Analysis of the global frequency and penetrance of *ATP7B* variants: implications for Wilson disease prevalence

Daniel F. Wallace[1*], James S. Dooley[2]

[1] Institute of Health and Biomedical Innovation and School of Biomedical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia.

[2.] UCL Institute for Liver and Digestive Health, Division of Medicine, University College London Medical School (Royal Free Campus), London, UK.

* Corresponding author:

Email: d5.wallace@qut.edu.au (DW)

Abstract

Wilson disease (WD) is a genetic disorder of copper metabolism. It can present with hepatic and neurological symptoms, due to copper accumulation in the liver and brain. WD is caused by compound heterozygosity or homozygosity for mutations in the copper transporting P-type ATPase gene *ATP7B*. Over 700 *ATP7B* genetic variants have been associated with WD. Estimates for WD population prevalence vary with 1 in 30,000 generally quoted. However, some studies have estimated much higher prevalence rates. The aim of this study was to estimate the population prevalence of WD by determining the frequency and evaluating the pathogenicity of *ATP7B* variants in a genomic sequence database. A catalogue of 732 WD-associated *ATP7B* variants was constructed using data from the WD Mutation Database and a literature review. A total of 231 WD-associated *ATP7B* variants were present in the gnomAD dataset giving an estimated population prevalence of around 1 in 2400 with a carrier rate of 1 in 25. Pathogenicity of the variants was assessed by (a) comparing gnomAD allele frequencies against the number of reports for variants in the WD literature and (b) using variant effect prediction algorithms. After exclusion of WD-associated *ATP7B* variants with predicted low penetrance, the revised estimates showed a prevalence of around 1 in 20,000, with higher rates in the Asian and Ashkenazi Jewish populations. *Conclusion:* We have calculated the prevalence of WD based on genomic sequencing data and our results highlight the importance of assessing penetrance when assigning causality to genetic variants. The high frequency of low penetrant *ATP7B* variants raises the possibility that these variants could contribute to abnormalities in copper homeostasis that do not manifest in a clear WD phenotype and diagnosis.

Author Summary

Wilson disease is a genetic disorder that causes copper accumulation in the liver and brain. It is caused by mutations in the *ATP7B* gene that encodes a protein involved in transporting copper across cell membranes. We used genomic sequencing data from more than 120,000 people from 8 global populations to estimate the prevalence of mutations that cause Wilson disease. From these data we calculated the predicted prevalence of Wilson disease and found that it is much higher than traditional estimates. Further analysis revealed that this high prevalence is likely due to several mutations that are too common to be a major cause of the disease and may only have mild effects on ATP7B protein function. After taking these mild mutations into account in our estimates of disease prevalence, we predict that Wilson disease has a population prevalence of around 1 in 20,000 with higher rates in East Asian and Ashkenazi Jewish populations. Our results suggest that some mutations in ATP7B may cause milder forms of Wilson disease.

Introduction

Wilson disease (WD) is a rare autosomal recessive disorder of copper metabolism, resulting in copper accumulation with, most characteristically, hepatic and/or neurological disease [1]. It is caused by mutations in the gene encoding ATP7B, a copper transporter which in hepatocytes not only transports copper into the transGolgi for association with apoceruloplasmin, but is fundamental for the excretion of copper into bile [1].

In WD copper accumulates in the liver, causing acute and/or chronic hepatitis and cirrhosis. Neuropsychiatric features are seen due to accumulation of copper in the brain. Other organs and tissues involved include the cornea (with the development of Kayser-Fleischer rings) and the kidneys.

There is a wide clinical phenotype and age of presentation. Early diagnosis and treatment are important for successful management. Diagnosis can be straightforward with a low serum ceruloplasmin associated with Kayser-Fleischer rings in the eyes, but may be difficult, requiring further laboratory tests, liver copper estimation and molecular genetic studies for *ATP7B* mutations.

Over 700 mutations in *ATP7B* have been reported as associated with WD. The majority of patients are compound heterozygotes, the minority being homozygous for a single mutation. Phenotype/genotype studies to date have shown a poor relationship [2, 3], and there have been studies and increasing interest in modifying genes and factors [4].

Currently treatment of WD is either with chelators (d-penicillamine or trientine) which increase urinary copper excretion or zinc salts which reduce intestinal copper absorption [1]. Liver transplantation may be needed for acute liver failure or decompensated liver disease unresponsive to treatment [1].

The prevalence of WD has been studied in several ways. In 1984 Scheinberg and Sternlieb [5] from their own data, the report from Bachmann et al [6] based on an accurately ascertained incidence, and data from Japan published by Saito [7], concluded that the worldwide prevalence of WD is around 30 per million. Screening studies using a low ceruloplasmin as the target have suggested that WD may be much more frequent [8, 9]. A molecular genetic study of 1000 control subjects in the UK, however, found an estimated potential prevalence of 1 in 7000 [10]. Next generation DNA sequencing (NGS) databases provide the opportunity to analyse the prevalence of WD mutations in large populations and sub-populations. The gnomAD database contains variant frequencies derived from the whole exome or whole genome sequencing of over 120,000 people, from eight ethnic subgroups. NGS datasets are valuable resources and have been used by us and others for estimating the population prevalence of genetic diseases, such as HFE and non-HFE hemochromatosis [11] and primary ubiquinone deficiency [12].

This study has: (1) collated reported variants in patients with WD; (2) searched a NGS dataset to define the prevalence of these variants, and (3) refined the prevalence data by analysing differences in variant penetrance.

The resulting prevalence derived from this study is intermediate between historical estimates and those from more recent studies, at approximately 1 in 19,500, with variation above and below this in specific populations.

Results

Wilson disease-associated *ATP7B* variants

The WDMD contained 525 unique *ATP7B* variants that have been reported in patients with WD and classified as disease causing (Supporting Table S1). A literature search (between 2010 and April 2017) revealed a further 207 unique *ATP7B* variants associated with WD since the last update of the WDMD (Supporting Table S2). Thus 732 *ATP7B* variants predicted to be causative of WD have been reported up until April 2017. For this study we refer to these 732 variants as WD-associated *ATP7B* (WD-*ATP7B*) variants.

The WD-*ATP7B* variants were categorized into their predicted functional effects, with the majority (400) being single base missense (non-synonymous) substitutions (Table 1). Variants predicted to cause major disruption to the protein coding sequence were further classified as loss of function (LoF). Variants were considered LoF if they were frameshift, stop gain (nonsense), start loss, splice donor, splice acceptor variants or large deletions involving whole exons. A total of 279 WD-*ATP7B* variants were categorized as LoF (Table 1) and their pathogenicity was considered to be high.

