Fish and chips: the origin of human gene families is a predictor of the location of GWAS signals.

Sara V. Good[1,2^], Ryan Gotesman, [3], Ilya Kisselev [1], Andrew D. Paterson[3,4^]

[1] Department of Biology, The University of Winnipeg, MB, [2] Department of Biological Sciences, The University of Manitoba, Winnipeg, MB, [3] Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada; [4] Divisions of Biostatistics and Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Author emails: s.good@uwinnipeg.ca (SVG), Ryan Gotesman (RG), Ilya Kisselev, Andrew D. Paterson Andrew.paterson@sickkids.ca

Author (SVG) ORCID: 0000-0003-0563-7724

Author (ADP) ORCID: 0000-0002-9169-118X

^**Joint corresponding authors**: (SVG)(ADP)

Running header: Ancestral Linkage Group influence on GWAS

Keywords: genome wide association studies, ancestral genome reconstruction, ancestral linkage groups, macrosynteny, Experimental Factor Ontology, ohnologs, gene-level constraint, GTEx.

Abstract

GWAS have identified thousands of loci associated with human complex diseases and traits. How these loci are distributed through the genome has not been systematically evaluated. We hypothesised that the location of GWAS loci differ between ancestral linkage groups (ALGs) related to the paralogy and function of genes. We used data from the NHGRI-EBI GWAS catalog to determine whether the density of GWAS loci relative to HapMap variants in each ALG differed, and whether ALG's were enriched for experimental factor ontological (EFO) terms assigned to the GWAS traits. In a gene-level analyses we explored the characteristics of genes linked to GWAS loci and those mapping to the ALG's. We find that GWAS loci were enriched or deficient in 9 and 7 of the 17 ALG's respectively, while there was no difference in the number of GWAS loci in regions of the human genome unassigned to an ALG. All but 2 ALG's were significantly enriched or deficient for one or more EFO terms. Lastly, we find that genes assigned to an ALG are under higher levels of selective constraint, have longer coding sequences and higher median expression in the tissue of highest expression than genes not mapping to an ALG. On the other hand, genes associated with GWAS loci have longer genomic length and exhibit higher levels of selective constraint relative to non-GWAS genes. Collectively, this suggests that understanding the location and ancestral origins of GWAS signals may be informative for the development of tools for variant prioritization and interpretation.

Introduction

Genome wide association studies (GWAS) have identified thousands of genetic variants that influence human complex traits including both disease and non-clinical phenotypes (Hirschhorn and Daly 2005; Hindorff et al. 2009; Visscher et al. 2017; Buniello et al. 2019). Collectively, these studies indicate that variation in complex traits is often controlled by a few to several dozen loci, i.e. core genes, while the rest, and the majority, of trait heritability is massively polygenic and driven by weak effects of trans-acting expression quantitative trait loci (eQTL) (Shi et al. 2016; Boyle et al. 2017; Liu et al. 2019). In order to prioritize the functional effects of GWAS variants, developing tools to assess the potential impact of variants on Mendelian and complex diseases has become an important area of research. Early tools to prioritize variants utilized orthologous gene conservation across taxa and/or information about the location of missense or nonsense mutations within the 3D structure of proteins (Siepel et al. 2005; Kumar et al. 2009; Adzhubei et al. 2013; Mistry et al. 2013; Shihab et al. 2013; Schwarz et al. 2014; Choi and Chan 2015; Ioannidis et al. 2016), while more recent tools incorporate other information such as population allele frequencies, functional genomics data, and other genome annotations (Davydov et al. 2010; Yandell et al. 2011; Hu et al. 2013; Kennedy et al. 2014; Kircher et al. 2014; Schwarz et al. 2014; Gulko et al. 2015; Kichaev et al. 2019).

In addition to variant prioritization methods, there are tools that use gene-level information to guide variant analysis or interpretation. For example, based on an empirical approach to calculate the expected number of *de novo* mutations per gene (Samocha et al. 2014), ExAC used information from 60,706 exomes to estimate the probability of loss-of-function (LOF) intolerance (pLI) for genes in the human genome (Lek et al. 2016). The most LOF-intolerant genes (pLI>0.9) predicted by this metric were expected to predict haploinsufficient

human dominant disease genes, but pLI conflates estimates of dominance and selection (*h* and *s* in population genetics theory), and is thus a measure of the strength of selection of deleterious alleles in heterozygotes (Fuller et al. 2019). Currently, the gnomAD consortium, the continuation of ExAC, recommends using a related metric known as the observed to expected number (*oe*) of LOF variants in a gene, which is a continuous statistic and provides a better estimate of gene-level purifying selection (Karczewski et al. 2020). Lastly, recently a novel approach to incorporate gene family information (i.e. paralogous genes) to facilitate variant interpretation in neurodevelopmental disorders was developed (Lal et al. 2020).

Few studies, however, have examined genome-wide patterns of the location of SNPs associated with diseases or traits from GWAS. Here, we explored the distribution of GWAS variants in the context of ancient conserved blocks of macrosynteny in the human genome. Ancestral genome reconstruction (AGR) has been used to reconstruct the karyotypes of extinct genomes including those of the early chordate (Putnam et al. 2008; Simakov et al. 2020) and vertebrate ancestors (Nakatani et al. 2007; Sacerdot et al. 2018; Nakatani et al. 2021). AGR views genomes as sets of macrosyntenic fragment adjacencies, and employs algorithms enforcing local or global parsimony to reconstruct ancestral karyotypes at internal nodes (Damas et al. 2021). AGR has uncovered that there are regions of strong macrosyntenic conservation across vertebrate genomes that can be traced to the vertebrate ancestor and which are shared with extant invertebrate deuterostome lineages such as scallop and amphioxus (Nakatani et al. 2007; Sacerdot et al. 2018; Simakov et al. 2020; Nakatani et al. 2021).

By employing AGR approaches, studies have inferred that there were 17 or 18 ancestral linkage groups (ALG's) in the proto-vertebrate and chordate ancestors (Sacerdot et al. 2018; Simakov et al. 2020; Nakatani et al. 2021). However, a major difference between invertebrate

and vertebrate genomes, is that early during vertebrate evolution two rounds (1R and 2R respectively) of whole-genome duplication (WGD) occurred at ~490 mya prior to the divergence of agnathans and at ~438 - 485 mya prior to the divergence of cartilaginous and bony fish (Simakov et al. 2020; Nakatani et al. 2021). This means that the post-2R vertebrate genome has undergone two tetraploidization events relative to the pre-vertebrate ancestor; the non-random distribution of paralogons harbouring subsets of these 2R-derived ohnologs could influence the distribution of GWAS variants. Overall, ~ 35% of all human genes are estimated to have one or more paralogous genes that were duplicated during 2R (so called ohnologs), (Makino et al. 2009; Singh et al. 2015), and ohnologs are enriched for monogenic diseases (Chen et al. 2013).

AGR has identified the variation in the length and conservation of homologous syntenic blocks (HSB's) across mammalian genomes, which are interspersed, by definition, with evolutionary breakpoint regions (EBRs) (Pevzner and Tesler 2003; Bourque et al. 2005; Peng et al. 2006; Alekseyev and Pevzner 2007; Becker and Lenhard 2007; Damas et al. 2021). HSB's are blocks of sequence with conserved gene order and orientation in two or more species, while EBR's are regions of nonalignment at a given resolution in between two HSB's (Larkin et al. 2009; Damas et al. 2021). To date, there has been no description of the evolutionary characteristics of conserved syntenic blocks in the human genome relative to the pre-vertebrate ALG's or to the amphioxus linkage groups, although such blocks are recognised (Marletaz et al. 2018; Simakov et al. 2020). Further, a new extensive genomic, methylation and transcriptomic database for the Mediterranean lancelet (*Branchiostoma lanceolatum*) has provided genome-wide insights into the genomic and regulatory changes associated with early vertebrate WGD events and identified the pre-2R duplicated orthologues of genes in amphioxus with those in diverse human species, including humans (Marletaz et al. 2018). These advances have made it

possible to more accurately identify conserved paralogons in the human genome and ohnologous gene family members.

Thus, we broadly hypothesized that loci contained within ancient (i.e. pre-2R) conserved linkage blocks in the human genome exhibit different evolutionary or functional characteristics from those not contained in ALGs. Secondly, we hypothesized GWAS variants underlying complex traits may be non-randomly distributed with respect to their location, and exhibit broad functional clustering by ALG. To this end, we examined the following hypotheses: 1) the distribution of GWAS-significant SNPs in the human genome is partially explained by the evolutionary relationship of the genomic regions in which the SNPs are located relative to the pre-2R ancestor, 2) there is enrichment for distinct ontological categories of traits residing in different ALG's, 3) the level of evolutionary constraint on genes will be significantly impacted by whether they are assigned an ALG and 4) the expression of genes associated with enriched ancestry/trait ontological combinations may exhibit clustering in the human genome.

Results

Identification of human genomic regions that can be conservatively traced to an ancestral chordate linkage group. We compared the ancestral origins of protein coding genes as predicted by Sacercdot et al. (2018) and Simakov et al. (2020) to develop a map of the human genome relative to those predicted to be in the pre-2R linkage groups. After adjusting for differences in linkage group names, we found a high concordance between the assigned linkage groups of genes in the two studies, as well as with those presented by Nakatani et al. (2021) (Table S1A). Starting with the genes in common between the two reconstructions (n=4781), an additional 3273 genes were added from the Simakov et al. data that were not included in Sacercdot et al.,

14 genes were removed because of inconsistencies between the two models, and another 1795 genes were added to the map from the Sacercdot data that were not included in Simakov et al. because they were predicted to be in the same ALG as the flanking genes in the Simakov et al. model. Finally, based on gene family membership information from Marletaz et al. (2018), 560 additional genes were assigned to an ALG, giving a total of 10,410 genes being assigned to an ALG (SuppFigS1, Table S1B). The number of genes mapping to each ALG ranged from 330 genes on ALG Q to 1162 genes on ALG A (Table S1C). Of the 10,410 genes with assigned ALG ancestry, 7 mapped to the human Y-chromosome and were not included in further analyses due to the lack of GWAS data. As expected, the fewest genes mapped to human chromosome 21 (n=73), and the most to chromosome 1 (n=1095), with a significant excess of ALG genes being found of human chromosomes 2, 10, 14, and 15, and a significant deficit on chromosomes 6, 11, 19, 21 and X (see Table S1C).

