Published in final edited form as:
Circulation. 2010 November 16; 122(20): 2016–2021. doi:10.1161/CIRCULATIONAHA.110.948828.

Identification of Genomic Predictors of Atrioventricular Conduction:

Using Electronic Medical Records as a Tool for Genome Science

Joshua C. Denny, MD, MS[*], **Marylyn D. Ritchie, PhD**[*], **Dana C. Crawford, PhD**, **Jonathan S. Schildcrout, PhD**, **Andrea H. Ramirez, MD**, **Jill M. Pulley, MBA**, **Melissa A. Basford, MBA**, **Daniel R. Masys, MD**, **Jonathan L. Haines, PhD**, and **Dan M. Roden, MD**
Office of Personalized Medicine and the Center for Human Genetics Research, and the Departments of Biomedical Informatics, Molecular Physiology and Biophysics, Biostatistics, Pharmacology, and Medicine Vanderbilt University School of Medicine, Nashville

Abstract

Background—Recent genome-wide association studies (GWAS) using selected community populations have identified genomic signals in *SCN10A* influencing PR duration. The extent to which this can be demonstrated in cohorts derived from electronic medical records is unknown.

Methods and Results—We performed a GWAS on 2,334 European-American patients with normal ECGs without evidence of prior heart disease from the Vanderbilt DNA databank, BioVU, which accrues subjects from routine patient care. Subjects were identified using combinations of natural language processing, laboratory, and billing code queries of de-identified medical record data. Subjects were 58% female, mean (±SD) age 54±15 years, and had mean PR intervals of 158±18 milliseconds. Genotyping was performed using the Illumina Human660W-Quad platform. Our results identify four single nucleotide polymorphisms (rs6800541, rs6795970, rs6798015, rs7430477) linked to *SCN10A* associated with PR interval ($p=5.73\times10^{-7}$ to 1.78×10^{-6}).

Conclusions—This GWAS confirms a gene heretofore-unimplicated in cardiac pathophysiology as a modulator of PR interval in humans. This study is one of the first replication GWAS performed using an electronic medical record-derived cohort, supporting their further use for genotype-phenotype analyses.

Keywords

electronic medical records; atrioventricular conduction; genome-wide association study; natural language processing

Introduction

Widely-used electrocardiographic (ECG) parameters, such as PR and QT intervals, display substantial variability when measured across large populations.[1,2] Some of this variability reflects underlying disease and concomitant drug therapy, but even after the correction for these, such variability persists. Thus, for example, we have recently reported that the 99%

Correspondence: Dan M. Roden, M.D., Professor of Medicine and Pharmacology, Director, Oates Institute for Experimental Therapeutics, Assistant Vice-Chancellor for Personalized Medicine Vanderbilt University School of Medicine, 1285 Medical Research Building IV, Nashville, TN 37232-0575, Telephone: 615-322-0067, Fax: 615-343-4522, dan.roden@vanderbilt.edu.
[*]contributed equally to this work

Conflict of Interest Disclosures
The authors have no conflicts of interest.

confidence interval for PR was 120–206 ms in a set of over 30,000 electrocardiograms studied in normal individuals in our electronic medical record (EMR).[3]

One extensively-studied electrocardiographic measurement is the QT interval. Variability in the QT interval is a marker for sudden cardiac death not only in patients with congenital and drug-associated long QT syndromes,[4,5] but also in more general settings, notably after myocardial infarction.[6] An initial genome-wide association study (GWAS) of variability in the QT interval across large populations identified variants at chromosome 1 near *NOS1AP* as contributors.[7] Subsequently, with accrual of larger numbers of subjects and meta-analyses, multiple loci at which genomic variants contribute to variability in QT interval have been identified.[8,9] Some of these are located in genes well known to modulate cardiac repolarization, such as those encoding cardiac ion channels, while the relationship of others to normal QT intervals is less clear. Thus, genome-wide approaches can uncover new pathways in pathophysiology. Indeed, variants in *NOS1AP* have now been implicated as predictors of sudden cardiac death in large populations.[10,11]

Most recently, three groups reported the results of GWAS evaluating atrioventricular conduction (the PR interval).[12–14] Multiple loci were implicated, and all 3 studies converged on variants in *SCN10A.* Interestingly, *SCN10A* is located on chromosome 3 adjacent to *SCN5A*, which encodes the major cardiac sodium channel gene in heart, but *SCN10A* had not, to date, been implicated as a modulator of cardiovascular physiology.

