SUPPLEMENT

Exome sequencing identifies novel AD-associated genes.

Authors:

Henne Holstege[1,2,3*#], Marc Hulsman[1,2,3*#], Camille Charbonnier[4*], Benjamin Grenier-Boley[5], Olivier Quenez[4], Detelina Grozeva[6], Jeroen G.J. van Rooij[7], Rebecca Sims[6], Shahzad Ahmad[8], Najaf Amin[8,57], Penny J. Norsworthy[9], Oriol Dols-Icardo[10], Holger Hummerich[9], Amit Kawalia[11], Philippe Amouyel[5], Gary W. Beecham[12], Claudine Berr[13], Joshua C. Bis[14], Anne Boland[15], Paola Bossù[16], Femke Bouwman[1], Dominique Campion[4], Antonio Daniele[17,18], Jean-François Dartigues[19], Stéphanie Debette[19], Jean-François – Deleuze[20], Nicola Denning[21], Anita L DeStefano[22,23,24], Lindsay A. Farrer[22,25,26,27,28], Nick C. Fox[29], Daniela Galimberti[30], Emmanuelle Genin[31], Jonathan L. Haines[32], Clive Holmes[33], M. Arfan Ikram[7,8,34], M. Kamran Ikram[7,8], Iris Jansen[1,35], Robert Kraaij[36], Marc Lathrop[37], Evelien Lemstra[1], Alberto Lleó[10,38], Lauren Luckcuck[6], Rachel Marshall[6], Eden R Martin[12,39], Carlo Masullo[40], Richard Mayeux[41], Patrizia Mecocci[42], Alun Meggy[21], Merel O. Mol[7], Kevin Morgan[43], Benedetta Nacmia[44], Adam C Naj[45,46], Pau Pastor[47], Margaret A. Pericak-Vance[12], Rachel Raybould[21], Richard Redon[48], Anne-Claire Richard[4], Steffi G Riedel-Heller[49], Fernando Rivadeneira[36], Stéphane Rousseau[4], Natalie S. Ryan[29], Salha Saad[6], Pascual Sanchez-Juan[50], Gerard D. Schellenberg[46], Philip Scheltens[1], Jonathan M. Schott[29], Davide Seripa[51], Gianfranco Spalleta[52], Betty Tijms[1], André G Uitterlinden[8,36], Sven J. van der Lee[1,2,3], Michael Wagner[53,54], David Wallon[4], Li-San Wang[46], Aline Zarea[4], Marcel J.T. Reinders[2], Jordi Clarimon[10], John C. van Swieten[7], John J. Hardy[55,29], Alfredo Ramirez[11,53], Simon Mead[9], Wiesje M. van der Flier[1,56], Cornelia M van Duijn[8,57], Julie Williams[21], Gaël Nicolas[4*#,] Céline Bellenguez[5*], Jean-Charles Lambert[5*#]

*Authors contributed equally to this work

[#] To whom correspondence should be addressed

— Henne Holstege: h.holstege@amsterdamumc.nl
— Marc Hulsman: m.hulsman@amsterdamumc.nl
— Gael Nicolas: gaelnicolas@hotmail.com
— Jean-Charles Lambert: jean-charles.lambert@pasteur-lille.fr

Affiliations:

(1) Alzheimer Center, Department of Neurology, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands; **(2)** Department of Clinical Genetics, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands; **(3)** Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands; **(4)** Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics and CNR-MAJ, F 76000, Normandy Center for Genomic and Personalized Medicine, Rouen, France; **(5)** Univ Lille, Inserm, CHU Lille, Institute Pasteur de Lille, U1167 - RID-AGE - Risk factors and molecular determinants of age-related diseases; Institute Pasteur de Lille, University of Lille, Lille Cedex, France; **(6)** Division of Psychological Medicine and Clinical Neuroscience, School of Medicine, Cardiff University, Cardiff, UK; **(7)** Department of Neurology, Erasmus Medical Centre, Rotterdam, The Netherlands; **(8)** Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands; **(9)** MRC Prion Unit at UCL; (**10)** Sant Pau Biomedical Research Institute, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain; **(11)** Division of Neurogenetics and Molecular Psychiatry, Department of Psychiatry and Psychotherapy, University of Cologne, Medical Faculty, 50937 Cologne, Germany; **(12)** John P Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami; **(13)** Univ. Montpellier, Inserm U1061, Neuropsychiatry: epidemiological and clinical research, PSNREC; **(14)** Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA (USA); (**15)** Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057, Evry, France; **(16)** IRCCS Fondazione Santa Lucia, Department of Clinical and Behavioral Neurology, Experimental Neuro-psychobiology Lab Via Ardeatina, 306, I-00179 Roma, Italy; **(17)** Department of Neuroscience, Università Cattolica del Sacro Cuore, Rome, Italy; **(18)** Neurology Unit, IRCCS Fondazione Policlinico Universitario A. Gemelli, Rome, Italy; (**19)** Univ. Bordeaux, Inserm U1219, Bordeaux Population Health Research Center, Bordeaux, France; CHU de Bordeaux, Department of Neurology, Bordeaux, France; **(20)** Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057, Evry, France; **(21)** UKDRI@ Cardiff, School of Medicine, Cardiff University, Cardiff, UK; **(22)** Department of Biostatistics, Boston University School of Public Health; **(23)** Department of Neurology, Boston University School of Medicine; **(24)** Framingham Heart Study; **(25)** Dept. of Medicine (Biomedical Genetics), Boston Univ. School of Med; **(26)** Department of Neurology, Boston University School of Medicine; **(27)** Department of Ophthalmology, Boston Univ. School of Medicine; **(28)** Department of Epidemiology, Boston Univ. School of Public Health; **(29)** Dementia Research Centre, UCL Queen Square Institute of Neurology, UK Dementia Research Institute; **(30)** University of Milan, Centro Dino Ferrari, CRC Molecular basis of Neuro-Psycho-Geriatrics diseases, Milan, Italy; **(31)** Univ Brest, Inserm, EFS, CHU Brest, UMR 1078, GGB, F-29200, Brest, France; **(32)** Population & Quantitative Health Sciences and Cleveland Institute for Computational Biology, School of Medicine, Case Western Reserve University, Cleveland, Ohio USA; **(33)** Clinical and Experimental Science, Faculty of Medicine, University of Southampton, Southampton, UK; **(34)** Department of Radiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands; **(35)** Complex Trait Genetics Lab, CNCR, VU University, Amsterdam; **(36)** Department of Internal Medicine, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands; **(37)** McGill University and Genome Quebec Innovation Centre, 740 Doctor Penfield Avenue, Montreal, QC, H3A 0G1, Canada; **(38)** Network Center for Biomedical Research in Neurodegenerative Diseases (CIBERNED), Madrid, Spain; **(39)** Department of Human Genetics , University of Miami Leonard M. Miller School of Medicine; **(40)** Istituto di Neurologia Policlinico Universitario A. Gemelli, 00168, Rome, Italy; **(41)** Columbia University; **(42)** Institute of Gerontology and Geriatrics, Department of Medicine and Surgery, University of Perugia, Italy; **(43)** Human Genetics, School of Life Sciences, University of Nottingham, UK NG7 2UH; **(44)** Department of Neuroscience, Psychology, Drug Research and Child Health , University of Florence, Italy; **(45)** Department of Biostatistics, Epidemiology, and Informatics; Perelman School of Medicine, University of Pennsylvania; **(46)** Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania; **(47)** Memory Disorders Unit, Department of Neurology, Hospital Universitari Mutua de Terrassa, Terrassa, Barcelona, Spain; **(48)** Université de Nantes, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000, Nantes, France; **(49)** Institute of Social Medicine, Occupational Health and Public Health, Medical Faculty, University of Leipzig, Leipzig, Germany; **(50)** Neurology Service and Centro de Investigación en Red de Enfermedades Neurodegenerativas (CIBERNED), Marques de Valdecilla University Hospital (University of Cantabria and IDIVAL), Santander, Spain; **(51)** Laboratory of Gene Therapy, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, FG, Italy; **(52)** Laboratory of Neuropsychiatry, IRCCS Santa Lucia Foundation, Rome, Italy; **(53)** Department of Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany; **(54)** DZNE, German Center of Neurodegenerative Diseases, Bonn, Germany; **(55)** Department of Neurodegenerative Disease, Reta Lila Weston Laboratories, Queen Square Genomics, UCL Dementia Research Institute, Wing 1.2 Cruciform Building, Gower Street, London WC1E 6BT; **(56)** Department of Epidemiology &Biostatistics, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands; **(57)** Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom.

INDEX

Supplemental tables

Table S1: Sample-characteristics per contributing study

	Samples		Gender (%female)		Case/Control							APOE genotype		Diagnostic validation		
			Cases	Controls		EOAD		LOAD		Controls		Cases	Controls			
study	before QC (#)	after QC (#)	%	%	% cases	#	AAO	#	AAO	#	ALS	% E4	% E4	Neuro-patho-logy	CSF	Clinical
France																
ADES-FR	3318	3254	63.1%	58.7%	61.4%	1068	59.0	930	78.2	1256	75.5	28.9%	11.6%	15	624	2615
Germany																
AgeCoDe-UKBonn	394	371	68.4%	-	99.7%	98	59.0	272	84.7	1	-	24.3%	50.0%	0	0	371
Spain																
Barcelona SPIN	60	59	44.0%	44.4%	84.7%	50	57.3	0	N/A	9	72.8	3.0%	16.7%	37	13	9
The Netherlands																
100-plus Study	276	254	84.4%	71.6%	25.2%	0	NA	64	101.5	190	102.9	7.0%	8.7%	0	0	274
ERF Study	1325	400	50.0%	57.3%	1.0%	1	-	3	75.8	396	48.1	50.0%	17.0%	0	0	400
AC-ERC	81	70	54.3%	NA	100.0%	57	57.0	13	69.1	0	NA	44.3%	NA	3	40	27
Rotterdam Study	2699	1891	68.9%	55.4%	19.4%	1	-	366	83.5	1524	82.7	25.8%	14.0%	0	0	1891
ADC-Amsterdam	518	483	55.1%	NA	100.0%	341	57.3	142	68.6	0	NA	30.6%	NA	0	483	0
United Kingdom																
PERADES[1]	4936	4140	58.3%	57.2%	83.3%	1265	58.1	2185	76.7	690	81.5	31.6%	12.0%	0	0	4140
CBC	471	363	54.1%	40.1%	30.6%	33	60.1	78	76.8	252	75.8	36.8%	18.7%	363	0	0
UCL-DRC EOAD	539	409	54.8%	NA	100.0%	389	54.9	20	76.6	0	NA	29.7%	N/A	7	35	367
Europe total																
ADES	*15088*	*12057*	*60.2%*	*55.7%*	*62.1%*	*3336*	*57.4*	*4151*	*79.0*	*4570*	*77.5*	*29.7%*	*13.5%*	*788*	*1195*	*10094*
United States																
ADSP[2]	*11365*	*9651*	*57.6%*	*58.3%*	*54.7%*	*757*	*62.4*	*4519*	*77.2*	*4375*	*86.5*	*23.7%*	*7.2%*	*0*	*0*	*9651*
Total	**25,982**	**21,345**	**59.2%**	**57.5%**	**59.3%**	**4060**	**58.8**	**8,592**	**77.9**	**8693**	**82.1**	**26.9%**	**10.1%**	**425**	**1195**	**19745**

Characteristics of the samples contributed by each study, grouped by country. A.A.O: mean age at onset; A.L.S. mean age at last screening. [1]The PERADES sample is UK-based, but also includes samples from Spain and Italy. [2]The ADSP cohort is composed of cohorts from the ADGC and CHARGE consortia.

Table S2: Capture kits used by the contributing studies.

Study	Capture kits (#samples, after QC)
France	
ADES-FR	Agilent V1: 6,Agilent V3: 10,Agilent V4: 119, Agilent V4UTR: 14,Agilent V5UTR: 789, Agilent V5: 1362, WGS: 954
Germany	
AgeCoDe-UKBonn	Nimblegen V2: 371
Spain	
Barcelona SPIN	Nimblegen V3: 59
The Netherlands	
100-plus Study	Agilent V6: 40,Nimblegen V3: 214
ERF Study	Agilent V4: 400
AC-ERC	Nimblegen V2: 70
Rotterdam Study	Nimblegen V2: 1891
ADC-Amsterdam	Agilent V6: 180,Nimblegen V3: 303
United Kingdom	
PERADES[1]	Nextera v1.2: 4140
CBC	Nimblegen V2: 63, Multiplex Illumina TruSeq v2: 100, Multiplex Illumina TruSeq: 200
UCL-DRC EOAD	Sureselect: 5, Haloplex: 404
United States	
ADSP-BCM	WGS: 11, Nimblegen VCRome v2.1: 2186
ADSP-Broad Institute	WGS: 16, Illumina Rapid Capture Exome: 4112
ADSP-WUGSC	WGS: 36, Nimblegen VCRome v2.1: 3290

WGS=whole genome sequencing.

Table S3: Single variant analysis.

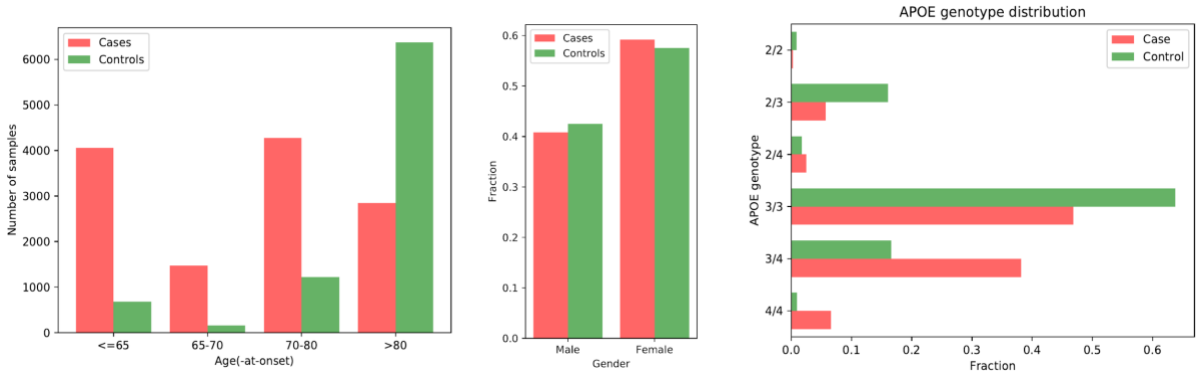
Gene Transcripts (canonical=bold)	SNP id	Burden outlier FDR	Case/Control		Protein change (per transcript) (bold: name in text)	Original reason for exclusion	REVEL	LOF	MAF	Refinement analysis bold=action *italic=reason for rejection*
			FDR	OR (95%CI)						
SORL1 **ENST00000260197: A** ENST00000525532: B ENST00000534286: C ENST00000532694: D ENST00000527934: E	rs2298813	NA	**4%**	1.16 (1.05-1.27)	A: **A528T**	MAF/REVEL	11		4.9%	*too common*
	rs140384365	NA	**20%**	2.49 (1.22-5.07)	A: **V1459I**, B: V403I, C: V369I, D: V305I, E: V74I	REVEL	9		0.08%	**added**
	rs143536682	**6.3%**	79%	0.53 (0.19-1.47)	A: **S2175R**, B: S1119R, C: S1085R, D: S1021R, E: S790R		81		0.04%	**removed**
TREM2 **ENST00000373113: A** ENST00000373122: B ENST00000338469: C	rs75932628	84%	**<<1%**	3.74 (2.84-4.92)	A,B,C: **R47H**		34		0.54%	kept
	rs143332484	NA	**<<1%**	1.58 (1.32-1.88)	A,B,C: **R62H**	MAF/REVEL	4		1.3%	*too common*
	rs142232675	NA	**1%**	2.63 (1.56-4.45)	A,B,C: **D87N**	REVEL	20		0.15%	**added**
	rs2234255	NA	**1%**	6.39 (2.68-15.2)	A,B,C: **H157Y**	REVEL	0		0.05%	**added**
	rs538447052	**11%**	20%	1.91 (0.71-5.08)	B: splice acceptor variant		NA	HC	0.04%	**removed**
	rs2234256	NA	**2%**	2.27 (1.34-3.86)	A: **L211P**	REVEL	0		0.15%	**added**
	rs2234258	NA	20%	2.28 (0.90-5.78)	C: **W191X** (stop gained)	REVEL	NA	LC	0.05%	*low OR*
ABCA7 **ENST00000263094: A** ENST00000433129: B ENST00000435683: C	rs546173555	**5.0%**	89%	1.09 (0.37-3.20)	A,B: **R19W**		54		0.04%	**removed**
	rs201665195	100%	**<1%**	3.67 (2.10-6.42)	A,B: **L101R**		28		0.18%	kept
	rs72973581	NA	**2%**	0.89 (0.81-0.97)	A,B: **G215S**, C: G77S	MAF/REVEL	16		5.6%	*too common*
	rs3764647	NA	**2%**	0.85 (0.76-0.95)	A,B: **H395R**, C:H257R	MAF/REVEL	18		3.5%	*too common*
	rs547447016	NA	NA	2.01 (1.35-3.01)	*A,B:* ***EEQ708-710X****, C: EEQ570-572X*	diff. miss	NA	HC	0.27%	*QC*
	rs117187003	**1.8%**	47%	0.84 (0.61-1.15)	A,B: **V1599M**, C: V1461M		58		0.41%	**removed**
	rs4147918	NA	**1%**	0.84 (0.76-0.94)	A,B: **Q1686R**, C:Q1548R	MAF/REVEL	15		3.6%	*too common*
	rs200538373	NA	NA	1.67 (1.23-2.28)	*A,B,C:* ***c.5570+5G>C***	diff.miss/REVEL	NA	Lit. [1]	0.43%	*QC*
ATP8B4 **ENST00000284509: A** ENST00000559829: B	rs74811880	99%	**5%**	3.14 (1.55-6.34)	A,B: **H987R**		26		0.08%	kept
	rs74012834	55%	**7%**	1.41 (1.05-1.91)	A,B: **C874R**		25		0.45%	kept
	rs138799625	99%	**<<1%**	1.83 (1.48-2.26)	A,B: **G395S**		86		0.92%	kept
ABCA1 **ENST00000374736: A**	rs2066715	NA	**17%**	0.93 (0.85-1.01)	A: **V825I**	MAF/REVEL	0		6.1%	*too common*
	rs2066714	NA	**2%**	0.91 (0.86-0.97)	A: **I883M**	MAF/REVEL	0		13.3%	*too common*
	rs140365800	**4.8%**	74%	0.81 (0.29-2.22)	A: **D1018G**		84		0.04%	**removed**
	rs143180998	NA	**13%**	0.49 (0.25-0.96)	A: **A1182T**	REVEL	17		0.09%	*protective*
	9:107565564:	**0.3%**	NA	0.94 (0.19-4.52)	A: splice donor variant		NA	HC	0.02%	**removed**
	rs150125857	NA	**13%**	2.75 (1.27-5.95)	A: **R1680Q**	REVEL	0		0.07%	**added**
	rs146292819	NA	**2%**	4.16 (2.02-8.56)	A: **N1800H**	REVEL	0		0.08%	**added**
CBX3 ENST00000409747: A	rs142550836	100%	**<<1%**	0.16 (0.08-0.34)	A: **N74S**		36		0.07%	kept
PRSS3 **ENST00000361005: A** ENST00000379405: B ENST00000342836: C ENST00000429677: D	rs143209949	**12%**	34%	0.75 (0.41-1.38)	A:**R125C**, B:R68C, C: R82C, D:R61C		53		0.11%	kept

See detailed gene discussion for explanation. Variants are shown that are i) included in the burden but considered outliers (outlier FDR < 20%, fisher exact test), ii) are a

missense or LOF variant and associated with AD (case/control FDR < 20%, logistic regression), iii) are mentioned in the text for other reasons. Refinement was only performed with variants from the first two categories. LOF: LC/HC: LOFTEE low/high confidence classification, Lit: based on literature this variant is known to be a LOF variant in ABCA7, but this was not recognized by LOFTEE. Refinement analysis: variants that were common (MAF > 1%), or had the opposite effect were not considered for the refined burden test.