In addition to the genes assigned to an ALG, we included 6,674 genes identified by Marletaz et al (2018) as co-orthologs of human genes in amphioxus ("amphioxus orthologs"), but which were not assigned to an ALG based on Sacerdote, Simakov or family information (i.e. genes in Table S1B with a family ID but no assigned ALG). Some of these genes pertain to tandem duplicates or regions of segmental duplication in the human genome (e.g. olfactory receptors) since such genes do not exhibit the requisite conserved synteny over deeper evolutionary time, but, in some cases one of them (the ancestral gene) will be included in the ALGs. Lastly, 2,288 genes were not assigned ancestry based on Sacerdot, Simakov nor did they exhibit orthology with genes in amphioxus and thus their location in the proto-vertebrate genome is unknown (Table S1B, SuppFig. S1).

Some regions in the human genome contain > 10Mb of contiguous sequence derived from a single ALG while others show evidence of high rates of fission/fusion over evolutionary time. For example, there are large blocks of ALG A on chromosomes 1, 2, 14 and 20; ALG B on chromosomes 2, 3, 5, 7 and 10; ALG D on chromosomes 16, 18 and 19; and ALG G on chromosome 7 and 12, while chromosomes 1, 5, 9 and 18/19 harbour discontinuous segments of ALGs C, L and/or M (Fig. 1a, Tables S1A, S1D, S1E), chromosomes 2 and 3 blocks of N/P and chromosomes 7 and 12 of E/O. These regions of smaller, intercalated, syntenic blocks of 2 or 3 ALG's correspond to the descendants of linkage groups that became fused post - 1R (see Nakatani et al. 2021, and Table S1A), and subsequently underwent local inversions etc. In addition to differences in the level of conservation of contiguous blocks derived from a single ALG's or a post-1R fused ALG, there are differences in the number of ohnolog descendants by ALG which varied from 1.61 for ALG H to 2.07 for ALG B (Table S1F). Given that ohnologs are, by definition, gene duplicates generated during 2R, it is not surprising that the odds of a gene assigned to an ALG being an ohnolog is 4.4 times greater than genes not assigned to an ALG (Odds ratio = 4.4, 95% CI: 4.12-4.71, z=43.38, p<0.0001).

Visualizing the gene families contained within blocks underscores the importance of 2R in generating global syntenic clusters of ancient gene family duplicates. For example, blocks of ALG B are conserved across parts of chromosomes 2, 3, 7, 12 and 17 with large blocks on chromosomes 2 and 7, in which HOX clusters A, B, C and D reside on chromosomes 7, 17, 12 and 2 respectively (Fig. 1b, Table S1b). Blocks derived from ALG A are found on chromosomes 1, 2, 11, 14, 15, 19 and 20, including almost the entirety of human chromosomes 14: the ohnologs of many gene families are distributed within these blocks (example *AKT, RYR, GREM, SPRED, FAM98, RASGRP, MTA, AKT, FERMT*, Fig 1c, for a full list see Table S4b). Regions of

the human genome derived from the post 1R fused blocks of ALG C and ALG L (Fig 1d, S4b) contain many other well-known ohnologous gene family members (*SKOR, SMAD, DOCK, JAK, PDE4, PK3R, ADAMTSL, RAB3, TARS*) even when the gene names belie otherwise (*INSR/INSRR/IGF1R*) (see Table S4b for a full list on putative ohnologs).

Mapping and comparing the distribution of GWAS and HapMap variants in the human genome relative to the ALG from which they are derived. The GWAS catalogue contained 128,322 entries with 72,041 unique rsID's of which 49,129 could be assigned to an ALG and 22,912 fell in regions of unknown ancestry (31.73%) (Fig. 2a, Table S2A). The density of GWAS significant variants was compared to that available from CEU phase II of the HapMap project, since this has been a common resource for GWAS imputation. To this end, 2,753,189 SNPs from the HapMap project were assigned to ancestral blocks in the human genome: 883,156 SNPs fell in regions of unknown ancestry (32.1%) and 1,870,033 to an ALG (Fig. 2a, SuppFig.2 for higher resolution). A Pearson's Chi-squared test with Yates' continuity correction was performed to assess individual deviation in the assignment of GWAS variants in each ALG relative to the HapMap expected proportion per ALG. All of the ancestral ALG's except ALG E showed either an excess (ALG's A, B, C, G, H, L, M, O and P) or deficit (ALG's E, F, I, J, K, N and Q) of GWAS variants relative to that expected based on HapMap diversity per ALG (Fig. 2b, Table S2b), except for variants falling into unassigned regions of the genome (Fig. 2b, Table S2b).

EFO-analysis of GWAS traits. Given that there is an excess of GWAS associated variants on some ALG's and a deficit on others relative to HapMap variation, we next asked whether the ontological terms assigned to the GWAS traits were non-randomly distributed with respect to ALG's. Of the ~72,000 GWAS associated SNPs, 50,538 were associated with a primary hierarchical term in the EFO classification (Tables 3Sa). EFO categories with fewer than 300

observations were removed or combined with other categories leaving 14 EFO categories: anthropometric, behavioural, biochemical, cardiovascular, digestive-endocrine, hematological, hepatobiliary, lipid, musculoskeletal, nervous, respiratory, immune-infectious, urinary, visual which were collectively associated with 42,099 SNPs. Using a global FDR, chi-square analysis indicated that all EFO traits except urinary system were enriched or underrepresented with at least one but not more than four ALG's (Table S3b). Odds ratio (+/- 95% confidence intervals) were calculated for the 29 associations exhibiting an excess or deficiency of GWAS variants for a given EFO x ALG combination (Table S3c, Figure 3). This revealed the ALG A is enriched for genes in nervous and visual traits, ALG B for musculoskeletal traits, ALG C for nervous system traits, ALG D for lipid and hematological traits, ALG I and Q for cardiovascular traits, and that genes that were unassigned to an ALG were enriched for endocrine and immune system related traits. (Figure 3, Table S3b)). The list of genes associated with each GWAS variants for each of the 14 EFO categories is given in Supplementary Tables 3.

Gene level analyses. Next, we employed data from gnomAD, GTEx (v8) and Ensembl to explore characteristics of genes: i) assigned to an ALG or ii) associated with GWAS hits (SuppFig. 1). For the analyses of genes assigned to an ALG, genes assigned as unknown and having orthologs in amphioxus were grouped. We identified 11,269 genes associated with the 72,041 unique GWAS variants: unsurprisingly, the number of associations per gene is highly skewed with 4091 genes having < 5 associations, and 199 having >30; each gene was counted once for the gene-level analyses (Table S4a). Of the 11,269 GWAS associated genes and the 10,410 genes assigned to an ALG, 11,055 and 10,366 were annotated in the gnomAD V2.1.1 database respectively (n=18,766 total genes at gnomAD, TableS4b). The odds of a GWAS associated gene being assigned to an ALG was 1.64 (CI: 1.54-1.73, z=16.4 p<0.00001, Table S4a), and

64.2% of the genes assigned to an ALG were associated with a GWAS, while 52% of the genes unassigned to an ALG were associated with a GWAS. In agreement, with the distribution of GWAS vs HapMap SNPs (Fig. 2b) some ALG's had more GWAS associated genes than expected (ALG's A, B, C, G, M, O and P), while others had fewer (ALG's F, J, K and N) (comparison of Figures 2b and 4A). The genes with the largest number of GWAS signals that were unassigned to an ALG included immune related genes (particularly HLA in the MHC region, and interleukin ligands and/or receptors), a cluster of overlapping genes involved in glucuronidation on chr 2 (*UGT1A4-UGT1A10*), and a variety of endocrine related genes (several related to obesity or lipid metabolism, including *FTO, LEPR, LPA*)(Table S4b). This is consistent with the EFO analysis that found that immune and endocrine functions are enriched for genes not assigned to an ALG. Notably, orthologs of the ancestral MHC locus have been difficult to identify and, as such, the HLA-loci are not included in the map presented here. However, current data suggests that the MHC locus was present in the ancestral vertebrate genome but acquired its unique function in antigen recognition after the divergence of agnathans (Flajnik and Kasahara 2010; Flajnik 2018; Ohta et al. 2019; Nakatani et al. 2021). The orthologs of these MHC genes would map to ALG M.

Gene level constraint. Genes that trace their ancestry to an ALG and those associated with a GWAS variant may exhibit constrained evolutionary change. Using pLI scores >0.9 as a proxy for haploinsufficiency, we find that genes that trace their ancestry to an ALG (OR=2.58; CI 2.37-2.81, z=22.00, $p<0.0001$) or are associated with a GWAS variant (OR = 2.02, CI 1.85-2.2, Z = 15.9, $p<0.0001$) (Figure 4b) are both enriched for genes that are intolerant to loss of function mutations. To further explore the characteristics of these genes, we performed logistic regression to explore the effect of six variables on ALG or GWAS association: i) gene length ii) coding

sequence (cds) length iii) number of coding exons iv) loss-of-function observed/expected upper bound fraction (LOEUF) v) the mean number of nonsynonymous mutations (dN) for a gene between human and mouse and vi) the median expression (in transcripts per million, TPM) in the highest tissue of expression listed in GTEx. After checking for collinearity and removing variables with low explanatory power, the logistic regression showed that genes assigned to any ALG had i) a lower rate of nonsynonymous substitutions (β=-6.08, χ^2 = 644.42, p<0.00001), ii) more coding exons (β=0.088, χ^2 = 558.43, p<0.00001), iii) longer cds length (β=-0.00035, χ^2 = 237.7, p<0.0001), iv) higher median gene expression (β=-0.00010, χ^2 = 14.1, p<0.0002), and iv) lower LOEUF (β=-0.26, χ^2 = 21.26, p<0.0002), (Fig 4b, c, SuppFig3, d, e, respectively). On the other hand, genes associated with GWAS variants had i) longer gene length (β=0.00000084, χ^2 = 644.3, p<0.0001), ii) lower LOEUF scores (β = -0.38, χ^2 = 92.35, p<0.0001), and iii) longer cds length (β=0.00012, χ^2 = 70.60, p<0.0001) (Figures 4 F, H, SuppFig.3 and 4). Collectively, this shows that genes mapping to ALGs evolve more slowly on both shorter (LOEUF) and deeper (dN mouse-human) evolutionary timescales, have more exons and longer coding sequences and higher levels of gene expression than genes not mapping to an ALG. On the other hand, genes associated with GWAS are longer than those not associated with a GWAS, and exhibit greater functional constraint as measured by both pLI and LOEUF but not dN.