Table 1. Predicted functional consequences of WD-*ATP7B* variants

Variant category	**Number of variants**	**Loss of function (LoF)**	**Number in gnomAD**
Missense (non-synonymous)	400 (55%)		158 (68%)
Frameshift deletions, insertions or substitutions	170 (23%)	Yes	23 (10%)
Stop gain (nonsense)	64 (9%)	Yes	22 (10%)
Splice donor or acceptor	43 (6%)	Yes	10 (4%)
Non-frameshift deletions, insertions or substitutions	26 (4%)		4 (2%)
Intronic variants	22 (3%)		13 (6%)
Promoter variants	2 (0.3%)		
5’ UTR variants	2 (0.3%)		1 (0.4%)
Large deletions	2 (0.3%)	Yes	
Stop loss	1 (0.1%)		
Total	732		231

Prevalence of WD-*ATP7B* variants in the gnomAD dataset

Of the 732 WD-*ATP7B* variants 231 were present in the gnomAD dataset derived from >120,000 individuals [13] (Table 1). There was a higher proportion of missense variants among the WD-*ATP7B* variants present in gnomAD compared to the total WD-*ATP7B* variants reported in the literature (68% compared to 55%; Fisher’s Exact test, p=0.0002). Consequently there were also fewer LoF variants among the WD-*ATP7B* variants present in gnomAD (24% compared to 38%; Fisher’s Exact test, p<0.0001).

Predicted prevalence of WD-associated genotypes in the gnomAD populations

Allele frequencies of all WD-*ATP7B* variants present in the gnomAD dataset were summed to give an estimate for the combined allele frequency of all WD-*ATP7B* variants in the general population, which we have termed the pathogenic allele frequency (PAF). This was done for the entire gnomAD population and also for the 8 subpopulations that make up this dataset (Table 2). Assuming Hardy-Weinberg equilibrium and using the Hardy-Weinberg equation, the PAFs were used to calculate the pathogenic genotype frequencies (being homozygous or compound heterozygous for WD-*ATP7B* variants), the heterozygous genotype frequencies (being heterozygous for WD-*ATP7B* variants) and the carrier rates for these genotypes, expressed as one per “n” of the population (Table 2). The PAF in the whole gnomAD dataset was 2.0%, giving a pathogenic genotype rate (PGR) of 1 in 2491 and heterozygous carrier rate of 1 in 25. The highest PAFs were seen in the Ashkenazi Jewish population (PAF 3.0%, PGR 1 in 1107) and the East Asian population (PAF 2.4%, PGR 1 in 1799) and the lowest in the African population (gnomAD: PAF 1.2%, PGR 1 in 7271).

Table 2. Combined WD-*ATP7B* variant allele frequencies, genotype frequencies and carrier rates in the gnomAD population

	gnomAD								
	All	**African**	**Ashkenazi Jewish**	**East Asian**	**European (non-Finnish)**	**European (Finnish)**	**Latino**	**South Asian**	**Other**
Pathogenic allele freq	0.02004	0.01173	0.03005	0.02358	0.02278	0.01744	0.01629	0.01546	0.02142
Pathogenic genotype freq	0.00040	0.00014	0.00090	0.00056	0.00052	0.00030	0.00027	0.00024	0.00046
Heterozygous genotype freq	0.03927	0.02318	0.05830	0.04604	0.04452	0.03427	0.03204	0.03044	0.04191
Pathogenic genotype carrier rate[a]	2491	7271	1107	1799	1927	3289	3770	4184	2180
Heterozygous carrier rate[a]	25	43	17	22	22	29	31	33	24

[a] Pathogenic genotype rate and heterozygous carrier rate are expressed as 1 in “n” of the population.

The above estimates do not account for WD variants that are present in the gnomAD populations but have not been reported in the literature. We made the assumption that *ATP7B* LoF variants would almost certainly be causative of WD when in the homozygous state or compound heterozygous state with other pathogenic *ATP7B* variants. We identified an additional 51 LoF variants present in the gnomAD dataset not reported in the literature as associated with WD (Supporting Table S3).

The ExAC database, a forerunner of gnomAD, that contains approximately half the number of genomic sequences, also reports copy number variants (CNVs) from 59,898 of the 60,706 exomes in the database [14]. We analysed the CNVs that intersected with the *ATP7B* gene and identified 10 deletions and 6 duplications that covered either all or part of the gene. The gnomAD database currently has no data on CNVs, however, as the ExAC database formed the basis for gnomAD we added frequency data derived from ExAC CNVs to our analysis. The CNV *deletions* were considered to be pathogenic and were included as large LoF deletions in subsequent PAF calculations (Supporting Table S3). As it was not straight forward to determine whether CNV *duplications* were pathogenic they were not included in the analysis.

The allele frequencies of the LoF variants and large LoF deletions were added to the gnomAD PAFs determined previously. The updated PAFs, genotype frequencies and carrier rates were calculated (Table 3). The additional LoF variants were only rarely encountered in the gnomAD populations and their inclusion did not contribute greatly to the overall PAFs and carrier rates, with only marginal increases (PAF 2.0%, PGR 1 in 2387).

Table 3. Combined WD-*ATP7B* plus LoF variant allele frequencies, genotype frequencies and carrier rates in the gnomAD population

	gnomAD								
	All	**African**	**Ashkenazi Jewish**	**East Asian**	**European (non-Finnish)**	**European (Finnish)**	**Latino**	**South Asian**	**Other**
Pathogenic allele freq	0.02055	0.01245	0.03005	0.02369	0.02335	0.01775	0.01664	0.01591	0.02298
Pathogenic genotype freq	0.00042	0.00016	0.00090	0.00056	0.00055	0.00031	0.00028	0.00025	0.00053
Heterozygous genotype freq	0.04026	0.02459	0.05830	0.04627	0.04562	0.03486	0.03273	0.03131	0.04491
Pathogenic genotype rate[a]	2367	6451	1107	1781	1833	3176	3610	3952	1893
Heterozygous carrier rate[a]	25	41	17	22	22	29	31	32	22

[a] Pathogenic genotype rate and heterozygous carrier rate are expressed as 1 in “n” of the population.

Identification of low penetrant or non-causative *ATP7B* variants

Our estimate for the population prevalence of WD-*ATP7B* variants and consequently the predicted prevalence of WD in the gnomAD population of around 1 in 2400 with heterozygous carrier rate of 1 in 25 is considerably higher than the often quoted prevalence of 1 in 30,000 with 1 in 90 heterozygous carriers. It is, however, closer to the estimates obtained from the analysis of ceruloplasmin measurements in the Japanese and Korean populations [8, 9] and the result obtained by a genetic study of 1000 controls in the UK population of 1 in 7000 [10]. These higher prevalence estimates and the estimate we obtained from the gnomAD population, however, do not appear to reflect the incidence of WD presenting to the clinic and suggest that either many WD patients remain undiagnosed or that some WD-*ATP7B* variants are not causative or have low penetrance.

We addressed the issue of variant penetrance using two approaches: firstly, by comparing the allele frequencies of individual variants in the gnomAD dataset with the frequency with which these variants have been reported in association with WD in the literature; and secondly by utilizing VEP algorithms.

In the first approach, if the allele frequency in the gnomAD dataset was such that more reports would have been expected in the literature (analysed broadly by number of references) then the variant was considered as a ‘*probable* low penetrant’ variant. Thus, when we ranked WD-*ATP7B* variants according to their allele frequencies in the gnomAD population we noticed that the p.His1069Gln variant, the most common WD-associated variant in European populations, was ranked number 6 in the entire gnomAD dataset, number 5 in the European (Finnish and non-Finnish) subpopulations and number 3 in the Ashkenazi Jewish subpopulation. Thus there were several WD-*ATP7B* variants with higher allele frequencies in these populations that would be expected to be detected regularly in WD patients. The 5 WD-*ATP7B* variants that ranked higher than p.His1069Gln in the gnomAD dataset were p.Val536Ala, p.Thr1434Met, p.Met665Ile, p.Thr991Met and p.Pro1379Ser. These variants have only been reported in a small number of cases of WD and hence their causality and/or penetrance is in question.