These results were obtained by analysis of large cohorts of patients, recruited from communities or specific clinics for epidemiologic study. We and others[15,16] have proposed an alternate strategy, in which DNA repositories are linked to electronic medical records (EMRs) across large healthcare systems. Theoretical advantages of this paradigm are rapid generation of patient sets for study (since electronic data are already in place), and the ability to study large numbers of subjects accrued without bias with respect to factors such as disease or age. However, despite these theoretical advantages, the utility of EMR-based approaches for validation or discovery of genotype-phenotype associations in populations remains largely unexplored.

The National Human Genome Research Institute's electronic MEdical Records and GEnomics (eMERGE) Network has as one of its primary goals, the evaluation of the utility of EMR systems coupled to DNA repositories as a tool for genome science. We report here the development of algorithms to identify subjects with normal PR intervals, lacking heart disease. A subset of this population has been accrued into BioVU, the Vanderbilt DNA databank,[15] and has undergone eMERGE-supported genome wide genotyping, allowing us to evaluate the extent to which recently reported PR signals can be replicated in an EMR environment.

Methods

BioVU design

The design of BioVU, which links DNA extracted from discarded blood samples to a de-identified image of the EMR, called the Synthetic Derivative, has been previously described.[15] BioVU accrues DNA samples extracted from blood remaining from routine clinical testing after they have been retained for three days and are scheduled to be discarded. The resource contains data and tissues that are de-identified in accordance with provisions of 45 CFR 46 that define criteria for investigations that are considered 'non-human subjects' research. Exclusion criteria are poor quality or insufficient DNA; age <18 years; absence of a signed consent to treatment form; an individual who has opted out; and duplicate samples. In addition, a small percentage of patients (~2%) is randomly excluded from BioVU so that

it is not possible to know whether any individual's sample is or is not included in the biobank. The project has been reviewed at multiple levels, including the Institutional Review Board, internal and external ethics committees, Community Advisory Board, legal department, and the federal Office of Human Research Protection, and this oversight is ongoing.[17]

As of May 10, 2010, BioVU included samples from 84,540 individuals. Preliminary studies genotyped the initial 10,000 subjects accrued at ~100 single nucleotide polymorphisms (SNPs) previously associated with common diseases; natural language processing (NLP) techniques were developed and validated to identify cases and controls for five common diseases, and previously-reported genotype-phenotype associations were replicated.[18] In the present report, we developed and validated NLP approaches to identify subjects with normal ECGs and no evidence of heart disease at the time of the ECG. A subset of these records is linked to eMERGE-supported Illumina 660W-Quad genotyping, and the results of the PR analysis are presented here.

Identifying the case population

Subjects were selected by analyzing records in the Synthetic Derivative, which is a de-identified image of the EMR linked to BioVU by research unique identifiers generated by hashing the medical record number.[15] The Vanderbilt EMR began accumulating data in the early 1990s and now includes all inpatient and outpatient billing codes, laboratory values, reports, and clinical documentation, almost all in electronic formats available for searching. It currently contains over 120 million documents on about 2 million patients. The Synthetic Derivative is refreshed regularly to add new clinical information from the EMR as it is accrued.

The study population consisted of subjects with a normal ECG without evidence of cardiac disease before or within one month following the ECG, concurrent use of medications that interfere with ventricular conduction, and who did not have abnormal potassium, calcium, or magnesium lab values at the time of the ECG. The algorithm combined NLP[19,20] to analyze unstructured text, billing code queries, and lab queries to exclude any subjects with evidence of arrhythmia, heart failure, cardiomyopathy, myocardial ischemia/infarct, or cardiac conduction defect. The algorithm considered all physician-generated clinical documentation, including clinical notes and cardiologist-generated ECG impressions. Patients with family histories of cardiac disease or allergic to restricted medications were allowed by the NLP algorithm. In addition, the algorithm selected only patients whose ECGs had normal Bazett's corrected QT intervals (<450ms), heart rates (between 50–100 bpm), and QRS (65–120 ms). To ensure sufficient record size, we required each subject to have at least one clinical note. The study included only individuals designated as "non-Hispanic white" European American; we have shown a high degree of concordance between EMR ancestry designation and genetic ancestry determined from ancestry-informative markers in BioVU.[21,22] We noted the presence or absence of concurrent medications that could affect atrioventricular conduction, including digoxin, non-dihydropyridine calcium channel blockers, tricyclic antidepressants, and beta-blockers. Measurements for the PR interval were taken from that stored in the EMR as produced by the ECG machine. All ECGs were processed using the Philips TraceMaster ECG software (Andover, MA).

