frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 29 April 2014
doi: 10.3389/fgene.2014.00105

# Assessing the functional consequence of loss of function variants using electronic medical record and large-scale genomics consortium efforts

***Patrick Sleiman[1,2] *, Jonathan Bradfield[1], Frank Mentch[1], Berta Almoguera[1], John Connolly[1] and Hakon Hakonarson[1,2] ****

[1] *Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, Philadelphia, PA, USA*
[2] *Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

***Edited by:***
*Mariza De Andrade, Mayo Clinic, USA*

***Reviewed by:***
*Lijun Ma, Wake Forest University Health Sciences, USA*
*Hui-Qi Qu, The University of Texas School of Public Health, USA*

****Correspondence:***
*Patrick Sleiman and Hakon Hakonarson, Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104-4318, USA*
*e-mail: sleimanp@email.chop.edu; hakonarson@email.chop.edu*

Estimates from large scale genome sequencing studies indicate that each human carries up to 20 genetic variants that are predicted to results in loss of function (LOF) of protein-coding genes. While some are known disease-causing variants or common, tolerated, LOFs in non-essential genes, the majority remain of unknown consequence. We explore the possibility of using imputed GWAS data from large biorepositories such as the electronic medical record and genomics (eMERGE) consortium to determine the effects of rare LOFs. Here, we show that two hypocholesterolemia-associated LOF mutations in the *PCSK9* gene can be accurately imputed into large-scale GWAS datasets which raises the possibility of assessing LOFs through genomics-linked medical records.

**Keywords: loss of function (LOF), imputation, PCSK9, eMERGE, biorepository**

## INTRODUCTION

Complete loss of function (LOF) variants are defined as variants expected to correlate with complete LOF of affected transcripts; i.e., nonsense mutations, splice site mutations, and insertion/deletion (indel) variants that result in downstream premature stop codons, or larger deletions removing either the first exon or more than 50% of the protein-coding sequence of the affected transcript (MacArthur et al., 2012). Partial LOF variants reduce gene activity but do not ablate it completely.

Data from the 1000 genomes project (1KGP), a large scale human genome sequencing study of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing data indicates that on average individuals carry ∼150 LOFs (Genomes Project Consortium et al., 2012). However, as detailed in **Table 1**, the majority of LOFs are common (>5%) and are distributed across a very small number (100–200) of genes. Genes containing common LOFs are strongly enriched for functional categories related to olfactory reception that are apparently unessential and do not result in any severe medical consequence. LOF enriched genes are typically depleted for genes implicated in protein-binding, transcriptional regulation, and anatomical development. Common LOFs are also enriched at the 3′ ends of genes as these mutations escape nonsense-mediated decay and are less subject to purifying natural selection. Finally, at the most highly conserved coding sites, more than 90% of stop-gain and splice-disrupting variants have a frequency below 0.5%. The population frequency of individual LOFs would therefore appear to correlate with their potential to adversely affect human health.

The 1KGP data indicates that each individual carries 10–20 LOF variants with a minor allele frequency (MAF) below 0.5% (**Table 1**). As these LOFs are under purifying selection they are less likely to be present in non-essential genes and at low conservation sites and therefore are likely to present pathological candidates.

The population frequency of rare variants differs considerably compared with common variation. Variants with frequencies above 10% were found in all of the populations studied in the 1KGP (Genomes Project Consortium et al., 2012), albeit with differences in MAF. Low-frequency variants in the 0.5–5% range were also largely shared between ancestral groups with only 17% of variants observed in a single ancestry group. For rare frequency variants with MAFs <0.5%, the majority (53%) were observed in a single population. Population stratification therefore represents a major confounder for rare variant analyzes which would ideally be controlled using principal component analysis from high-density GWAS arrays to select ancestrally matched cases and controls.

As a consequence of their rarity, LOFs will have largely been overlooked in GWAS studies which are best suited to the study of variants with minor alleles >3–5%. However, due to their rarity, very large, GWAS-type sample sets will be necessary to determine phenotypic association.

