The global prevalence of Wilson's Disease from next generation sequencing data

Short running title: The global prevalence of Wilson's Disease

Jiali Gao, BA (Hons)[a]*; Simon Brackley, BA (Hons)[a]*; Jake P. Mann, MBChB MRCP MRCPCH[b,c]

[a]University of Cambridge, School of Clinical Medicine, Cambridge, United Kingdom

[b]University of Cambridge, Department of Paediatrics, Cambridge, United Kingdom

[c]University of Cambridge, Institute of Metabolic Science-Metabolic Research Laboratories, Cambridge, United Kingdom

*these authors contributed equally to this work

Corresponding author:

Dr. Jake P. Mann
Department of paediatrics, Box 116, Addenbrooke's Hospital, Cambridge, CB2 0QQ
Telephone: +44 1223 763480
jm2032@cam.ac.uk

ABSTRACT

Purpose: Wilson Disease (WD) is an autosomal recessive disorder of copper metabolism, caused by mutations in *ATP7B*. We aimed to: 1) perform a meta-analysis of previous WD prevalence estimates, 2) estimate the prevalence of WD from population sequencing data, and 3) generate an *ATP7B* gene variant database.

Methods: MEDLINE and EMBASE were systematically searched. Previous prevalence estimates were subjected to meta-analysis. All previously reported pathogenic *ATP7B* variants were compiled and annotated with GnomAD allele frequencies. Pooled global and ethnicity-specific genetic prevalences for WD were generated using the Hardy-Weinberg equation.

Results: Meta-analysis of genetic studies of WD prevalence gave an estimate 12.7 per 100,000 (95% CI: 6.3-23.0). We developed a referenced, searchable *ATP7B* database comprising 11,520 variants including 782 previously reported disease variants, which can be found at http://www.wilsondisease.tk/. 216/782 of these were present in GnomAD, remained after filtering by allele frequency and met American College of Medical Genetics criteria. Based on these, the genetic prevalence of WD was 13.9 per 100,000 (95% CI: 12.9-14.9), or 1 per 7,194. Combining this with 60 predicted pathogenic variants gave a birth prevalence of 15.4 per 100,000 (95% CI: 14.4-16.5).

Conclusion: The genetic prevalence of Wilson disease may be greater than previous estimates.

Key words: Wilson Disease, ATP7B, Prevalence, Mutations, Database

INTRODUCTION

Wilson Disease (WD) is an autosomal recessive disorder of copper metabolism, in which there is defective transport of copper across the endoplasmic reticulum and biliary copper excretion. This manifests as hepatic, neurological and psychiatric symptoms. Diagnosis is based on a combination of clinical features, serum caeruloplasmin, urinary copper and hepatic copper[1], with early detection and effective treatment resulting in a normal life-span with minimal morbidity.

WD is known to be caused by mutations in the *ATP7B* gene. Although the possibility of a second WD gene has been discussed, no other genes have been identified[2]. Indeed, study sequencing 181 patients in the UK found an overall *ATP7B* mutation detection frequency of 98%[3], supporting the belief that WD is a classic monogenic disorder. However, the mutational spectrum is wide, with the WD Mutation Database by the University of Alberta (last updated in 2010) listing around 500 pathogenic variants [4] though the actual number may be greater. *ATP7B* mutation testing may be required for the diagnosis of WD but it is not yet routine[1].

The prevalence of WD is often quoted as 1 in 30,000, taken from a monograph written by Scheinberg and Sternlieb in 1984, before the discovery of the gene responsible[5]. This estimate was based on three studies that have been described as methodologically flawed[6] and it has been recognised some regions have a much higher prevalence. In accordance with this, recent estimates from sequencing studies have been much higher, with several papers raising the possibility of under-diagnosis[3,7,8]. Moreover, reports of WD in consecutive generations of families, producing a 'pseudodominant' inheritance pattern, suggest that heterozygous ATP7B mutation carriers may be more common than previously thought[3].

Therefore, we aimed to, firstly, perform a meta-analysis of previous prevalence estimates for WD and secondly, estimate the prevalence of WD across ethnicities from publicly available population sequencing data, using a validated methodology[9]. In addition, we aimed to compile a new database

of reported WD pathogenic variants by systematically searching the literature and cross-referencing with pre-existing databases.

MATERIALS AND METHODS

The protocol for this study was registered with PROSPERO (https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=74489)

Meta-analysis of previous prevalence estimates

Both MEDLINE and EMBASE were systematically searched for papers related to Wilson disease prevalence or epidemiology on 15/08/17 (see supplementary data for search terms). Two independent reviewers (JG & SB) screened abstracts to determine suitability for inclusion. Any disagreement was resolved through discussion with a third reviewer (JM). Foreign language reports; reviews, commentaries and editorials reporting non-original data; and in vitro or non-human studies were excluded; papers quoting an original estimate of WD prevalence or incidence were included.

Included papers were quality reviewed and assessed for risk of bias using the Appraisal tool for Cross-Sectional Studies (AXIS)[10] (Figure S3) and prevalence data was extracted. Where appropriate, prevalence was estimated from the number of cases reported in the manuscript and relevant population size data obtained from official statistics, as specified in the results. In addition, where multiple prevalence estimates over time were given, the most recent figure was taken.

Studies were subclassified into epidemiological studies, defined as those based on case frequencies within a population, and genetic studies, involving sequencing for WD pathogenic variants and calculation of mutant allele frequencies.

The degree of study heterogeneity was investigated by means of the Cochran Q test and its related metric I^2. Since in all cases there was significant heterogeneity, estimates of prevalence from

epidemiological studies and mutant allele frequencies from genetic studies were subjected to meta-analysis separately using random effects models[11]. Population size was taken to be the overall population of the catchment area for epidemiological studies and sample size for genetic studies. Where population size was unclear in the paper, population data for the appropriate year was obtained from WorldBank (data.worldbank.org, accessed 14/03/18), unless otherwise specified. From the overall mutant allele frequency an estimate of the prevalence of WD at birth was calculated using the Hardy Weinberg equation.

All statistical analysis was performed using the Meta-XL add-in for Microsoft Excel (www.epigear.com). Forest plots were generated using DistillerSR Forest Plot Generator from Evidence Partners (www.evidencepartners.com)

Identification of reported pathogenic variants and functional variants

A systematic search was performed to produce a list of all known disease-causing pathogenic variants in the *ATP7B* gene. Both MEDLINE and EMBASE were searched for papers relating to WD mutations or genetics on 15/08/17 (see supplementary data for search terms). As above, two independent reviewers screened abstracts to determine suitability for inclusion and disagreements were resolved through discussion with a third reviewer. Foreign language reports; reviews, commentaries and editorials reporting non-original data; and in vitro or non-human studies were excluded; papers reporting variants within the *ATP7B* gene were included. Variants reported as disease-causing were extracted.