Two physicians not associated with algorithm development reviewed algorithm results with access to the entire de-identified medical records of the patients. The results of the manual classification were then used to improve the algorithms, and the procedure was iterated until the positive predictive value reached the pre-designated target of ≥95% for a random selection of 100 subjects. In each iteration, the physicians reviewed a different set of 100 randomly selected subjects. The time between generation of the initial algorithm and its

finalization was about six weeks; identification of the final study set using the algorithm took six computational hours. Complete details of the algorithm are available from http://gwas.net.

Genotyping and data analysis

Genotyping was performed at the Center for Genotyping and Analysis at the Broad Institute, using the Illumina Human660W-Quadv1_A genotyping platform, consisting of 561,490 SNPs and 95,876 intensity-only probes. Data were cleaned using the quality control (QC) pipeline developed by the eMERGE Genomics Working Group. This process includes evaluation of sample and marker call rate, gender mismatch and anomalies, duplicate and HapMap concordance, batch effects, Hardy-Weinberg equilibrium, sample relatedness, and population stratification. Relatedness was determined based on IBD estimates generated from the genome-wide genotype data in PLINK. Individuals with an IBD estimate of greater than 0.0625 were identified (4 sib-pairs, 7 parent-child pairs, and 1 trio); one individual from each related pair and the child from the trio were removed from the analysis (12 total individuals removed).

After QC, 514,999 SNPs were used for analysis based on the following QC criteria: SNP call rate >99%, sample call rate >99%, minor allele frequency > 0.01, 99.99% concordance rate in duplicates, unrelated samples only, and individuals of European-descent only (based on STRUCTURE[23] analysis of >90% probability of being in the CEU cluster); our post-QC genomic control was $\lambda = 1.011$. We did not filter SNPs based on deviation from HWE as this can be an indication of either true association or genotyping error. As such, we flagged all markers with HWE $p<1\times10^{-4}$ for further evaluation post-analysis. All genotype data and a detailed QC report have been uploaded to dbGaP. The QC and data analysis was performed using a combination of PLINK and R.

Single-locus tests of association were performed using linear regression assuming an additive genetic model for all 514,999 SNPs in a total of 2,334 individuals with a measured PR interval. All analyses were performed unadjusted and adjusted for age, sex, exposure to medications that may alter the PR interval, and the first three principle components from Eigenstrat to adjust for potential population stratification.

Results

Our algorithm identified 2,334 European American individuals with genotyping data that passed QC (Table 1). The positive predictive value for the final algorithm was 97%, including absence of all exclusion criteria. Patients were a median of 55 years old (interquartile range 44–64, range 19–98), and had a median PR interval of 156 (interquartile range 145–170 milliseconds, range 112–224 milliseconds). Table 2 shows the results of linear regression analyses (including age, sex, PR-active drugs, and the first three principle components from Eigenstrat) that examined replication the three published GWAS[12–14] of PR interval. The top SNPs for each of these 3 published analyses (14 total SNPs combined) are presented and the corresponding P values in the present analysis; 6/14 of these had P values <0.02 (all SNPs had similar MAF, effect size, and effect direction compared to the associations in the literature; data not shown). In the full adjusted linear regression analysis, we identified four SNPs in LD with one another in the gene *SCN10A* that are associated with PR interval (Figure 1): intronic rs7430477 ($p=1.78\times10^{-6}$), missense (V1073A) rs6795970 ($p=7.26\times10^{-7}$), intronic rs6800541 ($p=5.73\times10^{-7}$), intronic rs6798015 ($p=1.36\times10^{-6}$). *SCN10A* was also associated in the unadjusted analysis (top SNP P value 5.18×10^{-7}). The effect size (beta) is reported in milliseconds (ms) per copy of the minor allele. For the primary adjusted analysis, the top SNP (rs6800541) has a 2.517 ms change in PR interval per copy of the minor allele. This is slightly smaller than the findings of Pfeufer et al.[12]

where beta is 3.7687 ms, as expected from the "winner's curse".[24] The percentage of variance in the PR interval explained by each of the four *SCN10A* SNPs above was about 1% in the unadjusted model. The adjusted model, including SNP, age, sex, use of PR-active medications, and the first three principle components, explained about 11% of the variance in PR duration.