Mapping of GWAS associated genes involved in lipid metabolism. The finding that there are blocks of genes with shared ancestry distributed throughout the genome which are enriched for ohnologs, evolve slowly and show evidence of clustering by functions based on the EFO analysis, suggests that genes that trace their ancestry to an ALG may be involved in core biological processes. As a test of this hypothesis, we took the list of putatively core genes involved in lipid metabolism (Dron and Hegele 2016) and placed them within their human

genome context coloured by ancestral ALG and labelled by the tissue of highest expression for each gene in the GTEx catalogue. This reveals that 18 of 19 of the genes involved in lipid metabolism emanate from 5 ALG's (one gene, APOC2 has unknown origin), and several of them are 2R ohnologs. For example, *APOA1*, *APOA5, APOE* and *APOC2* exist in two paralogons on chromosomes 11 and 19 that emanate from ALG A; *LPL*, *LIPC* and *LIPG* are 2R ohnologs on chromosomes 8, 15 and 18 respectively derived from ALG C; *LDLR* and *LDLRAD1* are paralogues derived from ALG L that now reside on chromosomes 1 and 19. Lastly, unrelated genes, *SAR1B* and *MTTP* reside on chromosomes 5 and 4 respectively, yet they are all derived from ALG Q in the pre-vertebrate ancestor. Although the remaining four genes – *LCAT* and *CETP* (ALG D), *PCSK9* and *ANGPTL3* (ALG L) are not paralogs, they share current (physical linkage) as well synteny in the pre-vertebrate genome.

Similarly, in the EFO analysis, we found that ALG A and C were enriched for nervous and/or visual traits and ALG E for behavioural traits. Given previous suggestions that genes within HSB are enriched for neurodevelopmental functions, we looked for evidence that genes on these three ALG's had brain as the tissue of highest gene expression. For genes without ALG ancestry, 18.5% (1680/9080) have the highest tissue of expression as brain versus 21.5% of those assigned to an ALG (2086/9685) with values ranging from 14.4% on ALG K to 25% on ALG Q. For the three ALG's exhibiting enrichment for brain-related traits, the proportion of genes with highest expression in brain were 24.1% (ALG A), 22.9% (ALGC) and 22.6% (ALG E). Visualization of the density of genes on hg38 (gray histogram), with the density of genes with highest expression in brain (overlaid red histogram), as well as the median TPM expression of brain-related genes on ALG A, C and E (blue dots to the left of the chromosomes) suggests that some regions of the human genome may harbor clusters of brain related traits. For example,

large blocks of genes with highest expression in brain derived from ALG A (dark blue bands on chromosomes) are present on chromosomes 1 and 2, from ALG C (burgundy bands) on chromosomes 5 and 15, and from ALG E (pink bands) on chromosome 3 (Figure 5B).

Discussion

Although the primary focus of GWAS is to identify causal loci for complex diseases and traits, an understanding of the population and evolutionary processes that maintain heritable variation is essential to their interpretation (Sella and Barton 2019). The analysis of highly studied GWAS traits has uncovered that the genetic basis of many complex phenotypes and diseases is extremely polygenic, highly pleiotropic and, at least in the case of height and body mass index, predominantly influenced by stabilizing (rather than directional) selection (Shi et al. 2016; Simons et al. 2018; Sella and Barton 2019). As more data are collected about the biological function of genes, other factors are coming to the forefront, particularly the role of allelic effects (cis and trans-regulatory elements) on gene expression (Boyle et al. 2017; Liu et al. 2019), which helps to explain the large number of GWAS loci that fall in regulatory regions (Visscher et al. 2012; Visscher et al. 2017). At the level of the gene, molecular evolutionary approaches to predicting the functional impact of variants has also revealed that i) biases and differences in the mutational process influence the probability of observing specific variants and ii) genes differ markedly in their average rate of evolutionary change, with slower evolving genes exhibiting greater functional constraint and greater susceptibility to deleterious mutations, a fact that is the basis for many variant interpretation tools (reviewed in (Eilbeck et al. 2017)).

316 At the genome level, approaches in comparative syntenic analyses and ancestral genome
317 reconstructions have identified so-called homologous syntenic blocks (HSBs) and evolutionary
318 breakpoint regions (EBRs) in amniote genomes (Peng et al. 2006; Alekseyev and Pevzner 2007;
319 Kemkemer et al. 2009; Larkin et al. 2009; Alekseyev and Pevzner 2010; Damas et al. 2018;
320 Farre et al. 2019), as well as the imprint of the early vertebrate WGD events (Nakatani et al.
321 2007; Putnam et al. 2008; Sacerdot et al. 2018; Simakov et al. 2020; Nakatani et al. 2021). The
322 location of breakpoints associated with interchromsomal exchange in mammals is non-random,
323 indicating that selection acts at the genome level to maintain combinations of genes and their
324 regulatory elements that are essential for biological organization (Damas et al. 2021). This is in
325 agreement with studies that show that the 3D organization of chromatin in the nucleus and the
326 GC peaks and valleys of topologically associated domains (TADs) strongly influence the number
327 of contacts and inter-chromosomal cross-over events (Berthelot et al. 2015; Jabbari et al. 2019).
328 In this study, we aimed to define the conserved 2-R derived paralogons in the human
329 genome and then examine the distribution of GWAS variants within them. To this end, we
330 generated a map of the human genome with respect to 17 inferred ancestral linkage groups
331 (ALGs) by comparing two reconstructions of the ancestral vertebrate (Putnam et al. 2008;
332 Sacerdot et al. 2018) and one of the ancestral chordate karyotype (Simakov et al. 2020). We
333 relied primarily on the Simakov et al. (2020) model of the ancestral chordate karyotype because
334 their reconstruction is based on a new chromosome level assembly of the amphioxus genome, a
335 basal deuterostome with similar linkage groups to the inferred chordate ancestor. A new
336 reconstruction of the proto-vertebrate karyotype by Nakatani et al., (2021) inferred there to be 18
337 linkage groups in the vertebrate ancestor; their proto-vertebrate chromosomes (pvc's) are highly
338 concordant with those described here, and similar to those present in scallop, an early diverging

deuterostome, as well as to the vertebrate ancestor genome inferred by Sacercdot et al., (2018). The additional linkage group they infer is a microchromosome that, based on the Simakov et al. (2020) model, corresponds to a small block of genes belonging to ALG A, and which lacked clear synteny with gnathostome or invertebrate genomes (Table S1a, pvc18). Although Nakatani et al. (2021) do not provide a list of the human genes that they infer to have putative co-orthologs in the pvc, they provide the linkage groups of 19,969 human genes in the proto-gnathostome genome, which we used to examine the concordance of their gene assignments to those presented here (Table S5a). Of the 10,410 genes assigned to an ALG in this study, 7616 (73.2%) were concordant with and 580 (5.6%) were not included in the Nakatani et al. (2021) study, while 2214 (21.2%) were assigned to either a different ALG or could have been assigned to the same or different ALG (Table S5b). The ambiguity in some gene assignments is due to fusion of some of the pvc's in the post-1R or post-2R genomes as presented in Nakatani et al. (2021) (see Table S1a).

Based on a thorough study of the orthology of genes between Mediterranean lancelet and vertebrate taxa, we find an additional 6,647 human genes that have a putative ortholog in amphioxus that are not present in conserved ALG's. Many of these genes belong to tandemly duplicated gene clusters which have likely duplicated in the post-2R genome, such that only one gene will map to an ALG. This underscores that the map we generated represents the conserved macrosyntenic clusters in the human genome, rather than all of the genes that originate from the pre-2R genome. On the other hand, some genes that are small and fast evolving may not be represented in the ALG's even if they diversified during 2R. For example, we previously performed detailed phylogenetic and syntenic mapping of the insulin superfamily peptides (Yegorov and Good 2012), and found that the ancestral INS-IGF linked genes trace to ALG O

and those of the relaxin like peptides to ALG L. However, these genes are not included in the Sacercdot or Simakov reconstructions. With the availability of improved genome assembly and annotation pipelines, such as that being developed by the Vertebrate Genome Project (Rhie et al. 2021), there will be more data with which to assess if genes are "ancient" or have arisen post 2-R by another process such as *de novo* origination, local duplication, and/or a retroposition event (Casola 2018; Van Oss and Carvunis 2019; Cosby et al. 2021). This will allow a deeper understanding of the nature of conserved versus adaptive regions of vertebrate genomes, and facilitate identification of regions in the human genome that would be susceptible to deleterious consequences from structural variants. A motivation for basing our analysis on the ancient pre-2R linkage groups, is that it also facilitates the identification of ohnologs in the human genome which could further help prioritize disease genes (Makino and McLysaght 2010; Chen et al. 2013; Lal et al. 2020)

We find that approximately half of the protein coding genes in the human genome can be reliably traced to their pre-vertebrate origins and that genes assigned to an ALG have a slower rate of evolution as measured on both deeper (dN) and more recent (LOEUF) timescales. Further, the paralogons harbouring the 2-R derived genes are found in blocks (in which gene memberships but not necessarily gene order is conserved) of varying size with some parts of the human genome showing large regions of conserved ancestry. By defining the chromosomal coordinates of the ALGs, we show that all but one of the ALG's (ALG E) and unassigned regions, are either enriched or depauperate in GWAS loci relative to that expected based on variation in the human genome identified by HapMap v2. Even though the selection of loci by the HapMap project was subject to ascertainment bias (Clark et al. 2005), the concordance between the GWAS and HapMap variants in unassigned regions, supports the inference that the

GWAS signals are non-randomly distributed with respect to ALGs. By further linking the GWAS loci to their closest gene and associated experimental factor ontology (EFO) terms, we determined that all of the EFO terms except one (Urinary) are enriched for ancestry of genes on one to four ALGs, while genes that were unassigned to an ALG are enriched for immune and endocrine system functions. This suggests that the human genome has retained blocks of genes associated with distinct functions that emanate from conserved linkage groups in the vertebrate ancestor. This is concordant with studies in amniotes showing that genes in HSBs are more likely to be involved in conserved processes such as development while those in EBR's are more likely to be involved in adaptive processes (Larkin et al. 2009; Damas et al. 2021).