The University of Alberta WD database [12] (www.wilsondisease.med.ualberta.ca/references.asp, accessed 02/03/18), ClinVar (www.ncbi.nlm.nih.gov/clinvar/, accessed 13/11/17) were screened for additional *ATP7B* variants with published reports of pathogenicity and these were added to our list. Variants derived from personal communications and unpublished data were excluded. Compound

mutations, defined as more than 1 non-overlapping mutation within the same *ATP7B* sequence, were also excluded.

A list of all known variants in *ATP7B*, including polymorphisms, was compiled. All above variants were combined with variants from the University of Alberta WD database and ClinVar without evidence of pathogenicity and additional variants from Ensembl (www.ensembl.org, accessed 26/01/18) and GnomAD (gnomad.broadinstitute.org/, accessed 26/01/18).

Annotation of variants with allele frequencies and functional predictions

Coding sequence nucleotide changes for each variant were identified and converted to Human Genome Variation Society (HGVS) format (hg38) using Mutalyzer (https://mutalyzer.nl/). The Ensembl Variant Effect Predictor (https://ensembl.org/Tools/VEP) was used to annotate variants with mutation consequences and SIFT/PolyPhen *in silico* predictions of pathogenicity. GnomAD allele frequency data was downloaded directly and added to each variant.

Frequency filtering

All variants with allele frequency data available from the GnomAD dataset were filtered using a method proposed by Whiffin *et al*[13]. A 'maximum credible population allele frequency' was calculated based on the equation:

$$Max\ credible\ population\ AF = \sqrt{prevalence} \times max\ allelic\ contribution \times \sqrt{max\ genetic\ contribution} \times \frac{1}{\sqrt{penetrance}}$$

Reliable estimates for these parameters were difficult to obtain, so an upper bound for the maximum credible AF was calculated. Prevalence was taken from the meta-analysis estimate from genetic studies, 1 in 7874, as these studies are most comparable to our current method. Maximum

allelic contribution, the maximum proportion of variation within a gene attributable to a single allele, was set at 30%, based on a variety of estimates of p.His1069Gln *ATP7B* variant prevalence in WD patients[3,8,14,15]. Maximum genetic contribution, the maximum proportion of disease attributable to variation within a gene, was set at 98% based on Coffey *et al*[3]. Lastly, the penetrance used for this calculation was selected as 50%, as suggested by the original authors' methods. This gave a 'maximum credible AF' of 0.473%.

'Filtering allele frequencies' were also computed for each variant based on GnomAD allele counts, using the R code provided by Whiffin *et al.* Variants with 'filtering AFs' greater than the 'maximum credible AF' were excluded from further analysis.

Disease variant classification

Variants reported as pathogenic with allelic frequency data were further screened and classified using American College of Medical Genetics (ACMG) criteria[16]. Variants reported in a peer-reviewed journal were labelled with PS4, PP4 and PP5 level evidence and classified as 'likely pathogenic' if they:

1) Were associated with at least 2 of the following criteria[5]:

    - low ceruloplasmin level <20 mg/dl
    - the presence of Kayser-Fleischer rings by slit-lamp examination
    - hepatic copper content of 250 mcg/g dry weight liver tissue
    - 24hr urinary copper >100mcg

    in the presence of hepatic or neurological manifestations consistent with WD

2) Had a significantly increased prevalence in affected individuals compared with controls

Prevalence estimation

Allele frequencies of relevant variants were extracted from our *ATP7B* variant database, pooled, and estimates of the prevalence of WD at birth were generated using the Hardy-Weinberg equation.

95% Confidence intervals for these estimates were calculated as Wilson interval scores by the Wilson Score method. Graphs were generated using Microsoft Excel (2016) for Windows.

RESULTS

Previous estimates of prevalence

1003 abstracts were identified as potentially eligible, of which 20 studies were included. 12/20 (60%) of these employed an epidemiological method and 8/20 (40%) used genetic sequencing (Figure S1).

From these, 22 estimates of prevalence were extracted, from various locations worldwide and reporting population prevalences ranging from 0.25-6667/100,000 (Table 1 and Table 2). Three epidemiological studies were excluded from meta-analysis: Dedoussis et al (2005) and Garcia-Villarreal et al (2000) describe outlier populations; and Lai et al (2010) due to geographical and temporal overlap with Tai et al (2017). Genetic studies with geographical overlap were not excluded since each study only sampled a small proportion of the target population, so overlap in study cohort was deemed unlikely.

Meta-analysis of epidemiological estimates of prevalence using the inverse variance method with a double arcsin transformation gave a pooled prevalence of 1.38 (95% CI: 0.85-2.05) per 100,000 (Figure 1A). Meanwhile, meta-analysis of mutant allele frequencies estimates from genetic sequencing studies gave a pooled allele frequency of 0.011 (95% CI: 0.008-0.015) (Figure 1B). Using the Hardy-Weinberg equation, this mutant allele frequency is equivalent to a prevalence at birth of 12.7 per 100,000 (95% CI: 6.25-23.0). The studies in both meta-analyses showed statistically significant heterogeneity (epidemiological: I^2= 98.0%, p<0.001; genetic: I^2= 76.9%, p<0.001).

We then proceeded to attempt to obtain a more reliable estimate of global WD prevalence, using publicly available sequencing data.

Identification of *ATP7B* Variants

Our systematic search for WD mutations returned 1558 abstracts, of which 245 papers were included, from which 771 pathogenic variants were extracted (Figure S2). Screening of the University of Alberta WD database and of ClinVar for variants reported as disease-causing in the published literature found an additional 10 variants and 1 variant, respectively, giving a total of 782 reported disease variants (Figure 2A). GnomAD allele frequency data was available for 238/782 (30.4%) of these disease-causing variants.

A list of all known variants in *ATP7B*, including non-pathogenic variants and polymorphisms, was also compiled. The above reported disease variants were combined with 10,199 variants from Ensembl, 2,301 from GnomAD, 372 from ClinVar and 654 from the University of Alberta WD database, giving a total of 11,520 variants (Figure 2A). The full database can be found at http://www.wilsondisease.tk/ and in Table S2. GnomAD allele frequency data was available for 2301/11520 (19.9%) of these variants.

