BioMed

MDPI

Article

Comprehensive Analysis of the Genetic Variation in the *LPA* Gene from Short-Read Sequencing

Raphael O. Betschart [1], Georgios Koliopanos [1], Paras Garg [2], Linlin Guo [3], Massimiliano Rossi [4], Sebastian Schönherr [5], Stefan Blankenberg [3,6,7], Raphael Twerenbold [3,6,7], Tanja Zeller [3,6,7] and Andreas Ziegler [1,3,6,8,*]

1 Cardio-CARE, Medizincampus Davos, Herman-Burchard-Str. 12, 7265 Davos, Switzerland
2 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Hess Center for Science and Medicine, New York, NY 10029, USA
3 Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany
4 Illumina Inc., San Diego, CA 92122, USA
5 Institute of Genetic Epidemiology, Medical University of Innsbruck, 6020 Innsbruck, Austria
6 Centre for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany
7 German Center for Cardiovascular Science (DZHK), Partner Site Hamburg/Kiel/Lübeck, 20246 Hamburg, Germany
8 School Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg 3209, South Africa
* Correspondence: ziegler.lit@mailbox.org; Tel.: +41-81-410-1800

Abstract: Lipoprotein (a) (Lp(a)) is a risk factor for cardiovascular diseases and mainly regulated by the complex *LPA* gene. We investigated the types of variation in the *LPA* gene and their predictive performance on Lp(a) concentration. We determined the Kringle IV-type 2 (KIV-2) copy number (CN) using the DRAGEN *LPA* Caller (DLC) and a read depth-based CN estimator in 8351 short-read whole genome sequencing samples from the GENESIS-HD study. The pentanucleotide repeat in the promoter region was genotyped with GangSTR and ExpansionHunter. Lp(a) concentration was available in 4861 population-based subjects. Predictive performance on Lp(a) concentration was investigated using random forests. The agreement of the KIV-2 CN between the two specialized callers was high (r = 0.9966; 95% confidence interval [CI] 0.9965–0.9968). Allele-specific KIV-2 CN could be determined in 47.0% of the subjects using the DLC. Lp(a) concentration can be better predicted from allele-specific KIV-2 CN than total KIV-2 CN. Two single nucleotide variants, 4925G>A and rs41272114C>T, further improved prediction. The genetically complex *LPA* gene can be analyzed with excellent agreement between different callers. The allele-specific KIV-2 CN is more important for predicting Lp(a) concentration than the total KIV-2 CN.

Keywords: genome-wide association study; lipoprotein(a); Kringle IV-2 repeat; KIV-2; short tandem repeat; whole genome sequencing

check for updates

Citation: Betschart, R.O.; Koliopanos, G.; Garg, P.; Guo, L.; Rossi, M.; Schönherr, S.; Blankenberg, S.; Twerenbold, R.; Zeller, T.; Ziegler, A. Comprehensive Analysis of the Genetic Variation in the *LPA* Gene from Short-Read Sequencing. *BioMed* **2024**, *4*, 156–170. https://doi.org/10.3390/biomed4020013

Academic Editor: Wolfgang Graier

Received: 19 March 2024
Revised: 29 May 2024
Accepted: 31 May 2024
Published: 4 June 2024
Corrected: 29 May 2026

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Lipoprotein (a) or Lp(a) molecules are a unique group of lipoproteins with unclearly deciphered function. Lp(a) is a lipoprotein consisting of a large apolipoprotein (a) molecule and a low-density lipoprotein (LDL) particle [1,2]. The apolipoprotein (a) is covalently bound to apolipoprotein B-100 through a disulfide bond [1,2]. Lp(a) shows a broad range of concentrations, ranging from 0.1 mg/dL to more than 200 mg/dL [3,4]. Lp(a) levels differ between populations of different geographic origin, and African populations show 2-3-fold higher Lp(a) levels than European and Asian populations [4].

High Lp(a) levels are associated with an increased cardiovascular disease (CVD) risk [5]. Several hypotheses exist as to why high Lp(a) levels may increase CVD risk.

One hypothesis is that Lp(a) inhibits the breakdown of blood clots by interfering with plasminogen. Interestingly, the *PLG* gene, which encodes plasminogen, and the *LPA* gene are thought to have evolved from the recent duplication of a common precursor gene [6–8]. It has been shown that Lp(a) acts as a modulator between blood clotting and fibrinolysis, but the physiological function remains unknown [3]. Furthermore, it was proposed that Lp(a) delivers cholesterol to sites of injury and wound healing when bound to fibrin [9].

Lp(a) levels are highly determined by underlying genetics, and the *LPA* gene explains up to 90% of the Lp(a) variation [3]. The main driver for the variability of Lp(a) levels is the Kringle IV (KIV) structure, which consists of 10 subtypes (Figure 1). The subtype KIV-2 is 5.5 kilobases long, highly variable, and the number of KIV-2 repeats ranges from 1 to >40 [4]. A high number of KIV-2 repeats is associated with low Lp(a) protein levels, and more than 95% of subjects are heterozygous for the number of KIV-2 repeats [3]. It is important to note that Lp(a) protein levels are primarily determined by the number KIV-2 repeats on the shorter allele, because a small number of KIV-2 repeats might facilitate secretion by the liver, and therefore leads to a higher Lp(a) protein levels [8,10,11].

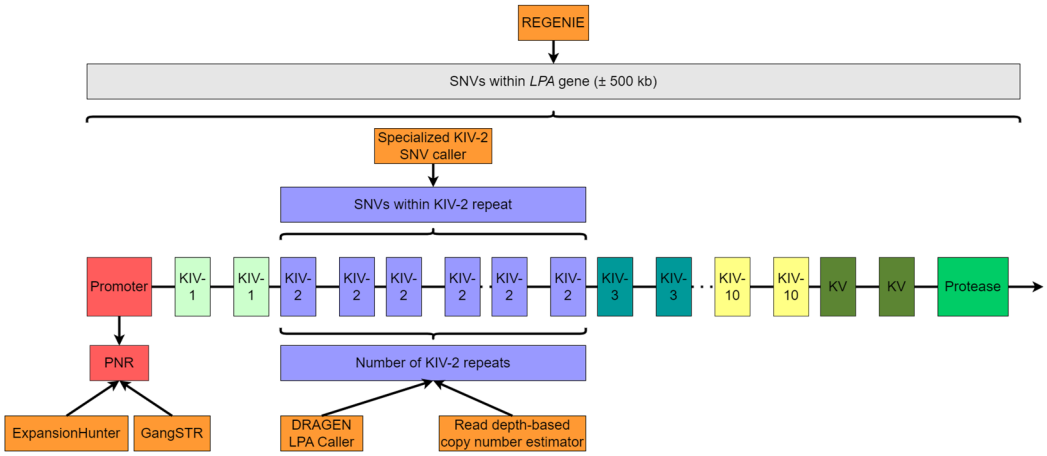


Figure 1. *LPA* gene structure, genetic variation, and different callers for short-read whole genome sequencing data. Figure not drawn to scale.

The number of KIV-2 repeats has been traditionally measured by laborious wet-lab experiments, such as immunoblotting and pulsed-field gel electrophoresis (PFGE) [12,13]. Because of the strong linkage disequilibrium (LD) between the KIV-2 repeat and several single nucleotide variations (SNVs) in the *LPA* gene, SNVs have been used as simple-to-use proxies for the KIV-2 repeat [4,14,15]. While this approach is accurate for European-like populations, it is diminishing for other populations, especially East Asian-like populations [15].

The use of long-read sequencing showed promising results to resolve complex genes such as *LPA* because of their high read length [16]. The high costs associated with this technology prevent its widespread application in both clinical applications and large-scale population-based studies [17]. Furthermore, even long-read sequencing technologies show non-unique alignments in the complex KIV-2 region [18]; the longer the reads, the lower the fraction of nun-unique alignments.

