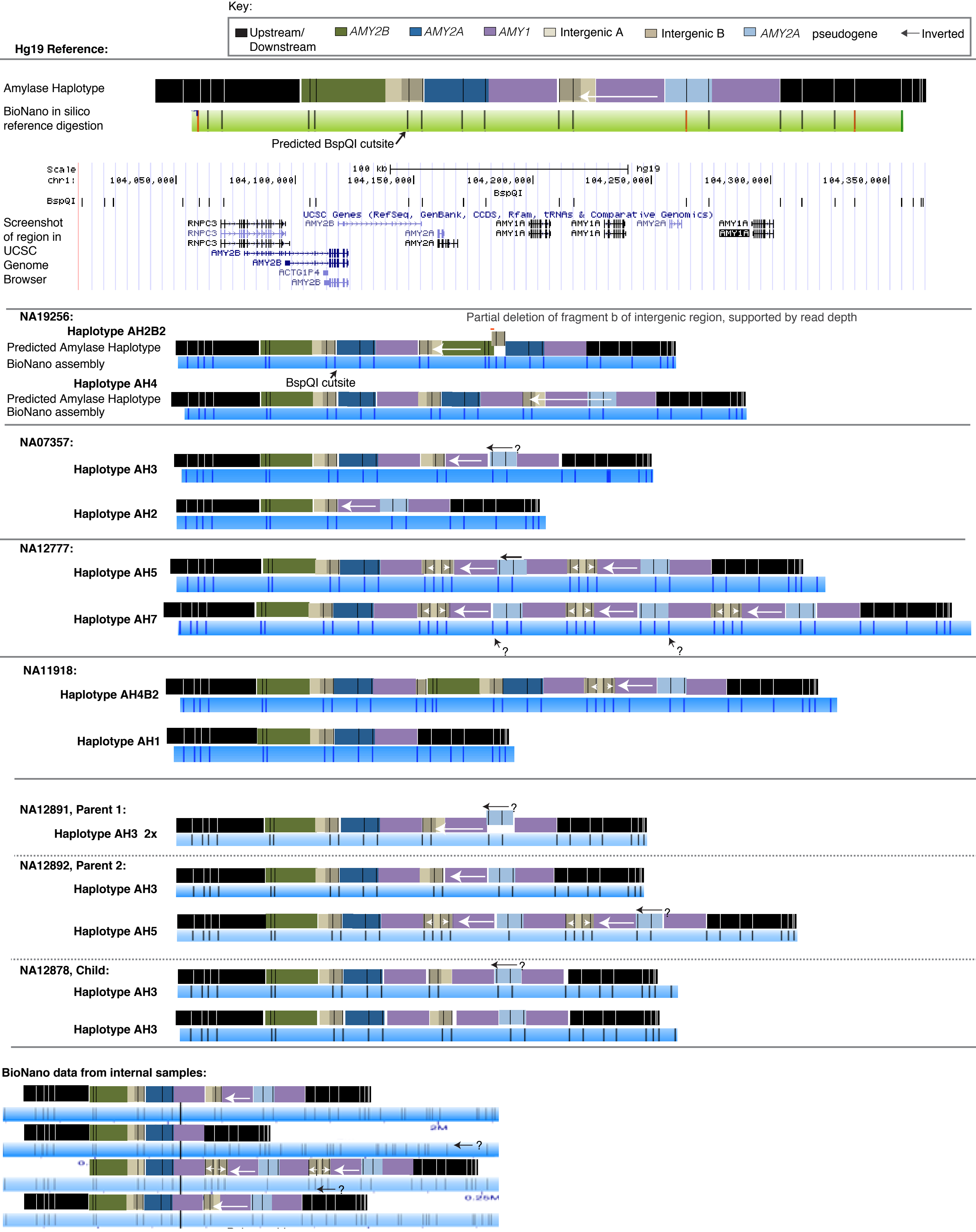
(C) ddPCR on amylase is sensitive to which control probe is used. Amylase is a late-replicating locus of the human genome, therefore there will be less of it than other parts of the genome in DNA isolated from replicating cells.

(D) *AMY1* copy number determined using 2 different assays. AMY1_assay1 and AMY1_assay2. Sequences are in **Supplementary Table 1.**

(E) *AMY2B* copy number determined using 2 different assays, AMY2B_assay1 and AMY2B_assay2. Sequences are in **Supplementary Table 1.**



Supplementary Figure 4. Population evidence for novel haplotypes. In addition to finding mother-father-offspring trios that can only be explained by invoking novel haplotypes (**Supplementary Table 3**), we also required there to be some evidence in other populations that support these haplotypes. Above, amylase copy number in 1000 Genomes samples (without YRI or CEU) is plotted and the evidence for supporting each haplotype is circled. The same process is repeated for GPC.