Analysis of allele frequencies

Known pathogenic variants

Out of 238 reported disease-causing variants with allele frequency data, 234 remained after frequency filtering and 216 of these were classified as 'likely pathogenic' under ACMG criteria (Figure 2A). Pooling of the allele frequencies of these variants gave a global mutant allele frequency of 0.0118, which is equivalent to a prevalence at birth of 13.9 per 100,000 (95% CI: 12.9-14.9), or 1 per 7,194 (Figure 2B and Table S1). The East Asian ethnicity had the highest estimated prevalence of

29.5 per 100,000 (95% CI: 23.6-36.8) compared to only 2.08 per 100,000 (95% CI: 1.43-3.03) in the Finnish population. There was insufficient variant-level penetrance data to adjust our estimate based on the cumulative penetrance of disease variants.

The most common mutation consequence was missense, accounting for 52% of all 'likely pathogenic' variants (including those without allele frequency data) and contributing 85% of the total mutant allele frequency (Figure 2C). After that, frameshift and stop gained/lost were the next most common.

50% of the total allele frequency was accounted for by the 9 most frequent variants, the top 5 of which are listed in Table 3. p.His1069Gln, the most frequently previously reported variant in the European population, ranks 3rd amongst these.

Predicted pathogenic variants

In order to estimate the true genetic prevalence of WD, including pathogenic variants that have not yet been identified in patients, we examined all variants reported in humans that caused major functional or structural changes (frameshift, premature stop codon, splice donor and splice acceptor variants). After frequency filtering, 113/11520 (0.98%) variants met these criteria and were found in the GnomAD dataset. Of these, 60 had not been previously identified in WD patients. When combined with the reported pathogenic variants above, the global mutant allele frequency of WD was 0.0124, equating to a birth prevalence of 15.4 per 100,000 (95% CI: 14.4-16.5), or 1 per 6,494 (Figure 2B and Table S1). The east Asian population had the highest prevalence of 29.7 per 100,000 (95% CI: 23.8-37.0) and the Finnish population had the lowest prevalence at 2.37 per 100,000 (95% CI 1.7-3.4).

DISCUSSION

In this study, we have produced an unbiased description of the genetic prevalence of Wilson disease, both globally and across 7 major ethnicities, and this was found to be higher than previous estimates. In addition, we have collated a publicly available, up-to-date database of *ATP7B* gene variants with robust classifications of pathogenicity.

Previous estimates of WD prevalence are extremely heterogeneous, which may be accounted for by differences in population, diagnosis and methodology. In particular, it should be remembered that the prevalence of WD in isolated populations, such as the Canary Islands and Crete, may be over 35 per 100,000. In contrast, using the large and diverse sample represented by the GnomAD dataset (comprising 123,136 exome sequences and 15,496 whole-genome sequences), we deduce that the global genetic prevalence of WD at birth is approximately 13.9 to 15.4 per 100,000. We have highlighted that patients of East Asian origin are at the greatest risk of WD, whereas those of Finnish origin have the lowest genetic prevalence.

In particular, our estimate of 13.9 per 100,000, derived from known pathogenic variants, is very similar to the prevalence from the meta-analysis of genetic studies (12.7 per 100,000), but both of these estimates are significantly higher than the prevalence from the meta-analysis of epidemiological studies (1.38 per 100,000). There are several possible reasons for this disparity.

Firstly, the genetic prevalence calculated in this study does not account for the incomplete penetrance of variants, as penetrance data are lacking. If the difference between epidemiological (1.4 per 100,000) and genetic (13.9 per 100,000) estimates were due to incomplete penetrance alone then the overall disease penetrance would be 10%. However, the genetic studies contributing to the meta-analysis were mostly based on a maximum of 6 common pathogenic variants, with the exception of Coffey *et al.* (2013)[3] and Gialluisi *et al*. (2012)[17]. Since these are mostly well-reported variants, which have repeatedly been screened for in control populations, it is unlikely that any of

these could have a penetrance low enough to account for the extent of disparity between genetic and epidemiological estimates.

Methodological differences may also contribute. WD commonly presents during the second and third decades of life, whereas the genotype is present from birth. Thus, the clinical phenotype quantified by epidemiological studies only exists for about 80% of an average 70 year lifespan.

Finally, this disparity may be evidence of under-diagnosis of WD on a population level, as has been repeatedly previously suggested[3,7], or of delayed diagnosis and consequent early deaths. With good compliance, the treatment for WD is highly effective and diagnostic failure has been reported to be the principal cause of death in WD patients[18]. Under-diagnosis may be due to milder disease phenotypes, or single-organ system phenotypes, or application of diagnostic criteria recently recognized as inappropriately narrow[1].

Thus, we have demonstrated the power and potential limitations of using rapidly expanding genomic databases such as GnomAD to estimate the prevalence of recessive diseases. These conclusions are, however, limited by the assumptions of the Hardy-Weinberg equation and many of the consequence-predicted variants are not yet supported by *in vitro* data. Moreover, pathogenic classification of our previously reported variants is reliant upon accurate diagnosis by the papers we screened.

We also implemented the method for filtering variants by frequency proposed by Whiffin *et al* (2017). As the authors note, a limitation of this procedure is the difficulty in obtaining reliable estimates of penetrance and maximum allelic contribution. Therefore, an upper bound for the maximum credible allele frequency was used here to avoid filtering out potentially pathogenic variants. Although not optimal, such a method is still more stringent than the standard practice of discarding variants more frequent than the arbitrary MAF of 5%. It should also be noted that the use of frequency filtering was only intended to remove variants too common to realistically be pathogenic, rather than to define pathogenic variants. Indeed, only 4 variants were filtered out here

and were we to assume 100% penetrance in this formula, only one additional variant would be filtered, giving a prevalence estimate of 11.6 per 100,000.

Our finding of 782 reported disease variants and 60 predicted pathogenic variants associated with WD is an update on previous lists of variants and highlights the mutational spectrum of the disease. The most common pathogenic variants are missense mutations (accounting for 85% of the prevalence), but large deletions and insertions have also been reported. This should be considered in efforts to develop an effective screening programme or in targeted genetic testing for the disease.

In summary, the genetic prevalence of WD is much higher than epidemiological estimates, potentially indicating underdiagnosis or the existence of less severe phenotypes. We have also produced more reliable global and ethnicity-specific estimates for WD genetic prevalence and in the process, a new up-to-date database of WD variants. These results provide important baseline data for clinical use, genetic counselling, and informing future research in *ATP7B*.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Patrick McKeirnan (Pittsburgh, USA) for his advice during this project. We are also grateful to Dr. Diane Cox, Dr. Georgina MacIntyre, and the whole team from the University of Alberta Wilson Disease Mutation Database for their support in this piece of work.