Table 3 lists all SNPs with a P value < 10^{-5} in the adjusted and unadjusted analyses including all subjects. Two of these (rs6795970, rs6800541) were also identified in the 3 previously-reported analyses.[12–14]

Discussion

We demonstrate here that a population identified by interrogation of a de-identified version of an electronic medical record system which has been organized for research purposes can be used to efficiently identify genomic determinants of PR interval. Although the number of subjects studied is relatively small compared to those in the recent GWAS,[12–14] the *SCN10A* signal, on which all three previous studies converged, is also seen in this analysis at significance levels of 10^{-6} to 10^{-7}. Subsequently, three additional loci implicated in the published PR GWAS replicated in this dataset at P<0.02. Although these do not meet a conventional definition of genome-wide significance, the significance threshold in a replication study is often relaxed especially when the replication sample size is significantly smaller than the discovery cohort. Based on the sample size in the published literature (10,000 individuals),[13] using the effect size and allele frequency in BioVU, we achieve 99.99% power to detect the same effect at $p<5\times10^{-8}$ in *SCN10A*, *ARGHGAP24*, *CAV1-CAV2*, and *TBX5*. Thus, we expect that if we had 10,000 individuals in the BioVU set using EMR-derived PR duration, we would have identified nearly all PR-associated variants that were previously discovered in population-based or community-based cohorts.

The major result was similar irrespective of whether age, sex, concomitant meds, and principle component adjustments were made. It was also similar in analyses that excluded patients on medications known to influence atrioventricular conduction. The latter result may seem surprising in light of the fact that PR intervals were longer in those on versus off medications (163±19 msec versus 157±17 msec); however, this can be explained by noting that the relationships between PR predictive polymorphisms and medication usage were very weak. That is, in these data, medication usage does not mediate the SNP - PR interval relationship in an obvious way. Accordingly, in cases such as these, analyses that include medicated patients for GWAS will often be more powerful than those that do not.

In this study, we used NLP approaches to identify a study set lacking potential confounders such as heart disease that would increase experimental noise. Despite this inclusion, we were readily able to identify sufficient samples for study, because the pool from which the samples were drawn is very large. Using electronic records in genomic research allows study of large populations without overt or hidden biases as to inclusion criteria that are often part of recruitment into clinical trials. Population-based epidemiologic cohorts lack such bias, but may have less data with respect to long-term outcomes across all disciplines. Nevertheless the attraction of EMRs as research tools in genomic science is only now beginning to be explored. In a previous study, we showed that well-studied SNPs associated with atrial fibrillation, rheumatoid arthritis, Crohn's disease, multiple sclerosis, and type 2 diabetes could be replicated in the initial 10,000 subjects accrued into BioVU.[18] The generalizability of any result such as this, from an EMR or other source such as a community cohort or a clinical trial, may be limited if the recruitment is in some fashion biased. The fact that we identify a locus implicated in recent GWA studies speaks further to the generalizability of this result.

One obstacle to this use of EMRs is the need to develop and validate algorithmic approaches to identify study subjects and controls, if needed. This study required a combined approach of NLP, lab queries, identification of prescribed medications, and billing code queries to achieve a high positive predictive value while capturing a sufficiently large case population. The availability of large EMR records, with long follow-up (a median of five years in BioVU), offers potential for reuse of genotype data for subsequent analyses of other phenotypes. Developing sets of validated phenotype selection algorithms and assessing the extent to which these apply across EMR systems is one goal of the eMERGE network.

This GWAS and previous ones identify a gene heretofore-unimplicated in cardiac pathophysiology as a modulator of PR interval in humans. Further studies are required to identify the role of *SCN10A* variants in normal and abnormal atrioventricular nodal function. Other genetic loci identified in previous studies, and this one, await further replication studies.

Three recent genome-wide association studies (GWAS) conducted in community populations have identified multiple loci contributing to variability in normal PR duration. This study sought to replicate these findings by performing a GWAS in subjects identified in BioVU, the Vanderbilt DNA databank that links accrues DNA from blood samples obtained routine patient careand links these linked to a de-identified copy of the electronic medical record (EMR). Subjects were identified using combinations of natural language processing, laboratory, and billing code queries of medical record data.The GWAS was conducted in 2,334 European-American patients with normal ECGs and no evidence of prior heart diseasein the EMR, and confirmed a signal in a sodium channel gene, SCN10A, previously-unimplicated in cardiac pathophysiology as a modulator of PR interval in humans. This study is one of the first to validate use of an EMR-derived cohort of genome-wide analysis, supporting their further use forgenotype-phenotype analysis.