A recent technological advancement is the development of specialized callers for estimating the number of KIV-2 repeats from short-read whole genome sequencing (WGS) data. Specifically, the DRAGEN *LPA* Caller allows determining the number of KIV-2 repeats [18]. Depending on the genotypes at two SNVs at position 296 and 1264 within the KIV-2 repeat unit, this caller may be able to determine the allele-specific KIV-2 repeat number [18]. A more general caller, termed read depth-based copy number estimator (CNE), to determine all multicopy genes was publicly released in late 2023 [19]. CNE uses the read depth from short-read WGS to estimate the total copy number for the genes of interest.

A number of genome-wide association studies (GWAS) have identified several SNVs which are associated with Lp(a) levels [12,20]. These SNVs are primarily located in the *LPA* gene, with some of them in the KIV-2 repeat. Because of its complex genetic structure, a specialized caller is needed for the KIV-2 repeat. This caller can determine if a subject carries a variant or not [13].

It was suggested that a pentanucleotide repeat (PNR) in the promotor region upstream of the *LPA* gene alters Lp(a) levels [4,21,22]. Specifically, the PNR alleles with 10 or 11 repeats were associated with a small number of KIV-2 repeats, but counterintuitively expressed very low Lp(a) concentrations [21]. However, it has been demonstrated that the association between a high number of KIV-2 repeats and low Lp(a) levels is mediated by an SNV located in the KIV-2 repeat [23]. Therefore, adding this SNV to a model predicting Lp(a) concentrations should abolish the association with the PNR.

The aim of this study was to investigate all types of genetic variation in the *LPA* gene and their association with Lp(a) levels. To this end, we used the 8351 short-read WGS from the GENEtic SequencIng Study Hamburg-Davos (GENESIS-HD) study for a comprehensive analysis of the *LPA* gene. Lp(a) protein level measurements were available for 4861 of these subjects. We compared both callers, DRAGEN *LPA* Caller and CNE, to determine the number of KIV-2 repeats against each other on their agreement in determining the total number of KIV-2 repeats. Furthermore, we compared both PNR callers ExpansionHunter and GangSTR and their agreement in determining the number of PNRs. We developed a random forest model to predict Lp(a) levels from all types of genetic variation. We hypothesized that the model incorporating all genetic variation would show the highest predictive performance because it has more predictive features than a sparse model. Furthermore, because each allele is separately transcribed and translated into an Lp(a) protein, we further hypothesized that models incorporating allele-specific KIV-2 repeat numbers would have higher predictive performance compared to models based solely on the total number of KIV-2 repeats. We also investigated whether the PNR is predictive for Lp(a) measurement if SNV 4925G>A located in the KIV-2 repeat is included in the model.

2. Materials and Methods

2.1. Cohort

The GENESIS-HD study is a collaborative effort. It was planned to sequence 9000 individuals in total, of which approximately 8000 were to come from the German population-based Hamburg City Health Study (HCHS); for a description of the HCHS design, see [24]. An additional set of approximately 1000 subjects were to be selected from patient-based clinical cohorts with distinct cardiovascular characteristics. The clinical cohorts included subjects with myocardial infarction at a young age, among others; for details, see Betschart et al. [25]. In this study, only Lp(a) levels measured in 4861 individuals from the HCHS were used.

The local ethics committee of the Landesärztekammer Hamburg (State of Hamburg Chamber of Medical Practitioners, PV5131) had no objections against the conduct of the study. The Data Protection Commissioner of the University Medical Center of the University Hamburg-Eppendorf and the Data Protection Commissioner of the Free and Hanseatic City of Hamburg approved the study. The study is registered at ClinicalTrial.gov (NCT03934957).

2.2. Measurement of Lp(a)

Lp(a) was centrally measured in thawed serum samples previously stored at −80° degrees Celsius with the molar-based Tina-quant® Gen 2 Lipoprotein (a) assay on a COBAS INTEGRA® 400 plus analyzer (Roche Diagnostics, Rotkreuz, Switzerland) at the University Medical Center Hamburg-Eppendorf. Concentrations were reported in nmol/L with a limit of detection (LOD) of 7 nmol/L. The assays inter-coefficient of variation (CV) was 5.86%, the inter-CV was 3.76%.

2.3. Sequencing

Core elements of the sequencing protocol for the GENESIS-HD study are as follows: After arrival of the DNA samples from the University Medical Center Eppendorf (UKE) in Hamburg at the University Medical Center Zurich (USZ), DNA concentration was measured with PicoGreen, followed by an automated DNA normalization with Hamilton Robotics. The library was constructed according to Illumina TruSeq DNA PCR Free Library Prep protocol HT (Illumina Inc., San Diego, CA, USA) for whole genome sequencing. Briefly, the protocol steps were: (1) fragmentation of 1 µg genomic DNA to 350 bp inserts by Covaris LE220-plus, (2) cleanup of fragmented DNA, (3) repair ends, (4) removal of large and small DNA fragments, (5) 3′-end adenylation, and (6) adapter ligation. The resulting library was quantified and quality assessed with the iSeq100 (Illumina). Samples were normalized according to the quantification values, and 54 samples were pooled for sequencing on an Illumina NovaSeq 6000 sequencer (Illumina Inc., San Diego, CA, USA). Samples were sequenced twice on S4 flow cells with 300 cycles (2 × 150 reads) with an estimated coverage of 15 × each, following Illumina protocols. The aimed sequencing depth was that ≥95% of all samples had a coverage of ≥30×.

2.4. Pre-Processing, Quality Control, and Multi-Sample Calling of WGS Data

Pre-processing and quality control (QC) of WGS data were described elsewhere by Betschart et al. [25]. In brief, QC was continuously performed during the conduct of the study, and approximately 600 samples were monthly processed on the used single NovaSeq 6000 sequencer. During normal operation, data were transferred in batches of approximately 250 subjects, and QC reports were generated for these batches. For pre-processing, we used DRAGEN version 3.8.4 on all samples for mapping and alignment and for single sample variant calling, using the hg38 human reference genome. Pre-processing was done without adapter trimming, and read length was 151 bp. During QC, PC coefficients were estimated using PCs provided by the 1000Genomes project phase 3 data [26]. The first two estimated PCs were used to define a genetically similar European-like population; for details, see Betschart et al. [25]. For the multi-sample calling (also termed joint-calling), the iterative gVCF genotyper was used.

2.5. Measurement of Number of KIV-2 Repeats

The total number of KIV-2 repeats was determined with two approaches, the CNE [19] and the DRAGEN *LPA* Caller [18].

2.5.1. Read Depth-Based Copy Number Estimator (CNE)

The CNE software divides the genome into non-overlapping bins with a size of 100 bp to estimate GC content and read depth for every sample. Read depth was estimated with mosdepth version 0.3.6 [27]. Because CNE produces relative copy numbers, the output was multiplied by six to account for the number of KIV-2 copies within the hg38.

2.5.2. DRAGEN *LPA* Caller

The Illumina DRAGEN *LPA* Caller is available with DRAGEN version 4.2 [18]. Specifically, we used the DRAGEN version 4.2.0-673-g9e903543. This caller counts reads which fall within the KIV-2 repeat. A total of 3000 additional regions, each with a length of 2000 bases are used for normalization. The read counts for all regions are normalized by their length and by GC-content. The number of KIV-2 repeats is derived from the normalized KIV-2 coverage multiplied by six to account for the number of KIV-2 copies within the reference genome.