PCSK9 is expressed primarily in the liver, it is a secreted protein that acts by reducing the amount of low density lipoprotein receptor (LDLR) at the cell surface. Structurally, the PCSK9 protein product is composed a signal peptide, a prodomain, a catalytic domain, and a C-terminal domain. Cleavage of the prodomain is required for PCSK9 maturation and secretion. Cleaved PCSK9 is transported along the secretory pathway, which ultimately

**Table 1 | Loss of function allele counts in 1,092 human genomes across three allele frequency bins.**

| | Allele frequency (%) | | |
|---|---|---|---|
| Variant type | <0.5 | 0.5–5 | >5 |
| Stop-gain | 3.9–10 | 5.3–19 | 24–28 |
| Stop-loss | 1.0–1.2 | 1.0–1.9 | 2.1–2.8 |
| Indel frameshift | 1.0–1.3 | 11–24 | 60–66 |
| Splice site donor | 1.7–3.6 | 2.4–7.2 | 2.6–5.2 |
| Splice site acceptor | 1.5–2.9 | 1.5–4.0 | 2.1–4.6 |

promotes LDLR degradation [for review see (Marais et al., 2012)]. One LOF missense mutation in PCSK9, Q152H, has been shown to impair cleavage and hence inhibit PCSK9 secretion (Mayne et al., 2011). The Q152H LOF mutation was shown to result in a 79% decrease in circulating PCSK9 and a 48% decrease in LDL-C in carriers compared with non-carriers (Mayne et al., 2011). The C679*X* mutation results in a processed, partially-folded protein that remains in the ER and is not secreted. As LDLR is degraded at the cell surface and endosomes, the C679*X* mutant has no activity toward the LDLR because of its inability to leave the ER and traffic to LDLR (Benjannet et al., 2006). R46L is also a LOF PCSK9 mutation, the R46L-PCSK9 undergoes near normal autocatalytic cleavage and is secreted, yet cells expressing the mutant displayed a 16% increase in of cell surface LDLR and a 35% increase in internalized LDL compared with WT-PCSK9, suggesting that R46L causes hypocholesterolemia through a decreased ability to degrade LDLR (Cameron et al., 2006).

Mutations in *PCSK9* were first identified in two French families with hypercholesterolemia that screened negative for mutations in both the LDLR and the apolipoprotein B (apoB) genes (Abifadel et al., 2003). The hypercholesterolemia *PCSK9* mutations were all missense variants that are thought to confer a gain of function as overexpression of *pcsk9* in the liver of mice produces hypercholesterolemia by reducing LDLR numbers (Lambert et al., 2006).

In 2005, causative LOF mutations in *PCSK9* were identified in individuals with low plasma LDL-C levels, the LOF variants were shown to be present in ∼2% of the African–American population but rare in European Americans (<0.1%; Cohen et al., 2005). LOF mutation carriers displayed reduced or no PCSK9 activity, and their plasma LDL-C levels were reduced by 40% compared with non-carriers. Further, coronary heart disease risk in those individuals was reduced by 88% compared to non-carriers (Cohen et al., 2006). This observation sparked interest in the biology of PCSK9 and led to the development of several LDL-reducing drugs (Stein et al., 2012).

While the cost of whole genome and exome sequencing experiments has dropped dramatically with improvements in yield from second generation sequencing technologies, very large scale studies remain prohibitively expensive. For sample sets with existing genotypes from dense whole-genome arrays, genotype imputation presents a viable alternative to direct sequencing. Data generated from large sequencing projects such as the 1KGP (Genomes Project Consortium et al., 2012) and the NHLBI exome sequencing project (ESP; Tennessen et al., 2012) is phased (Delaneau et al., 2012) and the haplotypes can be used as reference panel to impute missing variation into the sample genotype data (Howie et al., 2009). Recent improvements in imputation algorithms and the expansion of reference datasets have improved accuracy of imputation for even low MAF variants. Imputed data can then be annotated using tools developed for the annotation of sequencing data such as SnpEff (Cingolani et al., 2012) which determine the genomic location (i.e., exonic, intronic or intergenic, and the effects of variants, missense, nonsense etc. on known genes). Imputed LOF variants can then be assessed against binary phenotypes or quantitative laboratory values derived from patients electronic medical records (EMR).