REFERENCES

1. European Association for Study of Liver. EASL Clinical Practice Guidelines: Wilson's disease. *J Hepatol*. 2012;56(3):671-685. doi:10.1016/j.jhep.2011.11.007

2. Lovicu M, Dessì V, Lepori MB, et al. The canine copper toxicosis gene MURR1 is not implicated in the pathogenesis of Wilson disease. *J Gastroenterol*. 2006;41(6):582-587. doi:10.1007/s00535-006-1807-0

3. Coffey AJ, Durkie M, Hague S, et al. A genetic study of Wilson's disease in the United Kingdom. *Brain J Neurol*. 2013;136(Pt 5):1476-1487. doi:10.1093/brain/awt035

4. Bugbee D, Davies L, Kenney S, Cox DW. Wilson Disease Mutation Database. http://www.wilsondisease.med.ualberta.ca/database.asp. Accessed March 15, 2018.

5. Scheinberg I, Sternlieb I. *Wilson's Disease.* Philadelphia: WB Saunders; 1984.

6. Park RH, McCabe P, Fell GS, Russell RI. Wilson's disease in Scotland. *Gut*. 1991;32(12):1541-1545.

7. Jang J-H, Lee T, Bang S, Kim Y-E, Cho E-H. Carrier frequency of Wilson's disease in the Korean population: a DNA-based approach. *J Hum Genet*. May 2017. doi:10.1038/jhg.2017.49

8. Olivarez L, Caggana M, Pass KA, Ferguson P, Brewer GJ. Estimate of the frequency of Wilson's disease in the US Caucasian population: a mutation analysis approach. *Ann Hum Genet*. 2001;65(Pt 5):459-463. doi:doi:10.1017/S0003480001008764

9. 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):618-626. doi:10.1038/gim.2015.140

10. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open*. 2016;6(12):e011458. doi:10.1136/bmjopen-2016-011458

11. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ Publishing Group; :285-312.

12. Kenney SM, Cox DW. Sequence variation database for the Wilson disease copper transporter, ATP7B. *Hum Mutat*. 2007;28(12):1171-1177. doi:10.1002/humu.20586

13. Whiffin N, Minikel E, Walsh R, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med*. 2017;19(10):1151-1158. doi:10.1038/gim.2017.26

14. Ferenci P. Regional distribution of mutations of the ATP7B gene in patients with Wilson disease: impact on genetic testing. *Hum Genet*. 2006;120(2):151-159. doi:10.1007/s00439-006-0202-5

15. Krumina A, Keiss J, Sondore V, et al. From clinical and biochemical to molecular genetic diagnosis of Wilson disease in Latvia. *Genetika*. 2008;44(10):1379-1384.

16. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med Off J Am Coll Med Genet*. 2015;17(5):405-424. doi:10.1038/gim.2015.30

17. Gialluisi A, Incollu S, Pippucci T, et al. The homozygosity index (HI) approach reveals high allele frequency for Wilson disease in the Sardinian population. *Eur J Hum Genet EJHG*. 2013;21(11):1308-1311. doi:10.1038/ejhg.2013.43

18. Walshe JM. Cause of death in Wilson disease. *MovDisord Off J MovDisord Soc*. 2007;22(15):2216-2220. doi:10.1002/mds.21693

19. Poujois A, Woimant F, Samson S, Chaine P, Girardot-Tinant N, Tuppin P. Characteristics and prevalence of Wilson's disease: A 2013 observational population-based study in France. *Clin Res Hepatol Gastroenterol*. June 2017. doi:10.1016/j.clinre.2017.05.011

20. Tai C-S, Wu J-F, Chen H-L, Hsu H-Y, Chang M-H, Ni Y-H. Modality of treatment and potential outcome of Wilson disease in Taiwan: A population-based longitudinal study. *J Formos Med Assoc Taiwan Yi Zhi*. June 2017. doi:10.1016/j.jfma.2017.05.008

21. Lai C-H, Tseng H-F. Population-based epidemiologic study of Wilson's disease in Taiwan. *Eur J Neurol*. 2010;17(6):830-833. doi:10.1111/j.1468-1331.2009.02946.x

22. Cheng N, Wang K, Hu W, et al. Wilson disease in the South chinesehan population. *Can J Neurol Sci J Can Sci Neurol*. 2014;41(3):363-367.

23. Moller LB, Horn N, Jeppesen TD, et al. Clinical presentation and mutations in Danish patients with Wilson disease. *Eur J Hum Genet EJHG*. 2011;19(9):935-941. doi:10.1038/ejhg.2011.80

24. Giagheddu A, Demelia L, Puggioni G, et al. Epidemiologic study of hepatolenticular degeneration (Wilson's disease) in Sardinia (1902-1983). *Acta Neurol Scand*. 1985;72(1):43-55.

25. Adhami EJ, Cullufi P. Wilson's disease in Albania. *Panminerva Med*. 1995;37(1):18-21.

26. Reilly M, Daly L, Hutchinson M. An epidemiological study of Wilson's disease in the Republic of Ireland. *J NeurolNeurosurg Psychiatry*. 1993;56(3):298-300.

27. Bonne-Tamir B, Frydman M, Agger MS, et al. Wilson's disease in Israel: a genetic and epidemiological study. *Ann Hum Genet*. 1990;54(Pt 2):155-168.

28. García-Villarreal L, Daniels S, Shaw SH, 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. *HepatolBaltim Md*. 2000;32(6):1329-1336. doi:10.1053/jhep.2000.20152

29. Dedoussis GVZ, Genschel J, Sialvera T-E, et al. Wilson disease: high prevalence in a mountainous area of Crete. *Ann Hum Genet*. 2005;69(Pt 3):268-274. doi:10.1046/j.1529-8817.2005.00171.x

30. Kim G-H, Yang JY, Park J-Y, Lee JJ, Kim JH, Yoo H-W. Estimation of Wilson's disease incidence and carrier frequency in the Korean population by screening ATP7B major mutations in newborn filter papers using the SYBR green intercalator method based on the amplification refractory mutation system. *Genet Test*. 2008;12(3):395-399. doi:10.1089/gte.2008.0016

31. Zappu A, Magli O, Lepori MB, et al. High incidence and allelic homogeneity of Wilson disease in 2 isolated populations: a prerequisite for efficient disease prevention programs. *J Pediatr Gastroenterol Nutr*. 2008;47(3):334-338. doi:10.1097/MPG.0b013e31817094f6

32. Mak CM, Lam C-W, Tam S, et al. Mutational analysis of 65 Wilson disease patients in Hong Kong Chinese: identification of 17 novel mutations and its genetic heterogeneity. *J Hum Genet*. 2008;53(1):55-63. doi:10.1007/s10038-007-0218-2