To determine the allele-specific number of KIV-2 repeats, the DRAGEN *LPA* Caller measures the proportion of reads at two SNVs in linkage disequilibrium (LD) at positions 296 and 1264, which occur in every KIV-2 repeat. The DRAGEN *LPA* Caller reports the number of KIV-2 repeats for both the reference and the alternative allele. The shorter allele is termed KIV-2 short, and the longer allele is named KIV-2 long. We only used the number

of KIV-2 repeats determined by the DRAGEN *LPA* Caller in association analysis because of its ability to determine the number of allele-specific KIV-2 repeats.

2.6. Analysis of the Pentanucleotide Repeat (PNR)

The number of repeats of the PNR were determined with GangSTR version 2.5.0 [28] and ExpansionHunter version 5.0.0 [29], both with default parameters. The genomic location of the PNR is chr6:160665585-160665629 in hg38. Oketch et al. [30] demonstrated good agreement between these two short tandem repeat callers and showed fewer Mendelian inheritance errors of GangSTR in the analysis of Genome in a Bottle trios [31]. The resulting variant calling files (VCFs) were parsed for the REPCN field with bcftools query version 1.18 [32]. If a subject was heterozygous for the number of PNRs, the lower number (shorter allele) was assigned to the variable Allele 1, and the higher number was assigned to Allele 2 [23]. The polyserial correlation between the number of PNRs and the carrier status of KIV-2 SNV 4925G>A was estimated using the polyserial function from the polycor R package version 0.8-1.

2.7. Analysis of the KIV-2 Single Nucleotide Variations (SNVs)

SNVs within the KIV-2 repeat were genotyped with the pipeline termed exome-vntr-nf [13]. The pipeline takes aligned reads as input (BAM file) and extracts the reads that fall within the *LPA* genomic region, i.e., chr6:160530484-160665259 in hg38. These reads are then converted to a FASTQ file and realigned to a single KIV-2 repeat. In the last step, the variants are called with mutserve, a specialized caller originally developed for mitochondrial variant calling [33]. This caller only provides information on carrier status, which is defined as either carrier or non-carrier.

2.8. Statistical Analysis

2.8.1. Descriptive Statistics

Descriptive characteristics of important variables are provided with median and quartiles. For dichotomous variables, absolute and relative frequencies are given. Agreement between the two KIV-2 callers was determined using Pearson correlation, a scatterplot, and a Bland–Altman plot. Agreement between the two PNR callers was determined using Pearson correlation and dotplot.

2.8.2. Genome-Wide Association Study (GWAS)

GWAS for the Lp(a) concentration was performed with REGENIE version 3.2.5.3 [34]. A minor allele frequency (MAF) threshold of 0.005 and a Hardy–Weinberg equilibrium (HWE) threshold of 10^{-9} were used for filtering diallelic SNVs. Subjects showing a kinship coefficient of 3rd cousins or closer ($\geq$0.044; [35]) were partitioned to create an unrelated subset with the highest number of unrelated subjects using the pcairPartition function from the GENESIS R package version 2.28.0 [36].

For association analyses, Lp(a) measurements were log-transformed. An additive genetic model was used, and adjustments were conducted for age, sex, and the first five principal components (PCs). The genome-wide significance threshold was set to 5×10^{-8}. Genotypes of all statistically significant SNVs without missing values were extracted with the R package SeqVarTools version 1.40.0 [37].

2.8.3. Predictive Model for Lp(a) Levels Using Random Forests

Because of the high LD between the different genetic markers in the *LPA* gene, i.e., high multicollinearity, we used random forests by employing the ranger R package [38] to investigate the importance of multiple genetic markers on Lp(a) levels.

In the first step, we estimated two random forests. Random forest 1 (RF1) included all genetic variation plus the allele-specific number of KIV-2 repeats. Random forest 2 (RF2) included all genetic variation plus the total number of KIV-2 repeats. The genetic variation consisted in all SNVs from the GWAS, all SNVs from the specialized SNVs

caller in the KIV-2 repeat, and the number of PNRs estimated by GangSTR. To estimate the variables with a large contribution to the predictive performance, we estimated the conditional predictive impact (CPI) using the CPI R package. We used default values for the CPI function [39]. The CPI tests for conditional independence and measures the variable importance. Specifically, the CPI of a variable provides information on how much the predictive performance deteriorates if the variable is replaced by a non-informative variable. CPIs, their Cis, and test statistics were estimated from 10-fold cross-validation (CV). Variables with a CPI p-value < 0.05 were kept.

The direct comparison of the predictive performance between the number of allele-specific KIV-2 repeats and the total number of KIV-2 repeats was conducted with the following model:

1. Full model: inclusion of all available genetic variation plus
    a. the total number of KIV-2 repeats
    b. the allele-specific number of KIV-2 repeats
2. KIV-2 RF1: inclusion of genetic variation with CPI $p < 0.05$ from RF1 plus
    a. the total number of KIV-2 repeats
    b. the allele-specific number of KIV-2 repeats
3. KIV-2 RF2: inclusion of genetic variation with CPI $p < 0.05$ from RF2 plus
    a. the total number of KIV-2 repeats
    b. the allele-specific number of KIV-2 repeats

For all random forests, we tuned the following hyperparameters using 10-fold nested CV: minimal node size (min.node.size), percentage of included variables in each splitting step (mtry.ratio), and tree depth (max.depth). Default values were used for other parameters. Hyperparameter tuning was conducted with the mlr3 R package (version 0.17.2) [40]. The hyperparameters of the best performing model, i.e., the one with the lowest root mean square error (RMSE), was then used in the CPI calculations.

Since the random forests for the total number of KIV-2 repeats and the allele-specific number of KIV-2 repeats are estimated from the same CV data, they are directly comparable. Our hypotheses were:

- Hypothesis 1: The full model shows the highest predictive performance.
- Hypothesis 2: RF1 is sparser than RF2.
- Hypothesis 3: Model KIV-2 RF1 b performs better than KIV-2 RF1 a.
- Hypothesis 4: Model KIV-2 RF2 b performs better than KIV-2 RF2 a.
- Hypothesis 5: Full models a and b show similar performance, because proxy SNVs from the *LPA* gene should compensate for the additional information for the allele-specific number of KIV-2 repeats.

The global significance level was set to 0.05. To adjust for multiple testing, hypotheses were tested hierarchically, starting with Hypothesis 1. For Hypothesis 2, no formal statistical test was performed. For all models, we forced sex and age to be available for splitting in all trees and all splitting steps used the ranger option always.split.variable [38]. If the final model does not contain age or sex, the coefficient of determination (R^2) obtained from this model can be interpreted as locus-specific heritability. R^2 was estimated for all models.

2.8.4. Software and Hardware

The statistical analysis was performed with R version 4.3.2 [41] on our on-premises HPC equipped with 4 computing nodes. Each node is equipped with 2 AMD EPYC 7742 CPUs and a total of 2TB of RAM. The entire R workflow was run with targets version 1.3.2 [42]. Plots were generated with ggplot2 version 3.4.4 [43]. The code for the analyses is provided as a supplement.

3. Results

3.1. Study Characteristics

Characteristics of GENESIS-HD subjects with available Lp(a) measurements are displayed in Table 1. Figure 2 shows the Lp(a) concentration per binned total number of KIV-2 repeats of GENESIS-HD subjects with available Lp(a) measurements. Lp(a) concentration per binned number of KIV-2 repeats on the short and long allele are provided in Figures S1 and S2, respectively.

Table 1. Characteristics of GENESIS-HD subjects with available Lp(a) measurements. Continuous characteristics are provided as median and interquartile range (IQR), dichotomous variables as number of subjects and percentage.