We sought to determine if two PCSK9 LOF mutations that are present in the 1KGP data, the C679X nonsense mutation and the R46L missense mutation, could be imputed into our dataset and the previously reported association of the LOFs with decreased serum LDL-C replicated.

## MATERIALS AND METHODS

The Center for Applied Genomics (CAG) at The Children's Hospital of Philadelphia (CHOP) maintains a biorepository of over 160,000 genotyped samples, 60,000 of which are pediatric samples randomly recruited from CHOP with complete EMRs. As a proof of principle, we imputed the proprotein convertase subtilisin kexin type 9 (*PCSK9*; NM_174936) LOFs C679X (dbSNP:rs28362286) and R46L (dbSNP:rs11591147) into a random selection of 8,028 unrelated samples of Northern European ancestry genotyped on the Illumina HumanHap 550 array from the CAG biorepository. The study was approved by the Institutional Review Board at the CHOP, and written informed consent for sample collection and DNA genotyping/sequencing was provided by the parents of all participating children.

Genetic ancestry was determined by computing principal components on the dataset using smartpca, a part of the EIGENSTRAT package, on 100,000 random autosomal SNPs in linkage equilibrium. Samples were clustered into 4 Continental ancestry groups (Caucasian, African including admixed African–American, Asian, and native American/admixed Hispanic) by K-means clustering using the kmeans package in R. The European ancestry grouping in our dataset mapped most closely to the HapMap CEU population of Utah residents with Northern and Western European ancestry from the CEPH collection[1].

Duplicate samples and cryptic relatedness were assessed by pairwise IBD. IBD values were generated for all 8,028 samples of Northern European ancestry using the plink genome command. A random sample from any pair with a PI_HAT value exceeding 0.3 was excluded from further analysis.

Imputation of untyped markers (∼39 M) was carried out using IMPUTE2 after prephasing with SHAPEIT. Each chromosome was prephased separately. Reference phased cosmopolitan haplotypes and recombination rates were obtained from the 1000 genomes project (1000 Genomes Phase I integrated variant set b37 March 2012 release). Imputation was carried out in 5Mb intervals using an effective population size of 20000 as recommended. As

[1] http://hapmap.ncbi.nlm.nih.gov

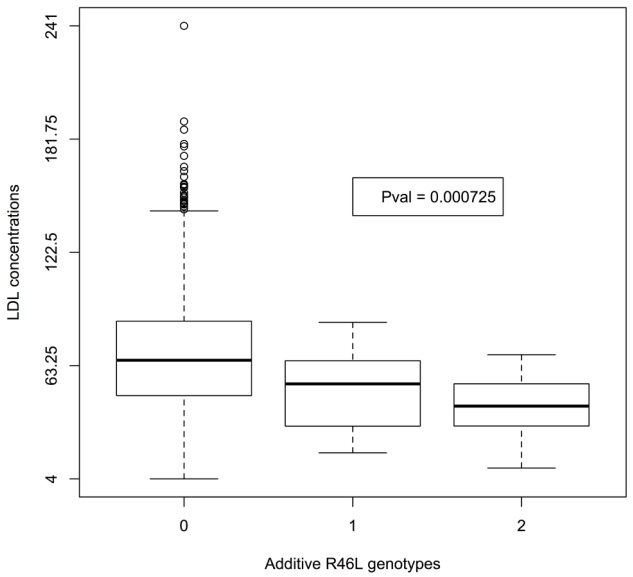


**FIGURE 1 | Box and whisker plot of additively encoded *PCSK9* R46L genotypes on the X-axis versus age corrected serum LDL concentrations on the Y-axis ($n = 1432$; min LDL concentration 4; max LDL concentration 241; mean LDL concentration 85).** Genotype 0 indicates no copies of the mutant allele, genotype 1 indicates one copy of the mutant allele and genotype 2 indicates 2 copies of the mutant allele.

