Recurrent rearrangements of human amylase genes create multiple independent CNV series

Nzar A.A. Shwan[1,3*], Sandra Louzada[2*], Fengtang Yang[2] and John A.L. Armour[1]

[* = co-first author]

1 School of Life Sciences, University of Nottingham, Medical School, Queen's Medical Centre, Nottingham NG7 2UH, UK

2 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

3 Scientific Research Centre, University of Salahaddin, Erbil, Kurdistan, Iraq

Corresponding author: John Armour (john.armour@nottingham.ac.uk)

Key words: genomic mutation; adaptation; genomic instability; CNV

Grant number: Wellcome Trust WT098051

Abstract

The human amylase gene cluster includes the human salivary (*AMY1,* MIM# 104700) and pancreatic amylase genes (*AMY2A*, MIM# 104650 and *AMY2B,* MIM# 104660), and is a highly variable and dynamic region of the genome. Copy number variation of *AMY1* has been implicated in human dietary adaptation, and in population association with obesity, but neither of these findings has been independently replicated. Despite these functional implications, the structural genomic basis of copy number variation (CNV) has only been defined in detail very recently. In this work we use high-resolution analysis of copy number, and analysis of segregation in trios, to define new, independent allelic series of amylase CNVs in sub-Saharan Africans, including a series of higher-order expansions of a unit consisting of one copy each of *AMY1*, *AMY2A* and *AMY2B*. We use fibre-FISH (fluorescence *in situ* hybridization) to define unexpected complexity in the accompanying rearrangements. These findings demonstrate recurrent involvement of the amylase gene region in genomic instability, involving at least five independent rearrangements of the pancreatic amylase genes (*AMY2A* and *AMY2B*). Structural features shared by fundamentally distinct lineages strongly suggest that the common ancestral state for the human amylase cluster contained more than one, and probably three, copies of *AMY1*.

Introduction

The adoption of agriculture was one of the most radical and pervasive innovations among the many changes introduced by humans to their own environments. In addition to a capacity to support higher population densities, agricultural food production has led to a shift in dietary composition, including increases in dietary starch as the result of reliance on starch-rich staples. Starch is initially digested by the enzyme amylase, present in humans in two tissue-specific isoenzymes: salivary amylase, encoded by the gene *AMY1*, and pancreatic amylase, encoded by *AMY2A* and *AMY2B*. These amylase genes are all found in a cluster on human chromosome 1, and early observations on pedigree segregation of protein electrophoretic variants demonstrated common and extensive copy number variation (CNV) in the salivary amylase gene *AMY1* [Pronk and Frants, 1979; Pronk et al., 1982]. These observations, coupled with detailed mapping of cloned genomic sequences, showed that there were common haplotypes containing odd numbers of *AMY1* genes, differing by pairs of genes in inverted orientation [Bank et al., 1992; Groot et al., 1989; Groot et al., 1991; Groot et al., 1990]. More recently, higher-resolution studies of the variation have demonstrated that most humans have an even number of *AMY1* copies, as predicted by the predominance of haplotypes containing odd numbers, with an overall copy number range of 2 to 18 copies per individual [Carpenter et al., 2015; Usher et al., 2015].

Primarily because of its early discovery and extensive range, most attention on amylase CNVs has focussed on the salivary amylase gene *AMY1*, but there have been reports of CNVs involving the *AMY2* genes [Conrad et al., 2010; Cooke Bailey et al., 2013; Groot et al., 1991; Sudmant et al., 2010]. Integration of information from read-depth analysis, segregation and direct typing of copy number demonstrated haplotypes harbouring even numbers of *AMY1* in conjunction with CNVs of the pancreatic amylase genes *AMY2A* and *AMY2B*. There are two common CNVs of *AMY2* genes in European populations – one carrying a deletion of the *AMY2A* gene, the other a duplication of both *AMY2A* and *AMY2B* [Carpenter et al., 2015; Usher et al., 2015]. Those investigations also implied that there were other rearrangements of the locus that could not be accounted for by the allelic series common in Europe.

The extensive variation in *AMY1* copy number has prompted studies exploring its functional significance, including the observation that populations with starch-rich diets appear to have significantly higher average *AMY1* copy number than populations with lower starch intake [Perry et al., 2007]. The implication that copy number expansion of *AMY1* represents an adaptation to dietary shifts following the adoption of agriculture fits with the observation that the gene is found as a single copy in chimpanzees [Perry et al., 2006], and in the genomes of archaic hominins [Lazaridis et al., 2014; Olalde et al., 2014]. More recently, the observation of a significant correlation between low *AMY1* copy number and higher body mass index (BMI) suggested that the CNV had considerable ongoing functional importance in modern humans [Falchi et al., 2014]. Although further studies have supported the association [Mejía-Benítez et al., 2015], doubt was cast on the medical importance of the association by the failure of a rigorous and well-powered study to reproduce the observation [Usher et al., 2015]. Most recently, a carefully calibrated study of *AMY1* copy number in East Asian samples also failed to demonstrate any association with BMI [Yong et al., 2016].

In this work we set out to understand more thoroughly the range of common genomic variation in amylase copy number found in humans, and in particular to define the potential range of CNVs of *AMY2* genes. We combine high-resolution DNA typing, fibre-FISH and SNP analysis to show that independent rearrangements of the *AMY2* genes have arisen on at least five occasions, and can include haplotypes containing up to 5 copies each of *AMY2A* and *AMY2B*. Although we cannot exclude neutral mutation processes at high frequency in this highly repetitive and unstable region, recurrent and human-specific rearrangements suggest the likelihood of adaptive value for these variants.

Materials and Methods

Amylase copy number determination

Previously published methods were used to measure relative representation of *AMY1*-coupled microsatellite alleles and ratios of *AMY2A*:*AMY2B* copy numbers [Carpenter et al., 2015]. *AMY1* copy number was measured by modified PRT approaches, in which distinctive sequence variants from the two terminal (centromeric) copies of *AMY1* ("*AMY1*C") were used as reference loci. Two fluorescent PCRs in a total volume of 10µl were done per sample, each using three primers at 1µM and 10ng genomic DNA in the buffer described [Carpenter et al., 2015], and switching the activities of primers using cycling conditions. The first PRT uses primers AMY1CF, HEX-AMY1CR and nested forward primer NF2, and the second contains AMY1CF, FAM-labelled AMY1CRB2 and nested forward primer NF5 (see Table 1).