Demographics	n = 4861
Age (years)	63 (56, 70)
Sex (male)	2425 (49.9%)
Lp(a) (nmol/L)	16.7 (7.50, 59.0)
Total number of KIV-2 repeats	39.0 (34.1, 43.9)
Number of KIV-2 repeats long allele	22.7 (20.0, 25.0)
Number of KIV-2 repeats short allele	16.6 (13.4, 19.9)
Allele-specific number of KIV-2 repeats available	2274 (47.0%)
4925G>A carrier	1124 (23.1%)
rs41272114 carrier	283 (5.8%)

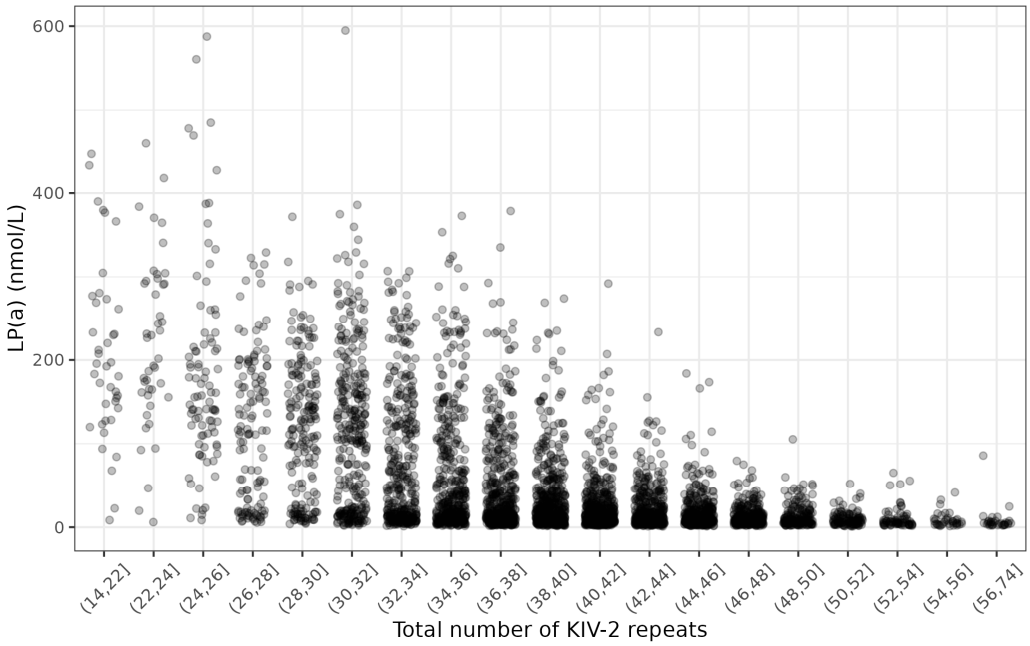


Figure 2. Scatterplot for Lp(a) and total number of KIV-2 repeats (n = 4861). KIV-2 CNs were binned so that there were at least 30 samples in each bin. Horizontal jittering was added to prevent overlapping between points.

3.2. Agreement between Specialized Variant Callers

Figure 3 displays the scatterplot for the total number of KIV-2 repeats estimated by the DRAGEN *LPA* Caller and the CNE. The correlation was 0.9966 (CI 0.9965–0.9968). Figure S3 shows the corresponding Bland–Altman plot.

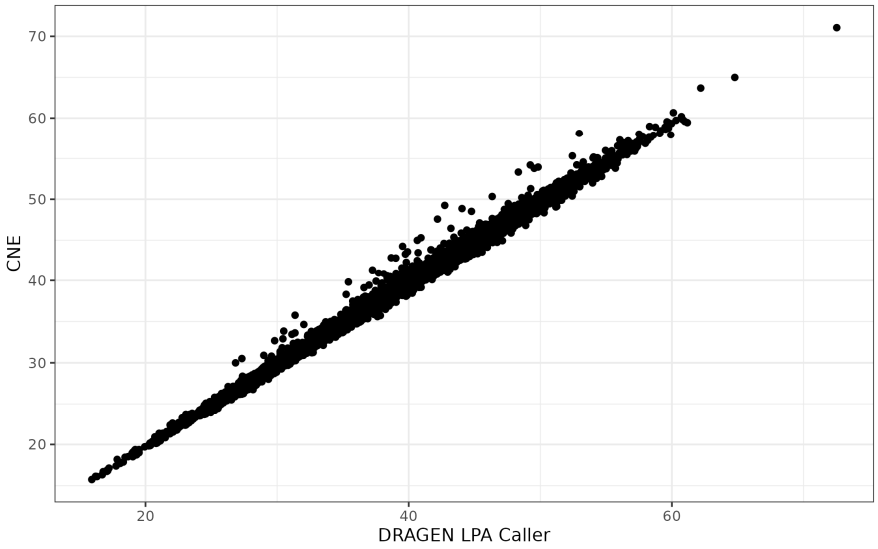


Figure 3. Scatterplot for total number of KIV-2 repeats estimated by the DRAGEN *LPA* Caller and read depth-based CN estimator (CNE) (n = 8351).

Both GangSTR and ExpansionHunter successfully estimated the genotypes of the PNR in the promoter region of the *LPA* gene on all 8351 subjects. The Pearson correlation between the callers was 0.9938 (CI 0.9936–0.9941) for allele 1 and slightly higher for allele 2 (r = 0.9961, CI 0.9960–0.9963). Figure S4 shows the dotplot for the number of subjects per genotype. In agreement with [23], we observed a high LD between the number of PNR and 4925G>A with a polyserial correlation of 0.991.

3.3. Agreement of Allele-Specific Number of KIV-2 Repeats with Results from 1000 Genomes Project

We compared the number of KIV-2 repeats of the GENESIS-HD dataset with the 1000 Genomes dataset (1KGP) analyzed by Behera et al. [18]. In 3925 out of the 8351 subjects (47.0%), the DRAGEN *LPA* Caller determined the allele-specific number of KIV-2 repeats. Excellent agreement between the GENESIS-HD dataset and the European ancestral population (633 samples) was observed (Figure 4).

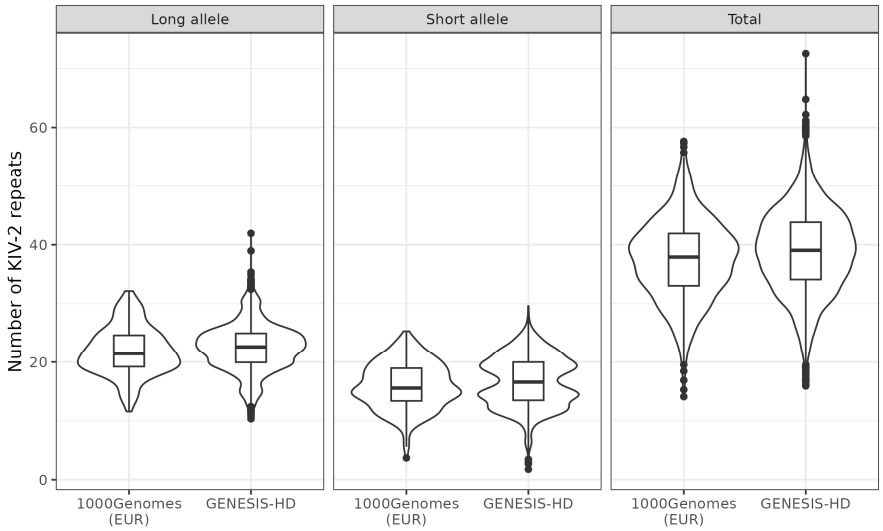


Figure 4. Violin and boxplots of number of KIV-2 repeats for GENESIS-HD samples and EUR-like samples (EUR) from 1000 Genomes project phase 3 (1KGP) [26] (n = 633). For the total number of KIV-2 repeats, 8351 subjects were included. For the number of KIV-2 repeats on the short and long allele, 3925 subjects were included.