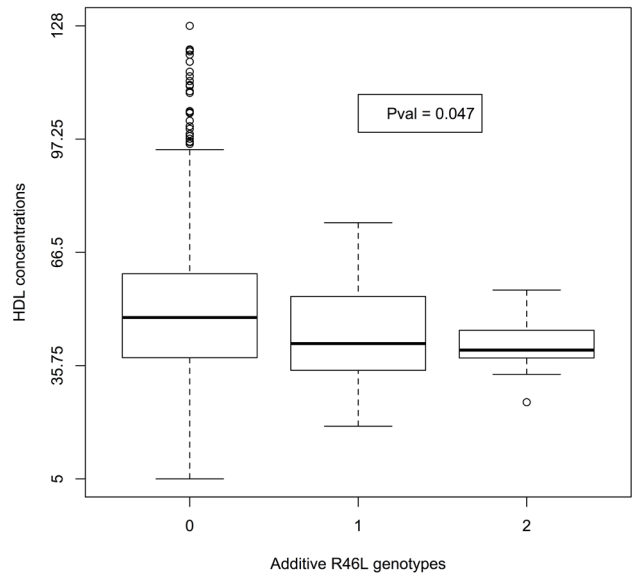


**FIGURE 2 | Box and whisker plot of additively encoded *PCSK9* R46L genotypes on the X-axis versus age corrected serum HDL concentrations on the Y-axis ($n = 1495$; min HDL concentration 5; max HDL concentration 128; mean HDL concentration 47).** Genotype 0 indicates no copies of the mutant allele, genotype 1 indicates one copy of the mutant allele and genotype 2 indicates 2 copies of the mutant allele.

a measure of the overall imputation accuracy we compared the concordance between the imputed and known genotypes in the subset of SNPs for which genotyping data was available. At a call threshold of 0.9, over 99% of the imputed genotypes were called and over 96% of those were concordant with the known genotypes.

## RESULTS

Following imputation using SHAPEIT[2] and IMPUTE2[3] and annotation using SnpEff[4], we extracted and additively re-encoded genotypes for C679X and R46L from the 8,028 European American samples from the CAG biorepository. Both variants were imputed with high confidence, info scores C679X = 0.9 and R46L = 1. The C679X mutation was previously reported to be present in 0.1% of European Americans (Cohen et al., 2005). We identified nine C679X carriers out of 8,028 samples for a frequency of 0.11%, consistent with previous reports. As the samples were randomly selected from the biorepository, not all contained serum lipid data in their EMR. Three of the nine C679X carriers had serum LDL data. The frequency of the R46L was also consistent with the NHLBI ESP data, homozygous wild-type R46L 0.98 (1432 unique individuals with lab values mean age 12.1 years); heterozygous R46L 0.02 (10 unique individuals with lab values mean age 13.5) and homozygous derived allele R46L 0.001 (12 unique individuals with lab values mean age 11.5). A total of twenty-two R46L carriers had LDL data in the EMR.

There was insufficient data to assess the statistical significance of C679X genotypes. Linear regression of EMR-derived age-corrected serum LDL concentrations against R46L genotypes was statistically significant ($P$-value $7 \times 10^{-4}$) and directions of effect consistent with the LOF allele reducing LDL cholesterol (**Figure 1**). Serum HDL concentrations also showed a trend toward association ($P$-value 0.04; **Figure 2**). By contrast, serum triglyceride levels showed no association with R46L genotype ($P$-value 0.58; **Figure 3**) as previously described (Kotowski et al., 2006). The mean age-adjusted LDL concentration for R46 wild-type homozygotes was 85.7, mean age-adjusted LDL concentration for R46L heterozygotes was 63 and 62.6 for R46L homozygotes which corresponds approximately to a 26% decrease of serum LDL consistent with the 23.5 mean LDL-C difference previously reported in European American R46L carriers (Kotowski et al., 2006).