Reactions started with 15 cycles of 95°C 30s/ 61°C 30s/ 65°C 2 minutes, during which AMY1CF and AMY1CR/RB2 anneal stably to make products specific to *AMY1*. The cycles then switched to 95°C 30s/ 54°C 30s/ 65°C 1 minute, for 14 (AMY1CR+NF2) or 13 (AMY1CRB2/NF5) cycles, before final extension at 72°C for 50 minutes; at the lower annealing temperature in the second phase the nested primers NF2 and NF5 anneal stably to make shorter products that are more readily resolved. PCR products were quantified after separation by capillary electrophoresis on an ABI3130xl Genetic Analyser 36 cm capillary, running the products from the two reactions in the same capillary. Before electrophoresis 2µl from reactions with AMY1CR/NF2 and 0.8µl from reactions with AMY1CRB2/NF5 were mixed in 10µl HiDi formamide containing 0.125µl ROX-500 markers. These samples were denatured at 96°C for 3 minutes before electrophoresis using POP- 7 polymer and an injection time of 30s at 1kV. GeneMapper software (Applied Biosystems) was used to extract peak area data.

In nearly all samples AMY1CR/NF2 amplify 436bp products from the two *AMY1C* copies and 427bp products from all other (*AMY1A/1B*) copies. AMY1CRB2/NF5 amplify 357bp products from typical copies of *AMY1C* and 344bp products from *AMY1A/1B*; a distinctive alternative product of 347bp is amplified from the variant *AMY1C* haplotype. Ratios of *AMY1A1B* to *AMY1C* can be used to deduce *AMY1* copy number, assuming that there are two copies of *AMY1C*, calibrating the data with integer clusters defined using k-means clustering. Further details and representative data can be found in the Supplementary Material and Supp. Figures S1-S5.

The assay for the *AMY2A*/*2B* duplication junction fragment [Carpenter et al., 2015] was modified to allow quantitative readout after capillary electrophoresis of fluorescent PCR products. PCRs of 10µl used 10ng genomic DNA in the buffer described [Carpenter et al., 2015], with final concentrations of 1µM of each of three primers AMY2B2D, FAM-AMY2B2R and AMY2B2F (Table 1).

PCRs used an initial denaturation stage of 95°C for 5 minutes, followed by 22 cycles of 95°C 30s/ 60°C 30s/ 65°C 1 minute, and final extension at 72°C for 50 minutes. Products of 192bp between AMY2BF and AMY2BR are made from all samples, and if it is present the duplication junction sequence produces a 176bp product between AMY2BD and AMY2BR. Before electrophoresis 1µl from PCRs was mixed with 10µl

HiDi formamide containing 0.125µl ROX-500 markers, and denatured and separated by capillary electrophoresis as above.

Copy number ratios for *AMY1* relative to *AMY2* (*AMY2A*+*AMY2B*) were determined by a PRT exploiting a consistent 4bp length difference in the paralogous products from just upstream of exon 4, using primers HEX- AMY1_2F and AMY1_2R (Table 1). The ratios of products from *AMY1* (169bp) to *AMY2A* + *AMY2B* (173bp) were used to infer the ratio of genomic copy numbers.

Fibre-FISH methods

The probes and general methods for fibre-FISH are given in detail in [Gribble et al., 2013] and [Carpenter et al., 2015]. In summary, DNA fibres were prepared from agarose-embedded cells by molecular combing (Genomic Vision), and probes were derived from one PCR product from the *AMY1* gene [Perry et al., 2007], and one each from the regions upstream of *AMY2A* and *AMY2B* [Carpenter et al., 2015].

Haplotype definition and database records

In the Leiden Open Variation Database format (LOVD, http://www.lovd.nl [Fokkema et al., 2011]), information defining structural allelic variants involving the *AMY1*, *AMY2A* and *AMY2B* genes is collected under the *AMY2B* locus-specific database (http://www.LOVD.nl/AMY2B). In this work we have given the *AMY2B* locus-specific database ID of each new or known haplotype – for example, the $(AMY1)_3(AMY2A)_1(AMY2B)_1$ haplotype found in the human reference assembly hg19 has the ID AMY2B_011111.

Results

Most haplotypes of the human amylase genes include one copy each of the *AMY2B* and *AMY2A* genes and an odd number of copies of *AMY1*. The differing *AMY1* copy numbers arise from variation in the numbers of a 95kb cassette including two copies of *AMY1* and one copy of the truncated *AMY2A* pseudogene "*AMYP1*" [Carpenter et al., 2015; Usher et al., 2015]. The arrangement of the sequence in the hg19 human reference assembly conforms to this pattern (locus-specific database (http://www.LOVD.nl/AMY2B) ID AMY2B_011111), with three copies of *AMY1*, and is illustrated in the upper panel of Figure 1. A common haplotype pattern not conforming to this structure has been described in recent work [Carpenter et al., 2015; Usher et al., 2015] (database ID AMY2B_022101); in this, *AMY2B*, *AMY2A* and one copy of *AMY1* are duplicated (via non-homologous rearrangement), creating a unique junction (shown as "J" in Figure 1) that can form the basis of a PCR assay for the structure [Carpenter et al., 2015]. Fibre-FISH confirmation of this structure for European sample GM12239 is shown in Supp. Figure S6. These *AMY2A2B* duplications are characteristically associated with haplotypes containing even numbers of *AMY1* (usually 4), and are common in European and African populations, but less so in East Asians [Carpenter et al., 2015; Usher et al., 2015]; in the nomenclature of Usher *et al*. [Usher et al., 2015], two examples are designated AH2B2 (AMY2B_022200) and AH4B2 (AMY2B_022211).

Higher-order expansions of pancreatic amylase genes

To understand the full scope of variation in human amylase genes, we aimed first to define the composition and structures of alleles containing more than two copies of each of the pancreatic amylase genes *AMY2A* and *AMY2B*. The gene content of haplotypes in Yoruban (YRI) trios from the HapMap phase 1 were determined first by measuring the gene copy numbers of *AMY1*, *AMY2A* and *AMY2B*, followed by analysis of segregation of *AMY1*-coupled microsatellite alleles [Carpenter et al., 2015], *AMY1*:*AMY2* ratios and *AMY2A*:*2B* ratios in trios. For most parental samples our direct measurements (Supplementary Dataset) were corroborated by read-depth measures [Carpenter et al., 2015]. For application in this work we developed a new PRT method to measure *AMY1* copy number based on the ratio between distinctive sequences at the centromeric (*AMY1*C) copy and the internal (*AMY1*A/B) copies; in practice, we found that this measure combined high levels of accuracy with the convenience of assigning most samples to integer classes with no more than two PCRs (Figure 2a and Supplementary Material). In parallel, we modified our assay for the junction sequence specific to the *AMY2A2B* duplication allele, to allow quantification of that sequence relative to the diploid genome (Figure 2b).