3.4. Genome-Wide Association Study for Lp(a) Concentration

Filtering and quality control (QC) left us with a total of 10.3 million (10,334,912) diallelic SNVs in 4803 unrelated subjects. The only genome-wide significant region for Lp(a) concentration was the *LPA* gene (Figure S5). A total of 855 SNVs were found to be statistically significant in the *LPA* gene (hg38: chr6:160531482-160664275) with a padding of 500 kb (Table S1). The lead SNV was rs56393506 (MAF: 0.16; per-allele change in Lp(a) concentration: 1.065 [95% CI 0.997–1.132]; $p = 1.74 \times 10^{-211}$; for locus zoom plot, see Figure S6). The scaled genomic inflation factor was 0.995.

3.5. Genetic Variants Selected by Conditional Predictive Impact (CPI)

Four and six variables were selected by conditional predictive impact for the allele-specific KIV-2 random forest model RF1 and the total KIV-2 random forest model RF2, respectively (Table 2). In both models, the number of KIV-2 repeats had the highest conditional predictive impact. In the allele-specific KIV-2 model, the shorter allele had a stronger effect on Lp(a) levels than the longer allele. Two SNVs, 4925G>A and rs41272114, had a significant conditional predictive impact in both models, and the contingency table is provided in Table 3.

Table 2. Variables selected by conditional predictive impact (CPI) for both random forest models RF1 and RF2. CPIs, their 95% confidence intervals (95% CI) and corresponding *p*-values are displayed. CPIs and 95% CIs were multiplied by 10^3.

	RF1 (Allele-Specific)			RF2 (Total)		
Variable	**CPI**	**95% CI**	***p*-Value**	**CPI**	**95% CI**	***p*-Value**
Short KIV-2	97.7	77.8–117.7	5.5×10^{-16}	–	–	–
Long KIV-2	22.6	14.3–30.9	3.8×10^{-6}	–	–	–
Total KIV-2	–	–	–	58.9	43.6–70.1	1.2×10^{-10}
4925G>A	9.9	4.3–15.5	1.8×10^{-3}	8.7	2.0–15.3	1.6×10^{-2}
rs41272114	3.6	1.0–6.3	1.3×10^{-2}	4.4	0.6–8.1	2.7×10^{-2}
rs9347465	–	–	–	9.6×10^{-4}	1.2×10^{-6}–1.9×10^{-3}	4.9×10^{-2}
rs2489959	–	–	–	8.6×10^{-4}	5.6×10^{-6}–1.7×10^{-3}	4.9×10^{-2}
rs2457567	–	–	–	8.4×10^{-4}	3.2×10^{-5}–1.6×10^{-3}	4.4×10^{-2}

Table 3. Contingency table for SNV 4925G>A carrier status and rs41272114.

	rs41272114			
4925G>A	**CC**	**CT**	**TT**	**Total**
GG	3490 (71.8%)	244 (5.0%)	3 (0.1%)	3737 (76.9%)
GA/AA	1088 (22.4%)	36 (0.7%)	0	1124 (23.1%)
Total	4578 (94.2%)	280 (5.8%)	3 (0.1%)	4861 (100%)

Non-carriers of both SNVs (4925GG and rs41272114CC; n = 3488) had a median Lp(a) concentration of 20.3 nmol/L (7.3, 83.0). Subjects carrying at least one A allele at SNV 4925 and genotype CC at rs41272114 (n = 1088) showed a lower Lp(a) concentration of 15.0 nmol/L (9.7–29.3). The lowest median Lp(a) concentration of 4.8 nmol/L (2.9–15.8) was observed in subjects carrying variant alleles at rs41272114, but not carrying an A allele at SNV 4925 (n = 247). Individuals carrying variations at both SNVs (n = 36) had an Lp(a) concentration of 8.4 nmol/L (6.3–12.7). It should be noted that the number of double carriers is low. The Lp(a) concentration of the genotype combination and short and total number of KIV-2 repeats is displayed in Figure 5.

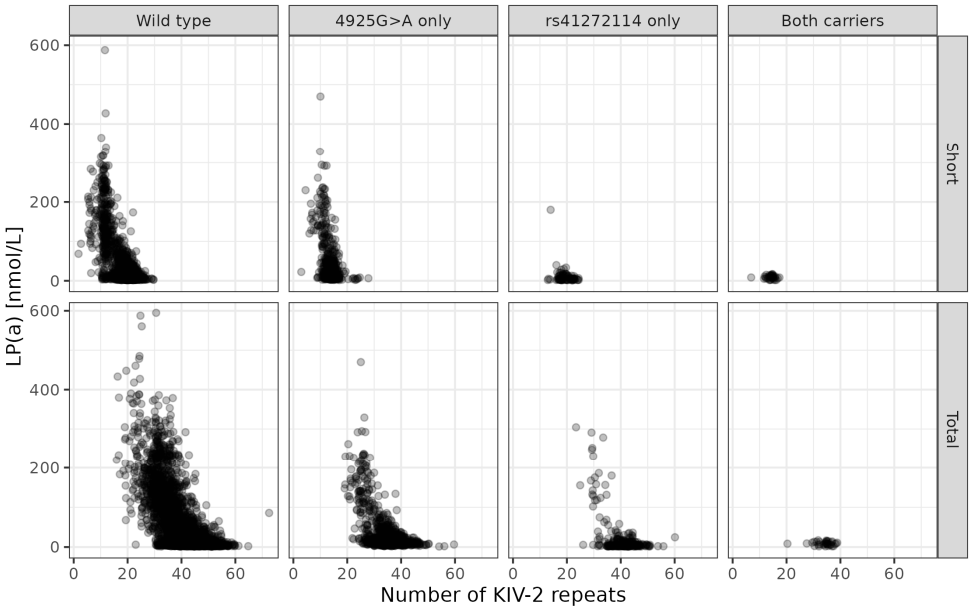


Figure 5. Lp(a) concentration by number of KIV-2 repeats for all single nucleotide variation combinations of 4925G>A and rs41272114 carriers of the T allele. (**Upper panels**): number of KIV-2 repeats of the short allele. (**Lower panels**): total number of KIV-2 repeats.

Only three subjects were homozygous TT at rs41272114; these three subjects did not harbor an A allele at SNV 4925. These subjects had Lp(a) concentrations of 2.2, 2.3, and 3.0 nmol/L.

The three SNVs from model RF2 rs9347465, rs2489959, and rs2457567 were neither found in the GWAS catalogue nor by a Medline search. As expected, the PNR did not have a significantly conditional predictive impact.

3.6. Comparison of Predictive Performances for Lp(a) Concentrations

Mean and standard deviation (SD) of the difference in R^2 between allele-specific number and total number of KIV-2 repeats are displayed in Table 4. Specifically, we estimated the difference in R^2 between the allele-specific number and total number of KIV-2 repeats. The full model showed the highest predictive performance for both the total number of KIV-2 repeats (R^2 = 0.6855, CI 0.6109–0.7601) and the allele-specific number of KIV-2 repeats for the total number of KIV-2 repeats (R^2 = 0.6953, CI 0.6414–0.7492). The performance of the full model was significantly better than that of RF1 and RF2 (both $p < 0.001$; Table S2). This confirms Hypothesis 1. For model RF1, four variables were selected, compared to six variables for model RF2. RF1 was thus sparser than RF2, which confirms Hypothesis 2.

Table 4. Coefficient of determination (R^2) and 95% confidence interval (CI) of different random forest (RF) models for Lp(a) concentrations.

Model	KIV-2	R^2	CI	Difference	CI of Difference	*p*-Value
Full	Total	0.6855	0.6109–0.7601	0.0098	−0.0163–0.0359	0.4620
	Allele-specific	0.6953	0.6414–0.7492			
RF1	Total	0.4290	0.3086–0.5494	0.0540	0.0116–0.0965	0.0126
	Allele-specific	0.4830	0.3806–0.5855			
RF2	Total	0.4788	0.3978–0.5599	0.0413	0.0000–0.0826	0.0496
	Allele-specific	0.5201	0.4054–0.6350			

Full model: All available genetic variation; RF1: inclusion of genetic variation with conditional predictive impact (CPI) $p < 0.05$ from allele-specific model, when allele-specific numbers of KIV-2 repeats were available; RF2: inclusion of genetic variation with CPI $p < 0.05$ from total model, when total number of KIV-2 repeats was available.