Acknowledgments

Source of Funding: This study was funded by U01 HG04603, “VGER: Vanderbilt Genome-Electronic Records Project”, a node in the National Human Genome Research Institute-supported eMERGE (Electronic Records and Genomics) Network

References

1. Hiss RG, Lamb LE, Allen MF. Electrocardiographic findings in 67,375 asymptomatic subjects. X. Normal values. Am J Cardiol. 1960; 6:200–231. [PubMed: 13855921]
2. Mason JW, Ramseth DJ, Chanter DO, Moon TE, Goodman DB, Mendzelevski B. Electrocardiographic reference ranges derived from 79,743 ambulatory subjects. J Electrocardiol. 2007; 40:228–234. [PubMed: 17276451]
3. Havens A, Schildcrout J, Masys D, Weinter J, Pulley J, Basford M, Roden D, Denny J. Abstract 2684: Modulators of Normal ECG Intervals Identified in a large Electronic Medical Record. Circulation. 2009; 120:S679.
4. Priori SG, Schwartz PJ, Napolitano C, Bloise R, Ronchetti E, Grillo M, Vicentini A, Spazzolini C, Nastoli J, Bottelli G, Folli R, Cappelletti D. Risk Stratification in the Long-QT Syndrome. N Engl J Med. 2003; 348:1866. [PubMed: 12736279]
5. Roden DM. Drug-induced prolongation of the QT Interval. N Engl J Med. 2004; 350:1013–1022. [PubMed: 14999113]
6. Schwartz PJ, Wolf S. QT interval prolongation as a predictor of sudden death in patients with myocardial infarction. Circulation. 1978; 56:1074–1077. [PubMed: 639227]

7. Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, Ikeda M, West K, Kashuk C, Akyol M, Perz S, Jalilzadeh S, Illig T, Gieger C, Guo CY, Larson MG, Wichmann HE, Marban E, O'Donnell CJ, Hirschhorn JN, Kaab S, Spooner PM, Meitinger T, Chakravarti A. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet. 2006; 38:644–651. [PubMed: 16648850]

8. Pfeufer A, Sanna S, Arking DE, Muller M, Gateva V, Fuchsberger C, Ehret GB, Orru M, Pattaro C, Kottgen A, Perz S, Usala G, Barbalic M, Li M, Putz B, Scuteri A, Prineas RJ, Sinner MF, Gieger C, Najjar SS, Kao WHL, Muhleisen TW, Dei M, Happle C, Mohlenkamp S, Crisponi L, Erbel R, Jockel KH, Naitza S, Steinbeck G, Marroni F, Hicks AA, Lakatta E, Muller-Myhsok B, Pramstaller PP, Wichmann HE, Schlessinger D, Boerwinkle E, Meitinger T, Uda M, Coresh J, Kaab S, Abecasis GR, Chakravarti A. Common variants at ten loci modulate the QT interval duration in the QTSCD Study. Nat Genet. 2009; 41:407–14. [PubMed: 19305409]

9. Newton-Cheh C, Eijgelsheim M, Rice KM, de Bakker PIW, Yin X, Estrada K, Bis JC, Marciante K, Rivadeneira F, Noseworthy PA, Sotoodehnia N, Smith NL, Rotter JI, Kors JA, Witteman JCM, Hofman A, Heckbert SR, O'Donnell CJ, Uitterlinden AG, Psaty BM, Lumley T, Larson MG, Ch Stricker BH. Common variants at ten loci influence QT interval duration in the QTGEN Study. Nat Genet. 2009; 41:399–406. [PubMed: 19305408]

10. Kao WH, Arking DE, Post W, Rea TD, Sotoodehnia N, Prineas RJ, Bishe B, Doan BQ, Boerwinkle E, Psaty BM, Tomaselli GF, Coresh J, Siscovick DS, Marban E, Spooner PM, Burke GL, Chakravarti A. Genetic Variations in Nitric Oxide Synthase 1 Adaptor Protein Are Associated With Sudden Cardiac Death in US White Community-Based Populations. Circulation. 2009; 119:940–51. [PubMed: 19204306]