We also attempted to identify variants that have questionable causality/penetrance by comparing them against a recent review article that analysed the geographic distribution of *ATP7B* variants that have been reported in WD patients [15]. This review lists the most commonly encountered *ATP7B* variants in WD patients from geographic regions around the world. Any variants reported in this article were considered to have high penetrance. Interestingly, the 5 variants with gnomAD allele frequencies higher than p.His1069Gln were not listed in the Gomes et al. review [15] suggesting that they are not commonly associated with WD.

We formalised this approach by analysing data from the WDMD. The WDMD lists all references that have reported particular variants. We counted the number of references associated with each WD-*ATP7B* variant (Supporting Table S1). The p.His1069Gln variant is listed against 46 references, the highest number for any variant in the WDMD. In contrast the 5 variants with higher gnomAD allele frequencies have only 1 or 2 associated references in the WDMD (Supporting Table S1), suggesting that their penetrance is low. We plotted gnomAD allele frequency against number of WDMD references for all WD-*ATP7B* variants and highlighted those variants that were reported by Gomes et al. [15] (Figure 1A). This analysis showed that there were a number of variants with relatively high allele frequencies in gnomAD, not reported in the Gomes et al. review paper and with few references in the WDMD. These variants are clustered towards the left-hand side of the graph in Figure 1A. On the basis of this analysis we classified 13 variants as having '*probable low penetrance*' (Table 4).

Table 4: WD-*ATP7B* variants (found in the gnomAD dataset) with probable or possible low penetrance

Coding DNA change	Protein change	Domain	gnomAD allele frequency	VEST3 score	Penetrance	References
c.406A>G	p.Arg136Gly	MBD1-2 linker	0.000313	0.182 *	Probable low	[16]
c.1555G>A	p.Val519Met	MBD5	0.000588	0.759	Probable low	[17]
c.1607T>C	p.Val536Ala	MBD5	0.003390	0.652	Probable low	[18]
c.1922T>C	p.Leu641Ser	MBD6-TMA linker	0.000462	0.893	Probable low	[19, 20]
c.1947-4C>T	.		0.000576		Probable low	[21, 22]
c.1995G>A	p.Met665Ile	TMA	0.001423	0.711	Probable low	[23]
c.2605G>A	p.Gly869Arg	A domain	0.000718	0.911	Probable low	[24-27]
c.2972C>T	p.Thr991Met	TM4	0.001259	0.96	Probable low	[19, 25]
c.3243+5G>A	.		0.000344		Probable low	[28]
c.3688A>G	p.Ile1230Val	P domain	0.000325	0.818	Probable low	[18]
c.4021+3A>G	.		0.000325		Probable low	[29]
c.4135C>T	p.Pro1379Ser	C-terminus	0.001063	0.864	Probable low	[19]
c.4301C>T	p.Thr1434Met	C-terminus	0.002060	0.249 *	Probable low	[30, 31]
c.122A>G	p.Asn41Ser	N-terminus	0.000224	0.149	Possible low	[32]
c.677G>A	p.Arg226Gln	MBD2-3 linker	0.000012	0.119	Possible low	WDMD
c.748G>A	p.Gly250Arg	MBD2-3 linker	0.000040	0.404	Possible low	[33]
c.997G>A	p.Gly333Arg	MBD3-4 linker	0.000004	0.124	Possible low	[29]
c.2183A>G	p.Asn728Ser	TM1-2	0.000032	0.203	Possible low	[34]
c.3490G>A	p.Asp1164Asn	N domain	0.000012	0.44	Possible low	[18]

c.3599A>C	p.Gln1200Pro	P domain	0.000020	0.299	Possible low	[35]
c.3886G>A	p.Asp1296Asn	P domain	0.000202	0.361	Possible low	[36, 37]
c.3971A>G	p.Asn1324Ser	TM5-6	0.000004	0.397	Possible low	[38]

* Probable low penetrant variants also classified as possible low penetrant variants based on a low VEST3 score.

Comparison of variant effect prediction algorithms

VEP algorithms are used extensively to predict whether amino acid substitutions (missense variants) are likely to alter protein function and hence contribute to disease. SIFT and Polyphen2 are two of the mostly widely used algorithms, however, in recent years newer algorithms have been developed. The output from wANNOVAR included results from 16 VEP algorithms. We tested the performance of these algorithms in discriminating between the 400 WD-*ATP7B* missense variants (identified in this study through literature review as associated with WD) and 786 missense variants (of 844 in total) that were identified in the gnomAD dataset but have not been previously reported in WD patients, termed non-WD-*ATP7B* missense variants. The scores for each of the algorithms were compared between the 2 groups (Supporting Figure S1) and their performance in discriminating between the 2 groups assessed using ROC curve analyses (Supporting Figure S2). Mean and median scores were compared between the two groups and the differences were statistically different for each algorithm (Supporting Figure S1, t-test p<0.01, Mann Whitney test p<0.0001). ROC curve analyses revealed area under the ROC curves that ranged between 0.5399 and 0.8821 (Supporting Figure S2).

The algorithm that performed the best at discriminating between WD and non-WD missense variants was VEST3 [39]. The median VEST3 score for WD missense variants was 0.957 compared with 0.404 for non-WD missense variants (Mann-Whitney test p<0.0001, AUROC 0.8821). None of the WD-*ATP7B* missense variants reported in the Gomes et al. review paper had VEST3 scores of less than 0.5 and only one variant with greater than 2 references in the WDMD had a VEST3 score of less than 0.5, indicating that the VEST3 score performs very well at discriminating between WD and non-WD *ATP7B* missense variants. We classified WD-*ATP7B* missense variants found in the WDMD and in our literature search as ‘*possible* low penetrance’ if they had a VEST3 score of <0.5 (Figure 1B). There were 11 such variants in the gnomAD dataset that were contributing to our initial estimates of WD prevalence (Table 4). Two of these variants were also classified as *probable* low penetrance in the previous analysis based on the number of publications.

Prevalence of WD-*ATP7B* variants in the gnomAD dataset after removing variants with probable or possible low penetrance

We recalculated the PAFs, genotype frequencies and carrier rates after excluding the variants we identified as having probable or possible low penetrance. Exclusion from the analysis of the 13 WD-*ATP7B* variants with *probable* low penetrance, based on relatively high allele frequencies but low numbers of reports in WD patients, resulted in a significant reduction in the predicted prevalence of WD. The updated PAF after exclusion of these variants was 0.76% in the gnomAD dataset, with PGR of 1 in 16,832. The updated PAFs, genotype frequencies and carrier rates, including the results for each subpopulation can be seen in Table 5.