Model RF1 b showed a significantly better predictive performance than model RF1 a ($p = 0.0126$), which confirms Hypothesis 3. Similarly, model RF2 b outperformed model RF2 a ($p = 0.0496$), which confirms Hypothesis 4. Finally, no statistically significant difference was observed between the total and allele-specific KIV-2 repeat models in the full model ($p = 0.4620$), which is in agreement with Hypothesis 5.

4. Discussion

In this study, we performed a comprehensive analysis of the genetic variation within the *LPA* gene by using data from whole genome short-read sequencing. Key to these analyses was the availability of specialized callers to estimate the number of KIV-2 repeats. There was excellent agreement between the DRAGEN *LPA* Caller and the CNE, which were both publicly released in 2023. The accuracy of the DRAGEN *LPA* Caller was previously confirmed by long-read sequencing [18], and the present analysis indirectly confirms the validity of the CNE caller for the *LPA* gene.

A weakness of both callers is the estimation of the number of allele-specific KIV-2 repeats. This information is important because the number of allele-specific KIV-2 repeats leads to better predictive performance for Lp(a) concentration. Of note, the CNE is able to estimate the total number of KIV-2 repeats only. In contrast, the DRAGEN *LPA* Caller is able to estimate the number of allele-specific KIV-2 repeats, depending on the genotypes of two SNVs within the KIV-2 region. In this study, the DRAGEN *LPA* Caller provided the number of allele-specific KIV-2 repeats in 47% of all subjects, which were of European ancestry. This percentage varies between subjects of different ancestry groups. For example, in AMR-like subjects, the percentage was the lowest at 38%, and highest in AFR-like individuals at 52% [18].

There was a high agreement between the two callers, ExpansionHunter and GangSTR, for estimating the number of PNRs. However, the genetic information of the PNR is not important for predicting Lp(a) concentrations because almost all the genetic information of the multiallelic PNR is captured by the dichotomous SNV 4925G>A, located within the KIV-2 repeat (correlation = 0.997). This is in agreement with previously published results [23].

The results of our GWAS confirmed previous reports that variants associated with Lp(a) concentrations are primarily located in the *LPA* gene region [12]. We identified 855 significant SNVs surrounding the *LPA* gene. The lead SNV rs56393506 has been previously reported as being associated with increased Lp(a) concentration [44].

Predicting Lp(a) concentrations based on genetic variation is challenging because of high LD in the *LPA* gene. This high multicollinearity between genetic variation invalidates classical regression models, such as linear regression or quantile regression. In contrast, some of the machine learning methods, such as random forests, are able to adequately deal with a high LD. The random forest models revealed that only a small number of genetic variations had a significant impact on the predictions of the Lp(a) concentrations. This is not surprising, because the *LPA* gene shows large LD blocks [45]. The full model incorporating all genetic variation from the *LPA* gene had a substantially higher predictive performance of Lp(a) concentrations when compared to models that covered the most important genetic markers (RF1 and RF2). Indeed, the locus-specific heritability (R) of the full model was 0.83 (CI 0.82–0.84), while it was only 0.68 (CI 0.65–0.70) and 0.71 (CI 0.69–0.72) for RF1 and RF2, respectively. The locus-specific heritability of the full model was thus in the same magnitude as that of other studies [14,15].

In line with the literature [14,15], the number of KIV-2 repeats had the highest conditional predictive impact, with the number of allele-specific KIV-2 repeats being more predictive than the total number of KIV-2 repeats. The two most important SNVs for predicting Lp(a) concentrations were 4925G>A and rs41272114, which confirms other work [14]. SNV 4925G>A is located on a splice site within the KIV-2 repeat and leads to a decrease in Lp(a) concentration [13,15,46]. rs41272114 is also located on a splice site, and T alleles at this SNV result in null alleles [12–14]. Such null alleles lead to extremely low Lp(a)

concentrations, which we also observed in this study. Three subjects were homozygous TT at rs41272114. The highest Lp(a) concentration in these individuals was 3.0 nmol/L, which is in the bottom five percent of the Lp(a) concentration levels among all subjects.

A limitation of our study is that analyses were based on a specific assay to measure the Lp(a) concentration. Specifically, we used the Roche assay, which reports the molar Lp(a) concentration (nmol/L). Various studies have reported differences between Lp(a) assays [47,48]. Furthermore, we restricted our study to EUR-like subjects only. Future research should investigate the predictive performance of all genetic variation within the *LPA* gene in different ancestry groups. Furthermore, determining not only the carrier status of SNVs within the KIV-2 repeats but also the genotype could lead to a better understanding of the complex *LPA* gene.

5. Conclusions

Technological advancements allow the accurate determination of the genetic variation within the complex *LPA* gene from short-read WGS data. The number of allele-specific KIV-2 repeats was more important for predicting Lp(a) protein levels than the total number of KIV-2 repeats. However, with current callers, the allele-specific number of KIV-2 repeats can only be determined in approximately 50% of European-like samples. The *LPA* gene-specific heritability of Lp(a) protein levels was approximately 70% in this population-based study. Due to high linkage disequilibrium between the PNR and the SNV 4925G>A, the PNR was not important for predicting Lp(a) protein levels. In contrast, SNVs 4925G>A and rs41272114C>T were of high importance in addition to the number of KIV-2 repeats.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/biomed4020013/s1, Figure S1: Scatterplot for LP(a) and number of KIV-2 repeats on the short allele (n = 4861). Number of KIV-2 repeats were binned so that there are at least 30 samples in each bin. Horizontal jittering was added to prevent overlapping between points; Figure S2: Scatterplot for LP(a) and number of KIV-2 repeats on the long allele (n = 4861). Number of KIV-2 repeats were binned so that there are at least 30 samples in each bin. Horizontal jittering was added to prevent overlapping between points; Figure S3: Bland-Altman plot with mean and 95% confidence interval for total number of KIV-2 repeats estimated by the DRAGEN *LPA* Caller and read depth based CN estimator (CNE) (n = 8351); Figure S4: Dotplot per number of subjects per genotype estimated by ExpansionHunter and GangSTR; Figure S5: Manhattan plot from genome-wide association analysis (GWAS) for LP(a) concentration; Figure S6: Locus-zoom plot for the *LPA* gene with 500kb padding; Table S1: Statistically significant SNVs in the *LPA* gene; Table S2: Performance statistics of the full random forest model and random forest models RF1 and RF2; R zip code S1: R code for analyses conducted in this work; instructions are provided README.md.

Author Contributions: Conceptualization, R.O.B. and A.Z.; data curation, R.O.B., L.G., R.T. and T.Z.; formal analysis, R.O.B.; funding acquisition, S.B., R.T., T.Z. and A.Z.; investigation, S.B., T.Z., R.T. and A.Z.; methodology, R.O.B. and A.Z.; project administration, A.Z.; resources, S.B., R.T., T.Z. and A.Z.; software, R.O.B., G.K., P.G., M.R. and S.S.; supervision, A.Z.; visualization R.O.B. and G.K.; writing—original draft, R.O.B. and A.Z.; writing—review and editing, All. All authors have read and agreed to the published version of the manuscript.

Funding: We gratefully acknowledge funding of the GENESIS-HD study by the Kühne Foundation and the measurement of Lp(a) in the HCHS by Amgen.

Institutional Review Board Statement: This study did not require consultation by an ethics committee. The ethics committee of the Landesärztekammer Hamburg (State of Hamburg Chamber of Medical Practitioners, PV5131 did not have objections against the conduct of the Hamburg City Health Study, see reference [24].