11. Eijgelsheim M, Newton-Cheh C, Aarnoudse AL, van NC, Witteman JC, Hofman A, Uitterlinden AG, Stricker BH. Genetic variation in NOS1AP is associated with sudden cardiac death: evidence from the Rotterdam Study. Hum Mol Genet. 2009; 18:4213–4218. [PubMed: 19643915]

12. Pfeufer A, van Noord C, Marciante KD, Arking DE, Larson MG, Smith AV, Tarasov KV, Muller M, Sotoodehnia N, Sinner MF, Verwoert GC, Li M, Kao WHL, Kottgen A, Coresh J, Bis JC, Psaty BM, Rice K, Rotter JI, Rivadeneira F, Hofman A, Kors JA, Stricker BHC, Uitterlinden AG, van Duijn CM, Beckmann BM, Sauter W, Gieger C, Lubitz SA, Newton-Cheh C, Wang TJ, Magnani JW, Schnabel RB, Chung MK, Barnard J, Smith JD, Van Wagoner DR, Vasan RS, Aspelund T, Eiriksdottir G, Harris TB, Launer LJ, Najjar SS, Lakatta E, Schlessinger D, Uda M, Abecasis GR, Muller-Myhsok B, Ehret GB, Boerwinkle E, Chakravarti A, Soliman EZ, Lunetta KL, Perz S, Wichmann HE, Meitinger T, Levy D, Gudnason V, Ellinor PT, Sanna S, Kaab S, Witteman JCM, Alonso A, Benjamin EJ, Heckbert SR. Genome-wide association study of PR interval. Nat Genet. 2010; 42:153–159. [PubMed: 20062060]

13. Holm H, Gudbjartsson DF, Arnar DO, Thorleifsson G, Thorgeirsson G, Stefansdottir H, Gudjonsson SA, Jonasdottir A, Mathiesen EB, Njolstad I, Nyrnes A, Wilsgaard T, Hald EM, Hveem K, Stoltenberg C, Lochen ML, Kong A, Thorsteinsdottir U, Stefansson K. Several common variants modulate heart rate, PR interval and QRS duration. Nat Genet. 2010; 42:117–122. [PubMed: 20062063]

14. Chambers JC, Zhao J, Terracciano CMN, Bezzina CR, Zhang W, Kaba R, Navaratnarajah M, Lotlikar A, Sehmi JS, Kooner MK, Deng G, Siedlecka U, Parasramka S, El-Hamamsy I, Wass MN, Dekker LRC, de Jong JSSG, Sternberg MJE, McKenna W, Severs NJ, de Silva R, Wilde AAM, Anand P, Yacoub M, Scott J, Elliott P, Wood JN, Kooner JS. Genetic variation in SCN10A influences cardiac conduction. Nat Genet. 2010; 42:149–152. [PubMed: 20062061]

15. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008; 84:362–369. [PubMed: 18500243]

16. Ginsburg GS, Burke TW, Febbo P. Centralized biorepositories for genetic and genomic research. JAMA. 2008; 299:1359–1361. [PubMed: 18349099]

17. Pulley JM, Clayton EW, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. Clin Transl Sci. 2009; 2:180–2. [PubMed: 20443890]

18. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, Haines JL, Roden DM. Robust replication of genotype-

phenotype associations across multiple diseases in an Electronic Medical Record. Am J Hum Genet. 2010; 86:560–72. [PubMed: 20362271]

19. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. Int J Med Inform. 2009; 78 (Suppl 1):S34–42. [PubMed: 18938105]
20. Denny JC, Spickard A III, Johnson KB, Peterson PA, Peterson JF, Miller RA. Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents. J Am Med Inform Assoc. 2009; 16:806–15. [PubMed: 19717800]
21. Ritchie MD, Dumitrescu L, Brown-Gentry K, Pulley J, Basford M, Denny JC, Masys D, Roden DM, Haines JL. Assessing the accuracy of ancestry reported in a biorepository linked to electronic medical records for genetic association studies. American Society for Human Genetics annual meeting. 2009
22. Dumitrescu L, Ritchie MD, Brown-Gentry K, Pulley JM, Basford M, Denny JC, Oksenberg JR, Roden DM, Haines JL, Crawford DC. Assessing the accuracy of reported ancestry in a biorepository linked to electronic medical records. Genet Med Genet Med. 2010 [Epub ahead of print].
23. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]
24. Kraft P. Curses--winner's and otherwise--in genetic epidemiology. Epidemiology. 2008; 19:649–651. [PubMed: 18703928]