33. 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. doi:10.1002/humu.9358

34. Lepori MB, Lovicu M, Dessi V, et al. Twenty-four novel mutations in Wilson disease patients of predominantly Italian origin. *Genet Test*. 2007;11(3):328-332. doi:10.1089/gte.2007.0015

35. Tanzi RE, Petrukhin K, Chernov I, et al. The Wilson disease gene is a copper transporting ATPase with homology to the Menkes disease gene. *Nat Genet*. 1993;5(4):344-350. doi:10.1038/ng1293-344

36. Zali N, Mohebbi SR, Esteghamat S, et al. Prevalence of ATP7B Gene Mutations in Iranian Patients With Wilson Disease. *Hepat Mon*. 2011;11(11):890-894. doi:10.5812/kowsar.1735143X.762

37. Figus A, Angius A, Loudianos G, et al. Molecular pathology and haplotype analysis of Wilson disease in Mediterranean populations. *Am J Hum Genet*. 1995;57(6):1318-1324.

38. Shah AB, Chernov I, Zhang HT, 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):317-328. doi:10.1086/514864

39. Kalinsky H, Funes A, Zeldin A, et al. Novel ATP7B mutations causing Wilson disease in several Israeli ethnic groups. *Hum Mutat*. 1998;11(2):145-151. doi:10.1002/(SICI)1098-1004(1998)11:2<145::AID-HUMU7>3.0.CO;2-I

40. Loudianos G, Dessì V, Lovicu M, et al. Further delineation of the molecular pathology of Wilson disease in the Mediterranean population. *Hum Mutat*. 1998;12(2):89-94. doi:10.1002/(SICI)1098-1004(1998)12:2<89::AID-HUMU3>3.0.CO;2-G

FIGURE LEGENDS

Figure 1: Forest plots of previous studies of WD prevalence

A:Prevalence estimates derived from epidemiological studies

B:Mutant allele frequencies derived from genetic studies

Error bars represent 95% CIs. Overall estimates are calculated using random effects models

PREV = prevalence, MAF = mutant allele frequency, LCL = lower 95% confidence interval, UCL = upper 95% confidence interval, WGHT=weight

Figure 2: ATP7B gene variants and GnomAD allele frequencies

A: Identification and classification of ATP7B disease variants

ATP7B variants were compiled into a new, comprehensive, publicly-available database. Additionally, these were filtered in order to obtain estimates of the prevalence of WD.

i) Reported disease variants were identified by a systematic search of the literature and cross-referencing with existing databases. Variants found in the GnomAD dataset were then filtered by frequency and ACMG criteria to give 'known disease variants'.

ii) Reported disease variants and variants from existing databases were combined to give a list of all known ATP7B variants. Those found in the GnomAD dataset were filtered by frequency and mutation consequence – variants causing major functional or structural changes (frameshift, premature stop codon, splice donor and splice acceptor variants) were classified as 'predicted disease variants'.

n = number of variants, Δnt = nucleotide change. UoA = University of Alberta, AF = allele frequencies

B: WD disease prevalence estimated from GnomAD allele frequencies

Estimates of WD prevalence for 7 ethnicities were calculated using the GnomAD allele frequencies of known and predicted disease variants (identified as per A). For comparison, estimates of prevalence from the meta-analyses of epidemiological and genetic studies are shown by the red and the yellow dotted lines, respectively. Error bars represent 95% CIs calculated using the Wilson score method

C: Mutation consequences of 'known disease variants'

i) Pie chart of the number of variants resulting in each mutation consequence

ii) Pie chart of the proportion of the total mutant allele frequency accounted for by each mutation consequence. del = deletion; ins = insertion; UTR = untranslated region