Examination of the distribution of the ALG's in the human genome, indicates high levels of intermixing of some ALG's, such as ALG's C/L/M, ALG's E/O, ALG's N/P and ALG's D/J. Based on the work of Nakatani et al's (2021), it is now clear that these associations are due to the gnathostome specific fusions of duplicates of these ALG's following 1R. They infer that following the 1R tetraploidization event, several of the resulting 36 chromosomes were fused giving rise to a post-1R karyotype of n=25 (see Table S1A), while after 2R, an additional few fusion events occurred leading to a karyotype of n=49, including several microchromosomes. Some of these post-1R fused chromosomes maintained high levels of conservation (e.g. C/M/L and E/O) in the human genome.

We found a strong signal that genes that map to an ALG evolve significantly more slowly than genes that do not. Using logistic regression, we find that the largest contributor to ALG vs non-ALG gene assignment was the rate of nonsynonymous substitution of the genes between mouse and human. Further, the level of evolutionary constraint measured LOEUF was higher for genes assigned to ALG's (median LOEUF = 0.38) than those associated with a GWAS loci (median LOEUF= 0.42) and they exhibited greater intolerance to loss of function mutations (pLI). Genes assigned to an ALG also had longer mean cds length and more coding exons

(which had a Pearson correlation coefficient of r^2=0.80), as well as a higher median transcript expression in the tissue of highest expression based on GTEx v8.0 data. These results are consistent with studies that show that highly expressed genes show greater evolutionary constraint and are more likely to exhibit dosage sensitivity (Duret and Mouchiroud 2000; Pal et al. 2001; Rocha and Danchin 2004; Gout et al. 2010), and further suggests that the genes with deep ancestry in the human genome are likely to be enriched for core genes, as per the conceptual framework of the omnigenetic model (Boyle et al. 2017; Liu et al. 2019)

On the other hand, genes associated with GWAS loci, had an approximately twofold higher odds of being haploinsufficient as measured by pLI (OR=2.02) and significantly lower LOEUF scores (median LOEUF=0.42) than non-GWAS genes (median LOEUF=0.57). The pLI statistic has been interpreted as a measure of haploinsufficiency, but this only holds when the selection coefficient is high (i.e. s=1), which occurs frequently for Mendelian disease genes, but not other genes (Fuller et al. 2019). When s <1, pLI measures the average effect of loss of function mutations in heterozygotes (i.e., hs). Apart from exhibiting high pLI scores, the best predictor of being a GWAS associated gene was gene length, which is unsurprising since GWAS loci are assigned to the nearest gene.

The finding that genes exhibiting deep ancestry in the human genome have strong functional constraint, are highly enriched for ohnologs (OR = 4.4) and have higher median gene expression, suggests that they should be enriched for disease associated loci. Overall, the odds of a gene assigned to an ALG being associated with a GWAS signal was 1.64 (CI:1.54-1.79). Studies have shown that genes associated with neurodevelopmental disorders (NDDs) exhibit greater evolutionary conservation of copy number, i.e. are intolerant to copy number variation, across mammals indicating that they are subject to greater functional constraint (Rice and

McLysaght 2017), and that most (perhaps 80%) known monogenic disease genes belong to gene families consisting of ohnologs and/or paralogs (Dickerson and Robertson 2012; Chen et al. 2013). Recently, Lal et al., (Lal et al. 2020) examined whether inclusion of gene family information could improve variant interpretation for missense mutations in patients with NDDs, given that NDDs are phenotypically and genetically heterogeneous. They found that by including information about whether variants changed paralog-conserved vs paralog non-conserved residues helped to prioritize missense variants with potential functional consequences. Using the same reasoning, information about membership of genes/ohnologs in ALGs may help prioritize genes influencing the same or similar phenotypic traits. For example, we demonstrated that even though the core genes underlying lipid metabolism are distributed on 11 chromosomes, they emanate from 5 ALG's, harbour 3 ohnologs, and have the highest tissue of expression in a logical tissue (e.g. liver, adrenal gland, spleen, adipose etc). Thus, by adding the information regarding the deep ancestry of genes into our understanding of gene and genome organization, this approach may help to identify regions of susceptibility to structural variants and/or to be another tool to inform variant prioritization.

The GWAS catalogue grows at an increasing pace and provides a valuable resource to summarise and interpret the results of GWAS. However, there are, of course, limitations of using the catalogue. Among the more salient caveats is that because the studies deposited into the catalogue were generated using many SNP arrays, different but highly correlated SNPs in LD with each other may appear as independent signals when a single SNP may have been identified if all studies used the same array. Secondly, traits that are easier to measure, such as height or weight, will have much larger study sizes than studies of other traits, and this will provide greater information about these traits than for rare phenotypes. These two caveats will compound each

other such that signals at specific genes could be driven by the large number of studies for a trait, measured on a variety of different panels across studies. We tried to minimize this effect by including genes associated with GWAS loci only once regardless of the number of associations. Additionally, we included SNPs in the study based on them having a genome-wide significant p-value ($< 5.0 \times 10^{-8}$). However, the p-value is a crude measure of effect and does not take into account sample size, effect size, nor minor allele frequency into account. Further, the use of the HapMap as a reference for SNP density is most relevant for early GWAS (since early array's typically genotyped variants based on HapMap, and there was limited imputation), but is less appropriate for more recent GWAS which use high-resolution imputation (e.g. TOPMed, (Taliun et al. 2021)) or even whole genome sequencing (WGS). But since each study used different GWAS arrays (often multiple) and in some cases different imputation panels it is not possible to come up with a simple solution for this. Despite these limitations, the data presented here indicate that there are genome-wide influences on the distribution of GWAS variants that can be, partially, explained by the deep and duplicated ancestry of our underwater ancestors.

Methods

We combined information from two studies to create a map of the human (*Homo sapiens*) genome with reference to hypothesized linkage groups in the pre-duplicated ancestral genome. Firstly, we used the reconstruction inferred by Sacercdot et al. (2018); they employed the Algorithm for Gene Order Reconstruction in Ancestors (AGORA) (Berthelot et al. 2015) based on local parsimony to identify contiguous ancestral regions (CARs) which they assembled into 17 hypothetical vertebrate ancestor chromosomes (VACs). As part of this, they employed the complete set of phylogenetic trees in Ensembl 69 from 45 extant amniotes to identify ohnologs – genes that were duplicated and retained following 2R in the Amniota ancestor – using criteria

such as putative ohnologs must be located at least 2N (90) genes apart on the same CAR or on separate CAR's, and that the estimated time of duplication of duplicates should fall within the 1R-2R window (~ 550mya). Secondly, we employed data from the study by Simakov et al. (2020) in which they reconstructed the ancestral chordate linkage groups (CLG's), i.e. an earlier timepoint, based on a new high-depth sequencing of the *Branchiostoma floridae* (Amphioxus) genome. To reconstruct the CLG's, they compared macro-syntenic blocks of genes between Amphioxus with an invertebrate (scallop) against three post-2R vertebrate genomes (chicken, frog and gar). Conservation of macro-synteny was estimated using a synteny discontinuity measure, D, based on the squared Euclidean norm of the difference between left and right windowed averages of the synteny indicator vector between Amphioxus and the four other taxa. In this way, they identified the set of genes and their linkage groups that were present in the chordate ancestor as well as the sets of genes, emanating from a single CLG, that were putatively duplicated in the post-2R vertebrate genome (i.e. ohnologous gene families). They infer that the chordate ancestral genome was also composed of 17 linkage groups. *B. floridae* has 19 chromosomes: they find that all of its' chromosomes, except *B. florida* (BFL) linkage groups 2, 3 and 4 (which exhibited 4, 1, and 1, breaks respectively) exhibited no discontinuity breaks with later diverging 2R taxa, indicating deeply conserved macro-synteny during chordate evolution, even though local synteny is not well conserved due to copious smaller scale inversions, gene loss and local duplications as well as subsequent chromosomal fission and fusion events during vertebrate evolution.

Construction of a map of the human genome based on the ancestral chordate genome. We downloaded the complete set of genes traced to the pre-vertebrate ancestral genome by Sacercdot et al. (2018) from (ftp://ftp.biologie.ens.fr/pub/dyogen/genomicus/69.10/ last accessed, October

2020). In total, 5,051 gene families were assigned to one of the ancestral linkage groups 1-17 and an additional 4773 were assigned to the ancestral genome but the linkage relationships were not defined. 4,968 of these families had at least one descendent in the human genome, with an average number of 1.69 paralogues per family, such that 8,389 protein coding genes from the human genome were assigned to one of the 17 VACs (number 1-17). Of these 8,389 genes, definitive gene names and locations were identified for 8,242 genes that were annotated by ENSEMBL based on build 38 of the human genome including all 22 autosomes and the X-chromosome (Table S1b).

Next, we downloaded the information from Simakov et al. (2020) who assigned ancestry of human genes relative to one of the 17 hypothetical CLG's (number A-Q) (https://bitbucket.org/viemet/public/src/master/ public, date accessed=November 2020). In total, they include 6082 gene families with representatives in the human genome, with an average number of 1.5 paralogues per family resulting in 9106 human protein coding genes. After removal of duplicates and selecting genes that were annotated in ENSEMBL, we obtained 5853 gene families representing 8065 genes. Using these two gene lists, we created a consensus map of the human genome relative to the ancestral linkage groups (ALGs) by using the Simakov-CLG ancestry assignment as a reference point and then adding or removing genes based on consistency with their assigned ancestral linkage group in Sacercdot et al. (2018). Specifically, we added genes included in the Sacercdot-VAC data if the genes fell within a contiguous stretch of the human genome mapping to equivalent regions in both studies. On the other hand, in cases where the two reconstructions differed in the assigned ancestry of a gene, gene ancestry was treated as unknown. For breakpoints, i.e. where there was a change in the ancestry of an ALG, the beginning and end of the block ID were assigned to be 10 kb from the nearest (start or end)

of a gene and intervening genomic regions were treated as having unknown ancestry. (Map: CLG_Map_locations_23-new.csv). To provide a third level of information, we used information about family membership based on orthologous gene comparisons between the human and amphioxus genomes as shown in Marletaz et al. (Marletaz et al. 2018). In cases in which the Marletaz study demonstrated that genes families with up to four members situated on different human chromosomes or more than 100Kb apart on the same chromosomes were putative ohnologs, the inferred mapped position based on other members of the gene family was inferred. This added additional ALG membership for 560 genes. Using the list of orthologous genes between human and amphioxus based in the Marletaz study (2018), genes with putative orthologs in amphioxus but which were not mapped to an ALG based on the above criteria were categorized as "amphioxus orthologs". Lastly, genes for which no ALG nor amphioxus ortholog could be identified were classified as having "unknown" ancestry.