In most cases (see Figure 3 and Table 2) measurement of copy numbers and ratios in Yoruban trios allowed deduction of the likely haplotype composition. The copy number data were consistent with analyses based on read-depth from the 1000 Genomes Project [Carpenter et al., 2015; Usher et al., 2015], and demonstrate that there are distinctive haplotypes associated with higher-order amplifications of *AMY2A* and *AMY2B*, including triplication, quadruplication and quintuplication (AMY2B_033201/044301/055401); in nearly all cases, alleles carrying higher-order expansions of *AMY2A* and *AMY2B* are predicted to carry equal numbers of *AMY1*, *AMY2A* and *AMY2B* genes, so that (for example) the untransmitted maternal quintuplication allele in family Y056 has the composition $(AMY1)_5(AMY2A)_5(AMY2B)_5$ (AMY2B_055401, Table 2). Quantification of product ratios showed that n-fold expansions of (*AMY2B*-*AMY2A*-*AMY1*) contained (n-1) copies of the junction sequence found in the *AMY2A+2B* duplication allele series ([Carpenter et al., 2015]; AMY2B_022101 above, or AMY2B_022200 and AMY2B_022211, equivalent to alleles AH2B2 and AH4B2 in [Usher et al., 2015]). This observation suggested that there is a new allelic series based on higher expansion of the repeat unit formed in the *AMY2A+2B* duplication allele, with the known junction sequence separating adjacent copies of an (*AMY2B-AMY2A-AMY1*) repeat unit.

We applied fibre-FISH to define the physical structure of the haplotypes we had defined on the basis of gene content; our previous observations demonstrated [Carpenter et al., 2015] that although the high level of sequence similarity between amylase gene sequences leads to cross-hybridization, especially between *AMY1* and *AMY2A*, it is nevertheless possible to distinguish the *AMY1* and *AMY2A* genes on the basis of hybridization patterns (Figure 3, top). By contrast, the sequence upstream of *AMY2B* is sufficiently distinct to give locus-specific hybridization. Fibre-FISH analysis of expanded alleles verified the prediction of a repeat unit containing one copy of each gene, but also showed that in all cases the first (telomeric) unit contained an inversion, to give the gene order (*AMY2B-AMY1-AMY2A*)-(*AMY2B-AMY2A-AMY1*)$_{(n-1)}$. This observation suggested the detailed structure for the triplication allele (AMY2B_033201) in family Y060 (Table 3) shown in Figure 3. The inversion of the first telomeric unit is also seen in fibre-FISH analysis of quadruplication and quintuplication alleles (Supp. Figures S7 and S8), but escaped detection by optical mapping [Usher et al., 2015]. We used long PCR and Sanger sequencing (Supplementary Material) to amplify a 9.8kb product across this inversion in the quintuplication (AMY2B_055401) carrier NA19159 (GenBank KX394682). This sequence verified the orientations shown in Figure 3, but demonstrated no further rearrangements or sequence variants unique to this structure.

Alleles containing higher-order (n ≥ 3) expansions of *AMY2A* and *AMY2B* were examined by fibre-FISH (4 examples), segregation (4 examples) and analysis of 1000 Genomes Project read-depth (12 examples of AFR individuals with more than 3 copies of both *AMY2A* and *AMY2B*); these alleles appeared to be coherent for general structure, gene content and SNP associations. All examples (AMY2B_033201/044301/055401) of higher-order amplifications of the unit (*AMY2B*-*AMY2A*-*AMY1*) were associated (D' = 1) in African populations with a common haplotype tagged by (for example) the derived allele rs12075086T, the same haplotype associated with simple duplication (AMY2B_022101) of *AMY2A* and *AMY2B* in worldwide populations [Carpenter et al., 2015; Usher et al., 2015]. Consistent with this conclusion of a single origin for all alleles containing amplifications of both *AMY2A* and *AMY2B*, all contained the same junction sequence (see Methods), and the deduced *AMY1* microsatellite allele content of expanded alleles resembled each other, and those of the duplication allele AMY2B_022101, with a predominance of microsatellite alleles yielding PCR products of 269bp (Supplementary Dataset and [Carpenter et al., 2015]).

Duplication of AMY2A

Our previous work and that of others demonstrated individuals (and therefore haplotypes) with higher numbers of *AMY2A* than *AMY2B* [Carpenter et al., 2015; Usher et al., 2015]. Such individuals are more frequent in African populations than others; for example, in our read-depth analysis of 1000 genomes samples, 13.6% of

African samples had more copies of *AMY2A* than *AMY2B*, compared with 2.51% of Asians and 0.55% of Europeans [Carpenter et al., 2015]. Segregation analysis in African (YRI) trios confirmed the prediction that the corresponding haplotypes carried a duplication of *AMY2A* unaccompanied by duplication of *AMY2B* (Table 3). As predicted for independently-arising duplications, they were not associated with the specific junction fragment characteristic of the *AMY2A+2B* duplication haplotype. Analysis of SNP associations in these individuals suggest that most examples of *AMY2A*-only duplications were found on a single haplotype background, but there was also evidence of heterogeneity, with (for example) NA19119, who has both haplotypes with an *AMY2A*-only duplication (AMY2B_012341 and AMY2B_012211) on two different SNP haplotypes. We undertook fibre-FISH analysis in family trio Y060 (Table 3), in which both haplotypes in the father NA19119 were predicted to have 2 copies of *AMY2A* and a single copy of *AMY2B* (Figure 3 and Supp. Figure S9).

The haplotypes characterised have structures that do not require the formation of new junctions, and can be created by new juxtapositions of sequences present in the reference haplotype AMY2B_011111. However, the structural differences indicate that these two haplotypes (AMY2B_012341 and AMY2B_012211) arose independently of one another,, and that the gene content feature common to these two haplotypes, amplification of *AMY2A* without amplification of *AMY2B*, appears coincidental rather than as the result of common ancestry. In particular, the shorter $(\textit{AMY1})_4(\textit{AMY2A})_2(\textit{AMY2B})_1$ haplotype (AMY2B_012211) has the duplicated copy of *AMY2A* in inverted orientation (Supp. Figure S9). Comparison of amylase copy number with flanking SNP haplotypes suggests that this $(\textit{AMY1})_4(\textit{AMY2A})_2(\textit{AMY2B})_1$ haplotype (AMY2B_012211) is the commonest type of $(\textit{AMY2A})_2(\textit{AMY2B})_1$ structure, whereas we found no evidence for other alleles corresponding to the longer (*AMY1*)$_8$(*AMY2A*)$_2$(*AMY2B*)$_1$ haplotype (AMY2B_012341) in NA19119.