Table 5. Combined WD-*ATP7B* plus LoF variant allele frequencies, genotype frequencies and carrier rates in the gnomAD population after exclusion of those variants with *probable* low penetrance.

	gnomAD								
	All	**African**	**Ashkenazi Jewish**	**East Asian**	**European (non-Finnish)**	**European (Finnish)**	**Latino**	**South Asian**	**Other**
Pathogenic allele freq	0.007708	0.003874	0.014093	0.015037	0.007884	0.004326	0.007724	0.006127	0.007902
Pathogenic genotype freq	0.000059	0.000015	0.000199	0.000226	0.000062	0.000019	0.000060	0.000038	0.000062
Heterozygous genotype freq	0.015297	0.007718	0.027790	0.029621	0.015645	0.008615	0.015328	0.012179	0.015680
Pathogenic genotype carrier rate[a]	16832	66625	5035	4423	16086	53427	16763	26636	16014
Heterozygous carrier rate[a]	65	130	36	34	64	116	65	82	64

[a] Pathogenic genotype rate and heterozygous carrier rate are expressed as 1 in “n” of the population.

The remaining 9 variants with *possible* low penetrance based on VEST3 score had lower allele frequencies and consequently their exclusion from the analyses had less effect on the predicted prevalence of WD. After exclusion of these variants the updated PAF decreased to 0.71% for the gnomAD dataset, with PGR of 1 in 19,457. The updated PAFs, genotype frequencies and carrier rates, including the results for each subpopulation can be seen in Table 6.

Table 6. Combined WD-*ATP7B* plus LoF variant allele frequencies, genotype frequencies and carrier rates in the gnomAD population after exclusion of those variants with *probable* and *possible* low penetrance.

	gnomAD								
	All	**African**	**Ashkenazi Jewish**	**East Asian**	**European (non-Finnish)**	**European (Finnish)**	**Latino**	**South Asian**	**Other**
Pathogenic allele freq	0.007169	0.003583	0.014093	0.013720	0.007390	0.002613	0.007694	0.005932	0.007438
Pathogenic genotype freq	0.000051	0.000013	0.000199	0.000188	0.000055	0.000007	0.000059	0.000035	0.000055
Heterozygous genotype freq	0.014235	0.007140	0.027790	0.027064	0.014672	0.005212	0.015269	0.011794	0.014765
Pathogenic genotype rate[a]	19457	77903	5035	5312	18309	146467	16893	28415	18076
Heterozygous carrier rate[a]	70	140	36	37	68	192	65	85	68

[a] Pathogenic genotype rate and heterozygous carrier rate are expressed as 1 in “n” of the population.

The mutational landscape of *ATP7B*

We noticed a difference in the distribution of WD-*ATP7B* missense variants across the coding region as seen in other studies [40, 41]. In particular, there appear to be very few WD-*ATP7B* missense variants located in the first one-third of the protein coding sequence, the region that encodes metal binding domains (MBDs) 1 to 4. This prompted us to measure the distribution of missense and LoF variants in the regions encoding the amino-terminal 480 amino acids (encompassing MBDs 1 to 4) and the carboxy-terminal 985 amino acids (encompassing MBDs 5 and 6, and the remainder of the ATP7B functional domains). The proportion of LoF variants in the amino-terminal portion of the *ATP7B* coding sequence (27%) was close to the expected 33% and did not significantly deviate from the expected ratio (Figure 2A; Fisher's Exact test: p=0.1254). The proportion of non-WD missense variants in the amino-terminal portion of the coding sequence (37%) was slightly higher than the expected 33% but did not quite reach statistical significance (Figure 2A; Fisher's Exact test: p=0.0507). In contrast the proportion of WD missense variants in the amino-terminal portion of the coding sequence was very low (17 out of 400 variants, 4.25%), significantly lower than the expected ratio (Figure 2A; Fisher's Exact test: p<0.0001).

We also analysed the distribution of variants across the various functional domains of the *ATP7B* coding region (Figure 2B) and found that, while the LoF and non-WD missense variants are distributed fairly uniformly across the coding region, the number of WD missense variants are lower in all amino-terminal domains up to and including MBD6. In most of the carboxy-terminal domains WD missense variants are over-represented and are particularly prevalent in the phosphorylation (P) and nucleotide (N) domains and transmembrane (TM) domains 1,2,4,5 and 6 (Figure 2B).

We also analysed the predicted pathogenicity of missense variants across the *ATP7B* coding sequence by plotting VEST3 score against coding sequence position for both WD and non-WD missense variants (Figure 2C). There was a striking difference in the pattern of VEST3 scores for both WD and non-WD missense variants across the protein coding sequence. WD missense variants in the C-terminal two-thirds of the coding sequence had higher VEST3 scores compared to those in the N-terminal one-third and were clustered into the main functional domains of the protein including the TM domains, A, P and N domains (Figure 2C). It is of note that the linker region between TM domains 3 and 4, the extended loop within the N domain and the carboxy-terminal tail of the protein are, similar to the amino-terminal one-third of the protein, relatively lacking in WD missense variants, and VEST3 scores in these regions are lower (Figure 2C).

Discussion

We have predicted the prevalence of WD in global populations using publically available NGS data. We used *ATP7B* variant data from the WDMD and updated this from a literature search done between 2010 and 2017. The 732 WD-associated variants identified were used to screen the NGS dataset.

Nearly one-third (32%) of the WD-*ATP7B* variants were present in the gnomAD dataset. It is of note that the majority of WD-*ATP7B* variants found in the gnomAD dataset were derived from the WDMD, and only 14% of the variants that contributed to our prevalence estimates were reported in the literature since the WDMD was last updated in 2010. The remaining two-thirds of WD-*ATP7B* variants not present in the gnomAD dataset were generally reported in fewer publications and hence we would predict them to be rarely encountered in the general population or limited to populations not represented in the gnomAD dataset.

Our initial estimates for population prevalence of WD included frequencies of all variants that had been reported as disease causing in the WDMD and more recent literature, with no adjustments for penetrance. We did include LoF variants that were present in the gnomAD dataset but had not been reported in WD patients. This initial estimate predicted that approximately 1 in 2400 people would have pathogenic genotypes and would be at risk of developing WD, with 1 in 25 people being carriers of pathogenic variants.

This initial prevalence estimate did not take into account variant penetrance that may lead to people carrying WD genotypes either not expressing the disease or having milder phenotypes. Further analysis of the data showed that this initial estimate was likely distorted by the presence of variants, with relatively high allele frequencies, that have been reported as disease causing in only a small number of WD patients. After removal of these ‘probable low penetrant’ variants from the analysis the predicted prevalence of WD fell dramatically to levels more consistent with traditional estimates. Review of the publications reporting the 13 variants classified as *probable* low penetrance confirm that given their frequencies in the gnomAD dataset (0.01% to 0.03% of ~240,000 chromosomes) the number of publications describing them in WD cohorts is much lower than expected [16-31]. The publications reporting these variants also include data suggesting that some have low penetrance. The c.1947-4C>T variant is reported as a polymorphism in two publications [21, 22] and appears to have been incorrectly classified as disease causing in the WDMD. The c.4021+3A>G [29] and p.Thr1434Met [30] variants were identified in WD patients who were also homozygous or compound heterozygous for other *ATP7B* variants that could account for their phenotypes. A publication reporting p.Gly869Arg suggests that it has a more benign clinical course [24], while p.Ile1230Val had an uncertain classification [18]. Publications reporting the remainder of the *probable* low penetrant variants do not give clinical details of the patients involved, so that it is difficult to assess their pathogenicity.