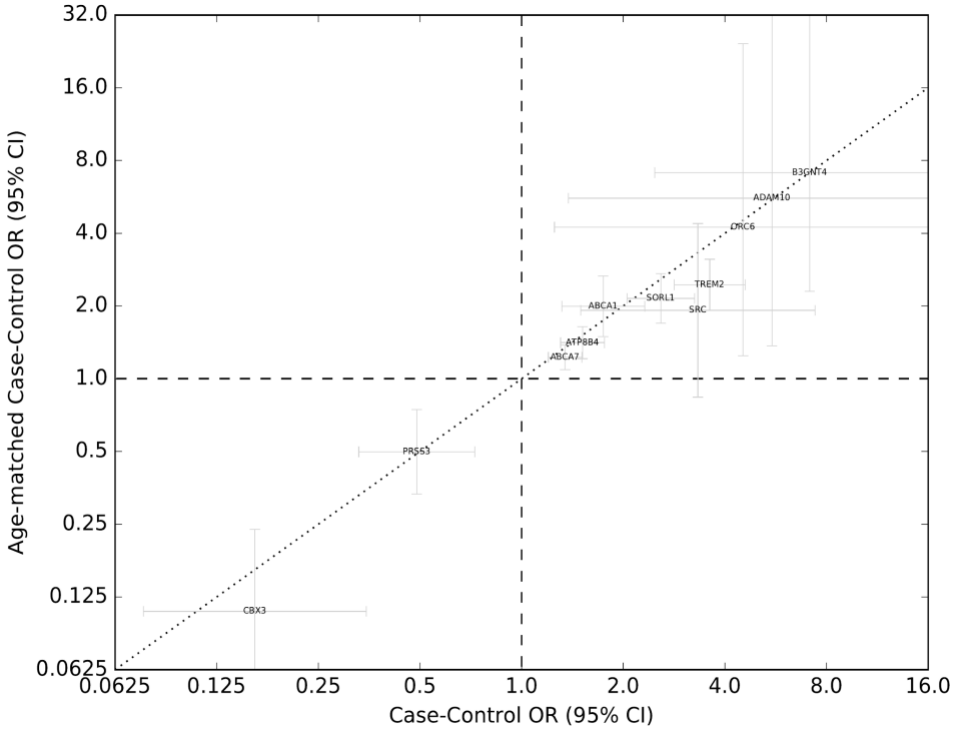
Supplemental figures

Figure S1: Age, gender, APOE genotype distribution



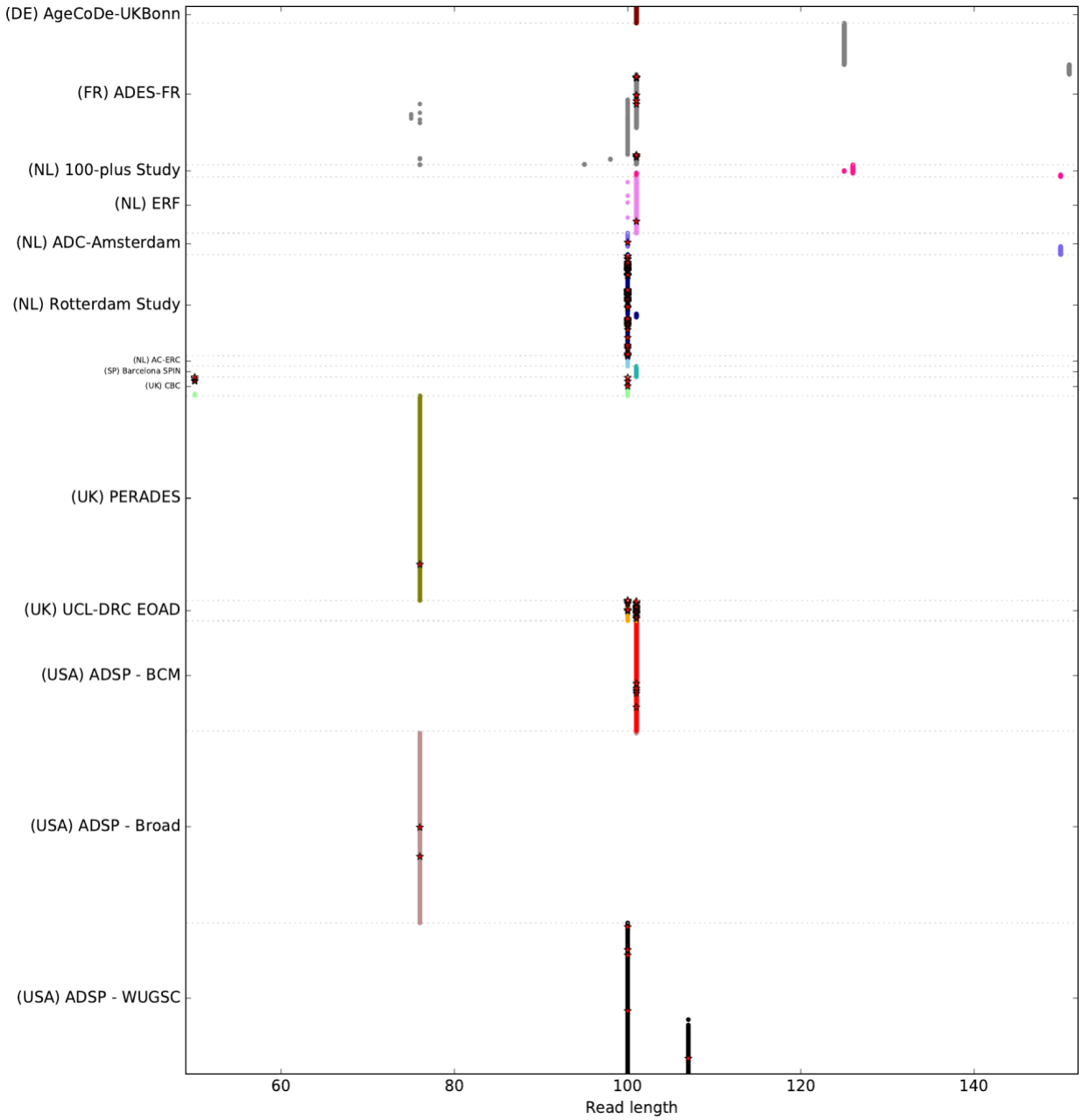
Age, gender and APOE genotype distribution of all samples, stratified by case/control status.

Figure S2: Sensitivity Analysis: AD vs Age association



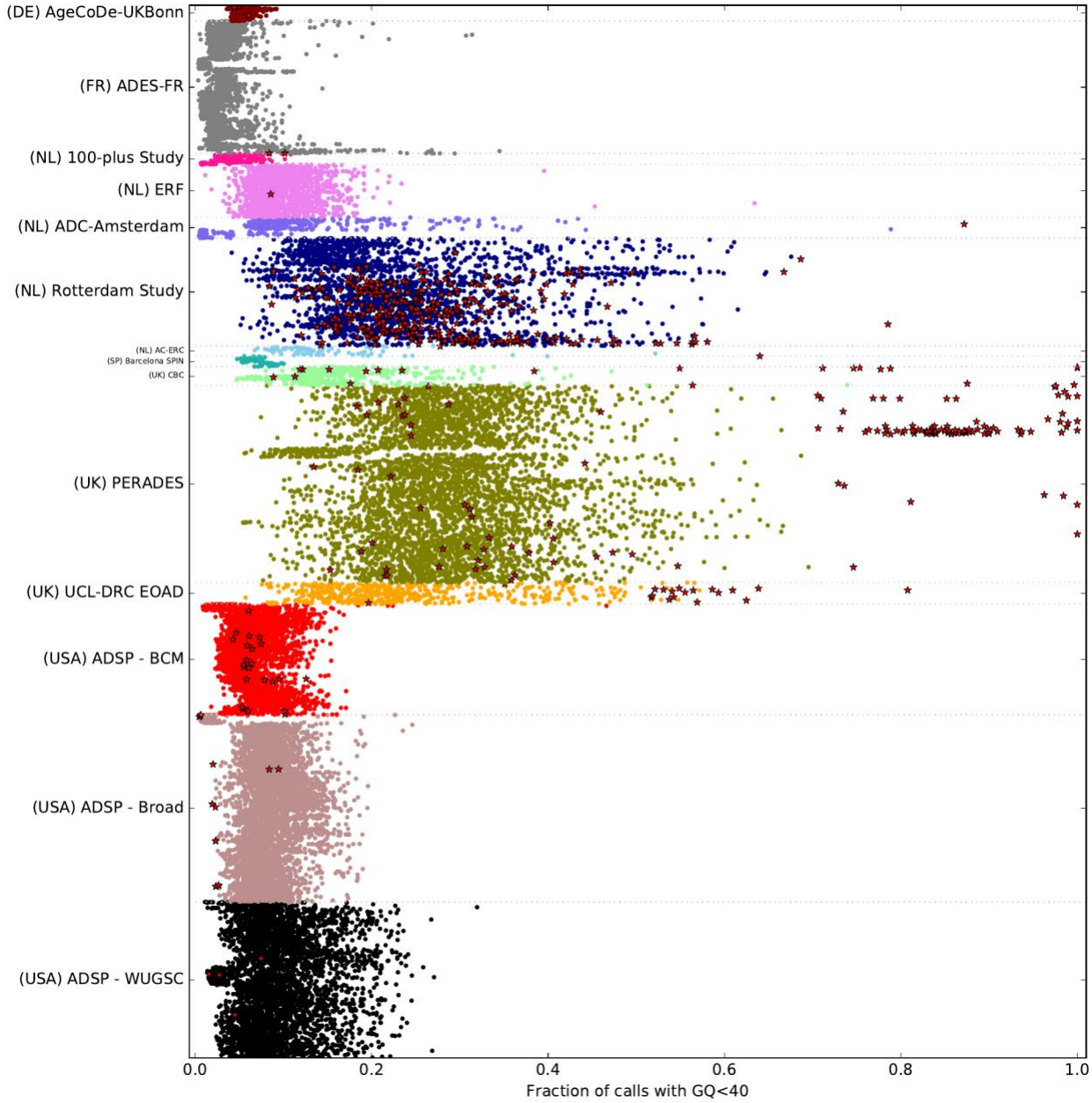
Sensitivity analysis of the gene burden tests (for the most significant deleteriousness thresholds, Table 1). Comparison of the case/control odds ratio of an age-matched and a normal analysis. Age-matching was performed as described in the methods.

Figure S3: Read length per study



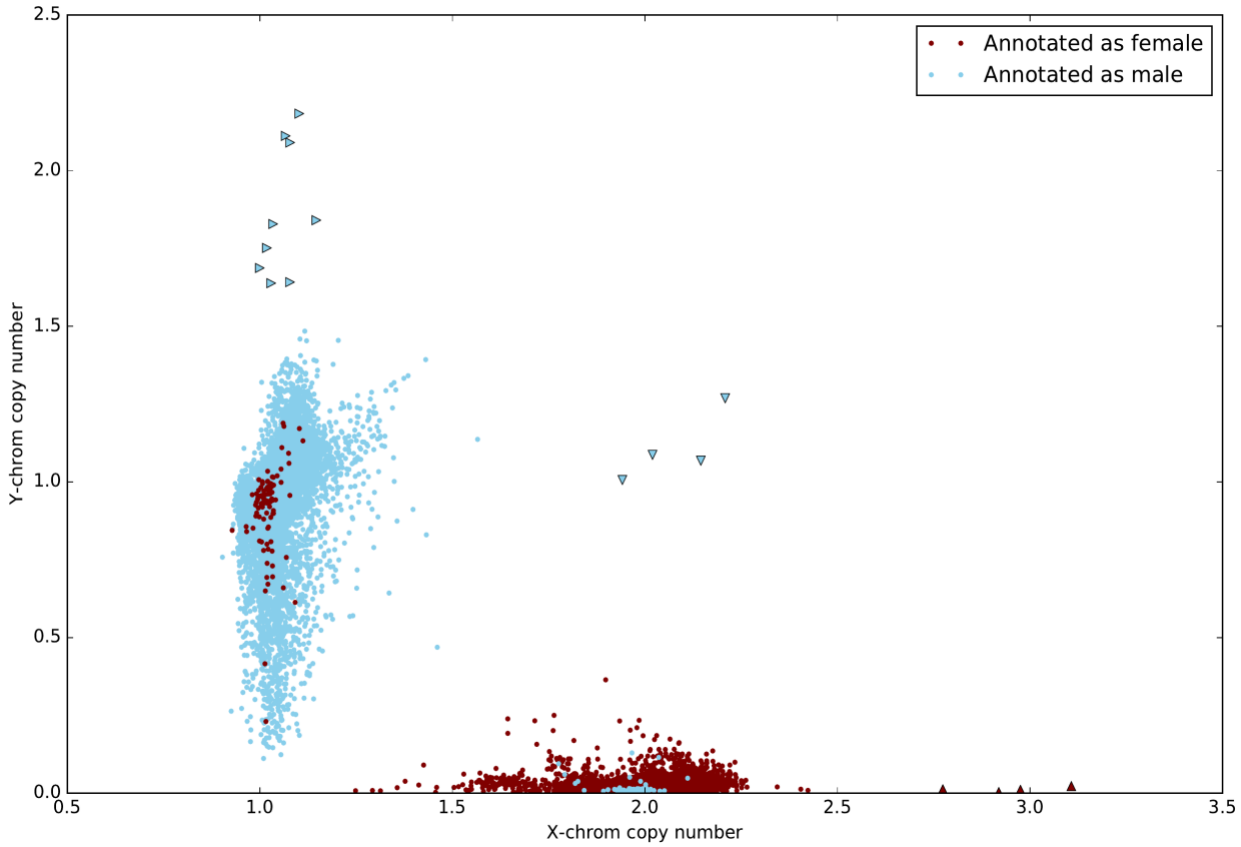
Read length.

Figure S4: Genotype Quality



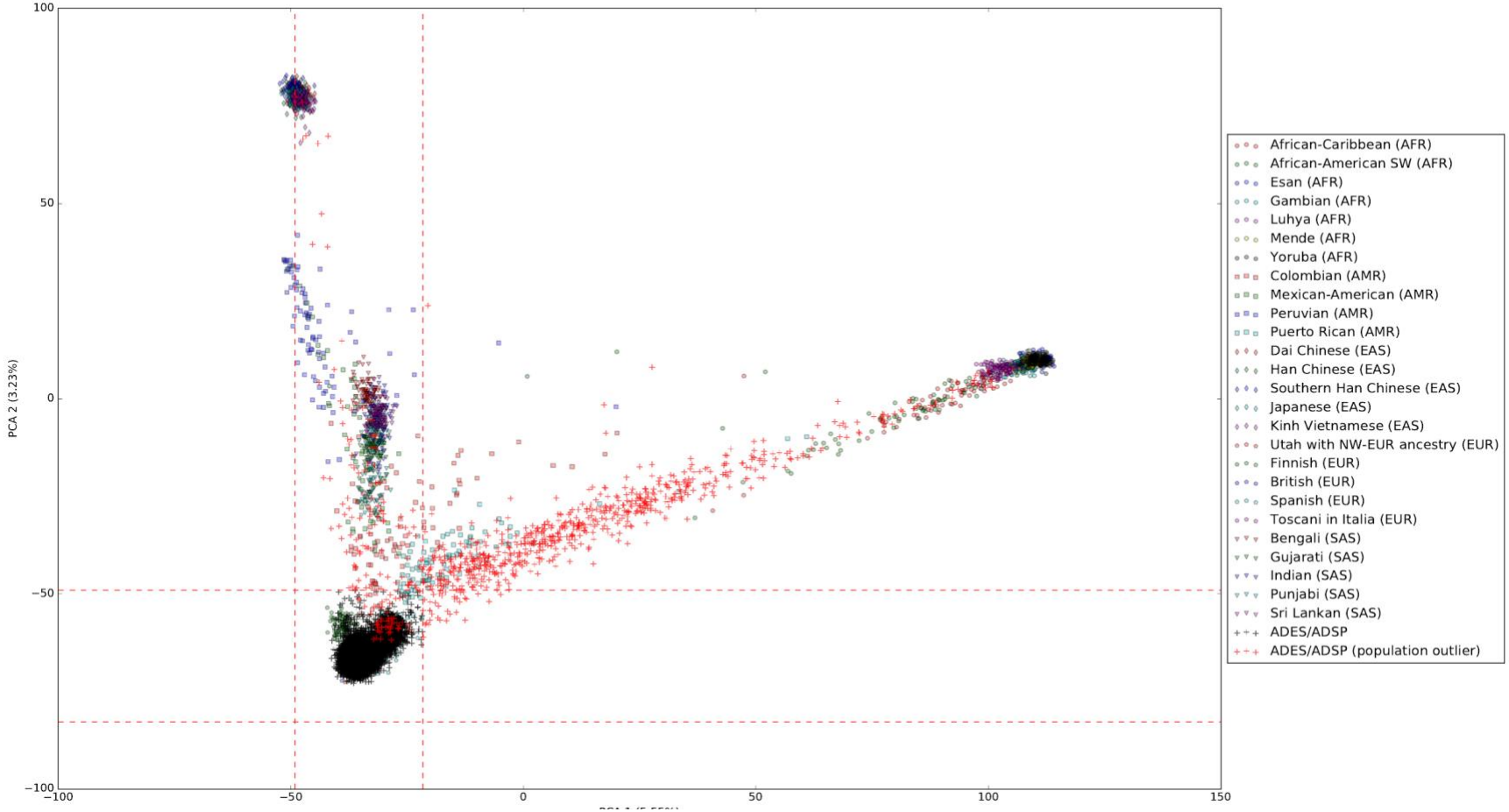
Fraction of genotype calls with a genotype quality < 40. Each sample is evaluated in context of its capture kit. Samples that are considered outliers due to missingness or contamination are indicates with a red '*' symbol.

Figure S5: Genetic sex



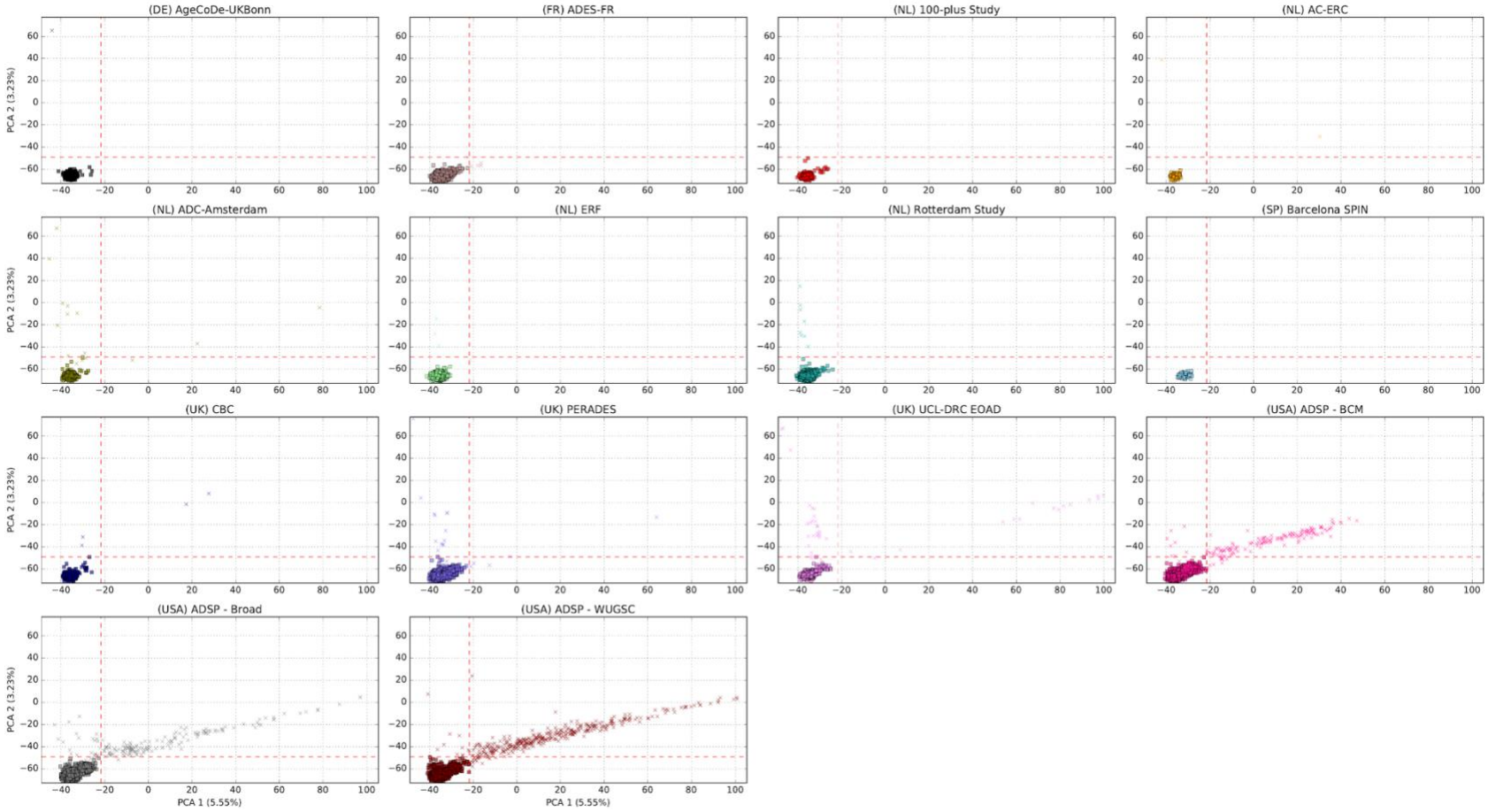
Sex chromosome copy number versus gender annotation. Samples that failed the sex check were plotted last. Samples that were classified as XXY, XXY and XXX are indicated by respectively right, down and upwards pointing triangle symbols.

Figure S6: PCA: Sample population compared to 1,000G population samples



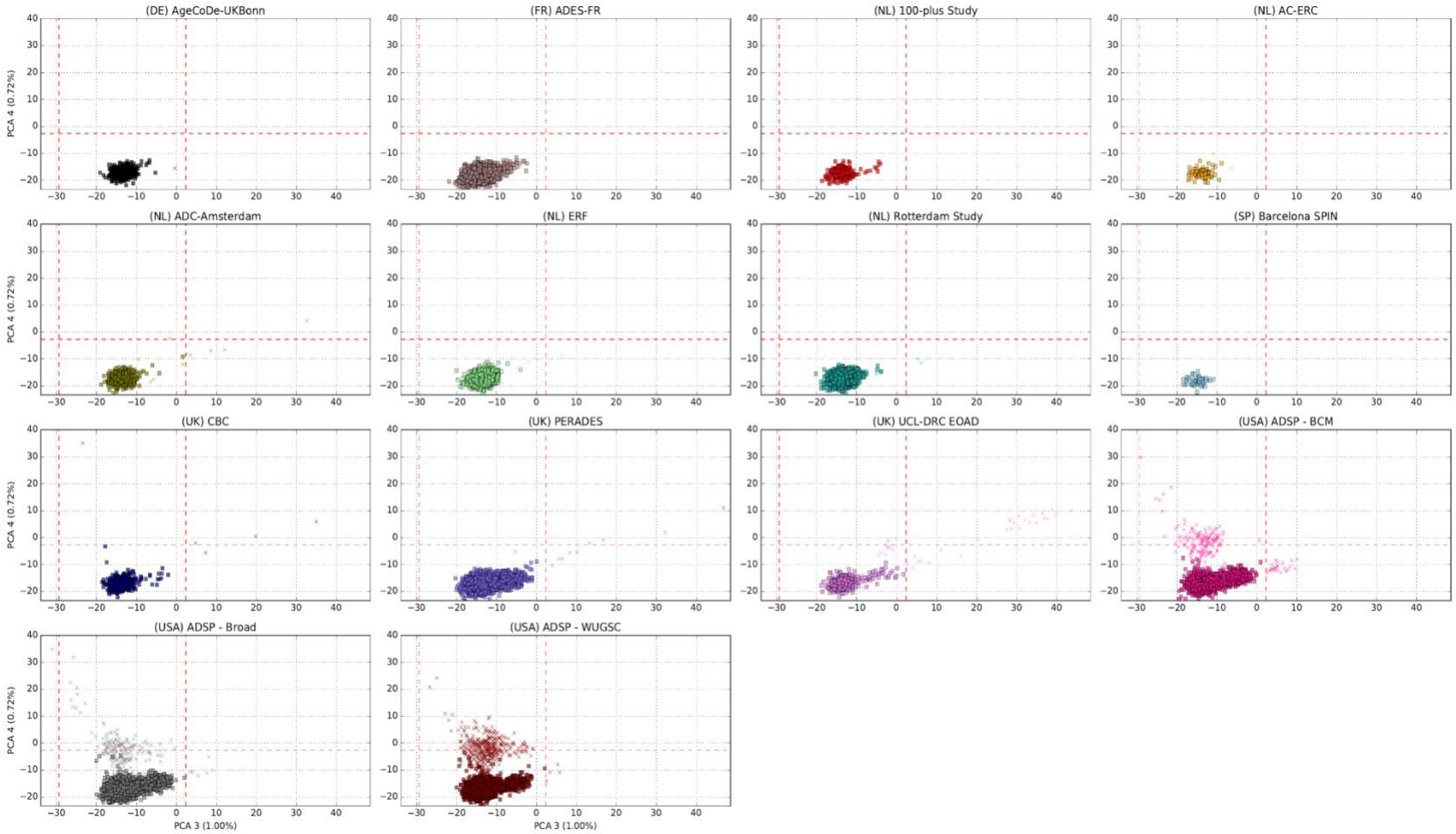
First two PCA components of the study samples, together with 1000 genome samples for reference. Samples in red are considered population outliers.