A new AMY1:AMY2A junction

Our analysis of copy number segregation in family Y072 consistently indicated ambiguous copy number of *AMY1* in the mother NA19152 and her child NA19154 (Table 3); measures of *AMY1* based on the microsatellite (upstream of the *AMY1* gene) indicated 3 copies in the transmitted maternal haplotype, but only 2 copies based on the (downstream) PRT, and intermediate values based on read depth analysis of 1000 Genomes Project reads from the mother NA19152 (estimates of 5.2 and 5.55 from [Carpenter et al., 2015] and [Usher et al., 2015] respectively). Segregation also indicated that this haplotype (AMY2B_023201) contained 3 copies of *AMY2A* and 2 copies of *AMY2B* (Table 3). Fibre-FISH analysis confirmed the overall composition of the haplotype, but also demonstrated a hybrid structure with a copy of *AMY2A* and its upstream sequence immediately interrupting one copy of *AMY1* (Supp. Figure S10). Examination of 1000 Genomes Project data from NA19152 showed a single read conforming to this hybrid junction, from which PCR primers were used to demonstrate that the new junction interrupted *AMY1* in exon 4, with 3bp microhomology at the breakpoint (GenBank KX230759).

Discussion

Our work shows that pancreatic amylase (*AMY2A*/*2B*) genes appear to have undergone at least five independent rearrangements to create new copy numbers in humans since the split from chimpanzees. The first is a seamless deletion of *AMY2A* (AMY2B_010011) common in Europeans and to a lesser extent in Africans, and generally found on a single SNP background [Carpenter et al., 2015; Usher et al., 2015]. The second is a duplication of *AMY2A* and *AMY2B* common in Europeans and Africans (AMY2B_022101/022200/022211), that results from a non-homologous duplication of an *AMY2B*/*AMY2A*/*AMY1* unit, again associated with a common SNP background [Carpenter et al., 2015; Usher et al., 2015]. Our data in this study show that this *AMY2A*/*2B* duplication rearrangement was the starting-point for higher-order homologous expansions of *AMY2A*/*2B* found in African populations, as exemplified by the triplication, quadruplication and quintuplication haplotypes (AMY2B_033201/044301/055401) we have characterised (Figure 3, Supp. Figures S7 and S8). There are at least two further lineages (AMY2B_012211/AMY2B_012341) with independent homologous exchanges resulting in duplication of *AMY2A* without concomitant duplication of *AMY2B* (Figure 3, Supp. Figure S9), and finally a fifth (non-homologous) rearrangement in which one copy of *AMY1* is interrupted at exon 4 by a duplication of *AMY2A* (AMY2B_023201, Supp. Figure S10).

In addition to these rearrangements involving *AMY2A/2B*, allelic series differing by the 95kb unit containing two repeats of *AMY1* create further overall structural diversity [Carpenter et al., 2015; Usher et al., 2015]. To summarise the different mechanisms that have operated in the generation of diversity at this locus in humans, there have been apparently homologous deletions or duplications of the

95kb (*AMY1A-AMY1B-AMYP1*) unit, and unequal recombination between homologous repeats 75kb apart is involved in the generation of the *AMY2A* deletion allele. By contrast, the duplication of the 116kb (*AMY2B-AMY2A-AMY1*) unit shows no evidence of being mediated by sequence similarity. Once the duplication is established, however, the generation of higher-order repeats of the (*AMY2B-AMY2A-AMY1*) unit could be generated by unequal exchanges between cognate sequences in 116kb repeat sequences. Without complete allele sequences, however, it is difficult to exclude the possibility that additional complexity is involved in some of the apparently simple exchanges between repeats. From a methodological standpoint it is noteworthy that some features of our findings, including the overall structure of the haplotypes, could not be defined using short-read sequencing alone. Long-read capabilities exceeding 10kb would be needed to resolve features, such as the inversion accompanying higher order expansion of *AMY2A* and *AMY2B*, which are clearly demonstrated by fibre-FISH (Figure 3, Supp. Figures S7 and S8), and even then it is unlikely that the overall spatial organisation of the 116kb (*AMY2B-AMY2A-AMY1*) units could be reconstructed unambiguously by primary read assembly, especially if both haplotypes in an individual were of unknown structure.

Where genomic rearrangements involve repeated sequences across scales refractory to direct characterisation by sequence assembly of short fragments, longer-range methods such as fibre-FISH or pulsed-field gel electrophoresis can be used to establish haplotype structures. As demonstrated here, fibre-FISH, using combed single-molecular DNA fibres, enabled us to resolve the order, orientation and copy number of amylase family genes on each haplotype unambiguously, even without the need to analyse all three members of each family trio. However, these methods do not provide detailed DNA sequence information. Large-insert (fosmid or BAC) cloning can be used to recover both DNA sequence and information about long-range spatial organisation; it still remains particularly difficult to reconstruct full haplotype sequences when there is population structural allelic variation, as in some disease-associated rearrangements at structurally variable sites, such that in any one sample *both* copies are of unknown structure, for example [Carvalho and Lupski, 2008; Yuan et al., 2015], and this study.

Including the well-established allelic series differing in copy numbers of *AMY1* (AMY2B_011111/AMY2B_011100/AMY2B_011122, etc.) [Carpenter et al., 2015; Groot et al., 1989; Groot et al., 1990; Usher et al., 2015], it is clear that structural diversity at the human amylase locus has arisen by both homologous and non-homologous events, and has involved rearrangements of both the salivary (*AMY1*) and pancreatic (*AMY2A* and *AMY2B*) amylase genes. The spread of independently-arising rearrangements of the locus can be seen as consistent with the proposal that higher copy-number alleles have been selectively advantageous specifically in recent human history, as suggested by apparently recent human-specific amplification from the single-copy state represented in modern chimpanzees and the genomes of archaic hominins [Lazaridis et al., 2014; Olalde et al., 2014; Prufer et al., 2014].

However, it is noteworthy that all the major allelic series of human amylase CNVs defined to date share evidence of the rearrangement that gave rise to the inverted copy of *AMY1* ("*AMY1B*") and the corresponding intergenic region ("18kb" in Figure 1), suggesting that the ancestral state for modern humans must have had multiple copies of *AMY1*. The amylase cluster is a region of late-replicating DNA, and is therefore predicted to be prone to frequent rearrangement [Koren et al., 2012; Usher et al., 2015]. Germline mutation to create new copy number alleles cannot be scored simply by observing copy number mismatch in family trios, and first requires enough segregation information to define the parental haplotypes unambiguously; we have nevertheless screened 440 microsatellite haplotype transmissions in three-generation (CEPH) pedigrees without observing any changes in copy number state, suggesting a germline mutation frequency below 0.7% (J.A. and Andrew Cubbon, unpublished work). Given the appearance of similar structures on diverse modern human haplotype backgrounds, the most recent common ancestral state of the locus for all humans is likely to have contained not one copy of each gene, as found in chimpanzees, but instead a sequence similar to the hg19 reference assembly structure $(\mathit{AMY1})_3(\mathit{AMY2A})_1(\mathit{AMY2B})_1$ (AMY2B_011111, equivalent to “H1” of Groot *et al.* [Groot et al., 1989; Groot et al., 1990], or “AH3” of Usher et al. [Usher et al., 2015]). This structure already contained both inverted and tandem-repeated sequences that could predispose to further recurrent rearrangement in the germline, and was itself the result of a non-homologous rearrangement.