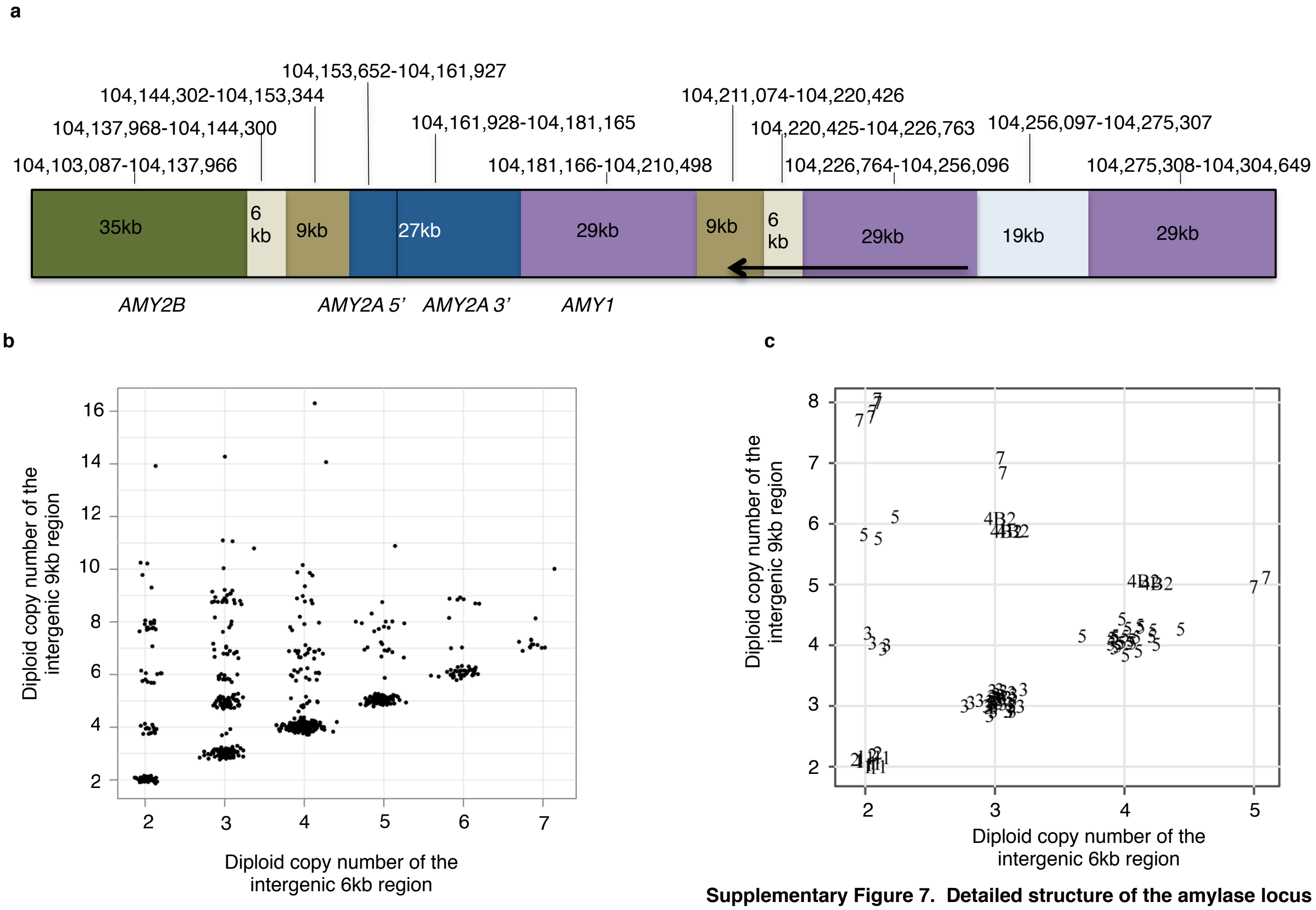


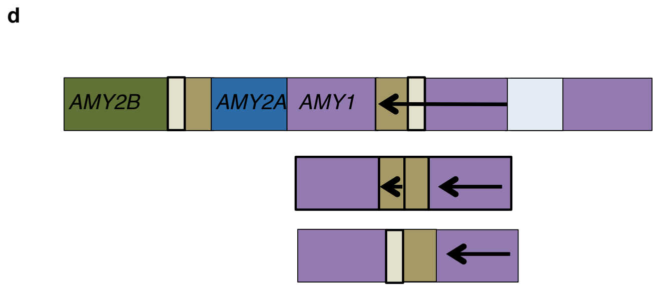
Supplemental Figure 5. Optical mapping assemblies for haplotypes of amylase.

For optical mapping, strands of whole, labeled HMW DNA (nick labeled and backbone stained with YoYo1) are electrophoresed through a NanoChannel array. The array straightens the DNA and the fluorescence left at the nick sites creates patterns that can be used to assemble a whole genome, or pieces thereof, in a manner similar to restriction mapping. Each amylase gene has its own restriction pattern, so we can determine the order and orientation of the genes from these patterns. Of note, in several haplotypes the AMY2A pseudogene is inverted. This feature appears to be stably inherited, but has not been confirmed using a second technology.

Supplementary Figure 6. SNPs tagging haplotypes. *P* values were obtained from 1 million permutations. Only a few associations and imputations replicated in G1000, probably due to a smaller sample size and fewer appearances of rarer haplotypes.

Haplotype (AH#)	Best Associated SNP	Minor Allele Freq.	r^2 of SNP	p-value	Replicated in 1000 Genomes?	r^2 of Imputation	p-value	Replicated in 1000 Genomes?
1	rs113922683	0.19	0.22	$<10^{-6}$	yes	0.23	$<10^{-6}$	yes
3	rs34848656	0.26	0.08	$<10^{-6}$	no	0.23	$<10^{-6}$	yes
5	rs1566154	0.24	0.07	$<10^{-6}$	no	0.03	$<10^{-3}$	no
7	s75133138	0.16	0.19	$<10^{-6}$	yes	0.09	$<10^{-5}$	yes
2	rs72694406	0.43	0.32	$<10^{-6}$	yes	0.43	$<10^{-6}$	yes
4	rs12740780	0.12	0.26	$<10^{-6}$	no	0.37	$<10^{-6}$	no
2B2	rs12076610	0.18	0.11	$<10^{-6}$	yes	0.44	$<10^{-6}$	no
4B2	rs79043596	0.17	0.27	$<10^{-6}$	yes	0.24	$<10^{-6}$	no