Mapping GWAS and HapMap variants to the ancestral linkage groups. Next, we downloaded the complete list of single nucleotide polymorphisms (SNPs) that have been identified in Genome Wide Association Studies (GWAS) and deposited into the GWAS catalogue from https://www.ebi.ac.uk/gwas/ (file: GWAS_catalog_v1.0.2-associations_e100_r2020-07-14.tsv, last accessed on July 14, 2020). Using their genomic coordinates on build 38, we mapped all GWAS SNPs with a unique rsID and an association p-value $< 5 * 10^{-8}$ to the ALG map generated above and then counted the number of SNPs falling into each ALG(Buniello et al. 2019). To compare the density of human genetic variants identified in GWAS to overall diversity in the genome, we used Phase II r24 of autosomal and X-chromosome variants from the HapMap Utah residents (CEPH- yes; Centre d'Etude du polymorphisme humain) with northern and western European ancestry (CEU) obtained from

ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2008-10_phaseII/ and employed liftover (genome.ucsc.org) to update the variant coordinates from hg36 to hg38. A total of 2,753,012 HapMap SNPs were available (for both autosomes and X, 275 did not lift over) from which we calculated i) the number of SNPs in 200 kb windows and ii) the number of HapMap SNPs mapping to each CLG.

Ontological classification of GWAS catalogue phenotypes. The GWAS catalogue reports associations for a wide range of phenotypes including both disease (e.g. Type 2 diabetes) and non-clinical phenotypes (e.g. hair color). All curated traits included in the catalogue are mapped to terms from the Experimental Factor Ontology (https://www.ebi.ac.uk/efo/). In most cases, traits are mapped to a singular phenotypic class, although some traits map to 2-3 EFO terms, with most mapping to one of the 27 higher level terms but sometimes to a sub-classification term. We selected only variants associated with the highest order EFO terms per category, and assessed association between EFO traits and ALG ancestry using a global chi-square analyses set with an FDR < 0.05. For those traits that rejected the null hypothesis based on the global FDR, a 2x2 Odd's ratio with 95% confidence intervals were calculated for those traits/ancestry combinations whose standardized residuals were > ± 3.84.

Gene-level analyses: Characteristics of genes with assigned ancestry or associated with a GWAS variant. We tabulated the possible protein-coding genes and the frequency of association for each GWAS variant from the GWAS catalogue. For SNPs that are intra-genic, the GWAS catalogue reports a single putative associated gene, the one flanking the focal SNP, while for inter-genic SNPs, the catalogue typically reports a minimum of two protein coding genes (those immediately up and down-stream of the variant), plus additional non-coding or splice variants. We retained a singular gene for intra-genic SNPs and no more than two protein-coding variants

per SNP, and counted the number of times associations were observed. To assess characteristics of genes associated with GWAS or that have an ALG ancestry, we merged this list and the list of genes that were assigned to an ALG (tabulated by ENSEMBL ID's), with a database of gene-level metrics of all human protein coding genes curated at gnomAD v 2.1.1 (gnomad.v2.1.1.lof_metrics.by_gene) (last accessed, July 15, 2020).

Gene-level Constraint. The degree of evolutionary constraint of GWAS-associated genes and those with ancestry assigned to an ALG was compared using two metrics of the relative intolerance to loss of function of mutations for genes i) associated with GWAS studies ii) mapping to an ALG and iii) those with unassigned ancestry. Firstly, we compared the distribution of the statistic, "loss-of-function observed/expected upper bound fraction (LOEUF)". This statistic is akin to a measure of haploinsufficiency on a continuous scale, in which genes with a lower value evolve under strong selection against predicted loss-of-function variation, and those with values = 1 are not constrained. Secondly, we employed a dichotomous categorization of genes as being haploinsufficient if the probability of being loss of function intolerant (pLi) was >0.9. Values of LOEUF and pLI were obtained from gnomAD V 2.1.1 (Karczewski et al. 2021) (last accessed, July 15, 2020).

Expression of genes on ALG's enriched for EFO terms. Given the finding that genes mapping to different ALG's were associated with different traits based on the EFO classification, we examined the relationship between ALG enrichment and gene expression in tissues relevant to EFO enrichment. To this end, we obtained the median transcripts per million (TPM) for all genes and tissues included in GTEx V8 (gtexportal.org, last date accessed, November 15, 2020), and recorded the tissue with the highest median TPM per gene. Tissues were grouped into biologically meaningful groups a) Adipose (sub-cutaneous and visceral (omentum)) b) Blood-

Vascular (Whole Blood, arteries (Coronary, Tibial), heart (left ventricle; atrial appendage) c) Brain (all sections) d) Cells (EBV-transformed lymphocytes and cultured Fibroblasts) e) Digestive System (Esophagous, stomach, colon, intestine, pancreas) f) Female Reproductive system (breast, uterus, cervix, vagina, fallopian tube) g) Lung h) Muscle i) Nerve Tibial, j) Skin (exposed and not-exposed) k) spleen l) Male Reproduction (Testis, prostate) m) Glands (Thyroid, Pituitary, Adrenal and Salivary) n) urinary (Bladder, Kidney). Overall patterns of gene expression by ALG and tissue/EFO terms were explored; in particular, the relationship between genes mapping to ALG's that were enriched for the EFO terms "brain" and "visual traits" were assessed for expression of genes in nervous tissues. Additionally, differences in gene expression for genes with and without ALG-ancestry were examined by GTEx grouped tissue. Lastly, the ALG ancestry and tissue of expression for genes identified as being core genes in lipid metabolism (Table 1) by Dron et al., (2016) were extracted (Dron and Hegele 2016).

Statistical and graphical Analysis. Fisher's Exact tests were used to compare associations of genes with different classifications (e.g. associated with a GWAS or not associated, mapping to an ALG or not-mapping). Odd's ratio (OR) and 95% confidence intervals calculated from 2x2 tables according to Altman (1991) for rows with complete data (no NA's in any cell counted for the 2x2 table). A global G-test and individual Pearson Chi-Square tests with Yates continuity correction were performed to assess the difference in observed (GWAS) vs expected (HapMap) proportion of SNPs mapping to the ALGs. Kolmogorv-Smirnov tests were performed to compare the distribution of continuous variables (e.g. LOEUF) between discrete classes. Logistic regression was used to assess the influence of six gene traits or measures of evolutionary constraint (all uncorrelated with pearson's $r^2 = 0.81$ or less) on being associated with an ALG or GWAS variant, and removing terms that had low explanatory value and using an FDR = 0.05. T-

tests or Mann-Whitney U tests were used to compare the means of normally and the ranks of non-normally distributed variables respectively. Chromosome plots were generated with R packages BioCircos (Cui et al. 2016) and ChromPlots (Orostica and Verdugo 2016).

Availability of data and materials

The source data and R and python codes used for all analyses are available at https://github.com/saravictoriagood/GWAS-origins

Competing interests.

The authors declare that they have no competing interests.

Acknowledgements
The authors would like to thank Jaanus Sarvoli for comments on an earlier version of this manuscript. The research was supported by a CIHR Strategic Training in Advanced Genetic Epidemiology (STAGE) fellowship to SVG, by an NSERC – DG to SVG, and CIHR project grant to ADP.

LITERATURE CITED

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**: Unit7 20.

Alekseyev MA, Pevzner PA. 2007. Are there rearrangement hotspots in the human genome? *PLoS Comput Biol* **3**: e209.

Alekseyev MA, Pevzner PA. 2010. Comparative genomics reveals birth and death of fragile regions in mammalian evolution. *Genome Biol* **11**: R117.

Becker TS, Lenhard B. 2007. The random versus fragile breakage models of chromosome evolution: a matter of resolution. *Mol Genet Genomics* **278**: 487-491.

Berthelot C, Muffato M, Abecassis J, Roest Crollius H. 2015. The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep* **10**: 1913-1924.

Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res* **15**: 98-110.

Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**: 1177-1186.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**: D1005-D1012.

Casola C. 2018. From De Novo to "De Nono": The Majority of Novel Protein-Coding Genes Identified with Phylostratigraphy Are Old Genes or Recent Duplicates. *Genome Biol Evol* **10**: 2906-2918.

Chen WH, Zhao XM, van Noort V, Bork P. 2013. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput Biol* **9**: e1003073.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**: 2745-2747.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**: 1496-1502.

Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, Feschotte C. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371**.

Cui Y, Chen X, Luo H, Fan Z, Luo J, He S, Yue H, Zhang P, Chen R. 2016. BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics* **32**: 1740-1742.

Damas J, Corbo M, Lewin HA. 2021. Vertebrate Chromosome Evolution. *Annu Rev Anim Biosci* **9**: 1-27.

Damas J, Kim J, Farre M, Griffin DK, Larkin DM. 2018. Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. *Genome Biol* **19**: 155.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**: e1001025.

Dickerson JE, Robertson DL. 2012. On the origins of Mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol* **29**: 61-69.

Dron JS, Hegele RA. 2016. Genetics of Lipid and Lipoprotein Disorders and Traits. *Curr Genet Med Rep* **4**: 130-141.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17**: 68-74.

Eilbeck K, Quinlan A, Yandell M. 2017. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet* **18**: 599-612.

Farre M, Kim J, Proskuryakova AA, Zhang Y, Kulemzina AI, Li Q, Zhou Y, Xiong Y, Johnson JL, Perelman PL et al. 2019. Evolution of gene regulation in ruminants differs between evolutionary breakpoint regions and homologous synteny blocks. *Genome Res* **29**: 576-589.