If an $(\mathit{AMY1})_3$ allele was the common ancestral structure for all modern humans, the initial amplification to higher gene copy number may have been selectively advantageous before the neolithic, consistent with a recent analysis of sequence data [Inchley et al., 2016]. Nevertheless, whether adaptive or neutral, a preneolithic expansion to higher copy number does not itself preclude subsequent adaptive value for copy number change after the neolithic [Perry et al., 2007].

We present no association data relevant to the potential influence of this CNV on obesity, but our results still have implications for the design and interpretation of such studies. Specifically, the expansions of *AMY2* genes we describe here suggest that any influence of amylase gene copy number on body fat is likely to have different genetic architecture in individuals of recent African ancestry. More generally, the extensive structural allelic diversity at the amylase CNV emphasises the extreme difficulty of imputing allelic diversity from SNP data, or of reconstructing structural alleles based on short-read sequence data.

Acknowledgements

We are grateful to Beiyuan Fu for technical assistance. This work was supported by the University of Nottingham, and by the Wellcome Trust (grant number WT098051). N.A.A.S. is supported by the Higher Committee for Education Development (HCED), Iraq. The authors have no conflicts of interest to declare.

References

Bank RA, Hettema EH, Muijs MA, Pals G, Arwert F, Boomsma DI, Pronk JC. 1992. Variation in gene copy number and polymorphism of the human salivary amylase isoenzyme system in Caucasians. Hum Genet 89(2):213-222.

Carpenter D, Dhar S, Mitchell L, Fu B, Tyson J, Shwan N, Yang F, Thomas MG, Armour JAL. 2015. 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.

Carvalho CMB, Lupski JR. 2008. Copy number variation at the breakpoint region of isochromosome 17q. Genome Res 18(11):1724-1732.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell PJ, Fitzgerald T, Hu M et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704-712.

Cooke Bailey JN, Lu L, Chou JW, Xu J, McWilliams DR, Howard TD, Freedman BI. 2013. The role of copy number variation in African Americans with type 2 diabetes-associated end stage renal disease. J Mol Genet Med 7:61.

Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, Sudmant PH, Dorajoo R, Al-Shafai MN, Bottolo L, Ozdemir E, So H-C et al. 2014. Low copy number of the salivary amylase gene predisposes to obesity. Nat Genet 46:492-497.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32(5):557-563.

Gribble SM, Wiseman FK, Clayton S, Prigmore E, Langley E, Yang F, Maguire S, Fu B, Rajan D, Sheppard O, Scott C, Hauser H et al. 2013. Massively Parallel Sequencing Reveals the Complex Structure of an Irradiated Human Chromosome on a Mouse Background in the Tc1 Model of Down Syndrome. PLoS ONE 8(4):e60482.

Groot PC, Bleeker MJ, Pronk JC, Arwert F, Mager WH, Planta RJ, Eriksson AW, Frants RR. 1989. The Human Alpha-Amylase Multigene Family Consists of Haplotypes with Variable Numbers of Genes. Genomics 5(1):29-42.

Groot PC, Mager WH, Frants RR. 1991. Interpretation of polymorphic DNA patterns in the human alpha-amylase multigene family. Genomics 10(3):779-785.

Groot PC, Mager WH, Henriquez NV, Pronk JC, Arwert F, Planta RJ, Eriksson AW, Frants RR. 1990. Evolution of the human alpha-amylase multigene family through unequal, homologous, and interchromosomal and intrachromosomal crossovers. Genomics 8(1):97-105.

Inchley CE, Larbey CDA, Shwan NAA, Pagani L, Saag L, Antão T, Jacobs G, Hudjashov G, Metspalu E, Mitt M, Eichstaedt CA, Malyarchuk B et al. 2016. Selective sweep on human amylase genes postdates the split with Neanderthals. Scientific Reports 6:37198.

Koren A, Polak P, Nemesh J, Michaelson Jacob J, Sebat J, Sunyaev Shamil R, McCarroll Steven A. 2012. Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. Am J Hum Genet 91(6):1033-1040.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, Berger B, Economou C et al. 2014.

Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513(7518):409-413.

Mejía-Benítez M, Bonnefond A, Yengo L, Huyvaert M, Dechaume A, Peralta-Romero J, Klünder-Klünder M, García Mena J, El-Sayed Moustafa J, Falchi M, Cruz M, Froguel P. 2015. Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican children. Diabetologia 58(2):290-294.

Olalde I, Allentoft ME, Sanchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, Prado-Martinez J, Rodriguez JA, Rasmussen S, Quilez J, Ramirez O, Marigorta UM et al. 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature 507(7491):225-228.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256-1260.

Perry GH, Tchinda J, McGrath SD, Zhang JJ, Picker SR, Caceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, Stone AC, Lee C. 2006. Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci USA 103(21):8006-8011.

Pronk JC, Frants RR. 1979. New genetic variants of parotid salivary amylase. Hum Hered 29(3):181-186.

Pronk JC, Frants RR, Jansen W, Eriksson AW, Tonino GJM. 1982. Evidence for duplication of the human salivary amylase gene. Hum Genet 60(1):32-35.

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505(7481):43-49.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE, Project G. 2010. Diversity of Human Copy Number Variation and Multicopy Genes. Science 330(6004):641-646.

Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, Cao H, Moon JE, Kashin S, Fuchsberger C, Metspalu A, Pato CN et al. 2015. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nat Genet 47(8):921-925.

Yong RYY, Mustaffa SAB, Wasan PS, Sheng L, Marshall CR, Scherer SW, Teo Y-Y, Yap EPH. 2016. Complex Copy Number Variation Of Amy1 Does not Associate With Obesity in Two East Asian Cohorts. Hum Mutat 37(7):669-678.

Yuan B, Liu P, Gupta A, Beck CR, Tejomurtula A, Campbell IM, Gambin T, Simmons AD, Withers MA, Harris RA, Rogers J, Schwartz DC et al. 2015. Comparative Genomic Analyses of the Human NPHP1 Locus Reveal Complex Genomic Architecture and Its Regional Evolution in Primates. PLoS Genet 11(12):e1005686.