CONFLICTS OF INTEREST

None to declare

FINANCIAL SUPPORT

None

Figure 1

Click here to access/download;Figure;Figure 1.tif

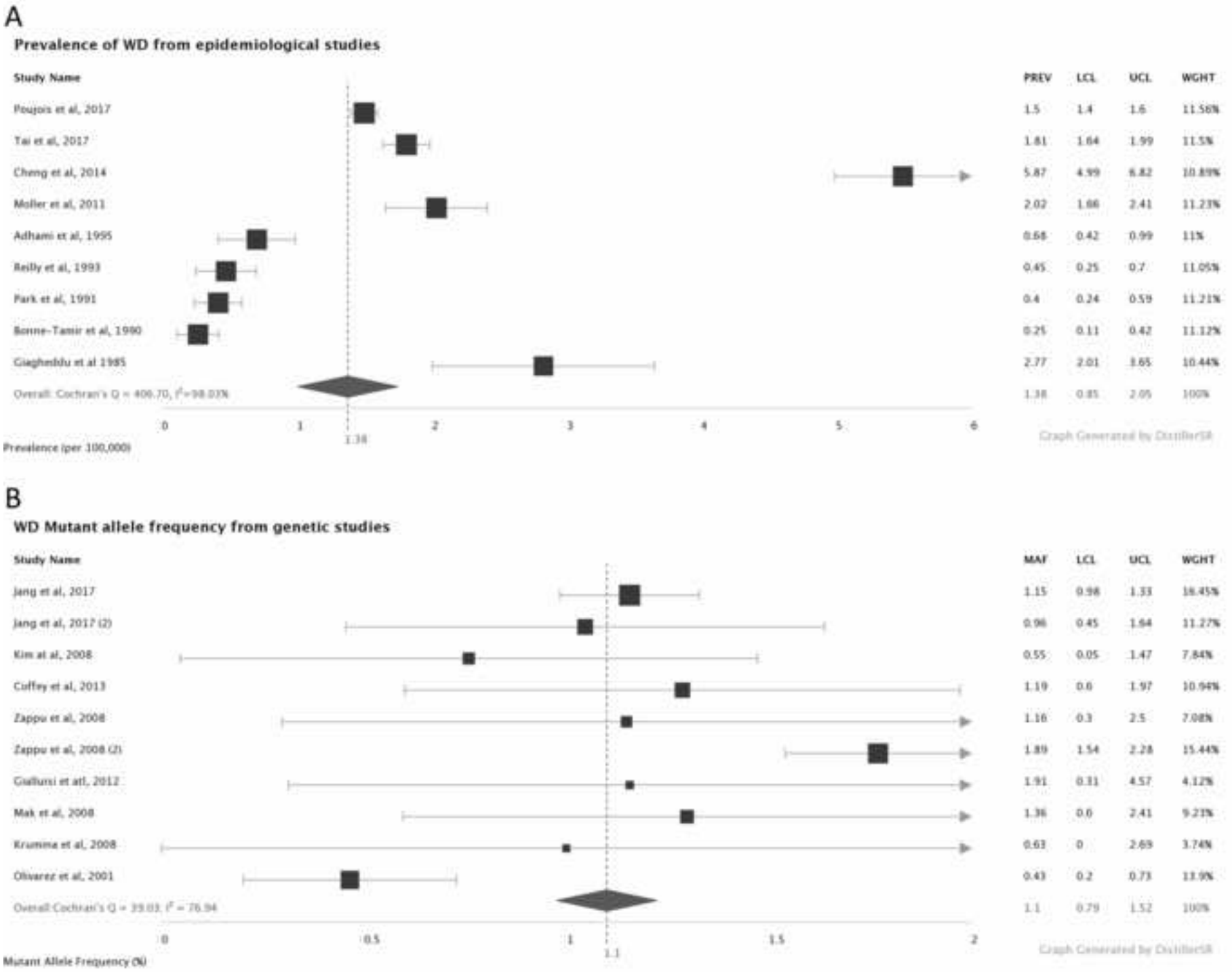


Figure 2

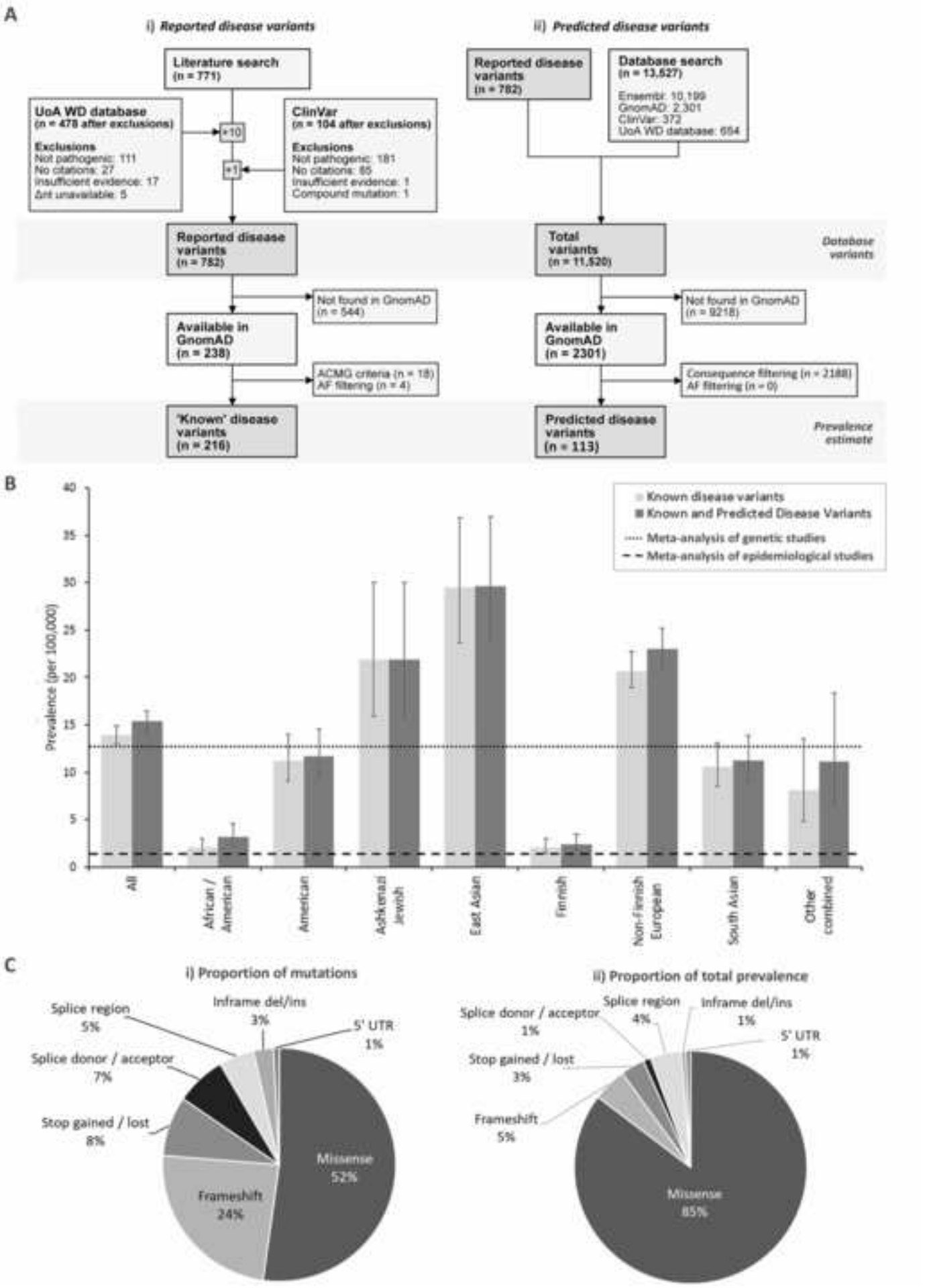


Table 1: Epidemiological studies estimating WD prevalence

Paper	Location	Age Range	Year	Population size (1000s)	Prevalence (per 100000)	AXIS asssessment
Poujois et al, 2017[19]	France	All	2013	58000	1.50	High quality, very low RoB
Tai et al, 2017[20]	Taiwan	All	2000-2011	23162[1]	1.81	High quality, very low RoB
Lai et al, 2010* [21]	Taiwan	All	2005	22770	1.60	High quality, very low RoB
Cheng et al, 2014[22]	Anhui, China	7-75	2008-2011	2700	5.87	Medium quality, risk of sampling bias
Moller et al, 2011[23]	Denmark	All	1990-2008	5494	2.02	High quality, very low RoB
Giagheddu et al 1985[24]	Sardinia	All	Unclear	74	2.77	High quality, low RoB
Adhami et al, 1995[25]	Albania	All	?-1991	3267	0.68	Low quality, unclear RoB, unclear methods
Reilly et al, 1993[26]	Ireland	All	1986	3541[2]	0.54	High quality, low RoB
Park et al, 1991[6]	Scotland	All	1989	5091	0.40	High quality, low RoB
Bonne-Tamir et al, 1990[27]	Israel	All	1958-1985	4106	0.25[3]	High quality, low RoB
Garcia-Villarreal et al, 2000*[28]	NE Canary Islands, Spain	All	1981	1586	38.50	Medium quality, risk of non-responder bias
Dedoussis et al 2005*[29]	Cretan village	All	1978-?	Unclear	6666.67	High quality, low RoB