Informed Consent Statement: Written informed consent was obtained from all study participants.

Data Availability Statement: The data used in this study may not be shared due to privacy issues. The targets file used for the analysis is available in the Supplementary Materials.

Acknowledgments: We are grateful to Patricia Bartoschek, Satya Bhowmik, Anna Lena Engels, Tim Hartmann, Yumi Hartmann, Lilia Kisselmann, Anna-Lena Post, and René Riedl for excellent laboratory work with DNA extraction, data management, and quality control.

Conflicts of Interest: M.R. is an employee of Illumina. R.T. holds a professorship in clinical cardiology at the University Medical Center Hamburg-Eppendorf, supported by the Kühne Foundation, and reports research support from the German Center for Cardiovascular Research (DZHK), the Joachim Herz Foundation, the Swiss National Science Foundation (Grant No. P300PB_167803), and the Swiss Heart Foundation as well as speaker honoraria/consulting honoraria from Abbott, Amgen, Astra Zeneca, Psyros, Roche, Siemens, Singulex, and Thermo Scientific BRAHMS. T.Z. is supported by the German Research Foundation, the EU Horizon 2020 programme, the EU ERANet and ERAPreMed Programmes, the German Centre for Cardiovascular Research (DZHK, 81Z0710102), and the German Ministry of Education and Research. S.B., R.T., T.Z., and A.Z. are listed as co-inventors of an international patent on the use of a computing device to estimate the probability of myocardial infarction (International Publication Number WO2022043229A1). R.T. and T.Z. are shareholders of the ART-EMIS Hamburg GmbH. A.Z. is the scientific director and R.O.B. and G.K. are employees of Cardio-CARE, which is a shareholder of the ART-EMIS Hamburg GmbH.

References

1. Koschinsky, M.L.; Marcovina, S.M. Structure-function relationships in apolipoprotein(a): Insights into lipoprotein(a) assembly and pathogenicity. *Curr. Opin. Lipidol.* **2004**, *15*, 167–174. [CrossRef] [PubMed]
2. Berg, K. A new serum type system in man—The Ld system. *Vox Sang.* **1965**, *10*, 513–527. [CrossRef] [PubMed]
3. Kronenberg, F.; Utermann, G. Lipoprotein(a): Resurrected by genetics. *J. Intern. Med.* **2013**, *273*, 6–30. [CrossRef] [PubMed]
4. Schmidt, K.; Noureen, A.; Kronenberg, F.; Utermann, G. Structure, function, and genetics of lipoprotein (a). *J. Lipid Res.* **2016**, *57*, 1339–1359. [CrossRef] [PubMed]
5. Kamstrup, P.R.; Tybjaerg-Hansen, A.; Steffensen, R.; Nordestgaard, B.G. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *JAMA* **2009**, *301*, 2331–2339. [CrossRef] [PubMed]
6. McLean, J.W.; Tomlinson, J.E.; Kuang, W.J.; Eaton, D.L.; Chen, E.Y.; Fless, G.M.; Scanu, A.M.; Lawn, R.M. cDNA sequence of human apolipoprotein(a) is homologous to plasminogen. *Nature* **1987**, *330*, 132–137. [CrossRef] [PubMed]
7. Hancock, M.A.; Boffa, M.B.; Marcovina, S.M.; Nesheim, M.E.; Koschinsky, M.L. Inhibition of plasminogen activation by lipoprotein(a): Critical domains in apolipoprotein(a) and mechanism of inhibition on fibrin and degraded fibrin surfaces. *J. Biol. Chem.* **2003**, *278*, 23260–23269. [CrossRef] [PubMed]
8. Nordestgaard, B.G.; Chapman, M.J.; Ray, K.; Boren, J.; Andreotti, F.; Watts, G.F.; Ginsberg, H.; Amarenco, P.; Catapano, A.; Descamps, O.S.; et al. Lipoprotein(a) as a cardiovascular risk factor: Current status. *Eur. Heart J.* **2010**, *31*, 2844–2853. [CrossRef]
9. Brown, M.S.; Goldstein, J.L. Plasma lipoproteins: Teaching old dogmas new tricks. *Nature* **1987**, *330*, 113–114. [CrossRef]
10. Gaubatz, J.W.; Ghanem, K.I.; Guevara, J.; Nava, M.L.; Patsch, W.; Morrisett, J.D. Polymorphic forms of human apolipoprotein[a]: Inheritance and relationship of their molecular weights to plasma levels of lipoprotein[a]. *J. Lipid Res.* **1990**, *31*, 603–613. [CrossRef]
11. Jawi, M.M.; Frohlich, J.; Chan, S.Y. Lipoprotein(a) the Insurgent: A New Insight into the Structure, Function, Metabolism, Pathogenicity, and Medications Affecting Lipoprotein(a) Molecule. *J. Lipids* **2020**, *2020*, 3491764. [CrossRef] [PubMed]
12. Mack, S.; Coassin, S.; Rueedi, R.; Yousri, N.A.; Seppala, I.; Gieger, C.; Schonherr, S.; Forer, L.; Erhart, G.; Marques-Vidal, P.; et al. A genome-wide association meta-analysis on lipoprotein (a) concentrations adjusted for apolipoprotein (a) isoforms. *J. Lipid Res.* **2017**, *58*, 1834–1844. [CrossRef] [PubMed]
13. Coassin, S.; Erhart, G.; Weissensteiner, H.; Eca Guimaraes de Araujo, M.; Lamina, C.; Schonherr, S.; Forer, L.; Haun, M.; Losso, J.L.; Kottgen, A.; et al. A novel but frequent variant in *LPA* KIV-2 is associated with a pronounced Lp(a) and cardiovascular risk reduction. *Eur. Heart J.* **2017**, *38*, 1823–1831. [CrossRef] [PubMed]
14. Coassin, S.; Kronenberg, F. Lipoprotein(a) beyond the kringle IV repeat polymorphism: The complexity of genetic variation in the *LPA* gene. *Atherosclerosis* **2022**, *349*, 17–35. [CrossRef] [PubMed]
15. Mukamel, R.E.; Handsaker, R.E.; Sherman, M.A.; Barton, A.R.; Zheng, Y.; McCarroll, S.A.; Loh, P.R. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **2021**, *373*, 1499–1505. [CrossRef] [PubMed]
16. Mahmoud, M.; Harting, J.; Corbitt, H.; Chen, X.; Jhangiani, S.N.; Doddapaneni, H.; Meng, Q.; Han, T.; Lambert, C.; Zhang, S.; et al. Closing the gap: Solving complex medically relevant genes at scale. *medRxiv* **2024**. [CrossRef] [PubMed]
17. Warburton, P.E.; Sebra, R.P. Long-Read DNA Sequencing: Recent Advances and Remaining Challenges. *Annu. Rev. Genomics. Hum. Genet.* **2023**, *24*, 109–132. [CrossRef] [PubMed]
18. Behera, S.; Belyeu, J.R.; Chen, X.; Paulin, L.F.; Nguyen, N.Q.H.; Newman, E.; Mahmoud, M.; Menon, V.K.; Qi, Q.; Joshi, P.; et al. Identification of allele-specific KIV-2 repeats and impact on Lp(a) measurements for cardiovascular disease risk. *bioRxiv* **2023**. [CrossRef] [PubMed]
19. Garg, P.; Jadhav, B.; Lee, W.; Rodriguez, O.L.; Martin-Trujillo, A.; Sharp, A.J. A phenome-wide association study identifies effects of copy-number variation of VNTRs and multicopy genes on multiple human traits. *Am. J. Hum. Genet.* **2022**, *109*, 1065–1076. [CrossRef]