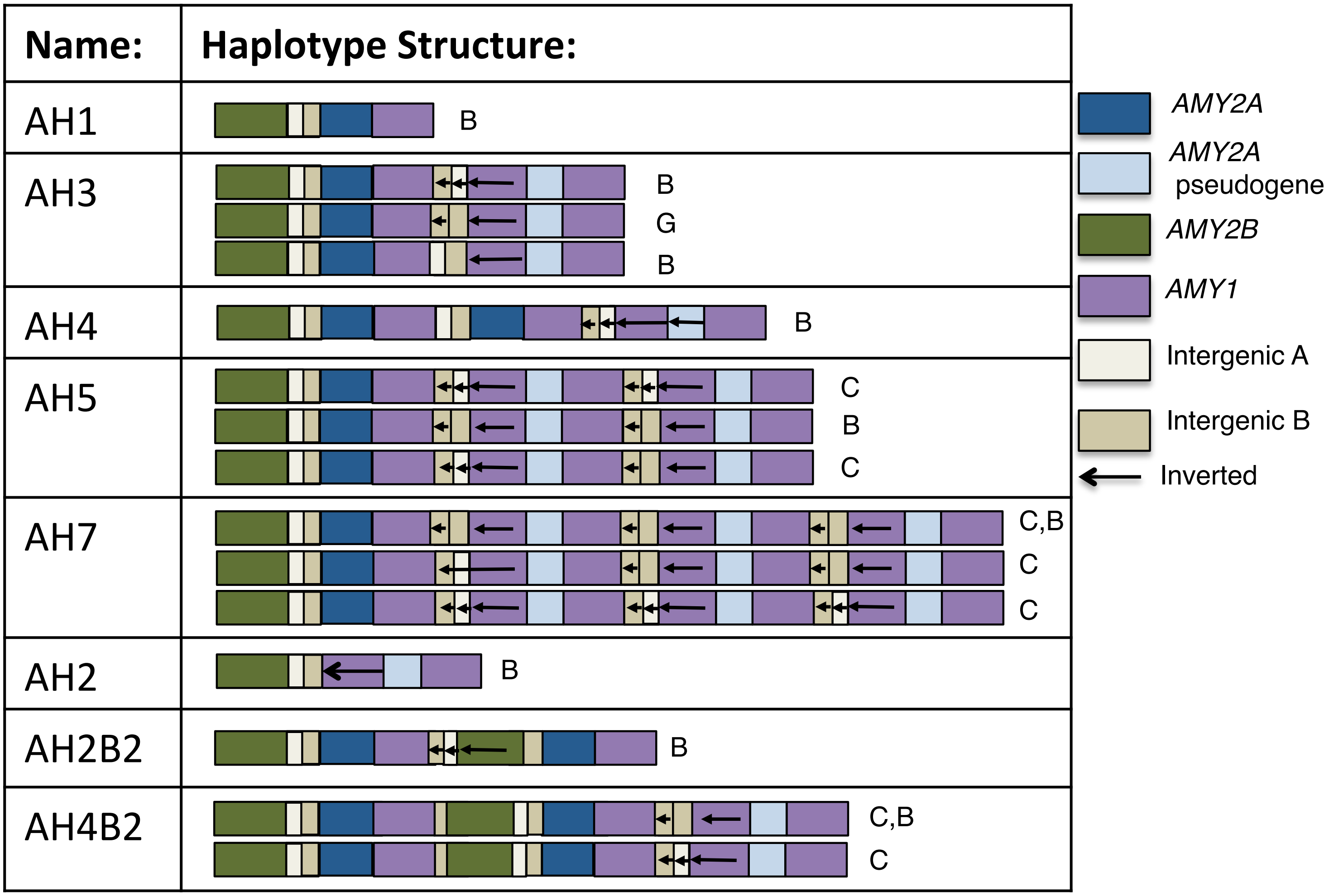
Supplementary Figure 7. Detailed structure of the amylase locus

(A) Hg19 genomic coordinates for the segments of the amylase locus. The colored representation of the locus is different from the main paper in that the intergenic region is displayed as two regions.

(B) The diploid copy number of the 9kb intergenic segment is plotted against the 6kb segment. Most individuals' copy numbers agree with the reference sequence arrangement, though some appear to have extra copies of the 9kb segment. This increase is offset by a concomitant decrease in 6kb copy number.

(C) Individuals with haplotype AH2 were selected out of the GPC cohort, and their other haplotype is plotted as a number (AH#). This shows more clearly that the 9kb region duplicates when the 6kb region is deleted. This may be indicative of the different breakpoints one would expect to see in a recurrently mutating locus.