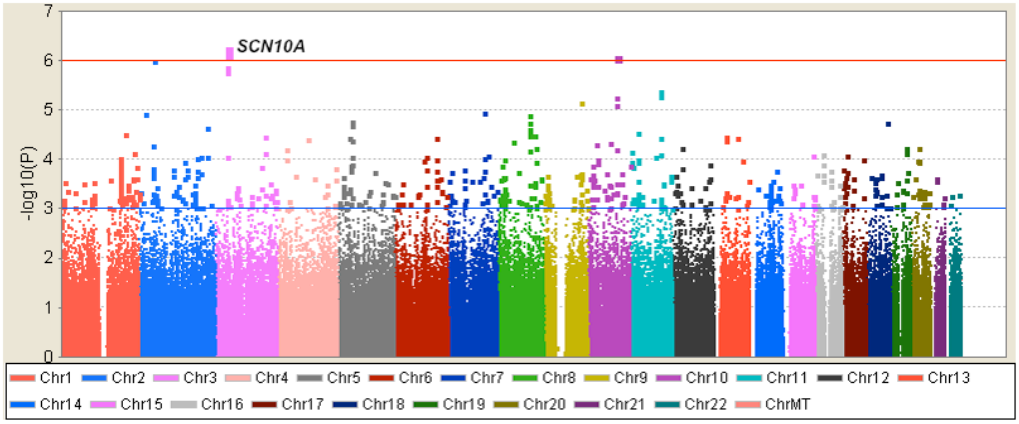

Figure 1.
Manhattan plot for PR interval adjusted for age, sex, and concomitant drug therapy, and first three principle components.

Table 1

Characteristics of subjects with normal ECGs

	Male	Female	All	P
N (%)	991 (42%)	1343 (58%)	2334	
Age (mean ± SD)	55 ± 14	53 ± 16	54 ± 15	0.001
Mean PR interval (mean ± SD)	161 ± 18	156 ± 18	158 ± 18	<0.0001
Number with PR altering medications (see text)	191	268	459	
Years of follow-up (median, inter-quartile range)	5.9 (2.3–9.8)	3.8 (1.0–8.5)	5.0 (1.5–9.2)	

Table 2

Effect estimates in the present study* of SNPs reported in previous GWAS

CHR	SNP	Gene	Minor Allele	MAF	BETA (ms)	SE (ms)	L95	U95	P	Publication
2	rs11897119	*MEIS1*	C	0.400	0.840	0.520	−0.179	1.860	1.06×10^{-1}	Pfeufer[12]
3	rs11708996†	*SCN5A*	-	-	-	-	-	-	-	Pfeufer[12]
3	rs6795970	*SCN10A*	A	0.407	2.505	0.504	1.517	3.493	7.26×10^{-7}	Holm[13], Chambers[14]
3	rs6800541	*SCN10A*	C	0.414	2.517	0.502	1.533	3.500	5.73×10^{-7}	Pfeufer[12], Chambers[14]
3	rs6599257	*SCN10A*	C	0.320	1.674	0.538	0.619	2.729	1.90×10^{-3}	Chambers[14]
4	rs7660702	*ARHGAP24*	C	0.289	−1.326	0.558	−2.419	−0.233	1.75×10^{-2}	Holm[13]
4	rs7692808†	*ARHGAP24*	-	-	-	-	-	-	-	Pfeufer[12]
5	rs251253†	*NKX2-5*	-	-	-	-	-	-	-	Pfeufer[12]
7	rs3807989	*CAV1-CAV2*	A	0.418	1.969	0.504	0.981	2.957	9.64×10^{-5}	Pfeufer[12], Holm[13]
11	rs4944092	*WNT11*	G	0.329	−0.037	0.531	−1.078	1.004	9.44×10^{-1}	Pfeufer[12]
12	rs11047543†	*SOX5*	-	-	-	-	-	-	-	Pfeufer[12]
12	rs1896312†	*TBX5-TBX3*	-	-	-	-	-	-	-	Pfeufer[12]
12	rs3825214	*TBX5*	G	0.189	1.481	0.634	0.238	2.724	1.96×10^{-2}	Holm[13]
14	rs365990	*MYH6*	G	0.367	−0.590	0.513	−1.596	0.416	2.50×10^{-1}	Holm[13]

*All analyses for this study are calculated by linear regression adjusted for age, sex, use of PR-active medications, and the first three principle components from Eigenstrat.