We also used VEP algorithms to assist in identifying further WD-associated variants that may have low penetrance. This analysis showed that the VEST3 algorithm performs very well in discriminating between WD and non-WD missense variants. After removing variants with low VEST3 scores the predicted prevalence of WD genotypes fell further but because these variants were relatively infrequent the reduction was marginal. While the removal of variants with a high gnomAD population prevalence not reflected in reports of WD patients is well justified, the removal based on VEP algorithms should be taken with some caution, since none of the algorithms are 100% accurate at discriminating between pathogenic and non-pathogenic variants.

We included LoF variants, that had not been previously reported in the literature as causing WD, in our prevalence calculations. While the majority of these are well justified for inclusion, it is possible that variants that disrupt the protein coding sequence close to the carboxy-terminus may not be pathogenic. However, the number of these variants and their frequencies were very small, and their inclusion does not greatly affect our final prevalence estimate.

Based on our analysis of WD-*ATP7B* variant frequencies and considering the above strategies to account for low penetrant variants our final prediction for the population prevalence of WD is in the range of 1 in 17,000 to 1 in 20,000 of the global population with 1 in 65 to 1 in 70 as heterozygous carriers. This gives a higher prevalence than the traditional estimate of 1 in 30,000 but is not as high as estimates from East Asia [8, 9] and the UK [10].

It is of note that the predicted prevalence was not uniform across the 8 gnomAD subpopulations. The highest prevalence was observed in the Ashkenazi Jewish and East Asian subpopulations, both being close to 1 in 5000 with 1 in 36 heterozygous carriers. In the Ashkenazi Jewish population the most prevalent mutation was p.His1069Gln. This was also the most prevalent mutation in the European population and reflects the likely origin of this mutation in the ancestors of Eastern Europeans [15]. In East Asians the most prevalent mutations were p.Thr935Met and p.Arg778Leu, both with similar allele frequencies. The lowest prevalence rate was in Africans, with around 1 in 78,000 predicted to carry WD-associated genotypes. This may represent a real low prevalence rate but may equally represent a lack of research into WD in the African continent and the consequent absence of African WD variants from our analysis.

Our population prevalence estimates are lower than two ceruloplasmin screening studies in children from Japan and Korea and a genetic study from the UK. The studies from Japan and Korea [8, 9] that predicted prevalences of 1 in 1500 and 1 in 3000 respectively, were relatively small pilot studies that identified only 1 or 2 children with WD. Hence the extrapolation of this data to the whole population may not be accurate. The genetic study from the UK [10] sequenced all *ATP7B* exons in over 1000 controls. The methodology was similar to the study presented here in that the frequencies of disease-causing variants present in the WDMD or detected by the local diagnostic genetics service were used to calculate prevalence rates. This study also used *in silico* analysis to identify further variants that may be disease causing and to exclude other variants that had questionable pathogenicity. However, their inclusion and exclusion criteria were slightly different to ours. Hence reanalysis of the Coffey et al. [10] data using our criteria would likely lead to a lower predicted prevalence of WD in the UK.

While this study was in preparation for publication, Gao et al. [42] reported a similar study where they estimated WD prevalence based on the frequency of variants in the gnomAD dataset. While their method for estimating prevalence was similar to our approach, their analysis of penetrance was different. Hence their final prevalence estimate of around 1 in 7000 is significantly higher than ours. To address the issue of low penetrant variants, they used an equation reported by Whiffin et al. [43] that calculates a maximum credible population allele frequency and filtered out all variants with allele frequencies higher than this. This method only removed 4 high frequency variants from their analysis. Our approach, which was based on a combination of high allele frequencies, the geographic distribution of WD-*ATP7B* variants and the number of reports of *ATP7B* variants in the WD literature was more stringent and removed 13 variants which were deemed to be low penetrant and at too high frequency to be contributing to the global prevalence of WD. We believe that our approach to address variant penetrance and the subsequent estimation of WD prevalence is more meaningful. For example, 3 of the top 5 variants that contributed to WD prevalence in the Gao et al. [42] study (p.Thr991Met, p.Pro1379Ser and p.Gly869Arg) are reported in very few WD publications, with some suggesting a benign clinical course in patients with these variants [19, 24-27]. Re-analysis of the Gao et al. [42] data with filtering of variants identified in this study as being probable low penetrant returns a predicted WD prevalence more similar to our estimate, at approximately 1 in 20,000. These predictions are more closely aligned with traditional estimates and suggest that reduced variant penetrance plays a much bigger role in the observed disparity in prevalence estimates between genetic and epidemiological studies [42].

Using the gnomAD and VEP data, we were also able to analyse the mutational landscape of *ATP7B* and clearly show that missense variants associated with WD cluster into the functional domains located in the carboxy-terminal two-thirds of the protein, with relative sparing of the amino-terminal MBDs. This indicates that the six MDBs are more permissive to mutations and that variants identified in these regions are less likely to be pathogenic.

This study emphasises the difficulty in assigning WD prevalence from population datasets. Accurate prevalence estimates depend upon an assessment of the penetrance of individual genetic variants, not a straightforward task. Studies to date have not clearly shown genotype/phenotype relationships, and with compound heterozygosity being the most frequent pattern and there being over 700 *ATP7B* variants this is not surprising. Other approaches will be needed to investigate the basis of the phenotype, some dependent on mutations but some on other features [3]. It is always a concern that the diagnosis of WD is not made or considered by clinicians. The higher prevalence of WD in some populations is confirmed here and should be used to emphasise their increased risk.

In conclusion, we have used NGS data to analyse the prevalence of WD in global populations, with a concerted approach to evaluating variant penetrance. This study highlights the importance of considering variant penetrance when assigning causality to genetic variants. Variants that have relatively high allele frequencies but low frequencies in patient cohorts are likely to have low penetrance. Other potential data to consider are VEP algorithm scores and the position of missense variants in the coding sequence. Large NGS datasets and improved VEP algorithms now allow us to evaluate with more accuracy the pathogenicity of genetic variants. The penetrance of *ATP7B* variants is likely to be on a spectrum: LoF variants are known to have high penetrance, whereas, some missense variants are thought to have lower penetrance [2]. WD-*ATP7B* missense variants are more likely to be in the carboxy-terminal two-thirds of the coding sequence, in regions encoding the functional domains of the protein. It would be valuable to determine the effects that low penetrant variants identified here have on ATP7B protein function and whether individuals carrying genotypes containing these variants have milder abnormalities of copper homeostasis, later onset or less severe forms of WD. Finally, this approach to predicting the prevalence of WD and penetrance of variants could be applied to other Mendelian inherited disorders.