Figure S7: first two population PCA components per study



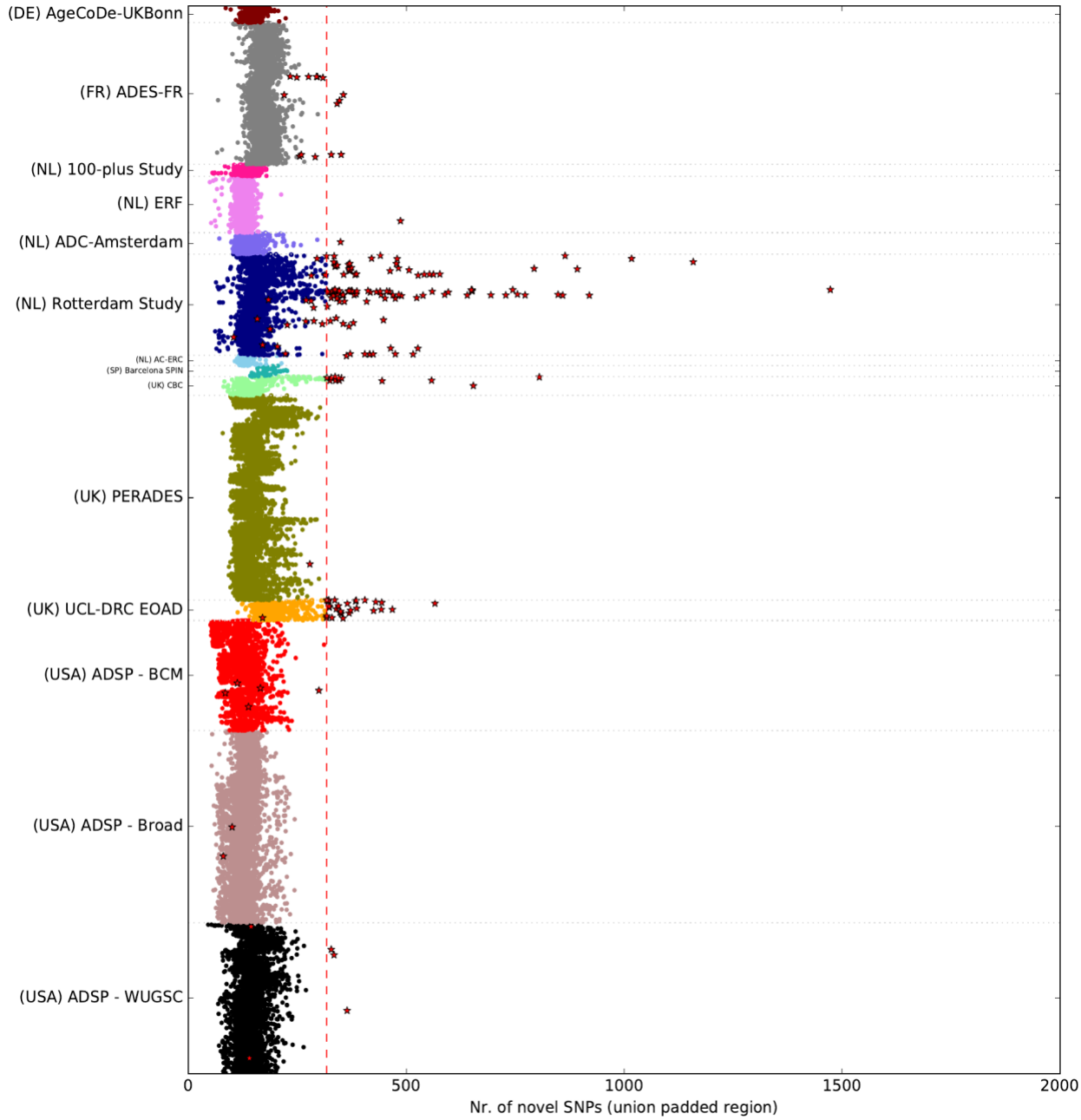
First two PCA components per study. Samples indicated as a ‘x’ are outliers.

Figure S8: Third and fourth population PCA components per study



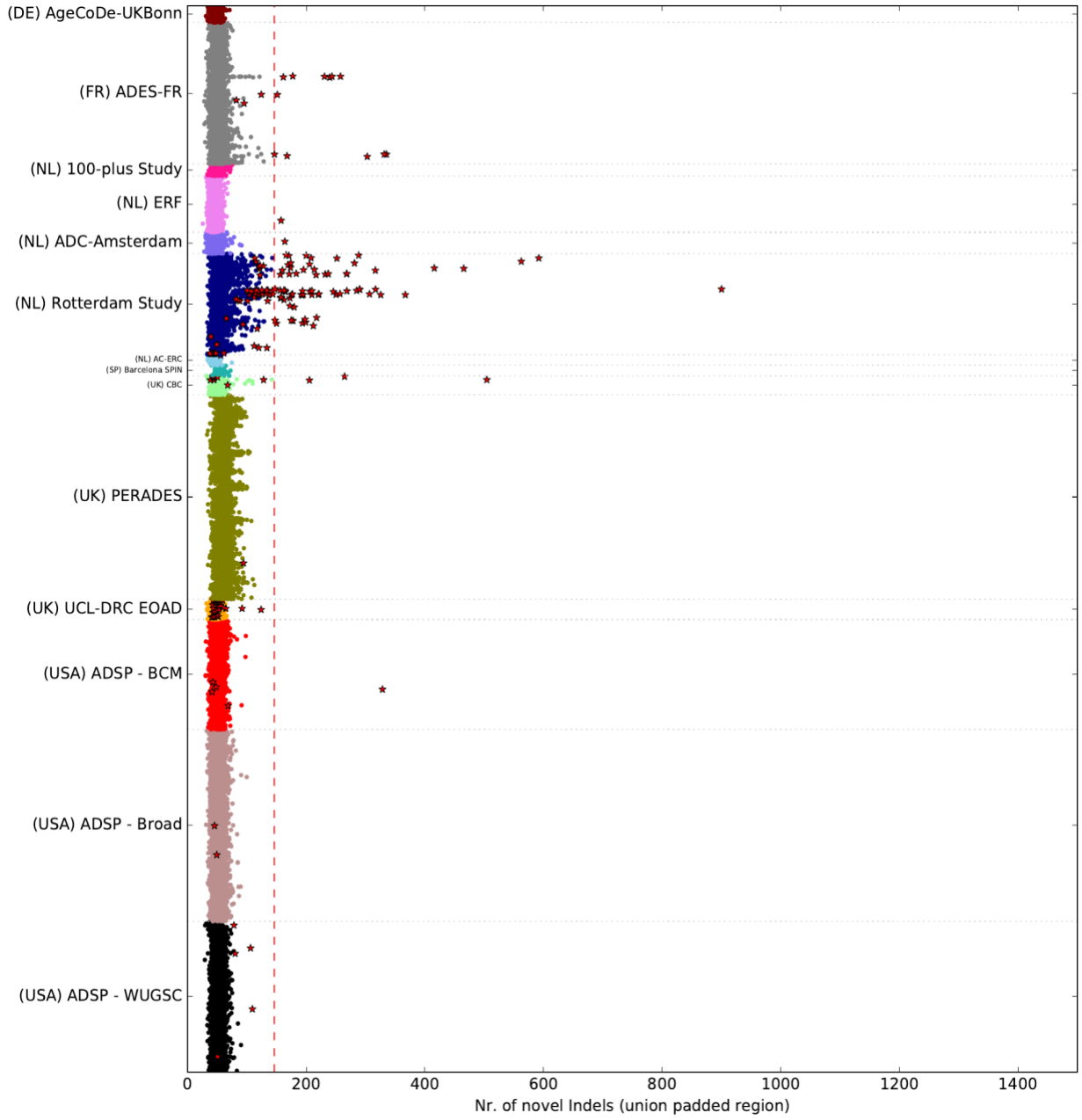
Third and fourth PCA component for each study. Samples indicated as a ‘x’ are outliers.

Figure S9: Number of novel SNPs (union of capture kits)



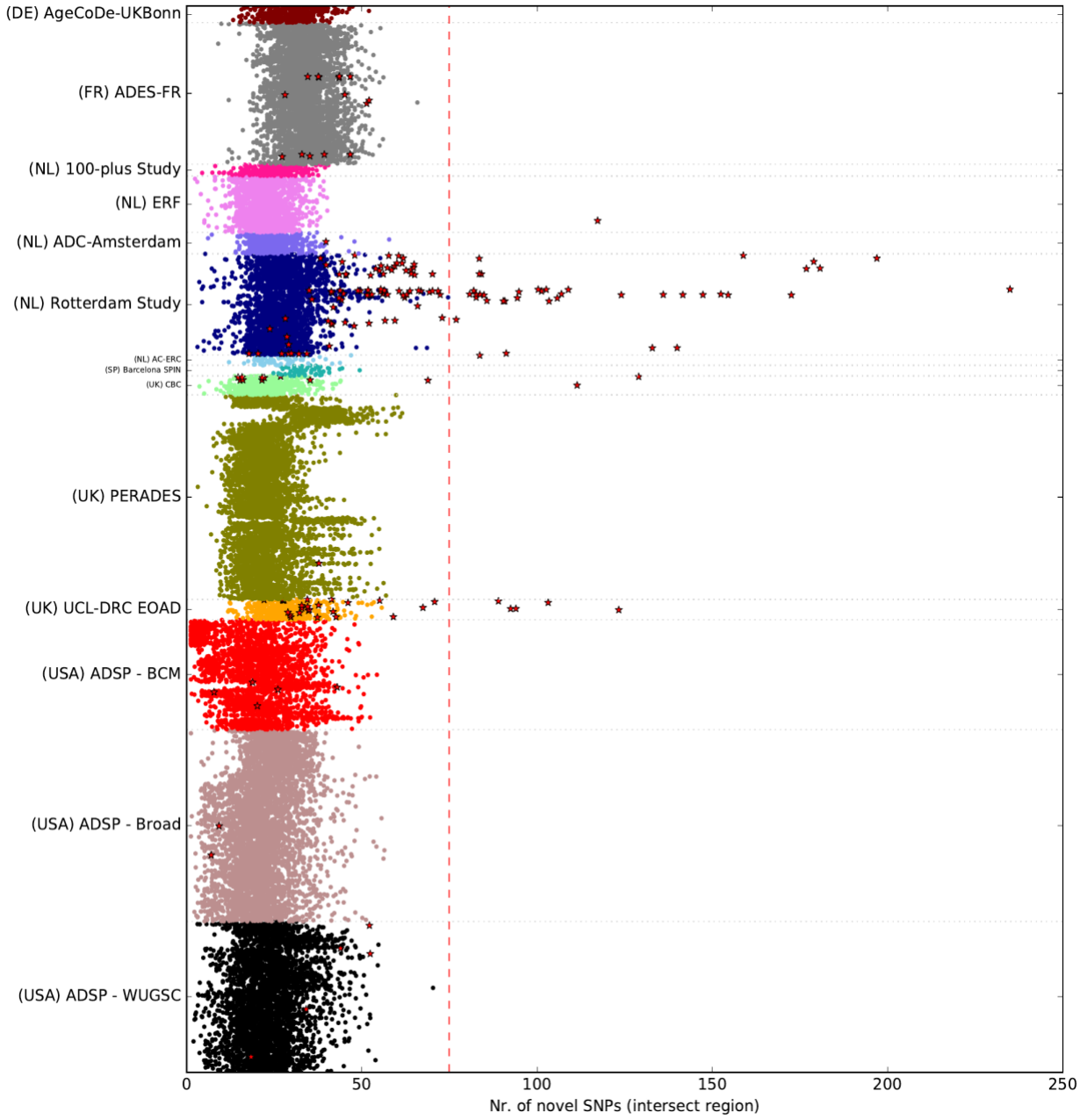
Nr. of novel SNPs in the region representing the union of all capture kits + 100bp padding. Sample QC outliers (step 5-8) are shown as red stars. Variants are classified as novel if they are not present in DBSNP v150. Per geographical region, the comprehensiveness of the annotation of local rare variants in DBSNP might vary.

Figure S10: Number of novel indels per study (union of capture kits)



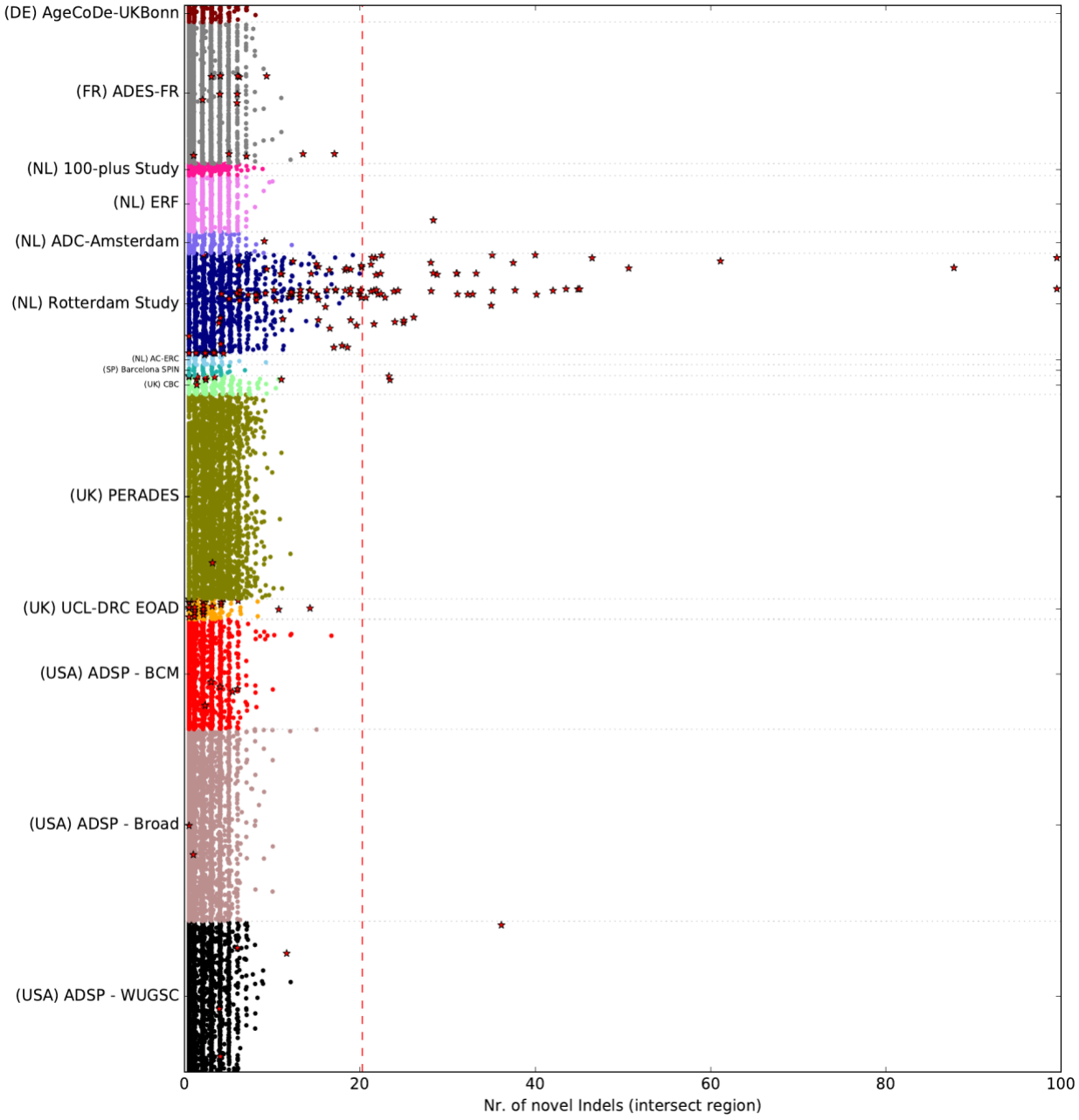
Nr. of novel indels in the region representing the union of all capture kits + 100bp padding. Sample QC outliers (step 5-8) are shown as red stars. Variants are classified as novel if they are not present in DBSNP v150. Per geographical region, the comprehensiveness of the annotation of local rare variants in DBSNP might vary.

Figure S11: Number of novel SNPs (intersection of capture kits)



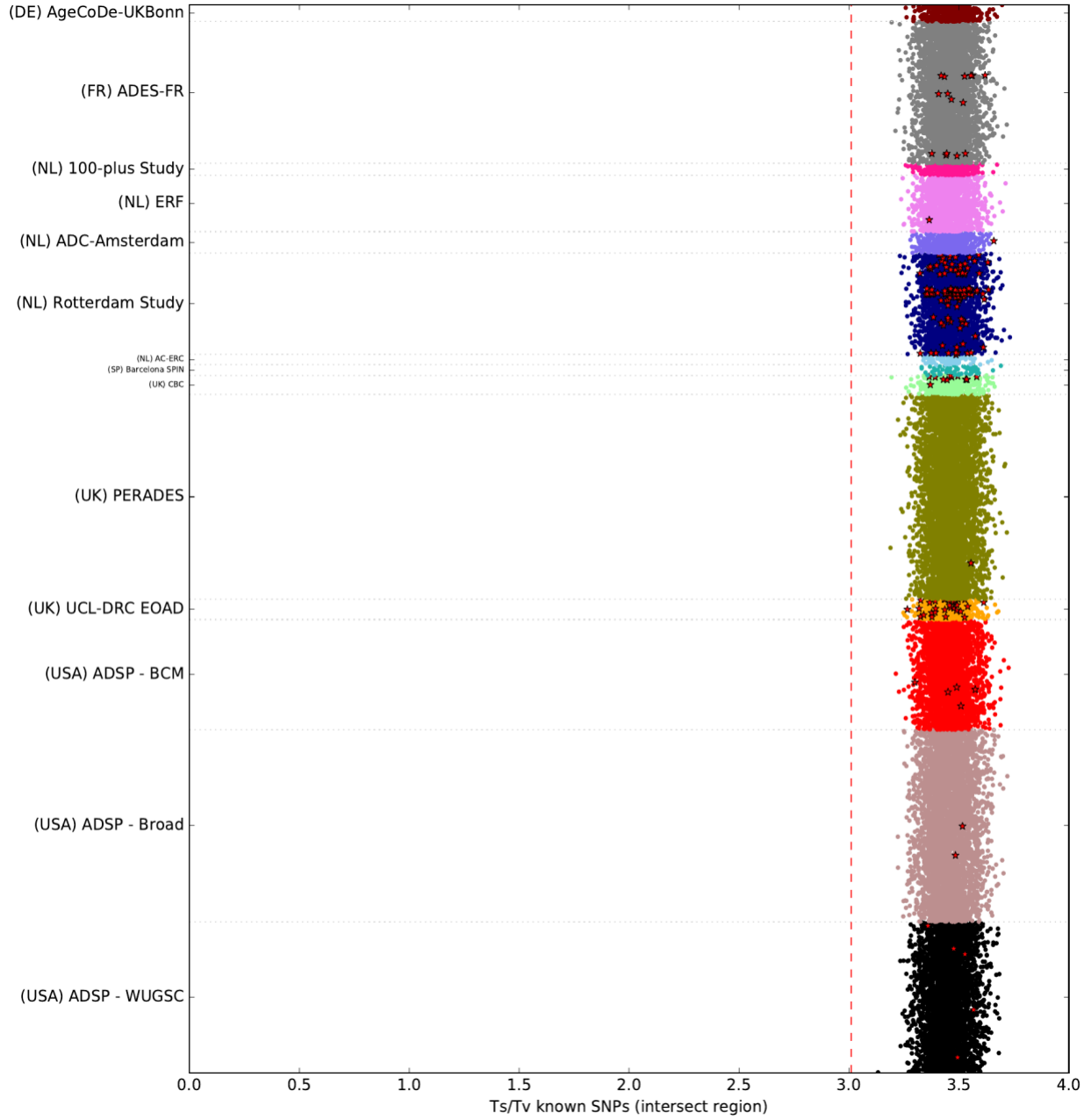
Nr. of novel SNPs in the intersection of all capture kits. Sample QC outliers (step 5-8) are shown as red stars. Variants are classified as novel if they are not present in DBSNP v150. Per geographical region, the comprehensiveness of the annotation of local rare variants in DBSNP might vary.