20. Lu, W.; Cheng, Y.C.; Chen, K.; Wang, H.; Gerhard, G.S.; Still, C.D.; Chu, X.; Yang, R.; Parihar, A.; O'Connell, J.R.; et al. Evidence for several independent genetic variants affecting lipoprotein (a) cholesterol levels. *Hum. Mol. Genet.* **2015**, *24*, 2390–2400. [CrossRef]
21. Mooser, V.; Mancini, F.P.; Bopp, S.; Petho-Schramm, A.; Guerra, R.; Boerwinkle, E.; Muller, H.-J.; H.Hobbs, H. Sequence polymorphisms in the apo(a) gene associated with specific levels of Lp(a) in plasma. *Hum. Mol. Genet.* **1995**, *4*, 173–181. [CrossRef] [PubMed]
22. Prins, J.; Leus, F.R.; Bouma, B.N.; van Rijn, H.J. The identification of polymorphisms in the coding region of the apolipoprotein (a) gene–association with earlier identified polymorphic sites and influence on the lipoprotein (a) concentration. *Thromb. Haemost.* **1999**, *82*, 1709–1717. [PubMed]
23. Grüneis, R.; Weissensteiner, H.; Lamina, C.; Schönherr, S.; Forer, L.; Di Maio, S.; Streiter, G.; Peters, A.; Gieger, C.; Kronenberg, F.; et al. The kringle IV type 2 domain variant 4925G>A causes the elusive association signal of the *LPA* pentanucleotide repeat. *J. Lipid Res.* **2022**, *63*, 100306. [CrossRef] [PubMed]
24. Jagodzinski, A.; Johansen, C.; Koch-Gromus, U.; Aarabi, G.; Adam, G.; Anders, S.; Augustin, M.; der Kellen, R.B.; Beikler, T.; Behrendt, C.A.; et al. Rationale and design of the Hamburg City Health Study. *Eur. J. Epidemiol.* **2020**, *35*, 169–181. [CrossRef] [PubMed]
25. Betschart, R.O.; Riccio, C.; Aguilera-Garcia, D.; Blankenberg, S.; Guo, L.; Moch, H.; Seidl, D.; Solleder, H.; Thalén, F.; Thiéry, A.; et al. Biostatistical Aspects of Whole Genome Sequencing Studies: Preprocessing and Quality Control. *Biom. J.* **2024**, *66*, e202300278.
26. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef] [PubMed]
27. Pedersen, B.S.; Quinlan, A.R. Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics* **2018**, *34*, 867–868. [CrossRef]
28. Mousavi, N.; Shleizer-Burko, S.; Yanicky, R.; Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **2019**, *47*, e90. [CrossRef]
29. Dolzhenko, E.; van Vugt, J.; Shaw, R.J.; Bekritsky, M.A.; van Blitterswijk, M.; Narzisi, G.; Ajay, S.S.; Rajan, V.; Lajoie, B.R.; Johnson, N.H.; et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **2017**, *27*, 1895–1903. [CrossRef]
30. Oketch, J.W.; Wain, L.V.; Hollox, E.J. A comparison of software for analysis of rare and common short tandem repeat (STR) variation using human genome sequences from clinical and population-based samples. *bioRxiv* **2023**. [CrossRef]
31. Zook, J.M.; McDaniel, J.; Olson, N.D.; Wagner, J.; Parikh, H.; Heaton, H.; Irvine, S.A.; Trigg, L.; Truty, R.; McLean, C.Y.; et al. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **2019**, *37*, 561–566. [CrossRef]
32. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, giab008. [CrossRef] [PubMed]
33. Weissensteiner, H.; Forer, L.; Fuchsberger, C.; Schöpf, B.; Kloss-Brandstätter, A.; Specht, G.; Kronenberg, F.; Schönherr, S. mtDNA-Server: Next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res.* **2016**, *44*, W64–W69. [CrossRef]
34. Mbatchou, J.; Barnard, L.; Backman, J.; Marcketta, A.; Kosmicki, J.A.; Ziyatdinov, A.; Benner, C.; O'Dushlaine, C.; Barber, M.; Boutkov, B.; et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **2021**, *53*, 1097–1103. [CrossRef] [PubMed]
35. Manichaikul, A.; Mychaleckyj, J.C.; Rich, S.S.; Daly, K.; Sale, M.; Chen, W.M. Robust relationship inference in genome-wide association studies. *Bioinformatics* **2010**, *26*, 2867–2873. [CrossRef] [PubMed]
36. Gogarten, S.M.; Sofer, T.; Chen, H.; Yu, C.; Brody, J.A.; Thornton, T.A.; Rice, K.M.; Conomos, M.P. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **2019**, *35*, 5346–5348. [CrossRef] [PubMed]
37. Gogarten, S.M.; Zheng, X.; Stilp, A. *SeqVarTools: Tools for Variant Data*, R package version 1.42.0; 2023. [CrossRef]
38. Wright, M.N.; Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
39. Watson, D.S.; Wright, M.N. Testing conditional independence in supervised learning algorithms. *Mach. Learn* **2021**, *110*, 2107–2129. [CrossRef]
40. Lang, M.; Binder, M.; Richter, J.; Schratz, P.; Pfisterer, F.; Coors, S.; Au, Q.; Casalicchio, G.; Kotthoff, L.; Bischl, B. mlr3: A modern object-oriented machine learning framework in R. *J. Open Source Softw.* **2019**, *4*, 1903. [CrossRef]
41. R Core Team. R: A Language and Environment for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 13 March 2024).
42. Landau, W.M. The targets R package: A dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *J. Open Source Softw.* **2019**, *6*, 2959. [CrossRef]
43. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer International Publishing: Basel, Switzerland, 2016. [CrossRef]
44. Solomon, T.; Smith, E.N.; Matsui, H.; Braekkan, S.K.; Consortium, I.; Wilsgaard, T.; Njolstad, I.; Mathiesen, E.B.; Hansen, J.B.; Frazer, K.A. Associations between common and rare exonic genetic variants and serum levels of 20 cardiovascular-related proteins: The Tromsø Study. *Circ. Cardiovasc. Genet.* **2016**, *9*, 375–383. [CrossRef]

45. Ronald, J.; Rajagopalan, R.; Cerrato, F.; Nord, A.S.; Hatsukami, T.; Kohler, T.; Marcovina, S.; Heagerty, P.; Jarvik, G.P. Genetic variation in LPAL2, LPA, and PLG predicts plasma lipoprotein(a) level and carotid artery disease risk. *Stroke* **2011**, *42*, 2–9. [CrossRef] [PubMed]
46. Schachtl-Riess, J.F.; Kheirkhah, A.; Grüneis, R.; Di Maio, S.; Schoenherr, S.; Streiter, G.; Losso, J.L.; Paulweber, B.; Eckardt, K.U.; Köttgen, A.; et al. Frequent *LPA* KIV-2 variants lower lipoprotein(a) concentrations and protect against coronary artery disease. *J. Am. Coll. Cardiol.* **2021**, *78*, 437–449. [CrossRef] [PubMed]
47. Kraft, H.G.; Lingenhel, A.; Pang, R.W.; Delport, R.; Trommsdorff, M.; Vermaak, H.; Janus, E.D.; Utermann, G. Frequency distributions of apolipoprotein(a) kringle IV repeat alleles and their effects on lipoprotein(a) levels in Caucasian, Asian, and African populations: The distribution of null alleles is non-random. *Eur. J. Hum. Genet.* **1996**, *4*, 74–87. [CrossRef]
48. van der Hoek, Y.Y.; Wittekoek, M.E.; Beisiegel, U.; Kastelein, J.J.; Koschinsky, M.L. The apolipoprotein(a) kringle IV repeats which differ from the major repeat kringle are present in variably-sized isoforms. *Hum. Mol. Genet.* **1993**, *2*, 361–366. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.