Flajnik MF. 2018. A cold-blooded view of adaptive immunity. *Nat Rev Immunol* **18**: 438-453.

Flajnik MF, Kasahara M. 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics* **11**: 47-59.

Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. 2019. Measuring intolerance to mutation in human genetics. *Nat Genet* **51**: 772-776.

Gout JF, Kahn D, Duret L, Paramecium Post-Genomics C. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet* **6**: e1000944.

Gulko B, Hubisz MJ, Gronau I, Siepel A. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* **47**: 276-283.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**: 9362-9367.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95-108.

Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M. 2013. VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* **37**: 622-634.

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D et al. 2016. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**: 877-885.

Jabbari K, Chakraborty M, Wiehe T. 2019. DNA sequence-dependent chromatin architecture and nuclear hubs formation. *Scientific Reports* **9**: 14646.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**: 434-443.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP et al. 2021. Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **590**: E53.

Kemkemer C, Kohn M, Cooper DN, Froenicke L, Hogel J, Hameister H, Kehrer-Sawatzki H. 2009. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC Evol Biol* **9**: 84.

Kennedy B, Kronenberg Z, Hu H, Moore B, Flygare S, Reese MG, Jorde LB, Yandell M, Huff C. 2014. Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Curr Protoc Hum Genet* **81**: 6 14 11-25.

Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL. 2019. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet* **104**: 65-75.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073-1081.

Lal D, May P, Perez-Palma E, Samocha KE, Kosmicki JA, Robinson EB, Moller RS, Krause R, Nurnberg P, Weckhuysen S et al. 2020. Gene family information facilitates variant interpretation and

identification of disease-associated genes in neurodevelopmental disorders. *Genome Med* **12**: 28.

Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res* **19**: 770-777.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.

Liu X, Li YI, Pritchard JK. 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**: 1022-1034 e1026.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. *Trends in genetics : TIG* **25**: 152-155.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A* **107**: 9270-9274.

Marletaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, Wyatt CDR, de la Calle-Mustienes E, Bertrand S, Burguera D et al. 2018. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**: 64-70.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**: e121.

Nakatani Y, Shingate P, Ravi V, Pillai NE, Prasad A, McLysaght A, Venkatesh B. 2021. Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. *Nature communications* **12**: 4489.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**: 1254-1265.

Ohta Y, Kasahara M, O'Connor TD, Flajnik MF. 2019. Inferring the "Primordial Immune Complex": Origins of MHC Class I and Antigen Receptors Revealed by Comparative Genomics. *J Immunol* **203**: 1882-1896.

Orostica KY, Verdugo RA. 2016. chromPlot: visualization of genomic data in chromosomal context. *Bioinformatics* **32**: 2366-2368.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158**: 927-931.

Peng Q, Pevzner PA, Tesler G. 2006. The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput Biol* **2**: e14.

Pevzner P, Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A* **100**: 7672-7677.

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064-1071.

Rhie A McCarthy SA Fedrigo O Damas J Formenti G Koren S Uliano-Silva M Chow W Fungtammasan A Kim J et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**: 737-746.

Rice AM, McLysaght A. 2017. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat Commun* **8**: 14366.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* **21**: 108-116.

Sacerdot C, Louis A, Bon C, Berthelot C, Roest Crollius H. 2018. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome biology* **19**: 166-166.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A et al. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**: 944-950.

772 Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the
773 deep-sequencing age. *Nat Methods* **11**: 361-362.
774 Sella G, Barton NH. 2019. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide
775 Association Studies. *Annual review of genomics and human genetics* **20**: 461-493.
776 Shi H, Kichaev G, Pasaniuc B. 2016. Contrasting the Genetic Architecture of 30 Complex Traits from
777 Summary Association Data. *Am J Hum Genet* **99**: 139-153.
778 Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013. Predicting
779 the functional, molecular, and phenotypic consequences of amino acid substitutions using
780 hidden Markov models. *Hum Mutat* **34**: 57-65.
781 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW,
782 Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
783 genomes. *Genome Res* **15**: 1034-1050.
784 Simakov O, Marletaz F, Yue JX, O'Connell B, Jenkins J, Brandt A, Calef R, Tung CH, Huang TK, Schmutz J et
785 al. 2020. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol*
786 **4**: 820-830.
787 Simons YB, Bullaughey K, Hudson RR, Sella G. 2018. A population genetic interpretation of GWAS
788 findings for human quantitative traits. *PLoS Biol* **16**: e2002985.
789 Singh PP, Arora J, Isambert H. 2015. Identification of Ohnolog Genes Originating from Whole Genome
790 Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS*
791 *Comput Biol* **11**: e1004394.
792 Taliun D Harris DN Kessler MD Carlson J Szpiech ZA Torres R Taliun SAG Corvelo A Gogarten SM Kang HM
793 et al. 2021. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*
794 **590**: 290-299.
795 Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet* **15**: e1008160.
796 Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* **90**:
797 7-24.
798 Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS
799 Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**: 5-22.
800 Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG. 2011. A probabilistic disease-
801 gene finder for personal genomes. *Genome Res* **21**: 1529-1542.
802 Yegorov S, Good S. 2012. Using Paleogenomics to Study the Evolution of Gene Families: Origin and
803 Duplication History of the Relaxin Family Hormones and Their Receptors. *Plos One* **7**: e32923.

804

805

Figure Legends

Fig. 1 Circular maps of the human genome (minus Y-chromosome) partitioned by ALG. Outermost circle, ideogram of C-banding patterns; second ring, blocks of the human genome coloured by the 17 ancestral ALGs; third ring, location of GWAS SNPs in the GWAS catalogue coloured by ancestral ALG ($p<10^{-08}$). Innermost, ribbons connecting regions of the human genome with > 2Mbp of contiguous sequence belonging to each ALG (top left). Circular Chromosome maps showing broad syntenic regions of ALG B (top right), ALG A (bottom left) and ALG's C & L (bottom right) with examples of ohnologous gene families membership indicated (outermost ring), ideogram with gene individual gene names, GWAS SNP density (third ring) and ribbons connecting regions with > 1Mb of contiguous sequence.

Fig. 2 a) Density of 49,129 GWAS significant variants coloured by ALG and number of HapMap SNPs in 200kb windows (gray line). The x-axis denotes the position in million of base pairs (Mbp) and the Y-axis the density of SNPs on a log-scale. b) Proportion of HapMap and GWAS significant SNPs by ALG, excluding those unassigned to a ALG.

Fig. 3 Odd's ratio and 95% confidence intervals for the association of the EFO traits assigned to 42,999 GWAS variants in the GWAS catalogue. Traits showing an excess (OR >1) or deficit (OR <1) of variants by ALG were those in which the standardized residuals from a chi-square analysis were = +/- 3.841.

Fig. 4 a) Weighted proportion of genes on a ALG that are GWAS. The dotted line marks 63.7, the overall percent of GWAS associated genes that map to any CLG. Total number of genes is 19,705. b) Odds ratio for genes that are associated with a GWAS variant or assigned to an ALG have a pLI > 0.9, a measure of haploinsufficiency c) the mean number of nonsynonymous mutations (dN) of genes between human and mouse (dN_{ALG} =0.071, $dN_{unassigned}$ = 0.132, t=-31.95, $p<2.2^{e-16}$) d) distribution of the number of coding exons for genes that map to an ALG or are unassigned (mean/median ALG: 12.5/10, mean/median unassigned: 7.5/5, W = 60266263, p-value < 2.2e-16) e) Median expression (in transcripts per million, TPM) of genes assigned to an ALG (median = 54.0) or unassigned (median = 33.3) in the tissue of highest expression obtained from GTEx (Mann-Whitney U -test W= 52949771, $p < 2.2^{e-16}$) f) Distribution of the length of genes associated with a GWAS vs those not associated (mean length of GWAS associated genes 94069.35 vs 33134 bp not associated, t = 36.819, df = 14156, p-value < 2.2e-16). g) Distribution of the loss-of-function observed/expected upper bound fraction (LOEUF) was significantly lower for genes that map ALG (Kolmogorov-Sminov D=0.186, p<2.2 e-16) and h) for genes associated with a GWAS variant (D=0.133, D=, $p<2.2^{e-16}$).

Fig. 5 a) Circular map depicting the chromosomes with ALG assignment (outer ring), ideogram (second ring), and tissue of highest expression (innermost ring) for 19 genes associated with core processes in lipid metabolism (see text for details). b) The human chromosomes housing blocks derived from ALG A, C and E (coloured by ALG), enriched for brain-related functions. Median expression (blue dots, left of chromosomes) of genes on ALG A, C or E whose highest tissue of expression in GTEx is brain. Overall gene density (gray histogram) and gene density in brain (red histogram) – right side of chromosomes.