## DISCUSSION

Recent genome sequencing studies have shown that each individual carries a significant number of variants that are predicted to result in a loss of protein function. The phenotypic effect of the majority of these LOFs remains to be determined. Here, we have shown a successful proof of concept that rare LOFs can be imputed into high density genotyping array data using data from large scale sequencing projects such as the 1KGP as a reference. While second generation sequencing remains prohibitively expensive in large numbers, high density genotyping data has been generated on hundreds of thousands of individuals. The eMERGE consortium biorepository includes ∼60,000 individuals

[2]http://www.shapeit.fr

[3]http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

[4]http://snpeff.sourceforge.net

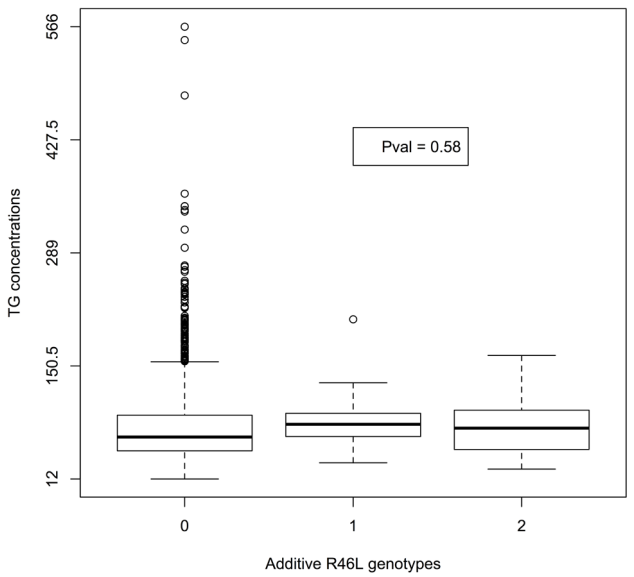


**FIGURE 3 | Box and whisker plot of additively encoded *PCSK9* R46L genotypes on the X-axis versus age corrected serum triglyceride concentrations on the Y-axis ($n = 1856$; min TG concentration 12; max TG concentration 566; mean TG concentration 82).** Genotype 0 indicates no copies of the mutant allele, genotype 1 indicates one copy of the mutant allele and genotype 2 indicates 2 copies of the mutant allele.

that have been genotyped on high-density GWA arrays (review at http://www.genome.gov/27540473), all of which has been linked with EMRs. As such eMERGE would be ideally suited for the assessment of rare LOF variants across multiple phenotypes either by direct assessment through single variant tests or through burden tests. For future analyses, in order to identify all possible association signals, the data would be analyzed using more than one statistical approach as detailed below.

Annotated, imputed variants, in vcf format[5], would be analyzed for association using both single point and agglomerative tests. Single variant tests for association against the EMR traits would be implemented in EMMAX (Kang et al., 2010), a mixed model algorithm that controls for both population substructure and relatedness between individuals in the test. In addition to the principal components for population stratification applicable covariates such as age could be included. For the agglomerative gene-based association tests, three complementary algorithms, the sequence kernel association test (SKAT; Ionita-Laza et al., 2013), the variable threshold test (Price et al., 2010) and the combined multivariate and collapsing (CMC) test which assess the burden of variation within the gene (Li and Leal, 2008) would be implemented. Gene-based association tests can achieve substantial increases in power to detect associations with rare variation compared with single variant tests (Ionita-Laza et al., 2013).

[5]http://www.1000genomes.org/wiki/analysis/variant-call-format/vcf-variant-call-format-version-42

We anticipate that for the single variant tests greatest power would be achieved against quantitative phenotypes such as lab values, however, gene burden scores could equally be applied using a pheWAS approach (Denny et al., 2010), i.e., EMR derived ICD9-based pseudo-case control analyzes for binary traits. These approaches will be validated on multiple LOF variants across the eMERGE networks in the near-future.