Methods

Catalogue of Wilson disease-associated *ATP7B* variants

Initially, details of all variants classified as "disease-causing variant" (DV) in the Wilson Disease Mutation Database (WDMD), hosted at the University of Alberta (http://www.wilsondisease.med.ualberta.ca/) were downloaded. As the WDMD has not been updated since 2010 a further literature search was carried out to identify WD-associated *ATP7B* variants that have been reported between the last update of the WDMD and April 2017, using the search terms ATP7B and mutation in the PubMed database (https://www.ncbi.nlm.nih.gov/pubmed). The Human Genome Variation Society (HGVS) nomenclature for each variant was verified using the Mutalyzer Name Checker program (https://mutalyzer.nl/) [44]. Duplicate entries were removed and any mistakes in nomenclature were corrected after comparison with the original publications. All HGVS formatted variants were then converted into chromosomal coordinates (Homo sapiens – GRCh37 (hg19)) using the Mutalyzer Position Converter program. A variant call format (VCF) file containing all of the WD-associated *ATP7B* variants was then constructed using a combination of output from the Mutalyzer Position Converter and Galaxy bioinformatic tools (https://galaxyproject.org) [45].

Prevalence of Wilson disease-associated *ATP7B* variants

All variants in the *ATP7B* gene (Ensembl transcript ID ENST00000242839) were downloaded from the gnomAD (http://gnomad.broadinstitute.org/) browser [13]. The WD-associated *ATP7B* variants (see above) were compared with the gnomAD *ATP7B* variants and allele frequency data were extracted for those variants with VCF descriptions that matched exactly. Allele frequency data were also extracted from the gnomAD dataset for variants that had not been previously reported in WD patients but were predicted to cause loss of function (LoF) of the ATP7B protein. These LoF variants included frameshift, splice acceptor, splice donor, start lost and stop gained mutations.

Pathogenic *ATP7B* allele frequencies were determined in the gnomAD dataset by summing all of the allele frequencies for variants classified as WD-associated. Predicted pathogenic *ATP7B* genotype frequencies, heterozygote frequencies and carrier rates were calculated from allele frequencies using the Hardy-Weinberg equation.

In silico analyses of variant pathogenicity

The functional consequence of WD-*ATP7B* missense variants and gnomAD-derived *ATP7B* missense variants (that had not been previously associated with WD) was assessed using the wANNOVAR program (http://wannovar.wglab.org/), which provides scores for 16 VEP algorithms. The performance of these 16 algorithms for predicting WD-associated variants was analysed using receiver operating characteristic (ROC) curve analysis. The best performing algorithm (VEST3)[39] was used, together with the gnomAD frequency data, data from the WDMD and other published data [15] to predict the pathogenicity of WD-associated *ATP7B* variants and refine the pathogenic genotype prevalence estimates.

Acknowledgements

We would like to acknowledge the curators of the Wilson Disease Mutation Database, University of Alberta for the use of data in this study.

References

1. Ala A, Walker AP, Ashkan K, Dooley JS, Schilsky ML. Wilson's disease. Lancet. 2007;369(9559):397-408.

2. Chang IJ, Hahn SH. The genetics of Wilson disease. Handb Clin Neurol. 2017;142:19-34.

3. Ferenci P, Stremmel W, Czlonkowska A, Szalay F, Viveiros A, Stattermayer AF, et al. Age,sex, but not ATP7B genotype effectively influences the clinical phenotype of Wilson disease. Hepatology. 2018.

4. Medici V, Weiss KH. Genetic and environmental modifiers of Wilson disease. Handb Clin Neurol. 2017;142:35-41.

5. Scheinberg IH, Sternlieb I. Wilson's disease. Philadelphia: Saunders; 1984. xii, 171 p., 2 leaves of plates p.

6. Bachmann H, Lossner J, Gruss B, Ruchholtz U. [The epidemiology of Wilson's disease in the German Democratic Republic and current problems from the viewpoint of population genetics]. Psychiatr Neurol Med Psychol (Leipz). 1979;31(7):393-400.

7. Saito T. An assessment of efficiency in potential screening for Wilson's disease. J Epidemiol Community Health. 1981;35(4):274-80.

8. Ohura T, Abukawa D, Shiraishi H, Yamaguchi A, Arashima S, Hiyamuta S, et al. Pilot study of screening for Wilson disease using dried blood spots obtained from children seen at outpatient clinics. J Inherit Metab Dis. 1999;22(1):74-80.

9. Hahn SH, Lee SY, Jang YJ, Kim SN, Shin HC, Park SY, et al. Pilot study of mass screening for Wilson's disease in Korea. Mol Genet Metab. 2002;76(2):133-6.

10. Coffey AJ, Durkie M, Hague S, McLay K, Emmerson J, Lo C, et al. A genetic study of Wilson's disease in the United Kingdom. Brain. 2013;136(Pt 5):1476-87.

11. Wallace DF, Subramaniam VN. The global prevalence of HFE and non-HFE hemochromatosis estimated from analysis of next-generation sequencing data. Genet Med. 2016;18(6):618-26.

12. Hughes BG, Harrison PM, Hekimi S. Estimating the occurrence of primary ubiquinone deficiency by analysis of large-scale sequencing data. Sci Rep. 2017;7(1):17744.

13. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.

14. Ruderfer DM, Hamamsy T, Lek M, Karczewski KJ, Kavanagh D, Samocha KE, et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat Genet. 2016;48(10):1107-11.

15. Gomes A, Dedoussis GV. Geographic distribution of ATP7B mutations in Wilson disease. Ann Hum Biol. 2016;43(1):1-8.

16. Mukherjee S, Dutta S, Majumdar S, Biswas T, Jaiswal P, Sengupta M, et al. Genetic defects in Indian Wilson disease patients and genotype-phenotype correlation. Parkinsonism Relat Disord. 2014;20(1):75-81.

17. Kroll CA, Ferber MJ, Dawson BD, Jacobson RM, Mensink KA, Lorey F, et al. Retrospective determination of ceruloplasmin in newborn screening blood spots of patients with Wilson disease. Mol Genet Metab. 2006;89(1-2):134-8.

18. Davies LP, Macintyre G, Cox DW. New mutations in the Wilson disease gene, ATP7B: implications for molecular testing. Genet Test. 2008;12(1):139-45.

19. Cox DW, Prat L, Walshe JM, Heathcote J, Gaffney D. Twenty-four novel mutations in Wilson disease patients of predominantly European ancestry. Hum Mutat. 2005;26(3):280.

20. Vrabelova S, Letocha O, Borsky M, Kozak L. Mutation analysis of the ATP7B gene and genotype/phenotype correlation in 227 patients with Wilson disease. Mol Genet Metab. 2005;86(1-2):277-85.

21. Kim EK, Yoo OJ, Song KY, Yoo HW, Choi SY, Cho SW, et al. Identification of three novel mutations and a high frequency of the Arg778Leu mutation in Korean patients with Wilson disease. Hum Mutat. 1998;11(4):275-8.

22. Okada T, Shiono Y, Hayashi H, Satoh H, Sawada T, Suzuki A, et al. Mutational analysis of ATP7B and genotype-phenotype correlation in Japanese with Wilson's disease. Hum Mutat. 2000;15(5):454-62.

23. Loudianos G, Dessi V, Lovicu M, Angius A, Nurchi A, Sturniolo GC, et al. Further delineation of the molecular pathology of Wilson disease in the Mediterranean population. Hum Mutat. 1998;12(2):89-94.