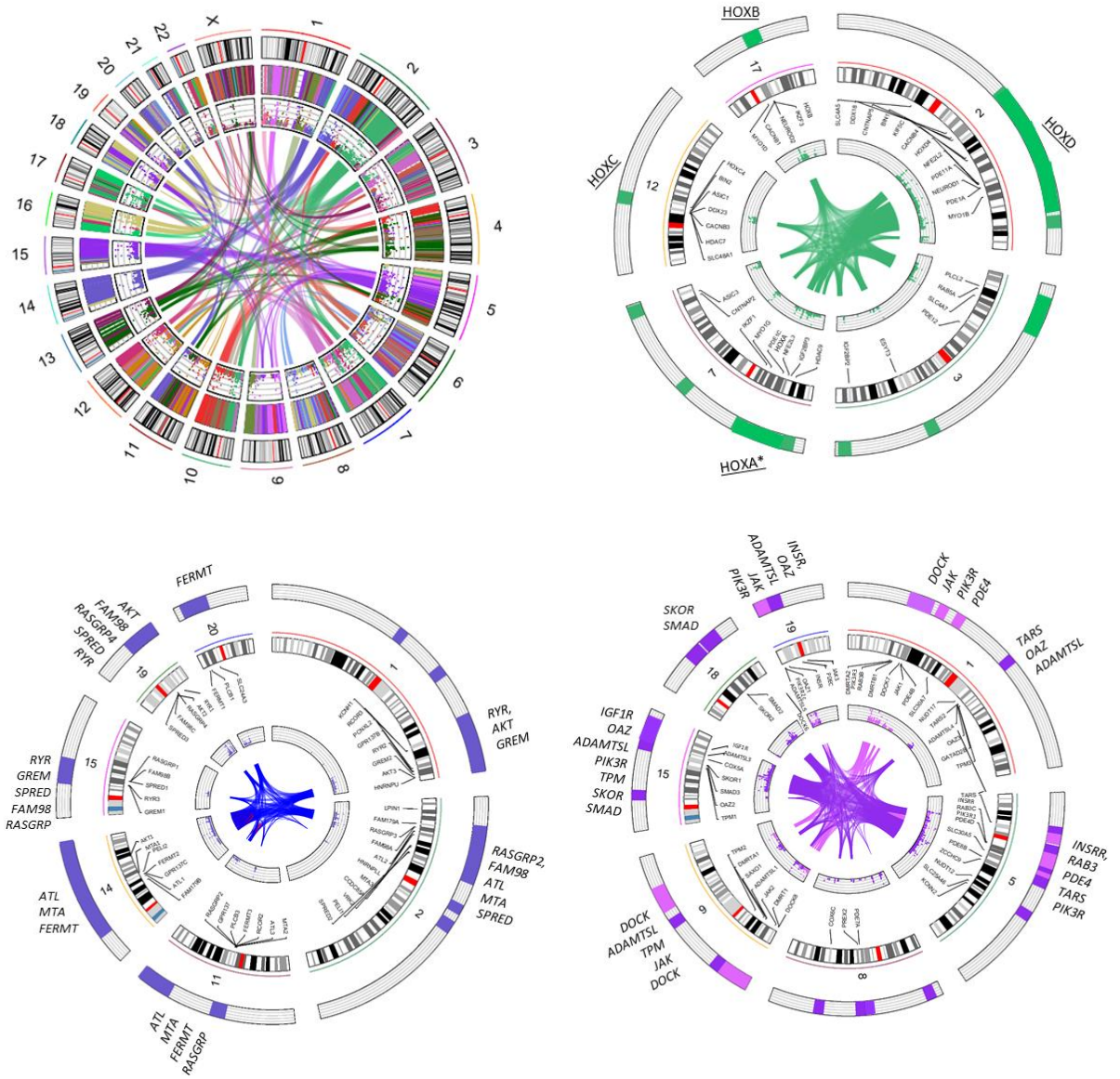

HOXB
HOXD
HOXC
HOXA*
FERMT
AKT
FAM98
RASGRP4
SPRED
RYR
RYR,
AKT
GREM
RYR
GREM
SPRED
FAM98
RASGRP
RASGRP2,
FAM98
ATL
MTA
SPRED
ATL
MTA
FERMT
ATL
MTA
FERMT
RASGRP
ADAMTSL
JAK
PIK3R
INSR,
OAZ
ADAMTSL
DOCK
JAK
PIK3R
PDE4
SKOR
SMAD
TARS
OAZ
ADAMTSL
IGF1R
OAZ
ADAMTSL
PIK3R
TPM
SKOR
SMAD
INSRR,
RAB3
PDE4
TARS
PIK3R
DOCK
ADAMTSL
TPM
JAK
DOCK
A B C D E F G H I J K L M N O P Q Un

Fig. 1

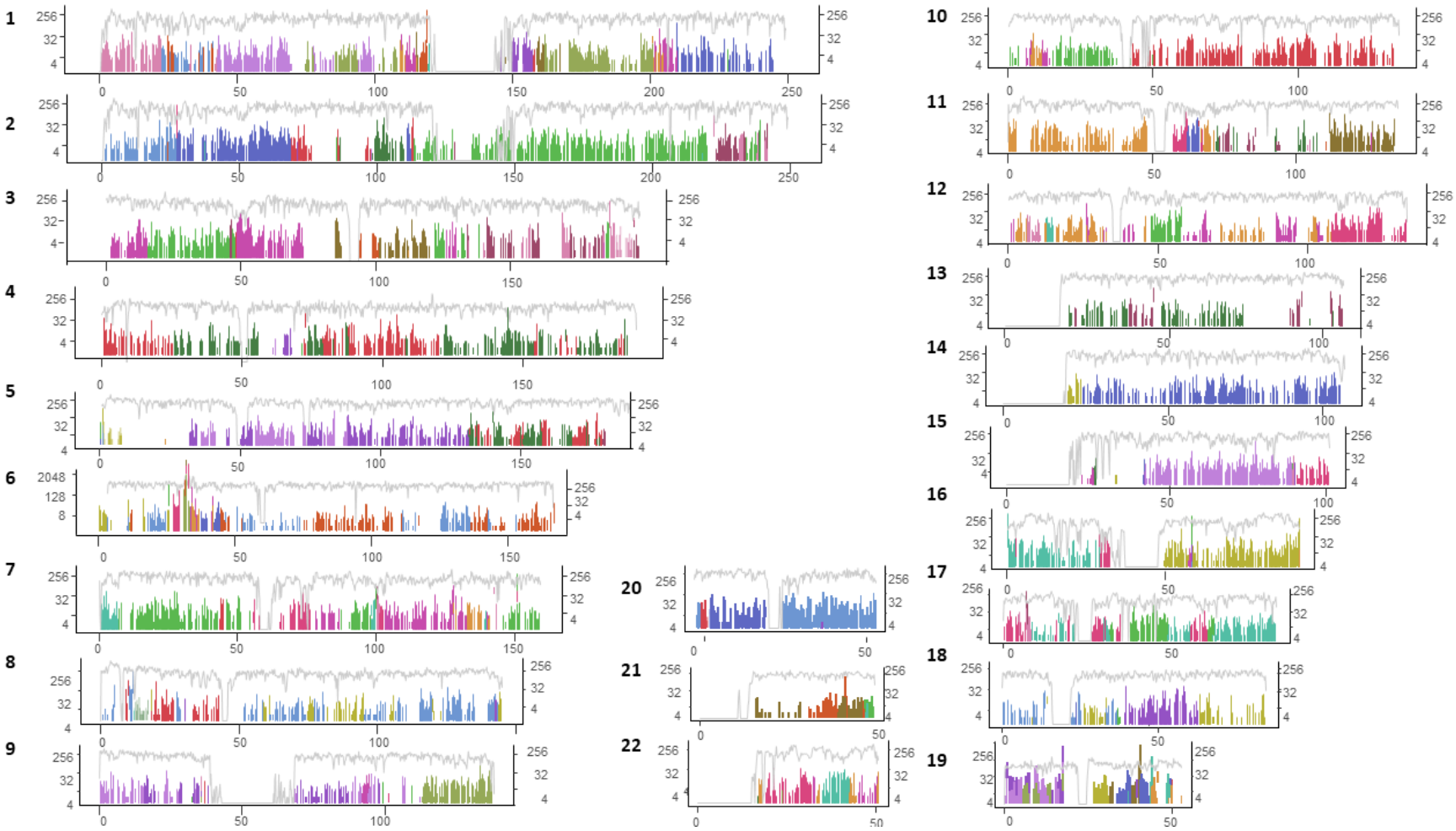

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
A B C D E F G H I J K L M N O P Q Un

Fig. 2A

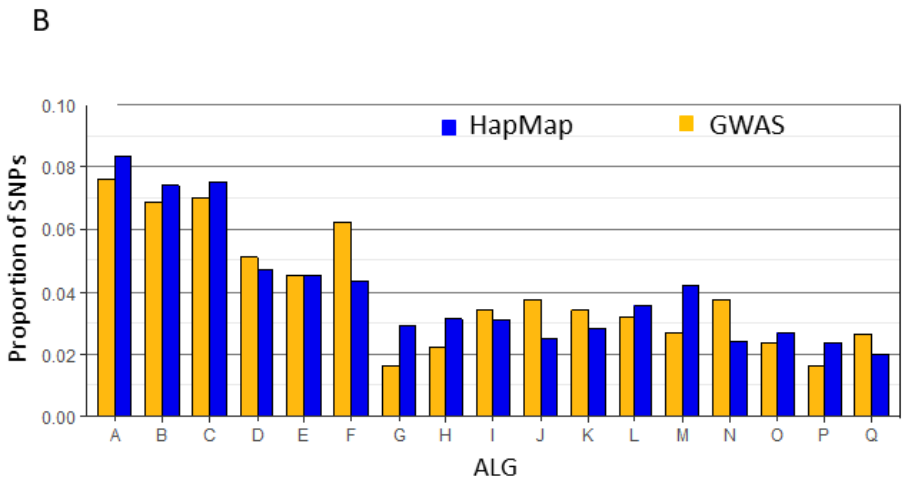

B
HapMap
GWAS
Proportion of SNPs
0.10
0.08
0.06
0.04
0.02
0.00
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
ALG