†SNP not present in our dataset.

CHR: chromosome

SNP: Single nucleotide polymorphism

MAF: Minor allele frequency

BETA: coefficient in ms

SE: standard error for BETA

L95: lower 95% confidence interval for BETA

U95: upper 95% confidence interval for BETA

Table 3a

SNPs with $P<10^{-5}$ in the adjusted analysis*

CHR	SNP	Gene	Allele	MAF	BETA (ms)	SE (ms)	L95	U95	P value
3	rs6800541	*SCN10A*	C	0.414	2.517	0.502	1.533	3.500	5.73×10^{-7}
3	rs6795970	*SCN10A*	A	0.407	2.505	0.504	1.517	3.493	7.26×10^{-7}
10	rs1937332	*Intergenic*[1]	G	0.466	2.453	0.497	1.479	3.426	8.45×10^{-7}
2	rs17034876	*Intergenic*[2]	C	0.292	2.707	0.552	1.625	3.788	1.00×10^{-6}
3	rs6798015	*SCN10A*	C	0.380	2.485	0.513	1.479	3.491	1.36×10^{-6}
3	rs7430477	*SCN10A*	T	0.473	−2.361	0.493	−3.328	−1.395	1.78×10^{-6}
11	rs1893057	*Intergenic*[3]	G	0.299	2.483	0.538	1.428	3.538	4.21×10^{-6}
11	rs1940163	*MAML2*	A	0.321	2.406	0.527	1.374	3.439	5.22×10^{-6}
10	rs2251050	*Intergenic*[1]	T	0.373	−2.316	0.510	−3.315	−1.317	5.80×10^{-6}
9	rs1889318	*ASTN2*	C	0.428	2.239	0.497	1.265	3.212	6.94×10^{-6}
10	rs1317361	*Intergenic*[1]	G	0.418	−2.233	0.499	−3.210	−1.256	7.84×10^{-6}

*adjusted for age, sex, use of PR-active medications, and the first three principle components from Eigenstrat.

Abbreviations as in Table 2.

[1]between FAS/ CH25H

[2]between PRKCE/ EPAS1

[3]between MTMR2/ MAML2

[4]*between C18orf20* and *LOC643448*

[5]*between RNF19A* and *ANKRD46*

Circulation. Author manuscript; available in PMC 2011 November 16.

Table 3b

SNPs with $P<10^{-5}$ in the unadjusted analysis

CHR	SNP	Gene	Allele	MAF	BETA (ms)	SE (ms)	L95	U95	P value
3	rs6800541	*SCN10A*	C	0.414	2.660	0.528	1.624	3.695	5.18×10^{-7}
10	rs1937332	*Intergenic*[1]	G	0.466	2.600	0.523	1.575	3.624	6.99×10^{-7}
3	rs6795970	*SCN10A*	A	0.407	2.637	0.531	1.597	3.677	7.25×10^{-7}
3	rs7430477	*SCN10A*	T	0.473	−2.526	0.519	−3.542	−1.509	1.19×10^{-6}
3	rs6798015	*SCN10A*	C	0.380	2.593	0.540	1.534	3.652	1.69×10^{-6}
18	rs470490	*Intergenic*[4]	G	0.391	2.450	0.533	1.405	3.496	4.59×10^{-6}
10	rs2251050	*Intergenic*[1]	T	0.373	−2.449	0.536	−3.500	−1.398	5.19×10^{-6}
2	rs7602460	-	A	0.396	−2.420	0.535	−3.469	−1.371	6.45×10^{-6}
10	rs1317361	*Intergenic*[1]	G	0.418	−2.342	0.524	−3.370	−1.315	8.22×10^{-6}
8	rs1371867	*Intergenic*[5]	C	0.421	2.321	0.523	1.297	3.345	9.33×10^{-6}

[*] adjusted for age, sex, use of PR-active medications, and the first three principle components from Eigenstrat.

Abbreviations as in Table 2.

[1] between FAS/ CH25H

[2] between PRKCE/ EPAS1

[3] between MTMR2/ MAML2

[4] *between C18orf20* and *LOC643448*

[5] *between RNF19A* and *ANKRD46*