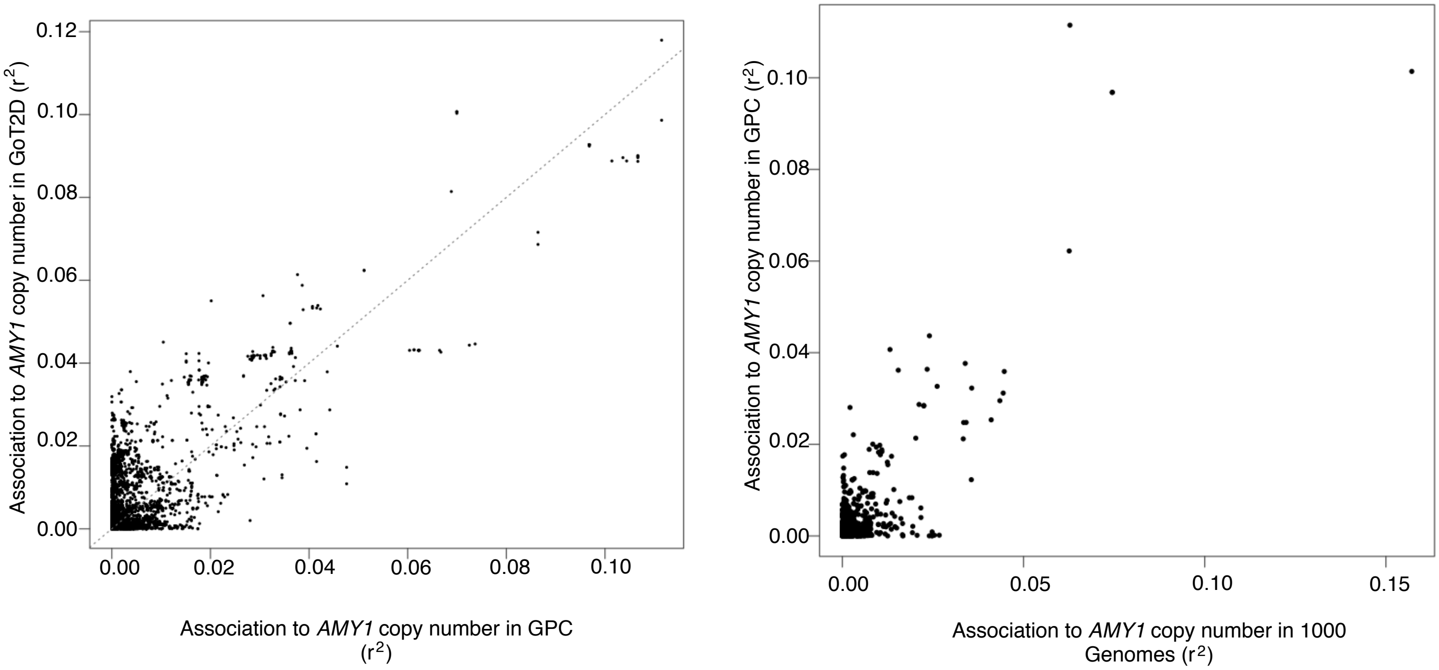
(D) The reference sequence is drawn above the proposed structure explaining the 9kb duplication and another structure we found. The structures are supported by the probing experiments of Groot *et al.* [1] and our optical mapping experiment with BioNano Genomics.



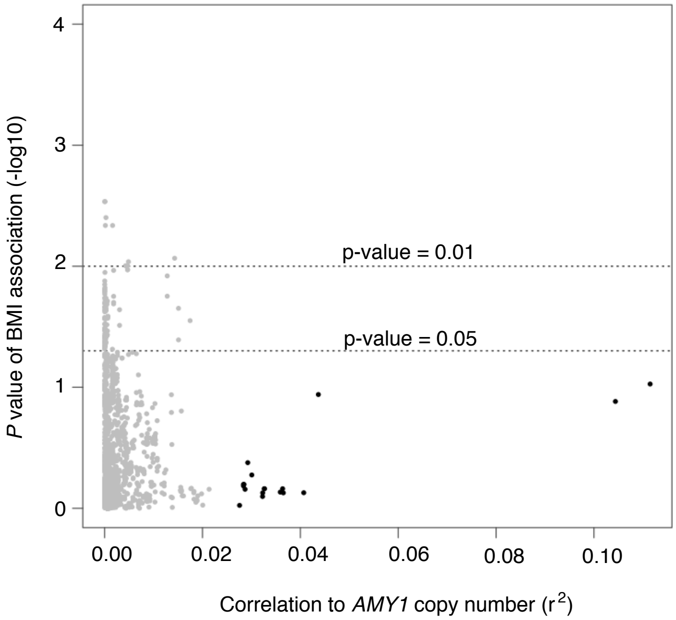
Supplementary Figure 8. Table of the haplotypes with the known diversity of the intergenic region listed. The evidence for each one's existence is listed to the right.

Structure assembled from:

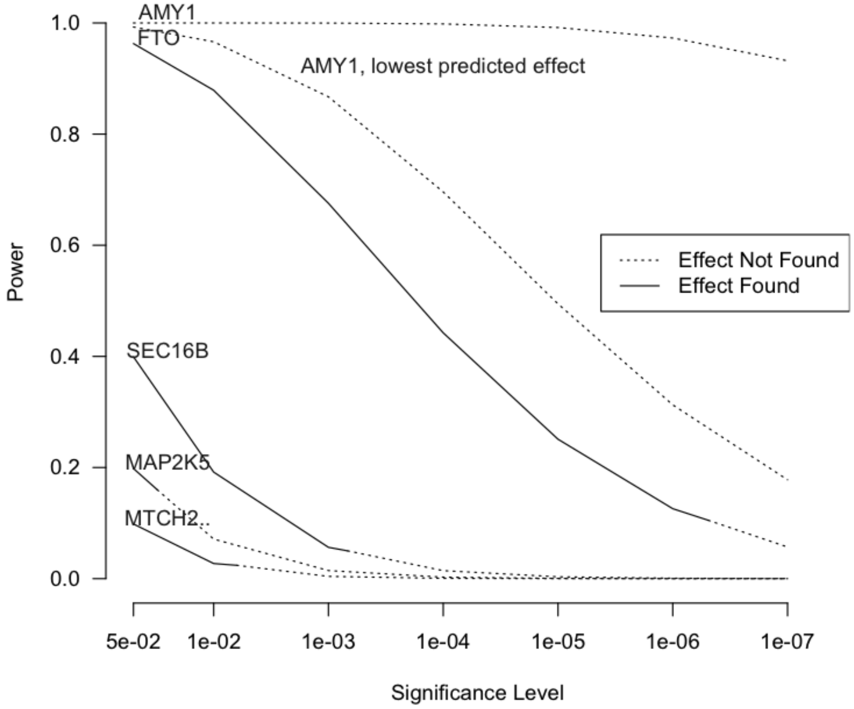
C=inference made with copy numbers
B=Bionano Optical Mapping
G=Groot *et al* [1] BAC assembly