Figure legends

Figure 1. Structures of the reference allele and an allele carrying the *AMY2A+2B* duplication

Most alleles at the human amylase locus conform to the general structure exemplified by the sequence in the human reference assembly (AMY2B_011111, upper diagram), with one copy each of *AMY2A* and *AMY2B*, and an odd number of copies of *AMY1* (in this case 3). Other members of this allelic series, with odd numbers of copies of *AMY1*, have different numbers (including none) of the 95kb unit shown, containing two copies of *AMY1* and the *AMY2A* pseudogene designated *AMYP1*. The lower diagram shows, on the same scale, the simplest example of a structure containing the *AMY2A+2B* duplication (AMY2B_022101); duplication of a 116kb sequence encompassing *AMY2B*, *AMY2A* and one copy of *AMY1* leads to the formation of a haplotype with 2 copies each of *AMY2B*, *AMY2A* and *AMY1*. Other members of this same allelic series can contain higher even numbers of *AMY1*, again differing in numbers of the 95kb (*AMY1*)$_2$–*AMYP1* unit shown above. Note that the non-homologous duplication is accompanied by the formation of a specific sequence junction between sequences upstream of *AMY2B* and the 18kb repeat sequence between *AMY1A* and *AMY1B*, indicated here by “J” [Carpenter et al., 2015; Usher et al., 2015].

Figure 2. New experimental methods for high-resolution measurement of *AMY1* copy number and *AMY2A*/*2B* duplication

(a) Results from 854 *AMY1* copy number assays, each with two measurements of *AMY1* copy number using the NF2 and NF5 variants of the *AMY1*C PRT assay (see Materials and Methods). The appearance of clear clusters allows the confident assignment of nearly all samples to integer copy numbers based on these two PCRs, especially at copy numbers below 10. (b) Quantification of the junction fragment for the *AMY2A*/*2B* duplication allele and its derivatives, measuring the representation of a PCR product from the specific duplication junction fragment (“dup”) relative to a control product present in two copies in every individual. Traces are shown from NA18854, NA18859, NA19116 and NA19200, with 0, 1, 2 and 3 copies of the duplication junction respectively.

Figure 3. Segregation of amylase haplotypes in family trio Y060 demonstrated by microsatellite and fibre-FISH analysis

The first table summarises measured copy numbers for *AMY1*, *AMY2A*, *AMY2B* and the junction sequence in this family (see also Table 3). The microsatellite allele profiles demonstrate the split of the total *AMY1* copy number between the different allele lengths (for example, the 12 copies of the father NA19119 are split 1 + 4 + 7). There are four possible segregation patterns for this trio logically compatible with the total copy numbers and whole-number splits. The untransmitted allele in the mother NA19116 carries one copy each of *AMY2A* and *AMY2B*, and is therefore strongly predicted to have an odd number of copies of *AMY1*. Only two of the four possible segregation patterns have an odd number of *AMY1* in the untransmitted maternal allele, and both of those involve transmission of 3 copies from the mother NA19116, and 8 copies from the father NA19119 (Table 3). One of those compatible segregation patterns is indicated here by the arrows and numbers. These analyses together suggest the haplotype segregation shown in the lower table, with transmitted alleles shown in orange (AMY2B_012341, paternal) and blue (AMY2B_033201, maternal).

In fibre-FISH analysis, the *AMY2B* probe employed (green) is specific to sequence upstream of *AMY2B*. The probe (red) for the sequence upstream of *AMY2A* cross-hybridizes with very similar sequence surrounding the ERV upstream of *AMY1,* and the *AMY1* gene probe (white) also cross-hybridizes with coding regions of *AMY2A* and *AMY2B*. In many locations, this additional cross-hybridization between similar amylase sequences provides useful confirmation of the type and orientation of the gene. Examples of hybridization observed with these three probes with *AMY1*, *AMY2A* and *AMY2B* are shown in the top panel. The orange box frames images from the haplotype AMY2B_012341 transmitted from father to child, with the composition $(AMY1)_8(AMY2A)_2(AMY2B)_1$, including a duplicated copy of *AMY2A* in the forward orientation preceded by 3 copies of *AMY1*. The reconstructed interpretation of the ≈490kb structure appears to be seamless, in that it includes no new short-range junctions, but the overall arrangement suggests that it arose independently of the untransmitted paternal $(AMY1)_4(AMY2A)_2(AMY2B)_1$ allele (AMY2B_012211) shown in Supp. Figure S9. In the blue box, full-length haplotype images from the triplication allele (AMY2B_033201) transmitted from mother to child (Table 3) are shown above the inferred gene arrangement, and finally the full (≈300kb) haplotype reconstruction. “J” shows the inferred positions of the duplication junction sequence, and the boxed region highlights the inversion of one copy each of *AMY1* and *AMY2A* relative to the reference assembly orientation.

Figure 1

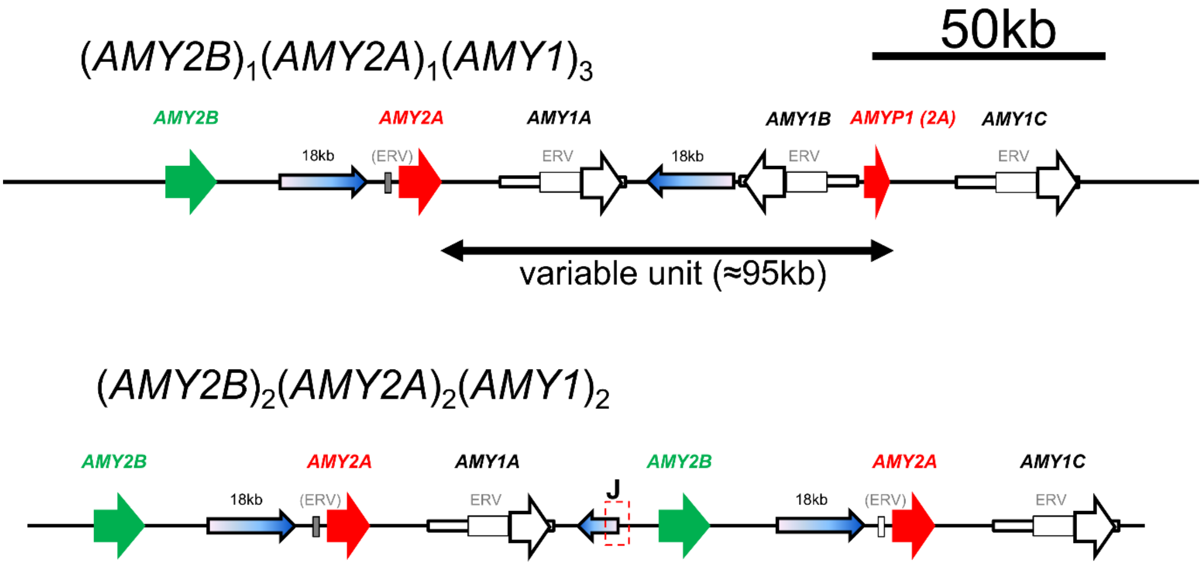

50kb
(AMY2B)1(AMY2A)1(AMY1)3
AMY2B
AMY2A
AMY1A
AMY1B
AMYP1 (2A)
AMY1C
18kb
(ERV)
ERV
18kb
ERV
ERV
variable unit (≈95kb)
(AMY2B)2(AMY2A)2(AMY1)2
AMY2B
AMY2A
AMY1A
AMY2B
AMY2A
AMY1C
18kb
(ERV)
ERV
J
18kb
(ERV)
ERV