## ACKNOWLEDGMENTS

We are grateful to the study volunteers for participating in the research studies and to the clinicians and support staff for enabling patient recruitment and blood sample collection. Informed consent was obtained from all participants or their parents or guardians. Sample genotyping was funded by an Institutional Development Award to the CAG from CHOP; imputation and association analyzes were funded by an eMERGE consortium 1U01HG006830-01 award from the NHGRI.

## REFERENCES

Abifadel, M., Varret, M., Rabès, J. P., Allard, D., Ouguerram, K., Devillers, M., et al. (2003). Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* 34, 154–156. doi: 10.1038/ng1161

Benjannet, S., Rhainds, D., Hamelin, J., Nassoury, N., and Seidah, N. G. (2006). The proprotein convertase (PC) PCSK9 is inactivated by furin and/or PC5/6A: functional consequences of natural mutations and post-translational modifications. *J Biol. Chem.* 281, 30561–30572. doi: 10.1074/jbc.M606495200

Cameron, J., Holla, Ø. L., Ranheim, T., Kulseth, M. A., Berge, K. E., and Leren, T. P. (2006). Effect of mutations in the PCSK9 gene on the cell surface LDL receptors. *Hum. Mol. Genet.* 15, 1551–1558. doi: 10.1093/hmg/ddl077

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695

Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr., and Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* 354, 1264–1272. doi: 10.1056/NEJMoa054013

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., Hobbs, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* 37, 161–165. doi: 10.1038/ng1509

Delaneau, O., Marchini, J., and Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 1205–1210. doi: 10.1093/bioinformatics/btq126

Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* 92, 841–853. doi: 10.1016/j.ajhg.2013.04.015

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kotowski, I. K., Pertsemlidis, A., Luke, A., Cooper, R. S., Vega, G. L., Cohen, J. C., et al. (2006). A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* 78, 410–422. doi: 10.1086/500615

Lambert, G., Jarnoux, A. L., Pineau, T., Pape, O., Chetiveaux, M., Laboisse, C., et al. (2006). Fasting induces hyperlipidemia in mice overexpressing proprotein convertase subtilisin kexin type 9: lack of modulation of very-low-density

lipoprotein hepatic output by the low-density lipoprotein receptor. *Endocrinology* 147, 4985–4995. doi: 10.1210/en.2006-0098

Li, B., and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321. doi: 10.1016/j.ajhg.2008.06.024

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823–828. doi: 10.1126/science.1215040

Marais, D. A., Blom, D. J., Petrides, F., Goueffic, Y., and Lambert, G. (2012). Proprotein convertase subtilisin/kexin type 9 inhibition. *Curr. Opin. Lipidol.* 23, 511–517. doi: 10.1097/MOL.0b013e3283587563

Mayne, J., Dewpura, T., Raymond, A., Bernier, L., Cousins, M., Ooi, T. C., et al. (2011). Novel loss-of-function PCSK9 variant is associated with low plasma LDL cholesterol in a French-Canadian family and with impaired processing and secretion in cell culture. *Clin. Chem.* 57, 1415–1423. doi: 10.1373/clinchem.2011.165191

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J., et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838. doi: 10.1016/j.ajhg.2010.04.005 +

Stein, E. A., Olsson, A. G., Scott, R., Kim, J. B., Xue, A., Gebski, V., et al. (2012). Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N. Engl. J. Med.* 366, 1108–1118. doi: 10.1056/NEJMoa1105803

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. doi: 10.1126/science.1219240

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 14 January 2014; accepted: 10 April 2014; published online: 29 April 2014.*
*Citation: Sleiman P, Bradfield J, Mentch F, Almoguera B, Connolly J and Hakonarson H (2014) Assessing the functional consequence of loss of function variants using electronic medical record and large-scale genomics consortium efforts. Front. Genet. **5**:105. doi: 10.3389/fgene.2014.00105*
*This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.*