Figure S12: Number of novel indels (intersection of capture kits)



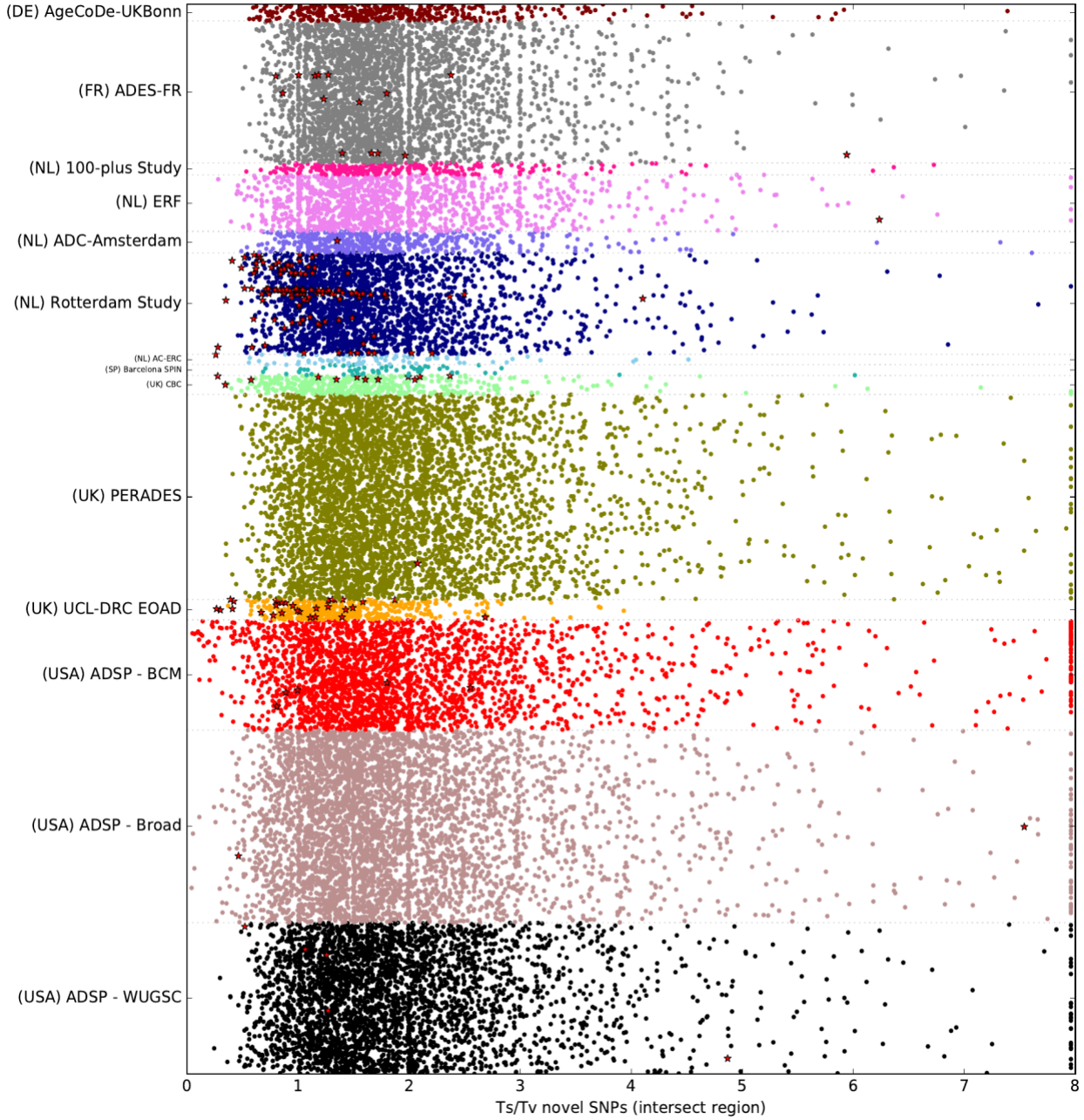
Nr. of novel indels in the intersection of all capture kits. Sample QC outliers (step 5-8) are shown as red stars. Variants are classified as novel if they are not present in DBSNP v150. Per geographical region, the comprehensiveness of the annotation of local rare variants in DBSNP might vary.

Figure S13: Ts/Tv ratio known variants (intersection capture kits)



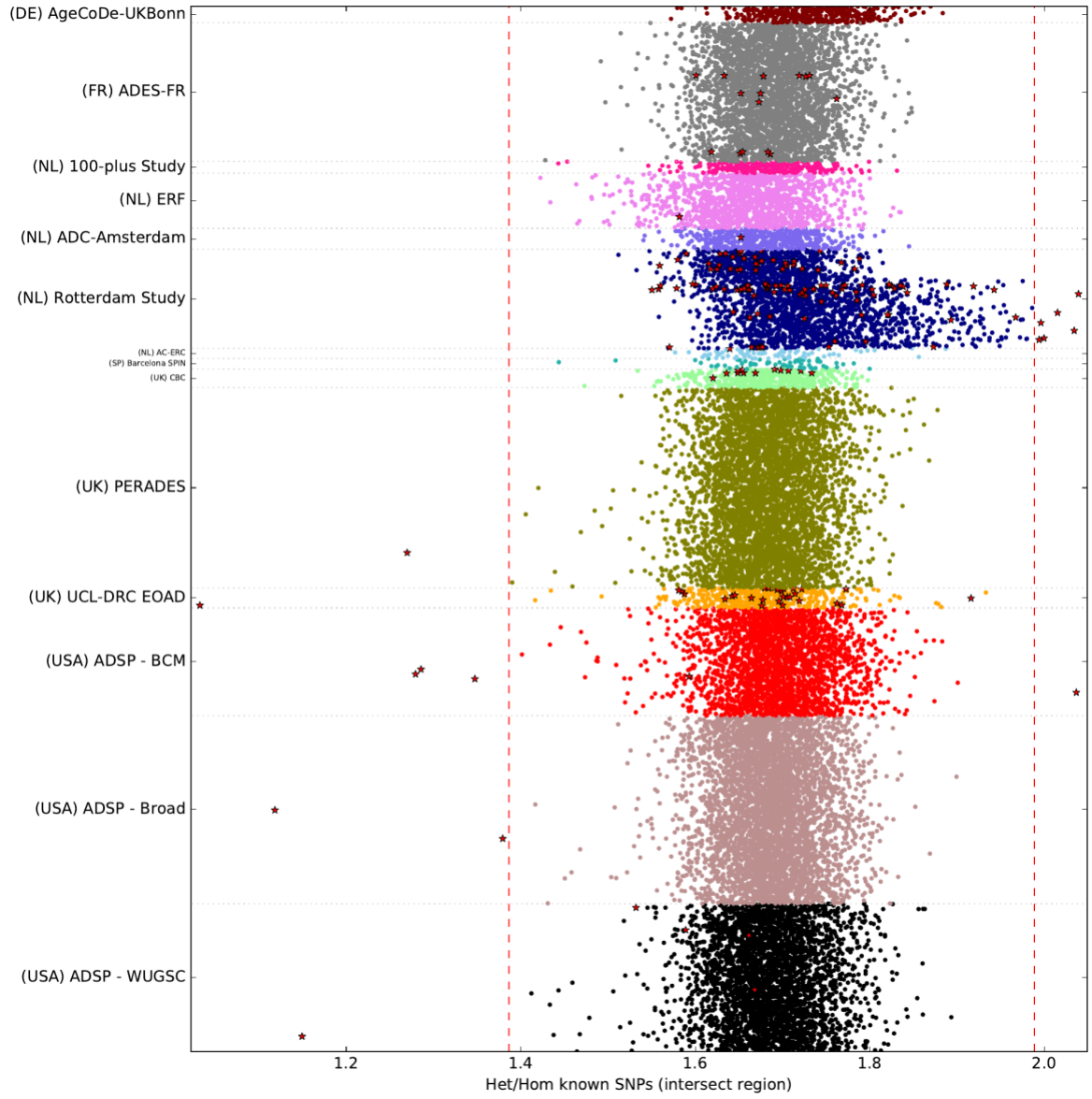
Ts/Tv of known variants in the region covered by all capture kits. Sample QC outliers (step 5-8) are shown as red stars. Variants are classified as known if they are present in DBSNP v150.

Figure S14: Ts/Tv ratio novel variants (intersection of capture kits)



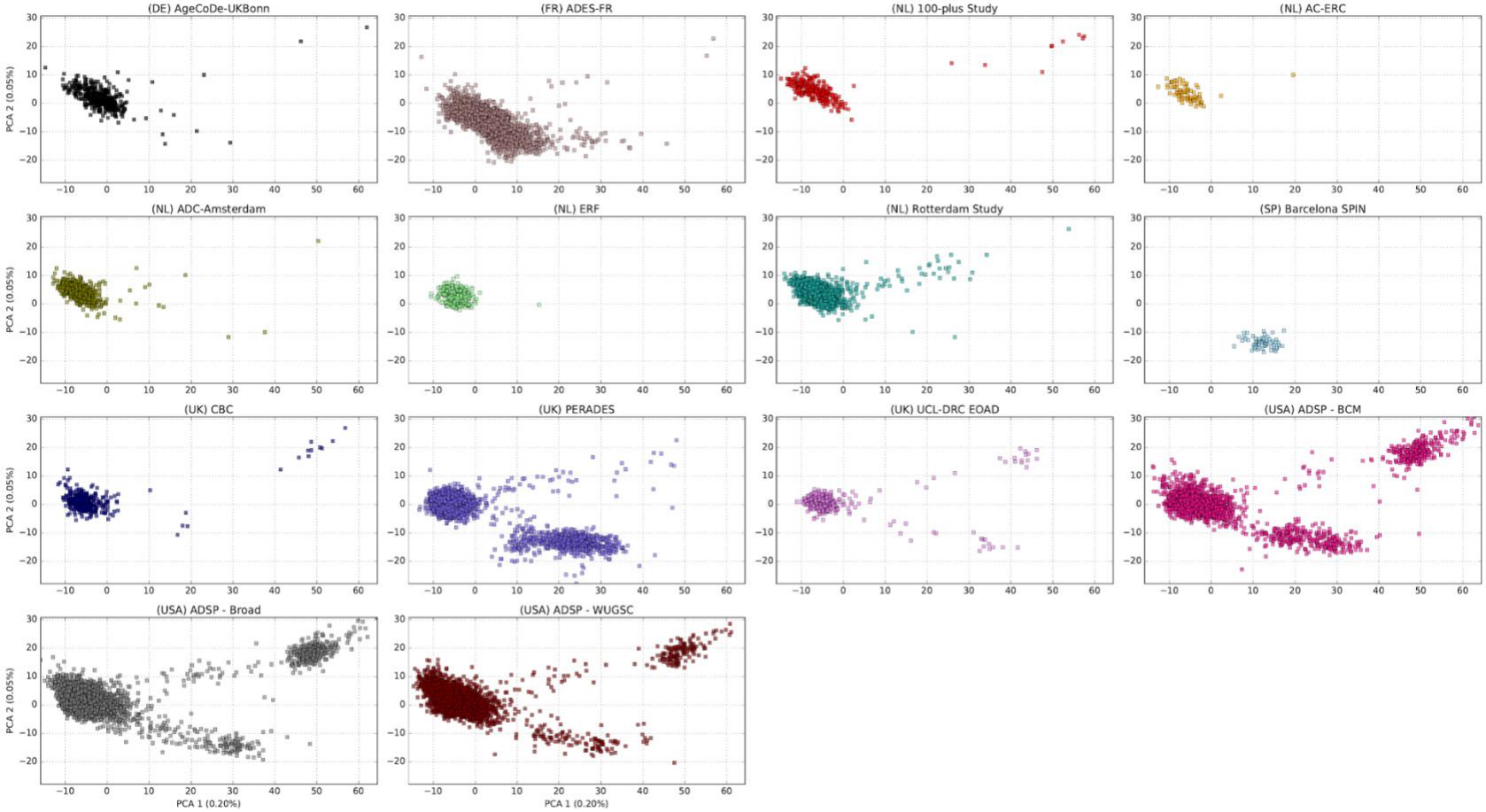
Ts/Tv of novel variants in the region covered by all capture kits. Sample QC outliers (step 5-8) are shown as red stars. The distribution is wide due to a low number of novel SNPs per sample (Figure S11). Ts/Tv values are for plotting purposes maximized at 8. Variants are classified as novel if they are not present in DBSNP v150.

Figure S15: Het/Hom ratio known variants (intersection capture kits)



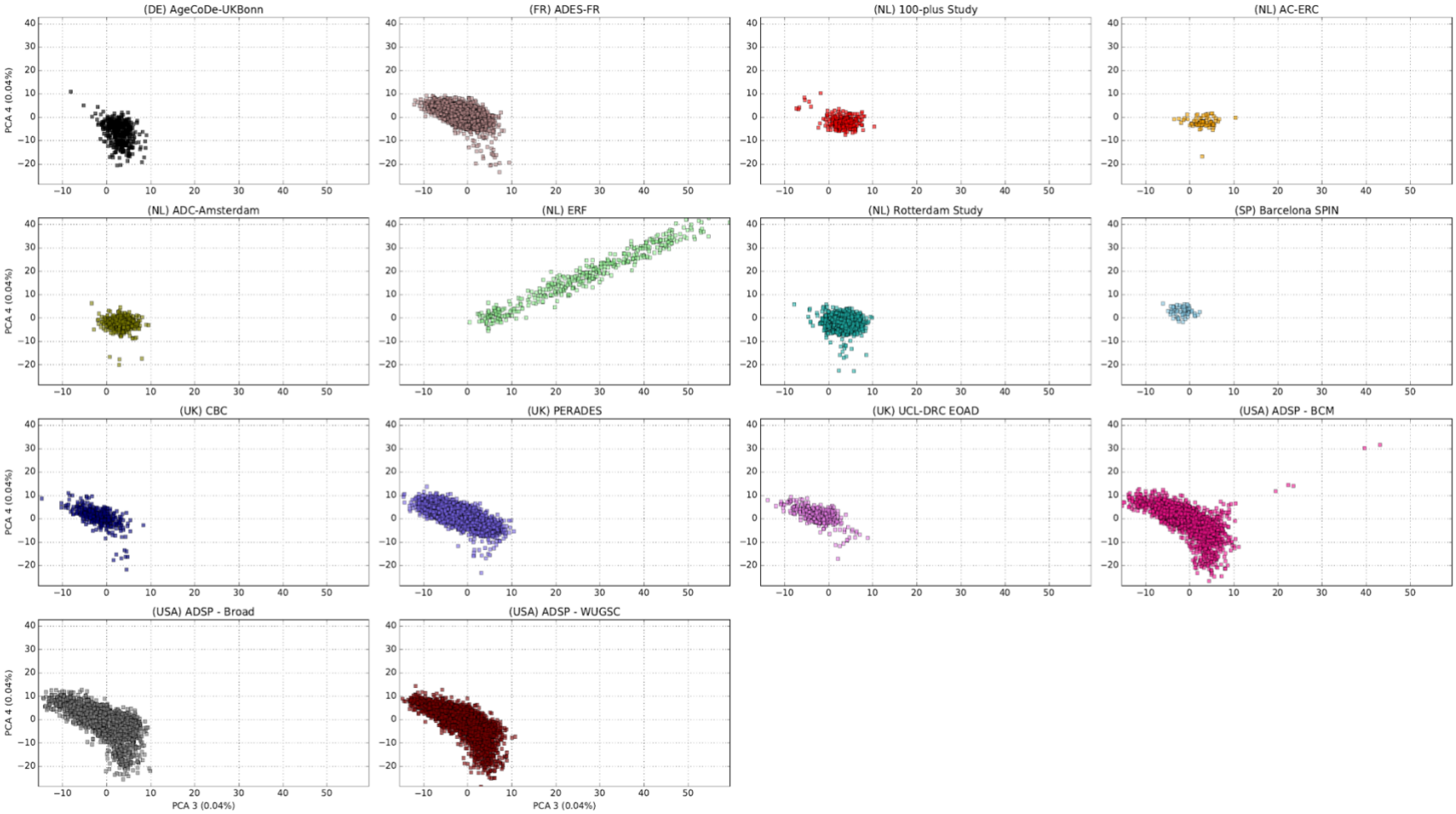
Het/Hom of known variants in the region covered by all capture kits. Sample QC outliers (step 5-8) are shown as red stars. Variants are classified as known if they are present in DBSNP v150. Low het/hom ratios can be an indication of inbreeding, while high het/hom ratios can be an indication of contamination. The problem of contamination is mostly limited to more common variants, and not the rare variants that are the focus of this study

Figure S16: First two PCA components per study, after sample QC.



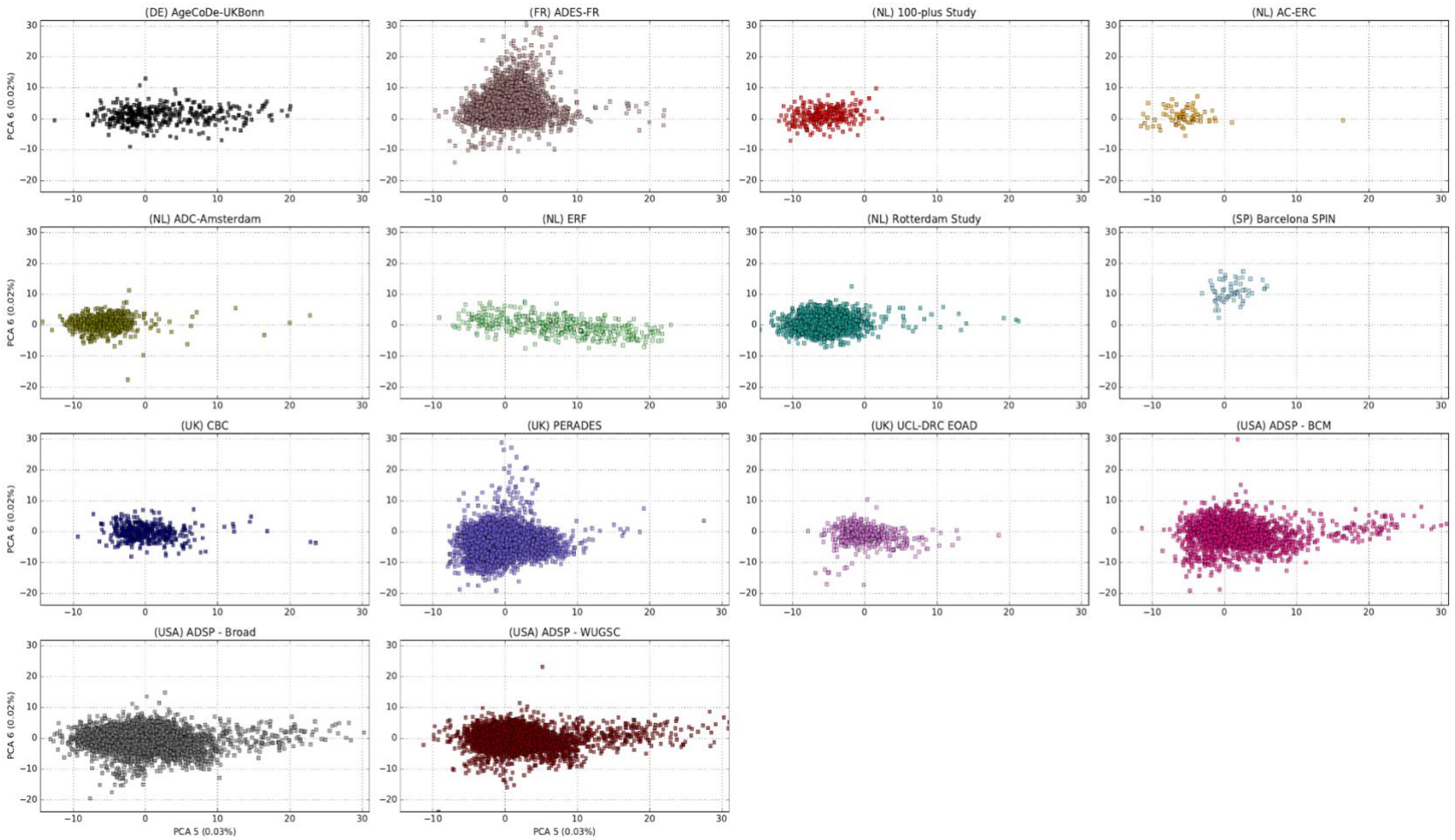
All analysis are corrected for the first 6 PCA components.

Figure S17: Third and fourth PCA components per study, after sample QC.



All analysis are corrected for the first 6 PCA components.

Figure S18: Fifth and sixth PCA components per study, after sample QC.



All analyses are corrected for the first 6 PCA components.

Supplemental methods

We analyzed a total sample of 25,982 individuals sequenced with Illumina technology. Of these, 15,088 individuals were collected as part of the Alzheimer Disease European Sequencing consortium (ADES), comprising 11 studies from Germany, France, The Netherlands, Spain, and the United Kingdom. All studies were approved by the ethics committees of respective institutes, and all participants provided informed consent for study participation. These samples were combined with 11,365 samples from the Alzheimer's Disease Sequencing Project (ADSP), which were described previously[2] (**Table S1**).

Across all studies, AD cases were defined according to NIAA criteria[3] for possible or probable AD or according to NINCDS-ADRDA criteria[4] depending on the date of diagnosis. When possible, supportive evidence for an AD pathophysiological process was sought (including CSF biomarkers) or the diagnosis was confirmed by neuropathological examination (**Table S1**). Cases were annotated with the age at onset or age at diagnosis (2014 samples), otherwise, samples were classified as late onset AD (366 samples). Controls were not diagnosed with AD. All contributing datasets were sequenced using a paired-end Illumina platform, but different exome capture kits were used, and a subset of the sample was sequenced using whole genome sequencing (**Figure S1, Table S2**).

Sample descriptions

ADES-FR

The ADES-FR project combines WES and WGS data from AD cases and controls from France[5]. Part of the patients are from the CNRMAJ-Rouen center (n=921) and patient ascertainment is described in detail in Nicolas et al.[6] including an update of the inclusions by the French National network CNR-MAJ (national reference center for young Alzheimer patients). Briefly, unrelated cases with early-onset AD (age at onset ≤65 years) from France were recruited among patients who fulfilled the NIAA criteria[3]. The clinical

examination included personal medical and family history assessment, neurologic examination, neuropsychological assessment, and neuroimaging. In addition, cerebrospinal fluid (CSF) biomarkers indicative of AD were available for 67% of the cases. Cases with CSF biomarkers not consistent with AD diagnostics were excluded. A positive family history (i.e., at least a secondary case among first- or second-degree relatives, whatever the age of onset) was present in 45% of cases. Patients were either screened by Sanger sequencing and QMPSF for pathogenic variants in *APP*, *PSEN1* or *PSEN2* prior to WES or by the interpretation of WES data or both. Carriers of pathogenic variants were not included for WES or were secondarily excluded following WES analysis so that none of the CNRMAJ-Rouen patients included in this work prior to shared analyses is a carrier of a pathogenic variant in *APP*, *PSEN1*, *PSEN2* as well as in a list of Mendelian dementia causative genes[7]. In addition, some controls were recruited directly from the CNRMAJ (n=30). A large part of the samples was from the European Alzheimer's Disease Initiative (EADI) dataset[8]. This study combined clinical prevalent and incident cases of AD (n=1,121) (i) from Lille cross-sectional studies and (ii) from the Three-City (3C) study, a population-based, prospective study with 12-years of follow-up[9]. Diagnoses were established according to the DSM-III-R and NINCDS-ADRDA criteria[4]. Controls were selected among the 3C individuals not diagnosed with dementia after a 12-year follow-up (n=670). In addition, other controls were obtained from the FREX consortium. These controls (n=576) were specifically designed from 6 French cities with the aim of studying and establishing the French population genetic structure of rare variants. Overall, the ADES-FR samples includes 2,042 AD cases (1,088 EOAD and 954 LOAD) and 1,276 controls. All patients and controls provided informed written consent for genetic analyses in a clinical and/or in a research setting, according to each study. In addition, the ethics committee of the Rouen University Hospital approved the use of retrospective data in the context of the ADES-FR project and with other ADES European and American partners (CERNI notifications 2017-015 and 2019-055).