Figure 2

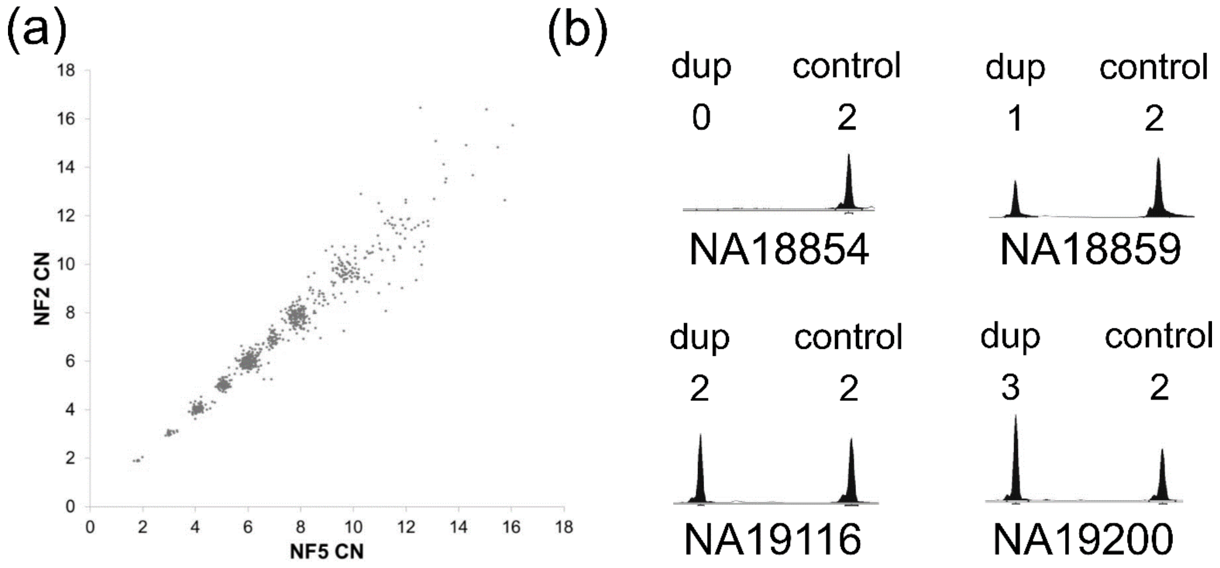
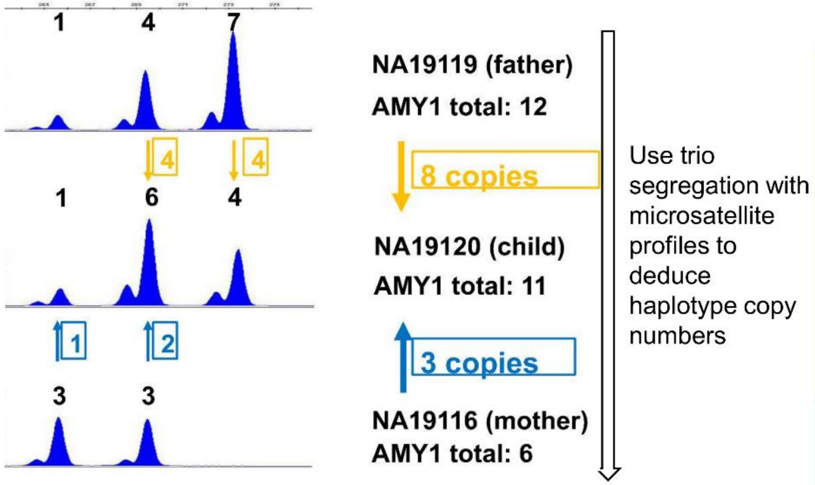

(a)
18
16
14
12
10
8
6
4
2
0
NF2 CN
0
2
4
6
8
10
12
14
16
18
NF5 CN
(b)
dup
control
0
2
NA18854
dup
control
1
2
NA18859
dup
control
2
2
NA19116
dup
control
3
2
NA19200

Figure 3

Diploid integer copy numbers:

component	*AMY1*	*AMY2A*	*AMY2B*	junction
Father (NA19119)	12	4	2	0
Child (NA19120)	11	5	4	2
Mother (NA19116)	6	4	4	2



NA19119 (father)
AMY1 total: 12

8 copies

NA19120 (child)
AMY1 total: 11

3 copies

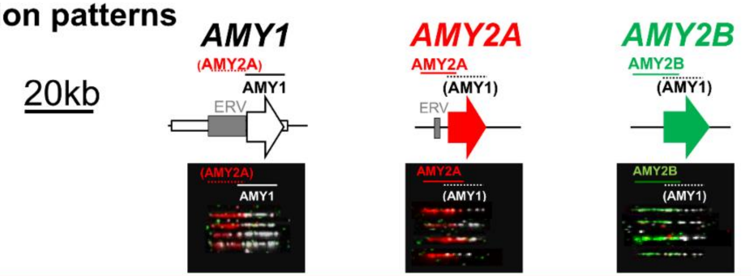
NA19116 (mother)
AMY1 total: 6

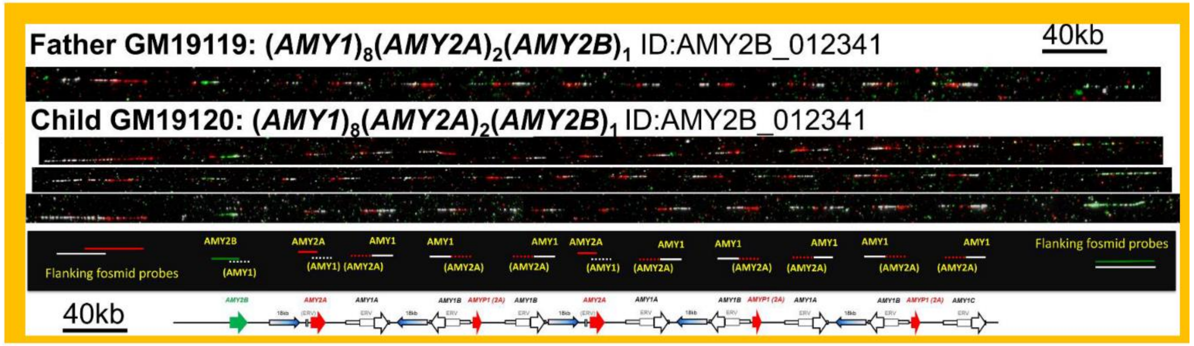
Use trio segregation with microsatellite profiles to deduce haplotype copy numbers