Fig. 2B

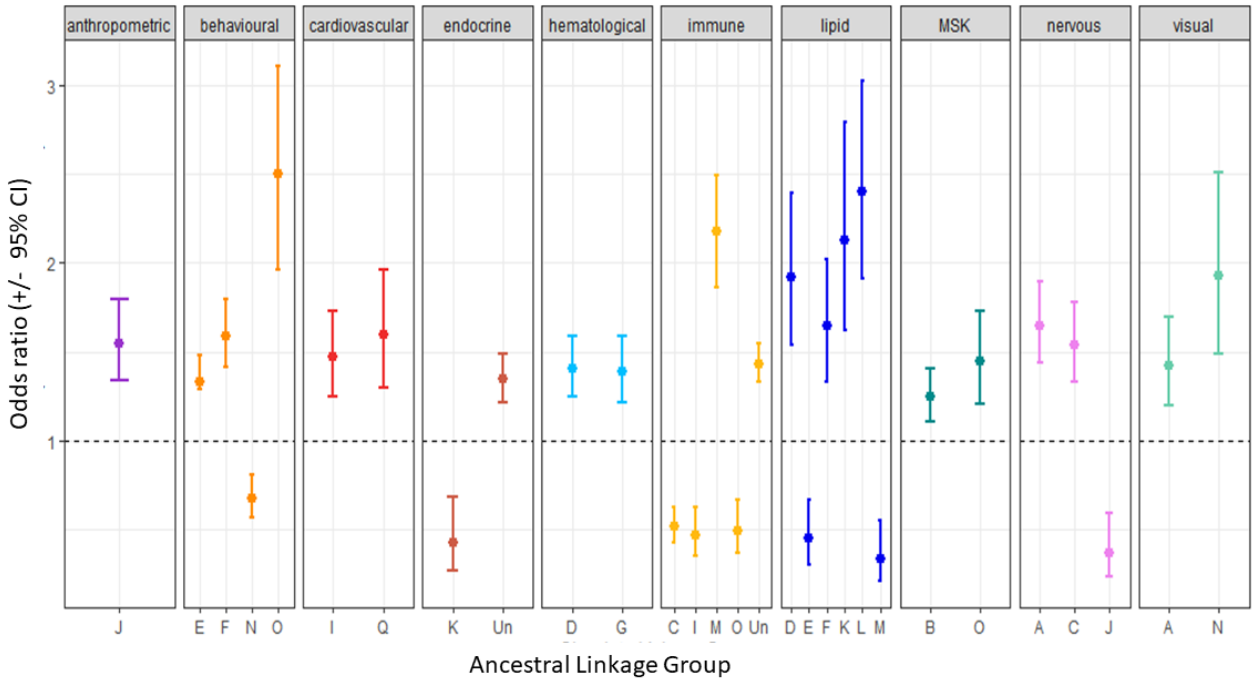

anthropometric
behavioural
cardiovascular
endocrine
hematological
immune
lipid
MSK
nervous
visual
Odds ratio (+/- 95% CI)
3
2
1
J
E F N O
I Q
K Un
D G
C I M O Un
D E F K L M
B O
A C J
A N
Ancestral Linkage Group

Fig. 3

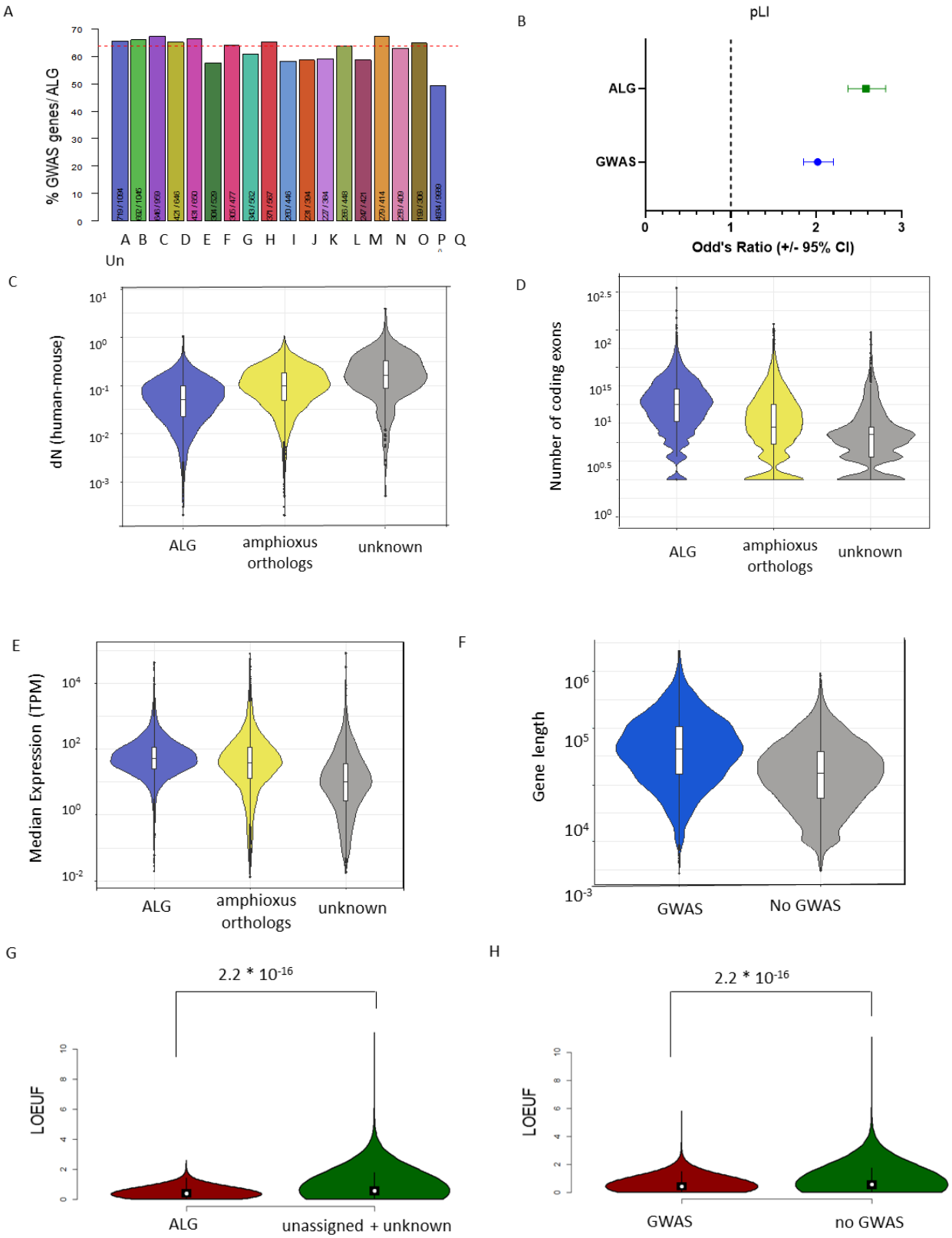
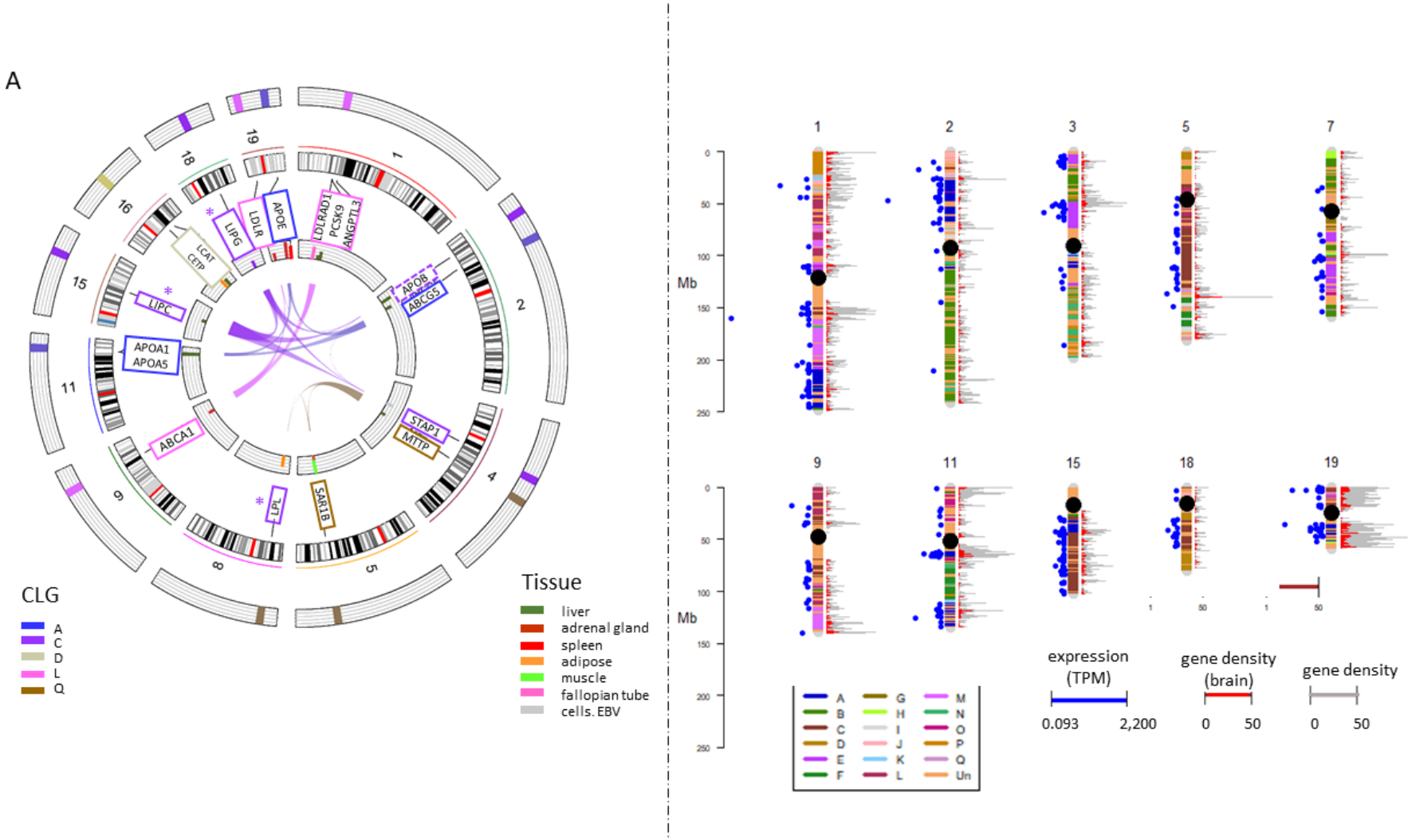

A
% GWAS genes/ ALG
A B C D E F G H I J K L M N O P Q
Un
B
pLI
ALG
GWAS
Odd's Ratio (+/- 95% CI)
C
dN (human-mouse)
ALG
amphioxus orthologs
unknown
D
Number of coding exons
ALG
amphioxus orthologs
unknown
E
Median Expression (TPM)
ALG
amphioxus orthologs
unknown
F
Gene length
GWAS
No GWAS
G
2.2 * 10-16
LOEUF
ALG
unassigned + unknown
H
2.2 * 10-16
LOEUF
GWAS
no GWAS

Fig. 4


A
CLG
A
C
D
L
Q
Tissue
liver
adrenal gland
spleen
adipose
muscle
fallopian tube
cells. EBV
APOE
LDLR
LIPG
LCAT
CETP
LIPC
APOA1
APOA5
ABCA1
LPL
SAR1B
MTTP
APOB
ABCG5
PCSK9
ANGPTL3
LDLRAD1
Mb
expression (TPM)
0.093
2,200
gene density (brain)
gene density
0
50

32

33 **Fig. 5**