24. Garcia-Villarreal L, Daniels S, Shaw SH, Cotton D, Galvin M, Geskes J, et al. High prevalence of the very rare Wilson disease gene mutation Leu708Pro in the Island of Gran Canaria (Canary Islands, Spain): a genetic and clinical study. Hepatology. 2000;32(6):1329-36.

25. Lepori MB, Lovicu M, Dessi V, Zappu A, Incollu S, Zancan L, et al. Twenty-four novel mutations in Wilson disease patients of predominantly Italian origin. Genet Test. 2007;11(3):328-32.

26. Margarit E, Bach V, Gomez D, Bruguera M, Jara P, Queralt R, et al. Mutation analysis of Wilson disease in the Spanish population -- identification of a prevalent substitution and eight novel mutations in the ATP7B gene. Clin Genet. 2005;68(1):61-8.

27. Shah AB, Chernov I, Zhang HT, Ross BM, Das K, Lutsenko S, et al. Identification and analysis of mutations in the Wilson disease gene (ATP7B): population frequencies, genotype-phenotype correlation, and functional analyses. Am J Hum Genet. 1997;61(2):317-28.

28. Aggarwal A, Chandhok G, Todorov T, Parekh S, Tilve S, Zibert A, et al. Wilson disease mutation pattern with genotype-phenotype correlations from Western India: confirmation of p.C271* as a common Indian mutation and identification of 14 novel mutations. Ann Hum Genet. 2013;77(4):299-307.

29. Santhosh S, Shaji RV, Eapen CE, Jayanthi V, Malathi S, Chandy M, et al. ATP7B mutations in families in a predominantly Southern Indian cohort of Wilson's disease patients. Indian J Gastroenterol. 2006;25(6):277-82.

30. Abdelghaffar TY, Elsayed SM, Elsobky E, Bochow B, Buttner J, Schmidt H. Mutational analysis of ATP7B gene in Egyptian children with Wilson disease: 12 novel mutations. J Hum Genet. 2008;53(8):681-7.

31. Loudianos G, Dessi V, Lovicu M, Angius A, Altuntas B, Giacchino R, et al. Mutation analysis in patients of Mediterranean descent with Wilson disease: identification of 19 novel mutations. J Med Genet. 1999;36(11):833-6.

32. Deguti MM, Genschel J, Cancado EL, Barbosa ER, Bochow B, Mucenic M, et al. Wilson disease: novel mutations in the ATP7B gene and clinical correlation in Brazilian patients. Hum Mutat. 2004;23(4):398.

33. Hua R, Hua F, Jiao Y, Pan Y, Yang X, Peng S, et al. Mutational analysis of ATP7B in Chinese Wilson disease patients. Am J Transl Res. 2016;8(6):2851-61.

34. Yuan ZF, Wu W, Yu YL, Shen J, Mao SS, Gao F, et al. Novel mutations of the ATP7B gene in Han Chinese families with pre-symptomatic Wilson's disease. World J Pediatr. 2015;11(3):255-60.

35. Bost M, Lachaux A, Accominotti M, Vandenberghe A. Mutation screening and genotype-phenotype correlation in 32 families with Wilson disease. J Trace Elem Exp Med. 1999;12(4):321-9.

36. Ohya K, Abo W, Tamaki H, Sugawara C, Endo T, Nomachi S, et al. Presymptomatic diagnosis of Wilson disease associated with a novel mutation of the ATP7B gene. Eur J Pediatr. 2002;161(2):124-6.

37. Owada M, Suzuki K, Fukushi M, Yamauchi K, Kitagawa T. Mass screening for Wilson's disease by measuring urinary holoceruloplasmin. J Pediatr. 2002;140(5):614-6.

38. Bost M, Piguet-Lacroix G, Parant F, Wilson CM. Molecular analysis of Wilson patients: direct sequencing and MLPA analysis in the ATP7B gene and Atox1 and COMMD1 gene analysis. J Trace Elem Med Biol. 2012;26(2-3):97-101.

39. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3.

40. Squitti R, Siotto M, Bucossi S, Polimanti R. In silico investigation of the ATP7B gene: insights from functional prediction of non-synonymous substitution to protein structure. Biometals. 2014;27(1):53-64.

41. Yu CH, Lee W, Nokhrin S, Dmitriev OY. The Structure of Metal Binding Domain 1 of the Copper Transporter ATP7B Reveals Mechanism of a Singular Wilson Disease Mutation. Sci Rep. 2018;8(1):581.

42. Gao J, Brackley S, Mann JP. The global prevalence of Wilson disease from next-generation sequencing data. Genet Med. 2018.

43. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med. 2017;19(10):1151-8.

44. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat. 2008;29(1):6-13.

45. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44(W1):W3-W10.

46. Gourdon P, Sitsel O, Lykkegaard Karlsen J, Birk Moller L, Nissen P. Structural models of the human copper P-type ATPases ATP7A and ATP7B. Biol Chem. 2012;393(4):205-16.

Figure Legends

Figure 1. Identification of probable and possible low penetrant *ATP7B* variants.

(A) The number of WDMD references was plotted against gnomAD allele frequency for WD-*ATP7B* variants identified in the gnomAD dataset. (B) VEST3 score was plotted against gnomAD allele frequency for WD-*ATP7B* variants identified in the gnomAD dataset. Variants reported in the Gomes et al [15] review as being the most common WD-*ATP7B* variants in various geographic regions are denoted by red dots and those not reported in the Gomes et al review by blue dots. In (A) 13 variants were classified as *probable* low penetrant based on relatively high allele frequencies, low numbers of WDMD references and not being reported in the Gomes et al. review (boxed). In (B) 11 variants were classified as *possible* low penetrant based on a VEST3 score of <0.5 (boxed).

Figure 2. The mutational landscape of *ATP7B*. (A) The number of WD missense, non-WD missense and LoF *ATP7B* variants located in the amino (N)-terminal one-third of the coding sequence (white bars; encompassing amino acids 1 to 480) was compared to the number of variants in the carboxy (C)-terminal two-thirds of the coding sequence (gray bars; encompassing amino acids 481 to 1465). The difference compared to the expected

number of variants, based on an even distribution across the coding sequence, was assessed using Fisher's Exact test (****, p<0.0001; ns, not significant). (B) The number of WD-missense, non-WD missense and LoF *ATP7B* variants located in the functional domains and linker regions of the *ATP7B* coding sequence (cyan boxes, metal binding domains; orange boxes, transmembrane domains; yellow box, actuator (A) domain; blue boxes, phosphorylation (P) domain; red boxes, nucleotide (N) domain; gray boxes, linker regions, N- and C-termini) were compared against the number of variants expected, based on an even distribution across the coding sequence, and expressed as percentage of variants observed minus percentage of variants expected. WD missense variants (red line), non-WD missense variants (gray line), LoF variants (blue line). (C) Coding sequence position was plotted against VEST3 score for WD missense variants (red dots) and non-WD missense variants (black dots). Positions of LoF variants are shown above the plot as blue triangles. Box and whisker plots show the median, quartiles and range of VEST3 scores for non-WD and WD missense variants. The ATP7B domains in panels B and C are as described in Gourdon et al. [46]. The exon structure of the *ATP7B* coding sequence is shown below the plots.

Supporting Information

Table S1. Disease causing variants identified in the Wilson Disease Mutation Database

Table S2. Disease causing variants identified by a literature search between 2010 and 2017.