AgeCoDe-UKBonn

The AgeCoDe-UKBonn sample was derived from the following two sources, the German study on Aging, Cognition, and Dementia in primary care patients (AgeCoDe, n=294) and the interdisciplinary Memory Clinic at the University Hospital of Bonn (UKBonn, n=100). **The German study on Aging, Cognition, and Dementia**: The AgeCoDe study is a multicenter prospective general practice-based cohort study since 2001, including community dwelling elderly aged 75 years or older that were recruited at six study sites (Bonn, Düsseldorf, Hamburg, Leipzig, Mannheim, and Munich). The AgeCoDe study was approved by the local ethics committees of the Universities of Bonn, Hamburg, Düsseldorf, Heidelberg/Mannheim, Leipzig, and Munich. Before participation written informed consents were collected from all subjects. The AgeCoDe study aims to identify risk factors and predictors of cognitive decline and dementia[10,11]. Participants were recruited from general practitioner (GP) registries. Inclusion criteria were an age of 75 and older, absence of dementia, one or more visits to the GP in the past year, no hearing or vision impairments and German as a native language. Exclusion criteria were only home-based GP consultations, severe illness with a fatal outcome within 3 months and a language barrier. The baseline assessment including 3,327 subjects was completed between 2002 and 2003. After the baseline assessment 70 subjects were excluded due to presence of dementia after standard assessment and 40 subjects were excluded with an age below 75 years. Participants were interviewed for follow up every 18 months. All assessments are performed at the participant's home by a trained study psychologist or physician. At all visits, assessment includes the Structured Interview for Diagnosis of Dementia of Alzheimer type, Multi-infarct Dementia, and Dementia of other etiology according to DSM-IV and ICD-10 (SIDAM)[12]. The SIDAM comprises: (1) a 55-item neuropsychological test battery, including all 30 items of the MMSE and assessment of several cognitive domains (orientation, verbal and visual memory, intellectual abilities, verbal abilities/ calculation, visual–spatial constructional abilities, aphasia/ apraxia); (2) a 14-item scale for the assessment of the activities of daily living (SIDAM-ADL-Scale); and (3) the Hachinski Rosen-Scale. Dementia was diagnosed according to DSM-IV criteria. AgeCoDe provided DNA from 294 persons who progressed to late onset AD dementia at any follow up.

UKBonn: The interdisciplinary Memory Clinic of the Department of Psychiatry and Department of Neurology at the University Hospital in Bonn provided early-onset AD patients (n=100). Diagnoses were assigned according the NINCDS/ADRDA criteria[4] and on the basis of clinical history, physical examination, neuropsychological testing (using the CERAD neuropsychological battery, including the MMSE), laboratory assessments, and brain imaging.

Barcelona- SPIN

Neuropathological samples were obtained from the Neurological Tissue Bank of the Biobanc-HospitalClinic-IDIBAPS, and disease evaluation was performed according to international consensus criteria. Clinical samples were recruited from the multimodal Sant Pau Initiative on Neurodegeneration (SPIN) cohort (https://santpaumemoryunit.com/our-research/spin-cohort/) [13], and were evaluated at the Memory Unit at Hospital de Sant Pau (Barcelona). The repository includes clinical data of more than 6,000 participants, >2900 plasma samples, genetic material (DNA and RNA) of >3,200 and >400 subjects, respectively, and >2,000 CSF samples. All controls had normal cognitive scores in the formal neuropsychological evaluation and normal core CSF AD biomarkers, based on previously published cut-offs[14]. AD patients fulfilled clinical criteria of "probable AD dementia with evidence of the AD pathophysiological process"[3] and therefore had abnormal core AD biomarkers (low Aβ1–42 and high t-Tau or p-Tau) in the CSF. The original protocol and the subsequent amendments were approved by our local Ethics Committee at the Sant Pau Research Institute as well as the Committee of the Neurological Tissue Bank. The SPIN cohort is based on blinded enrollment and only clinically relevant biomarker results are disclosed.

100-plus Study

The 100-plus Study, is a prospective cohort study of cognitively healthy centenarians that associated with the Alzheimer Center at the Amsterdam University Medical Center. Detailed participant recruitment and procedures were described previously[15]. Trained researchers visited the centenarians at their home residence annually, where they were

subjected to questionnaires regarding demographics, lifestyle, medical history, physical well-being and objective measurements of cognitive and physical functions. Cognitive function is tested by an extensive neuropsychological testing battery. Approximately 30% of the centenarians agreed to post mortem brain donation. For the current study, DNA samples 276 centenarians were included who completed at least one neuropsychological test at baseline, and exome sequencing from 254 centenarians passed QC (removal was mostly due to kinship). The 190 centenarians who scored >22 on the MMSE were regarded as controls, while 64 centenarians who scored <=22 were regarded as cases[16]. The Medical Ethics Committee of the Amsterdam UMC approved this study and informed consent was obtained from all participants. The study has been conducted in accordance with the declaration of Helsinki. All brain donors signed informed consent for brain donation.

ERF

The Erasmus Rucphen Family (ERF) study is a family-based cohort study that is embedded in the Genetic Research in Isolated Populations (GRIP) program in the South West of the Netherlands. The aim of this program was to identify genetic risk factors in the development of complex disorders. For the ERF study, 22 families that had at least five children baptized in the community church between 1850-1900 were identified with the help of genealogical records. All living descendants of these couples and their spouses were invited to take part in the study. Data collection started in June 2002 and was finished in February 2005.

Rotterdam Study

The Rotterdam Study is an ongoing prospective population-based cohort study, focused on chronic disabling conditions of the elderly 1, of which a random subset was exome sequenced. Participants were screened for dementia at baseline and at follow-up examinations using the Mini-Mental State Examination (MMSE) and the Geriatric Mental Schedule (GMS) organic level 2,3. Screen-positives (MMSE <26 or GMS organic level >0) underwent extensive examination 4. Finally, individuals were diagnosed in

accordance with standard criteria for dementia (Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised (DSM-III-R)) and Alzheimer's disease, NINCDS-ADRDA 5. Follow-up for incident dementia was complete until January 1st, 2014. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study). All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

AC-EMC

The Alzheimer Center Erasmus MC cohort (AC-EMC) includes patient referred to the Department of Neurology of the Erasmus Medical Center (Rotterdam, the Netherlands). DNA samples from 81 patients with probable AD were included in the current study. The average age at onset was 59 years (range 41-72). The majority of patients (64%) had a positive family history, defined as at least one first degree relative with dementia. All patients underwent clinical examination, neuropsychological assessment, neuroimaging, and if indicated, a lumbar puncture. The diagnosis was established according to the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria for AD[3].The study was approved by the Medical Ethical Committee of the Erasmus Medical Center, and written informed consent was obtained from all participants or their legal representatives.

ADC-Amsterdam

The ADC-Amsterdam cohort includes patients who visit the memory clinic of the Alzheimer Center at the Amsterdam University Medical Center, The Netherlands, and was described previously[17]. DNA samples from 518 patients with probable AD cases were included in the current study. Individuals in this cohort were extensively characterized to reduce the chance of misdiagnosis. Patients underwent an extensive standardized dementia assessment, including medical history, informant-based history, a physical examination, routine blood and CSF laboratory tests, neuropsychological testing,

electroencephalogram (EEG) and MRI of the brain. The diagnosis of probable AD was based on the clinical criteria formulated by the National Institute of Neurological and Communicative Disorders and Stroke—Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) and based on National Institute of Aging–Alzheimer association (NIA-AA). Clinical diagnosis is made in consensus-based, multidisciplinary meetings. All patients gave informed consent for biobanking and for the use of their clinical data for research purposes. Selection for whole exome sequencing was based on an early age-of-onset (age at diagnosis <70 years) and available CSF biomarkers.

PERADES

The PERADES sample (Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease) comprises individuals with Alzheimer's disease (AD) and healthy controls recruited across UK, Italy and Spain. The majority of the individuals are from the UK (n=4095 with samples recruited in Cardiff: n=2405), while the rest (n=841) were recruited in Spain and Italy. More specifically the recruitment centres were: MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK; Institute of Psychiatry, London, UK; University of Cambridge, Cambridge, UK; University of Southampton, Southampton, UK; University of Nottingham, Nottingham, UK; Catholic University of Rome, Rome, Italy; Santa Lucia Foundation, Rome, Italy; Instituto di Neurologia Policlinico Universitario, Rome, Italy; University of Milan, Milan, Italy; Laboratory of Gene Therapy, San Giovanni Rotondo, Italy; University of Perugia, Perugia, Italy; University of Cantabria and IDIVAL, Santander, Spain and the Regional Neurogenetic Centre (CRN), ASP Catanzaro, Lamezia Terme, Italy. The collection of the samples within the MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University was through national recruitment through multiple channels, including specialist NHS services and clinics, research registers and Join Dementia Research (JDR) platform. The participants were assessed at home or in research clinics along with an informant, usually a spouse, family member or close friend, who provided information about and on behalf of the individual with dementia. Established measures were used to ascertain the disease severity: Bristol activities of daily living (BADL), Clinical Dementia Rating scale (CDR), Neuropsychiatric Inventory (NPI) and Global Deterioration Scale

(GIDS). Individuals with dementia completed the Addenbrooke's Cognitive Examination (ACE-r), Geriatric Depression Scale (GeDS) and National Adult Reading Test (NART) too. Control participants were recruited from GP surgeries and by means of self-referral (including existing studies and Joint Dementia Research platform). For all other recruitment, all AD cases met criteria for either probable (NINCDS-ADRDA, DSM-IV) or definite (CERAD) AD. All elderly controls were screened for dementia using the Mini Mental State Examination (MMSE) or ADAS-cog, were determined to be free from dementia at neuropathological examination or had a Braak score of 2.5 or lower. Control samples were chosen to match case samples for age, gender, ethnicity and country of origin. Informed consent was obtained for all study participants, and the relevant independent ethical committees approved study protocols. The whole exome sequencing (WES) was performed in-house at the MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University. With the Nextera technology (Nextera Rapid Capture Exome v1.2), DNA was simultaneously fragmented and tagged with sequencing adapters in a single step. The enriched libraries were sequenced using the Illumina HiSeq 4000 (Illumina, USA) as paired-end 75 base reads according to manufacturer's protocols.

CBC: Control Brain Consortium

The Control Brain Consortium consists of 478 was previously described[18]. Whole-exome sequencing in 478 samples derived from several brain banks in the United Kingdom and the United States of America. Samples were included when subjects were, at death, over 60 years of age, had no signs of neurological disease and were subjected to a neuropathological examination, which revealed no evidence of neurodegeneration. The data was made publicly available at www.alzforum.org/exomes/hex.

UCL-DRC EOAD

University College London Dementia Research Centre (UCL-DRC) early-onset Alzheimer's disease cohort included patients seen at the Cognitive Disorders Clinics at

The National Hospital for Neurology and Neurosurgery (Queen Square), or affiliated hospitals. Individuals were assessed clinically and diagnosed as having probable Alzheimer's disease based on contemporary clinical criteria in use at the time, including imaging and neuropsychological testing where appropriate. All individuals consented for genetic testing and had causative mutations for Alzheimer's disease (*PSEN1, PSEN2, APP*) and prion disease (*PRNP*) excluded prior to entry into this study.

ADSP

Cases and controls were selected from over 30,000 non-Hispanic Caucasian subjects from multiple cohorts described in detail elsewhere[19]. All controls were greater than 60 years and were cognitively normal based on direct assessment. All cases met NINCDS-ADRDA criteria for possible, probably, or definite Alzheimer's disease. All cases had a documented age-at-onset, and for those with pathologically conformed AD, an age-at death. APOE genotypes were available for all. Cases were selected to have a minimal AD risk based on sex, age and APOE genotype. Controls were selected as those with the least probability of converting to AD by age 85. Controls were older (86.1 years, SD = 5.2) that cases (76.0 years, SD = 9.2). The selection criteria and the rationale for study design are described elsewhere[20]. We selected 5,096 cases and 4c,965 controls for exome sequencing by this protocol. In addition, we selected 682 additional cases from multiplex families with a strong AD family history. Because some of these subjects were Caribbean Hispanics, we also sequenced 171 cognitively normal Caribbean Hispanic controls.

Alignment and variant calling

Raw sequencing data from all studies were collected on a single site (Cartesius Supercomputer provided by SURF, in the Netherlands), and processed with a uniform pipeline. Reads were extracted from FastQ, BAM, CRAM or SRA files. For each lane/read group separately, paired reads were converted to SAM format using FastQToSam or picard RevertSam (Picard Tools version 2.10.5[21]), processed with Picard MarkIlluminaAdapters and subsequently transformed to interleaved fastq format with

Picard SamToFastq (while setting marked adapter regions to base quality 2). Next, reads were aligned to the human reference genome (build 37 with decoys) using the BWA MEM algorithm (BWA version 0.7.15-r1140)[22]. Alignments were processed with Samblaster (version 0.1.24) to add mate tags[23]. Read group alignments were then merged and duplicate reads were marked using Picard MarkDuplicates. We found that the presence of novel Indels and novel SNPs in certain samples correlated with the presence of larger amounts of soft-clipped reads, indicative of the presence of chimeric DNA fragments. Each sample for which the percentage of soft-clipped base alignments exceeded 0.5% was therefore processed with a custom tool (see below) which identified and removed parts of reads that were likely of chimeric origin. This tool was executed after the Picard MarkDuplicates step. Then, reads were sorted to chromosome order by samtools sort (version 1.8)[24]. We estimated contamination percentages using VerifyBamID2[25], retrieved 4 September 2018), while correcting for the 2 PCA components (default), and excluding common SNPs (allele frequency ≥0.01) present in the 1000-genomes dataset (phase3, version 5b)[26]. Base quality scores were recalibrated using GATK BQSR (version 3.8-1)[27]. on the sample capture kit region + 100bp padding. Known indels were obtained from the Mills and 1000G gold standard indels in the GATK resource kit[27]. Known SNPs were obtained from dbSNP (version 150) and gnomAD (version 2.0.2)[28]. Subsequently, variants were called on the sample capture kit region + 100bp padding using the HaplotypeCaller[27], while using the '-contamination' correction option, with the estimated contamination percentages. Ploidy was set to 1 for chromosome Y, and 2 for the other chromosomes, minPruning was set to 2, and the new quality model (--newqual) was used. Results were exported as gVCF format. Finally, gVCFs were combined per study in batches with a maximum size of 500 samples using GATK CombineGVCF. Then, variants were called using GATK GenotypeGVCF[29], using the new quality model and setting max-alternate-alleles to 20. Variants were then annotated with GATK variant score recalibrator (VQSR) using allele specific annotations, while for all other options the best practices were followed.

Chimeric read declipping

Chimeric fragments consist of multiple genomic sequences, joined together into one sequence. Sequencing of such fragments can result in reads that do not entirely align to the genome, and/or align at multiple locations. This results in so-called ‘soft-clipped’ alignments, where parts of the read sequence are not aligned. These soft-clipped regions cause issues for the variant caller, as it uses not just the aligned part of the reads, but also the unaligned soft-clipped regions during local reassembly and variant calling. The reason for this is that these clipped sequences can be an indication of an insertion variant. In case these clipped regions are caused by chimera’s, this is however not a correct strategy, and can cause false variant calls. To prevent their effect on variant calling, we i) estimate the extent of the chimera problem by quantifying the number of soft-clipped alignments, and ii) remove these soft-clipped sections for affected samples if they are (likely) caused by chimeras. To do this, the soft-clipped sections are turned into hard-clipped alignments, in which the underlying sequence is removed (the read is shortened), such that the variant caller cannot revive the clipped sequence during variant calling. In the following description, we assume paired end sequencing (in which both ends of the fragment are sequenced, resulting in two reads). We remove the following soft-clipped sequences:

i) One well-known type of artificial chimera occurs when the sequenced fragment is shorter than the read length. Fragments have adapters at the end, used as starting point for sequencing. In these cases, the 3’ end of read 1 will cover the adapter of read 2, and vice versa. Due to this, read 1 and 2 will have overlapping alignments with possibly soft-clipped 3’ends. Such read pairs can be detected based on their overlapping alignments. To remove the adapter sequence, we align the known adapter sequence to determine the clipping point, and hard-clip the identified sequences from there.

ii) A genomic chimera can have a join-point at different sites in the sequence fragment.

— If the chimeric join point occurs between read 1 and 2, or close to the end of read 1 or 2, then read 1 and 2 will (usually) be aligned at a distance from each other. If this

distance is >100kb, or one of the reads is unmapped, we remove the soft-clipped regions at the 3'end of both reads.

— If there are multiple, mostly non-overlapping, alignments for a read at different genomic locations, it is usually an indication that the chimeric join point occurs somewhere in the middle of that read. The overlapping parts of these alignments are pruned (in all alignments for that read). Then, soft-clipped sequences in the alignments that face each other are hard-clipped.
— in the above situation, it frequently occurs also that the fragment is short. The chimeric join point might then be present in both reads. If both reads have multiple alignments, we handle each read as described above.
— if the fragment is short, but not very short, read 1 might have multiple alignments, while read 2 has a soft-clipped 3'end (or vice versa). For example, for genomic region A and B, a chimeric fragment might read AABBB. Read 1 (AABB) might then have multiple alignments, one for the AA and one for the BB section. Read 2 (BBBA) however might have only an alignment for B, but not for A, as the sequence from A is too short to obtain an accurate alignment. The chimeric sequence A in read 2 will therefore be soft-clipped. We detect these situations based on overlapping alignments for fragment B, and hard-clip the soft-clipped 3'end of read 2.
— if the chimeric sequence consists of a very short piece at the 5' end of either read 1 or 2, this part might not be aligned as it is too short. It is in these situations unclear if the sequence has a chimeric origin, as such unaligned pieces can also be caused by indels. We find that in samples affected by chimeras, it is beneficial to remove these soft-clipped 5'ends. While this reduces the coverage of indels, in most cases many fragments still cover the complete indel. Also, differences in coverage between samples occurs commonly in exomes, where the covered regions are highly variable between capture kits, and handling this is part of the downstream pipeline (see posterior probabilities).
— After removal of the soft-clipped regions caused by chimera's, we unalign the alignments that are <= 1bp in length, we transform supplementary alignments to primary alignments if the primary alignment is unaligned, drop unaligned

supplementary alignments, update alignment tags, and validate the read records and cigar strings.

Sample QC

Before sample QC, we performed a pre-variant QC step, to remove bad quality variants (see Variant QC steps for details) that might impact sample quality statistics. In addition, we required that at least 25% of the samples had to have at least read depth 6. Next, sample QC was performed according to the steps described in Figure 1a, which are detailed below.

1. Missingness

We removed samples that had a contamination over 0.75, or a GQ<20 for 60% of the variants in its own exome kit, or a depth < 6 for 65% of the variants in its own exome kit. Additionally, we removed samples for which chromosomes were missing (GQ < 20 for 99% of the variants on a chromosome in the samples exome kit).

2. Contamination

Samples with a contamination percentage > 7.5% were removed.

3. Sex-check

We performed a sex-check, by comparing annotated sex with genetic sex (**Figure S5**). Genetic sex was determined based on the coverage of the sex chromosomes. Coverage was determined using off-target reads. Only coverage in regions outside capture kits (+500 bp padding), outside peaks in coverage called with MACS (version 1.4)[30] and outside segmental duplications (Segmental Dups track downloaded from UCSC which includes the PAR regions[31]. Coverage was determined in 20kb windows, and normalized for GC content using linear regression. Regions of 20kb with more than 100 N bases were discarded. X and Y chromosome coverage was normalized by dividing by the autosome coverage. Thresholds were set empirically, based on the distribution of male and female samples (see supplemental figure).

4. Population outliers

Next, we performed a PCA analysis to identify population outliers. Variants that were in the intersection region of all capture kits, and had a minor allele frequency ≥0.005 and a depth ≥6 for 90% of the variants, were used for this purpose. Variants were pruned with bcftools +prune tool (version 1.8)[24] with max LD set to 0.2 in 500kb windows. Only variants that were also in the 1000 genomes dataset (phase 3, v5b) were kept. PCA was performed on dosages (based on genotype calls for 1000G, and based on genotype probabilities for the study samples). Variant dosages were first normalized, as described[32], based on statistics obtained on the 1000G samples. Then, PCA was performed on the 1000G samples, and all ADES samples were mapped to this PCA space (**Figure S6**). Finally, we removed outliers for each of the first 4 PCA components (**Figure S7, Figure S8**), where outliers were defined as samples that fell outside the range *median(pca_component) ± 8 * mad(pca_component)*, where *mad* is the median absolute deviation and the *pca_component* vector only contains the ADES samples.

5,6: Excess novel SNPs or indels

We calculated and compared the number of novel SNPs and the number of novel indels per study, both in the union of the capture kits (**Figure S9 and Figure S10)** and the intersection of the capture kits (**Figure S11, Figure S12)**. Novel variants were defined as variants that were not present in DBSNP v150. These statistics were calculated based on posterior dosages (described below). Thresholds were set at the *median value + 6 * mad for novel SNPs and +12*mad for novel Indels*.

7. Het/hom and TsTV

Furthermore, we performed a per-sample QC on the following statistics (calculated on the intersection of the capture kits): Ts/Tv ratio of known variants (**Figure S13**), and Ts/Tv ratio of novel variants (**Figure S14),** Het/Hom rate of known SNPs (**Figure S15**). The acceptable range for Het/Hom was set to ±6 * mad. For Ts/tv measures, only a lower limit of -6 * mad was used.

8. IBD analysis

We performed an IBD analysis on the remaining samples using Seekin[33]. We kept variants with a minor allele frequency ≥0.005, and for which at least 90% of the samples had depth >=6. Variants were pruned with bcftools +prune tool (version 1.8), with max LD set to 0.2 in 500kb windows. Only variants that were also in the 1000G dataset were kept. We performed a PCA as described before. Using Seekin (version 1.0), we corrected for these PCA components using the options 'modelAF' and 'getAF', using 4 PCA components. Next, kinship was determined using all variants with the heterogeneous estimator of Seekin[33]. Duplicate samples with inconsistent annotation were removed (inconsistent status, *APOE* genotype, or gender, or more than 2 years difference in age at onset for cases). Otherwise, we kept the sample with the most complete annotation: we preferred samples with age (at onset), and *APOE* genotype over samples without. Also, we preferred whole genome sequenced samples over exomes, and samples with lower missingness over samples with higher missingness. For related samples up to 3rd degree (marked by the threshold of >9.4% shared identity by descent, which is the middle value between the expected value for 3rd-degree (12.5%) and 4th-degree (6.25%)), we preferred (in order) cases over controls, samples with more clinical data (age (at onset), apoe status), WGS samples, and samples with higher coverage.