AXIS = Appraisal tool for Cross-sectional studies[10], * = excluded from further analysis

[1] Source: www.ndc.gov.tw, accessed 14/3; [2] www.cso.ie, accessed 14/3, [3]Estimated from number of cases an study population detailed in paper

Table 2: Genetic studies estimating WD prevalence at birth

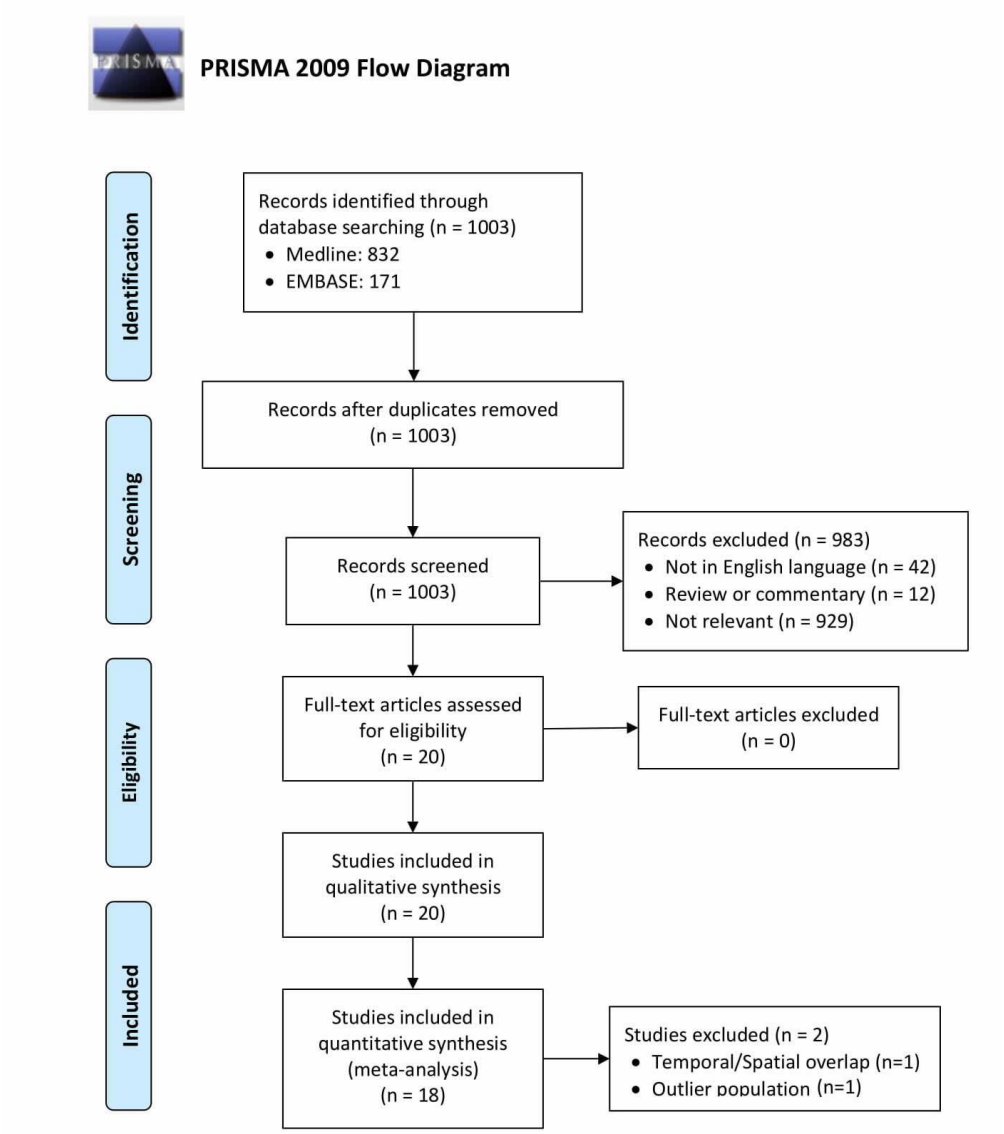
Paper	Method	Location	Sample size	Birth Prevalence (per 100000)	AXIS asssessment
Jang et al, 2017[7]	DNA analysis of neonatal DBSs for 6 mutations	Korea	14835	13.23	High quality, very low RoB
Jang et al, 2017[7]	Retrospective review of sequencing data for any WD disease variant	Korea	1090	9.22	High quality, very low RoB
Kim at al, 2008[30]	DNA analysis of neonatal DBSs for 3 mutations	Korea	476	3.05	High quality, very low RoB
Coffey et al, 2013[3]	DNA analysis of neonatal DBSs for any WD disease variant	UK	1000	14.23	High quality, very low RoB
Zappu et al, 2008[31]	DNA analysis of neonates for 2 mutations	Kalymnos, Greece	397	13.50	High quality, very low RoB
Zappu et al, 2008[31]	DNA analysis of neonates for 1 mutation	Sardinia	5290	35.75	High quality, very low RoB
Gialluisi et al, 2012[17]	Homozygosity index approach	Sardinia	178	36.60	High quality, very low RoB
Mak et al, 2008[32]	DNA analysis of healthy controls for 2 mutations	Hong Kong	660	18.52	High quality, low RoB
Krumina et al, 2008[15]	DNA analysis of healthy controls for 1 mutation	Latvia	157	3.91	High quality, low RoB
Olivarez et al, 2001[8]	DNA analysis of neonates for 1 mutation	New York state, US	2456	1.82	High quality, very low RoB

Jang et al, 2017 and Zappu et al, 2008 are represented twice as they each report two separate estimates based on different populations
AXIS = Appraisal tool for Cross-sectional studies[10]; DBS = Dried blood spot

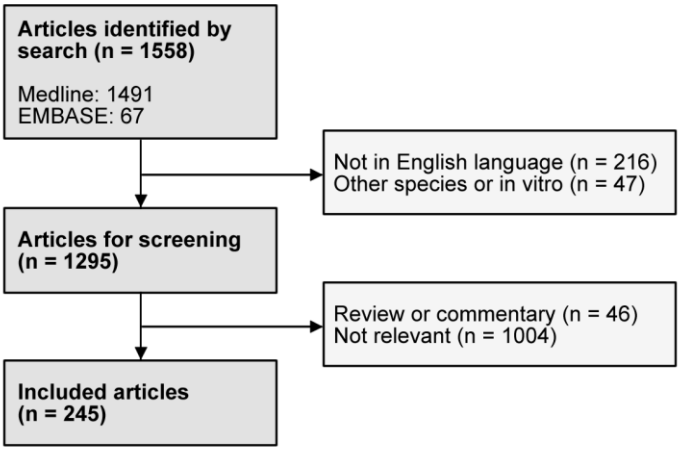
Table 3: Characteristics of the five most frequent 'known disease variants'