Table S3. *ATP7B* loss of function variants and CNV deletions identified in gnomAD and ExAC databases.

Figure S1. Comparison of non-WD missense and WD missense *ATP7B* variants using 16 VEP algorithm scores.

Figure S2. Receiver operating characteristic (ROC) curve analysis was used to assess the effectiveness of 16 VEP algorithms to discriminate between WD missense and non-WD missense *ATP7B* variants.

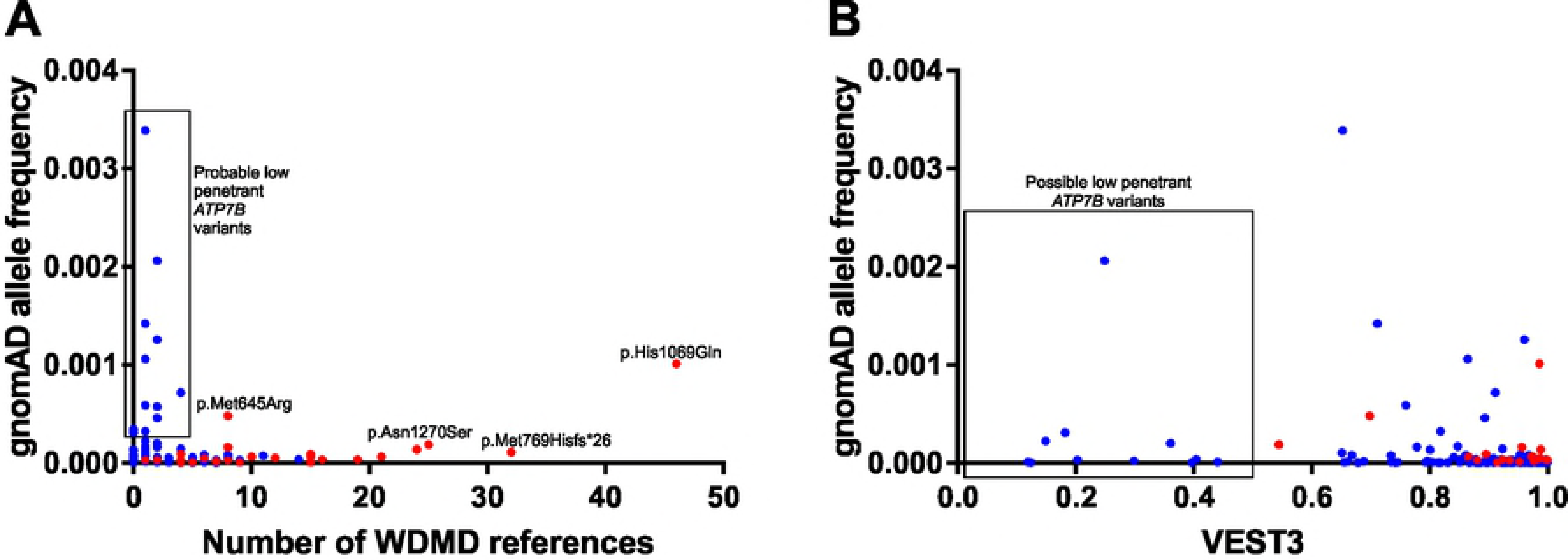

Figure 1

Figure 1

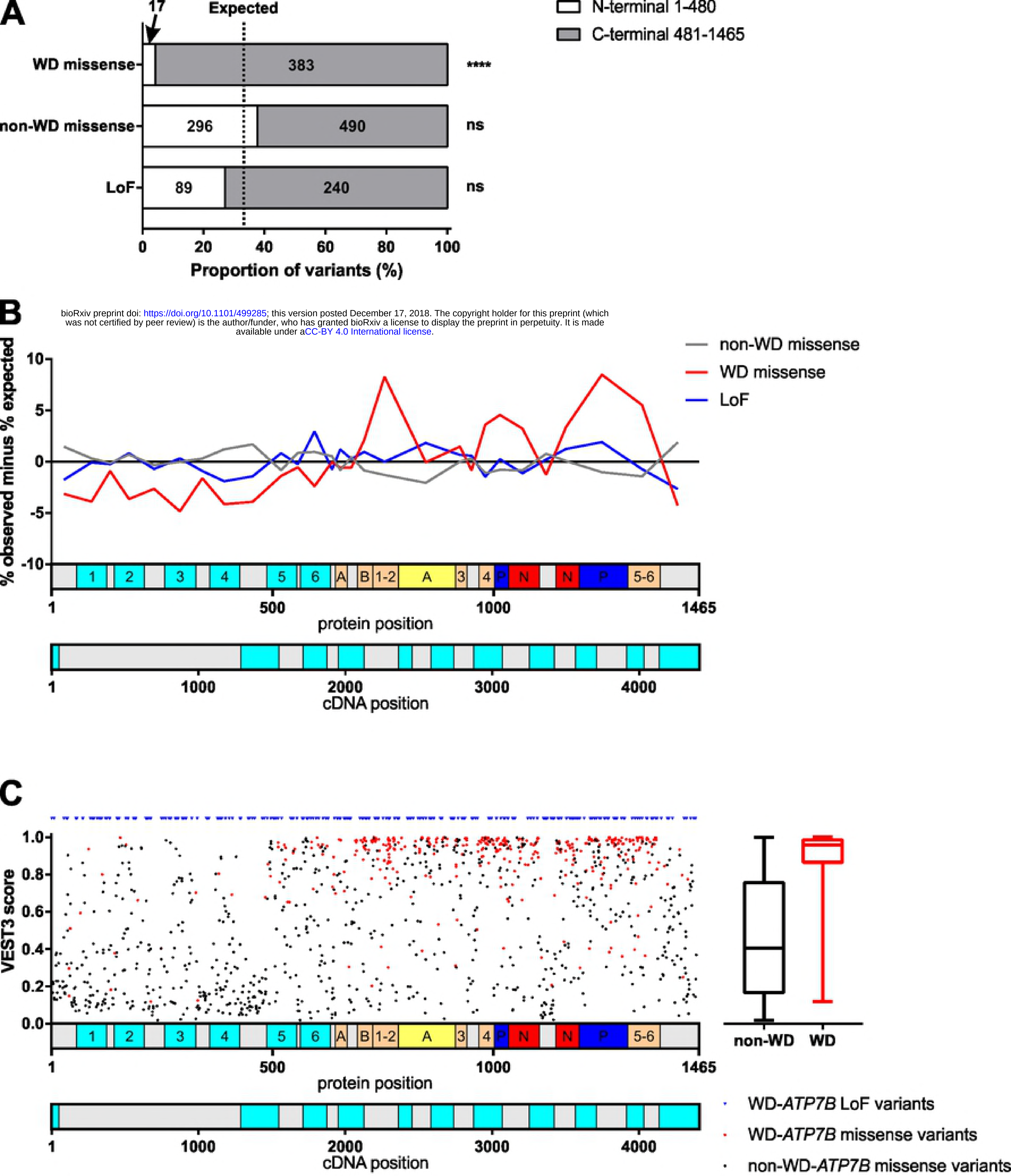

Figure 2

Figure 2