9. Bad PCR plates

We removed all samples on 3 PCR plates that were enriched with gender mismatches.

10. Removal of Mendelian AD-related variant-carriers

Next, we performed a manual curation of causative variants in a short list of Mendelian dementia genes. We extracted rare variants in the following two gene lists and interpreted them following the American College of Medical Genetics and Genomics and the Association for medical Pathology[34], (i) autosomal dominant AD genes: *APP*, *PSEN1*, *PSEN2* (autosomal dominant AD), *GRN*, *MAPT*, *FUS*, *TARDBP*, *VCP*, (fronto-temporal lobar degeneration spectrum), *NOTCH3* (CADASIL), *PRNP* (Prion diseases); (ii) autosomal recessive genes: *NPC1*, *NPC2* (Niemann-Pick type C disease), *TYROBP*,

TREM2 (*homozygous LOF: Nasu-Hakola disease, 1 carrier)*). Carriers of variants that reached enough evidence to be rated at least as likely pathogenic (class 4) were excluded from the analysis, whatever their disease status. Of note, for autosomal recessive genes, heterozygous carriers were not excluded, only carriers of bi-allelic pathogenic variants were excluded.

11. AD label

We excluded samples for which clinical information was indicative of non-AD dementia (e.g. vascular dementia). In addition, part of the case-control sample included minimal neuropathological information. Among them, we further excluded samples with discordant Braak stages, i.e. cases with stage <2 (n=265) and controls with stage >4 (n=43). Finally, 21,345 samples were available for analysis, constituting 12,652 cases (of which 4060 had early onset AD, onset ≤ 65 years) and 8693 controls.

Variant QC

Throughout an extensive QC, we attempted to find root causes for the presence of false variants. We identified two significant issues that were not handled by the default variant calling pipeline. After removal of samples excluded by the sample QC, variant statistics were recalculated. Then, we performed variant QC as described in (**Figure 1B)**.

1a. Multi allelic variants

First, multi-allelic variants were split into bi-allelic variants, and indels were normalized, using the bcftools norm tool. The tool was modified to also split the phased PGT fields, such that downstream variant merging was possible. Additionally, the splitting of the genotype likelihoods and read counts was modified (PL and AD fields), which is detailed in the next section. We removed bi-allelic variants that had as alternate allele '*' (which reflects overlap with a deletion variant), as well as multi-allelic variants for which the reference allele was lower in frequency than the frequency for at least two alternate alleles.

1b. Variant merging

Variants that were in close vicinity, in cis and always occurred together, were merged into single events, to account for for example nearby frameshifts that cancel each other out. Only indels with ≤10bp distance and snps with ≤2bp distance were considered for merging. We used the read-phasing output of GATK (PID/PGT) fields to determine which variants occurred in-phase.

2. Oxo-G

In some samples novel variants were enriched for G>T and C>A variants, caused by the oxygenation of G bases during sample processing[35]. Using a custom tool (see below), that uses per-sample statistics from Picard CollectSequencingArtifactMetrics, we identified and filtered variants and variant calls that could be attributed to this issue. We removed variants with an average OXO sensitivity > 1.5, or a remaining total dosage after OXO correction ≤0.1.

3. STR/LCR regions

STR and LCR regions were obtained respectively from the simple tandem repeats track by TRF from UCSC, and the LCRs as identified by the mdust program[36]. Variants in these regions were excluded.

4. Allele Balance

The balance between reference and alternate reads (allele balance) was determined both for heterozygous and homozygous calls. Allele balance was calculated based on posterior genotype probabilities (see below). Variants that had an average allele balance < 0.25 or > 0.75 for heterozygous calls, or < 0.9 for homozygous calls were removed.

5. Depth Fraction

The relative depth of heterozygous calls to other calls was determine, based on posterior genotype probabilities (see below). Variants for which the heterozygous depth was < 20% of the depth of other calls were removed.

6. Hardy Weinberg

Hardy-Weinberg scores (all samples and control samples: hw_all and hw_control) were calculated based on posterior genotype probabilities (see below). We removed variants for which the p-value for control samples was $< 5 * 10^{-8}$.

7. VQSR

Variants that were tagged by the variant quality score recalibration method from GATK were removed, for SNPs we removed variants from the VQSR > 99.5% tranche, while for indels we removed variants from the VQSR > 99.0% tranche.

Pre-variant QC versus final variant QC

For the pre-variant QC, which is performed prior to performing the sample QC, we performed all the above steps. Additionally, we removed variants with a missingness rate > 25%. Genotype calls which had a depth < 6 were considered missing. For the final variant QC, the missingness step was not performed, as it is included as part of the variant selection. Compared to the pre-variant QC, the final variant QC had variant batch detection as an additional step.

8. Variant Batch Detection

Finally, we developed a custom tool to remove variants that still presented batch effects that were not explainable by population structure or phenotype effects (see below). On variants identified to have a batch effect, we attempted variant batch correction, by setting batches that caused problems for a certain variant to missing. Afterwards, variants that still had a VBD score > 25, or a VBD score > 15 and MAF < 0.005 were removed from the analysis.

Genotype posterior probabilities

Due to the use of different capture kits and whole genome sequencing (WGS) data, the analysed dataset has highly variable coverage patterns across the samples. Many variants have as a consequence less than 100% coverage across the samples. In burden

testing, a missingness percentage of up to 20% is allowed. This requires an accurate handling of missing genotype calls in variants that contribute to the burden score. In cases of low and absent read coverage, direct calling of the genotype is not possible. Therefore instead, a probabilistic approach is used, in which each genotype is assigned a certain probability.

Genotype likelihoods

The GATK variant caller outputs the likelihood of each sample genotype in the PL field of the VCF. These likelihoods are based on the available sequencing reads for a sample. In case of missing data, each genotype is considered equally likely (i.e. p=⅓ in case of diploid chromosomes for ref/ref, ref/alt and alt/alt genotypes). These likelihoods cannot be used directly in a burden analysis, as by assuming equal likelihoods for each genotype the allele frequency of samples with missing coverage would effectively by 50%, and likely substantially differ from that of samples with coverage.

Posterior probability

This is solved by the use of posterior probabilities. Here the allele frequency in the study sample is used as a prior in assigning genotype probabilities. Using Bayes theorem, posterior genotype probabilities take the following form (assuming a diploid setting): $P(g) = \frac{L(g)\,\psi(g)}{\sum_i^G L(i)*\psi(i)}$, where *P(g)* is the posterior probability for genotype *g*, with *g* encoded as 0,1 or 2 for respectively the reference, heterozygous and homozygous alternate genotype. *L(g)* is the genotype likelihood as given by the variant caller. The genotype frequency $\psi(g) = \frac{2}{(2-g)!g!}\omega^g(1-\omega)^{2-g}$is derived from the allele frequency ω, assuming Hardy-Weinberg equilibrium. Notably, the allele frequency ω needs to be derived from the study sample, such that ω matches the allele frequency in samples with coverage, thereby preventing biases. A difficulty is that accurate estimation of this allele frequency requires posterior genotype probabilities. Here we follow the approach previously described by Li et al[37] using an EM-algorithm in which iteratively posterior probabilities and the allele frequency are estimated, until convergence (maximum difference in allele

frequency between iterations is $1e^{-7}$) is reached. Finally, posterior dosages in the diploid case were calculated as $d = P(1) + 2\,P(2)$.

Multi-allelic variants

As described in the previous section, variants with multiple alleles are split into bi-allelic variants prior to analysis. For this, the bcftools norm tool is used. However, splitting of the genotype likelihood was adapted from the default approach in bcftools. The standard REF/ALT interpretation of the resulting biallelic likelihoods was considered problematic for the analysis, as often the alleles would be neither REF nor ALT. Genotype probabilities would then not sum to 1. We adapted therefore to a NON_ALT/ALT interpretation of bi-allelic variants. Specifically, this meant that genotype likelihoods were converted to probabilities, and then summed to obtain the NON_ALT/NON_ALT, NON_ALT/ALT and ALT/ALT genotype probabilities (separately for each ALT in the multi-allelic variant to create multiple bi-allelic variants). Notably, in the absence of coverage, the variant caller considers each multi-allelic genotype equally likely. In this situation, the NON_ALT/NON_ALT genotype becomes the most likely genotype, as it sums more genotypes. As this causes biases, we correct for this, using an additional prior equal to *1 / (#summed multi-allelic genotypes)* for each bi-allelic genotype. Next to the genotype likelihood, the read count field (AD field) was also modified to follow the above described NON_ALT/ALT interpretation. To that end, read counts that contributed to the NON_ALT/NON_ALT and NON_ALT/ALT genotypes were summed during variant splitting.

Posterior sample QC-measures

Standard sample QC measures, when calculated on variant calls, are affected by samples with low or missing coverage. To prevent that, these measures were instead based on genotype posterior probabilities:

- ***Nr. of indels/SNPs:*** Determined by summing (across all samples) posterior dosages.
- ***Ts/Tv ratio:*** Determined by summing posterior dosages of transition variants and dividing them by the summer posterior dosages of transversion variants

— ***Het/Hom ratio:*** Determined by summing (across all samples) the posterior genotype probability of the heterozygous genotype, and dividing it by the summed posterior genotype probability of the homozygous genotype.

Posterior variant QC-measures

— ***Heterozygous allele balance:*** Defined as $\frac{\sum_i^N P_i(1)\, r_{ref}}{\sum_i^N P_i(1)\,(r_{ref}+r_{alt})}$, where $P_i(1)$ is the posterior genotype probability for the heterozygous genotype for sample *i*, *N* is the number of samples, and r_{ref} and r_{alt} are the number of reads carrying the reference or alternate genotype.

— ***Homozygous allele balance:*** Defined as $\frac{\sum_i^N P_i(2)\, r_{alt}}{\sum_i^N P_i(2)\,(r_{ref}+r_{alt})}$, where $P_i(2)$ is the posterior genotype probability of the homozygous genotype for sample *i*.

— ***Heterozygous depth ratio:*** Defined as $\frac{\frac{\sum_i^N P_i(1)\,(r_{ref}+r_{alt})}{\sum_i^N P_i(1)}}{(r_{ref}+r_{alt})/N}$.

— ***Hardy-Weinberg equilibrium:*** Posterior genotype probabilities assume Hardy-Weinberg equilibrium (HWE), thereby biasing variants with high rates of missingness towards HWE. Hardy-Weinberg equilibrium is therefore tested on non-probabilistic genotype calls, after filtering out samples with a read coverage < 6.

Oxo-G variant call filtering

During sample preparation, oxidation of G-nucleotides can lead to the generation of 8-oxoguanine lesions in DNA. These lesions lead to false positive G-T variants, and, dependent on the protocol step in which the oxidation occurs, also false positive C-A variants[35]. While this is primarily an issue for somatic variant calling, it also impacts germline rare-variant calls, in particular in exomes where coverage is variable. In modern protocols, these effects have mostly been mitigated, however, in older samples these false positive mutations can be a significant source of errors. Next to oxoG errors, similar problems are known to occur in DNA obtained from formalin-fixed samples. In these samples, deamination can occur, converting cytosine to uracil (C>U), thereby creating false positive C->T (and G->A) mutations. While the approach below handles these types

of errors as well, this problem was not encountered in a significant manner in the dataset. A modern variant caller such as GATK determines nucleotide-specific base error rates based on a comparison of the sequenced reads to the genome (in the case of GATK through base quality score recalibration (BQSR)). In GATK, this error rate is modelled on the observed nucleotide in the read (e.g. in case of a G->T mutation a T for reads aligned to the positive strand and an A for reads aligned on the negative strand). Although G-oxidation will lead to a somewhat higher base error rates in T and A nucleotides, the variant caller does not recognize that these errors occur mainly when the genomic reference contains respectively a G (or C in case of C->A mutations). This leads to underestimated error rates and, in the end, false positive variant calls. Briefly, our approach to detect and filter these oxo-G affected variant calls is therefore based on comparing i) the dosage as determined when considering a error model that does not consider oxoG errors ii) the dosage as determined with a model that does consider (sample-specific) oxoG errors. The ratio of these two dosages is considered a 'sensitivity' score, which is used to filter genotype calls and/or variants. Dosages are computed using a genotype likelihood calculation detailed below, and are 'posterior dosages' (see previous section): continuous numbers between 0 and 2, which take into account the confidence in the genotypes and the frequency of the variant in the study sample. In the variant QC pipeline, genotype calls with a sensitivity > 1.5 are set to missing, after which variant QC statistics are recalculated. Variants are flagged for exclusion if they have an average sensitivity > 1.5 or a summed dosage with the oxo-G error model < 0.1. The average sensitivity of a variant is here defined as the ratio of the summed normal dosages and the summed oxo-G-corrected dosages. In more detail, the method consists of the following steps:

Statistics

To determine the parameters for the base error model, we estimate for each sample the rate at which oxidation and other base errors occur, dependent also on different sequence contexts (neighboring bases affect the G-oxidation rates). These per-sample statistics are collected using Picard CollectSequencingArtifactMetrics. Next to base errors, we also

obtain summary error metrics per sample, based on measures available as part of the CollectSequencingArtifactMetrics. These consider two forms of the oxoG errors: pre-adapter (in this case G->T errors occur in forward reads, and C->A errors in backward reads) and bait-bias (in this case G->T errors occur in the exome template strand (often the positive strand), and C-A errors in the reverse strand).

Full error model

The error model describes mutation-specific error rates (in contrast to the usual read-nucleotide specific error rates). It takes into account sequence context (a single nucleotide before and after the variant). Strand-specific and forward/backward read specific error rates are averaged: although this information would be useful, it is not available per sample in the variant file (VCF), and a direct link between the original reads in the bam file and the read count in the VCF file is not straightforward to make due to the reassembly step performed by the variant caller.

Contrasting error model

A contrasting error model is created which exclusively models non-oxoG related errors. To this end, we select samples that are not affected by oxoG-related issues, based on the previously described summary metrics. As these summary metrics are sequence-context specific, we obtain a worst-case summary metric per sample, by taking the highest error value across all sequence contexts per sample. Samples with an error rate > 0.0001 for either pre-adapter or bait-bias errors are excluded. Using the remaining samples, regression models are trained which predicts (sequence context-specific) G->T and C->A mutation rates. These regression models are used to fill in G->T and C->A mutation rates for the samples that were excluded due to oxoG effects. Features for these regression models are the (sequence-context-specific) mutation rates for all mutations except G->T and C->A. To handle the extensive collinearity in these features, we reduce the feature space to 10 dimensions by using PCA, and make use of ridge regression.

Genotype likelihood calculation

For each sample, genotype likelihoods are calculated both using the contrasting and full error model. Read counts (r_{ref} and r_{alt} for respectively reads carrying the reference and the alternate allele) are obtained from the VCF file. Based on the error model, sequence context, and reference and alternate allele, ref->alt (e_{ra}) and alt->ref (e_{ar}) error rates are obtained. For a sample *s* (identifier omitted for brevity), and assuming a diploid setting, the likelihood of each genotype is calculated then as:

ref/ref: $(1 - e_{ra})^{r_{ref}} + {e_{ra}}^{r_{alt}}$

ref/alt: $\left(\frac{(1-e_{ra}) + e_{ar}}{2}\right)^{r_{ref}} + \left(\frac{(1-e_{ar}) + e_{ra}}{2}\right)^{r_{alt}}$

alt/alt: $(1 - e_{ar})^{r_{alt}} + {e_{ar}}^{r_{ref}}$

Likelihoods are normalized to sum to 1, and then converted to posterior probabilities ($p_{ref/ref}$, $p_{ref/alt}$ and $p_{alt/alt}$) as outlined in the previous section. The dosage per sample is then calculated as $d_s = p_{ref/alt,s} + 2\ p_{alt/alt,s}$ (where *s* refers to a specific sample) while sensitivity per sample is determined as: $s_s = d_{contrasting,s}/d_{full,s}$. Here, *full* and *contrasting* refer to the used error model to calculate the dosage. In practical use, we found that estimated oxoG-related errors are underestimated. This can be attributed to two factors: i) information loss as no information on read strand, and presence of mutations on forward and backward reads could be used. This could have diluted the estimated oxoG related-errors by a factor 2, ii) a selection bias, as false positive variants caused by this issue are likely sites that present more extreme oxoG-related errors, either by chance or due to (possibly unmodelled) sequence characteristics. To alleviate this issue, an error multiplication factor *f* was introduced, such that errors considered in the full model are rescaled according to $f\ (e_{full} - e_{contrasting}) + e_{contrasting}$. In practice, using $f = 5$led to an adequate filtering of oxoG related variants.

Genotype and variant filtering

Next to a genotype sensitivity measure, we also calculate a variant sensitivity measure: $s_{variant} = \frac{\sum_{samples} d_{contrasing,sample}}{\sum_{samples} d_{full,sample}}$. Variants were exluded from the analysis if $s_{variant} > 1.5$,

or if $\sum_{samples} d_{full,sample} < 0.1$. For variants with $s_{variant} > 1.1$ we performed genotype filtering, setting to missing all genotypes where the genotype sensitivity $s_s > 1.5$. Afterwards, variant QC measures (missingness, Hardy-Weinberg, allele balance, etc) are recalculated.

Variant batch detection and correction

For genetic studies, statistical power is a primary concern. This necessitates large-scale collaborations between sites, as well as the collection of samples that have been sequenced across a large time period. In such settings, it is often impossible to control which capture kits are used, if exome or WGS sequencing is performed, and many other relevant sequencing parameters such as read or fragment lengths. In the ADES consortium, this has resulted in the use of 17 different (versions of) capture kits, the use of both exome and WGS sequencing, read lengths that vary from 50 to 150 bp (**Figure S3**), and many other differences. Moreover, the different contributing studies also have very different case/control balances, ranging from exclusively cases to almost exclusively controls. When performing variant association, this presents a problem, as this step is highly sensitive to batch effects. Even after sample and variant QC, we found that certain variants still present batch effects that lead to spurious associations.

Examples of batch effects

It is not always immediately clear what the cause of such remaining batch effects is. Some examples which were encountered:

— Certain capture kit methods use restriction enzymes to cut sequence fragments before sequencing. We observe that mutations in these restriction sites can at some loci lead to an artificial loss of heterozygosity in the sequencing reads, resulting in a lower than expected allele frequency. Additionally, it is not possible to filter out PCR duplicates for these kits, leading to possible false positive mutations.

— For capture kits that fragment DNA at relatively ‘fixed’ positions in the genome we also observe an increase in batch effects. Explanations for this might include position-related biases in reads or mutations that affect the read coverage of one haplotype.

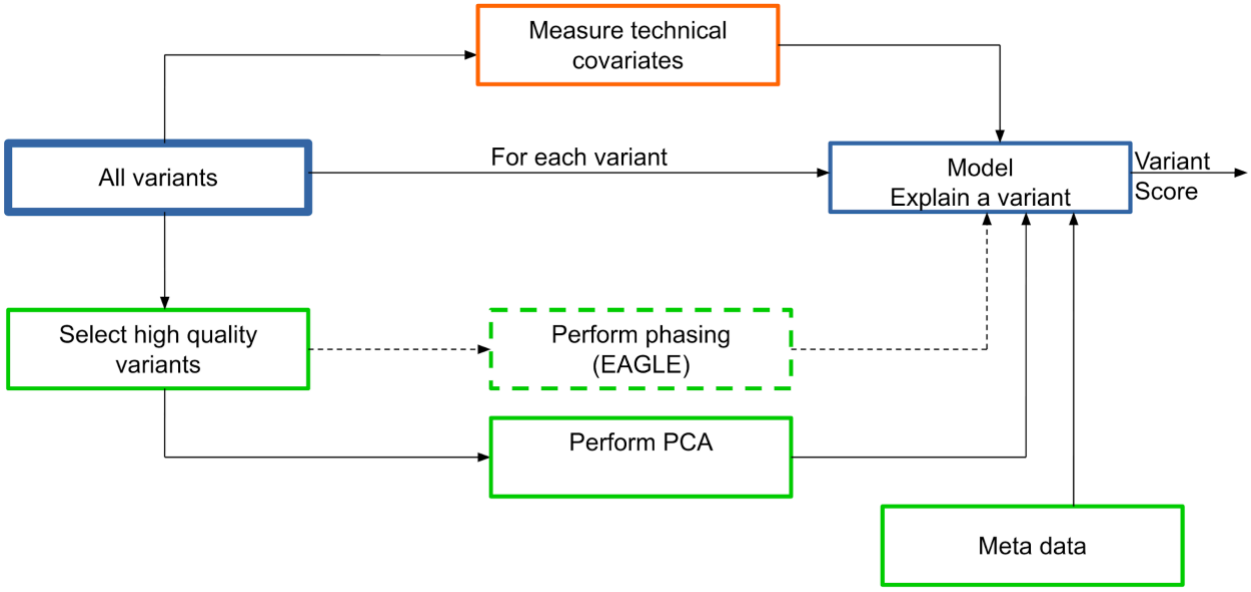
This is observed for capture kits that use restriction enzymes for fragmentation, but to a lesser extent also for those that use transposases, which can have tagmentation biases[38]. Finally, such batch effects are also present in probe-based kits for variants that in terms of read length are distant from a capture probe.

— Increased batch effects are also observed in WGS samples when compared to exome samples. A possible explanation might be that WGS samples have sequence reads originating from the whole genome, in contrast to exome capture kits. In some cases, this could result in sequences being misaligned at certain locations that are not present when using (certain) exome capture kits.

While not every batch effect can be easily be predicted based on causal mechanisms, the presence of many different batches in the dataset still enables the detection of these variants.

Algorithm overview

To this end, a method was developed to detect variants that are affected by such batch effects. The main challenge is to distinguish between non-technical effects that present as batch effects (such as a variant that is enriched in a certain country, and/or only in AD cases) and real batch effects that are caused by technical issues. This is solved by using a two-step approach. In the first step, the algorithm attempts to explain the presence of a variant in specific carriers only through population structure, presence of haploblocks, and/or phenotype effects. Secondly, it is determined if the explanation for the presence of a variant in specific carriers significantly improves if also technical covariates (membership of study batches, various sequencing parameters, etc.) are allowed. Variants for which this is the case are considered to be affected by technical issues, and are either corrected (detailed below) or not considered in the analysis. Below, we first detail the covariates that are used, the algorithm that is used to select the covariates, the regression model, how the presence of not-at-random missing genotypes (i.e. missingness depends on having a specific genotype) is detected, and finally how the algorithm is used in practice.