Genomic location	cDNA change	Protein change	SIFT	PolyPhen	GnomAD AF	References
51946372	c.2972C>T	p.Thr991Met	deleterious (0)	probably_damaging (0.999)	0.00126	[33,34]
51935019	c.4135C>T	p.Pro1379Ser	deleterious (0)	probably_damaging (0.978)	0.00106	[33]
51944145	c.3207C>A	p.His1069Gln	deleterious (0)	probably_damaging (1)	0.00101	[35–37] etc
51950132	c.2605G>A	p.Gly869Arg	deleterious (0)	probably_damaging (0.996)	0.00072	[28,34,38] etc
51961849	c.1934T>G	p.Met645Arg	tolerated (0.42)	benign (0)	0.00048	[38–40] etc

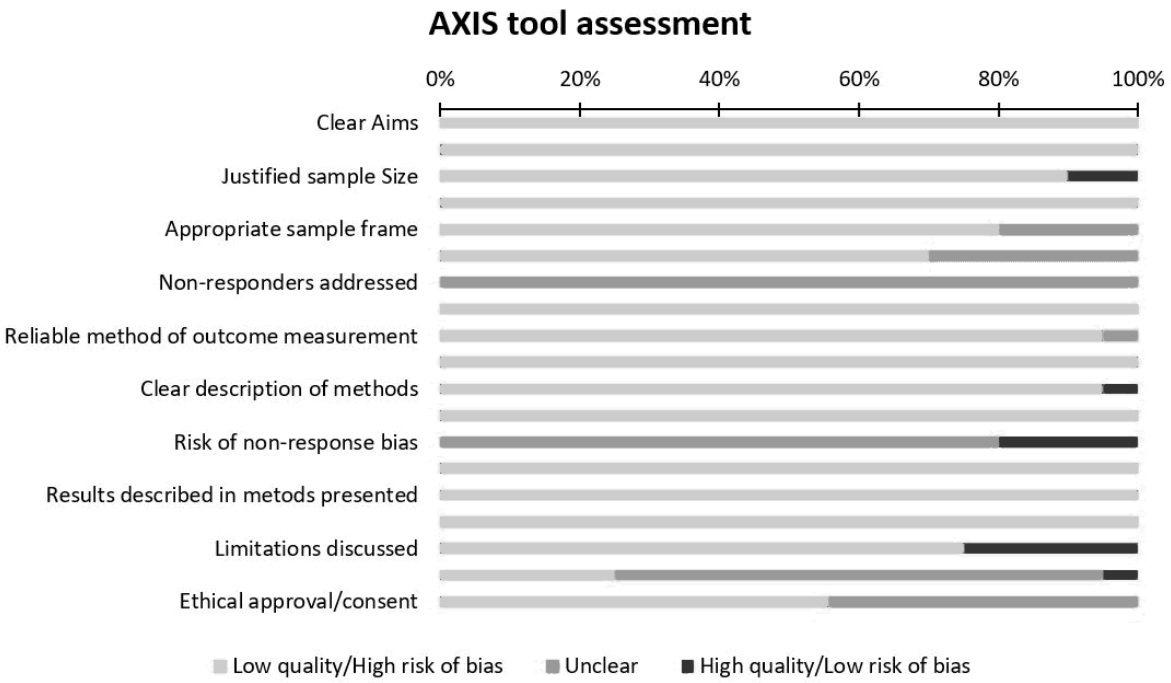
SUPPLEMENTARY DATA



Supplementary Figure 1: PRISMA flow chart for the selection of relevant papers estimating prevalence of WD (n = number of papers)



Supplementary Figure 2: Pathway for identification of relevant WD mutation papers (n = number of papers)



Supplementary Figure 3: AXIS tool assessment of study quality and RoB
Proportion of studies classified as low, unclear and high quality for each of the sections of the AXIS assessment

Supplementary table 1 - WD disease prevalence estimated from GnomAD allele frequencies

	PREVALENCE (per 100,000)								
	All	**African / American**	**American**	**Ashkenazi Jewish**	**East Asian**	**Finnish**	**Non-Finnish European**	**South Asian**	**Other combined**
Known DV	13.9	2.0	11.3	21.9	29.5	2.1	20.7	10.5	8.1
	(12.9-14.9)	(1.4-3.0)	(9.1-14.0)	(15.9-30.0)	(23.6-36.8)	(1.4-3.0)	(18.9-22.7)	(8.5-13.1)	(4.8-13.6)
Known +	15.4	3.2	11.7	21.9	29.7	2.4	23.0	11.2	11.2
Predicted DV	(14.4-16.5)	(2.2-4.6)	(9.5-14.5)	(15.9-30.0)	(23.8-37.0)	(1.7-3.4)	(21.0-25.2)	(9.0—13.9)	(6.8-18.4)

Estimates of WD prevalence for 7 ethnicities were calculated using the GnomAD allele frequencies of known and predicted disease variants (identified as per Figure 2). 95% Confidence intervals are shown in brackets

DV = disease variants

Search terms

Prevalence search

Full MEDLINE search term:

("Wilson Disease" OR "Wilson's Disease" OR ATP7B OR wilson disease[MeSH Terms]) AND (prevalence OR epidemiology OR occurrence OR distribution OR frequency OR prevalence[MeSH Terms] OR epidemiology [MeSH Terms]) NOT review[publication type]

Full EMBASE search term:

("Wilson Disease" OR "Wilson's Disease" OR ATP7B) AND (prevalence OR epidemiology OR occurrence OR distribution OR frequency)
Limit to (human and english language and exclude medline journals and articles)

Mutations search

Full MEDLINE search term:

("Wilson Disease" OR "Wilson's Disease" OR ATP7B OR wilson disease[MeSH Terms])
AND (mutation OR variant OR genetics OR mutation [MeSH Terms])
NOT review[publication type]

Full EMBASE search term:

("Wilson Disease" OR "Wilson's Disease" OR ATP7B) AND (mutation OR variant OR genetics)
limit to (human and english language and exclude medline journals and article)

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	2
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	4
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Supplementary data
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	4
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	4
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	4

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Supplementary data
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Figure S1 and S2
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Tables 1 and 2
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Tables 1 and 2
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Figure 1
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Figure 1
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Figure S3
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	11
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	11-13
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	11-13
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Conflicts of interest document

Supplementary data 2

Click here to access/download

Large Excel File

07-Supplementary data.xlsx