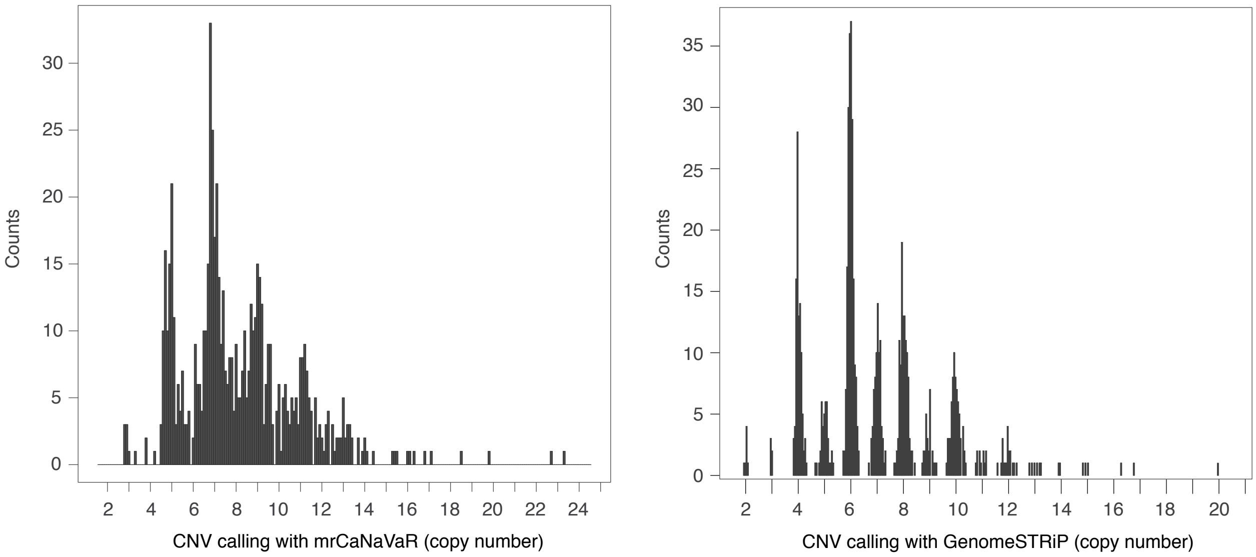
Supplementary Figure 9. Cross-cohort concordance of SNPs' association to *AMY1* copy number
AMY1 copy number was tested for association to SNPs in three cohorts, GPC, GoT2D, and the Europeans of 1000 Genomes. The three cohorts showed concordance for the r^2 values and effect sizes of overlapping SNPs..



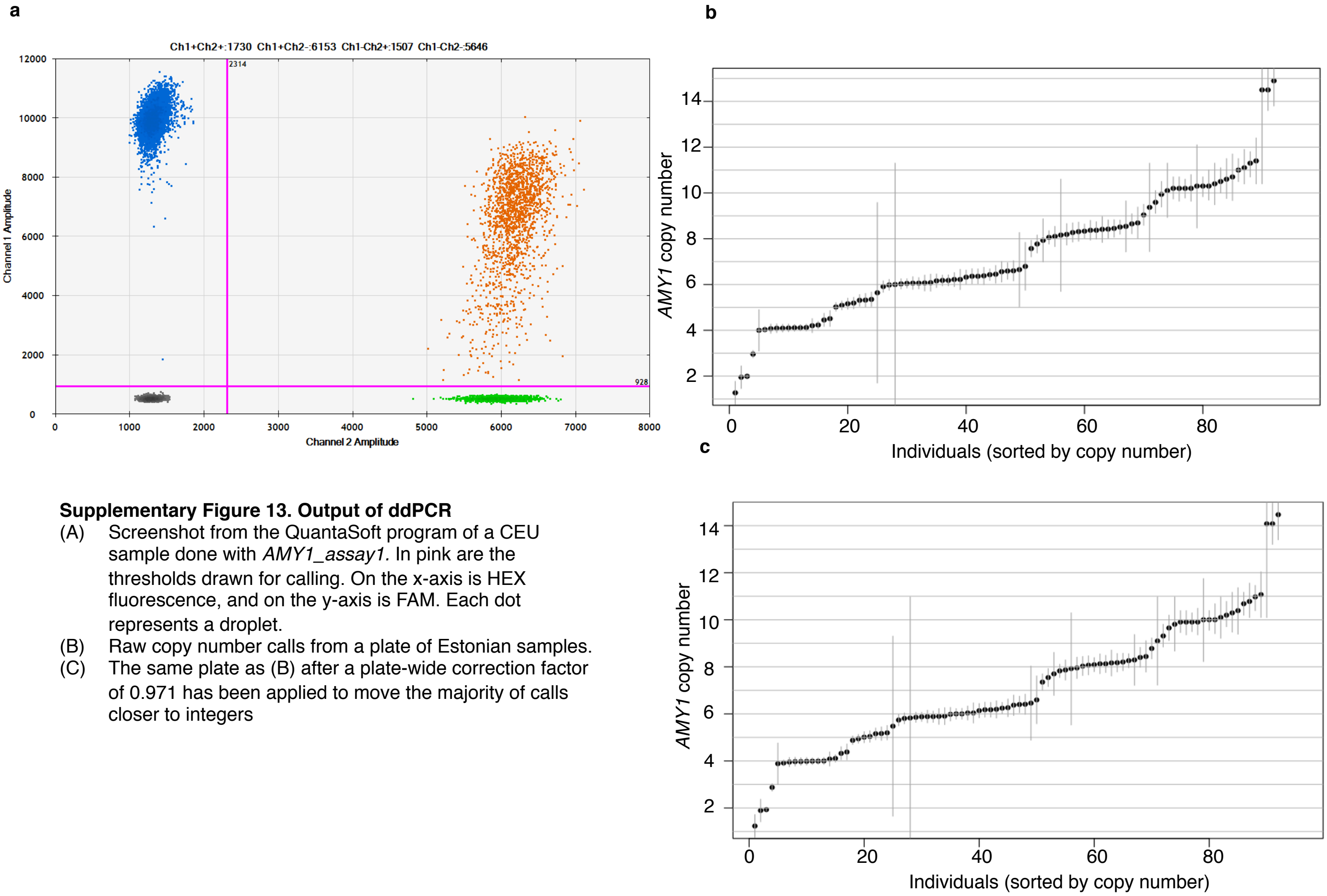
Supplementary Figure 10. SNPs' association to *AMY1* copy number and obesity
SNPs from the GIANT Consortium [2] were tested for *AMY1* correlation and plotted according to their *P* value in the GIANT BMI association. Correlation test one-tailed *P* value = 0.96. Points in black are significant for correlation to *AMY1* after Bonferroni correction.