Technical covariates

Statistics were generated with samtools[24], Picard[21], verifybamid2[25], and custom scripts. Covariates (which are vectors that contain for each sample a value) were defined for the following properties:

- **Batch, study, capture kit:** Covariates describing (for each sample) membership (no: 0, yes: 1) for each batch, study or (version of a) capture kit.
- **Read length, insert size:** Covariates describing read length and average fragment insert size. In addition, covariates were added describing the distance to the nearest capture probe (which differs across the samples due to the use of different kits), both in absolute terms, as well as relative to fragment size or read length (**Figure S3**). For WGS samples, 0 was used as the distance.
- **Contamination:** Contamination percentage as determined by Verifybamid2 (see sample QC).
- **Missingness:** Sample missingness (defined as genotype quality GQ < 40, for variants that are in the intersection of all capture kits, **Figure S4**)
- **Size selection:** The standard deviation of fragment insert-sizes divided by the average of fragment insert sizes. Indicative of the extent of size selection that was performed on the fragments.
- **Read error rate:** Error rate of the reads (mismatches / bases mapped).

- **GC ratio:** Depth of sequences with 35% GC / Depth of sequences with 50% GC
- **Mismapping ratio:** Fraction of fragments for which the two reads map to different chromosomes
- **Duplicate ratio:** Fraction of duplicated reads.
- **Not mapped ratio:** Fraction of reads that are not mapped.
- **Read quality variability:** Standard deviation of average Illumina quality scores across read cycles (a cycle corresponds to a single base position in each read).
- **Fraction of N nucleotides:** Percentage of bases being the N (unknown) nucleotide.
- **Insertion/deletion error fraction:** Nr. of insertions or deletions divided by the nr. of bases mapped.
- **Ts/tv rate, Het/Hom rate, Novel SNPs/Indels rate:** Sample statistics as defined in the sample QC.
- **Gender:** Genetic sex (**Figure S5**).
- **Supplementary reads / fraction of soft-clipped bases:** Fraction of reads with supplementary alignments, and fraction of mapped bases that are soft-clipped.
- **Pre-adapter/Bait oxo-G error pattern:** Phred-scaled error indicating the presence of an oxoG error pattern. 'Pre-adapter' indicates oxoG errors that occurred before adapter ligation, such that read 1 carries G->T mutations and read 2 carries C->A mutations, while 'Bait' indicates an oxoG pattern which is exome bait-specific.
- **Presence of illumina adapters or poly-A tails:** Fraction of reads with respectively Illumina adapters or poly-A tails.

Non-technical covariates

- **PCA covariates:** The top 10 PCA covariates, calculated after sample QC, using an approach described previously[32].
- **Age:** sample age (controls) or age-at-onset (cases). Missing values are imputed to the mean age.
- **AD status:** case or control status
- **Haploblock markers:** to obtain haploblock markers, we select nearby high-quality variants (passing variant QC, with minor allele frequency > 0.025% and a missingness < 10% (missingness defined as read depth < 6)). These variants were phased using

Eagle v2.4[39], with default settings. The resulting haploid genotype calls were used as covariates (algorithm detailed below). The region from which these 'nearby' variants are obtained was by default the 50kb up- and downstream from the variant that was tested for batch effects, with the exception of variants that were within 100bp (as there might be complex false positive events that present as multiple variants close together, which could present a false in-linkage signal). The region can be extended from 50kb up to a maximum of 250kb if there are too few variants (<25), or it can be reduced in size if too many are found (>1000).

— **Complex haploblock markers:** In addition, a search is performed for combination of these nearby variants to better mark the haploblock(s) in which the tested variant occurs (detailed below). Allowed Boolean operations are AND and NOT (e.g. a covariate can be defined which is true if variant 1 AND NOT variant 2 are present in a sample).

Forward-backward covariate search

The above covariates are used in a regression model (detailed below) to explain the tested variant. Covariates are selected using a greedy forward selection/backward elimination approach. First, all covariates are normalized to a range 0-1. A covariate set *E* is defined, which contains covariates that are excluded from the regression, that is, their regression parameter is clamped to 0. Furthermore, a covariate set *I* is defined, which contains covariates that are part of the regression: the parameters of these covariates are optimized using a maximum-likelihood approach. Initially, all covariates are in set *E*, and the regression model is fitted using only an intercept.

For all covariates in set *E*, the maximum likelihood gradient is determined. The covariate with the maximum gradient value is selected, and added to set *I*, after which the regression fit is reoptimized. If the AIC (Akaike Information Criterion[40]) score of the fit is improved, this step is accepted, and a new gradient search is performed to select the next covariate. If the AIC however decreases, the variant is removed from set *I.* The above steps are then repeated for the covariate with the next highest likelihood gradient. The forward search is stopped if none of the top 10 covariates improve the AIC metric. If more than 10 covariates are in set *I*, a backward elimination step is performed, in which each

covariate in set *I* is in turn dropped from the regression to determine if this improves the AIC score. This step is subsequently repeated every time when 5 new covariates have been added to set *I*.

Prioritizing non-technical covariates

To prioritize non-technical explanations for the presence of a variant, the above feature search is first performed using only non-technical covariates, until no model improvements can be found. The resulting AIC score is noted as the *non-technical score*. Next, technical covariates are added to the covariate set *E*, and the feature search is continued until no model improvements can be found anymore. The resulting score at that point is noted as the *technical score*. The final variant batch detection score is then calculated as the delta between these two scores, that is: *vbd score = technical score - non-technical score*.

Diploid logistic regression model

For haploid genotypes (chromosome Y), the above algorithm can be performed using a logistic regression model, in which $\gamma_j = lr(\alpha + \beta x_j)$ Here, *j* is the sample, *lr* is the logistic function, α is the intercept, x_j is the covariate vector for sample *j*, and β is the vector with covariate regression parameters. Normally, in a standard logistic regression, $\gamma_j \in \{0,1\}$. However, due to low coverage data, γ is adapted to represent for each sample the probability of the alternate genotype being present (note: not the posterior probability, but the probability given by the variant caller). Standard implementations of logistic regression usually perform a simplification of the maximum likelihood which assumes dichotomous labels. Therefore, a slightly more generic version of logistic regression was implemented which does not make this assumption. Let $p_j(a,\beta) = lr(a + \beta x_j)$. The log-likelihood then takes the following form: $LL(a,\beta) = \sum_j log(\gamma_j\, p_j(a,\beta) + (\mathbf{1} - \boldsymbol{\gamma_j})(\mathbf{1} - \boldsymbol{p_j}(\boldsymbol{a},\boldsymbol{\beta}))) - \boldsymbol{\lambda}\sum_i \boldsymbol{\beta_i^2}$. This function is maximized in terms of a and β. A small regularization term $\lambda = 0.005$ is added to prevent problems with singularities.

In case of diploid genotypes, this model does not suffice, as each sample can have either a reference, heterozygous or homozygous alternate genotype. The approach is to model

this as what can be seen as two coupled logistic regression models. Conceptually, in a simplified sense: $d_j = lr(\alpha + \beta g_{j,1} + \theta x_j) + lr(\alpha + \beta g_{j,2} + \theta x_j)$**,** where $\boldsymbol{d_j}$is a dosage for sample *j*, in the range [0,2], Here, $g_{j,i}$ is the matrix containing covariates that represent (complex combinations of) phased variants of sample *j* for haplotype *i*, and x_j is the vector with covariate values for sample *j* that are haplotype-independent, with vector θ containing the associated parameter values. Note that the two models share all parameters, but can differ (for phased variants) in their covariates.

More in detail, this is not modelled through dosages, but through genotype probabilities r_j, h_j and o_j, containing respectively the (non-posterior) genotype probabilities of the reference, heterozygous and homozygous alternate genotypes for sample *j*.

Let $p_{j,i}(\alpha, \beta, \theta) = lr(\alpha + \beta g_{j,i} + \theta x_j)$, which will be noted more shortly as $p_{j,i}$, then the maximum likelihood formulation takes the following form:

$$LL(a, \beta, \theta) = \sum_j log(r_j\,(1 - p_{j,1})(1 - p_{j,2}) + \boldsymbol{h_j}(p_{j,1}(1 - p_{j,2}) + (1 - p_{j,1})\,p_{j,2}) + \boldsymbol{o_j} p_{j,1} p_{j,2}) - \boldsymbol{\lambda}\left(\sum_{\boldsymbol{k}} \boldsymbol{\beta_k^2} + \sum_{\boldsymbol{l}} \boldsymbol{\theta_l^2}\right)$$

To optimize this likelihood (as well as for the logistic regression model above), gradients were derived, and the optimization was implemented using the SLSQP optimizer available through Scipy[41].

Tree search for complex haploblock-markers

Earlier, a forward selection-backward elimination algorithm was described to optimize the set of covariates. The main reason to use such an algorithm is clarified here. To tag a haploblock uniquely, the status of multiple SNPs is usually required to define an accurate marker (e.g. the marker is true if variant 1 is present, but not variant 2). Such markers are needed to define the haploblock(s) in which a tested variant occurs. Adding all possible combination of nearby variants would computationally be prohibitively expensive. Regular variant imputation algorithms have a similar problem, and solve this by using Hidden Markov Models on top of phased population haplotypes. It is however not immediately apparent how such an approach can be combined with a regular covariate regression framework as described above. Instead, to still enable the multi-variant haploblock markers, the forward-backward search is used to explore a tree of increasingly complex multi-variant haploblock markers.

The algorithm starts as described, with a set *E* of all covariates that are inactive, i.e. not part of the regression, and an empty set *I* which will contain all covariates that become 'active', i.e. that are selected to be part of the regression model. Next to the covariates that do not represent a genetic variant, set *E* contains at the start only single-variant haplotype markers and no complex multi-variant haplotype markers. That is, the haplotype marker set $Q \subseteq E$is equal to *M*, where *M* is the set of single-variant markers that are near the tested variant (see section on 'non-technical covariates' for how this set of markers is selected). Once a marker $q \in Q$ is moved to set *I*, we extend set *Q* (and thereby set *E*). For a positive association of *q* with the tested variant, we perform: $Q = Q \cup \{q \wedge m, q \wedge \neg m | m \in M\}$, while for a negative association of *q* we perform:$Q = Q \cup \{\neg q \wedge m, \neg q \wedge \neg m | m \in M\}$. Upon removal of marker *q* from set *I*, the reverse operation is performed. Note that usually in this case, one of the complex markers directly dependent on *q* has already been added to set *I.*

Detection of missing-not-at-random genotypes

While missing genotype calls are usually only observed due to lack of read coverage, this is not always the case. In certain situations, missingness was found to correlate with genotype status in certain batches (e.g. non-reference calls were more likely to be missing). This is not detected through the above algorithm, as for a missing genotype call all possible genotypes have the same probability, and therefore the sample has, as designed, no effect on the likelihood of the regression model. To detect these situations, the regression model optimized with the non-technical covariates (first step of algorithm) was used to impute the dosage of all samples. Then, a Fisher exact test was performed for each batch and contributing study, to detect possible allele frequency differences between samples for which the genotype call is missing, and for samples for which the genotype is not missing. More in detail, an imputed posterior dosage is determined using the maximum likelihood fit of the 'non-technical' regression model: $d_j = p_{j,1}(1 - p_{j,2}) + (1 - p_{j,1})\, p_{j,2} + \mathbf{2} p_{j,1} p_{j,2}$. Next, an allele-based Fisher exact test (number of alleles is 2 times number of samples) is performed for each batch and study separately, contrasting samples with a missing genotype call with samples with a non-missing genotype call. P-values < 1e-6 are considered indicative of a problematic batch effect.

Two-phase approach

In some cases, variants that were used as haploblock markers themselves carried large batch effects. Due to this, nearby variants with a similar batch effect pattern were not detected as having such a batch effect. To prevent this from occurring, a two-phase approach was adopted. In the first phase, VBD was run without any haploblock markers. This meant that the non-technical regression model only used the PCA and phenotype covariates. This results in a conservative scoring, as less of the variant is explained by non-technical covariates. Variants that scored a VBD score > 25 in this phase were excluded as haploblock marker in the second phase. In the second phase, the algorithm was then performed as described above, but without the haploblock markers that were excluded by the first phase.

Variant batch correction

For many variants, problematic technical effects were limited to certain batches. In such cases, exclusion of the whole variant seemed unwarranted. To correct these variants, we performed a batch correction step. Variants with a VBD score > 25, or a VBD score > 15 and a MAF < 0.05%, or a batch with a missing genotype batch p-value < $1e^{-6}$ were considered for correction. The correction process was performed iteratively, and continued until the VBD score < 10, and the minimum missing genotype batch p-value > $1e^{-4}$, or if the variant could not be corrected further. In each iteration, correction was performed in two steps. First, the correction process walked through the technical covariates in order of their addition to the regression model. If such a technical covariate described a batch, study or capture kit and led to an AIC score jump of at least 5, the genotypes for the variant under consideration were set to missing for all samples of such a batch, study or capture kit. This process was stopped once a covariate was encountered that did not fall under these criteria. Second, the correction process walked through all batches with a missing genotype batch p-value <$1e^{-4}$, which were set to missing as well. If no batches had a p-value <$1e^{-4}$, but there were contributing studies with a missing genotype p-value <$1e^{-4}$, then studies were considered instead. Variant were annotated both with VBD results before and after correction.

Variant filtering

Finally, variants were considered for analysis if after correction they had a VBD score < 25, or a VBD score < 15 if they had a MAF < 0.05%.

Variant selection and annotation

For the association tests, we performed variant selection (Figure 1c).

1. Protein coding transcripts.

We selected variants in autosomal protein-coding genes that were annotated by VEP (version 94.5[42])to affect the Ensembl basic set of protein coding transcripts (Gencode v19/v29 (liftover to build 37)[43]) of these genes. Transcripts of both Gencode versions were merged based on their identifier, with preference given to the v29-based annotation. Transcripts that passed our filter (protein coding + basic tag) in v19 but not in v29 were not considered.

2. Variant type.

We only kept variants that directly affected the protein (missense, stop_gained, splice_acceptor, splice_donor or frameshift annotation). For LOF annotations, we only kept those variants with a 'HIGH' VEP impact classification, while for missense annotations we required a 'MODERATE' VEP impact classification.

3. Variant prioritization.

We prioritized missense variants using REVEL (Rare Exome Variant Ensemble Learner)[44] (annotation obtained from DBNSFP4.0a[45] and only kept variants with a score ≥ 25 (score range 0 - 100). LOF variants were prioritized using LOFTEE[28] (version 1.0.2), and only LOF variants that had a LOFTEE 'high-confidence' flag were kept.

4. Variant frequency.

Of these, we only kept variants that were estimated to have at least one carrier, and had a minor allele frequency (MAF) of <1%.

5. Variant missingness.

Finally, we removed (5) variants with >20% genotyping missingness (genotypes with a read depth < 6 are considered missing), or that did not pass a filter for differential missingness between the EOAD, LOAD and control groups (Fisher-Exact test comparing EOAD cases versus controls and LOAD cases versus controls, p<1e-20).

6. Variant categorization.

Variants were divided in 4 deleteriousness categories: a LOF category, and 3 missense categories: REVEL ≥ 75, REVEL 50-75 and REVEL 25-50 (Figure 1c).

Analyses and statistical tests

Gene burden test

Based on previous findings in *SORL1*, *TREM2* and *ABCA7*[5] an enrichment can be expected of high impact rare risk variants in early onset cases compared to late onset cases. A regular case/control test (in which only a subset of the cases is EOAD) would be inefficient in picking up such signals. The alternative, performing an additional test that specifically tests for burden in EOAD cases, would however also be inefficient as (1) the additional signal from the LOAD cases would be excluded from the analysis and (2) adding such a test would lead to additional correction for multiple testing. Therefore, we combined both case-control and EOAD tests into one, through the use of ordinal logistic regression, where the genetic risk for AD is considered to increase EOAD > LOAD > control. This test is optimally suited for picking up differential variant loads between the sample categories (EOAD > LOAD > Control), but it can also pick up regular case-control signals for which genetic risk is equally distributed across EOAD and LOAD cases (EOAD ~ LOAD > Control) as well as EOAD-specific signals (EOAD > LOAD ~ Control). The burden test was implemented with the ordinal regression implementation available in the MASS package (version 7.3-51.5) for R (version 3.4.3). Six PCA population covariates

(calculated on the samples remaining after sample QC, using an approach described previously[32] were used, **Figure S16, Figure S17, Figure S18**), and p-values were calculated using a likelihood ratio test (‘*lrtest’* function from the lmtest package, version 0.9-35). An additive model was considered, by summing the dosages of the minor alleles of selected variants. To prevent biases due to missing or low coverage, we sampled the dosage of each variant call (i.e. 0,1 or 2) according to the posterior probabilities (see above) of the reference, heterozygous or homozygous genotypes. The burden test was performed multiple times with independently sampled dosages, to account for sampling uncertainty. P-values and beta values were averaged across these runs, while standard deviations were first converted to variances and then averaged. Repeated runs were performed until either the standard deviation of the mean of log10 transformed p-values became < 0.01, 100 runs were reached, or a mean p-value > 0.01 was obtained with at least 25 runs, or a mean p-value > 0.1 with at least 5 runs.

Variant impact thresholds

We tested the evidence for a differential burden for four sets of variants with incrementing levels of predicted deleteriousness: the LOF+REVEL≥25 threshold includes the variants from all deleteriousness categories, while the LOF+REVEL≥50 threshold and LOF+REVEL≥75 threshold condition on the variants with higher levels of predicted deleteriousness. Finally, the LOF threshold includes only variants that are predicted to lead to a complete loss-of-function. The rationale behind this is that for each gene, by concentrating maximum evidence for a differential burden-signal in one test, we maximize the power to identify a differential burden in this gene. Genes were only tested if the cumulative minor allele count (cMAC) of predicted damaging variants was ≥10. Multiple testing correction was performed across all performed tests (up to 4 per gene) using the False Discovery Rate procedure[46]. Genes were considered for replication if the false discovery rate was ≤20%. Additionally, we used family-wise correction using the Holm-Bonferoni procedure[47] to select genes that were significant in our discovery sample (corrected $p < 0.05$).

Carrier frequency and cumulative Minor Allele Frequency

A carrier of a set of variants was defined as a sample for which the summed dosage of those variants was ≥0.5. Carrier frequencies (CFs) were determined as #carriers / #samples. Confidence intervals for the CFs were assumed to be described through a Beta distribution (where *a*=#carriers, and *b*=#samples - #carriers). To accommodate situations for certain age-at-onset bins, in which the number of carriers was (close to) 0, a prior was added to *a* and *b* based on the carrier count in samples not included in the age-at-onset bin, scaled such that *a=0.1*. The cumulative Minor Allele Frequency (cMAF) for a set of variants and samples was defined as the sum of the minor allele frequencies (MAFs) of the included variants in those samples. When the summed frequency of these variants is <1%, the cMAF can be considered to have a similar uncertainty distribution as the MAF, which can be described using a Beta distribution, where *a=#cumulative Minor Allele Count (cMAC)* and *b=2 * #samples - cMAC*. Similar as for the CF, a prior was added based on the observed allele counts in non-included samples, scaled such that *a=0.1*.

Odds ratios

Effect sizes (odds ratios, ORs) of the ordinal logistic regression can be interpreted as weighted averages of the OR of being an AD case versus control, and the OR of being an early-onset AD case or not. Ordinal odds ratios were calculated for each test, as well as separately for the 4 variant categories REVEL 25-50, 50-75, 75-100 and LOF. Next to ordinal ORs, we estimated 'standard' ORs. This was done across all samples (case/control), as well as per age category (EOAD versus controls and LOAD versus controls), as well as for smaller age-at-onset categories: ≤65 (EOAD), (65-70], (70-80] and >80. Standard ORs were estimated using multinomial logistic regression, using the R nnet package (version 7.3-12), with correction for 6 PCA covariates. For low cMAC values, logistic regression has difficulties in obtaining accurate odds ratios and confidence intervals, as the normal distribution approximation for the log(OR) parameter starts to break down. For these situations (where cMAC≤10, or <3 for either cases or controls), the *OR* and its confidence intervals were estimated directly based on the cMAF of cases and controls: $OR = (cMAF_{case} / cMAF_{control}) / ((1 - cMAF_{case}) / (1 - cMAF_{control})$. While the uncertainty of this OR is difficult to evaluate directly, it is governed by the uncertainty in

cMAF$_{case}$ and *cMAF*$_{control}$. Confidence intervals were therefore estimated through the earlier described beta distribution approximation for the *cMAF*, by repeated sampling of possible *cMAF*$_{case}$ and *cMAF*$_{control}$ values.

Testing for an age-at-onset or a deleteriousness-category effect

We tested if enrichments of damaging variants increased (or decreased for protective variants) towards younger patients. To this end, an ordinal regression using only cases (no controls) was performed, in which cases were grouped in 4 age-at-onset bins: ≤65 , (65-70], (70-80] and <80+. A significant effect (FDR < 0.05) signaled that there was a difference in enrichment between young and older cases. To determine if there was a significant difference in effect sizes between the different deleteriousness categories (REVEL 25-50, 50-75, 75-100 and LOF), an ordinal logistic regression test was performed in which the H0 model included a single beta parameter for all deleteriousness categories, while the H1 model included 4 separate betas for the 4 deleteriousness categories (or <4 when missense deleteriousness categories with cMAC <5 were merged, see caption of Figure 4 for details). We tested if there was a trend effect, in which effect sizes increased with increasing predicted damagingness (REVEL 25-50 < REVEL 50-75 < REVEL 75-100 < LOF). To this end, we modified the ordinal logistic regression implementation, by adding a constraint on the beta parameters: $|b_{REVEL\ 25-50}| \leq |b_{REVEL\ 50-75}| \leq |b_{REVEL\ 75-100}| \leq |b_{LOF}|$ (or equivalent for genes where variant categories were merged). Subsequently, optimization was performed by first estimating **b** in an unconstrained model, followed by adding the model constraints. Likelihood-ratios in this setting follow a chi-bar-squared distribution. Significance (FDR < 0.05) was therefore determined through sample label permutation, based on the bootstrapping approach outlined in Garre et al [48]. The number of permutations was limited to 10.000.

Sensitivity analysis

A sensitivity analysis was performed to determine if effects were potentially due to age differences between cases and controls (**Figure S2**). An age-matched sample was constructed by dividing samples in strata based on age/age-at-onset, with each stratum covering 2.5 years. Case/control ratios in all strata were kept between 0.1 and 10 by

down-sampling respectively controls or cases. Subsequently, samples were weighted using the propensity weighting within strata method proposed by Posner and Ash[49]. Finally, a case-control logistic regression was performed both on the unweighted and weighted case-control labels, and estimated odds ratios and confidence intervals were compared.