Haplotype copy numbers:

ID	component	*AMY1*	*AMY2A*	*AMY2B*	junction
Father	transmitted	**8**	**2**	**1**	**0**
	untransmitted	4	2	1	0
Child	Paternal	**8**	**2**	**1**	**0**
	Maternal	**3**	**3**	**3**	**2**
Mother	transmitted	**3**	**3**	**3**	**2**
	untransmitted	3	1	1	0

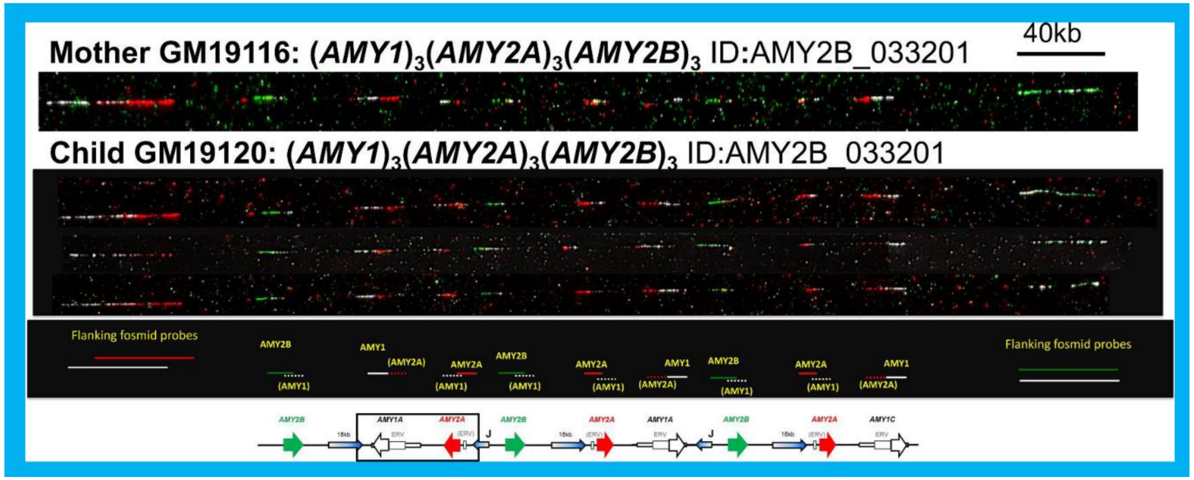
Hybridization patterns





Father GM19119: $(AMY1)_8(AMY2A)_2(AMY2B)_1$ ID:AMY2B_012341

Child GM19120: $(AMY1)_8(AMY2A)_2(AMY2B)_1$ ID:AMY2B_012341



Mother GM19116: $(AMY1)_3(AMY2A)_3(AMY2B)_3$ ID:AMY2B_033201

Child GM19120: $(AMY1)_3(AMY2A)_3(AMY2B)_3$ ID:AMY2B_033201

Table 1. Primers used in this work.

Primer	Sequence (5'-3')
AMY1CF	TTCTAAGGTGCCTTCTAGTC
AMY1CR	CATCTTCAAGCCTGCATTC
NF2	ATAGCTTAGAGTAGTTAAC
AMY1CRB2	AGTGAGATGAGGCATTGTG
NF5	GGCCTCTATACATGAG
AMY2B2D	GCCTGGCTAATTTGTTGTTAG
AMY2B2R	AAATTAACTCCATGCATCACC
AMY2B2F	TGCATAGAAATGGCACATAGT
AMY1_2F	ACAGTTGATTTTTGATCTTGTAGG
AMY1_2R	TACAGCATCCACATAAATACGAA

Table 2. Segregation of *AMY1*, *AMY2A* and *AMY2B* copy number in Yoruban trios Y045 and Y056

Family Y045	**ID**	**component**	***AMY1***	***AMY2A***	***AMY2B***	**junction**
Mother	NA19201	diploid	6	2	2	0
		transmitted haplotype	3	1	1	0
		untransmitted haplotype	**3**	**1**	**1**	**0**
Father	NA19200	diploid	7	5	5	3
		transmitted haplotype	**4**	**4**	**4**	**3**
		untransmitted haplotype	3	1	1	0
Child	NA19202	diploid	7	5	5	3
		Maternal haplotype	**3**	**1**	**1**	**0**
		Paternal haplotype	**4**	**4**	**4**	**3**
Family Y056	**ID**	**component**	***AMY1***	***AMY2A***	***AMY2B***	**junction**
Mother	NA19159	diploid	8	6	6	4
		transmitted haplotype	**3**	**1**	**1**	**0**
		untransmitted haplotype	5	5	5	4
Father	NA19160	diploid	6	2	2	0
		transmitted haplotype	**3**	**1**	**1**	**0**
		untransmitted haplotype	3	1	1	0
Child	NA19161	diploid	6	2	2	0
		Maternal haplotype	**3**	**1**	**1**	**0**
		Paternal haplotype	**3**	**1**	**1**	**0**

Table 3. Segregation of *AMY1*, *AMY2A* and *AMY2B* copy number in Yoruban trios Y060 and Y072

Family Y060	**ID**	**component**	***AMY1***	***AMY2A***	***AMY2B***	**junction**
Mother	NA19116	diploid	6	4	4	2
		transmitted haplotype	**3**	**3**	**3**	**2**
		untransmitted haplotype	3	1	1	0
Father	NA19119	diploid	12	4	2	0
		transmitted haplotype	**8**	**2**	**1**	**0**
		untransmitted haplotype	4	2	1	0
Child	NA19120	diploid	11	5	4	2
		Maternal haplotype	**3**	**3**	**3**	**2**
		Paternal haplotype	**8**	**2**	**1**	**0**
Family Y072	**ID**	**component**	***AMY1*[a]**	***AMY2A***	***AMY2B***	**junction**
Mother	NA19152	diploid	5/6	6	5	3
		transmitted haplotype	**2/3**	**3**	**2**	**1**
		untransmitted haplotype	3	3	3	2
Father	NA19153	diploid	8	2	2	0
		transmitted haplotype	**3**	**1**	**1**	**0**
		untransmitted haplotype	5	1	1	0
Child	NA19154	diploid	5/6	4	3	1
		Maternal haplotype	**2/3**	**3**	**2**	**1**
		Paternal haplotype	**3**	**1**	**1**	**0**

[a] Alternative values are shown for the *AMY1* copy numbers of the mother and child in family Y072; because of the partial copy of *AMY1* on the transmitted maternal haplotype, the copy number recorded depends on the precise location of the measure used.