Supplementary Figure 11: Power calculations for the detection of obesity-related SNPs in the Estonian cohort



Supplementary Figure 12. Comparison of CNV analysis programs using the same sequence data set. Copy number calling was done with GenomeSTRiP [3] in the same manner as in the GPC cohort, 1000 Genomes, and the GoT2D cohort. For mrCaNaVaR [4] calling was carried out as stated in the **Online Methods, alternative read-depth method**. Both programs were run on the same medium coverage sequencing data set (InCHIANTI). Out of necessary program compatibilities, GenomeSTRiP was run on BWA-aligned[10] data, and mrCaNaVaR was run on mrsFAST-aligned [5] data.

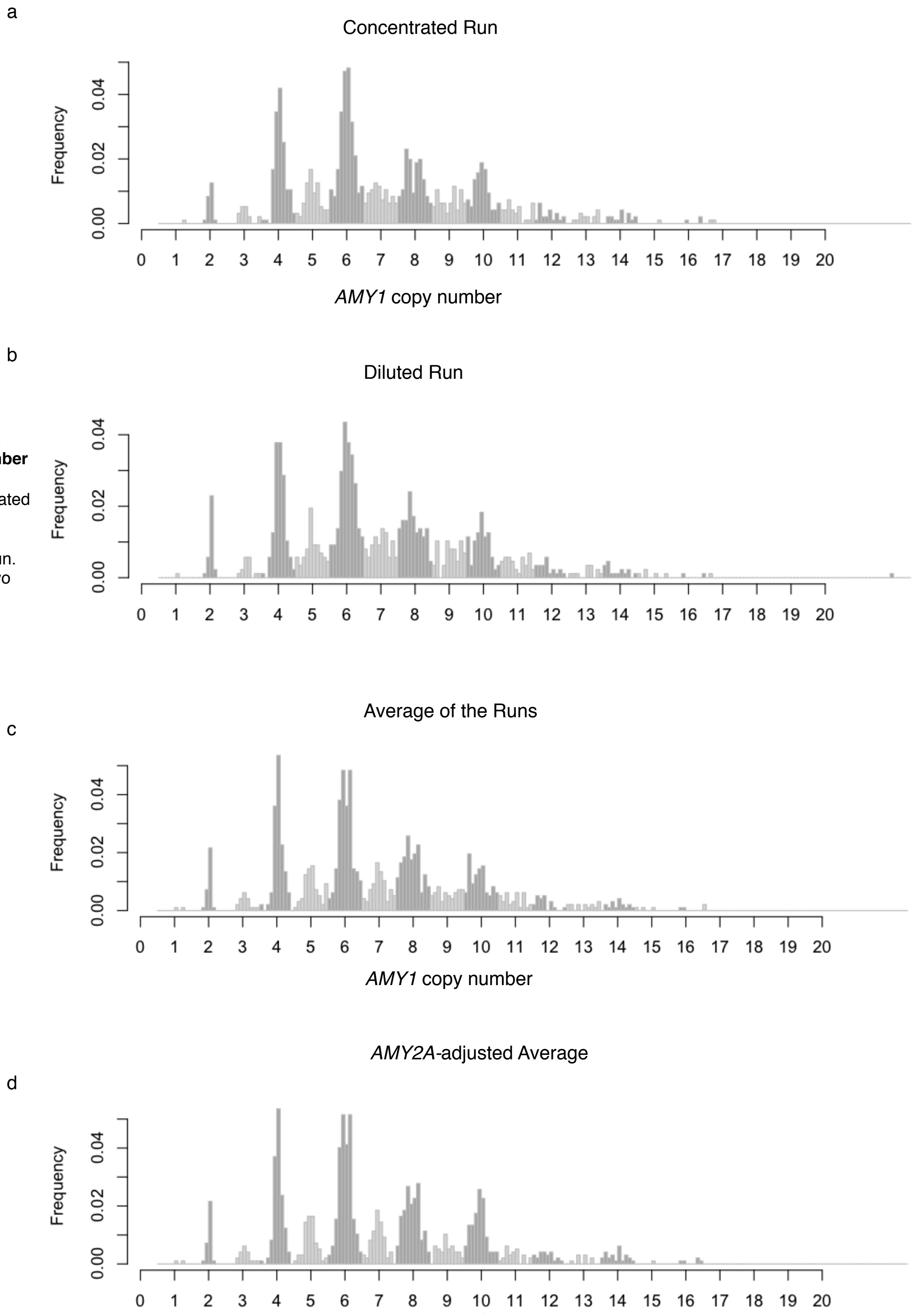


Supplementary Figure 13. Output of ddPCR

(A) Screenshot from the QuantaSoft program of a CEU sample done with *AMY1_assay1.* In pink are the thresholds drawn for calling. On the x-axis is HEX fluorescence, and on the y-axis is FAM. Each dot represents a droplet.

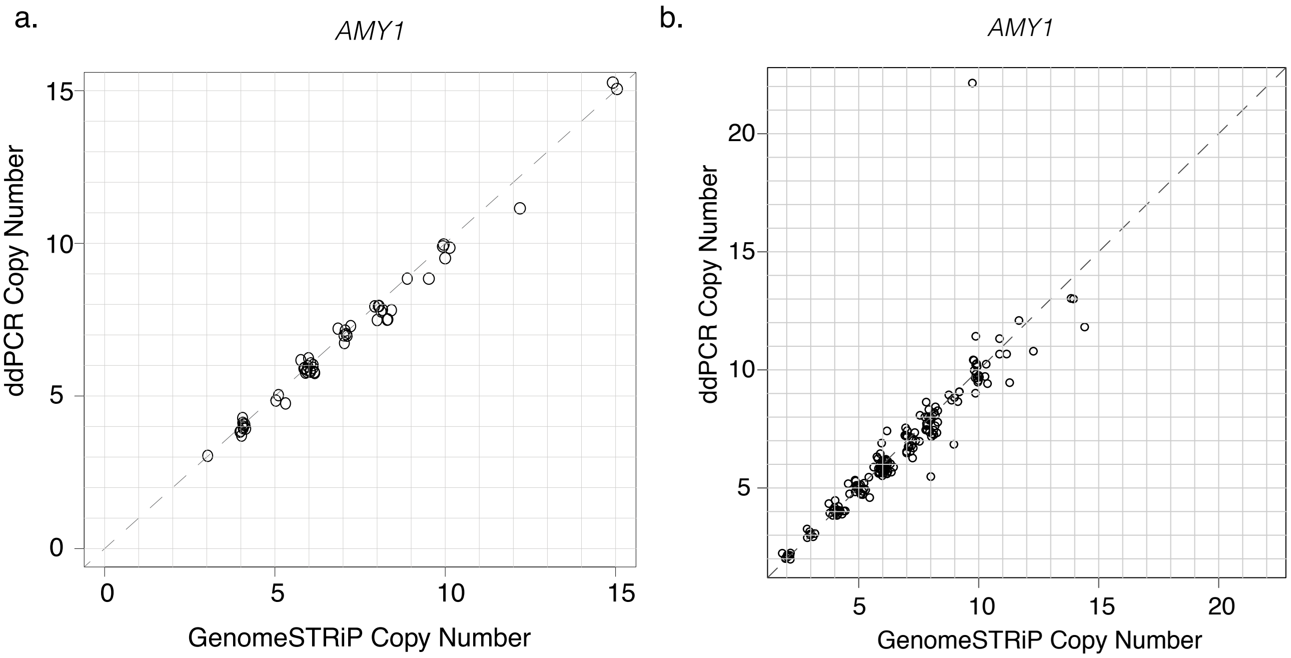
(B) Raw copy number calls from a plate of Estonian samples.

(C) The same plate as (B) after a plate-wide correction factor of 0.971 has been applied to move the majority of calls closer to integers