Variant-specific analysis

We performed a variant-specific analysis of the genes considered as significantly or suggestively associated with AD, to detect gene-specific idiosyncrasies not covered by our uniform exome-wide analysis. We checked for outlier variants among those that were included in the burden test, determining which ones had a significantly lower or opposite effect size (fisher exact test) compared to other included variants of the same category (missense or LOF). Furthermore, we determined which missense or potential LOF variants did associate with AD (logistic regression test, at least 15 carriers), irrespective of REVEL/LOFTEE or MAF thresholds. We performed corrections for multiple testing per gene using FDR, reporting only variants with a threshold of FDR < 0.2 **(Table S3).** We manually removed and added these variants to the burden tests, in order to calculate, next to standard odds ratios, also refined odds ratios.

Detailed gene discussion

Results of the variant-specific analysis can be found in **Table S3.** For each gene, we discuss 1) variants that were included in the burden test, but found to be outliers, 2) missense and potential LOF variants that associated with AD, 3) other variants of interest.

SORL1

We detected two variants that associated with AD: i) A528T (OR 1.16, 95% CI: 1.05-1.27, MAF 4.9%, FDR: 4%), a common variant, which therefore was not added to the (refined) burden test, ii) a suggestive association for the rare V1459I variant (OR 2.5, 95% CI: 1.22-5.07, FDR: 20%), which due to its REVEL score was not included in the burden test, but was added to the refined analysis. One missense SORL1 variant (S2175R) was detected

as outlier (FDR 6.3%) to the burden test, as it had OR of 0.53 (95% CI 0.19-1.47), lower than other missense variants. This variant was removed from the burden test in the refined analysis. The addition of V1459I and removal of S2175R in the refined burden analysis changed the SORL1 missense OR from 2.2 to 2.5 (**Table 1**).

TREM2

We detected 5 missense variants that associated with AD: i) R47H[50] (OR 3.7, 95%CI 2.8-4.9, FDR: 3.8e-10%), which was included in the burden test, ii) R62H[51] (OR 1.6, 95%CI 1.3-1.9, MAF 1.3%, FDR: 0.0006%), which is common, and was therefore not included in the (refined) burden test, iii) a new significant association for D87N (OR 2.6, 95%CI, 1.6-4.6, MAF:0.15%, FDR: 1%), iv) we confirmed the recently significantly associated H157Y[50] (OR 6.4, 95%CI: 2.7-15.2, MAF:0.05%, FDR:1%), and v) found a new significant association in L211P (OR 2.3, 95%CI: 1.3-3.9, MAF:0.05%, FDR:2%). The last three variants all had a low REVEL score, and were therefore not included in the burden test, but were added to the refined analysis. Notably, missense variant L211P affects only the canonical transcript, while the other mentioned missense variants affect all 3 protein-coding transcripts of TREM2. For LOF variants, we detected an outlier splice acceptor variant rs538447052 (OR: 1.9, 95%CI: 0.7-5.1, MAF: 0.04%), which only affected the non-canonical ENST00000373122 transcript. This variant had a significantly lower odds ratio (outlier FDR: 11%) compared to the other LOF variants that affect all transcripts. It was therefore removed in the refined analysis. Furthermore, we also note a suggestive association for a stop gained variant which only affected the soluble TREM2 transcript ENST00000338469 (OR: 2.3, 95%CI: 0.9-5.8, FDR: 20%). This variant was carried by 20 individuals (17 cases, 3 controls), and was not included in our burden test as it had a low-confidence classification from LOFTEE due to its location in the last exon. Given the different biological effect and the relatively lower OR compared to the other LOF variants that affect all transcripts, this variant was not added to the refined analysis. After refinement (inclusion of D87N, H157Y, L211P, and removal of the splice acceptor LOF variant for the non-canonical transcript), the LOF odds ratio of TREM2 was determined to be 10.8 (95% CI: 4.4-26.9), while the missense OR 3.5 (95% CI:3.1-6.1).

ABCA7

We associated 4 missense variants in ABCA7: i) L101R (OR: 3.7, 95%CI 2.1-6.4, FDR: 0.5%), which was included in the burden test, ii) a new significant association for a common protective variant G215S (OR: 0.89, 95%CI: 0.81-0.97, FDR: 2%), for which previously a suggestive protective association was found[52], iii and iv) a protective association in common variants H395R and Q1686R, which are (known to be) in tight linkage [53]. For H395R, a damaging association was previously found in African Americans[54], where the variant is much more common (25% vs. 3.5% in our study). These 4 variants did not lead to any changes in the refined burden analysis. Additionally, there were 2 missense variants detected as outlier in the burden test: i) R19W (outlier FDR: 5%), with an OR of 1.09 (95% CI: 0.4-3.2). We note that the OR in our study might be underestimated, as this variant was mainly present in young controls (median age 57). ii) V1599M (outlier FDR:1.8%), with an OR of 0.84 (95%CI: 0.61-1.15, MAF:0.4%). In the refinement analysis, these two variants were removed. The resulting missense OR in the burden test was 1.4 (95%CI: 1.3-1.6). Of note, our discovery analysis excluded two relatively often occurring LOF variants, flagged in our QC pipeline for differential missingness. However, for these variants, it was possible to reliably calculate a single-variant association (by excluding samples with low depth). The first variant is the splice-altering variant c.5570+5G>C, which maps outside the closest canonical splice site and hence did not fulfill our inclusion criteria for the exome wide burden tests. A loss of function effect was demonstrated in vitro for this variant[1]. In our study, we observed an OR of 1.67 (95%CI 1.2-2.3, p=0.002, MAF=0.43%). The second variant is the LOF frameshift variant 708-710:EEQ/X (earlier observed by de Roeck et al[55]) for which we report an OR of 2.0 (95%CI 1.36-3.01, p=0.003, MAF=0.27%). These odds ratios are in line with those obtained for the LOF burden test (OR 1.8, 95%CI: 1.2-2.6). Finally, we did not have the possibility to call an intronic variable number tandem repeat (VNTR) which was recently associated with an increased risk of developing AD, suggesting that the level of association of ABCA7 in AD is still likely underestimated in our study[56]. However, it is important to keep in mind that the real impact of some LOF mutations in ABCA7 may be restricted by a transcript rescue mechanism[55].

ABCA1

We associated 5 missense variants in ABCA1 with AD: i,ii) two common missense variants V825I and I883M, with a (suggestive) protective association with AD: OR 0.93 and OR 0.91 respectively. iii) A rare suggestive protective association of variant A1182T (OR 0.49, 95%CI: 0.25-0.95, FDR: 13%). iv) A rare suggestive association with increased AD risk of variant R1680Q (OR 2.75, 95% CI: 1.27-5.95, FDR: 13%). v) A significant association with 4.2-fold increased AD risk of variant N1800H (OR 4.2; 95%CI: 2.0-8.6, FDR:2%, MAF: 0.08%). This variant was not included in our burden test due to a low REVEL core. Furthermore, we detected 2 variants as outlier in the burden test: vi) a missense variant (outlier FDR: 4.8%) in D1018G (OR 0.81, 95%CI: 0.29-2.22, MAF:0.04%) and vii) a splice donor variant (outlier FDR: 0.3%, 9:107565564:C>A), which had an OR of 0.94 (95%CI: 0.19-4.52, MAF: 0.02%). In our refinement analysis, we removed the latter variants (vi and vii) and added variants (iv) R1680Q and (v) N1800H. The burden of all LOF associated with a 5-fold increased risk for AD (OR 4.9, 95%CI 2.1-11.4) and the burden of all missense mutations associated with a 2-fold increased risk of AD (OR 2.1, 95%CI: 1.6-2.7).

ADAM10

We note that one splice-acceptor LOF variant, carried by a single control, only affects transcripts ENST00000402627 and ENST00000561288. These transcripts, being 71 and 38 amino acids long, miss the majority of the canonical transcript (748 amino acids). This individual was last checked at age 89.

ACKNOWLEDGMENTS

ADES-FR

This study was funded by grants from the Clinical Research Hospital Program from the French Ministry of Health (GMAJ, PHRC, 2008/067), the CNR-MAJ, the JPND PERADES, and Equipe FRM DEQ20170336711. This research was supported by the Laboratory of Excellence GENMED (Medical Genomics) grant no. ANR-10-LABX-0013 managed by the National Research Agency (ANR) part of the Investment for the Future program. This work was also supported by Foundation Alzheimer, the Institut Pasteur de Lille, Inserm, the Haut-de-France and Lille Métropole Communauté Urbaine council, and the French government's LABEX (laboratory of excellence program investment for the future) DISTALZ grant (Development of Innovative Strategies for a Transdisciplinary approach to Alzheimer's disease). The 3C Study supports are listed on the Study Website (www.three-city-study.com).

AgeCoDe-UKBonn

The AgeCoDe cohort was funded in part by the German Federal Ministry of Education and Research (BMBF) (grants KNDD 01GI0710, 01GI0711, 01GI0712, 01GI0713, 01GI0714, 01GI0715, 01GI0716, 01ET1006B). Sequencing of AgeCoDe sample was in part funded by the German Research Foundation (DFG) grant RA 1971/6-1 to Alfredo Ramirez.

Barcelona- SPIN

Support for Jordi Clarimon provided by Maratón RTVE (Spain).Support for Oriol Dols provided by the Association for Frontotemporal Degeneration (Clinical Research Postdoctoral Fellowship, AFTD).

100-plus Study

Cohort collection and exome sequencing of the 100-plus Study cohort was supported by Stichting Alzheimer Nederland (WE.09-2014-03); Stichting Dioraphte; JPND-PERADES (ZonMw 733051022): HorstingStuit Foundation, VUmc Fund, and Dioraphte Project 17020403,

ERF

The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4- 2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043).

Rotterdam Study

The generation and management of the exome sequencing data for the Rotterdam Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE) (014-93-015; RIDE2), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO), the Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810, and by a Complementation Project of the Biobanking and

Biomolecular Research Infrastructure Netherlands (BBMRI-NL; www.bbmri.nl ; project number CP2010-41). We thank Mr. Pascal Arp, Ms. Mila Jhamai, Mr. and Marijn Verkerk, for their help in creating the RS-Exome Sequencing database.

AC-EMC

Exome sequencing was funded by Alzheimer Nederland.

ADC-Amsterdam

We thank all study participants and all personnel involved in data collection for the contributing studies. Research of Alzheimer center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience. Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland and Stichting VUmc fonds. The chair of Wiesje van der Flier is supported by the Pasman stichting. The clinical database structure was developed with funding from Stichting Dioraphte. This work was supported by Stichting Alzheimer Nederland (WE.09-2014-06, WE.05-2010-06); Stichting Dioraphte; Internationale Stichting Alzheimer Onderzoek (#11519); JPND-PERADES (ZonMw 733051022): Centralized Facility for Sequence to Phenotype analyses (ZonMW 9111025); Netherlands Consortium for Healthy Aging (NCHA 050-060-810); Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL CP2010-41); Netherlands Genomics Initiative (NGI)/NWO.

PERADES:

We thank all individuals who participated in the study. We also want to express our gratitude to the MRC Centre Core Team for the laboratory support and the Advanced Research Computing at Cardiff University (ARCCA) for the computational support. Cardiff University was supported by the Medical Research Council. Cardiff University was also supported by the European Joint Programme for Neurodegenerative Disease, Alzheimer's Research UK, the Welsh Assembly Government, and a donation from the

Moondance Charitable Foundation. Cardiff University acknowledges the support of the UK Dementia Research Institute, which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK. Cambridge University acknowledges support from the MRC. The University of Southampton acknowledges support from the Alzheimer's Society. ARUK provided support to Nottingham University. Join Dementia Research (JDR) is funded by the Department of Health and delivered by the National Institute for Health Research in partnership with Alzheimer Scotland, Alzheimer's Research UK and Alzheimer's Society

CBC: Control Brain Consortium

This work was supported by the UK Dementia Research Institute which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK, Medical Research Council (award number MR/N026004/1). Wellcome Trust Hardy (award number 202903/Z/16/Z), Dolby Family Fund; National Institute for Health Research University College London Hospitals Biomedical Research Centre; BRCNIHR Biomedical Research Centre at University College London Hospitals NHS Foundation Trust and University College London.

UCL-DRC EOAD

This work was supported by the Medical Research Council (UK), the Biomedical Research Centre at University College London Hospitals NHS Foundation Trust and charitable donations to the UCL Dementia Research Centre.

ADSP

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and

Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage, U54 AG052427 to Drs. Schellenberg and Wang, and R01 AG054060 to Dr Naj. The ADGC cohorts include: Adult Changes in Thought (ACT) (UO1 AG006781, UO1 HG004610, UO1 HG006375, U01 HG008657), the Alzheimer's Disease Centers (ADC) (P30 AG019610, P30 AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50 AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), the Chicago Health and Aging Project (CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 AG016574), Mayo Parkinson's Disease controls (NS039764, NS071674, 5RC2HG005605), University of Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, A2011048), the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE) (R01 AG09029, R01 AG025259), the National Cell Repository for Alzheimer's Disease (NCRAD) (U24 AG21886), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA- LOAD) (R01 AG041797), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), the Texas Alzheimer's Research and Care Consortium (TARCC) (funded by the Darrell K Royal Texas Alzheimer's Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU) (R01 AG019757, R01 AG021547, R01 AG027944, R01 AG028786, P01 NS026630, and Alzheimer's Association), the Washington Heights-Inwood Columbia Aging Project (WHICAP) (RF1

AG054023), the University of Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: R01NS069719), the Columbia University HispanicEstudio Familiar de Influencia Genetica de Alzheimer (EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded by Wellcome Trust, Medical Research Council, Canadian Institutes of Health Research), and Genetic Differences (GD) (R01 AG007584). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. R01 AG048927 for the Gwangju Alzheimer and Related Dementias Study, RF1 AG054080 to Dr. Beechem, U24 AG056270 to Dr. Mayeux, RF1 AG057519 to Dr. Farrer and Pericak-Vance, U01 AG062602 to Dr. Farrer, R01 AG067501 to Dr. Mayeux, and P30 AG13846 to Dr. Farrer

The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904),the EU Joint Programme - Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by

contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4- 2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services.

The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

Supplemental Authors

Miquel Aguilar, Ignacio Alvarez, Marina Arcaro, Nandini Badarinarayan, Carol Brayne, Keeley Brookes, Roberta Cecchetti, Monica Diez-Fairen, Chiara Fenoglio, Tamar Guetta-Baranes, Carmen Lage, Sara Lopez-Garcia, Simon Lovestone, John Powell, Eloy Rodriguez-Rodriguez, David Rubinsztein, Francesca Salani, Elio Scarpini, Sandro Sorbi, Elisa Toppi

References

1 Steinberg, S. *et al.* Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nature genetics* **47**, 445-447, doi:10.1038/ng.3246 (2015).

2 Bis, J. C. *et al.* Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Molecular psychiatry*, doi:10.1038/s41380-018-0112-7 (2018).

3 McKhann, G. M. *et al.* The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* **7**, 263-269, doi:10.1016/j.jalz.2011.03.005 (2011).

4 McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-944 (1984).

5 Bellenguez, C. *et al.* Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. *Neurobiol Aging* **59**, 220 e221-220 e229, doi:10.1016/j.neurobiolaging.2017.07.001 (2017).

6 Nicolas, G. *et al.* SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. *Molecular psychiatry*, doi:10.1038/mp.2015.121 (2015).

7 Nicolas, G. *et al.* Screening of dementia genes by whole-exome sequencing in early-onset Alzheimer disease: input and lessons. *European Journal of Human Genetics* **24**, 710-716, doi:10.1038/ejhg.2015.173 (2015).

8 Lambert, J. C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature genetics* **41**, 1094-1099, doi:10.1038/ng.439 (2009).

9 The 3C Study Group. Vascular Factors and Risk of Dementia: Design of the Three-City Study and Baseline Characteristics of the Study Population. *Neuroepidemiology* **22**, 316-325, doi:10.1159/000072920 (2003).

10 Deli, M. *et al.* Prediction of Dementia in Primary Care Patients. *PloS one* **6**, doi:10.1371/journal.pone.0016852 (2011).

11 Luck, T. *et al.* Mild Cognitive Impairment in General Practice: Age-Specific Prevalence and Correlate Results from the German Study on Ageing, Cognition and Dementia in Primary Care Patients (AgeCoDe). *Dementia and geriatric cognitive disorders* **24**, 307-316, doi:10.1159/000108099 (2007).

12 Zaudig, M. *et al.* SIDAM – A Structured Interview for the diagnosis of Dementia of the Alzheimer type, Multi-infarct dementia and dementias of other aetiology according to ICD-10 and DSM-III-R. *Psychological medicine* **21**, 225-236, doi:10.1017/s0033291700014811 (2009).

13 Alcolea, D. *et al.* The Sant Pau Initiative on Neurodegeneration (SPIN) cohort: A data set for biomarker discovery and validation in neurodegenerative disorders. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **5**, 597-609, doi:10.1016/j.trci.2019.09.005 (2019).

14 Alcolea, D. *et al.* Amyloid precursor protein metabolism and inflammation markers in preclinical Alzheimer disease. *Neurology* **85**, 626-633, doi:10.1212/wnl.0000000000001859 (2015).

15 Holstege, H. *et al.* The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. *European journal of epidemiology* **33**, 1229-1249, doi:10.1007/s10654-018-0451-3 (2018).
16 Kahle-Wrobleski, K., Corrada, M. M., Li, B. & Kawas, C. H. Sensitivity and specificity of the mini-mental state examination for identifying dementia in the oldest-old: the 90+ study. *Journal of the American Geriatrics Society* **55**, 284-289, doi:10.1111/j.1532-5415.2007.01049.x (2007).
17 van der Flier, W. M. & Scheltens, P. Amsterdam Dementia Cohort: Performing Research to Optimize Care. *J Alzheimers Dis* **62**, 1091-1111, doi:10.3233/JAD-170850 (2018).
18 Guerreiro, R. *et al.* A comprehensive assessment of benign genetic variability for neurodegenerative disorders. *BioXiv Preprint 10.1101/270686*, doi:10.1101/270686 (2018).
19 Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature genetics* **51**, 414-430, doi:10.1038/s41588-019-0358-2 (2019).
20 Beecham, G. W. *et al.* The Alzheimer's Disease Sequencing Project: Study design and sample selection. *Neurology Genetics* **3**, doi:10.1212/nxg.0000000000000194 (2017).
21 Broad Institute. PicardTools. *Broad Institute, GitHub repository* http://broadinstitute.github.io/picard/ (Accessed: 2018/02/21; version 2.17.8).
22 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
23 Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503-2505, doi:10.1093/bioinformatics/btu314 (2014).
24 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
25 Zhang, F. *et al.* Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome research* **30**, 185-194, doi:10.1101/gr.246934.118 (2020).
26 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
27 Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv preprint: doi: 10.1101/201178*, doi:10.1101/201178 (2018).
28 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).
29 Broad Institute. (2012).
30 Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome biology* **9**, doi:10.1186/gb-2008-9-9-r137 (2008).
31 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, doi:10.1126/science.1072047 (2002).
32 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909, doi:10.1038/ng1847 (2006).
33 Thornton, T. *et al.* Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS genetics* **13**, doi:10.1371/journal.pgen.1007021 (2017).
34 Richards, 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. *Genetics in Medicine* **17**, 405-423, doi:10.1038/gim.2015.30 (2015).
35 Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research* **41**, e67-e67, doi:10.1093/nar/gks1443 (2013).

36 Heng, L. *Low-complexity regions in hs37d5*, <https://figshare.com/articles/dataset/Low_complexity_regions_in_hs37d5/969685> (2014).
37 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).
38 Kia, A. *et al.* Improved genome sequencing using an engineered transposase. *BMC Biotechnology* **17**, doi:10.1186/s12896-016-0326-1 (2017).
39 Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics* **48**, 1443-1448, doi:10.1038/ng.3679 (2016).
40 Akaike, H. in *Selected Papers of Hirotugu Akaike Springer Series in Statistics* Ch. Chapter 15, 199-213 (1998).
41 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261-272, doi:10.1038/s41592-019-0686-2 (2020).
42 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome biology* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).
43 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **47**, D766-D773, doi:10.1093/nar/gky955 (2019).
44 Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* **99**, 877-885, doi:10.1016/j.ajhg.2016.08.016 (2016).
45 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human mutation* **37**, 235-241, doi:10.1002/humu.22932 (2016).
46 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
47 Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**, 65-70 (1979).
48 Garre, F. G., Vermunt, J. K. & Croon, M. A. Likelihood-ratio tests for order-restricted log-linear models: A comparison of asymptotic and bootstrap methods. *Metodologia de las Ciencias del Comportamiento* **4**, 325-337 (2002).
49 Posner, M. A. & Ash, A. S. Comparing weighting methods in propensity score analysis. *Unpublished working paper, Columbia University* **http://www.stat.columbia.edu/~gelman/stuff_for_blog/posner.pdf** (2012).
50 Zhou, S.-L. *et al.* TREM2 Variants and Neurodegenerative Diseases: A Systematic Review and Meta-Analysis. *Journal of Alzheimer's Disease* **68**, 1171-1184, doi:10.3233/jad-181038 (2019).
51 Jin, S. C. *et al.* Coding variants in TREM2 increase risk for Alzheimer's disease. *Human molecular genetics* **23**, 5838-5846, doi:10.1093/hmg/ddu277 (2014).
52 Sassi, C. *et al.* ABCA7 p.G215S as potential protective factor for Alzheimer's disease. *Neurobiol Aging* **46**, 235 e231-239, doi:10.1016/j.neurobiolaging.2016.04.004 (2016).
53 Nuchnoi, P., Nantakomol, D., Chumchua, V., Plabplueng, C. & Isarankura-Na-Ayudhya, C. The Identification of Functional Non-Synonymous SNP in Human ATPBinding Cassette (ABC), Subfamily Member 7 Gene: Application of Bioinformatics Tools in Biomedicine. *Journal of Bioanalysis & Biomedicine* **03**, doi:10.4172/1948-593x.1000039 (2011).
54 Logue, M. W. A Comprehensive Genetic Association Study of Alzheimer Disease in African Americans. *Archives of neurology* **68**, doi:10.1001/archneurol.2011.646 (2011).
55 De Roeck, A. *et al.* Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease. *Acta neuropathologica* **134**, 475-487, doi:10.1007/s00401-017-1714-x (2017).

56 De Roeck, A. *et al.* An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta neuropathologica*, doi:10.1007/s00401-018-1841-z (2018).