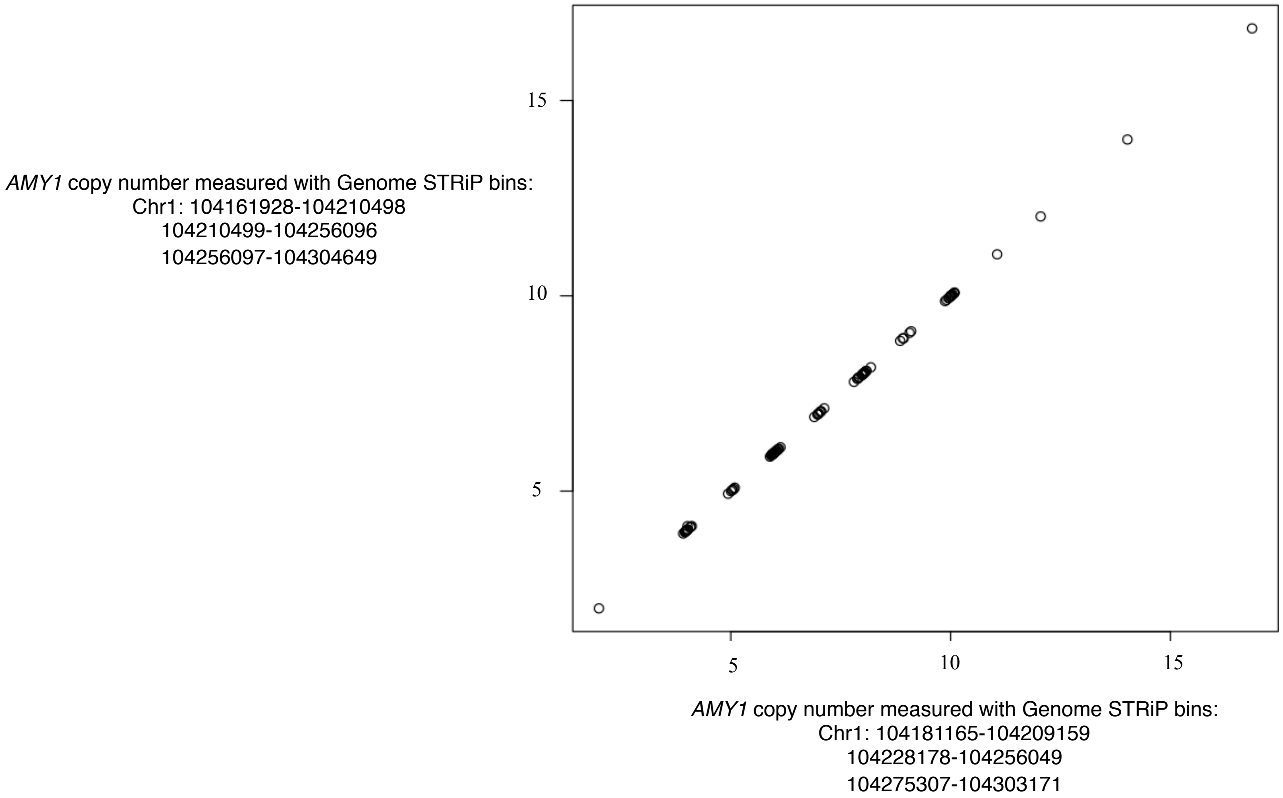


Supplementary Figure 14. Estonian *AMY1* copy number calls

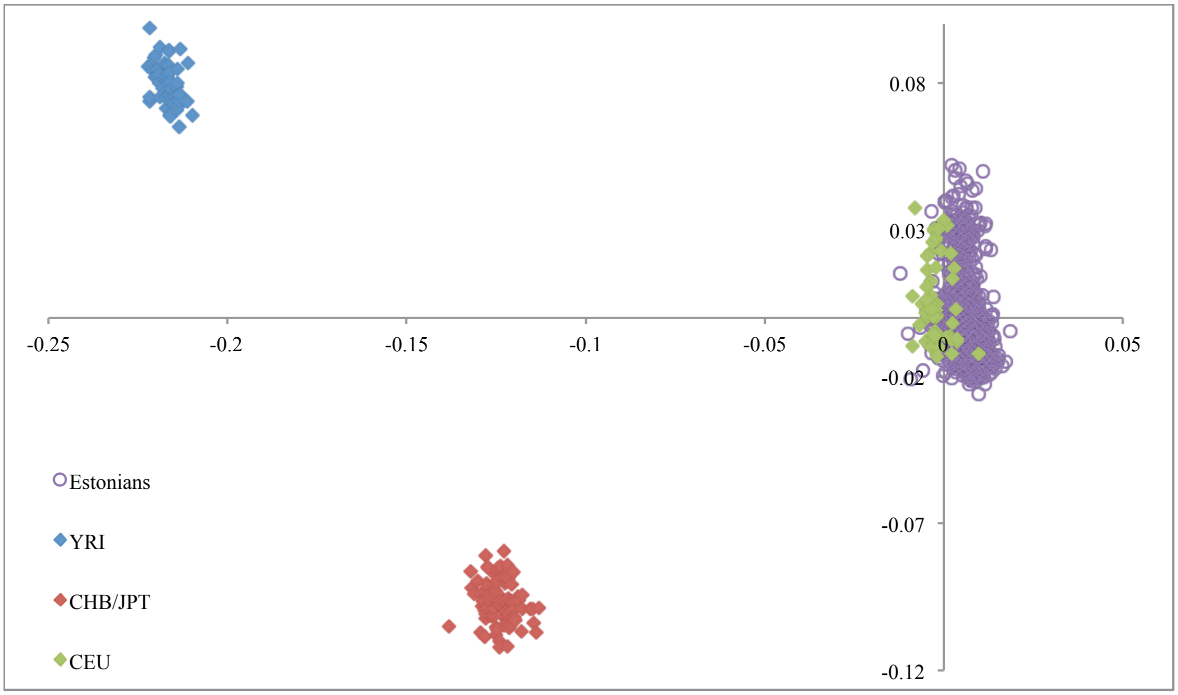
(A) Calls for the concentrated run (**Supplementary Note**).
(B) Calls for the diluted run.
(C) The average of the two runs.
(D) The *AMY2A*-adjusted average



Supplementary Figure 15: Concordance between ddPCR and GenomeSTRIP in InCHIANTI (a) and GoT2D (b) samples.



Supplementary Figure 16: Alternative read-depth bins for measuring *AMY1* copy number.
Shows that GenomeStrip selectively uses informative loci within each range



Supplementary Figure 17: Principle component analysis of the ancestry of the Estonians

References

1. Groot, P.C. *et al.* Evolution of the human alpha-amylase multigene family through unequal, homologous, and inter- and intrachromosomal crossovers. *Genomics* **8**, 97-105 (1990).
2. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
3. 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**, 269-76 (2011).
4. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-7 (2009).
5. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**, 576-7 (2010).